KISA METİN SINIFLANDIRMA İÇİN ÖZNİTELİK SEÇİMİ

Rasım ÇEKİK

Doktora Tezi

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Donanımı Bilim Dalı

Danışman: Doç. Dr. Alper Kürşat UYSAL

Eskişehir

Eskişehir Teknik Üniversitesi

Lisansüstü Eğitim Enstitüsü

Ağustos 2020

Bu tez çalışması BAP Komisyonu tarafından kabul edilen 20DRP040 no.lu proje kapsamında desteklenmiştir.

ÖZET

KISA METİN SINIFLANDIRMA İÇİN ÖZNİTELİK SEÇİMİ

Rasım ÇEKİK

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Donanımı Bilim Dalı

Eskişehir Teknik Üniversitesi, Lisansüstü Eğitim Enstitüsü, Ağustos 2020

Danışman: Doç. Dr. Alper Kürşat UYSAL

Kısa metin sınıflandırmada yüksek boyutluluk problemi sınıflandırıcıların işlem maliyetini ve performansını etkilediği için önemli bir yer tutmaktadır. Ayrıca kısa metinlerin seyrek, eksik ve tutarsız yapıda olmaları yüksek boyutluluk yanında uğraşılması gereken başka bir konudur. Tüm öznitelik uzayını en iyi temsil edecek alt öznitelik uzayını seçmek, yüksek boyutluluk problemine sunulan en etkili çözüm yollarından biridir. Bu yüzden bu alanda seyreklik probleminden en az etkilenecek ve etkili öznitelik seçecek yaklaşımlara ihtiyaç vardır. Bu amaç doğrultusunda kısa metin alanında etkili çalışacak iki öznitelik seçme yaklaşımı bu çalışmada sunulmuştur. Bu yaklaşımlardan ilki Orantılı Öznitelik Seçme (Proportional Rough Feature Selector-PRFS) adı verilen kaba kümeler tabanlı yaklaşımdır. PRFS yaklaşımı kaba kümeler yardımı ile terimlerin/özniteliklerin değer kümesine göre dokümanları belirli bölgelere ayırır. Bu bölgesel ayırma ile bir dokümanın bir sınıfa kesin ait olması veya ait olma olasılığında olduğu belirlenebilir. Ayrıca bir sınıfa ait olma olasılığında olma durumundaki dokümanlara bir ceza uygulamak için α adında bir katsayı ve terim vektör uzayındaki seyrekliğin etkisi hesaplanmıştır. Daha sonra PRFS metodu en iyi ve en çok bilinen öznitelik seçim yaklaşımlarından Gini katsayısı, bilgi kazanımı, ayırt edici öznitelik seçici ve son zamanlarda önerilmiş yöntemlerden max-min oranı ve normalleştirilmiş fark ölçüsü ile dört kısa metin veri kümesi üzerinde farklı öznitelik boyutlarında Makro-F1 sonuçlarının kıyaslanması yapılarak test edilmiştir. Deneysel sonuçlar, PRFS'nin Makro-F1 açısından diğer öznitelik seçim yöntemlerine göre daha iyi veya rekabetçi performans sunduğunu göstermiştir. Bu çalışma, kaba küme teorisi kullanılarak kısa metin sınıflandırması için yeni bir filtre öznitelik seçme yöntemi önerdiğinden, bu araştırma alanında öncü bir çalışma olabilir.

İkinci yaklaşım XY metot olarak tanıtılan yaklaşımdır. Bu yaklaşım, tüm ikili sınıf kombinasyonları için terimleri dokuman frekansına göre ikili koordinat düzleminde düşünür. Daha sonra terimin XY doğrusuna olan uzaklığı hesaplanır. Ayrıca λ gibi bir değer hesaplanır ve bu değere göre terimler pozitif, negatif ve üçüncü bölge diye farklı bölgelere ayrılır. XY metodunun amacı olabildiğince negatif bölgeden az terim seçmektir. Bu metot da çok iyi bilinen ki-kare, bilgi kazanımı, Poisson dağılımından sapma yaklaşımları ve son zamanlarda önerilmiş max-min oranı ve ayırt edici öznitelik seçici yaklaşımları ile dört farklı kısa metin verisi üzerinde farklı öznitelik boyutları için Makro-F1 sonuçları test edilmiştir. Deneysel sonuçlar, XY metodunun Makro-F1 açısından diğer öznitelik seçim yöntemlerine göre daha iyi veya rekabetçi performans sunduğunu göstermiştir.

Anahtar Sözcükler: Öznitelik Seçme, Kısa Metin Sınıflandırma, Metin Madenciliği, Kaba Kümeler.

ABSTRACT

FEATURE SELECTİON FOR SHORT TEXT CLASSIFICATION

Rasım ÇEKİK

Department of Computer Engineering

Programming Computer Hardware

Eskişehir Technical University, Institute of Graduate Programs, August 2020

Supervisor: Assoc. Prof. Dr. Alper Kürşat UYSAL

High dimensionality problem is an important concern for short text classification due to its effect on computational cost and accuracy of classifiers. Also, short text data, besides being high dimensional, has an incomplete, inconsistent and sparse structure. Selection of important features that provides a better representation is a solution for high dimensionality problem. However, it is a fact that in feature selection process, short texts need feature selection approaches that will be least affected by the sparse problem. In this study, two feature selection approaches that will work effectively in the short text field are presented for this purpose. Firstly, we developed a novel filter feature selection method called Proportional Rough Feature Selector (PRFS) which uses the rough set for a regional distinction according to the value set of term to identify documents that to be exact belong to a class and have a possibility for belonging to a class. Documents which are possible to belong to a class are penalized by multiplying with a coefficient named α. Additionally, the effect of sparsity in the term vector space is calculated with the help of rough set. The PRFS is compared with state-of-the-art filter feature selection methods such as Gini index, information gain, distinguishing feature selector, recently proposed max-min ratio and normalized difference measure methods. The comparison is carried out using various feature sizes on four different short text datasets with Macro-F1 success measure. Experimental results demonstrated that the PRFS offers either better or competitive performance with respect to other feature selection methods in term of Macro-F1. This study may be a pioneering study in this research field as it proposes a novel feature selection method for short text classification using rough set theory.

Secondly, this study presents a new filter feature selection method called XY method which represents the features on XY *line* and calculates the distance of a feature to the *XY line*. Also, a value like λ is calculated. According to this value, the terms are

divided into different regions such as negative, positive, and third. The XY method aims to select as few terms as possible in the negative region. The XY method is compared with well-known filter feature selection methods such as chi-square, information gain, deviation from Poisson distribution, recently proposed max-min ratio, and distinguishing feature selector methods. The comparison is carried out using various feature sizes in order to make a fair evaluation on four different short text datasets with Macro-F1 success measure. Experimental results demonstrate that the XY method offers either better or competitive performance with respect to other feature selection methods in term of Macro-F1.

Keywords: Short Text Classification, Rough Set, Feature Selection, Text Mining.

TEŞEKKÜR

Tez çalışmam boyunca benimle deneyimlerini ve bilgilerini koşulsuz paylaşan, desteğini en üst düzeyde hissettiğim gerek insanı tutumu olsun gerek yol göstericiliği ile hayatımda yeni pencereler açmama vesilen olan danışmanım Sayın Doç. Dr. Alper Kürşat UYSAL hocama teşekkürü bir borç bilir ve katkılarından ötürü en derin teşekkürlerimi sunarım.

Çalışmamı sağlıklı bir şekilde yürütmeme dayanak olan ve tez izleme süreçleri boyunca sabırla beni dinlediklerinden ve yol gösterdiklerinden tez izleme jürisinde yer alan hocalarım Sayın Prof. Dr. Serkan GÜNAL'a ve Sayın Dr. Öğr. Üyesi Şener AĞALAR'a verdikleri tüm fikir ve önerilerinden dolayı kendilerine çok teşekkür ederim.

Hayatımın her zorluğunda yanımda olan Sevgili Eşim Kevser ÇEKİK'e bu zorlu yolda yine tüm desteği ile bana güç verdiği için kendisine minnettarım.

Hayatımdaki tüm negatif enerjiyi varlıkları ile yok eden, hayatıma en güzel hisleri katan ve büyürken bana inanılmaz bilgi ve olgunluk kazandıran Sevgili Yavrularım Ahmet ve Aslı Miray'a varlıklarından ötürü Rabbime şükrümü ve onlara sevgilerimi sunuyorum.

Her daim güzel yüreği ile bana örnek olan ve destek veren Sevgili Ağabeyim Cemalettin ÇEKİK'e, duaları ve sevgisiyle beni her zaman kucaklayan Sevgili Anneme minnettarım.

Hayatımın her anında bana hep destek olan ve her şeyimi borçlu olduğum Sevgili Babamı saygıyla yad ediyorum.

Rasım ÇEKİK

06/08/2020

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.

Rasım ÇEKİK

İÇİNDEKİLER

Sayfa

BAŞLIK SAYFASI ..... **İ**

JÜRİ VE ENSTİTÜ ONAYI ..... **İİ**

ÖZET ..... **İİİ**

ABSTRACT ..... **V**

TEŞEKKÜR ..... **Vİİ**

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ ..... **Vİİİ**

İÇİNDEKİLER ..... **İX**

TABLOLAR DİZİNİ ..... **Xİİ**

ŞEKİLLER DİZİNİ ..... **Xİİİ**

SİMGELER VE KISALTMALAR DİZİNİ ..... **XV**

1. GİRİŞ **1**

1.1. Tezin Amacı ve Hedefleri **3**

1.2. Tez Organizasyonu **4**

2. METİN SINIFLANDIRMA AŞAMALARI VE ÖNBİLGİ **6**

2.1. Metin Sınıflandırma ve Süreçleri **6**

2.1.1. Önişlem (Pre-Processing) **7**

2.1.1.1. Tokenizasyon (Tokenization) ***8***

2.1.1.2. Durma sözcüklerin kaldırılması (Stop-word removal) ***8***

2.1.1.3. Küçük harf dönüşümü (Lowercase conversion) ***8***

2.1.1.4. Kök bulma (Stemming) ***8***

2.1.2. Öznitelik çıkarma (Feature extraction) **9**

2.1.2.1. Kelime çantası (The bag of words) ***9***

2.1.2.2. Word2Vec ***9***

2.1.3. Öznitelik Ağırlıklandırma (Feature Weighting) **9**

2.1.3.1. İkili ağırlıklandırma (Binary weighting) ***10***

2.1.3.2. Terim frekans-Ters doküman frekansı (TF-IDF) ***10***

2.1.4. Öznitelik seçme (Feature selection) **11**

2.1.4.1. Geliştirilmeye dayalı yaklaşımlar 12
2.1.4.2. Performans karşılaştırmaya ve seçim sürecini iyileştirmeye dayalı çalışmalar 14
2.1.5. Sınıflandırma 14
2.2. Öznitelik Seçme Teknikleri 15
2.2.1. Bilgi kazanımı (Information gain-IG) 15
2.2.2. Ki-kare (Chi-square-Chi2) 16
2.2.3. Gini katsayısı (Gini index-GI) 16
2.2.4. Poisson dağılımından sapma (Deviation from poisson distribution-DP) 17
2.2.5. Ayırt edici öznitelik seçici (Distinguishing feature selector-DFS) 18
2.2.6. Normalleştirilmiş fark ölçüsü (Normalized Difference Measure-NDM) 18
2.2.7. Max-min oranı (Max-min ratio-MMR) 18
2.3. Sınıflandırma Metotları 19
2.3.1. Destek vektör makineleri (Support vector machines -SVM) 19
2.3.2. K-en yakın komşular (K-nearest neighbors-KNN) 20
2.3.3. Karar ağacı (Decision tree-DT) 22
2.3.4. Naive bayes (NB) 22
2.4. Kaba Kümeler Teorisi 23
3. İLGİLİ ÇALIŞMALAR VE GENEL MOTİVASYON 26
4. KISA METİN SINIFLANDIRMA 31
4.1. Kısa Metinlerin Yapısal ve Yapısal Olmayan Özellikleri 32
4.1.1. Az kelime içerme 33
4.1.2. Anlık olma 33
4.1.3. Standartsız olma 33
4.1.4. Dengesiz dağılımlı olma 33
4.1.5. Az etiketli çok sayıda olma 34
4.2. Kısa Metin Problemleri 34
4.2.1. Seyreklik problemi 34
4.2.2. Yüksek boyutluluk problemi 34

5. ÖNERİLEN METOT: ORANTILI KABA ÖZNİTELİK SEÇİCİ (PROPORTIONAL ROUGH FEATURE SELECTOR- PRFS) 36

5.1. Motivasyon 36

5.2. PRFS'nin Çalışma Stratejisi 36

5.3. Deneysel Çalışmalar 42

5.3.1. Veri kümeleri 42

5.3.2. Başarı kriteri: Makro-F1 43

5.3.3. Başarı analizi 44

5.4. Değerlendirme 55

6. ÖNERİLEN METOT: XY METOT 56

6.1. Motivasyon 56

6.2. XY Metodunun Çalışma Stratejisi 57

6.3. Deneysel Çalışmalar 62

6.3.1. *Lamda* (λ) değerinin analizi 62

6.3.2. Başarı analizi 64

6.3.3. Zaman karmaşıklık analizi 73

6.4. Değerlendirme 74

7. SONUÇ VE TARTIŞMA 75

KAYNAKÇA 77

ÖZGEÇMİŞ

TABLOLAR DİZİNİ

Sayfa

Tablo 2.1. Bir karar tablosu örneği 23
Tablo 4.1. Kısa metin verilerinin özellikleri ve ilgili problemle ilişkisi 32
Tablo 5.1. Basit bir doküman koleksiyon örneği 41
Tablo 5.2. Deneylerde kullanılan veri kümelerinin karakteristiği 43
Tablo 5.3. İlgili veri kümelerinde SVM, KNN, DT ve NB sınıflandırıcıları kullanıldığında tüm öznitelik boyutları için ortalama Makro-F1 skorları 51
Tablo 5.4. Farklı deneysel setler için en yüksek Makro-F1 skorunu (tepe değeri-pik noktası) sağlayan öznitelik seçme yöntemleri 52
Tablo 5.5. İlgili öznitelik boyutlarında tüm veri kümeleri için SVM, KNN, DT ve NB sınıflandırıcıları kullanıldığında ortalama Makro-F1 skorları 53
Tablo 6.1. Boyut 500 öznitelik için XY metot ve diğer yaklaşımların negatif bölgede terim seçme oranları 64
Tablo 6.2. Her öznitelik boyutu için **SMS** veri kümesinde en yüksek Makro-F1 skorlarını veren öznitelik seçme yöntemleri 70
Tablo 6.3. Her öznitelik boyutu için **SLS** veri kümesinde en yüksek Makro-F1 skorlarını veren öznitelik seçme yöntemleri 71
Tablo 6.4. Her öznitelik boyutu için **AYSC** veri kümesinde en yüksek Makro-F1 skorlarını veren öznitelik seçme yöntemleri 71
Tablo 6.5. Her öznitelik boyutu için **Sentiments** veri kümesinde en yüksek Makro-F1 skorlarını veren öznitelik seçme yöntemleri 71
Tablo 6.6. Farklı deneysel setler için en yüksek Makro-F1 skorunu (tepe değeri-pik noktası) sağlayan öznitelik seçme yöntemleri 72
Tablo 6.7. İlgili öznitelik boyutlarında tüm veri kümeleri için SVM, KNN, DT ve NB sınıflandırıcıları kullanıldığında ortalama Makro-F1 skorları 73

ŞEKİLLER DİZİNİ

Sayfa

Şekil 2.1. Metin sınıflandırma genel şeması 7
Şekil 2.2. Öznitelik seçme metotları ile ilgili çalışmaların gruplandırılması 12
Şekil 2.3. Sınıflandırma sisteminin akış şeması 14
Şekil 2.4. Filtre yaklaşımlarının akış süreci 15
Şekil 2.5. Maks-marj hiper düzlem ve marjlar gösterimi 19
Şekil 2.6. Örnek bir KNN çizimi 21
Şekil 2.7. KNN yaklaşımının genel çalışma akışı 21
Şekil 2.8. Kaba kümelerde bölgeler ve küme yaklaşımları gösterimi 24
Şekil 5.1. Bölgeler ve MS ve NMS bölümleri gösterimi 38
Şekil 5.2. PRFS metodun mimarisi 41
Şekil 5.3. SMS veri kümesi için SVM (a), KNN (b), DT (c) ve NB (d) sınıflandırıcısı kullanıldığında Makro-F1 skorları 46
Şekil 5.4. SLS veri kümesi için SVM (a), KNN (b), DT (c) ve NB (d) sınıflandırıcısı kullanıldığında Makro-F1 skorları 47
Şekil 5.5. AYSC veri kümesi için SVM (a), KNN (b), DT (c) ve NB (d) sınıflandırıcısı kullanıldığında Makro-F1 skorları 48
Şekil 5.6. Sentiments veri kümesi için SVM (a), KNN (b), DT (c) ve NB (d) sınıflandırıcısı kullanıldığında Makro-F1 skorları 50
Şekil 5.7. Tüm veri kümeleri üzerinde SVM (a), KNN (b), DT (c) ve NB (d) sınıflandırıcıları kullanıldığında ortalama Precision ve Recall skorları 55
Şekil 6.1. XY metodun çalışma stratejisi 56
Şekil 6.2. Bir noktanın bir doğruya uzaklığı 57
Şekil 6.3. XY metoduna göre bölgeler gösterimi 59
Şekil 6.4. Basit örnek için her terimin koordinat düzleminde gösterimi 61
Şekil 6.5. Lamda değerlerine göre SVM sınıflandırıcı kullanılarak veri kümeleri üzerinde negatif bölge alt sınır belirleme 63
Şekil 6.6. SMS veri kümesi için SVM (a), KNN (b), DT (c) ve NB (d) sınıflandırıcısı kullanıldığında Makro-F1 skorları 66
Şekil 6.7. SLS veri kümesi için SVM (a), KNN (b), DT (c) ve NB (d) sınıflandırıcısı kullanıldığında Makro-F1 skorları 67

Şekil 6.8. AYSC veri kümesi için SVM (a), KNN (b), DT (c) ve NB (d) sınıflandırıcısı kullanıldığında Makro-F1 skorları 68
Şekil 6.9. Sentiments veri kümesi için SVM (a), KNN (b), DT (c) ve NB (d) sınıflandırıcısı kullanıldığında Makro-F1 skorları 70

SİMGELER VE KISALTMALAR DİZİNİ

α	:	Alfa
λ	:	Lamda
BoW	:	Kelime Çantası (Bag-of-Word)
CHI2	:	Ki-Kare (Chi Square)
DF	:	Doküman Frekansı (Document Frequency)
DFS	:	Ayırt Edici Öznitelik Seçici (Distinguishing Feature Selector)
DP	:	Poisson Dağılımından Sapma (Deviation from Poisson Distribution)
DT	:	Karar Ağacı (Decision Tree)
GI	:	Gini Katsayısı (Gini Index)
GR	:	Kazanım Oranı (Gain Ratio)
IDF	:	Ters Doküman Frekansı (Inverse Document Frequency)
IG	:	Bilgi Kazancı (Information Gain)
IR	:	Bilgi Erişimi (Information Retrieval)
KNN	:	K-En Yakın Komşu (K-Nearest Neighbor)
KKT	:	Kaba Kümeler Teorisi
MMR	:	Max-Min Oranı (Max-Min Ratio)
NB	:	Naive Bayes

NDM	:	Normalleştirilmiş Fark Ölçüsü (Normalized Difference Measure)
PRFS	:	Orantılı Kaba Öznitelik Seçici (Proportional Rough Feature Selector)
SMS	:	Kısa Mesaj Servisi (Short Message Service)
SVM	:	Destek Vektör Makinesi (Support Vector Machine)
TF	:	Terim Frekansı (Term Frequency)
TF-IDF	:	Terim Frekansı-Ters Doküman Frekansı (Term Frequency-Inverse Document Frequency)

1. GİRİŞ

Son yıllarda bilgi teknolojisinin çok hızlı ilerlemesi ile sosyal ağlar, lokasyon tabanlı hizmetler, sosyal medya ve e-ticaret gibi birçok yeni uygulama, bilgi toplama ve yayma merkezi ortaya çıkmıştır. Buna paralel olarak devasa büyüklükte bir kirli veri meydana gelmiştir. İnternette üretilen ve paylaşılan bu verilerin çok büyük bir çoğunluğu metin verileri biçimindedir. Örneğin, CNN ile ilgili haberler, Wall Street Journal web sitesinde listelenen makaleler, Twitter kullanıcıları tarafından gönderilen tweetler, Amazon'da yayınlanan ürünler hakkındaki yorumlar vb. daha birçok örnek sıralanabilir. Ancak bu bilgi toplama ve yayma merkezlerinde insanlar sadece ilgilendikleri web sayfalarını görmek ister, alakasız olanları görmek istemezler. Bu da metin verisinin içeriği içeren web sayfalarını (veya belgeleri) verimli bir şekilde endekslemek ve web sayfalarında aramayı kolaylaştırmak için konulara veya önceden tanımlanmış bazı kategorilere göre sınıflandırılması gerektiği anlamına gelir. Fakat metin verilerinin hacmi çok büyük olduğundan bu tür metin verilerinin manuel olarak işlenmesi imkansızdır. Bu zorluk ile başa çıkmak için belgeleri otomatik olarak sınıflandırmak için birçok teknik ve yöntem önerilmiştir. Bu sürece de metin kategorizasyonu veya sınıflandırması adı verilmiştir.

İnternetteki bu metin verilerinin çok büyük bir çoğunluğunu ise kısa metin belgeleri oluşturmaktadır. Özellikle de son yıllarda, elektronik ortamda genel olarak uzunluğu 200 karakteri geçemeyen kısa metin belgelerinin sayısı hızla artmaktadır. Sosyal medya, e-ticaret ve çevrimiçi iletişim gibi platformların kullanımındaki artışla kısa metin belgelerin varlığı çok önemli bir konu haline gelmiştir. Çevrimiçi sohbet günlükleri, anlık mesajlar, web günlüğü yorumları, bülten başlıkları, blog yorumları, soru-cevap web sitelerindeki sorular, internet haber yorumları, Twitter, Kısa Mesaj Servisi (Short Message Service-SMS) vb. aslında kısa metinler web'in her yerinde bulunmaktadır. Bu nedenle, bunları başarılı bir şekilde işlemek birçok Bilgi Erişimi (Information Retrieval-IR) ve web uygulamasında önemli bir yer tutmaktadır. Kısa metin belgelerinin sınıflandırılması bu tür verilerle ilgili en önemli görevlerin başında gelmektedir.

Kısa metinlerin yapısı nedeniyle sınıflandırmaları sürecinde bazı önemli problemlerle karşılaşılmaktadır. Kısa metin verilerindeki belgeler çok az kelime içerdiğinden, seyreklik (sparsity) sorunu ortaya çıkmaktadır. Bunun yanı sıra, tüm

belgelerden çıkarılan öznitelikler[1] çok yüksek boyuttadır. Bu problemler kısa metin sınıflandırmasında büyük bir dezavantaja neden olmaktadır. Başka bir ifade ile, yüksek boyutluluk ve seyreklik problemleri hem sınıflandırıcıların performansının azalmasına hem de hesaplama maliyetinin artmasına neden olmaktadır. Literatürde bu sorunların üstesinden gelmek ve özellikle yüksek boyutluluk sorununu çözmek için birçok çalışma bulunmaktadır. Örneğin Rao ve ark. (Rao, ve diğerleri, 2016) sosyal medyada kısa metin dokümanlarında duygu sınıflandırması yapmak için yeni bir model önermiştir. Önerilen modelde ilk olarak seyreklik sorununun etkisi en aza indirilerek, gizli konuları belirlemek için bir konu çıkarma işlemi gerçekleştirilmiştir. Daha sonra, kullanıcılar tarafından bildirilen çoklu duygu etiketlerini ilişkilendirmek için de Bernoulli modeli kullanılmıştır. Bu çalışmanın temel amacı, kısa metinlerde konu çıkarma ve çoklu duygu gibi bazı kavramları modelleyerek duyguları kullanılabilir bir yapıya dönüştürmektir. Li ve ark. (Yang, Q., & Li, 2013) kısa metin belgelerinde sınıflandırma için yeni bir yöntem önermiştir. Bu çalışmada, kısa metinlerin hem sözcüksel hem de anlamsal özellikleri kullanılarak belgeler sınıflandırılmıştır. Başka bir deyişle, çalışma hem sözcüksel özellikleri seçmek hem de arka plan bilgisini kullanarak anlamsal özellikleri ayıklamak için yeni bir yöntemi sunmaktadır. Ayrıca bu sözcüksel ve anlamsal öznitelikler birleştirilerek sınıflandırmada kullanılmıştır. Kim ve ark. (Kim, ve diğerleri, 2014) kısa metin dokümanları arasındaki benzerliği belirleyen yeni bir yaklaşım sunmuşlar. Önerilen yaklaşım, benzerliği hesaplamak için herhangi bir dilbilgisi etiketini ve sözcüksel veri kümesini kullanmaz. Yine (Rashid, Shah, & Irtaza, 2019), (Khan, Chowdhury, Ngo, & Apon, 2020), (Lochter, Zanetti, Reller, & Almeida, 2016) bu alanda yapılmış çalışmalardan bazılarıdır.

Yüksek boyutluluğa çözüm sağlamak için en etkili tekniklerden biri, öznitelik seçme yöntemlerini kullanarak tüm öznitelikler arasında ayırt ediciliği yüksek öznitelikleri seçmektir. Seçilen öznitelik kümesi tüm öznitelik kümesini en iyi temsil eden küme olması beklenir. Bunun için kullanılan öznitelik seçim yönteminin başarısı çok önemlidir. Daha genel bir ifade ile öznitelik seçme yöntemleri, sınıflandırıcı, makine öğrenimi gibi karar sistemlerinin performansını ve başarısını arttıran ve aynı zamanda onların yürütme zamanlarını da düşüren tekniklerdir. Özellik seçme teknikleri genellikle filtre (filter), gömme (embedded) ve sarmalayıcı (wrapper) yaklaşımları olmak üzere üç başlık altında

[1] Bu çalışma kapsamında "öznitelik" ve "terim" kavramları aynı anlamda kullanılmıştır. Dolayısıyla iki kavram tez boyunca birbirlerinin yerine kullanıldığı görülebilir.

toplanır. Filtre yaklaşımları her öznitelik için bir değer hesaplar ve en yüksek değerlere sahip belirli sayıda öznitelik seçer. Bu işlemler için herhangi bir sınıflandırıcı veya öğrenme algoritması kullanmazlar. Sarmalayıcı yaklaşımları, özniteliklerin en iyi alt kümesini seçmek için bir sınıflandırıcı kullanır. Gömülü yaklaşımlar, öznitelik seçme sürecinin sınıflandırıcıya entegre edilmesiyle oluşturulur ve bu tür öznitelik seçme yaklaşımları öğrenme algoritmalarına özgüdür.

Sonuç olarak, kısa metinler normal metinlerin özel halleri olarak düşünülebilir. Ancak kısa metinlerin normal metinlerden farkı metin uzunluklarının en çok 200-250 karakter olmasıdır. Metin sınıflandırma için yürütülen tüm işlemler kısa metin sınıflandırma için de yürütülür. Başka bir ifade ile metin sınıflandırma süreçlerinin tamamı kısa metin sınıflandırma için de uygulanır. Ancak yapılarından dolayı kısa metinleri sınıflandırmak normal metinlerden daha zordur. Bunun için ek işlemlerin yapılması gerekebilir.

1.1. Tezin Amacı ve Hedefleri

Bir önceki bölümde kısa metin dokümanların oluşumuna ve önemine değinilmişti. Ayrıca bu dokümanların analizi sırasında karşılaşılan temel problemler, seyreklik ve yüksek boyutluluk problemlerinden de söz edilmişti. Kısa metinlerin günlük hayatımızdaki varlığı düşünüldüğünde onlar üzerinde yapılan madenciliğin önemini her zaman güncel tutacağı açıktır. Literatürde yüksek boyutluluk sorununun üstesinden gelmek için birçok çalışma olmasına rağmen, daha iyi ve daha etkili çözümler önermek hala bir zorunluluktur. Bu yüzden daha iyi ve daha etkili çalışabilecek yaklaşımlar sunmak amacıyla bu tez çalışmasında aşağıdaki sorulara yanıt aranmıştır:

- İçerdiği dokümanların yeteri sayıda kelime içermediğinden vektör uzay model ile gösterilen kısa metin verisi boşluklu bir veridir. Bu yüzden kısa metin verisini sınıflandırmak diğer uzun metin verisinden daha zordur. Kısa metin verisinin bu yapısından dolayı metin sınıflandırma için kullanılan birçok öznitelik seçme metodu bu tip veriler üzerinde etkili çalışmamaktadır. Buna göre kısa metin verileri üzerinde başarılı bir şekilde çalışan yeni yaklaşımlar sunulabilir mi?
- Metin sınıflandırma için kullanılan birçok öznitelik seçme yaklaşımı sınıf bilgisini görmezden gelerek sadece terim frekanslarına veya doküman frekanslarına bakmaktadır. Halbuki sınıf bilgisi metin sınıflandırma için önemli ve kullanışlı bir bilgidir. Ayrıca bu metotlar sadece pozitif ve negatif sınıflardaki dokuman

frekanslarına dayalı çalışmaktadır. Bu durumları göz önünde bulundurarak yeni yaklaşımlar sunulabilir mi?

- Birçok yaklaşım terimleri seçerken çoğunlukla terimin sadece bir sınıfta geçmesine yoğunlaşmaktadır. Fakat bir terimin ayırt ediciliği birçok kritere bağlıdır. Farklı kriterlere göre çalışan yeni yaklaşımlar sunulabilir mi?
- Birçok yaklaşım terimleri seçerken terimlerin değer kümesini dikkate almamaktadır. Halbuki birçok sınıflandırıcı için bu değerler önemlidir. Dolasıyla bu terim değer kümesi öznitelik seçme sürecinde dikkate alınabilir mi?

Yukarıdaki amaçlar doğrultusunda kısa metin alanında seyreklik probleminin etkisinden minimize etkilenecek şekilde modellenmiş, etkili iki tane filtre öznitelik yaklaşımı önerilmiştir. Bu yaklaşımlardan ilki Orantılı Kaba Öznitelik Seçici (Proportional Rough Feature Selector-PRFS) adı verilen kaba kümeler tabanlı yaklaşımdır. İkincisi ise XY metot olarak tanıtılan istatiksel yaklaşımdır. Önerilen metotlar ile aşağıda belirtilen hedeflere ulaşılması planlanmıştır:

Kaba kümelerin eksik veriden etkili bir şekilde gizli bilgileri çıkarım yapma yeteneğinden faydalanarak sunulan PRFS ile;

- Terimin sadece bir sınıfta geçmesine yoğunlaşmaktansa terimin sınıf ile ilişkisini ortaya çıkarma
- Terimin vektör uzayında boşluklu yapının sınıf üzerindeki etkisini hesaplama
- Küme yaklaşımları ile terim bazında bölgeler oluşturup kritik bölgede olan dokümanlara bir ceza uygulamak ve buna bağlı daha etkin bir ayırt edicilik sunma
- Terim vektör uzayındaki değerleri dikkate alarak tüm terim uzayını en iyi ifade eden bir alt terim uzayını seçmeyi hedeflemektedir.

İstatiksel bir yaklaşım olan XY metodu ile;

- Sadece terim frekanslarına veya doküman frekanslarına bakmak yerine sınıf bilgisini de dikkate alarak seçim sunma
- Sadece pozitif ve negatif sınıflardaki dokuman frekanslarına bakmak yerine her sınıfın diğer tüm sınıflar ile ikili kombinasyonlarına bakarak daha anlamlı ve etkili bir sonuç çıkarmayı hedeflemektedir.

1.2. Tez Organizasyonu

Bu tez, organizasyon açısından toplam yedi bölümden oluşmaktadır. Çalışmanın genel anlatımı, amacı ve hedefleri 1. Bölümde yer almaktadır. 2. Bölümde çalışma

kapsamında faydalanan teknolojiler ve yaklaşımlar ile ilgili temel bilgiler yer almaktadır. Ayrıca bu bölümde kıyaslama için kullanılan yaklaşımlara ait bilgiler de bulunmaktadır. Çalışmanın genel motivasyonu ve ilgili çalışmalar 3. Bölümde verilmiştir. Kısa metin hakkında genel bilgi 4. Bölümde bulunmaktadır. 5. Bölüm çalışma kapsamında sunulan kaba küme tabanlı PRFS ve onun deneysel çalışmaları hakkındadır. Tezin deneysel çalışmalarının bulunduğu diğer bölüm 6. Bölümdür. Bu bölümde XY metodu tanıtılmıştır. Sonuç ve çalışma hakkındaki tartışmalara ise 7. Bölümde verilmiştir.

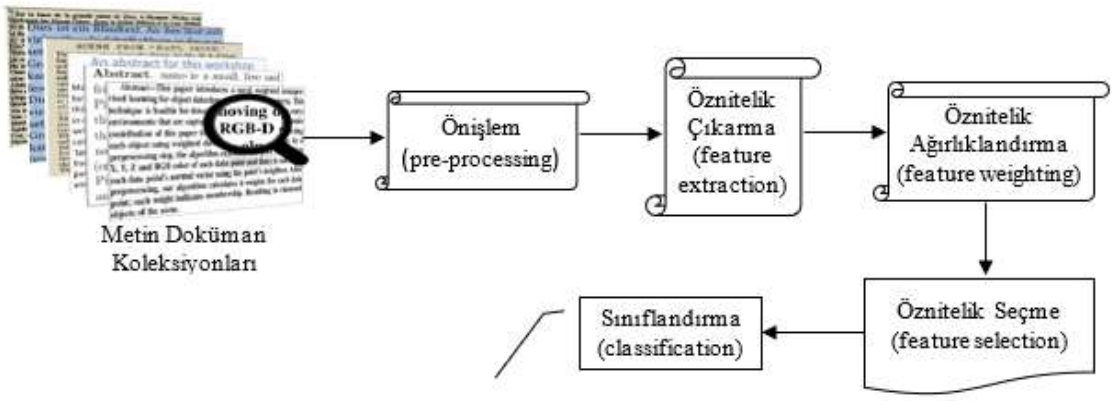
2. METİN SINIFLANDIRMA AŞAMALARI VE ÖNBİLGİ

2.1. Metin Sınıflandırma ve Süreçleri

Bir önceki bölümde dijital metinlerin kaynaklarının ne olduğu, ne tür veriler olduğu ve her gün sayısının nasıl hızla artığından bahsedilmişti. Bu bölümde ise bu veriler üzerinde yapılan analiz işlemlerden biri olan sınıflandırmadan bahsedilecektir.

Metin analitiği olarak da adlandırılan metin madenciliği belgelerdeki yapılandırılmamış metni doğal dil işlemeyi kullanarak, analize veya makine öğrenmesi algoritmaları için uygun hale getiren bir yapay zekâ teknolojisidir. Bilgi odaklı kuruluşlarda yaygın olarak kullanılan metin madenciliği, çok büyük metin koleksiyonlarından yeni bilgi çıkarmayı sağlayan veya özel araştırmalara yönelik cevaplar sunan bir inceleme sürecidir. Web siteleri, kitaplar, e-posta'lar, makaleler, online haberler gibi yazılı kaynaklardan çıkarım sağlayan metin madenciliği günlük hayatta kullanılan dilleri işleyerek, gelişmiş yaklaşımlar yardımı ile verileri yapılandırır. Bu veriler veri tabanlarına, veri ambarlarına veya iş zekâsı panellerine entegre edilebilir ve açıklayıcı, kuralcı veya tahmine dayalı analitik için kullanılabilir. Metin madenciliği; güvenlik, biyomedikal, online medya, duygu analizi, ticaret ve pazarlama, dijital beşeri bilimler ve hesaplamalı sosyoloji, bilimsel literatür madenciliği ve akademik uygulamalar gibi çok geniş bir uygulama alanına sahiptir. Metin sınıflandırma, metin kümeleme, kavram veya varlık çıkarımı, tanecikli taksonomi modelleme, duygu analizi, varlık ilişki modeli, doküman özet çıkarma vb. işlemler genel olarak metin madenciliğinin görevlerini oluşturmaktadır. Bu görevlerin başında metin sınıflandırma işlemi gelmektedir. Metin sınıflandırma çeşitli önişlemlerin uygulanmasından sonra metin dokümanlarına sınıflandırma algoritmalarının uygulandığı bir süreçtir. Amaç, metin belgesini belgenin içeriğine göre daha öncesinde kategorize edilmiş veya sınıflandırmış bir gruba atmaktır. Gerçek hayatta bir haberin; politika, spor, ticaret veya sağlık gibi kategorilerden hangisi ile ilgili olduğunu bulma işlemi metin sınıflandırmaya bir örnektir. Benzer şekilde bir mailin gereksiz (spam) olup olmadığını bulma ya da bir filmin, onun hakkında yapılmış yorumlara bakarak iyi, kötü veya ne iyi ne kötü olduğuna karar vermek metin sınıflandırmaya örnek gösterilebilir.

Metin sınıflandırma işlemi çeşitli işlemler ve süreçlerden oluşmaktadır. Aşağıda Şekil 2.1'de metin sınıflandırmanın içerdiği süreçlerin genel bir gösterimi verilmiştir.



Şekil 2.1. *Metin sınıflandırma genel şeması*

Yukarıdaki şemada da (Şekil 2.1) görüldüğü gibi metin sınıflandırma farklı süreçlerden oluşmaktadır. Bu süreçler sırasıyla önişlem, öznitelik çıkarma, öznitelik ağırlıklandırma, öznitelik seçme ve sınıflandırma işlemleridir. Her bir süreç ile ilgili açıklamalar aşağıdaki başlıklarda verilmiştir.

2.1.1. Önişlem (Pre-Processing)

Önişlem metin sınıflandırmanın çok önemli ve kritik bir adımını oluşturmaktadır. Önişlem aşaması, metin koleksiyonları üzerinde veri temizleme, sözcüklerin anlamsal değerlerini bulma, veri normalizasyonu ve veri bütünlüğü sağlama gibi birtakım işlemler dizini olarak da bilinir. Bu süreçte veya aşamada genel olarak aşağıdaki işlemler yapılır:

- İstenmeyen (noise) verileri temizleme (yazım hatalarından kaynaklı verilerin silinmesi veya düzeltilmesi vb.)
- Gereksiz kelimeleri (bağlaç, edat, zamir vb.) ayıklama
- Kelimelerin anlamsal değerlerini bulma (isim, fiil, zarf vb.)
- Noktalama işaretlerini kaldırma
- Metni bölümlere veya sözcüklere ayırma
- Küçük-büyük harf dönüşümünü yapma
- Kelimeleri köklerine ayırma (varsa kelimede ekin kaldırılması vb.).

Önişlem aşaması veri temizlemenin yanında veri bütünlüğü ve normalizasyonu da sağlayarak uygun formata getirmeyi amaçlamaktadır. Yaygın olarak tokenizasyon (tokenization), durma sözcüklerinin kaldırılması (stop-word removal), küçük harf dönüşümü (lowercase conversion) ve kök bulma (stemming) olmak üzere 4 başlıkta ele

alınmaktadır. Ayrıca bazı çalışmalarda öznitelik çıkarımı da önişlem süreci olarak işlenmektedir ancak bu çalışmada ayrı bir başlık altında ele alınacaktır.

2.1.1.1. Tokenizasyon (Tokenization)

Bir metni sözcüklere, ifadelere veya diğer anlamlı parçalara, yani belirteçlere (tokenlere) bölme prosedürüdür. Başka bir deyişle, tokenizasyon bir metini parçalama biçimidir. Yapısal olarak metin parçalama sadece alfa sayısal olmayan karakterler (örn. noktalama işaretleri, boşluk) ile sınırlanan alfabetik veya alfa sayısal karakterler dikkate alınarak gerçekleştirilir. Tokenizasyon ile metin belgeleri bölüm, paragraf, cümle, kelime veya hece gibi dizgelere ayırılabilir. Dizgelere ayırma işlemi metnin bağlı olduğu doğal dile göre farklılık gösterebilir (Manning, Raghavan, & Schütze, 2010).

2.1.1.2. Durma sözcüklerin kaldırılması (Stop-word removal)

Durma sözcükleri (stop-words) özel veya belirli bir konuyla ilgisi olmayan edat, zamir ve bağlaç gibi metin içinde çok sık geçen kelimelerdir. Bu tür kelimeler tek başına bir anlam ifade etmezler. Bundan dolayı durma sözcüklerinin genellikle metin sınıflandırma çalışmalarında alakasız olduğu varsayılır ve sınıflandırma işleminden önce kaldırılır. Durma sözcükleri, üzerinde çalışılan doğal dile özgüdür. Dolayısıyla durma sözcük listeleri dillere göre değişiklik göstermektedir. Örneğin, İngilizce için “the”, “an”, “a”, “he”, “on”... iken Türkçe için “ve”, “veya”, “onlar”, “ile”, “de”… şeklinde olabilir.

2.1.1.3. Küçük harf dönüşümü (Lowercase conversion)

Metin sınıflandırmada bütünlüğü sağlamak ve tutarlı bir bilgi sistemi elde etmek için kelimelerin büyük veya küçük harf formlarının hiçbir farkı olmadığı varsayılarak, tüm büyük harfli karakterler sınıflandırmadan önce genellikle küçük harfli biçimlere dönüştürülür veya tersi yapılır.

2.1.1.4. Kök bulma (Stemming)

Bu aşamada her kelimenin kökü bulunur. Amaç, farklı ek almış kelimelerin aynı kökle temsil edilmesini sağlamaktır. Bu şekilde hem öznitelik sayısı azaltılır hem de daha doğru bir kelime frekansı hesaplanması sağlanır. Kök bulma işlemi, her farklı doğal dil için sunulmuş farklı algoritmalar ile gerçekleştirilir. Ayrıca aynı dil için farklı yaklaşımlar da mümkündür. Örneğin, Zemberek (Akın & Akın, 2007) başta Türkçe olmak üzere Türk dilleri için tasarlanmış açık kaynaklı, platformdan bağımsız, genel amaçlı bir doğal dil işleme kütüphanesi ve araç takımıdır. Yine aynı amaçla Türkçe dili için oluşturulmuş

basit ama etkili yaklaşımlardan biri olan sabit önek algoritması (the fixed-prefix algorithm) Can ve ark. (Can, ve diğerleri, 2008) tarafında sunulmuştur. Aynı şekilde İngilizce dili için Porter (Porter, 1980) bir kök bulma yaklaşımı önermiştir.

2.1.2. Öznitelik çıkarma (Feature extraction)

Yapılandırılmamış metin belgeleri üzerinde gerekli önişlemler yapıldıktan sonra bu belgelere öznitelik çıkarma teknikleri uygulanabilir. Öznitelik çıkarmadaki amaç, metin belgelerinden bir kelime listesi oluşturmak ve bunu sınıflandırıcıların kullanabileceği hale dönüştürmektir. Bu amaç doğrultusunda yaygın olarak kullanılan teknikler sırasıyla kelime çantası (the bag of words-BoW) (Aggarwal & Zhai, 2012) ve Word2Vec yaklaşımlarıdır.

2.1.2.1. Kelime çantası (The bag of words)

N-gram olarak da bilinen kelime çantası (BoW) yöntemi diğer yöntemlere göre en yaygın kullanılan ve en basit yaklaşımdır. Bu yaklaşıma göre bir kelimenin dokümanda sıklığı ve sırasının bir önemi yoktur. Önemli olan tek şey dokümanda geçip geçmediğidir. Yöntem kaç kelimeyi birlikte bir öznitelik olarak alıyorsa o ismi almaktadır. Örneğin, her kelime bir öznitelik olarak alınırsa 1-gram (unigram), iki kelime alırsa 2-gram (bigram), üç kelime alırsa 3-gram (trigram) ve bu şekilde N-gram kadar devem eder.

2.1.2.2. Word2Vec

Word2Vec yaklaşımı kelime kalıplama (word embeddings) yapısını kullanmaktadır. İki katmanlı sinir ağı ile modellenen yöntem girdi olarak aldığı büyük metin verisindeki kelimeleri vektör uzayında ifade eder. Daha sonra her ayırt edici kelime, uzayda bir vektör ile temsil edilir ve ortak bağlamlara sahip kelimeler ise vektör uzayının yakınına yerleştirilir. Bu yaklaşımın iki alt yöntemi mevcuttur: CBOW (Continous Bag of Words) ve Skip-Gram.

2.1.3. Öznitelik Ağırlıklandırma (Feature Weighting)

Metin sınıflandırmanın önemli diğer bir aşaması terim ağırlıklandırma sürecidir. Bu aşamada dokümanların her bir terimi için bir terim ağırlıklandırma algoritması kullanılarak terimin ağırlığı hesaplanır. Amaç, belli dokümanlarda gecen sınıflandırmada ayırt edici özel bilgi sağlayan terimler ile tüm dokümanlarda yaygın geçen özel bilgi taşımayan terimlerin farkını ortaya çıkarmaktır. Bu amaç doğrultusunda farklı yaklaşımlar sunulmuştur. Bunlardan yaygın olarak kullanılanları: İkili Ağırlıklandırma,

TF-IDF (Jones, 1972), TF-PB (Liu, 2009), TF-IGM (Chen K. Z., 2016)ve TF-IDF-ICF (Ren, 2013) gibi terim ağırlıklandırma şemalarıdır. Ayrıca yakın zamanda sunulan TF-IGM $_{imp}$ ve SQRT_ TF-IGM$_{imp}$ (Dogan & Uysal, 2019) yöntemleri de etkili terim ağırlıklandırma şemalarından iki tanesidir. Tez kapsamında İkili Ağırlıklandırma ve TF-IDF terim ağırlıklandırma kullanıldığında, bunlar ile ilgili bilgi aşağıda alt başlıklar halinde verilmiştir.

2.1.3.1. İkili ağırlıklandırma (Binary weighting)

İkili ağırlıklandırma (Binary weighting) yönteminde terimin dokümanda geçme sıklığına bakmadan sadece dokümanda geçip geçmediğine bakılır. Ayrıca bu ağırlıklandırma yaklaşımında sınıf bilgisinin de bir önemi yoktur. Hesaplanması ve kullanımı çok basit bir metot olan ikili ağırlıklandırma yaygın olarak kullanılmaktadır. İkili ağırlıklandırma yönteminin matematiksel arka planı Eşitlik 2.1'de verilmiştir.

$$\text{İ}kili(t_i, d_k) = \begin{cases} 1 & eğer\ t_i\ terimi\ d_k dokümanında\ geçiyorsa \\ 0, & aksi\ takdirde \end{cases} \quad \textbf{(2.1)}$$

2.1.3.2. Terim frekans-Ters doküman frekansı (TF-IDF)

Metin verilerinde yüksek frekansa sahip kelimeler fazladan hesaplama maliyetine neden olabilir. Örneğin, İngilizcede "the" sözcüğü çok sık kullanılır. Bu da bu sözcüğün frekansının yüksek olması demek ve tüm dokümanlarda görülmesi anlamına gelir. Bu tür sözcükler metin sınıflandırmada fazla ayırt ediciliği olmayan terimlerdir ve atılması gerekebilir. Bu problemin üstesinden gelmek için terim frekans-ters doküman frekansı (term frequency-inverse document frequency-TF-IDF) (Sparck Jones, 2004) yöntemi, dokümanda kelimenin frekansını ve tüm dokümanlarda kelimenin skoruna bakmaktadır. Bu skor yardımı ile belirli bir belgedeki gerekli bilgileri temsil eden benzersiz kelimeler metinden çıkarılır. Dolayısıyla sık olmayan bir sözcüğün IDF'si yüksek iken sık bir sözcüğün ise düşüktür. Bu yüzden TF-IDF yönteminde ters doküman frekansı değerlerinin hesaplanması verilen metin koleksiyonunda sık geçen terime düşük, nadir geçen terime ise yüksek skor atanması sağlanır. Matematiksel olarak TF-IDF:

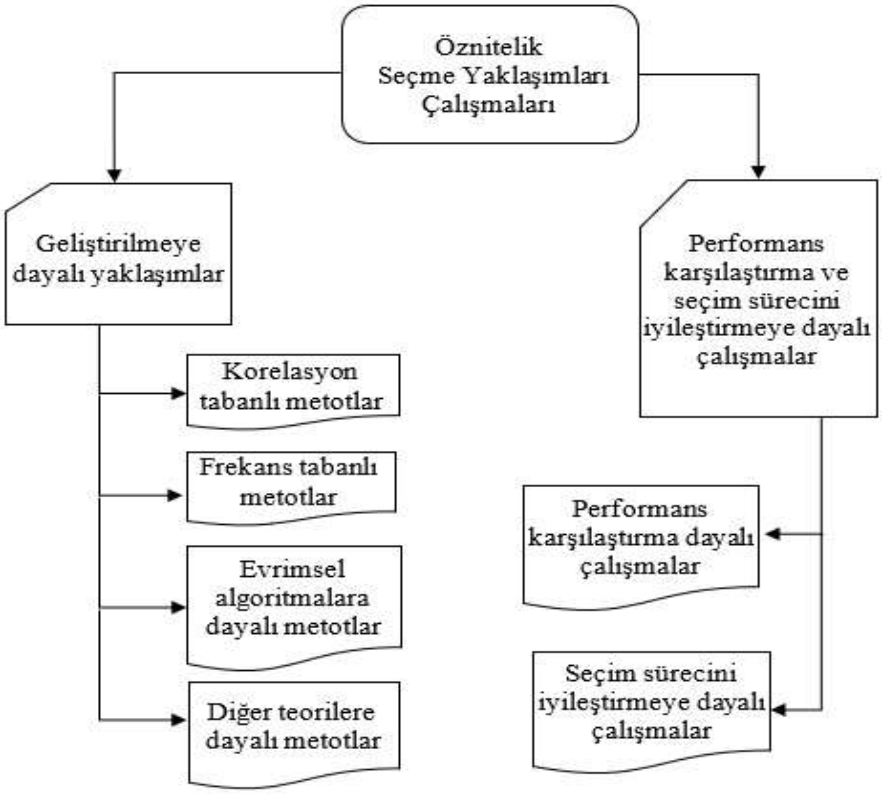
$$W_{TF.IDF}(t_i) = TF(t_i, d_k) * \log(\frac{D}{d(t_i)}) \quad \textbf{(2.2)}$$

Formülde yer alan D değeri koleksiyondaki toplam doküman sayısını, $d(t_i)$ ise t_i teriminin geçtiği toplam doküman sayısını ifade etmektedir. Eşitlik 2.2'de de anlaşıldığı gibi yöntem sınıf bilgisini kullanmamaktadır.

2.1.4. Öznitelik seçme (Feature selection)

Metin sınıflandırmanın en genel problemi terim uzayının çok yüksek boyutlu olmasıdır. Bu problem sınıflandırıcı, makine öğrenmesi gibi karar sistemlerin hem performansını hem de yürütme zamanını oldukça etkilemektedir. Bu probleme çözüm olarak ya terim uzayı boyutu indirgenir ya da bu uzayın bir alt kümesi seçilir. Öznitelik seçme terim uzayında tüm terim uzayını en iyi temsil eden alt kümeyi seçme işlemi olarak bilinir. Dolayısıyla öznitelik seçme metin sınıflandırmada çok önemli bir yer tutmaktadır. Bu yüzden öznitelik seçme ile ilgili çok sayıda çalışma olmasına rağmen bu alanda hala çalışmalar güncelliğini korumakta ve geliştirilmeye ihtiyaç duymaktadır.

Öznitelik seçme teknikleri yaygın olarak filtre (filter), sarmalayıcı (wrapper) ve gömülü (embedded) yaklaşımlar olmak üzere üç başlık altında toplanır. Filtre yaklaşımlar her bir terim için bir skor hesaplar ve skoru en yüksek n tane özniteliği seçer. Filtre yaklaşımları terim seçme işlemleri için herhangi bir sınıflandırıcı veya öğrenme algoritmasını kullanmaz. Sarmalayıcı yaklaşımları ise tamamıyla bir sınıflandırıcıya bağlı olarak tüm terim uzayını en iyi temsil eden terim alt kümesini seçmeyi amaçlamaktadır. Gömülü yaklaşımlar hem filter hem de sarmalayıcı yaklaşımlarının üstün özelliklerinin birleştirilmesiyle oluşturulmuş yaklaşımlardır. Ayrıca öznitelik seçme tekniklerinin yanında bu alanda yapılmış çalışmaları da gruplandırmak mümkündür. Literatürde var olan çalışmalara bakıldığında bu alanda yapılmış çalışmalar iki ana başlık altında toplanabilir (Wang & Hong, 2019). Şekil 2.2'de metin sınıflandırma alanındaki öznitelik seçme yaklaşımları ile ilgili yapılmış çalışmaların gruplandırması gösterilmiştir.



Şekil 2.2. *Öznitelik seçme metotları ile ilgili çalışmaların gruplandırılması*

2.1.4.1. Geliştirilmeye dayalı yaklaşımlar

Geliştirilmeye dayalı çalışmalarda amaç, daha etkili performans sergileyen veya daha anlamsal terimler seçen yeni yaklaşımlar sunmaktır. Bu tür çalışmalar temel seçme kriterine göre 4 grupta toplanabilir.

Korelasyon tabanlı metotlar: Bu tür çalışmalarda sunulan metotlar temel seçme kriteri olarak ya terim ile sınıf arasındaki korelasyonu ya da sadece terimler arası korelasyonu ele alırlar. Bu tür metotların büyük bir çoğunluğu terim ile sınıf arasındaki ilişkiye göre çalışmaktadır. Bu yaklaşımlara göre bir terimin ayırt ediciliğinin yüksek olması terim ile sınıf arasındaki ilişkinin güçlü, diğer terimler ile zayıf olmasına bağlıdır. Geliştirilmiş Gini katsayısı (Improved Gini index-IGI) (Shang, ve diğerleri, 2007), ayırt edici öznitelik seçici (distinguishing feature selector-DFS) (Uysal & Gunal, 2012), global bilgi kazanımı (global information gain-GIG) (Shang, Li, Feng, Jiang, & Fan, 2013), sınıf ayırıcı ölçümü (class discriminating measure-CDM) (Chen, Huang, Tian, & Qu, 2009), çok değişkenli göreceli ayrımcılık kriteri (multivariate relative discrimination criterion-MRDC) (Labani, Moradi, Ahmadizar, & Jalili, 2018), Fisher'ın ayırt edici oranı (Fisher's discriminant

ratio) tabanlı metot (Wang, Li, Wei, & Li, 2009), Normalleştirilmiş fark ölçüsü (normalized difference measure-NDM) (Rehman, Javed, & Babri, 2017) ve max-min oranı (max-min ratio-MMR) (Rehman A. , Javed, Babri, & Asim, 2018) gibi metotlar en iyi bilinen geliştirilmiş korelasyon tabanlı metotlardır. Bu çalışmada önerilen yaklaşımlar birer korelasyon tabanlı metotlardır.

Frekans tabanlı metotlar: Bu tür metotlarda genel olarak terim frekansı veya doküman frekansı dikkate alınır. Sınıf bilgisi dikkate alınmaz. Ancak sınıf bilgisi metin sınıflandırma için önemli bir faktördür. Bu alanda yapılmış çalışmalardan bazıları şunlardır: Gong ve arkadaşları (Gong, Zeng, & Zhang, 2011) tarafından terim frekanslarının ortalamasının karesi ile terim frekanslarının karesinin farkının kıyaslanmasıyla yeni bir model sunulmuş. Tutkan ve arkadaşları (Tutkan, Ganiz, & Akyokuş, 2016) ise Helmholtz prensibini kullanarak terim frekansları elde edilmiş ve buna bağlı yeni bir yaklaşım modeli oluşturmuşlar. Bharti ve Sing (Bharti & Singh, 2015) tarafından önerilen metotta dokuman frekanslarını ve terim varyansları kullanarak terim uzayında bir alt küme seçme modellemeye çalışılmış. Bu çalışmanın amacı ayırt ediciliği yüksek terimleri seçerek terim uzayının boyutunu indirgemektir. Ayrıca (Yao, Liu, Zhang, & Wang, 2017), (Wang & Hong, 2015), (Shamsinejadbabki & Saraee, 2012) yine bu alanda yapılmış çalışmalardan bazılarıdır.

Evrimsel algoritmalara dayalı metotlar: Bu yaklaşımların performansı genel olarak tekrarlı optimizasyon tekniklerini kullandıklarında korelasyon ve frekans tabanlı yaklaşımlara göre daha kötüdür. Çünkü optimizasyon süreci uzun süre alabilir. Bu alanda yapılmış çalışmalara genetik algoritma ve Kaos optimizasyonun entegrasyonuna dayalı Chen ve arkadaşlarının (Chen, Jiang, Li, & Li, 2013) sezgisel çalışması, parçacık sürü optimizasyonunu kullanılarak modellenmiş Lu ve arkadaşlarının (Lu, Liang, Ye, & Cao, 2015) yaklaşımı, metin kümeleme için genetik işlemlerle parçacık sürü optimizasyonun hibrit edilmesiyle Abualigah ve Khader'in (Abualigah & Khader, 2017) geliştirdiği metot ve yine parçacık sürü optimizasyonuna bağlı Kushwaha ve Pant (Kushwaha & Pant, 2018) tarafından geliştirdikleri yaklaşım birer örnektir.

Diğer teoriler dayalı metotlar: Bu bölümü oluşturan çalışmalar parametre tahmini, olasılık tahmini veya yardımcı bilgi gibi teorilere dayalı çalışmalardır. (Tommasel & Godoy, 2018a) ve (Li, Lu, Sun, & Xing, 2017) çalışmaları bu alanda yapılmış çalışmalara birer örnektir.

2.1.4.2. Performans karşılaştırmaya ve seçim sürecini iyileştirmeye dayalı çalışmalar

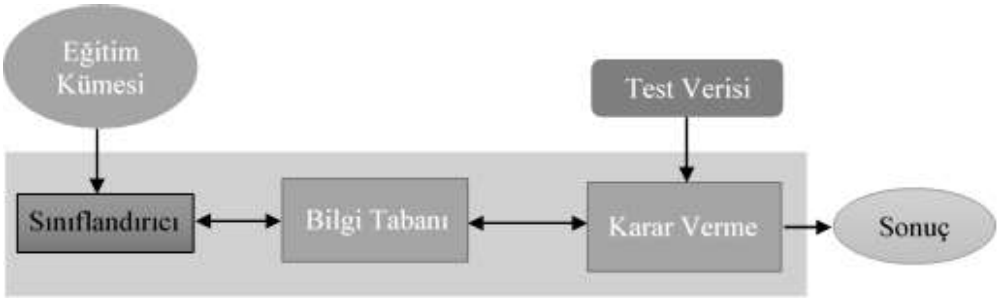
Bu yöndeki araştırmalar, diğer araştırmalarda önerilen öznitelik seçim yöntemlerinin performansı karşılaştırmak veya terimlerin seçim sürecini iyileştirmek için çaba sarf etmektedir. Bu yönde yapılmış çalışmaları iki başlık altında toplamak mümkündür.

Performans karşılaştırmaya dayalı çalışmalar: Bu tür çalışmalarda var olan metotların performans kıyaslaması yapılmıştır. Örneğin, Mesleh'in (Moh'd Mesleh, 2011) çalışmasında on yedi tane yöntem kıyaslaması yapılmıştır. Ogura ve ark. (Ogura, Amano, & Kondo, 2011) ise dengesiz veri kümeleri üzerine yöntemlerin performans kıyaslaması yapmıştır. Benzer şekilde Baccianella ve ark. (Baccianella, Esuli, & Sebastiani, 2013) öznitelik seçme üzerine terim frekansının etkisini araştırmıştır ve yine Tommasel ve Godoy (Tommasel & Godoy, 2018b) genel bir araştırma çalışması yapmıştır.

Seçim sürecini iyileştirmeye dayalı çalışmalar: Bu tür çalışmaların amacı terim seçme süresini indirgemektir. Başka bir ifade ile var olan metotlardan yürütme zamanı olarak daha hızlı yaklaşımlar sunmaktır. (Guru, Suhil, Raju, & Kumar, 2018) ve (Agnihotri, Verma, & Tripathi, 2017) bu alanda yapılmış çalışmalara birer örnektir.

2.1.5. Sınıflandırma

Sınıflandırma kavramı temel olarak veri kümesinde daha önce etiketlenmiş kategorilerden birine veri atama olarak tanımlanabilir. Sınıflandırmanın temel amacı, belirtilen veriler için hedef sınıfı doğru bir şekilde tahmin etmektir. Sınıflandırma sisteminin ana akışı Şekil 2.3'te gösterilmektedir.



Şekil 2.3. *Sınıflandırma sisteminin akış şeması*

Sınıflandırma işlemi için bir sınıflandırıcıya ihtiyaç vardır. Bu amaç doğrultusunda oluşturulmuş literatürde birçok sınıflandırıcı bulunmaktadır. Sınıflandırıcılar ile ilgili detaylı bilgi ilerleyen sayfalarda "Sınıflandırma Metotları" alt başlığında verilmiştir.

2.2. Öznitelik Seçme Teknikleri

Öznitelik seçme yöntemleri, mevcut tüm öznitelik arasında en yüksek ayırt ediciliğe sahip öznitelikleri seçerek yüksek boyutluluk sorununa bir çözüm sunmayı amaçlamaktadır (Forman, 2003). Bu amaç doğrultusunda, literatürde filtre, sargı ve gömülü gibi farklı çalışma mekanizmalarına sahip yaklaşımlar mevcuttur. Ancak bu bölümde, önerilen yöntemler filtre bazlı yaklaşım olduklarından sadece karşılaştırma için kullanılan filtre yaklaşımlarından bahsedilecektir. Bu tür yöntemlerde öznitelik seçimi; öznitelikler arası mesafe, öznitelikler arasındaki bağımlılık ölçümleri veya öznitelikler hakkındaki bilgi gibi istatistiksel kriterlere dayanan fonksiyonlar kullanılarak yapılır. Bu istatiksel fonksiyonlar aracılığıyla her öznitelik için bir puan[2] hesaplanır ve en iyi öznitelik alt kümesini oluşturmak için en yüksek puanlara sahip öznitelikler seçilir (Chandrashekar & Sahin, 2014). Filtre yöntemlerinin akışı süreci Şekil 2.4'de gösterilmektedir.

Şekil 2.4. *Filtre yaklaşımlarının akış süreci*

2.2.1. Bilgi kazanımı (Information gain-IG)

Özellikle veri ve metin madenciliğinde sıklıkla kullanılan bilgi kazanımı (information gain-IG) (Yang & Pedersen, 1997), en basit tanımı ile entropinin tersi olarak tanımlanabilir. Entropi kısaca bir sistemin düzensizliğini ifade eden terminolojidir. IG matematiksel olarak aşağıdaki gibi ifade edilir:

[2] Bu çalışma kapsamında "puan" ve "skor" kavramları aynı anlamda kullanılmıştır.

$$IG(t) = -\sum_{i=1}^{M} P(C_i) logP(C_i) + P(t) \sum_{i=1}^{M} P(C_i|t) logP(C_i|t) + P(\bar{t}) \sum_{i=1}^{M} P(C_i|\bar{t}) logP(C_i|\bar{t}) \quad \textbf{(2.2)}$$

Formülde M sınıf sayısını ve $P(C_i)$, C_i sınıfının olasılığını göstermektedir. $P(C_i|t)$ ve $P(C_i|\bar{t})$ sırasıyla t terim var olduğunda aynı anda C_i sınıfının olma ve t terim var olmadığı aynı anda C_i sınıfının olma koşullu olasılıklarını göstermektedir. Benzer şekilde $P(\bar{t})$ ve $P(t)$ ise t terimin geçme ve geçmeme olasılıklarını belirtmektedir.

2.2.2. Ki-kare (Chi-square-Chi2)

Yaygın ve popüler olan Ki-kare (Chi-square-Chi2) (Li & Chung, 2008) yaklaşımı, istatistiklere dayanan etkili bir öznitelik seçim aracıdır. Chi2, iki değişken arasındaki ilişkinin bağımlı mı yoksa bağımsız mı olduğunu belirler. Aslında Chi2 testi, istatistikteki iki bağımsız olayı analiz etmek için kullanılan bir tekniktir. Metin sınıflandırması için bağımsız olaylar, terimlerin ve sınıfların ortaya çıkmasıdır. Buna göre, Chi2 bilgisi aşağıdaki gibi hesaplanır:

$$CHI2(t, C) = \sum_{t \in \{0,1\}} \sum_{C \in \{0,1\}} \frac{(N_{t,C} - E_{t,C})^2}{E_{t,C}} \quad \textbf{(2.3)}$$

N ve *E* değerleri, sırasıyla *t* teriminin ve *C* sınıfının her durumu için gözlenen ve beklenen frekansı gösterir. Sonuç olarak Chi2 aşağıdaki gibi matematiksel olarak ifade edilir:

$$CHI2(t) = \sum_{i=1}^{M} P(C_i) * CHI2(t, C) \quad \textbf{(2.4)}$$

Formülde $P(C_i)$ sınıf olasılığını gösterir.

2.2.3. Gini katsayısı (Gini index-GI)

Gini katsayısı (Gini index-GI), Bilgi kazanımı ve kazanım oranı (gain ratio-GR) (Priyadarsini, Valarmathi, & Sivakumari, 2011) yaklaşımlarının eksikliklerine alternatif olarak sunulan bir metottur (Shang, ve diğerleri, 2007). İki yaklaşımdan farklı olarak

entropi kullanılmamaktadır. Yaklaşımın temel çalışma prensibi, m farklı sınıftan oluşan bir nesne kümesi sınıf sayısı kadar alt kümeye bölünebilir. Daha sonra her bir bölünen alt küme için gini katsayısı hesaplanır. Aynı şekilde tüm öznitelikler için de bu katsayı değeri hesaplanır. Daha sonra bu iki hesaplanan gini katsayısı kullanılarak her öznitelik için ayrı ayrı gini kazanım değerleri bulunur. Gini katsayısının az önce bahsedilen temel fikri üzerine geliştirilmiş versiyonları bulunmaktadır. Ancak genel gini katsayısı yaklaşımı, aşağıda matematiksel arka planı verilen halidir:

$$GI(t) = \sum_{i=1}^{M} P(t|C_i)^2 P(C_i|t)^2 \tag{2.5}$$

Formüldeki $P(t|C_i)$ gösterimi t terimin C_i sınıfında geçme olasılığını belirtmektedir.

2.2.4. Poisson dağılımından sapma (Deviation from poisson distribution-DP)

Poisson dağılımından sapma (deviation from Poisson distribution-DP) (Ogura, Amano, & Kondo, 2009) yöntemi, bilgi alma alanında etkili kelimeleri seçmek için yaygın olarak kullanılan bir yaklaşımdır ve Poisson dağılımından türetilmiştir. Poisson dağılımı, bilgisayar bilimi de dahil olmak üzere birçok mühendislik ve istatistik uygulamasında kullanılan bir dağılımdır. DP, metin sınıflandırması için öznitelik seçim problemlerine entegre edilmiş Poisson dağılımına dayalı bir yöntemdir. Nihayetinde, DP yönteminin matematiksel arka planı aşağıdaki gibidir:

$$\begin{gathered} DP(t,C) = \frac{(a-\hat{a})^2}{\hat{a}} + \frac{(b-\hat{b})^2}{\hat{b}} + \frac{(c-\hat{c})^2}{\hat{c}} + \frac{(d-\hat{d})^2}{\hat{d}} \\ \hat{a} = n(C)\{1-\exp(-\mu)\} \\ \hat{b} = n(C)\exp(-\mu) \\ \hat{c} = n(\overline{C})\{1-\exp(-\mu)\} \\ \hat{d} = n(\overline{C})\exp(-\mu) \\ \mu = \frac{F}{N} \end{gathered} \tag{2.6}$$

F ve N sırasıyla, tüm belgelerdeki t teriminin toplam sıklığını ve ayrılmış eğitim verisindeki doküman sayısını gösterir. $n(C)$ ve $n(\overline{C})$ sırasıyla, C sınıfına ait olan ve olmayan belge sayısını gösterir. DP yöntemi aşağıdaki gibi ifade edilir:

$$DP(t) = \sum_{i=1}^{M} P(C_i) * DP(t, C) \quad \textbf{(2.7)}$$

2.2.5. Ayırt edici öznitelik seçici (Distinguishing feature selector-DFS)

Metin sınıflandırmasında öznitelik seçimi için son zamanlarda kullanılan en etkili yöntemlerden biri DFS'dir. DFS yönteminin matematiksel arka planı aşağıda verilmiştir:

$$DFS(t) = \sum_{i=1}^{M} \frac{P(C_i|t)}{P(\bar{t}|C_i) + P(t|\overline{C_i}) + 1} \quad \textbf{(2.8)}$$

$P(t|\overline{C_i})$ ve $P(\bar{t}|C_i)$ sırasıyla t teriminin C_i sınıfı dışında diğer sınıflarda geçmesini ve C_i sınıfında geçmemeyi göstermektedir.

2.2.6. Normalleştirilmiş fark ölçüsü (Normalized Difference Measure-NDM)

Metin sınıflandırma için yeni bir öznitelik seçme yaklaşımı olarak sunulan normalleştirilmiş fark ölçüsü (NDM), NDM, dengeli doğruluk ölçüsüne (ACC2) (Forman, 2003) düzenleyici olarak minimum belge sıklığını sunar. Başka bir ifade ile NDM, ACC2 metodunun geliştirilmiş bir versiyonudur. Aşağıda matematiksel formülü verilmiştir:

$$NDM = \frac{|tpr - fpr|}{\min(tpr, fpr)} \quad \textbf{(2.9)}$$

Formülde tpr ve fpr sırasıyla $P(t|C_i)$ ve $P(t|\overline{C_i})$'yi gösterir. Eğer $\min(tpr, fpr)$ sıfıra eşit ise, o zaman payda 0.001 gibi çok küçük bir değer alır.

2.2.7. Max-min oranı (Max-min ratio-MMR)

MMR, bir terim bir sınıfta daha sıksa o terime yüksek bir puan atar, ancak terim birden fazla sınıfta aynı frekansa sahipse düşük bir puan atar. Buna ek olarak, büyük ve çok nadir verilerde NMD eksikliklerini tamamlar ve ilgili terime en yüksek değeri atar. MMR yönteminin matematiksel arka planı aşağıda verilmiştir:

$$MMR = \max(tpr, fpr) * \frac{|tpr - fpr|}{\min(tpr, fpr)} \quad \textbf{(2.10)}$$

tpr ve fpr sırasıyla $P(t|C_i)$ ve $P(t|\overline{C_i})$'yi gösterir. Eğer $\min(tpr, fpr)$ sıfıra eşit ise, o zaman payda 0.001 gibi çok küçük bir değer alır.

2.3. Sınıflandırma Metotları

En basit tanımıyla sınıflandırma kavramı, veri kümesinde daha önceden etiketlenmiş kategorilerden birine veri atama olarak tanımlanabilir. Sınıflandırmanın temel amacı, belirtilen veriler için hedef sınıfı doğru bir şekilde tahmin etmektir. Bu amacı gerçekleştirmek adına çeşitli sınıflandırma metotları geliştirilmiştir.

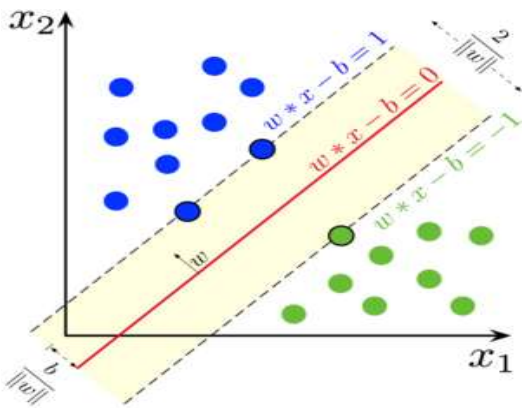
Bu tez kapsamında önerilen yöntemler öğrenme modeline bağlı değil, filtreye dayanan birer öznitelik seçim tekniğidir. Bu nedenle, seçilen özniteliklerin sınıflandırma doğruluğuna katkısını araştırmak için dört farklı sınıflandırıcı kullanılmıştır. Bu bölümde bu sınıflandırıcılar hakkında bilgi verilecektir.

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

SVM, literatürde etkili bir sınıflandırıcı olarak kabul edilir ve marj maksimizasyonu kavramına dayanır. Ayrıca, kullanılan çekirdek tipine bağlı olarak hem doğrusal hem de doğrusal olmayan versiyonlara sahiptir. Bu çalışmada SVM'in doğrusal versiyonu kullanılmıştır. SVM sınıflandırıcısının esas noktası marj kavramıdır (Joachims, 1998). Sınıflandırıcılar sınıfları ayırmak için hiper düzlemleri kullanır. Her hiper düzlem, yönü (w) ve uzaydaki tam konumu (w_0) ile karakterizedir. Böylece, doğrusal bir sınıflandırıcı basitçe şu şekilde tanımlanabilir:

$$w^T x + w_0 = 0. \tag{2.11}$$

Daha sonra iki sınıfa ayırım sağlayan $w^T x + w_0 = 1$ ve $w^T x + w_0 = -1$ hiper düzlemler arasındaki bölge marj olarak belirlenir. Şekil 2.5'da iki sınıftan (X_1 ve X_2) örneklerle eğitilmiş bir SVM için maks-marj hiper düzlem ve marjlar gösterimi verilmiştir.



Şekil 2.5. *Maks-marj hiper düzlem ve marjlar gösterimi*

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

$$J(w, w_0, \varepsilon) = \frac{1}{2}||w||^2 + K\sum_{i=1}^{N}\varepsilon_i \tag{2.12}$$

Buna göre;

$$\begin{gathered} w^T x + w_0 \geq 1 - \varepsilon_i \quad eğer \;\; x_i \in \; c_1 \\ w^T x + w_0 \leq -1 + \varepsilon_i \quad eğer \;\; x_i \in \; c_2 \\ \varepsilon_i \geq 0. \text{ şeklinde olur.} \end{gathered} \tag{2.13}$$

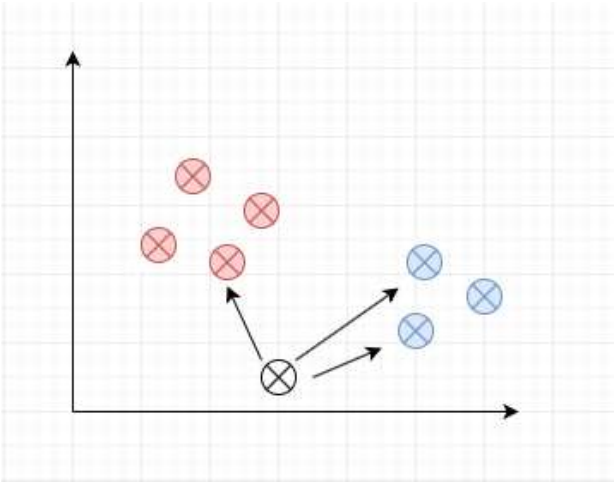
Burada K kullanıcı tanımlı bir sabittir ve ε ise marj hatasıdır. Eğer bir sınıfa ait veriler hiper düzlemin yanlış tarafında ise marj hatası oluşur. Bu nedenle maliyeti en aza indirmek, büyük bir marj ile az sayıda marj hatası arasında bir değişim ile ilgili bir konudur. Bu optimizasyon probleminin çözümü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

$$w = \sum_{i=1}^{N}\lambda_i\, y_i x_i \tag{2.14}$$

Burada sırasıyla, λ_i optimizasyon görevinin Lagrange çarpanı ve y_i ise bir sınıf etiketidir.

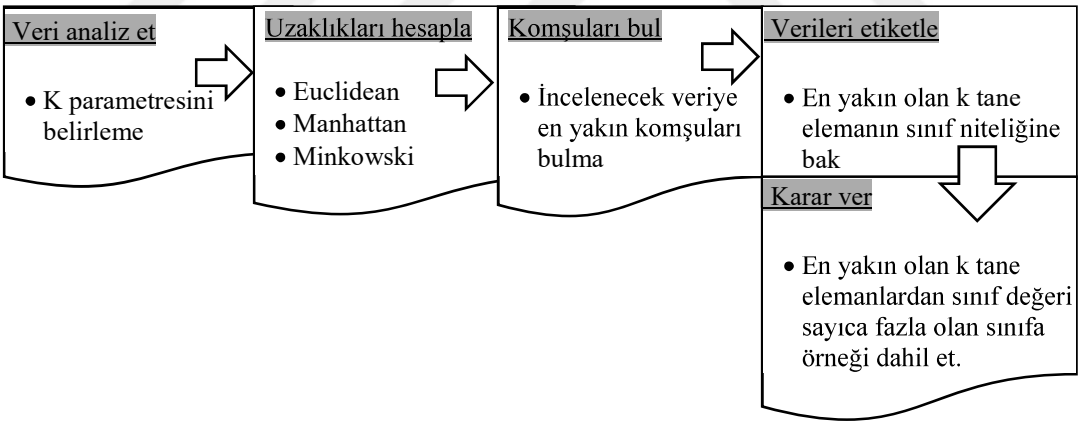
2.3.2. K-en yakın komşular (K-nearest neighbors-KNN)

K-en yakın komşular algoritması (KNN), sınıflandırma için kullanılan parametrik olmayan bir tekniktir (Kowsari, ve diğerleri, 2019). Bu metot birçok alanda olduğu gibi metin sınıflandırma alanında da yaygın olarak kullanılan bir tekniktir. Bu yaklaşımın çalışma prensibi, bir test belgesi x verildiğinde, eğitim setindeki tüm belgeler arasında x'in en yakın k komşusunu bulmak ve k adaylarının sınıfına göre kategori adaylarını puanlamaktır. x belgesinin ve her bir komşunun benzerliği, komşu belgelerin kategorisinin skoru olabilir. Şekil 2.6'de örnek bir KNN çizimi verilmiştir.



Şekil 2.6. *Örnek bir KNN çizimi*

Komşu hesaplama için Euclidean (Öklid), Manhattan ve Minkowski gibi uzaklık hesaplama yöntemlerinden birinden faydalanır. Yaklaşım belgenin komşu skorlarını hesapladıktan sonra bunlardan skoru en yüksek k tanesini alır. Aşağıda KNN yaklaşımının genel çalışma akışı Şekil 2.7'de verilmiştir.



Şekil 2.7. *KNN yaklaşımının genel çalışma akışı*

Minkowski uzaklık hesaplama fonksiyonu aşağıda verilmiştir:

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q} \tag{2.15}$$

Bu formüldeki q değeri eğer $q = 1$ ise Manhattan ve $q = 2$ ise Euclidean fonksiyonları elde edilir.

2.3.3. Karar ağacı (Decision tree-DT)

Bu tekniğin yapısı, veri alanının hiyerarşik bir ayrışması şeklinde gerçekleşir (Aggarwal & Zhai, 2012). Bir sınıflandırma görevi olarak karar ağacı D. Morgan (Magerman, 1995) tarafından tanıtılmış ve J.R. Quinlan (Quinlan, 1986) tarafından geliştirilmiştir. Ana fikir, kategorize edilmiş veri noktalarının niteliğine dayalı bir ağaç oluşturmaktır, ancak bir karar ağacının temel zorluğu, hangi niteliğin veya özniteliğin ebeveynler düzeyinde ve hangisinin çocuk düzeyinde olması gerektiğidir. Bu sorunu çözmek için De Mántaras (De Mántaras, 1991), ağaçta öznitelik seçimi için istatistiksel modelleme getirmiştir. p pozitiflik ve n negatiflik içeren bir eğitim seti için:

$$H\left(\frac{p}{n+p},\frac{n}{n+p}\right) = -(\frac{p}{n+p}log_2\frac{p}{n+p} + \frac{n}{n+p}log_2\frac{n}{n+p}) \quad \textbf{(2.16)}$$

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

$$BE(A) = \sum_{i=1}^{k}\frac{p_i + n_i}{p+n}H(\frac{p_i}{p_i+n_i},\frac{n_i}{p_i+n_i}) \quad \textbf{(2.17)}$$

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

$$A(BK) = H\left(\frac{p}{n+p},\frac{n}{n+p}\right) - BE(A) \quad \textbf{(2.18)}$$

Bilgi kazanımı en büyük olan öznitelik ebeveyn düğüm olarak seçilir.

2.3.4. Naive bayes (NB)

Naive Bayes sınıflandırcısı metin sınıflandırma alanında metinleri kategorize etmek için uzun zamandan beri yaygın olarak kullanılan bir yöntemdir. Yöntem teorik olarak Thomas Bayes tarafından formülüze edilen Bayes teoremine dayanmaktadır (Pearson, 1925). Bayes teoremi aşağıda verilmiştir:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \textbf{(2.19)}$$

Formülde sırasıyla $P(A|B)$, $P(B|A)$, $P(A)$ ve $P(B)$; B olayı gerçekleştiği durumda A olayının meydana gelme, A olayı gerçekleştiği durumda B olayının meydana gelme, A ve B olaylarının önsel olasılıklarıdır. Naive Bayes sınıflandırıcısının sırı Bayes teoremindeki durumların bağımsızlığında yatmaktadır. Yani öznitelikler bağımsızdır. Sınıflandırıcı her durumun olasılığını hesaplar ve olasılık değeri en yüksek olana göre sınıflandırır.

2.4. Kaba Kümeler Teorisi

Kaba Kümeler Teorisi (KKT) (Pawlak, 1998), (Zhang, Xie, & Wang, 2016) verideki gizli örüntüleri veya paternleri başarılı bir şekilde çıkaran, Pawlak tarafından önerilmiş bir matematiksel araçtır. KKT eksik, tutarsız ve belirsiz veri kümelerini düzenleyerek değerlendirme için uygun hale getirir. Çoğunlukla KKT yardımcı bir süreç olarak kullanılmasına rağmen, başka bir yaklaşım kullanmadan sınıflandırma, örüntü çıkarma, kural oluşturma, boyut indirgeme, öznitelik seçimi ve öznitelik çıkarma gibi işlemleri de tek başına gerçekleştirebilir (Cekik & Telceken, 2018), (Zhao, Wang, & Hu, 2016), (Zheng, Diao, & Shen, 2015). Kaba kümelerin temel kavramları ile ilgili kısa açıklamalar aşağıda verilmiştir.

Tablo 2.1. *Bir karar tablosu örneği*

$x \in U$	t_1	t_2	t_3	t_4	d
x_1	1	2	0	2	1
x_2	0	2	1	1	1
x_3	1	1	2	2	3
x_4	2	2	2	3	4
x_5	2	2	1	3	3
x_6	1	1	2	2	2
x_7	0	1	1	1	1
x_8	2	2	0	3	2

$S = (U, A, C)$ bir karar tablosunu veya bilgi sistemini göstermek üzere, burada $U = \{x_1, x_1, \dots, x_n\}$ nesnelerden oluşan evrensel kümeyi, A bir şartlı öznitelik kümesini ve C ise bir karar öznitelik kümesini göstermektedir. Herhangi bir $T \subseteq A$ şartlı öznitelik alt kümesi için gösterimi $IND(T)$ olan $T - ayırt\ edilmezlik\ ilişkisi$ aşağıdaki gibi tanımlanır:

$$IND(T) = \{(x_i, x_j) \in U^2 | \forall a \in T, a(x_i) = a(x_j)\} \quad \textbf{(2.20)}$$

Burada, $T - ayırt\ edilmezlik\ ilişkisinin$ denklik sınıfları $[x]_T$ olarak ifade edilir.

Alt (lower) ve üst (upper) küme yaklaşımları kaba kümelerde temel iki kavramı ifade etmektedir. Herhangi bir $X \subseteq U$ nesne alt kümesi ve belirtilen $T \subseteq A$ öznitelik alt kümesi üzerine X kümesinin $T - lower$ ve $T - upper$ yaklaşımlarının sırasıyla gösterimi $\underline{T}X$ ve $\overline{T}X$'dir. Ayrıca tanımlanmaları aşağıdaki gibidir:

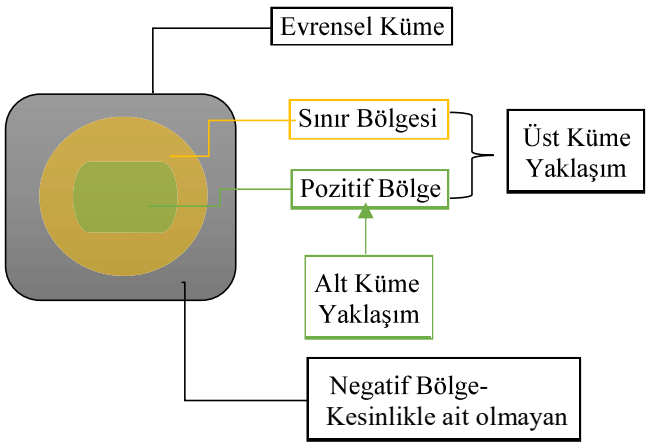
$$\begin{aligned} \underline{T}X &= \{x \mid [x]_T \subseteq X\}, \\ \overline{T}X &= \{x \mid [x]_T \cap X \neq \emptyset\} \end{aligned} \tag{2.21}$$

Bir kaba küme yaklaşım konsepti olarak bilinen ($\underline{T}X, \overline{T}X$) çifti bir kümenin kabaca belirlenip belirlenmeyeceğini söyler. Ayrıca bu çift kaba kümelerde bilinen bölgeleri de belirler. Eğer bir nesne $x \in \underline{T}X$ ise X kümesine ait olduğu bellidir. Fakat eğer $x \in \overline{T}X$ ise X kümesine ait olabilir, yani ait olduğu tam belli değildir. Örneğin, yukarıda Tablo 2.1 ile gösterilen karar tablosuna göre $T = \{t_1, t_4\}$ ve $X = \{x_1, x_2, x_5, x_7, x_8\}$ ise

$\underline{T}X = \{x_2, x_7\}$

$\overline{T}X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$ olur.

Burada $x_2 \; ve \; x_7$ elemanları X kümesine ait oldukları kesin belli iken diğer elemanlar ait olabilir, yani ait oldukları kesin değildir. ($\underline{T}X, \overline{T}X$) çiftine bağlı tanımlanan Pozitif (Positive-$POS_T X$), Negatif (Negative-$NEG_T X$) ve Sınır (Boundary-$BND_T X$) bölgelerin tanımı aşağıda ve temsili gösterimi de Şekil 2.8'de verilmiştir.



Şekil 2.8. *Kaba kümelerde bölgeler ve küme yaklaşımları gösterimi*

X ve T kümeleri üzerine bölgeler tanımlanması aşağıdaki gibidir:

$$POS_T X = \underline{T}X$$
$$NEG_T X = U - \overline{T}X \quad \textbf{(2.22)}$$
$$BND_T X = \overline{T}X - \underline{T}X$$

KKT'ye göre eğer sınır bölgesi boş küme değilse o zaman o küme kabaca belirlenir denir. Aksi durumda ise küme tam bellidir denilir. Örneğin, $T = \{t_1, t_4\}$ ve $X = \{x_1, x_2, x_5, x_7, x_8\}$ kümeleri için bölgeler şöyledir:

$POS_T X = \{x_2, x_7\}$

$NEG_T X = \emptyset$

$BND_T X = \{x_1, x_3, x_4, x_5, x_6, x_8\}$

3. İLGİLİ ÇALIŞMALAR VE GENEL MOTİVASYON

İnternet uygulamalarındaki teknolojik gelişmeler hızını artırarak gelişimini sürdürmektedir. Bu gelişim beraberinde yüksek sayıda kısa metin varlığını oluşturduğunu önceki bölümlerde belirtilmişti. İnternet teknolojileri geliştikçe ve insanların online işlem yapma alışkanlığı artıkça kısa metin sınıflandırmanın güncelliğini koruyacağı bir gerçektir. Bu yüzden kısa metin sınıflandırma alanında yapılmış birçok çalışma olmasına rağmen son yıllarda birçok çalışmanın yapılmış olması da bunun bir göstergesidir. Bu alanda yapılmış çalışmaların çoğunlukla odaklandıkları nokta seyreklik problemine ve özellikle de yüksek boyutluluk problemine çözüm sunmaktır. Bu alanda yapılmış çalışmaları motivasyonuna göre gruplandırmak, konuya daha geniş perspektiften bakmaya yardımcı olacaktır. Bu amaç doğrultusunda çalışmalar; duygu analizi yapma, yeni öğrenme yaklaşımı sunma, zayıflığı giderme ve gerçek zamanlı sınıflandırma yapma gibi kategorilere ayrılabilir (Song, Ye, Du, Huang, & Bie, 2014).

Duygu analizi yapan çalışmalarda, daha anlamlı bir yapı oluşturmak için metin içeriğindeki sözcüklerin iç yapı anlamsal düzeyine ve metinlerin korelasyonuna daha fazla dikkat edilir. Bu şekilde sözcüklerdeki eş-anlamlılık kavramı elimine edilir ve buna bağlı yüksek boyutluluk indirgenmiş olunur. Bu amaçla Kim ve ark. (Kim, ve diğerleri, 2014) tarafından dilbilgisi etiketleri ve sözcüksel veri tabanları kullanmadan kısa metinli belgeler arasındaki benzerliği etkili bir şekilde hesaplayabilen, dilden bağımsız semantik (LIS) çekirdek adı verilen yeni bir model önerilmiştir. Liang ve ark. (Liang, Xie, Rao, Lau, & Wang, 2018) etiketsiz kısa metinler üzerinden okuyucuların duygularını sınıflandırmak için evrensel etkili bir model (UAM) öneriyor. Geleneksel metin sınıflandırma modelinden farklı olarak, UAM yapısal olarak konu düzeyinde ve terim düzeyinde alt modellerden oluşur ve sosyal duyguları sosyal medyadaki okuyucular açısından algılar. Bu konuda yapılmış çalışma örnekleri bu şekilde çoğaltılabilir.

Kısa metin verileri yüksek oranda etiketsiz veriden oluşmaktadır. Bu yüzden eldeki az etiketlenmiş veri ile etiketlenmemiş veri birlikte kullanılarak daha etkili ve başarılı yeni öğrenme yaklaşımları (semi-supervised learning) sunulmuştur. Bu tür yaklaşımların amacı, etiketli ve etiketsiz verileri birlikte kullanarak aynı zamanda seyreklik probleminin öğrenme üzerindeki etkisini de azaltmaktır. (Dai & Le, 2015), (Duan, Luo, & Zeng, 2020) ve (Fu, ve diğerleri, 2019) bu alanda yapılmış çalışmalardan bir kaçıdır.

Özniteliklerin benzerliğinin sınıflandırılmasına dayalı tekli sınıflandırıcıların kısa metin alanında iyi bir sonuç elde etmeleri zordur. Çünkü seyreklik problemlerinden dolayı özniteliklerin benzerliklerini hesaplamak güçtür. Bu noktada bu zayıflığı gidermek için topluluk öğrenme algoritmaları (ensemble learning methods) devreye girmektedir. Bu tür metotlar her zayıf sınıflandırıcı için doğru ağırlığı hesaplayarak, her özniteliğin ağırlığını alır ve bu şekilde seyreklik problemine çözüm sunar. Örneğin, (Yan, Cao, & Li, 2009) çalışmasında yazarlar kısa metinlerdeki seyreklik probleminin ve dengesiz verilerin etkisini azaltmak için yeni bir dinamik yapı önermişler.

Kısa metinler anlık olma özeliğine sahip olduklarından sayıca çok yüksek bir orandadır. Bu denli yüksek sayıda verinin sınıflandırılması temel bir problemdir. Yüksek boyutluluktaki veriler aynı zamanda birçok gereksiz ve tutarsız veriyi de içermektedir. Bu gereksiz ve tutarsız veriler sınıflandırıcıların hem başarısını hem de performansını önemli ölçüde etkilemektedir. Bu yüzden bu tür verilerinin boyutlarının indirgenmesi gerekir. Boyut indirgemenin en etkili yollardan biri öznitelik seçmedir. Öznitelik seçme, bir öznitelik seçme algoritması kullanılarak var olan tüm öznitelik kümesini en iyi temsil eden ayırt ediciliği yüksek öznitelikleri seçme süreci olduğu daha önce belirtilmişti. Literatürde, bu amaçla öznitelik seçme yaklaşımı sunan birçok çalışma mevcuttur. Bu çalışmalardan bazıları IG, GR, GI, Chi2 (X^2), Karşılıklı Bilgi (Mutual Information-MI) (Sharmin, Shoyaib, Ali, Khan, & Chae, 2019) ve DP, vb. geleneksel yaklaşım örnekleri iken bazıları da DFS, NDM, MMR ve MRDC gibi son zamanlarda önerilmiş çalışma örnekleridir. Geleneksel yöntemler başarılarını kanıtlamıştır ve bu yöntemlerin çoğu metin sınıflandırması için de yaygın olarak kullanılmaktadır. IG özellikle veri ve metin madenciliği konularında sıkça kullanılır. Yaklaşım Shannon'ın bilgi teorisinden gelir ve temeli termodinamik konulara dayanır. Bir özniteliğin alabileceği farklı değerlerin sayısı yüksekse, IG yöntemi bu özniteliği ayırt ediciliği yüksek bir terim olarak seçerse sistemin ezberlemesine neden olabilir. Bu soruna bir çözüm olarak sunulan GR, her bir öznitelik için ayırma bilgilerini hesaplar ve elde edilen bilgi kazancını ayırma bilgilerine bölerek kazanç oranını hesaplar. GI, IG ve GR yöntemlerine alternatif olarak geliştirilen bir diğer özellik seçim yöntemidir. Bu yöntem, diğer iki yöntemin aksine entropi değerini kullanmaz. İstatistiğe dayanan Chi2 yöntemi, tüm özniteliklerin sınıflarına göre ki-karesini değerlendirir. MI, bir öznitelik ve bir sınıf etiketi arasındaki karşılıklı bağımlılığı hesaplamak için kullanılan bir yaklaşımdır. Bununla birlikte, nadir öznitelikleri tercih etme önyargısı ve olasılık tahminindeki hatalara duyarlılığı nedeniyle performansı diğer

yöntemlerden daha düşüktür. DP yöntemi, bilgi alma alanında etkili kelimeleri seçmek için yaygın olarak kullanılan bir yaklaşımdır. Bu yöntem daha sonra özellik seçim problemlerine uygulandı (Ogura, Amano, & Kondo, 2009).

Geleneksel yöntemlerin yanı sıra son zamanlarda önerilmiş birçok metodun da olduğunu belirtmiştik ve başarısını kanıtlamış olan DFS, bu yöntemlerden biridir. Terimin önemine göre her terim için 0,5 ile 1,0 arasında bir değer üretir. Benzer şekilde son zamanların çalışmalardan biri olan RDC, Rehman ve ark. (Rehman A. , Javed, Babri, & Saeed, 2015) tarafından metin sınıflandırma üzerine yeni bir öznitelik seçme yaklaşımı olarak sunulmuştur. Bu yönteme göre, yüksek ayırt edicilik özellik içeren terimler belirlenirken her terim için dikkate alınan terim sayısı, belge frekansları olarak alınır. Her terim için uygunluk değerini RDC yöntemini kullanarak hesaplayan MRDC yöntemi ise Labani ve ark. (Labani, Moradi, Ahmadizar, & Jalili, 2018) tarafından sunulmuştur. Yöntem, indirgenebilecek terimleri hesaplamak için Pearson korelasyonunu ve seçilen terim kümesini değerlendirmek için denetimli bir öğrenme algoritmasını kullanmaktadır. Yine Rehman ve ark. (Rehman, Javed, & Babri, 2017)'larının çalışmasında dengeli doğruluk ölçüsünü kullanarak metin sınıflandırması için yeni bir yöntem, NDM önerilmiş. Bu yaklaşım seyreklik oranı yüksek dengesiz veri kümelerinde etkili çalışmadığını yine aynı yazarlar tarafından belirtilmiş ve buna çözüm olarak MMR (Rehman A. , Javed, Babri, & Asim, 2018) adında yeni bir yaklaşım önerilmiştir. Wang ve Ming (Wang & Hong, 2019) metin sınıflandırması için yeni bir özellik seçme modeli, yani Hebb rule based feature selection (HRFS)'yi önermişlerdir. Model, sınıfı ve terimleri nöron olarak kabul eder ve ayırt ediciliği yüksek olan terimleri seçmeye çalışır.

Öznitelik seçme teknikleri genel olarak üç başlık altında ele alındığını daha önceki bölümlerde belirtmiştik. Yukarıda verilen metotlar birer filtre yaklaşım örnekleridir. Metin sınıflandırma için Linear Forward Search (LFS) (Gutlein, Frank, Hall, & Karwath, 2009), Span-Bound ve RW-Bound (Weston, ve diğerleri, 2001) yöntem birer sarmalayıcı yaklaşım örnekleri iken EGA (Ghareb, Bakar, & Hamdan, 2016), FSS (Bermejo, De la Ossa, G´amez, & Puerta, 2012), HybridBest ve HybridGreedy (Chou, Sinha, & Zhao, 2010) metotları ise birer gömülü yaklaşım örnekleridir. Burada dikkat edilmesi gereken nokta, metin sınıflandırma alanında sunulan filtre yaklaşım çalışma sayısı diğer yöntemlere göre çok daha fazladır. Bunun nedeni yüksek boyutlulukta bu tür yaklaşımların performans olarak daha hızlı çalışmasıdır. Sonuç olarak filtre başlığı

altında hala daha iyi ve etkili çözümler sunan çalışmalara büyük bir ihtiyaç vardır. Nitekim bu çalışmanın da temel motivasyonu, kısa metin sınıflandırma alanında etkili çalışabilen yeni filtre öznitelik seçme yaklaşımları sunmaktır.

Metin domaininde öznitelik seçme, öznitelik çıkarma gibi süreçlerde uygulamak üzere yapay sinir ağları, bulanık mantık, kaba kümler teorisi (KKT) vb. birçok farklı teknolojiler, teoriler veya yaklaşımlar kullanılmıştır. Bu yaklaşımlardan KKT bu çalışmada da kullanıldığından metin alanında bu teorinin kullanımı ile ilgili yapılmış çalışmalara yer verilmiştir. Örneğin, Chouchoulas ve Shen (Chouchoulas & Shen, 1999) metin sınıflandırmasındaki yüksek boyutluluğu hafifletmek için kaba kümeler kullanarak bir teknik önermişlerdir. Yaklaşımları, bu belgelerin ait olduğu sınıfları ayırt etmek için anahtar kelime vektörlerinin boyutluluğunu azaltan minimum bir koordinat anahtar kelime kümesi belirtir. Li ve ark. (Li, Shiu, Pal, & Liu, 2006) metin sınıflandırması için kaba kümelere dayanan yeni bir yaklaşım önermiştir. Önerilen yaklaşım; çıkarıcı, belge temsilcisi, kasiyer ve vaka alıcısı olmak üzere dört mekanizmadan oluşur. Önerilen yaklaşım, ilk olarak kaba küme yardımıyla belgedeki gereksiz terimlerinin sayısını azaltmaya çalışır. Daha sonra yaklaşım, belge seçimi için vaka muhakemesi kapsamı kullanılabilirliği kavramlarına dayanan yeni bir model kullanmaktadır. Qasem ve Ghufran (Al-Radaideh & Al-Qudah, 2017) verilerin yüksek boyutluluk etkisini azaltmak için kaba kümeler kullanılarak Arapça tweetler için duygu analizini içeren bir kısa metin sınıflandırma çalışması sunmuşlar. Miao ve ark. (Miao, Duan, Zhang, & Jiao, 2009) tarafından hem Rocchio hem de KNN sınıflandırıcılarının çalışmalarındaki güçlü yönlerini birleştirmek için değişken bir hassas kaba sete dayanan hibrit bir yaklaşım önerildi. Bu çalışmaların yanı sıra, kısa metin veya metin sınıflandırmasında KKT kullanan başka çalışmalar da vardır (Bekkali & Lachkar, 2019), (Gupta, Aha, & Moore, 2006), (La, Cao, & Xu, 2019), (Shi, Ma, Xi, Duan, & Zhao, 2011), (Singh & Dey, 2005).

Kaba kümeler elde yeteri veri olmadan eksik ve tutarsız veri üzerinde etkili çıkarım sağlayan matematiksel bir araç olduğundan birçok alanda kullanılmaktadır. Yukarı verilen çalışmalarda görüldüğü gibi kaba kümeler metin alanında veya domaininde de kullanılan bir yaklaşımdır. Ancak kaba kümeler bu alanda daha çok öznitelik çıkarma, öznitelik boyut indirgeme veya gereksiz özniteliklerin belirlenmesinde kullanılmıştır. Fakat literatür taramamıza göre metin veya kısa metin alanında kaba kümeler kullanılarak bir filtre öznitelik yaklaşımı sunulmamıştır. Ayrıca kaba kümelerin eksik ve tutarsız

veriden etkili çıkarım yapması kısa metin gibi eksik ve tutarsız verinin çok olduğu bir alanda etkili bir çıkarım sağlayacağı öngörüsü de bu çalışmanın diğer bir motivasyonunu oluşturmaktadır. Bu yüzden yukarıda bahsedilen kaba kümelerin metin veya kısa metin alanında kullanımı ile ilgili çalışmaların yanı sıra kaba kümelerin öznitelik seçme sürecinde kullanımı ile ilgili çalışmalardan da bahsetmek gerekir.

Temel olarak öznitelik seçme yaklaşımları öznitelikleri, bireysel (her özniteliği ayrı ayrı) veya alt küme (öznitelikleri birlikte) değerlendirme olarak iki farklı şekilde değerlendirir. Öznitelikleri bireysel değerlendirme olarak değerlendiren yöntemler genellikle daha az bilgi işlem maliyetine sahip olduğundan, öznitelikleri alt küme değerlendirme olarak değerlendiren yöntemlerden daha hızlı çalışırlar. Literatürde kaba kümeler kullanılarak sunulan öznitelik yöntemleri, öznitelikleri alt küme değerlendirme şeklinde ele alır. Bu yöntemler, olası tüm alt kümeleri kapsamlı bir şekilde oluşturmadan ileri öznitelik seçme mekanizmasından yararlanır veya ileri öznitelik seçim mekanizmasının aksine geriye doğru eliminasyon kullanır. Örneğin, Jensen ve Shen (Jensen & Shen, 2008) çalışmalarında kaba bir temel öznitelik seçme metriği, yani QuickReduct (QR) metodu önerilmiş. Algoritma boş bir kümeyle başlar ve öznitelikler tek tek eklenir. Her öznitelik eklendikten sonra, bağımlılık hesaplanır ve öznitelik bağımlılıkta maksimum bir artışla sonuçlanırsa o öznitelik kümeye dahil edilir. Algoritma, maksimum bağımlılık değeri elde edilene kadar devam eder. Herhangi bir aşamada, seçilen öznitelik kümesinin değeri tüm veri kümesinin değerine eşitse, algoritma seçili olan alt kümeyle "Reduct" olarak sonlanır. Öznitelik seçim sürecini optimize etmek amacıyla QR algoritması üzerine birçok geliştirilmiş yeni metotlar sunulmuş. Örneğin ReverseReduct (Jensen & Shen, 2008) metodu bunlardan biridir. Metrik geriye doğru eliminasyonu kullanır ve tüm öznitelik kümesini Reduct olarak değerlendirerek başlar. Ardından, her seferinde bir özniteliği kaldırır ve bağımlılığı hesaplar. Aslında kaba kümelerde bağımlılık ölçümü öznitelik seçimi için temel bir kriterdir. Bu yüzden var olan kaba kümeye dayalı öznitelik seçimi yaklaşımları onu optimize çalışmaktadır.

4. KISA METİN SINIFLANDIRMA

Çalışmanın başında belirtildiği gibi İnternetin hızlı gelişimiyle beraber e-ticaret, online iletişim, konum tabanlı hizmetler ve sosyal medya gibi platformlar ortaya çıkarmıştır. Bu tür platformlar da oldukça büyük sayıda elektronik dokümanların oluşmasına sebep olmuştur. Aşırı fazla sayıdaki bu dokümanların üzerine bir metin madenciliği uygulanması mecburiyeti açıktır. Çünkü insanlar her gün posta kutusuna gelen gereksiz maillerle muhatap olmak istemezler ya da sadece ilgilendikleri web sayfalarını görmek ister, alakasız olanları görmek istemezler. Bu durum metin verisinin içeriği içeren web sayfalarını (veya belgeleri) verimli bir şekilde endekslemek ve web sayfalarında aramayı kolaylaştırmak için konulara veya önceden tanımlanmış bazı kategorilere göre sınıflandırılması gerektiği anlamına gelir. Yani bir metin madenciliği olan metin sınıflandırma sürecine ihtiyaç vardır. Bu elektronik dokümanların çok büyük çoğunluğunu kısa metinler oluşturduğundan kısa metin sınıflandırmaya büyük bir ihtiyaç vardır.

Kısa metinler elektronik ortamda genel olarak uzunluğu 200 karakteri geçemeyen belgeleri veya dokümanları ifade etmektedir. Aşağıda bu dokümanların veya belgelerin bazı örnekleri verilmiştir:

- SMS mesajları
- Resim altyazıları
- Twitter mesajları (tweet)
- Forum gönderileri
- Ürün veya filim yorumları
- Blog ve haber yayınları
- Yazı başlıkları

Kısa metinlerin sınıflandırması normal metin sınıflandırmadan daha güçtür. Çünkü kısa metin dokümanların kaynağını oluşturan yapılar bir düzüne kelime veya birkaç cümle içerecek kadar kısa yazı, mesaj veya yorumdan ibarettir. Üstelik bu yorumlar veya yazıların çoğu eksik ya da anlamsız sözcük ve semboller içermektedir (Yan, Cao, & Li, 2009). Metinlerin uzunluğunun az olması metinler üzerinde çıkarım yapmayı ve iyi bir benzerlik ölçümü sağlamak için yeteri kadar bir içerik olanağı sunmamaktadır. Bu durum metinler üzerinde çalışan makine öğrenme veya sınıflandırıcılar gibi karar sitemlerinin performansını ciddi biçimde etkilemektedir. Kısa metin verileri ile ilgili problemleri

anlamak ve sunulan çözümleri analiz etmek için onları normal metinlerden daha zor kılan özelliklerine bakmak gerekir. Kısa metinler yapısal ve yapısal olmayan özelliklere sahiptir. Genel anlamda bu yapısal ve yapısal olmayan özellikler az kelime içerme (sparseness), gerçek zamanlı olmaya dayalı anlık olma (immediacy), standartsız olma, gereksiz ve dengesiz dağılımlı olma ve çok yüksek sayıda veri içermesine rağmen çok az sayıda etiketli veri içermesi gibi sıralanabilir (Song, Ye, Du, Huang, & Bie, 2014). Bu kavramların açıklamaları ve onlar hakkında gerekli bilgi ilgili alt başlıklarda verilmiştir.

Kısa metin verilerinin yukarıda bahsedilen özelliklerinden dolayı onlar ile ilgili yapılan çalışmalarda veya üzerinde çalışan karar sitemlerinin karşılaştıkları temel iki problem vardır:

- ✓ Seyreklik Problemi (Sparsity)
- ✓ Yüksek Boyutluluk Problemi (High Dimensionality)

Bu problemlerin tanımlaması ve açıklamaları ayrı bir başlık altında verilecektir.

Kısa metinlerin özellikleri göz önüne alındığında bu özelliklerin hangi probleme neden olduğu aşağıda Tablo 4.1'de verilmiştir.

Tablo 4.1. *Kısa metin verilerinin özellikleri ve ilgili problemle ilişkisi*

	az kelime içerme	anlık olma	standartsız olma	dengesiz dağılımlı olma	az etiketli çok sayıda olma
Seyreklik Problemi	√	√	√		
Yüksek Boyutluluk Problemi		√	√	√	√

Kısa metin alanında seyreklik ve yüksek boyutluluk problemlerine çözüm sunmak adına yapılmış birçok çalışma mevcuttur. Yüksek boyutluluk problemi normal metin sınıflandırma alanında da önemli bir yer tutuğundan özellikle bu probleme çözüm aramak için yoğun bir çalışma söz konusudur. Bu konuda yapılmış çalışmalar, 3. Bölümde ilgili başlık altında detaylı olarak verilmişti.

4.1. Kısa Metinlerin Yapısal ve Yapısal Olmayan Özellikleri

Kısa metinlerin karakteristik yapısına bağlı olarak sahip oldukları bazı özelikleri vardır. Bu özellikler hem yapısal hem de yapısal olmayan özellikler olabilir. Söz edilen

özellikler hakkında gerekli açıklamaya bölüm başında giriş yapılmıştı. Ancak detaylı açıklamalar ilgili alt başlıklarda verilmiştir.

4.1.1. Az kelime içerme

Az kelime içerme özelliği kısa metinlerin yapısal bir özelliğidir. Kısa metinler uzunlukları kısıtlı olduklarından çok az kelime içeren belgelerdir. Buna bağlı olarak yapıları seyrektir. Bu yapı aynı zamanda öznitelik vektör uzayının çok yüksek oranda boş olmasına neden olmaktadır. Nihayetinde bu yapı, bu tür veriler üzerinde etkili bir çıkarım yapmayı zorlaştırmaktadır. Ayrıca karar ve uzman sistemlerinin performansını da olumsuz yönde etkilemektedir.

4.1.2. Anlık olma

Anlık olma özelliği kısa metinlerin yapısı ile ilgili değil işleyişi ile ilgili bir özelliğidir. Kısa metinler hemen gönderilir ve gerçek zamanlı olarak alınır. Bu durum kısa metinlerin eksik, yetersiz bilgi içermesine neden olmaktadır. Hatta kısa metin olmasının temel sebeplerinden biri olarak da görülebilir. Aynı zamanda kısa metin sayısının çok fazla olmasına da etki yapan bir durumdur. Kısacası kısa metinlerin anlık olması onları hem seyrek yapıda olmasına hem de çok fazla sayıda veri topluluğu haline getirmektedir.

4.1.3. Standartsız olma

Standartsız olma özelliği kısa metinlerin yine yapısı ile ilgili bir özelliktir. Kısa metinlerin içeriğinin yazılması ile ilgili herhangi bir standarttın olmayışı bu metinlerde birçok yazım hatasına, standart olmayan terimlerin kullanılmasına veya eksik ve anlamsız sözcüklerle bulunmasına neden olmaktadır. Belirli bir standartta olmayan bu içeriklerin düzeltilmesi beraberinde ya seyrekliği ya da yüksek boyutluluğu getirmektedir. Eğer uygun olmayan bu veriler silinirse seyrekliğin artmasına, şayet olduğu gibi alırsa da öznitelik vektör uzayının büyümesine neden olacaktır.

4.1.4. Dengesiz dağılımlı olma

Bu özellik kısa metinlerin yapısal olmayan bir özelliğidir. Tamamıyla kısa metinlerin işleyişi ile ilgili bir durumdur. Kısa metinlerin devasa sayıda olmaları onları büyük ölçekli veriler arasında sadece küçük bir parçaya odaklanabiliriz. Bu yüzden kısa metinlerin kullanım örnekleri sınırlıdır ve dağılımı dengesizdir.

4.1.5. Az etiketli çok sayıda olma

Az etiketli çok sayıda olma özelliği kısa metinlerin yapısı ile ilgili değil sayısı ile ilgili bir durumdur. Tüm büyük ölçekli örnekleri manuel olarak etiketlemek zordur. Sınırlı etiketli örnekler yalnızca sınırlı bilgi sağlayabilir. Dolayısıyla, bu etiketli örneklerin ve diğer etiketlenmemiş örneklerin tam olarak nasıl kullanılacağı kısa metin sınıflandırmasında önemli bir sorun haline gelmiştir.

4.2. Kısa Metin Problemleri

Kısa metinlerin yapısal ve yapısal olmayan özellikleri, bu veriler sınıflandırılırken veya kategorize edilirken birtakım problemlere neden olmaktadır. Bu problemlerin en çok bilinenleri ve en yaygın olanları seyreklik ve yüksek boyutluluktur.

4.2.1. Seyreklik problemi

Kısa metinler çok az kelime içerdiklerinden seyrek yapıdalar. Ayrıca kısa metin belgelerinin içeriğinin yazılması herhangi bir yazım standartta bağlı olmadığında yazım hataları ve anlamsız kelimelerin varlığı da söz konusudur. Sonuç olarak kısa metinlerin yapısı ve işleyiş durumuna bağlı olarak kısa metin verileri çok yüksek oranda boş verilerdir. Verideki bu oran seyrekliğe işaret eder ve seyreklik problemi olarak da tanımlanır.

Seyreklik problemini daha somut olarak ifade etmek gerekirse eğer kısa metin belgesinde geçen her kelime bir öznitelik olarak alınırsa, öznitelik uzayı kelime sayısı kadar olur. Bu uzayın çok yüksek derecede olduğu bilindiğine göre her özniteliğin geçtiği doküman sayısı vektör uzayında boşluğu belirleyecektir. Örneğin, yüzbinlerce dokümandan oluşan ve bu dokümanların içeriğinde sadece yazı başlıklarının olduğu bir veri kümesi düşünüldüğünde, yüzbinlerce dokümanda geçen her bir kelime veya sözcüğün bir öznitelik olarak alındığı öznitelik uzayının çok büyük olacağı açıktır. Bir kelimenin çok az içerikli yüzbinlerce dokümandan çok çok azında geçeceğinden o kelimenin öznitelik vektör uzayı yüksek oranda boş olacaktır. Bu yüzden dokümanların satırda özniteliklerin sütunda olduğu bir matris çok küçük oranda dolu olacaktır.

4.2.2. Yüksek boyutluluk problemi

Kısa metinler de normal metinler gibi her bir kelimeyi veya sözcüğü bir öznitelik olarak alabilir. Bu alanda dokümanlarda geçen kelimelerden öznitelikler elde edildiğinden kelime sayısı kadar bir öznitelik uzayı oluşabilir. Çok çok büyük doküman

sayısının olduğu kısa metin alanında çok yüksek derecede öznitelik uzayı ile karşılaşmak kaçınılmazdır.

5. ÖNERİLEN METOT: ORANTILI KABA ÖZNİTELİK SEÇİCİ (PROPORTIONAL ROUGH FEATURE SELECTOR- PRFS)[3]

5.1. Motivasyon

Seyreklik problemi, kısa metinlerin az sayıda, yeteri kadar kelime içermemesinden kaynaklanan bir problem türü olduğu daha önce belirtilmişti. Çıkarım yapmak için elde yeteri kadar veri olmayan bu tür problemlere KKT ile etkili ve başarılı bir çözüm sunulabilir. KKT, bir bilgi sistemindeki özniteliklerin karar niteliğine olan bağlılığını ve her özniteliğin karar niteliği üzerindeki karakteristiğini başarılı bir şekilde ortaya çıkarmaktadır. Ayrıca, KKT bilgi sistemindeki gizli ve anlamlı bilgileri ortaya çıkardığı gibi bilgi sistemini tutarsız hale getiren tekrarlı ve anlamsız bilgiyi de ortaya çıkarmada oldukça başarılıdır. Kaba kümelerin bu başarısı ve özelliği bu alanda kullanmamızın temel nedenini oluşturmaktadır. KKT'nin Bulanık Mantık gibi bilinen yaklaşımlardan farklı olarak öncesinden hiçbir üyelik fonksiyonuna veya bilgiye ihtiyaç duymaması da bu alanda kullanılmasının başka bir nedenidir. Bunların yanı sıra, var olan birçok filtre öznitelik seçme yaklaşımı özniteliğin değer kümesini göz ardı etmektedir. Ancak bu değer kümesi birçok karar sistemi için önemli bir bilgidir. Yine kaba kümeler yardımı ile bu değer kümesi öznitelik seçme sürecinde dikkate alınmasını sağlanmak onu kullanımının diğer bir nedenidir.

5.2. PRFS'nin Çalışma Stratejisi

Bu çalışmada Orantılı Kaba Öznitelik Seçici (Proportional Rough Feature Selector-PRFS) adında bir filtre öznitelik seçme metodu önerilmiştir. Filtre öznitelik seçim yaklaşımlarında her terim için bir skor hesaplanır ve skor değeri en yüksek *n* tane öznitelik seçilir. İdeal bir filtre öznitelik seçme metodunda, ayırt ediciliği yüksek olan özniteliklere yüksek, ayırt ediciliği düşük olan özniteliklere ise düşük skor verilmesi beklenir. Bir terimin ayırt ediciliği ve skoru bazı durumların varlığına bağlıdır. Bu durumları açıklamadan önce bazı ifadelerin altını çizmek gerekebilir. Bilindiği gibi KKT'de alt ve üst küme yaklaşımları diye iki kavram vardır. Bu alt ve üst küme yaklaşımları ile bir dokumanın pozitif, sınır veya negatif bölgede olduğu tespit edilebiliyor. Buna bağlı olarak dokümanlardan oluşan bir kümenin;

- Tam veya kabaca belirlenip belirlenemeyeceği

[3] Bu bölüm *Expert Systems with Applications* adlı uluslararası dergide "A Novel Filter Feature Selection Method Using Rough Set for Short Text Data" başlığı ile kabul almıştır.

- Ya da belirsiz olduğu tespit edilebilir.

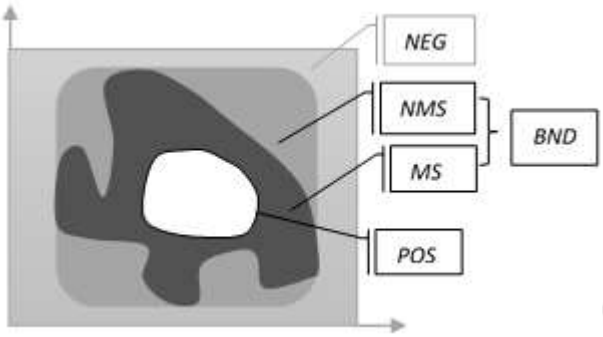
Bu bilgiler ışığında X, dokümanlardan oluşan bir sonlu küme olmak üzere bir terimin ayırt ediciliği ve skoru aşağıdaki durumlara bağlıdır:

1. Eğer bir terime göre X kümesinin eleman sayısı belirli çoğunlukta ve tüm elemanları pozitif bölgede kalıyorsa, X kümesinin elemanlarının hangi sınıfa ait olduğu tam bellidir ve X kümesi tam belirlenir denir. Dolayısıyla bu terimin ayırt ediciliği yüksektir ve skoru yüksek olması beklenir.
2. Eğer bir terime göre X kümesinin eleman sayısı belirli çoğunlukta ve elemanlarının çoğunluğu pozitif bölgede ve çok az miktarı da sınır bölgesinde kalıyorsa, X kümesi kabaca belirlenir denir. Dolayısıyla bu terimin ayırt ediciliği yüksektir ve skoru *nispeten* yüksek olması beklenir.
3. Eğer bir terime göre X kümesinin eleman sayısı belirli çoğunlukta ve elemanlarının bir kısmı pozitif bölgede ve bir kısmı da sınır bölgesinde kalıyorsa, X kümesi kabaca belirlenir denir. Dolayısıyla bu terimin ayırt ediciliği biraz daha düşüktür ve skoru *normal* olması beklenir.
4. Eğer bir terime göre X kümesinin eleman sayısı çok az sayıda ve elemanlarının tamamı pozitif bölgede kalıyorsa, X kümesi tam belirlenir denir. Ancak terimin ayırt ediciliği düşüktür ve skoru düşük olması beklenir.
5. Eğer bir terime göre X kümesinin eleman sayısı çok az sayıda ve elemanlarının çoğunluğu pozitif bölgede ve çok az miktarı da sınır bölgesinde kalıyorsa, X kümesi kabaca belirlenir denir. Ancak bu terimin ayırt ediciliği düşüktür ve skoru *nispetten* düşük olması beklenir.
6. Eğer bir terime göre X kümesinin eleman sayısı belirli çoğunlukta ve elemanlarının tamamı sınır bölgesinde kalıyorsa, X kümesi belirsizdir denir. Dolayısıyla bu terimin ayırt ediciliği yüksek veya düşük olabilir.

Yukarıdaki maddelere dikkat edildiğinde sınır bölgesindeki dokümanların belirsizliğe yol açtığı ve ayırt ediciliği düşürdüğü anlaşılmaktadır. Dolayısıyla, sınır bölgesindeki dokumanlar ile ilgili birtakım işlemlerin yapılması gerektiği açıktır. Bu çalışmada, bu probleme çözüm olarak BND bölgesi, Üye Bölüm (Member Section-MS) ve Üye Olmayan Bölüm (Non-Member Section-NMS) adında iki kısma ayırarak bir çözüm sunulmuştur:

- ✓ **MS:** Sınır bölgesindeki sınıfa ait dokumanlar kümesini
- ✓ **NMS:** Sınır bölgesindeki sınıfa ait olmayan dokumanlar kümesini göstermektedir.

Şekil 5.1'de bölgeler ve MS ve NMS bölümleri gösterimi verilmiştir.



Şekil 5.1. *Bölgeler ve MS ve NMS bölümleri gösterimi*

Sınır bölgesinin MS ve NMS olarak ayırımından sonra 6. madde yerine aşağıdaki maddeler eklenebilir:

6.1 Eğer bir terime göre X kümesinin eleman sayısı belirli çoğunlukta ve elemanlarının tamamı sınır bölgesinde kalıyorsa ve elemanlarının büyük çoğunluğu MS'ye aitse bu terimin ayırt ediciliği yüksektir ve skoru *yüksek* olması beklenir.

6.2 Eğer bir terimin göre X kümesinin eleman sayısı belirli çoğunlukta ve elemanlarının tamamı sınır bölgesinde kalıyorsa ve elemanlarının büyük çoğunluğu NMS'ye aitse bu terimin ayırt ediciliği çok düşüktür ve skoru *en düşük* olması beklenir.

PRFS; GI, DFS gibi geleneksel yöntemlerden farklı olarak sadece terimin dokümanda geçip geçmemesine değil onun değer kümesine bakarak değerlendirir. Örneğin, ağırlıklandırılmış bir veri kümesinde DFS gibi geleneksel yöntemler ağırlıkları göz ardı edebilir ancak PRFS bu ağırlıklara göre sınıflandırma yaparak değerlendirme yapmaktadır. Yöntem, bir terimin değer kümesindeki her farklı değer için birer küme oluşturur ve KKT kullanılarak bu kümeler üzerinde bölgesel ayırım yapmaktadır. Daha sonra her terim için bu bölgesel ayırımlar kullanılarak elde edilen değerlerin orantılanmasıyla bir skor üretmektedir.

Diyelim ki U tüm dokümanları içeren evrensel kümeyi, $X = \{x \mid x \in C_i\} \subset U$ ve C_i sınıf i'yi belirtmekte olsun. Buna göre $X = X_1 \cup X_0$ olmak üzere X_1, t adındaki terimin geçtiği dokümanlar kümesi, X_0 ise geçmediği dokümanlar kümesini gösterse X_1kümesi üzerinde aşağıdaki değerler hesaplanır:

$$\begin{aligned} &Lower = \underline{T}X_1, \qquad ve\ Upper = \bar{T}X_1 \\ &BND = Upper - Lower\ ve\ NEG = U - Upper \\ &MS = \{x|\ x \in BND\ \cap\ C_i\} \\ &NMS = \{x|x \in BND\ and\ x \notin C_i\}\ \ ve\ x\ \in\ X_1 \end{aligned} \tag{5.1}$$

PRFS'nin çalışmasının temelinde pozitif ve sınır bölgesindeki dokümanların ayrı ayrı değerlendirilmesi vardır. Sınır bölgesindeki dokümanlar α adında bir kat sayı ile çarpılmakta ve bu katsayının değeri aşağıdaki gibi hesaplanmaktadır:

$$\alpha: \mathrm{U} \rightarrow (0, 1]\ ve\ \alpha = \frac{abs(|MS| - |NMS|) + 1}{|MS| + |NMS| + 1} \tag{5.2}$$

Formülde $|MS|$ ve $|NMS|$ sırasıyla MS ve NMS kümelerinin eleman sayılarını göstermekte ve *abs* ise mutlak değeri ifade etmektedir.
α değeri sınır bölgesindeki dokümanlar için bir cezalandırma değeri olarak da düşünülebilir. MS ve NMS kümelerinin eleman sayıları birbirine yakın ise α değeri küçülür, tersi durumda ise büyür. Aslında α değeri bize terimin ayırt ediciliği hakkında bir ön bilgi sunmaktadır. Değerinin yüksek olması ya terim çoğunlukla tek sınıfta geçiyor veya tersi durumdur. Değerinin küçük olması ise tüm sınıflarda yakın sayıda geçtiğini belirtir.

Daha sonra X_0 kümesi üzerinde ise seyreklik pozisyonu (Sparsity Position-SP_0) hesaplanır:

$$SP_0 = [x]_{t=0} - [x]_{t=0}\ \cap\ X_0 \tag{5.3}$$

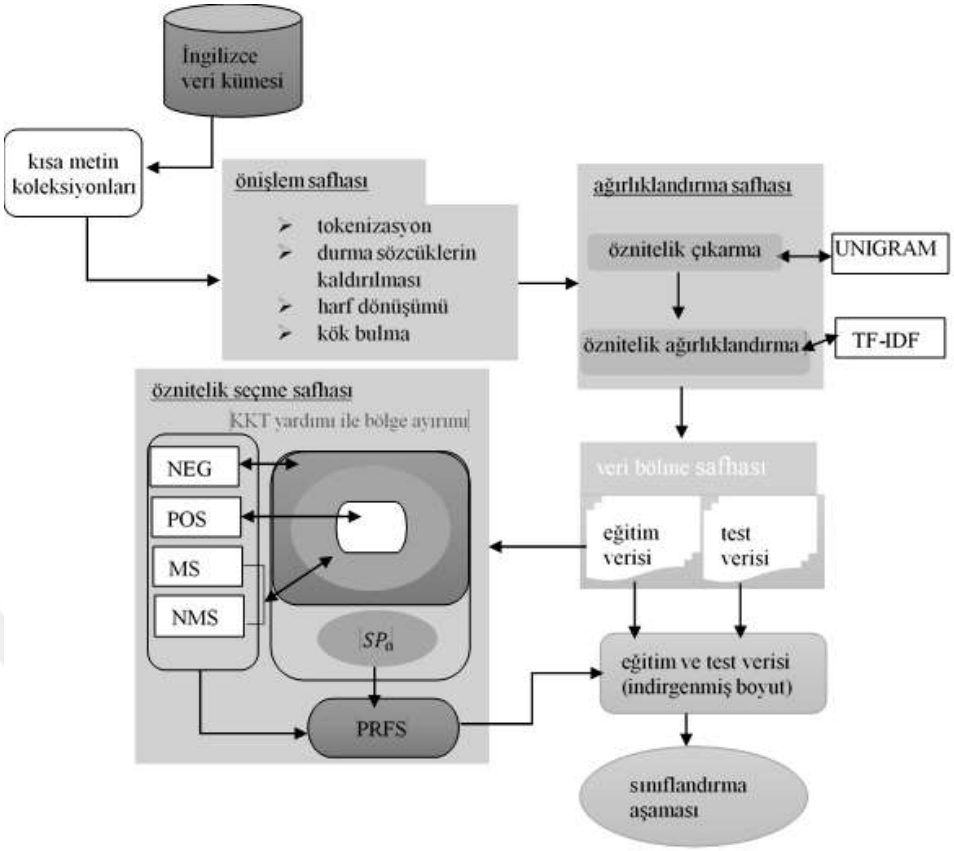
Formülde daha önce bahsedildiği gibi X_0, verilen C_i sınıfında t teriminin geçmediği doküman kümesini göstermektedir. $[x]_{t=0}$ ise *t-ayırt edilmezlik ilişkisi*'nin denklik sınıflarını göstermektedir. Yani $[x]_{t=0}$ t teriminin geçmediği dokümanlar kümesini belirtir. Örneğin, Tablo 5.1'de C_1 sınıfı verildiğinde köpek terimi için, $[x]_{t=köpek}$ kümesi

{D_1, D_4, D_5, D_6} kümesine denk ve X_0 ise {D_1} kümesine eşit olacaktır. Buna göre bu örnek için SP_0, {D_1, D_4, D_5, D_6} ve {D_1} kümelerinin küme farkına eşit olacak, yani SP_0 = {D_4, D_5, D_6}. Aynı örnekte, *MS* = {D_2} ve *NMS* = {D_3} olur. Nihayetinde gerekli hesaplamalardan sonra önerilen metot ile bir özniteliğin önem değeri ise aşağıda verilen formül ile hesaplanır:

$$PRFS(t) = \sum_{i=1}^{M} \frac{|Lower| + \alpha * |\mathrm{MS}|}{|NMS|/(|SP_0| + |\mathrm{NEG}| + 1) + 1} \quad \textbf{(5.4)}$$

Bu formülde m sınıf sayısını ve $|SP_0|$, SP_0 kümesinin eleman sayısını göstermektedir. Aynı durum $|\mathrm{NEG}|$ (Bkz. Eşitlik 5.1) gösterimi için de geçerlidir.

PRFS'nin nasıl çalıştığını göstermek için basit bir doküman koleksiyonu Tablo 5.1'de verilmiştir. Örneğin Tablo 5.1'de C_1 sınıfı verildiğinde köpek terimi için $\mathrm{NEG} = \emptyset$ iken aynı terim için C_3 sınıfı verildiğinde {D_2, D_3} kümesine denktir. $|Lower|$ pozitif bölgedeki dokümanları temsil etmektedir. Daha önce belirtildiği gibi pozitif bölge sadece tek sınıfta geçen dokümanları ifade etmektedir (Bkz. Eşitlik 2.21 ve 2.22). Aynı örnekte C_1 sınıfı verildiğinde köpek terimi için $Lower = \emptyset$ iken C_3 sınıfı verildiğinde ise $Lower$ {D_5, D_6} kümesine denktir. Özetle öznitelikler bu ağırlık değerine göre sıralanarak top-n tane öznitelik seçilir. Seçilen özniteliklere göre hem eğitim hem de test veri hazırlanır. Sonuç olarak artık eğitim ve test veri hazır olduktan sonra geriye onları bir sınıflandırıcı ile sınıflandırmak kalıyor.



Şekil 5.2. *PRFS metodun mimarisi*

Örnek:

Tablo 5.1'de PRFS'nin nasıl çalıştığını göstermek için basit örnek bir doküman koleksiyonu verilmiştir. Ayrıca bu tabloya göre önerilen metot ile her bir terimin skorunun hesaplanması da aşağıda verilmiştir.

Tablo 5.1. *Basit bir doküman koleksiyon örneği*

Doküman ismi	İçerik	Sınıf
D_1	kedi	C_1
D_2	kedi köpek	C_1
D_3	kedi köpek fare	C_2
D_4	kedi fare	C_2
D_5	kedi balık	C_3
D_6	kedi balık fare	C_3

$$PRFS(kadi) = \frac{0 + 2 * 3/7}{4/1 + 1} + \frac{0 + 2 * 3/7}{4/1 + 1} + \frac{0 + 2 * 3/7}{4/1 + 1} = 0.5143$$

$$PRFS(köpek) = \frac{0 + 1 * 1/3}{1/4 + 1} + \frac{0 + 1 * 1/3}{1/4 + 1} + \frac{0 + 0 * 1/1}{0/5 + 1} = 0.5333$$

$$PRFS(fare) = \frac{0 + 0 * 1/1}{0/5 + 1} + \frac{0 + 2 * 2/4}{1/4 + 1} + \frac{0 + 1 * 2/4}{2/3 + 1} = 1.1000$$

$$PRFS(balık) = \frac{0 + 0 * 1/1}{0/5 + 1} + \frac{0 + 0 * 1/1}{0/5 + 1} + \frac{2 + 0 * 1/1}{0/5 + 1} = 2.0000$$

Bu basit örnekte en yüksek skor "*balık"* terimi için hesaplanmıştır. Çünkü "*balık*" teriminin ayırt ediciliği 1 nolu madde ile belirtilen durumdur. Yani "*balık"* terimine göre dokümanların hangi sınıfa ait olduğu tam bellidir (sadece C_3 sınıfına aittir) ve tüm dokümanlar pozitif bölgededir. Ancak diğer terimlerin ayırt ediciliği ise 6 nolu madde ile belirtilen durumdur, yani belirsizdir ve dolayısıyla tüm dokümanlar sınır bölgesinde kalmaktadır. Fakat bu terimlerin sokurunun ne olacağına karar vermek için 6 nolu madde yerine eklenen maddelere bakmak gerekir. "*kedi"* terimin ayırt ediciliği 6.2 nolu madde ile belirtilen durumdur ve skorunun düşük olması beklenir aynı şekilde bu durum "*köpek"* terimi için de geçerlidir. *"fare"* terimin ayırt ediciliği ise 6.1 nolu madde ile belirtilen durumdur. Dolayısıyla ayırt ediciliği yüksek olması beklenir. Bu anlatılanları daha net anlamak için hesaplanan değerleri incelemek gerekir. Buna göre hesaplanan *"kedi"*, *"köpek"* ve *"fare"* skorlarına bakıldığında, *"fare"* skorunun *"kedi"* ve *"köpek"* skorlarından daha yüksek olduğu görülmektedir.

5.3. Deneysel Çalışmalar

Bu bölümde deneysel çalışmaların sonuçlarını göstereceğiz. İlk olarak, kullanılan veri kümeleri hakkında kısa bir bilgi verdikten sonra, Makro-F1 başarı ölçüsünü açıklıyoruz. Daha sonra PRFS tarafından seçilen terimlerin sınıflandırıcıların başarısı üzerindeki etkisini değerlendiriyoruz. Bu amaçla, PRFS'nin performansı önceki bölümlerde açıklanan öznitelik seçim yöntemleriyle karşılaştırılmıştır. Karşılık gelen öznitelik seçim algoritmaları tarafından seçilen öznitelikleri kullanan sınıflandırıcıların performansı bu bölümde gösterilmiştir. Ayrıca bu bölüme, PRFS'nin performans gelişiminin diğer yöntemlere göre istatistiksel olarak anlamlı olup olmadığını göstermek için tablo olarak bazı istatistiksel analizler dahil edilmiştir.

5.3.1. Veri kümeleri

Bu çalışmada, öznitelik seçme yöntemlerinin performansını değerlendirmek için dört farklı veri kümesi kullanılmıştır. İlk istihdam edilen veri kümesi, 425 spam ve 450

spam olmayan Kısa Mesaj Servisi (Short Message Service-SMS) metin mesajından oluşan SMS veri kümesidir (Nuruzzaman, Lee, & Choi, 2011), (Uysal A. K., Gunal, Ergin, & Gunal, 2013). Etiketli Duygu Cümleleri (Sentiment Labeled Sentences-SLS) olarak adlandırılan ikinci veri kümesi; film, restoran ve ürünler hakkındaki görüşlerden elde edilen olumlu ve olumsuz cümlelerin bir koleksiyonudur. Görüşler yelp.com, amazon.com ve imdb.com web sayfalarından elde edilmiştir (Kotzias, Denil, De Freitas, & Smyth, 2015). YouTube Spam Koleksiyonu olarak adlandırılan üçüncü veri kümesi, çeşitli YouTube videolarının spam ve spam olmayan incelemelerini içeren beş benzer veri kümesinin (Psy, KatyPerry, LMFAO, Eminem, Shakira) birleşimidir. Bu veri kümesi veri toplama döneminde en çok görüntülenen 10 video arasından beş YouTube videosunun incelemelerinden elde edilen 1.956 gerçek mesajdan oluşur (Alberto, Lochter, & Almeida, 2015). Çalışmada, oluşturulan veri kümesi Tüm YouTube Spam Koleksiyonu (All YouTube Spam Collection-AYSC) olarak adlandırılmıştır. Son veri kümesi Sentiment140 veri kümesinin bir alt kümesidir (Go, Bhayani, & Huang, 2009). Kullanılan veri kümeleri hakkında bilgi Tablo 5.2'de sunulmaktadır.

Tablo 5.2. *Deneylerde kullanılan veri kümelerinin karakteristiği*

Veri kümesi	Sınıf etiketi	Eğitim verisi	Test verisi	Öznitelik boyutu
SMS	Legitimate	315	135	1825
	Spam	297	128	1825
SLS	Pos	1000	500	3078
	Neg	1000	500	3078
AYSC	Legitimate	665	286	2380
	Spam	703	302	2380
Sentiments	Pos	13646	8876	10000
	Neg	13148	6124	10000

5.3.2. Başarı kriteri: Makro-F1

Önerilen metodun başarısını test etmek için çalışmada literatürde en iyi bilinen F1-ölçüsü başarı ölçütlerinden biri olan Makro-F1 kullanılmıştır. Makro ortalamasında, F-ölçüsü veri kümesindeki her sınıf için hesaplanır ve daha sonra tüm sınıfların ortalaması alınır. Bu nedenle, sınıf frekansından bağımsız olarak her sınıfa eşit ağırlık verilir. Makro-F1 matematiksel arka planı verilmeden önce değinilmesi gereken iki metrik var: Kesinlik (Precision) ve Duyarlılık (Recall).

Kesinlik metriği, pozitif olarak tahmin edilen değerlerin kaç tanesinin gerçekte pozitif olduğunu gösteren bir metriktir. Bu metrik değeri özellikle yanlış pozitif tahminlerin maliyeti yüksek olduğu durumlarda çok önem arz etmektedir. Duyarlılık metriği ise pozitif olarak tahmin edilmesi gereken durumların ne kadarının pozitif olarak tahmin edildiğini gösteren bir değerdir. Bu değer yanlış negatif durumların tahmin edilme maliyetinin yüksek olduğu durumlarda önem taşımaktadır. Buna göre, Makro-F1 şu şekilde hesaplanabilir:

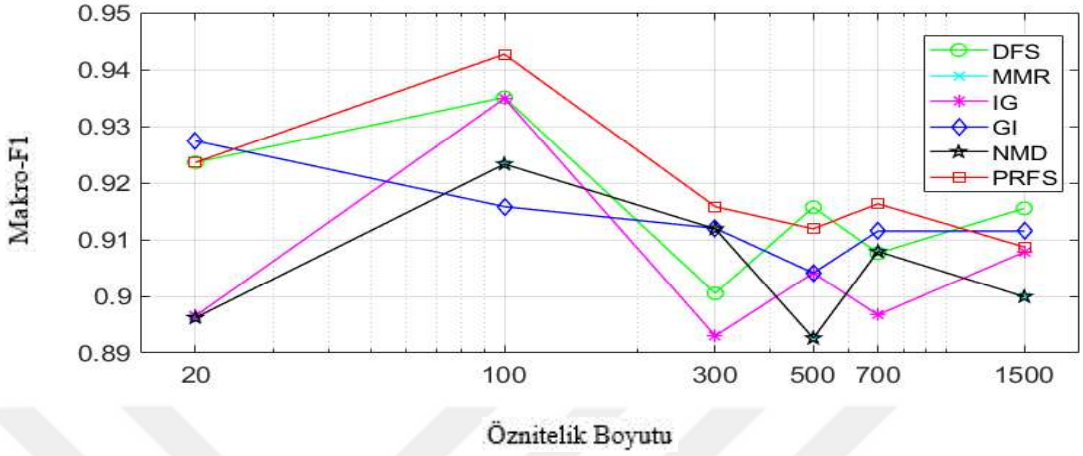
$$p_i = \frac{1}{C} * \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} TP_i + FP_i}, \qquad r_i = \frac{1}{C} * \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} TP_i + FN_i} \tag{5.5}$$

$$Makro - F1 = \frac{\sum_{i=1}^{C} F_i}{C}, F_i = 2 * \frac{p_i * r_i}{p_i + r_i} \tag{5.6}$$

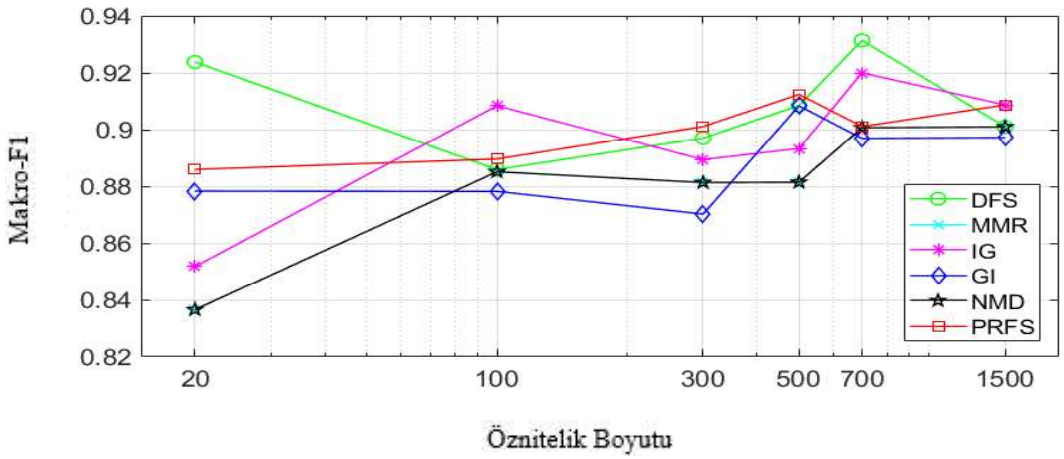
Formülde (p_i, r_i) ikilisi sırasıyla sınıf i için kesinlik ve duyarlılığa denk gelmektedir.

5.3.3. Başarı analizi

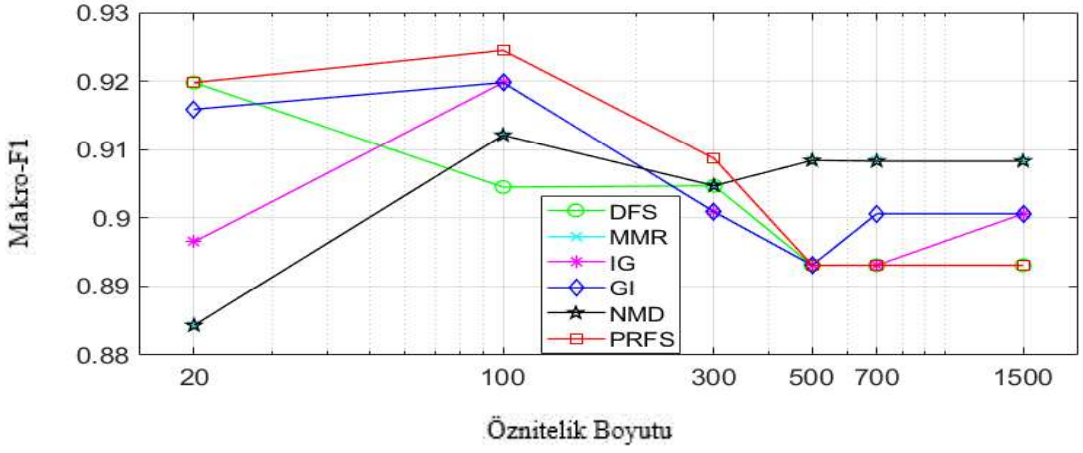
Öznitelik seçme yöntemlerinin performansını test etmek için çeşitli öznitelik boyutları kullanılır. Bu çalışmada her bir seçim yöntemi ile seçilen belirli sayıdaki yüksek dereceli 20, 100, 300, 500, 700 ve 1500 öznitelik boyutları kullanılarak SVM, KNN, DT ve BN sınıflandırıcıları ile test edilmiştir. Deneylerde elde edilen Makro-F1 skorları, her bir veri kümesi için Şekil 5.3-5.6'da listelenmiştir. En yüksek değerler göz önüne alındığında, PRFS karşılaştırma için kullanılan tüm diğer öznitelik seçim yöntemlerinden daha başarılıdır. Örneğin, SLS veri kümesinde, PRFS tarafından yüksek dereceli 1500 boyut hariç diğer tüm boyutlarda seçilen öznitelikler SVM, DT ve NB sınıflandırıcıları kullanıldığında en yüksek Makro-F1 skorlarını vermektedir. Benzer şekilde SMS veri kümesinde, PRFS tarafından seçilen öznitelik yüksek dereceli 100 boyutta SVM ve DT ve 300 boyutta NB kullanıldığında sağlanan en yüksek Makro-F1 skorlarıdır. Ayrıca, aynı veri kümesinde en yüksek değerlere boyut 100, 300 ve 700'de SVM ve boyut 300, 500 ve 1500'de KNN, boyut 20, 100 ve 300'de DT ve boyut 1500 hariç diğer tüm boyutlarda NB kullanıldığında ulaşılmıştır. Başka bir örnek olarak, Şekil 5.6'da PRFS'nin SVM ve NB sınıflandırıcılarını kullanarak çoğunlukla en yüksek Makro-F1 skorlarını ürettiğini gözlemledik.


0.95
0.94
0.93
0.92
0.91
0.9
0.89
Makro-F1
20
100
300
500
700
1500
Öznitelik Boyutu
DFS
MMR
IG
GI
NMD
PRFS

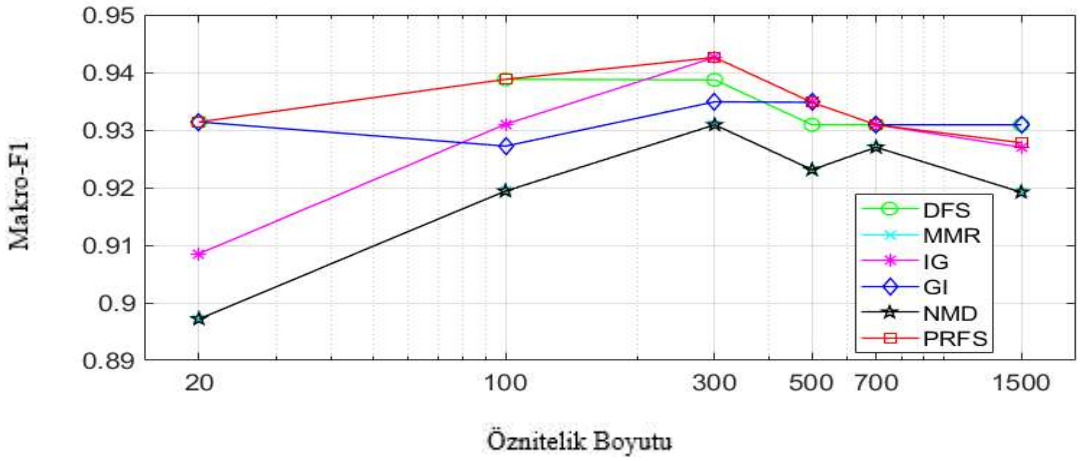
(a)


0.94
0.92
0.9
0.88
0.86
0.84
0.82
Makro-F1
20
100
300
500
700
1500
Öznitelik Boyutu
DFS
MMR
IG
GI
NMD
PRFS

(b)

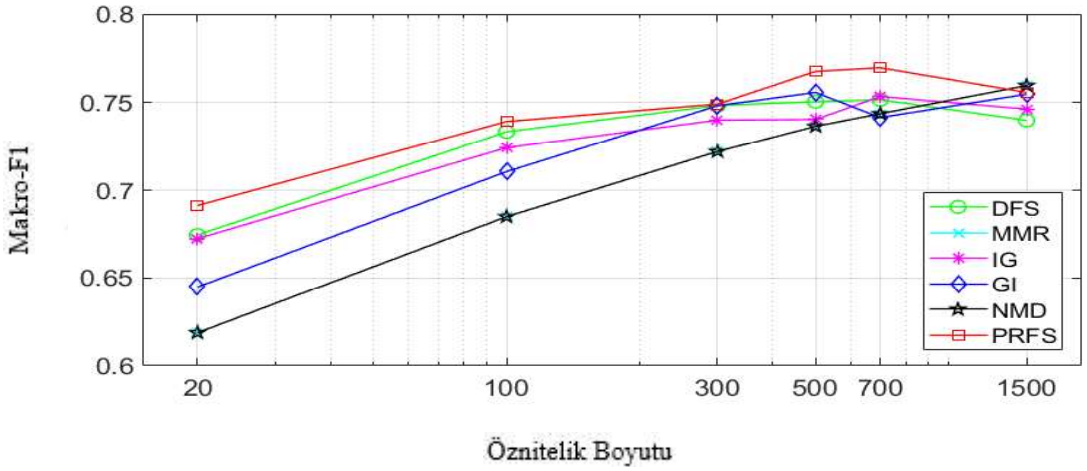

0.93
0.92
0.91
0.9
0.89
0.88
Makro-F1
20
100
300
500
700
1500
Öznitelik Boyutu
DFS
MMR
IG
GI
NMD
PRFS

(c)

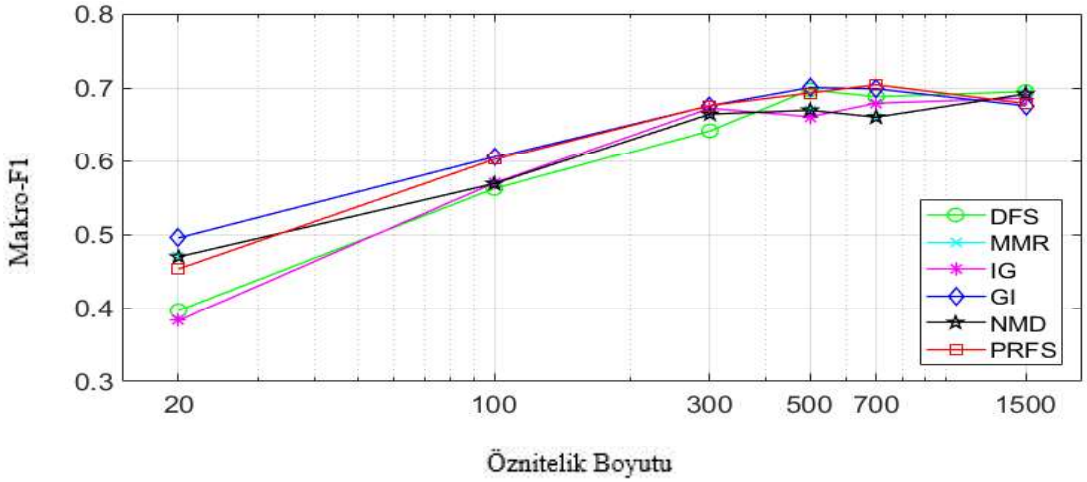


(d)

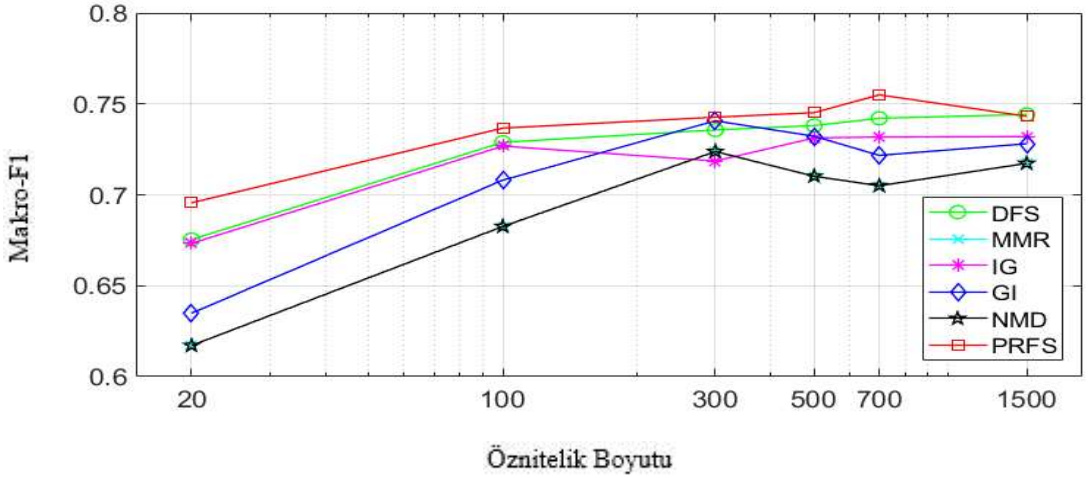
Şekil 5.3. SMS veri kümesi için SVM (a), KNN (b), DT (c) ve NB (d) sınıflandırıcısı kullanıldığında Makro-F1 skorları



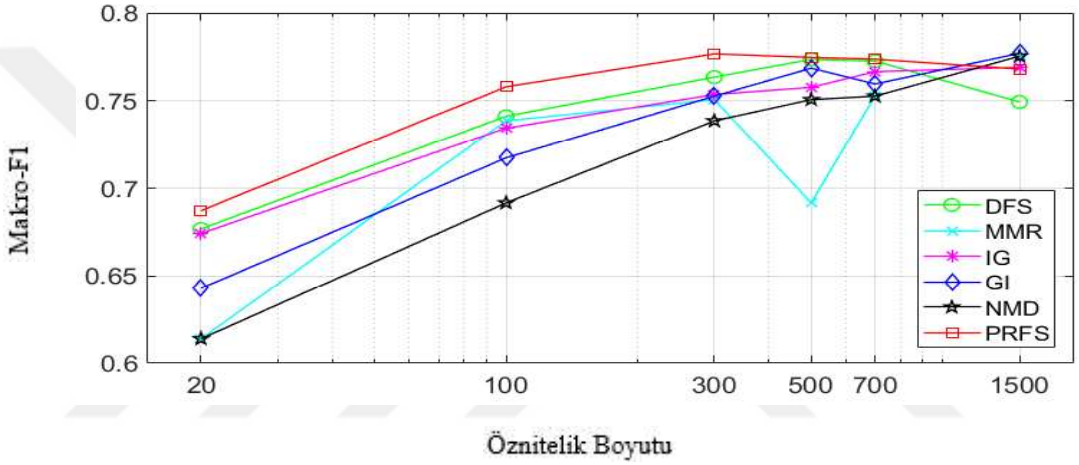
(a)



(b)

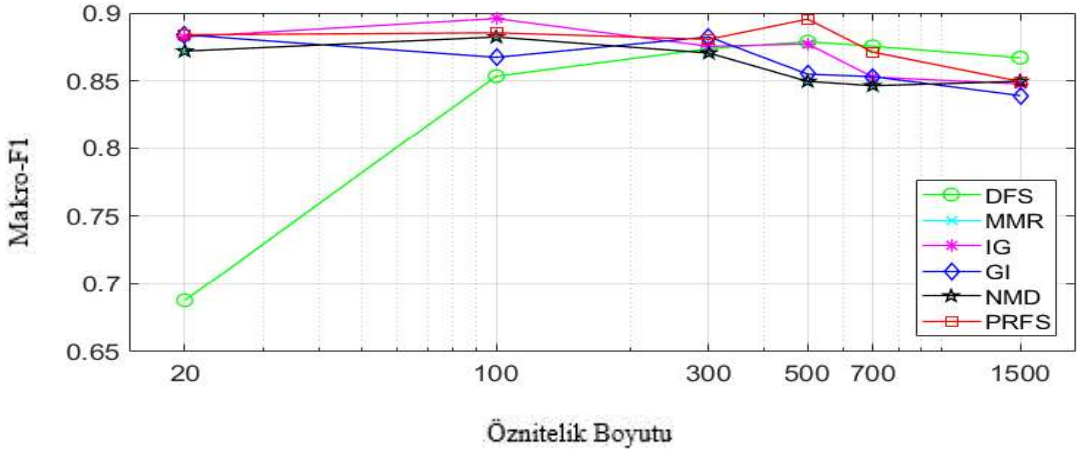


(c)

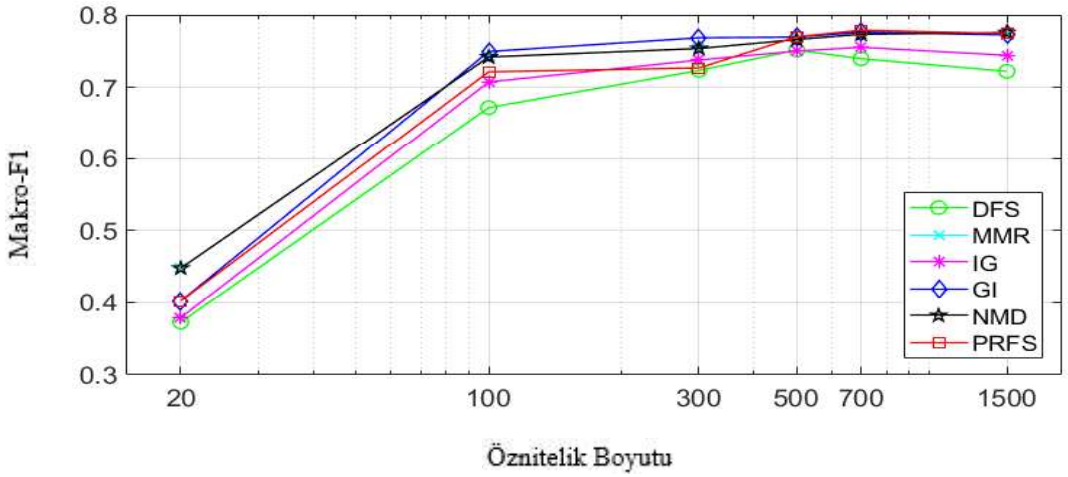


(d)

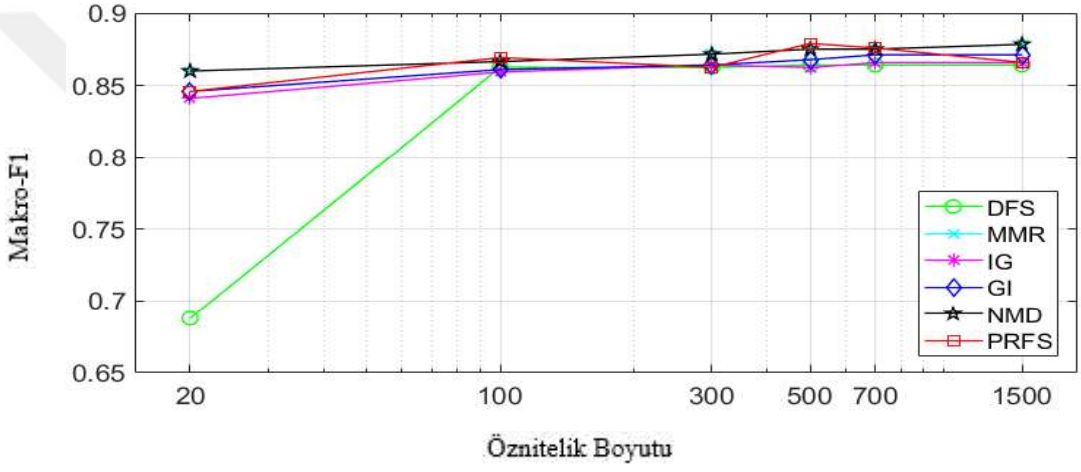
Şekil 5.4. ***SLS*** *veri kümesi için SVM (a), KNN (b), DT (c) ve NB (d) sınıflandırıcısı kullanıldığında Makro-F1 skorları*



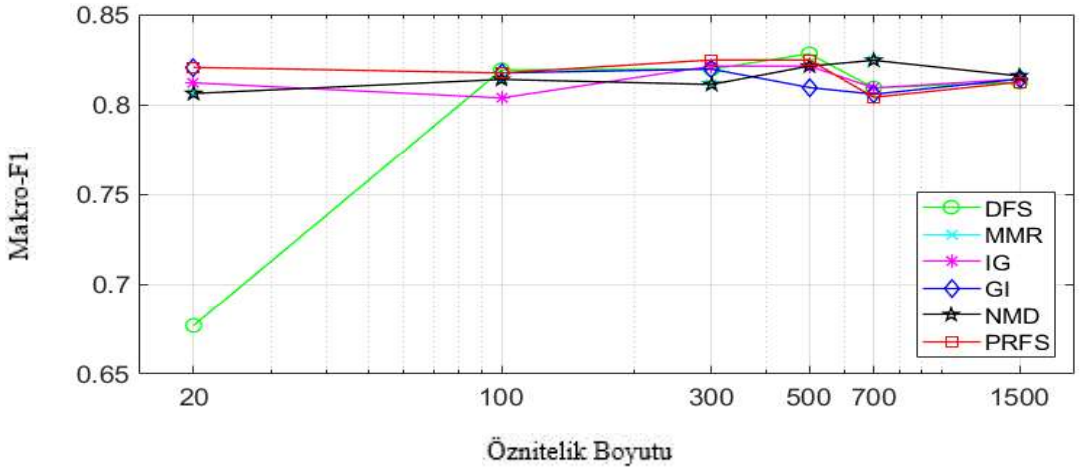
(a)



(b)

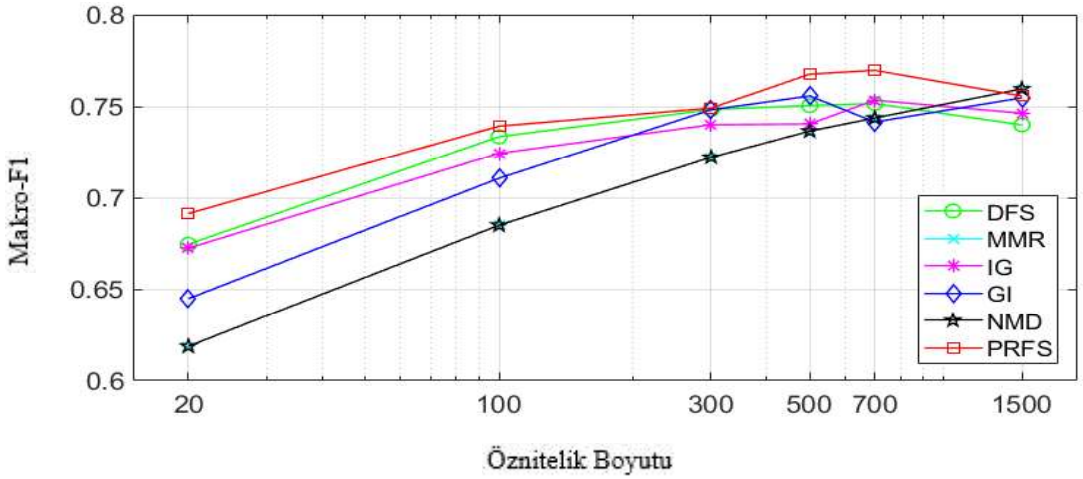


(c)

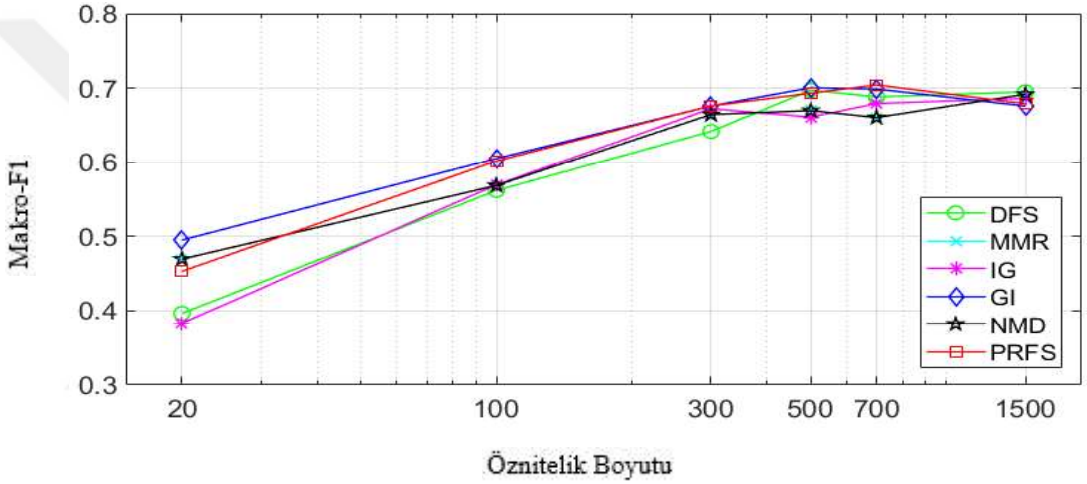


(d)

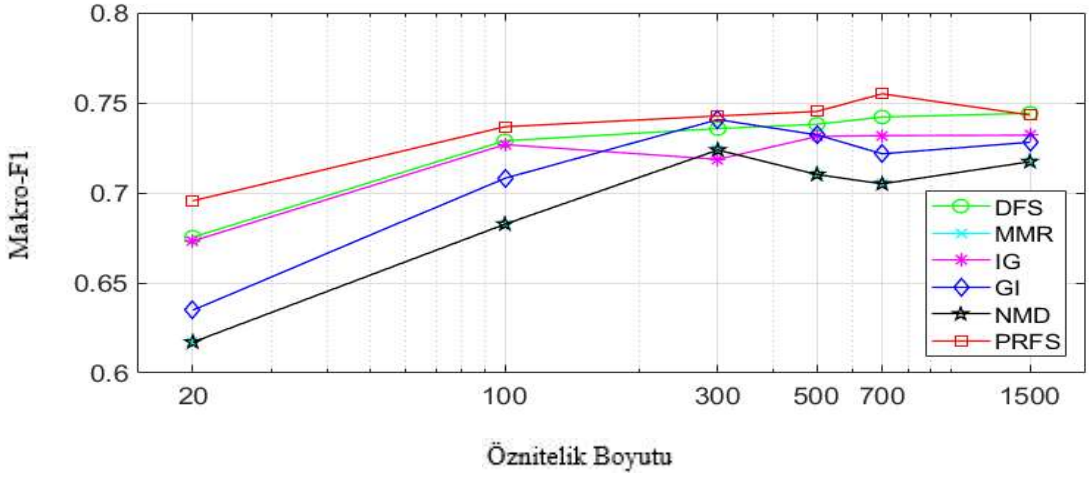
Şekil 5.5. ***AYSC*** *veri kümesi için SVM (a), KNN (b), DT (c) ve NB (d) sınıflandırıcısı kullanıldığında Makro-F1 skorları*


0.8
0.75
0.7
0.65
0.6
Makro-F1
20
100
300
500
700
1500
Öznitelik Boyutu
DFS
MMR
IG
GI
NMD
PRFS

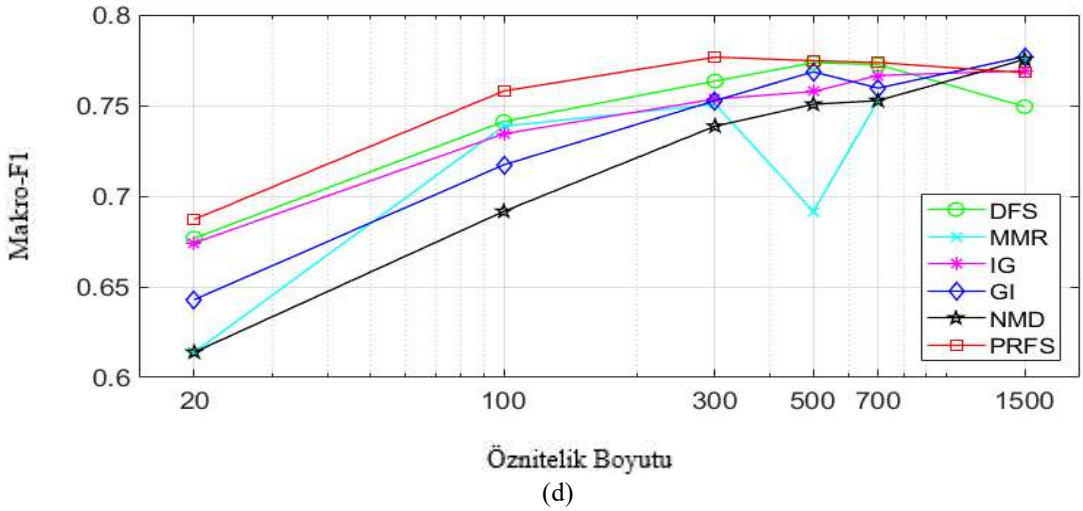
(a)


0.8
0.7
0.6
0.5
0.4
0.3
Makro-F1
20
100
300
500
700
1500
Öznitelik Boyutu
DFS
MMR
IG
GI
NMD
PRFS

(b)


0.8
0.75
0.7
0.65
0.6
Makro-F1
20
100
300
500
700
1500
Öznitelik Boyutu
DFS
MMR
IG
GI
NMD
PRFS

(c)



(d)

Şekil 5.6. ***Sentiments*** *veri kümesi için SVM (a), KNN (b), DT (c) ve NB (d) sınıflandırıcısı kullanıldığında Makro-F1 skorları*

Tablo 5.3, her bir sınıflandırıcı için Şekil 5.3-5.6'da daha önce sunulan Makro-F1 skorlarının ortalamasını göstermektedir. Burada amaç, her sınıflandırıcı için tüm veri kümelerinde çeşitli öznitelik boyutlarında, öznitelik seçim yöntemlerinin genel performansını göstermektir. Başka bir deyişle, Tablo 5.3 bize her yöntemin her veri kümesinde nasıl performans gösterdiğini göstermektedir. Bu nedenle aynı öznitelik boyutu için seçme yöntemi bazı veri kümelerinde iyi sonuçlar verirken diğer veri kümelerinde çok kötü sonuçlar verirse yöntemin performansının istatistiksel olarak iyi olduğu söylenemez. Benzer şekilde, bazı veri kümelerinde en iyi sonucu verirken ancak diğer veri kümelerinde ikinciyse yöntemin performansı istatistiksel olarak iyidir. Bu tablo bize bu durumu göstermektedir. Örneğin, AYSC veri kümesinde IG boyut 100'de SVM kullanıldığında en yüksek Makro-F1 skorlar elde edilirken, PRFS ikinci sıradadır. Bu nedenle Tablo 5.3 incelendiğinde, PRFS tüm veri kümeleri için SVM sınıflandırıcısını kullanıldığında en yüksek ortalama Makro-F1 skorlarını sunar. Benzer şekilde, aynı durum NB sınıflandırıcısı kullanıldığında geçerlidir. Ayrıca PRFS, SMS ve SLS veri kümeleri için DT sınıflandırıcısı kullanıldığında en yüksek ortalama Makro-F1 skorlarını sağlar. Kısacası bu veriler PRFS'nin ortalama Makro-F1 skorları açısından diğerlerinden daha iyi performans göstermediği durumlarda da ikinci sırada yer aldığını göstermektedir.

Tablo 5.3. *İlgili veri kümelerinde SVM, KNN, DT ve NB sınıflandırıcıları kullanıldığında tüm öznitelik boyutları için ortalama Makro-F1 skorları*

Sınıflandırıcı	Veri kümesi	Makro-F1					
		DFS	MMR	IG	GI	NDM	PRFS
SVM	SMS	91.63	90.53	90.55	91.37	90.53	**91.98**
	Sentiments	65.06	63.78	65.46	64.89	63.86	**65.54**
	AYSC	83.94	86.17	87.19	86.35	86.17	**87.77**
	SLS	73.30	71.09	72.95	72.59	71.09	**74.54**
KNN	SMS	**90.79**	88.10	89.53	88.82	88.10	89.98
	Sentiments	53.78	52.59	**54.09**	53.13	52.59	53.60
	AYSC	66.35	**70.97**	67.89	70.63	**70.97**	69.56
	SLS	**61.34**	62.08	60.86	**64.19**	62.08	63.46
DT	SMS	90.14	90.43	90.06	90.52	90.43	**90.54**
	Sentiments	62.62	60.98	**63.43**	61.86	60.99	62.23
	AYSC	83.42	**87.11**	85.98	86.36	**87.11**	86.64
	SLS	72.74	69.27	71.90	71.10	69.27	**73.64**
NB	SMS	93.36	91.95	92.91	93.17	91.95	**93.44**
	Sentiments	64.66	64.07	65.35	64.54	64.07	**65.41**
	AYSC	79.43	81.56	81.37	81.46	81.56	**81.74**
	SLS	74.62	72.05	74.26	73.64	72.05	**75.64**

Verilen grafikleri (Bkz. Şekil 5.3-5.6) analiz etmenin en önemli adımlarından biri öznitelik seçme yöntemleriyle elde edilen tepe (pik) değerlerini göstermektir. Bu amaç için Tablo 5.4 oluşturulmuştur. Tablo 5.4'deki satırlar, karşılık gelen sınıflandırıcıyı kullanarak veri kümelerinde pik değerlerini sağlayan öznitelik seçim yöntemlerini göstermektedir. Buna göre Tablo 5.4 dikkate alındığında, PRFS en iyisi olduğu görünmektedir.

Tablo 5.4. *Farklı deneysel setler için en yüksek Makro-F1 skorunu (tepe değeri-pik noktası) sağlayan öznitelik seçme yöntemleri*

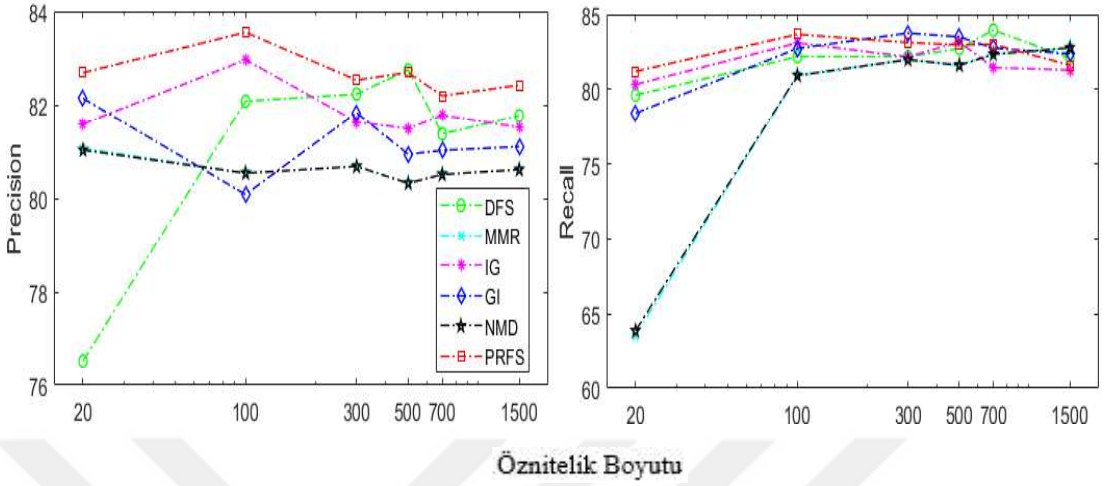
	Makro-F1			
Veri kümesi	SVM	KNN	DT	NB
SMS	PRFS	DFS	PRFS	PRFS, IG
Sentiments	PRFS	IG	IG	PRFS
AYSC	IG	PRFS	PRFS	DFS
SLS	PRFS	PRFS	PRFS	GI

Tablo 5.5 veri kümesini dikkate almayan ilgili sınıflandırıcı için öznitelik seçim yöntemlerinin ortalama performansını göstermektedir. Bu nedenle, her bir öznitelik seçim yönteminin farklı veri kümelerinde elde ettiği skorların ortalaması bu sunumda alınmıştır. Amaç, her öznitelik boyutu için yöntemlerin performansını istatistiksel olarak analiz etmektir. Başka bir deyişle bu deneysel çalışma, yöntemlerin tüm veri kümelerindeki her öznitelik boyutu için sınıflandırıcılar ile nasıl performans gösterdiğini görmek için yapılmıştır. Bu yolla ilgili sınıflandırıcılarla daha iyi sonuçlar veren yöntemleri daha net bir şekilde gözlemleyebiliriz. Örneğin, Tablo 5.5'deki PRFS için 76.04 skoru, SVM sınıflandırıcısı kullanıldığında boyut 20'de tüm veri kümelerinde Makro-F1 skorlarının ortalaması alınarak elde edilmiştir. Tablo 5.5 incelendiğinde, PRFS'nin neredeyse tüm kullanılan öznitelik boyutlarında diğer yaklaşımlardan daha iyi sonuçlar verdiği belirtilebilir. Sonuçlara göre, önerilen PRFS yönteminin tüm veri kümelerindeki farklı öznitelik boyutlarında kararlı olduğu söylenilebilir.

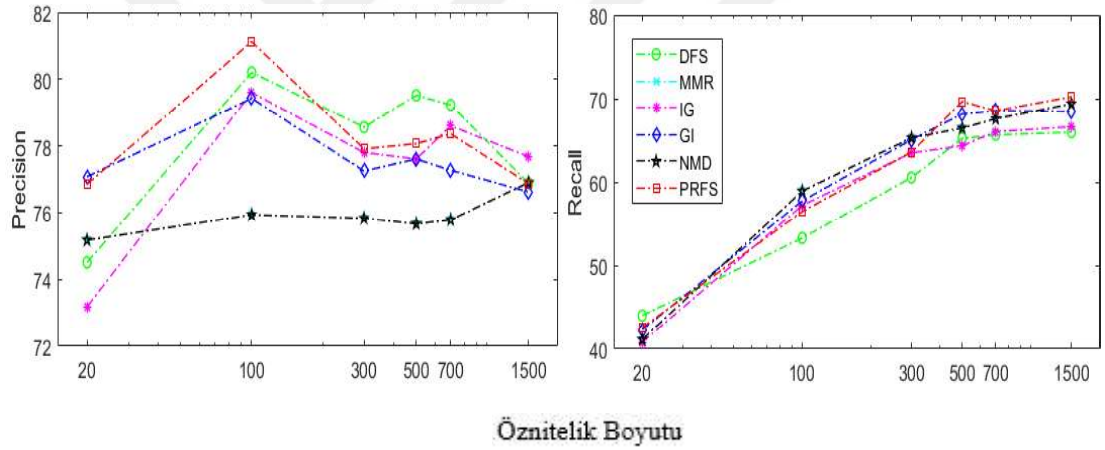
Tablo 5.5. *İlgili öznitelik boyutlarında tüm veri kümeleri için SVM, KNN, DT ve NB sınıflandırıcıları kullanıldığında ortalama Makro-F1 skorları*

Sınıflandırıcı	Öz. Boyut	Makro-F1					
		DFS	MMR	IG	GI	NDM	PRFS
SVM	20	70.50	72.61	75.12	74.86	72.74	**76.04**
	100	78.91	77.71	79.71	77.90	77.71	**80.04**
	300	79.96	79.14	79.69	80.46	78.14	**80.69**
	500	80.81	78.80	80.25	80.04	78.80	**81.84**
	700	80.53	79.39	79.72	79.80	79.39	**81.08**
	1500	**80.20**	79.72	79.73	79.75	79.71	80.08
KNN	20	54.69	55.85	52.97	**56.91**	55.85	55.85
	100	66.18	67.77	67.84	**68.80**	67.77	68.32
	300	70.14	70.93	71.15	**71.34**	70.93	71.22
	500	72.77	71.30	71.46	72.79	71.84	**73.23**
	700	72.85	71.84	72.69	72.79	71.84	**73.37**
	1500	71.79	**72.91**	72.44	72.35	**72.91**	**72.91**
DT	20	69.54	72.03	74.04	72.35	72.03	**74.86**
	100	78.30	76.76	78.52	77.86	76.77	**78.96**
	300	79.08	78.39	78.69	**79.23**	78.38	79.18
	500	78.92	78.25	78.66	78.60	78.25	**79.08**
	700	78.84	78.07	78.61	78.33	78.06	**79.12**
	1500	**78.70**	78.22	78.55	78.36	78.23	78.38
NB	20	69.60	71.04	73.37	72.54	71.04	**74.32**
	100	78.60	77.37	77.82	77.45	76.20	**78.98**
	300	79.97	78.84	79.87	79.61	78.53	**80.63**
	500	80.45	77.74	79.98	79.93	79.22	**80.53**
	700	**79.97**	79.61	79.79	79.50	79.61	79.88
	1500	79.53	79.84	80.02	**80.17**	79.84	80.01

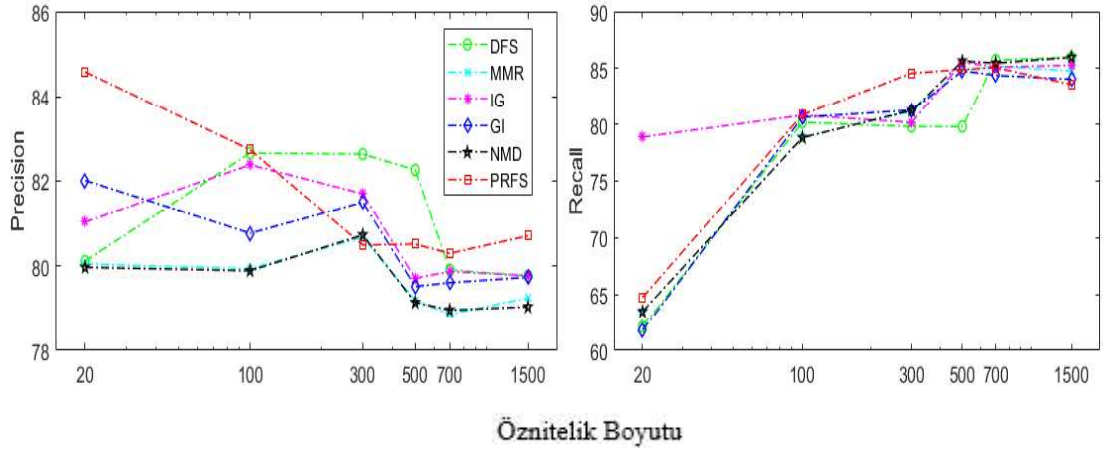
Ayrıca, Şekil 5.7'de, her bir sınıflandırıcı için öznitelik seçim yöntemlerinin ortalama Kesinlik (Precision) ve Duyarlılık (Recall) performansları gösterilmiştir.



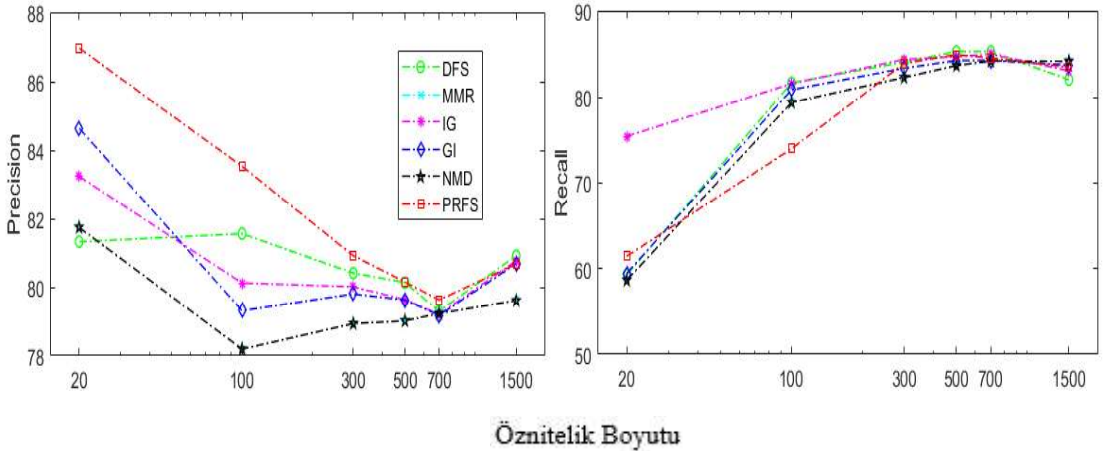
(a)



(b)



(c)



(d)

Şekil 5.7. *Tüm veri kümeleri üzerinde SVM (a), KNN (b), DT (c) ve NB (d) sınıflandırıcıları kullanıldığında ortalama Precision ve Recall skorları*

Kesinlik'te en yüksek değerler göz önünde bulundurulduğunda PRFS, karşılaştırma için kullanılan diğer tüm öznitelik seçim yöntemlerinden daha başarılı olduğu görülmektedir.

5.4. Değerlendirme

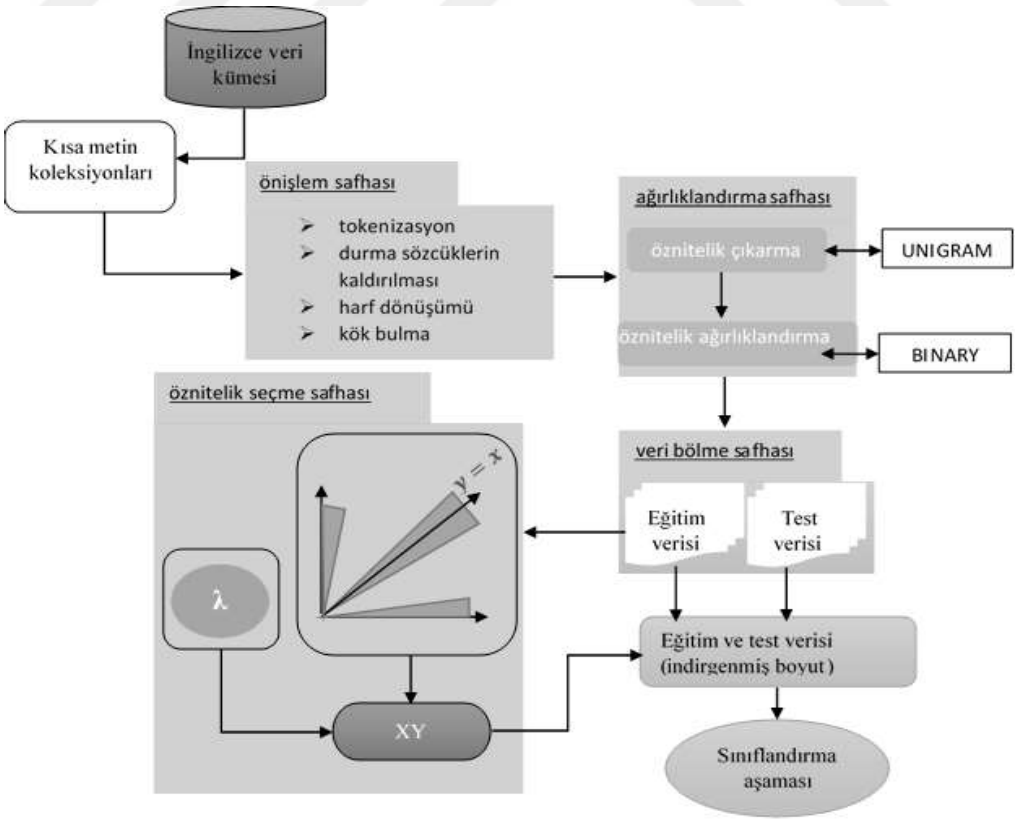
Çalışmanın bu bölümünde, PRFS olarak adlandırılan yeni bir filtre öznitelik seçim yöntemi önerilmiştir. Önerilen yöntem, KKT kullanarak terim / öznitelik değer kümesine göre bölgesel bir ayrımı temel alır. PRFS'de en önemli adımlardan biri doküman ve terimlerin/özniteliklerin ilişkilendirildiği vektör uzay modelinde her bir terimin vektör uzayındaki değerlere göre değerlendirilmesidir. Dahası PRFS, seyreklik sorununun etkisini azaltmak için oldukça başarılı bir performans sunmaktadır. Çünkü kaba kümeler verilerdeki eksiklik ve belirsizlikten etkili çıkarımlar yapmaktadır. Kaba kümeler, yeterli veri olmadığı durumlarda çok başarılı sonuçlar sunan bir yaklaşımdır.

Çalışma kapsamında sunulan yöntemde, kaba kümeler kullanılarak terim ve sınıf ilişkisi temelinde dokümanların durumu hakkında bilgi sunması bundan sonra bu alanda yapılacak çalışmalara bir ışık tutacağı kanısındayız. Bununla birlikte bildiğimiz kadarıyla literatürde, kaba kümeler kullanarak bir filtre özellik seçim yaklaşımı sunan bir çalışma bulunmamaktadır. Bu nedenle, kaba kümeler kullanarak bir filtre özellik seçim yöntemi öneren bu çalışma, bu araştırma alanında öncü bir çalışma olabilir.

6. ÖNERİLEN METOT: XY METOT

6.1. Motivasyon

Kısa metin sınıflandırma alanında yüksek boyutluluk ve seyreklik problemleri önemli bir yer tutmaktadır. Bu problemlere çözüm sunmak veya etkilerini en aza indirmek karar sistemlerinin performansını ve verimliliğini artırmak için önem arz etmektedir. Öznitelik seçme, yüksek boyutluluk problemine çözüm sunmanın en etkili yollardan birisidir. Ancak kısa metin sınıflandırma alanında kullanılan öznitelik seçme yaklaşımından hem seyreklik problemin etkisinden en az etkilenmesi hem de en önemli öznitelikleri seçmesi beklenir. Fakat literatürde metin sınıflandırma için kullanılan birçok öznitelik seçme yaklaşımı sınıf bilgisini görmezden gelerek sadece terim frekanslarına veya doküman frekanslarına bakmaktadır. Halbuki sınıf bilgisi metin sınıflandırma için önemli ve kullanışlı bir bilgidir. Ayrıca yine birçok metot sadece pozitif ve negatif sınıf dokuman frekanslarına dayalı çalışmaktadır. Dahası birçok yaklaşım terimleri seçerken çoğunlukla terimin sadece bir sınıfta geçmesine yoğunlaşmaktadır. Fakat bir terimin ayırt ediciliği birçok kritere bağlıdır. Sonuç olarak, bu durumları göz önünde bulundurarak yeni bir öznitelik seçme yaklaşımı önerilmiştir.

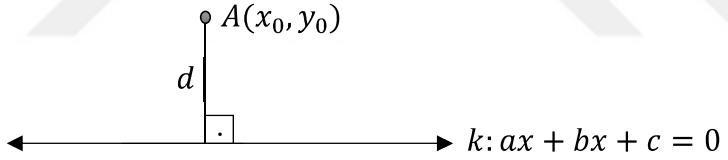


Şekil 6.1. *XY metodun çalışma stratejisi*

6.2. XY Metodunun Çalışma Stratejisi

Bu çalışmada XY Metot adında bir filtre öznitelik seçme metodu önerilmiştir. Önerilen metoda göre bir terimin XY doğrusuna[4] olan uzaklığı o terimin ayırt ediciliği için önemli bir faktör olduğundan bu metoda XY metodu ismi verilmiştir. Önerilen XY metot, bir filtre öznitelik seçme yaklaşımı olduğundan her terim için bir skor hesaplanır. XY metoduna göre bir terimin skoru hesaplanırken aşağıdaki adımlar sırasıyla uygulanır:

1. Her sınıfta terimin geçtiği tüm doküman sayısı o sınıf için koordinat değeri olarak alınır. İki sınıfın eksende olduğu bir koordinat düzleminde koordinat değerlerinden bir nokta elde edilir. Örneğin, *t* terimi C_1 sınıfında 5 ve C_2 sınıfında 4 dokümanda geçiyorsa C_1 için koordinat değeri 5, C_2 için 4 olur. Elde edilen nokta ise (5, 4) veya (4, 5) noktası olur.
2. Daha sonra elde edilen noktanın XY doğrusuna olan uzaklığı hesaplanır. Bunun için bir noktanın bir doğruya olan uzaklık formülünden yararlanır:



Şekil 6.2. *Bir noktanın bir doğruya uzaklığı*

A noktasının k doğrusuna olan uzaklığı,

$$d = \frac{|a * x_0 + b * y_0 + c|}{\sqrt{a^2 + b^2}} \quad \textbf{(6.1)}$$

Bir terimin ayırt ediciliğinin en kötü olduğu durum tüm sınıflarda aynı sayıda geçtiği durumdur. Buna göre her ekseninde birer sınıfın olduğu x ve y gibi iki eksenden oluşan bir koordinat düzleminde aynı terimin geçtiği doküman sayısı iki sınıfta da aynı ise bu terim bu koordinat düzleminde hangi doğru üzerinde olur? Bu sorunun cevabı bellidir: $x = y$ doğrusu, yani $x - y = 0$ doğrusudur. Buna

[4] Bu çalışma kapsamında XY doğrusu kavramı, $x = y$ *doğrusu* anlamında kullanılmıştır.

istinaden bu çalışmada $x = y$ doğrusu alınmıştır. Buna göre A noktasının $x - y = 0$ doğrusuna uzaklığı aşağıdaki gibi olur.

$$d = \frac{|x_0 - y_0|}{\sqrt{2}} \quad \textbf{(6.2)}$$

3. Bir sınıfın diğer tüm sınıflar ile ikili kombinasyonları için koordinat değerleri ve düzlemdeki yeri tespit edilir ve XY doğrusuna uzaklığı hesaplanır. Bu işlem tüm sınıflar için yapılır. Yani sınıf bazında tüm ikili kombinasyon düzlemi için 1. ve 2. adımlar uygulanarak XY doğrusuna tüm noktaların uzaklıkları hesaplanır.
4. Koordinat düzlemindeki her nokta için bir λ değeri hesaplanır. λ değeri aşağıdaki formül ile hesaplanır:

$$\lambda = \frac{min(x_0,\, y_0)}{max(x_0,\, y_0) + 1}, \qquad U{:}\,\lambda \rightarrow [0, 1) \quad \textbf{(6.3)}$$

5. Bir terimin tüm ikili koordinat düzlemlerdeki XY doğrusuna olan uzaklıkları λ değeri ile çarpılarak toplanır ve sonuç $P(t)$ değerine bölünür. XY metoduna göre bir terimin skoru aşağıdaki gibi hesaplanır:

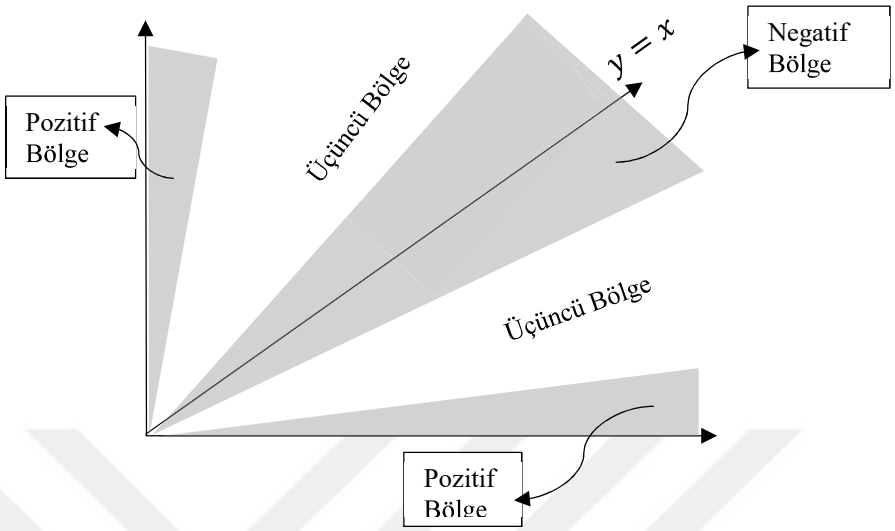
$$\mathrm{XY(t)} = \frac{\sum_{i=1}^{m*(m-1)/2}(1-\lambda) * d}{P(t) + 1} \quad \textbf{(6.4)}$$

Burada m sınıf sayısını ve $P(t)$, t terimin geçtiği sınıf olasılığını göstermektedir. Örneğin, üç sınıflı bir doküman koleksiyonu için *t* terimi bu üç sınıftan ikisinde geçiyorsa $P(t)$'te değeri 2/3 olur.

Hesaplanan skor değerinin yüksek olması o terimin yüksek ayırt ediciliğe, düşük olması ise düşük ayırt ediciliğe sahip olduğunu gösterir. Bu skor değeri aşağıdaki durumlara bağlı olarak değişmektedir. Diyelim ki $m = 2$ ve $A(x_t,\, y_t)$ t terimin koordinatı olsun. Buna göre;

- Eğer A noktası XY doğru üzerinde ise skoru sıfırdır. Dolayısıyla ayırt ediciliği çok düşüktür ve negatif bölgededir. Aynı zamanda skorunun sıfır olması o terimin *gereksiz* bir terim olduğu anlamına gelmektedir. Bu çalışmada XY doğrusu ve yakın bölgeleri negatif bölge, eksenler ve yakın bölgeleri de pozitif bölgeler ve bu

- iki bölge arasında kalan bölge ise üçüncü bölge olarak adlandırılmıştır. Aşağıda negatif, pozitif ve üçüncü bölgelerin temsili gösterimi verilmiştir.



Şekil 6.3. *XY metoduna göre bölgeler gösterimi*

Burada pozitif bölge bir terimin bir sınıfta geçerken diğer sınıfta geçmediğini veya çok az geçtiğini, negatif bölge ise her iki sınıfta da çoğunlukta geçtiğini göstermektedir. Pozitif ve negatif bölgeler arasında kalan bölgeler, üçüncü bölgeler, eksen yükseklikleri aynı olan noktaların pozitif bölgedekilerinden skorunun düşük, negatif bölgelerinkinden yüksek olduğunu göstermektedir. Yine aynı bölgede eksen yüksekliği aynı iki noktanın, XY doğrusuna uzaklığı fazla olanın skoru daha yüksek olur. Bu anlatılan bölgeleri belirlemek için bu çalışmada bir *lamda* (λ) değeri veya katsayısı elde edilmiştir. Şayet A bir terimin koordinat noktasını ve $2 < x_t,\ y_t$ olmak üzere bu noktanın koordinat değerlerini gösterirse, buna göre A noktası:

$$\begin{aligned} &Eğer\ \lambda \in [0, 0.1], && o\ zaman\ pozitif\ bölgede, \\ &Eğer\ \lambda \in [0.7, 1), && o\ zaman\ negatif\ bölgede, \\ &Eğer\ \lambda \in (0.1, 0.7), && o\ zaman\ üçüncü\ bölgede\ olur. \end{aligned} \qquad \textbf{(6.5)}$$

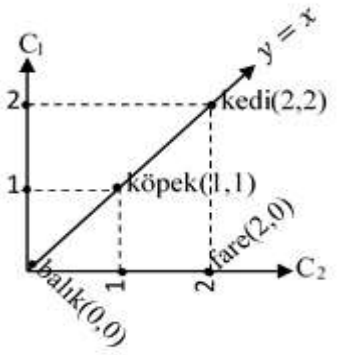
Burada "0.1" değeri, terimin %90 sadece tek bir sınıfta geçtiğini belirtir. Bu durum aynı zamanda bu terimin ayırt ediciliğinin iyi olduğu anlamına gelir. Bu yüzden bu tür terimler bu çalışmada pozitif bölgede olduğunu göstermiştir. "0.7"

değeri ise bir terimin %70 oranında her iki sınıfta geçtiğini belirtir. Bu durum terimin ayırt ediciliğinin iyi olmadığını gösterir. Bu yüzden bu terim negatif bölgede gösterilmişti. Son olarak "0" değeri terimin sadece tek sınıfta, "1" değeri ise her iki sınıfta aynı sayıda olduğunu göstermektedir. Not olarak şunun altını çizmekte gerekir: (0, 0), (1, 1) ve (2, 2) noktaları da negatif bölgededir. Ayrıca bir sonraki bölüm olan deneysel çalışma bölümünde λ ile ilgili bazı deneysel çalışmalar verilmiştir.

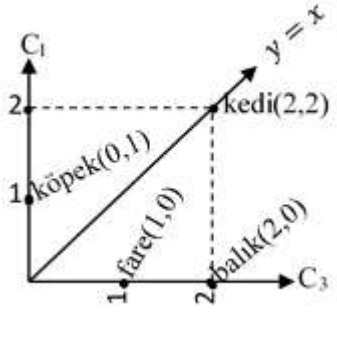
- Eğer A noktası eksenler üzerinde ve orijinden belli bir uzaklıkta ise terimin skoru yüksek olur. Dolayısıyla terimin ayırt ediciliği yüksektir ve terim pozitif bölgededir.
- Eğer A noktası eksenler üzerinde ancak orijine yakın bir uzaklıkta ise terimin skoru düşük olur. Dolayısıyla terimin ayırt ediciliği düşüktür ancak terim pozitif bölgededir.

Örnek:

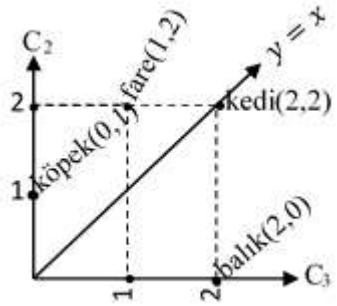
Bölüm 7'de verilen Tablo 5.1 temsili küçük bir doküman koleksiyonuydu. Bu basit örnek üzerine XY metodunun çalışması aşağıda gösterilmiştir (Bkz. Tablo 5.1). Ayrıca XY metoda göre bu basit örnek için her terimin skoru hesaplanmıştır. XY metoduna göre terimlerin skorunun hesaplanması aşağıda verilmiş ve koordinat düzleminde gösterimi de Şekil 6.4'de verilmiştir.



Düzlem 1



Düzlem 2



Düzlem 3

Şekil 6.4. *Basit örnek için her terimin koordinat düzleminde gösterimi*

$$XY(kedi) = \frac{\left(\frac{|2-2|}{\sqrt{2}}\right) * \left(1-\frac{2}{3}\right) + \left(\frac{|2-2|}{\sqrt{2}}\right) * \left(1-\frac{2}{3}\right) + \left(\frac{|2-2|}{\sqrt{2}}\right) * (1-\frac{2}{3})}{3/3+1} = 0.0000$$

$$XY(köpek) = \frac{\left(\frac{|1-1|}{\sqrt{2}}\right) * \left(1-\frac{1}{2}\right) + \left(\frac{|1-0|}{\sqrt{2}}\right) * \left(1-\frac{0}{2}\right) + \left(\frac{|1-0|}{\sqrt{2}}\right) * (1-\frac{0}{2})}{2/3+1} = 0.8485$$

$$XY(fare) = \frac{\left(\frac{|2-0|}{\sqrt{2}}\right) * \left(1-\frac{0}{3}\right) + \left(\frac{|1-0|}{\sqrt{2}}\right) * \left(1-\frac{0}{2}\right) + \left(\frac{|1-2|}{\sqrt{2}}\right) * (1-\frac{1}{3})}{2/3+1} = 1.5556$$

$$XY(balık) = \frac{\left(\frac{|0-0|}{\sqrt{2}}\right) * \left(1-\frac{0}{1}\right) + \left(\frac{|2-0|}{\sqrt{2}}\right) * \left(1-\frac{0}{3}\right) + \left(\frac{|2-0|}{\sqrt{2}}\right) * (1-\frac{0}{3})}{1/3+1} = 2.1213$$

Hesaplamalarda görüldüğü üzere, en yüksek değer *"balık"* terimi için hesaplanmıştır. Çünkü bu terim pozitif bölgede ve orijinden uzak koordinatlara sahiptir. Örneğin, Şekil 6.1 üzerindeki Düzlem 2 ve Düzlem 3'te görüldüğü gibi terim pozitif bölgede ve her iki düzlemde de orijinden belli uzaklıkta bir noktada kalmaktadır. Başka bir ifade ile *"balık"* terimi sadece bir sınıfta geçmiştir ve bu sınıf içinden de belli çoğunlukta dokümanda geçmektedir. Dolayısıyla sokurunun yüksek olması beklenir. Yapılan hesaplama da bunu göstermektedir. *"fare"* terimi hesaplanan en büyük ikinci skora sahiptir. Bu terim Düzlem 1'de pozitif bölgede ve orijinden uzak bir noktada yer almaktadır. Düzlem 2'de de yine pozitif bölgede ancak birinci düzlemdeki kadar orijinden uzak bir noktada değildir. Düzlem 3'te ise pozitif ve negatif bölge arasındaki bölgede ve

negatif bölgeye daha yakın bir noktada kalmaktadır. Dolayısıyla skor değeri *"balık"* terimine göre daha düşük olması beklenir. Çünkü bu terim iki sınıfta geçmektedir ancak daha çok C_2 sınıfında geçmektedir. *"köpek"* terimi Düzlem 1'de negatif bölgede iken Düzlem 2 ve Düzlem 3'te pozitif bölgede yer almaktadır. Dolayısıyla *"köpek"* terimi için hesaplanan skor değeri, *"balık"* ve *"fare"* terimlerinden az *"kedi"* teriminkinden fazladır. *"kedi"* teriminin hesaplanan skor değeri sıfırdır. Bu değer bu terimin gereksiz bir terim olduğunu göstermektedir. Düzlemler incelendiğinde de *"kedi"* terimin tüm düzlemlerde negatif bölgelerde ve XY doğrusu üzerinde olduğu görülmektedir.

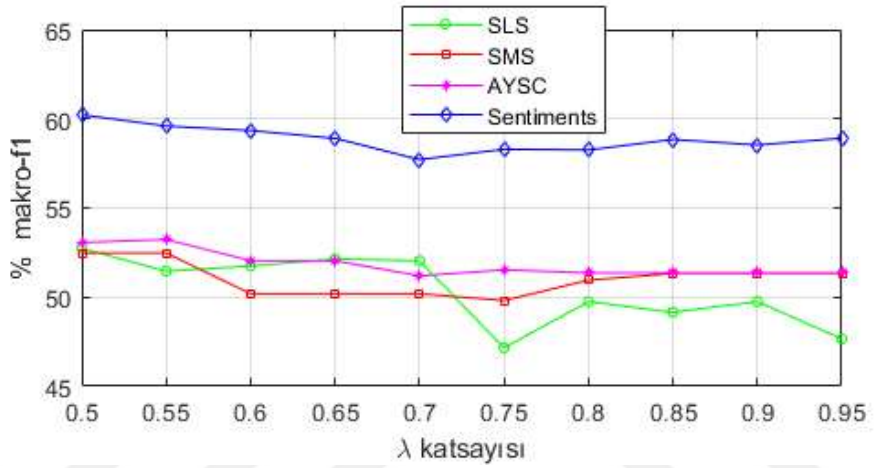
6.3. Deneysel Çalışmalar

Bu bölümde deneysel çalışmaların sonuçlarını göstereceğiz. İlk olarak kullanılan veri kümeleri hakkında kısa bir bilgi vermek gerekir. Bu bölümde kullanılan veri kümeleri bir önceki bölümde kullanılan veri kümeleri ile aynı veri kümeleri olduklarından bu bölümde veri kümeleri ile ilgili bir bilgi verilmeyecektir (Bkz. Tablo 5.2). Ayrıca başarı ölçütü bir önceki bölümde anlatılan Makro-F1 olduğundan bu ölçütten de bahsedilmeyecektir (Bkz. Eşitlik 5.5 ve 5.6). Bu bölümde öncelikle *lamda* değeri üzerine bazı analiz çalışmasına değineceğiz. Daha sonra XY yöntemi ile seçilen terimlerin sınıflandırıcıların başarısı üzerindeki etkisini değerlendiriyoruz. Bu amaçla, XY yönteminin performansı, önceki bölümlerde açıklanan öznitelik seçim yöntemleriyle karşılaştırılmıştır. Karşılık gelen öznitelik seçim algoritmaları tarafından seçilen öznitelikleri kullanan sınıflandırıcıların performansı bu bölümde gösterilmiştir.

6.3.1. *Lamda* (λ) değerinin analizi

Lamda değeri XY metodunda önemli bir faktördür. Terim bazında dokümanların kaldığı bölgeyi belirleme ve terimin ayırt ediciliğindeki etkisi çok önemlidir. Bir terim şayet tek bir sınıfta belirli çoğunlukta geçiyorsa ayırt ediciliğini belirlemek kolaydır. Ancak diğer durumlarda terimin ayırt ediciliğini belirlemek hassas bir hesaplama gerektirir. λ katsayısı bize bu hassasiyeti sağlamaktadır. Diyelim ki iki sınıflı bir metin koleksiyonunda $t_1$terimi, C_1sınıfında 100 ve C_2sınıfında ise 80 dokümanda geçsin. Benzer şekilde t_2 terimi, C_1 sınıfında 20 ve C_2 sınıfında ise 40 dokümanda geçsin. Buna göre bu iki terimin XY doğrusuna uzaklığı eşit ve yaklaşık olarak 14.14'tür. Ancak λ değerleri sırasıyla t_1 için 0.8 ve t_2 için 0.5'tir. λ değeri ne kadar küçük çıkarsa o terimin ayırt ediciliği o kadar yüksek olabilir. Bu yüzden örneğimizdeki XY doğrusuna uzaklığı eşit bu iki terimden t_2 terimin ayırt ediciliği daha yüksek olur.

Lamda değeri aynı zamanda bölgeleri belirleyen bir katsayıdır. Bu katsayı her bölge için belli değer aralığında bulunur. Çalışmanın önceki bölümlerinde bu aralıklardan bahsedilmişti (Bkz. Eşitlik 6.5). Bu çalışma kapsamında kullanılan veri kümeleri üzerinde negatif bölge alt sınırını belirlemek için SVM sınıflandırıcısı kullanılarak alt bölge sınırı belirlenmiştir. Bunun için $\lambda = [0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.8, 0.85, 0.90, 0.95]$ her bir değeri için negatif bölgede kalan terimler kullanılarak SVM sınıflandırıcısının Makro-F1 skorları değerlendirilmiştir. Elde edilen skorlar aşağıda Şekil 6.5'de gösterilmiştir.



Şekil 6.5. *Lamda değerlerine göre SVM sınıflandırıcı kullanılarak veri kümeleri üzerinde negatif bölge alt sınır belirleme*

Şekil 6.5 değerlendirildiğinde $\lambda = 0.7$ ve $\lambda = 0.75$ değerleri için en düşük skorlar elde edildiği görülüyor. Bunun anlamı ayırt ediciliği kötü en çok terimin aynı anda birlikte olduğu terim kümesinin oluşturulduğudur. Sonuç olarak bu çalışma kapsamında negatif bölge alt sınırı 0.7 alınmasının nedeni budur.

Bu çalışmada önerilen metodun amaçlarından biri de *Lamda* katsayısı yardımı ile elde edilen negatif bölgeden olabildiğince az sayıda terim seçmektir. Bu amaç doğrultusunda boyut 500 öznitelik için XY metot, IG, DP, Chi2, MMR ve DFS metotların bu öznitelik boyutunda seçtiği terimlerin negatif bölgede olma oranları aşağıdaki Tablo 6.1'de verilmiştir.

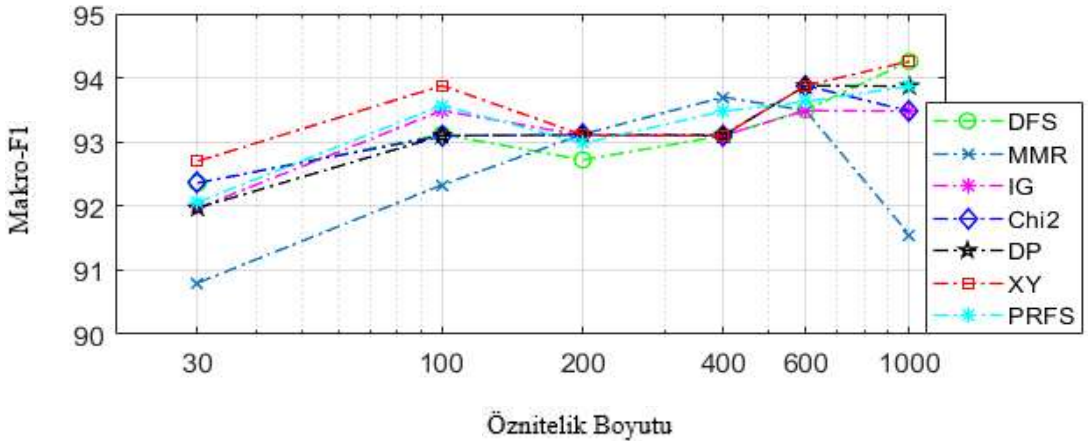
Tablo 6.1. *Boyut 500 öznitelik için XY metot ve diğer yaklaşımların negatif bölgede terim seçme oranları*

Veri kümesi	IG	XY	DP	Chi2	MMR	DFS
SMS	0.054	**0.042**	0.052	0.052	0.092	0.044
AYSC	0.112	**0.092**	0.112	0.104	0.190	0.098
SLS	0.188	**0.114**	0.164	0.164	0.408	0.126
Sentiments	0.744	**0.690**	0.750	0.760	0.986	0.728

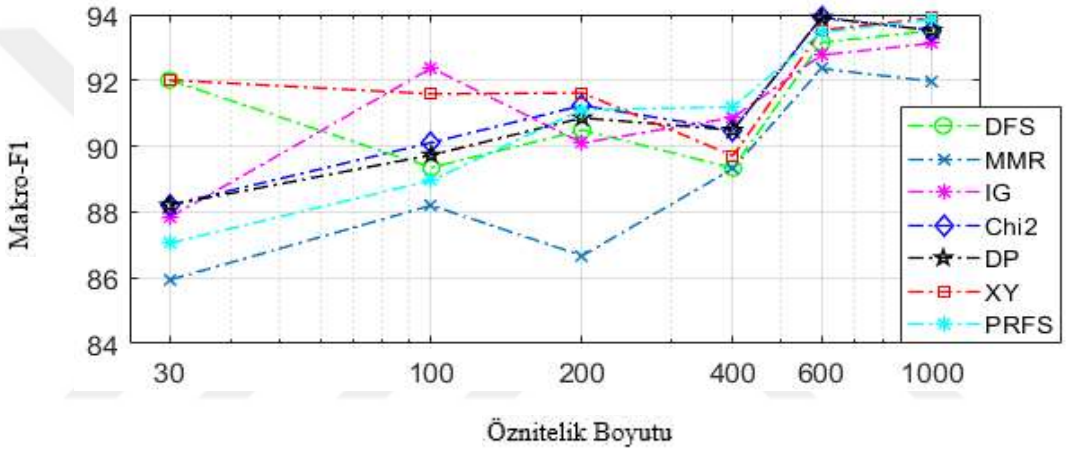
Tablo 6.1 incelendiğinde XY metodunun diğer metotlara göre negatif bölgede daha az terim seçtiği açıkça görülmektedir.

6.3.2. Başarı analizi

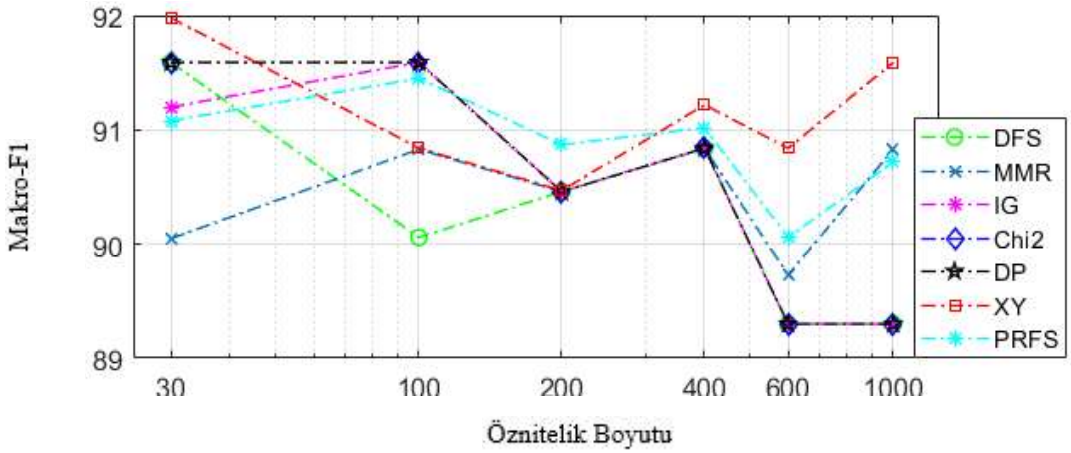
Öznitelik seçme yöntemlerinin performansını test etmek için çeşitli öznitelik boyutları kullanılır. Bu çalışmada her bir seçim yöntemi ile seçilen belirli sayıdaki yüksek dereceli 30, 100, 200, 400, 600 ve 1000 öznitelik boyutları kullanılarak SVM, KNN, DT ve BN sınıflandırıcıları ile test edilmiştir. Deneylerde elde edilen Makro-F1 skorları, her bir veri kümesi için Şekil 6.6-6.9'da listelenmiştir En yüksek değerler göz önüne alındığında, XY metot ve PRFS metodunun performansı, karşılaştırma için kullanılan tüm diğer özellik seçim yöntemlerinin performansından daha iyidir. Özellikle de burada XY metodun performansına bakmak gerekir. Örneğin, SMS ve AYSC veri kümelerinde, XY tarafından boyut 400 hariç diğer tüm öznitelik boyutlarında seçilen öznitelikler ile SVM sınıflandırıcısı kullanıldığında en yüksek Makro-F1 skorları elde edilmiştir. Ayrıca, Sentiments veri kümesinde, bu durum tüm öznitelik boyutları için geçerlidir. Benzer şekilde, SLS veri kümesinde en yüksek Makro-F1 skorları, boyut 30, 100 ve 200 öznitelik için NB kullanılarak sağlanır. Dahası, aynı sınıflandırıcı kullanıldığında AYSC veri kümesindeki boyut 30 ve 100'de, Sentiments veri kümesindeki boyut 100 ve 600 öznitelik için aynı durum geçerlidir. SMS veri kümesinde XY tarafından boyut 30, 200 ve 1000 öznitelik için KNN, boyut 100, 200, 600 ve 1000 öznitelikte DT kullanıldığında en yüksek Makro-F1 skorlarına ulaşılır. Ayrıca, Sentiments veri kümesinde, boyut 30 dışındaki tüm öznitelik boyutlarında en yüksek değerlere DT kullanıldığında elde edilir. Başka bir örnek olarak, Şekil 6.9'da, XY metodun SVM sınıflandırıcısını kullanarak çoğunlukla en yüksek Makro-F1 skorlarını ürettiğini gözlemledik.


95
94
93
92
91
90
Makro-F1
30
100
200
400
600
1000
Öznitelik Boyutu
DFS
MMR
IG
Chi2
DP
XY
PRFS

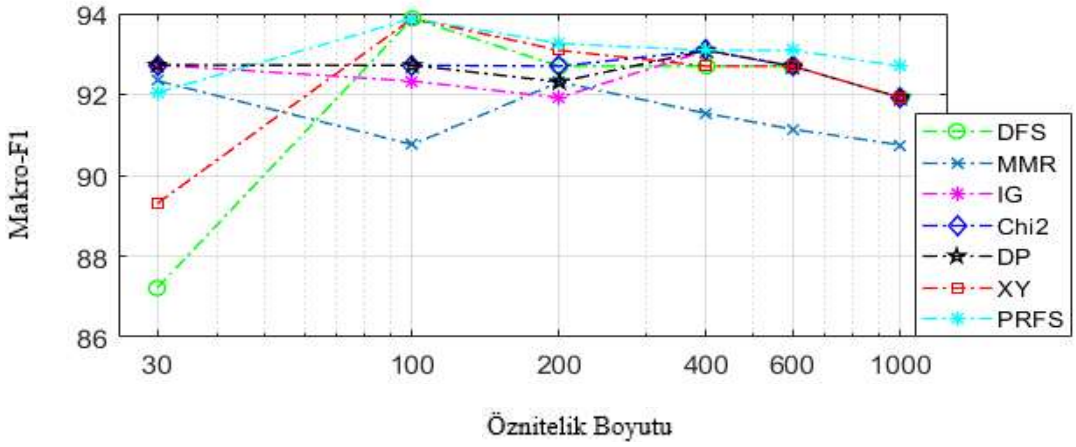
(a)


94
92
90
88
86
84
Makro-F1
30
100
200
400
600
1000
Öznitelik Boyutu
DFS
MMR
IG
Chi2
DP
XY
PRFS

(b)

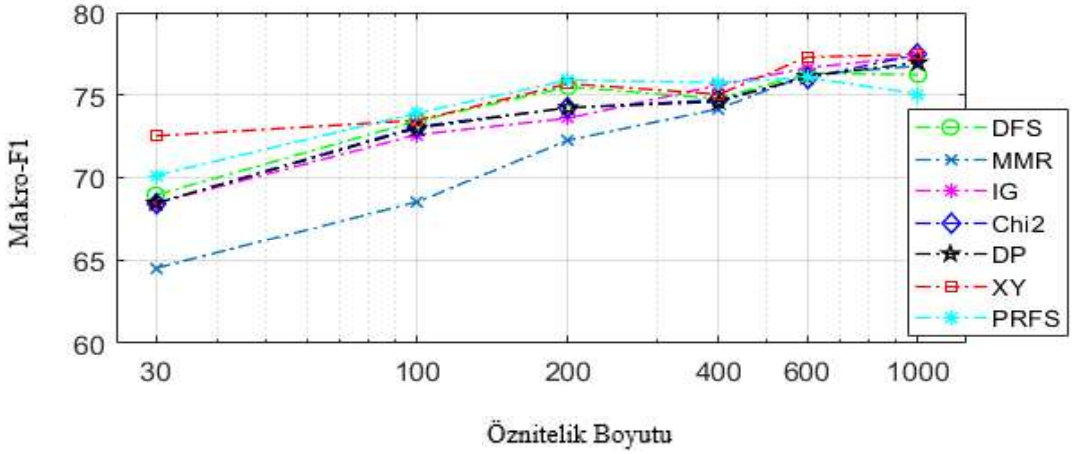

92
91
90
89
Makro-F1
30
100
200
400
600
1000
Öznitelik Boyutu
DFS
MMR
IG
Chi2
DP
XY
PRFS

(c)

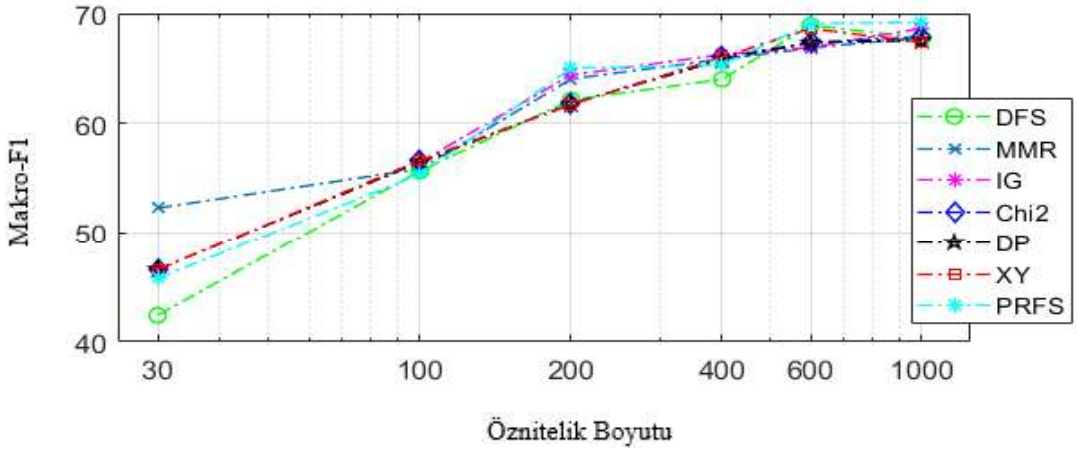


(d)

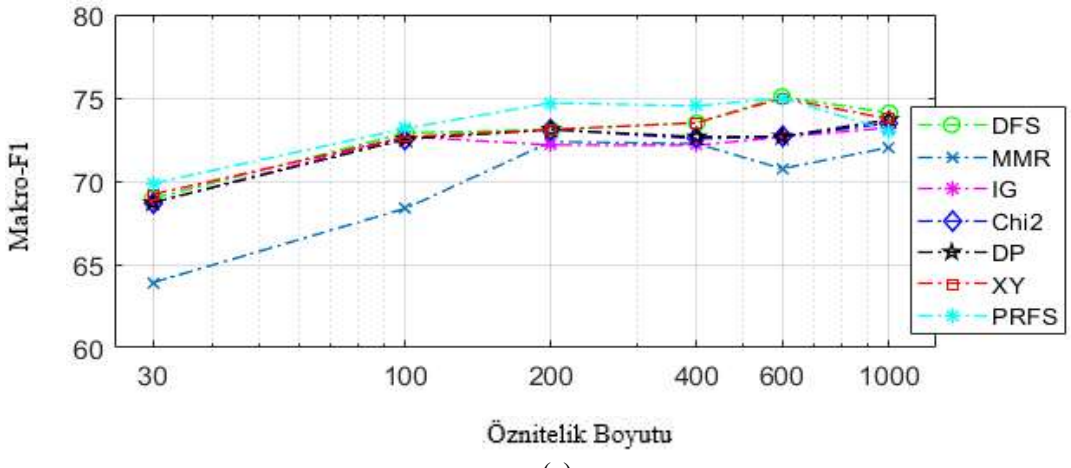
Şekil 6.6. ***SMS*** *veri kümesi için SVM (a), KNN (b), DT (c) ve NB (d) sınıflandırıcısı kullanıldığında Makro-F1 skorları*



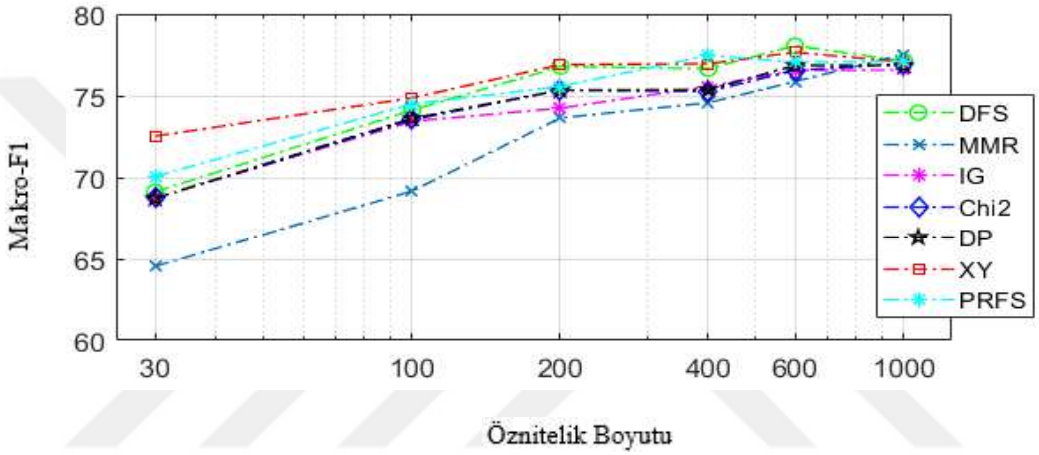
(a)



(b)

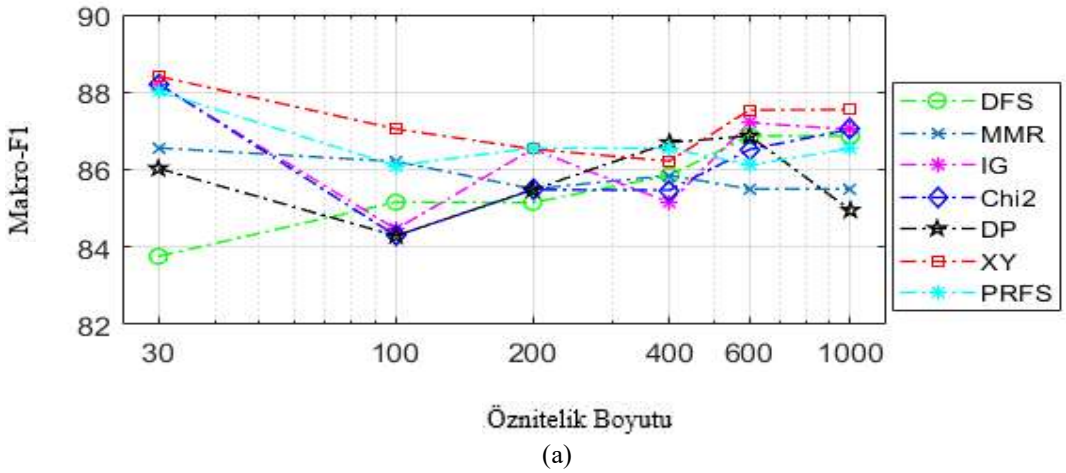


(c)

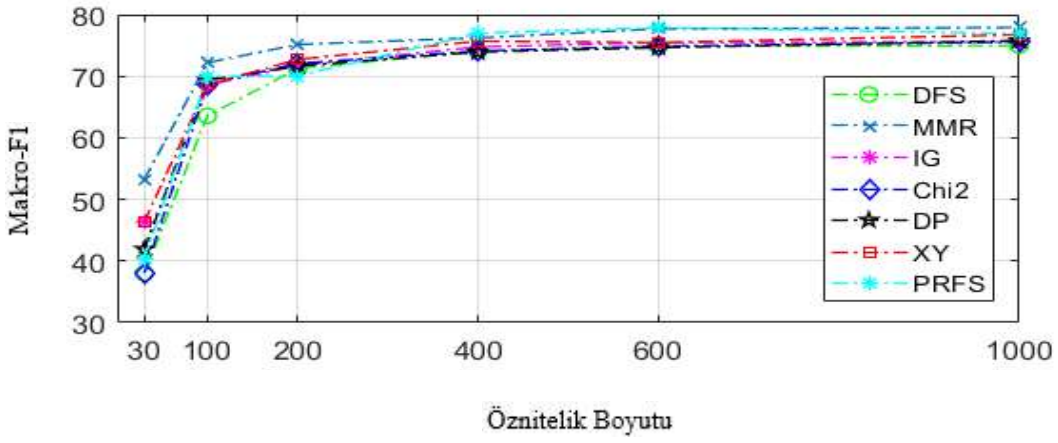


(d)

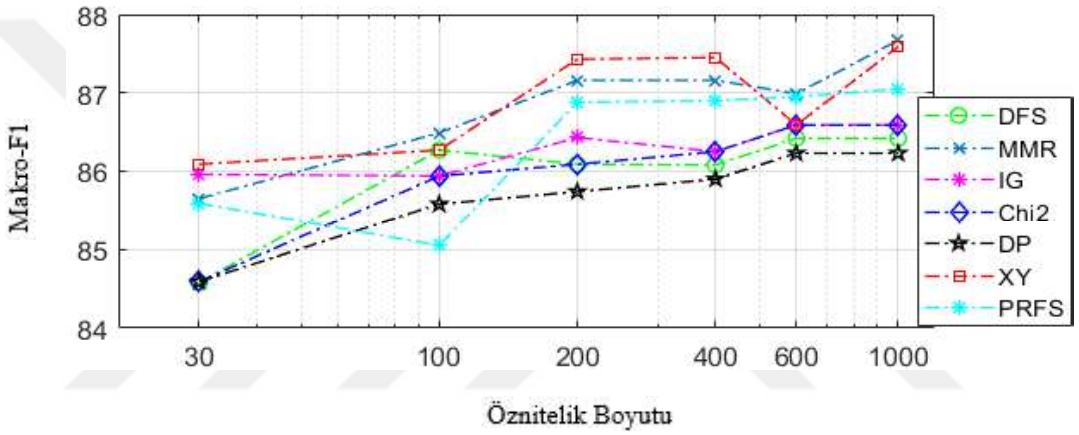
Şekil 6.7. ***SLS*** *veri kümesi için SVM (a), KNN (b), DT (c) ve NB (d) sınıflandırıcısı kullanıldığında Makro-F1 skorları*



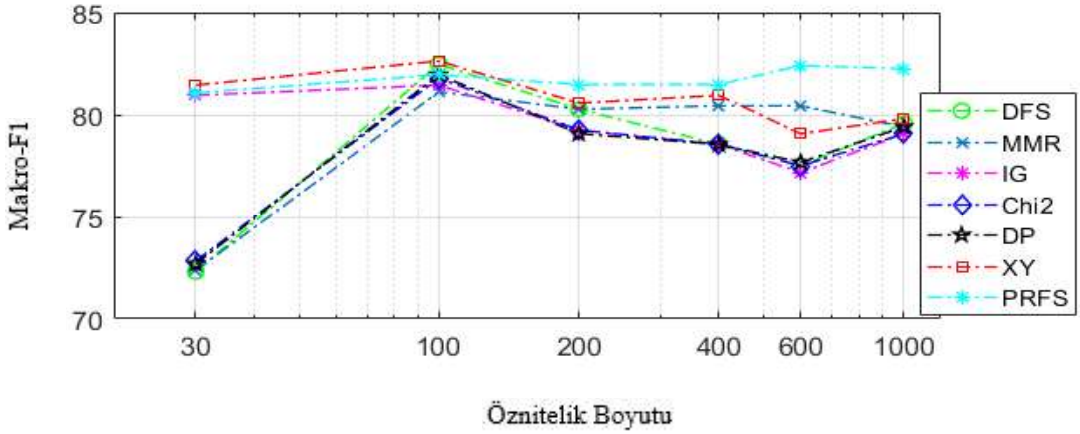
(a)



(b)

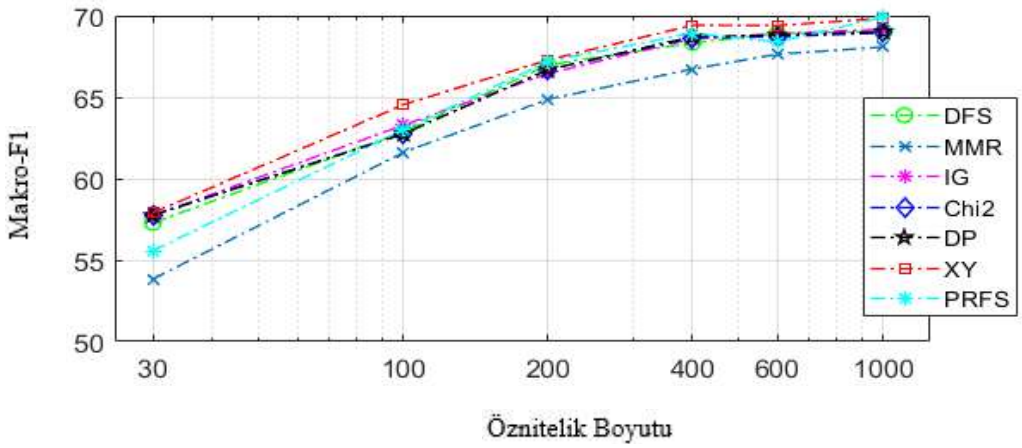


(c)

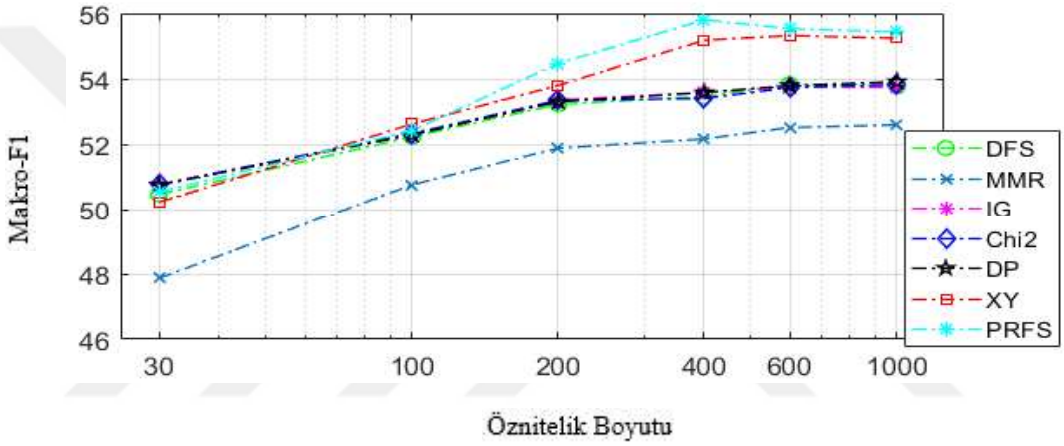


(d)

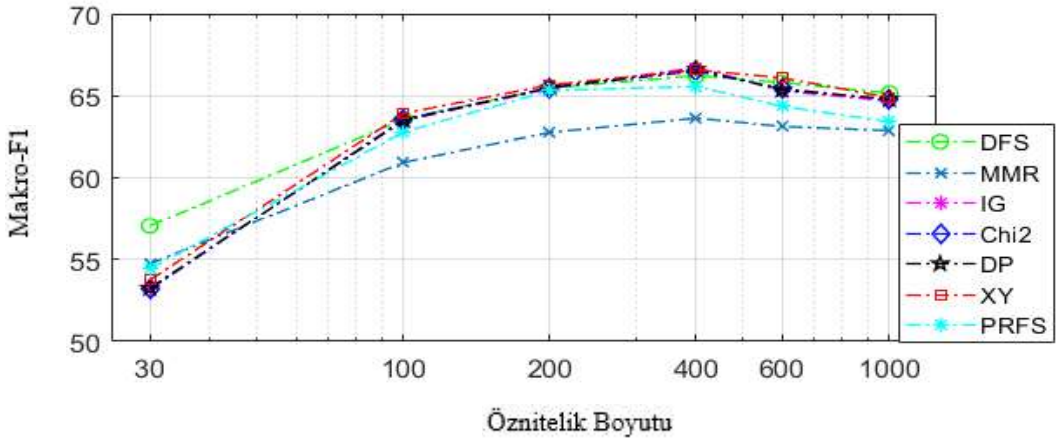
Şekil 6.8. ***AYSC*** *veri kümesi için SVM (a), KNN (b), DT (c) ve NB (d) sınıflandırıcısı kullanıldığında Makro-F1 skorları*


70
65
60
55
50
Makro-F1
30
100
200
400
600
1000
Öznitelik Boyutu
DFS
MMR
IG
Chi2
DP
XY
PRFS

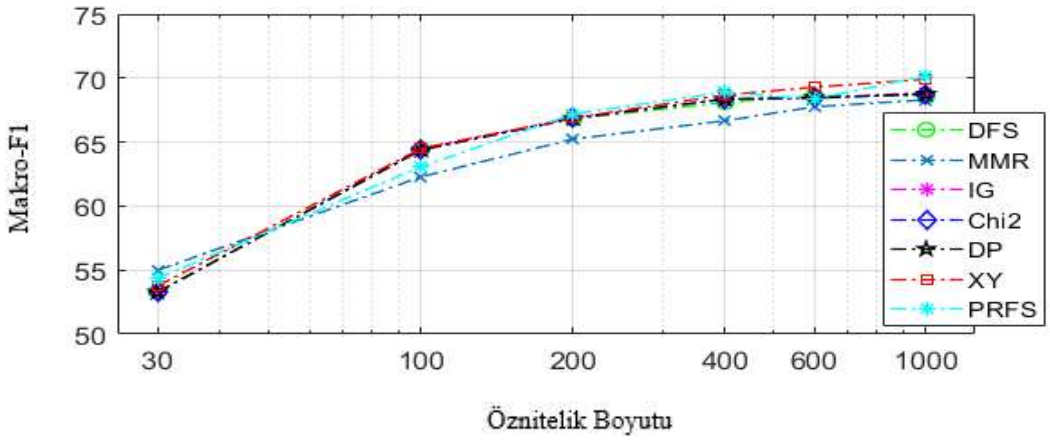
(a)


56
54
52
50
48
46
Makro-F1
30
100
200
400
600
1000
Öznitelik Boyutu
DFS
MMR
IG
Chi2
DP
XY
PRFS

(b)


70
65
60
55
50
Makro-F1
30
100
200
400
600
1000
Öznitelik Boyutu
DFS
MMR
IG
Chi2
DP
XY
PRFS

(c)



(d)

Şekil 6.9. ***Sentiments*** *veri kümesi için SVM (a), KNN (b), DT (c) ve NB (d) sınıflandırıcısı kullanıldığında Makro-F1 skorları*

Sonuç olarak, önceki şekillerde sunulan sonuçları açıklığa kavuşturmak amacıyla her veri kümesini temsil etmek için Tablo 6.2-6.5 verilmiştir. Bu tablolar, her bir üst öznitelik boyutu için tepe noktası veren yöntemlerin adlarını içerir. XY yönteminin öznitelik boyutları dikkate alınarak performans oranı tabloların son sütununda verilmiştir. Bu oran, XY yönteminin her veri kümesindeki her sınıflandırıcı için ulaştığı piklerin bir göstergesidir. Örneğin, Tablo 6.2'de XY yöntemi, SVM sınıflandırıcısı kullanılarak 6 en üst öznitelik boyutundan 5'inde en yüksek noktaya ulaşmıştır. Bu nedenle, Tablo 6.2'de SVM için XY oranı 0.8333'tür.

Tablo 6.2. *Her öznitelik boyutu için* ***SMS*** *veri kümesinde en yüksek Makro-F1 skorlarını veren öznitelik seçme yöntemleri*

	Öznitelik Boyutu						
Sınıflandırıcı	30	100	200	400	600	1000	XY oranı
SVM	XY	XY	~~DFS~~	MMR	XY, DP, Chi2	XY, DFS	0.8333
KNN	XY, DFS	IG	XY	IG	DP, Chi2	XY	0.5000
DT	XY	Chi2, DP	#	XY	XY	XY	0.8333
NB	DP,Chi2,IG	XY,PRFS	PRFS	Chi2,IG, DP,PRFS	PRFS	PRFS	0.1667

"~~DFS~~" gösterimi DFS yöntem dışında diğer tüm metotların olduğunu göstermektedir.
"# " gösterimi tüm metotların olduğunu göstermektedir.

Tablo 6.3. *Her öznitelik boyutu için **SLS** veri kümesinde en yüksek Makro-F1 skorlarını veren öznitelik seçme yöntemleri*

	Öznitelik Boyutu						
Sınıflandırıcı	30	100	200	400	600	1000	XY oranı
SVM	XY	PRFS	PRFS	PRFS	XY	XY, Chi2, IG	0.8333
KNN	MMR	XY, Chi2	IG	~~DFS~~	DFS	PRFS	0.3333
DT	PRFS	PRFS	PRFS	PRFS	DFS	DFS	0.0000
NB	XY	XY	XY	PRFS	DFS	MMR	0.5000

"~~DFS~~" gösterimi DFS yöntem dışında diğer tüm metotların olduğunu göstermektedir.

Tablo 6.4. *Her öznitelik boyutu için **AYSC** veri kümesinde en yüksek Makro-F1 skorlarını veren öznitelik seçme yöntemleri*

	Öznitelik Boyutu						
Sınıflandırıcı	30	100	200	400	600	1000	XY oranı
SVM	XY, IG	XY	XY, IG, PRFS	DP	XY	XY	0.8333
KNN	MMR	MMR	MMR	PRFS	PRFS	PRFS	0.0000
DT	XY	MMR	MMR	MMR	MMR	MMR	0.1666
NB	IG	XY	PRFS	PRFS	PRFS	PRFS	0.1667

Tablo 6.5. *Her öznitelik boyutu için **Sentiments** veri kümesinde en yüksek Makro-F1 skorlarını veren öznitelik seçme yöntemleri*

	Öznitelik Boyutu						
Sınıflandırıcı	30	100	200	400	600	1000	XY oranı
SVM	XY	XY	XY	XY	XY	XY, PRFS	1.0000
KNN	XY	Chi2	PRFS	PRFS	PRFS	PRFS	0.1667
DT	DFS	XY	XY	XY	XY	DFS	0.6666
NB	MMR	XY	XY, PRFS	XY, PRFS	XY	XY, PRFS	0.8333

Verilen grafikleri (Bkz. Şekil 6.6, 6.7, 6.8 ve 6.9) analiz etmek için en önemli adımlardan biri öznitelik seçme yöntemleriyle elde edilen tepe değerlerini göstermektir. Bu amaçla, Tablo 6.6 sunulmuştur. Tablo 6.6'daki satırlar, karşılık gelen sınıflandırıcıyı kullanarak veri kümelerindeki tepe değerlerini elde eden öznitelik seçim yöntemini gösterir. Tablo 6.6 göz önüne alındığında, önerilen XY yönteminin en yüksek Makro-F1 skorlarına göre en iyi yöntem olduğu anlaşılmaktadır. Ayrıca, SVM sınıflandırıcısını kullanan tüm veri kümelerinde en yüksek skorlara ulaştığı ve aynı şekilde SMS veri kümesinde tüm sınıflandırıcılar için tepe değerini sağlamıştır. Tablo 6.6'da satırlar öznitelik seçme yaklaşımlarının veri kümeleri üzerindeki performansını, sütunlar ise

sınıflandırıcılar ile performansını göstermektedir. Hem satır hem de sütunda yer alan XY metodunun oranı veri kümesi ve sınıflandırıcı bazında en yüksek skora çıkma performansının bir göstergesidir.

Tablo 6.6. *Farklı deneysel setler için en yüksek Makro-F1 skorunu (tepe değeri-pik noktası) sağlayan öznitelik seçme yöntemleri*

	Makro-F1				
Veri kümesi	SVM	KNN	DT	NB	XY oranı
SMS	DFS, XY	XY	XY	DFS, XY, PRFS	1.0000
Sentiments	XY, PRFS	PRFS	XY, DP	XY, PRFS	0.7500
AYSC	XY	MMR	MMR	XY	0.5000
SLS	XY	PRFS	DFS	DFS	0.2500
XY oranı	1.0000	0.5000	0.5000	0.7500	

Tablo 6.7'de, veri kümesini göz ardı ederek karşılık gelen sınıflandırıcı için önerilen PRFS dışında diğer öznitelik seçim yöntemlerinin ortalama performanslarını göstermekteyiz. Bu nedenle, her bir öznitelik seçim yönteminin farklı veri kümelerinde elde ettiği skorlarının ortalaması bu sunumda alınmıştır. Amaç, her öznitelik boyutu için yöntemlerin performansını analiz etmektir. Başka bir deyişle bu deneysel çalışma, yöntemlerin tüm veri kümelerindeki her öznitelik boyutu için sınıflandırıcılar ile nasıl performans gösterdiğini görmek için yapılmıştır. Bu şekilde sınıflandırıcılar ile birlikte daha iyi sonuçlar veren yöntemleri açıkça gözlemleyebiliriz. Ayrıca bu şekilde yöntemin istatiksel olarak istikrarlı olup olmadığı hakkında bir bilgi sahibi olunur. Örneğin, önerilen XY yöntemi için Tablo 6.7'deki 77.8850 puanı, SVM sınıflandırıcısı kullanıldığında boyut 30 öznitelikte tüm veri kümelerinde Makro-F1 puanlarının ortalaması alınarak elde edilmiştir. Tablo 6.7 incelendiğinde, önerilen XY yönteminin kullanılan neredeyse tüm özellik boyutlarındaki diğer yaklaşımlardan daha iyi sonuçlar verdiği söylenebilir. Sonuçlara göre, önerilen XY yönteminin tüm veri kümelerindeki farklı öznitelik boyutları için kararlı olduğu söylenebilir.

Tablo 6.7. *İlgili öznitelik boyutlarında tüm veri kümeleri için SVM, KNN, DT ve NB sınıflandırıcıları kullanıldığında ortalama Makro-F1 skorları*

Sınıflandırıcı	Öz. Boyut	Makro-F1					
		DFS	Chi2	IG	DP	MMR	XY
SVM	30	75.5925	76.7050	76.5925	76.0500	73.9275	**77.8850**
	100	78.6500	78.3150	78.4625	78.2875	77.1750	**79.7375**
	200	80.0875	79.8800	79.9225	79.8825	78.9425	**80.6550**
	400	80.5175	80.4800	80.6150	80.7800	80.1125	**80.9475**
	600	81.3975	81.2875	81.5600	81.4225	80.7200	**82.0300**
	1000	81.6025	81.7325	81.7650	81.2000	80.4725	**82.2775**
KNN	30	55.7150	55.9100	57.8825	56.8650	**59.8350**	58.7950
	100	65.2050	66.9375	**67.4375**	66.9525	66.7500	67.3550
	200	69.2725	69.6025	69.8925	69.3950	69.4375	**69.9575**
	400	70.2250	71.0000	71.3675	70.9525	70.8900	**71.6500**
	600	72.7000	72.4700	72.2275	72.4250	72.3500	**73.2375**
	1000	72.5050	72.6925	72.7725	72.6825	72.5775	**73.3100**
DT	30	**75.5450**	74.5175	74.7800	74.5350	73.5850	75.2525
	100	78.2025	78.4075	**78.4275**	78.3025	76.6575	78.4250
	200	78.7825	78.7875	78.6275	78.6975	78.1900	**79.1675**
	400	79.1625	79.0400	78.9950	79.0050	78.4725	**79.6875**
	600	79.1525	78.5025	78.4625	78.3950	77.6500	**79.6275**
	1000	78.7450	78.5350	78.4450	78.5050	78.3525	**79.4500**
NB	30	70.4850	71.8825	73.9225	71.8575	71.0575	**74.2575**
	100	78.7300	78.1200	77.9000	78.1725	75.8350	**78.9750**
	200	79.1425	78.5475	78.0550	78.4000	77.8650	**79.3625**
	400	79.0075	78.8200	78.8925	78.8450	78.3025	**79.8250**
	600	79.2175	78.7950	78.7125	78.9275	78.8025	**79.6800**
	1000	79.3100	79.1800	79.1100	79.2350	78.9975	**79.6750**

6.3.3. Zaman karmaşıklık analizi

Metin dokümanları ve öznitelik seçme matematiksel olarak aşağıdaki gibi ifade edilebilir (Wang & Hong, 2019):

- $|D|$ sayıdaki metin doküman kümesi; $D = \{D_i | i = 1, 2, \dots, |D|\}$ şeklinde

- D'nin terim kümesi; $T = \{t_j | j = 1, 2, \dots, |T|\}$ şeklinde
- D'nin sınıf kümesi; $C = \{c_k | k = 1, 2, \dots, |C|\}$ şeklinde ifade edilir.

Önerilen metotların zaman karmaşıklığı temel olarak yöntemin istatiksel formülündeki hesaplanmaları için döngü gerektiren fonksiyonlara bağlıdır. Buna göre kıyaslama için kullanılan DFS, Chi2, IG, DP, PRFS ve MMR yaklaşımların karmaşıklığı $O(|C| * |D| * |T|)$'dir. XY metodun karmaşıklığı ise $O(|T| * |D| * |C^2|)$'dir.

6.4. Değerlendirme

Kısa metin belgeleri çok az kelime içerdiğinden, sınıflandırmaları uzun metin belgelerinden daha zordur. Bu nedenle kısa metinlerde etkili bir şekilde çalışabilecek yöntemlere ihtiyaç vardır. Bu amaç doğrultusunda bu çalışmada, kısa metinler üzerinde etkili bir şekilde çalışan XY metot adında yeni bir öznitelik seçme yaklaşımı sunulmuştur. Önerilen XY metot yöntemi, iki boyutlu bir düzlemdeki bir noktanın XY çizgisine olan uzaklık mesafesinin hesaplanmasına dayanır. Bu nokta, terimin bir sınıfta geçtiği belge sayısına ve terimin başka bir sınıfta geçtiği belge sayısına dayanır. XY yöntemine göre, bir terimin XY çizgisine uzaklığı, ayırt edici özelliği hakkında önemli bilgiler sağlar. Ayrıca XY metot çalışma stratejisi kapsamında sunulan *lamda* değeri özniteliklerin ayırt edicilik yetenekleri hakkında bir önbilgi sunmaktadır. Aslında lamda değeri daha çok hangi özniteliğin ayırt ediciliğin kötü olduğu hakkında bilgi sunmaktadır. Bu bilgi bu alanda yapılacak çalışmalara bir yol gösterici olabilir. Dahası deneysel sonuçlar, XY yönteminin kısa metin sınıflandırması için karşılaştırılan diğer özellik seçim yöntemlerinden daha iyi performans elde ettiğini göstermektedir.

7. SONUÇ VE TARTIŞMA

Bu çalışma kapsamında, kısa metin sınıflandırma alanında sınıflandırıcı, makine öğrenmesi gibi karar sistemlerinin başarımını artırmak için ayırt ediciliği yüksek terimler ile temsil edilen vektör uzay modelinde mümkün olduğunca bu terimleri daha verimli bir biçimde yansıtabilmenin yolları araştırılmıştır. Bu bağlamda, vektör uzay modelinde tüm terimleri/öznitelikleri en iyi temsil eden alt terim vektör uzayını seçme alanına odaklanılmıştır. Tüm terim/öznitelik uzayını en iyi temsil edecek alt terim uzayını belirlemenin en etkili yollardan biri öznitelik seçme yaklaşımı kullanılarak etkili öznitelikleri seçmektir. Bu yüzden bu alanda yapılmış en iyi bilinen öznitelik seçme yaklaşımları tez kapsamında incelenmiştir. Bu inceleme ile var olan metotların seçme için izledikleri stratejileri analiz edilmiş ve buna bağlı avantaj ve dezavantajları hakkında bilgi elde edilmiştir. Ayrıca kısa metin koleksiyonun vektör uzay modeli ile temsil edildiğinde ortaya çıkan karar bilgi sisteminin çok az veri içerdiği görülmüştür. Dolayısıyla, metin sınıflandırma alanında öznitelik seçme yaklaşımı olarak sunulan mevcut birçok yöntemin bu alanda istenilen performansı göstermediği görülmüştür. Bu yüzden çalışma kapsamında, bu az veri içeren bilgi sistemleri üzerinde etkili çalışabilecek yeni yaklaşımları sunmanın yolları araştırılmıştır.

Yukarıdaki açıklamalar doğrultusunda belirtilen hedeflere uluşmanın en etkili yollardan biri az veriden etkili çıkarım yapacak bir araç kullanılmasıdır. Kaba kümeler bu çalışmada bu amaç için kullanılan bir matematiksel araçtır. Kaba kümelerin kendisi gibi çıkarım yapan diğer yaklaşımlardan farkı çıkarım için öncesinden herhangi bir bilgiye veya fonksiyona ihtiyaç duymamasıdır. Onun bu özelliği bu çalışmada kullanılmasının önemli bir etkenidir. Ayrıca az veriden oldukça etkili çıkarımlar sunması diğer önemli etkendir.

Tez çalışmasının amaçları doğrultusunda belirtilen hedeflere ulaşmak için yapılan ilk deneysel çalışma kaba kümeler tabanlı önerilen metodun, yani PRFS'nin ve mevcut en çok bilenen öznitelik yaklaşımlarının sunduğu alt terim uzayına bağlı ilgili veri kümeleri üzerinde bazı karar sistemlerinin performanslarındaki değişimler analiz edilmiştir. Analiz sonuçları önerilen metodun daha etkin bir terim alt uzayı seçtiğini göstermiştir. Ayrıca önerilen metodun öznitelikleri seçme stratejisi bundan sonra yapılacak çalışmalara ışık tutabilir. Çünkü önerilen metot istatistiksel bir formül sunmanın ötesinde öznitelikleri değerlendirme stratejisi ile ön plana çıkmaktadır.

Yöntemin öznitelikleri değerlendirme şekli mevcut yöntemlerden farklı ve daha anlamlı ilişkilere dayanmaktadır.

Metin sınıflandırma alanında çalışan mevcut öznitelik seçme yaklaşımlarının çoğu sınıf bilgisini hesaba katmadan sadece terim veya doküman frekansına dayalı çalışmaktadır. Bu yüzden bu tür yaklaşımların en iyi terim alt kümeyi seçme performansları düşüktür. Dahası sınıf bilgisini kullanan yaklaşımların ise büyük bir çoğunluğu terimleri pozitif ve negatif sınıflara göre değerlendirmektedir. Tezin ikinci deneysel çalışması, kısa metin sınıflandırma alanında çalışan matematiksel bir yaklaşım olan XY metot, mevcut yöntemlerin bu çalışma stratejilerine alternatif sunacak bir kapsama sahiptir. XY metot sınıf bilgisinin terimler üzerindeki etkisini hesaplayarak, terimleri sınıf bazında ikişerli tüm sınıf kombinasyonlarında değerlendirmektedir. Bu şekilde terim bazında her sınıfın diğer sınıflara göre ayırt ediciliği dikkate alınmıştır. Ayrıca bu yaklaşım stratejisi kapsamında sunulan *lamda* değeri ile hangi terimlerin ayırt ediciliği düşük veya gereksiz oldukları tespit edilebiliyor. Bu değer, bu konuda yapılacak çalışmalara bazı noktalarda önemli yol gösterici olabilir. Çünkü *lamda* değeri ile belirlenen negatif bölgede olabildiğince az terim seçmek öznitelik seçme yaklaşımlarının seçim başarılarının bir göstergesidir. Nitekim Bölüm 6'da "*Lamda* değerinin analizi" alt başlığı altında sunulan Tablo 6.1'deki değerler, özniteliklerin başarısı hakkında bize bazı ön bilgiler sunmaktadır. Bu bilgilere göre ilgili veri kümeleri üzerinde en etkili çözüm sunan XY metodudur. Aslında deneysel çalışmalara bakıldığında bu durumu destekler sonuçlar elde edildiği açıkça görülmektedir.

Sonuç olarak bu çalışma ile kısa metin sınıflandırma alanında kullanılmak üzere yani iki yaklaşım sunulmuştur. Bu iki yaklaşım kısa metin alanında temel iki ana probleme odaklı çözüm sunan birer yaklaşımdır.

KAYNAKÇA

Abualigah, L. M., & Khader, A. T. (2017). Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering. *The Journal of Supercomputing*, 73(11), 4773-4795.

Aggarwal, C. C., & Zhai, C. (. (2012). *Mining text data.* Londra: Springer Science & Business Media.

Aggarwal, C., & Zhai, C. (2012). *A survey of text classification algorithms.* In Mining Text Data.

Agnihotri, D., Verma, K., & Tripathi, P. (2017). Variable global feature selection scheme for automatic classification of text documents. *Expert Systems with Applications*, 81, 268-281.

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

Alberto, T. C., Lochter, J. V., & Almeida, T. A. (2015). Tubespam: Comment spam filtering on youtube. *Paper presented at the 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA).*

Al-Radaideh, Q. A., & Al-Qudah, G. Y. (2017). Application of rough set-based feature selection for Arabic sentiment analysis. *Cognitive Computation,*, 9(4), 436-445.

Baccianella, S., Esuli, A., & Sebastiani, F. (2013). Using micro-documents for feature selection: The case of ordinal text classification. *Expert Systems with Applications*, 40(11), 4687-4696.

Bekkali, M., & Lachkar, A. (2019). An effective short text conceptualization based on new short text similarity. *Social Network Analysis and Mining,*, 9(1), 1.

Bermejo, P., De la Ossa, L., G´amez, J., & Puerta, J. (2012). Fast wrapper feature subset selection in highdimensional datasets by means of filter re-ranking. *Knowl-Based Syst*, 25(1):35–44.

Bharti, K. K., & Singh, P. K. (2015). Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering. *Expert Systems with Applications*, 42(6), 3105-3114.

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

Cekik, R., & Telceken, S. (2018). A new classification method based on rough sets theory. *Soft Computing,*, 22(6), 1881-1889.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering,*, 40(1), 16-28.

Chen, H., Jiang, W., Li, C., & Li, R. (2013). A heuristic feature selection approach for text categorization by using chaos optimization and genetic algorithm. *Mathematical problems in Engineering*, 15-26.

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

Chen, K. Z. (2016). Turning from TF-IDF to TF-IGM for term weighting in text classification. *Expert Systems with Applications*, 66, 245-260.

Chou, C., Sinha, A., & Zhao, H. (2010). A hybrid attribute selection approach for text classification. *J Assoc Inf Syst*, 11(9):491.

Chouchoulas, A., & Shen, Q. (1999). A rough set-based approach to text classification. *Paper presented at the International Workshop on Rough Sets, Fuzzy Sets,Data Mining, and Granular-Soft Computing*.

Dai, A. M., & Le, Q. V. (2015). Semi-supervised sequence learning. *In Advances in neural information processing systems*, 3079-3087.

De Mántaras, R. (1991). *A distance-based attribute selection measure for decision tree induction.* Mach. Learn.

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

Duan, J., Luo, B., & Zeng, J. (2020). Semi-supervised Learning with Generative Model for Sentiment Classification of Stock Messages. *Expert Systems with Applications*, 113540.

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

Fu, X., Wei, Y., Xu, F., Wang, T., Lu, Y., Li, J., & Huang, J. Z. (2019). Semi-supervised aspect-level sentiment classification model based on variational autoencoder. *Knowledge-Based Systems*, 171, 81-92.

Ghareb, A., Bakar, A., & Hamdan, A. (2016). Hybrid feature selection based on enhanced genetic algorithm for text categorization. *Expert Syst Appl*, 49:31–47.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford,*, 1(12), 2009.

Gong, L., Zeng, J., & Zhang, S. (2011). Text stream clustering algorithm based on adaptive feature selection. *Expert Systems with Applications,*, 38(3), 1393-1399.

Gupta, K. M., Aha, D. W., & Moore, P. (2006). Rough set feature selection algorithms for textual case-based classification. *Paper presented at the European Conference on Case-Based Reasoning.*

Guru, D. S., Suhil, M., Raju, L. N., & Kumar, N. V. (2018). An alternative framework for univariate filter based feature selection for text categorization. *Pattern Recognition Letters*, 103, 23-31.

Gutlein, M., Frank, E., Hall, M., & Karwath, A. (2009). Large-scale attribute selection using wrappers. *In:IEEE symposium on computational intelligence and data mining, 2009. CIDM'09. IEEE,* ., (s. pp 332–339.).

Jensen, R., & Shen, Q. (2008). Computational intelligence and feature selection: rough and fuzzy approaches. *John Wiley & Sons.*, Vol. (8):.

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

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 11-21 (28).

Khan, S. M., Chowdhury, M., Ngo, L. B., & Apon, A. (2020). Multi-class twitter data categorization and geocoding with a novel computing framework. *Cities*, 96, 102410.

Kim, K., Chung, B. S., Choi, Y., Lee, S., Jung, J. Y., & Park, J. (2014). Language independent semantic kernels for short-text classification. *Expert Systems with Applications*, 41(2), 735-743.

Kim, K., Chung, B. S., Choi, Y., Lee, S., Jung, J. Y., & Park, J. (2014). Language independent semantic kernels for short-text classification. *Expert Systems with Applications,*, 41(2), 735-743.

Kotzias, D., Denil, M., De Freitas, N., & Smyth, P. (2015). From group to individual labels using deep features. *Paper presented at the Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). *Text classification algorithms: A survey.* Information.

Kushwaha, N., & Pant, M. (2018). Link based BPSO for feature selection in big data text clustering. *Future Generation Computer Systems*, 82, 190-199.

La, L., Cao, Q., & Xu, N. (2019). Dimensionality Reduction by Feature Co-Occurrence based Rough Set. *International Journal of Performability Engineering,*, 15(1).

Labani, M., Moradi, P., Ahmadizar, F., & Jalili, M. (2018). A novel multivariate filter method for feature selection in text classification problems. *Engineering Applications of Artificial Intelligence*, 70, 25-37.

Labani, M., Moradi, P., Ahmadizar, F., & Jalili, M. (2018). A novel multivariate filter method for feature selection in text classification problems. *Engineering Applications of Artificial Intelligence,*, 70, 25-37.

Li, Y. L., & Chung, S. M. (2008). Text clustering with feature selection by using statistical data. *IEEE Transactions on knowledge and Data Engineering*, 20(5), 641-652.

Li, Y., Shiu, S. C.-K., Pal, S. K., & Liu, J. N.-K. (2006). A rough set-based case-based reasoner for text categorization. *International journal of approximate reasoning*, 41(2), 229-255.

Li, Z., Lu, W., Sun, Z., & Xing, W. (2017). A parallel feature selection method study for text classification. *Neural Computing and Applications*, 28(1), 513-524.

Liang, W., Xie, H., Rao, Y., Lau, R. Y., & Wang, F. L. (2018). Universal affective model for Readers' emotion classification over short texts. *Expert Systems with Applications,*, 114, 322-333.

Liu, Y. L. (2009). Imbalanced text classification: A term weighting approach. . *Expert systems with Applications*, 36(1), 690-701.

Lochter, J. V., Zanetti, R. F., Reller, D., & Almeida, T. A. (2016). Short text opinion detection using ensemble of classifiers and semantic indexing. *Expert Systems with Applications*, 62, 243-249.

Lu, Y., Liang, M., Ye, Z., & Cao, L. (2015). Improved particle swarm optimization algorithm and its application in text feature selection. *Applied Soft Computing*, 35, 629-636.

Magerman, D. (1995). *Statistical decision-tree models for parsing.* Cambridge: In Proceedings of the 33rd Annual Meeting.

Manning, C., Raghavan, P., & Schütze, H. (2010). Introduction to information retrieval. *Natural Language Engineering*, 16(1), 100-103.

Miao, D., Duan, Q., Zhang, H., & Jiao, N. (2009). Rough set based hybrid algorithm for text classification. *Expert Systems with Applications,*, 36(5), 9168-9174.

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

Nuruzzaman, M. T., Lee, C., & Choi, D. (2011). Independent and Personal SMS Spam Filtering. *Paper presented at the 2011 IEEE 11th International Conference on Computer and Information Technology.*

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.

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.

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

Pawlak, Z. (1998). Rough set theory and its applications to data analysis. . *Cybernetics & Systems,* , 29(7), 661-688.

Pearson, E. (1925). *Bayes' theorem, examined in the light of experimental sampling.* Biometrika.

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

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. (1986). *Induction of decision trees.* Mach. Learn.

Rao, Y., Xie, H., Li, J., Jin, F., Wang, F. L., & Li, Q. (2016). Social emotion classification of short text via topic-level maximum entropy model. *Information \& Management*, 978-986.

Rashid, J., Shah, S. M., & Irtaza, A. (2019). Fuzzy topic modeling approach for text mining over short text. *Information Processing & Management*, 56(6), 102060.

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

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

Rehman, A., Javed, K., Babri, H. A., & Asim, M. N. (2018). Selection of the most relevant terms based on a max-min ratio metric for text classification. *Expert Systems with Applications*, 114, 78-96.

Rehman, A., Javed, K., Babri, H. A., & Saeed, M. (2015). Relative discrimination criterion–A novel feature ranking method for text data. *Expert Systems with Applications,*, 42(7), 3670-3681.

Ren, F. &. (2013). Class-indexing-based term weighting for automatic text classification. . *Information Sciences*, 236, 109-125.

Shamsinejadbabki, P., & Saraee, M. (2012). A new unsupervised feature selection method for text clustering based on genetic algorithms. *Journal of Intelligent Information Systems*, 38(3), 669-684.

Shang, C., Li, M., Feng, S., Jiang, Q., & Fan, J. (2013). Feature selection via maximizing global information gain for text classification. *Knowledge-Based Systems*, 54, 298-309.

Shang, W. H. (2007). A novel feature selection algorithm for text categorization. . *Expert Systems with Applications,*, 33(1), 1-5.

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.

Sharmin, S., Shoyaib, M., Ali, A. A., Khan, M. A., & Chae, O. (2019). Simultaneous feature selection and discretization based on mutual information. *Pattern Recognition*, 91, 162-174.

Shi, L., Ma, X., Xi, L., Duan, Q., & Zhao, J. (2011). Rough set and ensemble learning based semi-supervised algorithm for text classification. *Expert Systems with Applications,*, 38(5), 6300-6306.

Singh, S., & Dey, L. (2005). A new customized document categorization scheme using rough membership. *Applied Soft Computing,*, 5(4), 373-390.

Song, G., Ye, Y., Du, X., Huang, X., & Bie, S. (2014). Short text classification: A survey. *Journal of multimedia*, 9(5), 635.

Sparck Jones, K. (2004). A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation*, 28, 11-21.

Tommasel, A., & Godoy, D. (2018a). A Social-aware online short-text feature selection technique for social media. *Information Fusion*, 40, 1-17.

Tommasel, A., & Godoy, D. (2018b). Short-text feature construction and selection in social media data: a survey. *Artificial Intelligence Review*, 49(3), 301-338.

Tutkan, M., Ganiz, M. C., & Akyokuş, S. (2016). Helmholtz principle based supervised and unsupervised feature selection methods for text mining. *Information Processing & Management,*, 52(5), 885-910.

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

Uysal, A. K., Gunal, S., Ergin, S., & Gunal, E. S. (2013). The impact of feature extraction and selection on SMS spam filtering. *Elektronika ir Elektrotechnika,*, 19(5), 67-72.

Wang, H., & Hong, M. (2015). Distance variance score: an efficient feature selection method in text classification. *Mathematical Problems in Engineering*, 1-10.

Wang, H., & Hong, M. (2019). Supervised Hebb rule based feature selection for text classification. *Information Processing & Management*, 56(1), 167-191.

Wang, S., Li, D., Wei, Y., & Li, H. (2009). A feature selection method based on fisher's discriminant ratio for text sentiment classification. *In International Conference on Web Information Systems and Mining* (s. 88-97). Berlin, Heide: Springer,.

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., & Vapni, V. (2001). Feature selection for svms. *In: Advances in neural information processing systems,*, (s. pp 668–674.).

Yan, R., Cao, X. B., & Li, K. (2009). Dynamic Assembly Classification Algorithm for Short Text. *ACTA ELECTRONICA SINICA*, Vol. 37(5), pp. 1019-1024, 2009.

Yang, L. L., Q., L. L., & Li, L. (2013). Combining lexical and semantic features for short text classification. *Procedia Computer Science,*, 78-86.

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

Yao, H., Liu, C., Zhang, P., & Wang, L. (2017). A feature selection method based on synonym merging in text classification system. *EURASIP Journal on Wireless Communications and Networking*, 2017(1), 1-8.

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

Zhang, Q., Xie, Q., & Wang, G. (2016). A survey on rough set theory and its applications. *CAAI Transactions on Intelligence Technology,*, 1(4), 323-333.

Zhao, H., Wang, P., & Hu, Q. (2016). Cost-sensitive feature selection based on adaptive neighborhood granularity with multi-level confidence. *Information sciences,*, 366, 134-149.

Zheng, L., Diao, R., & Shen, Q. (2015). Self-adjusting harmony search-based feature selection. *Soft Computing,*, 19(6), 1567-1579.

ÖZGEÇMİŞ

ORCID NO: 0000-0002-7820-413X

E-Posta Adresi	:	rasimcekik@eskisehir.edu.tr
Telefon (İş)	:	2223213550-6569
Telefon (Cep)	:	5432198385
Adres	:	Eskisehir Teknik Üni. Mühendislik Fak. Bilgisayar Müh. Bölümü

Öğrenim Bilgisi

Yüksek Lisans 2012	ANADOLU ÜNİVERSİTESİ MÜHENDİSLİK FAKÜLTESİ/BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ/BİLGİSAYAR DONANIMI ANABİLİM DALI Tez adı: EKG Sinyallerin Kaba Kümeler Teorisi İle Sınıflandırılması Tez Danışmanı:(SEDAT TELÇEKEN)
Lisans 2005 2010	FIRAT ÜNİVERSİTESİ MÜHENDİSLİK FAKÜLTESİ/BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ/BİLGİSAYAR MÜHENDİSLİĞİ PR.

Görevler

ARAŞTIRMA GÖREVLİSİ 2018	ESKİŞEHİR TEKNİK ÜNİVERSİTESİ/MÜHENDİSLİK FAKÜLTESİ/BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ/BİLGİSAYAR DONANIMI ANABİLİM DALI)
ARAŞTIRMA GÖREVLİSİ 2012	ANADOLU ÜNİVERSİTESİ/MÜHENDİSLİK FAKÜLTESİ/BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ/BİLGİSAYAR DONANIMI ANABİLİM DALI)
ARAŞTIRMA GÖREVLİSİ 2010-2012	ŞIRNAK ÜNİVERSİTESİ/MÜHENDİSLİK FAKÜLTESİ/BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ/BİLGİSAYAR DONANIMI ANABİLİM DALI)

Projelerde Yaptığı Görevler:

1. Kömürcü Otomasyonu, DİĞER, Uzman, 2008-2008)

Eserler

Uluslararası hakemli dergilerde yayımlanan makaleler:

ÇEKİK RASIM,UYSAL ALPER KÜRŞAT (2020). A Novel Filter Feature Selection Method

1. Using Rough Set for Short Text Data. EXPERT SYSTEMS WITH APPLICATIONS
(Yayın No: 6364498)

2. ÇEKİK RASIM,TELÇEKEN SEDAT (2016). A new classification method based on rough sets theory. Soft Computing, 1-9., Doi: 10.1007/s00500-016-2443-0
(Yayın No: 3066540)

3. ÇEKİK RASIM,TELÇEKEN SEDAT (2014). EKG SİNYALLERİNİN KABA KÜMELER TEORİSİ KULLANILARAK SINIFLANDIRILMASI.
Anadolu University of Sciences & Technology-A: Applied Sciences & Engineering, 15(2), 125-135.
(Yayın No: 3428134)

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

1. ÇEKİK RASIM,TELÇEKEN SEDAT (2016). New Method Based on Rough Set for Filling Missing Value. World Conference on Soft Computing (WConSC) (Tam Metin Bildiri/Sözlü Sunum)
(Yayın No:3078173)

2. TURAN ABDULLAH,ÇEKİK RASIM (2016). KAMPANA FREN TASARIMI ve FRENLEME KUVVETLERİNİNMODELLENMESİ. Internation Conference Reseach in Education and Science (ICRES), 1372-1379. (Tam Metin Bildiri/Sözlü Sunum)
(Yayın No:3077975)

3. ÇEKİK RASIM,TELÇEKEN SEDAT (2016). ÖZNİTELİK SEÇME İÇİN KABA KÜMELER GENEL BİR BAKIŞ. Internation Conferance on Reseach in Education and Science(ICRES), 1258-1267. (Tam Metin Bildiri/Sözlü Sunum)
(Yayın No:3077687)

Üniversite Dışı Deneyim

Elmar Yazılım, Web tasarımı ve Masaüstü program üreten ve satan yazılım şireketi., (Diğer)