Journal of the Faculty of Engineering and Architecture of Gazi University 38:3 (2023) 1375-1387

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 Turkish language processing based method for classifying R&D projects

Serdar Koçak[1], Yusuf Tansel İç[2]*, Mustafa Sert[3], Berna Dengiz[2]

[1]The Scientific and Technological Research Council of Turkey, 06100 Kavaklıdere, Ankara, Türkiye
[2]Department of Industrial Engineering, Faculty of Engineering, Baskent University, 06810, Etimesgut, Ankara, Türkiye
[3]Department of Computer Engineering, Faculty of Engineering, Baskent University,06810, Etimesgut, Ankara, Türkiye

Highlights:

- Text mining methods were used in the classification of R&D projects
- Results of different text mining classification algorithms are compared with each other
- Natural language processing methods have been applied to a new field

Keywords:

- Text classification
- Text mining
- Natural language processing (NLP)
- Reviewer assignment
- Project selection

Article Info:

Research Article
Received: 02.03.2021
Accepted: 07.06.2022

DOI:

10.17341/gazimmfd.889395

Correspondence:

Author: Serdar Koçak
e-mail: serdar.kocak@tubitak.gov.tr
phone: +90 506 234 9494

Graphical/Tabular Abstract

In this study, in order to design the classifier for assigning reviewer to R&D Projects, approximately 500 R&D project abstracts, purpose and scope texts, as well as similar R&D articles were selected and the python programming language "tensorflow, numpy, os, csv" libraries are used. The words in the datasets consisting of selected R&D projects were separated into their roots, converted to lowercase letters, cleared of punctuation marks and symbols, meaningless unnecessary words were discarded, and stopped words and general expressions were removed. The obtained datasets were used in Word2Vec's CBoW and Skip-Gram model training datasets needed by CNN. It has been shown that high accuracy rates are obtained by trying to analyze the results by comparing the datasets and methods used in the studies. The general flow of the proposed system is given in Figure A.

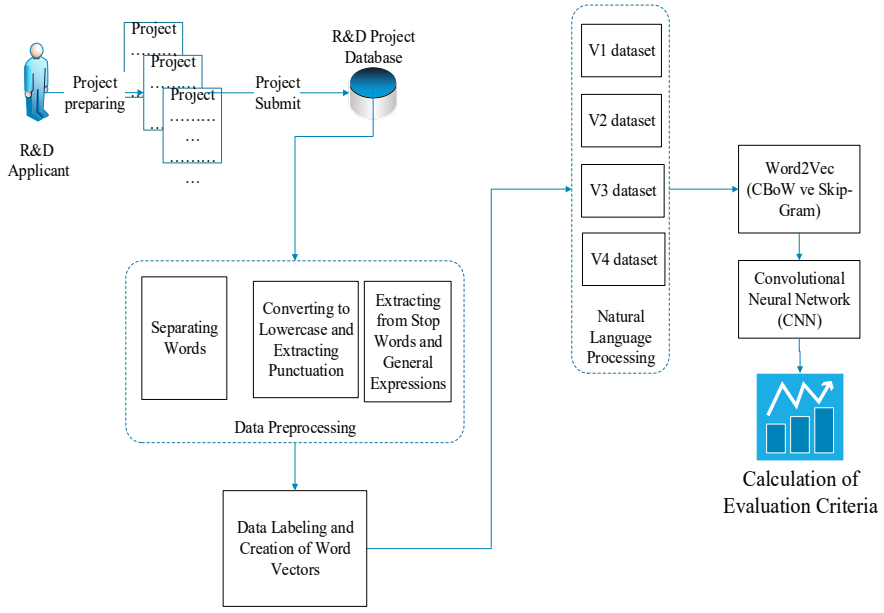


Figure A. Proposed system overview

Purpose: Classification procedures were made using text mining techniques and methods, and algorithms used and accuracy rates were compared in order to assign R&D projects applications to appropriate reviewer fields of activity.

Theory and Methods: Convolutional Neural Network (CNN) models were tried to be created in order to classify the features by using the word representation method and automatic feature learning approach, which is one of the natural language techniques "Word2Vec".

Results: The best results were obtained in the experimental studies conducted on the data set of R&D projects and articles with R&D project content selected from the TUBİTAK Dergipark website.

Conclusion: It has been ensured to eliminate the problems of reviewer assignments in the selection of R&D projects and to find the best project classification method. With the information obtained, the input of the decision support system to be created in the future was tried to be created.

Journal of the Faculty of Engineering and Architecture of Gazi University 38:3 (2023) 1375-1387

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

Ar-Ge projelerinin sınıflandırılması için doğal Türkçe dil işleme tabanlı yöntem

Serdar Koçak[1], Yusuf Tansel İç[2]*, Mustafa Sert[3], Berna Dengiz[2]
[1]Türkiye Bilimsel ve Teknolojik Araştırma Kurumu, 06100 Kavaklıdere, Ankara, Türkiye
[2]Başkent Üniversitesi, Mühendislik Fakültesi, Endüstri Mühendisliği Bölümü, 06810, Ankara, Türkiye
[3]Başkent Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 06810, Ankara, Türkiye

Ö N E Ç I K A N L A R

- Ar-Ge projelerinin sınıflandırılmasında metin madenciliği yöntemleri kullanılmıştır
- Farklı metin madenciliği sınıflandırma algoritmaları sonuçları birbiriyle karşılaştırılmıştır
- Doğal dil işleme yöntemleri yeni bir alana uygulanmıştır

Makale Bilgileri
Araştırma Makalesi
Geliş: 02.03.2021
Kabul: 07.06.2022

DOI:
10.17341/gazimmfd.889395

Anahtar Kelimeler:
Metin sınıflandırma, metin madenciliği, doğal dil işleme (DDİ), hakem atama, proje seçimi, evrişimsel sinir ağları, word2vec

ÖZ

Birçok farklı sektörde metin halinde bulunan verilerden istenilen bilgilerin elde edilmesi için doğal dil işleme, metin madenciliği ve derin öğrenme yöntemleri kullanılmaktadır. Son zamanlarda artan Ar-Ge proje sayıları ve farklılaşan proje faaliyet alanları ile birlikte Ar-Ge projelerinin ait olduğu araştırma alanlarının belirlenmesi ve bu araştırma alanlarına uygun hakemlerin tespitinde yaşanan sıkıntılar nedeniyle projelerin desteklenme süreçleri olumsuz etkilenebilmektedir. Bu makalede, Ar-Ge projelerinin sınıflandırılması amacıyla öncelikli olarak çalışmanın gerçekleştirildiği veri tabanındaki veriler temizlenmiş ve doğal dil tekniklerinden biri olan "Word2Vec" kelime temsili yöntemi ile otomatik özellik öğrenme yaklaşımı kullanılarak özelliklerin sınıflandırılması amacıyla Evrişimsel Sinir Ağları (CNN-Convolutional Neural Network) modelleri oluşturulmaya çalışılmıştır. TUBİTAK Dergipark sitesinden seçilen son on yılda başvurusu yapılmış, sınıfları belli olan Ar-Ge projeleri ve Ar-Ge proje içeriğine sahip makalelerden oluşan veri kümesi üzerinde yapılan deneysel çalışmalardan elde edilen değerlendirme sonuçları ile diğer klasik algoritmalar karşılaştırılmış ve özellikle Word2Vec modellerine sahip CNN'lerin daha etkili sonuçları ürettiği birçok performans parametresi ile gösterilmiştir.

Natural Turkish language processing based method for classifying R&D projects

H I G H L I G H T S

- Text mining methods were used in the classification of R&D projects
- Results of different text mining classification algorithms are compared with each other
- Natural language processing methods have been applied to a new field

Article Info
Research Article
Received: 02.03.2021
Accepted: 07.06.2022

DOI:
10.17341/gazimmfd.889395

Keywords:
Text classification, text mining, natural language processing (NLP), reviewer assignment, project selection, convolutional neural networks, word2vec

ABSTRACT

Natural language processing, text mining and deep learning methods are used to obtain the desired information from textual data in many different sectors. Recently increasing number of R&D projects and differentiated project activity areas, as well as the determination of research areas to which R & D projects belong and the determination of reviewer suitable for these research areas, the support processes of projects may be negatively affected. In this article, in order to classify R&D projects, the data in the database where the study was carried out was primarily cleared and in order to classify the features by using the word representation method and automatic feature learning approach "Word2Vec", which is one of the natural language techniques, Convolutional Neural Network (CNN) models have been tried to be created. The evaluation results obtained from experimental studies on the data set of R&D projects and articles with R&D project submitted in the last ten years content selected from the TUBITAK Dergipark site were compared with other classic algorithms and showed that CNNs with Word2Vec models in particular produce more effective results with many performance parameters.

*Sorumlu Yazar/Yazarlar / Corresponding Author/Authors : *serdar.kocak@tubitak.gov.tr, yustanic@baskent.edu.tr, msert@baskent.edu.tr, bdengiz@baskent.edu.tr / Tel: +90 506 234 9494

1. Giriş (Introduction)

Son yıllarda gelişen teknoloji ve rekabetçi piyasa koşulları ile birlikte, ülkelerin Araştırma-Geliştirme (Ar-Ge) projelerine destekleri ve Ar-Ge projelerine yapılan harcamaların Gayri Safi Yurtiçi Hasıla (çerisindeki payları gittikçe artmaktadır. Ülkemizde ise Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK) tarafından orta-uzun vadeli hedefler doğrultusunda teknoloji alanlarını öncelikli olmak üzere bu alanlardaki projeler çıktı odaklı olarak desteklemektedir. TÜBİTAK'a 2019 yılında üniversite ve sanayi kuruluşlarınca Ar-Ge desteği amacıyla yaklaşık 5.000 adet proje başvurusu yapılmıştır [1]. Bu projelerin değerlendirilmesi için Araştırma Bilgi Sistemi'nde (ARBİS) kayıtlı yaklaşık 1.000.020 kullanıcı içerisinden uygun olanlar projeleri değerlendirmek üzere seçilmektedir [1]. Ar-Ge projeleri değerlendiren hakemlerin belirlenmesinden önce proje konusu uygun olarak disiplinlerin ve faaliyet alanlarının belirlenmesi gerekmektedir. İleri teknoloji ve buna bağlı olarak verilerin büyük boyutlara ve miktarlara ulaşması herhangi bir araç kullanmadan insan yeteneğiyle verilerden anlamlı bilgilerin çıkartılmasını imkânsız hale getirmektedir. Özel sektör kuruluşlarının araştırma-teknoloji geliştirme ve yenilik faaliyetlerini desteklemek amacıyla ülkemizde bulunan destekleyici kurum ve kuruluşların her yıl binlerce Ar-Ge projesi değerlendirdiği düşünüldüğünde; Ar-Ge projelerini destekleyen kuruluşlar için Ar-Ge projelerinin değerlendirilmesi amacıyla uygun hakemlerin atanabilmesi için doğru şekilde sınıflandırılması ve değerlendirilmesi büyük önem kazanmaktadır.

Günümüzde verileri sınıflandırma ve verilerin birbirleri arasındaki ilişkilerin bulunması amacıyla yazılım araçları ve istatistik tabanlı yöntemler kullanılarak birçok analiz yapılabilmekte, bu süreçte anlamlı bilginin elde edilmesi için birçok yöntem kullanılmaktadır. Genel olarak veri sınıflandırma; özellik çıkarma, boyut küçültme, sınıflandırıcı seçimi ve değerlendirme bölümlerinden oluşan 4 farklı aşamada uygulanmaktadır [2]. Çoğu metin sınıflandırma ve belge sınıflandırma sistemi; doküman, paragraf, cümle ve alt cümle seviyesinde yapılandırılır. Özellik çıkarımı amacıyla Terim Frekansı-Ters Belge Frekansı (TF-IDF), Terim Frekansı (TF), Word2Vec ve Global Vektörler (GloVe) yöntemleri kullanılmaktadır. Kelime temsil (Word Embedding) teknikleri, sözcük dizilerinden oluşumlarını ve birlikte ortaya çıkma bilgilerini dikkate alarak öğrenmekte iken, ağırlıklı kelimeler (Weighted Words) belgelerdeki kelimeleri saymaya dayanır ve kelime gösteriminin basit bir puanlama şeması olarak kullanılabilir. Bunların yanı sıra hesaplama sürelerini iyileştirmek ve bellek karmaşıklığını azaltmak amacıyla Temel Bileşen Analizi (PCA), Doğrusal Diskriminant Analizi (LDA) ve Negatif Olmayan Matris Faktorizasyonu (NMF) gibi boyutsal küçültme teknikleri de kullanılmaktadır. Sınıflandırma yöntemlerine bakılacak olunursa Rocchio sınıflaması gibi geleneksel metin sınıflandırma yöntemleri ile birlikte en basit sınıflandırma algoritmalarından biri olarak görülen Lojistik Regresyon (LR) kullanılmaktadır. Naive Bayes Sınıflandırıcısı (NBC) hesaplama açısından pratik ve çok az bellek gerektiren bir algoritma olup parametrik olmayan teknikler içerisinden ise K-En Yakın Komşu (KNN) ve Destek Vektör Makinesi (SVM) gibi sınıflandırıcılar bulunmaktadır. Ayrıca Karar Ağaçları ve Rastgele Orman gibi ağaç temelli sınıflandırıcılar sınıflandırma probleminde kullanılmaktadır. Günümüzde ayrıca grafik sınıflandırmalar ve koşullu rastgele alanlar (CRF) gibi bir sınıflandırıcılar kullanılmaya başlanmış [2], son zamanlarda ise derin öğrenme yaklaşımları, görüntü sınıflandırması, doğal dil işleme, yüz tanıma vb. yöntemlerin kullanılmasıyla birlikte makine öğrenme algoritmalarına kıyasla daha başarılı sonuçlar elde edilmeye başlandığı görülmektedir [3]. Değerlendirme aşamasında, metin sınıflandırma işlemlerinde yapılan tahminlerin kalitesini ölçmek için bazı parametrik ölçütler kullanılmaktadır. Bu ölçütlerden hassasiyet (recall, ρ), belirli bir kategori için doğru sınıflandırılan dokuman yüzdesini, kesinlik (precision, π), belirli bir kategori için sınıflandırılan dokuman yüzdesini ifade etmektedir. F ölçütü (f score, F) ise hassasiyet ile kesinlik değerlerini birlikte ele alan bir değerlendirme kıstasıdır. ROC alanı veya diğer adıyla ROC eğrisi, sınıflandırıcı başarımı değerlendirmede kullanılan etkili bir ölçüttür [4]. Aşağıda özellik çıkarma, boyut küçültme, sınıflandırıcı seçimi ve değerlendirmeler ile ilgili olarak genel durum bir Tablo 1'de listelenmeye çalışılmıştır.

Doğal dil işleme metin madenciliği alanında metinlerin işlenmesinde yaygın olarak kullanılmaktadır. Morfolojik, fonolojik, sözcüksel, sözdizimsel, anlambilimsel, pragmatik düzeydeki metin ve konuşma dillerinden anlam çıkarmak için kullanılan farklı doğal dil seviyeleri bulunmakta olup, kelime temsil yöntemleri, sözcüklerin ve ifadelerin vektörlere dönüştürüldüğü doğal dil işlemede kullanılır ve performans artışı gösterdiği bilinmektedir [5]. 2013 yılında Mikolov tarafından kelime temsili tekniği olarak Skip-gram ve CBOW (continuous bag-of-word) isimli iki öğrenme algoritmasını içeren Word2Vec yöntemi önerilmiştir [6]. Naive Bayes, Random Forest, SVM, Karar Ağacı, vb. farklı metin sınıflandırmaları için yaygın olarak kullanılmış ancak son yıllarda geleneksel sınıflandırıcı algoritmaların yanı sıra başlangıçta bilgisayarlar için bulunan Evrişimsel Sinir Ağları (CNN-Convolutional Neural Network) modellerinin doğal dil işleme için çok daha etkili sonuçlar verdiği gösterilmiş ve semantik ayrıştırmada daha iyi sonuçlar elde edilebileceği ileri sürülmüştür [7].

Yapılan bu çalışma ile TUBİTAK Dergipark sitesinden seçilen son on yılda başvurusu yapılmış Ar-Ge proje içeriğine sahip makaleler, veri

Tablo 1. Metin Madenciliği Yöntemleri Genel Bakış (Text Mining Methods Overview)

Metin ve Doküman Özellik Çıkarım Yöntemleri	Boyut Küçültme Yöntemleri	Sınıflandırma Yöntemleri	Değerlendirme Yöntemleri
Weighted Words	Principal Component Analysis (PCA)	Rocchio Algorithm	F Ölçütü
TF-IDF	Linear Discriminant Analysis (LDA)	Bagging And Boosting	Matthew Correlation Coefficient (MCC)
Word2Vec	Non-Negative Matrix Factorization (NMF)	Logistic Regression (LR)	Receiver Operating Characteristics (ROC)
GloVe (Pre-Trained)	Random Projection	Naive Bayes Classifier (NBC)	Area under Curve (AUC)
GloVe (Trained)	Autoencoder	K-Nearest Neighbor (KNN)	
FastText Contextualized Word Representations	t-Distributed Stochastic Neighbor Embedding (t-SNE)	Support Vector Machine (SVM)	
		Decision Tree Classifier (DTC)	
		Random Forest	
		Conditional Random Field (CRF) Deep Learning	

kümesi üzerinde kelimelere ayrıma, küçük harfe çevirme, noktalama işaretlerinden arındırma, durak harf ve genel ifadelerden ayırma işlemleri yapılarak ön işlemeden geçirilmiş, elde edilen veriler çeşitli doğal dil işleme ve metin madenciliği algoritmaları kullanılarak sınıflandırılmıştır. Geleneksel sınıflandırıcılar ile yapılan denemelerin yanı sıra bu çalışmada Evrişimsel Sinir Ağları (CNN-Convolutional Neural Network) gibi derin sinir ağları eğitimi sağlanarak sınıflandırma yapılmıştır. Yapılan bu sınıflandırmalar geleneksel sınıflandırma algoritmaları ile elde edilen sonuçlar ile karşılaştırılmış ve Word2Vec + CNN metodunun Ar-Ge projeleri sınıflandırma işlemlerindeki performans değerleri bulunmaya çalışılmıştır. Ar-Ge projelerinin sınıflandırılmasına ilişkin geçmişte yapılan çalışmalarla bir karşılaştırma çalışması da yapılmış olup, önerilen metodun sınıflandırma işlemlerinde doğruluk oranlarının önemli seviyede iyileştirdiği gösterilmiştir. Ar-Ge projelerinin sınıflandırılması özelinde literatürde yapılan çalışmalar arasında word2vec+CNN temelli bir çalışmanın bulunmadığı görülmektedir. Yapılan bu çalışma sonucunda, Ar-Ge projeleri için hakem seçiminde kullanılmak üzere çok büyük metinsel veriler içerisinden anlamlı sonuçlar çıkartabilen, bu sonuçları hızlı ve etkin bir şekilde hakemlerin performanslarının ölçümüne yönelik olarak kullanılabilecek bir yapıya dönüştürebilecek bir model sunulmaktadır.

Makalenin ikinci bölümünde literatürde Ar-Ge projelerin sınıflandırma konularına benzer çalışmalar incelenmiş, üçüncü bölümde çalışmanın temelini oluşturan yöntem anlatılmış, dördüncü bölümde projelerin sınıflandırılması için önerilen model aktarılmış ve son bölümde ise çalışma ve denemelerin sonuçları paylaşılmıştır.

2. Literatür Taraması (Literature Survey)

Metin sınıflandırma, veri tabanından alınan metinsel verileri veya bilgilerden anlamlı sonuçlar alınmasını ele alan önemli bir yaklaşımdır. Metin madenciliği teknikleri, e-posta filtreleme, doküman yönetimi, müşteri ihtiyaçları belirleme gibi çeşitli uygulama alanlarında yaygın olarak kullanılmaktadır. Metin madenciliği tekniklerinin birçok farklı uygulama içerisinde ve farklı aşamalarda proje seçimi veya proje sınıflandırmalarında da (Tablo 2) kullanıldığı görülmektedir. Ar-Ge proje seçimi için [8] çalışmasında melez bir bilgi ve model sisteminin tasarımını ve uygulaması ortaya koymuştur. Karar modelleri ve bilgi kuralları bütün karar destek sürecini desteklemek için entegre edilmiş ve kullanılmıştır. Sistem, Çin'de bulunan Internet-based Science Information System (ISIS) isimli uygulamada çalıştırılmıştır. Cook vd. [9] çalışmasında hakemlerin değerlendirme sıralamaları yoluyla nasıl alt kümelere atandığı problemini ele almıştır. Önerilen çözüm yaklaşımı bir veya birden fazla hakem tarafından değerlendirilecek proje önerilerinin sayısını maksimize eden atama yapılmasıdır. Yaklaşımı uygulamak için bir tamsayı programlama seti-kaplama modeli ve bir sezgisel prosedür sunulmuştur. Hettich vd. [10] çalışmasında program direktörlerine yardımcı olması için görevlendirilecek hakemlerin tespit edilmesinde bir prototip uygulama önermiştir. Uygulama, program direktörlerinin projeleri gruplara ayırmasına ve projeler için hakemler bulmasına yardımcı olmaktadır. Bu görevleri yerine getirmek için, hem tekliflerin konularını, hem de hakemlerin uzmanlığını öğrenmek için projelerin içerisinde metinlerden bilgi çıkarılmaktadır.

Ar-Ge projelerinin tarama süreci, sezgisel ölçütlerden ziyade akıllı sistemlere dayanması gerektiği Choi vd. tarafından [11] çalışmasında belirtilmiş ve bu amaçla önceki veya devam eden projelerin kopyalarını belirlemek ve filtrelemek için kullanılabilecek akıllı ve sistematik bir Ar-Ge proje tarama sistemi önerilmiştir. Tarama algoritması, yeni bir Ar-Ge önerisi ile tamamlanmış ve mevcut Ar-Ge projeleri havuzu içerisinde içeriklerin ve maddenin tutarlılığına dayandığını bu amaçla metin madenciliği metodolojisi örnek bir gerçek veri kümesi üzerinde uygulamışlardır.

2008 yılında Sun vd. [12] makalesinde Ar-Ge proje seçimi hakemlerini değerlendirmek için bir grup karar destek yaklaşımı sunulmuştur. Analitik hiyerarşi sürecine (AHP), puanlama yöntemine ve bulanık dilbilimsel işlemeye dayanmakta ve önerilen yöntemin gösterimine yönelik bir grup karar destek sistemi tasarlanmış ve uygulanmıştır. Hakem atama problemi için karma bilgi ve model yaklaşımı yine Sun vd. [13] tarafından çalışmalarında önerilmiştir. Karar modeli ve bilgi kuralları kullanılarak hakem atama problemini optimize edilmiştir. 2009 yılında Fan vd. [14] projeler nitelikleri ve öneri sınıflarına göre gruplandırılır ve genetik algoritma ile belirlenen grupta arama geliştirmektedir. Ayrıca, önerilen yaklaşımı desteklemek için ilgili bir sistem tasarlanmış ve geliştirilmiştir. Araştırma alanlarında benzerliklerine dayanan araştırma önerileri için yeni bir ontoloji tabanlı metin madenciliği yaklaşımı Liu vd. [15] tarafından sunulmuştur. Yöntem, hem İngilizce hem de Çince metinlerle araştırma önerilerinin kümelenmesi için etkili çalıştığı görülmüştür. Ayrıca, yöntemde başvuru sahiplerini coğrafi bölgelere göre teklifleri dengeleme özelliklerini dikkate alan bir optimizasyon modelini de içermektedir. Xu vd. [16] çalışmasında temel fikir önce projeleri ve hakemleri tanımlama, projeleri kendi bölümlerine göre gruplama ve atama işlemidir. Böylelikle değerlendiricilere proje karar verme sürecini kolaylaştırmak için önerilen yaklaşıma dayalı bir sistem geliştirilmiştir. 2012 yılına gelindiğinde çok dilli desteklerle Ar-Ge proje yönetim sistemlerinin geliştirilmesine yönelik ontoloji temelli bir metin madenciliği metodoloji Ma vd. [17] tarafından sunulmuştur. Yapılan çalışmada araştırma alanlarındaki benzerliklerine dayanarak kümeleme önerilerine yönelik ontoloji temelli üç dili destekleyen yeni bir metin madenciliği yaklaşımı sunulmaktadır.

Bunun yanı sıra ontoloji tabanlı Metin Madenciliği Yöntemi, Araştırma Proje Önerilerinin yanı sıra, hakemleri sınıflandırmak ve daha sonra araştırma disiplini alanlarına göre gruplandırmak ve uygun grubu atamak için farklı bir çalışma Kaur vd. [18] tarafından sunulmuştur. Bu yaklaşım, projelerin ve hakemlerin sayısının artmasıyla araştırma projesi tekliflerinin seçilmesi için etkili bir yol sunmaktadır. Benzer şekilde Preethi vd.[19], Arunachalam vd. [20] ve Gunjal vd. [21] tarafından metin madenciliği yaklaşımı ile ontoloji tabanlı proje gruplandırma çalışmaları yapılmıştır. Anlamsal olmayan ve yaklaşımı olmayan metin kümeleme yöntemleri daha az doğruluk sağladığı görülmektedir. Yöntem ayrıca başvuruları coğrafi bölgelere göre dengelemek için bir optimizasyon modelini içermektedir. Gruplandırılmış projeler daha sonra sistem üzerinden kendi kendine inceleme için uygun araştırma uzmanlarına atanabilmektedir. 2014 yılında Silva vd. [22] karmaşık gözden geçirme atama sürecini yönetilebilir alt işlemlere ayırarak yeni bir karma işlem analizi yaklaşımı önermektedir. Yüksek operasyonel verimlilik ve yüksek kaliteli atama elde etmek için araştırma analitiği çerçevesi aracılığıyla üçgen bir perspektiften sistematik olarak veriye dayalı karar modelleri uygulanmıştır. Modelde projeleri ontoloji tabanlı metin madenciliği ve kümeleme yöntemleri kullanılarak gruplandırılmıştır.

Ayrıca, bilimsel veri tabanlarından gelen büyük verileri analiz edilerek ve etkili karar vermeyi desteklemek için görselleştirilmiş karar-hazır bilgi üretilmektedir. Chandre vd. [23] ile Madhuri vd. [24] makalelerinde projelerinin gruplandırılması içi ontoloji tabanlı bir metin madenciliği yöntemi önermektedir. Literatürde son yıllara bakıldığında hakem ataması için bir akıllı karar destek yaklaşımı Liu vd. [25] tarafından önerilmekte ve bir Atama Karar Destek Sistemi (ADSS) geliştirilmesi amaçlamaktadır. Bu yaklaşımda, hakem ataması ve operasyon araştırması teknikleri hakkında sezgisel bilgi entegre edilmiştir. Yaklaşım, tekliflere atanan hakemlerin toplam uzmanlık düzeyini en üst düzeye çıkaran hakemlik görevinin en iyi çözümünü belirlemek için karar modellerini kullanmaktadır. Ayrıca tekliflerin farklı derecelerde dağılımını dengeler ve hakemler ile projeler arasındaki çıkar çatışmaları çözülmektedir. Xu vd. [26]

Tablo 2. Literatüre Genel Bakış (Literature Overview)

Sıra	Kullanılan Yöntem	Kullanım Alanı	Kullanılan Veri Kümesi
[8]	Bilgi Tabanlı Sistemler	R&D Proje Seçimi	National Natural Science Foundation of China (NSFC) Projeleri
[9]	Tam Sayılı Programlama ve Sezgisel Yöntem	R&D Proje Seçimi	40 Proje Önerisi, 80 Hakem ve 4 Ayrı Sınıf
[10]	Anahtar Kelime Çıkarımı ve Kümeleme Algoritmaları	R&D Projelerine Hakem Belirleme	U.S. National Science Foundation'de 2.004 Adet Proje Önerisi
[11]	Önceki Projelerin Benzerlerini Belirlemek İçin Tarama Algoritması ve Metin Madenciliği	R&D Projeleri İzleme	-
[12]	Bulanık Dilbilimsel Analiz	R&D Hakem Değerlendirme	National Natural Science Foundation of China (NSFC) Projeleri 4 seviye
[13]	Bilgi Tabanlı Sistemler	R&D Proje Seçimi	National Natural Science Foundation of China (NSFC) Projeleri
[14]	Bilgi Tabanlı Kurallar ve Genetik Algoritma	R&D Proje Kümeleme	National Natural Science Foundation of China (NSFC) 168 Projesi
[15]	Ontoloji Tabanlı Metot	R&D Proje Yönetimi	Innovation and Technology Commissionun İngilizce, Çince dillerinde R&D Projeleri
[16]	Bilgi Tabanlı Kurallar	R&D Proje, Hakem Atama, Gruplama, Seçimi	National Natural Science Foundation of China (NSFC)'nin 100 Hakem ve 300 Proje Önerisi
[17]	Ontoloji Tabanlı Metin Madenciliği, Kümeleme	R&D Proje Seçimi	National Natural Science Foundation of China (NSFC) Projeleri
[18]	Ontoloji Tabanlı Metin Madenciliği, Kümeleme, K Means, Karar Ağacı	R&D Proje Seçimi	500 adet R&D Proje ve Hakem
[19]	Ontoloji Tabanlı Metin Madenciliği, Kümeleme, Latent Semantic Indexing (LSI), Fuzzy C Means Clustering Algorithm, Self-organized Mapping (SOM) Algorithm	R&D Proje Seçimi	1.000 R&D Projesi
[20]	Ontoloji Tabanlı Metin Madenciliği, Kümeleme, Bilgi Tabanlı Kurallar	R&D Proje Seçimi	-
[21]	Ontoloji Tabanlı Metin Madenciliği, Kümeleme, Self-organized Mapping (SOM) Algorithm Genetik Algoritma	R&D Projelerine Hakem Belirleme	1.000 R&D Projesi
[22]	Ontoloji Tabanlı Metin Madenciliği, Kümeleme	R&D Proje Seçimi	National Natural Science Foundation of China (NSFC) 40 Proje Önerisi
[23]	Ontoloji Tabanlı Metin Madenciliği, Kümeleme, Optimizasyon, Sıralama Algoritmaları, Self-organized Mapping (SOM)	R&D Proje Seçimi	-
[24]	Ontoloji Tabanlı Metin Madenciliği, Kümeleme, Self-organized Mapping (SOM), Latent semantic indexing (LSI), Genetik Algoritma	R&D Proje Seçimi	-
[25]	Bilgi Tabanlı Kurallar	R&D Hakem Atama	National Natural Science Foundation of China (NSFC) 500 Proje
[26]	Latent Dirichlet Allocation (LDA), Metin Madenciliği	R&D Hakem Atama	-
[27]	Doğal Dil İşleme, ATM (Author Topic Model), EM (Expectation Maximization) Algoritması, Tam Sayılı Programlama	Hakem Atama	30 Proje ve 4 Hakem

çalışmasında proje ve hakem profil bilgisini otomatik olarak sınıflandırır ve bilgileri farklı türlerine göre birkaç kategoriye entegre ederek her kategorinin proje ve hakem bilgilerinin farklı boyutlarını temsil etmesi sağlanmıştır. Ayrıca her bir kategorideki bilgilerin benzerliğini hesaplamakta ve aday hakemler olarak ilk 8 uzmanı seçmek için benzerliği sıralamakta ve birkaç hakemin projeye atanmasına karar vermek için aday hakemler için değerlendirme modeli oluşturulmaktadır. Projelerden konu kelimelerin çıkarılmasında LDA ve metin madenciliği tekniklerinden faydalanılmıştır. Grup başvurularına hakem tavsiye etmek için dikkate alarak konu, yazar ve araştırma alanı gibi farklı bakış açılarını dikkate alan ve tamsayılı programlamayı formülüze eden bir çalışma Jin vd. tarafından [27] yapılmıştır. Özellikle, konuyla ilgisi ve konu otoritesi, ilgili ve akredite edilmiş aday hakemleri sunumla ilgili konularda tavsiye etmek için kullanılırken, araştırma faaliyeti adayların bir sunumu gözden geçirme istekliliğini incelemektir. Önerilen yaklaşımın etkinliğini değerlendirmek için, iki büyük bilimsel veri kümesi üzerinde karşılaştırmalı deney kategorileri gerçekleştirilmiştir. Deneysel sonuçlar, kıyaslama yaklaşımları ile kıyaslandığında, önerilen yaklaşımın farklı değerlendirme ölçütlerinde kayda değer bir kayıp olmadan hakem adayların araştırma ilgisini yakalayabildiğini göstermektedir.

Son yıllarda Ar-Ge projelerinin sınıflandırılması konusu haricinde günlük aktivitelerin sınıflandırılması [28], duygu sınıflandırma [29], doküman dili sınıflandırma [30] gibi çalışmalar da yapılmıştır. Ancak literatürde yapılan çalışmalar incelendiğinde; derin öğrenme teknikleri ile yapay sinir ağlarını kullanarak Ar-Ge projelerinin teknolojik alanlarının sınıflandırmalarını yapan ve hakem performans değerlendirmesinde kullanılabilecek verileri büyük metin verileri içerisinden çekebilen bir sistemin bulunmadığı görülmektedir. Günümüze kadar yapılan çalışmaların hakem sınıflandırma, hakem gruplandırma, metin madenciliği, karar destek sistemi, matematiksel model, tam sayılı programlama, atama ve otomatik atama konu başlıkları bazında gruplandığı görülmektedir. Bu çalışma, Ar-Ge projeleri için hakem seçimi problemlerinde çok büyük metinsel veriler içerisinden anlamlı sonuçlar çıkartabilecek, hakemlerin performanslarının ölçümünde kullanılabilecek verileri derin öğrenme tekniklerini kullanarak hızlı ve etkin bir şekilde derleyebilecek bir sistem öneren ilk çalışma olacaktır.

3. Yöntemler (Methods)

Bu bölümde teknoloji geliştirme amacıyla yapılan Ar-Ge proje başvurularının sınıflandırılması amacıyla seçilen veri kümesi;

kelimelere ayırma, küçük harfe çevirme ve noktalama işaretlerinden arındırma, durak harf ve genel ifadelerden ayırma işlemleri ile temizlenerek ön işlemeden geçirilmiş, elde edilen veri kümesi Word2Vec+CNN metodu kullanılarak sınıflandırma işlemlerindeki performans değerleri bulunmaya çalışılmış ve klasik metin madenciliği algoritmaları kullanılarak elde edilen doğruluk oranları ile karşılaştırılmaya çalışılmıştır.

3.1. Veri Ön İşleme (Data Pre-Processing)

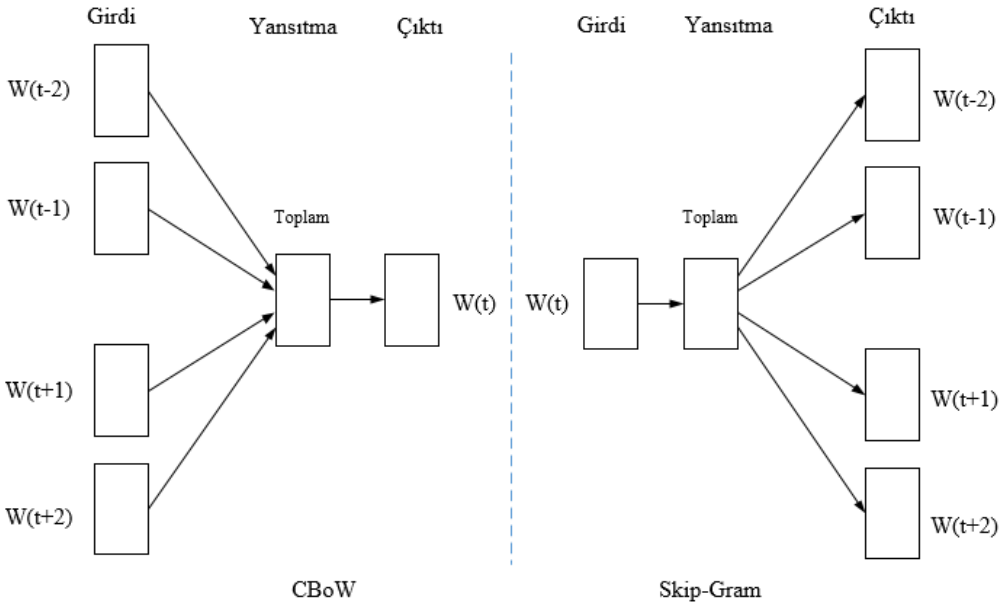
Ar-Ge projeleri başvurulan kuruluş bakımından çalışma alanı, projenin içeriği, uygulama yöntemleri vb. hususlara bağlı olarak çeşitlik gösterebilmekte ve buna bağlı olarak projelerin amaç, kapsam ve içerik gibi metinlerde farklılıklar bulunmaktadır. Metin sınıflandırma işlemlerinin yapılabilmesi ve anlamlı sonuçların üretilebilmesi için metinleri oluşturan kelimelerin ek, noktalama işareti, sembol, şekillerden arındırılması ve küçük-büyük harf farklılıklarının ortadan kaldırılması gerekmektedir. Yapılan bu sadeleştirme sayesinden metinlerin çözümlenmesi ve doğru sınıflandırma yapılabilmesi kolaylaşacak ve yüksek yüzdelerde doğru sınıflandırma yapabilme başarısı sağlanabilecektir. Bununla birlikte, metinleri oluşturan kelimelerin sayısallaştırılarak metinlerin frekans sayılarının bulunması sağlandıktan sonra söz konusu bu sayısal olarak ifade edilebilen metinlere yönelik çeşitli veri analizleri yapılabilecektir. Sınıflandırma işlemi yapılan her metin uzayda temsili bir vektör ile gösterilir ve çözümlenen metinler için kelime sözlüğünün oluşturulması gerekmektedir. Sınıflandırmalarda kullanılacak kelimelerin köklerinin bulunması zorlu bir süreçtir. Bu süreçte kelimeler üzerinden işlemler yapılması gerekmektedir. Türkçe metin gövdelerinin bulunması amacıyla Zemberek gibi yazılım kütüphaneleri bulunmaktadır. Kelime gövdelerinin morfolojik açıdan incelenerek analizinin yapılması gerekmektedir. Çalışmada zemberek yazılımı kullanılarak kelimeler köklerine ayrılmıştır. Eğitim verileri tamamı küçük harfe dönüştürülmüş ve noktalama işaretleri ve sembollerden arındırılmıştır. Elde edilen Türkçe metinler için ön işlem yapılması gerekmekte olup metinlerdeki kelimelerin gövdeleri bulunması ve herhangi bir kategorik anlamı olmayan gereksiz sözcükler atılması gerekmektedir. Kategorik anlamı ifade etmeyen sözcükler atılmış olduğundan üzerinde tekrar bir ön isleme yapılmamıştır. Oluşturulmak istenen vektör boyutuna göre özellik secimi yapılarak eğitim ve sınama metinlerine ait vektör kümesi oluşturulmaktadır. Öznitelikler vektörleri oluşturulurken, metinde çok sayıda geçmesine rağmen bir anlam ya da belirleyici olmayan "acaba, ama, belki, bile, birkaç, biz, bu, çok, defa, dolayı, eğer, gibi, henüz, hep, her, hiç, için, ile, olan, oysa, rağmen, sanki, üzere, yani, yine, yoksa, zaten" kelimeleri gibi edat, sayısal, zamir ve belirteçlerden oluşan durak kelimeler ve genel ifadelerden ayıklanmıştır.

3.2. Word2Vec Modeli (Word2Vec Model)

Derin öğrenme algoritmaları çok sayıda metin verisi içeren sınıflandırma çalışmalarında kullanılmaktadır. Daha önce yapılan çalışmalara bakıldığında kelime torbası ve TF-DF yöntemleri yaygın olarak kullanılmış ancak verinin büyük boyutlara ulaşmasından dolayı yeni çalışmalarda Word2Vec ve GloVe modeli vektör temsili için kullanılmaya başlandığı görülmektedir. Bu çalışmada sözcüklerin vektörlerini oluşturmak için kelime temsili olarak adlandırılan Word2Vec modeli uygulanmıştır. Son yıllarda yapay sinir ağları ile kapsamlı bir eğitim kümesi olmadan, kelimelerin denetimsiz vektör temsilini öğretimini mümkün olduğu anlaşılmıştır. Burada kullanılan kelime temsili yöntemi, kelimelerin söz dizimsel ve anlamsal yönlerini yakalamada önemli bir faktördür. Word2Vec, bir belgedeki kelimelerin anlamını, belirli bir bağlamda benzer anlamlara sahip kelimelerin yakın mesafeler sergilediği hipotezine dayanarak tespit eder ve vektörleştirir [6]. Böylelikle vektörler arasındaki mesafeler hesaplanarak aralarında benzerlikler bulunabilmektedir. Word2Vec model mimarisi, girdi olarak metin gövdesi ve çıktı olarak vektör alanı olan bir Sinir Ağı kullanılarak hesaplanır. Word2Vec modellerinden olan CBOW'un (continuous bag-of-word) temel prensibi, komşu kelimeleri analiz ederek belirli bir kelimenin ne zaman ortaya çıkacağını tahmin etmeyi sağlarken, diğer bir model olan Skip-gram'ın temel prensibi belirli bir kelimenin etrafında görünen diğer kelimeleri tahmin etmeyi çalışmaktır (Şekil 1) [31].

*3.3. Evrişimsel Sinir Ağları (CNN-*Convolutional Neural Networks*)*

Başlangıçta görüntü sınıflandırma ve bilgisayar görme sorunları için geliştirilmiş olan CNN, bilgileri kaybetmeden değerleri bir sonraki katmana aktarabilen, özellik çıkarım ve sınıflandırmak için kelimeler arasındaki anlamsal benzerlik gibi bilgileri kullanan bir sinir ağı



Şekil 1. CBoW ve Skip-gram Model Mimarisi (CBoW and Skip-gram Model Architecture)

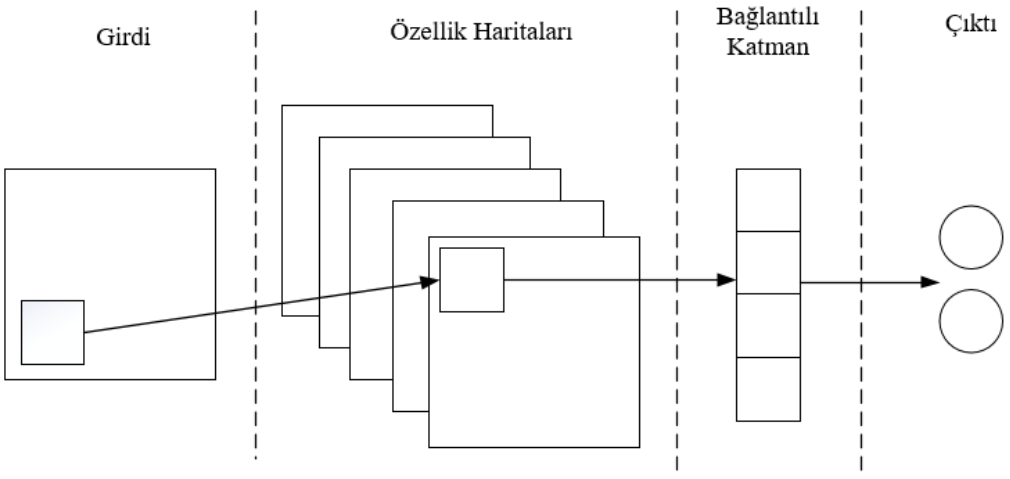
yöntemidir [31]. Anlamsal bilgileri kullanmak için uygun olan CNN girdi, çoklu gizli ve çıktı katmanlardan, katmanlarda özellik haritalar ile evrişimsel katmanlar ve havuzlama katmanlarıyla tam bağlantılı bir katmandan oluşur. Evrişimsel katman ve havuzlama katmanı, girdi değerlerinin özelliklerini çıkarır ve çıkarılan değerleri özellik haritasına eşler. Alt seviye detaylardan başlayarak üst seviye detaylara kadar etkili bir öğrenme gerçekleştirmek amacıyla bilişsel katmanda girdi filtrelenir ve öznitelik haritaları elde edilip havuzlama katmanında öznitelik haritaları örneklenerek ve ağın daha genel ve hızlı öğrenmesi sağlanır. Tamamen bağlantılı katmandaki her bir sinir bir önceki katmandan gelen tüm girişlere bağlı olarak çıktı üreterek her katman bir önceki katmanın sonucuna göre öznitelik çıkarmakta ve tüm katmanları birleştirip eğiterek öznitelik hiyerarşisini öğrenebilmektedir (Şekil 2) [32]. Sonuç olarak, evrişimsel katman ile filtreleme sonucunda giriş verisinden daha küçük bir matris elde edilmesi [33], havuzlama katmanı ile ağdaki veri miktarı düşürülerek ağdaki hesaplama miktarlarının ve kullanılacak bellek miktarının da azalması [34] ve tam bağlı katman ile tüm nöronlar bir dizi şeklinde görünerek tam bağlı katmanın özelliği önceki katmana bağlı olarak ortaya çıkartarak nesneyi belirleyecek olan özelliklerin hangi sınıfla ilişkili olduğu [3, 35] belirlenmiş olur. Başlangıçta CNN yöntemi cümleleri vektörlerle eşleştirmede kullanılırken akabinde cümleleri, belgeleri ve kitapları vektörlerle eşleştirmeye çalışılmıştır [36, 37, 38]. Word2Vec tarafından önceden eğitilmiş kelime vektörleri ile CNN kullanma yaklaşımının, otomatik özellik çıkarma, daha yüksek öğrenme verimliliği ve destek vektör makinesi ve gizli anlamsal analiz dahil geleneksel yöntemlere göre daha iyi sınıflandırma performansı gibi avantajları vardır [39].

3.4. Sınıflandırıcı Tasarımı (Classifier Design)

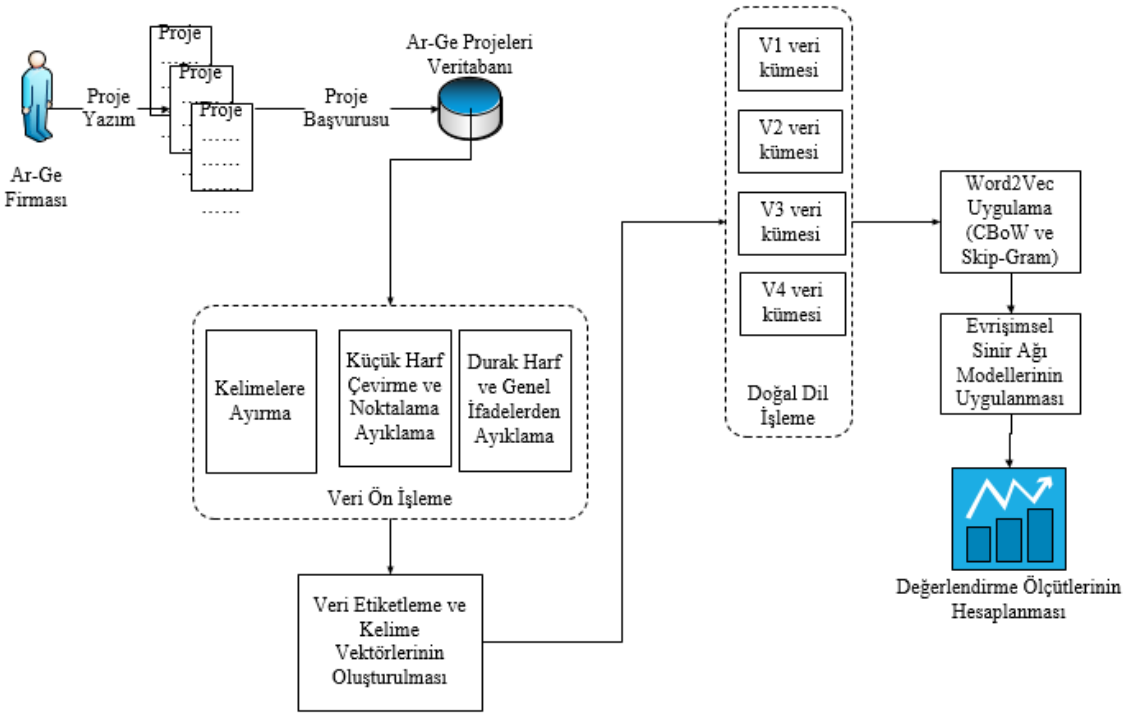
Bu çalışmada yaklaşık 500 adet Ar-Ge içerikli makale kullanılarak python programlama diline ait "tensorflow, numpy, os, csv" kütüphaneleri kullanılmıştır. Tablo 3'de kullanılan derin öğrenme bilgisayarının donanım özellikleri gösterilmiştir. Veri kümesi öncelikli olarak ön işlemden geçirilerek temizlenmiştir. Temizlenen veri kümesinde 4 farklı alt veri kümesi elde edilmiştir (Şekil 3). Kelime vektörleri elde edilme işlemleri ve 4 ayrı veri kümesinin oluşturulması işlemleri bütünleşik olarak yapılmıştır. Ham veri kümesi için veri ön işleme işlemleri yapılmazken, diğer 3 veri kümesinin oluşturulmasında veri ön işleme yapılmıştır. CNN yöntemi farklı operasyonlarla özellikleri yakalayan, onları sınıflandıran ve yüksek doğrulukla sınıflandırma yapabilen bir yöntemdir. Bir sinir ağındaki parametre sayısı, katman sayısındaki artışla birlikte hızla büyümektedir. Bu durum modelin eğitimin hesaplamasını zorlaştırmaktadır. Ar-Ge projelerinin oluşturduğu metinler büyük verilerdir ve CNN yöntemi ile modelin parametreleri ayarlamak için geçen süre performası çok olumlu etkileyecek şekilde düşürebilmektedir. Boyutsallık önemli bir konu olup boyut azaltmayı sağlayabildiği için klasik sınıflandırma algoritmaları yerine bu modelde CNN yöntemi kullanılmıştır. Elde edilen veriler CNN tarafından ihtiyaç duyulan Word2Vec'in CBoW ve Skip-Gram eğitim veri setlerinde kullanılmıştır. Veri kümesinin eğitminde oluşturulan modelde N boyutlu özellik haritalı evrişimsel sinir ağı katmanı, kayıp fonksiyonu olarak çapraz entrop kayıp fonksiyonu (croos-entropy loss), aktivasyon katmanı olarak SoftMax, öğrenme algoritması olarak denetimli ögrenme (supervised Learning) ve havuzlama işlemi olarak Max-Pooling kullanılmıştır.

Şekil 4'de, çalışmada kullanılan CNN mimarisinde CBOW ve Skip-Gram öğrenme algoritmaları ile gösterilmiştir. Ar-Ge proje metin kelimeleri önceden eğitilmiş CBOW ve Skip-Gram algoritması ile bir vektör değeri verilmiştir. CNN her bir sözcüğün vektör değerini giriş katmanına geçirdi ve giriş katmanı, belirli bir pencere boyutu içindeki komşu kelimelerdeki özellikleri ayıklanmıştır. Evrişimsel işlemle çıkarılan özellikler bir özellik haritası oluşturmak için kullanılmıştır. Akabinde CNN, zamanla çıkarılan özelliklerden en büyük değeri çıkaran maksimum bir havuzlama işlemi gerçekleştirmiştir. CNN tarafından çeşitli özelliklerin ayıklanması için pencere boyutu değiştirilerek tekrarlanmıştır. Elde edilen değerlere bağlı olarak, çıkış katmanı giriş metnini sınıfının bulunması sağlanmıştır. Yapılan çalışmada sınıfları belirlemek amacıyla şu tanımlamalar kullanılmıştır. Bunlar: "sequence_length" - Cümlelerimizin uzunluğu, "num_classes" - Çıktı katmanındaki sınıfların sayısı, kelime_boyutu - Kelime dağarcığının boyutu, "embedding_size" - düğümlern boyutluluğu. Katmanın mimarisi (Şekil 5) şu şekilde oluşturulmuştur.

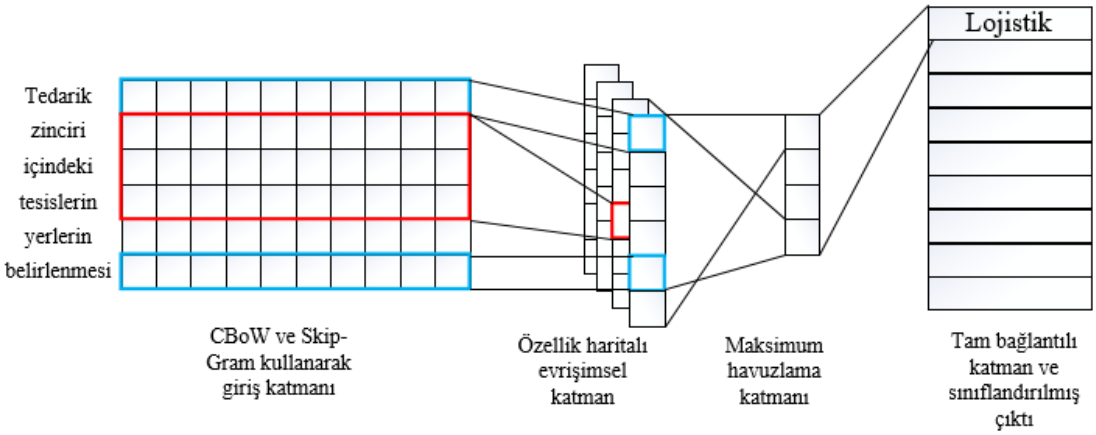
- Embedding Layer: Tanımlanılan ilk katman olup kelime indekslerini düşük boyutlu vektör temsillerine eşleyen gömme katmanıdır. Temelde verilerden öğrenilen bir arama tablosudur.
- Convolution and Max-Pooling Layers: Evrişimli katmanları ve ardından maksimumum havuzlama katmanları oluşturulmaktadır. Burada farklı boyutlarda filtreler kullanılmıştır. Sonuçların tek bir büyük özellik vektöründe birleştirilmektedir.
- Dropout Layer: Bırakma katmanı evrişimli sinir ağlarının düzenlemek için kullanılan yöntemdir. Nöronlarının bir kısmını stokastik olarak devre dışı bırakılır ve bu nöronların birlikte adapte olmasını engeller, böylece bireysel olarak yararlı özellikleri öğrenilmeye zorlanır.



Şekil 2. CNN-Evrişimsel Sinir Ağları Katmanları (*CNN*-Convolutional Neural Networks Layers)



Şekil 3. Önerilen Sisteme Genel Bakış (Proposed System Overview)



Şekil 4. CBoW ve Skip-Gram Algoritmalı CNN Yapısı (CBoW and CNN Structure with Skip-Gram Algorithm)

- Scores and Predictions: Maksimum havuzlamadan (bırakma uygulanmış) özellik vektörü kullanılarak bir matris çarpımı yapılır ve en yüksek puana sahip sınıf seçilerek tahminler oluşturulmaktadır.
- Loss and Accuracy: Puanlar kullanılarak kayıp fonksiyonlar tanımlanabilmektedir. Burada kayıp sinir ağının yaptığı hatanın bir ölçüdür ve amaç bu hatanın en aza indirilmedir.

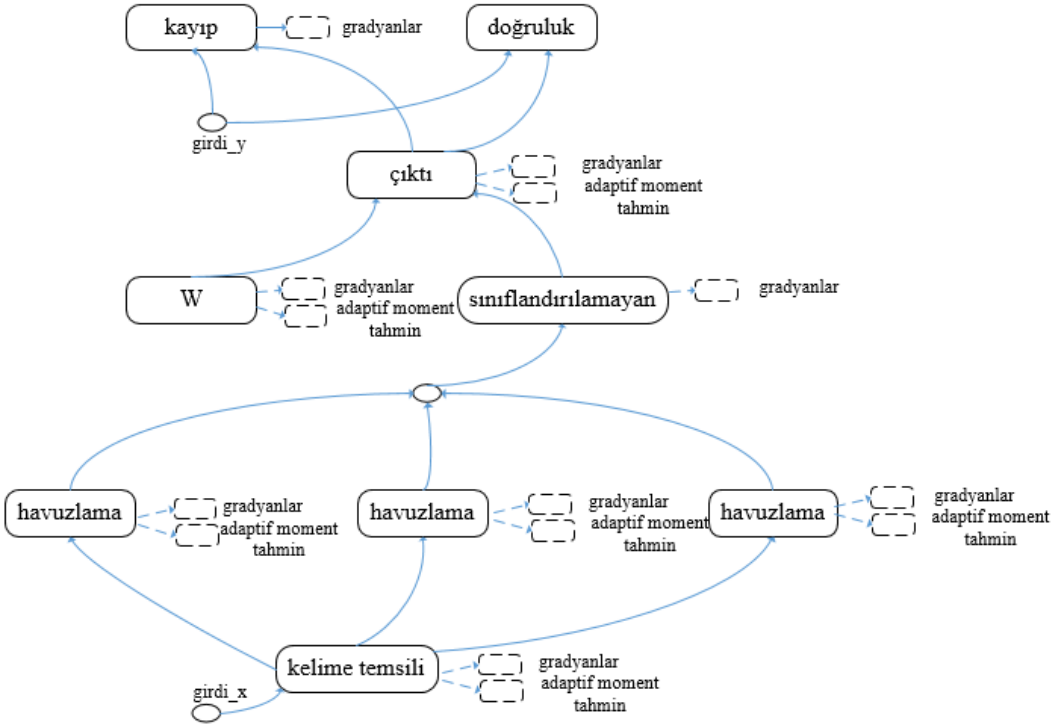
4. Deneysel Çalışmalar Ve Değerlendirmeler
(Experimental Studies And Evaluations)

4.1. Veri Kümesi (Dataset)

Bu çalışmada, TÜBİTAK Dergipark sitesinden seçilen Ar-Ge proje içeriğine sahip makalelerin konu ve kapsamlarında bulunan metinler derlenerek veri kümeleri elde edilmiştir. Elde edilen veri kümeleri içerisinde endüstri mühendisliği bilim dalı içerisinden "Stokastik", "Optimizasyon", "Simülasyon", "Üretim Planlaması", "Ergonomi", "Ekonomi", "Kalite", "Karar Destek", "Lojistik" teknoloji alanlarında yaklaşık 500 makale ve toplam 9.589 farklı öznitelik bulunmaktadır. Performansını daha iyi ölçebilmek amacıyla 4 farklı veri kümesi 100, 200, 300, 400 ve 500 adetlik veri kümelerine ayrılmıştır. Akabinde Türkçe metinler kullanılmadan önce ön islemeden geçirilerek kullanıma hazır hale getirilmiştir. Veri etiketleme işlemi için Word2Vec'in CBoW ve Skip-Gram yöntemleri kullanılmış, karşılaştırma amacıyla kullanılan geleneksel sınıflandırma metotlarında da hazır paket program olan Weka yazılımı kullanılmıştır.

Tablo 3. Veri Kümesi Özelikleri (Dataset Properties)

Sınıf Sayısı	9
Proje Sayısı	500
Öznitelik Sayısı	9.589
Eğitim Kümesi	%10
Test Kümesi	100/200/300/400/500



Şekil 5. Sınıflandırıcının Tasarımı (The Design of The Classifier)

4.2. Deneysel Kurulum (Experimental Setup)

Deneysel çalışma bölümünde yapılan çalışmalarda kullanılan donanımın özellikleri Tablo 4'de verilmiştir.

Tablo 4: Donanım Özellikleri (Hardware Spesification)

OS	Windows 10 (64 bit işletim sistemi)
CPU	İntel Core İ7-8750H@2.20 Ghz
RAM	16GB ddr4 Dual Channel 1330 Mhz
Framework	Microsoft
GPU	MVIDIA GeForce GTX 1060 6GB

4.3. Sonuçlar (Results)

CNN mimarisinin network/model yapısına ilişkin 4 ayrı veri seti içerisinde elde edilen en iyi train-validation loss grafiği aşağıda verilmiştir. Modelin doğru eğitilip eğitilmediği (overfit/underfit) göstermek amacıyla farklı epoch düzeylerinde denenmiştir. 1.400, 1.500, 2.000 adet olmak üzere farklı epochtan sonra train setlerinin 1.0 yakınsaması nedeniyle 1.400 epochtan sonra eğitim setin kesilmiştir. Böylelikle overfitting durumundan dolayı bu noktada test ve train set validationlarına bakılarak karar verilmiştir. Skip-Gram Algoritmalı CNN Yöntemi ile yapılan sınıflandırmalarda %90'nın üzerinde bir başarı sağlandığı görülmektedir.

Yapılan çalışmada sadece Ar-Ge proje metinleri ile oluşturulmuş ham veri kümesi kullanılmamış olup, ön işleme yapılarak oluşturulmuş 3 ayrı veri kümesi daha kullanılmıştır. Zemberek gibi Türkçe doğal dil işleme kütüphaneleri ya da farklı veri temizleme işlemleri yapılaması durumunda sınıflandırma modeline olumlu/olumsuz etkilerinin olup olmayacağı görülmeye çalışılmıştır. Oluşturulan Word2Vec + CNN modelinin Ar-Ge projeleri sınıflandırma işlemlerindeki performans değerleri kullanılan her 4 veri kümesi ile de olumsuz bir sonuç üretmediği tespit edilmiştir.

4.4. Değerlendirme Ölçütleri (Evaluation Criteria)

Sınıflandırma başarıları hesaplanırken doğru sınıflanan örneklerin yüzdesini veren birçok yöntem kullanılmakta olup karmaşıklık matrisinden (confusion matrix) de hesaplanabilmektedir. Karmaşıklık matrisi sınıflara ait verilerin, sınıflandırıcı tarafından hangi sınıflarda bulunduğunu gösteren bir matristir. Doğru sınıflandırılmış pozitif örnek (True Positive, TP), doğru sınıflandırılmış negatif örnek (True Negative, TN), yanlış sınıflandırılmış pozitif örnek (False Positive, FP) ve yanlış sınıflandırılmış negatif örnek (False Negative, FN) değerlerini vermektedir. Sınıflandırma performansını ölçmede yaygın olarak Eş. 1, Eş. 2, Eş. 3 ve Eş 4.'de gösterilen Doğruluk (Accuracy-ACC), Kesinlik (Precision), Anma (Recall) ve F Skoru kullanılmaktadır [32]. Bu çalışmada değerlendirmenin performans ölçütlerini göstermek için Doğruluk (Accuracy-ACC), Kesinlik (Precision), Anma (Recall) ve F Skoru değerleri hesaplanarak bir karşılaştırma yapılmıştır.

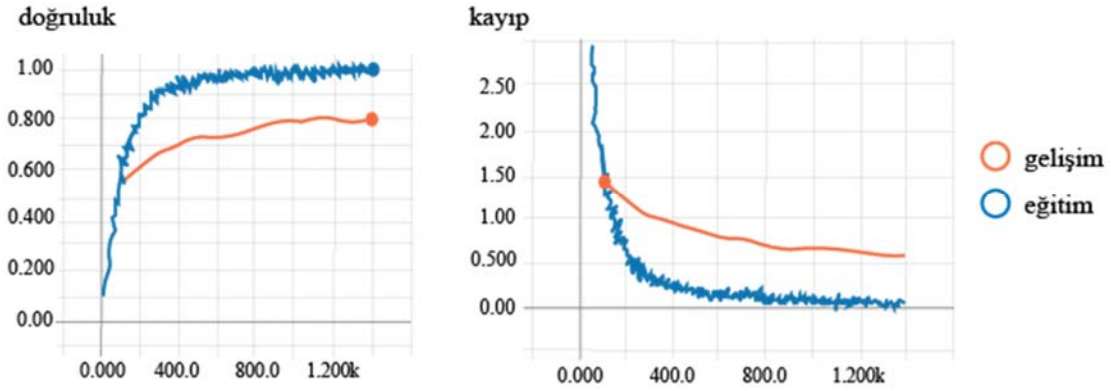
$$Doğruluk\ (Accuracy) = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Kesinlik\ (Recall) = \frac{TP}{TP+FN} \quad (2)$$

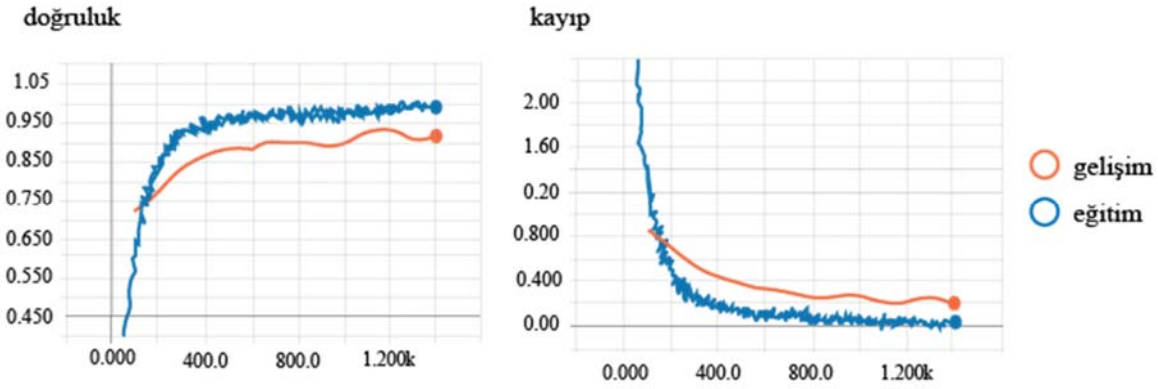
$$Anma\ (Precision) = \frac{TP}{TP+FP} \quad (3)$$

$$FSkoru = \frac{2*Kesinlik*Anma}{Kesinlik+Anma} \quad (4)$$

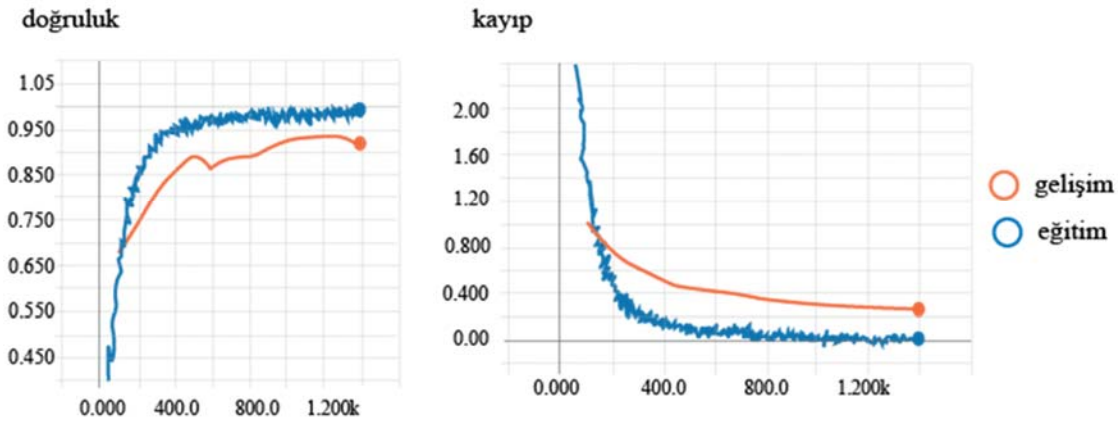
Çalışmanın bu bölümünde, önerilen CNN tabanlı derin mimarinin başarımını sınamak amacıyla değerlendirmeler yapılmıştır. Önerilen yöntemin etkinliğini gösterebilmek amacıyla, literatürdeki doğal dil



Şekil 6. Veri Seti 1'e Ait Doğruluk ve Kayıp Fonksiyonu (Accuracy and Loss Function of Data Set 1)



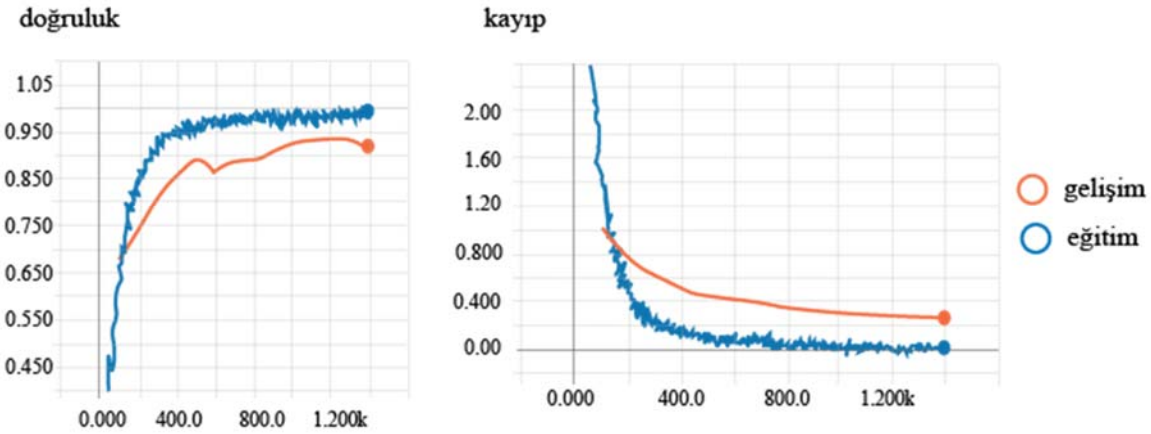
Şekil 7. Veri Seti 2'e Ait Doğruluk ve Kayıp Fonksiyonu (Accuracy and Loss Function of Data Set 2)



Şekil 8. Veri Seti 3'e Ait Doğruluk ve Kayıp Fonksiyonu (Accuracy and Loss Function of Data Set 3)

işleme ve metin madenciliği algoritmaları ile karşılaştırmalar sunulmuştur. Yapılan çalışma sonucunda Tablo 5'de bu sınıflandırmalar geleneksel sınıflandırma algoritmaları ile elde edilen sonuçlar ile karşılaştırılmış ve Word2Vec + CNN metodunun Ar-Ge projeleri sınıflandırma işlemlerindeki performans değerleri gösterilmiştir. Yapılan çalışmada farklı sayılarda örnek Ar-Ge proje metinleri ile oluşturulmuş veri kümeleri ile karşılaştırılmalar yapılmıştır. Bu veri kümeleri; herhangi bir işlem görmemiş ham veri kümesinden (V1), zemberek programı kullanılarak temizlenmiş veri kümesinden (V2), zemberek kullanılmış ve elle işlenerek temizlenmiş veri (V3) kümesi ile son olarak sadece elle işlenerek temizlenmiş veri kümesinden (V4) oluşmaktadır.

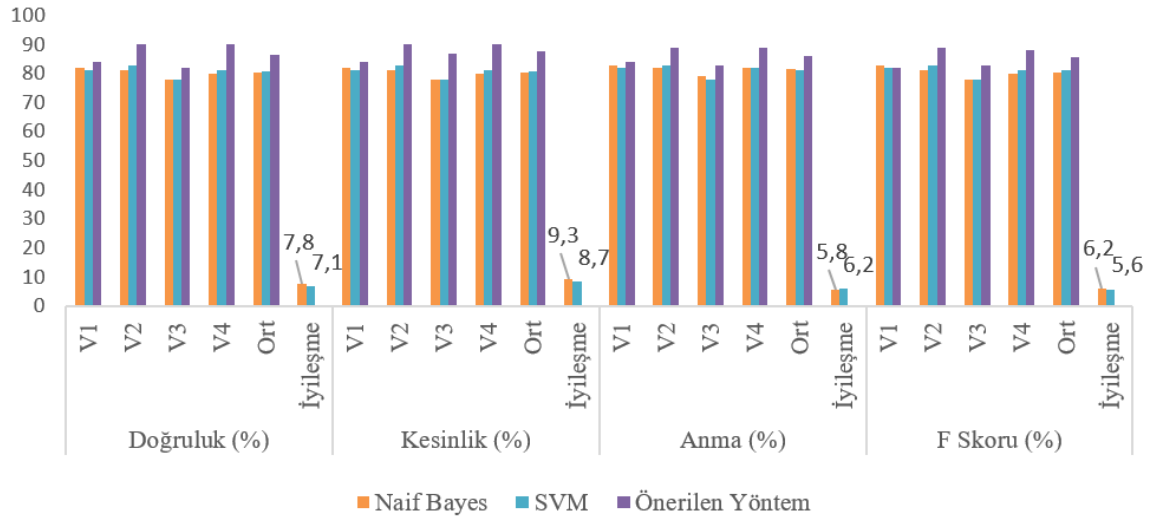
Çizelgelerdeki veriler ışığında klasik sınıflandırma algoritmaları kullanılarak Naive Bayes ve SMO ile %80'in üzerinde doğruluk oranıyla sınıflandırma yapabildiği görülebilmektedir. Ayrıca sınıflar içerisinden Ekonomi, Ergonomi ve Kalite sınıflarında daha yüksek doğruluk oranına ulaşılırken Lojistik, Stokastik ve Üretim Planlama



Şekil 9. Veri Seti 4'e Ait Doğruluk ve Kayıp Fonksiyonu (Accuracy and Loss Function of Data Set 4)

Tablo 5. Klasik Sınıflandırma Algoritmaları ile Yapılan Sınıflandırma Modellerinin Performans Göstergeleri
(Performance Metrics of Classification Models with Classical Classification Algorithms)

Model	Doğruluk (%)						Kesinlik (%)						Anma (%)						F Skoru (%)					
	V_1	V_2	V_3	V_4	Ort	İyileşme	V_1	V_2	V_3	V_4	Ort	İyileşme	V_1	V_2	V_3	V_4	Ort	İyileşme	V_1	V_2	V_3	V_4	Ort	İyileşme
Naif Bayes	82	81	78	80	80	*7,8*	82	81	78	80	80	*9,3*	83	82	79	82	82	*5,8*	83	81	78	80	81	*6,2*
SVM	81	83	78	81	81	*7,1*	81	83	78	81	81	*8,7*	82	83	78	82	81	*6,2*	82	83	78	81	81	*5,6*
Önerilen Yöntem	84	90	82	90	87		84	90	87	90	88		84	89	83	89	86		82	89	83	88	86	



Şekil 10. Klasik Sınıflandırma Algoritmaları İle Yapılan Sınıflandırma Modellerinin Performans Göstergeleri
(Performance Metrics of Classification Models with Classical Classification Algorithms)

sınıflarında daha düşük doğruluk oranlarına ulaşıldığı görülmektedir. Skip-Gram Algoritmalı CNN Yöntemi ile yapılan sınıflandırmalarda %90'ın üzerinde bir başarı sağlandığı görülmektedir. Şekil 10.'da ve Tablo 6.'da yer alan sonuçlar incelendiğinde, örnek Ar-Ge proje metinleri ile oluşturulmuş herhangi bir işlem görmemiş ham veri kümesinden (V1), diğer veri kümelerine göre daha az başarılı sınıflandırmalar yapıldığı görülmektedir. Bu açıdan bakıldığında Türkçe metinlerde veri ön işleme yapılarak elden edilen veriler yüksek başarım oranlarına sahiptir. Özellikle zemberek programı kullanılarak temizlenmiş veri kümeleri (V2) ve (V3), bu yüksek doğruluk başarım oranında daha fazla etkiye sahip olduğu görülecektir. Veri kümesinin eleman sayısı arttığında Türkçe dil yapısının özelliklerinde kaynaklı sorunları ortadan kaldırılması gerekmektedir.

Tablo 6. Word2Vec + CNN Yöntemi İle Yapılan Sınıflandırma Modellerinin Performans Göstergeleri
(Performance Metrics of Classification Models Made by Word2Vec + CNN Method)

Epoch	Veri Seti 1 İçin Doğruluk	Veri Seti 2 İçin Doğruluk	Veri Seti 3 İçin Doğruluk	Veri Seti 4 İçin Doğruluk
100	0,55	0,71	0,67	0,63
200	0,65	0,82	0,78	0,76
300	0,73	0,90	0,90	0,73
400	0,73	0,90	0,90	0,76
500	0,78	0,92	0,94	0,88
600	0,73	0,88	0,84	0,78
700	0,76	0,92	0,92	0,82
800	0,82	0,90	0,90	0,82
900	0,82	0,88	0,94	0,84
1000	0,80	0,90	0,94	0,88
1100	0,82	0,98	0,94	0,86
1200	0,82	0,94	0,94	0,90
1300	0,80	0,86	0,94	0,88
1400	0,82	0,94	0,90	0,92

5. Sonuçlar (Conclusions)

Bu çalışma ile araştırma ve teknoloji geliştirme amacıyla yapılan Ar-Ge başvurularının sınıflandırılması çalışmalarında kullanılmak üzere; seçilen uygun bir veri kümesi üzerinde kelimelere ayrıma, küçük harfe çevirme ve noktalama işaretlerinden arındırma, durak harf ve genel ifadelerden ayırma işlemleri yapılmış, elde edilen metinlerde bulunan kelimeler etiketleme ve vektörlerin oluşturulması işlemlerinden sonra veriler çeşitli doğal dil işleme yöntemleri ile işlenmiştir. Bunun yanı sıra, yüksek doğruluk oranlarını sağlayabilecek word2vec+CNN temelli bir sınıflandırıcı tasarlanmaya çalışılmıştır. Yapılan çalışmalarda kullanılan veri kümeleri ve yöntemleri karşılaştırılarak sonuçlar analiz edilmeye çalışılmış olup, yüksek başarım oranları elde edilmiştir. Sonrasında Ar-Ge projelerinin başlangıcından sonuçlanıncaya kadar geçen süreçte sınıflandırma, hakem değerlendirme ve hakem atama işlemlerini gerçekleştirmede kullanılabilecek bir sistem ortaya konmuştur. Bu çalışmada, çok büyük Ar-Ge proje içeriklerinden anlamlı sonuçlar çıkartabilen, hakemlerin performanslarını hiyerarşik bir yapıda kademe kademe değerlendirerek, hızlı ve etkin bir şekilde ölçebilecek tarzda verimli işleyen bir sürece dönüştürülmesine katkıda bulunabilen bir sistem ortaya konmuştur.

Yapılan çalışma geliştirilmesi planlanan performans odaklı Ar-Ge projelerine hakem seçim karar destek sisteminin giriş modülünün temelini oluşturan Ar-Ge projelerinin sınıflandırılması sağlanmıştır. Temelde amaç hızlı ve etkin bir bileşen oluşturulması ve performans odaklı diğer modüllere kolaylıkla entegre edilebilmesinin sağlanmasıdır. Bunun yanı sıra klasik sınıflandırma algoritmalarının öznitelik seçim algoritmaları, yetersiz öznitelik gösterimi ve kategoriler arasındaki ilişkinin kolayca göz ardı edilmesi gibi eksikliklerin ortadan kaldırılması hedeflenmiştir. Ar-Ge projelerinin metinlerinden oluşan veri kümesi yaşayan ver çok büyük boyutlu bir veri kümesi olduğu düşünüldüğünde klasik sınıflandırma metotlarının uzun vadede yetersiz kalacağı kaçınılmazdır.

Ar-Ge projelerini sınıflandırılmasına yönelik yapılan bu çalışmada elde edilen veriler çok ölçütlü karar verme yöntemleriyle kurulabilecek ve Ar-Ge projelerini değerlendirecek hakemlerin seçilmesine yönelik olarak kullanılabilecek yeni modellerle bütünleştirilebilir. Bu amaçla bu makalede ortaya konan sistem ile bütünleştirilerek geliştirilebilecek olan yeni modellerle, Ar-Ge projelerinin başarılı şekilde değerlendirilebilmesi için önceden belirlenmiş teknoloji sınıfları içerisinde bulunan hakemlerin performanslarının değerlendirilmesi ve en uygun hakemlerin ilgili projelere atanabilmeleri de sağlanabilecektir.

Word2Vec yöntemi ile metinlerin sınıflandırılması çalışmalarıyla birlikte son dönemde diğer bazı yeni birçok başarılı bağlamsal dil modelleri de kullanılmaya başlanmıştır. Bunun yanı sıra Google'un geliştirdiği BERT (Bidirectional Encode Representations from Transformers) modeliyle, belirli makine öğrenimi görevlerinde ince ayar yapmada kullanılabilen bir dil temsilini öğrenmek için birçok etiketlenmemiş metin verisi üzerinde ön eğitiminde çift yönlü bir dönüştürücü kullanılmıştır [40]. BERT gibi önceden eğitilmiş dil modelleri doğal dil işleme çalışmalarında büyük başarı elde etmesiyle birlikte daha farklı versiyonlar yeni çalışmaları beraberinde getirmiştir. ALBERT (a lite BERT) modeli ile daha hafif model eğitimi ve daha az özellik çıkarımı süresi için azaltılmış parametreler içeren bir sürüm sağlanmıştır. RoBERTa (Robustly optimized BERT) metodu ile sağlam bir şekilde optimize edilmiş BERT yaklaşımını daha fazla veri ve hesaplama gücü ile BERT'nin yeniden eğitimi sağlanmıştır. DistilBERT (distilled version of BERT) ile BERT'nin damıtılmış bir versiyonunun öğrenimi ve yüksek oranda performansı korurken parametre sayısının daha azı kullanılmıştır [41]. Türkçe metinler için de dönüştürücü tabanlı modellerde HuggingFace'in Transformers kütüphanesi kullanılarak başarılı sınıflandırmalar yapılan çalışmalar bulunmaktadır [42-44].

Gelecekte yapılabilecek çalışmalar için, Ar-Ge Projelerinin alt faaliyet alanlarına göre sınıflandırılmasına yönelik olarak, BERT ve Glove gibi dil modellerinin kodlayıcı-kod çözücü derin mimarilerinin kullanılması ve probleme uygulanması önerilebilir. Literatürde BERT gibi önceden eğitilmiş dil modelleri doğal dil işleme çalışmalarında büyük başarı elde etmesiyle birlikte BERT yönteminin daha farklı versiyonları yeni çalışmaları beraberinde getirmiştir. Gelecekte Ar-Ge projelerine hakem atama sürecinde BERT, ALBERT, RoBERTa, DistilBERT, ELECTRA, ERNIE, GPT gibi modellerin uygulanması çalışması yapılabilir.

Kaynaklar (References)

1. TÜBİTAK. Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK) Faaliyet Raporları (2019). https://www.tubitak.gov.tr/sites/default/files/18842/tubitak_2019_yili_faaliyet_raporu.pdf. Yayın tarihi Şubat 28, 2020. Erişim tarihi Kasım 09, 2022.
2. Kowsari K., Meimandi K.J., Heidarysafa M., Mendu S., Barnes L.E., Brown D.E., Text classification algorithms: a survey, Information, 10 (4), 2019.
3. LeCun Y., Bengio Y., Hinton G., Deep learning, Nature, 521, 436-444, 2015.
4. Szczepaniak P.S., Tomczyk A., Pryczek M., Supervised web document classification using discrete transforms, Active Hypercontours and Expert Knowledge, Lecture notes in Computer Science, 305-323, 1982.

5. Sharma A.K., Chaurasia S., Srivastava D.K., Sentimental short sentences classification by using CNN deep learning model with fine tuned word2vec, Procedia Computer Science, 167, 1139-1147, 2020.

6. Mikolov T., Chen K., Corrado G., Dean J., Efficient estimation of word representations in vector space, ArXiv Prepr ArXiv:1301.3781, 2013.

7. Kim Y., Convolutional neural networks for sentence classification, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1746-1751, Doha-Qatar, 2014.

8. Tian Q., Ma J., Liu O., A hybrid knowledge and model system for R&D project selection, Expert Systems with Applications, 23 (3), 265-271, 2002.

9. Cook W.D, Golany B., Kress M., Penn M., Optimal allocation of proposals to reviewers to facilitate effective ranking, Management Science, 51 (4), 655-661, 2005.

10. Hettich S., Pazzani M.J., Mining for proposal reviewers: lessons learned at the national science foundation, Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia-USA, 862-871, 2006.

11. Choi C., Park Y., R&D proposal screening system based on text-mining approach, International Journal of Technology Intelligence and Planning, 2, 61-72, 2006.

12. Sun Y.H., Ma J., Fan Z.P., Wang J., A group decision support approach to evaluate experts for R&D project selection, IEEE Transactions on Engineering Management, 55 (1), 158-170, 2008.

13. Sun Y.H., Ma J., Fan Z.P., Wang J., A hybrid knowledge and model approach for reviewer assignment, Expert Systems with Applications, 34, 817-824, 2008.

14. Fan Z.P., Chen Y., Ma J., Zhu Y., Decision support for proposal grouping: a hybrid approach using knowledge rule and genetic algorithm, Expert Systems with Applications, 36, 1004-1013, 2009.

15. Liu O., Ma J., A multilingual ontology framework for R&D project management systems, Expert Systems with Applications, 37 (6), 4626-4631, 2010.

16. Xu Y., Ma J., Sun Y., Hao G., Xu W., Zhao D., A decision support approach for assigning reviewers to proposals, Expert Systems with Applications, 37, 6948-6956, 2010.

17. Ma J., Xu W., Sun Y.H., Turban E., Wang S., Liu O., An ontology-based text-mining method to cluster proposals for research project selection, IEEE Transactions on Systems Man and Cybernetics - Part A Systems and Humans, 42 (3), 784-790, 2012.

18. Kaur P., Sapra R., Ontology based classıfıcatıon and clusterıng of research proposals and external research reviewers, International Journal of Computers & Technology, 5 (1), 2277-3061, 2013.

19. Preethi T., Lakshmi R., An implementation of clustering project proposals on ontology based text mining approach, Information Communication and Embedded Systems (ICICES), Chennai-India, 547-550, 2013.

20. Arunachalam N., Sathya E., Begum S.H., Makeswari M.U., An ontology based text mining framework for R&D project selection, International Journal of Computer Science & Information Technology, 5 (1), 161-170, 2013.

21. Gunjal S.N., Dange B.J., Brahamane A.V., A novel ontology based R&D project proposal classification using text mining approach, International Journal of Computer Applications, 108 (4), 23-28, 2014.

22. Silva T., Jian M., Chen Y., Process analytics approach for R&D project selection, ACM Transactions on Management Information Systems (TMIS), 5 (4), 21, 2014.

23. Chandre P., Vishe B., Vishe H., Lengule P., Shah A., Ontology in text mining to cluster research project proposals, International Journal of Emerging Technology and Advanced Engineering, 4 (4), 511-514, 2014.

24. Madhuri T., Chaitali N., Swapnali G., Seema M., Kadu N.B., Project paper selectıon using ontology based text-mining, Global Journal of Advanced Research, 2 (3), 595-602, 2015.

25. Liu O., Wang J., Ma J., Sun Y., An intelligent decision support approach for reviewer assignment in R&D project selection, Computers in Industry, 76, 1-10, 2016.

26. Xu Y.H., Zuo X.L., A LDA model based text-mining method to recommend reviewer for proposal of research project selection, 2016 13th International Conference on Service Systems and Service Management (ICSSSM), Kunming-China, 1-5, 2016.

27. Jin J., Niu B., Ji P., Geng Q., An integer linear programming model of reviewer assignment with research interest considerations, Annals of Operations Research, 1-25, 2018.

28. Metin İ.A., Karasulu B., A novel dataset of human daily activities: Its benchmarking results for classification performance via using deep learning techniques, Journal of the Faculty of Engineering and Architecture of Gazi University, 36 (2), 759-778, 2021.

29. Dönmez İ., Aslan Z., Document sentiment classification using hybrid wavelet methodologies, Journal of the Faculty of Engineering and Architecture of Gazi University, 36 (2), 701-714, 2021.

30. Noyan, T., Kuncan, F., Tekin, R., Yılmaz, K., 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), 1277-1292, 2022.

31. Jang B., Kim I., Kim J.W., Word2vec convolutional neural networks for classification of news articles and tweets, PloS one, 14 (8), 2019.

32. Acı Ç.İ., Çırak A., Türkçe haber metinlerinin konvolüsyonel sinir ağları ve word2vec kullanılarak sınıflandırılması, International Journal of Informatics Technologies, 12 (3), 219-228, 2019.

33. Ciresan D. C., Meier U., Masci J., Gambardella L.M., Schmidhuber J., Flexible, high performance convolutional neural networks for image classification, In IJCAI Proceedings-International Joint Conference on Artificial Intelligence, 22 (1), 2011.

34. Hinton G. E., A practical guide to training restricted boltzmann machines, Neural networks: Tricks of the trade, 599-619, Springer, Berlin, Heidelberg, 2012.

35. Lin M., Chen Q., Yan S., Network in network, arXiv preprint arXiv:1312.4400, 2013.

36. Le Q., Mikolov T., Distributed representations of sentences and documents, In International conference on machine learning, 1188-1196, 2014.

37. Mclean N., Davis J., Utilising semantically rich big data to enhance book recommendation engines, IEEE 18th International Conference on High Performance Computing and Communications, 1434-1441, 2016.

38. Mikolov T., Sutskever I., Chen K., Corrado G.S., Distributed representations of words and phrases and their compositionality, Advances in Neural Information Processing Systems, 26, 2013.

39. Hwang S., Shin J., Extending technological trajectories to latest technological changes by overcoming time lags, Technological Forecasting And Social Change 143, Elsevier, 142-153, 2019.

40. Devlin J., Chang M., Lee K., Toutanova K., BERT: pre-training of deep bidirectional transformers for language understanding, 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 4171-4186, 2018.

41. Tripathy J. K., Sethuraman S. C., Cruz M. V., Namburu A., Mangalraj P., Vijayakumar V., Comprehensive analysis of embeddings and pre-training in NLP, Computer Science Review, 42, 100433, 2021.

42. Masarifoglu M., Tigrak U., Hakyemez S., Gul G., Bozan E., Buyuklu A. H., Özgür A., Sentiment analysis of customer comments in banking using BERT-based approaches, In 2021 29th Signal Processing and Communications Applications Conference (SIU), 1-4, 2021.

43. Taşar D.E., Ozan S., Ozdil U., Akça M.F., Olmez O, Gulum S., Kutal S., Belhan C., Auto-tagging of Short Conversational Sentences using Transformer Methods, 2021 Innovations in Intelligent Systems and Applications Conference (ASYU), 1-6, 2021.

44. Cetiner M., Akgul Y. S., Emsal hukuk dokümanlarının otomatik belirlenmesi, Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 9 (6), 83-94, 2021.