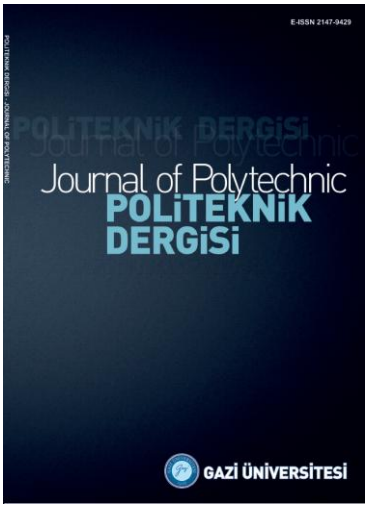


POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)

URL: http://dergipark.org.tr/politeknik

Patent dokümanlarının anlamsal benzerliğinin tespiti üzerine bir inceleme

A review on the determination of semantic similarity of patent documents

***Yazarlar (Authors)**: Ahmet KAYAKÖKÜ[1], Aslıhan TÜFEKCİ[2]*

ORCID[1]: 0000-0001-8946-1484

ORCID[2]: 0000-0002-8669-276X

To cite to this article: Kayakökü A. and Tüfekci A., "A review on the determination of semantic similarity of patent documents", ***Journal of Polytechnic***, *(*): *, (*).

Bu makaleye şu şekilde atıfta bulunabilirsiniz: Kayakökü A. ve Tüfekci A., "Patent dokümanlarının anlamsal benzerliğinin tespiti üzerine bir inceleme", ***Politeknik Dergisi***, *(*): *, (*).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: 10.2339/politeknik.1294789

Patent Dokümanlarının Anlamsal Benzerliğinin Tespiti Üzerine Bir İnceleme

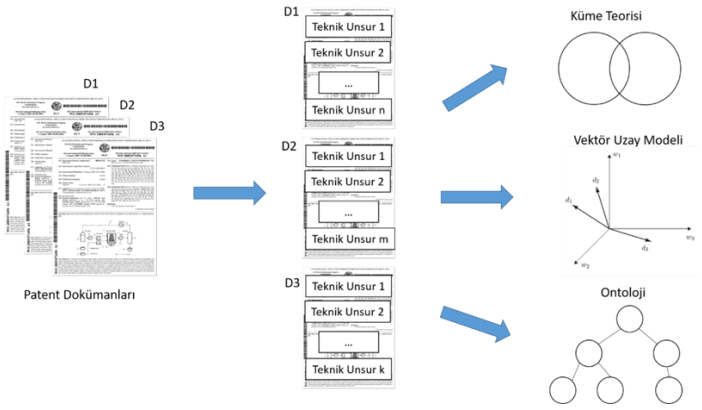
A Review on the Determination of Semantic Similarity of Patent Documents

Önemli noktalar (Highlights)

- *Patent madenciliğinde patentlenebilirlik tespiti / Patentability detection in patent mining*
- *Patentlenebilirlik tespitinde karşılaşılan zorluklar / Difficulties in patentability detection*
- *Patentlebilirlik tespitinde metin madenciliği yöntemleri / Text mining methods in patentability detection*
- *Patentlebilirlik tespitinde derin öğrenme yöntemleri / Deep learning methods in patentability detection*

Grafik Özet (Graphical Abstract)

Patent dokümanlarının anlamsal benzerliğini tespit ederken karşılaşılan zorlukların aşılması için kullanılan veri bilimine dayalı yöntemler açıklanmıştır. / The methods based on data science to overcome challenges in detecting semantic similarity of patent documents are explained.



Şekil. Anlamsal benzerlik tespit süreci / **Figure.** The process of semantic similarity detection

Amaç (Aim)

Patent dokümanlarının anlamsal benzerlik tespitinin önemine dikkat çekmek. / To draw attention to the importance of detecting the semantic similarity of patent documents.

Tasarım ve Yöntem (Design & Methodology)

Literatürde patent madenciliği ve anlamsal benzerlik alanındaki çalışmalar taranmıştır. / Studies on patent mining and semantic similarity in the literature have been reviewed.

Özgünlük (Originality)

Anlamsal benzerlik tespiti yöntemleri patentlenebilir tespitine odaklı olarak incelenmiştir. / Semantic similarity methods have been specifically examined in the context of patentability detection

Bulgular (Findings)

Patentlerin benzerlik tespitinde metin madenciliği ve derin öğrenme yöntemlerinin önemli bir rolü vardır. / Text mining and deep learning-based methods play an important role in detecting the semantic similarity of patents

Sonuç (Conclusion)

Patenlenebilirliğe karar verebilecek modeller geliştirilmesinin teknoloji yönetimi alanında ve Ar-Ge çalışmalarında önemli bir katkısı olacaktır. / Developing models that can determine their patentability would make a significant contribution to technology management and R&D.

Etik Standartların Beyanı (Declaration of Ethical Standards)

Bu makalenin yazarları çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler. / The authors of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Patent Dokümanlarının Anlamsal Benzerliğinin Tespiti Üzerine Bir İnceleme

Derleme Makalesi / Review Article

Ahmet KAYAKÖKÜ[1,2*], Aslıhan TÜFEKCİ[3]
[1]Gazi Üniversitesi Bilişim Enstitüsü, Yönetim Bilişim Sistemleri, Ankara, Türkiye
[2]Türk Patent ve Marka Kurumu, Ankara, Türkiye
[3]Gazi Üniversitesi Bilişim Enstitüsü, Yönetim Bilişim Sistemleri, Ankara, Türkiye
(Geliş/Received : 09.05.2023 ; Kabul/Accepted : 07.12.2023 ; Erken Görünüm/Early View :16.01.2024)

ÖZ

Teknik anlamda en güncel bilgileri barındıran, yüksek hacmiyle bilgi keşfi açısından müthiş bir potansiyele sahip olan ve teknoloji yönetimi alanında kilit bir rol üstlenen patent verisinin işlenmesinde patent madenciliği çalışmaları giderek önem kazanmaktadır. Patent verisi içerisinde bulunan yapısal veya yapısal olmayan verilerin hepsi önemli olsa da, patent madenciliği çalışmalarının en kritik hedefi patent dokümanlarının anlamsal benzerliğini tespit edebilmektir. Patentlerin anlamsal benzerlik tespiti ile patent başvuru sürecinin en zor ve en çok vakit alan safhası olan patentlenebilirlik kriterlerinin tespitinin otomatik olarak yapılabilmesi mümkün olacaktır. Bu makalede literatürde patent dökümanlarının anlamsal benzerliğini tespit etmek için yapılan çalışmalar incelenmiş ve kullanılan yöntemler metin madenciliği ve derin öğrenme tabanlı yöntemler olarak iki grupta kategorize edilmiştir. Patent metinlerinin, metin madenciliği yöntemleri ile yapısal hale getirilerek birbirine ne kadar benzediklerini tespit etmek için küme teorisi yaklaşımları, vektör uzay modeli yaklaşımları veya ontoloji vb. bilgi kaynaklarından faydalanılan yaklaşımlar mevcuttur. Ancak patent metinlerinin karmaşık yapısı ve kendine has terminolojisi sebebiyle bu yöntemlerden hedeflenen verim alınamamaktadır. Bu eksikliği gidermek için kullanıldığı her alanda büyük başarılar ortaya koyan yapay zeka yöntemlerinden, patent metinlerinin anlamsal olarak karşılaştırılmasında da faydalanılması gerekmektedir. Özellikle derin öğrenme yöntemlerinin yüksek kapasitesi ile karmaşık bir problem olan patentlenebilirlik kriterlerinin tespiti noktasında önemli bir mesafe kat edilebilecektir. Bu alanda çalışmalar yapılmasına rağmen etkin bir şekilde patentlenebilirlik tespiti yapabilen modeller henüz başlangıç aşamasındadır. Bu çalışmanın amacı teknoloji yönetimi alanındaki büyük ihtiyacın karşılanabilmesi adına patentlenebilirlik tespiti yaparak patent araştırma raporunun otomatik olarak hazırlanmasını sağlayacak nitelikli bir model geliştirilmesinin önemli bir adım olacağını ortaya koymaktır.
Anahtar Kelimeler: patent, patent madenciliği, metin madenciliği, anlamsal benzerlik, patentlenebilirlik, derin öğrenme.

A Review on the Determination of Semantic Similarity of Patent Documents

ABSTRACT

Patent mining studies are gaining importance in the processing of patent data, which contains the most up-to-date technical information, has a great potential in terms of information discovery with its high volume, and plays a key role in the field of technology management. Although all the structured or unstructured data in the patent data are important, the most critical goal of patent mining studies is to determine the semantic similarity of patent documents. With the semantic similarity detection of patents, it will be possible to automatically determine the patentability criteria, which is the most difficult and time-consuming phase of the patent application process. In this article, the studies carried out in the literature about determination the semantic similarity of patent documents were examined and the methods used were categorized into two groups: text mining and deep learning-based methods. Set theory approaches, vector space model approaches or ontology etc. are used to determine how similar patent texts are to each other by structuralizing them with text mining methods. There are approaches that make use of information sources. However, due to the complex structure and unique terminology of patent texts, the targeted efficiency cannot be obtained from these methods. In order to overcome this deficiency, artificial intelligence methods, which have shown great success in every field they are used, should also be utilized in the semantic comparison of patent texts. Especially with the high capacity of deep learning methods, a significant distance can be achieved in determining the patentability criteria, which is a complex problem. Although studies have been carried out in this area, models that can effectively detect patentability are still in their infancy. The objective of this study is to demonstrate that developing a qualified model capable of conducting patentability assessments and automatically generating patent search reports will be a significant step towards fulfilling the substantial needs in the field of technology management.
Keywords: patent, patent mining, text mining, semantic similarity, patentability, deep learning.

**Sorumlu Yazar (Corresponding Author)*
e-posta : ahmet.kayakoku@gazi.edu.tr

1. GİRİŞ (INTRODUCTION)

Her geçen gün hızla artan teknolojik gelişmeler ve geniş çevreler tarafından büyük yatırımlar yapılan Ar-Ge faaliyetleri, fikri mülkiyet haklarının önemini ve özellikle de patent hakkına gösterilen ilgiyi büyütmektedir. Patent başvuru ve tescil sayıları hem ülkemizde hem de tüm dünyada yüksek rakamlara ulaşmıştır. Patent başvuru ve tescil süreçlerinin bir gereği olarak yayınlanan her patent dokümanı ile kapsamlı bir teknik bilgi kamuoyu ile paylaşılmaktadır. Yayınlanan bu dokümanlarda yayın tarihi, başvuru sahibi, patent sınıfı, diğer patentlere atıflar gibi yapısal bilgilerin yanı sıra, patente konu olan buluşun detaylı teknik açıklamalarının yer aldığı yapısal olmayan metin bilgileri mevcuttur. Patent hakkını ilk başvuru yapan kişinin elde etmesi sebebiyle yayınlanan bu bilgilerin son derece güncel olması da patent verisine ayrı bir önem katmaktadır [1]. Bu çeşitli ve değerli bilgiler ile sürekli zenginleşen patent veritabanları, pek çok açıdan büyük bir potansiyele sahiptir. [2]

Patent veritabanlarının bu potansiyelinden faydalanmak için veri bilimi alanındaki gelişmelerin, patent dokümanları ve veritabanlarının karakteristik özellikleri çerçevesinde uygulanması ile patent madenciliği çalışmaları ortaya çıkmıştır. Patent madenciliği farklı uzmanlıkları içine alan bir çalışma alanı olup; hem patent dokümanlarını, veritabanlarını, patent süreçlerini kısaca patent sistemini tanımayı gerektirmekte, hem yapısal verilerin işlenmesi için veri madenciliği algoritmalarını içine almakta, hem de detaylı teknik açıklamaları ihtiva eden patent metinlerinin analiz edilebilmesi için doğal dil işleme, anlamsal analiz ve metin madenciliği yöntemlerinden faydalanmaktadır [3]. Derin öğrenme, makine öğrenmesi, yapay sinir ağları gibi yapay zeka yaklaşımları da her alanda olduğu gibi patent verisi üzerinde de çok ciddi sonuçlar ortaya koymaktadır.

Patent madenciliği çalışmaları ile çok çeşitli kazanımlar elde edilebiliyor olmakla birlikte bunların en önemlilerinin aşağıdaki amaçlar olduğu söylenebilir [4]:

- Patentlenebilirlik tespiti,
- Patent ihlal tespiti,
- Patentlerin teknik sınıflandırılması,
- Patenlerin özetlenmesi,
- Patentlerin teknolojik analizi ile stratejik teknoloji yönetimi
- Patent değerleme

Bu amaçlardan patenlenebilirlik tespiti, patent başvuru sürecinin en önemli aşamasıdır. Bir buluşun patent alabilmesi için sahip olması gereken 3 tane patentlenebilirlik kriteri vardır. Bunlar yenilik, buluş basamağı ve sanayiye uygulanabilirliktir. Yenilik kriteri tekniğin bilinen durumunda başvuru konusu buluşun teknik özelliklerinin tümüne birden sahip olan başka bir doküman bulunmadığı anlamına gelirken, buluş basamağı kriteri de başvuru konusu buluşun mevcut teknikteki en yakın benzerlerine göre öngörülemeyecek derecede önemli teknik etkilere sahip olduğunu ifade etmektedir. Bunlardan yenilik ve buluş basamağı kriterlerinin belirlenmesinde tekniğin bilinen durumunda araştırma yapılmakta ve aynı buluş konusundaki patent dokümanları incelenmektedir. Bu aşamada araştırılan patent dokümanı ile benzer patent dokümanları anlamsal olarak karşılaştırılmakta, patentlenebilirlik açısından kapsamlı teknik muhakemeler yapılmaktadır. Bir patent dokümanının patent alıp alamayacağı bu analizler ile anlaşılmaktadır.

Patentlenebilirlik araştırma neticesinde patent araştırma raporu hazırlanmaktadır. Ülkemizde ve başka bir çok ülkede yaygın olarak kullanılan rapor formatına göre mevcut teknikte yer alan ve buluşun yenilik veya buluş basamağı kriterine sahip olmadığı tek başına gösteren dokümanlar araştırma raporunda X kodu ile yer almakta, başvurunun buluş basamağı kriterinden yoksun olduğunu başka bir dokümanla birlikte değerlendirildiğinde gösteren dokümanlar Y kodu ile rapora yerleştirilmektedir. Rapordaki A kodlu dokümanlar ise buluş ile ilgili olan ancak buluşun teknik kapasitesine sahip olmayan dokümanları göstermektedir. Bir buluşun patentlenebilir olması için raporunda sadece A kodlu dokümanlar bulunmalıdır.

Ancak patent dokümanlarının karmaşık yapıda olması, patent literatürünün kendine özgü yapısı, benzer patent dokümanlarına ulaşabilmenin ciddi bir çaba gerektirmesi, patentlenebilirlik kriterlerinin tespit edilmesinde detaylı anlamsal analizlere ihtiyaç duyulması gibi sebepler patentlenebilirlik araştırmasını oldukça zorlaştırmaktadır. Ulusal patent ofislerinin uzmanları da bu araştırma süreci için önemli bir zaman ayırmak durumunda kaldığından patent süreci yavaşlamakta, çok sayıda patent başvurusu araştırma aşamasında beklemektedir. Tüm dünya genelinde patent tescil süreçlerinin oldukça uzun sürmesinin önemli bir sebebi patentlenebilirlik araştırmasının zaman alıcı bir işlem olmasıdır.

Patentlenebilirlik araştırması tescil otoritesince yapıldığı gibi başvuru öncesinde, başvuru sahibi tarafından buluşunu değerlendirmek amacıyla da yapılabilmektedir. Zira yayınlanmış tüm patent dokümanlarına patent arama motorları vasıtasıyla herkes tarafından erişilebilmektedir [5]. Ancak birçok başvuru sahibi nitelikli bir patentlenebilirlik araştırması yapamadığından, patentlenemeyeceği aşikar olan buluşlara patent almak için uğraşmaktadırlar. Böylece gereksiz masraflar yapılmakta, ciddi bir zaman ve emek israfı yaşanmaktadır. Patentlenebilirlik araştırmasının başvuru sahipleri tarafından etkin bir şekilde yapılabilmesi patent başvurularının kalitesini de yükseltecektir. Patentlenebilirlik araştırması neticesinde kendi başvurusuna en yakın dokümanları doğrudan görebilen başvuru sahipleri, buluşlarında geliştirmeler yapma fırsatına da sahip olacaklardır. Nitekim bir patent

başvurusu tescil otoritesine teslim edildikten sonra buluşun kapsamında herhangi bir geliştirme yapmak yasal olarak mümkün değildir. Tüm geliştirmelerin başvuru yapılmadan önce tamamlanması gerekmektedir.

Bu noktada bir patent başvurusunun patentlenebilirlik kriterlerine sahip olup olmadığını otomatik olarak tespit edebilecek bir yöntem ortaya konması teknoloji ve Ar-Ge dünyası açısından kritik bir gelişme olacaktır. Doğal dil işleme ve metin madenciliği yöntemleri ile anlamsal benzerlik tespitinde mesafe kat edilmiş olsa da anlamsal benzerlik tespitinde büyük işler başaran derin öğrenme yöntemlerine patent metinlerinin analizinde de büyük bir ihtiyaç duyulmaktadır. [6]

Bu çalışmanın amacı patentlerin anlamsal benzerlik tespiti alanında yapılan çalışmaları incelemek ve karmaşık bir problem olan patentlenebilirliğe karar verebilmek açısından gelinen noktanın yeterliliğini sorgulamaktır. Sonraki bölümde patentlenebilirlik tespiti problemi bir örnek üzerinden daha detaylı bir şekilde ortaya konacaktır. Ardından bu çalışmanın araştırma yöntemi anlatılacak ve elde edilen bulgular incelenerek patentlerin anlamsal benzerlik tespiti için kullanılan yöntemler açıklanacaktır.

2. PATENTLENEBİLİRLİK TESPİTİNDE KARŞILAŞILAN ZORLUKLAR (CHALLENGES FACED IN PATENTABILITY DETECTION)

Patentlenebilirlik tespiti yapılırken ilk olarak araştırılmakta olan patent dokümanında tarif edilen buluşu ifade edecek anahtar kelimelerin doğru bir şekilde belirlenmesi gerekmektedir. Tespit edilen bu anahtar kelimeler, patent arama motorlarında sorgulanarak ilgili dokümanlara ulaşılabilmektedir. Anahtar kelime tespiti de buluşun analiz edilerek esas teknik bileşenlerinin ortaya konması sonucu yapılabilmektedir. Bazen oldukça uzun olabilen teknik metinler ve buluşun çok çeşitli varyasyonlarının patent dökümanlarına has teknik tabirlerle anlatılması sebebiyle buluşun teknik karakterini ortaya koyabilmek kolay olmamaktadır. Patent dokümanlarında buluş tarif edilirken tüm detaylar yazılıyor olsa da bazı teknik unsurların araştırılmasına ihtiyaç olmadığı durumlar vardır. Buluşun olmazsa olmazı denilebilecek asli unsurlarıyla ilgili anahtar kelimelere odaklı olarak araştırma yapılmalı, buluşun çözümünü amaçladığı teknik probleme katkısı olmayacak unsurların anahtar kelimeleri araştırmanın odağında olmamalıdır..

Bir örnek üzerinden konuyu inceleyecek olursak WO2022115087 yayın numaralı uluslararası patent başvurusunun özet kısmı aşağıdaki gibidir:

“A SYSTEM FOR CREATING AUTOCONTENT IN VIDEO CONFERENCE INTERVIEWS

The present invention relates to a system (1) which enables to automatically create a content comprising meeting notes of a meeting and to submit these notes for at least one participant's approval, through the use of voice data in interviews that are carried out among participants in video conference interviews wherein a plurality of participants are included.”

Anılan başvuruda çok katılımcılı video konferans programlarında, toplantı içeriğinin otomatik olarak oluşturulmasına yönelik bir sistemden bahsedilmektedir. Bu buluşun 1. isteminde aşağıdaki teknik unsurlar için koruma talep edilmektedir:

“A system (1) which enables to automatically compile interview data created by mutual communication in a platform wherein a plurality of participants are included, and to share these data with other participant; characterized by

at least one electronic device (2) which is configured to run at least one application on it and to establish communication with at least one remote server by using any remote communication protocol;

at least one conference application (3) which is run on the electronic device (2), is configured to ensure that a voice and/or video interview can be carried out at least with other users;

at least one server (4) which is configured to establish communication with the electronic device (2) by using any remote communication protocol; to transfer voice and/or image data between the user who has started the voice and/or video conference interview to be carried out and participants who have joined the conference interview, over the conference application (3); to convert the voice data generated in the conference interview into text data; to process the text data by means of artificial intelligence models; to create at least one meeting note; to share the created meeting note with the participant who has started the interview, on the electronic device (2) and to share it with other participants upon the participant who has started the interview triggers.”

Görüldüğü üzere 1. istemde birçok teknik unsurdan bahsedilmekte, bazı unsurlar uzun ve karmaşık teknik tabirlerle tarif edilmektedir. Başvuru sahibi 1. istemde yer alan tüm unsurları içeren buluş için patent hakkı talep etmektedir. Bu buluşu araştırmak ve patentlenebilirlik tespiti yapabilmek için ilk olarak 1. istemde yer alan buluşu oluşturan asli unsurların neler olduğunun belirlenmesi gerekmektedir. Asli olmayan, yapısal detaylardan oluşan unsurların araştırmaya dahil edilmemesi daha isabetli olacaktır. Dikkatli bir analizden sonra buluşun asli unsurları şöyle karşımıza çıkmaktadır:

- Görüntülü görüşme yapılması
- Konuşmaların yazıya çevrilmesi
- Otomatik toplantı notu oluşturulması
- Yapay zeka kullanılması

İstemde geçen bir elektronik cihaz (one electronic device), görüntülü görüşme yapılan uygulama (one conference application which is run on the electronic device) gibi unsurların her biri için ayrı anahtar kelime belirlenmesine ihtiyaç bulunmamaktadır. Çünkü görüntülü görüşme yapılması özelliğine ulaşılan bir dokümanda zaten bir elektronik cihaz, mesela cep telefonu ya da bilgisayar ve görüntülü görüşme uygulaması yer alacaktır. Aynı şekilde, "server, database" gibi teknik unsurlar da görüntülü görüşmede bahsedilen dokümanlarda zaten mevcut olacaktır. Bazı teknik unsurlar ulaşılan bir dokümanda açıkça belirtilmemiş olsa da buluşun doğası gereği o buluşa dahil olarak kabul edilebilmektedir. Mesela "remote communication protocol" ifadesi görüntülü görüşme yapılmasını konu edinmiş bir dokümanda açıkça yazmıyor olsa da bir uzak haberleşme protokolü ile görüntü/ses verilerinin transfer edildiği bilinen bir gerçektir. Bu sebeple istemde yer alan her unsuru birebir olarak araştırmak mantıklı olmayacaktır.

Patentlenebilirlik tespitinin ilk basamağı araştırılan buluşun asli teknik bileşenlerinin belirlenmesidir. Bunun için de patent dokümanlarını ve bu dokümanlarda yer alan teknik tabirleri iyi tanımak, aynı zamanda buluşun ilgili olduğu teknik alanda da bilgi sahibi olmak gerekmektedir. Yoksa birebir bir teknik kelime karşılaştırması ile patentlenebilirlik tespitinde başarılı olmak çok güçtür.

Asli teknik bileşenlerin tespitinden sonra, bu bileşenleri tarif edecek anahtar kelimelerin belirlenmesi aşamasına geçilmektedir. Bu aşama teknik literatür hakkında geniş bir tecrübe ve bilgi gerektirmekte olup, eş anlamlı ve yakın anlamlı kelimelerin de araştırmaya dahil edilmesi sağlanmalıdır. Eş anlamlı ve yakın anlamlı kelimeler sadece sözlük ve dil bilgisi kuralları çerçevesinde düşünülmemeli, patent literatürü göz önüne alınmalıdır. Örneğin yukarıdaki istemde kullanılan elektronik cihaz tabiri cep telefonu ile aynı anlama gelmektedir. Bundan başka cep telefonu teknik unsurunun patent dokümanlarında aşağıdaki şekillerde ifade edilmesi mümkündür:

"Mobile device, mobile phone, mobile unit, mobile station, mobile terminal, smart device, smart phone, cell phone, cellular phone, portable device, portable phone, portable unit, portable terminal, wireless device, communication device, communication unit, electronic device, electronic equipment"

Görüleceği üzere dil bilgisi açısında eş anlamlı olmasa da patent literatürü açısından aynı anlama gelebilen çok sayıda farklı tabir söz konusu olabilmektedir. Nitelikli bir patentlenebilirlik araştırması yapılabilmesi için ilgili tüm kelimelerin hesaba katılması gerekmektedir. [7] Yukarıda bahsedilen video konferans programlarında otomatik olarak içerik oluşturulmasına yönelik buluşu oluşturan asli unsurların nasıl ifade edilebileceği hakkında aşağıdaki örnekler fikir vermektedir:

- Görüntülü görüşme yapılması: (video - conference, meeting, talk, chat, call, communication, conversation, phone)
- Konuşmaların yazıya çevrilmesi: (voice, speech, sound), (recognition, speech to text)
- Otomatik toplantı notu oluşturulması: (summary, note, report, draft, minutes, outline, keywords), (conference, meeting, conversation), (automatically, dynamically, simultaneously)
- Yapay zeka kullanılması: (artificial intelligence, deep learning, neural network, machine learning)

Anahtar kelimeler yukarıdaki şekilde belirlendikten sonra ilgili teknik bileşeni ifade edebilmek için bu kelimelerin birbiri ile yakınlık durumları da önem arz etmektedir. Örneğin görüntülü görüşme unsurunu ifade edebilmek için "video" kelimesi ile "conference" kelimesi birbirine yakın olmalıdır. Mesela aralarında en fazla 1 kelime yer almalıdır. Daha fazla kelimenin araya girdiği dokümanlarda konu farklılaşabilir. Ayrıca "video" kelimesi "conference" kelimesinden önce gelmelidir. Çünkü "video conference" ifadesi ile "conference video" ifadesi farklı teknik unsurları çağrıştırmaktadır. Diğer taraftan toplantı notu oluşturulması özelliğinde ise "summary" kelimesi ile "meeting" kelimesinin birbirine çok yakın olmalarına ihtiyaç yoktur. Bu iki kelimenin aynı cümlede geçtiği ya da aralarından en fazla 7 adet kelimenin yer aldığı dokümanlarda da toplantı notu oluşturulması özelliğinden bahsediliyor olabilir. "Summary" ve "meeting" kelimelerinin hangisinin önce hangisinin sonra geleceğinin de ayırt edici bir özelliği yoktur. Bu kelimelerin birbirlerine belli bir yakınlıkta olması aradığımız teknik unsura işaret edebilecektir. Benzer şekilde "voice" ve "recognition" kelimelerinin de sıralaması önemli olmadan birbirlerine en fazla 3 kelimelik bir mesafede bulunması gereklidir. Bu gibi değerlendirmelerden sonra profesyonel patent arama programlarında oluşturulan aşağıdaki gibi sorgu ifadeleri bizi en yakın dokümanlara ulaştırabilecektir:

- Görüntülü görüşme yapılması: video 1w (conferenc+ or meet+ or talk+ or chat or call or communicat+ or conversation or phone)
- Konuşmaların yazıya çevrilmesi: ((voice or speech or sound) 3d recogn+) or (speech 1w text)
- Otomatik toplantı notu oluşturulması: (summar+ or note? or report? or draft? or minutes or outline? or keyword?) 7d (conference? or meeting? or conversation?), (auto+, dynamic+, simultan+)
- Yapay zeka kullanılması: (artificial_intelligence, deep_learning, neural_network, machine_learning)

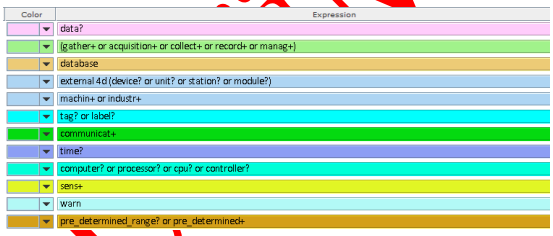
Yukarıdaki sorgu ifadelerinden yer alan 1w komutu iki kelime arasındaki en fazla 1 adet kelime olabileceğini ve kelimelerin sırasının değişmeyeceğini yani komutun solunda yazan kelimenin sağdaki kelimeden önce gelebileceğini belirlemektedir. 3d komutu ise kelimelerin sırasının değişebileceğini ve arada en fazla 3 adet kelime bulunabileceğini belirtmektedir. Kelime köklerinden sonra konan "+" operatörü o kökten gelen tüm kelimelerin sorguya dahil edileceğini "?" operatörü ise kelimenin çoğullarının da hesaba katılacağını ifade etmektedir. Nitelikli bir patent araştırması yapabilmek ve patentlenebilirlik kriterlerine doğru karar verebilmek için bu operatörlerin verimli şekilde kullanılması gerekmektedir. Böyle sorgular oluşturmak hem patent literatürüne hâkimiyet hem de teknik alanda uzmanlık gerektirdiğinden nitelikli bir patent araştırması ancak bu alanda profesyonelleşmiş araştırmacılar tarafından yapılabilmektedir. Hem de yukarıda örnekleri verilen operatörler ücretli ve özel geliştirilmiş yazılımlarda kullanılmakta olup, ücretsiz olarak herkesin erişebileceği patent arama motorlarında bu tip karmaşık arama sorguları yapılabilecek özellikler sunulmamaktadır.

Teknik bileşenleri ifade edecek anahtar kelimeler ve sorgu ifadeleri belirlendikten sonra sıra bu bileşenlerin patent dokümanının hangi kısımlarında aratılacağına gelmektedir. Tüm teknik bileşenler başvurunun özet kısmında aratıldığı takdirde çok az sayıda dokümana ulaşılma durumu olacağından böyle bir tercih verimli olmayacaktır. Diğer taraftan tüm teknik bileşenler tarifname içerisinde aratıldığı takdirde de çok fazla dokümana ulaşılabilecek ve aslında farklı konulardaki patent dokümanlarını incelemek için gereksiz bir zaman harcanacaktır. Bu noktada hangi teknik bileşenlerin özet kısmında hangi bileşenlerin tarifnamede aratılacağı hususundaki tercih de önemli bir tercihtir. Burada değinilen örnekte aşağıdaki gibi bir yaklaşım tercih edilebilir:

- Görüntülü görüşme yapılması => özet ya da başlık
- Konuşmaların yazıya çevrilmesi => tarifname
- Otomatik toplantı notu oluşturulması => özet ya da tarifname
- Yapay zeka kullanılması => tarifname

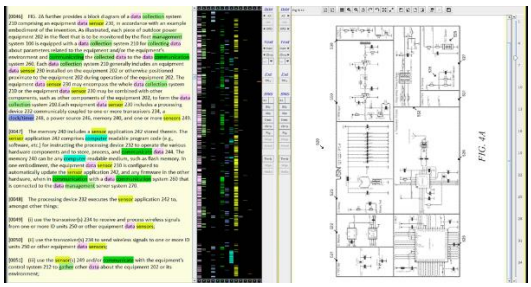
Sorgu senaryosu belirlendikten sonra yapılan araştırmalar neticesinde ilgili patent dokümanları karşımıza çıkacaktır. Bazı patent araştırmalarında 500 dökümana kadar bazılarında ise 1000 dökümana kadar incelenmesi gerekmektedir. Her bir patent dokümanının da onlarca sayfa uzunlukta olduğu düşünüldüğünde bu kadar büyük miktardaki teknik dokümanı incelemek oldukça zaman alıcı olmakta ve büyük bir iş gücü gerektirmektedir. Tüm dünyada patent tescil süreçlerinin oldukça uzun süren bir süreç olmasındaki temel sebep patentlenebilirlik araştırmasında karşılaşılan bu zorluklardır.

Yüzlerce patent dokümanını okumak mümkün olmadığı için patentlenebilirlik araştırması esnasında patent dökümanlarında ilgili teknik bileşenlerin yer aldığı kısımlara bakılmaktadır. Bunu gerçekleştirmek için profesyonel patent yazılımlarında her bir teknik bileşene farklı bir renk atanmakta ve tüm doküman içerisinde bu renklerin dağılımı bir bütün halinde gözlemlenebilmektedir. Bazı patent dokümanlarında asli unsurların sayısı çok fazla olduğundan ve bununla beraber asli olmayan unsurların da dokümandaki yerlerini görmek faydalı olabileceğinden atanacak renk sayısında bir sınır bulunmamaktadır. İstenildiği kadar farklı sayıdaki renkler oluşturulup teknik bileşenlere atanabilmektedir. **Şekil 1**'de farklı bir patent sorgusunda kullanılan kelimeler ve farklı teknik bileşenler için atanan çeşitli renkler mevcuttur:



Şekil 1. Teknik bileşenler ve atanan renkler (Technical components and assigned colors)

Atanan bu farklı renkler sayesinden patent dokümanı incelenirken odaklanılan teknik bileşenlere atanmış olan renklerin yoğunlaştığı yerlere bakılarak hızlıca dokümana konu edilen buluş hakkında fikir sahibi olunabilmektedir. **Şekil 2**'de bir doküman üzerinde yukarıdaki renklerin gösterildiği EPOQUE NET isimli profesyonel patent araştırma yazılımından bir ekran görüntüsü paylaşılmıştır:

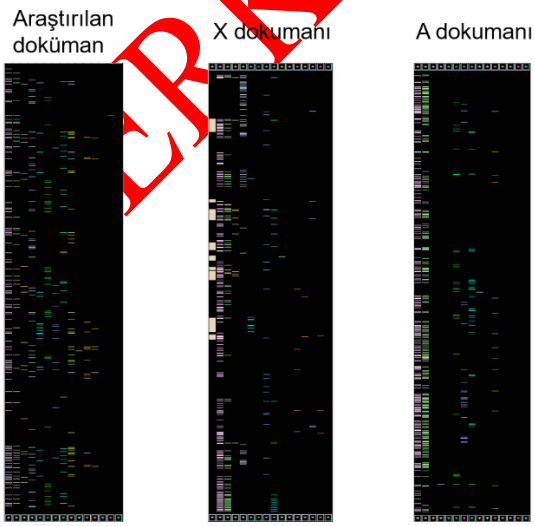


Şekil 2. Epoque Net yazılımı ekran görüntüsü (Epoque Net software screenshot)

Görüldüğü üzere doküman içerisinde ilgili teknik unsurlar renklendirilmiştir. Bu ekran görüntüsünde ekranın ortasında yer alan koyu renkli şerit şeklindeki bölgenin de patentlenebilirlik tespitinde önemli bir rolü bulunmaktadır. Bu şerit tüm doküman üzerinde belirlenen renklerin bulundukları yerleri ve dağılımını göstermektedir. Bu şerit sayesinde bir patent dokümanı incelenirken tüm dokümana bakmak yerine renklerin en çok yoğunlaştığı yerlere tıklayarak doğrudan dökümanın o kısmına ulaşılabilmektedir. İlgili teknik bileşenlerin dokümanda nasıl anlatıldığı hızlıca görülebilmektedir.

Araştırılan buluşta yer alan asli unsurların tamamının yer aldığı bir dokümana erişildiğinde, bu buluşun yenilik kriterine sahip olmadığı anlaşılmakta ve erişilen dökümanın X kategorisinde olduğu başvuru sahibine raporlanmaktadır. Eğer asli unsurların, diyelim 5 asli unsurun 4 tanesinin bir dökümanda yer aldığı bir durumda eğer geriye kalan 1 asli unsur, teknikte uzman bir kişi için aşikar olan ve beklenmedik bir teknik etki ortaya koymayan bir unsur ise erişilen dökümanın yine X kategorisinde olduğu anlaşılmaktadır. Eğer geriye kalan 1 asli unsur da önemli bir teknik etkiye sahipse ve bu 1 asli unsura da teknik alandaki başka bir dökümanda ulaşıldıysa o zaman bu iki doküman Y kategorisinde değerlendirilmekte ve araştırılan buluşun yeni olduğu ancak buluş basamağı kriterine sahip olmadığı anlaşılmaktadır. Eğer iki dökümanın birleştirilmesi suretiyle bile 5 asli unsur elde edilemiyorsa bu durumda araştırılan buluşun patentlenebilir olduğu anlaşılmakta ve en yakın dokümanlar A kategorisinde değerlendirilerek başvuru sahibine raporlanmaktadır.

Patentlenbilirlik tespitinde uzman görüşüne en çok ihtiyaç duyulan hususlardan birisi de erişilen dokümanların X, Y veya A kategorilerinden hangisine ait olduğuna karar verilmesidir. Bununla beraber yukarıda gösterilen koyu renkli şerit üzerindeki renklerin dağılımına bakarak bile patenlenebilirlik hakkında fikir sahibi olunabilmektedir. **Şekil 3**'te yer alan örnekte en soldaki şerit araştırılan buluşun tarif edildiği patent dokümanında teknik bileşenlerin dağılımını göstermektedir. Ortadaki şerit X kategorisinde olduğu raporlanan başka bir dokümandır. Görüldüğü üzere renklerin dağılımında ve yoğunluğunda bir benzerlik söz konusudur. En sağdaki şerit ise A kategorisindeki bir dokümana ait olup benzerlik oranı düşmüştür ve bazı renkler başvuruda hiç gözükmemektedir. Eğer burada farklı bir teknik alandaki yani tamamen ilgisiz bir başvurunun şeritine bakılsaydı muhtemelen tesadüfen tek tük renkler olsa da tamamen siyah renkli bir şerit ile karşılaşılabilirdi.



Şekil 3. Anahtar kelimelerin dağılımı (Distribution of keywords)

Yukarıda anlatılan ve patentlenebilirlik tespitinde ne kadar işlevsel olduğu ortaya konan teknik bileşenlere sınırsız sayıda renk atama, bu renklerin dokümandaki dağılımlarını bir şerit üzerinden görebilme ve doğrudan renklerin yoğunlaştığı bölümleri okuyabilme gibi fonksiyonlar profesyonel ve ücretli olan yazılımlarda mevcut olan özelliklerdir. Ücretsiz patent arama motorları bu işlevleri sunmamaktadır. Dolayısıyla bu işlevleri kullanırken bile uzmanlık bilgisi ve yoğun bir çaba gerektiren patentlenebilirlik tespiti, bu tip işlevlerin olmadığı durumda ise iyice zorlaşmaktadır.

Özetle patentlenebilirlik tespitinde aşağıdaki hususlar oldukça uzmanlık ve çaba gerektirmektedir:

- Başvuru konusu buluşun asli unsurlarının belirlenmesi
- Asli unsurları ifade edecek anahtar kelimelerin tespiti
- Anahtar kelimelerin birbirine yakınlıklarının belirlenmesi ve sorgunun oluşturulması
- Teknik bileşenlerin başvurunun hangi kısımlarında sorgulanacağına karar verilmesi
- Teknik bileşenlerin erişilen patent dokümanında mevcut oldukları yerlerde ulaşılabilmesi
- Teknik bileşenlerin erişilen dokümandaki dağılımını görülebilmesi
- Erişilen dokümanların X,Y veya A kategorisinden hangisinde yer aldığına karar verilebilmesi

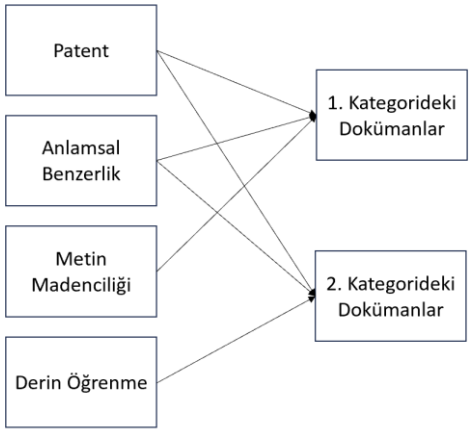
Yukarıda sıralanan bu zorlukların aşılması noktasında veri biliminde yaşanan gelişmelerin büyük bir rolü olacaktır. Hem doğal dil işleme ve metin madenciliği tekniklerinin gelişmesi ile hem de yapay zeka ve derin öğrenme tekniklerinin ortaya koyduğu potansiyel ile patentlenebilirlik tespitindeki bu zorluklar aşılabilecektir. Nitekim literatürde veri bilimindeki tüm bu gelişmeler patent dokümanları arasındaki benzerliğin tespit edilmesi için de uygulanmaktadır.

3. YÖNTEM (METHOD)

Patent dokümanlarının anlamsal benzerliğini tespit etmek amacıyla literatürde yapılan çalışmaları taramak için Scopus veri tabanında araştırma yapılmıştır. Yapılan araştırmalar iki kategoriye ayrılmıştır. Öncelikle patentlerin anlamsal benzerlik tespitini metin madenciliği ve veri madenciliği tabanlı yöntemlerle gerçekleştiren makaleleri incelenmiştir. Ardından patentlerin anlamsal benzerlik tespitini derin öğrenme ve diğer yapay zeka teknikleri kullanarak gerçekleştiren çalışmalar ele alınmıştır.

Bu konu ile ilgili dokümanlara ulaşabilmek için her iki kategorideki sorgu için de ortak olan araştırma bağlamları "patent" ve "anlamsal benzerlik" olarak belirlenmiştir. Metin madenciliği ve derin öğrenme

bağlamları da her kategorideki araştırmalar için üçüncü bağlam olarak araştırmaya dahil edilmiştir. **Şekil 4**'te çalışma yönteminin şematik gösterimi bulunmaktadır.



Şekil 4. Araştırma yöntemi (Search method)

Doğru dökümanlara ulaşabilmek için teknik bağlamları en iyi şekilde ifade edecek anahtar kelimelerin tespit edilmesi de çok önemlidir. Anahtar kelime tespitinde araştırma yapılan alana özel teknik terimlerin isabetli şekilde kullanılmasının rolü kritiktir. Bu noktada patent dokümanlarının anlamsal benzerliği üzerinde çalışıldığı ifade edecek anahtar kelimeler ile başka bir alandaki anlamsal benzerlik çalışmalarını ifade edecek kelimeler aynı olmayacaktır. Patent dokümanları için aşağıdaki sıralanmış anahtar kelimelerin hepsi anlamsal benzerlik ile ilgilidir:

patentability, novelty, inventive step, inventiveness, obviousness, prior art, search report, retrieval, semantic similarity, semantic relatedness

Dolayısıyla her iki araştırma kategorisinde ortak o larak bulunan "patent" ve "anlamsal benzerlik" kavramları için Scopus veritabanı sorgu operatörleri kullanılarak hazırlanan aşağıdaki sorgu doğru sonuçlara ulaştırabilecektir:

(TITLE-ABS-KEY (patent*) AND TITLE-ABS-KEY ((semantic* AND PRE/1 (related* OR similar*)) OR (novelty) OR (inventive PRE/1 step) OR obviousness OR inventiveness OR patentability OR retrieval OR (prior PRE/1 art*) OR (search PRE/1 report*))

Metin madenciliği alanında yapılan çalışmalara ulaşmak için yukarıdaki sorguya "mining" kelimesini AND operatörü ile eklemek yeterli olacaktır. Derin öğrenme alanındaki çalışmalar ulaşmak için de deep learning, neural network, artificial intelligence gibi kelimeleri sorguya dahil etmek gerekmektedir. Bu noktada metin madenciliği alanında yapılan çalışmalar içerisinde de derin öğrenme teknikleri kullanılabileceği için literatürün kategorizasyonunu netleştirmek adına metin madenciliğine ilişkin sorgu ifadesinde derin öğrenmeye dair anahtar kelimeleri AND NOT operatörü sorguya dahil etmek gerekmektedir. Böylece sadece metin madenciliği tekniklerine odaklanan çalışmalar ile derin öğrenme kullanan çalışmalar ayrılmış olacaktır.

Bu gerekçeler çerçevesinde bu çalışmada incelenen literatüre ulaşmak için aşağıdaki iki sorgu ifadesi kullanılmıştır:

1. kategorideki dokümanlar (Metin Madenciliği):

(TITLE-ABS-KEY (patent*) AND TITLE-ABS-KEY ((semantic* PRE/1 (related* OR similar*)) OR (novelty) OR (inventive PRE/1 step) OR obviousness OR inventiveness OR patentability OR retrieval OR (prior PRE/1 art) OR (search PRE/1 report*)) AND TITLE-ABS-KEY (mining*) AND NOT TITLE-ABS-KEY ((deep PRE/1 learning) OR (machine PRE/1 learning) OR (artificial PRE/1 intelligence) OR (neural PRE/1 network)))

2. kategorideki dokümanlar (Derin Öğrenme):

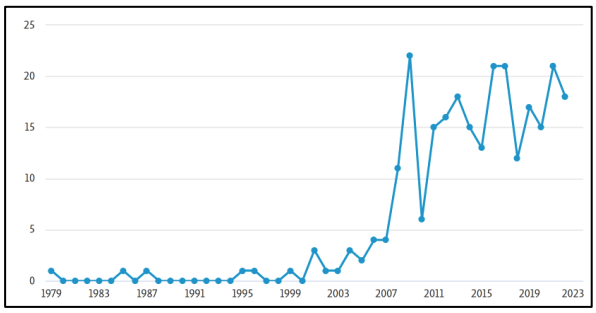
(TITLE-ABS-KEY (patent*) AND TITLE-ABS-KEY ((semantic* PRE/1 (related* OR similar*)) OR (novelty) OR (inventive PRE/1 step) OR obviousness OR inventiveness OR patentability OR retrieval OR (prior PRE/1 art) OR (search PRE/1 report*)) AND TITLE-ABS-KEY ((deep PRE/1 learning) OR (machine PRE/1 learning) OR (artificial PRE/1 intelligence) OR (neural PRE/1 network)))

Bir sonraki bölümde bu sorgular neticesinde elde edilen literatüre dair inceleme ve değerlendirmeler paylaşılacaktır.

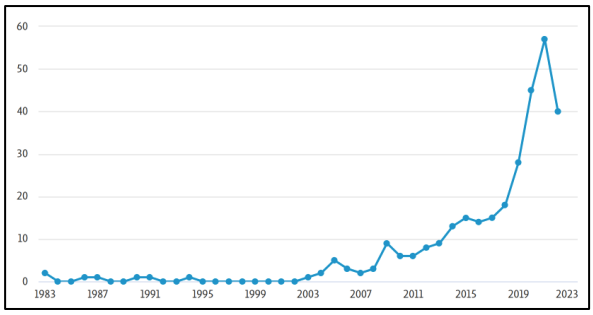
4. **BULGULAR** (FINDINGS)

Önceki bölümde açıklanan yöntemle literatürde ulaşılan ilgili dokümanlar **Hata! Başvuru kaynağı bulunamadı.**'de gösterilmiştir. Tablo incelendiğinde patent dokümanları üzerinde sadece metin madenciliği yöntemleri ile anlamsal analizler yapan çalışmaların 2019 tarihinden öncesinde yoğunlaştığı, derin öğrenme yöntemleri kullanan çalışmaların ise 2019 sonrasında yaygınlaştığı görülmektedir. Bu durum veri bilimi ve yapay zeka alanındaki gelişmelerin doğal bir sonucu olarak patent madenciliği çalışmalarına da yansımıştır.

Nitekim yöntem bölümünde belirlenen sorgular ile Scopus veritabanında yapılan araştırma sonucunda 1 Ocak 2023 tarihi itibariyle metin madenciliği alanında ulaşılan 265 dokümanın yıllara göre dağılımı **Şekil 5**'de ve yapay zeka alanında ulaşılan 306 dokümanların yıllara göre dağılımı da **Şekil 6**'da verilmiştir.



Şekil 5. Metin madenciliği yöntemleri ile patentlerin anlamsal benzerliğini tespit eden dokümanların yıllara göre dağılımı (Yearly distribution of documents about semantic similarity detection of patents with text mining mehtods)



Şekil 6. Yapay zeka yöntemleri ile patentlerin anlamsal benzerliğini tespit eden dokümanların yıllara göre dağılımı (Yearly distribution of documents about semantic similarity detection of patents with artificial intelligence methods)

Grafiklerde görülebileceği üzere veri biliminde yaşanan gelişmelere bağlı olarak metin madenciliği yöntemlerinin patent benzerlik analizlerinde kullanımı 2010'lu yıllardan sonra yükselişe geçerken yapay zeka yöntemleri 2020 lerden itibaren dramatik bir artış trendi içerisine girmiştir. Yükselişe sonra başlamasına rağmen yapay zeka yöntemleri patent benzerlik analizlerinde daha yoğun bir şekilde kullanılmaktadır.

Tablo 1 incelendiği zaman metin madenciliği alanındaki çalışmalar içerisinde özne-eylem-nesne yaklaşımı, konu modelleme teknikleri ve ontoloji kullanımı gibi yöntemlere başvurulduğu görülmektedir. Bu durumun klasik metin madenciliği yöntemlerinin patent metinlerinin karmaşık yapısı ile baş etmekte yetersiz kaldığının bir göstergesi olduğu düşünülmektedir. Patent metinleri içerisindeki anlamsal ilişkileri bulabilmek için araştırmacılar farklı yöntemleri bir arada kullanmışlardır. Yapay zeka ve derin öğrenme yöntemlerinin metin analizindeki yüksek başarısı ile çalışmalar büyük ölçüde bu alana kaymıştır.

Özellikle derin öğrenme yöntemleri benzer patenlerin tespitinde önemli bir gelişme sağlamış olsa da patentlenebilirlik tespiti noktasında henüz olması gereken düzeye gelinememiştir. **Hata! Başvuru kaynağı bulunamadı.**'de görüleceği üzere patentlerin anlamsal benzerliği ile ilgili yapılan çalışmaların içerisinde net olarak patenlenebilirlik tespitine odaklananlar az sayıdadır. Anlamsal benzerlik çalışmalarında patentlenebilirlik yaklaşımının ele alındığı çalışma oranı yaklaşık %20 olmuştur. Yani patentlerin anlamsal benzerliği ile ilgili çalışılmış ancak bu benzerlik tespitinden yola çıkarak patentlenebilirlik kararlarının verilebilmesi noktasındaki çalışmalar yetersiz kalmıştır. Bir patent dokümanına anlamsal olarak benzer olan bir çok dokümana ulaşılabilir ancak bu dokümanların yenilik ve buluş basamağı kriteri açısından değerlendirilmesi daha başka bir analiz gerektirmektedir. Literatürdeki çalışmalar iki benzer patenti tespit edebiliyor olsa da ilgili bir dokümanın X, Y veya A kategorisinden hangisine dahil olduğuna karar verecek yeterlilikte değildir.

Bu bölümde Tablo 1'de yer alan dokümanlar analiz edilerek öncelikli olarak metin madenciliği tabanlı yöntemlerin patent benzerlik analizinde nasıl kullanıldığı literatürden örneklerle incelecek ve ardından derin öğrenme yöntemleri ile elde edilen sonuçlardan bahsedilecektir.

A. METİN MADENCİLİĞİ TABANLI YÖNTEMLER İLE PATENTLERİN BENZERLİK TESPİTİ (SIMILARITY DETECTION OF PATENTS BY TEXT MINING-BASED METHODS)

Patentlerin meta verisi üzerinde patent sınıflarının veya atıflarının analizi gibi araştırmalar olsa da patentlerin metinlerine odaklanılarak benzerlik tespiti yapılması yönündeki çalışmalar daha başlangıç aşamasındadır [8]. Literatürde bir çalışmada yalnızca patent sınıflarına dayalı olarak yapılan analizlerin teknolojik olarak birbirine benzer patentlerin tespitinde yeterli olmayacağı, patent sınıfı farklı olsa da benzer teknolojik karakterlere sahip dokümanlar olabileceği yer almaktadır [9]. Anlamsal benzerlik için patent metinlerine odaklanmanın kaçınılmaz olduğu aşikârdır.

Patent dokümanlarının çok hızlı artması sonucu aşırı büyüklükte bir bilgi yığını oluşması ve patenlerin analizinin büyük bir çaba, zaman ve insan kaynağı gerektirmesi patentlerden faydalanabilmek için otomatize patent analiz araçlarına karşı büyük bir ihtiyaç doğurmaktadır. Bu araçların en kritik tarafı ise patent metinlerinin benzerliğine karar verebilmektir [10].

Patent dokümanlarının anlamsal benzerliğini tespit edebilmek için patent metinlerinin doğal dil işleme ve metin madenciliği teknikleri ile yapısallaştırılması gerekmektedir. Patent dokümanları şu özellikleri sayesinde metin madenciliği çalışmaları için elverişli bir bilgi kaynağı haline gelmiştir:

- Patent dokümanlarının yapısı kanunlarla düzenlendiğinden sabit ve belirlidir. Bu sebeple tüm dünyadaki patent dokümanları format olarak birbirinin aynıdır. Bilimsel makalelerde bile format yayınlanan dergiye göre değişmekteyken patent dokümanlarının formatı değişken değildir.

Çizelge 1. Literatür taraması neticesinde ulaşılan dokümanlar (Retrieved documents through literature search)

Yazarlar	Başlık	Yıl	Yöntem	Patentlenebilirlik İncelenmiş Mi?
AIndukuri K.V., Ambekar A.A. ve Sureka A.	Similarity Analysis of Patent Claims Using Natural Language Processing Techniques	2007	Metin Madenciliği Ontolojik benzerlik	Hayır
Bergmann I., Butzke D., Walter L., Fuerste J.P., Moehrle M.G., ve Erdmann V.A.	Evaluating the risk of patent infringement by means of semantic patent analysis: the case of DNA chips	2008	Metin Madenciliği özne-eylem-obje yapısı	Hayır
Moehrle M.G.	Measures for textual patent similarities: A guided way to select appropriate approaches	2010	Metin Madenciliği küme benzerliği	Hayır
Moehrle M.G. ve Gerken J.M.	Measuring textual patent similarity on the basis of combined concepts: Design decisions and their consequences	2012	Metin Madenciliği özne-eylem-obje yapısı	Hayır
Park H., Yoon J., ve Kim K.	Identifying patent infringement using SAO based semantic technological similarities	2012	Metin Madenciliği özne-eylem-obje yapısı	Evet
Sharma P., Tripathi R., Singh V.K. ve Tripathi R.C.	Automated patents search through semantic similarity	2015	Metin Madenciliği ontolojik benzerlik	Hayır
Cvitanić T., Lee B., Song H.I., Fu K.K. ve Rosen D.W.	LDA v. LSA: A Comparison of Two Computational Text Analysis Tools for the Functional Categorization of Patents	2016	Metin Madenciliği konu modelleme	Hayır
Sharma P., Tripathi R. ve Tripathi R.C.	Finding Similar Patents through Semantic Expansion	2016	Metin Madenciliği ontolojik benzerlik	Hayır
Walter L. Radauer A. ve Moehrle M.G.	The beauty of brimstone butterfly, novelty of patents identified by near environment analysis based on text mining	2017	Metin Madenciliği n-gram yöntemi	Evet
Arts S., Cassiman B., ve Gomez J.C.	Text matching to measure patent similarity	2018	Metin Madenciliği jaccard benzerliği	Hayır
Aras H., Türker R., Geiss D., Milbradt M, ve Sack H.	Get your hands dirty: Evaluating word2vec models for patent data	2018	Derin Öğrenme word2vec yöntemi	Hayır
Aristodemou L. ve Tietze F.	The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data	2018	Derin Öğrenme Litertür taraması	Hayır
Moehrle M.G.	Similarity measurement in times of topic modelling	2019	Metin Madenciliği Konu modelleme	Hayır
Shahmirzadi O., Lugowski A., ve Younge K.	Text Similarity in Vector Space Models: A Comparative Study	2019	Metin Madenciliği TF-TBF yöntemi	Hayır
Wang J. ve Chen Y.J.	A novelty detection patent mining approach for analyzing technological opportunities	2019	Metin Madenciliği konu modelleme	Hayır
Wang X.F., Ren H.C., Chen Y., Liu Y.Q., Qiao Y.L., ve Huang Y.	Measuring patent similarity with SAO semantic analysis	2019	Metin Madenciliği özne-eylem-obje yapısı	Hayır
Helmers L., Horn F., Biegler F., Oppermann T., ve Müller K.R.	Automating the search for a patent's prior art with a full text similarity search	2019	Derin Öğrenme doc2vec yöntemi	Evet
Krishna A.M., Jin Y., Foster C., Gabel G., Hanley B. ve Youssef A.	Query Expansion for Patent Searching using Word Embedding and Professional Crowdsourcing	2019	Derin Öğrenme fastText yöntemi	Hayır
Lee J.-S.	PatentTransformer: A Framework for Personalized Patent Claim Generation	2019	Derin Öğrenme Transformer Sinir Ağı	Hayır
Lei L., Q, J. ve Zheng K.	Patent Analytics Based on Feature Vector Space Model: A Case of IoT	2019	Derin Öğrenme Evrişimli Sinir Ağı	Hayır
Sarica S., Luo J. ve Wood K.L.	TechNet: Technology semantic network based on patent data	2020	Metin Madenciliği ontolojik benzerlik	Hayır

Çizelge 1'in devamı

Chen L., Xu S., Zhu L., Zhang J., Lei X., ve Yang G.	A deep learning based method for extracting semantic information from patent documents	2020	Derin Öğrenme LSTM ve GRU Sinir Ağları	Evet
Chung P. ve Sohn S.Y.	Early detection of valuable patents using a deep learning model: Case of semiconductor industry	2020	Derin Öğrenme CNN+bi-LSTM hibrit yöntemi	Hayır
Kim J., Yoon J., Park E., ve Choi S.	Patent document clustering with deep embeddings	2020	Derin Öğrenme doc2vec+AutoEncoder hibrit yöntemi	Hayır
Kim S., Park I. ve Yoon B.	SAO2Vec: Development of an algorithm for embedding the subject–action–object (SAO) structure using Doc2Vec	2020	Derin Öğrenme doc2vec + özne-eylem-obje yapısı	Hayır
Lee J.-S. ve Hsiang J.	Prior Art Search and Reranking for Generated Patent Text	2020	Derin Öğrenme Transformer Sinir Ağı	Hayır
Lu Y., Xiong X., Zhang W., Liu J., ve Zhao R.	Research on classification and similarity of patent citation based on deep learning	2020	Derin Öğrenme Evrişimli Sinir Ağı	Hayır
Whalen R., Lungeanu A., Dechurch L. ve Contractor N.	Patent Similarity Data and Innovation Metrics	2020	Derin Öğrenme doc2vec yöntemi	Evet
An X., Li J., Xu S., Chen L. ve Sun W.	An improved patent similarity measurement based on entities and semantic relations	2021	Metin Madenciliği özne-eylem-obje yapısı	Hayır
Jang H., Jeong Y., ve Yoon B.	TechWord: Development of a technology lexical database for structuring textual technology information based on natural language processing	2021	Metin Madenciliği ontolojik benzerlik	Hayır
Krestel R., Chikkamath R., Hewel C., ve Risch J.	A survey on deep learning for patent analysis	2021	Derin Öğrenme literatür taraması	Evet
Lo H.-C. ve Chu J.-M.	Pre-trained Transformer-based Classification for Automated Patentability Examination	2021	Derin Öğrenme Transformer Sinir Ağı	Evet
Setchi R., Spasić I., Morgan J., Harrison C. ve Corken R.	Artificial intelligence for patent prior art searching	2021	Yapay Zeka Literatür Taraması	Evet
Choi J., Lee J., Yoon J., Jang S., Kim J., ve Choi S.,	A two-stage deep learning-based system for patent citation recommendation	2022	Derin Öğrenme CSNet + CRNet	Hayır
Choi S., Lee H., Park E. ve Choi S.	Deep learning for patent landscaping using transformer and graph embedding	2022	Derin Öğrenme Transformer Sinir Ağı	Hayır
Hafner A., Damij N. ve Modic D.,	Augmented intelligence for state-of-the-art patent search	2022	Derin Öğrenme Literatür Tarması	Hayır
Jeon D., Ahn J.M., Kim J. ve Lee C.	A doc2vec and local outlier factor approach to measuring the novelty of patents	2022	Derin Öğrenme doc2vec yöntemi	Hayır
Nemani P. ve Vollala S.	A Cognitive Study on Semantic Similarity Analysis of Large Corpora: A Transformer-based Approach	2022	Derin Öğrenme Transformer Sinir Ağı	Hayır
Schellekens M.	Artificial Intelligence and the re-imagination of inventive step	2022	Yapay Zeka Literatür Taraması	Evet
Stamatis V.	End to End Neural Retrieval for Patent Prior Art Search	2022	Derin Öğrenme Transformer Sinir Ağı	Hayır
Villa A.M. ve Wirz M.	A sequential patent search approach combining semantics and artificial intelligence to identify initial State-of-the-Art documents	2022	Derin Öğrenme	Hayır
Li R., Yu W., Huang Q. ve Liu Y.	Patent Text Classification based on Deep Learning and Vocabulary Network	2023	Derin Öğrenme Transformer Sinir Ağı	Hayır
Vaish K., Rawat P., Kathuria S., Singh R., Joshi K. ve Verma A.,	Artificial Intelligence Reducing the Intricacies of Patent Prior Art Search	2023	Derin Öğrenme Literatür Taraması	Hayır

- Patent dokümanlarının sahip olduğu başlık, özet, tarifname, istemler bölümlerinin hangi bilgileri içereceği net bir şekilde tanımlanmıştır. Bu da dokümanların bölümlerinin analizini ve karşılaştırılmasını kolaylaştırmaktadır.

- Patent dokümanları tamamen teknik bir literatürden oluştuğundan duygusal veya kişisel ifadeler söz konusu değildir. [11]

Yapısallaştırma işleminde patent metinleri literatürde bilinen teknikler ile ön işlemeden geçirilerek metnin anlamına etki etmeyen ifadelerin kaldırılması, kelime köklerinin belirlenmesi gibi işlemlere tabi tutulmaktadır. Sonrasında patent dokümanın teknik karakterini ortaya koyabilmek için tarifname, özet veya istemler bölümlerinde yer alan teknik konseptler tespit edilmektedir. Bu noktada farklı uygulamalar karşımıza çıkmaktadır.

Moehrle vd. tarafından patent dokümanlarında basit konsept ve birleşik konsept olmak üzere iki tip konsept bulunduğu ifade edilmiştir. [10] Basit konseptler vites, debriyaj gibi temel konseptlerdir. Basit konseptlerin belirlenmesinde n-gram yöntemi kullanılmakta, 2 sözcükten, 3 sözcükten oluşan teknik unsurlar da ortaya konabilmektedir [8]. Ancak burada unsurlar tek tek ele alınmaktadır. Birleşik konseptler ise teknik unsurların birbiri ile ilişkisini de ortaya koyan Özne-Eylem-Obje (ÖEO) (Subject-Action-Object) yapısındaki, "güç dönüştüren vites" (gear transform power), "suyu hareket ettiren pompa" (pump moves water) gibi konseptlerdir.

Moehrle vd. bir başka çalışmalarında patent benzerlik ölçümlerinde bileşik konseptlerin basit konseptlerden daha önemli bir rol oynadığını ifade etmiştir. Basit konseptlerde tek bir teknik bileşen söz konusu olup bunların birbiri ile ilişkileri tanımlanmadığı için, bu konseptler dokümanların teknik karakterini yansıtmak için yeterli olmamaktadır. Bileşik konseptlerde ise teknik bileşenlerin birbiri ile ilişkileri hesaba katıldığından benzerlik ölçümleri daha net sonuç vermektedir. [11]

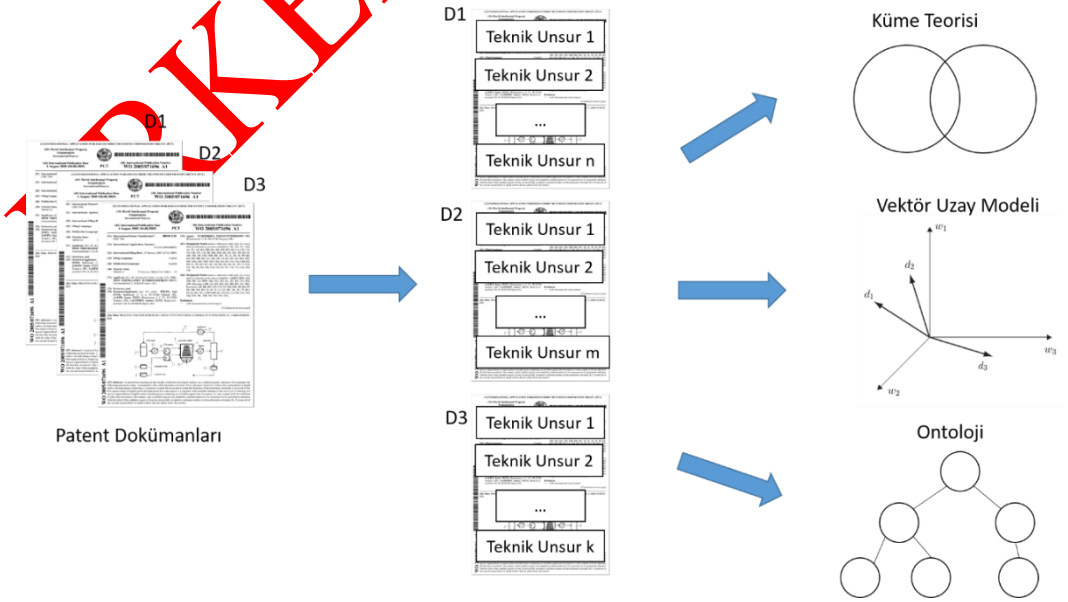
ÖEO analizlerinin incelendiği başka bir çalışmada da klasik kelime frekanslarına dayalı benzerlik yöntemlerin yetersiz olduğu, patent dokümanlarında ÖEO yapılarının daha doğru sonuç verdiği ifade edilmektedir. ÖEO yapılarının her biri için farklı ağırlıklar kullanılmış ve Farklı Ağırlıklı ÖEO yapıları (DWSAO) oluşturulmuştur [12, 13]. Literatürde, patent benzerlik analizlerinde ÖEO yapılarına sıkça rastlanmaktadır. [14]

Patent metinleri, sahip olduğu anlamı ifade edecek şekilde teknik unsurlara ayrıldıktan sonra dokümanların birbiri ile karşılaştırılması safhasına geçilir. Bu karşılaştırmaya göre dokümanlar arasındaki benzerlik durumuna karar verilecektir. Yapısallaştırılmış patent dokümanlarını birbiri ile karşılaştırmak için yine farklı yöntemler mevcuttur. Küme Teorisi yaklaşımıyla, Vektör Uzay Modeli yaklaşımıyla ve Ontoloji tabanlı yöntemlerle anlamsal benzerlik tespiti yapılmaktadır. Bu yöntemlerden sadece üzerinden çalışılan derlemden elde ettiği bilgileri işleyerek sonuca ulaşanlara Derlem Tabanlı Yöntemler; ontolojiler gibi önceden mevcut olan bilgi kaynaklarından faydalanarak sonuca ulaşanlara da Bilgi Tabanlı Yöntemler denilmektedir [6]. Bu yöntemleri gösteren bir şema **Şekil 7Hata! Başvuru kaynağı bulunamadı.**'de görülmektedir.

Anılan bu farklı yaklaşımların her birinde öğeler arasındaki benzerliği hesaplamak için çeşitli benzerlik metrikleri kullanılmaktadır. Benzerlik ölçmek için geliştirilen bir fonksiyonun benzerlik metriği olarak kabul edilebilmesi için aşağıdaki özellikleri sağlaması gerekmektedir.

x ve y öğeleri arasındaki benzerliği ölçen ve değeri [0,1] arasında değişebilen bir s(x,y) fonksiyonu için;

1. Yansıma Özelliği (reflexivity): s(x,x) = 1



Şekil 7. Patent dokümanlarının anlamsal benzerlik tespit süreci (The process of semantic similarity detection of patent documents)

2. Kimlik Özelliği (identity): s(x,y) = 1 => < x=y
3. Simetri Özelliği (symmetry): s(x,y) = s(y,x)
4. Üçgen Eşitsizliği (triangle inequality): s(x,y) ≥ s(x,z) + s(z,y)

Benzerlik ve uzaklık fonksiyonları birbirinin tümleyeni olup, x ve y öğeleri arasındaki uzaklığı hesaplamak için kullanılan d(x,y) fonksiyonu şöyle tanımlanabilir:

$$d(x,y) = 1 - s(x,y) \quad (1)$$

1) *Küme Teorisi (Set Theory)*

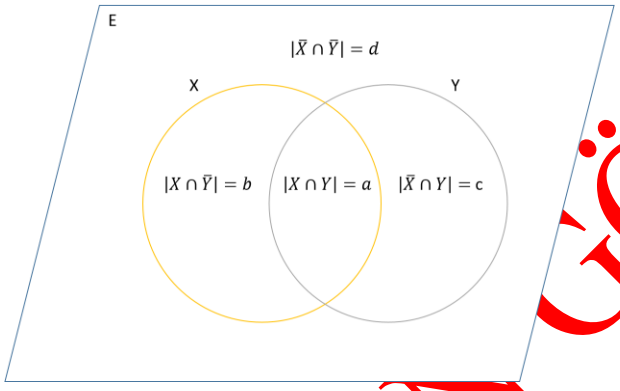
Küme teorisi yaklaşımında E evrensel kümesinde tanımlı X ve Y kümelerinde mevcut olan elemanların sayısı **Şekil 8**'de görüldüğü gibi şu şekilde verilmekte olsun:

a: Hem X hem de Y kümesinden bulunan elemanların sayısı

b: X kümesinde bulunup Y kümesinde bulunmayan elemanların sayısı

c: Y kümesinden bulunup X kümesinde bulunmayan elemanların sayısı

d: İki kümede de bulunmayan elemanların sayısı



Şekil 8. Benzerliği tespit edilecek olan x ve y kümeleri (The sets of x and y whose similarity will be detected)

İki küme arasındaki benzerliği hesaplamak için en genel ifadeyle Tverski tarafından öne sürülen aşağıdaki formül kullanılabilir [15]:

$$S_{Tverski}(X,Y) = \frac{|X \cap Y|}{|X \cap Y| + \alpha|X - Y| + \beta|Y - X|} \quad (2)$$

Burada α ve β değiştirilebilen katsayılardır. α = β = 0 durumu için en yaygın olarak kullanılan benzerlik metriği olan Jaccard benzerlik metriği elde edilir [15]:

$$S_{Jaccard}(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{a}{a+b+c} \quad (3)$$

Yine önemli bir benzerlik metriği olan Sorensen-Dice benzerliği de α = β = ½ olduğu duruma denk gelmekte olup şöyle hesaplanmaktadır [15]:

$$S_{Sorensen-Dice}(X,Y) = \frac{2 * |X \cap Y|}{|X| + |Y|} = \frac{2a}{2a+b+c} \quad (4)$$

Diğer benzerlik metrikleri olan Simple Matching ve Ochiai Benzerliği de şöyle hesaplanmaktadır [16]:

$$S_{Simple\ Matching}(X,Y) = \frac{|X \cap Y| + |\bar{X} \cap \bar{Y}|}{n} = \frac{a+d}{a+b+c+d} \quad (5)$$

$$S_{Ochiai}(X,Y) = \frac{|X \cap Y|}{\sqrt{|X||Y|}} = \frac{a}{\sqrt{(a+b)(a+c)}} \quad (6)$$

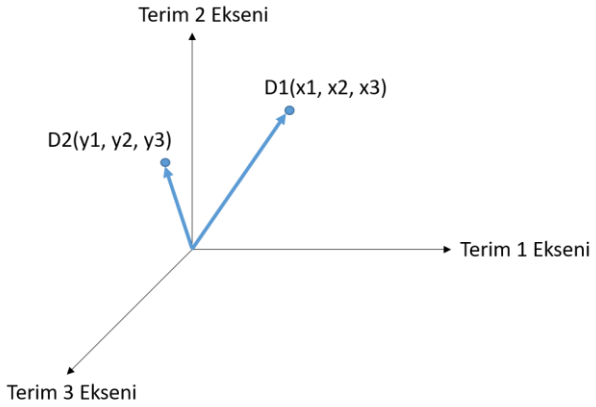
Bahsedilen bu ve benzeri metriklerin patentler arasındaki benzerlik durumunu hesaplamak için kullanılabilmesi için her bir patent çıkarılan teknik konseptlerden oluşan bir küme olarak ele alınır ve şu sayılar belirlenir: a: Her iki patentte de olan konseptlerin sayısı, b: Sadece 1. patentte olan ve 2 patentte olmayan konseptlerin sayısı, c: Sadece 2. patentte olan ve 1 patentte olmayan konseptlerin sayısı, d: İlgili teknik alanda mevcut fakat iki patentte de yer almayan konseptlerin sayısı. Bu sayılar belirlendikten sonra Jaccard, Sorensen, Inclusion gibi benzerlik ölçütleri ile patent dokümanları arasındaki benzerliğe karar verilebilmektedir [10].

Arts vd. 4.442.009 patent dokümanı üzerinden yaptıkları çalışmada toplamda 526.561 bir patent dokümanında da ortalama 37 benzersiz anahtar kelime tespit etmişlerdir. Ardından Jaccard indeksi ile patent dokümanlarının benzerliğini hesaplamışlardır. Ortalama olarak iki patent dokümanı arasında 14 ortak anahtar kelime olduğu sonucuna ulaşmışlardır. [9]

Patent ihlallerinin tespiti için yapılan bir çalışmada da patent metinlerinden ÖEO çıkarılmış, her patent sahip olduğu ÖEO yapılarından oluşan bir küme olarak ele alınmış ve Sorensen indeksine göre patentler arasındaki benzerlik tespit edilmiştir [17].

2) *Vektör Uzay Modeli (Vector Space Model)*

Bu yöntemde dokümanlar belirlenen teknik konseptlerden oluşan bir vektör olarak modellenir ve iki vektör arasındaki yakınlık hesaplanır. Ortaya çıkan sonuç dokümanlar arasındaki anlamsal benzerliği ifade eder. Vektör Uzay Modelini ifade eden bir şema **Şekil 9Hata! Başvuru kaynağı bulunamadı.**'da paylaşılmıştır.



Şekil 9. Vektör uzay modelinde patent dokümanları (Patent documents on vector space model)

Dokümanların vektörleştirilmesi için uygulanan en klasik yöntem kelime çantası (bag of words) yöntemidir. Buna göre analiz edilecek dokümanlardan oluşan derlemdeki benzersiz terimler ile derlemdeki dokümanlar arasında Doküman-Kelime matrisleri oluşturulur. Buradaki benzersiz kelimeler vektörlerin boyutlarını ifade eder. Derlemde doğal dil işleme yöntemleri ile n sayıda

benzersiz terim ortaya konulduğu düşünüldüğünde her bir doküman n boyutlu bir vektör olarak kabul edilir. Benzersiz terimlerin dokümanlardaki frekansına göre boyutlar şekillenir ve dokümanlar vektörleştirilmiş olur.

Dokümanlar vektörleştirildikten sonra iki vektör arasındaki farkı ölçmek için çeşitli formüller bulunmaktadır. D1 ve D2 dokümanlarının n terimden oluşan iki vektör olduğu düşünüldüğünde bunlar arasındaki uzaklık Minkowski uzaklık fonksiyonu ile şu şekilde ifade edilir [15]:

$$d_{Minkowski}(D1, D2) = \left(\sum_{i=1\ldots n}|D1_i - D2_i|^p\right)^{\frac{1}{p}} \quad (7)$$

Burada p = 1 durumu için Manhattan Uzaklık formülü:

$$d_{Manhattan}(D1, D2) = \sum_{i=1\ldots n}|D1_i - D2_i| \quad (8)$$

p=2 durumu için Euclidean (Öklid) Uzaklığı formülü:

$$d_{\text{Euclidean}}(D1, D2) = \sqrt{\sum_{i=1\ldots n}|D1_i - D2_i|^2} \quad (9)$$

p=∞ durumu için de Chebychev Uzaklığı formülü elde edilir.

$$d_{\text{Chebyschev}}(D1, D2) = \max_{i=1\ldots n}|D1_i - D2_i| \quad (10)$$

İki vektör arasındaki benzerliğin hesaplanması için en yaygın olarak kullanılan Cosinüs benzerliği ve Cosinüs uzaklığı ise aşağıdaki formüllerle hesaplanmaktadır:

$$S_{cosinüs}(D1, D2) = \frac{\sum_{i=1}^{n} D1_i D2_i}{\sqrt{\sum_{i=1}^{n} {D1_i}^2}\sqrt{\sum_{i=1}^{n} {D2_i}^2}} \quad (11)$$

$$d_{cosinüs}(D1, D2) = \frac{1}{2}\sum_{i=1}^{n}\left(\frac{D1_i}{\sqrt{\sum_{i=1}^{n} {D1_i}^2}} - \frac{D2_i}{\sqrt{\sum_{i=1}^{n} {D2_i}^2}}\right)^2 \quad (12)$$

Derlemdeki benzersiz kelimelerin aynı ağırlıkta olmasının önüne geçilmesi ve seyrek görülen kelimelerin daha ayırt edici olması düşüncesiyle yaygın olarak kullanılan bir yöntem de Terim Frekansı – Ters Belge Frekansı (TF-TBF) olarak bilinen yöntemdir. Doküman-Kelime matrisinde yer alan kelimelerin ağırlıkları TF-TBF formülü ile tekrar belirlenerek dokümanları ifade eden vektörler bu yeni ağırlıklara göre oluşturulur. Benzerlik ölçümleri yeni oluşturulan vektörlere göre yapılır. Patent metinlerinin benzerliğinin hesaplanmasında TF-TBF yöntemi nitelikli sonuçlar ortaya koymaktadır [18]. Klasik TF-TBF yönteminden başka TF-TBF yöntemlerinin bazı varyantları da patent-patent benzerlik analizinde kullanılmaktadır [18].

TF-TBF yöntemine göre her bir terimin (t), her bir dokümandaki (D) ağırlığını hesaplamak için aşağıdaki formül kullanılır:

$$w_{t,D=}\, tf_{t,D} * \log\left(\frac{N}{df_t}\right) \quad (13)$$

$tf_{t,D}$: *t teriminin d dokümanında görülme sayısı*

df_t: *t terimine sahip olan doküman sayısı*

N: derlemdeki toplam doküman sayısı

Kelime çantası ve TF-TBF gibi klasik matris yöntemlerinde derlemdeki tüm benzersiz kelimelerin kullanılması matrisin boyutunu çok büyütebilmekte, bu da veri üzerinde işlem yapmayı güç hale getirmektedir. Ayrıca görüntüde farklı ancak anlamsal olarak aynı olan eş anlamlı ve yakın anlamlı kelimelerin matriste farklı boyutlar olarak ele alınması verimi düşüren bir hususdur [19]. Bu dezavantajları ortadan kaldırmak için konu modelleme (topic modeling) teknikleri uygulanmaktadır. Bu yöntemlerde metin içerisinde aynı kısımlarda yer alan kelimelerin birbiri ile yakın anlamlı olduğu varsayımından hareketle, benzer kelimelerin matrislerin boyutunu küçültmekte ve eş anlamlı kelimelerin farklı boyut olarak ele alınmasının önüne geçilmektedir [19].

En yaygın olarak kullanılan konu modelleme teknikleri olan Gizli Anlamsal Analiz (Latent Semantic Analysis) ve Gizli Dirichlet Ayırımı (Latent Dirichlet Allocation) yöntemleri patent dokümanlarının benzerliğinin hesaplanmasında önemli katkılar yapmaktadır [20]. Teknolojik fırsatların tespiti alanında yapılan bir çalışmada patent dokümanları Gizli Anlamsal Analiz yoluyla işlenmekte, birbirine benzer dokümanlar kümelenmekte ve kümelere dâhil olmadığı için yenilik kriteri taşıdığı varsayılan patent dokümanları sapan veri (outlier) tespiti ile belirlenmektedir [19]. Patent dokümanlarının birbirine benzerlik durumuna göre gruplandırılması ile ilgili bir başka çalışmada da Gizli Anlamsal Analiz ve Gizli Dirichlet Ayırımı yöntemlerinin performansı, az sayıda patentten oluşan setlerde ve çok sayıda patentten oluşan setlerde ayrı ayrı karşılaştırılmıştır [21].

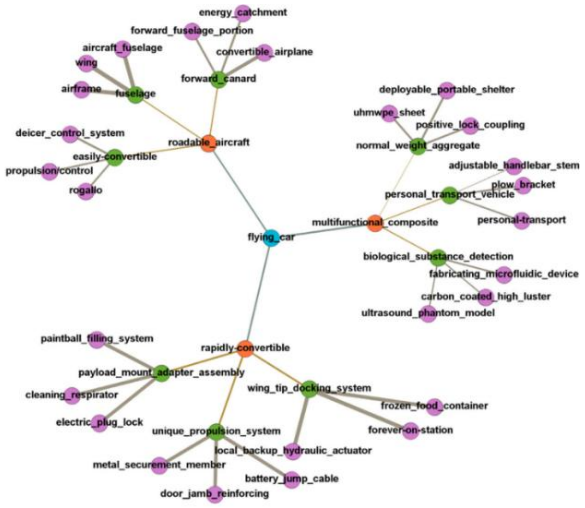
3) Bilgi Tabanlı Yöntemler (Knowledge Based Methods)

Anlamsal benzerlik analizinde anahtar kelimeler, konseptler arasındaki benzerliği belirlemek için kelimeler arası ilişkilerin tanımlandığı ontoloji yapılarının da önemli bir rolü vardır. Ontolojiler bir alandaki uzmanlar tarafından ilgili kelimeler veya kavramlar arasındaki eş anlamlı olma, kapsama, bağlı olma, parçası olma, sahip olma gibi ilişkileri tanımlayan şematik yapılardır. Özellikle anlamsal web alanındaki gelişmelere paralel olarak ontolojilerin önemi giderek artmaktadır. Ontolojiler sayesinde metinler de bilgisayarlar tarafından tanınabilen ve işlenebilen bir özellik kazanmaktadır [22].

En yaygın olarak kullanılan kelime ontolojilerinden birisi WordNet isimli kelimeler arası eş anlamlılık (synoym), alt anlamlılık (hyponym), zıt anlamlılık (antonym), meronim, troponim ilişkilerini tanımlayan ontolojidir. Literatürde ontoloji geliştirme yönündeki gayret de göze çarpmakta, patent verisi kullanılarak teknoloji alanında kullanılmak için geliştirilen bir semantik ontolojiden bahsedilmektedir [23]. Bu ontolojiden bir örnek **Şekil 10**'da paylaşılmıştır. Başka bir çalışmada ise WordNet ontolojisinin teknoloji alanında yapılan çalışmalarda yetersiz kaldığından bahsedilerek, patent dokümanları üzerinden yapılan ÖEO tabanlı analizlerle bir teknolojik anlamsal ağ oluşturulmuştur. Bu anlamsal ağın patent

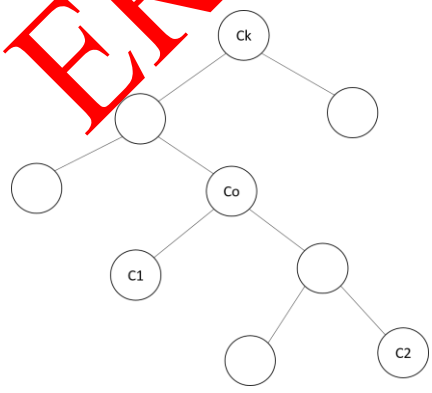
dokümanlarının benzerlik hesaplamalarında da faydalı olacağı belirtilmiştir [24].

Literatürde ontolojilerdeki unsurlar arasındaki anlamsal ilgililik (semantik relatedness) ve anlamsal benzerlik (semantik similarity) kavramlarının farklılıklarına değinilmiş olup, anlamsal ilgililik ölçümlerinde öğeler arasındaki meronim, eş anlamlılık, fonksiyonellik, aitlik, kapsama gibi çeşitli ilişkiler değerlendirilirken, anlamsal benzerlik ölçümlerinde sadece eş anlamlılık ve kapsama ilişkileri hesaba katılmaktadır. Bu bağlamda anlamsal benzerlik ölçümlerinin, anlamsal ilişki ölçümlerinin özel bir şekli olduğu söylenmiştir. Anlamsal uzaklık kavramı ise benzerlik kavramının zıttı olarak düşünülmektedir [25].



Şekil 10. Bir teknolojik ontolojide "uçan araba" konsepti (The "flying car" concept in a technological ontology)

Bir ontolojideki bu ölçütlerin hesaplanmasında esas olarak üç yöntem bulunmaktadır: ağ yapısındaki uzaklığa dayalı yöntemler, özellik tabanlı yöntemler ve bilgi içeriğine (information content) dayalı yöntemler. **Şekil 11**'de gösterildiği gibi bir ontolojide bu yöntemlere göre benzerlik hesaplamaları aşağıda açıklanmıştır.



Şekil 11. Bir anlamsal ontoloji örneği (An example of semantic ontology)

1- Ağ Yapısındaki Uzaklığa Dayalı Yöntemler (Kenar Sayma Yöntemi): İki öğe arasındaki en kısa yolda kaç tane kenar olduğunun sayıldığı benzerlik tespit yöntemidir. Rada formülü olarak bilinen bu yöntemde en yakın ortak ataları Co olan C1 ve C2 konseptleri arasındaki uzaklık şöyle hesaplanır:

$$d_{Rada}(C1, C2) = |path(C1, Co)| + |path(C2, Co)| \quad (14)$$

Bu yöntemde tüm kenarların ağırlığı birbirine eşit alındığı takdirde örneğin ontolojinin kök kısmında olup da birbirine 2 kenardan sonra ulaşılan unsurların yakınlığı ile ontolojinin en alttaki dallarında olup da aralarında 2 kenarlık bir uzaklık olan unsurların anlamsal yakınlığı aynı gibi gözükmektedir. Halbuki ontolojinin kök kısmına yaklaştıkça dallar arasındaki anlamsal farklılık giderek artar. Bu dezavantajı gidermek için kenarlara farklı ağırlıklar atanmaktadır: Kök kısmına yaklaştıkça kenarların ağırlığı büyür, en alt dala yaklaştıkça ağırlık küçülür. Bu mantığa göre geliştirilen Wu-Palmer anlamsal benzerlik ölçütünde benzerliği hesaplanmak istenen iki unsurun en küçük ortak atası olan diğer unsurun ontolojideki derinliği de hesaba katılmaktadır [25]. Bu formüle göre hiyerarşinin en üstünde yer alan konsept Ck ise, en yakın ortak ataları Co olan C1 ve C2 konseptleri arasındaki uzaklık şöyle hesaplanır:

$$d_{Wu-Palmer}(C1, C2) = \frac{2*|path(Ck,Co)|}{|path(C1,Co)|+|path(C2,Co)|+2*|path(Ck,Co)|} \quad (15)$$

2- Özelliğe Dayalı Ölçütler: Kenar sayma yönteminde konseptler arasındaki her bağlantının aynı uzaklığa sahip olduğu varsayımının getirdiği kısıtlamaları aşmak için özelliğe dayalı ölçütler kullanılmaktadır. Ontolojideki her konsepti sahip olduğu özelliklere göre ele alan bu ölçütlerde, konseptler arasındaki benzerlik, konseptlerin sahip olduğu ortak özelliklerin ve ortak olmayan özelliklerin bir fonksiyonu olarak hesaplanır. İki konsept ne kadar çok ortak özelliğe ve ne kadar az ortak olmayan özelliğe sahipse o kadar benzer oldukları sonucuna ulaşılır [22]. Buna göre C1 konseptinin sahip olduğu özelliklerden oluşan küme A kümesi, C2 konseptinin sahip olduğu özelliklerden oluşan küme B kümesi ise C1 ve C2 konseptlerinin benzerliği şöyle hesaplanır [26]:

$$s(C1, C2) = \frac{|A \cap B|}{|A \cap B| + \gamma(C1,C2) * |A \setminus B| + (1-\gamma(C1,C2)) * |B \setminus A|'} \quad (16)$$

$$\gamma(C1, C1) = \begin{cases} \frac{depth(C1)}{depth(C1)+depth(C2)}, & depth(C1) \le depth(C2) \\ 1 - \frac{depth(C1)}{depth(C1)+depth(C2)}, & depth(C1) > depth(C2) \end{cases} \quad (17)$$

3- Bilgi İçeriğine Dayalı Yöntemler: Bu yöntemde iki öğe ne kadar fazla bilgi paylaşıyorsa o kadar benzer oldukları kabul edilir. Benzerliği hesaplanmak istenen öğelerin (C1, C2) en yakın ortak ataları olan öğe (Co) belirlenerek, bu öğrenin sahip olduğu bilgi içeriği (information content) dikkate alınır. Buna göre geliştirilen Resnik benzerlik formülü şöyledir:

$$s_{Resnik}(C1, C2) = IC(Co) = -\log\big(p(Co)\big) \quad (18)$$

Burada p(Co) ise Co unsuruna bir derlemde rastlanma olasılığını gösterir. Yani Co unsurunun bir derlemdeki sıklığı arttıkça, C1 ve C2 unsurları birbirlerine daha az benzer olacaktır. Bu yaklaşımın bir dezavantajı aynı ortak ataya sahip öğelerden daha alt dallarda olanlarla ortak ataya daha yakın olanların aynı derecede benzer olduğu sonucuna ulaşılmasıdır. Bu dezavantajı ortadan kaldırmak için bu iki yöntem bir arada kullanılarak kenar sayısı hesabı yönteminin bir düzeltici faktörü olarak bilgi içeriği verisi de ölçüme dâhil edilmektedir [25]. Bu manada geliştirilen Jiang-Conrath Uzaklık ve Lin Benzerlik formülleri şöyledir:

$$d_{Jiang-Conrath}(C1, C2) = IC(C1) + IC(C2) - 2 * IC(Co) \quad (19)$$

$$s_{Lin}(C1, C2) = \frac{2*IC(Co)}{IC(C1)+IC(C2)} \quad (20)$$

Ontolojilerin bilgi keşfinde ve metinlerin anlamsal analizlerinde sahip oldukları potansiyelden patent dokümanlarının anlamsal benzerlik tespitinde de faydalanılmaktadır. Bazı çalışmalarda patent dokümanlarını oluşturan teknik konseptlerin belirlenmesinde ontolojilerden faydalanılabileceği belirtilmiştir [27]. Patent dokümanlarından çıkarılan ÖEO öğelerinin birbirine ne kadar benzediklerini tespit edebilmek için ontoloji kullanılması mümkündür [28]. Bir çalışmada patent istemleri analiz edilerek isim formundaki sözcükler belirlenmiş, bir istemdeki her ismin diğer istemdeki isimlerle ne kadar benzer olduğu Wordnet ontolojisi kullanılarak belirlenmiştir. Tüm isimler için yapılan bu işlemin neticesinde istemler arasındaki benzerlik ortaya konmuştur [27].

Sharma vd. [28] patentlerin benzerliğinde sadece doküman içerisinde geçen kelimelere göre sorgulama yapıldığında, sorgu kelimelerinin değil de eş anlamlılarının yer aldığı patent dokümanlarının gözden kaçabileceğini söylemiş ve bu problemi aşmak için sorgunun eş anlamlı kelimeleri de içerecek şekilde genişletilmesi gerektiğini belirtmiştir. Bunun için Wordnet ontolojisi kullanılarak, bir patentin özet kısmından çıkarılan kelimelerin eş anlamlıları ve yakın anlamlıları da sorguya dâhil edildiğinde benzerlik tespitinde daha yüksek bir verim elde edilmiştir. Sharma vd. [29] bir başka çalışmalarında ise sadece Wordnet ontolojisi yerine Wordnet ontolojisi ile Wiktionary isimli ontolojinin birlikte kullanıldığı takdirde verimin daha da arttığını vurgulamıştır.

B. DERİN ÖĞRENME TABANLI YÖNTEMLER İLE PATENTLERİN BENZERLİK TESPİTİ (SIMILARITY DETECTION OF PATENTS WITH DEEP LEARNING BASED METHODS)

Önceki bölümde tarif edilen yöntemler her ne kadar önemli sonuçlar ortaya koysa da patent metinlerinin karmaşık yapısı ve terminolojisi ile başa çıkmakta yeterince başarılı olamamaktadır. Bu yöntemlerin, yapay zekâ teknikleri ile bir arada kullanılması durumunda daha etkin sonuçlar elde edilebilmektedir. [30] Patent madenciliğinde yapay zeka tekniklerinin kullanımına dikkat çekilen bir araştırmada literatürde bu alanda yapılan 57 adet çalışma incelenmiş ve çalışmaların bilgi yönetimi, teknoloji yönetimi, ekonomik değer ve bilgi çıkarımı alanlarında yoğunlaştığı tespit edilmiştir. [31] Patentlerin benzerlik araştırmasında da yapay zeka tekniklerinin kullanımı araştırma kalitesini yükseltmektedir. [32]

Setchi vd. patent tekniğin bilinen durumu araştırmasının zorluklarından bahsetmiş ve bu zorlukların aşılmasında yapay zeka tekniklerinin önemine değinmiştir. Yapay zekanın patent araştırma işlemlerine dahil edilmesi ile buluşun patent sınıfının belirlenmesi noktasında büyük başarı elde edilmiş, ilgili patent dokümanlarına ulaşma süreci hızlanmıştır. Ancak patent araştırma sorgusunun doğru belirlenmesi noktasında yapay zekâ teknikleri de yetersiz kalmıştır. Sonuç olarak yapay zeka teknikleri de kullanılsa tekniğin bilinen durumu araştırmasını uzman görüşüne ihtiyaç duymadan tüm yönlerini ile gerçekleştirebilecek ve patentlenebilirliğe karar verebilecek nitelikte bir model ortaya koymanın kolay olmadığı ifade edilmiştir. [33] Benzer amaçla yapılan bir çalışmada da geleneksel patent arama motorları ile yapay zeka destekli araçlar kullanılarak tekniğin bilinen durumu araştırması yapılmış ve sonuçlar karşılaştırılmıştır. Sonuçlara göre yapay zekâ destekli araçlar büyük bir potansiyel ortaya koymuş olsalar da patent araştırmasının sadece yapay zeka destekli araçlar tarafından yapılabilmesi henüz mümkün gözükmemektedir. [34] Bu alanda yapılacak çalışmaların geleceği oldukça parlaktır. [35]

Bir başka çalışmada da patentlenebilirlik kriterlerinden buluş basamağı kriterine odaklanılmış ve buluş basamağı değerlendirmesinin çok yönlü ve incelikleri olan bir değerlendirme olduğuna değinilerek yapay zeka yöntemlerinin bile buluşçu bir yaklaşım ortaya koymakta yetersiz kaldığı ifade edilmiştir. [36]

Bu noktada çok katmanlı yapay sinir ağlarından oluşan ve makine öğrenmesinin özel bir türü olan derin öğrenme algoritmaları ve yöntemlerinin, veri analizinde şaşırtıcı derecede etkili sonuçlar ortaya koyduğu bilinmektedir. Son dönemde derin öğrenme tekniklerinde yaşanan gelişmeler, patent madenciliğine de yeni bir heyecan getirmiş, patent analizlerinde kullanılan derin öğrenme yöntemleri ile çeşitli yönlerden tatmin edici sonuçlara ulaşılmaya başlanmıştır. CNN (convolutional neural network), RNN ((simple) recurrent neural network,) LSTM (long short term memory network), GRU (gated recurrent unit network), SEQ2SEQ (sequence-to-sequence network) GAN (generative adversarial network) AE (autoencoder network), TRANS (transformer-based network) gibi derin öğrenme mimarilerinin hepsinden patent analizlerinde çeşitli amaçlarla faydalanılmaktadır [37].

Patent analizleri ile gerçekleştirmek istenilen tüm görevlerde etkin bir şekilde kullanılmakta olan derin öğrenme, benzer patentlere erişilmesi ve dolayısıyla patentlenebilirlik hakkında fikir edinilebilmesi için

önemli fırsatlar sunmaktadır. Literatürde patent dokümanlarının teknik literatür açısından karmaşık yapısından dolayı klasik metin madenciliği yöntemlerinde karşılaşılan zorlukların derin öğrenme ile aşılabileceği belirtilmiş, çalışmalar bu alana yoğunlaşmıştır.

Literatürde son dönemlerde çok yaygın olarak metin analizlerinde kullanılan ve derin öğrenme tabanlı olarak çalışan Word2vec isimli yöntem ve bu yöntemin benzerleri patent dokümanlarında yaygın olarak kullanılmaktadır. Kelime temsili (word embedding) olarak da adlandırılan bu yöntemde her bir kelime yapay sinir ağlarından faydalanılmak suretiyle vektörleştirilmektedir. Bu yöntemde sığ, iki katmanlı bir sinir ağı modeli büyük bir metin kümesi ile eğitilmekte ve bu ağ ile her bir kelime için benzersiz bir vektör oluşturulmaktadır. Word2Vec yönteminde sürekli kelime çantası (CBoW) ve atla-gram (Skip-gram) şeklinde 2 farklı yaklaşım vardır. CBoW bir kelimenin çevresinde yer alan kelimelerden yola çıkarak ortadaki kelimenin tahmin edilmesine dayanırken, Skip-gram ise ortadaki bir kelimeden yola çıkarak çevredeki kelimeleri tahmin eder. Word2Vec yöntemi ile oluşturulan kelime vektörlerinden, cümle, paragraf ve doküman vektörlerine ulaşılmakta, ardından Kelime Taşıyıcının Uzaklığı (KTU) gibi yöntemlerle metinler arasındaki benzerlik hesaplanmaktadır.

Word2vec yöntemi ve Word2vec yönteminin bir uzantısı olan Doc2vec yöntemi de patentlerin anlamsal benzerliğinin tespit edilmesinde geniş bir kullanım alanı bulmuştur [38]. Bir çalışmada patent dokümanının bölümleri olan tarifname, özet ve istemler kısımları ayrı ayrı benzerlik analizine tabi tutularak; TF-TBF, Gizli Anlamsal Analiz, Word2vec ile kombine edilmiş Kelime Çantası ve Doc2vec yöntemlerinin performansı karşılaştırılmıştır. Tarifname üzerinde TF-TBF, özet veya istemler üzerinde ise Doc2vec yöntemi daha yüksek performans göstermiştir [39]. Başka bir çalışmada da patent dokümanlarını oluşturan ÖEÖ yapıları tespit edilerek, bu yapılar Doc2vec yöntemi ile vektörleştirilmiş ve oluşan bu vektörlere göre patentlerin benzerliği hesaplanmıştır. Sao2vec ismi verilen bu yöntemin patent dokümanları üzerinde klasik Doc2vec yöntemine göre %3, ÖEÖ yapılarının frekanslarının hesaplanmasına dayalı yönteme göre %115 oranında daha yüksek performans sergilediği ortaya konmuştur [40]. Yine Doc2vec yöntemi ile patent dokümanlarının vektörleştirildiği bir makalede Local Outlief Factor yöntemi ile patentlenebilirlik kriterlerinden yenilik kriterinin tespiti üzerinde çalışılmıştır [41]. Ancak buluş basamağı kriteri hakkında bir yorum yapılmamıştır.

Aras vd. ise patent dokümanları arasındaki benzerliği tespit etmek için Word2vec modeli ile kelime vektörlerini elde etmiş, dokümanlardaki kelime vektörlerinin ortalamasını alarak da doküman vektörlerine ulaşmış ve kosinüs benzerliği ile benzerlik oranını hesaplamıştır. Bu noktada Word2vec modeli için iki alternatif kullanmış olup birincisi Google tarafından eğitilmiş olan genel Word2vec modeli iken diğeri ise patent dokümanları üzerinde eğitilmiş olan özel bir Word2vec modelidir. Bu iki modelden patent dokümanları üzerinde eğitilmiş olan model ile yapılan benzerlik ölçümlerinden çok daha yüksek bir başarı elde edilmiştir [42]. Patent dokümanlarında kullanılan literatürün kendine has olduğu göz önünde bulundurulduğunda ortaya çıkan bu sonuç oldukça normaldir.

Bir başka çalışmada da patent araştırmasında çok büyük önemi olan anahtar kelimelerin eş anlamlı ve yakın anlamlılarının tespit edilmesi için patent dokümanları üzerinde derin öğrenme algoritmaları olan FastText ve Word2Vec yöntemlerinden faydalanılmıştır. Ortaya çıkan sonuçlar patent uzmanlarının kontrolüne sunulmuş ve FastText yönteminin daha yüksek bir performans ortaya koyduğu ifade edilmiştir. Ayrıca çalışma neticesinde eşanlamlı ve yakın anlamlı kelimelerden oluşan bir veri seti elde edilmiştir. [7]

Kelime temsili olan Word2vec yönteminin gelişmişi olarak düşünülebilecek içerik temsil (context embedding) yöntemleri de patent analizinde kullanılmaktadır. Bir çalışmada patent metinlerinde çalıştırılan Word2vec yöntemi ile elde edilen vektörler LSTM yapısı ile içerik bilgilerine sahip olacak şekilde geliştirilmekte ve içerik temsil vektörleri oluşturulmaktadır. Bu vektörler CNN aracılığıyla işlenmekte ve patent metin vektörleri elde edilmektedir. Ardından bu vektörler arasında yapılan karşılaştırmalar ile patentlerin teknolojik benzerliğine karar verilmektedir. Vektörlerin karşılaştırılmasında çok katmanlı algılayıcı (MLP) ve Softmax sınıflandırıcı kullanılmıştır. Neticede kullanılan model ile başka benzerlik tespit yöntemleri karşılaştırılarak ortaya konan modelin daha iyi sonuç verdiği belirtilmiştir. Bu karşılaştırmalara göre klasik TF-TBF+Cosinüs benzerliği yöntemi ile yapılan benzerlik tespitinin doğruluk oranı %69 iken CNN+MLP yönteminin doğruluk oranı ise %94 olmuştur [43].

Jaeyoung vd. ise yapmış oldukları çalışmada patent dokümanlarının özet kısımlarında Doc2vec yöntemi ile çalışarak doküman vektörleri oluşturduktan sonra AutoEncoder tabanlı bir yöntem olan DEC(Deep Embedding Clustering) yöntemi ile vektörleri iyileştirmişler ve kümeleme işlemi yapmışlardır. Ortaya konan Doc2vec + DEC yönteminin doğruluğu TF-TBF +K-means, TF-TBF+GMM (Gaussian Mixture Modeling), Kelime Çantası + K-means, Kelima Çantası + GMM, Doc2vec+K-means, Doc2vec+GMM yöntemleri ile karşılaştırılmış ve %97.61 ile en yüksek oranda doğruluğa ulaşılmıştır. Doc2vec+Dec yönteminin ardından Doc2vec+GMM yöntemi de %97.55 ile yüksek bir performans gösterirken, Kelime Çantası+K-means yönteminin doğruluğu %55.47'de kalmıştır [44]. Bu çalışma da derin öğrenme yöntemlerinin patent madenciliğinde ne kadar verimli olduğunu ortaya koymaktadır.

Bir başka çalışmada öncelikle patent benzerliğinde önceki bölümde açıklanan vektör uzay modeli yaklaşımıyla patent dokümanlarının anahtar kelimelerden oluşan vektörler olarak modellenmesinin

dezavantajlarına değinilmiştir. Kelime çantası ve TF-TBF benzeri işlemlerle oluşturulan kelime matrislerinin hem çok büyük olması sebebiyle işlem yapılmasının zorlaştığından hem de kelimelere odaklı olunduğu için cümle anlamının kaybolduğundan ve dolayısıyla ortaya çıkan vektörlerin patentlerin içeriğini doğru yansıtamadığından bahsedilmiştir. Bu dezavantajları aşmak için anahtar kelimelerden değil de özellik çıkarımı (feature extraction) yapılarak elde edilen özellik matrislerinden oluşan vektörlerle benzerlik değerlendirilmesi yapılması önerilmiştir. Niteliklerin çıkarımı için de CNN kullanılmış ve resimlerden nitelik vektörleri çıkarımı yapılmasında yaygın olarak kullanılan CNN yapısının, metin dokümanlarından nitelik çıkarımı yapılmasında etkili olduğu ifade edilmiştir. Nitekim patentler arasındaki benzerlik ölçümlerinde, CNN tabanlı olarak oluşturulan doküman vektörleri ile yapılan işlemlerde doğruluk oranı %91 iken, klasik TF-TBF matrisleri yoluyla yapılan işlemlerde doğruluk oranı %82 olmuştur [45].

Hibrit bir yöntem uygulanan bir başka makalede de değerli patentlerin erken tespiti için patent dokümanlarının özet ve istem bölümleri CNN ve RNN modellerinin bir arada kullanımı ile sınıflandırılmıştır. Makale yazarları sadece CNN veya sadece RNN kullanımı ile ortaya çıkan dezavantajları, iki modelin bir arada kullanımı ile ortadan kaldırmaya çalışmışlardır. RNN modeli olarak bi-LSTM seçilmiş ve CNN+bi-LSTM modeli ile yapılan sınıflandırma işlemlerinde yüksek bir performans elde edilmiştir [46].

Başka bir çalışmada patent metinleri içerisinde anlamsal ilişki ve varlık çıkarımı yapabilmek için 1010 patent özeti üzerinde çalışılmıştır. Varlık belirleme amacıyla BiLSTM+CRF derin öğrenme modeli, anlamsal ilişki çıkarımı yapabilmek için BiGRU+HAN derin öğrenme modeli kullanılmıştır. RNN türevleri olan LSTM ve GRU modellerinin metinlerdeki kelimelerin dilbilgisi etiketini tahmin etmek için kullanılabilen CRF ve belge sınıflandırma gibi görevlerde etkin rol oynayan HAN yapıları ile hibrit kullanımı sonucu makalede ortaya konan yöntem, önceki yöntemlerden daha başarılı sonuçlar üreterek %45.8 kesinlik (precision), %58.8 duyarlılık (recall) ve %51.5 F-skor değerlerine ulaşmıştır. [47]

Tekniğin bilinen durumu araştırması için patent atıf önerisi yapmayı amaçlayan bir makalede de 110.000 patentin metin verisi ve meta verisinden oluşan bir veri seti üzerinde analiz yapmak için iki aşamalı bir derin öğrenme yapısı kurulmuştur. CSNet modeli ile aday dokümanlar seçilmiş ve CRNet modeli ile seçilen dokümanlar ilgililik açısından sıralanmıştır. Ortaya konan model 0.2506 puanlık MRR skoru ile bilinen yöntemlerden daha iyi sonuç vermiştir. [48]

Ayrıca son dönemde gelişmekte olan ve doğal dil işleme alanında güzel sonuçlar ortaya koyan Transformer Sinir Ağları olarak adlandırılan derin öğrenme yapılarından olan GPT, GPT-2 ve BERT isimli mimarilerden faydalanarak patent metinlerini analiz eden çalışmalar da literatürde mevcuttur. Uzun metin verilerini anlama ve analiz etme noktasında diğer derin öğrenme ağlarına göre yüksek bir başarı ortaya koyan Transformer sinir ağlarının patent metinlerini işlemek için kullanılması önemli sonuçlar doğuracaktır. [49]

Bir çalışmada en benzer patent dokümanın tespitine odaklanılarak öncelikle benzerlik araştırması yapılan patent metinleri kullanılarak GPT-2 modeli eğitilmiştir. Benzerlik tespiti yapmak için GPT-2 modelinin ürettiği patent metinleri incelenmiş ve BM25 yöntemi ile üretilen metinler benzerlik sıralamasına tabi tutulmuştur. Ardından BERT modeli ile metinler tekrar sıralanmış ve en yakın patent dokümanı tespit edilmeye çalışılmıştır. Transformer tabanlı bu yöntem oldukça güzel sonuçlar verse de makale yazarları uzun patent metinleri üzerinde yaşanan zorluklardan bahsetmişlerdir. [50] Benzer şekilde BM25 ve BERT modelinin iki aşamalı olarak kullanıldığı bir başka çalışmada da patent tekniğin bilinen durumu araştırmasında güzel sonuçlar elde edilmiştir. [51]

Patent metinlerini sınıflandırma üzerinde çalışılan bir makalede, patent metinlerinden özellik çıkarımının zorluklarına değinilmiş ve bir sözlük ağı oluşturularak bu zorluklar aşılmaya çalışılmıştır. Word2Vec, CNN, LSTM ve BERT modelleri ile sözlük ağı bir arada kullanılmış ve en başarılı sonuçlar BERT modeli ile elde edilmiştir. [52]

Patent haritalaması yaparken patent sınıf bilgisinin otomatik olarak tespit edilebilmesi için bir model geliştirilen bir çalışmada da patent özetlerinden Transformer mimarisi ile oluşturulan metin temsilleri (text embedding) ve patent sınıf kodlarından Grafik Sinir Ağları ile oluşturan grafik temsilleri (graph embedding) birlikte kullanılmıştır. Ortaya konan modelin performansı, en yüksek sonuç verdiği bilinen BERT tabanlı mevcut modeli geride bırakmıştır. [53].

Transformer derin öğrenme ağlarının sıralı olmayan verileri işleyebilme, kendi içinde dikkat (self attention) kabiliyeti gibi özellikleri sebebiyle metinlerin anlamsal benzerliğini tespit noktasında RNN ve LSTM yöntemlerine göre daha başarılı sonuçlar verdiği söylenen bir çalışmada da yazarlar patent metinlerinin benzerliğini tespit etmek için DeBERT modelinin varyantlarını kullanmışlardır. En yüksek performansa DeBERTa-Small modeli ile ulaşılmıştır. [54]

Karmaşık bir problem olan patentlenebilirlik problemine odaklanılan bir çalışmada da Amerikan Patent Ofisi verilerine odaklanılarak patent istemlerinin Amerikan patent kanununa göre patentlenemeyeceğini gösteren durumlar çok etiketli bir metin sınıflandırma problemi olarak ele alınmıştır. Patent istemleri transformer tabanlı bir yöntemler olan önceden eğitilmiş BERT-Base/Large, RoBERTa-Base/Large, XLNet modelleri ile analiz edilmiş ve mikro-duyarlılık, mikro-kesinlik, mikro-F1 skor ölçümlerine göre en yükse sonuç Roberta-Large modeli ile elde edilmiştir. Makale yazarlar geliştirdikleri yöntem ile mesafe kat etmiş olsalar da patentlenebilirlik probleminin kompleks yapısı sebebiyle elde edilen sonuçların henüz yeterli olmadığı belirtmişlerdir. [55]

5. SONUÇ (CONCLUSION)

Önceki bölümlerde detaylı şekilde anlatıldığı üzere patentlerin benzerlik tespiti üzerinde çok çeşitli çalışmalar yapılmış olsa da bunlar patentlenebilirlik tespiti açısından hedeflenen noktada değildir. Patentlenebilirlik tespiti gibi patent tescil sürecinin kilit noktası denilebilecek bir işlemi etkin ve verimli şekilde yapabilecek yeni yaklaşımlara büyük ihtiyaç duyulmaktadır.

Bu alanda yapılacak çalışmalarda hedeflenen nokta dökümanlara X, Y ve A kategorilerinin uygulanacak model neticesinde otomatik olarak atanabilmesi olmalıdır. Literatürde patentlerin benzerlik oranları ölçülmüş olsa da bu kategorizasyonu yapabilen bir yöntemden bahsedilmemiştir. Literatürdeki bu eksikliğin önemli bir sebebi, patent alanında yapılacak bu tür çalışmaların ciddi bir uzmanlık bilgisi gerektiriyor olmasıdır. Bahsedilen bu kodların dokümanlara atanabilmesi için patentlenebilirlik kriterleri olan yenilik ve buluş basamağı gibi patent alanında tecrübe gerektiren hususlara hâkim olunması gerekmektedir.

Bununla birlikte teknoloji yönetimi alanında kilit bir role sahip olan patent araştırma raporlarını otomatik olarak hazırlayacak böyle bir modelin geliştirilmesi öncelikle Ar-Ge ekosistemine büyük katkı sağlayacaktır. Buluş sahipleri geliştirdikleri buluşlarının patent alıp alamayacağını gösteren bir ön araştırma raporunu daha tescil işlemleri başlamadan görebilmiş olacaklardır. Hedeflenen modelin ortaya koyduğu raporda X ve Y kategorisinde dokümanlara rastlayan bir buluş sahibi, buluşunda geliştirmeler yapması gerektiğini fark edecek ve tescil sürecine daha nitelikli bir buluş ile başlayacaktır. Her ne kadar buluş sahipleri online patent arama motorlarından tekniğin bilinen durumu araştırması yapabiliyor olsalar da patent literatürü ve veri tabanlarında uzmanlık bilgisi olmadan X ve Y kategorisinden dokümanlara ulaşabilmek kolay olmamaktadır. Patentlenebilirlik tespiti alanında geliştirilebilecek nitelikli bir modelin bu zorluğun aşılmasında büyük yardımı olacaktır. Bundan başka patent tescil sürecinin en çok vakit alan ve en çok çaba gerektiren safhası olan patent araştırma raporunun hazırlanması işleminde büyük bir kolaylık ortaya çıkacaktır. Patent ofislerinde patent araştırma raporunun hazırlanması için uzun müddet beklemek durumunda kalan dosyaların işlemleri hızlanacak, hem yığılan iş yükü hafiflemiş olacak hem de yıllarca süren patent tescil süreci daha çabuk neticelenecektir.

ETİK STANDARTLARIN BEYANI (DECLARATION OF ETHICAL STANDARDS)

Bu makalenin yazarları çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler.

YAZARLARIN KATKILARI (AUTHORS' CONTRIBUTIONS)

Ahmet KAYAKÖKÜ: Literatür taraması yapmış ve makaleyi yazmıştır. / Performed literature search and wrote the manuscript.

Aslıhan TÜFEKCİ: Makaleyi organize etmiş ve gözden geçirmiştir. / Organize and review the article.

ÇIKAR ÇATIŞMASI (CONFLICT OF INTEREST)

Bu çalışmada herhangi bir çıkar çatışması yoktur. / There is no conflict of interest in this study.

KAYNAKLAR (REFERENCES)

[1] Bonino D., Ciaramella A., and Corno F., "Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics", ***World Patent Information*** , 32(1): 30-38, (2010).

[2] Schwander P. "An evaluation of patent searching resources: comparing the professional and free on-line databases", ***World Patent Information***, 22: 147-165, (2000).

[3] Madani F. and Weber C., "The evolution of patent mining: Applying bibliometrics analysis and keyword network analysis", ***World Patent Information***, 46: 32-48, (2016).

[4] Kayakökü A. and Akay D., "Patent Madenciliği", ***Journal of Polytechnic***, 24(2): 745-753, (2021).

[5] Kayakökü A. and Demirbaş Ş., "Patent Arama Motorlarının Kullanımı Üzerine Bir İnceleme", ***Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarim Ve Teknoloji***, 5(3):149-165, (2017).

[6] Chandrasekaran D. and Mago V., "Evolution of Semantic Similarity--A Survey", ***ACM Computing Surveys***, 54(2): 1-37, (2021).

[7] Krishna A.M., Jin Y., Foster C., Gabel G., Hanley B. and Youssef A., "Query Expansion for Patent Searching using Word Embedding and Professional Crowdsourcing", ***ArXiv***, (2019).

[8] Walter L. Radauer A. and Moehrle M.G., "The beauty of brimstone butterfly, novelty of patents identified by near environment analysis based on text mining", ***Scientometrics***, 111: 103-115, (2017).

[9] Arts S., Cassiman B., and Gomez J.C., "Text matching to measure patent similarity", ***Strategic Manage J***, 39(1): 62-84, (2018).

[10] Moehrle M.G., "Measures for textual patent similarities: A guided way to select appropriate approaches", ***Scientometrics***, 85(1): 95-109, (2010).

[11] Moehrle M.G. and Gerken J.M., "Measuring textual patent similarity on the basis of combined concepts: Design decisions and their consequences", ***Scientometrics***, 91(3): 805-826, (2012).

[12] An X., Li J., Xu S., Chen L. and Sun W., "An improved patent similarity measurement based on entities and semantic relations", ***Journal Informetrics***, 15,(2): 101-135, (2021).

[13] Wang X.F., Ren H.C., Chen Y., Liu Y.Q., Qiao Y.L., and Huang Y., "Measuring patent similarity with SAO semantic analysis", ***Scientometrics***, 121(1): 1-23, (2019)

[14] Bergmann I., Butzke D., Walter L., Fuerste J.P., Moehrle M.G., and Erdmann V.A., “Evaluating the risk of patent infringement by means of semantic patent analysis: the case of DNA chips”, ***R&D Management***, 38(5): 550-562, (2008).
[15] Ontañón S., “An overview of distance and similarity functions for structured data”, ***Artificial Intelligence Review***, 53(7): 5309-5351, (2020).
[16] Batyrshin I., Cross V., Kreinovich V., and Rifqi M., “Towards a general theory of similarity and association measures: Similarity, dissimilarity and correlation functions”, ***Journal of Intelligent & Fuzzy Systems***, 36(4): 2977-3004, (2019).
[17] Park H., Yoon J., and Kim K., “Identifying patent infringement using SAO based semantic technological similarities”, ***Scientometrics***, 90(2): 515-529, (2012).
[18] Shahmirzadi O., Lugowski A., and Younge K., “Text Similarity in Vector Space Models: A Comparative Study”, ***Book Text Similarity in Vector Space Models: A Comparative Study***, (2019) .
[19] Wang J. and Chen Y.J., “A novelty detection patent mining approach for analyzing technological opportunities”, ***Advanced Engineering Informatics***, 42: 100941, (2019).
[20] Moehrle M.G., “Similarity measurement in times of topic modelling”, ***World Patent Information***, 59: 101934, (2019).
[21] Cvitanić T., Lee B., Song H.I., Fu K.K. and Rosen D.W., "LDA v. LSA: A Comparison of Two Computational Text Analysis Tools for the Functional Categorization of Patents", ***ICCBR Workshops,*** (2016).
[22] Sánchez D., Batet M., Isern D. and Valls A., “Ontology-based semantic similarity: A new feature-based approach”, ***Expert Systems with Applications***, 39(9): 7718-7728, (2012).
[23] Sarica S., Luo J. and Wood K.L., “TechNet: Technology semantic network based on patent data”, ***Expert Systems with Applications***, 142: 112995, (2020).
[24] Jang H., Jeong Y., and Yoon B., “TechWord: Development of a technology lexical database for structuring textual technology information based on natural language processing”, ***Expert Systems with Applications***, 164, (2021).
[25] Cross V. and Youbo W., “Semantic Relatedness Measures in Ontologies Using Information Content and Fuzzy Set Theory”, ***The 14th IEEE International Conference on Fuzzy Systems***, (2005) .
[26] Gan M., Dou X. and Jiang R., “From ontology to semantic similarity: Calculation of ontology-based semantic similarity”, ***The Scientific Word Journal***, (2013).
[27] AIndukuri K.V., Ambekar A.A. and Sureka A., “Similarity Analysis of Patent Claims Using Natural Language Processing Techniques”, ***International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)***, 169-175, (2007).
[28] Sharma P., Tripathi R., Singh V.K. and Tripathi R.C., “Automated patents search through semantic similarity”, ***2015 International Conference on Computer, Communication and Control (IC4)***, Indore, 1-5, (2015).
[29] Sharma P., Tripathi R., and Tripathi R.C., “Finding Similar Patents through Semantic Expansion”, ***2016 International Conference on Computer Communication and Informatics***, India, (2016).
[30] Villa A.M. and Wirz M., “A sequential patent search approach combining semantics and artificial intelligence to identify initial State-of-the-Art documents”, ***World Patent Information,*** 68, (2022).
[31] Aristodemou L. and Tietze F., “The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data”, ***World Patent Information***, 55: 37-51, (2018).
[32] Genin B.L. and Zolkin D.S., “Similarity search in patents databases. The evaluations of the search quality”, ***World Patent Information***, 64, (2021).
[33] Setchi R., Spasić I., Morgan J., Harrison C., and Corken R., “Artificial intelligence for patent prior art searching”, ***World Patent Information***, 64, (2021).
[34] Hafner A., Damij N. and Modic D., “Augmented intelligence for state-of-the-art patent search”, ***2022 IEEE Technology and Engineering Management Conference***, Turkey, (2022).
[35] Vaish K., Rawat P., Kathuria S., Singh R., Joshi K. and Verma A., “Artificial Intelligence Reducing the Intricacies of Patent Prior Art Search”, ***2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions***, India, (2023).
[36] Schellekens M., “Artificial Intelligence and the re-imagination of inventive step”, ***Journal of Intellectual Property, Information Technology and E-Commerce Law***, 13(2): 89-98, (2022).
[37] Krestel R., Chikkamath R., Hewel C., and Risch J., “A survey on deep learning for patent analysis”, ***World Patent Information***, 65, (2021).
[38] Whalen R., Lungeanu A., Dechurch L. and Contractor N., “Patent Similarity Data and Innovation Metrics”, ***Journal of Empirical Legal Studies***, 17(3): 615-639, (2020).
[39] Helmers L., Horn F., Biegler F., Oppermann T., and Müller K.R., “Automating the search for a patent”s prior art with a full text similarity search”, ***Plos One***, 14(3), (2019).
[40] Kim S., Park I. and Yoon B., “SAO2Vec: Development of an algorithm for embedding the subject–action–object (SAO) structure using Doc2Vec”, ***Plos One***, , 15(2), (2020).
[41] Jeon D., Ahn J.M., Kim J. and Lee C., “A doc2vec and local outlier factor approach to measuring the novelty of patents”, ***Technol Forecast Soc***, 174, (2022).
[42] Aras H., Türker R., Geiss D., Milbradt M, and Sack H., “Get your hands dirty: Evaluating word2vec models for patent data”, ***Proceedings of the Posters and Demos Track of the International Conference on Semantic Systems (SEMPDF)***, 1-4, (2018).
[43] Lu Y., Xiong X., Zhang W., Liu J., and Zhao R., “Research on classification and similarity of patent citation based on deep learning”, ***Scientometrics***, 123(2): 813-839, 2020.
[44] Kim J., Yoon J., Park E., and Choi S., “Patent document clustering with deep embeddings”, ***Scientometrics***, 123(2): 563-577, (2020).
[45] Lei L., Q, J. and Zheng K., "Patent Analytics Based on Feature Vector Space Model: A Case of IoT" ***IEEE Access***, 7, 45705-45715. (2019).
[46] Chung P. and Sohn S.Y., “Early detection of valuable patents using a deep learning model: Case of semiconductor industry”, ***Technological Forecasting and Social Change***, 158: 120-146, (2020).
[47] Chen L., Xu S., Zhu L., Zhang J., Lei X., and Yang G., “A deep learning based method for extracting semantic

information from patent documents", ***Scientometrics***, 125(1): 289-312, (2020).
[48] Choi J., Lee J., Yoon J., Jang S., Kim J., and Choi S., "A two-stage deep learning-based system for patent citation recommendation", ***Scientometrics***, (2022).
[49] Lee J.-S., "PatentTransformer: A Framework for Personalized Patent Claim Generation", ***The 32nd International Conference on Legal Knowledge and Information Systems,*** Spain, (2019).
[50] Lee J.-S., and Hsiang J., "Prior Art Search and Reranking for Generated Patent Text", ***ArXiv***, (2020).
[51] Stamatis V., "End to End Neural Retrieval for Patent Prior Art Search", ***44th European Conference on IR Research,*** Norway, (2022).
[52] Li R., Yu W., Huang Q. and Liu Y., "Patent Text Classification based on Deep Learning and Vocabulary Network", ***International Journal of Advanced Computer Science and Applications***, 14(1), (2023).
[53] Choi S., Lee H., Park E. and Choi S., "Deep learning for patent landscaping using transformer and graph embedding", ***Technological Forecasting and Social Change***, 175: 121-413, (2022).
[54] Nemani P. and Vollala S., "A Cognitive Study on Semantic Similarity Analysis of Large Corpora: A Transformer-based Approach", ***2022 IEEE 19th India Council International Conference***, India, (2022).
[55] Lo H.-C. and Chu J.-M., "Pre-trained Transformer-based Classification for Automated Patentability Examination", ***2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering***, Australia, (2021).