KOCAELİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ
ANABİLİM DALI

DOKTORA TEZİ

AKADEMİK MAKALELERDE ANAHTAR KELİME ÇIKARIMI İÇİN YENİ YAKLAŞIMLAR

FURKAN GÖZ

KOCAELİ 2022

KOCAELİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ

ANABİLİM DALI

DOKTORA TEZİ

AKADEMİK MAKALELERDE ANAHTAR KELİME ÇIKARIMI İÇİN YENİ YAKLAŞIMLAR

FURKAN GÖZ

Dr. Öğr. Üyesi Alev MUTLU		
Danışman,	**Kocaeli Üniversitesi**	**.....**
Prof. Dr. Sevinç İLHAN OMURCA		
Jüri Üyesi,	**Kocaeli Üniversitesi**	**.....**
Dr. Öğr. Üyesi Burcu YILMAZ		
Jüri Üyesi,	**Gebze Teknik Üniversitesi**	**.....**
Prof. Dr. Yaşar BECERİKLİ		
Jüri Üyesi,	**Kocaeli Üniversitesi**	**.....**
Prof. Dr. Pınar KARAGÖZ		
Jüri Üyesi,	**Orta Doğu Teknik Üniversitesi**	**.....**

Tezin Savunulduğu Tarih: 02.12.2022

ETİK BEYAN VE ARAŞTIRMA FONU DESTEĞİ

Kocaeli Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırladığım bu tez/proje çalışmasında,

- Bu tezin/projenin bana ait, özgün bir çalışma olduğunu,
- Çalışmamın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamasında bilimsel etik ilke ve kurallara uygun davrandığımı,
- Bu çalışma kapsamında elde edilen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi,
- Bu çalışmamın Kocaeli Üniversitesi'nin abone olduğu intihal yazılım programı kullanılarak Fen Bilimleri Enstitüsü'nün belirlemiş ölçütlere uygun olduğunu,
- Kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- Tezin/projenin herhangi bir bölümünü bu üniversite veya başka bir üniversitede başka bir tez/proje çalışması olarak sunmadığımı,

beyan ederim.

☒ Bu tez/proje çalışmasının herhangi bir aşaması hiçbir kurum/kuruluş tarafından maddi/alt yapı desteği ile desteklenmemiştir.

☐ Bu tez/proje çalışması kapsamında üretilen veri ve bilgiler ..... tarafından ..... no'lu proje kapsamında maddi/alt yapı desteği alınarak gerçekleştirilmiştir.

Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçları kabul ettiğimi bildiririm.

Furkan GÖZ

YAYIMLAMA VE FİKRİ MÜLKİYET HAKLARI

Fen Bilimleri Enstitüsü tarafından onaylanan lisansüstü tezimin/projemin tamamını veya herhangi bir kısmını, basılı ve elektronik formatta arşivleme ve aşağıda belirtilen koşullarla kullanıma açma izninin Kocaeli Üniversitesi'ne verdiğimi beyan ederim. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin/projemin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanımı bana ait olacaktır.
Tezin/projenin kendi özgün çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin/projenin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.
Yükseköğretim kurulu tarafından yayınlanan ***"Lisanüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"*** kapsamında tezim aşağıda belirtilen koşullar haricinde YÖK Ulusal Tez Merkezi/ Kocaeli Üniversitesi Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

☐ Enstitü yönetim kurulu kararı ile tezimin/projemin erişime açılması mezuniyet tarihinden itibaren 2 yıl ertelenmiştir.
☐ Enstitü yönetim kurulu gerekçeli kararı ile tezimin/projemin erişime açılması mezuniyet tarihinden itibaren 6 ay ertelenmiştir.
☒ Tezim/projem ile ilgili gizlilik kararı verilmemiştir.

Furkan GÖZ

ÖNSÖZ VE TEŞEKKÜR

Bu tez çalışmasında, anahtar kelime çıkarma problemi için üç farklı yöntem geliştirilmiştir. Yöntem, akademik çalışmalardan oluşan farklı veri kümeleri ile test edilmiştir.

Doktora eğitimim süresince desteğini esirgemeyen, tezimin her aşamasında sorunlarımı dinleyerek, görüşleri ile çalışmalarıma katkıda bulunan ve yoğun akademik yaşamında değerli zamanını problemlerimi çözmeye ayıran tez danışmanım Dr. Öğr. Üyesi Alev MUTLU'ya sonsuz teşekkürlerimi sunarım.

Tez çalışmama değerli akademik bilgilerini ve yardımlarını esirgemeyen Prof. Dr. Sevinç İLHAN OMURCA'ya ve Dr. Öğr. Üyesi Burcu YILMAZ'a,

Tezime sunmuş olduğu katkılardan dolayı Prof. Dr. Pınar KARAGÖZ'e ve Prof. Dr. Yaşar BECERİKLİ'ye,

Akademik çalışmalarım sırasında beni destekleyen Bilgisayar Mühendisliği öğretim üyeleri ve araştırma görevlilerine teşekkür ederim.

Maddi ve manevi desteklerini her zaman hissettiğim sevgili anneme, babama ve kardeşim Büşra GÖZ'e sonsuz teşekkürlerimi sunarım.

Aralık - 2022 Furkan GÖZ

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

SİMGELER VE KISALTMALAR DİZİNİ

Kısaltmalar

AP	:	Average Precision (Ortalama Kesinlik)
AR	:	Autoregressive (Otoregresif)
BERT	:	Bidirectional Encoder Representations from Transformers (Çift Yönlü Kodlayıcı Temsilleri)
CRF	:	Conditional Random Field (Koşullu Rastgele Alan)
DDİ	:	Doğal Dil İşleme
FP	:	Frequent Pattern (Yaygın Örüntü)
GPT	:	Generative Pretrained Transformer (Üretken Ön İşlemeli Dönüştürücü)
IDF	:	Inverse Document Frequency (Ters Belge Sıklığı)
lasf	:	Least Allowable Seen Frequency (En Düşük Görülme Sıklığı)
LDA	:	Latent Dirichlet Allocation (Gizli Dirichlet Ayrımı)
LSA	:	Latent Semantic Analysis (Gizli Anlamsal İndeksleme)
LSTM	:	Long Short-Term Memory (Çift Yönlü Uzun Kısa Bellek)
MAP	:	Mean Average Precision (Ortalama Kesinlik Değerlerinin Ortalaması)
MLM	:	Masked Language Modeling (Maskeli Dil Modeli)
MPNet	:	Multimodal Pretrained Network (Çok Kipli Ön İşlemeli Ağ)
MRR	:	Mean Reciprocal Rank (Ortalama Karşılıklı Sıra)
NGD	:	Normalized Google Distance (Normalleştirilmiş Google Mesafesi)
NSP	:	Next Sentence Prediction (Sonraki Cümle Tahmini)
PF	:	Phrase Frequency (Öbek Frekansı)
PLM	:	Permuted Language Modeling (Permütasyon Tabanlı Dil Modeli)
RR	:	Reciprocal Rank (Karşılıklı Sıra)
RVA	:	Reference Vector Algorithm (Referans Vektör Algoritması)
TF	:	Term Frequency (Terim Frekansı)
T5	:	Text-To-Text Transfer Transformer (Metinden Metne Aktarım Dönüştürücü)
XLNet	:	Extra Large Neural Network (Ekstra Büyük Sinir Ağı)

AKADEMİK MAKALELERDE ANAHTAR KELİME ÇIKARIMI İÇİN YENİ YAKLAŞIMLAR

ÖZET

Anahtar kelimeler, bir metni en iyi tanımlayan kelime ya da kelime öbekleridir. Anahtar kelimeler birçok Doğal Dil İşleme (DDİ) probleminin çözümünde etkin bir şekilde kullanılmaktadır. Çevrim içi metin sayısındaki artışla beraber metinlerden anahtar kelimelerin otomatik olarak elde edilmesi problemi ortaya çıkmıştır.

Anahtar kelime çıkarma yöntemleri denetimli ve denetimsiz öğrenme yaklaşımları olmak üzere iki temel sınıfa ayrılmaktadır. Denetimsiz öğrenmeye dayalı yöntemler etki alanından bağımsız olması ve eğitim verisine ihtiyaç duyulmaması açısından öne çıkmaktadır. Denetimli öğrenmeye dayalı yöntemler denetimsiz öğrenmeye dayalı yöntemlere göre daha güçlü bir öğrenme modeli sunar ve genellikle daha yüksek başarıma sahiptir.

Bu tez kapsamında anahtar kelime çıkarma probleminin çözümü için üç farklı yöntem önerilmiştir. Geliştirilen ilk yöntemde denetimsiz öğrenmeye dayalı çizge tabanlı bir yaklaşım benimsenmiştir. MGRank olarak adlandırılan bu yöntem çok kenarlı tam çizge model yapısını kullanmaktadır. Çizgede kenar ağırlıkları aday anahtar kelimelerin arasındaki mesafeye, düğüm ağırlıkları aday anahtar kelimelerin metin içerisindeki konumlarına göre belirlenmektedir. SkyWords olarak adlandırılan ikinci yöntem denetimli ve denetimsiz öğrenme modellerini birleştiren hibrit bir anahtar kelime çıkarma yöntemidir. SkyWords, Skyline operatörü ve çoğunluk oylama prensibinden faydalanarak yüksek kalitede aday anahtar kelimelerin belirlenmesini sağlar. SkyWords metin ile aday anahtar kelimelerin arasındaki anlamsal benzerliğe göre anahtar kelimeleri belirler. SkyRank olarak adlandırılan üçüncü yöntem ise denetimsiz öğrenmeye dayalı istatistiksel bir yaklaşıma sahiptir. SkyRank girdi olarak bir metin alır ve Skyline operatörü yardımıyla aday anahtar kelimeleri tespit eder. SkyRank anahtar kelimeleri metne en çok benzeyen aday anahtar kelimelerden seçer.

Geliştirilen yöntemler akademik makalelerden oluşturulmuş veri kümeleri ile test edilmiştir. Yöntemlerin başarısı literatürde yer alan çeşitli yöntemlerle karşılaştırılmıştır. Karşılaştırmada kesinlik, duyarlılık, F1-Skor, MRR ve MAP ölçütleri kullanılmıştır. Geliştirilen yöntemlerin diğer yöntemlere göre başarılı olduğu görülmüştür.

Anahtar Kelimeler: Anahtar Kelime Çıkarma, Çizge, Skyline Operatörü.

NEW APPROACHES FOR KEYWORD EXTRACTION IN ACADEMIC ARTICLES

ABSTRACT

Keywords are words or phrases that describe a text. Keywords are used effectively in solving many Natural Language Processing (NLP) problems. With the increasing number of online texts, the problem of automatically extracting keywords from texts has emerged.

Keyword extraction methods are divided into two basic classes: supervised and unsupervised learning approaches. Unsupervised learning-based methods are characterized by the fact that they are independent of the domain and do not require training data. Supervised learning-based methods provide a stronger learning model and generally perform better than unsupervised learning-based methods.

In this thesis, three different methods were proposed to solve the problem of keyword extraction. In the first method developed, a graph-based approach based on unsupervised learning was adopted. This method, called MGRank, uses a parallel complete graph model structure. In the graph, edge weights are determined according to the distance between candidate keywords, and node weights are determined according to the positions of candidate keywords in the text. The second method, called SkyWords, is a hybrid keyword extraction method that combines supervised and unsupervised learning models. SkyWords uses the Skyline operator and the principle of majority voting to identify high-quality candidate keywords. SkyWords determines keywords based on semantic similarity between text and candidate keywords. The third method, SkyRank, uses a statistical approach based on unsupervised learning. SkyRank takes a text as input and identifies candidate keywords using the Skyline operator. SkyRank selects the keywords that are most similar to the text from the candidate keywords.

The developed methods were tested with datasets created from academic articles. The success of the methods was compared with different methods from the literature. Precision, recall, F1-Score, MRR and MAP criteria were used for comparison. The methods were found to be successful compared to other methods.

Keywords: Keyword Extraction, Graph, Skyline Operator.

1. GİRİŞ

Dijital devrimle birlikte ucuzlayan depolama alanları, artan işlem gücü ve gelişen veri tabanı sistemleri bilgi kaynaklarında artışa sebep olmuştur. Web'in gelişmesi, arama motorlarının ortaya çıkması ve internetin kolay erişilebilir hale gelmesiyle birlikte bilimsel çalışmaların, haberlerin ve buna benzer çeşitli metinlerin depolanması sağlanmıştır. Böylece ilk başlarda sadece bir iletişim aracı olarak görülen internet, sonrasında bilgiye erişimin en temel aracı haline gelmiştir. İnternet ortamındaki erişilebilir bilgideki devasa artış, arama ve sorgulamanın zaman alıcı ve pahalı olmasına yol açmıştır. Bu durum içeriklerin indekslenmesi gerekliliğini zorunlu kılmıştır. İçeriklerde önemli bir yere sahip metinlerin indekslenmesi sayesinde anahtar kelimelerin kullanımı artmış ve bilgiye erişim kolaylaşmıştır (Papagiannopoulou ve Tsoumakas, 2020).

Bir ya da birden fazla kelime grubundan oluşan anahtar kelimeler bir metin belgesinin en iyi şekilde temsil edilmesini sağlar. Anahtar kelimeler doğal dil işleme ve bilgi erişim sistemlerinde yaygın olarak kullanılmaktadır (Rose ve diğ., 2010). Metin sınıflandırma (Du ve Huang, 2018), metin kümeleme (Liu ve diğ., 2009; Shubankar ve diğ., 2011), metin özetleme (Hernández-Castañeda ve diğ., 2021), soru cevaplama sistemleri (Zhao ve diğ., 2019), sorgu genişletme (Lee ve diğ., 2018), öneri sistemleri (Simsek ve Karagoz, 2020) anahtar kelimelerin kullanım alanlarına örnek verilebilir. Aynı zamanda anahtar kelimeler bir kullanıcıya aradığı bir konu ya da çalışma ile ilgili hızlı bir ön değerlendirme fırsatı sunar.

Anahtar kelimelerin önemli bir problemi anahtar kelime tespitinin el ile yapılmasının öznel bir durum olmasıdır. Doğru şekilde belirlenmemiş anahtar kelimeler yanıltıcı olabilir. Bir diğer problem de anahtar kelimelerin indekslenmesi çok sayıda metin için zorlayıcı ve zaman alıcıdır. Bu nedenle anahtar kelimelerin otomatik araçlarla indekslenmesi önemli bir problem haline gelmiştir (Nasar ve diğ., 2019). Metinlerin anahtar kelimelerle indekslenmesinde iki farklı yaklaşım bulunmaktadır: anahtar kelime atama ve anahtar kelime çıkarma. Anahtar kelime atama yaklaşımlarında önceden belirlenen aday anahtar kelimelerden bir sözlük oluşturulur ve anahtar kelimeler bu

sözlükten seçilir (Beliga ve diğ., 2015). Anahtar kelime atama yaklaşımında anahtar kelimelerin metinde bulunması gerekmez. Anahtar kelime atamanın yeterli bilgiye ulaşılamayan metinler için daha uygun olduğu düşünülmektedir. Anahtar kelime çıkarma yaklaşımlarında aday anahtar kelimeler metinde bulunan kelimelerden seçilir ve anahtar kelimelerin en az bir kez metinde bulunması gerekmektedir. Otomatik sistemlere ihtiyaç duyulan problemlerde anahtar kelime çıkarma yaklaşımı anahtar kelime atama yaklaşımına göre daha uygundur (Csomai, 2008). Anahtar kelime çıkarma yöntemleri genellikle denetimli ve denetimsiz yaklaşımlar olarak iki sınıfta incelenmektedir.

Denetimsiz öğrenmeye dayalı yaklaşımlarda anahtar kelime çıkarma problemi aday anahtar kelimelerin sıralanması problemi olarak ele alınır. İstatistiksel ve çizge tabanlı yaklaşımlar denetimsiz öğrenmeye dayalı anahtar kelime çıkarma yöntemlerinde öne çıkmaktadır. İstatistiksel yaklaşımlarda kelimeleri temsil edecek çeşitli istatistiksel özellikler bir araya getirilerek anahtar kelimeler belirlenir. KPMiner (El-Beltagy ve Rafea, 2009), YAKE! (Campos ve diğ., 2020), TeKET (Rabby ve diğ., 2020), ELSKE (Knittel ve diğ., 2021) ve HAKE (Merrouni ve diğ., 2022) gibi yöntemler istatistiksel yaklaşımla geliştirilen yöntemlere örnek olarak verilebilir. Çizge tabanlı yaklaşımlarda aday anahtar kelimeler bir çizge modeliyle temsil edilir. Aday anahtar kelimeler çizge değerlendirme algoritmalarından biri ile sıralanır. Çizge tabanlı yaklaşımlarda TextRank (Mihalcea ve Tarau, 2004), PositionRank (Florescu ve Caragea, 2017), MultipartiteRank (Boudin, 2018), SingleRank (Wan ve Xiao, 2008) ve TopicRank (Bougouin ve diğ., 2013) gibi yöntemler bulunmaktadır.

Denetimli yaklaşımlar otomatik anahtar kelime çıkarma problemini anahtar kelime ve anahtar kelime değil şeklinde ikili sınıflandırma problemi olarak ele alır. Karar ağaçları ve karar destek sistemleri gibi farklı denetimli öğrenme yöntemleri sınıflandırma problemini çözmek için kullanılır. CeKE (Caragea ve diğ., 2014), KEA (Witten ve diğ., 1999), WINGNUS (Nguyen ve Luong, 2010) ve SurfKE (Florescu ve Jin, 2018) gibi yöntemler denetimli yaklaşımlara örnek olarak gösterilebilir.

Anahtar kelime çıkarma probleminde denetimli ve denetimsiz yaklaşımların farklı yönlerden tercih edilme sebepleri bulunmaktadır. Denetimsiz yaklaşımların önemli bir

avantajı eğitim verisine ihtiyaç duyulmamasıdır. Anahtar kelimeleri çıkarmak için veri etiketleme maliyeti bulunmaz. Eğitim gerekmediğinden verinin niteliği ve verinin etki alanı ya da metnin türü ile ilgili problemler ortadan kalkar. Ayrıca hesaplama maliyeti düşüktür. Öte yandan denetimli öğrenme yaklaşımları güçlü bir model sunduğu için tercih edilmektedir. Denetimli öğrenmeye dayalı yaklaşımlar denetimsiz öğrenmeye dayalı yaklaşımlara göre daha başarılı performans gösterme eğilimindedir (Hasan ve Ng, 2014).

1.1. Problem Tanımı ve Katkıları

Anahtar kelimeler birçok DDİ probleminin çözümünde yaygın olarak kullanılmaktadır. Anahtar kelimelerin otomatik olarak çıkarılması çevrimiçi metin sayısındaki devasa artışla beraber önem kazanmıştır. Anahtar kelime çıkarma problemi ile ilgili son yıllarda çok sayıda çalışma bulunmaktadır (Jain ve diğ., 2022; Merrouni ve diğ., 2022; Wang ve Li, 2022). Literatürde anahtar kelime çıkarma problemi ile ilgili birçok çalışma bulunmasına rağmen geliştirilmesi gereken yönler bulunmaktadır. Bu çalışmada anahtar kelime çıkarma problemi ele alınmıştır ve farklı yaklaşımlara sahip üç yeni anahtar kelime çıkarma yöntemi geliştirilmiştir. Geliştirilen yöntemler şu şekilde isimlendirilmiştir: MgRank, SkyWords ve SkyRank. Bu bölümde geliştiren yöntemlerin motivasyon ve katkıları açıklanmaktadır.

1.1.1. MGRank

Çizge tabanlı yaklaşımlar anahtar kelime çıkarma probleminde yaygın bir şekilde tercih edilmektedir. Çizge tabanlı yaklaşımlarda aday anahtar kelimeler bir çizge modeliyle temsil edilir. Çizge modelini oluşturmak genellikle basit çizge modelleri kullanılmaktadır (Florescu ve Caragea, 2017; Mihalcea ve Tarau, 2004).

Çizge tabanlı yaklaşımlarda çizge modelinde düğümlerin arasındaki kenar bağlantıları *d* boyutunda bir pencere boyutu parametresine göre belirlenir. *d* değerinin önceden belirlenmesi gerekmektedir. Metin içerisinde yan yana geçen *d* değeri kadar düğüm arasında kenar bağlantısı kurulur. *d* değerinin seçimi çizge modelini ve buna bağlı olarak bulunan anahtar kelimeleri değiştirmektedir. Bu nedenle *d* parametresinin alacağı değere

karar vermek çizge tabanlı anahtar kelime çıkarma yaklaşımlarında bir problem olarak ortaya çıkmaktadır. (Wan ve Xiao, 2008).

Literatürde anahtar kelime çıkarma problemi için çeşitli veri kümeleriyle *d* parametresinin ideal değeriyle ilgili çeşitli deneyler bulunmaktadır. Ancak ideal bir *d* değerinin daha iyi olduğunu belirleyen bir sonuç bulunmamaktadır. Aksine veri kümesine göre değişen *d* değerleri belirlenerek daha iyi sonuç alındığını gösteren bulgular bulunmaktadır. Örnek olarak farklı yöntemlerin akademik çalışmalardan oluşturulmuş Inspec veri kümesiyle yapmış oldukları deney sonuçları ele alınabilir. Literatürde TextRank (Mihalcea ve Tarau, 2004),YAKE! (Vega-Oliveros ve diğ., 2019), Key2Vec (Mahata ve diğ., 2018) ve SGRank (Danesh ve diğ., 2015) anahtar kelime çıkarma yöntemleri Inspec veri kümesiyle ile ilgili sonuçlar vermiştir. Sırasıyla bu yöntemlerin *d* değerinin 2, 3, 5 ve 10 için en iyi sonucu verdiği belirtilmiştir. Ayrıca Florescu ve diğ. yaptığı çalışmada (Florescu ve Caragea, 2017) değişken *d* değerleri için çeşitli veri kümeleri ile testler gerçekleştirilmiştir ve test edilen her bir veri kümesinde farklı *d* değeri ile daha iyi sonuç elde edilmiştir.

Pencere boyutu parametresini kullanmayan yöntemler bulunmaktadır. Pencere boyutu parametresini kullanmayan çizge tabanlı yaklaşımlarda genellikle çizge modelinde iki aday anahtar kelime tek bir kenarla bağlanmıştır. Ancak birden fazla bulunan aday anahtar kelimeler için iki aday anahtar kelimenin arasına tek bir kenar bağlantısı kurmak metnin iyi bir şekilde temsil edilmesini sağlamaz. Çeşitli çalışmalar birden fazla geçen aday anahtar kelimeler için (Bellaachia ve Al-Dhelaan, 2014; Kang ve diğ., 2021; Li ve Ning, 2021; Wu ve diğ., 2021; Zhang, Peng ve diğ., 2013), birden fazla bağlantı kurmanın anahtar kelime çıkarma yönteminin performansını iyileştirebileceğini göstermiştir. Ayrıca birden fazla özelliğin tek bir kenar bağlantısıyla temsil edilmesi mümkün değildir (Bellaachia ve Al-Dhelaan, 2014; Vega-Oliveros ve diğ., 2019).

Pencere boyutu problemini çözmek için çok kenarlı tam çizge modeli kullanan farklı anahtar kelime çıkarma yöntemleri önerilmiştir (Boudin, 2018; Bougouin ve diğ., 2013; Danesh ve diğ., 2015). Bu yöntemler pencere boyutu parametre gereksinimini ortadan kaldırmaktadır ancak konu tabanlı sistemlerdir. Bu tür sistemlerdeki temel problem konu

modelleme tekniğinin seçimidir. Ayrıca bu yöntemler aday anahtar kelimelerin arasındaki çoklu ilişkileri modelleyememektedir.

Çizge tabanlı yöntemlerde belirtilen kısıtlar yeni bir anahtar kelime çıkarma yönteminin geliştirilmesi gerekliliğini göstermektedir. Bu çalışmada bu kısıtları çözmek için MGRank olarak adlandırılan bir yöntem geliştirilmiştir. MGRank düğüm ve kenar bağlantıları ağırlıklı, çoklu kenar bağlantılarına izin veren, tam çizge modeli ile oluşturulmuş çizge tabanlı anahtar kelime çıkarma yöntemidir.

Bir çizge modelinde çoklu kenarlar basit çizgeleri iki düğüm arasında paralel kenarlara izin vererek genişletir. Öneri sistemleri (Jiang ve diğ., 2016) gibi farklı problemlerde çoklu kenarlar ile oluşturulan çizge modellerinin etkili olduğu görülmüştür. Ancak anahtar kelime çıkarma problemine uygulanmış bir örneği bulunmamaktadır.

MGRank'in tam çizge modelini kullanması pencere boyutu parametresi ihtiyacını ortadan kaldırmaktadır ve sorgu yapılan metnin küresel ölçekte değerlendirilebilmesini sağlamaktadır. Çoklu kenarlar birden fazla bağlantı gerektiren düğümler için bilgi kaybını engeller ve metindeki tüm ilişkilerin daha iyi temsil edilmesi sağlar. MGRank'te düğüm ağırlıkları kelimelerin metinde görüldüğü ilk konuma göre belirlenmektedir. Kenar ağırlıkları da birbirine bağlanan kelimelerin arasındaki uzaklıklarla ters orantılı olacak şekilde belirlenir.

MGRank'in çalışma mantığı şu şekildedir: Metinde bulunan isim ve sıfatlar aday anahtar kelime olarak belirlenir. Pagerank algoritması aracılığıyla aday anahtar kelimeler derecelendirilir. Daha sonra metinde yan yana geçen kelimelerden kelime öbekleri elde edilir. Aday anahtar kelimelerin dereceleri ile kelime öbeklerin önem değeri belirlenir ve en yüksek değere sahip kelime öbekleri anahtar kelime listesine eklenir.

MGRank'in katkıları şu şekilde sıralanabilir:

- MGRank ağırlıklı, çok kenarlı tam çizge model yapısını kullanan ilk anahtar kelime çıkarma yöntemidir.
- Önerilen çizge modeli pencere boyutu parametresini kullanmaz.
- Paralel kenarlar ile metinde birden fazla bulunan aday anahtar kelimelerin daha iyi

temsil edilmesi sağlanır.

- MGRank birbirine yakın konumda bulunan aday anahtar kelimelerin etkisini arttıran bir prensibe sahiptir. Bu prensip çizgede metin içerisinde daha yakın konumda bulunan düğümlerin kenar bağlantılarına daha yüksek ağırlıklandırma yaparak sağlanır.

1.1.2. SkyWords

Aday anahtar kelime kümesinin kalitesi anahtar kelime çıkarma yönteminin başarısını etkilemektedir (Abulaish ve diğ., 2022). Anahtar kelime çıkarma yöntemlerinde genellikle isim ve sıfatlar aday anahtar kelime olarak kabul edilir. Ayrıca bir metinde isim ve sıfatlar diğer sözcük türlerine göre daha çok bulunur (Kwary ve diğ., 2018). Kelime sayısının artması metindeki etkisiz ve anlamsız kelimelerin anahtar kelime olarak belirlenme ihtimalini arttırmaktadır. Bu problemi çözmek için kelime uzunluğu ve kelime frekansı gibi istatistiksel özellikleri kullanarak aday anahtar kelime kümesinin büyüklüğünü azaltmayı tercih eden anahtar kelime çıkarma çalışmaları bulunmaktadır (Papagiannopoulou ve Tsoumakas, 2018b; Yang ve diğ., 2018).

Literatürde bir anahtar kelime çıkarma yönteminin başarısını arttırmak için yüksek kalitede aday anahtar kelime kümesi oluşturan yöntemler bulunmaktadır. Ancak bu tür çalışmaları içeren anahtar kelime çıkarma yöntemleri genellikle belirli bir alan üzerine odaklanmıştır (Abulaish ve diğ., 2022; Yang ve diğ., 2018). Bu nedenle problemden bağımsız genel amaçlı anahtar kelime çıkarma problemi için yüksek kalitede aday anahtar kelime kümesi oluşturmak çalışılması gereken bir problemdir. Bu amaçla bu çalışmada SkyWords isimli bir yöntem geliştirilmiştir.

SkyWords denetimli ve denetimsiz öğrenme yaklaşımlarını içeren hibrit bir anahtar kelime çıkarma yöntemidir. Bu yöntemde yüksek kalitede aday anahtar kelime listesi oluşturmak için Skyline operatörüne dayalı yeni bir denetimli öğrenme adımı uygulanmaktadır (Borzsony ve diğ., 2001). Aday anahtar kelimeleri derecelendirmek için vektör temsiline dayalı denetimsiz bir adım benimsenmiştir. Böylece bir metnin aday anahtar kelimeleri farklı metinlerin bilinen anahtar kelimelerine ait özellikleri kullanılarak belirlenir ve belirlenen aday anahtar kelimelerin yüksek kalitede olması

amaçlanır.

SkyWords'ün çalışma mantığı şu şekildedir: Skyline operatörüyle bilinen anahtar kelimelerin özelliklerinden baskın olan özellik vektörleri tespit edilir. Belirlenen baskın özellik vektörleri yüksek kalitede aday anahtar kelimeleri bulmak için kullanılır. Bir kelimenin yüksek kalitede aday anahtar kelime listesine eklenebilmesi için o kelimenin en az belirlenen özellik vektörleri kadar iyi olması gerekmektedir. Bu ölçüm iki aşamalı çoğunluk oylama prensibi ile gerçekleştirilir. Oylamanın ilk adımında bir kelimenin özellik vektörü baskın özellik vektörleriyle karşılaştırılır ve kelimeye ait özellik vektörünün değerleri karşılaştırılan özellik vektörünün değerlerinin çoğundan daha iyiyse evet oyu alır ve aksi durumda hayır oyu alır. Tüm vektörlerle karşılaştırma yapıldıktan sonra son karar evet ve hayır oylarının sayısına göre verilir. Evet oylarının sayısı hayır oylarının sayısından yüksekse kelime aday anahtar kelime listesine eklenir. Evet oy sayılarına göre en yüksek oya sahip *n* sayıda kelime belirlenir ve belirlenen kelimeleri içeren 2 ve 3 uzunluklu isim tamlamaları nihai aday anahtar kelime kümesine dahil edilir. SkyWords'ün anahtar kelimeleri bulmak için uyguladığı temel prensip şu şekildedir. Bir metnin anahtar kelimeleri metne en çok benzeyen aday anahtar kelimeleri olmalıdır. Bu amaçla metin ve aday anahtar kelimelerin vektör temsilleri bulunur. Kosinüs benzerliği ile metne en çok benzeyen *n* sayıda aday anahtar kelime, anahtar kelime olarak belirlenir.

SkyWords'ün katkıları şu şekilde sıralanabilir:

- SkyWords anahtar çıkarma probleminde Skyline operatörünün kullanıldığı ilk yöntemdir.
- Geliştirilen aday anahtar kelime oluşturma algoritması tamamen istatistiksel özelliklere dayanmaktadır. Bu özellikler aday anahtar kelimelerin birçok farklı yönden değerlendirilmesini sağlamaktadır.
- Geliştirilen yöntem son işleme gerek duymaz. Anahtar kelimeler metne en çok benzeyen aday anahtar kelimelerden seçilir.

1.1.3. SkyRank

Papagiannopoulou ve arkadaşı anahtar kelimelerin dağılımıyla ilgili bir çalışma yapmıştır. Bu çalışmada metin içerisindeki kelimelerin 5 boyutlu vektör temsilleri elde edilmiştir ve aday anahtar kelimelerin metin içerisindeki dağılımı incelenmiştir. Elde edilen bulgular aday anahtar kelimelerin birbirine daha yakın konumda bulunduklarını ve diğer kelimelerden ayrıştıklarını göstermektedir (Papagiannopoulou ve Tsoumakas, 2018a).

SkyRank bu mantığa dayanarak geliştirilmiştir. SkyRank denetimsiz öğrenmeye dayalı istatistiksel bir yöntemdir. Bu yöntemde ayrışan aday anahtar kelimeleri bulmak için Skyline operatörü kullanılmaktadır.

SkyRank'in çalışma mantığı şu şekildedir: Metinde bulunan kelimelerin farklı özellikler için değerleri bulunarak özellik vektörleri oluşturulur. Daha sonra Skyline operatörü ile baskın özellik vektörlerine sahip kelimeler belirlenir ve aday anahtar kelime olarak alınır. Aday anahtar kelimeler metinde bulunan komşu kelimelerle birleştirilerek kelime öbekleri elde edilir. Aday anahtar kelimeler ile metin arasındaki anlamsal benzerliğe göre anahtar kelimeler belirlenir.

SkyWords ve SkyRank anahtar kelimeleri belirlemek için aynı istatistiksel özellikleri kullanmaktadır. İki yöntem de bu amaçla Skyline operatöründen yararlanmaktadır. SkyWords Skyline operatörünü model oluşturmak için kullanır ve çoğunluk oylama prensibiyle destekler. SkyRank çoğunluk oylama prensibini kullanmaz ve doğrudan metin içindeki baskın kelimeleri bulmayı amaçlar. Skyline operatörü SkyWords için önem eğiliminin tersi yönde, SkyRank için önem eğilimi ile aynı yönde çalıştırılmaktadır. Anahtar kelime bulma aşaması SkyWords ve SkyRank için ortaktır.

SkyRank'in katkıları şu şekilde sıralabilir:

- SkyRank aday anahtar kelimeleri metindeki diğer kelimelere göre daha iyi özelliklere sahip olan ve ayrışan kelimelerden seçer.
- Skyline operatörünü kullanmak için eğitime ihtiyaç duymaz.

- Geliştirilen yöntemde son işlem aşaması yoktur. Anahtar kelimeler metne en çok benzeyen aday anahtar kelimelerden seçilir.

1.2. Tezin Organizasyonu

Bu çalışma şu şekilde organize edilmiştir: 2. Bölümde literatür taraması verilmiştir. 3. Bölümde geliştirilen yöntemler anlatılmaktadır. 4. Bölümde deney ortamı, veri kümeleri, karşılaştırılan yöntemler değerlendirme ölçütleri ve deneysel sonuçlar sunulmaktadır. Son bölümde tez çalışması ile ilgili genel çıkarımlar Sonuçlar ve Öneriler başlığı altında verilmektedir.

2. LİTERATÜR

Bu bölümde ilk olarak anahtar kelime çıkarma problemi ve ilgili yapılmış çalışmalar anlatılmaktadır. Daha sonra Skyline sorguları açıklanmaktadır. Skyline sorguları bu çalışma kapsamında geliştirilen SkyWords ve SkyRank yöntemlerinde kullanılmaktadır. Son bölümde Transformers (Dönüştürücüler) modelleri anlatılmaktadır. SkyWords ve SkyRank yöntemlerinde anahtar kelimeleri bulmak için Transformers modellerinden all-mp-net-base-v2 (HuggingFace, 2022) modeli kullanılmıştır.

2.1. Anahtar Kelime Çıkarma

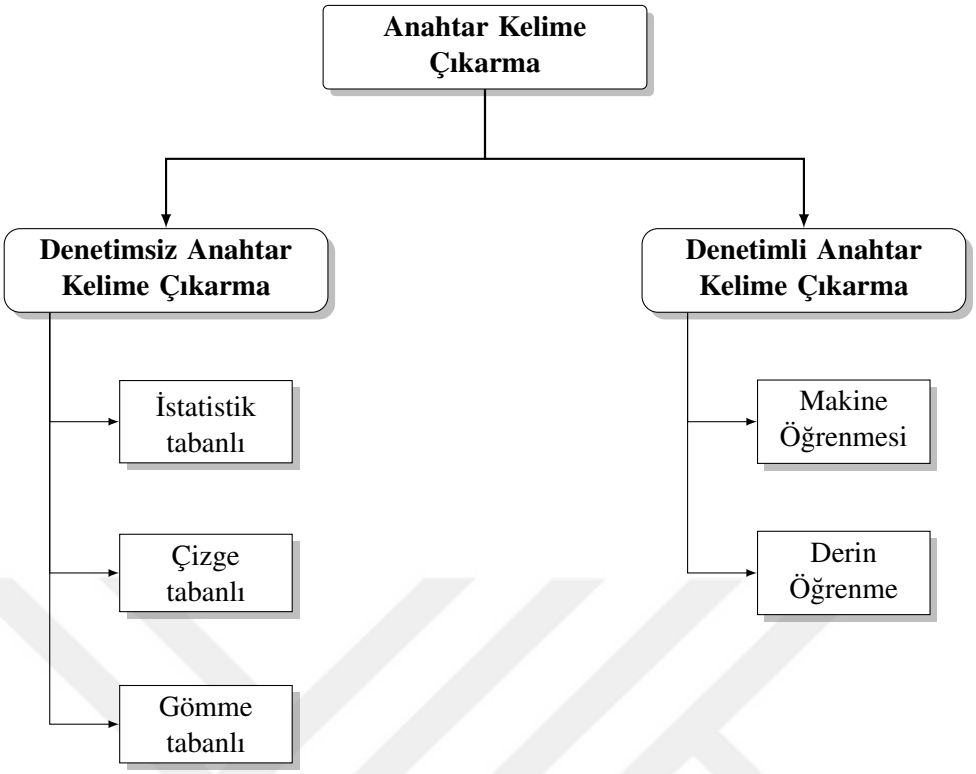
Anahtar kelimeler bir metni en iyi tanımlayan kelime ya da kelime öbekleridir. Anahtar kelimeler birçok DDİ problemin çözümünde önemli bir yer edinmektedir. Anahtar kelimelerin otomatik olarak elde edilmesinde literatürde farklı yaklaşımlar bulunmaktadır.

Anahtar kelime çıkarma yöntemleri denetimsiz anahtar kelime çıkarma yöntemleri ve denetimli anahtar kelime çıkarma yöntemleri olmak üzere iki sınıfa ayrılabilir (Papagiannopoulou ve Tsoumakas, 2020). Denetimsiz anahtar kelime çıkarma yöntemleri istatistik tabanlı, çizge tabanlı ve gömme tabanlı olmak üzere kendi içinde 3 sınıfta incelenebilir. Denetimli anahtar kelime çıkarma yöntemleri makine öğrenmesi ve derin öğrenme yöntemleri olmak üzere 2 sınıfta incelenebilir. Çalışmanın bu bölümünde anahtar kelime çıkarma yöntemleri ve yapılan çalışmalar incelenmektedir. Şekil 2.1 anahtar kelime çıkarma yöntemlerinin sınıflandırılması verilmiştir.

2.1.1. Denetimsiz Anahtar Kelime Çıkarma Yöntemleri

Denetimsiz yaklaşımlarda etiketli eğitim verilerine gerek duyulmaz. Aynı zamanda etki alanından bağımsızdır. Verilerin el ile olarak etiketlenmesinin zor olması ve verilerin doğru etiketlenememesinin öğrenme modelini sınırlandırması kaçınılmazdır. Denetimsiz öğrenme yaklaşımları bu problemleri ortadan kaldırmaktadır.

Denetimsiz yaklaşımlar anahtar kelime çıkarma problemini derecelendirme problemi olarak ele almaktadır. Denetimsiz anahtar kelime çıkarma yöntemleri üç temel aşamadan



Şekil 2.1. Anahtar Kelime Çıkarma Yöntemlerinin Sınıflandırılması

oluşmaktadır: veri ön işleme ve aday anahtar kelimelerin belirlenmesi, aday anahtar kelimelerin derecelendirilmesi, son işlem.

- Ön işlem ve aday anahtar kelimelerin belirlenmesi aşamasında bir metne öncelikle DDİ'nin temel ön işlem adımları uygulanır ve bunun sonucunda aday anahtar kelimeler tespit edilir. Bu ön işlem adımları sözcük türlerinin tespit edilmesi ve seçilmesi, kelime köklerinin bulunması, metnin en küçük birimlere dönüştürülmesi, cümlelerin belirlenmesi, etkisiz kelimelerin tespit edilmesi gibidir. Ayrıca bazı çalışmalarda konu tespiti gibi daha karmaşık adımlar bulunmaktadır. Literatürde yer alan yöntemler bu aşamalardan bazılarını kullanarak aday anahtar kelimeleri belirlemektedir.
- Aday anahtar kelimelerin derecelendirilmesi aşamasında belirlenen her bir aday anahtar kelimenin önem değerleri belirlenmektedir. Anahtar kelime çıkarma yöntemleri bu aşamada istatistiksel ya da çizge tabanlı yaklaşımları benimsemişlerdir. İstatiksel yaklaşımları benimseyen yöntemler aday anahtar kelimeleri TF, kelimelerin pozisyon bilgisi gibi istatistiksel özellikleri kullanarak

derecelendirir. Çizge tabanlı yaklaşımları benimseyen yöntemler Pagerank gibi düğüm derecelendirme algoritmaları kullanmaktadır.

- Son işlem aşaması anahtar kelimelerin bulunduğu adımdır. Bu amaçla öncelikle metin içinde yan yana bulunan aday anahtar kelimelerden kelime öbekleri elde edilir. Daha sonra kelimelerin önem değerleri ile kelime öbekleri için bir değer hesaplanır. Son adımda en yüksek değere sahip *n* sayıda kelime öbeği anahtar kelime olarak belirlenir.

İstatistik tabanlı anahtar kelime çıkarma yöntemleri uygulanması açısından diğer yöntemlere göre daha basit olan yöntemlerdir. Bu tip yöntemlerde metinlerden elde edilen TF, pozisyon, TF-IDF gibi istatistiksel özellikler ile aday anahtar kelimeler sıralanır. En yüksek değere sahip aday anahtar kelimeler anahtar kelime olarak alınır. İstatistik tabanlı yöntemler etki alanından ve konudan bağımsızdır.

İstatistik tabanlı yaklaşımlarda ilk temel yöntem TF-IDF olarak kabul edilmektedir (Jones, 1972). TF metindeki bir kelimenin toplam kelime sayısına oranını, IDF metin sayısının ilgili kelimenin bulunduğu metin sayısına oranının logaritması ile hesaplanır. TF-IDF ise TF ve IDF hesaplamalarının çarpımı ile elde edilir. Bu yöntem, bir metinden anahtar kelime çıkarmak için basit bir yaklaşıma sahiptir ancak sorgulanan metinden dışında büyük bir kelime sözlüğüne ihtiyaç duymaktadır.

KP-Miner TF-IDF yöntemi dışında yeni istatistiksel özellikler kullanarak geliştirilmiş istatistiksel tabanlı bir yöntemdir. KP-Miner aday anahtar kelimeleri iki aşamalı bir filtreleme yöntemi ile seçmektedir. İlk aşamada etkisiz kelimeler ve noktalama içermeyen kelime öbeklerini aday anahtar kelime listesine alır. Böyle çalışmasının sebebi kelime öbeklerinin noktalamalarla ayrılamayacağının ve kelime öbeklerinin çoğunda etkisiz kelimelerin bulunmayacağının düşünülmesidir. İkinci aşamada lasf ve cutoff olarak adlandırılan iki parametre kullanılmaktadır. lasf değeri bir metinde bir kelimenin aday anahtar kelime olabilmesi için gerekli minimum bulunma sayısını göstermektedir. cutoff değeri ise metinde arama yapılacak kelime sayısı gösteren bir parametredir. Bu parametre ile bir metnin başlangıcından itibaren cutoff değerine kadar olan kelime öbekleri alınmaktadır. Bu ölçüt uzun metinlerde metnin belirli bir

bölümünden sonra anahtar kelimelerin az olacağı düşüncesi ile aramayı sonlandırmak için kullanılmaktadır. Bu çalışmada kelime öbeklerinin kelimeler kadar birlikte bulunamayacağı için sadece TF-IDF ile anahtar kelimeleri bulmanın yeterli olmayacağı vurgulanmaktadır. Bu nedenle KP-Miner TF-IDF ile birlikte aday anahtar kelimelerin pozisyonu ve metindeki aday anahtar kelimelerin uzunluğunu dikkate alan bir ölçüt kullanarak anahtar kelimeleri bulmaktadır (El-Beltagy ve Rafea, 2009).

RAKE anahtar kelime çıkarma yönteminde etkisiz kelimeleri içeren bir liste ve kelime ile kelime öbeklerini ayırmak için kullanılan bir liste girdi olarak alınır. RAKE aday anahtar kelimeleri bir metinde yan yana bulunan kelimeleri bu listelerde yer alan elemanlara göre ayırarak bulur. Aday anahtar kelimelerin önem değerleri kelimelerin frekans ve derecelerine dayalı bir hesaplamayla elde edilir. Kelime frekansı bir kelimenin aday anahtar kelimelerde ne sıklıkta göründüğünü, kelime derecesi ise bir kelimenin daha uzun aday anahtar kelimelerde bulunma sıklığını ölçer. RAKE'de daha uzun aday anahtar kelimelerde geçen kelimelerin daha önemli olduğu varsayımı benimsenmiştir. Ayrıca birden fazla kelime içeren anahtar kelimelerin genellikle etkisiz kelimeleri içermediği ve genellikle DDİ problemlerinde değerlendirmeye alınmadığı vurgulanmıştır (Rose ve diğ., 2010).

YAKE! farklı istatistiksel özellikleri birleştiren anahtar kelime çıkarma yöntemidir. Bu yöntem kelime pozisyonu, büyük küçük harf ölçütü, kelime frekansı, kelime cümle farklılığı ve kelime çeşitliliği olarak adlandırılan istatistiksel özelliklerden faydalanmaktadır. YAKE! ilk olarak aday anahtar kelimeleri temel DDİ adımları ile bulmaktadır. İkinci aşamada bu aday anahtar kelimelerin 5 istatistiksel özellik için değerleri hesaplanır ve bu değerlerin dahil edildiği bir hesaplamayla her bir aday anahtar kelime için bir değer bulunur. Daha sonra aday anahtar kelimeler en fazla n uzunluklu olacak şekilde birleştirilir ve kelime öbeklerinden oluşan her bir aday anahtar kelime için bir değer bulunur. Son aşamada aday anahtar kelimeler arasında benzerlik hesaplanarak anahtar kelimelerin birbirine göre farklılığı arttırılır. Yöntemde kullanılan istatistiksel özelliklerden kelime pozisyonu bir metnin ilk cümlelerinde yer alan kelimelere daha sonra ortaya çıkan kelimelere göre daha yüksek ağırlık verir. Büyük-küçük harf ölçütü büyük harfli kelimelerin küçük harfli kelimelerden daha önemli olma eğiliminde olduğu

varsayımına dayanır ve cümle başlarında yer alan büyük harfli kelimeleri hesaba katmayarak büyük harfli kelimelere daha yüksek ağırlık verir. Kelime frekansı bir metin içerisinde daha sık bulunan kelimelere daha yüksek ağırlık verildiği ölçüttür ve uzun metinlerde yüksek sayıda geçen kelimelerin önemsiz olabileceği varsayımına dayalı olarak kelime frekanslarının ortalaması ile normalize edilir. Kelime cümle farklılığı özelliği bir aday anahtar kelimenin farklı cümlelerde ne sıklıkla göründüğünü ölçmektedir ve bir kelime ne kadar çok farklı cümlede bulunursa o kadar önemlidir varsayımını yansıtmaktadır. Kelime çeşitliliği bir kelimenin sağında ve solunda bulunan kelimelerin çeşitliliğini ölçer ve farklı kelimelerle yan yana bulunma sayısı arttıkça önem değeri artmaktadır (Campos ve diğ., 2020).

TeKET dil ve etki alanından bağımsız geliştirilen istatistiksel tabanlı bir yöntemdir. TekET ilk olarak bir metinden isim tamlamalarını bulur ve aday anahtar kelimelerin listesini oluşturur. Önceden belirlenen bir eşik değerinden daha az sayıda bulunan aday anahtar kelimeler ve alfabetik olmayan karakterleri içeren aday anahtar kelimeler filtrelenerek aday anahtar kelime listesi güncellenir. Daha sonra aday anahtar kelimelerden oluşan ikili ağaç yapısı oluşturulur ve bu ağaçta zayıf düğümler bulunarak budama yapılır. TeKET aday anahtar kelimelerin içerdiği her bir kelimenin diğer aday anahtar kelimelerde bulunma durumunu bir indeks ile ilişkilendirmektedir. Bu indeks aday anahtar kelimelerin içerdiği her bir kelimenin frekansları toplamı ile çarpılarak hesaplanır ve en yüksek değere sahip aday anahtar kelimeler anahtar kelime olarak alınır (Rabby ve diğ., 2020).

ELSKE çok sayıda metin içeren bir metin koleksiyonundan anahtar kelime çıkarmak için geliştirilmiş istatistik tabanlı yöntemdir. ELSKE ilk olarak kelime ve kelime öbeklerinden aday anahtar kelimeleri oluşturur. Aday anahtar kelimeler TF-IDF'den uyarlanan kelime öbeklerine dayalı PF-IDF ölçütü ile sıralanır. Daha sonra ELSKE anahtar kelimeleri bulmak için üç aşamalı filtreleme uygular. İlk olarak etkisiz kelimeleri hariç tutarak sadece bir kelime içeren kelime öbekleri aday anahtar kelime listesinden atılır. İkinci olarak etkisiz kelimelerden birini içeren uzun kelime öbekleri listeden atılarak aynı kelimeleri içeren kısa kelime öbekleri alınır. Son olarak kelime sayısına göre uzun bir kelime öbeğinin içinde bulunan kısa uzunluklu kelime öbekleri aday

anahtar kelime listesinden atılır ve anahtar kelime listesi oluşturulur (Knittel ve diğ., 2021).

HAKE bir metnin dil, yapı ve anlam bakımından çeşitli istatistiksel özelliklerini kullanarak anahtar kelimeleri bulmayı amaçlayan bir yöntemdir. HAKE anahtar kelime çıkarmak için ilk adımda temel DDİ adımlarını uygular. Bu adımlardan sonra ilgili metindeki cümleler bulunur ve her bir cümle için sözcük türüne dayalı parçalama ağacı oluşturulur. HAKE bu ağaç yapısına belirli kurallar uygulayarak aday anahtar kelimeleri belirler. Bu kurallara göre öncelikle aday anahtar kelimeler isim tamlaması olmalıdır, daha sonra isim tamlamaları etkisiz kelimelerden herhangi birisi ile başlamamalıdır ya da sonlanmamalıdır, ayrıca aday anahtar kelimeler 5'den fazla kelime içermemelidir. Ayrıca ortak içeriğe sahip isim tamlamalarından ilk bulunan aday anahtar kelime aday anahtar kelime listesine dahil edilmez. HAKE aday anahtar kelimeleri bu kurallara göre belirledikten sonra aday anahtar kelimelerin çeşitli istatistiksel, anlamsal ve yapısal özellikler için değerlerini hesaplar. İstatistiksel özellik ölçümü için metinler k-means ile kümelenir ve kelimeler için ki kare istatistiği kullanılır. Ki kare değeri bir küme ile bir kelime arasındaki ilişkinin derecesini gösterir. Anlamsal özellik olarak karşılıklı bilgi (mutual information) kullanılır. Karşılıklı bilgi ile kelimeler arasındaki ilişkinin gücü hesaplanır. Anlamsal özellik eş anlamlı ve çok anlamlı kelimelerden kaçınmak için kullanılmaktadır. Yapısal özellik olarak pozisyon bilgisi kullanılmaktadır. Bir kelime başlık ve özette bulunuyorsa kelimenin daha önemli olacağı varsayımı ile yüksek ağırlık alır. HAKE son aşamada hesaplanan değerleri birleştirerek her bir kelimenin önem değerini bulur ve anahtar kelimeler aday anahtar kelimelerin içerdiği her bir kelimenin değerleri toplamının kelime sayısına oranı ile bulunur (Merrouni ve diğ., 2022).

Zhuohao ve arkadaşları istatistiksel bir yaklaşımla SRP-TF-IDF olarak adlandırılan bir model önermiştir. TF-IDF modeli konudan bağımsız çalışmaktadır. Önerilen modelde metinlerin konularını hesaba katmak için TF-IDF uyarlanmıştır. Model TF-IDF'yi ağırlık dengeleme algoritması ile birleştirmektedir ve anahtar kelimeleri bu hesaplamaya göre bulmaktadır (Zhuohao ve diğ., 2021).

Li ve arkadaşları kapsam değeri ve anlamsal çeşitlilik adında iki yeni istatistiksel özellik

önerdiler ve kelimelerin pozisyon bilgisiyle birleştirip yeni bir istatistiksel yöntem geliştirdiler. TripleRank olarak adlandırdıkları yöntemde ilk olarak ön işlemler aracılığıyla aday anahtar kelimeler belirlenir. TripleRank her bir aday anahtar kelime için kapsam değeri, anlamsal çeşitlilik ve pozisyon bilgilerini hesaplar ve matematiksel bir hesaplamayla birleştirerek tek bir değer elde eder. Bu değer ile anahtar kelimeler belirlenir. TripleRank'te pozisyon bilgisi metinde önce bulunan kelimelere yüksek ağırlık verecek şekilde hesaplanmaktadır. Anlamsal çeşitlilik özelliğini hesaplamak için LDA modeli kullanılmaktadır. LDA ile aday anahtar kelimelerin konu dağılımlarına göre bir olasılık değeri hesaplanır. Bu yaklaşımla birbirine benzer aday anahtar kelimelerin bulunması ve anahtar kelime çeşitliliğinin arttırılması amaçlanmaktadır. Kapsam değerini hesaplamak için aday anahtar kelimelerin metin içerisinde yer alan diğer kelimelerle benzerliğini belirleyen bir hesaplama kullanılmaktadır. Bu değer Word2Vec modeli ile hesaplanmaktadır. Bu yönteme göre bir aday anahtar kelimenin metinde yer alan diğer kelimelerle olan benzerliğinin yüksek olması metnin daha iyi temsil edileceğini gösterir (Li ve diğ., 2021).

KP-Rank konu tabanlı geliştirilen anahtar kelime çıkarma yöntemlerinden birisidir. Yönteme göre anahtar kelimelerin metnin konularını temsil etmesi gerekmektedir. KP-Rank aday anahtar kelimeleri bulmak için cümle ayrıştırma ağacı kullanmaktadır. Bu ağacın kullanılması için ilk olarak metin cümlelere dönüştürülür ve bu cümleler sözcük türlerine göre modellenir. Kural tabanlı bir yaklaşımla (etkisiz kelime içermeyen, noktalama içermeyen vb.) aday anahtar kelimeler bulunur. Ayrıca bir cümlede yer alan birden fazla kelime öbeği aynı kelimeleri içeriyorsa aday anahtar kelime listesinden çıkarılır. Daha sonra aday anahtar kelimelerin LSA modeliyle anlamsal benzerlikleri bulunur. Hiyerarşik Kümeleme algoritması ile aday anahtar kelimeler konularına göre kümelenir. Bu aşamada kosinüs benzerliği aday anahtar kelimelerin konu kümelerini bulmak için kullanılır. Bir kümenin elemanları içerisinden diğer elemanlarına göre minimum ortalama değere sahip olan üye konu öbeği olarak seçilir. Daha sonra her bir konu öbeğinin önem derecelerine göre bölüm, paragraf ve cümle frekansları bulunur ve en yüksek değerli *n* sayıda aday anahtar kelime, anahtar kelime olarak belirlenir (Aman ve diğ., 2021).

Veisi ve arkadaşlarının önerdiği anahtar kelime çıkarma yönteminde varyans tabanlı yeni özellikler kullanılmıştır. Yöntemde başlangıçta temel ön işlem adımlarıyla kelimeler elde edilir. Kelimeler istatistiksel yöntemlerle ağırlıklandırılır ve en yüksek değere sahip kelimeler anahtar kelime olarak bulunur. Önerilen ağırlıklandırma yöntemleri şunlardır: TF varyans, IDF oranı; varyans, TF oranı; TF ile varyansın kesişimi ve varyans ile TF-IDF'nin kesişimi; varyans, TF-IDF oranı; varyans, TF oranı; TF varyans kesişimi ve varyans TF-IDF kesişimi (Veisi ve diğ., 2020).

Sharma ve arkadaşları metnin yapısı ve büyüklüğünden bağımsız yeni bir yaklaşım önermiştir. Yöntemde ilk aşamada aday anahtar kelimeler belirlenir. Aday anahtar kelimeler en fazla üç uzunluklu olacak şekilde büyük harf içeren kelimelerden ya da isim tamlamalarından oluşturulmaktadır. Daha sonra her bir aday anahtar kelimeden özellik çıkarımı yapılır. Elde edilen özelliklerle birlikte bir kelime vektörü de kullanılarak aday anahtar kelimeler için bir ağırlık bulunur. Bu yöntemde şu özellikler kullanılmaktadır: frekansa dayalı çeşitli özellikler, kelime uzunluğu, adayların farklı cümlelerde görünme sıklığı, kelimelerin birlikte bulunma sıklığı, pozisyon bilgisi, Wikipedia gibi farklı kaynaklardan elde edilen özellikler, kelime vektörü. Yapılan çalışmada kullanılan özellikler farklı kombinasyonlarla test edilerek anahtar kelime çıkarma problemi için başarımları ölçülmüştür. Yöntem anahtar kelimeleri bulmak için Hiyerarşik Kümeleme algoritması kullanır (Sharma ve diğ., 2021).

Mao ve arkadaşlarının önerdiği istatistik tabanlı yöntem üç aşamada gerçekleşmektedir. İlk aşamada bir metin temel ön işlem adımlarından geçirilir. Düzenli ifadeler yardımıyla en fazla 5 uzunluklu olacak şekilde aday anahtar kelimeler bulunur. İkinci aşamada kelimelerin önem değerleri bulunur. Bu amaçla kelimelerin metin içerisinde birlikte bulunma sayıları, WordNet, NGD ve önceden eğitilmiş GloVe modeli kullanılmaktadır. Bu özelliklerin her biri önce teker teker daha sonra farklı kombinasyonlarla birleştirilerek aday anahtar kelimeleri sıralamak için kullanılmıştır. Anahtar kelimeleri bulmak için son aşamada aday anahtar kelimeleri içeren kelimelerin ağırlıkları toplanarak önem değerleri bulunur. En yüksek değere sahip aday anahtar kelimeler, anahtar kelime olarak alınır. (Mao ve diğ., 2020).

Alrehamy ve arkadaşı SemCluster olarak adlandırdıkları kümeleme tabanlı bir yöntem geliştirmiştir. Bir anahtar kelime çıkarma yönteminde farklı bir kaynak ya da hazır bir sözlük kullanmak etki alanından bağımsız olması sebebiyle iyi sonuç vermeyebilir. Bu problemi çözmek için SemCluster bilgi kaynağı olarak WordNet'i kullanmaktadır. SemCluster ilk olarak bir metinden isim ve isim tamlamalarını tespit eder. Daha sonra elde edilen kelimelerin anlamsal ve sözlüksel ilişkilerine göre benzerlikleri tespit edilir ve benzer olan kelimeler kümelenir. Her bir küme bir metin için bir konuyu temsil etmektedir ve küme merkezine yakın kelimeler aday anahtar kelimeleri bulmak için kullanılır. Son aşamada düzenli ifadeler yardımıyla aday anahtar kelimelerden anahtar kelimeler seçilir (Alrehamy ve Walker, 2018).

Çizge tabanlı yöntemlerde bir metin, metinden elde edilen aday anahtar kelimelerin düğümlerle, aday anahtar kelimelerin arasındaki bağlantıların kenarlarla temsil edildiği çizge ile temsil edilmektedir. Çizgede düğüm ağırlıkları aday anahtar kelimelerin önemini, kenar ağırlıkları da aday anahtar kelimelerin arasındaki ilişkinin gücünü göstermektedir. Çizge tabanlı yöntemlerde amaç çizgedeki düğümleri bir düğüm derecelendirme algoritması kullanarak sıralamaktır. Çizgede en yüksek n sıralamasına sahip aday anahtar kelimeler metnin anahtar kelimeleri olarak kabul edilmektedir.

Birçok çizge tabanlı anahtar kelime çıkarma yönteminde düğüm derecelendirme algoritması olarak Google'ın Pagerank algoritması tercih edilmiştir. Pagerank algoritması çizgelerde düğümlerin önemini belirlemek için kullanılan özvektör merkeziliğine dayanmaktadır. Pagerank bir çizgede düğüm ağırlıklarını; bir düğüm için komşuları ve komşuların komşularını dikkate alarak günceller ve tüm ağı yinelemeli bir şekilde hesaplamaya dahil ederek çalışır. Pagerank'in formülü Denklem (2.1)'de verilmiştir:

$$S(V_i) = (1-\alpha) + \alpha \times \sum_{V_j \in in(V_i)} \frac{1}{out(V_j)} S(V_j) \tag{2.1}$$

Çizge teorisinde $G = (V, E)$ bir çizge olmak üzere V düğümleri E kenarları temsil eder. V_i düğümü için $in(V_i)$ diğer düğümlerden gelen bağlantıları, $out(V_i)$ diğer düğümlere giden bağlantıları ifade eder. α her iterasyonda bir önceki durumun etkisini belirlemek

için kullanılan sezgisel bir değerdir. Genellikle $0,85$ olarak kabul edilir.

TextRank çizge tabanlı anahtar kelime çıkarma yöntemlerinin ilki kabul edilmektedir. Yöntemde ilk olarak sözcük türleri bulunur ve bir metin içerisindeki isimler ve sıfatlar aday anahtar kelime olarak kabul edilir. Bir sonraki adım çizgenin oluşturulması aşamasıdır. Aday anahtar kelimeler çizgeye düğüm olarak eklenir. k sözcükten oluşan bir pencere boyutunda metinde bulunan kelimeleri içeren düğümlerin arasına bir kenar bağlantısı eklenir. Çizge modeli yönsüz ve basittir. TextRank'te başlangıçta metinde bulunan tüm kelimeler eşit öneme sahiptir varsayımı benimsenmiştir. TextRank'in çizge modeli bu varsayımla çizgede başlangıç düğüm ağırlıklarını 1 olarak alır. Daha sonra Pagerank algoritması çalıştırılır. Pagerank değerlerine göre en yüksek değere sahip n kelime alınır ve diğer kelimeler aday anahtar kelime listesinden çıkartılır. Aday anahtar kelimelerden metinde yan yana geçen kelimelerin değerleri toplanarak kelime öbeklerinin önem değerleri hesaplanır. Son adımda en yüksek değere sahip n adet kelime öbeği anahtar kelime olarak belirlenir (Mihalcea ve Tarau, 2004).

SingleRank TextRank'le aynı ön işlem adımlarını içerir ve çizge modeli benzerdir. SingleRank'in çizge modeline TextRank'ten farklı olarak kenar bağlantılarına ağırlık eklenir. Çizgede kenar ağırlığı birbiriyle bağlantı kurulan aday anahtar kelimelerin birlikte bulunma sayıları ile belirlenir. Daha sonra Pagerank algoritması uygulanır ve metinde yan yana geçen kelimelerin değerleri toplanarak kelime öbeklerinin önem değeri belirlenir. En yüksek değere sahip n kelime öbeği anahtar kelime olarak alınır. SingleRank metinde çok sayıda yan yana bulunan aday anahtar kelimeleri öne çıkartmaktadır (Wan ve Xiao, 2008).

PositionRank SingleRank ile aynı ön işlem, çizge oluşturma, düğüm sıralama ve son işlem adımlarını uygulamaktadır. PositionRank SingleRank'in çizge modelini her kelime için farklı başlangıç düğüm ağırlığı uygulayarak değiştirir. Çizgede düğümlerin başlangıç ağırlığı kelimelerin metin içerisindeki konumlarıyla ters orantılı olacak şekilde belirlenir. Metinde birden fazla geçen aday anahtar kelimelerin tüm pozisyonları hesaplamaya dahil edilir. Bu yöntem bir metnin başlangıcında bulunan kelimelerin daha değerli olduğu varsayımına dayanmaktadır (Florescu ve Caragea, 2017).

Vega-Oliveros ve arkadaşları anahtar kelime çıkarma problemi için farklı merkezilik ölçütlerin etkisini incelemiştir. Kullandıkları çizge modeli TextRank'te olduğu gibi pencere boyutuna göre belirlenmektedir. Aday anahtar kelimelerin önem değerlerini belirlemek için çizge merkezilik ölçütleri önce birer birer uygulanmıştır. Daha sonra merkezilik ölçütlerini birlikte kullanmanın etkisi ölçülmüştür. Bu amaçla kümeleme ve öznitelik seçme algoritmaları kullanılmıştır ve bu algoritmalar kendi kombinasyonlarını belirlemiştir. Kümeleme algoritmasının önerdiği kombinasyon, öznitelik seçme algoritmasının önerdiği kombinasyon ve merkezilik ölçütleri karşılaştırılmıştır. Kümeleme algoritmasına göre derece, arasındalık ve kümeleme merkeziliklerinin birlikte kullanımı en iyi sonucu vermiştir. Ancak öznitelik seçme algoritması kullanarak seçilen merkezilik ölçütleri diğerlerinden daha iyi sonuç elde etmiştir. Çalışma en iyi kombinasyonun derece, özvektör ve yapısal delikler merkeziliği ile elde edildiğini göstermektedir (Vega-Oliveros ve diğ., 2019).

TopicRank anahtar kelimeleri bulmak için metinlerin konu temsillerini kullanan çizge tabanlı bir yöntemdir. TopicRank ilk olarak metni ön işlemlerden geçirir ve isim tamlamalarını aday anahtar kelime olarak belirler. Daha sonra aday anahtar kelimeleri hiyerarşik kümeleme algoritması kullanarak konularına göre gruplandırır. Bir sonraki aşamada TopicRank yönsüz, ağırlıklı, tam çizge modeli oluşturur. Çizgede düğümler metindeki aday anahtar kelimelerin oluşturduğu konu gruplarını, kenarlar konu gruplarının arasındaki bağlantıları içerir. Çizgede kenar ağırlıkları konu grupları arasındaki anlamsal ilişkinin gücünü göstermektedir ve aday anahtar kelimelerin konumlarının arasındaki mesafeye göre hesaplanır. Yöntem düğümleri sıralamak için Pagerank algoritmasını kullanır. TopicRank son aşamada en önemli n konunun her birinden, metin içerisindeki ilk görülen aday anahtar kelimeyi anahtar kelime olarak belirler (Bougouin ve diğ., 2013).

SGRank metindeki aday anahtar kelimelerin istatistiksel özelliklerini kullanan çizge tabanlı bir yöntemdir. SGRank aday anahtar kelimeleri filtreleyerek azaltır. İlk olarak metindeki noktalamaları ve isim, sıfat, fiil dışında kalan tüm kelimeleri temizler. Daha sonra arama yapılan metin kısa, orta ve yüksek olarak belirlenen sınıflandırma etiketlerinden birisine atanır. Sınıflandırmaya göre bir metinden, metin kısa uzunluklu

ise 0, orta uzunluklu ise 2, yüksek uzunluklu ise 3 kez geçmeyen kelimeler aday anahtar kelime listesine dahil edilmez. Aday anahtar kelime listesinde bulunan diğer kelimelerden metinde yan yana bulunan kelime öbekleri aday anahtar kelime olarak belirlenir. Belirlenen aday anahtar kelimeler TF-IDF benzeri bir ölçütle sıralanır. Daha sonra aday anahtar kelimeler uzunluk ve ilk bulundukları konum bilgilerine göre önem değerleri yeniden hesaplanır. Bu aşamadan sonra aday anahtar kelimeler çizge ile modellenir ve Pagerank algoritması ile anahtar kelimeler bulunur (Danesh ve diğ., 2015).

MultipartiteRank TopicRank'te olduğu gibi konu tabanlı bir yöntemdir. MultipartiteRank TopicRank ile benzer aşamaları içerir. İlk olarak isim tamlamaları ile aday anahtar kelimeler oluşturulur ve hiyerarşik kümeleme algoritması ile konularına göre gruplandırılır. MultipartiteRank'in çizge modeli TopicRank'ten farklıdır. Çizgede konu grubu içinde yer alan her bir aday anahtar kelime için ayrı bir düğüm oluşturulur ve düğümlerin başlangıç ağırlığı 1 olarak kabul edilir. Kenar bağlantıları konu grubu dışında yer alan aday anahtar kelimeler arasında kurulur ve kenar ağırlığı aday anahtar kelimelerin konumlarının arasındaki mesafeye göre belirlenir. MultipartiteRank bu aşamadan sonra aday anahtar kelimeleri Pagerank algoritması ile derecelendirir. Son aşamada en yüksek değere sahip *n* sayıda aday anahtar kelime anahtar kelime olarak belirlenir (Boudin, 2018).

WeRank rastgele yürüyüşe dayalı anahtar kelime çıkarma yöntemidir. WeRank isim ve sıfatları aday anahtar kelime olarak kabul eder. Daha sonra kelime kelime, kelime konu ve konu konu ilişkilerini içeren bir çizge oluşturur. Kelime kelime çizgesi metinde bir pencere boyutunda yan yana bulunan kelimeler arasında kenar olacak şekilde kurulur. Kenar ağırlığı iki kelime arasında yer alan kelimelerin sayısıyla ters orantılı olacak şekilde belirlenir. Kelime konu çizgesi aynı konuya sahip kelimelerin arasında, konu konu çizgesi aynı kelimeyi içeren iki farklı konu arasında oluşturulur. Daha sonra oluşturulan bu hibrit çizgeden kelime vektörleri tespit edilir. Aday anahtar kelimeleri sıralamak için rastgele yürüyüş tabanlı bir yöntem benimsenmiştir. WeRank aday anahtar kelimeleri birleştirmek için GloVe modelini kullanır (Zhou ve diğ., 2022).

FLAKE bulanık mantık prensiplerini benimseyen bir yöntemdir. FLAKE her bir kelime

için metinde bulunduğu pozisyonu, bulunduğu cümledeki pozisyonu ve frekansını kullanarak bir ağırlık hesaplar. Yüksek ağırlığa sahip kelimeler aday anahtar kelime olarak belirlenir. Daha sonra WordNet kullanarak aday anahtar kelimelerden bulanık mantık prensiplerine dayalı bir çizge oluşturur. Oluşturulan çizgede kenar ağırlıkları kelimelerin birlikte bulunduğu cümle sayısını gösterir. FLAKE aday anahtar kelimeleri bulanık mantığa dayalı merkezilik ölçütleri (derece, yakındalık, arasındalık, Pagerank, HITS) ile sıralar (Jain ve diğ., 2022).

RDD-WRank kaba küme teorisine dayanan bir yöntemdir. RDD-WRank aday anahtar kelimeleri isim tamlamalarından oluşturur. Yöntem aday anahtar kelime listesini anlamsal benzerliklerine göre farklı sınıflara böler. Daha sonra farklı sınıfta yer alan her ikili aday anahtar kelime için kaba küme teorisine dayalı bir yaklaşımla bağlantısı güçlü aday anahtar kelimeleri belirler. Bu aşamadan sonra Wikipedia ile eğitilmiş Word2Vec modeli kullanılarak aday anahtar kelimelerin vektörleri tespit edilir ve Kmeans yöntemi ile bu vektörlerin önem değerleri bulunur. Daha sonra elde edilen bilgiler ile çizge oluşturulur ve Pagerank algoritması kullanılarak anahtar kelimeler tespit edilir (Zhou ve diğ., 2022).

ISKE cümleye dayalı çizge modeli kullanan bir anahtar kelime çıkarma yöntemidir. Bu yöntem komşu cümlelerde bulunan kelimelerin anlamsal olarak güçlü olduğu varsayımına dayanmaktadır. ISKE ilk adımda PositionRank'te önerilen pozisyon hesaplamasını kullanarak kelimelere değer atar. İkinci adımda her bir cümle, içerdiği kelimelerin ağırlıkları toplamı ile ağırlıklandırılır. Daha sonra cümlelerden oluşan bir çizge modeli kurulur. Çizgede kenar ağırlıkları cümlelerin arasında bulunan mesafenin harmonik ortalaması ile belirlenir. Pagerank ile cümle ağırlıkları bulunur ve cümlelerin ağırlıklarına göre cümlelerdeki kelimelerin değerleri güncellenir. Bu aşama değerler birbirine yakınsayana kadar devam eder ve anahtar kelimeler kelime ağırlıklarına göre bulunur(Chi ve Hu, 2021).

FNG-IE anahtar kelimeleri bulmak için Pagerank'e alternatif olarak sık bulunan yapılara dayalı bir yöntem önermiştir. Yöntemde en fazla 5 uzunluklu olacak şekilde kelime öbekleri aday anahtar kelime olarak belirlenir. Önerilen yöntem farklı uzunluklar için

belirlenen aday anahtar kelimeler için test edilmiştir. Testler en iyi sonucun en fazla 2 uzunluk için belirlenen aday anahtar kelimeler ile elde edildiğini göstermektedir (Tahir ve diğ., 2021).

Sung ve arkadaşı konu tabanlı anahtar kelime çıkarma yöntemi önermiştir. Yöntemde ilk olarak LDA ile metinlerden konu tespiti yapılmaktadır. Her bir konudan kelime öbeklerini en iyi temsil edecek aday anahtar kelimeler belirlenmiştir ve her bir konudan en yüksek değere sahip 30 kelime seçilmiştir. Daha sonra bu kelimeleri içeren isim tamlamaları bulunur. Konuların içerisinde bulunan aday anahtar kelimelerden hiyerarşik olarak ilgili konuyu en iyi temsil eden aday anahtar kelimeler seçilir. Bu amaçla yönlü bir çizge modeli oluşturulmaktadır. Çalışmada aynı metinde yan yana bulunan kelime öbeklerinin birbiriyle bağlantılı olduğu varsayılmaktadır. Bir sonraki aşamada hiyerarşik ilişkiye sahip kelime öbeklerinden alakasız olduğu belirlenen kelime öbekleri çıkartılır. Yöntemde kelime öbekleri düğümleri, kelime öbekleri arasındaki bağlantılar kenarları temsil eder. Kenar ağırlığı iki düğümün arasındaki bağlantının derecesi ile belirlenir. Sung ve arkadaşı aday anahtar kelimeleri sıralamak için ağırlıklı Pagerank, özvektör merkeziliği ve derece merkeziliği yöntemlerini kullanmıştır (Sung ve Kim, 2020).

Yeom ve arkadaşı istatistiksel ve çizge tabanlı modelleri birleştiren bir yöntem önermiştir. Yöntemde aday anahtar kelimeler isim, sıfat ve fiillerden (geçmiş zaman çekimli) seçilmektedir. İlk aşamada metinler ağırlıklandırılmış yönsüz çizge ile temsil edilir. Çizge modelinde düğümler kelimeleri, kenarlar birlikte bulunan kelimelerin bağlantılarını temsil eder. Kenar ağırlıkları birlikte bulunma sayılarıyla orantılı bir şekilde belirlenmektedir. Düğümler çizge derecelendirme algoritması ile sıralanır ve aday anahtar kelimeler kelime önem değerleriyle bulunur. İkinci aşamada aday anahtar kelimelerin önem değerini yeniden hesaplamak için C değeri kullanılır ve aday anahtar kelimeler bir kez daha sıralanır. Önerilen yöntemde aday anahtar kelimelerin önem değeri aday anahtar kelimelerin kelime sayısına bağlı olarak harmonik ortalamaya dayalı bir hesaplamayla bulunur. Çalışmada pozisyon bilgisinin kullanıldığı birlikte görünmeye dayalı çizge modellerinde anahtar kelime kayıplarının olduğu düşünülmektedir. Bu problemi çözmek için C değeri ile çözüm amaçlanmıştır (Yeom ve diğ., 2019).

Chen ve arkadaşları belirli kalıplar içeren alt çizgelere dayalı bir yöntem önermiştir. Yöntemde ilk olarak bir metinden etkisiz kelimeler ayıklanır ve kelimelerin kökleri bulunur. Bu kelimelerden bir çizge modeli oluşturulur. Çizgede kelimeler düğüm olarak temsil edilir, kenarlar aynı pencerede yan yana bulunan kelimelerin bağlantılarını gösterir. Kenar ağırlığı birlikte bulunma sayıları ile sağlanır. Ağırlığı belirli bir eşikten küçük olan kenarlar kesilir ve bağlantısız düğümler çizgeden silinir. Chen ve arkadaşları kalan düğümleri güçlü motifler olarak adlandırmıştır. Yöntemde motif sayısına karşılık gelen en yüksek değere sahip kelimeler anahtar kelime olarak belirlenmektedir (Chen ve diğ., 2019).

Duari ve arkadaşı anahtar kelime çıkarmak için sCAKE isimli çizge tabanlı parametresiz bir yöntem önermiştir. sCAKE isim ve sıfatları aday anahtar kelime olarak kabul eder. Çizgede aday anahtar kelimeler düğümlerle temsil edilir. Birbirini takip eden iki cümlede bulunan kelimeler arasında kenar bağlantıları oluşturulur. Böylece sCAKE pencere boyutu parametresini kullanmaz. Kenar ağırlığı birlikte bulunma sayısı ile sağlanır. sCAKE aday anahtar kelimeleri bir kelime için anlamsal ilişkiyi gösteren bir değer, pozisyon ve komşuları ile olan bağlantının gücünü ölçen bir değer kullanarak yapılan bir hesaplamayla sıralar. Bu çalışmada ayrıca LAKE isimli dilden bağımsız bir yöntem önerilmiştir. LAKE kelimelerin konumsal dağılımları ile ilgilenir. Birlikte bulunan kelimelerin arasındaki uzaklığın standart sapmasını kullanan bir filtreleme yöntemi ile aday anahtar kelimeleri belirler (Duari ve Bhatnagar, 2019).

Gömme tabanlı yöntemlerde önceden eğitilmiş kelime vektörleri kullanılarak anahtar kelimeler bulunur. Gömme tabanlı yöntemler aday anahtar kelimelerin belirlenmesi, aday anahtar kelimelerin vektör temsillerinin bulunması, sorgu yapılan metnin vektör temsilinin bulunması, metin ile aday anahtar kelimelerin vektör temsilleri arasındaki benzerliğin hesaplanması ve en yüksek benzerliğe sahip aday anahtar kelimelerin anahtar kelime olarak belirlenmesi aşamalarından oluşur.

Key2Vec bilimsel çalışmalardan anahtar kelimeleri çıkarmak için önerilmiş gömme tabanlı bir yöntemdir. Key2Vec aday anahtar kelimeleri bulmak için ilk olarak metne çeşitli filtrelemeler uygular. Daha sonra isim tamlamalarını aday anahtar kelime olarak

belirler. Key2Vec aday anahtar kelimelerin vektör temsillerini öğrenmek için Fasttext modelini kullanır. Fasttext ile kelimelerin arasındaki anlamsal ve morfolojik benzerliklerin bulunması amaçlanmıştır. Aday anahtar kelimeleri sıralamak için ağırlıklı Pagerank yöntemi uygulanmıştır (Mahata ve diğ., 2018).

EmbedRank gömme tabanlı yöntemlerden birisidir. EmbedRank aday anahtar kelimeleri bulmak için sözcük türlerinden yararlanır. İsim ve sıfat tamlamaları aday anahtar kelime olarak alınır. Daha sonra aday anahtar kelimelerin ve metnin ayrı ayrı vektörleri bulunur. Bu amaçla cümle vektörü (doc2vec veya sent2vec) kullanılmıştır. Son aşamada metin ile aday anahtar kelimelerin arasındaki kosinüs benzerliği hesaplanır ve sorgu yapılan metne en çok benzeyen aday anahtar kelimeler anahtar kelime olarak belirlenir (Bennani-Smires ve diğ., 2018).

SIFRank yöntemi önceden eğitilmiş dil modeli kullanan anahtar kelime çıkarma yöntemlerinden birisidir. SIFRank yönteminde ilk olarak isim tamlamaları bulunur ve bulunan isim tamlamaları aday anahtar kelime olarak alınır. Aday anahtar kelimeler için önceden eğitilmiş dil modeli olan ELMo ile vektör temsilleri bulunur. Bulunan vektörler SIF adı verilen bir cümle modeli ile güncellenir. Daha sonra metnin vektör temsili bulunur. Her bir aday anahtar kelimenin kelime vektörü ile metin vektörü arasındaki kosinüs mesafesi hesaplanır. Anahtar kelimeler metne en çok benzeyen n sayıda aday anahtar kelimeden oluşur (Sun ve diğ., 2020).

RVA makalelerden anahtar kelime çıkarmak için geliştirilmiş gömme tabanlı yöntemdir. RVA ilk olarak en fazla 3 uzunluklu olacak şekilde kelime öbeklerini belirler. RVA aday anahtar kelimeler için bazı kurallar belirlemiştir: Bir kelimenin aday anahtar kelime olabilmesi için rakam, özel karakter ve etkisiz kelimeleri içermemesi gerekmektedir. Aday anahtar kelimeler en az 2 uzunluklu kelimelerden oluşmalıdır. RVA aday anahtar kelimeleri GloVe modeli kullanarak belirler. Bunun için ilk olarak tam metinden GloVe modeli aracılığıyla kelime vektörleri bulunur. Daha sonra başlık ve özette bulunan aday anahtar kelimelerin vektör toplamları hesaplanır ve ortalaması bulunarak referans vektör olarak adlandırılan bir vektör temsili bulunur. Referans vektörü ile özet ve başlıkta yer alan aday anahtar kelimeler arasında kosinüs benzerliği hesaplanır ve böylelikle her bir

aday anahtar kelime için bir değer bulunmuş olur. 2 ve 3 uzunluklu aday anahtar kelimelerin önem değeri içerdiği her bir kelimenin değerleri toplanarak hesaplanır (Papagiannopoulou ve Tsoumakas, 2018b).

Khan ve arkadaşları gömme tabanlı bir yöntem önermiştir. Yöntemde ilk olarak Python diline ait CountVectorizer yöntemi kullanılmıştır. CountVectorizer ile yan yana bulunan *n* sayıda kelimeden oluşan aday anahtar kelimeler belirlenmiştir. Daha sonra BERT modeli ile metin ve aday anahtar kelimelerin vektör temsilleri bulunur. Son aşamada metin ile aday anahtar kelimelerin arasındaki kosinüs benzerliği hesaplanır ve metne en çok benzeyen aday anahtar kelimeler anahtar kelime olarak belirlenir (Khan ve diğ., 2022).

2.1.2. Denetimli Anahtar Kelime Çıkarma Yöntemleri

Denetimli öğrenme yaklaşımlarında anahtar kelimeler genellikle ikili sınıflandırma problemi olarak ele alınır. Geleneksel sınıflandırma algoritmaları ya da derin öğrenme yöntemleri kullanılarak aday anahtar kelimeler anahtar kelime ya da anahtar kelime değil şeklinde sınıflandırılır. İkili sınıflandırma probleminden başka etiketleme (sequence labeling), sıralamayı öğrenme (learning to rank) gibi farklı denetimli öğrenme yaklaşımları da bulunmaktadır. Denetimli anahtar kelime çıkarma yöntemleri makine öğrenmesi ve derin öğrenme yöntemleri olarak iki sınıfta incelenebilir.

KEA aday anahtar kelimelerin tespiti, özelliklerin çıkartılması, eğitim ve anahtar kelimelerin bulunması aşamalarından oluşur. KEA aynı cümle içerisinde yan yana bulunan kelime öbeklerini aday anahtar kelime olarak alır. Aday anahtar kelimelerin TF-IDF ve metin içerisinde ilk bulundukları pozisyon bilgisi özellikleri elde edilir. Eğitim aşamasında Naive Bayes sınıflandırıcısı kullanılmaktadır. Anahtar kelime bulma aşamasında eğitilen model sayesinde aday anahtar kelimeler anahtar kelime ya da anahtar kelime değil şeklinde sınıflandırılır (Witten ve diğ., 1999).

Nguyen ve arkadaşı bilimsel çalışmalardan anahtar kelime çıkarmayı amaçlayan bir yöntem geliştirmiştir. Önerilen yöntem anahtar kelimeleri çıkarmak için KEA ile aynı adımları içermektedir. Eğitim veri kümesinde aday anahtar kelimelerin TF-IDF ve metin

içerisinde ilk bulundukları pozisyon kullanılır. Bu özelliklerin yanı sıra KEA'dan farklı olarak aday anahtar kelimelerin morfolojik özellikleri (kısaltma durumu, son ek, sözcük türü) ve bilimsel çalışmaların yapısal özellikleri (bölüm bilgisi) kullanılmıştır (Nguyen ve Kan, 2007).

Maui aday anahtar kelimelerin seçimi ve makine öğrenmesi tabanlı filtrelemesi olmak üzere iki aşamadan oluşan anahtar kelime çıkarma yöntemidir. Maui aday anahtar kelimeleri metinde bulunan kelime öbeklerinden, başlangıcında ve bitişinde etkisiz kelimelerden birini içermeyen en fazla 3 uzunluklu olacak şekilde seçer. Maui eğitimde aday anahtar kelimelerin TF-IDF ve ilk pozisyon özelliklerini kullanır ve farklı özelliklerle genişletir. Bu özellikler anahtar kelime olabilirliği (aday anahtar kelimelerin eğitim kümesinde ne sıklıkta anahtar kelime olarak göründüğü), kelime öbeğinin kaç kelimeden oluştuğu, Wikipedia tabanlı anahtar kelime olabilirliği (bir kelime öbeğinin Wikipedia'da bağlantı içerme durumu), yayılım (bir metinde aday anahtar kelimenin ilk bulunduğu ve son bulunduğu pozisyonların arasındaki mesafe) ve Wikipedia'nın kaynak olarak kullanıldığı; istatistiksel düğüm derecesi, anlamsal ilintililik, ters Wikipedia bağlantısı özellikleridir. Maui anahtar kelimeleri sınıflandırmak için torbalı karar ağaçlarını kullanır (Medelyan ve diğ., 2009).

CeKE atıfa dayalı özelliklerin öğrenmeye dahil edildiği denetimli anahtar kelime çıkarma yöntemidir. CeKe aday anahtar kelimelerin TF-IDF, ilk pozisyon (ilk pozisyonun toplam kelime sayısına oranı) ve sözcük türleri özelliklerini kullanır. Bu özelliklere ilave olarak atıf durumu (bir kelime öbeğinin referans verilmiş ya da referans veren içeriklerde olup olmadığı), atıfa bağlı TF-IDF (atıf bağlantılarındaki TF-IDF değeri) ile TF-IDF ve pozisyon bilgilerinin modifiye edilmiş halleri (kelime öbeğinin başlangıçtan uzaklığı, kelime öbeğinin başlangıçtan uzaklığının belirli bir eşik değerine göre durumu, TF-IDF değerinin belirli bir eşik değerine göre durumu) özellikleri eğitime dahil edilmiştir. CeKE anahtar kelimeleri Naive Bayes sınıflandırıcısı kullanarak bulur (Caragea ve diğ., 2014).

WINGNUS akademik çalışmalar üzerinden anahtar kelime çıkarmayı amaçlayan bir yöntemdir. Aday anahtar kelimeler düzenli ifadeler yardımıyla bulunur. Denetimli

öğrenmede birçok özellik farklı kombinasyonlarla seçilerek eğitim aşamasına dahil edilmiştir. Eğitimde Naive Bayes algoritması kullanılmıştır. Yapılan testlerde anahtar kelime çıkarmak için en başarılı özelliklerin TF-IDF, kelime öbeklerini içeren kelimelerin sıklığı, ilk pozisyon ve kelime öbek uzunluğu olduğu vurgulanmıştır. Ayrıca anahtar kelimelerin akademik çalışmaların çeşitli bölümlerine (başlık, özet, giriş, ilgili çalışmalar, sonuç, ana metin, paragrafların ilk cümleleri) göre dağılımlarını incelemişler ve anahtar kelimelerin özet ile başlıkta akademik çalışmaların diğer bölümlerine göre daha yoğun olarak görüldüğü belirtilmiştir (Nguyen ve Luong, 2010).

Liu ve arkadaşları sıralı kalıp madenciliği (sequential pattern mining) ve konu modeli kullanarak denetimli öğrenmeye dayalı bir yöntem önermiştir. İlk olarak yöntemde LDA ile konularına göre kelime dağılımları bulunur. Daha sonra sıralı kalıp madenciliği ile sık kullanılan kalıplar elde edilir. Bir sonraki aşamada Naive Bayes algoritması ile anahtar kelime çıkarma problemi ikili sınıflandırma problemi olarak ele alınır. Aday anahtar kelimelerin TF-IDF, ilk pozisyon, genel özellikler (uzunluk vb.), derece merkeziliği ve çekirdek merkeziliği anahtar kelimeleri bulmak için kullanılmıştır (Liu ve diğ., 2019).

SurfKE denetimli anahtar kelime çıkarma yöntemlerinden birisidir. Denetimli öğrenme yöntemlerinde özellik mühendisliğinin düzgün bir şekilde yapılması önemlidir ve genellikle el ile yapılmaktadır. SurfKE özellik mühendisliğini otomatik olarak gerçekleştirmek için geliştirilmiştir. SurfKE ilk olarak aday anahtar kelimelerden SingleRank benzeri bir çizge modeli oluşturur. Daha sonra çizgeden özellikler rastgele yürüyüş (biased sampled random walk) yöntemiyle öğrenilir. Bir metinde bitişik konumda olan aday anahtar kelimeler birleştirilir. Eğitimde Gaus Naive Bayes modeli kullanılmıştır (Florescu ve Jin, 2018).

PCU-ICL harici bilgi kaynağı, kelime vektörü ve istatistiksel birçok özelliği birlikte kullanan anahtar kelime çıkarma yöntemidir. Her bir kelime için TF, IDF, TF-IDF, sözcük türleri; kelime öbekleri için uzunluk, bir önceki/sonraki kelime, bir önceki/sonraki kelime türü gibi özellikler elde edilir. Ayrıca özelliklere harici olarak İngilizce Wikipedia, IEEE bilgi kaynağı, farklı yöntemler ile elde edilen anahtar kelimeler (SGRank ve Textrank ile 20 anahtar kelime) ve kelime vektörü olarak önceden

eğitilmiş GloVe modeli dahil edilmiştir. Rastgele Orman (Random Forest) aday anahtar kelimeleri bulmak için kullanılmıştır (Wang ve Li, 2017).

Duari ve arkadaşları anahtar kelime çıkarmak için denetimli öğrenmeye dayalı bir yöntem önermiştir. Yöntemde ilk olarak metinde yer alan kelimelerden yönsüz ve ağırlıklı bir çizge oluşturulmaktadır. Çizgede düğümler kelimelerden oluşur ve iki ardışık cümledeki kelimelerin arasına kenar bağlantıları eklenmektedir. Çizgede kenar ağırlıkları, orijinal metinde bir kelimeyi temsil eden düğümün komşu düğümlerinin birlikte bulunma sayısına göre belirlenmektedir. Duari ve arkadaşları aday anahtar kelimelerin derece merkeziliği, özvektör merkeziliği, Pagerank, PositionRank, çekirdek ve kümeleme katsayısı değerlerini anahtar kelimeleri öğrenmek için seçmiştir. Öğrenme XGBoost ve Naive Bayes yöntemleri ile gerçekleştirilmektedir (Duari ve Bhatnagar, 2020).

Zhu ve arkadaşları bilimsel çalışmalardan anahtar kelimeleri çıkarmak için yapay sinir ağ tabanlı bir yöntem önermiştir. Yöntemde LSTM ve CRF yöntemleri kullanılmıştır. LSTM ile kelime temsilleri girdi olarak alınır ve metinde bulunan cümleler için özellikler çıkarılır. CRF ile cümlelerin etiket sıraları tahmin edilir (Zhu ve diğ., 2020).

Zhang ve arkadaşları anahtar kelimelerin sınıflandırılabilmesi için geleneksel LSTM yöntemine iki eklenti yapmışlardır. Önerilen TC-LSTM modeli bir kelimenin öncesi ve sonrasını her iki yönde hesaplamaya dahil eder. Verilen bir cümle için iki LSTM çalıştırılır. Birincisi cümlenin başından hedef kelimeye doğru, ikincisi cümlenin sonundan hedef kelimeye doğru işletilir (Zhang ve diğ., 2020).

Zhou ve arkadaşı bilimsel çalışmalardan anahtar kelime çıkarmak için MLM-CRF adlı çok seviyeli bir bellek ağı önermiştir. Çalışmada uzun bağlamsal bilgileri bulmak ve özellik mühendisliği ihtiyacını ortadan kaldırmak amaçlanmıştır. Uzun menzilli bağlamsal bilgileri bulmak için metin ve cümle seviyesinde bellek ağı kullanılmıştır. Anahtar kelime bulma problemi dizi etiketleme problemi (sequence labeling) olarak ele alınmıştır. Metinde komşu kelimelerin yapısal özelliklerini ve bir aday anahtar kelimenin anahtar kelime olup olmadığını belirlemek için CRF kullanılmıştır (Zhou ve diğ., 2020).

Alzaidy ve arkadaşları Bi-LSTM-CRF olarak adlandırdıkları çift yönlü LSTM ile CRF modelini birleştiren bir yöntem önermiştir. Yöntemde metin bağlamlarının anlamlarını ve komşu kelimelerin etiketleri arasındaki bağımlılıkları yakalamak amaçlanmıştır. Modelde ilk katmanda bir metnin (girdi) anlamsal yapısını öğrenmek için Bi-LSTM ağı kullanılmaktadır. Buradan elde edilen çıktı CRF katmanına aktarılır. Elde edilen dizinin etiket bağımlılıklarını kullanarak bir olasılık değeri elde edilir. En iyi diziyi bulmak için Viterbi algoritması kullanılmıştır. Bu çalışmada ayrıca her katmanın ayrı ayrı modele etkisi incelenmiştir (Alzaidy ve diğ., 2019).

charCNTN anahtar kelime çıkarma problemini Konvolüsyonel Sinir Ağlarını kullanarak ikili sınıflandırma problemi olarak çözen bir yöntemdir. Yöntem girdi olarak sadece metin ve kelimeleri alır. Önerilen modelde metinleri ve aday anahtar kelimelerin vektör temsillerini bulmak için bir metin modeli ve bir kelime modeli oluşturulmuştur. Bu iki modelden elde edilen çıktılar birleştirilmektedir (Lin ve Wang, 2019).

Basaldelal ve arkadaşları anahtar kelimeleri sınıflandırmak için Bi-LSTM yöntemini kullanmıştır. Çalışmada ilk olarak metindeki cümlelerden kelimeler elde edilir ve bu kelimelerin vektör temsilleri oluşturulur. Anahtar kelimeleri bulmak için kelime vektörleri Bi-LSTM modeline girdi olarak verilir (Basaldella ve diğ., 2018).

RankUp geri beslemeden yararlanan çizge tabanlı bir yöntemdir. Çizge tabanlı yaklaşımlarda metinle ilgisi olmayan kelimeler anahtar kelime olarak seçilebilmektedir. RankUp bu problemi çözmek için geliştirilmiştir. RankUp çizge tabanlı yaklaşımlarda kullanılan aday anahtar kelimelerin belirlenmesi, çizgenin oluşturulması ve düğümlerin sıralanması adımları içerir. Bu aşamalardan sonra hata tespit yaklaşımı ile her bir düğüm için doğru sıralamada olup olmadığı analiz edilir. Bu amaçla TF-IDF, IDF ve kümeleme ayrı ayrı uygulanmıştır ve test sonuçları karşılaştırılmıştır. Elde edilen sonuçlara göre kenar ağırlıkları geri beslemeye dayalı bir yaklaşımla güncellenir. Düğüm sıralama aşamasında TextRank ve RAKE tercih edilmiştir (Figueroa ve diğ., 2018).

2.2. Skyline Sorguları

Skyline sorguları veri madenciliği, veritabanı sistemleri ve veri yapıları alanlarında kullanılan bir optimizasyon işlemidir. Skyline sorguları ile veri kümesinde yer alan nesneler arasında ilişkiler belirlenebilmektedir. Skyline sorguları Skyline operatörü ile gerçekleştirilir.

Skyline sorguları, kullanıcıların çok sayıda veri içeren bir veri kümesinden hızlı seçim yapabilmelerini sağlar (Borzsony ve diğ., 2001). Skyline çok boyutlu verilerden diğer veri noktaları tarafından üstünlük sağlayamadığı veri noktalarıdır. Skyline operatörüne göre $a1$ ve $a2$ iki veri olmak üzere, $a1$ verisi tüm özelliklerde $a2$ verisinden daha kötü değilse ve en az bir özellikte $a2$'den daha iyiyse, $a1$ verisinin $a2$ verisine üstünlük sağladığı söylenebilir ve $a2$ verisi Skyline noktası olamaz.

d boyutlu bir nesne için p ve q, p bütün boyutlarda en az q kadar iyiyse ve en az bir boyutta daha iyiyse p q'ya baskın gelir. Baskınlıkla ilgili formül Denklem 2.2'de verilmiştir. >= en az iyi olma durumunu, > daha iyi olma durumunu gösterir.

$$\begin{aligned} p \prec q \leftrightarrow \forall i \in \{1, 2, \ldots, d\} : p_i \geq q_i \\ \wedge \exists j \in \{1, 2, \ldots, d\} : p_i > q_i \end{aligned} \tag{2.2}$$

Skyline operatörünün formülü Denklem 2.3'de gösterilmiştir.

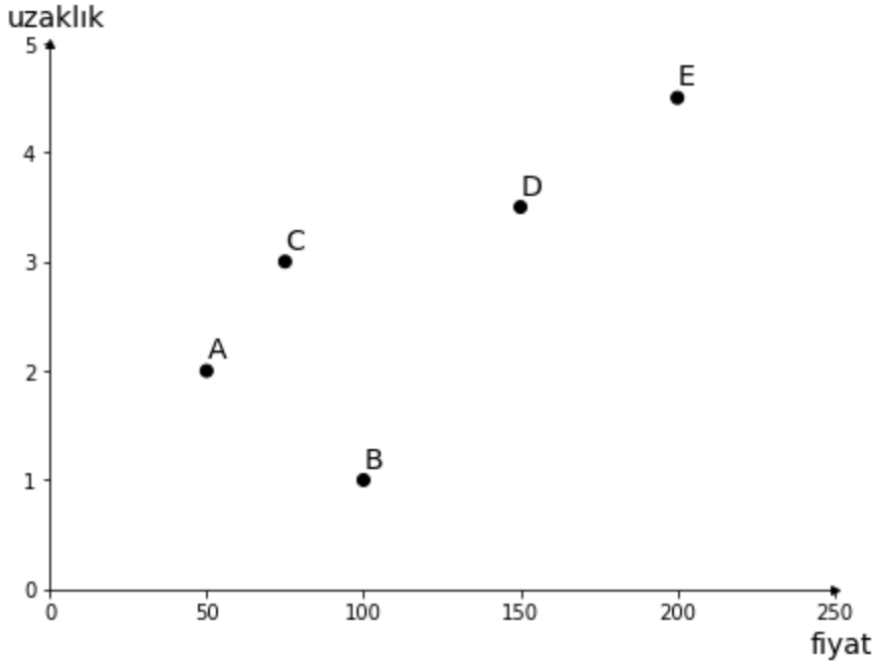
$$S_p = \{p \mid p \in P \wedge \nexists q \in P : q \prec p\} \tag{2.3}$$

Skyline noktalarının sağladığı kurallar şu şekildedir:

- Bir p verisi bir başka veri tarafından domine edilirse p Skyline noktası olamaz.
- p bir ya da birden fazla veriyi domine ederse, domine edilen veriler Skyline noktası olamaz.
- Skyline noktası olarak alınan verilerden herhangi biri diğer verileri hiçbir şekilde domine edemez.

Skyline sorgularıyla ilgili en çok karşılaşılan problemlerden birisi otel tercih problemidir. Bu problemde otel seçimi ile ilgili iki kıstas bulunmaktadır: denize olan

mesafe ve fiyat. Amaç tüm seçenekler arasından daha ucuz ve denize daha yakın olan otelleri belirleyebilmektir. Bu problem için örnek bir senaryo Şekil 2.2'de verilmiştir. Örnekte *A* otelinin fiyatı günlük 50 lira ve denize uzaklığı 2 kilometre, E oteli için ise



Şekil 2.2. Skyline Örneği

200 lira ve $4,5$ kilometredir. *E* otelinin fiyatı diğerlerinden pahalı ve denize uzaklığı herhangi bir otelden iyi olmadığı için baskın geldiği herhangi bir veri yoktur. *C* hem *D* hem de *E*'ye göre ucuz ve denize daha yakın bir otel olduğu için bu seçeneklere baskın gelmektedir. *C B*'ye göre daha ucuz ancak denize daha uzak mesafede olduğundan ikisi arasında bir baskınlık söz konusu değildir. Ancak *A* oteli *C* oteline göre iki özellik açısından da daha iyi olduğu için baskın gelir. *A* ile *B* kıyaslandığında biri daha ucuzken diğeri denize daha yakındır. Bu sebeple aralarında herhangi bir baskınlık yoktur. *A* ve *B* otelleri diğer otel verileriyle kıyaslandığında bu iki otel önerisinin en az bir özelliği diğerlerinden daha iyidir ve bundan dolayı bu iki veriye baskın gelen bir veri bulunmamaktadır. Skyline operatörü bu örnekte *A* ve *B* otellerini öneri olarak belirler.

Skyline operatörü öneri sistemleri (Zeng ve diğ., 2019), akan veri madenciliği (Ren ve diğ., 2019; Tao ve Papadias, 2006), veri sınıflandırma (Mutlu ve diğ., 2019) ve veri kümeleme (Mutlu ve Goz, 2022) gibi birçok problemde kullanılmıştır. Ancak Skyline

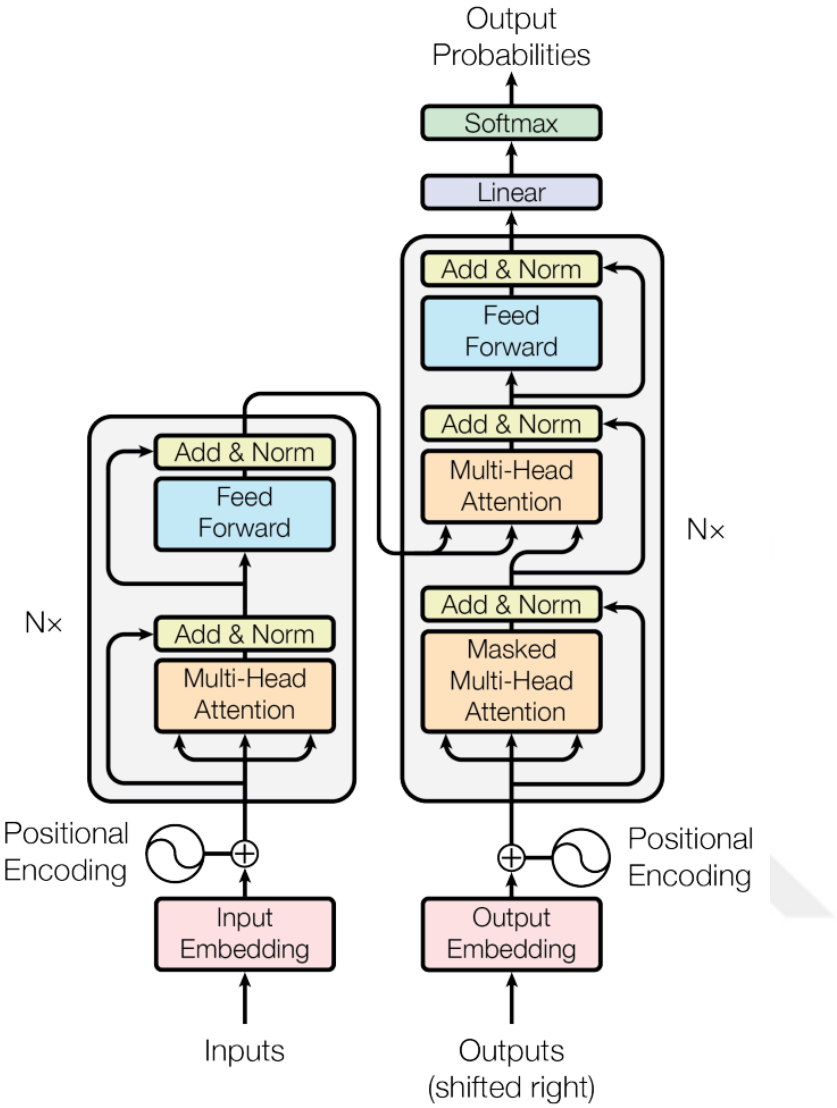
operatörünün anahtar kelime çıkarma probleminde kullanılmış bilinen bir örneği bulunmamaktadır.

2.3. Transformers

DDİ'de kelimelerin ve metinlerin anlamlı temsillerini oluşturmak önemli bir problemdir. Bu problemin çözümü için önceleri metin içinde kelimelerin frekansına dayalı bir yaklaşım olan kelime çantası kullanılmıştır (Wallach, 2006). Kelime çantası yaklaşımında kelimelerin anlamları doğru şekilde temsil edilmez. Daha sonra word2vec gibi kelime vektörleri ortaya çıkmıştır (Church, 2017). word2vec vektör temsillerini kelimelerin sabit bir pencere boyutunda yan yana bulunan diğer kelimelere göre benzerliklerini hesaplayarak bulur. Bu tip modeller her bir kelime için öğrendikleri anlamı bir düzlemde temsil eden vektörler üretir. Elde edilen vektör temsilleri farklı anlamları taşıyan kelimeler için değişmez. Bir kelime farklı anlamları olsa bile tek bir vektör ile temsil edilir. Bu nedenle bu temsillerde çok anlamlılık problemi bulunmaktadır. 2018 yılında ELMo modeli önerilmiştir. ELMo modeli iki katmanlı Bi-LSTM modeli kullanarak eğitilmiştir. ELMo Bi-LSTM ile giriş verilerini iki yönde öğrenmeyi sağlar (Peters ve diğ., 2018).

Son yıllarda Transformers modelleri popüler hale gelmiştir. Doğal dil işleme, görüntü işleme ve bilgisayarda görü gibi birçok problemde kullanılmaktadır. Transformers Attention (Dikkat) mekanizmasını kullanan bir modeldir. LSTM uzun süre bilgi hatırlama ve kullanma gücüne sahiptir. Ancak LSTM modellerinde girdi miktarı arttıkça hatırlama problemi çıkabilmektedir. Attention mekanizması bu problemi çözmektedir. Attention mekanizması yeterli hesaplama gücü ve kaynağı sağlandığında metin içerisinde yer alan tüm bağlamı öğrenebilir. Şekil 2.3 Transformers modellerinin yapısını göstermektedir (Vaswani ve diğ., 2017).

Metinden bilgi çıkarma amaçlı kullanılan bir Transformers modelinde ilk olarak metinlerden girdi dizisi (kelime dizisi) elde edilir. Transformers ağ yapısının girdi katmanına (Input Embedding) gönderilir. Bu katmanda önceden eğitilebilen özel bir vektör tablosuyla girdi dizisi için vektör temsilleri bulunur. Pozisyon kodlama (Position Encoding) ile girdi dizisine ait sıranın model tarafından anlaşılmasını sağlayan bir bilgi



Şekil 2.3. Transformers Model Yapısı

eklenir. Attention modülü girdi vektörleri arasındaki ilişkileri öğrenmeyi amaçlar ve Feed Forward modülü bu ilişkileri kullanarak girdi vektörlerinin özelliklerini daha da iyileştirmeyi amaçlar. Katmanların tümünden geçirilen vektörler çıktı katmanına gönderilir. Bu katman çıktı sınıfındaki özel bir gömme vektörü tablosuna göre vektör temsilleri bulur ve bu vektör temsilleri dil modelinin çıktısı olarak kullanılır. Add & Norm aşamasında modelin işlemlerinin sonucu olarak üretilen vektörleri bir araya getirir ve bu vektörleri normalize eder. Masked Multi-Head Attention aşamasında Attention modelin daha önceki çıktı öğelerini hesaba katmaması maskelenerek gerçekleştirilir (Vaswani ve diğ., 2017).

Transformers modellerine örnek olarak BERT, GPT, T5, XLNet ve MPNet gösterilebilir.

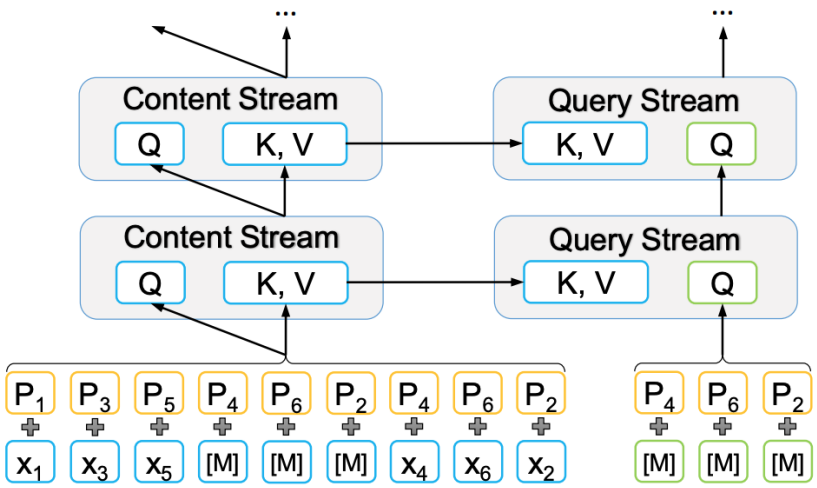
BERT 2018 yılında Google tarafından geliştirilen kodlayıcı kod çözücü yapısına sahip bir dönüştürücü modelidir (Devlin ve diğ., 2018). BERT mimarisi ince ayar ile eğitilerek birçok farklı DDİ probleminde kullanılmaktadır. Örneğin BioBERT biyomedikal metinler için eğitilmiş bir modeldir. BERT tabanlı modeller son yıllarda birçok problemin çözümünde başarılı bir şekilde uygulanmaktadır (Wang ve Kuo, 2020).

BERT modeli temsilleri elde etmek için değerlendirmeyi sağdan sola ve soldan sağa olacak şekilde çift yönlü yapar. BERT; MLM ve NSP adı verilen iki teknik kullanılarak eğitilmektedir. MLM tekniğinde cümlelerdeki kelimelerin yüzde 15'i seçilir. Bu kelimelerden yüzde 80'i maskelenir ve bir etiket ile ifade edilir. Diğer kelimelerin yarısı farklı bir kelime ile değiştirilir. Kalan diğer kelimeler orijinal halleriyle bırakılır. Bu teknikte maskelenen kelimeler tahmin edilir. NSP tekniğinde ise cümle çiftleri arasında karşılaştırma yapılır. Her bir cümle çifti için cümlelerden birisinin diğer cümlenin devamı olan bir cümle olup olmadığı tahmin edilir. Bu amaçla eğitime dahil edilen cümle çiftlerinden rastgele seçilen devam cümlelerinin yarısı ilk cümlelerle yer değiştirilir.

BERT, maskelenmiş konumlar arasındaki bağımlılığı ihmal eder. Bu nedenle ince ayar yapılmak istendiğinde problem çıkabilmektedir.

XLNet BERT'in olumlu ve olumsuz özelliklerini dikkate alarak AR modelini önermiştir. XLNet orijinal dönüştürücü mimarisini kullanır ve ince ayar ile problem olabilecek bazı uyumsuzlukları ortadan kaldırır. XLNet çift yönlü değerlendirmeyi özel bir PLM modeli ile yapar (Yang ve diğ., 2019). XLNet bir cümlenin tam konum bilgisinden yararlanmaz ve bu nedenle ince ayar yapıldığında konum uyuşmazlığı problemi ortaya çıkabilir.

MPNet XLNet ve BERT modellerinin olumlu yönlerinden esinlenip kısıtlamalarını çözüme kavuşturmayı amaçlayan Microsoft tarafından geliştirilmiş bir modeldir. MPNet MLM ve PLM yöntemlerini birlikte kullanır. Maskelenmiş belirteçler için PLM modelini tercih eder. Ayrıca bütün belirteçlerin konum bilgilerini saklar. MPNet 160GB'nin üzerinde veri ile eğitilmiştir ve aynı zamanda birçok farklı yöntem kullanarak ince ayar yapılmıştır. Şekil 2.4 MPNet ağ yapısını göstermektedir (Song ve diğ., 2020).



Şekil 2.4. MPNet Modelinin Yapısı

all-mpnet-base-v2 bir cümle dönüştürücü modelidir (HuggingFace, 2022). Cümleleri ve paragrafları 768 boyutlu vektör uzayına dönüştürür. Microsoft'un MpNet modeline ince ayar yaparak oluşturulmuştur (fine-tuning). all-mpnet-base-v2 1 milyon cümle çifti ile eğitilmiştir. 2015-2018 yılları arası Reddit'te yapılmış yorumlar, akademik makalelerin özet ve atıflarından oluşturulmuş S2ORC veri kümesi, WikiAnswers'da yer alan soru cevaplar ve bunun gibi birçok veri kaynağından yararlanılmıştır. İnce ayar yapılırken kümede yer alan olası her bir cümle çiftinden kosinüs benzerliği hesaplanmıştır ve doğru çiftlerle karşılaştırılarak çapraz entropi kaybı uygulanmıştır.

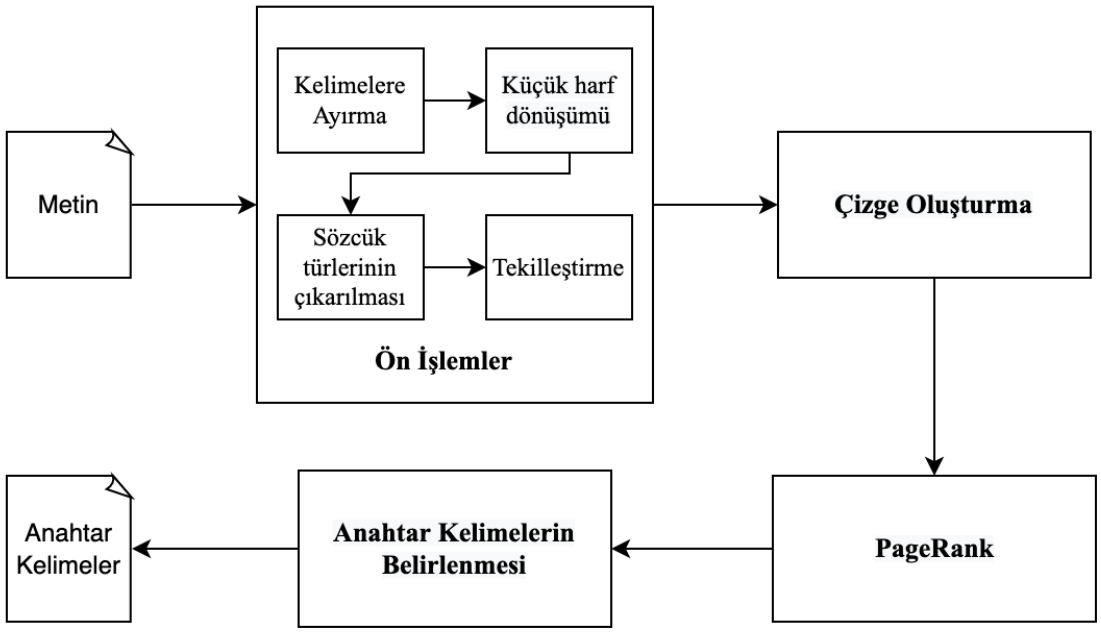
Bu çalışma kapsamında geliştirilen SkyWords ve SkyRank yönteminde all-mpnet-base-v2 modeli kullanılmıştır. Bu yöntemlerde all-mpnet-base-v2 modeli ile aday anahtar kelimelerin ve metinlerin vektör temsilleri oluşturulmaktadır. Daha sonra kosinüs benzerliği ile metne en çok benzeyen aday anahtar kelimeler belirlenmektedir.

3. GELİŞTİRİLEN YÖNTEMLER

Tez kapsamında anahtar kelime çıkarımı için 3 yeni yöntem önerilmiştir. Bu bölümde geliştirilen yöntemler açıklanmaktadır.

3.1. MGRank

Geliştirilen ilk yöntemde çizge tabanlı bir yaklaşım benimsenmiştir ve MGRank olarak adlandırılmıştır. Çizge tabanlı yöntemlerde pencere boyutu parametresinin belirlenmesi gerekmektedir. Parametrenin değeri anahtar kelime çıkarma yönteminin performansını etkiler. MGRank ile pencere boyutu parametresinin kullanımına gerek kalmaz. MGRank'in çizge modelinde düğümler ve kenarlar ağırlıklıdır. Çizgede iki düğümün arasına birden fazla kenar bağlantısı kurulabilir. MGRank tam çizge modeli kullanır. Yöntem ön işlemler, çizge oluşturma, aday anahtar kelimelerin sıralanması ve anahtar kelimelerin belirlenmesi olmak üzere 4 ana aşamadan oluşur. Şekil 3.1 MGRank'in genel çalışma mantığını göstermektedir.



Şekil 3.1. MGRank Yönteminin Genel Yapısı

3.1.1. Ön İşlem

Veri ön işlem adımı, aday anahtar kelimeleri belirlemek için uygulanmaktadır. Aday anahtar kelimeler bir algoritmanın etkinliğini ve başarısını doğrudan etkilemektedir. Bu aşamada bir metnin anlam ifade etmeyen, gürültülü ifadelerden arındırılarak önemli bölümleri belirlenir ve aday anahtar kelime olarak alınır.

MGRank veri ön işlemede bir metin için metni en küçük birimlerine dönüştürme (tokenization) (1.adım), büyük küçük harf dönüşümü (2. adım), tekilleştirme (3. adım) ve sözcük türlerinin bulunması adımlarını (4. adım) uygular. Anahtar kelime çıkarma yöntemleri ve diğer DDİ'nin birçok problemi için geliştirilmiş yöntemlerde sorgu yapılan metin en küçük birimlerine dönüştürülerek çözüm üretilmektedir (Campos ve diğ., 2020; Florescu ve Caragea, 2017; Mihalcea ve Tarau, 2004). Bu nedenle ilk olarak MGRank bir metni en küçük birimlerine dönüştürmektedir. Bu işlemle birlikte bir metin yalın hale getirilmiş olur ve metin gereksiz boşluklardan arındırılmış olur. Daha sonra elde edilen metin birimlerine büyük küçük harf dönüşümü uygulanır. Bir sonraki aşamada çoğul ifadeler içeren birimler tekil hallerine dönüştürülür.

MGRank'in uyguladığı bu ön işlem adımları metin madenciliğinde bir algoritmanın tahmin gücünü arttırmak ve anlamca birbiriyle ilintili kelimeleri bir araya getirmek amacıyla kullanılmaktadır (Hickman ve diğ., 2022). Ön işlem adımlarıyla aday anahtar kelimelerin sayısı azalır ve birbiriyle bağlantılı aday anahtar kelimelerin etkisi arttırılır. MGRank bir sonraki aşamada aday anahtar kelimeleri belirlemek için sözcük türlerini belirler. Bulunan sözcük türlerinden isim ve sıfat olanlar aday anahtar kelime olarak belirlenir. Anahtar kelimelerin genellikle isim ve sıfat tamlamalarından oluştuğu birçok çalışmada vurgulanmaktadır (Bougouin ve diğ., 2013; Mihalcea ve Tarau, 2004).

Tablo 3.1 bir örnek cümle için MGRank'in uyguladığı ön işlem adımlarını göstermektedir. Tabloda 2. ve 3. adımda bir önceki adıma göre oluşan farklılıklar koyu renkle gösterilmiştir. 4. adımda kelimelerin yanında yer alan ifadeler sözcük türlerini göstermektedir. Tabloda *J*, *N*, *V*, *R*, *I* ve *W* ile başlayan ifadeler sırasıyla sıfat, isim, fiil, zarf, edat ve belirteçleri temsil etmektedir. 5. adım MGRank'in bulduğu aday anahtar kelimeleri göstermektedir. Tekrar eden aday anahtar kelimeleri MGRank çizgede kenar

bağlantıları kurmak için kullanır. Bu nedenle tekrar eden aday anahtar kelimeler bu listede saklanır. Ayrıca aday anahtar kelimelerin metin içerisindeki bulunduğu sıra, konum bilgilerini kullanmak için saklanmaktadır.

Tablo 3.1. Örnek Metin ve Ön İşlem Adımları

Örnek metin	Collaborative document processing has been addressed by many approaches so far, most of which focus on document versioning and collaborative editing.
1. adım	["Collaborative", "document", "processing", "has", "been", "addressed", "by", "many", "approaches", "so", "far", ",", "most", "of", "which", "focus", "on", "document", "versioning", "and", "collaborative", "editing"]
2. adım	["**collaborative**", "document", "processing", "has", "been", "addressed", "by", "many", "approaches", "so", "far", ",", "most", "of", "which", "focus", "on", "document", "versioning", "and", "collaborative", "editing"]
3. adım	["collaborative", "document", "processing", "has", "been", "addressed", "by", "many", "**approach**", "so", "far", ",", "most", "of", "which", "focus", "on", "document", "versioning", "and", "collaborative", "editing"]
4. adım	[("collaborative", "JJ"), ("document", "NN"), ("processing", "NN"), ("has", "VBZ"), ("been", "VBN"), ("addressed", "VBN"), ("by", "IN"), ("many", "JJ"), ("approach", "NN"), ("so", "RB"), ("far", "RB"), (",", ","), ("most", "JJS"), ("of", "IN"), ("which", "WDT"), ("focus", "NN"), ("on", "IN"), ("document", "NN"), ("versioning", "NN"), ("and", "CC"), ("collaborative", "JJ"), ("editing", "NN")]
5. adım	["collaborative", "document", "processing", "many", "approach", "focus", "document", "versioning", "collaborative", "editing"]

MGRank örnek metnin aday anahtar kelimelerini *collaborative*, *document*, *processing*, *many*, *approach*, *focus*, *versioning*, *editing* olarak belirlemiştir.

3.1.2. Çizge Oluşturma

MGRank çizgeyi ağırlıklı, çok kenarlı tam çizge modeline dayalı olarak oluşturur. Çizgede düğümler benzersiz aday anahtar kelimelerden oluşur. Kenarlar her bir aday anahtar kelime için metinde aday anahtar kelimeyi takip eden diğer aday anahtar kelimelerle arasında oluşturulur. MGRank birden fazla yerde bulunan aday anahtar

kelimenin her biri için kendisinden sonra gelen aday anahtar kelimeler ile kenar bağlantısı oluşturur. MGRank iki düğüm arasında birden fazla kenar bağlantısına izin verir.

MGRank'te düğüm ağırlıkları aday anahtar kelimelerin başlangıçtaki önem değerlerini göstermektedir. MGRank'in düğüm ağırlıklandırması için bir metinde önce görünen aday anahtar kelimelerin, sonra görünen aday anahtar kelimelere göre daha önemli olduğunu varsayımı benimsenmiştir. Çizgede her bir aday anahtar kelime için başlangıç düğüm ağırlığı bir aday anahtar kelimenin metinde bulunduğu pozisyonun çarpmaya göre tersi hesaplanarak bulunur. Bir aday anahtar kelime metinde birden çok kez geçiyorsa sadece ilk bulunduğu pozisyon bilgisi hesaplamaya dahil edilir.

MGRank'in kenar ağırlıklandırması birbirine yakın konumda bulunan aday anahtar kelimeleri bağlayan kenarlara birbirinden uzak konumda bulunanlara göre yüksek ağırlık olacak şekilde belirlenmektedir. MGRank kenar ağırlıklarını birbiriyle kenar bağlantısı kurulacak her iki aday anahtar kelime için aralarındaki konumsal mesafe ile ters orantılı olacak şekilde belirler. Çizgede her bir düğümün kenar ağırlıkları kenarlarının ağırlıkları toplamı 1 olacak şekilde normalize edilmektedir. Konum bilgileri orijinal metin yerine aday anahtar kelimelerin konum bilgilerine göre alınır. Mihalcea ve arkadaşının yaptığı çalışmada kenar tipinin yönlü ya da yönsüz olmasının anahtar kelimelerin sıralamasını etkilemediği belirtilmiştir (Mihalcea ve Tarau, 2004). Bu nedenle MGRank'te çizge yönsüz çizge olarak modellenmiştir.

MGRank'in çizge modeli 5'li demet (tuple) yapısıyla şu şekilde temsil edilebilir: Çizge G'nin demet yapısı $G = (V, E, r, \ell_v, \ell_e)$ şeklindedir. V, aday anahtar kelimeleri temsil eden sonlu bir düğüm kümesidir. E bir kenarlar kümesidir. $r : E \rightarrow \{\{x, y\} : x, y \in V\}$ kenar bağlantılarını gösteren ifadedir. ℓ_v ve ℓ_e sırasıyla G'nin düğümleri ve kenarlarının ağırlıklarını gösterir.

Denklem 3.1 bir çizgede düğümlerin başlangıç düğüm ağırlığının belirlendiği hesaplamayı göstermektedir. Denklemde P_{v_i}, v_i düğümünün aday anahtar kelime

listesindeki ilk konum değeridir.

$$w_{v_i} = \frac{1}{P_{v_i}} \tag{3.1}$$

Denklem 3.2 çizgede bir kenarın ağırlık değerlerinin hesabını göstermektedir. Çizgede v_i ve v_j herhangi iki düğüm olmak üzere, d_{v_i,v_j}, v_i ve v_j düğümlerinin arasındaki konumsal mesafeyi gösterir. Denklem 3.3 lcm_{v_i} ifadesinin hesaplamasını göstermektedir. Çizge yönsüz olarak modellendiği için denklemde hesaplamaya $d_{v_i,v_t} > 0$ koşulu eklenmiştir. Bu ifade metinde bir aday anahtar kelimenin temsil ettiği v_t düğümünün bir diğer aday anahtar kelime v_i düğümünden sonra görüldüğünü belirtmektedir.

$$e_{v_i,v_j} = \frac{\frac{lcm_{v_i}}{d_{v_i,v_j}}}{\sum_{v_t \in adj(v_i):d_{v_i,v_t}>0} \frac{lcm_{v_i}}{d_{v_i,v_t}}} \tag{3.2}$$

Denklem 3.3'te lcm_{v_i}, metinde v_i aday anahtar kelimesinden sonra gelen tüm aday anahtar kelimelerin konumsal uzaklıklarının en küçük ortak katının ifadesidir.

$$lcm_{v_i} = lcm(d_{v_i,v_k} : v_k \in adj(v_i), d_{v_i,v_k} > 0) \tag{3.3}$$

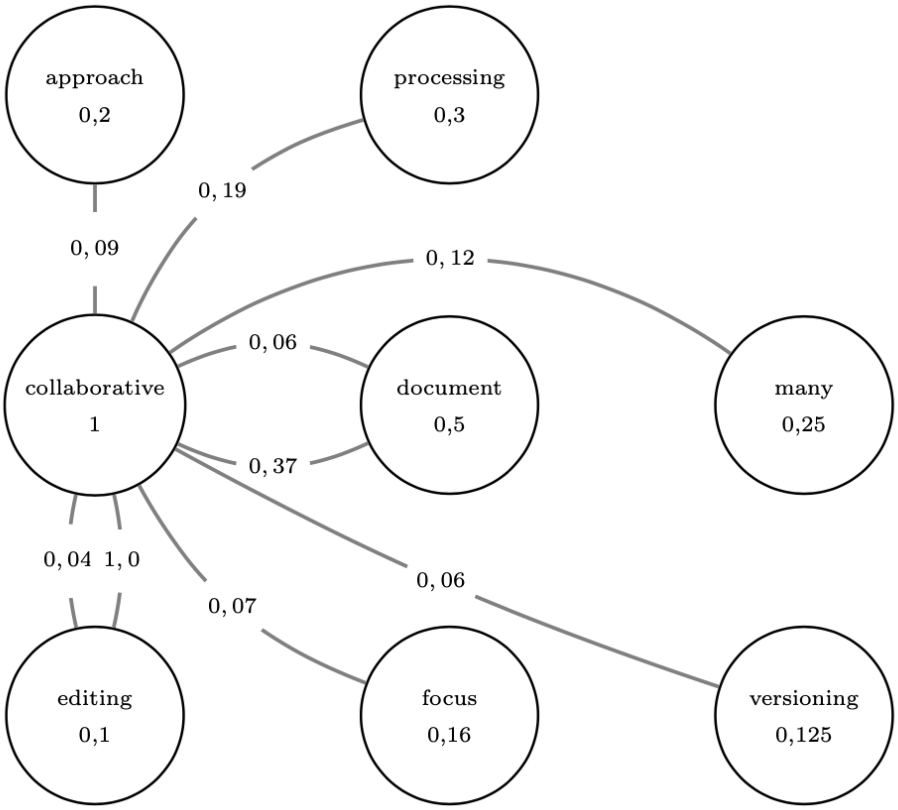
MGRank Tablo 3.1'de verilen örnek metnin çizge modelini Şekil 3.2'de verildiği gibi modellemektedir. Örnek çizgede şeklin anlaşılabilirliğini arttırmak için kenar bağlantıları ilk aday anahtar kelime olarak belirlenen *collaborative* kelimesine göre gösterilmiştir.

Örnek metne göre *collaborative* kelimesinden sonra gelen aday anahtar kelimeler sırasıyla *document*, *processing*, *many*, *approach*, *focus*, *document*, *versioning*, *editing* şeklindedir. *collaborative* kelimesinin diğer kelimelere göre konumsal mesafeleri sırasıyla *1, 2, 3, 4, 5, 6, 7, 9* şeklindedir. Denklem 3.2 ve Denklem 3.3'e göre bu mesafelerin ekok değeri hesaplanmaktadır. Bu mesafelerin ekok değeri *1260*'tır.

collaborative aday anahtar kelimesi ile *collaborative*'den sonra ilk görülen *document* aday anahtar kelimesinin arasındaki kenar ağırlığı şu şekilde hesaplanır:

$$\frac{1260/1}{1260/1+1260/2+1260/3+1260/4+1260/5+1260/6+1260/7+1260/9} = 0,37$$

collaborative aday anahtar kelimesi ile metinde son görülen *document* aday anahtar kelimesi arasındaki kenar ağırlığı şu şekilde hesaplanır:



Şekil 3.2. Örnek Çizge

$$\frac{1260/6}{1260/1+1260/2+1260/3+1260/4+1260/5+1260/6+1260/7+1260/9} = 0,06$$

collaborative metinde 2 defa bulunduğu için kenar ağırlıkları toplamı *2* olur. Metinde ikinci kez görülen *collaborative* kelimesinden sonra sadece *editing* kelimesi bulunmaktadır. Bundan dolayı bu iki kelime arasındaki kenar bağlantılarından birisinin ağırlığı *1* olmaktadır.

3.1.3. Aday Anahtar Kelimelerin Sıralanması

Aday anahtar kelimeleri sıralamak için Pagerank algoritması kullanılmıştır. Bu amaçla çizgede kenar ve düğüm ağırlıklarını hesaplamaya dahil eden yeni bir hesaplama yöntemi kullanılmıştır. Denklem aday anahtar kelimelerin derecelendirilmesi için kullanılan formülü göstermektedir.

$$S(v_i) = (1-\alpha) \times \ell_{v_i} + \alpha \times \sum_{v_j \in adj(v_i)} \frac{\ell_{e_{v_i,v_j}}}{\sum_{v_k \in adj(v_j)} \ell_{e_{v_i,v_j}}} S(v_j) \tag{3.4}$$

Denklemde α değeri farklı yöntemlerde kabul gören 0,85 olarak alınmıştır. Denklemde verilen hesaplama, aday anahtar kelimelerin önem değerlerinin değişimi arasındaki fark 0,001'den az olana kadar yinelemeli olarak yapılır. Bu aşamanın sonunda her bir aday anahtar kelime için önem değerini gösteren bir değer elde edilir.

3.1.4. Anahtar Kelimelerin Belirlenmesi

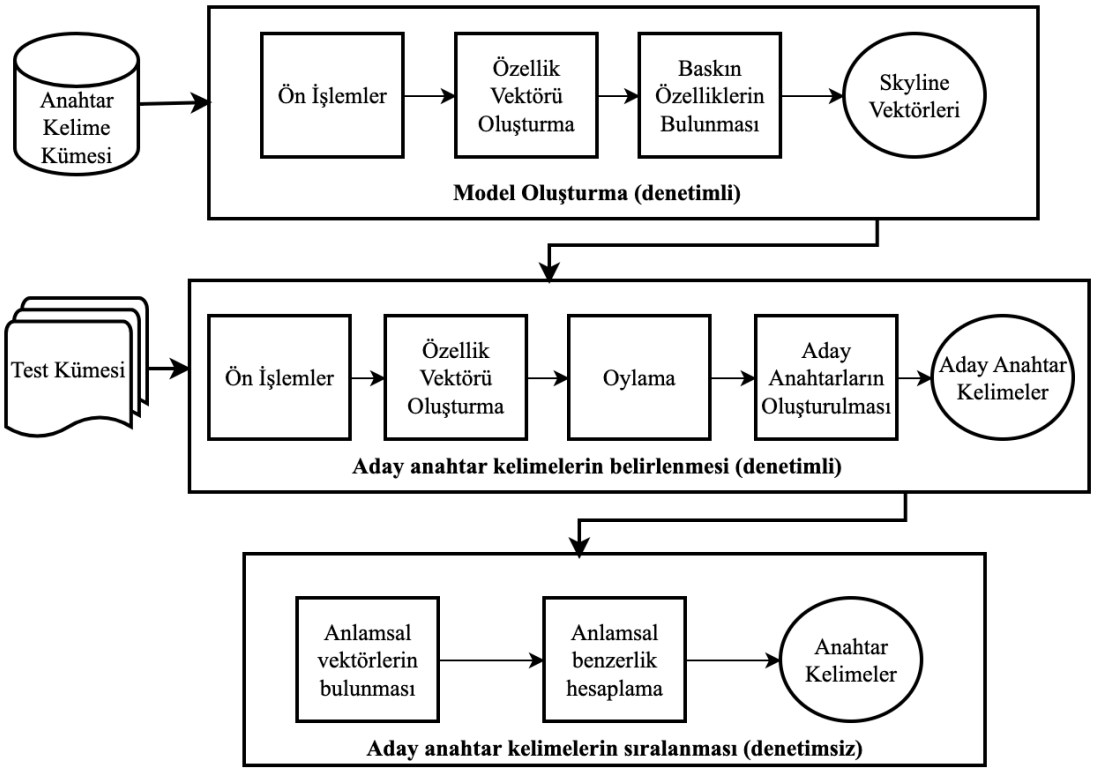
MGRank bir, iki ya da üç uzunluklu olacak şekilde isim tamlamalarını anahtar kelime olarak kabul eder. İsim tamlamalarını bulmak için $(adjective) * (noun)+$ biçimindeki düzenli ifade kullanılmaktadır. Bu aşamada orijinal metin taranır ve belirtilen düzenli ifadeyi içeren kelime yapıları bulunur.

Bir uzunluklu aday anahtar kelimelerin önem değeri aday anahtar kelimelerin sıralanması aşamasında bulunmaktadır. İki ve üç uzunluk kelime öbeklerinin değeri içerdiği her bir kelimenin değerleri toplamı ile elde edilir. Son aşamada elde edilen değerlere göre tüm aday anahtar kelimeler sıralanır. En yüksek değere sahip n adet aday anahtar kelime anahtar kelime olarak belirlenir.

3.2. SkyWords

Geliştirilen bir diğer yöntem denetimli ve denetimsiz öğrenme yaklaşımlarını birlikte uygulayan hibrit bir yöntemdir. SkyWords olarak adlandırılan bu yöntem Skyline operatörü ve çoğunluk oylama prensibi kullanılarak geliştirilmiştir. Skyline operatörünün anahtar kelime çıkarma problemine uygulandığı bilinen ilk yöntemdir. SkyWords'te bir metnin anahtar kelimelerini farklı metinlere ait bilinen anahtar kelimelerin özelliklerine göre daha iyi olacak şekilde tespit etmek amaçlanır. SkyWords denetimli model oluşturma, denetimli aday anahtar kelime seçimi, denetimsiz aday anahtar kelimelerin sıralanması olmak üzere 3 temel aşamadan oluşmaktadır. Şekil 3.3 SkyWords yönteminin genel çalışma yapısını göstermektedir.

SkyWords eğitim için seçilen metinlerden anahtar kelime kümesi oluşturur. Model oluşturma aşamasında farklı özellikler için anahtar kelimelerin özellik vektörleri bulunur. Her bir özelliğin önem eğilimi belirlenir. Bu eğilim bir kelimenin özellik vektörü için bazı özelliklerde düşük, bazı özelliklerde yüksek olması durumunda daha



Şekil 3.3. SkyWords Yönteminin Genel Yapısı

önemli olduğunu gösterir. Bu özellikler Skyline operatörüne alınarak baskın özellik vektörleri bulunur. Denetimli anahtar kelime seçme aşamasında baskın özellik vektörleri kadar iyi olacak şekilde bir metin için aday anahtar kelimeleri tespit etmek amaçlanır. Denetimsiz anahtar kelimelerin sıralanması aşamasında aday anahtar kelimelerin ait oldukları metinle anlamsal benzerlikleri bulunur ve aday anahtar kelimeler sıralanır. Bu bölümde SkyWords'ün adımları detaylıca açıklanmaktadır.

3.2.1. Ön İşlem ve Özellik Mühendisliği

Metin ön işleme ve özellik çıkarma adımları SkyWords'te denetimli model oluşturma ve denetimli anahtar kelime seçimi adımlarında kullanılmaktadır. Denetimli model oluşturma aşamasında metin koleksiyonu, anahtar kelime seçimi aşamasında tek bir metin kullanılır.

Metin ön işleme aşamasında metnin en küçük birimlerine dönüştürülmesi, büyük küçük harf dönüşümü, sözcük türlerinin bulunması ve isim ile sıfatların seçilmesi, tekilleştirme, tekrarlayan kelimelerin silinmesi ve cümlelerin bulunması adımları bulunur.

Anahtar kelime çıkarma problemi için anahtar kelime çıkarma problemlerinde kullanılan 8 özellik tercih edilmiştir. Bu özellikler 3 farklı kategoride incelenebilir: pozisyonel özellikler, anlamsal özellikler ve merkezilik özellikleri.

Pozisyonel özellikler anahtar kelime çıkarma çalışmalarında geniş bir şekilde kullanılmaktadır (Campos ve diğ., 2020; Danesh ve diğ., 2015; Florescu ve Caragea, 2017; Luo ve diğ., 2020; Sharma ve diğ., 2021). Pozisyonel özellikler bir metnin ilk kısımlarında bulunan kelimelerin sonlarında bulunan kelimelere göre daha önemli olduğu varsayımına dayanmaktadır. Bu varsayımdan dolayı pozisyonel özelliklerde metnin başlarında bulunan kelimelere daha fazla ağırlık verilir. Bu çalışmada kelime metin pozisyonu ve kelime cümle pozisyonu olmak üzere iki farklı pozisyonel özellik tercih edilmiştir. Kelime metin pozisyonu PositionRank, kelime cümle pozisyonu YAKE!'de tanımlanmış bir özelliktir.

Kelime metin pozisyonunda metnin başlangıcına yakın olan kelimelere daha yüksek ağırlık verilir. Kelime metin pozisyonunun hesabı Denklem 3.5'de verilmiştir. Denklemde p_{w_i}, w_i kelimesinin metin içinde bulunduğu ilk pozisyonu temsil eder.

$$WP(w_i) = \frac{1}{p_{w_i}} \tag{3.5}$$

Kelime cümle pozisyonu bir metnin ilk cümlelerinde yer alan kelimelerin son cümlelerde yer alan kelimelere göre daha önemli olduğunu ifade eder. Kelime cümle pozisyonu Denklem 3.6'da olduğu gibi hesaplanmaktadır. Sen_t w_i aday anahtar kelimesinin bulunduğu cümlelerin pozisyonlarını temsil eder.

$$SP(w_i) = \ln\left(\ln\left(3 + \text{Medyan}\left(\text{Sen}_t\right)\right)\right) \tag{3.6}$$

Anlamsal özellikler, bir metnin yapısının anlaşılmasını sağlar ve anahtar kelimeleri belirlemeyi kolaylaştırmaktadır (Abulaish ve diğ., 2022; Campos ve diğ., 2020; Gagliardi ve Artese, 2020). SkyWords'te üç farklı anlamsal özellik kullanılmıştır: kelime cümle çeşitliliği, normalize kelime frekansı ve ilintililik (relatedness).

Kelime cümle çeşitliliği bir kelimenin farklı cümlelerde kaç kez geçtiğinin bir ölçütüdür. Örneğin 2 farklı cümlede 3 kez bulunan bir kelime, 1 cümlede 3 kez bulunan kelimeden

daha yüksek değer alır. Bu ölçüt Sharma ve arkadaşlarının yaptığı çalışmadan esinlenilmiştir (Sharma ve diğ., 2021). Denklem 3.7 formülünü göstermektedir. Denklemde w_i aday anahtar kelimesi için $SF(w_i)$ bulunduğu farklı cümlelerin sayısını ifade etmektedir.

$$TDS(w_i) = \frac{SF(w_i)}{ToplamCumleSayisi} \tag{3.7}$$

TF anahtar kelime çıkarma çalışmalarında önemli bir yer edinmiştir. Bir metinde sayıca çok olan kelimelerin daha önemli olduğu varsayımını ifade eder. Bu çalışmada TF normalize edilmiş bir yaklaşımla kullanılmıştır. Bu ölçütte kelime sayısı tüm kelime frekanslarının ortalaması ve standart sapması hesaba katılarak normalize edilir (Wang ve Peng, 2005). Denklem 3.8 formülünü göstermektedir. Denklemde $TF(w_i)$, w_i kelimenin frekansını ve $MaksTF_D$, D olarak adlandırılan bir metindeki en çok bulunan kelimenin sayısını göstermektedir.

$$NTF(w_i) = \frac{TF(w_i)}{MaksTF_D} \tag{3.8}$$

İlintililik bir kelimenin metin içindeki dağılımına göre önemini belirleyen bir ölçüttür (Campos ve diğ., 2020; Sharma ve diğ., 2021). Bu ölçüte göre bir kelimenin sağında ve solunda bulunan kelimelerin çeşitliliği arttıkça önem değer düşer. SkyWords'te YAKE!'de kullanılan ilintililik yaklaşımı benimsenmiştir. Denklem 3.9 ve 3.10 hesaplamasını göstermektedir.

$$DL[DR] = \frac{\left|A_{w_i,t}\right|}{\sum_{k \in A_{t,},w_i} \mathrm{CoOccur}_{t,w_k}} \tag{3.9}$$

Hesaplamada, w_i aday anahtar kelime için t kadar sağında ve solunda görülen bir kelimenin bulunma sayısı $|A_{w_i,t}|$ ile ifade edilir. Sağında ve solunda bulunma durumu bir pencere boyutuna göre belirlenir ve t ile temsil edilir. DR ve DL sırasıyla bir kelimenin sağ ve sol tarafında bulunan dağılımları gösterir.

$$\mathrm{REL(w_i)} = 1 + (DL + DR) * \frac{\mathrm{TF(w_i)(t)}}{MaksTF_D} \tag{3.10}$$

Çizge teorisine göre merkezilik ölçütleri bir ağdaki düğümlerin önemini belirlemede kullanılır. Merkezilik ölçütleri anahtar kelime çıkarma yöntemlerinde yaygın olarak

kullanılmaktadır (Biswas ve diğ., 2018; Boudin, 2013; Lahiri ve diğ., 2014; Palshikar, 2007; Vega-Oliveros ve diğ., 2019). Merkezilik ölçütleri 3 farklı kategoride sınıflandırılabilir: yerel ölçekli, orta ölçekli ve küresel ölçekli. Yerel ölçekte yer alan merkezilik ölçütlerinde bir düğümün değerini belirlemek için sadece düğümün komşu düğümleri hesaplamaya dahil eder. Orta ölçekte bir düğümün değeri komşu düğümleri ve komşu düğümlerin komşularıyla belirlenir. Küresel ölçekte ise bir çizgede düğümlerin değeri tüm ağ yapısı hesaplamaya dahil edilerek hesaplanır.

Bu yöntemde 3 merkezilik ölçütü kullanılmıştır: derece merkeziliği, kümeleme katsayısı merkeziliği ve Pagerank merkeziliği. Bu ölçütler sırasıyla yerel, orta ve küresel ölçekte merkezilik ölçütleridir. SkyWords'te merkezilik ölçütlerini hesaplamak için Vega ve arkadaşlarının önerdiği çizge modeli tercih edilmiştir (Vega-Oliveros ve diğ., 2019). Geliştirilen modelde çizge yönsüz ve basittir. Çizgede düğümler aday anahtar kelimeleri temsil eder. Kenarlar ise birlikte bulunan aday anahtar kelimelerin arasındaki bağlantıları içerir. Çizgede pencere boyutu 3 alınmıştır.

Derece merkeziliği düşük hesaplama maliyeti ve verimliliği nedeniyle anahtar kelime çıkarma çalışmalarında tercih edilmektedir. Bir düğümün derece merkeziliği, v düğümünün derecesinin kenar sayısına oranıdır. $|V|$ kenarı olan bir çizge için v düğümünün derece merkeziliği Denklem 3.11'de gösterilmiştir.

$$DE(v) = \frac{deg(v)}{|V| - 1} \tag{3.11}$$

Kümeleme katsayısı değeri v düğümü için olası maksimum bağlantı sayısı üzerinden merkezlenen üçgen sayısı olarak tanımlanır. Denklem 3.12'de formülü verilmiştir. $T(v)$, v düğümünden geçen üçgenlerin sayısı, $deg(v)$, v düğümünün derecesidir.

$$CC(v) = \frac{2T(v)}{deg(v)(deg(v) - 1)} \tag{3.12}$$

Pagerank merkeziliğinin hesabı Denklem 3.13'te verilmiştir. $In(v_i)$, v_i düğümü ile bağlantı kuran düğümler kümesidir. v_j, v_i düğümünün komşu düğümlerini gösterir. $Out(v_j)$, v_j düğümünün bağlantı kurduğu komşu düğümlerin kümesidir.

$$S(v_i) = (1 - \alpha) + \alpha \times \sum_{v_j \in In(v_i)} \frac{1}{|Out(v_j)|} S(v_j) \tag{3.13}$$

3.2.2. Denetimli Model Oluşturma

Denetimli model oluşturma bölümü anahtar kelimelerin özellik değerlerine göre belirleyici olan özellik vektörlerinin belirlendiği bölümdür. Belirleyici özellik vektörlerini bulmak için Skyline operatörü kullanılır. Bu aşamada, girdi olarak eğitim için belirlenen metin kümesi alınır. Bu metin kümesindeki bir metin için öncelikle ön işlem adımları uygulanır daha sonra metindeki her kelime için özellik mühendisliği bölümünde belirlenen özelliklerin değerleri bulunur ve bir kelimeye ait değerler vektör haline getirilir. Bu vektörlerden önceden anahtar kelime olarak etiketlenmiş olanlar özellik vektörü veri tabanına eklenir. Bu adımlar diğer metinler için de gerçekleştirilir. Bu veri tabanı Skyline operatörüne girdi olarak verilmektedir. Skyline operatörünün çalışması için her bir özellik için önem eğilimi belirlenir. Skyline operatörü önem eğilimlerine göre baskın özellik vektörlerini belirlemek için çalıştırılır.

Algoritma 1'de denetimli model oluşturma aşaması özetlenmiştir. Algoritmada D metin kümesini; K, D kümesinin anahtar kelimelerini içeren kümeyi temsil etmektedir. Yöntem önce metin kümesinde yer alan her bir metin daha sonra metne ait anahtar kelimeler için ön işlem adımlarını uygular. *VeriOnİsleme* ve *AnahtarKelimeOnIsleme* bu adımların gerçekleştiği fonksiyonlardır. Tüm metinler için bu adımlar gerçekleştirilir. *VeriOnİsleme* fonksiyonunda kullanılan ön işlem adımlarının tümü *AnahtarKelimeOnIsleme* fonksiyonunda da ortak olarak kullanılmaktadır. *VeriOnİsleme* fonksiyonunda ayrıca metinden cümlelerin bulunması, sözcük türlerinin belirlenmesi ve isim ile sıfatların seçilmesi adımları bulunmaktadır. Bir kelimenin özellik değerlerini temsil eden özellik vektörü, v, D_i metninde bulunan her bir eşsiz anahtar kelime, k_j, için hesaplanır ve bu değerler özellik vektörü veri tabanına, V, eklenir. Bu aşama *Ekle(V,v)* bölümünde gerçekleşir. Tüm girdi metinleri için bu aşamalar gerçekleştirildikten sonra Skyline operatörü, belirleyici Skyline özellik vektörlerini, V_s, bulmak için kullanılır.

Bazı özelliklerin yüksek, bazı özelliklerin düşük değerde olması daha önemli olma eğilimi göstermektedir. SkyWords Skyline operatörünü yüksek öneme sahip özellikler için minimum, diğer özellikler için maksimum olacak şekilde çalıştırır. Böylece Skyline operatörü ile bulunan baskın özellik vektörleri anahtar kelimeler için eşik değerlerini

Algoritma 1 Denetimli model oluşturma

Girdi: D = $\{d_1, d_2, \cdots, d_n\}$, K = $\{kw_1, kw_2, \cdots, kw_n\}$
Çıktı: V_s: Skyline özellik vektörleri

```
for d_i in D do
    D_i ← VeriOnisleme(d_i)
    K_i ← AnahtarKelimeOnIsleme(kw_i)
    for k_j in K_i do
        if k_j in D_i then
            v ← OzellikVektoruOlustur(k_j, D_i)
            Ekle(V, v)
        end if
    end for
end for
V_s ← Skyline(V)
```

temsil eder. Tablo 3.2 SkyWords'te kullanılan özelliklerin önem eğilimini ve Skyline operatörünün parametre yönelimini göstermektedir. Kelime cümle pozisyonu ve ilintililik değerlerinin düşük değerde olması yüksek değerde olmasına göre daha önemlidir. Diğer özelliklerde değer arttıkça önem değeri artar. Skyline operatörünün parametre değerinin yönü özelliklerin önem eğilimlerinin tersi yönde ayarlanır.

Tablo 3.2. Özelliklerin Önem Eğilimi

Özellik	**Önem Eğilimi**	**Skyline Parametresi**
Kelime metin pozisyonu	↑	↓
Kelime cümle pozisyonu	↓	↑
Kelime cümle çeşitliliği	↑	↓
Normalize kelime frekansı	↑	↓
İlintililik	↓	↑
Derece merkeziliği	↑	↓
Kümeleme katsayısı merkeziliği	↑	↓
Pagerank	↑	↓

3.2.3. Denetimli Aday Anahtar Kelime Seçme

Bu aşamada bir metin belgesinden yüksek kalitede aday anahtar kelimelerin seçilmesi amaçlanmaktadır. Bu bölümde ilk olarak tek uzunluklu aday anahtar kelimeler bulunur daha sonra bulunan kelimeler metinde yan yana geçen diğer kelimelerle birleştirilerek kelime öbeklerinden oluşan aday anahtar kelimeler bulunur.

Algoritma 2 denetimli öğrenmeye dayalı aday anahtar kelimelerin seçilmesi aşamasını özetlemektedir. Bu aşamada bir metin belgesi, d, ve Skyline özellik vektör kümesi, V_s, girdi olarak alınır. Bu aşama aday anahtar kelime kümesini, CK, çıktı olarak verir. Algoritma bir metin belgesini işlemekle başlar. İşlenen belgenin, D, her bir kelimesi, w_j, için bir özellik vektörü, v, üretilir. Kelimenin kalitesini değerlendirmek için iki aşamalı bir çoğunluk oylama mekanizması uygulanmaktadır. İlk aşamada, bir kelimenin özellik vektörü, v, her bir Skyline özellik vektörü, v_j ile karşılaştırılır. Bir kelime için özelliklerinden en az yarısı Skyline özellik vektörünün değerlerinden iyiyse *evet* oyu verilir. Bu adımda Skyline özellik vektörlerine göre bir kelimenin iyilik durumu kelime cümle pozisyonu ve ilintililik değerleri için küçük, diğer özellikler için büyük olma durumunda sağlanır. İkinci aşamada *evet* ve *hayır* oylarının sayısına bakılır. Bir kelimenin *evet* oylarının sayısı *hayır* oylarının sayısından fazla ise aday anahtar kelime kümesine, UCK, eklenir. Elde edilen aday anahtar kelimeler tek uzunlukludur.

Algoritma 2 Aday anahtar kelimelerin seçilmesi

Girdi: d, V_s
Çıktı: CK

```
1:  D ← BelgeOnisleme(d)
2:  for w_j in D do
3:      v ← OzellikVektoruOlustur(w_j, D)
4:      for v_j in V_s do
5:          if iyilik(v, v_j) then
6:              evet ← evet + 1
7:          else
8:              hayir ← hayir + 1
9:          end if
10:     end for
11:     if evet > hayir then
12:         Ekle(UCK, v)
13:     end if
14:     evet ← 0
15:     hayir ← 0
16: end for
```

Algoritma 3 nihai aday anahtar kelimelerin bulunması aşamasını göstermektedir. Bu aşamada bir önceki aşamada elde edilen tek uzunluklu aday anahtar kelimeler olduğu gibi ya da farklı kelimelerle birleştirilerek alınır. Algoritmada ilk olarak aday anahtar kelimeler *evet* oylarının sayısına göre *AnahtarKelimeleriSirala* fonksiyonunda sıralanır.

En yüksek *evet* oyu alan n kadar kelime seçilir. Daha sonra *isimTamlamalariniBul* fonksiyonunda metinden l uzunluğa kadar kelime öbekleri bulunur ve aday anahtar kelimelerden herhangi birini içeren ifadeler nihai aday anahtar kelime kümesine, CK, eklenir. Aday anahtar kelime, herhangi bir kelime öbeğinde bulunmuyorsa nihai aday anahtar kelime kümesine eklenir. Bu aşamada aynı kelimeyi içeren birden fazla kelime ya da kelime öbeği varsa herhangi bir seçim yapmadan tümü kümeye alınır. Bu nedenle nihai aday anahtar kelime kümesinin boyutu n sayısından fazla olabilir.

Algoritma 3 Aday anahtar kelimelerin bulunması

Girdi: D, UCK, n, l, CK
Çıktı: CK

```
siraliUCK ← Aday AnahtarKelimeleriSirala(UCK)
TUCK ← siraliUCK[1..n]
NPs ← isimTamlamalariniBul(D, l)
for w_j in TUCK do
    icermeDurumu ← false
    for np_i in NPs do
        if varMi(np_i, w_j) then
            Ekle(CK, np_i)
            icermeDurumu ← true
        end if
    end for
    if not icermeDurumu then
        Ekle(CK, w_j)
    end if
end for
```

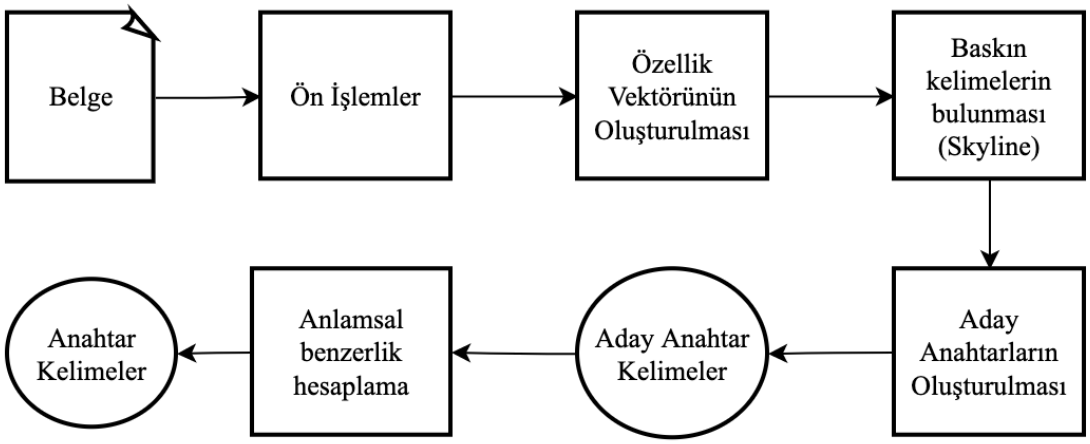
3.2.4. Denetimsiz Aday Anahtar Kelimelerin Sıralanması

Bu aşamada bir metnin anahtar kelimelerini bulmak için aday anahtar kelimeler sıralanmaktadır. Bu amaçla önce metnin daha sonra da aday anahtar kelimelerin vektör temsilleri önceden eğitilmiş bir modelle bulunur. Bu çalışmada eğitilmiş model olarak all-mpnet-base-v2 kullanılmıştır. Metnin vektör temsili her bir aday anahtar kelimenin vektör temsilleriyle karşılaştırılır ve kosinüs benzerliği hesaplanır. Metne en çok benzeyen n sayıda aday anahtar kelime anahtar kelime olarak belirlenir.

3.3. SkyRank

Geliştirilen bir diğer yöntem SkyRank olarak adlandırılmıştır. SkyRank istatistik tabanlı denetimsiz bir anahtar kelime çıkarma yöntemidir. SkyRank ile aday anahtar kelimelerin farklı istatistiksel özelliklerini çıkartarak metni en iyi en temsil eden anahtar kelimelerin bulunması amaçlanır. Bu amaçla Skyline operatörü aday anahtar kelimeleri belirlemek için kullanılır. Bulunan aday anahtar kelimeler metinde yan yana görüldüğü kelimelerle birleştirilir ve kelime öbekleri elde edilir. SkyRank ile ayrıca metne en benzer aday anahtar kelimelerin belirlenmesi ve anahtar kelime olarak alınması amaçlanır. Bu aşamada all-mpnet-base-v2 modeli kullanılmaktadır. Metin ve aday anahtar kelimelerin vektör temsilleri arasındaki benzerlik hesaplanarak anahtar kelimeler belirlenir.

SkyRank bir metin için ön işlem ve özellik mühendisliği, aday anahtar kelimelerin bulunması ve aday anahtar kelimelerin sıralanması aşamalarından oluşur. Şekil 3.4 SkyRank yönteminin genel çalışma yapısını göstermektedir.



Şekil 3.4. SkyRank Yönteminin Genel Yapısı

3.3.1. Ön İşlem ve Özellik Mühendisliği

SkyRank metin ön işleme aşamasında metnin en küçük birimlerine dönüştürülmesi, büyük küçük harf dönüşümü, sözcük türlerinin bulunması, tekilleştirme, tekrarlayan kelimelerin silinmesi ve cümlelerin bulunması adımlarını içermektedir. Metinden sadece isim ve sıfatlar alınır, diğer kelimeler arama uzayından çıkarılır.

Elde edilen kelimelerin SkyWords yönteminde kullanılan istatistiksel özellikler için değerleri hesaplanır. Daha sonra her bir kelime için özellik vektörü oluşturulur. Böylece bir metin pozisyonel özellikler, anlamsal özellikler ve merkezilik özellikler açısından temsil edilmiş olur.

3.3.2. Aday Anahtar Kelimelerin Belirlenmesi

Bu aşamada bir önceki aşamadan elde edilen kelimelere Skyline operatörü uygulanır. Skyline operatörü ile baskın kelimeler bulunur ve diğer kelimeler değerlendirmeye alınmaz. Bulunan kelimeler metinde birlikte geçen kelimelerle genişletilir ve aday anahtar kelime listesi elde edilir.

Tablo 3.3 her bir özellik için kullanılan Skyline parametresini göstermektedir. Skyline operatörü ilintililik ve kelime cümle pozisyonu için minimum, diğer özellikler için maksimum olacak şekilde çalıştırılır.

Tablo 3.3. Skyline Parametresi

Özellik	**Skyline Parametresi**
Kelime metin pozisyonu	↑
Kelime cümle pozisyonu	↓
Kelime cümle çeşitliliği	↑
Normalize kelime frekansı	↑
İlintililik	↓
Derece merkeziliği	↑
Kümeleme katsayısı merkeziliği	↑
Pagerank	↑

3.3.3. Aday Anahtar Kelimelerin Sıralanması

Aday anahtar kelimeler belirlendikten sonra all-mpnet-base-v2 modeli ile metin ve aday anahtar kelimelerin vektör temsilleri bulunur. Metne en benzer anahtar kelimeleri belirlemek için metin ile aday anahtar kelimelerin vektör temsilleri arasındaki kosinüs benzerlikleri hesaplanır. Daha sonra aday anahtar kelimeler sıralanır ve en yüksek değere sahip aday anahtar kelimeler anahtar kelime olarak alınır. Bu aşama SkyWords ile aynıdır.

4. DENEYLER

Bu bölümde tez kapsamında geliştirilen MGRank, SkyWords ve SkyRank yöntemlerinin performansını değerlendirmek için yapılan deneyler ayrıntılı bir şekilde anlatılmaktadır. Yöntemler literatürden seçilen farklı yöntemlerle çeşitli performans ölçütleri açısından karşılaştırılmıştır. Karşılaştırmada akademik çalışmalardan elde edilmiş veri kümeleri kullanılmıştır. Alt bölümlerde deney ortamı, veri kümeleri, karşılaştırma yapılan yöntemler, değerlendirme ölçütleri ve deneysel sonuçlar verilmiştir.

4.1. Deney Ortamı

Yöntemler Python programlama dili ile gerçekleştirilmiştir. Ön işlem aşamasında gerçekleştirilen adımlarda NLTK Kütüphanesi kullanılmıştır. MGRank'in çizge oluşturma ve düğüm derecelendirme aşamalarında NetworkX Kütüphanesinden faydalanılmıştır. SkyWords ve SkyRank'in kullandığı Skyline operatörü için paretoset kütüphanesi kullanılmıştır ve merkezilik ölçütlerin hesaplanması NetworkX kütüphanesi ile gerçekleştirilmiştir.

Yöntemlerin geliştirildiği kod geliştirme ortamı Google Colab'tır. Google Colab Python dilini tarayıcı üzerinden kodlamaya ve çalıştırmaya izin veren ücretsiz bir Google ürünüdür. Google Colab'ın herhangi bir kurulum gerektirmemesi, bulut ortamında çalışması ve Google'ın bilgi işlem kaynaklarına ücretsiz erişim sağlaması önemli özellikleridir.

4.2. Veri Kümeleri

Yöntemlerin performansını ölçmek için akademik çalışmalardan oluşturulmuş 7 farklı veri kümesi kullanılmıştır. Seçilen veri kümeleri şunlardır: KDD (Gollapalli ve Caragea, 2014), Nguyen (Nguyen ve Kan, 2007), Inspec (Hulth, 2003), Krapivin (Krapivin ve diğ., 2009), Sem2010 (Kim ve diğ., 2010), Sem2017 (Augenstein ve diğ., 2017) ve WWW (Gollapalli ve Caragea, 2014).

KDD ACM Conference on Knowledge Discovery and Data Mining'de 2004-2014 yılları arasında yayınlanmış 294 özet metinden oluşmaktadır. WWW veri kümesinde

2004-2014 yılları arasında World Wide Web konferansında yayınlanmış 666 özet metin bulunmaktadır. Nguyen bilimsel çalışmalardan oluşturulmuş ve gönüllü öğrencilerin etiketlediği 209 makale içermektedir. Inspec veri kümesinde 1998-2002 yılları arasında Bilgisayar Bilimleri alanında yayınlanmış çeşitli dergilerden seçilen 2000 çalışmanın özet metinleri bulunmaktadır. Krapivin veri kümesi 2003-2005 yılları arasında ACM tarafından yayınlanmış 2304 çalışmadan oluşmaktadır. Sem2010 ACM dijital kütaphenesinden elde edilmiş yapay zeka, dağıtık sistemler gibi konuları içeren 243 makaleden oluşturulmuştur. Sem2017 ScienceDirect'te yayınlanmış malzeme mühendisliği, fizik, bilgisayar mühendisliği alanlarının konularını içeren 493 çalışmanın paragraflarını içermektedir.

Krapivin, Nguyen ve Sem2017 veri kümeleri tam metin olarak oluşturulmuş veri kümeleridir. Deney aşamasında Krapivin, Nguyen ve Sem2017 veri kümelerinden özet metinler elde edilmiş ve yeniden düzenlenmiştir. KDD ve WWW veri kümelerinde içerikleri bulunmayan akademik çalışmalar bulunmaktadır. KDD ve WWW veri kümelerinden bu bölümler temizlenerek yeniden oluşturulmuştur.

Tablo 4.1'de sırasıyla her bir veri kümesinin içerdiği metin sayısı, orijinal metinlerde bulunan toplam kelime sayısı ve ön işlemlerden sonra elde edilen toplam kelime sayısı verilmiştir. Veri kümeleri ortalama kelime sayılarına göre 127 ile 175 arası kelime içermektedir. Metin başına en az kelime sayısı Inspec, en fazla kelime sayısı Sem2017 veri kümesinde bulunmaktadır.

Tablo 4.1. Veri Kümelerinin İstatistiksel Özellikleri

Veri Kümesi	Metin	Kelime	Kelime(Ön İşlem)	Ortalama
KDD	294	50961	22073	173
WWW	666	102504	44356	153
Nguyen	209	36200	15675	173
Inspec	2000	255499	112226	127
Krapivin	2304	375557	159764	163
Sem2017	493	86206	35910	175
Sem2010	243	41386	16974	170

Tablo 4.2'de veri kümelerinin içerdiği metinler için etiketlenmiş anahtar kelimelere ait bilgiler bulunmaktadır. Bilgiler sırasıyla şu şekildedir: metinlerin toplam anahtar kelime

sayısı, ortalama anahtar kelime sayısı, anahtar kelimelerin ortalama uzunluğu ve metinde bulunmayan anahtar kelimelerin yüzdesi.

Tablo 4.2. Anahtar Kelimelerin Özellikleri

Veri Kümesi	Toplam	Ortalama	Ortalama Uzunluk	Bulunmama Yüzdesi
KDD	1162	3,95	1,96	53
WWW	3052	4,58	1,91	55
Nguyen	2295	10,98	2,15	18
Inspec	27254	13,63	2,24	38
Krapivin	12296	5,34	2,05	15
Sem2017	8486	17,21	2,9	0
Sem2010	3762	15,48	2,16	11

Anahtar kelimelerin istatistiklerine bakıldığında KDD, WWW ve Krapivin veri kümeleri için ortalama anahtar kelime sayısı yaklaşık 5'tir. Diğer veri kümeleri ortalama 10 ve üzeri anahtar kelime sayısı içermektedir. Sem2017'de bu sayı yaklaşık 17'dir. Anahtar kelimelerin içerdiği ortalama kelime sayısı 1,91 ile 2,9 arasında değişkenlik göstermektedir. Etiketli anahtar kelimelerin bulunmama oranı yüzde 55'e kadar çıkmaktadır.

Tablo 4.3 KDD ve WWW veri kümelerinden seçilen örnek metinleri göstermektedir.

Tablo 4.3. Örnek Metinler

KDD: Feedback effects between similarity and social influence in online communities A fundamental open question in the analysis of social networks is to understand the interplay between similarity and social ties. People are similar to their neighbors in a social network for two distinct reasons: first, they grow to resemble their current friends due to social influence; and second, they tend to form new links to others who are already like them, a process often termed selection by sociologists. While both factors are present in everyday social processes, they are in tension: social influence can push systems toward uniformity of behavior, while selection can lead to fragmentation. As such, it is important to understand the relative effects of these forces, and this has been a challenge due to the difficulty of isolating and quantifying them in real settings.
WWW: A fault model and mutation testing of access control policies To increase confidence in the correctness of specified policies, policy developers can conduct policy testing by supplying typical test inputs (requests) and subsequently checking test outputs (responses) against expected ones. Unfortunately, manual testing is tedious and few tools exist for automated testing of access control policies. We present a fault model for access control policies and a framework to explore it. The framework includes mutation operators used to implement the fault model, mutant generation, equivalent-mutant detection, and mutant-killing determination. This framework allows us to investigate our fault model, evaluate coverage criteria for test generation and selection, and determine a relationship between structural coverage and fault-detection effectiveness. We have implemented the framework and applied it to various policies written in XACML. Our experimental results offer valuable insights into choosing mutation operators in mutation testing and choosing coverage criteria in test generation and selection.

4.3. Karşılaştırma Yapılan Yöntemler

Karşılaştırma yapılan yöntemler öğrenme modeline göre farklılık göstermektedir.

Karşılaştırma yapılan yöntemler şu şekildedir:

- TextRank, SingleRank ve PositionRank basit çizge model yapısı kullanan çizge tabanlı yöntemlerdir.
- TopicRank ve MultipartiteRank tam çizge model yapısı kullanan çizge tabanlı yöntemlerdir. TopicRank ve MultipartiteRank anahtar kelimeleri konu tabanlı bir yaklaşımla bulmayı amaçlar.
- RAKE ve YAKE! anahtar kelimeleri istatistik tabanlı bir yaklaşımla bulan yöntemlerdir. RAKE ve YAKE! metni temsil eden çeşitli istatistiksel özellikleri kullanarak anahtar kelimeleri tespit eder.
- SIFRank ve Key2Vec gömme tabanlı anahtar kelime çıkarma yöntemleridir.
- KEA ve WINGNUS denetimli öğrenmeye dayalı anahtar kelime çıkarma yöntemleridir. KEA ve WINGNUS öğrenme aşamasını birbirinden farklı özellikler kullanarak Naive Bayes algoritması ile gerçekleştirir.

MGrank çizge tabanlı denetimsiz öğrenmeye dayalı bir yöntem olduğu için istatistik tabanlı ve çizge tabanlı yöntemlerle karşılaştırılmıştır. Karşılaştırmada kullanılan yöntemler TextRank, SingleRank, PositionRank, TopicRank, MultipartiteRank ve YAKE!'dir.

SkyWords denetimli ve denetimsiz öğrenme yaklaşımlarını birleştiren hibrit bir öğrenme yöntemine sahiptir. SkyWords aday anahtar kelime belirleme ve model oluşturma aşamalarında denetimli, aday anahtar kelimeleri sıralama aşamasında denetimsiz öğrenmeye dayalı bir yaklaşıma sahiptir. SkyWords aday anahtar kelimeleri sıralamak için önceden eğitilmiş vektör modeli kullanmaktadır. SkyWords'ün performansını ölçmek için farklı öğrenme yöntemleri içerdiği için çeşitli yöntemlerle karşılaştırma yapılmıştır. SkyWords için karşılaştırmada kullanılan yöntemler şu şekildedir: PositionRank, SingleRank, TextRank, MultipartiteRank, TopicRank, YAKE!, RAKE, SIFRank, Key2vec, KEA ve WINGNUS.

SkyRank Skyline operatörü ve önceden öğrenilmiş kelime vektörü kullanan denetimsiz öğrenmeye dayalı bir yöntemdir. SkyRank SkyWords'e göre farklı öğrenme yöntemleri içermesine rağmen benzer araçları kullanmaktadır. Bu nedenle SkyRank'in performansını ölçmek için kullanılan yöntemler SkyWords'ün karşılaştırıldığı yöntemlerden seçilmiştir. Seçilen yöntemler F1 değeri açısından diğer yöntemlerden daha iyi sonuç vermektedir. Karşılaştırma yapılan yöntemler KEA, WINGNUS, SIFRank, YAKE!, MultipartiteRank, TopicRank, PositionRank ve SingleRank'tir.

Karşılaştırma yapılan tüm yöntemlerin deneyleri Google Colab ortamında gerçekleştirilmiştir. TextRank, SingleRank, PositionRank, MultipartiteRank, TopicRank, YAKE!, KEA ve WINGNUS yöntemlerinin sonuçları açık kaynak PKE anahtar kelime çıkarma kütüphanesi kullanılarak gerçekleştirilmiştir (Boudin, 2016). SIFRank yazarları tarafından paylaşılan açık kaynak kod kullanılarak test edilmiştir. Diğer yöntemlerin sonuçları Github'da açık kaynak olarak paylaşılmış en popüler kodlar kullanılarak elde edilmiştir (**Github**).

Parametre ayarlamasında YAKE!'nin (Campos ve diğ., 2020) önerdiği değerler dikkate alınmıştır. PositionRank, SingleRank ve TextRank için pencere boyutu sırasıyla 10, 10 ve 2 olarak belirlenmiştir. YAKE!'nin ilintililik değeri için pencere boyutu 2 alınmıştır. TopicRank ve MultipartiteRank için kümeleme parametresi olan minimum benzerlik değeri 0,74 olarak ayarlanmıştır. RAKE'nin aday anahtar kelime uzunluğunun değeri minimum 1, maksimum 3 alınmıştır. KEA ve WINGNUS parametresiz yöntemlerdir. SIFRank ve Key2vec için herhangi bir parametre ayarlaması yapılmamıştır. Tablolarda MultipartiteRank yöntemi MpRank olarak kısaltılmıştır.

4.4. Değerlendirme Ölçütleri

Anahtar kelime çıkarma yöntemlerinde performans ölçümü için iki farklı karşılaştırma tipi kullanılmaktadır: karmaşıklık matrisi tabanlı ölçütler ve sıralama tabanlı ölçütler. Karmaşık matrisi tabanlı ölçütlerde anahtar kelimelerin bulunma sırası dikkate alınmadan hesaplama yapılır. Sıralama tabanlı ölçütlerde anahtar kelimelerin elde edilme sıraları hesaplamaya dahil edilir. Geliştirilen yöntemlerin başarısını ölçmek için karmaşıklık matrisi tabanlı kullanılan ölçütler kesinlik, duyarlılık, F1 değeridir. Sıralama

tabanlı ölçüt olarak MRR ve MAP kullanılmıştır.

Tablo 4.4 Karmaşıklık matrisinin genel bir gösterimini içermektedir.

Tablo 4.4. Karmaşıklık Matrisi

		Tahmin	
		Yanlış	Doğru
Etiket	Yanlış	TN	FP
	Doğru	FN	TP

Anahtar kelime çıkarma problemi için karmaşıklık matrisi şu şekilde yorumlanabilir:

- TP (true-positive) anahtar kelime olarak etiketlenmiş ve anahtar kelime olarak bulunan kelime ya da kelime grubudur.
- TN (true-negative) anahtar kelime değil olarak etiketlenmiş ve anahtar kelime değil olarak bulunan kelime ya da kelime grubudur.
- FP (false-positive) anahtar kelime olmayan ancak anahtar kelime olarak belirlenmiş kelime ya da kelime grubudur.
- FN (false-negatif) anahtar kelime olarak etiketlenmiş ancak anahtar kelime olarak bulunamamış kelime ya da kelime gruplarıdır.

Denklem 4.1 kesinlik değerinin hesaplanmasını göstermektedir. Doğru bulunan anahtar kelime sayısının anahtar kelime olarak tahmin edilen aday anahtar kelime sayısına oranı kesinlik değerini verir.

$$P = \frac{TP}{TP + FP} \tag{4.1}$$

Denklem 4.2 duyarlılık değerinin hesaplanmasını göstermektedir. Doğru bulunan anahtar kelime sayısının anahtar kelime olarak etiketlenmiş anahtar kelime sayısına oranı duyarlılık değerini verir.

$$R = \frac{TP}{TP + FN} \tag{4.2}$$

F1 değerinin hesabı Denklem 4.3'te gösterilmiştir. F1 değeri kesinlilik ve duyarlılık değerlerine bağlı olarak hesaplanmaktadır. Bu değerlerin harmonik ortalaması F1

değerini verir.

$$F1 = \frac{2 \times P \times R}{P + R} \tag{4.3}$$

Sıralama tabanlı yöntemlerden MRR en yüksek sıralamada bulunan anahtar kelimeye göre hesaplanan bir ölçüttür. RR bulunan anahtar kelimeler içinden en yüksek sıralamalı anahtar kelimenin sırasının çarpımsal tersidir. MRR bir metin kümesi için her bir metnin RR değerinin ortalaması olarak tanımlanır. MRR'nin formülü Denklem 4.4'de verilmiştir. Formülde D metni, $rank_d$, d metninin sıralamasını göstermektedir.

$$MRR = \frac{1}{|d|} \sum_{d \in D} \frac{1}{rank_d} \tag{4.4}$$

MAP, anahtar kelimelerin ortalama kesinlik değerine göre anahtar kelime çıkarma yönteminin performansını ölçen bir değerdir. MAP bir sistem tarafından bulunan k adet anahtar kelime listesinden doğru bulunan anahtar kelimelerin sıralarını hesaba katan bir ölçüttür. Bir metin için AP'nin hesaplanması Denklem 4.5'te verilmiştir. Buna göre $|L|$ anahtar kelime listesinin eleman sayısını, $|LR|$ L listesinde yer alan doğru bulunan anahtar kelimelerin sayısını, $P(r)$ r. sıradaki kesinlik değerini (r. sıradaki doğru bulunan anahtar kelime sayısının yüzdesini), $rel(r)$ r. tahmin doğruysa 1 değilse 0 değerini alan durumu temsil eder. Denklem 4.5 ve Denklem 4.6'da formülleri verilmiştir.

$$AP = \sum_{k=1}^{|L|} Precision@k \times rel(k) \tag{4.5}$$

MAP her bir metnin AP değerlerinin ortalaması ile bulunur.

$$MAP = \frac{1}{|D|} \sum_{d \in D} AP(d) \tag{4.6}$$

Elde edilen deneysel sonuçların karşılaştırmalı istatistiksel analizi yapılmıştır. Bu amaçla bağımlı gruplar t-testi gerçekleştirilmiştir. T-testi normal dağılımlı veriler için ya da örneklem sayısı 30'dan fazla veri varsa gerçekleştirilebilmektedir. T-testinin sıfır hipotezi iki popülasyonun ortalamalarının eşit olduğunu varsaymaktadır. p değeri istatistiksel anlamlılık seviyesinin altında olduğunda sıfır hipotez reddedilir. Bu çalışmada istatistiksel anlamlılık değeri %95 olarak kabul edilmiştir. Deneysel sonuçların gösterildiği tablolarda geliştirilen yöntemlere göre istatistiksel olarak kötü olma durumu

▼ ile, istatistiksel olarak iyi olma durumu ▲ ile gösterilmektedir.

4.5. Deneysel Sonuçlar

Bu bölümde geliştirilen yöntemlerin sonuçları karşılaştırmalı olarak verilmektedir. Her bir yöntem için elde edilen sonuçlar kendi alt başlığı altında sunulmaktadır. Tüm yöntemler kesinlik, duyarlılık, F1 değeri, MAP, MRR ve toplam bulunan anahtar kelime sayıları açısından değerlendirilmektedir ve elde edilen sonuçlar tablolar halinde verilmektedir. Ayrıca niteliksel sonuçlar da örnek metinler üzerinden gösterilmektedir. Anahtar kelime sayısı belirtilmeyen sonuçlar 10 anahtar kelime için elde edilmiştir.

4.5.1. MGRank

Bu bölümde MGRank'le ilgili yapılmış deneyler ve sonuçları karşılaştırmalı olarak sunulmaktadır. Ayrıca MGRank'in çizge modeli başlangıç düğüm ağırlığı 1 alınarak farklı çizge modelleriyle karşılaştırılmıştır. Ek olarak MGRank farklı düğüm ağırlıklarıyla test edilmiş ve başlangıç düğüm ağırlığının yönteme etkisi incelenmiştir. Son olarak yöntemin zamansal analizi yapılmıştır.

Tablo 4.5'de MGRank için kesinlik sonuçlarını göstermektedir. MGRank KDD, WWW, Nguyen, Inspec, Krapivin ve Sem2010 veri kümelerinde en iyi kesinlik değeri elde etmiştir. Sem2017 veri kümesinde en iyi kesinlik değeri SingleRank yöntemi tarafından elde edilmiştir. Sem2017 veri kümesi için MGRank en iyi ikinci değere sahiptir. İstatistiksel olarak karşılaştırıldığında MGRank Nguyen ve Sem2017 veri kümeleri dışındaki veri kümelerinde diğer yöntemlere karşı üstünlük elde etmiştir. Nguyen için MGRank PositionRank hariç diğer yöntemlere karşı istatistiksel fark göstermiştir. Nguyen veri kümesinde MGRank PositionRank'ten daha iyi sonuç elde etmiştir ancak iki yöntem arasında istatistiksel olarak bir fark gözlemlenmemiştir. Sem2017 veri kümesi için MGRank ile SingleRank arasında istatistiksel olarak bir fark bulunmamaktadır. Sem2017'de MGRank diğer yöntemlere karşı istatistiksel olarak üstünlük sağlamıştır.

Tablo 4.6 MGRank için karşılaştırmalı duyarlılık sonuçlarını göstermektedir. MGRank Sem2017 veri kümesinde en iyi ikinci değeri, diğer veri kümelerinde en iyi değeri elde

Tablo 4.5. MGRank Karşılaştırmalı Sonuçlar: Kesinlik

Yöntem	KDD	WWW	Nguyen	Inspec	Krapivin	Sem2017	Sem2010
TextRank	0,054▼	0,056▼	0,104▼	0,151▼	0,059▼	0,174▼	0,087▼
SingleRank	0,067▼	0,062▼	0,135▼	0,263▼	0,082▼	**0,349**	0,113▼
PositionRank	0,078▼	0,077▼	0,158	0,244▼	0,093▼	0,327▼	0,112▼
TopicRank	0,05▼	0,058▼	0,124▼	0,204▼	0,072▼	0,28▼	0,099▼
MpRank	0,061▼	0,065▼	0,138▼	0,213▼	0,081▼	0,288▼	0,093▼
YAKE!	0,064▼	0,053▼	0,113▼	0,163▼	0,084▼	0,202▼	0,086▼
MGRank	**0,09**	**0,078**	**0,165**	**0,288**	**0,106**	0,348	**0,199**

etmiştir. Sem2017 veri kümesinde en iyi sonuç SingleRank'e aittir. Ancak istatistiksel olarak SingleRank ile MGRank arasında Sem2017 veri kümesi için istatistiksel olarak bir fark yoktur. Diğer veri kümelerinde MGRank SingleRank'ten istatistiksel olarak daha iyidir. PositionRank Nguyen veri kümesinde MGRank'ten daha düşük bir değere sahip olmasına rağmen istatistiksel olarak aralarında bir fark yoktur. Diğer veri kümelerinde MGRank PositionRank'ten istatistiksel olarak daha iyidir. MGRank duyarlılık ölçütü için diğer yöntemlerle karşılaştırıldığında tüm veri kümelerinde istatistiksel olarak daha iyi sonuç göstermiştir.

Tablo 4.6. MGRank Karşılaştırmalı Sonuçlar: Duyarlılık

Yöntem	KDD	WWW	Nguyen	Inspec	Krapivin	Sem2017	Sem2010
TextRank	0,139▼	0,12▼	0,099▼	0,098▼	0,111▼	0,102▼	0,054▼
SingleRank	0,18▼	0,149▼	0,144▼	0,215▼	0,158▼	**0,221**	0,075▼
PositionRank	0,212▼	0,184	0,169	0,201▼	0,184▼	0,204▼	0,074▼
TopicRank	0,131▼	0,134▼	0,13▼	0,167▼	0,139▼	0,172▼	0,064▼
MpRank	0,16▼	0,152▼	0,14▼	0,177▼	0,159▼	0,18▼	0,06▼
YAKE!	0,167▼	0,129▼	0,12▼	0,143▼	0,175▼	0,129▼	0,057▼
MGRank	**0,239**	**0,186**	**0,18**	**0,231**	**0,209**	0,219	**0,131**

Tablo 4.7 MGRank ve seçilen yöntemlerin F1 değerini göstermektedir. Tabloya göre MGRank KDD, WWW, Nguyen, Inspec, Krapivin ve Sem2010 veri kümelerinde en yüksek F1 değerine sahiptir. Bu veri kümeleri için istatistiksel olarak MGRank; TextRank, TopicRank, SingleRank, MultipartiteRank ve YAKE!'den üstündür. MGRank PositionRank ile kıyaslandığında WWW veri kümesinde aynı, Nguyen veri kümesinde daha iyi F1 değerine sahiptir. Nguyen ve WWW veri kümeleri için PositionRank ile arasında istatistiksel olarak bir fark gözlemlenmemiştir. Sem2017 veri kümesinde en

yüksek değer SingleRank'e aittir ve MGRank en yüksek ikinci değeri elde etmiştir. Aynı veri kümesi için MGRank SingleRank'ten daha düşük değer elde etmiş olmasına rağmen istatistiksel olarak aralarında bir fark yoktur.

Tablo 4.7. MGRank Karşılaştırmalı Sonuçlar: F1 Değeri

Yöntem	KDD	WWW	Nguyen	Inspec	Krapivin	Sem2017	Sem2010
TextRank	0,075▾	0,073▾	0,091▾	0,114▾	0,073▾	0,124▾	0,065▾
SingleRank	0,095▾	0,085▾	0,128▾	0,225▾	0,102▾	**0,261**	0,089▾
PositionRank	0,112▾	**0,106**	0,15	0,21▾	0,117▾	0,242▾	0,088▾
TopicRank	0,071▾	0,079▾	0,117▾	0,175▾	0,09▾	0,205▾	0,077▾
MpRank	0,086▾	0,088▾	0,128▾	0,184▾	0,102▾	0,213▾	0,072▾
YAKE!	0,091▾	0,073▾	0,107▾	0,144▾	0,109▾	0,151▾	0,068▾
MGRank	**0,128**	**0,106**	**0,16**	**0,244**	**0,133**	0,259	**0,156**

Kesinlik, duyarlılık ve F1 değeri sonuçlarına göre MGRank Sem2017 veri kümesi haricinde çizge tabanlı ve istatistiksel yaklaşımlardan daha iyi sonuç elde etmiştir. Sem2017 veri kümesinde ise en iyi ikinci değeri elde edilmiştir. Kabasakal ve arkadaşının çalışmasında anahtar kelimeler Sem2017 veri kümesinde metinlerin içinde dengeli olarak dağılmaktadır (Kabasakal ve Mutlu, 2021). MGRank metinde erken görünen kelimelere daha yüksek başlangıç ağırlığı atadığı için metnin sonraki bölümlerinde bulunan anahtar kelimeleri bulamayabilir. Bu nedenle en iyi sonuç pozisyon değerini kullanmayan SingleRank yöntemi tarafından elde edilmiş olabilir. MGRank gibi pozisyon tabanlı yöntem olan PositionRank Sem2017 veri kümesinde MGRank'ten daha kötü bir sonuç elde etmiştir.

Tablo 4.8 MGRank için MRR sonuçlarını göstermektedir. MGRank KDD, Inspec, Krapivin, Sem2017 ve Sem2010 veri kümelerinde en yüksek MRR değeri elde etmiştir. WWW ve Nguyen veri kümesinde en yüksek değer MultipartiteRank'e aittir ve MGRank en yüksek ikinci değeri elde etmiştir. Nguyen veri kümesinde ise en yüksek dördüncü değere sahip olmuştur. İstatistiksel olarak karşılaştırıldığında MGRank; TextRank ve SingleRank yöntemlerine göre tüm veri kümelerinde üstün sağlamıştır. PositionRank Nguyen veri kümesinde MGRank'ten daha iyi MRR değeri elde etmiştir ancak istatistiksel olarak MGRank'ten üstün değildir. MGRank ile PositionRank arasında KDD ve WWW veri kümelerinde istatistiksel olarak bir fark görülmemektedir.

Diğer dört veri kümesinde MGRank istatistiksel olarak PositionRank'ten daha iyidir. MGRank ile TopicRank arasında Nguyen haricinde diğer veri kümelerinde MGRank en iyi sonucu elde etmiştir ve bu veri kümelerinden üçünde istatistiksel olarak fark gözlemlenmiştir. Nguyen veri kümesinde TopicRank daha iyi sonuç elde etmesine rağmen istatistiksel olarak aralarında bir fark yoktur. MultipartiteRank Nguyen ve WWW veri kümelerinde MGRank'ten daha iyi sonuç elde etmiştir. MultipartiteRank Nguyen veri kümesinde MGRank'ten istatistiksel olarak üstün çıkmıştır, WWW veri kümesinde istatistiksel olarak aralarında bir fark gözlemlenmemiştir. MGRank diğer veri kümelerinde daha iyi sonuç elde etmiştir ve bu veri kümeleri arasından Inspec, Krapivin ve Sem2010'da istatistiksel olarak üstündür. MGRank YAKE!'ye göre tüm veri kümelerinde istatistiksel olarak anlamlı bir fark göstermiştir.

Tablo 4.8. MGRank Karşılaştırmalı Sonuçlar: MRR

Yöntem	KDD	WWW	Nguyen	Inspec	Krapivin	Sem2017	Sem2010
TextRank	0,156▼	0,155▼	0,259▼	0,333▼	0,171▼	0,368▼	0,229▼
SingleRank	0,158▼	0,147▼	0,281▼	0,549▼	0,193▼	0,583▼	0,249▼
PositionRank	0,288	0,242	0,4	0,583▼	0,287▼	0,606▼	0,288▼
TopicRank	0,217▼	0,235	0,402	0,518	0,256▼	0,61	0,317▼
MpRank	0,265	**0,255**	**0,462▲**	0,552▼	0,287▼	0,612	0,302▼
YAKE!	0,159▼	0,126▼	0,279▼	0,438▼	0,286▼	0,427▼	0,226▼
MGRank	**0,306**	0,242	0,397	**0,594**	**0,308**	**0,640**	**0,469**

Tablo 4.9 MGRank için MAP sonuçlarını göstermektedir. Tabloya göre MGRank 5 veri kümesi için en yüksek MAP değerine ve diğer 2 veri kümesi için ikinci en yüksek değere ulaşmıştır. TextRank ve SingleRank ile karşılaştırıldığında MGRank, tüm veri kümelerinden istatistiksel olarak daha iyi sonuç elde etmiştir. MGRank PositionRank ile karşılaştırıldığında, tüm veri kümelerinde daha iyi sonuç elde etmiştir. Bu veri kümelerinden Inspec, Krapivin, Sem2017 ve Sem2010'da istatistiksel olarak anlamlı bir fark oluşmuştur. MGRank, tüm veri kümeleri için TopicRank'ten daha iyi sonuç göstermiştir ve WWW, Nguyen ve Inspec haricindeki 4 veri kümesi için istatistiksel olarak önemli iyileşmeler elde etmiştir. MultipartiteRank, WWW ve Nguyen veri kümesi için MGRank'ten daha iyi performans göstermiştir ancak fark istatistiksel olarak anlamlı değildir. MGRank Inspec, Sem2017 ve Sem2010 veri kümelerinde istatistiksel olarak anlamlı ve diğer veri kümelerinde MultipartiteRank'ten daha iyi sonuç elde etmiştir.

YAKE! ile karşılaştırıldığında, MGRank tüm veri kümelerinde istatistiksel olarak daha anlamlı sonuçlar elde etmiştir.

Tablo 4.9. MGRank Karşılaştırmalı Sonuçlar: MAP

Yöntem	KDD	WWW	Nguyen	Inspec	Krapivin	Sem2017	Sem2010
TextRank	0,155▼	0,151▼	0,243▼	0,315▼	0,166▼	0,324▼	0,221▼
SingleRank	0,157▼	0,143▼	0,273▼	0,431▼	0,182▼	0,514▼	0,232▼
PositionRank	0,267	0,233	0,358	0,498▼	0,269▼	0,528▼	0,262▼
TopicRank	0,211▼	0,221	0,357	0,477	0,239▼	0,520▼	0,297▼
MpRank	0,246	**0,242**	**0,412**	0,478▼	0,268	0,517▼	0,283▼
YAKE!	0,156▼	0,123▼	0,256▼	0,383▼	0,267▼	0,376▼	0,201▼
MGRank	**0,276**	0,235	0,371	**0,508**	**0,283**	**0,551**	**0,416**

Tablo 4.10 MGRank ve diğer yöntemlerin doğru eşleşen anahtar kelimelerin toplam sayısını göstermektedir. MGRank Sem2017 dışındaki tüm veri kümelerinde en fazla sayıda doğru anahtar kelimeyi bulan yöntem olmuştur. SingleRank Sem2017 veri kümesi için MGRank'e göre 7 anahtar kelime daha fazla bulmuştur. Ancak MGRank toplamda daha fazla anahtar kelimeyi doğru çıkarmıştır. MGRank TextRank'e göre iki kat daha fazla anahtar kelime çıkarmıştır. En az iyileşme SingleRank ve PositionRank'e karşı elde edilmiştir.

Tablo 4.10. MGRank: Anahtar Kelime Sayıları

Yöntem	KDD	WWW	Nguyen	Inspec	Krapivin	Sem2017	Sem2010
TextRank	151	332	192	2508	1271	835	196
SingleRank	196	410	281	5213	1882	**1722**	274
PositionRank	230	515	330	4812	2142	1613	273
TopicRank	148	386	257	3999	1646	1379	240
MpRank	179	433	287	4195	1866	1421	225
YAKE!	187	355	236	3254	1932	997	210
MGRank	**265**	**519**	**344**	**5657**	**2428**	1715	**483**

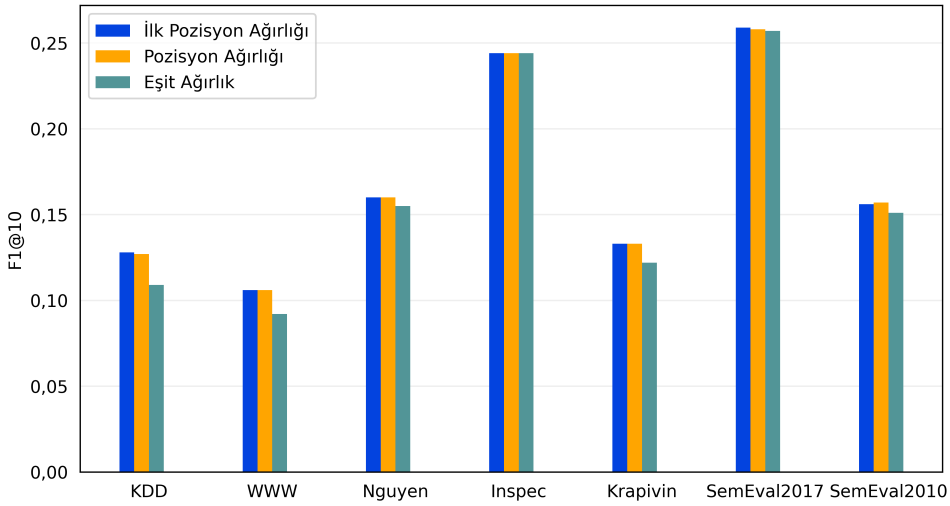
MGRank'te önerilen çizge modelin etkinliğini değerlendirmek için MGRank TextRank, TopicRank ve MultipartiteRank ile karşılaştırılmıştır. TextRank ile modelin basit çizge modeline göre, TopicRank ile modelin çok kenarlı çizge modeline göre, MultipartiteRank ile modelin tam ve çok parçalı çizge modeline göre karşılaştırma yapılmıştır. TextRank, TopicRank ve MultipartiteRank'in çizge modelinde başlangıç düğüm ağırlığı 1'dir. Benzer şartların oluşması için MGRank'in çizge modelinde

başlangıç düğüm ağırlığı 1 olarak güncellenmiştir. F1 değeri için sonuçlar Tablo 4.11'de verilmiştir. MGRank'in çizge modelinin TextRank, TopicRank ve MultipartiteRank'e göre WWWW haricindeki veri kümelerinde daha iyi F1 değeri ürettiğini göstermektedir. Inspec, Sem2017 ve Sem2010 veri kümesi için MGRank istatistiksel olarak fark göstermiştir. Krapivin ve KDD veri kümesi için MGRank TextRank ve TopicRank'ten istatistiksel olarak daha anlamlıdır ancak MultipartiteRank'ten istatistiksel olarak daha anlamlı değildir. Nguyen veri kümesinde MGRank TextRank'ten istatistiksel olarak daha iyi sonuç elde etmiştir. MGRank TopicRank'e göre istatistiksel fark göstermiştir. Nguyen'de MGRank ile MultipartiteRank arasında istatistiksel olarak bir fark oluşmamıştır. WWW veri kümesinde MultipartiteRank en iyi sonucu elde etmiştir ancak MGRank ile arasında istatistiksel olarak bir fark yoktur. WWW'de MGrank ile TopicRank arasında istatistiksel olarak bir fark oluşmamıştır. TextRank ile kıyaslandığında F1 değeri açısından istatistiksel olarak fark oluşmuştur. Bu sonuçlar önerilen çok kenarlı tam çizge modelinin karşılaştırılan çizge modellerinden daha üstün olduğunu göstermektedir.

Tablo 4.11. Önerilen Çizge Modelin Anahtar Kelime Çıkarmaya Etkisi

Yöntem	KDD	WWW	Nguyen	Inspec	Krapivin	Sem2017	Sem2010
TextRank	0,075▼	0,073▼	0,091▼	0,114▼	0,073▼	0,124▼	0,065▼
TopicRank	0,071▼	0,079	0,117▼	0,175▼	0,09▼	0,205▼	0,077▼
MpRank	0,086	**0,088**	0,128	0,184▼	0,102	0,211▼	0,072▼
MGRank	**0,095**	0,085	**0,133**	**0,223**	**0,106**	**0,248**	**0,095**

Başlangıç düğüm ağırlığının anahtar kelime çıkarmaya etkisi incelenmiştir. Bu amaçla MGRank literatürde yer alan başlangıç düğüm ağırlıklandırma yaklaşımlarıyla test edilmiştir. MGRank'in çizge modelinde başlangıç düğüm ağırlıkları belirlenen yaklaşımlarla ayarlanmıştır ve F1 değeri için sonuçlar elde edilmiştir. Bu yaklaşımlar her bir aday anahtar kelimesi için ilk konum bilgisi, tüm konum bilgileri (Awan ve Beg, 2021; Florescu ve Caragea, 2017) ve 1 (Mihalcea ve Tarau, 2004; Wan ve Xiao, 2008) olarak belirlenmiştir. Şekil 4.1 yapılan deneyleri F1 değeri açısından sonuçları içermektedir. İlk pozisyona dayalı düğüm ağırlığı ve tüm pozisyonlara dayalı düğüm ağırlığı benzer sonuçlar vermiştir ve diğer ağırlıklandırma yöntemine göre daha iyi sonuç elde etmiştir.



Şekil 4.1. Başlangıç Düğüm Ağırlığının Etkisi

Tablo 4.12 anahtar kelime çıkarma aşaması için ortalama süreleri göstermektedir. MGRank karşılaştırılan yöntemlere göre daha uzun sürede anahtar kelimeleri bulmaktadır. Bunun sebebi çizge modelinin tam ve çok kenarlı olmasından kaynaklanmaktadır.

Tablo 4.12. MGRank Anahtar Kelime Çıkarma Süreleri

Yöntem	KDD	WWW	Nguyen	Inspec	Krapivin	Sem2017	Sem2010
TextRank	0,01	0,01	0,01	0,01	0,01	0,01	0,01
SingleRank	0,01	0,05	0,01	0,01	0,01	0,03	0,04
PositionRank	0,04	0,07	0,01	0,01	0,01	0,01	0,01
TopicRank	0,01	0,01	0,01	0,09	0,01	0,01	0,01
MpRank	0,02	0,02	0,02	0,02	0,02	0,02	0,02
YAKE!	0,4	0,04	0,04	0,03	0,43	0,05	0,38
MGRank	1,22	0,96	1,53	0,72	1,1	1,1	1,1

MGRank örnek metinler üzerinden elde edilen anahtar kelimeler ile karşılaştırılmıştır. Karşılaştırma Tablo 4.3'te verilen örnek metinler ile yapılmıştır.

Tablo 4.13, 4.14 sırasıyla KDD ve WWW veri kümelerinden alınan metinler için MGRank ve diğer yöntemlerin bulduğu anahtar kelimeleri sıralarına göre göstermektedir.

KDD'den alınmış örnek metin 3 anahtar kelime içermektedir. MGRank, TextRank, PositionRank, MultipartiteRank ve YAKE! tüm anahtar kelimeleri doğru bir şekilde

Tablo 4.13. MGRank Karşılaştırmalı Sonuçlar: KDD Örneği

Anahtar Kelimeler: online communities, social influence, social networks
MGRank: *social influence*, *social networks*, social ties, social processes, feedback effects, relative effects, similarity, *online communities*, open question, real settings
TextRank: fundamental open question, social ties, *social networks*, *online communities*, *social influence*, feedback effects, effects, social, people, interplay
SingleRank: everyday social processes, current friends due, fundamental open question, real settings, relative effects, new links, distinct reasons, social network, social ties, *social networks*
PositionRank: *social influence*, everyday social processes, *social networks*, social ties, social network, feedback effects, fundamental open question, *online communities*, relative effects, similarity
TopicRank: *social influence*, similarity, selection, feedback effects, uniformity, present, people, systems, behavior, everyday social processes
MpRank: *social influence*, similarity, feedback effects, selection, process, *social networks*, *online communities*, uniformity, systems, tension
YAKE!: fundamental open question, *social influence*, *online communities*, fundamental open, open question, social, social ties, *social networks*, similarity, feedback effects

bulmuştur. MGRank ilk 2 anahtar kelimeyi; PositionRank, TopicRank ve MultipartiteRank yöntemleri ilk anahtar kelimeyi doğru bulmuştur.

Tablo 4.14. MGRank Karşılaştırmalı Sonuçlar: WWW Örneği

Anahtar Kelimeler: access control policies, fault model, mutation testing, test generation, testing tools
MGRank: *access control policies*, *fault model*, *mutation testing*, specified policies, policy developer, various policy, *test generation*, mutation operators, policy, choosing coverage criteria
TextRank: *access control policies*, typical test, policy testing, policy developers, *mutation testing*, *fault model*, fault, mutation, testing, test
SingleRank: typical test inputs, *access control policies*, valuable insights, experimental results, various policies, detection effectiveness, structural coverage, *test generation*, coverage criteria, mutant detection
PositionRank: *access control policies*, *fault model*, *mutation testing*, policy testing, fault, manual testing, mutation operators, *test generation*, various policies, typical test inputs
TopicRank: *access control policies*, *fault model*, *mutation testing*, framework, test outputs, coverage criteria, mutant generation, mutation operators, selection, equivalent
MpRank: *mutation testing*, *access control policies*, *fault model*, framework, coverage criteria, mutant generation, mutation operators, selection, *test generation*, equivalent
YAKE!: *access control policies*, conduct policy testing, access control, typical test inputs, checking test outputs, supplying typical test, subsequently checking test, control policies, *fault model*, policy developers

WWW'den alınmış örnek metin 5 anahtar kelime içermektedir. Bu anahtar kelimelerden 4 tanesini MGRank, PositionRank ve MultipartiteRank doğru bir şekilde bulmuştur. MGRank, PositionRank, TopicRank ve MultipartiteRank'in bulduğu ilk 3 anahtar kelime anahtar kelime listesinde yer almaktadır.

4.5.2. SkyWords

Bu bölümde SkyWords'le ilgili deneysel sonuçlar sunulmaktadır. Bu bölümde ilk olarak SkyWords ve diğer yöntemlerin ürettiği aday anahtar kelime sayıları verilmektedir. Daha sonra diğer ölçütler için deneysel sonuçlar sunulmaktadır. Ayrıca SkyWords farklı anahtar kelime sayıları için test edilmiştir. İlave olarak çoğunluk oylama prensibi farklı parametre değerleri için incelenmiştir.

Tablo 4.15 SkyWords ve diğer yöntemlerin elde ettiği aday anahtar kelime sayılarını göstermektedir.

Tablo 4.15. Aday Anahtar Kelime Sayıları

Yöntem	KDD	WWW	Nguyen	Krapivin	Sem2017	Sem2010	Ort.
KEA	163,49	144,07	148,18	137,71	146,62	145,96	147,67
WINGNUS	38,51	34,89	38,98	35,61	39,19	37,09	37,38
SIFRank	44,07	40,54	44,86	42,08	44,96	43,13	43,27
Key2vec	81,63	72,23	77,03	71,27	81,84	75,62	76,6
RAKE	53,68	48,8	54,71	52,37	57,94	54,79	53,72
YAKE!	102,9	92,32	98,06	91,36	101,38	100,21	97,71
MpRank	37,67	33,74	37,89	34,23	36,47	36,26	36,04
TopicRank	28,4	25,47	28,63	25,55	28,68	27,8	27,42
PositionRank	54,2	47,11	51,59	48,58	55,59	50,85	51,32
SingleRank	54,2	47,11	51,59	48,58	55,59	50,85	51,32
TextRank	54,2	47,11	51,59	48,58	55,59	50,85	51,32
SkyWords	13,3	12,94	13,52	13,55	12,55	13,36	13,2

KEA ortalama 147,67, WINGNUS 37,38, SIFRank 43,27, Key2Vec 76,6, RAKE 53,72, YAKE! 97,71, MultipartiteRank 36,04, TopicRank 27,42, PositionRank, SingleRank ve TextRank 51,32 aday anahtar kelime üretmiştir. SkyWords ortalama 12,84 aday anahtar kelime üretmiştir. KEA en fazla aday anahtar kelime sayısı üreten yöntemdir. Daha sonra sırasıyla YAKE! ve Key2vec en fazla aday anahtar kelime üreten yöntemlerdir. Çizge tabanlı yöntemlerden TextRank, SingleRank ve PositionRank aynı veri ön işlem adımlarını içerdikleri ve çizge modellerinin benzerlik içermesinden dolayı aynı sayıda aday anahtar kelime sayısı üretmiştir. TopicRank ve MultipartiteRank basit çizge modellerine göre daha düşük sayıda aday anahtar kelime üretmiştir. Konu tabanlı çizge modeline dayalı yöntemler aynı zamanda SkyWords hariç diğer yöntemlere göre daha az

sayıda aday anahtar kelime sayısına sahiptir. SkyWords ortalama 12,84 aday anahtar kelime sayısına sahiptir. SkyWords diğer yöntemlere göre önemli ölçüde az sayıda aday anahtar kelime ile çalışmaktadır.

Tablo 4.16, 4.17, ve 4.18 sırasıyla kesinlik, duyarlılık ve F1 değeri için SkyWords ve diğer yöntemlerin sonuçlarını göstermektedir. SkyWords ve denetimli öğrenme tabanlı yöntemlerin deneyleri 10 katlamalı çapraz doğrulama ile elde edilmiştir. Böylece tüm metinlerin test edilmesi sağlanmıştır. 10 katlamalı çapraz doğrulamada her bir eğitim ve test dizisi rastgele metinler seçerek belirlenmiştir. Tüm yöntemler aynı test ve eğitim kümeleri ile test edilmiştir.

Tablo 4.16. SkyWords Karşılaştırmalı Sonuçlar: Kesinlik

Yöntem	KDD	WWW	Nguyen	Krapivin	Sem2017	Sem2010
KEA	0,087▼	0,082▼	0,149▼	0,083▼	0,212▼	0,088▼
WINGNUS	0,07▼	0,071▼	0,148▼	0,09▼	0,303▼	0,103▼
SIFRank	0,058▼	0,057▼	0,128▼	0,082▼	**0,389▲**	0,125▼
Key2vec	0,041▼	0,04▼	0,079▼	0,043▼	0,187▼	0,074▼
RAKE	0,028▼	0,03▼	0,07▼	0,05▼	0,222▼	0,063▼
YAKE!	0,064▼	0,053▼	0,113▼	0,084▼	0,202▼	0,086▼
MpRank	0,061▼	0,065▼	0,138▼	0,081▼	0,288▼	0,093▼
TopicRank	0,05▼	0,058▼	0,124▼	0,072▼	0,28▼	0,099▼
PositionRank	0,078▼	0,077▼	0,158▼	0,093▼	0,327▼	0,112▼
SingleRank	0,067▼	0,062▼	0,135▼	0,082▼	0,349	0,113▼
TextRank	0,054▼	0,056▼	0,104▼	0,059▼	0,174▼	0,087▼
SkyWords	**0,104**	**0,089**	**0,179**	**0,116**	0,352	**0,216**

Tablo 4.16 SkyWords'ün kesinlik ölçütü için Sem2017 veri kümesi hariç diğer veri kümelerinde en iyi sonucu elde ettiğini göstermektedir. KDD, WWW, Nguyen, Krapivin ve Sem2010 veri kümelerinde SkyWords tüm yöntemlerde istatistiksel olarak fark göstermiştir. Sem2017 veri kümesinde SIFRank en iyi sonucu elde etmiştir. SIFRank Sem2017'de SkyWords'ten istatistiksel olarak üstündür. SkyWords Sem2017'de SIFRank ve SingleRank haricindeki yöntemlerden istatistiksel olarak daha anlamlıdır. SingleRank ile kıyaslandığında SkyWords daha iyi performans göstermiştir ancak istatistiksel bir fark oluşmamıştır.

Tablo 4.17'ye göre, SkyWords duyarlılık ölçütünde KDD, WWW, Nguyen, Krapivin ve

Tablo 4.17. SkyWords Karşılaştırmalı Sonuçlar: Duyarlılık

Yöntem	KDD	WWW	Nguyen	Krapivin	Sem2017	Sem2010
KEA	0,232▼	0,197	0,162▼	0,175▼	0,13▼	0,06▼
WINGNUS	0,184▼	0,167▼	0,16▼	0,175▼	0,193▼	0,067▼
SIFRank	0,157▼	0,137▼	0,14▼	0,153▼	**0,243▲**	0,082▼
Key2vec	0,112▼	0,091▼	0,085▼	0,08▼	0,117▼	0,048▼
RAKE	0,073▼	0,074▼	0,072▼	0,094▼	0,142▼	0,041▼
YAKE!	0,172▼	0,129▼	0,12▼	0,175▼	0,129▼	0,057▼
MpRank	0,16▼	0,152▼	0,14▼	0,159▼	0,18▼	0,06▼
TopicRank	0,131▼	0,134▼	0,13▼	0,139▼	0,172▼	0,064▼
PositionRank	0,212▼	0,184▼	0,169▼	0,184▼	0,204	0,074▼
SingleRank	0,18▼	0,149▼	0,144▼	0,158▼	0,221▼	0,075▼
TextRank	0,139▼	0,12▼	0,099▼	0,111▼	0,102▼	0,054▼
SkyWords	**0,269**	**0,204**	**0,191**	**0,222**	0,210	**0,138**

Sem2010 veri kümelerinde tüm yöntemlerde daha iyi performans elde etmiştir. Sem2017 veri kümesinde Skywords en iyi üçüncü sonucu elde etmiştir. Bu veri kümesi için en iyi sonuç SIFRank'e, en iyi ikinci sonuç PositionRank'e aittir. İstatistiksel olarak bakıldığında SkyWords WWW ve Sem2017 veri kümeleri dışında diğer veri kümelerinde karşılaştırılan yöntemlerden üstündür. WWW veri kümesinde SkyWords en iyi sonucu elde etmesine rağmen KEA ile arasında istatistiksel olarak bir fark oluşmamıştır. Sem2017'de SIFRank istatistiksel olarak SkyWords'ten üstündür. PositionRank Sem2017'de SkyWords'ten daha iyi duyarlılık değeri elde etmiştir ancak istatistiksel olarak bir fark oluşmamıştır.

Tablo 4.18. SkyWords Karşılaştırmalı Sonuçlar: F1 Değeri

Yöntem	KDD	WWW	Nguyen	Krapivin	Sem2017	Sem2010
KEA	0,124▼	0,112	0,143▼	0,108▼	0,155▼	0,071▼
WINGNUS	0,099▼	0,097▼	0,142▼	0,112▼	0,226▼	0,08▼
SIFRank	0,082▼	0,078▼	0,123▼	0,1▼	**0,288▲**	0,098▼
Key2vec	0,058▼	0,054▼	0,076▼	0,052▼	0,139▼	0,058▼
RAKE	0,039▼	0,041▼	0,065▼	0,061▼	0,167▼	0,049▼
YAKE!	0,091▼	0,073▼	0,107▼	0,109▼	0,151▼	0,068▼
MpRank	0,086▼	0,088▼	0,128▼	0,102▼	0,213▼	0,072▼
TopicRank	0,071▼	0,079▼	0,117▼	0,09▼	0,205▼	0,077▼
PositionRank	0,112▼	0,106▼	0,15▼	0,117▼	0,242▼	0,088▼
SingleRank	0,095▼	0,085▼	0,128▼	0,102▼	0,261	0,089▼
TextRank	0,075▼	0,073▼	0,091▼	0,073▼	0,124▼	0,065▼
SkyWords	**0,146**	**0,12**	**0,172**	**0,144**	0,253	**0,166**

Tablo 4.18'e göre, SkyWords F1 değeri açısından Sem2017 haricindeki tüm veri kümelerinde karşılaştırılan yöntemlerden daha iyi sonuç elde etmiştir. SkyWords Sem2017'de en iyi üçüncü değeri elde etmiştir. Sem2017'de, kesinlik ve duyarlılık sonuçlarında olduğu gibi en iyi sonuç SIFRank ve en iyi ikinci sonuç PositionRank tarafından elde etmiştir. İstatistiksel olarak, WWW veri kümesinde SkyWords KEA yöntemi haricindeki diğer yöntemlere üstünlük göstermiştir. WWW'de SkyWords ile KEA arasında istatistiksel olarak bir fark oluşmamıştır. Sem2017 veri kümesinde SIFRank ve PositionRank SkyWords'e karşı istatistiksel olarak bir fark göstermemiştir.

Kesinlik, duyarlılık ve F1 değeri ölçütlerinde elde edilen sonuçlar SkyWords'ün başarılı olduğunu göstermektedir. Kabasakal ve arkadaşının yapmış olduğu çalışmada (Kabasakal ve Mutlu, 2021), anahtar kelimeler Sem2017 veri kümesinde metnin geneline dengeli bir şekilde dağılmaktadır. Diğer veri kümelerinde ise metinlerin ilk kısımlarında anahtar kelimeler daha sık bulunmaktadır. SkyWords'te kullanılan pozisyon ve merkezilik ölçütleri bir metnin anahtar kelimelerinin dağılımını dikkate almaktadır. Bu nedenle SkyWords bir metinde eşit olarak dağılım gösteren Sem2017'de en iyi sonucu elde edememiştir.

Tablo 4.19 MRR sonuçlarını göstermektedir. SkyWords Nguyen hariç diğer veri kümelerinde en iyi MRR değerine sahiptir. SkyWords Nguyen veri kümesinde WINGNUS'dan sonra en iyi MRR değerini elde etmiştir. İstatistiksel olarak karşılaştırıldığında SkyWords KDD, Krapivin ve Sem2010'da karşılaştırılan yöntemlerden üstündür. WWW'de SkyWords KEA ve WINGNUS'dan daha iyi MRR değeri elde etmesine rağmen istatistiksel olarak üstünlük sağlayamamıştır. Ancak SkyWords diğer yöntemlerden istatistiksel olarak daha üstündür. Nguyen'de en iyi sonucu WINGNUS elde etmiştir ancak SkyWords'ten istatistiksel olarak üstün değildir. Nguyen'de MultipartiteRank ve KEA ile SkyWords arasında istatistiksel olarak bir fark gözlemlenmemiştir. SkyWords diğer yöntemlerden istatistiksel olarak üstündür. Sem2017'de en iyi MRR değeri SkyWords'te olmasına rağmen SIFRank'e istatistiksel olarak üstünlük yoktur. Sem2017'de SkyWords diğer yöntemlere göre istatistiksel fark göstermiştir. Bu sonuç SkyWords'ün ilk bulduğu anahtar kelimelerin doğru bulma başarısını göstermektedir.

Tablo 4.19. SkyWords Karşılaştırmalı Sonuçlar: MRR

Yöntem	KDD	WWW	Nguyen	Krapivin	Sem2017	Sem2010
KEA	0,294▼	0,307	0,433	0,301▼	0,521▼	0,308▼
WINGNUS	0,288▼	0,282	**0,474**	0,302▼	0,533▼	0,267▼
SIFRank	0,151▼	0,152▼	0,256▼	0,193▼	0,63	0,291▼
Key2vec	0,126▼	0,111▼	0,2▼	0,103▼	0,348▼	0,189▼
RAKE	0,055▼	0,059▼	0,135▼	0,098▼	0,374▼	0,142▼
YAKE!	0,159▼	0,126▼	0,279▼	0,286▼	0,427▼	0,226▼
MpRank	0,265▼	0,255▼	0,462	0,287▼	0,612▼	0,302▼
TopicRank	0,217▼	0,235▼	0,402▼	0,256▼	0,61▼	0,317▼
PositionRank	0,288▼	0,242▼	0,4▼	0,287▼	0,606▼	0,288▼
SingleRank	0,158▼	0,147▼	0,281▼	0,193▼	0,583▼	0,249▼
TextRank	0,156▼	0,155▼	0,259▼	0,171▼	0,368▼	0,229▼
SkyWords	**0,387**	**0,309**	0,454	**0,372**	**0,677**	**0,575**

Tablo 4.20 MAP sonuçlarını göstermektedir. Sonuçlar SkyWords'ün diğer yöntemlere göre tüm veri kümelerinde en iyi MAP değerini elde ettiğini göstermektedir. KDD, Krapivin, Sem2017 ve Sem2010'da SkyWords diğer yöntemlerden istatistiksel olarak daha üstündür. WWW'de SkyWords KEA ve WINGNUS'a istatistiksel olarak bir fark göstermemiştir. Nguyen'de SkyWords'ün KEA, WINGNUS ve MultipartiteRank'e göre istatistiksel bir üstünlüğü yoktur.

Tablo 4.20. SkyWords Karşılaştırmalı Sonuçlar: MAP

Yöntem	KDD	WWW	Nguyen	Krapivin	Sem2017	Sem2010
KEA	0,269▼	0,289	0,389	0,284▼	0,451▼	0,289▼
WINGNUS	0,269▼	0,265	0,415	0,277▼	0,464▼	0,247▼
SIFRank	0,142▼	0,149▼	0,233▼	0,185▼	0,555▼	0,267▼
Key2vec	0,124▼	0,107▼	0,183▼	0,1▼	0,314▼	0,177▼
RAKE	0,054▼	0,059▼	0,135▼	0,096▼	0,347▼	0,139▼
YAKE!	0,156▼	0,123▼	0,256▼	0,267▼	0,376▼	0,201▼
MpRank	0,246▼	0,242▼	0,412	0,268▼	0,517▼	0,283▼
TopicRank	0,211▼	0,221▼	0,357▼	0,239▼	0,52▼	0,297▼
PositionRank	0,267▼	0,233▼	0,358▼	0,269▼	0,528▼	0,262▼
SingleRank	0,157▼	0,143▼	0,273▼	0,182▼	0,514▼	0,232▼
TextRank	0,155▼	0,151▼	0,243▼	0,166▼	0,324▼	0,221▼
SkyWords	**0,36**	**0,295**	**0,42**	**0,346**	**0,599**	**0,507**

Tablo 4.21 SkyWords'ün ve diğer yöntemlerin doğru bulduğu toplam anahtar kelimele sayılarını göstermektedir. SkyWords Sem2017 ve WWW haricinde tüm veri kümeleri

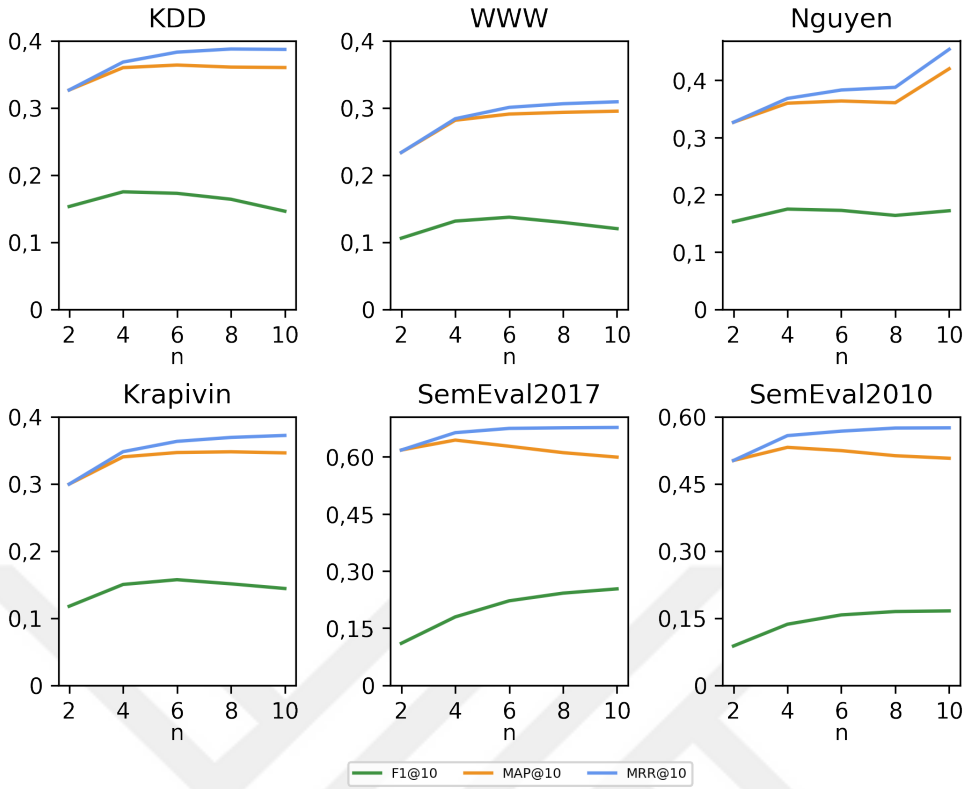
için diğer yöntemlerden daha çok sayıda anahtar kelimeyi doğru bulmuştur. Sem2017'de en çok anahtar kelimeyi SIFRank elde etmiştir. Sem2017'de SkyWords dördüncü sıradadır. WWW'de en çok anahtar kelimeyi KEA çıkarmıştır. SkyWords WWW'de KEA'dan 25 anahtar kelime daha az doğru bulmuştur ve ikinci sıradadır. SkyWords Sem2010 ve Krapivin'de diğer yöntemlere göre daha yüksek sayıda anahtar kelimeyi doğru bulmuştur. KEA SkyWords'ten KDD'de 11, Nguyen'de 15 tane eksik anahtar kelime bulmuştur ve en yüksek sayıda anahtar kelimeyi doğru bulan ikinci yöntem olmuştur.

Tablo 4.21. SkyWords Bulunan Anahtar Kelime Sayıları

Yöntem	KDD	WWW	Nguyen	Krapivin	Sem2017	Sem2010
KEA	285	**592**	348	2208	1174	240
WINGNUS	212	468	330	2280	1685	278
SIFRank	170	376	266	1879	**1921**	304
Key2vec	118	260	162	982	921	175
RAKE	91	229	175	1305	1182	155
YAKE!	187	355	236	1932	997	210
MpRank	179	433	287	1866	1421	225
TopicRank	148	386	257	1646	1379	240
PositionRank	230	515	330	2142	1613	273
SingleRank	196	410	281	1882	1722	274
TextRank	151	332	192	1271	835	196
SkyWords	**296**	567	**363**	**2569**	1615	**508**

SkyWords farklı anahtar kelime sayıları için test edilmiştir. Şekil 4.2 MAP, MRR ve F1 değeri bakımından 2, 4, 6, 8 ve 10 anahtar kelime sayısı için elde edilen sonuçları göstermektedir. Şekilde satır değerleri anahtar kelime sayılarını (n) sütun değerleri elde edilen sonuçları göstermektedir.

F1 değeri Sem2010 ve Sem2017 veri kümesinde anahtar kelime sayısı arttıkça artmıştır. Diğer veri kümelerinde önce artmış daha sonra azalmıştır. MRR değeri tüm veri kümelerinde Nguyen hariç anahtar kelime sayısı arttıkça artış eğilimi göstermektedir ancak 4 anahtar kelimeden sonra artış eğilimi azalmıştır ve daha sonra da sabitlenme eğilimi göstermiştir. Nguyen'de sonradan keskin bir artış gözlemlenmiştir. Anahtar kelime sayısının artışına bağlı olarak MRR ölçütünün sabitlenme değerinin erken olması bulunan anahtar kelimelerin ilk önerilerde bulunduğunu gösterir. MAP MRR ile benzer

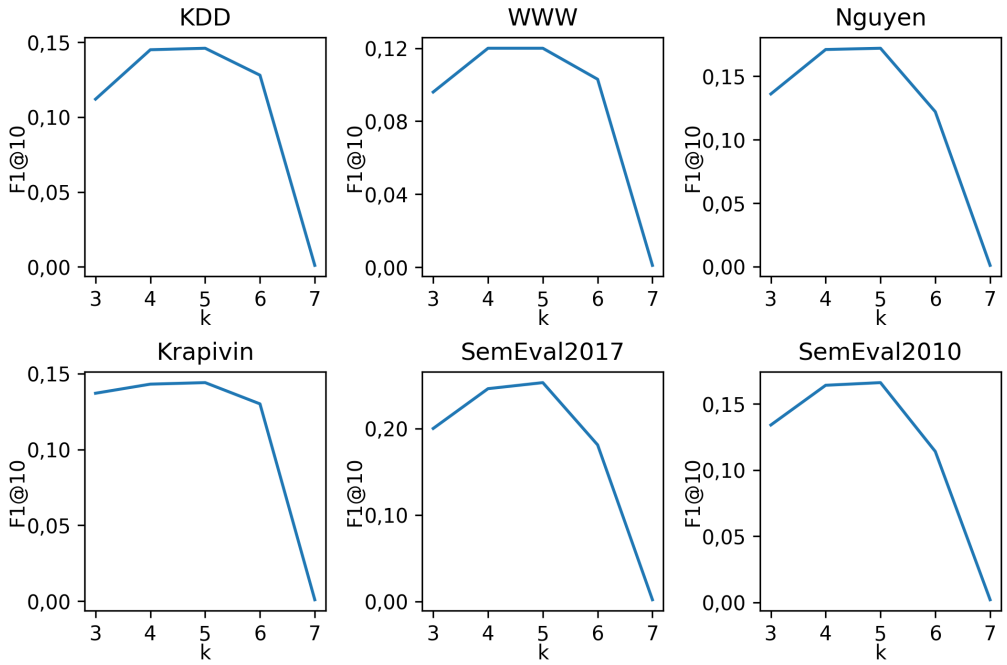


Şekil 4.2. Anahtar Kelime Sayılarına Göre Sonuçlar

sonuçları içermektedir.

Yapılan deneylerde çoğunluk oylama prensibi şu şekilde uygulanmıştır: Özellik mühendisliği aşamasında belirlenen 8 özellikten 5 özelliğin belirlenen eşik değerini sağlaması özelliği geçerli kılmaktadır. Bu bölümde bu parametre değerinin farklı değerlerle testi gerçekleştirilmiştir. Eşik değerinin sağlanması için gereken minimum özellik sayısı sırasıyla 3, 4, 5, 6 ve 7 değeri alınmıştır. Çoğunluk oylama prensibinde özelliklerin yarısından az sayıda şartın sağlanması durumunda anlamsız olmaktadır. Ancak genel eğilimi tespit etmek için 3 değeri için de test yapılmıştır. Şekil 4.2 belirlenen değerler için elde edilen sonuçların F1 değerini göstermektedir.

Parametre değeri k 3'ten 4 değerine geçildiğinde tüm veri kümelerinde F1 değerinin yükseldiği görülmektedir. 4 ve 5 değeri arasında önemli bir fark gözlemlenmemiştir. Ancak genelde 5 değeri ile yapılan deneyler 4 değeri ile yapılan deneylere göre daha iyi sonuç vermiştir. 6 değeri için düşüş gözlemlenmekle birlikte 7 parametresinde keskin bir düşüş görülmektedir. Bu parametrenin 0'a yaklaşması Skyline operatörünün etkisini



Şekil 4.3. Çoğunluk Oylama Prensibi: Parametre Analizi (k değeri)

ortadan kaldırmaktadır. 8'e yaklaşması da Skyline operatörünün etkisini arttırmaktadır. Parametre değeri 6 ile yapılan deneylerden sonra Skyline operatörü 0'a yakın aday anahtar kelime üretmiştir ve F1 değerinde keskin bir düşüş gözlemlenmiştir.

Tablo 4.22'de SkyWords ve karşılaştırılan yöntemlerin ortalama anahtar kelime çıkarma süreleri verilmiştir.

Tablo 4.22. SkyWords Anahtar Kelime Çıkarma Süreleri

Yöntem	**KDD**	**WWW**	**Nguyen**	**Krapivin**	**SemEval2017**	**SemEval2010**
KEA	0,11	0,51	0,18	0,17	0,1	0,11
WINGNUS	0,11	0,46	0,17	0,16	0,1	0,1
SIFRank	0,52	0,52	0,45	0,55	0,51	0,48
Key2vec	0,15	0,13	0,12	0,15	0,17	0,14
RAKE	0,01	0,01	0,01	0,01	0,01	0,01
YAKE!	0,4	0,04	0,04	0,43	0,05	0,38
MpRank	0,02	0,02	0,02	0,02	0,02	0,02
TopicRank	0,01	0,01	0,01	0,01	0,01	0,01
PositionRank	0,04	0,07	0,01	0,01	0,01	0,01
TextRank	0,01	0,01	0,01	0,01	0,01	0,01
SingleRank	0,01	0,05	0,01	0,01	0,03	0,04
SkyWords	0,31	0,30	0,28	0,34	0,29	0,30

Çizge tabanlı yöntemler, RAKE ve Key2Vec diğer yöntemlere göre daha hızlı çalışmaktadır. En yavaş çalışan yöntem SIFRank olmuştur. YAKE! SIFRank'ten sonra en yavaş çalışan yöntemdir. YAKE!, KEA ve WINGNUS'un ortalama anahtar kelime çıkarma süreleri veri kümesine göre değişiklik göstermektedir. WWW veri kümesinde SIFRank'ten sonra en yüksek sürede anahtar kelime çıkarma bulan yöntem sırasıyla KEA ve WINGNUS'tur. SkyWords'ün ortalama anahtar kelime çıkarma süreleri veri kümesinden bağımsız olarak belirli bir seviyede kalmaktadır. SkyWords SIFRank'ten daha iyi ortalama anahtar kelime çıkarma süresine sahiptir. YAKE! WINGNUS ve KEA'dan bazı veri kümelerinde daha iyi, bazı veri kümelerinde daha az iyi ortalama anahtar kelime çıkarma süresine sahiptir. Diğer yöntemler SkyWords'ten daha iyi ortalama anahtar kelime çıkarma süresi elde etmiştir.

SkyWords ve diğer yöntemler Tablo 4.3'te verilen örnek metinlerle test edilmiştir. Tablo 4.23 ve 4.24 elde edilen sonuçları göstermektedir.

Tablo 4.23. SkyWords Karşılaştırmalı Sonuçlar: KDD Örneği

Anahtar Kelimeler: online communities, social influence, social networks
SkyWords: *social influence*, *social networks*, similarity, *online communities*, social processes, feedback effect, social ties, relative effect, selection
KEA: *social influence*, social, influence, effects, similarity, feedback effects, understand, fundamental open, fundamental open question, open question
WINGNUS: *social influence*, fundamental open question, feedback effects, *online communities*, similarity, socialties, *social networks*, distinct reasons, interplay, current friends
SIFRank: *social influence*, everyday social processes, social ties, interplay, uniformity, similarity, social network, *social networks*, fragmentation, feedback effects
Key2Vec: social, *social influence*, influence, due, everyday social processes, distinct reasons, effects, real settings, social network, tend
RAKE: two distinct reasons, fundamental open question, form new links, current friends due, everyday social processes, challenge due, socialties, *social networks*, social network, *social influence*
YAKE!: fundamental open question, *social influence*, *online communities*, fundamental open, open question, social, social ties, *social networks*, similarity, feedback effects
MpRank: *social influence*, similarity, feedback effects, selection, process, *social networks*, *online communities*, uniformity, systems, tension
TopicRank: *social influence*, similarity, selection, feedback effects, uniformity, present, people, systems, behavior, everyday social processes
PositionRank *social influence*, everyday social processes, *social networks*, social ties, social network, feedback effects, fundamental open question, *online communities*, relative effects, similarity
SingleRank: everyday social processes, current friends due, fundamental open question, real settings, relative effects, new links, distinct reasons, social network, social ties, *social networks*
TextRank fundamental open question, social ties, *social networks*, *online communities*, *social influence*, feedback effects, effects, social, people, interplay

Tablo 4.23'e göre SkyWords tüm anahtar kelimeleri bulmuştur. Bulunan 3 anahtar

kelimeden 2'si ilk iki sırada, diğer anahtar kelime 5. sıradadır. WINGNUS, YAKE!, PositionRank, TextRank ve MultipartiteRank yöntemleri de tüm anahtar kelimeleri doğru bulmuştur. Bu yöntemlerden WINGNUS ilk tahmininde doğru anahtar kelimeyi bulmuştur ancak diğer anahtar kelimeler sırasıyla 5. ve 7. sıradadır. YAKE! 3., 4. ve 8. sıralarda doğru anahtar kelimeleri bulmuştur. Anahtar kelimeleri PositionRank 1., 3. ve 8.; TextRank 3., 4. ve 5. sırada doğru bulmuştur. MultipartiteRank'in ilk tahmini doğrudur ve diğer doğru tahminleri 6. ve 7. sıradadır. Her yöntem en az 1 tahmini doğru bulmuştur. Bu yöntemler arasından SingleRank tek doğru tahminini son sırada bulmuştur.

Tablo 4.24. SkyWords Karşılaştırmalı Sonuçlar: WWW Örneği

Anahtar Kelimeler: access control policies, fault model, mutation testing, test generation, testing tools
SkyWords: *mutation testing*, *access control policies*, *test generation*, automated testing, policy developer, manual testing, specified policies, evaluate coverage criteria, choosing coverage criteria, typical test input
KEA: *fault model*, mutation, fault, testing, policies, *access control policies*, control policies, access control, *mutation testing*, control
WINGNUS: *fault model*, *access control policies*, *mutation testing*, typical test inputs, specified policies, policy developers, policy testing, confidence, test outputs, correctness
SIFRank: *mutation testing*, typical test inputs, coverage criteria, *access control policies*, test outputs, policy testing, automated testing, mutation operators, mutant generation, mutant detection
Key2Vec: testing, policies, test, typical test inputs, tedious and few tools, structural coverage and fault detection effectiveness, detection, *access control policies*, policy testing, framework
RAKE: various policies written, *access control policies*, access control policies, access control policies, evaluate coverage criteria, choosing coverage criteria, choosing mutation operators, conduct policy testing, framework allows us, specified policies
YAKE!: *access control policies*, conduct policy testing, access control, typical test inputs, checking test outputs, supplying typical test, subsequently checking test, control policies, *fault model*, policy developers
MpRank *mutation testing*, *access control policies*, *fault model*, framework, coverage criteria, mutant generation, mutation operators, selection, *test generation*, equivalent
TopicRank *access control policies*, *fault model*, *mutation testing*, framework, test outputs, coverage criteria, mutant generation, mutation operators, selection, equivalent
PositionRank: *access control policies*, *fault model*, *mutation testing*, policy testing, fault, manual testing, mutation operators, *test generation*, various policies, typical test inputs
SingleRank: typical test inputs, *access control policies*, valuable insights, experimental results, various policies, detection effectiveness, structural coverage, *test generation*, coverage criteria, mutant detection
TextRank: *access control policies*, typical test, policy testing, policy developers, *mutation testing*, *fault model*, fault, mutation, testing, test

Tablo 4.24'e göre SkyWords 5 anahtar kelimeden 3'ünü doğru tahmin etmiştir ve bu tahminleri ilk 3 sırada yapmıştır. SkyWords gibi TopicRank ve WINGNUS yöntemleri 3 tahmini doğru yapmış ve ilk 3 sırada tamamlamıştır ancak doğru bulunan anahtar

kelimelerin sırası farklıdır. TextRank ve KEA yöntemleri sırasıyla 1., 5., 6. ve 1., 6., 9. sırada toplam 3 anahtar kelimeyi doğru bulmuştur. RAKE ve Key2Vec 1 anahtar kelimeyi, SIFRank ve YAKE! 2 anahtar kelimeyi doğru bulmuştur. MultipartiteRank ve PositionRank ilk tahminleri doğru olmak üzere 4 anahtar kelimeyi doğru bulmuştur. Diğer doğru tahminleri sırasıyla 8. ve 9. sıradadır. Tüm anahtar kelimeleri doğru bulan yöntem olmamıştır.

4.5.3. SkyRank

Bu bölümde SkyRank'le ilgili deneysel sonuçlar sunulmaktadır.

Tablo 4.25 kesinlik sonuçlarını göstermektedir. SkyRank tüm veri kümelerinde tüm yöntemlerde daha iyi kesinlik değeri elde etmiştir. İstatistiksel olarak SkyRank KDD, Nguyen, Krapivin ve Sem2010'da karşılaştırılan yöntemlerle arasında istatistiksel olarak fark oluşmuştur. Sem2017'de SIFRank ve SkyRank'in aralarında istatistiksel olarak bir fark yoktur. WWW'de KEA ile SkyRank arasında istatistiksel bir fark yoktur. Diğer veri kümelerinde SkyRank istatistiksel olarak daha üstündür.

Tablo 4.25. SkyRank Karşılaştırmalı Sonuçlar: Kesinlik

Yöntem	KDD	WWW	Nguyen	Krapivin	Sem2017	Sem2010
KEA	0,087▼	0,082	0,149▼	0,083▼	0,212▼	0,088▼
WINGNUS	0,07▼	0,071▼	0,148▼	0,09▼	0,303▼	0,103▼
SIFRank	0,058▼	0,057▼	0,128▼	0,082▼	0,389	0,125▼
YAKE!	0,064▼	0,053▼	0,113▼	0,084▼	0,202▼	0,086▼
MpRank	0,061▼	0,065▼	0,138▼	0,081▼	0,288▼	0,093▼
TopicRank	0,05▼	0,058▼	0,124▼	0,072▼	0,28▼	0,099▼
PositionRank	0,078▼	0,077▼	0,158▼	0,093▼	0,327▼	0,112▼
SingleRank	0,067▼	0,062▼	0,135▼	0,082▼	0,349▼	0,113▼
SkyRank	**0,106**	**0,09**	**0,185**	**0,118**	**0,391**	**0,239**

Tablo 4.26 duyarlılık sonuçlarını göstermektedir. Duyarlılık sonuçları kesinlik sonuçları ile benzerlik göstermektedir. SkyRank Sem2017 hariç tüm deneylerde diğer yöntemlere üstünlük sağlamıştır. Sem2017'de SIFRank ile aynı değeri elde etmiştir ve istatistiksel olarak aralarında bir fark yoktur. WWW'de SkyRank ile KEA arasında istatistiksel bir fark yoktur ve SkyRank diğer yöntemlerden istatistiksel olarak üstündür. Diğer yöntemlere karşı SkyRank KDD, Nguyen, Krapivin ve Sem2010'da istatistiksel fark

oluşturmuştur.

Tablo 4.26. SkyRank Karşılaştırmalı Sonuçlar: Duyarlılık

Yöntem	KDD	WWW	Nguyen	Krapivin	Sem2017	Sem2010
KEA	0,232▼	0,197	0,162▼	0,175▼	0,13▼	0,06▼
WINGNUS	0,184▼	0,167▼	0,16▼	0,175▼	0,193▼	0,067▼
SIFRank	0,157▼	0,137▼	0,14▼	0,153▼	**0,243**	0,082▼
YAKE!	0,172▼	0,129▼	0,12▼	0,175▼	0,129▼	0,057▼
MpRank	0,16▼	0,152▼	0,14▼	0,159▼	0,18▼	0,06▼
TopicRank	0,131▼	0,134▼	0,13▼	0,139▼	0,172▼	0,064▼
PositionRank	0,212▼	0,184▼	0,169▼	0,184▼	0,204▼	0,074▼
SingleRank	0,18▼	0,149▼	0,144▼	0,158▼	0,221▼	0,075▼
SkyRank	**0,279**	**0,214**	**0,203**	**0,234**	**0,243**	**0,157**

Tablo 4.27 F1 değeri için sonuçları göstermektedir. SkyRank tüm veri kümelerine karşı en iyi F1 değerini elde etmiştir. 3 veri kümesinde KEA, 2 veri kümesinde SIFRank ve 1 veri kümesinde PositionRank en iyi 2. değeri elde etmiştir. SkyRank ile SIFRank arasında Sem2017'de istatistiksel bir fark yoktur. WWW'de ise SkyRank ile KEA arasında istatistiksel bir fark oluşmamıştır. Diğer karşılaştırmalarda SkyRank SIFRank'ten ve KEA'dan istatistiksel olarak daha üstündür. SkyRank ile KEA ve SIFRank dışındaki diğer yöntemler arasında tüm veri kümelerinde istatistiksel olarak fark vardır.

Tablo 4.27. SkyRank Karşılaştırmalı Sonuçlar: F1 Değeri

Yöntem	KDD	WWW	Nguyen	Krapivin	Sem2017	Sem2010
KEA	0,124▼	0,112	0,143▼	0,108▼	0,155▼	0,071▼
WINGNUS	0,099▼	0,097▼	0,142▼	0,112▼	0,226▼	0,08▼
SIFRank	0,082▼	0,078▼	0,123▼	0,1▼	0,288	0,098▼
YAKE!	0,091▼	0,073▼	0,107▼	0,109▼	0,151▼	0,068▼
MpRank	0,086▼	0,088▼	0,128▼	0,102▼	0,213▼	0,072▼
TopicRank	0,071▼	0,079▼	0,117▼	0,09▼	0,205▼	0,077▼
PositionRank	0,112▼	0,106▼	0,15▼	0,117▼	0,242▼	0,088▼
SingleRank	0,095▼	0,085▼	0,128▼	0,102▼	0,261▼	0,089▼
SkyRank	**0,15**	**0,123**	**0,179**	**0,149**	**0,289**	**0,187**

Tablo 4.28 MRR sonuçlarını göstermektedir. SkyRank MRR değeri açısından KDD, Krapivin, Sem2017 ve Sem2010'da en iyi değeri elde etmiştir ve istatistiksel açıdan tüm yöntemlerle aralarında istatistiksel fark oluşmuştur. WWW'de SkyRank en iyi MRR

değerini elde etmiştir ancak KEA ile SkyRank arasında istatistiksel bir fark yoktur. WWW'de SkyRank WINGNUS'dan daha iyi MRR değeri elde etmiştir ancak aralarında istatistiksel bir fark yoktur. SkyRank diğer yöntemlerden WWW'de istatistiksel olarak daha üstündür. Nguyen'de en iyi MRR değeri WINGNUS'dadır ve SkyRank 2. sıradadır. Nguyen'de SkyRank MultipartiteRank ve WINGNUS yöntemlerine istatistiksel olarak üstünlük sağlayamamıştır ancak diğer yöntemlerden daha üstündür. SkyRank veri kümelerinin tümünde karşılaştırılan yöntemlerden daha iyi sonuç elde etmiştir.

Tablo 4.28. SkyRank Karşılaştırmalı Sonuçlar: MRR

Yöntem	KDD	WWW	Nguyen	Krapivin	Sem2017	Sem2010
KEA	0,294▼	0,307	0,433	0,301▼	0,521▼	0,308▼
WINGNUS	0,288▼	0,282	**0,47**	0,302▼	0,533▼	0,267▼
SIFRank	0,151▼	0,152▼	0,256▼	0,193▼	0,63▼	0,291▼
YAKE!	0,159▼	0,126▼	0,279▼	0,286▼	0,427▼	0,226▼
MpRank	0,265▼	0,255▼	0,464	0,287▼	0,612▼	0,302▼
TopicRank	0,217▼	0,235▼	0,402▼	0,256▼	0,61▼	0,317▼
PositionRank	0,288▼	0,242▼	0,4▼	0,287▼	0,606▼	0,288▼
SingleRank	0,158▼	0,147▼	0,281▼	0,193▼	0,583▼	0,249▼
SkyRank	**0,388**	**0,308**	0,458	**0,371**	**0,703**	**0,59**

Tablo 4.29 MAP sonuçlarını göstermektedir.

Tablo 4.29. SkyRank Karşılaştırmalı Sonuçlar: MAP

Yöntem	KDD	WWW	Nguyen	Krapivin	Sem2017	Sem2010
KEA	0,269▼	0,289	0,389▼	0,284▼	0,451▼	0,289▼
WINGNUS	0,269▼	0,265▼	0,415	0,277▼	0,464▼	0,247▼
SIFRank	0,142▼	0,149▼	0,233▼	0,185▼	0,555▼	0,267▼
YAKE!	0,156▼	0,123▼	0,256▼	0,267▼	0,376▼	0,201▼
MpRank	0,246▼	0,242▼	0,412	0,268▼	0,517▼	0,283▼
TopicRank	0,211▼	0,221▼	0,357▼	0,239▼	0,52▼	0,297▼
PositionRank	0,267▼	0,233▼	0,358▼	0,269▼	0,528▼	0,262▼
SingleRank	0,157▼	0,143▼	0,273▼	0,182▼	0,514▼	0,232▼
SkyRank	**0,359**	**0,29**	**0,42**	**0,342**	**0,607**	**0,507**

SkyRank MAP değeri açısından tüm veri kümelerinde karşılaştırılan yöntemlerden daha iyi sonuç elde etmiştir. SIFRank Nguyen'de en düşük ve Sem2017'de en yüksek 2. MAP değerine sahiptir. İstatistiksel olarak Nguyen'de SkyRank'in WINGNUS ve MultipartiteRank'ten bir farkı yoktur. Nguyen'de SkyRank diğer yöntemlerden daha

üstündür. WWW'de KEA ile SkyRank arasında istatistiksel bir fark oluşmamıştır. Karşılaştırmada kullanılan diğer veri kümeleri açısından SkyRank tüm deneylerde istatistiksel olarak daha üstündür.

Tablo 4.30 doğru bulunan anahtar kelimelerin toplam sayısını göstermektedir. SkyRank tüm veri kümelerinde en çok sayıda anahtar kelimeyi bulan yöntem olmuştur.

Tablo 4.30. SkyRank: Doğru Anahtar Kelime Sayıları

Yöntem	KDD	WWW	Nguyen	Krapivin	Sem2017	Sem2010
KEA	285	592	348	2208	1174	240
WINGNUS	212	468	330	2280	1685	278
SIFRank	170	376	266	1879	1921	304
YAKE!	187	355	236	1932	997	210
MpRank	179	433	287	1866	1421	225
TopicRank	148	386	257	1646	1379	240
PositionRank	230	515	330	2142	1613	273
SingleRank	196	410	281	1882	1722	274
SkyRank	**311**	**602**	**388**	**2727**	**1928**	**580**

Tablo 4.31 SkyRank ve karşılaştırılan yöntemlerin ortalama anahtar kelime çıkarma sürelerini göstermektedir. SkyRank en uzun sürede anahtar kelime çıkaran yöntem olmuştur. SkyRank'ten sonra SIFRank ve YAKE! yavaş çalışan yöntemler olmuştur. İstatistiksel ve önceden eğitilmiş kelime vektörü kullanan yöntemlerin diğer yöntemlere göre daha yavaş çalıştığı söylenebilir.

Tablo 4.31. SkyRank Anahtar Kelime Çıkarma Süreleri

Yöntem	KDD	WWW	Nguyen	Krapivin	SemEval2017	SemEval2010
KEA	0,11	0,51	0,18	0,17	0,1	0,11
WINGNUS	0,11	0,46	0,17	0,16	0,1	0,1
SIFRank	0,52	0,52	0,45	0,55	0,51	0,48
YAKE!	0,4	0,04	0,04	0,43	0,05	0,38
MpRank	0,02	0,02	0,02	0,02	0,02	0,02
TopicRank	0,01	0,01	0,01	0,01	0,01	0,01
PositionRank	0,04	0,07	0,01	0,01	0,01	0,01
SingleRank	0,01	0,05	0,01	0,01	0,03	0,04
SkyRank	2,38	2,34	2,55	2,34	2,4	2,16

SkyRank ve karşılaştırma yapılan yöntemler Tablo 4.3'de verilen örnek metinlerle test

edilmiştir. Tablo 4.32'ye göre SkyRank 3 anahtar kelimeden 2 anahtar kelimeyi doğru bulmuştur. Bu anahtar kelimeler ilk 2 sırada bulunmuştur. WINGNUS, YAKE!, PositionRank ve MultipartiteRank yöntemleri tüm anahtar kelimeleri bulmuştur. KEA, SingleRank ve TopicRank 1 anahtar kelimeyi doğru bulmuştur. SIFRank 2 anahtar kelimeyi doğru bulmuştur. Bu anahtar kelimelerden birisi 1. sırada, diğeri 8. sıradadır.

Tablo 4.32. SkyRank Karşılaştırmalı Sonuçlar: KDD Örneği

Anahtar Kelimeler: online communities, social influence, social networks
SkyRank: *social influence*, *social networks*, similarity, social processes, feedback effect, social ties, relative effect, selection, challenge, real settings
KEA: *social influence*, social, influence, effects, similarity, feedback effects, understand, fundamental open, fundamental open question, open question
WINGNUS: *social influence*, fundamental open question, feedback effects, *online communities*, similarity, socialties, *social networks*, distinct reasons, interplay, current friends
SIFRank: *social influence*, everyday social processes, social ties, interplay, uniformity, similarity, social network, *social networks*, fragmentation, feedback effects
YAKE!: fundamental open question, *social influence*, *online communities*, fundamental open, open question, social, social ties, *social networks*, similarity, feedback effects
MpRank: *social influence*, similarity, feedback effects, selection, process, *social networks*, *online communities*, uniformity, systems, tension
TopicRank: *social influence*, similarity, selection, feedback effects, uniformity, present, people, systems, behavior, everyday social processes
PositionRank *social influence*, everyday social processes, *social networks*, social ties, social network, feedback effects, fundamental open question, *online communities*, relative effects, similarity
SingleRank: everyday social processes, current friends due, fundamental open question, real settings, relative effects, new links, distinct reasons, social network, social ties, *social networks*

Tablo 4.33'e göre SkyRank 5 anahtar kelimeden 3 anahtar kelimeyi doğru bulmuştur. Bu anahtar kelimeler ilk 3 sırada bulunmuştur. TopicRank ve WINGNUS 3 anahtar kelimeyi ilk 3 sırada doğru bulmuştur. KEA 3 anahtar kelimeyi, SIFRank ve YAKE! 2 anahtar kelimeyi doğru etiketlemiştir. MultipartiteRank ve PositionRank 4 anahtar kelimeyi doğru bulmuştur.

4.5.4. Geliştirilen Yöntemlerin Karşılaştırılması

Bu bölümde tez kapsamında geliştirilen MGRank, SkyWords ve SkyRank yöntemleri karşılaştırılmaktadır. Karşılaştırma KDD, WWW, Nguyen, Krapivin, Sem2017 ve Sem2010 veri kümelerinde gerçekleştirilmiştir. Karşılaştırmada kesinlik, duyarlılık, F1 değeri, MRR ve MAP ölçütleri kullanılmıştır. Ayrıca ortalama anahtar kelime çıkarma süreleri değerlendirmeye dahil edilmiştir.

Tablo 4.33. SkyRank Karşılaştırmalı Sonuçlar: WWW Örneği

Anahtar Kelimeler: access control policies, fault model, mutation testing, test generation, testing tools
SkyRank: *mutation testing*, *access control policies*, *test generation*, automated testing, policy developers, manual testing, equivalent-mutant detection, specified policies, evaluate coverage criteria, choosing coverage criteria
KEA: *fault model*, mutation, fault, testing, policies, *access control policies*, control policies, access control, *mutation testing*, control
WINGNUS: *fault model*, *access control policies*, *mutation testing*, typical test inputs, specified policies, policy developers, policy testing, confidence, test outputs, correctness
SIFRank: *mutation testing*, typical test inputs, coverage criteria, *access control policies*, test outputs, policy testing, automated testing, mutation operators, mutant generation, mutant detection
YAKE!: *access control policies*, conduct policy testing, access control, typical test inputs, checking test outputs, supplying typical test, subsequently checking test, control policies, *fault model*, policy developers
MpRank *mutation testing*, *access control policies*, *fault model*, framework, coverage criteria, mutant generation, mutation operators, selection, *test generation*, equivalent
TopicRank *access control policies*, *fault model*, *mutation testing*, framework, test outputs, coverage criteria, mutant generation, mutation operators, selection, equivalent
PositionRank: *access control policies*, *fault model*, *mutation testing*, policy testing, fault, manual testing, mutation operators, *test generation*, various policies, typical test inputs
SingleRank: typical test inputs, *access control policies*, valuable insights, experimental results, various policies, detection effectiveness, structural coverage, *test generation*, coverage criteria, mutant detection

Tablo 4.34 geliştirilen yöntemlerin kesinlik, duyarlılık ve F1 değeri açısından sonuçlarını göstermektedir. SkyRank kesinlik, duyarlılık ve F1 Değeri ölçütünde tüm veri kümelerinde SkyWords ve MGRank'ten daha iyi performans göstermiştir. SkyWords ile MGRank karşılaştırıldığında Sem2017 veri kümesinde MGRank daha yüksek kesinlik, duyarlılık ve F1 değeri elde etmiştir. Diğer veri kümelerinde ise SkyWords üç ölçüt için MGRank'ten daha iyi sonuç elde etmiştir.

Tablo 4.34. Geliştirilen Yöntemlerin Karşılaştırması

Yöntem	Ölçüt	KDD	WWW	Nguyen	Krapivin	Sem2017	Sem2010
MGRank	Kesinlik	0,09	0,078	0,165	0,106	0,348	0,199
	Duyarlılık	0,239	0,186	0,18	0,209	0,219	0,131
	F1 Değeri	0,128	0,106	0,16	0,133	0,259	0,156
SkyWords	Kesinlik	0,104	0,089	0,179	0,116	0,352	0,216
	Duyarlılık	0,269	0,204	0,191	0,222	0,210	0,138
	F1 Değeri	0,146	0,12	0,172	0,144	0,253	0,166
SkyRank	Kesinlik	**0,106**	**0,09**	**0,185**	**0,118**	**0,391**	**0,239**
	Duyarlılık	**0,279**	**0,214**	**0,203**	**0,234**	**0,243**	**0,157**
	F1 Değeri	**0,15**	**0,123**	**0,179**	**0,149**	**0,289**	**0,187**

Tablo 4.35 geliştirilen yöntemlerin MRR ve MAP ölçütleri için elde edilen sonuçları

göstermektedir. MRR karşılaştırmasında SkyWords WWW ve Krapivin veri kümelerinde en iyi değeri elde etmiştir. SkyRank KDD, Nguyen, Sem2017 ve Sem2010 veri kümelerinde en iyi MRR değerine sahiptir. MGRank MRR açısından geliştirilen yöntemler arasında tüm veri kümelerinde en düşük değeri elde eden yöntemdir. MAP açısından geliştirilen yöntemler değerlendirildiğinde SkyWords KDD, WWW ve Krapivin veri kümelerinde en iyi değeri elde etmiştir. Nguyen ve Sem2010 veri kümelerinde SkyWords ve SkyRank aynı değeri elde etmiştir ve MGRank'ten daha iyi performans göstermişlerdir. Sem2017 veri kümesinde SkyRank en iyi MAP değerini elde etmiştir. MAP açısından MGRank SkyWords ve SkyRank'e göre tüm veri kümelerinde daha düşük değere sahip olmuştur.

Tablo 4.35. Geliştirilen Yöntemlerin Karşılaştırması: MRR, MAP

Yöntem	**Ölçüt**	**KDD**	**WWW**	**Nguyen**	**Krapivin**	**Sem2017**	**Sem2010**
MGRank	MRR	0,306	0,242	0,397	0,308	0,64	0,469
	MAP	0,276	0,235	0,371	0,283	0,551	0,416
SkyWords	MRR	0,387	**0,309**	0,454	**0,372**	0,677	0,575
	MAP	**0,36**	**0,295**	**0,42**	**0,346**	0,599	**0,507**
SkyRank	MRR	**0,388**	0,308	**0,458**	0,371	**0,703**	**0,59**
	MAP	0,359	0,29	**0,42**	0,342	**0,607**	**0,507**

Tablo 4.36 geliştirilen yöntemlerin ortalama anahtar kelime çıkarma sürelerini içermektedir. SkyWords MGRank ve SkyRank'e göre tüm veri kümelerinde en hızlı sürede anahtar kelime çıkaran yöntem olmuştur. MGRank'in ortalama anahtar kelime çıkarma süresi SkyRank'e göre daha düşüktür.

Tablo 4.36. Geliştirilen Yöntemlerin Anahtar Kelime Çıkarma Süreleri

Yöntem	**KDD**	**WWW**	**Nguyen**	**Krapivin**	**SemEval2017**	**SemEval2010**
MGRank	1,22	0,96	1,53	1,1	1,1	1,1
SkyWords	0,31	0,30	0,28	0,34	0,29	0,30
SkyRank	2,38	2,34	2,55	2,34	2,4	2,16

5. SONUÇLAR VE ÖNERİLER

Anahtar kelimeler bir metnin içeriğini en iyi şekilde temsil eden kelime ya da kelime gruplarıdır. Kümeleme, sınıflandırma, özetleme gibi birçok DDİ probleminde anahtar kelimeler önemli bir yer edinmiştir.

Anahtar kelimelerin el ile etiketlenmesinin öznel ve zorlayıcı olması anahtar kelimelerin otomatik araçlarla indekslenmesi ihtiyacını ortaya çıkarmıştır. İndekslemede kullanılan otomatik anahtar kelime çıkarma yöntemleri iki temel sınıfa ayrılabilir: denetimli öğrenmeye dayalı yaklaşımlar ve denetimsiz öğrenmeye dayalı yaklaşımlar. Anahtar kelime çıkarma problemini denetimli yaklaşımlar ikili sınıflandırma problemi olarak ele almaktadır. Denetimsiz öğrenmeye dayalı yaklaşımlar ise anahtar kelime çıkarma problemini sıralama problemi olarak ele alır. Denetimsiz yaklaşımların maliyetsiz olması denetimli yaklaşımların güçlü bir model sunması anahtar kelime çıkarma yöntemlerinde öne çıkmaktadır. Bu tez çalışmasında metinlerden otomatik olarak anahtar kelime çıkarmak için 3 yeni yöntem önerilmiştir.

Geliştirilen ilk yöntem, MGRank olarak adlandırılan çizge tabanlı denetimsiz öğrenmeye dayalı bir anahtar kelime çıkarma yöntemidir. MGRank düğümleri ve kenarları ağırlıklı, yönsüz, çok kenarlı tam çizge modeli kullanmaktadır.

MGRank ile şu katkılar elde edilmiştir:

- Çizge tabanlı birçok yöntem olmasına rağmen çok kenarlı tam çizge modelini benimseyen ilk anahtar kelime çıkarma yöntemidir.
- Çizge tabanlı yöntemlerde başlangıçta parametre olarak alınan pencere boyutu değerini ortadan kaldırmıştır.
- İki düğüm arasında birden fazla kenar bağlantısı kurarak kelimelerin daha iyi temsil edilmesini sağlamıştır.
- Çizgede önerilen kenar ağırlıklandırma prosedürü ile birbirine yakın kelimelerin birbirlerine olan etkisi arttırılmıştır.
- MGRank çizge tabanlı farklı yöntemlerle karşılaştırılmıştır ve birçok ölçütte daha iyi sonuç elde etmiştir. İstatistiksel analizler MGRank'in başarılı olduğunu

göstermektedir.

SkyWords olarak adlandırılan ikinci yöntem denetimli ve denetimsiz yaklaşımları birleştiren hibrit bir yöntemdir. Bu yöntem yüksek kalitede aday anahtar kelimeleri tespit etmeyi ve vektör temsiline dayalı bir yaklaşımla metne en çok benzeyen anahtar kelimeleri bulmayı amaçlamaktadır. Yüksek kalitede aday anahtar kelimeleri bulmak için Skyline operatörü ve çoğunluk oylama prensibi kullanılmıştır. all-mpnet-base-v2 cümle dönüşüm modeli ile elde edilen vektörler sayesinde bir metne en çok benzeyen anahtar kelimeler bulunmaktadır.

SkyWords ile şu katkılar elde edilmiştir:

- SkyWords Skyline operatörünün anahtar kelime çıkarma probleminde kullanıldığı ilk yöntemdir.
- SkyWords denetimli öğrenme yaklaşımı ile aday anahtar kelimelerin yüksek kalitede elde edilmesini sağlar ve aday anahtar kelime sayısını azaltarak arama uzayını daraltır.
- Aday anahtar kelimeleri bulmak için istatistiksel bir yaklaşım benimsenmiştir. Böylece hem çizge tabanlı merkezilik ölçütleri hem de anahtar kelime çıkarmada başarısı ispatlanmış birçok özellik birleştirilerek kelimelerin daha iyi temsil edilmesi sağlanmıştır.
- Anahtar kelime çıkarma yöntemlerinde aday anahtar kelimeler genellikle tek kelimelerden oluşur ve bu kelimelerden kelime öbeklerinin önem değerini elde etmek önemli bir problemdir. SkyWords kelime öbeklerinin önem değerini bulmak için vektör temsillerini kullanmaktadır ve metne en çok benzeyen kelime öbekleri anahtar kelime olarak alınmaktadır.
- Deneysel sonuçlar SkyWords'ün başarılı olduğunu göstermektedir.

Geliştirilen son yöntem SkyRank'tir. SkyRank istatistik tabanlı yaklaşıma sahip denetimsiz bir yöntemdir. SkyWords'te olduğu gibi Skyline operatörü ve all-mpnet-base-v2 cümle modelini kullanmaktadır.

- SkyRank Skyline operatörünü denetimsiz bir öğrenme modeli olarak

kullanmaktadır.

- SkyRank bir metnin kelimelerine ait çeşitli istatistiksel özellikler çıkararak metnin daha iyi temsil edilmesini sağlar.
- SkyWords'te olduğu gibi metin ve aday anahtar kelimelerin vektör temsilleri bulunur ve aralarındaki benzerlik hesaplanarak metne en çok benzeyen anahtar kelimeler elde edilir.
- SkyRank denetimsiz öğrenmeye dayalı bir yöntem olmasına rağmen denetimli öğrenme yöntemleri dahil, diğer yöntemlerden daha başarılı sonuçlar elde etmiştir.

Geliştirilen yöntemlerin başarısı anahtar kelime çıkarma problemi için akademik çalışmalardan oluşturulmuş veri kümelerinde test edilmiştir. Deneysel sonuçlar kesinlik, duyarlılık, F1 değeri, MRR ve MAP ölçütleri açısından incelenmiştir. Geliştirilen farklı öğrenme yaklaşımlarına sahip çeşitli anahtar kelime çıkarma yöntemleriyle karşılaştırılmıştır. Ayrıca örnek metinler üzerinden elde edilen sonuçlar ile niteliksel olarak karşılaştırma yapılmıştır. Sonuçlar geliştirilen yöntemlerin başarılı olduğunu göstermektedir.

İlerleyen çalışmalarda anahtar kelime listelerinde yer alan benzer kelime gruplarının azaltılması ve anahtar kelime çeşitliliğinin arttırılmasıyla ilgili çalışmalar yapılabilir.

KAYNAKLAR

Abulaish, M., Fazil, M., Zaki, M. J. (2022). Domain-Specific Keyword Extraction Using Joint Modeling of Local and Global Contextual Semantics. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *16*(4), 1–30.

Alrehamy, H., Walker, C. (2018). Exploiting Extensible Background Knowledge for Clustering-based Automatic Keyphrase Extraction. *Soft Computing*, *22*(21), 7041–7057.

Alzaidy, R., Caragea, C., Giles, C. L. (2019). Bi-LSTM-CRF Sequence Labeling for Keyphrase Extraction from Scholarly Documents. *The World Wide Web Conference*, San Francisco, USA, 13–17 May. 2019.

Aman, M., Abdulkadir, S. J., Aziz, I. A., Alhussian, H., Ullah, I. (2021). KP-Rank: A Semantic-based Unsupervised Approach for Keyphrase Extraction from Text Data. *Multimedia Tools and Applications*, *80*(8), 12469–12506.

Augenstein, I., Das, M., Riedel, S., Vikraman, L., McCallum, A. (2017). Semeval 2017 Task 10: ScienceIE-Extracting Keyphrases and Relations From Scientific Publications. *arXiv preprint arXiv:1704.02853*, 1–10.

Awan, M. N., Beg, M. O. (2021). TOP-Rank: A TopicalPostionRank for Extraction and Classification of Keyphrases in Text. *Comput. Speech Lang.*, *65,* 101116.

Basaldella, M., Antolli, E., Serra, G., Tasso, C. (2018). Bidirectional LSTM Recurrent Neural Network for Keyphrase Extraction. *Italian Research Conference on Digital Libraries*, Udine, Italy, 25–26 Ocak 2018.

Beliga, S., Meštrović, A., Martinčić-Ipšić, S. (2015). An Overview of Graph-based Keyword Extraction Methods and Approaches. *Journal of Information and Organizational Sciences*, *39*(1), 1–20.

Bellaachia, A., Al-Dhelaan, M. (2014). HG-Rank: A Hypergraph-based Keyphrase Extraction for Short Documents in Dynamic Genre. *4th Workshop on Making Sense of Microposts*, Seoul, Korea, 7–7 Nis. 2014.

El-Beltagy, S. R., Rafea, A. (2009). KP-Miner: A Keyphrase Extraction System for English and Arabic Documents. *Information Systems*, *34*(1), 132–144.

Bennani-Smires, K., Musat, C., Hossmann, A., Baeriswyl, M., Jaggi, M. (2018). Simple Unsupervised Keyphrase Extraction Using Sentence Embeddings. *arXiv preprint arXiv:1801.04470*,

Biswas, S. K., Bordoloi, M., Shreya, J. (2018). A Graph based Keyword Extraction Model Using Collective Node Weight. *Expert Systems with Applications*, *97,* 51–59.

Borzsony, S., Kossmann, D., Stocker, K. (2001). The Skyline Operator. *17th international conference on data engineering*, Heidelberg, Germany, 2–6 Nis. 2001.

Boudin, F. (2016). PKE: An Open Source Python-based Keyphrase Extraction Toolkit. *26th International Conference on Computational Linguistics: System Demonstrations*, Osaka, Japan, 11–16 Ara. 2016.

Boudin, F. (2013). A Comparison of Centrality Measures for Graph-based Keyphrase Extraction. *6th International Joint Conference on Natural Language Processing*, Nagoya, Japan, 14–18 Ekim 2013.

Boudin, F. (2018). Unsupervised Keyphrase Extraction with Multipartite Graphs. *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Louisiana, USA, 1–6 Haz. 2018.

Bougouin, A., Boudin, F., Daille, B. (2013). TopicRank: Graph-based Topic Ranking for Keyphrase Extraction. *Sixth International Joint Conference on Natural Language Processing*, Nagoya, Japan, 14–18 Ekim 2013.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., Jatowt, A. (2020). YAKE! Keyword Extraction from Single Documents Using Multiple Local Features. *Information Sciences*, *509,* 257–289.

Caragea, C., Bulgarov, F., Godea, A., Gollapalli, S. D. (2014). Citation-enhanced Keyphrase Extraction from Research Papers: A Supervised Approach. *2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, 26–28 Ekim 2014.

Chen, Y., Wang, J., Li, P., Guo, P. (2019). Single Document Keyword Extraction via Quantifying Higher-order Structural Features of Word Co-occurrence Graph. *Computer Speech and Language*, *57,* 98–107.

Chi, L., Hu, L. (2021). ISKE: An Unsupervised Automatic Keyphrase Extraction Approach Using the Iterated Sentences based on Graph Method. *Knowl. Based Syst.*, *223,* 107014.

Church, K. W. (2017). Word2Vec. *Natural Language Engineering*, *23*(1), 155–162.

Csomai, A. (2008). *Keywords in the Mist: Automated Keyword Extraction for Very Large Documents and Back of the Book Indexing*: University of North Texas.

Danesh, S., Sumner, T., Martin, J. H. (2015). SGRank: Combining Statistical and Graphical Methods to Improve the State of the Art in Unsupervised Keyphrase Extraction. *Fourth Joint Conference on Lexical ve Computational Semantics*, Denver, USA, 4–5 Haz. 2015.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 1–16.

Du, C., Huang, L. (2018). Text Classification Research with Attention-based Recurrent Neural Networks. *International Journal of Computers Communications & Control*, *13*(1), 50–61.

Duari, S., Bhatnagar, V. (2020). Complex Network Based Supervised Keyword Extractor. *Expert Systems with Applications*, *140,* 112876.

Duari, S., Bhatnagar, V. (2019). sCAKE: Semantic Connectivity Aware Keyword Extraction. *Information Sciences*, *477,* 100–117.

Figueroa, G., Chen, P.-C., Chen, Y.-S. (2018). RankUp: Enhancing Graph-based Keyphrase Extraction Methods with Error-feedback Propagation. *Computer Speech & Language*, *47,* 112–131.

Florescu, C., Caragea, C. (2017). Positionrank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. *55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, 30 Ağus.–4 Eyl. 2017.

Florescu, C., Jin, W. (2018). Learning Feature Representations for Keyphrase Extraction. *Thirty-Second AAAI Conference on Artificial Intelligence*, Louisiana, USA, 2–7 Şub. 2018.

Gagliardi, I., Artese, M. T. (2020). Semantic Unsupervised Automatic Keyphrases Extraction by Integrating Word Embedding with Clustering Methods. *Multimodal Technologies and Interaction*, *4*(2), 30.

Gollapalli, S. D., Caragea, C. (2014). Extracting Keyphrases from Research Papers Using Citation Networks. *Twenty-Eighth AAAI Conference on Artificial Intelligence*, Québec, Canada, 27–31 Tem. 2014.

Hasan, K. S., Ng, V. (2014). Automatic Keyphrase Extraction: A Survey of the State of the Art. *52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, 22–27 Haz. 2014.

Hernández-Castañeda, Á., García-Hernández, R. A., Ledeneva, Y., Millán-Hernández, C. E. (2021). Language-Independent Extractive Automatic Text Summarization Based on Automatic Keyword Extraction. *Comput. Speech Lang.*, 101267.

Hickman, L., Thapa, S., Tay, L., Cao, M., Srinivasan, P. (2022). Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations. *Organizational Research Methods*, *25*(1), 114–146.

HuggingFace (2022), https://huggingface.co/sentence-transformers/all-mpnet-base-v2 (Ziyaret Tarihi:25 Kas. 2022).

Hulth, A. (2003). Improved Automatic Keyword Extraction Given More Linguistic Knowledge. *2003 Conference on Empirical Methods in Natural Language Processing*, Sapporo, Japan, 11–12 Tem. 2003.

Jain, A., Mittal, K., Vaisla, K. S. (2022). FLAKE: Fuzzy Graph Centrality-based Automatic Keyword Extraction. *The Computer Journal*, *65*(4), 926–939.

Jiang, C., Liu, S., Lin, Z., Zhao, G., Duan, R., Liang, K. (2016). Domain-Aware Trust Network Extraction for Trust Propagation in Large-Scale Heterogeneous Trust Networks. *Knowledge-Based Systems*, *111,* 237–247.

Jones, K. S. (1972). A statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation*, *28*(1), 11–21.

Kabasakal, O., Mutlu, A. (2021). On the Effect of Word Positions in Graph-based Keyword Extraction. *Journal of Naval Sciences and Engineering*, *17*(2), 217–239.

Kang, H., Sun, X., Feng, T., Lin, S. (2021). Keyword Extraction Based on Semantic Similarity Metric and Multi-Feature Computing. *7th International Conference on Big Data Computing ve Communications (BigCom)*, Online, China, 13–15 Ağus. 2021.

Khan, M. Q., Shahid, A., Uddin, M. I., Roman, M., Alharbi, A., Alosaimi, W., Almalki, J., Alshahrani, S. M. (2022). Impact Analysis of Keyword Extraction Using Contextual Word Embedding. *PeerJ Computer Science*, *8,* e967.

Kim, S. N., Medelyan, O., Kan, M., Baldwin, T. (2010). SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles. *5th International Workshop on Semantic Evaluation*, Uppsala, Sweden, 15–16 Tem. 2010.

Knittel, J., Koch, S., Ertl, T. (2021). ELSKE: Efficient Large-Scale Keyphrase Extraction. *21st ACM Symposium on Document Engineering*, Limerick, Ireland, 24–27 Ağus. 2021.

Krapivin, M., Autaeu, A., Marchese, M. (2009). Large Dataset for Keyphrases Extraction, *University of Trento*, DISI-09-055, 1–6.

Kwary, D. A., Artha, A. F., Amalia, Y. S. (2018). Lexical Word-Class Distributions in Research Articles of Four Subject Areas. *Studies about Languages*, *33,* 108–118.

Lahiri, S., Choudhury, S. R., Caragea, C. (2014). Keyword and Keyphrase Extraction Using Centrality Measures on Collocation Networks. *arXiv preprint arXiv:1401.6571*, 1–11.

Lee, M.-C., Gao, B., Zhang, R. (2018). Rare Query Expansion Through Generative Adversarial Networks in Search Advertising. *24th ACM SIGKDD International Conference on Knowledge Discovery ve Data Mining*, London, United Kingdom, 19–23 Ağus. 2018.

Li, T., Hu, L., Li, H., Sun, C., Li, S., Chi, L. (2021). TripleRank: An Unsupervised Keyphrase Extraction Algorithm. *Knowledge-Based Systems*, *219,* 106846.

Li, Y., Ning, H. (2021). Multi-Feature Keyword Extraction Method Based on TF-IDF and Chinese Grammar Analysis. *2021 International Conference on Machine Learning ve Intelligent Systems Engineering (MLISE)*, Chongqing, China, 9–11 Tem. 2021.

Lin, Z.-L., Wang, C.-J. (2019). Keyword Extraction with Character-Level Convolutional Neural Tensor Networks. *Pacific-Asia Conference on Knowledge Discovery ve Data Mining*, Macau, China, 14–17 Nis. 2019.

Liu, H., Wang, L., Zhao, P., Wu, X. (2019). Document Specific Supervised Keyphrase Extraction with Strong Semantic Relations. *IEEE Access*, *7,* 167507–167520.

Liu, Z., Li, P., Zheng, Y., Sun, M. (2009). Clustering to Find Exemplar Terms for Keyphrase Extraction. *2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, 6–7 Ağus. 2009.

Luo, L., Zhang, L., Peng, H. (2020). An Unsupervised Keyphrase Extraction Model by Incorporating Structural and Semantic Information. *Progress in Artificial Intelligence*, *9*(1), 77–83.

Mahata, D., Kuriakose, J., Shah, R., Zimmermann, R. (2018). Key2vec: Automatic Ranked Keyphrase Extraction from Scientific Articles Using Phrase Embeddings. *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Louisiana, USA, 1–6 Haz. 2018.

Mao, X., Huang, S., Li, R., Shen, L. (2020). Automatic keywords extraction based on co-occurrence and semantic relationships between words. *IEEE Access*, *8,* 117528–117538.

Medelyan, O., Frank, E., Witten, I. H. (2009). Human-Competitive Tagging Using Automatic Keyphrase Extraction. *47th Annual Meeting of the Association for Computational Linguistics*, Singapore, 2–7 Ağus. 2009.

Merrouni, Z. A., Frikh, B., Ouhbi, B. (2022). HAKE: An Unsupervised Approach to Automatic Keyphrase Extraction for Multiple Domains. *Cognitive Computation*, *14*(2), 852–874.

Mihalcea, R., Tarau, P. (2004). TextRank: Bringing Order into Text. *2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, 25–26 Tem. 2004.

Nasar, Z., Jaffry, S. W., Malik, M. K. (2019). Textual Keyword Extraction and Summarization: State-of-the-Art. *Information Processing and Management*, *56*(6), 102088.

Nguyen, T. D., Kan, M. Y. (2007). Keyphrase Extraction in Scientific Publications. *10th International Conference on Asian Digital Libraries*, Hanoi, Vietnam, 10–13 Ara. 2007.

Nguyen, T. D., Luong, M.-T. (2010). WINGNUS: Keyphrase Extraction Utilizing Document Logical Structure. *5th International Workshop on Semantic Evaluation*, Uppsala, Sweden, 15–16 Tem. 2010.

Palshikar, G. K. (2007). Keyword Extraction from a Single Document Using Centrality Measures. *Pattern Recognition ve Machine Intelligence*, Kolkata, India, 18–22 Ara. 2007.

Papagiannopoulou, E., Tsoumakas, G. (2018a). Unsupervised Keyphrase Extraction Based on Outlier Detection. *arXiv preprint arXiv: 1808.03712*, 1–6.

Papagiannopoulou, E., Tsoumakas, G. (2018b). Local Word Vectors Guiding Keyphrase Extraction. *Information Processing & Management*, *54*(6), 888–902.

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

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *arXiv preprint arXiv:1802.05365*, *12,* 1–15.

Rabby, G., Azad, S., Mahmud, M., Zamli, K. Z., Rahman, M. M. (2020). Teket: A Tree-based Unsupervised Keyphrase Extraction Technique. *Cognitive Computation*, *12*(4), 811–833.

Ren, W., Lian, X., Ghazinour, K. (2019). Skyline Queries over Incomplete Data Streams. *The VLDB Journal*, *28*(6), 961–985.

Rose, S., Engel, D., Cramer, N., Cowley, W. (2010). Automatic Keyword Extraction from Individual Documents. *Text mining: applications and theory*, *1,* 1–20.

Sharma, S., Gupta, V., Juneja, M. (2021). Diverse Feature Set Based Keyphrase Extraction and Indexing Techniques. *Multimedia Tools and Applications*, *80*(3), 4111–4142.

Shubankar, K., Singh, A., Pudi, V. (2011). A Frequent Keyword-Set Based Algorithm for Topic Modeling and Clustering of Research Papers. *3rd Conference on Data Mining ve Optimization (DMO)*, Bangi, Malaysia, 28–29 Haz. 2011.

Simsek, A., Karagoz, P. (2020). Wikipedia Enriched Advertisement Recommendation for Microblogs by Using Sentiment Enhanced User Profiles. *Journal of Intelligent Information Systems*, *54*(2), 245–269.

Song, K., Tan, X., Qin, T., Lu, J., Liu, T.-Y. (2020). Mpnet: Masked and Permuted Pre-training for Language Understanding. *Advances in Neural Information Processing Systems*, *33,* 16857–16867.

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

Sung, Y. yeon, Kim, S. B. (2020). Topical Keyphrase Extraction with Hierarchical Semantic Networks. *Decision Support Systems*, *128,* 113163.

Tahir, N., Asif, M., Ahmad, S., Malik, M. S. A., Aljuaid, H., Butt, M. A., Rehman, M. (2021). FNG-IE: An Improved Graph-based Method for Keyword Extraction from Scholarly Big-data. *PeerJ Computer Science*, *7,* e389.

Tao, Y., Papadias, D. (2006). Maintaining Sliding Window Skylines on Data Streams. *IEEE Transactions on Knowledge and Data Engineering*, *18*(3), 377–391.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I. (2017). Attention is All you Need*. 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 4–9 Ara. 2017.

Vega-Oliveros, D. A., Gomes, P. S., Milios, E. E., Berton, L. (2019). A Multi-centrality Index for Graph-based Keyword Extraction. *Information Processing & Management*, *56*(6), 102063.

Veisi, H., Aflaki, N., Parsafard, P. (2020). Variance-based Features for Keyword Extraction in Persian and English Text Documents. *Scientia Iranica*, *27*(3), 1301–1315.

Wallach, H. M. (2006). Topic Modeling: Beyond Bag-of-words*. 23rd international conference on Machine learning*, Pittsburgh, USA, 25–29 Haz. 2006.

Wan, X., Xiao, J. (2008). Single Document Keyphrase Extraction Using Neighborhood Knowledge. İçinde: (13–17 Tem. 2008). 23th AAAI Conference on Artificial Intelligence. Chicago, Illinois, USA, ss. 855–860.

Wang, B., Kuo, C.-C. J. (2020). SBERT-WK: A Sentence Embedding Method by Dissecting BERT-based Word Models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28,* 2146–2157.

Wang, H., Li, J. (2022). Unsupervised Keyphrase Extraction from Single Document Based on Bert*. 2022 International Seminar on Computer Science ve Engineering Technology (SCSET)*, Indianapolis, IN, USA, 8–9 Ocak 2022.

Wang, J., Peng, H. (2005). Keyphrases Extraction from Web Document by the Least Squares Support Vector Machine. *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, Compiegne, France, 19–22 Eyl. 2005.

Wang, L., Li, S. (2017). PKU_ICL at SemEval-2017 Task 10: Keyphrase Extraction with Model Ensemble and External Knowledge. *11th International Workshop on Semantic Evaluation*, Vancouver, Canada, 3–4 Ağus. 2017.

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., Nevill-Manning, C. G. (1999). KEA: Practical Automatic Keyphrase Extraction. Berkeley, CA, USA, 11–14 Ağus. 1999.

Wu, C., Liao, L., Afedzie Kwofie, F., Zou, F., Wang, Y., Zhang, M. (2021). TextRank Keyword Extraction Method Based on Multi-feature Fusion. *Sixth International Congress on Information ve Communication Technology*, London, United Kingdom, 25 Şub. 2021–26 Şub. 2014.

Yang, M., Liang, Y., Zhao, W., Xu, W., Zhu, J., Qu, Q. (2018). Task-Oriented Keyphrase Extraction from Social Media. *Multimedia Tools and Applications*, *77*(3), 3171–3187.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., Le, Q. V. (2019). Xlnet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems*, *32,* 1–11.

Yeom, H., Ko, Y., 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.

Zeng, Y., Chen, G., Li, K., Zhou, Y., Zhou, X., Li, K. (2019). M-skyline: Taking Sunk Cost and Alternative Recommendation in Consideration for Skyline Query on Uncertain Data. *Knowledge-Based Systems*, *163,* 204–213.

Zhang, F., Peng, B. ve diğ. (2013). WordTopic-MultiRank: A New Method for Automatic Keyphrase Extraction. *Sixth International Joint Conference on Natural Language Processing*, Nagoya, Japan, 14–18 Ekim 2013.

Zhang, Y., Tuo, M., Yin, Q., Qi, L., Wang, X., Liu, T. (2020). Keywords Extraction with Deep Neural Network Model. *Neurocomputing*, *383,* 113–121.

Zhao, J., Guan, Z., Sun, H. (2019). Riker: Mining Rich Keyword Representations for Interpretable Product Question Answering. *25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, AK, USA, 4–8 Ağus. 2019.

Zhou, N., Shi, W., Liang, R., Zhong, N. (2022). TextRank Keyword Extraction Algorithm Using Word Vector Clustering Based on Rough Data-Deduction. *Computational Intelligence and Neuroscience*, *2022,* 1–19.

Zhou, T., Zhang, Y., Zhu, H. (2020). Multi-Level Memory Network with CRFs for Keyphrase Extraction. *Pacific-Asia Conference on Knowledge Discovery ve Data Mining*, Singapore, 11–14 May. 2020.

Zhu, X., Lyu, C., Ji, D., Liao, H., Li, F. (2020). Deep Neural Model with Self-Training for Scientific Keyphrase Extraction. *Plos one*, *15*(5), e0232547.

Zhuohao, W., Dong, W., Qing, L. (2021). Keyword Extraction from Scientific Research Projects Based on SRP-TF-IDF. *Chinese Journal of Electronics*, *30*(4), 652–657.

KİŞİSEL YAYIN VE ESERLER

Erdemir, M., **Göz**, **F.**, Mutlu, A., Karagoz, P. (2019). Comparison of Querying Performance of Neo4j on Graph and Hyper-graph Data Model. *11th International Joint Conference on Knowledge Discovery, Knowledge Engineering Knowledge Management*, Vienna, Austria, 17–19 Eyl. 2019.

Goz, **F.**, Mutlu, A. (2022). MGRank: A Keyword Extraction System Based on Multigraph GoW Model and Novel Edge Weighting Procedure. *Knowledge-Based Systems*, *251,* 109292.

Goz, **F.**, Şener, F., Mutlu, A., Küçük, K., Temur, M. (2021). Learning to Rank for Text Summarization: Revisiting the Features and Methods for Turkish Bank Documents. *2021 International Conference on Innovations in Intelligent Systems ve Applications (INISTA)*, Kocaeli, Turkey, 25–27 Ağus. 2021.

Goz, **F.**, Mutlu, A. (2021). Automatic Keyword Extraction From Text Documents. Gyamfi, A., Williams, I., (Ed.), *Digital Technology Advancements in Knowledge Management* (71–91). IGI Global.

Goz, **F.**, Kabasakal, O., Mutlu, A. (2020). Experimental Analysis of Keyword-based Social Network Similarity Approach for Document Classification. *2020 28th Signal Processing ve Communications Applications Conference (SIU)*, İstanbul, Turkey, 5–7 Ekim 2020.

Goz, **F.**, Mutlu, A. (2018). Learning Logical Definitions of n-ary Relations in Graph Databases. *International Conference on Hybrid Artificial Intelligence Systems*, Oviedo, Spain, 20–22 Haz. 2018.

Goz, **F.**, Mutlu, A., Akbulut, O. (2018). Analysis of Ramer-Douglas-Peucker Algorithm as a Discretization Method. *2018 26th Signal Processing Communications Applications Conference (SIU)*, İzmir, Turkey, 2–5 May. 2018.

Goz, **F.**, Mutlu, A. (2017). Concept Discovery in Graph Databases. *International Conference on Hybrid Artificial Intelligence Systems*, La Rioja, Spain, 21–23 Haz. 2017.

Goz, **F.**, Ozcan, M. (2017). An Entertaining Mobile Vocabulary Learning Application. *EPESS*, *7,* 63–66.

Göz, **F.**, Mutlu, A., Küçük, K., Temur, M., Gün, A. (2021). Effect of Centrality Measures for Keyword Extraction from Turkish Documents. *2021 29th Signal Processing Communications Applications Conference (SIU)*, İstanbul, Turkey, 9–11 Haz. 2021.

Kaya Gülağız, F., **Goz**, **F.**, Şahin, E., Albayrak, M. S., Kavak, A. (2016). Beacon Temelli Sanal Etiket Uygulaması. *Bilecik Şeyh Edebali Üniversitesi Fen Bilimleri Dergisi*, *3*(1), 1–7.

Kaya Gulagiz, F., **Goz**, **F.**, Kavak, A. (2016). A Study on Environmental Effect of Electromagnetic Waves. *International Journal of Computer Applications*, *140*(7), 35–41.

Mutlu, A., Goz, F. (2022). SkySlide: a hybrid method for landslide susceptibility assessment based on landslide-occurring data only. *The Computer Journal*, *65*(3), 473–483.

Mutlu, A., Abdisamad, M. A., Kabasakal, O., **Göz**, **F.**, Tüfekçi, Ö., Küçük, K. (2021). Dijital Kütüphanelerde Dokümanlardan Bilgi Geri Kazanımı için Kullanılan Güncel Teknolojiler: Derleme Çalışması. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, *9*(1), 79–91.

Mutlu, A., Goz, F., Koksal, K., Erener, A. (2019a). Landslide susceptibility assessment using skyline operator and majority voting. *Sakarya Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *23*(5), 782–787.

Mutlu, A., **Goz**, **F.**, Zeren, H. (2019b). Guiding Search in Relational Pathfinding-based Concept Discovery via Bivariate Statistical Methods. *International Journal of Intelligent Systems and Applications in Engineering*, *7*(3), 118–123.

Mutlu, A., **Göz**, **F.**, Akbulut, O. (2019c). Ifit: An Unsupervised Discretization Method Based on The Ramer-Douglas-Peucker Algorithm. *Turkish Journal of Electrical Engineering and Computer Sciences*, *27*(3), 2344–2360.

Ozcan, M., **Goz**, **F.** (2017). An Educational Mobile City Learning Application for Kids. *The Eurasia Proceedings of Educational and Social Sciences*, *7,* 24–29.

ÖZGEÇMİŞ

Furkan Göz ilk, orta ve lise öğrenimini İstanbul'da tamamlamıştır. Anadolu Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği bölümünden mezun olmuştur. Yüksek lisans eğitimini Kocaeli Üniversitesi Bilgisayar Mühendisliği bölümünde tamamlamıştır. Kocaeli Üniversitesi Bilgisayar Mühendisliği bölümünde Araştırma Görevlisi olarak görev yapmaktadır.