Fırat Üniversitesi Müh. Bil. Dergisi
35(1), 239-250, 2023

Araştırma Makalesi
https://doi.org/10.35234/fumbd.1195908

Filtre Tabanlı Öznitelik Seçim Yöntemleri Kullanılarak Metinlerde Duygu Sınıflandırması Üzerine Karşılaştırmalı Bir Çalışma

Ensar Arif SAĞBAŞ[1*]
[1] Bilişim Sistemleri Mühendisliği, Teknoloji Fakültesi, Muğla Sıtkı Koçman Üniversitesi, Muğla, Türkiye
[*1] arifsagbas@mu.edu.tr

(Geliş/Received: 28/10/2022; Kabul/Accepted: 15/02/2023)

Öz: Bir metin sınıflandırma problemi olarak duygu analizi, çevrimiçi metin belgelerinden öznel bilgi çıkarmanın kritik bir görevidir. Metin sınıflandırmanın önemli bir sorunu ise yüksek boyutluluktur. Boyut indirgeme, makine öğreniminde sınıflandırma performansını iyileştirmenin etkili bir yoludur. Alakasız özniteliklerin azaltılması eğitim süresini kısaltabilmekte ve sınıflandırma doğruluğunu artırabilmektedir. Farklı öznitelik seçim yöntemlerinin performansı, farklı veri kümelerinin özelliklerine bağlı olarak değişebilmektedir. Bu çalışmada filtre tabanlı 6 farklı öznitelik seçimi yönteminin (Korelasyon tabanlı öznitelik seçimi, Ki-kare, Kazanç oranı, Bilgi kazancı, OneR ve Simetrik belirsizlik katsayısı) performansı duygu sınıflandırmasında sıklıkla kullanılan 9 farklı veri kümesi üzerinde test edilmiş ve karşılaştırılmıştır. Bütün veri kümelerinde her bir öznitelik seçimi yöntemi için filtre puanları hesaplanmıştır. Elde edilen filtre puanları büyükten küçüğe sıralanmıştır. En yüksek filtre puanına sahip öznitelikten en düşük filtre puanına sahip özniteliğe doğru öznitelikler bir önceki alt kümeye eklenerek yeni alt kümeler oluşturulmuş ve sınıflandırılmıştır. Hesaplama sonuçları, önerilen yaklaşımın 9 genel duygu sınıflandırma veri kümesi için Çok terimli Naive Bayes sınıflandırıcısını kullanarak ortalama %94.34 doğruluk oranlarına ulaştığını göstermektedir. Arama uzayı dikkate alındığında, bu yaklaşımın geliştirilebilir ve mevcut yaklaşımlarla rekabet edebilir olduğu sonucuna varılabilir.

Anahtar kelimeler: Duygu sınıflandırma, öznitelik seçimi, makine öğrenmesi, Çok terimli Naive Bayes, doğal dil işleme.

A Comparative Study on Text Sentiment Classification by Using Filter-Based Feature Selection Methods

Abstract: Sentiment analysis as a text classification problem is a critical task of extracting subjective information from online text documents. An important problem of text classification is high dimensionality. Dimension reduction is an effective way to improve classification performance in machine learning. Reducing irrelevant features can reduce training time and improve classification accuracy. The performance of different feature selection methods may vary depending on the characteristics of different datasets. In this study, the performance of 6 different filter-based feature selection methods (Correlation-based feature selection, Chi-square, Gain ratio, Information gain, OneR, and Symmetric uncertainty coefficient) were tested and compared on 9 different datasets that are frequently used in sentiment classification. Filter scores were calculated for each feature selection method in all datasets. The obtained filter scores were sorted descendingly. New feature subsets were created and classified by adding features to the previous subset from the feature with the highest filter score to the feature with the lowest filter score. The computational results show that the proposed approach achieves average accuracy rates of 94.34% using the Multinomial Naive Bayes classifier for 9 general sentiment classification datasets. Considering the search space, it can be concluded that this approach can be improved and is competitive with existing approaches.

Keywords: Sentiment classification, feature selection, machine learning, Multinomial Naïve Bayes, natural language processing.

1. Giriş

Bilgi teknolojilerinin hızlı gelişimi ile kullanıcı tarafından oluşturulan içerikler kolaylıkla çevrimiçi olarak yayınlanabilmektedir [1]. Metin tabanlı sosyal medya, müşteriler ve işletmeler arasındaki önemli iletişim araçlarından biri haline gelmiştir. Sosyal medyada kullanıcılar ürün veya hizmetlerle ilgili görüş ve değerlendirmelerini rahatlıkla ifade edebilirler. Bu çevrimiçi kullanıcı deneyimleri, özellikle olumsuz değerlendirmeler, diğer tüketicilerin davranışlarını etkilemektedir. Sonuç olarak, müşterilerin duygularını etkin bir şekilde tespit etmek ve bu olumsuz yorumların girişimcilere büyük zarar vermesini önlemek kritik konulardan biri haline gelmiştir [2]. Fikir madenciliği olarak da bilinen duygu analizi, ana metin sınıflandırma yöntemlerinden biridir ve duygusal metinleri olumlu veya olumsuz etiketlere ayırma ile ilgilenmektedir. Duygu analizi genel olarak üç ayrıntı düzeyinde yapılabilir: [3] belge düzeyi, cümle düzeyi ve görünüm düzeyi [4,5].

* Sorumlu yazar: arifsagbas@mu.edu.tr Yazarların ORCID Numarası: [1] 0000-0002-7463-1150

Otomatik metin sınıflandırıcılar, spam filtreleme, duygu analizi ve haber sınıflandırması gibi birçok gerçek dünya sorununu ele almak için kullanılabilmektedir. Metinler genellikle yüksek boyutlu ve seyrek bir belge terim matrisi ile kelime sıklık sayımlarını içeren kelime dağarcığının boyutuna sahip bir uzayda temsil edilir. Yüksek boyutluluk, boyutluluğun laneti (curse of dimensionality) ve modelin aşırı öğrenmesi gibi bazı sorunlara neden olabilir. Öznitelik seçimi, boyutluluğu azaltmak, alakasız verileri kaldırmak ve öğrenme doğruluğunu artırmak için kullanılmaktadır. Öznitelik seçimi, belirli bir metnin sınıflandırılmasına en çok katkıda bulunan öznitelikleri otomatik veya manuel olarak seçme işlemidir. Metin sınıflandırma problemlerinde, öznitelikler genellikle kelimelerin bir alt kümesinin bir temsilidir. Metin külliyatından çıkarılan özniteliklerin önemli bir alt kümesi, metin sınıflandırma göreviyle ilgili olmayabilir. Bu alakasız öznitelikler, sınıflandırma modellerinin etkinliğini ve doğruluğunu bozabilmektedir [6]. Bu nedenle, metin sınıflandırması için öznitelik seçimi, yapay zekâ ve veri madenciliği çalışmalarında popüler bir araştırma konusu haline gelmiştir [7].

Öznitelik seçme yöntemleri kullanılan amaç fonksiyonlarına göre filtreler ve sarmalayıcılar olmak üzere iki kategoriye ayrılabilir. Filtre tabanlı yöntemler, belirli bir matematiksel ölçüte göre öznitelik alt kümelerini değerlendirirken, sarmalayıcı tabanlı yöntemler, değerlendirme için tahmin performansını (örneğin, doğruluk) kullanır. Uygun öğrenme modelini kullanarak sarmalayıcı tabanlı yöntemler, filtre tabanlı yöntemlerden daha etkili sonuçlar üretebilmektedir. Bu nedenle sınıflandırma için sarmalayıcı teknikleri yaygın olarak tercih edilmektedir. Sağlayabileceği yüksek tahmin performansına rağmen, öznitelik seçiminin ana dezavantajı, öznitelik sayısı arttıkça öznitelik alt kümesi arama uzayının katlanarak büyümesidir. Ayrıca, sarmalayıcı tabanlı yöntemler bir değerlendirme ölçütü olarak sınıflandırma modellerini kullandığından, filtre tabanlı yöntemlerden daha fazla hesaplama süresi gerektirmektedir [3].

Literatürde duygu sınıflandırması için daha verimli yöntemler arayan çeşitli çalışmalar bulunmaktadır. Wang et al. [8] konuşma parçası analizine dayalı duygu sınıflandırması için geliştirilmiş bir Rastgele alt uzay yöntemi, POS-RS önermiştir. Onan et al. [9] sınıflandırıcılara ve sınıflandırma algoritmalarının tahmin performansına dayalı her bir çıktı sınıfına uygun ağırlık değerleri atamak için çok amaçlı, eniyilemeye dayalı ağırlıklı oylama şeması geliştirmeye çalışmışlardır. Onan et al. [10] duygu sınıflandırması için kümeleme ve rastgele aramaya dayalı melez bir topluluk budama şeması önermişlerdir. Jalilvand and Salim [11] sınıflandırmada boyut indirgeme için öznitelik birleştirme adı verilen yeni bir yaklaşım önermiştir. Yang et al. [12] çoklu sınıflandırıcı sistemlerin bir topluluk yöntemi olarak çoğunluk oylamasını kullanmışlardır. Gokalp et al. [3] duygu sınıflandırması için yinelenen açgözlü metasezgisel tabanlı yeni bir sarmalayıcı öznitelik seçim algoritması önermiştir. Ayrıca, önerilen algoritmanın açgözlü yapı kısmı için önceden hesaplanmış filtre puanlarına dayanan bir seçim prosedürü geliştirmişlerdir. Onan [13] metin üzerinden duygu sınıflandırması için iki ayrı çift yönlü long short-term memory (LSTM) ve gated recurrent unit (GRU) katmanı kullanan çift yönlü bir evrişimsel tekrarlayan sinir ağı mimarisi önermiştir. Shao and Chen [14] finansal içerikli metinlerde duygu sınıflandırması gerçekleştirmek için derin öğrenme tabanlı bir yaklaşım önermiştir. Khan et al. [15] duygu sınıflandırmasının performansını iyileştirmek için geleneksel öznitelik oluşturma yöntemlerini deep neural network (DNN) tabanlı yöntemlerle entegre etmenin etkili bir yolunu araştırmıştır. Yang et al. [16] dinamik veri belirsizliğini sürekli olarak ele almak için dinamik metin duygu sınıflandırması için zamansal-uzaysal üç yollu çok parçalı öğrenme çerçevesi yürütmüştür. Ayetiran [17] belge ve boyut düzeyinde duygu verilerini ortaklaşa öğrenen yeni bir derin öğrenme tekniği sunmuştur. Karga vd. [18] COVID-19 salgınının yükseköğrenim üzerindeki etkisini analiz etmek için derin öğrenmeye dayalı bir duygu analizi yaklaşımı sunmuşlardır. Polat ve Ağca [19] kullanıcılarının Türkçe ve İngilizce yorumlarındaki duygusal eğilimlerin ortaya çıkarılması ve sınıflandırılmasında kullanılan duygu analizi yöntemlerini karşılaştırmıştır. Şahinaslan vd. [20] YouTube yorumları üzerinden Naive Bayes sınıflandırıcısı kullanarak çok dilli duygu analizi gerçekleştirmiştir. Dinçer vd. [21] Twitter verileri üzerinden siber zorbalığın tespiti üzerine çalışmışlardır. Salur ve Aydın [22] metinler üzerinden duygu sınıflandırmada derin öğrenme yöntemleriyle çıkartılan derin öznitelikler ile veri önişleme aşamasında silinen verilerden elle çıkartılan öznitelikleri birlikte kullanımına dayanan yeni bir model önermişlerdir.

Bu çalışmada, duygu analizi için filtre tabanlı öznitelik seçimi yöntemlerinin başarımlarının ölçülmesi amaçlanmaktadır. Böylece metin sınıflandırma için hem daha yüksek başarının elde edilmesi hem de boyutun azaltılması hedeflenmektedir. Bu amaçla, duygu sınıflandırması 6 farklı öznitelik seçimi algoritmasının (Korelasyon tabanlı öznitelik seçimi, Bilgi kazancı, Kazanç oranı, Ki-kare, OneR ve Simetrik belirsizlik) performansları değerlendirilmiştir. Sınıflandırma yöntemi olarak ise duygu sınıflandırmasındaki yüksek performansı nedeniyle, Çok terimli Naive Bayes (ÇTNB), kullanılmıştır. Öznitelik seçimi algoritmalarının performansını değerlendirmek için Whitehead and Yaeger'den [23] yaygın olarak kullanılan duygu sınıflandırma veri kümeleri üzerinde kapsamlı bir deneysel çalışma yapılmıştır. Etkili öznitelik alt kümeleri oluşturmak amacı ile öznitelik seçimi algoritmalarından elde edilen filtre puanlarını kullanan bir yaklaşım önerilmiştir. Önerilen en iyi N elemanlı alt küme yaklaşımı ile sarmalayıcı yöntemlere göre daha az yineleme ile başarılı sonuçlar

alınabileceği görülmüştür. Kapsamlı deneysel sonuçlar, önerilen filtre tabanlı öznitelik seçimi yaklaşımının, kullanılan 9 ortak veri kümesine dayalı duygu sınıflandırması için başarılı sonuçlar sergilediğini göstermektedir. Ek olarak, elde edilen sonuçlar çeşitli duygu analizi algoritmalarının en gelişmiş sonuçlarıyla karşılaştırılmıştır. Ortalama doğruluk oranı incelendiğinde, önerilen yaklaşımın ele alınan 6 son teknoloji yöntemin 4'ünden daha yüksek sınıflandırma başarısı yakaladığı görülmüştür.

Bu makalenin geri kalan bölümleri aşağıdaki gibi özetlenmiştir. Bölüm 2, öznitelik seçme yöntemlerini ve kullanılan sınıflandırma yöntemlerini özetlemektedir. Sonrasında Bölüm 3, ele alınan öznitelik seçimi algoritmalarının değerlendirilmesi için deneysel bir çerçeve sunmakta ve bunu öznitelik seçimi yöntemleri arasında ve son teknoloji çalışmalarla karşılaştırmaktadır. Son olarak, Bölüm 4 makaleyi sonuçlandırmakta ve gelecekteki olası çalışmaları tartışmaktadır.

2. Materyal ve Yöntem

2.1. Filtre tabanlı öznitelik seçimi

Filtre tabanlı öznitelik seçimi, tahmin modelleri kullanmak yerine öznitelik alt kümelerini bilgi içeriğine göre değerlendirir. Filtre tabanlı ölçümlerin kullanımı kolaydır, hızlıdır ve farklı sınıflandırıcılar için genelleştirilebilirler [3].

Bu alt bölüm, çalışmada kullanılan ve performansları karşılaştırılan ana filtre tabanlı öznitelik seçim yöntemlerini kısaca açıklamaktadır. Bu yöntemler; Ki-kare, Korelasyon tabanlı öznitelik seçimi, Kazanç oranı, Bilgi kazancı, OneR ve Simetrik belirsizlik katsayısıdır.

2.1.1. Korelasyon tabanlı öznitelik seçimi

Korelasyon tabanlı öznitelik seçimi, bir sınıflandırma işlemini başarıyla gerçekleştirebilen bir öznitelik alt kümesi oluşturan bir öznitelik filtreleme yöntemidir [24]. Korelasyon ölçümü, öznitelik ve sınıf arasındaki Pearson korelasyon katsayısını ölçerek bir özniteliğin değerini değerlendirir. Bu ölçü, aralarındaki ilişkinin gücünü temsil eder. Korelasyon katsayısı, +1 ile -1 arasında bir değere sahiptir; burada +1, pozitif doğrusal korelasyonu, 0 doğrusal korelasyonun olmadığını ve -1, negatif doğrusal korelasyonu gösterir ve Denklem (1)'de olduğu gibi tanımlanmaktadır [25].

$$R(i) = \frac{\sum_{k=1}^{m}(x_{k,i}-\bar{x}_i)(y_k-\bar{y})}{\sqrt{\sum_{k=1}^{m}(x_{k,i}-\bar{x}_i)^2 \cdot \sum_{k=1}^{m}(y_k-\bar{y})^2}} \quad (1)$$

Burada, x özniteliği, y sınıfı, m ise veri noktası sayısını göstermektedir.

2.1.2 Ki-kare

Ki-kare ölçüsü, sınıfa göre ki-kare istatistiğinin değerini hesaplayarak bir özniteliğin değerini değerlendirir. t ve c arasındaki bağımsızlık eksikliğini ölçmek için kullanılır (burada t terim ve c sınıftır) ve bir serbestlik dereceli $\chi 2$ dağılımı ile karşılaştırılır. Metin sınıflandırması için $\chi 2$ ölçümü Denklem (2)'de verildiği gibi tanımlanmıştır [26].

$$X_{t,c}^2 = \frac{D*(PN-MQ)^2}{(P+M)*(Q+N)*(P+Q)*(M+N)} \quad (2)$$

Burada D, toplam belge sayısıdır. P, t terimini içeren c sınıfı belgelerin sayısıdır. Q, c olmadan meydana gelen t içeren belge sayısıdır. M, t olmadan meydana gelen c sınıfı belge sayısıdır. N, t içermeyen diğer sınıfların belge sayısıdır [3].

2.1.3. Bilgi kazancı

Bilgi kazancı, sınıfa göre kazancı ölçerek bir özniteliğin değerini değerlendirir. Entropi bilgisine bağlıdır. Entropi, sistemdeki kaos veya rastgelelik derecesinin bir ölçüsüdür. Bilgi kazancı, belirsizliğin ortadan kaldırılmasından sonraki bilgi miktarını temsil eder [27]. Bilgi kazancı Denklem (3)'te tanımlanmaktadır.

$$\mathrm{BK}(\mathrm{X},\mathrm{Y}) = \mathrm{H}(\mathrm{X}) - \mathrm{H}(\mathrm{X}|\mathrm{Y}) \quad (3)$$

Burada X öznitelik, Y ise sınıfı göstermektedir.

2.1.4. Kazanç oranı

Bilgi kazanım oranı olarak da bilinen kazanç oranı, kararlı değerlendirmeye yönelik önyargıyı (bias) azaltır [28]. Öznitelik entropisi (içsel bilgi) tarafından kazanılan bilgilerin bölünmesiyle hesaplanır. Kazanç oranı Denklem (4)'teki formül ile hesaplanmaktadır.

$$\mathrm{KO} = \frac{\mathrm{BK}(\mathrm{X})}{\mathrm{İçselBilgi}(\mathrm{X})} \quad (4)$$

Burada X özniteliği göstermektedir.

2.1.5. OneR

Bu yöntem, 1R sınıflandırıcısını kullanarak her bir özniteliği ayrı ayrı değerlendirmektedir. Bu sınıflandırıcının kuralı sadece öznitelik değerlerine ve konuya bağlıdır. Böylece, her öznitelik ve özniteliğin her bir değeri için, veri kümesini sınıflandırmak için yalnızca o öznitelik kullanılacaksa üretilen hata hesaplanır. Daha sonra hata sayısı en az olan öznitelik seçilir. Bir öznitelik seçme yöntemi olarak, algoritma elde edilen hata oranına göre (her öznitelik tarafından bağımsız olarak) öznitelikleri azalan şekilde sıralar ve istenen ilk öznitelik sayısını korur [29].

2.1.6. Simetrik belirsizlik

Simetrik belirsizlik katsayısı, çok değerli özniteliklere yönelik önyargıyı azaltan bilgi kazancının bir modifikasyonudur. Sınıfa göre simetrik belirsizliği ölçerek bir özniteliğin değerini değerlendirir. Simetrik belirsizlik Denklem (5)'te gösterilmiştir [28].

$$\mathrm{SB}(\mathrm{X},\mathrm{Y}) = 2\frac{\mathrm{BK}(\mathrm{X},\mathrm{Y})}{\mathrm{H}(\mathrm{X})+\mathrm{H}(\mathrm{Y})} \in [0,1] \quad (5)$$

Burada X özniteliği, Y ise sınıfı göstermektedir.

2.2. Sınıflandırma

Sınıflandırma aşaması, etiketli verilere göre kalıba uygun bir kategoriye atar. Bu çalışmada Destek vektör makinesi (DVM), Çok terimli Naive Bayes (ÇTNB) ve Lojistik regresyon (LR) sınıflandırma modelleri kullanılmıştır. Bu öğrenme modelleri aşağıdaki alt bölümlerde kısaca açıklanmaktadır.

2.2.1. Çok terimli Naive Bayes

Çok terimli Naive Bayes modeli, çok değişkenli Bernoulli olay modeli yerine belge uzunluklarının belgelerdeki sınıftan bağımsız olduğunu varsayarak sözcük sıklığı bilgisini kullanan üretken bir modeldir (örneğin, kelimelerin uzayı üzerinde ikili vektör) [30]. Sınıflandırıcı, kelime sıklığı bilgisinden yararlandığı için metin sınıflandırma görevleri için çok uygundur.

2.2.2. Destek vektör makinesi

Destek vektör makinesi, hem doğrusal hem de doğrusal olmayan verileri sınıflandırmak için, sınıflandırma ve regresyon analizi için denetimli bir öğrenme algoritması oluşturur [31]. Destek vektör makinesi ile, orijinal veri kümesinin daha yüksek bir boyuta, yani verileri sınıflara bölmek için karar sınırı görevi gören bir hiper düzleme dönüştürülmesi için doğrusal olmayan bir eşleştirme yöntemi kullanılır [32]. Destek vektör makinesini kullanarak, amaç tipik olarak verileri farklı sınıflara bölmek için en uygun karar sınırını belirlemektir. Metin madenciliği,

yüksek boyutlu öznitelik uzayı, birkaç alakasız öznitelik ve doğrusal olarak ayrılabilir kategorizasyon sergilediği için Destek vektör makinesi metin sınıflandırması için uygun bir algoritmadır [33].

2.2.3. Lojistik regresyon

Lojistik regresyon, olayların meydana gelme olasılığını bir dizi öngörücü değişkenin doğrusal bir fonksiyonu olarak modelleyen ve bağımlı değişkenlerin değerini tahmin etmek için kullanılan bir doğrusal regresyon genellemesidir. Bağımlı değişkenin nokta tahminini tahmin etmek yerine, oluşma ihtimalini tahmin etmek için bir tahmin modeli oluşturur. Lojistik regresyon, sınıflandırma için basit ve etkili bir araç olarak da kullanılabilir [9,34].

3. Deneysel Çalışma

3.1. Deneysel kurulum

Öznitelik seçimi algoritmalarının performans analizi Java programla dili ve WEKA [35] kütüphanesi kullanılarak Intel® CoreTM i5 3210M 2.50 GHz CPU konfigürasyonlu bir bilgisayarda tek çekirdekli olarak gerçekleştirilmiştir. Bazı ön testlerden sonra, duygu sınıflandırma veri kümelerini performans sonuçları elde etmek için Destek vektör makinesi, Çok terimli Naive Bayes ve Lojistik regresyon olmak üzere üç sınıflandırıcı kullanılmıştır. Modeller için parametre değerleri WEKA'da varsayılan değerlere ayarlanmıştır. Duygu sınıflandırması veri kümelerini işlemek için kelime torbası çerçevesini (bag-of-words), tek terimli öznitelikler (unigram features) ve terim frekansı-ters belge frekansı (term frequency-inverse document frequency - tf-idf) ölçümü benimsenmiştir. Bu amaçla, dizgi özniteliklerini dizgilerdeki metinden sözcük oluşum bilgisini gösteren bir dizi sayısal özniteliğe dönüştürmek için StringToWordVector uygulanmıştır. Ayrıca güvenilir sonuçlar elde etmek için 10 katmanlı çapraz doğrulama uygulanmıştır. Bu işlemde veri seti rastgele 10 parçaya bölünür. Her parça bir test seti olarak kullanılırken, kalan diğer parçalar eğitim seti olarak kullanılır. İşlem 10 kez tekrarlanır ve her seferinde test için farklı katmanlar ele alınır. Daha sonra modelin ortalaması alınır ve sonlandırılır.

3.2. Duygu sınıflandırma veri kümeleri

Bu çalışmada, kullanılan öznitelik seçimi algoritmanın performanslarını değerlendirmek için Whitehead and Yaeger'den [23] 9 adet duygu analizi veri seti kullanılmıştır. 9 açık veri kümesinin adı camera, camp, doctor, drug, laptop, lawyer, music, radio ve tv'dir. Bu veri kümeleri yaklaşık %50 olumlu ve %50 olumsuz incelemeden oluşmaktadır. Veri kümelerinin özellikleri Tablo 1'de listelenmiştir. Fikir madenciliği/duygu madenciliği veri kümelerinin kısa açıklamaları Tablo 2'de gösterilmiştir.

Tablo 1. Kullanılan duygu sınıflandırması veri kümelerinin özellikleri

Veri kümesi	Öznitelik sayısı	Olumlu gözlem	Olumsuz gözlem	Gözlem sayısı
camera	1457	250	248	498
camp	1810	402	402	804
doctor	1811	739	739	1478
drug	1312	401	401	802
laptop	1840	88	88	176
lawyer	2123	110	110	220
music	1441	291	291	582
radio	1758	502	502	1004
tv	2423	235	235	470

Tablo 2. Kullanılan duygu sınıflandırması veri kümelerinin açıklamaları

Veri kümesi	Açıklama
camera	Amazon.com'dan dijital kamera incelemeleri
camp	CampRatingz.com'dan yaz kampı incelemeleri
doctor	RateMDs.com'dan doktor yorumları
drug	DrugRatingz.com'dan farmasötik ilaç incelemeleri
laptop	Amazon.com'dan dizüstü bilgisayar incelemeleri
lawyer	LawyerRatingz.com'dan avukatların yorumları
music	Amazon.com'dan müzik CD'si incelemeleri
radio	RadioRatingz.com'dan radyo programı incelemeleri
tv	TVRatingz.com'dan TV şovlarının değerlendirmeleri

3.3. Değerlendirme ölçümleri

Yöntemin performansını değerlendirmek için sınıflandırma doğruluk oranı, kesinlik, duyarlılık ve F-ölçütü dâhil olmak üzere 4 farklı değerlendirme ölçütü kullanılmıştır. Sınıflandırma doğruluğu, gerçek pozitiflerin (True pozitive - TP) ve gerçek negatiflerin (True negative - TN) toplamının toplam örnek sayısına bölünmesiyle hesaplanır. Doğruluk oranı hesaplanması Denklem (6)'da sunulmuştur.

$$\text{Doğruluk oranı} = \frac{\text{TN+TP}}{\text{Toplam gözlem sayısı}} \tag{6}$$

Kesinlik, pozitif tahmin değeridir. Gerçek pozitiflerin sayısının, gerçek pozitiflerin ve yanlış pozitiflerin (False positive - FP) toplamına bölünmesiyle hesaplanır. Kesinlik formülü Denklem (7)'de tanımlanmıştır.

$$\text{Kesinlik} = \frac{\text{TP}}{\text{TP+FP}} \tag{7}$$

Duyarlılık, gerçek pozitif oran veya isabet oranıdır. Gerçek pozitiflerin sayısının, gerçek pozitiflerin ve yanlış negatiflerin (False negative - FN) toplamına bölünmesiyle hesaplanır. Duyarlılık formülü Denklem (8)'de tanımlanmıştır.

$$\text{Duyarlılık} = \frac{\text{TP}}{\text{TP+FN}} \tag{8}$$

F-ölçütü, kesinlik ve duyarlılığın harmonik ortalamasıdır. Denklem (9)'da tanımlanmıştır.

$$\text{F} - \text{ölçütü} = 2 * \frac{\text{Kesinlik*Duyarlılık}}{\text{Kesinlik+Duyarlılık}} \tag{9}$$

3.4. Ön deneyler

Geçmiş çalışmalarda aynı veri kümesini kullanan çalışmalarda [8-12] Destek vektör makinesi (DVM) ve Lojistik regresyon (LR) yöntemlerinin başarılı sonuçlar sergilediği görülmüştür. Gökalp et al. [3] tarafından gerçekleştirilen çalışmada ise Bayes yaklaşımlarının performansı test edilmiş ve Çok terimli Naive Bayes (ÇTNB) öne çıkan yöntem olmuştur. Bu bilgiler ışığında, metinden duygu sınıflandırmasında kullanmak için DVM, LR ve ÇTNB olmak üzere üç sınıflandırıcının performansı analiz edip karşılaştırılmıştır. Bu nedenle, öznitelik seçimi olmadan 9 duygu sınıflandırması veri kümesi için sınıflandırıcı başına doğruluk oranları 10 katmanlı çapraz doğrulama ile hesaplanmıştır. Elde edilen doğruluk oranları Tablo 3'te sunulmuştur.

Tablo 3. Öznitelik seçimi olmadan DVM, LR ve ÇTNB sınıflandırıcılarının doğruluk oranları (%)

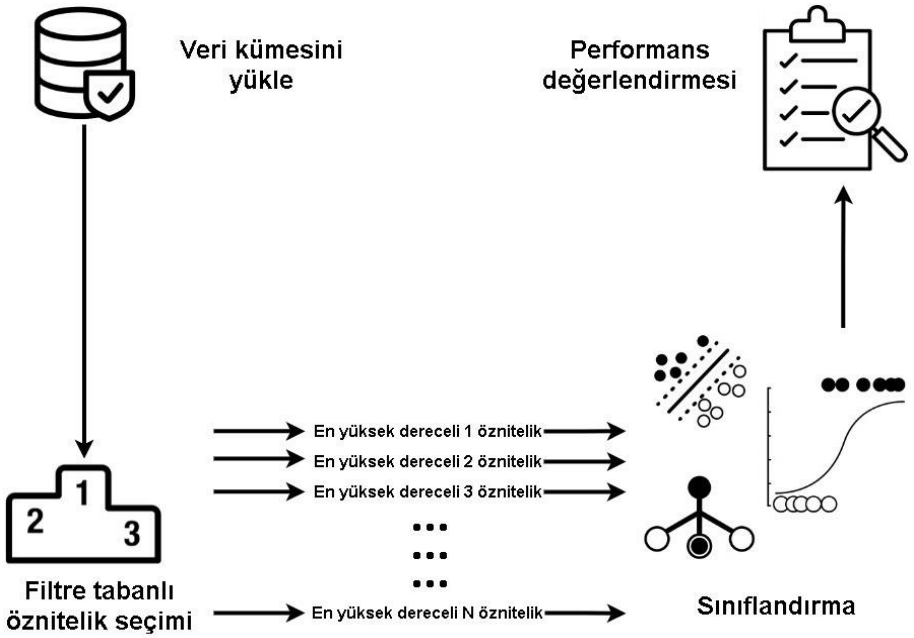
Veri kümesi	DVM	LR	ÇTNB
camera	75.30	77.11	81.33
camp	94.08	79.85	89.55
doctor	85.79	59.07	88.50
drug	70.32	56.23	73.57
laptop	80.11	85.23	88.64
lawyer	85.55	88.18	85.91
music	69.59	70.79	80.07
radio	72.01	69.02	77.79
tv	79.15	80.00	77.45
Ortalama	79.10	73.94	**82.53**

Ortalama doğruluk değerlerine göre en yüksek ortalama doğruluk değeri ÇTNB için %82,53 hesaplanırken, DVM için %79.10 ve LR için %73,94 hesaplanmıştır. Ortalama doğruluk oranları ele alındığında çalışmanın devamında ÇTNB yönteminin kullanılmasına karar verilmiştir.

3.5. Hesaplama sonuçları

Filtre tabanlı öznitelik seçimi algoritmaları öznitelikleri gerçekleştirdikleri hesaplamalar sonucunda derecelendirmektedir. Bu algoritmalardan elde edilen değerler ile en yüksek puana sahip N elemanlı alt kümeler oluşturulmuş ve sınıflandırma performansları test edilmiştir. Örnek ile açıklanacak olursa, öznitelik seçim

algoritması sonucunda öznitelik puanları büyükten küçüğe F30, F55, F61, F22, …, F198 gibi bir sıralamaya sahip olsun. Öncelikli olarak F30, devamında F30 ve F55, sonrasında F30, F55 ve F61 olacak şekilde bütün öznitelikler alt kümeye dâhil oluncaya kadar sınıflandırma performansları değerlendirilmiştir. Oluşturulan öznitelik seçimi yönteminin yapısı Şekil 1'de sunulmaktadır.

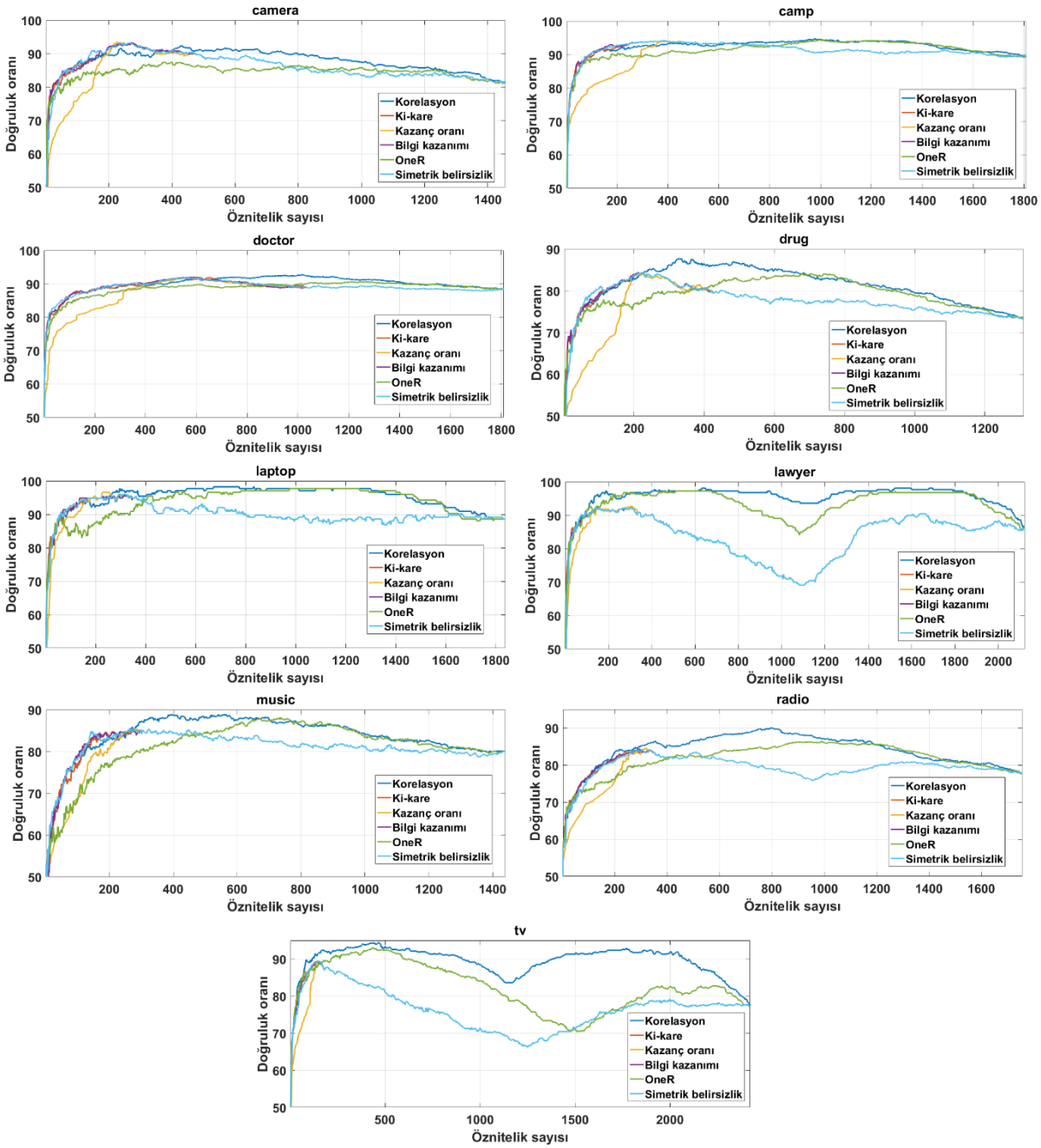


Şekil 1. En iyi N elemanlı alt küme seçimi yaklaşımı akış şeması

Şekil 1'de önerilen yaklaşımın algoritması Algoritma 1'de detaylandırılmıştır. Bu yaklaşım ile 9 veri kümesi, 6 farklı öznitelik seçimi yöntemi ile test edilmiştir. Veri kümelerine göre elde edilen doğruluk oranlarındaki değişim Şekil 2'de sunulmuştur.

Algoritma 1: En iyi N elemanlı alt küme yaklaşımı algoritması

```
1   V ← veri kümesini yükle;
2   filtreTabanliPuanlar[] ← filtre tabanlı bir yöntemle öznitelikler için puanları hesapla;
3   oznitelikSirasi[] ← sırala(V, filtreTabanliPuanlar);
4   oznitelikSayisi ← toplam öznitelik sayısı;
5   for i ← 1 to oznitelikSayisi do
6       seciliOznitelikler[i] ← false;
7   end
8   for j ← 1 to oznitelikSayisi do
9       seciliOznitelikler[oznitelikSirasi[j]] ← true;
10      basariMetrikleri ← siniflandir(V, seciliOznitelikler);
11      kaydet(basariMetrikleri);
12  end
```



Şekil 2. Veri kümelerinin öznitelik seçimi yöntemlerine göre sınıflandırma doğruluklarındaki değişim

Şekil 2 incelendiğinde bütün veri kümelerinde en yüksek sınıflandırma başarısına hemen hemen ilk 400 öznitelik kullanılarak ulaşıldığı görülmektedir. Öyle ki bazı veri kümelerinde ilk 50 öznitelik ile ciddi bir başarı artışının olduğu göze çarpmaktadır. Grafikler incelendiğinde tüm veri kümelerinde Korelasyon tabanlı öznitelik seçimi yaklaşımının başarılı sonuçlar sergilediği, Simetrik belirsizlik katsayısı ve OneR yöntemlerinin derecelendirdiği özniteliklerle oluşturulan alt kümelerde ise öznitelik sayısı arttıkça sınıflandırma başarısında düşüşün gerçekleştiği görülmektedir. Genel olarak Ki-kare, Bilgi kazancı ve Korelasyon tabanlı öznitelik seçimi yöntemleri hızlı bir şekilde en iyi sonuçlarını yakalarken, Kazanç oranı, OneR ve Simetrik belirsizlik katsayısından elde edilen alt kümelerde en iyi sonucu yakalamak için daha fazla sayıda özniteliğe ihtiyaç duyulmuştur. Şekil 2'de sunulan deneylerin en iyi sonuçlarına ait ayrıntılı sayısal veriler Tablo 4'te verilmiştir.

Tablo 4. Veri kümelerinden öznitelik seçimi yöntemlerine göre elde edilen en iyi sonuçlara ait sayısal veriler

	camera					camp				
Yöntem	# öz.	Doğruluk	Kesinlik	Duyarlılık	F-ölç.	# öz.	Doğruluk	Kesinlik	Duyarlılık	F-ölç.
Korelasyon	427	92.37	0.934	0.911	0.922	974	94.65	0.959	0.933	0.946
Ki-Kare	236	93.37	0.935	0.931	0.933	378	94.15	0.956	0.925	0.941
Kazanç Oranı	226	93.57	0.935	0.935	0.935	372	94.28	0.956	0.928	0.942
Bilgi Kazancı	236	93.37	0.935	0.931	0.933	378	94.15	0.956	0.925	0.941
OneR	376	87.55	0.894	0.851	0.872	994	94.40	0.952	0.935	0.944
Simetrik Belirsizlik	236	93.37	0.935	0.931	0.933	378	94.15	0.956	0.925	0.941

	doctor					drug				
Yöntem	# öz.	Doğruluk	Kesinlik	Duyarlılık	F-ölç.	# öz.	Doğruluk	Kesinlik	Duyarlılık	F-ölç.
Korelasyon	1014	92.76	0.939	0.915	0.927	327	87.78	0.867	0.893	0.880
Ki-Kare	563	92.02	0.933	0.905	0.919	211	84.41	0.842	0.848	0.845
Kazanç Oranı	555	91.88	0.932	0.904	0.918	209	84.41	0.842	0.848	0.845
Bilgi Kazancı	591	92.02	0.927	0.912	0.920	211	84.41	0.842	0.848	0.845
OneR	1249	90.66	0.908	0.905	0.907	679	84.16	0.819	0.878	0.847
Simetrik Belirsizlik	572	91.95	0.932	0.905	0.918	209	84.41	0.842	0.848	0.845

	laptop					lawyer				
Yöntem	# öz.	Doğruluk	Kesinlik	Duyarlılık	F-ölç.	# öz.	Doğruluk	Kesinlik	Duyarlılık	F-ölç.
Korelasyon	671	98.30	0.989	0.977	0.983	632	98.18	0.982	0.982	0.982
Ki-Kare	316	96.02	0.955	0.966	0.960	139	92.27	0.919	0.927	0.923
Kazanç Oranı	226	96.59	0.966	0.966	0.966	292	92.72	0.898	0.964	0.930
Bilgi Kazancı	316	96.02	0.955	0.966	0.960	139	92.27	0.919	0.927	0.923
OneR	975	97.72	0.967	0.989	0.978	449	97.27	0.973	0.973	0.973
Simetrik Belirsizlik	291	96.02	0.955	0.966	0.960	138	92.27	0.919	0.927	0.923

	music					radio				
Yöntem	# öz.	Doğruluk	Kesinlik	Duyarlılık	F-ölç.	# öz.	Doğruluk	Kesinlik	Duyarlılık	F-ölç.
Korelasyon	384	88.83	0.853	0.938	0.894	797	90.14	0.897	0.906	0.902
Ki-Kare	278	85.40	0.824	0.900	0.860	289	83.96	0.835	0.847	0.841
Kazanç Oranı	281	85.57	0.818	0.914	0.864	314	84.56	0.838	0.857	0.847
Bilgi Kazancı	323	85.40	0.832	0.887	0.859	287	84.36	0.835	0.857	0.846
OneR	732	88.14	0.868	0.900	0.884	987	86.35	0.863	0.865	0.864
Simetrik Belirsizlik	263	85.57	0.817	0.918	0.864	304	84.26	0.831	0.861	0.845

	tv				
Yöntem	# öz.	Doğruluk	Kesinlik	Duyarlılık	F-ölç.
Korelasyon	435	94.47	0.945	0.945	0.945
Ki-Kare	142	89.57	0.939	0.847	0.890
Kazanç Oranı	142	89.36	0.938	0.843	0.888
Bilgi Kazancı	134	89.15	0.938	0.838	0.885
OneR	429	92.98	0.931	0.928	0.930
Simetrik Belirsizlik	144	89.36	0.943	0.838	0.887

Tablo 4 incelendiğinde camera veri kümesi için OneR yaklaşımı dışında %90'ın üzerinde sınıflandırma başarısı yakalanmıştır. En başarılı sınıflandırma Kazanç oranı tarafından seçilen öznitelik alt kümesi ile sağlanmıştır. Bu sınıflandırma için 226 adet öznitelik kullanılmıştır. camp veri kümesinde bütün yaklaşımlar ile %94'ün üzerinde doğruluk oranı elde edilmiştir. Elde edilen en iyi sonuç %94.65 (974 öznitelik) ile Korelasyon tabanlı öznitelik seçimi ile sağlanmıştır. Kazanç oranı ile ise 372 adet öznitelik ile %94.28 doğruluk oranı yakalanmıştır. drug ve music veri kümelerinde ise en iyi sınıflandırma doğruluğu %87.78 (327 öznitelik) ve %88.83'te (384 öznitelik) sınırlı kalmıştır. laptop ve lawyer veri kümelerinde %98'in üzerinde bir başarı yakalanmıştır. Bu sınıflandırmalar için özniteliklerin 3'te 1'inden daha az sayıda özniteliğe ihtiyaç duyulmuştur. radio veri kümesinde %90, tv veri kümesinde %94 ve son olarak doctor veri kümesinde %92'lik bir sınıflandırma başarısı yakalanmıştır. Bu sınıflandırmalar için radio ve doctor veri kümelerinde özniteliklerin yaklaşık olarak yarısı, tv veri kümesinde ise yaklaşık %18'i kullanılmıştır. Veri kümelerine göre (sırasıyla camera, camp, doctor, drug, laptop, lawyer, music, radio ve tv) hangi öznitelik yönteminin kaçıncı sırada yer aldığının daha anlaşılır bir şekilde sunmak amacı ile Tablo 5 oluşturulmuştur. Tabloda ÖS öznitelik seçimi yöntemini, Doğ. sınıflandırma doğruluk oranını, #öz. öznitelik sayısını ifade etmektedir. Bunların dışında kullanılan kısaltmalarda KO Kazanç oranı, KK Ki-kare, BK Bilgi kazancı, SB Simetrik belirsizlik katsayısı, KR Korelasyon tabanlı öznitelik seçimi, 1R ise OneR yöntemlerini ifade etmektedir.

Tablo 5. Öznitelik seçimi yöntemlerinin başarı sırası

Veri	1. sıra			2. sıra			3. sıra			4. sıra			5. sıra			6. sıra		
kümesi	ÖS	Doğ.	#öz.	ÖS	Doğ.	#öz	ÖS	Doğ.	#öz	ÖS	Doğ.	#öz	ÖS	Doğ.	#öz	ÖS	Doğ.	#öz.
camera	KO	93.57	226	KK	93.37	236	BK	93.37	236	SB	93.37	236	KR	92.37	427	1R	87.55	376
camp	KR	94.65	974	1R	94.40	994	KO	94.28	372	KK	94.15	378	BK	94.15	378	SB	94.15	378
doctor	KR	92.76	1014	KK	92.02	563	BK	92.02	591	SB	91.95	572	KO	91.88	555	1R	90.66	1249
drug	KR	87.78	327	KO	84.41	209	SB	84.41	209	KK	84.41	211	BK	84.41	211	1R	84.16	679
laptop	KR	98.30	671	1R	97.73	975	KO	96.59	226	SB	96.02	291	KK	96.02	316	BK	96.02	316
lawyer	KR	98.18	632	1R	97.27	449	SB	92.27	138	KK	92.27	139	BK	92.27	139	KO	92.27	292
music	KR	88.83	384	1R	88.14	732	SB	85.57	263	KO	85.57	281	KK	85.40	278	BK	85.40	323
radio	KR	90.14	797	1R	86.35	987	KO	84.56	314	BK	84.36	287	SB	84.26	304	KK	83.96	289
tv	KR	94.47	435	1R	92.98	429	KK	89.57	142	KO	89.36	142	SB	89.36	144	BK	89.15	134

Korelasyon tabanlı öznitelik seçimi 8 veri kümesi için en iyi öznitelik alt kümesini sağlayan, OneR ise 6 veri kümesinde en iyi ikinci sonucu sağlayan öznitelik seçimi yöntemi olmuştur. camera ve camp veri kümelerinde Ki-kare, Simetrik Belirsizlik ve Bilgi kazancı yöntemlerinden sağlanan en iyi sonuçlar eşit çıkmıştır. doctor veri kümesinde ise Ki-kare ve Bilgi kazancı tarafından eşit doğrulukta sonuçlar elde edilmiştir. Fakat bu sonuç için Ki-kare 563, Bilgi kazancı yöntemi ise 591 adet öznitelik seçmiştir. Öznitelik sayıları göz önünde bulundurulduğunda bu veri kümesi için Ki-kare yöntemi daha başarılı olarak listelenmiştir. drug veri kümesinde ise 4 öznitelik seçimi yaklaşımından aynı sonuçlar elde edilmiştir. Kazanç oranı ve Simetrik belirsizlik 209, Ki-kare ve Bilgi kazancı ise 211 öznitelik ile bu sınıflandırmayı gerçekleştirmiştir. Bu nedenle Kazanç oranı ve Simetrik belirsizlik listede Ki-kare ve Bilgi kazancından önce yer almıştır. laptop %98.30 ile en yüksek başarının yakalandığı veri kümesidir. Bu veri kümesinde Bilgi kazancı, Ki-kare ve Simetrik belirsizlik tarafından sağlanan öznitelik alt kümelerinden eşit doğruluk oranında sınıflandırma gerçekleşmiştir. Sıralamada daha az öznitelik ile aynı sonucu yakalayan Simetrik belirsizlik önde yer almıştır. lawyer veri kümesinde OneR ve Korelasyon tabanlı öznitelik seçimi dışındaki yöntemlerin sağladığı sonuçlar eşit çıkmıştır. music veri kümesinde ise Simetrik belirsizlik ve Kazanç oranı, Ki-kare ve Bilgi kazancı yöntemlerinin sağladığı alt kümelerden elde edilen en iyi sonuçlar eşit çıkmıştır. Veri kümelerine göre elde edilen en başarılı sınıflandırmalara ait sayısal veriler Tablo 6'da sunulmaktadır.

Tablo 6. 9 duygu sınıflandırma veri kümesinden elde edilen en iyi sonuçlar

Veri kümesi	Yöntem	# öznitelik	Doğruluk oranı	Kesinlik	Duyarlılık	F-ölçütü
camera	Kazanç Oranı	226	93.57	0.935	0.935	0.935
camp	Korelasyon	974	94.65	0.959	0.933	0.946
doctor	Korelasyon	1014	92.76	0.939	0.915	0.927
drug	Korelasyon	327	87.78	0.867	0.893	0.880
laptop	Korelasyon	671	98.30	0.989	0.977	0.983
lawyer	Korelasyon	632	98.18	0.982	0.982	0.982
music	Korelasyon	384	88.83	0.853	0.938	0.894
radio	Korelasyon	797	90.14	0.897	0.906	0.902
tv	Korelasyon	435	94.47	0.945	0.945	0.945
Ortalama	-	640.56	94.34	0.942	0.946	0.944

Tablo 6 incelendiğine camera veri kümesi dışında bütün veri kümelerinde en yüksek sınıflandırma doğruluk oranının Korelasyon tabanlı öznitelik seçimi tarafından sağlanan öznitelik alt kümeleri ile elde edildiği görülmektedir. Ortalamada %94.34 doğruluk oranı yakalanmıştır. Bu sınıflandırmalar için kullanılan ortalama öznitelik sayısı ise 640 olarak hesaplanmıştır. Yani, ortalamada özniteliklerin yaklaşık olarak %65'i elenmiştir. Ayrıca ortalama kesinlik drug, music ve radio; ortalama duyarlılık drug ve ortama F-ölçütü değerleri drug ve music veri kümeleri dışında %90'ın üzerinde hesaplanmıştır. Sonuçlar duygu sınıflandırmada filtre tabanlı öznitelik seçiminin etkinliğini göstermektedir.

3.6. Duygu sınıflandırması için son teknoloji algoritmalarla karşılaştırma

Deneysel analizin bu bölümünde, filtre tabanlı öznitelik seçimi yöntemlerinin performansı, 9 genel veri kümesini kullanan diğer son teknoloji duygu sınıflandırma algoritmaları ile karşılaştırılmıştır. Seçilen son teknoloji algoritmalar şunlardır: Gökalp et al. [3] yinelemeli açgözlü (YA) algoritması, Onan et al. [10] hibrit bir budama (HB) yaklaşımı tabanlı algoritması; Onan et al. [9] çok amaçlı diferansiyel değerlendirme tabanlı ağırlıklı oylama topluluğu (AOT) algoritması; Jalilvand and Salim [11] öznitelik birleştirme (ÖB) algoritması; Wang et al. [8] rastgele alt uzay (RAU) yaklaşımı; ve Yang et al. [12] çoklu sınıflandırıcı sistemleri (ÇSS) tabanlı algoritması. Bu çalışma filtre tabanlı yaklaşım (FTY) olarak isimlendirilmiştir. Tablo 7, algoritma/veri kümesi çifti başına doğruluk oranlarını yüzde olarak bildirmektedir.

Tablo 7. 9 duygu sınıflandırması veri kümesi için filtre tabanlı öznitelik seçimi yaklaşımının literatürdeki son teknoloji algoritmalarla doğruluk oranları (%) karşılaştırılması. BY, bilgi yok anlamına gelmektedir.

Veri kümesi	YA	HB	FTY	AOT	ÖB	RAU	ÇSS
camera	97.15	95.92	**93.57**	92.87	79.80	76.49	BY
camp	97.99	96.58	**94.65**	93.74	86.00	85.26	82.89
doctor	95.64	95.65	**92.76**	91.05	86.10	85.03	83.87
drug	92.39	94.27	87.78	89.62	69.50	68.82	BY
laptop	99.89	98.92	98.30	98.86	78.86	79.79	BY
lawyer	99.59	97.90	**98.18**	97.87	80.91	83.86	BY
music	94.97	94.16	88.83	89.82	70.69	69.59	73.18
radio	93.05	93.37	**90.14**	88.60	75.30	70.66	67.75
tv	97.38	96.73	94.47	95.74	79.79	76.06	BY
Ortalama	96.45	95.94	**94.34**	93.13	78.55	77.28	-

Ortalama doğruluk değerlerine göre, ele alınan filtre tabanlı öznitelik seçimi yaklaşımı 6 rakibinden 4'ünü %94,34 ile geride bırakmıştır. Ayrıca, 9 veri kümesinden 5 tanesi için en iyi üçüncü, 1 tanesi en iyi ikinci sonucu sağlamıştır. Bu sonuçlar değerlendirilirken karşılaştırılan çalışmaların sarmalayıcı tabanlı yaklaşımlar olduğu, bu çalışmada ise en iyi N elemanlı alt küme yaklaşımı ile bu sonuçların elde edildiği göz önünde bulundurulmalıdır. Daha küçük ve sınırlı bir arama uzayında elde edilen bu sonuçlar ele alınan yaklaşımın umut verici olduğunu göstermektedir.

4. Sonuç ve Tartışma

Bu çalışmada duygu sınıflandırma için filtre tabanlı öznitelik seçimi yöntemlerinin performansları karşılaştırmalı bir şekilde değerlendirilmiştir. Yöntemlerden elde edilen öznitelik puanlarına göre en iyi N elemanlı alt kümeler oluşturulmuş ve Çok terimli Naive Bayes yöntemi ile sınıflandırılmıştır. Önerilen yaklaşım literatürde yaygın olarak bulunan 9 adet duygu analizi veri kümesi üzerinde test edilmiştir. Elde edilen ortalama sonuçlar son teknoloji algoritmaların 4'ünden daha yüksek sınıflandırma başarısı sunmuştur. Ayrıca, özniteliklerin büyük bir çoğunluğunun elendiği görülmektedir. Bu sayede hem bellekten hem de zamandan tasarruf sağlanmaktadır. Daha küçük bir arama uzayında başarılı sonuçlar sergileyen bu yaklaşımın gelecek çalışmalarda alt bir basamak olarak kullanılıp daha başarılı sonuçlara ulaşılabileceği düşünülmektedir. Gelecek çalışmalarda filtre tabanlı öznitelik seçimi yöntemlerinden elde edilen öznitelik puanlarının sezgi olarak kullanıldığı sarmalayıcı tabanlı öznitelik seçimi algoritmalarının geliştirilerek daha başarılı sonuçların elde edilmesi planlanmaktadır.

Kaynaklar

[1] A. Abbasi, H. Chen and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums", ACM Trans. Inf. Syst. 2018; 26(3): 1-34.

[2] J.R. Chang, H.Y. Liang, L.S. Chen and C.W. Chang, "Novel feature selection approaches for improving the performance of sentiment classification", J. Ambient Intell. Hum. Comput. 2020; 1-14.

[3] O. Gokalp, E. Tasci and A. Ugur, "A novel wrapper feature selection algorithm based on iterated greedy metaheuristic for sentiment classification", Expert Syst. Appl. 2020; 146: 113176.

[4] W. Medhat, A. Hassan and H. Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Eng. J. 2014; 5(4): 1093–1113.

[5] G. Wang, J. Sun, J. Ma, K. Xu and J. Gu, "Sentiment classification: The contribution of ensemble learning", Decis. Support Syst. 2014; 57: 77–93.

[6] P. Kumbhar and M. Mali, "A survey on feature selection techniques and classifcation algorithms for efcient text slassifcation", International Journal of Science and Research, 2013; 14(5): 2319–7064.

[7] J.T. Pintas, L.A. Fernandes and A.C.B. Garcia, "Feature selection methods for text classification: a systematic literature review", Artif. Intell. Rev. 2021; 54(8): 6149-6200.

[8] G. Wang, Z. Zhang, J. Sun, S. Yang and C.A. Larson, "POS-RS: A Random Subspace method for sentiment classification based on part-of-speech analysis", Inf. Process. Manage. 2015; 51(4): 458-479.

[9] A. Onan, S. Korukoğlu and H. Bulut, "A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification", Expert Syst. Appl. 2016; 62: 1-16.

[10] A. Onan, S. Korukoğlu and H. Bulut, "A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification", Inf. Process. Manage. 2017; 53(4): 814-833.

[11] A. Jalilvand and N. Salim, "Feature unionization: a novel approach for dimension reduction", Appl. Soft Comput. 2017; 52: 1253-1261.

[12] K. Yang, C. Liao and W. Zhang, "A sentiment classification model based on multiple multi-classifier systems", In International Conference on Artificial Intelligence and Security, 2019; 287-298.

[13] A. Onan, "Bidirectional convolutional recurrent neural network architecture with group-wise enhancement mechanism for text sentiment classification", Journal of King Saud University-Computer and Information Sciences, 2022; 34(5): 2098-2117.
[14] C. Shao and X. Chen, "Deep-learning-based financial message sentiment classification in business management", Comput. Intell. Neurosci. 2022; 3888675.
[15] J. Khan, N. Ahmad, A. Alam and Y. Lee, "Leveraging Semantic and Sentiment Knowledge for User-Generated Text Sentiment Classification", In Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022), 2022; 101-105.
[16] X. Yang, Y. Li, Q. Li, D. Liu and T. Li, "Temporal-spatial three-way granular computing for dynamic text sentiment classification", Inf. Sci. 2022; 596: 551-566.
[17] E. F. Ayetiran, "Attention-based aspect sentiment classification using enhanced learning through CNN-BiLSTM networks", Knowledge-Based Syst. 2022; 252: 109409.
[18] K. Karga, M. A. Toçoğlu ve A. Onan, "COVID-19 pandemi döneminde eğitimde derin öğrenmeye dayalı duygu analizi", Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, 2022; 24(72): 855-868.
[19] H. Polat ve Y. Ağca, "Tripadvisor kullanıcılarının Türkçe ve İngilizce yorumları kapsamında duygu analizi yöntemlerinin karşılaştırmalı analizi", Abant Sosyal Bilimler Dergisi, 2022; 22(2): 901-916.
[20] Ö. Şahinaslan, H. Dalyan ve E. Şahinaslan, "Naive Bayes sınıflandırıcısı kullanılarak Youtube verileri üzerinden çok dilli duygu analizi", Bilişim Teknolojileri Dergisi, 2022; 15(2): 221-229.
[21] E. Ş. Dinçer, D. Kayaoğlu ve S. Safarlı, "Metin madenciliği ve duygu analizi ile siber zorbalık tespiti", Eskişehir Türk Dünyası Uygulama Ve Araştırma Merkezi Bilişim Dergisi, 2022; 3(2): 38-45.
[22] M. U. Salur ve İ. Aydın, "Türkçe tweetler için derin özellik çıkarımı tabanlı yeni bir duygu sınıflandırma modeli", Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 2022; 34(1): 1-13.
[23] M. Whitehead and L. Yaeger, "Building a general purpose cross-domain sentiment mining model", In 2009 WRI world congress on computer science and information engineering 2019; 4: 472-476.
[24] A. S. Yuksel, F. A. Senel and I. A. Cankaya, "Classification of soft keyboard typing behaviors using Mobile device sensors with machine learning", Arabian J. Sci. Eng. 2019; 44(4): 3929-3942.
[25] X.W. Chen and M. Wasikowski, "Fast: a roc-based feature selection metric for small samples and imbalanced data classification problems", In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008; 124-132.
[26] S. Dey Sarkar, S. Goswami, A. Agarwal and J. Aktar, "A novel feature selection technique for text classification using Naive Bayes", International scholarly research notices, 2014; 2014: 717092.
[27] J. Ding and L. Fu, "A Hybrid Feature Selection Algorithm Based on Information Gain and Sequential Forward Floating Search", Journal of Intelligent Computing, 2018; 9(3): 93.
[28] W. Duch, "Filter methods. In Feature Extraction", Springer, Berlin, Heidelberg 2006; 89-117.
[29] D. Morariu, R. Cretulescu and M. Breazu, "Feature selection in document classification", In The fourth international conference in romania of information science and information literacy, 2013; ISSN-L. 2247-0255.
[30] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification", In AAAI-98 workshop on learning for text categorization 1998; 752(1): 41-48.
[31] V. Vapnik, "The nature of statistical learning theory", New York: Springer, 1995.
[32] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features", In European conference on machine learning, Springer, Berlin, Heidelberg 1998; 137-142.
[33] J. Han and M. Kamber, "Data mining: concepts and techniques", 2nd. University of Illinois at Urbana Champaign: Morgan Kaufmann, 2006.
[34] M. Kantardzic, "Data mining: concepts, models, methods, and algorithms", John Wiley & Sons, 2011.
[35] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, "The WEKA data mining software: an update", ACM SIGKDD explorations newsletter, 2009; 11(1): 10-18.