METİN SINIFLANDIRMA İÇİN TERİM AĞIRLIKLANDIRMA

Doktora Tezi

Turgut DOĞAN

Eskişehir 2019

METİN SINIFLANDIRMA İÇİN TERİM AĞIRLIKLANDIRMA

Turgut DOĞAN

DOKTORA TEZİ

Bilgisayar Mühendisliği Anabilim Dalı
Danışman: Dr. Öğr. Üyesi Alper Kürşat UYSAL

Eskişehir
Eskişehir Teknik Üniversitesi
Lisansüstü Eğitim Enstitüsü
Ağustos 2019

JÜRİ VE ENSTİTÜ ONAYI

Turgut DOĞAN'ın "Metin Sınıflandırma İçin Terim Ağırlıklandırma" başlıklı tezi 20/08/2019 tarihinde aşağıdaki jüri tarafından değerlendirilerek "Eskişehir Teknik Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği'nin ilgili maddeleri uyarınca, Bilgisayar Mühendisliği Anabilim dalında Doktora tezi olarak kabul edilmiştir.

Jüri Üyeleri	**Unvanı Adı Soyadı**	**İmza**
Üye (Tez Danışmanı)	: Dr. Öğr. Üyesi Alper Kürşat UYSAL	……
Üye	: Prof. Dr. Serkan GÜNAL	……
Üye	: Doç. Dr. Semih ERGİN	……
Üye	: Dr. Öğr. Üyesi Ahmet ARSLAN	……
Üye	: Dr. Öğr. Üyesi Mehmet KOÇ	……

Prof. Dr. Murat TANIŞLI
Lisansüstü Eğitim Enstitüsü Müdürü

ÖZET

METİN SINIFLANDIRMA İÇİN TERİM AĞIRLIKLANDIRMA

Turgut DOĞAN

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Donanımı Bilim Dalı

Eskişehir Teknik Üniversitesi, Lisansüstü Eğitim Enstitüsü, Ağustos 2019

Danışman: Dr. Öğr. Üyesi Alper Kürşat UYSAL

Metin sınıflandırma, metin dokümanlarının etiketleri önceden belirlenmiş sınıflara atanması işlevidir. İnternetin ve teknolojinin gelişimine paralel olarak elektronik ortama aktarılan metin dokümanlarının dramatik bir biçimde artması, söz konusu dokümanların hızlıca erişimi, organize edilmesi ve sınıflandırılması gibi işlevler metin sınıflandırmanın önemini daha da arttırmıştır. Metin sınıflandırmada efektif öznitelik vektör gösterimleri sınıflandırma performanslarını doğrudan etkileyebildiği için, metin içeriklerinden elde edilen özniteliklere (terimlere) uygun ağırlık değerlerinin atanması, önemli araştırma problemlerinden biridir. Literatürde bu araştırma problemine çözüm geliştirmeyi hedefleyen birçok terim ağırlıklandırma şeması önerilmiştir.

Bu tez çalışmasında, metin sınıflandırma için terim ağırlıklandırma problemleri ile popüler terim ağırlıklandırma şemalarının sundukları çözümler geniş kapsamlı olarak analiz edilmiş ve ağırlıklandırma problemlerine yönelik olarak çeşitli yeni çözümler önerilmiştir. Bu amaçla, ilk olarak yüksek terim frekansı değerlerinin ve çeşitli terim frekans faktörleri ile bu değerleri indirgemenin mevcut gözetimli terim ağırlıklandırma şemalarının performanslarına etkileri incelenmiştir. Bunun dışında, literatürde son yıllarda önerilmiş olan ters yer çekimi momentine bağlı olarak terim ağırlıklandırma şemasının bazı ekstrem senaryolara sahip terimlerin ayırt edicilik güçlerini daha makul bir biçimde yansıtabilen gelişmiş bir versiyonu önerilmiştir. Son olarak, metin sınıflandırma için, terimlerin geçmedikleri dokümanlardaki dağılım bilgilerini, ayırt ediciliklerini hesaplarken daha efektif bir biçimde kullanabilen; TF-MONO ve SRTF-MONO adında iki yeni ağırlıklandırma şeması önerilmiştir. Üç farklı popüler veri setinde iki farklı sınıflandırıcı kullanılarak, toplamda yedi farklı terim ağırlıklandırma şemasının kıyaslandığı deneylerden elde edilen sonuçlar; özellikle SRTF-MONO terim ağırlıklandırma şemasının diğerlerine nazaran daha başarılı olduğunu göstermiştir.

Anahtar kelimeler: Metin sınıflandırma, Terim ağırlıklandırma, Gözetimli terim ağırlıklandırma, Terim frekans faktörü, Koleksiyon frekansı faktörü.

ABSTRACT

TERM WEIGHTING FOR TEXT CLASSIFICATION

Turgut DOĞAN

Department of Computer Engineering

Program in Computer Hardware

Eskişehir Technical University, Institute of Graduate Programs, August 2019

Supervisor: Assist. Prof. Dr. Alper Kürşat UYSAL

Text classification is the process of assigning text documents to predefined categories. In parallel with rapid development of the Internet and technology, the volume of text documents which are transferred to electronic media has increased dramatically. Hence the importance of organization and classification of text documents and quick accessing to text documents have increased. Since effective vector representations can directly affect the classification performances in text classification, assigning appropriate weight values to the features extracted from text contents is one of the important research problems. Therefore, many term weighting schemes have been proposed in the literature aiming to develop solutions to this research problem.

In this thesis, general term weighting problems for text classification and proposed solutions with popular term weighting schemes are extensively analysed and various new solutions are proposed for weighting problems. For this aim, firstly, the effects of reducing high term frequency values with various term frequency factors on the performance of existing supervised term weighting schemes are investigated. In addition, an improved version of recently proposed term weighting approach based on inverse gravity moment has proposed for text classification. Proposed approach presents more reasonable representations for reflecting the discrimination power of terms on some extreme scenarios. Finally, two new term weighting schemes, namely TF-MONO and SRTF-MONO, are proposed for text classification. Proposed schemes can effectively use the distribution information of documents in which terms do not occur. The classification performances of proposed schemes are compared with five popular term weighting schemes by using two classifiers on the three benchmark datasets. Experiment results showed that SRTF-MONO has more successful classification results than other schemes.

Keywords: Text classification, Term weighting, Supervised term weighting, Term frequency factor, Collection frequency factor.

TEŞEKKÜR

Öncelikle, tez çalışması sürecinde, yurt dışında bulunduğu dönemler de dahil olmak üzere, deneyimlerini, bilgilerini ve desteğini en üst düzeyde hissettiğim çok iyi bir yol gösterici olduğuna inandığım danışman hocam Sayın Dr. Öğr. Üyesi Alper Kürşat UYSAL'a katkılarından ötürü en derin teşekkürlerimi sunarım.

Çalışmaya dayanak oluşturan tez izleme süreçleri boyunca tez izleme jürisinde yer alan hocalarım Sayın Prof. Dr. Serkan GÜNAL'a ve Sayın Doç. Dr. Semih ERGİN'e verdikleri fikirler ve öneriler yardımıyla yeni bakış açıları kazanmamı sağladıklarından dolayı çok teşekkür ederim.

Danışman hocamın yokluğunda, danışmanlık görevini üstlenerek gerek tecrübeleriyle gerekse desteğiyle motivasyonumu arttıran hocam Sayın Prof. Dr. Yaşar HOŞCAN'a da teşekkürü borç bilirim.

Tez yazım sürecinde verdikleri desteklerden ve gösterdikleri sabırdan ötürü aileme saygı ve sevgilerimi sunarım.

Turgut DOĞAN

Ağustos, 2019

20/08/2019

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 ilke 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.

Turgut DOĞAN

İÇİNDEKİLER

Sayfa

TABLOLAR DİZİNİ

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

SİMGELER VE KISALTMALAR DİZİNİ

BoW	:	Kelime Çantası (Bag-of-Word)
CHI2	:	Ki-Kare (Chi Square)
DF	:	Doküman Frekansı (Document Frequency)
DFS	:	Ayırt Edici Öznitelik Seçici (Distinguishing Feature Selector)
GI	:	Gini Katsayısı (Gini Index)
GR	:	Kazanım Oranı (Gain Ratio)
IDF	:	Ters Doküman Frekansı (Inverse Document Frequency)
ICF	:	Ters Sınıf Frekansı (Inverse Class Frequency)
ICSDF	:	Ters Sınıf Uzay Yoğunluk Frekansı (Inverse Class Space Density Frequency)
IG	:	Bilgi Kazancı (Information Gain)
IGM	:	Ters Yerçekimi Momenti (Inverse Gravity Moment)
IGM_{imp}	:	Gelişmiş Ters Yerçekimi Momenti (Improved Inverse Gravity Moment)
IR	:	Bilgi Erişimi (Information Retrieval)
KNN	:	K-En Yakın Komşu (K-Nearest Neighbor)
LOG_TF	:	Terim Frekansının Logaritması (Logarithm of Term Frequency)
MO	:	Maksimum Geçme

		(Maximum Occurrence)
MIDF	:	Modifiye Edilmiş Doküman Frekansı
		(Modified Inverse Document Frequency)
NN	:	Sinir Ağı
		(Neural Network)
NO	:	Geçmeme / Bulunmama
		(Non Occurrence)
PB	:	Olasılık Tabanlı
		(Probability-Based)
RF	:	İlgi Frekansı
		(Relevance Frequency)
SVM	:	Destek Vektör Makinesi
		(Support Vector Machine)
SQRT_TF	:	Terim Frekansının Kare-Kökü
		(Squared Root of Term Frequency)
TF	:	Terim Frekansı
		(Term Frequency)
TF-IDF	:	Terim Frekansı - Ters Doküman Frekansı
		(Term Frequency - Inverse Document Frequency)
TRR	:	Terim İlgi Oranı
		(Term Relevance Ratio)
VSM	:	Vektör Uzay Modeli
		(Vector Space Model)
WIDF	:	Ağırlıklandırılmış Ters Doküman Frekansı
		(Weighted Inverse Document Frequency)

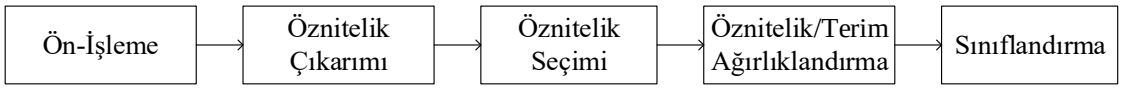
1. GİRİŞ

Teknolojideki hızlı gelişmelerin ve internetin de etkisiyle birlikte, elektronik ortama aktarılan verilerin hacmi dramatik bir biçimde artmaktadır. Bu verilerin büyük bir çoğunluğunu metinsel veriler oluşturmaktadır. Bu yüzden metin dokümanlarına etkin bir şekilde ulaşılabilmesini sağlamaya yönelik araştırmalar gün geçtikçe daha fazla önem kazanmaktadır. Bu araştırmalar içinde metinsel dokümanları organize etmek, indekslemek ve sınıflandırmak gibi işlevler önemli bir yere sahiptir. Bu durum araştırmacıların ilgisini metin sınıflandırma çalışmalarında yoğunlaştırmıştır.

Metin sınıflandırma, metinsel dokümanların, içeriklerine göre önceden belirlenmiş sınıflara atanmasıdır (Lan vd., 2009). Metin sınıflandırma kendi içinde çok geniş kapsamlı bir alan olup çeşitli alt çalışma alanlarına sahiptir. İstenmeyen e-posta filtreleme (Uysal vd., 2012), duygu analizi (Pak vd., 2011; Deng, Luo ve Yu, 2014; Singh ve Kumari, 2016), yazar tanıma (Sboev vd., 2016), medikal doküman sınıflandırma (Parlak ve Uysal, 2016) gibi çalışmalar metin sınıflandırma içerisinde yer alan araştırma alanlarına örnek olarak verilebilir.

1.1. Metin Sınıflandırma

Metin sınıflandırmanın metinsel içeriklere sahip dokümanların, içeriklerinin uygun bir biçimde işlenmesi yoluyla önceden belirlenmiş sınıflara atanma süreci olduğu giriş bölümünde belirtilmişti. Tipik bir metin sınıflandırma süreci; genel olarak, Şekil 1.1'de de gösterildiği gibi öznitelik çıkarımı, öznitelik seçimi, öznitelik (terim) ağırlıklandırma ve sınıflandırma aşamalarından oluşur. Metin sınıflandırma sürecinde yer alan tüm aşamalar ilerleyen alt bölümlerde detaylı olarak anlatılacaktır.



Şekil 1.1. *Metin Sınıflandırma Süreci*

Ön işleme aşaması; ham metin içeriklerinin, ait oldukları dilden bağımsız ve ilgili dile bağımlı birtakım önişleme metotları uygulanarak öznitelik çıkarma aşamasına hazır hale getirildiği aşamadır (Uysal ve Gunal, 2014).

Öznitelik çıkarımı aşamasında ham metin içerikleri sınıflandırıcıların işleyebileceği sayısal verilere dönüştürülürler. Bu aşamamın sonunda özgün öznitelikler elde edilir ve dokümanlar öznitelik vektörleri adı verilen sayısal değerlerle temsil edilirler.

Metin dokümanlarının doğasından dolayı, genellikle öznitelik çıkarımı aşaması sonucunda elde edilen öznitelik vektörlerinin boyutları çok yüksektir. Öznitelik seçimi aşamasında, ayırt ediciliği yüksek öznitelikler çeşitli öznitelik seçim yöntemleri ile seçilerek sınıflandırma aşamasından önce oluşturulan doküman-öznitelik vektörlerinin boyutu azaltılmaya çalışılır (Schneider, 2005; Uysal ve Gunal, 2012; Agnihotri, Verma ve Tripathi, 2017).

Vektör-Uzay-Modeli'nde; her eşsiz özniteliğin/terimin koleksiyondaki her bir doküman ile ilişkisi Kelime Çantası yaklaşımı (Aggarwal ve Zhai, 2012) ile temsil edilir. Metin sınıflandırma alanında bu ilişkileri gösteren sayısal değerlere ağırlık, ağırlık değerlerinin hesaplanması ve atanması işlemine ise terim ağırlıklandırma adı verilir. Doküman içeriklerinden elde edilen eşsiz terimlere uygun ağırlık değerlerinin atanması sınıflandırma performanslarını kayda değer bir biçimde etkileyebildiğinden, terim ağırlıklandırma, metin sınıflandırma üzerine çalışan birçok araştırmacı için, halen popüler bir çalışma alanı olarak önemini korumaktadır. Dolayısıyla da literatürde terim ağırlıklandırma için önerilmiş birçok yöntem/şema mevcuttur (Sparck Jones, 2004; Lan vd., 2009; Ren ve Sohrab, 2013; Chen vd., 2016; Dogan ve Uysal, 2019a).

Sınıflandırma aşamasında ise dokümanlar bir sınıflandırıcı yardımıyla kendilerine ait öznitelik vektörlerinin içerdikleri veriler ve etiket bilgileri kullanılarak uygun sınıflara atanırlar. Elektronik dokümanlar sayısal verilerle çalışabildiğinden, örüntü tanıma problemleri için kullanılabilen herhangi bir sınıflandırıcı metin sınıflandırma sürecine dahil edilebilir. Ayrıca uygun sınıflandırıcı seçimi de, sınıflandırma performansını önemli derecede arttırabilir.

1.2. Metin Sınıflandırmada Terim Ağırlıklandırma Problemleri

Metin sınıflandırmanın en önemli araştırma problemlerinden biri de çıkarılan özniteliklerin dokümanlarla ilişkisini vektör uzay modelinde (Manning, Raghavan ve Schütze, 2010) mümkün olduğu kadar iyi temsil edebilmektir. Bunun için öznitelik vektörlerinin makul bir biçimde yani dokümanları ayırt edebilecek şekilde

ağırlıklandırılması gerekir. Kullanılan ağırlıklandırma metodunun atadığı ağırlıklar ne kadar uygun olursa, metin sınıflandırma başarımı da o kadar arttırılabilir. Dolayısıyla efektif bir terim ağırlıklandırma yöntemi özniteliklerin taşıdığı bilgileri geniş kapsamlı bir şekilde analiz etmeli ve ağırlıklandırma sürecini bu analizlere bağlı olarak yönetmelidir. Ağırlıklandırma sürecinin en önemli basamağı her bir terim için hesaplanan, yer aldığı dokümanın ait olduğu sınıfı ayırt etme yeteneğini ifade eden ağırlık değerinin; o terimin gerçekte sahip olduğu ayırt etme potansiyelini mümkün olduğunca tutarlı bir biçimde yansıtmasıdır. Geçmişten günümüze konuyla alakalı olarak önerilen terim ağırlıklandırma şemalarının başarımları incelendiğinde; bunun yolunun, terimin yer aldığı dokümanın sahip olduğu sınıf bilgisini en iyi şekilde yansıtabilen ağırlıklandırma stratejisine sahip olmak olduğu söylenebilir.

Literatürde terim ağırlıklandırma ile ilgili son yıllarda başarılı şemalar önerilmiş olsa dahi, bu alanda halen daha yeni şemaların öneriliyor olması; terimlerin sahip oldukları ayırt etme potansiyelini daha iyi yansıtabilen ağırlıklandırma stratejileri olan yeni şemaların geliştirilebileceğinin ispatıdır. Terim ağırlıklandırma için önerilen yöntemler ne kadar güncel olursa olsun, her birinin ağırlık hesaplama sürecinde yetersiz kaldığı, göz ardı ettiği veya sahip olduğu ağırlıklandırma stratejisi yüzünden bazı ekstrem senaryolardaki terimler için makul ağırlıklar üretemediği durumlar mevcuttur.

1.3. Tezin Amacı ve Katkılar

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

I. Metin sınıflandırma performansını arttırmak için, terimler ile yer aldıkları dokümanlara ait sınıflar arasındaki ilişkileri daha iyi yansıtabilen bir gösterim nasıl elde edilebilir?
II. Metin sınıflandırma için geliştirilecek bir terim ağırlıklandırma şemasının sahip olması gereken özellikler neler olabilir?
III. Terim frekansı değerlerinin yüksek olduğu terimleri içeren veri setleri ile çalışırken, ağırlıklandırma sürecinde bu frekans değerlerinin indirgenmesinin metin sınıflandırma başarımına etkisi nedir?

IV. Terim ağırlıklandırma için önerilmiş bir şemanın metin sınıflandırma performansı, sahip olduğu ağırlıklandırma stratejisi geliştirilerek arttırılabilir mi?

V. Terimlerin geçmedikleri dokümanlara ait bilgiler ağırlıklandırma sürecinde daha efektif olarak kullanılabilir mi?

Yukarıda belirtilen araştırma problemlerine çözüm geliştirmek için hazırlanan bu tez çalışmasının ilk katkısı, literatürde mevcut olan gözetimli terim ağırlıklandırma şemalarında farklı terim frekans faktörlerinin kullanımının metin sınıflandırma başarımına etkisinin araştırılmasına yöneliktir. Bu çalışmada, gözetimli terim ağırlıklandırma şemalarının sınıflandırma performanslarının, terimlerin sınıf ayırt etme güçlerini verimli bir biçimde yansıtabilen yeni bir koleksiyon frekans faktörü geliştirip kullanmak kadar, uygun terim frekans faktörü seçimine de bağlı olduğu fikri savunulmuştur. 7 farklı gözetimli terim ağırlıklandırma şeması ile her birinin ağırlıklandırma sürecinde 3 farklı terim frekans faktörü kullanılarak en iyi kombinasyon belirlenmeye çalışılmıştır. Deneyde kullanılan yöntemlerden 6'sı literatürde hâlihazırda mevcut olup, biri ise ayırt edici öznitelik seçici (DFS) adlı öznitelik seçim yönteminden ilk kez terim ağırlıklandırmaya uyarlanmıştır. Biri dengeli biri de dengesiz yapıya sahip toplamda 2 metin veri seti üzerinde 2 farklı sınıflandırıcı ile gerçekleştirilen deneysel çalışmalardan elde edilen sonuçlar; terim frekans değerlerinin farklı fonksiyonlar (Logaritma fonksiyonu ve Karekök fonksiyonu) ile indirgenmesinin, neredeyse tüm terim ağırlıklandırma şemalarının performansını arttırdığını göstermiştir. Ayrıca terim frekans faktörü olarak ham terim frekansı değerlerinin kullanılmasının, bu değerleri indirgeyen diğer iki faktöre nazaran daha az verimli olduğu görülmüştür. Tüm bu analizlerin dışında; elde edilen performans farklarının olası nedenleri örnek bir koleksiyon üzerinde üç boyutlu vektör uzayında çeşitli görseller kullanılarak analiz edilmiştir.

Tez çalışmasının ikinci katkısı, terim ağırlıklandırma için önerilmiş gözetimli ve güncel yöntemlerden biri olan ters yerçekimi momentine (IGM) dayalı terim ağırlıklandırma yaklaşımının gelişmiş bir implementasyonuna yöneliktir. Bu çalışmada mevcut yöntemin kendi ağırlıklandırma yapısının az sayıdaki bazı ekstrem doküman frekansı özellikleri taşıyan terimler için yetersiz kaldığı fark edilmiş, yöntemin ağırlıklandırma formülüne bu yetersizliği en düşük seviyeye indirgemek için yeni bir oran (Ters Doküman Balans Frekansı, IDBF) eklenmiştir. Bu değişiklik yapılırken yöntemin

zaten başarılı ve makul olan (ekstrem olmayan) doküman frekansı özellikleri taşıyan terimleri ağırlıklandırma davranışındaki değişikliklerin de minimum düzeyde olması amaçlanmıştır. Söz konusu değişiklik ile gelişmiş ters yerçekimi momenti adıyla yeni bir koleksiyon frekans faktörü (IGM_{imp}) geliştirilmiş olup, bu faktöre bağlı olarak SQRT_TF-IGM_{imp} ve TF-IGM_{imp} adında 2 yeni terim ağırlıklandırma şeması önerilmiştir. Önerilen şemalar iki farklı dengeli ve bir dengesiz doküman dağılımına sahip olmak üzere toplamda 3 farklı metin koleksiyonu üzerinde 3 farklı sınıflandırıcı ile test edilmiştir. Önerilen şemalarının sınıflandırma performansları, standart IGM tabanlı şemalar haricinde literatürden 5 farklı terim ağırlıklandırma şemasıyla kıyaslanmıştır. Elde edilen sonuçlar, önerilen SQRT_TF-IGM_{imp} terim ağırlıklandırma şemasının standart IGM tabanlı şemalar (SQRT_TF-IGM ve TF-IGM) da dahil olmak üzere diğer tüm terim ağırlıklandırma şemalarına nazaran daha iyi sınıflandırma performansları sergilediğini göstermiştir. Ayrıca önerilen bir diğer yöntem olan TF-IGM_{imp}'de standart TF-IGM'den genel olarak daha üstün sınıflandırma başarımı değerlerine sahiptir. Ayrıca, her bir veri seti için seçilen öznitelik sayısı maksimum olduğunda her bir sınıflandırıcı ile tüm terim ağırlıklandırma şemalarından elde edilen sınıflandırma performansları da söz konusu deneysel çalışmada analiz edilmiştir. Bu analizin yapılmasındaki temel amaç önerilen şemaların yüksek boyutlardaki sınıflandırma performanslarını değerlendirmektir. Son olarak, söz konusu boyutlarda önerilen şemalardan en iyi performans değerlerine sahip olan SQRT_TF-IGM_{imp} ile literatürde mevcut olan diğer şemalar üzerinde elde edilen göreceli performans kazanımları, veri setleri ve sınıflandırıcılar bazında hesaplanmış ve analiz edilmiştir.

Tezin üçüncü ve en önemli katkısı ise terimlerin ayırt edicilik yeteneklerini hesaplarken; geçtikleri sınıf bilgilerini daha verimli biçimde kullanmanın yanı sıra, geçmediği doküman bilgisini de ağırlıklandırma sürecine etkin bir biçimde dahil etmeyi hedefleyen yeni bir ağırlıklandırma stratejisi geliştirmeye yöneliktir. Önerilen strateji; terimlere, maksimum geçme (Maksimum Occurrence, MO) ile geçmeme (Non Occurrence, NO) oranı adında iki orana bağlı olarak ağırlık ataması gerçekleştirmektedir. Çalışmada, metin sınıflandırma için MONO olarak adlandırılan bu stratejiyi temel alan TF-MONO ve SRTF-MONO isimli iki yeni terim ağırlıklandırma şeması önerilmiştir. Önerilen şemaların başarımları, 3 farklı veri seti üzerinde 2 farklı sınıflandırıcı kullanılarak literatürden 5 farklı terim ağırlıklandırma yönteminin başarımları ile

kıyaslanmıştır. Toplamda 7 farklı ağırlıklandırma şemasından elde edilen sınıflandırma performansları; SRTF-MONO terim ağırlıklandırma şemasının genel olarak diğerlerinden daha üstün performans gösterdiğini açıkça göstermektedir. Ayrıca TF-MONO terim ağırlıklandırma şeması da; Reuters-21578, 20-Newsgroups gibi geniş kapsamlı metin veri setleri üzerinde diğer beş terim ağırlıklandırma yöntemine nazaran umut verici sınıflandırma sonuçları sağlamıştır.

1.4. Tez Organizasyonu

Bu tez, organizasyon açısından toplam sekiz bölümden oluşmaktadır. 2. Bölüm metin sınıflandırma sürecini detaylı olarak olası tüm aşamalarıyla içermektedir. Konuyla alakalı olarak gerçekleştirilen çalışmalar ile deneylerde yararlanılan terim ağırlıklandırma şemaları sırasıyla 3. ve 4. Bölümlerde kabaca anlatılmıştır. 5. Bölümde bu tez çalışması kapsamında gerçekleştirilen ilk deneysel çalışma olan gözetimli terim ağırlıklandırma şemaları için farklı terim frekans faktörlerinin seçiminin, metin sınıflandırma performanslarına etkilerinin ölçüldüğü çalışma yer almaktadır. 6. Bölümde ise ters yerçekimi momentine dayalı koleksiyon frekans faktörünün (IGM) gelişmiş bir versiyonu olan IGM_{imp}, ve ona bağlı olarak önerilen şemalar kıyaslamalı performans sonuçları ile birlikte sunulmuştur. 7. Bölümde ise tez çalışmasının temelini oluşturan ve metin sınıflandırma için geliştirilen MONO adındaki yeni bir terim ağırlıklandırma yaklaşımı ve söz konusu yaklaşıma bağlı olarak önerilen iki farklı terim ağırlıklandırma şeması sunulmuştur. Son bölüm olan 8. Bölümde ise elde edilen sonuçlarla ilgili genel değerlendirmeler sunulmuştur.

2. METİN SINIFLANDIRMA SÜRECİNİN GENEL ÇERÇEVESİ

Tipik bir metin sınıflandırma sürecinin içerdiği aşamalar daha önceki bölümde genel olarak ifade edilmişti. Bu kısımda bahsi geçen aşamalar detaylı olarak açıklanmış ve kullanılan yaklaşımlar kabaca anlatılmıştır.

2.1. Ön İşleme

Metin sınıflandırma için en çok kullanılan ön işleme adımları, dokümanlarda yer alan metinsel içerikleri; dizgelere ayırma, küçük harfe dönüştürme, durak-kelimelerinden arındırma ve köklerine indirgeme şeklinde ifade edilebilir. Her bir ön işleme adımı ilerleyen alt bölümlerde detaylı olarak anlatılmıştır.

2.1.1. Dizgelere ayırma

Metin sınıflandırmada; dizgelere ayırma, metnin dizge (token) adı verilen kelimelere, öbeklere veya anlamlı parçalara ayrılması işlemidir. Bu ayırma işlemi genel olarak alfabetik veya alfa-numerik olarak gerçekleştirilir. Alfabetik dizgelere ayırmada metnin içerdiği nümerik ifadeler ayıklanırken, alfa-nümerik yaklaşım da ise nümerik terimler de öznitelik uzayına dahil edilip, boşluk, noktalama işaretleri gibi alfa-nümerik olmayan karakterler ayıklanırlar. Dizgelere ayırma işlemi, metnin ait olduğu dile bağımlı olarak değişebilir (Manning, Raghavan ve Schütze, 2010). Örneğin İngilizce dilinde hazırlanmış metin içeriklerini dizgelere ayırmak için ASCII olmayan karakterlerin ayıklanması yeterli olabilirken, Türkçe gibi evrensel olmayan kendine has karakterleri içeren metin dokümanları için farklı bir strateji izlemek gerekebilir. Tablo 2.1'de Türkçe olarak hazırlanmış basit bir cümlenin dizgelerine ayrılması gösterilmektedir.

Tablo 2.1. *Dizgelere ayırma ön işlemine bir örnek*

Dil	Cümle	Dizgeler (Virgüllerle ayrılmış)
Türkçe	Ali okula gitti.	Ali, okula, gitti

2.1.2. Durak-kelimeleri ayıklama

Durak-kelime kavramı, belirli bir sınıfa bağımlı olmayan, metinlerde sıklıkla karşılaşılabilecek terimleri veya kelimeleri (bağlaçlar, edatlar) ifade etmektedir. Bu yüzden genellikle metin sınıflandırma çalışmalarında ayırt edici olmayan terimler olarak değerlendirilir ve sınıflandırma aşamasından önce ayıklanırlar. Durak-kelime listeleri her ne kadar metinlerin ait oldukları dile bağımlı olsalar da, hiçbir dil için kesinleşmiş bir

durak-kelimeleri listesi mevcut değildir. Kullanılan durak-kelime listesi deneysel çalışmanın türüne veya amacına göre değişebilir. Türkçe ve İngilizce metin içeriklerinde yer alan bazı durak-kelimeler Tablo 2.2'de verilmiştir.

Tablo 2.2. *Örnek durak-kelime listeleri*

Dil	Durak-Kelimeler
Türkçe	ama, ancak, bile, böyle, dolayısıyla, her, ki, kim, olmak, sadece, ve, zaten
İngilizce	a, able, about, above, according, across, actually, after, are, at, before, then

2.1.3. Küçük harfe dönüştürme

Metin sınıflandırma için yaygın olarak kullanılan bir diğer ön işleme basamağı da küçük harfe dönüştürmedir. Terimlerin büyük veya küçük harfli biçimlerde bulunmasının ayırt edicilik hesabı için hiçbir farkının olmadığı düşünülen metin sınıflandırma çalışmalarında, dokümanların içerdiği tüm metinler küçük harfe dönüştürülür. Küçük harfe dönüştürme işlemi aynı kelimeleri grupladığından benzersiz özniteliklerin toplam sayısını azaltıcı etkiye sahiptir. Bu ön işleme basamağı da üzerinde çalışılan dilden dile farklılık gösterebilir. Aşağıdaki tablo bu tip olası farklılıklara örnek içermektedir.

Tablo 2.3. *İki farklı dil için bazı karakterlerin küçük harfe dönüştürülmüş biçimleri*

Özgün biçim	Küçük harfe dönüştürülmüş biçim (Türkçe)	Küçük harfe dönüştürülmüş biçim (İngilizce)
I	ı	i
U	u	u

2.1.4. Terimleri köklerine indirgeme

Metin dokümanlarının içerikleri, aynı kelime köküne sahip farklı şekilde türemiş çok sayıda terim içerebilirler. Metin sınıflandırma çalışmalarında bu tip terimler anlamsal olarak kök biçimlerine benzer olduğundan, çoğunlukla bir köke-indirgeme yaklaşımı ile köklerine indirgenirler. Köke indirgeme de üzerinden çalışılan dile göre farklılık gösterir. Bu sebeple her dil için uygulanan köke indirgeme algoritması farklıdır. Zemberek (Akın ve Akın, 2007) ve Porter (Porter, 1980) köke-indirgeme algoritmaları sırasıyla Türkçe ve İngilizce metinler için yaygın olarak kullanılan köke-indirgeme algoritmalarıdır.

Tablo 2.4. *Köke indirgemeye dair dil bazında örnekler*

Dil	Özgün Kelime Biçimi	Köke İndirgenmiş Biçim
Türkçe	gösterir	göster
İngilizce	indicates	indicate

2.2. Öznitelik Çıkarımı

Metin sınıflandırma çalışmalarının çoğunda, metin dokümanlarının geçme sıklıklarına dayanan, doküman içerisindeki terimlerin sırasının göz ardı edildiği kelime çantası (Bag Of Word, BoW) gösterimi kullanılır (Aggarwal ve Zhai, 2012). Bu gösterim biçiminde; bir doküman, içerdiği eşsiz terimlerle yani özniteliklerle temsil edilirler. Dolayısıyla bu gösterim biçiminde her bir doküman çok boyutlu öznitelik vektörlerine sahiptir. Dokümanların içeriklerinin çok boyutlu öznitelik vektörleri ile temsil edilmesi literatürde Vektör Uzay Modeli (Vector Space Model, VSM) olarak ifade edilir. (Manning, Raghavan ve Schütze, 2010). VSM, BoW gösteriminin gelişmiş versiyonu olarak kabul edilir. Doküman-öznitelik vektörleri oluşturulurken, her özniteliğin ile doküman arasındaki ilişki Eşitlik 2.1'deki gibi hesaplanan çeşitli ağırlık değerleri yardımıyla ifade edilir.

$$D_k = (W_{1,k}, W_{2,k,\ldots}, W_{i,k,\ldots}, W_{n,k}) \quad \textbf{(2.1)}$$

Burada $W_{n,k}$ değeri, *k.* dokümanın ayırt ediciliği için, *n* numaralı özniteliğin yaptığı katkının sayısal değerini ifade etmektedir. Bu ağırlık değerlerinin hesaplanması problemi metin sınıflandırma literatüründe terim ağırlıklandırma alt alanına girmektedir. Tezin esas amacı ve motivasyon alanı da terim ağırlıklandırma konusuyla ilgili olduğundan; bu konu ile ilgili detaylar, sırasıyla, öznitelik ağırlıklandırma alt alanı ve 3. Bölümde anlatılmıştır.

2.3. Öznitelik Seçimi

Öznitelik seçimi, genellikle metin sınıflandırma veya örüntü tanıma çalışmalarında; çıkarılan özniteliklerin çok yüksek boyutlara sahip olması, ayırt edici öznitelik alt kümelerinin ayıklanıp sınıflandırma performanslarının ve hesaplama sürelerinin arttırılması gibi hedefler söz konusu olduğunda sıklıkla başvurulan bir aşamadır. Literatürde öznitelik seçimi için yaygın olarak kullanılan yöntemlerden bazıları Ki-Kare (Manning, Raghavan ve Schütze, 2010), Bilgi Kazancı (Forman, 2003), Gini Katsayısı (Shang vd., 2013), Kazanım Oranı (Lan vd., 2009), Poisson dağılımından sapmalara dayalı öznitelik seçim metodu (Ogura, Amano ve Kondo, 2009) ve Ayırt Edici Öznitelik Seçici (Uysal ve Gunal, 2012), olarak ifade edilebilir. Deneysel çalışmalar esnasında yararlanılan Ki-Kare ile Ayırt Edici Öznitelik Seçici adlı öznitelik seçim yöntemleri ilerleyen alt bölümlerde daha detaylı bir biçimde anlatılmıştır.

2.3.1. Ki-Kare yöntemi (CHI2)

Ki-Kare (Chi-Square, CHI2) yöntemi bir öznitelik havuzunda yer alan çok sayıda öznitelik içinden, diğerlerine nazaran ait oldukları dokümanların sınıflarını daha ayırt edici olan belli sayıda özniteliği keşfedip ayıklamak için yaygın olarak kullanılan istatistiksel bir yöntemdir (Chen ve Chen, 2011). Metin sınıflandırma alanında; bu yöntem, bir öznitelik (terim) ile belirli bir sınıf arasındaki ilişkiyi ölçmek için sıklıkla kullanılır. Ölçülen değer ne kadar büyükse ilgili terimin söz konusu sınıfa bağımlılığı o kadar fazla olarak değerlendirilir. Metin sınıflandırma alanı için, bir t_i terimi ile C_j sınıfı arasındaki bağımlılığı ifade eden Ki-Kare değeri Eşitlik 2.2'deki gibi hesaplanır.

$$X^2(t_i, C_j) = \frac{D*(a_{ij}*d_{ij} + b_{ij}*c_{ij})^2}{(a_{ij}+b_{ij})(a_{ij}+c_{ij})(b_{ij}+d_{ij})(c_{ij}+d_{ij})} \quad \textbf{(2.2)}$$

Bu eşitlikte, a_{ij} ve c_{ij} sırasıyla t_i teriminin en az bir kez geçtiği C_j sınıfına ait olan ve C_j sınıfına ait olmayan toplam doküman sayılarını göstermektedir. Benzer şekilde b_{ij} ve d_{ij}, t_i teriminin geçmediği C_j sınıfına ait olan ve C_j sınıfına ait olmayan toplam doküman sayısını temsil etmektedir. D ile ifade edilen ise veri setindeki toplam doküman sayısıdır. Bu ifadeler ilerleyen bölümlerde anlatılacak öznitelik seçim yöntemleri ve terim ağırlıklandırma şemaları için de aynı anlamlara sahiptir.

2.3.2. Ayırt edici öznitelik seçici (DFS)

Deneysel çalışmalarda ayırt edici terimleri keşfetmek için kullanılan bir diğer öznitelik seçim yöntemi de DFS'tir. DFS, terimlerin koleksiyondaki sınıflarla ilişkilerini hesaplayarak, her bir terim-sınıf çifti için bir skor hesaplar ve atamasını gerçekleştirir (Uysal ve Gunal, 2012). DFS ile bir terim-sınıf ilişkisi yüksek ise bu durum terimin ilgili sınıfı temsil noktasında ayırt etme kabiliyetinin yüksek olduğu anlamına gelir, tam tersi durumlarda ise terimin ilgili sınıfı ayırt etme açısından yeteneğinin düşük olduğu anlamına gelmektedir. Metin sınıflandırma alanında, bir t_i terimi DFS ile ayırt edicilik skoru Eşitlik 2.3'deki gibi hesaplanır.

$$DFS(t_i) = \sum_{j=1}^{C} \left(\frac{\left(\frac{a_{ij}}{a_{ij} + c_{ij}} \right)}{\left(\frac{b_{ij}}{a_{ij} + b_{ij}} \right) + \left(\frac{c_{ij}}{c_{ij} + d_{ij}} \right) + 1} \right) \quad \textbf{(2.3)}$$

Bu eşitlikte yer alan a_{ij}, b_{ij}, c_{ij} *ve* d_{ij} sembollerinin anlam karşılıkları bir önceki alt bölümde açıklandığından, tekrar ifade edilmemiştir. Eşitlikte ayrıca yer alan *C*, veri setindeki toplam sınıf sayısını ifade etmektedir.

2.4. Öznitelik / Terim Ağırlıklandırma

Vektör Uzay Modelinde, doküman-öznitelik vektörleri oluşturulurken, her öznitelik ile doküman arasındaki ilişkinin çeşitli öznitelik ağırlıklandırma yöntemleri aracılığıyla hesaplanan çeşitli ağırlık değerleri yardımıyla temsil edildiği daha önce ifade etmişti.

Terimlere, yer aldıkları dokümanların sınıflarını ayırt edebilme açısından sahip oldukları potansiyeli en iyi şekilde yansıtabilen uygun ağırlık değerlerinin atanması; metin sınıflandırma performansını arttırma açısından hayati öneme sahiptir. Öyle ki, literatürde bu probleme yönelik olarak önerilmiş birçok terim ağırlıklandırma yaklaşımı mevcuttur.

Terim ağırlıklandırma şemaları (TAŞ, TWS) genel olarak gözetimsiz (Unsupervised) ve gözetimli (Supervised) yöntemler olarak iki gruba ayrılırlar (Sebastiani, 2002). Gözetimsiz yöntemler (Sparck Jones, 2004) ağırlık hesaplanması ve atanması sırasında terimlerin sınıf bilgilerini kullanmazken; gözetimli yöntemler (Ren ve Sohrab, 2013; Chen vd., 2016) ise sınıf bilgisini terim ağırlıklandırma sürecinde etkin biçimde kullanmaya çalışır.

Terim ağırlıklandırma şemaları, ikili sınıflandırmaya uygun ve çok-sınıflı sınıflandırmaya uygun yöntemler olarak da ikiye ayrılabilir. İkili-sınıflandırmaya uygun terim ağırlıklandırma şemaları (Debole ve Sebastiani, 2004; Liu, Loh ve Sun, 2009) çıkarılan her eşsiz terim için koleksiyondaki sınıf sayısı kadar ağırlık skoru üretirler. Ancak BoW yaklaşımında her terimin her bir doküman ile ilişkisi yalnızca tek bir ağırlık değeri ile temsil edilebildiğinden; bu tip şemaların ağırlıklandırma stratejilerinde, terim

için üretilen sınıf-bazlı ağırlık skorları, çeşitli globalleştirme metotları kullanılarak tek bir skora indirgeyen ekstra bir süreç mevcuttur (Dogan ve Uysal, 2018). Çoklu-sınıflandırmaya uygun terim ağırlıklandırma şemaları ise ağırlıklandırma yaparken, terimin koleksiyondaki tüm sınıflarla ilişkisini global bir biçimde temsil eden tek bir ağırlık skoru üretirler (Dogan ve Uysal, 2019a).

Bir terim ağırlıklandırma şeması (TAŞ) genellikle terim frekans faktörü, koleksiyon frekansı faktörü ve normalizasyon faktörü olmak üzere Eşitlik 2.4'te de belirtildiği gibi 3 farklı faktörün birleşiminden oluşur.

$$TAŞ = TerimFrekansFaktörü * KoleksiyonFrekansıFaktörü * NormalizasyonFaktörü \quad \textbf{(2.4)}$$

Terim frekans faktörü olarak, belirli bir terimin herhangi bir dokümandaki geçme sayısını esas alan terim frekansı (TF) yaygın olarak kullanılırken; koleksiyon frekansı, çoğunlukla ilgili terimin tüm veri koleksiyonundaki geçme bilgisini ifade eder. Normalizasyon faktörü ise daha çok Bilgi Erişimi (Information Retrieval, IR) çalışmalarında kullanılan, aşırı farklı boyutlarda dokümanlar içeren metin veri koleksiyonlarında çalışırken; terimlerin ağırlıklandırma sürecinde bu farklılıkları minimize etmeye yönelik bir ağırlık faktörü olarak ifade edilebilir (Yue-Heng, Pi-Lian ve Zhi-Gang, 2004).

Tablo 2.5'te, literatürdeki çeşitli terim ağırlıklandırma metotlarında kullanılan terim ve koleksiyon frekansı faktörlerine örnekler verilmiştir. Tabloda yer alan "----" ifadesi, söz konusu terim ağırlıklandırma metodunun ağırlıklandırma sürecinde o faktörün kullanılmadığını göstermektedir.

Tablo 2.5. *Literatürdeki bazı terim ağırlıklandırma şemaları*

TAŞ	Terim Frekans Faktörü	Koleksiyon Frekansı Faktörü
TF	Terim Frekansı (TF)	----
DF	----	Doküman Frekansı (DF)
TF-IDF	Terim Frekansı (TF)	Ters Doküman Frekansı (IDF)
TF-CHI2	Terim Frekansı (TF)	Globalleştirilmiş Ki-Kare Skoru (CHI2)
LogTF-RF	Terim Frekansının Logaritması	İlgi Frekansı (RF)
SqrtTF-IGM	Terim Frekansının Kare Kökü	Ters Yerçekimi Momenti (IGM)

2.5. Sınıflandırma

Sınıflandırma sürecinde, çıkarılan öznitelikler bir sınıflandırıcıya girdi olarak verilir. Metin sınıflandırmada en çok kullanılan sınıflandırıcılar Destek Vektör Makineleri (Support Vector Machines, SVM), K-En Yakın Komşu algoritması (K-Nearest Neighbor, KNN) ve Sinir Ağları (Neural Networks, NN) olarak sayılabilir. Bu bölümde deneylerde kullanılan SVM, KNN ve Rocchio isimli sınıflandırma algoritmaları anlatılmıştır.

2.5.1. Destek vektör makineleri (SVM)

SVM sınıflandırıcı, örüntü tanıma ve metin sınıflandırma problemlerinin çoğunda yaygın olarak tercih edilen başarılı bir sınıflandırma yaklaşımıdır (Sabbah vd., 2017). Tercih edilmesinin arkasında yatan en önemli sebeplerden biri de yüksek boyutlardaki öznitelik vektörlerini sınıflandırmada etkin bir performans sergileyebilmesidir. Öğrenme algoritması pozitif örnekleri negatif örneklerden ayırtmak için doğrusal veya doğrusal olmayan bir hiper-düzlem oluşturmaya dayanır. Bu hiper-düzlem, destek vektörleri isimli eğitim setindeki bazı örnekler sayesinde, negatif ve pozitif örnekler arasındaki mesafeyi maksimize eden konumda oluşturulur. Sınır-maksimizasyonu olarak adlandırılan bu konsept, SVM sınıflandırıcının karakteristik özelliklerini yansıtması açısından önemlidir.

Deneysel bölümde, çok-sınıflı sınıflandırmayı destekleyen LibSVM paketi varsayılan parametre değerleri ile kullanılmıştır (Chang ve Lin, 2011).

2.5.2. K-en yakın komşu algoritması (KNN)

KNN sınıflandırıcı, basit bir öğrenme algoritmasına sahip olan ve yaygın olarak tercih edilen bir sınıflandırma algoritmasıdır (Dogan ve Uysal, 2019a). Öğrenme algoritması, herhangi bir test dokümanının sınıfını, kendisine en yakın K sayıda komşu eğitim dokümanının sınıf bilgisine göre tahmin etmesine dayanmaktadır. Bu tahmin gerçekleştirilirken, ilgili test dokümanı ile komşuları arasındaki mesafe çeşitli mesafe ölçüm metrikleri ile ölçülür (Prasath vd., 2017). Test dokümanının hangi sınıfa atanacağına karar verilirken; söz konusu test dokümanına en yakın komşuların çoğunlukta olduğu sınıf dikkate alınır. Dolayısıyla K değerinin genellikle tek sayı olarak belirlenmesi beklenmektedir.

Deneylerde kullanılan mesafe ölçüm metrikleri ile K parametresi değerleri farklılık gösterdiğinden, her bir deney için açılan bölümlerde bu parametreler detaylı olarak verilmiştir.

2.5.3. Yapay sinir ağları (NN)

Yapay sinir ağları da örüntü tanıma (Fausett, 1994) ve metin sınıflandırma çalışmalarında yaygın bir biçimde başvurulan sınıflandırıcılardandır (Uysal ve Gunal, 2012). Yapay sinir ağlarının Perseptron ve çok-katmanlı yapay sinir ağları gibi birçok farklı tipi mevcuttur. Perseptron tipi çoğunlukla doğrusal sınıflandırma problemlerinde kullanılırken, çok-katmanlı yapay sinir ağları ise hem doğrusal hem de doğrusal olmayan sınıflandırma problemlerinde tercih edilmektedir. Basit bir sinir ağı genellikle giriş katmanı, gizli katman(lar) ve çıkış katmanı olmak üzere üç katmandan oluşur. Sinir hücresi olarak adlandırılan basit bir yapıya sahip olan işlem elemanları gizli katmanda bulunur ve birbirlerine ağırlıklandırılmış bağlarla bağlıdır. Gizli katman sayısı, yapay sinir ağları ile çözülecek sınıflandırma problemine göre değişkenlik gösterebilir.

2.5.4. Rocchio sınıflandırıcı

Rocchio sınıflandırıcı, dokümanları temsil eden öznitelik vektörlerinin ağırlık merkezlerinin hesaplanarak her bir test dokümanının o merkezlere yakınlığına göre sınıf atamasını gerçekleştiren bir algoritmaya sahiptir. Rocchio sınıflandırıcının sınıflandırma algoritması kabaca şu şekilde ifade edilebilir:

Öncelikle aynı sınıfa ait olan tüm eğitim dokümanlarının öznitelik vektörlerinin toplamı bulunur. Hesaplanan sınıf tabanlı bu toplam değerler, o sınıflara ait doküman sayılarına bölünerek her spesifik sınıf için ağırlık merkezi hesaplanır. Sınıfı belirlenecek olan test dokümanı ile tüm sınıf-spesifik ağırlık merkezleri arasındaki mesafeler hesaplanır. Son olarak ilgili test dokümanı kendisine en yakın (en düşük mesafeye sahip) ağırlık merkezinin ait olduğu sınıfa atanır.

3. İLGİLİ ÇALIŞMALAR

Terim ağırlıklandırma ile ilgili literatürde yer alan çalışmalar kronolojik bir biçimde sıralandığında; listenin ilk sırasında TF-IDF (Terim frekansı & Ters Doküman Frekansı) terim ağırlıklandırma şemasını görmek mümkündür. Bilgi Erişimi (Information Retrieval, IR) çalışmalarında efektif sorgu kelimelerinin keşfedilebilmesi için geliştirilmiş olan bu metot, daha sonradan Karen Spärck Jones tarafından metin sınıflandırma problemlerinde terim ağırlıklandırma için yeniden uyarlanmıştır (Sparck Jones, 2004). Karen, daha spesifik olan terimleri efektif bir biçimde kullanabilmek için, terimleri ağırlıklandırırken; terim frekansı dışında koleksiyon frekansı (IDF) değerlerinin de ağırlıklandırma sürecine dahil edilmesi gerektiğini, çeşitli metin koleksiyonları üzerinde gerçekleştirdiği test deneyleri ile savunmuştur. IDF tabanlı bir başka çalışmada ise metin sınıflandırma için Modifiye edilmiş ters doküman frekansı (Modified Inverse Document Frequency, MIDF) adında bir terim ağırlıklandırma şeması önerilmiş ve önerilen şemanın performansı, standart TF-IDF ve ağırlıklandırılmış IDF (Weighted IDF, WIDF) şemalarıyla karşılaştırılmıştır (Deisy vd., 2010). Aynı çalışmada, önerilen MIDF ağırlıklandırma şemasının genel olarak kıyaslanan diğer şemalara göre daha başarılı olduğu ve söz konusu önerilen şemanın ağırlıklandırma hesabının, yapısı itibariyle diğerlerinden daha kolay olduğu da vurgulanmıştır. Sabbah vd. web sayfası sınıflandırma için standart TF ve IDF şemalarından uyarlanan mTF, mTFIDF, TFmIDF ve mTFmIDF terim ağırlıklandırma şemalarını önermişlerdir (Sabbah vd., 2017). Önerilen şemalardan üçünün sınıflandırma başarımlarının; TF, TF-IDF ve Entropi gibi terim ağırlıklandırma şemalarınınkinden daha iyi olduğunu vurgulamışlardır.

Literatürde terimlerin IDF bilgisi yanı sıra sınıf bilgilerinin de terim ağırlıklandırma sürecine dahil edildiği çeşitli çalışmalar mevcuttur. Onlardan biri de Lertnattee ve Theeramunkong isimli araştırmacıların önerdiği Ters sınıf frekansı (Inverse Class Frequency, ICF) bilgisinin üç farklı biçiminin, standart TF-IDF ile birleştirildiği ve ICF bilgisinin uygun formlarının araştırıldığı çalışmadır. Araştırmacılar, en uygun formları bulabilmek için; deneylerini, farklı n-gram modelleri, az sayıda sınıf içeren metin koleksiyonları ve farklı sayıda eğitim setlerini kullandıkları çeşitli deneysel parametreler ile gerçekleştirmişlerdir (Lertnattee ve Theeramunkong, 2004). Ren ve Sohrab, metin sınıflandırma için, terimlerin TF-IDF bilgilerine ek olarak, sırasıyla ICF ile ters sınıf uzay yoğunluk frekansı (Inverse Class Space Density Frequency, ICSDF) bilgilerini de

ağırlıklandırma sürecinde kullanan TF-IDF-ICF ve TF-IDF-ICSDF isimli iki yeni terim ağırlıklandırma şeması önermiştir (Ren ve Sohrab, 2013). Bahsi geçen araştırmacılar, hem sık geçen hem de nadiren geçen terimler için pozitif bir ayırt edicilik sağması sebebiyle özellikle TF-IDF-ICSDF terim ağırlıklandırma şemasının unut vadeden sonuçlara sahip olduğunu vurgulamışlardır.

Öznitelik seçim yöntemlerinin, ayırt edici terimlerin keşfedilme, terimlerin sınıf bilgilerini verimli biçimde kullanabilme ve metin sınıflandırma başarımını dikkate değer bir biçimde arttırabilme yeteneğinin yapılan araştırmalarla fark edilmesi, terim ağırlıklandırma çalışan araştırmacıların ilgisini, bu metotları terim ağırlıklandırmaya uyarlama noktasında cezbetmiştir. Nitekim Debole ve Sebastiani Ki-Kare, Bilgi Kazancı, Kazanım Oranı gibi öznitelik seçim yöntemlerinden uyarladığı gözetimli TF-CHI2, TF-IG, and TF-GR terim ağırlıklandırma şemalarının standart TF-IDF'ten daha başarılı olduklarını göstermişlerdir (Debole ve Sebastiani, 2004). Sınıf bilgisini terim ağırlıklandırmada kullanan ve kullanmayan yöntemlerin kıyaslandığı bu çalışmada, "gözetimli ve gözetimsiz terim ağırlıklandırma şeması" kavramları ilk kez vurgulanmıştır. Benzer bir çalışma da Deng vd. tarafından yapılmış olup, öznitelik seçim tabanlı terim ağırlıklandırma şemalarının da içinde olduğu 4 farklı terim ağırlıklandırma şemasının (TF-IDF, TF-CRF, TF-OddsRatio ve TF-CHI2) performansları karşılaştırılmış ve genel olarak en iyi performansı TF-CHI2'in sağladığı ifade edilmiştir (Deng vd., 2004).

Bu kısımda derlenen terim ağırlıklandırma şemaları, ağırlıklandırma sürecinde, terimlerin pozitif ve negatif sınıflarda bulunma-geçme olasılıklarının hesabından yararlanmaktadırlar. Nitekim Lan vd. tarafından önerilen TF-RF terim ağırlıklandırma şeması, İlgi Frekansı (Relevance Frequency, RF) adında terimlerin pozitif ve negatif sınıflarda geçme olasılıklarının hesabına dayalı bir orana sahiptir (Lan vd., 2005; Lan, Tan ve Low, 2006). Söz konusu çalışmalarda, araştırmacılar, TF-RF ile elde edilen metin sınıflandırma performansının kıyaslanan diğer şemalara nazaran daha iyi olmasının; sınıf bilgisini kendinden önce önerilen öznitelik seçim tabanlı terim ağırlıklandırma şemalarından ve standart TF-IDF şemasından daha etkin biçimde kullanmasından kaynaklandığını iddia etmişlerdir. Xuan ve Quang ise TF değerlerini Logaritma fonksiyonu ile indirgediği çalışmasında LogTF-RF şemasını önermiş ve yüksek terim frekansı değerlerinin gürültüye sebep olduğunu ve terim ağırlıklandırma şemasının sahip

olduğu sınıfı ayırt edebilme yeteneğini tam manasıyla yansıtmadığını bunları indirgemenin metin sınıflandırma performansını arttırabileceğini göstermiştir (Xuan ve Le Quang, 2014). Ko ise sınıf bilgisini kullanan ve daha önce önerilmiş geleneksel şemalardan daha başarılı bir performans sergilediğini çeşitli deneylerle gösterdiği Log-TF-TRR terim ağırlıklandırma şemasını önermiştir (Ko, 2015). Emmanuel vd. Classic3 metin koleksiyonu üzerinde çeşitli deneyler gerçekleştirdiği çalışmasında, Pozitif Etki Faktörü (Positive Impact Factor, PIF) adında yeni bir terim ağırlıklandırma yaklaşımı önermişlerdir (Emmanuel, Khatri ve Babu, 2013). Araştırmacılar, PIF'in diğer kıyaslandığı terim ağırlıklandırma şemalarına nazaran metin sınıflandırma performansını arttırdığına ve hesaplama süresini azalttığına dikkat çekmişlerdir. Pozitif ve negatif sınıf olasılıkları hesabına dayalı olarak terim ağırlıklandırmayı temel alan bir diğer çalışmada (Liu, Loh ve Sun, 2009) ise Liu vd. dengesiz veri setleri için TF-PB terim ağırlıklandırma şemasını önermişlerdir. Söz konusu çalışmada, TF-PB'nin ağırlıklandırma sürecine, terimlerin sınıflar-arası sınıf dağılım bilgilerinin yanı sıra sınıf-içi dağılım bilgileri de eklendiğinden; daha az dokümana sahip azınlık sınıfların vektör uzay modelinde diğer 7 şemaya (gözetimsiz ve gözetimli) göre daha iyi temsil edilebildiği ifade edilmiştir.

Altınçay ve Erenel, metin sınıflandırma için önerilen terim ağırlıklandırma şemalarının performans farklılıklarını araştırdıkları çalışmada (Altınçay ve Erenel, 2010), bu farklılıkların, terim ağırlıkları üretmedeki terim-geçiş farklılıklarından ve farklı oran kullanımından kaynaklandığını ifade etmişlerdir. Bir başka çalışmada ise, aynı araştırmacılar, terim frekanslarının logaritmalarına dayalı bir koleksiyon frekans faktörü önermişlerdir (Erenel ve Altınçay, 2012). Terim frekanslarının küçük olduğu durumlarda, önerilen koleksiyon frekansının metin sınıflandırma için ağırlıklandırma sürecinde etkin bir rol oynadığını vurgulamışlardır. Terimlerin ayırt ediciliklerini daha iyi yansıtmaya yönelik olarak, daha etkin doküman vektörleri oluşturmaya dayanan bir başka çalışmada, Badawi ve Altınçay, tarafından BoW yaklaşımını temel alan bir gösterim biçimi önerilmiştir (Badawi ve Altınçay, 2014). Aynı araştırmacılar bir diğer çalışmalarında (Badawi ve Altınçay, 2017), kardinalite istatistiğine (Cardinality Statistics) bağlı olarak terim setlerini ağırlıklandıran bir terim ağırlıklandırma şeması önermişlerdir. İlgili çalışmada; n-terim setleri (n=2, 3, 4) ile ağırlıklandırma yapan önerilen şemanın, BoW tabanlı gösterim yaklaşımının metin sınıflandırmadaki performansını arttırdığını gösterilmiştir.

Önerilen terim ağırlıklandırma şemalarının bazıları, mevcut herhangi sınıflandırıcının sınıflandırma tarafındaki kısıtlarını aşarak performansını arttırmaya yönelik de olabilmektedir. Bu bağlamda, Jiang vd. öznitelik ağırlıklarını, şartlı olasılık dağılım tahminlerine dahil eden terim ağırlıklandırma çalışmasında, Naïve Bayes sınıflandırıcı için Derin Öznitelik Ağırlıklandırma (Deep Feature Weighting, DFW) adında bir öznitelik ağırlıklandırma yaklaşımı önermiştir (Jiang vd., 2016). Benzer şekilde Zhang vd. Naïve Bayes sınıflandırıcı için, kıyaslandığı diğer şemalara nazaran daha fazla sınıflandırma doğruluğuna, daha az yürütme zamanına ve daha basit bir yapıya sahip olan 2 farklı uyarlanabilir öznitelik ağırlıklandırma yaklaşımı önermişlerdir (Zhang vd., 2016). Kim ve Kim ise ikili Naïve Bayes model (Binary Naïve Bayes model) ve Çok-katlı terim modelini (Multinomial term model) temel alan model tabanlı bir terim ağırlıklandırma yaklaşımı önermişlerdir (Kim ve Kim, 2016). Nguyen vd. tarafından, sınıf öznitelik ağırlık-merkezi (class feature centroid CFC) sınıflandırıcının bazı problemli durumları ve kısıtlarıyla başa çıkabilmek için Kullback-Leibler ve Jensen-Shannon ıraksamasına bağlı olarak geliştirilmiş KL ve JS terim ağırlıklandırma şemaları önerilmiştir (Nguyen, Chang ve Hui, 2013). Söz konusu çalışmada, araştırmacılar; terim ağırlıklandırmanın, bu ıraksama metotlarına bağlı bir biçimde gerçekleştirilmesi sayesinde, CFC sınıflandırıcının performansını dikkate değer bir biçimde arttırdığını ifade etmişlerdir.

Chen vd. istatistiksel bir model olan Ters Yerçekimi Momentini (Inverse Gravity Moment, IGM) hesaplamaya dayalı olan TF-IGM terim ağırlıklandırma şemasını önermişlerdir (Chen vd., 2016). Önerilen TF-IGM şemasının, literatürden diğer 6 farklı şema (TF, TF-IDF, TF-IDF-ICSDF, TF-CHI, TF-PB ve TF-RF) ile performanslarının kıyaslandığı çalışmada; terimlerin sınıflar-arası dağılımları ile daha verimli hesaplandığı, ve sahip oldukları ayırt edicilik yeteneklerinin, VSM'de TF-IGM ile daha iyi yansıtıldığı vurgulanmıştır. Karşılaştırıldığı diğer şemaların aksine, TF-IGM'in ağırlıklandırma formülü, üzerinde çalışılan veri setinin dengeli veya dengesiz özelliklere sahip olmasına göre ayarlanabilen ekstra bir parametre içermektedir.

Terim ağırlıklandırma ile ilgili gerçekleştirilen diğer çalışmalar, şöyle kabaca özetlenebilir: Luo vd. WordNet kullanarak terimleri indeksleyen ve sınıfların semantik bilgilerini kullanan semantik bir terim ağırlıklandırma şeması önermişlerdir (Luo, Chen ve Xiong, 2011). Bahsi geçen çalışmada, önerilen şemanın sınıflandırma performansının,

farklı karakteristiklere sahip bazı eğitim setleri için TF-IDF'ten daha üstün olduğu gösterilmiştir. Fattah, otomatik duygu sınıflandırma için sınıf yoğunluk frekansı hesabına dayanan yeni terim ağırlıklandırma şemaları önermiştir (Abdel Fattah, 2015). Fattah, önerilen şemaların bazılarının nadir ve sık geçen terimler üzerinde diğer karşılaştırılan şemalara nazaran daha etkin bir ayırt edicilik sağladığını göstermiştir. Sabbah vd. çeşitli web sayfalarındaki terör faaliyetlerini bulma ve sınıflandırmada bireysel yöntemlerden daha etkin özelliklere sahip olan ve temel terim ağırlıklandırma şemalarını kullanmaya dayanan hibrit bir öznitelik seçim yaklaşımı önermişlerdir (Sabbah vd., 2016). Alsmadi ve Hoon ise iki farklı Twitter veri seti ile çeşitli deneyler gerçekleştirdiği çalışmalarında (Alsmadi ve Hoon, 2017), kısa metin sınıflandırma alanına yönelik olarak SW adlı gözetimli terim ağırlıklandırma yaklaşımını önermişlerdir. Araştırmacılar, SW şemasının kısıtlı kısa metin uzaylarındaki değerli bilgileri bulma ve işleme açısından, diğer gözetimsiz ve gözetimli şemalara nazaran daha iyi performans sağladığını vurgulamışlardır. Haddoud vd., çok sayıda farklı terim ağırlıklandırma metriklerini ele alarak, metin sınıflandırma için genişletilmiş terim gösterim ve sınıflandırıcı birleştirme metodolojisi önermişlerdir (Haddoud vd., 2016).

Terim ağırlıklandırma ile ilgili son yıllarda yayınlanan araştırma çalışmalarını ise birer cümle ile şu şekilde özetleyebiliriz: Rao vd. vuku bulan çeşitli olaylarla ilgili olarak ilk olay hikayesini belirlemeyi (First Story Detection) amaçlayan LGT isimli bir terim ağırlıklandırma şemasını önermişlerdir (Rao vd., 2017). Feng vd. kısa metin sınıflandırma için olasılıksal bir modele dayanan lrp isimli terim ağırlıklandırma şemasını önermişlerdir (Feng vd., 2018). Matsuo ve Ho ise klinik metinsel dokümanlar için; terimlere, semantik ilişkilerine göre ağırlıklar atayan iki aşamalı bir terim ağırlıklandırma yaklaşımı önermiştir (Matsuo ve Ho, 2018). Li vd. tarafından ise etkin topik modelleme (topic modelling) için mevcut iki terim ağırlıklandırma şemasının Entropi ağırlıklandırma ile bir tür kombinasyonu önerilmiştir (Li vd., 2018). Son olarak ise Santhanakumar vd., tüm dokümanlarda ortak geçen çoklu-terimleri ağırlıklandırmaya dayanan ortak-terim frekansı terim ağırlıklandırma metodunu önermiştir (Santhanakumar, Columbus ve Jayapriya, 2018).

4. MEVCUT TERİM AĞIRLIKLANDIRMA ŞEMALARI

Deneysel çalışmalar esnasında kullanılan, literatürde yer alan bazı terim ağırlıklandırma şemaları bu bölümde anlatılmıştır.

4.1. Geleneksel Terim Ağırlıklandırma Şemaları

4.1.1. Terim frekansı (TF)

Terim frekansına bağlı terim ağırlıklandırmada, her bir terimin ağırlık değeri dokümanda geçme/bulunma sayısına eşittir. Ayrıca, terimleri ağırlıklandırma sürecinde; terim frekans faktörü olarak TF değerlerini kullanma, literatürdeki terim ağırlıklandırma şemalarının en çok başvurduğu işlevler arasındadır. Terimlerin sınıf bilgisi ağırlıklandırma hesabında kullanılmadığından gözetimsiz terim ağırlıklandırma şemaları grubuna girer. TF şemasının ağırlıklandırma stratejisi Eşitlik 4.1'de verilmiştir.

$$TF(t_i, d_k) = \begin{cases} t_i \text{ teriminin } d_k \text{ dokümanında geçme sayısı}, & \text{eğer } t_i \text{ terimi } d_k \text{ dokümanında geçiyorsa} \\ 0, & \text{aksi takdirde} \end{cases} \quad \textbf{(4.1)}$$

4.1.2. Terim frekansı & ters doküman frekansı (TF-IDF)

TF-IDF, terim ağırlıklandırma şemaları arasında en yaygın olanların başında gelmektedir. Başlangıçta IR çalışmaları kapsamında efektif sorgu kelimelerini bulmak amacıyla önerilmiş olan bu yöntem (Sparck Jones, 2004) daha sonra terim ağırlıklandırmaya da uyarlanmıştır (Sparck Jones, 2004). Koleksiyon frekansı faktörü olarak terimlerin ters doküman frekansı değerlerinin hesaplandığı bu yöntemde koleksiyonda nadir geçen (az sayıda dokümanda geçen) terimlere yüksek skorlar atanır. Eşitlik 4.2'de herhangi bir t_i teriminin, TF-IDF şeması ile ağırlıklandırma formülü gösterilmektedir.

$$W_{TF.IDF}(t_i) = TF(t_i, d_k) * \log(\frac{D}{d(t_i)}) \quad \textbf{(4.2)}$$

Eşitlikte yer alan D değeri koleksiyondaki toplam doküman sayısını, $d(t_i)$ ise t_i teriminin geçtiği toplam doküman sayısını ifade etmektedir. TF-IDF ile ağırlıklandırma sürecinde terimlerin sınıf bilgilerini kullanmaz. Dolayısıyla, literatürde gözetimsiz terim ağırlıklandırma şeması olarak kabul edilir.

4.2. Öznitelik-Seçim Metriklerine Bağlı Terim Ağırlıklandırma

Bu alt bölümde, deneysel çalışma bölümlerinde CHI2 ve DFS öznitelik seçim yöntemleri kullanılarak türetilen ve kullanılan TF-CHI2 ve TF-DFS terim ağırlıklandırma şemaları anlatılmıştır. İki ağırlıklandırma yöntemi de terimlerin sınıf bilgilerini kullanarak ağırlıklandırma yaptığından gözetimli terim ağırlıklandırma şemaları olarak nitelendirilirler.

4.2.1. Ki-kare istatistiğine bağlı terim ağırlıklandırma (TF-CHI2)

Ki-Kare öznitelik seçim metodunun özellikleri, ikinci bölümde öznitelik seçim metotları kısmında kabaca anlatılmıştı. Bu bölümde bu istatistiksel seçim metodunu kullanarak terimlere ağırlık hesabı ve atamasının nasıl gerçekleştirildiği ifade edilmiştir (Debole ve Sebastiani, 2004; Deng vd., 2004). Metin sınıflandırma alanında (domaininde) bir t_i terimi için Ki-Kare yöntemine bağlı ağırlık hesabı Eşitlik 4.3'deki gibi gerçekleştirilir.

$$W_{TF.CHI2}(t_i) = TF(t_i, d_k) * \mathrm{D} * \max_{j=1}^{C} \left\{ \frac{\left(a_{ij} * d_{ij} - b_{ij} * c_{ij}\right)^2}{(a_{ij} + c_{ij})(b_{ij} + d_{ij})(a_{ij} + b_{ij})(c_{ij} + d_{ij})} \right\} \qquad \textbf{(4.3)}$$

Eşitlikte yer alan max ifadesi Ki-Kare metodu ile t_i terimine yönelik olarak koleksiyondaki her spesifik *j*. sınıf için hesaplanan ağırlık değerlerinden maksimum olanının atanacağını ifade etmektedir. Geriye kalan ifadelerin anlamları, daha önceki eşitliklerde geçtiğinden ve açıklandığından tekrar ifade edilmemiştir.

4.2.2. Ayırt edici öznitelik seçim metodu ile terim ağırlıklandırma (TF-DFS)

DFS metodu da daha önce, deneylerde kullanılan öznitelik seçim metotlarının anlatıldığı ikinci bölümde kabaca anlatılmıştı. Bu kısımda ifade edilmesi gereken bir diğer önemli konu ise, DFS adlı öznitelik seçim yönteminin ilk kez bu tez çalışmasında terim ağırlıklandırma için uyarlanmış ve kullanılmış olmasıdır (Dogan ve Uysal, 2019b). Bu uyarlama, terim frekans faktörü olarak TF değerlerinin, koleksiyon frekans faktörü olarak ise DFS skorlarının hesaplanmasını ve çarpımını temel almaktadır. Bir t_i teriminin üzerinde çalışılan tüm veri seti için ayırt ediciliğini ifade eden DFS tabanlı ağırlıklandırma formülü Eşitlik 4.4'te verilmiştir.

$$W_{TF.\mathrm{DFS}}(t_i) = TF(t_i, d_k) * \sum_{j=1}^{C} \left(\frac{\left(\frac{a_{ij}}{a_{ij} + c_{ij}} \right)}{\left(\frac{b_{ij}}{a_{ij} + b_{ij}} \right) + \left(\frac{c_{ij}}{c_{ij} + d_{ij}} \right) + 1} \right) \quad \textbf{(4.4)}$$

4.3. İkili (Binary) Yaklaşıma Sahip Terim Ağırlıklandırma Şemaları

Bu bölümde, terimlere ağırlık hesaplarken, ikili (Binary) sınıflandırma yaklaşımına sahip olan, deneylerde kullanılan TF-RF ve TF-PB terim ağırlıklandırma yöntemleri açıklanmıştır. İkili sınıflandırma yaklaşımından kasıt, ağırlık hesaplarken; bu iki yöntemin de terimlerin pozitif ve negatif dağılım bilgilerini içeren çeşitli oranlardan yararlanmalarıdır.

4.3.1. İlgi frekansına bağlı terim ağırlıklandırma (TF-RF)

TF-RF, ilgi frekansı (Relevance Frequency, RF) adında terimlerin sınıflar-arası dağılım bilgilerini hesaplamaya yönelik bir oran yardımıyla ağırlıklandırma yapan gözetimli bir terim ağırlıklandırma şemasıdır. RF'in terim ağırlıklandırma stratejisi, özellikle, terimin pozitif ve negatif sınıflarda geçme oranına odaklıdır. TF-RF tabanlı terim ağırlıklandırma Eşitlik 4.5'teki gibi gerçekleştirilmektedir.

$$W_{TF.RF}(t_i) = TF(t_i, d_k) * \max_{j=1}^{C} \left\{ \log \left(2 + \frac{a_{ij}}{\max(1, c_{ij})} \right) \right\} \quad \textbf{(4.5)}$$

Eşitlikte, logaritma parametreleri içinde yer alan 2 değeri ile, a_{ij}/c_{ij} oranından 0 gelmesi durumunda logaritmanın sonsuz değerini döndürmeyip RF'in minimum 1 değerini döndürmesini; dolayısıyla da terimlerin RF'den en düşük skorları alması halinde dahi TF değerleri ile ağırlıklandırılmasını sağlamak amacıyla yer almaktadır.

4.3.2. Olasılık dağılımlarına bağlı terim ağırlıklandırma (TF-PB)

TF-PB, terimleri ağırlıklandırırken, geçtikleri pozitif ve negatif sınıflardaki olasılık dağılımları hesabına dayanan iki önemli oranı (a_{ij}/b_{ij} and a_{ij}/c_{ij}) kullanan gözetimli bir terim ağırlıklandırma şemasıdır (Liu, Loh ve Sun, 2009). Bu oranlardan, a_{ij}/b_{ij} oranı terimin sınıf-içi dağılımını ifade ederken, a_{ij}/c_{ij} oranı ise terimin sınıflar-arası dağılımını ifade etmektedir. TF-PB ile terim ağırlıklandırma için kullanılan ağırlıklandırma formülü Eşitlik 4.6'da gösterilmiştir.

$$W_{TF.PB}(t_i) = TF(t_i, d_k) * \max_{j=1}^{C} \left\{ \log\left(1 + \frac{a_{ij}}{b_{ij}} * \frac{a_{ij}}{c_{ij}} \right) \right\} \quad \textbf{(4.6)}$$

Diğer şemalardan farklı olarak, TF-PB, terimlerin sınıf-içi dağılım bilgilerini de içerdiğinden, dengesiz metin veri setlerinde, dengeli yapıda olan metin veri setlerine nazaran daha başarılı bir performans sergilemektedir.

4.4. Sınıf Bilgisini İndekslemeye Dayalı Terim Ağırlıklandırma Şemaları

Bu bölümde sınıf bilgisini indekslemeye dayalı olarak ağırlıklandırma yapan, deneysel bölümde kullanılan TF-IDF-ICF ile TF-IDF-ICSDF terim ağırlıklandırma şemaları anlatılmıştır.

4.4.1. Ters sınıf frekansına dayalı terim ağırlıklandırma (TF-IDF-ICF)

IDF tabanlı olan bu gözetimli terim ağırlıklandırma şemasında koleksiyon frekans faktörü tarafında IDF haricinde, terimlerin Ters Sınıf Frekansı (Inverse Class Frequency, ICF) bilgisi de hesaplanıp ağırlıklandırma sürecine dahil edilir (Ren ve Sohrab, 2013). Her bir terim için koleksiyondaki sınıf sayısı kadar değil de, koleksiyondaki bütün sınıflarla ilişkisi hesaplanarak tek bir skor ürettiğinden, çoklu-sınıflandırmaya uygun terim ağırlıklandırma şemaları grubuna girer. TF-IDF-ICF ile herhangi bir t_i teriminin ağırlıklandırma hesabı Eşitlik 4.7'deki gibidir.

$$W_{TF.IDF.ICF}(t_i) = TF(t_i, d_k) * \left(1 + \log(\frac{D}{d(t_i)}) \right) * \left(1 + \log(\frac{C}{c(t_i)}) \right) \quad \textbf{(4.7.)}$$

Eşitlikte yer alan $c(t_i)$, t_i teriminin geçtiği sınıf sayısını, C ifadesi ise koleksiyondaki toplam sınıf sayısını göstermektedir. Parantezler içerisindeki 1 değerleri ise, logaritma fonksiyonuna ait kısımlardan 0 gelmesi durumunda, ağırlıklandırmanın ham TF değerleri ile yapılabilmesi için mevcuttur. Ağırlıklandırma stratejisi gereği, TF-IDF-ICF, az sayıda sınıf veya dokümanda geçen nadir kelimelere yüksek değerler atamaktadır.

4.4.2. Ters sınıf uzay yoğunluk frekansına dayalı terim ağırlıklandırma (TF-IDF-ICF)

Bu ağırlıklandırma şemasında ise ağırlıklandırma sürecinde; koleksiyon frekans faktörü olarak, IDF haricinde, terimin, Ters Sınıf Uzay Yoğunluk Frekansı (Inverse Class

Space Density Frequency, ICSDF) bilgisi de kullanılır (Ren ve Sohrab, 2013). Sınıf bilgisi kullandığından ve her bir terim için tek skor ürettiğinden dolayı çoklu-sınıflandırmaya uygun gözetimli terim ağırlıklandırma şeması olarak ifade edilebilir. Eşitlik 4.8'de TF-IDF-ICSDF ile bir t_i teriminin ağırlık hesaplaması için kullanılacak formül yer almaktadır.

$$W_{TF.IDF.ICSDF}(t_i) = TF(t_i, d_k) * \left(1 + \log(\frac{D}{d(t_i)})\right) * \left(1 + \log(\frac{C}{\sum_{j=1}^{M} \frac{df_{t_{ij}}}{D_j}})\right) \quad \textbf{(4.8)}$$

Eşitlikte yer alan $df_{t_{ij}}$, ifadesi t_i teriminin j. sınıftaki doküman frekansını, D_j ise ilgili sınıfta yer alan toplam doküman sayısını ifade etmektedir. TF-IDF-ICSDF, TF-IDF-ICF terim ağırlıklandırma şemasının aksine; terimlerin her bir sınıf içerisindeki doküman dağılımı bilgilerini de ağırlıklandırma aşamasında kullandığından, dengesiz metin veri setlerinde TF-IDF-ICF şemasına nazaran genel olarak daha iyi ağırlıklandırma temsili sunmaktadır.

4.5. Ters yer çekimi momentine dayalı terim ağırlıklandırma (TF-IGM)

TF-IGM, Ters Yerçekimi Momenti (Inverse Gravity Moment, IGM) adlı istatistiksel bir modele bağlı olarak ağırlıklandırma yapan yakın zamanda önerilmiş gözetimli bir terim ağırlıklandırma şemasıdır (Chen vd., 2016). Ağırlıklandırma stratejisi, terimlerin üzerinde çalışılan veri setindeki tüm sınıfları yansıtmasını amaçladığından, tek skor üreten çoklu sınıflandırmaya uygun ağırlıklandırma yapan şemalar grubuna girmektedir. Koleksiyon frekansı olarak kullanılan IGM, her bir sınıf için terimlerin doküman frekanslarına odaklanarak, onların sınıflar-arası dağılımlarını keşfetmeye çalışır. Terimler için TF-IGM tabanlı ağırlık skoru hesabı Eşitlik 4.9'da verilmiştir.

$$W_{TF.\mathrm{IGM}}(t_i) = TF(t_i, d_k) * \left(1 + \lambda * \overbrace{\left[\frac{f_{i1}}{\sum_{r=1}^{M} f_{ir} * r}\right]}^{IGM(t_i)}\right) \quad \textbf{(4.9)}$$

Eşitlikteki f_{ir} ifadesi büyükten küçüğe sıralanmış bir biçimde t_i teriminin sınıf-bazlı doküman frekanslarını temsil etmektedir. Başka bir ifade kullanılırsa, r sırasıyla büyükten küçüğe dizilmiş olan, r'nci sınıfta t_i terimini içeren metin dokümanı sayısını göstermektedir. IGM ağırlıklandırma stratejisinin diğer ağırlıklandırma stratejilerinden en büyük farkı, terimler için hesaplanan sınıf-bazlı doküman frekanslarını sıralaması ve ağırlıklandırma sürecinde bu sıralanmış formu kullanmasıdır. Ayrıca, IGM ağırlıklandırma stratejisi, kullanılan veri setinin dengeli veya dengesiz olma karakteristiğine göre global ve lokal ağırlık faktörleri arasında kısmi olarak bir denge kurabilmek için, ayarlanabilir bir λ denge katsayısı da içermektedir. Referans alınan çalışmada, bu değer 5.0-9.0 aralığında tanımlanmış olup, varsayılan değer olarak ise 7.0 değeri ifade edilmiştir.

5. GÖZETİMLİ TERİM AĞIRLIKLANDIRMA ŞEMALARINDA TERİM FREKANS FAKTÖRÜ SEÇİMİNİN ETKİLERİ

Bu çalışmada, metin sınıflandırma için önerilmiş literatürde mevcut olan yedi farklı gözetimli terim ağırlıklandırma şemasının sınıflandırma performansları üzerinde terim frekans seçiminin etkileri analiz edilmiştir. Bu amaçla literatürdeki terim ağırlıklandırma çalışmalarında ayrı ayrı kullanılmış olan üç farklı terim frekans faktörü, her bir gözetimli şemanın ağırlıklandırma sürecine dahil edilip sonuçlar incelenmiştir. İki farklı metin veri setinde iki farklı sınıflandırıcı kullanılarak gerçekleştirilen deneyler, bahsi geçen analizlerin farklı öznitelik boyutlarındaki etkilerini de görebilmek amacıyla farklı sayıda öznitelik setleri ile gerçekleştirilmiştir.

5.1. Motivasyon

Daha önce 3.Bölüm'de de özetlenen terim ağırlıklandırma ile ilgili yapılan çalışmalardan da anlaşılacağı üzere, araştırmacılar yeni terim ağırlıklandırma şemaları önermek amacıyla çoğunlukla yeni koleksiyon frekansı faktörleri geliştirmeye yönelmektedir. Dolayısıyla da metin sınıflandırma için özellikle son yıllarda önerilen terim ağırlıklandırma şemalarının çoğunda, terim frekans faktörü olarak ham terim frekansı (TF) kullanılıp, yeni geliştirilmiş bir koleksiyon frekans faktörü ile kombine edilmektedir. Ancak, tek başına terim frekans faktörü de gözetimli bir terim ağırlıklandırma şeması için sınıflandırma performansını arttırabilme potansiyeli taşıma açısından hayati öneme sahiptir. Bu çalışmada, herhangi bir terim ağırlıklandırma şemasının performansının sadece efektif bir koleksiyon frekansına sahip olmasına değil, aynı zamanda uygun bir terim frekans faktörü seçimine (kullanımına) de bağlı olduğu gösterilmeye çalışılmıştır (Dogan ve Uysal, 2019b).

Terim ağırlıklandırma sürecinde ham TF değerlerini kullanmak, özellikle koleksiyondaki terimler yüksek TF değerlerine sahip olduğunda, metin dokümanlarının VSM'deki temsilini daha karmaşık bir hale getirebilmektedir. Bu problem şu şekilde ifade edilebilir: Koleksiyonda çok yüksek TF değerlerine sahip terimler olması, farklı sınıfa ait dokümanların birbirine daha yakın, aynı sınıfa ait dokümanların ise birbirinden daha uzak konumlanmasına sebep olabilir. Bu durum, mevcut terim ağırlıklandırma şemalarının sahip oldukları asıl sınıf ayırt edicilik potansiyellerini tam olarak yansıtamaması nedeniyle, dokümanların vektör uzayında daha zayıf bir biçimde temsil

edilmelerine sebep olmaktadır. Dolayısıyla da söz konusu terim ağırlıklandırma şemaları gerçekte sahip olduğu sınıflandırma performansından daha düşük bir performans sergilemektedir. Bu problemi çözmek ve yukarıda bahsedilen hipotezi savunmak için, yedi farklı gözetimli terim ağırlıklandırma şeması için ham TF frekansı ile birlikte iki modifiye edilmiş TF versiyonu (LOG_TF ve SQRT_TF) olmak üzere üç farklı terim frekans faktörü kullanılmış ve performansları kıyaslanmıştır.

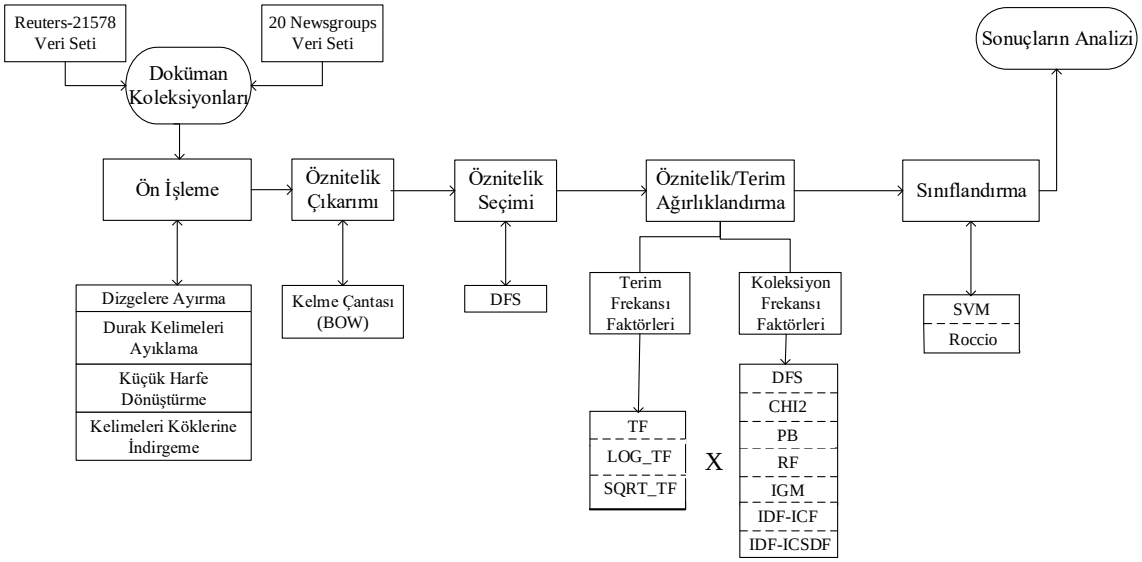
5.2. Terim Frekans Faktörleri ve Çalışmanın Genel Çerçevesi

Farklı terim frekans faktörleri kullanıldığında gözetimli terim ağırlıklandırma şemalarının sınıflandırma performanslarının nasıl değiştiğini analiz etmeyi amaçlayan bu çalışmada, deneysel bölümde yararlanılan üç farklı terim frekans faktörü Tablo 5.1'de gösterilmiştir.

Tablo 5.1. *Deneysel çalışmada kullanılan terim frekans faktörleri listesi*

TF Faktörü	Temsili	Açıklaması
tf	TF	Ham terim frekansı kullanımı (Bir terimin bir dokümanda geçme sayısı)
$\log_2(tf+1)$	LOG_TF	Yüksek terim frekansı etkisini düşürmek için terim frekansı değerlerinin logaritmasının kullanımı
sqrt(tf)	SQRT_TF	Yüksek terim frekansı etkisini düşürmek için terim frekansı değerlerinin karekökünün kullanımı

Yukarıda bahsedilen terim frekans faktörleri 7 gözetimli terim ağırlıklandırma şeması için iki farklı sınıflandırıcı ve iki farklı veri seti üzerinde kullanılarak gerçekleştirilen deneylerin genel çerçevesi ise Şekil 5.1'de verilmiştir.



Şekil 5.1. *Deneysel çalışmanın genel çerçevesi*

Deneysel çalışmayı gösteren büyük resim yukarıdaki gibi özetlenebilir. Kullanılan metotlar, veri setleri ve performans ölçütleri ise aşağıdaki bölümlerde detaylandırılmıştır.

5.3. Deneysel Çalışma Ortamı

Terim frekans faktörlerinin her bir şema için sergiledikleri performanslar, farklı karakteristiklere sahip Reuters-21578 ile 20-Newsgroups adlı iki metin koleksiyonu üzerinde gerçekleştirilen çeşitli deneylerle değerlendirilmiştir. Reuters-21578 veri seti literatürde metin sınıflandırma çalışmalarında yaygın olarak kullanılan ve Reuters-21578 ModApte (Asuncion ve Newman, 1994) olarak anılan bölümün ilk 10 sınıfını içermektedir. Her bir sınıfındaki doküman sayıları birbirinden oldukça farklı olduğundan dolayı Reuters-21578 metin veri seti dengesiz yapıya sahip (imbalanced) veri seti olarak değerlendirilir. Deneylerde kullanılan diğer veri seti ise biri hariç (997 adet) her sınıfında 1000'er dokümanı bulunan toplamda 20 sınıfa sahip olan 20-Newsgroups metin veri setinin ilk 10 sınıfını içermektedir. Sınıflarındaki doküman sayıları birbirine eşit olduğundan 20-Newsgroups veri seti dengeli (balanced) bir yapıya sahiptir.

Deneylerde, Reuters-21578 veri setinin kendine has önceden bölümlendirilmiş eğitim ve test dokümanları kullanılmış olup, 20-Newsgroups veri seti için ise eğitim ve test için kullanılmak üzere her sınıftan eşit sayıda doküman içeren (%50 ve %50) ayrı iki bölüm oluşturulmuştur. Kullanılan iki veri setine ait sınıf ve doküman sayısı bilgileri Tablo 5.2 ile 5.3'te sunulmuştur.

Tablo 5.2. *Reuters-21578 veri seti*

No	Sınıf Etiketi	Eğitim Dokümanı #	Test Dokümanı #
1	earn	2877	1087
2	acq	1650	719
3	money-fx	538	179
4	grain	433	149
5	crude	389	189
6	trade	369	117
7	interest	347	131
8	ship	197	89
9	wheat	212	71
10	corn	181	56

Tablo 5.3. *20-Newsgroups veri seti*

No	Sınıf Etiketi	Eğitim Dokümanı #	Test Dokümanı #
1	alt.atheism	500	500
2	comp.graphics	500	500
3	comp.os.ms-windows.misc	500	500
4	comp.sys.ibm.pc.hardware	500	500
5	comp.sys.mac.hardware	500	500
6	comp.windows.x	500	500
7	misc.forsale	500	500
8	rec.autos	500	500
9	rec.motorcycles	500	500
10	rec.sport.baseball	500	500

Ön işleme aşamasında, bahsi geçen koleksiyonlardan elde edilen doküman içeriklerine; sırasıyla, dizgelere ayırma, durak kelimeleri ayıklama, küçük harfe dönüştürme ve köklerine indirgeme gibi ön işlemler uygulanmıştır. Öznitelik seçim metodu olarak daha önceki bölümlerde kabaca anlatılan DFS adlı öznitelik seçim metodu kullanılmış olup, 300 ile 4000 arasında yedi farklı öznitelik boyutu için her bir şema ile bünyesinde kullanılan terim frekans faktörünün sınıflandırma performansları incelenmiştir.

Deneyler, TF-DFS, TF-CHI2, TF-PB, TF-RF, TF-IDF-ICF, TF-IDF-ICSDF ve TF-IGM terim ağırlıklandırma şemalarının ağırlıklandırma sürecinde, sırasıyla TF, LOG_TF ve SQRT_TF terim frekans faktörleri kullanılarak gerçekleştirilmiştir. TF-IGM için her iki veri seti üzerinde gerçekleştirilen deneylerde de, λ için 7.0 değeri atanmıştır. Sınıflandırma aşamasında, daha önceki bölümlerde çalışma stilleri kabaca anlatılan SVM ile Rocchio sınıflandırma algoritmaları kullanılmıştır. SVM sınıflandırıcı, gerçekleştirilen tüm deneylerde varsayılan parametrelerle çalıştırılmıştır.

5.4. Değerlendirme Ölçütleri

Metin sınıflandırma çalışmalarında, önerilen terim ağırlıklandırma metotlarının başarımları çoğunlukla Mikro-ortalama F ölçütü (Mikro-F1) ve Makro-ortalama F ölçütü (Makro-F1) değerlendirme ölçütleri ile değerlendirilmektedir.

Mikro-ortalamada, F ölçütü herhangi bir sınıf ayrımı olmadan genel olarak hesaplandığından, veri setinin tamamındaki veriler için tüm sınıflandırma kararları dikkate alınmaktadır. Dengesiz veri setleri ile çalışma yapılırken, yani sınıflardaki doküman sayılarında aşırı fark olan veri koleksiyonlarında, sınıflandırıcılar çok fazla dokümana sahip olan sınıflara atama yapmaya daha meyillidirler. Bu tip veri setlerinde büyük sınıflara ait Mikro-F1 başarımları küçük sınıflarınkini bastırabilmektedir. Mikro-F1 ölçütünün hesaplama formülü Eşitlik 5.1'deki gibidir.

$$Mikro - F1 = \frac{2 * \sum_{k=1}^{C} TP_{c_k}}{2 * \sum_{k=1}^{C} TP_{c_k} + \sum_{k=1}^{C} FP_{c_k} + \sum_{k=1}^{C} FN_{c_k}} \quad \textbf{(5.1)}$$

Makro-ortalamada ise, F ölçütü, veri seti içindeki her sınıf için hesaplanır ve ortalaması elde edilir. Bu ölçüt hesabında sınıf frekansları göz ardı edilerek her sınıfa eşit ağırlık atanmaktadır. Dengesiz veri koleksiyonları üzerinde terim ağırlıklandırma şemalarının performanslarını değerlendirirken Makro-F1 ölçütünü kullanmak, sınıflandırıcıların az sayıda dokümana sahip olan sınıfları ayırt edebilme yeteneklerini daha iyi gösterebilmeleri açısından daha adil bir seçim olabilir. Makro-F1 ölçütünün hesaplama formülü Eşitlik 5.2'deki gibidir.

$$Makro - F1 = \frac{1}{C} \sum_{k=1}^{C} F1_{c_k} \quad \textbf{(5.2)}$$

Yukarıdaki Eşitliklerde; *TP*, C_k sınıfına ait olan ve doğru olarak sınıflandırılan doküman sayısını; *FP*, C_k sınıfına ait olmadığı halde C_k sınıfına yanlış olarak sınıflandırılan doküman sayısını, *FN* ise aslında C_k sınıfına ait olduğu halde yanlış olarak sınıflandırılan doküman sayısını, *C* ise veri setindeki toplam sınıf sayısını göstermektedir.

Deneysel çalışma kısmında yukarıda bahsedilen üç farklı terim frekans faktörünün her bir terim ağırlıklandırma şemasının sınıflandırma performansına olan etkisi Mikro-F1 ölçütü kullanılarak ölçülmüş ve değerlendirilmiştir. Elde edilen sonuçlar ve değerlendirmeler ilerleyen alt bölümde sunulmuştur.

5.5. Sınıflandırma Sonuçları

Bu bölümde, çeşitli sayılarda öznitelikler kullanılarak Reuters-21578 ile 20-Newsgroups veri setleri üzerinde SVM ve Rocchio sınıflandırıcılarla elde edilen performans sonuçları sunulmuştur. Deneyler yukarıda bahsi geçen veri setleri ve sınıflandırıcılar ile yedi farklı gözetimli terim ağırlıklandırma şemasında üç farklı terim frekans faktörü kullanılarak gerçekleştirilmiştir. Deneysel sonuçlar içerisinde yer alan TF, LOG_TF ve SQRT_TF ifadeleri sırasıyla ham terim frekansını, logaritma fonksiyonu ile modifiye edilmiş terim frekansını ve kare-kök logaritma fonksiyonu ile modifiye edilmiş terim frekansını ifade etmektedir.

5.4-5.10 numaralı tablolar, Reuters-21578 veri seti üzerinde yedi farklı terim ağırlıklandırma şeması kullanılarak SVM ve Rocchio sınıflandırıcılar ile elde edilmiş Mikro-F1 skorlarını göstermektedir. Tablolarda aynı terim ağırlıklandırma şemasındaki her bir farklı terim frekans faktörü için maksimum değere sahip sonuçlar kalın punto ile gösterilmiş olup, sınıflandırıcı bazında ise her bir terim ağırlıklandırma şeması ile elde edilen en yüksek değerler de gri gölgeli bir biçimde gösterilmiştir.

Tablo 5.4. *Üç farklı terim frekans faktörü kullanılarak TF-DFS terim ağırlıklandırma yöntemi ile Reuters-21578 veri seti üzerinde elde edilen Mikro-F1 skorları (%)*

Öznitelik		SVM			Rocchio	
Sayısı	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	86.69	87.12	**87.44**	74.92	81.45	**81.74**
500	86.62	87.30	**87.33**	75.46	81.52	**81.92**
1000	86.33	87.51	**87.62**	75.67	81.63	**82.02**
1500	87.12	**87.80**	87.69	75.75	81.63	**82.17**
2000	86.98	87.77	**87.84**	75.85	81.81	**82.20**
3000	87.23	**87.77**	87.73	75.89	81.63	**82.13**
4000	87.30	**87.84**	87.69	75.92	81.67	**82.17**

Tablo 5.5. *Üç farklı terim frekans faktörü kullanılarak TF-CHI2 terim ağırlıklandırma yöntemi ile Reuters-21578 veri seti üzerinde elde edilen Mikro-F1 skorları (%)*

Öznitelik		SVM			Rocchio	
Sayısı	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	76.89	84.64	**84.93**	72.87	76.89	**77.93**
500	84.82	85.22	**85.33**	72.87	76.89	**77.97**
1000	85.47	86.55	**86.87**	72.87	76.89	**78.01**
1500	85.25	86.11	**86.19**	72.87	76.89	**78.01**
2000	85.22	86.19	**86.40**	72.87	76.89	**78.01**
3000	85.72	**86.47**	86.40	72.87	76.89	**78.01**
4000	85.58	86.47	**86.47**	72.87	76.89	**78.01**

Tablo 5.6. *Üç farklı terim frekans faktörü kullanılarak TF-PB terim ağırlıklandırma yöntemi ile Reuters-21578 veri seti üzerinde elde edilen Mikro-F1 skorları (%)*

Öznitelik		SVM			Rocchio	
Sayısı	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	85.94	**86.11**	85.79	70.97	74.63	**75.42**
500	86.04	**86.26**	86.01	70.97	74.67	**75.49**
1000	86.04	**86.33**	86.01	70.72	75.03	**75.49**
1500	86.08	**86.29**	85.97	70.72	75.03	**75.49**
2000	86.08	**86.29**	85.97	70.72	75.03	**75.49**
3000	86.08	**86.29**	85.97	70.72	75.03	**75.49**
4000	86.08	**86.29**	85.97	70.72	75.03	**75.49**

Tablo 5.7. *Üç farklı terim frekans faktörü kullanılarak TF-RF terim ağırlıklandırma yöntemi ile Reuters-21578 veri seti üzerinde elde edilen Mikro-F1 skorları (%)*

Öznitelik		SVM			Rocchio	
Sayısı	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	86.08	**86.87**	86.80	71.26	77.43	**78.08**
500	86.40	86.51	**86.62**	71.33	77.79	**78.22**
1000	86.44	86.76	**86.87**	71.44	77.83	**78.40**
1500	86.55	**87.23**	**87.23**	71.44	77.93	**78.58**
2000	86.44	86.98	**87.01**	71.44	78.08	**78.54**
3000	86.62	87.01	**87.08**	71.44	78.08	**78.47**
4000	86.87	**87.05**	86.98	71.48	78.08	**78.47**

Tablo 5.8. *Üç farklı terim frekans faktörü kullanılarak TF-IGM terim ağırlıklandırma yöntemi ile Reuters-21578 veri seti üzerinde elde edilen Mikro-F1 skorları (%)*

		SVM			Rocchio	
	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	85.54	**86.76**	86.58	75.64	79.98	**80.55**
500	85.76	**86.44**	86.37	76.39	80.91	**81.27**
1000	86.04	86.87	**87.05**	76.86	81.45	**81.77**
1500	86.01	87.12	**87.23**	77.40	81.63	**81.81**
2000	85.83	87.08	**87.30**	77.47	81.74	**81.84**
3000	85.94	87.08	**87.23**	77.47	81.77	**81.88**
4000	86.08	86.94	**87.23**	77.68	81.88	**82.13**

Tablo 5.9. *Üç farklı terim frekans faktörü kullanılarak TF-IDF-ICF terim ağırlıklandırma yöntemi ile Reuters-21578 veri seti üzerinde elde edilen Mikro-F1 skorları (%)*

Öznitelik		SVM			Rocchio	
Sayısı	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	85.76	86.62	**86.87**	78.90	81.59	**82.17**
500	85.72	86.22	**86.40**	79.37	81.38	**81.95**
1000	85.76	86.26	**86.58**	80.09	81.74	**82.17**
1500	85.94	86.80	**86.98**	80.66	81.92	**82.13**
2000	85.61	**86.69**	86.58	80.70	**81.92**	**81.92**
3000	85.61	86.44	**86.76**	80.88	**81.95**	81.77
4000	85.54	**86.69**	**86.69**	80.95	81.74	**81.74**

Tablo 5.10. *Üç farklı terim frekans faktörü kullanılarak TF-IDF-ICSDF terim ağırlıklandırma yöntemi ile Reuters-21578 veri seti üzerinde elde edilen Mikro-F1 skorları (%)*

Öznitelik	SVM			Rocchio		
Sayısı	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	85.25	86.40	**86.47**	80.19	82.78	**83.14**
500	84.97	85.15	**85.68**	80.34	82.24	**82.53**
1000	85.47	85.83	**86.01**	80.80	**81.63**	81.06
1500	85.33	85.94	**86.01**	81.02	**81.31**	81.09
2000	85.00	85.47	**85.65**	80.41	**80.91**	80.41
3000	85.25	85.68	**85.76**	79.55	**80.45**	80.30
4000	85.25	85.94	**85.97**	79.55	**80.52**	80.05

SVM sınıflandırıcı ile elde edilen sonuçlar değerlendirildiğinde, Mikro-F1 değerlerinin %85'ten yüksek olduğu ve en yüksek Mikro-F1 değerinin %87.84 ile LOG_TF-DFS ve SQRT_TF-DFS şemalarından elde edildiği görülmüştür. Benzer şekilde en düşük Mikro-F1 değeri, %76.9 ile TF-CHI2 şemasından 300 öznitelikle elde edilirken, SQRT_TF-CHI2 ve LOG_TF-CHI2 ile aynı boyutta elde edilen performans değerleri %84.6'dan yüksektir. Aradaki performans farkının (en az %7.7) bu kadar yüksek olmasının nedeni TF (ham terim frekansları) terim frekansı faktörü olarak söylenebilir. Çünkü muhtemelen yüksek değerler içeren bu faktörün diğer iki faktör ile indirgenmesiyle birlikte mevcut şemanın sınıflandırma performansı azımsanmayacak derece de (en az %7.7) artmıştır. Yani özetle, SVM sınıflandırıcı için TF terim frekans faktörünün modifikasyonu az sayıda öznitelikle yapılan deneylerde dahi sınıflandırma performansını pozitif bir şekilde etkilemiştir.

Rocchio sınıflandırıcı ile elde edilen sonuçlar değerlendirildiğinde ise, TF ile modifiye edilmiş versiyonları (LOG_TF ve SQRT_TF) arasında neredeyse %7'ye varan performans farklarının gözlendiği ifade edilebilir. Rocchio sınıflandırıcı ile elde edilen en düşük performans değeri TF-PB ile 1000 öznitelikle %70.72 olarak gerçekleşirken, en yüksek performans değeri ise %83.14 olarak SQRT_TF-ICSDF ile öznitelik boyutu 300'iken ölçülmüştür.

Reuters-21578 veri seti üzerinde elde edilen yukarıdaki sonuçları genel olarak değerlendirecek olursak; SQRT_TF ve LOG_TF terim frekans faktörleri ile tüm öznitelik boyutlarında, neredeyse tüm terim ağırlıklandırma şemaların performansları TF terim frekans faktörüne nazaran daha da artmıştır. LOG_TF ile SQRT_TF arasında bir kıyaslama yapılacak olunursa da; LOG_TF terim frekans faktörü terim ağırlıklandırma şemalarında daha çok tercih edilmesine rağmen, SQRT_TF terim frekans faktörünün

LOG_TF faktöründen genel olarak daha başarılı sınıflandırma sonuçları sağladığı görülmektedir.

5.11-5.18 numaralı tablolar, 20-Newsgroups veri seti üzerinde yedi farklı terim ağırlıklandırma şeması kullanılarak SVM ve Rocchio sınıflandırıcılar ile elde edilmiş Mikro-F1 skorlarını göstermektedir.

Tablo 5.11. *Üç farklı terim frekans faktörü kullanılarak TF-DFS terim ağırlıklandırma yöntemi ile 20-Newsgroups veri seti üzerinde elde edilen Mikro-F1 skorları (%)*

	SVM			Rocchio		
	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	97.36	98.30	**98.42**	88.94	97.88	**98.24**
500	97.34	98.04	**98.22**	88.70	97.90	**98.26**
1000	97.36	98.26	**98.54**	89.20	97.76	**98.18**
1500	97.26	98.32	**98.50**	89.20	97.78	**98.18**
2000	97.44	98.30	**98.44**	89.34	97.80	**98.16**
3000	97.48	98.30	**98.44**	89.46	97.90	**98.20**
4000	97.40	98.26	**98.38**	89.50	97.86	**98.18**

Tablo 5.12. *Üç farklı terim frekans faktörü kullanılarak TF-CHI2 terim ağırlıklandırma yöntemi ile 20-Newsgroups veri seti üzerinde elde edilen Mikro-F1 skorları (%)*

	SVM			Rocchio		
	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	96.34	97.74	**97.90**	84.00	96.50	**96.94**
500	96.82	**97.98**	97.94	84.00	96.52	**96.94**
1000	97.08	97.84	**98.10**	83.72	96.48	**96.96**
1500	97.30	97.88	**98.00**	83.72	96.48	**96.96**
2000	97.38	97.84	**98.04**	83.72	96.48	**96.96**
3000	97.32	97.82	**98.06**	83.72	96.48	**96.96**
4000	97.38	97.82	**98.06**	83.72	96.48	**96.96**

Tablo 5.13. *Üç farklı terim frekans faktörü kullanılarak TF-PB terim ağırlıklandırma yöntemi ile 20-Newsgroups veri seti üzerinde elde edilen Mikro-F1 skorları (%)*

	SVM			Rocchio		
	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	**77.34**	77.20	77.12	49.02	55.44	**58.54**
500	**77.94**	77.90	77.28	49.04	55.48	**58.56**
1000	**77.64**	76.84	76.16	41.00	50.78	**53.64**
1500	**77.76**	77.18	76.30	41.00	50.78	**53.70**
2000	**77.76**	77.32	76.32	41.00	50.78	**53.68**
3000	**77.68**	77.30	76.36	41.00	50.78	**53.70**
4000	**77.84**	77.32	76.38	41.00	50.78	**53.68**

Tablo 5.14. *Üç farklı terim frekans faktörü kullanılarak TF-RF terim ağırlıklandırma yöntemi ile 20-Newsgroups veri seti üzerinde elde edilen Mikro-F1 skorları (%)*

	SVM			Rocchio		
	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	96.82	98.02	**98.16**	61.82	69.34	**73.90**
500	96.58	97.76	**98.02**	62.10	69.84	**74.48**
1000	96.80	**98.18**	**98.18**	46.04	61.30	**66.70**
1500	97.04	98.32	**98.38**	46.38	61.60	**66.86**
2000	97.22	98.28	**98.40**	46.60	62.02	**67.08**
3000	97.12	**98.30**	**98.30**	46.84	62.24	**67.28**
4000	97.28	98.22	**98.30**	46.86	62.42	**67.42**

Tablo 5.15. *Üç farklı terim frekans faktörü kullanılarak TF-IGM terim ağırlıklandırma yöntemi ile 20-Newsgroups veri seti üzerinde elde edilen Mikro-F1 skorları (%)*

	SVM			Rocchio		
	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	96.68	97.74	**97.90**	87.94	96.78	**97.46**
500	97.14	**97.90**	**97.90**	88.10	96.82	**97.30**
1000	97.32	98.02	**98.18**	88.38	96.72	**97.24**
1500	97.24	97.88	**98.08**	88.02	96.70	**97.12**
2000	96.98	98.00	**98.08**	88.20	96.62	**97.02**
3000	97.10	98.00	**98.12**	88.22	96.54	**96.98**
4000	97.10	97.98	**98.14**	88.16	96.56	**96.94**

Tablo 5.16. *Üç farklı terim frekans faktörü kullanılarak TF-IDF-ICF terim ağırlıklandırma yöntemi ile 20-Newsgroups veri seti üzerinde elde edilen Mikro-F1 skorları (%)*

	SVM			Rocchio		
	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	96.78	97.86	**98.12**	81.92	92.66	**93.02**
500	96.46	97.82	**97.92**	80.34	92.36	**92.90**
1000	96.64	98.04	**98.16**	78.76	92.04	**92.52**
1500	96.30	98.02	**98.18**	78.50	91.78	**92.16**
2000	96.40	97.88	**98.02**	78.48	91.74	**91.90**
3000	96.10	**98.06**	**98.06**	78.44	91.40	**91.76**
4000	96.06	**97.98**	**97.98**	78.36	91.30	**91.64**

Tablo 5.17. *Üç farklı terim frekans faktörü kullanılarak TF-IDF-ICSDF terim ağırlıklandırma yöntemi ile 20-Newsgroups veri seti üzerinde elde edilen Mikro-F1 skorları (%)*

	SVM			Rocchio		
	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	96.04	97.50	**97.76**	78.02	89.62	**90.28**
500	95.94	97.40	**97.60**	76.14	89.66	**90.00**
1000	95.48	97.72	**97.92**	74.28	89.14	**89.50**
1500	95.12	97.66	**97.90**	74.02	**88.70**	**88.70**
2000	94.72	97.48	**97.74**	74.12	88.14	**88.50**
3000	94.00	96.94	**97.26**	74.30	87.46	**87.82**
4000	93.42	96.82	**97.12**	73.76	87.14	**87.38**

SVM sınıflandırıcı ile 20-Newsgroups veri setinden yedi farklı terim ağırlıklandırma şeması için elde edilen sonuçlar değerlendirildiğinde, TF-PB haricindeki tüm şemaların modifiye edilmiş terim frekans faktörleri (LOG_TF ve SQRT_TF) performanslarının TF terim frekans faktörü performanslarından daha yüksek olduğu görülmektedir. Terim frekans faktörleri kullanılarak terim ağırlıklandırma şemalarından elde edilen maksimum performans artışı %3.7 olarak TF-IDF-ICSDF üzerinde 4000 öznitelikle ölçülmüştür. SVM sınıflandırıcı için, SQRT_TF terim frekans faktörünün TF-PB haricindeki diğer altı şema üzerindeki başarım değerleri genel olarak LOG_TF ile TF terim frekans faktörlerininkilerden daha yüksek gözlenmiştir.

Rocchio sınıflandırıcı ile 20-Newsgroups veri setinde elde edilen sonuçlara göre ise, en yüksek performans farkı %20.67 ile SQRT_TF-RF (%66.7) ile TF-RF (%46.04) terim ağırlıklandırma şemaları arasında ölçülmüştür. En düşük Mikro-F1 değeri %41 olarak 1000 öznitelikle TF-PB şeması ile ölçülürken, en yüksek değer ise 98.26 ile SQRT_TF-DFS ile 500 öznitelik kullanılarak elde edilmiştir. Ayrıca tablolarda Rocchio sınıflandırıcı bölmesinde sunulan değerler incelenirse, öznitelik sayısı belirli bir eşiği aştıktan sonra bu sınıflandırıcı ile elde edilen performans sonuçlarının daha stabil bir hal aldığı değerlendirmesi yapılabilir.

20-Newsgroups veri seti üzerinde elde edilen yukarıdaki sonuçları genel olarak değerlendirecek olursak da; Reuters-21578 veri setinden elde edilen sonuçlara benzer bir değerlendirme yapabiliriz. Yani SQRT_TF ve LOG_TF terim frekans faktörleri ile genel olarak biri hariç (TF-PB) tüm terim ağırlıklandırma şemaların performansları TF terim frekans faktörüne nazaran daha yüksektir. 20-Newsgroups veri seti üzerinde, TF-PB terim ağırlıklandırma şamasının tüm terim frekans faktörleri (TF, LOG_TF ve SQRT_TF) ile performansı genel olarak diğer terim ağırlıklandırma şemalarının performanslarından daha düşük gözlenmiştir. TF-PB terim ağırlıklandırma şemasının sınıflandırma için dengesiz veri setlerinde dengeli veri setlerine nazaran daha uygun bir şema olduğunu göz önünde bulundurulursa, bu performans düşüklüğü 20-Newsgroups veri setinin dengeli bir yapıya sahip olmasıyla ilişkilendirilebilir. Nitekim TF-PB performansları veri seti düzeyinde kıyaslandığında da, Reuters-21578 veri seti üzerindeki Mikro-F1'değerlerinin dengeli bir yapıya sahip olan 20-Newsgroups üzerindekilerden daha yüksek olması da bu duruma ışık tutmaktadır. Ayrıca TF-IDF-ICF ve TF-IDF-ICSDF terim ağırlıklandırma şemalarının TF terim frekans faktörü ile performansları öznitelik boyutu arttıkça dramatik

bir biçimde azalırken, LOG_TF ve SQRT_TF ile ise bu değişim daha tutarlı bir biçimde gerçekleşmiştir. Bu durum, modifiye edilmiş terim frekans faktörlerinin (LOG_TF ve SQRT_TF) terim ağırlıklandırma şemalarını daha istikrarlı bir hale getirmiştir.

Her iki veri setinden de elde edilen sonuçlara göre, Reuters-21578 veri seti üzerindeki performans değişimlerinin aralığının, 20-Newsgroups veri seti üzerindekilerden genel olarak daha düşük olduğu söylenebilir. Bu durum, her bir veri setindeki dokümanların sınıf dağılımlarından kaynaklı olabilir. 20-Newsgroups veri seti dengeli veri seti yapısına sahip iken, Reuters-21578 ise aşırı derecede asimetrik dağılım yapısına sahiptir. Ayrıca SQRT_TF terim frekans faktörünü kullanan terim ağırlıklandırma şemaları, genel olarak, LOG_TF ile TF kullanan versiyonlarından daha üstün bir başarı grafiği çizmiştir. Deneysel sonuçlar, TF terim frekans faktörünün, deneyde kullanılan üç terim frekans faktörü arasında, terim ağırlıklandırma şemalarının potansiyelini yansıtma açısından en az etkin olanı olduğu görülmektedir.

Tablo 5.18 yedi farklı terim ağırlıklandırma üzerinde üç farklı terim frekans faktörü ile elde edilen maksimum sınıflandırma performanslarını göstermektedir. Mikro-F1 skorlarının altında yer alan parantez içinde belirtilmiş ifadeler, söz konusu terim ağırlıklandırma şemasında, ilgili maksimum Mikro-F1 değerinin hangi terim frekans faktörü ile elde edildiğini göstermektedir. Örneğin, Reuters-21578 veri seti üzerinde SVM sınıflandırıcı kullanılarak TF-DFS için elde edilen maksimum sınıflandırma performansı %87.84'tür ve SQRT_TF terim frekans faktörü ile elde edilmiştir.

Tablo 5.18. *Üç farklı terim frekans faktörü ile yedi terim ağırlıklandırma şemasından elde edilen maksimum Mikro-F1 skorları (%)*

Veri Seti	Sınıflandırıcı	TF-DFS	TF-CHI2	TF-PB	TF-RF	TF-IGM	TF-IDF-ICF	TF-IDF-ICSDF
Reuters-21578	SVM	**87.84** (SQRT_TF)	86.87 (SQRT_TF)	86.33 (LOG_TF)	87.23 (SQRT_TF)	87.30 (SQRT_TF)	86.98 (SQRT_TF)	86.47 (SQRT_TF)
	Rocchio	82.20 (SQRT_TF)	78.01 (SQRT_TF)	75.49 (SQRT_TF)	78.58 (SQRT_TF)	82.13 (SQRT_TF)	82.17 (SQRT_TF)	**83.14** (SQRT_TF)
20-Newsgroups	SVM	**98.54** (SQRT_TF)	98.10 (SQRT_TF)	77.94 (TF)	98.40 (SQRT_TF)	98.18 (SQRT_TF)	98.18 (SQRT_TF)	97.92 (SQRT_TF)
	Rocchio	**98.26** (SQRT_TF)	96.96 (SQRT_TF)	58.56 (SQRT_TF)	74.48 (SQRT_TF)	97.46 (SQRT_TF)	93.02 (SQRT_TF)	90.28 (SQRT_TF)

Ayrıca, tabloda her bir veri seti ve her bir sınıflandırıcı bazında hangi terim ağırlıklandırma şemasının en iyi performansı sağladığı ise kalın puntolu değerler ile ifade

edilmiştir. Sonuçlara bakıldığında ise, bir öznitelik seçim yönteminden ilk kez bu çalışma ile terim ağırlıklandırmaya uyarlanan TF-DFS terim ağırlıklandırma şemasının diğer altı şemaya nazaran genel anlamda daha üstün bir performans gösterdiği görülmüştür. Özellikle SQRT_TF-DFS terim ağırlıklandırma şemasının, metin sınıflandırma için son yıllarda önerilmiş olan TF-IGM, TF-IDF-ICF ile TF-IDF-ICSDF şemalarının performanslarını geride bırakmış olması, terim ağırlıklandırma alanında da, öznitelik seçim alanı kadar umut vadettiği düşüncesini doğurmuştur.

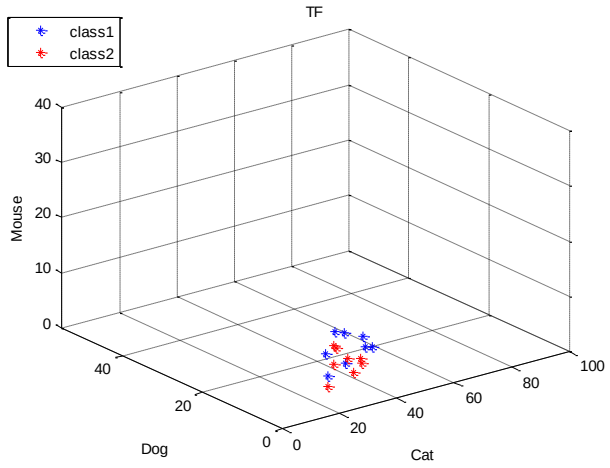
5.6. Tartışma

Bu bölümde, deneysel kısımda sunulan performans kazanımlarının nedenleri Tablo 5.19'da gösterilen küçük bir örnek veri kümesi ile tartışılmıştır. Metin sınıflandırmada metin dokümanlarının çok boyutlu öznitelik vektörleri ile temsil edildiği daha önceki bölümlerde belirtilmişti. Ancak, boyutu 3'ü aşan çok boyutlu öznitelik vektörlerini görselleştirmek oldukça zordur. Bu nedenle, Tablo 5.19'da yer alan örnek veri kümesi 3 öznitelik içerecek şekilde sunulmuştur. Örnek veri kümesinde 2 ayrı sınıfa ait toplam 16 dokümanın yer aldığını varsayalım. Bu dokümanlarda mevcut olduğunu düşündüğümüz 3 farklı özniteliğin terim frekansları da Tablo 5.19'daki gibi olsun.

Tablo 5.19. *Üç farklı öznitelik ve 16 dokümandan oluşan örnek veri kümesi*

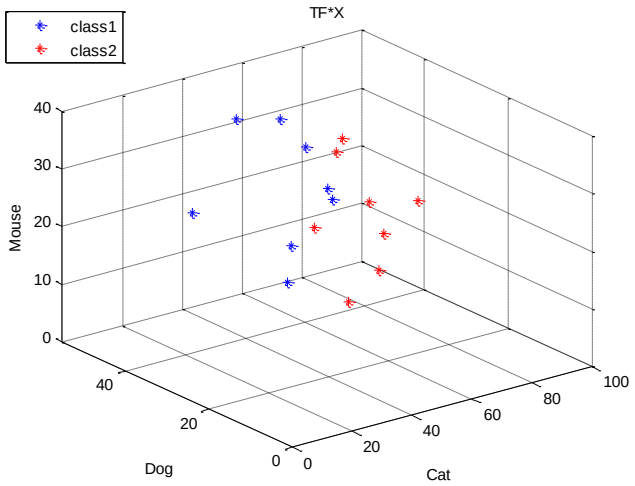
Sınıf-1	Cat	Dog	Mouse	**Sınıf-2**	Cat	Dog	Mouse
Doc1	43	20	1	Doc9	20	3	4
Doc2	30	10	2	Doc10	30	2	8
Doc3	51	14	3	Doc11	36	6	5
Doc4	49	15	5	Doc12	29	8	5
Doc5	47	13	4	Doc13	27	6	9
Doc6	40	13	2	Doc14	32	5	4
Doc7	41	14	7	Doc15	25	5	10
Doc8	42	17	6	Doc16	30	5	7

Üç boyutlu uzayda, tabloda belirtilen terim frekanslarına sahip bu üç özniteliği içeren doküman vektörleri Şekil 5.2'deki gibidir.



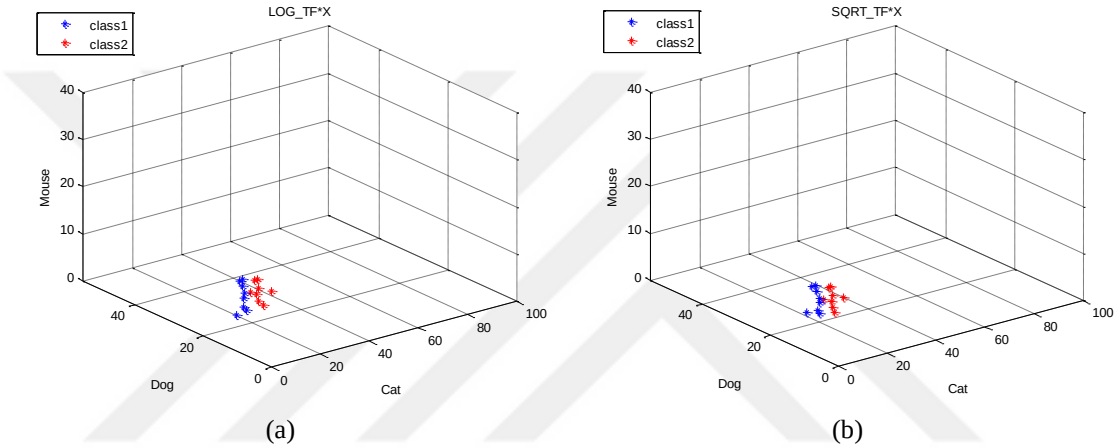
Şekil 5.2. *Üç boyutlu uzayda doküman vektörlerinin görünümü*

Şekil 5.2'deki doküman vektörleri incelendiğinde, iki farklı sınıfa ait doküman noktalarının iç içe girdiği ve mevcut haliyle sınıflandırılmasının hayli zor olduğu görülmektedir. İyi bir terim ağırlıklandırma şemasının, sınıflandırıcının performansını arttırmak için vektör uzayındaki metin dokümanlarının konumlarını uygun ağırlıklarla ayarlaması beklenir. Örneğin, X isimli bir terim ağırlıklandırma şeması yardımıyla 'Cat', 'Dog', ve 'Mouse' isimli özniteliklere sırasıyla 1.76, 3.91 ve 4.1 ağırlık değerlerini atadığını varsayalım. Söz konusu ağırlıklarla, bu üç özniteliğe ait doküman vektörleri üç boyutlu uzayda Şekil 5.3'teki gibi konumlanmıştır.



Şekil 5.3. *Üç boyutlu uzayda ağırlıklandırılmış doküman vektörleri*

Şekil 5.3'e bakıldığında, herhangi bir sınıflandırıcı için, X adlı terim ağırlıklandırma şeması ile ağırlıklandırıldıktan sonra iki sınıfa ait doküman vektörlerini ayırt etmenin nispeten daha kolay olduğunu söylemek mümkündür. Örnekte de görüldüğü üzere, ham terim frekansı değerlerini kullanmak, yüksek frekans değerlerinin etkisi nedeniyle çoğu durumda sınıflandırıcının performansını arttırmak için yeterli olmayabilir. Modifiye edilmiş terim frekansı değerlerinin etkisini göstermek için, LOG_TF-X ve SQRT_TF-X terim ağırlıklandırma şemaları ile ağırlıklandırılmış doküman vektörleri Şekil 5.4 ile üç boyutlu uzayda gösterilmiştir.



Şekil 5.4. *Üç boyutlu uzayda LOG_TF-X ve SQRT_TF-X terim ağırlıklandırma şemaları ile ağırlıklandırılmış doküman vektörlerinin gösterimi*

Şekil 5.4'te de görüldüğü üzere, çeşitli fonksiyonlarla ham TF değerlerinin modifiye edilerek yüksek terim frekansı değerlerinin etkisinin düşürülmesi, sınıflar arasında daha iyi bir ayrım yapılmasına yol açabilir. Örnekte de açıkça görüldüğü gibi, aynı sınıfa ait doküman vektörlerinin uygun bir biçimde ağırlıklandırılıp birbirlerine daha da yakınlaşması sağlanarak, sınıflandırma süreçlerine katkı sağlanabilir.

Özetle, bu bölümde, bir terim ağırlıklandırma şeması, terimlerin ağırlıklandırma süreçlerinde ham TF, LOG_TF ve SQRT_TF terim frekans faktörlerini kullandığında, doküman vektörlerinin vektör uzay modelinde nasıl konumlandığı gösterilmeye çalışılmıştır. Bu amaçla; ham terim frekanslarını, logaritmik fonksiyon ile indirgenmiş terim frekanslarını ve karekök fonksiyonu ile indirgenmiş terim frekanslarını kullanan üç farklı terim frekans faktörü vektör uzay modelinde örnek bir ver kümesi yardımıyla temsil edilmiştir. Şekil 5.3'te de görüldüğü gibi, terimlerin frekans değerleri yüksek olduğunda, terim ağırlıklandırma şemasının gerçekte sahip olduğu sınıflandırma potansiyelini tam

olarak yansıtamama ihtimali açısından, ham terim frekansı değerlerini kullanmak pek de efektif bir seçim olmayabilir. Nitekim Şekil 5.4, LOG_TF ve SQRT_TF terim frekans faktörleri kullanılarak ham TF değerlerinin indirgenmesinin, vektör uzay modelinde doküman vektörlerinin daha iyi temsil edilmesi bakımından daha etkili bir çözüm olduğunu göstermiştir. Doküman vektörlerinin ağırlık değerleri makul biçimde güncellenerek daha iyi temsil edilmesi, alt bölüm 5.5'te de rapor edildiği gibi terim ağırlıklandırma şemalarının daha başarılı sınıflandırma sonuçları elde etmesini sağlamaktadır.

5.7. Değerlendirmeler

Bu çalışmada, gözetimli terim ağırlıklandırma şemalarında terim frekans faktörü seçiminin etkileri geniş kapsamlı bir biçimde analiz edilmiştir. Bu analiz, üç farklı terim frekans faktörü, yedi farklı terim ağırlıklandırma şeması, iki sınıflandırıcı ile farklı karakteristiklere sahip iki metin veri seti üzerinde gerçekleştirilmiştir. Sınıflandırıcı olarak, yaygın olarak kullanılan iki vektör tabanlı sınıflandırma algoritması olan SVM ve Rocchio sınıflandırıcı kullanılmıştır. Gözetimli terim ağırlıklandırma şemalarında uygun terim frekans faktörü seçimi, terim ağırlıklandırma şemalarının ve sınıflandırıcıların sahip oldukları sınıflandırma performanslarını daha iyi yansıtmalarını sağlayabilir. Özellikle, yüksek terim frekansı değerleri söz konusu olduğunda, doküman vektörlerinin vektör uzay modelinde daha iyi temsil edilebilmeleri için, bu değerlerin tekrar düzenlenmesi gerekebilmektedir. Bu amaçla, gözetimli terim ağırlıklandırma şemalarında kullanılan yüksek terim frekansı değerleri uygun bir biçimde indirgenmelidir. Deneysel sonuçlar, gözetimli terim ağırlıklandırma şemalarında kullanılan terim frekans faktörünün modifikasyonunun, hem dengeli hem de dengesiz dağılıma sahip metin koleksiyonları üzerinde neredeyse tüm terim ağırlıklandırma şemalarının sınıflandırma performansını arttırdığını göstermiştir. Bu bağlamda, terim frekans faktörü uygun bir biçimde modifiye edilmesinin, belirli bir sınıfa ait olan doküman vektörlerinin konumlarının yeniden birbirlerine daha da yaklaşacak biçimde düzenlenmesine, dolayısıyla da terim ağırlıklandırma şemasının performansını pozitif olarak etkileyebileceği ifade edilmiştir. Bu durum sınıflandırma sürecini daha da kolaylaştırarak sınıflandırma performansını arttırabilmektedir. Deneysel bulgular göz önünde bulundurulduğunda, LOG_TF terim frekans faktörü SQRT_TF terim frekans faktöründen daha popüler olmasına rağmen,

SQRT_TF terim frekans faktörünün sınıflandırma performansı genel olarak LOG_TF terim frekans faktörünün performansından daha üstün gözlenmiştir. Özetlersek, bu çalışmada var olan ağırlıklandırma şemasının genel sınıflandırma başarımının yalnızca yeni geliştirilmiş bir koleksiyon frekansı kullanmaya değil, aynı zamanda da, üzerinde çalışılan metin veri setine göre uygun bir terim frekans faktörünün seçimine bağlı olduğu gösterilmeye çalışılmıştır. Nitekim deneysel sonuçlar da, terim ağırlıklandırma şemalarında terim frekansı değerlerinin logaritmik ve karekök fonksiyonları ile dönüştürülmesinin ham terim frekanslarının kullanıldığı duruma göre daha iyi sınıflandırma sonuçları sağlayabileceği fikrini desteklemiştir.

Deneysel kısımda kullanılan TF-DFS terim ağırlıklandırma şeması, literatürde mevcut olan ayırt edici öznitelik seçici isimli öznitelik seçim metodundan ilk kez terim ağırlıklandırma için uyarlanmıştır. Metin sınıflandırma için efektif bir öznitelik seçim metodu olan bu metot, literatürde terim ağırlıklandırma için daha önce kullanılmamıştır. Dolayısıyla bu çalışma bu açıdan da bir ilk olma özelliği taşımaktadır. Deneysel sonuçlar, SQRT_TF-DFS şemasının, TF, LOG_TF ve SQRT_TF terim frekans faktörlerini kullanan diğer altı terim ağırlıklandırma şemasından çoğunlukla daha üstün performans sergilediğini göstermiştir.

6. TERS YER ÇEKİMİ MOMENTİNE DAYALI TERİM AĞIRLIKLANDIRMA ŞEMASININ GELİŞMİŞ BİR İMPLEMENTASYONU

Bu çalışmada, metin sınıflandırma için daha önce önerilmiş olan ters yer çekimi momentine (IGM) dayalı terim ağırlıklandırma stratejisinin (Chen vd., 2016) bazı ekstrem durumlarda yetersiz kaldığı ağırlıklandırma süreçleri analiz edilmiştir. Söz konusu IGM ağırlıklandırma stratejisi çeşitli modifikasyonlarla güncellenerek bahsi geçen ekstrem durumlar için daha makul ağırlıklar üretebilmeyi hedefleyen, gelişmiş ters yer çekimi momenti (IGM_{imp}) adıyla yeni bir ağırlıklandırma stratejisi geliştirilmiştir (Dogan ve Uysal, 2019a). Geliştirilen IGM_{imp} ağırlıklandırma stratejisine bağlı olarak $SQRT\_TF\text{-}IGM_{imp}$ ve $TF\text{-}IGM_{imp}$ adında iki yeni terim ağırlıklandırma şeması önerilmiştir. Önerilen $SQRT\_TF\text{-}IGM_{imp}$ ve $TF\text{-}IGM_{imp}$ terim ağırlıklandırma şemalarının performansları, ikisi standart IGM tabanlı terim ağırlıklandırma şemaları olmak üzere literatürden toplamda yedi farklı terim ağırlıklandırma şemasının performanslarıyla kıyaslanmıştır. Bu kıyaslamalar, SVM, KNN ve NN sınıflandırıcılar kullanılarak, ikisi dengeli doküman dağılıma sahip metin veri seti (20 Mini Newsgroups ve 20-Newsgroups) ve biri dengesiz doküman dağılımına sahip metin veri seti (Reuters-21578) olmak üzere toplamda üç veri seti üzerinde gerçekleştirilmiştir. Ayrıca değerlendirme ölçütü olarak daha önceki bölümde ifade edilen hem Mikro-F1 hem de Makro-F1 metrikleri beraber tercih edilmiştir.

6.1. Motivasyon

Metin sınıflandırmada bir terim ağırlıklandırma şemasının başarısı, belirli bir terimin geçtiği sınıf hakkında mümkün olduğunca fazla bilgi içermesine bağlıdır. Eğer literatürdeki mevcut şemalardan daha başarılı yeni bir terim ağırlıklandırma şeması geliştirilecekse, mevcut olanların avantajlı ve dezavantajlı olduğu yanları, ağırlıklandırmada yetersiz kaldığı problematik durumları iyi analiz etmek gerekir. Bu yüzden araştırmacılar, yeni koleksiyon frekans faktörleri geliştirmeye çalışırken, genellikle mevcut terim ağırlıklandırma şemalarının ağırlıklandırma davranışlarını araştırmaya meyillidir.

Bu çalışmada, TF-IGM adıyla terim ağırlıklandırma için yakın zamanda önerilmiş olan ters yerçekimi momentine dayalı terim ağırlıklandırma stratejisinin özellikle bazı ekstrem senaryolara sahip terimler için uyguladığı ağırlıklandırma süreci detaylı olarak

incelenmiştir. İnceleme sonunda, IGM ağırlıklandırma formülünün, söz konusu senaryolardaki terimlere atanan ağırlıkları daha fazla bilgi içerebilecek biçimde düzenlenebileceği keşfedilmiştir. IGM_{imp} adıyla yeniden düzenlenen ağırlıklandırma formülü kullanılarak metin sınıflandırma için SQRT_TF-IGM_{imp} ve TF-IGM_{imp} adında iki yeni terim ağırlıklandırma şeması önerilmiştir. Çalışmada özellikle ağırlıklandırma süreci geliştirilmeye çalışılan üç senaryo mevcuttur. Birinci senaryoda, tek bir sınıfa ait bir ya da daha fazla dokümanda geçen terimler, ikinci senaryoda iki ya da daha fazla sınıfa ait eşit sayıda dokümanda geçen terimler, üçüncü senaryoda ise iki ya da daha fazla sınıfa ait farklı sayılarda dokümanlarda geçen terimler incelenmiştir. Bu üç senaryonun da ortak özelliği standart IGM koleksiyon frekansının bu üç senaryodaki terimlere doküman frekansları farklı dahi olsa aynı ağırlık değerlerini atamasıdır. Başka bir deyişle, bu üç senaryoya sahip terimlere, ayırt edicilikleri farklı olmasına rağmen standart IGM formülü ile eşit ağırlık değerleri hesaplanmaktadır. Ayrıca, her ne kadar bu çalışmada üç senaryo üzerinde durulduysa da, daha da derinlemesine araştırıldığında bu tip senaryoların sayısını arttırmak da mümkündür. Standart TF-IGM şemasının bu tip senaryolar için yetersiz kaldığı ağırlıklandırma davranışı, mevcut IGM koleksiyon frekansı formülünün paydasına eklenen yeni bir oran ile çözülmeye çalışılmıştır. IGM formülünün yeniden düzenlenmesi sadece yukarıda belirtilen problematik senaryolar için etkin bir çözüm oluşturmamış, aynı zamanda problematik olmayan, daha makul ağırlıklandırma gerçekleştirdiği diğer durumları da göz ardı edilebilir ölçüde etkilemiştir. Belirtilen senaryoların ve yeniden düzenlenen IGM formülünün detayları ilerleyen alt bölümlerde daha detaylı bir biçimde ifade edilmiştir.

6.2. TF-IGM ile Terim Ağırlıklandırmaya Genel Bakış

Bu alt bölümde, standart TF-IGM ile terim ağırlıklandırma hesabı, referans alınan çalışmadaki (Chen vd., 2016) iki örnek ile anlatılmıştır. İlk örnekte, 5 farklı sınıfa ait her sınıfında 10'ar dokümanı olan bir metin koleksiyonu için t_1 ve t_2 adında, doküman frekansları, sırasıyla, {4, 2, 2, 2, 2} ve {4, 8, 0, 0, 0} olan iki farklı terim olduğunu varsayalım. Eşitlik 4.9'da belirtilen standart IGM(t_i) formülüne göre, t_1 ve t_2 terimlerinin hesaplanan lokal IGM değerleri 0.125 ve 0.5 olacaktır. Yine aynı formülde yer alan λ katsayısı 7 olarak set edilirse, t_1 ve t_2 terimleri için hesaplanan global IGM ağırlıkları ise sırasıyla 1.875 ve 4.5 olacaktır. Hesaplanan ağırlık değerleri göz önüne alındığında, eşit

terim frekansı değerlerine sahip olmaları durumunda t_2'nin ağırlık değerinin t_1'den daha yüksek olduğunu, yani w (t_2) > w (t_1) olduğunu söylemek mümkündür.

İkinci örnek için ise, 6 farklı sınıfa ait her sınıfında 10'ar dokümanı olan bir metin koleksiyonu olduğunu varsayalım. Bu koleksiyonda geçen t_1 ve t_2 terimlerinin doküman frekansları ise, sırasıyla, {8, 7, 6, 6, 0, 0} ve {9, 2, 2, 2, 0, 0} olsun. Bu durumda, t_1 ve t_2 terimlerinin hesaplanan lokal IGM değerleri, sırasıyla, 0.125 ve 0.333, λ=7 alındığında da global IGM tabanlı ağırlık değerleri de, sırasıyla, 1.875 ile 3.333 olarak hesaplanır. Dolayısıyla bu doküman dağılımına sahip t_1 ve t_2 terimlerine atanan ağırlık değerleri için, ağırlık sıralaması da w (t_2) > w (t_1) şeklindedir.

6.3. Bazı Ekstrem Senaryolar için Standart IGM Faktörünün Ağırlıklandırma Davranışları

Bu bölümde IGM faktörünün üç farklı senaryoya sahip 13 farklı terim için ağırlıklandırma davranışı tasvir edilmiştir. Her bir senaryo ve içerdiği terimler aşağıdaki gibidir:

Senaryo 1: 5 farklı sınıfa ait dokümanlar içeren, her sınıfında 10'ar dokümanı olan bir metin koleksiyonu içinde; doküman frekansları sırasıyla {10, 0, 0, 0, 0}, {8, 0, 0, 0, 0}, {5, 0, 0, 0, 0}, {3, 0, 0, 0, 0} ve {1, 0, 0, 0, 0} olan beş farklı terim (t_1, t_2, t_3, t_4 ve t_5), yer aldığını varsayalım. Söz konusu terimlerin (sezgisel olarak) sınıf ayırt edicilik düzeyleri $t_1 > t_2 > t_3 > t_4 > t_5$ şeklinde bir sıralamaya sahipken, standart IGM formülüne (Eşitlik 4.9) göre tamamının ağırlık değerleri 1'e eşittir.

Senaryo 2: Bu senaryoda, 6 farklı sınıfı olan ve her sınıfında 100'er dokümanı olan bir metin koleksiyonumuz olduğunu düşünelim. Bu koleksiyondaki beş farklı terim ise (t_6, t_7, t_8, t_9 ve t_{10}) sırasıyla {100, 100, 0, 0, 0, 0}, {40, 40, 0, 0, 0, 0}, {23, 23, 0, 0, 0, 0}, {11, 11, 0, 0, 0, 0} ve {2, 2, 0, 0, 0, 0} biçiminde doküman frekanslarına sahip olsun. Böyle bir ortamda, terimlerin (sezgisel olarak) sınıf ayırt edicilik düzeyleri $t_6 > t_7, > t_8, > t_9 > t_{10}$ şeklinde sıralanmalıdır ancak standart IGM formülü, (Eşitlik 4.9) tamamına aynı ağırlık skorunu (0.333) atamaktadır.

Senaryo 3: Bu kez, 3 farklı sınıfımız ve her birinde 1000'er dokümanınız olsun. Bu koleksiyon içinde t_{11}, t_{12} ve t_{13} ile gösterilen üç farklı terim yer alsın. Söz konusu terimlerin doküman frekansları ise sırasıyla {924, 476, 112, 0, 0, 0, 0}, {231, 119, 28, 0,

0, 0, 0}, {33, 17, 4, 0, 0, 0, 0} biçiminde olsun. Doküman frekanslarına bakıldığında; bir sınıfa ait 1000 adet dokümanın 924'ünde geçtiği için, bu üç terim içerisinde en ayırt edici olan terimin t_{11} olduğu görülmektedir. Benzer biçimde t_{12} terimi de t_{13}'ten daha ayırt edicidir. Ancak Eşitlik 4.9'daki standart IGM formülü bu üç ayırt edicilik gücü farklı terime eşit ağırlık değeri (0.418) hesaplamaktadır.

Özetleyecek olursak, yukarıdaki üç senaryoda bahsedilen terimler (t_1-t_{13}), standart IGM ile ağırlıklandırıldığında; söz konusu terimlerin sahip oldukları gerçek ayırt edicilik kabiliyeti vektör uzay modeline tam olarak yansıtılamamaktadır. Senaryolar içerisindeki terimlerin her biri, farklı ayırt etme yeteneklerine sahip olmalarına rağmen; standart IGM hepsine eşit ağırlık skoru atamaktadır. Daha da önemlisi, bu tip senaryoları çoğaltmak da mümkündür. Bu problemli ağırlıklandırma davranışı yeni bir oran kullanılarak, standart IGM ağırlıklandırma formülünün yeniden organize edilmesiyle çözülmeye çalışılmıştır. Yeni ağırlıklandırma formülünün detayları bir sonraki alt bölümde mevcuttur.

6.4. Önerilen Koleksiyon Frekansı Faktörü: Geliştirilmiş Ters Yer Çekimi Momenti (IGM_{imp})

Daha önceki alt bölümde de belirtildiği gibi, problematik ağırlıklandırmaya sahip yukarıdaki senaryoları çoğaltmak mümkündür. Ve bu yüzden bu tip senaryolardaki ağırlıklandırma problemlerinin çözülmesi, terimlerin daha verimli ağırlıklandırılması açısından oldukça önemlidir. Önerilen çözüm yaklaşımı, Eşitlik 4.9'daki standart IGM ağırlıklandırma formülünün paydasına $\text{Log}_{10}[\, D_{total}(t_{i\_\max}) / D_{t_{i\_\max}}\,]$ oranının Eşitlik 6.1'deki gibi eklenerek, söz konusu senaryolar için daha makul ağırlık değerleri atanmasını amaçlamaktadır.

$$IGM_{imp}(t_i) = \frac{f_{i1}}{\sum_{r=1}^{M} f_{ir} * r + Log_{10}\left[\frac{D_{total}(t_{i\_\max})}{D_{t_{i\_\max}}}\right]} \qquad \textbf{(6.1)}$$

Eşitlikte yer alan $D_{t_{i\_\max}}$ ifadesi t_i teriminin en çok geçtiği sınıftaki, geçtiği doküman sayısını, $D_{total}(t_{i\_\max})$ ise, t_i teriminin en çok geçtiği sınıftaki toplam doküman sayısını göstermektedir.

Ters Doküman Balans Frekansı (Inverse Document Balance Frequency, IDBF) olarak adlandırılan bu oran, yukarıdaki senaryolarda bahsedilenlere benzer ekstrem doküman frekanslarına sahip terimler için daha makul ağırlıklar atamaktadır. Söz konusu formül, standart TF-IGM formülündeki ifadelerle kombine edildiğinde, IGM_{imp} koleksiyon frekansı faktörünün son hali Eşitlik 6.2'deki gibidir.

$$IGM_{imp}(t_i) = \frac{f_{i1}}{\sum_{r=1}^{M} f_{ir} * r + Log_{10}\left[\frac{D_{total}(t_{i\_\max})}{f_{i1}}\right]} \qquad \textbf{(6.2)}$$

$D_{t_{i\_\max}}$ ifadesi, yani t_i teriminin en çok geçtiği sınıftaki, geçtiği doküman sayısı, standart IGM formülünde f_{i1} ifadesine eşit olduğundan, önerilen IGM_{imp} formülü de bu yönde güncellenerek son halini almıştır.

Tablo 6.1'de, yukarıda bahsedilen üç senaryoya sahip 13 terime, standart IGM ve önerilen IGM_{imp} metotları ile verilen ağırlık değerleri gösterilmiştir.

Tablo 6.1. *Bazı ekstrem senaryolar için standart IGM ve önerilen IGM_{imp} metotları ile terim ağırlıklandırma*

	Terimler	**Sınıf ve Doküman Sayıları**	**Doküman Frekansları**	**Ayırt Etme Gücü Sıralamaları (Sezgisel Olarak)**	**IGM Değerleri**	**IGM_{imp} Değerleri**
Senaryo 1	t_1	5 / 10	{10, 0, 0, 0, 0}	$t_1 > t_2 > t_3 > t_4 > t_5$	1	**1**
	t_2		{8, 0, 0, 0, 0}		1	**0.988**
	t_3		{5, 0, 0, 0, 0}		1	**0.943**
	t_4		{3, 0, 0, 0, 0}		1	**0.852**
	t_5		{1, 0, 0, 0, 0}		1	**0.5**
Senaryo 2	t_6	6 / 100	{100, 100, 0, 0, 0, 0}	$t_6 > t_7 > t_8 > t_9 > t_{10}$	0.333	**0.333**
	t_7		{40, 40, 0, 0, 0, 0}		0.333	**0.332**
	t_8		{23, 23, 0, 0, 0, 0}		0.333	**0.330**
	t_9		{11, 11, 0, 0, 0, 0}		0.333	**0.323**
	t_{10}		{2, 2, 0, 0, 0, 0}		0.333	**0.259**
Senaryo 3	t_{11}	7 / 1000	{924, 476, 112, 0, 0, 0, 0}	$t_{11} > t_{12} > t_{13}$	0.418	**0.418**
	t_{12}		{231, 119, 28, 0, 0, 0, 0}		0.418	**0.417**
	t_{13}		{33, 17, 4, 0, 0, 0, 0}		0.418	**0.410**

Tablo 6.1'den de açıkça görüldüğü gibi, problemli senaryolara sahip terimler için, IGM_{imp} sahip oldukları ayırt edicilik güçlerini daha tutarlı bir biçimde yansıtan ağırlıklar üretmektedir. Dolayısıyla da IGM_{imp} ile söz konusu senaryolara sahip terimler için standart IGM'den daha makul ağırlıklandırma gerçekleştirilmektedir.

Her ne kadar, önerilen IGM_{imp} koleksiyon frekansı ile problematik durumlardaki terimlere yukarıdaki gibi daha verimli ağırlıklandırma çözümleri elde edilmiş olsa da, problematik olmayan durumlarda standart IGM formülünün zaten makul olan genel ağırlıklandırma davranışını yeni önerilen IGM_{imp} ile mümkün olduğu kadar değiştirmemek gereklidir. Dolayısıyla, önerilen IGM_{imp} formülü ile problematik olmayan durumlardaki ağırlıklandırma davranışının değişip değişmediğini; daha önce standart IGM formülünün genel ağırlıklandırma sürecinin anlatıldığı bölüm 5.2'deki ilk örnek ile test edelim. Hatırlarsak, bahsi geçen örnekte, 5 farklı sınıfa ait her sınıfında 10'ar dokümanı olan bir metin koleksiyonu için t_1 ve t_2 adında, doküman frekansları, sırasıyla, {4, 2, 2, 2, 2} ve {4, 8, 0, 0, 0} olan iki farklı terim olduğunu varsaymıştık. Standart IGM formülüne göre, t_1 ve t_2 terimlerinin hesaplanan lokal IGM değerleri sırasıyla 0.125 ve 0.5 olarak hesaplanmıştı. Önerilen IGM_{imp} formülüne göre ise bu değerler 0.12346 ve 0.49698 olarak hesaplanır. Standart IGM ve önerilen IGM_{imp} ağırlıklandırma formülü ile karşılaştırma yapılırsa, $0.12346 \cong 0.125$ ve $0.49698 \cong 0.5$ olduğundan; önerilen formülün genel ağırlıklandırma davranışını pek etkilemediğini söylenebilir. Yani zaten makul olan genel ağırlıklandırma davranışının önerilen IGM_{imp} ile de sürdürüldüğü ifade edilebilir. Özetle, önerilen IGM_{imp} terim frekans faktörü, problemli senaryodakilere benzer bir dağılıma sahip terimler için sadece daha verimli bir biçimde ağırlıklandırma yapmakla kalmayıp, aynı zamanda da standart IGM terim frekans faktörünün zaten iyi olan genel ağırlıklandırma davranışını sergilemeye devam etmektedir.

Geliştirilen IGM_{imp} faktörüne bağlı olarak, önerilen TF-IGM_{imp} ve SQRT_TF-IGM_{imp} terim ağırlıklandırma şemaları sırasıyla Eşitlik 6.3 ile 6.4'te belirtilmiştir.

$$W_{TF.IGM_{imp}}(t_i) = TF(t_i, d_k) * \left(1 + \lambda * IGM_{imp}(t_i)\right) \quad \textbf{(6.3)}$$

$$W_{SQRT\_TF.IGM_{imp}}(t_i) = SQRT\_TF(t_i, d_k) * \left(1 + \lambda * IGM_{imp}(t_i)\right) \quad \textbf{(6.4)}$$

Deneysel kısımda önerilen bu iki terim ağırlıklandırma şeması, standart versiyonları ile ve 5 farklı terim ağırlıklandırma şeması ile birlikte kullanılmış olup, farklı

veri setleri ve sınıflandırıcılar ile performansları kıyaslanmıştır. Deneysel çalışma ortamına ait bütün parametreler, deneysel sonuçlar ve genel değerlendirmeler ilerleyen alt bölümlerde anlatılmıştır.

6.5. Deneysel Çalışma Ortamı

Önerilen TF-IGM$_{imp}$ ve SQRT_TF-IGM$_{imp}$ terim ağırlıklandırma şemalarının sınıflandırma başarımları, Reuters-21578, 20 Mini Newsgroups ve 20-Newsgroups metin veri setleri üzerinde SVM, KNN ve NN sınıflandırıcılar kullanılarak test edilmiştir. 3 farklı veri setinin tercih edilmiş olmasının sebebi, hem dengeli hem de dengesiz bir dağılıma sahip literatürde çok sık başvurulan veri koleksiyonları üzerinde önerilen şemaların performanslarını analiz etmektir.

Reuters-21578 veri setine ait dokümanlar işlenirken, ilk 10 sınıfa ait ModApte (Asuncion ve Newman, 1994) kümesi içinde yer alan iki ya da daha fazla sınıf etiketine sahip dokümanlar ayıklanmıştır. Ayıklama sonucunda "wheat" ve "corn" adlı sınıflarda doküman kalmadığı için, bu sınıflar silinmiş ve kalan 8 sınıf üzerinde deneyler gerçekleştirilmiştir. Deneysel kısımda kullanılan Reuters-21678 veri setinin sınıf sayısı ve doküman dağılımı bilgileri Tablo 6.2'deki gibidir.

Tablo 6.2. *Reuters-21578 veri seti*

No	Sınıf Etiketi	Eğitim Dokümanı #	Test Dokümanı #
1	earn	2840	1083
2	acq	1596	696
3	money-fx	206	87
4	grain	41	10
5	crude	253	121
6	trade	251	117
7	interest	190	75
8	ship	108	36

20 Mini Newsgroups veri seti, metin sınıflandırma çalışmalarında sıklıkla kullanılan popüler 20 Newsgroups (Asuncion ve Newman, 1994) metin veri setinin mini bir alt kümesidir. Toplamda 20 sınıfa sahip olan ve her sınıfında 100'er doküman bulunan bu veri seti toplamda 2000 doküman yer almaktadır. Dengeli bir dağılım yapısına sahip bu veri setinde yer alan 20 sınıfa ait dokümanların yüzde 70'i eğitim, yüzde 30'u da test için kullanılmıştır. Deneysel kısımda kullanılan 20 Mini Newsgroups veri seti ile ilgili bilgiler Tablo 6.3'te sunulmuştur.

Tablo 6.3. *20 Mini Newsgroups veri seti*

No	Sınıf Etiketi	Eğitim Dokümanı #	Test Dokümanı #
1	alt.atheism	70	30
2	comp.graphics	70	30
3	comp.os.ms-windows.misc	70	30
4	comp.sys.ibm.pc.hardware	70	30
5	comp.sys.mac.hardware	70	30
6	comp.windows.x	70	30
7	misc.forsale	70	30
8	rec.autos	70	30
9	rec.motorcycles	70	30
10	rec.sport.baseball	70	30
11	rec.sport.hockey	70	30
12	sci.crypt	70	30
13	sci.electronics	70	30
14	sci.med	70	30
15	sci.space	70	30
16	soc.religion.christian	70	30
17	talk.politics.guns	70	30
18	talk.politics.mideast	70	30
19	talk.politics.misc	70	30
20	talk.religion.misc	70	30

20-Newsgroups veri seti 20 sınıftan oluşmaktadır. Biri haricinde geriye kalan tüm sınıflarında 1000'er doküman bulunan bu veri setinde, toplamda 19997 doküman yer almaktadır (Asuncion ve Newman, 1994). 20-Newsgroups veri seti ile yapılan deneylerde, veri setinin sahip olduğu 20 sınıf ve bu sınıflara ait tüm dokümanlar eğitim ve test için yarı yarıya bölümlendirilerek kullanılmıştır. Sınıflara ve doküman sayılarına ait bilgiler daha Tablo 6.4'te sunulmuştur.

Tablo 6.4. *20-Newsgroups veri seti*

No	Sınıf Etiketi	Eğitim Dokümanı #	Test Dokümanı #
1	alt.atheism	500	500
2	comp.graphics	500	500
3	comp.os.ms-windows.misc	500	500
4	comp.sys.ibm.pc.hardware	500	500
5	comp.sys.mac.hardware	500	500
6	comp.windows.x	500	500
7	misc.forsale	500	500
8	rec.autos	500	500
9	rec.motorcycles	500	500
10	rec.sport.baseball	500	500
11	rec.sport.hockey	500	500
12	sci.crypt	500	500
13	sci.electronics	500	500
14	sci.med	500	500
15	sci.space	500	500
16	soc.religion.christian	500	497
17	talk.politics.guns	500	500
18	talk.politics.mideast	500	500
19	talk.politics.misc	500	500
20	talk.religion.misc	500	500

Ön işleme aşamasında, bahsi geçen üç metin koleksiyonundan elde edilen doküman içeriklerine, sırasıyla, dizgelere ayırma, durak kelimeleri ayıklama, küçük harfe dönüştürme ve köklerine indirgeme gibi ön işlemler bu çalışmada da uygulanmıştır. Ayrıca tüm veri setinde sadece bir defa geçen terimler de bu deneysel çalışmada ayıklanmıştır. Öznitelik seçim metodu olarak, daha önceki bölümlerde kabaca anlatılan CHI2 istatistiği (Chen ve Chen, 2011) kullanılmıştır. CHI2 ile Reuters-21578, 20 Mini Newsgroups, ve 20-Newsgroups veri setleri için seçilen, sırasıyla {500, 1000, 2000, 3000, 4000, 5000, 6000, 7000 ve 8000}, {500, 1000, 2000, 4000, 6000, 8000, 10000, 12000 ve 14000} ve {500, 1000, 2000, 4000, 6000, 9000, 12000, 16000, 20000 ve 25000} öznitelik üzerinde terim ağırlıklandırma şemalarının performansları test edilmiştir. Seçilen terimlerin skor globalleştirmesi ve sıralaması $CHI2_{max}$ ile gerçekleştirilmiştir.

Deneyler, literatürde hali hazırda önerilmiş olan 7 farklı terim ağırlıklandırma şeması (TF-IDF, TF-PB, TF-RF, TF-IDF-ICF, TF-IDF-ICSDF, TF-IGM, SQRT_TF-IGM) ve önerilen iki farklı terim ağırlıklandırma şeması (TF-IGM_{imp} ve SQRT_TF-IGM_{imp}) ile gerçekleştirilmiştir. TF-PB ve TF-RF için, skor globalleştirmede, sırasıyla, TF-PB_{max} and TF-RF_{max} adlı maksimum sınıf-bazlı koleksiyon frekans faktörleri kullanılmıştır. TF-IGM için, Reuters-21578, 20 Mini Newsgroups ve 20-Newsgroups

veri setleri üzerinde gerçekleştirilen deneylerde de λ katsayısı için, sırasıyla, 6.0, 7.0 ve 7.0 değerleri atanmıştır.

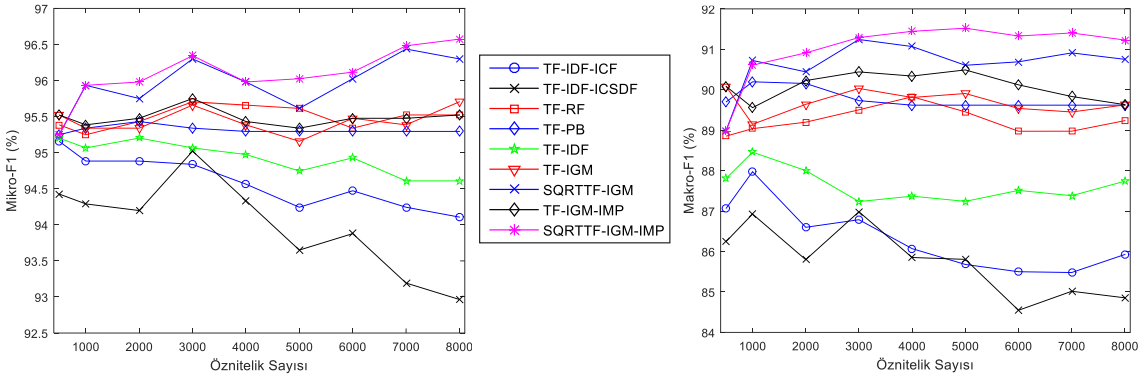
Sınıflandırma aşamasında, daha önceki bölümlerde çalışma stilleri kabaca anlatılan SVM, KNN, ve NN sınıflandırma algoritmaları kullanılmış olup, SVM sınıflandırıcı, gerçekleştirilen tüm deneylerde varsayılan parametrelerle çalıştırılmıştır. KNN için ise bütün veri setlerinde k için 15 değeri atanmıştır. KNN algoritmasında benzerlik ölçütü olarak Kosinüs benzerlik metriği (Prasath vd., 2017) kullanılmıştır. NN sınıflandırıcının kullanıldığı tüm deneyler için seçilen parametreler şöyledir: Öncelikle, her biri 20 sinir hücresine sahip 2 gizli katmandan oluşan çok katmanlı bir yapay sinir ağı yapısı kullanılmıştır. Geri-beslemeli özelliklere sahip bu yapay sinir ağı sınıflandırıcı için, öğrenme algoritması olarak; momentumlu ve adaptif öğrenme oranlı gradyan azalma algoritması (Gradient descent with momentum and adaptive learning rate) kullanılmıştır. Gizli katmanlar için transfer fonksiyonu olarak sigmoid fonksiyon seçilirken, çıktı katmanı için ise doğrusal transfer fonksiyonu tercih edilmiştir. Durdurma kriteri olarak, maksimum iterasyon sayısı 1000 ve minimum hata ise 0.01 olarak belirlenmiştir. Ayrıca rastsallığın etkisini minimuma düşürmek amacıyla, NN sınıflandırıcı her bir deney için 5 defa çalıştırılmış ve elde edilen sonuçların ortalama değerleri alınmıştır.

6.6. Sınıflandırma Sonuçları

Bu bölümde, 3 farklı veri seti ve 3 farklı sınıflandırıcı kullanılarak 9 farklı terim ağırlıklandırma şemasının sınıflandırma performansları, Bölüm 5.4'te daha önce bahsedilen Mikro-F1 ve Makro-F1 değerlendirme ölçütleri yardımıyla ölçülmüştür.

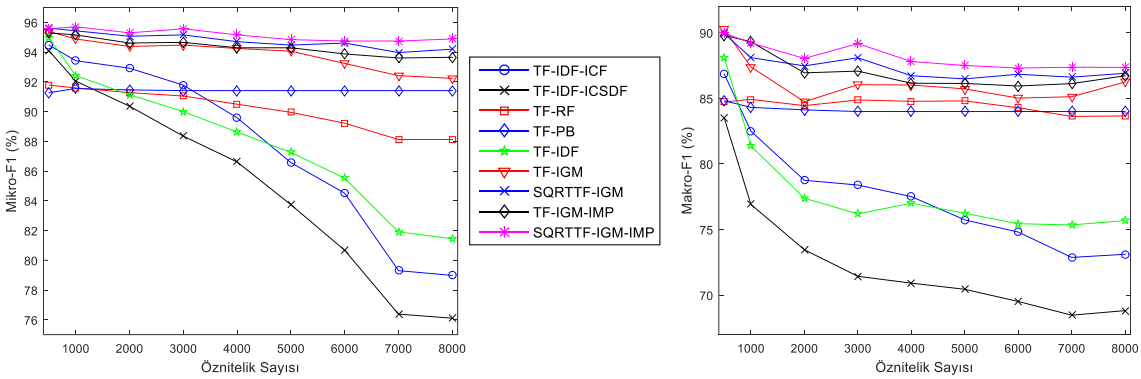
6.6.1. Tüm terim ağırlıklandırma şemaları için performans kıyaslamaları

Reuters-21578 veri seti üzerinde, SVM, KNN ve NN sınıflandırıcılar kullanılarak Mikro-F1 ve Makro-F1 cinsinden elde edilen sınıflandırma sonuçları, sırasıyla, Şekil 6.1, 6.2 ve 6.3'te sunulmuştur.



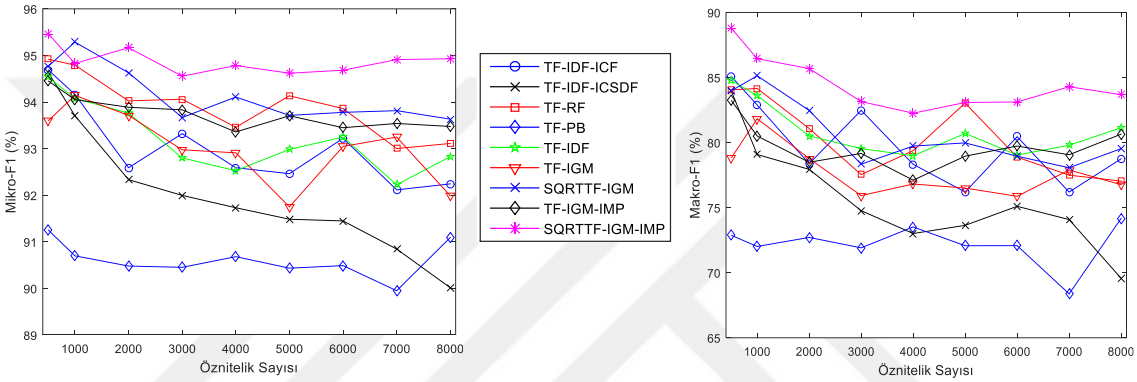
Şekil 6.1. *Reuters-21578 veri seti üzerinde 9 farklı terim ağırlıklandırma şeması ile SVM sınıflandırıcı kullanılarak elde edilen Mikro-F1 ve Makro-F1 sonuçları*

Şekil 6.1 incelendiğinde, Reuters-21578 veri seti üzerinde SVM sınıflandırıcı ile tüm şemalardan elde edilen Mikro-F1 değerleri, tüm boyutlar için genellikle %93'ten yüksek olarak gözlenmiştir. En yüksek Mikro-F1 ve Makro-F1 değerleri, sırasıyla 8000 ve 5000 öznitelik boyutunda, SQRT_TF-IGM$_{imp}$ terim ağırlıklandırma şeması ile elde edilmiştir. Terim ağırlıklandırma şemaları içerisinde en düşük sınıflandırma performansları ise, TF-IDF-ICF ve TF-IDF-ICSDF terim ağırlıklandırma şemaları ile sergilenmiştir. Ayrıca dikkat edilmesi gereken bir diğer nokta ise, TF-PB'nin Mikro-F1 ve Makro-F1 performanslarının, öznitelik sayısı arttıkça diğerlerine nazaran çok daha az değişim göstermesi hatta neredeyse sabitlenmiş olmasıdır. Söz konusu şekilde dikkat edilmesi gereken bir diğer husus ise, önerilen şemaların sınıflandırma performanslarının (SQRT_TF-IGM$_{imp}$ ile TF-IGM$_{imp}$) standart emsallerinden (SQRT_TF-IGM ile TF-IGM) Reuters-21578 veri seti üzerinde SVM sınıflandırıcı ile daha iyi düzeyde ölçülmüş olmasıdır.



Şekil 6.2. *Reuters-21578 veri seti üzerinde 9 farklı terim ağırlıklandırma şeması ile KNN (k=15) sınıflandırıcı kullanılarak elde edilen Mikro-F1 ve Makro-F1 sonuçları*

Reuters-21578 veri seti üzerinde KNN sınıflandırıcı ile TF-IDF, TF-IDF-ICF ve TF-IDF-ICSDF terim ağırlıklandırma şemalarından elde edilen Şekil 6.2'deki Mikro-F1 ve Makro-F1 değerleri incelendiğinde, öznitelik sayısı arttıkça sınıflandırma performanslarının dramatik bir biçimde azaldığı görülmektedir. Bu yüzden ters doküman frekansının yüksek boyutlu öznitelik vektör uzaylarında yetersiz kaldığı yorumu yapılabilir. KNN sınıflandırıcı ile de, en iyi performansın SQRT_TF-IGM$_{imp}$ ile gösterildiğini söylemek mümkündür. TF-PB, KNN sınıflandırıcı ile de sabite yakın bir performans göstermiş, öznitelik sayısının artmasından pek etkilenmemiştir.

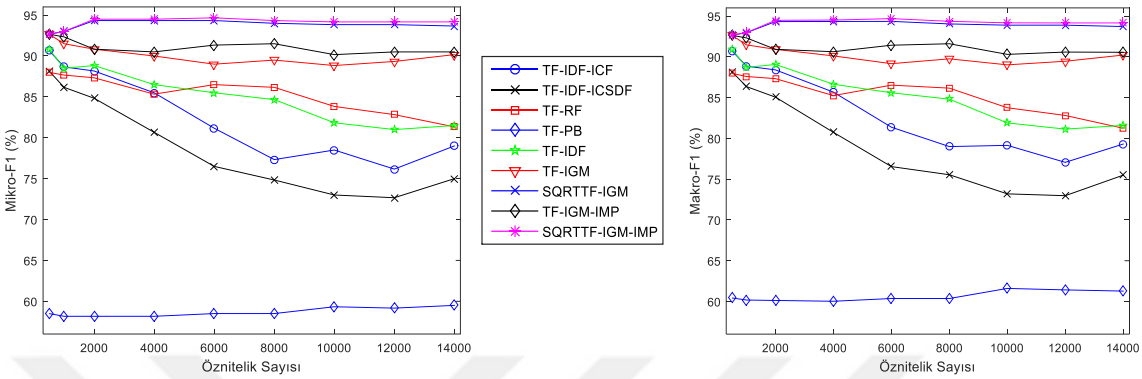


Şekil 6.3. *Reuters-21578 veri seti üzerinde 9 farklı terim ağırlıklandırma şeması ile NN sınıflandırıcı kullanılarak elde edilen Mikro-F1 ve Makro-F1 sonuçları*

NN sınıflandırıcı için de, Reuters-21578 veri seti üzerinde genel olarak en iyi sınıflandırma performansının SQRT_TF-IGM$_{imp}$ ile en kötü performansın ise TF-PB ile elde edildiği Şekil 6.3'te açıkça görülmektedir. Ayrıca, NN sınıflandırıcı ile terim ağırlıklandırma şemalarından elde edilen sınıflandırma başarımlarının farklı öznitelik boyutlarında, diğer sınıflandırıcılara kıyasla daha fazla değiştiği görülmüştür.

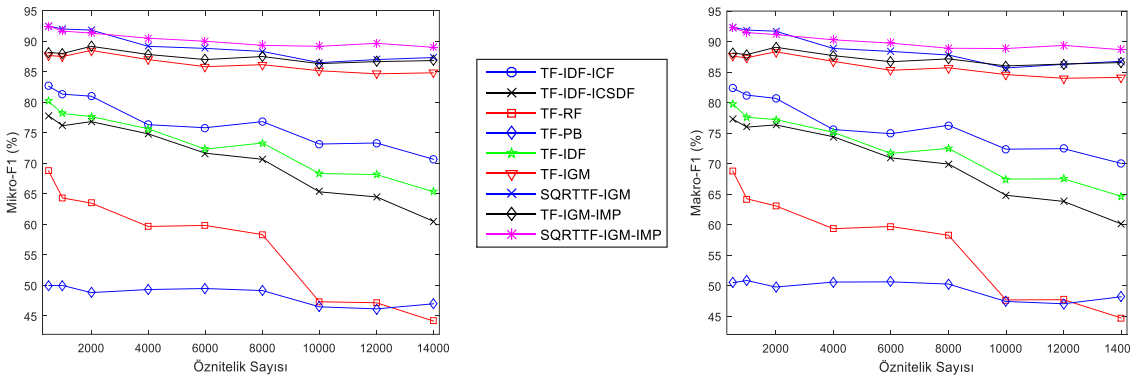
Reuters-21578 veri seti üzerinde tüm sınıflandırıcılar bazında önerilen terim ağırlıklandırma şemalarının performansları göz önüne alındığında, önerilen şemalar (SQRT_TF-IGM$_{imp}$ ile TF-IGM$_{imp}$) ile standart olanları (SQRT_TF-IGM ile TF-IGM) arasındaki Makro-F1 cinsinden performans farklarının Mikro-F1 cinsinden olanlara nazaran daha dikkate değer olduğu gözlenmiştir. Bu bağlamda, KNN, SVM ve NN sınıflandırıcılar ile söz konusu şemalar arasında ölçülen maksimum farklar, sırasıyla, %2.2, %0.9 ve %6.2'dir.

Şekil 6.4, 6.5 ve 6.6'da, 20 Mini Newsgroups veri seti üzerinde, sırasıyla, SVM, KNN ve NN sınıflandırıcılar kullanılarak Mikro-F1 ve Makro-F1 cinsinden elde edilen sınıflandırma sonuçları sunulmuştur.



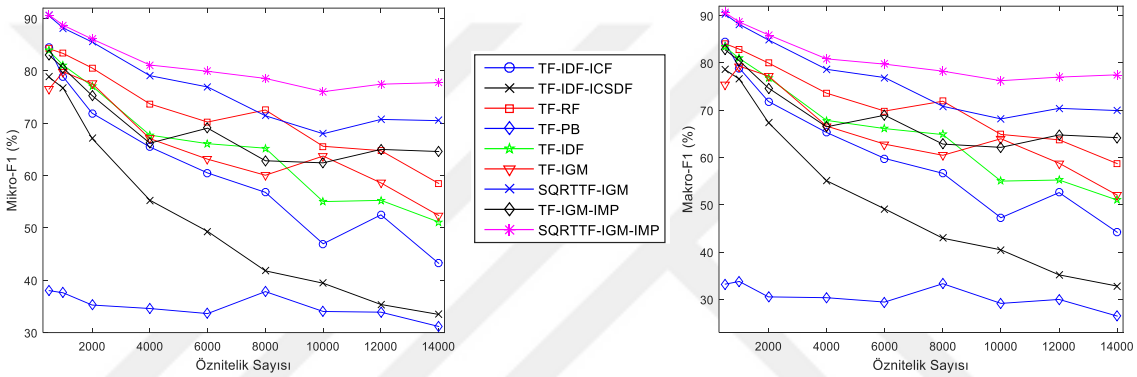
Şekil 6.4. *20 Mini Newsgroups veri seti üzerinde 9 farklı terim ağırlıklandırma şeması ile SVM sınıflandırıcı kullanılarak elde edilen Mikro-F1 ve Makro-F1 sonuçları*

Şekil 6.4 incelendiğinde, SVM sınıflandırıcı ile 20 Mini Newsgroups veri seti üzerinde de, SQRT_TF-IGM$_{imp}$ terim ağırlıklandırma şemasının sınıflandırma performanslarının hem Makro-F1 hem de Mikro-F1 cinsinden diğer şemalara nazaran daha üstün olduğu görülmektedir. Önerilen şemalar ile standart versiyonları arasındaki sınıflandırma başarımları dikkate alındığında, TF-IGM$_{imp}$ ile TF-IGM arasındaki performans farklarının, SQRT_TF-IGM$_{imp}$ ile SQRT_TF-IGM arasında ölçülen performans farklarından daha yüksek olduğu gözlenmiştir. Ayrıca, 20 Mini Newsgroups veri seti üzerinde SVM sınıflandırıcı ile en kötü sınıflandırma performansı TF-PB'ye aittir.



Şekil 6.5. *20 Mini Newsgroups veri seti üzerinde 9 farklı terim ağırlıklandırma şeması ile KNN (k=15) sınıflandırıcı kullanılarak elde edilen Mikro-F1 ve Makro-F1 sonuçları*

KNN sınıflandırıcı için, önerilen şemaların 20 Mini Newsgroups veri seti üzerindeki sınıflandırma performanslarının (SQRT_TF-IGM$_{imp}$ ile TF-IGM$_{imp}$) standart emsallerinden (SQRT_TF-IGM ile TF-IGM) her iki değerlendirme ölçütü cinsinden de daha başarılı olduğu Şekil 6.5'te görülmektedir. Ayrıca IDF tabanlı terim ağırlıklandırma şemalarının sınıflandırma başarımlarının, genel olarak TF-RF terim ağırlıklandırma şemasınınkinden daha üstün olduğu; yine aynı şekli inceleyerek söylenebilir. IDF tabanlı terim ağırlıklandırma şemalarının sınıflandırma performansları, öznitelik sayısı arttıkça hem SVM sınıflandırıcı ile hem de KNN sınıflandırıcı ile genel olarak bir düşüş eğilimi göstermiştir.



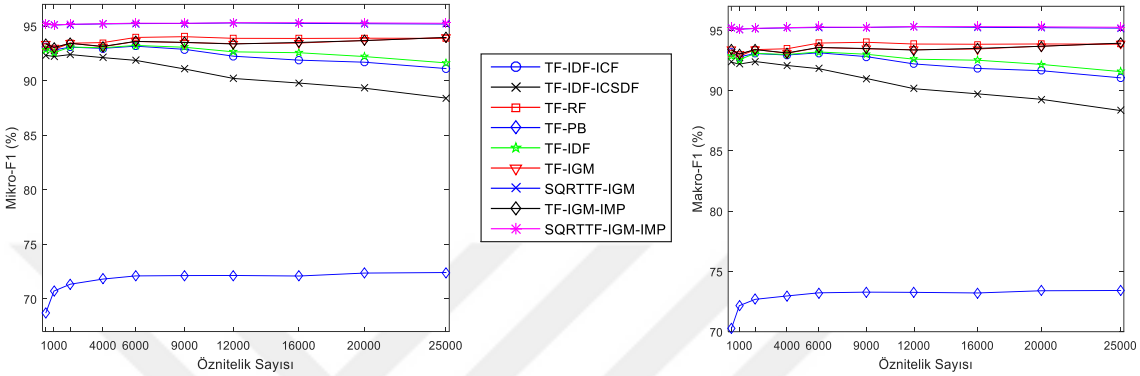
Şekil 6.6. *20 Mini Newsgroups veri seti üzerinde 9 farklı terim ağırlıklandırma şeması ile NN sınıflandırıcı kullanılarak elde edilen Mikro-F1 ve Makro-F1 sonuçları*

20 Mini Newsgroups veri seti üzerinde NN sınıflandırıcı ile tüm terim ağırlıklandırma şemalarından elde edilen Şekil 6.6'daki Makro-F1 ve Mikro-F1 değerleri göz önüne alındığında, öznitelik sayısı arttıkça genel olarak tüm şemaların sınıflandırma başarımlarının dramatik bir şekilde düştüğü görülmektedir. Ayrıca öznitelik boyutları arttıkça, SQRT_TF-IGM$_{imp}$ ile SQRT_TF-IGM şemaları arasındaki performans farkları da artmıştır. 20 Mini Newsgroups veri seti üzerinde tüm sınıflandırıcılar ile diğer şemalara nazaran en düşük performansı gösteren TF-PB için, NN sınıflandırıcı performansının diğer sınıflandırıcılara göre daha düşük olduğu gözlenmiştir.

20 Mini Newsgroups veri seti üzerinde, bütün sınıflandırıcılar için geçerli olmak üzere, önerilen şemaların (SQRT_TF-IGM$_{imp}$ ile TF-IGM$_{imp}$) sınıflandırma performanslarının genel anlamda standart olanlardan (SQRT_TF-IGM ile TF-IGM) hem Mikro-F1 hem de Makro-F1 cinsinden daha üstün olduğu gözlenmiştir. Söz konusu veri seti için, TF-RF ile TF-PB şemalarının KNN sınıflandırıcı ile sergiledikleri sınıflandırma

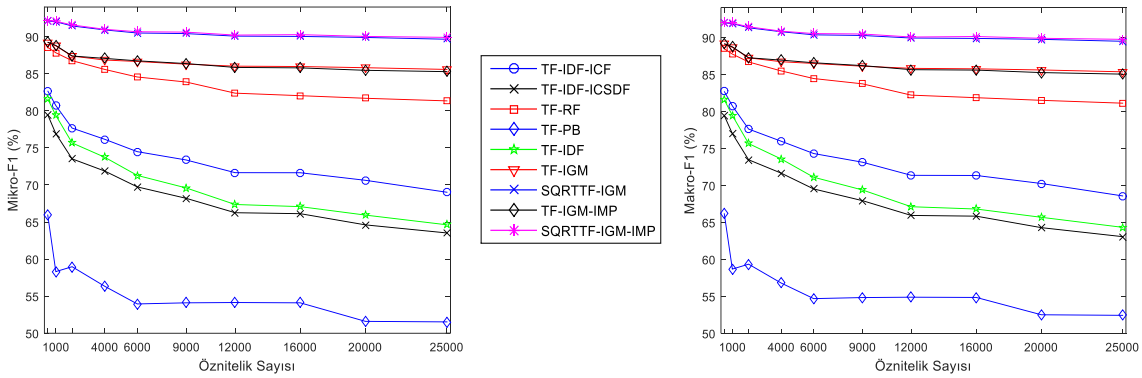
başarımları, SVM sınıflandırıcı ile sergiledikleri sınıflandırma başarımlarından daha düşüktür.

20-Newsgroups veri seti üzerinde, sırasıyla, SVM, KNN ve NN sınıflandırıcılar kullanılarak Mikro-F1 ve Makro-F1 cinsinden elde edilen sınıflandırma sonuçları, sırasıyla Şekil 6.7, 6.8 ve 6.9'da gösterilmiştir.



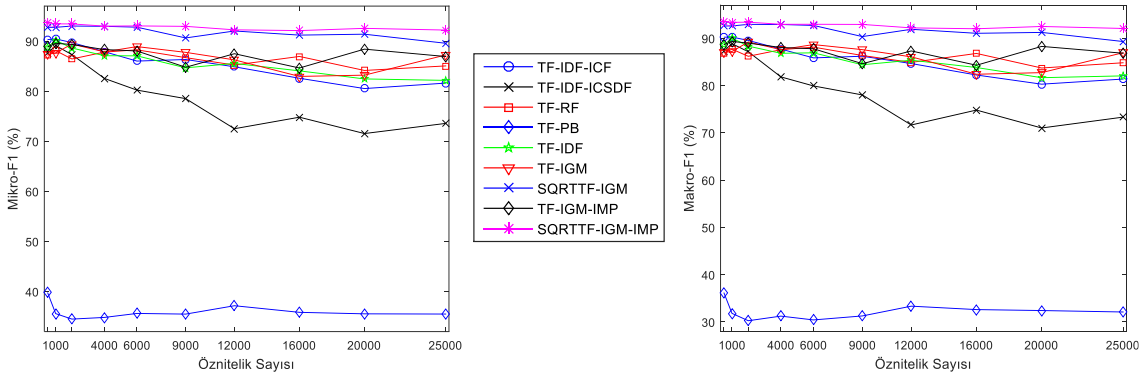
Şekil 6.7. *20-Newsgroups veri seti üzerinde 9 farklı terim ağırlıklandırma şeması ile SVM sınıflandırıcı kullanılarak elde edilen Mikro-F1 ve Makro-F1 sonuçları*

Şekil 6.7'de yer alan 20-Newsgroups veri seti ile SVM sınıflandırıcıdan elde edilen Makro-F1 ve Mikro-F1 skorları incelendiğinde, önerilen SQRT_TF-IGM$_{imp}$ ile standart SQRT_TF-IGM terim ağırlıklandırma şemalarının sınıflandırma sonuçlarının birbirine oldukça yakın olduğu görülmektedir. Burada dikkat edilmesi gereken husus, önerilen SQRT_TF-IGM$_{imp}$ şemasının performansının SQRT_TF-IGM şemasının performansına nazaran çok az farkla daha iyi olmasıdır (Tablo 6.11). Farkların bu denli az olması veri setinin büyüklüğüyle ilişkilendirilebilir. Bu veri seti üzerinde, SVM sınıflandırıcı ile TF-PB hariç tüm şemalardan elde edilen sınıflandırma sonuçlarının, genel olarak %87-96 aralığında olduğunu söylemek mümkündür. Söz konusu veri seti ve sınıflandırıcı için terim ağırlıklandırma şemalarının performansları, genel anlamda, SQRT_TF-IGM-tabanlı şemalar > TF-RF > IGM- tabanlı şemalar > IDF- tabanlı şemalar > TF-PB şeklinde sıralanabilir.



Şekil 6.8. *20-Newsgroups veri seti üzerinde 9 farklı terim ağırlıklandırma şeması ile KNN (k=15) sınıflandırıcı kullanılarak elde edilen Mikro-F1 ve Makro-F1 sonuçları*

Şekil 6.8'de de görüldüğü gibi, önerilen SQRT_TF-IGM$_{imp}$ ile standart SQRT_TF-IGM terim ağırlıklandırma şemalarının 20-Newsgroups veri seti üzerinde KNN sınıflandırıcı ile elde edilen Makro-F1 ve Mikro-F1 skorları da birbirlerine oldukça yakındır. Benzer durum TF-IGM$_{imp}$ ile TF-IGM için de geçerlidir. Her ne kadar önerilen şemalar ile standart versiyonlarının sınıflandırma başarımları birbirine çok yakın olsalar da; önerilen SQRT_TF-IGM$_{imp}$ ve TF-IGM$_{imp}$ şemalarının standart versiyonlarına (SQRT_TF-IGM ve TF-IGM) nazaran nispeten daha üstün performans gösterdiği Tablo 6.12 üzerinden incelenebilir. KNN sınıflandırıcı ile 20-Newsgroups üzerinde tüm şemaların gösterdikleri Mikro-F1 ile Makro-F1 başarımlarını, IGM-tabanlı şemalar > TF-RF > IDF- tabanlı şemalar > TF-PB biçiminde sıralamak mümkündür.



Şekil 6.9. *20-Newsgroups veri seti üzerinde 9 farklı terim ağırlıklandırma şeması ile NN sınıflandırıcı kullanılarak elde edilen Mikro-F1 ve Makro-F1 sonuçları*

20-Newsgroups veri seti üzerinde, önerilen SQRT_TF-IGM$_{imp}$ ve standart SQRT_TF-IGM terim ağırlıklandırma şemalarının NN sınıflandırıcı ile elde edilen

Makro-F1 ve Mikro-F1 başarımları arasındaki farklar, SVM ve KNN ile elde edilenlere nazaran Şekil 6.9 yardımıyla daha net görülebilmektedir. Yani söz konusu veri seti üzerinde, SQRT_TF-IGM$_{imp}$ terim ağırlıklandırma şemasının SQRT_TF-IGM terim ağırlıklandırma şeması üzerindeki üstünlüğü NN sınıflandırıcı ile bariz bir şekilde ortaya çıkmıştır. TF-PB, 20-Newsgroups veri seti üzerinde, yine NN sınıflandırıcı ile SVM ve KNN sınıflandırıcılara nazaran daha kötü bir sınıflandırma performansı sergilemiştir.

20-Newsgroups veri seti için, tüm sınıflandırıcılar düzeyinde bir değerlendirme yapılırsa, öznitelik sayısı arttıkça NN sınıflandırıcı ile diğer iki sınıflandırıcıya nazaran genel olarak tüm şemaların performanslarının daha keskin değişimler gösterdiği söylenebilir. Ayrıca en iyi sınıflandırma başarımlarını gösteren ve performansları SVM ve KNN sınıflandırıcılar ile birbirine oldukça yakın iki terim ağırlıklandırma şeması olan SQRT_TF-IGM$_{imp}$ ve SQRT_TF-IGM arasındaki performans farklarının NN ile daha net görülebildiği yorumu yapılabilir.

Sınıflandırma başarımları test edilen 9 terim ağırlıklandırma şeması tüm veri setleri düzeyinde değerlendirildiğinde, SQRT_TF-IGM$_{imp}$ terim ağırlıklandırma şemasının tüm sınıflandırıcılar ve öznitelik boyutları için genel olarak diğer 8 şemadan daha üstün performans sergilediği açıktır. Ayrıca, TP-PB terim ağırlıklandırma şemasının 20 Mini Newsgroups ve 20 Newsgroups veri setleri üzerindeki performansı Reuters-21578 veri seti üzerindeki performansından açık ara daha düşüktür. Bu durumun, TF-PB'nin dengesiz veri setlerinde daha verimli bir temsili sağlayan kendi ağırlıklandırma yapısından kaynaklı olduğu düşünülmektedir.

6.6.2. Standart IGM ve IGM$_{imp}$ arasında performans kıyaslamaları

Bu kısımda, Reuters-21578, 20 Mini Newsgroups ve 20-Newsgroups veri setleri üzerinde, SVM, KNN ve NN sınıflandırıcılar kullanılarak; önerilen IGM$_{imp}$ ve standart IGM koleksiyon frekansı faktörleri tabanlı terim ağırlıklandırma şemalarının, Mikro-F1 ve Makro-F1 değerleri, sırasıyla, 6.5-6.13 numaralı tablolarda verilmiştir. Söz konusu tablolarda, her bir öznitelik boyutu için elde edilen en iyi performans, koyu olarak işaretlenmiştir.

Tablo 6.5. *Reuters-21578 veri seti üzerinde SVM sınıflandırıcı kullanılarak, gelişmiş IGM tabanlı ve standart IGM tabanlı terim ağırlıklandırma şemalarından elde edilen sınıflandırma performansları*

	Mikro-F1 (%)				Makro-F1 (%)			
Terim Sayısı	TF-IGM	TF-IGM_{imp}	SQRTTF-IGM	SQRTTF-IGM_{imp}	TF-IGM	TF-IGM_{imp}	SQRTTF-IGM	SQRTTF-IGM_{imp}
500	95.52	95.52	95.25	95.25	90.09	90.09	88.98	88.98
1000	95.34	**95.39**	95.93	95.93	89.15	**89.56**	**90.72**	90.61
2000	95.34	**95.48**	95.75	**95.98**	89.64	**90.23**	90.44	**90.91**
3000	95.66	**95.75**	96.30	**96.35**	90.03	**90.44**	91.24	**91.29**
4000	95.39	**95.43**	95.98	95.98	89.81	**90.34**	91.07	**91.45**
5000	95.16	**95.34**	95.61	**96.03**	89.91	**90.49**	90.61	**91.52**
6000	95.48	95.48	96.03	**96.12**	89.54	**90.12**	90.69	**91.33**
7000	95.39	**95.48**	96.44	**96.48**	89.45	**89.83**	90.91	**91.40**
8000	**95.71**	95.52	96.30	**96.57**	89.63	**89.63**	90.75	**91.23**

Tablo 6.6. *Reuters-21578 veri seti üzerinde KNN (k=15) sınıflandırıcı kullanılarak, gelişmiş IGM tabanlı ve standart IGM tabanlı terim ağırlıklandırma şemalarından elde edilen sınıflandırma performansları*

	Mikro-F1 (%)				Makro-F1 (%)			
Terim Sayısı	TF-IGM	TF-IGM_{imp}	SQRTTF-IGM	SQRTTF-IGM_{imp}	TF-IGM	TF-IGM_{imp}	SQRTTF-IGM	SQRTTF-IGM_{imp}
500	**95.39**	95.30	**95.61**	95.57	**90.27**	89.81	89.94	**89.97**
1000	94.88	**95.16**	95.43	**95.71**	87.40	**89.32**	88.10	**89.22**
2000	94.38	**94.61**	95.07	**95.30**	84.74	**86.95**	87.47	**88.04**
3000	94.47	**94.66**	95.16	**95.57**	86.04	**87.08**	88.10	**89.18**
4000	94.24	**94.29**	94.70	**95.16**	86.02	**86.15**	86.73	**87.81**
5000	94.06	**94.29**	94.47	**94.84**	85.72	**86.14**	86.48	**87.51**
6000	93.24	**93.88**	94.61	**94.75**	85.02	**85.94**	86.84	**87.31**
7000	92.42	**93.60**	93.97	**94.75**	85.13	**86.12**	86.62	**87.37**
8000	92.23	**93.65**	94.20	**94.88**	86.24	**86.72**	86.91	**87.36**

Tablo 6.7. *Reuters-21578 veri seti üzerinde NN sınıflandırıcı kullanılarak, gelişmiş IGM tabanlı ve standart IGM tabanlı terim ağırlıklandırma şemalarından elde edilen sınıflandırma performansları*

	Mikro-F1 (%)				Makro-F1 (%)			
Terim Sayısı	TF-IGM	TF-IGM_{imp}	SQRTTF-IGM	SQRTTF-IGM_{imp}	TF-IGM	TF-IGM_{imp}	SQRTTF-IGM	SQRTTF-IGM_{imp}
500	93.60	**94.45**	94.77	**95.47**	78.81	**83.25**	83.98	**88.80**
1000	**94.14**	94.05	**95.29**	94.83	**81.83**	80.50	85.13	**86.45**
2000	93.71	**93.89**	94.63	**95.17**	**78.77**	78.49	82.47	**85.69**
3000	92.97	**93.83**	93.67	**94.56**	75.93	**79.17**	78.34	**83.15**
4000	92.91	**93.36**	94.11	**94.78**	76.82	**77.11**	79.73	**82.26**
5000	91.75	**93.71**	93.71	**94.62**	76.50	**78.96**	79.98	**83.08**
6000	93.05	**93.45**	93.78	**94.68**	75.87	**79.74**	78.95	**83.12**
7000	93.25	**93.54**	93.82	**94.91**	77.90	**79.07**	78.07	**84.29**
8000	91.99	**93.48**	93.63	**94.93**	76.77	**80.65**	79.53	**83.69**

Tablo 6.8. *20 Mini Newsgroups veri seti üzerinde SVM sınıflandırıcı kullanılarak, gelişmiş IGM tabanlı ve standart IGM tabanlı terim ağırlıklandırma şemalarından elde edilen sınıflandırma performansları*

	Mikro-F1 (%)				Makro-F1 (%)			
Terim Sayısı	TF-IGM	TF-IGM_{imp}	SQRTTF-IGM	SQRTTF-IGM_{imp}	TF-IGM	TF-IGM_{imp}	SQRTTF-IGM	SQRTTF-IGM_{imp}
500	92.67	92.67	92.67	92.67	92.63	92.63	92.66	92.66
1000	91.50	**92.33**	93	93	91.46	**92.32**	92.97	92.97
2000	90.83	90.83	94.33	**94.50**	90.92	90.92	94.32	**94.48**
4000	90	**90.50**	94.33	**94.50**	90.12	**90.63**	94.34	**94.51**
6000	89	**91.33**	94.33	**94.67**	89.16	**91.42**	94.35	**94.68**
8000	89.50	**91.50**	94	**94.33**	89.76	**91.61**	94.04	**94.36**
10000	88.83	**90.17**	93.83	**94.17**	89.02	**90.30**	93.88	**94.17**
12000	89.33	**90.50**	93.83	**94.17**	89.46	**90.59**	93.89	**94.16**
14000	90.17	**90.50**	93.67	**94.17**	90.22	**90.56**	93.72	**94.16**

Tablo 6.9. *20 Mini Newsgroups veri seti üzerinde KNN (k=15) sınıflandırıcı kullanılarak, gelişmiş IGM tabanlı ve standart IGM tabanlı terim ağırlıklandırma şemalarından elde edilen sınıflandırma performansları*

	Mikro-F1 (%)				Makro-F1 (%)			
Terim Sayısı	TF-IGM	TF-IGM_{imp}	SQRTTF-IGM	SQRTTF-IGM_{imp}	TF-IGM	TF-IGM_{imp}	SQRTTF-IGM	SQRTTF-IGM_{imp}
500	87.67	**88.17**	**92.50**	92.50	87.59	**88.12**	**92.35**	92.30
1000	87.50	**88**	**92**	91.67	87.37	**87.88**	**91.89**	91.48
2000	88.50	**89.17**	**91.83**	91.33	88.34	**89.07**	**91.71**	91.14
4000	87	**87.83**	89.17	**90.50**	86.77	**87.70**	88.88	**90.30**
6000	85.83	**87**	88.83	**90**	85.31	**86.71**	88.40	**89.75**
8000	86.17	**87.50**	88.33	**89.33**	85.71	**87.17**	87.81	**88.91**
10000	85.17	**86.33**	86.50	**89.17**	84.61	**86.02**	85.68	**88.87**
12000	84.67	**86.67**	87	**89.67**	83.99	**86.32**	86.28	**89.38**
16000	84.83	**86.83**	87.33	**89**	84.14	**86.57**	86.77	**88.68**

Tablo 6.10. *20 Mini Newsgroups veri seti üzerinde NN sınıflandırıcı kullanılarak, gelişmiş IGM tabanlı ve standart IGM tabanlı terim ağırlıklandırma şemalarından elde edilen sınıflandırma performansları*

	Mikro-F1 (%)				Makro-F1 (%)			
Terim Sayısı	TF-IGM	TF-IGM_{imp}	SQRTTF-IGM	SQRTTF-IGM_{imp}	TF-IGM	TF-IGM_{imp}	SQRTTF-IGM	SQRTTF-IGM_{imp}
500	76.47	**83.07**	90.57	**90.73**	75.48	**82.83**	90.33	**90.66**
1000	79.83	**80.57**	88.20	**88.73**	79.32	**80.44**	88.15	**88.65**
2000	**77.60**	75.30	85.53	**86.10**	**77.24**	74.70	84.86	**85.94**
4000	**67.07**	66.20	79.10	**81.13**	**66.65**	66.49	78.65	**80.87**
6000	63.13	**69.07**	76.93	**79.93**	62.83	**68.95**	76.90	**79.75**
8000	60.07	**62.83**	71.43	**78.57**	60.47	**62.84**	70.83	**78.26**
10000	**63.73**	62.43	68.00	**76.03**	**63.97**	62.15	68.17	**76.25**
12000	58.60	**65.00**	70.73	**77.47**	58.80	**64.77**	70.40	**77.01**
14000	52.37	**64.60**	70.50	**77.77**	52.04	**64.20**	69.93	**77.49**

Tablo 6.11. *20-Newsgroups veri seti üzerinde SVM sınıflandırıcı kullanılarak, gelişmiş IGM tabanlı ve standart IGM tabanlı terim ağırlıklandırma şemalarından elde edilen sınıflandırma performansları*

	Mikro-F1 (%)				Makro-F1 (%)			
Terim Sayısı	TF-IGM	TF-IGM_{imp}	SQRTTF-IGM	SQRTTF-IGM_{imp}	TF-IGM	TF-IGM_{imp}	SQRTTF-IGM	SQRTTF-IGM_{imp}
500	93.38	**93.39**	95.18	**95.19**	93.39	**93.40**	95.19	**95.20**
1000	93.04	**93.05**	95.11	95.11	93.02	**93.03**	95.11	95.11
2000	**93.44**	93.43	95.16	**95.18**	**93.42**	93.41	95.15	**95.18**
4000	93.12	**93.15**	95.21	**95.23**	93.10	**93.13**	95.21	**95.23**
6000	93.60	**93.61**	95.24	**95.28**	93.58	**93.59**	95.24	**95.28**
9000	93.50	**93.52**	95.25	**95.27**	93.47	**93.50**	95.25	**95.27**
12000	**93.40**	93.37	95.30	**95.32**	**93.40**	93.37	95.30	**95.31**
16000	93.46	**93.51**	95.26	**95.32**	93.46	**93.51**	95.26	**95.31**
20000	**93.69**	93.68	95.23	**95.30**	**93.69**	93.68	95.23	**95.30**
25000	93.96	**93.97**	95.20	**95.27**	93.93	**93.94**	95.19	**95.26**

Tablo 6.12. *20-Newsgroups veri seti üzerinde KNN (k=15) sınıflandırıcı kullanılarak, gelişmiş IGM tabanlı ve standart IGM tabanlı terim ağırlıklandırma şemalarından elde edilen sınıflandırma performansları*

	Mikro-F1 (%)				Makro-F1 (%)			
Terim Sayısı	TF-IGM	TF-IGM_{imp}	SQRTTF-IGM	SQRTTF-IGM_{imp}	TF-IGM	TF-IGM_{imp}	SQRTTF-IGM	SQRTTF-IGM_{imp}
500	89.25	**89.26**	92.13	**92.18**	89.16	**89.17**	91.98	**92.03**
1000	88.77	**88.82**	92.02	**92.08**	88.69	**88.74**	91.90	**91.96**
2000	87.39	**87.40**	91.45	**91.59**	87.27	87.27	91.32	**91.46**
4000	86.86	**87.08**	90.92	**91.00**	86.73	**86.95**	90.79	**90.87**
6000	86.66	**86.75**	90.50	**90.67**	86.50	**86.60**	90.37	**90.54**
9000	86.30	**86.38**	90.42	**90.62**	86.13	**86.21**	90.28	**90.49**
12000	**86.03**	85.86	90.08	**90.22**	**85.84**	85.66	89.94	**90.07**
16000	**85.98**	85.80	90.04	**90.28**	**85.77**	85.60	89.89	**90.14**
20000	**85.82**	85.47	89.90	**90.04**	**85.62**	85.27	89.76	**89.90**
25000	**85.59**	85.29	89.66	**89.91**	**85.37**	85.06	89.50	**89.76**

Tablo 6.13. *20-Newsgroups veri seti üzerinde NN sınıflandırıcı kullanılarak, gelişmiş IGM tabanlı ve standart IGM tabanlı terim ağırlıklandırma şemalarından elde edilen sınıflandırma performansları*

	Mikro-F1 (%)				Makro-F1 (%)			
Terim Sayısı	TF-IGM	TF-IGM_{imp}	SQRTTF-IGM	SQRTTF-IGM_{imp}	TF-IGM	TF-IGM_{imp}	SQRTTF-IGM	SQRTTF-IGM_{imp}
500	87.34	**88.86**	92.76	**93.67**	87.02	**88.66**	92.59	**93.54**
1000	87.65	**89.56**	92.76	**93.45**	87.20	**89.32**	92.61	**93.36**
2000	89.42	**89.34**	93.02	**93.44**	**89.21**	88.98	92.94	**93.41**
4000	87.86	**88.32**	92.97	**93.00**	87.60	**88.05**	92.92	**92.95**
6000	**88.91**	88.04	92.79	**93.04**	**88.67**	87.78	92.72	**92.96**
9000	**87.76**	84.79	90.64	**92.96**	**87.59**	84.63	90.36	**92.93**
12000	86.26	**87.49**	92.03	**92.26**	86.07	**87.32**	91.90	**92.20**
16000	82.96	**84.64**	91.21	**92.11**	82.39	**84.29**	91.05	**92.01**
20000	83.21	**88.41**	91.39	**92.56**	82.73	**88.28**	91.24	**92.50**
25000	**87.24**	86.93	89.58	**92.21**	**86.95**	86.83	89.31	**92.07**

Tablolardaki verilerden yola çıkılarak, yeni geliştirilen IGM_{imp} stratejisine dayanarak önerilen terim ağırlıklandırma şemalarının genel olarak standart IGM tabanlı terim ağırlıklandırma şemalarından daha üstün performans sergilediklerini söylemek mümkündür. Öznitelik sayısı arttıkça, ekstrem senaryoya sahip terimlerin sayısı da arttığından ve yüksek terim frekansı değerlerinin yarattığı gürültü etkisi karekök terim frekans fonksiyonu ile indirgendiğinden dolayı; önerilen SQRT_TF-IGM_{imp} terim ağırlıklandırma şeması, genel olarak standart IGM tabanlı şemalardan yüksek boyutlu vektör uzaylarında daha iyi bir başarım göstermiştir. Reuters-21578 veri seti üzerinde önerilen şemalar ile standart versiyonları arasındaki maksimum performans farkı %6.2 olarak NN sınıflandırıcı ve 7000 öznitelik ile ölçülürken; söz konusu değerler yine aynı sınıflandırıcı ile 20 Mini Newsgroups ve 20-Newsgroups veri seti üzerinde 14000 ile 20000 öznitelikle sırasıyla %12.23 ile %5.5 olarak ölçülmüştür. Bu yüzden, SQRT_TF-IGM_{imp} terim ağırlıklandırma şemasının yüksek sayıda ayırt edici öznitelik içeren sınıflandırma işlevlerinde de daha etkin bir seçimdir.

6.6.3. Maksimum sayıda öznitelik seçildiğinde önerilen şemalarla göreceli olarak diğer şemalara nazaran edinilen performans kazanımları

Her bir veri seti, sınıflandırıcı ve ölçüm metriği için, seçilen öznitelik sayısı maksimum olduğunda; tüm terim ağırlıklandırma şemalarının sergilediği sınıflandırma başarımları Tablo 6.14'te gösterilmiştir. Ayrıca, benzer kıstaslar ile en iyi performansı gösteren SQRT_TF-IGM_{imp} terim ağırlıklandırma şemasının literatürde mevcut olan diğer 7 terim ağırlıklandırma şeması üzerindeki göreceli sınıflandırma performansı kazanımları da Tablo 6.15'te verilmiştir. Her iki tabloda da yer alan maksimum performans değerleri koyu olarak işaretlenmiştir.

Tablo 6.14. *Seçilen öznitelik sayısı maksimum olduğunda 9 farklı terim ağırlıklandırma şemasından elde edilen sınıflandırma performansları*

Veri Seti (Maximum Terim #)	Sınıflandırı.	Metrik	TF-IDF	TF-IDF-ICSDF	TF-IDF-ICF	TF-PB	TF-RF	TF-IGM	TF-IGM_{imp}	SQR TTF-IGM	SQRT TF-IGM_{imp}
Reuters-21578 (8000)	SVM	Mikro-F1	94.61	92.97	94.11	95.30	95.52	95.71	95.52	96.30	**96.57**
		Makro-F1	87.74	84.86	85.93	89.62	89.24	89.63	89.63	90.75	**91.23**
	KNN	Mikro-F1	81.45	76.11	78.99	91.41	88.12	92.23	93.65	94.20	**94.88**
		Makro-F1	75.69	68.82	73.11	84.00	83.66	86.24	86.72	86.91	**87.36**
	NN	Mikro-F1	92.83	90.01	92.24	91.09	93.11	91.99	93.48	93.63	**94.93**
		Makro-F1	81.14	69.57	78.75	74.18	77.04	76.77	80.65	79.53	**83.69**
20 Mini Newsg. (14000)	SVM	Mikro-F1	81.50	75.00	79.00	59.50	81.33	90.17	90.50	93.67	**94.17**
		Makro-F1	81.60	75.49	79.31	61.28	81.24	90.22	90.56	93.72	**94.16**
	KNN	Mikro-F1	65.33	60.50	70.67	47.00	44.17	84.83	86.83	87.33	**89.00**
		Makro-F1	64.67	60.15	70.06	48.21	44.74	84.14	86.57	86.77	**88.68**
	NN	Mikro-F1	51.13	33.47	43.23	31.17	58.50	52.37	64.60	70.50	**77.77**
		Makro-F1	51.00	32.81	44.19	26.47	58.72	52.04	64.20	69.93	**77.49**
20 Newsgroups (25000)	SVM	Mikro-F1	91.64	88.41	91.12	72.41	93.89	93.96	93.97	95.20	**95.27**
		Makro-F1	91.58	88.35	91.06	73.42	93.86	93.93	93.94	95.19	**95.26**
	KNN	Mikro-F1	64.63	63.53	69.02	51.54	81.34	85.59	85.29	89.66	**89.91**
		Makro-F1	64.34	63.08	68.59	52.44	81.13	85.37	85.06	89.50	**89.76**
	NN	Mikro-F1	82.17	73.58	81.62	35.51	85.06	87.24	86.93	89.58	**92.21**
		Makro-F1	82.05	73.32	81.37	32.11	84.82	86.95	86.83	89.31	**92.07**

Tablo 6.15. *Seçilen öznitelik sayısı maksimum olduğunda SQRT_TF-IGM$_{imp}$ terim ağırlıklandırma şeması ile diğer 7 şema üzerinde sağlanan sınıflandırma performansı kazanımları (%)*

Veri Seti, (Max. Terim Sayısı)	Sınıflandırıcı	Metrikler	TF-IDF	TF-IDF-ICSDF	TF-IDF-ICF	TF-PB	TF-RF	TF-IGM	SQRTTF-IGM
Reuters-21578 (8000)	SVM	Mikro-F1	2.07	**3.87**	2.61	1.33	1.10	0.90	0.28
		Makro-F1	3.98	**7.51**	6.17	1.80	2.23	1.79	0.53
	KNN	Mikro-F1	16.49	**24.66**	20.12	3.80	7.67	2.87	0.72
		Makro-F1	15.42	**26.94**	19.49	4.00	4.42	1.30	0.52
	NN	Mikro-F1	2.26	**5.47**	2.92	4.22	1.95	3.20	1.39
		Makro-F1	3.14	**20.30**	6.27	12.82	8.63	9.01	5.23
20 Mini Newsg. (14000)	SVM	Mikro-F1	15.55	25.56	19.20	**58.27**	15.79	4.44	0.53
		Makro-F1	15.39	24.73	18.72	**53.66**	15.90	4.37	0.47
	KNN	Mikro-F1	36.23	47.11	25.94	89.36	**101.49**	4.92	1.91
		Makro-F1	37.13	47.43	26.58	83.95	**98.21**	5.40	2.20
	NN	Mikro-F1	52.09	132.36	79.90	**149.50**	32.94	48.50	10.31
		Makro-F1	51.94	126.18	75.36	**192.75**	31.97	48.91	10.81
20-Newsgroups (25000)	SVM	Mikro-F1	3.96	7.76	4.56	**31.57**	1.47	1.39	0.07
		Makro-F1	4.02	7.83	4.61	**29.75**	1.50	1.42	0.08
	KNN	Mikro-F1	39.12	41.52	30.27	**74.45**	10.53	5.05	0.28
		Makro-F1	39.51	42.30	30.86	**71.17**	10.64	5.14	0.29
	NN	Mikro-F1	12.22	25.32	12.97	**159.67**	8.41	5.60	2.94
		Makro-F1	12.21	25.57	13.15	**186.73**	8.55	5.89	3.09

Terim ağırlıklandırma şemalarından elde edilen Tablo 6.15'deki göreceli performans kazanımları incelendiğinde, SVM ile elde edilen kazanım yüzdelerinin KNN ve NN sınıflandırıcı ile elde edilenlerden daha düşük olduğu gözlenmiştir. Bunun sebebi, SVM sınıflandırıcının metin sınıflandırma da dahil olmak üzere çoğu sınıflandırma işlevinde KNN ve NN sınıflandırıcılara nazaran daha başarılı bir sınıflandırma algoritmasına sahip olması ile açıklanabilir. Yine aynı tablodaki değerler veri setleri bazında değerlendirilirse; 20 Mini Newsgroups veri seti üzerinde, SQRT_TF-IGM$_{imp}$ ile diğer terim ağırlıklandırma şemaları üzerinde sağlanan göreceli performans kazanımlarının diğer iki veri seti üzerinde (20-Newsgroups ve Reuters-21578) elde edilenlerden genel anlamada daha yüksek olduğu yorumu yapılabilir.

Reuters-21578 veri seti üzerinde en fazla göreceli performans kazanımı, TF-IDF-ICSDF üzerinde elde edilirken, 20-Newsgroups veri seti üzerinde ise TF-PB üzerinde elde edilmiştir. 20 Mini Newsgroups veri seti için ise, en fazla kazanım SVM ve NN sınıflandırıcılar kullanıldığında TF-PB üzerinde, KNN sınıflandırıcı kullanıldığındaysa

TF-RF üzerinde sağlanmıştır. Ayrıca, TF-PB ve TF-RF üzerinde sağlanan performans kazanımlarının TF-IDF-ICSDF üzerinde sağlananlardan açık ara daha yüksek olduğu gözlenmiştir. Bu durum TF-RF ve TF-PB terim ağırlıklandırma şemalarının karakteristiğinden kaynaklı olabilir. Diğer bir deyişle, TF-RF ve TF-PB'nin, terimlere her bir sınıf için, ikili sınıflandırma yaklaşımına göre ağırlık hesabı ve ataması gerçekleştirmesinin; çoklu-sınıflandırma yaklaşımı ile terim ağırlığı hesaplayan TF-IDF-ICSDF gibi şemalara nazaran daha düşük performanslar göstermesinde etkisi olabilir.

6.7. Değerlendirmeler

Bu çalışmada, standart IGM tabanlı terim ağırlıklandırma şemalarının ağırlıklandırma davranışları, geniş kapsamlı bir biçimde analiz edilmiş ve yeni geliştirilmiş IGM_{imp} faktörünü temel alan iki yeni terim ağırlıklandırma şeması (SQRT_TF-IGM_{imp} ve TF-IGM_{imp}) önerilmiştir. Önerilen şemalar, mevcut TF-IGM terim ağırlıklandırma şemasının ekstrem senaryolar için ağırlıklandırma sürecini geliştirirken; genel ağırlıklandırma işlevinde ise yok denecek kadar az (göz ardı edilebilir) değişikliklere yol açmıştır. Önerilen iki yeni terim ağırlıklandırma şeması, standart IGM ağırlıklandırma eşitliğinin paydasına Ters Doküman Balans Frekansı (IDBF) oranı eklenerek mevcut IGM formülünün yeniden düzenlenmesi ile elde edilmiştir.

Önerilen şemaların sınıflandırma performansları, IGM faktörü tabanlı iki standart terim ağırlıklandırma şeması dahil olmak üzere literatürden toplamda yedi farklı şemanın sınıflandırma performansları ile kıyaslanmıştır. Deneyler, üç popüler sınıflandırıcı kullanılarak farklı karakteristiklere sahip üç farklı veri seti üzerinde gerçekleştirilmiştir. Sınıflandırıcı olarak, literatürde yaygın olarak tercih edilen vektör tabanlı sınıflandırıcılar olan SVM, KNN ve NN sınıflandırma algoritmaları tercih edilmiştir. Ağırlıklandırma başarımında öznitelik boyutunun da etkilerini görmek için, deneyler farklı sayılarda özniteliklerle de gerçekleştirilmiştir.

Deneysel sonuçlar, önerilen SQRT_TF-IGM_{imp} terim ağırlıklandırma şemasının, standart TF-IGM ile SQRT_TF-IGM terim ağırlıklandırma şemaları da dahil, tüm diğer şemalardan genel anlamda daha üstün bir sınıflandırma performansı sergilediğini göstermiştir. Bunun haricinde, önerilen bir diğer şema olan TF-IGM_{imp} terim ağırlıklandırma şeması, standart TF-IGM terim ağırlıklandırma şemasından çoğunlukla daha iyi sınıflandırma başarımlarına sahiptir. Ayrıca, özellikle çok yüksek boyutlu öznitelik vektörleri ile yapılan deneylerde, SQRT_TF-IGM_{imp} şemasının diğerlerine

nazaran üstünlüğü dikkate değerdir. Reuters-21578 ve 20-Newsgroups veri setleri üzerinde, maksimum sayıda öznitelik seçildiğinde; en iyi sınıflandırma performanslarına sahip SQRT_TF-IGM$_{imp}$ terim ağırlıklandırma şeması ile en fazla performans kazanımları; sırasıyla, TF-IDF-ICSDF ve TF-PB üzerinde elde edilmiştir. 20 Mini Newsgroups veri seti için söz konusu şema ve deneysel şartlarda sağlanan performans kazanımları ise, SVM ve NN sınıflandırıcılar ile çalışırken TF-PB üzerinde; KNN ile çalışırken ise TF-RF terim ağırlıklandırma şeması üzerinde elde edilmiştir.

7. METİN SINIFLANDIRMA İÇİN YENİ BİR TERİM AĞIRLIKLANDIRMA YAKLAŞIMI: MONO

Bu bölümde, metin sınıflandırma için, MONO adında yeni bir terim ağırlıklandırma yaklaşımı önerilmiştir. Önerilen yöntemde; terimler, iki farklı oranın hesabına ve kullanımına dayanarak ağırlıklandırılmaktadır. Bu oranlardan ilki; terimin maksimum sayıda geçtiği sınıfa ait, geçtiği doküman sayısının söz konusu sınıfın toplam doküman sayısına oranını ifade eden maksimum geçme oranı (Maximum Occurence, MO) iken; diğeri ise geriye kalan sınıflarda, terimin geçmediği dokümanların yine o sınıflara ait toplam doküman sayılarına oranını gösteren geçmeme oranıdır (Non-Occurence, NO). Önerilen koleksiyon frekansı faktörü, MO ve NO oranlarının çarpımına dayandığından; önerilen yeni ağırlıklandırma stratejisi, MONO olarak adlandırılmıştır. Ayrıca, geliştirilen MONO koleksiyon frekansı faktörünü temel alan, TF-MONO ve SRTF-MONO adında iki terim ağırlıklandırma şeması önerilmiştir. Önerilen şemaların sınıflandırma performansları, Reuters-21578, 20-Newsgroups ve WebKB gibi üç farklı metin veri seti üzerinde, SVM ve KNN sınıflandırıcılar ile test edilmiştir. Önerilen TF-MONO ve SRTF-MONO terim ağırlıklandırma şemalarının performansları, TF-IDF, TF-RF, TF-IDF-ICF, TF-IDF-ICSDF ve TF-IGM gibi literatürde mevcut olan beş farklı geleneksel ve popüler ağırlıklandırma şeması ile kıyaslanmıştır.

7.1. Motivasyon

Efektif bir terim ağırlıklandırma şemasından beklenen, ayırt edici terimlere yüksek ağırlık değerleri ataması; diğerlerine ise düşük ağırlık değerleri atamasıdır. Literatürde gözetimli terim ağırlıklandırma şeması olarak önerilen şemalarda; terimlerin sınıf bilgilerini uygun bir biçimde kullanabilen ağırlıklandırma stratejilerinin, sınıflandırma performanslarını dikkate değer bir şekilde arttırabildiği gösterilmektedir. Diğer bir çıkarım ise terimlere her sınıf için binary (ikili) yaklaşımla birden fazla ağırlık üreten TF-RF gibi yöntemlerin performanslarının, terimlere veri setindeki tüm sınıflarla ilişkilerine bağlı olarak tek bir ağırlık skoru üreten TF-IGM gibi yöntemlerin gerisinde kaldığını göstermektedir. Yani tek skor üreten yöntemlerin çoklu skor üreten yöntemlere göre genel olarak daha başarılı olduğu söylenebilir.

Literatürde mevcut olan geleneksel veya popüler terim ağırlıklandırma yöntemlerinin çoğunun ortak özelliği; terimlere skor atarken, koleksiyondaki sınıflarda

veya dokümanlarda bulunma sayılarını veya oranlarını temel alan bir ağırlıklandırma stratejisine sahip olmalarıdır. Genel eğilim sınıf bilgisini etkin bir biçimde kullanmak olduğundan, terim ağırlıklandırma çalışmalarında bu konu en çok araştırılan problemlerden biri durumundadır. Sınıf bilgisini etkin bir biçimde kullanma düşüncesi, terimlerin ayırt etme yeteneğini yansıtmak için mantıklı bir seçim olmasına rağmen tek başına yeterli değildir. Terimlerin gerçekte sahip olduğu ayırt etme potansiyelini yansıtmak için; en çok geçtiği sınıftaki doküman oranı bilgisi dışında, geriye kalan sınıflardaki geçmediği dokümanların dağılım bilgisi de makul bir biçimde ağırlıklandırma hesabına dahil edilmelidir.

7.2. Yeni Terim Ağırlıklandırma Stratejisi: MONO

Motivasyon kısmında da belirtildiği gibi, literatürde önerilen gözetimli terim ağırlıklandırma şemalarının sınıflandırma performansları; terimlerin ağırlıklandırılırken yer aldıkları dokümanların sınıf bilgilerini etkin bir biçimde yansıtabilecek bir ağırlıklandırma stratejisine sahip olmalarıyla doğru orantılıdır. Çoklu sınıflandırmaya uygun bir biçimde terim ağırlıklandıran yöntemlerin ikili yaklaşımla terim ağırlıklandıran yöntemlere nazaran daha başarılı olduğu son yıllarda önerilen şemaların çoğunlukla çoklu sınıflandırmaya uygun şemalar oluşundan anlaşılmaktadır.

Geleneksel veya popüler diye nitelendirilebilecek son yıllarda önerilen terim ağırlıklandırma şemalarının çoğunun ortak özelliği terimlerin geçtikleri sınıf/doküman bilgilerini ağırlıklandırma sürecinde mümkün olduğunca efektif bir biçimde kullanmalarıdır. Örneğin TF-RF terim ağırlıklandırma şeması ile herhangi bir terim ağırlıklandırılırken, terimin geçtiği pozitif ve negatif sınıflardaki doküman sayıları oranı baz alınmaktadır. Benzer şekilde TF-IDF, TF-IDF-ICF ile TF-IDF-ICSDF terim ağırlıklandırma şemaları da ağırlıklandırma sürecinde terimlerin geçtikleri sınıf/doküman bilgilerini farklı biçimlerde kullanmaktadır. Yakın zamanda önerilen TF-IGM terim ağırlıklandırma şeması da terimlerin sınıf bilgilerini makul bir biçimde ağırlıklandırma sürecine dahil etmeye çalışsa da; bir önceki bölümde belirtildiği gibi bazı uç senaryolara sahip terimlere uygun ağırlık atamasında yetersiz kalmaktadır. Efektif terim ağırlıklandırma stratejileri geliştirmeye çalışırken; terimlerin ayırt ediciliklerini daha iyi yansıtmak için sınıf bilgilerini kullanmak etkili bir yoldur ancak tek başına yeterli değildir. Herhangi bir terimin ağırlığı hesaplanırken, geçtiği sınıf veya sınıflardaki doküman dağılımları bilgisi kadar; geçmediği sınıf veya sınıflardaki doküman dağılımı

bilgileri de önemlidir. Çünkü bir terimin ayırt edicilik gücü, tek sınıfa ait dokümanlarda yoğun olarak geçmesi dışında; aynı zamanda da geriye kalan diğer sınıflardaki dokümanlarda da geçmemesine (bulunmamasına) göre artar. Bu durumu iki farklı senaryoya sahip bir örnekle açıklayalım: Bir x teriminin tek bir sınıfa ait tüm dokümanlarda geçtiğini, geriye kalan diğer sınıflara ait dokümanların çok azında geçtiğini düşünelim. Böyle bir durumda x terimi için, ayırt edicilik gücü yüksektir değerlendirmesi yapılabilir. Başka bir senaryoda ise bir y teriminin tek bir sınıfa ait tüm dokümanlarda geçtiğini, geriye kalan diğer sınıflara ait dokümanların hiçbirinde geçmediğini düşünelim. Yani sadece tek sınıfa ait bütün dokümanlarda geçen bir terim olsun y terimi. Bu durumda y teriminin sınıf ayırt edicilik gücü x teriminin ayırt edicilik gücünden daha yüksektir; dolayısıyla da y teriminin ağırlığının, x teriminin ağırlığından daha yüksek olması gerekir. Sadece bu örnek bile, terimlerin geçmediği doküman bilgisinin de ağırlıklandırma sürecinin bir parçası olması gerektiğini göstermektedir.

Yukarıdaki bilgileri derlenip özetlenirse; etkin bir terim ağırlıklandırma şemasının, aşağıdaki özelliklere sahip olması gerektiği söylenebilir:

- Terimlerin sınıf ayırt edebilme güçlerini mümkün olabildiğince yansıtabilecek ağırlıklar hesaplamalı ve atamalıdır.
- Terimlerin sınıf bilgilerini mümkün olabildiğince efektif bir biçimde ağırlıklandırma aşamasında kullanabilmelidir.
- Terimlerin ayırt ediciliklerini daha doğru yansıtabilmek için en çok geçtikleri sınıf veya sınıflardaki doküman dağılımlarına özellikle odaklanmalıdır.
- Terimlerin ayırt ediciliklerini daha doğru yansıtabilmek için, geçmedikleri sınıf veya sınıflardaki doküman dağılımı bilgilerini de ağırlık hesabı sürecine dahil etmelidir.
- Tek bir sınıfta geçen terimlerin ayırt edicilik yetenekleri, söz konusu sınıfın kaç dokümanında geçtiğiyle doğru orantılı olarak arttırılmalıdır.
- Terimlere, veri setindeki tüm sınıflarla olan ilişkisini gösterecek şekilde, tek bir skor ataması gerçekleştirmelidir.

Önerilen terim ağırlıklandırma yaklaşımı da yukarıda belirtilen düşüncelerin ve yine yukarıda da bahsedilen, etkin bir terim ağırlıklandırma şemasının sahip olması gereken özellikler ışığında geliştirilmiştir. Bu stratejinin detaylarını anlatırken; j sayıda

sınıftan oluşan bir metin koleksiyonu için t_i teriminin doküman frekanslarının Eşitlik-7.1'deki gibi gösterildiğini varsayalım.

$$df_{t_i} = \{d_{i1}, d_{i2}, d_{i3}, ..., d_{ij-1}, d_{ij}\} \quad \textbf{(7.1)}$$

Bir t_i teriminin önerilen ağırlıklandırma yaklaşımı ile ağırlıklandırma süreci sırasıyla aşağıdaki aşamaları içermektedir:

1) Öncelikle, t_i teriminin sınıf-tabanlı doküman frekansları büyükten küçüğe olmak üzere sıralanır. Bu sıralamanın Eşitlik 7.2'deki gibi olduğunu varsayalım.

$$sıralanmış\_df_{t_i} = \{d_{i3}, d_{i1,} d_{i4,}, ..., d_{ij}, d_{ij-1}\} \quad \textbf{(7.2)}$$

2) Sıralanan sınıf-tabanlı doküman frekansları; t_i teriminin maksimum sayıda geçtiği sınıf ile geriye kalan sınıflar olmak üzere Eşitlik 7.3'teki gibi iki gruba ayrılır.

$$sıralanmış\_df_{t_i} = \{\ \overset{C_{t_i\_\text{maks}}}{d_{i3}}\ |, \overbrace{\overline{d_{i1}}, \overline{d_{i4}}, ..., \overline{d_{ij}}, \overline{d_{ij-1}}}^{C_{diğer}}\} \quad \textbf{(7.3)}$$

İlk grup Maksimum-Geçme bilgisi (Max-Occurrence, MO) ile ifade edilirken, ikinci grup Geçmeme/Bulunmama bilgisi (Non-Occurrence, NO) ile ifade edilmektedir.

2.1) Maksimum-Geçme (MO) bilgisi hesaplanırken; t_i teriminin maksimum sayıda geçtiği sınıfta yer alan, ilgili terimin geçtiği doküman sayısı ile o sınıftaki toplam doküman sayısı oranı Eşitlik-7.4'deki gibi hesaplanır.

$$MO_{t_i} = \frac{D_{t_i\_\text{maks}}}{D_{total(t_i\_\text{maks})}} \quad \textbf{(7.4)}$$

2.2) Geçmeme/Bulunmama (NO) bilgisi hesaplanırken ise; t_i teriminin maksimum sayıda geçtiği sınıf haricindeki sınıflarda yer alan ilgili terimin geçmediği doküman sayısı ile o sınıftaki toplam doküman sayısı oranı Eşitlik-7.5'teki gibi hesaplanır.

$$NO_{t_i} = \frac{D_{\overline{t_i}}}{D_{total(\overline{t_i})}} \quad \textbf{(7.5)}$$

3) t_i teriminin lokal ağırlığı; ilgili terim için yukarıda hesaplanan *MO* ile *NO* oranlarının çarpımıyla elde edilir (Eşitlik 7.6).

$$MONO_{Lokal}(t_i) = \overbrace{\left[\frac{D_{t_i\_\text{maks}}}{D_{total(t_i\_\text{maks})}}\right]}^{MO_{t_i}} * \overbrace{\left[\frac{D_{\bar{t}_i}}{D_{total(\bar{t}_i)}}\right]}^{NO_{t_i}} \quad \textbf{(7.6)}$$

4) t_i teriminin global ağırlık değeri ise Eşitlik 7.7'deki gibi hesaplanmaktadır.

$$MONO_{Global}(t_i) = \left[1 + \alpha * MONO_{Lokal}(t_i)\right] \quad \textbf{(7.7)}$$

$MONO_{Global}$ ağırlıklandırma stratejisinde yer alan α denge katsayısı, kullanılan veri setine göre global ağırlıklar üretilebilmesi amacıyla formülde yer almakta olup değer aralığı 5.0-9.0 olarak tanımlanmıştır. Söz konusu α denge katsayısı için kullanılacak varsayılan değer ise 7.0 olarak belirlenmiştir.

MONO terim ağırlıklandırma stratejisine bağlı olarak önerilen TF-MONO ile SRTF-MONO terim ağırlıklandırma şemalarının terim ağırlıklandırma formülleri ise Eşitlik 7.8 ile 7.9'daki gibidir.

$$TF - MONO = TF(t_i, d_k) * \left[MONO_{Global}(t_i)\right] \quad \textbf{(7.8)}$$

$$SRTF - MONO = SRTF(t_i, d_k) * \left[MONO_{Global}(t_i)\right] \quad \textbf{(7.9)}$$

Burada $SRTF(t_i, d_k)$, $TF(t_i, d_k)$ değerlerinin karekök fonksiyonu ile indirgenmiş halidir.

7.3 Deneysel Çalışma Ortamı

Önerilen TF-MONO ve SRTF-MONO terim ağırlıklandırma şemalarının sınıflandırma başarımları; Reuters-21578, ve 20-Newsgroups ve WebKB metin veri setleri üzerinde, SVM ve KNN sınıflandırma algoritmaları ile test edilmiştir. Bu çalışmada bir öncekinden farklı olarak iki farklı dengesiz yapıya sahip veri seti ile bir dengeli dağılıma sahip veri seti kullanılmıştır. Amaç, önerilen şemaların farklı dengesiz veri setlerinde üzerinde de etkinliğini göstermektir.

Reuters-21578 veri seti üzerinde, bir önceki çalışmada da olduğu gibi en fazla dokümana sahip 10 sınıfın yer aldığı ModApte (Asuncion ve Newman, 1994) kümesi içerisinde yer alan iki ya da daha fazla sınıf etiketine sahip dokümanlar ayıklanmıştır. Söz konusu dokümanlar ayıklanınca, "wheat" ve "corn" adlı sınıflarda doküman kalmadığından, bu sınıflar silinmiştir. Deneyler kalan 8 sınıf ile gerçekleştirilmiştir.

Deneysel bölümde kullanılan Reuters-21578 veri setine ait bilgiler Tablo 6.2 ile daha önceki bölümde belirtilen bilgilerle eşdeğerdir.

20-Newsgroups veri seti ile yapılan deneylerde, her bir sınıfa ait tüm dokümanlar eğitim ve test için yarı yarıya bölümlendirilerek; daha önceki bölümde Tablo 6.4'te ifade edildiği biçimde kullanılmıştır. Toplamda eğitim için kullanılan eğitim dokümanı sayısı 10000, test dokümanı ise 9997'dir.

WebKB veri seti, 4 farklı üniversitenin Bilgisayar Bilimi bölümleri tarafından derlenen, toplamda 7 ayrı sınıftan web sayfalarını içermektedir (Craven vd., 1998). Deneysel kısımda, WebKB veri setinde yer alan "course", "faculty", "student" ve "project" sınıflarına ait toplam 4199 dokümanın 2803'ü eğitim, 1396'sı ise test için kullanılmıştır. WebKB veri setine ait bilgiler Tablo 7.1'de mevcuttur.

Tablo 7.1. *WebKB veri seti*

No	Sınıf Etiketi	Eğitim Dokümanı #	Test Dokümanı #
1	course	620	310
2	faculty	750	374
3	project	336	168
4	student	1097	544

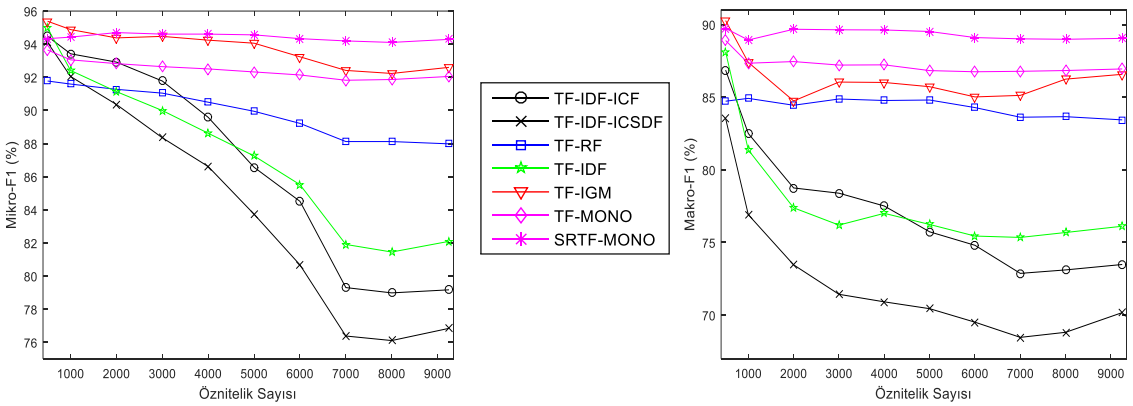
Yukarıda bahsedilen üç metin veri setlerinden elde edilen doküman içeriklerine; sırasıyla, dizgelere ayırma, durak kelimeleri ayıklama, küçük harfe dönüştürme ve köklerine indirgeme gibi ön işlemler uygulanmıştır. Ayrıca tüm veri setlerinde, sadece bir defa geçen terimler ayıklanmıştır. Öznitelik seçimi için ise Ki-Kare (Chen ve Chen, 2011) öznitelik seçim yöntemi kullanılmıştır. Reuters-21578 veri seti için, Ki-Kare (CHI2) ile seçilen, $CHI2_{max}$ globalleştirmesi ile sıralanmış, en yüksek skorlara sahip, sırasıyla, ilk 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000 ve 9237 öznitelik ile deneyler gerçekleştirilirken; 20-Newsgroups için ise yine aynı seçim metodu ve globalleştirme yöntemi ile seçilen ve sıralanmış en yüksek skorlara sahip, ilk 500, 1000, 2000, 4000, 6000, 8000, 10000, 12000, 14000 ve 16000 terim kullanılmıştır. WebKB veri seti üzerindeki deneyler ise, söz konusu metodoloji ile seçilen ilk 100, 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000 terim ile gerçekleştirilmiştir. Farklı öznitelik boyutları ile deneyler gerçekleştirmedeki temel amaç, düşük ve yüksek boyutlarda her bir terim ağırlıklandırma yönteminin, terimlerin ayırt ediciliklerini yansıtma gücünü ve sınıflandırma performanslarını incelemek ve kıyaslamaktır.

Deneyler, önerilen TF-MONO ile SRTF-MONO terim ağırlıklandırma şemalarıyla beraber literatürden mevcut 5 farklı terim ağırlıklandırma şeması (TF-IDF, TF-RF, TF-IDF-ICF, TF-IDF-ICSDF, TF-IGM) da dahil olmak üzere toplamda 7 farklı şema ile gerçekleştirilmiştir. TF-RF için skor globalleştirmede, $TF\text{-}RF_{max}$ ile gösterilen maksimum sınıf-bazlı koleksiyon frekans faktörü kullanılmıştır. TF-IGM ve MONO tabanlı terim ağırlıklandırma şemaları için, Reuters-21578, WebKB dengesiz veri setleri üzerinde gerçekleştirilen deneylerde λ ve α katsayıları için 6.0 değeri kullanılırken, 20-Newsgroups veri seti için ise söz konusu katsayılara varsayılan değer olan 7.0 değeri atanmıştır.

Sınıflandırma aşamasında, SVM ve KNN sınıflandırıcılar kullanılmış olup; SVM sınıflandırıcı, gerçekleştirilen tüm deneylerde varsayılan parametrelerle çalıştırılmıştır. KNN için ise Reuters-21578 ile 20-Newsgroups veri setleri üzerinde gerçekleştirilen deneylerde k parametresine 15 değeri, WebKB üzerinde gerçekleştirilen deneylerde ise 11 değeri atanmıştır. KNN sınıflandırıcının öğrenme algoritmasında ise benzerlik ölçütü olarak Kosinüs (Cosine) benzerliği (Prasath vd., 2017) tercih edilmiştir. Terim ağırlıklandırma şemalarının sınıflandırma başarımlarını ölçmek için ise Makro-F1 ve Mikro-F1 değerlendirme metrikleri kullanılmıştır.

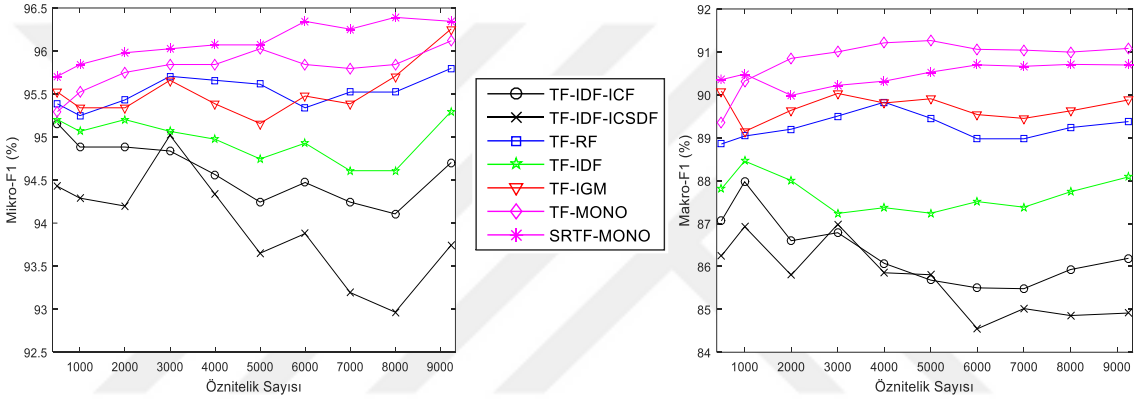
7.4. Sınıflandırma Sonuçları

Reuters-21578 veri seti üzerinde, KNN ve SVM sınıflandırma algoritmaları kullanılarak Mikro-F1 ve Makro-F1 cinsinden elde edilen sınıflandırma sonuçları, sırasıyla, Şekil 7.1 ve 7.2'de sunulmuştur.



Şekil 7.1. *Reuters-21578 veri seti üzerinde 7 farklı terim ağırlıklandırma şeması ile KNN (k=15) sınıflandırıcı kullanılarak elde edilen Mikro-F1 ve Makro-F1 sonuçları*

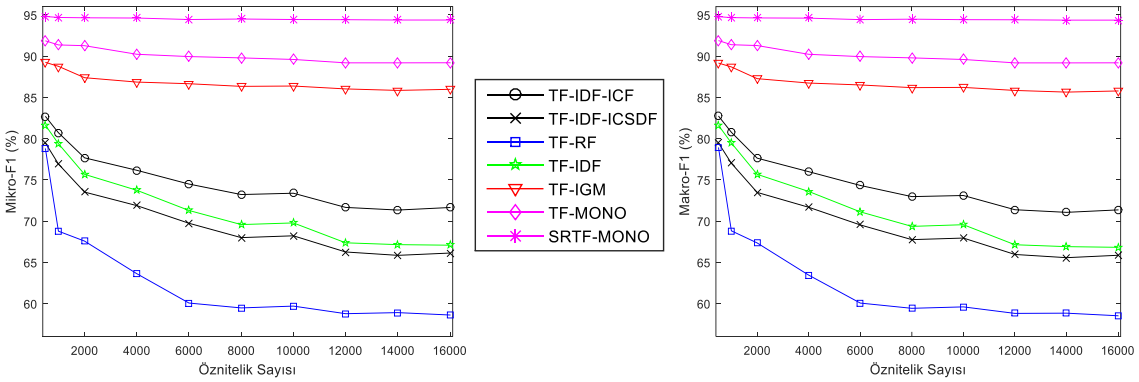
Reuters-21578 veri seti üzerinde *KNN* sınıflandırıcı ile elde edilen Şekil 7.1'deki sınıflandırma sonuçları incelendiğinde; önerilen SRTF-MONO terim ağırlıklandırma şemasının, diğerlerine nazaran performansının daha iyi olduğu açıkça görülmektedir. Aynı veri seti ve sınıflandırıcı için Makro-F_1 sonuçlarına bakılırsa da; önerilen diğer terim ağırlıklandırma yöntemi olan TF-MONO şemasının, güncel TF-IGM terim ağırlıklandırma şemasına kıyasla daha iyi sınıflandırma başarımı gösterdiği açıktır. Söz konusu şekil incelendiğinde; öznitelik sayısı arttıkça, IDF tabanlı terim ağırlıklandırma şemalarının Mikro-F1 ve Makro-F1 değerlerinin dramatik bir biçimde azaldığı görülmüştür.



Şekil 7.2. *Reuters-21578 veri seti üzerinde 7 farklı terim ağırlıklandırma şeması ile SVM sınıflandırıcı kullanılarak elde edilen Mikro-F1 ve Makro-F1 sonuçları*

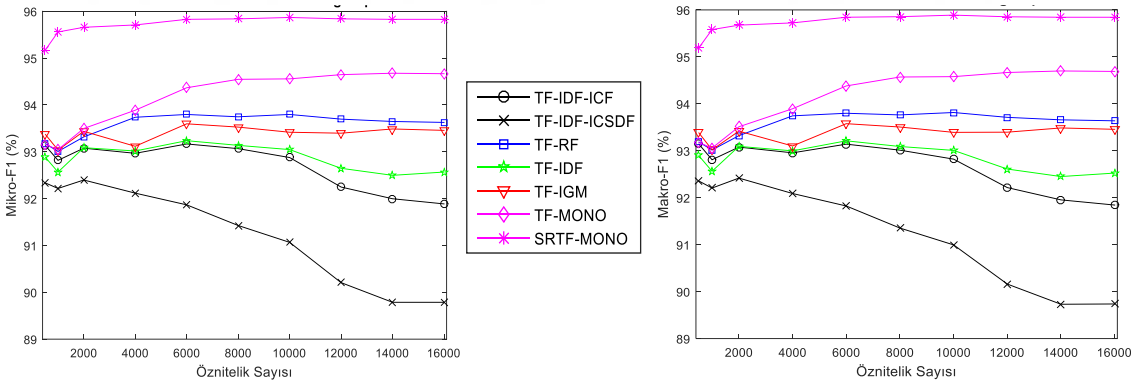
Reuters-21578 veri seti üzerinde *SVM* sınıflandırıcı ile elde edilen Şekil 7.2'deki sınıflandırma sonuçları incelendiğinde; önerilen iki yöntemin de (TF-MONO ve SRTF-MONO) diğer 5 terim ağırlıklandırma yöntemine nazaran daha iyi performans gösterdikleri söylenebilir. Burada SVM sınıflandırıcı ile özellikle TF-IDF-ICF ve TF-IDF-ICSDF yöntemlerinden elde edilen sınıflandırma performanslarının; KNN sınıflandırıcı ile elde edilenlere nazaran genel olarak daha yüksek olduğu ve öznitelik sayısı arttıkça daha tutarlı bir biçimde değiştiğim gösterdiği de görülmektedir. TF-IGM ile TF-RF terim ağırlıklandırma şemalarının, SVM sınıflandırıcı ile genel olarak sınıflandırma başarımları birbirlerine daha yakındır.

20-Newsgroups veri seti üzerinde, KNN ve SVM sınıflandırma algoritmaları kullanılarak, Mikro-F1 ve Makro-F1 cinsinden elde edilen sınıflandırma sonuçları, sırasıyla, Şekil 7.3 ve 7.4'te gösterilmektedir.



Şekil 7.3. *20-Newsgroups veri seti üzerinde 7 farklı terim ağırlıklandırma şeması ile KNN (k=15) sınıflandırıcı kullanılarak elde edilen Mikro-F1 ve Makro-F1 sonuçları*

20-Newsgroups veri seti üzerinde, KNN sınıflandırıcı ile 7 terim ağırlıklandırma şemasından elde edilen Şekil 7.3'teki Mikro-F_1 ve Makro-F_1 performansları; SRTF-MONO > TF-MONO > TF-IGM > TF-IDF-ICF > TF-IDF > TF-IDF-ICSDF > TF-RF şeklinde sıralanabilir. Burada TF-RF'in performansının diğer 6 yöntemin gerisinde kaldığı görülmektedir. Söz konusu şekilde IDF tabanlı şemalardan olan TF-IDF ile TF-IDF-ICSDF terim ağırlıklandırma şemalarının sınıflandırma başarımları birbirlerine oldukça yakın gözlenmiştir.

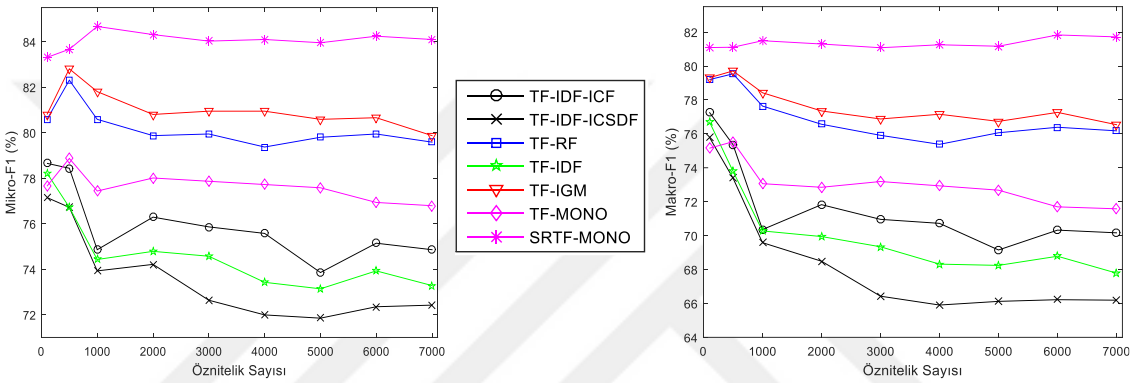


Şekil 7.4. *20-Newsgroups veri seti üzerinde 7 farklı terim ağırlıklandırma şeması ile SVM sınıflandırıcı kullanılarak elde edilen Mikro-F1 ve Makro-F1 sonuçları*

20-Newsgroups veri seti üzerinde SVM sınıflandırıcı ile 7 terim ağırlıklandırma şemasından elde edilen mikro-F_1 ve makro-F_1 performansları ise; SRTF-MONO > TF-MONO > TF-RF > TF-IGM > TF-IDF > TF-IDF-ICF > TF-IDF-ICSDF şeklinde sıralanabilir. KNN sınıflandırıcı ile diğer yöntemlere nazaran en düşük sınıflandırma performansı gösteren TF-RF, SVM sınıflandırıcı ile diğer 4 yönteme kıyasla daha iyi performans göstermiştir. Bu durum terim ağırlıklandırma yöntemlerinin

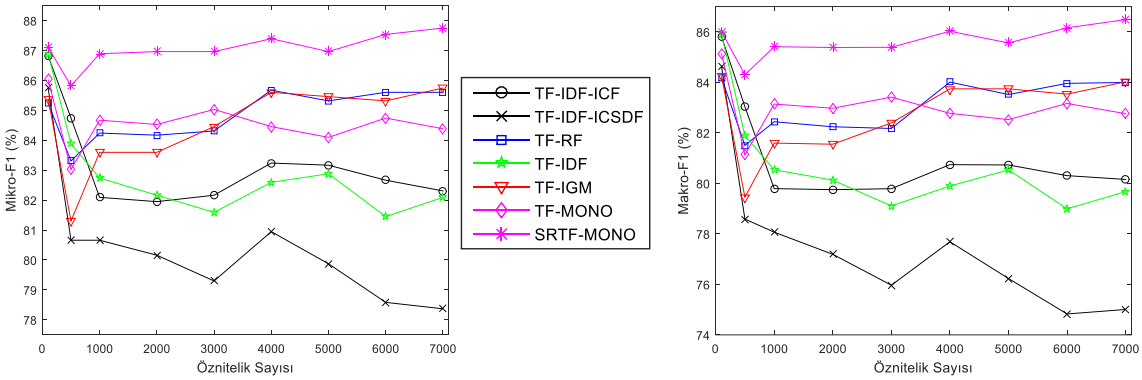
performanslarında, sınıflandırıcıların kabiliyetlerinin de önemli ölçüde etkili olduğunu göstermektedir. Reuters-21578 ve 20-Newsgroups veri setlerinde SVM ve KNN ile yedi farklı şemadan elde edilen sonuçlar değerlendirilirse; özellikle yüksek boyutlu öznitelik vektörlerini sınıflandırma da SVM sınıflandırıcının KNN sınıflandırıcıya göre daha başarılı bir grafik sergilediğini söylemek mümkündür.

WebKB veri seti üzerinde, KNN ve SVM sınıflandırma algoritmaları kullanılarak Mikro-F1 ve Makro-F1 cinsinden elde edilen sınıflandırma sonuçları, sırasıyla, Şekil 7.5 ve 7.6'da gösterilmektedir.



Şekil 7.5. *WebKB veri seti üzerinde 7 farklı terim ağırlıklandırma şeması ile KNN (k=11) sınıflandırıcı kullanılarak elde edilen Mikro-F1 ve Makro-F1 sonuçları*

WebKB veri seti üzerinde KNN sınıflandırıcı ile 7 terim ağırlıklandırma şemasından elde edilen Mikro-F_1 ve Makro-F_1 sonuçları incelendiğinde; önerilen SRTF-MONO terim ağırlıklandırma şemasının sınıflandırma performansının, diğer 6 yönteme göre açık ara daha iyi olduğu görülmektedir. Yüksek TF değerlerinin, SRTF terim frekans faktörü ile indirgenmesinin; SRTT-MONO şemasının diğerleriyle performans farklarını açmasında etkili olduğu söylenebilir. Önerilen diğer yöntem olan TF-MONO terim ağırlıklandırma şemasının performansının ise TF-IGM ve TF-RF terim ağırlıklandırma yöntemlerinin gerisinde kaldığı görülmüştür.



Şekil 7.6. *20-Newsgroups veri seti üzerinde 7 farklı terim ağırlıklandırma şeması ile SVM sınıflandırıcı kullanılarak elde edilen Mikro-F1 ve Makro-F1 sonuçları*

WebKB veri seti üzerinde Şekil 7.6'daki SVM sınıflandırıcı sonuçlarına bakıldığında ise; burada da önerilen SRTF-MONO terim ağırlıklandırma şemasının, diğer 6 terim ağırlıklandırma şemasına göre daha üstün sınıflandırma başarımları sergilediği görülmüştür. Düşük öznitelik boyutlarında, TF-MONO ağırlıklandırma şemasının TF-IGM ve TF-RF şemalarından nispeten daha iyi performans gösterdiği söylenebilir. Genel olarak TF-RF ile TF-IGM terim ağırlıklandırma şemalarının SVM ile sınıflandırma performansları birbirine daha yakın gözlenmiştir. WebKB veri seti üzerinde her iki sınıflandırıcı ile de en düşük sınıflandırma performansı TF-IDF-ICSDF terim ağırlıklandırma şeması ile gözlenmiştir.

7.5. Değerlendirmeler

Bu çalışmada, metin sınıflandırma için, terimlerin hem sınıf bilgilerini daha etkin olarak kullanmaya yönelik, hem de geçmedikleri doküman bilgilerini ağırlıklandırma sürecine dahil ederek ayırt edicilik güçlerini daha iyi yansıtabilmeyi amaçlayan; MONO adında yeni bir ağırlıklandırma stratejisi geliştirilmiştir. MONO ağırlıklandırma stratejisi, terimlerin en çok geçtikleri sınıftaki doküman dağılımına odaklanan MO oranı ile söz konusu terimlerin geriye kalan sınıflarda geçmediği dokümanlardaki dağılımlarını temel alan NO oranının hesabına dayanmaktadır. Terimlerin yerel ağırlık değerlerinin bu iki oranın çarpımı ile hesaplandığı bu çalışmada; TF-MONO ve SRTF-MONO adında iki yeni terim ağırlıklandırma şeması önerilmiştir. Önerilen şemalar, literatürden geneleksel ve popüler beş farklı terim ağırlıklandırma şeması (TF-IDF, TF-RF, TF-IDF-ICF, TF-IDF-ICSDF ve TF-IGM) ile Reuters-21578, 20-Newsgroups ve WebKB veri setleri üzerinde SVM ve KNN sınıflandırma algoritmaları kullanılarak kıyaslanmıştır. Elde

edilen sınıflandırma sonuçları, önerilen SRTF-MONO terim ağırlıklandırma şemasının performansının, diğer tüm şemalardan genel olarak daha üstün olduğunu göstermiştir. Ayrıca önerilen bir diğer terim ağırlıklandırma şeması olan TF-MONO'nun ise, Reuters-21578 ve 20-Newsgroups gibi popüler geniş kapsamlı veri setleri üzerinde, literatürden diğer beş şemaya nazaran daha yüksek Mikro-F1 ve Makro-F1 değerlerine sahip olduğu görülmüştür. Önerilen ağırlıklandırma stratejisinin son yıllarda önerilen yöntemlerde dahil olmak üzere kıyaslandığı tüm şemalardan genel olarak daha başarılı bir performans sergilemesi; bir terimin gerçek ayırt edicilik gücünün, sadece yoğun olarak geçtiği sınıftaki doküman dağılımlarına değil, aynı zamanda daha az geçtiği sınıflardaki geçmediği doküman dağılımlarına da bağlı olduğu fikrini desteklemiştir.

8. SONUÇ VE TARTIŞMA

Bu tez çalışmasında, metin sınıflandırma başarımını arttırmak için, terimlerin sahip oldukları ayırt etme güçlerini vektör uzay modelinde mümkün olduğunca daha verimli bir biçimde yansıtabilmenin yolları araştırılmıştır. Bu bağlamda; terimlerle, yer aldıkları dokümanların ait oldukları sınıflar arasındaki ilişkilerin belirlendiği terim ağırlıklandırma alt alanına yoğunlaşılmıştır. Dolayısıyla, metin sınıflandırma için terim ağırlıklandırma alanına yönelik olarak, literatürde mevcut olan popüler terim ağırlıklandırma şemaları incelenmiştir. İnceleme sırasında, terimleri ağırlıklandırırken izledikleri stratejiler, güçlü oldukları ve yetersiz kaldığı durumlar tespit edilmeye çalışılmıştır. Mevcut terim ağırlıklandırma şemalarının ağırlıklandırma davranışları ve sınıflandırma performansları analiz edilerek; terimlere, söz konusu şemalardan daha verimli bir biçimde ağırlık değerleri hesaplayıp atayabilecek yeni terim ağırlıklandırma şemaları geliştirmenin yolları araştırılmıştır.

Yukarıda bahsedilen hedefler doğrultusunda; tezin ilk geniş kapsamlı deneysel çalışmasında, gözetimli terim ağırlıklandırma şemaları için üç farklı terim frekans faktörü kullanılarak, sınıflandırma performanslarında meydana genel değişimler analiz edilmiştir. Yüksek terim frekansı değerlerine sahip olan terimlerin sayılarının fazla olması, özellikle gözetimli terim ağırlıklandırma şemalarının sahip oldukları sınıflandırma potansiyelini tam olarak yansıtmalarına izin vermeyebilir. Bu durumda, bu değerlerin çeşitli fonksiyonlar yardımıyla makul bir biçimde indirgenmesi gerekmektedir. Deneysel çalışmada, logaritma (LOG_TF) ve karekök (SQRT_TF) gibi farklı terim frekans faktörleri kullanılarak indirgenen ham TF değerlerinin, gözetimli terim ağırlıklandırma şemalarının sınıflandırma performanslarını dikkate değer bir biçimde arttırdığı görülmüştür. Bu bilgi, tezin diğer bölümlerinde önerilen yeni terim ağırlıklandırma şemaları için biz araştırmacılara ışık olmuştur. Yapılan deneysel çalışma; farklı üç terim frekans faktörünün beraber kullanıldığı ve literatürden mevcut popüler şemalar üzerinde performanslara etkisinin incelendiği ilk çalışmadır. Ayrıca kullanılan şemalardan biri olan TF-DFS de, popüler ayırt edici öznitelik seçici (DFS) yönteminden, terim ağırlıklandırmaya ilk defa bu çalışmada uyarlanmıştır. Özellikle SQRT_TF-DFS şemasının diğer şemalara nazaran çoğunlukla daha üstün bir sınıflandırma başarımı göstermesi, DFS yönteminin terim ağırlıklandırma için de etkin bir biçimde kullanılabileceğini göstermiştir.

Tezin ikinci deneysel çalışmasında, son yıllarda önerilmiş popüler bir terim ağırlıklandırma stratejisi olan ters yer çekimi momentine (IGM) dayalı terim ağırlıklandırma (TF-IGM) şeması ele alınmıştır. IGM stratejisinin genel olarak diğer çoğu şemaya nazaran daha başarılı ve makul olan terim ağırlıklandırma davranışlarının, bazı ekstrem senaryolara sahip terimlerin ayırt etme güçlerini yansıtmada yetersiz kaldığı durumlar belirlenmiş ve detaylı bir biçimde örnek senaryolarla gösterilmiştir. Bu kapsamda, söz konusu ağırlıklandırma probleminin, IGM ağırlıklandırma stratejisinin mevcut formülünde yapılabilecek modifikasyonlarla çözülüp çözülemeyeceği araştırılmıştır. Detaylı analizler sonucunda, IGM formülünün paydasına ters doküman balans frekansı (IDBF) adında bir oran eklenerek, mevcut IGM koleksiyon frekans faktörü IGM_{imp} adıyla güncellenmiştir. Yeni geliştirilen IGM_{imp} koleksiyon frekans faktörü ile problemli senaryolara sahip terimlere daha makul ağırlıklar atandığı tablolarla gösterilmiş, ve standart IGM faktörünün, zaten makul olan genel ağırlıklandırma davranışını da pek değiştirmediği örnekle gösterilmiştir. Geliştirilen IGM_{imp} ağırlıklandırma stratejisine bağlı olarak, SQRT_TF-IGM_{imp} ve TF-IGM_{imp} adında iki yeni terim ağırlıklandırma şeması önerilmiştir. Önerilen SQRT_TF-IGM_{imp} ve TF-IGM_{imp} terim ağırlıklandırma şemalarının performansları, ikisi standart IGM tabanlı terim ağırlıklandırma şeması olmak üzere, literatürden toplamda yedi farklı terim ağırlıklandırma şemasının performanslarıyla kıyaslanmıştır. Üç farklı metin veri seti ve üç farklı sınıflandırıcı ile yapılan deneylerden hem Mikro-F1 hem de Makro-F1 cinsinden elde edilen sonuçlar, önerilen stratejinin daha makul olan ağırlıklandırma başarımını sınıflandırma tarafında da desteklemiştir. Özellikle önerilen SQRT_TF-IGM_{imp}, standart IGM tabanlı şemalar da dahil olmak üzere, kıyaslanan 8 şemadan daha üstün bir sınıflandırma performansı sergilemiştir. Ayrıca önerilen TF-IGM_{imp} da standart versiyonundan (TF-IGM) genel olarak daha iyi sınıflandırma başarımlarına sahiptir. Ayrıca en yüksek başarıma sahip olan SQRT_TF-IGM_{imp} şeması, yüksek boyutlu doküman vektörlerinin sınıflandırılmasında da diğer şemalardan daha üstündür. Bu deneysel çalışma, terim ağırlıklandırma için yakın zamanda önerilmiş olan başarılı bir ağırlıklandırma stratejisinin sadece terim frekans faktörü tarafında değil, aynı zamanda koleksiyon frekansı faktörü tarafında da geliştirilebileceğini göstermesi bakımından önemlidir.

Tezin üçüncü deneysel çalışmasında ise, daha önceki deneysel çalışmalar, yapılan analizler ve elde edilen birikimler doğrultusunda; yeni bir terim ağırlıklandırma yaklaşımı geliştirilmiştir. Geliştirme esnasında, popüler TF-IGM şemasının yüksek sınıflandırma başarımları sağlayan, terimlerin sınıf-spesifik doküman dağılımlarının sıralanmasına ve en çok geçtikleri sınıfa odaklanılan ağırlıklandırma stratejisinden ilham alınmıştır. Bunun yanı sıra bir terimin ayırt ediciliğinin yalnızca tek sınıfa ait dokümanlarda sıkça geçmesiyle değil; aynı zamanda diğer sınıflarda geçmemesi durumunda çok yüksek olarak nitelendirilebileceği fikri de geliştirilen yaklaşımda kullanılmıştır. Önerilen yaklaşım, terimlerin en çok geçtikleri (maksimum-occurence, MO) sınıftaki doküman dağılımına ve geriye kalan sınıflardaki geçmedikleri doküman dağılımına (non-occurence, NO) dayanan iki oranın çarpımına bağlı olarak ağırlıklandırma yapmaktadır. MONO olarak adlandırılan bu ağırlıklandırma stratejisine bağlı olarak TF-MONO ve SRTF-MONO adında iki yeni terim ağırlıklandırma şeması önerilmiştir. Önerilen şemaların sınıflandırma performansları, üç farklı veri seti ve iki farklı sınıflandırıcı kullanılarak; TF-IDF, TF-RF, TF-IDF-ICF, TF-IDF-ICSDF ve TF-IGM gibi beş farklı geleneksel ve popüler terim ağırlıklandırma şemasının sınıflandırma performanslarıyla kıyaslanmıştır. Elde edilen sonuçlar, SRTF-MONO terim ağırlıklandırma şemasının, diğer tüm şemalardan kullanılan veri setleri ve sınıflandırıcıların tamamında, daha üstün Mikro-F1 ve Makro-F1 değerlerine sahip olduğunu; TF-MONO'nun ise metin sınıflandırma için önemli olan iki büyük veri setinde diğer beş şemaya göre daha iyi performans sergilediğini göstermiştir. Bu çalışma, terimlerin gerçek ayırt edicilik güçleri, hem koleksiyon frekans faktörü hem de terim frekans faktörü tarafında ne kadar çok yansıtılabilirse, mevcut şemaların metin sınıflandırma başarımlarının o kadar yükselebileceğini göstermiştir.

KAYNAKÇA

Abdel Fattah, M. (2015). New term weighting schemes with combination of multiple classifiers for sentiment analysis. Neurocomputing, 167, 434-442.

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

Agnihotri, D., Verma, K. ve Tripathi, P. (2017). Variable Global Feature Selection Scheme for automatic classification of text documents. Expert Systems with Applications, 81, 268-281.

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

Alsmadi, I. ve Hoon, G. K. (2017). Term weighting scheme for short-text classification: Twitter corpuses. Neural Computing and Applications, 1-13.

Altınçay, H. ve Erenel, Z. (2010). Analytical evaluation of term weighting schemes for text categorization. Pattern Recognition Letters, 31(11), 1310-1323.

Asuncion, A. ve Newman, D. (1994). UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2007).

Badawi, D. ve Altınçay, H. (2014). A novel framework for termset selection and weighting in binary text classification. Engineering Applications of Artificial Intelligence, 35, 38-53.

Badawi, D. ve Altınçay, H. (2017). Termset weighting by adapting term weighting schemes to utilize cardinality statistics for binary text categorization. Applied Intelligence, 47(2), 456-472.

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

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

Chen, Y.-T. ve Chen, M. C. (2011). Using chi-square statistics to measure similarities for text categorization. Expert Systems with Applications, 38(4), 3085-3090.

Craven, M., McCallum, A., PiPasquo, D., Mitchell, T. ve Freitag, D. (1998). Learning to extract symbolic knowledge from the World Wide Web. In: Carnegie-mellon univ pittsburgh pa school of computer Science.

Debole, F. ve Sebastiani, F. (2004). Supervised term weighting for automated text categorization. In Text mining and its applications (pp. 81-97): Springer.

Deisy, C., Gowri, M., Baskar, S., Kalaiarasi, S. ve Ramraj, N. (2010). A novel term weighting scheme MIDF for text categorization. Journal of Engineering Science and Technology, 5, 94-107.

Deng, Z.-H., Luo, K.-H. ve Yu, H.-L. (2014). A study of supervised term weighting scheme for sentiment analysis. Expert Systems with Applications, 41(7), 3506-3513.

Deng, Z.-H., Tang, S.-W., Yang, D.-Q., Zhang, M., Li, L.-Y. ve Xie, K. Q. (2004). A comparative study on feature weight in text categorization. In APWeb (pp. 588-597): Springer.

Dogan, T. ve Uysal, A. K. (2018). The Effects of Globalization Functions on Feature Weighting for Text Classification. In 2018 International Conference on Artificial Intelligence and Data Processing (IDAP) (pp. 1-6): IEEE.

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

Dogan, T. ve Uysal, A. K. (2019b). On Term Frequency Factor in Supervised Term Weighting Schemes for Text Classification. Arabian Journal for Science and Engineering, 1-16.

Emmanuel, M., Khatri, S. M. ve Babu, D. R. R. (2013). A Novel Scheme for Term Weighting in Text Categorization: Positive Impact Factor. In 2013 IEEE International Conference on Systems, Man, and Cybernetics (pp. 2292-2297).

Erenel, Z. ve Altınçay, H. (2012). Nonlinear transformation of term frequencies for term weighting in text categorization. Engineering Applications of Artificial Intelligence, 25(7), 1505-1514.

Fausett, L. V. (1994). *Fundamentals of neural networks: architectures, algorithms, and applications* (Vol. 3): prentice-Hall Englewood Cliffs.

Feng, G., Li, S., Sun, T. ve Zhang, B. (2018). A probabilistic model derived term weighting scheme for text classification. Pattern Recognition Letters, 110, 23-29.

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

Haddoud, M., Mokhtari, A., Lecroq, T. ve Abdeddaïm, S. (2016). Combining supervised term-weighting metrics for SVM text classification with extended term representation. Knowledge and Information Systems, 49(3), 909-931.

Jiang, L., Li, C., Wang, S. ve Zhang, L. (2016). Deep feature weighting for naive Bayes and its application to text classification. Engineering Applications of Artificial Intelligence, 52, 26-39.

Kim, H. K. ve Kim, M. (2016). Model-induced term-weighting schemes for text classification. Applied Intelligence, 45(1), 30-43.

Ko, Y. (2015). A new term-weighting scheme for text classification using the odds of positive and negative class probabilities. Journal of the Association for Information Science and Technology, 66(12), 2553-2565.

Lan, M., Sung, S.-Y., Low, H.-B. ve Tan, C.-L. (2005). A comparative study on term weighting schemes for text categorization. In Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on (Vol. 1, pp. 546-551): IEEE.

Lan, M., Tan, C. L. ve Low, H.-B. (2006). Proposing a new term weighting scheme for text categorization. In AAAI (Vol. 6, pp. 763-768).

Lan, M., Tan, C. L., Su, J. ve Lu, Y. (2009). Supervised and traditional term weighting methods for automatic text categorization. IEEE transactions on pattern analysis and machine intelligence, 31, 721-735.

Lertnattee, V. ve Theeramunkong, T. (2004). Analysis of inverse class frequency in centroid-based text classification. In Communications and Information Technology, 2004. ISCIT 2004. IEEE International Symposium on (Vol. 2, pp. 1171-1176): IEEE.

Li, X., Zhang, A., Li, C., Ouyang, J. ve Cai, Y. (2018). Exploring coherent topics by topic modeling with term weighting. Information Processing & Management, 54(6), 1345-1358.

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

Luo, Q., Chen, E. ve Xiong, H. (2011). A semantic term weighting scheme for text categorization. Expert Systems with Applications, 38(10), 12708-12716.

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

Matsuo, R. ve Ho, T. B. (2018). Semantic term weighting for clinical texts. Expert Systems with Applications, 114, 543-551.

Nguyen, T. T., Chang, K. ve Hui, S. C. (2013). Supervised term weighting centroid-based classifiers for text categorization. Knowledge and information systems, 35(1), 61-85.

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

Pak, A., Paroubek, P., Fraisse, A. ve Francopoulo, G. (2011). Normalization of term weighting scheme for sentiment analysis. In Language and Technology Conference (pp. 116-128): Springer.

Parlak, B. ve Uysal, A. K. (2016). The impact of feature selection on medical document classification. In Information Systems and Technologies (CISTI), 2016 11th Iberian Conference on (pp. 1-5): IEEE.

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

Prasath, V., Alfeilat, H. A. A., Lasassmeh, O. ve Hassanat, A. (2017). Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier-A Review. arXiv preprint arXiv:1708.04321.

Rao, Y., Li, Q., Wu, Q., Xie, H., Wang, F. L. ve Wang, T. (2017). A multi-relational term scheme for first story detection. Neurocomputing, 254, 42-52.

Ren, F. ve Sohrab, M. G. (2013). Class-indexing-based term weighting for automatic text classification. Information Sciences, 236, 109-125.

Sabbah, T., Selamat, A., Selamat, M. H., Al-Anzi, F. S., Viedma, E. H., Krejcar, O. ve Fujita, H. (2017). Modified frequency-based term weighting schemes for text classification. Applied Soft Computing, 58, 193-206.

Sabbah, T., Selamat, A., Selamat, M. H., Ibrahim, R. ve Fujita, H. (2016). Hybridized term-weighting method for dark web classification. Neurocomputing, 173, 1908-1926.

Santhanakumar, M., Columbus, C. C. ve Jayapriya, K. (2018). Multi term based co-term frequency method for term weighting in information retrieval. International Journal of Business Information Systems, 28(1), 79-94.

Sboev, A., Litvinova, T., Gudovskikh, D., Rybka, R. ve Moloshnikov, I. (2016). Machine learning models of text categorization by author gender using topic-independent features. Procedia Computer Science, 101, 135-142.

Schneider, K.-M. (2005). Weighted average pointwise mutual information for feature selection in text categorization. In European Conference on Principles of Data Mining and Knowledge Discovery (pp. 252-263): Springer.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34, 1-47.

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

Singh, T. ve Kumari, M. (2016). Role of Text Pre-processing in Twitter Sentiment Analysis. Procedia Computer Science, 89, 549-554.

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

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

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

Uysal, A. K., Gunal, S., Ergin, S. ve Gunal, E. S. (2012). A novel framework for SMS spam filtering. In Innovations in Intelligent Systems and Applications (INISTA), 2012 International Symposium on (pp. 1-4): IEEE.

Xuan, N. P. ve Le Quang, H. (2014). A New Improved Term Weighting Scheme for Text Categorization. In Knowledge and Systems Engineering (pp. 261-270).

Yue-Heng, S., Pi-Lian, H. ve Zhi-Gang, C. (2004). An improved term weighting scheme for vector space model. In Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.04EX826) (Vol. 3, pp. 1692-1695 vol.1693).

Zhang, L., Jiang, L., Li, C. ve Kong, G. (2016). Two feature weighting approaches for naive Bayes text classifiers. Knowledge-Based Systems, 100, 137-144.

ÖZGEÇMİŞ

Adı Soyadı : Turgut DOĞAN

Yabancı Dil : İngilizce

Doğum Yeri ve Yılı : Hekimhan / 12.06.1989

E-Posta : turgutdogan@eskisehir.edu.tr

Eğitim Geçmişi:

- Doktora (2018-) : Eskişehir Teknik Üniversitesi, Fen Bil. Ens., Bilgisayar Mühendisliği A.B.D.
- Doktora (2014-2018) : Anadolu Üniversitesi, Fen Bil. Ens., Bilgisayar Mühendisliği A.B.D.
- Yüksek Lisans (2011-2014) : Trakya Üniversitesi, Fen Bil. Ens., Bilgisayar Mühendisliği A.B.D.
- Lisans (2006-2010) : Trakya Üniversitesi, Bilgisayar Mühendisliği Bölümü
- Lise (2003-2006) : Yalova Lisesi

Meslek Geçmişi:

- Araştırma Görevlisi (2018-) : Eskişehir Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü
- Araştırma Görevlisi (2014-2018) : Anadolu Üniversitesi, Bilgisayar Mühendisliği Bölümü
- Araştırma Görevlisi (2011-2014) : Trakya Üniversitesi, Bilgisayar Mühendisliği Bölümü

SCI ve SCI-Expanded kapsamındaki dergilerde yayınlanan makaleler:

****Dogan, T.**, & Uysal, A. K. (2019). On Term Frequency Factor in Supervised Term Weighting Schemes for Text Classification. *Arabian Journal for Science and Engineering*, 1-16.

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

Uluslararası bilimsel toplantılarda sunulan ve bildiri kitaplarında basılan bildiriler:

****Doğan, T.** , Uysal, A.. (2018). Tıbbi Metin Dokümanlarının Sınıflandırılmasında Terim Ağırlıklandırma Yöntemlerinin Başarımlarının Kıyaslanması. *Academic Perspective Procedia,* 1 (1), 253-262. DOI: 10.33793/acperpro.01.01.49

Doğan, T., & Uysal, A. K. (2018, September). The Effects of Globalization Functions on Feature Weighting for Text Classification. In *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)* (pp. 1-6). IEEE.

Uluslararası hakemli dergilerde basılan çalışmalar:

Dogan, T., & Uysal, A. K. (2018). The Impact of Feature Selection on Urban Land Cover Classification. *International Journal of Intelligent Systems and Applications in Engineering*, *6*(1), 59-64.

Ulusal bilimsel toplantılarda sunulan ve bildiri kitaplarında basılan bildiriler:

***Doğan, T**., Sert, E., & Taşkın, D. (2013). Araç Destek Sistemleri İçin Kuş Bakışı Görüntü Dönüşümü. *Akademik Bilişim*.

**: Doktora tezinden üretilmiş yayınlar.
*: Yüksek lisans tezinden üretilmiş yayınlar.