FATİH Projesine Yönelik Görüşlerin Metin Madenciliği Yöntemleri İle Otomatik Değerlendirilmesi

Hanife GÖKER[1], Hakan TEKEDERE[2]

[1]Bilişim Enstitüsü, Gazi Üniversitesi, Ankara, Türkiye
[2]Sağlık Hizmetleri MYO, Gazi Üniversitesi, Ankara, Türkiye
hanifegoker@gmail.com, tekedere@gazi.edu.tr
(Geliş/Received:02.02.2017; Kabul/Accepted:01.06.2017)
DOI: 10.17671/gazibtd.331041

Özet— Bu çalışmada, FATİH projesine yönelik internet ortamında yer alan görüşlerin metin madenciliği yöntemleri ile otomatik tespitinin yapılması amaçlanmaktadır. Çalışma iki temel kısımdan meydana gelmektedir. İlk basamakta, internet ortamındaki yapısal olmayan veri kümelerinin yapısal veri haline dönüştürülmesini sağlamak amacıyla metin madenciliği veri önişleme yazılımı geliştirilmiştir. İkinci basamakta ise geliştirilen metin madenciliği veri önişleme yazılımı ile yapısal veri kümesine dönüştürülen veriler üzerinde makine öğrenmesi algoritmaları uygulanarak yorumlar otomatik sınıflandırılmaktadır. Geliştirilen metin madenciliği veri önişleme yazılımının en önemli ayırt edici özelliği, yazılımın sadece FATİH projesine yönelik görüşlerinin veri önişleme basamağında değil, istenilen amaca yönelik metin sınıflandırma işleminin veri önişleme basamağında konudan bağımsız bir şekilde kullanılabilir olmasıdır. Çalışmada FATİH projesine yönelik 444 görüş içeren metin dosyasındaki metinler tf-idf ağırlıklandırma yöntemi ile vektörel olarak temsil edilerek sınıflandırma algoritmalarının model başarım ölçütleri karşılaştırılmıştır. Performansı karşılaştırılan algoritmalardan en yüksek başarımın Ardışık Minimal Optimizasyon Algoritmasına ait olduğu (%88,73) gözlemlenmiştir.

Anahtar Kelimeler— Metin madenciliği, FATİH projesi, metin sınıflandırma, fikir madenciliği, ardışık minimal optimizasyon algoritması.

Automatic Evaluation of Opinions Concerning FATİH Project with Text Mining Methods

Abstract— In this study, it is aimed to make automatic determination of views towards the FATIH project in internet environment by using text mining methods. The study is based on two main parts. In the first step, text mining data preprocessing software was developed to convert non-structural data sets on the internet into structured data. In the second step, interpretation is automatically classified by applying machine learning algorithms on the data converted into the structural data set with the developed text mining data preprocessing software. The most important distinguishing feature of the developed text mining data preprogramming software is that its views at the data preprocessing step are not only available for the FATIH project but it is available at the data preprocessing step of all the desired text classification purposes. In the study, the texts containing 444 visions for the FATIH project were represented as vectors by the tf-idf weighting method and the model performance criteria of the classification algorithms were compared. Highest achievement for performance comparison algorithms is detected that Sequential Minimal Optimization Algorithm (88.73%).

Keywords— Text mining, FATİH project, text classification, opinion mining, sequential minimal optimization algorithm.

1. GİRİŞ (INTRODUCTION)

Son yıllarda dijital ortamda kayıt altına alınan bilgi miktarı her geçen gün artmaktadır. Dijital ortamdaki metin dosyaları, forum verileri ve eposta içerikleri gibi yapısal olmayan bu verilerin analiz edilmesi kolay değildir. Bu verilerin analizinde klasik sorgu metotları yetersiz kalmakta, bu nedenle metin yığınları içinden değerli bilginin çıkarılması için metinlerin sınıflandırılması gerekmektedir. İnternet ortamında bulunan forum verilerinin analiz edilmesi, özellikle incelenecek veri sayısı büyük ise, çok fazla emek, maliyet ve zaman isteyen bir süreçtir. Bu bağlamda forum verilerinin metin madenciliği ile otomatik tespitinin yapılması, istatistiksel olarak değerlendirilmesinde önemlidir. Günümüzde yapılan veri madenciliği ve makine öğrenmesi çalışmaları genellikle veri tabanı veya veri ambarlarındaki veriler gibi yapısal veriler üzerinde odaklanmaktadır. Fakat günlük hayatta karşılaşılan; metin dosyaları, web sayfalarında yer alan forum verileri, e-posta içerikleri, makaleler, bloglar ve açık uçlu anket cevapları gibi verilere bakıldığında çoğunlukla verilerin yapısal olmayan metinsel veriler olduğu görülmektedir.

Yapılandırılmış veri tipi, belirli kurallar ve sistemler doğrultusunda depolandıkları için kolay erişilebilir, düzenlenebilir ve kategorize edilebilir [1]. Yapısal olmayan veriler ise standart kurallara sahip değildir ve veri analiz süreci yapısal olan verilere göre daha karmaşıktır. Metinsel verilerin yazımında standart kurallar olmadığından bilgisayar bunları anlayamamaktadır. Her bir metnin dili ve içerdiği anlam amaca bağlı olarak çeşitlilik göstermektedir. Yapısal olmayan bilgiden içerik çıkarmak için kullanılan geleneksel yöntemler içeriği açıklayıcı sonuçlar elde edemez [2]. Bu nedenle metin verisindeki anlamın ortaya çıkarılabilmesi yani otomatik veri analizi için kullanılan yöntem metin madenciliğidir.

Metin madenciliği; önceden bilinmeyen ve önemli olan bilgilerin keşfedilmesi amacıyla çok sayıda dokümanı analiz eden bir teknolojidir [3]. Metin madenciliği yöntemleri arasında metin sınıflandırma yaygın olarak kullanılmaktadır. Metin sınıflandırma bir dokümanın özelliklerine bakılarak önceden belirlenmiş belli sayıdaki kategorilerden hangisine dahil olacağını belirleme işlemidir [4].

Metin sınıflandırma için belge sınıflandırma, metin kategorilerinin belirlenmesi gibi farklı isimler de kullanılmaktadır. Metin sınıflandırma uygulamalarında Naive Bayes, Karar Ağaçları, Yapay Sinir Ağları, Destek Vektör, Örnek Tabanlı ve İstatistiksel Dil Modeli Tabanlı Sınıflandırıcılar yaygın olarak tercih edilmektedir [5].

Literatürde birçok farklı problem alanları için metin madenciliği yöntemleri kullanılmıştır. Bu çalışmalara örnek olarak; metin içerikli Türkçe dokümanların sınıflandırılması [6], müşteri memnuniyetlerinin belirlenmesinde metin madenciliği tekniklerinin kullanılması [7], akademik belgelerin otomatik olarak sınıflandırılması [8], web ortamındaki verilerden siyasi görüşlerin modellenmesi ve tahmin edilmesi [9], Türkçe dokümanlarda yapay sinir ağları ile yazar tanıma [10] ve sosyal medyada tüketicilerin markalara bakış açılarının tespiti [11] gösterilebilir.

Bu çalışma ile;

a) FATİH projesine yönelik internet ortamında yer alan görüşlerin metin madenciliği yöntemleri ile otomatik tespitinin yapılması,
b) Metin madenciliği çalışmalarının veri önişleme basamağında kullanılmak üzere uyarlanabilir "metin madenciliği veri önişleme" yazılımının geliştirilmesi,
c) Konu ile ilgili görüşlerin bulunduğu bu veri seti üzerinde sınıflandırma algoritmalarının performanslarının karşılaştırılması amaçlanmaktadır.

2. METİN MADENCİLİĞİ (TEXT MINING)

Metin madenciliği, özel amaçlar için metinden değerli bilgileri çıkarmak adına, metnin analiz edilmesi işlemidir [12]. Bilgisayarın metinsel bilgileri anlayabileceği seviyeye getirmek için metin madenciliği teknikleri kullanılır [13].

Metin madenciliği yöntemlerinin temelinde matematiksel ve istatistiksel yöntemler bulunmaktadır. Metin madenciliği, yazar tanıma, metin sınıflama, fikir madenciliği, duygu analizi, anahtar kelime çıkartımı, başlık çıkartımı gibi farklı alanlarda da kullanılmaktadır [14]. Metin madenciliği teknikleri sınıflandırma, birliktelik analizi, bilgi çıkarım ve kümeleme olmak üzere dört temel kategoriye ayrılır [15]. Bu çalışma kapsamında eldeki metinlerin önceden belirlenen sınıflardan hangisine ait olduğunun öngörülmesi işlemi yapıldığından metin sınıflandırma tekniği kullanılmıştır.

Sınıflandırma, nesnelerin daha önceden bilinen sınıflara atanması işlemidir. Metin sınıflandırma iki alt basamağa indirgenebilir [16]:

1. Metinlerin nasıl temsil edileceği
2. Sınıflandırma işleminin yapılması

Metinlerin temsil edilmesi kısmında yapısal olmayan metinsel veriler yapısal bir hale dönüştürülür. Böylelikle metinsel veriler veri madenciliği tekniklerinin uygulanabileceği formata dönüştürülmüş olur [17]. Sınıflandırma işleminde daha önceden belirlenmiş sınıflara veriyi yerleştirmek için kullanılacak fonksiyonun öğrenilmesi sağlanır. Çıktılar önceden bilindiği için sınıflama veri kümesini denetimli olarak öğrenir. Öğrenmenin amacı bir sınıflandırma modelinin oluşturulmasıdır. Tüm veriler kullanılarak eğitme işi yapılmaz [18].

Sınıflandırma algoritmaları kullanılırken veriler; eğitim kümesi ve test kümesi olmak üzere ikiye bölünür. Algoritma eğitim kümesi ile eğitilirken, test kümesi ile

kontrol edilir [19]. Sınıflandırıcının başarımının doğru biçimde ölçülebilmesi için hem eğitim kümesinin hem de test kümesinin yeterli temsil yeteneğine sahip olması yani tipik metinleri içermesi gerekir. Bunu sağlamak üzere k katlı çapraz doğrulama (k-fold cross-validation) yöntemi yaygın olarak kullanılmaktadır. Bu yöntemde elle etiketlenmiş toplam veri kümesi, ortak eleman içermeyen k tane farkı gruba ayrılır. Yöntemin çalışma mantığında k adımda eğitim ve sınama gerçekleştirildikten sonra başarımların ortalaması alınarak son başarım değeri olarak belirlenir. Sıklıkla k=10 ya da k=5 değerleri kullanılmaktadır [5]. Metin sınıflandırma temelde beş adımdan oluşmaktadır [20]:

1. Veri kümesinin toplanması
2. Veri önişleme aşaması
   - Dönüştürme
   - Tarama ve işaretleme
   - Durak kelimelerin çıkarılması
   - Kök bulma
3. Özellik seçimi
4. Vektör uzay model seçimi
5. Sınıflandırma

2.1. Veri Kümesinin Toplanması (Collection of Dataset)

Veri toplama basamağında, kullanılacak veri ele alınan probleme uygun bir biçimde temin edilir. Veri toplama aşamasında farklı kaynaklardan yararlanılabilmektedir. Verinin toplama süreci, veriyi anlama sürecidir.

Veri toplama süreci; a) başlangıç verilerinin toplanması, b) toplanan verinin tanımlanması ve ihtiyaçları karşılayıp karşılayamayacağının değerlendirilmesi, c) çalışmanın gerçekleştirilebilmesi için veri anlamında eksikliklerinin tespit edilmesi, yani verinin keşfinin yapılması ve d) çalışmada kullanılacak olan verilerin tam mı, doğru mu, eksik var mı, hata içeriyor mu gibi verinin kalitesinin belirlenmesi şeklinde tanımlanabilir [21].

2.2. Veri Önişlem Aşaması (Data pre-processing stage)

Metin madenciliğinin en büyük sorunu işleyeceği veri kümesinin yapısal olmamasıdır. Genellikle doğal dil kullanılarak yazılmış dokümanlar üzerinde çalışılan metin madenciliği alanında ön işleme aşaması veri temizlemenin yanında veriyi uygun formata getirme işlemini de gerçekleştirmektedir [22]. Veri önişleme aşağıdaki basamaklardan oluşmaktadır:

a) Dönüştürme: İnternette dokümanlar genellikle HTML, XML gibi çeşitli tiplerde tutulduğundan bunları düz metin haline dönüştürmek gerekmektedir. Bu aşamada metinler HTML ve XML etiketlerden temizlenir [23].

b) Tarama ve işaretleme: Metin içindeki terimleri; simgelere, noktalama işaretlerine veya kelimelere ayırılması işlemidir. Belgeler bölüm, paragraf, cümle, kelime ve hecelere ayrılabilir. En sık rastlanan durum ise kelimelere ayrılmasıdır [20].

c) Durak kelimelerin çıkarılması: Metin içerisinde çok sık geçen fakat sınıflandırmada bir anlam ifade etmeyen edat, bağlaç ve zamir gibi kelimeler metinden çıkartılır [23].

d) Kök bulma: Aynı kökten gelen farklı ek almış kelimelerin doküman içerisindeki kelime sıklıklarına bakılırken aynı kelime olarak algılanması için köklerinin bulunması gerekmektedir [23].

2.3. Özellik Seçimi (Feature Selection)

Özellik seçimi; veriye ait birçok özellikten verinin kümesinin veya sınıfının değerlerini belirleyen özelliklerin belirlenmesidir [24]. Özelliklerin alt kümeleri seçilirken doğruluktan ödün vermeden seçilmelidir. Özellik seçiminde ilgisiz ve gereksiz veriler silinerek yüksek boyutu indirgemek hedeflenmektedir [25]. Özellik seçimi yalnızca arama uzayını küçültmekle kalmayıp, sınıflama işleminin kalitesini de artırmaktadır [26].

2.4. Vektör Uzay Model Seçimi (Vector Space Model)

Vektör uzay modelinde her nesne, vektör yapısında tanımlanmaktadır. Nesnelerin sahip olduğu farklı özellikler, vektör uzayının eksenlerini oluşturmakta ve her nesne sahip olduğu özelliklere göre vektör uzayında belli bir konuma sahip olmaktadır [27]. Doküman sınıflandırma çalışmalarında kullanılan özellik vektör uzayı sözcüklerin dokümanlardaki görüntülenme sıklıklarına dayanmaktadır. Yani her bir doküman içindeki sözcüklerin dokümanlardaki frekansları hesaplanıp sözcük vektör uzayı oluşturulur [6]. Bir metnin, vektör uzay modelinde temsili için üç farklı yöntem kullanılmaktadır:

a) Binary vektör: Bu yöntem ile metinsel veriler 1 ve 0'lar ile ifade edilmektedir. Veri içinde barındırdığı kelimelerin sözlükteki varlıklarına göre bu değerleri almaktadırlar [28]. Veri setindeki kelimelerin alacağı değerler binary vektör temsilinde {1,0,0,1…} şeklinde olmaktadır.

b) Frekans vektör: Binary tanımlamadan farklı olarak veri içinde bulunan kelime köklerinin kaç defa geçtiği bilgisinin de tutularak yapıldığı bir tanımlama biçimidir [28]. Veri setindeki kelimelerin alacağı değerler frekans vektör temsilinde {2,0,3,1…} şeklinde olmaktadır.

c) TF-IDF vektör: Tf-idf ağırlıklandırmasında her bir dokümandaki kelimelerin frekansı rol oynamaktadır. TF (terim frekansı) değeri frekans bilgisini yani terimin veri setinde kaç kez geçtiğini tutar. IDF ise tüm dokümanlarda seyrek görülen kelimeler ile ilgili bir ölçü verir. Eğer kelime tüm eğitim dokümanları

incelendiğinde sadece o dokümanda geçiyor ise o doküman için belirleyici özelliği vardır [29].

Aşağıda 2. 1' de TF ve IDF hesaplanması, 2. 2' de ise ağırlıklandırma hesaplanması gösterilmiştir.

$$TF_{ij} = \frac{n_{ij}}{|d_i|} \quad , \quad IDF_{ij} = \log_2\left(\frac{n}{n_j}\right) \qquad \textbf{(2.1)}$$

$$\text{Ağırlıklandırma} = TF_{ij} \text{ x } IDF_{ij} \qquad \textbf{(2.2)}$$

TF değerinin hesaplanmasında kullanılan n değeri, *j nci* kelime kökünün toplanan *i nci* veri seti içinde kaç defa geçtiği sayıdır. *d* değeri ise veri seti içinde yer alan tüm kelime köklerinin sayısıdır. Formül içinde yer alan *i* değeri ise eposta içinde yer alan kelimelerin sayısıdır. IDF değerinin hesaplanmasında kullanılan *n* değeri toplam belge sayısını *nj* ise *j*. terimin göründüğü belgelerin sayısını gösterir. Ağırlıklandırma ise bu iki değerin çarpımı ile elde edilir [28].

2.5. Sınıflandırma (Classification)

Sınıflandırma; sınıfı tanımlanmış mevcut verilerden yararlanarak sınıfı belli olmayan verilerin sınıfını tahmin etmek için kullanılır [30]. Burada sınıfların sayısı ve tanımlamaları daha önceden bilinmektedir. Sınıfı bilinmeyen yeni veri örneklerinin bu sınıflara atanması gerçekleştirilir [31].

Verilerin sınıflandırılması için belirli bir süreç izlenir. Öncelikle var olan veritabanının bir kısmı eğitim amacıyla kullanılarak sınıflandırma kurallarının oluşturulması sağlanır. Daha sonra bu kurallar yardımıyla yeni bir durum ortaya çıktığında nasıl karar verileceği belirlenir [32]. Her veri kümesinde mükemmel çalışan bir algoritma olmadığından birçok sınıflandırma algoritması geliştirilmiştir. Bu çalışmada en yüksek performansı gösteren algoritma Ardışık Minimal Optimizasyon (SMO) Algoritmasıdır.

2.5.1. Ardışık Minimal Optimizasyon Algoritması (Sequential Minimal Optimization Algorithm)

SMO, esas itibariyle destek vektör kullanan bir algoritmadır. Çok terimli kernel kullanarak destek vektör sınıflandırıcıyı eğitmek için SMO Algoritmasını uygular [33]. SMO algoritması, her aşamada mümkün olan en küçük optimizasyon sonucuna ulaşmayı amaçlar [34].

Bu algoritma, destek vektör makinelerinin eğitiminde ortaya çıkan ikinci dereceden programlama problemini iteratif bir şekilde bir dizi daha küçük alt problemlere bölmektedir. Destek vektör makinelerindeki ikinci dereceden programlama problemi 2. 3'deki gibi ifade edilmektedir.

$$max_{\propto} = \sum_{i=1}^{n} \propto_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j K(x_i x_j) \propto_i \propto_j \qquad \textbf{(2.3)}$$

2. 4' de *i' nin 1'den n'* e kadar tüm değerleri için α_i değerinin *0 ve C* değeri aralığında olması gerekmektedir. Ayrıca 2. 5' de *i'den n'* e kadar tüm α_i *ve* y_i değerlerinin çarpımlarının toplamları 0'a eşit olmalıdır. Elde edilen küçük alt problemler iki adet Lagrange çarpanı içermektedirler. Problem α_1*' in 0'dan* büyük eşit ve α_2'nin C değerinden küçük eşit olduğu durumda denklem 2. 6'daki hale dönüşmektedir.

$$0 \leq \propto_i \leq C \qquad \textbf{(2.4)}$$

$$\sum_i^n y_i \propto_i = 0 \qquad \textbf{(2.5)}$$

$$y_1 \propto_1 + y_2 \propto_2 = k \qquad \textbf{(2.6)}$$

Bu aşamadan sonra yapılması gereken tek boyutlu ikinci dereceden fonksiyonun minimumunun bulunmasıdır. Bunun için öncelikle optimizasyon problemi için uygun bir α_1 çarpanı seçilmektedir. Ardından bir α_2 çarpanı seçilmekte ve (α_1, α_2) çifti optimize edilmektedir. Bu iki aşama yakınsayana kadar devam etmektedir [35].

3. UYGULAMA VE BİLEŞENLERİ (IMPLEMENTATION AND COMPONENTS)

Geliştirilen metin madenciliği modelinin çalışma adımları aşağıda belirtildiği gibidir:

Adım 1. FATİH projesine yönelik görüşlerin metin madenciliği yöntemleri ile otomatik değerlendirilmesinin yapılması amacıyla uygulamada öncelikle internet ortamından alınan görüşler toplanarak bir veri seti oluşturulmuştur.

Adım 2. Bu veri seti çok sayıda metin dosyalarından yani yapısal olmayan veriden oluşmaktadır. Metin madenciliği yöntemlerinin bu veriler üzerinde uygulanabilmesi için bu verilerin yapısal veri haline dönüştürülmesi gerekmektedir. Bu bağlamda elde edilen veriler veri önişleme basamaklarından geçirilmiştir. Metin madenciliği veri önişleme bölümü; veri dönüştürme, tarama ve işaretleme, durak kelimelerin çıkarılması ve kök bulma olmak üzere dört aşamadan oluşmaktadır.

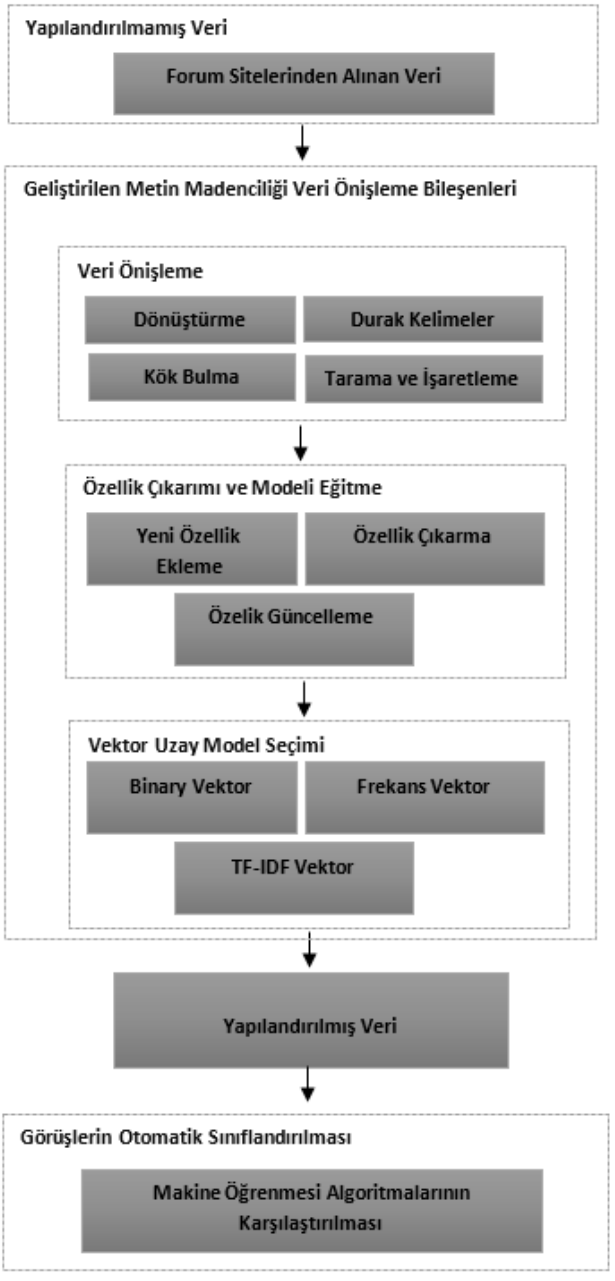
Adım 3. "Modelin eğitilmesi" bölümünde kullanıcının özellik çıkarmasına, eklemesine ve güncellemesine fırsat verilmektedir.

Adım 4. Her bir kelime vektör uzay modelinde temsil edilir. Kelimeler vektör uzay modelinde temsil edilirken binary vektör, frekans vektör ve tf-idf vektör olmak üzere üç yöntem kullanılmaktadır. Geliştirilen modelde kelimeler 3 ayrı vektör modelinde de temsil edilebilir. Yapısal hale

dönüştürülen veriler txt veya arff formatında kaydedilebilir.

Adım 5. Literatürde en fazla kullanılan makine öğrenmesi algoritmaları çalıştırılmış ve FATİH projesine yönelik görüşleri içeren metin dosyaları sınıflandırılarak algoritmaların bu veri seti üzerinde performansları karşılaştırılmıştır.

Uygulamanın akış şeması Şekil 1'de gösterilmiştir:



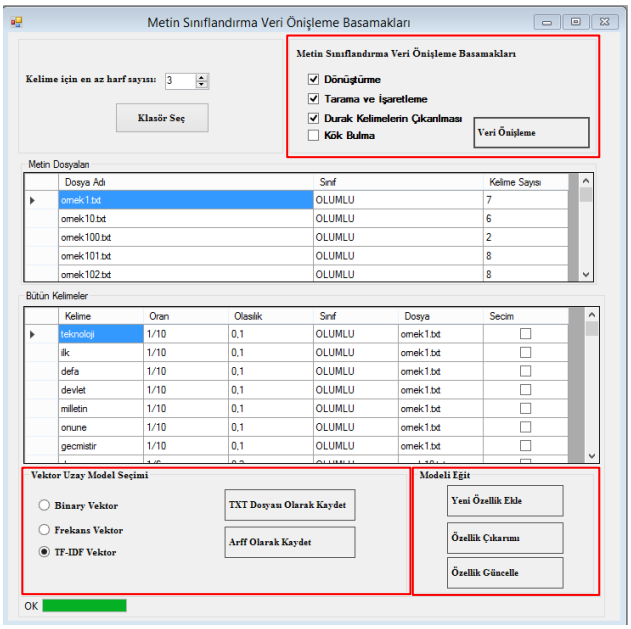
Şekil 1. Geliştirilen metin madenciliği modelinin akış şeması (Improved text mining model flowchart)

3.1. Verilerin Elde Edilmesi (Obtaining Data)

FATİH projesine yönelik internet ortamında bulunan görüşler alınarak bir veri seti oluşturulmuştur. Veri setinde 444 adet görüş bulunmaktadır. Bu veri seti pozitif ve negatif olarak gruplandırılmış ve sınıflandırıcıyı eğitmek üzere metin dosyalarına kaydedilmiştir. İnternet ortamından alınan bu yapılandırılmamış verilerin metin madenciliği ile sınıflandırılabilmesi için yapısal hale dönüştürülmesi gerekmektedir. Bu nedenle geliştirilen metin madenciliği veri önişleme yazılımı kullanılarak veri seti yapısal hale dönüştürülmüştür.

3.2. Geliştirilen Metin Madenciliği Veri Önişleme Yazılımının Bileşenleri (Improved Text Mining Data Pre-processing Software Components)

Metin madenciliği sürecinde yer alan basamaklar doğrultusunda modelin bileşenleri oluşturulmuştur. Uygulama kodlanırken Visual Studio C# programı kullanılmıştır. Yazılım çalıştırıldığında kullanıcıların karşısına Şekil 2' de yer alan form gelmektedir.

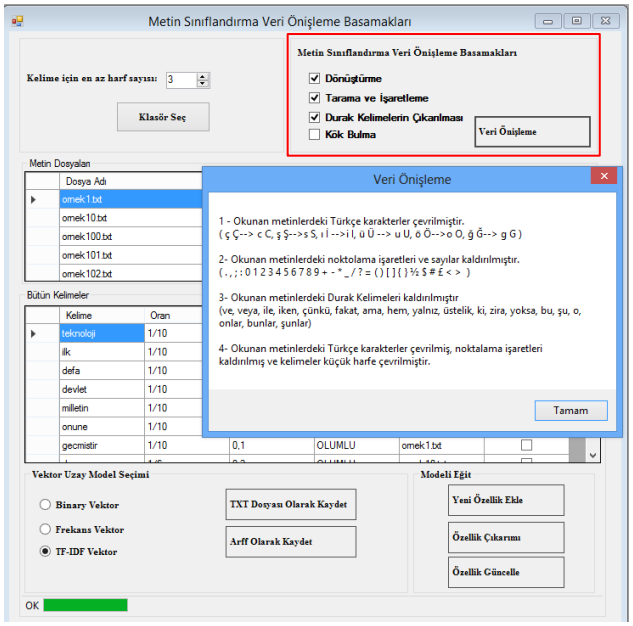


Şekil 2. Geliştirilen metin madenciliği veri önişleme yazılımı *(Improved Text Mining Data Pre-processing Software)*

Şekil 2'de verilen metin madenciliği veri önişleme yazılımı veri önişleme, modeli eğitme ve vektör uzay modelinin seçimi olmak üzere üç ana kısımdan oluşmaktadır. Yazılımda "Klasör Seç" butonu ile FATİH projesine yönelik görüşlerin bulunduğu metin dosyaları seçilmektedir. Metin dosyası içinde bulunan metinlerin kelime sayılabilmesi için gerekli harf sayısı belirlenebilmektedir.

3.2.1. *Veri önişleme (Data pre-processing)*

Metin dosyalarındaki verilerin yapısal veri haline dönüştürülmesinde geliştirilen "Metin Madenciliği Veri Önişleme Basamakları" yazılımı kullanılmıştır. Şekil 3' de veri önişleme basamaklarının bulunduğu kısım gösterilmektedir:



Şekil 3. Veri önişleme basamakları bölümü *(Data pre-processing)*

Alınan metin dosyalarındaki veriler aşağıdaki veri önişleme basamaklarından geçirilir:

a) Dönüştürme: FATİH projesine yönelik internet ortamından alınan metinler incelendiğinde Türkçe ve İngilizce karakterlerin karışık olarak kullanıldığı görülmüştür. Bu nedenle uygulamada standardizasyonun sağlanması amacıyla metinlerdeki Türkçe karakterler İngilizce karakterlere dönüştürülmüştür.

b) Durak kelimelerin çıkarılması: Uygulamada metin dosyaları içerisinde yer alan ancak sınıflandırma işleminde bir anlamı olmayan "ve, veya, ile, iken, çünkü, fakat, ama, hem, yalnız, üstelik, ki, zira, yoksa, bu, şu, o, onlar, bunlar, şunlar" gibi edat, bağlaç ve zamir gibi kelimeler metinden çıkartılmıştır.

c) Tarama ve işaretleme: Metin içindeki noktalama işaretleri, simgeler ve sayılar gibi terimler ayıklanmıştır. Metindeki cümleler kelimelere ayrılmış, tüm kelimeler küçük harfe çevrilmiştir.

d) Kök bulma: Çalışmada Türkçe doğal dil işleme kütüphanesi olan Zemberek programı kullanılarak kelimelerin kökleri tespit edilmiştir.

Uygulamada 1224 kelimeyi (özelliği) ve "Pozitif veya Negatif" değerini alabilen sınıf (class) bilgisini içeren bir veri seti oluşturulmuştur.

3.2.2. *Özellik seçimi ve modeli eğitme (Feature selection and model training)*

Modeli eğitme bölümünde; yeni özellik ekleme, özellik çıkarma ve özellik güncelleme olmak üzere 3 işlem yapılabilmektedir. Kullanıcı isterse bu bölümde manuel olarak veri setine özellik girip, çıkartabilir veya veri setini güncelleyebilir. "Yeni özellik ekleme" butonuna tıklandığında Datagrid nesnesi üzerine girilen yeni kayıt veritabanına kaydedilmektedir. Datagrid nesnesi üzerinde "Secim" kolonu içinde seçilen kayıt üzerinde özellik çıkarma ve özellik güncelleme işlemleri yapılabilmektedir. Bu sayede kullanıcı modeli eğiterek, modelin daha iyi çalışmasını sağlayabilmektedir.

3.2.3. Vektör uzay modeli seçimi (Vector Space Model)

Veri önişleme basamaklarından geçirilerek kelimelere ayrılan metin dosyasındaki bilgilerin üzerinde metin madenciliği yöntemlerinin uygulanabilmesi için vektör uzay modelinde temsil edilmeleri gerekmektedir. Kullanıcı bu bölümde vektör tanımlama yöntemlerinden (binary vektör, frekans vektör ve tf-idf vektör) birini seçerek kelimeleri vektör uzay modelinde temsil edebilmektedir.

Uygulamada özellikler (kelimeler) tf-idf vektör tanımlama yöntemi kullanılarak kelimeler vektör uzay modelinde temsil edilmiştir. Vektör uzay modelinde temsil edilen veri seti kullanıcının isteğine göre txt veya arff dosyası formatında kaydedilebilmektedir.

Geliştirilen metin madenciliği veri önişleme yazılımının en önemli ayırt edici özelliği; sadece FATİH projesine yönelik görüşlerinin veri önişleme basamağında değil, istenilen amaca yönelik metin sınıflandırma işleminin veri önişleme basamağında konudan bağımsız bir şekilde kullanılabilir olmasıdır. Yani yazılım; "Klasör Seç" butonu ile konudan bağımsız bir şekilde görüşlerin bulunduğu metin dosyalarının seçilmesine fırsat tanımaktadır.

3.3. FATİH Projesine Yönelik Görüşlerin Otomatik Değerlendirilmesi (Automatic Evaluation of Opinions Concerning FATİH Project)

Çalışmada veri önişlemi basamağından sonra, veri seti üzerinde makine öğrenmesi algoritmalarının uygulanması aşaması gerçekleştirilmiştir. Makine öğrenmesi algoritmalarının uygulanmasında açık kaynak kodlu WEKA programı kullanılmıştır. Veri seti 1224 özellik ve 1 sınıf (pozitif/ negatif) bilgisini içeren toplam 444 kayıttan oluşmaktadır. Veri seti; eğitim ve test verisi olarak ayrılır iken K- Kat çapraz geçerleme yöntemi kullanılmıştır.

K- Kat çapraz geçerleme tekniği; birkaç bin veya daha az satırdan meydana gelen küçük veri tabanlarında, verilerin k gruba ayrıldığı k katlı çapraz geçerlilik yöntemi kullanılabilir. Veri seti rastgele k adet gruba ayrılır. Bu yöntemde, ilk aşamada birinci grup test, diğer gruplar öğrenim için kullanılır. Bu süreç her defasında bir grubun test, diğer grupların öğrenim amaçlı kullanılması ile sürdürülür. Sonuçta elde edilen on hata oranının ortalaması, kurulan modelin tahmini hata oranı olacaktır. Literatürü incelediğimizde, genellikle k değerinin 10 olarak seçildiği görülmektedir [36]. Bu nedenle bu çalışmada da K-Kat çapraz geçerleme tekniğindeki k değeri 10 olarak belirlenmiştir.

Veri seti eğitim verisi ve test verisi olarak gruplandırıldıktan sonra, veri seti üzerinde literatürde en çok kullanılan makine öğrenmesi algoritmalarından Naive Bayes, K-En yakın komşu (k-NN, IBk), Karar Ağaçları (J48), SMO ve RBF Network [37] algoritmaları uygulanarak oluşturulan modellerin başarım ölçütleri karşılaştırılmıştır:

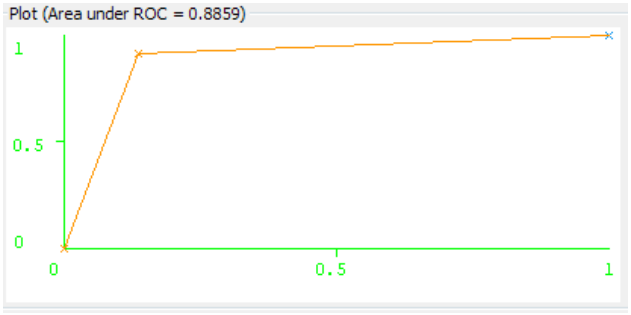
Tablo 1. Sınıflandırma algoritmalarının karşılaştırmaları
(Comparison of classification algorithms)

	Sınıflandırma Algoritmaları				
	Naive Bayes	**k-NN (IBk k=1)**	**J48**	**SMO**	**RBF Network**
Doğru Olarak Sınıflandırılan Örnek Sayısı	370	362	357	394	377
Yanlış Olarak Sınıflandırılan Örnek Sayısı	74	82	87	50	67
Doğru Pozitif (TP)	179	187	152	180	180
Yanlış Negatif (FN)	30	22	57	29	29
Yanlış Pozitif (FP)	44	60	30	21	38
Doğru Negatif (TN)	191	175	205	214	197
ROC Area değeri	0.896	0.915	0.859	0.8859	0.861
Recall değeri	0.833	0.815	0.804	0.887	0.849
Precision değeri	0.835	0.827	0.807	0.888	0.85
F-Ölçütü	0.833	0.815	0.803	0.887	0.849
Kappa İstatistiği	0.666	0.633	0.603	0.773	0.697
Başarı Yüzdesi (Doğruluk)	83.33	81.53	80.40	88.73	84.90

Tablo 1' de verilen sınıflandırma algoritmaları incelendiğinde, doğru olarak sınıflandırılan örnek sayısı (doğruluk yüzdesi) en yüksek olan algoritmanın SMO algoritması olduğu (%88,73) görülmektedir. Sınıflandırma algoritmalarında başarı yüzdesinin genel olarak %80 ve üzerinde olması beklenir. Karşılaştırılan sınıflandırma algoritmalarının başarı yüzdelerinin %80 ve üzerinde olduğu tespit edilmiştir.

Tablo1' de TP (True Positive) ve TN (True Negative) ile gösterilen sayılar sınıfları doğru tahmin edilen, FP (False Positive) ve FN (False Negative) ile gösterilen sayılar ise sınıfları yanlış tahmin edilen örneklerin sayılarını göstermektedir. En yüksek başarımı gösteren SMO algoritmasının karışıklık matrisi (TP, TN, FP ve FN değerleri) incelendiğinde doğru pozitif sınıflandırılan örnek sayısı 180, yanlış pozitif sınıflandırılan örnek sayısı 21, doğru negatif sınıflandırılan örnek sayısı 214 ve yanlış negatif sınıflandırılan örnek sayısı 29 olarak bulunmuştur. Toplamda doğru sınıflandırılan örnek sayısı 394 ve yanlış sınıflandırılan örnek sayısı 50 olarak tespit edilmiştir.

Çalışmada en yüksek başarımı gösteren SMO algoritmasına ait ROC eğrisi Şekil 4' te verilmiştir. SMO algoritmasına ait ROC eğirişi incelendiğine bu değerin 0,8859 olduğu görülmektedir. Bu değerin 1' e yakın olması algoritmanın tesadüfi bir tahminde bulunmadığını göstermektedir.



Şekil 4. SMO algoritmasına ait ROC eğrisi (SMO algorithm of ROC curve)

Model başarım ölçütlerinden recall, precision ve f ölçüt değerlerinin, ROC area değeri gibi 1' e yakın olması istenir. Karşılaştırma tablosu incelendiğinde tüm algoritmaların Roc area, recall, precision ve f ölçüt değerlerinin 0,80'den büyük olduğu görülmektedir. Ayrıca kappa istatistik değeri 0,6 ile 0,8 arasında ise önemli derecede bir uyum olduğunu, sınıflandırıcının tesadüfi bir tahminde bulunmadığını göstermektedir [38].

Literatürde konu ile ilgili benzer çalışmalar incelendiğinde otomatik değerlendirilmesi istenen metinlerin sınıflandırılmasında metin madenciliği yöntemlerinin kullanıldığı görülmektedir. Metin madenciliği yöntemleri; reklam içerikli epostaların otomatik tespiti [28], e-ticaret sitelerinin belirlenmesi [23], metinlerin otomatik sınıflandırılması [39], ürün yorumlarının otomatik değerlendirilmesi [18] ve metin madenciliği tekniklerinin kullanılarak öğrencilerin mesajlarından akademik durumlarının çıkarılması [40] gibi birçok dokümanın otomatik analizinde kullanılarak etkili sonuçlar elde edildiği rapor edilmiştir. Yorumları, görüş ve önerileri içeren metin dosyalarının sınıflandırılmasında paralel bir şekilde metin madenciliği yöntemlerinin kullanıldığı görülmektedir.

4. SONUÇ VE ÖNERİLER (CONCLUSION AND SUGGESTIONS)

Bu çalışmada; metin madenciliği yöntemleri ile FATİH projesine yönelik internet ortamında bulunan görüşler otomatik analiz edilmiş ve bu metin dosyalarını içeren veri seti üzerinde metin sınıflandırma algoritmalarının performansları karşılaştırılmıştır.

Çalışma kapsamında elde edilen sonuçlar aşağıda listelenmiştir:

- FATİH projesine yönelik görüşleri içeren metin dosyaları metin madenciliği yapılabilecek veri formatına dönüştürülebilmesi için "Metin Madenciliği Veri Önişleme" yazılımı geliştirilmiştir. Geliştirilen veri önişleme yazılımı kullanılarak yapısal veri olmayan metin dosyaları yapısal veri haline dönüştürülmektedir.

- Geliştirilen yazılımda; veri setine yeni özellikler eklenebilmekte, mevcut özellikler silinip, düzenlenebilmekte yani modelin performansının iyileştirilebilmesi içim modelin eğitilebilmesine fırsat verilmektedir. Başka bir ifadeyle kullanıcı tarafından modele dinamiklik katılarak modelin performansı artırılabilmektedir.

- Veri seti üzerinde literatürde en çok kullanılan metin sınıflandırma algoritmalarının model başarım ölçütlerine göre karşılaştırmalı analizi yapılmıştır.

- Kullanılan metin sınıflandırma algoritmalarının tümünün başarı yüzdelerinin 0.80' den büyük olduğu tespit edilmiştir. Performansı karşılaştırılan algoritmalardan en yüksek başarımın SMO algoritmasına ait olduğu (%88,73) gözlemlenmiştir.

- Model başarım ölçütlerinden ROC area, recall, precision ve f ölçüt değerlerinin 1' e yakın olması durumu sınıflandırıcının tesadüfi bir tahminde bulunmadığına işaret eder [5, 38]. Tablo 1 incelendiğinde söz konusu değerlerinin 1'e yakın olduğu görülmektedir. Bu bulgu önemli derecede bir uyum olduğunu, sınıflandırıcının tesadüfi bir tahminde bulunmadığını göstermektedir.

- Çalışmada FATİH projesine yönelik görüşlerin olumlu ya da olumsuz yargı içerdiğinin metin madenciliği yöntemleri ile %88,73 doğruluk oranı ile otomatik tespiti yapılmıştır.

Uygulamada; FATİH projesine yönelik 444 adet görüş toplanarak bir veri seti oluşturulmuş ve özellikler tf-idf vektör tanımlama yöntemi kullanılarak vektör uzay modelinde temsil edilmiştir. Daha sonraki çalışmalarda FATİH projesine ait görüşlerin bulunduğu doküman sayısının artırılarak örneklemin genişletilmesi ve diğer vektör tanımlama yöntemleri (binary vektör, frekans vektör) uygulanarak performanslarının birbirleriyle karşılaştırılması önerilebilir. Özellikle günlük hayatta karşılaşılan verilerin yaklaşık %85'nin yapısal olmayan veri olduğu [41] ve bu verilerin belli bir amaç doğrultusunda işlendiği zaman bir anlam ifade edeceği düşünüldüğünde [42], mevcut çalışmanın bu konu ile ilgili çalışacak araştırmacılara ve FATİH projesini yürüten konu uzmanlarının karar alma süreçlerine destek sağlayacağı düşünülmektedir.

KAYNAKLAR (REFERENCES)

[1] K. Doğan, S. Arslantekin, "Büyük Veri: Önemi, Yapısı ve Günümüzdeki Durum", *DTCF Dergisi*, 56 (1), 15-36, 2016.

[2] M. Ö. Dolgun, T. G. Özdemir, D. Oğuz, "Veri Madenciliğinde Yapısal Olmayan Verinin Analizi: Metin ve Web Madenciliği", *İstatikçiler Dergisi*, 2(2), 48-58, 2009.

[3] A. Karadağ, H. Takçı, "Metin Madenciliği ile Benzer Haber Tespiti", **Akademik Bilişim Konferansı Bildirileri**, Muğla Üniversitesi, 10-12 Şubat, 2010.

[4] M. F. Amasyalı, B. Diri, F. Türkoğlu, "Farklı Özellik Vektörleri ile Türkçe Dokümanların Yazarlarının Belirlenmesi", **15. Türkiye Yapay Sinir Ağları Sempozyumu**, Muğla, 21- 24 Haziran, 2006.

[5] A. C. Tantuğ, "Metin Sınıflandırma (Text Classification)", *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, *5*(2), 2012.

[6] Aşlıyan, R., Günal, K., "Metin İçerikli Türkçe Doküman Sınıflandırılması", **Akademik Bilişim Konferansı Bildirileri**, Muğla Üniversitesi, 10-12 Şubat, 659-665, 2010.

[7] L. Kuzucu, **Müşteri memnuniyeti belirlemek için metin madenciliği tabanlı bir yazılım aracı**, Yüksek Lisans Tezi, Maltepe Üniversitesi, Fen Bilimleri Enstitüsü, 2015.

[8] H. Núñez, E. Ramos, "Automatic Classification of Academic Documents Using Text Mining Techniques", **In Informatica (CLEI) XXXVIII Conferencia Latinoamericana**, Medelin-Colombia, 1-5 October, 1-7, 2012.

[9] P. Sobkowicz, M. Kaschesky, G. Bouchard, "Opinion Mining in Social Media: Modeling, Simulating, and Forecasting Political Opinions in the Web", *Government Information Quarterly*, 29, 470-479, 2012.

[10] V. Levent, B. Diri, "Türkçe Dokümanlarda Yapay Sinir Ağları İle Yazar Tanıma", **Akademik Bilişim'14 Konferansı Bildirileri**, Mersin Üniversitesi, 5-7 Şubat, 735-741, 2014.

[11] M. M. Mostafa, "More Than Words: Social Networks' Text Mining For Consumer Brand Sentiments", *Expert Systems with Applications, 40*(10), 4241-4251,2013.

[12] A. Visa, "Technology of text mining", **In International Workshop on Machine Learning and Data Mining in Pattern Recognition**, Springer-Verlag London, July 25-27, 1–11, 2001.

[13] F. Erten, **Metin madenciliği tabanlı bir web sitesi sınıflandırma aracı tasarımı**, Yüksek Lisans Tezi, Maltepe üniversitesi, Fen bilimleri Enstitüsü, 2015.

[14] D. Kılınç, E. Borandağ, F. Yücalar, V. Tunalı, M. Şimşek, A. Özçift, "KNN Algoritması ve R Dili ile Metin Madenciliği Kullanılarak

Bilimsel Makale Tasnifi", *Marmara Fen Bilimleri Dergisi*, 28(3), 89-94. 2016.

[15] V. Tunalı, **Metin madenciliği için iyileştirilmiş bir kümeleme yapısının tasarımı ve uygulaması**", Doktora Tezi, Marmara Üniversitesi, Fen Bilimleri Enstitüsü, 2011.

[16] M. Çetin, M. F. Amasyalı, "Active Learning for Turkish Sentiment Analysis", **IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA)**, Albena-Bulgaria, 19-21 June, 1–4, 2013.

[17] D. Kılınç, F. Bozyiğit, A. Özçift, F. Yucalar, E. Borandağ, "Metin Madenciliği Kullanılarak Yazılım Kullanımına Dair Bulguların Elde Edilmesi", **9. Ulusal Yazılım Mühendisliği Sempozyumu**, Yaşar Üniversitesi, 9-11 Eylül, 2015.

[18] K. Ergün, **Metin madenciliği yöntemleri ile ürün yorumlarının otomatik değerlendirilmesi**, Doktora Tezi, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, 2012.

[19] H. Göker, H. I. Bülbül, "Improving an Early Warning System to Prediction of Student Examination Achievement", **13th International Conference on Machine Learning and Applications (ICMLA'13),** Detroit-USA, 3-5 December, 568-573, 2014.

[20] A. Haltaş, A. Alkan, M. Karabulut, "Metin Sınıflandırmada Sezgisel Arama Algoritmalarının Performans Analizi", *Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, 30(3), 417-427, 2015.

[21] Y. Argüden, B. Erşahin, **Veri Madenciliği: Veriden Bilgiye, Masraftan Değere**, ARGE Danışmanlık Yayınları, İstanbul, 2008.

[22] A. Güven, **Türkçe belgelerin anlam tabanlı yöntemlerle madenciliği**, Doktora Tezi, Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, 2007.

[23] T. Kaşıkçı, H. Gökçen, "Metin Madenciliği İle E-Ticaret Sitelerinin Belirlenmesi", *Bilişim Teknolojileri Dergisi*, 7(1), 25-32, 2014.

[24] E. Alpaydın, **Introduction to Machine Learning,** The MIT Press, London, 2004.

[25] A. Tunç, İ. Ülger, "Veri Madenciliği Uygulamalarında Özellik Seçimi İçin Finansal Değerlere Binning ve Five Number Summary Metotları İle Normalizasyon İşleminin Uygulanması", **XVIII. Akademik Bilişim Konferansı**, Adnan Menderes Üniversitesi, 30 Ocak- 5 Şubat, 2016.

[26] H. Almuallim, T. G. Dietterich, **Learning with many Irrelevant Features**, AAAI Pres, California, 1991.

[27] S. İlhan, N. Duru, Ş. Karagöz, M. Sağır, "Metin Madenciliği ile Soru Cevaplama Sistemi", **Elektronik ve Bilgisayar Mühendisliği Sempozyumu (ELECO)**, Bursa, 26-30 Kasım, 356-359, 2008.

[28] K. Çalış, O. Gazdağı, O. Yıldız, "Reklam İçerikli Epostaların Metin Madenciliği Yöntemleri ile Otomatik Tespiti", *Bilişim Teknolojileri Dergisi*, 6(1), 1-7, 2013.

[29] P. Soucy, W. Mineau, "Beyond TFIDF Weighting for Text Categorization in the Vector Space Model", **IEEE International Conference**,Edinburgh-Scotland, July 30- August 05, 1130-1135, 2005.

[30] J. Han, M. Kamber, J. Pei, **Data Mining: Concepts and Techniques**, Morgan Kaufmann Publishers, USA, 2011.

[31] C. Zhang, S. Zhang, **Association Rule Mining, Models and Algoritms**, Springer, USA, 2002.

[32] Y. Özkan, **Veri Madenciliği Yöntemleri**, Papatya Yayıncılık Eğitim, İstanbul, 2008.

[33] E. Ardıl, **Esnek hesaplama yaklaşımı ile yazılım hata kestirimi**, Yüksek Lisans Tezi, Trakya Üniversitesi, Fen Bilimleri Enstitüsü, 2009.

[34] M. Elmas, **Destek vektör makineleri ile fiyat tahminleri ve kuyumculuk sektöründe bir uygulama**, Yüksek Lisans Tezi, İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, 2012.

[35] J. C. Platt, 12 fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods*, 185-208, 1999.

[36] H. Akpınar, "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", *İstanbul Üniversitesi İşletme Fakültesi Dergisi*, 29(1), 1-22, 2000.

[37] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh,Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. Zhou, M. Steinbach, D. J. Hand, D. Steinberg, "Top 10 Algorithms in Data Mining", *Knowledge and Information Systems*, 14(1), 1-37, 2008.

[38] H. Göker, **Üniversite giriş sınavında öğrencilerin başarılarının veri madenciliği yöntemleri ile tahmin edilmesi**, Yüksek Lisans Tezi, Gazi Üniversitesi, Bilişim Enstitüsü, 2012.

[39] H. K. Mohamed, "Automatic documents classification", **Computer Engineering & Systems, ICCES'07. International Conference on IEEE,** 27- 29 November, 33-37, 2007.

[40] W. He, "Examining Students' Online Interaction In A Live Video Streaming Environment Using Data Mining and Text Mining". Computers in Human Behavior, 29(1), 90-102, 2013.

[41] F. S. Gharehchopogh, Z. A. Khalifelu, "Analysis and Evaluation of Unstructured Data: Text Mining Versus Natural Language Processing", 5th International Conference on Application of Information and Communication Technologies (AICT), Baku-Azerbaijan, 12-14 October, 1-4, 2011

[42] S. Savaş, N. Topaloğlu, M. Yılmaz, "Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri", *İstanbul. Ticaret Üniversitesi Fen Bilimleri Dergisi*, 11(21), 1–23, 2012