YAPAY BAĞIŞIKLIK SİSTEMİ İLE SPAM FİLTRELEME

Cüneyt ÖZDEMİR

Yüksek Lisans Tezi

Bilgisayar Mühendisliği Yazılım Anabilim Dalı

Danışman: Doç. Dr. Ahmet Bedri ÖZER

OCAK-2013

T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YAPAY BAĞIŞIKLIK SİSTEMİ İLE SPAM FİLTRELEME

YÜKSEK LİSANS TEZİ

Cüneyt ÖZDEMİR

(102129101)

Anabilim Dalı: Bilgisayar Mühendisliği

Programı: Yazılım

Danışman: Doç. Dr. Ahmet Bedri ÖZER

Tezin Enstitüye Verildiği Tarih: 31 Ocak 2013

OCAK-2013

T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YAPAY BAĞIŞIKLIK SİSTEMİ İLE SPAM FİLTRELEME

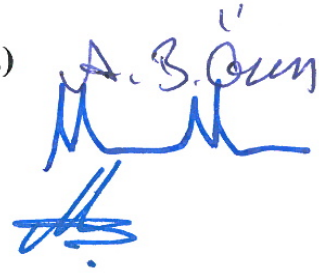
YÜKSEK LİSANS TEZİ

Cüneyt ÖZDEMİR

(102129101)

Tezin Enstitüye Verildiği Tarih: 31 Ocak 2013
Tezin Savunulduğu Tarih: 11 Ocak 2013

Tez Danışmanı :	Doç. Dr. Ahmet Bedri ÖZER (F.Ü.)
Diğer Jüri Üyeleri :	Doç. Dr. Mehmet KAYA (F.Ü.)
	Yrd. Doç. Dr. Musa ATAŞ (S.Ü.)



OCAK-2013

ÖNSÖZ

Çalışmam boyunca ilgi ve desteğini hiç esirgemeyen, bilgi ve deneyimlerini benimle paylaşarak beni yönlendiren ve bana büyük katkı sağlayan danışman hocam Doç. Dr. Ahmet Bedri ÖZER'e sonsuz teşekkür ederim.

Akademik hayatta bana yön gösteren ve bu camiada örnek insanın nasıl olacağını kendi yaşantısı ile bana gösteren, hiçbir yardımını benden esirgemeyen ve bütün konularda düzeyli bir şekilde tartışabildiğim değerli hocam ve arkadaşım Yrd.Doç.Dr. Musa ATAŞ'a teşekkürü bir borç bilirim.

Çalışmalarım sırasında desteğini benden esirgemeyen Yrd.Doç.Dr. Yılmaz Kaya ve Yrd.Doç.Dr. Murat UYAR'a çok teşekkür ederim.

Çalışmam boyunca maddi manevi her konuda bana destek olan eşime, çocuklarım Berfin, Emine, Şeyma Nur ve Muhammed Yasin'e, ayrıca burada adını yazmadığım tüm değerli dost ve arkadaşlarıma teşekkür ederim.

Cüneyt ÖZDEMİR

ELAZIĞ - 2013

İÇİNDEKİLER

Sayfa No

ÖZET

Günümüzde elektronik postalar (e-posta), hızlı ve maliyet olarak ucuz olduklarından dolayı hızlı reklam veya propaganda yapmak isteyen kişi veya şirketlerin yoğun ilgisini çekmektedir. Kişi veya şirketler hedeflerine ulaşmak için milyonlarca sayıda ve aynı içerikte gereksiz ve uygunsuz e-postayı tanımadıkları e-posta hesaplarına yollarlar. Birçok kişi bu şekilde e-posta saldırısına maruz kalmaktadır. Elektronik postalar rahatsız edici içeriklere sahip olmalarının yanında, internet trafiğini de çok fazla meşgul etmekte, e-posta hizmeti veren firmalara büyük bir maliyet getirmekte ve internet kullanıcılarının zamanlarının israf olmasına sebebiyet vermektedir.

Bu tez çalışması istenmeyen e- postaların filtrelenmesinde yürütülen kararlı mücadeleye katkıda bulunmayı hedeflemiştir. Bu çalışmada Türkçe istenmeyen e- postaların tespiti amaçlanmıştır. Bu işlem için Bayes sınıflandırıcısı ve farklı yapay bağışıklık sistemi (YBS) algoritmaları kullanılmış ve en iyi sınıflandırmayı yapan algoritmalar tespit edilmeye çalışılmıştır. Tez çalışmasında YBS algoritmalarından AIRS1, AIRS2, AIRS2PARALLEL, CLONALG ve CSCA algoritmaları incelenmiştir.

Tez çalışması için Türkçe spam ve normal e-postalar toplanmıştır. Toplanan e-postalar ile ilgili herhangi bir işlem yapmadan önce bu e-postalar 2 gruba ayrılmıştır. Özniteliklerin tespiti için Fisher ayrımsallık analizi ve Öklid uzaklığı kullanılmıştır. Öznitelik vektöründen veri kümesi oluşturmak için her gruptaki e-postalar işlenmiştir. Böylelikle 2 farklı veri kümesinde 2 farklı öznitelik çıkarımı ile 4 farklı veri kümesi elde edilmiştir. Sınıflandırma algoritmaları bu 4 farklı veri kümesi üzerinde çalıştırılmış, elde edilen sonuçların ortalaması alınmıştır.

Fisher ayrımsallık analiz yöntemine göre seçilen özniteliklerin, Öklid uzaklığına göre seçilen özniteliklerden daha iyi sınıflandırma oranına sahip olduğu görülmüştür. Bu çalışmada YBS algoritmalarından CSCA algoritmasının spam e-postaları en iyi sınıflandıran algoritma olduğu tespit edilmiştir.

Anahtar Kelimeler: Yapay Bağışıklık Sistemi, İstenmeyen Elektronik Posta, Veri Kümesi Oluşturma, Fisher Ayrımsallık Analizi, CSCA, AIRS

SUMMARY

ARTIFICIAL IMMUNE SYSTEM WITH SPAM FILTER

Since today, e-mails are fast and cheap enough, they engages the attentions of individuals and companies who want to advertise and make promotions. Person or companies send millions of unnecessary and unsolicited e-mails with the same content to unidentified e-mail accounts in order to reach their aims. Many people are subjected to this type of e-mails every day. These e-mails occupy internet network bandwidth too much along with having disturbing contents, loading a huge cost to the companies providing e-mail services and wasting the internet user's times.

This thesis aims to make contribution on the stable struggle about filtering unsolicited e-mails. In this work, it is aimed to detect unsolicited e-mails written in Turkish language. Bayes classifier and different algorithms of artificial immune system (AIS) are utilized and best AIS algorithms are determined for this study. In this thesis, AIRS1, AIRS2, AIRS2PARALLEL, CLONALG and CSCA algorithms of AIS, have been investigated.

Turkish spam and Non-spam e-mails are collected. Initially these e-mails are splitted into two groups. Fisher dicriminant analysis and Eulidean distance methods are utilized for feature extraction. E-mails of each group are processed in order to build dataset from the feature vector. Thus, four different datasets are obtained from two types of feature extraction on the original two datasets. Classification algorithms are performed on these four datasets and then mean of the results are taken. It is observed that, classification rates of the Fisher based features outperform to Euclidean based features. CSCA algorithm which is a significant type of the AIS algorithms is determined as a best spam e-mail classifier.

Key Words: Artificial Immune System, Unsolicited E-Mail, Dataset Generation, Fisher Discriminant Analysis, CSCA, AIRS

ŞEKİLLER LİSTESİ

Sayfa No

TABLOLAR LİSTESİ

Sayfa No

SİMGELER VE KISALTMALAR

APC : Antijen Sunan Hücreler – Antigen presenting cells

BCC : Gizli Bilgi Kopyası - Blind Carbon Copy

CC : Bilgi Kopyası - Carbon Copy

DH : Dendrit Hücre – Dendritic Cell

DHA : Dentrit Hücre Algoritması – Dendritic Cell Algorithm

DNS : Alan Adı Sistemi - Domain Name System

E-Posta : Elektronik Posta

FAA : Fisher ayrımsallık analizi

GA : Genetik Algoritma

GPL : Genel Kamu Lisansı - General Public Licence

HLA : İnsan Lokosit Antijenleri - Human Leukocyte Antigens

IMAP : Internet Mesaj Kabul Protokolü - Internet Message Access Protocol

IP : Internet Protokolü – Internet Protocol

Makrofaj : Büyük Yiyiciler - Macrophage

MHC : Majör Doku Uyuşma Kompleks - Major Histokompatibility Complex

MIME : Çok Amaçlı İnternet Posta Uzantıları - Multipurpose Internet Mail Extensions

NK : Doğal Öldürücü - Natural Killers

NON-SELF : Yabancı Antijenler – Non-Self Antigens

PAMPs : Desen İlişkili Moleküler Desenler - Pattern Associated Molecular Patterns

POP3 : Postane Protokolü 3 - Post Office Protokol 3

PRRs : Örüntü Tanıma Reseptörleri - Pattern Recognition Receptors

SELF : Vücut Antijenleri – Self Antigens

SMTP : Basit Posta İletim Protokolü – Simple Mail Transfer Protocol

Spam : İstenmeyen Elektronik Posta

UBE : Talep Edilmemiş Kitlesel E-posta - Unsolicited Bulk e-mail

UCE : Talep Edilmemiş Ticari E-posta - Unsolicited Commercial e-mail

YBS : Yapay Bağışıklık Sistemleri – Artificial Immune System

YSA : Yapay Sinir Ağları – Artificial Neural Network

1. GİRİŞ

Belirsizlik insan yaşamındaki güç problemlerden biridir. Bu problem bilinmeyen kaynaklardan gelen, öngöremediğimiz ve etkisinin ne olacağını bilmediğimiz problemlerdir. Canlı vücudu çevreden gelen, hiç tanımadığı, bilmediği organizmaların saldırısına maruz kalmaktadır. Canlı vücudundaki bağışıklık sistemi, vücudun hiç tanımadığı hücreleri kısa sürede tanımasını sağlayarak onları imha etmeye çalışan, çok gelişmiş, yüksek performanslı bir sistemdir.

Bağışıklık sistemi; vücuda giren yabancı maddelerin etkisiz hale getirilmesi, dışarıya atılması veya kimyasal bir değişime tabii tutulması için vücudun oluşturduğu fizyolojik mekanizmalara denir [1]. Bağışıklık sistemi, vücut hücreleriyle yabancı hücreleri ayırt etme konusunda çok gelişmiş bir yeteneğe sahiptir.

Dünya üzerinde oluşan birçok olay birbirine benzerlik göstermektedir. Örneğin bütün dünyada devletlerin veya kişilerin düşmanları var olduğu sürece bireyler veya devletler bu savaş anında kendilerini muhafaza etmek ve gerekli teçhizata sahip olmak için gerekli güvenlik önlemlerini alırlar. Canlı vücudunda buna benzer olarak canlının dışarıdan gelen tehlikelere karşı kendini muhafaza etmesi için insan vücudu içindeki mekanizmanın sorunsuz çalışmasını sağlayan bir bağışıklık sistemi yaratılmıştır.

İnsanoğlu yaşamını kolaylaştırmak için doğada gerçekleşen olaylardan esinlenerek birçok problemi çözmüştür. Bilgisayar sistemlerindeki birçok karmaşık problemde de doğadan esinlenerek geliştirilen benzer yöntemler kullanılarak problemler çözülmeye çalışılmıştır. Bilgisayar sistemlerine uygulanan bu çözüm yöntemlerine yapay yöntemler denir. Bu yöntemler arasında en fazla ilgi uyandıran insan beynini ve insan vücudunun çalışma sistemini taklit eden ve öğrenme, kavrama, yorumlama gibi birçok özelliği taklit ederek geliştirilen yapay zekâ teknikleridir. Yapay zekâ tekniklerine örnek olarak Yapay Sinir Ağları (YSA), Genetik Algoritmalar (GA) ve Yapay Bağışıklık Sistemleri (YBS) [2] verilebilir.

YBS canlı vücudunun bağışıklık sisteminin çalışma prensibinden yola çıkılarak geliştirilmiş bir sistemdir. Bağışıklık sistemi beyin gibi öğrenebilme yeteneğine sahiptir.

Günlük yaşantımızda haberleşme aracı olarak kullanılan araçlardan biri ve en önemlisi

elektronik postalardır (e-posta). E-postaların hızlı ve maliyet olarak ucuz olması popüler haberleşme aracı olmasını sağlamıştır. TÜİK 2012 verilerine göre Türkiye'de her gün internete giren kullanıcıların %66,8 'i e-posta sunucularına gelen postaları okumak veya e-posta yollamak için internete girmektedir [3]. E-postalar hızlı ve maliyet olarak ucuz yoldan reklam, propaganda yapmak isteyen kişi veya şirketlerin ilgisini her gün biraz daha fazla çekmektedir. Bu kişi veya şirketler hedeflerine ulaşmak için milyonlarca sayıda ve aynı içerikte gereksiz ve uygunsuz e-postaları tanımadıkları e-posta hesaplarına yollarlar. Birçok kişi bu şekilde e-posta saldırısına tabi tutulmaktadır. Bu e-postalar rahatsız edici içerik veya görüntüye sahip ve internet trafiğini çok fazla meşgul etmektedirler. E-postaların güçlü haberleşme aracı olması güvenlik açısından risk getiren bir durumdur.

Günümüzde istenmeyen e-postaların (spam) engellenmesi için veri madenciliği teknikleri ile önlemler alınmaya çalışılmaktadır. Bu amaçla karar destek sistemleri, YSA, genetik algoritma veya karınca kolonisi, karar destek vektör makineleri, Bayes sınıflandırıcısı gibi algoritmalar kullanılarak otomatik istenmeyen e-posta filtreme araçları geliştirilmiştir [4].

Z. Chuan ve arkadaşları çalışmalarında spamassassin veri kümesini kullanmışlardır. Bu veri kümesinde 580 spam e-posta ve 420 normal e-posta bulunmaktadır. Bütün örnekler İngilizce e-postalardan oluşmaktadır. Öncelikle e-postaları içeriklerine göre (politik, alışveriş vs.) alt sınıflara ayırmışlardır. Daha sonra bu alt sınıflara ayrılan e-postaları doğru sınıflandırmak için linear vector quantization (LVQ) YSA ile bu sınıflar birleştirilmiştir. Çalışmada YSA deney sayısı filtrelemenin performansını etkilemiştir. Yeterli deney yapılmadığında sınıflandırmanın kötü olduğu görülmüştür. Bunun nedeni ise YSA'nın her bir özniteliğin diğer öznitelik kelimeleri arasındaki ilişki avantajıdır. Diğer bir sınıflandırma olarak Bayes yöntemi denenmiş ve YSA'nın Bayes yöntemine göre daha iyi sınıflandırdığı görülmüştür. Bayes sınıflandırmada özniteliklerin hepsi birbirinden bağımsızdır [5].

Burim Sirisanyalak ve Ohm Sornil çalışmalarında, spam e-postaları algılamak için öznitelik çıkarım yöntemi olarak YBS kullanmışlardır. Bu çalışmada birçok istenmeyen ve normal e-posta öncelikle toplanmıştır. Ardından html ayırma işlemi ve gereksiz kelimelerin atılması (stopwords) yapılmıştır. Negatif seçim yönteminden hareketle bu kelimeler arasından öznitelik çıkarımı yapılmıştır. Çıkarılan özniteliklere göre geriye yayılım sinir ağı kullanılmış ve %92,4 oranında doğru sınıflandırma elde edilmiştir [6].

Jiujun Chen ve ark.'larının yaptığı çalışmada bir veri kümesi oluşturulmuştur. Veriler üç kısımda toplanmıştır.

- Spam e-posta ve normal e-postalar toplanarak bu e-postalarda bulunan kelimeler ve cümleler birinci bölümde toplanmış
- Spam mesajlarda geçen telefon numarası, web adresleri ve iletişim adresleri ikinci bölümde toplanmış
- E-posta üst bilgisinde yer alan bilgiler üçüncü bölümde toplanmıştır.

Oluşturulan veriler geliştirilen algoritma ile sınıflandırılmaya çalışılmıştır. Bu algoritmada bir e-postanın istenmeyen olup olmadığını anlamak için Bayes yöntemine benzer bir sistem geliştirilmiştir. Bu sisteme göre kelimelerin benzerlik değerleri hesaplanmış ve bir eşik değerin üstünde veya altında olmasına göre e-postanın istenmeyen e-posta veya normal e-posta olduğu belirlenmeye çalışılmıştır. Test e-postalarında spam ve normal e-postaların yüksek oranda doğru bir şekilde sınıflandırıldığını söylemektedirler [7].

Ismaila Idris, spam e-postaları ayırt etmek için negatif seçim algoritmasını kullanmıştır. 57 özniteliğe sahip Corpus veri kümesi kullanarak yapılan çalışmada e-postaları sınıflandırma için YBS algoritmalarından negatif seçim algoritması kullanılmıştır. Örnek sınıf ve eğitim veri kümesi arasında mesafe Öklid uzaklığı formülü kullanılarak hesaplanmıştır. Yapılan testlerde %94,05 doğruluk oranına ulaşılmıştır [8].

Jing Zhang ve ark.'ları kısa e-postaların filtrelenmesinde benzerlik oranını kullanmışlardır. Buna göre yeni bir e-posta geldiğinde kısa e-posta özellik vektörü (VSM) e-posta içindeki verileri kelimelere ayırır, ardından bu e-postanın benzerlik değeri ile eşik değeri kıyaslanır. Benzerlik değeri $1 - d(i,j)$ formülü ile hesaplanır [9].

$$\text{Öklid mesafe uzaklığı } d(i,j) = \sqrt{\sum_{t=1}^{n} \left(X_{it} - Y_{jt}\right)^2} \tag{1.1}$$

Formül 1.1'de verilen X_{it} antijen, Y_{jt} antikorları temsil etmektedir. Gelen kısa e-postanın benzerlik değeri eşik değerden yüksek ise bu e-posta spam, değilse normal e-postadır. Bu işlemin ardından bu e-posta eğitim kümesine dâhil edilir ve sistem yeniden eğitilir [9].

Spam e-postaların getirdiği maliyet ve iş yükünün fazlalığı bu e-postalar ile

mücadelenin kararlı bir şekilde sürdürülmesi gerektiğini göstermektedir. İstatistiklere bakıldığında e-postalar arasında spam e-postalar 2010 yılında %89,1 oranında iken 2011 yılında bu oran %71'e düşmüştür [10,11]. Bu veriler spam e-postalarda yürütülen kararlı mücadeleyi göstermektedir. Spam e-postalar ile yürütülen mücadelede çalışmalar İngilizce veri kümeleri üzerine yoğunlaşmış durumdadır. Bu tez çalışması özellikle Türkçe spam e-postaların içerik olarak filtrelemesine katkıda bulunmayı amaçlamaktadır. Bu çalışmada Türkçe spam e-postaların karakteristik özellikleri çıkarılmış ve bu özelliklere göre filtreleme çalışmaları yapılmıştır. Çalışmada Türkçe e-postalardan oluşan spam ve normal (Non-spam) e-postaları içeren veri kümeleri elde edilmiştir. Elde edilen veri kümeleri için özniteliklerin çıkarılmasında Fisher ayrımsallık analizi (FAA) ve Öklid uzaklık yöntemi kullanılmış, sınıflandırma algoritması olarak YBS algoritmaları ve Bayes sınıflandırıcısı kullanılmıştır. Kullanılan YBS algoritmaları AIRS1, AIRS2, AIRS2PARALLEL, CLONALG ve CSCA algoritmalarıdır. Bu çalışmada farklı veri kümeleri için kullanılan sınıflandırma algoritmaları arasında en yüksek sınıflandırma ayrımsallığını veren algoritma tespit edilmeye çalışılmıştır.

Tez çalışması 7 bölümden oluşmaktadır. Birinci bölümde tezin giriş bölümü anlatılmıştır. İkinci bölümde e-posta ve istenmeyen e-postalar (spam) ile ilgili detaylı bilgiler verilmiştir.

Üçüncü bölümde bağışıklık sistemi detaylı bir şekilde anlatılmış, dördüncü bölümde bağışıklık sisteminden esinlenerek bilgisayar sistemlerine uyarlanan YBS, YBS'de kullanılan algoritmalar ve Bayes yöntemi anlatılmıştır.

Beşinci bölümde bu tez çalışması sırasında deneylerde kullanacağımız Türkçe veri kümesini oluşturma adımları detaylı olarak anlatılmıştır. Altıncı bölümde deneylerde elde edilen bulgular paylaşılmıştır. Son bölümde ise yapılan deneyler ve deney sonuçları paylaşılmıştır.

2. ELEKTRONİK POSTA SİSTEMİ

2.1. Elektronik Posta Sistemi

İnsanlar arasındaki haberleşme araçlarından biri ses diğeri ise yazıdır. Ses insanoğlunun yaradılışından itibaren kullanılan bir iletişim aracıdır. Yazı ise Sümerler ile birlikte yüzyıllardır kullanılan bir iletişim aracıdır. Yüzyıllardır kullanılan ve günümüzde eskisi kadar olmasa da halen kullanılan haberleşme araçlarından biri mektuptur. Bilgisayarın gelişmesi, internetin keşfi ile birlikte mektuptan esinlenerek bilgisayar kullanıcılarının haberleşmesi için e-postalar geliştirilmiştir. Günümüzde en fazla kullanılan haberleşme araçlarından biri e-postalardır. TÜİK verilerine göre [3] 2012 hane halkı bilişim teknolojileri kullanım araştırması verilerine göre internet kullanıcılarının %66,8 i interneti e-posta gönderme/alma için kullanmaktadır.

E-postalar normal düz metin postalardan oluşacağı gibi video, resim, ses vb. eklentiler ile birlikte gönderilebilir.

E-posta mesajların gönderilmesi ve alınması için e-posta kullanıcı hesaplarına ihtiyaç duyulmaktadır. E-posta kullanıcı hesapları oluşturulurken "@" simgesinin kullanılması gerekmektedir. "@" simgesi bir hesabın e-posta adresi olduğunu belirtmek için kullanılan zorunlu bir simgedir. Bir e-posta hesabı oluşturulurken @ simgesinin ön tarafına tekil kullanıcı adı yazılır, sonuna ise bu e-posta hizmetini veren domain yazılır. Örneğin cozdemir@siirt.edu.tr e-posta hesabı için "@" simgesinden sonraki siirt.edu.tr bu e-posta hizmetini veren domain adresini, "@" simgesinden önceki cozdemir ise siirt.edu.tr domaini içindeki tekil kullanıcı adını belirtmektedir.

2.2. Elektronik Posta Tarihçesi

E-postalarda kullanımı zorunlu olan "@" simgesinin kâşifi Ray Tomlinson'dur. Ray Tomlinson 1971 yılında, ARPANET üzerinden insanların birbirleriyle iletişim kurmalarını sağlayan "SNDMSG" adlı projeyi geliştirdi. Projede kişilerin birbirlerine attıkları mesajlar bir önceki mesaja eklenmekteydi. Böylelikle önceki mesajların silinmesini

engellemekteydi. Ray Tomlinson “SNDMSG” projesini bir adım ileriye götürerek “CYPNET” projesine önderlik etti. “CYPNET” projesi "SNDMSG" projesine ek olarak yollanan e-postalara ekli dosyaların iliştirilmesi sağlandı [12].

E-Postanın icadından iki yıl sonra yapılan bir araştırmada ARPANET deki e-postaların, internet trafiğinin %75’ini oluşturduğu belirtildi. 1988 yılında sadece e-posta yollama işlemlerini yürüten ilk bilgisayar programı "Eurora" yazıldı. 1995 yılında kullanıcıların internet gezginini kullanarak ulaşılabilecekleri ilk posta servisi "Hotmail"’ kullanıcılara sunuldu. Bu gelişmelerden ve e-posta kullanım oranında görülen büyük artıştan sonra diğer büyük bilgisayar şirketleri (Yahoo, G-Mail vs) internet tabanlı posta servisi vermeye başladı [13].

Günümüzde e-postaları okumak ve yollamak için kullanılan farklı programlar mevcuttur. Bu programlara örnek olarak Mozilla Thunderbird, Microsoft Outlook, Microsoft Outlook Express, Eudora ve Pine gösterilebilir. Web tarayıcılarını kullanarak e-posta yollamak için kullanılan bazı ücretsiz e-posta servisleri bulunur. Bu e-posta servislerinin bazıları; Google Gmail, Yahoo Mail, Microsoft Hotmail, Yandex Mail, Mynet vb. sitelerdir.

2.3. Elektronik Posta Yapısı

E-posta mesajları başlık ve gövde kısımlarından oluşur. Başlık kısmında gönderici (FROM), alıcı kullanıcıların kimliği (TO), konu başlığı (SUBJECT), tarih (DATE), alındı (RECEIVED) ve içerik numarası (Message-ID) bulunur. Gövde kısmında ise mesajın içeriği ve eklerin (ATTACHMENT) yapılacağı kısım bulunur. Gönderici, alıcı ve tarih kısımları zorunlu, başlık ve içerik numarası kısımları isteğe bağlıdır. Birçok e-posta hizmeti veren programlarda sadece başlık ve içerik zorunlu tutulmaktadır. RECEIVED, e-postanın yolculuğu sırasında uğradığı sunucuları tutmak için kullanılır. E-postanın geriye dönük izlenmesi için sunucular mesaj başlığına RECEIVED satırını eklerler.

E-postada diğer bir kısım, alıcı kullanıcıların yazıldığı TO kısmından farklı olarak, karşı tarafa mesaj vermek amacıyla kullanılan Bilgi Kopyası (Carbon Copy – CC) ve Gizli Bilgi Kopyası (Blind Carbon Copy - BCC) kısımlarıdır. CC ve BCC, TO kısmına yazılan e-posta adreslerinin girildiği kısımdır. TO kısmından farklı olarak CC kısmına yazılan kullanıcı kimlikleri ile CC kısmına yazılan e-posta sahiplerine: *“bu e-posta TO kısmında yazılan*

kişiye yazılmış olup sizlere bilgi amaçlı yazılmıştır" mesajı vermektedir.

BCC kısmı gizli alıcılar için kullanılır. BCC kısmına eklenen e-posta sahipleri sadece kendi e-posta adreslerini görebilir, e-postanın yollandığı diğer kullanıcıları göremezler.

Çok amaçlı internet posta uzantıları (MIME - Multipurpose Internet Mail Extensions) olarak adlandırılan internet standardı, mesajların nasıl biçimlendirileceğini ve eklentilerin mesajdan nasıl ayrılacağını belirler. MIME, Aktarım Kodlaması (Content-Transfer-Encoding) ve İçerik Türü (Content-Type) satırlarından oluşur. MIME kodlaması hakkındaki bilgi, Request for Comments (RFC 2822) de belirtilen üstbilgi (DATE, TO, FROM) alanlarından sağlanabilir. Content-type üstbilgisi bir bileşenin veya gövdenin tamamının içeriğini tanımlamada kullanılır [14,15].

```
From: cozdemir@siirt.edu.tr
To: uyuksel@hotmail.com
Subject: Picture.
MIME-Version: 1.0
Content-Transfer-Encoding: base64
Content-Type: image/jpeg
 base64 encoded data .....
.....
.....base64 encoded data
```

Şekil 2.1. Çok amaçlı internet e-posta uzantısı [12].

Şekil 2.1'de, "Content-Transfer-Encoding" veriyi çözmek için kullanılan metodu, "Content-Type" satırı ise mesaj gövde ve ek veri tipinin image/jpeg olduğunu belirtmekte olup, "Base64 encoded data…" satırı çözülmüş veriyi ifade etmektedir [14].

Farklı bir content-type örneği aşağıda verilmiştir.

```
Content-type: text/html; charset=euc-kr;
Content-Type: application/zip; name="testfile.zip"
```

Şekil 2.2. Çok amaçlı internet e-posta uzantısı 2 [15].

İlk satır, mesajın Kore dili karakter kümesi kullanılarak HTML formatında olacağını, ikinci satır bileşenin bir zip dosyası olduğunu ve testfile.zip olarak kaydedildiğini belirtir. İkilik sayı düzenindeki (binary) dosyalar eklenti olarak gönderilmek istenildiğinde binary dosyaları kodlanmak zorundadır [15].

2.3. Elektronik Posta Protokolleri

2.3.1. SMTP (Simple Mail Transport Protocol)

Basit Posta İletim Protokolü (Simple Mail Transport Protocol – SMTP), internet protokollerinde bulunan uygulama katmanı protokollerinden en çok kullanılan standartlardan biridir.

E-posta mesaj formatı, 1982 yılında oluşturulan RFC 822 standardı ile sadece ASCII metin içerikli kullanımı desteklemiştir. 1992 yılında RFC 1341 standardı ile oluşturulan ve 1996 yılında RFC 2045 standardı ile güncellenen çok amaçlı internet posta uzantıları (MIME) ile birlikte, içinde ekli dosyaların olduğu e-postaların kullanımı mümkün hale gelmiştir [16].

SMTP, e-postaların gönderilmesini sağlayan protokoldür. SMTP protokolü ile kullanıcılar e-postalarını gönderebilirler. SMTP protokolü 25 numaralı portu kullanmaktadır. Türkiye de TTNET spam ile mücadele edebilmek için SMTP protokolünün kullandığı port numarasını 587 olarak değiştirmiştir. Bu değişim ile bilgisayarlara sızan ve kullanıcıların haberi olmadan e-posta yollayan programların 25 numaralı port üzerinden e-posta yollanması engellenmiştir.

2.3.1.1. SMTP modeli

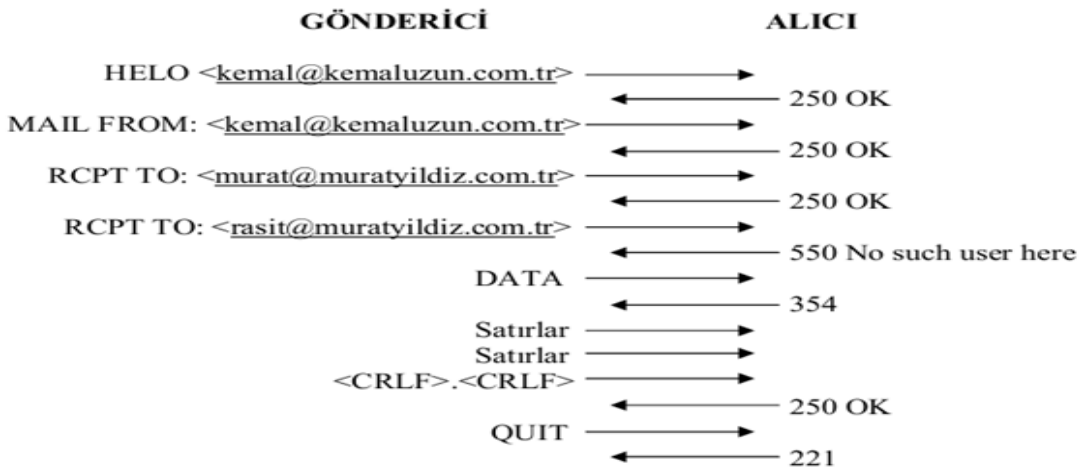
Kullanıcılar e-posta yollamak istediği zaman SMTP iki yönlü bir iletişim kanalı kurar. SMTP göndericisinin sorumluluğu mesajı bir veya daha fazla SMTP alıcısına aktarmak ve hata varsa bunu raporlamaktır [17].

SMTP protokolü e-postayı aynı domain içindeki kullanıcıya yollarsa, e-posta domain

içendeki sunucuya direk verilir. Farklı domainlerde bulunan kullanıcılar haberleştiği vakit SMPT protokolleri şu şekilde iletişim kurarlar [12,14].

1. SMTP e-posta yollanmadan önce iki SMTP protokolü yetkili işaretleri yollamak için karşı tarafın 25 numaralı SMTP portuna bağlanır.
2. Ardından gönderici SMTP, e-posta alışverişi için gerekli bilgilerini (gönderici, alıcı hesap adını, mesajı) içeren MAIL komutunu alıcı SMTP sunucusuna iletir.
3. Gönderici SMTP tarafından yollanan bilgilere bakılır herhangi bir problem yaşanmaz ise alıcı SMTP bir onay komutu olan "250 OK" komutunu yollar. Bu kısımda gönderici SMTP sunucu, alıcı SMTP sunucuyu bulmak için domain name system (DNS) sunucuları kullanır.

Örnek olarak Şekil 4'de iki SMTP kullanıcısının basit bir e-posta alış veriş işlemi gösterilmiştir. Gönderici HELO komutu ile alıcıya iletişim kurma isteği yollar. MAIL FROM komutu alıcı tarafta bulunan muratyildiz.com.tr alan adı için yeni bir e-posta başlangıcı olduğunu belirtir. Alıcı bu komutu tamponlarını temizlemek, konum tablolarını sıfırlamak ve mesaja hazırlanmak için kullanır. Sonra RCPT komutu alıcıya murat@muratyildiz.com.tr ve rasit@muratyildiz.com.tr adreslerini verir. murat@muratyildiz.com.tr 250 OK cevabı ile iletişimi kabul ettiğini onaylar. Ancak rasit@muratyildiz.com.tr adresi alıcı sunucusunda olmadığı için 550 cevabı ile iletişimi reddeder. DATA komutu alıcıya, kendisini mesaj içeriğinin takip edeceğini bildirir. 354 cevabı, e-posta girişini başlat ve bunu CRLF CRLF komutu ile sonlandır anlamındadır [12,14].



Şekil 2.3. SMTP E-posta transferi [17].

SMTP sunucuda en yaygın kullanılan komutlar şunlardır [15]:

- HELO: Sunucu kendisini tanıtır.
- EHLO: Sunucu kendisini tanıtır ve genişletilmiş modu talep eder.
-MAIL FROM: Göndereni tanımlar
-RCPT TO: Alıcıyı tanımlar
-DATA: Mesajın gövdesini tanımlar.
-RSET: Bağlantıyı yeniden başlatır.
-QUIT: Oturumu sonlandırır.

Aşağıdakiler alıcı SMTP tarafından gönderilen bazı yanıt kodlarıdır.

• 211 System Status or system help reply
• 220 domain Service ready
• 221 domain Service closing transmission channel
• 250 Requested action OK and completed
• 354 Start mail input; end with .
• 421 Domain service not available, closing connection
• 450 Mailbox unavailable, requested mail action not taken

2.3.2. POP3

Postane Protokolü (Post Office Protocol 3 – POP3) istemci kullanıcı e-postalarının sunuculardan alınmasını sağlayan protokoldür. SMTP protokolü ile yollanan e-postalar

POP protokolü tarafından çekilir.

2.3.3. IMAP

Internet Mesaj Kabul Protokolü (Internet Message Access Protocol - IMAP) kullanıcıların e-posta sunucusuna erişmesini sağlayan ve bu e-posta sunucularından kullanıcıların e-postalarını çekmeye yarayan bir uygulama katmanı protokolüdür. Bu protokolün POP3 protokolünden farkı; POP3 istemci kullanıcının bütün e-postalarını sunucudan çekerken, IMAP protokolü istemci kullanıcının sunucuda bulunan e-postalarının sadece başlık bilgilerini getirir. IMAP 143 numaralı portu kullanır.

2.4. Elektronik Posta Sunucuların Çalışma Prensibi

E-posta sunucularının çalışma prensibi temel olarak şu şekilde çalışır [18].

1. E-posta sunucu, istemciden e-posta gönderme isteği geldiğinde bunu kabul eder.
2. E-posta sunucusu, alıcının alan adına (domain 'ine) ait MaileXchange kayıtlarını DNS sunucudan ister.
3. DNS sunucu, cevabını gönderir.
4. E-posta sunucu, öğrenmiş olduğu MX kayıtlarına ait IP kayıtlarına bağlanır. Bu IP adresleri karşı sunucunun IP adresleridir.
5. Bağlantı kurulduktan sonra e-posta isteği gönderilir. Karşı taraf alıcı gerçekten varsa bunu kabul eder. Yoksa ya da kota aşımı gibi durumlar söz konusuysa bunu hata kodu olarak geri gönderir.

2.4.1. DNS

Internet sayfaları web sunucular sayesinde hizmet vermektedir. Internet kullanıcılarının bir web sayfasına ulaşabilmeleri için o hizmeti sunan web sunucunun IP adresini bilmeleri ve bu IP adresi ile web sayfalarına ulaşması gerekmektedir. IP adresleri genel olarak IPv4 standartlarına göredir. IPv4 32 bit, IPv6 128 bitten oluşan adreslerdir. Günümüzde IPv4 kullanılmakta ancak teknolojinin çok hızlı gelişmesi ve internet kullanıcılarının artması IPv4 aralığındaki IP adreslerini tükenme noktasına getirmiştir. Bu problemi öngören kuruluşlar IPv6 standardını geliştirmiştir.

Internet kullanıcıları bir web sayfasını görüntülemek için o web sayfasının yayında olduğu web sunucu ip adresini bilmeleri gerekmektedir. Internet kullanıcılarının 32 bitlik IP adreslerini ezberlemeleri çok zordur. Bu problemi gidermek ve internet sayfalarına daha rahat erişebilmek için alan adı sistemi Domain Name System (DNS) geliştirilmiştir. DNS sunucular internet adreslerinin IP adres karşılığını tutarlar. Örneğin web tarayıcımızdan www.siirt.edu.tr yazdığımız zaman DNS sunucular web tarayıcımıza siirt.edu.tr adresinin IP karşılığı olan 194.24.146.18 IP adresini geri döndürecektir. Tarayıcımız DNS sunucudan aldığı bu ip adresinin olduğu web sunucuya giderek o internet sayfasını ekrana getirecektir.

IP adresini çözümleme iki şekilde yapılır; özyinelemeli çözümleme ve özyinelemeli olmayan çözümleme. Sorgu özyinelemeli ise, doğrudan sorulan adrese karşılık gelen IP adresi ya da "makine bulunamadı" cevabı verilebilir. Fakat özyinelemeli olmayan (yinelemeli) bir sorguda cevabı bulmak için başka bir isim sunucusunun IP'si verilebilir [19].

2.5. İstenmeyen Elektronik Posta Sistemi

Tanımadığımız kişilerden gelen ve içerik olarak tanıtım, kamuoyu oluşturmak adına gönderilmiş ilgilenmediğimiz e-postalara istenmeyen e-postalar ya da spam e-postalar denilmektedir.

Spam e-postalar merak uyandırma, reklam, korku, eğlence, propaganda, yardım, yasal olmayan duyurular gibi konularla karşımıza çıkmaktadır [15,20]. Spam e-postalar, bilgisayar kullanıcıları tarafından önemli bir güvenlik sorunu olarak kabul edilir.

Spam e-postalar bilinçsiz internet kullanıcılarının çok fazla zamanını çalmakta ve kurum veya kuruluşların sistem kaynaklarının tüketilmesini sağlamaktadır. Spam e-postalar kullanıcıların ilgisini çekecek içeriğe sahip olduklarından özellikle çocuk yaştaki kullanıcıların ahlaki olarak olumsuz etkilenmesine sebep olabilmektedir.

Spam e-postalar sürekli olarak güncel olaylardan hareketle yazılır ve gönderilirler. Spam e-postalar içerik olarak bölgesel ve kültürel alanlarda değişim gösterebilirler. Dünya üzerinde farklı kültürlerin ve dillerin olması spam e-postaların tespit edilmesini zorlaştıran

unsurlardan biridir. Bazen dünya kamuoyunu ilgilendiren bir olayda bütün internet kullanıcılarını hedef alabilen spam içerikler yollanabilir.

İstatistiklere bakıldığında 2012 yılında gerçekleştirilen Euro2012 ve Londra yaz olimpiyatlarının yapılmasından dolayı 2012 yılının ikince çeyreğinde gönderilen spam e-postaların birçoğunun bilgisayarlar kullanıcılarına sahte bilet satmayı hedef aldıkları görülmüştür [21]. Görüldüğü üzere spam e-postalar zamana uygun olarak geliştiricileri tarafından güncellenmektedir.

Spam, 1960'lı yıllarda Amerika'da Geo. A. Hormel and Company şirketi tarafından üretilen, baharatlı domuz eti konservesinin adıdır. 1970 yılında oynanan bir skeç'te, lokantaya yemek yemeğe giden ancak spam yiyeceğinden başka yemek bulamayıp zorunlu bu yemeği yiyen bir çiftin yaşadıkları konu edinmiştir. Bu skeç'ten esinlenerek istenmeyen reklam amaçlı e-postalara da spam adı verilmeye başlanmıştır [22].

Talep Edilmemiş Ticari E-posta (UCE – Unsolicited commercial e-mail), istenmediği halde gönderilen, bir ürünü ya da hizmeti tanıtıcı e-postalardır [23]. UCE'ler gönderici açısından çok az bir harcama ile yollanırlar. Ticari reklam niteliğindeki e-postalardır. UCE'ler servis sağlayıcı kuruma çok büyük maliyetler getirebilirler [24].

Talep Edilmemiş Kitlesel E-posta (UBE - Unsolicited bulk e-mail), aynı anda yüz binlerce kişiye gönderilen içerik olarak ticari olması gerekmeyen e-postalardır. Bu e-postalar ticari içeriklidir. Propaganda yapmak veya kamuoyu oluşturmak amacıyla gönderilen e-postalardır [23].

Spam e-postaların yıllık ortalama maliyeti yaklaşık 200 milyon $'dır. Spam e-postaların küçük bir işletmeye getirdiği maliyeti küçük bir örnek ile açıklayalım.

Küçük bir işletmede çalışan sayısı 10 ve her bir çalışanın ortalama maaşı 1,500 TL olsun. Her bir çalışanın günlük aldığı e-posta sayısı 50 olsun. Alınan e-postaların ortalama %70 i spam e-posta olduğunu kabul edersek , her bir çalışan günlük 35 spam e-posta almış olur. Her bir spam e-postayı silme süresi 5 saniye olarak kabul edelim. Buna göre:

Her bir çalışan için spam postaların silinmesi günde 35 saniye, 360 günde (1 yılda) 17,5 saat olacaktır. Her bir çalışanın günlük 9 saat çalıştığı düşünülürse bu zaman bir çalışanın yaklaşık 2 gününe karşılık gelmektedir ve her bir çalışan için yıllık 97,2 TL maliyet getirmektedir. Küçük işletmemizde 10 çalışan olduğundan spam e-postaların bu işletmeye

yıllık maliyeti 972 TL olacaktır. İşletmemiz e-posta hizmetini kiralık sunuculardan veya kendi sunucusundan veriyor ise bu spam e-postaların internet trafiğine getireceği bir maliyet olacaktır. Spam e-postaların küçük bir işletme için getirdiği maliyete baktığımızda dünya üzerinde hizmet veren irili ufaklı milyarlarca şirketi bu hesaba kattığımızda getirdiği maliyet çok fazla olmaktadır. Bu nedenle spam e-postalar ile mücadelenin kararlı bir şekilde sürdürülmesi gerekmektedir.

2.5.1. İstenmeyen Elektronik Posta Özellikleri

Spam e-postaların ayırt edici tipik özellikleri aşağıdaki şekilde özetlenebilir [14].

- Birden fazla alıcıya aynı içerik gönderilir.
- Tanıtım amacıyla gönderilirler.
- Genellikle içerikleri yanıltıcıdır.
- Dini inanç ya da insani duyguları konu edebilirler ve e-postanın belli bir sayıda kişiye iletilmesini isteyebilirler.
- Gönderen, kime gibi adres bilgileri uygun biçimde değildir ve genellikle rastgele sahteleri üretildiğinden harf hatalarına oldukça sık rastlanır.
- E-posta mesaj başlık bilgileri tahrip edilmiş olur ve bu sebeple geriye dönük izleme zor olur.
- Alıcıların bu dağıtımdan e-posta almak istemediklerini belirtebilecekleri geçerli veya fonksiyonel bir geri dönüş adresi bulunmaz.
- Genel olarak içerikleri güncel konulardır.

2.5.1.1. Elektronik Posta Adresleri Nasıl Elde Edilir?

Spam e-posta yollamak için öncelikle e-posta adreslerine ihtiyaç duyulur. Bu e-posta adreslerini temin etmek için spam yollayanlar tarafından tercih edilen en önemli yol web

robotlarıdır. Web robotları internet sayfalarını arama motoru şeklinde tarayarak internet siteleri içinde bulunan e-posta adreslerini bulup bunları kendi veri tabanlarına kaydederler.

Diğer bir yöntem zincir e-postalardır. Spam yollayıcıların belirli bir konu veya önemli günlerde insanların duygularını ajite edecek şekilde yolladıkları ve kişilerin bundan etkilenerek diğer arkadaşlarına yolladıkları zincirleme e-postalardır. Bu yöntemi kullanan spam geliştiricileri özellikle insanların hassasiyetlerinden hareket ederek kişilerin duygularını istismar ederler. Örneğin dini konularda bazı rivayetlerde bulunarak bu iletilerin yönlendirme edilmesi gerektiği aksi takdirde başına kötü olayların geleceğini söyleyerek, Photoshop grafik programı ile yüzü yanmış bir çocuk için para yardımında bulunarak, kan bağışı talebinde bulunarak veya vatanperver insanların hassasiyetlerini kullanarak bu e-postaların kişinin e-posta listesinde bulunan diğer kullanıcılara iletmesini sağlarlar. Verilen örnekler genel olarak ülkemize özgüdür. Spam geliştiricileri yaşadıkları toplumun hassasiyetlerine göre spam içeriklerini oluştururlar.

E-Posta elde etme yöntemlerinden bir diğeri alan kayıtlarıdır. Kişilerin sahip oldukları internet siteleri alan kaydı yaptıkları sırada zorunlu olarak e-posta adresini vermek zorundadır. Bunu bilen spam yollayıcılar bu e-postaları elde etmek için whois yöntemini kullanarak internet sayfası sahiplerinin e-posta adreslerine rahatlıkla ulaşabilirler.

Bu yöntemlerin dışında e-posta sunucularına saldırarak e-posta hesapları elde edilir veya kullanıcı bilgisayarına bulaşan virüsler ile bu bilgiler elde edilebilir.

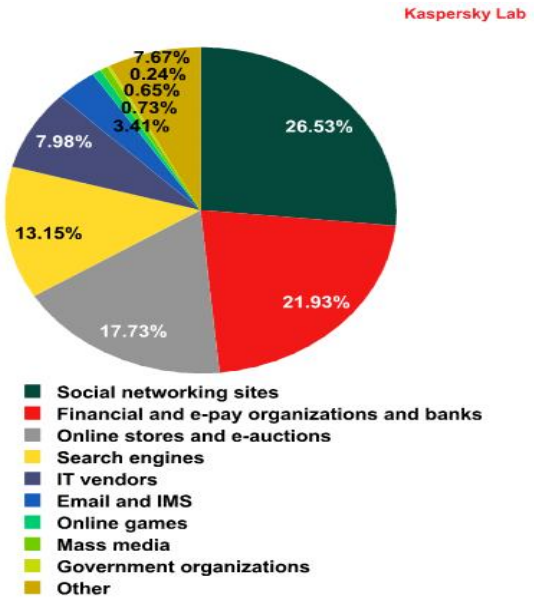
2.5.1.2. Phishing

Phishing spam şeklinde yayılan bir tuzak sistemidir. Teknik olarak spam e-postalara benzemekle birlikte içerik olarak farklı olduğundan spam e-postaya göre daha ağır sonuçları olmaktadır [25].

Phising daha çok banka ve e-bay gibi alışveriş sitelerini taklit ederek, resmi e-posta izlenimi vererek, e-posta alıcısından siteye giriş bilgilerinin teyit edilmesini ister. Alıcıda e-postayı gerçek zannederek siteye yönlenir ve giriş bilgilerini girer. Böylece saldırgan kurbanın bütün bilgilerini elde ederek kullanıcı hesaplarında dilediği işlemi yapabilecektir[26]. Ülkemizde geçmiş yıllarda Phising yöntemi kullanılarak çok fazla

insan mağdur edilmiştir. Günümüzde internet bankacılığına getirilen SMS mesaj kodları sayesinde internet bankacılık işlemleri çok daha güvenli bir hale gelmiştir.

Anti virüs firması Kaspersky Lab tarafından Kasım 2012 yılında yayınlanan analize göre Phishing yöntemiyle hedeflenen ilk 100 şirket aşağıda gösterilmiştir [26].



Şekil 2.4. Phishing yöntemi ile 2012 yılında hedeflenen ilk 100 şirket [26].

Phising yöntemi sadece bankacılık işlemlerinde kullanılan bir yöntem değildir. Saldırganlar kullanıcı hesapları ile ilgilendikleri için bu hesaplar sadece parasal hesaplar olmayabilir. Örneğin sosyal ağlar içerisinde en popüler konumda olan Facebook hesabını ele geçirmek isteyenler, kurbanına Facebook domain adından değil de Facabook domaininden bir e-posta yollayarak sisteme giriş yapmasını sağlayacak bir spam e-posta yollayabilirler. Kurbanın linke tıklaması ile Facebook web sitesinin birebir kopyası olan Facabook web sayfasına yönlendirme işlemi gerçekleşir. Bu ince noktayı fark edemeyen ve açılan sayfaya hesap bilgilerini giren kurbanların hesap bilgileri saldırganlar tarafından ele geçirilecektir. Ele geçirilen hesaptan saldırgan hedeflerine göre kişinin arkadaşlarından kontör, borç para veya farklı taleplerde bulunabilir.

PWG (Anti phishing work group) tarafından Kasım 2005 de yayımlanan bildiri göre Phishing yapılırken kullanılan sahte siteler ortalama 5,5 gün çevrimiçi kalmakta ve sonra kendilerini yok etmektedirler. Bunun nedeni kurbanların hukuki olarak hak aramalarını engellemektir [25].

2.5.2. İstenmeyen elektronik posta istatistikleri

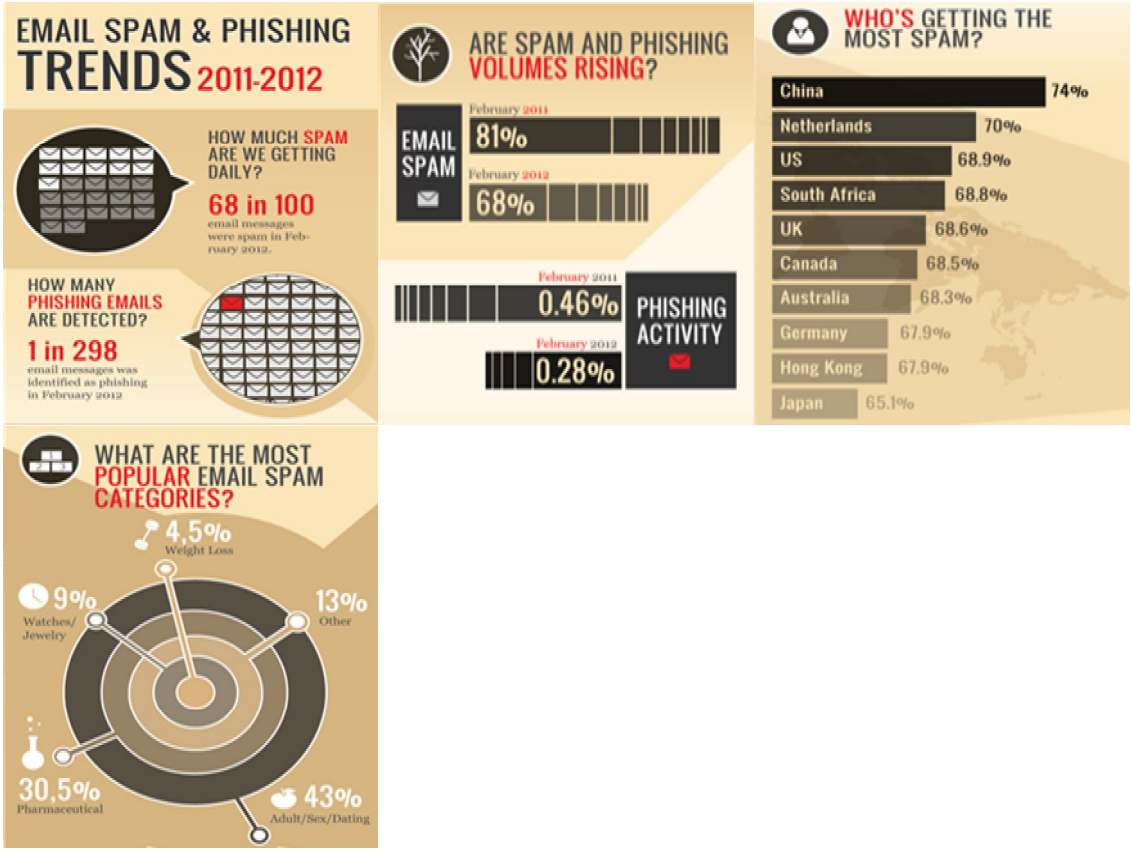
Spam e-postalar gelen kutusuna düştüğü vakit bu e-postaların okunması ve silinmesi zaman kaybına neden olmaktadır. Spam e-postaların birçoğu güncel konuları içerdiği için bu e-postalar kullanıcıların dikkatini çekmektedir. Kullanıcılar, dikkatini çeken e-postalarda bulunan linklere tıklayarak daha fazla detay öğrenmek isteyebilir. Bu durumda tıkladıkları web sitesi içerisindeki zararlı programlara maruz kalabilirler. Internet kullanıcıları için düşünüldüğünde spam e-postalar zaman kaybına neden olmakla birlikte kişilerin zihinlerinin farklı konular ile meşgul olmasına sebep olduğundan özellikle işyerlerinde performansın düşmesine neden olmaktadırlar.

Symantec Şubat 2012 istihbarat raporuna göre en fazla spam'a maruz kalan ülke Çin'dir. Çin'li kullanıcılar tarafından alınan e-postaların %74'ü spam e-postalardan oluşmaktadır. Hollanda ve ABD sırasıyla %70 oranında spam e-postalara maruz kalmaktadır. Spam e-postalar içerisinde yetişkinlerle ilgili mesajlar en fazla ilaç, güzellik ürünleri, kilo kaybı konularını içermekte ve bu spam e-postaların %43'ünü oluşturmaktadır [27].

Genel olarak istatistiklere bakıldığı zaman spam e-posta oranında geçen yıllara göre bir düşüş görülmektedir. Bu bize spam ile mücadelede yürütülen kararlılığı göstermektedir. Şekil 1.5'te görüldüğü gibi [27]:

1. 2011-2012 yılları arasında her 100 e-postadan 68'i spam e-postalardan oluşur.
2. Spam ve Phishing değer riski 2011 Şubat ayında %81 iken 2012 Şubat ayında %68'e gerilemiştir.
3. Phishing aktiviteleri 2011 Şubat ayında %0,46 iken 2012 Şubat ayında %0,28 oranına düşmüştür.
4. Spam e-postalara en fazla maruz kalan ülkeler sırasıyla Çin, Hollanda, ABD, Kuzey Afrika şeklindedir.

5. En popüler spam e-posta kategorileri sırasıyla %43 oranıyla Pornografik, %30,5 oranında ilaç, %13 oranında Güzellik, %4,5 oranında kilo kaybı ve %13 diğer şeklinde sıralanabilir.



Şekil 2.5. Spam e-posta istatistiği [27].

2.5.3. Spam e-postaları engelleme teknikleri

2.5.3.1. Eğitim

Bu yöntem spam e-postaları engellemek için teknik bir yöntem değildir ancak spam e-postaları engelleme yöntemleri arasında bulunması gereken en önemli yöntemlerden biridir. Bu yöntemde bilgisayar kullanıcılarının spam e-postalar konusunda bilgilendirilmesi gerekmektedir. Spam e-postaların zararları ve spam e-postaların engellenmesi için basit yöntemlerin anlatılması özellikle spam geliştiricilerin zincirleme e-postalardan istediği verimi alamamalarına neden olacaktır. Bu yöntem ile:

- Spam geliştiricilerini, insanların hassasiyetlerini kullanma yönteminden vazgeçirebilir.
- Kullanıcılara güvenlik noktasında yardımcı olabilir.
- Internet kullanıcılarının ahlaki olarak zarar görmesini engelleyebilir.
- Zaman kaybını azaltabilir.
- İnsan zihninin farklı konular üzerinde meşgul olmasını engelleyebilir.
- Özellikle çalışanlar için performans düşmelerini engelleyebilir.

Spam e-postalardan kaçınmak için aşağıdaki kurallara dikkat etmek gerek [28]:

- Spam e-postalar yanıtlanmamalı
- Spam e-posta üzerindeki linkler tıklanmamalı
- E-postalar text olarak okunmalı (bu sayede resim ya da link olarak gelen tehlikelerden koruna bilinir.)
- Outlook tarzı bir e-posta client kullanılıyor ise tüm e-postayı açmadan önce spam ya da bilinmeyen kişilerden gelen e-postaların özeti okunmalı,
- Gelen e-postanın header bilgisine bakıldığında "FROM" ve "REPLY TO" adreslerinin aynı olup olmadığına dikkat edilmeli
- E-postanın, yönlendir yolu ile zincir e-posta olmamasına dikkat edilmeli, iletilmek istendiğinde önceki e-posta adresleri silinmeli
- Sözlük tipteki spammer saldırılarından korunmak için mümkünse zor e-posta adresleri seçmeli
- Herhangi bir e-posta hesabını açmak istediğinizde sağlayıcının bildirim ve şartları dikkatlice okunmalı
- Forum vs. yerlerde kullanılmak üzere bir e-posta hesabını kullanarak gerçek e-posta adresinizin korunmasına dikkat edilmeli
- E-posta adresinizi forum, portal vs. yerlerde kullanıcı adı olarak kullanmamalı,
- E-posta adresinizin belirtilmesi gereken durumlarda isimATdomainDOTcom şeklinde bir yöntem ya da bir link üzerinden verilmesi gibi bir yöntem seçerek e-posta adresini doğrudan bir yazılım vasıtası ile kolayca alınmasının önüne geçilmeli
- Bilgisayarda muhakkak bir spam filtresi/spam blocker ile Anti-spam kullanılmalı.

2.5.3.2. Ters DNS kaydı

Ters DNS kaydı gönderilen bir e-postanın kaynak IP'si üzerinde yapılan DNS sorgulamasıdır. Bu yöntem e-postanın gerçek kullanıcısından gelip gelmediğini belirlemek içindir. Bu yöntemde gelen e-postanın kaynak IP'sinin hangi alan adına ait olduğu belirlenir ve bu alan adının, gelen e-postadaki FROM adresi ile eşleşip eşleşmediği kontrol edilir. Eğer eşleşme varsa DNS bilgileri doğrulanmış olur ve e-posta kabul edilir. Aksi takdirde e-posta göndericisinin DNS bilgileri geçersiz kabul edilerek e-posta reddedilir. Ters DNS sorgulaması, iki makine arasındaki SMTP aktarımının selamlaşma aşamasında gerçekleşir [14].

2.5.3.3. DNS MX araması

DNS MX araması, Ters DNS arama yönteminden farklı olarak gelen e-posta adresi içerisindeki FROM kısmında yazılı olan e-posta hesabının geçerliliğini kontrol etmek için e-posta adresi içindeki domain adresi üzerinden MX sorgulaması yapar. Eğer domain adresi geçerli bir MX kaydına sahip değilse gelen e-postanın geçersiz olduğu tespit edilir.

2.5.3.4. Kara listeler

Kara listeler yaygın kullanılan ve basit bir spam engelleme yöntemidir. Bu yöntemde spam e-posta üretildiği bilinen veya spam e-postaların üretildiği IP adreslerinin bazı kurumlar tarafından bir listeye dâhil edilmesi işlemidir. Bu şekilde bu IP adreslerinden gelen e-postalar engellenmiş olur.

İstenmeyen e-postacılar düzenli olarak IP adreslerini değiştirdiğinden ve geniş aralıkta IP adresleri kullandığından, kara listeler kısa zaman aralıklarında, küçük istenmeyen e-posta miktarlarını engellemede çok etkindir, fakat genel bir istenmeyen e- posta filtrelemede etkin değildir [15].

Kara listelerin tersi beyaz listelerdir. Bu yöntem kara listede belirtilen avantaj ve dezavantajlara sahip olduğundan her iki yöntemde çok sağlıklı değildir.

2.5.3.5. Gerçek zamanlı kara listeler

Gerçek zamanlı kara listeler (RBL - Realtime Black List), DNSBL (DNS-based Blackhole List) olarak bilinir. Bu yöntemin kara liste yönteminden farkı spam yollayan IP adreslerinin belli bir süreliğine kara listede tutulması ve daha sonra serbest bırakılmasıdır.

Organizasyon veya kurumlar tarafından tutulan RBL'ler, gelen e-postanın IP adresini ters DNS kontrolüne benzer biçimde kendi listelerinde sorgulayarak, listede olması durumunda gelen e-postayı reddetmektedir. Sürekli güncellenen RBL servisleri bazen hatalı çıkarımlar yapabilir. Örneğin, tüm dünyadaki IP adreslerini listelerinde tuttukları için genellikle IP adreslerini tek tek tutmayan birçok RBL servisi, 256'lık bloklar halinde bulunan bir IP sınıfı (C sınıfı) içerisinde çok sayıda IP adresi üzerinden spam gönderildiğini tespit ederse, o IP sınıfını tümüyle kara listelerine alacaktır. Buda IP sınıfı içerisinde, spam olmayan normal e-posta gönderimlerinin yapıldığı IP adreslerinin engellenmesine sebebiyet verecektir [29].

2.5.3.6. Kelime filtreleme

Kelime filtreleme yöntemi spam e-postaların engellenmesinde kullanılan en etkili yöntemlerden biridir. Bu yönteme göre istenmeyen kelimelerden oluşan bir kara liste tanımlanır. Gelen e-postalar içerisindeki kelimeler bu kara liste ile karşılaştırılır. Eğer benzer kelimeler var ise gelen e-postanın spam olduğu tespit edilir.

Yukarda anlatılan kelime filtreleme yöntemi bu şekilde kullanıldığında çok sağlıklı çalışmayacağı aşikârdır. Örnek olarak webcam kelimesini istenmeyen kelimeler grubuna eklediğimiz vakit bunu engellemek için spam geliştiricileri "web cam", "WebCAM", "W eb Cam", "web C A M" şeklinde değiştirerek bu yöntemi atlatabilirler. Bu yöntemin zayıf tarafını bilen spam engelleme yazılımları bu kelimelerin farklı formatlarını düzenli olarak güncelleyerek bu yöntemin etkili kullanmasını sağlayabilirler. Ancak bu yöntem tek başına etkili olamaz.

2.5.3.7. Kural tabanlı engelleme

Bu yöntem puanlama sistemine dayanarak geliştirilen bir yöntemdir. Bu yönteme göre gelen e–postanın başlık kısmı, konu başlığı ve içeriği önceden tanımlanmış bir kurallar dizisine tabi tutulur. Her bir adımın bir puanı mevcuttur.

Bu yöntemi en fazla kullanan spamassassin yazılımında örneğin e-postanın konu başlığı boş veya tamamı büyük harflerle yazılmış ise, e-posta içeriğinde çok fazla html etiketi mevcut ise, e-posta birden fazla kişiye yollanmış ise, Ters DNS, DNS MX araması sorgularından dönen sonuca göre ve farklı kurallara göre farklı puanlar verilir. Bu puanların toplamı belli bir eşik değerin üzerinde veya altında olmasına göre e-postanın spam olup olmadığına karar verilir.

3. BAĞIŞIKLIK SİSTEMİ

Bağışıklık sistemi; Vücuda giren yabancı maddelerin etkisiz hale getirilmesi, dışarıya atılması veya kimyasal bir değişime tabii tutulması için vücudun oluşturduğu fizyolojik mekanizmalara denir [1]. Bağışıklık sistemi, vücut hücreleriyle yabancı hücreleri ayırt etmek konusunda çok gelişmiş bir yeteneğe sahiptir.

İnsan vücudu, çevresinde bulunan birçok organizmanın saldırısına uğrar. Bağışıklık sisteminin temel görevi bu organizmaların vücuda girmesini engellemektir. Şayet bu zararlı organizmalar vücuda girmiş ise bağışıklık sistemi bu organizmaları vücuda girdikleri yerde yutarak, yayılmalarını engellemek veya vücutta yayılmalarını geciktirmektir. Bağışıklık sistemi insan vücudunu enfeksiyona yol açabilen mikroorganizmaların zarar verici etkilerine karşı korur. Bağışıklık sistemi bu görevini, canlının yaşamı boyunca devam ettirir.

Bağışıklık sistemi; kendisine yabancı farklı yapıdaki çeşitli düşmanları tanıyıp ayırt edebilme yeteneğine sahiptir. Bu özellik bağışıklık sisteminin hafıza özelliğinden gelmektedir. Hafıza özelliği sayesinde bağışıklık sistemindeki hücreler, ilk karşılaştığı yabancı hücreyi görür, belleğine kaydeder ve daha sonra gördüğünde savunma sistemini hızlıca devreye sokar.

Bağışıklık sistemimizin vücudumuzu savunmasındaki başarısının sırrı; vücudumuzun çok farklı özellikteki ve dinamik yapıdaki iletişim ağına sahip olmasıdır. Milyonlarca hücre, bilgileri arkadan ileriye doğru iletir. Uyarıyı alan hücreler uyarının tipine göre şekillenerek çok güçlü kimyasallar üretirler. Bu maddeler hücrelerin kendi büyüme ve hareketlerini düzenlemelerine izin vererek vücut savunmasını başlatır [1].

Canlı öldüğünde bağışıklık sistemi çöker. Bu nedenden dolayı vücudumuza çok çeşitli bakteriler, parazitler ve mikroplar istila etmeye başlar. Şayet canlı yaşarken bağışıklık sistemi bozulur veya çökerse canlı vücudu; birçok alerji, enfeksiyon gibi ölümcül hastalığın gündeme geldiği durumla karşılaşabilir.

3.1. Bağışıklık Sisteminin Çalışma Prensibi

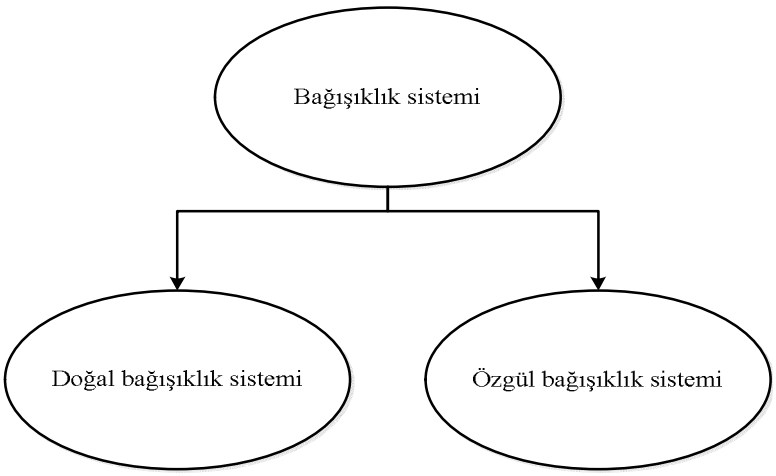
Canlı vücudunu farklı savunma sistemleri korumaktadır. Dış dünyadan vücuda giren ve vücudun tanımadığı moleküllere antijen denilir. Antijenler canlı vücudundaki bağışıklık

sistemini harekete geçirirler. Bu tarz bir saldırıda öncelikle antijenlerin vücuda girmesini engelleyici yüzey engelleri devreye girer. Yüzey engelleyicilere örnek olarak göz, deri, solunum, sindirim sistemi verilebilir. Yüzey engellerini aşan antijenler ikinci savunma sistemi olan doğal bağışıklık sistemi ile karşılaşırlar.

Doğal bağışıklık sisteminde fagositler, makrofajlar ve lenfositler devreye girer. Fagositler ve makrofajlar öncü hücreler olduğundan ilk aşamada antijenleri yok etmeye çalışırlar. Vücuda girebilen birçok organizma bu yöntem ile yok edilmeye çalışılır [30].

İkinci koruma sistemini aşan antijenler üçüncü savunma sistemi olan özgül bağışıklık sistemi ile karşılaşırlar. Üçüncü savunmada B ve T lenfositler devreye girer. Antijen saldırısını haber alan T hücreleri harekete geçerek diğer savunma hücrelerini tetiklerler [31].

Bağışıklık sistemi dış yüzey hariç 2 kısımda incelenir. Bu kısımlardan ilki doğal bağışıklık sistemi, diğeri ise özgül bağışıklık sistemidir.



Şekil 3.1. Bağışıklık sistemi

3.2. Doğal Bağışıklık Sistemi

Doğal bağışıklık sistemi vücudumuzdaki ilk yabancı materyali karşılayan savunma sistemidir. Bu sistem doğuştan gelir. Doğal bağışıklık sistemi doğuştan itibaren birçok yabancı maddeyi tanıma özelliğine sahiptir. İlk karşılaşmada birçok mikroorganizmayı yok

edebilir. Erken savunma sistemi olarak adlandırılabilir. Özgül bağışıklık sistemi, aktivasyon için antijenin kaynağı ve verilecek yanıtın tipi konusunda bilgiye gereksinim duyar. Doğal bağışıklık sistemi özgül bağışıklık sisteminin ihtiyaç duyduğu parametreleri yollar. Kendisinden sonra devreye girecek olan özgül bağışıklık sistemini hem uyarmakta, hem de yönlendirmektedir. Doğal bağışıklık sisteminde bellek özelliği yoktur.

Doğuştan gelen bağışıklık sistemi, birçok canlıda konağın korunmasını sağlar [34].

Doğal bağışıklık sisteminin önemli bir bileşeni "complement" olarak bilinen bir kan proteini sınıfıdır. Complement antikorların aktif olmasına yardım etmektedir. Doğal bağışıklık, pattern associated molecular patterns (PAMPs) olarak adlandırılan mikrobik patojenlerle birleştirilmiş moleküler örüntüleri tanımak için pattern recognition receptors (PRRs) olarak bilinen germinal merkezde kodlanmış alıcıların bir kümesi esasına dayanmaktadır. PAMPs sadece mikroplar tarafından üretilir ve asla organizma tarafından üretilmez bu yüzden PRRs ile onların tanınması, patojenik maddelerin bulunduğuna işaret eden bir sinyalle sonuçlanabilir. Bu şekilde bağışıklık tanıması ile ilgili yapılar bulunduğu vücuda zarar vermekten kaçınmak için özel hücre ve moleküllerden tamamen bağımsız olmak zorundadır. Bu mekanizmanın sonucu olan doğal bağışıklık aynı zamanda self/nonself ayırım yeteneğine sahiptir ve özgül bağışıklığın desteklenmesinde rol oynar [2].

Patojenler vücuda dışardan giren zararlı organizmalardır. Antikorlar patojenlere karşı vücudu koruyan proteinlerdir.

Doğal bağışıklık tanımasının en önemli aşaması, özgül bağışıklık tepkisinin başlamasını sağlayan T hücrelerinin aktivasyonunda rol oynayan antijen sunan hücrelerde (ASH'ler) uyarıcı sinyal ifade etmesidir. Bu şekilde doğal bağışıklık tanıması olmaksızın özgül bağışıklık tanıması, özgül tanımada reseptörleri ifade eden lenfositlerin negatif seçimi ile sonuçlanabilir [2].

3.3. Özgül Bağışıklık Sistemi

Bu bağışıklık sisteminin birçok farklı ismi mevcuttur. Özgül, edinsel, kazanılmış bağışıklık sistemi olarak farklı isimlerde tanımlanır. Biz bu çalışmada özgül bağışıklık adını kullanacağız. Özgül sistem yabancı madde vücuda girdiğinde harekete geçer. O

maddeye göre uyarlanarak gelişir. Özgül bağışıklık sistemi uyarı sonrası gelişir ve etki gösterir. Temel özelliği özgünlük ve bellektir. Bu nedenle aynı madde ile tekrar karşılaştığında çok hızlı bir şekilde etki gösterir. Kendinden olanı olmayandan ayırma özelliği mevcuttur.

Özgül bağışıklık daha önce saldırganla hiç karşılaşmamışsa bile vücudun her hangi bir mikrobu tanımasını ve tepki vermesini sağlar [2].

Özgül bağışıklık sistemi karşılaştığı yabancı maddeye karşı vücudu 3 aşamada savunur [35].

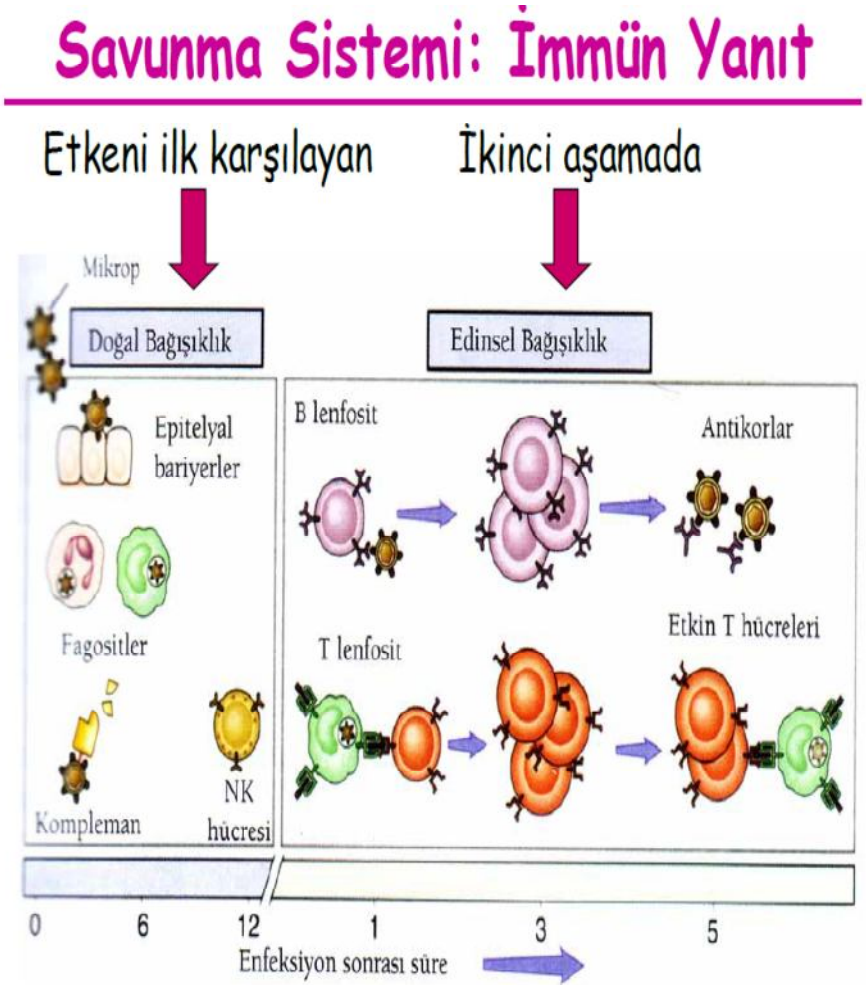
1. Yabancı olduğunu algılar.
2. Bunlara karşı savunmayı sağlar.
3. Bir daha unutmamak üzere belleğine kaydeder.

Vücut her farklı antijene karşı özel antikor üretir. Üretilen antikorlar bu farklı antijenleri tanır ve aynı antijen vücuda tekrar girdiğinde bu antikorlar devreye girer. Antijenler için üretilen özel antikorların yanıt verecek düzeye gelmesi bazen günler alabilir. Bu antikorlar antijenlere yaptıkları saldırıları ve sonraki enfeksiyonları önlemek için belleğe kaydederler.

Tablo 3.1. Doğal ve özgül bağışıklık sisteminin özellikleri

Doğal bağışıklık sistemi	**Edinilmiş bağışıklık sistemi**
Yanıt özel ve tekil değil.	Patojen ve antijene özgü yanıt verilir.
Hızlı ve en büyük seviyede yanıt verilir	Daha önce karşılaştıysa hızlı cevap verir. Hiç karşılaşmadığı durumlarda gecikmeli cevap verir.
Belleği yoktur	Belleği vardır.
Bütün canlılarda vardır.	Çenelilerde vardır.

Canlı vücudunda bulunan ve bağışıklık sisteminde kendinden olmayan (non-self) moleküllerden ayrılabilen bileşenlere kendinden olan moleküller (self) denir. Doğuştan bağışıklık sistemi ve özgül bağışıklık sistemi self ve non-self moleküllerin ayrımına bağlıdır. Kendinden olmayan moleküllerin bir sınıfı antijendir ve bunlar yabancı moleküller olarak adlandırılır. Antijenler özgül bağışıklık almaçlarına bağlanıp bir bağışıklık yanıtının oluşmasına neden olan maddelerdir [36,37].



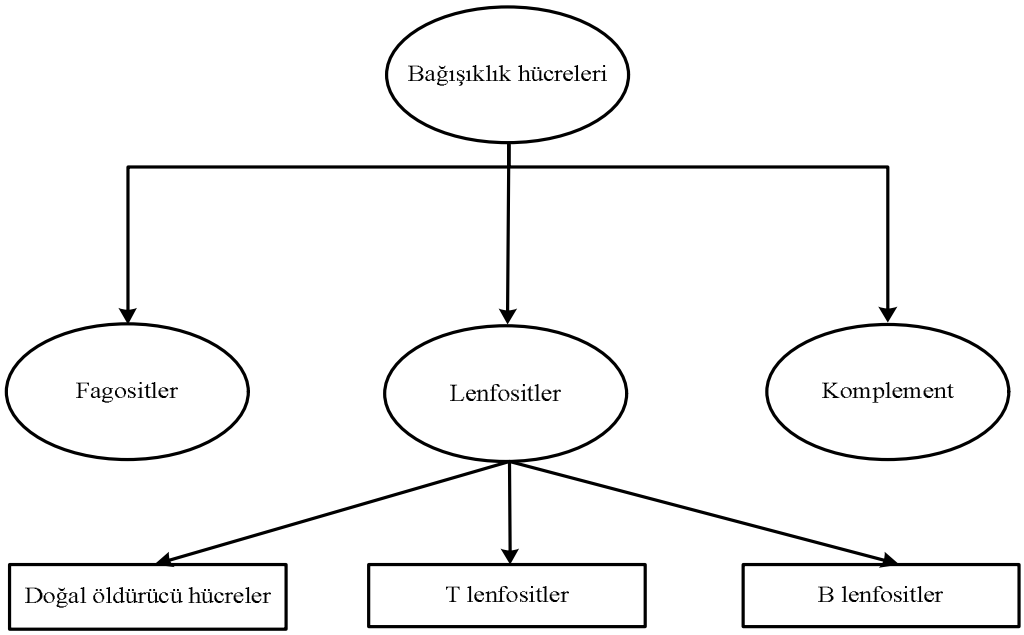
Şekil 3.2. Doğal ve özgül savunma sistemi

3.4. Bağışıklık Sistemindeki Temel Birimler

Bağışıklık sisteminde hücreler 3 kısma ayrılır.

1. Lenfositler
2. Fogositler
3. Komplement

Bağışıklık hücrelerinin sınıflandırılması Şekil 2.3 te gösterilmiştir.



Şekil 3.3. Bağışıklık hücreleri

Lenfositler vücuda giren yabancı maddelerin (Antijen) uyarısıyla farklılaşarak bağışıklık cevabını oluşturur ve antijenleri yok eder. Fagositler patojenleri yutmalarından veya emmelerinden dolayı bu şekilde adlandırılmışlardır. Genellikle vücutta patojenleri ararlar. Sitokonlar tarafından özelleşmiş bölgelere çağrıla bilirler [37].

3.4.1. Lenfositler

Lenfositler vücuda giren antijenlerin uyarısıyla farklılaşarak bağışıklık cevabını oluşturan asıl birimlerdir.

Lenfositler genel olarak T,B ve Doğal öldürücüler şeklinde sınıflandırılır. Dokuya göre değişmekle birlikte genel olarak T hücreler lenfositlerin %80'ini geriye kalan %20'sini eşit bir şekilde B hücreler ve doğal öldürücüler oluşturur [38].

3.4.1.1. T lenfositler

T hücresi oluşturmak için bağışıklık hücreleri kemik iliğindeki olgunlaşmamış kök hücrelerinden salındıktan sonra timus'a geçer ve timus'ta olgunlaşırlar. Olgunlaşan

hücreler timus'tan ayrılarak kanda ve lenf sistemlerinde dolaşır. Farklı görevleri yerine getirmek için çeşitli T hücreleri vardır. T hücreleri çeşitlerinden öldürücü T lenfositler virüs taşıyan hücreleri tespit edip o hücreleri yok etmeye çalışır. Diğer T hücreleri benzer şekilde öldürücü lenfositleri destekleyen görevleri yerine getirirler [39].

T hücreleri timüste negatif seçim aşamasından geçer. Bu aşamada vücudun kendi hücreleri ile eşleşmeyen T hücreleri kemik iliğinden serbest bırakılır. Geriye kalan hücreler ise burada yok edilir. Bu şekilde hayatta kalan bütün T hücreleri kendinden olmayan hücreleri tanıyacaktır. T hücrelerinin bu şekilde elenmesi vücudun bağışık sistemini kendi proteinlerinin saldırılarından korur. Bu şekilde T lenfositler timusta geliştikleri sırada kendisine yabancı olanı ayırt edebilme yeteneğine sahip olurlar. Bu yapıdan yararlanılarak T hücrelerinin negatif seçimine dayalı bir kural dışı tespit algoritması geliştirilmiştir [40].

T lenfositler immünolojik bellekten ve hücresel bağışık yanıttan sorumludurlar. T lenfositler taşıdıkları belirleyicilere göre CD4 veya CD8 olmak üzere 2 alt gruba ayrılır. CD4 yardımcı lenfositler, bağışık yanıtı başlatır veya yardım eder, CD8 ise sitotoksik etkilidir.

3.4.1.2. B lenfositler

B lenfositler antijene özel antikor üretir. Vücuda herhangi bir antijen girdiğinde, B lenfositler milyonlarca antikor üreterek çoğalır [39].

Antikor diğer adıyla immünglobulin serum ve doku sıvılarında bulunan bir grup glikoproteindir. Antikor belirli bir antijene yönelik immünglobulindir. Antikorlar B lenfositleri tarafından antijene karşı yapılırlar ve yapılmalarına neden olan antijene özgül bir şekilde bağlanırlar [41].

İmmünglobulinler (Ig, antikor) vücutta immün sistem tarafından (B lenfositler, plazma hücreleri) virüsler, bakteriler, mantarlar, kanser hücreleri gibi yabancı maddelere karsı üretilen proteinlerdir. Bu yabancı maddelere saldırarak immün sistemin onları yok etmesini sağlar ve vücuda zarar vermesini önlerler. Her antijene (vücuda yabancı madde) karşı, ona özgü Ig üretilir. Ayrıca alerjik olaylarda (IgE'ler) rol oynadıkları gibi bazen de otoimmün hastalıklarda olduğu gibi vücudun kendi doku ve organlarına da saldırabilirler. Tüm Ig'ler birbirlerine disülfid bağıyla bağlı iki ağır ve iki hafif zincirden oluşur. 5 çeşit ağır zincir (Gama, mü, alfa, delta ve epsilon) bulunmasına karşılık yalnızca 2 çeşit hafif zincir vardır

[42]:

Kappa ve Lambda. Ağır zincirler Ig tipini belirler, adlarını da onlardan alırlar.

Monoklonal gamopatilerde genellikle bir hafif zincir yüksek bulunur ve kappa/lambda hafif zincir oranı bozulur. Vücut ne kadar az Ig üretirse direnç düşeceğinden tekrarlayan enfeksiyonlar o kadar sık görülecektir. Az üretim doğuştan olabileceği gibi, kanserlerde olduğu üzere sonradan da olabilmektedir. İnsan vücudunda 5 tip major Ig tipi mevcuttur [42].

3.4.1.2.1. IgA

Burun, trake gibi solunum yolları, sindirim sistemi mukozası, kulak, tükürük bezi, göz yaşı, kan ve vajinada bulunur. Bebeklerde 6 aylığa ulaşıncaya dek yeterince IgA yoktur. Kandaki Ig'lerin yaklaşık %10-15'i IgA'dır. Görevleri vücut yüzeylerini dış ortamdaki yabancı cisimlere karşı savunmaktır [42].

3.4.1.2.2. IgG

Tüm vücut sıvılarında ve kanda bulunur. Vücuttaki en küçük ama en sık bulunan (tüm Ig'lerin %75-80'i) antikorlardır. IgG1, IgG2, IgG3 ve IgG4 olmak üzere 4 subtipi bulunmaktadır. Özellikle bakteri ve virüslerde savaşta çok önemli rol oynarlar. Gebelikte anneden fetüse geçebilen tek antikor IgG'dir. Böylece bebeği koruma rolleri de vardır [42].

3.4.1.2.3. IgM

En büyük boyutlu ve enfeksiyona karşı ilk önce üretilen antikorlardır. Vücuttaki tüm Ig'lerin yaklaşık % 5-10'u IgM'dir. Kan ve lenfoid sıvıda bulunurlar. Hümoral immünitede görevlidirler. Diğer immün sistem hücrelerinin de yabancı cisimleri yok etmesini sağlarlar [42].

3.4.1.2.4. IgE

Akciğer, deri ve müköz membranlarda bulunurlar. Polen, mantar sporları, hayvan epitelleri gibi yabancı cisimlerle savaşta rol oynarlar. Süt ve ilaç alerjilerinde, parazit

enfeksiyonlarında, bazı zehirlenmelerde de ortaya çıkarlar. Alerjik bireylerde kanda genellikle seviyesi yüksektir [42].

3.4.1.2.5. IgD

Karın ve göğüs gibi dokularda çok az miktarda bulunur. Kanda minimal düzeydedir (Kandaki tüm Ig'lerin % 0,25'i). Olgunlaşmamış B lenfositlerinin yüzeyinde bulunurlar. Görevi tam olarak çözülememiştir. B lenfositlerinin plazma hücrelerine dönüşmesi için bir antijenik reseptör görevi görmektedir. Hümoral immün yanıtın artırılmasında rol oynarlar. Son zamanlarda yapılan çalışmalarda bazofil ve mast hücrelerine bağlanıp onları aktive ederek antimikrobiyal madde salınımlarını uyardığı ve respiratuvar immün savunma sistemini güçlendirdiği saptanmıştır [42].

3.4.1.3. Doğal öldürücüler (NK)

Doğal öldürücü hücreler kan kemik iliği ve dalakta bulunur. En önemli görevleri, normal hücrelere zarar vermeden tümör hücrelerini ve virüs taşıyan hücreleri öldürmektir [39].

3.4.2. Fagositler

Fagositler hücre yiyiciler olarak ta bilinir. Bazı fagositlerin lenfositlere antijenleri sunma (APC - Antigen Presenting Cells) özelliği vardır. En önemli fagositler, monosit ve makrofajlardır. Monositler kanda dolaşarak dokulara yerleşir ve makrofaj (Büyük Yiyiciler) haline gelirler. Makrofajlar antijenleri parçalayıp sindirdikten sonra T lenfositlerine sunarlar. Bağışıklık reaksiyonun başlamasında önemli rol oynarlar [43].

3.4.3. Komplement

Vücuda giren canlı hücrelerin erimesine yol açan antikorların daha etkili bir şekilde çalışmasına yardımcı olmakla görevlidirler. İstilacı organizmanın yüzeyine bağlanarak lezyon oluşturarak fagositlere yardımcı olur [44].

Komplement sistem, kanla birlikte dolaşan ve antikorların işlevini tamamlayıcı özellikler gösteren bir plazma proteinleri gurubudur. Tamamlayıcı sistem istilacı bir organizma tespit ettiğinde, her bileşeni bir zincirleme tepkimeyi körükleyici davranış gösterirler. Bunun sonucunda istilacı organizmanın hücre duvarına bağlanan protein kompleksleri lezyonlar oluşmasına sebep olur. Bu işler, yutar-hücrelerin sindirme işlerini kolaylaştırır. Beden genelinde uyur halde gezinen yaklaşık 25 farklı protein molekülünden oluşur [45].

3.5. Lenfoidler organlar

Bağışıklık sistemini meydana getiren dokular ve organlar vücudun her tarafına dağıtılmıştır. Lenfositlerin üretimi, büyümesi ve gelişmesi ile ilgilenen bu organlar lenfoid organlar olarak bilinirler [46]. Lenfoid organlar 2 kısma ayrılır.

Santral lenfoid organlar

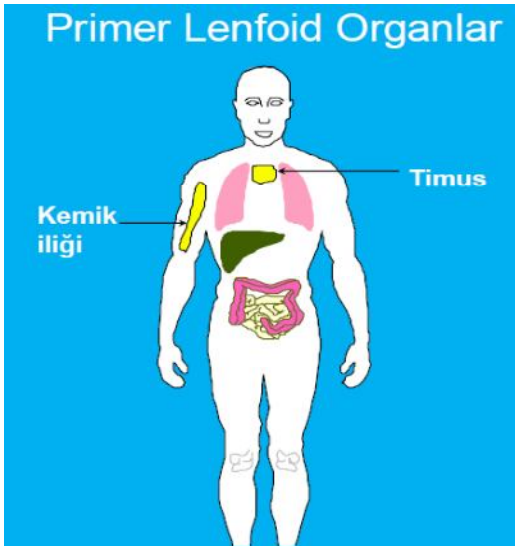
- Timus
- Kemik iliği

Periferik lenfoid organlar

- Lenf nodları
- Dalak
- Mukozal lenfoid doku

3.5.1. Periferik lenfoid organlar

İmmun sistem hücrelerinin, kök hücrelerden itibaren erginleşme sürecini geçirdikleri bölgelerdir [47].



Şekil 3.4. Primer lenfoid organlar

3.5.1.1. Kemik iliği

B lenfositlerin ve miyeloid seri hücrelerinin olgunlaşma yeridir [47]. Kan üreten dokudur. Bazı kemiklerin iç kısımlarında yer alır. Kemik iliği kök hücreler tüm kan hücresi gruplarının kaynağıdır.

3.5.1.2.Timüs

T Lenfositlerin olgunlaşma bölgesidir. Antikor sentezi yapılmayan tek lenfoid organdır[47].

3.5.2. Sekonder lenfoid organlar

Olgun ve göreve hazır hale gelmiş immünositlerin görev bekleme bölgeleridir. Burada organizmaya yabancı antijenler işlemden geçirilir, antijenlere özgül antikorlar üretilir [47].

3.5.2.1. Dalak

Lökositlerin kan akışına saldıran organizmayı yok ettiği yerdir. Kandaki antijenlerin işlem bölgesidir.

3.5.2.2. Bademcik ve lenf bezi

Solunum sistemine saldırılara karşı vücudu koruyan bağışıklık hücreleri içeren özelleşmiş lenf düğümleridir.

3.5.2.3. Lenf damarları

Bağışıklık organları ve kan için lenf taşıyan kanalların ağından oluşmaktadır.

3.5.2.4. Lenf düğümleri

Her düğümün B ve T bağışıklık hücrelerini depoladığı lenf damarlarının yakınında bulunur.

3.5.2.5. Apandis

Bağışıklık hücreleri içeren özelleşmiş lenf düğümleri sindirim sistemini korumak için ayrılır.

3.6. Bağışıklık Sistemi İletişim Ağı

Bağışıklık sistemi antijenlerin vücuda girmesi ile harekete geçer. Harekete geçme işlemi ile bağışıklık yanıtı oluşur. Antijenler bir mikroorganizma veya bir molekül olabilir. Bağışıklık hücreleri kemik iliğindeki olgunlaşmamış kök hücreleridir. B hücreleri özel ve tekil antikor üretmeye programlanmışlardır. Örneğin bir B hücresi difteri hastalığını yok etmeye çalışan antikorlar üretmeye programlanmışken, bir diğer B hücresi kolera hastalığını yok etmek için antikor üretmeye programlanmıştır [48].

Hücre yüzeylerinde bulunan MHC (major histokompatibility compleks) molekülleri kendinden olmayan antijenleri T hücrelerinin fark edebileceği bir yapıya dönüştürür. T hücreleri, antijeni, yüzeyinde bulunan MHC molekülüyle bağlanmış ise fark eder [48].

Yardımcı T hücreleri fagosit hücreleri veya diğer T hücrelerini uyararak bağışıklık cevabını başlatır veya B hücrelerini uyarıp antikor üretiminin başlamasını sağlar. Öldürücü T hücreleri yüzeylerinde yabancı ya da normal olmayan molekülleri taşıyan hücrelere direkt saldırır [48].

4. BAYES VE YAPAY BAĞIŞIKLIK SİSTEMİ

Bu çalışmada e-postaları sınıflandırmak için Bayes sınıflandırıcılarından Naive (Yalın) Bayes sınıflandırıcısı ve YBS algoritmaları kullanılmıştır.

4.1. NAIVE BAYES

Naive Bayes bir olasılıksal sınıflandırmadır. Herhangi bir kelimenin sınıfını bulmak için o kelimenin sınıfının koşullu olasılıklarını kullanır. Bu yaklaşımın "Yalın" kısmı içindeki kelime bağımsızlığı varsayımından kaynaklanmaktadır. Kelime kombinasyonlarının olasılıklarını tahminci olarak kullanmaz. Bu yüzden karar ağacı gibi algoritmaların üstel karmaşıklığından öte verimli bir yaklaşımdır. Belge sınıflandırma ve spam filtreleme gibi bazı öğrenme problemlerinde yaygın olarak kullanılan pratik bir yaklaşımdır [14].

4.2. YAPAY BAĞIŞIKLIK SİSTEMİ

YBS insan bağışıklık sistemini model alarak geliştirilen bir sistemdir. Bağışıklık sistemi yabancı antijenlerle (non-self), kendi vücut antijenlerinin (self) ayrılmasını sağlayan ve yabancı maddeleri yok eden karmaşık bir mekanizmadır. Verilerin işlenmesi boyutundan olaya baktığımızda öğrenme yeteneğine sahip paralel ve dağıtık akıllı bir sistemdir [49]. YBS, bağışıklık sistemini model alarak geliştirilen makine öğrenmesi algoritmasıdır.

YBS, örüntü tanıma, hesapsal güvenlik, anormali tespiti, optimizasyon, makine öğrenmesi, robotik, hata teşhisi gibi alanlarda ve bunların alt dallarında, ayrıca ekoloji, üretim sistemleri, akıllı evler, açık web sunucu koordinasyonu, protein yapısı tahmini gibi alanlarda başarıyla kullanılmış ve etkili sonuçlar alınmıştır [50].

4.2.1. YBS Algoritmaları

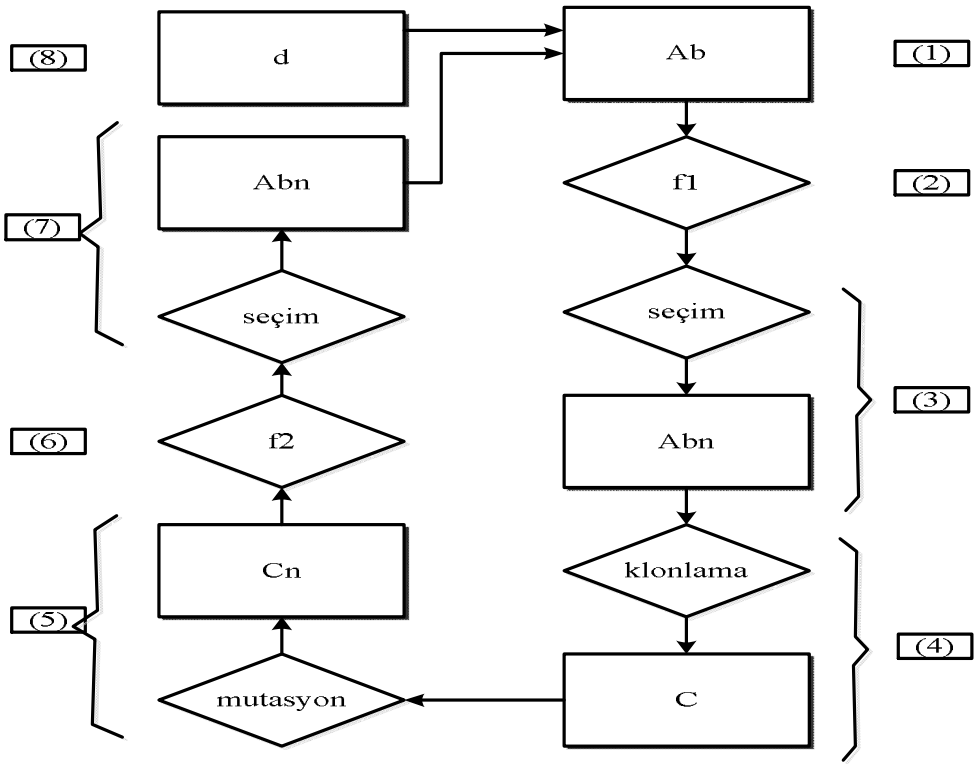
4.2.1.1. Klonal seçim algoritması

Klonal seçim mekanizması, yabancı hücrenin B hücresi tarafından tanındığı zaman verdiği bağışıklık yanıtını açıklamak için kullanılır [50].

Klonal seçim algoritması sürekli değişen ortamlarda bilinmeyeni keşfetme problemlerinin çözümü için kullanılmaktadır [51].

Bu algoritma; biçim tanıma, optimizasyon gibi kompleks problemlerin çözümüne uygulanmıştır [37]. Bağışıklık sisteminin, hafıza hücrelerine sahip olması, en fazla benzerlik gösteren hücreleri seçerek klonlaması, etkinlik gösteremeyen hücrelerin ölümü, benzerlik derecesini arttırmak için olgunlaştırma ve yeniden seçme, üretim ve çeşitliliğin korunması gibi özellikler bu algoritmada uygulanmıştır [40]. Klonal seçim algoritmasının adımları ve notasyonu, aşağıda verilmiştir [52];

1. Vücut antikor repertuarını oluşturan antikorlar, başlangıç çözüm kümesini oluştururlar (Ab).
2. Antikorların benzerlik dereceleri hesaplanır (f1).
3. n adet en yüksek benzerlikli antikor seçilir (Ab_n).
4. Seçilen n adet antikorun benzerlik dereceleri ile doğru orantılı olarak, yüksek benzerlik dereceli antikorun daha fazla olacak şekilde klonlanır (C).
5. Antikorların benzerlik derecesi yüksek olan, daha az olacak şekilde mutasyona uğratılır (Cn).
6. Mutasyona uğramış klonların benzerlik dereceleri belirlenir (f2).
7. n yüksek benzerlik dereceli antikorlar yeniden seçilir (Ab_n).
8. d adet en düşük benzerlik derecesindeki antikorlar, yeni üretilen antikorlarla değiştirilir (Ab_n).



Şekil 4.1. Klonal seçim algoritmasının blok diyagramı [53].

4.2.1.2. Negatif seçim algoritması

İnsan vücudunda herhangi bir antijenin vücut antijeni mi yoksa yabancı antijen mi olduğunun tespiti çok önemlidir. Şayet bağışıklık sistemi bu ayırt edici özelliği yanlış kullanırsa kendi vücudunu yabancı bir antijen olarak görmeye başlayabilir ve bu durumda vücutta hastalıklar meydana gelebilir. Bağışıklık sistemi vücuttaki antijenleri tespit etmek için negatif seçim yöntemini kullanır.

Negatif seçim algoritması YBS sistemleri alanında geliştirilen ilk algoritmalardan biridir. Bu algoritma T hücrelerinin gelişimini örnek alarak geliştirilmiştir.

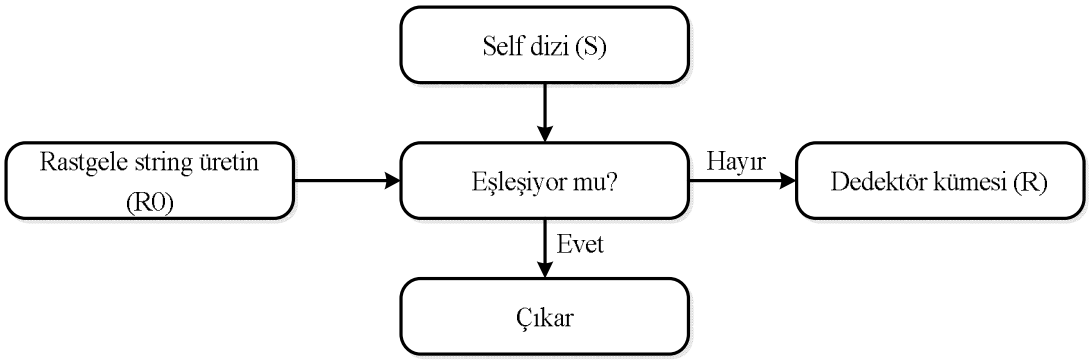
Bilgisayar saldırıları, ağ saldırıları, zaman serileri tahmini, şekil denetimi ve donanım hata toleransı gibi anormallik belirleme problemlerinde negatif seçim algoritması önerilmiştir [40]. Dasgupta ve Forrest self/nonself ayrımını yapan negatif seçim algoritmasını kullanarak değişikliği algılama algoritması geliştirmişlerdir. Bu algoritma 2 aşamadan oluşmaktadır [40].

1. Dedektörler bir küme oluşturur. Her dedektör (dikkatli bir "EŞLEŞME" için) korumalı

verilere uymayan bir string'dir. Bu aşamaya algılama safhası denir. Bu aşama, Şekil 4.2 de gösterilmiştir.

Bu aşamanın işlem adımları aşağıda listelenmiştir:

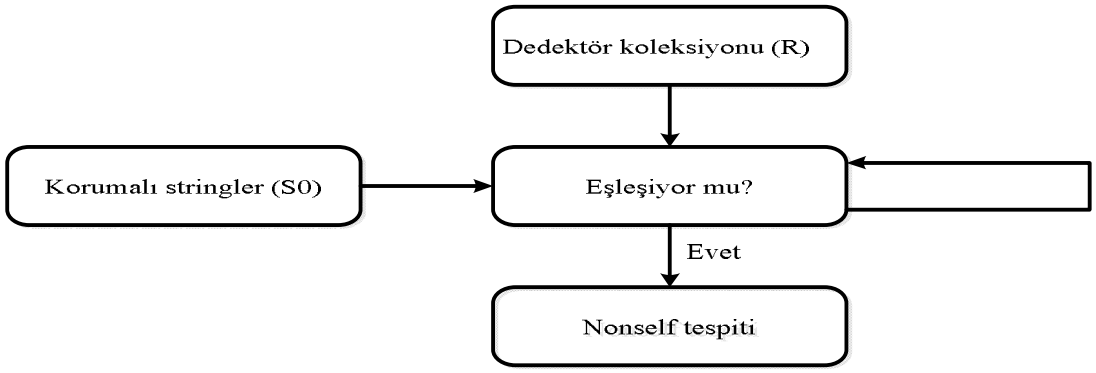
a. Korunacak veriler kümesi / self dizi (S) belirlenir.
b. Rastgele aday elemanları R0 string'leri şeklinde üretilir.
c. R0 kümesine ait olmayan elemanları tanımak için dedektör kümesi olan R kümesi oluşturulur.
d. S kümesiyle R0 kümesindeki elemanlar XOR işlemi ile eşleştirilir. Elemanlar eşleşiyorsa bu eleman yabancı bir antijendir denir. Bu durumda R0 kümesindeki eleman çıkarılır. Eşleşmiyorsa şayet R0 elemanı dedektör küme (R) içine alınır.



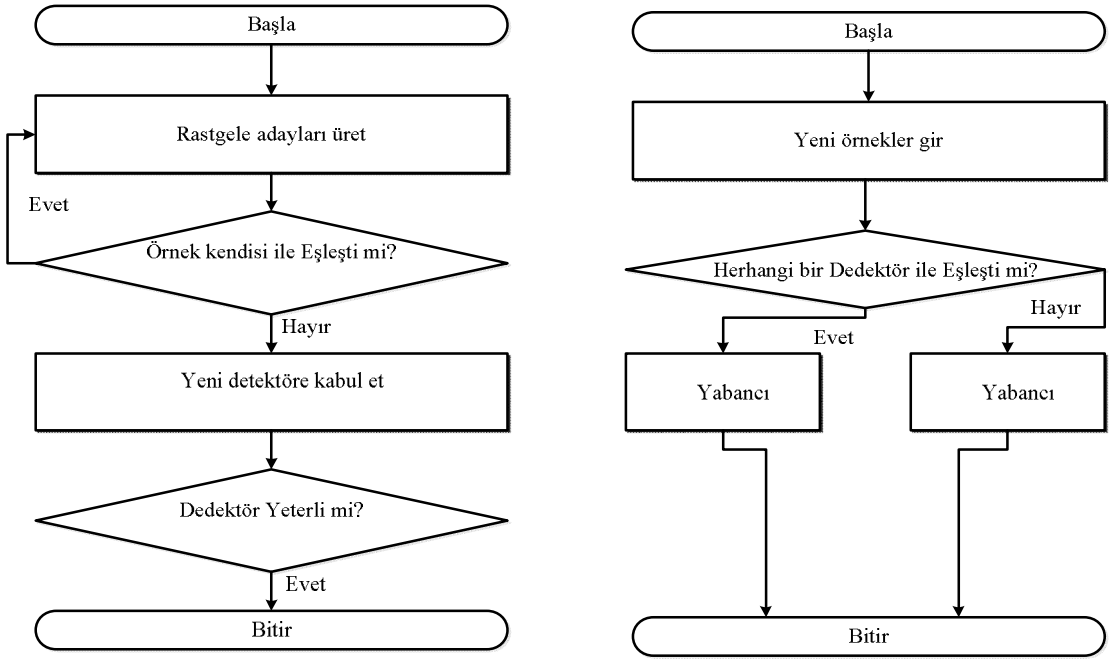
Şekil 4.2. Geçerli dedektör kümesi üretimi [40].

2. Korunan veriler dedektörler ile karşılaştırılarak izlenir. Bir dedektör bugüne kadar etkinse, bir değişiklik meydana gelmiş demektir. Bu aşamaya gözetleme safhası denir.

Algoritmanın diğer bir aşaması yabancı elemanların varlığını tespit etmek için dedektör koleksiyonu (R) ile yabancı maddeleri karşılaştırarak sistemi izlemekten oluşur. Bu durumda korumalı stringler (S0) oluşturulur. Dedektör kümesinin tüm elemanları için S0 kümesinin bir elemanını tanıyıp tanımadığı kontrol edilir. Eşleşme var ise bir non-self tespiti yapılmış demektir. Eşleşme yoksa diğer elemanlar için eşleşme devam eder. Bu aşama Şekil 4.3 te gösterilmiştir [40].



Şekil 4.3. Değişim için monitör korumalı stringler [40].



Şekil 4.4. Negatif seçim algoritmasının akış diyagramı

4.2.1.3. AIRS1 (Artificial Immune Recognition System) algoritması

AIRS1 2001 yılında Watkins tarafından oluşturulmuş denetimli öğrenme algoritmasıdır.

AIRS1 algoritmasında 2 popülasyon vardır. Bunlar Artificial Recognition Ball (ARB) ve hafıza hücreleri. ARB aynı antikora sahip çok sayıdaki hücrelerdir. Bir eğitim hücresi sunulduğunda, antijenle uyuşan ARB'ler seçilir ve onlara daha fazla kaynak ayrılır. Bu uyum, mutasyon ve seçim süreci boyunca yeterli bilgiye sahip olan aday hafıza hücresi,

genel hafıza hücresi kümesine yerleştirilir. Bu işlem her eğitim elemanı için tekrarlanır. En sonunda sınıflandırma en yakın hafıza hücresi K-en yakın komşuluk algoritmasıyla bulunarak gerçekleştirilir [43].

AIRS1 algoritması aşağıdaki özelliklere sahiptir [61].

Genelleştirme: Genelleme için tüm veri kümesi kullanılmaz.

Parametre kararlılığı: Kullanıcı tanımlı parametreler optimize edilmese bile, performansındaki değişim çok küçüktür.

Performans: Yüksek performans yeteneğine sahiptir.

Öz denetim (self-regulatory): Problem YSA'ya uygulandığında bir topoloji seçmek gerekiyor ancak AIRS1 algoritmasında topoloji seçmeye gerek yoktur, eğitim sırasında belirlenir. Eğitimden önce topoloji seçmeye gerek yoktur.

AIRS1 algoritması karmaşık olduğundan AIRS2 algoritması geliştirilmiştir.

4.2.1.4. AIRS2 algoritması

AIRS2 algoritmasında örneğe, antijen denilmektedir. Algoritmanın temel amacı, eğitim kümesini modellemek için tanıma hücreleri havuzu oluşturmak ve bilinmeyen veriyi bu havuza göre sınıflamaktır. AIRS2 Algoritması 5 adımdan oluşur. Algoritmada 1. ve son adımlar bir kez uygulanırken; 2., 3., 4. adımlar eğitim kümesindeki her örnek için uygulanır. Algoritma aşamaları [61,62]:

1. Aşama (Başlangıç): Veri kümesi [0-1] aralığına normalize edilir. Antijenlerin (özniteliklerin) benzerlik eşik değerleri, mesafe değeri kullanılarak hesaplanır. Mesafe, Öklid uzaklığı ile hesaplanır. İki antijen arasındaki benzerlik değeri, mesafenin en büyük mesafeye bölünmesi ile elde edilir.

2. Aşama (Antijen eğitimi): Antijenler bellek havuzuna birer birer sunulur. Bellek havuzundaki tanıma hücreleri uyarılarak bu hücrelere uyarılma değeri atanır. En yüksek uyarılma değerine sahip olan tanıma hücresi, en iyi uyuşan bellek hücresi olarak seçilir ve affinity olgunluk sürecinde kullanılır. Bu hücre klonlanır ve mutasyona tabi tutulur. Klonlar, yapay tanıma küreleri havuzuna (YTK) eklenir. Formül 4.2 ile klonların sayısı hesaplanır. Formül 4.1'de stim (uyarılma) değerinin nasıl hesaplanacağını gösterilmiştir. ClonalRate ve hyperMutationRate kullanıcı tanımlı parametrelerdir.

$$stim = 1 - affinity \tag{4.1}$$

$$numClones = stim * conRate * hyperMutRate \tag{4.2}$$

3. Aşama (Sınırlı kaynak için yarış): YTK havuzuna mutasyona uğramış klonlar eklendiği zaman yarışma başlar. YTK havuzu antijen ile uyarılır ve uyarılma değerine göre sınırlı kaynak ataması yapılır. Yeterli kaynağı olmayan YTK'ler havuzdan atılır. Durma kriteri sağlanınca, işlem durur. Aksi halde, YTK'nin mutasyona uğramış klonları oluşturulur ve bu döngü durma kriteri sağlanana kadar devam eder.

4. Aşama (Bellek hücre seçimi): Bir önceki adımda, durma kriteri sağlandığı zaman, en yüksek uyarılma değerine sahip YTK aday bellek hücre olarak seçilir. YTK'nin uyarılma değeri orjinal en iyi uyuşan hücreden daha iyiyse, YTK bellek hücre havuzuna kopyalanır.

5. Aşama (Sınıflandırma): Bellek hücre havuzunun son hali çapraz geçerlemede kullanılır. K-en yakın komşu yaklaşımına göre sınıflandırma gerçekleştirilir.

4.2.1.5. AIRS1 algoritması ile AIRS2 algoritması arasındaki farklar

Her iki algoritma benzer adımlardan oluşur. AIRS1 algoritmasının karmaşık yapısından AIRS2 geliştirilmiştir. AIRS1 algoritmasının karmaşık yapısı performansının düşük olmasına neden olmuştur. AIRS2 algoritmasının daha basit yapısı performans olarak daha iyi olmasını sağlamıştır [63].

Algoritmaların temel farkları [63]:

1. YTK havuzu AIRS1 için kalıcı, AIRS2 için geçici bir kaynak olarak kullanılır. YTK ile aynı sınıf etiketine sahip antijenlerin yüksek benzerliğe sahip olmasında AIRS2 daha doğru bir yaklaşımdır. Diğer YTK iyileştirme geçişlerinden gelen önceki YTK'ler AIRS1 gibi daha fazla değerlendirme için gerekli değildir.
2. AIRS1 algoritmasında klonların sınıfları mutasyon işleminden sonra değiştirilebilir ama AIRS2 algoritmasında buna izin verilmez.
3. AIRS1 algoritması kullanıcı tanımlı mutasyon parametresini kullanır, AIRS2 somatik hiper mutasyon kavramını kullanır. Klon Mutasyon miktarı affinity orantılıdır.

4.2.1.6. AIRS2PARALLER algoritması

AIRS2PARALLEL algoritması paralel programlama yaklaşımı ile gerçekleştirilmiştir. Yaklaşım adımları aşağıdaki adımlardan oluşur [61,63]:

1. Eğitim veri kümesi np bölümlere ayırmak.
2. Eğitim bölümlerini süreçlere ayırmak.
3. Bellek havuzları np sayısını birleştirmek.
4. Kombine bellek havuzu oluşturmak için birleştirme yaklaşımı kullanmak.

Veri kümesi çok küçük alt kümelere ayrılmamışsa, bu yaklaşım performansı muhafaza ederek daha hızlı bir çözüm sağlar. Kötü veri azaltma yüzdesi ve daha fazla araştırmaya neden olan birleştirme teknikleri bu sorunu çözmek için yapılmalıdır [64].

4.2.1.7. Klonal Seçim Algoritması (CLONALG)

CLONALG algoritması Klonal seleksiyon teorisinden esinlenerek geliştirilmiş bir algoritmadır. Bu algoritma aşağıdaki elementlerden esinlenmiştir [65].

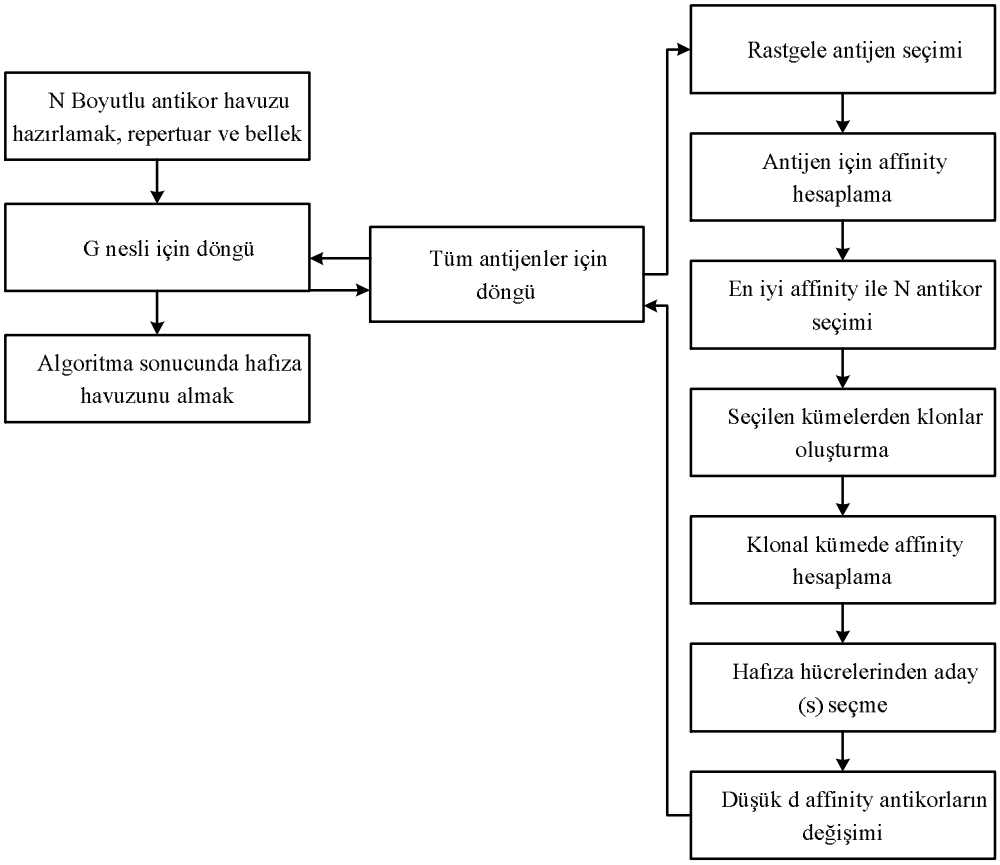
- Belirli bellek kümesi bakımı
- En çok uyarılmış antikorların seçimi ve klonlaması
- Uyarılmış olmayan antikorların yok edilmesi
- Benzerlik Mutasyonu
- Antijen ile yakınlık orantılı klonların yeniden seçimi
- Çeşitliliğin üretimi ve bakımı

Bu algoritmasının amacı mühendislik problemlerinin çözümünü temsil edecek antikor havuzunu geliştirmektir. Bu durumda bir antikor, soruna tekil bir çözüm veya çözüm elemanı gösterir ve bir antijen, problem uzayının bir ögesini veya değerlendirmesini gösterir. Algoritma bellek antikorlarının arzu edilen nihai hafızaya göre araması için iki mekanizma sağlar. İlk mekanizma klonlanmış antikorun affinity olgunlaşması yolu ile yerel arama, ikincisi bir yerel sorunu önlemek için rastgele oluşturulan antikorların nüfus ekleme noktasını içeren genel bir arama mekanizması. Bu algoritmanın her adımı aşağıda açıklanmıştır [65].

- **Başlatma:** CLONALG için ilk adım sabit boyutlu antikor havuzu hazırlamaktır. Daha

sonra, bu havuz iki bölüme ayrılır. Birinci bölüm çözümü temsil eder, kalan ise sistem havuzu içinde çeşitliliği tanıtmak için kullanılır.

- **Döngü:** Başlangıç aşamasından sonra kullanıcı tarafından özelleştirilebilir G yineleme sayısı kadar algoritma yürütülür. Sistem tüm bilinen antijenler için bu aşamada ortaya çıkar.
    a. **Antijen seç:** Bir antijen havuzdan rastgele seçilir.
    b. **Ortaya çıkarma:** Seçilen antijen antikorları gösterilmiştir ve antikor-antijen arasındaki benzerlik gösteren benzeşme değerleri hesaplanır. Genellikle Hamming veya Öklid mesafe ile affinity değeri hesaplanır.
    c. **Seçim:** Affinity en yüksek değerler tespit edilir ve bu yüksek değerler n antikorlar kümesinden seçilir.
    d. **Klonlama:** Seçilen antikorlar kendi benzerlik değerleri ile orantılı olarak klonlanır.
    e. **Affinity olgunlaşma (mutasyon):** Yinelenen antijenleri ebeveynlerinin benzerlik değerleri ile orantılı ters mutasyona uğratılır.
    f. **Klon ortaya çıkarma:** Klon için antijen sunulur ve benzeşme değerleri yeniden hesaplanır.
    g. **Aday:** Yüksek benzeşime sahip klonlar aday bellek antikorları olarak işaretlenir. Bu aday bellek hücre benzeşimi bu antijen yüksek olması ile aday bellek antikor bellek havuzundan en yüksek uyarılan antijen ile değiştirilir.
    h. **Değişim:** D antijenleri içinde kalan antijen havuzu r en düşük benzeşme değerlerine sahip yeni rastgele antikorlar ile değiştirilir.
- **Bitir:** Eğitim aşamasından sonra antijen havuzu bellek bölümü soruna çözüm kabul edilir.



Şekil 4.7. CLONALG algoritmasına genel bakış [65].

Bu algoritmada AIRS algoritmasından daha düşük karmaşıklık ve daha az sayıda kullanıcı parametresi vardır. CLONALG algoritmasında bulunan yedi kullanıcı tanımlı parametre aşağıdaki gibi açıklanabilir [65].

- **Antikor havuz boyutu (N):** Sistem içinde antikorların toplam sayısı, bu daha sonra bir hafıza havuzu ve kalan havuza ayrılır.
- **Klonal faktör (b):** Klonlar için seçilen her bir antikor için oluşturulan klonların sayısını büyütmek için kullanılan çarpan.
- **Nesiller sayısı (G):** Aday nesillerin toplam sayısı.
- **Kalan havuz oranı:** Toplam antikorların oranı (N) kalan antikor havuzu için ayrılır.
- **Rastgele sayı üreteci çekirdeği:** Rastgele sayı üreteci için çekirdek
- **Seçim havuzu boyutu (n):** Antikorların her antijene exposure için tüm antikor havuzundan seçilecek toplam sayısı.

- **Toplam değiştirmeleri (d):** Antikor nüfus çeşitliliği üzerinde Yeni rastgele antikorların her antijene güçlü bir etki ile maruz kalmasıyla antikor değişiminden kalan sayıdır.

4.2.1.8. CSCA algoritması

CSCA algoritması CLONALG algoritmasının iyileştirilmiş halidir. Veri giriş verileri ve karşılık gelen sınıfı gösteren bir etiket özniteliklerini temsil eden d boyutlu vektör tarafından oluşan antikorlar tarafından temsil edilir. Algoritma aşağıda açıklanan dört ana aşamadan oluşur [66]:

- **Başlatma:** Öncelikle içinde rastgele seçilen veri vektörlerinin antikorları bulunan bir ilk popülasyon antikorları oluşturmak için kullanılan bir başlangıç tohum aşaması vardır.
- **Döngü:** Bazı adımlar nesillerin kararlı sayısı için tekrarlanır. Her döngüde antikorlar nüfusun (antijenler denir) tüm eğitim veri öğeleri için sunulur ve buda hangi antijen hangi antikor tarafından (sınıflandırılmış) doğrulandığını belirtir. Doğru ve yanlış sınıflama numarası sadece en iyi eşleşen birimler için (antikor belirli bir antijen için en büyük affinity ile) hesaplanır. Bu puanlamaya dayanarak, daha az uygunluk veya bir eşik değere eşit olan antikorları muaf tutarak uygunluk hesaplanır. Geriye kalanlar ve sınıflandırılmayan antikorlar tamamen tanımlandığında kabul edilir ve ana nüfusa (AN) eklenir. Ana nüfusu içinde kalan diğer oluşumlar (ANDK) klonlanır ve mutasyona uğratılır. Sonuçta elde edilen antikorlar rasgele seçilerek ana nüfus içine dahil edilir. Hiçbir antikor klonlama ve mutasyona uğratılmak için seçilmediğinde döngü biter, bunun dışında döngü devam eder.
- **Son Ayıklama:** Nüfus bu durumda son kez antijen nüfusuna maruz bırakılır, uygunluk puanları hazırlanır ve ayıklama işlemi 2. adımda gerçekleştirilir.
- **Sınıflandırma:** Kalan antikorlar kNN tabanlı işlemler için girdi olarak hafıza hücre kümeleri oluşturur. Bu aşama sürecinde bir sınıf hafıza temelli veri ögesi algoritma tarafından hala görülememiş öğeleri üretir. Her yeni örnek ögelerin bir mesafe ölçümünü kullanarak karşılaştırılır ve yakın var olan örnekler için yeni bir sınıf atamak için kullanılır. Çoğunluk sınıfı (kullanıcı tarafından tanımlanan) en yakın k komşuları için yeni örnek olarak atanır.

ilk üç adım algoritma öğrenme aşamasını içerir ve sonuncusu sınıflandırma aşamasını içerir. Öğrenme aşamasında CSCA optimizasyonu, seçilen antikor nüfusuna bağlıdır ve sunulan eğitim verisi evrimleştirilerek maksimize edilir. Sadece en yakın komşular sınıflandırma aşaması olarak kullanılmış olduğundan, hali hazırda tanınan veri ögelerinde yalnızca her bir temel antikor için bir uygunluk değeri için algoritma hesaplanır. Antikorların uygunluk değeri (4.3) numaralı formül tarafından hesaplanır: $nr_{sameClass}$ formül içinde aynı sınıftaki antikorların tanımıdır, $nr_{otherClass}$ ise farklı sınıftaki antikorların tanımıdır. Geleneksel olarak $nr_{otherClass}$=0 olduğunda uygunluk(antikor)=$nr_{sameClass}$ olur.

$$\text{Uygunluk(antikor)} = \frac{nr_{sameClass}}{nr_{otherClass}} \tag{4.3}$$

Adım 2'de, ANDK ($an^{'}$) klonlanır ve yeni bir popülasyon olan C nüfusu üretilir. AN tarafından tanımlanmış (4.4) numaralı formül ile her bir antikor için oluşturulmuş klonların sayısı hesaplanır. Burada λ varsayılan olarak isteğe bağlı ölçek bir faktör, n antijenlerin toplam sayısı ve r(ab) 4.5. formül tarafından tanınan bağıl uygunluğudur.

$$numClones(an) = r(an) * n * \lambda \tag{4.4}$$

$$r(an) = \frac{uygunluk_{an}}{\sum_{i=1}^{|AN|}(uygunluk(an_j))} \tag{4.5}$$

Klonlamadan sonra, C nüfusundaki her an antikor, nitelik vektöre göre her bir att değeri değiştiren bir mutasyon işleminden geçer (4.6), burada rnd(an) rastgele reel sayı aralığını [a,n] üreten bir fonksiyondur ve δ ($range_{att}$) olası aralığına göre (4.7) tarafından hesaplanan att mutasyon oranıdır, antikor suffering mutasyon göreli uygunluk ile problem tarafından tanımlanır.

$$att_{new} = rnd(att_{old} - \delta, att_{old} + \delta) \tag{4.6}$$

$$\delta = \frac{range_{\text{att}}}{2} * r(an) \tag{4.7}$$

5. VERİ KÜMESİ OLUŞTURMA VE ÖZNİTELİK ÇIKARIMI

5.1. Veri Kümesi Oluşturma

E-posta hizmet sağlayıcıları tarafından spam e-postalar en fazla 30 gün tutulmaktadır. 30 gün sonunda spam e-postalar hizmet sağlayıcı tarafından otomatik olarak silinir. Bu nedenle veri kümesini oluşturmak için düzenli olarak belli aralıklarla e-posta hesabına gelen spam e-postalar ayrı bir klasöre kaydedilmelidir. Normal e-postalar hizmet sağlayıcı tarafından otomatik silinmezler. Bu e-postalara kullanıcı haricinde herhangi bir müdahale söz konusu değildir.

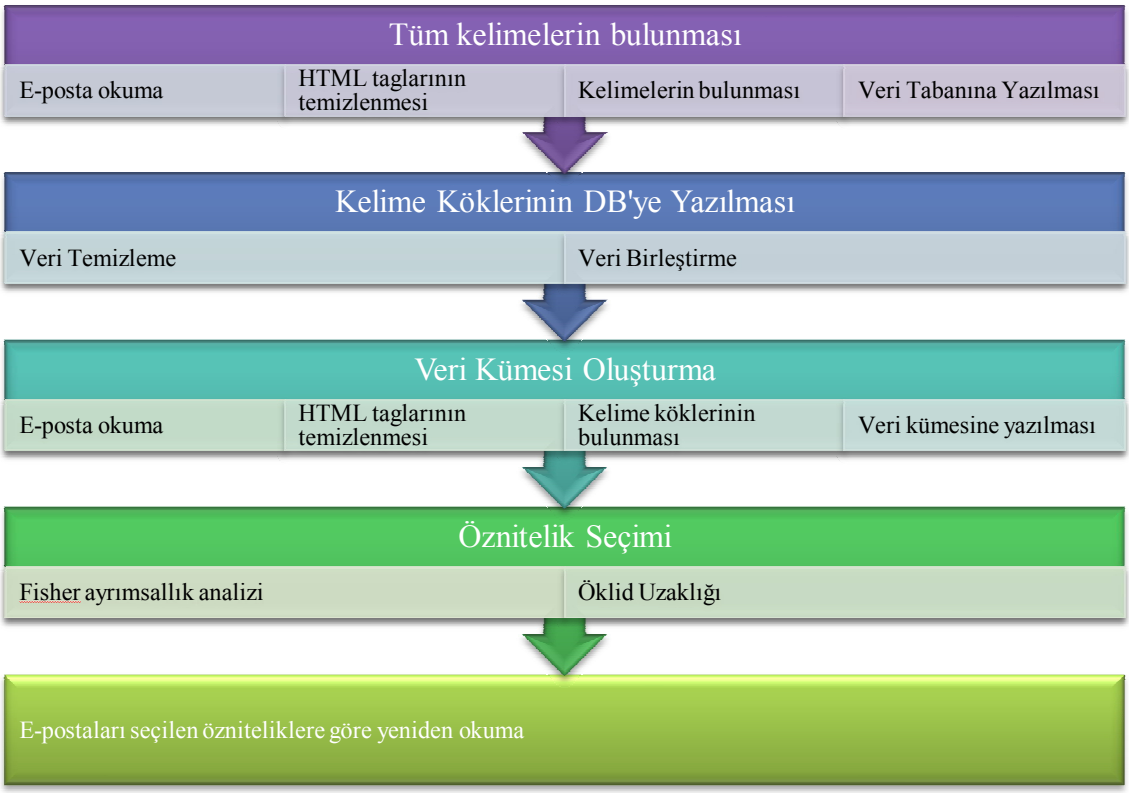
Veri kümesi Siirt Üniversitesi resmi web sayfasının giriş ekranına eklenen iletişim e-posta adresine gelen spam e-postalardan ve 1 yıldan fazla süredir e-posta hizmet sağlayıcı hesaplarıma düşen spam ve 5 yıldır kullandığım Gmail ve Hotmail e-posta adreslerime gelen normal e-postalardan oluşmaktadır. Bu e-postalar arasında sadece Türkçe e-postalar dikkate alınmış geri kalan e-postalar silinmiştir. Tez çalışması için 539 adet spam, 601 adet normal olmak üzere toplam 1140 adet e-posta kullanılmıştır. E-postalar iki farklı gruba ayrılmıştır. Tablo 5.1 her bir grup için ayrılan e-postaların sayısı verilmiştir.

E-postaların 2 farklı gruba ayrılmasının nedeni: gruplardan biri öznitelik seçmek için kullanılırken diğer gruptaki e-postalar seçilen özniteliklerden veri kümesi elde etmek için kullanılmıştır. Bu işlem için öncelikle 1. gruptaki e-postalar okunarak bir veri kümesi oluşturulmuştur ve oluşturulan bu veri kümesinden özniteliklerin belirlenmesi sağlanmıştır. Veri kümesinden öznitelik seçmek için FAA ve Öklid yöntemi kullanılmıştır. 1. grupta elde edilen özniteliklere göre 2. gruptaki e-postalar okunmuş ve 2 farklı öznitelik seçim yöntemi kullanıldığı için 2 farklı veri kümesi elde edilmiştir. 1. gruptaki işlemler bittikten sonra 2. gruptaki e-postalar içinde öznitelikler tespit edilmiş ardından 1.grupdaki e-postalar bu özniteliklere göre yeniden okunarak veri kümeleri elde edilmiştir. Bu işlem sonucunda 2 farklı grupta 2 farklı öznitelik seçme yöntemi kullanıldığından toplam 4 farklı veri kümesi elde edilmiştir. Bu veri kümeleri ile algoritmaların sınıflandırma başarımları ölçülmüştür. 2 farklı grup olduğundan dolayı sınıflandırma algoritmalarının FAA ve Öklid yöntemlerinden elde edilen veri kümeleri üzerindeki sınıflandırma başarılarının ortalaması alınmıştır.

Tablo 5.1. Veri kümesi oluşturmak için kullanılan e-posta sayıları

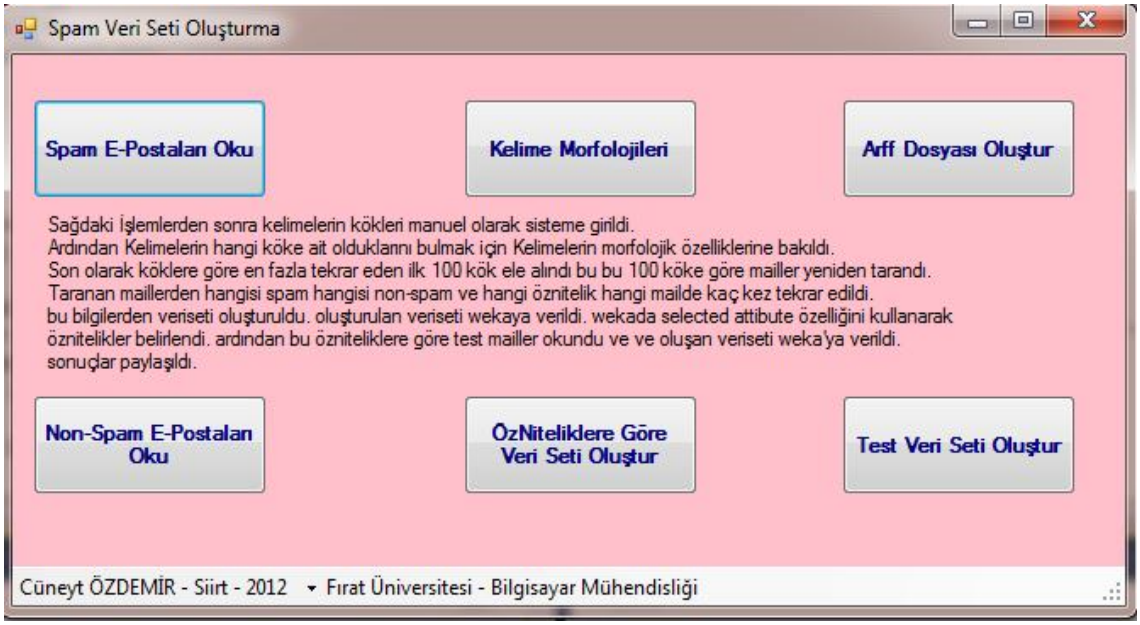
	1. Grup	2. Grup	Toplam
Spam	269	270	539
Normal	302	299	601
Toplam	571	569	1140

Bütün e-postalar Mozilla Thunderbird programı kullanılarak bilgisayara dosya şeklinde kaydedilmiştir. Mozilla Thunderbird açık kaynak kodlu olduğundan dolayı tercih edilmiştir. E-postalar bilgisayara indirildiği sırada spam e-postalar ve normal e-postalar ayrı klasörlere kaydedilmiştir. Spam ve Normal e-postalar ayrı klasörlere dosyalandıktan sonra spam ve normal e-postalar 1. grup ve 2. grup e-postalar olmak üzere yarı yarıya oranında ayrı klasörlere kaydedilmiştir (Bak. Tablo 5.1). E-postaların dosya şeklinde klasörlere düzenli olarak kaydedilmesinden sonra program kodlama aşamasına geçilmiştir.



Şekil 5.1. Veri kümesi oluşturma aşamaları

Veri kümesi oluşturmak için yazılan program C#.NET programlama dili ile geliştirilmiştir. Geliştirilen yazılımın grafiksel kullanıcı ara yüzü (GUI) Şekil 5.1.'de gösterilmiştir.



Şekil 5.2. Veri kümesi oluşturma programının açılış ekranı

Program kodlamanın ilk aşamasında 1. gruptaki tüm e-postaların okunması gerçekleştirilmiştir. Bu işlem ile e-postalarda geçen kelimeler tespit edilerek bu kelimelerin kullanım sıklıkları tespit edilmeye çalışılmıştır.

Okuma işleminin ilk adımında e-postalarda geçen tüm html etiketleri ("<html>,
,<tr>" vb.) silinmiştir. Okuma işleminin ardından her bir e-postada geçen tüm kelimeler birbirinden ayrılarak bir liste haline getirilmiştir. E-postalarda geçen kelimelerin birbirinden ayrılması işlemi aşağıdaki karakterlere göre gerçekleştirilmiştir.

```
"", "/", "\\", "^", "$", "]", "[", "|", "?", "+", "*", ",", " ", "!", "-", ")", "(", " ",
"\r\n", """, ":", "...", "'"
```

Şekil 5.3. E-postalardaki kelimeleri birbirinden ayırmak için kullanılan karakterler

E-postaların okunması ve e-postalarda geçen kelimelerin bulunması işlemi aşamasında

birçok e-postanın base64 algoritması ile şifrelendiği görülmüştür. Bu problemi gidermek için şifrelenmiş e-postalara de-şifreleme metodu uygulanmış ve e-postaların anlamlı bir şekilde açılması sağlanmıştır. Böylelikle bu e-postalarda geçen kelimelerin okunması sağlanmıştır.

E-postalarda geçen kelimelerin bulunması işlemi sırasında birçok Türkçe kelimenin anlamsız bir şekilde okunduğu görülmüştür. Anlaşılmaz olan bu kelimelerin Türkçe karakterler olan "ş,Ş,ç,Ç,ö,Ö,ü,Ü,ğ,Ğ,ı,İ" gibi kelimelerden kaynaklandığı ve bu karakterler yerine bu karakterlerin ASCII karşılıklarının yazıldığı tespit edilmiştir. Bu sorunu gidermek için e-postalar okunurken iso-8859-9 standardına göre okuma işlemi gerçekleştirilmiş, ardından Türkçe harflerin ACSII kod dönüşümleri yapılmıştır.

Yukarıda söz edilen problemler giderildikten sonra e-posta içerisinde geçen kelimeler bulunmuş ve bu kelimeler veri tabanında bulunan ilgili grup ile alakalı tabloya yazılmıştır. Veri tabanı tasarlanırken 1. grupta bulunan e-postalar içinde geçen kelimeler ve kullanım sıklıkları için "spambirgrup" tablosu, 2. grupta bulunan e-postalar içinde geçen kelimeler ve kullanım sıklıkları için "spamikigrup" tablosu oluşturulmuştur. Veri tabanı olarak MS SQL SERVER 2008 kullanılmıştır. Şekil 5.4'te veri tabanında oluşturulan "spambirgrup" tablosu yer almaktadır. "spamikigrup" tablosu "spambirgrup" tablosu ile aynı şekilde oluşturulmuştur.

```
USE [SPAMVeriKümesi] GO SET ANSI_NULLS ON GO SET QUOTED_IDENTIFIER ON
GO SET ANSI_PADDING ON GO
CREATE TABLE [dbo].[spambirgrup](
        [ID] [int] IDENTITY(1,1) NOT NULL,
        [Kelime] [varchar](999) NULL,
        [SPAM] [int] NULL,
        [NONSPAM] [int] NULL,
        [Kok] [varchar](50) NULL
) ON [PRIMARY] GO SET ANSI_PADDING OFF GO
```

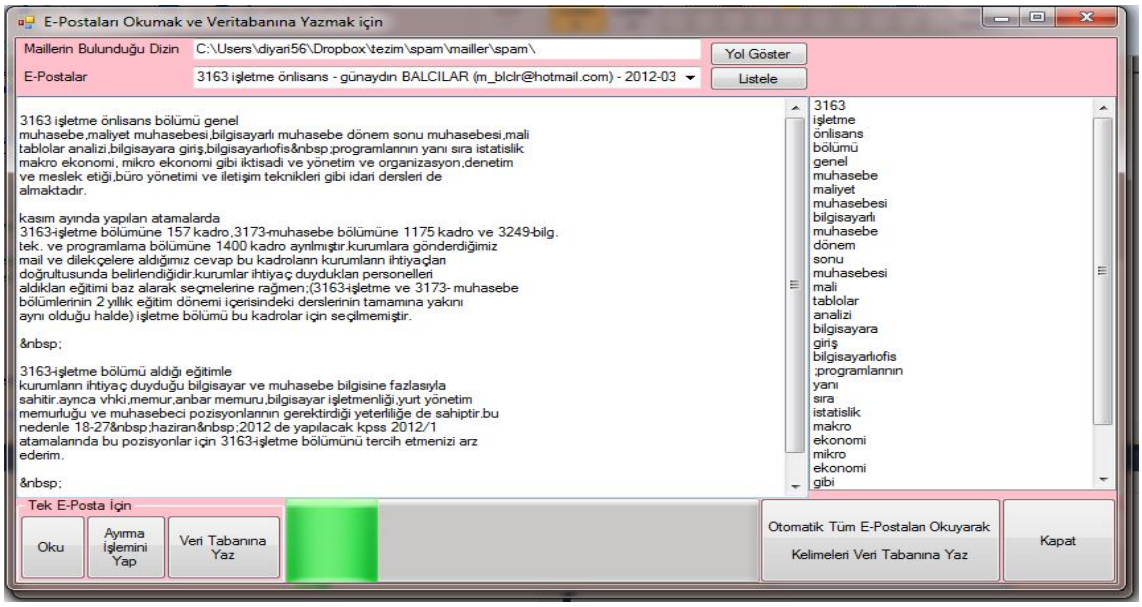
Şekil 5.4. Oluşturulan veri tabanı tablosu

Oluşturulan veri tabanı tablosu kolonları

- **ID** : Otomatik artan bir sayı.
- **Kelime** : Spam ve normal e-postalarda geçen ve tekrar etmeyen kelimeler
- **SPAM** : Kelimelerin spam e-postalarda tekrar etme sayısı
- **NONSPAM:** Kelimelerin normal e-postalarda tekrar etme sayısı
- **Kok** : Kelimelerin Kökleri

Veri tabanına yazım işleminde bütün spam ve normal e-postalar okunmuştur. Her bir e-posta okuma sürecinde, öncelikle e-postada geçen tüm kelimeler bulunmuş ardından bu kelimeler veri tabanına ilgili tabloya yazılmıştır. Kelimeler veri tabanına yazılırken o kelimenin daha önce veri tabanında olup olmadığı kontrol edilmiştir. Kelime veri tabanında yer alıyor ise o kelimenin ilgili alanı okunan e-postaya göre (spam veya nonspam sütunu) 1 artırılmıştır. Kelime daha önce veri tabanında hiç yoksa şayet kelime veri tabanına eklenmiş ve ilgili alana okunan e-postaya göre (spam veya nonspam sütununa) 1, ilgisiz alana 0 yazılmıştır. Bu işlemin sonunda 1. grup e-postalar için 35.547 adet kelimenin, 2. grup e-postalar için 34.446 adet kelimenin veri tabanına yazıldığı görülmüştür. Bu işlemin ardından veri temizleme işlemi gerçekleştirilmiştir. Türkçe olmayan kelimeler, web adresleri, sayılar, Romen rakamları ve herhangi bir anlam içermeyen kelimeler (of, in=, y=, t=, vb.) veri tabanından silinmiştir. 1. grupta bulunan toplam 35.547 olan kelime sayısı bu işlemin ardından 19.557 kelimeye düşmüştür. 2. grupta bulunan toplam 34.446 olan kelime sayısı bu işlemin ardından 12.670 kelimeye düşmüştür.

Veri tabanında bulunan kelimelerin kökleri manuel olarak (elle) ilgili tablolardaki kök sütununa girilmiştir. “spambirgrup” tablosunda bulunan kök kelimelerin sayısı 2.351 olarak belirlenirken “spamikigrup” tablosunda bulunan kök kelimelerin sayısı ise 1.997 olarak belirlenmiştir.

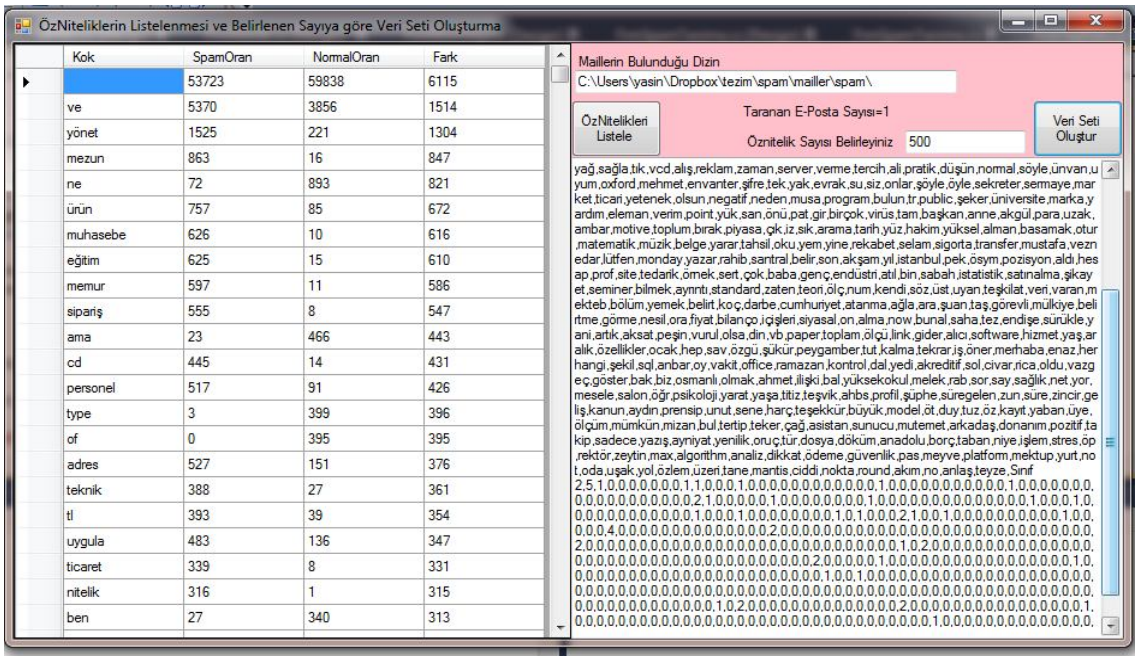


Şekil 5.5. E-postaların kelimelere ayrılması ve veri tabanına yazılması

Kök kelimelerin veri tabanına kaydedilmesi ve kelimelerin temizlenip, birleştirilmesinden sonra veri tabanında bulunan kök kelimelerin (özniteliklerin) kullanım sıklık değerleri belirlenmeye çalışılmıştır. Bu işlem için "spambirgrup" tablosunda bulunan kök kelimelerin kullanım sıklık değerlerini bulmak için 1. grupta bulunan bütün e-postalar yeniden okunmuştur. E-postaların okunma işlem adımları:

1. İlk adım olarak e-postalarda geçen html etiketleri silinmiş ardından e-posta içerisinde geçen kelimeler ayrı listelere alınmıştır.
2. Kelime listelerinde bulunan kelimelerin veri tabanında bulunan kök değerleri ile eşleşmesi durumunda özniteliğin (kök kelimenin) sütun değeri 1 artırılmıştır.
3. E-posta içerisinde geçen bütün kelimeler için aynı işlem uygulanmıştır. Kök kelimelerin bir e-postadaki kullanım sıklık değerleri tespit edildikten sonra veri kümesinin en son sütununa eklenen sınıf sütununa okunan e-posta spam ise 1, normal ise 0 yazılmıştır.
4. Bu işlem 1. grup için ayrılan bütün spam ve normal e-postalara uygulanarak kök kelimelerin sıklık değerleri bulunmuş ve 1. grup için bir veri kümesi elde edilmiştir. Bu işlemler Şekil 5.6 da gösterilmiştir.

Yukarıda anlatılan işlemler 2. grup e-postalara da aynı şekilde uygulanmış ve 2. grup için bir veri kümesi elde edilmiştir.



Şekil 5.6. Özniteliklere göre e-postaların okunması

Veri tabanında bulunan ilgili tablolarda bulunan tüm özniteliklerin 2 farklı grupta bulunan e-postalardaki kullanım sıklıklarına göre veri kümeleri oluşturulmuştur. Şekil 5.7'de oluşturulan veri kümesinin küçük bir örneği gösterilmiştir.

```
eğitim,muhasebe,yönet,sipariş,mezun,işletme,kurum,kadro,ürün,Sınıf
0,0,5,1,0,0,0,1,0,0,1
2,2,24,0,1,3,0,3,0,0,1
2,2,24,0,1,3,0,3,0,0,1
5,3,5,1,1,2,0,1,0,0,1
1,0,11,0,1,2,0,3,0,0,1
3,2,24,0,1,3,1,3,0,0,0
2,2,24,0,1,3,0,3,0,0,0
0,0,3,0,1,3,0,0,2,0,0
2,5,11,1,2,19,3,6,3,0,0
```

Şekil 5.7. E-postalar okunduktan sonra oluşan örnek veri kümesi

Çalışmada Türkçe kelimelerin morfolojik analizi yapılmamıştır. Bunun yerine e-postalardaki kelimeler okunmuş ve veri tabanındaki kelimeler ile kıyaslanarak kök kelimeleri bulunmuştur. Bu kıyaslama işlemi SQL komutları ile gerçekleştirilmiştir.

5.2. Özniteliklerin Çıkarımı

Her iki farklı grup için özniteliklerin kullanım sıklıkları elde edildikten sonra her iki farklı grup için öznitelik çıkarım işlemine geçilmiştir. Öznitelik çıkarımı için kullanılan yöntemler her iki farklı grup için elde edilen veri kümesine uygulanmıştır. Öznitelik çıkarımı için FAA ve Öklid uzaklığı kullanılmıştır.

5.2.1. Fisher ayrımsallık analizi

Fisher ayrımsallık analizi (FAA) ilk olarak 1936 yılında Fisher tarafından önerilmiştir [67]. FAA çeşitli uygulamalarda kullanılmıştır. Örneğin Wang [68] yüksek lisans tezinde,

Ataş ve ark.,'ları pul biberlerde aflatoksin tespiti için yaptıkları çalışmada [69] FAA yöntemini, öznitelik seçiminde başarılı olarak uygulamışlardır.

FAA ile en iyi ayrımsallık veren öznitelikler belirlenmeye çalışılmıştır. Bu özniteliklerin belirlenmesi için 5.1'de verilen FAA formülü elde edilen veri kümesinde bütün özniteliklere uygulanmıştır.

$$F_{da} = \frac{|m1-m2|^2}{s1^2+s2^2} \tag{5.1}$$

m1= Spam verilerin ortalama değeri

m2= Normal verilerin ortalama değeri

s1= Spam verilerin standart sapma değeri

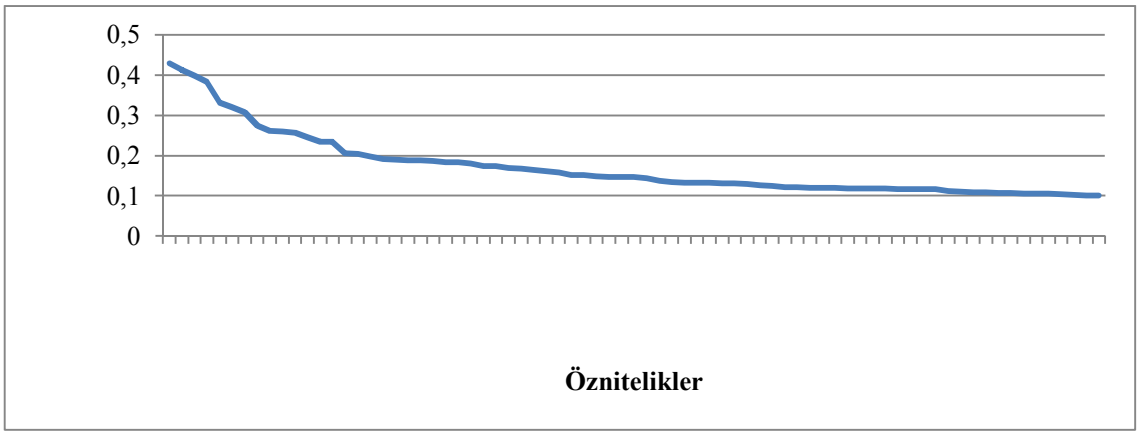
s2= Normal verilerin standart sapma değeri

Veri kümesine 5.1'de verilen FAA formülü uygulandıktan sonra bütün öznitelikler yeni sonuç değerlerine göre en yüksek sonuçtan en düşük sonuca göre sıralanmıştır. FAA sonuç değeri 0,1 ve üzerinde olanlar yeni öznitelikler olarak kabul edilmiş geriye kalan öznitelikler ayırt edilmiştir. Buna göre 1. grupta 75 adet öznitelik, 2. grupta ise 69 adet yeni öznitelik tespit edilmiştir. Tespit edilen öznitelikler Tablo 5.2 de gösterilmiştir.

Tablo 5.2. Özniteliklerin FAA yöntemine göre listesi

1. grupta bulunan FAA öznitelikleri				**2. grupta bulunan FAA öznitelikleri**			
Öznitelik Adı	FAA değeri	Öznitelik Adı	FAA değeri	Öznitelik Adı	FAA değeri	Öznitelik Adı	FAA değeri
yetki	0,429	ödeme	0,147	olarak	0,551	bilgi	0,254
kurum	0,413	hazır	0,143	talep	0,506	birçok	0,252
sipariş	0,398	ben	0,138	kurum	0,488	ülke	0,251
eğitim	0,383	print	0,133	saygı	0,457	utan	0,247
ürün	0,331	format	0,133	veysi	0,436	alım	0,247
tl	0,320	arkadaş	0,132	en	0,418	bu	0,244
mezun	0,308	dvd	0,131	alma	0,413	arz	0,244
ünvan	0,274	var	0,131	hizmet	0,410	girişim	0,243
iktisat	0,260	şirket	0,130	bölüm	0,404	artık	0,237
teslim	0,259	iletişim	0,129	amaç	0,399	bilim	0,237

adres	0,257	bana	0,125	dağ	0,362	temel	0,232
cd	0,245	başlık	0,124	için	0,360	eğitim	0,229
kargo	0,234	mağdur	0,121	memur	0,356	nişan	0,229
fiyat	0,234	kadar	0,121	uzman	0,356	unsur	0,225
işlet	0,206	konu	0,120	yer	0,340	nitelik	0,223
memur	0,204	onu	0,120	konu	0,324	değiş	0,222
video	0,198	ıslat	0,119	mezun	0,322	oluş	0,222
kdv	0,191	oksit	0,118	ad	0,321	oldu	0,222
özde	0,189	eğer	0,117	olmak	0,320	hem	0,221
kavram	0,188	personel	0,117	gerek	0,319	merkez	0,221
aşağı	0,188	kod	0,117	sahip	0,318	yak	0,219
doküman	0,186	ayrıntı	0,117	büyük	0,317	arttır	0,215
telefon	0,183	oldu	0,116	iktisat	0,315	olan	0,213
işletme	0,183	vergi	0,116	bir	0,307	hakkı	0,211
kadro	0,181	geliş	0,116	kadro	0,307	iş	0,211
ne	0,174	toplam	0,111	belirt	0,302	genel	0,210
alım	0,173	yeter	0,110	yap	0,298	ekonomi	0,210
nitelik	0,168	kal	0,109	maliye	0,296	sabır	0,208
telafi	0,168	gün	0,108	idare	0,286	ambar	0,2057
firma	0,164	http	0,106	devlet	0,282	mekteb	0,201
daha	0,161	eğitmen	0,106	al	0,279	şikayet	0,200
atama	0,158	kitap	0,105	atama	0,278	peşin	0,2
temel	0,151	ters	0,105	ilgi	0,278		
ihtiyaç	0,151	olsa	0,105	ediyor	0,276		
hoca	0,148	ambar	0,104	diğer	0,275		
bölüm	0,147	deney	0,102	kod	0,266		
oluş	0,147	can	0,101	etti	0,257		



Şekil 5.8. 1. grup özniteliklerin FAA yöntemine göre grafiksel gösterimi

5.2.2. Öklid uzaklığı

Öklid uzaklığı iki nokta arasındaki lineer uzaklıktır. n boyutlu Öklid uzayında $P=(P_1,P_2,P_3,\ldots.P_N)$, $Q=(Q_1,Q_2,Q_3,\ldots Q_N)$ noktaları arasındaki mesafe aşağıdaki formülle bulunur.

$$\text{Mesafe} = \sqrt{\sum_{k=1}^{n}(P_k - Q_k)^2} \tag{5.2}$$

Bu çalışmada spam ve normal değerleri tek boyutlu olduğundan özniteliklerin seçimi için Öklid bağlantısı uzunluk kavramına eş değer şekilde kullanılmıştır.

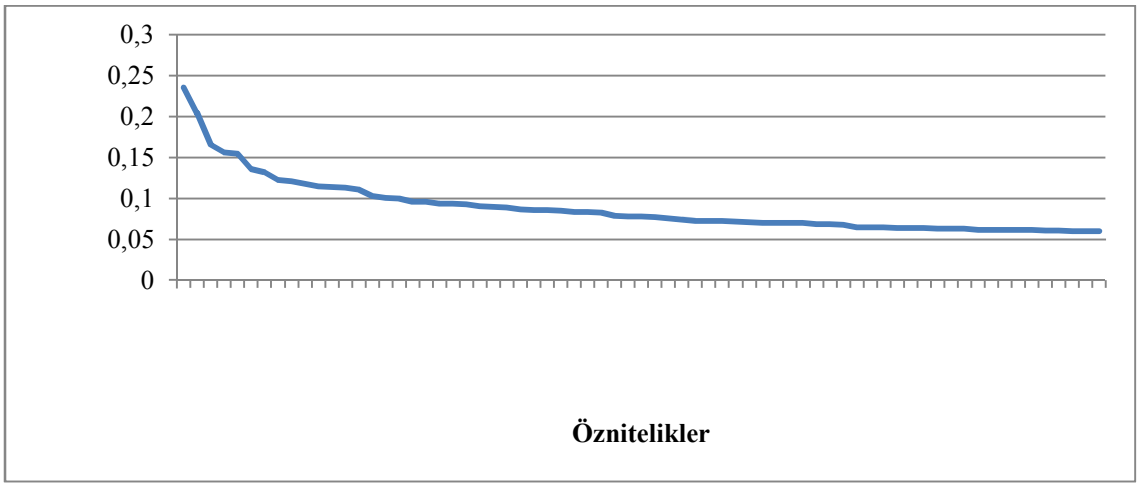
$$Mesafe = |P_{spam} - P_{nonspam}| \tag{5.3}$$

Veri kümesine 5.3 numaralı Öklid uzaklık formülü uygulandıktan sonra bütün öznitelikler yeni sonuçlara göre büyükten küçüğe doğru sıralanmıştır. Sıralama işleminden sonra Öklid sonuç değeri 0,05 değerinden büyük olanlar yeni öznitelikler olarak kabul edilmiş geriye kalan öznitelikler çıkarılmıştır. Buna göre 1. grupta 74 adet öznitelik, 2. grupta ise 67 adet yeni öznitelik listesi tespit edilmiş ve Tablo 5.3 de gösterilmiştir.

Tablo 5.3. Özniteliklerin Öklid uzaklık yöntemine göre listesi

1. gruptaki Öklid öznitelikleri				**2. gruptaki Öklid öznitelikleri**			
Öznitelik	Değeri	Öznitelik	Değeri	Öznitelik	Değeri	Öznitelik Adı	Değeri
yetki	0,236	maliye	0,074	birçok	0,215	etti	0,127
sipariş	0,204	video	0,073	saygı	0,198	bin	0,125
ürün	0,166	ödeme	0,073	olmak	0,198	çoğunluk	0,124
eğitim	0,156	temel	0,073	mekteb	0,182	sürükle	0,119
cd	0,155	yeter	0,072	alma	0,173	fakülte	0,119
kurum	0,136	idare	0,071	amaç	0,173	nitelik	0,119
teslim	0,132	kdv	0,070	maliye	0,170	ad	0,118
ünvan	0,123	özet	0,070	peşin	0,166	utan	0,118
adres	0,121	tanı	0,070	ediyor	0,164	oluş	0,117
telefon	0,118	fiyat	0,070	kadro	0,164	şuan	0,116
ayrıntı	0,115	hoca	0,069	sahip	0,162	ambar	0,112
kavram	0,114	diler	0,069	devlet	0,160	olarak	0,110

işlet	0,113	doldur	0,068	uzman	0,155	sefer	0,109
oluş	0,111	vergi	0,065	diğer	0,154	orjinal	0,109
telafi	0,103	diksiyon	0,065	lisans	0,153	şükür	0,109
dvd	0,101	vade	0,065	aksat	0,148	satınalma	0,107
ambar	0,100	ters	0,064	iktisat	0,148	girişim	0,107
konu	0,096	bölüm	0,064	istifa	0,146	çağ	0,107
psikoloji	0,096	personel	0,064	ülke	0,145	veznedar	0,107
aşağı	0,094	teknik	0,063	yetiş	0,145	bölüm	0,106
doküman	0,094	akreditif	0,063	belirt	0,141	artık	0,106
memur	0,093	alım	0,063	gerek	0,141	memur	0,106
ihtiyaç	0,091	mağdur	0,062	vurul	0,140	uyum	0,105
kargo	0,090	tahvil	0,062	veysi	0,137	titiz	0,104
iletişim	0,089	nitelik	0,062	teşkilat	0,137	donanım	0,103
format	0,087	alma	0,062	kriter	0,136	kasım	0,102
işletme	0,086	muhasebe	0,062	darbe	0,133	kod	0,101
mezun	0,086	alan	0,061	sabır	0,133	zor	0,101
eğitmen	0,085	düzey	0,061	bunal	0,133	hep	0,100
iktisat	0,084	fon	0,060	ciddi	0,133	talep	0,100
tl	0,084	alıcı	0,060	siyasal	0,133		
amaç	0,083	değer	0,060	vazgeç	0,133		
çıkmak	0,079	hedef	0,060	bilim	0,131		
firma	0,078	yönet	0,060	büyük	0,129		
toplam	0,078	kazan	0,060	endişe	0,129		
özde	0,077	motive	0,060	dağ	0,127		
geliş	0,076	teşvik	0,060	al	0,127		



Şekil 5.9. 1. grup özniteliklerin Öklid uzaklığına göre grafiksel gösterimi

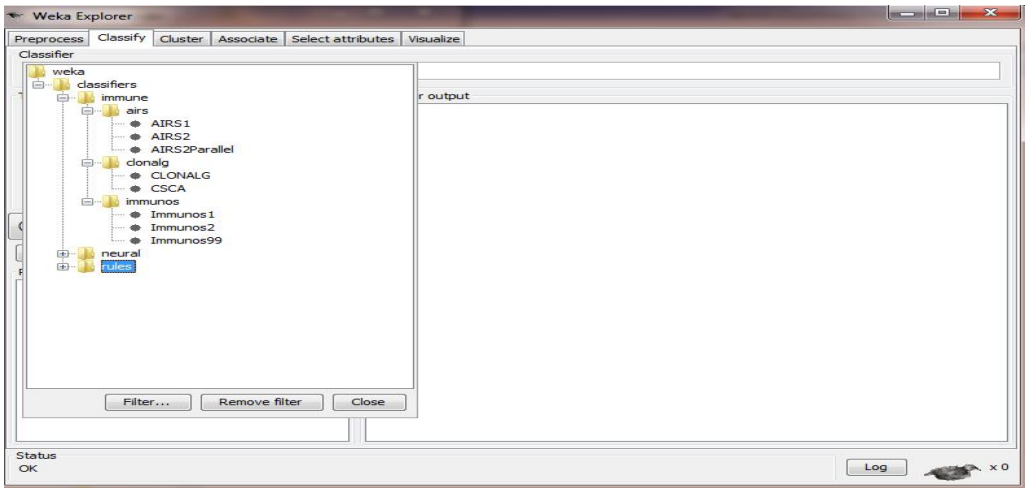
5.2. Kullanılan YBS programı

WEKA Waikato Üniversitesi'nde geliştirilmiş açık kaynak kodlu bir uygulamadır. Bu uygulama makine öğrenmesi ve veri madenciliği algoritmalarını içermektedir. WEKA, platformdan bağımsız olarak çalışmaktadır.

WEKA uygulaması içinde YBS algoritmalarını kullanmak için, Jason Brownlee tarafından WEKA ya eklenti olarak yazılan ve YBS algoritmalarını içeren yazılım http://sourceforge.net/projects/wekaclassalgos adresinden indirilmiştir.

İndirilen WEKA eklentisinde bulunan YBS algoritmaları:

1) AIRS
    (a) AIRS1
    (b) AIRS2
    (c) AIRS2PARALLEL
2) CLONALG
    (a) CLONALG
    (b) CSCA
3) IMMUNOS
    (a) IMMUNOS1
    (b) IMMUNOS2
    (c) IMMUNOS99



Şekil 5.10. Weka programında kullanılan YBS algoritmaları

5.3. Performans Ölçütleri

Kullanılan sınıflandırıcıların performansını ölçmek için aşağıdaki ölçütler kullanılmıştır.

Tablo 5.3. Performans ölçütleri

Gerçek Sistemde / Tahmini Sistemde	**Spam (Evet- 1)**	**Normal (Hayır- -1)**
Spam E-Posta (1)	DP (Doğru Pozitif)	YP (Yanlış Pozitif)
Normal E-Posta (-1)	YN (Yanlış Negatif)	DN (Doğru Negatif)
Toplam	P	N

Doğru Pozitif: Gerçekte spam olarak bilinen e-postaların, filtre tarafından spam olarak belirlendiği e-posta sayısı

Doğru Negatif: Gerçekte Normal E-Posta olarak bilinen e-postaların, filtre tarafından normal e-posta olarak belirlendiği e-posta sayısı

Yanlış Pozitif: Gerçekte normal e-posta olarak bilinen e-postaların, filtre tarafından spam e-posta olarak belirlendiği e-posta sayısı

Yanlış Negatif: Gerçekte spam e-posta olarak bilinen e-postaların, filtre tarafından normal e-posta olarak belirlendiği e-posta sayısı

P= Toplam spam e-posta sayısı

N= Toplam normal e-posta sayısı

$$\text{Özgünlük} = \frac{DN}{\text{DN+YP}} * 100 \qquad \text{Duyarlılık} = \frac{DP}{\text{DP+YN}} * 100$$

$$\text{Doğruluk} = \frac{DP+D}{\text{DP+YP+YN+DN}} * 100 \qquad \text{Kesinlik} = \frac{DP}{\text{DP+YP}} * 100$$

$$\text{Hata} = \frac{YP+YN}{\text{DP+YP+YN+DN}} * 100$$

Yukarıda verilen bilgileri bir örnek üzerinde açıklayalım.

Tablo 5.4. Performans ölçütleri örneği

		Gerçek Sınıflandırma	
		Spam	Normal
Tahmini Sınıflandırma	Spam (Pozitif)	15	5
	Normal(Negatif)	3	17
TOPLAM		18 (P)	22 (N)

Bu verilere göre 18 adet spam e-postadan 15 tanesi doğru sınıflandırılmış, 3 tanesi ise yanlış sınıflandırılmıştır. Normal 22 e-postadan 17 tanesi doğru, 5 tanesi ise yanlış sınıflandırılmıştır.

$$\text{DP Oranı} = \frac{15}{18} = 0{,}83 \qquad \text{DN Oranı} = \frac{17}{22} = 0{,}77$$

$$\text{YN Oranı} = \frac{3}{18} = 0{,}16 \qquad \text{YP Oranı} = \frac{5}{22} = 0{,}23 \qquad \text{Hata} = \frac{5+3}{40} = 0{,}2$$

$$\text{Doğruluk Oranı} = \frac{DP+DN}{\text{DP+DN+YN+YP}} = \frac{15+17}{18+22} = 0{,}8 \qquad \text{Kesinlik} = \frac{15}{15+5} = 0{,}75$$

Duyarlılık =DP oranı, özgünlük= DN oranıdır.

$$\text{F} - \text{Ölçütü} = 2 * \frac{Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık}$$

Tablo 5.5. Detaylı doğruluk sınıflandırma ölçütleri

DP Oranı	YP Oranı	Kesinlik	Duyarlılık	F-Ölçütü	ROC Eğrisi	Class
0,83	0,16	0,75	0,83	0,789	0,8	1 (Spam)
0,77	0,23	0,85	0,77	0,808	0,8	0 (Normal)

6. BAYES VE YAPAY BAĞIŞIKLIK SİSTEM ALGORİTMALARININ SINIFLANDIRMA PERFORMANSLARI

Tezin bu bölümünde Naive Bayes ve YBS algoritmalarının sınıflandırma performansları ele alınmıştır. YBS sınıflandırma algoritmalarından 4. bölümde anlatılan CLONALG, CSCA, AIRS1, AIRS2 ve AIRS2Parallel algoritmaları kullanılmıştır. Yapılacak deneylerde kullanılacak veri kümelerini oluşturma işlemi 5. bölümde anlatılmıştır.

Bu bölümde sınıflandırma algoritmalarının başarılarını ölçmek için iki farklı deney yapısı kullanılmıştır. Birinci deney grubunda, kullanılan algoritmaların parametrelerinde herhangi bir değişiklik yapılmamıştır. Varsayılan parametre değerleri kullanılmıştır. İkinci deney grubunda, sınıflandırma için kullanılan algoritmaların parametre değerleri değiştirilmiş ve en iyi sonucu veren parametre değerleri tespit edilmeye çalışılmıştır. Yapılan tüm deneylerde dört farklı veri kümesi kullanılmıştır. Veri kümeleri FAA yöntemi ile edilen 1. grup veri kümesi ile 2. grup veri kümesi, Öklid yöntemi ile elde edilen 1. grup veri kümesi ile 2. grup veri kümelerinden oluşmaktadır. Yapılan bütün deneylerde veri kümesine 10 kat çapraz doğrulama işlemi uygulanmıştır. Yapılan çapraz doğrulama işlemine göre veri kümesi 10 kümeye bölünmüştür. Bir çapraz doğrulamada 10 kümenin 1 kümesi test, 9 kümesi ise eğitim için kullanılır. Her küme 1 kez test için geriye kalan 9 küme ise eğitim için kullanılır.

Birinci deney grubunda farklı veri kümeleri için YBS algoritmalarının sınıflandırma başarı oranları Tablo 6.1, Tablo 6.2, Tablo 6.3, Tablo 6.4 ve Tablo 6.5'te gösterilmiştir. Bu işlemlerden sonra ikinci deney grubuna geçilmiş ve YBS algoritmalarının ikinci deney grubunda sınıflandırma başarı oranları Tablo 6.8, Tablo 6.9, Tablo 6.10, Tablo 6.11 ve Tablo 6.12'de gösterilmiştir. Her bir öznitelik seçme yönteminden 2 farklı veri kümesi elde edildiğinden YBS algoritmalarının bu veri kümeleri üzerindeki sınıflandırma başarılarının ortalaması alınmış ve bu ortalama değerlerin standart sapma değerleri hesaplanmıştır. Naive Bayes sınıflandırma başarı oranı Tablo 6.13'de gösterilmiştir.

Tablo 6.1. AIRS1 sınıflandırma algoritması sonuçları

			DP Oranı	YP Oranı	Kesinlik	Duyarlılık	F-Ölçütü	ROC Eğrisi
Spam	FAA	Grup 1	0,593	0,241	0,69	0,593	0,637	0,676
		Grup 2	0,611	0,342	0,616	0,611	0,613	0,634
	Öklid	Grup 1	0,589	0,194	0,733	0,589	0,653	0,697
		Grup 2	0,552	0,312	0,613	0,552	0,581	0,62
	Ortalama	FAA	0,602±1					
		Öklid	0,570±1,5					
Normal	FAA	Grup 1	0,759	0,407	0,674	0,759	0,714	0,676
		Grup 2	0,658	0,389	0,653	0,658	0,656	0,634
	Öklid	Grup 1	0,806	0,411	0,685	0,806	0,74	0,697
		Grup 2	0,688	0,448	0,631	0,688	0,658	0,62
	Ortalama	FAA	0,708±5					
		Öklid	0,747±6					

Tablo 6.2. AIRS2 sınıflandırma algoritması sonuçları

			DP Oranı	YP Oranı	Kesinlik	Duyarlılık	F-Ölçütü	ROC Eğrisi
Spam	FAA	Grup 1	0,596	0,094	0,852	0,596	0,702	0,751
		Grup 2	0,678	0,123	0,832	0,678	0,747	0,777
	Öklid	Grup 1	0,544	0,043	0,919	0,544	0,684	0,75
		Grup 2	0,63	0,233	0,708	0,63	0,667	0,699
	Ortalama	FAA	0,637±4					
		Öklid	0,587±4,5					
Normal	FAA	Grup 1	0,906	0,404	0,713	0,906	0,798	0,751
		Grup 2	0,877	0,322	0,752	0,877	0,81	0,777
	Öklid	Grup 1	0,957	0,456	0,699	0,957	0,808	0,75
		Grup 2	0,767	0,37	0,698	0,767	0,731	0,699
	Ortalama	FAA	0,891±1,5					
		Öklid	0,862±9,5					

Tablo 6.3. AIRS2Parallel sınıflandırma algoritması sonuçları

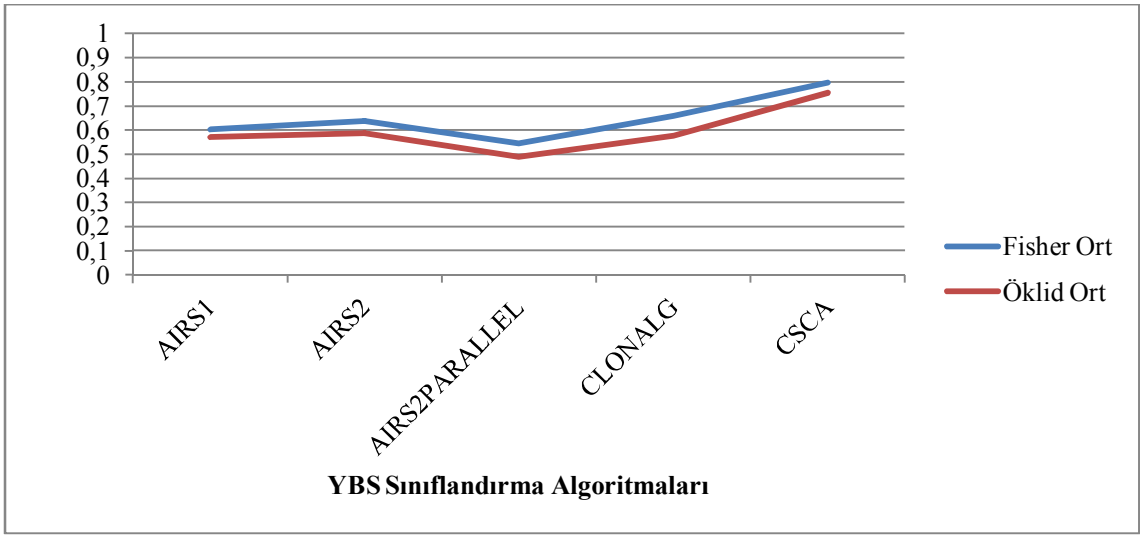
			DP Oranı	YP Oranı	Kesinlik	Duyarlılık	F-Ölçütü	ROC Eğrisi
Spam	FAA	Grup 1	0,578	0,07	0,881	0,578	0,698	0,754
		Grup 2	0,515	0,056	0,891	0,515	0,653	0,729
	Öklid	Grup 1	0,5	0,027	0,944	0,5	0,654	0,737
		Grup 2	0,478	0,113	0,791	0,478	0,596	0,682
	Ortalama	FAA	0,546±3					
		Öklid	0,489±1,5					
Normal	FAA	Grup 1	0,93	0,422	0,709	0,93	0,805	0,754
		Grup 2	0,944	0,485	0,684	0,944	0,793	0,729
	Öklid	Grup 1	0,973	0,5	0,683	0,973	0,803	0,737
		Grup 2	0,887	0,522	0,654	0,887	0,753	0,682
	Ortalama	FAA	0,937±0,5					
		Öklid	0,93±4,5					

Tablo 6.4. CLONALG sınıflandırma algoritması sonuçları

			DP Oranı	YP Oranı	Kesinlik	Duyarlılık	F-Ölçütü	ROC Eğrisi
Spam	FAA	Grup 1	0,641	0,194	0,749	0,641	0,691	0,723
		Grup 2	0,678	0,309	0,663	0,678	0,67	0,684
	Öklid	Grup 1	0,622	0,227	0,712	0,622	0,664	0,697
		Grup 2	0,533	0,342	0,583	0,533	0,557	0,596
	Ortalama	FAA	0,659±1,5					
		Öklid	0,577±4,5					
Normal	FAA	Grup 1	0,806	0,359	0,713	0,806	0,757	0,723
		Grup 2	0,691	0,322	0,705	0,691	0,698	0,684
	Öklid	Grup 1	0,773	0,378	0,694	0,773	0,731	0,697
		Grup 2	0,658	0,467	0,611	0,658	0,634	0,596
	Ortalama	FAA	0,748±5,5					
		Öklid	0,715±6					

Tablo 6.5. CSCA sınıflandırma algoritması sonuçları

			DP Oranı	YP Oranı	Kesinlik	Duyarlılık	F-Ölçütü	ROC Eğrisi
Spam	FAA	Grup 1	0,767	0,08	0,896	0,767	0,826	0,843
		Grup 2	0,826	0,073	0,91	0,826	0,866	0,876
	Öklid	Grup 1	0,726	0,054	0,925	0,726	0,813	0,836
		Grup 2	0,785	0,103	0,872	0,785	0,827	0,841
	Ortalama	FAA	0,796±3					
		Öklid	0,755±3					
Normal	FAA	Grup 1	0,92	0,233	0,814	0,92	0,863	0,843
		Grup 2	0,927	0,174	0,856	0,927	0,89	0,876
	Öklid	Grup 1	0,946	0,274	0,793	0,946	0,863	0,836
		Grup 2	0,897	0,215	0,823	0,897	0,859	0,841
	Ortalama	FAA	0,923±0					
		Öklid	0,921±2,5					



Şekil 6.1. 1. deneyde YBS algoritmalarının spam e-postaları sınıflandırma grafiği

Yapılan birinci deney grubu sonuçlarında Öklid uzaklığı ve FAA yöntemi ile elde edilen özniteliklerden oluşan veri kümesi üzerinde yapılan sınıflandırma oranları verilmiştir. Bu oranlara bakıldığında spam ve normal e-postaların sınıflandırılmasında FAA yöntemine göre seçilen öznitelikler, Öklid uzaklığı yöntemine göre seçilen özniteliklere göre daha başarılıdır denilebilir.

Birinci deney sonuçlarında CSCA sınıflandırma algoritması en iyi sonucu veren algoritma olarak tespit edilmiştir. CSCA algoritması, FAA yöntemi ile elde edilen özniteliklerden oluşan veri kümesindeki spam e-postaları yaklaşık %80 oranında doğru sınıflandırırken, normal e-postaları yaklaşık %76 oranında doğru sınıflandırmıştır.

Bu çalışmada kullanılan algoritmaların parametre değerleri değiştirilerek sınıflandırma başarısına etkileri incelenmiştir. İkinci deney grubunda algoritmaların parametre değerleri değiştirilerek en iyi sınıflandırma sonucunu veren parametre değerleri tespit edilmeye çalışılmıştır. Tablo 6.6'da sadece CSCA algoritması için kullanılan parametrelerin farklı değerleri ve sonuçları paylaşılmış ve en iyi sonucu veren parametre değerleri gösterilmiştir. Diğer algoritmalar için bütün deney sonuçları yazılmamış bunun yerine en iyi sonucu veren parametre değerleri yazılmıştır.

YBS algoritmaları ve bu algoritmaların parametrelerindeki değer değişimlerine göre en iyi sınıflandırma oranını veren parametreler ve parametre değerleri Tablo 6.7'de verilmiştir.

Tablo 6.6. CSCA sınıflandırma algoritması parametre değişimlerine göre sonuçlar

KNN	İnitial PopulationSize	numPartitions	seed	total Generations	Başarı	TP %	TN %
1	50	1	1	5	0,831	0,75	0,904
1	50	1	1	6	0,803	0,69	0,904
1	50	1	1	7	0,808	0,72	0,886
1	50	1	1	8	0,803	0,69	0,904
…	…	…	…	…	…	…	…
1	50	1	1	19	0,859	0,81	0,904
…	…	…	…	…	…	…	…
1	30	1	1	7	0,859	0,81	0,904
Diğer parametre değişimlerinde kırmızı ile gösterilen değerden yüksek olan sonuçlar çıkmadığı için geri kalan sonuçlar yazılmamıştır.							

Tablo 6.7. YBS algoritmalarının en iyi sonucu veren parametre değerleri

Algoritmalar	**Parametre**	**Değer**
AIRS1,AIRS2,AIRS2PARALLEL	memInitialPoolSize	50
	numInstancesAffinityThreshold	1
	stimulationValue	0,5
CLONALG	antibodyPoolSize	50
	selectionPoolSize	10
	totalReplacement	2
CSCA	initialPopulationSize	30
	totalGenerations	5

Algoritmaların parametre değerlerinin değişiminden sonra Öklid ve FAA yöntemine göre elde edilen veri kümelerinin YBS algoritmaları ile sınıflandırılmasında, sonuçların daha başarılı çıktığı Tablo 6.8, Tablo 6.9, Tablo 6.10, Tablo 6.11 ve Tablo 6.12'de verilen tablolarda gösterilmiştir.

Tablo 6.8. AIRS1 sınıflandırma algoritması parametre değişim sonuçları

			DP Oranı	YP Oranı	Kesinlik	Duyarlılık	F-Ölçütü	ROC Eğrisi
Spam	FAA	Grup 1	0,674	0,144	0,809	0,674	0,735	0,765
		Grup 2	0,659	0,12	0,832	0,659	0,736	0,77
	Öklid	Grup 1	0,593	0,06	0,899	0,593	0,714	0,766
		Grup 2	0,615	0,13	0,81	0,615	0,699	0,743
	Ortalama	FAA	0,666±1					
		Öklid	0,604±1					
Normal	FAA	Grup 1	0,856	0,326	0,744	0,856	0,796	0,765
		Grup 2	0,88	0,341	0,742	0,88	0,805	0,77
	Öklid	Grup 1	0,94	0,407	0,719	0,94	0,814	0,766
		Grup 2	0,87	0,385	0,716	0,87	0,786	0,743
	Ortalama	FAA	0,868±1,5					
		Öklid	0,905±3,5					

Tablo 6.9. AIRS2 sınıflandırma algoritması parametre değişim sonuçları

			DP Oranı	YP Oranı	Kesinlik	Duyarlılık	F-Ölçütü	ROC Eğrisi
Spam	FAA	Grup 1	0,752	0,154	0,815	0,752	0,782	0,799
		Grup 2	0,793	0,133	0,843	0,793	0,817	0,83
	Öklid	Grup 1	0,726	0,161	0,803	0,726	0,763	0,783
		Grup 2	0,759	0,183	0,788	0,759	0,774	0,788
	Ortalama	FAA	0,772±2					
		Öklid	0,742±1,5					
Normal	FAA	Grup 1	0,846	0,248	0,791	0,846	0,817	0,799
		Grup 2	0,867	0,207	0,823	0,867	0,845	0,83
	Öklid	Grup 1	0,839	0,274	0,772	0,839	0,804	0,783
		Grup 2	0,817	0,241	0,791	0,817	0,804	0,788
	Ortalama	FAA	0,856±1					
		Öklid	0,828±1					

Tablo 6.10. AIRS2Parallel sınıflandırma algoritması parametre değişim sonuçları

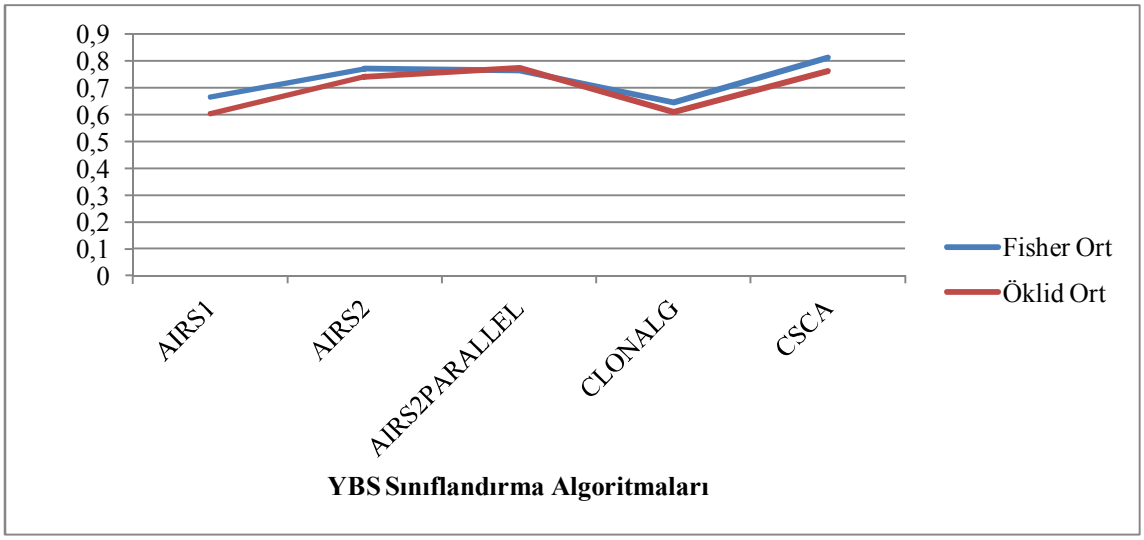
			DP Oranı	YP Oranı	Kesinlik	Duyarlılık	F-Ölçütü	ROC Eğrisi
Spam	FAA	Grup 1	0,748	0,164	0,805	0,748	0,775	0,792
		Grup 2	0,781	0,116	0,858	0,781	0,818	0,833
	Öklid	Grup 1	0,763	0,147	0,824	0,763	0,792	0,808
		Grup 2	0,785	0,166	0,809	0,785	0,797	0,81
	Ortalama	FAA	0,764±2					
		Öklid	0,774±1					
Normal	FAA	Grup 1	0,836	0,252	0,786	0,836	0,81	0,792
		Grup 2	0,884	0,219	0,818	0,884	0,85	0,833
	Öklid	Grup 1	0,853	0,237	0,799	0,853	0,825	0,808
		Grup 2	0,834	0,215	0,812	0,834	0,823	0,81
	Ortalama	FAA	0,86±2,5					
		Öklid	0,843±1					

Tablo 6.11. CLONALG sınıflandırma algoritması parametre değişim sonuçları

			DP Oranı	YP Oranı	Kesinlik	Duyarlılık	F-Ölçütü	ROC Eğrisi
Spam	FAA	Grup 1	0,652	0,221	0,727	0,652	0,688	0,716
		Grup 2	0,641	0,306	0,653	0,641	0,647	0,668
	Öklid	Grup 1	0,67	0,211	0,742	0,67	0,704	0,73
		Grup 2	0,552	0,286	0,634	0,552	0,59	0,633
	Ortalama	FAA	0,646±0,5					
		Öklid	0,611±6					
Normal	FAA	Grup 1	0,779	0,348	0,713	0,779	0,744	0,716
		Grup 2	0,694	0,359	0,683	0,694	0,689	0,668
	Öklid	Grup 1	0,789	0,33	0,726	0,789	0,756	0,73
		Grup 2	0,714	0,448	0,64	0,714	0,675	0,633
	Ortalama	FAA	0,736±4					
		Öklid	0,751±3,5					

Tablo 6.12. CSCA sınıflandırma algoritması parametre sınıflandırma sonuçları

			DP Oranı	YP Oranı	Kesinlik	Duyarlılık	F-Ölçütü	ROC Eğrisi
Spam	FAA	Grup 1	0,785	0,07	0,91	0,785	0,843	0,857
		Grup 2	0,841	0,09	0,894	0,841	0,866	0,876
	Öklid	Grup 1	0,733	0,064	0,912	0,733	0,813	0,835
		Grup 2	0,793	0,13	0,846	0,793	0,818	0,832
	Ortalama	FAA	0,813±3					
		Öklid	0,763±3					
Normal	FAA	Grup 1	0,93	0,215	0,827	0,93	0,876	0,857
		Grup 2	0,91	0,159	0,864	0,91	0,887	0,876
	Öklid	Grup 1	0,936	0,267	0,795	0,936	0,86	0,835
		Grup 2	0,87	0,207	0,824	0,87	0,847	0,832
	Ortalama	FAA	0,92±1					
		Öklid	0,903±1					



Şekil 6.2. 2. deneydeki YBS algoritmalarının spam e-postaları sınıflandırma grafiği

Öklid ve FAA yöntemine göre elde edilen veri kümelerinin sınıflandırma başarısını Naive Bayes yöntemini kullanarak incelediğimizde Tablo 6.13'te verilen sonuçlar elde edilmiştir.

Tablo 6.13. Naive Bayes sınıflandırma algoritması parametre sınıflandırma sonuçları

			DP Oranı	YP Oranı	Kesinlik	Duyarlılık	F-Ölçütü	ROC Eğrisi
Spam	FAA	Grup 1	0,685	0,047	0,93	0,685	0,789	0,893
		Grup 2	0,756	0,063	0,915	0,756	0,828	0,916
	Öklid	Grup 1	0,681	0,06	0,911	0,681	0,78	0,891
		Grup 2	0,626	0,086	0,867	0,626	0,727	0,873
	Ortalama	FAA	0,720					
		Öklid	0,653					
Normal	FAA	Grup 1	0,953	0,315	0,77	0,953	0,852	0,861
		Grup 2	0,937	0,244	0,81	0,937	0,869	0,921
	Öklid	Grup 1	0,94	0,319	0,766	0,94	0,844	0,878
		Grup 2	0,914	0,374	0,731	0,914	0,812	0,876
	Ortalama	FAA	0,945					
		Öklid	0,927					

İkinci deney grubundaki sınıflandırma sonuçları birinci deney grubundaki sınıflandırma sonuçlarından daha yüksek çıktığı görülmüştür.

Yapılan ikinci deneyde CSCA algoritması, FAA yöntemi ile elde edilen veri kümesindeki spam e-postaları ortalama %81 oranında doğru sınıflandırırken normal e-postaları %92 oranında doğru sınıflandırmıştır. CSCA algoritması Öklid yöntemi ile oluşturulan veri kümesindeki e-postaları %76 oranında doğru sınıflandırmıştır. CSCA algoritması spam e-postaları en yüksek oranda doğru sınıflandıran YBS algoritması olarak tespit edilmiştir.

AIRS2PARALLEL algoritması Öklid yöntemi ile elde edilen özniteliklerden oluşan veri kümesindeki spam e-postaları %77 oranında, FAA yöntemi ile elde edilen özniteliklerden oluşan veri kümesindeki e-postaları %76 oranında doğru sınıflandırmıştır.

Naive Bayes yönteminin sınıflandırma başarısına bakıldığında FAA yöntemi ile elde edilen özniteliklerden oluşan veri kümesindeki spam e-postaları ortalama %72 oranında doğru sınıflandırırken, normal e-postaları %94 oranında doğru sınıflandırmıştır. Öklid yöntemi ile elde edilen özniteliklerden oluşan veri kümesindeki spam e-postaları %65 oranında doğru sınıflandırırken normal e-postaları %92 oranında doğru sınıflandırmıştır. Tablo 6.16 da sınıflandırma algoritmalarının spam ve normal e-postalardaki başarılarının ikinci deney grubundaki ortalama değerleri verilmiştir.

Tablo 6.14. Sınıflandırma algoritmalarının ortalama sınıflandırma sonuçları

		FAA Ortalama DP Oranı	Öklid Ortalama DP Oranı
Spam	AIRS1	0,666	0,604
	AIRS2	**0,772**	0,742
	AIRS2PARALLEL	0,764	**0,774**
	CLONALG	0,646	0,611
	CSCA	**0,813**	**0,763**
	Naive Bayes	0,720	0,653
Normal	AIRS1	0,868	0,905
	AIRS2	0,856	0,828
	AIRS2PARALLEL	0,86	0,843
	CLONALG	0,736	0,751
	CSCA	**0,92**	**0,903**
	Naive Bayes	**0,945**	**0,927**

7. SONUÇLAR VE ÖNERİLER

Günümüzde en fazla kullanılan haberleşme araçlarından biri e-postalardır. TÜİK verilerine göre hane halkı bilişim teknolojileri kullanım araştırması 2012 verilerine göre internet kullanıcılarının %66,8 i interneti e-posta gönderme/alma için kullanmaktadır. E-postaların bu kadar yoğun kullanımı, maliyet olarak ucuz ve hızlı olması reklam, propaganda, yardım, korku, eğlence, kamuoyu oluşturmak isteyen kullanıcıların ilgisini her gün biraz daha çekmektedir.

EmailTray firması 2011-2012 yılları arasında yaptığı araştırmada her 100 e-postadan 68'inin spam e-postalarda oluştuğunu bildirmektedir [26]. Symantec Şubat 2012 İstihbarat raporunda en fazla spam e-postalara maruz kalan ülkenin Çin olduğu ve Çinli kullanıcıların e-postalarına gönderilen mesajların %74'ünün spam e-postalardan oluştuğu açıklanmıştır.

Spam e-postaların istatistiklerine bakıldığında, oluşturduğu maliyet ve zaman kaybı göz önüne alınırsa bununla mücadelenin çok kararlı bir şekilde sürdürülmesi gerektiği görülmektedir.

Tez çalışmasında Türkçe e-postalardan oluşan bir veri kümesi oluşturulmuştur. Bu veri kümesi için veri oluşturma, veri temizleme işlemleri yapılmıştır. Veri kümesi oluşturulduktan sonra öznitelik seçme işlemine geçilmiştir. Öznitelik seçme işleminde Öklid uzaklık ve FAA yöntemleri kullanılmıştır. Öznitelik seçme yöntemi ile elde edilen veri kümelerinin sınıflandırma başarımlarını belirmek için Naive Bayes ve YBS sınıflandırma algoritmaları kullanılmıştır. YBS algoritmalarının sınıflandırma sonuçları 2 bölüme ayrılmıştır. Birinci bölüm deneylerde algoritmaların varsayılan değerleri ile sınıflandırma sonuçları elde edilmiştir. İkinci bölüm deneyde ise algoritmaların parametre değerleri değiştirilerek sınıflandırma sonuçları elde edilmiştir. Bu karşılaştırma neticesinde kullanılan veri kümeleri için sınıflandırma algoritmalarının varsayılan değerlerinin sınıflandırma başarımında en iyi sonucu vermediği görülmüştür. Bu çalışmada kullanılan veri kümeleri için en iyi sınıflandırma sonucunu elde etmek için 5 farklı YBS algoritmasının en ideal parametre değerleri tespit edilmiştir.

Kullanılan sınıflandırıcı algoritmalarının sonuçlarına bakıldığında FAA yöntemine göre seçilen özniteliklerden elde edilen veri kümelerinin, Öklid yöntemine göre seçilen özniteliklerden elde edilen veri kümelerine göre e-postaları daha iyi sınıflandırdığı görülmüştür. Farklı veri kümelerinin ortalama sınıflandırma başarısına bakıldığında FAA

yönteminin öznitelik seçmek için Öklid yöntemine göre daha başarılı olduğu tespit edilmiştir.

YBS algoritmalarının parametre değerlerinin değişimi ile sınıflandırma başarısının arttığı görülmüştür.

FAA yöntemine göre seçilen özniteliklerden elde edilen veri kümesindeki normal e-postaları sınıflandırıcılardan Naive Bayes %94, CSCA algoritması %92 ve AIRS1 algoritması %86 oranında doğru sınıflandırmıştır.

Öklid yöntemine göre seçilen özniteliklerden elde edilen veri kümesindeki normal e-postaları sınıflandırıcılardan Naive Bayes %92, CSCA algoritması %90 ve AIRS1 algoritması %90 oranında doğru sınıflandırmıştır.

FAA yöntemine göre seçilen özniteliklerden elde edilen veri kümesindeki spam e-postaları, en iyi CSCA algoritması sınıflandırırken AIRS algoritmaları da CSCA algoritmasına yakın oranda doğru sınıflandırma başarısı göstermiştir.

FAA yöntemine göre seçilen özniteliklerden elde edilen veri kümesindeki spam e-postaları sınıflandırıcılardan Naive Bayes %72, CSCA algoritması %81 ve AIRS2 algoritması %77 oranında doğru sınıflandırmıştır.

Öklid yöntemine göre seçilen özniteliklerden elde edilen veri kümesindeki spam e-postaları sınıflandırıcılardan Naive Bayes %65, CSCA algoritması %76 ve AIRS2PARALLEL algoritması %77 oranında doğru sınıflandırmıştır.

Çalışmamızda kullanılan algoritmalar arasında CSCA algoritması spam e-postaları diğer YBS algoritmalarından ve Naive Bayes sınıflandırıcısından daha yüksek oranda doğru sınıflandırdığından en iyi sınıflandırma algoritması olarak seçilmiştir.

Tez çalışmasında yapılan deneylerde YBS algoritmalarından negatif seçim algoritması test edilmemiştir. Gelecek çalışmalarda Türkçe spam e-postaların tespiti için negatif seçim algoritması kullanılarak sınıflandırma sonuçlarına etkisi incelenebilir.

KAYNAKLAR

1. **Baysal, B.,** 2001. Temel İmmünoloji Ders Notları, Selçuk Üniversitesi Tıp Fakültesi Mikrobiyoloji Anabilim Dalı, Konya.

2. **De Castro, L.N., Von Zuben F.J.,** 1999. Artificial Immune Systems:Part I – Basic Theory And Applications, Technical Report, Tr – Dca 01/99 December.

3. www.tuik.gov.tr/PreHaberBultenleri.do?id=10880 Hanehalkı Bilişim Teknolojileri Kullanım Araştırması 2012. 14 Ekim 2012

4. **Kaya, Y., Yeşilova, A., Tekin, R.,** 2011, İstenmeyen Elektronik Postaların (SPAM) Filtrelenmesinde Kaba Küme Yaklaşımının Kullanılması, Elektrik-Elektronik-Bilgisayar Sempozyumu 5-7 Ekim 2011, Elazığ.

5. **Z. Chuan, L. Xianliang, H. Mengshu, Z. Xu**, 2005. A LVQ-based neural network anti-SPAM email approach. ACM SIGOPS Operating Systems Review. Volume 39, Issue 1, 34-39

6. **Burim, S., Ohm, S**., 2007. Artificial Immunity-based Feature Extraction for SPAM Detection. In: Eighth ACIS Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, Qingdao, P.R. China, 359–364.

7. **Jiujun, C., Gang X., Fei G., Yuanming Z.,** 2008. SPAM Filtering Method Based on an Artificial Immune System, MultiMedia and Information Technology, MMIT '08. International Conference, 169–171.

8. **Idris, I.,** 2011, Negative Selection Algorithm In Artificial Immune System For SPAM Detection, 2011 5th Malaysian Conference in Software Engineering(MySEC), 379-382.

9. **Zhang, J., Li, X.M., Xu, W., et al.:,** 2011. Filtering algorithm of SPAM short messages based on artificial immune system, In: 2011 International Conference on Electrical and Control Engineering, ICECE 2011 Proceedings, Yichang China, 195–198.
10. http://royal.pingdom.com/2011/01/12/internet-2010-in-numbers/ Internet 2010 in numbers. 30 Kasım 2012.

11. http://royal.pingdom.com/2012/01/17/internet-2011-in-numbers/ Internet 2010 in numbers. 30 Kasım 2012.

12. publib.boulder.ibm.com/infocenter/iseries/v5r4/topic/cl/sndmsg.htm, 11 Kasım 2011.

13. **Şahinaslan, Ö.,** 2007. Bilgisayar Ağlarında Açık Kaynak Kodlu Güvenlik Yazılımları ile Anti-SPAM Modülünün Geliştirilmesi, *Yüksek Lisans Tezi*, Maltepe Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

14. **Uzun, K.,** 2009. SPAM iletilerin internet omurgasına yan etkileri ve engellenmesi, *Yüksek Lisans Tezi*, Beykent Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

15. **Altunyaprak, C.,** 2006. Bayes yöntemini kullanarak istenmeyen elektronik postaların filtrelenmesi, *Yüksek Lisans Tezi*, Muğla Üniversitesi Fen Bilimleri Enstitüsü, Muğla.

16. **Freed, N., Borenstein, N**., 1996. Multipurpos e Internet Mail Extensions, RFC 2045.

17. **Postel, J.B.,** 1982. Simple Mail Transfer Protokol, RFC 821.

18. **Evlice, E.,**2012. SMTP Server Kullanmadan E-posta Gönderme (.NET). http://www.csharpnedir.com/makalegoster.asp?MId=119 , 19 Haziran 2012.

19. http://tr.wikipedia.org/wiki/DNS DNS. 11 Haziran 2012.

20. **Şahinaslan, Ö., Borandağ E., Can, E.,Şahinaslan, E**., 2009. Posta sunucularda SPAM önleme teknikleri, IX. Akademik Bilişim Konferansı, Bil. No:81.

21. http://blog.trendmicro.com/trendlabs-security-intelligence/london-2012-olympics-scams-spotted/ London 2012 Olympics Scams Spotted. 28 Kasım 2012.

22. http://bidb.pau.edu.tr/SPAMmail.aspx SPAM Postalar. 20 Haziran 2012.

23. http://www.spam.org.tr, 21 Haziran 2012.

24. http://www.cauce.org, 22 Haziran 2012.

25. **Alataş, Ş.,** 2006. SMTP Sorununa Yeni Bir Yaklaşım: Sistemin Yeniden İnşası, Akademik Bilişim Konferansı , Denizli.

26. http://www.securelist.com/en/analysis/204792251/Spam_in_Q3_2012#14 Spam in Q3 2012. 30 Kasım 2012.

27. http://www.emailtray.com/blog/infographic-email-spam-phishing-trends-2011-2012/ INFOGRAPHIC: Email Spam and Phishing Trends 2011-2012. 24 Haziran 2012.

28. www.spamlaws.com SPAM and etc. 24 Haziran 2012.

29. http://karaliste.ttnet.net.tr/2.html Türk Telekom A. Ş. Anti-SPAM & Anti-abuse Grubu. 25 Haziran 2012.

30. http://pathmicro.med.sc.edu/ghaffar/innate.htm Immunology - chapter one inna (non-specific) immunity. 19 Kasım 2011.

31. http://pathmicro.med.sc.edu/bowers/immune%20cells.htm Immunology - Chapter Nine Cells Involved In Immune Responses And Antigen Recognition. 19 Kasım 2011.

32. http://pathmicro.med.sc.edu/mayer/antigens2000.htm Immunology - Chapter Three Antigens. 19 Kasım 2011.

33. **Çırakoğlu, B.,** 2003. Bağışıklık Sistemi, Bilim ve Teknik, Mart 2003 eki, İstanbul, Türkiye.

34. **Litman G, Cannon J, Dishaw L.,** 2005. Reconstructing immune phylogeny: new perspectives, Nat Rev Immunol 5 (11): 866-79. PMID 16261174.

35. www.istanbul.edu.tr/itf/itfogrenci/attachments/079_immun.sistem.bagisiklik.sistemi1.pdf immun sistem "Bağışıklık Sistemi". 20 Kasım 2011.

36. **Alberts, B., Johnson, A., Lewis J., Raff, M., Roberts, K., Walter P.,** 2002. Moleculer Biology of The Cell, Garland Science ISBN 0-8153-3218-1.

37. **Smith, A.D.,** 1997. (Ed) Oxford dictionary of biochemistry and molecular biology, Oxford University Press. ISBN 0-19-854768-4.

38. **Rich, RR.,** 2008. The human immune response. Clinical Immunology, Principles and Practice. Editörler: Rich RR, Fleisher TA, Shearer WT, Schroeder HW, Frew AJ, Weyand CM. Mosby Elsevier, ABD, say. 3 – 17.

39. **Abbas, A.K., Lichtman, A.H.,** 2007. Temel immünoloji, immün sistemin işlev ve bozuklukları. Editörler: Y. Camcıoğlu, G. Deniz, İstanbul Medikal Yayıncılık, İstanbul.

40. **Forrest, S., Perelson,A., Allen, L.,** 1994. Self-Nonself Discrimination in a Computer, Proceedings of the IEEE Symposium on Research in Security and Privacy: 202-212.

41. **Yeğin, O.,** 1992. Temel İmmünoloji Ve İmmün Eksiklik Hastalıkları, Akdeniz Üniversitesi Yayınları, Antalya.

42. **Aslan, K.,** 2011. Immunglobulinler (IgG-A-M-D-E). http://www.burtom.com.tr/images/image/makale/IMMUNGLOBULINLER.pdf, 18 Kasım 2011.

43. **Kaban, Z.,** 2008. Yapay bağışıklık sistemlerini kullanarak türkçe metinlerde tür, yazar ve cinsiyet tanıma, *Yüksek Lisans Tezi*, Y.T.Ü. Fen Bilimleri Enstitüsü, İstanbul.

44. **İnce, S.,** Sindirim Kanalı. "Temel Histoloji . Çev. Ed. Aytekin, Y., Solakoğlu S., Barış Kitabevi. İstanbul.1993." (Junqueira LC, Carneiro J, Kelley RO. Basic Histology 7th ed. Appleton and Lange, 1993'den çeviri.)

45. **Kuşluoğlu, Hasan E.,** 2007. http://web.itu.edu.tr/orencik/AgGuvenligi2007Sunumlari/saldiri_bagisikligi.pdf SALDIRI-BAĞIŞIK AĞLAR- Ağ Güvenliği Dersi. 15 Şubat 2012.

46. **De Castro, L.N., Von Zuben, F.J.,** 1999. Artifical Immune Systems:Part I- Basic Theory and Application, Teknik Rapor-RT DCA 01/99,95 sayfa.

47. **Arıcan, E.,** 2011. immunogenetiğe giriş. http://www.istanbul.edu.tr/fen/notlar/1299849984.pdf, 18 Kasım 2011.

48. **İkinci, Ö.,** 2010. Bağışıklık Sistemi: Dost mu, Düşman mı?, Bilimsel Programlar Uzmanı, TÜBİTAK Bilim ve Teknik Dergisi.

49. **Y. Diao, K.M. Passino,** 2000. Immunity-based hybrid learning methods for structure and parameter adjustment, Submitted to IEEE Trans. on Syst., Man, and Cybern.,July.

50. **De Castro,L.N. and Timmis J.,** 2002. A Novel approach to pattern recognition. In L Alonso,J Corchado, and C Fyfe, editors, Artificial Neural Networks in pattern Recognition, University of Paisley: 67-84.

51. **De Castro, L.N. ve Von Zuben, F.J.,** 2000. The Clonal Selection Algorithm with Engineering Applications, Workshop on Artificial Immune Systems and Their Applications, Las Vegas, USA, 36-37.

52. **Hunt, J. ve Cooke, D.E.,** 1996. Learning Using an Artificial Immune System, Journal of Network and Computer Applications , 19, 189-212.

53. **Engin, O., Döyen, A.,** 2004. Artificial Immune Systems and Applications in Industrial Areas, G.U. Journal of Science 17(1): 71-84.

54. **De Castro L.N and Timmis J.,** 2002. An Artificial Immune Network for Multimodal Function Optimisation, Proceedings of IEEE World Congress on Evolutionary Computation. 669- 674.

55. **De Castro, L. N. ve Timmis, J.,** 2002. Artificial Immune systems: A New Computational Intelligence Approach, Springer Verlag.

56. **Agiza., H.N., Hassan, A.E., Salah, A.M**., 2011. An Improved Version of opt-aiNet Algorithm (I-opt-aiNet) for Function Optimization, IJCSNS International Journal of Computer Science and Network Security, VOL.11 No.3.

57. **Sönmez, M.,** 2005. Dendritik Hücre İmmünobiyolojisi, Moleküler hematoloji ve sitogenetik alt komitesi temel moleküler hematoloji kursu.

58. **Greensmith,J., Aickelin, U.,Tedesco, G.,** 2010. Information Fusion for Anomaly Detection with the Dendritic Cell Algorithm, Information Fusion, 11(1):21-34.

59. **Greensmit,J.,** 2007. The Dendritic Cell Algorithm, PhD thesis, Nottingham University.

60. http://www.artificial-immune-systems.org/algorithms.shtml Basic Immune Inspired Algorithms. 25 Aralık 2011.

61. **Brownlee, J.**, 2005. Artificial Immune Recognition System: A Review and Analysis, Teknik Rapor, No 1-02, Swinburne University of Technology.

62. **Çatal, Ç., Diri, B.,** 2007. Yazılım Kalite Sınıflandırma Probleminde Yeni Yaklaşımlar: Yapay Bağışıklık Sistemleri, UYMS-2. Ulusal Yazılım Mühendisliği Konferansı, Ankara.

63. **Watkins, A.,** 2005. Exploiting Immunological Metaphors in the Development of Serial, Parallel and Distributed Learning Algorithms, PhD Thesis, Mississippi State University.

64. **Çatal, Ç., Diri, B**., 2007. Application and Benchmarking of Artificial Immune Systems to Classify Fault-Prone Modules for Software Development Projects, IADIS International Conference on Applied Computing, Salamanca, İspanya, 18-20 Şubat 2007, sf. 347-354.

65. **Brownlee, J.,** 2005. Clonal selection theory & CLONALG. The clonal selection classification algorithm, Technical Report 2-02, Swinburne University of Technology.

66. **Oliveira, L.O.V.B., Mota, R.L.M.; Barone, D.A.C.,** 2012. Clonal Selection Classifier with Data Reduction: Classification as an optimization task, Evolutionary Computation (CEC), 2012 IEEE Congress on,sf. 1-7.

67. **Fisher, R.A.,** 1936. The use of multiple measurements in taxonomic problems, Annals of Eugenics 7, 179–188.

68. **Wang, R.,** 2006. Feature selection strategies for spam e-mail filtering, *Masters thesis*, Concordia University, Canada.

69. **Ataş, M., Yardimci, Y., Temizel, A.,** 2012. A new approach to aflatoxin detection in chili pepper by machine vision,Computers and Electronics in Agriculture, Volume 87, issue (September, 2012), p. 129-141.

ÖZGEÇMİŞ

Cüneyt ÖZDEMİR

Siirt Üniversitesi
Siirt Meslek Yüksekokulu
56100, Siirt
Tel : 484 – 2341212 – 1087
E.posta : cuneytozdemir@hotmail.com

1980	: Siirt'te doğdu.
1995 – 1998	: Mersin Atatürk Lisesini tamamladı.
1998 – 2004	: Fırat Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümünden mezun oldu.
2004 – 2009	: Barla Bilgisayar Müh. Ltd.Şti. Şirketi bünyesinde Mersin ve Ankara ODTÜ Teknokent'te yazılım mühendisi ve proje koordinatörü olarak farklı projelerde çalıştı.
2009 – …	: Siirt Üniversitesi Meslek Yüksekokulunda Öğretim görevlisi kadrosunda çalışmaktadır.