Makine Öğrenme Yöntemleri ve Kelime Kümesi Tekniği ile İstenmeyen E-Posta/ E-Posta Sınıflaması Spam/ Ham E-Mail Classification using Machine Learning Methods based on Bag of Words Technique

Esra ŞAHİN
Ulaşım, Güvenlik, Enerji ve
Otomasyon Sistemleri Sektör Başkanlığı (UGES)
ASELSAN A.Ş., Ankara, Türkiye
esrasahin@aselsan.com.tr

Murat AYDOS
Bilgisayar Mühendisliği
Hacettepe Üniversitesi
Ankara, Türkiye
maydos@hacettepe.edu.tr

Fatih ORHAN
COMODO Group
New Jersey,USA
fatih.orhan@comodo.com

***Özetçe* —Günümüzde elektronik ortamda sıkça kullandığımız iletişim yollarından birisi olan e-postalar; kişisel iletişimler, iş odaklı aktiviteler, pazarlama, reklam gibi birçok nedenden dolayı hayatımızda önemli bir yer kaplamaktadır. Birçok farklı konudaki iletişim ihtiyacı için hayatımızı kolaylaştıran e-postalar, kötü amaçla kullanıldıklarında alıcıların zamanını çalması, maddi ve manevi kayıplara sebep olması nedeniyle hayatı oldukça zorlaştırabilmektedir. Tanımadığımız ya da güvenilmeyen adreslerden, reklam ya da güvenlik tehdidi oluşturma amaçlı gönderilen e-postalar, bilgi güvenliği açısından önemli tehlikeler oluşturabilir. Bu türdeki e-postaları, tespit edip önleyebilmek ise ayrı bir çalışma konusu olmuştur. Bu çalışmada, e-posta içeriğinde yer alan linklerin metinleri ele alınarak, makine öğrenmesi yöntemleri ve Kelime Kümesi Tekniği ile istenmeyen e-posta/e-posta sınıflaması yapılmıştır. Yapılan çalışmada doğruluk metriği kullanılarak Kelime Kümesi Tekniği sonucu oluşturulan farklı N-Gramların sınıflandırma başarısına olan etkisi ve farklı makine öğrenme tekniklerinin istenmeyen e-posta sınıflandırılmasındaki başarısı analiz edilmiştir.**

Anahtar Kelimeler—Spam Filtreleme; Kelime Kümesi Tekniği; Makine Öğrenmesi

***Abstract*—Nowadays, we use frequently e-mails, one of the communication channels, in electronic environment. It play an important role in our lives because of many reasons such as personal communications, business-focused activities, marketing, advertising etc. E-mails make life easier because of meeting many different types of communication needs. On the other hand they can make life difficult when they are used outside of their purposes. Spam emails can be not only annoying receivers, but also dangerous for receiver's information security. Detecting and preventing spam e-mails has been a separate issue. In this study, the texts of the links which is in the e-mail body are handled and classified by the machine learning methods and Bag of Word Technique. We analyzed the effect of different N-Grams on classification performance and the success of different machine learning techniques in classifying spam e-mail by using accuracy metric**

Keywords—Spam Filtering; Bag of Word Technique; Machine Learning

I. GİRİŞ

Elektronik ortamda birçok farklı konudaki iletişim ihtiyacı için hayatımızı kolaylaştıran e-postalar, amacı dışında kullanıldıklarında alıcıların zamanını çalması, maddi ve manevi kayıplara sebep olması nedeniyle hayatı oldukça zorlaştırabilmektedir. Tanımadığımız ya da güvenilmeyen adreslerden, reklam ya da güvenlik tehdidi oluşturma amaçlı gönderilen e-postalara, "istenmeyen e-posta" (Spam) denilmektedir [1]. Özellikle pazarlama ve reklam yöntemi olarak çok sayıda işletme potansiyel müşterilerinin mail adreslerini toplamayı denemektedir. Bu işletmelerin bazıları kötü niyetli davranarak topladığı bu bilgileri diğer ticari işletmelerle paylaşmaktadır [1]. Bunun sonucunda binlerce insan istenmeyen içerikli maillerle karşı karşıya kalmaktadır. Stephen tarafından yapılan durum çalışmasında istenmeyen e-posta ekonomisinin getirdiği kar, gerçek hayattaki şu örnekle anlatılmıştır [2] Geleneksel reklam yaklaşımı ile potansiyel bir müşteriye gönderdiğiniz her broşür, en az 1.00 USD değerindedir. 5.000 broşürü postaladığımızı düşünürsek sadece posta maliyeti 5.000 dolar etmektedir. Toplu e-posta yolu ile reklam yapmak ise durumu oldukça farklı bir hale getirir. Wall Street Journal'ın (11/13/02) yaptığı bir araştırma, e-postaları kullanırken % 0.001 gibi düşük bir getiri oranının karlı olabileceğini göstermektedir. 5 milyon mesajın gönderildiği bu örnekte, ilk hafta %0.0023 gibi bir oranla 81 satışla sonuçlanmış, toplamda 1.500 doları gelir elde edilmiştir. Buradaki 5000 iletiyi gönderme maliyeti çok düşüktür. 56 Kbps modem üzerinde bile saatte binlerce mesaj gönderebilir. Bu durum çalışmasında da görüldüğü gibi istenmeyen e-postalar, düşük maliyetle reklam ve pazarlama ile yüksek kazançlar sağlamaktadır. İstenmeyen e-postalar bilgi güvenliği açısından da önemli tehlikeler oluşturabilir. Bu tehlikeyi insanlara zarar vermeden önce tespit edip önleyebilmek ise ayrı bir çalışma konusu olmuştur. 2017 yılı İnternet güvenliği tehdit raporuna bakıldığında e-postaların yarısından fazlasının (%53) istenmeyen e-posta, bu e-postaların artan bir kısmının da kötücül amaçlı yazılımlar (malware) içerdiği görülmektedir [3]. 2015 yılında 220 e-postadan 1 tanesinin kötücül amaçlı yazılım içerdiği görülürken, 2016 yılında bu oran 131 e posta da 1'e yükselmiştir [3]. E-posta ile bulaşan kötücül amaçlı yazılımlardaki bu artış, bu tür yazılımların büyük

oranda profesyonelleştiğini göstermektedir. Bir diğer yandan ise saldırganların bu tür saldırılardan önemli kazançlar sağladıklarını ve bu tarz istenmeyen e-postaların etkilerinin devam etmesinin muhtemel olduğunu göstermektedir [3] İstenmeyen e-postaların burada bahsedilen tehlikelerinden korunmak için iyi bir filtreleme ihtiyacı doğmuştur. İstenmeyen e-postaları sınıflandırmak için yapılan çalışmalar incelendiğinde farklı farklı teknikler ile e-postaların başlık(header), içerik (body) bölümlerindeki çeşitli bilgiler kullanılarak sınıflandırma yapıldığı görülmüştür. Örneğin Sah ve arkadaşları [4] tarafından yapılan çalışmada sırasıyla veri toplama işlemi, verileri ön işleme, verilerden özellik çıkarma, verileri sınıflandırma ve sonuçları analiz etme işlemi yapılarak metin(body) tabanlı sınıflandırma çalışması yapılmıştır. Çalışmada eğitim için 702 e-posta, test için 260 e-posta kullanılarak Destek Vektör Makineleri (Support Vector Machine (SVM)) ve Naive Bayes algoritmaları ile sınıflandırma yapılmıştır. Çalışma sonucunda Naive Bayes ve SVM algoritmalarının benzer başarı gösterdiği sonucuna varılmıştır. Benzer bir diğer çalışma Renuka ve arkadaşları [5] tarafından Hadoop ortamında yapılan sınıflandırma çalışmasında istenmeyen e-postaları ayırt etmek için Gradient Boost ve Naive Bayes sınıflandırma tekniği kullanılmıştır. Çalışma temelde eğitim ve test olmak üzere iki aşamada ele alınmıştır. Çalışma da sınıflandırma performansını iyileştirmek için tek düğüm Hadoop ortamı kullanılarak farklı sınıflandırma tekniklerinin bir arada kullanıldığı karma bir model çıkarılmıştır. Duyarlılık(precision), doğruluk(accuracy) ve hassasiyet(recall) metrikleri ile çalışmanın performansı ölçülmüştür. Literatürdeki çalışmalar ve buradaki çalışmanın katkısı aşağıdaki gibi özetlenebilir.

- Kötü amaçlı istenmeyen e-posta araştırmalarının çoğunda araştırmacılar, istenmeyen e-postaları ayırt etmek için makine öğrenme algoritmalarının eğitiminde kullanılacak olan içerik temelli yaklaşımlara yönelmişlerdir. Bu yaklaşımlar 4 ana kategoriye ayrılabilir. Bunlar; başlık (header) özellikleri, konu (subject) özellikleri, e-posta metni(body) özellikleri ve e-posta eki(attachment) özellikleridir [4]. Çalışmalar genel olarak içerik metinlerinden özellik çıkarma ve bunları sınıflandırma üzerine odaklanmıştır. Buradaki çalışma bu anlamda literatürdeki çalışmalara benzerlik gösterse de, e-postaların bağlantı metinlerine odaklanması ve bağlantı metinlerine göre sınıflandırma yapılabileceğini göstermiş olması açısından literatüre yeni bir bakış açısı katmıştır.
- Çalışmalarda çok kullanılan makine öğrenmesi teknikleri; SVM, ANN, RF, Naive Bayes ve AdaBoost olmuştur [4]. Bazı araştırmalarda ise bilinen makine öğrenme tekniklerinin yanı sıra Rough setler gibi kural bazlı yaklaşımları da konu alan karma yaklaşımlar sergilenmiştir [4]. Buradaki çalışmada literatürde yer alan yaklaşımlar ve diğer makine öğrenme yöntemleri ele alınarak karşılaştırma yapılmıştır.
- Çalışmalarda genel olarak kullanılan e-posta verileri Spambase(1999), Spam Assassin(2006), TREC(2007)'dir [4]. Bu çalışmada ise literatürdeki çalışmalardan farklı olarak daha güncel ve gerçek veri seti ile sınıflandırma yapılmıştır.

Literatürdeki çalışmalardan farklı olarak bu çalışmanın temel motivasyonu; e-postaların içeriklerinde yer alan bağlantıların metinlerine odaklanmasıdır. Çalışmada veri setlerinin oluşturulması aşamasında Kelime Kümesi Tekniği (BOW), sınıflandırma işlemleri için ise makine öğrenmesi teknikleri kullanılmıştır. BOW sonucu oluşan farklı uzunluktaki N Gram'ların sınıflandırma performansına etkisi ve farklı makine öğrenme tekniklerinin istenmeyen e-posta sınıflandırması için başarısı analiz edilmiştir. Çalışma sonucunda Bayes, SVM algoritmalarının en iyi performansı sergilediği görülmüştür.

TABLO I: Çalışmada kullanılan makine öğrenmesi teknikleri

Karar Ağaçları	Decision Tree [6]
	Gradient Booested Tree
	Decision Stump
	Random Tree
	Random Forest
Bayes	Naive Bayes cite[14]
	Naive Bayes Kernel [7]
Sinir Ağları	Perceptron [8]
Destek Vektör Makineleri	Lib SVM [9]
	Linear SVM [10]
	S-Pegasos [11]
En Yakın Komşu	K-NN [12]

TABLO II: İşlenmemiş Veri Formatı

Benzersiz E-Posta Numarası ||| E-Posta Bağlantısı(URL) ||| Bağlantı Metni
Örnek:
00000460-6b36-41c3-aeaf 35bb32c02766|||http://www.bigbv.top/5D899TU358EM3L...649B3249141020.php|||Click Here

II. METHODLAR

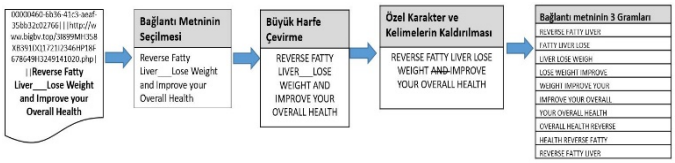
Çalışma kapsamında ele alınan sınıflandırma algoritmaları Tablo I'deki gibidir; Sayfa sayısındaki sınırlama nedeniyle bu algoritmaların ayrıntıları okuyuculara bırakılmıştır.

III. VERİSETİ

İstenmeyen e-postalar ve e-postalarda yer alan bağlantı(link) bilgileri ve bu bağlantılarda yer alan metin bilgileri benzersiz bir e-posta numarası ile birlikte Tablo II'deki formatta sunulmuştur.

İşlenmemiş veriler üzerinde şu işlemler yapılmıştır:

- Her bir satır verinin benzersiz e-posta numarası, bağlantı ve bağlantı metni olarak parçalanmıştır.
- Bağlantı metinlerinde yer alan rakamların ve tanımlanamayan karakterlerin silinmiştir.
- İki karakter olan kelimeler performansı artırmak amacıyla silinmiştir.
- Bazı özel anlam içeren kelimelerin ayırt edici niteliği olmadığı için çıkarılmıştır. Bunlar, Türkçe ve İngilizce haftanın günleri, günlerin kısaltması, Türkçe ve İngilizce yılın ayları, ayların kısaltması, para birimi kısaltmaları, "WWW, HTTP, HTTPS, COM, AND, WITH, THE, İLE, VIA, ANY, YOU, THEY, ARE vs." gibi kelimelerdir. Bu kelimelere veri seti incelenerek karar verilmiştir.
- Türkçe ve İngilizce alfabelerinin farklılığından dolayı oluşan hataları en aza indirmek için tüm harfler büyük harfe çevrilmiştir. Türkçe de yer alan noktalı harfler,



Şekil 1: Veri Ön İşleme ve Gramların Oluşturulması.

TABLO III: Veri Setinin Oluşturulması

Bağlantı Metni	Etiket	Your Overall Health	Fatty Liver Rem.	Liver Rem. Click	Rem. Click Here	Reverse Fatty Liver
REVERSE FATTY LIVER LOSE WEIGHT AND IMPROVE YOUR OVERALL HEALTH	İst. EP.	1	0	0	0	2

noktasız hale dönüştürülmüştür. Bu durum aynı kelime olup birinde noktalı, diğerinde noktasız yazılmış karakterler nedeniyle kelimelerin farklı algılanmasının önüne geçmektedir.

Ön işlemden geçirilmiş olan verilerin, makine öğrenmesi yöntemlerinde kullanılması için eğitim ve test işlemlerinde kullanılacak matris yapıdaki veri seti haline getirilmesi gerekmektedir. Bu amaçla Kelime Kümesi Tekniği (Bag of Word (BOW)) [13] kullanılarak verilerin 1 Gram, 2 Gram, 3 Gram, 4 Gram ve 5 Gram olmak üzere özellikleri(feature) çıkarılmıştır.

Oluşturulan bu gramlar veri setindeki sütunları oluşturmaktadır. İstenmeyen e-postalar ve e-postalar için ayrı ayrı özellik çıkarım işlemi yapılmıştır. Ayrı ayrı oluşturulan bu özellik kümeleri birleştirilerek ortak olan Gram'lardan tek bir tanesi özellik olarak sayılmıştır. Oluşturulan her özellik, diğer bir deyişle her bir gram özellik vektörünün bir elemanını oluşturmaktadır. Şekil 1'de veri ön işleme ve gramların oluşturulması bir örnekle gösterilmiştir.

Gramlar oluşturulduktan sonra eğitim ve test işlemlerinin performansını artırmak amacıyla özellik vektörünün boyutu 30, 40 ve 50 limiti belirlenerek azaltılmıştır. Buradaki limitler bir Gram'ın ham veri seti içerisindeki tekrar sayısını göstermektedir. Tekrar sayısı 30'un altında olan gramlar özellik vektörüne dahil edilmemiştir. Özellik kümesinin boyutunun bu şekilde farklı limitlere göre azaltılması aynı zamanda özellik sayısının makine öğrenme teknikleri başarısına olan etkisini gösterecektir.

Özellik vektörü oluşturulduktan sonra veri setinin oluşturulması aşaması Tablo III'deki örnekteki gibidir. Burada bağlantı metninin her bir gramın özellik vektörü ile karşılaştırılmakta ve tekrar sayısı sayılmaktadır.

IV. DENEYLER VE SONUÇLARI

Şekil 2'de istenmeyen e-posta tespiti için önerilen prosedüre ait iş akışı sunulmuştur. Burada veri ön işleme aşamasının ardından N-Gramlar oluşturulmuş, özellik seçimi yapılmış ve oluşan özellik vektörü sınıflandırma algoritmalarına girdi olarak sunulmuştur. Toplamda 15 adet veri seti elde edilmiştir. Veri setlerinin boyutları Tablo IV'deki gibidir. Veri setlerinin boyutları [Bağlantı metni sayısı]X[Özellik Vektörü Boyutu] formatında verilmiştir. Veri setleri oluşturulduktan sonra RapidMiner Studio aracı ile Tablo I'de belirtilen makine öğrenmesi yöntemleri ile deneyler yapılmıştır. Tüm deneyler 10 katlamalı çapraz doğrulama (cross validation 10 folds) ile gerçekleştirilmiştir.

TABLO IV: Veri Setlerinin Boyutları

Limit	1 Gram	2 Gram	3 Gram	4 Gram	5 Gram
30	50000x3335	50000x4332	44864x4286	33818x3902	28802x3487
40	50000x2658	50000x3239	43045x3131	32613x2825	25673x2525
50	50000x2188	50000x2550	17324x2448	32368x2192	23232x1941

Deneyler yapılırken öncelikle 50 limiti ile sınırlandırılmış 3 gramlık veri seti ile eğitim yapılmıştır. Bu veri setinin seçilmesinin nedeni özellik sayının diğer veri setlerine göre daha az olması ve gram olarak ortalama bir değerde olmasıdır. Bu eğitimin sonunda başarısı %90'ının üzerinde olan algoritmalar için N gramlara göre performans değişimi incelenmiştir. %90 başarının altında kalan algoritmalarla ise sadece tek bir deney gerçekleştirilmiştir.

A. Deney sonuçları

Her bir makine öğrenmesi algoritması için doğruluk metriği ile performans ölçümü yapılmıştır. Tablo V ve Tablo VI'deki deney sonuçlarına göre şu sonuçlar elde edilmiştir:

- Naive Bayes Kernel, Linear SVM ve Pegasos SVM %99.8 başarı oranı ile en istenmeyen e-postaların sınıflandırılmasında en başarılı algoritmalar olmuşlardır. İstenmeyen e-postaların sınıflandırılması için genel bir bakış açısıyla algoritmalara bakıldığından Bayes Algoritmaları, SVM'ler, K-NN ve Perceptron algoritmalarının oldukça iyi performans gösterdiği görülmektedir. Öte yandan bir diğer sınıflandırma algoritması olan Karar Ağaçları'nın bu konuda yeterince iyi olmadığı görülmüştür. Bu nedenle Karar Ağaçları için deneyler sadece 50 limit 3 Gram veri seti ile yapılmıştır. N gramların performansa etkisi incelenmemiştir.
- Deney sonuçları genel olarak değerlendirildiğinde 30 limit ile hazırlanan veri setlerinin performansının oldukça düşük olduğu görülmektedir. Bu durum özellik vektörü boyutunun fazla uzun olmasının performansı kötü etkilediğini göstermektedir.
- Gram sayılarının performansa etkisi incelendiğinde neredeyse tüm algoritmalarda gram sayısının artırılmasının doğruluk oranına iyi yönde katkı yaptığı görülmüştür.
- Naive Bayesian sınıflandırıcıları, tüm yöntemler arasında en az eğitim süresine ulaşmıştır. Öte yandan karar ağaçları ve en yakın komşu algoritması oldukça uzun sürelerde sonuçlanmıştır.

V. SONUÇ

Bu çalışma kapsamında farklı N-Gramların performansa etkisi, özellik vektörü boyutunun performansa etkisi ve farklı makine öğrenmesi methodlarının istenmeyen epostaları sınıflamadaki performansı incelenmiştir. Çalışma öncelikle 1,2,3

TABLO V: Deney Sonuçları

Limit		NB	NBK	LibS	LinS	Peg	Perc	K-NN
	1 Gram	78.03	89.31	85.89	91.04	90.8	86.54	90.2
	2 Gram	86.16	95.48	94.45	95.91	95.67	92.95	93.64
30	3 Gram	95.84	98.6	89.24	98.6	98.66	95.6	96.75
	4 Gram	97.91	99.38	81.16	99.41	99.38	98.34	98.7
	5 Gram	98.8	99.89	99.76	99.89	99.88	99.63	99.72
	1 Gram	77.97	89.19	85.97	90.97	90.55	85.38	90.27
	2 Gram	85.69	95.5	94.47	95.88	95.76	93.18	93.62
40	3 Gram	95.67	98.73	90.17	98.2	98.76	96.83	96.79
	4 Gram	97.71	99.39	92	99.38	99.37	98.01	98.71
	5 Gram	98.67	99.89	90.89	99.89	99.89	99.7	99.7
	1 Gram	78.37	89.16	86.62	90.78	90.29	86.4	90.16
	2 Gram	85.54	95.59	94.63	95.93	95.85	93.26	93.22
50	3 Gram	95.67	98.79	91.24	98.81	98.8	97.59	96.74
	4 Gram	97.66	99.37	91.81	99.37	99.36	97.94	98.73
	5 Gram	98.53	99.88	91.94	99.88	99.88	99.66	99.66

TABLO VI: Deney Sonuçları (Başarı oranı %90 altında olan algoritmalar)

Decision Tree	Boosted Trees	Decision Stump	Random Tree	Rnd. Forest
86.49	79.19	63.26	59.10	59.41

Gramlık Özellik Vektörleri ve Support Vector Machines, SVM-Pegasos ve Naive Bayes [14] algoritmaları ile yapılmış, daha sonra ise farklı makine öğrenme tekniklerinin karşılaştırılması ve 4 ve 5 Gram ile hazırlanan özellik vektörlerinin performansa etkisinin incelenmesi amacı ile geniş kapsamda ele alınmıştır.

VI. BİLGİLENDİRME

Bu çalışma ASELSAN A.Ş. tarafından desteklenmiş olup, çalışma boyunca sağladığı imkânlardan ve katkılarından dolayı teşekkür ederiz. Ayrıca veri setindeki desteklerinden dolayı COMODO Group Türkiye'ye teşekkür ederiz.

KAYNAKLAR

[1] B. Richardson, "Aggreting emaıl." US Patent 20,170,279,756, Tech. Rep., 2017.

[2] S. Cobb, "The economics of spam," *ePrivacy Group, http://www. spamhelp. org/articles/economics of spam. pdf*, 2003.

[3] "Internet security threat report." Tech. Rep., 2017.

[4] U. K. Sah and N. Parmar, "An approach for malicious spam detection in email with comparison of different classifiers," 2017.

[5] D. K. Renuka and P. V. S. Rajamohana, "An ensembled classifier for email spam classification in hadoop environment," *Appl. Math*, vol. 11, no. 4, pp. 1123–1128, 2017.

[6] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE transactions on systems, man, and cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[7] P. Breheny, "Kernel density classification," *STA*, vol. 621.

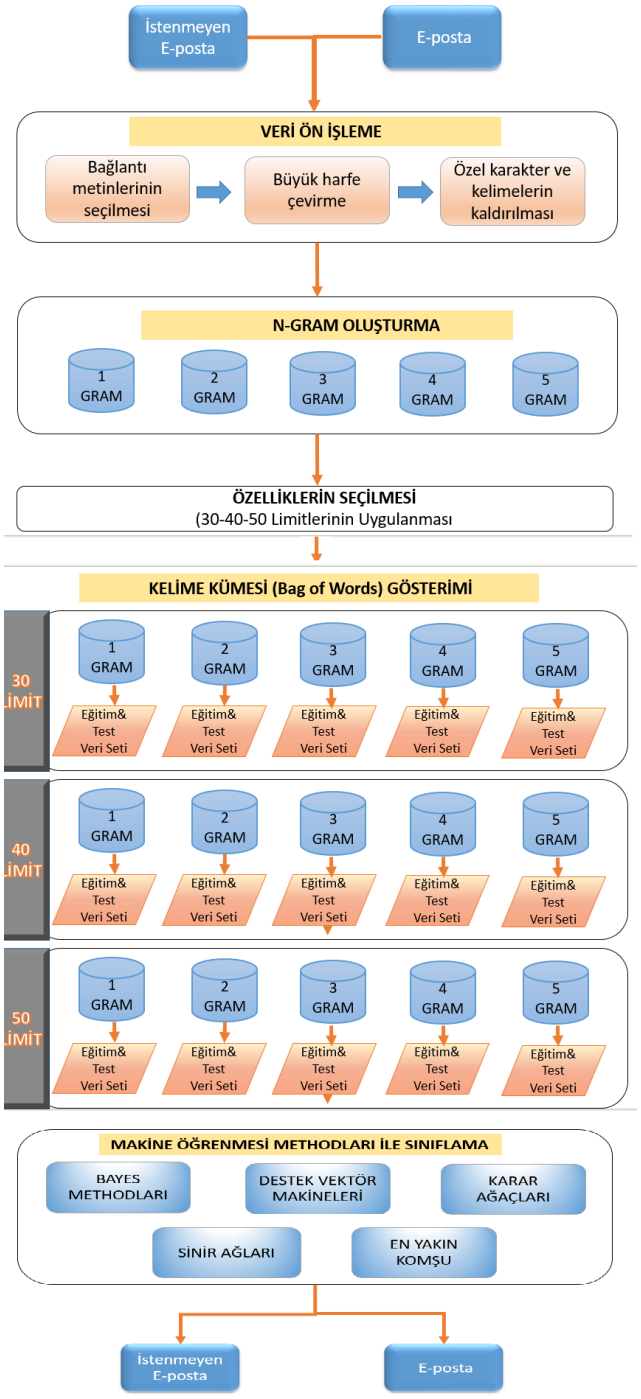
[8] . Elmas, *Yapay zeka uygulamaları:(yapay sinir ağı, bulanık mantık, genetik algoritma)*. Seçkin Yayıncılık, 2011.

[9] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[10] S. AYHAN and Ş. ERDOĞMUŞ, "Destek vektör makineleriyle sınıflandırma problemlerinin çözümü için çekirdek fonksiyonu seçimi," *Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi*, vol. 9, no. 1, 2014.

[11] S. Shwartz, Y. Singer, and N. Pegasos, "Primal estimated subgradient solver for svm," *Mathematical Programming*, vol. 27, no. 1, pp. 807–814, 2011.

[12] E. Taşcı and A. Onan, "K-en yakın komşu algoritması parametrelerinin sınıflandırma performansı üzerine etkisinin incelenmesi," *Akademik Bilişim*, 2016.



Şekil 2: İstenmeyen E-posta Sınıflandırma Yöntemi.

[13] C. W. Kim, "Ntmaldetect: A machine learning approach to malware detection using native api system calls," *arXiv preprint arXiv:1802.05412*, 2018.

[14] A. S. Bozkir, E. Sahin, M. Aydos, and F. Orhan, "Spam e-mail classification by utilizing n-gram features of hyperlink texts."