T.C.

MALTEPE ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

METİN MADENCİLİĞİ TABANLI BİR WEB SİTESİ SINIFLANDIRMA ARACI TASARIMI

Filiz Erten

Yüksek Lisans Tezi

Tez Danışmanı: Yrd. Doç. Dr. Şenol Zafer Erdoğan

İstanbul, 2015

T.C.
MALTEPE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

METİN MADENCİLİĞİ TABANLI BİR WEB SİTESİ SINIFLANDIRMA ARACI TASARIMI

Filiz Erten

Yüksek Lisans Tezi

Tez Danışmanı: Yrd. Doç. Dr. Şenol Zafer Erdoğan

İstanbul, 2015

Bu tez çalışması, Maltepe Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 21/01/2015 tarih ve 2015/02 sayılı kararıyla oluşturulan jüri tarafından ***Bilgisayar Mühendisliği Yüksek Lisans Tezi*** olarak kabul edilmiştir.

JÜRİ

Yrd.Doç.Dr. Şenol Zafer ERDOĞAN
Danışman

Yrd.Doç.Dr. Volkan TUNALI
Üye

Yrd.Doç.Dr. Fatih YÜCALAR
Üye

ÖZET

Teknoloji her geçen gün şaşırtıcı biçimde ilerlemekte ve hayatımızı kolaylaştıran pek çok yeniliği beraberinde getirmektedir. Bu gelişimlere paralel olarak, analizi mümkün görünmeyen veriler kullanılır hale gelmektedir.

Veritabanında kategorize edilerek saklanan yapısal verilerin yanı sıra, yapısal olmayan metin halindeki veriler de sınıflandırma yöntemleri kullanılarak kategorilere ayrılabilir ve dolayısıyla sorgulanabilir bir yapıya dönüştürülebilir. Metin verisi üzerinden analizler yapılabilir ve aranan bilgiye rahatlıkla ulaşım sağlanabilir.

Bu tez çalışmasında, ortaya çıkarılan sınıflandırma aracı ile Google arama motoru üzerinde yazılan anahtar kelime sonucunda çıkan bağlantıların gerçekten o anahtar kelime ile uyumlu olup olmadığı ortaya çıkarılmaktadır. Metin madenciliği sınıflandırma algoritmalarından olan Naïve Bayes metodu kullanılarak sonuçlar analiz edilmiş ve arama sonucunda çıkan bağlantıların o anahtar kelime ile uyumlu olma olasılıkları hesaplanmıştır.

Bu tez 2015 yılında yapılmıştır ve 37 sayfadan oluşmaktadır.

Anahtar Kelimeler: Metin Madenciliği, Naïve Bayes Olasılık Tespiti, İnternet Sitesi Analizi

SUMMARY

Technology is improving surprisingly every day and bringing many other innovations with it. In parallel with these developments, data that seems impossible to analyze becomes usable.

Besides structural data kept with categories, other text data can also be categorized by classification methods and becomes usable to examine. Text data can be analyzed and the searched information can be achieved easily.

In this study, by the classification method developed, it is possible to see if the results of a Google search are really compatible with the key word. The results are analyzed by using Naïve Bayes, one of text mining classification algorithms, and the compatibility possibilities between the results and key words are calculated.

This thesis has been completed in 2015 and consists of 37 pages.

Keywords: Text Mining, Naïve Bayes Detection Probability, Web Site Analysis

TEŞEKKÜR

Tez sürecimin başlangıcından itibaren desteğini ve yardımını esirgemeyen tez danışmanı hocam Sayın Yrd. Doç. Dr. Şenol Zafer ERDOĞAN'a teşekkürlerimi sunarım.

Tez konusu seçimi sonrasında, tezimi ilerletme yönünde verdiği fikir ve yorumlardan dolayı hocam Sayın Yrd. Doç. Dr. Volkan TUNALI'ya teşekkür eder, saygılarımı sunarım.

Yüksek Lisans eğitim sürecimin başından beri beni sürekli destekleyen ve yalnız bırakmayan aileme teşekkür eder, saygılarımı sunarım.

İÇİNDEKİLER

KISALTMALAR

Kısaltma	İngilizcesi	Türkçesi
HTML	Hyper Text Markup Language	Zengin Metin İşaret Dili
XML	Extensible Markup Language	Genişletilebilir İşaretleme Dili
CRM	Customer Relationship Management	Müşteri İlişkileri Yönetimi
DNA	Deoxyribose Nucleic Acid	Deoksiribo Nükleik Asit
SVM	Support Vector Machine	Destek Vektör Makineleri
SQL	Structured Query Language	Yapılandırılmış Sorgu Dili
KNN	K Nearest Neighborhood	K En Yakın Komşu

ŞEKİLLER

Sayfa

ÇİZELGELER

Sayfa

DENKLEMLER

1. GİRİŞ

Günümüzde internet, intranet ve veritabanı gibi ortamlarda çok büyük miktarda veri depolanmaktadır. Verinin tipi, direkt analiz işlemini yapmaya uygun olabilir ya da metin şeklinde tutulan bir veriyse, analiz yapabilmek için verinin buna uygun bir yapıya getirilmesi gerekir. Direkt analiz yapılabilen veriler yapısal veri olarak adlandırılır. Veritabanında bir satır ve sütun olarak tutulur. Metin verileri ise yapısal olmayan verilerdir. Bu bilgiyi kullanabilmek için öncesinde metnin bazı işlemlerden geçmesi gerekir.

Metin madenciliği işte tam burada devreye girer. Metin verileri genelde çok büyük boyuta sahiptir. İstenilen bilgiye ulaşabilmek için süzme, kategorilere ayırma gibi işlemler uygulanır. Bu işlemlerle veride kullanılmayacak bilgiler atılır, değerli bilgi denilen ve analiz için kullanılacak veriler kalır.

Süzme ve kategorizasyon işlemleri sonrasında yapısal bir veri elde edilir. Veri artık analiz işlemlerine hazırdır.

1.1. Tezin Amacı

Bu tez çalışmasında amaç, internet üzerinden yapılan aramalarda, listelenen sonuçlardaki sitelerin anahtar kelime ile uyumlu site olma olasılıkları hesaplanarak kullanıcıyı doğru bağlantılara yönlendirmektir. Bunun için Visual Studio. NET platformu üzerinde bir Windows Forms uygulaması geliştirilmiştir. Geliştirilen uygulamada Visual Basic kodlama dili kullanılmıştır.

Uygulamanın çalışma adımları şu şekildedir;

1. Gerçekte e-ticaret sitesi olma özelliğine sahip 10 kadar internet sitesinin ana sayfa html kaynakları otomatik olarak alındı ve içindeki kelimeler ayrıştırılarak veritabanına yazıldı.

2. Benzer şekilde 10 kadar e-ticaret sitesi olmayan sitenin ana sayfa içeriği kelimelere ayrılarak veritabanına yazıldı.
3. Oluşan iki kategoriye ait veritabanında bulunan her kelime için, Naïve Bayes algoritmasındaki formül uygulanarak oran hesaplaması yapıldı.
4. Bu ön çalışma ile kelime için e-ticaret sitesi olma veya olmama durumundaki oran hesaplanmış oldu.
5. Kullanıcı, Windows Forms uygulaması olarak hazırlanan uygulamayı çalıştırdığında kelime araması yapabilir. Arama Google arama motoru üzerinden gerçekleştirilir. 20 arama sonucu üzerinde analiz işlemleri yapılarak her site için yapılan oransal sonuçlar liste şeklinde kullanıcıya listelenir. Kullanıcı, listelenen oranları değerlendirerek gerçekte e-ticaret sitesi olan siteleri görebilir.

1.2. Tezin Kapsamı

Tez çalışmasında kullanılan veri, yapısal türde olmayan bir veridir. Bu tür verilerin analiz, sorgulama gibi işlemlerde kullanılabilmesi için yapısal veri şekline dönüştürülmesi gerekir. Bunun için çeşitli tekniklere ihtiyaç duyulur. Bunlardan biri Metin Madenciliği teknikleridir. Metin Madenciliği ise temelde Veri Madenciliği'ne dayanır.

Tezin genel bilgileri kapsayan ilk bölümünde Veri Madenciliği'nden bahsedilmektedir. Tez çalışmasının dayanağı olan Metin Madenciliği konusu ise ikinci bölümünde detaylı olarak anlatılmıştır. Metin Madenciliği, XML, HTML, ofis dokümanları (Word, Excel), metin dokümanları gibi yapısal olmayan birçok kaynaktan veri toplamaktadır. Bu veriler boyut olarak çok büyük değerlere sahip olabilirler.

2. GENEL BİLGİLER

Veri, çeşitli formatlarda olabilir. Bazı veriler hazır, kullanılabilir durumda iken bazı verilerin ise kullanım öncesinde bir dizi işlemden geçmesi ve uygun formata dönüşmesi gerekir. Örneğin metin verisi, direkt olarak kullanılamaz. Öncesinde işlenmesi, kullanıma uygun formata dönüştürülmesi gerekir. Bunun için çeşitli yöntemler mevcuttur.

2.1. Veri Madenciliği

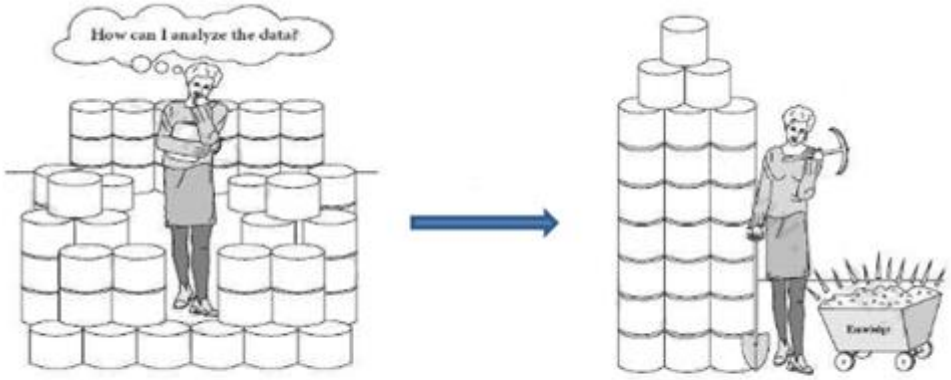
Saklı olan her türlü kaynak gün ışığına çıkarılmayı bekler. Elmas, altın, kömür gibi değerli madenler araştırma ve kazı yöntemleri ile gün ışığına çıkar.

Yapısal olmayan veri de aslında bir tür madendir ve keşfedilmeyi bekler. Veri madenciliği de işte tam bu aşamada devreye girer. Veriyi keşfederek ortaya çıkarır ve analiz yapmaya uygun bir yapıya getirir.

2.1.1. Veri Madenciliği Nedir?

Veri Madenciliği, büyük miktardaki veri kaynağı içerisinden, gelecekle ilgili tahmin yapmamızı sağlayacak kural ve bağlantıların aranması olarak özetlenebilir [1]. Veri üzerinde bir anlamda bilgi keşfi yapılır. Veriyi çözümleyip bilgiye ulaşabilmek, veri üzerinde çözümlemeler yapabilmek amacıyla veri madenciliği yöntemi ortaya çıkmıştır. Veri madenciliği, bir sorgulama işlemi veya istatistik programlarıyla yapılmış bir çalışma değildir. Veri madenciliği milyarlarca veri ve çok fazla değişken ile ilgilenir. Teknolojik gelişmeler dünyada gerçekleşen birçok işlemin elektronik olarak kayıt altına alınmasını, bu kayıtların kolayca saklanabilmesini ve gerektiğinde erişilebilmesini hem kolaylaştırıyor, hem de bu işlemlerin her geçen gün daha ucuza mal edilmesini sağlıyor. Bununla beraber ilişkisel veri tabanlarında saklanan birçok veriden, karar alma noktasında anlamlı

çıkarımlar yapabilmek için bu verilerin bilinçli uzmanlarca analiz edilmesi gerekir. Veri miktarı çok büyük olduğu için bazı özel analiz algoritmaları geliştirilmiştir [2]. Şekil 2.1.'de veri madenciliğinin analiz işlemi için veri üzerinde yaptığı örnek bir düzenleme görülmektedir.



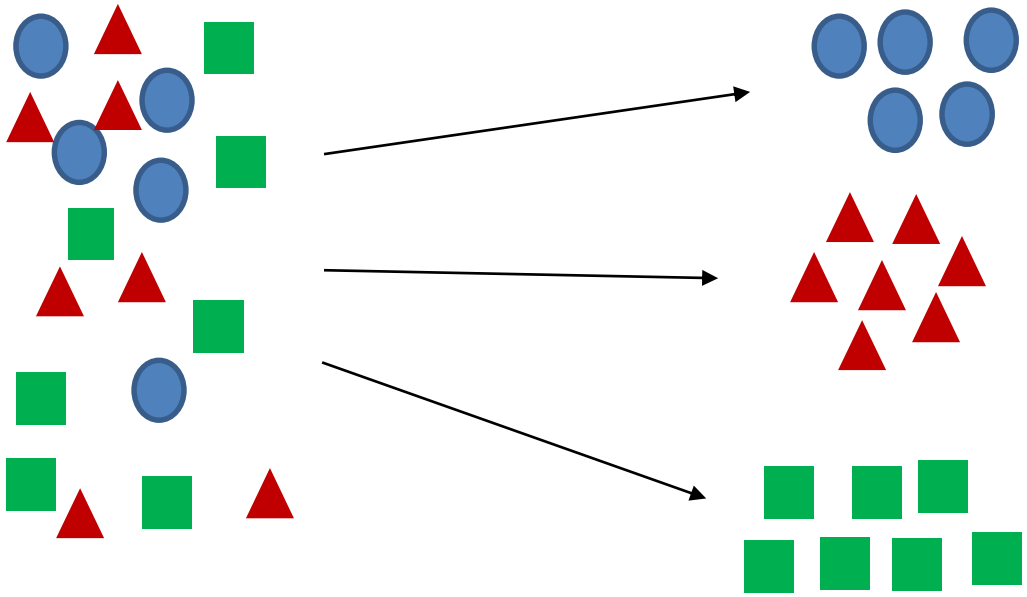
Şekil 2.1. Veri madenciliğinde bilgiye ulaşma [3]

2.1.2. Veri Madenciliğinde Kullanılan Yöntemler

İşlenmemiş veriden bilgi elde etmenin birçok yolu vardır. İstenilen bilgiye ulaşmak için öncelikle uygulama alanını doğru olarak belirlemek gerekir. Bu uygulama alanına uygun veri kümesi oluşturularak, veri üzerinde ön işlemler uygulanır ve analiz yapılacak sisteme uygun yapıya getirilir. Bu yöntemlerle veri daha küçük boyutlara ulaşır, gereksiz ilgiler ayıklanmış olur. Bu aşamadan sonra veri madenciliği tekniklerinden biri uygulanarak sonuç elde edilebilir.

Sınıflandırma (Classification): Veri madenciliğinde oldukça kullanılan popüler bir tekniktir. Veri kümesindeki nesnenin nitelikleri incelenerek bir sınıfa atanır. Burada önemli olan konu, sınıflandırmanın net olarak belirtilmiş olmasıdır. Şekil 2.2.'de verinin sınıflandırılması gösterilmektedir.

Sınıflandırma algoritmalarında belirlenen eğitim kümesinden dağılım şekli öğrenilir ve edinilen test verisi bu sette yakın olduğu veri grubu üzerinden sınıflandırılır. Şekil 2.2.'de üçgen daire ve karelerin bulunduğu dağınık bir veri kümesi görülmektedir. Veri, şekillere göre sınıflara ayrılır ve eğitim için set hazırlanmış olur. Bu aşamada, test için yeni bir şekil geldiğinde bu üç sınıftan hangisine yakınsa o sınıfa atama yapılabilir.

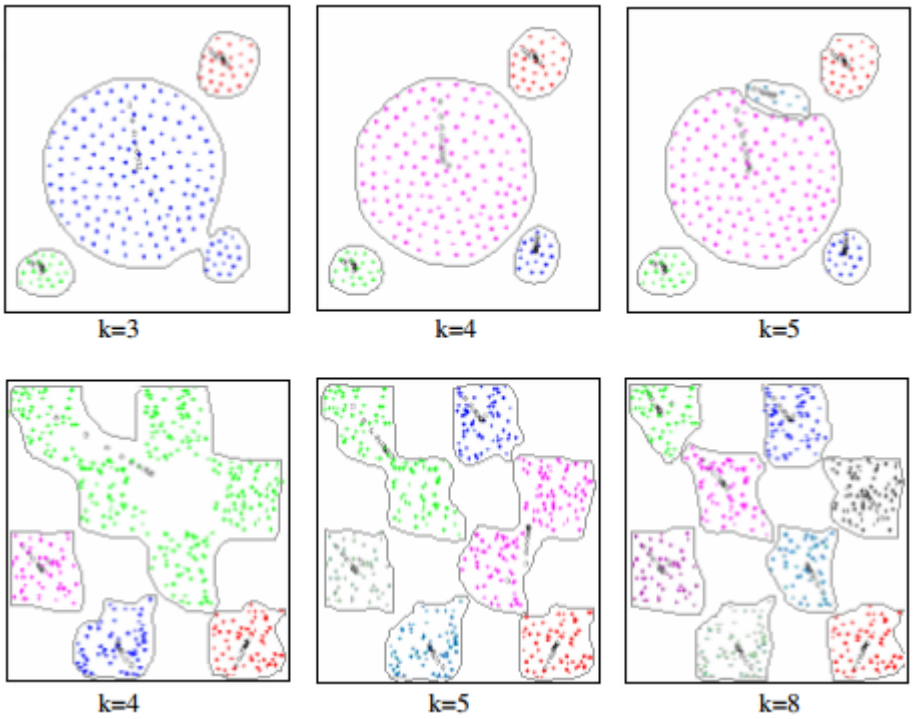


Şekil 2.2. Verinin sınıflandırılması

Kümeleme (Clustering): Veri madenciliğinde kullanılan bir diğer teknik de kümelemedir. Bu teknikte, belirli bir yapı içinde geçen terimlere göre gruplar oluşturulur. Bu gruplar içinde en çok geçen verilerden yararlanılarak benzerlik grupları oluşturulur ve buna göre kümeleme işlemi yapılır. Örneğin, bir ürün alımında alıcılar yaş gruplarına göre ayrılmak istendiğinde, kümeler belirlenir (genç-orta-yaşlı veya yaş ölçeğine göre grup yapılabilir) ve veri hangi gruba uygunsa bu gruba atanır.

Küme sayısındaki artış, verinin daha detaylı analizini sağlamaktadır. Sayı arttıkça, veri setinde daha özel bilgiye ulaşılır. Şekil 2.3.'de görüldüğü gibi, küme sayısındaki artış daha detay veriye ulaşmayı sağlar. Küme sayısı 4 olduğunda, 4

farklı özelliğe sahip veriye ulaşılır. Küme sayısı özellik gibi düşünülebilir. 4 küme 4 farklı özellik anlamında kullanılabilir.



Şekil 2.3. Küme sayısındaki artışa bağlı veri çeşitliliği [4]

***Birliktelik (Association)*:** Bu teknikte, verinin diğer veri ile yakınlığı dikkate alınarak birlikte olma durumları üzerine tahmin yapılır. Örneğin, çocuk maması satın alan müşterinin, çocuk bezi alma olasılığı tespit edilerek bu iki ürün arasında bağlantı kurulabilir. Bu veri ile müşterinin çocuk sahibi olduğu saptanabilir ve müşteriye çocuk ürünleri ile ilgili çıkan kampanyaların mesajları gönderilebilir. Bu tarz bir çalışma potansiyel müşteriye ulaşmada önemli bir araç olarak ortaya çıkmaktadır.

Birliktelik kuralına ilişkin olarak geliştirilen bazı algoritmalar şunlardır; AIS, SETM, Apriori, Partition, RARM - Rapid Association Rule Mining, CHARM. Bu algoritmalar içerisinde, ilk olanı AIS, en bilineni ise Apriori algoritmasıdır [5].

Veri üzerinde her zaman bilinen özellikler üzerinden analiz yapmak gerekmez. **Regresyon** yönteminde, bir veri başka bir veri türü üzerinden tahmin yürütülerek saptanabilir. Örneğin, bir ziraatçı buğday verimi ve gübre miktarı arasındaki ilişkiyi, bir mühendis, basınç ve sıcaklık, bir ekonomist gelir düzeyi ve tüketim harcamaları, bir eğitimci öğrencilerin devamsızlık gösterdiği gün sayıları ve başarı dereceleri arasındaki ilişkiyi bilmek isteyebilir. Regresyon, iki (ya da daha çok) değişken arasındaki doğrusal ilişkinin fonksiyonel şeklini, biri bağımlı diğeri bağımsız değişken olarak bir doğru denklemi olarak, göstermekle kalmaz, değişkenlerden birinin değeri bilindiğinde diğeri hakkında kestirim yapılmasını sağlar [6].

Tahminleme (Forecasting) metodu, adından da anlaşılacağı üzere veri üzerinde geleceğe dair tahmin yürütme yöntemidir [7]. Nüfus artışı, satış rakamları tahminleri yapmak için kullanılır.

2.2. Metin Madenciliği

Her bilgi aslında bir veridir. İlk bakıldığı anda bir anlam taşımıyor olsa da, bilgi işlenerek anlamlı hale getirilebilir. Metin bilgileri de bu tür verilerdir ve işlenerek kullanılabilir hale getirilebilirler.

Metin Madenciliği, genellikle yapısal olmayan veriden bilgi keşfi yaparak dokümanları analiz etmeyi sağlayan bir teknolojidir [8]. Metin yazımında standart kurallar bulunmadığından, bilgisayarın bunu anlaması mümkün değildir. Bilgisayarın bunu anlayabileceği seviyeye getirmek için ise metin madenciliği teknikleri kullanılır.

Metin madenciliği çalışmaları metni veri kaynağı olarak kabul eden veri madenciliği (data mining) çalışmasıdır. Diğer bir tanımla metin üzerinden yapısallaştırılmış veri elde etmeyi amaçlar [9]. Örneğin metinlerin sınıflandırılması, kümelenmesi (clustering), metinlerden konu çıkarılması (concept/entity extraction),

sınıf taneciklerinin üretilmesi (production of granular taxonomy), duygu analizi (sentimental analysis), metin özetleme (document summarization), varlık ilişki modellemesi (entity relationship modelling) gibi çalışmaları hedefler [9].

Metin Madenciliği teknikleri dört temel kategoriye ayrılır: **sınıflandırma** (classification), **birliktelik analizi** (association analysis), **bilgi çıkarım** (information extraction) ve **kümeleme** (clustering). Sınıflandırma işlemi nesnelerin daha önceden bilinen sınıflara ya da kategorilere dâhil edilmesidir. Birliktelik analizi, sıklıkla birlikte yer alan ya da gelişen sözcük ya da kavramların belirlenmesi ve böylece doküman içeriğinin ya da doküman kümelerinin anlaşılmasını amaçlamaktadır. Bilgi çıkarım teknikleri ile dokümanların içerisindeki kullanışlı veri ya da ifadeler bulunmaya çalışılmaktadır. Kümeleme analizi, doküman kümelerinin temelini oluşturan yapıların keşfedilmesi amacıyla uygulanmaktadır [10]. Şekil 2.4.'de metin madenciliğinde veri işleme akışı görülmektedir.

Şekil 2.4. Metin madenciliği veri işleme akışı

Çalışma öncelikle veri toplama işlemi ile başlar. Bunun için genellikle arama motoru olarak Google kullanılır ve internet sitelerinden HTML (Hyper Text Markup Language) kaynak kodları çekilerek veri seti oluşturacak şekilde depolanır.

HTML kaynak kodlarının yanı sıra, bilgisayarlarda bulunan doküman dosyaları da kaynak olarak kullanılabilir. Şekil 2.4.'de görülen ikinci aşamada ise doküman işleme işlemi uygulanır. Bu işlem sırasında kelimelerin anlamsal değerleri

bulunur (fiil, yüklem, zamir, sıfat vb.), yazım hataları tespit edilir, anlamsal değeri bulunmayan kelimeler filtrelenir.

Metin madenciliğinin en büyük sorunu işleyeceği veri kümesinin yapısal olmamasıdır. Genellikle doğal dil kullanılarak yazılmış dokümanlar üzerinde çalışılan metin madenciliği alanında ön işleme aşaması veri temizlemenin yanında veriyi uygun formata getirme işlemini de gerçekleştirmektedir [11].

Yapısal veri dönüşüm aşamasında, doğal dil işleme işlemleri uygulanarak veri temizleme ve özet çıkarma işlemleri yapılır.

Sorgu ve analiz işlemleri aşamasında ise veri madenciliği yöntemleri devreye girer. Bu aşamada veri, yapısal veriye dönüştürülmüş durumdadır. Karar ağaçları, kümeleme, genetik algoritmaları gibi veri madenciliği teknikleri ile veri analiz işlemi gerçekleştirilir.

Sonuç aşamasına gelindiğinde veri kullanıcıya sunuma hazır hale gelmiş demektir. Bu aşamada, son kullanıcıya, kullanıcının yorumlayabileceği şekilde grafik, tablo formatında veri sunulur.

2.2.1. Metin Madenciliği Kullanım Alanları

Metin madenciliği çalışmaları genelde devlet seviyesi, bilimsel araştırma ve iş dünyası ihtiyaçları için çeşitli çözümler sunmaktadır. Kullanım alanları aşağıdaki şekilde sıralanabilir [12,13,14]:

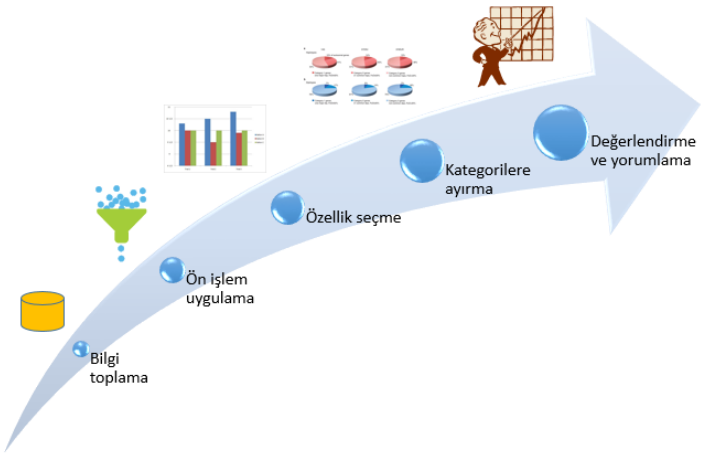
- Müşteri İlişkileri Yönetimi (Customer Relationship Management, CRM): Bütün müşterilerin e-posta, işlem, çağrı merkezi ve anket gibi erişim noktalarından elde edilen metin bilgilerinden nitelikli bilgi çıkarılır. Bu

nitelikli bilgi müşterinin terk etme ve çapraz satışlarını tahmin etmek üzere kullanılır.

- Sahtekârlık (Fraud) Tespiti: Sağlık, sigorta ve hükümet tarafında toplanan büyük çaptaki metin verilerinde kalıplar ve anormallikler aranarak sahtekârlıklar tespit edilir.
- Bilimsel ve Medikal Araştırmalar: Hasta raporları, makale başlıkları, yayınlanmış araştırma sonuçları ve diğer yayınlar gibi metin materyallerinden çıkarım yapılır.
- Güvenlik/İstihbarat: Organizasyonlar ve bireyler arasındaki kalıplar ve bağlantılar, terörist tehlikeleri ve kriminal davranışları tahmin etmek ve engelleyebilmek için büyük çaptaki metin içerisinde aranır.
- Pazar Araştırması: Yayınlanmış belgeler, basın bültenleri ve web sayfaları pazar etkisinin ölçülmesi için aranır ve izlenir. Metin madenciliği kantitatif yöntemler ile açık uçlu anket soruları ve mülakatların değerlendirilmesinde kullanılabilmektedir.

2.3. Metin Madenciliği Adımları

Metin madenciliği temelde beş adımdan oluşmaktadır. Şekil 2.5' te metin madenciliğinin adımları gösterilmektedir.



Şekil 2.5. Metin madenciliği adımları

2.3.1. Metin Bilgilerini Toplama

Metin bilgileri toplama, seçilen konularda bilgiye erişim sistemleri kullanılarak metin koleksiyonu oluşturma sürecidir. Bu süreç, günümüzde genel olarak internet üzerinden, özellikle Google vb. arama motorları kullanılarak gerçekleştirilmektedir. Çevrim içi veritabanlarının yanı sıra veritabanlarında ya da kişisel bilgisayarlarda bulunan metin türü veriler ile oluşturulan koleksiyonlar da metin madenciliğinde kullanılmaktadır [15,16].

2.3.2. Metin Topluluğu Üzerinde Ön İşlem Uygulama

Ön işlem uygulama, metni kelimelere ayırma, kelimelerin anlamsal değerlerini bulma (isim, sıfat, fiil, zarf, zamir vb.), kelimeleri köklerine ayırma ve gereksiz kelimeleri ayıklama, yazım kurallarına uygunluğunu tespit etmek ve var olan hataları düzeltmek gibi metin belgelerin yapıtaşı olan kelimelerle ilgili işlemleri içeren süreçtir [17].

Metin madenciliğinin en büyük sorunu işleyeceği veri kümesinin yapısal olmamasıdır. Genellikle doğal dil kullanılarak yazılmış dokümanlar üzerinde çalışılan metin madenciliği alanında ön işleme aşaması veri temizlemenin yanında veriyi uygun formata getirme işlemini de gerçekleştirmektedir [11].

2.3.3. Metin Üzerinde Özellik Seçme

Metin madenciliği uygulamalarında her zaman gürültülü ve önemsiz bilgi içeren metin koleksiyonlarıyla uğraşılma ihtiyacı bulunmaktadır. İlgili verilerin saptanması üzerine odaklanan özellik seçme, büyük miktarlardaki veriler üzerinde işlem yapılırken iş yükünü azaltmada yardımcı olmaktadır [18]. Özellik seçme aşamasında, ön işlemden geçen metinlerdeki önemli kelimeleri (varlıkları) belirleme (isimler, tamlamalar, bileşik kelimeler, kısaltmalar, sayılar, tarihler, para birimleri

vb.) ve ilişkili olmayan özelliklerin çıkarılması (sadece birkaç dokümanda gözlemlenen özelliklerin çıkarılması, birçok dokümanda gözlemlenen özellikleri azaltma vb.) işlemleri yapılmaktadır [10].

2.3.4. Metinlerin Kategorilere Ayrılması

Bu aşama, analiz süreci olarak da adlandırılabilir. Veri madenciliği yöntemleri uygulanarak veri gruplara ayrılır. Bu aşamada genellikle istatistik, yapay zekâ ve genetik algoritmaları gibi yapılar kullanılmaktadır. Değerlendirme sürecine bağlı olarak bu aşamada kullanılacak yöntem değişiklik gösterir [19].

2.3.5. Değerlendirme ve Yorumlama

Veri bu aşamada sunuma hazır duruma gelmiştir. Kullanıcının anlayacağı şekilde grafik, tablo veya şekil olarak bilgi sunumu yapılır. Veri bu aşamada yapısal veriye dönüştüğü için analiz işlemi kolaylıkla yapılabilir durumdadır. Verinin dönüşümü basit bir örnekle açıklanacak olursa, bir kavanozda farklı renk ve boylarda misketler olduğu varsayıldığında, misketlerin renk ve boylarına göre gruplara ayrılmış hali, verinin artık kullanıcı tarafından değerlendirme ve yorumlama yapabilir seviyeye gelmiş halidir.

2.4. Metin Sınıflandırma

Çok sayıda bilgi varlığının getirdiği sayısız fayda ile beraber ortaya çıkan bazı sorunların da çözülmesi gerekmektedir. Bilgiye erişmenin (aranılan bilginin bulunabilirliğinin) kolay olması gereklidir. Bu bağlamda ortaya çıkan sorunlardan bir tanesi de elektronik ortamdaki metinlerin sınıflandırılması sorunudur. Metin sınıflandırma sorunu, en genel anlamı ile eldeki bir metnin önceden belirlenen

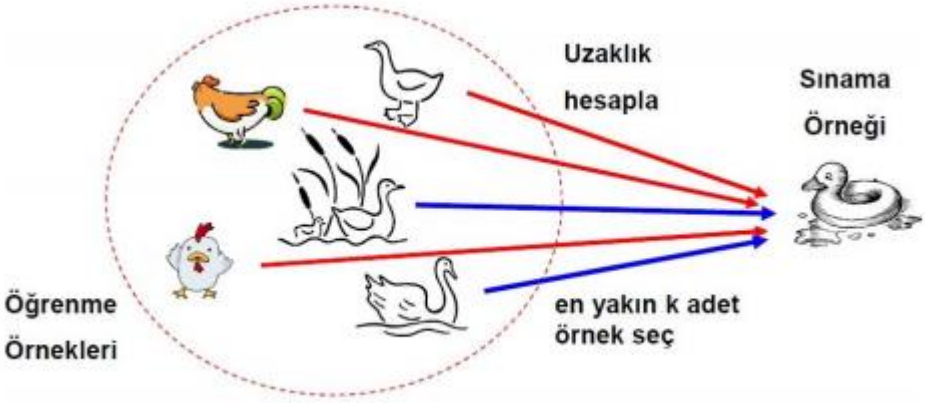
sınıflardan hangisine ya da hangilerine girdiğinin belirlenmesi demektir [20]. Metin sınıflandırma işlemi çeşitli algoritmalar kullanılarak yapılır.

2.4.1. Metin Sınıflandırma Algoritmaları

Sınıflandırma özetle, bir veri kümesi üzerinde tanımlı olan sınıflara veriyi dağıtma işlemidir. Sınıflandırma algoritmaları, verilen eğitim kümesinden bu dağılım şeklini öğrenirler ve daha sonra sınıfının belirli olmadığı test verileri geldiğinde doğru şekilde sınıflandırmaya çalışırlar [11]. Bu işlemler için çeşitli algoritmalar kullanılmaktadır.

2.4.1.1. KNN Algoritması (K-Nearest Neighborhood)

K-NN olarak ifade edilen K-Nearest Neighbor (K-En Yakın Komşu) sınıflandırma algoritmasının temelinde "birbirine yakın olan nesneler muhtemelen aynı kategoriye aittir" mantığı yatar. Algoritmanın amacı, yeni bir nesneyi özelliklerinden faydalanarak önceden sınıflandırılmış örnekler yardımıyla sınıflandırmaktır [21]. Hangi sınıfa ait olduğu bilinmeyen nesne sınama örneği, önceden sınıflandırılmış nesneler ise öğrenme örnekleri olarak adlandırılır. K-NN algoritmasında sınama örneğinin öğrenme örneklerine olan uzaklıkları hesaplanır ve en yakınındaki k adet örnek çoğunlukla hangi sınıfa aitse sınama örneğinin de o sınıfa ait olduğu düşünülür [22]. Şekil 2.6.'da KNN sınıflandırma algoritması görülmektedir. Şekil 2.6.'da, örnekte gelen sınama örneğinin hangi sınıfa ait olduğu bulunmaya çalışılmaktadır. Şekilde okların renkleri üzerinden bakıldığında iki adet sınıflama yapıldığı görülmektedir, k sayısı 2 olarak seçilmiştir. Her bir şekle olan uzaklık hesaplanarak en yakın olduğu şeklin sınıfına dahil kabul edilir.



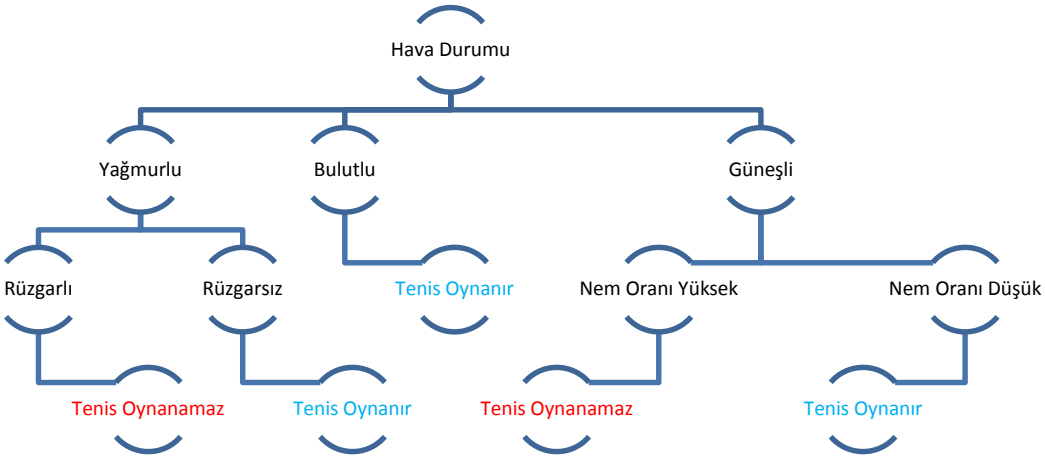
Şekil 2.6. K-NN sınıflandırma algoritması [23]

2.4.1.2. Karar Ağacı (Decision Tree)

Belirli bir soruna çözüm bulmak amacıyla elde bulunan verilerin belirli bir kurala göre ağaç dalları şeklinde haritalanması ve bu ağaç yapısına göre tahmini değere ulaşılmasını sağlayan yapıdır.

Metin sınıflandırmada kullanılan karar ağaçlarının, iç düğümleri terimleri, yaprak düğümleri (en alttaki düğümler) ise sınıfları göstermektedir. Her iç düğümden, yani terimden, çıkan iki dallanma bulunmaktadır. Bu dallanmalardan bir tanesi o terimin ilgili belgede bulunduğu durumu, diğeri de bulunmadığı durumu göstermektedir. Şekil 2.7.'deki sınıflandırma kurallarına ilişkin karar ağacı görülmektedir.

Bir belge sınıflandırılırken ilgili terimlerin o belgede yer alıp almamasına göre bu ağaç üzerinde gezilir ve sonuçta ulaşılan yaprak düğümün etiketine göre sınıflandırma yapılır [20]. Örnekteki ağaç, hava durumuna göre tenis oynama durumunu gösterir. Hava durumu, olasılıklara göre üç gruba ayrılır. Yine her hava durumu, olasılıklarına göre kırılımlara ayrılır.



Şekil 2.7. Karar ağacı örneği

2.4.1.3. Naïve Bayes

Naïve Bayes Sınıflandırıcı adını İngiliz matematikçi Thomas Bayes'ten (yak. 1701 - 7 Nisan 1761) alır. Naïve Bayes Sınıflandırıcı Örüntü tanıma problemine ilk bakışta oldukça kısıtlayıcı görülen bir önerme ile kullanılabilen olasılıkçı bir yaklaşımdır. Bu önerme, örüntü tanıma da kullanılacak her bir tanımlayıcı nitelik ya da parametrenin istatistik açıdan bağımsız olması gerekliliğidir. Her ne kadar bu önerme Naïve Bayes Sınıflandırıcısının kullanım alanını kısıtlandırsa da, genelde istatistik bağımsızlık koşulu esnetilerek kullanıldığında da daha karmaşık yapay sinir ağları gibi metotlarla karşılaştırabilir sonuçlar vermektedir [24].

Her özellik için sınıflar içinde bulunma olasılıkları ve sınıfların veri üzerinde görülme olasılıklarını hesaplayarak karar veren bir modeldir. “koşullu bağımsızlık kabulü” ile bir özelliğin bir sınıfta belirli bir olasılıkla geçmesi, bir başka özelliğin aynı sınıfta geçiş olasılığından etkilenmez ve o olasılığı etkilemez [19]. Bu özelliğinden dolayı metod Naïve adını alır.

Tez için geliştirilen uygulamada kullanılan Naïve bayes sınıflandırmasında aşağıdaki denklemler kullanılır.

$$p(d|c_j) = p(|d|)|d|!\prod_{t=1}^{|V|}\frac{p(w_t|c_j)^{xt}}{x_t!}$$ Denklem 1.

$$p(w_t|c_j) = \frac{1+N_{jt}}{|V|+N_J}$$ Denklem 2.

$$M(C) = P(word1|C)^{n1}P(word2|C)^{n2}...P(wordv|C)^{nv}P(|C)$$ Denklem 3.

Yukarıdaki denklemlerdeki d kategori sayısını, x_t kelimenin sıklığını, |V| kelime sayısını, N_{jt} j sınıfındaki dokümanlarda t kelimesinin görülme sıklığını, N_j j sınıfındaki toplam kelime sayısını, P (|d|) kategori olasılığını, x_t kelimenin sıklığını, w_t işlem gören kelimeyi, c_j sınıfı ifade eder. Denklem 3.'te bulunan formül uygulandığında çıkan M(C) değeri en büyük olan sınıfa atama yapılır [16].

Bu sınıflandırma işleminde veri kaynağı üzerinde mutlaka bir sınıflandırma-kategori tanımlamasının bulunması gerekir. Test edilecek veri, öğretilmiş veri seti üzerindeki olasılık değerlerine göre hesaplanır. Bu oranlamaya göre, test setinin hangi kategoriye daha yakın olduğu bulunur. Öğretilmiş veri sayısı arttıkça, test verisinin bulunduğu kategoriyi saptamak kolaylaşır.

Naïve Bayes kullanım ile ilgili basit bir örnek aşağıdaki Çizelge 1. de verilmiştir. Bu denklemin kullanılmış olmasının nedeni, bu tür uygulamalarda başarı oranının yüksek olmasından kaynaklanmaktadır.

Gıda ürünlerinin bulunduğu bir veri seti olduğunu varsayalım. Set içeriği;

Çizelge 1. Veri seti dağılımı

Grup No	**Gruba ait veri**	**Kategori**
1	Et, Su, Yumurta, Yoğurt	P
2	Et, Et, Su	P
3	Ekmek, Su, Et	K
4	Yoğurt, Su, Et	P

P(K) = K kategorisinin veri setindeki adedi / tüm setteki kayıt sayısı = 1/4 = 0.25
P(P) = P kategorisinin veri setindeki adedi / tüm setteki kayıt sayısı = 3/4 = 0.75

Bu işlemden sonra, kelimelerin bağlı olduğu gruptaki ağırlıkları hesaplanır. Bunun için öncelikle bir liste oluşturulur.

P kategorisi için;

x kelimesinin ağırlığı = (P kategorisindeki x kelimesinin sayısı + 1) / (P kategorisindeki toplam kelime adedi + tüm setteki tekrarlanmayan veri sayısı) Buna göre bu kategorideki her kelime için hesaplama yaptığımızda;

Et = (4 + 1) / (10 + 6) = 0.3125
Su = (3 + 1) / (10 + 6) = 0.25
Yumurta = (1 + 1) / (10 + 6) = 0.125
Yoğurt = (2 + 1) / (10 + 6) = 0.1875

K kategorisi için hesaplama yaptığımızda;

Et = (1 + 1) / (3 + 6) = 0.2222
Su = (1 + 1) / (3 + 6) = 0.2222
Ekmek = (1 + 1) / (3 + 6) = 0.222

Bu işlemlerden sonra veri seti oluşmuş oldu. Yeni bir test veri setini ele aldığımızda;

Et, Et, Yoğurt, Zeytin, Su Ait olduğu veri seti hangisi?

Yapılacak hesaplamada, test veri setindeki her kelimenin, test edildiği gruptaki oranı birbiri ile çapılır. İki grupta da bulunmayan veri işleme dâhil edilmez. Buna göre;

P grubu oranı = 0.75 * 0.3125 * 0.3125 * 0.25 = 0.18310547
K grubu oranı = 0.25 * 0.2222 * 0.2222 * 0.2222 = 0.002742661

Yoğurt iki grupta birden bulunduğundan işleme dâhil edilmemiştir. Zeytin, iki grupta da bulunmadığından işleme dâhil edilmemiştir. Buna göre sonuçları yüzdelik değere çevrildiğinde;

P grubuna dâhil olma oranı = (0.18310547 * 100) / (0.18310547 + 0.002742661)
= % 98,524 olarak bulunur.

2.4.1.4. Destekçi Vektör Makineleri (Support Vector Machine - SVM)

SVM, sınıflandırma (classification) konusunda kullanılan oldukça etkili ve basit yöntemlerden birisidir. Sınıflandırma için bir düzlemde bulunan iki grup arasında bir sınır çizilerek iki grubu ayırmak mümkündür. Bu sınırın çizileceği yer ise iki grubun da üyelerine en uzak olan yer olmalıdır. İşte SVM bu sınırın nasıl çizileceğini belirler [25]. Bu yöntem nesne tanıma, el yazısı tanıma, zaman serisi tahmin testleri gibi uygulamalarda kullanılır.

3. METİN MADENCİLİĞİ TABANLI BİR WEB SİTESİ SINIFLANDIRMA ARACI TASARIMI

Metin madenciliğinde kullanılan sınıflandırma tekniği ile web siteleri üzerinde sınıflandırma işlemi yapılabilir. Bir kategori seçimi yapılarak, ilgili sitenin o kategoriye dâhil olup olmadığı saptanabilir.

3.1. Giriş

İnternet alanındaki gelişmeler ve internet kullanımındaki artış firmaları, ürünlerini internet üzerinden daha geniş kitlelere sunma imkanı sağlamıştır. Bu durum, aynı zamanda tüketiciler açısından da çok çeşitli avantajlar sağlar. Ürünü kendi bulunduğu çevredeki mağazalardan temin edemeyen bir tüketici, internet üzerinden bu ürüne ulaşıp temin edebilir. Mağazalarda daha yüksek fiyata sahip olan ürünü, internetten daha çok mağaza üzerinde arama yaparak daha uygun fiyata bulabilir. Aradığı ürünü, başka ürünlerle kıyaslama, ürünü deneyen kullanıcıların yorumlarına ulaşabilir.

Ürün arama işlemi, sürekli alışveriş yapılan, bilindik bir site yoksa yeni adres bulabilmek için genellikle Google arama motoru üzerinden yapılır. Ancak bu aramalarda, ürünün satışını yapan e-ticaret sitesi yerine, listeye farklı siteler de gelebilir.

Bu tez çalışmasında bir web sitesi sınıflandırma aracı tasarımı yapılmaktadır. Bu yazılım aracının sayesinde, arama sonucu listesinde yer alan sitelerin e-ticaret sitesi olma oranlarını bularak kullanıcıyı e-ticaret ile ilgili gerçek siteleri göstermektir.

3.2. Makine Öğrenmesi İçin Veri Toplama

Veri toplama öncesinde, metin sınıflandırması için kullanılacak yöntem seçilir. Bunun üzerinden kategorilere ayrılarak veri toplamı işlemi yapılır.

Bu tez çalışmasında Naïve Bayes sınıflandırma yöntemi seçildi. Bu yöntemin seçilmesinin nedeni, bu tür uygulamalarda ortaya çıkan başarı yüzdesinin yüksek olmasıdır. Sınıflandırmada kullanmak üzere iki grup belirlendi. Birince grupta e-ticaret sitesi olan adreslere ait veriler, ikinci grupta e-ticaret sitesi olmayan sitelere ait veriler toplandı.

Visual Studio. Net platformu üzerinde bir Windows Forms uygulaması geliştirildi. Uygulama geliştirme için Visual Basic .NET programlama dili kullanıldı. Veritabanı olarak Microsoft SQL Server kullanıldı.

İlk olarak internet sitelerinden toplanan kelimelerin veritabanı üzerinde yazılacağı bir tablo oluşturuldu. Şekil 3.1.'de Datamin isimli tablonun yapısı görülmektedir.

Column Name	Data Type	Allow Nulls
id_no	decimal(15, 0)	☐
kelime	varchar(50)	☑
ana_cumle	varchar(250)	☑
site_adres	varchar(500)	☐
kategori	varchar(1)	☑
olasilik	decimal(15, 13)	☑
		☐

Şekil 3.1. Datamin tablo yapısı

Gerçekte e-ticaret sitesi kategorisine dâhil olan on bir adet site seçildi. Sitelerin listesi Çizelge 2.'de verilmiştir. Listede bulunan siteler, e-ticaret konusunda Türkiye'de popüler olan siteler arasından rastgele seçildi.

http://www.teknosa.com
http://www.sahibinden.com
http://www.mediamarkt.com.tr
http://www.vatanbilgisayar.com
http://www.bimeks.com.tr
http://www.hepsiburada.com
http://www.ykm.com.tr
http://www.boyner.com.tr
http://www.e-bebek.com
http://www.gold.com.tr
http://www.istanbulbilisim.com.tr

Çizelge 2. E-ticaret sitesi olan sitelerin listesi

E-ticaret kategorisine dâhil olmayan siteler için de dokuz adet örnek site seçilmiştir. Bu siteler Çizelge 3.'te görülmektedir. Site seçimi yapılırken Google arama motoruna e-ticaret yazılıp arama yapılmış ve çıkan listedeki bağlantılardan rastgele aşağıdaki listedeki bağlantılar seçilmiştir.

http://www.noramedya.com
http://www.proticaret.org
http://www.ticimax.com
http://www.eticaretkur.com
http://www.akinsofteticaret.com
http://www.ideasoft.com.tr
http://www.platinmarket.com
http://www.tsoft.com.tr
http://www.neticaret.com.tr

Çizelge 3. E-ticaret olmayan sitelerin listesi

Datamin isimli tablodaki kategori alanına göre veriler iki gruba göre sınıflandırılır. İnternet sitesi adresleri belirlendikten sonra geliştirilen uygulama ile sitelerin açılış sayfalarındaki html uzantılı dosyaların kaynak içerikleri kelimelere ayrılarak, veritabanındaki datamin isimli tabloya yazılır.

Kelimeler için bazı filtrelemeler uygulanır. Örneğin, kelime sayısal olamaz, sayı değeri içeremez, ilk karakteri http, ?, -, + ile başlayamaz, www, html, %, #, (), amp;, script, (,) gibi karakterleri içeremez, uzunluğu 3'ten küçük ve 26'dan büyük olamaz. Bu sınırlandırmaları aşan kelimeler datamin tablosuna yazılır.

Bu çalışma ile birlikte, sitelerde geçen kelimeler ve kelime adetleri datamin tablosunda toplanır. Aşağıdaki Şekil 3.2.'te beyaz kelimesinin sitelere göre dağılımı listelenmektedir.

Results | Messages

	id_no	kelime	ana_cumle	site_adres	kategori
19	11406	beyaz	Mobee Nett S900 E 7inc 16GB Beyaz Tablet PC	http://www.gold.com.tr	E
20	11410	beyaz	Mobee Nett S900 E 7inc 16GB Beyaz Tablet PC	http://www.gold.com.tr	E
21	11432	beyaz	dan cep telefonuna fotoğraf makinesinden be...	http://www.gold.com.tr	E
22	11508	beyaz	Beyaz Eşya	http://www.gold.com.tr	E
23	11510	beyaz	Beyaz Eşya	http://www.gold.com.tr	E
24	11846	beyaz	Beyaz Eşya	http://www.teknosa.com	E
25	11848	beyaz	Beyaz Eşya Setleri	http://www.teknosa.com	E
26	11851	beyaz	Beyaz Eşya Setleri	http://www.teknosa.com	E
27	12255	beyaz	SAMSUNG I9301Q GALAXY S3 NEO BEYAZ...	http://www.teknosa.com	E
28	12261	beyaz	SAMSUNG I9301Q GALAXY S3 NEO BEYAZ...	http://www.teknosa.com	E
29	12286	beyaz	Beyaz Eşyada Kargo Bedava	http://www.teknosa.com	E
30	12290	beyaz	Beyaz Eşyada Kargo Bedava	http://www.teknosa.com	E
31	12635	beyaz	Parrot Rolling Spider Beyaz	http://www.teknosa.com	E
32	12639	beyaz	Parrot Rolling Spider Beyaz	http://www.teknosa.com	E
33	12668	beyaz	TO WATCH T 1000W BLUETOOTH AKILLI ...	http://www.teknosa.com	E
34	12673	beyaz	TO WATCH T 1000W BLUETOOTH AKILLI ...	http://www.teknosa.com	E
35	13903	beyaz	SAMSUNG N9000Q GALA...	http://www.vatanbilgisayar.c...	E
36	13909	beyaz	SAMSUNG N9000Q GALA...	http://www.vatanbilgisayar.c...	E
37	13980	beyaz	SAMSUNG GALAXY S5 G...	http://www.vatanbilgisayar.c...	E
38	13985	beyaz	SAMSUNG GALAXY S5 G...	http://www.vatanbilgisayar.c...	E

Şekil 3.2. Örnek kelime listesi

3.3. Kategorilere Göre Veri Havuzunun Analiz Edilmesi

Kategorilere göre havuz oluşturmak için seçilen sitelere ait kelimeler datamin tablosunda toplanır. Bu aşamadan sonra yapılacak işlem, her kelimenin bulunduğu kategorideki olası ağırlık oranının hesaplanmasıdır. Bu hesaplama için "indirim" kelimesi örnek kelime olarak kullanılacaktır. Hesaplama aşağıdaki şekilde yapılır;

Denklem için kullanılacak kısaltmalar **x, y, z** olarak tanımlanmış olsun.

x: Eğitilmiş veri setinde bulunan kelime sayısını ifade eder. Yalnız buradaki sayıda, kelime birden fazla sayıda geçiyorsa biri hesaplanır. Oluşturulan veritabanında bu sayı 4934 olarak çıkmaktadır.

y: E-ticaret kategorisindeki kelime adedini gösterir. Uygulama içinden elde edilen ve kaydedilen bu sayı 17801 adettir.

z: Bu oran hesaplanacak kelimenin e-ticaret kategorisinde kaç adet geçtiğini ifade eder. ‘indirim’ kelimesi bu kategoride 88 adettir.

$$\text{Oran} = \frac{(\text{kelimenin kategorideki sayısı} + 1)}{(\text{öğretilmiş kelime sayısı} + \text{kategorideki toplam kelime adedi})} \qquad \text{Denklem 4.}$$

Oran = (z + 1) / (x + y)
Oran = (88 + 1) / (4934 + 17801)
Oran = 0,0039146690125 olarak bulunur.

Bu oran, ‘indirim’ kelimesinin e-ticaret sitesi kategorisi içindeki oranıdır. Aynı işlem, e-ticaret sitesi olmayan gruptaki ‘indirim’ kelimesi için de yapılır. Buradaki sonuçlarda;

x : 4934
y : 10664
z : 12 olarak bulunur. Oran için formül hesaplandığında;
Oran = (12 + 1) / (4934 + 10664)
Oran = 0,0008334401846 olarak hesaplanır.

Bu sonuçlara göre ‘indirim’ kelimesinin e-ticaret sitesindeki oranı 0,0039146690125, e-ticaret olmayan sitedeki oranı ise 0,0008334401846 olarak hesaplanır. Çıkan bu orandaki değerler, çok düşük değerler olduğu için, veritabanına kaydedilirken 1000 ile çarpılır. Bu kelimenin veritabanındaki örnek görüntüsü Şekil 3.3.’te verilmiştir.

	id_no	kelime	ana_cumle	site_adres	kategori	olasilik
1	34252	indirim	%50 indirim kampanyasını kaçırmayın	http://www.platinmarket.com/	H	0.8334401846000
2	9053	indirim	Baby Me Yenidoğan ıslak mendillerinde indirim	http://www.e-bebek.com/	E	3.9146690125000
3	9058	indirim	Baby Me Yenidoğan ıslak mendillerinde indirim	http://www.e-bebek.com/	E	3.9146690125000
4	33071	indirim	Bayiye ouml;zel indirim uygulama	http://www.noramedya.com/	H	0.8334401846000
5	33075	indirim	Bayiye ouml;zel indirim uygulama	http://www.noramedya.com/	H	0.8334401846000
6	594180	indirim	Çoklu İndirim Modülü Kullanımlarında Simge Eklenmiştir	http://www.tsoft.com.tr/	H	0.8334401846000
7	3885	indirim	Hem boyner com tr'de hem de Boyner Mağazaları'nd...	http://www.boyner.com.tr/	E	3.9146690125000
8	3907	indirim	Hem boyner com tr'de hem de Boyner Mağazaları'nd...	http://www.boyner.com.tr/	E	3.9146690125000
9	772	indirim	Hem ykm com tr'de hem de Ykm Mağazaları'nda size...	http://www.ykm.com.tr/	E	3.9146690125000
10	794	indirim	Hem ykm com tr'de hem de Ykm Mağazaları'nda size...	http://www.ykm.com.tr/	E	3.9146690125000
11	10270	indirim	indirim	http://www.sahibinden.com/	E	3.9146690125000

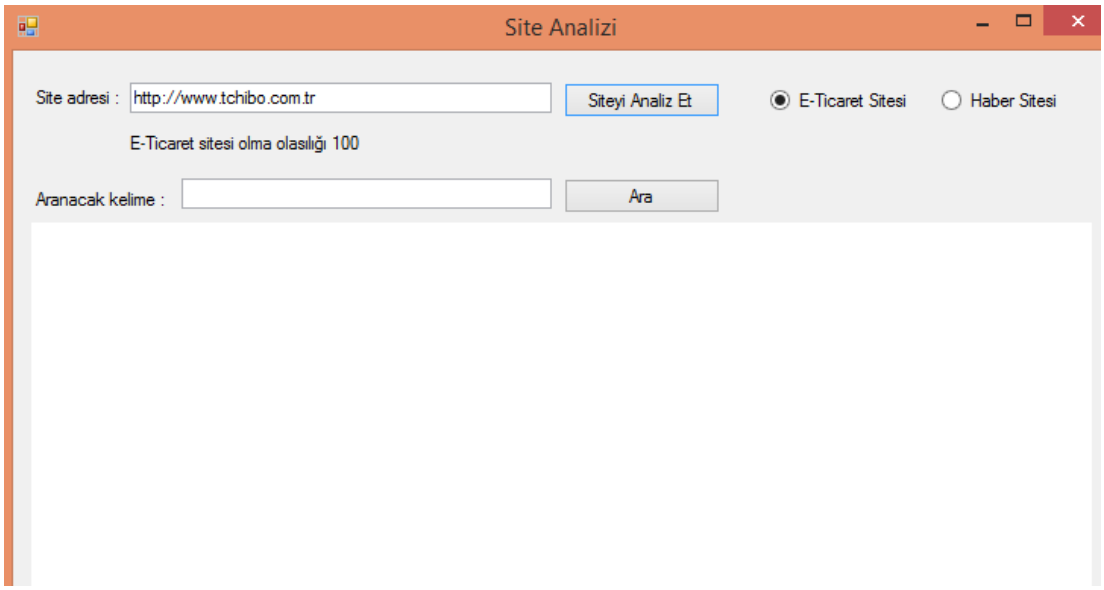
Şekil 3.3. İndirim kelimesinin kategorilerdeki oranları

Bu şekilde yer alan tüm kelimeler için oran hesaplaması yapılır. İşlemlerin sonucunda makine öğrenmesi için gerekli bir veri havuzu elde edilmiş olur.

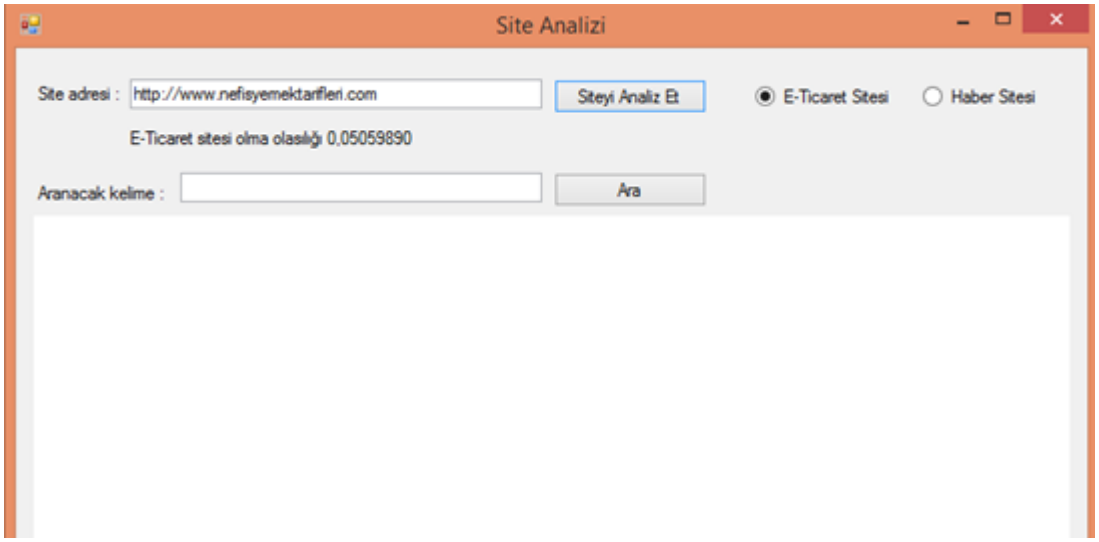
3.4. Arama Sonuçlarındaki Bağlantıların Analizi

Makine öğrenmesi gerçekleştikten sonra, bu veriler üzerinden olasılık hesaplaması yapılarak, bulunan sitenin e-ticaret sitesi olma oranı hesaplanabilir.

Geliştirilen uygulamada, istenirse direkt olarak site adresi girilip, 'Bu site hangi oranda e-ticaret sitesi olma özelliği taşıyor?' sorusunun cevabı bulunabilir. Kelime araması yapılarak, listelenen sonuçlardaki her bir site yine analiz işlemi uygulanarak oran hesaplaması yapılabilir. Şekil 3.4.'te uygulamanın ekran görüntüsü yer almaktadır. Bu ekranda Site Adresi alanına www.tchibo.com.tr adresi girilmektedir ve bu adresin bir e-ticaret sitesi olma olasılığı belirlenmektedir. Site adresi girilerek, siteyi analiz et butonuna basılır. Şekil 3.5.'de ise başka bir site için olasılık hesaplaması görüntülenmektedir.



Şekil 3.4. Site Analizi-1



Şekil 3.5. Site Analizi-2

Kelime arama işleminde, arama Google arama motoru üzerinden yapılır. Google'daki ilk 20 arama sonucu site bağlantı adresleri dikkate alınır. Bu sitelerin her biri için olasılık hesaplanır. Tüm hesaplamalar Naïve Bayes sınıflandırma formülleri üzerinden gerçekleştirilir.

Hesaplamaları yapabilmek için öncelikle özet bir veritabanı tablosu oluşturuldu. Şekil 3.6.'de bu tablonun yapısı görülmektedir. Veri havuzuna ait özet veriler bu tabloya SQL üzerinden eklenmektedir.

Column Name	Data Type	Allow Nulls
kelime	varchar(50)	☑
kelime_sayisi	int	☑
kategori	varchar(1)	☑
olasilik	decimal(15, 13)	☑
		☐

Şekil 3.6. Datamin_ozet tablosunun yapısı

Datamin_ozet tablosundaki kelimeler, kategorilere göre (e-ticaret olma-olmama) gruplanarak tutulmakta, kelime sayısı ise o kategoride ilgili kelimenin kaç adet olduğunu ifade etmektedir. Bu özet tablo, hesaplamadaki sorgularda sorgu süresini kısaltmak amacıyla kullanılmaktadır. Şekil 3.7.'de özet kelimeler tablosu görülmektedir.

	kelime	kelime_sayisi	kategori	olasilik
273	tasarlanmıştır	2	H	0.1923323503000
274	aksesuar	51	E	2.2872223444000
275	imalat	2	E	0.1319551353000
276	ıso	2	E	0.1319551353000
277	gezdiğiniz	2	E	0.1319551353000
278	kibo	2	H	0.1923323503000
279	batman	1	H	0.1282215669000
280	belirli	2	E	0.1319551353000
281	otomatik	12	H	0.8334401846000
282	dahili	2	H	0.1923323503000
283	googlestore	2	H	0.1923323503000

Şekil 3.7. Özet kelime tablosu

Google arama sonuçlarında listelenen veya site ismi girilerek yapılan hesaplamalarda, veri havuzu oluşturulurken kullanılan yöntem uygulanır. Arama sonuçlarında gelen kod kaynağı üzerinde bağlantı araştırması yapılır. Bağlantı olan

adresler tespit edilerek arama sonucunda çıkan internet sitesi adreslerine ulaşılır. Bu adresler üzerinden tek tek hesaplama işlemi gerçekleştirilir.

Öncelikle internet sitesinin html kaynak kod içeriğindeki kelimeler, havuz oluşturulurken kullanılan metot ve filtreleme işlemlerinden geçerek, veritabanındaki datamin isimli tabloya eklenir. Bu ekleme işleminde, sitenin kategorisi T olarak belirlenir. Bu kategori ayırımı bize bu verinin test verisi olduğunu ifade eder. Kategorisi T olan bu test verisinde kelimelerle ilgili olasılık hesabı yapılmaz. Havuzdaki hesaplanmış olasılık değerleri üzerinden işlem yapılır. T kategorisinde bulunan kelimelere filtreleme işlemi uygulanır. Bu işlemde, ilgili kelimelerin e-ticaret olan ve e-ticaret olmayan kategorilerinin ikisinde de bulunması gerekir. Örneğin, test verisi olarak http://www.nefisyemektarifleri.com/ sitesi örnek alındığında, siteden toplamda 1394 adet veri gelmiş. Bu kelimeler arasında, her iki grupta da bulunan kelime sayısı 124 olarak çıkmaktadır. Hesaplamada bir döngü oluşturulur. Bu döngü, test edilen sitedeki kelimelerden her iki grupta da bulunan kelime adedi kadardır. Bu sayı 124 olarak çıkmıştı. Her döngüde, iki adet oran hesabı yapılır. Oran1 kelimenin e-ticaret sitesindeki oranı, Oran2 kelimenin e-ticaret olmayan sitedeki oranını ifade eder. Oran1 ve Oran2 döngüye girmeden önce 1 değeri ile başlatılır. Döngüde, ilgili kelimenin iki gruptaki oran karşılığı (datamin isimli tablodan çekilen olasılık alanı) ait olduğu kategorinin oranı ile çarpılır.

Örneğin, test edilen sitedeki ilk kelime olan ‘ana’ kelimesi ile ilgili yapılan hesaplamada, Oran1 ve Oran2’nin ilk değerleri 1 olarak başlanır. ‘ana’ kelimesi test edilen sitede iki kez geçmektedir. Dolayısıyla yapılacak 124 adet döngüden 2’si bu kelime için yapılır. E-ticaret sitesindeki bu kelimenin oransal değeri 0.5718055861 olarak bulunur. E-ticaret olmayan sitedeki oran ise 0.5128862675’tir. Hesaplamadaki ilk döngüde değer;

Oran1 = 1 * 0.5718055861 = 0.5718055861
Oran2 = 1 * 0.5128862675 = 0.5128862675

olarak bulunur. İkinci döngüye gelindiğinde;

Oran1 = 0.5718055861 * 0.5718055861 = 0.326961628
Oran2 = 0.5128862675 * 0.5128862675 = 0.263052323

olarak hesaplanır. Döngüdeki bir sonraki kelime ‘çıkan’ kelimesi olarak bulunur. Bu kelimenin her iki gruptaki ağırlığı bulunur ve Oran1, Oran2 değerleri ile çarpılır. Bu değerler aşağıdaki gibidir:

Oran1 = 0.326961628 * 0.1319551353 (kelimenin bu gruptaki ağırlığı) = 0.043144266
Oran2 = 0.263052323 * 0.448775484 (kelimenin bu gruptaki ağırlığı) = 0.118051434

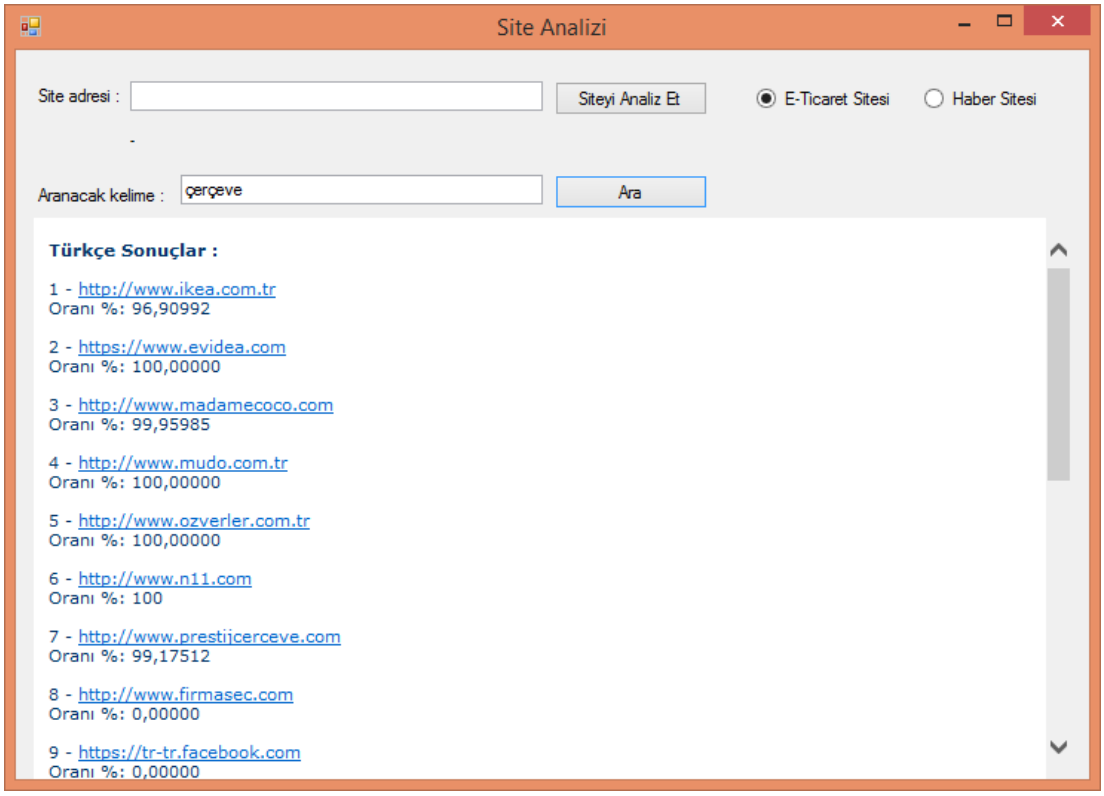
olarak bulunur. Döngünün ilerleyen bölümlerinde değerler sıfıra yaklaştığı için, döngü içinde Oran1’in değeri kontrol edilerek 0.00001’in altına düşmesi durumunda Oran1 ve Oran2 alanlarına 1000 ile çarpım uygulanır.

Bu şekilde hesaplama 124 adet kelime üzerinden yapıldığında, elde edilen sonuçta;

Oran1 = 0.00159133751653225462771907 61
Oran2 = 2100.92811353253382740842620 67

olarak bulunur. Bu oranlar yüzde hesabına çevrildiğinde bu sitenin e-ticaret sitesi olma olasılığı %0.00007574444365480016190373 73 olarak bulunur.

Google üzerinden, kelime yazarak yapılan aramada ise arama sonuçlarındaki bağlantı adresleri bulunarak, yukarıda örneği belirtilen şekilde her adres için hesaplama işlemi yapılır. Sonuçlar, uygulamanın ara yüzünde bulunan bir tarayıcı üzerinden oranlarını gösterir biçimde listelenir. Kullanıcı isterse ilgili bağlantıya tıklayarak site adresine erişebilir. Şekil 3.8.’de kelime olarak ‘çerçeve’ aranmış, arama sonucunda çıkan site adresleri analiz edilmiş ve sonuçları kullanıcıya listelenmiş haliyle görüntülenmektedir.



Şekil 3.8. Kelime arama sonuçları

Örnek Uygulama 2:

Uygulamadaki benzer aramalar, aranan kelimenin ait olduğu sitenin haber sitesi olup olmadığını bulmak için de yapılabilir. Bu işlem için benzer şekilde bir havuz oluşturulur. Haber sitesi olan ve olmayan birkaç sitenin içindeki kelimeler toplanır, ağırlık oranları hesaplanır. Bu çalışmada aşağıdaki listede bulunan haber sitelerinin içerikleri alındı. Çizelge 4.'de bu sitelerin listesi yer almaktadır. Listede bulunan bağlantılar, popüler olan internet haber sitelerinin içinden rastgele seçilmiştir.

http://www.ensonhaber.com
http://www.fanatik.com.tr
http://www.haber3.com
http://www.haberturk.com
http://www.haberx.com
http://www.hurriyet.com.tr/anasayfa
http://www.internethaber.com
http://www.milliyet.com.tr
http://www.posta.com.tr
http://www.radikal.com.tr
http://www.sondakika.com.tr
http://www.sozcu.com.tr

Çizelge 4. Haber siteleri havuzu

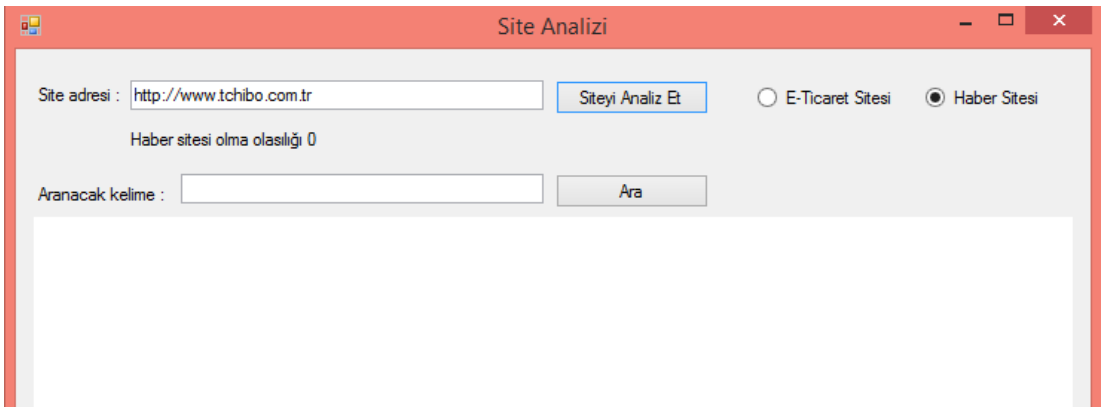
Haber sitesi olmayan birkaç site de bu havuza aktarıldı. Bu kategoride kullanılan sitelerin listesi aşağıdaki Çizelge 5.'de gösterilmiştir. Listedeki internet sitesi adresleri Google arama motoru üzerinden haber sitesi kelimesi aratılarak çıkan sonuçlar içinden rastgele seçim yapılarak elde edildi.

http://turkiyemix.com
http://www.doviz.com
http://www.eticaretuzmani.com.tr
http://www.habersitesikur.com
http://www.habersitesikurulumu.com
http://www.hepsidijital.com
http://www.istanbulsaglik.gov.tr
http://www.mgm.gov.tr
http://www.tumeva.com

Çizelge 5. Haber sitesi olmayan siteler

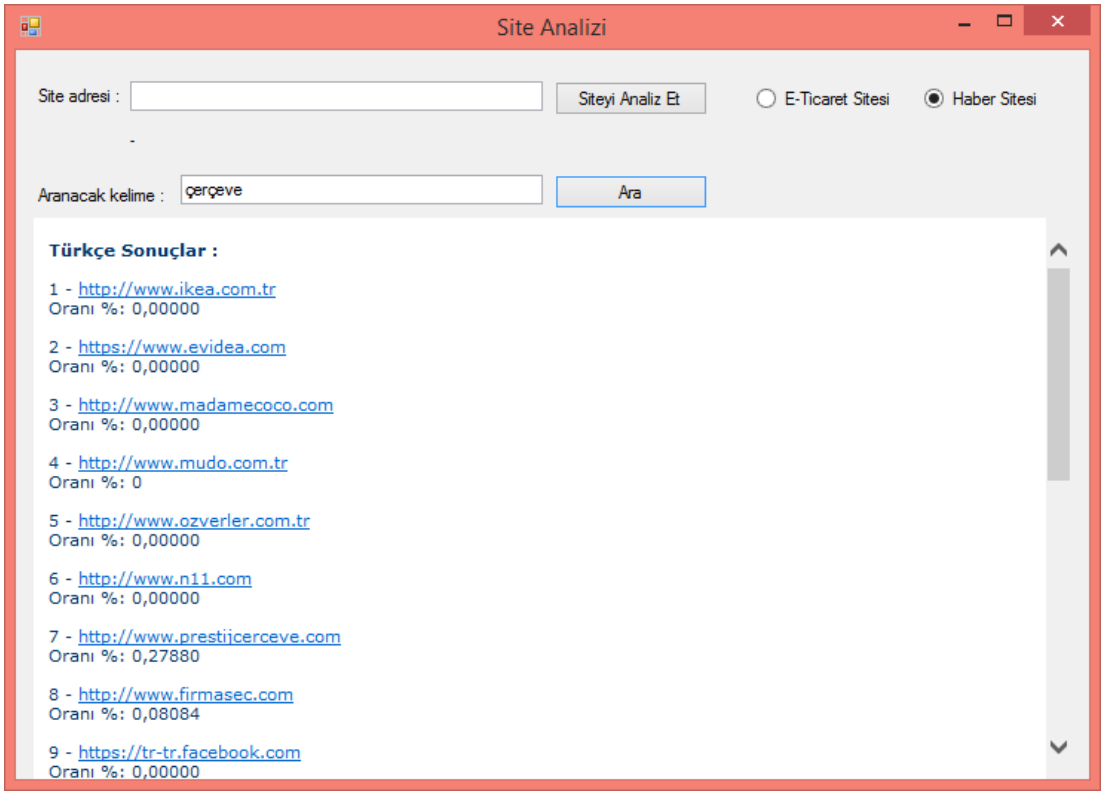
Haber sitesi olan ve olmayan kategorideki kelimelerin ağırlıkları yine Naïve Bayes yöntemi kullanılarak hesaplandı. İşlem sonrasında iki kategoriyi içeren kelime havuzundaki kelimelerin ağırlıkları, ait oldukları kategoriler üzerinden hesaplanmış oldu.

Şekil 3.4.'de e-ticaret sitesi için analizi bulunan www.tchibo.com.tr adresi, haber kategorisi için sorgulandığında, sonuç Şekil 3.9.'daki gibi çıkmaktadır.



Şekil 3.9. Haber sitesine göre analiz

Şekil 3.8.'de e-ticaret sitesi kategorisine göre sonuçları listelenen "çerçeve" kelimesi, haber sitesi kategorisinde değerlendirildiğinde sonuçlar Şekil 3.10.'daki gibi çıkmak tadır.



Şekil 3.10. Çerçeve kelimesinin Haber sitesi analizi

Şekil 3.8.'e göre, çerçeve kelimesi için Google'da çıkan arama listesindeki siteler üzerinden yapılan analiz işlemlerinde, sonuçların çoğunluğunda e-ticaret sitesi oranı yüksek çıkmış, ancak haber sitesi üzerinden yapılan analizlerde sonuçlar düşük çıkmıştır. Sonuçların düşük çıkması aslında kelimelerdeki oranların düşük olmasına bağlıdır. Şöyle ki, örnek bir kelime olarak "diğer" kelimesi ele alındığında, haber sitesi olan bir site için bu oran 0.3687740567000 iken, haber sitesi olmayan bir site için 1.3704343681000 olarak hesaplanmıştır. Sitedeki toplam kelime sayısı ve bu kelimenin sıklığı dikkate alındığında, haber sitelerinde bu kelimenin ağırlığı daha düşük çıkmaktadır.

4. SONUÇ VE ÖNERİLER

İnternetin yoğun olarak kullanıldığı günümüz teknolojisinde, kullanıcılar, reklam amaçlı siteleri aşarak, aradıkları e-ticaret, haber ya da benzer şekillerde kategorize edilebilecek sitelere daha kolay ulaşabilirler.

Bunun için, kategori seçimi yapıldıktan sonra, kategoriye ait olan ve olmayan sitelerden bir havuz oluşturup, sitelerin ana sayfalarındaki kelimeler havuzda toplanarak, Naïve Bayes yöntemi ile kelime ağırlıkları hesaplanabilir. Bu işlem sonrasında eğitilmiş bir makine bilgisi elde edilir. Sonrasında, istenilen sitenin analiz işlemi yapılarak, hangi oranda bu kategoriye yakın olduğu hesaplanır.

Yapılan bu tez çalışmasında, e-ticaret sitesi ve haber sitesi kategorileri ele alınmıştır. Örnek havuz oluşturma işleminde bu kategorilerdeki popüler siteler arasından rastgele seçim yapılmıştır. Bu iki kategori için de ayrı ayrı kelime havuzları oluşturulmuş ve kelime ağırlıkları hesaplanmıştır.

Analiz işlemi, direkt olarak bir web sitesi adresi girilerek bunun üzerinden yapılabileceği gibi, kelime girilerek Google arama motoru üzerinden bu kelime için çıkan arama sonuçlarındaki sitelerin adresleri üzerinden de yapılabilmektedir. Bu sayede kullanıcı, listelenen oranlara göre, aradığı e-ticaret veya haber sitesine daha kısa yoldan ulaşabilir. Kullanıcıya bu yönden avantaj sağlar.

Hazırlanan uygulama için farklı kategoriler eklenebilir. Uygulamaya yabancı dil desteği eklenerek yabancı dildeki internet siteleri üzerinden de analiz işleminin yapılması mümkün olabilir.

Tez çalışmasında ortaya konulan yazılım aracı içerisindeki analiz aşaması sırasında ana sayfalarla beraber sitenin iç sayfaları da dâhil edilirse daha kesin sonuçların alınabileceği düşünülmektedir.

5. KAYNAKLAR

[1] ALPAYDIN E., "Zeki Veri Madenciliği Sunumu", Boğaziçi Üniversitesi Bilgisayar Mühendisliği Bölümü, 2000.

[2] Akademik Bilişim Konferansları, Akdeniz Üniversitesi, "Veri Madenciliği", http://ab.org.tr/ab13/bildiri/175.pdf, 23-25 Ocak 2013.

[3] Çözümpark, "Data Warehouse", Çevrimiçi: http://www.cozumpark.com, erişim tarihi 31.12.2014.

[4] DEMİRALAY M., ÇAMURCU A.Y., "Cure, Agnes ve K-Means Algoritmalarındaki Kümeleme Yeteneklerinin Karşılaştırılması", İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, Yıl:4 Sayı:8, 2005/2.

[5] ÇELİKYAY E.K., "Metin Madenciliği Yöntemiyle Türkçede En sık Kullanılan ve Birbirini Takip Eden Harflerin Analizi ve Birliktelik Kuralları", Beykent Üniversitesi FBE, Yüksek Lisans Tezi, 2010.

[6] http://tr.wikipedia.org/wiki/Regresyon_analizi, erişim tarihi 04.01.2015.

[7] http://datawarehouse.gen.tr/veri-madenciligi-nedir/, erişim tarihi 04.01.2015.

[8] BÜYÜKKAVUT M., "Metin Madenciliği", https://prezi.com/kerijj7d07uq/metin-madenciligi/ , erişim tarihi 04.01.2015.

[9] http://tr.wikipedia.org/wiki/Metin_madenciliği, erişim tarihi 28.12.2014.

[10] TUNALI V.," Metin Madenciliği İçin İyileştirilmiş Bir Kümeleme Yapısının Tasarımı ve Uygulaması", Marmara Üniversitesi FBE, Doktora Tezi, 2011.

[11] GÜVEN A., "Türkçe Belgelerin Anlam Tabanlı Yöntemlerle Madenciliği", Yıldız Teknik Üniversitesi FBE, Doktora Tezi, 2007.

[12] Introduction to Text Mining (2008), SPSS Inc.

[13] FAN W., WALLACE L., RICH S., ZHANG Z., "Tapping into the power of text mining", Communications of ACM, 49(9), 76-82, 2006.

[14] DOLGUN M. Ö., ÖZDEMİR T. G., OĞUZ D., "Veri Madenciliğinde Yapısal Olmayan Verinin Analizi: Metin ve Web Madenciliği", İstatistikçiler Derneğinin İstatistikçiler Dergisi, No: 48-58, 2009.

[15] OĞUZ B., "Metin Madenciliği Teknikleri Kullanılarak Kulak Burun Boğaz Hasta Bilgi Formlarının Analizi", Akdeniz Üniversitesi Sağlık Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2009.

[16] PİLAVCILAR İ.F., "Metin Madenciliği İle Metin Sınıflandırma", Yıldız Teknik Üniversitesi FBE, Yüksek Lisans Tezi, 2007.

[17] ERGÜN K., "Metin Madenciliği Yöntemleri İle Ürün Yorumlarının Otomatik Değerlendirilmesi", Sakarya Üniversitesi FBE, Doktora Tezi, 2012.

[18] TD D., SC H., "Fong ACM. Associative Feature Selection for Text Mining", International Journal of Information Technology, No: 12(4): 59-68, 2006.

[19] DİRİ B., "Doküman Sınıflandırma Sunumu", Yıldız Teknik Üniversitesi Bilgisayar Mühendisliği Bölümü, 2014

[20] TANTUĞ A.C., "Metin Sınıflandırma", İTÜ Bilgisayar ve Bilişim Fakültesi ,Bilgisayar Mühendisleri Bölüm Başkanlığı Kurulu Dergisi, http://www.bmbb.info/dosya/dergi/6_4.pdf, erişim tarihi 20.01.2014.

[21] TEKNOMO K., "K Nearest Neighbors Tutorial Online Edition", 2012.

[22] ÖZKAN H., "K-Means Kümeleme ve K-NN Sınıflandırma Algoritmalarının Öğrenci Notları ve Hastalık Verilerine Uygulanması", İTÜ Fen Edebiyat Fakültesi, Bitirme Ödevi, 2013.

[23] ÖĞÜDÜCÜ G. Ş., "Veri Madenciliği Farklı Sınıflandırma Yöntemleri Proje Sunumu", İTÜ, 2007.

[24] http://tr.wikipedia.org/wiki/Naive_Bayes_sınıflandırıcı, erişim tarihi 11.12.2014.

[25] ŞEKER Ş.E., Kişisel internet sayfası, "Veri Madenciliği Sınıflandırma(Classification)", Yayın tarihi 31.03.2013.

6. ÖZGEÇMİŞ

Filiz ERTEN Gaziantep'te doğdu. İlk, orta ve lise öğrenimini İstanbul'da tamamladı. 1998 yılında Gaziantep Üniversitesi MYO Bilgisayar Programcılığı Bölümünden mezun oldu. Mezuniyeti sonrasında özel sektörde "İnsan Kaynakları" üzerine yazılım geliştiren bir yazılım firmasında çalışmaya başladı. Anadolu Üniversitesinde lisans tamamladı. 2013 yılında Maltepe Üniversitesi Fen Bilimleri Enstitüsü'nde Bilgisayar Mühendisliği Bölümü'nde yüksek lisansa başladı. 2015 yılında bu bölümden mezun oldu. Halen aynı firmada Proje Yöneticisi olarak görevini sürdürmektedir.