Ankara Üniversitesi
SBF Dergisi,
Cilt 77, No.2, 2022, s. 427-450

Araştırma Makalesi
DOI: 10.33630/ausbf.1103952

MELEZ YÖNTEMLER İLE ANKARA ÜNİVERSİTESİ SİYASAL BİLGİLER FAKÜLTESİ DERGİSİNDE YAYIMLANAN BİLİMSEL MAKALELERİN SINIFLANDIRILMASI *

Doç. Dr. Mehmet Özçalıcı
Kilis 7 Aralık Üniversitesi
İktisadi ve İdari Bilimler Fakültesi
ORCID: 0000-0003-0384-6872

Dr. Öğr. Üyesi Aslı Boru İpek
A. Alparslan Türkeş Bil. Tek. Üni.
Mühendislik Fakültesi
ORCID: 0000-0001-6403-5307

Doç. Dr. Ayşe Tuğba Dosdoğru
A. Alparslan Türkeş Bil. Tek. Üni.
Mühendislik Fakültesi
ORCID: 0000-0002-1548-5237

Doç. Dr. Mustafa Göçken
A. Alparslan Türkeş Bil. Tek. Üni.
Mühendislik Fakültesi
ORCID: 0000-0002-1256-2305

● ● ●

Öz

Teknoloji, sosyal bilimler ve diğer alanlarda yapılan çalışmaların sayısı hızla artmaktadır. Bu nedenle dergilerde bulunan makalelerin sayısı da her geçen gün artış göstermektedir. Dergide bulunan makaleleri manuel olarak sınıflandırmak çok zaman almaktadır. Bu nedenle, belge seviyesinde sınıflandırma, günümüzde farklı uygulama alanlarında çok sayıda metin belgesi bulunması nedeniyle her zaman önemli bir araştırma konusu olmuştur. Bu noktada, yapılandırılmamış metin analizi yapılmalı ve sınıflandırmak için uygun yöntemler tasarlanmalıdır. Verilerin hızlı artışı nedeniyle, sınıflandırma yapmak için güçlü yöntemlere ihtiyaç duyulmaktadır. Bundan dolayı, araştırmacılar güçlü yöntemler ve algoritmalar geliştirmeye çalışmaktadırlar. Yöntemlerin ve algoritmaların başarısı, uygulanan dil, verilerin yapısı, analiz edilecek verinin uzunluğu gibi birçok faktöre bağlıdır. Çalışmamızda destek vektör makinesi (DVM), k-en yakın komşu algoritması (KNN), karar ağacı (KA) ve genetik algoritma (GA) tabanlı melez yöntemler kullanılarak Ankara Üniversitesi Siyasal Bilgiler Fakültesi Dergisi'nde bulunan bilimsel makaleler sınıflandırılmıştır. Ayrıca farklı veri kümeleri kullanılarak önerilen yöntemler karşılaştırılmıştır. Çalışmanın sonuçları önerilen GA tabanlı yöntemlerin minimum %82.5 doğruluk oranı ile belge sınıflandırılmasında başarıyla kullanılabileceğini göstermiştir.

Anahtar Sözcükler: Metin madenciliği, Belge sınıflandırması, Destek vektör makinesi, K-En yakın komşu algoritması, Karar ağacı, Genetik algoritma

Classification of Scientific Articles Published in Ankara University Journal of the Faculty of Political Science with Hybrid Methods

Abstract

The number of studies in technology, social sciences and other fields is increasing rapidly. For this reason, the number of articles in journals is increasing day by day. It takes a lot of time to manually classify the articles in the journal. Therefore, classification at the document level has always been an important research topic because of the large number of text documents in different application areas. At this point, unstructured text analysis should be used and appropriate methods should be designed to classify. Due to the rapid increase of data, strong methods are needed to classify. Hence, researchers are trying to develop powerful methods and algorithms. The success of the methods and algorithms depends on many factors such as the applied language, the structure of the data, and the length of the data to be analyzed. In our study, scientific articles in Ankara University Journal of the Faculty of Political Science are classified using hybrid methods based on support vector machine (DVM), k-nearest neighbor algorithm (KNN), decision tree (KA) and genetic algorithm (GA). In addition, the proposed methods were compared using different data sets. The results of the study showed that the proposed GA based methods can be successfully used in document classification with a minimum accuracy of 82.5%.

Keywords: Text mining, Document classification, Support vector machine, K-Nearest neighbor algorithm, Decision tree, Genetic algorithm

---

* Makale geliş tarihi: 10.07.2020
Makale kabul tarihi: 02.08.2021
Erken görünüm tarihi: 15.04.2022

Melez Yöntemler ile Ankara Üniversitesi Siyasal Bilgiler Fakültesi Dergisinde Yayımlanan Bilimsel Makalelerin Sınıflandırılması

Giriş

Teknoloji ilerledikçe elektronik cihazların kullanımı ve veri miktarı gün geçtikçe artmaktadır. Verilerin hızla ve büyük miktarda artması sonucunda üretilen verilerden değerli bilgilerin çıkarılması gerekmektedir. Doğal dil işleme araçları, büyük miktarda dijital metin verisinden otomatik olarak bilgi alma veya çıkarma gibi metin işleme uygulamaları için kullanılmaktadır. Bu uygulamaların yardımıyla insanlar, kişisel web siteleri, bloglar ve mikro bloglar, kurumsal web sitesi, haber siteleri, forumlar vb. bilgi kaynaklarının yaygınlaşması ile birlikte hızla artan belgelerini düzenlemenin ve sınıflandırmanın daha kolay yollarını bulabilmektedirler. Ayrıca, otomatik metin sınıflandırma yöntemleri, veri kümesini ilgili alt kategorilere ayırmaya veya sınıflandırmaya yardımcı olmaktadır (Deniz ve Kiziloz, 2017: 655). Genel olarak, metin sınıflandırma sistemi dört farklı seviyede incelenmektedir. Belge seviyesinde, algoritma tüm belgeyi ilgili kategorilere ayırabilmektedir. Paragraf seviyesinde algoritma tek bir paragrafın (belgenin bir kısmı) ilgili kategorilerini incelemektedir. Cümle düzeyinde tek bir cümlenin (paragrafın bir bölümü) ilgili kategorileri incelenmektedir. Son olarak alt cümle seviyesinde algoritma bir cümle içinde (cümlenin bir kısmı) ilgili alt ifade kategorilerini incelemektedir (Kowsari vd., 2019: 2).

İnternetin kullanılmaya başlanmasıyla birlikte, sosyal yaşamda, yaşam tarzında ve insanların kararlarında büyük bir değişiklik olmuştur. İnsanlar bir konu hakkındaki görüşlerini internet üzerinden ifade etmeye başlamıştır. Dolayısıyla internet uygulamaları yoluyla üretilen belgelerin miktarı günden güne artmaktadır. Belgeler genellikle yapılandırılmamış veri olduğundan, bu verilere erişmek ve bunları yönetmek neredeyse imkânsızdır. Bu nedenle, otomatik belge sınıflandırma işlemine ihtiyaç duyulmaktadır.

Belge sınıflandırma işleminde kullanılan metin verileri, çoğu senaryoda üretilebilecek en basit veri biçimlerinden biri olan yapılandırılmamış bilgilere iyi

bir örnektir. Yapılandırılmamış metin verilerinin manuel analizi pratik değildir ve metin verilerinin çeşitli uygulamalarda etkili bir şekilde işlenmesi için güçlü yöntemler ve algoritmalara ihtiyaç duyulmaktadır. KA, Bayes yöntemleri, KNN algoritması, DVM ve yapay sinir ağları gibi hemen hemen tüm sınıflandırma teknikleri metin verilerinin sınıflandırılmasında kullanılabilir (Brindha vd., 2016: 4). Yöntemlerin başarısı, verilerdeki karmaşık modelleri ve doğrusal olmayan ilişkileri anlama kapasitelerine dayanmaktadır. Örneğin, makalelerin dergilere veya araştırma düzeyine göre sınıflandırılmasında kullanılacak yöntemin karmaşık metin verilerini analiz edebiliyor olması gerekmektedir. Bu nedenle, araştırmacılar güçlü yöntemler ve algoritmalar geliştirmeye çalışmaktadırlar. Çalışmamızda makaleleri özet bölümüne göre ilgili kategorilere sınıflandırmak için farklı yöntemler tasarlanmıştır. DVM, KNN algoritması, KA ve GA tabanlı melez yöntemler kullanılarak makaleler önceden tanımlanmış sabit kategorilere sınıflandırılmaktadır.

1. Önceki Çalışmalar

Metin madenciliği, veri madenciliği alanındaki bir disiplindir ve metin sınıflandırmasını gerçekleştirmek için algoritmalar sunmaktadır (Kılınç, 2016: 215). Metin madenciliği uygulamaları, büyük metin belgelerinden bilgi elde etmek için gereken çabaları azaltmaktadır. Metin madenciliğinde metin özetleme, bilgi çıkarma, anlamsal ilişkileri keşfetme vb. sorunlar bulunduğu için belgeyi belirli bir etki alanında sınıflandırmak önemlidir. Metin sınıflandırma tekniklerinde mevcut olan sorunların farkına varmak çok önemlidir, böylece farklı yöntemlerin performanslarını değerlendirmek daha kolay olacaktır.

Nanba vd. (2000), atıf bağlantıları ve atıf türlerini kullanarak araştırma belgelerini otomatik olarak veritabanında sınıflandırmak için çeşitli yöntemler sunmuşlardır. Sonuçlar, bibliyografik birlikteliğe dayanarak oluşturulan yöntemin diğerlerinden daha etkili olduğunu göstermiştir. Türkoğlu vd. (2007) çalışmalarında WEKA'da bulunan beş farklı sınıflandırma algoritmasını (Naïve Bayes, DVM, rastgele orman, KNN ve çok katmanlı algılayıcı) metin sınıflamasında kullanmışlardır. Tüm algoritmalar WEKA'nın varsayılan parametreleriyle çalıştırılmıştır. Kullanılan yöntemlerin etkinliği, 10-kat çapraz doğrulama kullanılarak değerlendirilmiştir.

Prabowo ve Thelwall (2009) çalışmalarında kural tabanlı sınıflandırma, denetimli öğrenme ve makine öğrenimini yeni bir melez yöntemde birleştirmişlerdir. Önerilen yöntem film değerlendirmelerinde, ürün değerlendirmelerinde ve MySpace yorumlarında test edilmiştir. Ayrıca, ID3 ve RIPPER olmak üzere iki indüksiyon algoritmasının performansı değerlendirilmiştir. Küçük ve Yazıcı (2010) çalışmalarında Türkçe metinler için varlık tanıyıcı adlı bir melez yöntem sunmuşlardır. Önerilen yöntemler, haber

metinleri, finansal haber metinleri, çocuk hikayeleri ve tarihsel metinler olmak üzere dört farklı veri kümesinde değerlendirilmiştir. Performans değerlendirmeleri melez yöntemin daha iyi olduğunu göstermiştir.

Torunoğlu vd. (2011), metin sınıflandırmasında önişlemenin Türkçe metin üzerindeki etkisini geniş veri kümesi kullanarak analiz etmişlerdir. Etkisiz kelimeler, kök bulma ve kelime ağırlıklandırmada kullanılan yaygın ön işleme yöntemlerini kullanarak kapsamlı analiz yapmışlardır. Farklı sınıflandırma algoritmalarının performansları karşılaştırılmıştır. Çalışmada, kök bulma yöntemi olarak Zemberek ve sabit önek kök bulma yöntemi kullanılmıştır. Çalışmada sınıflandırıcı olarak Naïve Bayes, Naïve Bayes Multinomial, DVM ve KNN kullanılmıştır. Bu algoritmaları çalıştırmak için WEKA makine öğrenimi yazılımı kullanılmıştır.

Çetin ve Amasyalı (2012), metinleri temsil etmek için sınıfların dağılımını dikkate alan Delta1 yöntemi adı verilen denetimli terim ağırlıklandırma yöntemini kullanmışlardır. Çalışmalarında telekomünikasyon şirketine ait veri kümesi kullanılmıştır. Ayrıca, denetimli terim ağırlıklandırma yönteminin, geleneksel denetimsiz yöntemlerden daha uygun olduğu bulunmuştur. Doddi vd. (2014), yapılandırılmamış büyük hacimli veriyi değerlendirmek için DVM kullanmışlardır. DVM'de, iki sınıf arasındaki hiper düzlemin iki sınıf arasındaki mesafeyi en üst düzeye çıkaracak şekilde iki sınıfa sınıflandırılması istenmektedir. Bu, görülmeyen noktaların daha iyi sınıflandırılmasını sağlamıştır.

Yıldırım (2014) çalışmalarında beş farklı özellik seçim yöntemiyle dört farklı makine öğrenme algoritması uygulamıştır. Çalışmada dört farklı özellik alanı içinde dört farklı terim ağırlıklandırma şeması da kullanılmıştır. Uysal ve Gunal (2014), tokenizasyon, etkisiz kelimeler filtresi, küçük harf dönüşümü ve kök bulma dahil olmak üzere, metin sınıflandırmasının dört yaygın önişleme adımının etkilerini analiz etmişlerdir. Önişleme yöntemlerini değerlendirmek için iki farklı alan (e-posta ve haber) ve iki farklı dil (Türkçe ve İngilizce) dikkate alınmıştır. Sonuçlar, metin sınıflandırmasındaki önişleme adımının, özellik çıkarma, özellik seçimi ve sınıflandırma adımları kadar önemli olduğunu göstermiştir. Küçük (2015), Vikipedi makale başlıklarından Türkçe'de adlandırılmış varlık tanıma için dil kaynaklarını derlemeye yönelik otomatik bir prosedür sunmuştur. Başlıkları otomatik olarak sınıflandırmak için KNN algoritması kullanılmıştır. Kılınç (2016) topluluk öğrenme modellerinin Türkçe metin sınıflandırmasına etkisini incelemiştir. Dört sınıflandırıcı (Naïve Bayes, DVM, KNN, J48 Karar ağacı) ve üç topluluk modeli (Bagging, Boosting ve Rotation Forest) deneysel olarak değerlendirilmiştir. Deneysel sonuçlar, topluluk öğrenme modellerinin temel sınıflandırıcıların başarısını artırarak genellikle daha doğru sonuçlar verdiğini göstermektedir. Tüm deneysel değerlendirmeler WEKA veri madenciliği aracı kullanılarak yapılmıştır. Her sınıflandırıcı model,

performans tahmini için iyi bilinen bir strateji olan 10-kat çapraz doğrulama kullanılarak değerlendirilmiştir. Kilimci vd. (2016), topluluk algoritmalarının denetimli metin kategorizasyonu üzerindeki etkinliğini gözlemlemişlerdir. Sınıflandırma performansları Naïve Bayes modelinin iki versiyonu, DVM, KA sınıflandırıcıları ve dört homojen topluluk sınıflandırıcısı dikkate alınarak İngilizce ve Türkçe metin belgelerde değerlendirilmiştir.

Deniz ve Kızıloz (2017) çalışmalarında Türkçe belgelerin yazar, tür ve cinsiyet açısından metin sınıflandırması için hem karakter hem de kelime düzeyi n-gram modellerini uygulamışlardır. Bu sınıflandırmayı gerçekleştirmek için Naive Bayes, DVM ve Rastgele Orman olmak üzere üç makine öğrenimi tekniğinden yararlanmışlardır. Teknikler WEKA'nın varsayılan parametreleriyle oluşturulmuştur. Ayrıca çalışmada önişleme tekniklerinin Türkçe metin kategorizasyonu üzerindeki etkileri de incelenmiştir. Kök bulma işlemi için Zemberek-NLP1 kütüphanesini kullanmıştır. Gezici ve Yanıkoğlu (2018), film değerlendirmelerinin Türkçe sınıflandırılmasında farklı özelliklerin etkinliğini araştırmışlardır. Daha sonra sözlük veri boyutunun aynı veri kümesindeki yorumların genel duygusunu tespit etme üzerindeki etkisi araştırılmıştır. Önerilen yöntem denetimli öğrenme ve sözlük tabanlı yaklaşımları birleştirmektedir. Dönmez ve Adalı (2018), sınıflandırma ve özellik seçimi algoritmaları için WEKA yazılım paketini kullanmışlardır. Ayrıca, cümleleri ayrıştırmak için doğal dil işleme aracı kullanılmıştır. Cümleler İTÜ NLP web servisi tarafından ön işlemden geçirilmiştir. Çalışmada, belge sınıflandırması için Türkçe cümlelerin anlamsal matris gösterimini kullanan yeni bir yöntem sunulmaktadır. Urologin (2018) çalışmasında BBC haber makalelerinde metin özetleme ve duygu analizini ele almıştır. Çalışmada metin özetleme yöntemi geliştirilmiş ve duygu bilgisini belirlemek için VADER (for Valence Aware Dictionary forsEntiment Reasoning) kullanılmıştır. Ayrıca VADER'den elde edilen duygu bilgilerinin 3 boyutlu görselleştirmeleri verilmiştir.

Kilimci ve Akyokus (2018), heterojen topluluk sistemi oluşturmak için temel sınıflandırıcılar olarak çok değişkenli Bernoulli Naïve Bayes, çok terimli Naïve Bayes, DVM, rastgele orman, evrişimsel sinir ağı kullanmışlardır. Çalışmada, sekiz farklı veri kümesinde üç farklı belge sunum yöntemi kullanılarak yöntemlerin performansı değerlendirilmiştir. Gürbüz ve Aydın (2018), belge sınıflarını otomatik olarak belirlemek için Naïve Bayes'i kullanmışlardır. Bulut bilişim altyapısı, Türkçe bilimsel belgelerini analiz etmek için kullanılmıştır. Apache Mahout kütüphanesi ve Zemberek uygulanmıştır. TÜBİTAK Dergipark web sitesinde verilen kategorilerin her biri bir sınıf olarak kabul edilmiş ve beş sınıf elde edilmiştir.

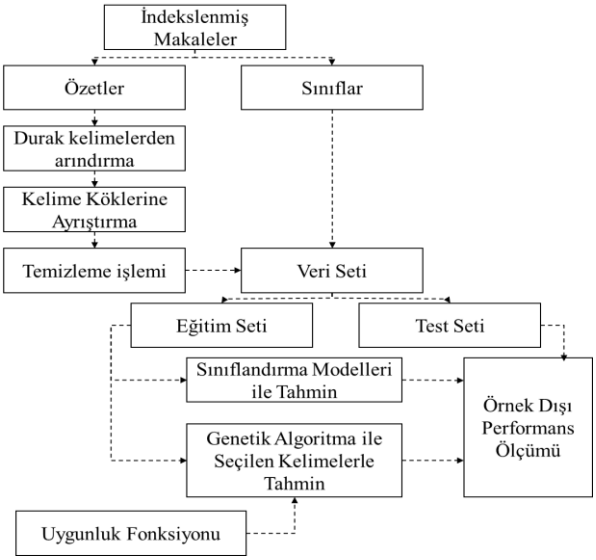
Literatürde yapılan çalışmalar incelendiğinde, çalışmamızın genel olarak katkıları: (1) özet bölümü kullanılarak dergide yayımlanan makalelerin önceden tanımlanmış sabit kategorilere sınıflandırılması; (2) DVM, KNN algoritması ve

KA yöntemi kullanılması; (3) genetik algoritma tabanlı DVM (GA-DVM), genetik algoritma tabanlı KNN (GA- KNN) ve genetik algoritma tabanlı KA yöntemi kullanılarak yeni yöntemler tasarlanması; (4) önerilen yöntemlerin farklı veri kümelerinde kıyaslanmasıdır.

2. Önerilen Metodoloji

Metin sınıflandırması, metin belgelerini otomatik olarak bir veya daha fazla önceden tanımlanmış sınıfa atama işlemi olarak tanımlanabilir. Basit bir metin sınıflandırma sistemi üç aşamaya göre oluşturulabilir. Bu aşamalar, metin önişleme, sınıflandırma ve değerlendirmedir. Metin önişleme, gereksiz gürültüyü gidermek için metin verilerinin temizlenmesi işlemidir. Çoğu metin veri kümesi birçok gereksiz kelime içermektedir. Bu gereksiz kelimelerin sistem performansı üzerinde olumsuz etkileri olabilmektedir. Sınıflandırma, bir konuda karar almayı desteklemek için metin bilgilerini kullanılabilir hale getirmek olarak tanımlanabilir. Metin sınıflandırma modellerinin performansı büyük ölçüde toplanan veri setlerine ve verilerin kalitesine bağlıdır. Bu nedenle, çalışmamızda önerilen yöntemler farklı veri setlerine de uygulanmıştır. Ankara Üniversitesi Siyasal Bilgiler Fakültesi Dergisi (Ankara Üniversitesi SBF Dergisi)'nde bulunan 1943-2011 yılları arasında makaleleri sınıflandırmak için DVM, KNN, KA, GA-DVM, GA-KNN ve GA-KA kullanılmıştır. Önerilen yöntemlerin genel yapısı Şekil 1'de verilmiştir.

Şekil 1: Çalışmanın modeli



2.1. Kullanılan Yöntemler

2.1.1. Destek Vektör Makinesi (DVM)

DVM, doğrusal iki sınıflı sınıflandırıcıya bir örnektir. Doğrusal bir sınıflandırıcı, aşağıda bulunan denklemde ifade edilen doğrusal ayırma işlevine dayanmaktadır.

$$f(x) = w^T x + b \quad (1)$$

w vektörü, ağırlık vektörü olarak tanımlanır ve b ise sapmadır. Doğrusal bir sınıflandırıcı tanımlamak için, iki vektör arasındaki nokta çarpımı gereklidir. Bu da $w^T x = \sum_i w_i x_i$ olarak tanımlanmaktadır. x_i işareti, bir veri kümesindeki $i.$ vektörünü temsil etmektedir. DVM ile ilgili detaylar Ben-Hur ve Weston (2010) çalışmalarında mevcuttur.

DVM tekniği ikili sınıflandırmaya olanak sağlamaktadır. Çalışmada 5 gruptan oluşan bir sınıflandırma işlemi gerçekleştirilmek istenmektedir. Bu nedenle de Error-Correcting Output Codes (ECOC) isimli ve DVM tabanlı sistem kullanılmıştır. ECOC modeli, ikili sınıflandırma modellerinden oluşan ve çok sınıflı öğrenmeye olanak sağlayan bir sistemdir (Mathworks, 2020). DVM ise sadece ikili sınıflandırmaya olanak sağlamaktadır.

2.1.2. K-en Yakın Komşu (KNN) Algoritması

KNN algoritması, veri madenciliğinin en basit yöntemlerinden biridir. KNN algoritmasının uygulanması için izlenecek genel yol aşağıdaki gibidir (Şekil 2).

Şekil 2: KNN algoritması (Imandoust ve Bolandraftar, 2013: 606)

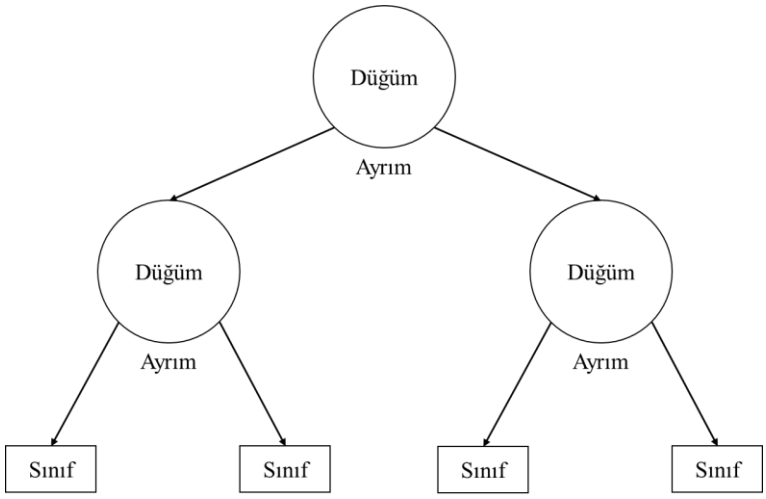
for tüm bilinmeyen örneklem ($BMÖ_{(i)}$)
  for tüm bilinen örneklem ($BÖ_{(j)}$)
    $BMÖ_{(i)}$ve $BÖ_{(j)}$ arasındaki uzaklığı hesapla
  end for
  k en kısa mesafeyi bul
  İlgili numuneyi (Örneklem (j_1),…, Örneklem (j_k) yerleştir
  $BMÖ_{(i)}$'yi daha sık görünen sınıfa ata
end for

Sınıflandırmanın performansı, komşu büyüklüğünün seçiminden etkilenmektedir. K çok küçük seçilirse, gürültülü, belirsizlik veya yanlış etiketlenmiş noktalar nedeniyle yerel tahmin kötü olabilir. Diğer yandan, K değerinin çok arttırılması da, sınıflandırma performansını düşürebilmektedir. Bu nedenle, uygun K değerinin seçilmesi tasarlanacak algoritmanın sınıflandırma performansını büyük ölçüde etkileyen önemli bir konudur (Imandoust ve Bolandraftar, 2013). KNN algoritması ile ilgili detaylar Imandoust ve Bolandraftar (2013) çalışmasında mevcuttur.

2.1.3. Karar Ağacı

Karar ağacı, örnek uzayının özyinelemeli bölümü olarak ifade edilen bir sınıflandırıcıdır. Karar ağacı, köklü bir ağaç oluşturan düğümlerden oluşur. Bir karar ağacında, her iç düğüm, girdi değerlerinin belirli ayrık fonksiyonuna göre örneklem alanını iki veya daha fazla alt alana ayırır. Karar ağaçları parametrik olmayan bir yöntem olarak kabul edilir. Bu, karar ağaçlarının alan dağılımı ve sınıflandırıcı yapısı hakkında hiçbir varsayımı olmadığı anlamına gelir. Karar ağaçları oluşturulurken takip edilmesi kolaydır. Karar ağaçları hem nominal hem de sayısal girdi niteliklerini işleyebilmektedir. Karar ağacı gösterimi, herhangi bir ayrık değer sınıflandırıcısını temsil edecek kadar zengindir. Karar ağacı ile ilgili ayrıntılı bilgiler Rokach and Maimon (2005) çalışmasında mevcuttur. Basit bir karar ağacı yapısı Şekil 3'te verilmiştir.

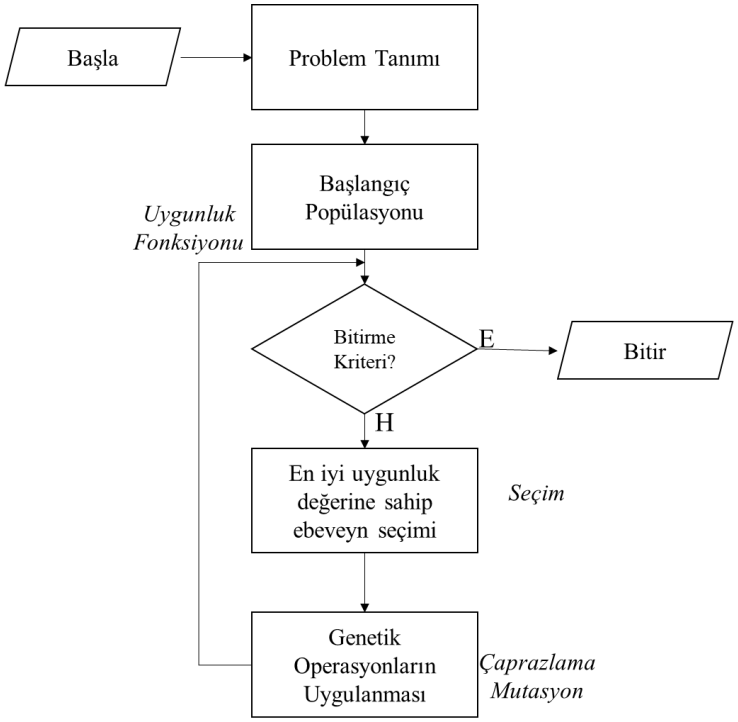
Şekil 3: Basit bir karar ağacı yapısı



2.1.4. Genetik Algoritma

GA, hem problemleri çözmek hem de evrimsel sistemleri modellemek için kullanılabilecek bir arama yöntemidir. GA'nın temel fikri çok basittir. Birincisi, bilgisayarda bir popülasyon yaratılır ve daha sonra bu popülasyon varyasyon, seçim ve kalıtım ilkeleri kullanılarak geliştirilir. Genetik algoritmalar popülasyon genetiğinden gelen fikirlere dayanmaktadır; bellekte depolanan genotip popülasyonları (bir bireyin genetik materyali), bu genotiplerin farklı şekilde çoğaltılması, mutasyon ve çaprazlamanın biyolojik süreçlerine benzer süreçler tarafından oluşturulan varyasyonlar içermektedir (Forrest, 1993). GA çalışma prensibi Şekil 4'te verilmiştir.

Şekil 4: GA Çalışma Prensibi



2.2. Veri Kümesi

Ankara Üniversitesi Siyasal Bilgiler Fakültesi Dergisi (Ankara Üniversitesi SBF Dergisi), siyaset bilimi, kamu yönetimi, uluslararası ilişkiler, iktisat, işletme, maliye, çalışma ekonomisi ve endüstri ilişkileri konu alanlarında özgün araştırma makaleleri, yorum yazıları ve kitap eleştirileri yayımlayan

hakemli bilimsel bir dergidir (https://dergipark.org.tr/tr/pub/ausbf/aim-and-scope). Dergi, Türkiye'deki sosyal bilimlerdeki en eski dergilerden biridir ve 1943 yılından itibaren bilimsel makale yayımlamaktadır. Dergi üç ayda bir yayımlanmaktadır.

Ankara Üniversitesi SBF Dergisi için 1943-2011 yılları arasında dergide yayımlanan makalelerin dizinleri yayımlanmıştır (Tellal vd., 2012). Endeks belgesi, 1943-2010'un tüm sayılarını ve 2011'in ilk iki sayısını kapsamaktadır. 1943-2010 arasında yayımlanan makaleler altı sınıfta ((1) Çalışma Ekonomisi ve Endüstri İlişkileri, (2) İktisat, (3) Maliye, (4) İşletme, (5) Siyaset Bilimi ve Kamu Yönetimi ve (6) Uluslararası İlişkiler) değerlendirilmiştir. Sınıflandırma süreci uzmanlar tarafından gerçekleştirilmiştir. Daha açık bir ifade ile bir grup uzman, dergide yayımlanan makaleleri yukarıda bahsi geçen altı sınıfa atamıştır.

1999'un ilk sayısından itibaren makalelerde özet bölümü bulunmaktadır. 1999'dan önce yayımlanan makaleler açık olarak özet bölümü içermemektedir. Sınıflandırma süreci özet dikkate alınarak yapıldığından, 1999'dan önceki çalışmalar dikkate alınmamıştır. Makalelerin bazıları İngilizce, Almanca veya Fransızca gibi yabancı bir dilde yayımlanmaktadır. Derginin yazım kurallarına göre, yazar(lar) yabancı dilde makale yazmayı seçerse, makalenin Türkçe özetini hazırlamak zorundadır. Yabancı dil yayınlarının Türkçe özetleri kullanılmıştır.

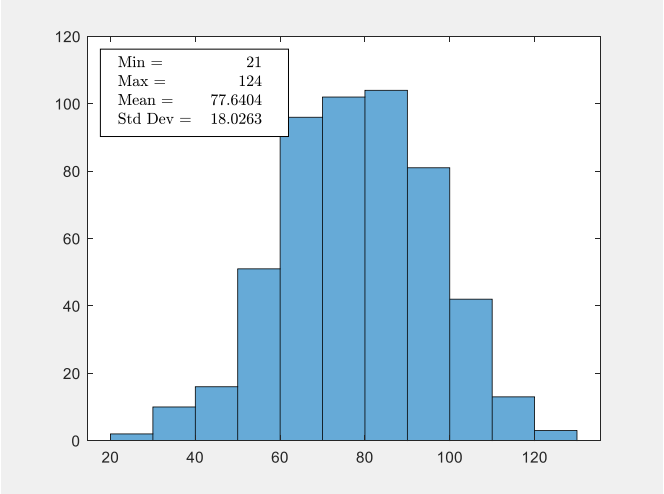
Bu dönemler arasında 432 makale bulunmaktadır. Özetlerden bazıları derginin internet sitesinde txt formatında bulunmaktadır. Ancak, bazı durumlarda yalnızca taranmış PDF belgesi mevcuttur. Bu gibi durumlarda, özetler yazarlar tarafından txt formatında yazılarak veri kümesine girdi olarak eklenmiştir.

Makalelerin sadece beşinin maliye alanında olduğu tespit edilmiştir. Maliye kategorisi, yayımlanan makalelerin az olması nedeniyle çalışmadan çıkarılmıştır. Sonuç olarak, analizde kullanılmak üzere hazır bulunan $432 - 5 = 427$ tane özet mevcuttur. Bu 427 tane özetin her birisinin hangi kategoride olduğu uzmanlar tarafından belirlenmiştir. Bu veri seti analizi gerçekleştirmek ve analizin performansını ölçmek için kullanılmıştır. Bununla birlikte 2011 yılının son iki sayısında ve 2012-2014 yılları arasındaki sayılarda yayımlanan 93 tane makalenin hangi kategoriye ait olduğu uzmanlar tarafından belirlenmemiştir. Çıktı değişkeni söz konusu değildir. Çalışmanın amacı doğrultusunda eğitim setinde geliştirilen model, bu 93 tane özeti otomatik olarak sınıflandırmak için kullanılacaktır. Söz konusu özetlerin hangi kategoriye ait oldukları bilinmediğinden, performans ölçümü gerçekleştirilemeyecektir. Bununla birlikte model yardımıyla belirlenen kategorileri kullanmak suretiyle uzmanların makaleleri sınıflandırma faaliyetlerinin önemli derecede kolaylaşacağı tahmin edilmektedir.

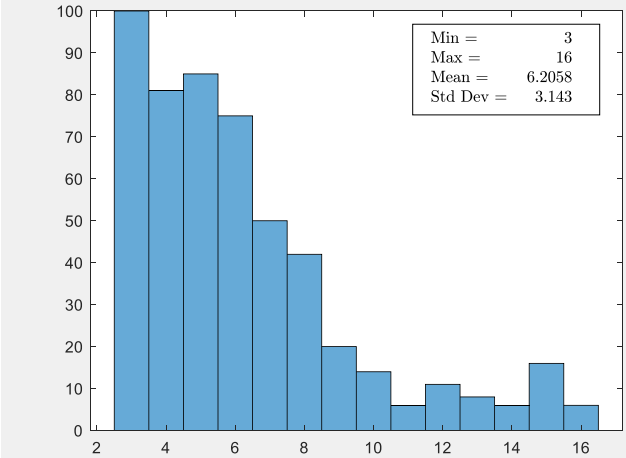
Özetlerdeki kelime sayısının histogramı Şekil 5'te gösterilmiştir. Bazı önemli tanımlayıcı istatistikler de Şekil 5'te yer almaktadır. Özetlerde bulunan

cümle sayısı Şekil 6'da verilmiştir. Özetlerde ortalama yaklaşık 6 cümle bulunmaktadır.

Şekil 5: Özetlerdeki kelime sayısının dağılımı



Şekil 6: Özetlerdeki cümle sayısının dağılımı



2.3. Veri Temizleme

Önceki bölümde belirtildiği gibi, 427 makale özeti vardır ve bu özetler analize hazırdır. Veri temizleme sürecinin ilk adımı olarak, her bir özet tüm

karakterlerin küçük harflerle yazılması için düzenlenmiştir. Özetlerin içerisinde toplamda 15436 kelime olduğu tespit edilmiştir. Bir sonraki adımda etkisiz kelimeler özetlerden çıkarılmıştır. Etkisiz kelimeleri, veri madenciliği uygulamaları için çok önemli olmayan ve bir dilde sıkça ortaya çıkan kelimelerdir (Aggrawal, 2015:431). Örneğin, Türkçede "hepsi", "bu", "ve", "veya" kelimeleri de etkisiz kelimeler olarak düşünülebilir. Türk edebiyatı araştırılarak 371 maddeden oluşan etkisiz kelime listesi oluşturulmuştur. Çalışmamızda 15436 kelimeden 121'i etkisiz kelime olarak kabul edilmiş ve veri kümesinden çıkarılmıştır. Bu adımdan sonra, veri kümesi 15436-121 = 15315 kelime içermektedir.

Bir sonraki adım kök bulmadır. Kök bulma, kelimelerden yaygın kök çıkarmayı ifade eder ve çıkarılan kök kendi başına bir kelime bile olmayabilir (Aggrawal, 2015:431). Örneğin, çoğul ya da tekil kelime biçimleri aynı kelimeyi ifade etmektedir. Kök bulma işlemi ile aynı kelimeleri ortadan kaldırmak, veri kümesinin karmaşıklığını azaltacaktır.

Bu çalışmada Osman Tunçelli'nin Python programlama dilinde hazırladığı bir kütüphane kök bulma işlemi amacıyla kullanılmıştır (https://github.com/otuncelli/turkish-stemmer-python). Veri kümesinde 15315 kelime vardı, ancak kök bulma işlemi uygulanarak, 5114 kelimenin aslında diğer kelimelerle aynı köke sahip olduğu belirlenmiştir. Dolayısıyla, bu 5114 kelime veri kümesinden çıkarılmış ve böylece veri kümesinde 15315-5114 = 10201 farklı kelime kalmıştır. En sık bulunan kökler Tablo 1'de verilmiştir.

Tablo 1: En Sık Bulunan Kökler

Kök	Sayı	Kök	Sayı	Kök	Sayı
Çalışma	405	Sonuç	229	Analiz	185
Ara	295	Konu	226	Etki	180
Temel	252	Alan	204	Politika	170
Orta	239	Ülke	200	Son	146
Önem	238	Süreç	196	Devlet	140
El	232	Yıl	185		

2.4. Doküman-Terim Matrisi (DTM)

DTM, belgede bir kelimenin bulunma sıklığını tanımlayan bir matristir. DTM'de, satırlar koleksiyondaki belgelere ve sütunlar kelimelere karşılık gelir. DTM gösterimi, belgeleri sayısal bir yapı olarak göstermenin oldukça basit bir

yoludur. Metni sayısal bir yapı olarak göstermek, metin madenciliği için önemlidir. DTM oluşturmak analizdeki ana adımdır.

Çalışmamızda veri temizleme işleminden sonra, DTM oluşturulmuştur. DTM'de 427 satır (özet sayısı) ve 10201 sütun (kelime sayısı) vardır. Varsayımsal bir DTM örneği aşağıdaki tabloda verilmiştir. Sayılar gerçek DTM'de farklı olabilir.

DTM (Tablo 2) rastgele iki alt kümeye ayrılmıştır. Eğitim seti 347 özetten oluşmaktadır ve geriye kalan 80 özet ise test amacıyla kullanılmıştır. Eğitim ve test setinin tanımlayıcı istatistikleri Tablo 3'te verilmiştir.

Tablo 2: DTM örneği

Doküman-Terim	**1**	**2**	**3**	**…**	**10199**	**10200**	**10201**
	Yüzyıl	**Son**	**İhtimal**	**…**	**Örf**	**Fiyat**	**Bono**
D001	1	0	1	…	1	0	0
D002	0	0	0	…	0	1	1
…	…	…	…	…	…	…	…
D427	1	1	0		1	0	0

Tablo 3: Eğitim ve Test Setindeki Kategori Sayısı

		ÇEKO	**IKT**	**İŞL**	**SBKY**	**UAİ**
Eğitim Seti	**Sayı**	30	50	65	139	63
	Yüzdesi	8.65	14.41	18.73	40.06	18.16
Test Seti	**Sayı**	8	10	20	32	10
	Yüzdesi	10.00	12.50	25.00	40.00	12.50

Dergide bulunan bilimsel makaleler beş sınıfa ((1) Çalışma Ekonomisi ve Endüstri İlişkileri (ÇEKO), (2) İktisat (IKT), (3) İşletme (İŞL), (4) Siyaset Bilimi ve Kamu Yönetimi (SBKY) ve (5) Uluslararası İlişkiler (UAİ)) ayrılmıştır.

3. Bulgular ve Tartışma

Her gün üretilen metin miktarı önemli ölçüde artmaktadır. Bu nedenle, araştırmacılar metinlerden anlamlı bilgiler elde edebilmek için metin madenciliğine yönelmişlerdir. Üretilen verileri analiz edebilmek için güçlü teknikler ve algoritmalar gerekmektedir. Çalışmamızda dergide yayımlanan makalelerin özet bölümü kullanılarak makaleler sınıflandırılmıştır. Bilimsel dergilerin arşivleri, her sayıda yeni makale eklendiğinden önemli ölçüde

büyümektedir. Bu büyüme bir taraftan araştırmacıların daha fazla bilimsel bilgiye kolayca erişmesini sağlarken diğer taraftan ilgi alanlarına uygun makaleleri tanımlamalarını oldukça zorlaştırmıştır. Bu nedenle, çalışmamızda farklı yöntemler tasarlanarak sınıflandırma performansı iyileştirilmeye çalışılmıştır.

3.1. Hata Matrisi Sonuçları ve Modellerinin Karşılaştırmalı Analizi

Kullanılan yöntemlerin performansının karşılaştırılabilmesi için farklı performans ölçütleri kullanılabilir. Hata matrisi, sınıflandırma yöntemlerinin performansının özetlenmesi için kullanılan performans ölçütlerinden biridir. Hata matrisinde doğru ve yanlış tahminlerin sayısı verilmektedir. Sonuçların daha hızlı analiz edilmesi ve daha kolay okunmasını sağlamak için kullanılabilir. Hata matrisinde ilk olarak, veri kümesini beklenen sonuç değerleriyle test etmeniz gerekmektedir. Daha sonra, her sınıf için doğru tahmin sayısı ve her sınıf için yanlış tahmin sayısı belirlenmektedir. Çalışmamızda kullanılan yöntemlerin hata matrisi aşağıda verilmektedir (Tablo 4-6).

Önerilen DVM yöntemi parametrik olmayan bir sınıflandırma tekniğidir. İkiden fazla gruplarda sınıflandırmaya olanak sağlayan DVM yöntemini oluşturmak için Matlab'da yer alan çok sınıflı model kullanılmıştır. DVM eğitiminde doğrusal çekirdek fonksiyonu (linear kernel function) kullanılmıştır. DVM'nin en güçlü yanlarından biri eğitiminin nispeten daha kolay olmasıdır. DVM, karmaşık veri setlerinde başarılı olan yöntemlerden biridir. DVM'nin performansı hata matrisi kullanılarak değerlendirilmiştir. Sınıflandırılmış veri kümesi için genel doğruluk hesaplanmış ve sonuçlar karşılaştırılmıştır.

Tablo 4: DVM'nin Sınıflandırma Sonuçları (Hata Matrisi)

		Gerçek Değerler					
		ÇEKO	**IKT**	**İŞL**	**SBKY**	**UAİ**	**Toplam**
Modelin Tahmin Ettiği Değerler	**ÇEKO**	2	3	1	3	0	8
	IKT	0	4	2	4	0	10
	İŞL	0	0	14	6	0	20
	SBKY	1	1	0	30	0	32
	UAİ	0	0	0	6	4	10
	Toplam	3	8	17	49	4	80

Tablo 5: KNN Sınıflandırma Sonuçları (Hata Matrisi)

		Gerçek Değerler					
		ÇEKO	**IKT**	**İŞL**	**SBKY**	**UAİ**	**Toplam**
Modelin Tahmin Ettiği Değerler	**ÇEKO**	3	3	0	1	1	8
	IKT	0	3	4	2	1	10
	İŞL	2	1	16	1	0	20
	SBKY	1	3	0	25	3	32
	UAİ	0	0	0	2	8	10
	Toplam	6	10	20	31	13	80

Sınıflandırma için kullanılan KNN, parametrik olmayan bir yöntemdir. KNN yönteminde, birden fazla sınıf aynı en düşük değere sahipse, bağlı gruplar arasındaki en küçük dizin kullanılır. Çalışmamızda komşu sayısı 6'ya sabitlenmiştir. Bu değerin anlamı bakılacak eleman sayısını ifade etmektedir. Daha sonra, mesafe fonksiyonu kullanılarak sıralanır ve uygun sınıfa atama yapılır. Mesafe fonksiyonu, 1 eksi gözlemler arasındaki açının kosinüsü olarak seçilmiştir. KNN yönteminde parametreler, deneme yanılma yöntemi ile belirlenmiştir.

KA'nın parametreleri aşağıdaki gibidir: Maksimum kategori seviyesi sayısı 10 ile minimum dal düğümü sayısı 10'dur. KA'nın parametreleri, deneme yanılma yöntemi ile belirlenmiştir. Sınıflandırma doğruluğunu ölçebilmek için hata matrislerinden genel doğruluk değerleri hesaplanmıştır.

GA kullanılarak DVM, KNN ve KA'nın parametreleri optimize edilmiştir. GA göstergeleri (uygunluk fonksiyonu değerlendirme sayısı, nesil sayısı, CPU zamanı, ilk değişken sayısı, seçilen kelime sayısı, indirgeme faktörü) Tablo 7'de verilmiştir. Tüm modellerin kelime sayısı yarıya indirilmiş ve veri kümesinin karmaşıklığı azaltılmıştır. İndirgeme Faktörü (%) aşağıdaki denklem kullanılarak hesaplanmıştır.

$$\text{İ}ndirgeme\ Faktörü\ (\%) = \frac{Seçilen\ Kelime\ Sayısı}{\text{İ}lk\ Değişken\ Sayısı} \quad (2)$$

Tablo 6: KA Sonuçları (Hata Matrisi)

		Gerçek Değerler					
		ÇEKO	**IKT**	**İŞL**	**SBKY**	**UAİ**	**Toplam**
Modelin Tahmin Ettiği Değerler	**ÇEKO**	2	0	2	2	2	8
	IKT	1	3	2	4	0	10
	İŞL	1	0	14	3	2	20
	SBKY	2	1	5	21	3	32
	UAİ	0	1	1	4	4	10
	Toplam	6	5	24	34	11	80

Tablo 7: GA Göstergeleri

	Uygunluk fonksiyonu değerlendirme sayısı	**Nesil Sayısı**	**CPU Zamanı (Saniye)**	**İlk Değişken Sayısı**	**Seçilen Kelime Sayısı**	**İndirgeme Faktörü (%)**
DVM	70501	140	36265	10201	5185	50.83
KNN	76501	154	84548	10201	5131	50.30
KA	75001	149	19748	10201	5108	50.07

Önerilen melez yöntemlerin hata matrisleri Tablo 8-10'da verilmektedir. GA'nın parametreleri deneme yanılma yöntemi ile belirlenmiştir: çaprazlama olasılığı 0.85, elit birey sayısı 3, nesil sayısı 200'de sabittir ve popülasyon büyüklüğü 500'dür.

Tablo 8: GA-DVM Sonuçları

		Gerçek Değerler					
		ÇEKO	**IKT**	**İŞL**	**SBKY**	**UAİ**	**Toplam**
Modelin Tahmin Ettiği Değerler	**ÇEKO**	4	3	0	1	0	8
	IKT	0	8	0	2	0	10
	İŞL	0	0	18	2	0	20
	SBKY	0	0	0	32	0	32
	UAİ	0	0	0	1	9	10
	Toplam	4	11	18	38	9	80

Tablo 9: GA-KNN Sonuçları

		Gerçek Değerler					
		ÇEKO	**IKT**	**İŞL**	**SBKY**	**UAİ**	**Toplam**
Modelin Tahmin Ettiği Değerler	**ÇEKO**	5	2	0	1	0	8
	IKT	0	10	0	0	0	10
	İŞL	0	0	19	1	0	20
	SBKY	0	0	0	32	0	32
	UAİ	0	0	0	0	10	10
	Toplam	5	12	19	34	10	80

Tablo 10: GA-KA Sonuçları

		Gerçek Değerler					
		ÇEKO	**IKT**	**İŞL**	**SBKY**	**UAİ**	**Toplam**
Modelin Tahmin Ettiği Değerler	**ÇEKO**	2	1	1	4	0	8
	IKT	0	8	0	1	1	10
	İŞL	1	0	16	3	0	20
	SBKY	0	0	1	31	0	32
	UAİ	0	0	0	2	8	10
	Toplam	3	9	18	41	9	80

Tablo 11: Önerilen Yöntemlerin Performans Göstergeleri

	DVM	**KNN**	**KA**	**GA-DVM**	**GA-KNN**	**GA-KA**
Doğruluk	0.65	0.738	0.55	0.875	0.95	0.825
Kesinlik	0.5	0.5	0.333	1	1	0.571
Duyarlılık	0.125	0.375	0.25	0.5	0.75	0.5
Özgünlük	0.986	0.958	0.944	1	1	0.958
F Skoru	0.2	0.429	0.286	0.667	0.857	0.533
AUC	0.774	0.819	0.653	0.886	0.947	0.843

Performans göstergeleri Tablo 11'de gösterilmektedir. En iyi performans gösteren model %95 doğruluk oranına sahip GA-KNN modelidir. Eğri Altında Kalan Alan (AUC) tüm olası karşılaştırmaların ortalamasıdır. AUC değeri 0 ile 1 arasında değişmektedir. Mükemmel bir model, 1'e yakın AUC'ye sahip olmalıdır.

Doğruluk (Accuracy), tüm doğru tahminlerin sayısının veri kümesinin toplam sayısına bölünmesiyle hesaplanır. En iyi doğruluk 1, en kötüsü 0'dır. Kesinlik (Precision), doğru pozitif tahmin sayısının toplam pozitif tahmin sayısına bölünmesiyle hesaplanır. En iyi kesinlik 1, en kötü kesinlik 0'dır. Duyarlılık (Recall), doğru pozitif tahmin sayısının toplam pozitif sayısına bölünmesiyle hesaplanır. En iyi duyarlılık 1, en kötü duyarlılık 0'dır. Özgünlük (Specificity), doğru negatif tahmin sayısının toplam negatif sayısına bölünmesiyle hesaplanır. En iyi özgünlük 1, en kötüsü 0'dır. F Skoru, hem kesinlik hem de duyarlılığı dikkate almaktadır.

3.2. Hangi Kategoriye Ait Olduğu Bilinmeyen Makalelerin Model Tarafından Otomatik Olarak Sınıflandırılması

Model geliştirildikten ve performansı ölçüldükten sonra, 427 tane özeti kullanmak suretiyle model bir kere daha çalıştırılmıştır. Çalıştırılan bu model, metinde de ifade edildiği gibi 93 tane sınıfı henüz belirlenmemiş makalelerin otomatik olarak sınıflandırılmasında kullanılmıştır.

Tablo 12: Model tarafından sınıfı doğru bir şekilde belirlenen makale özetlerinden örnekler

Yazar	*Makalenin Başlığı*	*Yıl*	*Cilt*	*Sayı*	*Model tarafından atanan sınıf*
'Onur KOYUNCU'	*Tek özelliği test edilen ürünlerin test sürecinde ekipman seçimi için kısıt programlama temelli optimizasyon*	*2014*	*69*	*3*	*İşletme*
'Akın USUPBEYLİ'	*Döviz kurunun doğrusal olmayan yapısı ve bir modelleme önerisi*	*2012*	*67*	*4*	*Ekonomi*
'Özkan AGTAŞ'	*Ceza yasasının gölgesinde siyaset*	*2012*	*67*	*4*	*Siyaset Bilimi ve Kamu Yönetimi*
'Umut BEKCAN'	*Devrimden sonra: Bolşeviklerin zorunlu dış politikası 1917-1925*	*2013*	*68*	*4*	*Uluslararası İlişkiler*

93 tane özetin gerçekte hangi kategoriye ait olduğu bilinmemektedir. Bu nedenle bu set üzerinde performans ölçümünü gerçekleştirmek mümkün olmayacaktır. Bununla birlikte bazı makalelerin hangi kategoride olduğu, sosyal bilimler konusunda uzman olmayan kişiler tarafından bile kolayca belirlenebilir. Tablo 12'de modelin otomatik olarak doğru bir şekilde sınıflandırdığı makalelerden örnekler yer almaktadır. Bu örnekler, hangi kategoriye ait oldukları kolayca belirlenebilen makalelerden seçilmiştir. Bununla birlikte bazı

makalelerin hangi kategoriye ait olduğunun belirlenebilmesi için mutlaka uzman bilgisine ihtiyaç duyulacaktır. Bu çalışmada geliştirilen model, uzmanlar için karar destek sistemi olarak değerlendirilebilir.

3.3. Kıyaslama (Benchmarking)

Önerilen GA tabanlı yöntemlerin başarısını kanıtlamak için çeşitli veri kümelerine de uygulanmıştır. Kıyaslamada kullanılan veri kümeleri http://www.kemik.yildiz.edu.tr/?id=28 adresinden temin edilmiştir. Bu web sitesinde çeşitli Türkçe veri setleri bulunmaktadır. Çalışmamızda haber, cinsiyet, film yorumları ve ruh hali olmak üzere dört farklı veri kümesi kullanılarak önerilen yöntemler karşılaştırılmıştır (Tablo 14-17).

Tablo 13: Hata matrisi

	Gerçek Pozitif	**Gerçek Negatif**
Tahmin Edilen Pozitif	Doğru Pozitif (tp)	Yanlış Pozitif (fp)
Tahmin Edilen Negatif	Yanlış Negatif (fn)	Doğru Negatif (tn)

Hata matrisi Tablo 13'te verilmiştir. Performans göstergeleri de kısaca aşağıdaki gibi tanımlanabilir.

$$Doğruluk = \frac{tp + tn}{tp + fp + fn + tn} \tag{3}$$

$$Kesinlik = \frac{tp}{tp + fp} \tag{4}$$

$$Duyarlılık = \frac{tp}{tp + fn} \tag{5}$$

$$Özgünlük = \frac{tn}{tn + fp} \tag{6}$$

$$F\ Score = 2 * \frac{Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık} \tag{7}$$

Kategori sayısının iki olduğu durumlarda yukarıdaki formüller kullanılmaktadır. Kategori sayısının üç veya daha fazla olduğu durumlarda ise MATLAB yazılımında yer alan classperf fonksiyonu (https://www.mathworks.com/help/bioinfo/ref/classperf.html) yardımıyla performans ölçümü gerçekleştirilmiştir. AUC değeri tüm olası karşılaştırmaların ortalamasıdır. AUC değeri 0 ile 1 arasında değişkenlik göstermektedir. 1'e yakın

AUC mükemmel bir modeli ifade etmektedir. Yöntemlerin farklı veri kümelerinde farklı performans gösterdikleri belirlenmiştir. Ayrıca, GA tabanlı yöntemler genellikle tüm veri kümelerinde daha iyi performans göstermiştir.

Haber: Bu veri kümesi beş kategoride (ekonomi, dergi, sağlık, politika, spor) haberleri içerir. Her kategoride 230 farklı haber bulunmaktadır.

Cinsiyet: Bu veri kümesinde erkek ve kadın yazarlar tarafından gazetelerde yayımlanan makaleler yer almaktadır. Her cinsiyette 200 makale bulunmaktadır.

Film yorumları: Bu veri kümesinde film değerlendirmeleri vardır. Pozitif, negatif ve nötr değerlendirmeler olmak üzere üç kategoriye ayrılmıştır. Her kategoride 35 yorum bulunmaktadır.

Ruh Hali: Bu veri kümesinde günlük yazılar bulunmaktadır. Dört kategori mevcuttur (karışık, mutlu, gergin ve üzgün). Her kategoride 40 yazı mevcuttur.

1150 haber veri seti birçok sınıflandırma çalışmasında kullanılmıştır. Bu çalışmalar aşağıdaki gibi özetlenebilir. Dönmez ve Adalı (2016) çalışmalarında bir cümledeki ifadeleri bulmak, ifadelerin kavramlara ilişkisini tanımlamak için bir model önermektedirler. Güran vd. (2013), orijinal vektör uzayını ve boyutu azaltmak için negatif matris çarpanlarına ayırma yöntemini kullanmışlardır. Belge kümeleme için k-ortalamalar kümeleme algoritması kullanılmıştır. Mesafe ölçüsü olarak kosinüs benzerliği/mesafesi kullanılmıştır. Farklı özetleme oranlarına ve veri dönüşüm süreçlerine sahip iki Türkçe veri kümesi üzerinde kapsamlı analiz yapılmıştır. Önerdikleri sistemin Türkçe belgelerin sınıflandırılmasından önce ön-işleme amacıyla kullanılabileceğini ifade etmektedirler.

Tablo 14: Haber Veri Kümesinin Performansı

	DVM	KNN	KA	GA-DVM	GA-KNN	GA-KA
Doğruluk	0.9	0.95	0.76	1	1	0.97
Kesinlik	1	0.889	0.923	1	1	0.95
Duyarlılık	0.75	1	0.6	1	1	0.95
Özgünlük	1	0.917	0.988	1	1	0.988
F Skoru	0.857	0.941	0.727	1	1	0.95
AUC	0.958	0.958	0.85	1	1	0.983

Tablo 15: Cinsiyet Veri Kümesinin Performansı

	DVM	KNN	KA	GA-DVM	GA-KNN	GA-KA
Doğruluk	0.96	0.95	0.93	0.99	0.99	0.99
Kesinlik	0.95	0.889	0.867	0.976	0.976	0.976
Duyarlılık	0.95	1	0.975	1	1	1
Özgünlük	0.967	0.917	0.9	0.983	0.983	0.983
F Skoru	0.95	0.941	0.918	0.988	0.988	0.988
AUC	0.958	0.958	0.938	0.992	0.992	0.992

Tablo 16: Film Yorumları Veri Kümesinin Performansı

	DVM	KNN	KA	GA-DVM	GA-KNN	GA-KA
Doğruluk	0.433	0.433	0.433	1	1	0.767
Kesinlik	0.2	0.4	0.333	1	1	0.667
Duyarlılık	0.1	0.4	0.4	1	1	1
Özgünlük	0.8	0.7	0.6	1	1	0.75
F Skoru	0.133	0.4	0.364	1	1	0.8
AUC	0.697	0.68	0.617	1	1	0.835

Tablo 17: Ruh Hali Veri Kümesinin Performansı

	DVM	KNN	KA	GA-DVM	GA-KNN	GA-KA
Doğruluk	0.5	0.5	0.45	0.925	0.95	0.775
Kesinlik	0.5	0.474	0.333	0.818	0.833	1
Duyarlılık	0.2	0.9	0.4	0.9	1	0.8
Özgünlük	0.933	0.667	0.733	0.933	0.933	1
F Skoru	0.286	0.621	0.364	0.857	0.909	0.889
AUC	0.654	0.684	0.663	0.875	0.934	0.915

Poyraz vd. (2012) çalışmalarında metin sınıflandırmasını geliştirmek için Türkçe Vikipedi kullanmışlardır. Çalışmada, sözcükler arasındaki anlamsal ilişki göz önüne alınmıştır. Ayrıca, çok terimli Naïve Bayes ve doğrusal çekirdekli DVM kullanılmıştır. Önerilen modellerin performansı, Türk gazetelerinden alınan dört veri kümesi kullanılarak test edilmiştir. Kilimci vd. (2017) heterojen topluluk algoritmaları kullanmışlardır. Yöntemleri değerlendirmek için dört farklı Türkçe veri kümesi kullanılmıştır. Çalışmada heterojen bir topluluk sistemi oluşturmak için Naïve Bayes modelinin iki versiyonu, DVM ve Rastgele Orman algoritması kullanılmıştır. Çalışmanın sonuçları Rastgele Orman algoritmasının

performansının diğer sınıflandırma tekniklerinin performansından daha iyi olduğunu göstermiştir.

Yukarıda adı geçen çalışmalarda önerilen sistemlerin performansları diğer veri setlerinin yanı sıra 1150 haber veri seti üzerinde de ölçülmüştür. Söz konusu çalışmalarda raporlanan en yüksek sınıflandırma performansları ve bu performansın erişildiği teknik Tablo 16'da özetlenmiştir. Bu çalışmada 1150 haber veri setinde %100 oranında sınıflandırma performansı elde edilmiştir. Söz konusu performans GA ile kelime sayısının azaltıldığı KNN modelinde ortaya çıkmıştır.

Tablo 16: Diğer Çalışmalarla Karşılaştırılması

	Kullanılan Teknik	**Kesinlik**
Kilimci vd. (2017)	Heterojen topluluk algoritmaları	%97.16
Güran vd. (2013)	k-ortalamalar algoritması	%86.00
Dönmez ve Adalı (2016)	Naïve Bayes	%97.12
Poyraz vd. (2012)	Naïve Bayes	%93.13
Önerilen çalışma	GA-KNN	%100.00

Sonuçlar

Sınıflandırma problemlerinde verilerin yapısını incelemek ve anlamak çok önemlidir. Büyük boyuttaki belgelerin analizinde daha iyi performans elde edebilmek için, genellikle boyutu önemli ölçüde azaltmak gerekmektedir. Bu noktada, veri kümesinin karmaşıklığını azaltmak için GA kullanılabilir. Çalışmamızda sosyal bilimlerde yayımlanan makalelerin Türkçe özetleri dikkate alınarak GA tabanlı yöntemler (GA-DVM, GA-KNN, GA-KA), DVM, KNN ve KA kullanılarak sınıflandırılma yapılmıştır. Dergide bulunan makalelerin özetlerinden oluşan veri kümesi, dengesiz bir veri kümesidir. Kategoriler arasındaki makale sayısı değişiklik göstermektedir. Bununla birlikte, GA ile değişken seçimi, sınıflandırma performansını önemli ölçüde iyileştirmiştir. Çalışmanın sonuçları, makalelerin sınıflandırılmasında GA tabanlı yöntemlerin performans göstergelerinin 0.5 ile 1 arasında değiştiğini göstermektedir. Kıyaslamada kullanılan veri kümelerinde ise performans göstergeleri 0.667 ile 1 arasında değişmektedir. Sonuçlar, kullanılan veri kümelerinde, GA tabanlı yöntemlerin sınıflandırma performansının oldukça iyi olduğunu göstermektedir. Sonraki çalışmalarda sınıflandırma sürecinde tam metin veya makalelerin giriş bölümleri kullanılabilir. Ayrıca, melez yapay sinir ağları kullanılarak sınıflandırma performansı artırılabilir.

Kaynakça

Aggrawal, Charu (2015), *Data Mining the Textbook* (New York: Springer).

Çetin, Mahmut ve Fatih Amasyalı (2013), "Active Learning for Turkish Sentiment Analysis", *IEEE INISTA* (New York: IEEE): 1-4.

Ben-Hur, Asa ve Jason Weston (2010), "A User's Guide to Support Vector Machines", Carugo, Oliviero ve Frank Eisenhaber (Der.) *Data Mining Techniques for the Life Sciences, Methods in Molecular Biology* (New York: Springer): 223−239.

Brındha, Senthil Kumar, K. Arun Prabha ve Sandeep Sukumaran (2016), "A Survey on Classification Techniques for Text Mining", *2016 3rd International Conference on Advanced Computing and Communication Systems* (New York: IEEE): 1-5.

Deniz, Ayça ve Hakan Ezgi Kiziloz (2017), "Effects of Various Preprocessing Techniques to Turkish Text Categorization using n-gram Features", *2017 International Conference on Computer Science and Engineering* (New York: IEEE): 655-660.

Doddi, Kiran Shriniwas, Y. V. Haribhakta ve Parag Kulkarni (2014), "Sentiment Classification of News Article", *International Journal of Computer Science and Information Technologies,* 5 (3): 4621-4623.

Dönmez, İlknur ve Eşref Adalı (2018), "Turkish Document Classification with Coarse-Grained Semantic Matrix", *International Conference on Intelligent Text Processing and Computational Linguistics* (Berlin: Springer): 472-484.

Forrest, Stephanie (1993), "Genetic Algorithms: Principles of Natural Selection Applied to Computation", *Science*, 261: 872-878.

Gezici, Gizem ve Berrin Yanıkoglu (2018), "Sentiment analysis in Turkish", Oflazer, Kemal ve Murat Saraçlar (Der.), *Turkish Natural Language Processing. Theory and Applications of Natural Language Processing* (Berlin: Springer): 255-271.

Gurbuz, Selen ve Galip Aydin (2018), "Büyük Veri Teknolojileri ile Bilimsel Makalelerin Sınıflandırılması", *2nd International Conference on Computer Science and Engineering* (New York: IEEE): 697-701.

Güran, Aysun, Murat Can Ganiz, Hamit Selahattin Naiboğlu ve Halil Oğuz Kaptıkaçtı (2013), "NMF Based Dimension Reduction Methods for Turkish Text Clustering", *2013 IEEE INISTA* (New Jersey: IEEE): 1-5.

Imandoust, Sadegh Bafandeh ve Mohammad Bolandraftar (2013), "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background", *International Journal of Engineering Research and Applications*, 3(5): 605-610.

Kılınç, Deniz (2016), "The Effect of Ensemble Learning Models on Turkish Text Classification", *Celal Bayar Üniversitesi Fen Bilimleri Dergisi*, 12(2): 215-220.

Kilimci, Zeynep Hilal, Selim Akyokus ve Sevinc Ilhan Omurca (2016), "The Effectiveness of Homogenous Ensemble Classifiers for Turkish and English Texts", Badica, Costin, Mirel Coşulshchi, Adina Modga Florea, Petia Koprinkova-Hristova ve Tülay Yıldırım (Der.), *2016 International Symposium on Innovations in Intelligent Systems and Applications* (Danvers: IEEE): 1-7.

Kilimci, Zeynep Hilal, Selim Akyokus ve Sevinc Ilhan Omurca (2017), "The Evaluation of Heterogeneous Classifier Ensembles for Turkish Texts. Jedrzejowicz, Piotr, Tülay Yildirim ve Ireneusz Czarnowski (Der.), *2017 IEEE International Conference on Innovations in Intelligent Systems and Applications* (Danvers: IEEE Publishing): 307-311.

Kilimci, Zeynep Hilal ve Selim Akyokus (2018), "Deep Learning-and Word Embedding-Based Heterogeneous Classifier Ensembles for Text Classification", *Complexity*, https://doi.org/10.1155/2018/7130146.

Küçük, Dilek (2015), "Automatic Compilation of Language Resources for Named Entity Recognition in Turkish by Utilizing Wikipedia Article Titles", *Computer Standards & Interfaces*, 41: 1-9.

Küçük, Dilek ve Adnan Yazıcı (2010), "A Hybrid Named Entity Recognizer for Turkish with Applications to Different Text Genres", Gelenbe, Erol, Ricardo Lent, Georgia Sakellari, Ahmet Sacan, Hakkı Toroslu, Adnan Yazıcı (Der.), *Computer and Information Sciences Lecture Notes in Electrical Engineering* (Dordrecth: Springer): 113-116.

Kowsari, Kamran, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes ve Donald Brown (2019), "Text Classification Algorithms: A Survey", *Information*, 10 (150): 1-68.

Mathworks (2020), https://www.mathworks.com/help/stats/classificationecoc.html (01.04.2020).

Nanba, Hidetsugu, Noriko Kando, ve Manabu Okumura (2000), "Classification of Research Papers using Citation Links and Citation Types: Towards Automatic Review Article Generation", Soergel, Dagobert, Padmini Srinivasan ve Barbara H. Kwasnik (Der.), *11th ASIS SIG/CR Classification Research Workshop*, 11(1): 117-134.

Poyraz, Mitat, Murat C. Ganiz, Selim Akyokuş, Burak Görener ve Zeynep Hilal Kilimci (2012), "Exploiting Turkish Wikipedia as a Semantic Resource for Text Classification", *2012 International Symposium on Innovations in Intelligent Systems and Applications*, 1-5.

Prabowo, Rudy ve Mike Thelwall (2009), "Sentiment Analysis: A Combined Approach", *Journal of Informetrics*, 3(2): 143-157.

Rokach, Lior ve Oded Maimon (2005), *Decision Trees Data Mining and Knowledge Discovery Handbook* (Boston: Springer).

Tellal, Erel, Murat Baskıcı, Bülent Duru, Bilge Sevim Çimen Aytekin, Şenay Sabah Kıyan, Bilge Kağan Şakacı, Murat Yaman ve Sevgi Eda Tuzcu (2012), "SBF Dergisi Dizin 1943–2011" https://sbfdergi.ankara.edu.tr/dosyalar/SBF-Dergisi-Dizini-28-Mayis-2012.pdf (01.04.2020).

Torunoğlu, Dilara, Erhan Çakırman, Murat Can Ganiz, Selim Akyokuş ve M. Zahid Gürbüz (2011), "Analysis of Preprocessing Methods on Classification of Turkish Texts", *2011 International Symposium on Innovations in Intelligent Systems and Applications* (New York: IEEE): 112-117.

Türkoğlu, Filiz, Banu Diri ve M. Fatih Amasyalı (2007), "Author Attribution of Turkish Texts by Feature Mining", Türkoğlu, Filiz, Banu Diri ve M. Fatih Amasyalı (Der.), *International Conference on Intelligent Computing* (Berlin: Springer): 1086-1093.

Urologın, Siddhaling (2018), "Sentiment Analysis, Visualization and Classification of Summarized News Articles: A Novel Approach*", IJACSA0 International Journal of Advanced Computer Science and Applications,* 9(8): 616- 625.

Uysal, Alper Kursat ve Serkan Gunal (2014), "The Impact of Preprocessing on Text Classification", *Information Processing & Management*, 50: 104-112.

Yıldırım, Savaş (2014), "A Knowledge-Poor Approach to Turkish Text Categorization", Gelbukh, Alexander (Der.), *International Conference on Computational Linguistics and Intelligent Text Processing* (Berlin: Springer): 428-440.