Reklam İçerikli Epostaların Metin Madenciliği Yöntemleri ile Otomatik Tespiti

Kübra ÇALIŞ, Oya GAZDAĞI, Oktay YILDIZ

Bilgisayar Mühendisliği, Mühendislik Fakültesi, Gazi Üniversitesi, Ankara, Türkiye
kcaliscalis@gmail.com, oyagazdagi@hotmail.com, oyildiz@gazi.edu.tr
(Geliş/Received: 14.10.2012; Kabul/Accepted: 21.03.2013)

Özet— Elektronik posta (eposta), internet kullanımının yaygınlaşması, basit ve kolay erişilebilir olması sebebiyle son kırk yıl içinde ciddi oranda artarak, günümüzde en yaygın kullanılan iletişim aracı olmuştur. Artan eposta kullanımı birtakım sorunları da beraberinde getirmiştir. Ortaya çıkan önemli sorunlardan biri istenmeyen, reklam içerikli elektronik postalardır. Bunlar, eposta kullanıcılarını rahatsız etmekte, ayrıca gereksiz kaynak israfına yol açmaktadırlar. Reklam içerikli epostalar ile legal veya illegal pek çok ürünün tanıtımı yapılmakta, pek çok kaynaktan farklı amaçla yollanan milyonlarca istenmeyen eposta internet kullanıcılarının posta kutularını doldurmaktadır. Giderek büyük bir sorun haline gelen reklam epostaları, hem internet trafiğini hem de posta sunucularını meşgul etmektedir. İstenmeyen epostaların filtrelemesi üzerine pek çok çalışma yapılmış olmasına rağmen Türkçe içerikli reklam epostalarının filtrelenmesi üzerine yapılan çalışma çok azdır. Bu çalışmalar incelendiğinde ya başarı istenen düzeyde değildir ya da önerilen algoritmalar oldukça karmaşıktır. Bu çalışmada metin madenciliği yöntemleri kullanılarak Türkçe içerikli reklam epostalarının tespiti gerçekleştirilmiştir. Bu amaçla Destek Vektör Makinesi, k En Yakın Komşu ve Naive Bayes sınıflandırma algoritmaları kullanılmıştır. Çalışmada reklam içerikli eposta binary, frekans ve TF-IDF ağırlıklandırma yöntemleri ile vektörel olarak ifade edilmiştir. Yapılan çalışmada Reklam epostalarının tespit edilmesi için Türkçe içerikli 400'ü normal, 400'ü de reklam içerikli olmak üzere 800 eposta kullanılmıştır. Yapılan deneysel çalışmalarda reklam epostaları, kNN ile %96,5 doğrulukta sınıflandırma başarısı elde edilmiştir.

Anahtar Kelimeler— Spam, reklam eposta tespiti, metin madenciliği, sınıflandırma

Automatic Detection of Advertising Email Using Text Mining

Abstract— Today, electronic mail (email) is one of the widely used communication tool because it is simple and easily accessible. With increasing number of internet usage, e-mail users have been increased dramatically in the last four decade. By the way, it has brought many problems. Unwanted email issue is one of the biggest problem for internet users. This type of emails often contains malicious codes and consumes redundant internet resources. At the same time, user's run out of mail quota due to the legal or illegal content. In literature, many unwanted email filter approaches are proposed, however neither them are successfully applicable. In recent years, researchers try to find best, simple and feasible method. For that reason, one of the promising research field emerges to overcome this problem which is called text mining for filtering unwanted email. On the other hand, globalization is another concern for non English based email filtering such as Turkish. In this study, three different classification algorithms (Support Vector Machine, k Nearest Neighbor and, Naive Bayes) were used to determine unwanted Turkish contents. Our dataset contains 800 samples as 400 normal and 400 unwanted emails. In order to achieve these tasks, emails were transformed into binary, frequency and TF-IDF vectors for proper classification. The best accuracy was obtained with k-nearest-neighbor algorithm with respect to the 96.5% accuracy rate.

Keywords— Spam, advertising email detection, text mining, classification

1. GİRİŞ (INTRODUCTION)

Elektronik posta, günümüzde en yaygın kullanılan iletişim araçlarının başında gelmektedir. İnternet kullanımının yaygınlaşması, elektronik posta yoluyla rahat ve ucuz iletişim kurulması, eposta kullanıcılarının sayısını son kırk yıl içinde ciddi oranda artırmıştır.

Eposta, ortaya çıktığı ilk zamanlarda sadece metin gönderme amacıyla kullanılırken, günümüzde geliştirilen tekniklerle, çeşitli medya öğeleri ve çalışabilir program uygulamalarının iletimini de sağlayacak duruma gelmiştir. Bu teknolojik gelişmeler beraberinde birtakım sorunları da getirmiştir. Ortaya çıkan önemli sorunlardan biri

istenmeyen, reklam içerikli epostalardır. Reklam spamları ile internet kullanıcılarına legal veya illegal pek çok ürünün tanıtımı yapılmakta, pek çok kaynaktan farklı amaçlarla yollanan milyonlarca istenmeyen elektronik posta internet kullanıcılarının posta kutularını doldurmaktadır. Giderek büyük bir sorun haline gelen bu konu, hem internet için kullanılan iletişim hatlarını boşa meşgul etmekte hem de sunucuların yükünü arttırmaktadır.

İstenmeyen elektronik postaların belirlenmesi amacıyla yapılmış çok sayıda çalışma literatürde yer almaktadır. Bunlar içinde makine öğrenme yöntemleri ve metin madenciliği spam belirlemede başarılı bir şekilde kullanılmaktadır.

Sasaki ve Shinnou [1] yaptıkları spam algılama çalışmasında, Navïe Bayes (NB), Destek Vektör Makineleri (DVM) ve birçok makine öğrenmesi yaklaşımlarını kullanmışlardır. Destek vektör makinesi ile %97 oranında spam epostalarını doğru tespit etmişlerdir. Clark ve arkadaşları [2] çalışmalarında yapay sinir ağlarını, sıklık, terim sıklığı- devrik belge sıklığı (TF-IDF), mantıksal ağırlıklandırmalar ile birlikte kullanmışlardır. En iyi sonuca sıklık ve TF-IDF ağırlıklandırma ile ulaşmışlardır. Sakkis ve arkadaşları [3] yaptıkları çalışmada Navïe Bayes ve k En Yakın Komşu (kNN) sınıflandırıcılarının diğer yöntemlere göre sınıflandırma işleminde daha iyi performans sergilediğini ortaya koymuşlardır. kNN sınıflandırıcı ile %98,8, NB ile %99,5 başarı elde etmişlerdir. Bayes yaklaşımını kullanan bir diğer çalışmayı Sahami ve arkadaşları yapmışlardır [4]. Bu çalışmada istenmeyen epostaların sınıflandırılmasında %99,9 başarı elde edilmiştir. Cohen [5] çalışmasında Önermeli Öğrenici Ripper (Repeated Incremental Pruning to Produce Error Reduction) ile öğrenme kurallarını kullanarak eposta sınıflandırması yapmıştır. Wang ve arkadaşları [6] hazırladıkları spam algılama çalışmasında DVM kullanmışlardır. Yapılan bu çalışmada sınıflandırma işlemi için gerekli öznitelik seçimi genetik algoritma ile gerçekleştirilmiştir. Spam etiketlemede %87,89 başarı oranı elde edilmiştir. Kun-Lun Li ve arkadaşları [7] öğrenen bir sistem temeline oluşturdukları çalışmalarında DVM sınıflandırıcısı ile epostaları hızlı ve rahat bir şekilde kategorize etme yollarını araştırmışlardır. Çalışma sonucunda teorik bir sonuç elde edilmiş olmakla birlikte spam algılamada DVM yönteminin NB yönteminde daha başarılı olduğu sonucuna varmışlardır. Chih-Chin Lai ve Ming-Chi Tsai [8] tarafından gerçekleştirilen çalışmada ise farklı makine öğrenmesi yaklaşımları birbiri ile karşılaştırılmış, istenmeyen epostaların algılanmasında verinin başlık kısmının incelenmesinin sınıflandırma başarımını artırmada büyük önem arz ettiğini ortaya koymuşlardır. Pantel ve Lin'in [9] gerçekleştirdikleri çalışmada NB ile eposta sınıflandırmasında %92 oranında sınıflandırma başarısı elde edilmiştir. Metsis ve arkadaşları [10] 2006 yılında hazırladıkları spam filtreleme çalışmasında NB versiyonlarından Multinominali'nin binary vektör ile kullanımında %97,53 oranında sınıflandırma başarısı elde etmişlerdir. Androutsopoulos ve arkadaşları [11] üzerinde çalıştıkları spam filtrelemede, NB yöntemini ile %99.993 başarım elde etmişlerdir. Blanzieri ve Bryl [12] çalışmalarında daha önce yapılmış spam filtreleme çalışmalarını incelemişlerdir. Bu inceleme sonucunda Kelime Çantası vektör yönteminin NB, kNN, DVM ve Boosting sınıflandırıcılarla uyumlu olduğunu görmüşlerdir. McCue [13] çalışmasında DVM ile eposta sınıflandırmada %96,6 oranında sınıflandırma başarısı elde etmiştir.

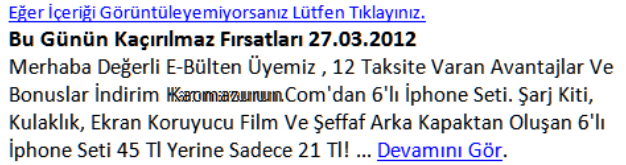
Yapılan literatür çalışmalarında Türkçe içerikli epostaların tespit edilmesi ile ilgili çalışmaların çok az olduğu, var olan çalışmaların da beklenen sınıflandırma başarısını göstermediği ya da çok karmaşık olduğu görülmüştür. Bu çalışmada metin madenciliği yöntemleri kullanılarak Türkçe içerikli reklam epostalarının tespiti gerçekleştirilmiştir. Bu amaçla Destek Vektör Makinesi, k En Yakın Komşu ve Naive Bayes sınıflandırma algoritmaları kullanılmıştır. yapılan çalışmada reklam içerikli eposta metinleri, binary, frekans ve TF-IDF ağırlıklandırma yöntemleri ile vektörel olarak ifade edilmiştir.

Yapılan çalışmada Türkçe reklam epostalarının tespit edilmesi için 400'ü normal, 400'ü de reklam epostası olan toplam 800 eposta kullanılmıştır. Bu epostalardan rastsal olarak belirlenmiş 200 eposta test, 600 eposta eğitim amaçlı kullanılmıştır. Çalışma sonucunda yapılan deneylerde en yüksek başarı kNN ile %96,5 olarak elde edilmiştir.

2. MATERYAL VE METOT (MATERIAL AND METHOD)

2.1. Reklam İçerikli Eposta Veri Kümesi (Advertising Email Dataset)

Reklam içerikli epostaların belirlenmesi için Türkçe içerikli 800 eposta toplanmıştır. Bu epostaların 400'ü normal, 400'ü reklam içerikli epostadır. Elde edilen eposta dört ana başlık altında bilgi içermektedir. Bunlar epostanın başlık bilgisi, gönderen bilgisi, adres bilgisi ve eposta içeriğidir. Şekil 1'de reklam eposta örneği görülmektedir.


Eğer İçeriği Görüntüleyemiyorsanız Lütfen Tıklayınız.
Bu Günün Kaçırılmaz Fırsatları 27.03.2012
Merhaba Değerli E-Bülten Üyemiz , 12 Taksite Varan Avantajlar Ve Bonuslar İndirim Kapımızın[illegible]Com'dan 6'lı İphone Seti. Şarj Kiti, Kulaklık, Ekran Koruyucu Film Ve Şeffaf Arka Kapaktan Oluşan 6'lı İphone Seti 45 Tl Yerine Sadece 21 Tl! ... Devamını Gör.

Şekil 1. Reklam eposta örneği
(Sample of advertising email)

2.2. Metin Madenciliği (Text Mining)

Veri madenciliği, istatistik, veri tabanı yönetim sistemleri ve makine öğrenmesi gibi birçok disiplini kullanarak daha önce keşfedilmemiş ve açık bir şekilde ortada olmayan bilgiyi çıkarmak için kullanılan veri analiz metodudur [14]. Veri madenciliği disiplini içerisinde yer alan ve bu disiplinin alt bir dalı olarak isimlendirilen Metin Madenciliği kavramı ise metinsel dokümanlar içerisinde keşfedilmemiş bilgileri bilgisayar sistemleri sayesinde

ortaya çıkartılmasında kullanılan bir bilim dalıdır [15]. Doğal dil işleme metotları ile metinsel dokümanlardan bilgi elde edilmesinde kullanılan bu disiplinin temel amacı yapısal olmayan yani kesin formatlarla tanımlanamayan yapılardan bir çıkarım yapmaktır. Metin madenciliği disiplini ile metinlerin hızlı bir şekilde sınıflandırılması, özetlenmesi, benzer metinler ile aralarındaki ilişkilerin ortaya konması başarılı bir şekilde yapılabilmektedir. Bu çalışmada metin madenciliğinde başarılı bir şekilde kullanılan sınıflandırma algoritmalarından Naive Bayes (NB), k En Yakın Komşu (kNN) ve Destek Vektör Makinesi (DVM) kullanılmıştır.

2.2.1. Sınıflandırma (Classification)

Sınıflandırma, etiket (sınıf) değerleri önceden bilinen verikümesinden elde edilen model ile, sınıfı bilinmeyen yeni verinin etiketinin tahmin edilmesi işlemidir. Bu yöntemde, eğitim kümesi olarak adlandırılan ve önceden etiket değeri bilinen bir verikümesine ihtiyaç duyulur [16,17]. Sıklıkla kullanılan sınıflandırma yöntemleri şunlardır:

Naive Bayes (Naive Bayes)
Naive Bayes, verilerin etiketlenmesi için kullanılan gözetimli bir sınıflandırma algoritmasıdır. Verilerin sınıflandırılmasında sıklıkla tercih edilen bu algoritmanın kullanımı oldukça kolaydır. Genel olarak Naive Bayes sınıflandırıcısı her kriterin sonuca olan etkilerinin olasılık değerlerinin hesaplanmasına dayanır. Naive Bayes, bir belgenin ya da verinin bulunduğu sınıfının olasılığını tahmin etmek için verilen kelime ya da verinin ait olma ihtimalinin bulunduğu sınıfının koşullu olasılıklarını hesaplar [18]. Bayes teoremi aşağıdaki eşitlik ile ifade edilebilir:

$$P(A/B)=(P(B/A)*P(A))/P(B) \quad (1)$$

Burada P(A), A olayının bağımsız olasılığı (öncül olasılık), P(B), B olayının bağımsız olasılığı, P(B|A), A olayının olduğu bilindiğinde B olayının olasılığı (şartlı olasılık), P(A|B), B olayının olduğu bilindiğinde A olayının olasılığıdır (şartlı olasılık). Böylece P(A|B)'yi maksimum yapan durumlar bulunarak yeni gelen örneğin sınıfı tahmin edilebilir.

k En Yakın Komşu (k Nearest Neighbor)
kNN algoritması, öğrenme temelli bir algoritmadır. Eğitim verisinden öğrenmiş olduğu modeli test verisinde kullanarak sınıf etiketi ataması yapar. Her yeni gelenin sınıf etiketi, belli olan bu modele göre sınıf etiketi atanmaktadır. Algoritmanın temel yapısı ve işleyişi her yeni gelen örnek için, onun en yakınında bulunan k komşuya bakarak bir sınıf etiketi öngörüsünde bulunmaktır. kNN algoritmasının herhangi bir sınıf etiketi belli olmayan veri için şu yol izlenir. kNN sınıflandırıcısında öncelikle k değeri belirlenir. Belirlenen k değerlerine göre öğrenme kümesinde bulunan tüm verilere göre yeni verinin uzaklıkları hesaplanır ve minimum uzaklığa sahip verilerin bulunduğu sınıf etiketlerine bakılarak yeni örnek için bir sınıf etiketi öngörüsünde bulunulur [19].

Destek Vektör Makinesi (Support Vektor Machine)
Destek Vektör Makinesi doğrusal ve doğrusal olmayan verilerin sınıflandırmasında kullanılan başarılı bir sınıflandırma algoritmasıdır. Vapnik tarafından geliştirilen bu algoritmanın temel çalışma prensibi, etiket olarak atanacak sınıflar arasında en büyük ayrımı sağlayan hiper düzlemi bulmak ve oluşturulan modele göre sınıfı bilinmeyen örneğin sınıflandırmasını sağlamaktır [20].

2.2.2. Metnin vektörel olarak ifade edilmesi (Vectorial ezpression of the text)

Metinsel verilerin sınıflandırması, yapısal olmayan ve doğal dil işleme kuralları ile oluşturulan metinler üzerinde yapılmaktadır. Sınıflandırma algoritmaları tarafından işlem yapılabilmesi için bu verilerin uygun formlara çevrilmesi gerekmektedir. Bu işlem sonucunda sınıflandırıcılar için giriş olacak vektör uzayları elde edilmektedir. Vektör uzayı modelinde her bir veri uzayda bir nokta olarak ifade edilmektedir. Verinin sahip olduğu öznitelik değerleri noktanın koordinatları olarak ifade edilmektedir. Yapılan literatür çalışmasında, vektör oluşturmada öznitelik seçiminde "BoW-Kelime Çantası" yönteminin başarılı sonuçlar verdiği görülmüştür. Bu yöntem ile vektörü oluşturacak öznitelikler, veri kümesi üzerindeki kelimelerin köklerine ayrıştırmasıyla veya en sık rastlanan kelimelerin göz ardı edilmesi ile belirlenir [21].

Çalışmada üç farklı vektör tanımlama yöntemi kullanılmıştır. Bunlar:

Binary Vektör oluşturma (Formation Binary Vector)
Bu yöntem ile metinsel veriler 1 ve 0' lar ile ifade edilmektedir. Veri içinde barındırdığı kelimelerin sözlükteki varlıklarına göre bu değerleri almaktadırlar. Bu şöyle örneklenebilir:

Veri: "Ali bugün yağmur yağacak şemsiye almayı unutmamalısın."
Sözlük: {Bugün, Akşam, Şemsiye, Güneş}
Binary Tanımlama: {1, 0, 1, 0}

Frekans Sıklığı ile Vektör oluşturma (Formation Vector with Term Frequency)
Binary tanımlamadan farklı olarak veri içinde bulunan kelime köklerinin kaç defa geçtiği bilgisinin de tutularak yapıldığı bir tanımlama biçimidir.

Veri: "Bugün hava çok güneşli. Bugün pikniğe gidelim mi?"
Sözlük: {Bugün, Akşam, Şemsiye, Güneş}
Frekans Tanımla: {2, 0, 0, 1}

TF–IDF Ağırlıklandırma ile Vektör oluşturma(Formation Vector with TF-IDF Weighting)

Bu yöntemde birinci adım ağırlıklandırma çalışması ile tüm dokümanda geçen verilerin, aranan verilere oranının belirlenmesidir. Böylece veri havuzu içinde belirleyici yani ayırt edici özelliği fazla olan ya da dokümanlar içerisinde sürekli geçen ve bu sebepten ötürü ayırt edici özelliği bulunmayan kelimeler belirlenir. Ağırlıklandırma içerisinde kullanılan TF değeri, kelime köklerinin belirlenen veri içeriğinde kaç kere geçtiği yani frekans bilgisini tutmaktadır. IDF değeri ise aranan kelime sözcüğün tüm veri havuzu içinde kaç kez geçtiği bilgisi ile ilgili değeri vermektedir. Bu değer sayesinde tüm veri havuzu içinde bulunan ve ayırt edici özelliği fazla olan kökler ile her veri içeriğinde geçip ayırt edici özelliği olmayan kelime kökleri rahatlıkla belirlenebilmektedir.

Aşağıda Eşitlik 2'de TF ve IDF hesaplanması, Eşitlik 3'te ise ağırlıklandırma hesaplanması gösterilmiştir.

$$TF_{ij} = \frac{n_{ij}}{|d_i|} \quad , \quad IDF_{ij} = \log_2\left(\frac{n}{n_j}\right) \tag{2}$$

Ağırlıklandırma = TF_{ij} x IDF_{ij} (3)

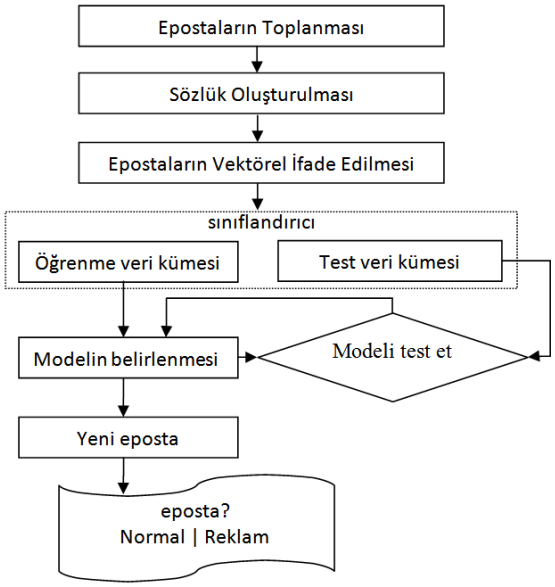
TF değerinin hesaplanmasında kullanılan *n* değeri, *j* nci kelime kökünün toplanan *i* nci veri seti içinde kaç defa geçtiği sayıdır. *d* değeri ise veri seti içinde yer alan tüm kelime köklerinin sayısıdır. Formül içinde yer alan *i* değeri ise eposta içinde yer alan kelimelerin sayısıdır. IDF değerinin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

3. REKLAM İÇERİKLİ EPOSTALARIN BELİRLENMESİ (DETERMINING OF ADVERTISING EMAILS)

Şekil 2'de reklam içerikli epostaların belirlenmesinde kullanılan modelin yapısı görülmektedir. Reklam içerikli epostaların belirlenmesi amacıyla yapılması gereken ilk iş toplanan epostalardan sözlük oluşturulması işlemidir. Bu işlem farklı eposta kullanıcılarından toplanan postaların içerik kısımlarının Zemberek Kütüphanesi yardımıyla kelime köklerine ayrılarak bir sözlük oluşturulmasıdır. Sonraki adım ise bu epostaların vektörel olarak ifade edilmesidir. Son olarak reklam içerikli epostaların sınıflandırma algoritmaları kullanımıyla tespiti gerçekleştirilir.

3.1. Ön işlem Aşaması (*Pre-processing Stage*)

Reklam içerikli epostaların belirlenmesinde içerik bilgisi dikkate alınmıştır. Bu sebeple öncelikle içerik bilgisinin elde edilmesi ve ayrıca eposta içeriğinde yer alan kelimelerin belirlenmesi gerekmektedir. eposta içeriğine yer alan kelimelerin ayrıştırılmasında Türkçe içerikli metinlerde kelime kök ve gövdesini başarılı bir şekilde bulan zemberek kütüphanesinden faydalanılmıştır. Eposta eğitim veri kümesinin tamamı Zemberek Kütüphanesi kullanımı ile oluşturulmuş ve köklerine ayrılmış 4263 kelime içeren bir sözlük elde edilmiştir.



Şekil 2. Gerçekleştirilen çalışmanın yapısı
(Structure of the study)

Eposta sözlüğünün oluşturulmasından sonra geçilen aşama eposta verisinin vektör uzayında tanımlanmasıdır. Kökleri bulunup sözlük haline getirilen tüm epostaların vektör uzayında tanımlanması üç farklı yöntem ile gerçekleştirilmiştir. Vektör oluşturmada öznitelik seçiminde "BoW-Kelime Çantası" yönteminden yararlanılmıştır.

Eğitim setinde yer alan her bir eposta içeriğinde geçen kelime kökleri öznitelikleri oluşturmaktadır. Var olan tüm epostaların içeriğine ait kelime kökleri ise sözlüğü oluşturmaktadır. Bu çalışmada 2 tür öznitelik yapısı kullanılmıştır. İlkinde sözlüğün tüm elemanları yani 4263 kelime kökü öznitelik olarak kabul edilirken, ikinci yapıda ise sık karşılaşılan yani frekans (eposta içeriklerinde görülme sıklığı fazla olan) sıklığı değeri yüksek olan 144 kelime kökü öznitelik olarak ele alınmıştır.

Şekil 3'te eposta içerisinde geçen kelime kökleri, Şekil 4'te belirlenen 144 kelime ve Şekil 5'te ise eposta içerisinde geçen kelimelerin vektör ifadesi yer almaktadır.

eğer-eğ-içerik-içer-görüntüle-görüntü-lütfen-tıkla-bu-günü-gün-kaçır-fırsat-merhaba-değer-üye-taksit-varan-var-avantaj-ve-indir-in-set-şarj-kit-kulak-ekran-koru-film-ve-şeffaf-arka-ark-kapak-oluş-ol-set-yerin-yer-sadece-sade-devam-deva-gör-indir-in-koru-deri-ve-tablet-türkiye-ücret-kargo-deri-ve-tablet-kılıf-yerin-yer-sadece-sade-devam-deva-gör-indir-fiyat-saç-çıkart-çıkar-lazer-tarak-saç-dök-son-saç-çıkart-çıkar-lazer-tarak-yerin-yer-sadece-sade-devam-deva-gör-üye-çık-için-iç-lütfen-tıkla

Şekil 3. Eposta içerisinde geçen kelime kökleri
(Base of words in email)

tl-ve-yer-bir-yerin-ol-in-indir-iç-için-il-ile-sade-sadece-gün-bu-fiyat-et-gör-al-konak-konakla-kişi-fırsat-gece-de-deva-devam-o-tur-da-siz-kahvaltı-baş-özel-tüm-yap-adres-iste-bugün-var-bak-her-kadar-tıkla-güzel-en-ya-gel-ara-geçer-değer-başla-nisan-ürün-ye-çok-ne-dahil-bakım-ver-türkiye-ay-tatil-eğit-eğitim-tarih-ek-oda-ev-alışveriş-keyif-menü-temiz-kullan-kul-saat-çocuk-bul-sağ-kargo-otel-bakı-yeni-mail-detay-lezzet-ben-daha-ekle-aç-çık-oluş-ankara-set-günlük-yemek-ücret-veya-an-kişilik-sen-yıl-gibi-kur-günü-kendi-zaman-ürü-el-boy-yen-takip-dünya-edin-yaz-hizmet-ulaş-ula-kat-kalite-açık-çal-tarihi-eş-çek-arkadaş-sev-çalış-biri-merkez-bil-sınır-muhteşem-uygula-merkezi-görüntü-bek-defter-bekle-istanbul-iş-sağla-paket

Şekil 4. Belirlenen 144 kelime kökü
(determined 144 the root words)

0-0-1-0-0-1-0-1-0-0-0-0-0-0-1-0-0-0-0-0-0-0-0-0-
1-1-0-0-0-0-1-1-0-1-1-0-0-0-0-0-0-1-1-1-0-0-0-0-
0-0-0-1-1-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-1-
0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-
0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-
0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0

Şekil 5. Epostanın vektörel ifadesi
(Vectorial expression of the email)

Binary tanımlama ile vektör oluşturmada incelenen eposta içerisinde, oluşturulan eposta sözlüğü yapısında bulunan kökleri barındırıp barındırmadığına göre yapılan bir atama, vektör oluşturma şeklidir. Bulunan her kelime kökü için 1 değeri, bulunmayan kelime kökleri için 0 değeri atanarak oluşturulur. Frekans sıklığı ile vektör oluşturulurken eposta içinde bulunan kelime köklerinin eposta içinde kaç defa geçtikleri bilgisi kullanılmaktadır.

3.2. *Epostaların sınıflandırılması (Classification of emails)*

Bu çalışmada Naive Bayes, k En Yakın Komşu ve Destek Vektör Makinesi sınıflandırıcıları kullanılmıştır. Doğruluk, Duyarlılık ve Hassasiyet değerleri Şekil 6'da gösterildiği gibi ifade edilmiştir.

		Tahmin Edilen Sınıf	
Gerçek Sınıf		Reklam	Normal
	Reklam	DP	YN
	Normal	YP	DN
	Toplam	P	N

Şekil 6. Hata matrisi
(Confusion matrix)

Performans ölçütleri aşağıda gösterildiği gibi hesaplanır:

$$\text{Doğruluk} = (DP + DN)/(P + N) \quad (4)$$
$$\text{Duyarlılık} = DP/(DP + YN) \quad (5)$$
$$\text{Hassasiyet} = DP/(DP + YP) \quad (6)$$

Burada P, pozitif, N, negatif, DP, doğru pozitif, DN, doğru negatif, YP, yanlış pozitif ve YN, yanlış negatif olarak adlandırılmıştır.

4. DENEYSEL ÇALIŞMA (EMPIRICAL STUDY)

Deneysel çalışmada vektör oluşturma yöntemlerinin, sınıflandırma yöntemlerinin ve nitelik seçimini yöntemlerinin 18 farklı kombinasyonunda ortaya çıkan sonuçlar karşılaştırılmıştır. Yapılan deneysel çalışmalar Çizelge 1'de gösterildiği gibidir.

Çizelge 1. Deneysel çalışmada kullanılan yöntem kombinasyonları
(Method combinations used in empirical study)

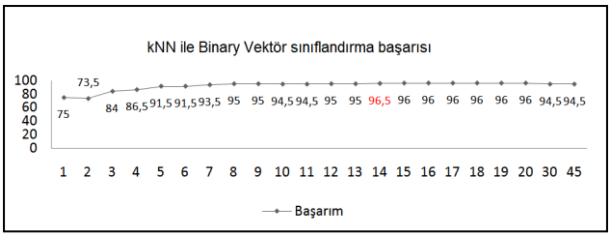
	Binary V.	Frekans V.	TF-IDF V.
144 Nitelikli	DVM	DVM	DVM
	kNN	kNN	kNN
	NB	NB	NB
4263 Nitelikli	DVM	DVM	DVM
	kNN	kNN	kNN
	-	-	-

Oluşturulan eposta veri kümesinin kendi içinde dağılımı şu şekilde gerçekleştirilmiştir: eposta kümesinin 600'ü öğrenme, geriye kalan 200'ü ise test için kullanılmıştır. Öğrenme kümesi içinde verilerin 300 tanesini normal epostalardan alınan bilgiler, geriye kalan 300 tanesi ise reklam epostalardan alınan verilerden oluşturulmuştur. Çizelge 2'de kullanılan eposta veri setinin dağılımı gösterilmiştir.

Çizelge 2. Eposta veri setinin dağılımı
(The distribution of dataset)

	Öğrenme Kümesi	Test Kümesi
Reklam Eposta	300	100
Normal Eposta	300	100
Toplam	600	200

Şekil 7'de farklı k değerleri için kNN ile Binary vektör sınıflandırma başarısı grafiği görülmektedir.



Şekil 7. kNN ile Binary vektör sınıflandırma başarısı grafiği
(The graphic of classification success with using Knn and binary vector)

Şekil 7'de de görüldüğü gibi en yüksek sınıflandırma başarısı %96,5 ile k değeri 14 iken elde edilmiştir.

Çizelge 3. k=14 için hata matrisi
(Confusion matrix for k=14)

		Tahmin Edilen Sınıf	
		Reklam	Normal
Gerçek Sınıf	Reklam	94	6
	Normal	1	99
	Toplam	95	105

Çizelge 3'te k 14 için hata matrisi görülmektedir. Çizelge 4'te yapılan deneylerin sonuçlarına göre kNN yönteminin maksimum hassasiyet değerleri verilmiştir.

kNN için farklı k değerlerine göre 4263 Nitelik- Binary Vektör üzerinden sınıflandırılmasında k=14, 4263 Nitelik- Frekans Vektör üzerinden sınıflandırılmasında k=45, 4263 Nitelik - TF-IDF Vektör üzerinden sınıflandırılmasında k=15 değerlerinde en yüksek hassasiyet elde edilmiştir.

Çizelge 4. Elde edilen maksimum hassasiyet değerleri
(The peak sensitivity values)

		4263 Nitelikli	144 Nitelikli
Binary Vektör	DVM	1	0,59
	kNN	0,99	0,99
	NB	-	0,89
Frekans Vektör	DVM	1	0,58
	kNN	1	0,98
	NB	-	0,95
TF-IDF Vektör	DVM	1	0,53
	kNN	0,96	0,95
	NB	-	0,90

Kelime frekanslarına göre 144 Nitelik Binary Vektör üzerinden sınıflandırılmasında k=45, 144 Nitelik Frekans Vektör sınıflandırılmasında k=20, 144 Nitelik TF-IDF Vektör sınıflandırılmasında k=45 değerlerinde en yüksek hassasiyet elde edilmiştir.

Çizelge 5'te 4263 nitelik için elde edilen deneysel bulgular, Çizelge 6'da 144 nitelik için elde edilen deneysel bulgular verilmiştir.

Çizelge 5. 4263 kelime için sınıflandırma başarı oranları
(Success rate of classification for 4263 words vector)

%		Binary	Frekans	TF-IDF
NB		X	X	X
KNN	k=1	68	75	64,5
	k=2	67	73,5	64,5
	k=3	76	84	86,5
	k=4	81,5	86,5	87
	k=5	89	91,5	83
	k=6	90,5	91,5	83,5
	k=7	93	93,5	76
	k=8	93,5	95	80,5
	k=9	94	94	76,5
	k=10	95	94,5	77
	k=11	96	94,5	73
	k=12	96	95	73
	k=13	96	95	71
	k=14	**96,5**	95,5	71
	k=15	95,5	96	70
	k=16	95,5	96	70
	k=17	94,5	96	66,5
	k=18	94,5	96	67
	k=19	94	96	62,5
	k=20	94	96	68
	k=30	93,5	94,5	54,5
	k=45	89	94,5	51,5

Çizelge 5 ve Çizelge 6'da da görüldüğü gibi binary vektör kullanımı diğer vektör tanımlama yöntemlerine göre daha başarılı sonuçlar vermiştir. Ayrıca ortalama olarak bakıldığında Naive Bayes ve K-En Yakın Komşu Yönteminin büyük ölçüde başarı sağladığı gözlemlenmiştir. Tüm bu deneyler sonucu en yüksek başarıma 4263 kelime için kNN algoritması, k=14 iken binary vektör ile %96,5 oranıyla ulaşılmıştır.

Çizelge 6. 144 kelime için sınıflandırma başarı oranları
(Success rate of classification for 144 words vector)

		Binary	Frekans	TF-IDF
NB		91	95,5	94
KNN	k=1	83,5	84	91
	k=2	83	84	91
	k=3	80	80	88
	k=4	84	83,5	88,5
	k=5	89,5	89,5	89
	k=6	91,5	89,5	91,5
	k=7	91	91,5	91,5
	k=8	92	90,5	92,5
	k=9	92	91	91
	k=10	92	90,5	92
	k=11	92	92	91
	k=12	93	92,5	93
	k=13	93,5	92	92,5
	k=14	95	93,5	92,5
	k=15	95	93,5	92,5
	k=16	95,5	94	93,5
	k=17	95	93,5	93,5
	k=18	95	94	93,5
	k=19	95,5	94	93
	k=20	92	95	94
	k=30	95	93,5	91
	k=45	91	92,5	92
DVM		60	61	53

5. SONUÇ VE DEĞERLENDİRME (CONCLUSION AND DISCUSSION)

Yaşanan teknolojik gelişmeler ve Internet kullanımının yaygınlaşması ile eposta kullanıcı sayısı her geçen gün daha da artmaktadır. Kolaylığı ve basit erişilebilir olması sebebiyle her gün milyarlarca eposta sunuculardan gönderilip alınmaktadır.

Eposta, günlük hayatımızda haberleşmede önemli bir yere sahiptir. Ancak bununla birlikte spam ve reklam içerikli epostalar da büyük bir problem olarak karşımıza çıkmaktadır. Spam ve reklam epostaları hem internet kullanıcılarının vaktini çalmakta, hem de internet kaynaklarının boşa harcanmasına neden olmaktadır.

Yapılan bu çalışmada, Türkçe içerikli reklam epostalarının metin madenciliği yöntemleri ile otomatik tespiti gerçekleştirilmiştir. Bu amaçla toplanan 800 eposta üç ayrı sınıflandırma algoritması ile sınıflandırılmıştır. Yapılan deneysel çalışmalarda görülmüştür ki en yüksek sınıflandırma başarısı Naive Bayes ve kNN ile elde edilirken, binary vektör diğer yöntemlere göre daha yüksek sınıflandırma başarısı sağlamaktadır.

KAYNAKLAR (REFERENCES)

[1] M. Sasaki, H.Shinnou, **Spam Detection Using Text Clustering** *International Conference on CYBERWORLDS*, Singapore, 2005.

[2] J. Clark, I. Koprinska ve J. Poon, **A Neural Network Based Approach To Automated E-mail Classification**, *Web Intelligence IEEE/WIC International Conference*, 2003.

[3] G. Sakkis, I. Androutsopoulos, V. Karkaletsis, C. Spyropoulos ve P. Stamatopoulos, **Stacking Classifiers For Anti-Spam Filtering of E-mail**, *Empirical Methods In Natural Language Processing-EMNLP*, 2001, pp 44–50.

[4] M. Sahami, S. Dumais, D. Heckerman ve E. Horvitz E, **A Bayesian Approach To Filtering Junk E-mail**, *AAAI Technical Report WS-98-05*, 1998.

[5] W. Cohen, **Learning Rules that Classify E-mail**, *AAAI Spring Symposium on Machine Learning in Information Access MLIA '96,* 1996

[6] Huai-bin Wang, Ying Yu, ve Zhen Liu, **SVM Classifier Incorporating Feature Selection Using GA for Spam Detection**, *Embedded And Ubiquitous Computing – EUC 2005 Lecture Notes in Computer Science*, 2005, Volume 3824/2005.

[7] L. Kun-Lun, L. Kai, H. Hou-Kuan, T. Sheng-Feng **Active Learning With Simplified SVMs for Spam Categorization**, *Machine Learning and Cybernetics International Conference*, 2002.

[8] C-C Lai ve M-C Tsai, **An Empirical Performance Comparison of Machine Learning Methods for Spam E-mail Categorization**, *Fourth International Conference on Hybrid Intelligent Systems*, Japan, 2004.

[9] P. Pantel ve D. Lin, **SpamCop: A Spam Classification & Organization Program**, *AAAI Technical Report WS-98-05*, 1998.

[10] V. Metsis, I. Androutsopoulos, G. Paliouras, **Spam Filtering with Naive Bayes** – Which Naive Bayes?", 2006

[11] I. Androutsopoulos, G. Paliouras, V. Karkaletsis, G. Sakkis, C. Spyropoulos ve P. Stamatopoulos, **Learning to Filter Spam E-mail: A Comparison of A Naive Bayesian And A Memory-Based Approach**, *4th European Conference On Principles And Practice Of Knowledge Discovery İn Databases- PKDD*, France, 2000.

[12] E. Blanzieri ve A. Bryl, **Evaluation of The Highest Probability SVM Nearest Neighbor Classifier With Variable Relative Error Cost**, *Fourth Conference On E-mail And Anti-Spam, CEAS'2007*, 2007.

[13] R. McCue, **A Comparison of the Accuracy of Support Vector Machine and Naive Bayes Algorithms In Spam Classification**, *University of California at Santa Cruz*, 2009.

[14] U.M. Fayyad, G. Piatetsky-Shapiro ve P. Smyth, **From Data Mining to Knowledge Discovery: An Overview**, *AAAI Press/ MIT Press*, Cambridge, 1996.

[15] A-H. Tan, **Text Mining: The State of The Art and The Challenges**, *PAKDD'99 workshop on Knowledge Discovery from Advanced Databases*, 1999.

[16] J. Han, M. Kamber, **Data Mining: Concepts and Techniques**, Second Edition, *Morgan Kaufmann Publisher*, San Francisco, 2006.

[17] E. Alpaydın, **Yapay Öğrenme**, *Boğaziçi Üniversitesi Yayınevi*, 2011.

[18] G. F. COOPER, **A Bayesian Method for the Induction of Probabilistic Networks from Data**, *Kluwer Academic Publishers*, Boston Manufactured in the Netherlands, 1992.

[19] D. Wettschereck, W. Aha, T. Mohri**, A Review and Comparative Evaluation of Feature Weighting Methods for Lazy Learning Algorithms,** *Technical Report AIC-95-012 Naval Research Laboratory*, Washington-USA, 1995.

[20] H. Drucker, D. Wu, V. Vapnik, **Support Vector Machines For Spam Categorization,** *IEEE Transaction On Neural Networks*, 10(5): 1048–1054,1999.

[21] E. Blanzieri ve A. Bryl, **A Survey Of Learning-Based Techniques of E-mail Spam Filtering**, *Artificial Intelligence Review*, 2008.