Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi
Afyon Kocatepe University Journal of Science and Engineering

AKÜ FEMÜBİD **23** (2023) 025102 (370-379) AKU J. Sci. Eng. **23** (2023) 025102 (370-379)
DOI: 10.35414/akufemubid.1172637

Araştırma Makalesi / Research Article

Dengesiz Metin Sınıflandırmada Öznitelik Seçim Yöntemlerinin Etkililiği

Hande TİRYAKİ[1], Alper Kürşat UYSAL[2]
[1] *Eskişehir Teknik Üniversitesi, Lisansüstü Eğitim Enstitüsü, Bilgisayar Mühendisliği Bölümü, Eskişehir.*
[2]*Alanya Alaaddin Keykubat Üniversitesi, Rafet Kayış Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Alanya.*

Sorumlu yazar e-posta: handetiryaki@eskisehir.edu.tr ORCID ID: http://orcid.org/ 0000-0002-1533-6901
e-posta: alper.uysal@alanya.edu.tr ORCID ID: https://orcid.org/0000-0002-4057-934X

Geliş Tarihi: 10.09.2022 Kabul Tarihi: 14.03.2023

Öz

Metin verilerinin sınıflar arasında dağılımı genellikle eşit değildir. Bu durum, metin sınıflandırma işleminde sınıflandırıcıların performansına olumsuz yansımaktadır. Dengesiz metin sınıflandırma ile ilgili birçok çalışma yapılmıştır. Metin sınıflandırma işleminin önemli aşamalarından olan öznitelik seçim aşaması, dengesiz metin probleminde de kritik öneme sahiptir. Öznitelik seçme metotlarının dengesiz metinlerin sınıflandırılması üzerindeki etkisi bu çalışmada etraflıca araştırılmıştır. Bu doğrultuda, iki farklı veri seti üzerinde üç farklı sınıflandırıcı ve dokuz farklı öznitelik seçim metodu ile birçok deney yapılmıştır. Ayrıca öznitelik seçim yöntemlerinin başarıları farklı öznitelik sayılarında da gözlemlenmiştir. NDM, DFSS, PFS, POISSON, CHI2, IG, GINI, DFS ve MDFS olarak adlandırılan 9 farklı öznitelik seçim metodu değerlendirilmiştir. Destek Vektör Makinesi (SVM), Karar Ağacı (DTREE) ve Basit Bayes (MNB) sınıflandırıcıları ile deneysel sonuçlar elde edilmiştir. Reuters-21578 veri setinde DFS ve CHI2 öznitelik seçim yöntemleri Makro-F1 değerlendirme metriği üzerinden yaklaşık en yüksek 80 değerini alırken, SPAM SMS veri setinde, DFS öznitelik seçim yöntemi en yüksek skor olarak 95 ve CHI2 öznitelik seçim yöntemi 94 değerlerini almıştır. Öznitelik seçme metotlarından DFS ve CHI2'nin dengesiz metin sınıflandırmada daha başarılı olduğu görülmektedir.

Anahtar kelimeler
Dengesiz metin sınıflandırma; Boyut indirgeme; Öznitelik seçimi; Terim seçimi

The Effectiveness of Feature Selection Methods for Imbalanced Text Classification

Abstract

The distribution of text data across classes is often imbalanced. This situation has a negative impact on the performance of classifiers in the text classification process. Many studies have been performed on imbalanced text classification. The feature selection stage, which is one of the important stages of the text classification process, is also critical in the imbalanced text classification problem. The effect of feature selection methods on the classification of imbalanced texts has been thoroughly investigated in this study. In this direction, many experiments were carried out with three different classifiers and nine different feature selection methods on two different data sets. In addition, the success of feature selection methods has been observed employing different number of features. Nine different feature selection methods called NDM, DFSS, PFS, POISSON, CHI2, IG, GINI, DFS and MDFS were evaluated. Experimental results obtained with Support Vector Machines (SVM), Decision Tree (DTREE), and Naïve Bayes (MNB) classifiers. On the Reuters-21578 dataset, DFS and CHI2 feature selection methods obtained approximately 80 as the highest Macro-F1 score. On the SPAM SMS dataset, DFS feature selection method obtained 95 and CHI2 feature selection method obtained 94 as the highest Macro-F1 score. It is seen that feature selection methods DFS and CHI2 are more successful than the others for imbalanced text classification.

Keywords
Imbalanced text classification; Dimension reduction; Feature selection; Term selection

1. Giriş

Dengesiz veride, bir sınıftaki örnek sayısı diğer sınıflardaki örnek sayılarından önemli ölçüde azdır (He and Garcia 2009, Kamalov *et al.* 2022). Dengesiz veri problemini çözmek için örnekleme yöntemleri, algoritmik yöntemler ve öznitelik seçme yöntemleri

dahil olmak üzere çok sayıda yaklaşım ortaya konmuştur (Ogura *et al*. 2011, Maldonado *et al*. 2014). Algoritmik yöntemler tek sınıflı öğrenme, topluluk ve maliyete duyarlı öğrenme yöntemleri olmak üzere üç başlık altında incelenebilir (Galar *et al*. 2011). Örnekleme yöntemleri ve algoritmik yöntemler, yüksek boyutlu dengesiz verilerde her zaman yüksek performans vermezler (Chen and Wasikowski 2008). Öznitelik seçimi, dengesizlik sorununu çözmenin yöntemlerinden biri olarak kabul edilmektedir (Ogura *et al*. 2011). Yüksek boyutlu dengesiz metin sınıflandırmanın başarımını iyileştirmek için örnekleme yöntemleri, öznitelik seçimi ve her iki yöntemin birleştirilmesinin sonuçları, öznitelik seçme yöntemlerinin etkisinin örnekleme yöntemlerinden daha fazla olduğunu göstermiştir (Liu and Yu 2005).
Metin sınıflandırma, metin içerikli belgeleri önceden tanımlanmış kategorilere ayırma işlemidir. Metin sınıflandırma işlemi genel olarak öznitelik çıkarma, öznitelik seçimi, öznitelik (terim) ağırlıklandırma ve sınıflandırma aşamalarından oluşur.

Metin sınıflandırma çalışmalarının çoğunda olduğu gibi, öznitelik çıkarma işlemi için kelime çantası (bag of words) yaklaşımı kullanılabilir (Uysal and Gunal 2012, Gasparetto *et al.* 2022). Bu yaklaşımda, belgelerdeki terimlerin sırası göz ardı edilir ve bunların dokümanlarda görülme sıklıkları kullanılır (Uysal *et al*. 2012). Bu nedenle, bir metin koleksiyonundaki benzersiz kelimenin her biri farklı bir öznitelik olarak kabul edilir. Sonuç olarak, bir doküman çok boyutlu bir öznitelik vektörü (Uysal and Gunal 2012) ile temsil edilir. Bir öznitelik vektöründe, her boyut, terim frekansı (TF), terim frekansı-ters doküman frekansı (TF-IDF) vb. ile ağırlıklandırılmış bir değere karşılık gelir (Schütze *et al*. 2008, Hajibabaee *et al.* 2022).

Metinden öznitelik çıkarımı sırasında "dizgelere ayrıştırma", "küçük harfe dönüştürme", "durak kelimeleri ayıklama" ve "köklere indirgeme" ön işleme adımları uygulanmaktadır. Literatürde dengesiz veri setleri üzerinde çalışan öznitelik seçme metotları araştırıldığında, bu konuda pek fazla çalışmaya rastlanmamıştır. Bulunan çalışmalar şu şekilde özetlenebilir:

Dengesiz verilerde öznitelik seçimi için kullanılan FAST (Chen and Wasikowski 2008), özniteliklerin sınıflandırıcıdaki eşik değerini değiştirerek AUC değerini ölçüp öznitelik ayrımını değerlendirir. Gömülü ve çok adımlı yöntemler de araştırmacılar tarafından kullanılmaktadır.

Yüksek boyutlu dengesiz verilerin sınıflandırmasında kullanılan bir diğer öznitelik seçme metodu olan SYMON, sarmalayıcı öznitelik seçme metodu kategorisine girmektedir. Harmony arama algoritmasını kullanmaktadır. SVM-RFE, SVM-BFE, D-HELL, SMOTE-RLF, SMOTE-PCA ile karşılaştırıldığında G-Ortalama (G-Mean) ve F-skor performansı olarak SYMON iyi sonuçlar vermiştir (Moayedikia *et al*. 2017).

PFS ise dengesiz metin sınıflandırma için önerilen bir diğer öznitelik seçme metodudur. Öznitelik seçme alt kategorilerinden filtre kategorisinde yer alıp olasılıksal bir metottur (Pouramini *et al*. 2018).

Bu çalışmada, dengesiz metin sınıflandırma problemi öznitelik seçimi açısından ele alınmıştır. Ayrıca yapılan deneylerde "dizgelere ayrıştırma", "küçük harfe dönüştürme", "durak kelimeleri ayıklama" ve "köklere indirgeme" ön işleme adımları uygulanmıştır. Terim ağırlıklandırma için TF-IDF kullanılmıştır. Değerlendirme metriği olarak özellikle Makro-F1 tercih edilmiştir. Makro-F1, dengesiz veri setlerinde sınıflar bazındaki sınıflandırma başarımını daha iyi temsil etmektedir. Daha kapsamlı analizler yapabilmek amacıyla, 30 ile 500 arasında 5 farklı öznitelik boyutunda Makro-F1 cinsinden sonuçlar karşılaştırılmıştır. Böylece, çok sayıda güncel öznitelik seçim yönteminin veri dağılımının sınıflar bazında farklılaştığı noktalarda ne şekilde performans sağladığı detaylı olarak araştırılmıştır. Bu sayede, öznitelik seçim yöntemleri arasında dengesiz metin sınıflandırma problemi açısından performansı yüksek olanlar tespit edilmiştir.

2. Materyal ve Metot

2.1 Öznitelik seçimi

Öznitelik seçim teknikleri genellikle üç kategoriye ayrılır: filtreler, sarmalayıcılar ve gömülü yöntemler. Filtre teknikler hesaplama açısından hızlıdır; ancak genellikle öznitelik bağımlılıklarını dikkate almazlar (Uysal and Gunal 2012). Filtre tabanlı yöntemler özellikle metin sınıflandırma alanı için yaygın olarak tercih edilmektedir. Metin sınıflandırmada ayırt edici özniteliklerin seçimi için çok sayıda filtre

tabanlı teknik vardır. Bu çalışmada dokuz farklı filtre tabanlı öznitelik seçim yöntemi kullanılmıştır. Bu yöntemler aşağıda detaylı olarak açıklanmıştır. Öznitelik seçim yöntemlerinin formüllerini ortak bir gösterim ile ifade edebilmek için Çizelge 1'deki bilgilerden faydalanılması söz konusu olacaktır. Çizelge 1'de, a değeri sınıf ve ilgili terimin aynı anda bulunduğu durumların sayısını belirtmektedir. d değeri ise hem sınıfın hem de ilgili terimin aynı anda bulunmadığı durumların sayısını temsil etmektedir. c ve d değerleri de bu mantıksal çerçevede hesaplanmaktadır.

Çizelge 1. İkili Olasılık Çizelgesi için Genel Gösterim

	t_i	$\overline{t_i}$
C_j	a	b
$\overline{C_j}$	c	d

2.1.1 Normalleştirilmiş Fark Ölçütü (NDM)

Normalleştirilmiş fark ölçütü, göreli doküman frekanslarını hesaba katan yeni bir öznitelik sıralama ölçüsüdür (Rehman *et al*. 2017). NDM'nin formülü aşağıda verilmiştir:

$$NDM = \frac{\left|\frac{a}{a+b} - \frac{c}{c+d}\right|}{\min\left(\frac{a}{a+b}, \frac{c}{c+d}\right)} \quad (1)$$

$\min\left(\frac{a}{a+b}, \frac{c}{c+d}\right) = 0$ olduğu durumda, NDM değerinin sonsuz çıkmaması için payda küçük bir değerle değiştirilir.

2.1.2 Ayrımcı Öznitelik Seçimi (DFSS)

DFSS, dokümanlardaki öznitelikleri ayırt edici güçle seçen, ardından öznitelikler ve belgeler arasındaki anlamsal benzerliği hesaplayan yeni bir öznitelik seçim yöntemidir (Zong *et al*. 2015). Formülü aşağıda belirtilmiştir:

$$DFSS = \frac{\frac{a\_tf}{a}}{\frac{c\_tf}{c}} * \frac{a}{a+c} * \frac{a}{a+b} * \left|\frac{a}{a+c} - \frac{b}{b+d}\right| \quad (2)$$

*a*_tf : t teriminin o sınıfta geçme sayısı

c_tf : t terimin o sınıf hariç tüm sınıflarda geçme sayısı

2.1.3 Olasılıksal Öznitelik Seçimi (PFS)

Dengesiz metin sınıflandırma için önerilen tek-taraflı öznitelik seçme metodu PFS, öznitelik dağılımını hesaba katan olasılıksal bir metottur (Pouramini *et al*. 2018). PFS formülü aşağıdaki gibidir:

$$PFS = \frac{\left(\frac{a}{a+b}\right) + \left(\frac{a}{a+c}\right)}{(a+b)/N} + \frac{\left(\frac{d}{d+c}\right) + \left(\frac{d}{d+b}\right)}{(c+d)/N} \quad (3)$$

$\frac{a}{a+b}$: O sınıfta t teriminin olma olasılığı

$\frac{a}{a+c}$: t teriminin geçtiği durumlarda t'nin o sınıfta olma olasılığı

$\frac{d}{d+c}$: O sınıf hariç tüm sınıflarda t teriminin olmama olasılığı

$\frac{d}{d+b}$: t teriminin olmadığı durumlarda o sınıfta olmama olasılığı

$(a+b)/N$: t teriminin var olma olasılığı

$(c+d)/N$: t terimin var olmama olasılığı

2.1.4 Poisson Dağılımı (POISSON)

Her bir terimin olasılık dağılımının standart Poisson dağılımından ne kadar saptığına bağlı olarak terimin önemini tahmin eden bir öznitelik seçme yöntemidir (Ogura *et al*. 2009). Formül adımları aşağıda gösterilmektedir:

$$lambda = F / N \quad (4)$$

$$\hat{a} = (a+b) * \left(1 - e^{(-lambda)}\right) \quad (5)$$

$$\hat{b} = (a+b) * e^{(-lambda)} \quad (6)$$

$$\hat{c} = (c+d) * \left(1 - e^{(-lambda)}\right) \quad (7)$$

$$\hat{d} = (c+d) * e^{(-lambda)} \tag{8}$$

$$Poisson = \frac{(a-\hat{a})^2}{\hat{a}} + \frac{(b-\hat{b})^2}{\hat{b}} + \frac{(c-\hat{c})^2}{\hat{c}} + \frac{(d-\hat{d})^2}{\hat{d}} \tag{9}$$

F : t teriminin tüm dokümanlardaki toplam frekansı

N : Bütün sınıflardaki toplam doküman sayısı

a+b : O sınıfa ait olan doküman sayısı

c+d :O sınıfa ait olmayan doküman sayısı

2.1.5 Ki-kare (CHI2)

Öznitelik seçim yöntemleri içinde istatistik dağılımı baz alan ki-kare istatistiği, terim ve sınıf arasındaki bağımlılığın varlığını değerlendirir (Chen and Chen 2011). Ki-Kare ile elde edilen skorun yüksek olması, belirtilen terim ile ilgili sınıf arasındaki bağlantının kuvvetli olduğunu gösterir. Ki-kare formülü aşağıda belirtilmiştir:

$$CHI_2 = N * \frac{\left(a*d - b*c\right)^2}{\left(a+c\right)*\left(b+d\right)*\left(a+b\right)*\left(c+d\right)} \tag{10}$$

2.1.6 Bilgi Kazanımı (IG)

Bir kelimenin varlığının veya yokluğunun herhangi bir sınıf için uygun sınıflandırma seçiminin yapılmasını sağladığı bilgi miktarı IG ile ölçülür (Forman 2003). Bilgi kazanımı formülü aşağıda verilmiştir:

$$IG(t) = \frac{a}{N} * \log_2 * \frac{a*N}{(a+c)*(a+b)} + \frac{b}{N} * \log_2 * \frac{b*N}{(b+d)*(b+a)} + \frac{c}{N} * \log_2 * \frac{c*N}{(a+c)*(c+d)} + \frac{d}{N} * \log_2 * \frac{d*N}{(b+d)*(c+d)} \tag{11}$$

$\frac{a}{N}$: t teriminin o sınıfta olma olasılığı

$\frac{b}{N}$: t teriminin o sınıfta olmama olasılığı

$\frac{c}{N}$: t teriminin o sınıf hariç tüm sınıflarda olma olasılığı

$\frac{d}{N}$: t teriminin o sınıf hariç tüm sınıflarda olmama olasılığı

2.1.7 Gini Katsayısı (GINI)

GINI, karar ağaçlarında en iyi öznitelik dağılımını bulmak için orijinal olarak kullanılan yöntemin geliştirilmiş bir versiyonudur. Doğru, hesaplama karmaşıklığı düşük ve hızlı bir yöntemdir (Shang *et al*. 2007). Formülü aşağıdaki gibidir:

$$GI(t) = \sum_{i=1}^{M} \left(\frac{a}{a+b}\right)^2 * \left(\frac{a}{a+c}\right)^2 \tag{12}$$

Burada (a/a+b), o sınıfta t teriminin olma olasılığıdır. (a/a+c) ise t teriminin geçtiği durumlarda t'nin o sınıfta olma olasılığıdır.

2.1.8 Ayırtedici Öznitelik Seçici (DFS)

DFS, belirtilen kriterlere göre bilgilendirici olmayanları atarken ayırt edici öznitelikleri seçmeyi amaçlayan metin sınıflandırma için başarılı öznitelik seçim algoritmalarından biridir (Uysal and Gunal 2012). DFS aşağıdaki formülle ifade edilebilir:

$$DFS(t) = \sum_{j=1}^{M} \frac{(a/(a+c))}{(b/(a+b)) + (c/(c+d)) + 1} \tag{13}$$

M, sınıfların toplam sayısıdır, (a/a+c), t teriminin geçtiği durumlarda t'nin C_j sınıfında olma olasılığıdır. (b/a+b), C_j sınıfında t teriminin olmama olasılığıdır ve (c/c+d) C_j hariç tüm sınıflarda t teriminin olma olasılığıdır.

2.1.9 Değiştirilmiş Ayırtedici Öznitelik Seçici (MDFS)

Başarılı, ayırt edici bir öznitelik seçici olan DFS'in analizine ve modifikasyonuna dayanarak önerilen değiştirilmiş ayırt edici öznitelik seçici MDFS, özellikle metin sınıflandırmanın terim ağırlıklandırma alanında TF-MDFS olarak uygulanmıştır. TF-MDFS'in basit, etkin ve verimli olduğu gözlemlenmiştir (Chen *et al*. 2021).

Bu çalışmada, MDFS yöntemi öznitelik seçme metodu olarak kullanılmıştır ve DFS'e çarpan olarak (d/(b+d)) eklenerek MDFS formülü elde edilmiştir :

$$MDFS(t) = \sum_{j=1}^{M} \frac{(a/(a+c))*(d/(b+d))}{(b/(a+b))+(c/(c+d))+1} \quad (14)$$

(d/(b+d)), t teriminin olmadığı durumlarda C_j hariç tüm sınıflarda olmama olasılığıdır.

2.2 Sınıflandırıcılar

2.2.1 Destek vektör makineleri (Support vector machines-SVM)

SVM, verileri iki veya daha fazla kategoriye ayırabilen yaygın bir sınıflandırma tekniğidir. Sınıflandırma yapmak, yani iki sınıflı bir örnek uzayında pozitif ve negatif örnekler arasında ayrım yapmak için bir hiper düzlem kullanır. Çalışmalarda libSVM paketinin varsayılan değerleri kullanılmıştır (Chang and Lin 2011).

2.2.2 Karar ağaçları (Decision tree-DT)

Karar ağaçları doğrusal olmayan sınıflandırıcılardır. Karar ağaçları algoritmasının temel amacı, öznitelikleri kategorilere denk gelecek şekilde farklı bölgelere ayırmaktır (Quinlan 1986). İkili sınıflandırma ağacı en çok tercih edilen karar ağacıdır.

2.2.3 Çok terimli Basit Bayes (Multinomial naive bayes-MNB)

Basit Bayes, metin sınıflandırma konusunda sıklıkla kullanılan bir sınıflandırıcıdır. Modelleme ve durum geçişlerini belirtmek için kullanılan yöntemlerden biri olan Basit Bayes (Witten and Frank 2002), genellikle çok terimli verilerin ayrık ve sürekli değişkenlerini modellemek için kullanılır. Çok terimli Basit Bayes ise metin sınıflandırma için tasarlanmış bir Basit Bayes türüdür. Basit Bayes doküman üzerinde belirli kelimelerin varlığı ya da yokluğu ile ilgili modelleme yaparken, çok terimli Basit Bayes kelime sayılarını modelleme ve hesaplamalara öncelik verir.

2.3 Veri setleri

Reuters-21578: Reuters-21578 veri seti, haber makaleleri içeren bir doküman koleksiyonudur. 135 kategoriden oluşmaktadır. Dengesiz metin sınıflandırma için ilk 8 kategori kullanılarak deneysel çalışmalar gerçekleştirilmiştir. Çizelge 2'de Reuters-21578 veri seti için eğitim ve test doküman sayıları verilmiştir.

Spam SMS: SMS Spam Koleksiyonu, SMS Spam araştırması için toplanan bir dizi SMS etiketli mesajdır. Yasal (legitimate) veya spam olmasına göre etiketlenmiş 5.574 mesajdan oluşan bir İngilizce SMS seti içerir. Deneysel çalışmada orijinal versiyon kullanılmıştır. Çizelge 3'te ise SPAM SMS veri seti için eğitim ve test doküman sayıları verilmiştir.

Bu çalışmada, tüm ön işleme adımları uygulanmıştır. Kök bulma işlemi için Porter kök çıkarma algoritması (Moayedikia *et al*. 2017) ve ağırlıklandırma yaklaşımı olarak TF-IDF (Schütze *et al*. 2008) kullanılmıştır. Her iki veri setinde de öznitelik sayıları olarak sırayla 30, 50, 100, 300 ve 500 seçilerek sınıflandırma başarıları ölçülmüştür.

Çizelge 2. Reuters-21578 Veri Seti (R8) Özellikleri

Sınıf adı	Eğitim Doküman Sayısı	Sınıf Dağılımı (%)	Test Doküman Sayısı	Toplam Doküman Sayısı
earn	2877	42	1087	3964
acq	1650	25	719	2369
money-fx	538	8	179	717
grain	433	6	149	582
crude	389	6	189	578
trade	369	5	117	486
interest	347	5	131	478
ship	197	3	89	286
Toplam	**6800**	**100**	**2660**	**9460**

Reuters-21578 Veri Seti (R8), earn ve acq adında sırayla %42 ve %25 dağılımları olan iki majör sınıf ve altı minör sınıf içermektedir.

Çizelge 3. Spam SMS Veri Seti Özellikleri

Sınıf adı	Eğitim Doküman Sayısı	Sınıf Dağılımı (%)	Test Doküman Sayısı	Toplam Doküman Sayısı
spam	522	14	225	747
legitimate	3378	86	1449	4827
Toplam	**3900**	**100**	**1674**	**5574**

Spam SMS Veri Seti ise, bir majör (legitimate) ve bir minör (spam) sınıf içeren binary (ikili) bir veri setidir.

2.4 Değerlendirme metriği (Makro-F1)

Özellikle dengesiz veri setlerinin kullanıldığı deneylerde değerlendirme metriği olarak Makro-F1 tercih edilmektedir. Makro-F1 formülü aşağıda verilmiştir:

$$Makro - F1 = \frac{1}{M} * \sum_{i=1}^{M} \left\{ \frac{2 * TP_{M_i}}{2 * TP_{M_i} + FP_{M_i} + FN_{M_i}} \right\} \quad (15)$$

TP: M_i sınıfında olan ve doğru şekilde sınıflandırılan doküman sayısı
FP: M_i sınıfında olmayan ve yanlış şekilde sınıflandırılan doküman sayısı
FN: esasen M_i sınıfında olmadığı halde yanlış şekilde sınıflandırılan doküman sayısı
M: toplam sınıf sayısı

3. Bulgular

İki farklı veri setinde gerçekleştirilen deneyler sonucu elde edilen sonuçlar Çizelge 4 ve Çizelge 5'te sunulmaktadır.

Çizelge 4. Reuters-21578 Veri Seti için SVM (a), DTREE (b) ve MNB (c) sınıflandırıcıları Makro-F1 değerleri

Öznitelik Seçme Metodları	Öznitelik Sayıları				
	30	**50**	**100**	**300**	**500**
NDM	44,328	49,26	53,71	68,2	72,13
DFSS	77,195	79,959	78,93	79,6	**80,58**
PFS	14,449	14,449	14,27	14,5	20,304
POISSON	76,23	79,26	79,73	**80,8**	80,48
CHI2	74,242	78,806	80,3	78,5	79,79
IG	70,315	77,357	80,03	80,5	80,04
GINI	74,425	79,361	79,73	80,1	**80,86**
DFS	76,771	79,794	80,3	79,1	79,66
MDFS	77,109	77,041	77,32	77,3	78,03

(a)

Öznitelik Seçme Metodları	Öznitelik Sayıları				
	30	**50**	**100**	**300**	**500**
NDM	45,574	50,115	54,28	71,2	71,53
DFSS	74,587	75,013	74,97	76,9	77,35
PFS	15,409	15,409	15,54	15,5	16,723
POISSON	73,301	72,873	74,47	77,7	77,65
CHI2	74,333	77,479	75,36	77,9	**78,37**
IG	71,726	72,999	76,1	**78,4**	77,54
GINI	71,875	74,926	74,63	76,3	77,68
DFS	74,95	76,684	76,65	77,9	**78,43**
MDFS	75,157	76,044	76,3	76,3	75,81

(b)

Öznitelik Seçme Metodları	Öznitelik Sayıları				
	30	**50**	**100**	**300**	**500**
NDM	44,405	49,844	54,93	73,2	74,32
DFSS	74,137	78,56	78,46	77,6	77,13
PFS	15,361	15,354	12,96	13,4	20,353
POISSON	75,88	77,652	78,28	**79,5**	77,8
CHI2	75,356	77,424	**79,5**	**79,5**	78,97
IG	70,659	76,52	76,27	76,9	75,75
GINI	74,458	79,212	78,98	78,8	77,96
DFS	73,69	77,45	78,81	**79,3**	77,59
MDFS	74,769	75,773	76,54	75,8	74,73

(c)

Çizelge 5. SPAM SMS Veri Seti için SVM (a), DTREE (b) ve MNB (c) sınıflandırıcıları Makro-F1 değerleri

Öznitelik Seçme Metodları	Öznitelik Sayıları				
	30	**50**	**100**	**300**	**500**
NDM	88,15	88,925	91,043	94,153	94,028
DFSS	89,802	92,319	93,742	94,747	93,106
PFS	89,482	89,206	91,252	91,134	91,560
POISSON	90,786	93,493	94,684	94,182	94,035
CHI2	90,786	91,645	93,501	94,453	94,059
IG	90,747	92,95	93,815	94,584	**94,787**
GINI	90,904	91,679	93,692	**95,34**	94,828
DFS	89,322	93,035	94,106	**95,216**	**95,482**
MDFS	87,338	89,206	91,252	91,134	91,56

(a)

Öznitelik Seçme Metodları	Öznitelik Sayıları				
	30	**50**	**100**	**300**	**500**
NDM	87,666	87,835	88,679	91,365	89,912
DFSS	89,685	90,394	90,734	89,802	88,554
PFS	88,293	89,295	91,043	91,229	91,529
POISSON	90,823	**91,995**	**92,382**	91,298	91,298
CHI2	90,823	90,393	**92,382**	91,298	91,298
IG	90,979	91,124	91,495	91,48	91,413
GINI	90,979	89,82	90,507	91,48	91,149
DFS	89,823	91,662	91,529	**91,825**	91,645
MDFS	87,054	89,02	90,624	90,739	90,813

(b)

Öznitelik Seçme Metodları	Öznitelik Sayıları				
	30	**50**	**100**	**300**	**500**
NDM	89,423	89,186	91,576	94,807	94,95
DFSS	90,972	92,734	93,621	94,97	94,97
PFS	88,113	88,075	90,959	93,176	93,176
POISSON	91,731	93,166	94,011	94,869	94,869
CHI2	91,731	93,043	94,011	94,869	94,869
IG	92,8	92,771	**95,034**	94,89	94,89
GINI	92,193	93,466	93,986	**95,178**	**95,178**
DFS	90,904	93,221	93,661	**95,428**	**95,428**
MDFS	87,91	86,711	88,016	90,713	90,754

(c)

Çizelge 4'e göre Reuters-21578'in kullanıldığı deneyde SVM Makro-F1 değeri olarak 80 civarında en yüksek değerlerini vermiştir. Onun ardından sırasıyla 79 ve 78 civarında MNB ve DTREE sınıflandırıcılarının sonuçları gelmektedir. Reuters-21578 veri setinde SVM > MNB > DTREE sıralamasını yapmak mümkündür. Makro-F1 değerleri arasında fark çok açık değildir. Yüksek değerler genellikle 300, 500 öznitelik sayılarında görülmüştür. Özellikle DTREE ve MNB sınıflandırıcılarında CHI2 ve DFS 78 ve 79 Makro-F1 değerleriyle diğer öznitelik seçme yöntemlerine göre ön plandadır. PFS ise tüm sınıflandırıcılarda 20 ve altı Makro-F1 değerleriyle en düşük Makro-F1 sonuçlarını vermiştir. PFS yöntemini 74 ve altı Makro-F1 değerleriyle NDM isimli yöntem takip etmektedir.

Çizelge 5'teki SPAM SMS veri seti üzerinde yapılan deneylerde sınıflandırıcı başarı sıralaması SVM ≈ MNB > DTREE şeklindedir. SVM ve MNB 95 Makro-F1 civarında sonuçlar verirken DTREE 91-92 değerleri aralığındadır. Ancak, Makro-F1 başarı düzeyleri Reuters-21578 veri setine göre yüksektir. Öznitelik sayılarına göre bakıldığında SVM ve MNB yüksek öznitelik sayılarında düşük öznitelik sayılarına göre yüksek sonuçlar vermektedir. Buna karşın, DTREE sınıflandırıcı 50, 100, 300 öznitelik boyutlarında diğer öznitelik boyutlarına göre yüksek makro-F1 sonuçları vermiştir. DFS, diğer öznitelik seçme yöntemlerine göre SVM ve MNB'de 95 ve üzeri değerlerle daha başarılı sonuçlar vermiştir. PFS, Reuters-21578 veri setindeki kadar düşük değerler vermemiştir ve diğer öznitelik seçme yöntemleriyle çok yakın Makro-F1 değerleri elde edilmiştir.

Çizelge 4 ve Çizelge 5'te verilen sonuçların daha iyi yorumlanabilmesi için Çizelge 6 ilave olarak sunulmuştur. Çizelge 6'da, deneyler sonucunda elde edilen Makro-F1 değerlerine göre en iyi skorları üreten öznitelik seçme yöntemleri görülmektedir.

Çizelge 6. Öznitelik seçim metotlarının en iyi Makro-F1 değerlerini ürettiği durumların sayıları

DFS	**CHI2**	**POISSON**	**GINI**	**IG**	**DFSS**
6	4	3	3	3	1

Çizelge 6'dan görüleceği üzere dengesiz metin sınıflandırmada öznitelik seçimi noktasında başarı açısından DFS ve CHI2 yöntemleri diğer yöntemlere göre önde görünmektedir.

4. Tartışma ve Sonuç

Bu çalışmada Reuters-21578 ve SPAM SMS veri setleri üzerinde SVM, DTREE ve MNB sınıflandırıcıları kullanılarak 9 farklı öznitelik seçme yönteminin başarıları kıyaslanmıştır. 2 farklı veri seti üzerinde yapılan deneyler sonucunda, genel olarak DFS ve CHI2'nin iyi performans verdiği, PFS ve NDM'nin ise özellikle Reuters-21578 veri seti üzerinde kötü performans gösterdiği gözlemlenmiştir. Bu çalışmanın devamı olarak spesifik alanlardaki dengesiz metin sınıflandırma problemi çerçevesinde öznitelik seçim yöntemlerinin performansı araştırılabilir.

5. Kaynaklar

Chang, C.C. and Lin, C.J., 2011. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, **2(3)**, 1-27.

Chen, L., Jiang, L. and Li, C., 2021. Modified DFS-based term weighting scheme for text classification. *Expert Systems with Applications,* **168**, 114438.

Chen, X.W. and Wasikowski, M., 2008. 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,* 124-132.

Chen, Y.T. and Chen, M.C., 2011. Using chi-square statistics to measure similarities for text categorization. *Expert systems with applications,* **38(4)**, 3085-3090.

Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, **3(Mar)**, 1289-1305.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. and Herrera, F., 2011. A review on ensembles for the class imbalance problem: bagging-, boosting-,and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* **42(4)**, 463-484.

Gasparetto, A., Marcuzzo, M., Zangari, A. and Albarelli, A., 2022. A Survey on Text Classification Algorithms: From Text to Predictions. *Information,* **13(2)**, 83.

Hajibabaee, P., Malekzadeh, M., Ahmadi, M., Heidari, M., Esmaeilzadeh, A., Abdolazimi, R. and James Jr, H., 2022. Offensive language detection on social media based on text classification. *In 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC),* 0092-0098. IEEE.

He, H. and Garcia, E.A., 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering,* **21(9)**, 1263-1284.

Kamalov, F., Thabtah, F. and Leung, H.H., 2022. Feature Selection in Imbalanced Data. *Annals of Data Science,* 1-15.

Liu, H. and Yu, L., 2005. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering,* **17(4)**, 491-502.

Maldonado, S., Weber, R. and Famili F., 2014. Feature selection for high-dimensional class-imbalanced data sets using support vector machines. *Information sciences,* **286**, 228-246.

Moayedikia, A., Ong, K. L., Boo, Y. L., Yeoh, W.G. and Jensen, R., 2017. Feature selection for high dimensional imbalanced class data using harmony search. *Engineering Applications of Artificial Intelligence*, **57**, 38-49.

Ogura, H., Amano, H. and Kondo M., 2009. Feature selection with a measure of deviations from Poisson in text categorization. *Expert Systems with Applications,* **36(3)**, 6826-6832.

Ogura, H., Amano, H. and Kondo, M., 2011. Comparison of metrics for feature selection in imbalanced text classification. *Expert Systems with Applications,* **38(5)**, 4978-4989.

Pouramini, J., Minaei-Bidgoli, B. and Esmaeili, M., 2018. A novel feature selection method in the categorization of imbalanced textual data. *KSII Transactions on Internet and Information Systems (TIIS),* **12(8),** 3725-3748.

Quinlan, J.R., 1986. Induction of decision trees. *Machine learning,* **1(1)**, 81-106.

Rehman, A., Javed, K. and Babri, H.A., 2017. Feature selection based on a normalized difference measure for text classification. *Information Processing & Management,* **53(2)**, 473-489.

Schütze, H., Manning, C.D. and Raghavan, P., 2008. Introduction to information retrieval. Vol. **39**. Cambridge University Press Cambridge.

Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y. and Wang, Z., 2007. A novel feature selection algorithm for text categorization. *Expert Systems with Applications,* **33(1)**, 1-5.

Uysal, A.K. and Gunal, S., 2012. A novel probabilistic feature selection method for text classification. *Knowledge-Based Systems*, **36**, 226-235.

Uysal, A.K., Günal, S., Ergin, S. and Günal, E.Ş., 2012. Detection of SMS spam messages on mobile phones. *In 2012 20th Signal Processing and Communications Applications Conference (SIU),* 1-4. IEEE.

Witten, I.H. and Frank, E., 2002. Data mining: practical machine learning tools and techniques with Java implementations. *Acm Sigmod Record,* **31(1)**, 76-77.

Zong, W., Wu, F., Chu, L. K. and Sculli, D., 2015. A discriminative and semantic feature selection method for text categorization. *International Journal of Production Economics,* **165**, 215-222.