DENGESİZ METİN SINIFLANDIRMA İÇİN YENİ YAKLAŞIMLAR

Doktora Tezi

Hande TİRYAKİ

Eskişehir 2023

DENGESİZ METİN SINIFLANDIRMA İÇİN YENİ YAKLAŞIMLAR

Hande TİRYAKİ

Doktora Tezi

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Bilimleri Bilim Dalı

Danışman: Doç. Dr. Alper Kürşat UYSAL

Eskişehir

Eskişehir Teknik Üniversitesi

Lisansüstü Eğitim Enstitüsü

Ağustos 2023

.

JÜRİ VE ENSTİTÜ ONAYI

Hande TİRYAKİ'nin DENGESİZ METİN SINIFLANDIRMA İÇİN YENİ YAKLAŞIMLAR başlıklı çalışması 01/08/2023 tarihinde aşağıdaki jüri tarafından değerlendirilerek "Eskişehir Teknik Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği"nin ilgili maddeleri uyarınca, Bilgisayar Mühendisliği Anabilim dalında Doktora Tezi olarak kabul edilmiştir.

Jüri Üyeleri	**Unvan Adı Soyadı**	**İmza**
Üye	: Prof. Dr. Gürkan ÖZTÜRK	
Üye	: Doç. Dr. Alper Kürşat UYSAL	
Üye	: Doç. Dr. Ahmet ARSLAN	
Üye	: Doç. Dr. Mehmet ŞİMSEK	
Üye	: Dr. Öğr. Üyesi Rasim ÇEKİK	

Prof. Dr. Semra KURAMA

Lisansüstü Eğitim Enstitüsü Müdürü

01/08/2023

DANIŞMAN ONAYI

Danışmanlığını yürüttüğüm Doktora öğrencisi Hande TİRYAKİ, DENGESİZ METİN SINIFLANDIRMA İÇİN YENİ YAKLAŞIMLAR başlıklı tez çalışmasını tamamlamıştır. Hazırlamış olduğu tez tarafımca incelenmiş ve öğrencinin tez savunma sınavına alınması bilimsel ve etik açıdan uygun görülmüştür.

Tez Danışmanı

Doç. Dr. Alper Kürşat UYSAL

ÖZET

DENGESİZ METİN SINIFLANDIRMA İÇİN YENİ YAKLAŞIMLAR

Hande TİRYAKİ

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Bilimleri Bilim Dalı

Eskişehir Teknik Üniversitesi, Lisansüstü Eğitim Enstitüsü, Ağustos 2023

Danışman: Doç. Dr. Alper Kürşat UYSAL

Metin verilerinin sınıflar arasında genellikle dengesiz bir dağılımı vardır. Bu durum, sınıflandırıcıların dengesiz veri kümelerinde küçük kategoriler üzerinde kötü performansa sahip sınıflandırma eğilimi göstermelerine neden olmaktadır. Bunun sonucu olarak, metin sınıflandırma dengesiz sınıf probleminden oldukça etkilenen bir süreçtir. Literatürde, dengesiz metin sınıflandırma üzerine birçok çalışma yapılmıştır ve bu konu halen popüler bir araştırma alanıdır. Metin sınıflandırma sürecinin önemli aşamalarından biri olan öznitelik seçimi aşaması, dengesiz metin sınıflandırma problemi için de önemlidir. Bu tez çalışmasında, metin sınıflandırma için öznitelik seçme problemleri ile popüler öznitelik seçme yöntemlerinin sundukları çözümler geniş kapsamlı olarak analiz edilmiş ve öznitelik seçme aşamasına yönelik olarak çeşitli çözümler önerilmiştir. Bu amaçla, ilk olarak öznitelik seçme yöntemlerinin dengesiz metinlerin sınıflandırılması üzerindeki etkisi ayrıntılı olarak incelenmiştir. Bu doğrultuda, iki farklı veri setinde üç farklı sınıflandırıcı ve dokuz farklı öznitelik seçme yöntemi ile birçok deney gerçekleştirilmiştir. Ayrıca, farklı öznitelik sayıları kullanılarak öznitelik seçme yöntemlerinin başarısı gözlemlenmiştir. Aynı zamanda dengesiz metin sınıflandırma için iki yeni öznitelik seçme yöntemi (EFS_IMP1 ve EFS_IMP2) önerilmiştir. Bu yöntemler, Kapsamlı Öznitelik Seçici (EFS) adlı yeni bir öznitelik seçme yönteminden türetilmiştir. EFS_IMP1 ve EFS_IMP2 yöntemlerinin performanslarının karşılaştırması, filtre tabanlı altı öznitelik seçme yöntemli ile gerçekleştirilmiştir. Üç referans dengesiz metin veri seti, Destek Vektör Makineleri (SVM), Karar Ağacı (DT), Rastgele Orman (RF) ve K-En Yakın Komşular (kNN) sınıflandırıcıları ile kullanılmıştır. Deneysel sonuçlar, EFS_IMP1 ve EFS_IMP2'nin dengesiz metin sınıflandırma için Makro-F1'e göre diğer öznitelik seçme yöntemleri ile karşılaştırıldığında üstün veya karşılaştırabilir performans sunduğunu göstermiştir.

Anahtar Sözcükler: Dengesiz metin sınıflandırma, Boyut indirgeme, Öznitelik seçimi, Terim seçimi

ABSTRACT

NEW APPROACHES TO IMBALANCED TEXT CLASSIFICATION

Hande TİRYAKİ

Department of Computer Engineering

Programme in Computer Science

Eskişehir Technical University, Institute of Graduate Programs, August 2023

Supervisor: Assoc. Prof. Dr. Alper Kürşat UYSAL

The distribution of text data across classes is often imbalanced. This condition leads to classifiers tending to perform poorly on smaller categories within imbalanced data sets. As a result, text classification is a process significantly affected by the imbalanced class problem. The feature selection stage, one of the crucial stages of the text classification process, is also important for the imbalanced text classification problem. In this thesis, the problems of feature selection for text classification and the solutions offered by popular feature selection methods are extensively analyzed, and various solutions are proposed for the feature selection stage. To this end, firstly, the effect of feature selection methods on the classification of imbalanced texts is thoroughly examined. In this direction, many experiments were carried out with three different classifiers and nine different feature selection methods on two different data sets. Additionally, the success of feature selection methods has been observed using different numbers of features. Also, two new feature selection methods (EFS_IMP1 and EFS_IMP2) were proposed for imbalanced text classification. These methods are derived from a recent feature selection method called Extensive Feature Selector (EFS). The performance comparison of EFS_IMP1 and EFS_IMP2 methods was carried out with six filter-based feature selection methods. Three benchmark imbalanced text data sets were employed with Support Vector Machines (SVM), Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbors (kNN) classifiers. Experimental results showed that EFS_IMP1 and EFS_IMP2 offer superior or comparative performance compared with other feature selection methods based on Macro-F1 for imbalanced text classification.

Keywords: Imbalanced text classification, Dimension reduction, Feature selection, Term selection

TEŞEKKÜR

Doktora tezimin tamamlanma sürecinde emeği geçen herkese en içten teşekkürlerimi sunuyorum. İlk olarak, bana eşsiz bilgi ve deneyimlerini aktarmaktan çekinmeyen değerli danışmanım Sayın Doç.Dr. Alper Kürşat UYSAL'a derin bir minnettarlık duyuyorum. Bilgelikleri, sabırlı rehberliği ve desteği olmasaydı bu çalışma bu şekliyle mevcut olamazdı. Teşekkürlerimi aynı zamanda, çalışmamın ilerlemesi ve kalitesi üzerindeki kritik etkileri için jüri üyesi hocalarıma da sunuyorum. Bilgi ve deneyimlerini benimle paylaştıkları için Sayın Prof.Dr. Gürkan ÖZTÜRK ve Sayın Doç.Dr. Ahmet ARSLAN'a teşekkürlerimi sunuyorum.

Bunun yanında, hayatımın her döneminde bana sonsuz sevgi, destek ve anlayış gösteren annem ve babama teşekkür ederim. Çalışmalarımın çoğu zaman uzun ve zorlu olduğu dönemlerde bile beni hep cesaretlendirdiler ve motivasyonumu kaybetmememi sağladılar. Onların güçlü karakterleri ve sürekli desteği, kişisel ve akademik başarılarımda büyük bir etkiye sahiptir.

Son olarak, bu süreçte benim yanımda olan ve beni her zaman destekleyen sevgili Nurcan'a teşekkür ederim. Nurcan'ın sevgisi, anlayışı ve sürekli desteği, bu tezin tamamlanmasında önemli bir rol oynadı. Her zaman bana güven verdi ve başarabileceğime inandı. Bu yolculukta bana eşlik ettiği için ona minnettarım.

Bu tez, başından sonuna kadar beni destekleyen, bana yardımcı olan ve beni teşvik eden herkese adanmıştır. Sizler olmadan, bu başarı mümkün olmazdı. Kalbimin en derinlerinden teşekkür ederim.

Hande TİRYAKİ

01/08/2023

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.

Hande TİRYAKİ

İÇİNDEKİLER

Sayfa

TABLOLAR DİZİNİ

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

SİMGELER VE KISALTMALAR DİZİNİ

BA	:	Dengeli Doğruluk Ölçütü (Balanced Accuracy Measure)
BoW	:	Kelime Çantası (Bag-of-Words)
CHI2	:	Ki-kare (Chi-Square)
CICI	:	Sınıf-Indeks Korpus-Indeks Ölçütü (Class-Index Corpus-Index Measure)
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)
DT	:	Karar Ağacı (Decision Tree)
EFS	:	Kapsamlı Öznitelik Seçici (Extensive Feature Selector)
EFS_IMP1	:	Kapsamlı Öznitelik Seçici Geliştirme 1 (Extensive Feature Selector Improvement 1)
EFS_IMP2	:	Kapsamlı Öznitelik Seçici Geliştirme 2 (Extensive Feature Selector Improvement 2)
GR	:	Kazanım Oranı (Gain Ratio)
GINI	:	Gini Katsayısı (Gini Index)
GSS	:	Galavotti-Sebastiani-Simi Öznitelik Seçici (Galavotti-Sebastiani-Simi Feature Selector)
IDF	:	Ters Doküman Frekansı (Inverse Document Frequency)
IG	:	Bilgi Kazancı (Information Gain)
kNN	:	K-En Yakın Komşular (K-Nearest Neighbors)
MDFS	:	Değiştirilmiş DFS (Modified DFS)
MMR	:	Maks-min Oranı (Max-min Ratio)
MNB	:	Çok Terimli Naïve Bayes (Multinomial Naïve Bayes)
NDM	:	Normalleştirilmiş Fark Ölçütü (Normalized Difference Measure)
OR	:	Olasılık Oranı (Odds Ratio)
PFS	:	Olasılıksal Öznitelik Seçimi (Probabilistic Feature Selection)
POISSON	:	Poisson Dağılımından Sapma (Deviation from Poisson Distribution)
RF	:	Rastgele Orman (Random Forest)

SVM	:	Destek Vektçr Makinesi (Support Vector Machine)
TF	:	Terim Frekansı (Term Frequency)
TF-IDF	:	Terim Frekansı – Ters Doküman Frekansı (Term Frequency – Inverse Document Frequency)

1. GİRİŞ

İnternetin gelişmesi ve yaygınlaşması, hayatımızı büyük ölçüde kolaylaştırmakla birlikte, hayatımıza bazı zorluklar da getirdiği görülmektedir. İnternet ortamında veri miktarındaki hızlı artış, bilgi edinme ve veri analizi süreçlerinde çeşitli problemlerin ortaya çıkmasına neden olmaktadır. İnternet üzerindeki verilerin büyük bir kısmı metin verilerinden oluşmaktadır. Bu veriler yararlı ve yararsız olmak üzere iki ana kategoriye ayrılabilir. Yararlı veriler, insanlar için değerli bilgiler içerirken, yararsız veriler ise önemsiz ya da gereksiz bilgiler barındırmaktadır. Veri miktarının sürekli olarak büyümesi ve hızla gelişen teknoloji, manuel olarak metin verilerinin ayrıştırılmasını neredeyse imkânsız hale getirmektedir. Bu nedenle otomatik metin sınıflandırma yöntemleri, veri analizi ve bilgi işlem süreçlerinde önemli bir yere sahip olmuştur. Otomatik metin sınıflandırma yöntemlerinin geliştirilmesi ve kullanılması, veri analizi süreçlerini hızlandırmanın yanı sıra, daha doğru ve tutarlı sonuçlara ulaşılmasını sağlamaktadır.

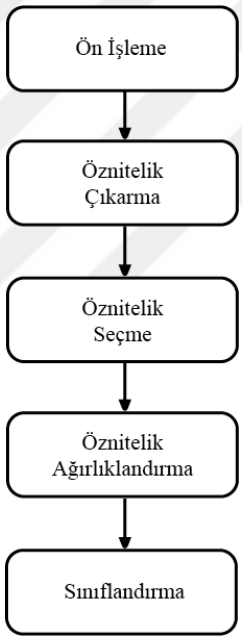
Otomatik metin sınıflandırma yöntemleri, farklı sektörler ve uygulama alanları için büyük önem taşımaktadır. Örneğin, internet üzerindeki haber ve blog yazılarının sınıflandırılması, kullanıcılara daha özelleştirilmiş ve ilgi alanlarına yönelik içerik sunulmasına olanak tanımaktadır. Ayrıca, sosyal medya platformlarında paylaşılan metin verilerinin sınıflandırılması, duygu analizi (Uysal & Uysal, 2022) ve kullanıcı eğilimlerinin belirlenmesi gibi amaçlarla kullanılmaktadır. E-posta filtreleme (Bhowmick & Hazarika, 2018)ve spam engelleme (Sjarif, ve diğerleri, 2019) uygulamaları da otomatik metin sınıflandırma yöntemlerinden faydalanmaktadır. Bu sayede, kullanıcıların önemli e-postalarını daha rahat ve hızlı bir şekilde yönetmeleri mümkün olmaktadır. Otomatik metin sınıflandırma aynı zamanda müşteri hizmetleri, pazar araştırması ve tıbbi metin analizi gibi alanlarda da doğru, güvenilir ve etkili sonuçlar vermektedir. Otomatik metin sınıflandırma yöntemlerinin geliştirilmesi ve uygulanması, zaman ve maliyet açısından büyük avantajlar sağlamaktadır. Ayrıca, bu yöntemlerin kullanılması, insan hatalarının ve önyargıların minimize edilmesine de yardımcı olmaktadır.

İnternetten elde edilen bu metin verilerinden oluşturulan veri setleri çoğunlukla dengesiz sınıf dağılımına sahiptir. Bir veri setindeki dengesizlik, bazı sınıfların büyük sayıda örnekle (çoğunluk sınıfları), diğer sınıfların ise sadece birkaç örnekle (azınlık sınıfları) temsil edildiği bir durumu ifade eder. Dengesizlik, makine öğrenme modelleri

için zorluklar yaratmaktadır. Çoğu standart model, sınıfların kabaca eşit olarak temsil edildiği zaman en iyi performansı göstermek üzere tasarlanmıştır ve bu durum söz konusu olmadığında kötü performans göstermektedir. Bu nedenle, metin sınıflandırma sürecinde dengesiz veri setleri problemi kritik öneme sahiptir.

1.1. Metin Sınıflandırma

Metin sınıflandırma, genellikle metin verilerinin çeşitli kullanım durumları için önemli olan belgelerin bir dizi önceden tanımlanmış kategoriye atanmasını içermektedir. Bir metin sınıflandırma işlemi, Şekil 1.1'de de belirtildiği üzere, ön işleme, öznitelik çıkarma, öznitelik seçme, öznitelik (terim) ağırlıklandırma ve sınıflandırma evrelerini içerir. Bu süreç boyunca gerçekleşen her adım, sonraki alt bölümlerde ayrıntılı bir şekilde ele alınacaktır.



Şekil 1.1. *Metin Sınıflandırma Süreci*

Bu sürecin ilk adımı, belgeleri ön işlemeye tabi tutmaktır. Ön işleme, metin verilerinin daha sonraki analiz için hazırlanmasına yardımcı olmaktadır.

Bu aşama izleyen adımları içermektedir:

- Belgeleri ayrıştırma (tokenization)
- Metinleri küçük harfe dönüştürme (case normalization)
- Durak kelimeleri ayıklama (stop word removal)

- Kelimeleri köklerine indirgeme (stemming)

İkinci adım olan öznitelik çıkarma aşamasında ise, genellikle kelime çantası modeli kullanılmaktadır (Gasparetto, Marcuzzo, Zangari, & Albarelli, 2022). Bu modelde her belge için, belgenin içinde bulunan her bir benzersiz kelime için bir öznitelik oluşturulmaktadır.

Bir sonraki aşama olan öznitelik seçiminde, hangi özniteliklerin metni en iyi şekilde temsil ettiğine karar verilir. Genellikle, çok sayıda öznitelik bulunabileceğinden, bu öznitelikler arasından seçim yapmak ve en etkili olanları belirlemek önemlidir.

Öznitelik ağırlıklandırma aşamasında ise, belirli bir belgeye ilişkin her özniteliğin ne kadar önemli olduğu belirlenmektedir. Bu genellikle Terim Frekansı (TF) veya Terim Frekansı-Ters Doküman Frekansı (TF-IDF) gibi yöntemlerle yapılmaktadır (Hajibabaee, ve diğerleri, 2022). TF, bir belgedeki bir terimin sıklığını belirtirken, TF-IDF, terimin belgedeki sıklığını ve tüm belge koleksiyonundaki nadirliğini dikkate almaktadır.

Sınıflandırma aşamasında, belirlenen öznitelikler kullanılarak belgeler önceden tanımlanmış kategorilere ayrılmaktadır. Bu aşamada genellikle makine öğrenmesi algoritmaları, özellikle denetimli öğrenme algoritmaları kullanılmaktadır. Algoritmanın seçimi genellikle veri setine, belgenin özelliklerine, belirlenen hedeflere ve kullanılabilir işlem gücüne bağlıdır.

Metin sınıflandırma, belgelerin semantik içeriklerini çözümlemek ve daha sonra belgeleri önceden belirlenmiş kategorilere ayırmak için karmaşık bir süreçtir. Bu süreç, çeşitli endüstrilerde ve uygulama alanlarında, özellikle büyük miktarda metin verisinin hızlı ve etkin bir şekilde işlenmesi gerektiğinde kritik öneme sahiptir.

1.2. Dengesiz Metin Sınıflandırmada Öznitelik Seçme Problemleri

Dengesiz metin sınıflandırma, gerçek dünyada karşılaştığımız bir durumdur ve çoğu zaman çeşitli sınıflandırma problemlerinin temelini oluşturmaktadır. Ancak, bu alanda önerilen öznitelik seçim yöntemlerinin sayısı yetersizdir.

Öznitelik seçimi, makine öğrenmesi ve metin madenciliği uygulamalarının önemli bir parçasıdır. Özellikle metin sınıflandırmada, yüksek boyutluluğu yönetmek ve en iyi sonuçları elde etmek için öznitelik seçimine önem verilmektedir. Ancak, dengesiz metin kümeleri söz konusu olduğunda, öznitelik seçim süreci daha da karmaşıklaşır. Çünkü

dengesiz veri setlerinde, bir sınıfın örnekleri diğer sınıflara göre çok daha azdır ve bu durum, sınıflandırma modelinin performansını olumsuz yönde etkilemektedir. Ayrıca, dengesiz veri setleriyle çalışırken karşılaşılan diğer zorluklar arasında, yüksek boyutluluk, veri seyrekliği ve gürültü bulunur. Bu faktörler, öznitelik seçimi yöntemlerinin etkinliğini daha da zorlaştırmaktadır.

Bu nedenle, dengesiz metin sınıflandırma için önerilen öznitelik seçim yöntemlerinin sayısının az olması, bu alanda daha fazla araştırma ve geliştirme yapılması gerektiğine işaret etmektedir. Dengesiz metin sınıflandırma, daha fazla dikkat ve kaynak gerektiren önemli bir araştırma alanıdır.

1.3. Tezin Amacı ve Katkılar

Tezin literatüre kazandırdığı katkılar ve çözümler, aşağıdaki araştırma problemlerine cevap olma niteliği taşımaktadır:

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

Literatürdeki geleneksel ve güncel öznitelik seçim metotlarının önemli veri kümeleri üzerinde uygulanması sınıflandırma başarımlarını nasıl etkiler?

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

Yukarıda maddeler halinde ortaya konulan araştırma problemlerine çözüm geliştirmek için hazırlanan bu tez çalışmasının iki tane katkısı vardır;

Tezin ilk katkısı, literatürde mevcut olan öznitelik seçim metotlarının kullanımının dengesiz metin sınıflandırma performansına etkisinin araştırılmasına yöneliktir. Bu tez çalışmasında güncel ve başarılı dokuz farklı öznitelik seçim metodunun iki farklı veri setinde üç farklı sınıflandırıcıyla farklı öznitelik sayıları kullanılarak performansları ayrıntılı bir şekilde analiz edilmiştir.

Tezin diğer katkısı, dengesiz metin sınıflandırma problemi için öznitelik seçim metodu olarak EFS_IMP1 ve EFS_IMP2 metotlarının önerilmesidir. Önerilen EFS_IMP1 ve EFS_IMP2 metotları, literatürde öne çıkan popüler altı öznitelik seçim metodu ile karşılaştırılmıştır. Deney sonuçlarına göre, önerilen EFS_IMP1 ve EFS_IMP2 metotları,

öznitelik boyutlarının çoğunluğu için sınıflandırma doğruluğu açısından diğer öznitelik seçim metotlarından daha iyi performans göstermektedir.

1.4. Tez Organizasyonu

Sekiz bölümden oluşan bu tezin birinci bölümünde dengesiz metin problemi, dengesiz metin sınıflandırmada öznitelik seçme problemi, tezin amacı ve katkıları anlatılmıştır. İkinci bölümde metin sınıflandırma sürecinin tüm aşamaları, tez kapsamında kullanılan öznitelik metotları ve sınıflandırıcıların detaylı açıklamalarıyla birlikte verilmiştir. Tez konusuyla ilgili çalışmalar üçüncü bölümde yer almaktadır. Dördüncü bölümde dengesiz metinle ilgili genel bilgiler, düzenlenen öznitelik seçme ve düzenlenen öznitelik ağırlıklandırma metotları ve dengesiz metin veri setleri hakkında bilgiler bulunmaktadır. Beşinci bölümde ise dengesiz metin sınıflandırmada öznitelik seçim yöntemlerinin etkililiği deneysel çalışmalarla analiz edilmiş ve sonuçları paylaşılmıştır. EFS_IMP1 ve EFS_IMP2 metotlarını önermeden önce yapılan deneysel çalışmalar altıncı bölümde verilmiştir. Yedinci bölümde önerilen EFS_IMP1 ve EFS_IMP2 öznitelik seçim metotları ile ilgili deneyler ve sonuçları sunulmuştur. Tez kapsamında yapılan çalışmalar ile ilgili sonuçlara ve tartışmalara ise son bölüm olan sekizinci bölümde yer verilmiştir.

2. METİN SINIFLANDIRMA SÜRECİ

Metin sınıflandırma, metin içerikli belgeleri önceden tanımlanmış kategorilere ayırma işlemidir. Metin sınıflandırma işlemi genel olarak ön işleme, öznitelik çıkarma, öznitelik seçme, öznitelik (terim) ağırlıklandırma ve sınıflandırma aşamalarından oluşmaktadır.

2.1. Ön İşleme Adımları

Metin sınıflandırmanın yaygın olarak kullanılan ön işleme aşamaları; dizgelere ayırma, durak sözcükleri ayıklama, büyük/küçük harfe dönüştürme, özel karakterleri ayıklama ve kelimeleri köklerine indirgeme aşamalarıdır. Bu ön işleme aşamalarının her biri, sonraki bölümlerde açıklanmıştır.

2.1.1. Dizgelere ayırma

Dizgelere ayırma, bir metni daha küçük parçalar olan dizgelere (token) bölme işlemidir. Dizgelere ayırma, metin verilerini analiz etmek ve işlemek için gerekli olan önemli bir adımdır. Bu işlem, genellikle bir metni kelimelere veya cümlelere ayırmayı içermektedir. Metni kelimelere veya cümlelere ayırmak, metin üzerinde daha detaylı analizler yapmayı mümkün kılmaktadır. Ayrıca, metni dizgelere ayırmak, durak sözcükleri ayıklama gibi diğer ön işleme adımları için de gerekli olan bir adımdır.

2.1.2. Durak sözcükleri ayıklama

Durak sözcükler genellikle bir dildeki en yaygın kullanılan sözcüklerdir ve genellikle bir metnin genel anlamına çok az katkıda bulunurlar. İngilizce'de "the", "is", "and" gibi sözcükler durak sözcük örneği olarak verilebilir. Türkçe'de ise "ve", "ama", "ile", "de" gibi sözcükler durak sözcüktür.

Metin sınıflandırmada, genellikle belirli bir konu veya duyguya işaret eden anahtar sözcükler tespit edilmeye çalışılmaktadır. Bu nedenle, bu tür durak sözcükler genellikle gürültü olarak kabul edilmektedir ve analizden çıkarılmaktadır.

2.1.3. Büyük/küçük harfe dönüştürme

Metin sınıflandırma sürecinde ön işleme aşamasında büyük/küçük harfe dönüştürme (case normalization) önemli bir adımdır. Bu işlem, genellikle metindeki tüm kelimeleri küçük harfe çevirir, böylece "Elma", "elma" ve "ELMA" gibi ifadelerin hepsi aynı kelime olarak kabul edilir. Bu, metin analizinde tutarlılık sağlar ve kelimelerin yanlışlıkla farklı olarak tanımlanmasını önler.

Ancak, bazı durumlarda, büyük harf kullanımı belirli bir anlam veya duygu ifade edebilmektedir, bu nedenle büyük harfe dönüştürme işlemi her durumda uygulanmamaktadır.

2.1.4. Özel karakterleri ayıklama

Özel karakterler, genellikle bir metindeki anlamı etkilemeyen veya metin sınıflandırma gibi bir görev için gereksiz olan karakterlerdir. Bu karakterler genellikle noktalama işaretleri, sayılar veya diğer non-alfabetik sembollerdir.

Özel karakterlerin ayıklanması, genellikle metnin daha temiz ve daha tutarlı hale getirilmesini sağlamaktadır. Ancak, her durumda özel karakterleri ayıklamanın uygun olup olmadığına dikkatli bir şekilde karar verilmelidir. Bazı durumlarda, örneğin Twitter verilerini analiz ederken, hashtagler (#) veya kullanıcı adları (@) gibi özel karakterler önemli bilgiler içerebilmektedir.

2.1.5. Sözcükleri köklerine indirgeme

Metin sınıflandırma sürecinde ön işleme aşamasında kullanılan tekniklerden biri de kök bulma (stemming) dır. Bu işlem, sözcükleri köklerine indirger, böylece "geliyor", "geldi" ve "gelecek" gibi ifadelerin hepsi aynı kök olan "gel" olarak kabul edilir. Bu, metin analizinde tutarlılık sağlar ve benzer anlamlı kelimelerin yanlışlıkla farklı olarak tanımlanmasını önler. Fakat Türkçe dilinde kök bulma işlemi biraz daha karmaşıktır, çünkü Türkçe sondan eklemeli bir dil olduğu için çok sayıda ek ve çekim formu vardır. Kök bulma algoritmaları, Türkçe ve İngilizce metinler için yaygın olarak kullanılan yöntemlerdir. Zemberek algoritması (Akın & Akın, 2007), Türkçe metinlerdeki kelimelerin köklerini bulmak için kullanılmaktadır. İngilizce için ise Porter algoritması (Porter, 1980) tercih edilmektedir. Bu kök bulma algoritmaları, dilin yapısını anlamak ve metinlerin daha etkili bir şekilde işlenmesini sağlamak için temel adımlardır.

2.2. Öznitelik Çıkarma

Metin sınıflandırma çalışmalarında sıkça kullanılan bir yöntem olan kelime çantası (bag of words) yaklaşımı, öznitelik çıkarma aşamasında terimlerin sırasını göz ardı etmekte ve terimlerin dokümanlardaki sıklığına odaklanmaktadır (Uysal, Günal, Ergin, & Günal, 2012). Bu sayede, bir metin koleksiyonundaki her bir farklı kelime, bir öznitelik olarak kabul edilmektedir. Böylece, her doküman, çok boyutlu bir öznitelik vektörüyle temsil edilmektedir (Uysal & Gunal, 2012). Öznitelik vektöründe, terim frekansı (TF),

terim frekansı-ters doküman frekansı (TF-IDF) gibi ağırlıklandırılmış değerler bulunmaktadır (Schütze, Manning, & Raghavan, 2008). Metinlerden öznitelik çıkarılırken, "dizgelere ayırma", "durak sözcükleri ayıklama", "küçük harfe dönüştürme", "özel karakterleri ayıklama" ve "sözcükleri köklerine indirgeme" ön işleme adımları uygulanmaktadır.

2.3. Öznitelik Ağırlıklandırma

Öznitelik ağırlıklandırma aşamasında dokümanların her bir terimi için bir terim ağırlıklandırma metodu kullanılır ve terimin ağırlığı hesaplanır. Belli dokümanlarda geçen sınıflandırmada ayırt edici özel bilgi sağlayan terimler ile tüm dokümanlarda çokça geçen özel bilgi sağlamayan terimler arasındaki farkı göstermek hedeflenir.

Buradaki ana amaç, belirli bir dokümanda geçen ve bu dokümanı veya dokümanın içeriğini tanımlamada ayırt edici bir rol oynayan terimler ile genel olarak tüm dokümanlarda sıkça geçen ve bu nedenle belirli bir özel bilgi sağlamayan terimler arasında bir fark oluşturmaktır. Yani, bir terim, belirli bir dokümanın içinde ne kadar önemliyse, bu terim o doküman için o kadar ağırlıklıdır.

Ağırlıklandırma, genellikle terim frekansı-ters doküman frekansı (TF-IDF) gibi istatistiksel teknikler kullanılarak gerçekleştirilmektedir. TF-IDF, bir terimin belirli bir dokümanda ne kadar sık geçtiğini TF ve aynı terimin diğer tüm dokümanlarda ne kadar nadir olduğunu IDF ile ölçmektedir (Manning, Raghavan, & Schutze, 2008). Bu teknik, bir terimin belirli bir dokümanda ne kadar önemli olduğunu belirlemeye yardımcı olmaktadır. Eşitlik 2.1'de *ti* teriminin TF-IDF değerini hesaplamak için kullanılan formül gösterilmektedir.

$$W_{TF-IDF}(t_i) = \mathrm{TF}(t_i, d_k) * \log\left(\frac{D}{d(t_i)}\right) \quad \textbf{(2.1)}$$

D değeri koleksiyondaki toplam doküman sayısını, $d(t_i)$ ise t_i teriminin geçtiği toplam doküman sayısını belirtmektedir.

2.4. Öznitelik Seçme

Öznitelik çıkarma aşamasından gelen özniteliklerin boyutunun çok yüksek olması durumunda hesaplama karmaşıklığını önemli ölçüde artırabilmektedir. Bu durum, işlem süresini yavaşlatmakta ve modelin genel performansını olumsuz etkilemektedir.

Öznitelik seçimi, bu sorunu hafifletmeye yardımcı olmaktadır. Bu süreç, özellikle yüksek boyutlu öznitelik uzaylarında, en bilgilendirici ve ayırt edici özniteliklerin alt kümelerini belirlemeye yardımcı olmaktadır. Bu, modelin genel performansını artırmakta ve aynı zamanda hesaplama sürelerini azaltmaktadır, çünkü modelin işlemesi gereken öznitelik sayısı önemli ölçüde azaltılmış olmaktadır.

Öznitelik seçiminin bir diğer avantajı, modelin anlaşılabilirliğini ve yorumlanabilirliğini artırmasıdır. Öznitelik seçimi hangi özniteliklerin modelin kararlarına en çok etki ettiğini belirlememize yardımcı olmaktadır. Bu durum, modelin kararlarını daha şeffaf ve anlaşılır hale getirmektedir, böylece modelin neden belirli bir tahminde bulunduğu daha net anlaşılmaktadır.

Öznitelik seçim metotları genel olarak üç ana gruba ayrılır: Filtre yöntemler, sarmalayıcılar ve gömülü teknikler. Filtre yöntemleri hesaplama hızı bakımından avantajlıdır, ancak genellikle öznitelikler arası bağımlılıkları göz ardı ederler (Uysal & Gunal, 2012). Filtre tabanlı teknikler metin sınıflandırma alanında sıklıkla kullanılan seçenekler arasındadır. Metin sınıflandırma için, belirgin özniteliklerin seçiminde kullanılan bir dizi filtre tabanlı yöntem bulunmaktadır.

Öznitelik seçme metot formülleri için notasyonlar ve açıklamaları Tablo 2.1'de verilmiştir.

Tablo 2.1. *Öznitelik Seçme Metotları için Notasyonlar*

Notasyon	Anlamı
a	Cj sınıfında t terimi geçen doküman sayısı
b	Cj sınıfında t terimi geçmeyen doküman sayısı
c	Diğer sınıflarda $(\overline{C_j})$ t terimi geçen doküman sayısı
d	Diğer sınıflarda $(\overline{C_j})$ t terimi geçmeyen doküman sayısı
e	Cj sınıfında t teriminin sayısı
f	Diğer sınıflarda $(\overline{C_j})$ t teriminin sayısı
N	Tüm sınıflardaki doküman sayısı
M	Toplam sınıf sayısı
$P(t)$	t teriminin olma olasılığı
$P(\bar{t})$	t teriminin olmama olasılığı
$P(C_j)$	Cj sınıfının olma olasılığı
$P(\overline{C_j})$	Cj sınıfının olmama olasılığı
$P(t\|C_j)$	Cj sınıfının t terimi ile olma olasılığı
$P(\bar{t}\|C_j)$	Cj sınıfının t terimi ile olmama olasılığı
$P(t\|\overline{C_j})$	Cj sınıfı dışındaki diğer sınıfların t terimi ile olma olasılığı
$P(\bar{t}\|\overline{C_j})$	Cj sınıfı dışındaki diğer sınıfların t terimi ile olmama olasılığı
$P(C_j\|t)$	t terimi mevcut olduğunda, Cj sınıfının olasılığı
$P(\overline{C_j}\|t)$	t terimi mevcut olduğunda, Cj sınıfının olmama olasılığı
$P(C_j\|\bar{t})$	t terimi mevcut olmadığında, Cj sınıfının olasılığı
$P(\overline{C_j}\|\bar{t})$	t terimi mevcut olmadığında, Cj sınıfının olmama olasılığı
$P(t,C_j)$	Cj sınıfı mevcut olduğunda, t teriminin olasılığı
$P(\bar{t},C_j)$	Cj sınıfı mevcut olduğunda, t teriminin olmama olasılığı
$P(t,\overline{C_j})$	Cj sınıfı mevcut olmadığında, t teriminin olasılığı
$P(\bar{t},\overline{C_j})$	Cj sınıfı mevcut olmadığında, t teriminin olmama olasılığı

2.4.1. Galavotti-Sebastiani-Simi öznitelik seçici

Galavotti-Sebastiani-Simi (GSS) metodu, basitleştirilmiş bir Ki-kare yaklaşımını kullanan bir öznitelik seçme algoritmasıdır. Bu metot, özniteliklere benzersiz bir şekilde puan verir; negatif puanlar üyeliği olmayanı (öznitelik sınıflandırma görevi için ilgili veya yardımcı olmayanı) ve pozitif puanlar üyeliği (öznitelik ilgili olduğu ve modele dahil edilmesi gerektiği) gösterir.

GSS metodu özellikle pozitif özniteliklere odaklanır. Yani, ilgilenilen sınıfla pozitif korelasyona sahip özniteliklere daha çok ağırlık verir (Moh'd Mesleh, 2011).

GSS metodunun formülasyonu Eşitlik 2.2'de verilmiştir:

$$GSS\left(t, C_j\right) = \frac{a}{N} * \frac{d}{N} - \frac{c}{N} * \frac{b}{N} = \frac{a * d - b * c}{N^2} \tag{2.2}$$

2.4.2. Sınıf-indeks korpus-indeks ölçütü

CICI (Class-Index Corpus-Index) belirli bir öznitelik için hem sınıf içi bilgisi hem de tüm korpus genelindeki bilgiyi dikkate almayı mümkün kılar. Bu, özellikle dengesiz veri setlerinde, bazı sınıfların diğerlerinden çok daha fazla belgeye sahip olduğu durumlarda yararlı olabilir. Bu tür durumlarda, sınıf bazlı bir yaklaşım, belirli bir sınıfa özgü olan ancak genel korpus genelinde nadir olan öznitelikleri göz ardı edebilir. CICI'nin korpus bazlı yaklaşımı, bu tür özniteliklerin belirlenmesine yardımcı olabilir (Parlak, 2022). CICI'nin formülü aşağıda belirtilmiştir:

$$CiCi\left(t\right) = \frac{\left(a * d - b * c\right)^2}{\left(a + b\right) * \left(c + d\right) * N^2} \tag{2.3}$$

2.4.3. Normalleştirilmiş fark ölçütü

NDM (Normalized Difference Measure), bir öznitelik için göreceli belge frekansını belirler, yani bir özniteliğin belirli bir sınıf içerisindeki belgelerde ne kadar sıklıkla yer aldığını, bu belgelerin genel sayısına göre hesaplar. Bu, bir özniteliğin belirli bir sınıf için ne kadar önemli olduğunu ortaya koyar. Bu değer, özniteliğin tüm belge setindeki etkisini saptamak için daha sonrasında normalleştirilir (Rehman, Javed, & Babri, 2017). NDM'nin formülü Eşitlik 2.4'de verilmiştir:

$$NDM = \frac{\left| \frac{a}{a+b} - \frac{c}{c+d} \right|}{\min\left(\frac{a}{a+b}, \frac{c}{c+d} \right)} \quad \textbf{(2.4)}$$

Payda sıfır olduğu durumda, NDM değerinin sonsuz çıkmaması için payda küçük bir değerle değiştirilir.

2.4.4. Max–min oranı

BA (Balanced Accuracy - Dengeli Doğruluk Ölçütü) (Forman, 2003) genellikle sınıflandırma görevlerinde, özellikle de sınıfların dengesiz olduğu durumlarda sıkça kullanılan bir istatistiksel metriktir. Bu ölçüm, doğru pozitif tahminlerin oranının ve doğru negatif tahminlerin oranının ortalamasını hesaplar, etkin bir şekilde hem pozitif hem de negatif sınıflara eşit ağırlık verir. Ancak, belirli durumlarda, örneğin doğru pozitiflerin ve yanlış pozitiflerin oranlarının eşit farklılık gösterdiği durumlarda, BA hem ilgili hem de ilgisiz özniteliklere eşit puan verebilir.

NDM (Normalized Difference Measure - Normalleştirilmiş Fark Ölçütü) (Rehman, Javed, & Babri, 2017), öznitelik seçimi için kullanılabilecek başka bir metottur. Ancak, büyük ve aşırı dengesiz verilerde, NDM tamamen ilgisiz seyrek özniteliklere daha yüksek puanlar verebilir.

Bu sorunları ele almak için MMR (Max-Min Ratio) geliştirilmiştir. MMR, hem BA'nın hem de NDM'nin zayıf yanlarını dikkate alır ve özniteliklerin gerçek relevansını başarılı bir şekilde belirleyebilir. MMR puanı, verilerdeki dengesizlikten etkilenmez ve özniteliklerin önemine dair daha güvenilir bir ölçüm sağlar (Rehman A. , Javed, Babri, & Asim, 2018). MMR'nin formülasyonu aşağıda verilmiştir:

$$MMR(t) = \sum_{j=1}^{M} \frac{\left| \frac{a}{a+b} - \frac{c}{c+d} \right|}{\min\left(\frac{a}{a+b}, \frac{c}{c+d} \right)} * \max\left(\frac{a}{a+b}, \frac{c}{c+d} \right) \quad \textbf{(2.5)}$$

2.4.5. Kapsamlı öznitelik seçim ölçütü

CMFS (Comprehensively Measure Feature Selection), bir özniteliğin önemini belirlemek için kullanılan hibrit bir tekniktir. Bu yaklaşım, bir özniteliğin önemini değerlendirirken hem kategori içi (intra-category) hem de kategoriler arası (inter-category) bilgileri dikkate alır.

Kategori içi bilgi, belirli bir kategorideki belgeler arasında bir özniteliğin ne kadar yaygın olduğunu belirler. Örneğin, bir kelimeden bahsedersek, bu kelimenin belirli bir kategorideki belgelerde ne sıklıkla geçtiği bu bilgiyi oluşturur.

Kategori arası bilgi ise bir özniteliğin farklı kategoriler arasında ne kadar ayrı olduğunu belirler. Yine bir kelime örneği üzerinden gidersek, bir kelimenin bir kategoride sıkça görülmesi ancak diğer kategorilerde nadiren karşılaşılması, bu kelimenin kategoriler arasında önemli bir ayrım sağladığı anlamına gelir.

Bu iki tür bilgi bir arada kullanıldığında, CMFS bir özniteliğin genel önemini belirleyebilir. Bir öznitelik hem kendi kategorisi içinde yaygınsa hem de farklı kategoriler arasında belirgin bir ayrım sağlıyorsa, bu öznitelik genellikle önemli kabul edilir. Bunun tersi durumda, bir öznitelik hem kendi kategorisi içinde nadir görülüyorsa hem de farklı kategoriler arasında belirgin bir ayrım sağlamıyorsa, bu öznitelik genellikle önemsiz kabul edilir (Yang J. , Liu, Zhu, Liu, & Zhang, 2012). CMFS'nin formülasyonu aşağıdaki gibidir:

$$CMFS(t) = \sum_{j=1}^{M} \frac{(a+b)}{N} * \frac{a}{a+b} * \frac{a}{a+c} \tag{2.6}$$

2.4.6. Ayrımcı öznitelik seçimi

DFSS (Discriminative and Semantic Feature Selection) yöntemi, geleneksel teknikleri yenilikçi stratejilerle birleştirerek kapsamlı ve etkili bir sonuç elde etmeyi hedefleyen gelişmiş bir öznitelik seçme yaklaşımıdır. Bu yöntem, geleneksel öznitelik seçme modellerine dayanır ancak öznitelik seçimini daha da rafine etmek için iki aşamalı bir süreç sunar.

DFSS yönteminin ilk aşaması ayırt edici güce odaklanır. Bu aşamada, yöntem, kategoriler veya sınıflar arasında ayırt etme yeteneğine sahip öznitelikleri belirler. Bunu, her bir özniteliğin istatistiksel özelliklerini, örneğin varyansını veya hedef değişkenle olan korelasyonunu, farklı kategoriler boyunca değerlendirerek yapar. Buradaki amaç, kategoriler arasında önemli ölçüde değişen, yüksek ayırt edici güce sahip öznitelikleri bulmaktır.

İkinci aşamada, DFSS yöntemi odak noktasını semantik karşılaştırmaya kaydırır. Sadece özniteliklerin istatistiksel özelliklerini dikkate almak yerine, aynı zamanda onların orijinal materyalle bağlamında semantik anlamını da göz önünde bulundurur. Bu,

yöntemin sadece yüksek ayırt edici güce sahip olan değil, aynı zamanda semantik olarak da orijinal materyalle alakalı veya ona karşılaştırılabilir olan öznitelikleri seçtiği anlamına gelir. Bu, seçilen özniteliklerin sadece istatistiksel olarak önemli olmalarını değil, aynı zamanda ele alınan problem bağlamında anlamlı olmalarını da sağlamaktadır.

Hem ayırt edici gücü hem de semantik alakayı dikkate alarak, DFSS yöntemi daha bütünsel bir öznitelik seçme yaklaşımı sunar (Zong, Wu, Chu, & Sculli, 2015). DFSS'nin formülü aşağıda verilmiştir:

$$DFSS\left(t,C_j\right)=\left(\frac{\frac{e}{a}}{\frac{f}{c}}\right)*\frac{a}{a+c}*\frac{a}{a+b}*\left|\frac{a}{a+c}-\frac{b}{b+d}\right| \tag{2.7}$$

2.4.7. Kapsamlı öznitelik seçici

EFS (Extensive Feature Selector) (Parlak & Uysal, 2023) yöntemi, sınıf ve koleksiyon bazlı olasılıkları bir araya getirerek öznitelik seçimi yapar. Öznitelik seçiminde, her bir özniteliğin sınıf temelli skoru, belirli bir sınıf veya kategoriye ait veri örneklerinin ne kadar iyi tanımlandığını belirler. Öte yandan, koleksiyon bazlı skor, bir özniteliğin tüm veri koleksiyonunda ne kadar etkili olduğunu değerlendirir. EFS, bu iki skoru çarparak bir özniteliğin nihai skorunu hesaplar. Bu karma yaklaşım, özniteliklerin hem belirli bir sınıftaki etkinliklerini hem de genel koleksiyon üzerindeki etkilerini dikkate alır. EFS'nin formülasyonu aşağıdaki gibidir:

$$EFS\left(t\right)=\sum_{j=1}^{M}\left(\frac{a/\left(a+b\right)}{b/\left(a+b\right)+c/\left(c+d\right)+1}\right)*\left(\frac{a/\left(a+c\right)}{c/\left(a+c\right)+b/\left(b+d\right)+1}\right) \tag{2.8}$$

2.4.8. Olasılıksal öznitelik seçimi

Bir belgenin hangi sınıfa dahil olduğunun tespitinde, belgedeki öznitelikler hayati bir rol üstlenir. Sınıflandırma sürecinde pozitif ve negatif öznitelikler belirleyici unsurlar oluşturur. Pozitif bir öznitelik, belgenin azınlıktaki bir sınıfa, yani genelde daha az rastlanan bir sınıfa ait olduğunu gösterir. Diğer taraftan, negatif bir öznitelik, belgenin çoğunluğun temsil ettiği, yani genellikle negatif bir sınıfa dahil olduğunu gösterir.

Dengesiz metin sınıflandırma meselesine bir çözüm getiren PFS (Probabilistic Feature Selection), bu durumu ele almak için geliştirilmiştir. Bu stratejinin ana hedefi,

pozitif bir öznitelik tespit etmek ve bu özniteliklerin daha hassas ve verimli bir sınıflandırmayı gerçekleştirebilmesini sağlamaktır.

PFS'nin diğer tekniklerden farklılaştığı nokta, pozitif öznitelik ölçümü için ekstra parametreler kullanma çabasıdır. Bu, pozitif özniteliklerin daha iyi tanımlanması ve genel sınıflandırma doğruluğunun artırılması amacını taşır (Pouramini, Minaei-Bidgoli, & Esmaeili, 2018). PFS formülü Eşitlik 2.9'daki gibidir:

$$PFS = \frac{\left(\frac{a}{a+b}\right)+\left(\frac{a}{a+c}\right)}{(a+b)/N} + \frac{\left(\frac{d}{d+c}\right)+\left(\frac{d}{d+b}\right)}{(c+d)/N} \quad \textbf{(2.9)}$$

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

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

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

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

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

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

2.4.9. Poisson dağılımından sapma

POISSON (Deviation from Poisson Distribution), bir olayın belirli bir zaman aralığında kaç kez gerçekleşeceğini tahmin etmek için kullanılan bir olasılık dağılımıdır. Bu durumda, terimlerin bir metinde kaç kez tekrarlandığına dair bir model oluşturmak için kullanılabilir.

Her bir terimin olasılık dağılımının standart Poisson dağılımından ne kadar saptığı, terimin önemini belirleyen bir faktör olabilir (Ogura, Amano, & Kondo, 2009). Yani, bir terim metinde beklenenden daha fazla veya daha az tekrarlanıyorsa, bu terimin belki de metnin genel anlamı üzerinde daha büyük bir etkisi olduğunu düşünülebilir.

Örneğin, bir haber makalesinde "ekonomi" kelimesi beklenenden çok daha fazla geçiyorsa, bu makalenin muhtemelen ekonomiye odaklandığını söyleyebiliriz. Standart

Poisson dağılımından sapma, bu kelimenin makaledeki önemini belirlemede bir ölçüttür. Formül adımları aşağıda gösterilmektedir:

$$\lambda = F / N \tag{2.10}$$

$$\hat{a} = (a+b) * \left(1 - e^{(-\lambda)}\right) \tag{2.11}$$

$$\hat{b} = (a+b) * e^{(-\lambda)} \tag{2.12}$$

$$\hat{c} = (c+d) * \left(1 - e^{(-\lambda)}\right) \tag{2.13}$$

$$d = (c+d) * e^{(-\lambda)} \tag{2.14}$$

$$Poisson = \frac{(a-\hat{a})^2}{\hat{a}} + \frac{(b-\hat{b})^2}{\hat{b}} + \frac{(c-\hat{c})^2}{\hat{c}} + \frac{(d-\hat{d})^2}{\hat{d}} \tag{2.15}$$

F : t teriminin tüm dokümanlardaki toplam frekansı

N : Bütün sınıflardaki toplam doküman sayısı

a+b : O sınıfa ait olan doküman sayısı

c+d :O sınıfa ait olmayan doküman sayısı

2.4.10. Ki-kare

CHI2 (Chi-Square), çok sayıda öznitelik arasından dokümanların sınıflarını daha net belirleyen öznitelikleri tespit etmek için kullanılan istatistiksel bir tekniktir.

Metin sınıflandırmada, bu metot genellikle bir sınıf ve bir öznitelik arasındaki ilişkiyi belirlemek için kullanılır (Chen & Chen, 2011). Ki-kare değeri ne kadar büyükse, o özniteliğin sınıfa bağımlılığı da o kadar yüksek kabul edilir. Ki-kare formülü Eşitlik 2.16'da belirtilmiştir:

$$CHI_2 = N * \frac{\left(a*d - b*c\right)^2}{(a+c)*(b+d)*(a+b)*(c+d)} \tag{2.16}$$

2.4.11. Bilgi kazanımı

Entropi, bir sistemin düzensizliğini veya belirsizliğini ölçen bir terimdir. Bilgi teorisi alanında bilginin belirsizliğini ifade eder. IG (Information Gain), bir özniteliğin bir veri setini sınıflara ayırma yeteneğini ölçer (Forman, 2003) ve genellikle veri ve metin madenciliği alanlarında kullanılır (Yang & Pedersen, 1997). IG, entropinin tersi olarak tanımlanabilir çünkü bir özniteliğin bilgi kazanımı ne kadar yüksekse, o öznitelik o kadar düşük entropiye sahip olur. IG formülü Eşitlik 2.17'de verilmiştir:

$$IG(t) = \frac{a}{N} * \log_2 * \frac{a*N}{(a+c)*(a+b)} + \frac{b}{N} * \log_2 * \frac{b*N}{(b+d)*(b+a)} + \frac{c}{N} * \log_2 * \frac{c*N}{(a+c)*(c+d)} + \frac{d}{N} * \log_2 * \frac{d*N}{(b+d)*(c+d)} \quad \textbf{(2.17)}$$

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

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

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

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

2.4.12. Gini katsayısı

GINI (Gini Index), IG ve GR (Kazanım Oranı - Gain Ratio) (Priyadarsini, Valarmathi, & Sivakumari, 2011) alternatifi olarak ortaya çıkmıştır (Shang, ve diğerleri, 2007). GINI, entropi yerine bir özniteliğin "saf olmayanlık" (impurity) oranını ölçer. Bir öznitelik değerinin saf olmayanlık oranı, o değerin ait olduğu sınıfın diğer sınıflarla ne kadar karışık olduğunu gösterir. GINI, bir özniteliğin tüm değerlerinin saf olmayanlık oranlarının ağırlıklı ortalamasını alır. Bu şekilde, en düşük GINI değerine sahip öznitelik, en iyi öznitelik olarak seçilir.

GINI, IG ve GR ye kıyasla daha az hesaplama karmaşıklığına sahip olmasıyla bilinir. Bu yüzden, büyük veri setleri üzerinde çalışırken, GINI kullanmak genellikle daha

hızlı ve etkilidir. Ayrıca, GINI özellikle sürekli değerli özniteliklerde, IG ve GR ye göre daha iyi sonuçlar verir. Formülü Eşitlik 2.18'deki gibidir:

$$GI(t) = \sum_{i=1}^{M} \left(\frac{a}{a+b} \right)^2 * \left(\frac{a}{a+c} \right)^2 \quad \textbf{(2.18)}$$

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

2.4.13. Ayırtedici öznitelik seçici

DFS (Distinguishing Feature Selector), metin sınıflandırma işlemleri için kullanılan bir öznitelik seçim algoritmasıdır. DFS, belirli kriterlere dayanarak, özniteliklerin önemini belirler ve önemsiz olanları ayıklar (Uysal & Gunal, 2012). Algoritma, bir koleksiyonda yer alan terimlerin sınıflarla olan ilişkilerini değerlendirir ve bu ilişkilere göre her bir terim-sınıf çifti için bir skor belirler. Yüksek skorlu bir terim-sınıf ilişkisi, terimin belirli bir sınıfı tanımlama ve ayırt etme yeteneğinin yüksek olduğunu gösterirken, düşük skorlu bir ilişki, terimin belirli bir sınıfı ayırt etme yeteneğinin düşük olduğunu gösterir. Bu öznitelik seçim yöntemi, metin sınıflandırma doğruluğuna büyük katkı sağlamıştır. DFS, Eşitlik 2.19'deki formülle ifade edilebilir:

$$DFS(t) = \sum_{j=1}^{M} \frac{(a/(a+c))}{(b/(a+b)) + (c/(c+d)) + 1} \quad \textbf{(2.19)}$$

M: sınıfların toplam sayısı

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

$\frac{b}{a+b}$: C_j sınıfında t teriminin olmama olasılığı

$\frac{c}{c+d}$: C_j hariç tüm sınıflarda t teriminin olma olasılığı

2.4.14. Değiştirilmiş ayırtedici öznitelik seçici

DFS, metin sınıflandırma için kullanılan etkili bir öznitelik seçim algoritmasıdır. Ancak, DFS'in geliştirilmiş bir versiyonu olan MDFS (Modified Distinguishing Feature Selector), metin sınıflandırma görevlerinde terim ağırlığına özel bir odaklanma sunar. Bu yenilik, TF-MDFS olarak adlandırılır ve her bir terimin metindeki sıklığına dikkat çeker.

Bu çalışma, TF-MDFS'in basit, etkili ve verimli bir algoritma olduğunu göstermiştir (Chen, Jiang, & Li, 2021). Deneyler kapsamında, MDFS yöntemi öznitelik seçme metodu olarak kullanılmıştır ve DFS'e çarpan olarak (d/(b+d)) eklenerek MDFS formülü elde edilmiştir, metodun formülü Eşitlik 2.20'te verilmiştir:

$$MDFS(t) = \sum_{j=1}^{M} \frac{(a/(a+c)) * (d/(b+d))}{(b/(a+b)) + (c/(c+d)) + 1} \quad \textbf{(2.20)}$$

$\frac{d}{b+d}$: t teriminin olmadığı durumlarda C_j hariç tüm sınıflarda olmama olasılığı

2.5. Sınıflandırma

Özniteliklerin çıkarıldığı sınıflandırma sürecinde, bu öznitelikler bir sınıflandırıcıya girdi olarak sunulur. Sınıflandırma algoritmaları, veri setinin eğitim kısmından gelen verinin özelliklerini öğrenir ve sınıfı bilinmeyen bir örnek geldiğinde bu örneği eğitim verisinde öğrendiği sınıflardan birine sınıflandırır. Sınıflandırmanın esas amacı, verilen veriler için doğru hedef sınıfı başarılı bir şekilde tahmin etmektir. Bu bölümde Destek Vektör Makineleri, K-En Yakın Komşu Algoritması, Karar Ağacı, Rastgele Orman Algoritması ve çok terimli Naïve Bayes hakkında bilgi verilecektir.

2.5.1. Destek vektör makineleri

SVM (Support Vector Machines), metin sınıflandırma gibi birçok uygulamada yüksek doğruluk sağlayabilen denetimli bir makine öğrenmesi algoritmasıdır. Metin dokümanları, sınıflandırma için özniteliklere dönüştürülür. Örneğin, bir metin belgesindeki kelimeler, birer öznitelik olarak düşünülerek bir vektör uzayında temsil edilir. Her bir kelime veya öznitelik, vektörün boyutunu temsil eder ve metnin bir noktasını belirler. Bu noktaların kümesi, dokümanların farklı sınıflarını temsil eder.

SVM sınıflandırıcısı, verilen öznitelik vektörleri arasında en iyi şekilde ayrılan bir hiper düzlem bulmaya çalışır. Bu hiper düzlem, sınıfları birbirinden mümkün olduğunca net bir şekilde ayıran bir karar sınırını ifade eder. SVM, sınıflar arasındaki ayrımı maksimize etmek için en uygun hiper düzlemi bulur.

Sınıfı bilinmeyen bir metin dokümanı, bu hiper düzlemle ayrılmış bölgeye göre sınıflandırılır. Yani, dokümanın hangi tarafta bulunduğuna bağlı olarak bir sınıfa atanır. SVM, özellikle doğrusal olmayan veri kümelerinde de iyi performans gösterebilir, çünkü

verileri daha yüksek boyutlu bir öznitelik uzayına dönüştürerek doğrusal olarak ayrılabilir hale getirebilir.

SVM'nin başarısı, en uygun maksimum marjın ve hiper düzlemin tespit edilmesine bağlıdır. Marj, sınıf noktaları ve hiper düzlem arasındaki en geniş boşluğu ifade eder. SVM, maksimum marjı olan bir hiper düzlemi bulmaya çalışarak, daha iyi genelleme yeteneği olan ve yeni verilere daha iyi uygulanabilen bir model oluşturur.

SVM metin sınıflandırma gibi birçok alanda kullanılabilir ve genellikle yüksek doğruluk elde etme potansiyeline sahiptir. Ancak, veri setinin büyüklüğüne, özniteliklerin seçimine ve modelin hiper parametrelerine bağlı olarak performansı değişebilir. SVM'nin avantajlarından biri de, özellikle öznitelik uzayında yüksek boyutluluk olduğunda etkili çalışabilmesidir.

SVM hem doğrusal hem de doğrusal olmayan formlarda mevcuttur ve hem iki sınıflı kategorizasyon hem de çok sınıflı kategorizasyon açısından sınıflandırma yapabilir (Joachims, 1998). Deneylerimizde, SVM'nin doğrusal versiyonu ve varsayılan parametrelerle birlikte libSVM kütüphanesi kullanılmıştır (Chang & Lin, 2011).

2.5.2. K-en yakın komşular

Metin sınıflandırma araştırmalarında, çeşitli sınıflandırma algoritmaları kullanılır. Bu algoritmalar arasında, KNN (K-Nearest Neighbors), basit öğrenme algoritmasından dolayı geniş ölçüde tercih edilmektedir (Dogan & Uysal, 2019). KNN algoritması, bir test belgesinin sınıfını belirlerken, belgenin k en yakın komşusuna olan benzerliğine dayanır.

KNN algoritmasının komşuları belirlemek için bir mesafe hesaplama yöntemine ihtiyacı vardır. Bu yöntemler arasında Öklidyen (Euclidean), Manhattan, Kosinüs ve Minkowski mesafe hesaplama yöntemleri bulunur. Bu yöntemler, bir belgenin diğer belgelerle olan benzerliğini veya farklılığını ölçmek için kullanılır.

Deneysel çalışmalarda genellikle Kosinüs benzerliği kullanılır. Kosinüs benzerliği, belgeler arasındaki açıyı ölçerek benzerlik belirler. Bu özellikle metin koleksiyonları arasındaki benzerliği belirlemek için oldukça etkilidir.

KNN algoritmasında, k parametresi veya komşu sayısı oldukça önemlidir. k, bir belgenin hangi sınıfa ait olduğunu belirlemek için dikkate alınacak komşu sayısını belirler. Deneylerimizde, en iyi sonuçları almak için k'nın değeri 1 olarak belirlenmiştir.

Bu, her test belgesinin en yakın komşusunun sınıfına dayanarak sınıflandırılacağı anlamına gelir.

2.5.3. Karar ağacı

DT (Decision Tree), doğrusal olmayan sınıflandırıcılar kategorisine girerler ve genellikle karmaşık veri setlerinin analizi için kullanılırlar. DT algoritmasının ana hedefi, özniteliklerin her birini belirli kategorilere karşılık gelen ayrı bölgelere ayırmaktır (Quinlan, 1986). Bu ayrım, genellikle bir dizi "eğer-ise(if-then)" kuralına dayalıdır ve bu kurallar, özniteliklerin ve hedef sınıfların ilişkisini tanımlar.

C4.5 adı verilen algoritma, karar ağacı oluşturma yöntemlerinin arasında çok popülerdir ve en başarılı karar ağacı sınıflandırma tekniklerinden biri olarak kabul edilir. C4.5, belirli özniteliklerin hedef sınıfı tahmin etmedeki etkinliğini belirlemek için entropi tabanlı bir ölçüm kullanır.

Ancak, birçok uygulamada, en çok tercih edilen karar ağacı türü ikili sınıflandırma ağacıdır. İkili sınıflandırma ağaçları, her bir düğümde yalnızca iki karar yolunun olmasını sağlar. Bu, genellikle modelin anlaşılmasını ve hesaplanmasını daha basit ve etkili hale getirir.

2.5.4. Rastgele orman

RF (Random Forest), çok sayıda DT kullanan birleşik tabanlı bir sınıflandırıcıdır. Sınıflandırma işlemi, karar ağacı sayısı belirlendikten sonra başlar, genellikle yüzlerce hatta binlerce ağaç içerebilir. Bu süreçte, her DT bağımsız olarak eğitilir, yani her bir ağaç kendi özel veri setini alır. Bu veri setleri, orijinal veri setinden bootstrap örnekleme yöntemiyle, yani değiştirme ile seçilmiş örneklerden oluşur.

Her ağaç, veri setinin belirli bir alt kümesini kullanarak eğitildiği için, her biri farklı bir model oluşturur. Bu, modelin genel esnekliğini artırır ve aşırı uyum problemini azaltır.

Sınıflandırma sonucunda, her ağaç bir tahminde bulunur ve final sınıfı, her DT'nin çoğunluk oylaması ile belirlenir (Jasti, ve diğerleri, 2022). Bu, tüm ağaçların tahminlerini birleştirerek, genellikle tek bir DT'den daha doğru ve istikrarlı bir tahmin sağlar.

2.5.5. Çok terimli naïve bayes

Naïve Bayes, metin sınıflandırma konusunda sıklıkla kullanılan bir sınıflandırıcıdır. Naïve Bayes, modelleme ve durum geçişlerini belirtme yeteneğine sahiptir (Witten &

Frank, 2002). Genellikle, bu yöntem ayrık ve sürekli değişkenlerden oluşan çok terimli verileri modellemek için kullanılır.

Multinomial Naïve Bayes (MNB), metin sınıflandırma amaçlı olarak tasarlanmış bir Naïve Bayes türüdür. Metin belgelerini sınıflandırmak için belge içindeki belirli kelimelerin varlığını ve sayısını inceler. Örneğin, bir belgenin 'spor' sınıfına ait olup olmadığını belirlemek için, belgedeki 'futbol', 'basketbol' gibi sporla ilgili kelimelerin sayısını değerlendirebilir.

Naïve Bayes, bir doküman üzerinde belirli kelimelerin varlığı ya da yokluğu ile ilgili modelleme yaparken, MNB, belirli kelimelerin belgedeki sayısını modellemeye ve hesaplamalara öncelik verir. Bu, belirli bir belgedeki kelime dağılımının, belgenin hangi sınıfa ait olduğunu belirlemede önemli bir faktör olabileceği anlamına gelir. Bu nedenle, MNB'nin, metin belgelerini sınıflandırmak için Naïve Bayes'ten daha doğru sonuçlar verebileceği düşünülür. Bu özelliği, metin sınıflandırma için bir araç olarak MNB'yi oldukça güçlü ve etkili kılar.

3. İLGİLİ ÇALIŞMALAR

Dengesiz metin sınıflandırma sorunu için üretilen çözümler üç ana başlık altında incelecektir. Dengesiz metin sınıflandırma konusunda geniş bir literatür mevcuttur. Bu literatürün genel sonuçları şunları içerir: Birincisi, dengesiz metin verilerinde, öğrenme algoritmasının kendisinden daha çok, öznitelik seçim işleminin kritik olduğudur. İkinci olarak, yüksek boyutlu dengesiz veri setlerinde örnekleme stratejileri ve algoritmik yaklaşımların sürekli olarak yüksek performans gösteremeyebileceği belirtilmiştir. Üçüncü olarak ise, öznitelik seçimi ve örnekleme tekniklerinin etkilerinin entegrasyonunun, yüksek boyutlu dengesiz metin sınıflandırmayı geliştirdiği ve öznitelik seçiminin örnekleme yöntemlerine kıyasla daha önemli olduğunu göstermiştir. Günümüzde, öznitelik seçimi, dengesizlik probleminin çözümüne yardımcı olacak teknikler arasında genel kabul görmüştür. Dengesiz metin sınıflandırma için birkaç yöntem geliştirilmiştir, ancak bu konu halen aktif bir araştırma konusu niteliğindedir. Bu konu ile ilgili en son çalışmalar üç başlık altında incelenecektir.

3.1. Örnekleme Metotları

Chawla ve takımı, dengesiz veri setleri için sınıflandırma algoritmalarının oluşturulması konusunda yepyeni bir aşırı örnekleme stratejisi geliştirmişlerdir (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). Bu, dengesiz metin sınıflandırma için ilk defa uygulanan bir örnekleme tekniğidir. Bu deney sürecinde, C4.5, Ripper ve Naïve Bayes gibi sınıflandırıcılar kullanılmıştır. Yapılan deneyler, dokuz farklı dengesiz veri seti üzerinde gerçekleştirilmiş ve SMOTE (Synthetic Minority Oversampling Technique) ile birlikte azınlık (minor) sınıfının aşırı örnekleme ve çoğunluk (major) sınıfının alt örnekleme yapılması kombinasyonunun, sadece çoğunluk sınıfının alt örnekleme yapılmasına kıyasla, ROC (Receiver Operating Characteristic – Alıcı Çalışma Karakteristik Eğrisi) altında kalan alan ile daha iyi bir sınıflandırma performansı elde edebileceğini göstermiştir. Chen ve arkadaşları, metin belgelerinin semantik anlamını kullanarak metin kategorizasyonunun sınıf dengesizliği altında nasıl geliştirilebileceğine dair benzersiz bir yaklaşım önermişlerdir (Chen, Lin, H., Luo, & Maa, 2011). DECOM (data re-sampling with probabilistic topic models) ve DECODER (data re-sampling with probabilistic topic models after smoothing) adlı iki yeniden örnekleme tekniği, olasılığa dayalı konu modelleri üzerinde kurulmuştur. DECOM, azınlık sınıfların yeni örneklerini oluşturarak sınıf dengesizliğini gidermek için tasarlanmıştır. DECODER, gürültülü örnekler ve azınlık sınıflar içeren veri kümeleri için, olasılık temelli konu modellerini

kullanarak her sınıfın tüm örneklerini yeniden oluşturur ve azınlık sınıfların boyutlarını genişletir. 20 Newsgroups, Reuters-21578, WebACE ve RCV1 veri setlerinde yapılan deneyler, DECOM ve DECODER'ın azınlık sınıflarında daha iyi sınıflandırma sonuçları sağladığını belgelemiştir. Barua ve takımı, dengesiz öğrenme problemlerine etkili bir çözüm sağlamak için MWMOTE (Majority Weighted Minority Oversampling Technique - Çoğunluk Ağırlıklı Azınlık Aşırı Örnekleme Tekniği) adlı yeni bir yöntem önermiştir (Barua, Islam, Yao, & Murase, 2012). MWMOTE dört yapay ve yirmi gerçek dünya veri seti üzerinde detaylı bir şekilde incelenmiştir. Sonuç olarak, önerilen teknik genellikle daha iyi bir G-mean (Geometrik ortalama) ve ROC altında daha yüksek alan göstermiştir. BorajoIglesias ve ekibi, COS-HMM (Content-based Over-Sampling HMM) isimli yeni bir aşırı örnekleme yöntemi tanıtmışlardır (Iglesias, Vieira, & Borrajo, 2013). Bu yöntemin etkinliğini ispatlamak adına COS-HMM, OHSUMED ve TREC Genomics adlı iki medikal metin veri seti üzerinde SVM sınıflandırıcısı ile denenmiştir ve sonrasında ROS (Random Over Sampling -Rastgele Aşırı Örnekleme) ve SMOTE teknikleriyle karşılaştırılmıştır. Deneysel ve istatistiksel analizler sonucunda, bu yeni yaklaşımın açık bir şekilde ROS'u geride bıraktığı ve çoğu durumda SMOTE'dan daha iyi performans sergilediği görülmüştür.

3.2. Algoritmik Metotlar

Manevitz ve Yousef, bilgi çıkarımı bağlamında tek sınıflı sınıflandırma için uygulanabilir olan SVM versiyonlarını kullanmışlardır (Manevitz & Yousef, 2002). Standart Reuters veri seti üzerinde çeşitli deneyler gerçekleştirmişlerdir. Sonuçlar, bir sınıflı SVM, dışa doğru SVM ve Manevitz ve Yousef'in çalışmasındaki dört diğer algoritmanın (Prototip (Rocchio'nun) algoritması, K-En Yakın Komşular, Naive Bayes, Compression Neural Network) F1 skorlarına göre üstünlük kurduğunu belirtmiştir. Galar ve ekibi, her bir önerinin hangi iç metodolojiye dayandığına bağlı olarak sınıflandırılabileceği bir taksonomi sunarak sınıf dengesizliği sorununu ele alan topluluk tabanlı metotlara odaklanmışlardır (Galar, Fern´andez, Barrenechea, Bustince, & Herrera, 2012). Deneylerini, sınıf dengesizliği ile ilgili 44 çift sınıflı gerçek dünya problemi üzerinde yapmışlar ve bu veriler KEEL veri seti depolarından alınmıştır. Ayrıca, çalışmalarında C4.5 sınıflandırıcısını kullanmışlardır. Sonuçlar, RUSBoost veya UnderBagging'in, birçok daha karmaşık diğer algoritmalara kıyasla daha yüksek performanslar sunduğunu göstermiştir. Özellikle örnekleme tekniklerinin (örneğin, alt örnekleme veya SMOTE) ve Bagging ensemble öğrenme algoritması ve RUSBoost

arasındaki sinerjinin, en iyi performans gösteren yaklaşımlar arasında hesaplama açısından en az karmaşıklığa sahip olduğu belirlenmiştir. Bir dizi çalışma, mevcut topluluk algoritma yöntemlerini geliştirmeye odaklanmıştır. Bunlardan bir tanesi, Wu ve ekibi tarafından geliştirilen FORESTEXTER (Wua, Ye, Zhang, Ng, & Ho, 2014) isimli yöntemdir. Dengesiz metin veri setlerini (Reuters-21578, Ohsumed ve 20 Newsgroup) kullanarak bu yöntemi uygulamışlardır. Sonuçta, önerdikleri yaklaşımın standart rastgele ormanlar ve çeşitli SVM varyantlarına (standart SVM, alt örnekleme SVM ve aşırı örnekleme SVM) kıyasla rekabet edebildiğini göstermişlerdir.

3.3. Öznitelik Seçme ve Öznitelik Ağırlıklandırma Metotları

Chen ve Wasikowski, FAST (Chen & Wasikowski, 2008) isimli yeni bir öznitelik seçim tekniği önermiştir. Bu metot, beş veri setinde (CNS, LYMPH, OVARY, PROST, NIPS) denenmiş ve lineer SVM ve 1-NN sınıflandırıcılarıyla test edilmiştir. Araştırmanın sonucunda, yeni metotlarının GINI, MI gibi diğer tekniklerden daha verimli olduğunu belirtmişlerdir. Ayrıca FAST'ın, AUC (Area under the ROC Curve) açısından RELIEF ve korelasyon katsayısı yöntemlerine kıyasla daha iyi performans sergilediği görülmüştür. Uysal ve Gunal, metin sınıflandırma için DFS (Uysal & Gunal, 2012) adında yeni bir filtre-tabanlı olasılık öznitelik seçim metodu geliştirmişlerdir. CHI2, IG, GINI ve POISSON'ı DFS ile karşılaştırmışlardır. Deneylerini hem dengeli bir veri seti (20 Newsgroups) hem de dengesiz veri setleri (Reuters-21578 ModApte split, Kısa Mesaj Servisi (SMS) mesaj koleksiyonu ve Enron1 Spam e-posta koleksiyonu) üzerinde gerçekleştirmişlerdir. Sonuçlar, yeni yaklaşımlarının sınıflandırma doğruluğu, boyut azaltma oranı ve işlem süresi bakımından en iyi performansı gösterdiğini ortaya koymuştur. Rehman ve takımı tarafından geliştirilen ve belgelerin göreceli sıklıklarını kullanan NDM isimli yeni bir öznitelik sıralama metriği önerilmiştir (Rehman, Javed, & Babri, 2017). WebACE, Reuters, 20 NewsGroups ve Concept drift (e-posta spam) veri setleri üzerinde MNB ve SVM sınıflandırıcılarıyla bu yaklaşımı denemişlerdir. Deneyler, yeni yaklaşımlarının genellikle Mikro-F1 ve Makro-F1 başarı ölçütlerinde CHI2, IG, OR, DFS ve GINI gibi diğer yaygın öznitelik seçim yöntemlerine göre daha üstün olduğunu göstermiştir. Kok ve arkadaşları, harmony search teknolojisi kullanılarak yüksek boyutlu dengesiz sınıf verileri için SYMON (Moayedikia, Ong, Boo, Yeoh, & Jensen, 2017) adlı yeni bir sarmalayıcı tabanlı öznitelik seçim metodu sunmuşlardır. Deneylerini DNA microarray ve Olivetti Faces veri setleri üzerinde yapmışlardır. Bu yöntem, SVM - Recursive Feature Elimination (SVM-RFE), SVM - Backward Feature Elimination

(SVM-BFE), Hellinger tabanlı öznitelik seçim algoritması (D-HELL), SMOTE-RLF (ReliefF), SMOTE-PCA (Principal Component Analysis) ile karşılaştırılmıştır. Sonuçlar, önerilen metodun daha yüksek G-mean ve F-measure performansı sağladığını göstermiştir. Pouramini ve ekibi, dengesiz metin sınıflandırma için PFS adı verilen yeni bir tek taraflı öznitelik seçim yöntemi önermişlerdir. Bu yöntem, öznitelik dağılımını olasılık hesaplamasına dayandırmaktadır (Pouramini, Minaei-Bidgoli, & Esmaeili, 2018). Diğer yöntemlere göre PFS'nin daha çok parametresi bulunmaktadır. Reuters-21875 ve WebKB veri setlerindeki F-measure test sonuçları, bu önerilen öznitelik seçim tekniğinin C4.5 ve Naïve Bayes sınıflandırıcılarının performansını ciddi anlamda artırdığını ortaya koymaktadır. Parlak, mevcut yöntemlerin sadece sınıf içindeki doküman frekansına dayalı olarak geliştirildiğini öne sürerek sınıf ve korpus içindeki özniteliğin durumunu dikkate alan yeni bir yöntem önermiştir. Dengesiz metin sınıflandırma için CICI (Parlak, 2022) adlı yeni bir öznitelik seçimi yöntemi sunmuştur. CICI, sınıf ve korpus içindeki öznitelik dağılımını kullanarak hesaplama yapan olasılıksal bir yöntemdir.

Daha fazla belge içeren kategoriler, daha az belge içerenlere göre daha az temsil edilmekte ve sınıflandırıcılar genellikle örnekleri daha çok belge içeren sınıfa yerleştirmektedir. Bu durumu ele almak adına, Liu ve ekibi, küçük kategorilere ait belgelerin daha iyi bir şekilde ayrılabilmesi için basit bir olasılığa dayalı terim ağırlıklandırma düzeni önermiştir (Liu, Loh, & Sun, 2009). Reuters-21578 ve MCV1 veri setlerinde hem SVM hem de NB sınıflandırıcılarıyla yapılan deneysel çalışmaların, küçük kategoriler için ciddi bir iyileşme sağladığını göstermiştir. Okkalioglu ve Okkalioglu'nun önerdiği yöntem AFE-MERT (Okkalioglu & Okkalioglu, 2022) tanınmış sınıflandırıcılarla karşılaştırıldığında rekabetçi bir sınıflandırıcı durumundayken, en son teknolojiye sahip terim ağırlıklandırma şemaları ve öznitelik seçim teknikleriyle deneysel olarak karşılaştırıldığında ise, terim ağırlıklandırma şemasında seçilen öznitelik sayısı sınıf sayısıyla sınırlı olduğunda, iyi sonuçlar almaktadır.

4. DENGESİZ METİN SINIFLANDIRMA

Bir veri setindeki dengesizlik, bazı sınıfların büyük sayıda örnekle (çoğunluk sınıfları), diğer sınıfların ise sadece birkaç örnekle (azınlık sınıfları) temsil edildiği bir durumu ifade etmektedir. Dengesizlik içsel ve dışsal dengesizlik olarak incelenebilir:

İçsel dengesizlik, verilerin doğası gereği bazı sınıfların diğerlerinden daha fazla örnekle temsil edildiği durumlara işaret etmektedir. Örneğin, dolandırıcılık tespitinde, dolandırıcılığa dayalı işlemler doğal olarak meşru olanlara göre daha nadirdir. Benzer şekilde, tıbbi teşhislerde, belirli bir hastalığa (örneğin kanser) sahip hastaların sayısı, genellikle sağlıklı hastaların sayısından çok daha azdır. Bu, veri setinde içsel bir dengesizlik yaratır. Öte yandan, dışsal dengesizlik, veri setindeki dengesizliğin, verinin kendisinden bağımsız faktörler tarafından yaratıldığı durumlara işaret etmektedir. Bunlar, veri toplamanın zorluğu veya maliyeti, belirli verilerin toplanmasını engelleyen gizlilik veya yasal sorunlar veya verinin nasıl işlendiği veya etiketlendiği ile ilgili sorunları içerebilir. Bu durumlarda, dengesizlik, verinin doğal bir özelliği olmayıp, dış koşullar tarafından belirlenmiştir (Yang P. , Liu, Zhou, Chawla, & Zomaya, 2013).

Dengesiz metin sınıflandırmada öznitelik seçimi, verinin doğası ve modelin sınıflandırma yeteneği açısından çeşitli sorunları beraberinde getirmektedir. Bu problemler aşağıdaki gibi sıralanabilir:

Yüksek Boyutluluk: Metin verileri genellikle yüksek boyutludur çünkü her bir kelime bir öznitelik olarak kabul edilmektedir. Bu durum, modelin eğitim süresini uzatabilir.

Dilbilimsel Gürültü: Metin verileri genellikle önemsiz kelimeler (stop words), noktalama işaretleri, yazım hataları ve dilbilgisel tutarsızlıklar gibi gürültüler içermektedir. Bu gürültüler, ön işleme aşamasında çıkarılmazsa, modelin performansını olumsuz etkileyebilir.

Seyrek Veri: Metin sınıflandırmada öznitelik matrisi genellikle seyrektir. Yani, birçok öznitelik (kelime), birçok belgede (metinde) nadiren görünmektedir. Bu, dengesiz sınıflandırmada bir sorun oluşturabilir çünkü azınlık sınıfındaki metinler, çoğunluk sınıfındaki metinlerden önemli öznitelikleri içerebilir ancak bu öznitelikler genellikle göz ardı edilir.

Özniteliklerin Dağılımı: Özellikle dengesiz veri kümelerinde, azınlık sınıfını temsil eden öznitelikler genellikle çoğunluk sınıfına göre daha az yaygındır. Bu, çoğunluk sınıfını tanımayı kolaylaştırırken, azınlık sınıfını tanımayı zorlaştırabilir.

Bu sorunları çözmek için bir dizi teknik kullanılabilir. Bunlar arasında öznitelik seçme (en önemli öznitelikleri belirleme) ve öznitelik ağırlıklandırma (azınlık sınıfını temsil eden özniteliklere daha yüksek ağırlık verilmesi) önde gelen tekniklerdir.

Bu bağlamda, varolan öznitelik seçme ve öznitelik ağırlıklandırma metotlarının içeriğinin değiştirilerek geliştirilmesine çalışılmıştır. Yapılan çalışmalara, 4.1 ve 4.2 bölümlerinde yer verilmiştir.

4.1. Düzenlenmiş Öznitelik Seçme Metotları

4.1.1. MGSS

GSS metodunun değiştirilmiş halidir. MGSS formülü aşağıdaki gibidir:

$$GSS\left(t,C_j\right)=\left(\frac{a*d-b*c}{N^2}\right)\rightarrow MGSS\left(t,C_j\right)=\left(a*d-b*c\right) \tag{4.1}$$

4.1.2. MCICI

CICI metodunun değiştirilmiş halidir. MCICI'nin formülü Eşitlik 4.2'de verilmiştir:

$$CiCi\left(t\right)=\frac{\left(a*d-b*c\right)^2}{\left(a+b\right)*\left(c+d\right)*N^2}\rightarrow MCICI\left(t\right)=\left(\frac{a*d-b*c}{\left(a+c\right)}\right) \tag{4.2}$$

4.1.3. MMMR

MMR metodunun değiştirilmiş halidir. MMMR'nin formülü aşağıda belirtilmiştir:

$$MMR\left(t\right)=\sum_{j=1}^{M}\frac{\left|\frac{a}{a+b}-\frac{c}{c+d}\right|}{\min\left(\frac{a}{a+b},\frac{c}{c+d}\right)}*\max\left(\frac{a}{a+b},\frac{c}{c+d}\right)$$

$$\downarrow \tag{4.3}$$

$$MMMR\left(t\right)=\sum_{j=1}^{M}\frac{\left|\frac{a}{a+b}-\frac{c}{c+d}\right|}{\max\left\{1,\min\left(\frac{a}{a+b},\frac{c}{c+d}\right)\right\}}*\max\left(\frac{a}{a+b},\frac{c}{c+d}\right)$$

4.1.4. MCMFS

CMFS metodunun değiştirilmiş halidir. MCMFS metodunun formülasyonu Eşitlik 4.4'te verilmiştir:

$$CMFS(t)=\sum_{j=1}^{M}\frac{(a+b)}{N}*\frac{a}{a+b}*\frac{a}{a+c}\rightarrow MCMFS(t)=\sum_{j=1}^{M}\frac{a^2}{(a+b)^2*(a+c)} \quad \textbf{(4.4)}$$

4.1.5. MDFSS

DFSS metodunun değiştirilmiş halidir. MDFSS, Eşitlik 4.5'teki formülle ifade edilebilir:

$$DFSS(t,C_j)=\left(\frac{\frac{e}{a}}{\frac{f}{c}}\right)*\frac{a}{a+c}*\frac{a}{a+b}*\left|\frac{a}{a+c}-\frac{b}{b+d}\right|$$

$$\downarrow \quad \textbf{(4.5)}$$

$$MDFSS(t,C_j)=\left(\frac{\frac{e*(a+b)}{a}}{\frac{f*(c+d)}{\max(1,c)}}\right)*\frac{a}{(a+b)(a+c)}*\frac{a}{a+b}$$

4.1.6. MEFS1

EFS metodunun değiştirilmiş halidir. MEFS1'nin formülasyonu aşağıdaki gibidir:

$$EFS(t)=\sum_{j=1}^{M}\left(\frac{a/(a+b)}{b/(a+b)+c/(c+d)+1}\right)*\left(\frac{a/(a+c)}{c/(a+c)+b/(b+d)+1}\right)$$

$$\downarrow \quad \textbf{(4.6)}$$

$$MEFS1(t)=\sum_{j=1}^{M}\left(\frac{a/(a+b)}{(a+c)*((c/(c+d))+1)}\right)*\left(\frac{a/(a+b)}{((c*(a+c))/(c+d))+1}\right)$$

4.1.7. MEFS2

EFS metodunun değiştirilmiş halidir. MEFS2 formülü Eşitlik 4.7'deki gibidir:

$$EFS(t)=\sum_{j=1}^{M}\left(\frac{a/(a+b)}{b/(a+b)+c/(c+d)+1}\right)*\left(\frac{a/(a+c)}{c/(a+c)+b/(b+d)+1}\right)$$

$$\downarrow \quad \textbf{(4.7)}$$

$$MEFS2(t)=\sum_{j=1}^{M}\left(\frac{a/(a+b)}{(a+c)*((c/(c+d))+1)}\right)*\left(\frac{a/(a+b)}{(c/((a+c)*(c+d)))+1}\right)$$

4.2. Düzenlenmiş Öznitelik Ağırlıklandırma Metotları

4.2.1. TF-MGSS

Terim Frekansı (TF) ile MGSS değerleri çarpılarak terimlerin ağırlık değerlerine ulaşılır. TF-MGSS'nin formülasyonu aşağıdaki gibidir:

$$W_{TF-MGSS(t_i)} = TF(t_i, d_k) * MGSS(t_i, C_j) = TF(t_i, d_k) * (a * d - b * c) \quad \textbf{(4.8)}$$

4.2.2. TF-MCICI

Terim Frekansı (TF) ile MCICI değerleri çarpılarak terimlerin ağırlık değerleri hesaplanır. TF-MCICI formülü Eşitlik 4.9'da belirtilmiştir:

$$W_{TF-MCICI(t_i)} = TF(t_i, d_k) * MCICI(t_i) = TF(t_i, d_k) * \left(\frac{a * d - b * c}{(a + c)} \right) \quad \textbf{(4.9)}$$

4.2.3. TF-MMMR

Terim Frekansı (TF) ile MMMR değerleri çarpılarak terimlerin ağırlık değerleri elde edilir. TF-MMMR'nin formülasyonu aşağıdaki gibidir:

$$W_{TF-MMMR(t_i)} = TF(t_i, d_k) * MMMR(t_i)$$

$$W_{TF-MMMR(t_i)} = TF(t_i, d_k) * \sum_{j=1}^{M} \frac{\left| \frac{a}{a+b} - \frac{c}{c+d} \right|}{\max\left\{1, \min\left(\frac{a}{a+b}, \frac{c}{c+d} \right)\right\}} * \max\left(\frac{a}{a+b}, \frac{c}{c+d} \right) \quad \textbf{(4.10)}$$

4.2.4. TF-MCMFS

Terim Frekansı (TF) ile MCMFS değerleri çarpılarak terimlerin ağırlık değerleri belirlenir. TF-MCMFS formülü Eşitlik 4.11'de verilmiştir:

$$W_{TF-MCMFS(t_i)} = TF(t_i, d_k) * MCMFS(t_i)$$

$$W_{TF-MCMFS(t_i)} = TF(t_i, d_k) * MCMFS(t_i) = \sum_{j=1}^{M} \frac{a^2}{(a+b)^2 * (a+c)} \quad \textbf{(4.11)}$$

4.2.5. TF-MDFSS

Terim Frekansı (TF) ile MDFSS değerleri çarpılarak terimlerin ağırlık değerlerine ulaşılır. TF-MDFSS, Eşitlik 4.12'deki formülle ifade edilebilir:

$$W_{TF-MDFSS(t_i)} = TF\left(t_i, d_k\right) * MDFSS\left(t_i, C_j\right)$$

$$W_{TF-MDFSS(t_i)} = TF\left(t_i, d_k\right) * \left(\frac{\frac{e*(a+b)}{a}}{\frac{f*(c+d)}{\max(1,c)}} \right) * \frac{a}{(a+b)(a+c)} * \frac{a}{a+b} \quad \textbf{(4.12)}$$

4.2.6. TF-MEFS1

Terim Frekansı (TF) ile MEFS1 değerleri çarpılarak terimlerin ağırlık değerleri elde edilir. TF-MEFS1'in formülasyonu aşağıda verilmiştir:

$$MEFS1\left(t_i\right) = \sum_{j=1}^{M} \left(\frac{a/(a+b)}{(a+c)*\left(\left(c/(c+d)\right)+1\right)} \right) * \left(\frac{a/(a+b)}{\left(\left(c*(a+c)\right)/(c+d)\right)+1} \right) \quad \textbf{(4.13)}$$

$$W_{TF-MEFS1(t_i)} = TF\left(t_i, d_k\right) * MEFS1\left(t_i\right)$$

4.2.7. TF-MEFS2

Terim Frekansı (TF) ile MEFS2 değerleri çarpılarak terimlerin ağırlık değerleri hesaplanır. TF-MEFS2'nin formülasyonu aşağıdaki gibidir:

$$MEFS2\left(t_i\right) = \sum_{j=1}^{M} \left(\frac{a/(a+b)}{(a+c)*\left(\left(c/(c+d)\right)+1\right)} \right) * \left(\frac{a/(a+b)}{\left(c/\left((a+c)*(c+d)\right)\right)+1} \right) \quad \textbf{(4.14)}$$

$$W_{TF-MEFS2(t_i)} = TF\left(t_i, d_k\right) * MEFS2\left(t_i\right)$$

4.3. Dengesiz Metin Veri Setleri

Dengesiz veri setlerinde, her sınıfın ve veri setinin bir bütün olarak dengesizlik oranını (IR – Imbalance Ratio) ölçmek mümkündür.

Dengesizlik oranı, Eşitlik 4.15'teki gibi formüle edilmiştir:

$$IR = \frac{N_{C_j}}{N - N_{C_j}} \quad \textbf{(4.15)}$$

burada N_{C_j} , C_j sınıfının doküman sayısını ve N toplam doküman sayısını temsil etmektedir.

4.3.1. Reuters – 21578

Reuters-21578 veri seti, tutarlı bir şekilde 135 kategoriye ayrılmış bir Reuters haber ajansı veri setidir (Asuncion & Newman, 2007). Dengesiz metin sınıflandırma çalışmalarında ilk 8 kategori (R8) seçilmiştir ve Tablo 4.1, Reuters-21578 veri seti için eğitim ve test doküman sayılarını, öznitelik sayısını, dengesizlik oranını ve sınıf dağılımlarını göstermektedir.

Tablo 4.1. *Reuters-21578 veri setinin (R8) özellikleri*

Sınıf Adı	Eğitim Doküman Sayısı	Dengesizlik Oranı	Sınıf Dağılımı (%)	Öznitelik Sayısı	Test Doküman Sayısı	Toplam Doküman Sayısı
earn	2877	0.733	42	8614	1087	3964
acq	1650	0.320	25	8378	719	2369
money-fx	538	0.085	8	3662	179	717
grain	433	0.068	6	3697	149	582
crude	389	0.060	6	4296	189	578
trade	369	0.057	5	3869	117	486
interest	347	0.053	5	2633	131	478
ship	197	0.029	3	2988	89	286
Toplam	**6800**	**0.175**	**100**	**16955**	**2660**	**9460**

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

R8 MM (Minör Majör) ise, R8 veri setinde en az (“ship”) ve en çok (“earn”) doküman sayılarına sahip iki sınıftan oluşan ikili (binary) dengesiz bir veri setidir. Tablo 4.2’de R8 MM verisetinin özellikleri sunulmuştur.

Tablo 4.2. *Reuters-21578 veri setinin (R8 MM) özellikleri*

Sınıf Adı	Eğitim Doküman Sayısı	Dengesizlik Oranı	Sınıf Dağılımı (%)	Öznitelik Sayısı	Test Doküman Sayısı	Toplam Doküman Sayısı
earn	2877	14.604	94	7742	1087	3964
ship	197	0.068	6	3860	89	286
Toplam	**3074**	**7.336**	**100**	**9804**	**1176**	**4250**

4.3.2. WebKB

WebKB, dört farklı üniversite web sitesinden alınmış web sayfalarının bir koleksiyonudur ve 8282 web sayfası içermektedir. Tüm web sayfaları eşit olmayan bir şekilde yedi kategoriye ayrılmıştır. Deneylerimizde dört kategori ("course", "faculty",

"project", "student") kullanılmıştır. Veri seti, iki minör ("course", "project") ve iki majör ("student", "faculty") kategori içermektedir (Cardoso Cachopo, 2014). WebKB veri setine ait özellikler Tablo 4.3'te gösterilmektedir:

Tablo 4.3. *WebKB veri setinin özellikleri*

Sınıf Adı	Eğitim Doküman Sayısı	Dengesizlik Oranı	Sınıf Dağılımı (%)	Öznitelik Sayısı	Test Doküman Sayısı	Toplam Doküman Sayısı
student	1149	0.640	39	8065	492	3964
faculty	787	0.365	27	13224	337	2369
course	651	0.284	22	13301	279	717
project	353	0.136	12	8388	151	582
Toplam	2940	**0.356**	**100**	**25658**	**1259**	**4199**

4.3.3. Enron1

İyi bilinen bir karşılaştırma veri seti olan Enron Email veri setinde ise "ham" ve "spam" olmak üzere iki sınıf ve 6 kategori bulunmaktadır (Metsis, Androutsopoulos, & Paliouras, (2006, July)). Deneylerimizde Enron Email veri setindeki kategorilerden birinin kullanılması tercih edilmiştir.

Eğitim/test doküman sayıları, hedef sınıf olarak eğitim setini dikkate alırken Enron1 veri setindeki her sınıfın dengesizlik oranı, öznitelik sayısı ve sınıf dağılımları Tablo 4.4'te gösterilmektedir.

Tablo 4.4. *Enron1 veri setinin özellikleri*

Sınıf Adı	Eğitim Doküman Sayısı	Dengesizlik Oranı	Sınıf Dağılımı (%)	Öznitelik Sayısı	Test Doküman Sayısı	Toplam Doküman Sayısı
ham	2570	2.447	71	27340	1102	3672
spam	1050	0.408	29	8605	450	1500
Toplam	**3620**	**1.427**	**100**	**31388**	**1552**	**5172**

4.3.4. Spam SMS

SMS Spam veri seti, SMS Spam çalışmasının çıktılarını (SMS etiketli mesajları) içermektedir. 5.574 İngilizce SMS mesajından oluşmaktadır ve bu mesajlar spam veya ham (legal) olarak kategorize edilmektedir (Almeida, Hidalgo, & Yamakami, (2011, September)). Deneylerimizde SMS Spam veri setinin orijinal versiyonu kullanılmıştır.

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

Spam SMS veri setinin eğitim/test doküman sayıları, hedef sınıf olarak eğitim setini dikkate alırken her sınıfın dengesizlik oranı, öznitelik sayısı ve sınıf dağılımları Tablo 4.5'te gösterilmiştir.

Tablo 4.5. *Spam SMS veri setinin özellikleri*

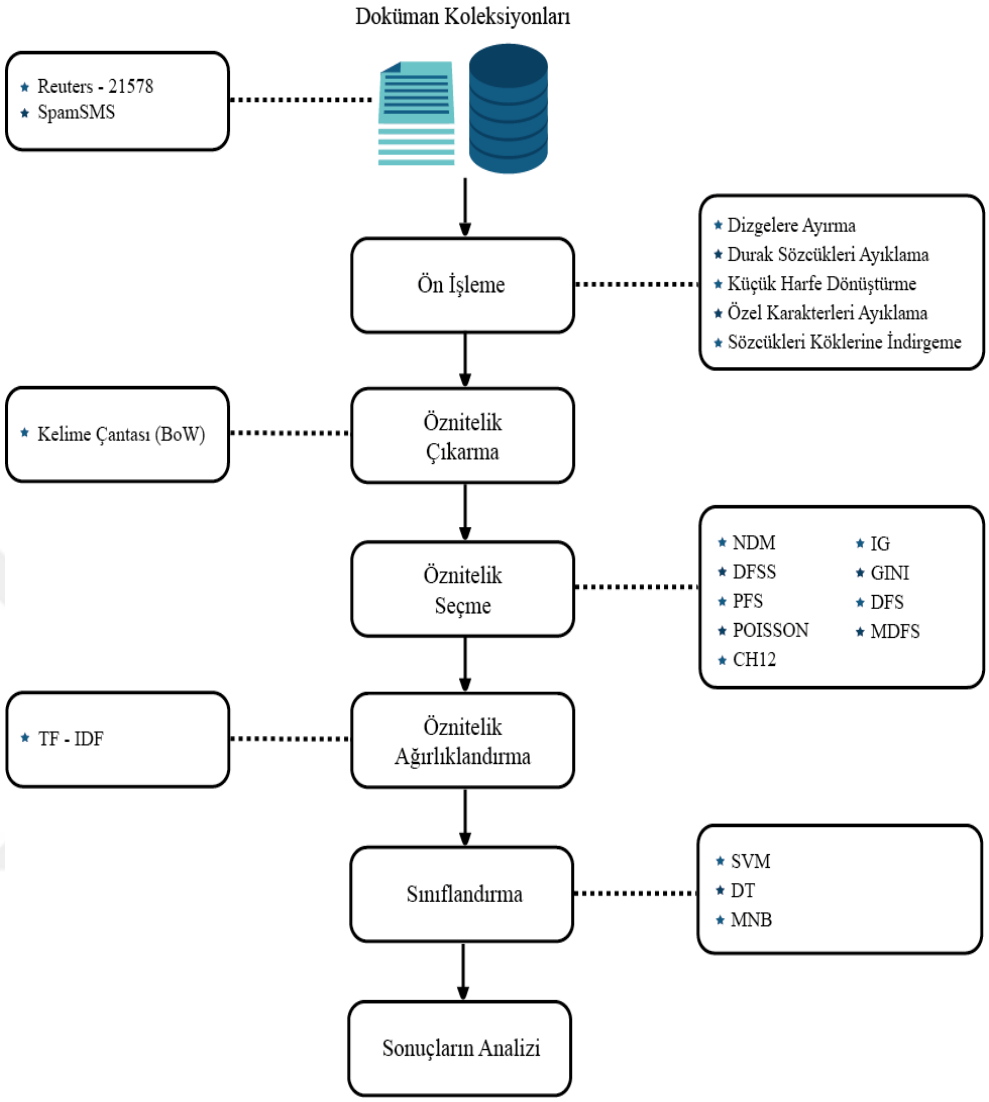
Sınıf Adı	Eğitim Doküman Sayısı	Dengesizlik Oranı	Sınıf Dağılımı (%)	Öznitelik Sayısı	Test Doküman Sayısı	Toplam Doküman Sayısı
legitimate	3378	6.471	87	4439	1449	4827
spam	522	0.154	13	1399	225	747
Toplam	**3900**	**3.312**	**100**	**5185**	**1674**	**5574**

5. DENGESİZ METİN SINIFLANDIRMADA ÖZNİTELİK SEÇİM YÖNTEMLERİNİN ETKİLİLİĞİ

Öznitelik seçiminin, metin sınıflandırma sürecinin kritik unsurlarından biri olduğu ve dengesiz metin konusunda da önemli bir rol oynadığı bilinmektedir. Bu çalışmada, dengesiz metinlerin sınıflandırılmasında öznitelik seçim tekniklerinin etkisi ayrıntılı olarak incelenmiştir. Bu bağlamda, iki ayrı veri kümesinde, üç çeşit sınıflandırıcı ve dokuz farklı öznitelik seçim stratejisi kullanarak bir dizi deney gerçekleştirilmiştir. Bunun yanı sıra, farklı öznitelik boyutlarındaki başarılar da öznitelik seçim metotları açısından incelenmiştir.

5.1. Motivasyon

Bu çalışmada, öznitelik seçiminin dengesiz metin sınıflandırma problemi üzerindeki rolü incelenmiştir. NDM, DFSS, PFS, POISSON, CHI2, IG, GINI, DFS ve MDFS isimlerindeki dokuz farklı öznitelik seçme stratejisi değerlendirilmiştir. Reuters-21578 (R8) ve SpamSMS veri setlerinde, SVM, DT ve MNB sınıflandırıcıları kullanılarak deneyler gerçekleştirilmiştir. Geniş kapsamlı analizler için, 30 ile 500 arasında değişen beş farklı öznitelik boyutunda sonuçlar karşılaştırılmıştır. Bu yaklaşım, bir dizi modern öznitelik seçme yönteminin, sınıflar bazında veri dağılımı varyasyonlarında nasıl bir performans sergilediğini detaylı bir şekilde ortaya koymaktadır. Bunun sonucunda, dengesiz metin sınıflandırma sorununda en iyi performansı sergileyen öznitelik seçme yöntemleri belirlenmiştir (Tiryaki & Uysal, 2023). Deney süreçlerinin özetini sunan metin sınıflandırma diyagramı Şekil 5.1'de verilmiştir.



Şekil 5.1. *Deney çalışmasının genel şeması*

5.2. Deneysel Çalışma

Deney sürecinde, dengesiz metin sınıflandırma konusu öznitelik seçme perspektifinden incelenmiştir. Dengesiz niteliğe sahip Reuters-21578 (R8) ve SpamSMS adlı iki veri seti üzerinde deneyler gerçekleştirilmiştir. "dizgelere ayrıştırma", "küçük harfe dönüştürme", "durak kelimeleri ayıklama" ve "köklere indirgeme" olarak adlandırılan ön işlem adımları da bu deneylerde uygulanmıştır. Kökleri belirlemek için Porter stemmer algoritması (Moayedikia, Ong, Boo, Yeoh, & Jensen, 2017) ve terim ağırlığını belirlemek için TF-IDF (Schütze, Manning, & Raghavan, 2008) kullanılmıştır. Değerlendirme metriği olarak Makro-F1 seçilmiştir çünkü dengesiz veri setlerinde

sınıflandırma performansını daha iyi yansıtmaktadır. Her iki veri setinde de sırasıyla 30, 50, 100, 300 ve 500 öznitelik sayısı seçilmiş ve sınıflandırma başarılarına bakılmıştır.

5.3. Değerlendirme Ölçütü

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

$$Makro - F1 = \frac{1}{M} * \sum_{i=1}^{M} \left\{ \frac{2 * TP_{M_i}}{2 * TP_{M_i} + FP_{M_i} + FN_{M_i}} \right\} \quad \textbf{(5.1)}$$

TP: M_i sınıfında olan ve doğru şekilde sınıflandırılan doküman sayısı

FP: M_i sınıfında olmayan ve yanlış şekilde sınıflandırılan doküman sayısı

FN: esasen M_i sınıfında olmadığı halde yanlış şekilde sınıflandırılan doküman sayısı

M: toplam sınıf sayısı

5.4. Performans Analizi

İki farklı veri setinde gerçekleştirilen deneylerde elde edilen sonuçlar Tablo 5.1 ve Tablo 5.2'de sunulmaktadır. En yüksek Makro-F1 skoruna sahip üç değer kalın yazı tipiyle vurgulanmıştır.

Tablo 5.1. *Reuters-21578 (R8) veri seti için SVM (a), DT (b) ve MNB (c) sınıflandırıcıları Makro-F1 değerleri*

Öznitelik Seçme Metotları	Öznitelik Sayıları				
	30	**50**	**100**	**300**	**500**
NDM	44,328	49,260	53,710	68,200	72,130
DFSS	77,195	79,959	78,930	79,600	**80,580**
PFS	14,449	14,449	14,270	14,500	20,304
POISSON	76,230	79,260	79,730	**80,800**	80,480
CHI2	74,242	78,806	80,300	78,500	79,790
IG	70,315	77,357	80,030	80,500	80,040
GINI	74,425	79,361	79,730	80,100	**80,860**
DFS	76,771	79,794	80,300	79,100	79,660
MDFS	77,109	77,041	77,320	77,300	78,030

(a)

Öznitelik Seçme Metotları	Öznitelik Sayıları				
	30	**50**	**100**	**300**	**500**
NDM	45,574	50,115	54,280	71,200	71,530
DFSS	74,587	75,013	74,970	76,900	77,350
PFS	15,409	15,409	15,540	15,500	16,723
POISSON	73,301	72,873	74,470	77,700	77,650
CHI2	74,333	77,479	75,360	77,900	**78,370**
IG	71,726	72,999	76,100	**78,400**	77,540
GINI	71,875	74,926	74,630	76,300	77,680
DFS	74,950	76,684	76,650	77,900	**78,430**
MDFS	75,157	76,044	76,300	76,300	75,810

(b)

Öznitelik Seçme Metotları	Öznitelik Sayıları				
	30	**50**	**100**	**300**	**500**
NDM	44,405	49,844	54,930	73,200	74,320
DFSS	74,137	78,560	78,460	77,600	77,130
PFS	15,361	15,354	12,960	13,400	20,353
POISSON	75,880	77,652	78,280	**79,500**	77,800
CHI2	75,356	77,424	**79,500**	**79,500**	78,970
IG	70,659	76,520	76,270	76,900	75,750
GINI	74,458	79,212	78,980	78,800	77,960
DFS	73,690	77,450	78,810	**79,300**	77,590
MDFS	74,769	75,773	76,540	75,800	74,730

(c)

Tablo 5.2. *Spam SMS veri seti için SVM (a), DT (b) ve MNB (c) sınıflandırıcıları Makro-F1 değerleri*

Öznitelik Seçme Metotları	Öznitelik Sayıları				
	30	**50**	**100**	**300**	**500**
NDM	88,150	88,925	91,043	94,153	94,028
DFSS	89,802	92,319	93,742	94,747	93,106
PFS	89,482	89,206	91,252	91,134	91,560
POISSON	90,786	93,493	94,684	94,182	94,035
CHI2	90,786	91,645	93,501	94,453	94,059
IG	90,747	92,950	93,815	94,584	**94,787**
GINI	90,904	91,679	93,692	**95,340**	94,828
DFS	89,322	93,035	94,106	**95,216**	**95,482**
MDFS	87,338	89,206	91,252	91,134	91,560

(a)

Öznitelik Seçme Metotları	Öznitelik Sayıları				
	30	**50**	**100**	**300**	**500**
NDM	87,666	87,835	88,679	91,365	89,912
DFSS	89,685	90,394	90,734	89,802	88,554
PFS	88,293	89,295	91,043	91,229	91,529
POISSON	90,823	**91,995**	**92,382**	91,298	91,298
CHI2	90,823	90,393	**92,382**	91,298	91,298
IG	90,979	91,124	91,495	91,480	91,413
GINI	90,979	89,820	90,507	91,480	91,149
DFS	89,823	91,662	91,529	**91,825**	91,645
MDFS	87,054	89,020	90,624	90,739	90,813

(b)

Öznitelik Seçme Metotları	Öznitelik Sayıları				
	30	**50**	**100**	**300**	**500**
NDM	89,423	89,186	91,576	94,807	94,950
DFSS	90,972	92,734	93,621	94,970	94,970
PFS	88,113	88,075	90,959	93,176	93,176
POISSON	91,731	93,166	94,011	94,869	94,869
CHI2	91,731	93,043	94,011	94,869	94,869
IG	92,800	92,771	**95,034**	94,890	94,890
GINI	92,193	93,466	93,986	**95,178**	**95,178**
DFS	90,904	93,221	93,661	**95,428**	**95,428**
MDFS	87,910	86,711	88,016	90,713	90,754

(c)

Reuters-21578 (R8) veri seti üzerinde gerçekleştirilen deneyde, Tablo 5.3'e göre SVM'nin Makro-F1 değeri yaklaşık 80 ile en yüksek sonucu verdiği görülmektedir. Bu değeri yaklaşık 79 ve 78 ile MNB ve DT sınıflandırıcıları izlemektedir. Bu durumda Reuters-21578 (R8) veri seti için SVM > MNB > DT sıralaması yapabiliriz. Ancak Makro-F1 değerlerinde çok belirgin bir fark bulunmamaktadır. Yüksek değerler genel olarak 300 ve 500 öznitelik sayıları için kaydedilmiştir. CHI2 ve DFS özellikle DT ve MNB sınıflandırıcılarında 78 ve 79 Makro-F1 değerleriyle diğer öznitelik seçme tekniklerinden ayrılmaktadır.

Diğer yandan, PFS tüm sınıflandırıcılarda 20 ve altındaki Makro-F1 değerleriyle en düşük sonuçları vermiştir. PFS'yi, 74 ve altındaki Makro-F1 değerleri ile NDM yöntemi takip etmektedir. SPAM SMS veri setinde gerçekleştirilen deneylere bakıldığında, Tablo 5.4'te gösterildiği gibi, sınıflandırıcı performans sıralaması SVM > MNB > DT şeklinde belirlenmiştir. SVM ve MNB, yaklaşık 95'lik Makro-F1 değeri ile sonuçları sunarken, DT'nin değeri 91-92 arasında değişmektedir. Fakat bu Makro-F1 başarı seviyeleri Reuters-21578 (R8) veri seti ile karşılaştırıldığında daha yüksek çıkmıştır.

Öznitelik sayısı bakımından incelendiğinde, SVM ve MNB, az öznitelik sayılarına kıyasla daha fazla öznitelik sayılarında daha yüksek sonuçlar vermektedir. Ancak, DT sınıflandırıcı, 50, 100 ve 300 öznitelik boyutlarında diğer öznitelik boyutlarına nazaran daha yüksek Makro-F1 sonuçları sunmuştur. DFS, SVM ve MNB'de 95 ve üzeri değerler ile diğer öznitelik seçme yöntemlerine kıyasla daha etkili sonuçlar sağlamıştır. PFS, Reuters-21578 (R8) veri setinde elde edilen düşük sonuçların aksine daha yüksek değerler vermiştir ve diğer öznitelik seçme yöntemleriyle benzer Makro-F1 sonuçları kaydedilmiştir.

Tablo 5.1 ve Tablo 5.2'de verilen sonuçların daha iyi yorumlanabilmesi için Tablo 5.3 ilave olarak sunulmuştur. Tablo 5.3'te, deneyler sonucunda elde edilen Makro-F1 değerlerine göre en iyi skorları üreten öznitelik seçme yöntemleri görülmektedir.

Tablo 5.3. *Öznitelik seçme metotlarının en iyi Makro-F1 değerlerini ürettiği durumların sayıları*

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

Tablo 5.3’ten görüleceği üzere dengesiz metin sınıflandırmada öznitelik seçimi noktasında başarı açısından DFS ve CHI2 yöntemleri diğer yöntemlere göre önde görünmektedir.

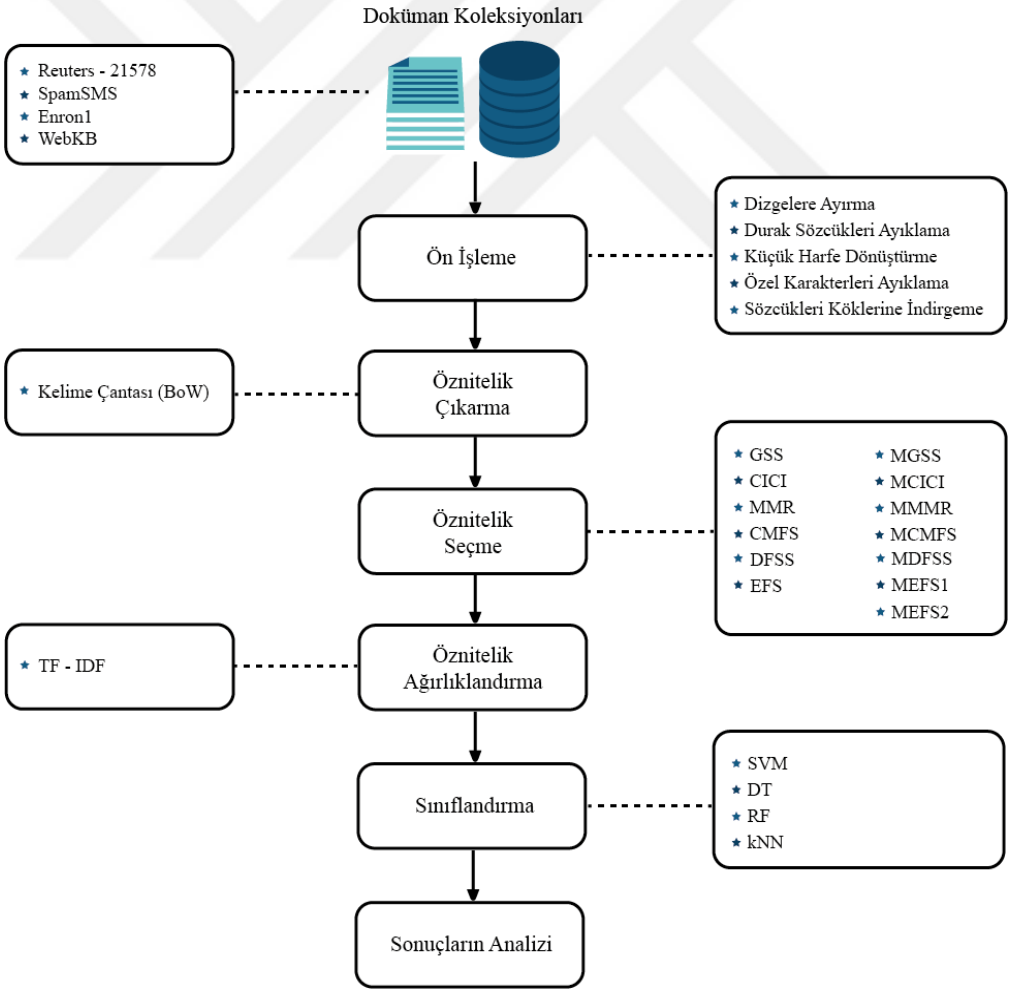
5.5. Sonuçlar

Bu çalışmada, Reuters-21578 (R8) ve SPAM SMS veri setlerinin üzerinde, SVM, DT ve MNB sınıflandırıcılarıyla birlikte dokuz farklı öznitelik seçme stratejisinin etkinliği karşılaştırılmıştır. Yapılan deneylerde, DFS ve CHI2'nin genel anlamda güçlü sonuçlar sergilediği, öte yandan PFS ve NDM'nin özellikle Reuters-21578 (R8) veri seti üzerinde zayıf performans ortaya koyduğu belirlenmiştir. Bu çalışmanın bir uzantısı olarak, belirli alanlarda dengesiz metin sınıflandırma sorunlarına yönelik olarak öznitelik seçme tekniklerinin etkinliğinin incelenmesi planlanabilir.

6. DENEYSEL ÇALIŞMALAR : DÜZENLENMİŞ METOTLAR

6.1. Öznitelik Seçme Deneyleri

Yapılan deneylerde, GSS, MGSS, CICI, MCICI, MMR, MMMR, CMFS, MCMFS, DFSS, MDFSS, EFS, MEFS1, MEFS2 adlarında toplamda on üç öznitelik seçme yaklaşımı incelenmiştir. Reuters-21578, Enron1, WebKB ve SpamSMS gibi veri kümeleri üzerinde, SVM, DT, kNN ve MNB sınıflandırma algoritmaları ile testler yapılmıştır. Öznitelik ağırlıklandırma için TF-IDF kullanılmıştır. Detaylı incelemeler adına, 50'den 1000'e kadar olan beş değişik özellik uzunluğunda sonuçlar değerlendirilmiştir. Deney çalışmalarının sonunda, dengesiz metin sınıflandırmada iki sınıflı ve çok sınıflı veri setlerinde en üstün sonuçları veren öznitelik seçme metotları saptanmıştır. Yapılan deneylerin özeti Şekil 6.1'deki metin sınıflandırma şemasında sunulmuştur.



Şekil 6.1. *Deney çalışmasının genel şeması*

Deneyler sonucunda elde edilen Makro-F1 değerlerine göre, iki ve çok sınıflı veri setlerinde en iyi skorları üreten öznitelik seçme yöntemleri Tablo 6.1 ve Tablo 6.2'de gösterilmiştir.

Tablo 6.1. *Öznitelik seçme metotlarının iki sınıflı veri setlerinde en iyi Makro-F1 değerlerini ürettiği durumların sayıları*

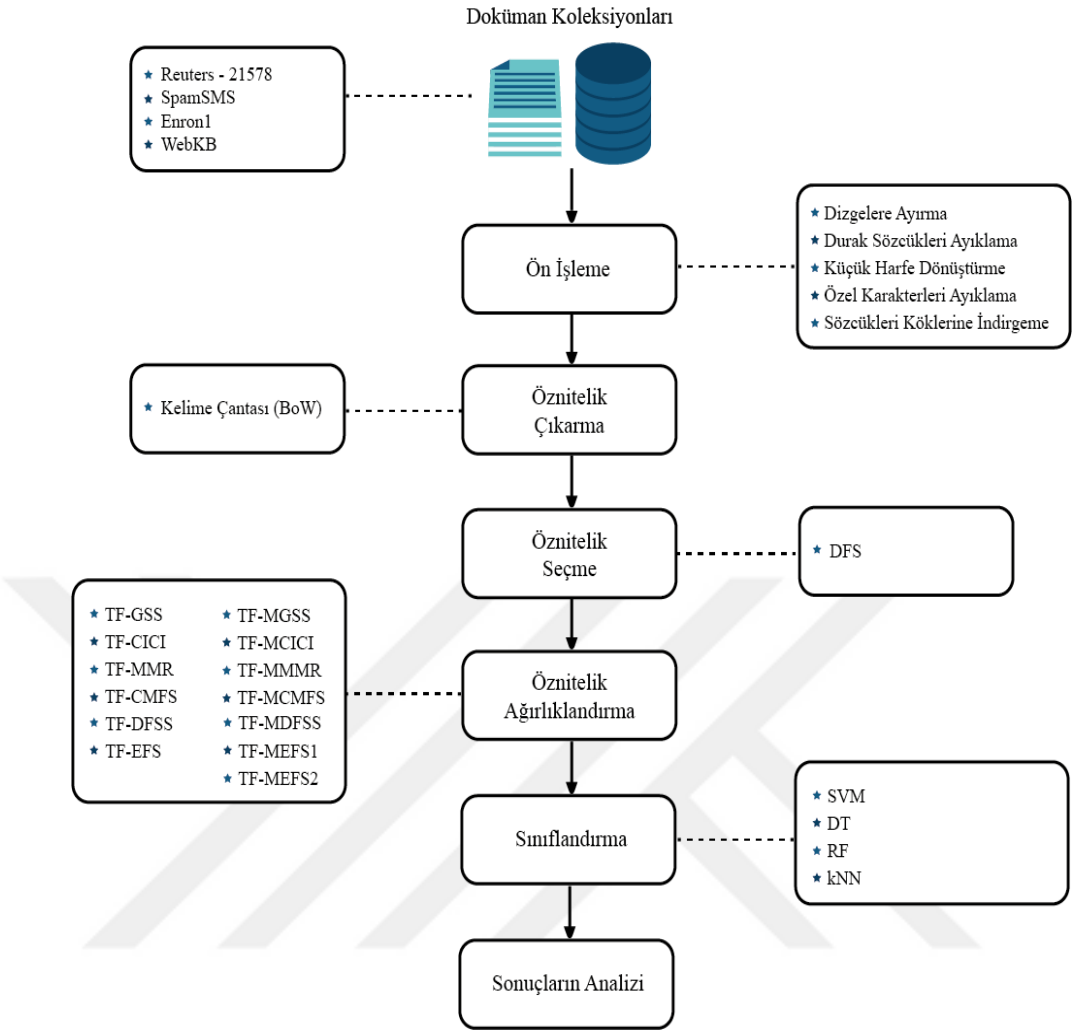
MEFS1	MEFS2	MGSS	MMMR	MCMFS	CICI	CMFS	GSS
6	2	1	1	1	1	1	1

Tablo 6.2. *Öznitelik seçme metotlarının çok sınıflı veri setlerinde en iyi Makro-F1 değerlerini ürettiği durumların sayıları*

CICI	MMMR	DFSS	EFS	MMR
3	2	1	1	1

6.2. Öznitelik Ağırlıklandırma Deneyleri

Bu deneylerde ise dengesiz metin sınıflandırmada öznitelik ağırlıklandırmanın etkisi gözlemlenmiştir. TF-GSS, TF-MGSS, TF-CICI, TF-MCICI, TF-MMR, TF-MMMR, TF-CMFS, TF-MCMFS, TF-DFSS, TF-MDFSS, TF-EFS, TF-MEFS1 ve TF-MEFS2 isimlerindeki on üç farklı öznitelik ağırlıklandırma yaklaşımı ele alınmıştır. Reuters-21578, Enron1, WebKB ve SpamSMS veri setleri üzerinde, SVM, DT, kNN ve MNB sınıflandırma teknikleriyle deneyler yapılmıştır. DFS öznitelik seçme stratejisi tercih edilmiştir. Analizler için, 50 ile 1000 arasındaki beş farklı öznitelik boyutu üzerinde sonuçlar karşılaştırılmıştır. Sonuç olarak, dengesiz metin sınıflandırmada, hem iki sınıflı hem de çok sınıflı veri setleri için en etkili öznitelik ağırlıklandırma yöntemleri tanımlanmıştır. Deneyin kapsamlı bir özeti, Şekil 6.2'de metin sınıflandırma şemasında gösterilmektedir.



Şekil 6.2. *Deney çalışmasının genel şeması*

Deneyler sonucunda elde edilen Makro-F1 değerlerine göre, iki ve çok sınıflı veri setlerinde en iyi skorları üreten öznitelik ağırlıklandırma yöntemleri Tablo 6.3 ve Tablo 6.4'te sunulmuştur.

Tablo 6.3. *Öznitelik ağırlıklandırma metotlarının iki sınıflı veri setlerinde en iyi Makro-F1 değerlerini ürettiği durumların sayıları*

TF_MMR	TF_MEFS2	TF_EFS	TF_MGSS	TF_GSS	TF_CICI	TF_CMFS
5	2	2	1	1	1	1

Tablo 6.4. *Öznitelik ağırlıklandırma metotlarının çok sınıflı veri setlerinde en iyi Makro-F1 değerlerini ürettiği durumların sayıları*

TF_MMR	TF_MMMR	TF_MEFS1	TF_MEFS2
3	1	1	1

6.3. Sonuçlar

İki ve çok sınıflı veri setleri, düzenlenmiş öznitelik seçme ve ağırlıklandırma metotları göz önünde bulundurularak değerlendirmeler yapıldığında, çok sınıflı veri setlerinde düzenlenmiş öznitelik seçme metotları istenen başarıyı gösterememiştir. Öznitelik ağırlıklandırma kategorisinde ise hem iki sınıflı hem çok sınıflı veri setlerinde ön plana çıkan düzenlenmiş öznitelik ağırlıklandırma metodu görülmemektedir.

Düzenlenmiş öznitelik seçme yöntemlerinin iki sınıflı veri setlerindeki performanslarına bakıldığında ise, MEFS1 ve MEFS2 metotları çoğunlukla yüksek Makro-F1 değerleri elde etmiştir. Bu sayede, MEFS1 ve MEFS2, EFS_IMP1 ve EFS_IMP2 adlarını alarak, tez kapsamında filtre bazlı iki öznitelik seçme metodu olarak önerilmiştir.

7. ÖNERİLEN METOTLAR : EFS_IMP1 VE EFS_IMP2

7.1. Motivasyon

Literatürde, dengesiz metin sınıflandırma için önerilen az sayıda öznitelik seçme yöntemi bulunması bu alanın halen geliştirmeye açık olduğuna işaret etmektedir. Bu çalışmanın ana katkısı, özellikle iki sınıflı veri setlerinde başarılı sonuçlar veren EFS_IMP1 ve EFS_IMP2 adlı iki yeni filtre bazlı öznitelik seçme yönteminin dengesiz metin sınıflandırma için geliştirilmesidir.

7.2. Çalışma Stratejisi

Birçok sınıflandırıcı, büyük sınıfı (major) seçip küçük olanı (minor) göz ardı etmektedir (Kübler, Liu,, & Sayyed, 2018). Önerilen iki öznitelik seçme yöntemi (EFS_IMP1 ve EFS_IMP2) küçük sınıfın dokümanlarını vurgulamayı amaçlamaktadır.

EFS, terimlerin ayrım kapasitesini belirleyebilir, ancak küçük sınıfın ilgili öznitelikleri EFS tarafından yeterince ayırt edilemez. Yani, EFS küçük sınıftaki her ilgili özniteliğe aynı ağırlıkları vermektedir.

EFS_IMP1 ve EFS_IMP2, terimlerin küçük ve büyük sınıflara olan ilgililik derecesini hesaplarken küçük ve büyük sınıf dağılımlarını dikkate almaktadır. Önerilen yöntemler doküman temsilinde kullanıldığında, küçük ve büyük sınıf dokümanları arasındaki farklılık artmakta ve sınıflandırıcıların performansları iyileşmektedir. Sonuçlar, öznitelik seçme yöntemlerinin terimlerin ayırt edici gücünden ziyade ilgili olma gücünün incelemesi gerektiğini göstermektedir (Naderalvojoud & Sezer, 2020).

EFS'nin formülü, özniteliklerin sınıflarda geçme sayıları (a, b, c, d) ve olasılıklar açısından sırasıyla Eşitlik 7.1 ve Eşitlik 7.2'de verilmiştir.

$$EFS(t) = \sum_{j=1}^{M} \frac{a/(a+b)}{(b/(a+b))+(c/(c+d))+1} * \frac{a/(a+c)}{(c/(a+c))+(b/(b+d))+1} \quad \textbf{(7.1)}$$

$$EFS(t) = \sum_{j=1}^{M} \left(\frac{P(t|C_j)}{P(\bar{t}|C_j)+P(t|\overline{C_j})+1} \right) * \left(\frac{P(C_j|t)}{P(\overline{C_j}|t)+P(C_j|\bar{t})+1} \right) \quad \textbf{(7.2)}$$

EFS formülünün sınıf tabanlı ve koleksiyon tabanlı ayrımı gösteren iki parçası vardır. EFS'nin modifiye edilmiş versiyonlarında bu formülün her iki bölümünü de geliştirilmiştir.

Aşağıdaki bölümlerde, önerilen geliştirilmiş EFS versiyonlarındaki değişiklikler ve eklemeler açıklanacaktır.

EFS_IMP1 için yapılan değişiklikler şunlardır:

EFS_IMP1'in sınıf tabanlı bölümü için EFS'nin sınıf tabanlı bölümünün paydasından (b/(a+b)) çıkarılmıştır. Koleksiyon tabanlı bölümü için ise, EFS_IMP1'in sınıf tabanlı bölümünün payı (a+c) ile bölünmüş ve EFS_IMP1'in sınıf tabanlı bölümünün paydası (a+c) ile çarpılmıştır.

Sınıf tabanlı bölümde;

$\frac{b}{a+b} = \frac{P\left(\bar{t}, C_j\right)}{P\left(C_j\right)} = P\left(\bar{t} \mid C_j\right)$ EFS'nin sınıf tabanlı bölümünün paydasından çıkarılmıştır.

Koleksiyon tabanlı bölümde;

$\frac{a}{a+b} = P\left(t \mid C_j\right)$ EFS_IMP1'in sınıf tabanlı bölümünün payı $(a+c) = P\left(t\right) * N$ ile bölünmüştür.

$\frac{c}{c+d} = P\left(t \mid \overline{C_j}\right)$ EFS_IMP1'in sınıf tabanlı bölümünün paydası $(a+c) = P\left(t\right) * N$ ile çarpılmıştır.

EFS_IMP1'in formülü, özniteliklerin sınıflarda geçme sayıları (a, b, c, d) ve olasılıklar açısından sırasıyla Eşitlik 7.3 ve Eşitlik 7.4'te verilmiştir.

$$EFS\_IMP1\left(t\right) = \sum_{j=1}^{M} \left(\frac{a/\left(a+b\right)}{\left(a+c\right) * \left(\left(c/\left(c+d\right)\right)+1\right)} \right) * \left(\frac{a/\left(a+b\right)}{\left(\left(c*\left(a+c\right)\right)/\left(c+d\right)\right)+1} \right) \quad \textbf{(7.3)}$$

$$EFS\_IMP1(t) = \sum_{j=1}^{M} \frac{P\left(t \mid C_j\right)}{P(t) * N * \left(P\left(t \mid \overline{C_j}\right)+1\right)} * \frac{P\left(t \mid C_j\right)}{P\left(t \mid \overline{C_j}\right) * P\left(t\right) * N + 1} \quad \textbf{(7.4)}$$

$P\left(t \mid C\right)$: Terim t'nin ilgili C_j sınıfıyla ilişkili olma olasılığı artar; bu nedenle, yüksek puan alması gerekmektedir.

$P(t \mid \overline{C})$: Terim t'nin ilgili C_j sınıfıyla ilgisiz olma olasılığı düşer; bu nedenle, düşük puan alması gerekmektedir.

$P(t)$: düşük puan alması gerekmektedir.

EFS_IMP2 için yapılan değişiklikler şunlardır:

EFS_IMP2'in sınıf tabanlı bölümü için EFS'nin sınıf tabanlı bölümünün paydasından (b/(a+b)) çıkarılmıştır. Koleksiyon tabanlı bölümü için ise, EFS_IMP2'nin sınıf tabanlı bölümünün payı (a+c) ile bölünmüş ve EFS_IMP2'nin sınıf tabanlı bölümünün paydası (a+c) ile bölünmüştür.

Sınıf tabanlı bölümde;

$\frac{b}{a+b} = \frac{P(\bar{t}, C_j)}{P(C_j)} = P(\bar{t} \mid C_j)$ EFS'nin sınıf tabanlı bölümünün paydasından çıkarılmıştır.

Koleksiyon tabanlı bölümde;

$\frac{a}{a+b} = P(t \mid C_j)$ EFS_IMP2'nin sınıf tabanlı bölümünün payı $(a+c) = P(t) * N$ ile bölünmüştür.

$\frac{c}{c+d} = P(t \mid \overline{C_j})$ EFS_IMP2'nin sınıf tabanlı bölümünün paydası $(a+c) = P(t) * N$ ile çarpılmıştır.

EFS_IMP2'nin formülü, özniteliklerin sınıflarda geçme sayıları (a, b, c, d) ve olasılıklar açısından sırasıyla Eşitlik 7.5 ve Eşitlik 7.6'da verilmiştir.

$$EFS\_IMP2(t) = \sum_{j=1}^{M} \left(\frac{a/(a+b)}{(a+c)*((c/(c+d))+1)} \right) * \left(\frac{a/(a+b)}{((c/((a+c)*(c+d)))+1} \right) \quad \textbf{(7.5)}$$

$$EFS\_IMP2(t) = \sum_{j=1}^{M} \frac{P(t \mid C_j)}{P(t \mid \overline{C_j})+1} * \frac{P(t \mid C_j)}{P(t \mid \overline{C_j})+P(t)*N} \quad \textbf{(7.6)}$$

$P(t \mid C)$: Terim t'nin ilgili C_j sınıfıyla ilişkili olma olasılığı artar; bu nedenle, yüksek puan alması gerekmektedir.

$P(t \mid \overline{C})$: Terim t'nin ilgili C_j sınıfıyla ilgisiz olma olasılığı düşer; bu nedenle, düşük puan alması gerekmektedir.

$P(t)$: düşük puan alması gerekmektedir.

Tablo 7.1'de, EFS_IMP1 ve EFS_IMP2 yöntemlerinin nasıl çalıştığını göstermek için bir örnek koleksiyon verilmiştir.

Tablo 7.1. *Örnek koleksiyon*

Doküman adı	**İçerik**				**Sınıf**
D1	brown	red	green		X
D2	brown	yellow	blue	green	Y
D3	brown	yellow	blue		Y
D4	brown	yellow	green		Y

$$EFS\_IMP1('brown') = \left(\frac{1/(1+0)}{(3/(3+0))+1}\right) * \left(\frac{1/((1+0)*(1+3))}{((3*(1+3))/(3+0))+1}\right) + \left(\frac{3/(3+0)}{(1/(1+0))+1}\right) * \left(\frac{3/((3+0)*(3+1))}{((1*(3+1))/(1+0))+1}\right) = 0.050$$

$$EFS\_IMP1('red') = \left(\frac{1/(1+0)}{(0/(0+3))+1}\right) * \left(\frac{1/((1+0)*(1+0))}{((0*(1+0))/(0+3))+1}\right) + \left(\frac{0/(0+3)}{(1/(1+0))+1}\right) * \left(\frac{0/((0+3)*(0+1))}{((1*(0+1))/(1+0))+1}\right) = 1.000$$

$$EFS\_IMP1('yellow') = \left(\frac{0/(0+1)}{(3/(3+0))+1}\right) * \left(\frac{0/((0+1)*(0+3))}{((3*(0+3))/(3+0))+1}\right) + \left(\frac{3/(3+0)}{(0/(0+1))+1}\right) * \left(\frac{3/((3+0)*(3+0))}{((0*(3+0))/(0+1))+1}\right) = 0.333$$

$$EFS\_IMP1('blue') = \left(\frac{0/(0+1)}{(2/(2+1))+1}\right)*\left(\frac{0/((0+1)*(0+2))}{((2*(0+2))/(2+1))+1}\right)+$$

$$\left(\frac{2/(2+1)}{(0/(0+1))+1}\right)*\left(\frac{2/((2+1)*(2+0))}{((0*(2+0))/(0+1))+1}\right)=0.222$$

$$EFS\_IMP1('green') = \left(\frac{1/(1+0)}{(2/(2+1))+1}\right)*\left(\frac{1/((1+0)*(1+2))}{((2*(1+2))/(2+1))+1}\right)+$$

$$\left(\frac{2/(2+1)}{(1/(1+0))+1}\right)*\left(\frac{2/((2+1)*(2+1))}{((1*(2+1))/(1+0))+1}\right)=0.084$$

$$EFS\_IMP2('brown') = \left(\frac{1/(1+0)}{(3/(3+0))+1}\right)*\left(\frac{1/((1+0)*(1+3))}{(3/((1+3)*(3+0)))+1}\right)+$$

$$\left(\frac{3/(3+0)}{(1/(1+0))+1}\right)*\left(\frac{3/((3+0)*(3+1))}{(1/((3+1)*(1+0)))+1}\right)=0.200$$

$$EFS\_IMP2('red') = \left(\frac{1/(1+0)}{(0/(0+3))+1}\right)*\left(\frac{1/((1+0)*(1+0))}{(0/((1+0)*(0+3)))+1}\right)+$$

$$\left(\frac{0/(0+3)}{(1/(1+0))+1}\right)*\left(\frac{0/((0+3)*(0+1))}{(1/((0+1)*(1+0)))+1}\right)=1.000$$

$$EFS\_IMP2('yellow') = \left(\frac{0/(0+1)}{(3/(3+0))+1}\right)*\left(\frac{0/((0+1)*(0+3))}{(3/((0+3)*(3+0)))+1}\right)+$$

$$\left(\frac{3/(3+0)}{(0/(0+1))+1}\right)*\left(\frac{3/((3+0)*(3+0))}{(0/((3+0)*(0+1)))+1}\right)=0.333$$

$$EFS\_IMP2('blue') = \left(\frac{0/(0+1)}{(2/(2+1))+1}\right)*\left(\frac{0/((0+1)*(0+2))}{(2/((0+2)*(2+1)))+1}\right)+$$

$$\left(\frac{2/(2+1)}{(0/(0+1))+1}\right)*\left(\frac{2/((2+1)*(2+0))}{(0/((2+0)*(0+1)))+1}\right)=0.222$$

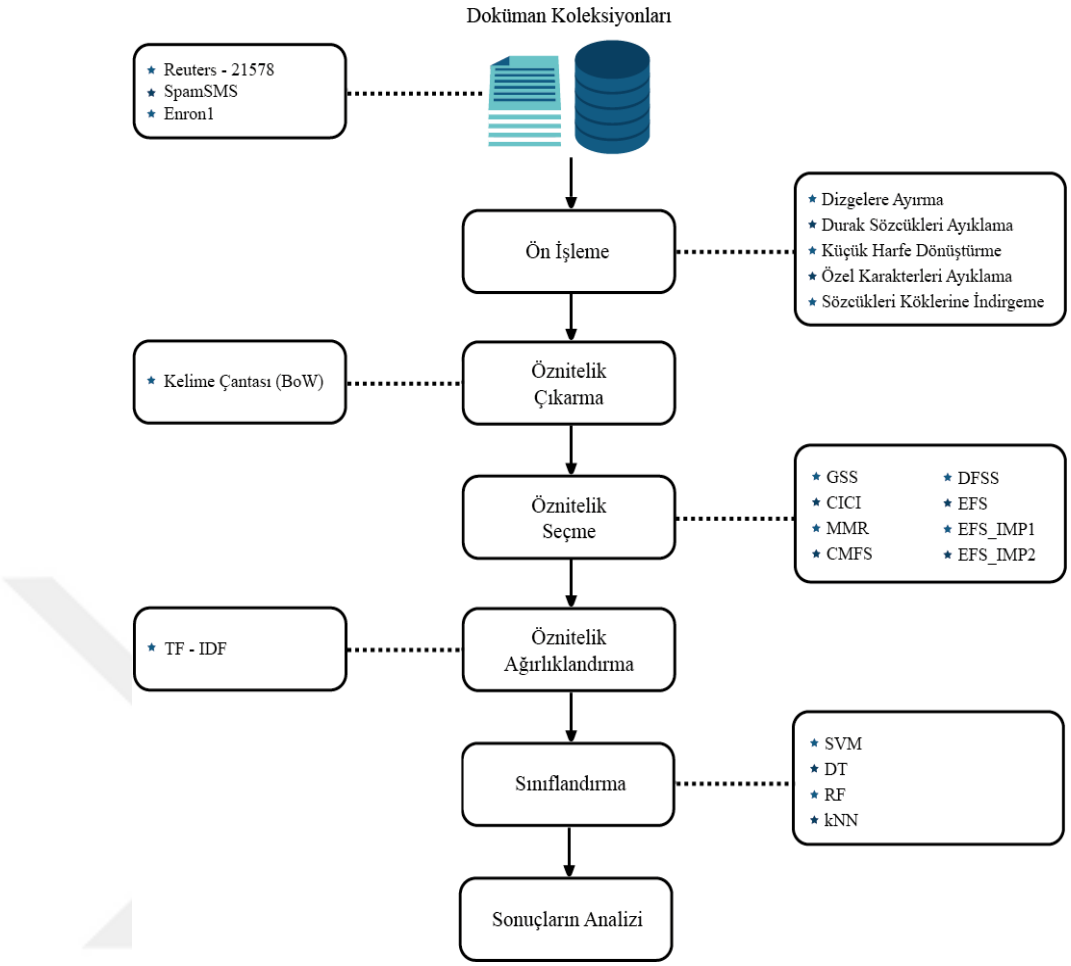
$$EFS\_IMP2('green') = \left(\frac{1/(1+0)}{(2/(2+1))+1}\right) * \left(\frac{1/((1+0)*(1+2))}{(2/((1+2)*(2+1)))+1}\right) +$$
$$\left(\frac{2/(2+1)}{(1/(1+0))+1}\right) * \left(\frac{2/((2+1)*(2+1))}{(1/((2+1)*(1+0)))+1}\right) = 0.218$$

Bu örnek koleksiyonda EFS_IMP1 ve EFS_IMP2 için, 'red' terimi küçük sınıf (minor) X'te bulunmaktadır ve tüm terimler arasında en yüksek puanı almaktadır. 'yellow' ve 'blue' terimlerini içeren dokümanlar aynı sınıfta, büyük sınıf (major) Y'dedir. Ancak bu iki terim aynı puanı alamaz çünkü farklı doküman frekanslarına sahiptirler: 'yellow' terimi 3 ve 'blue' terimi 2'dir. Puana bakıldığında 'yellow' terimi, 'blue' teriminden daha ilgilidir. 'green' terimi hem küçük (minor) hem de büyük (major) sınıflarda bulunmaktadır. Dolayısıyla, 'blue' terimi, 'green' teriminden daha ilgilidir. Hem küçük (minor) hem de büyük (major) sınıftaki her doküman, koleksiyondaki en ilgisiz terim olan 'brown' terimini içermektedir. EFS_IMP1'de 'red' terimi için en büyük puan olan 1.0 hesaplanırken, 'brown' terimi için en düşük puan olan 0.050 hesaplanmıştır. EFS_IMP2'de ise 'red' terimi için en büyük puan olan 1.0 hesaplanırken, 'brown' terimi için en düşük puan olan 0.200 hesaplanmıştır.

Kısacası EFS_IMP1 ve EFS_IMP2 terimleri ayırt edici güç yerine, ilgili güçlerine göre 'red', 'yellow', 'blue', 'green' ve 'brown' olarak sıralamaktadır.

7.3. Deneysel Çalışmalar

Deneylerin akışını gösteren metin sınıflandırma şeması Şekil 7.1'de gösterilmiştir.



Şekil 7.1. *Deney çalışmasının genel şeması*

7.3.1. Veri setleri

Bu çalışmada, metin sınıflandırma için iki sınıflı (binary) veri setleri test edilmiştir. Deneysel çalışmalarda Spam SMS, Enron1 ve Reuters-21578 (R8 MM) veri setleri dengesiz veri setleri olarak kullanılmıştır.

7.3.2. Başarı kriteri : Makro-F1

Bu başlık altında anlatılacak olan Makro-F1 bilgilerine 5.3.Değerlendirme Ölçütü başlığı altında yer verilmiştir.

7.3.3. Benzerlik analizi

Üç veri seti için Tablo 7.2, Tablo 7.3 ve Tablo 7.4'te görüleceği üzere, önerilen yöntemlerin en yüksek puan alan on özniteliği, rekabetçi diğer yöntemlerin en yüksek puan alan on özniteliği ile karşılaştırılmıştır. Sadece belirtilen öznitelik seçme yöntemi tarafından seçilen öznitelikler kalın yazı tipiyle vurgulanmıştır. Dolayısıyla, vurgulanan

öznitelikler tek bir öznitelik seçme yöntemine özgüdür, bu durumda önerilen EFS_ IMP1 ve EFS_IMP2 yöntemlerinin karşılaştırma için kullanılan altı diğer öznitelik seçme yönteminde seçilen ortak öznitelikleri seçtiği sonucuna varılabilmektedir. Sonuç olarak, diğer öznitelik seçme yöntemlerinden çok farklı öznitelikler içeren bir öznitelik seti oluşturmak için yeni bir yöntem gerekli değildir. Ancak, EFS_ IMP1 ve EFS_IMP2 tarafından oluşturulan öznitelik setlerinin EFS ile aynı olmadığı görülebilmektedir.

Tablo 7.2. *Enron1 veri seti için seçilen ilk 10 öznitelik*

Metotlar	**Öznitelikler**
GSS	enron, hpl, daren, hou, mmbtu, **xls, sitara, pat, med, teco**
CICI	cc, gas, enron, ect, forward, pm, http, meter, **corp,** attach
MMR	meter, nom, ect, cc, gas, volum, pm, weight, attach, **medic**
CMFS	subject, enron, cc, hpl, forward, gas, ect, daren, hou, pm
DFSS	ect, subject, cc, pm, deal, volum, meter, http, **width, height**
EFS	subject, enron, cc, hpl, gas, forward, ect, daren, hou, pm
EFS_IMP1	enron, hpl, daren, hou, meter, cc, ect, gas, nom, mmbtu
EFS_IMP2	subject, enron, cc, forward, hpl, gas, ect, deal, daren, hou

Tablo 7.3. *Spam SMS veri seti için seçilen ilk 10 öznitelik*

Metotlar	**Öznitelikler**
GSS	claim, prize, uk, tone, guarantee, **ppm, cs, pobox, land, voucher**
CICI	call, â, txt, free, text, mobil, stop, ur, **www,** repli
MMR	www, txt, box, nokia, award, â, service, mobil, rate, **landlin**
CMFS	call, ll, â, good, love, time, day, txt, ur, home
DFSS	â, txt, free, mobil, call, www, nokia, service, **box, award**
EFS	â, txt, call, free, claim, mobil, www, prize, uk, servic
EFS_IMP1	claim, â, txt, prize, uk, www, call, tone, free, box
EFS_IMP2	call, â, free, txt, ur, text, claim, mobil, stop, repli

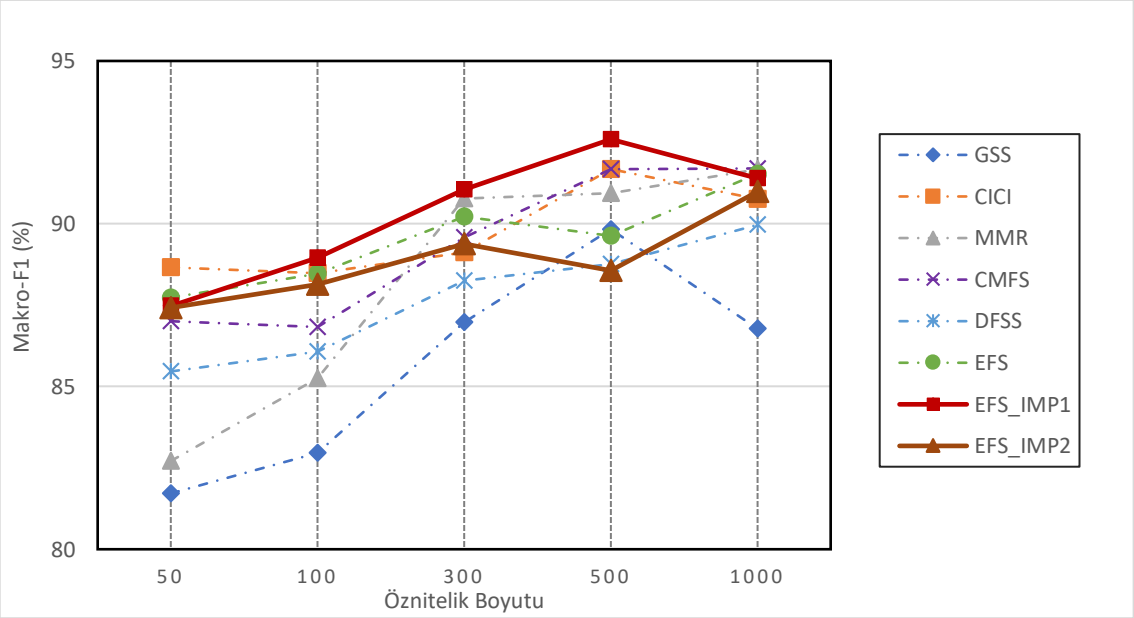
Tablo 7.4. *Reuters-21578 (R8 MM) veri seti için seçilen ilk 10 öznitelik*

Metotlar	**Öznitelikler**
GSS	shr, qtr, rev, vessel, dividend, div, **avg, qtly, iran, labour**
CICI	cts, ship, net, mln, shr, port, note, qtr, **loss,** dlr
MMR	cts, c**argo,** tanker, ship, net, port, **minist, coast,** strike, attack
CMFS	march, cts, net, mln, shr, year, dlr, qtr, note, rev
DFSS	ship, port, strike, cts, **grain,** tanker, **tonn,** net, attack, **offici**
EFS	ship, cts, net, march, mln, shr, port, vessel, qtr, year
EFS_IMP1	cts, shr, net, qtr, rev, note, dividend, **profit,** div, **record**
EFS_IMP2	march, cts, net, mln, year, shr, dlr, qtr, note, rev

7.3.4. Doğruluk analizi

Öznitelikler seçildikten sonra SVM, DT, RF ve kNN sınıflandırıcıları kullanılmıştır. Öznitelik alt kümelerini oluşturmak için 50, 100, 300, 500 ve 1000 öznitelik boyutları tercih edilmiştir. Elde edilen Makro - F1 değerleri Tablo 7.5 - Tablo 7.16'da verilirken, en yüksek puanlar kalın yazı tipiyle vurgulanmıştır. Tüm öznitelik boyutlarında ve veri setlerinde en iyi sonucu veren tek bir yöntem yoktur. Bununla birlikte, ikili (binary) veri setlerinde EFS_IMP1 ve EFS_IMP2, en yüksek Makro-F1 puanlarını alarak, diğer tüm yöntemleri geride bırakmaktadır.

Enron1 veri seti için SVM sınıflandırıcısı kullanıldığında oluşan sınıflandırma performansları Şekil 7.2'de gösterilmektedir.



Şekil 7.2. *Enron1 veri seti için SVM sınıflandırıcısı kullanıldığında Makro-F1 skorları*

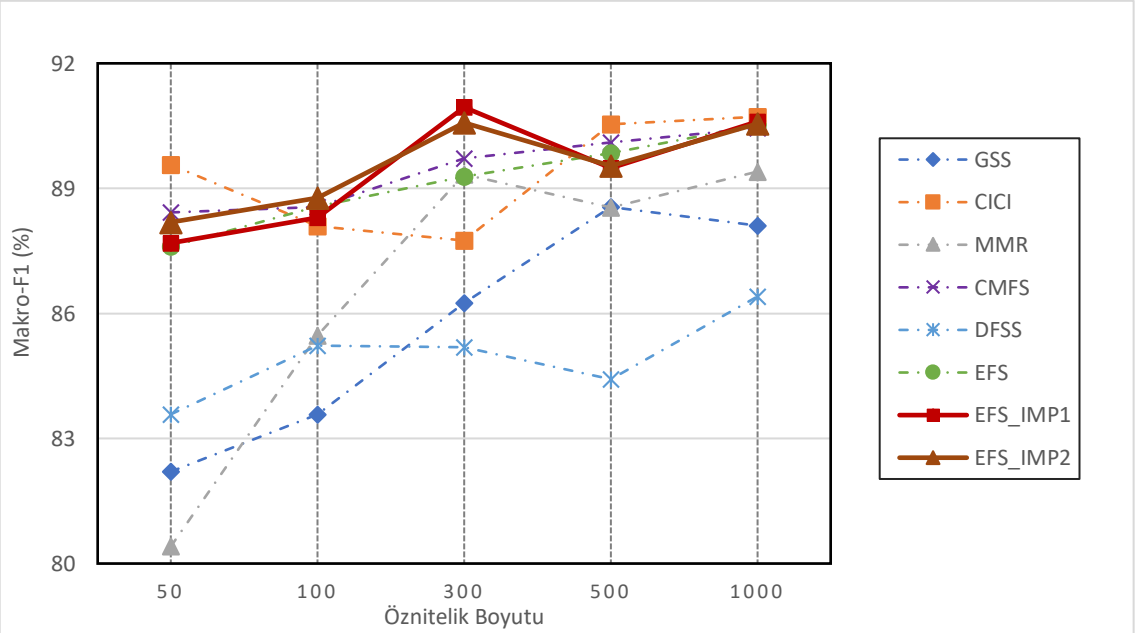
EFS_IMP1, SVM sınıflandırıcısı kullanılarak Enron1 veri setinde 92,594 ile en yüksek Makro-F1 puanını 500 öznitelik boyutunda almıştır. EFS _IMP1'in ayrıca 100 ve 300 öznitelik boyutlarında da başarılı olduğunu belirtmek önemlidir.

Enron1 veri seti için SVM sınıflandırıcı kullanıldığında elde edilen Makro-F1 değerleri Tablo 7.5'te verilmiştir. EFS_IMP1 metodunun Makro-F1 değerleri birçok öznitelik boyutunda diğer öznitelik seçme metodundan daha başarılıdır.

Tablo 7.5. *Enron1 veri seti için Makro-F1 skorları (a) SVM*

Öznitelik Seçme Metotları	Öznitelik Sayıları				
	50	**100**	**300**	**500**	**1000**
GSS	81,717	82,960	86,967	89,829	86,770
CICI	88,667	88,479	89,122	91,677	90,754
MMR	82,713	85,258	90,779	90,937	91,668
CMFS	87,012	86,827	89,573	91,668	91,698
DFSS	85,463	86,065	88,250	88,762	89,963
EFS	87,731	88,457	90,221	89,623	91,532
EFS_IMP1	87,472	88,947	91,046	**92,594**	91,388
EFS_IMP2	87,420	88,141	89,381	88,569	90,983

Önerilen metotlar ve mevcut metotların Enron1 veri seti için DT sınıflandırıcı kullanıldığında elde edilen sınıflandırma performansları Şekil 7.3'te sunulmuştur.



Şekil 7.3. *Enron1 veri seti için DT sınıflandırıcısı kullanıldığında Makro-F1 skorları*

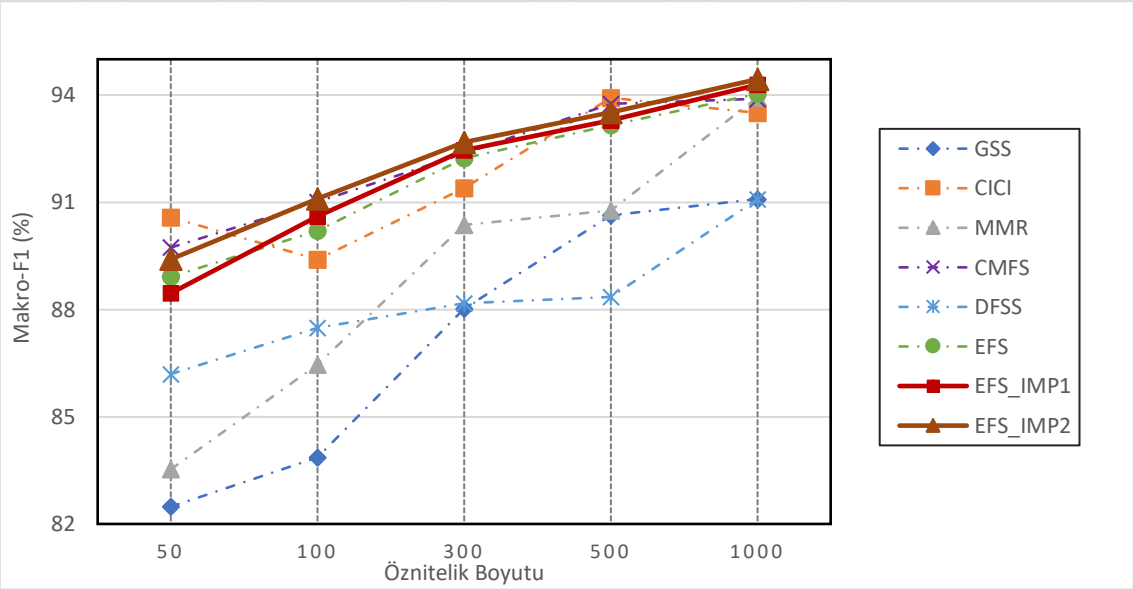
DT sınıflandırıcısı kullanılarak Enron1 veri setinde, 300 öznitelik boyutunda EFS_IMP1, 90,946 ile en yüksek Makro-F1 puanını alır. CICI ve EFS _IMP2'nin de çeşitli öznitelik boyutlarında başarılı olduğunu belirtmek gerekir.

Enron1 veri seti için DT sınıflandırıcı kullanıldığında elde edilen Makro-F1 değerleri Tablo 7.6'da verilmiştir.

Tablo 7.6. *Enron1 veri seti için Makro-F1 skorları (b) DT*

Öznitelik Seçme Metotları	Öznitelik Sayıları				
	50	**100**	**300**	**500**	**1000**
GSS	82,206	83,580	86,250	88,553	88,103
CICI	89,566	88,090	87,749	90,536	90,719
MMR	80,415	85,465	89,337	88,542	89,402
CMFS	88,424	88,551	89,720	90,101	90,445
DFSS	83,581	85,232	85,187	84,427	86,405
EFS	87,603	88,569	89,288	89,849	90,516
EFS_IMP1	87,690	88,305	**90,946**	89,486	90,600
EFS_IMP2	88,189	88,777	90,576	89,543	90,551

Enron1 veri setinde RF sınıflandırıcısı kullanıldığında hesaplanan sınıflandırma performansları Şekil 7.4'te gösterilmektedir.



Şekil 7.4. *Enron1 veri seti için RF sınıflandırıcısı kullanıldığında Makro-F1 skorları*

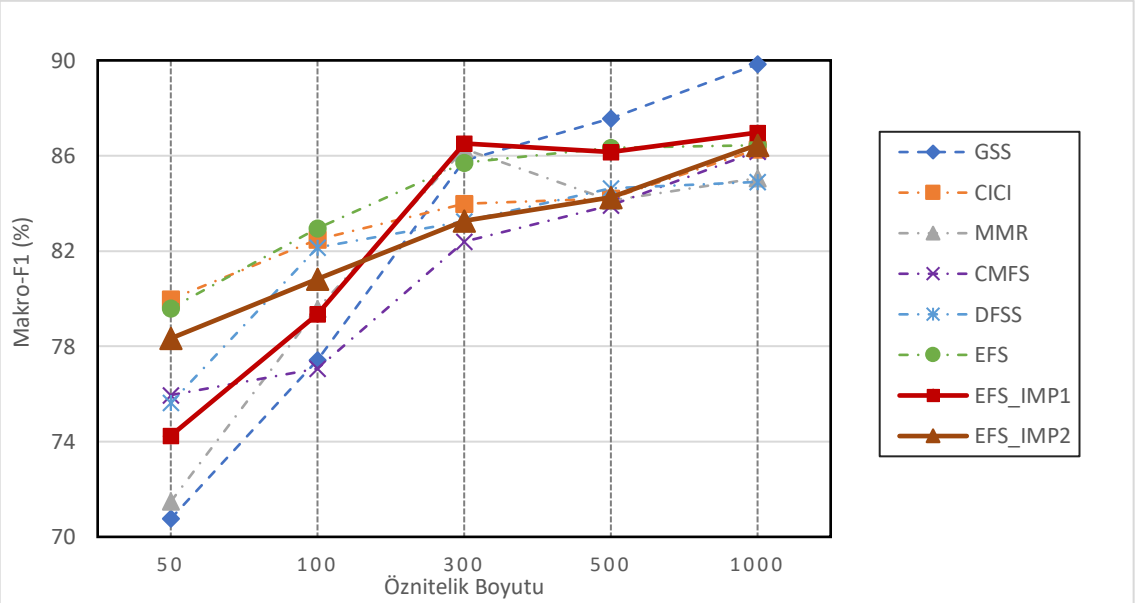
RF sınıflandırıcısı kullanılarak Enron1 veri setinde, 1000 öznitelik boyutunda EFS_IMP2, 94,441 ile en yüksek Makro-F1 puanını alır. CICI'nin de çeşitli öznitelik boyutlarında başarılı olduğunu belirtmek gerekir.

Enron1 veri seti için RF sınıflandırıcı kullanıldığında elde edilen Makro-F1 değerleri Tablo 7.7'de verilmiştir. EFS_IMP2 metodunun Makro-F1 değerleri birçok öznitelik boyutunda diğer öznitelik seçme metodundan daha başarılıdır.

Tablo 7.7. *Enron1 veri seti için Makro-F1 skorları (c) RF*

Öznitelik Seçme Metotları	Öznitelik Sayıları				
	50	**100**	**300**	**500**	**1000**
GSS	82,482	83,860	88,021	90,636	91,089
CICI	90,569	89,401	91,392	93,921	93,493
MMR	83,531	86,451	90,373	90,766	93,929
CMFS	89,733	91,006	92,367	93,749	93,894
DFSS	86,194	87,488	88,167	88,348	91,079
EFS	88,914	90,199	92,237	93,159	94,030
EFS_IMP1	88,457	90,598	92,455	93,289	94,285
EFS_IMP2	89,407	91,096	92,684	93,519	**94,441**

Enron1 veri setinde kNN sınıflandırıcısı kullanıldığında hesaplanan sınıflandırma performansları Şekil 7.5'te verilmiştir.



Şekil 7.5. *Enron1 veri seti için kNN sınıflandırıcısı kullanıldığında Makro-F1 skorları*

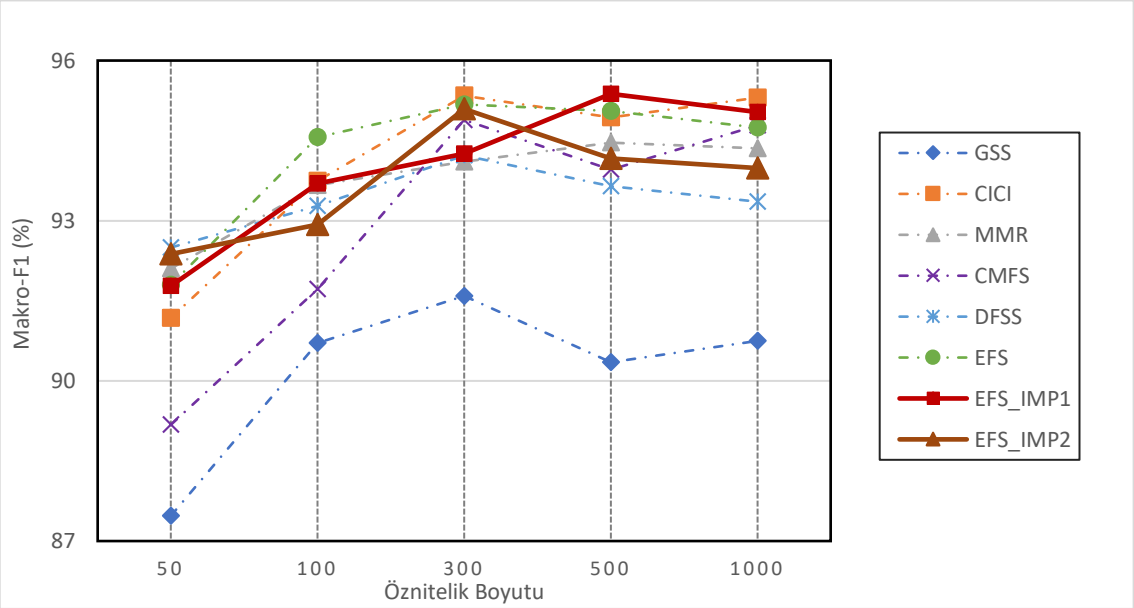
kNN sınıflandırıcısı kullanılarak Enron1 veri setinde, 1000 öznitelik boyutunda GSS, 89,837 ile en yüksek Makro-F1 puanını alır. CICI, EFS ve EFS_IMP1'in de çeşitli öznitelik boyutlarında başarılı olduğunu belirtmek gerekir.

Enron1 veri seti için kNN sınıflandırıcı kullanıldığında elde edilen Makro-F1 değerleri Tablo 7.8'de verilmiştir. RF sınıflandırıcısının, en yüksek puanlara göre Enron1 veri setinde SVM, kNN ve DT sınıflandırıcılarını geride bıraktığı belirtilebilir.

Tablo 7.8. *Enron1 veri seti için Makro-F1 skorları (d) kNN*

Öznitelik Seçme Metotları	Öznitelik Sayıları				
	50	**100**	**300**	**500**	**1000**
GSS	70,770	77,425	85,788	87,558	**89,837**
CICI	79,976	82,490	83,978	84,206	86,268
MMR	71,490	79,562	86,283	84,110	85,066
CMFS	75,953	77,060	82,384	83,921	86,180
DFSS	75,624	82,157	83,240	84,635	84,899
EFS	79,601	82,950	85,706	86,344	86,438
EFS_IMP1	74,253	79,349	86,516	86,153	86,970
EFS_IMP2	78,334	80,827	83,260	84,259	86,455

Spam SMS veri setinde SVM sınıflandırıcısı kullanıldığında hesaplanan sınıflandırma performansları Şekil 7.6'da gösterilmiştir.



Şekil 7.6. *Spam SMS veri seti için SVM sınıflandırıcısı kullanıldığında Makro-F1 skorları*

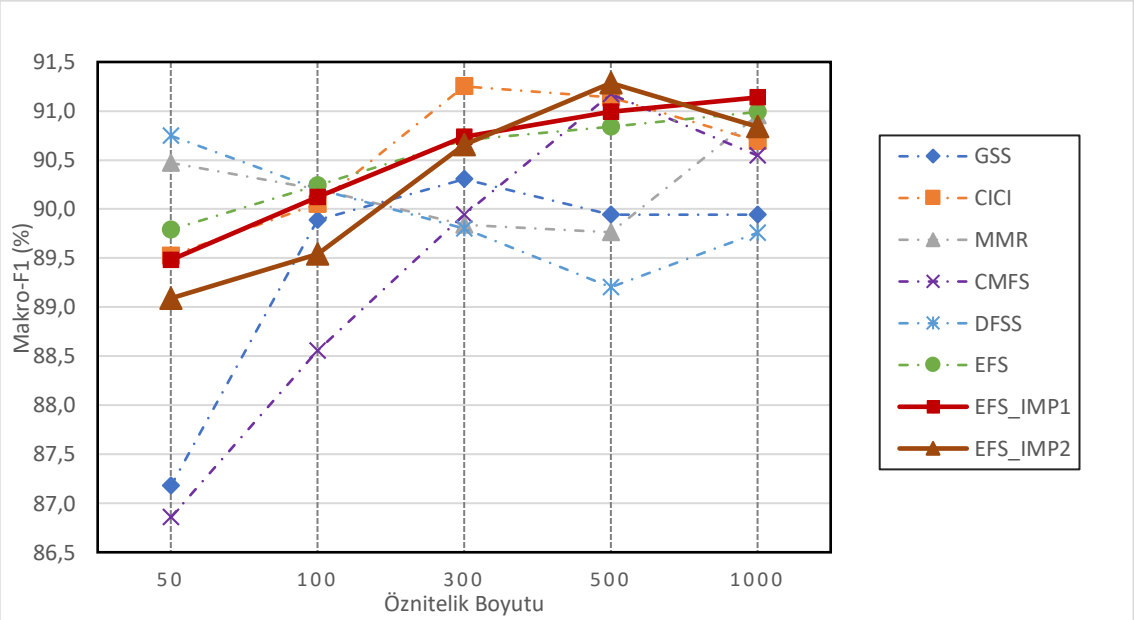
SVM sınıflandırıcısı kullanılarak Spam SMS veri setinde, 500 öznitelik boyutunda EFS_IMP1, 95,376 ile en yüksek Makro-F1 puanını alır. CICI, EFS ve DFSS'nin de çeşitli öznitelik boyutlarında başarılı olduğunu belirtmek gerekir.

Spam SMS veri seti için SVM sınıflandırıcı kullanıldığında elde edilen Makro-F1 değerleri Tablo 7.9'da verilmiştir.

Tablo 7.9. *Spam SMS veri seti için Makro-F1 skorları (a) SVM*

Öznitelik Seçme Metotları	Öznitelik Sayıları				
	50	**100**	**300**	**500**	**1000**
GSS	87,475	90,712	91,591	90,350	90,750
CICI	91,180	93,742	95,340	94,930	95,303
MMR	92,105	93,667	94,101	94,462	94,352
CMFS	89,180	91,713	94,888	93,958	94,787
DFSS	92,498	93,276	94,221	93,646	93,356
EFS	91,799	94,562	95,178	95,054	94,745
EFS_IMP1	91,780	93,692	94,252	**95,376**	95,034
EFS_IMP2	92,377	92,922	95,093	94,169	93,984

Spam SMS veri setinde DT sınıflandırıcısı kullanıldığında hesaplanan sınıflandırma performansları Şekil 7.7'de sunulmuştur.



Şekil 7.7. *Spam SMS veri seti için DT sınıflandırıcısı kullanıldığında Makro-F1 skorları*

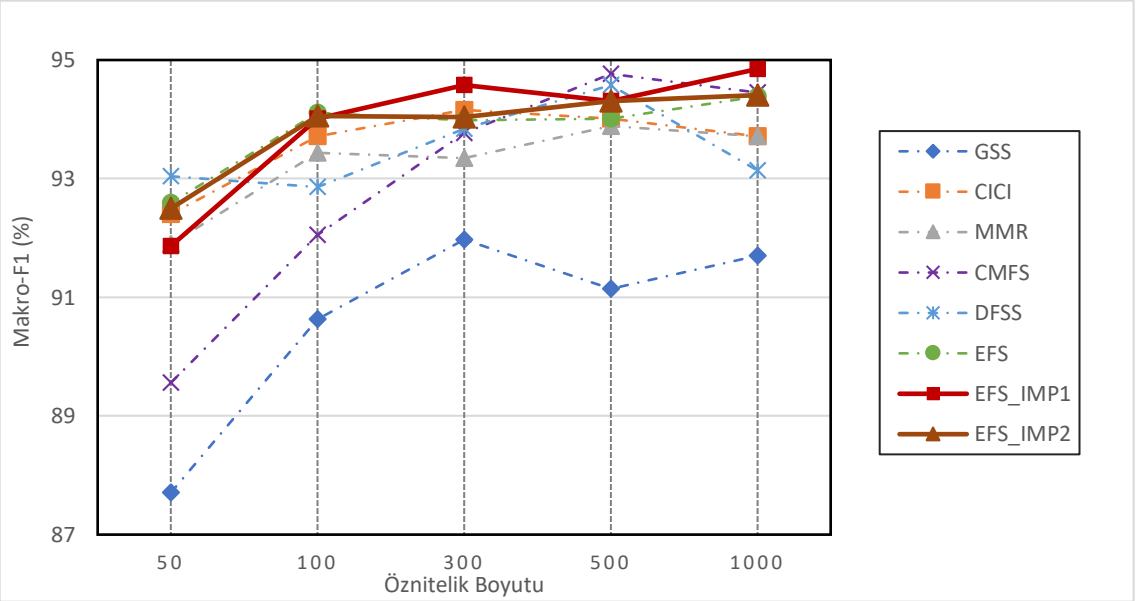
DT sınıflandırıcısı kullanılarak Spam SMS veri setinde, 500 öznitelik boyutunda EFS_IMP2, 91,285 ile en yüksek Makro-F1 puanını alır. CICI, DFSS, EFS ve EFS_IMP1'in de çeşitli öznitelik boyutlarında başarılı olduğunu belirtmek gerekir.

Spam SMS veri seti için DT sınıflandırıcı kullanıldığında elde edilen Makro-F1 değerleri Tablo 7.10'da verilmiştir.

Tablo 7.10. *Spam SMS veri seti için Makro-F1 skorları (b) DT*

Öznitelik Seçme Metotları	Öznitelik Sayıları				
	50	**100**	**300**	**500**	**1000**
GSS	87,177	89,887	90,305	89,943	89,943
CICI	89,526	90,053	91,251	91,138	90,693
MMR	90,471	90,202	89,840	89,763	90,955
CMFS	86,857	88,556	89,945	91,172	90,544
DFSS	90,747	90,202	89,802	89,203	89,756
EFS	89,791	90,241	90,702	90,842	90,990
EFS_IMP1	89,482	90,123	90,739	90,990	91,138
EFS_IMP2	89,090	89,540	90,657	**91,285**	90,842

Spam SMS veri setinde RF sınıflandırıcısı kullanıldığında hesaplanan sınıflandırma performansları Şekil 7.8'de gösterilmektedir.



Şekil 7.8. *Spam SMS veri seti için RF sınıflandırıcısı kullanıldığında Makro-F1 skorları*

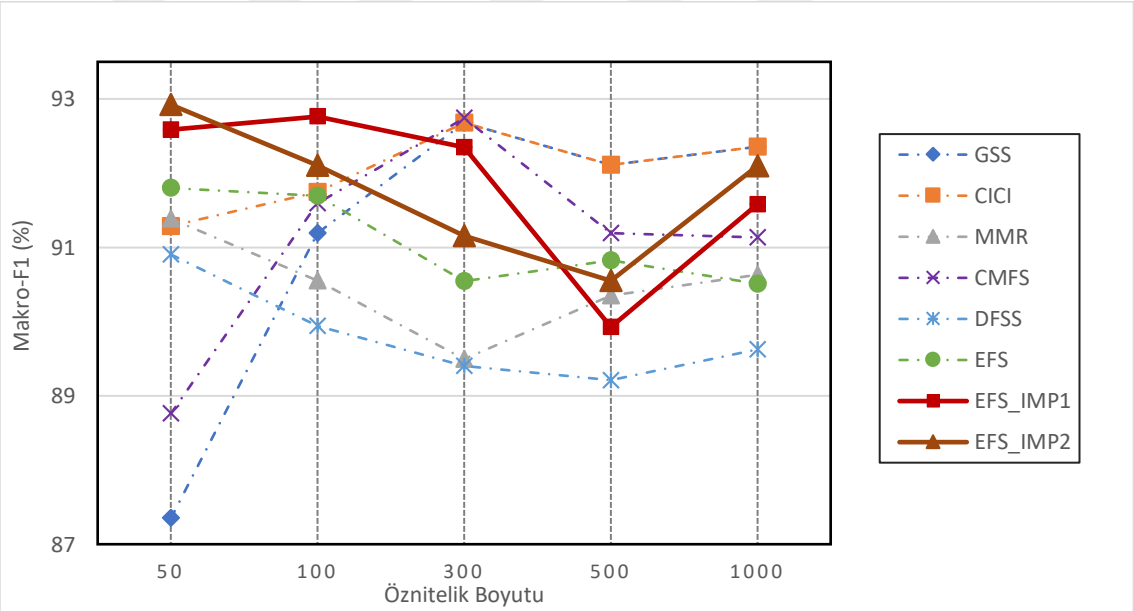
RF sınıflandırıcısı kullanılarak Spam SMS veri setinde, 1000 öznitelik boyutunda EFS_IMP1, 94,848 ile en yüksek Makro-F1 puanını alır. EFS _IMP1'in ayrıca 300 öznitelik boyutunda da başarılı olduğunu belirtmek önemlidir.

Spam SMS veri seti için RF sınıflandırıcı kullanıldığında elde edilen Makro-F1 değerleri Tablo 7.11'de verilmiştir.

Tablo 7.11. *Spam SMS veri seti için Makro-F1 skorları (c) RF*

Öznitelik Seçme Metotları	Öznitelik Sayıları				
	50	**100**	**300**	**500**	**1000**
GSS	87,709	90,631	91,973	91,143	91,697
CICI	92,404	93,713	94,159	94,011	93,713
MMR	91,899	93,439	93,343	93,887	93,713
CMFS	89,560	92,050	93,764	94,766	94,453
DFSS	93,043	92,858	93,837	94,577	93,138
EFS	92,589	94,106	93,986	94,011	94,385
EFS_IMP1	91,866	94,008	94,577	94,306	**94,848**
EFS_IMP2	92,498	94,059	94,035	94,306	94,408

Spam SMS veri setinde kNN sınıflandırıcısı kullanıldığında elde edilen sınıflandırma performansları Şekil 7.9'da verilmiştir.



Şekil 7.9. *Spam SMS veri seti için kNN sınıflandırıcısı kullanıldığında Makro-F1 skorları*

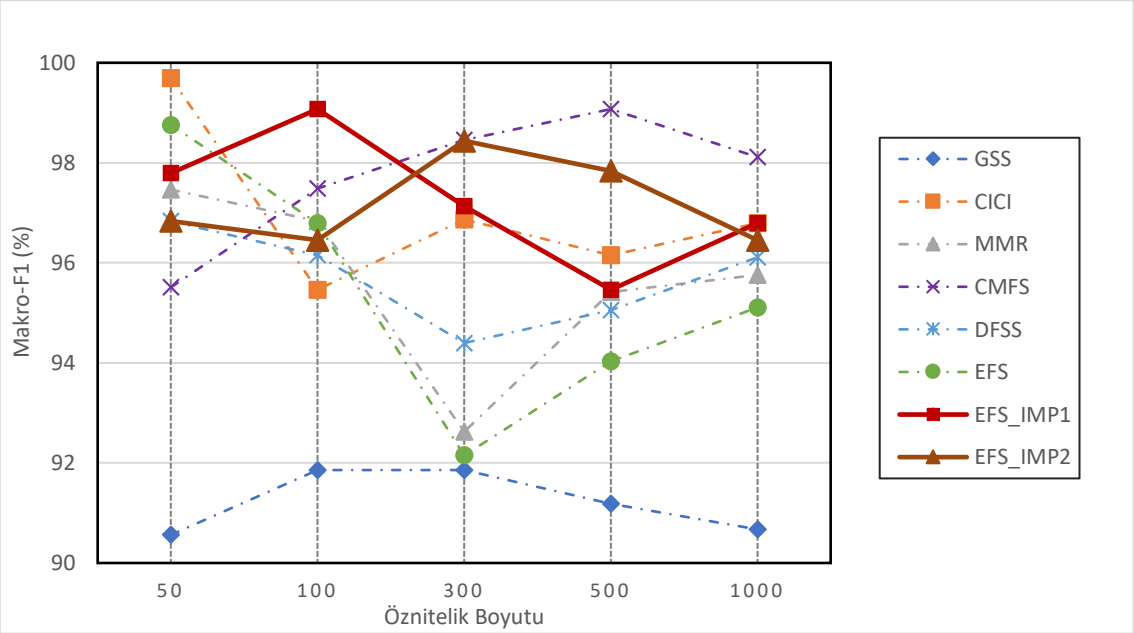
kNN sınıflandırıcısı kullanılarak Spam SMS veri setinde, 50 öznitelik boyutunda EFS_IMP2, 92,922 ile en yüksek Makro-F1 puanını alır. EFS_IMP1, CMFS, GSS ve CICI'nin de çeşitli öznitelik boyutlarında başarılı olduğunu belirtmek gerekir.

Spam SMS veri seti için kNN sınıflandırıcı kullanıldığında elde edilen Makro-F1 değerleri Tablo 7.12'de verilmiştir. SVM sınıflandırıcısının, en yüksek puanlara göre Spam SMS veri setinde RF, kNN ve DT sınıflandırıcıları geride bıraktığı belirtilebilir.

Tablo 7.12. *Spam SMS veri seti için Makro-F1 skorları (d) kNN*

Öznitelik Seçme Metotları	Öznitelik Sayıları				
	50	**100**	**300**	**500**	**1000**
GSS	87,351	91,189	92,679	92,113	92,353
CICI	91,288	91,747	92,679	92,113	92,353
MMR	91,384	90,550	89,495	90,347	90,631
CMFS	88,761	91,595	92,741	91,189	91,134
DFSS	90,904	89,943	89,403	89,211	89,628
EFS	91,799	91,695	90,548	90,828	90,510
EFS_IMP1	92,589	92,762	92,346	89,927	91,576
EFS_IMP2	**92,922**	92,105	91,152	90,550	92,095

Reuters-21578 (R8 MM) veri setinde SVM sınıflandırıcısı kullanıldığında hesaplanan sınıflandırma performansları Şekil 7.10'da verilmiştir.



Şekil 7.10. *Reuters-21578 (R8 MM) veri seti için SVM sınıflandırıcısı kullanıldığında Makro-F1 skorları*

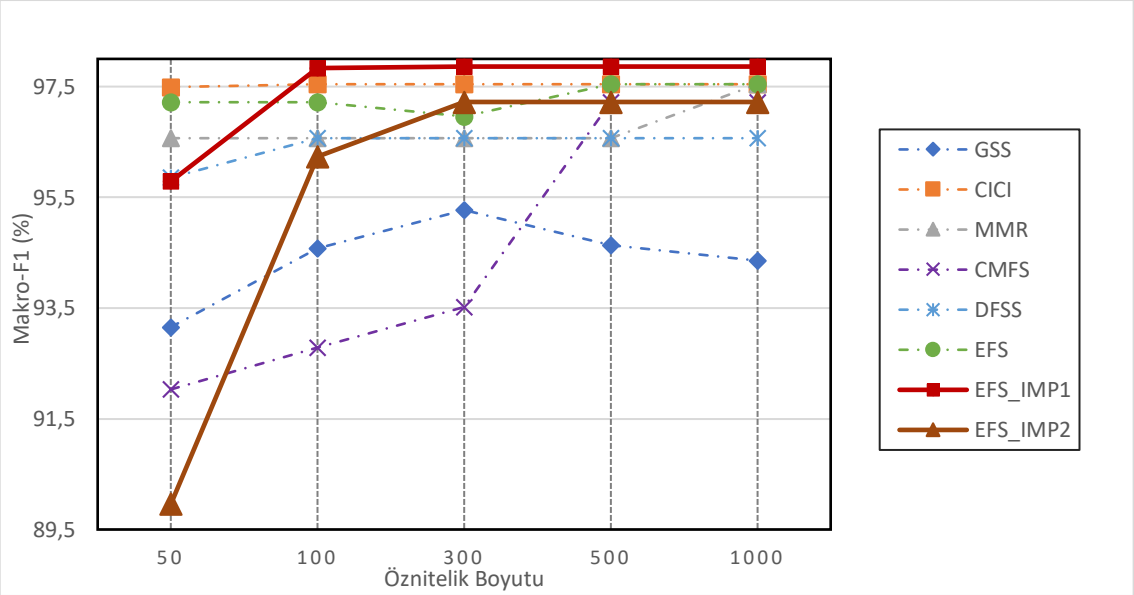
SVM sınıflandırıcısı kullanılarak Reuters-21578 (R8 MM) veri setinde, 50 öznitelik boyutunda CICI, 99,695 ile en yüksek Makro-F1 puanını alır. CMFS ve EFS_IMP1'in de çeşitli öznitelik boyutlarında başarılı olduğunu belirtmek gerekir.

Reuters-21578 (R8 MM) veri seti için SVM sınıflandırıcı kullanıldığında elde edilen Makro-F1 değerleri Tablo 7.13'te verilmiştir.

Tablo 7.13. *Reuters-21578 (R8 MM) veri seti için Makro-F1 skorları (a) SVM*

Öznitelik Seçme Metotları	Öznitelik Sayıları				
	30	**50**	**100**	**300**	**500**
GSS	90,563	91,858	91,858	91,180	90,676
CICI	**99,695**	95,463	96,864	96,154	96,795
MMR	97,464	96,830	92,623	95,412	95,763
CMFS	95,513	97,491	98,457	99,074	98,118
DFSS	96,830	96,154	94,397	95,056	96,111
EFS	98,759	96,795	92,151	94,033	95,112
EFS_IMP1	97,793	99,074	97,131	95,463	96,795
EFS_IMP2	96,830	96,455	98,440	97,839	96,455

Reuters-21578 (R8 MM) veri setinde DT sınıflandırıcısı kullanıldığında elde edilen sınıflandırma performansları Şekil 7.11'de verilmiştir.



Şekil 7.11. *Reuters-21578 (R8 MM) veri seti için DT sınıflandırıcısı kullanıldığında Makro-F1 skorları*

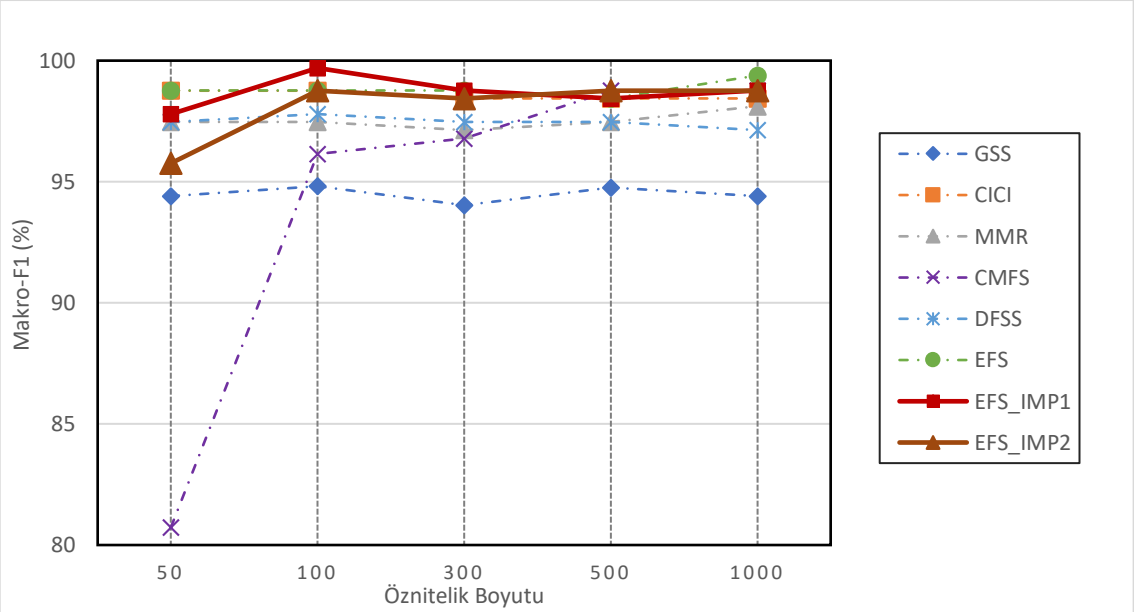
DT sınıflandırıcısı kullanılarak Reuters-21578 (R8 MM) veri setinde, 300, 500 ve 1000 öznitelik boyutunda EFS_IMP1, 97,862 ile en yüksek Makro-F1 puanını alır.

Reuters-21578 (R8 MM) veri seti için DT sınıflandırıcı kullanıldığında elde edilen Makro-F1 değerleri Tablo 7.14'te verilmiştir. EFS_IMP1 metodunun Makro-F1 değerleri birçok öznitelik boyutunda diğer öznitelik seçme metodundan daha başarılıdır.

Tablo 7.14. *Reuters-21578 (R8 MM) veri seti için Makro-F1 skorları (b) DT*

Öznitelik Seçme Metotları	Öznitelik Sayıları				
	30	**50**	**100**	**300**	**500**
GSS	93,156	94,581	95,270	94,640	94,355
CICI	97,491	97,543	97,543	97,543	97,543
MMR	96,568	96,568	96,568	96,568	97,543
CMFS	92,031	92,790	93,518	97,222	97,222
DFSS	95,856	96,568	96,568	96,568	96,568
EFS	97,222	97,222	96,961	97,543	97,543
EFS_IMP1	95,789	97,839	**97,862**	**97,862**	**97,862**
EFS_IMP2	89,964	96,236	97,222	97,222	97,222

Reuters-21578 (R8 MM) veri setinde RF sınıflandırıcısı kullanıldığında elde edilen sınıflandırma performansları Şekil 7.12'de gösterilmektedir.



Şekil 7.12. *Reuters-21578 (R8 MM) veri seti için RF sınıflandırıcısı kullanıldığında Makro-F1 skorları*

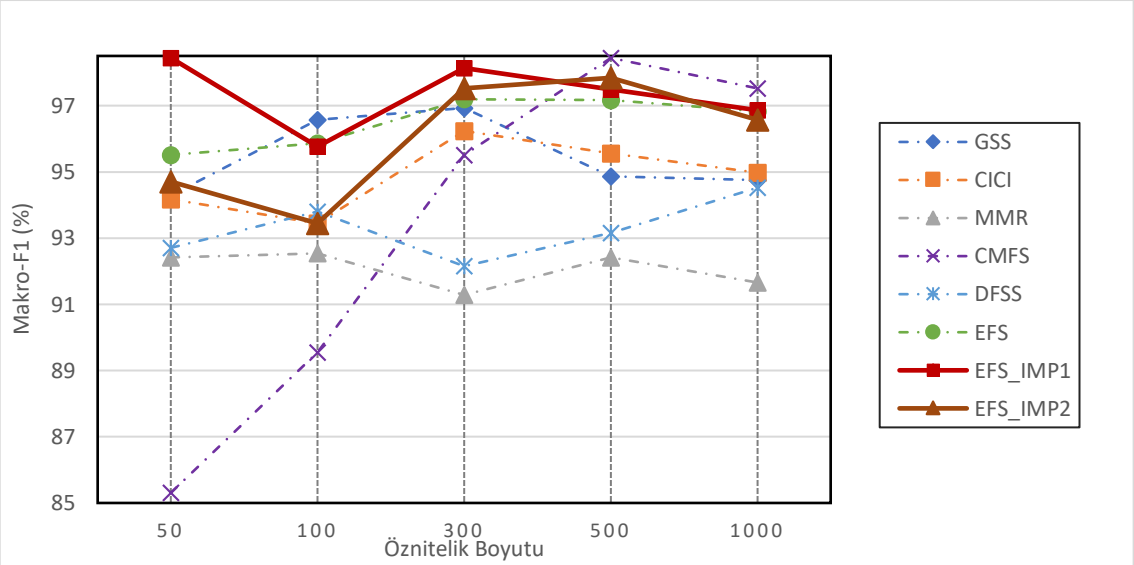
RF sınıflandırıcısı kullanılarak Reuters-21578 (R8 MM) veri setinde, 100 öznitelik boyutunda EFS_IMP1, 99,695 ile en yüksek Makro-F1 puanını alır. CICI, CMFS, EFS ve EFS_IMP2'nin de çeşitli öznitelik boyutlarında başarılı olduğunu belirtmek gerekir.

Reuters-21578 (R8 MM) veri seti için RF sınıflandırıcı kullanıldığında elde edilen Makro-F1 değerleri Tablo 7.15'te verilmiştir.

Tablo 7.15. *Reuters-21578 (R8 MM) veri seti için Makro-F1 skorları (c) RF*

Öznitelik Seçme Metotları	Öznitelik Sayıları				
	30	**50**	**100**	**300**	**500**
GSS	94,397	94,815	94,033	94,756	94,397
CICI	98,759	98,759	98,440	98,440	98,440
MMR	97,464	97,464	97,131	97,464	98,118
CMFS	80,737	96,154	96,795	98,759	98,759
DFSS	97,464	97,793	97,464	97,464	97,131
EFS	98,759	98,759	98,759	98,440	99,386
EFS_IMP1	97,793	**99,695**	98,759	98,440	98,759
EFS_IMP2	95,763	98,759	98,440	98,759	98,759

Reuters-21578 (R8 MM) veri setinde kNN sınıflandırıcısı kullanıldığında elde edilen sınıflandırma performansları Şekil 7.13'te sunulmuştur.



Şekil 7.13. *Reuters-21578 (R8 MM) veri seti için kNN sınıflandırıcısı kullanıldığında Makro-F1 skorları*

kNN sınıflandırıcısı kullanılarak Reuters-21578 (R8 MM) veri setinde, 50 öznitelik boyutunda EFS_IMP1 ve 500 öznitelik boyutunda CMFS, 98,440 ile en yüksek Makro-F1 puanını alır. EFS _IMP1 ayrıca 300 öznitelik boyutunda da başarılıdır.

Reuters-21578 (R8 MM) veri seti için kNN sınıflandırıcı kullanıldığında elde edilen Makro-F1 değerleri Tablo 7.16'da verilmiştir. SVM ve RF sınıflandırıcılarının performansının neredeyse eşit olduğu ve en yüksek puanlara göre Reuters-21578 (R8 MM) veri setinde kNN ve DT sınıflandırıcılarını geride bıraktıkları belirtilebilir.

Tablo 7.16. *Reuters-21578 (R8 MM) veri seti için Makro-F1 skorları (d) kNN*

Öznitelik Seçme Metotları	Öznitelik Sayıları				
	30	**50**	**100**	**300**	**500**
GSS	94,293	96,568	96,929	94,860	94,752
CICI	94,167	93,445	96,236	95,562	94,982
MMR	92,419	92,537	91,282	92,419	91,666
CMFS	85,311	89,545	95,513	**98,440**	97,517
DFSS	92,708	93,808	92,151	93,156	94,521
EFS	95,513	95,856	97,192	97,162	96,830
EFS_IMP1	**98,440**	95,763	98,138	97,491	96,864
EFS_IMP2	94,696	93,445	97,517	97,839	96,568

Eşleşen sınıflandırıcı ile veri setlerinde en yüksek puanları alan öznitelik seçme metotları Tablo 7.17'de gösterilmiştir. Ayrıca, Tablo 7.17'den görüleceği üzere, en yüksek Makro-F1 puanlarına göre, EFS_IMP1 ve EFS_IMP2 çoğu durumda en iyi değerleri almıştır.

Tablo 7.17. *En yüksek Makro-F1 değerlerine göre en başarılı öznitelik seçme metotları*

	Makro-F1			
Veri seti	SVM	DT	RF	kNN
Enron1	EFS_IMP1	EFS_IMP1	EFS_IMP2	GSS
Spam SMS	EFS_IMP1	EFS_IMP2	EFS_IMP1	EFS_IMP2
Reuters-21578 (R8 MM)	CICI	EFS_IMP1	EFS_IMP1	EFS_IMP1, CMFS

7.4. Değerlendirmeler

Literatürdeki veri setlerinin çoğu genellikle dengesizdir (Pouramini, Minaei-Bidgoli, & Esmaeili, 2018). Bu durum, metin sınıflandırma performansını düşürmektedir. Dengesiz veri problemi için bazı çözümler önerilmiştir, ancak öznitelik seçimi alanında hâlâ boşluklar bulunmaktadır. Bu çalışmanın ana katkısı, dengesiz metinlerin sınıflandırılması için özellikle ikili veri setlerinde etkili olan iki yeni filtre bazlı öznitelik seçme yöntemi önermektir.

EFS _IMP1 ve EFS_IMP2'nin performansları, çeşitli sınıflandırıcılar ve iki sınıflı dengesiz veri setleri kullanılarak altı filtre bazlı öznitelik seçme yöntemi ile karşılaştırılmıştır. Sonuç olarak, EFS_IMP1 ve EFS_IMP2, çoğu durumda iki sınıflı dengesiz veri setlerinde en yüksek Makro-F1 puanlarına göre rekabetçi yöntemlere göre üstündür. Sunulan yöntemler, gelecekteki çalışmalarda farklı alanlarda çeşitli veri setlerine uygulanabilir.

8. SONUÇ VE TARTIŞMA

Bu tez kapsamında, dengesiz metin sınıflandırma problemlerine çözüm üretebilmek için çeşitli çalışmalar yapılmıştır. Öncelikli olarak, dengesiz metin sınıflandırma için öznitelik seçme alanına yönelik olarak, literatürde mevcut olan popüler öznitelik seçme metotları incelenmiştir. İnceleme sırasında, öznitelik seçerken izledikleri stratejiler, güçlü oldukları ve yetersiz kaldığı durumlar tespit edilmeye çalışılmıştır. Mevcut öznitelik seçme metotlarının davranışları ve sınıflandırma performansları analiz edilerek; öznitelikleri, söz konusu metotlardan daha verimli bir biçimde seçebilecek yeni öznitelik seçme metotları geliştirmenin yolları araştırılmıştır.

Tezin ilk aşamasında, öznitelik seçme metotlarının dengesiz metinlerin sınıflandırılması üzerindeki etkisi etraflıca araştırılmıştır. Bu doğrultuda, iki farklı veri seti üzerinde üç farklı sınıflandırıcı ve dokuz farklı öznitelik seçme metodu ile birçok deney yapılmıştır. Ayrıca öznitelik seçme metotlarının başarıları farklı öznitelik sayılarında da gözlemlenmiştir. NDM, DFSS, PFS, POISSON, CHI2, IG, GINI, DFS ve MDFS olarak adlandırılan dokuz farklı öznitelik seçme metodu değerlendirilmiştir. SVM, DT ve MNB sınıflandırıcıları ile deneysel sonuçlar elde edilmiştir. Öznitelik seçme metotlarından DFS ve CHI2'nin dengesiz metin sınıflandırmada daha başarılı olduğu gözlemlenmiştir.

Literatürdeki veri setlerinin çoğu genellikle dengesizdir. Birçok sınıflandırıcı, ana sınıfı (majör) tercih ederken küçük olanı(minör) görmezden gelmektedir. Bu durum, metin sınıflandırma performansını düşürmektedir. Dengesiz veri sorunu için bazı çözümler önerilmiş olsa da, öznitelik seçimi alanında hâlâ boşluklar mevcuttur. Bu çalışmanın ana katkısı, dengesiz metinlerin sınıflandırılması için özellikle ikili veri setlerinde etkili olan EFS_IMP1 ve EFS_IMP2 adında iki yeni filtre bazlı öznitelik seçme yöntemini önermektir.

EFS, terimlerin ayırt edici kapasitesini belirleyebilir, ancak küçük sınıfın ilgili terimleri EFS tarafından yeterince ayırt edilemez. Yani EFS, küçük sınıftaki her ilgili terime tamamen aynı ağırlığı vermektedir. EFS_IMP1 ve EFS_IMP2, terimlerin küçük ve büyük sınıflara olan ilgililik derecesini hesaplarken küçük ve büyük sınıf dağılımını göz önünde bulundurmaktadır. Önerilen yöntemler metin sınıflandırmada kullanıldığında, küçük ve büyük sınıf dokümanları arasındaki ayrım artmaktadır ve sınıflandırıcıların da performansı artmaktadır. Sonuçlar, öznitelik seçme yöntemlerinin

terimlerin ayırt etme gücünden ziyade ilgililik gücünü incelemesi gerektiğini göstermektedir.

Tezin ikinci aşamasında önerilen EFS _IMP1 ve EFS_IMP2'nin performansları, çeşitli sınıflandırıcılar ve iki sınıflı dengesiz veri setlerini kullanarak altı filtre bazlı öznitelik seçme yöntemiyle karşılaştırılmıştır. Sonuç olarak, EFS_IMP1 ve EFS_IMP2, çoğu durumda iki sınıflı dengesiz veri setlerinde en yüksek Makro-F1 değerlerine göre rekabetçi yöntemlerden üstündür. Sunulan yöntemler, gelecek çalışmalarda farklı alanlardan çeşitli veri setlerine uygulanabilir.

KAYNAKÇA

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

Almeida, T. A., Hidalgo, J. M., & Yamakami, A. ((2011, September)). Contributions to the study of SMS spam filtering: new collection and results. . *In Proceedings of the 11th ACM symposium on Document engineering*, (pp. 259-262).

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

Barua, S., Islam, M. M., Yao, X., & Murase, K. (2012). MWMOTE--majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on knowledge and data engineering*, 26(2), 405-425.

Bhowmick, A., & Hazarika, S. M. (2018). E-mail spam filtering: a review of techniques and trends. *Advances in Electronics, Communication and Computing: ETAEERE-2016*, 583-590.

Cardoso Cachopo, A. (2014). Datasets for single-label text categorization .

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

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.

Chen, E., Lin, Y., H., X., Luo, Q., & Maa, H. (2011). Exploiting probabilistic topic models to improve text categorization under class imbalance. *Information Processing & Management*, 47(2): p. 202-214.

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

Chen, X.-w., & Wasikowski, M. (2008). FAST: a roc-based feature selection metric for small samples and imbalanced data classification problems. *, in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, (s. 124-132). Las Vegas, Nevada, U.S.A.

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

Dogan, T., & Uysal, A. K. (2019). Improved inverse gravity moment term weighting for text classification. *Expert Systems with Applications*, 130, 45-59.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3 , 1289–1305.

Galar, M., Fern´andez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):p.463-484.

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

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

Iglesias, E., Vieira, A. S., & Borrajo, L. (2013). An HMM-based over-sampling technique to improve text classification. *Expert Systems with Applications* , 40(18): p. 7184-7192.

Jasti, V. D., Kumar, G. K., Kumar, M. S., Maheshwari, V., Jayagopal, P., Pant, B., & Muhibbullah, M. (2022). Relevant-based feature ranking (RBFR) method for text classification based on machine learning algorithm. *Journal of Nanomaterials*, 1-12.

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

Kübler, S., L. C., & Sayyed, Z. A. (2018). To use or not to use: Feature selection for sentiment analysis of highly imbalanced data. *Natural Language Engineering*, 24(1), 3-37.

Liu, Y., Loh, H., & Sun, A. (2009). Imbalanced text classification: A term weighting approach. *Expert Systems with Applications*, 36(1): p. 690-701.

Manevitz, L. M., & Yousef, M. (2002). One-class SVMS for document classification . *The Journal of Machine Learning Research* , 2, 139–154.

Manning, C., Raghavan, P., & Schutze, H. (2008). *Introduction to information retrieval.* NewYork, USA: Cambridge University Press.

Metsis, V., Androutsopoulos, I., & Paliouras, G. ((2006, July)). Spam filtering with naive bayes-which naive bayes? *In CEAS*, (Vol. 17, pp. 28-69).

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

Moh'd Mesleh, A. (2011). Feature sub-set selection metrics for Arabic text classification. *Pattern Recognition Letters*, 32(14), 1922-1929.

Naderalvojoud, B., & Sezer, E. A. (2020). Term evaluation metrics in imbalanced text categorization. *Natural Language Engineering*, 26(1), 31-47.

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

Okkalioglu, M., & Okkalioglu, B. D. (2022). AFE-MERT: imbalanced text classification with abstract feature extraction . *Applied Intelligence*, 1-17.

Parlak, B. (2022). Class-index corpus-index measure: A novel feature selection method for imbalanced text data. *Concurrency and Computation: Practice and Experience*, 34(21), e7140.

Parlak, B., & Uysal, A. K. (2023). A novel filter feature selection method for text classification: Extensive Feature Selector. *Journal of Information Science*, 49(1), 59-78.

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

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

Priyadarsini, R. P., Valarmathi, M. L., & Sivakumari, S. (2011). Gain ratio based feature selection method for privacy preservation. *ICTACT J Soft Comput* , 1(04), 2229-6956.

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

Rehman, A., Javed, K., & Babri, H. (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.

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

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.

Tiryaki, H., & Uysal, A. K. (2023). Dengesiz Metin Sınıflandırmada Öznitelik Seçim Yöntemlerinin Etkililiği. *Afyon Kocatepe Üniversitesi Fen Ve Mühendislik Bilimleri Dergisi*, 23(2), 370-379.

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

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

Uysal, D., & Uysal, A. K. (2022). AUTOMATIC CLASSIFICATION OF EFL LEARNERS'SELF-REPORTED TEXT DOCUMENTS ALONG AN AFFECTIVE CONTINUUM . *Advanced Education*, 4-14.

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

Wua, Q., Ye, Y., Zhang, H., Ng, M., & Ho, S. (2014). FORESTEXTER: An efficient random forest algorithm for imbalanced text categorization. *Knowledge-Based Systems*, 67 , 105–116.

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.

Yang, P., Liu, W., Zhou, B. B., Chawla, S., & Zomaya, A. Y. (2013). Ensemble-based wrapper methods for feature. *springer,Advances in Knowledge Discovery and Data Mining* , vol. 7818, pp. 544-555.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. *Paper presented at the Icml*.

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

ORCID NO: 0000-0002-1533-6901

Ad Soyad **:** Hande TİRYAKİ

Yabancı Dil **:** İngilizce

Doğum Yeri ve Yılı **:**

E-Posta **:**

Eğitim Geçmişi:

- Doktora (2018-2023) : Eskişehir Teknik Üniversitesi, Lisansüstü Eğitim Enstitüsü, Bilgisayar Mühendisliği A.B.D.
- Doktora (2013-2018) : Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği A.B.D.
- Yüksek Lisans (2010-2013) : Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği A.B.D.
- Lisans (2002-2007) : Başkent Üniversitesi, Mühendislik Fakültesi/ Bilgisayar Mühendisliği

Meslek Geçmişi:

- Yazılım Mühendisi (2020-) : Anadolu Üniversitesi, B.A.U.M.
- Öğretim Görevlisi (2018-2020) : Eskişehir Teknik Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü
- Öğretim Görevlisi (2015-2018) : Anadolu Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü
- Yazılım Mühendisi (2011-2015) : Anadolu Üniversitesi, B.A.U.M.
- Yazılım Mühendisi (2009-2011) : Orta Doğu Teknik Üniversitesi, Bilgi İşlem Daire Başkanlığı

- Yazılım Mühendisi (2007-2009) : Bilkent Üniversitesi,
  Cyberplaza/ Tepe International A.Ş.

Tez Kapsamındaki Yayınlar:

Tiryaki, H., & Uysal, A. K. (2023). Dengesiz Metin Sınıflandırmada Öznitelik Seçim Yöntemlerinin Etkililiği.*Afyon Kocatepe Üniversitesi Fen Ve Mühendislik Bilimleri Dergisi*, *23*(2), 370-379.

Tiryaki, H., & Uysal, A. K. (2023). Two Novel Feature Selection Methods for Imbalanced Text Classification. *Expert Systems with Applications.*

(İncelemede)

Diğer yayınlanan makaleler:

Tiryaki, H., & Doğan, M. (2015). Solo Test Oyunu Üzerinde Paralel Önce-Derine Arama Algoritmasının İşlemci Performans Değerlendirmesi. AB2015, 17. Akademik Bilişim Konferansı, Eskişehir