

**METİN MADENCİLİĞİ VE MAKİNE ÖĞRENMESİ İLE
İNTERNET SAYFALARININ SINIFLANDIRILMASI**

**WEB PAGE CLASSIFICATION USING TEXT MINING
AND MACHINE LEARNING**

İLKER ŞAHİN

DOÇ. DR. OUMOUT CHOUSEINOLOU

Tez Danışmanı

Hacettepe Üniversitesi

Lisansüstü Eğitim – Öğretim ve Sınav Yönetmeliğinin

Endüstri Mühendisliği Anabilim Dalı için Öngördüğü

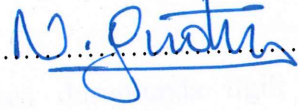
YÜKSEK LİSANS TEZİ olarak hazırlanmıştır.

2019

İLKER ŞAHİN'nin hazırladığı “METİN MADENCİLİĞİ VE MAKİNE ÖĞRENMESİ İLE İNTERNET SAYFALARININ SINIFLANDIRILMASI” adlı bu çalışma aşağıdaki jüri tarafından ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI'nda YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

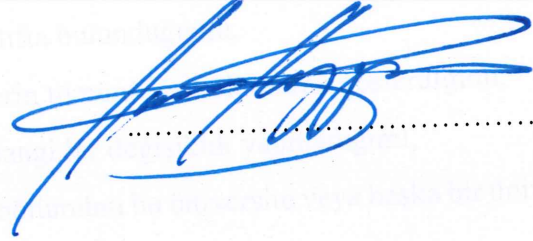
Doç. Dr. Necla GÜNDÜZ TEKİN

Başkan



Doç. Dr. Oumout CHOUSEINOLOU

Danışman




Prof. Dr. Özlem Müge TESTİK

Üye



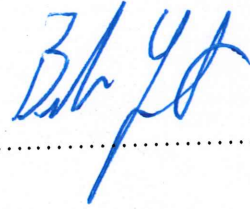
Dr. Öğr. Üyesi Ceren TUNCER ŞAKAR

Üye



Dr. Öğr. Üyesi Barbaros YET

Üye



Bu tez Hacettepe Üniversitesi Fen Bilimleri Enstitüsü tarafından YÜKSEK LİSANS TEZİ olarak / / tarihinde onaylanmıştır.

Prof. Dr. Menemşe GÜMÜŞDERELİOĞLU

Fen Bilimleri Enstitüsü Müdürü

ETİK

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmada,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversite veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

06 / 09 / 2019

ILKER ŞAHİN

YAYINLANMA FİKRİ MÜLKİYET HAKKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanması zorunlu metinlerin yazılı izin alarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan “*Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge*” kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H. Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

- ☐ Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir.
- ☐ Enstitü / Fakülte yönetim kurulu gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren ay ertelenmiştir.
- ☐ Tezim ile ilgili gizlilik kararı verilmiştir.

06 / 09 / 2019

İLKER ŞAHİN

ÖZET

METİN MADENCİLİĞİ VE MAKİNE ÖĞRENMESİ İLE İNTERNET SAYFALARININ SINIFLANDIRILMASI

İlker ŞAHİN

Yüksek Lisans, Endüstri Mühendisliği Bölümü

Tez Danışmanı: Doç. Dr. Oumout CHOUSEINOLOU

Eylül 2019, 50 sayfa

Alan adı bir web sitesinin İnternet ortamındaki adresidir. Bu alan adları kullanılarak, istenilen adres ziyaret edilebilir ve istenilen bilgiye ulaşılabilir. Günümüz dünyasında İnternet sitelerinin sayısı üstel artmakta ve bu sitelerin içeriğindeki zararlı içeriği engellemek ya da yararlı bilgilere daha kolay ulaşmak için, İnternet sayfalarını sınıflandırmak gerekmektedir. İnternet sitelerini sınıflandırmak için hem akademik çalışmalarda hem de özel şirketlerde, yöntemler ve algoritmalar geliştirilmektedir. Bu sayede İnternet kullanıcısının, içeriğine göre sınıflanan sitelerde, varsa sahtekârlık içeren unsurlara maruz kalmaması veya önceden belirlenen sınıfa sahip İnternet sitelerine erişimin engellenmesi hedeflenir.

İnternet sitelerinin filtreleneşmesi sayesinde, belirli sitelerin erişimine izin vermeye veya erişimi engellemek için kurallar oluşturmaya olanak tanınır. İnternet kullanıcıları için kurallar oluşturulabilir, belirlenen kurala göre kullanıcı, belirlenen sınıftaki İnternet sitelerine erişemez. Bu özelliğı sayesinde, sınıflandırma hem ev hem de iş ortamları için önem arz eder. Örneğın, ebeveynler, çocuklarının uygunsuz web sitelerini ziyaret etmelerini engelleyebilirken, şirketler çalışma saatlerinde çalışanların sosyal ağ vb. siteleri ziyaret etmelerini engelleyebilir. Bu sınıflandırma çalışması için, son yıllarda hızla yeni yöntemlerin ve algoritmaların geliştirilğı; istatistik, yazılım, endüstri mühendisliğı, matematik gibi farklı disiplinleri arasında bulunduran veri bilimi

kullanılmıştır. Veri biliminin alt dallarından makine öğrenmesi ve derin öğrenme algoritması ile bu sınıflandırma işlemi otomatize edilmiştir.

Tezin amacının uygulama kısmı ise İnternet sayfalarını sınıflandırmaktır. Çalışmanın sonunda girdi olarak bir alan adı verildiğinde, oluşturulan model sayesinde bu alan adına ilişkin bir sınıf bilgisinin geri dönüş olarak alınması amaçlanmıştır. Bu sınıflandırma işlemi için öncelikle İnternet sayfası-sınıfı şeklinde veriler çıkartılıp öğrenme seti ve test verileri oluşturulmuştur. Bu çalışmada, farklı makine öğrenmesi yöntemleri ve yapay sinir ağları kullanılarak İnternet sitesi sınıflandırma problemi incelenmiştir. Bu sınıflandırma probleminin çözümü için, İkili Sınıflandırma ve Çok Sınıflı Sınıflandırma olarak iki farklı yaklaşım uygulanmış, her iki yaklaşım da çalışma kapsamında toplanan İnternet siteleri üzerinde test edilip, performansları karşılaştırılmıştır. Başarıma bakıldığında ikili sınıflandırıcılar için en iyi performans gösteren algoritma Lojistik Regresyon olmuştur. Çok Sınıflı Sınıflandırma yaklaşımında uygulanan algoritmalar arasından ise en yüksek başarıma sahip yöntem Destek Vektör Makineleri (Support Vector Machines, SVM) olmuştur. Ayrıca, Çok Sınıflı Sınıflandırma problemi için farklı kelime vektörleştirme yöntemleri denenmiş ve performansları karşılaştırılmıştır. İkili ve Çok Sınıflı sınıflandırma yaklaşımlarında kullanılan algoritmaların ayrı ayrı ve farklı vektörleştirme yöntemleri ile denenmesi, İnternet sayfalarının sınıflandırılması ve içerik filtrelenmesi problemlerinin birlikte ele alınmasını sağlamış olup, alandaki benzer çalışmalardan farkı ortaya konmuştur. Öğrenme yöntemlerinin öğrenme ve test setlerinin yanlılığını araştırmak için performans araçlarından F1 Skoru, hata matrisi gibi teknikler kullanılmıştır. Tüm deneysel sonuçlar göz önüne alındığında, İkili Sınıflandırma sadece istenilen bir İnternet site sınıfının filtrelenmesi görevini yerine getirmek için kullanıldığında daha etkili olacağı tespit edilmiştir. İkili Sınıflandırmada kullanılan yöntemlerin, analizin süreçleri boyunca işlemsel performans (süre) göz önüne alındığında, Lojistik Regresyon ve Bernoulli Naive Bayes sınıflandırıcılarının, yapay sinir ağlarına göre 150 kat daha hızlı sonuçlandığı gözlenmiştir.

Anahtar Kelimeler: İnternet Sayfa Sınıflandırması, Metin Madenciliği, Doğal Dil İşleme, Derin Öğrenme, Makine Öğrenmesi

ABSTRACT

WEB PAGE CLASSIFICATION USING TEXT MINING AND MACHINE LEARNING

İlker ŞAHİN

Master of Science, Department of Industrial Engineering

Supervisor: Assoc. Prof. Dr. Oumout CHOUSEINOLOU

September 2019, 50 pages

The domain name is the address of a website on the Internet. By using these domain names, the desired address can be visited and the desired information can be accessed. In today's world, the number of Internet sites are increasing exponentially and in order to prevent accessing possible harmful content in these web sites or to find useful information more easily it is necessary to classify the web pages. Methods and algorithms for website classification are proposed by both academic studies and private companies. Hence, it is intended that the Internet user is not exposed to fraudulent elements in any of the sites classified according to their content or the access to predetermined websites is prevented.

Filtering Internet sites allows us to set rules to allow or block access to certain sites. Rules can be created for specific users of the computer, the user cannot access the Internet sites in the specified class according to the specified rule. For this feature, classification is important for both household and work environments. For example, while parents can prevent children from visiting inappropriate web sites, companies might also prevent their employees to visit social media websites during work hours. For this classification study the approaches in the domain of data science, which includes several disciplines such as statistics, software engineering, industrial engineering and mathematics, and where new methods are continuously developed and proposed in the last years, have been employed. The classification process has been automated with machine learning and deep learning algorithms, which are sub-branches of data science.

The technical part of this thesis is the classification of web pages. The aim is that at the end of the study, when a web domain name is given as input, a class value should be returned for this web domain with respect to the developed model. For this classification process, firstly the data was extracted in the form of web page-class, and accordingly the learning set and test data were created. In this study, web site classification problem is investigated by using different machine learning methods and artificial neural networks. In order to solve this classification problem, two different approaches have been employed, namely Binary Classification and Multi-Class Classification. Both approaches have been tested on web sites collected in the study and their performance has been compared. In terms of performance, it has been observed that for binary classifiers Logistic Regression is the best performing algorithm. Among the algorithms applied in the Multi-Class Classification approach, Support Vector Machines (SVM) is the most successful method. Furthermore, different word vectorization methods have been employed and their performances have been compared in the Multi-Class Classification problem. The use of algorithms in Binary and Multi-Class Classification approaches by employing different vectorization methods, is a combined approach to the problems of classification of web pages and content filtering, and this puts forward the difference of the current study from similar studies in the field. In order to investigate the bias of learning methods and test sets, techniques such as F1 score of performance and error matrix were used. Considering all experimental results, it has been found that Binary Classification will be more effective only when used to fulfill the task of filtering a desired Internet site class. In the analysis, Logistic Regression and Bernoulli Naive Bayes classifiers have been found to be 150 times faster than artificial neural networks when computing performance (time) of the methods used in Binary Classification has been taken into account.

Keywords: Web Page Classification, Text Mining, Natural Language Processing, Deep Learning, Machine Learning

TEŞEKKÜR

Lisansüstü eğitimimin tez konusu seçme, araştırma yapma ve yön tayin etme gibi tüm aşamaları boyunca, desteğini hiç esirgemeyen ve sadece bilimsel olarak da değil hayata bakış açısı ve pozitif duyguları hep ağır basan yönüyle bana hep yol gösteren saygıdeğer hocam Doç. Dr. Oumout CHOUSEINOLOU'na

Tez aşamasında sürekli destekçim olan, eksikliğini hiç hissettirmeyen, her türlü teknik ve motivasyon sorunumda yanımda olan Comodo'dan proje arkadaşım Menekşe KUYU'ya

Sadece lisansüstü eğitimim sırasında değil tüm hayatım boyunca tüm sevgi ve sabırlarıyla destek veren, bugünlere gelmemde en büyük katkıları ile koşulsuz şartsız yanımda olan ve karşılığını hiçbir zaman ödeyemeyeceğim canım annem Leyla KOÇ'a ve ablam İlkay Şahin KARSLI'ya

Sonsuz Teşekkürler...

İlker ŞAHİN

Eylül 2019, Ankara

İÇİNDEKİLER

ÖZET	i
ABSTRACT	iii
TEŞEKKÜR.....	v
İÇİNDEKİLER	vi
ÇİZELGELER DİZİNİ.....	viii
ŞEKİLLER DİZİNİ	ix
SİMGELER VE KISALTMALAR.....	x
1. GİRİŞ	1
1.1. Metin Sınıflandırması	2
1.2. Kelime Vektörleştirme Algoritmaları	3
1.3. Sınıflandırma Sırasında Kullanılan Algoritmalar	4
1.4. Araştırma Soruları.....	6
2. BENZER ÇALIŞMALAR	8
2.1. Metin Sınıflandırma Çalışmaları	8
2.2. İnternet Sitesi Sınıflandırma Çalışmaları	11
3. METODOLOJİ.....	18
3.1. Veri Tanımı ve Toplanması.....	18
3.2. Verinin Ön İşleme Süreci (Veri Temizleme)	21
3.3. Kelime Vektörleştirme Yöntemleri (Word Embedding)	21
3.3.1. Terim Sıklığı-Ters Metin Sıklığı (Term Frequency – Inverse Document Frequency, TF-IDF).....	22
3.3.2. Kelime Torbası (Bag Of Words - BOW)	23
3.3.3. Word2Vec.....	23
3.4. Sınıflandırma Sırasında Kullanılan Algoritmalar	23
3.4.1. Rastgele Orman Algoritması (Random Forest)	26

3.4.2. Naive Bayes Sınıflandırıcı.....	26
3.4.3. Destek Vektör Makineleri (Support Vector Machine- SVM)	28
3.4.4. Lojistik Regresyon	29
3.4.5. Tam Bağlantılı Yapay Sinir Ağları	30
3.5. Sınıflandırıcı Performans Ölçme Yöntemleri	35
4. ANALİZ VE VAKA ÇALIŞMASI.....	38
4.1. İkili (Binary - Binominal Class) Sınıflandırma	39
4.2. Çok Sınıflı Sınıflandırma	44
5. SONUÇLAR VE ÖNERİLER	45
5.1. İnternet Sitelerinin Sınıflandırma Probleminde İkili Sınıflandırma Yaklaşımı ...	45
5.2. İnternet Sitelerinin Sınıflandırma Probleminde Çok Sınıflı Sınıflandırma	47
5.3. Araştırma Soruları Üzerine Öneriler	48
6. KAYNAKLAR.....	51
7. EKLER.....	55
7.1. EK - 1	55
7.2. Tezden Türetilmiş Bildiriler	58
ÖZGEÇMİŞ	59

ÇİZELGELER DİZİNİ

Çizelge 1. Ön işleme süreci sonrası en fazla toplam kelime sayısı sahip olan ilk 10 sayfa sınıfı ve toplam kelime sayıları (kısmi örnek)	19
Çizelge 2. Ön işleme süreci sonrası elde edilen kelime kökleri (kısmi örnek)	21
Çizelge 3. İnternet sınıfları ve kelimelerin TF-IDF skorları (kısmi örnek)	24
Çizelge 4. Örnek çekilmiş İnternet sınıfları ve kelimelerin kelime torbası istatistikleri (kısmi örnek)	25
Çizelge 5. Hata Matrisi Gösterimi.....	37
Çizelge 6. Örnek sınıfa ait örnek ham veri (kısmi örnek)	38
Çizelge 7. Sınıflandırıcı Algoritmaların Tüm Sınıflar Üzerindeki Sonuçları.....	41
Çizelge 8. Çok Sınıflı Sınıflandırıcıların Farklı Kelime Vektörleştirme Yöntemlerine Göre Başarım Oranları.....	44
Çizelge 9. İkili Sınıflandırıcı Algoritmaların Performansları	46
Çizelge 10. Çok Sınıflı Sınıflandırıcı Algoritmaların Başarım Oranları.....	48

ŞEKİLLER DİZİNİ

Şekil 1. Yıllara göre İnternet sayfalarının ve İnternet kullanıcı sayılarının dağılımı [1].	1
Şekil 2. Sınıflandırma Süreçleri: (a) İkili Sınıflandırma Süreci (b) Çok Sınıflı Sınıflandırma Süreci.....	6
Şekil 3. Çalışmada Uygulanan Sınıflandırma Süreci	20
Şekil 4. İki sınıflı örnek bir model için maksimum marjlı hiper düzlemi	29
Şekil 5. Sinir ağındaki bir düğümün ilerleme süreci.	32
Şekil 6. Tam Bağlantılı YapaySinir Ağları Mimarisi.....	33
Şekil 7. Sınıflı Erişkin İçerik ve Diğerleri olan Bir İkili Sınıflandırıcı FCNN' in Hata Matrisi.....	40
Şekil 8. İkili Sınıflandırıcıların Ölçüm Ortalama Değerleri	47
Şekil 9. Sınıflandırıcı Algoritmalar ile Kelime Vektörleştirme Kombinasyonlarının Başarım Oranları(%)	48

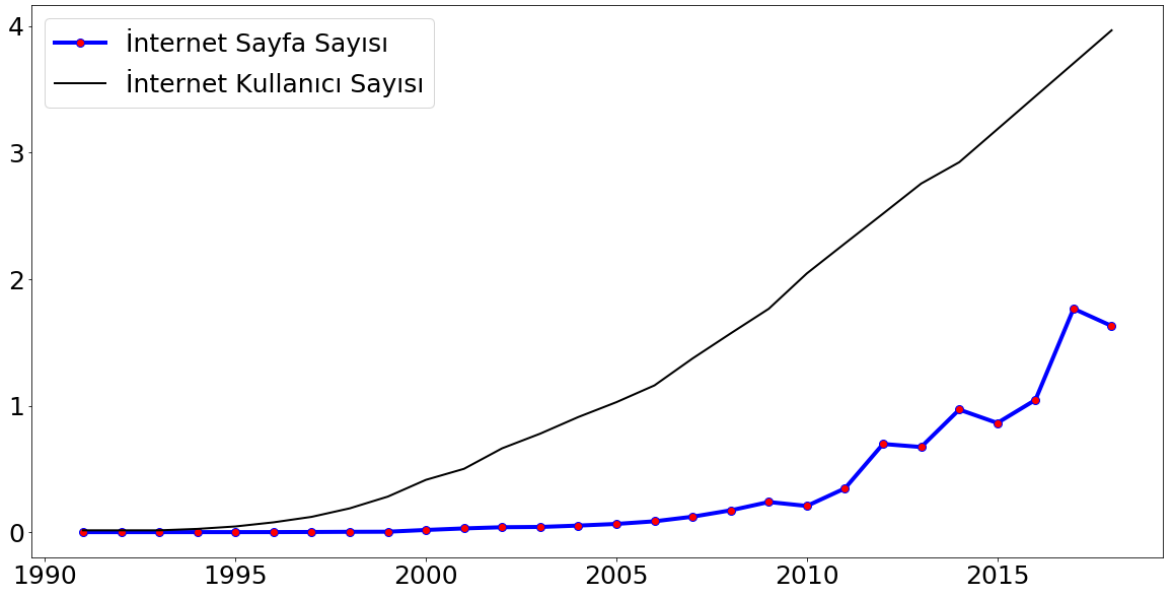
SİMGELER VE KISALTMALAR

Kısaltmalar

BOW	Kelime Torbası (Bag of Words)
BPNN	Geri Yayılım Sinir Ağı (Back Propagation Neural Networks)
FCNN	Tam Bağlantılı Yapay Sinir Ağları (Fully Connected Neural Network)
GA	Genetik Algoritma
HTML	Hiper Metin İşaretleme Dili (Hypertext Markup Language)
KNN	k -En Yakın Komşu (k Nearest Neighbour)
LSA	Gizli Anlambilimsel Analiz (Latent Semantic Analysis)
LSTM	Uzun ve Kısa Vadeli Hafıza Ağları (Long-Short Term Memory)
MBPNN	Değiştirilmiş Geri Yayılma Sinir Ağı (Multilayer Back Propagation Neural Networks)
M-C	Eşleşme-Yakınsama (Mapping-Convergence)
ML-KNN	Çok Etiketli KNN (Multi-Label KNN)
NLP	Doğal Dil İşleme (Natural Language Processing)
PEBL	Olumlu Örnek Tabanlı Öğrenme (Positive Example Based Learning)
RNN	Özyineli Sinir Ağı (Recurrent Neural Network)
SSO	Basitleştirilmiş Sürüler Optimizasyonu (Simplified Swarm Optimization)
SVM	Destek Vektör Makineleri (Support Vector Machine)
TF-IDF	Terim Sıklığı-Ters Metin Sıklığı (Term Frequency – Inverse Document Frequency)
URL	Birörnek Kaynak Konumlayıcı (Uniform Resource Locator)
VG-RAM WNN	Sanal Genelleme Rasgele Erişim Belleği Ağırlıksız Sinir Ağları (Virtual Generalizing Random Access Memory Weightless Neural Networks)
VSM	Vektör Uzay Modeli (Vector Space Model)

1. GİRİŞ

Günümüz dünyasında bilgiye en kolay ve en hızlı şekilde erişme imkan sağlayan başlıca alan İnternet'tir ve bu ortamda bulunan sayfalara alan adları kullanılarak ulaşılır. Alan adı; bir İnternet sitesinin İnternet ortamındaki adresidir. Bu alan adları ile İnternet tarayıcısı üzerinden, istenilen adres ziyaret edilebilir ve istenilen bilgiye ulaşılabilir. Bu ihtiyaç doğrultusunda, sanal ortamda depolanan bilgiler de İnternet sitelerin sayısı ile orantılı olarak hızlı bir şekilde artmaktadır [1]. Bugün dünya çapında yaklaşık 1,5 milyar İnternet sitesi bulunmaktadır. Bunlardan 200 milyona yakını aktiftir [1] [2]. NetCraft tarafından yürütülen Ekim 2018 Web Sunucusu Anketi'nde onaylandığı ve ilk olarak İnternet Live Stats tarafından tahmin edildiği ve duyurulduğu üzere; ilk olarak Eylül 2014'de 1 milyar İnternet sitesine ulaşılmıştır. İnternet sitelerinin sayısı Aralık 2016'da 1.7 milyara yükselmiştir. Şekil 1'de yıllara göre toplam [1] İnternet sitelerinin sayısı gösterilmektedir.



Şekil 1. Yıllara göre İnternet sayfalarının ve İnternet kullanıcı sayılarının dağılımı [1].

İnternet siteleri ve sayfalarının sayısının artması ile bilgi kaynaklarının erişimi sırasında, zararlı içeriği engellemek ya da yararlı bilgilere daha hızlı ve kolay erişmek için, İnternet sayfalarını sınıflandırma ihtiyacı doğmuştur [3]. Metin yoğunluklu bir sitenin içeriğini insan gözüyle anlamak için çok fazla zaman ve efor harcanabilir. İnternet sitesi

sınıflandırması ile kişilerin yaptıkları bir araştırma sırasında konu hakkında gereksiz bilgilere yönelmekten, konunun üst başlığı ön görülüp yararlı olduğu düşünülen bilgiye daha hızlı ve etkili ulaşılabilirken, zararlı içeriklere ise kişilerin erişimi engellenebilmektedir. İnternet sitelerini sınıflandırmak için hem araştırmacılar hem de siber güvenlik sağlayan özel şirketler sınıflandırma yöntemleri ve algoritmaları geliştirmektedir. Bu sayede İnternet kullanıcısının, içeriğine göre sınıflanan sitelerde, varsa sahtekârlık içeren unsurlara maruz kalmaması veya önceden belirlenen sınıflara sahip İnternet sitelerine erişimin engellenmesi hedeflenmektedir. Bu zararlı içerikler arasında oltalama (phishing), yalan haber (fake news) gibi siber güvenliğin büyük kısmını oluşturan kullanıcı tehditleri bulunmaktadır. Mevcut çalışmada bu güvenlik unsurlarına değinilmeyecek olsa da, İnternet sitelerinin sınıflandırılması ardından ek çalışmalarla bu unsurlar da aynı yöntemlerle sağlanabilir [4] [5]. İnternet sitelerini sınıflandırmak için hem akademik çalışmalarda hem de özel şirketlerde, yöntemler ve algoritmalar geliştirilmektedir. Bu sayede İnternet kullanıcısının, içeriğine göre sınıflanan sitelerde, varsa sahtekârlık içeren unsurlara maruz kalmaması veya önceden belirlenen sınıfa sahip İnternet sitelerine erişimin engellenmesi hedeflenmektedir.

Bu sınıflandırma işlemi için öncelikle İnternet sayfası-sınıfı şeklinde veriler oluşturulup makine ve derin öğrenme tekniklerinde kullanılacak öğrenme seti ve test seti oluşturulmuştur. Çalışmanın sonunda girdi olarak bir alan adı verildiğinde, bu alan adına ilişkin bir İnternet sınıfı tahmini yapılması amaçlanmıştır. Bu çalışmada İnternet sitesi sınıflandırma probleminin çözümü; verinin tanımı ve toplanması, verinin ön işleme süreci, kelime vektörleştirme ve algoritmaların eğitim, test ve analiz süreci olmak üzere dört alt aşama altında ele alınmıştır. İnternet sayfalarının sınıflandırılması bir metin sınıflandırma problemi olarak kabul edilmiş, bu aşamaların her biri için metin sınıflandırmasında kullanılan yöntemler kullanılmıştır.

1.1. Metin Sınıflandırması

Metin sınıflandırması; önceden belirlenen sınıflardan faydalanarak, ele alınan bir metnin, belgenin ya da bir cümlemin hangi sınıfa dâhil olacağının otomatik olarak hesaplanması işlemidir [6] [7]. Başka bir deyişle **içeriğine göre metne etiket veya kategori atama işlemi** olarak tanımlanabilir. Metin sınıflandırması, matematiksel olarak Denklem 1 [8] şeklinde ifade edilebilir.

$$Y = f(X, \theta) + \varepsilon \quad (1)$$

Bu denklemde,

f : eğitim verilerini kullanarak tahmin eden sınıflandırıcı veya sınıflandırma modeli

Y : metnin belirli bir sınıfa üyeliğini belirten sayısal değer

X : kelimeler veya kelime gruplarını içeren bir metin vektörü

θ : f fonksiyonuyla ilişkilendirilen bilinmeyen parametrelerin kümesi

ε : sınıflandırma hatasıdır.

Bu çalışmada metin sınıflandırma yaklaşımına model (f) olarak; makine öğrenmesi ve derin öğrenme literatüründe önerilmiş olan yaygın birkaç yaklaşım kullanılarak, başarımlar ve etkinlikleri uygun değerlendirme yöntemleri ile ölçülmektedir. Sınıflandırma modeli oluşturulurken Çok Sınıflı Sınıflandırma ve İkili Sınıflandırma yaklaşımları kullanılmıştır. Model oluşturulduktan sonra yeni bir metin verildiğinde bu eğitim sürecinde eğitilmiş sınıflardan birine dâhil olması planlanmıştır [9].

Metin biçiminde yapılandırılmamış ve sınıflandırılmamış veriler günlük hayatta sıklıkla karşımıza çıkmaktadır. E-postalar, sohbetler, İnternet sayfaları, sosyal medya, support ticket (şikâyet ve destek bildirimi), anket yanıtları gibi birçok örnek verilebilir. Metinsel veriler, son derece zengin bir bilgi kaynağı olabilir, ancak yapılandırılmamış doğası nedeniyle fayda sağlayacak bilgileri elde etmek zor ve zaman alıcı olabilir. Bu sebeplerden dolayı kişiler, kuruluşlar veya işletmeler, karar verme sürecini kısaltma ve süreçleri otomatikleştirme yoluna gitmektedir. Bunun için kullanılacak hızlı ve uygun maliyetli yöntem metnin otomatik olarak sınıflandırılmasıdır.

Bu süreç; duygu analizi, konu etiketleme, spam (istenmeyen mesajlar) tespiti gibi geniş uygulamalarda kullanılabilir. Bu uygulamaların hepsi, metnin içerisindeki kelime kombinasyonlarının bir istenmeyen e-posta, anket sonucunda memnuniyet gibi bir sınıfı veya kategoriye temsil edecek şekilde makinenin eğitilmesidir [9] [10] [11].

1.2. Kelime Vektörleştirme Algoritmaları

Kelime vektörleştirme, makine ve derin öğrenme süreçlerinden önce uygulanan bir dönüşüm işlemidir. Bu dönüşüm ile birlikte metin farklı yöntemlere göre kendi sınıfı

içerisinde sayısal veriler içeren vektörlere dönüştürülür [12]. Bu işlem kelimelerin birbirine yakınlığı ya da kelimelerin sınıf içerisindeki sıklığı gibi farklı yöntemler kullanılarak yapılabilir. Bu işlemin sonucunda, metin verisi algoritmalar tarafından eğitilebilecek ve analiz edilebilecek sayısal değerlere dönüşmüş olur. Bu yöntemlerin bazılarını kullanırken dikkatli olunması gerekir. Çünkü kategorilere farkında olmadan birbirleri arasında üstünlük verebilecek değerler atanabilir. Bu yüzden metin sınıflandırmasının başarımını etkileyecek en önemli süreçlerden biri de kelime ve kelime gruplarının vektörleştirilmesidir [13].

Bu çalışmada kelime vektörleştirme işlemi için Kelime Torbası (Bag of Words, BOW) [14], Terim Sıklığı-Ters Metin Sıklığı (Term Frequency – Inverse Document Frequency, TF-IDF) ve Word2Vec [12] yöntemleri bu tez çalışmasındaki her bir uygulamada ayrı ayrı kullanılmış, devamında da kendi aralarındaki başarımlarını kıyaslanmıştır.

1.3. Sınıflandırma Sırasında Kullanılan Algoritmalar

Tez kapsamında, sınıflandırılması istenen metin İnternet sayfasının içeriği olup, kullanılacak olan sınıflar da önceden belirlenen ve şu anda yayında bulunan İnternet sitelerin büyük çoğunluğunu temsil edebilecek sınıflardır. Bu kapsamda, sınıflandırma işlemini yerine getirmek için önceden belirlenen sınıflara ait ham veriler (içerik metni) analiz edilebilir hale getirdikten sonra makine ve derin öğrenme sınıflandırıcıları kullanılarak bir metin sınıflandırma modeli oluşturulmuştur.

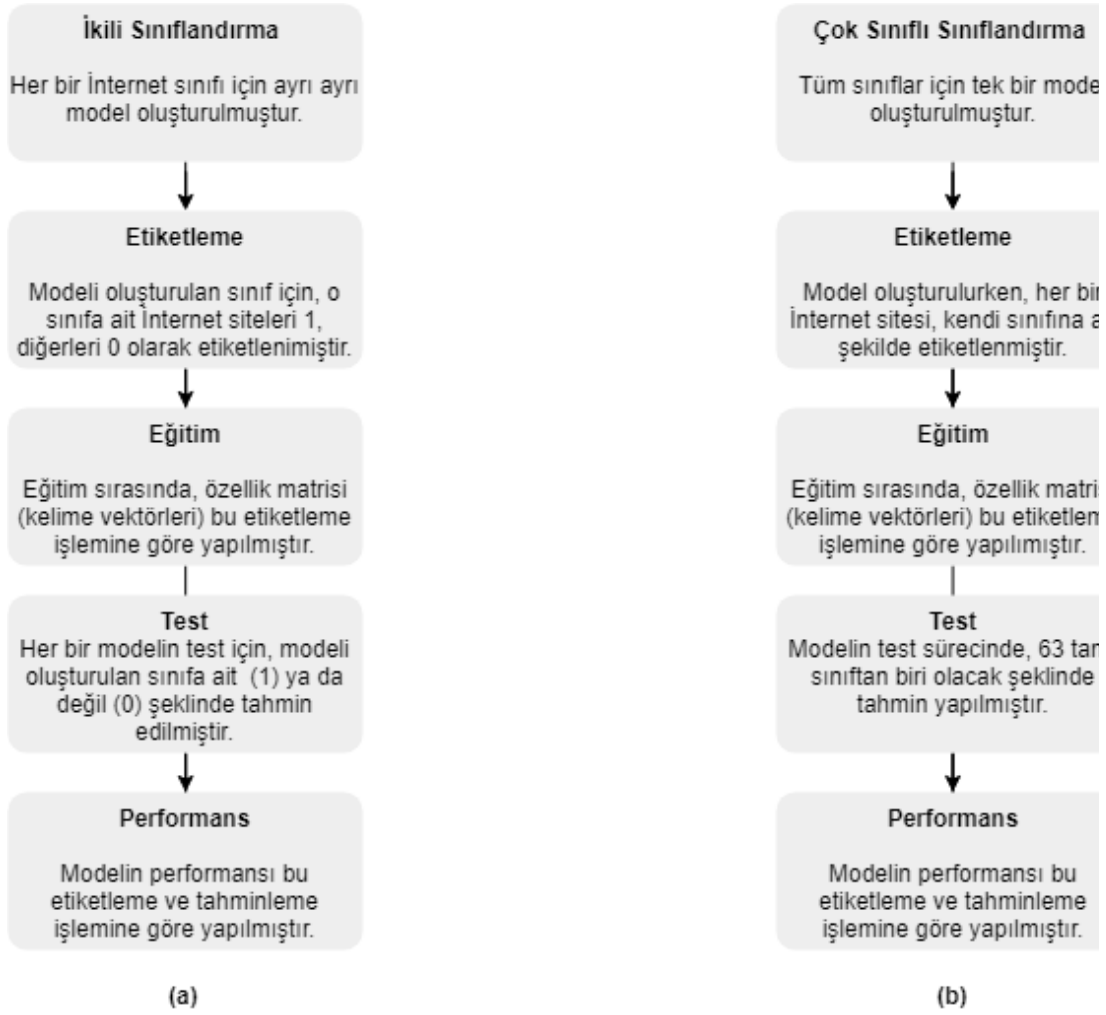
Bu çalışmada, gözetimli öğrenme algoritmaları kullanılmıştır [15]. Gözetimli öğrenme algoritmaları, önceden etiketlenmiş verileri girdi olarak almaktadır. Metin sınıflandırma probleminde, İnternet siteleri ait olduğu sınıflarla birlikte modele beslenmektedir. Böylelikle model eğitim setindeki sınıflandırmayı öğrenebilmektedir.

Bu süreçte; üzerinde çok defa farklı alanlarda da çalışılmış makine öğrenmesi ve derin öğrenme algoritmaları seçilmiş ve bu algoritmalar, İkili ve Çok Sınıflı sınıflandırıcı modellerini oluşturmak için kullanılmıştır.

- **İkili Sınıflandırma:** İkili veya binom sınıflandırma, belirli bir veri setinin öğelerini bir sınıflandırma kuralı temelinde iki gruptan (her birinin hangi gruba

ait olduğunu tahmin ederek) birine atama işlemidir. Bir tür gözetimli öğrenme olan bu yöntemde, eğitim sırasında özellik seti ile birlikte etiketlenen değişken, modeli oluşturulan sınıf için 1, geri kalan örnekler için 0 değeri alır [16]. Örneğin “Araba” sınıfı için bir ikili sınıflandırma modeli oluşturulmak istendiğinde, “Araba” sınıfına ait İnternet sayfaları 1, geri kalan sayfalar 0 olarak etiketlenir. Bu çalışmada, tanımlanmış sınıf sayısı kadar (63 tane) sınıflandırıcı oluşturulmuş ve bunların her birinin başarımları, hata matrisleri ve işlem süreleri çıkarılmıştır. İkili sınıflandırma sürecinde yapılan her bir işlem Şekil 2a’da gösterilmiştir. Tam Bağlantılı Yapay Sinir Ağları (Fully Connected Neural Network, FCNN), Lojistik Regresyon (Logistic Regression) ve Bernoulli Naive Bayes algoritmaları ikili sınıflandırma için ayrı ayrı kullanılıp her birinin başarımları incelenmiştir [17].

- **Çok Sınıflı Sınıflandırma:** Çok sınıflı sınıflandırma yöntemi bir gözetimli sınıflandırma yöntemi olup, bu sınıflandırma içinde bulunan algoritmalar üç ve daha fazla sınıfı olan veriler için kullanılmaktadır. Çalışmada sınıflandırma modeli oluşturulurken, her bir İnternet sitesi sınıfı, kendi özellik setinde (kelime vektörü) temsil edilecek şekilde eğitilmiştir. Eğitim sonrası algoritmaların testleri sürecinde de yine tahmin edilmesi beklenen çıktı yine her bir 63 sınıftan biridir. İkili sınıflandırmadan farklı olarak; her bir sınıf için ayrı ayrı sınıflandırıcı yerine, tek bir sınıflandırıcı oluşturulup, tüm testler bu model üzerinden yapılmıştır. Çok sınıflı sınıflandırma sürecinde yapılan her bir işlem Şekil 2b’de gösterilmiştir. Çok sınıflı sınıflandırma modelini oluşturmak için, Multinomial Naive Bayes, Rastgele Orman (Random Forest) ve Destek Vektör Makineleri (Support Vector Machines – SVM) sınıflandırıcıları ayrı ayrı kullanılıp her birinin başarımları incelenmiştir [18] [19] [20].



Şekil 2. Sınıflandırma Süreçleri: (a) İkili Sınıflandırma Süreci (b) Çok Sınıflı Sınıflandırma Süreci

1.4. Araştırma Soruları

Mevcut çalışmada, İnternet site sınıflandırma problemi ele alınırken, çalışmayı dolaylı veya dolaysız etkileyecek bazı cevaplanması gereken sorular ön görülmüştür. Bu kapsama uygun olarak aşağıdaki araştırma soruları (AS) tanımlanmıştır:

AS1: Bir İnternet sayfa sınıfının filterelenmesi probleminde; aynı veri setinde, her bir sınıf için kendi içerisinde ikili (binary) sınıflandırıcı ve tüm sınıfları içeren çok sınıflı sınıflandırıcının kullanılması mümkün müdür?

AS2: İkili sınıflandırmada kullanılan yöntemler arasında, analiz süreçleri boyunca işlemsel performans (süre) olarak nasıl bir fark bulunmaktadır?

AS3: Herhangi bir sınıfa ait veri setinin dengesiz yani eşit sayıda olmaması, eğitim sırasında o sınıfın temsil edilmemesine sebep olmuş mudur?

AS4: Kullanılan makine öğrenmesi sınıflandırıcılarının performansları arasında belirgin bir fark olmuş mudur?

Tezin devamı aşağıdaki gibi biçimlendirilmiştir. Bölüm 2’de yapılmış olan literatür taramasında tanımlanan benzer çalışmalar özetlenmiş, Bölüm 3’te bu tez çalışmasında takip edilen metodoloji verilmiştir. Tezde yapılan vaka çalışması ve sonuçları Bölüm 4’te, tezin sonuçları ve önerileri de Bölüm 5’te sunulmaktadır.

2. BENZER ÇALIŞMALAR

Literatürde, metin sınıflandırma için kullanılan tekniklerin çok farklı problemlerde farklı yöntemlerle denendiği görülmektedir. Önerilen yöntemler siber güvenlikten sağlık alanına kadar çok geniş bir alanda kullanılıp, uygulanan alana göre maliyeti azaltıp, hata payını küçülttüğü gözlenmiştir. Ek olarak sürecin hızlandığı, çoğu çalışma kapsamında vurgulanmıştır. Bu hızlanma el ile yapılan analizlerden, metnin baştan sona okunmasına kadar olan kısmın otomatize edilmesidir. Bu sayede insan hatası da minimize edilmektedir.

2.1. Metin Sınıflandırma Çalışmaları

Rong, Yuancheng ve Xiangqian [21] yaptığı çalışmada, İnternetin hızlı gelişimiyle birlikte, sanal ortamda arama motorlarını kullanan kullanıcıları etkileyen istenmeyen uygulama ve içeriklerin sayısının arttığını vurgulamıştır. Bu durumu azaltmak ve önlemek için yapılan web spam sınıflandırmasını geliştirmek adına **ilk kez derin inanç ağları (deep belief networks)** kullanılmıştır. Çok yönlü araştırma ve değerlendirme sonrasında bu çalışmanın etkinliğini artırmak için Sentetik Azınlık Aşırı Örnekleme Tekniği ve Gürültü Giderici Otomatik Kodlayıcı algoritmaları etkili bir şekilde birleştirilmiştir. WEBSPAM-UK2007 veri kümesi ile elde edilen test sonuçlarına göre; bu çalışmada önerilen sınıflandırma yönteminin sınıflandırma performansını gelecek sınıflandırma çalışmalarına katkı sağlayacak bir iyileştirme yaptığı tespit edilmiştir.

Agren ve Agren [10], çalışmasında yalan haberin nasıl yapıldığı incelenmiş ve insanların yatırım hareketlerinde ve gelecek planlamalarında yanlış karar vermeye yol açmasına ve New York Times gibi basın organlarında “aldatmak amacıyla uydurma hikâye” başlığıyla, bu sahtekârlık yönteminin yaygınlaştığını belirtmişlerdir. Yapay zekâ ve makine öğrenme yöntemleri ile bu yalan haberlerin tespit edilmesi amaçlamıştır. **Doğal dil işleme (Natural Language Process – NLP) ile metnin yazarının; yanlış, yanlış ya da tarafsız olup olmadığını; haber başlığının haber içeriği ile ilişkisi ile ilgili çıkarımlarda bulmaya çalışılmıştır.**

Cardoso ve Silva [5] yürüttüğü çalışmada, yalan ve yanlış görüşler nedeniyle müşterilerin bazı şirket veya ürünler hakkında yanıltıcı fikirlere yöneldiği ve günlük hayatlarındaki tüketim hareketlerinde olumsuz yönde değişikliğe neden olduğu vurgulamıştır. Çalışmada, kullanılacak olan sınıflandırma yöntemlerinin performansını artırmak için; gerçek dünyada çevrimiçi öğrenme gerektirip gerektirmediği tartışılmış ve yalan veya yanlış görüş tespiti için metin bazlı kapsamlı sınıflandırma analizi kullanılmıştır.

Kudugunta ve Ferrara [22] çalışmalarında, bir ülkede seçimleri etkileyebilecek, sağlık çalışmalarında sahte gelişmelerle insanlar üzerinde psikolojik ve maddi etkiler yaratabilecek ve bir yazılım ile otomatik oluşturulan tweetleri tespit etmek amaçlamıştır. Çalışma kapsamında tekrarlayan sinir ağlarının bir türü olan uzun-kısa vadeli hafıza ağları kullanılmıştır.

Figueira [4] çalışmasında da; basılı ve İnternet ortamında yayınlanan haberlerin her geçen gün arttığını ve özellikle sosyal ağlarda paylaşılan ve haber niteliği taşıyan yazılar kısa bir sürede çok fazla kullanıcıya ulaşır, çoğu kullanıcının üzerinde etki yarattığı vurgulamıştır. Bu çalışmada ele alınan problem haberin kaynağına ve haberin yayılmasına göre ikiye ayrılmış, yalan haberlerin tespitinde ise gizli Markov modelleri kullanılmıştır.

Yao ve Zhi-Min [23], modern yazı sınıflandırma metotlarının konu bazlı yazı sınıflandırmasında eksik kalabildiğinin altını çizmekte ve DNB adını verdikleri metodunu Destek Vektör Makineleri (Support Vector Machine - SVM), k -En Yakın Komşu (k Nearest Neighbour – KNN), Naive Bayes gibi yöntemler ile kıyaslamakta olup, DNB'nin sınıflandırma sonuçlarının başarısını arttırdığını vurgulamaktadır.

Trstenjak, Mikac ve Donko yaptıkları çalışmada [24], KNN algoritmasının metin sınıflandırma için çok popüler bir yöntem olduğunu vurgulayarak, TF-IDF ile KNN algoritmasını denemiştir. Metin sınıflandırma yöntemi ve çerçevesi, çeşitli parametrelere, analiz sonuçları ve ölçümlere göre oluşturulmuştur. Çalışmanın değerlendirilmesi sırasında, sınıflandırma hızına ve kalitesine odaklanılmıştır. Test aşamasında, önerilen algoritmanın iyi ve kötü yönleri tespit edilmiş ve gelecek çalışmalara fayda sağlayabileceği belirtilmiştir.

Liang ve arkadaşları [25], kimya konulu İnternet sayfalarını verimli bir şekilde sınıflandırmak için yeni bir sözlük temelli metin sınıflandırma yaklaşımı önermiştir. Bir kimya sözlüğü kullanma yaklaşımı ile web sayfalarından ilgili bilgilerin daha kesin olarak çıkarabileceği vurgulanmıştır. Sözlük terimlerini bulmak için belgelerde otomatik bölümlendirme işlemi gizli anlamsal indeksleme kullanılarak yapılmış ve ardından metin sınıflandırması algoritması olarak KNN kullanılmıştır. Bu çalışmada, ek bir sözlük kullanmanın etkileri ve verimliliği ele alınmış, ayrıca farklı veri toplama yöntemleri denenmiş ve sınıflandırma performansını geliştirmede rol oynayan yeni bir oylama yöntemi önerilmiştir. Deneysel sonuçlar; önerilen yaklaşımın, geleneksel sınıflandırma yöntemine göre üstün bir performans gösterdiğini ve kimya konulu İnternet sayfalarının sınıflandırılmasında uygulanabileceği göstermektedir.

Kotevska, Padi ve Ll bath [26] günümüzde Twitter, Facebook ve benzeri sosyal ağların ve diğer mikro-blog sitelerinin günlük yaşamda veri üretimi üzerinde büyük bir etkisi olduğunu belirtmiştir. Mikro-bloglama aracılığıyla yayınlanan değerli veriler, zamanında yakalanır ve doğru şekilde analiz edilirse farklı durumlara faydalı bilgiler sağlayabileceği vurgulanmıştır. Yazarlar, akıllı şehir problemleri için ise, Twitter üzerinden iletilen mesajları otomatik olarak belirleme işlemi, şehirle ilgili durum farkındalığına katkıda bulunabilir ve aynı zamanda şehir hakkında bilgi arayanlar için birçok yararlı bilgi ortaya çıkarabileceğini belirtmişlerdir. Bu makalede, özellikle Twitter verileri kullanılarak NLP ve Rastgele Orman sınıflandırıcısının birlikte kullanıldığı bir otomatik sınıflandırma yöntemi önerilmiştir. Twitter mesajlarının işlenmesinin zor bir iş olduğu belirtilip, twitter mesajlarının otomatik olarak ön işleme tabi tutulması için bir algoritma önerilmiştir. Bunun için bir coğrafi konuma göre on altı farklı kategori için Twitter mesajları toplanıp, mesajları ön işleme koymak için önerilen algoritma kullanılmıştır. Rastgele Orman sınıflandırıcısını kullanarak bu tweetler otomatik olarak önceden tanımlanmış kategorilere ayrıştırılmıştır. Rastgele Orman sınıflandırıcısının SVM ve Naive Bayes sınıflandırıcılarından daha iyi performans ortaya koyduğu gösterilmiştir.

Chen, Cheng ve Cheng [27] tarafından bir İnternet sitesinde farklı başlıklara sahip büyük haber raporlarının bulunmasının yaygın bir durum olduğu ve gıda güvenliği haberlerinin raporlarının daha açıklayıcı bir analizi için kullanılması gerekliliği ve bunun bir metin sınıflandırma problem olduğu belirtilmiştir. Bu çalışmada, toplu öğrenme temelli Uzun ve Kısa Vadeli Hafıza Ağları (Long-Short Term Memory-LSTM) kullanarak bir gıda

güvenliği belge sınıflandırma yöntemi önerilmiştir. İlk olarak, insan eliyle yapılan işlemin yüksek maliyeti nedeniyle, gıda güvenliği doküman metni sadece bir sınıf örnek içerir ve böyle bir metin üzerinde yapılan gıda güvenliği dokümanı sınıflandırması tek sınıf sınıflandırma problemidir. Bu çalışmada, negatif örnek olarak bildirilen çok sayıda etiketsiz haber (gıda güvenliği ile ilgili olmayan belgeler) metin genişletme yaklaşımında kullanılmıştır. Bu sayede gıda güvenliği ile ilgili bir haber metni, ikili sınıfı temel alan hem pozitif örnekler hem de negatif örnekler olan bir metne dönüştürülmüştür. Bu dönüşüm veri gürültüsü ve veri dengesizliği gibi problemler yaratmıştır. Belge sınıflandırma işlemi toplu öğrenme tabanlı LSTM modeline dayanmaktadır. Genel olarak, LSTM tabanlı topluluk öğrenme yöntemine dayalı sınıflandırma metodu ile İnternet sitelerinden gıda güvenliği içeren metni yüksek performansla otomatik olarak saptayabildiği vurgulanmıştır.

Qazi ve Guadar [28], İnternet belgesinin sınıflandırılması işinin, İnternet sayfalarının sayılarının büyüklüğündeki yüksek artış nedeniyle oldukça önemli bir araştırma problem olduğunu dile getirmiştir. Bu makale, İnternet sayfalarının sınıflandırılmasında yeni ve etkili olan ontoloji temelli bir terim ağırlıklandırma tekniği önermektedir. Ontoloji, herhangi bir alan için bilgi temsiliinin merkezini oluşturur. Önerilen yaklaşım etki alanı ontolojisini oluşturur ve tahmin performansını önemli ölçüde artıran özellikleri seçmektedir. Alan tabanlı İnternet sayfalarında deneyler yapılmış ve sınıflandırma performansı, teknik sınıflandırma algoritmalarıyla hesaplanmıştır. Deneysel analizler, önerilen yaklaşımın geleneksel anahtar kelime temelli yaklaşımlara kıyasla çok daha iyi sonuçlar verdiğini göstermektedir.

2.2. İnternet Sitesi Sınıflandırma Çalışmaları

Cuzzola ve arkadaşları [11] çalışmalarında, alışveriş sitelerine ilişkin bir sınıflandırma yapmayı amaçlamıştır. Bunun için İnternet sitelerini sınıflandırırken, Çok Sınıflı sınıflandırma yerine, ikili sınıflandırma teknikleri kullanmışlardır. Bu teknikler; Sinir ağları sınıflayıcısı ve destek vektör makineleridir. Yazarlar, verdikleri deneysel sonuçlara göre sinir ağlarının daha iyi sınıflandırma performansı göstermektedir.

Hernández ve arkadaşları [29] çalışmalarında, İnternet sitesi sınıflandırma problemi için gözetimsiz sınıflandırma teknikleri kullanılmıştır. Bu sayede önceden belirlenen bir sınıf

veya bir örnek olmadan site içerisindeki metinden kendi bir küme oluşturmaları beklenmiştir. Araştırmacılar farklı gözetimsiz sınıflandırma tekniklerini denemiş ve yorumlamıştır.

Li ve arkadaşları [30] hızla büyüyen İnternet siteleri ve içerikleri sonrasında, kullanıcı ilgileri ile İnternet sayfasının içeriği ile ilgili olan eşleşme sınıflandırması yöntemleri zorlaştığı vurgulamıştır. İnternet sayfasındaki kelimeler (“N-gram with Wikipedia Entity Words” veri seti kullanılarak) ana metinden, başlıktan çıkarılıp Bayes sınıflandırıcısı ile tahmin edilmeye çalışılmıştır.

Lee, Yeh ve Chuang [31] tarafından yürütülen çalışmada da, İnternet sayfalarındaki yararlı bilgileri hızlı bir şekilde almak için sınıflandırma yöntemleri kullanılmıştır. Bunun için eğitim veri kümesindeki her özellik için en iyi ağırlıkları öğrenmek ve test veri setindeki yeni İnternet sayfalarını sınıflandırmak için en iyi ağırlıkları benimsemeyen yeni bir basitleştirilmiş sürüler optimizasyonu (Simplified Swarm Optimization – SSO) önerilmektedir. Ayrıca, parametre ayarları, SSO'nun güncelleme mekanizmasında önemli bir rol oynar, böylece parametre ayarlarını belirlemek için bir Taguchi yöntemini kullanılmıştır. Algoritmanın etkililiğini göstermek için, performansını iyi bilinen Genetik Algoritma (GA), Bayes sınıflandırıcısı ve KNN sınıflandırıcıları ile dört veri kümesine göre karşılaştırılmıştır.

Rekik ve arkadaşları [3] bu çalışmalarında, bir İnternet sitesinin içeriğinin sınıflandırılması yapılmadan kullanıcıya ihtiyaçları sağlayamayacağı vurgulamıştır. Çalışma iki aşamalı olup ilk aşamasında sistematik literatür taraması yapılmıştır. Çalışmanın devamında sayfanın sınıfı ile değerlendirme ölçütleri arasında ilişkiyi bulmak için önsel algoritma denilen metin madenciliği yöntemi kullanılmıştır.

Gali, Meriescu-Istodor ve Fränti [32], İnternet sayfalarını sınıflandırmak adına yapılan çalışmaların büyük bir çoğunluğu, Hiper Metin İşaretleme Dili (Hypertext Markup Language – HTML) yapılarından çıkarılan yazı ve yapısal bilgilere dayalı yapılmakta olduğunu, ancak bu yaklaşımın İnternet sayfasında görüntülenen reklamlardan ötürü yanıltıcı olabileceğini öne sürmüştür ve bu sorunun çözümü için istatistiksel özellikler, dil bilgisi, metin bölümlendirme metotlarını belirtmiştir. Naive Bayes, KNN, SVM gibi

yöntemlerle değerlendirilen bu yaklaşımın, metin analizi ve gruplandırılması açısından önem taşıdığı vurgulanmaktadır.

Shen, Yang ve Chen [33], İnternet sayfalarına yerleştirilmiş olan çok çeşitli gürültülü (noise) bilgiler nedeniyle, İnternet sayfası sınıflandırmasının saf metin sınıflandırmasından çok daha zor bir problem haline geldiğini belirtmiştir. Bu çalışmada, özetleme teknikleri ile gürültüyü ortadan kaldırarak İnternet sayfası sınıflandırma performansını iyileştirme önerilmiştir. Öncelikle, insan editörler tarafından oluşturulan ideal İnternet sayfası özetlerinin, İnternet sayfası sınıflandırma algoritmalarının performansını gerçekten artırabileceğine dair ampirik kanıtlar sunulmuştur. Daha sonra, İnternet sayfası düzenine dayanan yeni bir İnternet sayfası özetleme algoritması ortaya koyulup, LookSmart Web dizinindeki diğer birçok modern metin özetleme algoritması ile birlikte değerlendirilmiştir. Deneysel sonuçlar, herhangi bir özetleme yaklaşımının arttırdığı sınıflandırma algoritmalarının (Naive Bayes veya SVM), saf metin bazlı sınıflandırma algoritmalarına kıyasla % 5'ten daha fazla bir gelişme sağlayabildiğini göstermektedir. Ayrıca farklı özetleme algoritmalarını birleştirmek için bir topluluk yöntemi sunulmuştur. Topluluk özetleme yöntemi, saf metin tabanlı yöntemlere göre %12'den daha fazla gelişme sağlandığı gözlenmiştir.

De Souza ve arkadaşları [34], otomatik çok etiketli metin kategorizasyonunda, otomatik bir kategorizasyon sistemi, analiz edilen her belge için boyutu önceden bilinmeyen bir etiket seti çıkarılmasını gerektiğini belirtmiştir. Birçok makine öğrenmesi tekniğinde, bu tür otomatik metin sınıflandırma sistemlerinin oluşturulması için birçok teknik kullanıldığı vurgulanmıştır. Bu araştırmada Sanal Genelleme Rasgele Erişim Belleği Ağırlıksız Sinir Ağlarını (Virtual Generalizing Random Access Memory Weightless Neural Networks - VG-RAM WNN) incelenmektedir. Bu algoritma; otomatik çoklu etiketli metin sınıflandırma sistemleri oluşturmak için bir araç olarak basit uygulama ve hızlı eğitim ve test sunan etkili makine öğrenme tekniğidir. VG-RAM WNN'in iki gerçek dünya problemindeki performansını değerlendirildiğinde;

- (i) ekonomik faaliyetlerin serbest metin açıklamalarının sınıflandırılması
- (ii) Web sayfalarının kategorilendirilmesi

ile elde edilen sonuçları çok etiketli tembel öğrenme yaklaşımı (Multi-Label KNN, ML-KNN) ile karşılaştırılmıştır. Bu deneysel karşılaştırmalı analiz sonucunda, ortalama

olarak, kategorizasyon performansının VG-RAM WNN'nin ML-KNN'den daha iyi performans gösterdiğini veya benzer olduğunu göstermiştir.

Yu, Xu ve Li [35] yapmış oldukları çalışmada, yeni metin kategorizasyon modellerinde kullanılan Geri Yayılım Sinir Ağı (Back Propagation Neural Networks – BPNN) ve Değiştirilmiş Geri Yayılma Sinir Ağı (Multilayer Back Propagation Neural Networks – MBPNN) önermişlerdir. Etkili bir özellik seçme yöntemi, boyutlandırmayı azaltmak ve performansı iyileştirmek için kullanılır. Temel BPNN öğrenme algoritması yavaş eğitim hızının dezavantajına sahiptir, bu yüzden eğitimi hızlandırmak için temel BPNN öğrenme algoritmasını değiştirmeyi önermişlerdir. Sonuç olarak kategorizasyon başarımları da geliştirilmiştir. Geleneksel kelime eşleme tabanlı metin sınıflandırma sistemi, dokümanı temsil etmek için vektör uzay modelini (VSM) kullanır. Ancak, dokümanı temsil etmek için yüksek boyutlu bir alana ihtiyaç duyar ve terimler arasındaki anlamsal ilişkiyi hesaba katmaz, bu da zayıf sınıflandırma problemine yol açabilir. Gizli anlambilimsel analiz (Latent Semantic Analysis – LSA) ile tek tek kelimeler yerine istatistiksel olarak elde edilen kavramsal endeksleri kullanarak ortaya çıkan problemlerin üstesinden gelebileceği düşünülmüştür. Her terim veya belgenin uzayda bir vektör olarak temsil edildiği kavramsal bir vektör uzayı oluşturmaktadır. Sadece boyutsallığı büyük ölçüde azaltmaz, aynı zamanda terimler arasındaki ilişkisel önemli olanları da keşfeder. Kategorizasyon modelleri 20 haber grubu veri setinde test edilmiştir. Deneysel sonuçlarda, MBPNN kullanan modeller temel BPNN'den daha iyi performans göstermiştir. Sistem için LSA'nın uygulanması, iyi sınıflandırma sonuçları elde ederken önemli ölçüde boyutsallık azalmasına neden olduğu gözlenmiştir.

Lin ve Zongpeng çalışmalarında [36], İnternet sayfalarının sayısı keskin bir şekilde arttıkça, İnternet madenciliği ve bilgi alma gibi bazı alanlarda İnternet sayfası sınıflandırması daha da önem kazandığını vurgulamışlardır. Ancak, geleneksel metinsel sınıflandırıcılar genellikle birçok el yapımı özelliğe dayandığı tatmin edici sonuçlar vermediği belirtilmiştir. Hedef HTML belgesinin basitleştirilmiş versiyonuna dayanarak İnternet sayfası sınıflandırma problemi için nispeten derin bir kalıntı sinir ağı (deep residual neural network) önerilmiştir. Birkaç gelişmiş öğrenme tekniğini birleştiren optimal model, parametrelerle birlikte 20 sinir katmanına sahiptir ve uçtan uca farklılaştırılabilir. Ayrıca, akrobaların Web sayfalarından gelen sınıf bilgisini kullanmak için bir üst özyineli sinir ağı (Recurrent Neural Network – RNN) sınıflandırıcı

sunulmuştur. Bunu göstermek için iki büyük ölçekli veri seti oluşturulmuştur. ResNet-20 ve üst RNN tasarımı, birkaç temel yöntemle kıyasla en iyi veya gelecek vaat eden sonuçları elde edebileceği vurgulanmıştır.

Chen ve Hsieh [19] araştırmalarında, geleneksel bilgi alma yöntemini uygulamışlardır. Bu yöntemle belgelerin sınıfını belirlemek için belgelerde bulunan anahtar kelimeler kullanıldığında, genellikle ilgisiz İnternet sayfalarının ortaya çıkacağı belirtilmiştir. Eş anlamlı anahtar kelime problemini çözen web sayfalarını etkili bir şekilde sınıflandırmak için ağırlıklı oy şeması kullanan SVM modeline dayanan İnternet sayfa sınıflandırması önerilmiştir. Bu sistem hem gizli anlamsal analiz, hem İnternet sayfası özellik seçme eğitimi hem de SVM modeli ile tahmin yapabilmektedir. Anahtar sözcükler ve belgeler arasındaki anlamsal ilişkileri bulmak için gizli anlam analizi kullanılmıştır. Gizli anlambilimsel çözümleme, anahtar kelimeler arasındaki ve belgeler arasındaki anlamsal ilişkileri bulmak için kullanılır. Gizli anlambilimsel analiz yöntemi, belgede gizli bilgileri bulmak için terimleri ve bir belgeyi vektör uzayına yansıtır. Aynı zamanda, İnternet sayfasının içeriğinden de metin özellikleri çıkarılmıştır. Metin özellikleri sayesinde, web sayfaları uygun bir kategoride sınıflandırılır. Bu iki özellik sırasıyla eğitim ve test için SVM'ye gönderilir. SVM'nin çıktısına dayanarak, web sayfasının kategorisini belirlemek için bir oylama şeması kullanılır. Deneysel sonuçlar, çalışmada kullanılan yöntemin geleneksel yöntemlerden daha etkili olduğunu göstermektedir.

Yu, Han ve Chang [37] çalışmalarında, ilgili bir İnternet sayfası sınıfı için bir sınıflandırıcı oluşturmak, pozitif ve negatif eğitim örnekleri toplamak gibi zahmetli bir ön işleme gerektirdiğini vurgulamışlardır. Örneğin, bir “anasayfa” sınıflandırıcısı oluşturmak için bir tane bir anasayfa örneği (pozitif örnekler) ve anasayfa dışı bir örnek (negatif örnekler) toplanmalı ve özellikle, olumsuz eğitim örneklerinin toplanmasındaki zorluluktan ve yanlışlık yaratmaktan kaçınmak için dikkatli olmak gerektiği belirtilmiştir. Bu çalışmada, Olumlu Örnek Tabanlı Öğrenme (Positive Example Based Learning – PEBL) ile manuel olarak negatif toplama ihtiyacını ortadan kaldıran İnternet sayfası sınıflandırması için önışlemede eğitim örnekleri bir çerçeve sunulmuştur. PEBL, sınıflandırmada yüksek başarımla elde etmek için (pozitif ve etiketsiz verilerle), geleneksel bir SVM'ninkinden daha yüksek (pozitif ve negatif verilerle) Eşleşme – Yakınsama (Mapping-Convergence M-C) adlı bir algoritma uygulanmıştır. Eşleşme aşamasında, algoritma “güçlü” negatif verinin ilk yaklaşımını çizen zayıf bir sınıflandırıcı

kullanılmıştır. İlk yaklaşıma göre, yakınsama aşaması, negatif verinin yakınlaştırılmasını kademeli olarak iyileştirmek için marjları maksimize eden içsel bir sınıflandırıcı (örneğin SVM) olarak yinemeli çalıştığı ve sonunda, özellik alanındaki pozitif sınıfın gerçek sınırına yakınlığı belirtilmiştir. M-C algoritmasını teorik ve deneysel gerekçeler baz alınarak sunulmuştur. Yapılan deneyler, aynı pozitif örnekler verildiğinde, M-C algoritmasının tek sınıf SVM'lerden daha iyi performans gösterdiği ve neredeyse geleneksel SVM'ler kadar başarımlı olduğunu göstermektedir.

Riboni [38] yaptığı çalışmada, İnternet sayfasının sınıflandırması probleminin, HTML yapısı ve köprüler tarafından sağlanan bazı ek bilgilerin varlığı nedeniyle geleneksel metin sınıflandırmalarından önemli ölçüde farklı olduğu vurgulanmıştır. Bu çalışmada kategorizasyon başarımlarını arttırmak için, bu özellikleri analiz edilip, web sayfalarını temsil etmek için kullanmaya çalışılmıştır. Top Yahoo! kategorilerine ait 8.000 belgeden oluşan bir grup üzerinde Kernel Perceptron ve Naive Bayes sınıflandırıcıları kullanarak çeşitli deneyler yapılmıştır. Deneyler boyutsallık azaltmanın ve yeni yapıya yönelik bir ağırlıklandırma tekniğinin kullanışlılığını göstermektedir. Bağlantılı sayfaları temsil etmek için hipermetin kategorisini gerçek zamanlı uygulamalar için uygun kılan yerel bilgileri kullanarak yeni bir yöntem de sunulmuştur. Son olarak, web sayfalarının normal temsili için yerel sözcükleri kullanarak köprü metni ile bir arada kullanılmasının, sınıflandırma performansını artırabileceğini gözlemlenmiştir.

Mladen [39] çalışmasında, Yahoo hiyerarşisine dayanan otomatik İnternet sayfası sınıflandırmasına yönelik bir yaklaşım önermiştir. Metin verilerini öğrenmek için geliştirilen makine öğrenme teknikleri burada hiyerarşik sınıflandırma anlayışıyla kullanılmaktadır. Hiyerarşik yapı göz önünde bulundurularak ve bilgi alımından bilinen yöntemeye dayanan özellik alt kümesi seçimi kullanıldığında yüksek sayıda özellik azaltılır. Belgeler, metin verilerini öğrenirken yaygın olarak kullanılan tek kelimeleri (unigramlar) dahil etmek yerine n-gram içeren özellik vektörleri olarak gösterilir. Hiyerarşik yapıya dayanarak ele alınan problem, her biri Yahoo hiyerarşisinde yer alan kategorilerden birini temsil eden alt problemlere bölünür. Öğrenmenin sonucu, her biri yeni bir örneğin ilgili kategorideki bir üye olma ihtimalini tahmin etmek için kullanılan bir dizi bağımsız sınıflandırıcıdır. Gerçek veriler üzerine yapılan deneysel değerlendirme, önerilen yaklaşımın iyi sonuçlar verdiğini göstermiştir.

Shih ve Karger [40] alışmalarında, İnternet sayfasını sınıflandırma işini otomatikleştirmek için içerik önerisi ve reklam engelleme gibi görevlerini yerine getiren yeni özellikler ve algoritmaları vurgulamıştır. İnternet sayfalarının otomatik sınıflandırmasının, metin içeriğine bakmak yerine, her bir bağlantının birörnek kaynak konumlayıcısını (Uniform Resource Locator – URL) ve bu bağlantıların sayfadaki görsel yerleşimini göz önünde bulundurmasını önermiştir. Bu özellikler kelime sayıları gibi skaler ölçümlere göre olağan dışıdır. Bu ağaç yapılı özellikleri kullanan makine öğrenmesi algoritma için bir model geliştirilmiştir. Bu alışmada bir okuyucuya “ilginç” haber hikayeleri önermek ve İnternet reklamlarını engellemek için otomatik araçlar kullanılmıştır. Deneysel sonuçlara göre, geliştirilen algoritmanın metin içeriğine göre yapılan sınıflama işleminden hem daha hızlı hem de daha yüksek başarımlı olduğunu göstermiştir.

Osanyin, Oladipupo ve Afolabi [41] alışmalarında, İnternet’te depolanan dijital belgelerin ve kaynakların artması nedeniyle, bu kaynakların düzgün bir şekilde sınıflandırılmasına ihtiyaç duyulduğunu belirtmişlerdir. Bu tarz doğru bir sınıflandırmanın, bireylerin aradıkları kaynakları çok daha kolay filtreleyip analiz edebilmesine olanak sağladığını dile getirmişlerdir. Bu kaynakların doğru sınıflandırmasının, ancak kategorilerine doğru ayrılmış bir eğitim kümesi kullanılarak mümkün olduğunu söylemişlerdir ve bu veri setinin kalitesi, ön işleme tekniklerine doğrudan bağlıdır. Ayrıca, metin dökümanlarının gösterim teknikleri de sınıflandırma fonksiyonunu etkileyen faktörlerden biridir. Bu alışmada, farklı döküman gösterim teknikleri ve sınıflandırma algoritmaları kullanılarak, sınıflandırma performansı incelenmiş ve çok boyutlu verilerin gösterimi için özüm üretilmiştir.

Ren ve arkadaşları [42] yaptıkları alışmada, derinlikli SVM’nin sınıflandırma algoritmasını yüksek boyutlu hesaplamalar için incelemişlerdir. Bu hesaplamalar için, iki farklı kernel fonksiyonunu hedef alan gelişmiş bir SVM önerilmiştir. Önerilen kernel fonksiyonları; global ve lokal kernel kullanılarak yeni bir kernel fonksiyonu yaratılmış ve bu fonksiyonun parametrelerini optimize etmek için genetik algoritmalarından yararlanılmıştır. Bu sayede, tek kernel fonksiyonunun sınırlandırıcı etkisi düzeltilmiştir ve metin sınıflandırma problemi için oluşturulan SVM modelinin genelleştirme ve öğrenme yetenekleri artmıştır. Toplanan benzetim verileriyle de önerilen algoritmanın metin sınıflandırma üzerindeki olumlu etkisi kanıtlanmıştır.

3. METODOLOJİ

3.1. Veri Tanımı ve Toplanması

Çalışma için, önceden belirlenen ve EK1’de verilen sınıflara ait İnternet sayfalarındaki kelime ve kelime grupları veri olarak kabul edilip, bu çerçevede veri toplama işlemi gerçekleştirilmiştir. Sınıf belirleme süreci, algoritma sonunda denenebilecek tüm İnternet sitelerine en iyi şekilde temsil edebilecek bir sınıf atanabilmesi için uzman görüşü ve detaylı çalışmalar ile olabilecek en geniş sınıf sayısı (63 İnternet sayfa sınıfı) belirlenmiştir.

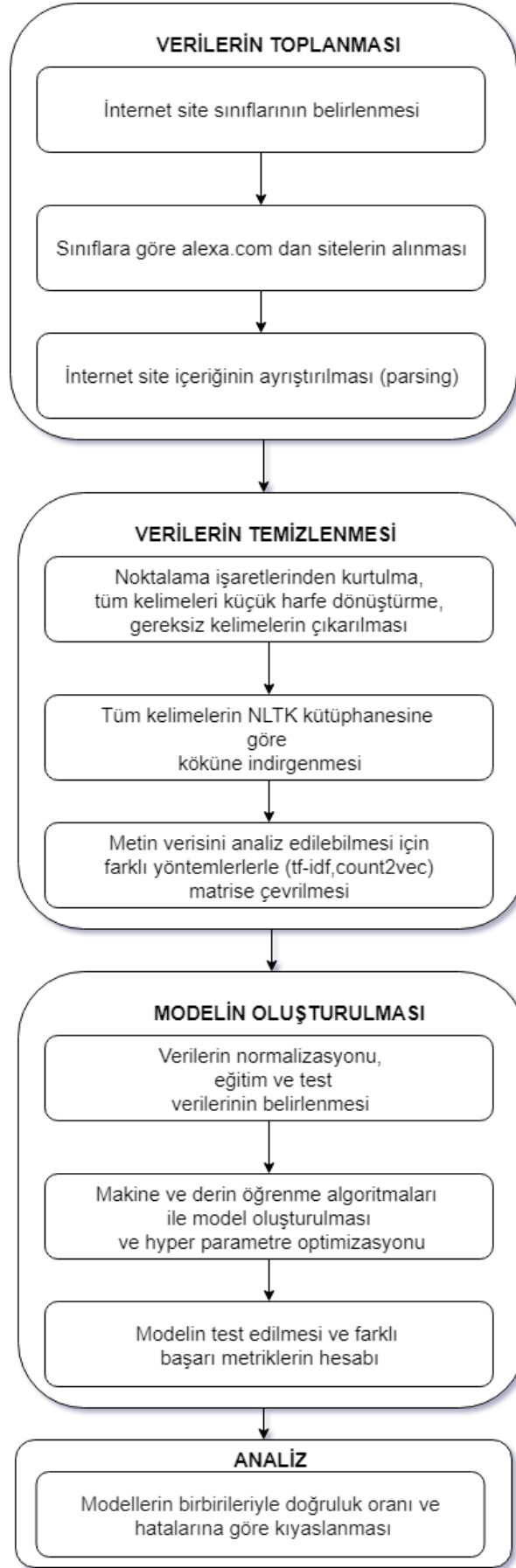
Veri toplama işlemi için Amazon’a ait Alexa Top Sites¹ hizmetinden yararlanılmıştır. Alexa Top Sites, Alexa Traffic Rank algoritmasına göre en yüksek performanslı İnternet sitelerinin listelerini sunan bir Amazon Web Hizmetidir (Amazon Web Service). Önceden belirlenen 63 tane sınıfa ait 45.543 adet İnternet sayfası Alexa’dan alınıp bu sayfalardaki kelime ve kelime grupları veri olarak kabul edilmiş, çalışma kapsamında geliştirilmiş Python betikleri ile veri toplama işlemi otomatik olarak gerçekleştirilmiştir. Çizelge 1’de, tanımlanmış olan 63 sınıf arasından en fazla kelime sayısına sahip olan ilk 10 İnternet sayfa sınıfı, o sınıf içinde örnek olarak kullanılmış olan İnternet site sayıları ve toplam kelime sayıları ile birlikte verilmiştir.

Hem ikili sınıflandırma, hem de çok sınıflı sınıflandırma yaklaşımlarında, eğitim setinde, veri setinin yaklaşık %83’lük kısmı olan 37.814 İnternet sitesi kullanılmışken, test sırasında veri setinin yaklaşık %17’lük kısmı olan 7.729 tane İnternet site verisi (içerik metni) kullanılmıştır. Her model için aynı eğitim ve test örnekleri kullanılmıştır İkili sınıflandırma sürecinde, eğitim ve test veri setleri 10 kez karıştırılıp tekrar model oluşturulmuştur. Sınıflandırma sürecinde takip edilmiş olan adımlar detaylı olarak Şekil 3’de verilmiştir.

¹<https://www.alexa.com/topsites>

Çizelge 1. Ön işleme süreci sonrası en fazla toplam kelime sayısı sahip olan ilk 10 sayfa sınıfı ve toplam kelime sayıları (kısmi örnek)

	İnternet Site Sayısı	Toplam Kelime Sayısı
Political Issues	1.000	534.330
Community and Society	1.000	420.280
Education-Reference	1.000	411.976
Food-Drink	1.000	393.388
News	1.000	384.178
Religion-Spirituality	1.000	384.135
Paranormal	1.000	377.745
Adult Content	1.000	372.913
Travel	1.000	352.077
Software-Hardware	1.000	350.071



Şekil 3. Çalışmada Uygulanan Sınıflandırma Süreci

3.2. Verinin Ön İşleme Süreci (Veri Temizleme)

Makine ve derin öğrenme teknikleri uygulanmadan önce veriyi temizleme ve ön işleme süreçleri yapılmıştır. Sırası ile noktalama işaretlerinden kurtulma, etkisiz kelimelerin çıkarılması, tüm kelimeleri küçük harfe dönüştürme ve kelimenin kök öbeğine ulaşma işlemleri yapılmıştır. Bu çalışmada etkisiz kelimeler, o dilde içerikten bağımsız kelimeler, bağlaçlar, imleçler, sayılar, kalıplaşmış kısaltmalardır. Uygulama sadece İngilizce dili içeren İnternet siteleri üzerinde olacağı için İngilizce’de “stopwords” olarak geçen etkisiz “the”, “a”, “an”, “of” ve benzeri kelimeler çıkartılmıştır. Bu ön temizleme ve işleme sürecinde Python içerisinde bir Doğal Dil İşleme Kütüphanesi olan Natural Language Tool Kit (NLTK) paketi kullanılmıştır [43].

Ham verinin ön temizleme işlemi olan etkisiz kelimeler ve noktalama işaretlerden kurtulduktan sonra kelimelerin köküne ulaşılan örnek veri seti Çizelge 2’deki gibidir.

Çizelge 2. Ön işleme süreci sonrası elde edilen kelime kökleri (kısmi örnek)

'abil'	'academi'	'access'	'accessori'	'accommod'	'achiev'	'action'	'activ'	'addit'
'address'	'admin'	'administr'	'adult'	'advanc'	'advantag'	'adventur'	'advertis'	'advic'
'affili'	'afford'	'africa'	'age'	'agent'	'ago'	'agre'	'air'	'airlin'
'america'	'american'	'analysi'	'android'	'angel'	'airport'	'alaska'	'amaz'	...

Ön işleme sürecinden sonra, sınıfları bilinen İnternet sitelerine ait veriler (kelime ve kelime grupları) öğrenme ve test verisi setleri olarak paylaştırılmıştır. İkili sınıflandırma sürecinde, bu işlem eğitim setine aşırı uyumu (over-fiting) en aza indirmek için farklı oranlar denenerek belirlenmiştir. Eğitim ve test veri setleri aynı paylaşım oranı kullanılarak rastgele seçimlerle 10 tekrar ile farklı modeller oluşturulmuştur. Her bir model çapraz doğrulamaya benzer bu yöntem ile başarımları karşılaştırılmıştır.

3.3. Kelime Vektörleştirme Yöntemleri (Word Embedding)

Kelime vektörleştirme yöntemleri, makine ve derin öğrenme süreçlerinden önce uygulanan bir dönüşüm işlemidir. Bu dönüşüm ile birlikte metin farklı yöntemlere göre kendi sınıfı içerisinde sayısal veriler içeren vektörlere dönüştürülür [12]. Bu işlemin

amacı yöntemine göre hem kelimeleri kendi arasında yakınlığına göre sayısal değer atamak, hem de sadece sadece kelimenin sınıf içerisinde sıklığını belirlemektir. Sonuç olarak, metin verisi algoritmalar tarafından eğitilebilecek ve analiz edilebilecek sayısal değerlere dönüştürülmektedir.

Kelime Çantası, TF-IDF ve Word2Vec kelime vektörleştirme algoritmaları, Çok Sınıflı Sınıflandırma yöntemlerinin her birinde ayrı ayrı modellenmiş ve her birinin başarımları kıyaslanmıştır.

3.3.1. Terim Sıklığı-Ters Metin Sıklığı (Term Frequency – Inverse Document Frequency, TF-IDF)

Ağırlık faktörleri kullanılarak bir kelimenin herhangi bir İnternet sitesindeki sıklığı pozitif etkili ve diğer sitelerdeki sıklığı negatif etkili olacak şekilde bir katsayı oluşturur. Başka bir deyişle; bir kelime bir İnternet sitesinde ne kadar çok bulunuyorsa, bu kat sayı o İnternet sayfa sınıfı için önemi artacak iken aynı kelime diğer İnternet sayfa sınıflarında da bulunması durumunda kelimenin İnternet sayfa sınıfları temsili için önemi azalacaktır [44]. Çizelge 3’de örnek olarak bazı İnternet site sınıfları ve kelimelere ait TF-IDF sonuçları mevcuttur.

$$\text{normalleştirilmiş terim sıklığı} = \frac{tf(t, d)}{nd}$$

t : Veri setindeki bir kelime

$tf(t, d)$: d sınıfındaki kelime sayısı.

nd : d sınıfındaki toplam kelime sayısı.

$$\text{ters metin sıklığı (idf)} = \log\left(\frac{nd}{nd(t)}\right)$$

nd : Toplam sınıf sayısı.

$nd(t)$: t kelimesinin geçtiği toplam sınıf sayısı.

Son olarak bu iki terimin çarpılmasıyla Denklem 2’de bulunan TF-IDF Skoruna ulaşılır.

$$Tf\text{idf} = \frac{tf(t,d)}{nd} \times \log\left(\frac{nd}{nd(t)}\right) \quad (2)$$

3.3.2. Kelime Torbası (Bag Of Words - BOW)

İnternet sitelerinin içerisindeki tüm benzersiz kelimelerin sayısı ve İnternet sitelerini toplam sınıf sayısı olacak şekilde bir matris oluşturur. Hangi sınıfta o kelime var ise matristeki ilgili hücreye 1 değeri, olmayan durumda ise 0 değeri atanır [14]. Bu işlemle kelimenin ilgili sınıftaki sıklığı gösterilmektedir. Torba olarak adlandırılmasının sebebi ise, bu algoritmanın kelimenin İnternet sitesi içeriğinde bulunma sırasıyla ilgilenmemesidir. Sadece kelimenin sıklığı ile ilgilenir. İsteğe göre tekli ikili ve daha fazla kelime grupları (N-gram) oluşturulabilir. Çizelge 4’de de görüldüğü gibi çalışmada sadece tekli kelimeler kullanılmıştır.

3.3.3. Word2Vec

Word2Vec; Google tarafından geliştirilmiş, Google News’daki 100 milyar kelime ön-eğitim ile eğitilmiş NLP paketidir. Kelimeler arasındaki anlamsal uzaklık göz önüne alınarak oluşturmuş vektör temsilleri kümesidir. Bu vektör temsilleri kullanılarak bir İnternet sitesi içerisindeki kelimelerin matematiksel gösterim vektörü öğrenim setinde kullanılmıştır [12].

Bu yaklaşım kullanılmadan önce her bir İnternet sayfa sınıfı temsil edecek 100 kelime çıkarılmıştır. İnternet site içerisindeki her bir kelime; Word2vec’den yararlanılarak vektörleştirilmiştir. Bu çıkarılan 100 kelime ile olan uzaklıklık (Euclidean Distance-Cosine Similarity) ortalamaları atanmıştır.

3.4. Sınıflandırma Sırasında Kullanılan Algoritmalar

Tüm ön-işleme uygulamaları yapıldıktan sonra ham veri analiz edilebilir hale getirilmiştir. Bu süreçten sonra; üzerinde çok defa farklı alanlarda da çalışılmış makine öğrenmesi ve derin öğrenme algoritmaları seçilmiş ve bu algoritmalar, İkili ve Çok Sınıflı Sınıflandırıcı modellerini oluşturmak için kullanılmıştır.

Çizelge 3. İnternet sınıfları ve kelimelerin TF-IDF skorları (kısmi örnek)

	abil	academ	academi	access	accessori	achiev	act	action	activ	addit	...
Adult Content	0,003468	0,000874	0,001573	0,028476	0,040607	0,006590	0,020703	0,020924	0,016497	0,012595	...
Alcohol-Tobacco	0,001677	0,000136	0,000868	0,013931	0,017733	0,003383	0,005026	0,002945	0,009980	0,011390	...
Anime-Manga-Comics	0,004780	0,000452	0,003943	0,011872	0,005330	0,003045	0,014356	0,034390	0,013317	0,016696	...
Books-Literature	0,005859	0,007322	0,007559	0,014240	0,001037	0,009064	0,013548	0,011253	0,014644	0,014156	...
Business	0,009075	0,005781	0,004826	0,042176	0,036033	0,016430	0,008314	0,011226	0,024627	0,037444	...
Celebrity Fan-Gossip	0,002053	0,000764	0,005922	0,018643	0,007764	0,004171	0,012801	0,008623	0,008357	0,008489	...
Chat-Messaging	0,011653	0,000685	0,000000	0,047820	0,003479	0,001869	0,003623	0,012483	0,026503	0,027506	...
Community and society	0,008430	0,007786	0,003199	0,036485	0,000501	0,021295	0,048013	0,099540	0,052405	0,019709	...
Computing-Technology	0,011360	0,008471	0,004888	0,055343	0,015427	0,014193	0,005510	0,011693	0,025782	0,025476	...
Cryptocurrency	0,004645	0,000000	0,002010	0,031419	0,000000	0,006606	0,001829	0,014408	0,018328	0,012411	...
Diet-Exercise	0,007066	0,004931	0,038792	0,032691	0,003808	0,022071	0,010261	0,023799	0,044473	0,018373	...
Education-Reference	0,005674	0,148186	0,022957	0,035307	0,002164	0,015086	0,008480	0,008604	0,042508	0,012660	...
Family-Parenting	0,006817	0,005344	0,006511	0,018871	0,001810	0,007367	0,008996	0,008927	0,048233	0,014878	...
...

Çizelge 4. Örnek çekilmiş İnternet sınıfları ve kelimelerin kelime torbası istatistikleri (kısmi örnek)

	abil	academ	academi	access	accessori	achiev	act	action	activ	addit	...
Adult Content	31	7	13	271	320	58	188	193	157	118	...
Alcohol-Tobacco	69	5	33	610	643	137	210	125	437	491	...
Anime-Manga-Comics	59	5	45	156	58	37	180	438	175	216	...
Books-Literature	109	122	130	282	17	166	256	216	290	276	...
Business	836	477	411	4136	2926	1490	778	1067	2415	3615	...
Celebrity Fan-Gossip	3	1	8	29	10	6	19	13	13	13	...
Chat-Messaging	19	1	0	83	5	3	6	21	46	47	...
Community and society	191	158	67	880	10	475	1105	2327	1264	468	...
Computing-Technology	274	183	109	1421	328	337	135	291	662	644	...
Cryptocurrency	5	0	2	36	0	7	2	16	21	14	...
Diet-Exercise	40	25	203	197	19	123	59	139	268	109	...
Education-Reference	473	11063	1769	3133	159	1238	718	740	3772	1106	...
Family-Parenting	94	66	83	277	22	100	126	127	708	215	...
Fashion-Beauty	63	17	322	395	6838	128	122	112	617	469	...
Finance-Investment	101	5	8	692	9	153	180	168	276	216	...
...

3.4.1. Rastgele Orman Algoritması (Random Forest)

Rastgele orman algoritması; karar ağaçlarının çoklu hali olan yönetimsel bir makine öğrenmesi yöntemi, regresör ve sınıflandırıcısıdır. Her bir karar ağacı, eğitim setinde verilen girdiye göre bütün rastgele düğümleri gezerek ve bu düğümlerdeki koşullara bağlı olarak ağacın en altındaki yaprağa ulaşır. Ağacın son düğümü olan yaprakta hedef olan cevap değişkeni bulunur [18]. Rastgele orman algoritmasında, test setindeki hatayı objektif olarak hesaplamak için ekstra bir çapraz validasyona gerek yoktur. Algoritma çalışma esnasında her ağacı, orijinal veriden farklı örnekler seçerek oluşturur. Oluşturulacak olan ağaçların genelde üçte birinde, rastgele seçilen örnek ağacın oluşturulma aşamasında dışarıda bırakılır. Bu sayede, ağaçlar oluşturulurken otomatik olarak test edilmiş olur. Bu rastgele örnekleme yöntemi, çeşitliliği artırarak ağaç içindeki düğümlere ait varyansı azaltır. Bu süreç aynı zamanda “özellik torbalama” (feature bagging) olarak da bilinmektedir.

Bu çalışma kapsamında, karar ağaçlarındaki düğümlerde kelime ve kelime grupları, yaprak düğümünde ise İnternet sayfasının sınıfı ile her bir ağaç eğitilmiştir. Bu eğitim sırasında rastgele orman algoritmasının parametreleri olan ağaç derinliği, ağaç sayısı, gini-entropy gibi ölçüt değişkenlerini ızgara araması (grid search) ile optimize edilmiştir. Çok sınıflı sınıflandırma probleminin çözümünde kullanılmıştır.

3.4.2. Naive Bayes Sınıflandırıcı

Popüler Bayes olasılık teoremine dayanan bir sınıflayıcı olan Naive Bayes sınıflandırıcıları, özellikle belge sınıflandırma ve istenmeyen e-posta tespit etmekte kullanılan, basit ama iyi performans gösteren bir algoritmadır [45].

Naive Bayes sınıflandırıcıları basit ama çok etkili oldukları bilinen doğrusal sınıflayıcılardır. Naive Bayes sınıflandırıcılarının olasılıksal modeli, Bayes teoremine dayanmaktadır ve “naive” sıfatı, bir veri setindeki özelliklerin birbiri arasında bağımsız olduğu varsayımından gelmektedir. Uygulamada, bağımsızlık varsayımı genellikle ihlal edilmektedir, ancak Naive Bayes sınıflandırıcıları bu gerçekçi olmayan varsayım altında çok iyi performans gösterme eğilimindedir. Özellikle küçük örneklem büyüklükleri için Naive Bayes sınıflandırıcıları daha güçlü alternatiflerden daha iyi performans gösterebilir [6]. Bu çalışmada, metnin

sınıflandırılmasına en uygun olan Çoklu Nominal Naive Bayes ve Çok Değişkenli Bernoulli Naive Bayes kullanılmıştır.

3.1.1.1 Bernoulli Naive Bayes Algoritması

Multinomial modeline bir alternatif, çok değişkenli Bernoulli modeli veya Bernoulli modelidir. Her bir İnternet sitesi için bir gösterge oluşturan, ikili bağımsızlık modeline eşdeğer olan Bernoulli Naive Bayes sınıflandırıcılar ile bir İnternet sitenin hangi sınıfa ait olduğunu göstermesi 1 ile aksi durum ise 0 ile gösterilir.

$P(t/c)$ olarak gösterilen Bernoulli modeli, İnternet sitesi içerisinde geçen bir ‘t’ teriminin ‘c’ İnternet sitesi sınıfına ait olma durumunu tahmin eder. Bu model ikili sınıflandırma yöntemi olduğu için, sadece terimin bir İnternet sitesi sınıfına ait olup olmama bilgisi önemlidir. Modeller ayrıca sınıflandırmada oluşmayan terimlerin nasıl kullanıldığına göre farklılık gösterir. Multinomial modelinde sınıflandırma kararını etkilemezler ama Bernoulli modelde terimlerin oluşmama olasılığı hesaplanırken katsayılandırılır. Bunun nedeni, sadece Bernoulli Naive Bayes modelinin, terimlerin bulunmamasını modellemesidir [17].

3.1.1.2 Multinomial Naive Bayes Algoritması

Multinomial Naive Bayes, metin belgeleri için tasarlanmış olan Naive Bayes'in özel bir şeklidir. Basit Naive Bayes, bir belgeyi belirli kelimelerin varlığı ve yokluğu olarak modellerken, matematiksel gösterimi Denklem 3’de olan Multinomial Naive Bayes kelimenin sıklığını göz önünde bulundurarak modeli tasarlar [45].

$$P(x_i|w_j) = \frac{\sum f(x_i, d \in w_j) + \alpha}{\sum N_{d \in w_j} + \alpha \cdot V} \quad (3)$$

x_i : Bir kelimeyi ifade eden vektör

$\sum f(x_i, d \in w_j)$ w_j sınıfına ait eğitim setinde x_i kelimesinin metin sıklığının toplamı

$\sum N_{d \in w_j}$: w_j sınıfı için eğitim setindeki bütün kelime sıklıklarının toplamı

w_j : j. İnternet sınıfı

d : İnternet sınıfları kümesi

α : Ek düzgünleştirme parametresi

V: Eğitim setindeki farklı kelime sayısı

3.4.3. Destek Vektör Makineleri (Support Vector Machine- SVM)

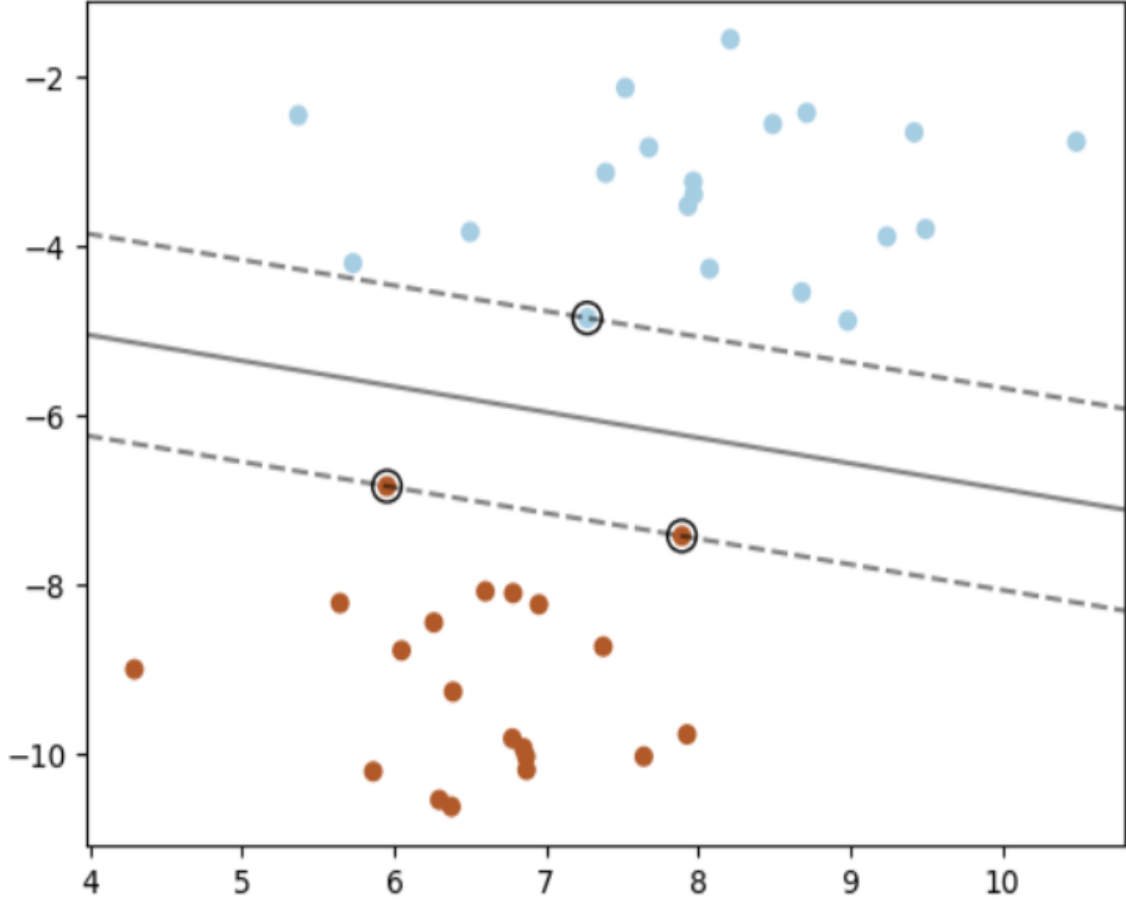
Destek Vektörü Makinesi ilk defa Boser, Guyon ve Vapnik tarafından tanıtılmıştır [46]. SVM ayırıcı bir hiper düzlem tarafından biçimsel olarak tanımlanmış, ayırt edici bir sınıflandırıcıdır [19]. Başka bir deyişle, sınıfı bilinen eğitim verileri (gözetimli öğrenme) verildiğinde, algoritma yeni örnekleri kategorize eden en uygun hiper düzlemi çıkarır. İki boyutlu boşlukta bu hiper düzlem, bir düzlemi, her bir sınıfın iki tarafında da bulunduğu iki parçaya bölen bir çizgidir. Bu hiper düzlem Şekil 4’de de görüldüğü gibi, iki sınıflı bir model iken sadece bir çizgiyi, üç sınıflı bir model olduğunda iki boyutlu bir alanı ifade eder. N boyutlu bir Öklid uzayındaki bir hiper düzlem, uzayı iki ayrı bölüme ayıran o alanın N-1 boyutlu alt kümesidir. İnce ayar parametreleri çekirdek (kernel), düzenleme (regularization), gamma ve kenar boşluğu (margin) parametreleridir. Bu parametrelerin en iyi kombinasyon ayarlanarak modelin doğru tahmin etme oranı artırılabilir.

SVM algoritmasına göre, her iki sınıftan çizgiye en yakın noktalar bulunur. Bu noktalara destek vektörleri denir. Destek vektörleri arasındaki mesafe olan marj hesaplanır. Amaç, marjı maksimize etmektir. Marjinin maksimum olduğu hiper düzlem, optimum hiper düzlemdir.

Bir doğrusal SVM modelinde hiper düzlemin öğrenilmesi düz bir fonksiyon olacak şekilde kolay olmayabilir bu durumda, çekirdek (kernel) yardımı ile veri üzerinde bir miktar lineer cebir kullanılarak dönüşüm işlemi yapılır ve model oluşturulur.

Gama parametresi tek bir eğitim örneğinin etkisinin ne kadar olduğunu tanımlar. Eğer düşük bir değere sahipse bu etki, her bir noktanın ulaşabileceği bir yer anlamına gelir ve bunun aksine, yüksek bir gama değeri, her noktanın yakın erişime sahip olduğu anlamına gelir.

Eğer gama çok yüksek bir değere sahipse, karar sınırı sadece karar sınırından çok uzak olan bazı noktaları görmezden gelmesiyle sonuçlanan çizgiye çok yakın olan noktalara bağlı olacaktır. Bunun nedeni, daha yakın noktaların daha fazla ağırlık alması ve kıpırdama eğrisine yol açmasıdır.



Şekil 4. İki sınıflı örnek bir model için maksimum marjlı hiper düzlemi

3.4.4. Lojistik Regresyon

İkili lojistik regresyon modeli, bir veya daha fazla tahmin edici değişkenine (özellikler matrisi) dayalı ikili cevap değişkenini tahmin etmek için kullanılır. Matematiksel gösterimi Denklem 4’de [47] verilen Lojistik regresyon, kategorik bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki ilişkiyi, kümülatif lojistik dağılım olan bir lojistik fonksiyon kullanarak olasılıkları tahmin ederek ölçmektedir [48].

$$P(y|x) = \frac{1}{1 + \exp(-y\theta^T x)} \quad (4)$$

x : Kelime vektörü

y : İkili sınıflandırıcısı oluşturulan İnternet sınıfına ait olup, olmama durumu (1 ya da 0)

θ : Bağımsız değişkenlerin (kelime vektörleri) regresyon çarpanlarını oluşturan Parametre vektörü (B_0, B_1, \dots)

Regresyonun bir alt türü olan Lojistik Regresyon, Regresyon ile aynı parametrelere sahiptir. Bu parametreler;

- Bağımlı değişkenler
- Bağımsız değişkenler
- Katsayılar vektörü

Regresyondan en önemli ayırt edici özelliği çıktının 0 veya 1 olmasıdır. Eğitim setindeki bağımlı ve bağımsız değişkenler (gözetimli makine öğrenme) kullanılarak katsayılar vektörü ile lojistik fonksiyon elde edilir. (Hataları en aza indirgeyen katsayıları bulmak için; maksimum olabilirlik kullanılır.)

Lojistik Regresyon, sınıflandırmanın daha çok olasılıksal bir görüşünü içerir. Lojistik regresyonun sonucu kesikli olmalı yani süreksiz olmalıdır, ancak algoritma çok boyutlu özellik (kategorik veya sürekli) uzayı üzerinde çalışabilir. Bu algoritma ikili sınıflandırıcı modellemek için kullanılmıştır.

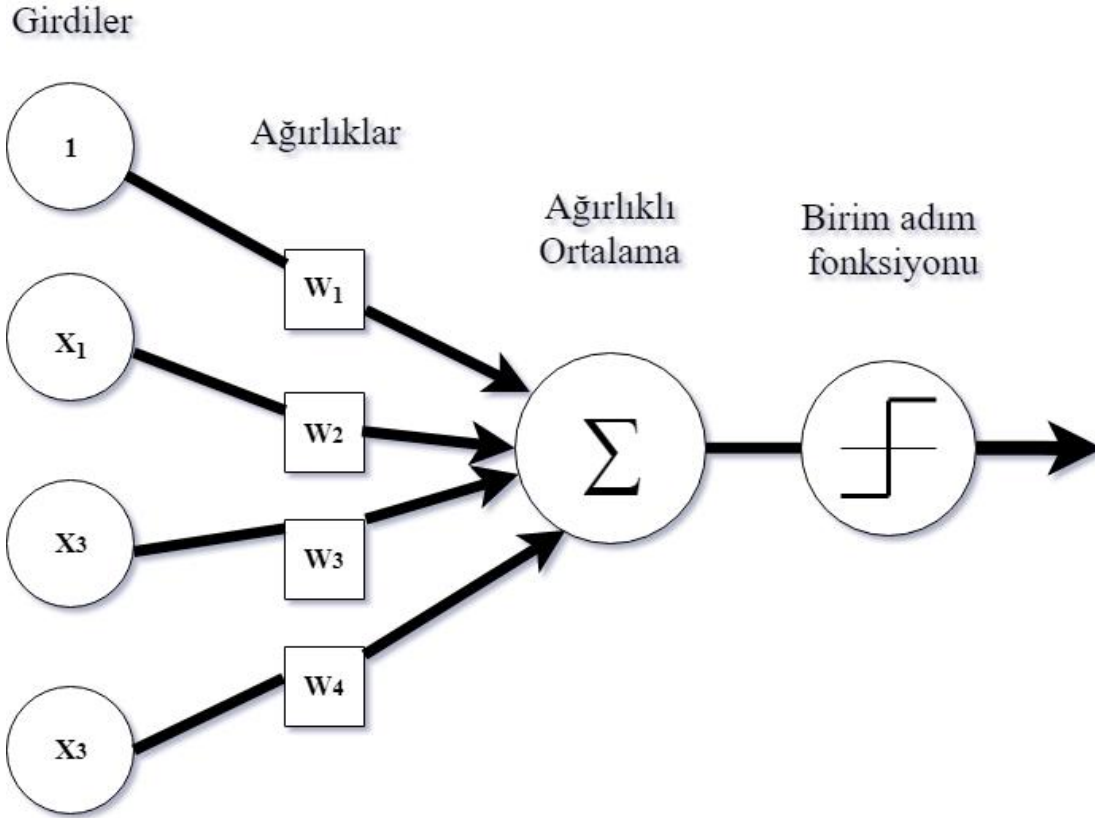
3.4.5. Tam Bağlantılı Yapay Sinir Ağları

Yapay sinir ağları, insan beyninde bulunan sinir ağlarından ilham alınarak geliştirilmiş, insan beyninin belirli kalıpları tanıma ve öğrenme örüntüsüne benzer olarak, kalıpları tanımak için tasarlanmış bir algoritmalar kümesidir. Yapay sinir ağlarının öğrenme yetisi sayesinde veriler kümelenebilir ve sınıflandırılabilir. Bu mekanizma, etiketlenmemiş verileri (gözetimsiz öğrenme) örnek girdiler arasında benzerlik kurarak gruplarken, üzerinde çalışacak etiketli bir veri kümesi olduğunda (gözetimli öğrenme) verileri sınıflandırır. Yapay sinir ağları, kümeleme ve sınıflandırma için kullanılan diğer algoritmalar için özellik seti (feature set) çıkarmak için kullanılabilir. Derin öğrenme, yapay sinir ağlarını temel alan takviye öğrenme, sınıflandırma ve regresyon için yöntemler içeren daha büyük makine öğrenme uygulamalarının bileşenleri olarak düşünülebilir [35].

Derin öğrenme “yığılmış sinir ağları” için kullanılan bir terimdir. Bu terim, bir kaç katmandan oluşan sinir ağlarını temsil eder. Sinir ağı katmanları düğümlerden meydana gelir.

Şekil 5’de görüldüğü gibi bir düğüm, hesaplamanın gerçekleştiği yer olup veri girişini, bu girişi genişleten veya azaltan bir dizi katsayı veya ağırlık ile birleştirir. Bu katsayılar, algoritmanın öğrenmeye çalıştığı görevle ilgili ve ilgili olmayan kısımların ayrıştırılmasını sağlar. Bir düğüm katmanı ise, girdi ağdan beslendikçe açılıp kapanan nöron benzeri anahtarların bir sırasıdır. Her bir katmanın çıktısı, verileri alan başlangıç katmanından başlayarak sonraki katmanın girdisidir. Modelin ayarlanabilir ağırlıklarını girdi özellikleriyle eşleştirmek, sinir ağının, girdiyi nasıl sınıflandırdığı ve kümelediğine ilişkin olarak bu özelliklere önem atamaktır.

Örneğin, bir veri kümesini hatasız olarak sınıflandırılmasını sağlayan en yararlı olan özellik bulunmak istendiğinde, bu özelliğin hesaplanmış ağırlıkları toplanır ve daha sonra bu toplam, sinyalin ağ üzerinden daha fazla ilerlemesi gerekip gerekmediğini belirlemek ve bir sınıflandırma eylemini etkilemek için düğümdeki aktivasyon fonksiyon aracılığıyla iletilir. Sinyaller aktivasyon fonksiyonundan geçerse, nöron aktif hale gelir ve bir sonraki katmana ilerler.



Şekil 5. Sinir ağındaki bir düğümün ilerleme süreci.

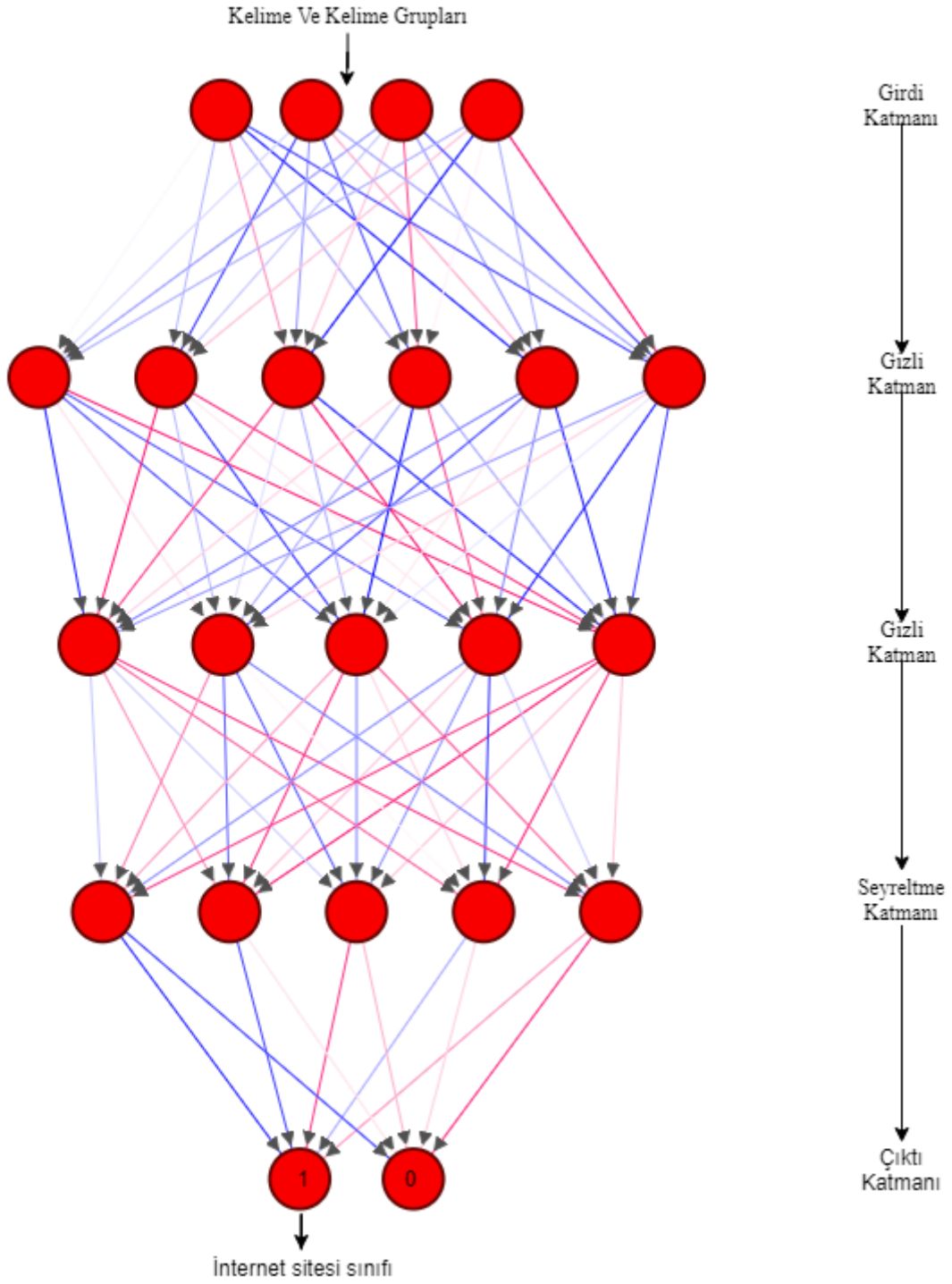
Her bir girdi ağırlıkları ile çarpılıp bu gizli katmanın düğümünde toplanır ve bu düğüm için bir değer bulunur. Gizli katmana ait düğümler, önceden tanımlanmış bir aktivasyon fonksiyona sahiptir. Aktivasyon fonksiyonları bu toplama göre düğümün aktif olup olmayacağına karar verir.

Aktivasyon fonksiyonları, çok katmanlı yapay sinir ağlarında doğrusal olmayan dönüşüm işlemleri için kullanılmaktadır. Gizli katmanlarda geri türev alınabilmesi için bu katmanların çıktılarını aktivasyon fonksiyonları ile normalize etmeye yarar. Bu çalışma kapsamında ReLU aktivasyon fonksiyonu, çıktı katmanı içerisinde ise Softmax aktivasyon fonksiyonu kullanılmıştır [49].

Aktivasyon Fonksiyonları:

- Identity
- Binary Step
- Sigmoid
- Tanh

- ReLU
- Leaky ReLU
- Softmax



Şekil 6. Tam Bağlantılı YapaySinir Ağları Mimarisi

Şekil 6’da mimarisi bulunan Yapay Sinir Ağı 3 tip ana katmandan oluşmaktadır:

- Giriş katmanı - sinir ağı için ilk veriler. X vektörü olarak gösterilir.
- Gizli katmanlar - giriş ve çıkış katmanı ve tüm hesaplamaların yapıldığı yer arasındaki ara katman.
- Çıkış katmanı - verilen girdiler için sonuç üretir.

Parametreler:

- Eğitim tur sayısı (epoch) bir veri kümesindeki örneklerin tamamının bütün ağdan bir kere geçtiğini ifade eden bir parametredir ve probleme göre değişiklik göstermektedir.
- Parça (batch) ise, modelin ağırlıkları güncellenmeden önce ağdan geçen örnek sayısını ifade etmektedir. Model eğitimi sırasında verinin tümü eş zamanlı modele sokulmaz. Parçalar halinde eğitimi katılır ve her bir parça ağdan geçtiğinde geriye yayılım algoritması kullanılarak ağırlıklar güncellenir. Sinir ağı modeli ilk oluşturulduğunda, modele başlangıç ağırlıkları atanması gerekmektedir. Bu ağırlıklar, veriler ağ üzerinden geçtikçe güncellenmektedir. Başlangıç ağırlıkları literatürde rastgele atama, normal dağılıma göre atama gibi farklı yöntemlerde denendiği görülmüştür. Başlangıç ağırlıklarının düzgün atanması, öğrenme aşamasında modelin performansını büyük ölçüde etkilemektedir.
- Düşümlerin seyretilmesi (dropout) parametresi, tam bağlı katmanlarda belirlenen bir eşik değerinin altında olan düşümlerin eğitim aşamasında rastgele deaktif edilmesini sağlamaktadır. Bu seyretilme işleminin, doğru sınıflandırma oranını arttırdığı gözlenmiştir. Zayıf veya gereksiz bilgilerin öğrenim setinden çıkarılması süreci iyileştirmektedir.
- Katman ve nöron sayıları iyileşen sayı ızgara arama gibi tekniklerle en iyi başarımlar oranını verenler için seçilebilir. Ancak katman sayısı arttıkça geriye yayılım sırasında bilgi aktarımı zorlaşabilir.
- Yapay sinir ağları, model hatasını hesaplamak için Denklem 5’de tanımlanan bir maliyet fonksiyonu (loss function) optimizasyon işlemi kullanılarak eğitilir [35]. Çapraz entropi ve ortalama karesel hata, sinir ağı modellerini eğitirken kullanılan iki ana maliyet fonksiyonudur.
- Maksimum olabilirlik, sinir ağlarını ve genel olarak makine öğrenme modellerini eğitirken bir kayıp fonksiyonunu seçmek için bir çerçeve sağlar.

Maliyet Fonksiyonu Eşitliği:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m |h_{\theta}(x^{(i)}), y^{(i)}| \quad (5)$$

m : Örnek veri sayısı

$h_{\theta}(x^{(i)})$: Tahmin edilen İnternet sınıfı (sayısal gösterim)

$y^{(i)}$: Gerçekte olan İnternet Sınıfı (sayısal gösterim)

İleri yayılım süreci; verilen bir girdiyi temel alarak sinir ağı çıkış değerini (tahmin) elde etmeye yarar. Bu algoritma maliyet değerini hesaplamak için kullanılır.

Geri yayılımda, θ (ağırlıklar) için en uygun değer kümesini ayarlanarak ve kullanılarak maliyet fonksiyonunu $J(\theta)$ en aza indirmek amaçlanır. Ağırlıklar güncellenirken Olasılıksal Dereceli Azaltma (Stochastic Gradient Descent), Adagrad, Adam, Adadelat gibi optimizasyon algoritmaları kullanılır. Çalışma kapsamında Adam algoritması kullanılmıştır.

3.5. Sınıflandırıcı Performans Ölçme Yöntemleri

Çalışmada kullanılmış olan makine öğrenme yöntemlerinin performansı değerlendirmek için, ölçme yöntemlerinden olan ve yaygın olarak kullanılan hata matrisinin bileşenlerinden üretilmiş ölçüm değerleri kullanılmıştır. Bu ölçüm değerleri; Başarım Oranı (Accuracy), F1 Skoru (F1 Score), Kesinlik (Precision) ve Duyarlılık (Recall) hesaplamalarıdır [50]. Çizelge 5’de verilmekte olan hata matrisinde negatif (0) gösterimi İnternet sayfasının ilgili (sınıflandırıcısı oluşturulan) sınıfa ait olmaması durumu iken; pozitif (1) gösterimi İnternet sayfasının ilgili sınıfa ait olması durumudur.

Performans ölçüm değerleri:

- Başarım Oranı (Denklem 6); tüm test seti içerisinde doğru sınıflandırılmış İnternet sitesi sayısı oranıdır.

$$Başarım Oranı = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

- Kesinlik (Denklem 7); pozitif tahmin edilen İnternet sitelerinin kaçının doğru olduğunun oranıdır. Yanlış pozitif tahminin maliyeti yüksek olduğu zaman kullanılabilecek bir ölçümdür. Bu çalışma özelinde Kesinlik, filtrelenecek istenen bir İnternet sitesi sınıfına ait olmayan bir sitenin filtrelenmesinin maliyeti olarak düşünülebilir.

$$Kesinlik = \frac{TP}{TP + FP} \quad (7)$$

- Duyarlılık (Denklem 8); gerçekte pozitif sınıfa ait olan İnternet sitelerinin ne kadarının pozitif tahmin edildiği oranıdır. Yanlış negatif tahminin maliyeti yüksek olduğu zaman kullanılabilecek bir ölçümdür. Bu çalışmada özelinde Duyarlılık, filtrelenecek istenen bir İnternet site sınıfına ait olan bir sitenin filtrelenmemesinin maliyeti olarak düşünülebilir.

$$Duyarlılık = \frac{TP}{TP + FN} \quad (8)$$

- F1 Skoru (Denklem 9); kesinlik ve duyarlılığın harmonik ortalamasıdır. Her iki ölçümü aynı anda göz önünde bulundurmak açısından önemli bir göstergedir.

$$F1Skoru = 2x \frac{Duyarlılık * Kesinlik}{Duyarlılık + Kesinlik} \quad (9)$$

Çizelge 5. Hata Matrisi Gösterimi

		Gerçekte Olan	
		Negatif (0)	Pozitif (1)
Tahmin Edilen	Negatif (0)	Doğru Negatif (True negative, TN): İkili sınıflandırma probleminde, 0 olarak tahmin edilen sınıfın gerçekte de 0 olması durumudur.	Yanlış Negatif (False negative, FN): İkili sınıflandırma probleminde, 0 olarak tahmin edilen sınıfın gerçekte 1 olması durumudur.
	Pozitif (1)	Yanlış Pozitif (False positive, FP): İkili sınıflandırma probleminde, 1 olarak tahmin edilen sınıfın gerçekte 0 olması durumudur.	Doğru Pozitif (True positive, TP): İkili sınıflandırma probleminde, 1 olarak tahmin edilen sınıfın gerçekte de 1 olması durumudur.

Diğer yandan bu çalışmada uygulanmış olan ikinci yöntem olan Çok Sınıflı Sınıflandırmanın performansını ölçmek için başarımlı ölçüm değeri kullanılmıştır. Bu yaklaşımda başarımlı değeri; toplam doğru sınıflandırılmış İnternet sayfası sayısının, test setinde kullanılan toplam İnternet sayfasına oranı ile elde edilmiştir (Denklem 6).

4. ANALİZ VE VAKA ÇALIŞMASI

63 tane sınıfa ait 45.543 adet İnternet sayfası Alexa'dan alınıp bu sayfalardaki kelime ve kelime grupları veri olarak kabul edilmiş, çalışma kapsamında geliştirilmiş Python betikleri ile veri toplama işlemi otomatik olarak gerçekleştirilmiştir. Bu işlem sonucunda ham veriye ulaşılmıştır. Ham veri içerisinde; sınıflar (tüm sınıflar Ekler kısmında, örnek bir sınıfa ait ham veri de kısmi örnek olarak Çizelge 6'da verilmektedir) ve o sınıflara ait Alexa dan alınan İnternet site içerisindeki tüm kelimeler bulunmaktadır.

Çizelge 6. Örnek sınıfa ait örnek ham veri (kısmi örnek)

Bilim Sınıfına Ait Kelimeler		
mathematics	membership	datasets
statistics	committees	employment
information	meetings	opportunities
nav	special	print
navigation	lectures	mission
home	iamg	iamg
home	publications	promote
information	student	worldwide
iamg	affairs	advancement
...

Sınıflandırma işleminden önce veriyi temizleme ve ön işleme süreçleri yapılmıştır. Kelimenin kök öbeğine ulaşma, noktalama işaretlerinden kurtulma, tüm kelimeleri küçük harfe dönüştürme, içerikten bağımsız kelimeler, bağlaçlar, imleçler, sayılar, kalıplaşmış kısaltmalardır gibi etkisiz kelimelerin çıkarılması gibi işlemler uygulanmıştır.

Kelime vektörleştirme işleminde, algoritmaların uygulanmasın önce uygulanan dönüştürme işlemi uygulanmıştır. Bu dönüştürme ile birlikte metin farklı yöntemlere göre kendi sınıfı içerisinde sayısal veriler içeren vektörlere dönüştürülmüştür. Bu yöntemler: TF-IDF, Kelime Torbası, Word2Vec dönüştürmeleridir. Bu dönüştürmelerden sonra 45543 İnternet sitesi test ve

eğitim seti olarak ayrılmıştır. Eğitim setinde 37814 İnternet sitesi kullanılmışken, test sırasında 7729 (tüm verinin % 16'lık kısmı) tane İnternet sitesi kullanılmıştır.

Vaka çalışmalarında, makine öğrenmesi ve yapay sinir ağları eğitim sırasında tahmin edilmesi istenen sınıfın da eğitim testinde verilen gözetimli sınıflandırıcılar kullanılmıştır. Kullanılan sınıflandırıcılar çok değişkenli Bernoulli ve Multinomial Naive Bayes, Rastgele Orman, Destek Vektör Makineleri, Tam Bağlantı Yapay Sinir Ağları kullanılmıştır.

Sınıflandırıcıların her biri kendi içinde farklı hiper parametreleri ile ve her biri farklı kelime vektörleştirme yöntemleri ile test edilmiştir. Test etme sırasında modelin başarısını öğrenmek için farklı metrikler kullanılmıştır. Bu metrikler; 3.5. *Sınıflandırıcı Performans Ölçme Değerleri* bölümünde detaylandırılan Karışıklık (hata) Matrisi, Başarım Oranı, F1 Skoru, Kesinlik ve Duyarlılık metrikleridir.

Çalışmada iki farklı sınıflandırma şekli kullanılmıştır. Bu sınıflandırmalar İkili Sınıflandırma ve Çok Sınıflı Sınıflandırma yöntemleri kullanılmıştır.

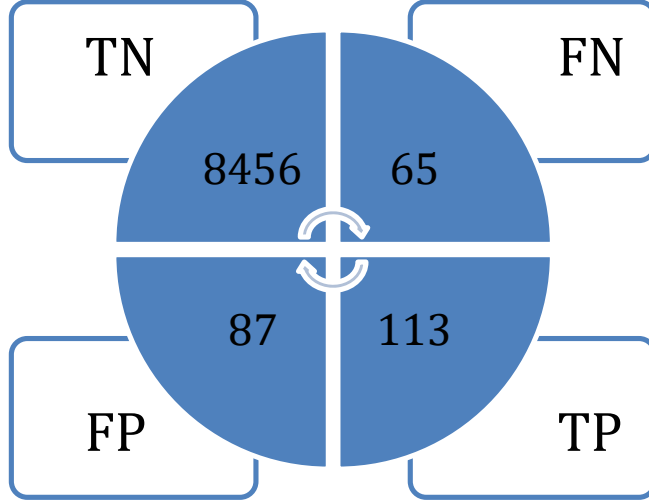
4.1. İkili (Binary - Binominal Class) Sınıflandırma

İkili veya binom sınıflandırma, belirli bir veri setinin öğelerini bir sınıflandırma kuralı temelinde iki gruptan (her birinin hangi gruba ait olduğunu tahmin ederek) birine atama işlemidir. Bir tür gözetimli öğrenme olan bu yöntem, eğitim sırasında özellik seti birlikte etiketlenen değişken, modeli oluşturulan sınıf için 1, geri kalan örnekler için 0 değeri alır [16]. Test sırasında da test setinden bir özellik matrisi verildiğinde 1 veya 0 dönmesi beklenir.

Bu çalışmada her bir sınıf için ayrı ayrı sınıflandırıcılar oluşturulmuştur. Yani bu çalışmada tanımlanmış sınıf sayısı kadar (63 tane) sınıflandırıcı oluşturulmuş ve bunların her birinin başarım oranları, hata matrisleri ve işlem süreleri çıkarılmıştır. Bu sayede 1 değeri atanan sınıfın filterenebilmektedir.

Bu aşamada; Tam Bağlantılı Katmanlı Sinir Ağları (Fully Connected Layer Network FCNN), Lojistik Regresyon (Logistic Regression) ve Bernoulli Naive Bayes (Bernoulli

Naive Bayes) algoritmaları ikili sınıflandırma için ayrı ayrı kullanılıp her birinin Başarım, Kesinlik, Duyarlılık ve F1 skor oranları Çizelge 7’de incelenmiştir.



Şekil 7. Sınıfı Erişkin İçerik ve Diğerleri olan Bir İkili Sınıflandırıcı FCNN' in Hata Matrisi

Şekil 7’deki karışıklık matrisi; FCNN sınıflandırıcısının, İnternet sınıfı erişkin içerik ve diğerleri olan bir test sonrasında 8456 tane doğru negatif, 65 tane yanlış negatif, 87 tane yanlış pozitif ve 113 tane doğru pozitif tahminde bulunduğunu göstermektedir.

Çizelge 7. Sınıflandırıcı Algoritmaların Tüm Sınıflar Üzerindeki Sonuçları

	Tam B. Y. Sinir Ağları					Lojistik Regresyon					Bernoulli Naïve Bayes				
İnternet Site Sınıfları	Başarım (%)	Kesinlik (%)	Duyarlılık (%)	F1 (%)	Skor	Başarım (%)	Kesinlik (%)	Duyarlılık (%)	F1 (%)	Skor	Başarım (%)	Kesinlik (%)	Duyarlılık (%)	F1 (%)	Skor
Alcohol-Tobacco	0,990	0,709	0,915	0,799		0,992	0,824	0,845	0,835		0,765	0,083	0,930	0,152	
Anime-Manga-Comics	0,975	0,464	0,620	0,531		0,986	0,819	0,475	0,601		0,656	0,057	0,910	0,107	
Books-Literature	0,974	0,456	0,680	0,546		0,986	0,748	0,595	0,663		0,927	0,168	0,560	0,258	
Business	0,907	0,175	0,835	0,289		0,923	0,201	0,800	0,321		0,534	0,044	0,950	0,084	
Celebrity Fan-Gossip	0,997	0,600	0,400	0,480		0,997	1,000	0,033	0,065		0,971	0,033	0,267	0,058	
Chat-Messaging	0,995	0,679	0,594	0,633		0,995	1,000	0,344	0,512		0,982	0,021	0,031	0,025	
Community and society	0,964	0,356	0,750	0,483		0,982	0,602	0,590	0,596		0,895	0,123	0,595	0,204	
Computing-Technology	0,952	0,251	0,570	0,349		0,972	0,397	0,465	0,429		0,602	0,050	0,930	0,096	
Diet-Exercise	0,992	0,731	0,664	0,696		0,992	0,905	0,479	0,626		0,951	0,103	0,345	0,159	
Education-Reference	0,960	0,330	0,755	0,460		0,977	0,502	0,690	0,581		0,903	0,141	0,645	0,231	
Family-Parenting	0,985	0,677	0,670	0,673		0,989	0,891	0,570	0,695		0,935	0,156	0,430	0,229	
Fashion-Beauty	0,969	0,401	0,710	0,513		0,979	0,532	0,590	0,559		0,673	0,055	0,830	0,103	
Finance-Investment	0,984	0,626	0,720	0,670		0,988	0,845	0,600	0,702		0,931	0,215	0,780	0,337	
Fine Arts	0,978	0,505	0,805	0,620		0,987	0,696	0,755	0,724		0,603	0,050	0,925	0,095	
Food-Drink	0,978	0,508	0,825	0,629		0,987	0,706	0,755	0,729		0,970	0,404	0,725	0,519	
Forums	0,998	0,200	0,125	0,154		0,998	0,210	0,540	0,302		0,982	0,055	0,830	0,103	
Gambling	0,990	0,764	0,810	0,786		0,992	0,978	0,680	0,802		0,979	0,520	0,665	0,583	
Games	0,960	0,321	0,675	0,435		0,982	0,593	0,620	0,606		0,932	0,205	0,695	0,317	
Government-Legal	0,990	0,744	0,753	0,749		0,989	0,897	0,500	0,642		0,967	0,327	0,644	0,434	
Health-Medicine	0,975	0,477	0,885	0,620		0,987	0,660	0,835	0,737		0,945	0,266	0,815	0,401	
Home-Garden	0,986	0,664	0,750	0,704		0,990	0,919	0,620	0,740		0,846	0,115	0,870	0,204	
Humor	0,974	0,413	0,310	0,354		0,980	0,909	0,150	0,258		0,944	0,144	0,295	0,194	
Hunting	0,989	0,757	0,765	0,761		0,990	0,838	0,670	0,744		0,689	0,065	0,950	0,122	
Information Security	0,988	0,773	0,573	0,658		0,986	0,811	0,410	0,545		0,945	0,211	0,629	0,316	

	Tam B. Y. Sinir Ağları				Lojistik Regresyon				Bernoulli Naïve Bayes			
İnternet Site Sınıfları	Başarım (%)	Kesinlik (%)	Duyarlılık (%)	F1 Skor (%)	Başarım (%)	Kesinlik (%)	Duyarlılık (%)	F1 Skor (%)	Başarım (%)	Kesinlik (%)	Duyarlılık (%)	F1 Skor (%)
İnternet Telephony	0,998	0,474	0,600	0,529	0,998	0,500	0,133	0,211	0,985	0,054	0,970	0,103
Job Search-Career Development	0,983	0,614	0,685	0,648	0,987	0,812	0,540	0,649	0,916	0,164	0,660	0,263
Kids Sites	0,989	0,256	0,373	0,303	0,995	0,923	0,203	0,333	0,978	0,034	0,085	0,049
Mobile Communications	0,971	0,398	0,575	0,470	0,982	0,676	0,375	0,482	0,904	0,138	0,620	0,225
Motor Vehicles	0,974	0,458	0,765	0,573	0,985	0,663	0,690	0,676	0,617	0,054	0,970	0,103
Movies	0,975	0,472	0,710	0,567	0,988	0,795	0,660	0,721	0,932	0,189	0,605	0,288
Music	0,972	0,437	0,780	0,560	0,982	0,594	0,665	0,627	0,643	0,057	0,950	0,108
News	0,990	0,724	0,890	0,798	0,993	0,840	0,840	0,840	0,931	0,180	0,580	0,275
Nudity	1,000	0,600	0,750	0,667	1,000	0,450	0,650	0,532	0,989	0,011	0,250	0,022
Online Services	0,987	0,561	0,548	0,554	0,989	0,800	0,349	0,486	0,961	0,102	0,222	0,140
Online Storage	1,000	0,714	0,714	0,714	0,999	0,620	0,540	0,577	0,994	0,149	0,810	0,252
Paranormal	0,989	0,778	0,735	0,756	0,991	0,984	0,625	0,765	0,945	0,211	0,520	0,300
Peer-to-Peer	0,998	0,429	0,316	0,364	0,998	1,000	0,053	0,100	0,986	0,111	0,440	0,177
Personals-Dating	0,995	0,776	0,559	0,650	0,995	0,935	0,426	0,586	0,976	0,066	0,162	0,094
Pets-Animals	0,968	0,397	0,800	0,531	0,985	0,650	0,760	0,700	0,619	0,055	0,970	0,103
Political Issues	0,979	0,525	0,675	0,591	0,986	0,786	0,495	0,607	0,907	0,111	0,440	0,177
Pornography	0,993	0,817	0,588	0,684	0,992	0,980	0,430	0,598	0,938	0,136	0,693	0,227
Proxies	0,998	0,583	0,350	0,438	0,998	1,000	0,150	0,261	0,982	0,014	0,100	0,024
Real Estate	0,986	0,667	0,770	0,715	0,991	0,895	0,685	0,776	0,891	0,149	0,810	0,252
Religion-Spirituality	0,962	0,357	0,865	0,506	0,976	0,488	0,885	0,629	0,941	0,252	0,815	0,385
Science	0,919	0,197	0,835	0,319	0,952	0,283	0,740	0,409	0,629	0,053	0,910	0,100
Search Engines-Portals	0,989	0,311	0,173	0,222	0,991	0,750	0,037	0,071	0,964	0,008	0,025	0,012
Sex Education	0,998	0,727	0,667	0,696	0,998	1,000	0,333	0,500	0,983	0,215	0,780	0,337
Shopping	0,920	0,191	0,780	0,307	0,951	0,283	0,770	0,413	0,689	0,065	0,950	0,122
Social Networking	0,995	0,250	0,054	0,089	0,996	0,650	0,350	0,455	0,970	0,017	0,108	0,029

	Tam B. Y. Sinir Ağları				Lojistik Regresyon				Bernoulli Naïve Bayes			
İnternet Site Sınıfları	Başarım (%)	Kesinlik (%)	Duyarlılık (%)	F1 Skor (%)	Başarım (%)	Kesinlik (%)	Duyarlılık (%)	F1 Skor (%)	Başarım (%)	Kesinlik (%)	Duyarlılık (%)	F1 Skor (%)
Software-Hardware	0,945	0,259	0,770	0,388	0,968	0,373	0,620	0,466	0,844	0,112	0,855	0,198
Sports	0,963	0,361	0,805	0,498	0,979	0,519	0,770	0,620	0,685	0,065	0,965	0,122
Stock Trading	0,989	0,674	0,655	0,664	0,990	0,818	0,497	0,618	0,968	0,264	0,517	0,350
Tasteless-Offensive	1,000	0,401	0,710	0,513	1,000	0,340	0,500	0,405	0,998	0,010	0,048	0,016
Technical Information	1,000	0,197	0,835	0,319	1,000	0,380	0,600	0,465	0,997	0,199	0,405	0,266
Television-Video	0,981	0,584	0,485	0,530	0,982	0,823	0,255	0,389	0,942	0,167	0,390	0,234
Travel	0,969	0,404	0,775	0,531	0,981	0,556	0,715	0,626	0,945	0,244	0,670	0,357
Weapons	0,993	0,654	0,631	0,642	0,995	0,863	0,524	0,652	0,980	0,238	0,524	0,327
cryptocurrency	0,999	1,000	0,429	0,600	0,999	1,000	0,381	0,552	0,986	0,010	0,048	0,016
online meetings	0,999	1,000	0,474	0,643	0,999	1,000	0,368	0,538	0,988	0,010	0,048	0,016
personal vpn	0,999	0,525	0,675	0,591	1,000	0,540	0,450	0,491	0,996	0,054	0,970	0,103
url shortener	0,999	0,175	0,835	0,289	0,999	0,450	0,560	0,499	0,996	0,065	0,965	0,122
web hosting	0,983	0,618	0,670	0,643	0,986	0,734	0,620	0,672	0,675	0,061	0,935	0,115

4.2. Çok Sınıflı Sınıflandırma

Çok sınıflı sınıflandırma yöntemi bir gözetimli sınıflandırma yöntemi olup, bu sınıflandırma içinde bulunan algoritmalar üç ve daha fazla sınıfı olan bir veriler için kullanılır. Çalışmada sınıflandırma modeli oluşturulurken, her bir İnternet sitesi sınıfı, kendi özellik setinde (kelime vektörü) temsil edilecek şekilde eğitilmiştir. Eğitim sonrası algoritmaların testleri sürecinde de; tahmin edilmesi beklenen çıktı yine her bir sınıf olacak şekilde modellenmiştir.

İkili sınıflandırma farklı olarak, her bir sınıf için bir sınıflandırıcı değil; tek bir sınıflandırıcı oluşturulup, tüm testler bu model üzerinden yapılmıştır. Bu durumdan dolayı hata matrisi 2X2 boyutluk bir matris değil, 63X63 boyutlu bir matristir. Bu matriste sütunlar sınıfların gerçekteki verisi iken satırlar tahmin edilen sınıflardır. Bu matris sayesinde hata yapılan ve birbirine en çok karışmış yani kelime yakınlığı en yakın sınıflar çıkarılmasına yardımcı olmuştur.

Çok sınıflı sınıflandırma model oluşturmak için, Multinomial Naive Bayes, Rastgele Orman Algoritması ve Destek Vektör Makineleri sınıflandırıcı algoritmalar kullanılmıştır. Çok Sınıflı Sınıflandırma sırasında, kelime farklı kelime vektörleştirme yöntemleri denenip, Çizelge 8’ de da görülen performans ölçüm değerleri kıyaslanmıştır.

Çizelge 8. Çok Sınıflı Sınıflandırıcıların Farklı Kelime Vektörleştirme Yöntemlerine Göre Başarım Oranları

		Çok Sınıflı Sınıflandırma Algoritmaları (Başarım Oranları)		
		Multinomial Naive Bayes	Rastgele Orman	Destek Vektör Makineleri
Kelime Vektörleştirme	TF-IDF	0,637	0,516	0,713
	BOW	0,694	0,503	0,643
	Word2Vec	0,621	0,495	0,610

5. SONUÇLAR VE ÖNERİLER

Çalışmanın sonunda, İnternet sitelerinin sınıflandırılması problemi için her bir sınıf için oluşturulan Çok Sınıflı Sınıflandırıcı modelleri ile farklı kelime vektörleştirme yöntemleri denenmiştir. İkili sınıflandırıcılar için ise kelime vektörleştirme yöntemi olarak sadece TF-IDF kullanılmıştır. Bu modeller oluşturulduktan sonra, test seti kullanılarak sonuçlar alınmış ve kıyaslanmıştır. Elde edilen deneysel sonuçlar, ikili sınıflandırma ve çok sınıflı sınıflandırma yaklaşımı olmak üzere iki farklı başlıkta incelenmiş ve aşağıda verilmektedir

5.1. İnternet Sitelerinin Sınıflandırma Probleminde İkili Sınıflandırma Yaklaşımı

Elde edilen deneysel sonuçlarda; İnternet sitelerini sınıflandırma probleminde; ikili sınıflandırma modellerinin, her İnternet sınıfı için oluşturulan sınıflandırıcıların kendi test set verileri üzerinde (sınıf etiketleri; ilgili sınıfa ait olma durumu: 1, diğerleri: 0) yüksek başarı oranı verdiği görülmüştür. Çok sınıflı sınıflandırma yöntemleri yerine, ikili sınıflandırma yöntemlerini kullanmak, sadece filtrelenmesi istenilen İnternet sınıfının bir sınıflandırıcısı oluşturulurak, daha etkili olacağı tespit edilmiştir.

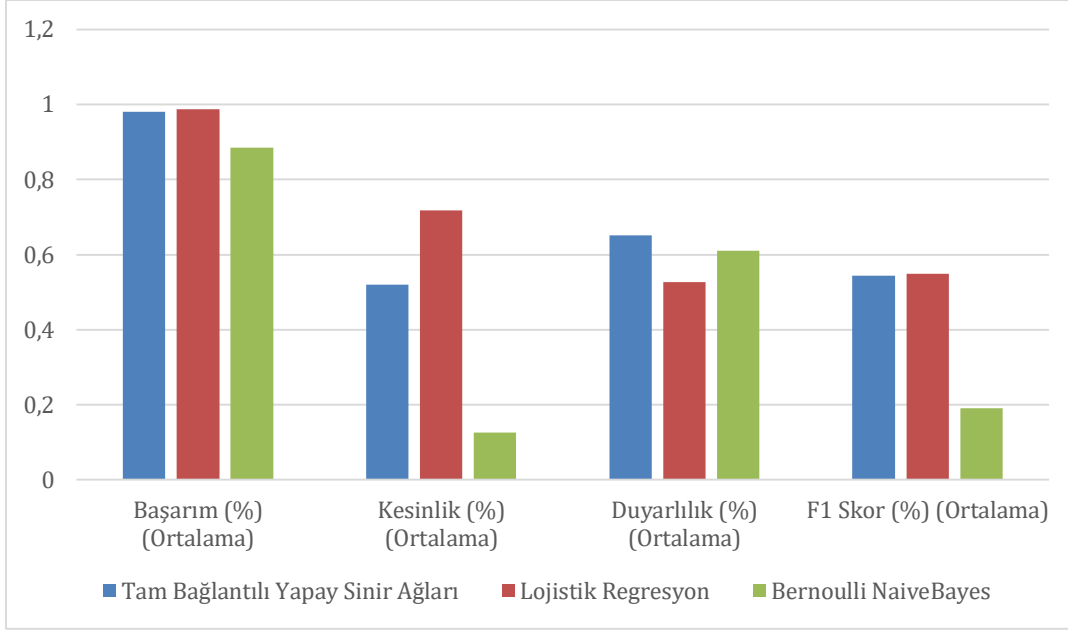
İkili sınıflandırma yöntemlerinin her biri için kendi içinde performans ölçüm değerleri ve ortalama test süreleri süreleri hesaplanmıştır. Çizelge 9 ve Şekil 8’de gösterilen ortalama başarı ölçümlerine bakıldığında, kesinlik ve başarı metriklerine göre en başarılı sınıflandırıcı Lojistik Regresyon’dur. Tam Bağlantılı Yapay Sinir Ağları ise, diğer sınıflandırıcılardan daha yüksek duyarlılığa sahiptir. İşlem süreleri incelendiğinde Bernoulli Naive Bayes ve Lojistik Regresyon sınıflandırıcılarının, Tam Bağlantılı Yapay Sinir Ağlarına göre çok daha hızlı sonuç verdiği görülmüştür. Bernoulli Naive Bayes sınıflandırıcısı ise F1 skoruna bakıldığında oldukça düşük performans göstermiştir.

Çizelge 9. İkili Sınıflandırıcı Algoritmaların Performansları

	İkili Sınıflandırma Algoritmaları (Başarım Oranları)		
	Tam Bağlantılı Yapay Sinir Ağları	Lojistik Regresyon	Bernoulli NaiveBayes
Başarım (%) (Ortalama)	0,980	0,987	0,884
Kesinlik (%) (Ortalama)	0,519	0,717	0,125
Duyarlılık (%) (Ortalama)	0,651	0,527	0,610
F1 Skor (%) (Ortalama)	0,543	0,548	0,190
İşlem Süresi (Sn.) (Ortalama)	64,7 sn	0,37 sn	0,45 sn

Ortalama performans ölçüm değerlerinden başarımlar oranı göz önüne alındığında (Çizelge 9), İkili Sınıflandırmada kullanılan algoritmalarından Lojistik Regresyon için daha yüksek olduğu tespit edilmiştir. Çizelge 8’de bulunan tüm sınıflara ait başarımlar oranları üzerinden; Lojistik Regresyon ile Tam Bağlantılı Yapay Sinir Ağları başarımlar ortalama değerleri ve Lojistik Regresyon ile Bernoulli Naive Bayes başarımlar ortalama değerleri arasında anlamlı bir fark olup olmadığı hipotezleri araştırılmıştır. Lojistik Regresyon ile Tam Bağlantılı Yapay Sinir Ağları başarımlar ortalama değerleri arasında fark yoktur hipotezi t-testi p-değeri: 0,0324 ile

reddedilirken, Lojistik Regresyon ile Bernoulli Naive Bayes başarımlar ortalama deęerleri arasında fark yoktur hipotezi t-testi p-deęeri: 0,000 ile reddedilmiřtir.



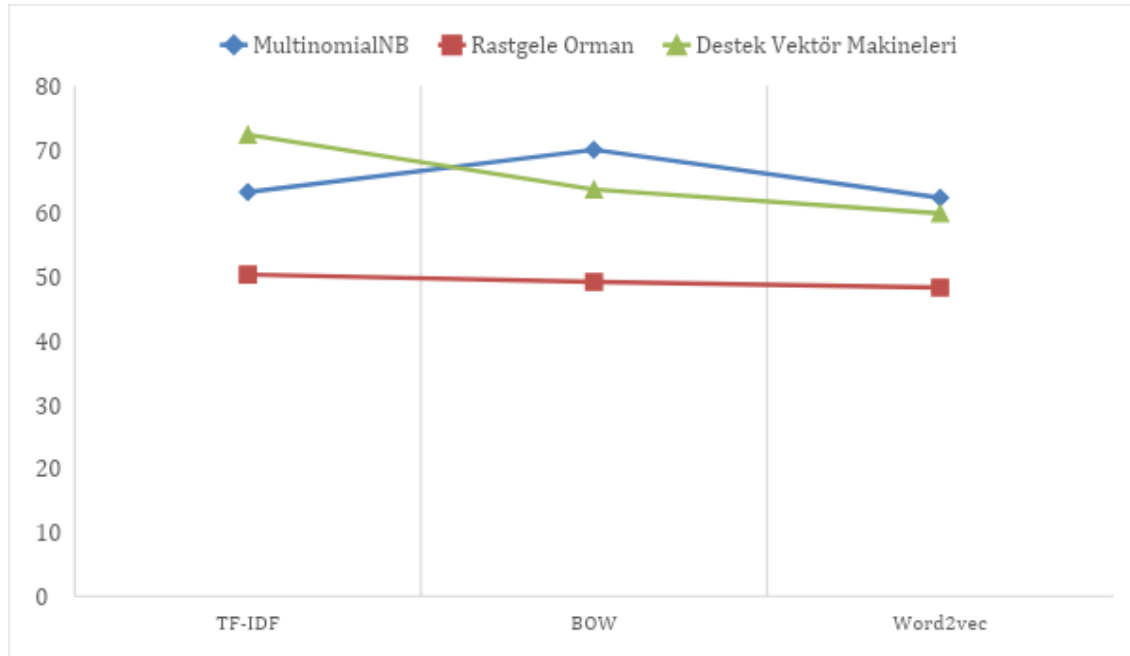
řekil 8. İkili Sınıflandırıcıların Ölçüm Ortalama Deęerleri

5.2. İnternet Sitelerinin Sınıflandırma Probleminde Çok Sınıflı Sınıflandırma

Elde edilen deneysel sonuçlarda, çok sınıflı sınıflandırma problemi için bařarım oranlarını içeren Çizelge 10 ve řekil 9’da da görüldüęü gibi farklı sınıflandırıcı ve kelime vektörleştirme yöntemlerinin kullanımı, sınıflandırmanın bařarılı olmasında bir etki yarattıęı görülmüřtür. Bu algoritmalarından Destek Vektör Makineleri çok sınıflı sınıflandırıcısı ve TF-IDF kelime vektörleştirme yöntemi en bařarılı sonucu vermiřtir. Deneysel sonuçların tümü göz önüne alındıęında BOW yönteminin ortalamada en iyi sonucu verdięi söylenebilir. Sırasıyla TF-IDF ve Word2Vec bu kelime vektörleştirme yöntemini izlemiřtir.

Çizelge 10. Çok Sınıflı Sınıflandırıcı Algoritmaların Başarım Oranları

		Çok Sınıflı Sınıflandırma Algoritmaları (Başarım Oranları)		
		Multinomial Naive Bayes	Rastgele Orman	Destek Vektör Makineleri
Kelime Vektörleştirme	TF-IDF	0,637	0,516	0,713
	BOW	0,694	0,503	0,643
	Word2Vec	0,621	0,495	0,610



Şekil 9. Sınıflandırıcı Algoritmalar ile Kelime Vektörleştirme Kombinasyonlarının Başarım Oranları(%)

5.3. Araştırma Soruları Üzerine Öneriler

Bu çalışmada bir metin sınıflandırma veri seti oluşturulmuştur ve bu veri seti kullanılarak, İkili Sınıflandırma ve Çok Sınıflı Sınıflandırma temelli iki farklı İnternet sitesi sınıflandırma sistemi geliştirilmiştir. Çizelge 10 ve 11’de verilmekte olan başarım oranları göz önüne alındığında AS1’de önerilen soruya cevap olarak İkili Sınıflandırma sadece istenilen bir İnternet site sınıfının filtrelenmesi görevini yerine getirmek için kullanıldığında daha etkili olacağı tespit edilmiştir.

İkili Sınıflandırmada kullanılan yöntemlerin, Çizelge 10'da bulunan analizin süreçleri boyunca işlemsel performansları (süre olarak) göz önüne alındığında, AS2'de önerilen soruya cevap olarak Lojistik Regresyon ve Bernoulli Naive Bayes sınıflandırıcılarının, Tam Bağlantılı Yapay Sinir Ağlarına göre 150 kat daha hızlı sonuçlandığı gözlenmiştir.

Şekil 8'de görüldüğü gibi ikili sınıflandırma algoritmaları, yüksek başarımla sonuçlanmıştır. Ancak diğer performans ölçüm değerlerinde bu başarıyı yakalayamama sebeplerinden biri de; İkili Sınıflandırma algoritmaları kullanılırken herhangi bir sınıfa ait veri setinin dengesiz yani eşit sayıda olmaması, eğitim sırasında o sınıfı daha az temsil etme sorununu doğurmuştur. AS3'de öngörülen bu problem, Çizelge 10'da tanımlanan kesinlik, duyarlılık ve F1 skoru gibi ölçümleri ile tespit edilmiştir. Gelecek çalışmalarda bu sorunun çözümü için veri setini genişletme veya veri üretme gibi tekniklere başvurulabilir.

AS4'de önerilen soruya cevap olarak, kullanılan makine öğrenme sınıflandırıcılarının performansları arasında belirgin bir fark olmuştur. Bu durum sınıflandırıcıların metin üzerindeki kapasite ve yetenekleri ile doğrudan alakalı olabilir. İkili sınıflandırma yöntemlerinden Lojistik Regresyon, Tam Bağlantılı Yapay Sinir Ağları ve Bernoulli Naive Bayes sınıflandırıcılarından, çok sınıflı sınıflandırma yöntemlerinden Destek Vektör Makineleri, Rastgele Orman ve Naive Bayes sınıflandırıcılarına göre daha yüksek başarımla olduğu tespit edilmiştir. Ayrıca ikili ve çok sınıflı sınıflandırma algoritmalarının ayrı ayrı ve farklı vektörleştirme yöntemleri ile denenmesi; İnternet sayfalarının sınıflandırılması ve filtrelenmesi problemlerini birlikte ele alınmasını sağlamış olup, benzer çalışmalardan farkı ortaya konmuştur.

Veri setinin son hali oluşturulmadan önce, İnternet sayfa sınıfları üzerine detaylı bir çalışma yapılmıştır. Çalışmanın zorlayıcı tarafı, İnternet sayfa sınıfının literatürdeki diğer çalışmalardan daha fazla olması ve bir İnternet sayfa sınıfının diğer İnternet sayfa sınıfının bir alt kümesi olacak kadar yakın olmasıdır. Bu çalışma kapsamında, geliştirilen sınıflandırıcılar tarafından sınıflandırılan İnternet sitelerinin bir biriyle karışması yani yakın sınıflara ait olması üzerine geribesleme ve uzman görüşü ile Alexa'dan 45.543 adet her bir İnternet sayfası kelime dağılımları ve diğer sınıfla etkileşimleri de çıkarılarak özenle seçilmiştir. Bu çalışma; sonraki benzer çalışmalara, veri seti oluşturulmadan önce, İnternet

sayfa sınıfı belirleme kısmına da yardımcı olacakken, alıřmacıların zamanını daha ok sınıflandırıcı algoritmalar zerine geirmelerine katkı saėlayacaktır.

6. KAYNAKLAR

- [1] Anonim, «<https://www.internetlivestats.com/total-number-of-websites/>,» 13 Mayıs 2019. [Çevrimiçi]. [Erişildi: 2019].
- [2] Anonim, «<https://news.netcraft.com/archives/category/web-server-survey/>,» 16 Mayıs 2019. [Çevrimiçi]. [Erişildi: 2019].
- [3] R. Rekika, A. M. Alimia, J. Casillas ve I. Kallelac, «Assessing web sites quality: A systematic literature review by text and association rules mining,» *International Journal of Information Management*, cilt 38, pp. 201-216, 2018.
- [4] Á. Figueira, «The current state of fake news: challenges and opportunities,» *Procedia Computer Science*, cilt 121, pp. 817-825, 2017.
- [5] E. Cardoso ve R. Silva, «Towards automatic filtering of fake reviews,» *Neurocomputing*, cilt 309, pp. 106-116, 2018.
- [6] J. Hartmann, J. Huppertz ve C. Scham, «Comparing automated text classification methods,» *International Journal of Research in Marketing*, cilt 36, no. 1, pp. 20-38, 2018.
- [7] Y. Li ve A. Jain, «Classification of Text Documents,» *Computer Journal*, 1998.
- [8] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, B. L. ve B. D. , «Text classification algorithms: A survey. Information,» 2019.
- [9] C. C. Aggarwal ve C. X. Zhai, *Mining Text Data*, 2012.
- [10] A. Ågren ve Å. Christian, «Combating Fake News with Stance Detection using Recurrent Neural Networks,» 2018.
- [11] J. Cuzzola, J. Jovanovic, D. Gašević ve E. Bagheria, «Automated Classification and Localization of Daily Deal Content from the Web,» *Applied Soft Computing*, cilt 31, pp. 41-256, 2015.
- [12] R. A. Stein, J. Patricia ve J. F. Valiati, «An analysis of hierarchical text classification using word embeddings,» *Information Sciences*, cilt 471, pp. 216-232, 2019.
- [13] F. Elghannam, «Text representation and classification based on bi-gram alphabet,» *Journal of King Saud University - Computer and Information Sciences*, 2019.

- [14] R. A. Sinoara, J. Camacho-Collados, G. R. Rossi, R. Navigli ve S. O. Rezende, «Knowledge-enhance document embeddings for text classification,» *Knowledge-Based Systems*, cilt 163, pp. 955-971, 2018.
- [15] S. Borah ve R. Panigrahi, «Classification and Analysis of Facebook Metrics Dataset Using Supervised Classifiers,» pp. 1-19, 2019.
- [16] T. Takenouchi ve S. Ishii, «Binary classifiers ensemble based on Bregman divergence for multi-class classification,» *Neurocomputing*, cilt 273, pp. 424-434, 2017.
- [17] C. Manning, P. Raghavan ve H. Schütze, Introduction to Information Retrieval Text classification and Naive Bayes, Cambridge University Press, 2008.
- [18] A. Onan, S. Korukoğlu ve H. Bulut, «Ensemble of keyword extraction methods and classifiers in text classification,» *Expert Systems with Applications*, cilt 57, pp. 232-247, 2016.
- [19] R.-C. Chen ve C.-H. Hsieh, «Web page classification based on a support vector machine using a weighted vote schema,» *Expert Systems with Applications*, cilt 31, pp. 427-435, 2006.
- [20] S. Xu, Y. Li ve Z. Wang, «Bayesian multinomial Naïve Bayes classifier to text classification.,» *Advanced multimedia and ubiquitous engineering*, pp. 347-352, 2017.
- [21] R. Huang, X. Nie ve Y. Li, «Web Spam Classification Method Based on Deep Belief Network,» *Expert Systems With Applications*, cilt 96, pp. 261-270, 2017.
- [22] S. Kudugunta ve E. Ferrara, «Deep neural networks for bot detection,» *Information Sciences*, pp. 312-322, 2018.
- [23] Z. Yao ve C. Zhi-Min, «An Optimized NBC Approach in Text Classification,» %1 içinde *International Conference on Applied Physics and Industrial Engineering 2012*, 2012.
- [24] B. Trstenjak, S. Mikac ve D. Donko, «KNN with TF-IDF Based Framework for Text Categorization,» *Procedia Engineering*, cilt 69, pp. 1356-1364, 2014.
- [25] C.-Y. Liang, G. Li , Z.-J. Xia, Z.-Y. Yang, L. Su, X.-X. Li ve F.-G. Nie, «Dictionary-based text categorization of chemical web pages,» *Information Processing and Management*, cilt 42, pp. 1017-1029, 2006.
- [26] O. Kotevska, S. Padi ve A. Lbath, «Automatic Categorization of Social Sensor Data,» *Procedia Computer Science*, cilt 98, pp. 596-603, 2016.

- [27] Y. Chen, B. Cheng ve X. Cheng, «Food Safety Document Classification Using LSTM-based Ensemble Learning,» *Revista Técnica de la Facultad de Ingeniería Universidad del Zulia*, cilt 39, no. 10, pp. 172-178, 2016.
- [28] A. Qazi ve R. Goudar, «An Ontology-based Term Weighting Technique for Web Document Categorization,» cilt 133, pp. 75-81, 2018.
- [29] I. Hernández, C. Rivero, R. Corchueloc ve R. Corchuelo, «CALA: An unsupervised URL-based web page classification system,» *168-180*, cilt 57, pp. 168-180, 2014.
- [30] H. Li, K.-K. R. Choo, G. Sun ve T. Li, «An optimized approach for massive web page classification using entity similarity based on semantic network,» *Future Generation Computer Systems*, cilt 76, pp. 510-518, 2016.
- [31] J.-H. Lee, C. Chuangab ve W.-C. Yeh, «Web page classification based on a simplified swarm optimization,» *Applied Mathematics and Computation*, cilt 270, pp. 13-24, 2015.
- [32] N. Gali, P. Fränti ve R. Marinescu-Istodor, «Using linguistic features to automatically extract internet page title,» *Expert Systems with Applications*, cilt 79, pp. 296-312, 2017.
- [33] D. Shen, Q. Yang ve Z. Chen, «Noise reduction through summarization for Web-page classification,» *Information Processing & Management*, cilt 43, pp. 1735-1747, 2007.
- [34] A. F. De Souza, F. Pedroni , E. Oliveia, P. M.Ciarelli, C. Badue ve W. F. Henriquea, «Automated multi-label text categorization with VG-RAM weightless neural networks,» *Neurocomputing*, cilt 72, pp. 2209-2217, 2009.
- [35] B. Yu, Z.-B. Xu ve C.-h. Li, «Latent semantic analysis for text categorization using neural network,» *Knowledge-Based Systems*, cilt 21, pp. 900-904, 2008.
- [36] Y. Lin ve L. Zongpeng , «RNN-Enhanced Deep Residual Neural Networks,» 2016.
- [37] H. Yu, J. Han ve . K. Chang, «Web Page Classification without Negative Examples,» *IEEE Transactions on Knowledge and Data Engineering*, cilt 16, pp. 70 - 81, 2004.
- [38] D. Riboni, «Feature Selection for Web Page Classification,» 2003.
- [39] D. Mladeni, «Turning Yahoo into an Automatic Web Page Classifier,» 1998.
- [40] L. Shih ve D. Karger, «Using URLs and Table Layout for Web Classification Tasks,» 2004.

- [41] A. Osanyin, O. Oladipupo ve I. Afolabi, «A Review on Web Page Classification,» *Covenant Journal of Informatics & Communication Technology*, 2018.
- [42] X. Ren, C. Shi, W. Wang ve D. Zhang, «An improved SVM web page classification algorithm,» *Journal of Physics: Conference Series*, cilt 1187, 2019.
- [43] E. Loper ve S. Bird, «NLTK: The Natural Language Toolkit,» 2014.
- [44] D. Kim, D. Seo, P. Kang ve S. Cho, «Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec,» *Information Sciences*, pp. 15-29, 2019.
- [45] W. Zheng, Y. Li ve S. X. Yan, «Bayesian Multinomial Naïve Bayes Classifier to Text Classification,» %1 içinde *Advanced Multimedia and Ubiquitous Engineering: MUE/FutureTech 2017*, 2017, pp. 347-352.
- [46] B. E. Boser, I. M. Guyon ve V. N. Vapnik, «A Training Algorithm for Optimal Margin Classifiers,» %1 içinde *Proceedings of the fifth annual workshop on Computational learning theory, ACM*, 1992.
- [47] X. Zhu, «Text Categorization with Logistic Regression,» 2007.
- [48] J. Hilbe, *Logistic Regression Models*, 2011.
- [49] M.-L. Zhang ve Z.-H. Zhou, «Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization,» *IEEE Transactions on Knowledge and Data Engineering*, cilt 18, pp. 1338 - 1351, 2006.
- [50] E. Beauxis-Aussalet ve L. Hardman, «Visualization of Confusion Matrix for Non-Expert Users,» *CWI - Information Access Group*, cilt 6, 2014.
- [51] R. A. Sinoara, J. Camacho-Collados, R. G. N. R. Rossi ve S. O. Rezende, «Knowledge-enhanced document embeddings for text classification,» *Knowledge-Based Systems*, cilt 163, pp. 955-971, 2019.

7. EKLER

7.1. EK - 1

	Category (İngilizce)	Sınıf (Türkçe)
1	Adult Content	Erişkin İçerik
2	Alcohol-Tobacco	Alkol-Tütün
3	Anime-Manga-Comics	Çizgi Roman
4	Books-Literature	Kitap Edebiyat
5	Business	İşletme
6	Celebrity Fan-Gossip	Eğlence
7	Chat-Messaging	Mesajlaşma
8	Community and society	Topluluk ve toplum
9	Computing-Technology	Bilgisayar Teknolojileri
10	Cryptocurrency	Dijital Para
11	Diet-Exercise	Diyet-Egzersiz
12	Education-Reference	Eğitim-Referans
13	Family-Parenting	Aile Ebeveyn
14	Fashion-Beauty	Moda
15	Finance-Investment	Finans-Yatırım
16	Fine Arts	Güzel Sanatlar
17	Food-Drink	Yiyecek-İçecek
18	Forums	Forumlar
19	Gambling	Kumar
20	Games	Oyun
21	Government-Legal	Devlet-Hukuk
22	Health-Medicine	Sağlık-Tıp
23	Home-Garden	Ev-Bahçe
24	Humor	Komik
25	Hunting	Avcılık
26	Information Security	Bilgi Güvenliği
27	İnternet Telephony	İnternet Telefonu
28	Job Search - Career Development	İş Arama - Kariyer Gelişim
29	Kids Sites	Çocuk Siteleri

30	Mobile Communications	Mobil İletişim
31	Motor Vehicles	Motorlu Taşıtlar
32	Movies	Film
33	Music	Müzik
34	News	Haber
35	Nudity	Çıplak İçerik
36	Online Meetings	Çevrimiçi Toplantı
37	Online Services	Çevrimiçi Servis
38	Online Storage	Çevrimiçi Bellek
39	Paranormal	Normal ötesi
40	Peer-to-Peer	Birebir İletişim
41	Personal vpn	Kişisel Vpn
42	Personals-Dating	Kişisel Buluşma
43	Pets-Animals	Evcil Hayvanlar
44	Political Issues	Politik
45	Pornography	Açık Saçık Yayın
46	Proxies	Proxy
47	Real Estate	Gayri Menkul
48	Religion-Spirituality	Din Ruhani
49	Science	Bilim
50	Search Engines-Portals	Arama Motoru
51	Sex Education	Cinsel Eğitim
52	Shopping	Alışveriş
53	Social Networking	Sosyal Ağ
54	Software-Hardware	Yazılım-Donanım
55	Sports	Spor
56	Stock Trading	Borsa
57	Tasteless-Offensive	Saldırgan
58	Technical Information	Teknik Bilgi
59	Television-Video	TV-Video
60	Travel	Gezi
61	Url shortener	Birörnek Kaynak Konumlayıcı Kısaltıcı
62	Weapons	Silah

7.2. Tezden Türetilmiş Bildiriler

İ. Şahin ve O. Chouseinoglou, «6th International Management Information Systems Conference-Data Mining and Machine Learning Session» İstanbul, 2019.



HACETTEPE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
YÜKSEK LİSANS/DOKTORA TEZ ÇALIŞMASI ORJİNALLİK RAPORU

HACETTEPE ÜNİVERSİTESİ
FEN BİLİMLER ENSTİTÜSÜ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI BAŞKANLIĞI'NA

Tarih: 06/09/2019

Tez Başlığı: Metin Madenciliği ve Makine Öğrenmesi İle İnternet Sayfalarının Sınıflandırılması

Yukarıda başlığı gösterilen tez çalışmamın a) Kapak sayfası, b) Giriş, c) Ana bölümler d) Sonuç kısımlarından oluşan toplam 50 sayfalık kısmına ilişkin, 06/09/2019 tarihinde benim/tez danışmanım tarafından Turnitin adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı % 3 'tür.

Uygulanan filtrelemeler:

- 1- Kaynakça hariç
- 2- Alıntılar hariç/dâhil
- 3- 5 kelimedenden daha az örtüşme içeren metin kısımları hariç

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü Tez Çalışması Orjinallik Raporu Alınması ve Kullanılması Uygulama Esasları'nı inceledim ve bu Uygulama Esasları'nda belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini saygılarımla arz ederim.

Tarih ve İmza

Adı Soyadı: İlker ŞAHİN

Öğrenci No: N16121595

Anabilim Dalı: Endüstri Mühendisliği

Programı: Endüstri Mühendisliği Tezli Yüksek Lisans

Statüsü: ☒ Y.Lisans ☐ Doktora ☐ Bütünleşik Dr.

06/09/2019

DANIŞMAN ONAYI

UYGUNDUR.

Doç. Dr. Oumout CHOUSEINOLOU

ÖZGEÇMİŞ

Adı Soyadı : İlker Şahin
Doğum yeri : Ankara
Doğum tarihi : 08.09.1988
Medeni hali : Bekar
Yazışma adresi : Comodo Ankara (Ar-Ge) Üniversiteler Mh. İhsan Doğramacı
Bulvarı Bina No:29 Gümüş Bloklar Bk:10 06800 ODTÜ
Teknokent Ankara/TÜRKİYE
Telefon : 0 (554) 274 79 03
Elektronik posta adresi : ilker.sahin@hacettepe.edu.tr
Yabancı dili : İngilizce

EĞİTİM DURUMU

Lisans : Gazi Üniversitesi-Fen Fakültesi-İstatistik Bölümü-2013
Y. Lisans : Hacettepe Üniversitesi-Mühendislik Fakültesi-Endüstri
Mühendisliği Bölümü-2019

İş Tecrübesi

Temmuz 2015 - Ağustos 2017 Aria Telekom-Veri Analisti
Ağustos 2017 -Devam Comodo Türkiye-Veri Bilimci

Diğer

Programlama Dilleri: Python, R

Teknolojiler: PostgreSQL, Pandas, Numpy, NLP, Sci-kit, NLKT, Elastic Search API,
PySpark (Başlangıç), Tensorflow (Başlangıç), DASK (Başlangıç)

Programlar: PyCharm, Jupyter Notebook, Kibana, Minitab Statistical Analysis Program,
SPSS Statistical Analysis Program, Win QSB Integer Programming and Operational
Research, Bayes Network (SAMIAM) SpamAssassin with Perl Regex, Jira

İşletim Sistemleri: Linux, Window

İlgi Alanları: Derin Öğrenme, Makine Öğrenmesi, Veri-Metin Madenciliği, Doğal Dil
İşleme, Veri Bilimi, İstatistiksel Analiz

Kurslar: Udemy ve Lynda Online Veri Bilimi Kursları