METİN SINIFLANDIRMA İÇİN ÖZNİTELİK SEÇİMİ VE GLOBALLEŞTİRMENİN ETKİSİ

Bekir PARLAK

Doktora Tezi

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Bilimleri Bilim Dalı

Danışman: Doçent Doktor Alper Kürşat UYSAL

Eskişehir

Eskişehir Teknik Üniversitesi

Lisansüstü Eğitim Enstitüsü

Şubat 2021

ÖZET

METİN SINIFLANDIRMA İÇİN ÖZNİTELİK SEÇİMİ VE GLOBALLEŞTİRMENİN ETKİSİ

Bekir PARLAK

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Bilimleri Bilim Dalı

Eskişehir Teknik Üniversitesi, Lisansüstü Eğitim Enstitüsü, Şubat 2021

Danışman: Doçent Doktor Alper Kürşat UYSAL

Günümüzde internet hizmetlerinin artmasıyla her geçen gün metinsel veri üssel olarak artış göstermektedir. Bu metinlerin daha anlamlı ve kullanışlı hale gelebilmesi için metinlerin içeriklerine göre sınıflandırılması gerekmektedir. Bu sebeple otomatik metin sınıflandırma yaklaşımları oldukça önem kazanmıştır. Metin sınıflandırma yaklaşımlarının temel görevi metinleri içeriklerine göre sınıflara atamaktır. Metin içerikli dokümanları içeriklerine uygun sınıflara atayabilmek için birçok işlem adımları bulunmaktadır. Bunlar; öznitelik çıkartma, öznitelik seçimi, öznitelik ağırlıklandırma ve sınıflandırma işlemleridir. Metin sınıflandırma başarımını artırabilmek için bu aşamaların her biri ayrı bir öneme sahiptir. Ancak öznitelik seçimi son yıllardaki çalışmalarda daha popüler hale gelmiştir. Bu tez çalışmasında, metin sınıflandırma için kullanılan lokal öznitelik seçim metotları üzerinde farklı globalleştirme (maksimum, toplam, ağırlıklı toplam) teknikleri kullanılarak performans karşılaştırması yapılmış ve literatürde var olan güncel öznitelik seçim metotlarının performansından daha yüksek performansa sahip yeni bir öznitelik seçim metodu önerilmiştir. Bu amaçla farklı karakteristiğe sahip veri kümeleri üzerinde globalleştirme tekniklerinin başarımı nasıl değiştirdiğini gözlemlemiş olduk. Ayrıca, özniteliğin koleksiyon bazlı ve sınıf bazlı skorlarını göz önünde bulundurarak, Ayrıntılı Öznitelik Seçimi (EFS) adında yeni bir öznitelik seçim metodu önerilmiştir.

Anahtar Sözcükler: Metin sınıflandırma, Globalleştirme teknikleri, Boyut indirgeme, Öznitelik seçimi.

ABSTRACT

FEATURE SELECTION FOR TEXT CLASSIFICATION AND THE EFFECT OF GLOBALISATION

Bekir PARLAK

Department of Computer Engineering

Programme in Computer Science

Eskişehir Technical University, Institute of Graduate Programs, February 2021

Supervisor: Associate Professor Alper Kürşat UYSAL

Nowadays, with the increase of internet services, textual data increases exponentially with day by day. In order to make these texts more meaningful and useful, the texts should be classified according to their content. For this reason, automatic text classification approaches have gained importance. The main task of text classification approaches is to assign texts to classes according to their content. There are many steps to assign text-containing documents to classes suitable for their content. These are feature extraction, feature selection, feature weighting and classification processes. In order to increase the text classification performance, each of these stages has a special importance. However, feature selection has become more popular in recent years. In this thesis, performances were compared using different globalisation techniques (maximum, sum, weighted sum) on local feature selection methods used for text classification and a novel feature selection method with higher performance than the current feature selection methods in the literature are proposed. For this purpose, we have observed how globalisation techniques change performance on datasets with different characteristics. Also, considering the corpus-based and class-based scores of the feature, a new feature selection method is proposed, called Extensive Feature Selector(EFS).

Keywords: Text classification, Globalisation techniques, Dimension reduction, Feature selection.

TEŞEKKÜR

Tez çalışmam boyunca tecrübelerini ve bilgilerini her zaman paylaşan, desteğini üst düzeyde hissettiğim, gerek samimi tutumuyla olsun gerek yol göstericiliği ile akademik hayata farklı açılardan bakmama vesile olan danışmanım Sayın Doç. Dr. Alper Kürşat UYSAL hocama teşekkürü bir borç bilir ve katkılarından dolayı en derin teşekkürlerimi sunarım.

Çalışmaya dayanak oluşturan tez izleme süreçlerinde beni sabırla dinlediklerinden ve her çalışmamda farklı katkılar sunmaya çalışan, tez izleme jürisinde yer alan hocalarım Sayın Prof. Dr. Serkan GÜNAL'a ve Sayın Doç. Dr. Semih ERGİN'e verdikleri fikirler ve öneriler yardımıyla yeni bakış açıları kazanmamı sağladıklarından dolayı çok teşekkür ederim.

İster akademik hayatta ister günlük hayatta karşılaştığım her zorlukta yanımda olan Sevgili Eşim Esra PARLAK'a bu desteklerinden ötürü minnettarım.

Hayatımı varlığıyla daha da anlamlandıran, güleç yüzü, bitmeyen enerjisi ile üzerimdeki bütün yorgunluğumu bitiren Sevgili Yavrum İbrahim Eymen'i yuvamıza bahşeden Rabbime şükrümü ve kendisine sevgilerimi sunuyorum.

Babama beni hayatın zorluklarıyla karşılaştırıp bir taraftan yol gösterirken, bu zorlukların yanında beni her daim destekleyen Anneme yürekten teşekkür ediyorum. Bana olan özveri ve çabalarından dolayı minnettarım.

Bekir PARLAK

18/02/2021

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Bu tezin bana ait, özgün bir çalışma olduğunu; çalışmamın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamalarında bilimsel etik ve kurallara uygun davrandığımı; bu çalışma kapsamında elde edilen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi; bu çalışmanın Eskişehir Teknik Üniversitesi tarafından kullanılan "bilimsel intihal tespit programı"yla tarandığını ve hiçbir şekilde "intihal içermediğini" beyan ederim. Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçları kabul ettiğimi bildiririm.

Bekir PARLAK

İÇİNDEKİLER

Sayfa

TABLOLAR DİZİNİ

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

SİMGELER VE KISALTMALAR DİZİNİ

AVG	:	Ağırlıklı Toplam (Average)
BA	:	Dengeli Doğruluk Ölçütü (Balanced Accuracy Measure)
BCB	:	İkili-sınıf Dengeli (Binary-class Balanced)
BoW	:	Kelime Çantası (Bag-of-Words)
BCU	:	İkili-sınıf Dengesiz (Binary-class Unbalanced)
CDM	:	Sınıf Ayırıcı Ölçütü (Class Discriminating Measure)
CHI2	:	Ki-kare (Chi Square)
CMFS	:	Kapsamlı Öznitelik Seçim Ölçütü (Comprehensively Measure Feature Selection)
DF	:	Doküman Frekansı (Document Frequency)
DFS	:	Ayırt Edici Öznitelik Seçici (Distinguishing Feature Selector)
DFSS	:	Ayrımcı Öznitelik Seçimi (Discriminative Feature Selection)
DPM	:	Ayrımcı Güç Ölçütü (Discriminative Power Measure)
DT	:	Karar Ağacı (Decision Tree)
EFS	:	Kapsamlı Öznitelik Seçici (Extensive Feature Selector)
IDF	:	Ters Doküman Frekansı (Inverse Document Frequency)

KNN	:	K-En Yakın Komşular (K-Nearest Neighbors)
LÖS	:	Lokal Öznitelik Seçimi
MAX	:	Maksimum
MCB	:	Çoklu-sınıf Dengeli (Multi-class Balanced)
MCU	:	Çoklu-sınıf Dengesiz (Multi-class Unbalanced)
MMR	:	Mak-min Oranı (Max-min Ratio)
MNB	:	Çok Terimli Naive Bayes (Multinomial Naïve Bayes)
MS	:	Metin Sınıflandırma
NDM	:	Normalleştirilmiş Fark Ölçütü (Normalized Difference Measure)
OR	:	Olasılık Oranı (Odds Ratio)
ÖS	:	Öznitelik Seçimi
SUM	:	Toplam
SVM	:	Destek Vektör Makineleri (Support Vector Machine)
TF	:	Terim Frekansı (Term Frequency)
TF-IDF	:	Terim Frekansı - Ters Doküman Frekansı (Term Frequency - Inverse Document Frequency)

1. GİRİŞ

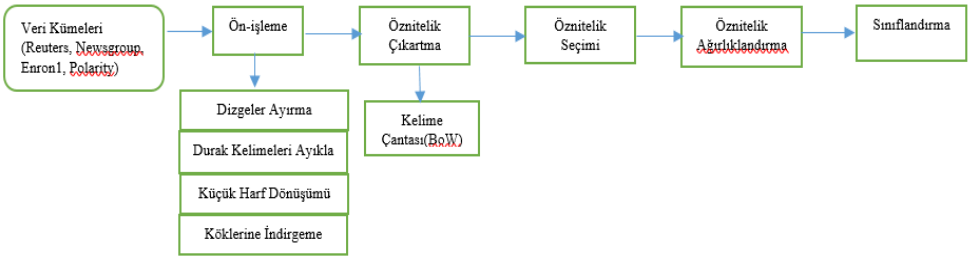
Günümüzde İnternet hizmetlerinin artmasıyla her geçen gün metinsel veri üssel olarak artış göstermektedir. Bu metinlerin daha anlamlı ve kullanışlı hale gelebilmesi için metinlerin alanlara göre sınıflandırılması gerekmektedir. Bu sebeple otomatik metin sınıflandırma yaklaşımları oldukça önem kazanmıştır. Metin sınıflandırma yaklaşımlarının temel görevi metinleri içeriklerine göre uygun sınıflara atamaktır.

Metin belgelerinin en küçük bileşeni metin belgelerinin sınıflandırılmasında önemli rol oynayan kelimedir (Agnihotri, Verma, & Tripathi, 2017; Uysal, 2016). MS geniş kapsamlı bir alan olmakla birlikte çeşitli alt çalışma alanları içermektedir. MS, özellikle bilgisayar bilimleri, bilgi erişimi ve bilgi bilimleri alanlarında çalışılmış, tür sınıflandırması (Onan, 2018), spam e-posta filtreleme (Bhowmick & Hazarika, 2018), SMS spam filtreleme (Sjarif, ve diğerleri, 2019), yazar tanıma (Zhang, Wu, Niu, & Ding, 2014), konu tespiti (Chang, Hsieh, Chen, & Hsu, 2017), tıbbi doküman sınıflandırması (Parlak & Uysal, 2020) ve web sayfalarının sınıflandırılması (Hashemi, 2020) gibi çeşitli alanlarda farklı çalışmalar yapılmıştır.

1.1. Metin Sınıflandırma

MS'nın önceden tanımlanmış sınıflara dokümanların atanma işlemi olduğu giriş bölümünde tanımlanmıştı. Genel olarak, MS Şekil 1.1'de de gösterildiği gibi önişleme, öznitelik çıkartma, ÖS, öznitelik ağırlıklandırma ve sınıflandırma olmak üzere birkaç adım içerir.

MS çalışmalarında gerçekleştirilen tüm aşamalar ilerleyen alt bölümlerde detaylı olarak anlatılacaktır.



Şekil 1.1. *Metin sınflandırma akış diyagramı*

Ön işleme aşamaları; dizgeciklere ayırma, durak kelimelerin çıkarılması, küçük harf dönüşümü, kök bulma gibi teknikler sınıflandırma performansını geliştirmek için

kullanılır (Uysal & Gunal, 2014a). Dokümanlar, ön işleme aşamasında belirtilen birkaç teknik uygulanarak öznitelik çıkartma aşamasına hazır hale getirilir.

İşlenmiş dizgecikler metinsel verilerin özniteliklerinin bir listesini oluşturmak için kullanılır. Öznitelik çıkartma adımını gerçekleştirmek için kelime çantası (BoW) ve vektör uzay modeli kullanılmaktadır (Forman, 2003). Bu aşamada, ham veri kümesif içerikleri sınıflandırma aşamasında işlenmek üzere sayısal verilere dönüştürülmektedir. Tekil öznitelikler elde edilir ve bütün dokümanlar öznitelik vektör uzayı adıyla bilinen sayısal değerlere dönüştürülmüş olur. Bu vektör uzayında, her terim bir boyut oluşturur.

Sıradan bir metin veri kümesi bile yüz binlerce öznitelikten oluşur. Yüksek boyutlu vektör uzayı sınıflandırma işlemlerini zorladığından, bu vektör uzayı büyüklüğünün azaltılması gerekmektedir. Bu işlemi yaparken, vektör alanına ayırt ediciliği yüksek öznitelikler eklemek için bir ÖS metodu gereklidir (Agnihotri, Verma, Tripathi, & Singh, 2019). Ayırt edici öznitelikler sınıflandırma başarımını doğrudan etkilediğinden, öznitelik seçimi halen MS alanında çalışan birçok araştırmacı için popüler bir çalışma alanı olarak önemini korumaktadır. Literatürde, MS için ÖS metodu ya da şeması üzerine birçok çalışma mevcuttur (Uysal & Gunal, 2012; Uysal, 2016; Agnihotri, Verma, & Tripathi, 2017; Agnihotri, Verma, Tripathi, & Singh, 2019).

Vektör uzayında; her tekil özniteliğin koleksiyon içindeki her bir doküman ile ilişkisi BoW (Aggarwal & Zhai, 2012) ile gösterilir. MS için bu ilişkileri gösteren değerlere ilgili özniteliğin ağırlığı, bu değerlerin hesaplanması işlemine ise öznitelik ağırlıklandırma denir.

Sınıflandırma kısmında ise sınıflandırılmamış dokümanlar önceden tanımlanmış sınıflara atanır. MS, etiketli dokümanlardan öğrenme modelleri oluşturmayı ve bu modelleri kullanarak etiketlenmemiş metin dokümanları üzerinde sınıflandırma yapmayı amaçlamaktadır. Metin dokümanları sayısal hale getirildiğinden, örüntü tanıma için kullanılabilen sınıflandırıcılar MS için de kullanılabilir. Veri kümelerinin özelliklerine göre uygun sınıflandırıcının seçimi, sınıflandırma performansını önemli oranda arttırabilmektedir (Parlak & Uysal, 2020).

1.2. Metin Sınıflandırmada Öznitelik Seçim Problemleri

MS çalışmalarında en önemli problemlerden bir tanesi de ön işleme aşaması bittikten sonra çıkarılan özniteliklerin dokümanlarla ilişkisini vektör uzay modelinde en

iyi biçimde temsil edebilmektir. Bu işlemi başarılı bir şekilde gerçekleştirmek için ayırt edici özniteliklere yüksek skorlar atanması, benzer şekilde ayırt edici olmayan özniteliklere düşük skorlar atanması gerekmektedir. Kullanılan öznitelik seçim metodunun atadığı skorlar ne kadar uygun ve mantıklı olursa, dokümanları temsil edecek vektörler de iyi olacağından MS başarımı da bir o kadar iyi olacaktır. Dolayısıyla etkili bir öznitelik seçim metodu özniteliklerin taşıdığı bilgileri kapsamlı bir şekilde analiz etmeli ve öznitelik seçme sürecini bu bilgilere bağlı olarak gerçekleştirmelidir. Öznitelik seçim sürecinin en önemli kısmı ilgili özniteliğin sınıf bazlı ve koleksiyon bazlı skorları göz önüne alınarak, sahip olduğu ayırt edici gücünü en iyi biçimde yansıtmasıdır. Literatürde, bugüne kadar önerilen öznitelik seçim metotlarının performansları incelendiğinde, özniteliğin yer aldığı sınıf bazlı ve koleksiyon bazlı değerlerinin öznitelik seçim sürecinde çok önemli olduğu söylenebilir.

Literatürde öznitelik seçim metodu ile ilgili son yıllarda yüksek performans gösteren metotlar önerilmiş olsa dahi, bu alanda halen daha yeni metotların öneriliyor olması; özniteliklerin sahip oldukları ayırt etme gücünü daha iyi yansıtabilen öznitelik seçim stratejileri olan yeni metotların geliştirilebileceğinin ispatıdır. Öznitelik seçim metodu için önerilen metotlar ne kadar güncel olursa olsun, her birinin skor hesaplama sürecinde yetersiz kaldığı, göz ardı ettiği veya sahip olduğu skor atama stratejisi yüzünden bazı olağan dışı senaryolardaki öznitelikler için makul skorlar üretemediği durumlar mevcuttur.

1.3. Amaçlar ve Katkılar

Bu tez çalışmasında bir önceki alt bölümde belirtilen öznitelik seçim problemlerine çeşitli çözümler önerilmiştir. Tezin literatüre kazandırdığı katkılar ve çözümler, aşağıdaki araştırma problemlerine cevap olma niteliği taşımaktadır:

I. Metin sınıflandırma performansını arttırmak için, öznitelikler ile yer aldıkları dokümanlara ait sınıflar arasındaki ilişkileri daha iyi yansıtabilen bir gösterim nasıl elde edilebilir?

II. Literatürdeki geleneksel ve güncel öznitelik seçim metotlarını önemli veri kümeleri üzerinde uygulayarak sınıflandırma başarımları üzerindeki etkisi nedir?

III. Lokal öznitelik seçim metotlarını farklı tekniklerle global hale getirerek, farklı karakteristiğe sahip veri kümeleri üzerinde sınıflandırma başarımı nasıl değişmektedir?

IV. Literatürdeki mevcut yöntemlerin performansından daha iyi performans gösterebilecek yeni bir öznitelik seçim metodu geliştirilebilir mi?

Yukarıda belirtilen araştırma problemlerine çözüm geliştirmek için hazırlanan bu tez çalışmasının ilk katkısı, literatürde mevcut olan öznitelik seçim metotlarının kullanımının metin sınıflandırma performansına etkisinin araştırılmasına yöneliktir. Bu tez çalışmasında güncel ve başarılı 9 farklı öznitelik seçim metodu farklı karakteristiğe sahip veri kümeleri üzerindeki performansları ayrıntılı bir şekilde analiz edilmiştir.

Tezin ikinci katkısı, lokal öznitelik seçim metotlarını farklı tekniklerle global hale getirerek, farklı karakteristiğe sahip veri kümeleri üzerinde sınıflandırma başarımı nasıl değiştiğini gözlemlemektir. Bu çalışmada 3 farklı lokal öznitelik metodu kullanılmıştır. Bu metotları global hale getirirken ise literatürde mevcut bulunan 3 farklı globalleştirme metodu kullanılmıştır. Ayrıca, kapsamlı bir analiz çalışması yapmak için 4 farklı karakteristiğe sahip veri kümeleri kullanılmıştır. Dört farklı karakteristiğe sahip veri kümelerinde iki farklı sınıflandırıcı kullanılarak gerçekleştirilen deneyler, altı farklı öznitelik boyutlarındaki etkilerini de görebilmek amacıyla farklı boyutlarda gerçekleştirilmiştir.

Tezin en önemli katkısı ise, metin sınıflandırması için yeni bir filtre temelli terim ve sınıf bazlı olasılıklardan oluşan, Kapsamlı Öznitelik Seçici (EFS) adıyla yeni bir öznitelik seçim metodu geliştirilmiştir. EFS kapsamlı bir şekilde geliştirilmiştir. Hem sınıf bazlı olasılıklar hem de koleksiyon bazlı olasılıklar dikkate alınarak özniteliğin daha ayırt edici hale gelmesini sağlayacak bir skor ataması yapılmıştır. Nihai skor, sınıf bazlı ve koleksiyon bazlı skorlar çarpılarak elde edilmiştir. EFS, literatürdeki öznitelik seçme yöntemlerinden farklı olarak birçok olasılığı göz önünde bulundurarak ayırt edici öznitelikleri seçmektedir. Önerilen metot, 9 farklı öznitelik seçim metodu ile karşılaştırılmıştır. Deneylerde 4 farklı veri kümesi kullanılmıştır. Sınıflandırma aşamasında ise 3 farklı sınıflandırıcı kullanılmıştır. Deneysel sonuçlar, EFS metodunun en yüksek skor açısından diğer metotlara göre daha iyi performans elde ettiğini göstermiştir.

1.4. Tez Organizasyonu

Bu tez toplam yedi bölümden oluşmaktadır. 2. Bölüm metin sınıflandırmanın temel bileşenlerini detaylı olarak açıklamaktadır. Konu ile alakalı olarak literatürde yapılan çalışmalar ve deneylerde kullanılan öznitelik seçim metotları 3. ve 4. Bölümlerde kısaca anlatılmıştır. 5. Bölümde ise bu tez çalışması kapsamında yapılan ilk deneysel çalışma olan globalleştirme tekniklerinin öznitelik seçim metodu üzerindeki etkileri kapsamlı bir şekilde analiz edilmiştir. 6. Bölümde ise Kapsamlı Öznitelik Seçici(EFS) metodu sunulmuştur. Ayrıca deneysel çalışmalarla EFS metodunun verimliliği gösterilmiştir. 7. Bölümde ise deneysel çalışmalardan elde edilen sonuçlarla ilgili genel yorumlar sunulmuştur.

2. METİN SINIFLANDIRMANIN TEMEL BİLEŞENLERİ

Metin sınıflandırma sürecinin içerdiği aşamalar bir önceki bölümde genel olarak ifade edilmişti. Bu bölümde ise, ilgili aşamalar detaylı olarak açıklanmış ve kullanılan yöntemler kabaca anlatılmıştır.

2.1. Ön-işleme

Genel olarak ön işleme aşaması, metin sınıflandırma çalışmalarında dört bölümden oluşur. Bunlar dizgelere ayırma, durak kelimeleri ayıklama, küçük harf dönüşümü, kelimeleri köklerine indirgemedir (Uysal & Gunal, 2014a). Ön işleme aşaması, dizgelere ayırma aşamasıyla başlar. Bu adımda, bir metin belgesi, kelimeler veya terimler olarak bilinen küçük parçalara dönüştürülür. Daha sonra, alfabetik olmayan belirli karakterler kaldırılır. Bir sonraki adım, tüm dizgeleri küçük harfe dönüştürme işlemidir. Bu adımdan sonra gerçekleştirilen iki adım vardır: gereksiz kelimeleri ayıklama ve kelimeleri köklerine indirgemedir. Her adım aşağıdaki alt bölümlerde ifade edilmiştir.

2.1.1. Dizgelere ayırma (Tokenization)

Dizgelere ayırma, bir cümleyi simge olarak ifade eden kelimelere veya diğer anlamlı parçalara bölme görevidir. Sözcükler veya tümcecikler genellikle boşluklar, noktalı virgül, virgül ve sınırlayıcı olarak tırnak işaretleri olan boşluklarla birbirinden ayrılır. Tipik olarak dizgelere ayırma, kelime düzeyinde gerçekleşir. İlk olarak, dizgeleri "\ r", "\ n", "\ t" gibi sınırlayıcı kümelerle belirtmek için basit bir java dizgeleme uygulanır ve ardından noktalama listesi ".,;: " '()?! & - # 0123456789 + / <> $ ^% [] / = " ”, alakasız dizgeleri kaldırmak için kullanılır.

Dizgelere ayırma dillere göre farklılık gösterebilir (Schütze, Manning, & Raghavan, 2008). ASCII olmayan karakterlerin kaldırılması metin dokümanlarının İngilizce dilinde belirtilmesi için yeterli olabilirken, Türkçe dilinde metin dokümanları için yeterli olmayabilir. Tablo 2.1, Türkçe ve İngilizce bir cümlenin dizge haline getirilmesine ilişkin bir örneği göstermektedir.

Tablo 2.1. *Örnek dizgeler*

Dil	Cümle	Dizgeler
Türkçe	Ali ata bak	Ali, ata, bak
İngilizce	I will go to Amasya	I, will, go, to, Amasya

2.1.2. Durak kelimeleri ayıklama (Stop-word removal)

Zamirler, çekimler, sıfatlar, zarflar ve edatlar olan kelimelere durak kelimesi denir. Durak kelimeleri metnin kapsamıyla ilgili değildir ve sınıflandırmadan önce kaldırılır. Metin sınıflandırmasında artan sistem doğruluğu nedeniyle durak kelimelerinin kaldırılması önemli bir adımdır. Bu süreç, İngilizce dilinde "a", "an", "the", "above" gibi bazı yaygın sözcüklerin kaldırılmasını içerir. Türkçede buna örnek "ne", "nerede", "ama", "böylece" vb. kelimelerdir. Durak kelimeleri, kök bulma aşamasında olduğu gibi çalışılan dile göre spesifiktir. Örnek durak kelimeleri Tablo 2.2'de Türkçe ve İngilizce dilleri için gösterilmiştir.

Tablo 2.2. *Örnek durak kelimeler*

Dil	Biçim
Türkçe	acaba, ama, bana, bazen, çok, çünkü, diğer, elbette, fakat, hangi
İngilizce	above, again, best, better, can, currently, definitely, every, has

2.1.3. Küçük harf dönüşümü (Lowercase conversion)

Küçük harf dönüşümü, metin sınıflandırmadaki önemli ön-işleme adımlarından biridir. Kelimelerin büyük ve küçük hallerini ayrı ayrı ele alırsak, aynı kelime için farklı öznitelikler kullanmış oluruz. Bu nedenle, tüm büyük harfli karakterler, kök bulma adımından önce küçük harf formlarına dönüştürülür. Küçük harf dönüşümü, toplam öznitelik sayısını azaltır. Küçük harf dönüşümü, Türkçe ve İngilizce dilinin özelliğine bağlı olarak bazı durumlarda değişiklik gösterebilir. Aynı karakterlerin küçük harfe dönüştürülmesine bir örnek Türkçe ve İngilizce için Tablo 2.3'te gösterilmektedir.

Tablo 2.3. *Örnek Dizgeler*

Dil	Orijinal form	Küçük harf dönüşümü
Türkçe	U, I	u, ı
İngilizce	U, I	u, i

2.1.4. Kelimeleri köklerine indirgeme (Stemming)

Kök bulma işlemi, kelime hakkında gramer veya sözcük bilgisi sunan son ekleri kaldırarak, morfolojik kök olan bir kelime veya terimin kökünü elde etmek için uygulanır. Türkçe'nin sondan eklemeli bir dil olması ve bir kelimenin kökünden yüzlerce veya binlerce farklı kelime türetilebilmesi nedeniyle, kök bulma, metin sınıflandırması

yapmadan önce önemli bir adımdır. Kök bulma algoritmaları çalışılan dile göre değişir. Türkçe dili için sabit önek algoritması (fixed prefixed stemming=FPS) (Can, ve diğerleri, 2008) ve Zemberek (Akın & Akın, 2007) adlı dizin tabanlı bir kök bulma algoritması geliştirilmiştir. Öte yandan, Porter (Porter, 1980) kök bulma algoritması, araştırmacılar tarafından İngilizce dili için yaygın olarak kullanılmaktadır. FPS, metin dokümanlarında ilk "n" karakterini tanıyan sözde bir kök bulma algoritmasıdır. Ancak, Zemberek genel amaçlı bir açık kaynak doğal dil işleme(NLP) araç takımıdır ve kök bulma için oluşturulmuş bir son-ek sözlüğü içerir. Kök bulmaya bir örnek, sırasıyla hem Türkçe hem de İngilizce dilleri için Tablo 2.4'te gösterilmektedir.

Tablo 2.4. *Bazı kelimelerin kök hali*

Dil	Orijinal form	Köküne indirgenmiş hali
Türkçe	kararlaştırmak	karar
İngilizce	decison	decide

2.2. Öznitelik Çıkartma

MS alanında, öznitelikler kelime çantası (Bag-of-words=Bow) olarak bilinen teknikle gösterilir. Kelimeler, bilgi erişim araştırma alanındaki dokümanları sınıflandırmak için de kullanılmaktadır. Her farklı kelime, Bow yaklaşımında kelimenin dokümanda geçme sayısı ile ilişkilendirilen belli ağırlığa sahip bir özelliğe karşılık gelir. Sonuç olarak, bir doküman vektör uzay modeli gibi çok boyutlu bir öznitelik vektörü ile temsil edilir (Salton, Wong, & Yang, 1975).

2.2.1. Öznitelik ağırlıklandırma

Metin sınıflandırma çalışmalarında dokümanlar vektör olarak temsil edilir. Bir dokümandaki her sözcüğün terim frekansı, sözcüğün dağılımına bağlı olan ve dokümandaki sözcüğün önemini ifade eden bir ağırlıktır.

Metin veri kümesinde K dokümanları, k ise tek bir dokümanı temsil ettiğini varsayarsak, $f_{k,d}$ ya da TF “d” dokümanında “k” teriminin görünme sayısıdır, t_k k terimini temsil eder ve metin dokümanlarında k_i defa geçer.

$$TF(k,d) = \begin{cases} d \text{ dokümanında } k \text{ teriminin geçme sayısı}, & \text{Eğer k terimi d dokümanında geçiyorsa} \\ 0, & \text{Diğer durumda} \end{cases} \quad \textbf{(2.1)}$$

TF-IDF ağırlıklandırma, sınıflandırma sisteminin başarısını veya başarısızlığını belirleyen önemli bir adımdır (Salton & Buckley, 1988). TF faktörü ve IDF faktörü, bir dokümanda ilgili terimin önemini belirler. Her kelime veya terimin ters doküman frekansı, dokümandaki her kelimenin dağılımına bağlı bir ağırlıktır. TF-IDF tekniği, bir terimin ağırlığını belirlemek için hem TF hem de IDF değerini kullanır. TF-IDF terim ağırlıklandırma tekniği, metin sınıflandırma alanında yaygın olarak kullanılmaktadır ve diğer terim ağırlıklandırma şemaları bu şemanın türevleridir.

Sezgisel olarak, TF-IDF yöntemi, terimin belirli bir metin dokümanıyla ne kadar alakalı olduğunu belirler. Ters doküman frekansı formülü aşağıdaki gibidir:

$$IDF(t_k) = log\frac{K}{k_i} \quad \textbf{(2.2)}$$

TF-IDF formülü, metin dokümanındaki bir terimin alakalı veya anlamlı değerini ölçmek için kullanılır. TF-IDF formülü aşağıdaki gibidir:

$$\text{TF-IDF(k,d)} = \text{TF(k,d)} * \text{IDF}(t_k) = f_{k,d} * \log\frac{K}{k_i} \quad \textbf{(2.3)}$$

2.3. Öznitelik Seçimi

Öznitelik seçimi, boyutu azaltmak ve metin sınıflandırma alanındaki önemsiz öznitelikleri kaldırmak için çok önemli bir adımdır. Bu adım, bazı sezgisel kurallara göre tüm öznitelik kümesinden bir alt küme seçer.

Öznitelik seçimi, metinsel verilerde gereksiz kelimelerin atılıp, kelimelerin kökleri bulunduktan sonra yapılan işlemdir. Bu ön işlem adımları, metinlerin dilleri bazında farklılık göstermektedir. Özniteliklerin belirlenmesi aşamasında, bu alandaki çalışmaların hemen hemen hepsinde kullanılan kelime çantası (bag-of-words) yaklaşımı kullanılmıştır. Bu yaklaşımda, veri kümesindeki her bir terim ayrı bir özniteliktir ve terimlerin dokümanda görülme sırası dikkate alınmaz. Öznitelikler, veri kümesindeki tekil terimlerin birleşiminden oluşmaktadır. Bu yüzden, az sayıda doküman barındıran bir veri kümesindeki öznitelik sayısı dahi binler veya yüzbinler ile ifade edilebilir. Öznitelik

boyutunun yüksek olması başarımı kötü etkileyebilmekte ve işlem zamanını artırabilmektedir. Bu sebeple metin sınıflandırma problemlerinin genelinde öznitelik seçimi üzerinde yoğunlaşılmaktadır. Öznitelik seçiminin başarılı bir şekilde yapılmasının işlem zamanını düşürebileceği ve/veya başarımı artırabileceği bilinmektedir. Bu tez çalışması kapsamında literatürde önerilmiş birçok geleneksel ve güncel yöntem kullanılmıştır.

2.4. Sınıflandırma

Genel olarak, metin sınıflandırması, kategorize edilmemiş dokümanların önceden tanımlanmış kategorilere göre sınıflandırılması işlemidir. Makine öğrenimi açısından, metin sınıflandırmasının amacı, etiketli dokümanlardan sınıflandırıcıları öğrenmek ve etiketlenmemiş metin dokümanları üzerinde sınıflandırmayı tamamlamaktır. Metin sınıflandırma alanında yaygın olarak kullanılan sınıflandırıcılardan bazıları Çok terimli naive bayes (MNB), Destek vektör makineleri (SVM), Karar ağaçları (DT), K-en yakın komşular (KNN).

2.4.1. Çok terimli naive bayes (Multinomial naive bayes-MNB)

Multinomial Naive Bayes, daha çok metin dokümanları için tasarlanmış ve Naive Bayes'in özel bir versiyonudur. Naive Bayes bir dokümanı belirli kelimelerin varlığı ve yokluğu olarak modellerken, çok terimli naive bayes, kelime sayılarını açık bir şekilde modeller ve arka plandaki hesaplamaları ele almak için ayarlar.

Sınıf seti C ve kelime sözlüğümüzün boyutu ise N ile temsil edilmektedir. Multinomial Naive Bayes, en yüksek olasılığa $P(c|t_i)$ sahip sınıfa Bayes kuralı kullanılarak t_i dokümanını atar:

$$P(c|t_i) = \frac{P(c)\cdot P(t_i|c)}{P(t_i)}, \quad c \in C \tag{2.4}$$

P (c), c sınıfına ait doküman sayısının toplam doküman sayısına bölünmesiyle elde edilebilir. $P(t_i|c)$ c sınıfında t_i terimini içeren dokümanı elde etme olasılığıdır ve aşağıdaki şekilde hesaplanır:

$$P(t_i|c) = (\sum_n f_{ni})! \prod_n \frac{P(w_n|c)^{f_{ni}}}{f_{ni}!}, \tag{2.5}$$

f_{ni}, test dokümanı t_i'deki n kelimesinin sayısıdır ve $P(w_n|c)$ c sınıfındaki n kelimesinin olasılığıdır. İkinci olasılık eğitim dokümanlarından şu şekilde tahmin edilir:

$$\mathrm{P}(w_n|\mathrm{c}) = \frac{1 + F_{nc}}{N + \sum_{x=1}^{N} F_{xc}}\,, \tag{2.6}$$

F_{xc} c sınıfına ait tüm eğitim dokümanlarındaki x kelimesinin sayısıdır ve Laplace tahmincisi, sıfır frekans probleminden kaçınmak için her kelimenin sayısını bir ile başlatmak (McCallum & Nigam, 1998) için kullanılır. Normalleştirme faktörü $\mathrm{P}(t_i)$ kullanılarak hesaplanabilir:

$$\mathrm{P}(t_i) = \sum_{k=1}^{|C|} P(k) \cdot \mathrm{P}(t_i|\mathrm{k})\,, \tag{2.7}$$

Hesaplama açısından maliyetli terimlerin $(\sum_n f_{ni})!$ ve $\prod_n f_{ni}!$ sonuçlarda herhangi bir değişiklik yapılmadan silinebilir, çünkü hiçbiri c sınıfına bağlı değildir ve şu şekilde yazılabilir:

$$\mathrm{P}(t_i|\mathrm{c}) = \alpha \prod_n P(w_n|c)^{f_{ni}}, \tag{2.8}$$

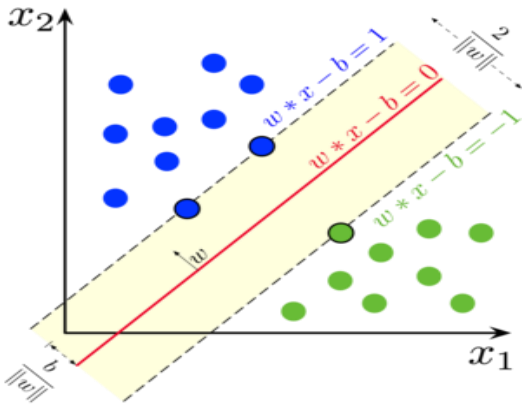
burada α, normalizasyon adımı nedeniyle kullanılan sabittir.

2.4.2. Destek vektör makineleri (Support vector machines-SVM)

SVM, MS çalışmalarında en verimli sınıflandırıcılardan biridir. SVM sınıflandırıcısının doğrusal ve doğrusal olmayan iki versiyonu vardır (Joachims, 1998). Bu çalışmada SVM'nin lineer versiyonunu kullandık. SVM sınıflandırıcısının önemli konusu marjin kavramıdır. Lineer çekirdekli SVM sınıflandırıcısı için LibSVM kütüphanesi kullanılmıştır. SVM sınıfları ayırmak için hiper düzlemleri kullanır. Her hiper düzlem, yönü (w) ve uzaydaki tam konumu (w_0) ile karakterize edilir. Böylece, doğrusal bir sınıflandırıcı en basit haliyle şu şekilde tanımlanabilir:

$$w^T x + w_0 = 0. \tag{2.9}$$

Daha sonra iki sınıf için ayrımı sağlayan $w^T x + w_0 = 1$ ve $w^T x + w_0 = -1$ hiper düzlemler arasındaki bölge marj olarak işaretlenir. Şekil 2.1'de görüldüğü üzere iki sınıftan (X_1 ve X_2) örneklerle eğitilmiş bir SVM için maks-marj hiper düzlem ve marjlar vardır.



Şekil 2.1. *Maks-marj hiper düzlem ve marjlar gösterimi*

Marj genişliği $2/||w||$ değerine eşittir. SVM algoritmasının temel amacı, mümkün olan en yüksek marjı elde etmektir. Marjın maksimize edilmesi;

$$J(w, w_0, \varepsilon) = \frac{1}{2}||w||^2 + K\sum_{i=1}^{N} \varepsilon_i \tag{2.10}$$

Buna göre;

$$\begin{gathered} w^T x + w_0 \geq 1 - \varepsilon_i \quad eğer \;\; x_i \in \; c_1 \\ w^T x + w_0 \leq -1 + \varepsilon_i \quad eğer \;\; x_i \in \; c_2 \\ \varepsilon_i \geq 0. \end{gathered} \tag{2.11}$$

Burada K kullanıcı tanımlı bir sabittir ve ε ise marj hatasıdır. Eğer bir sınıfa ait veriler hiper düzlemin yanlış tarafında ise marj hatası oluşur. Bu nedenle maliyeti en aza indirmek, büyük bir marj ile az sayıda marj hatası arasında bir değişim ile ilgili bir konudur. Bu optimizasyon probleminin çözümü eğitim özniteliklerinin ağırlıklı ortalamasıdır ve şu şekilde elde edilir:

$$w = \sum_{i=1}^{N} \lambda_i \, y_i x_i \tag{2.12}$$

Burada, λ_i optimizasyon probleminin Lagrange çarpanı ve y_i ise bir sınıfın etiketidir.

2.4.3. Karar ağaçları (Decision tree-DT)

DT lineer(doğrusal) olmayan bir sınıflandırıcıdır. DT'ler, sınıfların uygun bir sınıf tespit edilene kadar art arda reddedildiği çok aşamalı karar sistemleridir. J. R. Quinlan (Quinlan, 1986) tarafından geliştirilmiştir. Amacı, öznitellikler sınıflara karşılık gelen

farklı bölgelere ayrılır. İkili sınıflandırma ağacı en yaygın kullanılan DT türüdür. Bu nedenle, bir sınıfa Evet / Hayır karar dizisi aracılığıyla bilinmeyen bir öznitelik vektörü atanır.

De Mántaras (De Mántaras, 1991) ağaçta öznitelik seçimi için istatistiksel modelleme getirmiştir. p pozitiflik ve n negatiflik içeren bir eğitim seti için:

$$H\left(\frac{p}{n+p},\frac{n}{n+p}\right) = -(\frac{p}{n+p}log_2\frac{p}{n+p} + \frac{n}{n+p}log_2\frac{n}{n+p}) \quad \textbf{(2.13)}$$

Buna göre eğer k farklı değer ile öznitelik A seçilirse ve eğitim verisi E, $\{E_1, E_2, \dots, E_k\}$ gibi alt kümelere bölünürse o zaman beklenen entropi (BE):

$$BE(A) = \sum_{i=1}^{k}\frac{p_i + n_i}{p+n}H(\frac{p_i}{p_i+n_i},\frac{n_i}{p_i+n_i}) \quad \textbf{(2.14)}$$

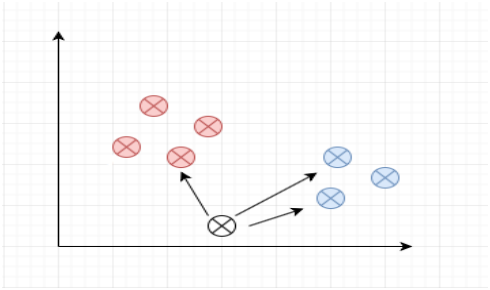
Bu öznitelik için bilgi kazanımı (BK) ise şu şekilde belirtilir:

$$A(BK) = H\left(\frac{p}{n+p},\frac{n}{n+p}\right) - BE(A) \quad \textbf{(2.15)}$$

2.4.4. K-en yakın komşular (K-nearest neighbors-KNN)

KNN sınıflandırıcı, etkili ve basit bir öğrenme algoritmasıdır (Tan, 2006). MS alanında, eğitim dokümanlarının en yakın komşularının kategori bilgilerine göre test dokümanlarının sınıfını tahmin etmeyi amaçlamaktadır. Test dokümanları ve komşulara olan mesafe, farklı ölçülerle hesaplanabilir. Farklı değerler arasından k belirlenir. KNN sınıflandırıcısı, tüm veri kümeleri için k değerini 7 olarak ayarlanıp deneyler gerçekleştirilmiştir.

Bu metot metin sınıflandırma alanında yaygın olarak kullanılan bir sınıflandırıcıdır. Bu yaklaşımın çalışma prensibi, bir test belgesi x verildiğinde, eğitim setindeki tüm belgeler arasında x'in en yakın k komşusunu bulmak ve k adaylarının sınıfına göre kategori adaylarını puanlamaktır. x dokümanın ve her bir komşunun benzerliği, komşu belgelerin kategorisinin skoru olabilir. Şekil 2.2'de örnek bir KNN çizimi verilmiştir.



Şekil 2.2. *Örnek KNN gösterimi*

Komşu hesaplama için Euclidean (Öklid), Manhattan ve Minkowski gibi uzaklık hesaplama yöntemlerinden birinden faydalanır. Yaklaşım belgenin komşu skorlarını hesapladıktan sonra bunlardan skoru en yüksek k tanesini alır.

3. İLGİLİ ÇALIŞMALAR

ÖS yöntemlerinin amacı, ilgili ayırt edici özniteliğe yüksek bir skor atamaktır. MS dışında farklı araştırma alanlarına ilişkin ÖS hakkında da birçok çalışma var (Guyon & Elisseeff, 2003). Metinsel veriler binlerce öznitelik içerdiğinden, genel örüntü tanıma sorunları için geliştirilen ÖS yöntemlerinden bazıları MS'da etkili veya başarılı olmayabilir. Bu nedenle ÖS, MS alanındaki sınıflandırıcıların yanı sıra en önemli adımlardan biridir.

Forman (Forman, 2003) birçok farklı ÖS yöntemini analiz etmiş ve lokal politikaları dikkate almıştır. Ancak, Forman oldukça dengesiz veri kümelerinde özellikle başarılı olan bi-normal seperation yöntemini geliştirmiştir. Debole ve Sebastiani (Debole & Sebastiani, 2004) bir özniteliğin entropisini kullanarak bilgi kazanma skorunu normalleştirmeye dayanan bir yaklaşım önermiştir. Bu çalışmada, hem dengeli hem de dengesiz olanlar da dahil olmak üzere birçok veri kümesi kullanılmış ve sınıflandırma aşamasında SVM kullanılmıştır. Özgür ve arkadaşları (Özgür, Özgür, & Güngör, 2005) lokal ve global ÖS politikalarını SVM sınıflandırıcısını kullanarak karşılaştırmışlardır. Ancak, tek bir veri kümesi kullanılır ve ÖS yaklaşımlarının rolü ayrıntılı olarak analiz edilmiştir. Shang ve arkadaşları, Gini endeksinin MS alanındaki ÖS için nasıl verimli bir şekilde kullanılabileceğini araştırdı (Shang, ve diğerleri, 2007). Pinheiro ve arkadaşları (Pinheiro, Cavalcanti, Correa, & Ren, 2012) ÖS metodu olarak yeni bir filtre tabanlı ALOFT (at least one feature) adında yöntem önermişlerdir. ALOFT yöntemi, her dokümanın eğitim setindeki en az bir öznitelik ile gösterilmesini garanti eder. Ayrıca, iki sınıflandırıcılı üç genel veri kümesinde oldukça başarılıdır. Taşcı ve Güngör (Taşcı & Güngör, 2013) MS'da kullanılan ÖS politikalarını karşılaştırdı. ÖS politikalarının değerlendirilmesine ek olarak Forman tarafından geliştirilen Acc2 yönteminin ileri formları olarak kabul edilebilecek bazı yeni ÖS yöntemleri önerilmiştir (Forman, 2003). Bu yöntemler, ÖS işlemi sırasında hem olumsuz(negatif) öznitelikleri hem de olumlu(positif) öznitelikleri dikkate aldıkları için iki taraflı ÖS yöntemleridir. Bu yöntemler aynı zamanda koleksiyondaki bir özniteliğin bir sınıfta varlığına ve yokluğundaki dağılımına dayanır. Pinheiro ve arkadaşları MS için iki filtre temelli ÖS yöntemi önerdi (Pinheiro, Cavalcanti, & Ren, 2015) Önerilen yöntemler, üç öznitelik değerlendirme fonksiyonu ile Naive Bayes sınıflandırıcısını kullanan dört kıyaslama veri kümesinde çok başarılıdır. Sonuç olarak, lokal öznitelik seçme yöntemleri literatürde

farklı şekillerde globalleştirilmiştir. Farklı globalleşme tekniklerinin amacı performansı arttırmaktır.

MS alanında filtre tabanlı ÖS metotları, ÖS şemaları ve filtre tabanlı ile sarıcı tabanlı tekniklerin birleşimini içeren iki aşamalı şemalar vardır. Uysal ve Gunal (Uysal & Gunal, 2012) DFS adlı yeni bir ÖS metodu önermiştir. DFS ayırtedici özniteliklere yüksek skorlar atarken, bazı önceden tanımlanmış ölçütlere göre alakasız özniteliklere düşük skorlar atamaktadır. Farklı veri kümeleri, sınıflandırma algoritmaları ve başarı ölçütleri kullanılarak DFS'nin etkinliği araştırılmış ve iyi bilinen filtre teknikleriyle karşılaştırılmıştır. Kapsamlı bir deneysel analizin sonuçları, DFS'nin doğruluk, boyut küçültme oranı ve işlem süresi açısından oldukça başarılı bir performans sunduğunu açıkça göstermiştir. Yang ve arkadaşları (Yang, Liu, Zhu, Liu, & Zhang, 2012) CMFS adlı yeni bir ÖS tekniği sundular. CMFS, sınıf içi ve sınıflar arası dağılımları dikkate alarak bir özniteliğin önemini hesaplar. Deneysel çalışmalar, CMFS'nin iki sınıflandırma algoritması kullanarak üç önemli veri kümesinde altı ÖS tekniğinden daha başarılı olduğunu göstermiştir. Zong ve arkadaşları (Zong, Wu, Chu, & Sculli, 2015) hem terimler ve sınıflar arasındaki korelasyonu hem de terimler ve dokümanlar arasındaki anlamsal benzerliği dikkate alan ayrımcı öznitelik seçimi (DFSS) adlı yeni bir ÖS metodu önermiştir. Deneysel çalışmalar, DFSS'nin performansının diğer ÖS teknikleri ile rekabetçi olduğunu göstermiştir. Rehman ve arkadaşları, MS alanı için birkaç ÖS yöntemi önermiştir (Rehman, Javed, Babri, & Saeed, 2015; Rehman, Javed, & Babri, 2017; Rehman, Javed, Babri, & Asim, 2018). Rehman ve arkadaşları (Rehman, Javed, Babri, & Saeed, 2015) öznitelik sayılarını dikkate alan RDC adlı yeni bir ÖS yöntemi önermişlerdir. RDC, iki sınıflandırıcı kullanılarak dört genel veri kümesinde dört farklı ÖS metodundan daha başarılı performans göstermiştir. Diğer bir çalışmada (Rehman, Javed, & Babri, 2017), bağıl doküman frekanslarını dikkate alan NDM adlı yeni bir ÖS yöntemi önerdiler. NDM, bazı sınıflandırıcılar kullanarak farklı veri kümelerindeki geleneksel ÖS algoritmalarından daha başarılı performans göstermiştir. Bu çalışmaya ek olarak, Rehman ve arkadaşları yüksek derecede dengesiz kategoriler içeren veri kümelerinde bile daha bilgilendirici özniteliklerin daha küçük alt kümelerini seçen MMR (Rehman, Javed, Babri, & Asim, 2018) adı verilen farklı bir yeni ÖS tekniği önerdiler. Kim ve Zzang (Kim & Zzang, 2019), bağıl doküman frekanslarını dikkate alarak Trigonometric Comparison Measure (TCM) adı verilen yeni bir filtre tabanlı ÖS tekniği önerdiler. Yöntem, sınıflandırma performansını iyileştirmek ve bilgilendirici terimleri

belirlemek için gerçek pozitif oran ve yanlış pozitif oran kullanır. Deneylerde yöntem, iki sınıflandırıcılı on veri kümesinde sekiz başarılı filtre tabanlı ÖS yöntemi ile karşılaştırılmıştır. Deneysel çalışmalar TCM'nin MS için diğer yöntemlerden daha başarılı olduğunu göstermiştir. Pinheiro ve arkadaşları (Pinheiro, Cavalcanti, Correa, & Ren, 2012), MS alanının spesifik özelliklerini dikkate alan ALOFT (At Least One FeaTure) adlı filtre tabanlı bir ÖS yöntemi önermişlerdir. Yöntem, her dokümanın eğitim setinde en az bir özniteliğinin olmasını garanti eder. Ayrıca ALOFT, veri güdümlü bir şekilde optimum öznitelikleri bulabilir. Deneysel çalışmalar, yöntemin performansının iki sınıflandırıcılı üç genel veri kümesinde beş ÖS yönteminden daha başarılı olduğunu göstermiştir. ALOFT'a ek olarak, Pinheiro ve arkadaşları (Pinheiro, Cavalcanti, & Ren, 2015) MS çalışma alanı için Maximum f Features per Document (MFD) ve Maximum f Features per Document-Reduced (MFDR) olarak adlandırılan iki filtre tabanlı ÖS yöntemi önermişlerdir. Bu yöntemler, her dokümanın son öznitelik kümesine katkıda bulunduğunu garanti eder. Deneysel çalışmalar, önerilen yöntemlerin Naive Bayes sınıflandırıcısı kullanılarak dört karşılaştırma veri kümesinde üç ÖS yöntemiyle Değişken Sıralama (Variable Ranking) ve ALOFT algoritmalarından daha iyi performans verdiğini göstermiştir.

Filtre tabanlı ÖS yöntemlerine ek olarak, literatürde birkaç ÖS şeması vardır (Uysal, 2016; Agnihotri, Verma, & Tripathi, 2017; Agnihotri, Verma, Tripathi, & Singh, 2019). Filtre tabanlı ÖS şemalarında, özniteliklere ayırt edici güçlerine göre bir skor atanır. Bu öznitelikler daha sonra skorlarına göre azalan sırada sıralanır. Son aşamada, en iyi N adet öznitelik seçilir. Uysal (Uysal, 2016) her sınıfı eşit olarak temsil eden daha bilgilendirici öznitelik kümesi elde etmek için standart ÖS şemasını modifiye ederek geliştirilmiş bir global ÖS şeması (IGFSS) önermiştir. Yerel bir ÖS yöntemi kullanılarak, sınıflardaki özniteliklerin ayırt edici güçleri dikkate alınarak her özniteliğe bir etiket atanır. Bazı sınıflar seçilen öznitelikler tarafından iyi bir şekilde temsil edilirken, bazı sınıflar iyi temsil edilmeyebilir. Bu durumu düzeltmek için IGFSS, her sınıfı temsil eden eşit sayıda öznitelik seçerek öznitelik kümesini oluşturur. Bu çalışmaya dayanarak, Agnihotri ve arkadaşları (Agnihotri, Verma, & Tripathi, 2017) tarafından öznitelik kümesi oluşturma sürecinde her sınıfı temsil eden değişken sayıda öznitelik kullanan yeni bir şema önerilmiştir. Eşit sayıda öznitelik kullanmanın çok sınıflı dengesiz veri kümelerinde sorun yarattığını belirtmişlerdir. Eğer her sınıftan eşit sayıda öznitelik alınırsa, sınıflardaki bazı önemli öznitelikler seçilemeyebilir. Bu sorunu çözmek için,

sınıflardaki özniteliklerin dağılımına bağlı olarak, her sınıftan değişken sayıda öznitelik seçilmiştir. Agnihotri ve arkadaşları (Agnihotri, Verma, Tripathi, & Singh, 2019) IGFSS'nin bazı eksikliklerinin üstesinden gelmek için yeni bir Soft Voting Tekniğini (SVT) önerdi. SVT, Odds Ratio (OR), Korelasyon Katsayısı (CC) ve GSS Katsayısı (GSS) gibi üç yöntemin ağırlıklı ortalamasını kullanmıştır. SVT, beş kriter veri kümesine uygulanan dört sınıflandırıcı ile karşılaştırıldı. Deneysel çalışmalar SVT'nin standart tekniklere kıyasla sınıflandırma performansında önemli bir iyileşme elde ettiğini göstermiştir.

Ayrıca, filtre ve sarmalama tekniklerini birleştirerek birçok hibrit yöntem önerilmiştir (Ghareb, Bakar, & Hamdan, 2016; Uysal, 2018; Uysal & Gunal, 2014b). Uysal ve Gunal, ÖS yöntemleri ve öznitelik dönüşüm aşamalarından oluşan GALSF adlı bir yaklaşım önerdi (Uysal & Gunal, 2014b). İlk olarak, özniteliklerin bir alt kümesini seçmek için filtre tabanlı yöntemler uygulanmıştır. İkinci olarak, genetik algoritma ile güçlendirilmiş gizli anlamsal indeksleme (LSI=Latent Semantic Indexing) kullanılmıştır. Deneysel çalışmalar, GALSF'nin, farklı öznitelik boyutları için genel veri kümelerinde hem LSI hem de filtre tabanlı ÖS tekniklerinden daha iyi performans elde ettiğini göstermiştir. Ghareb ve arkadaşları (Ghareb, Bakar, & Hamdan, 2016) genetik algoritmaya (GA) dayanan hibrit bir ÖS tekniği önerdi. Teknik, yüksek öznitelik boyutunu çözmek ve sınıflandırma performansını artırmak için hem filtre tabanlı yöntemlerden hem de geliştirilmiş GA'dan yararlanan hibrit bir teknik kullanır. Deneysel sonuçlar, hibrit yaklaşımların boyut indirgeme aşaması için sadece filtre tabanlı yöntemlerden daha iyi performans verdiğini göstermiştir. Uysal (Uysal, 2018) lokal ÖS yöntemleri, öznitelik dönüşümü ve sarıcı tabanlı ÖS yöntemlerinden oluşan iki aşamalı ÖS yöntemlerini analiz etti. Bu çalışmada, maksimum küreselleşme tekniği (MAX), ağırlıklı ortalama küreselleşme tekniği (AVG) ve her sınıf (EQ) için eşit sayıda öznitelik seçilmesi olarak adlandırılan üç öznitelik yapım tekniği ile birlikte dört yerel ÖS yöntemi kullanılmıştır. Daha sonra temel bileşen analizi (PCA), LSI ve GA kullanıldı. Deney sonuçlarına göre, AVG ve EQ yöntemleri genellikle MAX yönteminden daha iyi performans elde etti. Ayrıca, PCA, öznitelik dönüştürme aşamasında LSI ve GA yöntemlerinden daha başarılı olmuştur.

4. MEVCUT ÖZNİTELİK SEÇİM METOTLARI

MS alanında birçok filtre tabanlı ÖS yöntemi vardır. Bu tez çalışmasında ÖS aşaması için dokuz farklı filtre yöntemi kullanılmıştır. Bu metotlar; Chi-Square (CHI2), Class Discriminating Measure (CDM), Discriminative Power Measure (DPM), Odds Ratio (OR), Distinguishing Feature Selector (DFS), Comprehensively Measure Feature Selection (CMFS), Discriminative Feature Selection (DFSS), Normalized Difference Measure (NDM) ve Max-Min Ratio (MMR).

Tablo 4.1 ve Tablo 4.2, MS için ÖS yöntemleri hakkında bazı yaygın notasyonları ifade etmektedir.

Tablo 4.1. *C_j sınıfının ve t teriminin olasılık tablosu*

	t terimini içeren (t)	**t terimini içermeyen ($\bar{t}$)**
Sınıf içi (C_j)	a	c
Sınıf dışı ($\overline{C_j}$)	b	d

Tablo 4.2. *ÖS metotları için ön gösterimler*

Notasyon	**Değer**	**Anlamı**
a	*count(t,C_j)*	C_j sınıfında *t* terimini içeren doküman sayısı
b	*count(t,$\overline{C_j}$)*	Diğer sınıflarda ($\overline{C_j}$) *t* terimini içeren doküman sayısı
c	*count($\bar{t}$,C_j)*	C_j sınıfında *t* terimini içermeyen doküman sayısı
d	*count($\bar{t}$,$\overline{C_j}$)*	Diğer sınıflarda ($\overline{C_j}$) *t* terimini içermeyen doküman sayısı
e	*frequency(t, C_j)*	C_j sınıfındaki *t* teriminin sayısı
f	*frequency(t, $\overline{C_j}$)*	C_j sınıfı dışındaki diğer sınıflarda ($\overline{C_j}$) *t* teriminin sayısı
N	*(a+b+c+d)*	Tüm sınıflarda *t* terimini içeren toplam doküman sayısı
M	*count(C_j)*	Toplam sınıf sayısı
p(t)	*(a+b)/N*	*t* teriminin olma olasılığı
p($\bar{t}$)	*(c+d)/N*	*t* teriminin olmama olasılığı
p(C_j)	*(a+c)/N*	C_j sınıfının olma olasılığı
p($\overline{C_j}$)	*(b+d)/N*	C_j sınıfının olmama olasılığı
p(t, C_j)	*a/N*	C_j sınıfının *t* terimi ile olma olasılığı
p(t, $\overline{C_j}$)	*b/N*	C_j sınıfı dışındaki diğer sınıfların *t* terimi ile olma olasılığı
p($\bar{t}$, C_j)	*c/N*	C_j sınıfının *t* terimi ile olmama olasılığı
p($\bar{t}$, $\overline{C_j}$)	*d/N*	C_j sınıfı dışındaki diğer sınıfların *t* terimi ile olmama olasılığı
p(t|C_j)	*a/(a+c)*	C_j sınıfı mevcut olduğunda, *t* teriminin olasılığı

$p(\bar{t}\|C_j)$	$c/(a+c)$	C_j sınıfı mevcut olduğunda, t teriminin olmama olasılığı
$p(t\|\bar{C}_j)$	$b/(b+d)$	C_j sınıfı mevcut olmadığında, t teriminin olasılığı
$p(\bar{t}\|\bar{C}_j)$	$d/(b+d)$	C_j sınıfı mevcut olmadığında, t teriminin olmama olasılığı
$p(C_j\|t)$	$a/(a+b)$	t terimi mevcut olduğunda, C_j sınıfının olasılığı
$p(\bar{C}_j\|t)$	$b/(a+b)$	t terimi mevcut olduğunda, C_j sınıfının olmama olasılığı
$p(C_j\|\bar{t})$	$c/(c+d)$	t terimi mevcut olmadığında, C_j sınıfının olasılığı
$p(\bar{C}_j\|\bar{t})$	$d/(c+d)$	t terimi mevcut olmadığında, C_j sınıfının olmama olasılığı

4.1. Ki-kare (Chi-square-CHI2)

Popüler bir ÖS yöntemi olarak, CHI2 yöntemi, t özniteliğinin meydana gelmesi açısından C_j sınıfından bağımsız olması durumunda beklenen dağılımdan sapmayı hesaplar (Schütze, Manning, & Raghavan, 2008). CHI2'nin matematiksel formülü aşağıdaki gibi hesaplanabilir:

$$CHI2(t, c_j) = \frac{N \cdot [p(t, c_j) \cdot p(\bar{t}, \bar{c_j}) - p(t, \bar{c_j}) \cdot p(\bar{t}, c_j)]^2}{p(t) \cdot p(\bar{t}) \cdot p(C_j) \cdot p(\bar{C_j})} \quad \textbf{(4.1)}$$

Bir özniteliğin CHI2 skoru her sınıf için hesaplanır ve her öznitelik için tek bir skor elde etmek için sınıf temelli skorların global hale gelmesi gerekir. Bu skoru global hale getirmek için toplam, ağırlıklı toplam ve maksimum olarak adlandırılan üç yöntem vardır. Bu çalışmada CHI2 skorunu globalleştirmek için ağırlıklı toplam tercih edilmiştir.

$$CHI2(t) = p(c_j) \cdot CHI2(t, c_j) \quad \textbf{(4.2)}$$

4.2. Sınıf Ayırıcı Ölçütü (Class discriminating measure-CDM)

CDM, Chen ve arkadaşları (Chen, Huang, Tian, & Qu, 2009) tarafından önerilen Odds Ratio metodundan (OR) türetilmiştir. CDM, performansa göre Çok-Sınıflı Odds Ratio metodundan daha iyidir. Yöntemin formülü aşağıdaki gibidir:

$$CDM(t) = \sum_{j=1}^{M} \left| log \frac{P(t|\, c_j)}{P(t|\bar{c_j})} \right| \quad \textbf{(4.3)}$$

4.3. Ayrımcı Güç Ölçütü (Discriminative power measure-DPM)

DPM hem pozitif hem de negatif ayırt edici özniteliklere odaklanır. Ayrıca, yöntem düşük hesaplama süresine sahiptir (Chen, Lee, & Chang, 2009). DPM yönteminin amacı, sınıflar arasında daha büyük farklılıkları ortaya çıkaran öznitelikleri seçmektir. DPM, MS için daha yüksek ayırt edici özellikleri seçer. DPM metodunun formülü aşağıdaki gibi tanımlanmaktadır:

$$DPM(t) = \sum_{j=1}^{M} \left| P\left(t \middle| c_j\right) - P\left(t \middle| \overline{c_j}\right) \right| \tag{4.4}$$

4.4. Olasılık Oranı (Odds ratio-OR)

Deneyler çok sınıflı veri kümelerinde gerçekleştirildiğinden, bu çalışmada orijinal Odds Ratio (Chen, Huang, Tian, & Qu, 2009) yerine Çok-Sınıflı Odds Ratio(MOR) kullanılmıştır. MOR yöntemi sadece pozitif öznitelikleri değil, negatif öznitelikleri de seçer. MOR yönteminin formülü aşağıdaki gibidir:

$$MOR(t) = \sum_{j=1}^{M} \left| log \frac{P\left(t \middle| c_j\right) \cdot \left(1 - P\left(t \middle| \overline{c_j}\right)\right)}{P\left(t \middle| \overline{c_j}\right) \cdot \left(1 - P\left(t \middle| c_j\right)\right)} \right| \tag{4.5}$$

4.5. Ayırtedici Öznitelik Seçici (Distinguishing feature selector-DFS)

DFS (Uysal & Gunal, 2012), önceden tanımlanmış dört kritere dayanır. DFS, öznitelik karakterleriyle ilgili belirli gereksinimleri dikkate alarak önemsiz öznitelikleri ortadan kaldırırken, bilgilendirici öznitelikleri seçer.

$$DFS(t) = \sum_{j=1}^{M} \frac{P(C_j|t)}{P\left(\overline{t} \middle| C_j\right) + P\left(t \middle| \overline{C_j}\right) + 1} \tag{4.6}$$

4.6. Kapsamlı Öznitelik Seçim Ölçütü (Comprehensively measure feature selection-CMFS)

CMFS, bir özniteliğin hem sınıf içi hem de sınıflar arası önemini kapsamlı bir şekilde araştırmaktadır (Yang, Liu, Zhu, Liu, & Zhang, 2012). CMFS yönteminin formülü aşağıdaki gibidir:

$$CMFS(t) = \sum_{j=1}^{M} P\left(c_j\right) \cdot P\left(t \middle| c_j\right) \cdot P(C_j|t) \tag{4.7}$$

4.7. Ayrımcı Öznitelik Seçimi (Discriminative feature selection-DFSS)

DFSS (Zong, Wu, Chu, & Sculli, 2015) öznitelikleri bazı kriterlere göre seçer. Bunlar, tüm dokümanlarda daha yüksek terim frekansı, daha yüksek oluşum oranı ve gözardı edilen öznitelikleri seçmektir. Ayrıca, DFSS, sınıflar arasında ayrım yapabilen öznitelikleri seçer. Bu kriterlere göre, DFSS yöntemi aşağıdaki gibi gösterilmiştir:

$$DFSS(t, c_j) = \frac{e/a}{f/b} \cdot P(C_j|t) \cdot P(t|\,c_j) \cdot |P(C_j|t) - P(C_j|\bar{t})| \quad \textbf{(4.8)}$$

4.8. Normalleştirilmiş fark ölçütü (Normalized difference measure-NDM)

NDM, bağıl dokümanların frekansları göz önünde bulundurularak Rehman ve arkadaşları (Rehman, Javed, & Babri, 2017) tarafından geliştirilmiştir. BA(Balanced Accuracy Measure=Dengeli Doğruluk Ölçütü), pozitif sınıf ve negatif sınıftaki doküman frekansının farkını alarak öznitelik skorunu hesaplar. NDM, BA metodu geliştirilerek oluşturulmuştur. Yöntemin formülü aşağıdaki gibidir:

$$NDM(t) = \sum_{j=1}^{M} \frac{|P(t|\,c_j) - P(t|\overline{C_j})|}{\min(P(t|\,c_j), P(t|\overline{C_j}))} \quad \textbf{(4.9)}$$

4.9. Max-min Oranı (Max-min ratio-MMR)

BA metodu, pozitif sınıftaki ve negatif sınıftaki doküman frekansı arasında aynı farka sahip iki özniteliğe aynı skoru atamaktadır. Ayrıca NDM, büyük ve yüksek oranda dengesiz metin veri kümelerindeki son derece bilgilendirici olmayan seyrek özniteliklere yüksek skor atayabilir. Bununla birlikte MMR, BA ve NDM'nin geliştirilmiş versiyonudur. MMR, yüksek dengesiz sınıfları da içeren veri kümelerinde daha iyi performans verir (Rehman, Javed, Babri, & Asim, 2018). Yöntemin formülü aşağıdaki gibidir:

$$MMR(t) = \sum_{j=1}^{M} \frac{|P(t|\,c_j) - P(t|\overline{C_j})|}{\min(P(t|\,c_j), P(t|\overline{C_j}))} \cdot \max(P(t|\,c_j), P(t|\overline{C_j})) \quad \textbf{(4.10)}$$

5. METİN SINIFLANDIRMA İÇİN GLOBALLEŞTİRME TEKNİKLERİNİN ÖZNİTELİK SEÇİMİNE ETKİLERİ

Bu çalışmada, çoklu-sınıf dengesiz (MCU), çoklu-sınıf dengeli (MCB), ikili-sınıf dengesiz (BCU) ve ikili-sınıf dengeli (BCB) gibi farklı karakteristiğe sahip veri kümeleri kullanılarak lokal öznitelik seçim metotları üzerinde çeşitli globalleştirme tekniklerinin etkisi kapsamlı bir şekilde analiz edilmiştir. Bu amaçla literatürdeki globalleştirme tekniklerinin kullanıldığı çalışmalarda ayrı ayrı kullanılmış olan üç farklı teknik, üç farklı öznitelik seçim tekniğine uygulanıp sonuçlar irdelenmiştir. Dört farklı karakteristiğe sahip veri kümelerinde iki farklı sınıflandırıcı kullanılarak gerçekleştirilen deneyler, altı farklı öznitelik boyutlarındaki etkilerini de görebilmek amacıyla farklı sayıda öznitelik kümeleri ile gerçekleştirilmiştir.

5.1. Motivasyon

Daha önce 3.Bölüm'de de özetlenen global ve lokal öznitelik seçim teknikleri ile ilgili yapılan çalışmalardan da anlaşılacağı üzere, araştırmacılar yeni öznitelik seçim tekniği ya da şeması önermek amacıyla farklı şekillerde globalleştirme teknikleri kullanmaktadırlar. Metin sınıflandırma için özellikle son yıllarda kullanılan globalleştirme teknikleri toplam (SUM), ağırlıklı toplam (AVG) ve maksimum (MAX)'dur. Bu çalışmada, globalleştirme tekniklerinin performansını kapsamlı bir şekilde analiz ettik. Dört farklı önemli veri kümesinde DFSS, OR ve CHI2 olmak üzere üç LÖS yöntemi ve SUM, MAX ve AVG olmak üzere üç globalleştirme tekniği kullandık. Bu veri kümeleri Reuters-21578, 20Newsgroups, Enron1 ve Polarity'dir. Ayrıca, bu veri kümelerinin dört farklı özelliği vardır. Bu veri kümeleri sırasıyla MCU, MCB, BCU ve BCB'dir. Sınıflandırma aşamasında SVM ve DT olmak üzere iki başarılı tanınmış sınıflandırıcı kullanılmıştır (Parlak & Uysal, 2020).

5.2. Globalleştirme Teknikleri

Bu çalışmada SUM, AVG ve MAX olarak adlandırılan üç farklı globalleştirme tekniği kullanılmıştır. Burada, $f(t_i, C_j)$ C_j sınıfındaki t_i teriminin skoruna karşılık gelir. Bu teknikler aşağıdaki gibi hesaplanabilir:

$$SUM = \sum_{j=1}^{M} f(t_i, C_j) \qquad \textbf{(5.1)}$$

SUM globalleştirme tekniğinde, her sınıf için hesaplanan tüm skorlar toplanır. Skor bu şekilde global hale gelmektedir.

$$AVG = \sum_{j=1}^{M} P(C_j) * f(t_i, C_j) \quad \textbf{(5.2)}$$

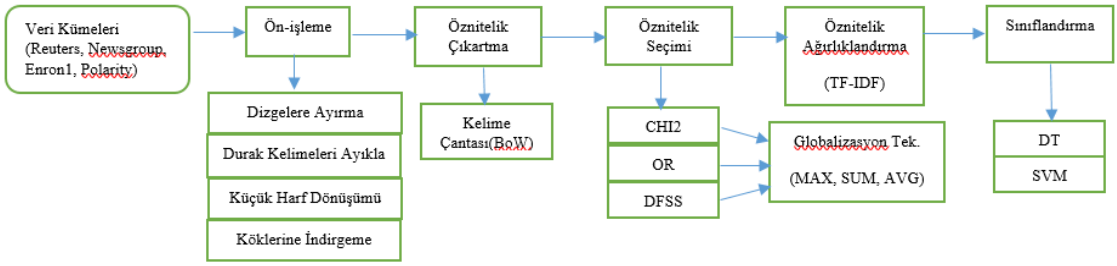
AVG globalleştirme tekniğinde, sınıf bazında hesaplanan skorlar sınıf olasılıkları kullanılarak globalleştirilir. Böylece, her sınıfın olasılığı da dikkate alınır.

$$MAX = max_{j=1}^{M}(f(t_i, C_j)) \quad \textbf{(5.3)}$$

MAX globalleştirme tekniğinde, sınıf bazında hesaplanan skorlardan en yüksek olanı alınır. Diğer sınıflarda hesaplanan skorlar dikkate alınmaz.

5.3. Deneysel Çalışma

Deneylerimizde, Reuters-21578, 20Newsgroups, Enron1 ve Polarity adlı dört veri kümesinde MS için üç globalleşme tekniğini test ettik. Tüm bu veri kümelerinin farklı özellikleri vardır. Bu veri kümeleri performans değerlendirmesi için MS alanında yaygın olarak kullanılmaktadır. İlk veri kümesi, ilk 10 sınıfı içeren Reuters-21578 ModApte bölünmesidir (Asuncion, 2007). Reuters-21578 veri kümesi literatürde MS çalışmalarında yaygın olarak kullanılan ve Reuters-21578 ModApte olarak anılan bölümün en çok doküman içeren ilk 10 sınıfını içermektedir. Her sınıftaki doküman sayıları birbirinden farklı olduğundan dolayı Reuters-21578 metin veri kümesi dengesiz yapıya sahip veri kümesi olarak değerlendirilir. Deneylerde kullanılan ikinci veri kümesi ise biri hariç (997 adet) her sınıfında 1000'er dokümanı bulunan toplamda 20 sınıfa sahip olan 20Newsgroups'tur (Asuncion, 2007). Bu metin veri kümesinin de ilk 10 sınıfı deneylerde kullanılmıştır. Her sınıftaki doküman sayısı eşit olduğundan 20Newsgroups veri kümesi dengeli bir yapıya sahiptir. Üçüncü veri kümesi, Enron1 adlı bir spam e-posta koleksiyonudur (Uysal & Gunal, 2012). Bu veri kümesi ise sadece iki sınıftan oluşup, sınıflardaki doküman sayısı birbirine eşit olmadığından dengesiz veri kümesi olarak değerlendirilir. Son veri kümesi, film yorumlarını içeren Polarity veri kümesidir (Parlak & Uysal, 2020). Bu veri kümesi de iki sınıf içermekte olup, her sınıftaki doküman sayısı eşit olduğundan dengeli bir yapıya sahiptir. Bu şekilde, MCU, MCB, BCU ve BCB olmak üzere 4 farklı özelliğe sahip veri kümelerini kullandık. Yapılan deneysel çalışmaları Şekil 5.1.'de gösterdik.



Şekil 5.1. *Deneysel çalışma*

Deneylerde, Reuters-21578 veri kümesinin kendine has önceden bölümlendirilmiş eğitim ve test dokümanları kullanılmış olup, diğer veri kümeleri için ise eğitim ve test için kullanılmak üzere her sınıftan belli oranda doküman içeren (%70 ve %30) ayrı iki bölüm oluşturulmuştur. Kullanılan dört veri kümesine ait sınıf ve doküman sayısı bilgileri Tablo 5.1, Tablo 5.2, Tablo 5.3 ile Tablo 5.4'te gösterilmiştir.

Tablo 5.1. *Reuters-21578 veri kümesi*

No	Sınıf Etiketi	Eğitim Dokümanları	Test Dokümanları
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 5.2. *20Newsgroups veri kümesi*

No	Sınıf Etiketi	Eğitim Dokümanları	Test Dokümanları
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 5.3. *Enron1 veri kümesi*

No	Sınıf Etiketi	Eğitim Dokümanları	Test Dokümanları
1	Legitimate	2570	1102
2	Spam	1050	450

Tablo 5.4. *Polarity veri kümesi*

No	Sınıf Etiketi	Eğitim Dokümanları	Test Dokümanları
1	Positive	700	300
2	Negative	700	300

Bu çalışmada ön işleme aşamasında, bahsi geçen veri kümelerinden elde edilen doküman içeriklerine; sırasıyla, dizgelere ayırma, gereksiz kelimeleri ayıklama, küçük harf dönüşümü ve kelimeleri köklerine indirgeme gibi ön işlemler uygulanmıştır. Öznitelik çıkartma için yaygın olarak bilinen bir metin temsil yöntemi olan BoW (kelime çantası) yaklaşımı kullanılmıştır. Öznitelik seçim metodu olarak ise DFSS, OR ve CHI2 metotları kullanılmıştır. Her kelime, bu gösterimdeki bir özniteliğe karşılık gelir. Öznitelikleri sayısal olarak göstermek için, öznitelik ağırlıklandırma aşamasında TF-IDF ağırlıklandırma yöntemini kullandık. Sınıflandırma aşamasında, daha önceki bölümlerde çalışma stilleri kabaca anlatılan SVM ile DT sınıflandırma algoritmaları kullanılmıştır. SVM sınıflandırıcı, yapılan tüm deneylerde lineer versiyonu ile uygulanmıştır.

5.4. Değerlendirme Ölçütleri

MS çalışmalarında, önerilen öznitelik seçim metotlarının başarımları çoğunlukla Mikro-ortalama F ölçütü (Mikro-F1) ve Makro-ortalama F ölçütü (Makro-F1) değerlendirme ölçütleri ile değerlendirilmektedir. Değerlendirme ölçütlerinin ifade edilmesine yardımcı olmak için bir olasılık tablosu Tablo 5.5'te sunulmaktadır.

Tablo 5.5. *c sınıfı için olasılık tablosu*

	Etiket 'Yes'	Etiket 'No'
Sınıflandırılmış 'Yes'	*A*	*C*

Sınıflandırılmış 'No'	*B*	*D*

Bu iki kriteri hesaplamak için c sınıfı için kesinlik (P), duyarlılık (R) ve F-ölçütü (F) olmak üzere üç kriter hesaplanmalıdır. Bu kriterlerin formülü aşağıda gösterilmiştir:

$$R(c) = \frac{A}{\mathrm{A} + \mathrm{B}} \tag{5.4}$$

$$P(c) = \frac{A}{\mathrm{A} + \mathrm{C}} \tag{5.5}$$

$$F(c) = \frac{2 * P(c) * R(c)}{P(c) + R(c)} \tag{5.6}$$

c sınıfı için duyarlılık (R(c)), hedef sınıfa atanan tüm dokümanlar arasındaki doğru atamaların oranıdır. c sınıfı için kesinlik (P(c)), hedef sınıfa atanan tüm dokümanlar arasındaki gerçek atamaların oranıdır. c sınıfı için F-ölçütü (F (c)), P (c) ve R(c) 'nin harmonik ortalaması kullanılarak hesaplanır. Mikro-F1 ölçütünde, tüm sınıflandırma kararları, sınıf bilgileri tamamen göz ardı ederek dikkate alınır. Dengeli olmayan veri kümelerinde, sınıflandırıcılar doküman sayısı fazla olan sınıflara atama yapmaya daha yatkın olabilmektedirler. Bu yüzden, bir koleksiyondaki sınıflar içindeki doküman sayısı dengeli değilse, büyük sınıflar diğer küçük sınıfları domine edebilir.

Ancak, Macro-F1 ölçütü veri kümesindeki her sınıf için ayrı ayrı hesaplanır. Daha sonra, tüm sınıflar üzerinden ortalama bir skor alınır. Bu metrik, sınıf yoğunluğunu göz ardı ederek her sınıfa eşit ağırlık atar. Dengesiz veri kümelerinde öznitelik seçim metotlarının performanslarını değerlendirirken Makro-F1 ölçütünü kullanmak, sınıflandırıcıların daha az dokümana sahip olan sınıfları ayırt edebilme yeteneklerini daha iyi gösterebilmeleri açısından daha adil bir seçim olabilir. (p_j, r_j) çifti sırasıyla j. sınıfın kesinlik ve duyarlılık değerlerine karşılık gelir.

Micro-F ve Macro-F ölçütleri aşağıdaki gibi formüle edilebilir:

$$R_{micro} = \frac{\sum_{j=1}^{M} A_j}{\sum_{j=1}^{M} (A_j + B_j)} \tag{5.7}$$

$$P_{micro} = \frac{\sum_{j=1}^{M} A_j}{\sum_{j=1}^{M} (A_j + C_j)} \quad \textbf{(5.8)}$$

$$F_{micro} = \frac{2R_{micro} x P_{micro}}{R_{micro} + P_{micro}} \quad \textbf{(5.9)}$$

$$R_{macro} = \frac{\sum_{j=1}^{M} (\frac{A_j}{(A_j + B_j)})}{c} \quad \textbf{(5.10)}$$

$$P_{macro} = \frac{\sum_{j=1}^{M} (\frac{A_j}{(A_j + C_j)})}{c} \quad \textbf{(5.11)}$$

$$F_{macro} = \frac{2R_{macro} x P_{macro}}{R_{macro} + P_{macro}} \quad \textbf{(5.12)}$$

Deneysel çalışma kısmında yukarıda bahsedilen üç globalleştirme tekniklerinin her bir öznitelik seçim metodu için sınıflandırma performansına olan etkisi hem Mikro-F1 hem de Makro-F1 ölçütü kullanılarak ölçülmüş ve değerlendirilmiştir. Sonuçlar ve değerlendirmeler ilerleyen alt bölümde sunulmuştur.

5.5. Terim Benzerlik Analizi

Seçilen öznitelikler ÖS yöntemine ve kullanılan globalleştirme tekniğine göre farklılık gösterir. Seçilen öznitelikler bir ÖS yönteminin performansını değerlendirmek için iyi bir göstergedir. Ayırt edici öznitelikler yüksek bir skor atanarak seçilirken, ayırt edici olmayan öznitelikler düşük bir skor atanarak seçilmeyebilir. Belirtilen globalleştirme tekniklerini kullanarak her ÖS yöntemi tarafından seçilen ilk 10 öznitelik Tablo 5.6.-Tablo 5.9.'da listelenmiştir. Bu tablolarda, bireysel seçim yöntemine ve globalleşme tekniğine özgü öznitelikler koyu renkle gösterilmiştir. Tablolar, seçilen özniteliklerin yalnızca kullanılan ÖS yöntemine göre değil, aynı zamanda kullanılan globalleştirme tekniğine göre de değiştiğini göstermektedir.

Tablo 5.6. *Reuters-21578 veri kümesinde en iyi 10 öznitelik*

Metotlar	Öznitelikler
DFSS(MAX)	shr, rev, qtr, barrel, ct, opec, oil, crude, bpd, net
DFSS(SUM)	shr, rev, qtr, barrel, ct, opec, oil, crude, bpd, net
DFSS(AVG)	shr, rev, qtr, ct, net, barrel, loss, opec, **payout,** oil
OR(MAX)	wheat, rev, **taupo, payout,** shr, **zeebrugg,** qtr, **bpd, dump, fernando**
OR(SUM)	net, ct, agriculture, **prior,** wheat, grain, **england, corn,** qtr, **profit**
OR(AVG)	qtr, ct, rev, agriculture, monetary, grain, **shortage,** shr, **net, economist**
CHI2(MAX)	ct, barrel, oil, net, shr, wheat, **crude,** qtr, **ship, trade**
CHI2(SUM)	ct, wheat, net, tonn, oil, shr, **agriculture,** barrel, **corn, bank**
CHI2(AVG)	ct, net, shr, qtr, rev, loss, acquir, **profit, note, dividend**

Tablo 5.7. *20Newsgroups veri kümesinde en iyi 10 öznitelik*

Metotlar	Öznitelikler
DFSS(MAX)	atheism, motorcycle, baseball, forsal, auto, religion, atheist, bike, sport, dod
DFSS(SUM)	atheism, motorcycle, baseball, forsal, auto, religion, atheist, bike, sport, dod
DFSS(AVG)	atheism, motorcycle, baseball, forsal, auto, religion, atheist, bike, sport, dod
OR(MAX)	cantaloup, cs, id, newsgroup, path, subject, date, messag, cmu, srv
OR(SUM)	cantaloup, cs, id, newsgroup, path, subject, date, messag, cmu, srv
OR(AVG)	cantaloup, cs, id, newsgroup, path, subject, date, messag, cmu, srv
CHI2(MAX)	atheism, baseball, motorcycle, forsal, auto, sport, **os, ms, mac, graphic**
CHI2(SUM)	baseball, forsal, atheism, auto, motorcycle, rec, sport, comp, hardwar, misc
CHI2(AVG)	baseball, forsal, atheism, auto, motorcycle, rec, sport, comp, hardwar, misc

Tablo 5.8. *Enron1 veri kümesinde en iyi 10 öznitelik*

Metotlar	Öznitelikler
DFSS(MAX)	ect, meter, cc, nom, pm, volum, ga, attach, corp, deal
DFSS(SUM)	ect, meter, cc, nom, pm, volum, ga, attach, corp, deal
DFSS(AVG)	ect, meter, cc, nom, pm, volum, ga, attach, corp, deal
OR(MAX)	subject, **enron,** meter, **hpl, daren, hou,** nom, **mmbtu,** ect, weight
OR(SUM)	subject, meter, nom, ect, weight, cc, dealer, gra, tel, ali
OR(AVG)	subject, meter, nom, ect, weight, cc, dealer, gra, tel, ali
CHI2(MAX)	enron, http, cc, hpl, ga, ect, **daren, hou,** pm, meter
CHI2(SUM)	http, cc, ga, ect, pm, meter, enron, forward, offer, **corp**
CHI2(AVG)	http, cc, enron, ga, ect, pm, meter, hpl, forward, offer

Tablo 5.9. *Polarity veri kümesinde en iyi 10 öznitelik*

Metotlar	Öznitelikler
DFSS(MAX)	bad, wast, worst, stupid, bore, life, great, outstand, perform, world
DFSS(SUM)	bad, wast, worst, stupid, bore, life, great, outstand, perform, world
DFSS(AVG)	bad, wast, worst, stupid, bore, life, great, outstand, perform, world
OR(MAX)	outstand, magnific, **atroci, darker, incoher,** uninvolv, **justin, misfir, lovingly, joli**
OR(SUM)	outstand, magnific, uninvolv, ludicr, plod, rivet, themat, numb, seagal, flawless
OR(AVG)	outstand, magnific, uninvolv, ludicr, plod, rivet, themat, numb, seagal, flawless
CHI2(MAX)	wast, bad, stupid, worst, bore, suppose, great, outstand, ridicule, life
CHI2(SUM)	wast, bad, stupid, worst, bore, suppose, great, outstand, ridicule, life
CHI2(AVG)	wast, bad, stupid, worst, bore, suppose, great, outstand, ridicule, life

5.6. Performans Analizi

Bu bölümde, farklı karakteristiğe sahip 4 veri kümesi, 3 ÖS yöntemi, 3 globalleştirme tekniği ve 2 sınıflandırıcı kullanılarak kapsamlı bir analiz çalışması yapılmıştır. DFSS, OR ve CHI2 yöntemleri bu çalışmada LÖS yöntemlerine örnek olarak

kullanılmıştır. Ancak, bu çalışmada kullanılan 2 örüntü sınıflandırıcısı SVM ve DT'dir. Ayrıca, bu çalışmada kullanılan 3 farklı globalleştirme tekniği MAX, SUM ve AVG'dir. Tüm veri kümeleri için Porter kök bulma algoritması ve öznitelik ağırlıklandırma için TF-IDF kullanılmıştır. Deneylerde, sınıflandırma algoritmalarının performansını analiz etmek için iyi bilinen Mikro-F1 ölçütü ve Makro-F1 ölçütü kullanılmıştır. MCB ve MCU olan veri kümeleri için sırasıyla ilk 10 ve en çok doküman içeren 10 sınıfı kullandık. BCB ve BCU olan veri kümeleri yalnızca 2 sınıf içerir. Reuters-21578 dışındaki veri kümeleri, adil bir değerlendirme için eğitim (% 70) ve test (% 30) olarak ikiye ayrılmıştır. Reuters-21578 veri kümesinde önceden oluşturulmuş bir eğitim ve test bölümü zaten vardır.

Her seçim yöntemi tarafından seçilen farklı boyutlardaki öznitelikler SVM ve DT sınıflandırıcılarına girdi olarak gönderilmiştir. Öznitelik kümeleri, 50, 100, 300, 1000, 3000 ve 5000 gibi çeşitli öznitelik boyutları kullanılarak oluşturulmuştur. Toplam öznitelik sayısı Reuters-21578, 20Newsgroups, Enron1 ve Polarity veri kümeleri için sırasıyla 16867, 50419, 31238 ve 21875'tir. Ortaya çıkan Mikro-F1 ve Makro-F1 ölçüt skorları Reuters-21578 veri kümesi için Tablo 5.10.-Tablo 5.13.'te listelenmiştir.

Tablo 5.10. *SVM ile Reuters-21578 veri kümesinden elde edilen Mikro-F1 skorları (%)*

	DFSS			**OR**			**CHI2**		
Boyut	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**
50	90.36	90.79	91.03	83.44	88.56	84.45	91.54	91.28	90.83
100	91.58	91.66	92.06	85.83	91.07	87.12	91.85	92.02	92.27
300	91.79	91.96	92.64	87.46	91.72	91.07	91.96	92.37	92.31
500	92.39	92.16	92.29	88.03	91.35	91.35	92.35	92.16	92.39
1000	92.52	92.49	92.52	89.14	91.34	92.12	92.31	92.41	92.62
3000	92.54	92.60	**92.73**	91.68	92.39	**92.41**	92.56	92.56	**92.66**

Tablo 5.11. *SVM ile Reuters-21578 veri kümesinden elde edilen Makro-F1 skorları (%)*

	DFSS			OR			CHI2		
Boyut	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**
50	58.23	59.67	59.65	39.50	51.31	40.53	61.39	60.71	56.20
100	61.35	60.66	61.92	46.06	60.13	48.68	61.79	61.93	60.70
300	62.92	64.74	65.06	50.87	**64.74**	62.05	63.61	66.32	**64.75**
500	65.55	65.56	**65.94**	52.19	63.44	62.64	64.16	64.45	64.57
1000	64.36	63.81	64.50	58.32	63.09	63.19	63.26	63.79	64.25
3000	63.45	63.58	63.46	64.04	63.65	63.29	64.16	63.56	64.12

Tablo 5.12. *DT ile Reuters-21578 veri kümesinden elde edilen Mikro-F1 skorları (%)*

	DFSS			OR			CHI2		
Boyut	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**
50	89.71	89.65	89.69	83.29	88.16	84.24	89.76	89.45	88.52
100	90.28	**90.45**	90.66	85.26	90.25	86.78	**90.47**	90.19	89.58
300	89.69	89.62	89.99	87.03	89.73	90.08	90.15	90.12	89.97
500	89.36	89.36	89.73	87.53	89.60	89.43	89.95	89.91	89.36
1000	90.30	90.15	89.97	88.79	88.94	89.80	89.58	89.71	89.69
3000	89.89	89.93	89.73	**90.34**	90.25	89.08	89.45	89.41	89.34

Tablo 5.13. *DT ile Reuters-21578 veri kümesinden elde edilen Makro-F1 skorları (%)*

	DFSS			OR			CHI2		
Boyut	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**
50	57.14	57.04	56.63	36.72	53.45	38.58	57.05	56.91	52.98
100	60.05	**61.06**	58.56	44.33	57.49	45.36	58.62	**58.71**	56.67
300	57.92	57.44	58.16	50.02	**61.02**	56.97	57.75	57.57	57.42
500	57.23	57.17	60.98	51.31	57.66	55.71	58.21	58.31	57.19
1000	58.47	58.20	58.01	55.44	56.67	56.79	57.44	58.08	58.38
3000	58.23	58.17	58.12	58.52	58.17	56.44	57.48	57.37	57.25

Reuters-21578 veri kümesi için en yüksek Mikro-F1 ölçüt skoru 92.73'tür. DFSS yöntemi, AVG globalleştirme tekniği, 3000 öznitelik kullanılarak ve SVM sınıflandırıcısının kombinasyonu ile elde edilmiştir. Bununla birlikte, en yüksek Makro-F1 skoru 65.94'tür. DFSS yöntemi, AVG globalleştirme tekniği ve 500 öznitelik kullanılarak SVM sınıflandırıcısının kombinasyonu ile elde edilmiştir. DT sınıflandırıcı

çoğu durumda ikincil sınıflandırma algoritmasıdır. SVM sınıflandırıcısı için AVG globalleştirme tekniği diğer globalleştirme tekniklerinden daha başarılı olsa da, DT sınıflandırıcısı için SUM globalleştirme tekniği diğer globalleştirme tekniklerinden daha başarılı performans göstermiştir.

Micro-F1 ve Macro-F1 ölçüt skorları 20Newsgroups veri kümesi için Tablo 5.14.-Tablo 5.17.'te listelenmiştir.

Tablo 5.14. *SVM ile 20Newsgroups veri kümesinden elde edilen Mikro-F1 skorları (%)*

	DFSS			OR			CHI2		
Boyut	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**
50	98.03	98.32	98.32	98.56	**98.63**	**98.63**	98.41	**98.68**	**98.68**
100	98.32	98.61	98.61	98.53	98.43	98.43	98.58	98.53	98.53
300	98.56	**98.67**	**98.67**	98.58	97.98	97.98	98.31	98.13	98.13
500	98.32	98.32	98.32	98.60	97.77	97.77	98.24	97.92	97.92
1000	98.06	98.29	98.29	98.49	97.63	97.63	98.17	98.17	98.17
3000	97.94	97.77	97.77	97.92	97.75	97.75	97.79	97.77	97.77

Tablo 5.15. *SVM ile 20Newsgroups veri kümesinden elde edilen Makro-F1 skorları (%)*

	DFSS			OR			CHI2		
Boyut	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**
50	96.11	96.66	96.66	97.17	**97.30**	**97.30**	96.84	**97.40**	**97.40**
100	96.67	97.25	97.25	97.09	96.89	96.89	97.19	97.08	97.08
300	97.18	**97.36**	**97.36**	97.18	96.02	96.02	96.67	96.33	96.33
500	96.70	96.68	96.68	97.22	95.60	95.60	96.54	95.90	95.90
1000	96.19	96.61	96.61	97.02	95.37	95.37	96.40	96.39	96.39
3000	95.96	95.62	95.62	95.93	95.58	95.58	95.64	95.62	95.62

Tablo 5.16. *DT ile 20Newsgroups veri kümesinden elde edilen Mikro-F1 skorları (%)*

	DFSS			OR			CHI2		
Boyut	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**
50	98.77	98.99	98.99	98.75	**99.06**	**99.06**	98.89	98.94	98.94
100	98.97	**99.06**	**99.06**	98.87	98.87	98.87	**98.99**	98.92	98.92
300	98.99	98.99	98.99	98.97	98.92	98.92	98.91	98.92	98.92
500	98.87	98.87	98.87	98.87	98.92	98.92	98.94	98.92	98.92
1000	98.97	98.96	98.96	99.02	98.92	98.92	98.94	98.96	98.96
3000	98.99	98.97	98.97	98.94	98.94	98.94	98.97	98.96	98.96

Tablo 5.17. *DT ile 20Newsgroups veri kümesinden elde edilen Makro-F1 skorları (%)*

	DFSS			OR			CHI2		
Boyut	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**
50	97.56	98.00	98.00	97.54	**98.12**	**98.12**	97.80	97.89	97.89
100	97.97	**98.14**	**98.14**	97.76	97.77	97.77	**98.00**	97.86	97.86
300	98.00	98.00	98.00	97.98	97.87	97.87	97.82	97.87	97.87
500	97.75	97.75	97.75	97.77	97.86	97.86	97.89	97.87	97.87
1000	97.97	97.93	97.93	98.06	97.87	97.87	97.89	97.94	97.94
3000	98.00	97.97	97.97	97.91	97.90	97.90	97.96	97.94	97.94

20Newsgroups veri kümesi için en yüksek Mikro-F1 skoru 99.06'dır. DFSS yöntemi, SUM ve AVG globalleştirme tekniği, 100 öznitelik kullanılarak ve DT sınıflandırıcısının kombinasyonu ile elde edilmiştir. Bu veri kümesi için, SUM ve AVG globalleştirme teknikleriyle elde edilen skorlar, eşit sınıf olasılıkları nedeniyle tüm durumlar için eşittir. Bununla birlikte, en yüksek Makro-F1 skoru 98.14'tür. DFSS yöntemi, SUM ve AVG globalleştirme teknikleri, 100 öznitelik kullanılarak ve DT sınıflandırıcısının kombinasyonu ile elde edilmiştir. SVM sınıflandırıcı çoğu durumda ikincil sınıflandırma algoritmasıdır. SUM ve AVG globalleştirme teknikleri, SVM ve DT sınıflandırıcıları için MAX globalleştirme tekniğinden daha başarılıdır.

Mikro-F1 ve Makro-F1 ölçüm skorları, Enron1 veri kümesi için Tablo 5.18.-Tablo 5.21.'te listelenmiştir.

Tablo 5.18. *SVM ile Enron1 veri kümesinden elde edilen Mikro-F1 skorları (%)*

	DFSS			OR			CHI2		
Boyut	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**
50	92.56	92.56	92.48	91.09	89.50	91.81	94.10	94.13	93.99
100	93.77	93.77	94.10	93.55	90.70	92.60	95.45	95.24	95.14
300	95.31	95.31	95.06	95.49	93.99	94.39	95.56	95.52	94.67
500	95.10	**95.63**	95.35	95.38	95.06	95.31	95.77	95.77	96.15
1000	94.57	94.17	94.13	96.08	95.66	95.10	96.36	96.36	**96.71**
3000	94.60	94.53	94.60	96.08	96.08	**96.22**	95.80	96.29	96.19

Tablo 5.19. *SVM ile Enron1 veri kümesinden elde edilen Makro-F1 skorları (%)*

	DFSS			OR			CHI2		
Boyut	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**
50	84.15	84.15	84.00	82.30	72.20	83.25	87.25	87.25	87.00
100	86.50	86.50	87.05	86.30	75.80	84.55	89.95	89.50	89.25
300	89.45	89.45	88.95	89.90	86.85	87.60	90.05	89.90	88.10
500	88.80	**89.95**	89.30	89.50	88.80	89.35	90.40	90.45	91.25
1000	87.70	86.85	86.70	91.00	90.00	88.85	91.45	91.45	**92.30**
3000	87.75	87.60	87.75	90.95	90.90	**91.20**	90.20	91.30	91.10

Tablo 5.20. *DT ile Enron1 veri kümesinden elde edilen Mikro-F1 skorları (%)*

	DFSS			OR			CHI2		
Boyut	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**
50	91.85	91.85	92.00	89.89	86.51	91.85	94.42	94.71	94.64
100	91.96	91.96	92.71	92.52	86.51	92.07	**95.25**	93.70	94.17
300	**92.78**	92.56	92.11	**94.35**	92.82	93.63	94.31	93.15	93.08
500	91.92	92.19	92.71	92.93	92.45	93.95	92.97	93.26	94.21
1000	91.81	91.81	91.81	93.26	94.10	93.33	93.26	93.33	93.44
3000	92.07	92.07	92.18	93.44	94.10	94.06	93.44	93.22	94.21

Tablo 5.21. *DT ile Enron1 veri kümesinden elde edilen Makro-F1 skorları (%)*

	DFSS			OR			CHI2		
Boyut	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**
50	83.25	83.25	83.45	80.35	75.00	83.35	87.80	88.35	88.25
100	83.05	83.05	84.30	84.50	75.00	83.75	**89.55**	86.45	87.25
300	**84.60**	84.10	83.15	**87.85**	84.95	86.40	87.65	85.10	85.35
500	82.75	83.30	84.45	84.90	83.75	86.85	84.70	85.40	87.40
1000	82.65	82.65	82.60	85.25	87.20	85.50	85.50	85.60	85.80
3000	82.90	82.90	83.05	85.35	86.90	86.80	85.85	85.35	87.20

Enron1 veri kümesi için, en yüksek Mikro-F1 ölçüt skoru 96.71'dir. CHI2 öznitelik seçim yöntemi, AVG globalleştirme tekniği, 1000 öznitelik kullanılarak ve SVM sınıflandırıcı kombinasyonu ile elde edilmiştir. Bununla birlikte, en yüksek Makro-F1 skoru 65.94'tür. CHI2 özellik seçim yöntemi, AVG globalleştirme tekniği, 1000 öznitelik kullanılarak ve SVM sınıflandırıcı kombinasyonu ile elde edilmiştir. DT sınıflandırıcı çoğu durumda ikincil sınıflandırma algoritmasıdır. AVG globalleştirme tekniği SVM sınıflandırıcısı için diğer globalleştirme tekniklerinden daha başarılı olmakla birlikte, MAX globalleştirme tekniği DT sınıflandırıcısı için diğer globalleştirme tekniğinden daha başarılıdır.

Polarity veri kümesi için Mikro-F1 ve Makro-F1 ölçüt skorları Tablo 5.22.-Tablo 5.25.'te listelenmiştir.

Tablo 5.22. *SVM ile Polarity veri kümesinden elde edilen Mikro-F1 skorları (%)*

	DFSS			OR			CHI2		
Boyut	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**
50	86.47	87.22	87.22	73.55	75.39	75.39	86.90	86.90	86.90
100	88.27	87.01	87.01	76.03	75.78	75.78	**89.20**	**89.20**	**89.20**
300	87.64	**88.68**	**88.68**	78.05	82.35	82.35	87.32	87.43	87.43
500	86.36	88.27	88.27	79.76	82.47	82.47	86.15	88.16	88.16
1000	88.58	88.06	88.06	83.04	84.17	84.17	86.58	87.11	87.11
3000	88.06	88.06	88.06	85.93	**87.32**	**87.32**	87.22	87.85	87.85

Tablo 5.23. *SVM ile Polarity veri kümesinden elde edilen Makro-F1 skorları (%)*

	DFSS			OR			CHI2		
Boyut	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**
50	76.15	77.25	77.25	53.65	58.25	58.25	76.80	76.80	76.80
100	78.95	77.00	77.00	59.65	60.30	60.30	**80.45**	**80.45**	**80.45**
300	78.00	**79.70**	**79.70**	63.85	70.00	70.00	77.50	77.65	77.65
500	76.00	79.00	79.00	66.20	70.15	70.15	75.65	78.80	78.80
1000	79.50	78.70	78.70	71.00	72.65	72.65	76.35	77.15	77.15
3000	78.65	78.65	78.65	75.30	**77.50**	**77.50**	77.35	78.35	78.35

Tablo 5.24. *DT ile Polarity veri kümesinden elde edilen Mikro-F1 skorları (%)*

	DFSS			OR			CHI2		
Boyut	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**
50	80.24	77.43	77.43	69.00	69.99	69.99	83.16	83.16	83.16
100	80.72	80.83	80.83	69.99	70.83	70.83	**83.50**	**83.50**	**83.50**
300	82.12	**82.81**	**82.81**	70.83	81.31	81.31	81.77	82.01	82.01
500	81.77	82.01	82.01	78.67	81.31	81.31	80.60	82.23	82.23
1000	80.12	78.67	78.67	81.66	82.58	82.58	82.58	82.12	82.12
3000	81.42	79.64	79.64	81.31	**83.72**	**83.72**	81.42	81.54	81.54

Tablo 5.25. *DT ile Polarity veri kümesinden elde edilen Makro-F1 skorları (%)*

	DFSS			OR			CHI2		
Boyut	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**	**MAX**	**SUM**	**AVG**
50	66.95	63.05	63.05	39.75	43.30	43.30	71.10	71.10	71.10
100	67.60	67.80	67.80	43.30	46.30	46.30	**71.65**	**71.65**	**71.65**
300	69.56	**70.65**	**70.65**	46.30	67.55	67.55	69.15	69.45	69.45
500	69.00	69.25	69.25	62.25	67.55	67.55	67.50	69.85	69.85
1000	66.80	64.75	64.75	68.25	70.10	70.10	70.30	69.70	69.70
3000	68.65	66.15	66.15	68.35	**71.95**	**71.95**	68.70	68.85	68.85

Polarity veri kümesi için en yüksek Mikro-F1 ölçüt skoru 89.20'dir. CHI2 yöntemi, MAX, SUM ve AVG globalleştirme teknikleri ve 100 öznitelik kullanılarak SVM sınıflandırıcı kombinasyonu ile elde edilmiştir. Bu veri kümesinde, SUM ve AVG skorları eşit sınıf olasılıkları nedeniyle tüm durumlar için eşittir. Ancak, en yüksek

Makro-F1 skoru 80.45'tir. CHI2 yöntemi, MAX, SUM ve AVG globalleştirme teknikleri ve 100 öznitelik kullanılarak SVM sınıflandırıcı kombinasyonu ile elde edilmiştir. DT sınıflandırıcı çoğu durumda ikincil sınıflandırma algoritmasıdır. SUM ve AVG globalleştirme teknikleri, SVM ve DT sınıflandırıcısı için MAX globalleştirme tekniğinden daha başarılıdır.

5.7. Sonuçlar

Genel olarak, SUM ve AVG globalleştirme teknikleri 20Newsgroups ve Polarity olan dengeli veri kümelerinde MAX'dan daha başarılıdır. Dengeli veri kümelerinde, sınıf olasılığı her sınıf için eşittir. Böylece, SUM ve AVG globalleştirme teknikleri özniteliklerin aynı alt kümesini seçer. Bununla birlikte, globalleştirme tekniklerinin performansı, dengesiz veri kümelerinde sınıflandırıcıya göre değişir. Reuters veri kümesinde, AVG globalleştirme tekniği SVM sınıflandırıcısı ile en iyi performansı elde ederken, SUM globalleştirme tekniği DT sınıflandırıcısı ile en iyi performansı gösterdi. Ayrıca, AVG globalleştirme tekniği SVM sınıflandırıcısı ile en iyi performansı gösterirken, MAX globalleştirme tekniği Enron1 veri kümesinde DT sınıflandırıcısı ile en iyi performansı göstermiştir.

6. KAPSAMLI ÖZNİTELİK SEÇİCİ (EXTENSIVE FEATURE SELECTOR-EFS)

Literatürde çok sayıda çalışma olmasına rağmen, öznitelik seçim konusu metin sınıflandırma alanı için halen devam eden bir araştırma konusudur (Rehman, Javed, & Babri, 2017; Rehman, Javed, Babri, & Asim, 2018). Bu alanda çalışan araştırmacılar, daha ayırt edici öznitelikleri seçmek için yeni teknikler aramaktadırlar. Çünkü ayırt edici öznitelikler hem sınıflandırma doğruluğunu iyileştirebilmekte hem de işlem süresini azaltabilmektedir. Bu amaçla, metin sınıflandırması için yeni bir filtre temelli terim ve sınıf bazlı olasılıklardan oluşan, Kapsamlı Öznitelik Seçici (EFS) adıyla yeni bir öznitelik seçim metodu önerilmiştir (Parlak & Uysal, 2021). EFS'nin teorik altyapısı ve ilgili deneyler aşağıdaki alt bölümlerde verilmiştir.

6.1.Teorik Altyapı

İdeal bir filtre tabanlı öznitelik seçim metodu; ayırt edici özniteliklere yüksek skor atarken, ayırt edici olmayan özniteliklere düşük skor atamalıdır. Metin sınıflandırma çalışmalarında her terim farklı bir özniteliğe karşılık gelir. Bir özniteliğe karşılık gelen skor belirli bir formüle göre hesaplanır. Bu çalışmada, özniteliğin sınıf bazlı ve koleksiyon bazlı ayırt ediciliği göz önünde bulundurularak yeni bir öznitelik seçim metodu önerilmiştir. Bu, önerilen metodun hem sınıf bazlı olasılıklardan hem de koleksiyon bazlı olasılıklardan bir terimin önemini hesapladığı anlamına gelmektedir.

Özniteliğin sınıf bazlı ayırt ediciliği göz önüne alındığında;

1) Terimin sınıf içindeki tüm dokümanlarda bulunması daha muhtemel ise, terimin ilgili sınıf için ayırt ediciliği artar; bu nedenle yüksek bir skor atanmalıdır.
2) Eğer terim ilgili sınıfın bazı dokümanlarında bulunmuyorsa, terimin ilgili sınıf için ayırt ediciliği azalır; bu yüzden düşük skor atanmalıdır.
3) Terimin ilgili sınıftan başka diğer sınıflarda olması muhtemel ise, ilgili sınıf için terimin ayırt ediciliği azalır; bu nedenle düşük bir skor atanmalıdır.

Yukarıdaki gereksinimlere dayanarak, terimin sınıf bazlı ayırt edicilik skoru aşağıdaki gibidir:

$$EFS_{sınıf-bazlı} = \left(\frac{P(t|C_j)}{P(\bar{t}|C_j) + P(t|\bar{C_j}) + 1}\right) \quad \textbf{(6.1)}$$

Bu nedenle ilgili sınıfta yer alan dokümanlarda terim yer alıyorsa yüksek skor verilmesi gerekmektedir. Benzer şekilde ilgili sınıftaki bazı dokümanlarda terim yer almıyorsa düşük skor verilmesi gerekmektedir. Ayrıca terim ilgili sınıfın dışında başka sınıflarda geçiyorsa, düşük skor verilmelidir. Adil bir skor ataması için, hesaplanan skorları pay ve paydaya ekledik. Eğer hesaplanan skor, terimin ayırt ediciliğini arttıracaksa paya, azaltacaksa paydaya yazdık. $P(t|C_j)$, $P(\bar{t}|C_j)$, ve $P(t|\overline{C_j})$ olan bu üç olasılık, sırasıyla yukarıdaki üç gereksinimi karşılar. Bu formülden, terim sadece bir sınıfın tüm belgelerinde mevcutsa, en yüksek skor olarak 1.0 skoru atanacaktır. Bazı durumlarda paydanın 0 olmasını önlemek için formülde paydaya "+1" ekledik.

Özniteliğin koleksiyon bazlı ayırt ediciliği göz önüne alındığında;

1) Terimin olması durumunda ilgili sınıfın olma olasılığı yüksekse, koleksiyon bazlı ayırt edicilik artar; bu yüzden yüksek skor atanmalıdır.
2) Terimin olması durumunda ilgili sınıfın olma olasılığı biraz düşükse, koleksiyon bazlı ayırt edicilik azalır; bu yüzden düşük skor atanmalıdır.
3) Terimin olmaması durumunda ilgili sınıfın olması muhtemel ise, koleksiyon bazlı ayırt edicilik azalır; bu yüzden düşük skor atanmalıdır.

Yukarıdaki gereksinimlere dayanarak, özniteliğin koleksiyon bazlı ayırt edicilik skoru aşağıdaki gibidir:

$$EFS_{koleksiyon-bazlı} = \left(\frac{P(C_j|t)}{P(\overline{C_j}|t) + P(C_j|\bar{t}) + 1}\right) \quad \textbf{(6.2)}$$

Bu nedenle, öznitelik olması durumunda ilgili sınıfın olma olasılığı artarsa yüksek skor atanması gerekir. Benzer şekilde, terim olması durumunda ilgili sınıfın olasılığı düşükse, düşük skor verilmesi gerekir. Ayrıca, özniteliğin olmaması durumunda ilgili sınıfın olasılığı yüksekse, düşük skor atanır. Yüksek bir skor atamak için, hesaplanan değeri formülün payına eklerken, düşük bir skor atamak için formülün paydasına ekledik. $P(C_j|t)$, $P(\overline{C_j}|t)$, and $P(C_j|\bar{t})$ olan bu üç olasılık, sırasıyla yukarıdaki üç gereksinimi karşılar. Bu formülden, bir terimin tüm oluşumları ilgili sınıfta bulunursa, en yüksek skor olan 1.0 atanır. Bazı durumlarda paydanın 0 olmasını önlemek için formülde paydaya "+1" ekledik.

Özniteliğin sınıf bazlı ve koleksiyon bazlı ayırtediciliği göz önüne alındığında; önerilen yöntemin Kapsamlı Özellik Seçici (EFS=Extensive Feature Selector) formülü aşağıdaki gibidir:

$$EFS(t) = \sum_{j=1}^{M} \left(\left(\frac{P(t|C_j)}{P(\bar{t}|C_j) + P(t|\overline{C_j}) + 1} \right) \cdot \left(\frac{P(C_j|t)}{P(\overline{C_j}|t) + P(C_j|\bar{t}) + 1} \right) \right) \quad \textbf{(6.3)}$$

EFS metodunun formülü sınıf bazlı ve koleksiyon bazlı iki bölümden oluşur. Her iki parçanın değer aralıkları 0 ile 1 arasındadır. Dolayısıyla özniteliğin nihai skoru, 0.0 ile 1.0 arasında bir değer olacaktır. Öznitelik yalnızca tek bir sınıfın tüm dokümanlarında varsa, formülün her iki bölümünün değerleri de "1.0" olur. Dolayısıyla, formülün iki parçasını çarparak öznitelik için en yüksek skor 1.0 olur. Ancak, öznitelik bir sınıfın tüm dokümanlarında ve bazı sınıflarda ortaya çıkarsa, hem sınıfa dayalı hem de terime dayalı skorlar azalır. Ayrıca, sınıf bazlı değer 0 olduğunda, koleksiyon bazlı değer 0'dan farklı olamaz veya tam tersi de geçerlidir. Dolayısıyla, sınıf bazlı ve koleksiyon bazlı skorların birbirinin sönümlenmesi söz konusu değildir.

EFS, bazı yönlerden CMFS ve DFS'ye benzer. Yöntem, sınıf temelli parçanın ve koleksiyon bazlı parçanın payını çarparak CMFS'e benzer. Ayrıca, yöntem sınıf temelli parçanın paydası ve terim tabanlı parçanın payı ile DFS'e benzer. Ancak, EFS formülü kapsamlı çalışmalardan sonra açığa çıkarılmıştır.

EFS kapsamlı bir şekilde incelenmiştir. Hem sınıf bazlı olasılıklar hem de koleksiyon bazlı olasılıklar dikkate alınarak özniteliğin daha ayırt edici hale gelmesini sağlayacak bir skor ataması yapılmıştır. Nihai skor, sınıf temelli ve koleksiyon bazlı skorlar çarpılarak elde edilmiştir. EFS, literatürdeki öznitelik seçme yöntemlerinden farklı olarak birçok olasılığı göz önünde bulundurarak ayırt edici öznitelikleri seçmektedir.

EFS'nin nasıl çalıştığını göstermek için Tablo 6.1.'de bir örnek koleksiyon sunulmaktadır. Ancak, EFS tarafından atanan skorlar ve öznitelik oluşumları hakkında bilgi Tablo 6.2.'de verilmiştir.

Tablo 6.1. *Örnek koleksiyon*

Sınıf etiketi	Doküman adı	İçerik
C1	Doc1	cat dog
C2	Doc2	cat fish dog mouse
C2	Doc3	cat fish mouse
C3	Doc4	cat mouse
C3	Doc5	cat

Tablo 6.2. *Öznitelik oluşum bilgisi ve atanan skorlar*

Öznitelik	Doküman frekansı	Sınıf oluşum sayısı	Skor
fish	2	1	1.000
mouse	3	2	0.410
dog	2	2	0.341
cat	5	3	0.305

$$EFS('fish') = \left(\frac{0/1}{1/1+2/5+1} * \frac{0/2}{2/2+1/4+1}\right) + \left(\frac{2/2}{0/2+0/3+1} * \frac{2/2}{0/2+0/3+1}\right) + \left(\frac{0/2}{2/2+2/3+1} * \frac{0/2}{2/2+2/3+1}\right) = 1.000$$

$$EFS('mouse') = \left(\frac{0/1}{1/1+3/4+1} * \frac{0/3}{3/3+1/2+1}\right) + \left(\frac{2/2}{0/2+1/3+1} * \frac{2/3}{1/3+0/2+1}\right) + \left(\frac{1/2}{1/2+2/3+1} * \frac{1/3}{2/3+2/3+1}\right) = 0.410$$

$$EFS('dog') = \left(\frac{1/1}{0/1+1/4+1} * \frac{1/2}{1/2+0/3+1}\right) + \left(\frac{1/2}{1/2+1/3+1} * \frac{1/2}{1/2+1/3+1}\right) + \left(\frac{0/2}{2/2+2/3+1} * \frac{0/2}{2/2+2/3+1}\right) = 0.341$$

$$EFS("cat') = \left(\frac{1/1}{0/1+4/4+1} * \frac{1/5}{4/5+0/0+1}\right) + \left(\frac{2/2}{0/2+3/3+1} * \frac{2/5}{3/5+0/0+1}\right) + \left(\frac{2/2}{0/2+3/3+1} * \frac{2/5}{3/5+0/0+1}\right) = 0.305$$

Bu örnek koleksiyonda, yalnızca tek bir sınıfın, yani C1'in tüm dokümanlarında yer alan "fish" özniteliğine maksimum skor atanır. Ayrıca, "mouse" ve "dog" öznitelikleri iki sınıfta bulunur. Özelliklerin doküman frekansları aynı olmadığı için skorları da eşit değildir. Bu nedenle, "mouse" özniteliği, "dog" özniteliğinden daha ayırt edicidir. "cat" özniteliği, üç sınıftaki tüm dokümanlarda görüldüğü için en önemsiz özniteliktir. "fish"

özniteliği tek bir sınıfın tüm dokümanlarında bulunurken, "cat" bütün sınıfların tüm dokümanlarında bulunur. Bu örnek senaryoda "fish" özniteliğine 1.00 olan en yüksek skor atanırken, "cat" özniteliğine en düşük skor olarak 0.305 atanmıştır. Özetlemek gerekirse, EFS mantıksal olarak öznitelikleri sınıf ve koleksiyon bazlı ayırt etme gücüne göre "fish", "mouse", "dog" ve "cat" olarak sıralar. Tablo 6.3.'te görüldüğü gibi, CHI2 yöntemi EFS'ye benzer şekilde öznitelikleri sıraladı. Dolayısıyla, EFS, özniteliklere adil skor atar ve EFS, skor atama açısından iyi bir öznitelik seçme yöntemidir.

Tablo 6.3. *Her ÖS yöntemi için öznitelik skorları*

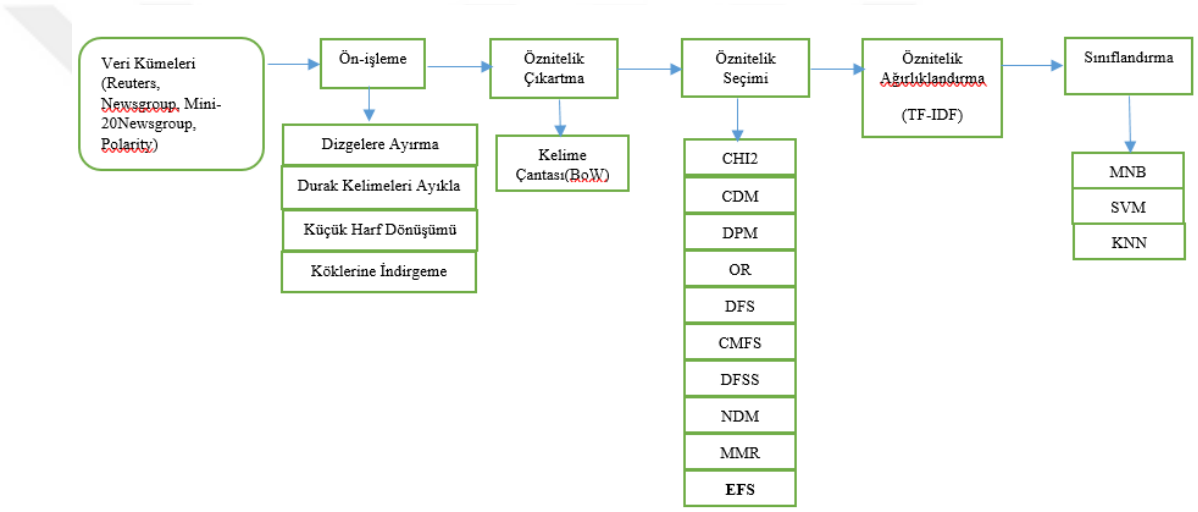
Feature	CHI2	CDM	DPM	MOR	DFS	CMFS	DFSS	NDM	MMR	EFS
fish	6.94	6.91	1.00	6.91	1.000	0.400	0.00	1000.0	5.00	1.000
mouse	4.82	1.38	0.83	6.50	0.654	0.333	1.33	2.333	2.22	0.410
dog	2.93	1.79	0.92	6.62	0.673	0.200	1.50	3.500	3.25	0.341
cat	0.00	0.00	0.00	200.0	0.500	0.360	0.00	0.000	0.00	0.305

Bu yöntemde sınıf bazlı ve koleksiyon bazlı ayırtediciliği birleştirdik. Böylece daha önemli öznitelikler elde ettik. EFS yöntemimizi diğer yöntemlerle karşılaştırdığımızda hem terim tabanlı hem de sınıf tabanlı olasılıkları ayrı ayrı hesaplayarak daha ayırt edici özellikler seçtiği söylenebilir. Elbette önerilen EFS yönteminin formülünü oluştururken kullanılan bazı olasılıklar, diğer yöntemlerin bazılarında kısmen ayrı ayrı kullanılmıştır. Bununla birlikte, önerilen yöntem EFS, formülünde belirtilen tüm olasılıkları dikkate alarak öznitelikleri seçer. Örnek koleksiyondaki ilgili skorlar, EFS yönteminin nasıl çalıştığını göstermektedir. EFS'nin performansı, deneysel çalışmalar bölümünde farklı kıyaslama veri kümelerinde test edilmiştir.

6.2.Deneysel Çalışma

EFS, öznitelik benzerliği ve doğruluk analizi açısından yukarıda belirtilen dokuz farklı ÖS teknikleriyle karşılaştırılmıştır. Bu amacı gerçekleştirmek için dört farklı veri kümesi ve iki farklı başarı ölçütü kullanılmıştır. Deneylerde kullanılan veri kümeleri ve başarı ölçütleri aşağıdaki alt bölümlerde açıklanmıştır. Daha sonra, tüm ÖS yöntemlerinin doğruluk skorları verilmeden önce seçilen ilk 10 özniteliğin benzerlik durumları sunulmuştur. Gereksiz kelimelerin atılması (Uysal & Gunal, 2014a), kök bulma (Porter, 1980) ön işleme adımları olarak gerçekleştirildi.

Bir ÖS yönteminin performansını göstermek için farklı özelliklere sahip veri kümelerinin kullanılması kaçınılmazdır. Bu nedenle deneyler, tek etiketli dört farklı veri kümesi kullanılarak gerçekleştirildi. İlk veri kümesi, MS alanında en çok kullanılan veri kümelerinden biri olan Reuters-21578'dir (Asuncion, 2007). Deneylerde ilk 10 sınıfı kullandık. İkinci veri kümesi 20Newsgroups (Asuncion, 2007) veri kümesidir ve bu veri kümesinin 10 sınıfı kullanılmıştır. Kullanılan bir diğer veri kümesi, 20Newsgroups alt kümesi kullanılarak oluşturulan Mini-20Newsgroups veri kümesidir. Son olarak kullanılan veri kümesi ise, film yorumlarını içeren Polarity veri kümesidir (Parlak & Uysal, 2020). Bu dört veri kümesi hakkında daha fazla bilgi sırasıyla Tablo 6.4.-Tablo 6.7.'de verilmektedir. Deneysel çalışmayı ise Şekil 6.1.'de gösterdik.



Şekil 6.1. *Deneysel çalışma*

Tablo 6.4.'de görüldüğü gibi, Reuters-21578 dengesiz bir veri kümesi iken, 20Newsgroups, Mini-20Newsgroups ve Polarity dengeli veri kümeleridir. Ayrıca, Reuters veri kümesi için eğitim ve test bölümleri belli olduğundan, deneylerde 20Newsgroups, Mini-20Newsgroups ve Polarity veri kümeleri manuel olarak eğitim (% 70) ve test (% 30) bölümlerine ayrılmıştır. Reuters, 20Newsgroups ve Mini-20Newsgroups çoklu-sınıf veri kümeleri iken, Polarity ise ikili-sınıf veri kümesidir. Böylece deneylerde, hem çoklu-sınıf hem de ikili-sınıf veri kümeleri kullanılmıştır.

Tablo 6.4. *Reuters-21578 veri kümesi*

No	Sınıf Etiketi	Eğitim Dokümanları	Test Dokümanları
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 6.5. *20Newsgroups veri kümesi*

No	Sınıf Etiketi	Eğitim Dokümanları	Test Dokümanları
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 6.6. *Mini-20Newsgroups veri kümesi*

No	Sınıf Etiketi	Eğitim Dokümanları	Test Dokümanları
1	alt.atheism	70	30
2	comp.graphics	70	30
3	comp.os.ms-windows.misc	70	30
4	comp.sys.ibm.pc.hardware	70	30
5	comp.sys.mac.hardware	70	30
6	comp.windows.x	70	30
7	misc.forsale	70	30
8	rec.autos	70	30
9	rec.motorcycles	70	30

10	rec.sport.baseball	70	30

Tablo 6.7. *Polarity veri kümesi*

No	Sınıf Etiketi	Eğitim Dokümanları	Test Dokümanları
1	Positive	700	300
2	Negative	700	300

Bu çalışmada ÖS yöntemlerinin performans değerlendirmesi için Makro-F1 ve Mikro-F1 ölçütü olmak üzere iki başarı ölçütü kullandık. F1 ölçütü hem kesinlik hem de duyarlılık değerlerini dikkate alır. F1-ölçütü makro-ortalamada her sınıf için ayrı ayrı hesaplanır ve daha sonra tüm sınıflar için ortalaması alınarak hesaplanır. Bu durumda, her sınıf aynı ağırlığa sahiptir. Makro-F1 aşağıdaki gibi hesaplanabilir:

$$Macro - F1 = \frac{\sum_{j=1}^{M} F_j}{M}, \qquad F_j = \frac{2 \cdot p_j \cdot r_j}{p_j + r_j} \tag{6.4}$$

Bu formülde, p_j ve r_jsırasıyla j sınıfının kesinlik ve duyarlılık değerleridir.

Bununla birlikte, F1-ölçütü mikro-ortalamada sınıf bilgisi olmadan hesaplanmaktadır. Böylece, metin koleksiyonundaki tüm sınıflandırma kararları dikkate alınır. Dengesiz veri kümelerini değerlendirirken, mikro değeri hesaplarken büyük kategoriler küçük kategorilere hükmedebilir. Bu yüzden dengesiz verilerde Mikro-F1 ölçütü adil olmayabilir. Mikro-F1 aşağıdaki gibi hesaplanabilir:

$$Mikro - F1 = \frac{2 \cdot p \cdot r}{p + r} \tag{6.5}$$

Bu formülde, p ve r, tüm sınıflar için kesinlik ve duyarlılık değerlerine karşılık gelir.

Mikro-F1 ölçütü, büyük sınıflara hakim olması nedeniyle tüm durumlar için adil bir değerlendirme sağlamayabilir. Bu nedenle, deneylerde Micro-F1'in yanında Macro-F1 ölçütü kullanılmıştır.

6.3.Öznitelik Benzerlik Analizi

Her ÖS yöntemi formülüne göre bir öznitelik alt kümesi seçer. Her özniteliğe, ayırtedici gücüne göre bir skor atanır. ÖS algoritmaları tarafından seçilen öznitelik alt kümelerinin profili, sınıflandırma performanslarıyla test edilebilir. Ayırtedici

özniteliklere herhangi bir ÖS yöntemi ile yüksek skorlar atanırsa, bu öznitelikler tarafından elde edilen sınıflandırma performansı daha yüksek olacaktır. Benzer şekilde, bilgilendirici olmayan özniteliklere herhangi bir ÖS yöntemi ile düşük skorlar atanırsa, bu öznitelikleri içermeyen öznitelik alt kümesinin sınıflandırma doğruluğu muhtemelen daha yüksek olacaktır. Aksine, ayırtedici olmayan özniteliklere herhangi bir ÖS algoritması tarafından yüksek skorlar atanırsa, bu öznitelikler tarafından elde edilen sınıflandırma performansı muhtemelen düşecektir. İlk olarak, önerilen yöntemle seçilen ilk 10 öznitelik, her bir veri kümesi için Tablo 6.8-Tablo 6.11'de diğer yöntemlerle karşılaştırılmıştır. Bu tablolarda, bir öznitelik aynı anda başka bir ÖS yöntemi tarafından seçilmediyse, bu durumda ilgili öznitelik kalın olarak gösterdik. Dolayısıyla, kalın olarak gösterilen özniteliklerin ilgili ÖS yöntemine özgü olduğu söylenebilir. Tablolardan, önerilen EFS yönteminin genel olarak karşılaştırma için kullanılan diğer 9 ÖS yöntemleriyle seçilen ortak öznitelikleri seçtiğini çıkarabiliriz. Karşılaştırma için kullanılan bu 9 yöntemin çoğunlukla metin sınıflandırması için en gelişmiş ÖS yöntemleri olduğuna dikkat edilmelidir. Bu nedenle, diğer güncel ve başarılı ÖS yöntemlerinden çok farklı olan birçok farklı öznitelik içeren bir öznitelik kümesi oluşturmak yeni bir yöntem için anlamlı olmayabilir.

Tablo 6.8. *Reuters-21578 veri kümesinde en iyi 10 öznitelik*

Metotlar	Öznitelikler
CHI2	ct, net, shr, qtr, rev, loss, profit, **note, dividend,** div
CDM	net, ct, prior, agriculture, england, **monetari,** qtr, **jan,** grain, wheat
DPM	net, ct, **export, bank, share,** wheat, ton, **company,** agriculture, **market**
OR	net, ct, agriculture, prior, wheat, grain, england, corn, qtr, profit
DFS	ct, wheat, net, oil, shr, ton, corn, barrel, qtr, agricultur
CMFS	ct, net, shr, march, qtr, **mln,** rev, **year,** loss, **dlr**
DFSS	shr, rev, qtr, barrel, ct, opec, bpd, **crude,** net, oil
NDM	rev, qtr, **taupo, payout,** ct, **zeebrugg,** wheat, bpd, net, opec
MMR	div, rev, **qtly,** qtr, ct, shr, **mth, bbl,** net, barrel
EFS	ct, wheat, net, shr, oil, barrel, qtr, march, tonn, **trade**

Tablo 6.9. *20Newsgroups veri kümesinde en iyi 10 öznitelik*

Metotlar	Öznitelikler
CHI2	basebal, forsal, atheism, auto, motorcycle, rec, sport, comp, hardwar, misc
CDM	auto, dod, forsal, **team,** basebal, rec, sport, **ride,** window, alt
DPM	comp, rec, misc, window, hardwar, sy, **refer,** ms, mac, alt
OR	cantaloup, cs, id, newsgroup, path, subject, date, messag, cmu, srv
DFS	atheism, basebal, motorcycle, forsal, auto, sport, os, ms, mac, hardwar
CMFS	cantaloup, cs, id, newsgroup, path, subject, date, messag, cmu, srv
DFSS	atheism, motorcycl, basebal, forsal, auto, **religion, atheist, bike,** sport, dod
NDM	atheism, **kmr, manti,** basebal, **powerbook, harlei, bobb, automot, schneider,** forsal
MMR	atheism, basebal, motorcycl, forsal, **pitcher, islam,** rec, auto, **livesei, solntz**
EFS	atheism, basebal, motorcycl, forsal, auto, sport, hardwar, os, sy, rec

Tablo 6.10. *Mini-20Newsgroups veri kümesinde en iyi 10 öznitelik*

Metotlar	Öznitelikler
CHI2	atheism, motorcycl, forsal, basebal, auto, sport, hardwar, sy, rec, comp
CDM	window, ibm, **mac,** rec, sy, hardwar, **game,** comp, os, misc
DPM	comp, rec, window, misc, hardwar, sy, **write,** ibm, **refer, pc**
OR	cmu, cantaloup, cs, id, newsgroup, **line,** messag, srv, subject, path
DFS	atheism, basebal, motorcycl, forsal, auto, sport, graphic, alt, os, ms
CMFS	cmu, cantaloup, cs, id, newsgroup, messag, srv, subject, path, **date**
DFSS	motorcycl, auto, forsal, basebal, bike, sport, **moral,** graphic, atheist, religion
NDM	atheism, bike, forsal, motorcycl, basebal, christian, **fido, asd,** atheist, religion
MMR	atheism, motorcycl, forsal, basebal, bike, auto, sport, atheist, religion, christian
EFS	atheism, basebal, forsal, motorcycl, auto, sport, graphic, alt, os, ms

Tablo 6.11. *Polarity veri kümesinde en iyi 10 öznitelik*

Metotlar	Öznitelikler
CHI2	bad, life, perform, great, bore, wast, world, script, worst, **plot**
CDM	magnific, outstand, uninvolv, plod, ludicr, rivet, themat, numb, seagal, tribut
DPM	bad, wast, bore, life, great, perform, worst, stupid, world, script
OR	outstand, magnific, uninvolv, ludicr, plod, rivet, themat, numb, seagal, **flawless**
DFS	wast, bad, stupid, worst, bore, **suppos,** outstand, **ridicul,** great, **excel**
CMFS	film, movi, make, time, charact, plai, scene, good, stori, end
DFSS	bad, wast, worst, stupid, bore, life, great, outstand, perform, world
NDM	magnific, outstand, uninvolv, plod, ludicr, rivet, themat, numb, seagal, tribut
MMR	outstand, wast, stupid, magnific, worst, ludicr, bore, bad, uninvolv, **balanc**
EFS	film, movi, make, time, charact, plai, scene, good, stori, end

6.4.Doğruluk Analizi

Her ÖS tekniği ile seçilen farklı sayıda öznitelik kümelerinden oluşan vektör uzayları, MNB, SVM ve KNN sınıflandırıcıları ile performansları analiz edilmiştir. Öznitelik alt kümeleri 10, 30, 50, 100, 300 ve 500 boyutları kullanılarak oluşturulmuştur. Toplam öznitelik sayısı sırasıyla Reuters-21578, 20Newsgroups, Mini-20Newsgroups ve Polarity için 16867, 50419, 14244 ve 21875'dir. Deney sonuçları Tablo 6.12-Tablo 6.22'de gösterilmektedir. En yüksek skor, karşılık gelen öznitelik boyutu için kalın yazıyla yazılırken, tüm tablo için en yüksek skoru hem kalın hem de altı çizili olarak gösterdik. Tablo 6.12-Tablo 6.22'de görüldüğü gibi, farklı veri kümelerine sahip tüm boyutlar için en iyi performansa sahip olan tek bir yöntem yoktur. Ancak, EFS tablolardaki en yüksek skor açısından diğer tüm yöntemlerden daha başarılıdır.

Reuters-21578 veri kümesi için, MNB, SVM ve KNN sınıflandırıcıları kullanılarak Mikro-F1 ve Makro-F1 açısından EFS öznitelik seçim yöntemi ile en yüksek skor elde edilmiştir. Ancak, EFS yöntemi yalnızca 20Newsgroups ve Mini-20Newsgroups veri kümeleri için MNB ve KNN sınıflandırıcıları ile en yüksek skoru elde etti. Ek olarak, Polarity very kümesi için SVM ve KNN sınıflandırıcıları için en yüksek skoru EFS metodu ile elde edilmiştir.

Tablo 6.12. *MNB sınıflandırıcısı kullanan Reuters-21578 veri kümesi için Micro-F1 ve Macro-F1 skoru*

	Micro-F1						**Macro-F1**					
Boyut	**10**	**30**	**50**	**100**	**300**	**500**	**10**	**30**	**50**	**100**	**300**	**500**
CHI2	0.553	0.830	0.874	0.912	0.922	0.926	0.074	0.422	0.527	0.634	0.681	0.689
CDM	0.612	0.792	0.831	0.897	0.900	0.905	0.219	0.419	0.487	0.651	0.652	0.655
DPM	**0.787**	0.867	0.887	0.913	0.926	**0.927**	**0.379**	0.586	0.637	0.671	0.698	0.697
OR	0.596	0.798	0.876	0.899	0.907	0.907	0.220	0.433	0.597	0.655	0.669	0.660
DFS	0.646	0.830	0.895	0.917	0.927	0.926	0.296	0.621	0.670	0.689	0.703	0.695
CMFS	0.661	0.786	0.816	0.912	0.923	0.923	0.157	0.312	0.378	0.630	0.680	0.685
DFSS	0.596	0.787	0.853	0.904	0.919	0.926	0.127	0.524	0.631	0.685	0.692	0.698
NDM	0.605	0.654	0.675	0.731	0.880	0.881	0.146	0.338	0.385	0.684	0.620	0.613
MMR	0.590	0.796	0.889	0.907	0.919	0.926	0.122	0.524	0.654	0.688	0.692	**0.702**
EFS	0.668	**0.883**	**0.901**	**0.919**	**0.928**	**0.927**	0.319	**0.649**	**0.673**	**0.693**	**0.705**	0.696

MNB sınıflandırıcısı kullanılarak Reuters-21578 veri kümesi için Mikro-F1 skorlarının aralığı 0.553 ile 0.928 arasında iken, Makro-F1 skorlarının aralığı 0.122 ile 0.705 arasındadır. MNB sınıflandırıcısı için, Reuters-21578 veri kümesindeki 300 özellik kullanılarak EFS yöntemi ile elde edilen en yüksek Mikro-F1 skoru ve Makro-F1 skoru sırasıyla 0.928 ve 0.705'tir. EFS yönteminin 30, 50, 100, 300 gibi çoğu öznitelik boyutu için de en iyi performansı gösterdiğine dikkat edilmelidir.

Tablo 6.13. *SVM sınıflandırıcısı kullanan Reuters-21578 veri kümesi için Micro-F1 ve Macro-F1 skoru*

	Micro-F1						**Macro-F1**					
Boyut	**10**	**30**	**50**	**100**	**300**	**500**	**10**	**30**	**50**	**100**	**300**	**500**
CHI2	0.761	0.843	0.880	0.919	**0.925**	0.925	0.150	0.396	0.484	0.584	0.646	0.645
CDM	0.802	0.834	0.857	0.907	0.911	0.911	0.282	0.337	0.413	0.589	0.624	0.619
DPM	0.820	0.887	0.903	0.915	**0.925**	0.926	0.341	0.526	0.583	0.618	**0.656**	0.665
OR	0.790	0.840	0.886	0.911	0.919	0.914	0.250	0.361	0.513	0.601	0.647	0.634
DFS	0.819	**0.903**	**0.914**	**0.925**	0.923	0.922	0.315	**0.587**	**0.614**	**0.631**	0.655	0.648
CMFS	0.767	0.809	0.832	0.908	0.922	0.926	0.189	0.333	0.375	0.570	0.651	0.668
DFSS	0.794	0.871	0.899	0.916	0.919	0.925	0.230	0.493	0.570	0.614	0.629	0.656
NDM	0.793	0.827	0.836	0.856	0.896	0.898	0.242	0.349	0.375	0.440	0.561	0.582
MMR	0.785	0.860	0.900	0.917	0.923	0.922	0.220	0.458	0.573	0.622	0.643	0.649
EFS	**0.836**	0.899	0.909	0.919	0.922	**0.927**	**0.377**	0.578	0.593	0.623	0.653	**0.670**

SVM sınıflandırıcısı kullanılarak Reuters-21578 veri kümesi için Mikro-F1 skor aralığı 0.761 ile 0.927 arasında iken, Makro-F1 skor aralığı 0.150 ile 0.670 arasındadır. SVM sınıflandırıcısı için, 500 öznitelik kullanılarak EFS yöntemi ile elde edilen en yüksek Mikro-F1 skoru ve Makro-F1 skoru, Reuters-21578 veri kümesi için sırasıyla 0.927 ve 0.670'tir. DFS'nin 30, 50, 100 gibi çoğu öznitelik boyutu için de en iyi performansı gösterdiğine dikkat edilmelidir. Önerilen EFS yöntemi, 10 ve 500 gibi diğer öznitelik boyutlarında en iyi performansları elde etmiştir. Ayrıca MNB sınıflandırıcısının Reuters-21578 veri kümesindeki SVM sınıflandırıcısından en yüksek performans açısından daha verimli olduğu sonucuna varılabilir.

Tablo 6.14. *KNN sınıflandırıcısı kullanan Reuters-21578 veri kümesi için Micro-F1 ve Macro-F1 skoru*

	Micro-F1						**Macro-F1**					
Boyut	**10**	**30**	**50**	**100**	**300**	**500**	**10**	**30**	**50**	**100**	**300**	**500**
CHI2	0.758	0.852	0.867	0.887	0.877	0.847	0.157	0.436	0.466	0.523	0.495	0.460
CDM	0.803	0.841	0.867	0.895	0.881	0.870	0.265	0.384	0.467	0.550	0.508	0.487
DPM	0.822	0.883	0.883	0.889	0.873	0.834	0.357	0.535	0.547	0.554	0.504	0.436
OR	0.793	0.847	0.886	0.899	0.887	0.870	0.253	0.404	0.547	0.554	0.534	0.493
DFS	0.827	**0.901**	0.900	0.901	0.897	0.878	0.322	**0.602**	0.591	0.574	0.556	0.497
CMFS	0.756	0.807	0.831	0.878	0.868	0.859	0.222	0.331	0.375	0.494	0.498	0.502
DFSS	0.795	0.870	0.901	0.908	**0.898**	0.884	0.229	0.507	0.590	**0.600**	**0.558**	0.505
NDM	0.790	0.824	0.832	0.849	0.882	0.881	0.239	0.341	0.372	0.434	0.515	0.513
MMR	0.785	0.866	0.899	**0.902**	0.896	**0.886**	0.219	0.501	0.590	0.595	0.545	**0.519**
EFS	**0.852**	0.889	**0.905**	0.893	0.879	0.840	**0.444**	0.536	**0.603**	0.552	0.518	0.455

KNN sınıflandırıcısı kullanılarak Reuters-21578 veri kümesi için Mikro-F1 skor aralığı 0.756 ile 0.905 arasında iken, Makro-F1 skor aralığı 0.157 ile 0.603 arasındadır. KNN sınıflandırıcısı için, 50 öznitelik kullanılarak EFS yöntemi ile elde edilen en yüksek Mikro-F1 skoru ve Makro-F1 skoru, Reuters-21578 veri kümesi için sırasıyla 0.905 ve 0.603'tür. DFS, DFSS ve MMR'nin 30, 100, 300 ve 500 gibi farklı özellik boyutları için de en iyi performansı gösterdiğine dikkat edilmelidir. Ayrıca, MNB sınıflandırıcısının, en yüksek performans açısından Reuters-21578 veri kümesinde SVM ve KNN sınıflandırıcısından daha verimli olduğu sonucuna varılabilir.

Tablo 6.15. *MNB sınıflandırıcısı kullanan 20Newsgroups veri kümesi için Micro-F1 ve Macro-F1 skoru*

	Micro-F1						**Macro-F1**					
Boyut	**10**	**30**	**50**	**100**	**300**	**500**	**10**	**30**	**50**	**100**	**300**	**500**
CHI2	**0.881**	0.974	0.969	0.957	0.950	0.950	**0.733**	0.948	0.939	0.917	0.905	0.905
CDM	0.715	0.920	0.959	0.943	0.938	0.930	0.466	0.834	0.920	0.891	0.883	0.867
DPM	0.846	0.971	0.970	0.955	0.939	0.931	0.694	0.944	0.941	0.914	0.885	0.870
OR	0.182	0.961	0.961	0.946	0.939	0.931	0.018	0.923	0.925	0.898	0.883	0.871
DFS	0.876	0.974	0.968	0.951	0.955	0.949	0.727	0.948	0.938	0.903	0.913	0.903
CMFS	0.182	0.972	**0.975**	0.970	0.947	0.942	0.018	0.946	**0.951**	0.942	0.898	0.890
DFSS	0.655	0.872	0.944	0.967	0.957	0.954	0.422	0.719	0.882	0.937	0.918	0.912
NDM	0.478	0.663	0.808	0.927	**0.971**	**0.959**	0.254	0.430	0.616	0.852	**0.943**	**0.920**
MMR	0.656	0.925	**0.975**	**0.973**	0.964	0.957	0.423	0.844	**0.951**	**0.948**	0.930	0.918
EFS	0.809	**0.976**	**0.975**	0.971	0.945	0.944	0.606	**0.953**	**0.951**	0.943	0.895	0.893

MNB sınıflandırıcısı kullanılarak 20Newsgroups veri kümesi için Mikro-F1 skorlarının aralığı 0.182 ile 0.976 arasında iken, Makro-F1 skorlarının aralığı 0.018 ile 0.953 arasındadır. MNB sınıflandırıcısı için, 20Newsgroups veri kümesinde 30 öznitelik kullanılarak EFS yöntemi ile elde edilen en yüksek Mikro-F1 skoru ve Makro-F1 skoru sırasıyla 0.976 ve 0.953'tür. EFS'nin en iyi performansı yalnızca 30 ve 50 olan iki öznitelik boyutunda gerçekleşmiştir.

Tablo 6.16. *SVM sınıflandırıcısı kullanan 20Newsgroups veri kümesi için Micro-F1 ve Macro-F1 skoru*

	Micro-F1						**Macro-F1**					
Boyut	**10**	**30**	**50**	**100**	**300**	**500**	**10**	**30**	**50**	**100**	**300**	**500**
CHI2	0.879	**0.986**	**0.987**	0.984	0.979	0.977	0.735	0.972	0.974	0.971	0.963	0.959
CDM	0.820	0.970	0.986	0.983	0.978	0.977	0.623	0.941	0.973	0.969	0.961	0.960
DPM	0.878	**0.986**	0.985	0.981	0.976	0.974	**0.757**	**0.973**	0.971	0.967	0.958	0.955
OR	0.182	0.983	0.986	0.983	0.978	0.975	0.018	0.966	0.973	0.969	0.960	0.956
DFS	**0.935**	0.985	**0.987**	0.985	0.980	0.975	0.857	0.971	**0.976**	**0.974**	0.965	0.956
CMFS	0.182	0.985	0.986	0.985	0.978	0.978	0.018	0.970	0.971	0.971	0.956	0.957
DFSS	0.746	0.935	0.977	0.981	0.984	0.981	0.530	0.878	0.956	0.967	0.972	0.967
NDM	0.589	0.757	0.873	0.935	**0.987**	**0.988**	0.358	0.555	0.739	0.887	**0.977**	**0.979**
MMR	0.749	0.934	**0.987**	0.985	0.986	0.982	0.535	0.873	0.975	**0.974**	0.975	0.969
EFS	0.876	**0.986**	0.986	**0.986**	0.980	0.979	0.741	0.972	0.971	0.972	0.960	0.958

SVM sınıflandırıcısı kullanılarak 20Newsgroups veri kümesi için Mikro-F1 skor aralığı 0,182 ile 0,988 arasında iken, Makro-F1 skor aralığı 0,018 ile 0,979 arasındadır. 20Newsgroups veri kümesinde SVM sınıflandırıcısı için, 500 öznitelik kullanılarak NDM yöntemi ile elde edilen en yüksek Mikro-F1 skoru ve Makro-F1 skoru sırasıyla 0.988 ve 0.979'dur. CHI2, DPM, DFS, NDM, MMR ve EFS gibi farklı yöntemlerin de farklı öznitelik boyutları için en iyi performansı gösterdiğine dikkat edilmelidir. Ayrıca, SVM sınıflandırıcısının 20Newsgroups veri kümesinde MNB sınıflandırıcısından en yüksek performans açısından daha verimli olduğu sonucuna varılabilir.

Tablo 6.17. *KNN sınıflandırıcısı kullanan 20Newsgroups veri kümesi için Micro-F1 ve Macro-F1 skoru*

	Micro-F1						**Macro-F1**					
Boyut	**10**	**30**	**50**	**100**	**300**	**500**	**10**	**30**	**50**	**100**	**300**	**500**
CHI2	0.882	0.865	0.841	0.834	0.771	0.723	0.766	0.769	0.743	0.721	0.630	0.577
CDM	0.830	0.957	0.949	0.905	0.806	0.752	0.658	0.917	0.902	0.826	0.678	0.607
DPM	0.866	0.951	0.853	0.651	0.485	0.448	0.740	0.905	0.745	0.484	0.327	0.295
OR	0.182	0.857	0.845	0.813	0.759	0.715	0.100	0.756	0.737	0.690	0.620	0.564
DFS	0.926	0.953	**0.964**	0.931	0.829	0.761	0.859	0.928	**0.932**	0.872	0.711	0.622
CMFS	0.182	0.801	0.730	0.477	0.472	0.460	0.100	0.672	0.580	0.329	0.319	0.309
DFSS	0.747	0.925	0.955	**0.959**	0.924	0.872	0.529	0.844	0.932	0.924	0.860	0.776
NDM	0.589	0.755	0.884	0.939	0.957	**0.939**	0.356	0.551	0.759	0.884	**0.918**	**0.886**
MMR	0.751	0.939	0.958	**0.959**	**0.945**	0.861	0.541	0.884	0.927	**0.929**	0.897	0.762
EFS	**0.931**	**0.962**	0.806	0.693	0.605	0.577	**0.863**	**0.936**	0.683	0.531	0.450	0.424

KNN sınıflandırıcısı kullanılarak 20Newsgroups veri kümesi için Mikro-F1 skorları aralığı 0.182 ile 0.962 arasında iken, Makro-F1 skorları aralığı 0.100 ile 0.936 arasındadır. KNN sınıflandırıcısı için, 30 öznitelik kullanılarak EFS yöntemi ile elde edilen en yüksek Mikro-F1 skoru ve Makro-F1 skoru, Reuters-21578 veri kümesi için sırasıyla 0.962 ve 0.936'tür. DFS, DFSS, NDM, MMR and EFS metotları farklı öznitelik boyutları için de en iyi performansı gösterdiğine dikkat edilmelidir. Ayrıca, SVM sınıflandırıcısının, en yüksek performans açısından 20Newsgroups veri kümesinde MNB ve KNN sınıflandırıcısından daha verimli olduğu sonucuna varılabilir.

Tablo 6.18. *MNB sınıflandırıcısı kullanan Mini-20Newsgroups veri kümesi için Micro-F1 ve Macro-F1 skoru*

	Micro-F1						**Macro-F1**					
Boyut	**10**	**30**	**50**	**100**	**300**	**500**	**10**	**30**	**50**	**100**	**300**	**500**
CHI2	**0.810**	0.964	0.955	0.953	0.951	0.942	0.607	0.931	0.913	0.910	0.907	0.889
CDM	0.814	0.957	0.940	0.936	0.905	0.897	**0.661**	0.916	0.886	0.878	0.824	0.811
DPM	0.800	0.955	0.955	0.944	0.921	0.895	0.645	0.914	0.913	0.891	0.853	0.809
OR	0.182	0.944	0.957	0.917	0.905	0.889	0.018	0.891	0.917	0.845	0.825	0.798
DFS	0.802	0.964	0.953	0.953	0.946	0.936	0.612	0.931	0.910	0.910	0.896	0.880
CMFS	0.182	0.932	**0.967**	0.949	0.934	0.919	0.018	0.848	**0.937**	0.903	0.876	0.850
DFSS	0.745	0.913	0.955	0.946	0.946	0.927	0.512	0.819	0.914	0.896	0.893	0.862
NDM	0.561	0.667	0.687	0.783	**0.960**	**0.947**	0.322	0.443	0.470	0.617	**0.923**	**0.898**
MMR	0.661	0.967	0.964	**0.960**	0.946	0.944	0.422	0.937	0.931	**0.924**	0.896	0.893
EFS	0.802	**0.971**	**0.967**	0.947	0.951	0.932	0.612	**0.944**	**0.937**	0.900	0.907	0.873

MNB sınıflandırıcısı kullanılarak Mini-20Newsgroups veri kümesi için Mikro-F1 skor aralığı 0.182 ile 0.971 arasında iken, Makro-F1 skor aralığı 0.018 ile 0.944 arasındadır. Mini-20Newsgroups veri kümesinde MNB sınıflandırıcısı için, 30 öznitelik kullanılarak EFS yöntemi ile elde edilen en yüksek Mikro-F1 skoru ve Makro-F1 skoru, sırasıyla 0.971 ve 0.944'tür. EFS'nin yalnızca 30 ve 50 olan iki öznitelik boyutu için en iyi performansı gösterdiğine dikkat edilmelidir.

Tablo 6.19. *SVM sınıflandırıcısı kulanan Mini-20Newsgroup veri kümesi için Micro-F1 ve Macro-F1 skoru*

	Micro-F1						**Macro-F1**					
Boyut	**10**	**30**	**50**	**100**	**300**	**500**	**10**	**30**	**50**	**100**	**300**	**500**
CHI2	0.817	0.973	**0.973**	0.973	0.976	0.971	0.653	0.946	0.946	0.946	0.953	0.944
CDM	0.853	0.966	0.971	0.971	0.964	0.953	0.724	0.932	0.942	0.943	0.930	0.910
DPM	0.810	0.971	0.971	0.971	0.955	0.949	0.671	0.943	0.942	0.943	0.913	0.903
OR	0.182	0.955	0.971	0.966	0.960	0.955	0.018	0.912	0.943	0.933	0.924	0.913
DFS	**0.889**	0.973	**0.973**	0.969	0.971	0.967	**0.744**	0.946	0.946	0.939	0.943	0.938
CMFS	0.182	0.934	**0.973**	**0.979**	0.973	0.951	0.018	0.877	**0.947**	**0.959**	0.946	0.907
DFSS	0.783	0.966	0.965	0.966	0.964	0.955	0.588	0.932	0.933	0.932	0.930	0.913
NDM	0.667	0.753	0.765	0.807	**0.981**	**0.980**	0.429	0.551	0.576	0.661	**0.963**	**0.960**
MMR	0.745	**0.978**	0.971	0.973	0.973	0.971	0.529	**0.956**	0.943	0.946	0.946	0.944
EFS	**0.889**	0.976	0.971	0.976	0.976	0.958	**0.744**	0.953	0.943	0.953	0.953	0.920

SVM sınıflandırıcısı kullanılarak Mini-20Newsgroups veri kümesi için Mikro-F1 skor aralığı 0.182 ile 0.981 arasında iken, Makro-F1 skor aralığı, 0.018 ile 0.963 arasındadır. SVM sınıflandırıcısı için, Mini-20Newsgroups veri kümesinde 300 öznitelik kullanılarak NDM yöntemi ile elde edilen en yüksek Mikro-F1 skoru ve Makro-F1 skoru, sırasıyla 0.981 ve 0.963'tür. CHI2, DFS, CMFS, NDM, MMR ve EFS gibi farklı yöntemlerin farklı öznitelik boyutları için de en iyi performansı gösterdiğine dikkat edilmelidir. Ayrıca, Mini-20Newsgroups veri kümesinde SVM sınıflandırıcısının MNB sınıflandırıcısından daha verimli olduğu sonucuna varılabilir.

Tablo 6.20. *KNN sınıflandırıcısı kullanan Mini-20Newsgroup veri kümesi için Micro-F1 ve Macro-F1 skoru*

	Micro-F1						**Macro-F1**					
Boyut	**10**	**30**	**50**	**100**	**300**	**500**	**10**	**30**	**50**	**100**	**300**	**500**
CHI2	0.824	0.905	0.907	0.885	0.687	0.526	0.693	0.825	0.830	0.795	0.542	0.371
CDM	0.823	0.925	0.887	0.747	0.533	0.485	0.675	0.859	0.798	0.602	0.366	0.320
DPM	0.773	0.895	0.684	0.547	0.465	0.395	0.625	0.809	0.517	0.393	0.328	0.275
OR	0.182	0.855	0.881	0.793	0.585	0.409	0.018	0.735	0.783	0.653	0.415	0.254
DFS	0.885	0.947	0.940	0.917	0.768	0.608	0.748	0.897	0.884	0.845	0.647	0.464
CMFS	0.182	0.821	0.768	0.481	0.408	0.385	0.018	0.672	0.633	0.347	0.267	0.233
DFSS	0.778	0.932	0.932	0.932	**0.885**	0.812	0.582	0.873	0.872	0.871	**0.787**	0.687
NDM	0.664	0.747	0.729	0.753	0.872	**0.872**	0.425	0.547	0.531	0.582	0.771	**0.765**
MMR	0.747	0.957	**0.946**	**0.938**	0.848	0.529	0.529	0.916	**0.895**	**0.883**	0.747	0.372
EFS	**0.887**	**0.958**	0.855	0.462	0.425	0.404	**0.751**	**0.918**	0.748	0.433	0.321	0.298

KNN sınıflandırıcısı kullanılarak Mini-20Newsgroups veri kümesi için Mikro-F1 skorları aralığı 0.182 ile 0.958 arasında iken, Makro-F1 skorları aralığı 0.018 ile 0.918 arasındadır. KNN sınıflandırıcısı için, 30 öznitelik kullanılarak EFS yöntemi ile elde edilen en yüksek Mikro-F1 skoru ve Makro-F1 skoru, Reuters-21578 veri kümesi için sırasıyla 0.958 ve 0.918'tür. DFSS, NDM, MMR ve EFS metotları farklı öznitelik boyutları için de en iyi performansı gösterdiğine dikkat edilmelidir. Ayrıca, SVM sınıflandırıcısının, en yüksek performans açısından Mini-20Newsgroups veri kümesinde MNB ve KNN sınıflandırıcısından daha verimli olduğu sonucuna varılabilir.

Tablo 6.21. *MNB sınıflandırıcısı kullanan Polarity veri kümesi için Micro-F1 ve Macro-F1 skoru*

	Micro-F1						**Macro-F1**					
Boyut	**10**	**30**	**50**	**100**	**300**	**500**	**10**	**30**	**50**	**100**	**300**	**500**
CHI2	0.852	0.851	0.875	**0.895**	0.889	0.885	**0.741**	0.740	0.778	**0.810**	0.800	0.794
CDM	0.710	0.726	0.749	0.769	0.824	0.828	0.459	0.522	0.568	0.616	0.699	0.706
DPM	0.843	0.856	0.882	0.882	0.887	0.887	0.728	0.748	0.788	0.788	0.797	0.797
OR	0.707	0.726	0.749	0.762	0.795	0.810	0.457	0.522	0.568	0.588	0.656	0.677
DFS	0.774	**0.868**	0.882	0.893	0.892	0.876	0.615	**0.767**	0.788	0.806	0.805	0.780
CMFS	0.712	0.817	0.826	0.852	0.861	0.842	0.553	0.690	0.703	0.728	0.735	0.715
DFSS	**0.844**	0.862	**0.883**	0.885	**0.893**	0.887	0.730	0.757	**0.790**	0.794	0.807	0.797
NDM	0.710	0.725	0.742	0.783	0.814	0.817	0.459	0.514	0.556	0.634	0.685	0.690
MMR	0.708	0.805	0.875	0.887	0.881	0.884	0.456	0.661	0.779	0.797	0.787	0.792
EFS	0.712	0.820	0.852	0.883	**0.893**	**0.889**	0.553	0.695	0.692	0.790	**0.808**	**0.800**

MNB sınıflandırıcısı kullanılarak Polarity veri kümesi için Mikro-F1 skor aralığı 0.707 ile 0.895 arasında iken, Makro-F1 skor aralığı 0.456 ile 0.810 arasındadır. Polarity veri kümesinde MNB sınıflandırıcısı için, 100 öznitelik kullanılarak CHI2 yöntemi ile elde edilen en yüksek Mikro-F1 skoru ve Makro-F1 skoru, sırasıyla 0.895 ve 0.810'dur. EFS'nin yalnızca 300 ve 500 olan iki öznitelik boyutu için en iyi performansı gösterdiğine dikkat edilmelidir.

Tablo 6.22. *SVM sınıflandırıcısı kulanan Polarity veri kümesi için Micro-F1 ve Macro-F1 skoru*

	Micro-F1						**Macro-F1**					
Boyut	**10**	**30**	**50**	**100**	**300**	**500**	**10**	**30**	**50**	**100**	**300**	**500**
CHI2	0.835	0.853	0.863	0.882	0.874	0.882	0.711	0.743	0.759	0.785	0.767	0.788
CDM	0.711	0.729	0.754	0.755	0.799	0.829	0.464	0.528	0.583	0.607	0.665	0.708
DPM	0.837	0.863	0.869	0.881	0.873	0.885	**0.715**	0.758	0.769	0.787	0.773	0.793
OR	0.707	0.729	0.754	0.740	0.806	0.813	0.459	0.528	0.583	0.587	0.675	0.685
DFS	**0.837**	**0.867**	0.869	0.881	0.867	0.856	**0.715**	**0.764**	0.768	0.787	0.765	0.748
CMFS	0.704	0.813	0.819	0.849	0.861	0.872	0.521	0.685	0.693	0.725	0.738	0.761
DFSS	0.834	0.859	0.865	**0.883**	**0.876**	0.864	0.709	0.753	0.762	**0.790**	0.780	0.760
NDM	0.711	0.728	0.747	0.765	0.794	0.815	0.464	0.521	0.571	0.612	0.656	0.688
MMR	0.810	0.858	**0.874**	**0.883**	0.872	0.850	0.670	0.750	**0.777**	**0.790**	0.773	0.738
EFS	0.704	0.815	0.825	0.869	0.873	**0.886**	0.521	0.688	0.702	0.769	**0.775**	**0.795**

SVM sınıflandırıcısı kullanılarak Polarity veri kümesi için Mikro-F1 skor aralığı 0.704 ile 0.886 arasında iken, Makro-F1 skor aralığı, 0.464 ile 0.795 arasındadır. SVM sınıflandırıcısı için, Polarity veri kümesinde 500 öznitelik kullanılarak EFS yöntemi ile elde edilen en yüksek Mikro-F1 ve Makro-F1 skorları, sırasıyla 0.886 ve 0.795'dir. DPM, DFS, DFSS, MMR, ve EFS gibi farklı yöntemlerin farklı öznitelik boyutları için de en iyi performansı gösterdiğine dikkat edilmelidir.

Tablo 6.23. *KNN sınıflandırıcısı kullanan Polarity veri kümesi için Micro-F1 ve Macro-F1 skoru*

	Micro-F1						**Macro-F1**					
Boyut	**10**	**30**	**50**	**100**	**300**	**500**	**10**	**30**	**50**	**100**	**300**	**500**
CHI2	0.821	0.801	0.819	0.827	0.768	0.771	0.697	0.668	0.693	0.705	0.602	0.591
CDM	0.710	0.725	0.738	0.755	0.784	**0.824**	0.459	0.511	0.547	0.583	0.668	**0.693**
DPM	**0.837**	0.801	0.821	0.828	0.763	0.709	**0.720**	0.668	0.697	0.705	0.582	0.441
OR	0.707	0.725	0.738	0.719	0.792	0.759	0.457	0.511	0.547	0.495	0.634	0.565
DFS	0.832	0.834	**0.844**	0.819	0.788	0.788	0.712	**0.715**	0.730	0.693	0.636	0.622
CMFS	0.701	0.703	0.737	0.795	0.788	0.721	0.537	0.539	0.579	0.656	0.642	0.586
DFSS	0.834	0.817	0.851	0.824	0.772	0.740	0.715	0.690	**0.740**	0.700	0.606	0.527
NDM	0.710	0.712	0.723	0.759	0.763	0.781	0.459	0.487	0.518	0.586	0.585	0.625
MMR	0.808	**0.835**	0.843	0.846	0.787	0.774	0.668	**0.715**	0.729	0.733	0.644	0.604
EFS	0.701	0.725	0.788	**0.853**	**0.793**	0.775	0.537	0.564	0.689	**0.742**	**0.682**	0.520

KNN sınıflandırıcısı kullanılarak Polarity veri kümesi için Mikro-F1 skorları aralığı 0.701 ile 0.853 arasında iken, Makro-F1 skorları aralığı 0.441 ile 0.742 arasındadır. KNN sınıflandırıcısı için, 100 öznitelik kullanılarak EFS yöntemi ile elde edilen en yüksek Mikro-F1 skoru ve Makro-F1 skoru, Polarity veri kümesi için sırasıyla 0.853 ve 0.742'dir. CDM, DPM, DFS, DFSS, MMR ve EFS metotları farklı öznitelik boyutları için de en iyi performansı gösterdiğine dikkat edilmelidir. Ayrıca, MNB sınıflandırıcısının, en yüksek performans açısından Polarity veri kümesinde SVM ve KNN sınıflandırıcısından daha verimli olduğu sonucuna varılabilir. Ayrıca, Polarity veri kümesinde, EFS metodu KNN ve SVM sınıflandırıcı için en yüksek skoru elde etmiştir.

6.5.Sonuçlar

Bu çalışmada, Kapsamlı Öznitelik Seçici (EFS=Extensive Feature Selector) olarak adlandırılan yeni bir filtre tabanlı ÖS metodu önerdik. EFS metodunun performansı, iyi bilinen dokuz adet ÖS metodlarıyla, üç farklı sınıflandırıcı kullanılarak 4 temel veri kümeleri üzerinde ve farklı öznitelik boyutları için gerçekleştirilmiştir. Deneysel sonuçlar Mikro-F1 ve Makro-F1 skorlarına göre çoğu durumda EFS metodunun performansının diğer yöntemlerden daha başarılı olduğunu göstermektedir.

7. SONUÇLAR VE GELECEK ÇALIŞMA

Bu tez çalışmasında, metin sınıflandırma problemlerinin başarımını arttırmak için çeşitli çözümler önerilmiştir.

Literatürdeki üç farklı globalleştirme tekniklerinin (MAX, SUM ve AVG) farklı özellikli veri kümeleri kullanılarak LÖS yöntemleri üzerindeki etkilerini analiz etmek için kapsamlı deneyler gerçekleştirdik. Deneylerde iki başarılı sınıflandırıcı kullanılmıştır. MCU ve MCB özellikli veri kümelerinde DFSS yöntemi OR ve CHI2 yöntemlerinden daha iyi performans sergilerken, BCB ve BCU özellikli veri kümelerinde CHI2 yöntemi OR ve DFSS yöntemlerinden daha iyi performans göstermiştir. Tüm sonuçlar dikkate alındığında en başarılı globalleştirme tekniği AVG'dir. SVM sınıflandırıcısı çoğu durumda DT sınıflandırıcısından daha başarılıdır.

MS alanı için EFS adı verilen yeni bir filtre tabanlı boyut indirgeme yaklaşımı önerdik. Ayrıca, güncel ve başarılı ÖS yönteminin kapsamlı bir analizini sunduk. Önerilen EFS yöntemi, öznitelikleri, özniteliklerin ayırt edici güçlerinin belirlenmesine katkıda bulunan sınıf tabanlı ve terim tabanlı olasılıklara göre seçer. Deney sonuçlarına göre, tüm öznitelik boyutları için en yüksek performansı elde eden bir yöntem yoktur. Bununla birlikte, EFS, çeşitli veri kümelerindeki en yüksek skorlar açısından diğer ÖS yöntemlerinden daha başarılıdır. EFS genel olarak küçük boyutlarda daha yüksek performans sergilemesi de diğer yöntemlerden daha başarılı olduğunun bir göstergesidir. EFS'nin performansı, farklı veri kümeleri, sınıflandırıcılar ve başarı ölçütleri kullanılarak, iyi bilinen dokuz filtre tabanlı ÖS tekniği ile karşılaştırılmıştır. Sonuç olarak, her veri kümesindeki en yüksek skorlar açısından EFS diğerlerinden daha başarılıdır.

Gelecekteki çalışmalar olarak, MS alanı için sınıflandırma performansını artıran yeni globalleştirme tekniği ya da teknikleri bulmaya çalışacağız. Ayrıca, farklı öznitelik seçim şemalarının performansının farklı globalleştirme teknikleriyle nasıl değiştiğini gözlemleyeceğiz. Bunlara ek olarak, EFS metin madenciliği problemlerindeki diğer alanlara uygulanacaktır.

KAYNAKÇA

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. *Mining text data*, 163-222.

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.

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.

Akın, A. A., & Akın, M. D. (2007). Zemberek, an open source nlp framework for turkic languages. *Structure, 10*, 1-5.

Asuncion, A. &. (2007). *UCI machine learning repository*.

Bhowmick, A., & Hazarika, S. M. (2018). E-mail spam filtering: a review of techniques and trends. *Advances in Electronics, Communication and Computing*, 583-590.

Can, F., Kocberber, S., Balcik, E., Kaynak, C., Ocalan, H. C., & Vursavas, O. M. (2008). Information retrieval on Turkish texts. *Journal of the American Society for Information Science and Technology, 59*(3), 407-421.

Chang, Y. C., Hsieh, Y. L., Chen, C. C., & Hsu, W. L. (2017). A semantic frame-based intelligent agent for topic detection. *Soft Computing, 21*(2), 391-401.

Chen, C. M., Lee, H. M., & Chang, Y. J. (2009). Two novel feature selection approaches for web page classification. *Expert systems with Applications, 36*(1), 260-272.

Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications, 36*(3), 5432-5435.

De Mántaras, R. L. (1991). A distance-based attribute selection measure for decision tree induction. *Machine learning, 6*(1), 81-92.

Debole, F., & Sebastiani, F. (2004). Supervised term weighting for automated text categorization. *Text mining and its applications*, 81-97.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research, 3*(Mar), 1289-1305.

Ghareb, A. S., Bakar, A. A., & Hamdan, A. R. (2016). Hybrid feature selection based on enhanced genetic algorithm for text categorization. *Expert Systems with Applications, 49*, 31-47.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research, 3*(Mar), 1157-1182.

Hashemi, M. (2020). Web page classification: a survey of perspectives, gaps, and future directions. *Multimedia Tools and Applications*, 1-25.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *European conference on machine learning* (s. 137-142). Springer, Berlin, Heidelberg.

Kim, K., & Zzang, S. Y. (2019). Trigonometric comparison measure: A feature selection method for text categorization. *Data & Knowledge Engineering, 119*, 1-21.

Kim, K., & Zzang, S. Y. (2019). Trigonometric comparison measure: A feature selection method for text categorization. *Data & Knowledge Engineering, 119*, 1-21.

McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. *AAAI-98 workshop on learning for text categorization, 752*(1), 41-48.

Onan, A. (2018). An ensemble scheme based on language function analysis and feature engineering for text genre classification. *Journal of Information Science, 44*(1), 28-47.

Özgür, A., Özgür, L., & Güngör, T. (2005). Text categorization with class-based and corpus-based keyword selection. *International Symposium on Computer and Information Sciences* (s. 606-615). Berlin, Heidelberg: Springer.

Parlak, B., & Uysal, A. K. (2020). On classification of abstracts obtained from medical journals. *Journal of Information Science, 46*(5), 648-663.

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

Parlak, B., & Uysal, A. K. (2021). A novel filter feature selection method for text classification: EFS. *Journal of Information Science*, 0165551520930897.

Pinheiro, R. H., Cavalcanti, G. D., & Ren, T. I. (2015). Data-driven global-ranking local feature selection methods for text categorization. *Expert Systems with Applications, 42*(4), 1941-1949.

Pinheiro, R. H., Cavalcanti, G. D., Correa, R. F., & Ren, T. I. (2012). A global-ranking local feature selection method for text categorization. *Expert Systems with Applications, 39*(17), 12851-12857.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130-137.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130-137.

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

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

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.

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.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management, 24*(5), 513-523.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613-620.

Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval* (Cilt 39). Cambridge: Cambridge University Press.

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

Sjarif, N. N., Azmi, N. F., Chuprat, S., Sarkan, H. M., Yahya, Y., & Sam, S. M. (2019). SMS Spam Message Detection using Term Frequency-Inverse Document Frequency and Random Forest Algorithm. *Procedia Computer Science, 161*, 509-515.

Tan, S. (2006). An effective refinement strategy for KNN text classifier. *Expert Systems with Applications, 30*(2), 290-298.

Taşcı, Ş., & Güngör, T. (2013). Comparison of text feature selection policies and using an adaptive framework. *Expert Systems with Applications, 40*(12), 4871-4886.

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

Uysal, A. K. (2018). On two-stage feature selection methods for text classification. *IEEE Access, 6*, 43233-43251.

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

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

Uysal, A. K., & Gunal, S. (2014b). Text classification using genetic algorithm oriented latent semantic features. *Expert Systems with Applications, 41*(13), 5938-5947.

Yang, J., Liu, Y., Zhu, X., Liu, Z., & Zhang, X. (2012). A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. *Information Processing & Management, 48*(4), 741-754.

Yürekli, A. (2019). Lisansüstü Eğitim Enstitüsü Tez Yazım Kılavuzu. *Eskişehir Teknik Üniversitesi Lisansüstü Eğitim Enstitüsü*, 1-12.

Zhang, C., Wu, X., Niu, Z., & Ding, W. (2014). Authorship identification from unstructured texts. *Knowledge-Based Systems, 66*, 99-111.

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.

ÖZGEÇMİŞ

Öğrenim Bilgisi

Doktora (2016-)	ESKİŞEHİR TEKNİK ÜNİVERSİTESİ MÜHENDİSLİK FAKÜLTESİ/BİLGİSAYAR MÜHENDİSLİĞİ/BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI Tez adı: Metin Sınıflandırma İçin Boyut İndirgeme Tez Danışmanı:(Doç. Dr. Alper Kürşat UYSAL)
Yüksek Lisans (2014-2016)	ANADOLU ÜNİVERSİTESİ MÜHENDİSLİK FAKÜLTESİ/BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ/BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI Tez adı: Tıbbi Dokümanların Hastalıklara Göre Sınıflandırılması Tez Danışmanı:(Doç. Dr. Alper Kürşat UYSAL)
Lisans (2007-2012)	KOCAELİ ÜNİVERSİTESİ MÜHENDİSLİK FAKÜLTESİ/BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ/BİLGİSAYAR MÜHENDİSLİĞİ PR.

Görevler

ARAŞTIRMA GÖREVLİSİ 2020	AMASYA ÜNİVERSİTESİ/TEKNOLOJİ FAKÜLTESİ/BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ/BİLGİSAYAR BİLİMLERİ ANABİLİM DALI)
ARAŞTIRMA GÖREVLİSİ 2018-2020	ESKİŞEHİR TEKNİK ÜNİVERSİTESİ/MÜHENDİSLİK FAKÜLTESİ/BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ/BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI)
ARAŞTIRMA GÖREVLİSİ 2014-2018	ANADOLU ÜNİVERSİTESİ/MÜHENDİSLİK FAKÜLTESİ/BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ/BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI)
ARAŞTIRMA GÖREVLİSİ 2012-2014	AMASYA ÜNİVERSİTESİ/TEKNOLOJİ FAKÜLTESİ/BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ/BİLGİSAYAR BİLİMLERİ ANABİLİM DALI)

Eserler

A. Uluslararası hakemli dergilerde yayımlanan makaleler:

1. PARLAK BEKİR, UYSAL ALPER KÜRŞAT (2021). A Novel Filter Feature Selection Method For Text Classification: Extensive Feature Selector. JOURNAL OF INFORMATION SCIENCE

2. PARLAK BEKİR, UYSAL ALPER KÜRŞAT (2020). The Effects Of Globalisation Techniques On Feature Selection For Text Classification. JOURNAL OF INFORMATION SCIENCE

3. PARLAK BEKİR, UYSAL ALPER KÜRŞAT (2019). On Classification Of Abstracts Obtained From Medical Journals. JOURNAL OF INFORMATION SCIENCE

B. Ulusal/Uluslararası bilimsel toplantılarda sunulan ve bildiri kitaplarında (proceedings) basılan bildiriler:

1. PARLAK BEKİR, UYSAL ALPER KÜRŞAT (2015). Classification of medical documents according to diseases. 23nd signal processing and communications applications conference (siu) (Tam Metin Bildiri/Poster Sunum)

2. PARLAK BEKİR, UYSAL ALPER KÜRŞAT (2016). The impact of feature selection on medical document classification. 11th Iberian Conference on Information Systems and Technologies (CISTI), 1-5. (Tam Metin Bildiri/Sözlü Sunum)

Click or tap here to enter text.