Kurumsal Kolektif Süreçler için E-Posta İletilerinden Görev Keşfi ve Gerçek Zamanlı Görev Yönetim Sisteminin Geliştirilmesi

Halil ARSLAN[1]*, Oğuz KAYNAR[2], Ahmet Gürkan YÜKSEK[1]

[1]Bilgisayar Mühendisliği, Cumhuriyet Üniversitesi, Sivas, Türkiye
[2]Yönetim Bilişim Sistemleri, Cumhuriyet Üniversitesi, Sivas, Türkiye
harslan@cumhuriyet.edu.tr, okaynar@cumhuriyet.edu.tr, agyuksek@cumhuriyet.edu.tr
(Geliş/Received:27.01.2017; Kabul/Accepted:03.09.2017)
DOI: 10.17671/gazibtd.281713

Özet— E-posta sistemleri, kurumsal iletişim ve işbirliği için kullanılan en yaygın araçlardan birisidir. Başta planlama, kaynak veya proje yönetimi olmak üzere neredeyse tüm kurumsal işlemler e-posta üzerinden gerçekleştirilmektedir. Bu nedenle e-posta sistemleri, şirketler için değerli bilgiler içeren hizmet depoları ve işlerin yönetildiği vazgeçilmez araçlardan biri haline gelmiştir. Özellikle müşteri odaklı kuruluşlar için e-posta servisleri üzerinden iş akışlarının yönetilebilir olması son derece önemlidir. Kurumsal sistemler ile müşteri düzeyindeki iş listelerinin her ne kadar kullanılan yazılımlar üzerinden iletilme imkanı olsa da bu talepler çoğunlukla e-postalar ile iletilmektedir. Ancak bu durum çalışanın e-posta kutusundaki işleri planlamaması, unutması, kaybolması, önem düzeyini doğru belirleyememesi gibi sonuçlara yol açabilmektedir. Yapılan bu çalışma ile personelin kurumsal e-posta hesaplarına gelen mesajlar incelenerek, metin madenciliği ve sınıflama teknikleri yardımıyla iş taleplerini etiketleyen ve etiketlenen bu mesajları geliştirilen yapılacaklar listesi (todo list) uygulamasına girdi olarak gönderen bir yöntem önerilmektedir. Önerilen yöntem ve geliştirilen uygulamanın, genişletilebilir mesajlaşma ve durum protokolü üzerine entegre edilmesiyle gerçek zamanlı işbirlikçi çalışma olanağı sunulmaktadır. Böylece kurumsal e-postaların iş listelerine dönüştürülme süreci, kurumsal anında mesajlaşma sistemleri üzerinden gerçek zamanlı durum yönetim yaklaşımı ile işbirlikçi çalışmaya uygun hale getirilmiştir.

Anahtar Kelimeler— İşbirlikçi süreçler, todo list yönetimi, e-posta analizi, metin madenciliği

Task Exploration from E-mail Messages for Corporate Collaborative Processes and Development of a Real-Time Task Management System

Abstract— Email systems are one of the most common tools used for corporate communication and collaboration. Almost all the corporate tasks, especially planning, resource or project management are carried out via e-mail systems. For this reason, email systems have become service stores containing valuable information for companies, and very important tools where the tasks are managed. It is particularly important for customer-focused organizations to manage workflows via e-mail services. Although it is possible to create work lists on customer level via enterprise systems over central software, these requests are mostly transferred by employees' e-mail accounts. This case may cause such results as the employee's not planning the tasks in his email address, forgetting, losing, not defining its level of importance. In this study, a method is proposed that labels the requests to be managed coming from the corporate e-mail account by using text mining and classification techniques and provides input for the to-do list application developed by examining the massages in corporate e-mail accounts of employees. The proposed method and the developed application offer real-time collaborative working environment with the extensible messaging and its integration with the status protocol. Thus, the conversion process of corporate e-mails to work lists has been adapted to collaborative work though the real-time status management approaches over the corporate instant messaging systems.

Keywords— Collaborative processes, ToDo list management, e-mail analysis, text mining

1. GİRİŞ (INTRODUCTION)

E-posta sistemleri, bilgisayar tabanlı iletişim açısından geliştirilmiş en önemli ve popüler yazılım sistemleri olarak gösterilebilir [1-6]. Günümüzde gerek kişisel, gerekse de iş yaşamında e-posta yoluyla işlerimizi organize etmekteyiz. Ancak profesyonel iş ortamında çoğu e-posta mesajları, ortak görevler ve işbirlikçi çalışmalar gerektiren durumlar ortaya çıkarmaktadır [2]. Şirketlerde iş akışlarının e-posta iletileri üzerinden yapılıyor oluşu, e-posta sistemlerinin kişinin gelen kutusundan ibaret olmaması gerektiğini göstermektedir[4]. Bu noktada, iş süreçlerinin yönetimi ve kontrol edilmesi ile bu süreçlerin iyileştirilmesi ancak etkili bir e-posta yönetim sistemi ile sağlanabilecektir. Ancak pek çok şirket, mevcut iş süreçlerine, e-posta sistemlerini nesnel anlamda konumlandıramamaktadır. Özellikle planlama, kaynak yönetimi veya proje yönetimi gibi temel iş akışları ile ilgili neredeyse tüm kurumsal görevler için bilgi akışının e-posta iletileri üzerinden gerçekleşmesi, e-postaların değerli bilgiler içeren veri depoları olarak işlenmesini gerektirmektedir[2,5].

E-posta sistemleri ve iletiler, kuruluşlar için işlerin alındığı ve yönetildiği çok önemli bir merkezi alan haline gelmiştir [3]. Çalışanlara, iş süreçleri kapsamında, özellikle müşteri düzeyinde bir e-posta iletisi geldiğinde, kullanıcı, bu iletinin gereksinimini ve takibini, gelen kutusunun özerkliği içerisinde gerçekleştirmektedir. Bu durum, çalışanın e-postasının gelen kutusunda bekleyen işleri doğru planlayamaması, unutması, kaybolması, önem düzeyini doğru belirleyememesi ve işbirliği gerektiren etkileşimli bir iş süreci gereksiniminin sağlanamaması gibi durumları ortaya çıkarmaktadır. Ayrıca, çalışanların e-posta istemcisi, planlama, todo list yönetimi vb. iş süreçlerinin yürütülmesi için ihtiyaç duyduğu yazılımlar, çoğunlukla bağımsız altyapılar üzerinde koşmaktadır. Bu yazılımların birbirleri ile haberleşememeleri, gelen kutusundaki bir iletinin planlama ve todo list yönetimi gibi işbirlikçi altyapılar sunan yazılımlara manuel olarak taşıması gibi ekstra iş yüklerini doğurmaktadır. Bunlara ek olarak e-posta sistemi üzerindeki bir iş, doğru sonuçlandırılsa dahi geri dönüşü, kişinin inisiyatifine kalmaktadır. Ayrıca e-posta iletisi sonucu, todo list ve planlama gibi sistemlere aktarılmadan sonuçlandırılan iş, unutulan hizmet faturalamaları ve tahsilat sorunları gibi maliyet sorunlarını da beraberinde getirmektedir. Tüm bu durumlar müşteri düzeyinde ya da personeller arası yoğun e-posta trafiğinin yaşandığı şirketler açısından iş süreçlerine ve kurumsallığa uygun olmayan süreç dışı sonuçlar ortaya çıkarmaktadır [7].

Şirketler, çalışanlarının iş listelerini, todo list olarak isimlendirilen yazılımlar üzerinden takip etmektedirler. Bu yazılımlar, çoğunlukla merkezi/bulut veritabanları üzerinden kişilerin iş listelerini görüntüleyebildikleri ve yönetebildikleri çeşitli fonksiyonlar ve arayüzler sunmaktadır. Todo list yazılımları, çalışanlar için günlük yerine getirmeleri ve kontrol etmeleri gereken rutin bir iş faaliyetleri olmak zorundadırlar. Aksi takdirde bu yazılımlar sadece proje yöneticilerinin raporlama amacıyla kullandıkları genel taslaktan öteye geçememektedirler. Bu çalışma ile todo list uygulaması, XMPP [8] protokolü üzerinden kurum içi anında mesajlaşma ve iş yönetim uygulaması olarak işbirlikçi çalışmaya uygun bir yapı teşkil edecek şekilde sunulmaktadır. Todo list gibi iş süreçleri için işbirliği gerektiren uygulamalar, erişilebilirlik ve kullanılabilirlik noktasında, kullanıcıya rutin faaliyet izlenimi vermeyen, memnuniyet düzeyi yüksek deneyimler sunmalıdır. Önerilen sistem, metin madenciliği teknikleri yardımıyla, gelen e-postaları analiz ederek, bunlardan iş ile ilgili olan e-postaları, yapılacaklar listesine dönüştürerek, todo list yazılımları ile entegrasyon sağlamakta ve yukarıda bahsedilen sorunların üstesinden gelmektedir.

Çalışma şu şekilde organize edilmiştir. Birinci bölümünde, literatürde daha önce benzer alanlarda yapılmış çalışmalar değerlendirilmekte, ikinci bölümde, çalışmada kullanılan metodoloji ve yöntemler sunulmakta, üçüncü bölümde, kurumsal e-posta sistemleri üzerinden elde edilen veri setleri, metin madenciliği ve makine öğrenmesi teknikleri kullanılarak sınıflandırılmakta ve geliştirilen işbirlikçi uygulama modeli sunulmakta, son bölümde elde edilen sonuçlar değerlendirilmektedir.

2. İLGİLİ ÇALIŞMALAR (RELATED WORK)

Çalışmanın temel araştırma alanları olarak, (1) E-Posta iletilerinin, metin madenciliği teknikleri ile sınıflandırılabilmesi için gerçek ortamda eğitim ve test veri setlerinin hazırlanması, (2) E-Posta iletilerinin metin madenciliği ve makine öğrenmesi teknikleri kullanılarak sınıflandırılması ve iş süreçlerine dönüştürülmesi, (3) Todo list uygulaması için gerçek zamanlı ve işbirlikçi çalışmaya uygun uygulama yazılımının geliştirilmesi, şeklinde sıralanabilir. Bu bağlamda yapılan literatür taramasında, e-posta iletilerinin makine öğrenmesi teknikleri ile analiz edildiği çalışmalara rastlanılmaktadır [6]. Ancak çalışmaların pek çoğu spam ileti tespiti ve e-posta doğruluğu üzerine odaklanmaktadır [9]. Son zamanlarda özellikle e-posta iletilerinin kurumsal uygulamalar için girdi teşkil edebilecek şekilde kullanıldığı çalışmalar ve değerlendirmeler giderek artmaktadır [10]. Stuit vd.[2] çalışmalarında iş süreçleri için e-posta iletilerinin analizi ve bilgi keşfine yönelik çalışamalar yapmışlar ve e-posta etkileşimli veri madenciliği yöntemleri önermişlerdir. Önerilen yöntemle analiz ettikleri e-posta odaklı iş süreçlerini TALL adı verilen bir modelleme dili ile görselleştirmişlerdir. Soares vd.[11] bilgi yönetimi odaklı iş süreçleri için e-postalar üzerine odaklanarak, bilginin yarı otomatik keşfi için yöntem tarif etmektedirler. Bu yöntemde, e-posta iletileri ile işbirlikçi süreçler üzerinden mevcut bilginin alışverişini tanımlarken, iletilerin ilişkilerini tanımlayan kavramsal bir harita ortaya koymuşlardır. Çalışma sonucu ortaya konan bulguların bir organizasyon yapısının oluşturulması ve değerlendirilmesi için kullanılabileceğini

ifade etmişlerdir. Dey vd.[4] metin madenciliği, ağ analizi ve veri analitiği teknikleri kullanarak yaptıkları etkinlik yönetimi ve bilgi keşfi için e-posta analizi çalışmalarında, içeriğe odaklı akıllı gruplama, e-posta kümelerinde zamansal analiz ve metin analizleri yapmışlardır. Méndez vd.[12] açık kaynak kodlu anti-spam projesi olan SpamAssassin için e-posta iletilerinin sınıflandırılmasında performansı artırmaya yönelik bir optimizasyon aracı geliştirmişlerdir. Yaptıkları çalışmada bilinen dört farklı veri madenciliği tekniği (Naïve Bayes, Flexible Bayes, Adaboost ve Support Vector Machines) ile kıyaslamışlar ve önerdikleri aracın bu teknikleri maliyet ve performans olarak geride bıraktığını ifade etmişlerdir. Koprinska vd.[13] e-posta iletilerinin sınıflandırılmasında kullanılan denetimli ve yarı-denetimli öğrenme tekniklerini ele almışlardır. Rastgele Orman, Karar Ağaçları, Destek Vektör Mekanizmaları ve Naïve Bayes gibi popüler algoritmaları kullandıkları çalışmalarında e-posta sınıflandırması için Rastgele Orman algoritmasının daha verimli olduğunu ifade etmişlerdir. Pankaj vd.[14] çalışmalarında iş uygulamalarının teorik temellerini ortaya koymuşlar, uygun iş modelleri önermişler ve bu uygulamaların eşler arası iletişim bağlamında uygun altyapılar sunabileceğini vurgulamışlardır. Ragavan vd.[15] XMPP protokolü kullanarak çeşitli endüstriyel uygulamaları kontrol eden gerçek zamanlı veri toplama yeteneğine sahip bir uygulama geliştirmişlerdir. Bu çalışmalarında XMPP protokolünün gerçek zamanlı iş uygulamaları için önemli bir altyapı sunabileceğini göstermişlerdir.

E-posta iletilerinin sınıflandırılması üzerine yapılan çalışmaların, gerek akademik gerekse de ticarileşme potansiyeli açısından önemli bir araştırma alanı olmayı sürdüreceği görülmektedir. Ancak burada ifade edilmeyen fakat literatürde sıkça karşılaşılan çalışmaların pek çoğu, anti-spam üzerine odaklansa da bu noktada kullanılan teknikler e-posta iletileri üzerinden başkaca sınıflama ve analiz gerektiren problemlere yeni ufuklar açmaktadır [16]. Ancak literatür kapsamında bakılan çalışmalarda bu tür analiz teknikleri sonucunda ortaya konulan ve iş uygulamaları için işbirlikçi bir uygulama çalışmasına rastlanmamıştır.

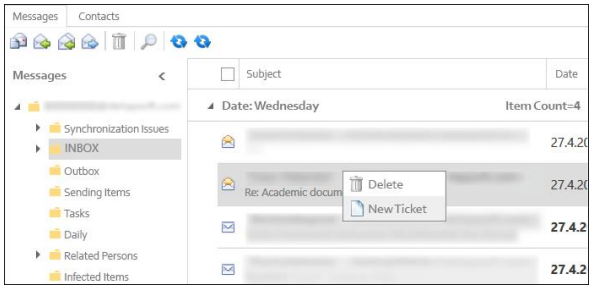
3. METODOLOJİ VE YÖNTEM (METHODOLOGY AND METHOD)

Kullanıcıların gelen kutusundaki iletilerden görev keşfi süreci iki aşamada gerçekleşmektedir. Birinci aşamada gelen kutusundaki iletilerin manuel olarak iş listesine dönüştürülebilmesi sağlanmaktadır. Bu aşamada, kullanıcıların tanımladıkları iş listeleri ikinci aşamada kullanılacak metin madenciliği ve makine öğrenmesi teknikleri için veri seti olarak kullanılmaktadır.

3.1. Veri Setlerinin Oluşturulması (Collect Data Sets)

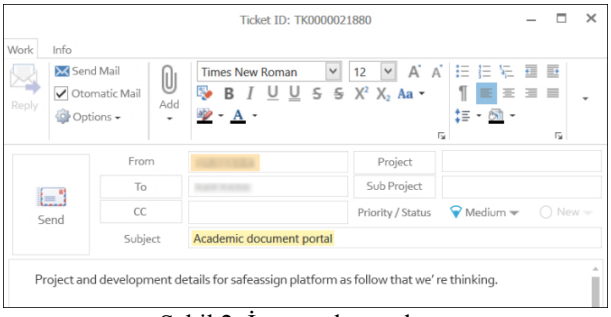
Veri setlerinin oluşturulabilmesi için ileti alma, gönderme, yanıtlama gibi temel e-posta fonksiyonlarını karşılayan ve gelen kutusundaki bir iletiyi iş listesine aktarabilen bir e-posta istemcisi geliştirilmiştir (Şekil 1).

Geliştirilen yazılım ile kullanıcı, kurumsal e-posta hareketlerini gerçekleştirebilirken, istediği e-posta iletisini iş listesine dönüştürebilmektedir. Ayrıca kurumsal kullanıcılarda yaygın olarak kullanılan Microsoft Outlook yazılımı için geliştirilen bir eklenti ile kullanıcının e-posta iletilerini iş listelerine aktarımı sağlanmıştır. Hazırlanan bu eklenti, bir web servis aracılığı ile istenilen işlevi yerine getirmektedir. Böylelikle geliştirilen e-posta istemcisi haricinde Outlook gelen kutusunda yer alan işle ilgili e-postaların da veri setine aktarımı gerçekleştirilmiştir. Bu sayede, e-posta iletilerinin sınıflandırılmasında kullanılan metin madenciliği ve makine öğrenmesi teknikleri için gerçek veri setleri toplanmıştır.



Şekil 1. E-Posta istemci ekranı ve iş listesine aktarımı
(E-Mail client screen and transfer of messages to the ToDo list)

Şekil 1'de sunulan ekran üzerinden kullanıcı gelen kutusundaki iletiyi, todo list uygulamasına gönderebilmektedir. Bu fonksiyon, kullanıcının istediği bir iletiyi hızlıca iş listesine dönüştürmesine olanak sağlayarak uygulamalar arası işbirliğini kolaylaştırmaktadır. Şekil 2'de ise e-posta istemcisi üzerinden todo list uygulamasına aktarılan bir iletinin yeni iş tanımlama ekranı görülmektedir.



Şekil 2. İş tanımlama ekranı
(Task definition screen)

3.2. Ön İşlem (Pre-processing)

E-posta istemcisi üzerinden tanımlanan işlerle, işe dönüştürülen iletiler ilişkilendirilmiş ve "ToDo" sınıfına ait ileti olarak ayrılmıştır. Bu yöntemle elde edilen 175 e-posta iletisi "ToDo" sınıfında, diğer 200 ileti ise "Normal" sınıfında yer almaktadır. Veri setlerinin eğitim

ve sınıflandırma süreçlerine hazırlanması amacıyla çeşitli veri ön işleme adımlarından geçirilmesi gerekmektedir. Ön işlem aşamaları aşağıdaki şekilde özetlenebilir.

- HTML ve XML etiketlerin temizlenmesi,
- İletideki konu ve içerik kısmını oluşturan bölümlerin düz metin haline dönüştürülmesi,
- İletideki simge ve noktalama işaretlerinin temizlenmesi ve karakterlerin küçük harfe çevrimi,
- Her bir kelimenin köklerinin bulunması ve terim listelerinin oluşturulması,
- Metin içerindeki edat, bağlaç ve zamirlerden oluşan durak kelimelerin kaldırılması
- Uzunluğu 3 harften kısa olan kelimelerin temizlenmesi,
- Terim frekansları ve ters doküman frekansları yardımıyla vektör uzay modelinin oluşturulması.

3.3. Sınıflandırma Yöntemleri (Classification Methods)

Çalışmada, e-posta iletilerinin sınıflandırılabilmesi için Merkez Tabanlı Sınıflayıcı, Çok Katmanlı Yapay Sinir Ağları ve Destek Vektör Makineleri kullanılmıştır.

a. Merkez tabanlı sınıflayıcı (Centroid-based classifier)

Merkez tabanlı sınıflayıcı, vektör uzay modeli tabanlı olup, oldukça basit ve performansı yüksek bir algoritmadır [17]. Algoritmada her bir doküman terim uzayında bir vektör olarak ele alınır. Vektörün her bir boyutu, dokümanda geçen bir terimin ters doküman frekansıyla ağırlıklandırılmış frekansını tutar. N toplam doküman sayısını gösterirken, d_{fi}, i teriminin geçtiği doküman sayısını göstermektedir. Böylelikle dokümanlarda sık geçen kelimeler için daha küçük bir ağırlık değeri elde edilirken daha seyrek geçen kelimeler için daha büyük bir ağırlık değeri elde edilir. Terim frekansı ve ters doküman frekansı çarpılarak ağırlıklandırılmış TF-IDF değeri elde edilir (Eşitlik 1).

$$d_{tf\_idf} = \left(tf_1 \log\left(\frac{N}{d_{f1}}\right), tf_2 \log\left(\frac{N}{d_{f2}}\right), \dots tf_n \log\left(\frac{N}{d_{fn}}\right)\right) \quad (1)$$

Vektör uzay modelinde d_i ve d_j dokümanlarının benzerliği, Eşitlik 2'deki kosinüs fonksiyonu yardımıyla ölçülür.

$$\cos(d_i, d_j) = \frac{d_i . d_j}{\|d_i\| * \|d_j\|} \quad (2)$$

"." vektörlerin skaler çarpımını gösterir. Merkez tabanlı sınıflandırmada, sınıflar merkez adı verilen vektörlerle sunulur. Merkez, sınıf elemanlarını temsil eden ortalama bir değerdir ve bu orta değerin bütün sınıfı temsil ettiği kabul edilir. Eğitim seti k farklı sınıf içeriyorsa bu eğitim verilerinden k adet merkez vektörü elde edilir.

$$C_k = \frac{1}{|S|} \sum_{d \in S} d \quad (3)$$

Eşitlik 3'te, S değeri, ilgili sınıftaki dokümanların kümesini göstermektedir. İlgili sınıftaki ortalama vektör o sınıftaki vektörlerin skaler toplamının, sınıftaki doküman sayısına bölümüyle elde edilir. Merkez tabanlı sınıflayıcının çalışma mantığının arkasında, vektörlerin benzerlik ilkesi yatar. Bir dokümanın hangi sınıfta olduğuna karar vermek için, ilgili dokümana ait vektörün merkez vektörlerin her biriyle kosinüs benzerliğine bakılır. Test vektörü daha çok hangi merkez vektöre benziyorsa ilgili vektörün o sınıfta olduğuna karar verilir.

b. Çok katmanlı yapay sinir ağları (Multilayer artificial neural networks - MLP)

İnsan sinir sistemi örnek alınarak tasarlanan yapay sinir ağları (YSA), genelleme yapabilme, veriden öğrenebilme, sınırsız sayıda değişkenle işlem yapabilme gibi farklı özeliklere sahip denetimli bir makine öğrenmesi yöntemidir [18]. En küçük YSA birimine nöron adı verilmektedir. Nöronlar birleşerek katmanları, katmanlar ise bir ağı meydana getirmektedir. Bir YSA girdi katmanı ve çıktı katmanı olmak üzere en az iki katman bulundurmak zorundadır. Girdi katmanı, çözülmesi istenilen probleme ait verilerin ağ tarafından okunmasını sağlayan katmandır ve öznitelik sayısı kadar nöron içermektedir. Çıktı katmanı ise ağ tarafından işlenerek üretilen verinin dışarıya aktarıldığı katmandır. Bu katman tek bir nöron içerebildiği gibi tahmin edilecek probleme ait sınıf sayısı kadar da nöron içerebilmektedir. MLP ise bu iki katman arasında bulunan, bir ya da daha fazla sayıda gizli katman içermektedir. Bu katmanlardaki nöron sayısı tam olarak belli olmamakla birlikte ağın performansını etkileyen önemli parametrelerden biridir. YSA ilgili örnekler yardımı ile bu katmanları eğiterek genelleme yapmayı hedeflemektedir [19]. YSA'da eğitim işlemi ağın sahip olduğu ağırlıkların, seçilen eğitim algoritması yardımı ile güncellenmesi ile yapılmaktadır. MLP ağlarının eğitiminde, kolay anlaşılabilir ve matematiksel olarak ispatlanabilir olmasından dolayı geri yayılım algoritması kullanılmaktadır. Bu algoritma eşitlik 4 kullanılarak hesaplanan hata değerini en aza indirecek şekilde ağırlık değerlerini güncelleyerek ağı eğitmektedir [20].

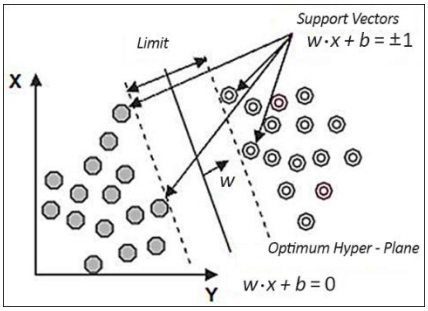
$$\mathrm{E} = \frac{1}{2} \sum_{k=1}^{m} (y_k - t_k)^2 \quad (4)$$

Eşitlikte y_k ağ tarafından üretilen sonucu, t_k gerçek sınıf değerini, m ise toplam örnek sayısını temsil etmektedir. Hatayı en aza indirgemek için bağlantı ağırlıkları yeniden düzenlenerek güncellenir. Böylece sınıf değerlerinin en az hata ile tahmin edilmesi amaçlanır.

c. Destek vektör makineleri (Support vector machine)

Destek vektör makineleri (DVM), istatistiksel öğrenme teorisi ve yapısal riski en aza indirme prensibinden

faydalanan, sınıflandırma ve eğri uydurma problemlerinin çözümü amacıyla geliştirilmiş bir öğrenme yöntemidir [21]. Lineer olarak ayrıştırılabilen sınıfların belirlenmesinde sıkça kullanılan yöntem, kernel fonksiyonları sayesinde doğrusal olarak ayrıştırılamayan girdi uzayını daha yüksek boyutlu lineer olarak ayrıştırılabilen bu uzaya taşıyarak, doğrusal olmayan verilerin sınıflandırılmasında başarıyla kullanılmaktadır. Eğitim için kullanılacak N elemandan oluşan verinin $\theta = \{x_i, y_i\}$, i = 1, 2, N olduğu varsayılırsa. x_i özellik vektörünü, $y_i \in \{-1,1\}$ ise sınıf değerlerini gösterir. Lineer olarak ayrılma durumunda, bu iki değerli veriler direkt olarak bir ayırıcı düzlem ile ayrılabilir. Veri setini sınıflara ayırabilecek sonsuz sayıda çoklu düzlem çizilebilmesine karşın, amaç, bilinmeyen sınıflama hatasını en küçük yapacak hiper düzlemi seçmektir. Şekil 3'te görüleceği üzere $f(\vec{x}) = \vec{w}^T\vec{x} + b \geq 1$ durumu birinci sınıfı, $(y_i = 1)$ ve $f(\vec{x}) = \vec{w}^T\vec{x} + b \leq -1$ durumu ise ikinci sınıfı $(y_i = -1)$ temsil eder.



Şekil 3. Destek Vektör Makineleri ve Hiper Düzlem Seçimi
(Support vector machine and selection of the hyper-plane)

İki sınır arasındaki uzaklık $\lambda = 2/\vec{w}^2$ formülü ile ifade edilir. Amaç, λ değerini maksimum yapmak olduğu için $1/\lambda$ ifadesi minimum olmalıdır. Buna bağlı sınırlama ise $y_i(w^T x_i + b) - 1 \geq 0, \quad y_i \in \{-1, +1\}$'dir. İlgili problemin duali, Eşitlik 5'te verilmiştir. Eşitlikteki problem, Lagrange denklemleri, Eşitlik 6 ve Eşitlik 7'de verilen "Karush–Kuhn–Tucker (KKT)" ın kısıtları yardımıyla çözülür.

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \sum_{i=1}^{N} \alpha_i \left[y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) - 1\right], \quad \alpha_i \geq 0, \forall i \tag{5}$$

$$\frac{\partial L}{\partial w_j} = 0, \forall j \tag{6}$$

$$\frac{\partial L}{\partial w_0} = 0 \tag{7}$$

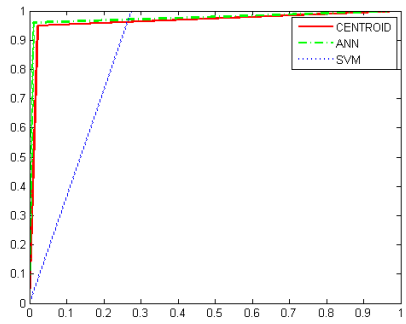
Verilerin doğrusal olarak ayrılamadığı durumlarda, girdi uzayı kernel fonksiyonları yardımıyla daha yüksek boyutlu bir uzaya taşınarak doğrusal olarak ayrıştırılabilir hale getirilerek sınıflandırılabilir. En çok kullanılan kernel fonksiyonları arasında Gauss, Polinomial ve Sigmoid fonksiyonları sayılabilir.

4. E-POSTA İLETİLERİNDEN GÖREV KEŞFİ (TASK EXPLORATION FROM E-MAIL MESSAGES)

Çalışmada, 200 adet "Normal" ileti ile uygulama üzerinden "ToDo" olarak işaretlenmiş 175 adet iletiden oluşan toplam 375 adet e-posta içeren veri seti kullanılmıştır. Veri setlerinin eğitim ve sınıflandırma süreçlerine hazırlanması için Bölüm 3.2'de sunulan ön işlem aşamaları uygulanmıştır. Ön işlem aşaması sonrasında, her bir ileti için TF-IDF değerlerinden oluşan vektör uzay modeli elde edilmiştir. Kelimelere ilişkin terimlerin köklerinin elde edilmesinde Nzemberek [22] doğal dil işleme kütüphanesinden faydalanılmıştır. Sınıflayıcıların eğitimi için veri setinin %75'i eğitim kümesi olarak ayrılmıştır. Kalan %25'lik bölüm ise test amacıyla kullanılmıştır. Sınıflayıcı olarak merkez tabanlı, yapay sinir ağları ve destek vektör makinaları kullanılmıştır. Üç sınıflayıcı için de eğitim ve test kümelerinde, aynı veriler kullanılmış ve böylece örneklemeden doğacak performans farklılıklarının önüne geçilmiştir. Yapay sinir ağı modelinde ileri beslemeli çok katmanlı algılayıcı tercih edilmiştir. Oluşturulan modelde gizli katmanda 10 adet nöron, çıkış katmanında ise 2 sınıf olduğundan dolayı 2 nöron kullanılmıştır. Giriş katmanında kullanılan nöron sayısı ise veri setinin özellik vektörü tarafından belirlenmektedir. DVM sınıflayıcı için kernel fonksiyonu olarak Gauss fonksiyonu, genişlik değeri sigma=1 ve cezalandırma katsayısı c=1 olarak belirlenmiştir. Sınıflandırma sonuçlarını karşılaştırmak amacıyla her üç sınıflayıcı için doğruluk (accuracy), duyarlılık (precsion), hassasiyet (recall or sensitivy) ve f-metrik değerlerinden faydalanılmıştır. Bu değerlerin hesaplanmasında kullanılan eşitlikler Tablo 1'de sunulmuştur. Tablodaki eşitliklerde kullanılan TP, TN, FP, FN değerleri sırasıyla;

- TP (True Pozitif): Gerçekte "ToDo" olup, "ToDo" olarak sınıflandırılan e-posta sayısı
- TN (True Negatif): Gerçekte "Normal" olup, "Normal" olarak sınıflandırılan e-posta sayısı
- FP (False Pozitif): Gerçekte "Normal" olup, "ToDo" olarak sınıflandırılan e-posta sayısı
- FN (False Negatif): Gerçekte "ToDo" olup, "Normal" olarak sınıflandırılan e-posta sayısıdır.

Tablo 2'de her üç veri seti ve sınıflandırma yöntemi için TP, TN, FP ve FN değerlerini gösteren karmaşıklık matrisleri verilmiştir. Karmaşıklık matrisi yardımıyla hesaplanan doğruluk (A), duyarlılık (P), hassasiyet (R), f-metrik ve alıcı işletim karakteristiği (Receiver Operating Characteristic - ROC) eğrisi altında kalan alanın (Area Under Curve - AUC) değerini gösteren sonuçlar Tablo 3'te sunulmuştur. Performans ölçütleri ve grafiklerden görüleceği üzere her üç veri setinde de en iyi sınıflama performansı yapay sinir ağlarında elde edilmiştir. Şekil 4'te verilen ROC eğrilerinde bu durum açıkça görülmektedir. ROC eğrilerinde sol üst köşeye en yakın grafik, performansı en yüksek sınıflayıcıya ait grafiktir.



Şekil 4. ROC eğrisi
(ROC curve)

Çalışma sonucunda en iyi performans elde edilen eğitilmiş yapay sinir ağına ait ağırlıklar, kurumsal e-posta sistemleri üzerinde çalışan e-posta istemcisi ve kurumsal anında mesajlaşma ve iş yönetim uygulaması ile entegre edilmiştir. Geliştirilen yazılım, XMPP tabanlı gerçek zamanlı durum ve mesaj yönetim protokolü üzerinden işbirlikçi bir çalışma altyapısı sunmaktadır. Yapılan çalışmaya ait mimari diagram Şekil 5'te sunulmuştur.

Table 1. Sınıflayıcıları karşılaştırmak üzere kullanılan metrikler
(Metrics used to compare the classifiers)

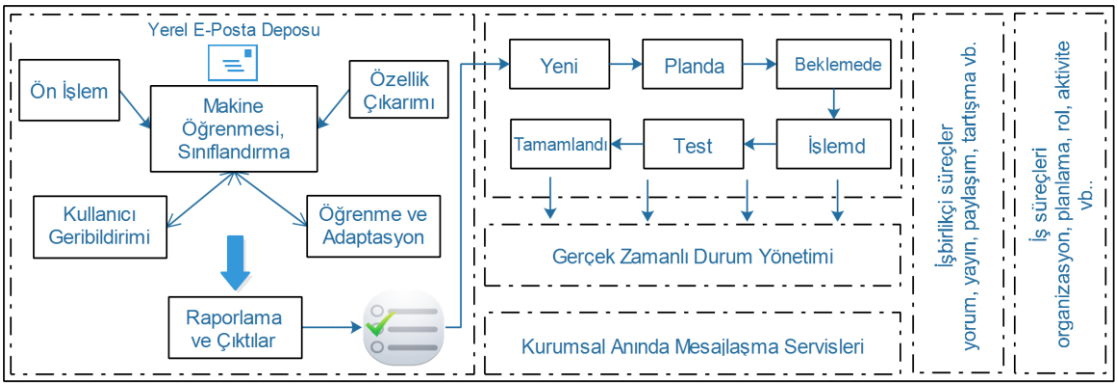
Doğruluk(Accuracy)	Keskinlik(Precision)	Duyarlılık(Sensivity)	F-metric
$A = \frac{TP + TN}{TP + TN + FN + FP}$	$P = \frac{TP}{TP + FP}$	$R = \frac{TP}{TP + FN}$	$F = \frac{2 * P * R}{P + R}$

Tablo 2. Sınıflayıcılara ilişkin karmaşıklık matrisleri
(Complexity matrices for classifiers)

Centroid				ANN				SVM			
Tahmin	Gerçek (Actual)			Tahmin	Gerçek (Actual)			Tahmin	Gerçek (Actual)		
		ToDo	Normal			ToDo	Normal			ToDo	Nomal
	ToDo	171	10		ToDo	173	8		ToDo	127	0
	Normal	4	190		Normal	2	192		Normal	48	200

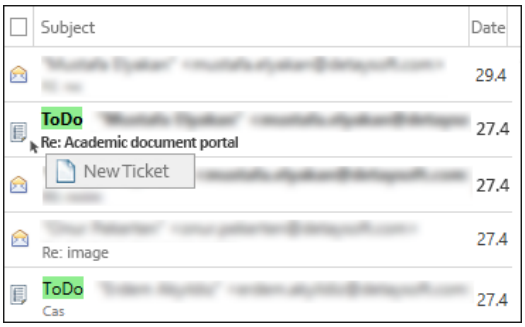
Tablo 3. Sınıflayıcılara ilişkin performans ölçütleri
(Performance criteria for classifiers)

Sınıflayıcı	A	P	R	F-metrik	AUC
Centroid	0.9627	0.9448	0.9771	0.9607	0.9636
ANN	0.9733	0.9558	0.9886	0.9717	0.9795
SVM	0.8720	1.0000	0.7257	0.8411	0.8629

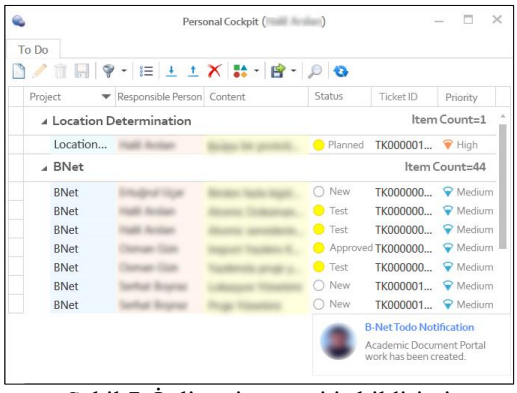


Şekil 5. Uygulama mimarisi
(Application architecture)

Geliştirilen uygulamanın e-posta istemcisinin gelen kutusundaki iletiler, Şekil 6'da gösterildiği gibi "ToDo" ve "Normal" ileti olarak etiketlenmiştir. "ToDo" etiketine sahip iletiler, kullanıcının onayından sonra iş listesine aktarılmaktadır.



Şekil 6. Sınıflandırılmış gelen kutusu ekranı
(Classified inbox screen)



Şekil 7. İş listesi ve yeni iş bildirimi
(Task list and new task notification)

İş listesine aktarılan ileti, gerçek zamanlı bildirim (push notification) olarak bu işle ilişkili diğer kullanıcılara bildirilmektedir. Ayrıca, iş listesine dönüştürülen e-posta iletisi üzerinden alınan tüm cevaplar ve iş değişiklikleri todo list uygulaması ile takip edilebilmektedir. İş listesine eklenen bir e-postanın ilgili kullanıcılara XMPP altyapısı kullanılarak yapılan bildirimi Şekil 7'de gösterilmiştir.
Kullanıcı iş listesine eklenen e-posta iletisine, todo list uygulaması üzerinden erişilebilmektedir. Bu e-postanın cevap iletileri görüntülenebilmekte ve bu iş üzerinde oluşan tüm hareketler, ilgili kişiye bilgi e-postası olarak gönderilmektedir.

5. SONUÇ VE DEĞERLENDİRME (RESULTS AND EVALUATIONS)

E-posta sistemleri, kurumsal iş akışlarının yürütüldüğü önemli bir iletişim ortamıdır. Dolayısıyla, e-posta iletileri, todo list, planlama vb. iş uygulamaları için önemli girdiler sağlamaktadır. Bu çalışma ile işbirlikçi çalışmaya uygun, kurumsal anında mesajlaşma altyapısı üzerinde, akıllı e-posta istemcisi ve todo list yönetim uygulaması geliştirilmiştir. E-posta istemcisi, kullanıcının gelen kutusunundaki iletileri sınıflandırabilmekte ve bu iletileri todo list uygulamasına aktarabilmektedir. Bu entegrasyon sayesinde dikkatsizlik nedeniyle e-posta üzerinden iletilen müşteri talepleri ve şikayetleri anında todo list uygulamasına aktarılarak yaşanacak iş kayıpları, taleplere zamanında cevap verememe gibi sorunlar büyük oranda azalacak, sonuçlandırılmayan işlerden dolayı geciken hizmet faturalamaları ve tahsilat sorunları ortadan kalkacak, personelin daha verimli bir şekilde çalışması sağlanacaktır. Sistem iş kayıplarını azaltarak firma karlılığı artırırken, diğer yandan firmanın daha iyi bir müşteri ilişkileri yönetimi geliştirmesine katkıda bulunacaktır.

Çalışmada sınıflayıcı olarak merkez tabanlı sınıflayıcı, yapay sinir ağları ve destek vektör makinaları kullanılmış, ilgili sınıflayıcıların karşılaştırılması için çeşitli metrikler hesaplanmış ve ROC grafikleri verilmiştir. Bu metrikler incelendiğinde, gerek doğru sınıflandırma oranı, gerek f-metrik değeri, gerekse AUC oranları açısından en yüksek değerler yapay sinir ağlarında elde edilmiştir. Bu nedenle, bu çalışma için sınıflandırma performansı en yüksek sınıflayıcı olarak yapay sinir ağları öne çıkmıştır. Çalışmada en düşük performans değerlerine sahip sınıflandırıcı ise destek vektör makinaları olmuştur. Merkez tabanlı sınıflayıcı ise çok basit bir sınıflayıcı olmasına karşın neredeyse yapay sinir ağlarına yakın performans değerlerine sahip olduğu görülmektedir.

Çalışmada kullanılan veri setinin sınıflandırılması sonucunda en yüksek başarıma sahip olan yapay sinir ağlarına ait ağırlıklar, geliştirilen kurumsal mesajlaşma ve iş yönetim uygulamasına entegre edilmiştir. Bu entegrasyon sonucunda tasarlanan uygulama, kullanıcının gelen kutusundaki e-posta iletilerini, "ToDo" ya da "Normal" olarak etiketleyebilmekte ve bu iletilerin iş listesine aktarımı için fonksiyonlar sunmaktadır.

Sonuç olarak, çalışma ile e-posta sistemlerinin kurumsal uygulama ve iş süreçlerine katılımı için yeni bir yöntem önerilmektedir. Sonraki çalışmalarda gelen e-postaların içeriği analiz edilerek, iletinin hangi çalışma grubu ya da personelle ilgili olduğu tespit edilmeye çalışılacak ve görev bağlama işi tamamen otomatikleştirilecektir. Ayrıca gelen e-postalar, önceki postalarla karşılaştırarak müşteriler için otomatik cevaplama sistemleri, şirket personeli için ise, daha önce karşılaşılan aynı sorunların çözümüne ilişkin öneri ve tavsiye sistemi geliştirilmesi düşünülmektedir. Böylelikle kurum hafızanısının oluşturulmasına ve geliştirilmesine katkıda bulunulacaktır.

TEŞEKKÜR (ACKNOWLEDGEMENT)

Bu çalışma, TÜBİTAK tarafından desteklenen ve Detaysoft Ar-Ge Merkezi bünyesinde yürütülen 3150617 no'lu projenin bir sonucudur. Test ortamı ve desteklerinden ötürü teşekkür ederiz.

KAYNAKLAR (REFERENCES)

[1] G. Tang, J. Pei, W. S. Luk, "Email Mining: Tasks, Common Techniques, and Tools", *Knowledge and Information Systems*, 41(1), 1-31, 2014.

[2] M. Suit, H. Wortmann, "Discovery and analysis of e-mail-driven business processes", *Information Systems*, 37(2), 142-168, 2012.

[3]K. Coussement, D. V. Poel, Improving customer complaint management by automatic email classification using linguistic style features as predictors, *Decision Support Systems*, 44(4), 870-882, 2008.

[4] L. Dey, S. Bharadwaja, G. Meera, G. Shroff, Email Analytics for Activity Management and Insight Discovery, **IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT)**, 557-564, 2013.

[5] S. S. Weng, C. K. Liu, "Using text classification and multiple concepts to answer e-mails", *Expert Systems with applications*, 26(4), 529-543, 2004.

[6] S. Appavu, R. Rajaram, M. Muthupandian, G. Athiappan, K. S. Kashmeera, "Data mining based intelligent analysis of threatening e-mail", *Knowledge-Based Systems*, 22(5), 392-393, 2009.

[7] M. F. Wan, M. F. Tsai, S. L. Jheng, C. H. Tang, "Social feature-based enterprise email classification without examining email contents", *Journal of Network and Computer Applications*, 35(2), 770-777, 2012.

[8] Internet: Openfire XMPP Server, a real time collaboration community, http://www.igniterealtime.org, 08.01.2016.

[9] B. Yu, D. Zhu, "Combining neural networks and semantic feature space for email classification", *Knowledge-Based Systems*, 22(5), 376-381, 2009.

[10] I. Alsmadi, I. Alhami, "Clustering and classification of email contents", *Journal of King Saud University - Computer and Information Sciences*, 27(1), 46-57, 2015.

[11] D. C. Soares, F. M. Santoro, F. A. Baiao, "Discovering collaborative knowledge-intensive processes through e-mail mining", *Journal of Network and Computer Applications*, 36(6), 1451-1465, 2013.

[12] J. R. Méndez, M. Reboiro-Jato, F. Díaz, E. Díaz, F. Fdez-Riverola, "Grindstone4Spam: An optimization toolkit for boosting e-mail classification", *The Journal of Systems and Software*, 85(12), 2909-2920, 2012.

[13] I. Koprinska, J. Poon, J. Clark, J. Chan, "Learning to classify e-mail", *Information Sciences*, 177(10), 2167-2187, 2007.

[14] P. Pankaj, M. Hyde, J. A. Rodger, "P2P Business Applications: Future and Directions", *Communications and Network*, 4, 248-260, 2012.

[15] S. V. Ragavana, I. K. Kusnantoa, V. Ganapathyb, "Service Oriented Framework for Industrial Automation Systems", *Procedia Engineering*, 41, 716-723, 2012.

[16] M. G. Armentano, A. A. Amandi, "Enhancing the experience of users regarding the email classification task using labels", *Knowledge-Based Systems*, 71, 227-237, 2014.

[17] E. H. S. Han, G. Karypis, "Centroid-Based Document Classification: Analysis and Experimental Results", **European conference on principles of data mining and knowledge discovery**, 424-431, Springer Berlin Heidelberg, Eylül 2000.

[18] L. Fausett, **Fundamentals of Neural Networks: Architectures, Algorithms and Applications**, Prentice Hall, Inc., 1994.

[19] J. Clark, I. Koprinska, J. Poon, A Neural Network Based Approach to Automated E-mail Classification, **International Conference on Web Intelligence (WI'03)**, 702-705, 2003.

[20] O. Kaynar, F. Demirkoparan, "Forecasting Industrial Production Index with Soft Computing Techniques", *Economic Computation and Economic Cybernetics Studies and Research*, 46(3), 113-138, 2012.

[21] C. Cortes, V. Vapnik, "Support-Vector Networks", *Machine Learning*, 20(3), 273-297, 1995.

[22] Internet: A. A. Akin, M. D. Akin, NLP library, NZemberek 0.1.0, http://www.nuget.org/packages/NZemberek, 11.02.2016.