Uluslararası Sosyal Araştırmalar Dergisi / The Journal of International Social Research
Cilt: 11 Sayı: 55 Şubat 2018 Volume: 11 Issue: 55 February 2018
www.sosyalarastirmalar.com Issn: 1307-9581
http://dx.doi.org/10.17719/jisr.20185537249

KARAR AĞACI ALGORİTMASI İLE METİN SINIFLANDIRMA: MÜŞTERİ YORUMLARI ÖRNEĞİ

TEXT CLASSIFICATION VIA DECISION TREES ALGORITHM: CUSTOMER COMMENTS CASE

Çiğdem AYTEKİN•
Cem Sefa SÜTCÜ**
Umut ÖZFİDAN***

Öz

Günümüzde mevcut olan verinin büyük çoğunluğunun metin tabanlı olması, onların analizi için birtakım yöntemlerin geliştirilmesini zorunlu hale getirmiştir. Zira bu metinlerin manuel olarak incelenmesi çok zordur, hatta çoğu durumda imkânsızdır. Metin verilerden bilgi çıkarımının zorunlu hale gelmesi otomatik olarak bilgi çıkarımına yönelik çalışmaları tetiklemiş ve metin sınıflandırma yöntemleri ortaya çıkmıştır. Ancak metin veriler yapısal olmadığından analizleri de geleneksel makine öğrenmesi uygulamalarından farklı olmaktadır.

Bu çalışmada bir işletme veri tabanında yer alan müşteri yorumlarından örneklem seçilerek, onları şikâyet-talep-teşekkür sınıflarına atayacak bir karar ağacı modeli oluşturulmuştur. Algoritma, entropi ve bilgi kazanımı hesaplama yöntemlerini esas almaktadır. Bu yolla önce müşteri yorumlarından onları temsil edebilecek nitelikteki öznitelikler –kelimeler– çıkarılmış ve düğümler belirlenerek ilgili sınıf etiketleri tespit edilmiştir.

Anahtar Kelimeler: Metin Sınıflandırma, Karar Ağacı Algoritması, Müşteri Yorumları, Yapısal Olmayan Veri, Entropi.

Abstract

Most of the data available today are text based. This necessitates developing some methods for their analysis. Because inspecting these text is very difficult, even impossible most of the time. Necessity of extracting knowledge from text data has triggered works about automatically extracting knowledge out of text data and text classifications methods have emerged. But since text data are not structural, their analysis are different than traditional machine learning applications.

In this study, by selecting a sample from customer comments in a firm's database, a decision tree model is constructed which can assign these comments into complaint-request-acknowledgement classes. Algorithm is based on entropy and knowledge gain calculations. This way, first attributes -words- that can represent customer comments have extracted and by defining nodes class labels.

Keywords: Text Classification, Decision Trees Algorithm, Customer Comments, Unstructured Data, Entropy.

1. Giriş

Günümüzde pek çok alanda kullanılan metin formatının gittikçe büyüyen hacme sahip yığınlar oluşturması ve bu metin yığınlarını farklı amaçlara yönelik olarak otomatik bir biçimde kategorize etme ihtiyacı, metin sınıflandırmanın temelini oluşturmuştur. Metin sınıflandırmada amaç, en basit şekliyle metin yığınlarının oluşturulacak bir program aracılığı ile sınıflara ayrılmasıdır.

Metin sınıflandırma çalışmaları en temelde bilgi yönetimi çerçevesinde değerlendirilebilir. Bu tür sistemlerde veri-enformasyon-bilgi-bilgelik piramidinin (Fricke, 2009:131-142; Ahsan ve Shah, 2006:1-7) mantığını açıklamak yerinde olacaktır: Fricke'ye göre, bir kişi dünyadaki ülkelerle ilgili pek çok ansiklopedik bilgiye sahip olabilir. Fakat bu bilgi onu "akıllı" yapmaz. Bu geniş bilginin kişinin nasıl hareket edeceği ile ilgili karmaşık, etik ve pratik problemlere çözüm bulması gerekir (Fricke, 2009:145). Dolayısıyla verinin enformasyona dönüşmesi, enformasyonun bilgiye dönüşmesi, bilginin de aklı oluşturması sürecinde yaşanan döngüsellikte, kişinin doğru kararlar verip hedeflerini gerçekleştirebilmesi için veriden bilgiyi elde edecek araç ve yöntemlere ihtiyaç vardır. Bu döngüsellik piramidin ters dönmesi anlamına da gelir (Ahsan ve Shah, 2006:6-7).

• Doç. Dr., Marmara Üniversitesi İletişim Fakültesi Bilişim Anabilim Dalı
** Prof. Dr., Marmara Üniversitesi İletişim Fakültesi Bilişim Anabilim Dalı
*** Uygulama Mimarı-Danışman, Fore Teknoloji

Veriden bilgi çıkarma süreci, veriler yapısal bir formatta olduğunda veri madenciliği teknikleri ile gerçekleşir. Hant ve diğerlerine göre (2001:2), veri madenciliği büyük veri kümelerinden veya veri tabanlarından faydalı bilginin çıkarılmasına ilişkin bir bilimdir. Bu alan istatistik, makine öğrenmesi, veri yönetimi ve veri tabanları, desen tespit etme, yapay zekâ vb. alanların kesişiminde yer alır. Bütün bu alanlar veri analizinin bir boyutu ile ilgilidir ve pek çok ortak noktası vardır. Veri madenciliği analizi genellikle, çok büyük miktardaki veri setleri için ilişkilerin aranması ve anlaşılır bir şekilde özetlenmesi esasına dayanır. Bu ilişkiler modeller veya desenler olarak tanımlanır. Bu tür veriler genellikle veri madenciliği analizi dışında başka amaçlarla toplanır. Örneğin, bir bankanın müşteri hesapları ile ilgili tuttuğu kayıtlar veri madenciliği amacıyla tutulmaz, ama bu veriler üzerinde veri madenciliği yapılarak müşterilerin harcama alışkanlıkları ve ihtiyaçları tespit edilebilir.

Veri madenciliği teknikleri, son dönemlere kadar işletmelerin sahip olduğu *yapısal* verilerin analiz edilerek karar süreçlerinde değerlendirilmesi amacıyla kullanılagelmiştir. Bununla birlikte, 2000'li yılların ortalarından itibaren sosyal medya platformları ortaya çıkmış ve müşteriler bu mecralarda işletmelerle ve diğer kullanıcılarla etkileşim halinde kalarak *yapısal olmayan* ama anlamlı metin veriler üretmişlerdir. Böylelikle işletmeler reklamcılık, müşteri ilişkileri yönetimi, potansiyel müşteri analizi ve kurumsal itibar yönetimi gibi alanlar için bu mecralara yönelmiş ve takibe başlamışlardır (Sütcü ve Aytekin, 2013:7-15).

Bu çalışmada, bir işletme veri tabanında yer alan gerçek müşteri yorumlarının (yapısal olmayan) otomatik olarak nasıl sınıflandırılacağı üzerine odaklanılmıştır. Sınıflar şikâyet-talep-teşekkür sınıfları olarak belirlenmiştir. Böylelikle işletme, veri tabanında örneğin, yüzde kaç oranında şikâyet ya da talep ya da teşekkür yorumu barındırdığını tespit edebilir ve karar süreçlerinde bu bilgileri kullanarak rekabet avantajı sağlayabilir. Diğer yandan, işletme bir sonraki aşamada örneğin, şikâyet sınıfına ait yorumların "hangi nedenden kaynaklandığı" sorusuna cevap aramak isteyebilir. Bu kez metin sınıflandırma değil, metin kümeleme tekniklerinden söz etmek gerekir. Aşağıdaki kısımda metin sınıflandırmaya ilişkin literatür ve metin kümelemeden farklılıklar ele alınmıştır.

2. Metin Sınıflandırma

Metin sınıflandırma bir metin madenciliği görevidir. Metin Madenciliği, metin formatında bulunan verilerin yapılandırılmış hale getirilerek madencilik teknikleri ile analiz edilmesi ve değerli bilginin elde edilmesi süreci olarak tanımlanabilir. Soucy ve Mineau da metin sınıflandırmayı, "*önceden belirlenmiş kategorilere göre doğal dil metinlerinin sınıflandırılması*" olarak tanımlamışlardır (Soucy ve Mineau, 2001:647-648).

Metin sınıflandırma, diğer bir metin madenciliği görevi olan metin kümelemeden farklıdır. Metin sınıflandırmada metin veriler önceden belirlenmiş etiket olarak da adlandırılan sınıflara atanırken, metin kümelemede sınıflar etiketsizdir ve burada metin veriler kendi sınıflarını otomatik olarak oluşturur.

Amasyalı ve diğerlerine göre, metin sınıflandırmada amaç bir metnin özelliklerine bakılarak önceden belirlenmiş belli sayıdaki kategorilerden hangisine dahil olacağını belirlemektir. Metin sınıflandırma bilgi getirme, bilgi çıkarma, doküman indeksleme, doküman filtreleme, otomatik olarak meta data elde etme ve web sayfalarını hiyerarşik olarak düzenleme gibi pek çok alanda önemli bir rol oynamaktadır. Metin sınıflandırma sistemlerinin ilk örnekleri 70'li yıllarda karşımıza otomatik doküman indeksleme olarak çıkmıştır. Belirli bir konu için özel sözlükler oluşturulmuş ve bu sözlük içerisindeki kelimeler birer kategori gibi algılanarak dokümanlar sınıflanmıştır (Amasyalı ve diğerleri, 2006).

Metin sınıflandırma ile ilgili literatür incelendiğinde çalışmaların 19 yıl öncesine dayandığı görülebilir. Bayes algoritması ile metin sınıflandırmada vaka modellerinin karşılaştırılması (Mccallum ve Nigam, 1998:41-48), Metin sınıflandırma için sözcüklerin dağılımsal kümelemesi (Baker ve Mccallumzy, 1998:96-103), Metin sınıflandırmada maksimum entropi kullanımı (Nigam ve diğerleri, 1999:61-67), Destek vektör makineleri kullanılarak metin sınıflandırma için transdüktif çıkarım (Joachims, 1999:200-209), Beklenen maksimizasyon yoluyla etiketli ve etiketsiz belgelerden metin sınıflandırma (Nigam ve diğerleri, 2000:103-104), Metin sınıflandırma için uygulamalarla destek vektör makinesi aktif öğrenmesi (Tong ve Koller, 2001:45-66), Metin sınıflandırma için yüksek performanslı özellik seçimi (Rogati ve Yang, 2002:659-661), Metin sınıflandırma için özellik seçim metriklerinin kapsamlı ampirik bir çalışması (Forman, 2003:1289-1305), Bilgi filtrelemeyi iyileştirmek için Twitter'da kısa metin sınıflandırması (Sriram ve diğerleri, 2010:841-842) konulu çalışmalar metin sınıflandırma alanında yapılan kronolojik sıradaki bazı çalışmalardandır.

Doğal dil işleme ise (natural language processing), metin madenciliğinin çalışma alanlarından birisidir. Bu aşama tüm metin madenciliği aşamalarında kullanılmasa bile genelde özellik çıkarımı ve metinden bazı anlamsal bilgilerin elde edilmesinde sıklıkla başvurulan aşamadır. Örneğin, konuşma parçalarının etiketlenmesi veya cümlebilimsel parçalama veya diğer dilbilimsel işlemler doğal dil işleme

aşamasında yapılır (Şeker, 2015:32). Bu alan yapay zekâ, biçimsel diller kuramı, kuramsal dilbilim ve bilgisayar destekli dilbilim, bilişsel psikoloji gibi çok değişik alanlarda geliştirilmiş kuram, yöntem ve teknolojileri bir araya getirir (Çakıroğlu ve Özyurt, 2006:7-9)

Yazım denetimi, doğal dil işleme alanının metin sınıflandırmadaki en önemli konularından birisidir. Ancak Oflazer çalışmasında, yazım düzeltmesi amacıyla diğer diller için geliştirilen ve sonlu bir sözcük dağarcığı kabulüne dayanan tekniklerin Türkçe için uygun olmadığını ifade etmektedir (Oflazer, 2012). Çünkü Türkçe bitişken bir dil yapısına sahiptir ve yapım ekleri, bileşik kelimeler gibi birçok nedenle analizi diğer dillerden farklıdır.

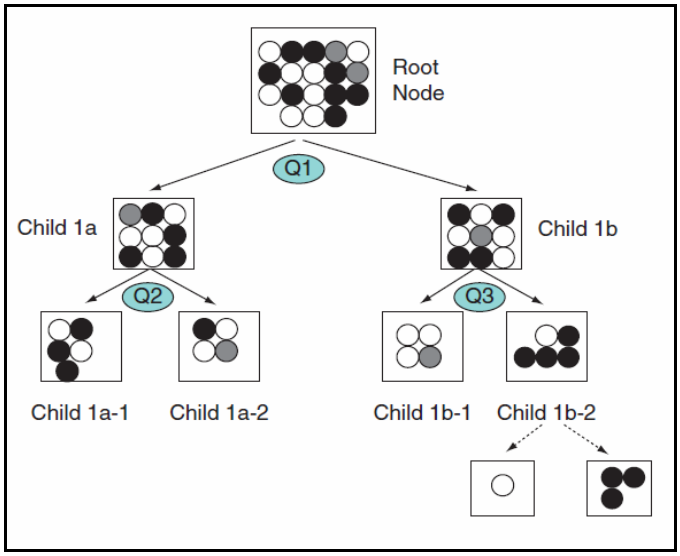
Metin sınıflandırma Medikal alanda sıkça sorulan sorularda (sss), İş alanında müşteri desteği için sıkça sorulan sorularda, Eğitim alanında bir konu üzerine araştırma ve sıkça sorulan sorularda geniş bir kullanım alanı bulmaktadır (Fan ve diğerleri, 2006). Metin verilerin büyük bir kısmının işletme veri tabanlarında saklandığı düşünüldüğünde; örneğin, müşterilere ait şikâyet ve memnuniyet içerikli metinlerden elde edilen bilgiler ürün geliştirme, hata izleme, garanti süresi gibi konularda işletmeye girdi oluşturacaktır (Delen ve Crossland, 2008).

Diğer yandan, metin sınıflandırma uygulamaları belli alanlarda daha gelişmiş olmakla beraber, büyük miktarda metin verinin bulunduğu her ortamda (eğer bilgi çıkarımı önemli bir katma değere sahipse) sınıflandırmaya ihtiyaç duyulduğu söylenebilir. Özellikle sosyal medya olarak adlandığımız ortamlar böylesine ihtiyaçlar için önemli bir zemin oluşturmaktadır.

Metin sınıflandırmada kullanılan bazı algoritmalar Destek Vektör Makinesi (Support Vector Machine), Naive Bayes (Naive Bayes), Karar Ağaçları (Decision Trees) ve K-En Yakın Komşu (KNN, K-nearest neighborhood) algoritmalarıdır. Her bir algoritmanın sınıflandırmanın farklı özelliklerine göre farklı performanslar gösterdiği söylenebilir. Bu çalışmada metin sınıflandırma için müşteri yorumları üzerinden bir araştırma "karar ağaçları algoritması" ile gerçekleştirilmiştir.

3. Metin Sınıflandırmada Karar Ağaçları Algoritması

Bir karar ağacı, kök düğüm adı verilen bir değişkenden başlayan ağaç benzeri bir yapıda hiyerarşik bir ilişki grubudur. Bu kök düğüm, kök düğümün ayrı sınıflarını veya düğümün ölçeği boyunca belirli aralıkları temsil eden çok sayıda dalda iki bölüme ayrılır. Her bölüntüde, bölünen değişkenin sınıfları veya aralığı bakımından yanıtı olan bir soru sorulmaktadır. Bu soru örneğin, "erkek mi kadın mı?" olabilir. Bunun gibi sorular, ikiye bölünmüş karar ağacı oluşturmak için kullanılır. Karar ağaçları birden çok bölme ile de oluşturulabilir. Her bir bölünmede sorulan sorular, sonuçta ortaya çıkan vakaların bölünmelerde ne kadar üniform olması gerektiğini yansıtan bazı belirsizlik ölçüleri açısından tanımlanır. Her dal, diğer değişkenlerin sınıfları veya aralıkları kullanılarak daha da bölünür. Her bölüntüde bölünen düğüme ana düğüm, bölünmüş olduğu düğümlere de alt düğüm adı verilir. Bu işlem, kesme kuralı gerçekleşinceye kadar devam eder (Nisbet ve diğerleri, 2009:241). Şekil 1, renkli topları sınıflandırmak için bir karar ağacı örneğini göstermektedir.



Şekil 1: Basit Bir İkili Karar Ağacı Yapısı

Kaynak: Nisbet R., Elder J., Miner G. (2009). "Handbook of Statistical Analysis and Data Mining Applications". Burlington: Elsevier.

Karar ağaçları evreni temsil edeceği düşünülen belli bir sayıdaki örnek üzerine kurulur ve elde edilen sonuçlar evren olarak adlandırılan örnekler üzerinde test edilerek kontrol otomasyon yöntemiyle en uygun düğümlere ulaşmaya çalışılır. Hant ve diğerlerine göre, bu tamamen sezgisel bir işlemdir ve model oluşturma sırasında çoklu rastgele örneklemenin bir sonucu olarak ağaçta önemli bir değişken ihmal edilmiş olabilir. Bununla birlikte, makul bir süre zarfında bu, anlamlı sonuçlar elde etmek için pratikte sıklıkla gerekli olan "veri mühendisliğinin" oldukça tipik bir örneğidir. Rastgele örneklemenin temel fikrine, örneğin "verinin genel görünüşü" hakkında bir fikir edinmek için başlangıçta küçük bir örneği almak, daha sonra bu örneği bazı otomatik yöntemlerle rafine etmek gibi çok sayıda iyileştirme söz konusudur (Hant ve diğerleri, 2001:252).

Karar ağaçlarında kullanılan birçok algoritma mevcuttur. Tablo 1'de bazı karar ağacı algoritmaları ve özellikleri görülmektedir.

Tablo 1 Bazı Karar Ağacı Algoritmaları ve Özellikleri

Karar Ağacı Algoritması	Özellikler
CART	Düğüm eğer bir terminal ise her biri iki dala ayrılır. Budama işlemi ağacın komplex yapısına göre temellendirilir. Sınıflandırma ve regresyonu destekleyici bir yapıya sahiptir. Dallar sürekli olarak değişkenleri hedefler. Verinin hazırlanmasına gereksinimi vardır.
C4.5 ve C5.0 (ID3 karar ağacı algoritmasının ileri versiyonları)	Her bir düğümden çıkan dallar ağacı oluşturur. Dalların sayısı oluşturulmak istenen sınıf sayısına eşittir. Ayırma işlemi bilgi kazancı esasına göre yapılır.
CHAID (Chi-Squared Automatic Interaction Detector)	Ki-kare testleri kullanılarak bölme işlemi gerçekleştirilir. Dalların sayısı iki ile oluşturulmak istenen sınıf sayısı arasında değişir.
SLIQ (Supervised Learning In Quest)	Hızlı bir sınıflayıcıdır. Ağaç budama algoritması hızlıdır.
SPRINT (Scalable Parallelizable Induction of Decision Trees)	Büyük veri kümelerinde kullanılır. Bölme işlemi tek bir nitelik değerini esas alır.

Kaynak: Bounsaythip C., Rinta-Runsala E. (2001). "Overview of Data Mining For Customer Behavior Modeling". VTT Information Technology Research Report. Ver 1. ss. 21.

Entropi ise, algoritmada kullanılması gerekli olan parametrelerden birisidir, bir sistemin düzensizliğini ifade eder. Entropi kavramı bilişim teorisine Claude Shannon tarafından uyarlanmıştır (Shannon, 1948: 392). Shannon'a göre bir olayın gerçekleşme olasılığı P(A) ise, A olayını tanımlayan entropi $-log_x P(A)$ şeklinde ifade edilir. Bir olay ile ilgili ortaya çıkabilecek birden fazla olası durum varsa, bu olay için gerekli olan enformasyon miktarı tüm olasılıklara bağlıdır ve bu durumda entropi aşağıdaki şekilde ifade edilir:

$$H = -\sum_{i=1}^{n} P(s_i) \log_2 P(s_i)$$

Entropi bir olayda kesinsizliğin ölçüsü iken; başka bir deyişle, bir olayın sonucunun kesin olarak belirlenmesi için gerekli olan enformasyon miktarını ifade ederken, bilgi kazanımı onun tersi olarak formülllendirilmektedir. Bilgi kazanımı 0 ile 1 arasında bir değer alır ve iki entropi arasındaki fark olarak ifade edilir:

Bilgi Kazanımı (öznitelik) = H (veri seti) – H (öznitelik)

Bilgi kazanımı yüksek olan öznitelikler algoritmada en değerli öznitelikler anlamına gelir. Özniteliklerin bilgi kazanımları en büyükten en küçüğe doğru sıralanır. Böylelikle en değerli öznitelik(ler) tespit edilerek ve diğerleri o adımda elenerek iterasyon gerçekleştirilir.

4. Metin Sınıflandırmada Karar Ağacı Algoritması ile Müşteri Yorumları Üzerine Bir Araştırma

4.1. Araştırmanın Amacı ve Önemi

Araştırmanın amacı, müşteri yorumlarını otomatik olarak "şikâyet", "talep", "teşekkür" sınıf etiketlerine atayacak bir karar ağacı modeli geliştirmektir. Bu modele ilişkin yazılacak kod ile yorumların otomatik olarak sınıflara atanması sağlanabilir. Bu yolla önce müşteri yorumlarından onları temsil edebilecek nitelikteki kelimeler çıkarılmış ve düğümler belirlenerek etiketlendirilmiştir.

Metin sınıflandırmada Naive Bayes, KNN, K-Ortalama, Karar Ağaçları gibi algoritmalar kullanılmaktadır. Yapılan araştırma karar ağaçları algoritması üzerine kuruludur. Yorumlardaki belli kelimelerin sınıf etiketleri için yorumları temsil edebilecek şekilde belirlenmesi ve algoritmada bu şekilde kullanılması sözlük temelli bir teknik olarak değerlendirilebilir.

4.2. Araştırmanın Örneklemi

Araştırmanın örneklemini bir işletme veri tabanında yer alan dördü "şikayet", dördü "talep" ve dördü "teşekkür" tipinde olan 12 adet müşteri yorumu oluşturmaktadır. Örnekler, kararsal örnekleme yoluyla seçilmiş ve yazım denetimi manuel yolla gerçekleştirilmiştir. Aşağıda her tipe ilişkin örnek bir yorum verilmiştir:

Şikâyet: "*23349 nolu siparişimin havalesini yapmama rağmen sipariş sayfasında havale bekleniyor yazıyor, acil cevap bekliyorum. Ayrıca benimle aynı gün alışveriş yapan arkadaşlara teslimat yapılmış, ancak bana dönülmedi de, sorun olup olmadığı konusunda bilgilendirilmek istiyorum.*" (Tablo 2, Yorum 1).

Talep: "*564 nolu ürünün stok durumu nedir? Kampanya dahilinde mi öğrenmek istiyorum. Beğenmezsem değişim yapabilir miyim? Acil dönebilirseniz sevinirim.*" (Tablo 2, Yorum 8).

Teşekkür: "*23266 nolu siparişimi 24 saat içerisinde gönderdiğiniz için teşekkür ediyorum, bunu oğluma hediye olarak aldım, perşembe vereceğim, olası bir numara değişikliğinde aynı gün haber vermem gerektiğini biliyorum, bana cumaya kadar süre tanırsanız sevinirim, ürün resimlerde görülenden daha güzel, teşekkürler.*" (Tablo 2, Yorum 10).

4.3. Araştırmanın Yöntemi

Araştırma pozitivist bir yaklaşımı benimsemektedir. Örnekleme ilişkin gerçek sınıf etiketleri semantik yolla belirlenmiş ve otomatik metin sınıflandırma için Karar Ağaçları Algoritması kullanılmıştır. Analizler işletme veri tabanından alınan örnek veriler üzerinde gerçekleştirildiğinden araştırmanın verileri ikincil veri kategorisindedir. Karar ağacı modelinin başarı ölçümü için ise kesinlik (precision) kriteri önerilmiştir.

4.4. Araştırmanın Bulguları

Öncelikle araştırmanın örneklemini oluşturan 12 adet yorum manuel yolla "şikâyet", "talep" ve "teşekkür" adlı tip sınıflarına atanmış ve her birini temsil edeceği öngörülen öznitelikler –kelimeler–belirlenmiştir. Tablo 2'de görüldüğü gibi kelimeler isim/sıfat/zarf temelli bir yapıda seçilmişlerdir ki, bu kelimeler karar ağacının düğümlerini oluşturacaktır. Ağacın birinci düğümü için Tablo 2'de görülen 22 kelimenin her biri için bilgi kazanımları hesaplanmalıdır. En fazla bilgi kazanımı değerine sahip olan kelime, ağacın birinci düğümünü oluşturacaktır.

Tablo 2 Karar Ağacının Birinci Düğümü İçin Kelimeler

Yorum	Yorumları Temsil Eden Kelimeler	Tip
1	Rağmen, havale, acil, sorun, ancak	Şikâyet
2	Değişim, hala, rağmen, kimse, net	Şikâyet
3	Soyulma, garantili, rağmen	Şikâyet
4	Kampanya, ancak	Şikâyet
5	Fiyat	Talep
6	Kampanya, çok, mutlu, teşekkür	Talep
7	İade, sorun, bilgi	Talep
8	Stok, kampanya, acil	Talep
9	Teşekkür, kampanya	Teşekkür
10	Teşekkür, güzel	Teşekkür
11	İade	Teşekkür
12	Fiyat, tavsiye, uygun, acil, güzel	Teşekkür

Tablo 3'te karar ağacının birinci düğümü için örnek 10 kelimeye ilişkin bilgi kazanımları görülmektedir. 22 kelimeden oluşan veri kümesi bir kelime torbası olarak düşünüldüğünde, her bir kelime torbada bulunup bulunmama durumuna göre "var" veya "yok" değerlerini almaktadır.

Tablo 3 Karar Ağacının Birinci Düğümü İçin Bilgi Kazanımı (Örnek 10 kelime)

	VARSA				YOKSA				
kelime	**şikayet**	**talep**	**teşekkür**	**entropi**	**şikayet**	**talep**	**teşekkür**	**entropi**	**bilgi kazanımı**
rağmen	**3**	**0**	**0**	**0,000**	**1**	**4**	**4**	**1,392**	**0,541**
ancak	2	0	0	0,000	2	4	4	1,522	0,317
güzel	0	0	2	0,000	4	4	2	1,522	0,317
teşekkür	0	1	2	0,918	4	3	2	1,530	0,208
çok	0	1	0	0,000	4	3	4	1,573	0,143
tavsiye	0	0	1	0,000	4	4	3	1,573	0,143
uygun	0	0	1	0,000	4	4	3	1,573	0,143
mutlu	0	1	0	0,000	4	3	4	1,573	0,143
stok	0	1	0	0,000	4	3	4	1,573	0,143
garantili	1	0	0	0,000	3	4	4	1,573	0,143

Tablo 3'te yapılan hesaplama adımları aşağıdaki şekildedir:

1. Örnek olarak "rağmen" kelimesi 3 adet şikâyet tipindeki yorumda yer almakta, ancak talep ve teşekkür tipindeki yorumlarda hiç yer almamaktadır: 3,0,0. Diğer yandan, her bir tipin yorum sayısı 4 olduğundan "yoksa" kısmındaki tipler bu kelimeye ilişkin olarak sırasıyla 1,4,4 değerlerini alacaktır.
2. Entropi değerleri varsa/yoksa durumlarına göre ayrı ayrı hesaplanmıştır. Örnek olarak rağmen kelimesine ilişkin entropiler:

Entropi (rağmen,varsa)

$=-3/(3+0+0) * \log_2(3/3+0+0)-0/(3+0+0) * \log_2(0/3+0+0)-0/(3+0+0) * \log_2(0/3+0+0)$

=0,000

Entropi (rağmen,yoksa)

$=-1/(1+4+4) * \log2(1/1+4+4)-4/(1+4+4) * \log2(4/1+4+4)-4/(1+4+4) * \log2(4/1+4+4)$

=1,392

3. Örnek olarak "rağmen" kelimesine ilişkin bilgi kazanımı için öncelikle veri setinin entropisi hesaplanmalıdır. Aşağıdaki formülde her bir 4/12 değeri, tipteki yorum sayısı/toplam yorum sayısını göstermektedir.

Entropi (veri seti)

$=-(4/12 * \log_2(4/12))-(4/12 * \log_2(4/12))-(4/12 * \log_2(4/12))$

=2

Bilgi kazanımı (rağmen)

= entropi (veri seti) – varsa (şikâyet + talep + teşekkür)/ toplam yorum sayısı * entropi (varsa,rağmen) – yoksa (şikayet + talep + teşekkür)/ toplam yorum sayısı * entropi (yoksa,rağmen)

=0,541

Bu durumda bilgi kazanımı en yüksek kelime "rağmen" olduğundan ağacın ilk düğümü olacaktır. *Başka bir deyişle, algoritma "rağmen" kelimesini içeren bir yorumun mutlaka şikâyet tipinde bir yorum olacağı kararını vermiştir.* Bu durumda ikinci düğüm kararı için "rağmen" kelimesinin geçtiği tüm yorumlar Tablo 2'den çıkarılarak iterasyona devam edilecektir. Tablonun yeni hali artık Tablo 4'te görüldüğü gibidir: "rağmen" kelimesi sadece şikâyet tipindeki yorumda geçtiğinden yorum sayısı 9'a düşmüştür.

Tablo 4 Karar Ağacının İkinci Düğümü İçin Kelimeler

Yorum	Yorumları Temsil Eden Kelimeler	Tip
4	Kampanya, ancak	Şikâyet
5	Fiyat	Talep
6	Kampanya, çok, mutlu, teşekkür	Talep
7	İade, sorun, bilgi	Talep
8	Stok, kampanya, acil	Talep
9	Teşekkür, kampanya	Teşekkür
10	Teşekkür, güzel	Teşekkür
11	İade	Teşekkür
12	Fiyat, tavsiye, uygun, acil, güzel	Teşekkür

Tablo 5'te karar ağacının ikinci düğümü için örnek 10 kelimeye ilişkin bilgi kazanımları görülmektedir. 14 kelimeden oluşan veri kümesi yine bir kelime torbası olarak düşünüldüğünde, her bir kelime torbada bulunup bulunmama durumuna göre "var" veya "yok" değerlerini almaktadır.

Tablo 5 Karar Ağacının İkinci Düğümü İçin Bilgi Kazanımı (Örnek 10 kelime)

	VARSA				YOKSA				
kelime	**şikayet**	**talep**	**teşekkür**	**entropi**	**şikayet**	**talep**	**teşekkür**	**entropi**	**bilgi kazanımı**
ancak	**1**	**0**	**0**	**0,000**	**0**	**4**	**4**	**1,000**	**0,503**
güzel	0	0	2	0,000	1	4	2	1,379	0,320
kampanya	1	2	1	1,500	0	2	3	0,971	0,186
bilgi	0	1	0	0,000	1	3	4	1,406	0,143
mutlu	0	1	0	0,000	1	3	4	1,406	0,143
sorun	0	1	0	0,000	1	3	4	1,406	0,143
stok	0	1	0	0,000	1	3	4	1,406	0,143
tavsiye	0	0	1	0,000	1	4	3	1,406	0,143
uygun	0	0	1	0,000	1	4	3	1,406	0,143
çok	0	1	0	0,000	1	3	3	1,406	0,143

Bu durumda bilgi kazanımı en yüksek kelime "ancak" olduğundan ağacın ikinci düğümü olacaktır. *Başka bir deyişle, algoritma bir yorumun "rağmen" varsa şikâyet, yoksa ama "ancak" varsa yine şikâyet tipinde bir yorum olacağı kararını vermiştir.* Bu durumda üçüncü düğüm kararı için "ancak" kelimesinin geçtiği 4 nolu yorum Tablo 4'den çıkarılarak iterasyona devam edilecektir. Tablonun yeni hali artık Tablo 6'da görüldüğü gibidir: "ancak" kelimesi tek kalan şikâyet tipindeki yorumda geçtiğinden yorum sayısı 8'e düşmüş ve şikâyet yorumları ayrılmıştır.

Tablo 6 Karar Ağacının Üçüncü Düğümü İçin Kelimeler

Yorum	Yorumları Temsil Eden Kelimeler	Tip
5	Fiyat	Talep
6	Kampanya, çok, mutlu, teşekkür	Talep
7	İade, sorun, bilgi	Talep
8	Stok, kampanya, acil	Talep
9	Teşekkür, kampanya	Teşekkür
10	Teşekkür, güzel	Teşekkür
11	İade	Teşekkür
12	Fiyat, tavsiye, uygun, acil, güzel	Teşekkür

Tablo 7'de karar ağacının üçüncü düğümü için örnek 10 kelimeye ilişkin bilgi kazanımları görülmektedir. 13 kelimeden oluşan veri kümesi yine bir kelime torbası olarak düşünüldüğünde, her bir kelime torbada bulunup bulunmama durumuna göre "var" veya "yok" değerlerini almaktadır.

Tablo 7 Karar Ağacının Üçüncü Düğümü İçin Bilgi Kazanımı (Örnek 10 kelime)

	VARSA				YOKSA				
kelime	**şikayet**	**talep**	**teşekkür**	**entropi**	**şikayet**	**talep**	**teşekkür**	**entropi**	**bilgi kazanımı**
güzel	**0**	**0**	**2**	**0,000**	**0**	**4**	**2**	**0,918**	**0,311**
bilgi	0	1	0	0,000	0	3	4	0,985	0,138
mutlu	0	1	0	0,000	0	3	4	0,985	0,138
sorun	0	1	0	0,000	0	3	4	0,985	0,138
stok	0	1	0	0,000	0	3	4	0,985	0,138
tavsiye	0	0	1	0,000	0	4	3	0,985	0,138
uygun	0	0	1	0,000	0	4	3	0,985	0,138
çok	0	1	0	0,000	0	3		0,985	0,138
kampanya	0	2	1	0,918	0	2	3	0,971	0,049
teşekkür	0	1	2	0,918	0	3	2	0,971	0,049

Bu durumda bilgi kazanımı en yüksek kelime "güzel" olduğundan ağacın üçüncü düğümü olacaktır. *Başka bir deyişle, algoritma bir yorumun "rağmen" varsa şikâyet, yoksa ama "ancak" varsa yine şikâyet, yoksa ama "güzel" varsa teşekkür tipinde bir yorum olacağı kararını vermiştir.* Bu durumda dördüncü düğüm kararı için "güzel" kelimesinin geçtiği 10 ve 12 nolu yorum Tablo 6'dan çıkarılarak iterasyona devam edilecektir. Tablonun yeni hali artık Tablo 8'de görüldüğü gibidir: "güzel" kelimesi iki teşekkür tipindeki yorumda geçtiğinden yorum sayısı 6'ya düşmüştür.

Tablo 8 Karar Ağacının Dördüncü Düğümü İçin Kelimeler

Yorum	Yorumları Temsil Eden Kelimeler	Tip
5	Fiyat	Talep
6	Kampanya, çok, mutlu, teşekkür	Talep
7	İade, sorun, bilgi	Talep
8	Stok, kampanya, acil	Talep
9	Teşekkür, kampanya	Teşekkür
11	İade	Teşekkür

Tablo 9'da karar ağacının dördüncü düğümü için örnek 10 kelimeye ilişkin bilgi kazanımları görülmektedir. 10 kelimeden oluşan veri kümesi yine bir kelime torbası olarak düşünüldüğünde, her bir kelime torbada bulunup bulunmama durumuna göre "var" veya "yok" değerlerini almaktadır.

Tablo 9 Karar Ağacının Dördüncü Düğümü İçin Bilgi Kazanımı (Örnek 10 kelime)

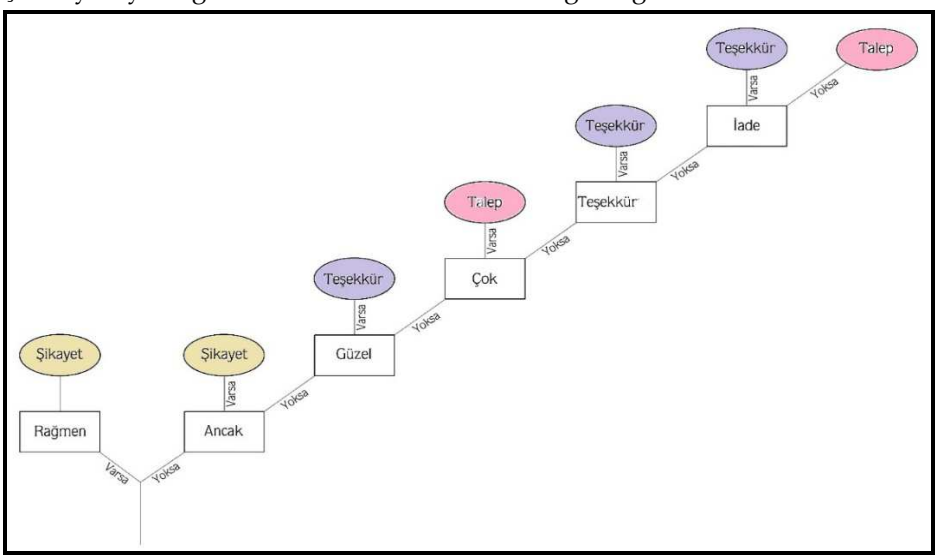
	VARSA				YOKSA				
kelime	**şikayet**	**talep**	**teşekkür**	**entropi**	**şikayet**	**talep**	**teşekkür**	**entropi**	**bilgi kazanımı**
çok	**0**	**1**	**0**	**0,000**	**0**	**3**	**2**	**0,442**	**0,550**
acil	0	1	0	0,000	0	3	2	0,971	0,109
bilgi	0	1	0	0,000	0	3	2	0,971	0,109
fiyat	0	1	0	0,000	0	3	2	0,971	0,109
mutlu	0	1	0	0,000	0	3	2	0,971	0,109
sorun	0	1	0	0,000	0	3	2	0,971	0,109
stok	0	1	0	0,000	0	3	2	0,971	0,109
iade	0	1	1	1,000	0	3	1	0,811	0,044
teşekkür	0	1	1	1,000	0	3	1	0,811	0,044
kampanya	0	2	1	0,918	0	2	1	0,918	0,000

Bu durumda bilgi kazanımı en yüksek kelime "çok" olduğundan ağacın dördüncü düğümü olacaktır. *Başka bir deyişle, algoritma bir yorumun "rağmen" varsa şikâyet, yoksa ama "ancak" varsa yine şikâyet, yoksa ama "güzel" varsa teşekkür, yoksa ama "çok" varsa talep tipinde bir yorum olacağı kararını vermiştir.* Bu durumda beşinci düğüm kararı için "çok" kelimesinin geçtiği 6 nolu yorum Tablo 8'den çıkarılarak

iterasyona devam edilmiş ve "teşekkür" ağacın beşinci düğümü, son olarak da "iade" ağacın altıncı düğümü olarak hesaplanmıştır.

4.5. Karar Ağacı Modeli ve Değerlendirme

Yukarıda açıklanan iterasyonlarla oluşturulan Karar Ağacı Modeli Şekil 2'de görüldüğü gibidir. Ancak oluşturulan model seçilen örneklem ile sınırlıdır. Daha büyük bir örnek kitle ile oluşturulan modelin kesinlik ölçütü de daha büyük olacaktır. Burada amaç, "yorumları temsil edebilecek nitelikteki özniteliklerin –kelimelerin– seçimi" yoluyla algoritmanın nasıl kullanılabileceğinin gösterilmesidir.



Şekil 2 Karar Ağacı Modeli

Bu durumda modele ilişkin olarak aşağıdaki kural çerçevesinde kod yazılabilir:

```
switch (HücreDeğeri){
        case "rağmen":
                tip = "şikayet";
                break;
        case "ancak":
                tip = "şikayet";
                break;
        case "güzel":
                tip = "teşekkür";
                break;
        case "çok":
                tip = "talep";
                break;
        case "teşekkür":
                tip = "teşekkür";
                break;
        case "iade":
                tip = "teşekkür";
                break;
        default:
                tip = "talep";
                break;
}
```

Sınıflandırma metodu tarafından oluşturulan modelin değerlendirilmesi için ne kadar iyi performans gösterdiğinin ölçülmesi gerekir. Bir modelin performansı sınıfı doğru tahmin etme yeteneğidir. Sınıflandırma modellerinin başarı ölçümünde genellikle kesinlik (precision) kriteri kullanılır.

Kesinlik=
$$\frac{\text{Gerçek değeri pozitif olup pozitif olarak sınıflandırılan örnek sayısı}}{\text{Gerçek değeri pozitif olan örnek sayısı}}$$

Aşağıdaki gerçek yorumları oluşturulan modele göre şikâyet, talep, teşekkür sınıflarından birine atamak isteyelim:

Yorum a: "Acaba kampanya ne zaman bitecek, yani ne zaman puan kazanmaya başlayacağım?"

Yorum b: "Ben de 1+1 logolu 2 tane ürün ekledim ama ikisinin de fiyatı hesaplandı bir sorun var sanırım."

Yorum c: "23732 nolu siparişimin iptal edilmesini istiyorum. Çünkü siparişime 1 ilave daha olacak. İki kargo ödemek istemiyorum."

Yorum d: "Ürün kodu 5289 nolu ayakkabıyı almak istiyorum ancak stoklarda bulunmuyor ve ben sizlerden yardımı olmanızı rica ediyorum."

Yorum e: "23899 numaralı siparişimin iptal edilmesini istemiştim, siz de iptal edildiğini söylemiştiniz ama bugün kargo geldi."

Tablo 10'da görüldüğü gibi oluşturulan model 5 yorumdan 3'ünü doğru sınıfa atamıştır. Bu durumda kesinlik ölçütü %60 olarak hesaplanır. Doğru bir kesinlik ölçütü hesabı yapabilmek için örnek sayısının artırılması gerekir. Diğer yandan, yorumların doğru sınıfa atanamamasının farklı nedenleri olabilir. Örneğin;

1. Yorumları temsil eden kelimeler yanlış/az/çok seçilmiş olabilir, bu durum metin sınıflandırmanın en zor yanıdır. Sınıfları temsil edebilecek şekilde iyi bir sözlük alt yapısı gerektirir. Sözlük oluşturma, karar ağaçlarında ön budama işlemi olarak da değerlendirilebilir.
2. Yorumlarda geçen kelimeler yazım yanlışları içeriyorsa, onları temsil eden kelimelerle örtüşmeyecek ve hesaplamalar yanlış yapılabilecektir.
3. Yorumlar kısaltmalar içerebilir, dolayısı ile yine onları temsil eden kelimelerle örtüşmeyecek ve hesaplamalar yanlış yapılabilecektir.
4. Kelimelerin sesteşlik özelliği de yorumların yanlış sınıfa atanmasına neden olabilir.

Tablo 10 Karar Ağacı Modeli Üzerinde Bir Uygulama

	Gerçek Sınıf	Modelin Atadığı Sınıf
Yorum a	Talep	Talep
Yorum b	Şikâyet	Talep
Yorum c	Talep	Talep
Yorum d	Şikâyet	Şikâyet
Yorum e	Şikâyet	Talep

5. Sonuç ve Öneriler

İletişimin ve enformasyonu işlemenin son derece önemli olduğu bir çağda yaşıyoruz. Bu çerçevede teknoloji, enformasyon iletişimini daha süratli ve eksiksiz yapabilmek için sürekli geliştirilmekte. Dünyada yaşanmakta olan büyük bir dönüşüm ile işletmeler de kendi sistemlerini bilgi ve iletişim teknolojilerinin sağladığı yeni ortamlarla uyumlu olacak şekilde yeniden düzenlemekte.

Günümüzde metin verilerin logaritmik artışı, metin sınıflandırma yöntemlerini kullanmayı zorunlu hale getirmiştir. Bu veriler işletme veri tabanlarındaki müşteri yorumları olabileceği gibi, sosyal medyada paylaşılan metin formatındaki içerikler de olabilir. Metin sınıflandırma eğer işletmeye katma bir değer sağlıyorsa, tüm ortamlar için söz konusu edilebilir. Böylelikle işletmeler, müşterileriyle ilgili sahip oldukları verileri analiz ederek onların ihtiyaç ve tercihleri doğrultusunda uygun stratejiler geliştirebilir, ürün/hizmetleri hakkında genel bir görünüş elde ederek rekabet avantajı sağlayabilirler.

Bu çalışmada kullanılan metin veriler, bir işletme veri tabanında yer alan müşteri yorumlarının oluşturduğu verilerdir. Binlerce sayıdaki bu verinin işletme tarafından manuel olarak sınıflara atanması imkânsız olduğundan, atama işleminin otomatik olarak yapılması gerekir. Bu çalışmada böyle bir program oluşturmak üzere bir model geliştirilmiştir. Modeller, metin sınıflandırmada kullanılan başlıca araçlardandır

ve kaçınılmaz olarak basittirler. Bununla birlikte model geliştirmenin bir amaç değil, bir araç olduğu unutulmamalıdır.

Model, "yorumları temsil edebilecek nitelikteki kelimelerin seçimi" yoluyla algoritmanın nasıl kullanılacağını göstermesi bakımından önemlidir. Bu da sözlük temelli bir alt yapı gerektirir ki, sözlük oluşturma metin sınıflandırmada başvurulan yöntemlerden biridir. Ancak oluşturulacak sözlüklerde yer alan kelimeler sınıflandırma yapılacak mecraya göre farklılık gösterebileceği gibi, sınıf etiketlerine, sektöre ve bunların çeşitli kombinasyonlarına göre de farklılık gösterebilir. Örneğin, "değişim" kelimesinin ürün bazında faaliyet gösteren bir işletmenin görece olarak talep etiketli sınıflarında yer alması ihtimali yüksekken, hizmet bazında faaliyet gösteren bir işletmenin aynı sınıf etiketinde yer alması ihtimali düşük olabilir. Bu nedenle sözlük temelli bir alt yapı oluşturma kontrol otomasyonu gerektirir.

Gelecek çalışmalar için; örneklem ve test verisinin genişletilmesi, yorumları temsil eden kelimelerin yeniden düzenlenmesi, bunun için sınıf etiketleri ile ilgili bir sözlük oluşturulması, sözlüğün sektörel bazda da değerlendirilmesi, bu sözlüğün inşasında sınıf etiketlerine yönelik olarak çok sayıda yorumun manuel değerlendirmesinden yararlanılması ve böylelikle sözlük için bir kontrol otomasyon sisteminin kurulması önerilebilir.

KAYNAKÇA

Ahsan, Syed; Shah, Abad. (2006). Data, Information, Knowledge, Wisdom: A Doubly Linked Chain?. Proceedings of the 2006 International Conference on Information & Knowledge Engineering IKE 2006. June 26-29. Las Vegas, Nevada.

Amasyalı, Fatif; Diri, Banu; Türkoğlu, Filiz. (2006). Farklı Özellik Vektörleri İle Türkçe Dokümanların Yazarlarının Belirlenmesi. 15. Turkish Symposium On Artificial Intelligence And Neural Network. 21-24 June. Muğla, Türkiye.

Baker, Douglas; McCallum, Andrew Kachites. (1998). Distributional Clustering of Words for Text Classication. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. August 24-28. Melbourne, Australia. pp. 96-103.

Bounsaythip, Catherine; Rinta-Runsala, Esa. (2001). Overview of Data Mining For Customer Behavior Modeling. *VTT Information Technology Research Report*. Ver 1. ss. 21.

Çakıroğlu, Ünal; Özyurt, Özcan. (2006). Türkçe Metinlerdeki Yazım Yanlışlarına Yönelik Otomatik Düzeltme Modeli. Elektrik-Elektronik-Bilgisayar Mühendisliği Sempozyumu Ve Fuarı-ELECO'2006. Aralık 06-10. Bursa. Türkiye. ss.7-9.

Delen, Dursun; Crossland, Martin D. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications*. Vol 34. Issue 3. pp. 1707-1720.

Fan, Weiguo; Wallace, Linda; Rich, Stephanie; Zhang, Zhongju. (2006). Tapping into the Power of Text Mining. *Communications of the ACM*. Vol 49. Issue 9. pp. 76-82.

Forman, George. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*. Issue 3. pp. 1289-1305.

Fricke, Martin. (2009). The knowledge pyramid: a critique of the DIKW hierarchy. *Journal of Information Science*. Vol 35. pp. 131-142.

Hand, David; Mannila, Heikki; Smyth, Padhraic. (2001). *Principles of Data Mining*. Cambridge: MIT Press.

Joachims, Thorsten. (1999). Transductive Inference for Text Classification using Support Vector Machines. Proceedings of the Sixteenth International Conference on Machine Learning. June 27-30. Bled, Slovenia. pp. 200-209.

McCallum, Andrew; Nigam, Kamal. (1998). A Comparison of Event Models for Naive Bayes Text Classification. Learning for Text Categorization: Papers from the 1998 AAAI Workshop. July 26-27. Madison, Wisconsin. pp. 41-48.

Nigam, Kamal; Lafferty, John; McCallum, Andrew. (1999). Using Maximum Entropy for Text Classification. In IJCAI-99 Workshop on Machine Learning for Information Filtering. August 1st. Stockholm, Sweden. pp. 61-67.

Nigam, Kamal; Mccallum, Andrew Kachites; Thrun, Sebastian; Mitchell, Tom. (2000). Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*. Vol. 39. Issue 2-3. pp. 103-134.

Nisbet, Robert; Elder, John; Miner, Gary. (2009). *Handbook of Statistical Analysis and Data Mining Applications*. Burlington: Elsevier.

Oflazer, Kemal. (2012). Türkçe Doğal Dil İşleme. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri Ve Mühendisliği Dergisi*. Cilt 5. Sayı 2. ss. 1-12.

Rogati, Monica; Yang, Yiming. (2002). High-performing feature selection for text classification. CIKM '02 Proceedings of the eleventh international conference on Information and knowledge management. November 04 – 09. McLean, Virginia. pp. 659-661.

Shannon, Claude. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*. Vol. 27. ss.379–423.

Soucy, Pascal; Mineau, Guy. (2001). A Simple KNN Algorithm For Text Categorization". IEEE International Conference. 11-14 June. Helsinki. Finland. pp: 647-648.

Sriram, Bharath; Fuhry, David; Demir, Engin; Ferhatosmanoğlu, Hakan; Demirbaş, Murat. (2010). Short text classification in twitter to improve information filtering. Proceedings of The 33rd İnternational ACM SIGIR Conference On Research And Development in Information Retrieval. July 19 – 23. Geneva. Switzerland. pp. 841-842.

Sütcü, Cem Sefa; Aytekin, Çiğdem. (2013). *Elektronik Ticaretten Sosyal Ticarete Dönüşüm Sürecinde Ölçümleme*. İstanbul: Der'in Yayınevi.

Şeker, S. Evren. (2015). Metin Madenciliği (Text Mining). *YBS Ansiklopedi*. (c.2, s.3, ss. 30-32).

Tong, Simon; Koller, Daphne. (2001). Support Vector Machine Active Learning with Applications to Text Classification. *Journal of Machine Learning Research*. Issue 11. pp. 45-66.