Journal of the Faculty of Engineering and Architecture of Gazi University 38:3 (2023) 1879-1892

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University
Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Natural language processing for the Turkish Academic texts in the engineering field and development of a decision support system: the case of TUBITAK project proposals

Bora Kat*
The Scientific and Technological Research Council of Türkiye (TUBITAK), 06530, Çankaya, Ankara, Türkiye

Highlights:

- Key term retrieval from the academic engineering documents in Turkish
- Similarity algorithm that detects almost all of the revised project proposal pairs
- Supervised machine learning classification algorithm that predicts the subfields of the proposals

Keywords:

- Key term extraction
- Feature extraction
- Natural language processing
- Supervised machine learning
- Naïve Bayes classifier
- Conceptual similarity
- Decision support system

Article Info:

Research Article
Received:06.07.2022
Accepted: 07.09.2022

DOI:

10.17341/gazimmfd.1132053

Acknowledgement:

The author gratefully acknowledges TÜBİTAK for providing this opportunity and thanks Prof. Dr. Lale Özbakır for her comments.

Correspondence:

Author: Bora Kat
e-mail: bora.kat@tubitak.gov.tr
phone: +90 312 298 1231

Graphical/Tabular Abstract

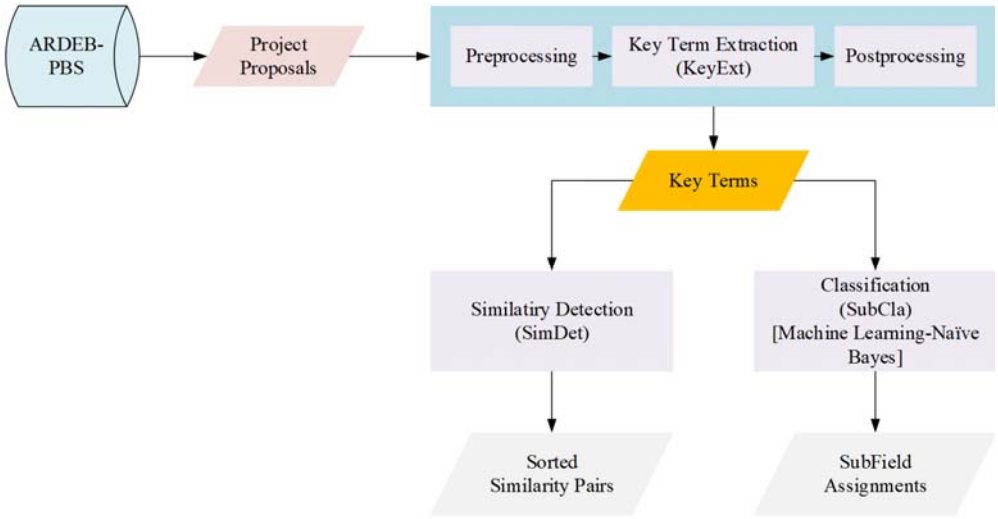


Figure A. Task flow of NLP and Machine Learning steps in the study

Purpose:
This study proposes a decision support system (as illustrated in Figure A) based on NLP applications and machine learning algorithm. Three modules (key term extraction, similarity detection and subfield assignment) are developed that would automatically index academic engineering documents, calculate their conceptual similarities and assign them to the most appropriate subfield over 31 subfields.

Theory and Methods:
Tailored preprocessing procedures are applied to the texts and the initial key terms are extracted. After a post-processing step, final versions of the term-frequency vectors are obtained. These vectors are used in the proposed similarity detection algorithm and as an input to the Naïve Bayes classifiers.

Results:
The proposals submitted to TUBITAK Academic Research Funding Program Directorate (ARDEB) are analyzed as a case study. The results indicate that the proposed similarity algorithm correctly detects almost all of the revised proposals while the accuracy of the Naïve Bayes classifier is more than 80% over a sample of 1255 proposals. The accuracy level exceeds 95% based on the best three predictions.

Conclusion:
NLP studies conducted in this study and the proposed algorithms are the first attempt to classify Turkish academic texts. Current study focuses on engineering; further studies on classifying other disciplines are needed. Moreover, the success of the machine learning in classification would pave the way for other applications such as reviewer identification.

Journal of the Faculty of Engineering and Architecture of Gazi University 38:3 (2023) 1879-1892

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Mühendislik alanındaki Türkçe akademik metinler için makine öğrenmesi destekli doğal dil işleme çalışmaları ve bir karar destek sisteminin geliştirilmesi: TÜBİTAK projeleri örneği

Bora Kat*
Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK), 06530, Çankaya, Ankara, Türkiye

ÖNEÇIKANLAR

- Mühendislik alanındaki Türkçe akademik metinlerden anahtar terimlerin çıkarımı
- Revize projelerin tamamına yakınını tespit eden metinsel benzerlik algoritması
- Bir metnin hangi alana ait olduğunu tahmin eden gözetimli makine öğrenmesi algoritması

Makale Bilgileri
Araştırma Makalesi
Geliş: 06.07.2022
Kabul: 07.09.2022

DOI:
10.17341/gazimmfd.1132053

Anahtar Kelimeler:
Anahtar terim çıkarımı, öznitelik çıkarımı, doğal dil işleme, gözetimli makine öğrenmesi, Naïve Bayes sınıflayıcısı, TÜBİTAK, kavramsal benzerlik

ÖZ

Makale, bildiri, tez ve proje önerisi gibi akademik metinlerin, gelişen doğal dil işleme araçları ve algoritmaları ile işlenmesi sonucunda elde edilen bilgi farklı amaçlar için kullanılabilmektedir. Mevcut çalışmanın ilk aşamasında, mühendislik alanında kullanılan kelime ve kelime gruplarının içerikleri ve yapıları dikkate alınarak bir kütüphane oluşturulmuş; ilgili metni en uygun ve kapsamlı şekilde tanımlayacak anahtar terimlerin/özniteliklerin çıkarımı gerçekleştirilmiştir. Bu işlem sonucunda elde edilen terim vektörleri kullanılarak farklı dokümanların benzerliğinin tespit edilmesine yönelik bir algoritma geliştirilmiştir. Son olarak ise, gözetimli makine öğrenmesi kapsamında Naïve Bayes sınıflandırıcısı kullanılarak TÜBİTAK Araştırma Destek Programları Başkanlığı'na (ARDEB) sunulan proje önerilerinin 31 farklı mühendislik alt alanından hangisine ait olduğunun tespitine yönelik bir analiz gerçekleştirilmiştir. 1255 proje önerisi ile gerçekleştirilen vaka çalışmasında, önerilen benzerlik algoritmasının revize proje önerilerinin benzerlik tespitinde %100'e yakın, sınıflama algoritmasının ise alt alan belirlemede ilk tahminde %83,3, ilk iki tahminde %92,5 ve ilk üç tahminde %96.4'lük doğruluk sağladığı gözlenmiştir.

Natural language processing for the Turkish Academic texts in the engineering field and development of a decision support system: the case of TUBITAK project proposals

HIGHLIGHTS

- Key term retrieval from the academic engineering documents in Turkish.
- Similarity algorithm that detects almost all of the revised project proposal pairs
- Supervised machine learning classification algorithm that predicts the subfields of the proposals

Article Info
Research Article
Received:06.07.2022
Accepted: 07.09.2022

DOI:
10.17341/gazimmfd.1132053

Keywords:
Key term extraction, feature extraction, natural language processing, supervised machine learning, Naïve Bayes classifier, TUBITAK, conceptual similarity

ABSTRACT

The information retrieved from the academic texts such as articles, proceedings, thesis and project proposals are used for a wide range of purposes. In the first phase of this study; a library, that can transform the raw text into a standard form, is created by considering the key terms/features in the engineering field. Then, the key terms that can best represent the document are retrieved and a similarity detection algorithm is developed using these terms. Finally, the Naïve Bayes Classifier in machine learning is used to assign the documents to the appropriate engineering sub-fields. The project proposals submitted to TUBITAK Academic Research Funding Program Directorate (ARDEB) are analyzed as a case study. The results indicate that the proposed similarity algorithm correctly detects almost all of the revised proposals while the accuracy of the classifier is 83.3% in the first prediction and reaches up to 96.4% in the first three predictions over a sample of 1255 proposals.

*Sorumlu Yazar/Yazarlar / Corresponding Author/Authors : *bora.kat@tubitak.gov.tr / Tel: +90 312 298 1231

1. Giriş (Introduction)

Bilginin üretimine ve kullanımına ilişkin gereksinimin öne çıktığı günümüz dünyasında, her alanda hızla artan dijitalleşme ve bunun sonucunda üretilen veri boyutundaki muazzam genişleme, bu verinin yönetilmesi ve işlenerek anlamlı bilgi elde edilmesi hususlarında destek sağlayacak mekanizmaların geliştirilmesini elzem kılmaktadır. Bu noktada, doğal dil işleme (DDİ) çalışmaları, yapısal olmayan metinlerden öznitelik çıkarımı, benzerlik tespiti ve sınıflama konularında farklı alanlarda pek çok sürece destek sağlamaktadır. Özellikle makine öğrenmesi tabanlı metin sınıflama çalışmaları [1] son yıllarda hız kazanmış; duygu analizinden haber metinlerinin indekslenmesine, sosyal medya paylaşımlarının gruplanmasından hukuki kararların sınıflamasına pek çok uygulama alanı bulmuştur. Ancak, Türkçe akademik metinlerin konularına göre tasnifi konusunda önemli bir eksiklik bulunmaktadır. Mevcut çalışma; bu eksikliği, Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK) Araştırma Destek Programları Başkanlığı'na (ARDEB) sunulan proje önerisi metinleri üzerinden gerçekleştirilen analizler sonucunda geliştirilen algoritmalarla gidermeyi amaçlamaktadır.

Dijital dönüşüm sonucu artan veri yükünden, ülkemizin araştırma-geliştirme (Ar-Ge) çalışmalarına fon sağlayan en önemli aktör olan TÜBİTAK da payını almıştır. Son yıllarda, Ar-Ge'ye ayrılan fon miktarları ve beraberinde proje öneri sayıları önemli ölçüde artmış; paydaşlara, proje başvurularına ve desteklenen projeler kapsamında üretilen makale, bildiri, tez, patent vb. proje çıktılarına yönelik verilerin etkin bir şekilde takip edilmesi ihtiyacı ortaya çıkmıştır. Bu ihtiyacı karşılamak üzere; TÜBİTAK, başvuru, değerlendirme, destekleme ve izleme aşamalarında kullanılmak üzere iş uygulama yazılımları geliştirmiştir. Bu yazılımların en yaygın kullanılanları: Araştırmacı Bilgi Sistemi (ARBİS), Proje Başvuru Sistemi (ARDEB-PBS), Panel Yönetim Sistemi (PYS) ve Proje Takip Sistemi'dir (ARDEB-PTS). Diğer taraftan, bu uygulamalarla işletilen süreçlerin daha etkin, verimli ve nitelikli bir şekilde yürütülmesine yardımcı olacak karar destek sitemlerine ihtiyaç da gün geçtikçe artmaktadır. Bu ihtiyaçlardan birisi olan başvuru konularına en uygun değerlendiricilerin seçimi problemi [2]'de ele alınmıştır. Mevcut çalışmada ise ARDEB'e sunulan proje önerileri için ihtiyaç duyulan aşağıdaki üç temel probleme çözüm sunacak yaklaşım ve algoritmalar geliştirilmiş olmakla birlikte; önerilen algoritmalar, benzer yapıdaki Türkçe akademik metinlerin tamamında kullanılabilecek niteliktedir.

i. Proje öneri metinlerinden anahtar terimlerin/özniteliklerin çıkarımı,
ii. Proje önerileri arasındaki kavramsal benzerlik seviyelerinin belirlenmesi,
iii. Bir proje önerisinin konusunun öntanımlı bir konu kümesinden hangisine ait olduğunun belirlenmesi.

İlk sırada yer alan madde, mevcut çalışma için kritik bir öneme sahiptir ve farklı amaçlara hizmet etmektedir. Bunlardan ilki; ARDEB'e 2017 yılından önce sunulan proje önerilerinde anahtar kelime bilgilerinin sistematik bir şekilde alınmamış olması nedeniyle konu bazlı bir sınıflandırmanın ancak kaba bir şekilde yapılabilmesidir. Bu durum, hem kurum içinde gerçekleştirilen planlama çalışmalarında ihtiyaç duyulan analizlerin, hem de farklı kurum ve birimlerden talep edilen veri ve istatistik taleplerinin hızlı, sağlıklı ve tekrar edilebilir bir şekilde gerçekleştirilmesini güçleştirmektedir. ARDEB'in farklı birimleri, desteklenen projeler kapsamında geriye dönük güncellemeleri gerçekleştirmiş olsa da, destek kriterlerini sağlayamayan çok sayıda proje önerisi bulunmaktadır. Bu projelere ait veriler de hem istatistik oluşturmak hem de planlama yapmak açısından önem taşımaktadır. Sayısı yüz binler mertebesinde olan bu öneriler için akıllı bir yaklaşımın geliştirilmesi elzemdir. Anahtar terim çıkarımının diğer önemli katkısı ise sonrasında sunulan iki madde kapsamında girdi oluşturacak olmasıdır; ilk aşamada elde edilecek kelime/kavram vektörü, benzerlik analizlerinin ve makine öğrenmesinin temel girdisi olarak kullanılmaktadır. İkinci sırada yer alan çalışma (ii) ise hem proje ekipleri için hem de ARDEB için pek çok avantaj sağlayacaktır. Bunları kısaca özetlemek gerekirse:

- Proje yürütücülerine başvuru aşamasında, önerilen projenin daha önce ARDEB'e sunulup sunulmadığı sorulmaktadır. Daha önce sunulup değerlendirmeye alınan ancak desteklenmesi uygun bulunmayan önerilerde, proje ekibinin değerlendirme raporunda iletilen hususlara cevap vermesi ve açıklığa kavuşturması için "Proje Önerisi Değişiklik Bildirim Formu"nun doldurulması zorunludur. Revize olan bir proje önerisinde bu durumun beyan edilmemesi iade nedenidir. Geliştirilen benzerlik algoritması sayesinde, proje ekibinin daha önce reddedilen projeleri hızlı bir şekilde taranarak benzerlik oranları belirlenecektir. Bu aşamada, belirli bir eşiğin üzerindeki projeler için başvuru sahiplerine uyarı notu sunulabilir.
- Başvuru sisteminde (ARDEB-PBS), önerilen projenin sadece reddedilen projelerle değil, desteklenmesine karar verilmiş, yürürlükte veya sonuçlanmış olan projelerle ilişkisi de sorgulanmaktadır. Başvuru sisteminde; proje ekiplerinden, mevcut projelerinin bu proje ile ilişkili olabilecek diğer projelerinden (öneri durumunda, desteklenmesine karar verilmiş, yürürlükte veya sonuçlanmış) farklarını açıklamaları istenen bir bölüm bulunmaktadır. Bu bölüme yönelik eksiklik iade sebebi olmamakla birlikte, değerlendirme aşaması öncesinde tamamlatılmaktadır. Revize projeler için önerilen hususa benzer şekilde, bu aşamada da belirli bir benzerlik aralığında olan projeler için başvuru sahiplerine uyarı notu sunulabilir.
- Yukarıdaki iki husus başvuru sahiplerine yönelik olmakla birlikte; bir proje önerisi revize olarak belirtilmemişse veya ilişkili olabileceği projeler için gerekli bilgiler doldurulmamışsa, bu durumların tespiti ARDEB tarafında oldukça güç olabilmekte ve zaman almaktadır. Ayrıca, belirli bir sistematiğin olmayışı, uzmandan bağımsız standart bir sürecin işleyişini güçleştirmektedir. Geliştirilen benzerlik algoritması, proje önerisi için ön değerlendirme yapan ARDEB Uzmanı'na proje ekibinin diğer projeleri ile benzerliklerini gösteren bir arayüzle önemli ölçüde zaman kazandıracak ve sürecin daha etkin bir şekilde işletilmesine olanak sağlayacaktır. Özellikle kalabalık ekiplerin yer aldığı proje önerilerinde, kontrol edilmesi gereken proje sayısı onlarca olabilmektedir.
- Daha önce de değinildiği üzere, başvurular için en uygun değerlendiricilerin belirlenmesine yönelik PaneLIST karar destek sitemi geliştirilmiştir [2]. PaneLIST, araştırmacıların yayınlarında belirttiği anahtar kelimeleri ve ARBİS hesaplarına girmiş oldukları bilgileri kullanarak bir uygunluk skoru oluşturmaktadır. Ancak, kişilerin ARDEB tarafından desteklenmiş projelerindeki uzmanlık alanları henüz bu sisteme entegre edilmemiştir. Önerilen benzerlik algoritması ile, daha önce benzer konularda desteklenmiş olan proje ekiplerinin de potansiyel değerlendirici olarak tespit edilmesi mümkün olacaktır.
- Dönemli veya çağrılı programlarda çok sayıda proje önerisi aynı anda değerlendirme sürecine girmektedir. Bu projeler konularına göre gruplanarak panel değerlendirme süreçleri işletilmektedir. Bu aşamada, ilgili dönemde sunulan önerilerin birbiri ile olan kavramsal benzerliklerinin bilinmesi durumunda, taslak bir gruplama oluşturularak karar vericilerin süreci daha etkin ve nitelikli işletmesi sağlanacaktır.

Son maddede (iii) ele alınan problem bir önceki maddeyle kısmen örtüşmekle birlikte, makine öğrenmesine dayalı akıllı bir çözüm

sunmakta ve daha kapsamlı bir vizyona hizmet etmektedir. ARDEB'de, farklı spesifik programlar yürüten üç grup (kamu araştırmaları, savunma ve güvenlik teknolojileri, uzay araştırmaları) dışında, disiplinler bazında oluşturulmuş bir organizasyon yapısı (Mühendislik, Çevre-Yer-Atmosfer, Elektrik-Elektronik-Enformatik, Kimya-Biyoloji, Matematik-Fizik, Tarım-Orman-Gıda, Sosyal ve Beşeri, Sağlık) bulunmaktadır. Başvurular öncelikle bu ana başlıklar altında gruplanmakla birlikte, her bir ana alan için çok sayıda alt alan mevcuttur. Örneğin, mühendislik alanı 45 alt alana sahiptir. Çalışmanın bu aşamasında, mühendislik alanına sunulan bir projenin hangi alt alana ait olduğunu belirleyen gözetimli makine öğrenmesine dayalı bir algoritma geliştirilmiştir. Algoritmanın aşağıda sunulan ihtiyaçları karşılaması beklenmektedir:

- İstatistik oluşturmada ve planlama çalışmalarında sadece ilk aşamada belirlenen anahtar kelimeler ve kavramlar yeterli olmamaktadır. Aynı kavramın çok sayıda farklı disiplinde yer alması mümkündür. Geliştirilen algoritma ile belirli alt alanlara giren proje önerileri hızlı bir şekilde tespit edilebilecektir. Bu durum, özellikle geçmişe yönelik projelerin indekslenmesinde kullanılacaktır.
- Başvuru aşamasında yürütücülerin doğru alan ismini seçmelerine yardımcı olacaktır.
- Geliştirilen algoritmanın başarısı, değerlendirici belirleme gibi diğer karar noktalarında kullanılabilecek ve yapay zekâ tabanlı yaklaşımlar için yol gösterici olacaktır.

Makalenin sonraki bölümünde literatür taraması sunulacaktır. Bölüm 3'te, ele alınan üç temel problem için izlenen metodoloji açıklanacaktır. Ardından, vaka çalışmasına yönelik bilgiler verilecek ve geliştirilen algoritmaların gerçek veri kümesi üzerinden elde ettiği sonuçlar ve performans analizi sunulacaktır. Son olarak, sonuçların değerlendirilmesi ve bundan sonra gerçekleştirebilecek çalışmalara yönelik tartışma bölümü yer almaktadır.

2. Literatür Taraması (Literature Review)

Projede ele alınacak ilk aşama olan anahtar terim çıkarımı, sonrasında geliştirilecek olan algoritmalara temel girdi teşkil edeceğinden, bu aşamanın başarımı oldukça önemlidir. Bu konuda en kritik husus metinde yer alan kelime ve kelime gruplarının standart bir yapıya dönüştürülmesine yardımcı olacak ön işleme prosedürleridir. Doğal dil işleme algoritmalarında, öncelikle metinlerin noktalama işaretleri ile birlikte bağlaç, edat, zamir gibi durak kelimelerinden arındırılması önem taşımaktadır. Metinlerin sınıflandırılmasında; sözcüklerin temel köklerini elde etmek üzere kullanılan gövdeleme (stemming) ve sözcüklerin morfolojilerini dikkate alarak yalın hale getirmeye çalışan "kök çözümleme" (lemmatization) ön işleme adımları da kullanılmaktadır. Literatürde, gövdeleme işleminin Türkçe metinlerde konu sınıflandırması için etkisinin sınırlı olduğu görülen çalışmalar [3, 4] bulunmakla birlikte; kök çözümleme algoritmalarının bilgi çıkarımında olumlu etki yaptığı görülmektedir [5] ve bu hususa yönelik güncel algoritmaların geliştirilmesini konu alan [6], mevcut algoritmaları kıyaslayan veya bir arada kullanan çalışmalar [7, 8] bulunmaktadır. Mevcut çalışmada, standart ön işleme adımlarına ek olarak gerçekleştirilen asıl ayırt edici katkı, ele alınan mühendislik alanı kapsamında sık kullanılan terimlerin öngörülmesi ve bu terimleri ortak bir yapıya dönüştürecek şablonların tanımlanmasıdır.

İki metnin veya bu metinlerden elde edilen vektörel yapıların benzerliğinin araştırılması literatürde ve uygulamada karşılık bulan problemlerdir. Konu, intihal tespitine yönelik çalışmalar/yazılımlar [9-11] ve kavramsal benzerliğin tespitine yönelik çalışmalar [12] olmak üzere iki genel çerçevede ele alınabilir. Ayrıca, benzerlik düzeyinin tespitine yönelik farklı metrikler tanımlanmıştır [9, 12]. Mevcut çalışma kapsamında, ARDEB'in ihtiyaçları da dikkate alınarak hem metinsel hem de kavramsal benzerliğin tespit edilebileceği bir benzerlik algoritması geliştirilmiştir. Analizler; öngörülen algoritmanın revize önerilerin, kavramsal olarak yakın önerilerin ve etik ihlal şüphesi olan proje önerilerinin tespitinde 100%'e yakın doğruluk sağladığını göstermiştir.

Bir metnin konusuna göre sınıflandırılması literatürde sıkça çalışılan ve özellikle makine öğrenmesi temelli algoritmaların geliştirildiği bir araştırma alanı olarak karşımıza çıkmaktadır [1, 15, 18]. Ağırlıkla İngilizce metinler üzerine yapılan çalışmalarla [13, 14] birlikte Türkçe metinleri ele alan çalışmalar da bulunmaktadır [16-18]. Sınıflandırma, tanımlı bir sınıf kümesine atama şeklinde olabileceği gibi, metinlerin herhangi bir sınıf kümesi olmadan belirli sayıda gruba ayrılması şeklinde de olabilir. İlk durum için genellikle gözetimli makine öğrenmesi algoritmaları kullanılmaktadır ki mevcut çalışma da bu kapsama girmektedir. Literatürde Türkçe metinlere yönelik sınıflama çalışmaları da özellikle son yıllarda önem kazanmıştır. Bu çalışmalarda derin öğrenme [19, 20]; Gizli Dirichlet Ayırımı [21]; Naïve Bayes, Destek Vektör Makinelerini ve J48'in de yer aldığı melez yaklaşımlar [23] kullanılmıştır. Ayrıca, varlık isimleri ile konu başlıklarının ilişkisinin ele alındığı bir çalışma da bulunmaktadır [22]. Türkçe metinlere yönelik çalışmalar; film [23, 27], otel [23] ve restoran [7] yorumları, web sitesi içerikleri [26], tweet içerikleri [23, 28], e-posta içerikleri [29], masallar [25] vb. veri kümeleri üzerinde test edilse de, test verisi olarak ağırlıkla haber metni veri kümeleri [3, 5, 8, 16, 17, 21, 24, 29, 30] kullanılmaktadır. Diğer taraftan, literatürde özellikle İngilizce akademik metinlerin sınıflandırılmasına yönelik kapsamlı çalışmalar yer almaktayken [14, 31-33], Türkçe metinler için yapılmış bir çalışmaya rastlanmamıştır. Uluslararası çalışmalara ek olarak, Kılınç vd. [34] K-En Yakın Komşu (KNN) algoritmasını kullanarak İngilizce özetlerini kullandıkları akademik metinleri, kavramsal olarak görece uzak olan iki farklı alana ("Materials Science & Engineering" and "Social Sciences & Humanities") atamaya yönelik bir çalışma gerçekleştirmiştir.

Mevcut çalışmada, sınıflandırma yaklaşımı olarak, Naïve Bayes [34, 35] algoritması kullanılmıştır. Bu algoritma, literatürde benzer konular için sıkça kullanılmakta ve başarılı sonuçlar vermektedir; ayrıca, ön işleme, eğitim kümesi büyüklüğü vb. pek çok husus açısından da detaylı bir şekilde incelenerek önemli çıkarımlar elde edilmiştir [37-39].

3. Tanımlar ve Metot (Definitions and Methodology)

Çalışma kapsamında gerçekleştirilen aşamaların akışı Şekil 1'de gösterilmiştir. Geliştirilen algoritma ve prosedürlerde kullanılan indis, parametre, küme ve değişkenlerin listesi ve tanımları ise Tablo 1'de sunulmuştur. Tüm aşamalarda Python programlama dili kullanılmıştır.

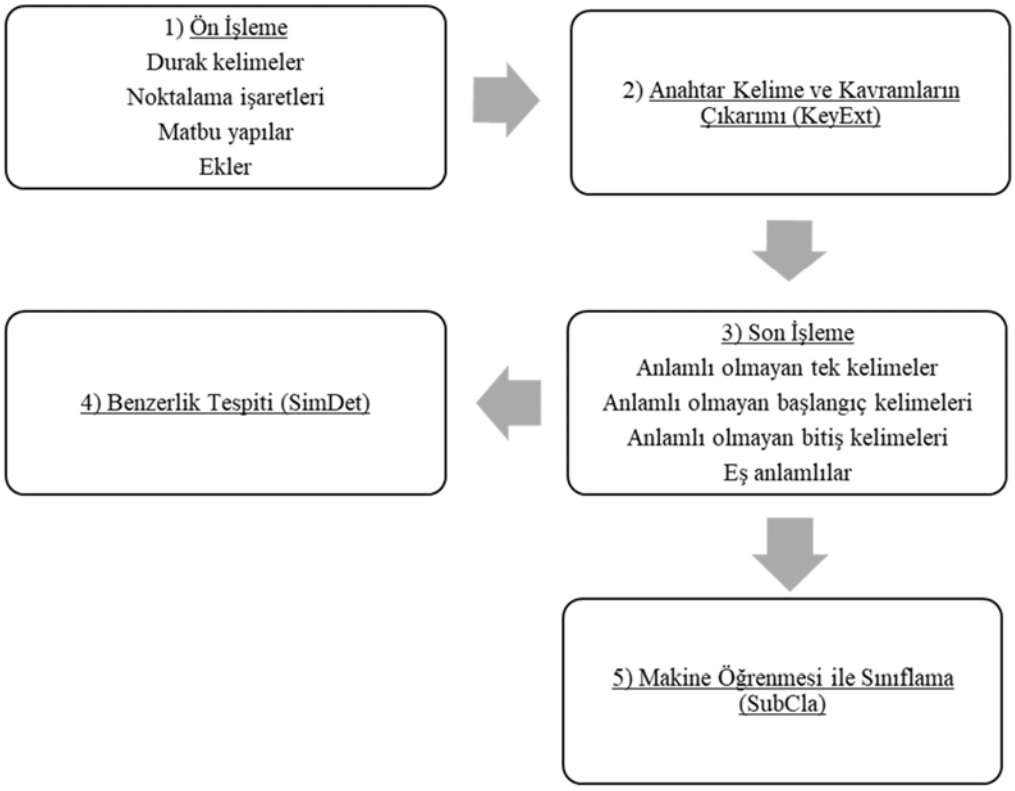
3.1. Tanımlar ve Notasyon (Definitions and Notation)

3.2. Ön İşleme (Preprocessing)

Bu çalışmadaki ön işleme aşamalarının bir kısmı literatürde ve uygulamada takip edilen standart yöntemlerle gerçekleştirilmiş, bir kısmında ise ele alınan metinlere (mühendislik alanındaki akademik proje önerileri) özgü düzenlemeler yapılmıştır. Gerçekleştirilen ön işleme adımları aşağıda sırasıyla sunulmuştur:

- *Noktalama işaretlerinin temizlenmesi*
- *Metnin küçük harflere dönüştürülmesi*
- *Durak kelimelerinin temizlenmesi*

Bu aşamada standart durak kelimelerine (ve, veya, için, ile, gibi, da, sırasında, itibariyle, örneğin, bkz, …) ek olarak iki farklı grup durak kelimelerine daha odaklanılmıştır. Bunlardan ilki, proje öneri



Şekil 1. Çalışma kapsamında izlenen aşamalar (Stages followed within the scope of the study)

Tablo 1. Kümeler ve parametreler (Sets and parameters)

Sembol	Tanım
D, d	Doküman kümesi ve kümede yer alan her bir eleman
D_e	Eğitim kümesi
S, s	Sınıfların kümesi ve kümedeki her bir eleman
K^d, k^d	d dokümanından çıkarımı yapılan terimlerin kümesi ve kümede yer alan her bir terim
$GRAM_n^d$	d dokümanında yer alan n-gramların kümesi, n=1, 2, 3, 4
$\widehat{GRAM_1^d}$	d dokümanında yer alan tekli gramların en yüksek frekansa sahip beş gramın frekans ortalaması
$xGRAM_n^d$	d dokümanında yer alan n-gram vektöründe yer alan en yüksek frekans değeri
$tf^d(k)$	d dokümanında yer alan k teriminin frekans değeri
$atf^d(k)$	d dokümanında yer alan k teriminin frekans değeri - uyarlanmış
γ_n	n-gram'lar için ölçeklendirme parametresi – üs değeri
δ	Çoklu gramlar için ölçeklendirme parametresi – metindeki toplam kelime sayısına oran
θ	Benzerlik tespiti algoritmasında dikkate alınacak doküman skor eşik değeri
σ	Benzerlik tespiti algoritmasında, düşük sayıdaki örtüşmeleri cezalandırma parametresi
τ	Benzerlik algoritmasında yüksek frekanslı örtüşmelere yönelik ikinci eşik değeri
π, C, μ	Benzerlik analizinde kullanılan kümelerin ve kesişim kümesinin eleman sayıları üzerinden hesaplanan düzeltme çarpanı ve çarpan hesabında kullanılan parametreler
$sim(x, y)$	x ve y dokümanları arasındaki benzerlik skoru - [0,1]
s_{ML}	En yüksek olasılığa sahip sınıf
$m(s)$	Eğitim kümesinde yer alan s sınıfına ait doküman sayısı
α	Düzeltme parametresi
V	Eğitim kümesindeki tüm dokümanlarda yer alan terimlerden oluşan küme

formlarında sıklıkla geçen ancak konu içeriği ile ilişkili olmayan "burs, "bursiyer", "proje ekibi", "dr. öğretim üyesi", "doç. dr.", "prof. dr.", "iş-zaman çizelgesi", "tübitak", "yaygın etki" vb. TÜBİTAK proje başvuruları jargonuna ait olan sözcükler ve sözcük gruplarıdır. Bu bölümde, ayrıca herhangi bir terim ile karıştırılma ihtimali olmayan özel isimler de ("Ayça", "Berk", "Öztürk", "Özdemir" …) durak kelimeleri olarak belirlenmiştir. Diğer grupta ise, gerçekleştirilen ön analizler sonucunda yüksek frekansa sahip olan ancak konu sınıflandırması açısından ayırt edici bir özelliğe sahip olmayan sözcük/sözcük grupları (örn. "-mektedir/-maktadır", "-cektir/-caktır", "-diği/-dığı/-düğü/-duğu", "-miştir/-mıştır/-müştür/-muştur" ile biten kelimeler, "değildir", "vardır", "yoktur", "zorunludur", "şekliyle" vb. kelimeler) yer almaktadır. Diğer bir deyişle, metinden silindiğinde kavramların belirlenmesini etkilemeyecek nitelikte ve düşük IDF (ters belge frekansı) değerlerine sahip olan ifadeler metinden çıkarılmıştır.

- *Sık kullanılan ve içeriği belirleyebilecek kelimelerin standart hale dönüştürülmesi*

Daha önce de değinildiği üzere Türkçe metinlerde sözcük köklerinin bulunması ve sınıflandırmada bu köklerin kullanılmasının sınıflandırma performansına etkisinin sınırlı olduğu görülmektedir. Mevcut çalışmada, amaçlardan birisi de akademik metinlerdeki anahtar terimleri çıkarmak olduğundan, bu terimlerin mümkün olduğunca kullanıldığı haliyle elde edilmesi önemlidir. Yapılan ön çalışmalar, akademik terimlerin ve metin konularının çoğunlukla sözcüklerin yalın hali ile (örn. "sentez", "rota", "ısı", "çimento", "dinamik") veya belirtisiz isim tamlaması şeklinde (örn. "polimer sentezi", "araç rotası", "vücut ısısı", "portland çimentosu", "akışkanlar dinamiği") kullanıldığını göstermiştir. Bu nedenle, sözcüklerin yapım eklerine değil, çekim eklerine odaklanılmış; bu eklerden arındırılmış sözcüklerin elde edilmesine yönelik Tablo 2'de örnekleri verilen dönüştürme şablonları kullanılmıştır. Örneklerden görüldüğü üzere; şablon oluşturmada, sert sessiz yumuşaması ve ünlü düşmesi gibi dilbilgisi kuralları da dikkate alınmıştır. Bazı kelimeler için boş olan satırın sebebi, bu kelimelerin "i hali" ile üçüncü tekil şahıs iyelik eki almış hallerinin aynı olmasından kaynaklı sorunu engellemektir. Bu şablonlar, akademik metinlerde sık geçen terimleri, sözcüklerin yalın hallerine veya üçüncü tekil şahıs iyelik eki almış hallerine dönüştürmektedir. Ayrıca, bu şablonlar kullanılarak, tek tek terimler yerine ortak ekleri olan sözcüklerin ("-yon", "-tör", "-izm", "-lilik", "-lenme", "-loji", "-grafi", …) toplu halde standart forma dönüştürülmesi de sağlanmıştır.

3.3. Anahtar Kelime ve Kavramların Çıkarımı - KeyEx
(Keyword and Key Concept Extraction - KeyEx)

Ön işleme aşamasının ardından elde edilen metinlerden anahtar terimlerin çıkarımı için öncelikle metinlerin dizgeciklere (token) ayrılması gerekir. Bu ayırma işlemin sonucunda gram denilen n (n=1, 2, 3, …) tane elemandan oluşan ardışık diziler elde edilir. Mevcut çalışmada tekli (uni-gram), ikili (bi-gram), üçlü (tri-gram) ve dörtlü (4-gram) diziler oluşturulduktan sonra, metin içinde geçme frekansları belirlenerek bir terim-frekans vektörü oluşturulmuştur. Ek olarak, eğer metin içerisinde metin yazarları tarafından tanımlanmış anahtar kelimeler varsa (hemen hemen tüm akademik makalelerde ve proje öneri formlarında özetin altında yer almaktadır), geliştirilen kod bu kavramları da ayırt edebilmekte ve diğer n-gram'lardan elde edilen en yüksek frekansla terim-frekans vektörüne eklemektedir. Vektörde yer alan terimler, uzman gözüyle daha detaylı analiz edilmek üzere bir kelime bulutu görseline dönüştürülmüştür (Şekil 4). Bu aşamada, hem bulut görselinde daha anlaşılır bir yapı elde edebilmek, hem de sonrasında çalıştırılacak olan benzerlik algoritmasında daha anlamlı sonuçlar almak için frekanslar üzerinde bir ölçeklendirme gerçekleştirilmiştir. Kullanılan ölçeklendirme fonksiyonu Eş. 1 ve Eş. 2'de tanımlanmıştır.

$$atf^d\left(k^d \middle| k^d \epsilon GRAM_1^d\right) = \left[\frac{tf^d(k^d)}{\widehat{GRAM_1^d}}\right]^{\gamma_1} \quad \forall\, d \in D \tag{1}$$

$$atf^d(k^d \middle| k^d \epsilon GRAM_n^d) = \left[\frac{tf^d(k^d)}{\max\{xGRAM_n^d, \delta \cdot |GRAM_1^{d|}\}}\right]^{\gamma_n} \quad \forall\, d \in D, n = 2,3,4 \tag{2}$$

Bu iki denklemdeki ilk amaç, öncelikle frekans değerlerini normalize ederek 0-1 aralığına getirmektir. İlk denklemde, tekli gramların frekans değerleri, en yüksek beş değerin ortalamasına bölünmektedir. İkinci denklemde ise, çoklu gramların frekansı, ilgili gram kümesindeki en yüksek frekansa veya metindeki toplam kelime sayısının belirli bir oranına bölünmektedir. Bu hesaplamanın ardından 1'den büyük olan tüm değerler 1'e eşitlenmektedir. Diğer amaç ise, tekli ve çoklu dizilerin frekans farkından kaynaklanan uyumsuzluğu gidermektir. Tekli gramlar bir metin içerisinde farklı amaçla yazılmış ifadelerin içerisinde farklı kelime ve kelime grupları ile birleşerek yüksek frekansa sahip olabilirler. Frekanslarının yüksek olması, o kelimelerin ilgili metin için tanımlayıcı veya ayırt edici nitelikte olduğunu söylemek için yeterli olmayabilir. Ancak, çoklu dizilerin metinde tekrar etmesi, bu dizilerin projeyi betimleyen bir kavram olma ihtimalini güçlendirmektedir. Örneğin "makine" ve "teori" kelimeleri ilgili alandaki bir metinde yüksek frekanslara sahip olabilir. "Makine" kelimesi, kullanılan bir cihazın, analiz yapılacak

Tablo 2. Sözcüklerin standart yapılara dönüştürülmesi (Converting words to standard structures)

motorundaki >>> motoru	sürecindeki >>> süreci	algoritmasındaki >>> algoritması	beynindeki >>> beyni
motorundan >>> motoru	sürecinden >>> süreci	algoritmasından >>> algoritması	beyninden >>> beyni
motorunda >>> motoru	sürecinde >>> süreci	algoritmasında >>> algoritması	beyninde >>> beyni
motoruna >>> motoru	sürecine >>> süreci	algoritmasına >>> algoritması	beynine >>> beyni
motoruyla >>> motoru	süreciyle >>> süreci	algoritmasıyla >>> algoritması	beyniyle >>> beyni
motorunun >>> motoru	sürecinin >>> süreci	algoritmasının >>> algoritması	beyninin >>> beyni
motorunu >>> motoru	sürecini >>> süreci	algoritmasını >>> algoritması	beynini >>> beyni
motordaki >>> motor	süreçteki >>> süreç	algoritmadaki >>> algoritma	beyindeki >>> beyin
motordan >>> motor	süreçten >>> süreç	algoritmadan >>> algoritma	beyinden >>> beyin
motorda >>> motor	süreçte >>> süreç	algoritmada >>> algoritma	beyinde >>> beyin
motora >>> motor	sürece >>> süreç	algoritmaya >>> algoritma	beyine >>> beyin
motorla >>> motor	süreçle >>> süreç	algoritmayla >>> algoritma	beyinle >>> beyin
motorun >>> motor	sürecin >>> süreç	algoritmanın >>> algoritma	beynin >>> beyin
---	---	algoritmayı >>> algoritma	---
motorlarındaki >>> motoru	süreçlerindeki >>> süreci	algoritmalarındaki >>> algoritması	beyinlerindeki >>> beyni
motorlarından >>> motoru	süreçlerinden >>> süreci	algoritmalarından >>> algoritması	beyinlerinden >>> beyni
motorlarında >>> motoru	süreçlerinde >>> süreci	algoritmalarında >>> algoritması	beyinlerinde >>> beyni
motorlarına >>> motoru	süreçlerine >>> süreci	algoritmalarına >>> algoritması	beyinlerine >>> beyni
motorlarıyla >>> motoru	süreçleriyle >>> süreci	algoritmalarıyla >>> algoritması	beyinleriyle >>> beyni
motorlarının >>> motoru	süreçlerinin >>> süreci	algoritmalarının >>> algoritması	beyinlerinin >>> beyni
motorlarını >>> motoru	süreçlerini >>> süreci	algoritmalarını >>> algoritması	beyinlerini >>> beyni
motorlardaki >>> motor	süreçlerdeki >>> süreç	algoritmalardaki >>> algoritma	beyinlerdeki >>> beyin
motorlardan >>> motor	süreçlerden >>> süreç	algoritmalardan >>> algoritma	beyinlerden >>> beyin
motorlarda >>> motor	süreçlerde >>> süreç	algoritmalarda >>> algoritma	beyinlerde >>> beyin
motorlara >>> motor	süreçlere >>> süreç	algoritmalara >>> algoritma	beyinlere >>> beyin
motorlarla >>> motor	süreçlerle >>> süreç	algoritmalarla >>> algoritma	beyinlerle >>> beyin
motorların >>> motor	süreçlerin >>> süreç	algoritmaların >>> algoritma	beyinlerin >>> beyin
motorları >>> motoru	süreçleri >>> süreci	algoritmaları >>> algoritması	beyinleri >>> beyni
motorlar >>> motor	süreçler >>> süreç	algoritmalar >>> algoritma	beyinler >>> beyin

laboratuvarın bağlı bulunduğu bölümün veya proje kapsamında iş birliği yapılan bir kuruluşun isminde geçebilir. Benzer şekilde, sunulan literatür veya yöntem bölümlerinde, çalışmada yapılacak pek çok farklı teoriye atıfta bulunurken "teori" kelimesi kullanılabilir. Ancak, "makine teorisi" ikili dizisinin tekrarlı bir şekilde metinde yer alması; frekansı, tekli gramlara göre az bile olsa, metin için daha anlamlı bir kavram olduğuna işaret edebilir. Bu noktada, önemli bir husus, kelimelerin IDF değerleri ile ağırlıklandırılması fikri olabilir. Ancak, bir metni belirli bir doküman kümesine bağlı kalmadan kendi içinde gerçekte kullanılan haline yakın bir şekilde kavramsallaştırma hedefi ve ARDEB özelinde ihtiyaç duyulan benzerlik tespitinde kavramsal benzerlik kadar metinsel benzerliğin de önemli olması nedeniyle IDF ağırlıkları kullanılmamıştır.

3.4. Son İşleme (Post-processing)

Ön işleme aşamasında gerçekleştirilen işlemler, metni mümkün olduğunca standart bir yapıya getirmeye yöneliktir. Son işleme aşamasında ise artık metinde değil, metinden elde edilmiş terim-frekans vektöründe güncellemeler yapılmıştır. DDİ çalışmalarında pek karşılaşılmayan son işleme aşamasının mevcut çalışmada kullanılmasındaki amaç, daha önce de değinildiği üzere, elde edilen kavramların mümkün olduğunca ek bir uzman değerlendirme sürecine ihtiyaç duymadan metinleri indeksleyebilecek niteliğe kavuşturulmasıdır. Bu bağlamda, detaylı ön analizler ve frekans analizleri sonucunda, aşağıdaki üç grupta yer alan kavramlar, terim-frekans vektöründen çıkarılmıştır:

- Tek başına anlam ifade etmeyen tekli kavramlar. Bu kelimeler, genel anlamları olan kelimeler ("amaç", "ortak", "deney", …) veya genellikle çoklu dizilerin bir parçası olan kelimelerdir ("kendiliğinden", "enerjisi", "sulu", …).
- İlk kelimesi, dilbilgisi açısından anlamlı bir kavram oluşturulmasına olanak vermeyen kelimeler. Bu kelimeler de genellikle iyelik eki almış olan ve çoklu dizilerde kavramın sonraki bölümlerinde yer alan kelimelerdir ("tabanlı", "dayalı", "tokluğu", …).
- Kavramın son kelimesi olarak anlam ifade edemeyecek kelimeler. Bu kelimeler de genellikle isim veya sıfat tamlamalarında sıfat veya tamlayan niteliğinde olup aslında başka dizilerde farklı konumda yer alan kelimelerdir ("yeni", "bölgesel", "karbonca",…).

3.5. Benzerlik Tespiti - SimDet (Similarity Detection - SimDet)

Çalışma kapsamında, literatürdeki benzerlik tanımları da (cosine benzerliği, jaccard benzerliği) dikkate alınarak, hem kavramsal hem de metinsel benzerliğin bir arada ele alındığı, çalışmaya özgü bir benzerlik algoritması geliştirilmiştir. Bu kapsamda, bir metin ($x \in D$) için oluşturulan uyarlanmış terim-frekans (adjusted term-frequency) vektörü atf^x ile diğer bir metin ($y \in D$) için tanımlanmış olan ve uyarlanmış terim-frekans vektörü atf^y üzerinden hesaplanan benzerlik aşağıda sunulan denklemler kapsamında elde edilmektedir.

İlk olarak, her bir dokümandan oluşturulan terim kümelerinde yer alan ve atf değerleri belirli bir eşik seviyesinin üzerinde olan terimlerden oluşan kesişim kümesi oluşturulmuştur, Eş. 3-Eş. 6'da ise; sırasıyla her bir doküman için eşik seviyesi üzerinde skora sahip terimlerin sayısı ve kesişim kümesindeki terim sayısı hesaplanmaktadır. Eş. 7'de, kesişim kümesinin her bir dokümandaki terimlerin ne kadarını kapsadığını dikkate alan bir benzerlik çarpanı hesaplanmıştır.

$$INT(x,y) = set(x|atf^x(k^x) > \theta) \cap set(y|atf^y(k^y) > \theta) \quad (3)$$

$$len(x) = |set(x|atf^x(k^x) > \theta)| \quad (4)$$

$$len(y) = |set(y|atf^y(k^y) > \theta)| \quad (5)$$

$$lenINT(x,y) = |INT(x,y)| \quad (6)$$

$$\pi = \left[C \cdot \frac{lenINT(x,y)}{len(x)} \cdot \frac{lenINT(x,y)}{len(y)}\right]^{\mu} \quad (7)$$

Benzerlik tespitine yönelik bir skor oluşturmak için öncelikle kesişim kümesindeki terimler için, her bir dokümandaki atf değerlerinin geometrik ortalaması alınmıştır, Eş. 8. Daha sonra, Eş. 9'da, elde edilen geometrik ortalamaların aritmetik ortalamaları alınmaktadır. Ancak, ilgili denklemde de görüleceği üzere; aritmetik ortalama alınırken, geometrik ortalamaların toplamı, kesişim kümesinin belli bir değerin altında olması durumda sabit bir sayıya bölünmektedir. Ön çalışmalar sonucunda yapılan gözlemlere dayalı bu uyarlama, çok az sayıda ama yüksek skorlu eşleşmelerden kaynaklı yanıltıcı sonuçları engellemektedir. Benzer şekilde; daha önce hesaplanan π çarpanı, metinlerdeki terimlerin ne ölçüde örtüştüğü bilgisini benzerlik skoruna yansıtmak için kullanılmaktadır. Ayrıca, hesaplanan aritmetik ortalama için eklenen üst terimi, görece yüksek frekanslı örtüşmelerin etkisini arttırmak ve özellikle revize proje önerilerin tespitini kolaylaştırmak amacıyla kurgulanmıştır. Algoritmada kullanılan parametreler kapsamlı bir ön çalışma sonucunda belirlenmiş ve Tablo 3'te sunulmuştur.

$$atf^z(k^z) = \sqrt{atf^x(k^z) \cdot atf^y(k^z)} \quad \forall k^z \in INT(x,y) \quad (8)$$

$$sim(x,y) = \pi \cdot \left[\frac{\sum_{\forall k^z \in setU(x,y)} atf^z(k^z)}{\max\{\sigma, lenINT(x,y)\}}\right]^{\frac{2}{(1+|INT(x,y)|atf^z(k^z)>\tau|)}} \quad (9)$$

Tablo 3. SimDet kapsamında kullanılan parametre değerleri
(Parameters used in SimDet)

θ = 0,60	C = 100	μ = 0,50	σ = 50	τ = 0,65

Bu tabloda yer alan parametrelerden θ daha önce de belirtildiği üzere her iki dokümanda da belirli bir sayının üzerinde yer alan ortak terimleri belirlemek için kullanılan eşik değerdir. Dokümanlarda yer alan anahtar terimler frekanslarına göre sıralanmakta; en yüksek frekansa sahip terimin skoru 1 olacak şekilde, dizi büyüklüğü de (tekli, ikili, üçlü veya dörtlü) dikkate alınarak, Bölüm 3.3'te detaylı bir şekilde açıklandığı üzere bu terimlere skor değerleri atanmaktadır. Tabloda yer alan 0.60 değeri, hem proje başvurusunda ekipler tarafından sunulan anahtar kelimeler hem de grup uzmanlarının görüşleri dikkate alınarak belirlenmiş bir eşik değeridir. Bu değerin üstünde yer alan terimlerin tamamı ya proje ekipleri tarafından başvuru aşamasında anahtar terim olarak girilmiş ya da grup uzmanı tarafından kesinlikle anahtar kavram olarak dikkate alınması gerektiği tespit edilmiştir. Eşiğin altında kalan terimlerin önemli bir kısmının proje özelinde tanımlayıcı nitelikte olmadığı, daha çok benzer alandaki projelerde yer alabilecek genel ve sık kullanılan terimlerden oluştuğu gözlenmiştir. Eş. 7'de yer alan benzerlik katsayısı hesabında kullanılan C ve μ parametreleri ise temel olarak ölçeklendirme amaçlı kullanılmaktadır; benzerlik skorunun nihai değerinin 0-100 ve revize olarak işaretlenecek projelerin skorunun 90-100 arasında olmasını sağlamak için taranan parametre kümeleri arasından seçilmiştir. İlgili denklemde de görüldüğü üzere; C değeri dışındaki ifadeler dikkate alındığında, yüksek frekanslı terimlerin yer aldığı kesişim kümesinin, her bir dokümandaki ilgili terim kümelerine oranları hesaplanarak geometrik ortalamaları alınmaktadır. Kesişim kümeleri ilgili dokümanların kümesine ne kadar yakınsa, benzerlik çarpanları da o kadar yüksek olmaktadır. σ değeri de benzer şekilde ölçeklendirme amaçlı kullanılan bir parametredir. Daha önce de belirtildiği üzere çok az sayıda ama yüksek skorlu eşleşmelerden kaynaklı yanıltıcı sonuçları engelleme amaçlı kullanılan bu parametre gerçekte revize olan projelerde yer alan kesişim kümelerinin ortalama eleman sayıları baz alınarak belirlenmiştir. Kullanılan 50 kelime değeri az sayıda ama yüksek frekanslı kesişimleri olan proje çiftlerinin benzerlik

skorlarının düşük olmasını sağlamaktadır. Bu parametre için 30'un altındaki değerler aslında revize olmadıkları halde yüksek benzerlik skoruna sahip olan proje çiftlerine neden olmaktadır. 30 ve üzerindeki σ değerlerinde revize projelerin tamamı tespit edilebilmekle birlikte, daha güvenli tarafta kalabilmek adına nihai çalışmada 50 değeri kullanılmıştır. Son olarak, τ için kullanılmış olan 0.65 değeri, kesişim kümelerinde yer alan skor değerlerinin 0,60-0,65 aralığında kümelendiği gözlemi sonucunda belirlenmiştir. Bu değerin üzerindeki eşleşme sayısının yüksek olduğu durumlarda benzerlik oranını güçlendirmek amacıyla kullanılmıştır.

3.6. Makine Öğrenmesi ile Alt Alan Sınıflama - SubCla (Subfield Classification by Machine Learning - SubCla)

Çalışmada kullanılan Naïve Bayes sınıflandırıcısı; sınıfı tahmin edilecek bir metnin her bir sınıfa ait olma ihtimalini, eğitim kümesinde yer alan metinler üzerinden belirlenen olasılıklar kapsamında hesaplamak için kullanılan gözetimli bir makine öğrenmesi yaklaşımıdır. Naïve Bayes sınıflandırıcısını kullanabilmek için öncelikle sınıfları belirlenmiş bir eğitim kümesine, D_e, ihtiyaç vardır. Eş. 10'da, d dokümanının s sınıfına ait olma olasılığının Bayes teorisine göre eşit olduğu ifade görülmektedir. Bu ifade, önsel (prior) olasılık ile koşullu (conditional) olasılığın çarpımının kanıt değerine (evidence) bölümünü göstermektedir. Sonuç olarak, ara aşaması Eş. 11'de verilen ilişki, Eş. 12'de sunulduğu şekliyle; metninden elde edilen terim-frekans vektörüne sahip bir dokümanın hangi sınıfa ait olduğuna yönelik olasılığı maksimize eden sınıfı tespit etmektedir.

$$p(s|d) = \frac{p(d|s)\cdot p(s)}{p(d)} \tag{10}$$

$$s_{ML} = argmax_{s\in S}\, p(s|d) = argmax_{s\in S}\, \frac{p(d|s)\cdot p(s)}{p(d)} \tag{11}$$

$$s_{ML} = argmax_{s\in S}\{\prod_{k^d\in K^d} p(k^d|s)\cdot p(s)\} \tag{12}$$

Hedef sınıfın belirlenebilmesi için Eş. 12'de yer alan $p(s)$ ve $p(k|s)$ değerlerinin hesaplanması gerekmektedir. Bu değerlerden ilki, eğitim kümesinde s sınıfına ait olan doküman sayısının toplam eğitim dokümanı sayısına bölünmesiyle elde edilmektedir, Eş. 13. Bir metnin s sınıfına ait olduğu bilindiği durumda, ilgili metinde k teriminin yer alma olasılığı ise Eş. 14'te sunulmuştur; payda yer alan ifade ilgili sınıftaki tüm dokümanlarda yer alan k terimlerinin toplamını, payda ise s sınıfında yer alan tüm dokümanlardaki toplam terim sayısını vermektedir. Pay ve paydaya eklenen +1 değerleri, Eş. 12'de yer alan çarpımın 0 olmasını engellemek ve tahmin performansını iyileştirmek üzere kullanılan bir düzeltme parametresidir [34, 35]; 1 olduğunda Laplace, daha küçük olduğunda Lidstone düzeltmesi olarak adlandırılmaktadır [35].

$$p(s) = \frac{m(s)}{|D_e|} \tag{13}$$

$$p(k|s) = \frac{[\sum_{d\in s} tf^d(k)]+1}{\sum_{v\in V, d\in s}[tf^d(v)+1]} = \frac{[\sum_{d\in s} tf^d(k)]+1}{\sum_{v\in V, d\in s}[tf^d(v)]+|V|} \rightarrow \frac{[\sum_{d\in s} tf^d(k)]+\alpha}{\sum_{v\in V, d\in s}[tf^d(v)]+\alpha\cdot|V|} \tag{14}$$

Algoritmanın işleyişini sağlamak ve hem hız hem de tahmin performansını artırmak üzere aşağıdaki uyarlamalar gerçekleştirilmiştir:

- Çok sayıda terimin yer aldığı Eş. 12'de, çok küçük değerlere sahip olan bu terimlerin çarpımının sıfıra yakınsaması problemini aşmak için bu terimlerin çarpımı yerine, doğal logaritmalarının toplamı şeklinde ele alınmıştır (yapılan düzenleme, $\ln(AxB) = \ln(A) + \ln(B)$ ilişkisi nedeniyle elde edilen sonuçları etkilememektedir).
- Her bir metin için frekansı 3 ve daha fazla olan terimler dikkate alınmış; frekans değerleri de doğal logaritmaları alınarak ölçeklendirilmiştir.

4. Vaka Çalışması (Case Study)

4.1. Veri (The Data)

Başvuru sahipleri, ARDEB'e proje önerisi sunarken, projelerinin hangi destek grubuna ve bu destek grubu altında yer alan panel alanlarından hangisine uygun olduğunu işaretlemektedir. MAG'da yer alan mühendislik alanları ve alt alanlar Tablo 4'de sunulmuştur. Görüldüğü üzere, "Endüstri Mühendisliği", "Havacılık ve Uzay Mühendisliği" ve "Petrol ve Doğal Gaz Mühendisliği" sadece tek bir ana alandan oluşmaktayken, diğer mühendislikler için detaylı alt alanlar ve her biri için "diğer" opsiyonu tanımlanmıştır. Liste, toplam 45 seçenekten oluşmaktadır.

Konu ile ilgili uzman deneyimleri ve ön analizler sonucunda, "diğer" seçenekleri ile birlikte seçilen zaman aralığı ve program kapsamında birkaç öneri yer alan "Havacılık ve Uzay Mühendisliği", "Petrol ve Doğal Gaz Mühendisliği" ve Mimarlık kapsam dışında tutulmuştur. Birbirine yakın içeriğe sahip olabilen ve alt alanlar dikkate alındığında yeterli sayıda öneri elde edilemeyen Tekstil Mühendisliği ise tek bir alan olarak ele alınmıştır. Sonraki analizler için kolaylık olması açısından, vaka çalışmasında kullanılan alanlar için tabloda görüldüğü üzere 4 harften oluşan birer kısaltma tanımlanmıştır.

Çalışmada 2015-2021 yılları arasında MAG'a "1001-Bilimsel ve Teknolojik Araştırma Projelerini Destekleme Programı" kapsamında sunulmuş olan 1255 proje önerisi analiz edilmiştir. Bu projelerin alanlara ve alt alanlara dağılımı Şekil 2 ve Şekil 3'te sunulmuştur.

4.2. Algoritmaların Performansları (Performance of the Algorithms)

4.2.1. KeyExt

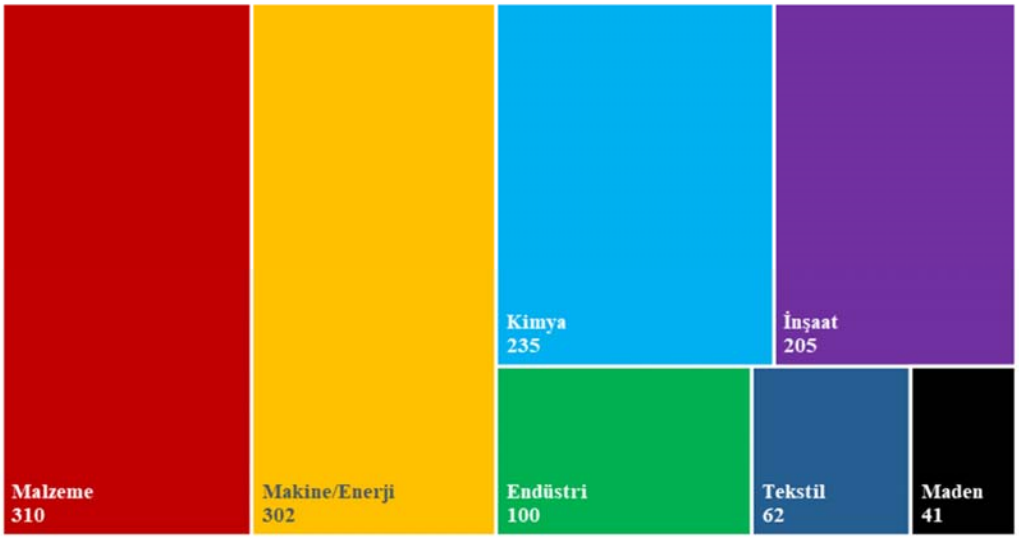
Durak kelimelerinin, ARDEB önerilerinde yer alan matbu bölümlerin, noktalama işaretleri ve rakamların temizleme işlemleri sadece akademik metinler için değil DDİ uygulamalarında ele alınan tüm uygulama alanları için benzer şekilde işlemektedir. Dolayısıyla, anahtar terimlerin çıkarımı aşamasının performansı; ağırlıkla, metinlerde yer alan mühendislik kavramlarının ne ölçüde standart hale dönüştürüldüğüne bağlıdır. Bu nedenle, detaylı bir ön çalışma gerçekleştirilmiş ve kelimeleri benzer formlara dönüştüren şablonlar alandaki terimleri mümkün olduğunca kapsayacak şekilde uygulanmıştır. Ayrıca, benzer çalışmalardan farklı olarak, son işleme aşamasının kurgulanmış olması da önemli katkı sağlamıştır. Elde edilen terim kümelerinde yer alan yüksek frekanslı terimlerin gösterildiği kelime bulutları dört farklı mühendislik alanından rasgele seçilen projeler için Şekil 4'te sunulmuştur. Bu şekilden de görüldüğü üzere; anahtar terimler, tekrar bir uzman güncellemesi gerektirmeksizin kullanılabilecek niteliktedir. Diğer taraftan, her bir metinde yüzlerce terim çıkarılmakla birlikte, özellikle düşük frekanslı terimlerin projeleri tanımlayacak içerikte olmadığı gözlenmiştir. Bu nedenle, sonraki aşamalar için geliştirilen algoritmalarda frekans filtrelemeleri denenmiş ve algoritma performansında olumlu etki yaptığı görülmüştür.

4.2.2. SimDet

İlk bölümde de belirtildiği üzere; öneri metinlerinin benzerliğinin belirlenmesi iki önemli amaca hizmet etmektedir. Bunlardan ilki, kavramsal olarak benzer projelerin belirlenmesi; diğeri ise, benzer oldukları öngörülen projeler için durumu bir adım ileri götürerek revize proje önerilerinin tespit edilmesidir. Reddedilen proje

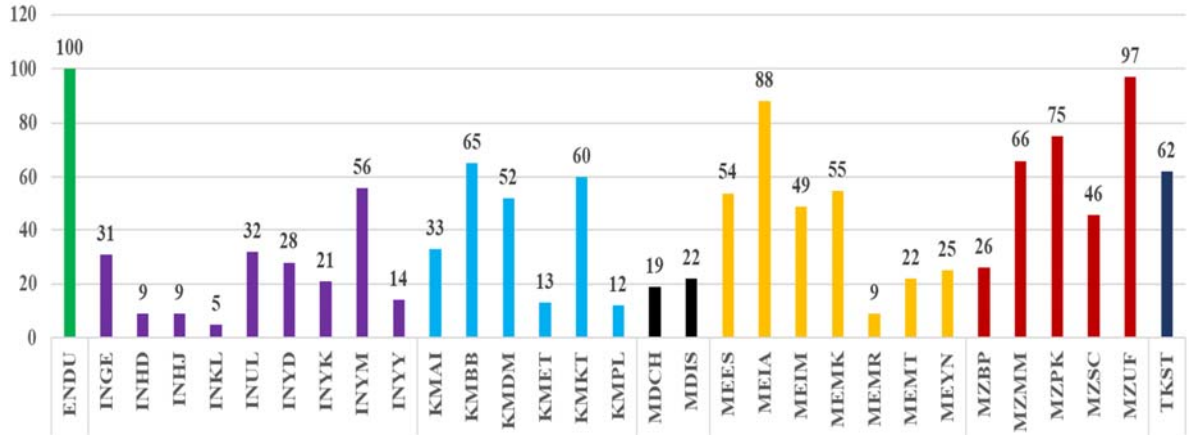
Tablo 4. MAG'da yer alan mühendislik alanları ve alt alanlar (Engineering fields and sub-fields in MAG)

Endüstri Mühendisliği (ENDU)	Makine/Enerji Mühendisliği
Havacılık ve Uzay Mühendisliği	Enerji Sistemleri (MEES)
İnşaat Mühendisliği	İmalat-üretim teknolojileri (MEIM)
Geoteknik (INGE)	Isı-Akışkan (MEIA)
Hidrolik (INHD)	Makine Teorisi, Dinamiği ve Tasarımı (MEMT)
Hidroloji (INHJ)	Mekanik (MEMK)
Kıyı Liman (INKL)	Mekatronik-Robotik (MEMR)
Mimarlık	Yanma (MEYN)
Ulaştırma (INUL)	Diğer
Yapı- Deprem (INYD)	Malzeme Bilimi ve Mühendisliği
Yapı Malzemeleri (INYM)	Batarya / Pil Teknolojileri (MZBP)
Yapı Mekaniği (INYK)	Metalurji / Metal Alaşım ve Kompozitleri (MZMM)
Yapım Yönetimi (INYY)	Polimer Kompozit (MZPK)
Diğer	Seramik /Cam (MZSC)
Kimya Mühendisliği	Uygulamalı Fizik (MZUF)
Ayırma İşlemleri (KMAI)	Diğer
Biyoteknoloji – Biyomühendislik (KMBB)	Petrol ve Doğal Gaz Mühendisliği
Polimer (KMPL)	Tekstil Mühendisliği (TEKS)
Doku Mühendisliği (KMDM)	Konfeksiyon Teknolojisi
Enerji Teknolojileri (KMET)	Teknik Tekstiller
Katalizör (KMKT)	Tekstil Makineleri
Diğer	Tekstil Malzemeleri
Maden Mühendisliği	Tekstil Teknolojileri
Cevher Hazırlama (MDCH)	Tekstil Terbiyesi ve Kimyası
İşletme (MDIS)	Diğer
Diğer	



Şekil 2. Vaka çalışmasında yer alan projelerin ana alanlara dağılımı: sayı.
(Break-down of proposals in the case study by main fields: number)

önerilerinin revize edilerek tekrar sunulması, araştırmanın doğası gereği beklenen ve sıkça karşılaşılan bir durumdur. TÜBİTAK, akran değerlendirmesi kapsamında uzun yıllardır edindiği bilgi birikimi ve tecrübe sonucunda, revize edilerek sunulan projeler için iyi tanımlanmış bir sürece sahiptir ve bu sürecin sağlıklı bir şekilde işletilmesine özen göstermektedir. Proje yürütücüleri, başvuru aşamasında proje önerilerinin revize olup olmadığını beyan etmekle yükümlüdür. Revize olduğu beyan edilen projeler için, bir önceki değerlendirme raporunda değinilen hususlara cevap verilmesi ve açıklık getirilmesi beklenmektedir. Ayrıca, eğer proje ekiplerinin; mevcut önerileri ile ilişkili olabilecek, öneri durumunda, desteklenmesine karar verilmiş, yürürlükte veya sonuçlanmış projeleri varsa, mevcut önerileri ile bu projelerin farklarını açıklamaları gerekmektedir. Çalışma kapsamında, öncelikle MAG'da bir pilot ön çalışma yapılmış ve benzerlik skoru için hangi aralıkların revize projeleri işaret ettiği, hangi aralıkların ise revize olmasa da



Şekil 3. Vaka çalışmasında yer alan projelerin alt alanlara dağılımı: sayı.
(Break-down of the proposals in the case study by sub-fields: number)

Şekil 4. Proje önerilerinden elde edilen anahtar terimlerin kelime bulutu ile gösterimi
(Word-clouds of the key terms extracted from project proposals)

kavramsal olarak benzer projeler için kullanılabileceği belirlenmiştir. Sonuç olarak, benzerlik skorunun 0.90'ın üzerinde olduğu proje ikililerinin revize olarak, 0.70-0.90 arasında olanların ise kavramsal olarak yakın projeler olarak değerlendirilmesi uygun bulunmuştur. SimDet algoritmasını performansı bu değerler dikkate alınarak gerçekleştirilmiştir.

Vaka çalışması kapsamında ele alınan 1255 proje önerisinin revizyon durumu bilgisi, ARDEB-PBS üzerinden kontrol edilmiş ve sistemdeki kayıtlara göre 148 revize ikilisi, 12 revize üçlüsü ve bir revize dörtlüsünün yer aldığı belirlenmiştir. Her bir revize üçlüsünde 3 çift revize ikilisi, revize dörtlüsünün ise 6 çift revize ikilisi olduğu dikkate alındığında, toplam 190 revize ikilisi bulunmakta, herhangi bir şekilde revizesi olan proje sayısı ise 148x2+12x3+1x4=336 olmaktadır. Vaka çalışmasında, 1255 projenin ikili kombinasyonları ($C(1255,2) = 786.885$) için benzerlik skorları hesaplanmıştır. Skorlar üzerinden gerçekleştirilen kontrollerde, sistemde revize olduğu beyan edilmeyen ancak revize olduğu tespit edilen 10 çift proje önerisi daha belirlenmiştir. Bu durumda, vaka çalışması kapsamında yapılacak ikili karşılaştırmalar sonucunda; 786.885 çift arasından 200 proje çiftinin ve toplam 1255 proje içinden 356 projenin tespit edilmesi beklenmektedir.

SimDet algoritmasının çalıştırılması sonucunda elde edilen sonuçlar ve performans göstergeleri Tablo 5, Tablo 6 ve Tablo 7'de sunulmuştur. Tablo 5'te görüldüğü üzere herhangi bir revize versiyonu da proje kümesinde yer alan 356 projeden 342'si doğru bir şekilde tespit edilmiştir. Gerçekte revize olmayan ama benzerlik skoru 0.90'ın üzerinde olan 11 çift (22 tekil) proje önerisi bulunmaktadır. Ancak, bu öneriler daha detaylı incelendiğine, bu çiftlerden 7'sinin ekiplerinde ortak isimler bulunduğu; diğer 4'ünün ise, konu, içerik ve hatta proje başlığı olarak çok yakın olduğu ve aynı/benzer panellerde değerlendirildiği görülmüştür. Gerçekte revize olmasına rağmen benzerlik skoru 0.90'ın altında kalan 7 çift projede ise, değerlendirme raporunda iletilen hususlar doğrultusunda kapsamlı değişikliklerin yapıldığı görülmüştür. Bu örneklerden birinde, ilk değerlendirmenin ardından literatürde yayımlanan yeni bir çalışma nedeniyle projede ele alınan konunun neredeyse tamamen değiştirildiği tespit edilmiştir. Gerçekte revize olan tüm proje çiftleri dikkate alındığında benzerlik skoru ortalaması 0,972 olarak elde edilmiştir.

Tablo 5. SimDet algoritmasının revize projeleri belirlemedeki performansı: proje sayıları üzerinden
(Performance of SimDet for identifying revised manuscripts: over the number of proposals)

	Gerçekte revize olan projeler	Gerçekte revize olmayan projeler
Revize olduğu tahmin edilen projeler	342	22
Revize olmadığı tahmin edilen projeler	14	877

Tablo 5'te proje önerileri bazında verilen sonuçlar, Tablo 6'da proje ikilileri üzerinden sunulmuştur. Bunun nedeni, SimDet'in aslında ikili karşılaştırmalar sonucunda benzerlik tespitinde bulunmasıdır. Sonuç olarak, toplamda 800.000'e yakın karşılaştırma yapılarak önemli bir çoğunluğu arasında kavramsal bir benzerlik olmadığı tespit edilmektedir ki, bu da algoritmanın bir başarısı olarak kabul edilebilir. Algoritmanın performans göstergeleri, sözü geçen iki yaklaşımı da dikkate alarak Tablo 7'de sunulmuştur. Görüldüğü üzere, ikililer üzerinden gerçekleştirilen analizde, doğruluk oranı %100'e yakındır.

Tablo 6. SimDet algoritmasının revize projeleri belirlemedeki performansı: proje ikilileri üzerinden
(Performance of SimDet for identifying revised manuscripts: over the number of proposal pairs)

	Gerçekte revize olan proje çiftleri	Gerçekte revize olmayan proje çiftleri
Revize olduğu tahmin edilen proje çiftleri	193	12
Revize olmadığı tahmin edilen proje çiftleri	7	786.673

Tablo 7. SimDet algoritmasının revize projeleri belirlemedeki performans göstergeleri
(Performance indicators of SimDet for identifying revised manuscripts)

	Proje sayılarına göre	Proje çiftlerine göre
Doğruluk	97,131%	99,998%
Kesinlik	93,956%	94,146%
Duyarlılık	96,067%	96,500%
F1 Skoru	95,000%	95,309%

4.2.3. SubCla

Vaka çalışmasında kullanılan 1255 proje önerisi, eğitim ve test kümelerine %70-%30 oranında pay edilmiştir. Sonuç olarak, 1255 proje önerisinin 895'i eğitim kümesinde, 360'ı ise test kümesinde kullanılmıştır. Eğitim ve test kümelerinde yer alan proje sayıları ve yüzdelerinin alan bazında dağılımı Şekil 5'te verilmiştir. Görüldüğü üzere, tam sayıya yuvarlama etkisinden dolayı bazı alanlarda %70-%30 dağılımından minör sapmalar bulunmaktadır.

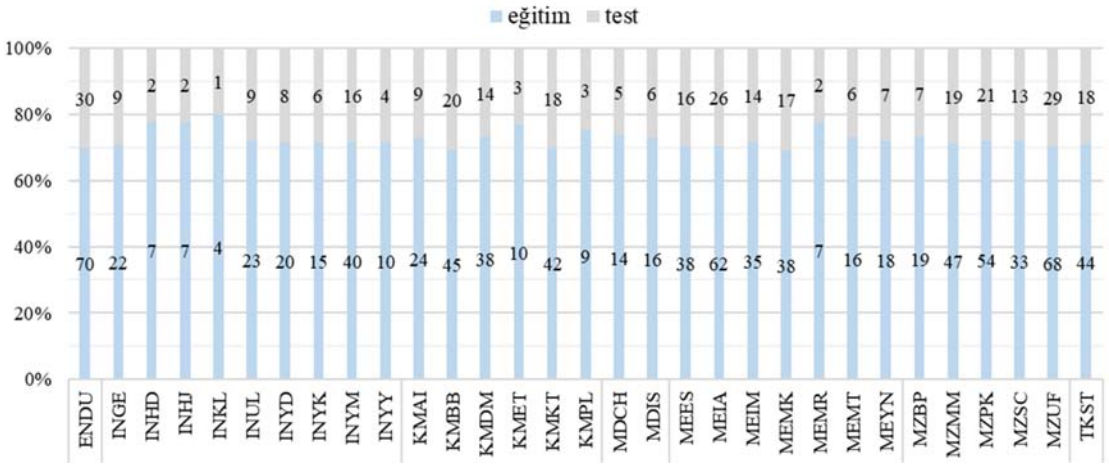
895 proje metni ile eğitilen algoritma ile 360 test proje önerisinin her bir alt alan sınıflarına ne ölçüde uygun olduğuna yönelik uygunluk değerleri hesaplanmıştır. Her proje için, en yüksek üç değer dikkate alınarak hazırlanan doğruluk yüzdeleri Şekil 6 ve Şekil 7'de görülebilir. Sadece ilk tahminler üzerinden hazırlanan performans göstergeleri ise Tablo 8'de sunulmuştur. Bu analiz gerçekleştirilirken, elde edilen sonuçlar ARDEB iş uygulama yazılımlarındaki verilerin detaylı incelenmesiyle çapraz kontrol edilmiştir. Bu kontrollerde, ilk başta yanlış tahmin gibi görünen bir kısım sonucun aslında sistemdeki kayıtlardaki uyuşmazlıklardan kaynaklandığı fark edilmiş ve düzeltilmiştir.

Elde edilen sonuçlar; SubCla algoritmasının, bu kadar çok sayıda tanımlanmış olan alt alan sınıflarını tahminde oldukça başarılı olduğunu göstermektedir (Şekil 6). Görece düşük olan performans gösterilen alt alanların bir kısmı için (INHD, INYK, MEMR) ilgili alandaki veri kümesinin sınırlı olması, bir kısmı için (MZSC, MZUF) ise, konu olarak disiplinler arası projelerin yoğun olması veya aslında kendi içinde de ayrı alt alanlara bölünme ihtiyacı olan alanlar olması hususları dikkate alınabilir. Diğer taraftan, ana alanlar bazında bir değerlendirme yapıldığında, algoritma performansının, beklenildiği üzere, anlamlı bir biçimde daha yüksek olduğu görülmektedir. Bunun nedeni, örneğin INHD alt alanında bulunan bir projenin tahminlemesinde INHD sınıfı yer almasa da inşaat mühendisliğinin başka bir alanının yer alıyor olabilmesidir.

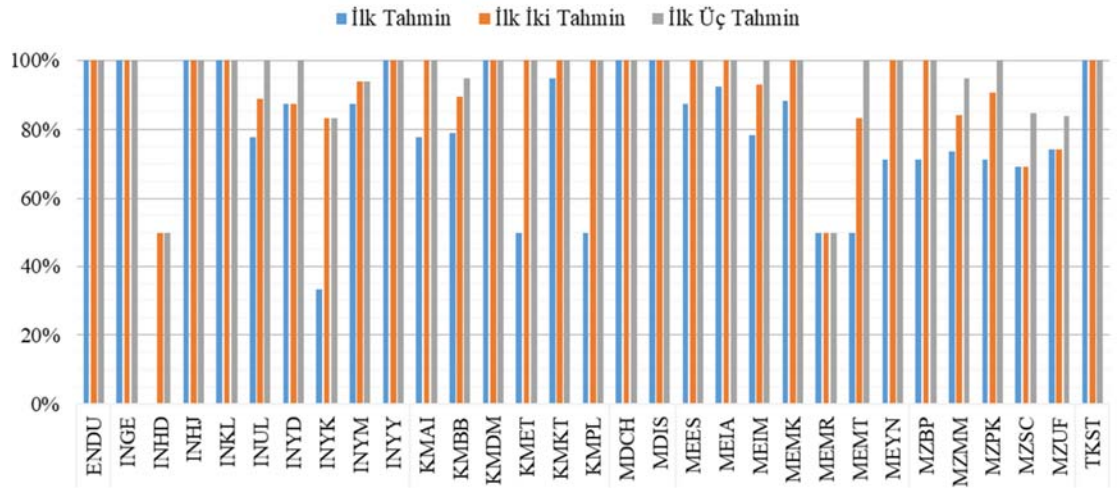
5. Simgeler (Symbols)

5.1. Kısaltmalar (Abbreviations)

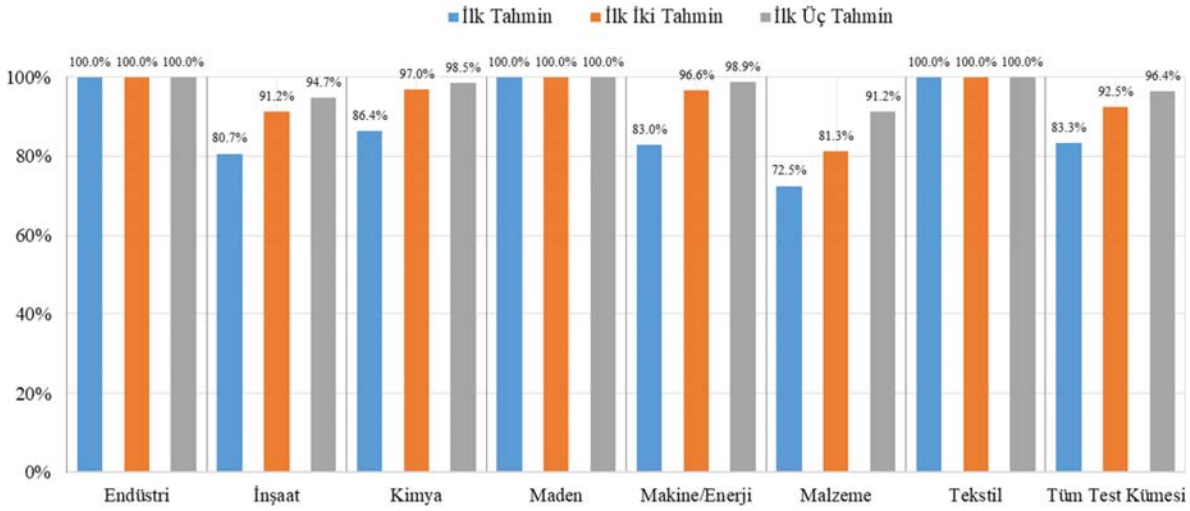
TÜBİTAK Kurumu : Türkiye Bilimsel ve Teknolojik Araştırma
ARDEB : Araştırma Destek Programları Başkanlığı
MAG : Mühendislik Araştırma Destek Grubu



Şekil 5. Makine öğrenmesi için kullanılan eğitim ve test verisinin alan bazında dağılımı: sayı ve yüzde.
(Break-down of the training and test data used in machine learning by main fields: number and percent)



Şekil 6. Naive Bayes sınıflandırıcısının alt alanlar bazında performansı. (Performance of the Naive Bayes classifier by sub-fields)



Şekil 7. Naive Bayes sınıflandırıcısının ana alanlar bazında performansı (Performance of the Naive Bayes classifier by main fields)

Tablo 8. SubCla algoritmasının sınıflama performansı göstergeleri (Classifier performance of SubCla algorithm)

	Doğruluk	Duyarlılık	F1		Doğruluk	Duyarlılık	F1
ENDU	100,0%	96,8%	98,4%	MDCH	100,0%	100,0%	100,0%
INGE	100,0%	90,0%	94,7%	MDIS	100,0%	100,0%	100,0%
INHD	0,0%	-	-	MEES	87,5%	82,4%	84,8%
INHJ	100,0%	66,7%	80,0%	MEIA	92,3%	72,7%	81,4%
INKL	100,0%	100,0%	100,0%	MEIM	78,6%	91,7%	84,6%
INUL	77,8%	100,0%	87,5%	MEMK	88,2%	75,0%	81,1%
INYD	87,5%	77,8%	82,4%	MEMR	50,0%	100,0%	66,7%
INYK	33,3%	50,0%	40,0%	MEMT	50,0%	50,0%	50,0%
INYM	87,5%	100,0%	93,3%	MEYN	71,4%	100,0%	83,3%
INYY	100,0%	100,0%	100,0%	MZBP	71,4%	83,3%	76,9%
KMAI	77,8%	77,8%	77,8%	MZMM	73,7%	73,7%	73,7%
KMBB	78,9%	88,2%	83,3%	MZPK	71,4%	71,4%	71,4%
KMDM	100,0%	71,4%	83,3%	MZSC	69,2%	90,0%	78,3%
KMET	50,0%	100,0%	66,7%	MZUF	74,2%	88,5%	80,7%
KMKT	94,7%	81,8%	87,8%	TKST	100,0%	89,5%	94,4%
KMPL	50,0%	100,0%	66,7%				

ARBİS : Araştırmacı Bilgi Sistemi
ARDEB-PBS : ARDEB Proje Başvuru Sistemi
ARDEB-PTS : ARDEB Proje Takip Sistemi
PYS : Panel Yönetim Sistemi
NLP : DDİ, Doğal Dil İşleme
TF-IDF : Terim Frekansı - Ters Belge Frekansı

6. Sonuçlar (Conclusions)

Ar-Ge'ye ayrılan fonlarda ve bilim ve teknoloji ekosistemindeki paydaş sayısındaki büyüme, bu alanda ülkemizdeki ana aktör olan TÜBİTAK tarafından sunulan destek programlarının çeşitliliğini artırmış ve toplam başvuru sayısında da önemli bir artışı beraberinde getirmiştir. TÜBİTAK, artan iş yükünü yönetmek ve tüm paydaşların ihtiyaçlarını etkin bir şekilde karşılayabilmek üzere bir takım bilgi yönetim sitemleri geliştirmiştir. Bu sistemler; başvuru, değerlendirme ve izleme süreçlerinin işletilmesi ile birlikte iş zekâsı modülleri ile sorgulanarak talep edilen istatistik ve raporların hazırlanmasında kullanılmaktadır. Ancak, gerçekleştirilen işlemlerin hacminin önemli derecede artmış olması, kullanıcılara ve karar vericilere her aşamada yardımcı olabilecek akıllı karar destek sistemlerinin geliştirilmesini elzem kılmaktadır. Bu çalışma, ilk adımı [2]'de atılan bu girişimlerin devamı olarak proje öneri metinlerinden anahtar terimlerin çıkarımı, elde edilen kavram vektöründen benzerlik tespiti ve konu sınıflandırması yapılması problemlerini ele almaktadır.

Çalışma kapsamında, Türkçe akademik metinler için anahtar kavramların çıkarımı, benzerlik tespiti ve konu sınıflandırması bağlamında literatürdeki önemli bir boşluğun doldurulduğu düşünülmektedir. Literatür bölümünde de değinildiği üzere; Türkçe metinlerin (film/otel/restoran sitelerinde yer alan yorumlar, web sitesi/tweet/e-posta içerikleri veya haber metinleri) sınıflandırılmasına yönelik çalışmalar olsa da akademik metinlerin sınıflandırılmasına yönelik bir çalışmaya rastlanmamıştır. Böyle bir çalışma, akademik metinlerin içeriklerine özgü bir ön işleme kütüphanesi gerektirmektedir. Mevcut çalışmada, TÜBİTAK ARDEB altında yer alan Mühendislik Araştırma Destek Grubu'na sunulmuş olan projeler incelenmiş, gerçekleştirilen ön analizler sonucunda mühendislik alanında kullanılan kelime ve terimler dikkate alınarak bir normalizasyon kütüphanesi oluşturulmuştur. Ardından, 1255 proje önerisi içeren bir veri seti hazırlanarak ilk aşamada öneri metinlerinden anahtar terimlerin çıkarımı yapılmış ve vektörel formda elde edilen bu veri kullanılarak bir benzerlik algoritması geliştirilmiştir. Sonrasında ise, bir makine öğrenmesi algoritması olan Naïve Bayes sınıflandırma yaklaşımı kodlanarak, yine aynı vektörel yapı üzerinden projelerin 31 farklı mühendislik alt alanından hangi kategoriye ait olduğunun tahmin edilmesi sağlanmıştır. Geliştirilen algoritmanın hem revize hem de içerik olarak birbirine yakın önerilerin tespitinde ihtiyacı tamamen karşıladığı belirlenmiştir. Sınıflama açısından ise ilk tahminde %83,3, ilk iki tahminde %92,5 ve ilk üç tahminde %96.4'lük bir başarım sağlanmıştır.

Sonraki aşamada, işlerliği açık bir şekilde ortaya konulan makine öğrenmesi yaklaşımının değerlendirici belirleme süreçlerinde de kullanılmasına yönelik algoritmaların geliştirilmesi planlanmaktadır.

Teşekkür (Acknowledgement)

Bu çalışmanın gerçekleştirilmesi için imkân sağlayan TÜBİTAK'a; değerli görüşleri ve katkılarından dolayı ARDEB MAG çalışanlarına ve Prof. Dr. Lale Özbakır'a teşekkür ederim.

Kaynaklar (References)

1. Khan A., Baharudin B., Lee L., Khan K., A review of machine learning algorithms for text-documents classification. Journal of Advances in Information technology, 1 (1), 4–20, 2010.
2. Kat B., An Algorithm and a Decision Support System for the Panelist Assignment Problem: The Case of TUBITAK. Journal of the Faculty of Engineering and Architecture of Gazi University, 36 (1), 69–88, 2021.
3. Çağataylı M., Çelebi E., The effect of stemming and stop-word-removal on automatic text classification in Turkish language. Arik S, Huang T, Lai WK, Liu Q, editors. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9489, 168–76, 2015.
4. Deniz A., Kiziloz H.E., Effects of various preprocessing techniques to Turkish text categorization using n-gram features. 2nd International Conference on Computer Science and Engineering, UBMK 2017, 655–60, 2017.
5. Öztürkmenoğlu O., Alpkoçak A., Comparison of different lemmatization approaches for information retrieval on Turkish text collection. INISTA 2012 - International Symposium on INnovations in Intelligent SysTems and Applications, 1–5, 2012.
6. Tahiroğlu B.T., Lemmatization and a lemmatization application for Turkish: elemanTR. RumeliDE Journal of Language and Literature Research, 24, 475–86, 2021.
7. Salur M., Aydın İ., Jamous M., An ensemble approach for aspect term extraction in Turkish texts. Pamukkale University Journal of Engineering Sciences, 28 (5), 769-776, 2021.
8. Yıldırım S., Yıldız T., A comparative analysis of text classification for Turkish language. Pamukkale University Journal of Engineering Sciences, 24 (5), 879–86, 2018.
9. Kat B., Bilimsel Çalışmaların Benzerliğinin Tespitinde Kullanılan Araçların Araştırılması: ARDEB'e Sunulan Proje Önerileri İçin Uygun Modelin Ve Uygulama Yol Haritasının Belirlenmesi, TÜBİTAK, 2015.
10. Vrbanec T., Mestrovic A., The struggle with academic plagiarism: Approaches based on semantic similarity. 40th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2017 - Proceedings. 870–5, 2017.
11. Chong M., Specia L., Mitkov R., Using natural language processing for automatic detection of plagiarism, Proceedings of the 4th International Plagiarism Conference (IPC-2010), 2010.
12. Gomaa W.H., Fahmy A.A., A Survey of Text Similarity Approaches. International Journal of Computer Applications, 68 (13),975–8887, 2013.
13. Dharmadhikari S.C., Ingle M., Kulkarni P., Empirical Studies on Machine Learning Based Text Classification Algorithms. Advanced Computing, 2 (6), 161, 2011.
14. Kandimalla B., Rohatgi S., Wu J., Giles C.L., Large Scale Subject Category Classification of Scholarly Papers With Deep Attentive Neural Networks, Frontiers in Research Metrics and Analytics. 5, 600382, 2021.
15. Kadhim A.I., Survey on supervised machine learning techniques for automatic text classification, Artificial Intelligence Review, 52 (1), 273–92, 2019.
16. Gurcan F., Multi-Class Classification of Turkish Texts with Machine Learning Algotirthms, ISMSIT 2018 - 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies, Proceedings, 1–5, 2018.
17. Koksal O., Tuning the Turkish Text Classification Process Using Supervised Machine Learning-based Algorithms, International Conference on INnovations in Intelligent SysTems and Applications (INISTA), 2020.
18. Küçük D., Arıcı N., A Literature Study on Deep Learning Applications in Natural Language Processing, International Journal of Management Information Systems and Computer Science, 2 (2),76–86, 2018.
19. Kilimci Z.H., Akyokus S., The Evaluation of Word Embedding Models and Deep Learning Algorithms for Turkish Text Classification, UBMK 2019 - Proceedings, 4th International Conference on Computer Science and Engineering, 548–53, 2019.
20. Kilimci Z.H., Akyokus S., Deep learning- and word embedding-based heterogeneous classifier ensembles for text classification, Complexity, 2018.
21. Aydin G., Hallaç İ.R., Türkçe Metinlerde Otomatik Konu Tespiti, Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 33 (2), 599–606, 2021.
22. Güran A., Akyokuş S., Güler Bayazıt N., Gürbüz M.Z., Turkish Text Categorization Using N-Gram Words, International Symposium on INnovations in Intelligent SysTems and Applications (INISTA), 369–73, 2009.
23. Erşahin B., Aktaş Ö., Kilinç D., Erşahin M., A hybrid sentiment analysis method for Turkish. Turkish Journal of Electrical Engineering & Computer Sciences, 27, 1780–93, 2019.
24. Kaya M., Fidan G., Toroslu I.H., Sentiment analysis of Turkish political news, Proceedings - 2012 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2012, 174–80, 2012.
25. Boynukalın Z., Emotion Analysis of Turkish texts by using machine learning method, METU, 2012.
26. Kaşıkçı T., Gökçen H., Metin Madenciliği ile E-Ticaret Sitelerinin Belirlenmesi, Journal of Information Technologies, 7 (1), 25–32, 2014.

27. Kaynar O., Görmez Y., Yildiz M., Albayrak A. Sentiment Analysis with Machine Learning Techniques, International Artificial Intelligence and Data Processing Symposium, 2016.

28. Coban O., Ozyer B., Ozyer G.T., Sentiment analysis for Turkish Twitter feeds, 23rd Signal Processing and Communications Applications Conference, SIU 2015, 2388–91, 2015.

29. Uysal A.K., Gunal S., The impact of preprocessing on text classification. Information Processing & Management. 50 (1), 104–12, 2014.

30. Aydin G., Hallac I.R., Document Classification Using Distributed Machine Learning. arXiv preprint arXiv:180203597, 166–9, 2018.

31. Yau C.K., Porter A., Newman N, Suominen A., Clustering scientific documents with topic modeling, Scientometrics, 100 (3), 767–86, 2014.

32. Kim S.W., Gil J.M., Research paper classification systems based on TF-IDF and LDA schemes, Human-centric Computing and Information Sciences, 9 (1), 2019.

33. Suominen A., Toivanen H., Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification, Journal of the Association for Information Science and Technology, 67 (10), 2464–76, 2016.

34. Kılınç D., Borandağ E., Yücalar F., Tunali V., Şimşek M., Özçift A., KNN Algoritması ve R Dili ile Metin Madenciliği Kullanılarak Bilimsel Makale Tasnifi, Marmara Fen Bilimleri Dergisi, 28 (3), 89–94, 2016.

35. Raschka S., Naive Bayes and Text Classification I - Introduction and Theory, arXiv preprint arXiv:14105329, 2014.

36. Jurafsky D., Martin J.H., Speech and Language Processing, 3rd edn, Prentice Hall, US, 2019.

37. Huang Y., Li L., Naive Bayes classification algorithm based on small sample set, CCIS2011 - Proceedings: 2011 IEEE International Conference on Cloud Computing and Intelligence Systems, 34–9, 2011.

38. Chandrasekar P., Qian K., The impact of data preprocessing on the performance of a naive bayes classifier, IEEE 40th Annual Computer Software and Applications Conference (COMPSAC), 2, 618-619, 2016.

39. Noyan T., Kuncan F., Tekin R., Kaya Y., A new content-free approach to identification of document language: Angle Patterns, Journal of the Faculty of Engineering and Architecture of Gazi University, 37 (3), 1277–92, 2022.