BŞEÜ Fen Bilimleri Dergisi
9(2), 672-682, 2022

BSEU Journal of Science
https://doi.org/10.35193/bseufbd.993833

e-ISSN:2458-7575 (https://dergipark.org.tr/tr/pub/bseufbd)

Araştırma Makalesi - Research Article

Metin Sınıflandırma için Öznitelik Ağırlıklandırma Metotlarının Lokal Öznitelik Seçim Metotları Üzerindeki Rolü

The Role of Feature Weighting Methods on Local Feature Selection Methods for Text Classification

Bekir Parlak[1*]

Geliş / Received: 10/09/2021 ***Revize / Revised: 11/05/2022*** ***Kabul / Accepted: 15/08/2022***

ÖZ

İnternet teknolojilerinin gelişimiyle birlikte metinsel verilerde ciddi bir artış yaşanmıştır. Bu metinsel verilerin anlamlı hale gelebilmesi için otomatik metin sınıflandırma yaklaşımları önemli hale gelmiştir. Otomatik metin sınıflandırma yaklaşımlarında öznitelik seçimi ve öznitelik ağırlıklandırma önemli bir yer tutar. Bu çalışmada, öznitelik ağırlıklandırma metotlarının lokal öznitelik seçim metotları üzerindeki etkisi ayrıntılı bir şekilde incelenmiştir. Çalışmada iki farklı ağırlıklandırma metodu, üç farklı lokal öznitelik seçim metodu, üç farklı kriter veri kümesi ve iki sınıflandırıcı kullanılmıştır. En yüksek Mikro-F1 ve Makro-F1 skoru, Reuters-21578 veri kümesi için 92.88 ve 65.55, 20Newsgroup veri kümesi için 99.02 ve 98.15, Enron1 veri kümesi için 97.19 ve 93.40'tır. Deneysel sonuçlar, OddsRatio (OR) öznitelik seçim metodu, Terim Frekansı (TF) öznitelik ağırlıklandırma ve Destek Vektör Makinesi (DVM) sınıflandırıcı kombinasyonu ile daha iyi sonucun elde edildiğini göstermektedir.

Anahtar Kelimeler- Metin Sınıflandırma, Öznitelik Seçimi, Öznitelik Ağırlıklandırma

ABSTRACT

With the development of internet technologies, there has been a significant increase in textual data. Automatic text classification approaches have become important in order for these textual data to become meaningful. Feature selection and feature weighting have an important place in automatic text classification approaches. In this study, the effect of feature weighting methods on local feature selection methods is examined in detail. Two different weighting methods, three different local feature selection methods, three different criteria datasets, and two classifiers were used in the study. The highest Micro-F1 and Macro-F1 scores were 92.88 and 65.55 for the Reuters-21578 dataset, 99.02 and 98.15 for the 20Newsgroup dataset, and 97.19 and 93.40 for the Enron1 dataset. Experimental results show that better results are obtained with the combination of Odds Ratio (OR) feature selection method, Term Frequency (TF) feature weighting and Support Vector Machine (SVM) classifier.

Keywords- Text Classification, Feature Selection, Feature Weighting

[1]*Sorumlu yazar iletişim: bekir.parlak@amasya.edu.tr (https://orcid.org/0000-0001-8919-6481)
Bilgisayar Mühendisliği Bölümü, Amasya Üniversitesi, Yeşilırmak Yerleşkesi, Amasya, Türkiye

BŞEÜ Fen Bilimleri Dergisi
9(2), 672-682, 2022

BSEU Journal of Science
https://doi.org/10.35193/bseufbd.993833

e-ISSN:2458-7575 (https://dergipark.org.tr/tr/pub/bseufbd)

I. GİRİŞ

İnternet erişiminin artmasıyla birlikte, metinsel veriler her alanda ciddi artış göstermiştir. Bu metinsel verilerin daha anlamlı hale gelebilmesi için bu verilerin organizasyonu da önemli hale gelmiştir. Metin sınıflandırma(MS) metinsel verilerin içeriklerine göre önceden tanımlanmış sınıflara atanmasında önemli bir rol oynar [1]. MS birkaç adımdan oluşmaktadır. Bunlar önişleme, öznitelik çıkartma, öznitelik seçimi, öznitelik ağırlıklandırma ve sınıflandırma işlemleridir. Önişleme adımları [2], metin dokümanlarını daha kompakt ve uygulanabilir hale getirmek için kullanılır. Önişleme adımında dizgelere ayırma, gereksiz kelimelerin atılması, küçük harf dönüşümü, kök bulma gibi teknikler uygulanmaktadır. Daha sonra önişlemden geçirilen kelimeler öznitelik çıkartma aşamasında kelime-çantası tekniği ile sayısal hale dönüştürülür. Ancak küçük veri kümelerinde bile yüzbinlerce öznitelik bulunabilir. Bu özniteliklerin tamamını kullanmak hem performansı hem de işlem zamanını arttıracaktır. Bu yüzden kelime-çantasındaki bazı önemli özniteliklerin seçilmesi sınıflandırma performansı açısından oldukça önemlidir. Öznitelik seçiminin yanında bu seçilen özniteliklere atanan ağırlık değerleri de sınıflandırma performansını ciddi oranda etkilemektedir.

Öznitelik seçimi, kuşkusuz MS çalışmalarının en önemli aşamalarından bir tanesidir. Seçilen öznitelikler vektör uzay modelinde gösterilirken her dokümanda bulunduğu frekansa göre bir ağırlık atanır. Bu işlem öznitelik ağırlıklandırma olarak adlandırılır. Öznitelik ağırlıklandırma da öznitelik seçimine benzer şekilde sınıflandırma performansına etki etmektedir. Son aşama olarak sınıflandırma, ana bileşenlerden bir tanesidir. Bu aşamada, önceden tanımlanan dokümanlardan öğrenme fonksiyonu uygular ve sınıf etiketleri bilinmeyen dokümanları sınıflandırmak için kullanılır.

Metinsel verilerde öznitelik seçimi çoğunlukla, tüm özniteliklerin önemlerini tahmin eden bir ölçü ile değerlendirildiği ve ardından en yüksek puanlara sahip ilk N özniteliklerin seçildiği öznitelik sıralamalarına dayanır[3]. Öznitelik seçme yöntemleri lokal ve global olmak üzere iki ana gruba ayrılmaktadır. Lokal politikada her sınıf farklı bir öznitelik kümesi ile temsil edilirken, global politikada öznitelik kümesi global olarak oluşturulur ve tüm sınıflar için aynıdır.

Yerel öznitelik seçim yöntemleri söz konusu olduğunda, birden fazla lokal skoru küresel bir skora dönüştürmek için bir küreselleştirme tekniği gerekir. Bu teknikler TOPLAM, AĞIRLIKLI TOPLAM, MAKSİMUM literatürde sıklıkla kullanılmaktadır [3]. Öte yandan, global öznitelik seçim yöntemleri söz konusu olduğunda, skorlar doğrudan öznitelik sıralaması için kullanılabilir. Öznitelikler azalan düzende sıralanır ve ilk N öznitelikleri, öznitelik alt kümesine dahil edilir. N genellikle deneysel olarak belirlenmiş bir sayıdır. Metin sınıflandırması için global öznitelik seçim yöntemlerinin örnekleri DF (Document Frequency), IG (Information Gain), GI (Gini Index), DFS (Distinguishing Feature Selector) iken, yerel öznitelik seçim yöntemlerinin örnekleri ise MI (Mutual Information), OR (Odds Ratio), CHI2 (Chi-Square)'dir [4].

Eğitim dokümanlarında çok ciddi sayıda öznitelik açığa çıkmaktadır. Ayrıca bu özniteliklerden bazıları önemsiz olabilmektedir. Bundan dolayı, eğitim dokümanlarından çıkartılan özniteliklerden optimum sayıda öznitelik seçmek için öznitelik seçim algoritmaları çok önemlidir. Öznitelik ağırlıklandırma metotları, MS çalışmalarında bir takım avantajlar sağlamaktadır [1]. Her öznitelik bulunduğu dokümanlarda eşit sayıda bulunmamaktadır. Bu yüzden her özniteliğe bulunduğu dokümanlara göre farklı ağırlık değerleri atanır.

Literatürde birçok lokal öznitelik seçim çalışmaları ile ilgili çalışmalar olsa da, metin sınıflandırması için öznitelik seçimi halen devam eden bir araştırma konusudur [5]. Ancak lokal öznitelik seçim yöntemleri üzerinde öznitelik ağırlıklandırma tekniklerinin etkisi konusunda literatürde çok fazla çalışma bulunmamaktadır. Bu çalışmada, metin sınıflandırma alanında temel olarak kullanılan Terim Frekansı (TF) ve Terim Frekansı-Ters Doküman Frekansı (TF-IDF) tekniklerinin üç farklı lokal öznitelik seçim metotları üzerindeki etkisi ayrıntılı olarak analiz edilmiştir. Ayrıca çalışmanın etkinliğini arttırmak için farklı öznitelik boyutları, farklı sınıflandırıcılar ve farklı başarı kriterleri ile farklı veri kümeleri kullanılmıştır. Deneysel sonuçlara bakıldığında, TF tekniği rekabetçi bir performans göstermiştir.

Makalenin geri kalanı şu şekilde organize edilmiştir: Literatürde lokal ve global öznitelik seçim teknikleri ve terim ağırlıklandırma ile ilgili yapılan çalışmalar Bölüm 2'de kısaca açıklanmıştır. Bölüm 3, bu çalışmadaki deneysel aşamaları açıklamaktadır. Son olarak, Bölüm 4'te bazı açıklamalar sunulmuştur.

BSEU Journal of Science
https://doi.org/10.35193/bseufbd.993833

e-ISSN:2458-7575 (https://dergipark.org.tr/tr/pub/bseufbd)

II. İLGİLİ ÇALIŞMALAR

Öznitelik seçimi ve ağırlıklandırma, çok sayıda öznitelik içeren veri kümelerinin yaygın olduğu sınıflandırma görevlerinde önemli bir adımdır. Metin sınıflandırmaya odaklanan çok sayıda öznitelik seçim ve ağırlıklandırma çalışması vardır [1-3]. Metin veri kümeleri binlerce öznitelik içerdiğinden, metin sınıflandırma performansını düşürmektedir. Bu nedenle literatürde önerilen birçok öznitelik seçim yöntemi ve öznitelik ağırlıklandırma çalışmaları mevcuttur [1,3-6].

Metin sınıflandırma çalışmalarında sınıflandırma aşamaları kadar öznitelik seçimi de önemli bir konudur. Literatürde araştırmacılar tarafından tasarlanan öznitelik seçim yöntemlerini değerlendiren birçok çalışma bulunmaktadır. Öznitelik seçim yöntemleri, öznitelikleri değerlendirmek için lokal ve global olmak üzere ikiye ayrılır. Lokal politikada, bir öznitelik sınıf bazında değerlendirilir, daha sonra toplam, ağırlıklı toplam ve maksimum olan globalleştirme teknikleri ile nihai skor hesaplanır [3]. Ancak, bir özniteliğin skoru, global politikada global olarak hesaplanır. Global yöntemlerde özniteliklere tek bir skor atanır. Bu alanda araştırmacılar tarafından öznitelik seçme yöntemlerini ve globalleştirme politikalarını değerlendirmeye yönelik birçok çalışma yapılmıştır [1,2]. Ancak bu çalışmaların bulgularını genellemek oldukça zordur. Çünkü veri kümeleri, ön işleme, öznitelik ağırlıklandırma yöntemleri, globalleştirme politikası gibi deneysel ayarların çeşitliliği çok farklıdır.

Forman [7] lokal politikayı dikkate alarak öznitelik seçim yöntemleri hakkında kapsamlı bir çalışma yapmıştır. Deneylerde dengeli ve dengesiz birçok veri kümesi kullanılmıştır. Ayrıca sınıflandırma algoritması olarak SVM kullanılmıştır. Debole ve Sebastiani [8], öznitelik seçim politikalarını değerlendirdi. Öznitelik seçim skorlarını kullanan yeni bir öznitelik ağırlıklandırma şemasına odaklansalar da, globalleştirme politikaları hakkında ayrıntılı bilgi vermezler. Özgür ve arkadaşları [9], sınıflandırma algoritması olarak SVM kullanarak lokal ve global teknikleri karşılaştıran bir çalışma yaptılar. Ayrıca lokal ve global teknikler sırasıyla sınıf tabanlı ve derlem tabanlı anahtar kelime seçimi olarak adlandırılır. Taşçı ve Güngör [10], metin sınıflandırmada güncel öznitelik seçme yöntemlerini kullanarak öznitelik seçim politikalarını karşılaştırdılar. Deneylerde farklı özelliklere sahip veri kümelerini kullanmışlardır. Politikaların değerlendirilmesinin yanı sıra yeni öznitelik seçim yöntemleri önerdiler. Uysal [4], metin sınıflandırma alanı için geliştirilmiş bir global öznitelik seçim şeması (IGFSS) önerdi. IGFSS, global ve tek taraflı lokal öznitelik seçim yöntemini birleştiren bir topluluk yöntemidir. Öznitelik alt kümesi, bu yöntemleri birleştirerek sınıfları neredeyse eşit olarak temsil eder. IGFSS, öznitelik seçim yöntemlerinin tekli performansından daha başarılıdır. Daha sonraki bir çalışmada, Uysal [11], metin sınıflandırma için filtre tabanlı lokal öznitelik seçme yöntemleri ile öznitelik dönüştürme ve sarmalayıcı tabanlı öznitelik seçme yöntemlerinin birleştirilmesiyle oluşturulan iki aşamalı öznitelik seçim yöntemlerini kapsamlı bir şekilde analiz etmiştir. Öncelikle bazı lokal öznitelik seçim yöntemleri ve üç adet öznitelik alt küme oluşturma yöntemi kullanılmıştır. İkinci olarak, temel bileşen analizi (PCA), gizli anlamsal indeksleme (LSI) veya genetik algoritmalar kullanılmıştır. Micro-F1 ve Macro-F1 ölçüleri kullanılarak Reuters ve Ohsumed olmak üzere iki genel veri kümesi üzerinde lineer bir SVM sınıflandırıcı ile farklı ayarlar yapılmıştır. Parlak ve Uysal [3], farklı küreselleşme tekniklerinin üç lokal öznitelik seçme yöntemi üzerindeki etkilerini dört kıyaslama veri kümesi kullanarak kapsamlı bir şekilde analiz ettiler. Deneylerde, DVM sınıflandırıcı DT sınıflandırıcıdan daha başarılıdır. DFSS (Discriminative Feature Selection) yöntemi çok sınıflı veri kümelerinde OR ve CHI2 yöntemlerinden daha iyi performans gösterirken, ikili sınıf veri kümelerinde CHI2 yöntemi OR ve DFSS yöntemlerinden daha başarılıdır. Kou ve arkadaşları [12], küçük örnek veri kümeleriyle öznitelik sıralamasını değerlendirmek için Çoklu-Kriter Karar Verme (MCDM) tabanlı yaklaşımları kullanır. MCDM yöntemlerinin performansı, on öznitelik seçim yöntemiyle gösterilmektedir. Deneysel sonuçları üç sınıflandırıcı, beş MCDM yöntemi ve on veri kümesi arasında karşılaştırırlar. Agnihotri ve arkadaşları [13], metin sınıflandırması için Değişken Global Öznitelik Seçim Şeması (VGFSS) adlı yeni bir öznitelik seçim şeması önerdi. Yöntem, kategorilerdeki öznitelik sayısına göre her kategoriden farklı sayıda öznitelik seçer. VGFSS yöntemi, genel veri kümelerinde diğer yedi güncel yöntemden daha başarılıdır. Daha sonraki bir çalışmada [14], metin sınıflandırma performansını iyileştirmek için yeni bir Yumuşak Oylama Tekniği (SVT) önerdiler. SVT yöntemi, beş genel veri kümesi üzerinde dört sınıflandırıcı kullanılarak test edilir. Deneysel çalışmalar, SVT yönteminin standart yöntemlere göre sınıflandırıcıların performansında önemli bir gelişme sağladığını göstermektedir. Parlak ve Uysal [15] öznitelik seçimi ve öznitelik ağırlığının tıbbi doküman sınıflandırması üzerindeki etkisi, MEDLINE dokümanlarını içeren iki veri kümesi kullanılarak analiz etmiştir. Gini indeksi(GI) ve ayırt edici özellik seçici(DFS) olmak üzere iki farklı öznitelik seçim yöntemi ile terim frekansı (TF) ve terim frekansı-ters doküman frekansı (TF-IDF) olmak üzere iki farklı terim ağırlıklandırma yönteminin performansları iki örüntü sınıflandırıcı kullanılarak analiz edilmiştir.

III. DENEYSEL ÇALIŞMALAR

İki farklı öznitelik ağırlıklandırma metodu olan TF ve TF-IDF, iki sınıflandırıcı, iki performans metriği ve üç veri kümesi kullanarak test ettik. Bu veri kümeleri, öznitelik seçim metotları, öznitelik ağırlıklandırma metotları, sınıflandırma algoritmaları, performans ölçütleri ve performans analizleri alt bölümlerde açıklanmıştır.

A. Veri Kümeleri ve Önişleme

Bu çalışmada veri kümeleri olarak Reuters-21578, 20Newsgroup ve Enron1 kullanılmıştır. Bu veri kümeleri için sınıf etiketleri ve doküman sayıları Tablo 1-3' te ayrıntılı olarak gösterilmiştir. Tüm veri kümeleri için Porterstemmer algoritması [16] kullanıldı. Gereksiz kelimeler listesine göre veri kümelerinde geçen gereksiz kelimeler çıkarıldı. Ayrıca, ön işleme adımı olarak küçük harfe dönüştürme ve alfabetik olmayan karakterlerin kaldırılması kullanılmıştır [2]. Çok sınıflı veri kümeleri için Reuters-21578, 20Newsgroup veri kümesini kullanırken, ikili sınıf veri kümeleri için Enron1 veri kümesini kullandık. Çok sınıflı veri kümeleri için ilk 10 ve en çok doküman içeren 10 sınıfı kullandık. Reuters-21578 dışındaki tüm veri kümeleri, adil değerlendirme için eğitim (%70) ve test (%30) bölümlerine ayrılmıştır. Reuters-21578 zaten eğitim ve test bölümlerine ayrılmış bir şekilde oluşturulmuştur.

Tablo 1. Reuters-21578 veri kümesi

No.	Sınıf etiketi	Doküman sayısı (eğitim)	Doküman sayısı (test)
1	earn	2877	1087
2	acq	1650	719
3	money-fx	538	179
4	grain	433	149
5	crude	389	189
6	trade	369	117
7	interest	347	131
8	ship	197	89
9	wheat	212	71
10	corn	181	56

Tablo 2. 20Newsgroup veri kümesi

No.	Sınıf etiketi	Doküman sayısı (eğitim)	Doküman sayısı (test)
1	alt.atheism	700	300
2	comp.graphics	700	300
3	comp.os.ms-windows.misc	700	300
4	comp.sys.ibm.pc.hardware	700	300
5	comp.sys.mac.hardware	700	300
6	comp.windows.x	700	300
7	misc.forsale	700	300
8	rec.autos	700	300
9	rec.motorcycles	700	300
10	rec.sport.baseball	700	300

Tablo 3. Enron1 veri kümesi

No.	Sınıf etiketi	Doküman sayısı (eğitim)	Doküman sayısı (test)
1	yasal	2570	1102
2	Yasal olmayan	1050	450

B. Öznitelik Seçim Metotları

Deneysel çalışmalarda öznitelik ağırlıklandırma tekniklerini değerlendirmek için üç farklı lokal öznitelik seçim yöntemi kullanıldı. Bu yöntemler alt bölümlerde açıklanmıştır**.** Ayrıca globalleştirme tekniği olarak MAXIMUM metodu kullanılmıştır.

Tablo 4. Öznitelik Seçim yöntemleri için ön gösterimler

Notasyon	Değer	Anlam
a(TP)	$df(t,C_j)$	C_j sınıfında *t* terimini içeren doküman sayısı
b(FP)	$df(t,\bar{C}_j)$	Diğer sınıflarda ($\bar{C}_j$)*t* terimini içeren doküman sayısı
c(TN)	$df(\bar{t},C_j)$	C_j sınıfında *t* terimini içermeyen doküman sayısı
d(FN)	$df(\bar{t},\bar{C}_j)$	Diğer sınıflarda ($\bar{C}_j$)*t* terimini içermeyen doküman sayısı
e	$tf(t,C_j)$	C_j sınıfında *t* teriminin frekansı
f	$tf(t,\bar{C}_j)$	Diğer sınıflarda ($\bar{C}_j$)*t* teriminin frekansı
N	*(a+b+c+d)*	Tüm sınıflardaki toplam doküman sayısı
M	$count(C_j)$	Toplam sınıf sayısı
$p(t)$	*(a+b)/N*	t teriminin olasılığı
$p(\bar{t})$	*(c+d)/N*	t teriminin bulunmama olasılığı
$p(C_j)$	*(a+c)/N*	C_jsınıfının olasılığı
$p(\bar{C}_j)$	*(b+d)/N*	C_jsınıfının olmaması olasılığı
$p(t,C_j)$	*a/N*	*t* teriminin C_j sınıfında olma olasılığı
$p(t,\bar{C}_j)$	*b/N*	*t* teriminin diğer sınıflarda ($\bar{C}_j$) olma olasılığı
$p(\bar{t},C_j)$	*c/N*	*t* teriminin C_j sınıfında olmama olasılığı
$p(\bar{t},\bar{C}_j)$	*d/N*	*t* teriminin diğer sınıflarda($\bar{C}_j$) olmama olasılığı
$p(t\|C_j)$	*a/(a+c)*	C_jsınıfı varken*t* teriminin olma olasılığı
$p(\bar{t}\|C_j)$	*c/(a+c)*	C_jsınıfı varken *t* teriminin olmama olasılığı
$p(t\|\bar{C}_j)$	*b/(b+d)*	C_jsınıfı yokken ($\bar{C}_j$)*t* teriminin olma olasılığı
$p(\bar{t}\|\bar{C}_j)$	*d/(b+d)*	C_jsınıfı yokken ($\bar{C}_j$)*t* teriminin olmama olasılığı
$p(C_j\|t)$	*a/(a+b)*	*t* terimi varken C_jsınıfının olma olasılığı
$p(\bar{C}_j\|t)$	*b/(a+b)*	*t* terimi varken C_jsınıfının olmama olasılığı
$p(C_j\|\bar{t})$	*c/(c+d)*	*t* terimi yokken C_jsınıfının olma olasılığı
$p(\bar{C}_j\|\bar{t})$	*d/(c+d)*	*t* terimi yokken C_jsınıfının olmama olasılığı

1) DFSS (Discriminative Feature Selection): DFSS, metin sınıflandırması için filtre tabanlı bir öznitelik seçim yöntemidir [17]. Yöntem, önemli öznitelikleri seçmek için bazı kriterler içerir. Bunlar, (i) terim frekansı daha yüksek olan özniteliklerin seçilmesi, (ii) daha yüksek doküman frekansına sahip özniteliklerin seçilmesi, (iii) tüm sınıflardaki dokümanların çoğunda geçen özniteliklerin göz ardı edilmesidir. Bu kriterler kullanılarak DFS skoru şu şekilde hesaplanır:

$$DFSS(t,c_j) = \frac{tf(t,c_j)/df(t,c_j)}{tf(t,\bar{c}_j)/df(t,\bar{c}_j)} \, x \, P(t|c_j) \, x \, P(c_j|t) \, x \, |P(t|c_j) - P(t|\bar{C}_j)| \quad (1)$$

*2) OR (Odds Ratio):*OR formülünden dolayı tek taraflı bir lokal öznitelik seçim yöntemidir. Yöntem hem pozitif hem de negatif öznitelikleri seçer. OR skoru aşağıdaki formül kullanılarak hesaplanabilir:

$$OR(t,c_j) = log \, \frac{P(t|c_j)(1-P(t|\bar{c}_j))}{(1-P(t|c_j))P(t|\bar{c}_j)} \quad (2)$$

*3) CHI2 (Chi-Square):*CHI2, metin sınıflandırması için etkili bir öznitelik seçme yöntemidir. CHI2 yöntemi, t özniteliği ile C sınıfı arasındaki bağımsızlık farkını hesaplar. Olaylar olan A ve B, aşağıdaki durumlarda bağımsız kabul edilir:

$$p(XY) = p(X)p(Y) \quad (3)$$

Öznitelik seçme sürecinde, bu iki olay sırasıyla belirli bir özniteliğin ve sınıfın oluşumuna karşılık gelir. CHI2 puanı şu şekilde hesaplanır:

$$CHI2(t_i, c_j) = \frac{N*(TP*TN-FP*FN)^2}{(TP+FN)*(TP+FP)*(FN+TN)*(FP+TN)} \quad (4)$$

C. Öznitelik Ağırlıklandırma Metotları

Bu çalışmada öznitelik ağırlıklandırma metodu olarak TF(Terim Frekansı) ve TF-IDF(Terim Frekansı-Ters Doküman Frekansı) kullanılmıştır. Bu iki metot metin sınıflandırma çalışmalarında yoğun olarak kullanılan metotlardandır. Ayrıca hem basit olması hem de metin sınıflandırma alanında başarılı performans göstermeleri sebebiyle bu iki temel metot tercih edilmiştir.

Metin veri kümesinde K dokümanları, k ise tek bir dokümanı temsil ettiğini düşünürsek, $f_{k,d}$ ya da TF "d" dokümanında "k" teriminin geçme sayısıdır. t_k k terimini karşılık gelir ve metin dokümanlarında k_i defa geçer.

$$\text{TF(k, d)} = f_{k,d} \{d \text{ dokümanında } k \text{ teriminin geçme sayısı, } k \text{ terimi } d \text{ dokümanında varsa }\} \quad (5)$$

TF-IDF tekniği, bir özniteliğin ağırlığını belirlemek için hem TF hem de IDF değerini kullanır. TF-IDF öznitelik ağırlıklandırma tekniği, metin sınıflandırma alanında yaygın olarak kullanılmaktadır ve diğer terim ağırlıklandırma teknikleri bu tekniğin türevleridir. Sezgisel olarak, TF-IDF tekniği, özniteliğin ilgili metin dokümanı için ne kadar önemli olduğunu belirler. Ters doküman frekansı formülü aşağıdaki gibi hesaplanmaktadır:

$$IDF(t_k) = log\frac{K}{k_i} \quad (6)$$

TF-IDF formülü, metin dokümanındaki bir özniteliğin önemini değerlendirmek için kullanılır. TF-IDF skoru aşağıdaki gibi hesaplanmaktadır:

$$\text{TF-IDF(k,d)} = \text{TF(k,d)} * \text{IDF}(t_k) = f_{k,d} * log\frac{K}{k_i} \quad (7)$$

D. Sınıflandırma Algoritmaları

Bu çalışmada, Destek Vektör Makineleri (DVM) ve Karar Ağacı (KA) olmak üzere iki sınıflandırıcı kullandık. Bu sınıflandırıcıların temel amaçları aşağıda açıklanmıştır:

DVM [18], metin sınıflandırma alanındaki en önemli sınıflandırıcılardan biridir. Ayrıca DVM, çekirdek tipine göre lineer ve lineer olmayan versiyonlar olabilir. Bu çalışmada, metin sınıflandırma alanında yaygın olarak kullanılan DVM'nin doğrusal versiyonu kullanıldı. Bir karar yüzeyi, DVM sınıflandırıcısı tarafından belirlenir. Herhangi bir veri noktasından maksimum derecede uzaktadır. İki sınıflı sınıflandırmayı çoklu sınıflandırmaya modifiye etmek için bütüne karşı bir ve bire karşı bir olan iki yaygın yaklaşımdan biri tercih edilebilir.

KA'lar, uygun bir sınıf tespit edilene kadar sınıfların sırasıyla reddedildiği çok aşamalı karar sistemleridir [19]. Bu amaca ulaşmak için, öznitelikler ilgili sınıflara göre farklı bölgelere ayrılır. İkili sınıflandırma ağacı, en yaygın kullanılan KA türüdür.

E. Performans Ölçütleri

Bu çalışmada öznitelik ağırlıklandırma tekniklerinin lokal öznitelik seçim yöntemleri üzerindeki performansını değerlendirmek için önemli bir başarı ölçütü olan Micro-F1 ve Macro-F1 skorları kullanılmıştır. F-skoru hem kesinlik hem de duyarlılık dikkate alınarak hesaplanır. Makro ortalamada, F-skoru her sınıf için ayrı ayrı hesaplanır ve ardından tüm sınıfların ortalaması hesaplanır. Macro-F1 skorunun hesaplanması aşağıdaki gibi gösterilebilir:

$$Macro - F1 = \frac{\sum_{j=1}^{M} F_j}{M}, \; F_j = \frac{2 \cdot p_j \cdot r_j}{p_j + r_j} \quad (8)$$

Bu formülde, p_j and r_j, sırasıyla j sınıfının kesinlik ve duyarlılık skorlarına karşılık gelir.

Mikro ortalamada F-skoru ise sınıf bilgisi dikkate alınmadan hesaplanmaktadır. Bu nedenle, tüm sınıflandırma kararları tüm dokümanlarda dikkate alınır. Dengesiz veri kümelerini değerlendirirken, mikro

ortalamayı hesaplarken büyük sınıflar küçük sınıflara hükmedebilir. Micro-F1 skorunun hesaplanması aşağıdaki gibi gösterilebilir:

$$Micro - F1 = \frac{2 \cdot p \cdot r}{p + r} \tag{9}$$

Bu formülde p ve r, tüm sınıflar için kesinlik ve duyarlılık değerlerine karşılık gelir. Micro-F1 skoru, daha fazla doküman içeren büyük sınıfların baskın olması nedeniyle tüm durumlar için adil bir değerlendirme sağlamayabilir. Bu nedenle, deneylerde Micro-F1 skorunun yanında Macro-F1 ölçütü kullanılmıştır.

F. Performans Analizi

Bu bölümde, üç lokal öznitelik seçim yöntemi, farklı karakteristiklere sahip üç veri kümesi ve iki adet başarılı sınıflandırıcı kullanılarak kapsamlı bir analiz yapılmıştır. Ayrıca iki farklı öznitelik ağırlıklandırma tekniklerini iyi bilinen lokal öznitelik seçim teknikleri üzerindeki performanslarını karşılaştırdık. Öznitelik seçimi aşamasında lokal öznitelik seçim yöntemlerine örnek olan DFSS, OR ve CHI2 öznitelik seçim yöntemleri kullanılmıştır. Ancak bu çalışmada SVM ve DT olmak üzere iki farklı başarılı sınıflandırıcı kullanılmıştır. Globalleştirme tekniği olarak MAXIMUM tekniği kullanılmıştır. Tüm veri kümeleri için Porterstemmer algoritması [20] kullanıldı. Gereksiz kelimeler, belirlenen gereksiz kelime listesine göre kaldırıldı ve öznitelik ağırlıklandırma aşaması için TF ve TF-IDF [21] kullanıldı. Ayrıca, ön işleme adımı olarak küçük harf dönüşümü ve alfabetik olmayan karakterlerin kaldırılması kullanıldı [2]. Öznitelik ağırlıklandırma tekniklerinin performansını iki farklı sınıflandırıcı üzerinde değerlendirmek için Micro-F1 skoru ve Macro-F1 skoru kullanılmıştır.

Üç lokal öznitelik seçim yöntemi ile seçilen farklı öznitelik boyutları SVM ve DT sınıflandırıcıları ile beslenmiştir. Öznitelik alt kümeleri, 50, 100, 300, 500, 1000 ve 3000 olmak üzere farklı boyutlarda oluşturulmuştur. 16867, 50419 ve 31238 sırasıyla Reuters-21578, 20Newsgroup ve Enron1 [22] veri kümeleri için toplam öznitelik sayısıdır. Micro-F1 ve Macro-F1 [23] skorlarının sonuçları, Reuters-21578 veri kümesi için Tablo 5-10'da gösterilmektedir. Tablolarda öznitelik seçim metoduna göre en yüksek skor kalın şekilde, tablo bazında en yüksek skor hem kalın hem de altı çizili iken, veri kümesi bazında en yüksek skor hem kalın, hem altı çizili, hem de gölgelendirme yapılarak gösterilmiştir.

Tablo 5. DVM ile Reuters-21578 veri kümesinden elde edilen Micro-F1 ve Macro-F1 skorları (%)

	Mikro-F1 skorları (%)						**Makro-F1 skorları (%)**					
	DFSS		**OR**		**CHI2**		**DFSS**		**OR**		**CHI2**	
Dimension	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**
50	90.75	90.36	83.37	83.44	91.66	91.54	58.79	58.23	38.86	39.50	61.36	61.39
100	91.75	91.58	85.76	85.83	92.14	91.85	61.02	61.35	45.64	46.06	61.81	61.79
300	92.46	91.79	87.67	87.46	92.41	91.96	63.04	62.92	50.78	50.87	62.86	63.61
500	92.53	92.39	88.28	88.03	92.46	92.35	64.40	**65.55**	52.71	52.19	63.70	64.16
1000	92.75	92.52	89.61	89.14	92.72	92.31	63.65	64.36	57.29	58.32	63.31	63.26
3000	**92.88**	92.54	**92.45**	91.68	**92.84**	92.56	64.87	63.45	**64.41**	64.04	**64.47**	64.16

Tablo 6. KA ile Reuters-21578 veri kümesinden elde edilen Micro-F1 ve Macro-F1 skorları (%)

	Mikro-F1 skorları (%)						**Makro-F1 skorları (%)**					
	DFSS		**OR**		**CHI2**		**DFSS**		**OR**		**CHI2**	
Dimension	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**
50	89.91	89.71	83.58	83.29	90.28	89.76	58.37	57.14	37.85	36.72	58.32	57.05
100	90.58	90.28	85.47	85.26	90.58	90.47	59.26	**60.05**	44.80	44.33	59.06	58.62
300	90.66	89.69	87.53	87.03	90.21	90.15	59.00	57.92	50.75	50.02	57.93	57.75
500	90.28	89.36	87.96	87.53	**90.96**	89.95	58.28	57.23	52.44	51.31	**59.65**	58.21
1000	**90.79**	90.30	88.85	88.79	90.45	89.58	59.00	58.47	55.56	55.44	58.22	57.44
3000	90.49	89.89	**90.68**	90.34	90.17	89.45	58.77	58.23	**59.49**	58.52	58.30	57.48

Reuters-21578 veri kümesi için DFSS öznitelik seçim yöntemi, TF ve TF-IDF öznitelik ağırlıklandırma metodu, 3000 ve 500 öznitelik kullanan DVM sınıflandırıcı kombinasyonundan elde edilen en yüksek Mikro-F1 ve Makro-F1skoru sırasıyla 92.88, 65.55'dir. DVM sınıflandırıcı çoğu durumda KA'na göre daha başarılıdır. Ayrıca DVM sınıflandırıcıda en yüksek skorlar daha çok yüksek boyutlarda elde edilirken, KA sınıflandırıcıda daha çok düşük boyutlarda elde edilmiştir.

Tablo 7. DVM ile 20Newsgroup veri kümesinden elde edilen Micro-F1 ve Macro-F1 skorları (%)

	Micro-F1skorları (%)						**Makro-F1skorları (%)**					
	DFSS		**OR**		**CHI2**		**DFSS**		**OR**		**CHI2**	
Dimension	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**
50	97.75	98.03	98.61	98.56	98.27	98.41	96.03	96.11	97.54	97.17	96.93	96.84
100	98.21	98.32	98.67	98.53	98.46	**98.58**	96.82	96.67	97.64	97.09	**97.26**	97.19
300	98.52	**98.56**	**98.74**	98.58	98.13	98.31	**97.36**	97.18	**97.76**	97.18	96.70	96.67
500	98.21	98.32	98.69	98.60	97.90	98.24	96.83	96.70	97.67	97.22	96.30	96.54
1000	98.31	98.06	98.53	98.49	98.15	98.17	97.01	96.19	97.41	97.02	96.74	96.40
3000	98.29	97.94	98.29	97.92	98.17	97.79	96.97	95.96	96.96	95.93	96.77	95.64

Tablo 8. KA ile 20Newsgroup veri kümesinden elde edilen Micro-F1 ve Macro-F1 skorları (%)

	Mikro-F1skorları (%)						Makro-F1skorları (%)					
	DFSS		OR		CHI2		DFSS		OR		CHI2	
Dimension	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**
50	98.70	98.77	98.13	98.75	98.88	98.89	97.63	97.56	98.10	97.54	98.00	97.80
100	98.93	98.97	98.86	98.87	98.95	**98.99**	**98.10**	97.97	97.96	97.76	**98.14**	98.00
300	98.88	98.99	98.91	98.97	98.84	98.91	97.99	98.00	98.05	97.98	97.92	97.82
500	98.78	98.87	98.89	98.87	98.88	98.94	97.82	97.75	98.03	97.77	97.99	97.89
1000	98.86	98.97	98.93	**99.02**	98.86	98.94	97.96	97.97	98.09	98.06	97.95	97.89
3000	98.80	**98.99**	98.97	98.94	98.78	98.97	97.87	98.00	**98.15**	97.91	97.83	97.96

20Newsgroup veri kümesi için OR öznitelik seçim yöntemi, TF-IDF ve TF öznitelik ağırlıklandırma metodu, 1000 ve 3000 öznitelik kullanan KA sınıflandırıcı kombinasyonundan elde edilen en yüksek Mikro-F1 ve Makro-F1skoru sırasıyla 99.02, 98.15'tir.KA sınıflandırıcı çoğu durumda DVM'e göre daha başarılıdır. Ayrıca DVM sınıflandırıcıda en yüksek skorlar daha çok düşük boyutlarda elde edilirken, KA sınıflandırıcıda daha çok yüksek boyutlarda elde edilmiştir.

Tablo 9. DVM ile Enron1 veri kümesinden elde edilen Micro-F1 ve Macro-F1 skorları (%)

	Mikro-F1 skorları (%)						Makro-F1 skorları (%)					
	DFSS		OR		CHI2		DFSS		OR		CHI2	
Dimension	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**
50	92.56	92.56	91.05	91.09	94.67	94.10	84.20	84.15	82.20	82.30	88.35	87.25
100	93.81	93.77	92.82	93.55	95.45	95.45	86.60	86.50	85.10	86.30	89.95	89.95
300	95.45	95.31	95.14	95.49	95.66	95.56	89.80	89.45	89.15	89.90	90.35	90.05
500	95.49	95.10	95.42	95.38	95.73	95.77	89.70	88.80	89.55	89.50	90.35	90.40
1000	**95.66**	94.57	96.08	96.08	**96.84**	96.36	**90.05**	87.70	91.00	91.00	**92.65**	91.45
3000	95.49	94.60	**97.19**	96.08	96.29	95.80	89.65	87.75	**93.40**	90.95	91.35	90.20

Tablo 10. KA ile Enron1 veri kümesinden elde edilen Micro-F1 ve Macro-F1 skorları (%)

	Mikro-F1 skorları (%)						**Makro-F1 skorları (%)**					
	DFSS		**OR**		**CHI2**		**DFSS**		**OR**		**CHI2**	
Dimension	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**	**TF**	**TF-IDF**
50	91.85	91.85	89.97	89.89	94.42	94.42	83.25	83.25	80.00	80.35	87.80	87.80
100	91.96	91.96	92.74	92.52	95.24	**95.25**	83.05	83.05	84.95	84.50	**89.55**	**89.55**
300	**92.78**	**92.78**	93.55	**94.35**	94.31	94.31	**84.60**	**84.60**	86.35	**87.85**	87.65	87.65
500	91.92	91.92	93.30	92.93	92.96	92.97	82.75	82.75	85.80	84.90	84.69	84.70
1000	91.81	91.81	93.15	93.26	93.26	93.26	82.65	82.65	85.05	85.25	85.50	85.50
3000	92.07	92.07	94.24	93.44	94.17	93.44	83.00	82.90	87.10	85.35	87.10	85.85

Enron1 veri kümesi için OR öznitelik seçim yöntemi, TF öznitelik ağırlıklandırma metodu, 3000 öznitelik kullanan DVM sınıflandırıcı kombinasyonundan elde edilen en yüksek Mikro-F1 ve Makro-F1 skoru sırasıyla 97.19, 93.40'tır. KA sınıflandırıcı çoğu durumda DVM'e göre daha başarılıdır. Ayrıca DVM sınıflandırıcıda en yüksek skorlar daha çok düşük boyutlarda elde edilirken, KA sınıflandırıcıda daha çok yüksek boyutlarda elde edilmiştir.

Genel olarak, TF ağırlıklandırma metodu TF-IDF metoduna göre daha yüksek performans göstermiştir. Ayrıca, OR öznitelik seçim metodu ise DFSS ve OR metotlarına kıyasla daha yüksek performans göstermiştir. Ayrıca öznitelik boyutu arttıkça sınıflandırma performansı genel olarak artmıştır. Bazı durumlarda küçük boyutlar en yüksek performansı göstermiştir.

IV. SONUÇLAR

Bu çalışmada, yaygın olarak bilinen iki farklı öznitelik ağırlıklandırma metotlarının iyi bilinen iki sınıflandırıcı, üç farklı lokal öznitelik seçim yöntemi kullanılarak kapsamlı bir şekilde analiz edilmiştir. Bu analiz, metin sınıflandırma alanında yaygın olarak kullanılan üç farklı veri kümesi üzerinde gerçekleştirilmiştir. Deneylerde, ön işleme adımları olarak gereksiz kelimeleri kaldırma ve kök bulma metodu uygulanmıştır. Deneysel sonuçlar, DVM sınıflandırıcının Reuters-21578 ve Enron1 veri kümesinde en yüksek performansı gösterirken, KA sınıflandırıcısı ise 20Newsgroup veri kümesinde en yüksek performansı göstermiştir. Buradan DVM sınıflandırıcının dengesiz veri kümelerinde başarılı performans gösterirken, KA sınıflandırıcının ise dengeli veri kümesinde daha başarılı olduğu gözlenmiştir. Ayrıca, TF ağırlıklandırma metodu genel olarak lokal öznitelik seçim metotları üzerinde TF-IDF ağırlıklandırma metoduna göre daha yüksek performans sergilemiştir. Bunun sebebi ise Ters Doküman Frekansı (IDF)'nın lokal öznitelik seçim metotlarının performansına katkı sağlamadığı görülmektedir. IDF skoru bütün dokümanlar üzerinden hesaplanırken, lokal metotlar sadece ilgili sınıf için skor üretmektedir. Bu sebepten ötürü TF metodu daha başarılı performans sergilediği gözlenmiştir. Gelecekteki bir çalışma olarak, iyi bilinen bu ağırlıklandırma metotlarını global öznitelik seçim metotları üzerine uygulamak ve performans analizi yapmak olacaktır.

KAYNAKLAR

[1] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR), 34*(1), 1-47.

[2] Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management, 50*(1), 104-112.

[3] Parlak, B., & Uysal, A. K. (2020). The effects of globalisation techniques on feature selection for text classification. *Journal of Information Science*, 0165551520930897.

[4] Uysal, A. K. (2016). An improved global feature selection scheme for text classification. *Expert Systems with Applications*, *43*, 82-92.

[5] Parlak, B., & Uysal, A. K. (2021). A novel filter feature selection method for text classification: Extensive Feature Selector. *Journal of Information Science*, 0165551521991037.
[6] Rehman, A., Javed, K., Babri, H. A., & Asim, M. N. (2018). Selection of the most relevant terms based on a max-min ratio metric for text classification. *Expert Systems with Applications*, *114*, 78-96.
[7] Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, *3*(Mar), 1289-1305.
[8] Debole, F., & Sebastiani, F. (2004). Supervised term weighting for automated text categorization. *In Text mining and its applications*, 81-97.
[9] Özgür, A., Özgür, L., & Güngör, T. (2005). Text categorization with class-based and corpus-based keyword selection. *In International Symposium on Computer and Information Sciences*, 606-615.
[10] Taşcı, Ş., & Güngör, T. (2013). Comparison of text feature selection policies and using an adaptive framework. *Expert Systems with Applications*, *40*(12), 4871-4886.
[11] Uysal, A. K. (2018). On two-stage feature selection methods for text classification. *IEEE Access*, *6*, 43233-43251.
[12] Kou, G., Yang, P., Peng, Y., Xiao, F., Chen, Y., & Alsaadi, F. E. (2020). Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. *Applied Soft Computing*, *86*, 105836.
[13] Agnihotri, D., Verma, K., & Tripathi, P. (2017). Variable global feature selection scheme for automatic classification of text documents. *Expert Systems with Applications*, *81*, 268-281.
[14] Agnihotri, D., Verma, K., Tripathi, P., & Singh, B. K. (2019). Soft voting technique to improve the performance of global filter based feature selection in text corpus. *Applied Intelligence*, *49*(4), 1597-1619.
[15] Parlak, B., & Uysal, A. K. (2018). On feature weighting and selection for medical document classification. In *Developments and advances in intelligent systems and applications*, 269-282.
[16] Porter, M. F. (1980). An algorithm for suffix stripping. *Program*.
[17] Zong, W., Wu, F., Chu, L. K., & Sculli, D. (2015). A discriminative and semantic feature selection method for text categorization. *International Journal of Production Economics*, *165*, 215-222.
[18] Joachims, T. (1998, April). Text categorization with support vector machines: Learning with many relevant features. *In European conference on machine learning*, 137-142.
[19] Theodoridis, S., Pikrakis, A., Koutroumbas, K., & Cavouras, D. (2010). *Introduction to pattern recognition: a matlab approach*. Academic Press.
[20] Rehman, A., Javed, K., Babri, H. A., & Saeed, M. (2015). Relative discrimination criterion–A novel feature ranking method for text data. *Expert Systems with Applications*, *42*(7), 3670-3681.
[21] Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval*, 39, 234-265. Cambridge: Cambridge University Press.
[22] Parlak, B. (2022). Class- index corpus- index measure: A novel feature selection method for imbalanced text data. *Concurrency and Computation: Practice and Experience*, *34*(21), e7140.
[23] Parlak, B., & Uysal, A. K. (2020). On classification of abstracts obtained from medical journals. *Journal of Information Science*, *46*(5), 648-663.