©2018 Published in 6th International Symposium on Innovative Technologies in Engineering and Science 09-11 November 2018 (ISITES2018 Alanya – Antalya - Turkey)

Tıbbi Metin Dokümanlarının Sınıflandırılmasında Terim Ağırlıklandırma Yöntemlerinin Başarımlarının Kıyaslanması

Comparing the Performances of Term Weighting Methods on Medical Document Classification

*[1]Turgut Doğan, [1]Alper Kürşat Uysal
Eskişehir Teknik Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Eskişehir, Türkiye

Özet

Günümüzde metin sınıflandırma çalışmalarında elde edilen başarılar, araştırmacıların belirli bir alana yönelik metin dokümanlarının sınıflandırılması konusundaki motivasyonlarını da arttırmaktadır. Tıbbi dokümanların sınıflandırılması problemi bu kapsamdaki araştırma çalışmalarına örnek olarak verilebilir. Tıbbi metin dokümanları denince akla literatürde yaygın olarak kullanılan, tıbbi dergilere ait, çok sınıflı ve çok etiketli dokümanları içeren bibliyografik bir veri tabanı olan MEDLINE gelmektedir. Bu çalışmada, MEDLINE veri tabanında yer alan 10 sınıfa ait dokümanlar kullanılarak bir yüksek lisans tezi kapsamında oluşturulmuş oldukça dengesiz bir yapıya sahip iki adet veri seti kullanılmıştır. Literatürde Türkçe dokümanların sınıflandırılmasına yönelik çalışmaların az olması sebebiyle; deneyler bu veri setlerinin içerdiği hem İngilizce hem de Türkçe dokümanlar ile gerçekleştirilmiştir. Öğrenme algoritması olarak SVM ve KNN olmak üzere 2 farklı sınıflandırıcının kullanıldığı deneylerde, ağırlıklandırma aşamasında ise 7 farklı terim ağırlıklandırma şeması kullanılmış ve sınıflandırma başarımları hem düşük hem de yüksek boyutlarda kıyaslanmıştır. TF-IDF, TF-PB, TF-RF, TF-TRR, TF-IGM, TF-IDF-ICF ve TF-IDF-ICSDF terim ağırlıklandırma yöntemleri ile elde edilen sonuçlar, TF-IGM terim ağırlıklandırma yönteminin genel olarak diğerlerine nazaran daha iyi performans sergilediğini göstermektedir. İki-sınıflı sınıflandırmaya uygun olarak ağırlıklandırma yapan TF-PB ve TF-RF gibi yöntemlerin sınıflandırma başarımları ise çok-sınıflı sınıflandırmaya uygun olarak ağırlıklandırma yapan IDF tabanlı yöntemlerin oldukça gerisinde kalmıştır.

Anahtar Kelimeler: Metin sınıflandırma, Tıbbi doküman sınıflandırma, Medline veritabanı, Terim ağırlıklandırma

Abstract

Today's achievements obtained from text classification studies has increased the motivation of researchers to classify text documents belonging to a specific field. The problem of classification of medical documents can also be given as an example of research studies in this context. MEDLINE is a popular bibliographic database containing multi-class and multi-label documents of medical journals. In this study, two unbalanced datasets constructed as a part of a master thesis including MEDLINE documents belonging to 10 categories are used. As there are low number of studies in the literature about the classification of Turkish documents, the experiments were conducted with not only English documents but also Turkish documents in these two datasets. On experiments, seven different term weighting schemes and well-known SVM and kNN classifiers are used and their performances are compared on low and high dimensional feature spaces. The results obtained from TF-IDF, TF-PB, TF-RF, TF-TRR, TF-IGM, TF-IDF-ICF and TF-IDF-ICSDF show that the performance of TF-IGM weighting method is generally better than the others. The classification performances of the weighting methods which is suitable for binary classification such as TF-PB and TF-RF are worst than the performances of term weighting methods which is suitable for multi-class classification based on IDF.

Key words: Text classification, medical document classification, Medline database, term weighting

*Corresponding author: Address: Faculty of Engineering, Department of Computer Engineering Eskisehir Technical University, 26555, Eskisehir TURKEY. E-mail address: turgutdogan@anadolu.edu.tr, Phone: +902223213550-6583

1. Giriş

Metin sınıflandırma metinsel içeriklere sahip dokümanların bu içeriklere göre etiketleri önceden belirlenmiş sınıflara atanmasıdır. Sınıflandırma işlemi genellikle öznitelik çıkarma, öznitelik seçimi, öznitelik (terim) ağırlıklandırma ve sınıflandırma aşamalarından oluşur. Öznitelik çıkarma aşamasında içerikleri dizgelere ayırma (tokenization), ayrılan dizgeleri küçük harfe dönüştürme (lowercase conversion), dizgelerin içinden durak kelimeleri ayıklama (stopword removal) ve dizgelerin köklerine indirgenmesi (stemming) gibi ön işlemlerin tamamı veya çalışmanın türüne göre birkaçı gerçekleştirilir [1]. Öznitelik seçimi ise özellikle metin dokümanlarının kullanıldığı veri seti çok büyük olduğunda sınıflandırıcının performansını arttırmak için daha ayırt edici özniteliklerin seçilmesi işlemidir [2]. Literatürde öznitelik seçimi için önerilmiş çok sayıda yöntem mevcuttur [3-5]. Başarılı bir metin sınıflandırma için uygun öznitelik seçim yönteminin kullanılması kadar, seçilen özniteliklerin/terimlerin uygun ağırlık değerleriyle ağırlıklanması da çok önemlidir. Bu noktada terim ağırlıklandırma yöntemleri devreye girmektedir. Terim ağırlıklandırma, dokümanlardan ayıklanan terimlere, o dokümanı ve dokümanın ait olduğu sınıfı ayırt edebilme kaabiliyetinin hesaplanıp belirli ağırlık değerlerinin atanmasıdır. Bu ağırlık değerlerinin aralığı seçilen yöntemlerin hesaplama biçimine göre değişmektedir. Bazıları 0-1 aralığında değerler atayabilirken, bazıları ise daha geniş değer aralıklarına sahip olabilmektedir. Bazıları ağırlıklandırma yaparken terimlerin sınıf bilgilerini kullanırken (supervised term weighting methods), bazıları ise (unsupervised term weighting methods) bu bilgiyi göz ardı etmektedir. [6]

Bir terim ağırlıklandırma şeması; terim frekansı faktörü, koleksiyon frekansı faktörü ve normalizasyon faktörü gibi üç temel bileşenden oluşur. Ancak literatürde terim ağırlıklandırma için önerilen yeni yöntemler çoğunlukla ilk iki bileşene yoğunlaşmaktadır. TF-IDF'de bu iki bileşeni baz alan, başta Bilgi Erişimi (Information Retrieval) çalışmaları için önerilen ancak sonrasında terim ağırlıklandırmaya uyarlanan en temel ve eski yöntemlerden biridir [7]. Debole ve Sebastiani ise terim ağırlıklandırma için yeni koleksiyon faktörleri önermenin haricinde koleksiyon faktörü olarak öznitelik seçim yöntemlerinin de kullanılabileceğini savunmuştur [8]. Yapılan bu çalışma sınıf bilgisinin terim ağırlıkladırmada kullanımını teşvik etmesi bakımından önemli bir yere sahiptir. Sun ve arkadaşları geleneksel Vektör Uzay Modeli'nin (Vector Space Model) doküman vektörlerini temsil etmede yetersiz kaldığını ifade ederek, vektör uzay modeli için Bilgi Kazancı'na (Information Gain) dayanan bir terim ağırlıklandırma yöntemi önermişlerdir [9]. Lan ve arkadaşları sınıf bilgisini kullanan Bağıntı Frekansı (Relevance Frequency) olarak adlandırdıkları yeni koleksiyon frekansını içeren TF.RF yöntemini önermişlerdir [10]. Bu çalışmada TF-RF'in, TF-IDF de dahil Chi-square ve Information Gain gibi öznitelik seçim yöntemi tabanlı ağırlıklandırma yöntemlerine kıyasla daha üstün performans sergilediğini göstermişlerdir. Bir diğer çalışmada ise dengesiz bir dağılıma sahip olan veri setlerine yönelik olarak 2 farklı olasılık dağılımından faydalanan TF-PB terim ağırlıklandırma yöntemi önerilmiştir [11]. Bu yöntemin diğerlerinden farkı terimlerin sınıf içi dağılımlarını da ağırlıklandırma hesabına dahil edilmesi olarak vurgulanmıştır. Altınçay ve arkadaşları negatif ve pozitif kategorilerde terimlerin geçme olasılıklarından yararlanarak 6 farklı terim ağırlıklandırma şemasının ağırlıklandırma davranışlarını analiz etmişlerdir [12]. Yapılan analizler sonucunda, farklı şemalarının performans farklarının, ağırlıklandırma yapılırken terim geçme olasılıklarının farkından ve oran kullanma biçimlerinden kaynaklı olduğu ifade edilmiştir. Emmanuel ve arkadaşları bir özniteliğin bir

kategoriye olan pozitif katkısı onun diğer kategorilere ait negatif katkısı hesaplanarak elde edilebileceğini ifade etmiş ve terim ağırlıklandırma için PIF yöntemini önermiştir [13]. PIF yönteminin, aralarında TF-IDF, TF-PB ve TF-RF bulunan 7 yönteme kıyasla sınıflandırma doğruluğu ve sınıflandırma zamanı açısından daha üstün olduğunu göstermişlerdir. Ko pozitif ve negatif sınıf dağılımları bilgisinden yararlanarak sınıf bilgisini kullanan TF.TRR terim ağırlıklandırma yöntemini önermiş, TF.IDF'in birkaç varyasyonu ve TF-RF'ten tutarlı bir şekilde daha üstün performans gösterdiğini ifade etmiştir [14]. Bir başka çalışmada ise Sabbah ve arkadaşları doğru web sayfası sınıflandırma için mTF, mTF-IDF, TF-mIDF ve mTF-mIDF adında 4 farklı terim ağırlıklandırma yöntemi önermiş ve Reuters-21578, 20Newsgroups ve WebKB gibi ünlü metin-sınıflandırma veri setleri üzerinde SVM ve KNN de dahil 4 farklı sınıflandırıcı ile performansları test edilmiştir [15]. Deneysel sonuçlarda önerdikleri şemaların TF, TF-IDF ve Entropi gibi diğer şemalara nazaran önemli ölçüde üstün olduğunu göstermişlerdir.

Metin sınıflandırma bünyesinde özellikle terim ağırlıklandırmada farklı amaçlara yönelik çalışmalar da yapılmaktadır. Bunlardan bazıları duygu analizi ve yazar tanıma olarak ifade edilebilir[16, 17]. Kullanılan veri setinin tipi de bu alanda yapılan çalışmaları alt alan olarak çeşitlendirebilmektedir. Tıbbi dokümanların sınıflandırılması konusu da metin sınıflandırma alanındaki alt başlıklardan biri olarak değerlendirilebilir. Özellikle Türkçe için olmak üzere farklı dillerdeki metin dokümanlarının sınıflandırılmasında terim ağırlıklandırmasına yönelik literatürdeki çalışma sayısının az olması bizim bu çalışmayı yapmamızda motivasyon kaynağı olmuştur. Bu çalışmada Medline veritabanından elde edilen İngilizce ve Türkçe metin dokümanları 7 farklı terim ağırlıklandırma yöntemi kullanılarak sınıflandırılmış, her birinin sınıflandırma performansına etkisi karşılaştırmalı olarak analiz edilmiştir. Bu amaçla, 2. Bölümde kullanılan terim ağırlıklandırma yöntemleri, 3. Bölümde ise sınıflandırıcılar kısaca özetlenmiştir. Veri setleri, öznitelik seçimi, değerlendirme ölçütü ve deneysel sonuçlara ilişkin bilgiler 4. Bölümde verilmiş olup, genel değerlendirme ise 5. Bölümde gerçekleştirilmiştir.

2. Terim Ağırlıklandırma Metotları

Bu bölümde deneylerde kullanılan terim ağırlıklandırma metotları anlatılmıştır. İki-sınıflı (binary) sınıflandırmaya yönelik ağırlıklandırma yapan yöntemlerin çoğunda kullanılan bazı ifadeler ortak olduğundan bu ifadeler aşağıdaki tabloda gösterilmiştir.

Tablo 1. Binary sınıflandırmada bir t_i terimi ile C_j kategorisi arasındaki ilişkinin durumsallık tablosu

Doküman Sayısı	t_i terimini içeren	t_i terimini içermeyen
C_j kategorisine ait olan	a_{ij}	b_{ij}
C_j kategorisine ait olmayan	c_{ij}	d_{ij}

Deneylerde literatürden geleneksel ve güncel olarak toplamda 7 farklı terim ağırlıklandırma şeması kullanılmıştır. Her birinin kısaca tanıtımı aşağıda mevcuttur.

2.1. TF-IDF

TF-IDF, terimlerin terim frekansı (TF) ile ters doküman frekansı (IDF) değerlerinin çarpımıyla elde edilir [7]. Bu ağırlıklandırmada terimlerin sınıf bilgileri kullanılmaz. Dolayısıyla gözetimsiz (unsupervised) bir ağırlıklandırma yöntemidir. Herhangi bir terimin TF-IDF skoru Eşitlik-1'deki gibi hesaplanır:

$$W_{TF-IDF}(t_i) = TF(t_i, d_k) * \log\left(\frac{D}{d(t_i)}\right) \tag{1}$$

Burada $TF(t_i,d_k)$, t_i teriminin d_j dokümanındaki geçme sayısını göstermektedir. D koleksiyondaki toplam doküman sayısını ifade ederken, $d(t_i)$ *ise* t_i teriminin geçtiği doküman sayısını ifade eder.

2.2. TF-PB

Terimlerin sınıf-içi ve sınıflar-arası olasılık dağılımlarından yola çıkılarak hesaplanana 2 orana bağlı olarak gerçekleştirilen bu ağırlıklandırma yöntemi gözetimli ağırlıklandırma yöntemleri grubuna girmektedir [11]. Özellikle dengesiz (unbalanced) veri setlerine ve iki-sınıflı (binary) sınıflandırmaya yönelik olarak ağırlıklandırma yapmak için geliştirilmiş olan bu yöntemde ağırlıklandırma aşağıdaki Eşitlik-2'deki gibi hesaplanır.

$$W_{TF-PB}(t_i) = TF(t_i, d_k) * \max_{j=1}^{C}\left\{\log\left(1 + \frac{a_{ij}}{b_{ij}} * \frac{a_{ij}}{c_{ij}}\right)\right\} \tag{2}$$

Bu formülde C, koleksiyondaki toplam sınıf sayısını ifade etmektedir. Diğer ifadeler Tablo-1 ile Eşitlik-1'de de belirtildiğinden tekrar açıklamaya gerek duyulmamıştır.

2.3. TF-RF

İki-sınıflı sınıflandırmaya uygun olarak ağırlıklandırma yapan ve gözetimli grubuna giren TF-RF yöntemi, pozitif ve negatif sınıflarda terimlerin geçiş oranlarına odaklanır [10]. TF-RF ile ağırlıklandırma formülü Eşitlik-3'de gösterilmiştir.

$$W_{TF-RF}(t_i) = TF(t_i, d_k) * \max_{j=1}^{C}\left\{\log\left(2 + \frac{a_{ij}}{c_{ij}}\right)\right\} \tag{3}$$

Bu eşitlikte ve bir önceki eşitlikte belirtilen max ifadesi, her bir t_i terimi için koleksiyondaki sınıf sayısı kadar hesaplanan ağırlık değerlerinden maksimum olanının atanacağını göstermektedir.

2.4. TF-TRR

TF-TRR, pozitif ve negatif sınıf dağılımlarını kullanarak iki-sınıflı sınıflandırmaya uygun bir

şekilde ağırlıklandırma yapan terim ağırlıklandırma yöntemidir [14]. Ağırlıklandırma formülü Eşitlik-4'teki gibidir.

$$W_{TF-RR}(t_i) = \log\left(TF(t_i, d_k)\right) * \max_{j=1}^{C} \left\{ \log\left(\left(\frac{a_{ij} / (a_{ij} + b_{ij})}{c_{ij} / (c_{ij} + d_{ij})} \right) + 2 \right) \right\} \tag{4}$$

Önerilen şemanın orijinal formülünde terim frekansının logaritma fonksiyonu ile indirgenmiş değerleri kullanıldığından bu çalışmada da aynı şekilde kullanılmıştır.

2.5. TF-IDF-ICF

TF-IDF-ICF, terimlerin geçtiği toplam doküman sayısı bilgisinin yanı sıra geçtiği toplam sınıf sayısı bilgisini de kullanmaya dayanan gözetimli bir yöntemdir. Bu ağırlıklandırma yönteminde; her terim için TF-IDF ile hesaplanan ağırlık değerleri, o terimin ters sınıf frekansı (ICF) değerleriyle çarpılarak terimlerin ağırlık değerlerine ulaşılır [18]. Eşitlik-5'te bu ağırlıklandırmanın formülüzasyonu ifade edilmiştir.

$$W_{TF-IDF-ICF}(t_i) = TF(t_i, d_k) * \left(1 + \log(\frac{D}{d(t_i)}) \right) * \left(1 + \log(\frac{C}{c(t_i)}) \right) \tag{5}$$

Burada $c(t_i)$, t_i teriminin geçtiği toplam sınıf sayısını, C ise veri setindeki toplam sınıf sayısını göstermektedir.

2.6. TF-IDF-ICSDF

Bu ağırlıklandırma yönteminde her terimin TF-IDF ağırlık değeri o terimin ters sınıf uzay yoğunluk frekansı (ICSDF) değeri ile çarpılarak terimlerin ağırlık değerleri elde edilir [18]. Bu yöntemin formülünde bir önceki yönteme göre yer alan en temel fark, ağırlıklandırma yapılırken her terimin geçtiği toplam sınıf sayısı yerine, her bir sınıf için geçtiği doküman sayısı ve o sınıftaki toplam doküman sayısı oranını içermesidir. İlgili ağırlıklandırma formülü Eşitlik-6'da verilmiştir.

$$W_{TF.IDF.ICSDF}(t_i) = TF(t_i, d_k) * \left(1 + \log(\frac{D}{d(t_i)}) \right) * \left(1 + \log(\frac{C}{\sum_{j=1}^{C} \frac{df_{t_{ij}}}{D_j}}) \right) \tag{6}$$

Bu formülde D_j, j'nci sınıftaki toplam doküman sayısını $df_{t_{ij}}$ ise t_i teriminin o sınıfta geçtiği toplam doküman sayısını temsil etmektedir.

2.7. TF-IGM

TF-IGM çok-sınıflı (multi-class) sınıflandırmaya yönelik olarak ağırlıklandırma yapan son yıllarda önerilmiş gözetimli bir ağırlıklandırma yöntemidir. Terimlerin, Ters Yerçekimi Momenti (IGM) hesabına dayanır [19]. Bu yöntemde IGM hesabı yapılırken; herhangi bir terimin her bir sınıf için en az bir kez geçtiği doküman sayıları hesaplanır. Daha sonra bu sayılar büyükten küçüğe sıralanır. Bir terimin IGM değerinin hesaplanması Eşitlik-7'de gösterilmektedir.

$$IGM(t_i) = \frac{f_{i1}}{\sum_{r=1}^{C} f_{ir} * r} \tag{7}$$

Bu formülde f_{ir} *(r=1,2, ..., C)* frekansı terimin sınıf-bazlı doküman frekansını göstermektedir. Yani *r* sırasıyla büyükten küçüğe dizilmiş olan, *r*'nci kategoride t_i terimini içeren metin dokümanı sayısını göstermektedir. Bir terimin *TF-IGM* ağırlığı Eşitlik-8'deki gibi hesaplanır.

$$W_{TF.\mathrm{IGM}}(t_i) = TF(t_i, d_k) * \left(1 + \lambda * IGM(t_i)\right) \tag{8}$$

Burada λ ayarlanabilir sabit bir değer olup referans alınan çalışmada 5.0-9.0 değer aralığında tanımlanmıştır. Varsayılan λ değeri 7.0 olup veri setinin dengesiz bir yapıya sahip olmasından dolayı deneylerde bu değer 6.0 olarak set edilmiştir.

3. Kullanılan Sınıflandırıcılar

Deneylerde sınıflandırma için 2 farklı sınıflandırma algoritması kullanılmış olup, ilgili sınıflandırıcılar bu bölümde özetle anlatılmıştır.

3.1. Destek Vektör Makineleri (SVMs)

SVM, hem iki-sınıflı hem de çok-sınıflı sınıflandırmaya uygun olarak sınıflandırma yapabilen popüler bir sınıflandırma algoritmasıdır. İki-sınıf içeren bir örneklem uzayında; sınıflandırmayı, gerçekleştirmek yani pozitif ve negatif örnekleri birbirinden ayırmak için bir hiper-düzlem kullanır. Deneylerde varsayılan değerler ile libSVM paketinden yararlanılmıştır [20].

3.2. K-En Yakın Komşu Algoritması (kNN)

Metin sınıflandırma araştırmalarında yaygın olarak kullanılan bir diğer sınıflandırıcı da basit bir öğrenme algoritmasına sahip olan k En-yakın Komşu algoritmasıdır. Bu algoritmada bir test dokümanının sınıfı kendisine en yakın k adet komşusuna benzerliğine göre belirlenir. Benzerlik için Öklid (Euclidean Similarity) veya Kosinus (Cosine Similarity) benzerliği gibi çeşitli benzerlik metotları kullanılır. Bu çalışmada gerçekleştirilen deneylerde benzerlik için Kosinus benzerliği kullanılmıştır [21].

4. Deneysel Bilgiler

Bu bölümde deneylerde kullanılan metin veri setleri, öznitelik seçim yöntemi ile öznitelik boyutları, deneysel parametreler, değerlendirme ölçütleri ve deneysel sonuçlar kısaca anlatılmıştır.

4.1. İngilizce ve Türkçe Özet Veri setleri (English and Turkish Abstract Datasets)

Deneylerde, daha önce Anadolu Üniversitesi'nde gerçekleştirilmiş bir yüksek lisans tezi kapsamında oluşturulmuş iki adet veri setinden faydalanılmıştır [22]. Deneylerde kullanılan İngilizce ve Türkçe özet veri setleri, MEDLINE veri seti içerisinde yer alan toplamda 23 kategori içinden en fazla dokümana sahip ilk 10 kategoriye ait toplam 1235 dokümanı içermektedir. Bu dokümanların 788'i eğitim, 344'ü ise test için kullanılmıştır. Giriş bölümünde öznitelik çıkarımı için anlatılan kök bulma haricindeki tüm ön işlemler öznitelik çıkarımı esnasında gerçekleştirilmiştir. Öznitelik seçimi için DFS [23] yöntemi kullanılarak sınıflandırma performansları hem İngilizce hem de Türkçe özet veri setlerinde sırasıyla ilk 100, 300, 500, 1000, 1500, 2000, 3000 ve 4000 öznitelik için elde edilmiştir.

4.2. Değerlendirme Ölçütü (Macro-F1)

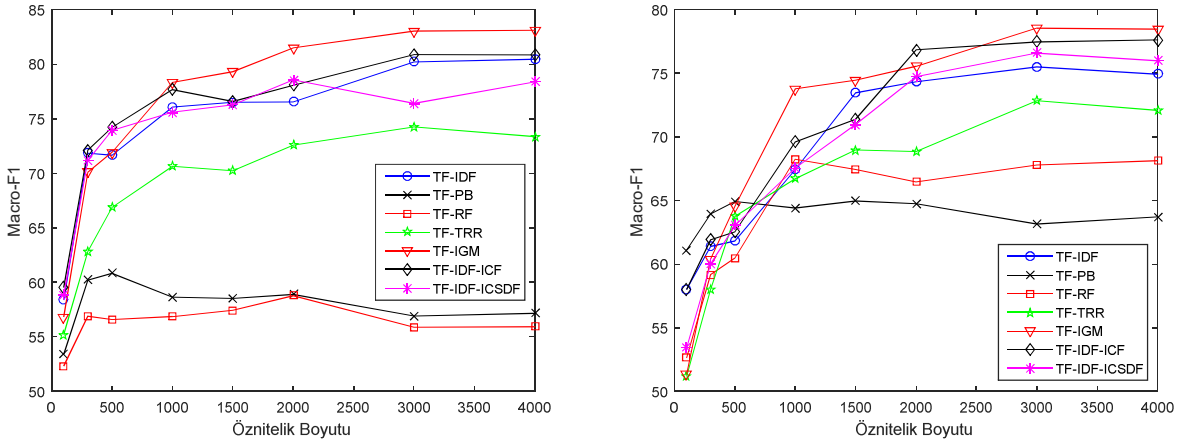
Terim ağırlıklandırma yöntemlerinin sınıflandırma başarımlarını ölçmek için *macro-F1* ölçüm metriği kullanılmıştır. Bu metriğin seçilmesinin sebebi özellikle dengesiz bir yapıya sahip olan yukarıda bahsettiğimiz veri setlerinde daha adil bir performans değerlendirmesi gerçekleştirebilmektir. Eşitlik-9, bir C_k sınıfı için *macro-F1* değerinin hesaplama formülünü göstermektedir.

$$Macro - F1 = \frac{1}{C}\sum_{k=1}^{C}\left\{\frac{2 * TP_{c_k}}{2 * TP_{c_k} + FP_{c_k} + FN_{c_k}}\right\} \tag{9}$$

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

4.3. Sonuçlar

Şekil 1 İngilizce özet dokümanlarının yedi farklı terim ağırlıklandırma metodu ve iki farklı sınıflandırıcı ile edinilen sınıflandırma sonuçlarını göstermektedir. *Macro-F1* cinsinden verilen sonuçlar hem düşük hem de yüksek boyutlarda değerlendirme yapılabilmesi için 100 ile 4000 arasında 8 farklı öznitelik boyutunu içermektedir.

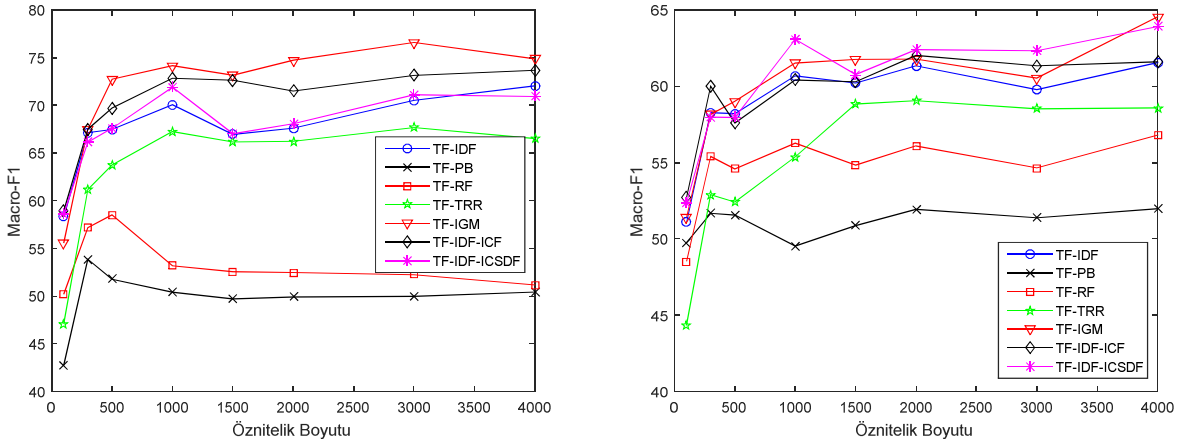


(a) (b)

Şekil 1. İngilizce özet veri seti üzerinde *kNN (k=1)* (a) ve SVM (b) sınıflandırıcı ile 7 farklı terim ağırlıklandırma şeması için elde edilen *macro-F1* sonuçları

İngilizce özet veri seti üzerinde her iki sınıflandırıcı ile elde edilen sonuçlar incelendiğinde genel olarak *TF-IGM* ağırlıklandırma yönteminin diğerlerine oranla daha iyi performansa sahip olduğu görülmektedir. *IDF* tabanlı ağırlıklandırma yöntemlerinin *macro-F1* sonuçları ise ilgili veri seti için iki sınıflandırıcı da da birbirine yakın değerler üretmiştir. *TF-PB* ve *TF-RF,* diğer ağırlıklandırma yöntemlerine nazaran daha düşük *macro-F1* değerlerine sahiptir. Sınıflandırıcılar arasında kıyaslama yapılacak olunursa TF-IGM ve IDF tabanlı terim ağırlıklandırma şemalarının *kNN* sınıflandırıcı ile TF-PB ve TF-RF terim ağırlıklandırma şemalarının ise *SVM* sınıflandırıcı ile daha iyi performanslar gösterdikleri söylenebilir.

Şekil 2 ise Türkçe özet dokümanlarının yedi farklı terim ağırlıklandırma metodu ve iki farklı sınıflandırıcı ile sınıflandırma sonuçlarını göstermektedir.



(a) (b)

Şekil 2. Türkçe özet veri seti üzerinde *kNN (k=1)* (a) ve *SVM* (b) sınıflandırıcı ile 7 farklı terim ağırlıklandırma şeması için elde edilen *macro-F1* sonuçları

Şekil 2'deki *macro-F1* sonuçları; Türkçe özet veri seti için 7 terim ağırlıklandırma şemasının

performanslarının, İngilizce özet veri seti için elde edilen sonuçlardan (Şekil 1) her iki sınıflandırıcı için de genel olarak daha düşük değerlere sahip olduğunu göstermektedir. Şekil 1-a baz alındığında genel performans sıralamasının TF-IGM, IDF tabanlı ağırlıklandırma şemaları, TF-TRR, TF-RF ve TF-PB şeklinde olduğu ifade edilebilir. Aynı veri seti için SVM sınıflandırıcı ile elde edilen Şekil 2-b'deki sonuçlar kNN ile elde edilenlere oranla daha düşük değerler olup bazı öznitelik boyutlarında TF-IDF-ICSDF ağırlıklandırma yönteminin TF-IGM'den ve diğerlerinden daha iyi performans sergilediğini göstermektedir.

5. Tartışma

Bu çalışmada hem Türkçe hem de İngilizce özet dokümanlarını içeren medikal bir veritabanı olan MEDLINE veritabanı üzerinden elde edilen her iki dile ait 2 farklı veri seti üzerinde kNN ve SVM sınıflandırıcı kullanılarak 7 farklı terim ağırlıklandırma şemasının sınıflandırma performansları kıyaslanmıştır. *Macro-F1* cinsinden elde edilen sonuçlar, çok-sınıflı sınıflandırmaya uygun biçimde ağırlıklandırma yapan TF-IGM ya da IDF tabanlı ağırlıklandırma yöntemlerinin; TF-TRR, TF-PB ve TF-RF gibi iki-sınıflı sınıflandırmaya uygun biçimde ağırlıklandırma yapan terim ağırlıklandırma yöntemlerine nazaran her iki veri seti üzerinde de genel olarak daha iyi performans sergilediğini göstermektedir. Genel olarak en iyi performansın TF-IGM'e ait olduğu, en kötü performansı ise TF-PB'in gösterdiği ifade edilebilir. TF-PB'nin dengesiz veri setlerinde daha başarılı bir ağırlıklandırma şeması olmasına rağmen performansının bu çalışmada diğerlerine kıyasla düşük olmasının, kullanılan veri setlerinin özellikleri ile ilgili olduğu düşünülmektedir. Yani veri setindeki sınıflara ait dokümanların içeriklerinin ve sayılarının az olması temsil noktasında TF-PB'nin performansını negatif yönde etkilemiş olması muhtemeldir. Son olarak dil bazında kıyaslama yapılacak olunursa; İngilizce doküman sınıflandırma başarımlarının genel olarak Türkçe dokümanların sınıflandırma başarımlarından daha iyi olduğu değerlendirmesi yapılabilir. Türkçe'nin sondan eklemeli çekimli bir dil olması yani kendine has morfolojik özelliklere sahip olması buna sebep olmuş olabilir.

References

[1] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," Information Processing & Management, vol. 50, no. 1, pp. 104-112, 2014.

[2] G. Feng, J. Guo, B.-Y. Jing, and T. Sun, "Feature subset selection using naive Bayes for text classification," Pattern Recognition Letters, vol. 65, pp. 109-115, 2015.

[3] D. Agnihotri, K. Verma, and P. Tripathi, "Variable Global Feature Selection Scheme for automatic classification of text documents," Expert Systems with Applications, vol. 81, pp. 268-281, 2017.

[4] H. Ogura, H. Amano, and M. Kondo, "Feature selection with a measure of deviations from Poisson in text categorization," Expert Systems with Applications, vol. 36, no. 3, pp. 6826-6832, 2009.

[5] J. Yang, Z. Qu, and Z. Liu, "Improved Feature-Selection Method Considering the Imbalance Problem in Text Categorization," The Scientific World Journal, vol. 2014, p. 17, 2014, Art. no. 625342.

[6] F. Sebastiani, "Machine learning in automated text categorization," ACM computing surveys (CSUR), vol. 34, no. 1, pp. 1-47, 2002.
[7] K. Sparck Jones, "A Statistical Interpretation of Term Specificity and Its Application in Retrieval," Journal of Documentation, vol. 28, no. 1, pp. 11-21, 2004.
[8] F. Debole and F. Sebastiani, "Supervised term weighting for automated text categorization," in Text mining and its applications: Springer, 2004, pp. 81-97.
[9] S. Yue-Heng, H. Pi-Lian, and C. Zhi-Gang, "An improved term weighting scheme for vector space model," in Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.04EX826), 2004, vol. 3, pp. 1692-1695 vol.3.
[10] M. Lan, C. L. Tan, J. Su, and Y. Lu, "Supervised and traditional term weighting methods for automatic text categorization," IEEE transactions on pattern analysis and machine intelligence, vol. 31, no. 4, pp. 721-735, 2009.
[11] Y. Liu, H. T. Loh, and A. Sun, "Imbalanced text classification: A term weighting approach," Expert Systems with Applications, vol. 36, no. 1, pp. 690-701, 2009.
[12] H. Altınçay and Z. Erenel, "Analytical evaluation of term weighting schemes for text categorization," Pattern Recognition Letters, vol. 31, no. 11, pp. 1310-1323, 2010.
[13] M. Emmanuel, S. M. Khatri, and D. R. R. Babu, "A Novel Scheme for Term Weighting in Text Categorization: Positive Impact Factor," presented at the 2013 IEEE International Conference on Systems, Man, and Cybernetics, 2013.
[14] Y. Ko, "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, vol. 66, no. 12, pp. 2553-2565, 2015.
[15] T. Sabbah et al., "Modified frequency-based term weighting schemes for text classification," Applied Soft Computing, vol. 58, pp. 193-206, 2017.
[16] Q. Luo, E. Chen, and H. Xiong, "A semantic term weighting scheme for text categorization," Expert Systems with Applications, vol. 38, no. 10, pp. 12708-12716, 2011.
[17] Z.-H. Deng, K.-H. Luo, and H.-L. Yu, "A study of supervised term weighting scheme for sentiment analysis," Expert Systems with Applications, vol. 41, no. 7, pp. 3506-3513, 2014.
[18] F. Ren and M. G. Sohrab, "Class-indexing-based term weighting for automatic text classification," Information Sciences, vol. 236, pp. 109-125, 2013.
[19] K. Chen, Z. Zhang, J. Long, and H. Zhang, "Turning from TF-IDF to TF-IGM for term weighting in text classification," Expert Systems with Applications, vol. 66, pp. 245-260, 2016.
[20] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," ACM transactions on intelligent systems and technology (TIST), vol. 2, no. 3, p. 27, 2011.
[21] V. Prasath, H. A. A. Alfeilat, O. Lasassmeh, and A. Hassanat, "Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier-A Review," arXiv preprint arXiv:1708.04321, 2017.
[22] B. Parlak, "Classification of Medical Documents According to Diseases," Master's Thesis, Computer Engineering, Anadolu University Graduate School of Sciences, 2016.
[23] A. K. Uysal and S. Gunal, "A novel probabilistic feature selection method for text classification," Knowledge-Based Systems, vol. 36, pp. 226-235, 2012.