SPAM İÇERİKLİ E-POSTALARIN TESPİTİ İÇİN BİR METİN MADENCİLİĞİ UYGULAMASI: TERİMLERİN GAMA İLİŞKİ KATSAYISINA DAYALI POLARİZASYONU

***

A TEXT MINING APPLICATION ON THE DETERMINATION OF SPAM-CONTENTED E-MAIL: POLARIZATION OF TERMS BASED ON THE GAMA RELATIONSHIP COEFFICIENT

Ahmet YÜCEL[*]

Meltem KESKİN KÖYLÜ[**]

Öz

Teknolojinin gelişimi, iletişimin düzey ve şeklini de değiştirmiştir. İki nokta arası kapalı devre iletişim (telefon, mektup, telgraf, vb.) modellerinin yerini daha çok, tek noktadan tüm dünyaya açılan (Facebook, Twitter, Instagram, vb.) iletişim modelleri almıştır. Bu durum iletişimin sınırlarını kişisel olarak belirlememizi imkânsız hale getirirken, gizlemesi mümkün olmayan (E-mail, Whatsapp numarası, vb.) birçok kişisel iletişim yolunu da dünyaya açık hale getirmektedir. Basit bir e-mail yoluyla, bilgisayarda kayıtlı özel verilerin istenmeyen kişilerin eline geçmesi gibi, mevcut durum birçok risk taşımaktadır. Buna engel olmak amacıyla birçok virüs yazılımı geliştirilmekte ve elektronik ortamda karşılaşılan riskli unsurların tespitinde yardımcı olmaktadır. Ancak bazı riskli unsurlar virüs formatından uzak, normal bir metin olarak karşımıza çıkmaktadır. Bu tarz durumlarda ilgili metnin içerik olarak incelenip, riskli olup olmadığına karar vermek gerekmektedir. Bu çalışmada, istenen ve istenmeyen içeriğe sahip e-postaların bir metin madenciliği algoritması ile tespit edilip sınıflandırılması işlemi yapılmaktadır. Bu amaçla, gama ilişki katsayısına dayalı kompozit bir polarite değişkeni oluşturulmuş ve bu değişken üzerine genelleştirilmiş lineer modeller kurulmuştur. Modellerin sınıflandırma başarısı ortalama ise yaklaşık % 81,2'dir.

Anahtar Kelimeler: *Metin Madenciliği, Veri Madenciliği, Genelleştirilmiş Lineer Model, Polarite, Gama İlişki Katsayısı, Sınıflandırma, İletişim, İstenmeyen İçerik.*

Abstract

The development of technology has also changed the level and form of communication. Two-ended closed-circuit communication (telephone, letter, telegraph, etc.) models have been replaced by communication models that are originated from a single point and opens to the world (Facebook, Twitter, Instagram, etc.). While this makes it impossible for us to determine the limits of communication personally, it also makes a lot of personal communication paths that cannot be hidden (E-mail, Whatsapp number, etc.). The current situation carries many risks, such as by a simple e-mail, that private data stored on the computer gets into the hands of undesirable people. In order to prevent this, many virus software is being developed and it helps to detect the risky elements encountered in electronic environment. However, some risky elements appear as a normal text rather than a virus format. In such cases it is necessary to examine the relevant text as content and decide whether it is risky or not. In this study, e-mails with spam and ham content are determined and classified by a text mining algorithm. For this purpose, a composite polarity variable based on the gamma relationship coefficient was created and generalized linear models were established on this variable. The average classification success of the models is approximately 81.2%.

Keywords: *Text Mining, Data Mining, Generalized Linear Model, Polarity, Gamma Relationship Coefficient, Classification, Communication, Spam Content.*

[*] Dr. Öğr. Üyesi Ankara Yıldırım Beyazıt Üniversitesi, Şereflikoçhisar Uygulamalı Bilimler Fakültesi, Bankacılık ve Finans Bölümü, ayucel@ybu.edu.tr.
ORCID: 0000-0002-2364-9449

[**] Dr. Öğr. Üyesi Ankara Yıldırım Beyazıt Üniversitesi, Şereflikoçhisar Uygulamalı Bilimler Fakültesi, Uluslararası Ticaret ve Lojistik Bölümü, mkeskinkoylu@ybu.edu.tr
ORCID: 0000-0002-8536-4940

Geliş Tarihi / Received: 29.11.2018
Kabul Tarihi / Accepted: 26.12.2018

1. GİRİŞ

En sık kullanılan iletişim aracının ilk sırasında yer alan e-postalar, kişisel iletişimden, ticarete, eğitimden sağlığa hayatın her alanında kullanılmaktadır. Elektronik iletişim araçları kullanıcıları, internet bağlantısı sayesinde mekân ve zamandan bağımsız olarak iletişim kanalları kullanılabilmektedir. Bu durum kullanıcılarına pek çok avantaj sağlarken dezavantajlarla da karşı karşıya kalmasına neden olabilmektedir.

Haberleşmede pek çok sosyal iletişim araçları kullanmakla birlikte yaygın olarak kullanılan bir diğer kişisel iletişim aracı e-postalardır. Bu postalar hızlı, maliyetsiz ve pek çok kişiye aynı anda uluşabilmektedir. Bu nedenledir ki dünyada en yaygın kullanılan iletişim araçlarından biridir. İnternet kullanıcısının sahip olduğu e-posta adresleri, kişiye özel görülmesine rağmen dünyaya açık ve kolayca ulaşılabilir bilgilerdir. Yaygın olarak kullanılan e-posta ağları, her türlü saldırıya açık olabilmektedir. Kötü amaçlı kişiler, kötü yazılımlı içerikleri kullanma yoluyla bilgisayar kullanıcısının güvenliğini tehdit edebilmektedir. Bu kötü yazılımlara "spam" denmektedir. Bilgisayar kullanıcıları için temel problemlerden birini spam oluşturmaktadır. Kapsamları ve nitelikleri sürekli gelişen spam'ları belirlemek için yöntemler geliştirilmektedir. Ancak geliştirilen yöntemlerde zaman zaman yetersiz kalabilmektedir. Kimi kötücül yazılımlar virüs içerikleri taşımadığı için tehlikesiz gibi algılanabilmektedirler. Bu aşamada normal metin içeriği barındıran yazılımın içeriğinin araştırılması önemli olmaktadır. Bu nedenle kullanıcıların e-posta adreslerine düşen postaların istenmeyen içerikler barındırıp barındırmadığı da postanın bizzat kendisinin istenmeyen mi yoksa istenen e-posta olduğunun tespiti önemli olmaktadır.

Bu çalışmada, metin madenciliği algoritması aracılığı ile önceden öngörülemeyen bilgiye ulaşmanın ve elde edilen sonuçları karar aşamasında kullanmanın çok aşamalı süreç yönetilmiştir. Çalışmada, toplamı oluşturan5574 adet e-mail verileri "Public Domain Dedication" lisansına sahip olup paylaşım platformu olan Kaggle.com'dan sağlanmıştır. Elde edilen veriler incelendikten sonra 740'ı spam ve 740'ı spam olmayan e-mail olmak üzere 1480 e-mail değerlendirilmiştir. Gama ilişki katsayısı ile değişken oluşturulmuş ve elde edilen değişkenler üzerinde genelleştirilmiş lineer model kullanılmıştır. Gama parametresine dayalı polar değişkenin modellerde sağladığı başarı oranı en düşük %79 ve en yüksek %84 arasında oluşmaktadır. Çalışmada kurgulanan modelin sınıflandırma başarısı ortalama %81,2 ile en düşük ve en yüksek değerin arasında yer almayı başarmıştır.

2. LİTERATÜR VE İLGİLİ ÇALIŞMALAR

Sosyal ağları kullanarak haberleşme dünyanın her yerinde yaygın olarak kullanılan yöntemdir. İlk elektronik haberleşme sistemi olarak görülen Elektronik Enformasyon Takas Sistemi New Jersey Teknoloji Enstitüsünde grubun kendi arasında e-posta gönderebilmesi ile 1978 yılında başlamıştır (Hambrick, 2012). Günümüzde ise tüm bilgisayar kullanıcılarının e-posta hizmeti ücretsiz olarak hizmet sağlayıcılarından alarak kullanmaktadırlar. İşletmeler pazarlama stratejileri dâhilinde müşterilerinin e-posta adreslerini sahip olmak için gerekli çabaları göstermektedirler. Böylece elde edilen e-posta adresleri diğer pek çok işletme ile de paylaşılması mümkün olabilmektedir. Sonuç olarak e-posta kullanıcısı istenmeyen e-postalara maruz kalabilmektedir. Bu postaların içerilerinde güvenliği tehdit edebilecek içeriklerde barındırma olasılıkları yüksektir. İşte böyle postalara "istenmeyen posta" (spam, junk e-mail) denilmektedir (Richardson, (2017)'den aktaran, Şahin, 2018: 1).Cobb (2003) ise istenmeyen postaların ekonomiye getirdiği yükleri spam ekonomisi adlı çalışmasında örnekleri ile açıklamıştır. Renukavd (2017)tarihli çalışmalarında istenmeyen postaların sınıflama algoritmaları üzerinde durmuşlar ve 1990'lü yıllardan başlayarak istenmeyen postaların hızla yayıldığını ve kullanıcılara büyük zararları olduğunu vurgulamışlardır.

Spam belirlemek için Integro, Pajek, ReDites, SybilRank gibi pek çok harici virüs yazılım yöntemleri bulunmaktadır. Ayrıca spam virüs tespit etmek amacı ile yapılmış çalışmalara literatürde rastlanmaktadır. Bunlardan biri Çıtlak, 2018 yılında yaptığı çalışmadır. Çalışmada, spam hesap tespit modelinin veri kümesinde ve canlı sosyal ağlar üzerindeki sorgulanmıştır. Sosyal ağlara

yönelik öğrenmeye dayalı bir spam hesap tespit modeli ile spam hesapların Twitter üzerinde tespit edilmesine yönelik çalışma yapılmış. Spam hesaplamaları çalışmada, Link analizi, makine öğrenmesi ve metin analizi yöntemleri kullanılmıştır (Çıtlak, 2018). Zararlı yazılımları erken uyarı yöntemlerinden biri olan "Honeypot" ağı ile uyarı sistemi 2004 yılında Dagon vd. leri tarafından araştırılmıştır (Dagon vd.,2004). Bu çalışmadan sonrada pek çok farklı yöntemle spam belirlemede sistemleri geliştirilmiştir.

Metin Madenciliği, verilerin açık parçası olmayan büyük çaplı verilerin içinden özel unsurları açığa çıkararak gelecekle ilgili tahminleri bilgisayar programları aracılığı ile açığa çıkarma sürecidir (Kalikov, 2006). Sarıkoz, yaptığı çalışmasında Veri Madenciliğinin temellerinin 1950'li yıllardaki yapay zekâ ve makine öğrenimi çalışmalarına dayandığını ama asıl pratikte kullanımının 1990'lı yıllara dayandığını vurgulamıştır (Sarıkoz, 2010. 4-5).Yirmi birinci yüzyıla gelindiğinde sanayiden finansa, sağlıktan eğitime kadar geniş çalışma alanı bulan Veri Madenciliği hem yurt içi hem de yurt dışı pek çok çalışmada kullanılmıştır. Verin madenciliği bir dizi süreçten oluşur. Bu süreçler; problemi belirleme, verileri hazırlama, madencilik algoritmalarını uygulama, modeli kullanma ve sonuçları izleme değerlendirme aşamalarını kapsar (Shearer, 2000). Veri madenciliği, bankacılık, pazarlama, sigortacılık ve sağlık gibi sektörler başta olmak üzere pek çok alanda kullanılmaktadır (Akman, 2010). Sarıkoz, spam filtrelemek için veri madenciliğinin kullanılması ile ilgili çalışmasında, kullanıcılara ve e-posta hizmeti servis sağlayıcılarına spam mesajlardan korunmak için görevler düştüğüne çalışmasında yer vermiştir. Yapay sinir ağları ve kümeleme metotları ile veri setleri üzerinde deneme yaparak 'spam' ve 'spam olmayan' e-postaların belirlemiştir (Sarıkoz, 2010).

3. YÖNTEM VE UYGULAMA

3.1. Gama (γ) İlişki katsayısı

Gama katsayısı, her iki değişken de sıralama (*ordinal*) düzeyinde ölçüldüğünde, derlenen istatistiksel verileri tanımlamak için kullanılır. Katsayının hesaplanması için derlenen veriler ilk önce bir çapraz tablo formatında gösterilmesi gerekir. Katsayı hesabı aşağıdaki formüle göre yapılır.

$$\gamma = \frac{U-T}{U+T} \tag{1}$$

burada U değeri, çapraz tablodaki her gözlemsel frekansın, frekansın kendisinin sağ alt kısmındaki gözlemlerdeki frekansların toplamı ile çarpılmasıyla elde edilir. Benzer şekilde T değeri, çapraz tablodaki her bir gözlemsel frekansın, frekansın kendisinin sol alt kısmındaki gözlemlerdeki frekansların toplamı ile çarpılmasıyla elde edilir.

Gama parametresi -1 ile 1 arasında değişir. -1 ve +1 her biri mükemmel ilişkileri gösterir ve 0 hiçbir ilişki olmadığı anlamına gelir. Parametre değeri 0'dan uzaklaştıkça, ilişkinin gücü artar. Parametre -1 ile 0 arasındaysa, ilgili değişkenler zıt yönlü bir ilişkiye sahiptir ve benzer şekilde parametre 0 ile +1 arasındaysa, değişkenler aynı yönlü bir ilişkiye sahiptir denir. Katsayıya dayanan hipotezi test etmek için klasik z testi uygulanır.

$$z = \frac{\gamma - E(\gamma)}{\sqrt{\frac{n(1-\gamma^2)}{U+T}}} \sim N(0,1) \tag{2}$$

Burada $E(\gamma) = 0$(başlangıç hipotezine bağlı olarak) beklenen değerdir ve parametrenin varyansı $V(\gamma) = \frac{n(1-\gamma^2)}{U+T}$ 'dır. Ayrıca belirlenen anlamlılık düzeyi$\alpha = 0.05$ ve istatistiksel olarak anlamlı ilişkilerin ilgili tablolarda kırmızı ile vurgulandığı görülmektedir (Ünver vd., 2016: 288-289).

3.2. Polarite ve Matematiksel İfadesi

Polarite, kelime anlamı olarak zıt kutuplu olma halidir. Veri madenciliği alanında ise polarite, veri içinde yer alan iki farklı eğilim olarak ifade edilebilir. Bu çalışmada polaritenin birinci eğilimini 'spam', ikinci eğilimini ise 'not spam' olarak isimlendiriyoruz. Polaritenin matematiksel hesabı kontrol grubuna dayalı olarak yapılmaktadır ve elde edilen matematiksel değerler test grubu üzerinde uygulanmaktadır.

Buna göre, kontrol grubunda her bir terim ile bağımlı değişken arsındaki ilişki düzeyi gama katsayısı olarak hesaplanır. Burada, m satır (e-mail/doküman) sayısı ve n sütun (terim/değişken) sayısı olmak üzere,

$$G_{n\times 1} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_n \end{bmatrix} \tag{3}$$

öyle ki, $g_i i.(i = 1,2,\dots,n)$terimin bağımlı değişkenle olan gama katsayısıdır.

$$A_{m\times n} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \tag{4}$$

ilgili metin veriden elde edilmiş terim-doküman ikili (*binary*) frekans dağılım tablosu olsun. Buna göre,

$$A_{m\times n} \cdot G_{n\times 1} = \begin{bmatrix} a_{11}\gamma_1 + a_{12}\gamma_2 + \cdots + a_{1n}\gamma_n \\ a_{21}\gamma_1 + a_{22}\gamma_2 + \cdots + a_{2n}\gamma_n \\ \vdots \\ a_{m1}\gamma_1 + a_{m2}\gamma_2 + \cdots + a_{mn}\gamma_n \end{bmatrix}_{m\times 1} \tag{5}$$

$A_{m\times n} \cdot G_{n\times 1}$matris çarpımı $m \times 1$ boyutunda yeni bir matris oluşturuyor öyle ki j.satır,j.dokümanın polariteye dayalı toplam gama değerini vermektedir. Bu durumda,

$$P_j = \frac{1}{w_j}\left(\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}g_i\right) \tag{6}$$

öyle ki $w_j j.(j = 1,2,\dots,m)$dokümanda seçilen terim sayısıdır.

3.3. Genelleştirilmiş Lineer Model (GLM)

Y bağımlı değişken ve $X_1, X_2, \dots, X_n$ rasgele değişkenler olmak üzere, $P(Y \mid X_1, X_2, \dots, X_n)$ şartlı olasılık dağılımına bağlı genelleştirilmiş lineer model, şu şekilde ifade edilebilir:

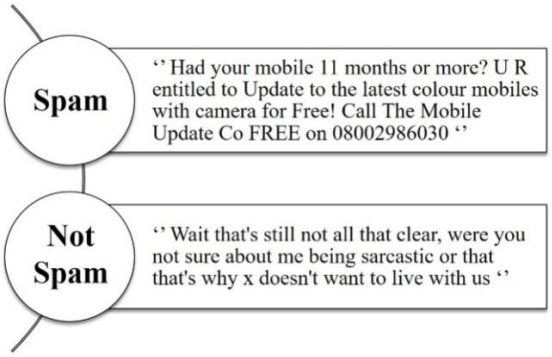
1- X_i'lerin Y üzerindeki etkileri g fonksiyonu ile ifade edilir öylekigX_i'lerden oluşan lineer bir kombinasyondur.
2- Model bir link fonksiyonu üzerine kurulur.
3- $l: g \to E(Y) = \mu$ tersinir bir fonksiyon olmak üzere,

$$g = \beta_0 + \sum_{i=1}^{n}\beta_i X_i \tag{7}$$

ifadesi bir genelleştirilmiş lineer modeldir öyle ki $g = l(\mu)$ link fonksiyonudur (Levy, 2012: 107-108).

3.4. Veri

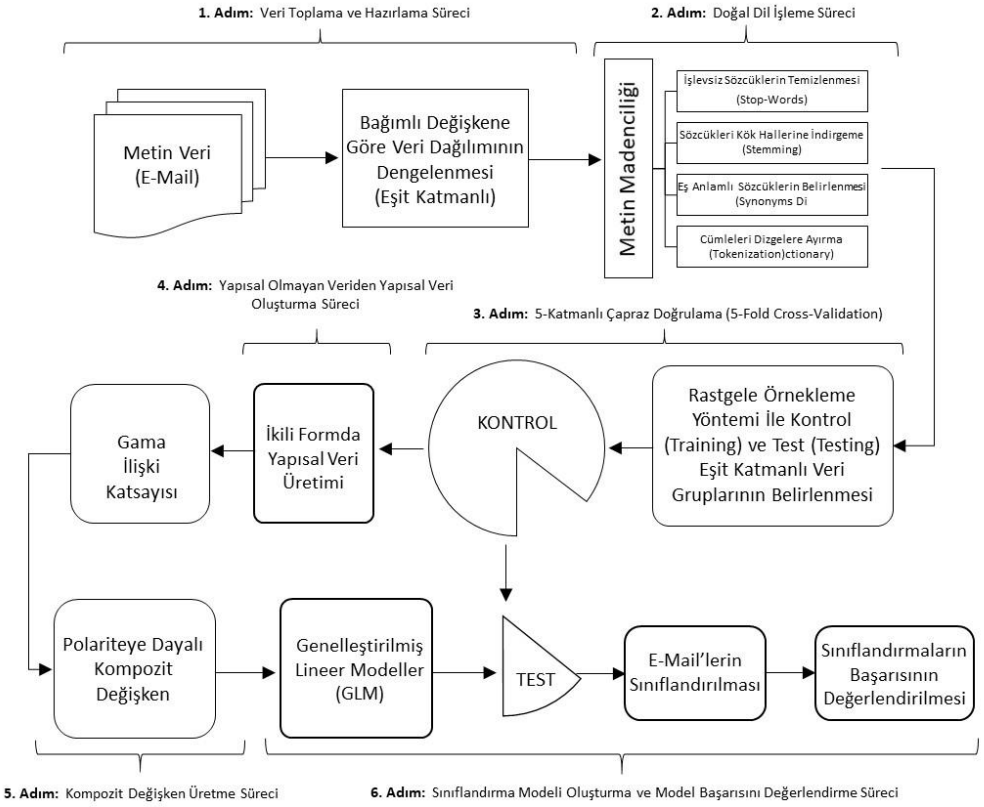
Çalışma için gerekli olan veri, dünya çapında bir veri paylaşım platformu olan Kaggle.com'dan temin edilmiştir. Veri *CC0 1.0 Universal (CC0 1.0) "Public Domain Dedication"* lisansına sahiptir. Verinin orijinal adı '*Spam_Massages*' (İstenmeyen Mesajlar) olup, orijinal halinde %13 (747 adet) '*spam*' ve %87 (4827 adet) '*not spam (ham)*' olmak üzere, toplam 5574 e-mail dokümanı bulunmaktadır. Veri temizleme ve dengeleme işleminden sonra, 1480 e-mail (740 'spam' ve 740 'not spam') incelemeye alınmıştır. Veri içinde yer alan e-maillerden ikisine örnek olarak Şekil 1'de paylaşılmıştır.



Şekil 1: Veride yer alan, istenen (not spam)ve istenmeyen (spam) e-postalara örnekler

3.5. Yöntem

Çalışmanın, veri toplamadan model oluşturmaya kadar yöntem ve uygulamalara dair genel adımlar Şekil 2'da ifade edilmiştir.



Şekil 2: Yöntem ve uygulama adımları

İfade edilen yöntem genel olarak 6 adımdan oluşmaktadır. Birinci adımda, veri depolandığı kaynakdan, web tarayıcı (web-crawling) bir algoritma yazılımı ile doğrudan veya bir veri tabanından hazır paket olarak elde edilir. Daha sonra, veri içinde yer alan, çalışmayla alakasız

kısımlar (link veya dosya uzantıları, veri kaynağına ait özel isimler, vb.) veriden temizlenir ve oluşturulacak olan modelin başarısının doğru bir şekilde değerlendirilebilmesi için verinin kategorik dağılımı bağımlı değişkenin kategorilerine göre (her kategoriye eşit sayıda doküman olacak biçimde) dengelenir. İkinci adımda, metin madenciliğinin en önemli adımı olan doğal dil işleme süreci uygulanır. Bu amaçla, veriye her hangi bir şekilde anlamsal katkı sağlamayan işlevsiz terimler (stop-words) bir liste halinde programa (StatSoft Statistica 12) yükleniyor ve metinlerden çıkarılır. Bunun yanı sıra eş anlamlı (synonyms) terimlerde listelenerek programa yüklenir. Bu noktada, online olarak hizmet veren 'Cambridge English Dictionary' (https://dictionary.cambridge.org/dictionary/essential-american-english) online sözlük aracından (Amerikan İngilizcesi kategorinde) faydalanılmıştır. Ayrıca Statistica 12 İngilizce için kelimeleri köklerine indirgeyen (stemming) ve cümleleri dizgelerine ayıran (tokenization) algoritmaları saptanmıştır. Bu algoritmalar yardımıyla metinler içinde yer alan kelime ve kelime grupları yapısal olarak ifade edilecek forma hazır hale getirilmiştir. Üçüncü adımda, veri kontrol ve test gruplarına ayrılmıştır. Burada veri rastgele ve dengeli olarak beş parçaya (%20) ayrılmıştır (5-katmanlı çapraz-doğrulama (5-fold cross-validation) öyle ki sırayla her bir parça test ve geriye kalan dört parça ise kontrol grubu olarak işlev görmektedir. Dördüncü adımda mevcut kontrol veriden ikili (binary) formda terim-doküman frekans matrisi elde edilir. Elde edilen matristen beşinci adımda, bağımlı değişken ve terimler arsında yer alan gama ilişki düzeyleri, ilgili kısımda verilen yöntem ile hesaplanır ve daha sonra, ilgili kısımda ifade edildiği gibi polariteye dayalı kompozit değişken oluşturulmuştur. Altıncı ve son adımda, elde edilen kompozit değişkenle genelleştirilmiş lineer bir model kurularak test veri üzerine uygulanmış ve test veri üzerine uygulanan sınıflama modelinin doğruluk oranı hesaplanmıştır. Modelin doğruluk oranı, doğru sınıflama sayısının toplam birim sayısına oranıyla elde edilmiştir. Modellerin tanımsal kodlaması Tablo 1'de ve Tahmin Modeli İşaretleme Dili GLM algoritması Tablo 2'de paylaşılmıştır.

Tablo 1: GLM tanımsal kodlama

```
Distribution : BINOMIAL
Link function: LOGIT
Response variable: "SPAM_Category (0:NO,1:YES)"

Codes of dependentvariable
0: Primary code   (1)
1: Secondary code (0)
Design Effects:
Continuous effects:  "Gamma_Polarite1"
Categorical effects:

Model specification:
GLZ;
  RESPONSE = "SPAM_Category(0:NO,1:YES)" (0 1);
  GROUPS = none;
  COVARIATE = "Gamma_Polarite1";
  DESIGN = "Gamma_Polarite1";
  INTERCEPT = include;
  PARAM = sigma;
  SDELTA = 7;
  SURFACE = none;
  MIXTURE = none;
  SAMPLE = "Testing_Set1" ( 102.);
  COUNTV = none;
  MBUILD = all;
  CONVERGE = 7;
  MAXITER = 100;
  INITIALS = none;
  OFFSET = none;
  OUTPUT = none;
```

Tablo 2: StatSoftStatistica GLM algoritması (StatSoftStatistic PMML (Predictive Model Markup Language)

```
<?xmlversion="1.0" encoding="windows-1254" ?>
<PMML version="2.0">
<Header copyright="STATISTICA Data Miner, Copyright (c) StatSoft, Inc., www.statsoft.com."/>
<Data Dictionary number Of Fields="2">
        <DataField name="SPAM_Category(0:NO,1:YES)" optype="categorical">
                <Value value="0" Numeric Value="0"/>
                <Value value="1" Numeric Value="1"/>
        </DataField>
        <DataField name="Gamma_Polarite1" optype="continuous"/>
</Data Dictionary>
<GeneralizedLinearModel
        Function Name="classification"
        Model Name="Generalized linearregression"
        Model Type="general Linear"
        Target VariableName="SPAM_Category(0:NO,1:YES)">
<Extension name="Distribution" value="binomial"/>
<Extension name="LinkFunction" value="logit"/>
<Parameter List>
        <Parameter name="p1" label="Intercept"/>
        <Parameter name="p2" label="Gamma_Polarite1"/>
</Parameter List>
<Factor List>
</Factor List>
<Covariate List>
        <Predictor name="Gamma_Polarite1"/>
</Covariate List>
<PPMatrix>
        <PPCellvalue="1" predictor Name="Gamma_Polarite1" parameter Name="p2"/>
</PPMatrix>
<Extension name="Correct Dummy Code" value="1"/>
<Extension name="Incorrect Dummy Code" value="-1"/>
<ParamMatrix>
<PCelltarget Category="0" parameter Name="p1" beta="2.39698409648957e+000"/>
<PCelltarget Category="0" parameter Name="p2" beta="-6.95747291468540e+000"/>
</Param Matrix>
        </Generalized Linear Model>
         </PMML>
```

4. BULGULAR

Veride toplam 1480 doküman (e-mail) bulunduğu için, her bir test veri (%20) için toplam 296 doküman, rastgele ve doküman kategorilerine göre (‘spam’/’not spam’(0: hayır, 1: evet)) eşit (148 adet) şekilde dağıtılmıştır. Doğrusal modeller (GLM) binomial dağılıma sahip bir veri üzerine uygulanmıştır ve logit olasılık fonksiyonu kullanılmıştır. Her bir katman için Tablo 3’de GLM’lere ait ‘Olabilirlik Oranı Tip 1Test’ ve Tablo 4’de, ‘Parametre Tahmin’ sonuçları paylaşılmaktadır. Sonuçlar modellerin tamamının $\alpha = 0.05$ anlamlılık düzeyinde, istatistiksel olarak anlamlı sonuç verdiğini göstermektedir.

Tablo 5’de modellerin sınıflamada doğru ve yanlış tahmin sayıları ve modellerin doğruluk oranları verilmektedir. Ek olarak, Şekil 3‘de modellerin doğruluk oranları sütun grafiğiyle gösterilmiştir. Tabloda yer alan ‘Eksik’ sütunu kontrol veriden test veriye geçişteki bilgi kaybının sayısal bir ifadesidir. Yöntem kısmında da ifade edildiği üzere, modeller oluşturulmadan önce, veri %80 kontrol ve %20 test grubu olarak iki parçaya ayrılmıştır. Daha sonra kontrol grubunda yer alan metinlerden oluşturulan sözlük ve terimlere ait polar oran ile test grubunda yer alan metinler yapısal hale getirilmiştir. Bu süreçte, kontrol grubu içinde yer alan ancak test grupta bulunmayan veya tam tersi, kontrol grupta bulunmayıp test grupta bulunan bilgiler polariteye dayalı kompozit değişken hesabına dâhil edilmemiştir. Bu durum, test grubunda yer alan bazı dokümanların mevcut bilgiyle anlamlandırılmasını imkânsız kılıyor. Modellerin doğruluk oranları, ‘eksik’ tahminlerin de yanlış kabul edilerek hesaplanmıştır.

Gama parametresine dayalı polar değişkenin modellerde sağladığı başarı oranı en düşük %79 ve en yüksek %84 arasında değişmektedir. Sonuçlar, modellerin tamamıyla otomatik süreçte üretilen değişkenlere dayandığı dikkate alındığında, oldukça başarılı bir noktadadır.

Tablo 3: Olabilirlik Oranı Tip 1Test **(**LikelihoodType 1 Test)

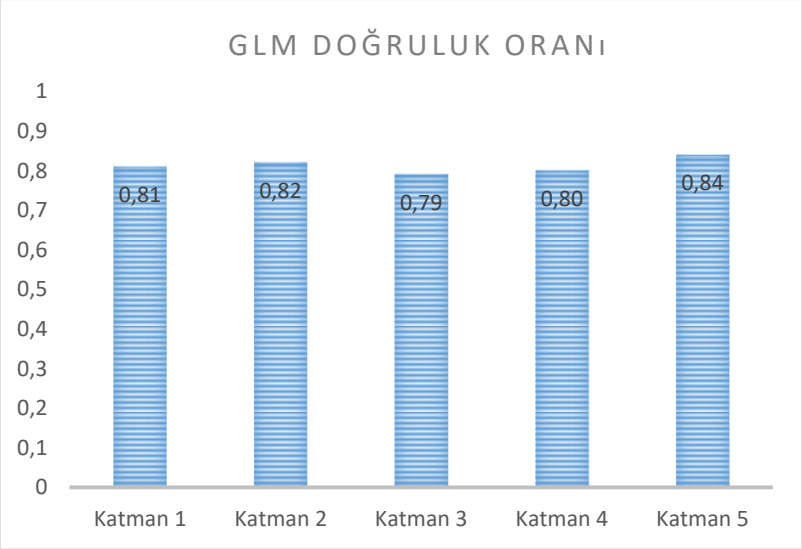
Modeller		Log-Likelihood	Chi-Square	P-Value
Katman 1	**Intercept1**	-742,752		
	Gamma_Polarite1	-269,536	946,431	0,0001
Katman 2	**Intercept2**	-748,114		
	Gamma_Polarite2	-284,523	927,181	0,0001
Katman 3	**Intercept3**	-750,898		
	Gamma_Polarite3	-281,020	939,756	0,0001
Katman 4	**Intercept4**	-746,238		
	Gamma_Polarite4	-282,198	928,079	0,0001
Katman 5	**Intercept5**	-745,010		
	Gamma_Polarite5	-296,196	897,628	0,0001

Tablo 4: Parametre Tahminleri

Modeller		Estimate	Standard Error	Wald Stat.	P-Value
Katman 1	**Intercept1**	2,396	0,202	140,143	0,0001
	Gamma_Polarite1	-6,957	0,406	292,306	0,0001
Katman 2	**Intercept2**	2,300	0,196	136,858	0,0001
	Gamma_Polarite2	-6,532	0,379	296,147	0,0001
Katman 3	**Intercept3**	2,387	0,201	140,465	0,0001
	Gamma_Polarite3	-6,698	0,391	293,474	0,0001
Katman 4	**Intercept4**	2,249	0,191	138,252	0,0001
	Gamma_Polarite4	-6,625	0,382	299,658	0,0001
Katman 5	**Intercept5**	2,308	0,196	138,261	0,0001
	Gamma_Polarite5	-6,417	0,371	297,752	0,0001

Tablo 5: Her bir katman için, GLM tahminlerine göre doğru veya yanlış sınıflandırma sayıları ve modellerin doğruluk oranları

Doğrusal Modeller		Tahmin 0	Tahmin 1	Eksik	GLM Doğruluk Oranı
Katman 1	**Gözlem 0**	105	26	17	0,81
	Gözlem 1	7	137	4	
Katman 2	**Gözlem 0**	109	12	27	0,82
	Gözlem 1	14	134	0	
Katman 3	**Gözlem 0**	104	15	29	0,79
	Gözlem 1	15	131	2	
Katman 4	**Gözlem 0**	104	21	23	0,80
	Gözlem 1	12	134	2	
Katman 5	**Gözlem 0**	110	15	23	0,84
	Gözlem 1	7	141	0	



Şekil 3: Modellerin doğruluk oranları (sütun grafiği)

5. SONUÇ

E-ticaretin gelişmesi ile birlikte hızlı ve maliyetsiz haberleşme ihtiyacı ortaya çıkmıştır. Elektronik posta tamda bu aşamada devreye girmiş. Sadece ticaret de değil hayatın her alanında kullanılan vazgeçilmez unsurlardan biri olmuştur. Ancak ister reklam amaçlı ister kurum ve kuruluşlardan bilinen içerikli gelen e-postaların içeriklerinde yer alan metin içeriklerinde istenmeyen kötü yazılımlar bulunabilmektedir. Zaman zaman ise spam içerik barındırmayan e-postaların spam kutusuna düşmesi. Ya da spam niteliği taşıyan e-postaların bazılarının gelen postalar kutusu içine yerleşebilmektedir. Google istenmeyen içerikli sitelerin, arama sonuçlarında sansürlenmesi için, Outlook/Hotmail/Yahoo gibi e-mail servis sağlayıcılar ise istenmeyen içerikli e-postaların tespiti için bu tarz metin madenciliği algoritmaları kullanmaktadırlar. Buna benzer dünya çapında birçok virüs yazılım firması da bu alanda çalışmalarını sürdürmektedir.

E-postaların içerisinde yer alan istenmeyen postaların belirlenmesi, spam olmayan postaların ise spam kutusuna geldiği anda tespit etmek önemli bir unsurdur. Çalışmada, hazırlanan algoritma ile bu sorunu çözmek amaçlanmıştır. Çalışmada, gama ilişki katsayısı kullanılarak ve tamamıyla otomatik (sezgisel müdahaleden uzak) üretilen değişkenlere dayalı olarak kurgulanan genelleştirilmiş lineer model ile sınıflandırma başarısı ortalama %81,2'ye ulaşılmıştır. Bu algoritma, veri boyutundan bağımsız olarak yapısal olmayan veride metin sınıflandırması yapmaya imkân sağlaması açısından önemli bir katkıdır. Böylece de e-postaların içeriklerinde bulunan bağlantı belgeleri ile spam e-postalarının ayıklanması sağlanmıştır.

KAYNAKÇA

Akman,M., (2010).“Veri Madenciliğine Genel Bakış ve Random Forests Yönteminin İncelenmesi: Sağlık Alanında Bir Uygulama”, Yüksek Lisans Tezi, Ankara Üniversitesi, ,Sağlık Bil. Enstitüsü, Ankara.

Cambridge English Dictionary, (2018). https://dictionary.cambridge.org/dictionary/essential-american-english adresinden 27.11.2018 tarihinde erişilmiştir.

Cobb, S., (2003). “The Economics of Spam”.ePrivacy Group, https://www.cobbsblog.com/spam/economics_of_spam.pdf adresinden 27.11.2018 tarihinde erişilmiştir.

Çıtlak, O. (2018). “Sosyal Ağlara Yönelik Öğrenmeye Dayalı Bir Spam Hesap Tespit Modeli ve Uygulaması”Gazi Üniversitesi Fen Bilimleri Enstitüsü. Yayınlanmış Yüksek Lisans Tezi.

Dagon, D.,Qin, X., Gu, G., Lee, W., Grizzard, J., Levine, J. andOwen, H. (2004).Honeystat: Local worm detection using honey pots. In International Workshop on Recent Advances in Intrusion Detection, Springer, Berlin, Heidelberg, 39-58.ttps://link.springer.com/chapter/10.1007/978-3-540-30143-1_3 adresinden 27.11.2018 tarihinde erişilmiştir.

Hambrick, M. E. (2012). Six degrees of information: Using social network analysist of explore the spread of information within sport social networks. International Journal of Sport Communication, 5(1), 16-34.

Kalıkov, A., (2006), "Veri Madenciliği ve Bir E-Ticaret Uygulaması", Yayınlanmış Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü. Ankara.

Levy, R. (2012). "Probabilistic Models in the Study of Language" , ch. 6, November 6. http:// idiom.ucsd.edu/~rlevy/pmsl_textbook/book_draft.pdf adresinden 27.11.2018tarihinde erişilmiştir.

Renuka, D. K.,Visalakshi P.and. Rajamohana, S (2017). "An Ensembled Classifier for Email Spam Classification in Hadoop Environment". Appl.Math, 2017. 11(4): p. 1123-1128.http://www.naturalspublishing.com/files/published/7ttcu333pd8l38.pdf adresinden 27. 11.2018 tarihinde erişilmiştir.

Richardson, B.B., (2017)."Aggregating Email".US Patent.

Sarıkoz, K. (2010). "Veri Madenciliği Yöntemleri İle Spam Filtreleme". Gazi Üniversitesi Bileşim Enstitüsü. Yayınlanmış Yüksek Lisans Tezi.

Shearer, C., (2000), "The Crisp-DM Model: The New Blue print for Data Mining "Journal of Data Warehousing, Cilt 5 No 4, 13-23.

Şahin, E., (2018). "Makine Öğrenme Yöntemleri ve Kelime Kümesi Tekniği İle İstenmeyen E-Posta / E-Posta Sınıflaması". Hacettepe Üniversitesi Yayınlanmış Yüksek Lisans Tezi. Ankara.

Ünver, Ö.,Gamgam, H. Ve Altunkaynak, B., (2016). SPSS Uygulamalı Temel İstatistik Yöntemler Olasılık – Hipotez Testleri – Regresyon Analizi. Seçkin Yayıncılık. Ankara.