KOCAELİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ
ANABİLİM DALI

YÜKSEK LİSANS TEZİ

ANAHTAR KELİME ÇIKARIMI İÇİN KELİME VEKTÖRLERİ: KARŞILAŞTIRMALI BİR DEĞERLENDİRME

IRMA DIBRA

KOCAELİ 2022

KOCAELİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ

ANABİLİM DALI

YÜKSEK LİSANS TEZİ

ANAHTAR KELİME ÇIKARIMI İÇİN KELİME VEKTÖRLERİ: KARŞILAŞTIRMALI BİR DEĞERLENDİRME

IRMA DIBRA

Dr. Ögr. Üyesi ALEV MUTLU		
Danışman,	**Kocaeli Üniv.**	**.....**
Doç. Dr. ORHAN AKBULUT		
Jüri Üyesi,	**Kocaeli Üniv.**	**.....**
Dr. Öğr. Üyesi BURCU YILMAZ		
Jüri Üyesi,	**Gebze Teknik Üniv.**	**.....**

Tezin Savunulduğu Tarih: 13.06.2022

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 aşamalarında 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 çalışmanın Kocaeli Üniversitesi'nin abone olduğu intihal yazılım programı kullanılarak Fen Bilimleri Enstitüsü'nün belirlemiş olduğu 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.

(İmza)
Irma DIBRA

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.

(İmza)
Irma DIBRA

ÖNSÖZ VE TEŞEKKÜR

Bu tez ve araştırma için, her zaman dürüst görüşleri ile bu yolda yardımcı olan ve beni doğru yola yönlendiren Kocaeli Üniversitesi Bilgisayar Mühendisliğinde danışmanım Dr. Öğr. Üyesi Alev MUTLU ve Arş. Gör. Furkan GÖZ'e büyük destekleri için sonsuz teşekkürlerimi sunarım. Ayrıca okul yıllarımda ve sonrasında beni her zaman destekleyen ve her zaman yanımda olan, bana olan güvenini ve yeteneğimi hiçbir zaman yitirmeyen aileme ve Ahmet İLGİN'e teşekkürü bir borç bilirim.

Ocak – 2022 Irma DIBRA

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

SİMGELER VE KISALTMALAR DİZİNİ

Kısaltmalar

AI	: Artificial Intelligence (Yapay Zeka)
AKE	: Automatic Keyword Extraction (Otomatik Anahtar Kelime Çıkarımı)
BERT	: Bidirectional Encoder Representations from Transformer (Transformatörlerden Çift Yönlü Kodlayıcı Beyanı)
CBOW	: Continuous Bag Of Words (Sürekli Kelime Torbası)
GloVe	: Global Vectors (Global Vektörler)
KEA	: Keyphrases Extraction Algorithm (Anahtar Kelime Çıkarma Algoritması)
LSA	: Latent Semantic Analysis (Gizli Semantik Analiz)
MAP	: Mean Average Precision (Orantılı Ortalama Hassasiyeti)
ML	: Machine Learning (Makine Öğrenimi)
MRR	: Mean Reciprocal Rank (Ortalama Karşılıklı Sıralama)
NLM	: Neural Language Models (Sinirsel Dil Modelleri)
NLP	: Natural Language Processing (Doğal Dil İşleme)
PV-DBOW	: Paragraph Vector Distributed Bag of Words (Paragraf Vektör Dağıtık Kelime torbası)
PV-DM	: Paragraph Vector Distributed Memory (Paragraf Vektör Dağıtık Bellek)
RVA	: Reference Vector Algorithm (Referans Vektör Algoritması)
VSM	: Vector Space Model (Vektör Uzay Modeli)

ANAHTAR KELİME ÇIKARIMI İÇİN KELİME VEKTÖRLERİ: KARŞILAŞTIRMALI BİR DEĞERLENDİRME

ÖZET

Anahtar kelimeler, bir belgenin içeriğini en iyi tanımlayan söz veya söz öbekleridir. Çevrimiçi dokümanların sayısı yüksek hızda artması ve anahtar kelimelerin metin sınıflandırma ve kümeleme gibi çeşitli metin işleme problemlerinde uygulanması nedeniyle daha da zorlu ve ilgi çekici bir görev haline geldi. Son zamanlarda, kelimelerin sayısal temsilleri (vektör), sözcüklerin anlamsal benzerliğini tahmin etmek için faydalı araçlar sağladığından, otomatik anahtar kelime çıkarma alanında uygulama bulmuştur. Bu çalışmada, farklı kelime vektör modellerinin tam metinden anahtar kelime çıkarmadaki başarımları irdelenmiştir. Bu amaçla Referans Vektör Algoritması (RVA) farklı kelime vektör modelleri ile çalışacak şekilde değiştirilmiş ve FAO30, Krapivin, Nguyen, Schutz, ve SemEval2010 veri seti üzerinde deneyler yapılmıştır. Elde edilen sonuçlar, GloVe ve Scibert'in her biri iki veri kümesi için en yüksek F1 @ 10 puanına ve bir veri kümesi için Word2Vec'ye ulaştığını göstermektedir. Ayrıca GloVe, üç veri kümesi için en yüksek ortalama karşılıklı sıralamaya (MRR) ve orantılı ortalama hassasiyete (MAP) ulaştı.

Anahtar Kelimeler: Kelime Temsilleri/Vektörleri, Otomatik Anahtar Kelime Çıkarma, RVA.

WORD EMBDEDDINGS FOR AUTOMATIC KEYWORD EXTRACTION: A COMPARATIVE ASSESSMENT

ABSTRACT

Keywords are salient words / phrases that best describe the content of a text document. Automatic keyword extraction from text documents has become a challenging task as the amount of online text documents increases in high speed and keywords have applications in several text processing problems such as text classification and clustering. Recently, word embeddings have found application in automatic keyword extraction as these representations provide means to estimate semantic similarity of words. In this study, the performances of different word vector models in extracting keywords from full text are examined. To this aim, the Reference Vector Algorithm (RVA) was modified to work with different word vector models and experiments were conducted on the FAO30, Krapivin, Nguyen, Schutz, and SemEval2010 datasets. The experimental results show that GloVe and SciBERT achieved the highest F1@10 score for two datasets each and Word2Vec for one dataset. Moreover, GloVe achieved the highest mean reciprocal rank (MRR) and mean average precision (MAP) for three datasets.

Keywords: Word Embeddings, Automatic Keyword Extraction, RVA.

1. GİRİŞ

Anahtar kelimeler bir metnin içeriğini en iyi şekilde ifade eden söz veya söz öbekleridir. Anahtar kelimeler kümeleme, sınıflandırma ve özetleme gibi çeşitli metin işleme problemlerinde ve arama işlemlerinde sıklıkla kullanılmaktadır (Nasar ve diğ., 2019). Düz yazılar, bilimsel makaleler ve sosyal medya gönderileri gibi metin dokümanlarının miktarı günümüzde katlanarak artmaktadır. Çokluğu nedeniyle bu metinlere elle anahtar kelime atanması hem imkansız hem de öznel bir işlem haline gelmiştir. Bu nedenle metinlerden otomatik anahtar kelime çıkarma problemi önem kazanmıştır.

Metinden otomatik anahtar kelime çıkarma yöntemleri farklı bakış açılarına göre farklı şekillerde sınıflandırılabilmektedir. İlk sınıflandırma anahtar kelimelerin metin içinde geçip geçmemesine göre çıkarımsal ve atayıcı olarak sınıflandırılabilir. Etiketli veri ihtiyacına göre metinden otomatik anahtar çıkarma yöntemleri gözetimli, gözetimsiz ve melez olarak sınıflandırılabilir. Kullanılan yönteme göre ise metinden otomatik anahtar kelime çıkarma yöntemleri istatistiksel, dilbilimsel, makine öğrenmesine dayalı ve melez olarak sınıflandırılabilir. Odaklanılan doküman sayısına göre metinden otomatik anahtar kelime çıkarma yöntemleri tek doküman ve çoklu doküman yöntemi şeklinde sınıflandırılabilir. Diğer yandan, anahtar kelimelerin çıkarıldığı metin parçasına göre otomatik anahtar kelime yöntemleri özetten anahtar kelime çıkarma ve tam metinden anahtar kelime çıkarma yöntemi olarak sınıflandırılabilir (Göz ve diğ., 2021).

Temel olarak otomatik anahtar kelime çıkarma yöntemleri üç aşamadan oluşmaktadır: metin önişleme, anahtar kelime çıkarma ve ardıl işleme. İlk aşamada metin bilgisayarın okuyabileceği formata çevrilir ve kullanılacak anahtar kelime çıkarma yöntemine göre kelimeler için özellikler çıkarılır. İkinci aşamada çıkarılan özelliklerden de faydalanılarak metindeki anahtar kelimeler bulunur. Son aşamada ise çıkarılan anahtar kelimeler kullanıcıların anlayabileceği formata çevrilir. Otomatik anahtar kelime çıkarma yöntemlerinin çoğu metin ön ve ardıl işlemede benzer işlemler yaparken özellik çıkarma aşamasında farklılık göstermektedir. Literatürdeki pek çok çalışma terim frekansı, terim frekansı – ters doküman frekansı, birlikte geçme sıklığı gibi istatistiksel özellikler, türü gibi dil bilimsel özellikler ve kelime pozisyonu gibi uzamsal özellikler kullanmaktadır. Yakın zamanda kelimelerin sayısal temsili olarak kullanılan kelime vektörleri de otomatik anahtar kelime çıkarma sistemlerinde kullanılmaya başlanmıştır.

Kelime vektörleri kelimelerin makine öğrenmesi yöntemlerinde kullanılabilmesi için geliştirilen sayısal gösterimlerdir. Bu temsiller kelimelerin anlamsal yakınlıkları hakkında bilgi vermedikleri için pek çok problemde kullanılmış ve farklı yaklaşımlara dayalı kelime vektörleri oluşturulmuştur (Khattak ve diğ., 2019). Kelime vektörleri otomatik anahtar kelime çıkarma probleminde de sıklıkla kullanılmış ve başarılı sonuçlar elde edilmiştir.

Bu tezin amacı farklı kelime vektörlerinin anahtar kelime çıkarımındaki başarısını karşılaştırmalı olarak değerlendirmektir. Bu amaçla, gözetimsiz, kelime vektörlerine dayalı ve tam metinden anahtar kelime çıkarma sistemi olan Reference Vector Algorithm (RVA) (Papagiannopoulou ve diğ., 2018) farklı kelime vektörleri ile çalışacak şekilde yeniden kodlanmıştır. Otomatik anahtar kelime çıkarmada sıklıkla kullanılan Word2Vec, Doc2Vec, GloVe, fastText, BERT ve SciBERT kelime vektör temsillerinin anahtar kelime çıkarımındaki başarısı yeniden kodlanan RVA algoritması ile beş tane tam metin veri seti üzerinde karşılaştırılmıştır. Elde edilen sonuçlar hem karmaşıklık matrisine dayalı olan F1 değeri hem de sırlamaya dayalı olan Mean Average Precision (MAP) ve Mean Recoprical Rank (MRR) ölçütleri kullanılarak değerlendirilmiştir. Ayrıca elde edilen sonuçlar istatistiksel olarak değerlendirilmiştir.

Yapılan testler sonucunda, GloVe ve SciBERT ikişer Word2Vec ise bir veri kümesinde en yüksek F1 değerini elde etmiştir. GloVe beş veri kümesinden dördü için en yüksek MAP değerini elde ederken BERT bir veri kümesi için en yüksek MAP değerini elde etmiştir. Benzer sonuçlar MRR ölçütü için de elde edilmiştir. GloVe üç veri seti için en yüksek MRR değerini elde etmiş, Word2Vec ve BERT ise birer veri kümesi için en yüksek MRR değerini elde etmiştir. İstatistiksel analiz kelime temsillerinin birbirine göre üstünlükleri konusunda kesin yargılar oluşmasına olanak sağlamamıştır.

2. GENEL BILGILER

2.1. Tez Çalışmasının Amacı ve Başlatılma Sebebi

Her gün gelişen teknoloji ve büyük veri patlaması ile veri biliminin önemi sürekli artmaktadır. Bu, büyük verileri anlamlı ve yönetilebilir hale getirmek için süreçler ve sistemlerle ilgilenen disiplinler arası bir alandır. Farklı konulardaki birçok sayfalık belge, makale ve haberlerin anlaşılmasında zamanı azaltmak ve verimliliği artırmak için birçok teknik geliştirilmiş ve çalışma yapılmıştır.

Bu çalışmada, bilimsel makalelerde ve kongrelerde yayınlanan anahtar kelime çıkarma yöntemlerindeki farklı kelime temsilerinin performansları karşılaştırılmıştır. Bu kapsamda makaleler sayısal tabanlı verilere dönüştürülmüştür. Dönüşüm hem tahminlere dayalı hem de sayıma dayalı farklı temsiller kullanılarak yapılmıştır. Makalelerin özet kısmında ve başlığında yer alan kelimeler/cümleler anahtar kelime çıkarma yöntemi ile değerlendirilmiş ve sonuçlar karşılaştırılmıştır. Bu çalışmada verilen sonuçlara göre, kelime temsillerinin önemi ve farklı veri kümelerindeki performansı hakkında bir çıkarım yapılması amaçlanmıştır.

2.2. Tez Çalışmasının Katkıları

Çalışma, kelime temsillerinin önemi ve bu temsillerin anahtar kelime çıkarımına nasıl etki ettiği hakkında bir çıkarım yapmayı amaçlamaktadır. Ayrıca, her bir temsilin farklı veri kümelerindeki performansı gösterilmiştir.

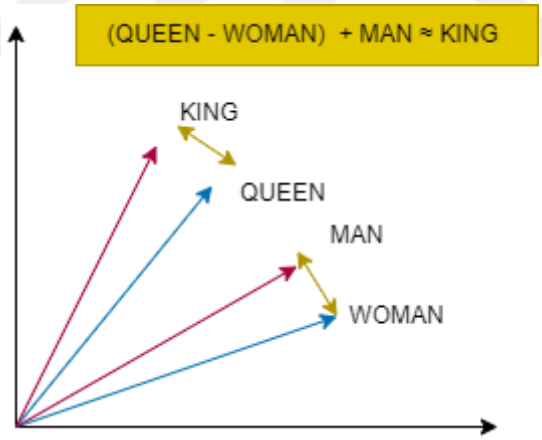
2.3. Literatür Taraması

Bu bölümde, bilimsel makalelerden otomatik anahtar kelime çıkarmada kelime temsillerinin kullanımı ve gelişimi üzerine literatür çalışması sunulmaktadır.

NLP görevlerinin en temel problemi, kelimelerin analitik anlamlarının çıkarılmasıdır. Bunun için kelimelerdeki benzerliğin, kompozisyonun vb. gibi özelliklerin net bir şekilde temsil edilmesi gerekmektedir. İhtiyacı ve nedenini anlamak önemlidir, kelime temsilleri otomatik anahtar kelime çıkarmada ve birçok NLP problemlerinde önemli rol almıştır. Genel olarak, kelime temsilleri kullanılmadan önce, metinden anlamlı veriler oluşturmak için iki farklı model uygulanılmıştır.

Bilgi erişimi alanında, Vektör Uzay Modeli önerilen ilk model olmuştur (Salton, 1975). Vektör Uzay Modeli (VSM), sözcükleri ve belgeleri vektörler olarak temsil eden ilk ve en etkili modellerden biridir. Bu vektörler, kelimeler arasındaki bağımlılıkları yakalamaya izin verir ve bilgi çıkarma, sohbet robotları programlama ve daha fazlası gibi birçok uygulamada kullanılır. Vektör uzayı modellerinin kullanımında amaç metindeki her kelimenin göreceli anlamını yakalamaktır ve bu her kelimenin bağlamı belirlenerek yapılır. Genel olarak, bu model iki vektörü geometrik bir perspektifte karşılaştırmaya ve aralarındaki benzerliği anlamaya izin verir. VSM, kelime-kelime ve kelime-belge olmak üzere iki farklı mimariden oluşur (Almeida, 2019).

Kelime-kelime tasarım, sabit bir k mesafesi için iki kelimenin birlikte meydana gelme sayısını hesaplar. Diğer yandan, belgeye göre kelime tasarımı, belirli kategorilere ait belgelerde belirli bir kelime dağarcığından kelimelerin kaç kez göründüğünü sayar. Örnek olarak "Queen" ve "Woman" arasındaki benzerliği biliyorsak "King ve Man" arasındaki ilişkiyi belirlemek için vektör gösterimini kullanabiliriz (Mikalov, 2013). Şekil 2.1'de modelin görsel bir temsili gösterilmektedir.



Şekil 2.1. Uzayda vektörlerin benzerliği

En erken aşamalarda sunulan bir diğer model ise İstatistiksel Dil Modelidir (SLM) (Bahl, 1983). Bu model, önceki kelimeleri de dikkate alarak bir sonraki kelimeyi tahmin eder. SLM olasılıksal bir modeldir. Model, belgeye dayalı olarak kelimenin oluşma olasılığını öğrenir. Bu çalışmada, ses sinyallerinden gelen doğal konuşma otomatik olarak yorumlanmış ve gürültüler ile yanlış alarmlar filtrelenmiştir. Diğer çalışmalar, bu yöntemin kullanımının sadece maksimum 3-gram kelime için iyi sonuçlar verdiğini göstermiştir (Goodman, 2001). Bu nedenle, yöntemler, eğitim amaçlı kullanılan

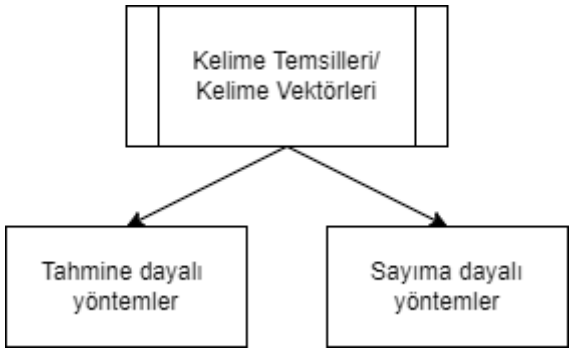
belgelerde bulunmayan büyük kelime dağarcığı ve kelimelerle uğraşırken zorluklarla karşılaşmıştır (Bengio, 2003). Genellikle bu modeller tek başına bulunmaz, model, dil modellerine ihtiyaç duyan farklı algoritmaların önüne veya sonuna eklenir.

N-gramlar en çok kullanılan istatistiksel modellerden biridir. Bu modellerin en basit yaklaşım oldukları söylenebilir. Burada n, olasılık dağılımında dikkate alınacak kelime sayısıdır. N boyutuna göre modeller isimlendirilir, unigrams, bigrams, trigrams vb. gibi.

Üstel model, n-gram ve özellik işlevinin bir kombinasyonunu değerlendirmek için kullanılır. Modelin prensibi entropide yatmaktadır. Bu modeller daha az istatistiksel varsayıma sahip oldukları için daha doğrudur.

Sürekli uzay modeli, sinir ağı olarak da bilinir. Ağırlıkların doğrusal olmayan bir kombinasyonudur. Kelimeye ağırlık verilmesine kelime temsili denir (Bengio, 2003). Sinir ağları boyunca log-lineer modeller de kullanılmıştır (Mnih ve Hinton, 2007). Son yöntemler daha iyi bir etkiye ve saygın sonuçlara sahiptir.

Kelime temsilleri yaygın olarak (Baroni, 2014, Pennington, 2014, Li ,2015) iki türe ayrılır. Tahmine dayalı adlandırılmış yöntemler, yerel verileri kullanan yöntemlerdir; genel bilgileri kullanan yöntemler ise sayıma dayalı yöntemler olarak adlandırılır. Aşağıdaki, Şekil 2.2'de kelime temsil modelleri gösterilmektedir.



Şekil 2.2. Kelime temsil modelleri

Tahmine dayalı yöntemlerin başlangıcı, kelime temsil yöntemlerinde iz bırakan Nöral Dil Modelleri (NLM'ler) ile başlar (Bengio, 2003). Bu çalışmada, büyük korpus verilerindeki performans, karmaşıklık ve basitlik arasındaki değişimler analiz edilmiştir. Bu çalışma, önceden yapılan sinirsel dil modelleri ve diğer çözümlerin performans karşılaştırması sonuncularından yola çıkarak yapılmıştır. Bu çalışmanın hesaplama maliyeti çok

yüksektir. Bu sorun Bengio ve Senegal (2003) tarafından yapılan çalışmada aşılmıştır. Yardımcı dağıtım kullanarak sinir ağındaki gradyanları tahmin etmişlerdir. Bu adımlar takip edilerek eğitim süresini ve çıktısını iyileştirmek için birçok teknik geliştirilmiştir.

Collobert ve Weston (2008) çalışmalarında sadece temsilleri öğrenmeyi amaçlayan bir model önerilmiştir. Tahminlerde kelimelerin tam bağlamının kullanılması ve çalışmada en önemli iki gelişmenin belirtildiği temsiller üretmek için etiketlenmemiş verilerden yararlanılmıştır. Veri setini değiştirip daha fazla yanlış örnek eklenmiştir ve pozitif olanları söyleyebilecek bir model eğitilmiştir.

Ayrıca, Mikolov ve diğerlerinin 2009-2010 yıllarında yapmış oldukları katkılarından bahsetmek gerekir. 2009'da iki adımlı bir yöntem önerilmiştir. İlk adım, bağlam olarak tek bir kelime kullanarak modeli eğitmek ve ikinci adım, ilk adımın çıktısını kullanarak tam modeli eğitmektir. Daha sonra, 2010 yılında, dil modellerini eğitmek için Tekrarlayan Sinir Ağları (RNN) önerilmiştir (Mikolov ve diğ., 2010).

2013 yılında Mikolov ve diğerleri, tekrarlayan sinir ağı modelinin eğitimi ile elde edilen temsillerin araştırma ve analizinde (Mikalov ve diğ., 2010), kodlanmış vektörler arasında belki de sözdizimsel düzenliliklere sahip olabileceği tahmin edilmiştir. Sonunda, temsiller arasında sadece sözdizimsel değil, aynı zamanda anlamsal ilişkilerin de var olduğu görülmüştür.

Daha sonra 2013'te Mikalov, öğrenme yerleştirmeleri için CBOW ve SG adlı iki model sunmuştur. Bu modeller log-linear modellerdir ve eğitim amacıyla 2009 yılında sunulan 2 adımlı yöntemi kullanırlar.

Tahmin modeli için yapılan çalışmalarda iki makale daha öne çıkmaktadır. FastText modelinin eğitiminde sadece kelime temsilleri değil, aynı zamanda n-gramlar da dahil edilmiştir (Bojanowski ve diğ., 2016).

Sayıma Dayalı Modeller'de ise, kelimeleri temsil etmek için farklı bir mantık kullanılmıştır. Bu yöntemde, kelime bağlamı ve birlikte oluşumu dikkate alınmıştır. Turney ve Patrick Pantel (2010) kelime-bağlam matrislerini sunan ilk kişilerdir.

Daha sonra yeni bir yöntem önerilmiştir. Yöntem, kelime dağarcığındaki her kelimenin analizinden ve hedef kelime ile bağlamın birlikte oluşumunun hesaplanmasından oluşur. Yöntem sekiz pencere boyutu ile sınırlı olsa da başarılı sonuçlar gözlemlenmiştir. Metodun adı HAL (Hyperspace Analog to Language) olarak adlandırılmıştır (Lund, 1996). Çalışmada, normalizasyon ve kelimelerin birlikte bulunma durumlarına göre bir filtreleme yapılmadığı için orantısızlık sorunu ortaya çıkmıştır.

Orantısızlık sorununa bir çözüm olarak COALS yöntemi önerilmiştir (Rohde ve diğ., 2006). COALS yönteminde bazı normalleştirme yöntemleri önerilmiştir. Sadece kelimenin sayımı yerine koşullu birliktelik sayımı önerilmiştir.

Sayım temelli modellere katkı yapan arasında Lebret ve Collobert (2013) yaptığı çalışmada yer almaktadır. Kelime bağlamının Hellinger PCA17 dönüşümü ile değiştirilmesi önerilmiştir. Bu yöntem, (Collbert ve Weston, 2008) ve (Mnih ve Hinton, 2008) tarafından sunulan sayıma dayalı modellere göre çok daha iyi sonuç vermiştir.

Sonunda, sayıma dayalı en başarılı yöntemlerden biri Pennington'ın ünlü GloVe yöntemidir. Model, denetimsiz bir öğrenme algoritmasıdır. Model, kelime vektörleri üretmek için küresel istatistiklere dayanır. Bu yöntemlerin temeli, Gizli Anlamsal Analize (LSA) kadar uzanmaktadır. Gizli Anlamsal Analiz doğal dil işlemede ve dağıtımsal anabiliminde kullanılmaktadır. LSA, Metinlerin otomatik sınıflandırılmasında ve anahtar kelimeleri belirlenmesi gibi çalışmalarda kullanılabilir (Landauer ve dig., 1997).

2.4. Tezin Yapısı

Bu çalışma altı bölüme ayrılmıştır. Birinci bölümde giriş ve genel bilgiler verilmiştir. İkinci bölüm literatür taraması ile ilgili olup, çalışmanın amacı ve genel anahtar kelimeler anlatılmıştır. Üçüncü bölümde temel kavramlar ve kelime vektörü çıkarma yöntemleri yer almaktadır. Dördüncü bölümde denetimsiz yaklaşım yöntemi olan Referans Vektör Algoritması anlatılmıştır. Beşinci ve altıncı bölümde yapılan deneyler ve sonuçlar gösterilmiştir.

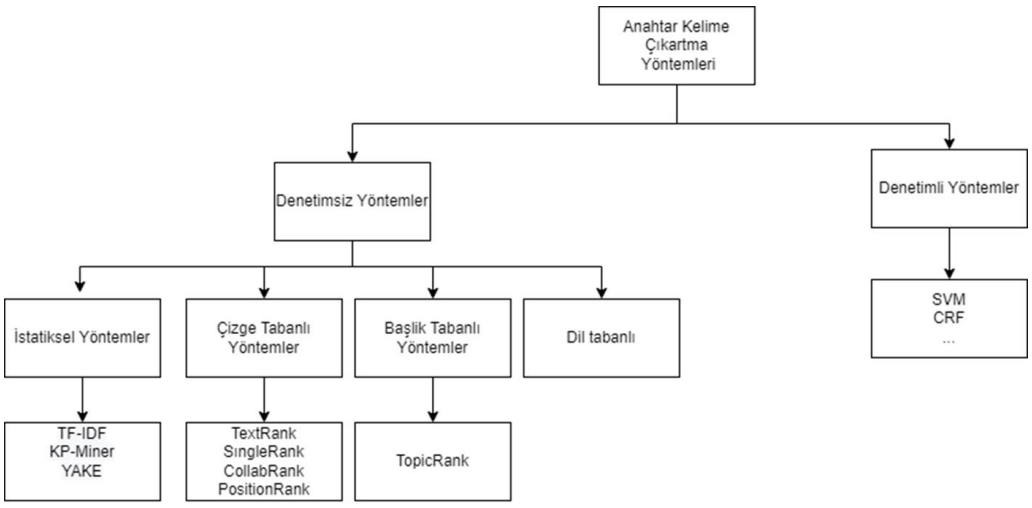
3. TEMEL KAVRAMALAR

3.1. Anahtar Kelime Çıkarma Yöntemleri

Anahtar kelime çıkarma (KE), belgenin ne hakkında olduğu ile ilgili fikir sağlanabilecek özet kelime veya kelime öbeklerinin çıkarılmasıdır. Otomatik Anahtar Kelime Çıkarımı (AKE), anahtar kelimelerin algoritmalar tarafından çıkarılması işlemine denir. Anahtar Kelime Çıkarma yöntemleri Soyut Eğilimli (Abstractive) yöntemler ve Çıkarım (Extractive) yöntemleri olmak üzere ikiye ayrılır (Bharti ve diğ., 2017). Çıkarım yöntemleri dokümanlarda geçen kelimeler üzerinden anahtar kelimelerin çıkarılmasını hedeflemektedir. Soyut eğilimli yöntemler ise dokümanda bulunmayan ama doküman hakkında fikir verebilecek kelimeleri de bulmayı hedeflemektedir. Bu bölümde çıkarım (Extractive) yöntemleri anlatılacaktır.

Otomatik anahtar kelime çıkarma yöntemleri denetimli (Supervised) ve denetimsiz (Unsupervised) olmak üzere ikiye ayrılır. Denetimli olan yöntemlerde makine öğrenmesi ve derin öğrenme yöntemleri ön plana çıkmaktadır. Denetimsiz anahtar kelime çıkarma yöntemleri dört farklı kategoride incelenmektedir. Bunlar istatistiksel tabanlı yöntemler, çizge tabanlı yöntemler, başlık tabanlı yöntemler ve dil tabanlı yöntemlerdir. Aşağıdaki

Şekil 3.1'de anahtar kelime çıkarma yöntemleri gösterilmiştir.



Şekil 3.1. Anahtar kelime çıkarma yöntemleri

Denetimli anahtar kelime çıkarma yöntemlerinde ana fikir aday anahtar kelimeleri anahtar kelime veya anahtar kelime değil şeklinde ikili sınıflandırma haline getirmektir.

Denetimli anahtar kelime çıkarma yöntemlerinde kelimenin yeri ve geçme sıklığı çok önemlidir (Beliga ve Slobodan, 2014).

Sınıflandırma yapılırken, bir aday anahtar kelimenin bir metni temsil etmeye uygun olup olmadığı çıkarımı yapılır. Bu işlemleri yapabilecek birçok algoritma bulunmaktadır; Naive Bayes, Decisin Trees, Support Vector Machines ve literatürde en çok yer alan sistemler GenEx (Turney, 1999) ve KEA'dır (Witten ve diğ., 1999).

Denetimli anahtar kelime çıkarma yöntemleri arasında Anahtar Çıkarım Algoritması (KEA) literatürde çalışılan en popüler yöntemlerden biridir. KEA (Witten ve diğ., 1999) tarafından önerilmiştir. Bu çalışma, kelimenin Terim Frekansı -Ters Metin Frekansı (TF-IDF) ve oluşum bilgisi ele alınarak kelime çıkarımı için Naïve-Bayes sınıflandırma algoritmasını kullanır. Çalışmada 1800 rapor kullanılmıştır ve her raporun önceden anahtar sözcükleri çıkarılmıştır. 500 rapor test veri seti olarak kullanılmıştır. Sunulan yaklaşım NZDL (New-Zealand Digital Library) tarafından CSTR'de (Computer Science Technical Reports) değerlendirilmiştir. Değerlendirmek için yazar tarafından atanan ve ayıklanan anahtar kelimeler arasında eşleşme sayısına bakılmıştır. KEA'nın en az 20 belgeden oluşan bir eğitim seti için iyi çalıştığı gözlemlenmiştir. Başlıklar ve özetler yerine belgenin tamamı kullanıldığında KEA en iyi sonucu vermiştir (Witten ve diğ., 2005).

KEA algoritmasının bir uzantısı olarak KEA++ tanıtılmıştır. KEA++ eş anlamlı kelimeleri içeren sözlük kullanmıştır (Medelyan ve Witten, 2006). KEA++'da Naïve-Bayes kullanılmıştır. Naïve Bayes basittir ve iyi sonuçlar verir (Medelyan ve Witten, 2006). Sonuçların değerlendirilmesinde Rolling ölçütü (Rolling, 1981) kullanılmıştır. KEA++ iyi bir performans sergilemiştir.

Denetimsiz anahtar kelime çıkarımı için İstatistik tabanlı, Çizge tabanlı, Dil tabanlı ve Başlık tabanlı olmak üzere dört farklı yaklaşım yapılmıştır. Bütün yaklaşımlar genel olarak kelimelerin filtrelenmesi, skor hesaplaması ve top-rank olan kelimelerin anahtar kelime olarak çıkarılması adımlarından oluşur.

Çizge tabanlı yöntemler arasında literatürde en yaygın ve öncü çalışmalardan biri TextRank algoritmasıdır (Rada, 2004). Bu modelde kelimeler grafın düğümleri olarak

kabul edilmektedir ve kenarlarla birbirilerine bağlanmaktadırlar. Kenarlar kelimelerin bir arada bulunma durumlarına göre bağlanırlar. Bir arada bulunma durumlarına bir pencere gezdirilerek bakılır eğer kelimeler pencere içerisinde ise bu iki kelime arasında bağlantı vardır. Pencere boyutu 2 ila 10 arasında seçilebilir. İki kelime arasında başka yabancı kelimelerin olması o kelimelerin arasındaki bağlantı derecesini düşürmektedir. En son düğümlerin aldığı puanlar büyükten küçüğe sıralanır ve en çok puan alan kelimelerin dokümanın anahtar kelimeleri olduğuna karar verilir. Bu çalışmada TextRank en yüksek keskinliği (Precision) ve f-skor (F-score) değerini elde etmiştir. Ancak duyarlılık (Recall) denetimli yöntemlere göre daha düşüktür.

Çizge tabanlı yöntemler arasında literatürde bulunan bir diğer yöntem PositionRank algoritmasıdır (Florescu ve Caragea, 2017). Üç aşamadan oluşur. Kelimeler filtrelemeden geçirilir ve aday kelimeler grafın düğümü olacak şekilde graf oluşturulur, kelimelerin birlikte bulunma durumuna göre ağırlıklandırma yapılır ve alınan puanlara göre en yüksek puanı alan aday anahtar kelimeler seçilir.

İstatistik tabanlı kelime çıkarımı yöntemleri genelde en basit yöntemlerdendir. Bu modelde dil tabanlı özellikleri kullanılmadan kelimenin pozisyonu veya geçme sıklığı gibi özelliklere bakılarak anahtar kelime olup olmadığı yönünde bir sonuca varılır. İstatistik tabanlı yöntemler arasında en öne çıkanlar TF-IDF ve KP-Minnerdir.

TF-IDF (Term Frequency-Inverse Document Frequency), kelimenin bulunduğu dokümanı ne kadar temsil ettiğini gösteren bir değerdir. TF (Terim Sıklığı) İlgili kelimenin dokümandaki frekansıdır. Kelimenin dokümanda geçme sayısını, dokümandaki toplam kelime sayısına bölerek elde edilir. DF (doküman sıklığı) TF ile benzemektedir ama bu kez diğer dokümanlara odaklanılır. İlgili kelimenin geçtiği doküman sayısının toplam doküman sayısına bölünmesi ile hesaplanır. IDF (Ters Doküman Sıklığı) DF değerinin logaritması alınarak hesaplanır.

KP-Miner denetimsiz anahtar kelime çıkarma yöntemlerinden biridir (El-Beltagy ve Rafea, 2009). Üç aşamadan oluşmaktadır; aday anahtar kelimelerin seçilmesi, aday anahtar kelimelerin ağırlıklarının bulunması ve aday kelimelerden anahtar kelimelerin bulunması. Aday kelimelerin seçilmesi için önce n-gram bulunur ardından etkisiz kelimeler (Stop-Words) filtrelenir. TF-IDF ve kelimenin pozisyonuna göre

ağırlıklandırma işlemi yapılır. Puanlamaya göre en yüksek puanı alan aday kelimelerin anahtar kelimeler olduğuna karar verilir. Bu çalışmanın, bilimsel makaleler üzerinde iyi sonuçlar verdiği görülmüştür.

Dil tabanlı denetimsiz yaklaşım dilin özelliklerine göre anahtar kelimelerin çıkarılması yöntemidir. Sözcüksel ve sözdizimsel analizler yapılır. Kelimenin pozisyonu, paragrafın neresinde bulunduğu gibi özellikleri dikkate alınır. Bu yaklaşım, kelimelerin dilsel özelliklerini kullanır.

Başlık tabanlı denetimsiz model dokümandaki konuşulan bütün konular için bir anahtar kelime bulmayı hedeflemektedir (Yangve ve diğ., 2016). Bu konuda araştırılan en önemli tekniklerden biri olan KeyCluster metodu Wikipedia ve birlikte oluşum istatistiklerini kullanarak benzer anahtar kelimeleri kümelemektedir. Her kümenin, belgenin belirli bir konusuna karşılık gelmesi ve her kümeden aday anahtar kelimelerin seçilmesiyle tüm konuların ele alınması amaçlanmıştır.

Otomatik anahtar çıkarımı yapılabilmesi için ve kelimeler arasındaki ilişkilerin, benzerliklerin bilgisayarlar tarafından anlaşılabilmesi için kelimelerin bilgisayarın anlayabileceği şekilde ifade edilmesi gerekmektedir. Kelimelerin bilgisayarlar tarafından anlaşılabilir olması için kelimeler vektörlerle ifade edilmiştir. Anahtar kelime çıkarmak için kelimelerin vektörlerinin kullanılmasıyla ilgili birçok çalışma yapılmıştır.

Her kelimenin geçtiği belgeye göre yerel kelime vektörü (Papagiannopoulou ve diğ., 2018) oluşturulmuştur. Kelime vektörleri yardımıyla kelimelerin anlamları arasında bir ilişki yakalanabilmektedir. Anahtar kelime çıkarmada kelime vektörlerinin çok iyi sonuçlar çıkardığı görülmüştür. Bilimsel makalelerin anlaşılabilmesi için kelime vektörleri kullanılarak anahtar kelimelerin tanımlanması amacıyla yapılan çalışmanın (Wang ve dig., 2015) diğer algoritmalara göre iyi performans sergilediği görülmüştür.

3.2. Kelime Vektörleri

Kelime vektörü/temsili, kelimeleri bilgisayarın anlayacağı şekilde temsil etme işlemidir. Benzer sözcüklerin benzer temsillere sahip olduğu bir metin için öğrenilmiş bir temsildir. Bu yöntem, doğal öğrenme sürecinin en önemli bir parçası olarak kabul edilmektedir. Kelime temsilleri, kelimeleri bir vektör uzayında sayısal bir vektör olarak temsil etmeyi

amaçlayan farklı teknikler kullanılarak oluşturulur. Sözcükler vektörlerle temsil edilir ve bu vektörler sinir ağına benzeyen yöntemlerle öğrenilir.

Kelime temsili yöntemleri, bir metin külliyatından önceden tanımlanmış bir kelime hazinesi için vektörü öğrenir. Bu öğrenme süreçleri, bir sinir ağı kullanarak veya istatistiksel yöntemler kullanarak olabilir. Aşağıdaki Şekil 3.2'de bir kelime temsilinin temel matematiksel denklemi gösterilmektedir.

$$f_\theta(W_n) = \theta_n$$

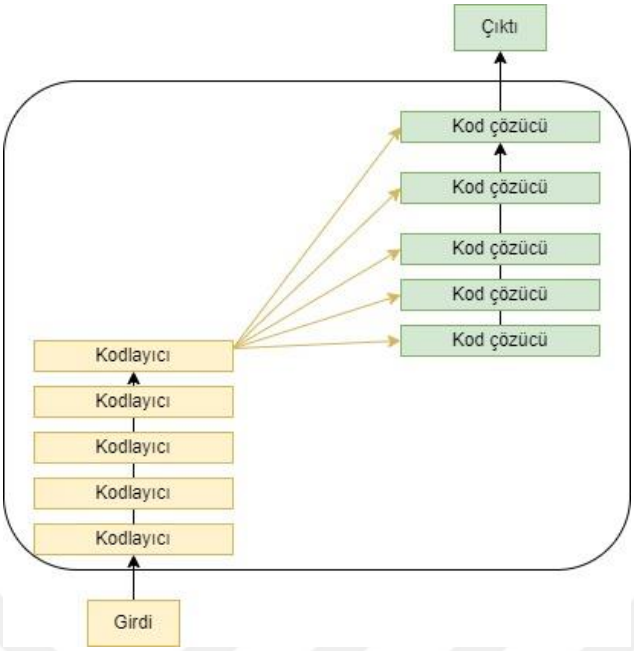
Şekil 3.2. Kelime temsilinin matematiksel gösterimi

Bu çalışmada sadece kelime temsillerinin yapısı değil ayrıca transformatör yöntemi de incelenmiştir.

Transformatörler, sinir ağı mimarisini kullanan yöntemlerdir. Bir belgedeki belirteçler için yoğun, bağlam duyarlılığı temsillerini hesaplar ve analiz ederler. Transformatörler, büyük kodlama-kod çözme mekanizmasını kullanarak vektörler üretir. Uzun menzilli bağımlılıkları öğrenmelerine izin veren tüm dizi üzerinde çalışırlar ve algoritmanın parçaları, eğitim süresini azaltmak için paralel olarak işlenebilir. Mekanizma, kodlayıcı ve kod çözücü yığınından oluşur ve Şekil 3.3'de gösterilmiştir.

Kodlayıcılar, giriş olarak kelime temsilini ve kelimenin pozisyonunu alırlar ve bu girişlere belirteç denir. Girdiyi işledikten sonra, kodlayıcının ilk katmanı "öz-dikkat" katmanıdır, burada aynı sıradaki ilgili belirteçler algılanır. Mesafe sorun değildir. Sonraki "Ekle ve Normalize Et" katmanı, öz-dikkat çıkışını giriş dizisiyle ekler ve normalize eder. Bu katmandan sonra belirteçler ileri beslemeli sinir ağı tarafından işlenir. Adım paralel olarak işlenebilir. Son olarak, sinir ağının çıkışı normalleştirilmiş ve kod çözücüleri beslemeye hazır olarak eklenir.

Kod çözücü yığını da aynı mimariye sahiptir. Bunlar; öz-dikkat katmanı, 3 adet Ekle ve Normalize Et katmanı ve ileri besleme katmanlarıdır. Burada ilk iki "Ekle ve Normalize Et" katmanı arasına bir katman daha eklenmiştir. Kodlayıcı-Kod çözücü dikkat katmanı, mekanizma, giriş dizisinin hangi belirteçlerinin mevcut çıkış belirteci ile daha alakalı olduğu hakkında bilgi verir.



Şekil 3.3. Transformatör hattı

3.2.1. Word2Vec

Word2Vec, yaklaşık olarak 2013 yılında Google tarafından tanıtılan (Mikolov ve diğerleri, 2013), kelime temsilleri oluşturmak için kullanılan bir yöntemdir. Kelimelerin yüksek kaliteli, dağıtılmış ve sürekli yoğun vektör temsilini hesaplamak ve oluşturmak için sinir ağı mimarisini kullanır. Her temsil, bağlamsal ve anlamsal benzerliği yakalar. Büyük bir veri kümesini işleyen denetimsiz bir modeldir.

Word2Vec'in iki sinir ağı mimarisi vardır: Sürekli Atlama Gram (SG–Skip Grams) modeli ve Sürekli Kelime Çantası (CBOW-Continuous Bag of Words) modeli. Bu modellerdeki fark kayıp fonksiyonundadır. Kayıp fonksiyonu, eğitimin her döneminde modeli güncellemek için kullanılan fonksiyondur.

Ancak, Word2Vec, büyük verilerle uğraşırken bazı sorunlarla karşı karşıyadır. Bu sorunların çözümü için pek çok çözüm önerilmiştir. Bunlardan bir tanesi diğerlerine göre daha iyi sonuçlar vermektedir. Bu çözümde benzer verileri kümelemek için bir strateji geliştirilmiştir. Bu strateji veri boyutunu azaltmak için de kullanılabilir.

Bunu başarmak için iki yöntem kullanıyoruz. İlk yöntemde, kelime vektörlerini oluşturmak için eğitim verilerimiz ile Word2Vec'i besliyoruz ve her adımda doğrusal hesaplama kullanarak her bir kelime vektörünün anlamsal ilişkili kelimelerini buluyoruz.

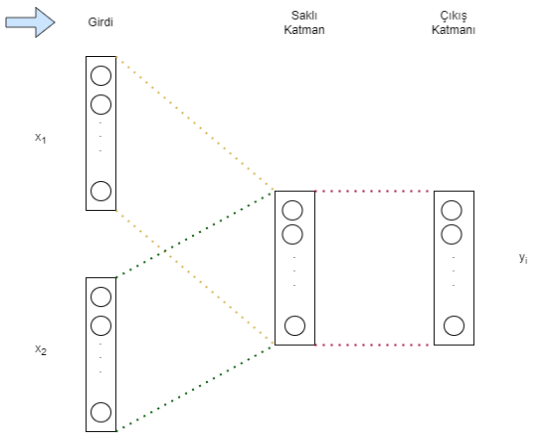
Diğer yöntemde, benzer kelimeleri K-means kümeleme algoritmasını kullanarak benzerleri gruplandırıyoruz. K'ye verilen değerlerle K kümeleri oluşturuyoruz. Böylece, kelime dağarcığı aracılığıyla kelime vektörleri oluşturmak yerine, her bir kümedeki içeriklere dayalı kelime vektörleri yapıyoruz (Zhang ve diğ., 2013). Bu çalışmada, kullanılan kelime vektörlerinin boyutu 100'dür ve 3.2.1.1'de anlatılan CBOW modeli kullanılmıştır.

3.2.1.1. CBOW Modeli

CBOW yöntemi, modeldeki kelimelerin girdilerine dayalı olarak mevcut hedef kelimeyi tahmin etmeyi amaçlar. Bu modelde, pencerenin merkezinde bulunan kelime hedef kelimedir ve pencerenin merkezinde olmayan kelimeler girdi olarak kullanılır.

Örneğin, "*the quick brown fox jumps over the lazy dog*" cümlesinde, ([bağlam_penceresi], hedef_kelime) türü çiftleri yapılabilir. Pencere boyutunun iki olduğunu kabul ederek şu şekilde sonuçlar çıkarılabilir: ([quick, fox], brown), ([the, brown], quick), ([the, dog], lazy)**.** Bu nedenle modelin bağlam penceresi kelimelerinden yola çıkarak hedef kelimeyi tahmin etmeye çalıştığını söyleyebiliriz. Birçok kütüphane CBOW modelin uygulanmasına sahiptir. Modelin uygulanması dört adımdan oluşur. Aşağıdaki Şekil 3.4'te modelin temel tasarımı gösterilmektedir.

Adım 1. Külliyat oluşturulur ve tanımlayıcıyla eşleştirilir
Adım 2. CBOW üreteci oluşturulur
Adım 3. Veriler eğitilir
Adım 4. Kelime temsilleri üretilir

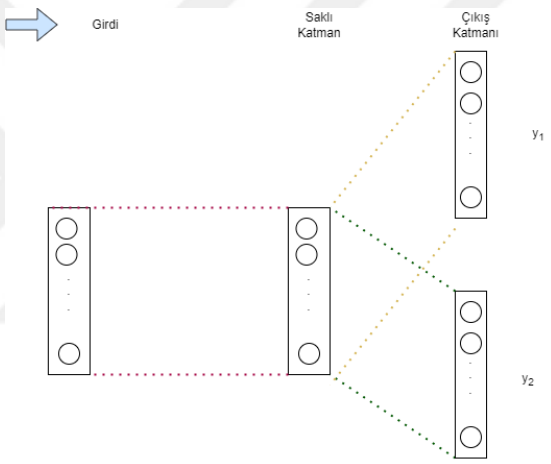


Şekil 3.4. CBOW modelinin gösterimi

3.2.1.2. Skip Grams Modeli

Bu model, CBOW yönteminin aksine, hedef kelimeden yola çıkarak çevresindeki kelimeleri tahmin ederek öğretilir. Algoritmadan tahmin edilen kelimeler, cümledeki mevcut kelimeden önce ve sonra bulunur. Model dört adımda tarif edilebilir ve Şekil 3.5'te modelin tasarımı gösterilmektedir.

Adım 1. Hedef ve bağlam kelimelerinden yoğun vektörler üretilir. Bağlam kelimeler, algoritmaya ayrı katmanlarda girdi olarak verilir
Adım 2. İki vektörün nokta çarpımı hesaplanır
Adım 3. Çıktı bir sigmoid katmana girilir ve çıktı 1 veya 0 olur
Adım 4. Sonuç gerçek değerle karşılaştırılır ve kayıp fonksiyonu güncellenir



Şekil 3.5. Skip grams modelinin gösterimi

3.2.2. Doc2Vec

Denetimsiz ve tahmine dayalı bir modeldir. Sinir ağı mimarisi kullanılmıştır. Bir dökümanın uzunluğundan bağımsız olarak dökümanı temsil eden bir vektörün çıkarılması amaçlanmıştır. Doc2Vec kelime ile dökümanlar arasındaki bağlantıyı anlamsal olarak çıkarmakta kullanılır ve dökümana ait temsilin üretilmesini sağlar. Doc2Vec ait iki model bulunmaktadır. Bunlar; PV-DM (Dağıtık Bellek) ve PV-DBOW (Dağıtık Kelime Torbası). Bu çalışmada PV-DBOW modeli kullanılmıştır ve vektör boyutu 24 seçilmiştir.

3.2.3. GloVe

GloVe, denetimsiz bir öğrenme algoritmasıdır. Word2Vec modelinin yerellik sorununa çözüm üretmek için geliştirilmiştir. Global korpus istatistiği yakalandığından yöntem

Global Vector olarak adlandırılır. Kelimelerin birlikte bulunma durumları dahil edilerek kelime vektörleri üretilir. Birlikte bulunan kelimelerin hangi sıklıkta bir arada bulunduklarına gösteren bir matris oluşturulur. Bu çalışmada, GloVe modelinde kullanılan kelime vektörünün boyutu 50'dir.

3.2.4. fastText

fastText modeli, skip gram yönteminden elde edilen vektör temsillerini iyileştirmek için önerilmiştir. Modelin arkasındaki mantık, n-gram uzunluğunda alt kelimeler oluşturmak ve alt kelimelerin kelime temsillerini üretmektir. Örnek olarak "eating" kelimesini alırsak. Kelimenin 3 gramı şu şekilde "ea eat ati tin ing ng" gösterilebilir. Amaç doğrudan metinde bulunmayan sözcükleri elde etmektir. Bu çalışmada, fastText modelinde kullanılan kelime vektörünün boyutu 24'tür.

3.2.5. BERT

BERT, kod çözücüler yerine kodlayıcıların kullanıldığı transformatör tabanlı bir modeldir. Amaç, kelimelerin sabit temsilini diğer kelime temsillerine geçirmektir. Önceden eğitilmiş modeller kullanırlar, bu nedenle eğitim diğerlerine göre çok daha hızlıdır. BERT, etraflarındaki kelimeler tarafından dinamik olarak bilgilendirilen kelime temsilleri üretir. BERT ile gelen iki farklı model var. Temel modelde 12 adet transformatör, 768 adet gizli katman ve 12 adet dikkat başlığı bulunmaktadır. Büyük model 24 trafo bloğuna, 1024 gizli katmana ve 16 dikkat başlığına sahiptir. Bu çalışmada, temel model kullanılmıştır.

3.2.6. SciBERT

SciBERT, bilimsel yayınlardan oluşan çok alanlı geniş bir külliyat üzerinde eğitilmiş BERT tabanlı bir modeldir. Bu çalışmada, SciBert modelinden oluşturulan kelime vektörlerinin boyutu 768'dir.

3.3. Veri Kümeleri

Bu çalışmada, altı farklı kelime yerleştirme davranışını karşılaştırmak ve analiz etmek için beş farklı veri seti kullanılmıştır. Veri kümeleri FAO30, Krapivin, Nguyen, Schutz ve SemEva2010'dur. Tablo 3.1'de bu veri seti hakkında detaylı bilgiler gösterilmektedir.

Tablo 3.1. Veri kümelerinin istatistikleri

#Adı	#Doküman	#Anahtar Sözcük/Doküman	#Belirteçler/Doküman	#İlgi Alanı
FAO30	30	33,23	4777	Ziraat
Krapivin	2304	6,34	8040	Bilgisayar Bilimleri
Nguyen	209	11,3	5201	Bilgisayar Bilimleri
Schutz	1231	44,69	3901	Genel
SemEval2010	243	16,47	8332	Bilgisayar Bilimleri

4. DENETİMSİZ YAKLAŞIM YÖNTEMİ

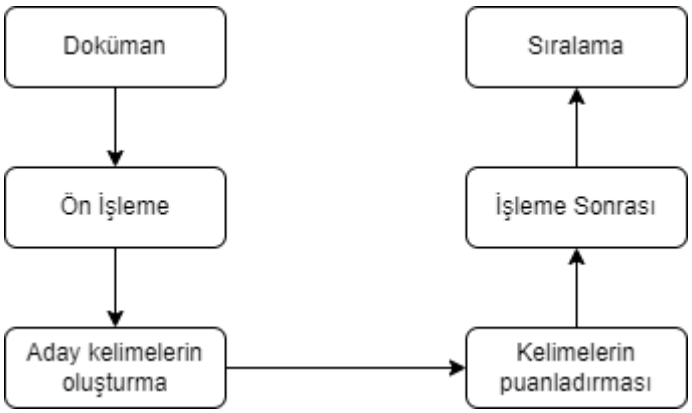
Bu bölümde Referans Vektör Algoritması anlatılacaktır. Bu algoritma denetimsiz öğrenme algoritmasıdır. Yöntem, orijinal olarak Glove ile RVA'yı uygulayan Papagiannopoulou ve Tsoumaka'nın (2018) çalışmasında kullanılmıştır. Burada, kelime temsilinde farklı kelime temsilleri kullanılarak algoritmadan yararlanmak için değişiklik yapılmıştır.

Otomatik anahtar kelime çıkarmada, denetimsiz algoritmalar, metinden istatistiksel özellikler kullanır. Bu anahtar kelime çıkarmada herhangi bir eğitim yapılması gerekmediği anlamına gelir (Sun ve diğ., 2020).

Genel olarak, denetimsiz yaklaşımlar üç adımlık bir standardı takip eder (Papagiannopoulou ve Tsoumakas, 2018). Bilimsel makalede sunulan metodoloji, bir kelime-kelime ortak oluşum matrisinde sadece sıfır olmayan öğeler üzerinde eğitim vererek istatistiksel bilgileri verimli bir şekilde kullanır ve son olarak anlamlı bir kelime vektör uzayı oluşturur (Papagiannopoulou ve Tsoumakas, 2018).

İlk adımda aday sözcük birimi bazı kurallara göre seçilir, örneğin aday sözcük bir durma sözcüğü değil bir isim veya bir sıfat olmalıdır. İkinci adımda, elde edilen kelimelerin önemi ölçülmüş ve son adım olarak en üst sıradaki kelimeler belge için anahtar kelime olarak seçilmiştir.

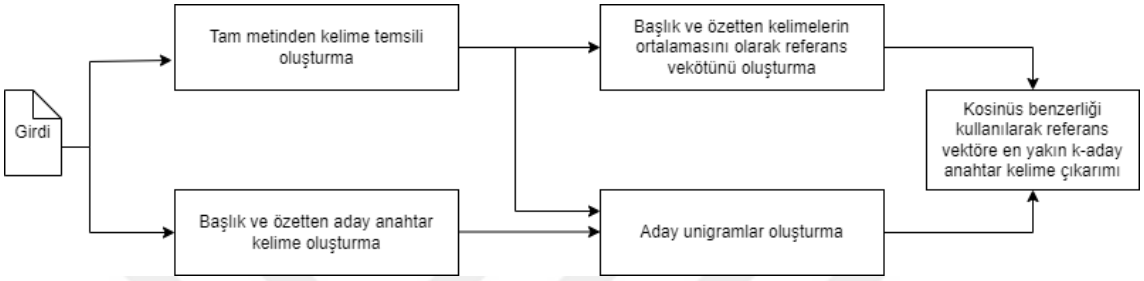
Aşağıdaki Şekil 4.1'de denetimsiz yaklaşımların nasıl çalıştığına ilişkin model gösterilmektedir.



Şekil 4.1. Denetimsiz yaklaşımların ardışık düzeni

4.1. Referans Vektör Algoritması

RVA'nın ana nedeni, yerel kelimenin vektör temsilinin anahtar kelime öbeği çıkarma sürecini iyileştirebileceği inancıdır (Papagiannopoulou ve Tsoumakas, 2018). Ayrıca, algoritma, anahtar kelimelerin her belgenin özetinde ve başlığında bulunabileceğini varsayar; bu, tüm belgeleri hesaplamak için fazladan çaba sarf edilmemesi gerektiği anlamına gelir. Şekil 4.2'de. RVA'nın arkasındaki temel konsept gösterilmektedir.



Şekil 4.2. RVA algoritmasının çalışma prensibi

4.1.1. Ön İşleme

Anahtar kelime oluşturmaya yönelik ilk adım olarak, belgenin her kelimesi ve sıklığı bir .txt dosyasına kaydedilir. Her belgenin başlığı ve özeti de .txt dosyasına kaydedilir. Belgelerdeki kelimeler, durma kelimeleri gibi düşük bilgilendirici kelimeleri silmek ve anlamlı unigramlar elde etmek için bir filtre algoritmasından süzülür.

Filtre algoritması 2'den küçük ve 36'dan büyük tüm kelimeleri siler ve herhangi bir sayı veya istenmeyen karakterler (!, @, #, $, *, =, +, . ?, > , < , &, (,), {, }, [,], |) silinir.

4.1.2. Aday Anahtar Sözcükleri Oluşturma

Filtre algoritması, belgeden tüm aday unigramları üretir. İkinci olarak bigram ve trigram adayları üretilir. Burada kullanılan algoritma, başlıkta ve özette belirli bir sırayla gelen tüm kelimeleri seçer. n= {1, 2, 3} n-gram üretilmesinin nedeni, farklı çalışmalara göre yapılan gözlemlerle ilgilidir (Gollapalli ve diğ.,2017).

4.1.3. Aday Anahtar Sözcüklerini Puanlama

Başlık ve özet dosyasında yer alan her kelime aday kelime vektörü ile temsil edilmektedir. Bu kelime vektörlerinin ortalaması alınarak referans vektörü oluşturulur. Her aday kelimenin vektörü referans vektörü ile karşılaştırılır ve en çok benzeyen anahtar kelimeler

belgenin gerçek anahtar kelimesi ile karşılaştırılır. Karşılaştırma yapmak için kullanılan yöntem kosinüs benzerliğidir ve ifadesi denklem (4.1)'de gösterilmiştir. Bigram ve trigramların hesaplanması için kelime puanlarının toplamı tercih edilir.

$$\text{Benzerlik} = \cos(d_i, d_j) = \frac{V_{d_i}^T V_j^T}{|V_{d_i}|_2 \times |V_{d_j}|_2} \quad (4.1)$$

5. DENEYLER

Bu bölümde, ilk etapta, farklı kelime temsilleri ile RVA performansının analizi anlatılacaktır. Bu metrikler, çıkarılan doğru veya yanlış anahtar kelime öbeklerinin sayısının oranını hesaplayarak performansı analiz eder (Chengyu Sun ve diğ., 2020). Ayrıca, dil tabanlı metrikler de analiz edilir. Ortalama Karşılıklı Sıra (MRR) ve Orantılı Ortalama Hassasiyet (MAP) çıktıları kullanılır. Bu ölçümler, anahtar kelimelerin bağımsız olduğunu ve daha önemli bir anahtar kelime öbeğinin daha üst konumda sıralanması gerektiğini varsayar.

Kesinlik, anahtar sözcük çıktı algoritmasının doğruluğunu gösteren, ayıklanmış olandaki gerçek anahtar sözcük sayısıdır. Kesinlik denklemi (5.1)'de gösterilmiştir.

$$\text{Kesinlik} = \frac{\text{tp}}{\text{tp+fp}} = \frac{\text{true anahtar sözcükleri}}{\text{çıkarılmıs anahtar sözcükleri}} \tag{5.1}$$

Duyarlık, tüm doğru etiketli anahtar sözcükler arasındaki gerçek doğru anahtar sözcük miktarıdır. Duyarlık denklemi (5.2)'de gösterilmiştir.

$$\text{Duyarlık} = \frac{\text{tp}}{\text{tp+fn}} = \frac{\text{true eşlenen anahtar sözcükleri}}{\text{atanmış anahtar sözcükleri}} \tag{5.2}$$

F1-skoru, kesinlik ve duyarlık arasındaki ilişkinin çıktısıdır. İdeal olarak bu değerlerin her ikisi de yüksek olmalıdır, ancak kesinlik yüksek olduğunda duyarlık düşüktür ve bunun tersi de geçerlidir. F1-skor denklemi (5.3)'te gösterilmiştir.

$$\text{F1} - \text{skor} = 2\,\text{x}\frac{\text{kesinlik x duyarlık}}{\text{kesinlik+duyarlık}} \tag{5.3}$$

Ortalama Karşılıklı Sıra (MRR), sorgular için karşılıklı sonuç sıralarının ortalamasıdır. MRR denklemi (5.4)'te gösterilmiştir.

$$\text{MRR} = \frac{\sum_{\text{d}\in\text{D}}\frac{1}{\text{rank}_\text{d}}}{|\text{D}|} \tag{5.4}$$

Orantılı Ortalama Hassasiyet (MAP), çıktı listesi anahtar sözcüklerinin sıralaması göz önüne alındığında, MAP, AP'nin ortalamasıdır. AP şu şekilde tanımlanır; N liste uzunluğudur, LN ilgili öğe sayısıdır, p(N) kesinliktir ve gd(n) öğe altın bir anahtar sözcükse 1'dir, aksi takdirde 0'dır. MAP denklemleri (5.5) ve (5.6)'de gösterilmiştir.

$$AP = \frac{\sum_{n=1}^{|N|} P(n)gd(n)}{|LN|} \quad (5.5)$$

$$MAP = \frac{1}{n}\sum_{I=1}^{n} AP_i \quad (5.6)$$

5.1. Veri Kümesine Göre Sonuçlar

Bu çalışmada 5 farklı veri seti kullanılmıştır. Nguyen (Kleinberg ve diğ., 2011), SemEval2010 (Brin ve Page, 1998), Krapivin (Krapivin ve diğ., 2009), Schutz (Schütz, Einhäuser, 2018), Fao30 (Witten ve diğ., 2008). Aşağıdaki tablolarda, ilk 10 anahtar kelimeye dayalı deneysel sonuçlar gösterilmektedir.

Tablo 5.1'de F1@10 sonucunu bildirir. Gösterildiği gibi, GloVe iki veri kümesi için en iyi sonuçları elde etti, iki veri kümesi için SciBert ve bir veri kümesi için Word2Vec en iyi sonuçları elde etti. GloVe ve Word2Vec, sırasıyla bir ve üç veri kümesi için ikinci en yüksek puanı elde etti.

Tablo 5.1. Karşılaştırma sonucu: F1@10

Veri Kümeleri	**FAO30**	**Krapivin**	**Nguyen**	**Schutz**	**SemEval2010**
Glove	0,212	**0,311**	**0,354**	0,332	0,341
Word2Vec	**0,215**	0,303	0,344	0,319	0,353
Doc2Vec	0,192	0,275	0,316	0,329	0,344
fastText	0,200	0,289	0,327	0,310	0,343
BERT	0,194	0,274	0,308	0,333	0,342
SciBERT	0,196	0,286	0,315	**0,344**	**0,361**

Tablo 5.2'de MRR karşılaştırmasını gösterilmiştir. Sonuçların gösterdiği gibi, SciBERT dışındaki tüm modeller yüksek puanlar almaktadır. Ancak Glove, iki veri kümesi için en yüksek ikinci puanı alarak öne çıkıyor.

Tablo 5.2. Karşılaştırma sonucu: MRR

Veri Kümeleri	FAO30	Krapivin	Nguyen	Schutz	SemEval2010
Glove	0,469	**0,272**	**0,311**	**0,735**	0,307
Word2Vec	0,524	0,254	0,303	0,702	**0,566**
Doc2Vec	0,490	0,250	0,304	0,674	0,286
fastText	0,485	0,260	0,319	0,683	0,294
BERT	**0,600**	0,254	0,289	0,703	0,303
SciBERT	0,446	0,246	0,296	0,684	0,306

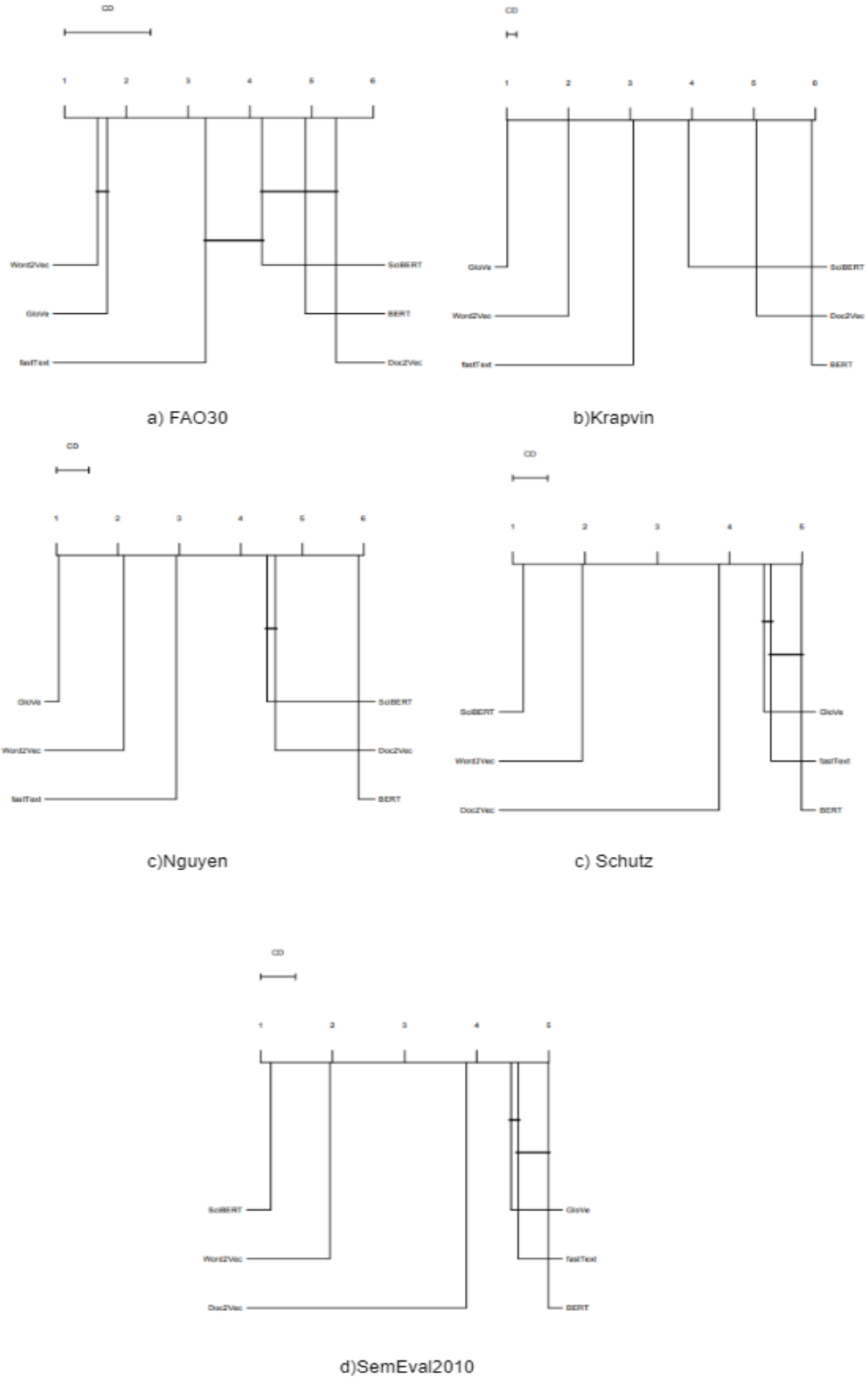
Tablo 5.3'te MAP karşılaştırma sonuçlarını gösterilmiştir. GloVe, üç veri seti için en yüksek puanı almıştır, BERT, üç veri seti için en yüksek puanı almıştır. GloVe ayrıca kalan veri kümeleri için ikinci en yüksek puanı almıştır.

Tablo 5.3. Karşılaştırma sonucu: MAP

Veri Kümeleri	FAO30	Krapivin	Nguyen	Schutz	SemEval2010
Glove	0,369	**0,358**	**0,448**	**0,642**	**0,410**
Word2Vec	0,369	0,334	0,431	0,587	0,404
Doc2Vec	0,326	0,318	0,398	0,562	0,391
fastText	0,347	0,336	0,406	0,578	0,390
BERT	**0,380**	0,146	0,390	0,552	0,405
SciBERT	0,326	0,327	0,387	0,568	0,394

5.2. Sonuçlar

Friedman testi, kelime temsillerine ait F1-skorunun, farklı veri setlerinde gösterdiği rolün ve farklılığının daha iyi anlaşılmasını sağlamak için yapılmıştır. Friedman testi, tek yönlü tekrarlanan varyans ölçümlerinin parametrik olmayan bir test analizidir. İstatistiksel analizden elde edilen en düşük sıra BERT'tir. SciBert, Doc2Vec kelime temsilleri, bu çalışmada en kötü modeller arasında yer aldılar. fastText modeli, üç kez ilk üç sıralamada yer aldı, Word2Vec ve Glove iyi bir performans sergiledi. Aşağıdaki Şekil 5.1'de Friedman testin sonuçları gösterilmiştir.



Şekil 5.1. Friedman testin sonuçları

6. SONUÇLAR VE ÖNERİLER

Bu çalışma, farklı kelime temsil modellerinin otomatik anahtar kelime çıkarımına katkısını karşılaştırmayı amaçlamaktadır. Bu çalışmada, Referans Vektör Algoritmasını kullanarak, farklı kelime temsillerini analiz ederek, beş farklı veri setindeki altı modelin performansı değerlendirildi. Sonuçların değerlendirilmesi için kesinlik, duyarlık, F1@10 ve MRR, MAP dil istatistik metrikleri kullanıldı. Ayrıca, F1@10 metriğini daha iyi anlamak için Friedman testi yapıldı. Bu çalışmanın sınırlamaları dahilinde, GloVe ve Word2Vec'in anahtar kelime çıkarma için önceden eğitilmiş modellerden daha iyi performans gösterdiği sonucuna varabiliriz.

KAYNAKLAR

Almeida, F., Xexeo, G. (2019). Word Embeddings: A Survey. *arXiv preprint arXiv:1901.09069*, Rio de Janeiro, Brazil, 25 Jan 2019.

Bahl, LR., Jelinek, F., Mercer RL. (1983). A Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE transactions on pattern analysis and machine intelligence, 5*(2),179-90. DOI: 10.1109/tpami.1983.476737

Baroni, Marco, Dinu G., Kruszewski G. (2014). Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics,* 238-247.

Beliga, Slobodan. (2014). Keyword extraction: a review of methods and approaches. *University of Rijeka, Department of Informatics, Rijeka* 1, No:9.

Bengio, Y., Sen'ecal, J.S. (2003). Quick Training of Probabilistic Neural Nets by Importance Sampling. *International Workshop on Artificial Intelligence and Statistics,* 17-24.

Bharti, Kumar S., Babu. (2017). Automatic keyword extraction for text summarization: A survey. *arXiv:1704.03242*.

Brin, S., Page, L. (1998). The Anatomy of a Large-scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, *30*(1-7), 107-117.

Florescu, C., Caragea, C., (2017). Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long paper)*. 1105-1115.

Gollapalli, S.D., Li, L., Yang, P. (2017). Incorporating Expert Knowledge into Keyphrase. In *Proceedings of the AAAI Conference on Artificial Intelligence, 31*(1).

Goodman, J. (2001). Classes for Fast Maximum Entropy Training. *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, vol. 1, 561-56. DOI: 10.1109/ICASSP.2001.940893

Göz, F., Mutlu, A. (2021). Automatic Keyword Extraction From Text Documents. In *Digital Technology Advancements in Knowledge Management* ,71-91. IGI Global.

Khattak, F. K., Jeblee, S., Pou-Prom, C., Abdalla, M., Meaney, C., & Rudzicz, F. (2019). A survey of word embeddings for clinical text. *Journal of Biomedical Informatics*, 100, 100057.

Kleinberg, J. M., Newman, M., Barabási, A. L., Watts, D. J. (2011). Authoritative Sources in a Hyperlinked Environment, Princeton University Press.

Landauer, T. K., Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211–240. DOI: 10.1037/0033-295X.104.2.211

Lebret, R., Collobert, R. (2013). Word Emdeddings through Hellinger. *arXiv preprint arXiv:1312.5542*.

Medelyan, O., Witten, I.H. (2006). Thesaurus based automatic keyphrase indexing. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, 296-297.

Mikolov, T., Chen, K., Corrado, G., Dea, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. DOI: 10.48550/arXiv.1301.3781.

Mikolov, T., Karafiat, M, Bur- get, L., Cernocky, Jan., Khudanpur, S. (2010). Recurrent Neural Network Based Language Model. In *Interspeech*, vol. 2, no. 3, 1045-1048, 2010.

Mikolov, T., Kopecky, J., Burget, L., Glembek, O., Cernocky, J. (2009). Neural Network Based Language Models for Highly Inflective Languages. *IEEE international conference on acoustics, speech and signal processing*, 4725-4728. *DOI:* 10.1109/ICASSP.2009.4960686

Mnih, A., Hinton, G. (2007). Three New Graphical Models for Statistical Language Modelling. In*Proceedings of the 24th international conference on Machine learning*, 641-648. DOI: 10.1145/1273496.1273577

Nasar, Z., Jaffry, S. W., & Malik, M. K. (2019). Textual keyword extraction and summarization: State-of-the-art. *Information Processing & Management*, *56*(6), 102088.

Papagiannopoulou, E., Tsoumakas, G. (2018). Local Word Vectors Guiding Keyphrase Extraction. *Information Processing & Management*, *54*(6), 888-902. DOI: 10.1016/j.ipm.2018.06.004

Pennington, J., Socher, R., Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532-1543.

Rada, M. (2004). Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 170–173, Barcelona, Spain.

Rohde, D.L.T., Gonnerman, L.M., Plaut, D.C. (2006). An Improved Model of Semantic Similarity Based on Lexical Co-occurence. *Communications of the ACM, 8*(627-633),116.

Rolling, L. (1981). Indexing consistency, quality and efficiency. *Information Processing & Management*, *17*(2), 69-76.

Salton, G., Wong, A., and Yang, C.S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM, 18*(11), 613-620. DOI: 10.1145/361219.361220

Sun C., Hu, L., Li, S., Li, T., Li, H, Chi, L. (2020). A Review of Unsupervised Keyphrase Extraction Methods Using Within-Collection Resources. *Symmetry*, 12(11), 1864. DOI: 10.3390/sym12111864.

Turney, D. (1999). Learning to Extract Keyphrases from Text. *Institute for Information Technology, technical report ERB-1057*.

Turney, P.D., Pantel, P. (2010). From frequency to meaning: Vector Space Models of Semantics. *Journal of artificial intelligence research*, 141-188. DOI: 10.1613/jair.2934

Wang, R., Liu, W., McDonald, C., (2015). Using word embeddings to enhance keyword identification for scientific publications. In *Australasian Database Conference*, 257-268.

Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C. Nevill-Manning, C.G., (2005). Kea: Practical automated keyphrase extraction. In *Design and Usability of Digital Libraries: Case Studies in the Asia Pacific*, 129-152.

Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., Nevill-Manning, C.G. (1999). KEA: practical automatic keyphrase extraction. *Proceedings of the fourth ACM conference on digital libraries*, 254-255.

Yang, K., Chen, Z., Cai, Y., Huang, D., Leung, H.F. (2016). Improved automatic keyword extraction given more semantic knowledge. In *International conference on database systems for advanced applications*, 112-125.

Zhang, Y., Long, M. (2015). Using Word2Vec to Process Big Text Data. In *2015 IEEE International Conference on Big Data (Big Data)*, 2895-289. DOI: 10.1109/BigData.2015.7364114

KİŞİSEL YAYIN VE ESERLER

Dibra, I., Göz, F., Mutlu, A. Anahtar Kelime Çıkarımı için Kelime Vektörleri-Karşılaştırmalı bir Değerlendirme. *IMASCON*, Kocaeli, Türkiye, 13-14 May 2022.

ÖZGEÇMİŞ

Irma Dibra, liseyi İşkodra/Arnavutluk'ta Hasan Rıza Paşa Koleji'nde bitirdi. 2013 yılında Kocaeli Üniversitesi'nde Bilgisayar Mühendisliği okumak için Türkiye'ye taşındı. Bir yıl Türkçe eğitimi aldıktan sonra, 2014'te üniversiteye başladı ve iyi bir ortalama ile 2018'de mezun oldu. Aynı yıl, halen devam etmekte olan yüksek lisansa başladı. 2018 yılından itibaren farklı firmalarda yazılım mühendisi olarak çalışan Irma, hala deneyim kazanmaktadır. Şu an savunma sanayi için çalışan bir şirkette yazılım departmanının takım liderdir.