Medline Veritabanı Üzerinde Bulunan Tıbbi Dokümanların Kanser Türlerine Göre Otomatik Sınıflandırılması

Ahmet HALTAŞ[1], Ahmet ALKAN[2]

[1]Gaziantep Üniversitesi, Teknik Bilimler Meslek Yüksekokulu, Gaziantep, Türkiye.
[2]KSÜ, Elektrik Elektronik Mühendisliği Bölümü, Kahramanmaraş, Türkiye.
ahaltas18@gmail.com , aalkan05@gmail.com
(Geliş/Received: 31.01.2016; Kabul/Accepted: 24.04.2016)
DOI: 10.17671/btd.27567

Özet— Tıp araştırmacıları tarafından sık kullanılan bir arama motoru olan Pubmed, MEDLINE veri tabanında üzerinde sorgulama yapmaktadır. MEDLINE medikal, biyoloji ve genetik alanındaki çalışmaları içeren ve sürekli güncel tutulan bibliyografik bir veri tabanıdır. İçerdiği yüksek hacimdeki yapısal olmayan metinler sebebiyle, MEDLINE veri tabanı veya belli bölümleri üzerinde pek çok metin sınıflandırma çalışmaları mevcuttur. Bu çalışmada kanser türleri hakkında yazılmış makale özetlerini inceleyerek makalenin hangi kanser türüyle ilgili olduğunu otomatik bulan bir metot geliştirilmiştir. Metodu eğitmek ve test etmek için MEDLINE veri tabanı üzerinde 25962 makale özeti, Pubmed arama motoru üzerinden ayrıca geliştirilen bir program (crawler) üzerinden toplanmıştır.

Elde edilen veri seti üzerinde iki ayrı çalışma yürütülmüştür. Birinci çalışmada, geliştirilen metot özellik seçim yöntemi uygulamadan ve Ki-Kare ve Bilgi Kazancı özellik seçim yöntemlerini uygulayarak, Naif Bayes ve Destek Vektör Makinelerinin sınıflandırma performans ve işlem süreleri analiz edilmiştir. Makalelerin hangi kanser türüne ait olduğunu bulmaya çalışılmış ve oldukça yüksek bir başarım elde edilmiştir. İkinci çalışmada ise, elde edilen metinlerdeki kilit anahtar kelimeler çıkartılarak, veri seti, analiz edilmesi daha zor bir hale dönüştürülmüştür. Bu ikinci veri seti üzerinde aynı metot tekrar test edilmiştir. Çalışma sonunda, çıkartılan anahtar kelimelerin sınıflandırma başarımında kilit rol oynadığı gözlemlenmiştir. Her iki durumda da, önerilen metot makul bir sınıflandırma başarımı göstermiştir.

Anahtar Kelimeler— Metin Sınıflandırma, Metin/Doküman İşleme, Tıbbı Metin Sınıflandırma, Kanser Sınıflandırma, Naif Bayes, Destek Vektör Makineleri

Automatic Classification of the Medical Documents on the Medline Database into Relevant Cancer Types

Abstract— Pubmed, which is a search engine that is frequently used by medical researchers, is a tool to perform queries over the MEDLINE database. MEDLINE is a bibliographic database updated regularly to cover recent studies in the fields of medical, biology and genetics. Since, it includes large volume of unstructured data, i.e., texts, several text classification studies have been conducted over the MEDLINE database or some of its parts. In this study, a method has been developed that examines abstracts of articles written on several types of cancers and automatically detects the type of cancer mentioned in the text. In order to train and test the proposed method, 25962 article abstracts have been collected over the MEDLINE database by the help of a software (crawler) that is specifically developed in the scope of this study to query Pubmed search engine.

Two different studies have been applied to the obtained data set. In the first study, classification performance and processing time of Naïve Bayes and Support Vector Machines are analyzed on the data without any preprocessing and with Chi-Square and Information Gain feature selection. It is tried to find out what type of cancer types are explained in the articles, and obtained quite high success rates.

In the second study, some of the key words are removed from the text so that classifying them became harder than the first case. Same methods are trained and tested over this second version of the dataset. As a result, it is observed that the removed key words play an important role in classifying the texts. In both cases, the proposed methodology has shown reasonable performance classification.

Keywords— Text Classification, Text/Document Processing, Medical Text Classification, Cancer Classification, Naive Bayes, Support Vector Machines

1. GİRİŞ (INTRODUCTION)

Bilgi teknolojilerindeki gelişmeler ile birlikte bilgisayar ortamında üretilen belge sayısında da her geçen yıl artan bir ivme göstermektedir. Merrill Lynch, bütün verilerin yaklaşık %85'inin yapısal olmayan veri halinde olduğunu ifade etmiştir [1]. Yapısal olmayan veriler sınırlandırılmamış detaylı bilgi tutabilmesi özelliği ile daha önemli bilgileri barındırabilir. Bu büyük yapıdaki verilerden nitelikli bilgileri bulup çıkarmak olarak adlandırılan metin madenciliğinin, bu verilerin analizindeki önemi her geçen gün artmaktadır[2].Metin sınıflandırmanın temel uygulamaları olarak istenmeyen e-posta (spam) filtrelenmesi[3,4], metinden yazarı veya metin dilini tanıma[5], haberlerin konu/kategori ataması yapılması [6-8]gibi birçok uygulama gösterilebilir.Bu çalışmalar gösteriyor ki, çok sayıdaki detaylı teknik dokümanları, hızlı bir şekilde organize ve analiz etme ihtiyacı metin sınıflandırmanın önemini ortaya koymaktadır.

Medikal metin sınıflandırma çalışmaları incelendiğinde derlem oluşturma yönüyle iki farklı çalışma yapıldığı görülür. Bunlardan birincisi Google ve Pubmed gibi arama motorları yardımıyla bilgi toplamayarak derlem oluşturma, ikincisi ise klinik veri tabanlarından elde edilen veriyi kullanarak derlem oluşturma şeklindedir.MEDLINE Amerikan Ulusal Tıp Kütüphanesi tarafından 1960'lardan bu yana geliştirilen tıp ile ilgili makalelerden oluşan veri tabanıdır. Bu veri tabanında yaklaşık 5600 tıbbi dergiden yaklaşık 21 milyonun üzerinde makale barındırmaktadır. Tıp araştırmacıların yoğun olarak kullandığı Pubmed arama motoru MEDLINE veri tabanından üzerinde sorgulama yapmak için kullanılmaktadır. MEDLINE veri tabanında veya belli bölümlerinde birçok metin sınıflandırma çalışmaları mevcuttur[9-21].

MEDLINE medikal, biyoloji ve genetik alanındaki çalışmaları içeren ve sürekli güncel tutulan bibliyografik bir veri tabanıdır. İçerdiği yüksek hacimdeki yapısal olmayan metinler sebebiyle, MEDLINE veri tabanını veya belli bölümleri üzerinde pek çok metin sınıflandırma çalışmaları mevcuttur. Bu çalışmada kanser türleri hakkında yazılmış makale özetlerini inceleyerek makalenin hangi kanser türüyle ilgili olduğunu otomatik bulan bir metot geliştirilmiştir. Metodu eğitmek ve test etmek için MEDLINE veri tabanında bulunan 25962 makale özeti, Pubmed arama motoru üzerinden otomatik sorgulama yapmak için ayrıca geliştirilen bir program (crawler) üzerinden toplanmıştır. Metodun başarımını daha ileri seviyede test etmek için elde edilen veri setinden ikinci bir set daha üretilmiştir. Bu üretilen kümede, ilk kümede bulunan önemli bazı anahtar kelimeler silinmiş ve metnin ayrıştırılması daha zor hale getirilmiştir. Çalışmamızın genel amacı, öncelikle metin madenciliği yöntemleriyle kanser üzerine yazılmış makaleleri birbirinden otomatik sınıflandırma amaçlı geliştirilmiş metodun başarımını analiz etmektir. İkinci olarak, metnin içinde bulunan bazı anahtar kelimeler ile sınıflandırma başarımı arasındaki ilişkiyi ortaya çıkarmaktır. Dolayısıyla, çalışmamızın metin madenciliğindeki salt kelime sıklığı tabanlı yaklaşımın genel başarımı ile ilgili araştırmacılara ciddi bir fikir vermesi amaçlanmaktadır.

2. MATERYAL (MATERIAL)

Deneysel çalışmayı yapmak için gerekli veri seti, Pubmed arama motoru üzerinden otomatik arama yapan bir "internet crawler" program yardımıyla yapılmıştır. Geliştirilen bu program, istenen kanser türüne ait dokümanları Pubmed aracılığıyla arayıp, gelen dokümanları veri tabanına kaydetmektedir. Bu program aracılığıyla elde edilen derlemde toplam 13 kanser türü başlığında toplam 25962 makale özeti bulunmaktadır. Oluşturulan derlemin kanser türleri arasındaki dağılımı Tablo 1'de gösterilmektedir. Makalenin özellikle özet kısmının seçilmesi, tüm makalenin özünü vermesi, bilginin yoğun ve kısa olarak elde edilebilmesi özellikleri göz önüne alınarak tercih edilmiştir. Ayrıca çoğu okuyucunun, her dergiye aboneliği bulunmadığı için sadece makale özetlerine erişebilmesi de özetleri kullanarak metin sınıflandırma yapma konusundaki fikri desteklemektedir. Tüm kullanılan özetler internette her kullanıcıya açık şekilde paylaşılan bilgilerdir. Derlem oluşturulurken her bir makale özeti metin dosyasına kaydedilerek ilgili kanser türü dizinine kaydedilmiştir. Tablo 1 kullanılan veri setindeki sınıfların dağılımı hakkında özet bilgi vermektedir.

Tablo 1. Oluşturulan veri seti özellikleri (The generated dataset properties)

Sıra	Kanser Türü	Makale Sayısı
1	Bladder Cancer	1445
2	Breast Cancer	4830
3	Cervical Cancer	1316
4	Colorectal Cancer	1379
5	Endometrial Cancer	1066
6	Gastric Cancer	1134
7	Leukemia Cancer	4428
8	Lung Cancer	2855
9	Ovarian Cancer	1479
10	Pancreatic Cancer	1042
11	Prostate Cancer	3141
12	Renal Cancer	1149
13	Thyroid Cancer	698
	Toplam	**25962**

3. METİN ÖNİŞLEMEYÖNTEMLERİ(TEXT PREPROCESSING METHODS)

Metin dokümanları yapısal olmayan veri olduğu için doğrudan veri madenciliği teknikleri kullanılamaz. Yapısal olmayan bu veriyi yapısal hale çevirmek için önişleme kullanılır. Bunun için sık kullanılan yöntemler, işaretlere ayırma, kök bulma, durak kelime filtreleme gibi adımlardır. İşaretlere ayırma yöntemi metin içindeki terimleri hece, kelime, cümleler gibi bölümlere ayırma işlemidir. En sık kullanılan yöntem, bizim de bu çalışmada kullandığımız kelimelere ayırma yöntemidir. **Ek almış kelimelerin basit temel hallerine dönüştürerek, metin sınıflandırmada problem olan yüksek boyut sorun çözümüne de katkı sağlar. Çalışmamızda snowball kök bulma algoritması kullanılmıştır**[22]. Özellik azaltma işlemlerinden yaygın olarak kullanılan adımlardan biride edat, bağlaç ve zamir gibi çok sık kullanılan ve tek başına anlam ifade etmeyen kelimelerin çıkarımı işlemidir. Çalışmada, (http://www.unine.ch/Info/clef/) adresinde bulunan stop-word-list, etkisiz kelime listesi veri setinden çıkarılmıştır.

4. ÖZELLİK SEÇME(FEATURE SELECTION)

Metin madenciliğinde en önemli sorun çok sayıda özelliğin olmasıdır. Bu önemsiz özelliklerin ayrılması, önemli özelliklerinde belirlenmesi işlemidir. Bu aşamada metinler yapılandırılmış formata dönüştürülür. Bu amaç için bu çalışmada ki-kare ve bilgi kazancı yöntemleri kullanılmıştır.

4.1. Kİ-KARE (CHI SQUARE)

Değişik uygulama alanları mevcuttur. İki değişken arasında bir bağımlılığı ortaya çıkarmak için kullanılır. Eğer özellik sınıftan bağımsız olursa sıfır değerini alır, yüksek değerler aldığında ise daha tanımlayıcıdır. İki bağımsız olayı incelemek için kullanılır. Eşitlik 1'de verildiği üzere X ve Y'nin olasılıkları çarpımı XY nin çarpımının olasılığına eşitse, X ve Y'nin birbirinden bağımsız olduğu varsayılır [23].

$$p(XY) = p(X)p(Y) \quad (1)$$

Chi2 hesaplamak için eşitlik 2 kullanılır.

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

Eşitlik 2'de kullanılan N gözlenen frekansı, C sınıfı, t süreyi, E beklenen frekansı temsil edilir.

4. 2. BİLGİ KAZANCI(INFORMATION GAIN)

IG değişken etkinlik ölçüsü olarak tanımlanan istatistiki bir değer olarak hesaplanır [24]. Bilgi kazancı hesaplanırken birinci olarak alt bölümlere bölünmeden ki halinin entropisi bulunur, daha sonra tüm alt bölümlerin entropisi bulunarak iki değer arasındaki farkın büyük olduğu değişken en iyi kriter olarak seçilir. Bu teknik terim azaltma işlemlerinde yaygın olarak kullanılır [25].

Sınıf tahmini için olası kümeler $\{c_1, c_2, \ldots, c_j\}$ olmak üzere her terim (t) için eşitlik 3 ile hesaplanır.

$$IG(t,c_i) = \sum_{c\in\{c_i,\bar{c}_i\}} \sum_{t'\in\{t,\bar{t}\}} P(t'|c).\log\frac{P(t'|c)}{P(t').P(c)} \quad (3)$$

$P(t, c_i)$=t'nin c_i'ye üyelik olasılığı,
$P(t, \bar{c}_i)$=t'nin c_i'ye üye olmama olasılığı,
$P(\bar{t}, c_i)$=t'nin değilinin c_i 'ye olma olasılığı,
$P(\bar{t}, \bar{c}_i)$= t'nin değilinin c_i 'ye olmama olasılığı olarak ifade edilir.

Eşik değer altında kalan terimler belirlenerek özellik uzayından elenir, eşik değer üstünde kalan nitelikli terimler ise özellik uzayı içinde kalır. **Çalışmada eşik değeri parametresi için ön tanımlı değer olan -1.797 kullanılmıştır.**

5. ÖRÜNTÜ SINIFLANDIRICILARI(THE PATTERN CLASSIFIER)

Yapılandırılmış formata dönüştürülen veriler artık veri madenciliği teknikleri ile analiz edilebilir. Metin sınıflandırmada yaygın olarak kullanılan algoritmalar Naive Bayes, Destek Vektör Makinesi, K-En Yakın Komşuluk ve C 4.5 algoritmalarıdır[26]. İlgili literatürde yaygın kullanılan ve başarımları genelde yüksek olarak raporlanan[9]algoritmalar olmalarından dolayı, bu çalışmada Naive Bayes ve Destek Vektör Makinesi algoritmaları kullanılmıştır.

5.1. NAİF BAYES (NAİVE BAYES)

Metin sınıflandırma alanında yaygın ve başarılı olmasından dolayı en çok kullanılan sınıflandırıcılardan biri olmuştur**[27, 28]**. Temeli istatistikteki Bayes teoremine dayanan,her özelliğin bağımsız olması gerektiği önermesinin basitleştirilmiş halidir.

Sınıflandırma amaçlı kullanılan bayes teoremi muhtemel çıkış durumlarından en büyük olasılıklı durum hedef sınıf olarak seçilirve eşitlik 4 ile gösterilir.

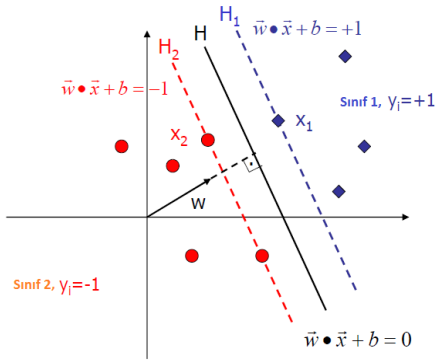
$$Y' = \arg\max_{yj \in Y} \mathrm{P}(Y = yj|\mathrm{X}) \quad (4)$$

Eşitlik 4'de kullanılan terimlerden hedef sınıfı *Y'* ifadesi, *yj*her olası sınıf için *j*.çıkış durumu, *X* tespit edilecek giriş sınıfını temsil eder.

5.2. DESTEK VEKTÖR MAKİNELERİ(SUPPORT VECTOR MACHINES)

1990'ların sonlarına doğru çoğu kişi tarafından kullanılan DVM, günümüzde de etkili ve basit kullanımı sayesinde popüler sınıflandırıcılar arasına girmiştir. Yüksek boyutta

doğrusal sınıflandırma yapabilir. Bir düzlemde bulunan iki grubu ayırmak için bu iki grup üyelerine en uzak çizilen sınırın nasıl çizileceğini belirler. **Sınıfları ayırmak için en büyük uzaklığı olan doğrusal fonksiyonu arar**. Doğrusal olarak ayrılamıyorsa daha yüksek boyutlu üst uzaya taşıyarak sınıflandırma yapar [29].



Şekil 1. İki sınıflı bir veri seti örneği

DVM ile sınıflandırma yapılacak iki grup arasında sınır çizilerek ayırmak mümkündür. Ayırım işlemi yapılması için iki gruba da yakın ve birbirine paralel iki sınır çizgisi çizilir ve sınır çizgileri birbirine yaklaştırılarak ortak sınır çizgisi oluşturur. Sınıf çizgileri arasında oluşan aralık tolerans olarak adlandırılır.

Düzlemde her bir nokta eşitlik 5'de gibi tanımlanabilir.

$$D=\{(x_i,c_i)|x_i\in\mathbb{R}^p,\ c_i\in\{-1,1\}\}_{i=1}^{n} \quad (5)$$

Noktaların hiperdüzlem üzerinde olduğu varsayılırsa her bir nokta eşitlik 6'deki gibi tanımlanabilir.

$$w * x - b = 0 \quad (6)$$

w=Hiperdüzleme dik olan normal vektörü

b=kayma oranı

İki farklı hiperdüzlem olasılığı bulunmasına karşılık DVM yönteminde bu olasılıklardan en büyük toleransa sahip olan alınır.

Sınıflandırma aşamasında Destek Vektör Makinası çekirdek fonksiyonu olarak polinom çekirdek fonksiyonu kullanılmıştır.

6. PERFORMANS DEĞERLENDİRME KRİTERLERİ(PERFORMANCE EVALUATION CRITERIA)

Genel olarak metin sınıflandırma başarımlarını değerlendirmek için veri sınıflandırmada da kullanılan duyarlılık, anma ve F-ölçütü kullanılmaktadır. Bu değerlendirmelerde geçen kısaltma anlamları şunlardır; TP(True Positive) doğru sınıflandırılmış pozitif örnek sayısı, TN(True Negative) doğru sınıflandırılmış negatif örnek sayısı, FP(False Positive) yanlış sınıflandırılmış pozitif örnek sayısı, FN(False Negative) yanlış sınıflandırılmış negatif örnek sayılarını ifade eder.

$$\text{Kesinlik} = \frac{TP}{TP + FP} \quad (7)$$

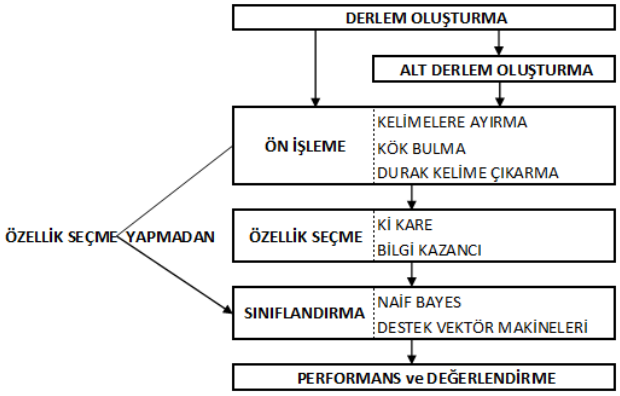
$$\text{Anma} = \frac{TP}{TP + FN} \quad (8)$$

$$F - \text{Ölçütü}, F = \frac{2 \text{ x Kesinlik x Anma}}{\text{Kesinlik} + \text{ Anma}} \quad (9)$$

Kesinlik(Eşitlik 7) ve anma (Eşitlik 8) ölçütleri, beraber kullanıldığında her iki ölçütün etkisini görebileceğimiz ve daha doğru sonuçları verebilecek F-ölçütü kullanılır. F-ölçütü Eşitlik 9'te görüldüğü gibi kesinlik ve anma değerlerinin harmonik ortalaması olarak hesaplanır.

7. DENEYSEL ÇALIŞMA(EXPERIMENTAL STUDY)

Derlem oluşturulduktan sonra WEKA veri madenciliği çatısı kullanılarak deneyler modellenmiştir.



Şekil 2. Önerilen çalışmanın iş akışı

Önerilen çalışmanın iş akışı Şekil 2'de özetlenmiştir. Önerilen çalışmada, özellik seçimi yapılmadan ve özellik seçim yapıldıktan sonra sınıflandırıcı algoritmalar ile sonuçlar bulunmuştur. Özellik seçimi için sırasıyla Ki-kare ve Bilgi kazanımı algoritmaları, sınıflandırma içinse destek makinesi (DVM) ve Naive Bayes (NB) sınıflandırma algoritmaları 10 katlı çapraz geçerleme ile kullanılmıştır. Algoritma performanslarına ait tüm sonuçlar F-ölçüt başarım endeksine göre ölçülmüştür. Bahsedilen şekilde yapılan deneylerdeki başarım sonuçları Tablo 2'de verilmiştir.

Tablo 2. Orijinal veri seti F-ölçüt sonuçları
(Results for original datasets of F-measure)

Özellik Çıkarım	NB	DVM
Ki-kare	0.961	0.968
Bilgi Kazanımı	0.954	0.967
Özellik Seçme Olmadan	0.780	0.966

Tablo 2 incelendiğinde DVM'nin NB sınıflandırıcısından daha başarılı olduğu gözlenmektedir. Özellik seçimi olmadan sınıflandırma yapıldığında NB sınıflandırıcının başarımının oldukça düştüğü fakat DVM sınıflandırıcısının sonuçlarının çok düşmediği görülmektedir. Dolayısıyla DVM algoritmasının NB'e göre özellik seçiminden daha az etkilendiği söylenebilmektedir. Özellik seçim algoritmalarını kıyaslamak istediğimizde ise,Ki-kare algoritmasının nispeten daha başarılı olduğu gözlenmektedir. Genel olarak baktığımızda, özellik seçimi ve ilgili sınıflandırıcılar beraber kullanıldığında, önerilen yöntem F-ölçüt kıstasına göre ortalamada %95 üzeri,yani oldukça iyi bir başarım göstermektedir.

Tablo 3. Anahtar kelimeleri çıkarıldığı veri seti F-ölçüt sonuçları (Results for removed keywords datasets of F-measure)

Özellik Çıkarım	NB	DVM
Ki-kare	0.753	0.795
Bilgi Kazanımı	0.723	0.785
Özellik Seçme Olmadan	0.605	0.803

Çalışmamızda metin sınıflandırma başarımının doğru anahtar kelimelerin özellik seçici algoritmalar tarafından seçilmesine oldukça bağımlı olduğunu gösterme amaçlı ikinci bir deney çalışması daha yapılmıştır. Bu ikinci deneyin amacı, kelime sıklığı tabanlı metin sınıflandırma yaklaşımında doğru anahtar kelimeler metnin içinde bulunmadığı durumlarda da sınıflandırıcının yine belli bir düzeyde başarım elde edebileceğini göstermektir. Fakat bu anahtar kelimeler sınıflandırıcı için kritik bir öneme sahiptir, bu yüzden bu kelimeler çıkartıldığında **Tablo 3' de görüldüğü gibi sınıflandırıcı performans başarımı düşmektedir**.

Tablo 4. Veri setinden çıkartılan anahtar kelimeler (Keywords extracted from the data set)

Sıra	Kanser Türü	Çıkartılan Anahtar Kelimeler
1	Mesane Kanseri	bladder
2	Meme Kanseri	breast
3	Rahim Ağzı Kanseri	cervical
4	Kolon Kanseri	colorectal
5	Rahim Kanseri	endometrial
6	Mide Kanseri	gastric
7	Lösemi Kanseri	leukemia
8	Akciğer Kanseri	lung
9	Yumurtalık Kanseri	ovarian
10	Pankreas Kanseri	pancreatic
11	Prostat Kanseri	prostate
12	Böbrek Kanseri	renal cell
13	Tiroid Kanseri	thyroid

Anahtar kelimelerin çıkarıldığı veri setimizde Tablo 4'de genel olarak başarının düştüğü gözlenmektedir. Tablo 2'den farklı olarak özellik seçme yapılmadan sınıflandırılan anahtar kelimeleri çıkartılan veri setinde DVM algoritmasında daha başarılı çıkmıştır.

Tablo 5. Orijinal veri setinin sınıflandırma işlem süresi (sn)(The classification of the original data set the duration of the transaction)

Özellik Çıkarım	NB	DVM
Ki-kare	14	65
Bilgi Kazanımı	15	75
Özellik Seçme Olmadan	522	1415

İşlem süreleri olarak NB algoritmasının daha hızlı olduğu Tablo 5'de görülmektedir. Özellik seçme yapılmadığı durumlarda işlem süresinin uzadığı izlenmektedir. Sınıflandırıcılarda NB algoritmasının, özellik çıkarımda ise ki-kare algoritmasının daha hızlı olduğu gözlenmektedir.

Tablo 6. Anahtar kelimeleri çıkartılan veri setinin sınıflandırma işlem süreleri (sn) (The classification of keywords extracted from the data set the duration of the process)

Özellik Çıkarım	NB	DVM
Ki-kare	17	93
Bilgi Kazanımı	19	108
Özellik Seçme Olmadan	523	2481

Tablo 6'de anahtar kelimelerin çıkarılması ile işlem süresinin arttığı gözlenmiştir. Tablo 5'te benzer sonuçlar çıkmıştır. Bunun temel sebebi, sınıflandırıcı ayıracağı veri yeterince ayrık olmadığında, daha fazla adım kullanarak sınıflandırma yapmak zorunda kalmasıdır. Böylece, daha fazla adım çalışmakta ve bu da işlem süresini uzatmaktadır.

8. SONUÇLAR(CONCLUSIONS)

Bu çalışmada tıbbı makale özetlerinin anahtar kelime çıkarılarak ve orijinal haliyle kanser türlerine göre sınıflandırma başarımları ve işlem süreleri analiz edilmiştir. Analiz yapılırken NB ve DVM sınıflandırıcıları kullanılmış DVM sınıflandırıcının daha başarılı, ancak NB sınıflandırıcısının ise daha hızlı olduğu gözlenmiştir. Orijinal ve anahtar kelimeleri çıkartılan veri setlerinin, sınıflandırma başarımları karşılaştırıldığında orijinal veri setinin sınıflandırma başarımları yüksek ve hızlı olduğu görülmektedir. Özellik çıkarım işleminde ise orijinal veri setinde ki-kare algoritmasının daha başarılı ve hızlı olduğu görülmektedir. Anahtar kelimelerin çıkarıldığı durumlarda işlem süresinin arttığı ve başarımın düştüğü ancak özellik seçme işlemi yapılmadan DVM algoritması ile sınıflandırıldığında bu veri setinde yüksek başarı sağlandığı gözlenmektedir. Çalışmanın devamında veri setinin geliştirilerek daha farklı kanser türleri eklenmesi ve sınıflandırma başarımının değerlendirilmesi planlanmaktadır. Ayrıca, ilgili literatürde Türkçe dili üzerinde bu tarz bir çalışma pek bulunmadığı için, benzer

bir çalışmanın Türkçe dili için yapılması da literatüre önemli bir katkı yapacağı düşünülmektedir.

KAYNAKLAR (REFERENCES)

[1] R. Blumberg ve S. Atre, "The problem with unstructured data", *DM Rev.*, c. 13, s. 62, sayı 42–49, 2003.

[2] A.-H. Tan ve P. S. Yu, "Guest Editorial: Text and Web Mining", *Appl. Intell.*, c. 18, s. 3, ss. 239–241, May. 2003.

[3] I. Idris, A. Selamat, N. T. Nguyen, S. Omatu, O. Krejcar, K. Kuca, ve M. Penhaker, "A combined negative selection algorithm–particle swarm optimization for an email spam detection system", *Eng. Appl. Artif. Intell.*, c. 39, ss. 33–44, 2015.

[4] A. Çıltık ve T. Güngör, "Time-efficient spam e-mail filtering using n-gram models", *Pattern Recognit. Lett.*, c. 29, s. 1, ss. 19–33, 2008.

[5] C. Zhang, X. Wu, Z. Niu, ve W. Ding, "Authorship identification from unstructured texts", *Knowl.-Based Syst.*, c. 66, ss. 99–111, 2014.

[6] H. K. Yıldız, M. Gençtav, N. Usta, B. Diri, ve M. F. Amasyalı, "A new feature extraction method for text classification", içinde **Signal Processing and Communications Applications**, 2007. SIU 2007. IEEE 15th, 2007, ss. 1–4.

[7] Karadağ, A., Takçı, H., "Metin Madenciliği ile Benzer Haber Tespiti", **Akademik Bilişim**, 2010.

[8] M. F. Amasyali ve T. Yildirim, "Automatic text categorization of news articles", içinde **Signal Processing and Communications Applications Conference**, 2004. Proceedings of the IEEE 12th, 2004, ss. 224–226.

[9] A. Haltaş, A. Alkan, ve M. Karabulut, "Metin Sınıflandırmada Sezgisel Arama Algoritmalarının Performans Analizi", *Gazi Üniversitesi Mühendis.-Mimar. Fakültesi Derg.*, c. 30, sayı 3, ss. 417–427, 2015.

[10] M. Karabulut, "Fuzzy unordered rule induction algorithm in text categorization on top of geometric particle swarm optimization term selection", *Knowl.-Based Syst.*, c. 54, ss. 288–297, Ara. 2013.

[11] H. Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm", *Knowl.-Based Syst.*, c. 24, sayı 7, ss. 1024–1032, 2011.

[12] M. Yetisgen-Yildiz ve W. Pratt, "The effect of feature representation on MEDLINE document classification", içinde **AMIA annual symposium proceedings**,Washington, s. 849, 2005.

[13] A. K. Uysal ve S. Gunal, "Text classification using genetic algorithm oriented latent semantic features", *Expert Syst. Appl.*, c. 41, s. 13, ss. 5938–5947, 2014.

[14] B. Parlak ve A. K. Uysal, "Classification of medical documents according to diseases", içinde *Signal Processing and Communications Applications Conference (SIU)*, 2015 23th, ss. 1635–1638, 2015.

[15] R. B. Dollah ve M. Aono, "Ontology based approach for classifying biomedical text abstracts", *Int. J. Data Eng. IJDE*, c. 2, sayı 1, ss. 1–15, 2011.

[16] O. Frunza, D. Inkpen, S. Matwin, W. Klement, ve P. O'blenis, "Exploiting the systematic review protocol for classification of medical abstracts", *Artif. Intell. Med.*, c. 51, sayı 1, ss. 17–25, 2011.

[17] K. Yi ve J. Beheshti, "A hidden Markov model-based text classification of medical documents", *J. Inf. Sci.*, c. 35, sayı 1, ss. 67–81, Oca. 2009.

[18] G. L. Poulter, D. L. Rubin, R. B. Altman, ve C. Seoighe, "MScanner: a classifier for retrieving Medline citations", *BMC Bioinformatics*, c. 9, sayı 1, s. 108, 2008.

[19] F. Camous, S. Blott, ve A. F. Smeaton, "**Ontology-based MEDLINE document classification**", Bioinformatics Research and Development, Springer, ss. 439–452, 2007.

[20] S. Spat, B. Cadonna, I. Rakovac, C. Gutl, H. Leitner, G. Stark, ve P. Beck, "*Multi-label text classification of German language medical documents*",**Proceedings of the 12th World Congress on Health (Medical) Informatics**, 2007.

[21] R. Rak, L. A. Kurgan, ve M. Reformat, "Multilabel associative classification categorization of MEDLINE articles into MeSH keywords", *IEEE Eng. Med. Biol. Mag.*, c. 26, s. 2, s. 47, 2007.

[22] "weka - Stemmers". [Çevrimiçi]. Available at: **http://weka.wikispaces.com/Stemmers**. [Erişim: 22-Kas-2015].

[23] M. Güngör ve Y. Bulut, "Ki-Kare Testi Üzerine", *Doğu Anadolu Bölgesi Araştırmaları*, c. 7, s. 1, ss. 84–89, 2008.

[24] V. V. KÖK ve N. KULOĞLU, "Sollama Esnasında Taşıt Ve Yol İle İlgili Faktörlerin Karar Ağacı Yöntemiyle İrdelenmesi", *Erciyes Üniversitesi Fen Bilim. Enstitüsü Derg.,* 21(1–2), ss. 180–188, 2005.

[25] Cover, Thomas M., and Joy A. Thomas, "Elements of information theory", John Wiley & Sons, Canada, 2012.

[26] F. Sebastiani, "Machine Learning in Automated Text Categorization", *ACM Comput Surv,* c. 34, s. 1, ss. 1–47, Mar. 2002.

[27] E. Alpaydin, Introduction to machine learning, 2nd ed. Cambridge, Mass: MIT Press, 2010.

[28] C. C. Aggarwal ve C. Zhai, "A Survey of Text Classification Algorithms", içinde Mining Text Data, C. C. Aggarwal ve C. Zhai, Ed. Springer US, ss. 163–222, 2012.

[29] Vapnik, V.N., "The Nature of Statistical Learning Theory", Springer-Verlag, New York, 1995.