T.C.
MANİSA CELAL BAYAR ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ
YAZILIM MÜHENDİSLİĞİ ANABİLİM DALI
YAZILIM MÜHENDİSLİĞİ BİLİM DALI

MAKİNE ÖĞRENMESİ İLE E-TİCARET ÜRÜN YORUMLARININ ANALİZİ

Müjdat ÇABUK

Danışman
Dr. Öğretim Üyesi Fatih YÜCALAR

II. Danışman
Dr. Öğretim Üyesi Mansur Alp TOÇOĞLU



MANİSA-2023

TAAHÜTNAME

Bu tezin Manisa Celal Bayar Üniversitesi Hasan Ferdi Turgutlu Teknoloji Fakültesi Yazılım Mühendisliği Bölümü'nde, akademik ve etik kurallara uygun olarak yazıldığını ve kullanılan tüm literatür bilgilerini referans gösterilerek tezde yer aldığını beyan ederim.

Müjdat ÇABUK

İÇİNDEKİLER

İÇİNDEKİLER I
SİMGELER VE KISALTMALAR DİZİNİ III
ŞEKİLLER DİZİNİ IV
TABLO DİZİNİ VI
TEŞEKKÜR VII
ÖZET VIII
ABSTRACT IX
1. GİRİŞ 1
2. LİTERATÜR TARAMASI 6
3. GENEL BİLGİLER 10
3.1. Makine Öğrenmesi 10
3.1.1. Makine Öğrenmesi Yöntemleri 11
3.1.1.1. Denetimli Öğrenme 11
3.1.1.2. Denetimsiz Öğrenme 13
3.1.1.3. Takviyeli Öğrenme 14
3.1.2. Makine Öğrenmesi Algoritmaları 14
3.1.2.1. Lojistik Regresyon 15
3.1.2.2. Destek Vektör Makineleri 15
3.1.2.3. Rastgele Orman 16
3.1.2.4. Naive Bayes 17
3.2. Derin Öğrenme 18
3.2.1. Derin Öğrenme Mimarileri 19
3.2.1.1. Tekrarlayan Sinir Ağları 19
3.2.1.2. Uzun Kısa Süreli Bellek Ağları 20
3.2.1.3. Geçitli Tekrarlayan Birim 21

3.2.1.4. Evrişimli Sinir Ağları 22
3.3. Metin Madenciliği 23
3.4. Duygu Analizi 24
3.4.1. Duygu Analizi Yöntemleri 25
3.4.1.1. Kural Tabanlı Yöntemler 25
3.4.1.2. Makine Öğrenmesi Yöntemleri 26
3.4.1.3. Hibrit Yöntemler 27
3.4.1.4. Derin Öğrenme Yöntemleri 27
3.5. Doğal Dil İşleme 28
4. MATERYAL VE YÖNTEMLER 30
4.1. Manuel Veri Setinin Hazırlanması 30
4.1.1. Veri Toplama 31
4.1.2. Veri Ön İşleme 33
4.1.3. Verilerin Etiketlenmesi 34
4.2. Otomatik Veri Setinin Hazırlanması 36
4.3. Deneyler 38
5. UYGULAMA 51
5.1. Uygulama Modülleri 51
5.1.1. Veri Çekme Modülü 52
5.1.2. Veri Ön İşleme Modülü 54
5.1.3. Veri Etiketleme Modülü 55
6. SONUÇLAR 59
KAYNAKLAR 61
ÖZGEÇMİŞ 65

SİMGELER VE KISALTMALAR DİZİNİ

ANN	Artificial Neural Network (Yapay Sinir Ağı)
CNN	Convolutional Neural Network (Evrişimli Sinir Ağı)
CRM	Customer Relations Management (Müşteri İlişkileri Yönetimi)
CSV	Comma Separated Values (Virgülle Ayrılmış Değerler, Bir Dosya Formatı)
DBSCAN	Density-Based Spatial Clustering of Applications With Noise (Gürültü İçeren Uygulamaların Yoğunluğa Dayalı Mekânsal Kümelenmesi)
DT	Decision Tree (Karar Ağacı)
ETBİS	E-Ticaret Bilgi Sistemi
GRU	Gated Recurrent Unit (Geçitli Tekrarlayan Birim)
KNN	K-Nearest Neighbours (K En Yakın Komşuluk)
LR	Logistic Regression (Lojistik Regresyon)
LSTM	Long-Short Term Memory (Uzun-Kısa Süreli Bellek)
MLP	Multilayer Perceptron (Çok Katmanlı Algılayıcı)
MVC	Model-View-Controller (Uygulama Geliştirmede bir Mimari Desen)
NB	Naive Bayes
NLP	Natural Language Processing (Doğal Dil İşleme)
RF	Random Forest (Rastgele Orman)
RNN	Recurrent Neural Network (Tekrarlayan Sinir Ağı)
SVM	Support Vector Machine (Destek Vektör Makinesi)
TF	Term Frequency (Terim Frekansı)
TF-IDF	Term Frequency – Inverse Document Frequency (Terim Frekansı – Ters Doküman Frekansı)
TP	Term Presence (Terim Varlığı)

ŞEKİLLER DİZİNİ

Şekil 1.1. 2019 Yılında E-ticaret Faaliyetinde Bulunan İşletme Sayıları 2
Şekil 1.2. 2022 Yılında E-ticaret Faaliyetinde Bulunan İşletme Sayıları 3
Şekil 1.3. 2019-2022 Yılları Arasında E-ticaret Hacmi 3
Şekil 3.1. Makine Öğrenmesi Süreci 10
Şekil 3.2. Makine Öğrenmesi Yöntemleri 11
Şekil 3.3. Sınıflandırma Yöntemi 12
Şekil 3.4. Regresyon Yöntemi 12
Şekil 3.5. Kümeleme Yöntemi 13
Şekil 3.6. Takviyeli Öğrenme Yöntemi 14
Şekil 3.7. Lojistik Regresyon Yöntemi 15
Şekil 3.8. Destek Vektör Makineleri Yöntemi 16
Şekil 3.9. Rastgele Orman Yöntemi 17
Şekil 3.10. Yapay Zekâ Alt Kırılımı 18
Şekil 3.11. RNN Mimarisi 20
Şekil 3.12. LSTM Mimarisi 21
Şekil 3.13. GRU Mimarisi 22
Şekil 3.14. CNN Mimarisi 23
Şekil 3.15. Metin Madenciliği Süreci 24
Şekil 3.16. Duygu Analizi Süreci 25
Şekil 3.17. Doğal Dil İşleme 29
Şekil 4.1. Çalışma Akış Diyagramı 30
Şekil 4.2. Manuel Veri Seti Ürün Yorumlarının Kategori Dağılımı 33
Şekil 4.3. Otomatik Veri Seti Ürün Yorumlarının Kategori Dağılımı 36
Şekil 4.4. Derin Öğrenme Algoritmaları Ortalama Doğruluk Sonuçları 45
Şekil 4.5. Derin Öğrenme Algoritmaları Ortalama F1-Puanı Sonuçları 45
Şekil 4.6. Geleneksel Makine Öğrenmesi Algoritmaları N-Gram Metotları Ortalama Doğruluk Sonuçları 49
Şekil 4.7. Geleneksel Makine Öğrenmesi Algoritmaları Ağırlıklandırma Metotları Ortalama Doğruluk Sonuçları 49
Şekil 4.8. Otomatik Etiketlenmiş Veri Seti Kullanılarak Elde Edilen Ortalama Sonuçlar 50
Şekil 5.1. Uygulama Veri Çekme Modülü 53

Şekil 5.2. Veri Çekme Modülü ile Oluşturulan Dosyadan Bir Kesit 53
Şekil 5.3. Uygulama Veri Ön İşleme Modülü 54
Şekil 5.4. Veri Ön İşleme Modülü ile Oluşturulan Dosyadan Bir Kesit 55
Şekil 5.5. Uygulama Veri Etiketleme Modülü 56
Şekil 5.6. Veri Etiketleme Modülü ile Etiketlenmiş Veri Seti 56
Şekil 5.7. Veri Etiketleme Modülü Verilerin Sınıf Dağılımı Çıktısı 57
Şekil 5.8. Veri Etiketleme Modülü Seçilen Algoritmaların Doğruluk Sonuçları Çıktısı 57

TABLO DİZİNİ

Tablo 4.1. Örnek Veriler 32
Tablo 4.2. Etiketleme Sonucu Örnek Veriler 35
Tablo 4.3. Sınıflardaki Yorum Sayısı 36
Tablo 4.4. LSTM Modelinde Kullanılan Parametreler 37
Tablo 4.5. Çapraz Tahmin 39
Tablo 4.6. Manuel Veri Seti Kullanılarak Derin Öğrenme Algoritmaları Tarafından Elde Edilen Doğruluk Sonuçları 40
Tablo 4.7. Manuel Veri Seti Kullanılarak Derin Öğrenme Algoritmaları Tarafından Elde Edilen Geri Çağırma Sonuçları 41
Tablo 4.8. Manuel Veri Seti Kullanılarak Derin Öğrenme Algoritmaları Tarafından Elde Edilen Kesinlik Sonuçları 41
Tablo 4.9. Manuel Veri Seti Kullanılarak Derin Öğrenme Algoritmaları Tarafından Elde Edilen F1-Puanı Sonuçları 42
Tablo 4.10. Otomatik Veri Seti Kullanılarak Derin Öğrenme Algoritmaları Tarafından Elde Edilen Doğruluk Sonuçları 42
Tablo 4.11. Otomatik Veri Seti Kullanılarak Derin Öğrenme Algoritmaları Tarafından Elde Edilen Geri Çağırma Sonuçları 43
Tablo 4.12. Otomatik Veri Seti Kullanılarak Derin Öğrenme Algoritmaları Tarafından Elde Edilen Kesinlik Sonuçları 43
Tablo 4.13. Otomatik Veri Seti Kullanılarak Derin Öğrenme Algoritmaları Tarafından Elde Edilen F1-Puanı Sonuçları 44
Tablo 4.14. Otomatik Veri Seti Kullanılarak Geleneksel Makine Öğrenmesi Algoritmaları Tarafından Elde Edilen Doğruluk Sonuçları 46
Tablo 4.15. Otomatik Veri Seti Kullanılarak Geleneksel Makine Öğrenmesi Algoritmaları Tarafından Elde Edilen Geri Çağırma Sonuçları 47
Tablo 4.16. Otomatik Veri Seti Kullanılarak Geleneksel Makine Öğrenmesi Algoritmaları Tarafından Elde Edilen Kesinlik Sonuçları 47
Tablo 4.17. Otomatik Veri Seti Kullanılarak Geleneksel Makine Öğrenmesi Algoritmaları Tarafından Elde Edilen F1-Puanı Sonuçları 48

TEŞEKKÜR

Çalışmamın her aşamasında tüm bilgi ve deneyimleriyle yol gösteren danışman hocam Sayın Doktor Öğretim Üyesi Fatih YÜCALAR'a, desteğini hiç esirgemeyen ve tecrübeleriyle beni aydınlatan ikinci danışman hocam Sayın Doktor Öğretim Üyesi Mansur Alp TOÇOĞLU'na ve bu çalışma boyunca bana her zaman destek olan sevgili eşim Suzan ÇABUK'a yürekten teşekkür ederim.

Müjdat ÇABUK
Manisa, 2023

ÖZET

Yüksek Lisans Tezi

MAKİNE ÖĞRENMESİ İLE E-TİCARET ÜRÜN YORUMLARININ ANALİZİ

Müjdat ÇABUK

Manisa Celal Bayar Üniversitesi
Fen Bilimleri Enstitüsü
Yazılım Mühendisliği Anabilim Dalı

Danışman: Dr. Öğretim Üyesi Fatih YÜCALAR

II. Danışman: Dr. Öğretim Üyesi Mansur Alp TOÇOĞLU

Günümüzde teknolojinin gelişmesiyle birlikte e-ticaret sektörü hızlı bir büyüme sağlamıştır. E-ticaretin gelişmesiyle ürünlere yapılan yorumlar da muazzam bir şekilde artmış ve bunları tek tek inceleyip analiz etmek zorlaşmıştır. Satıcılar yorumların hepsini değerlendiremez duruma gelmiştir. Ayrıca yararlı yorumlar gözden kaçabilmekte ve yorumların tam olarak ne hakkında olduğu da tespit edilememektedir. Bu problemler karşısında yapay zekâ destekli bir otomasyon ihtiyacı ortaya çıkmıştır.

Bu çalışmada, otomasyonun ilk basamağı olarak görülen ürün yorumlarının sınıflandırılması için bir yöntem geliştirilmiştir. Öncelikle ülkemizde faaliyet gösteren e-ticaret platformlarından 15 bin 170 adet ürün yorumu toplanarak bu yorumlar olumlu, olumsuz ve nötr olarak farklı zamanlarda iki kere etiketlenmiş ve nötr yorumlar çıkarılarak iki sınıflı bir veri seti hazırlanmıştır. Bu veri seti ile Uzun-Kısa Süreli Bellek (Long-Short Term Memory – LSTM) algoritması kullanılarak bir model geliştirilmiş ve bu modelle yine e-ticaret platformlarından toplanan 203 bin 274 adet yorumun yer aldığı veri seti otomatik olarak etiketlenmiştir. Bu sayede manuel ve otomatik etiketlenmiş olarak elde edilen iki veri seti kullanılarak derin öğrenme ve geleneksel makine öğrenmesi algoritmaları ile deneyler yapılmıştır.

Çalışma sonucunda, derin öğrenme algoritmalarının makine öğrenmesi algoritmalarına kıyasla daha başarılı olduğu gözlemlenmiş ve geliştirilen modelle otomatik etiketlenen veri setinin kullanıldığı deneylerde yüksek sonuçlar elde edilmiştir. Oluşturulan sınıflandırma modelinin kullanıldığı bir web tabanlı uygulama geliştirilerek ürün yorumlarının otomatik sınıflandırılması ve akademik çalışmalarda otomatik etiketlenmiş veri seti oluşturma problemlerine çözüm sunulmuştur.

Anahtar Kelimeler: Makine Öğrenmesi, Derin Öğrenme, Metin Madenciliği, Türkçe Metinler, E-ticaret, Ürün Yorumları, Doğal Dil İşleme

2023, 65 sayfa

ABSTRACT

Master Thesis

ANALYSIS OF E-COMMERCE PRODUCT REVIEWS WITH MACHINE LEARNING

Müjdat ÇABUK

Manisa Celal Bayar University
Graduate School of Natural and Applied Sciences
Software Engineering Department

Supervisor: Asst. Prof. Fatih YÜCALAR

Co-Supervisor: Asst. Prof. Mansur Alp TOÇOĞLU

Today, with the development of technology, the e-commerce sector has grown very rapidly. With the development of the e-commerce sector, the comments made on the products have increased tremendously. As the number of reviews of the products increased, it became very difficult to examine and analyse them one by one, and the sellers became unable to evaluate all the comments made on their products. In the face of these problems, the need to create an artificial intelligence supported automation has emerged.

In this study, a method has been developed for the classification of product reviews, which is seen as the first step of automation. First of all, 15,170 product reviews were collected from e-commerce platforms operating in our country, and these comments were labelled as positive, negative and neutral twice at different times, and a two-class data set was prepared by removing neutral comments. With this data set, a model was developed using the Long-Short Term Memory algorithm, and with this model, the data set containing 203,274 comments collected from e-commerce platforms was automatically labelled. In this way, experiments were carried out with machine learning and deep learning algorithms using two manually and automatically labelled data sets.

As a result of the study, very high results were obtained in the experiments in which the automatically labelled data set with the developed model was used. By developing a web-based application in which the classification model we have created is used, we have provided solutions to the problems of automatic classification of product reviews and automatic labelled dataset creation in academic studies.

Keywords: Machine Learning, Deep Learning, Text Mining, Turkish Texts, E-commerce, Product Reviews, Natural Language Processing

2023, 65 pages

1. GİRİŞ

Günümüzde teknolojik gelişmeler tüm hızıyla devam etmektedir. Tüm sektörler bu teknolojik gelişmelerden faydalandığı gibi ticaret sektörü de büyük ölçüde faydalanmaktadır. Ticaret sektörü hem perakende hem de toptan satışlar için artık dijital kanalları tercih etmektedir. Ticaret sektöründe faaliyet gösteren firmaların ürünlerini internet üzerinden satmaları ve verdikleri hizmeti internet üzerinden vermelerine elektronik ticaret ya da kısa bilinen adıyla e-ticaret denilmektedir.

E-ticaret internet üzerinden ürün veya hizmetlerin satışını gerçekleştiren işletme faaliyetidir. İşletmeler online mağaza veya platformlar aracılığıyla müşterilere ürünlerini tanıtırlar, sipariş alırlar, ürünlerini teslim ederler. E-ticaret müşterilere çeşitli ürünlerin geniş bir yelpazesine erişim sağlar ve fiziksel bir mağazaya gitmeye gerek kalmadan kolaylıkla alışveriş yapmalarına olanak sağlar. E-ticaretin avantajları arasında daha geniş bir müşteri kitlesi, 7/24 açık işletme avantajı, düşük işletme maliyetleri, müşteri verilerinin kolaylıkla toplanması ve işlem maliyetlerinin azalması yer almaktadır.

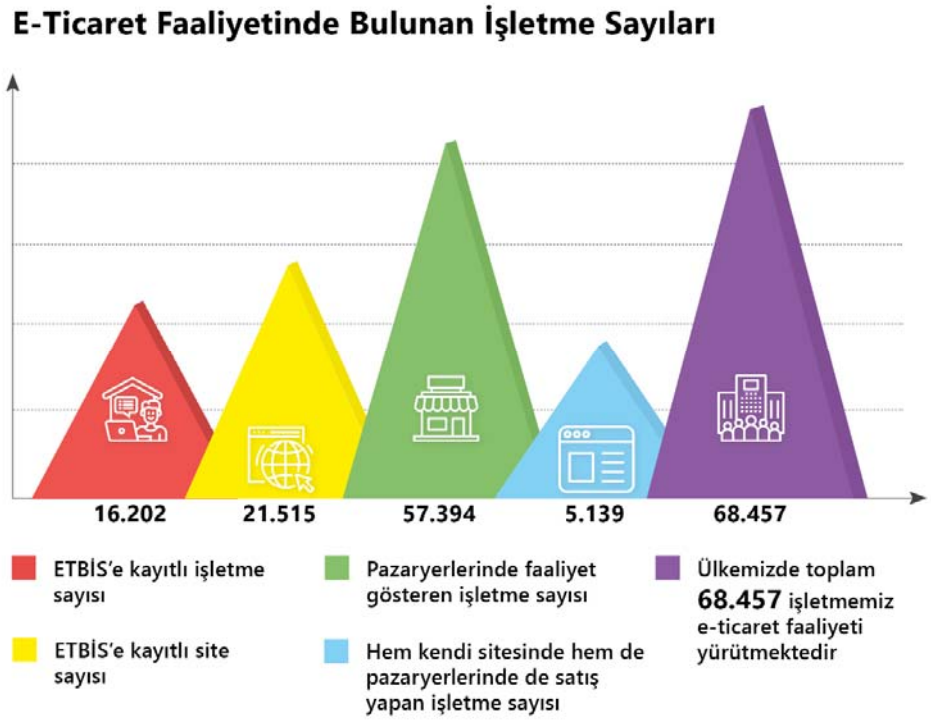
Temeli daha eskiye dayansa da 1995 yılında Amazon ve eBay şirketlerinin kurulması e-ticaret sektörünün hızla gelişmesi ve yaygınlaşmasını sağlamıştır [1]. Önceleri bilgisayar ekipmanları gibi ürünlerin satışı için kullanılan e-ticaret artık ürün yelpazesini geliştirerek kitap, müzik, kıyafetler gibi ürün kategorilerinde de varlık göstermiştir. Farklı e-ticaret firmalarının da kurulumuyla birlikte 2000'li yılların başlarında birçok insan internet üzerinden alışveriş yapmaya başlamıştır. Aynı zamanda ticaret sektöründe faaliyet gösteren birçok firma da e-ticaret siteleri üzerinden hizmet vermeye başlamıştır. Bu dönemde PayPal ve diğer dijital ödeme araçları da ortaya çıkarak e-ticaret alışverişlerinin güvenliği arttırılmıştır.

Ülkemizde ise 2000'li yılların başına dayanan e-ticaret sektörü 2007 yılında n11.com platformunun kurulmasıyla tam anlamıyla başlamıştır. 2010'larda e-ticaret sektörü hızla büyümeye devam etmiştir. Bu dönemde Hepsiburada, Trendyol, GittiGidiyor gibi birçok e-ticaret platformu açılmıştır [2].

Mobil cihazların ortaya çıkması ve yaygınlaşmasıyla e-ticaret sektörü adeta bir devrim yaşamıştır. İnsanlar artık kolaylıkla mobil cihazlarıyla e-ticaret platformlarına ulaşabilmekte ve saniyeler içerisinde ihtiyacı olan ürünleri alabilmektedirler. E-ticaret işletmeleri sosyal medya pazarlaması ve yapay zekâ gibi teknolojileri kullanarak müşteri deneyimini daha da geliştirmeye odaklanmaktadırlar. Günümüzde e-ticaret sektörü hala hızlı bir şekilde büyümekte ve daha önemli bir pazar haline gelmektedir. İnsanlar online alışveriş yapmak için fiziksel mağazalara gitmek yerine interneti tercih etmektedirler ve birçok geleneksel mağazada bu trende ayak uydurarak online satışlara odaklanmaktadır.

Ticaret Bakanlığı'nın son açıkladığı e-ticaret verileri ile 2019 yılının verilerini karşılaştırdığımızda hem işletme sayılarında hem de e-ticaret hacminde büyük bir artış olduğu görülmektedir. Ülkemizde 2019 yılında e-ticaret platformlarında hizmet veren işletme sayısı 68 bin 457 iken günümüzde bu sayı 548 bin 688'e ulaşmıştır. 2022 yılı e-ticaret hacmi ise 2019 yılına göre yaklaşık 7 kat artarak 800,7 milyar TL olduğu görülmüştür [3].

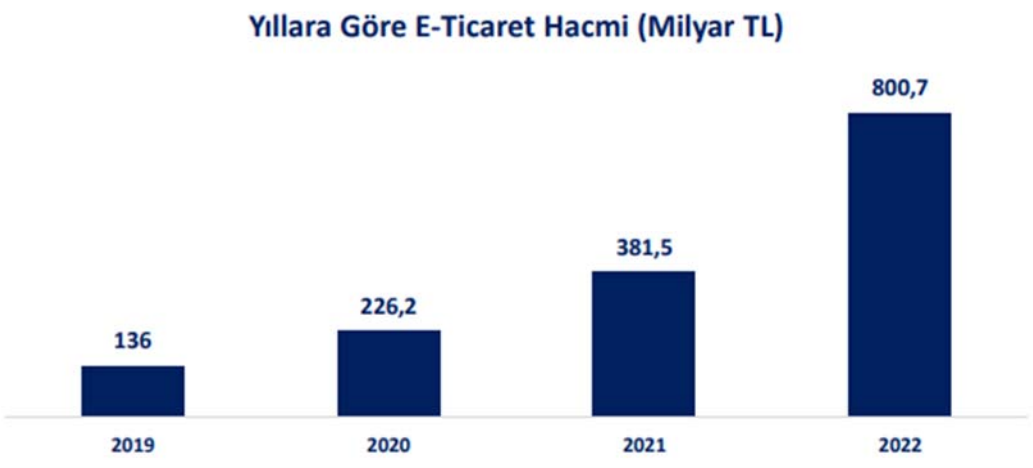
Şekil 1.1, Şekil 1.2 ve Şekil 1.3'te Ticaret Bakanlığı'nın E-Ticaret Bilgi Platformunda paylaştığı istatistikler grafik olarak gösterilmektedir.



Şekil 1.1. 2019 Yılında E-ticaret Faaliyetinde Bulunan İşletme Sayıları



Şekil 1.2. 2022 Yılında E-ticaret Faaliyetinde Bulunan İşletme Sayıları



Şekil 1.3. 2019-2022 Yılları Arasında E-ticaret Hacmi

E-ticaret sektörü son yıllarda yapay zekâ teknolojilerinin kullanımıyla önemli bir dönüşüm yaşamaktadır. Yapay zekâ, e-ticaret şirketlerinin müşterileri daha iyi anlamalarına ve onların ihtiyaçlarına uygun ürünler sunmalarına yardımcı olmaktadır. Aşağıda yapay zekâ teknolojilerinin e-ticaret sektöründe kullanıldığı bazı alanlardan bahsedilmiştir.

- *Kişiselleştirilmiş Öneriler:* E-ticaret platformları müşterilerin önceki alışveriş deneyimlerine ve arama terimlerine dayanarak kişiselleştirilmiş ürün önerileri sunabilmektedirler. Bu öneriler müşterilerin ilgi alanlarına ve satın alma geçmişlerine göre yapılandırılmaktadır.

- *Chatbotlar:* Chatbotlar, müşterilerin sorularına hızlı ve doğru yanıtlar vermek için kullanılır. Yapay zekâ teknolojileri chatbotların daha akıllı hale gelmesini sağlayarak müşteri memnuniyetini arttırır.
- *Dolandırıcılık Önleme:* E-ticaret şirketleri dolandırıcılık faaliyetlerini önlemek için yapay zekâ teknolojilerini kullanmaktadırlar. Bu teknolojiler dolandırıcılık şüphesi olan işlemleri tespit ederek önlem alınmasını sağlar.
- *Stok Yönetimi:* Yapay zekâ teknolojileri ürün taleplerini tahmin ederek stok yönetimini optimize etmeye çalışırlar. Bu e-ticaret şirketlerinin stoklarını daha iyi yönetmelerini ve müşteri memnuniyetini arttırmaktadır.

Yapay zekâ teknolojileri e-ticaret işletmelerine müşteri memnuniyetini arttırmanın yanı sıra işlemlerini daha verimli hale getirme imkânı da sağlar. Bu teknolojiler e-ticaret şirketlerinin rakipleri arasında fark yaratmalarına yardımcı olmaktadır.

E-ticarette ürün yorumları müşterilerin bir ürün hakkında görüşlerini paylaştığı ve diğer müşterilerin de bu görüşleri inceleyerek ürün hakkında bir fikir sahibi olabilecekleri önemli bir unsurdur. Ürün yorumları e-ticaret şirketleri için çok önemli bir geri bildirim mekanizmasıdır ve hem müşterilerin memnuniyetini arttırmaya hem de şirketlerin pazarlama stratejilerini belirlemeye yardımcı olur.

Ürün yorumları müşterilerin ürünler hakkında doğru ve açıklayıcı geri bildirimler sunmasını sağlar. Bu yorumlar müşterilerin ürünleri nasıl kullandığı, hangi özelliklerinin beğenilip beğenilmediği, ürünün kalitesi gibi unsurlar hakkında bilgi vermektedir. Bu geri bildirimler şirketlerin ürünlerini geliştirmelerine yardımcı olur ve müşteri memnuniyetini arttırır. Ayrıca ürün yorumları diğer müşterilerin de ürün hakkında bilgi sahibi olmalarını sağlamaktadır. Müşterilerin gerçek deneyimlerine dayalı olarak yazdığı yorumlar diğer müşterilerin ürün hakkında karar vermelerine yardımcı olmaktadır. Bu nedenle ürün yorumları e-ticaret şirketleri için önemli bir pazarlama aracıdır.

E-ticaret sektörünün gelişmesiyle ürünlere yapılan yorumlar da muazzam bir şekilde artmıştır. Ürünlerin yorum sayıları arttıkça bunları tek tek inceleyip analiz etmek oldukça zorlaşmıştır. Satıcılar ürünlerine yapılan yorumların hepsini

değerlendirememekte ve yararlı olabilecek yorumlar gözden kaçabilmektedir. Ayrıca yapılan yorumların tam olarak ne hakkında olduğu da manuel olarak tespit edilemez hale gelmiştir. Müşteriler için de aynı problem geçerlidir. Müşteriler ürün hakkındaki yorumların en fazla iki üç sayfasını okuyabilmektedir. Tüm bu problemlerin karşılığında e-ticaret ürün yorumlarının otomatik analiz edilebilmesi gerekmektedir.

Bu çalışmada e-ticaret ürün yorumlarının duygu analizi yöntemi üzerine çalışılmıştır. Çalışmanın ilk aşamasında ülkemizde faaliyet gösteren e-ticaret platformlarından Türkçe yorumlar toplanarak ham bir veri seti oluşturulmuştur. Ürün yorumlarını toplamak için bir uygulama geliştirilmiş ve bu uygulama içerisinde Web Scraping yöntemi kullanılmıştır. Ürün yorumları farklı zamanlarda iki kere olumlu, olumsuz ve nötr olarak manuel sınıflandırılmıştır. Sonrasında nötr yorumlar ve kararsız kalınan yorumlar çıkarılarak iki sınıflı bir veri seti oluşturulmuştur. Çalışmanın ikinci aşamasında Uzun-Kısa Süreli Bellek Ağları (Long-Short Term Memory – LSTM) derin öğrenme yöntemi kullanılarak otomatik bir veri seti oluşturulmuştur. Sonraki aşamada, oluşturulan manuel ve otomatik veri setleri üzerinde Tekrarlayan Sinir Ağları (Recurrent Neural Network – RNN), Uzun Kısa Süreli Bellek Ağları (Long-Short Term Memory – LSTM), Geçitli Tekrarlayan Birim (Gated Recurrent Unit – GRU) ve Evrişimli Sinir Ağları (Convolutional Neural Network – CNN) gibi derin öğrenme algoritmaları kullanılarak duygu analizi test sonuçları elde edilmiştir. Testler sonucunda otomatik yöntemle elde edilen veri setiyle eğitilen makine öğrenmesi modellerinin manuel veri setiyle eğitilen modellere yakın sonuçlar verdiği gözlemlenmiştir. Sonrasında otomatik etiketlenmiş veri seti ile Rastgele Orman (Random Forest – RF), Lojistik Regresyon (Logistic Regression – LR), Naive Bayes (NB), Destek Vektör Makineleri (Support Vector Machine – SVM) gibi temel geleneksel makine öğrenmesi algoritmaları kullanılarak testler yapılmıştır. Bunun sonucunda da derin öğrenme mimarilerinin geleneksel makine öğrenmesi algoritmalarından daha iyi sonuçlar verdiği gözlemlenmiştir. Ürün yorumlarını çekmek için geliştirilen uygulamaya geliştirdiğimiz sınıflandırma modelleri de entegre edilmiştir. Ayrıca uygulamaya veri ön işleme modülü de ekleyerek makine öğrenmesi süreçlerinin tamamını içeren bir uygulama geliştirilmiştir.

2. LİTERATÜR TARAMASI

Bu çalışma kapsamında öncesinde literatürde yer alan çalışmalar incelenmiştir. Bu aşamada öncelikle Türkçe metinler ile yapılan çalışmalar ve ağırlıklı olarak müşteri yorumlarının yer aldığı veri setleri ile yapılan çalışmalar incelenmiştir. İncelenen çalışmalar arasında İngilizce metinlerden oluşan veri setleri ile yapılan çalışmalar da yer almaktadır. Bu çalışmalar kullanılan teknikler için tercih edilmiştir.

Rumelli ve ark. [5], bir çalışmalarında e-ticaret ürün yorumları kullanarak duygu analizi modeli geliştirmişlerdir. Bu çalışmada hepsiburada.com e-ticaret sitesi üzerinden 272 bin adet ürün yorumu çekilerek bir veri seti hazırlanmıştır. Bu veri setinde 1 ve 2 yıldız olan yorumlar olumsuz, 3 yıldız olanlar nötr, 4 ve 5 yıldız olanlar da olumlu olarak etiketlenmiştir. Son olarak 13 bin olumlu ve 13 bin olumsuz yorum seçilerek çalışmada kullanılacak veri seti hazırlanmıştır. Çalışmada NB, RF, SVM ve K – En Yakın Komşuluk (K-Nearest Neighbours – KNN) algoritmaları kullanılarak sonuçlar alınmıştır. Çalışma sonucunda ortalama %73 doğruluk oranı yakalanmıştır.

Topçu ve Erkaya [6], Müşteri İlişkileri Yönetimi (Customer Relations Management – CRM) yazılımlarında toplanan talep mesajlarının makine öğrenmesi kullanarak kategorize edilmesi amaçlı bir çalışma yapmışlardır. Çalışmada vatandaşlar tarafından gönderilen 3 bin farklı konudan oluşan 220 bin talep mektubu kullanılmıştır. Veriler ön işlemlerden geçirildikten sonra TF-IDF algoritması kullanılarak özellik seçimi yapılmıştır. Veri seti 5 farklı algoritma kullanılarak test edilmiştir. En verimli algoritmanın %85 doğruluk oranıyla LSTM olduğu görülmüştür. Çalışma sonucunda oluşturulan model örnek olarak "1314. sokakta bozuk yollar var. Bir an önce asfalt atılmasını ve yolun onarılmasını talep etmekteyim. Gereğini arz ederim." şeklinde gelen talebi "Asfalt Bakım Onarım" olarak kategorilendirmiştir.

Tuzcu [7], kitap yorumlarının duygu analizi yöntemleri ile sınıflandırılması çalışması yapmıştır. Popüler bir çevrimiçi kitap platformundan topladıkları okuyucu yorumlarından bir veri seti hazırlamıştır. Bu çalışmada Çok Katmanlı Algılayıcı (Multilayer Perceptron – MLP) algoritmasını uygulamıştır. Sonrasında NB, SVM ve LR gibi geleneksel makine öğrenmesi algoritmaları ile de testler yapmış ve sonuçları

karşılaştırmıştır. Karşılaştırma sonucunda %89 başarı oranı ile MLP algoritmasının en yüksek başarı elde ettiği görülmüştür.

Mengutayci ve Temurtas [8], Türkçe otel yorumlarının yapay sinir ağları ile sınıflandırılması konulu bir çalışma yapmıştır. Çalışmasında çevrimiçi otel rezervasyonu yapılan bir platformda yer alan müşteri yorumlarını incelemiştir. Bu yorumların 4-5 yıldız olanları olumlu, 1-2 yıldız olanları olumsuz olarak etiketlenmesini sağlayarak iki sınıflı bir çalışma yapmıştır. Veri setini gerekli ön işlemlerden geçirdikten sonra TF-IDF modeli ile yorumlar bir matrise dönüştürülmüş ve her terimin sınıflandırmadaki önem derecesinin belirlenmesini sağlamıştır. Veri setini 2 farklı yöntemle eğitim ve test verilerine ayırmış ve bu veri setlerinin başarısı test edilmiştir. Testlerini yapay sinir ağı mimarileriyle yapmıştır.

Mayda ve Korkmaz [9], ortak çalışmalarında müşteri yorumlarının sınıflandırılması için sözlük tabanlı bir yaklaşım kullanmışlardır. Yazarlar tarafından bir sıfat sözlüğü hazırlanmıştır. Kitap satışının yapıldığı bir e-ticaret platformundan topladıkları okuyucu yorumlarından bir veri seti hazırlamışlar ve sıfat sözlüğü kullanarak olumlu, olumsuz, nötr sınıfları için bir duygu analizi çalışması yapmışlardır. Ayrıca çalışmalarında Türkçe diline özgü kurallar geliştirerek de çalışmanın başarısını arttırmayı başarmışlardır. Çalışma sonucunda yaklaşık %62'lik doğruluk oranı elde etmişlerdir.

Gezici ve Yanıkoğlu [10], bir çalışmalarında makine öğrenmesi algoritmalarıyla Türkçe film incelemelerinin duygu analizini yapmışlardır. Bu çalışmalarında denetimli öğrenme ve sözlük tabanlı yaklaşımları birleştirerek bir yöntem sunmuşlardır. Çalışmalarında Beyazperde adlı film sitesinden önceden oluşturulan bir veri seti kullanılmış ve bu veri setindeki incelemelerin 4 ve 5 yıldızlı olanları olumlu, 1 ve 2 yıldız olanları olumsuz olarak etiketleyerek ikili sınıflandırma çalışması yapmışlardır. NB ve SVM algoritmalarının sonuçları karşılaştırılmıştır. Bu karşılaştırma esnasında sözlük tabanlı yaklaşım ile denetimli öğrenme yaklaşımlarının sonuçları da göz önünde bulundurulmuştur. Sonuçlarda NB algoritmasının doğruluk oranı %75 olduğu görülmüştür.

Pervan ve Keleş [11], RF algoritması ile Türkçe yorumlarda duygu analizi çalışması yapmışlardır. Popüler bir e-ticaret platformundan elektronik kategorisi altında yer alan ürünlere yapılan müşteri yorumlarından bir veri seti hazırlamışlar ve etiketleme işlemini yıldız puanına göre otomatik yapmışlardır. RF algoritmasıyla yaklaşık %85 doğruluk oranı elde etmişlerdir.

Onan [12], bir çalışmasında denetimli ve denetimsiz terim ağırlıklandırma şemaları modellerini karşılaştırmıştır. Veri seti olarak Türkçe Twitter mesajlarını kullanmıştır. Mesajları pozitif ve negatif olarak iki kişi tarafından etiketlenmiş ve gerekli ön işlemlerden geçirilmiştir. Sonrasında KNN, SVM, LR ve NB algoritmalarıyla testler yapmıştır. Bu testlere AdaBoost, Bagging ve Random SubSpace topluluk öğrenme yöntemleriyle devam etmiştir. Testler sonucunda denetimli terim ağırlıklandırma modellerinin denetimsiz terim ağırlıklandırma modellerine göre daha başarılı olduğu gözlemlenmiştir.

Demircan ve ark. [13], bir çalışmalarında Türkçe metinlerde duygu analizi yöntemi sunmuşlardır. Veri seti olarak e-ticaret ürün yorumlarını kullanmışlar ve veri etiketlenmesi için ürün yorumlarındaki puanlardan faydalanmışlardır. SVM, RF, DT, LR ve KNN algoritmalarıyla modeller geliştirip testlerini yapmışlardır. Testler sonucunda SVM ve RF modellerinin daha iyi sonuçlar verdiği gözlemlenmiştir.

Toçoğlu ve ark. [14], Türkçe tweetlerden duygu analizi çalışması yapmışlardır. Çalışma için oluşturdukları veri setini sözlük tabanlı bir yaklaşım ile her tweete altı duygu kategorisi için otomatik olarak açıklama eklemişlerdir. Çalışmalarında ANN, CNN, LSTM ve RNN mimarileri inceleyerek bunlar üzerinden testler yapmışlardır. Testler sonucunda sözlük tabanlı yaklaşım ile otomatik eklenen açıklamaların etkisini gözlemlemişlerdir. Doğruluk puanları karşılaştırıldığında en yüksek puanı %74 ile CNN mimarisinin verdiğini gözlemlemişlerdir.

Çataltaş ve ark. [15], bir çalışmalarında e-ticaret olumsuz ürün yorumlarından ürünlerin kusurlu özelliklerini bulmak için bir metin analizi yöntemi önermişler ve bulunan ürün kusurlarından da yapılandırılmış bir özet oluşturmaya odaklanmışlardır. Amazon.com üzerinden ayakkabı kategorisine ait 10 bin olumsuz yorum çekerek bir veri seti oluşturmuşlardır. Ön işlemlerin ardından veri sayısı 6190'a inmiş ve çalışmayı

bu verilerle yapmışlardır. Ürün kusurlarını bulmak için DBSCAN algoritması kullanılarak yorumlar kümelere ayrılmıştır. Bulunan her kusur daha önceden tanımlanmış Part-of-Speech (Konuşma Parçası) kalıplarını arayarak kendisinin görüş sözcüklerini bulmak için kullanılmıştır. Bulunan kusurlar ve onların görüş sözcükleri kullanılarak yapılandırılmış bir özet oluşturulmuştur.

Literatürde Türkçe metinlerle yapılan çalışmalar günümüzde gittikçe artmaktadır. İncelenen çalışmalarda çoğunlukla çevrimiçi müşteri yorumlarından oluşan veri setleri kullanıldığı görülmüştür. Çalışmalarda öncelikle veri etiketleme işlemleri karşılaştırılmış ve manuel etiketlemenin daha doğru sonuçlar verdiği görülmüştür.

3. GENEL BİLGİLER

Bu başlık altında çalışmada kullanılan makine öğrenmesi ve derin öğrenme kavramlarına değinilerek, bu kavramların altında yer alan algoritmalar açıklanmıştır. Ayrıca çalışmanın amacı olarak görülen metin madenciliği, duygu analizi ve doğal dil işleme kavramlarına da değinilmiştir.

3.1. Makine Öğrenmesi

Makine öğrenmesi, yapay zekanın bir alt dalıdır. Bilgisayar sistemlerinin insan müdahalesi olmadan çeşitli algoritma ve teknikler kullanarak verileri analiz etmesini ve tahminler yapmasını amaçlamaktadır. Makine öğrenmesi, yüksek boyutlu ve karmaşık verileri işleyerek sonuçlar çıkarmak ve gelecekteki olayları öngörmek için kullanılabilir [16].

Makine öğrenmesi genellikle öğrenme, test ve kullanma olmak üzere üç aşamadan oluşur. Öğrenme aşamasında belirli bir veri kümesi üzerinde algoritmalar kullanılarak model eğitilir. Test aşamasında eğitilen model başka bir veri kümesi üzerinde test edilir. Bu test sonucunda eğitilen modelin başarısı ölçülür. Kullanma aşamasında, eğitilen ve test edilen modeller gerçek dünya verileri ile kullanılarak sonuçlar üretilir ve bu karşılaşılan problemleri çözmeye fayda sağlar. Şekil 3.1'de makine öğrenmesi süreci görülmektedir.



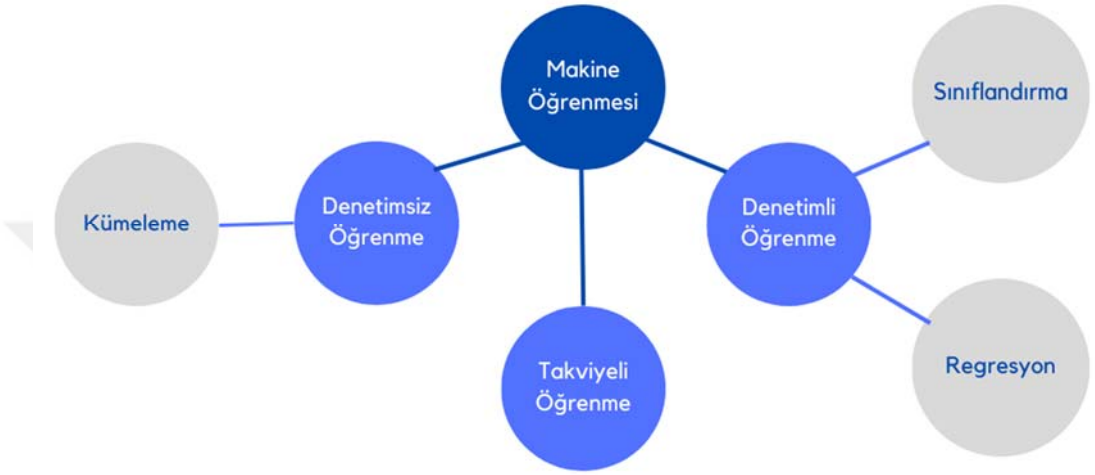
Şekil 3.1. Makine Öğrenmesi Süreci

Makine öğrenmesi, günümüzde eğitim, sağlık, finans, otomotiv ve e-ticaret gibi alanlarda kullanılmaktadır. Örneğin eğitim alanında öğrenci performanslarını izlemek, öğrencilerin davranışlarını analiz etmek ve eğitim materyalleri oluşturma gibi konularda makine öğrenmesi teknikleri kullanılmaktadır. Finans sektöründe ise makine öğrenmesi, yatırım yönetimi, risk yönetimi ve sahtecilik tespiti gibi konularda

kullanılmaktadır. Bu çalışmada ise makine öğrenmesi, e-ticaret sektöründeki müşteri yorumlarının analizi yapılarak sektöre fayda sağlamak amaçlanmıştır.

3.1.1. Makine Öğrenmesi Yöntemleri

Şekil 3.2'de görüldüğü üzere makine öğrenmesi literatürde denetimli öğrenme, denetimsiz öğrenme ve takviyeli öğrenme olarak üç farklı kategoriye ayrılır [16].



Şekil 3.2. Makine Öğrenmesi Yöntemleri

3.1.1.1. Denetimli Öğrenme

Makine öğrenmesinin bu kategorisinde algoritmaların belirli bir sonuca varması için önceden etiketlenmiş veri setleri kullanılmaktadır. Denetimli öğrenme algoritmaları etiketlenmiş veri setlerini kullanarak yeni verilerin doğru bir şekilde sınıflandırılması veya tahmin edilmesi için eğitilmektedir [17]. Örneğin bir algoritmanın bir cümlenin olumlu veya olumsuz olduğuna karar verebilmesi için önceden etiketlenmiş olumlu ve olumsuz cümlelerin yer aldığı bir veri seti ile eğitilmesi gerekmektedir. Denetimli öğrenme yöntemini kullanan algoritmalar arasında destek vektör makineleri, lojistik regresyon gibi algoritmalar yer almaktadır.

Denetimli öğrenme, birçok uygulamada kullanılabilir. Örneğin, doğal dil işleme, görüntü işleme, fiyat tahminleme gibi alanlarda kullanılabilmektedir. Doğal dil işlemede, belirli bir metnin ne anlama geldiğini tahmin etmek için kullanılabilir. Görüntü işlemede ise belirli bir görüntüdeki nesneleri tanımlamak için kullanılabilir.

Denetimli öğrenme algoritmalarının sınıflandırma ve regresyon olmak üzere iki temel türü vardır [18]. Şekil 3.3'te görüldüğü üzere sınıflandırma algoritmalarında verileri farklı sınıflara ayıran bir model oluşturma süreci vardır. Örneğin, bir e-posta mesajının spam olup olmadığının kontrolü için sınıflandırma modelleri kullanılabilir. Şekil 3.4'te görülen regresyon algoritmalarında ise verilere dayanarak bir sonucu tahmin etme süreci vardır. Örneğin, bir reklam kampanyasının başarısını ölçmek için regresyon modelleri kullanılabilir.



Şekil 3.3. Sınıflandırma Yöntemi

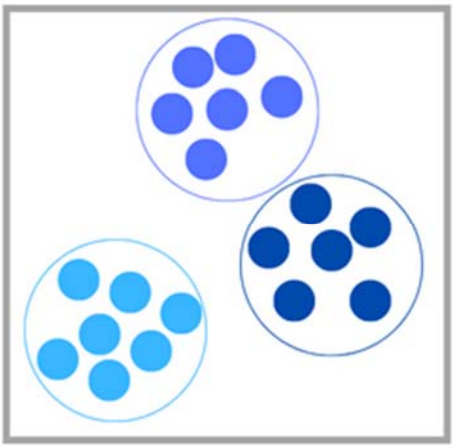


Şekil 3.4. Regresyon Yöntemi

3.1.1.2. Denetimsiz Öğrenme

Denetimsiz öğrenme, denetimli öğrenmenin aksine etiketlenmemiş veri setleri kullanılmaktadır. Bu yaklaşımda algoritmaların belirli bir sonuca varması beklenmez, verilerin gruplanması veya kümelenmesi beklenir [18].

Denetimsiz öğrenme algoritmaları, özellikle büyük veri kümelerinde yapısal olarak bilinmeyen kalıpların ve yapıların tanınması gerektiğinde kullanılır. Denetimsiz öğrenme algoritmalarının birkaç türü vardır. Bunlardan en yaygını ise kümeleme algoritmalarıdır. Şekil 3.5'te temel yapısı görülen kümeleme yönteminde benzer özelliklere sahip veriler gruplanır. Bu algoritma ile bir müşteri veri setinde yer alan veriler müşterilerin demografik özelliklerine göre kümelenebilir.



Şekil 3.5. Kümeleme Yöntemi

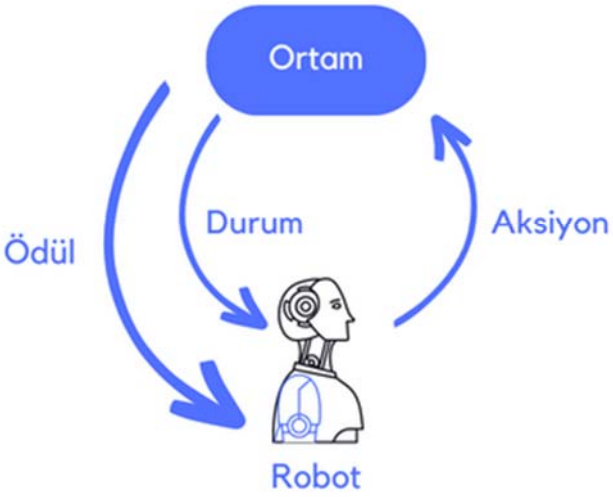
Denetimsiz öğrenme yöntemi birçok farklı alanda kullanılmaktadır. Örneğin finansal piyasalardaki dalgalanmaların veya sosyal medya verilerindeki trendlerin analizi gibi durumlarda denetimsiz öğrenme yöntemleri kullanılabilir. Başka bir örnek olarak, sağlık sektöründe benzer genlerin birbirine yakın olduğu bölgeleri belirlemek için kullanılabilir. Ayrıca bu yöntem denetimli öğrenme algoritmalarının veri etiketleme gereksinimleri nedeniyle de kullanılamadığı durumlarda kullanılabilir.

3.1.1.3. Takviyeli Öğrenme

Takviyeli öğrenmede bir modelin doğru kararlar verebilmesi için belirli bir görevde çalışması ve bu sayede öğrenmesi sağlanır. Bu öğrenme yöntemi denetimli ve denetimsiz öğrenme yöntemlerinin bir kombinasyonu olarak düşünülebilir.

Takviyeli öğrenme yöntemi oyun veya robotik sistemler gibi etkileşimli uygulamalarda kullanılmaktadır. Örneğin oyun oynayan bir bilgisayar programında rakiplerine karşı daha iyi oynayabilmesi için takviyeli öğrenme yöntemleri kullanılabilir. Bu durumda program bir hamle yaptığında sonucunun ne kadar iyi veya kötü olduğunu belirleyen bir ödül alır. Her zaman daha iyi ödüller almak için daha iyi hamleler yapmayı öğrenir [19].

Takviyeli öğrenme yöntemi çevresel faktörlerin ve belirsizliklerin olduğu durumlarda da kullanılmaktadır. Örneğin robotların çevresel faktörleri anlamaları ve bunlara göre hamleler yapması gerektiğinde takviyeli öğrenme yöntemleri kullanılır. Bu yapı Şekil 3.6'da görülmektedir.



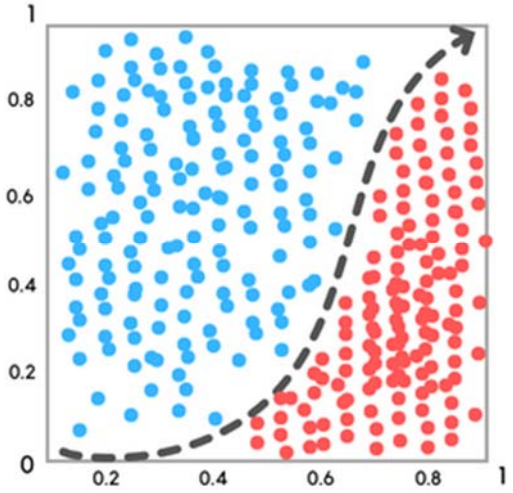
Şekil 3.6. Takviyeli Öğrenme Yöntemi

3.1.2. Makine Öğrenmesi Algoritmaları

Bir önceki başlıkta bahsedilen makine öğrenmesi yöntemlerinde kullanılan birçok algoritma bulunmaktadır. Bu algoritmalar kullanılan veri seti ve çalışma sonucunda istenilen sonuçlara göre değişkenlik göstermektedir. Bu başlık altında tez çalışması içerisinde ele alınan makine öğrenmesi algoritmaları açıklanmıştır.

3.1.2.1. Lojistik Regresyon

Lojistik Regresyon (Logistic Regression – LR) algoritması, veri setindeki değişkenler arasındaki ilişkiyi tanımlamak için kullanılır ve sınıflandırılmış bir değer olarak çıktı verir. Bir bağımlı değişkenin iki veya daha fazla kategorik veya sayısal bir değer alabileceği veri setlerinde kullanılabilir. Bu algoritma öncelikle bir doğrusal regresyon yöntemi kullanarak bir çıktı verir ve sonra bu çıktı değerini bir olasılık değerine dönüştürür. Şekil 3.7'de görüldüğü üzere bu olasılık değeri, bir sınıfa ait olma olasılığını temsil eder [20].



Şekil 3.7. Lojistik Regresyon Yöntemi

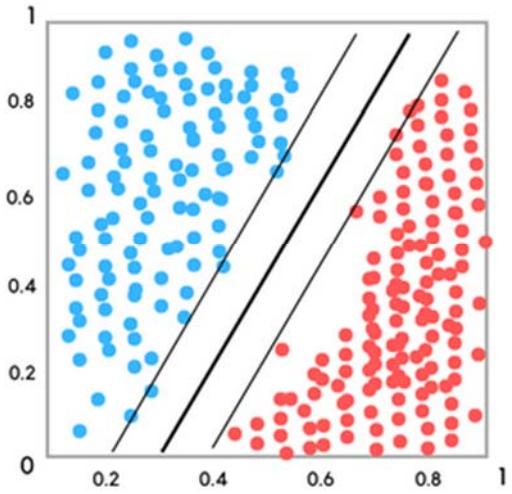
Lojistik regresyon algoritması, pazarlama, finans, tıp gibi birçok alanda kullanılır. Örneğin, tıbbi bir teşhis koyulurken, hastanın belirli bir hastalığa sahip olma olasılığı gibi durumlarda lojistik regresyon kullanılabilir.

3.1.2.2. Destek Vektör Makineleri

Destek Vektör Makineleri (Support Vector Machine – SVM) algoritması sınıflandırma ve gruplandırma problemlerinde kullanılmaktadır. Veri noktalarını birbirinden ayıran bir çizgi oluşturarak sınıflandırma işlemini yapmaktadır. SVM algoritması genellikle çözümlenmesi zor olan büyük veri setlerinde etkili olduğu görülmüştür [21].

SVM algoritması, verileri birbirinden ayıran en iyi hiper düzlemi bulmak için çalışır. Hiper düzlem verilerin tam olarak ayrılamadığı durumlarda da kullanılabilir. Bu durumda bir hata payı belirlenir ve hiper düzlem en düşük hata payını sağlayacak

şekilde oluşturulur. Bu algoritma özellikle doğrusal olmayan veri setleri için etkilidir. Bu veri setlerinde verileri daha yüksek boyutlu bir uzaya taşır ve orada daha kolay sınıflandırılabilir hale getirir. Bu işlem çekirdek fonksiyonları kullanılarak yapılır.



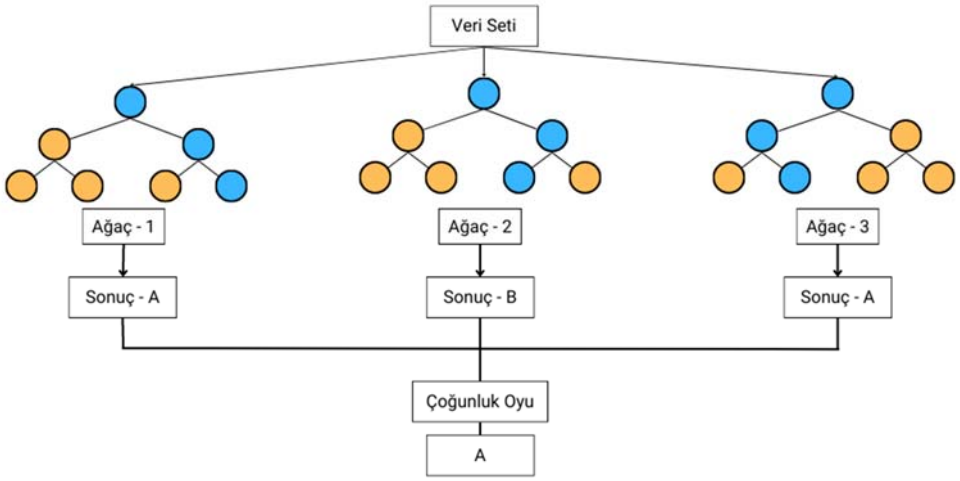
Şekil 3.8. Destek Vektör Makineleri Yöntemi

Şekil 3.8'de görüldüğü üzere SVM algoritması sınıflandırma işlemi yaparken sınıfların ötesinde kalan verileri de dikkate alarak daha başarılı sonuçlar verir. Bu algoritma özellikle veri boyutu büyük olduğunda diğer algoritmalara göre daha hızlı çalışır.

3.1.2.3. Rastgele Orman

Rastgele Orman (Random Forrest – RF) algoritması, bir ağaç yapısı kullanarak bir sınıflandırma modeli oluşturur ve veri kümesindeki öznitelikleri test ederek sonuçlar elde eder. Bu algoritma veri setini daha küçük parçalara ayırarak her bir parçayı ayrı ayrı öğrenir ve sonrasında bu parçaları birleştirerek tahminler yapar [22].

RF algoritması yüksek boyutlu verilerin sınıflandırılmasında kullanılır ve doğrusal olmayan ilişkileri tespit edebilir. Bu algoritma her bir karar ağacının kendine özgü bir eğitim seti ve farklı özellikler ile eğitildiği bir topluluk öğrenmesi yöntemidir. Bu sayede RF algoritması öğrenme aşamasında daha esnek ve doğru sonuçlar vermektedir. Şekil 3.9'da RF algoritması genel yapısı görülmektedir.



Şekil 3.9. Rastgele Orman Yöntemi

Bu algoritma geniş bir veri setinde doğru bir şekilde sınıflandırma yapmak için kullanılabilir. Ayrıca modelin eğitilmesi aşamasında aşırı öğrenmeyi önlemek için karar ağacı sayısını arttırmak ve bir ağaçta kullanılan özellik sayısını sınırlamak da mümkündür. Bu sayede doğru bir sınıflandırma yapan model geliştirilebilir.

3.1.2.4. Naive Bayes

Naive Bayes (NB) algoritması, bayes teoremini kullanarak veri setindeki özellikleri ve sınıfları birbirine bağlar. Özellikle veri setindeki özelliklerin birbirinden bağımsız olduğunu varsayar ve "naive" adı da buradan gelmektedir.

NB algoritması sınıflandırılacak verinin özelliklerini kullanarak her sınıfın olasılık dağılımını hesaplar. Bu dağılımlar eğitim verileri üzerinde hesaplanarak öğrenilir. Sonrasında yeni bir veri geldiğinde bu verinin her bir sınıfta olma olasılığı hesaplanır ve sonucunda en yüksek olasılığa sahip sınıf seçilir [21].

NB algoritması genellikle sınıflandırma problemlerinde kullanılan bir algoritmadır. Örneğin doğal dil işleme, duygu analizi ve spam filtreleme gibi problemlerde sıklıkla kullanılmaktadır.

$$P(A/B) = \frac{P(B/A)*P(A)}{P(B)} \qquad (1)$$

Denklem (1)'de Bayes Teoreminin formülü gösterilmektedir. Denklemde $P(A/B)$ değeri B olayı gerçekleştiğinde A olayının gerçekleşme ihtimalini gösterir. $P(B/A)$ değeri ise tam tersi olarak A olayı gerçekleştiğinde B olayının gerçekleşme ihtimalini gösterir. $P(A)$ değeri A olayının gerçekleşme ihtimalini ve $P(B)$ değeri ise B olayının gerçekleşme ihtimalini gösterir.

3.2. Derin Öğrenme

Şekil 3.10'da görüldüğü üzere derin öğrenme, makine öğrenmesinin bir alt dalıdır. Yapay sinir ağları ve makine öğrenimi tekniklerini kullanarak karmaşık veri yapılarını öğrenmeyi ve anlamayı amaçlar. Yapay sinir ağları, insan beyninin nöronlarının işleyişini taklit eden matematiksel modellerdir. Derin öğrenme algoritmaları, yüksek hacimli, yüksek boyutlu verilerdeki kalıpları tanımlayarak verileri sınıflandırmak, tanımak veya tahmin etmek için kullanılır [23].



Şekil 3.10. Yapay Zekâ Alt Kırılımı

Derin öğrenmenin temelinde çok katmanlı sinir ağları vardır. Bu ağlar birbiriyle bağlantılı çok sayıda nörondan oluşur. Bu nöronlar veri setinden girdiler alır ve bu girdileri işleyerek sonuç üretirler. Buradan üretilen sonuç bir sonraki katmana iletilir ve bu işlem tekrarlanır. Bu sayede daha karmaşık özellikler elde edilir ve verilerin daha doğru bir şekilde sınıflandırılması sağlanır.

Derin öğrenme modelleri çoğunlukla büyük miktarda veri kullanılarak eğitilir. Eğitim süreci boyunca modelin hataları azaltılır ve performansı iyileştirilir. Sonuç olarak model belirli bir görevi yapmak için daha verimli hale getirilir. Derin öğrenme

mimarileri çoğunlukla sınıflandırma, regresyon, kümeleme, doğal dil işleme ve görüntü tanıma gibi problemlerde kullanılır. Bu problemlerde görüntü, ses ve metin gibi farklı veri türlerinde çalışabilmektedir.

Derin öğrenme uygulamaları arasında ses tanıma, yüz tanıma, nesne tanıma, doğal dil işleme gibi uygulamalar vardır. Bu teknoloji her geçen gün yaygınlaşmakta ve gelecekte daha birçok alanda kullanılacaktır.

3.2.1. Derin Öğrenme Mimarileri

Literatürde birçok farklı derin öğrenme mimarisi bulunmaktadır. Bu mimariler Tekrarlayan Sinir Ağları, Uzun Kısa Süreli Bellek Ağları, Geçitli Tekrarlayan Birim ve Evrişimli Sinir Ağları'dır.

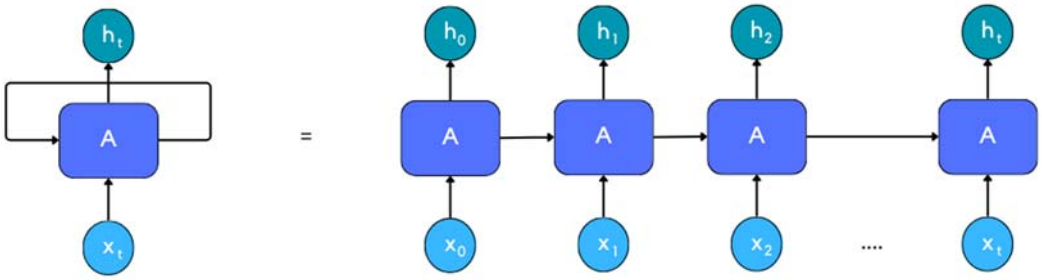
3.2.1.1. Tekrarlayan Sinir Ağları

Tekrarlayan Sinir Ağları (Recurrent Neural Network – RNN), zaman serilerinde veya sıralı verilerde kullanılan bir derin öğrenme mimarisidir. Girdilerin sıralı olarak işlendiği bir yapıya sahiptir. Önceki adımlarda hesaplanan çıktıların ve saklanan bilgilerin bir sonraki adımda tekrar kullanılmasına olarak tanır. RNN modelleri doğal dil işleme, metin üretimi ve ses tanıma gibi uygulamalarda kullanılır [23].

RNN mimarisinde girdi verilerini işleyen gizli bir durum hücresi bulunmaktadır. Gizli durum hücresi önceki adımdaki bilgileri içerir ve sonraki adımda tahmin yapmak için kullanılır. Yeni bir veri geldiğinde her adımda bir girdi alır ve çıktı üretir. Bu çıktı bir sonraki adımda girdi olarak kullanılır ve bu şekilde süreç devam eder. RNN modelleri geriye doğru yayılma ve veri setindeki örneklerin sırasına göre eğitilir. Bu eğitim sırasında model, girdilerin ardışık adımlarında beklenen çıktılarla gerçek çıktıları karşılaştırarak hata oranı azaltmaya çalışır.

RNN mimarisinin bilinen bir sorunu vardır. Bellekteki hafıza hücrelerinin sınırlı kapasitesinden dolayı uzun süreli bağımlılıkları ele almakta zorlanır. Bu sorunun çözümü için de RNN tabanlı farklı mimariler geliştirilmiştir. Bunlar LSTM

(Uzun Kısa Süreli Bellek) ve GRU (Geçitli Tekrarlayan Birim) mimarileridir. Bu mimariler uzun süreli bağımlıkları ele alarak daha doğru sonuçlar verebilir.



Şekil 3.11. RNN Mimarisi

Şekil 3.11'de görüldüğü üzere RNN mimarisi tekrarlayan bir yapıya sahiptir. Bir önceki adımda çıkan sonuç bir sonraki adımda kullanılmaktadır. Örneğin burada h_1 çıktısı için önceki adımda x_0 girdisi ile elde edilen h_0 çıktısı ve x_1 girdisi birlikte girdi olarak kullanılmaktadır.

$$h_t = f(h_{t-1}, x_t) \tag{2}$$

Denklem (2)'de RNN modelinin matematiksel formülü verilmiştir. Formülde h_t değeri şu anki çıktı değerini, h_{t-1} değeri bir önceki çıktı değerini ve x_t değeri ise şu anki girdi değerini göstermektedir.

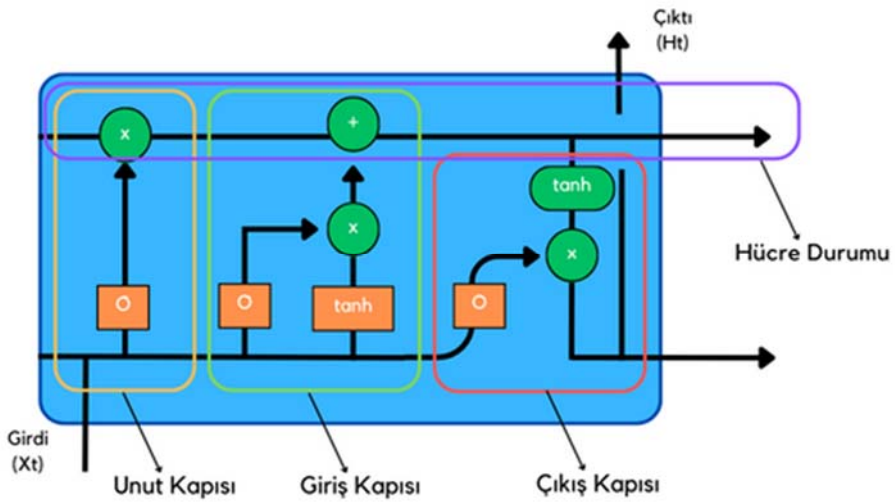
3.2.1.2. Uzun Kısa Süreli Bellek Ağları

Uzun Kısa Süreli Bellek Ağları (Long-Short Term Memory – LSTM) mimarisi, aslında RNN'nin bir sorununa çözüm olarak literatüre girmiş ve RNN'nin alt türü olarak görülmektedir. RNN mimarisindeki "uzun süreli bellek kaybı" sorununu çözmek için gizli durum hücresine ek olarak birçok kapı mekanizması içermektedir. Bu kapılar ağın hangi bilgilerin saklanıp hangilerinin saklanmayacağına karar vermesine yardımcı olur [24].

LSTM mimarisinde üç temel kapı vardır. Bu kapılar unutma (forgot) kapısı, giriş (input) kapısı ve çıkış (output) kapısıdır. Unutma kapısı, hafıza hücresindeki bilginin ne kadarını unutacağını belirler. Bu kapı önceki hafıza hücresindeki bilgi ve şimdiki girdiye dayanarak belirlenir. Giriş kapısı, hafıza hücresine ne kadar yeni bilgi ekleneceğine karar verir. Bu kapı şimdiki girdi ve önceki saklı durum arasındaki

ilişkiye dayanarak belirlenir. Çıkış kapısı ise, hafıza hücresindeki bilgiyi kullanarak ne kadar çıktı üretileceğini belirler. Bu kapı hafıza hücresindeki bilgi ve şimdiki girdiye dayanarak belirlenir.

LSTM modelleri metin sınıflandırma, zaman serileri, ses tanıma, çeviri ve görüntü işleme gibi uygulamalarda kullanılmaktadır. Zaman serileri uygulamasına örnek olarak hisse senetleri, döviz kurları ve faiz oranları gibi finansal verilerin analizi ve gelecek tahminleri yapmak için kullanılabilir. Bu durumda LSTM modeli geçmiş verileri ve mevcut piyasa koşullarını girdi olarak kullanabilir.



Şekil 3.12. LSTM Mimarisi

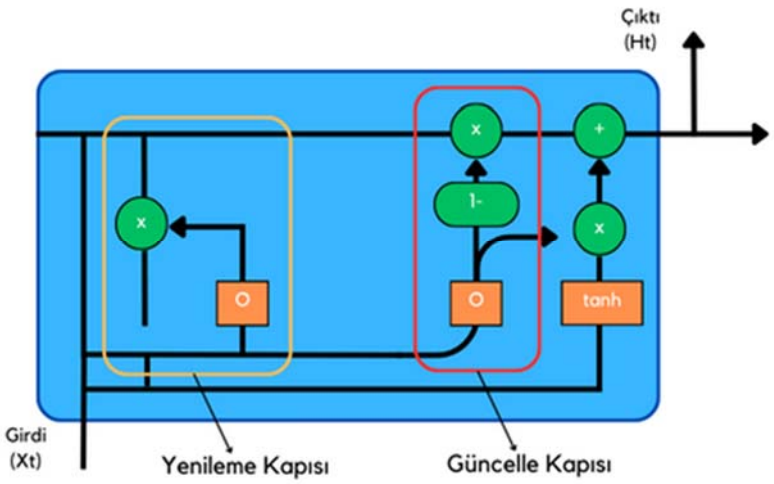
Şekil 3.12'de görüldüğü üzere LSTM mimarisi birbirini takip eden sıralı yapılardan oluşmaktadır. Bu sıralı yapılar unut kapısı, giriş kapısı ve çıkış yapısı olarak tanımlanmaktadır. İlk adım olarak giriş değerleri unut kapısına gelmektedir ve bu kapı hangi verilerin silineceğine hangi verilerin hafızada tutulacağına karar vermektedir. Ardından buradan aktarılan veriler giriş kapısına gitmektedir ve giriş kapısında depolanacak bilgiler belirlenmektedir. Son olarak çıkış kapısında bir çıktı üretilir. Bu çıktı hafızadaki bilgiler ve giriş kapısındaki bilgiler ile belirlenmektedir.

3.2.1.3. Geçitli Tekrarlayan Birim

Geçitli Tekrarlayan Birim (Gated Recurrent Unit – GRU) mimarisi, LSTM ile benzer şekilde RNN mimarisinin alt türüdür. Özellikle doğal dil işleme gibi sıralı verilerin modellenmesi için kullanılır. GRU, LSTM ile benzer bir kapı mekanizması kullanır. Ancak, LSTM'e göre daha az parametreye sahip ve daha hızlıdır.

GRU mimarisinde LSTM mimarisinden farklı olarak 2 temel kapı bulunmaktadır. Bu kapılar yenileme (reset) ve güncelleme (update) kapılarıdır. Yenileme kapısı önceki durumdaki hafıza hücresinin ne kadarının unutulacağını belirler. Bu kapı önceki durumdaki saklı durum ve şimdiki durumdaki girdi arasındaki bir ilişkiye dayanarak belirlenir. Güncelleme kapısı hafıza hücresinin ne kadarının güncelleneceğini belirler. Bu kapı önceki durumdaki saklı durum ve şimdiki durumdaki girdi arasındaki ilişkiye göre belirlenir [24].

GRU modelleri LSTM modelleri ile benzer başarı oranlarına sahip olabilir ve daha az sayıda parametreye sahip olduğu için daha hızlı eğitilebilirler. Bu nedenle daha çok tercih edilebilir. Şekil 3.13'te GRU mimarisinin yenileme ve güncelleme kapıları görülmektedir.



Şekil 3.13. GRU Mimarisi

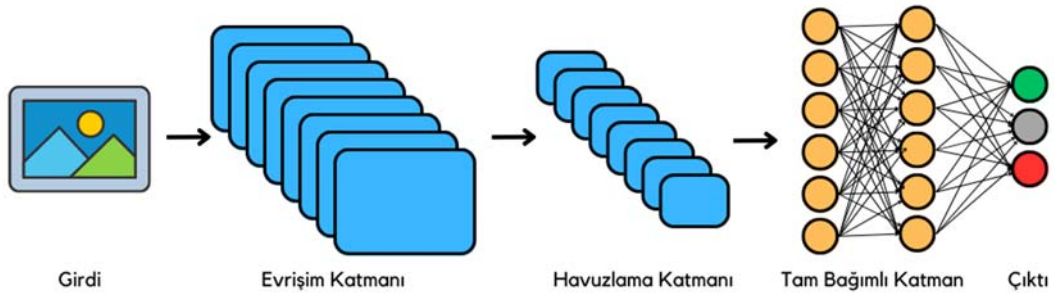
3.2.1.4. Evrişimli Sinir Ağları

Evrişimli Sinir Ağları (Convolutional Neural Network – CNN) mimarisi, özellikle görüntü işleme alanında kullanılan bir derin öğrenme mimarisidir. CNN mimarisi evrişim katmanlarından oluşur ve her katmanda filtreler kullanılarak verilerin özelliklerinin çıkarılmasına dayanır [23].

CNN modelleri girdi olarak alınan görüntülerin özelliklerini çıkartır ve bu işlemi yaparken filtreler kullanır. Örneğin bir görüntü için bu filtreler kenarlar, köşeler veya şekiller olabilir. Ardından bir veya birden fazla tam bağımlı katman kullanarak

bu özellikleri sınıflara ayırır. Bu katmanlar son çıktıyı hesaplayarak modelin tahmin ettiği sınıfın olasılığını hesaplar.

CNN mimarinin avantajı verilerin özelliklerini otomatik öğrenebilmesidir. Bu nedenle verilerin özelliklerinin elle çıkarılmasına gerek yoktur. Ayrıca eğitim verilerindeki küçük değişikliklere dayanıklıdır ve genelleştirme yeteneği yüksektir.



Şekil 3.14. CNN Mimarisi

Şekil 3.14'te görülen CNN mimarisinde girdi olarak bir görüntü verilmektedir. Evrişim katmanında bu girdinin özellikleri çıkartılmaktadır. Havuzlama katmanında ise girdi görüntüsünün boyutu küçültülerek ağırlığı azaltılır. Son olarak tam bağımlı katmanda modelin sinir ağı yöntemiyle öğrenme işlemi gerçekleştirilir. Çıktı olarak da örneğin bu modelimizde olumlu, nötr veya olumsuz çıktısını verir.

3.3. Metin Madenciliği

Metin madenciliği birçok veri madenciliği ve makine öğrenmesi tekniklerini kullanarak, doğal dil işleme teknolojileri ile metin verilerinin analiz edilmesi işlemidir. Metin madenciliği, büyük miktarda yapısal olmayan metin verileriyle çalışarak, bu verilerden anlamlı bilgiler elde etmek için oldukça güçlü bir teknolojidir [25].

Metin madenciliği, sağlık, finans, dijital pazarlama başta olmak üzere birçok alanda kullanılmaktadır. Bu teknoloji kullanıldığı alanlarda belgelerin sınıflandırılmasına, büyük metinlerden özetler çıkarılmasına, metinlerdeki duygu analizini yapmaya fayda sağlamaktadır. Örneğin dijital pazarlama alanında işletmeler müşteri geri bildirimlerini analiz etmek, rakiplerini izlemek ve pazarlama kampanyalarını optimize etmek için metin madenciliği kullanabilirler. Sağlık alanında ise hastane kayıtları ve sağlık verileri gibi büyük miktarda metinsel verileri analiz etmek için metin madenciliği yöntemleri kullanılabilir.

Metin madenciliği süreçleri her uygulama için değişkenlik gösterebilmektedir. Şekil 3.15'te görüldüğü üzere temelde bu süreçler veri toplama, veri ön işleme, modelleme, değerlendirme ve uygulama adımlarından oluşmaktadır.

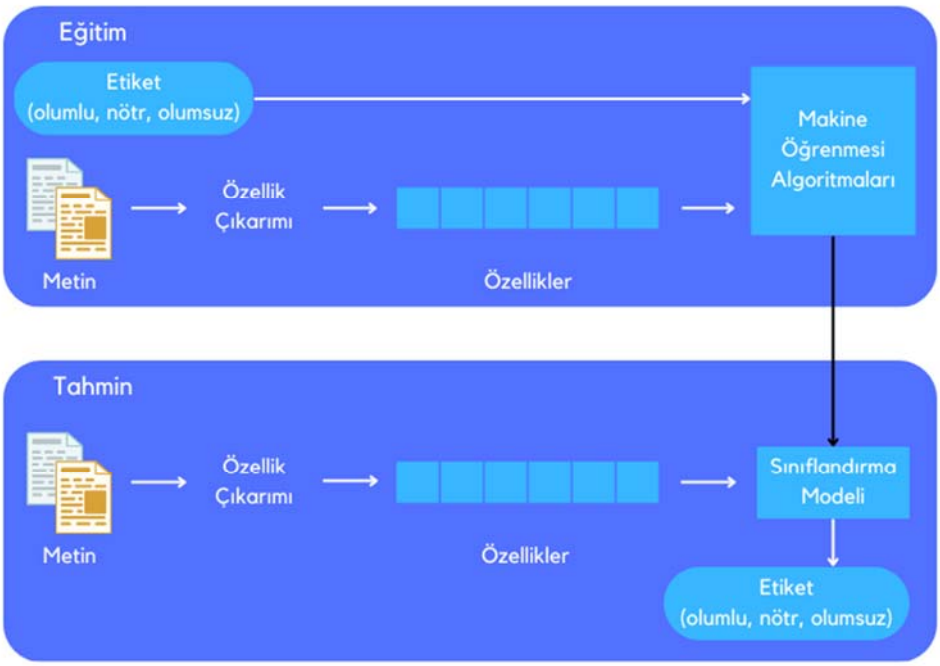
- *Veri Toplama:* Metin verilerinin çeşitli kaynaklar kullanılarak toplandığı adımdır.
- *Veri Ön İşleme:* Toplanan verilerin daha kolay analiz edilebilmesi için çeşitli işlemlerin yapıldığı adımdır. Belge temizleme, ayrıştırma, bazı kelimelerin çıkarılması gibi işlemler yapılabilir.
- *Modelleme:* Makine öğrenimi teknikleri kullanılarak verinin modellenmesinin yapıldığı adımdır. Metin sınıflandırılması, duygu analizi, metin özetleme gibi işlemler için model oluşturulur.
- *Değerlendirme:* Modelin performansının ölçüldüğü adımdır. Doğruluk, F1 skoru gibi performans ölçüm metrikleri kullanılabilir.
- *Uygulama:* Modelin kullanılacağı bir uygulamanın oluşturulduğu adımdır.



Şekil 3.15. Metin Madenciliği Süreci

3.4. Duygu Analizi

Duygu analizi, makine öğrenmesi yöntemleri kullanılarak metinlerdeki duygusal içeriği otomatik tanımlama işlemidir. İnsan doğası gereği birçok duyguya sahiptir ve bu duyguları yazdığı içeriklere, metinlere yansıtmaktadır. Duygu analizi işlemi çoğu duygu için yapılsa da temelde olumlu, olumsuz veya nötr olarak sonuçlar vermesi için kullanılmaktadır. Bir metnin olumlu, olumsuz veya nötr olarak otomatik bir şekilde sınıflandırılması için metindeki kelimeler, cümleler, dil yapıları kullanılmaktadır [26]. Şekil 3.16'da duygu analizi süreci görülmektedir.



Şekil 3.16. Duygu Analizi Süreci

Duygu analizi büyük veri kümeleri içindeki duygu durumlarını tanımlayarak insanların manuel olarak yapabileceği işleri daha hızlı ve etkili bir şekilde gerçekleştirir. Bu nedenle birçok farklı sektörde kullanılmaktadır. Duygu analizi uygulamaları, sosyal medya analizi, müşteri geri bildirimleri, ürün incelemeleri gibi metinsel verilerin oldukça fazla olduğu alanlarda kullanılmaktadır. Özellikle pazarlama ve müşteri hizmetleri alanlarında kullanımı yaygındır.

3.4.1. Duygu Analizi Yöntemleri

Duygu analizi çalışmalarında kural tabanlı yöntemler, makine öğrenmesi yöntemleri, hibrit yöntemler, derin öğrenme yöntemleri gibi çeşitli yöntemler kullanılmaktadır. Bu yöntemler yapılan çalışmada daha doğru sonuçlar elde etmek için farklı avantajlara ve dezavantajlara sahiptir.

3.4.1.1. Kural Tabanlı Yöntemler

Kural tabanlı yöntemler, metinlerin duygu durumunu belirlemek için önceden tanımlanmış kurallar kullanmaktadır. Bu yöntem doğal dil işleme teknikleri kullanılarak ve belirli bir dilde yazılmış metinleri analiz etmek için tasarlanmıştır.

Bu yöntemde metinde belirli kelimeler ve kelime öbekleri aranır, bu kelimelerin olumlu, olumsuz veya nötr olduğuna karar verir. Örneğin; “güzel”, “çok iyi” vs. gibi kelimeleri olumlu olarak sınıflandırılabilirken, “çok kötü”, “çirkin” vs. gibi kelimeleri olumsuz olarak sınıflandırılabilir.

Kural tabanlı yöntemlerde duygu analizi için kullanılan kurallar manuel olarak belirlenir ve bu kurallar belirli bir sektör veya belirli bir metin türü için optimize edilir. Bu yöntemin avantajı kullanıldığı sektörde veya kullanıldığı belirli bir metin türünde doğruluk ve güvenilirlik sağlamasıdır. Dezavantajı ise kullanılması gereken tüm kuralların önceden belirlenmesi gerektiğidir. Bu nedenle bu yöntemler yeni veya beklenmedik bir duygu durumunda yeterince esnek değildir. Ayrıca bu yöntemin doğru çalışabilmesi için de kural setinin metnin türüne göre kaliteli ve doğru belirlenmesi gerekmektedir.

3.4.1.2. Makine Öğrenmesi Yöntemleri

Makine öğrenmesi algoritmalarını ve tekniklerini kullanarak metinlerin duygu durumu belirlenir. Bu yöntem kural tabanlı yöntemlere kıyasla daha esnektir ve yeni veya beklenmedik duygu durumlarının belirlenmesi için daha uygun olabilir.

Makine öğrenmesi, denetimli öğrenme ve denetimsiz öğrenme olmak üzere iki temel sınıflandırma yönteme dayanır. Denetimli öğrenmede önceden etiketlenmiş veriler kullanılarak bir model oluşturulur. Doğru sonuçlar elde edilebilmesi için iyi etiketlenmiş yüksek kalitede eğitim verileri gerektirir. Denetimsiz öğrenmede ise etiketlenmemiş veri kümesi kullanılır ve veriler benzerlik ölçütleri, kümeleme gibi teknikler kullanılarak gruplara ayrılır.

Makine öğrenmesi yöntemleri metnin özelliklerini tanımlamak için birçok farklı teknik kullanır. Bu özellikler kelime frekansları, kelime öbekleri, kelime vektörleri, morfolojik özellikler, semantik özellikler vs. gibi özelliklerdir. Metnin duygu durumunu belirlemek için bu özellikler kullanılmaktadır.

Makine öğrenmesi yöntemleri kural tabanlı yöntemlere göre daha yüksek bir doğruluk oranına sahip olabilirler. Ancak bu yöntemde doğru sonuçlar elde edebilmek için yeterli miktarda yüksek kaliteli eğitim verileri gerektirirler.

3.4.1.3. Hibrit Yöntemler

Hem kural tabanlı hem makine öğrenmesi yöntemlerinin bir arada kullanıldığı duygu analizi yöntemidir. Bu yöntemler farklı duygu analizi yöntemlerinin avantajlarını bir araya getirerek daha iyi sonuçlar elde etmeyi amaçlar.

Bir hibrit yöntemde, kural tabanlı yaklaşım doğru duygu etiketlemesi için kelime öbeklerini belirleyebilir, makine öğrenmesi yaklaşımı ise daha geniş bir bağlamda duygusal anlamı çıkarabilir. Bu sayede bir amaç için kullanılan farklı türdeki veri kümeleri de bir arada analiz edilebilir. Örneğin sosyal medya analizi için paylaşılan mesajlar, görüntüler, videolar ve diğer medya ögeleri farklı veri türleridir. Duygu analizi için bu farklı türdeki veri kümelerini kullanmak istediğimizde hibrit yöntemleri kullanabiliriz.

Hibrit yöntemler farklı duygu analizi tekniklerinin birleştirilmesi sayesinde daha doğru sonuçlar çıkarabilir. Fakat bu yöntemlerin yüksek kaynak ve daha fazla işlem gücü gerektirdiği de göz önünde bulundurulmalıdır.

3.4.1.4. Derin Öğrenme Yöntemleri

Derin öğrenme algoritmalarının, mimarilerinin kullanıldığı duygu analizi yöntemidir. Bu yöntemde büyük veri kümesi ve hesaplama gücü gerektiren bir öğrenme aşaması vardır.

Duygu analizi için kullanılan derin öğrenme yöntemleri genellikle bir veri kümesindeki metinlerin duygusal içeriğini öğrenmek için tasarlanmıştır. Bu yöntemler genellikle sinir ağları kullanarak işlem adımlarını önceden otomatik olarak gerçekleştirirler. Ayrıca metinleri analiz etmek için özel olarak tasarlanmış sinir ağı mimarileri kullanırlar [26].

Duygu analizinde yaygın kullanılan derin öğrenme mimarileri CNN ve RNN mimarileridir. CNN mimarisinde, metin özellikleri belirli boyutlara sahip kernel adı verilen filtrelerle analiz edilir ve sonuç olarak her kelime veya kelime öbeği için bir vektör çıktısı elde edilir. Bu vektörler daha sonra duygu etiketi atamak için kullanılır. RNN mimarisinde de ise, bir metnin duygu analizi için metnin önceki kısımları bellekte tutularak metin adım adım işlenir. Bu yöntem bir kelimenin anlamının bir önceki kelimeyle birlikte değerlendirilerek belirlenmesine olanak sağlar.

Derin öğrenme yöntemleri, büyük veri setlerindeki karmaşık duygu yapılarını anlayabilme yeteneğine sahiptir. Ancak daha fazla işlem gücü gerektirdiğinden eğitim ve uygulama için daha fazla kaynak ihtiyacı olmaktadır.

3.5. Doğal Dil İşleme

Doğal dil, insanlar tarafından günlük hayatta kullanılan dildir. Doğal dil işleme, bir makine öğrenmesi uygulamasıdır ve doğal dilin bilgisayarlar tarafından anlaşılmasını, işlenmesini amaçlamaktadır. Doğal dil işleme oldukça karmaşık ve esnek bir yapıya sahip olan insan diliyle alakalı birçok sorunu çözmeye çalışır [27].

Doğal dil işleme genellikle makine öğrenimi, derin öğrenme ve yapay sinir ağları gibi teknolojilerle birlikte kullanılır. Bu teknolojiler büyük veri kümelerindeki kalıpları tanımlayarak verileri analiz etmek için kullanılır.

Şekil 3.17'de görüldüğü üzere makine öğrenmesi doğal dil işleme için birçok farklı teknik sağlar. Örneğin kelime gömme (word embedding) yöntemi, her kelimenin bir sayısal vektörle temsil edilmesini sağlar. Bu sayısal vektörler doğal dil işleme modellerinde kullanılabilir ve kelimenin anlamını temsil eder. Ayrıca sınıflandırma ve regresyon gibi makine öğrenmesi teknikleri de doğal dil işlemede sıklıkla kullanılır. Örneğin bir metindeki duygu analizi için sınıflandırma yöntemi kullanılabilir. Bu yöntem belirli bir metnin olumlu, olumsuz ve nötr olarak sınıflandırılmasına olanak sağlar.



Şekil 3.17. Doğal Dil İşleme

Doğal dil işleme, çeviri, metin analizi, metin özet çıkarma, konuşma tanıma gibi birçok alanda kullanılmaktadır. Örneğin makine çevirisi, bir dilde yazılmış olan metnin başka bir dile otomatik çevrilmesini sağlar. Bu uygulama özellikle uluslararası iş yapan şirket ve kuruluşlar için büyük önem taşımaktadır.

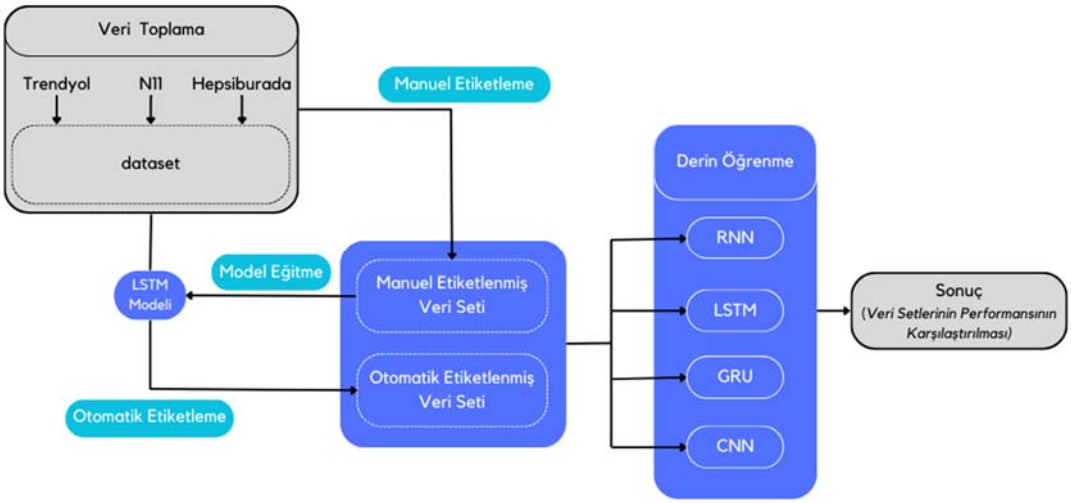
Doğal dil işleme uygulamaları günümüzde oldukça kullanılmaktadır ve gittikçe de yaygınlaşmaktadır. Örneğin müşteri geri bildirimlerinin analizinde, sosyal medya takibinde doğal dil işleme uygulamaları kullanılabilmektedir. Bunların dışında özellikle haberlerin, makalelerin ve diğer tüm yazılı materyallerin otomatik olarak sınıflandırılması veya bu metinlerden özetler çıkarılması alanlarında doğal dil işleme kullanılabilmektedir.

Sonuç olarak doğal dil işleme uygulamaları, insan dilinin anlaşılması ve işlenmesiyle ilgili birçok sorunu çözmek için tasarlanmış bir teknolojidir. Bu teknolojinin kullanımı gittikçe yaygınlaşmaktadır.

4. MATERYAL VE YÖNTEMLER

Çalışma kapsamında manuel ve otomatik etiketlenmiş olmak üzere iki veri seti hazırlanmıştır. Bu bölümde öncelikle çalışmada kullanılan veri setlerinin nasıl hazırlandığı açıklanmaktadır. Sonrasında ise bu veri setleri kullanılarak yapılan deneysel çalışmalar ele alınmaktadır.

Şekil 4.1'deki akış diyagramında görüldüğü üzere öncelikli olarak veri toplama işlemi yapılmıştır. Toplanan veriler manuel etiketleme işlemi ile etiketlenmiştir. Bu işlemin ardından oluşan manuel etiketlenmiş veri seti kullanılarak model eğitme işlemi yapılmış ve bir LSTM modeli oluşturulmuştur. Sonrasında tekrar bir veri toplama işlemi yapılmış ve bu toplanan veriler LSTM modeli sayesinde otomatik olarak etiketlenmiştir. Bu sayede manuel ve otomatik etiketlenmiş iki veri seti elde edilmiştir. Son olarak derin öğrenme algoritmaları kullanılarak veri setlerinin performansının karşılaştırılması yapılmıştır.



Şekil 4.1. Çalışma Akış Diyagramı

4.1. Manuel Veri Setinin Hazırlanması

Çalışmalarda veri seti önemli bir rol oynamaktadır. İyi bir veri seti çalışmanın sonucuna direk olarak etki etmektedir. Bazı çalışmalarda hazır veri setleri kullanılırken bazı çalışmalarda ise özgün veri setleri hazırlanmaktadır.

Bu çalışmada e-ticaret ürün yorumlarını içeren bir veri seti hazırlanmıştır. Öncelikle ülkemizde faaliyet gösteren popüler e-ticaret platformlarından ürün yorumları toplanmıştır. Veri ön işleme işlemlerinin ardından ürün yorumları olumlu, olumsuz ve nötr olarak etiketlenmiştir. Bu etiketlemenin ardından nötr ürün yorumları çıkarılarak olumlu ve olumsuz olarak iki sınıflı bir veri seti oluşturulmuştur.

4.1.1. Veri Toplama

Veri toplama, veri seti hazırlama sürecinin ilk adımıdır. Çalışmada kullanılacak veriler belirlendikten sonra ilk olarak veri toplama işlemi yapılmaktadır.

Bu çalışmada kullanılacak veri seti için ülkemizde faaliyet gösteren Hepsiburada [37], Trendyol [38], N11 [39] gibi popüler e-ticaret platformlarından ürün yorumları toplanmıştır. Öncelikle manuel etiketlemek için 15 bin 170 adet yorum toplanarak bir veri seti oluşturulmuştur. Bu veri seti için ağırlıklı olarak tıraş makinesi, çalışma koltuğu ve oyun konsolu kategorileri seçilmiştir.

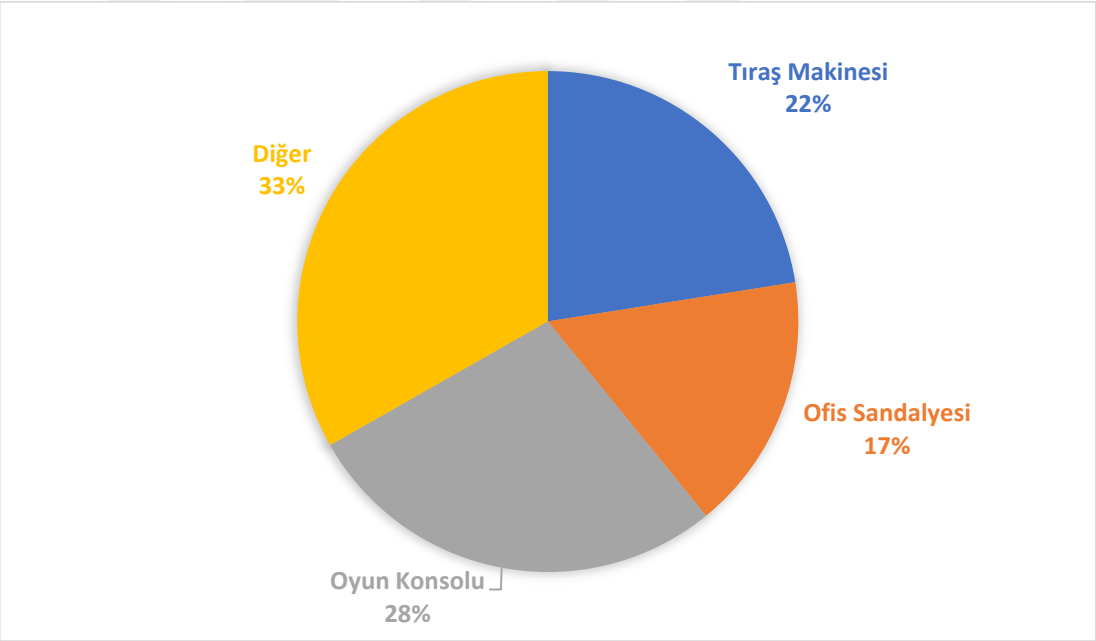
E-ticaret platformlarından ürün yorumlarını toplamak için web tabanlı bir uygulama geliştirilmiştir. Uygulamanın geliştirilmesinde C# programlama dili, .Net framework'ü ve MVC (Model – View – Controller) mimarisi kullanılmıştır. Veri toplama modülünde Selenium kütüphanesi ile Web Scraping yöntemi kullanılmıştır. Her e-ticaret platformu için farklı yöntemler geliştirilmiştir. Uygulamanın arayüzü ile ilgili detaylı bilgiler 5. Bölümde yer almaktadır.

Veri toplama sırasında hiçbir kişisel veri ile ürünlerin marka ve satıcı gibi bilgileri toplanmamıştır. Sadece ürünler üzerindeki yorum metni toplanmıştır. Tablo 4.1'de görülen örnek ürün yorumları aynen toplandığı gibi hiç dokunulmadan ele alınmıştır. Şekil 4.2'de ise ürün yorumlarının ürün kategorilerine göre dağılımına yer verilmiştir.

Tablo 4.1. Örnek Veriler

Siparişim çok özensiz bir şekilde paketlenmiş olup, satıcı firma ne kargo ile ne de mail üzerinden fatura gönderimi yapmadı. Defalarca fatura talebi iletmeme rağmen, mail üzerinden bile fatura gönderilmedi. Hakkınızda N11 üzerinden şikayet talebinde bulunacağım.
Çok güzel paketleme ile hızlı bir şekilde gönderdi mağaza teşekkürler daha kullanma fırsatı olmadı ama yorumları çok iyi diye aldım zaten...
1 tane aldım beklentimin üstünde çıktı 1 tane daha sipariş verdim.
Kesinlikle çok başarılı eşime aldım çok beğendi.tasarım olarak da çok şık
Maalesef gereksiz övülen bir ürün. Sakallarımı kısaltırken çok olmasada kıllarımı çekti. Ayrıca yukarıdan aşağı doğru traş etmedi. Performansı vasat, bence bir F/P ürünü değil. Bu arada bu ürünle saç kesimi yapmayı düşününenler için ; Yapar ama sıkıntı çekersiniz. Hem uzun sürer hem de kısa ve yumuşak kıllları almak ta zorlanırsınız.
Ürünü bugün kargodan teslim aldım. Kontrol etme fırsatım oldu . Şarjz konusunda iyi değil malesef . Çabuk bitiyor . Uç kısımlarına takılan aparatların sayısı az . Açıkçası aldığıma pişman oldum
güzel ve şık bir ürün beklentilerimi karşılar
hayir kotuydu tavsiye atmem kargo 11 gunde geldi
Evet ürünü beğendim sıfır kesim İcin almıştım işimi görecek umarım uzun ömürlü olur
Cihazı sıfır sakal kesmek için aldım bekletimin çok üstünde. Berberlerdeki pahalı makineler gibi direk kesiyor ve çekme vs. yapmıyor. Ses ise eski makinemin yarısı kadar. Bakalım dayanıklı olacak mı göreceğiz. Mağaza sorulara yanıt veriyor.
Ense ve kirli sakal için oldukça iyi bir makine. Bu amaçlar için kullanılır görüşündeyim.
Ürün sakalı çok fazla çekiyor hatalı olduğunu düşünüyorum
Ürünü şarjdan çıkarınca çalışmıyo o yüzden şarjdayken oynadık bu seferde ilk günden bataryası şişti atmak zorunda kaldık ve bazı oyunlar açılmıyo tavsiye etmiyorum
Ürün çok basit ama eğlenceli
Tuşlar çok iyi basmıyor beklediğim gibi gelmedi.

ÜRÜN ALDIK BUGÜN MONTAJINI YAPTIK,SABİT DURUYOR İNİP KALKMIYOR AMORTİSÖR ÇALIŞMİYOR
nasil bi sandaliye bu...!! sanki dag gibi bi sandaliye, hergun tas ustunde oturuyorum, sirtima hergun mesaj yapiyorum su kalitesiz sandaliye uzunden
kızım için aldık ve fiyatına göre süper
15 gunde anca elime ulasti.deri kisminin alti yırtikti.oturmasi kismin yanlardan tam oturtulmamis bosluk var resmen.cok bekledigimiz icin cocuk üzülmesin diye iade etmedim.ama pisman oldum.tavsiye etmem.
Ürün geldi üst kısmını alt kısma bağlayacak parça yok kızım denemek için bir oturdu yere düştü
fiyatına göre gayet güzel kargo iyi
ürün gayet güzel ve aynen görüntü deki gibi



Şekil 4.2. Manuel Veri Seti Ürün Yorumlarının Kategori Dağılımı

4.1.2. Veri Ön İşleme

Makine öğrenmesinde veri ön işleme adımları, analiz ve modelleme çalışmalarında en uygun hale getirmek için yapılan işlemlerdir. Veri ön işleme süreci, çalışmanın etkili ve doğru sonuçlar üretebilmesi için oldukça önemlidir. Veri ön işleme sürecinde veri temizleme, veri normalizasyonu, özellik seçimi gibi işlemler yapılmaktadır.

Bu çalışma kapsamında toplanan veriler veri temizleme adımından geçirilmiştir. Bu kapsamda sadece noktalama işareti veya emoji içeren veriler çıkarılmıştır. Metinlerin arasında yer alan noktalama işaretleri ve emojiler de kaldırılarak veri setinin sadece metinlerden oluşması sağlanmıştır. Metinlerdeki yazım yanlışları da olabildiğinde düzeltilmiştir. Ayrıca alakasız bir yorum içeren metinler de veri setinden çıkarılarak veri temizleme adımı tamamlanmıştır.

Tüm harflerin küçük harfe çevrilmesi, sayıların kaldırılması ve StopWords olarak sınıflandırılan kelimelerin de metinlerin içerisinden çıkarılması gibi standart veri ön işleme adımları da veri seti üzerinde model eğitiminden önce uygulanmıştır.

4.1.3. Verilerin Etiketlenmesi

Makine öğrenmesinde denetimli öğrenme yöntemleri için verinin etiketlenmesi gerekmektedir. Çalışmada kullanılacak sınıflandırma tipine göre toplanan verilerin etiketlenmesi gerekmektedir.

Ürün yorumlarının sınıflandırılması için genelde olumlu, olumsuz ve nötr sınıflandırılmaları kullanılmaktadır. Bu etiketleme için genellikle ürün yorumlarının değerlendirme puanı kullanıldığı görülmüştür. Örneğin incelenen çalışmalarda 4 ve 5 yıldız puanına sahip yorumlar olumlu, 1 ve 2 yıldız puanına sahip yorumlar olumsuz ve 3 yıldız puanına sahip yorumlar ise nötr olarak etiketlendiği görülmüştür. Ancak bu etiketleme yöntemi birtakım sorunları ortaya çıkarmaktadır. Kullanıcıların verdiği yıldız puanı ile yazdığı yorum metni bazı durumlarda çok zıt olabilmektedir. Örneğin kullanıcı verdiği yorumun yukarılarda görülmesi için 5 yıldız verdiği fakat yazdığı yorumun tamamen olumsuz olduğu tespit edilmiştir. Bu çalışmada toplanan veriler tek tek okunarak bir etiketleme işlemi yapılmıştır.

E-ticaret platformlarından toplanan yorumlar tek tek okunarak olumlu, olumsuz ve nötr olarak etiketlenmiştir. Ardından farklı bir zamanda önceki etiketler gizlenerek veri seti tekrar etiketlenmiştir. Bu sayede iki farklı etiketleme sonucu ortaya çıkmıştır. Bu sonuçlar karşılaştırılarak etiketi farklı olan yorumlar veri setinden çıkarılmıştır. Örneğin öncekinde olumlu dediğimiz bir yorumu sonraki etiketleme işleminde nötr olarak etiketlediysek bu eğiteceğimiz modeli etkileyeceği için bu

yorum veri setinden çıkarılmıştır. Bu işlemin sonucunda olumlu yorumların olumsuz yorumlardan çok fazla olduğu tespit edilmiş ve e-ticaret platformlarından olumsuz yorumlar çekilmeye çalışılmıştır. Geliştirdiğimiz uygulamanın veri çekme modülünde yıldız puanı filtresi kullanılarak kolaylıkla olumsuz yorumlar toplanmıştır. Sonrasında tekrar çekilen veriler tek tek okunarak olumlu yorumlar çıkarılmış ve sadece olumsuz yorumlar veri setine eklenmiştir. Etiketleme işleminin ardından 6818 adet olumlu, 6926 adet olumsuz ve 1393 adet nötr yorumun olduğu bir seti hazırlanmıştır. Nötr yorumların az olması ve tespit edilmesinin de zor olması nedeniyle nötr yorumlar veri setinden çıkarılmıştır. Son olarak dengeli ve iki sınıflı bir veri seti çalışmaya hazır hale getirilmiştir. Tablo 4.2'de etiketleme sonucu elde edilen örnek yorumlar görülmektedir. Tablo 4.3'te ise yorumların etiket dağılımına yer verilmiştir.

Tablo 4.2. Etiketleme Sonucu Örnek Veriler

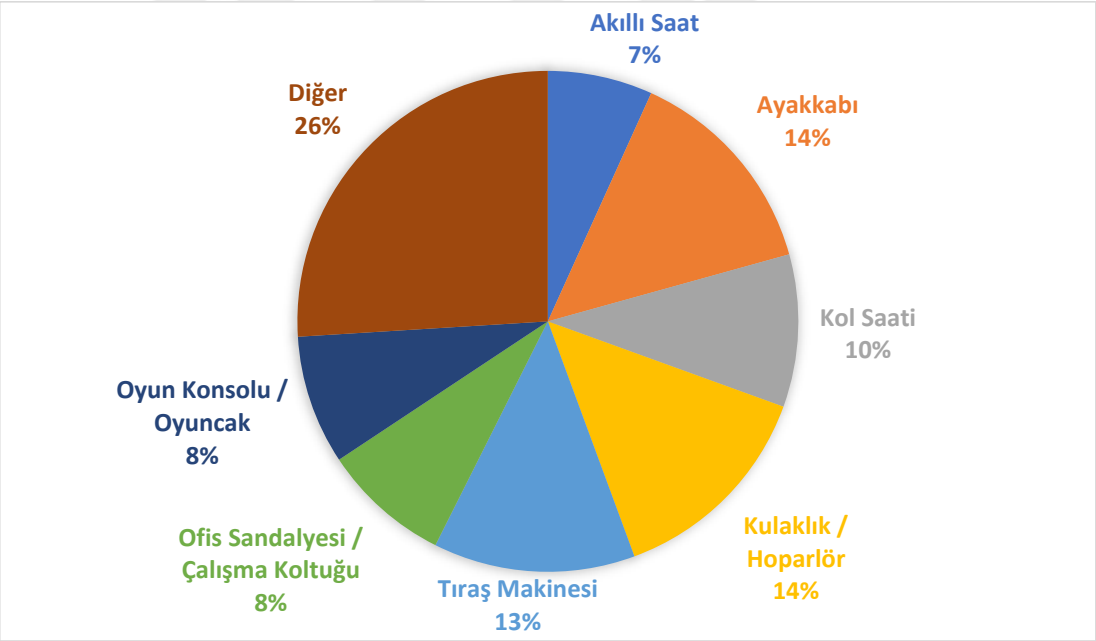
Yorum Metni	**Etiket Sınıfı**
ürün çok şık fakat içi boş bir ürün. biraz daha bütçe zorlayıp iyi bir makine almanızı tavsiye ederim. sıfırı çok kötü	0
emsallerine göre fiyatı düşük olunca bi çekinceyle aldım ama gerçekten kaliteli ve kullanışlı. 5 yıldızı sonuna kdar hakediyor. mağaza hızı da cok iyi	1
Ne yalan söyleyeyim, bu kadar iyi bekle miyordum.	1
üründeki 3mm tarağı bildiğimiz 3 numaralı traş diye aldım. meğer 0 ile 1 arasında bişeymiş. bir defa neme kesimi yaptığım için iade yada değişim olmuyormuş. boşa almış oldum çöpe para attık.	0
urunu kullanmadim ama elime ulasti cok guzel gozukuyor . urun Orta boydadir.	1
Ürün zaten aşırı sağlam kaliteli bir şey değil, oyunlar güzel tatmin edici olsa da ilk şarj edişimde çok dikkatli olmama rağmen usb girişi kırıldı.	0
kutu kırılmış ve iki tarafı açık durumda geldi içerisinde fatura ve gameboy vardı fakat pil ve kablolar yoktu üstelik kargo elime yaklaşık 1 hafta içerisinde ulaştı iade edeceğim.	0
Beğenmedim çok kötü tuşlara iyice basmadan çalışmıyo parmaklarım ağrıdı oynarken kalitesi kötü parasını haketmiyo gereksizmiş pişman oldum aldığım için :(	0
Kargosu hızlıydı güzel geldi eğlenceli bir oyun	1

Tablo 4.3. Sınıflardaki Yorum Sayısı

Sınıf	Yorum Sayısı
Olumlu	6818
Olumsuz	6926
Toplam	13744

4.2. Otomatik Veri Setinin Hazırlanması

Çalışmada kullanılan bir diğer veri seti olan otomatik veri seti için e-ticaret platformlarından 203 bin 274 adet ürün yorumu toplanmıştır. Bu veri setinde bir önceki veri setinde kullanılan ürün kategorileri dışında ayakkabı, kol saati, kulaklık, oyuncak, akıllı saat ve hoparlör gibi birçok ürün kategorisi kullanılmıştır. Ürün yorumlarının kategori dağılımı Şekil 4.3'te görülmektedir.



Şekil 4.3. Otomatik Veri Seti Ürün Yorumlarının Kategori Dağılımı

Toplanan verilerin otomatik etiketlenmesi için LSTM algoritmasının kullanıldığı bir model geliştirilmiştir. Bu modelin eğitilmesinde manuel etiketlenen veri seti kullanılmıştır.

LSTM modelinde kullanılan parametreler Tablo 4.4'te yer almaktadır. Bu parametrelerden Embedding Dimension Size (Gömme Boyutu), kelime gömme

yönteminde her kelimenin temsil edildiği vektörün boyutunu belirler ve örneğin 200 seçildiğinde her kelime için 200 boyutlu vektör kullanılır. Embedding Padding Size (Gömme Dolgu Boyutu) parametresi, metin verilerinin gömme işlemi öncesinde veya sonrasında yapılacak olan dolgu işlemidir ve örneğin 100 seçildiğinde metin verileri 100 kelimeye tamamlanır. Neuron Size (Nöron Boyutu) parametresi, LSTM modellerinde hücrelerin boyutunu temsil eder ve daha büyük bir nöron boyutu daha fazla bilgiyi depolayabilme ve karmaşık ilişkileri daha iyi öğrenebilme yeteneği sağlar. Dropout Rate (Bırakma Oranı) parametresi, modelin aşırı öğrenmesini azaltmak için kullanılan bir parametredir. Number of Epochs (Dönem Sayısı) parametresi, modelin veri setini kaç kez tamamladığını belirler ve yeterli değer verildiğinde daha iyi sonuçlar elde edilmesi sağlanır. Batch Size (Grup Boyutu) parametresi, verilerin kaç alt gruba ayrılacağını belirler ve bu alt gruplar sırasıyla modele sunularak eğitim süreci tamamlanır. Optimizer (Optimize Edici) parametresi, modelin ağırlıklarını güncellemek ve kayıp fonksiyonunu minimize etmek için kullanılır ve Adamax algoritması büyük veri setleri ve büyük parametre uzayları için etkili bir algoritmadır. Activation Function (Aktivasyon Fonksiyonu) parametresi ise sinir ağlarının katmanlarında kullanılan non-lineer dönüşüm işlevini temsil eder ve her bir sinir hücresi aktivasyon fonksiyonu tarafından çıktıya dönüştürülür. Relu aktivasyon fonksiyonu negatif girişleri 0'a dönüştürür, pozitif girişleri olduğu gibi bırakır. Aktivasyon fonksiyonu modelin başarısını olumlu yönde etkilemektedir. Bu parametrelerin en başarılı değerleri deneme yanılma yoluyla seçilmiştir.

Tablo 4.4. LSTM Modelinde Kullanılan Parametreler

Parametre Adı	**Parametre Değeri**
Gömme Boyutu (Embedding Dimension Size)	200
Gömme Dolgu Boyutu (Embedding Padding Size)	100
Nöron Boyutu (Neuron Size)	100
Bırakma Oranı (Dropout Rate)	0.4
Dönem Sayısı (Number of Epochs)	10
Grup Boyutu (Batch Size)	16
Optimize Edici (Optimizer)	Adamax
Aktivasyon Fonksiyonu (Activation Function)	Relu

4.3. Deneyler

Bu bölümde ürün yorumlarından hazırlanan veri setleri ile olumlu ve olumsuz olarak iki sınıflı duygu analizi deneyleri yapılmıştır. Deneylerin ilk aşamasında manuel ve otomatik etiketlenmiş iki veri seti kullanılarak derin öğrenme algoritmaları tarafından elde edilen sonuçlar karşılaştırılmıştır. Sonraki aşamada otomatik etiketlenmiş veri seti ile geleneksel makine öğrenmesi algoritmaları ile sonuçlar alınıp derin öğrenme algoritmalarıyla alınan sonuçlar karşılaştırılmıştır.

Çeşitli sınıflandırma algoritmalarının performansını ölçmek için deneylerde Keras ve Python kullanılmıştır. 10 katlı bir çapraz doğrulama yaklaşımı ile verinin 10 eşit parçaya bölünmesi ve modelin 9 parça ile eğitilip geri kalan parça ile test edilmesi ile sonuçlar alınmıştır. Ayrıca bu işlem testte kullanılacak her parça için 10 kez tekrarlanmaktadır.

Duygu analizi sonuçları doğruluk (accuracy), geri çağırma (recall), kesinlik (precision) ve F1-Puanı (F1-Score) performans ölçütleri ile değerlendirilmiştir. Doğruluk değeri, algoritmanın doğru sınıflandırdığı örnek sayısının toplam örnek sayısına oranıdır. Geri çağırma değeri, algoritmanın pozitif örnekleri sınıflandırma başarısını ölçmektedir. Kesinlik değeri, algoritmanın pozitif olarak sınıflandırdığı örneklerin gerçekten pozitif olma olasılığını ölçmektedir. F1-Puanı değeri, geri çağırma ve kesinlik değerlerinin harmonik ortalaması olarak görülmektedir. F1-Puanı değerinin yüksek olabilmesi için her iki değerin de yüksek olması gerekmektedir [40]. Tablo 4.5'te bu ölçütlerde kullanılan çapraz tahmin değerleri gösterilmektedir. Bu değerlerden Gerçek Pozitif (True Positive – TP) değeri, gerçek değerin ve tahmin edilen değerin pozitif (1) olduğu durumdur. Gerçek Negatif (True Negative – TN) değeri, gerçek değerin ve tahmin edilen değerin negatif (0) olduğu durumdur. Yanlış Pozitif (False Positive – FP) değeri, gerçek değerin negatif (0), tahmin edilen değerin pozitif (1) olduğu durumdur. Yanlış Negatif (False Negative – FN) değeri ise, gerçek değerin pozitif (1), tahmin edilen değerin negatif (0) olduğu durumdur [41].

Tablo 4.5. Çapraz Tahmin

		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahmini Sonuçlar	Pozitif (1)	TP	FP
	Negatif (0)	FN	TN

Denklem (3), Denklem (4), Denklem (5) ve Denklem (6)'da sırasıyla doğruluk, geri çağırma, kesinlik ve F1-Puanı ölçütlerinin hesaplaması gösterilmektedir [42].

$$Doğruluk\ (Acc)\ = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Geri\ Çağırma\ (R) = \frac{TP}{TP\ + FN} \quad (4)$$

$$Kesinlik\ (P) = \frac{TP}{TP\ + FP} \quad (5)$$

$$F1 - Puanı\ = \frac{2\ x\ R\ x\ P}{R + P} \quad (6)$$

Yapılan deneylerin temel amacı manuel ve otomatik olarak etiketlenen veri setlerini kullanarak çeşitli sınıflandırma algoritmalarının performanslarını karşılaştırmaktır. Bu karşılaştırma sonucunda da otomatik etiketleme modelimizin gerçek hayatta kullanılabilirliğini göstermektir.

Derin öğrenme algoritmaları ile yapılan deneylerde Gömme Boyutu (Embedding Size) ve Katman Birimi (Layer Unit) parametreleri kullanılmıştır. Gömme Boyutu parametresi, kelime gömmeleri gibi dizi veya metin verilerini temsil etmek için kullanılmaktadır. Örneğin bir kelimenin 100 boyutlu bir gömme vektörüyle

temsil edilmesi durumunda bu parametrenin değeri 100 olacaktır. Katman Birimi parametresi, derin öğrenme modellerindeki katmanların birim sayısını temsil etmektedir. Derin öğrenme modellerinde her bir katmanın kendine ait birimleri vardır. Deneylerde en uygun parametre değerini bulmak genellikle deneme yanılma yoluyla olmaktadır. Biz deneylerimizde her iki parametre için 100, 200, 300 değerleri kullanılarak kombinasyonlar ile en yüksek sonuçları almaya çalıştık.

Tablo 4.6, Tablo 4.7, Tablo 4.8 ve Tablo 4.9 manuel etiketlenmiş veri seti kullanılarak derin öğrenme algoritmaları ile elde edilen sırasıyla doğruluk, geri çağırma, kesinlik ve F1-Puanı sonuçlarını göstermektedir. Doğruluk sonuçları incelendiğinde 0,96 ile en yüksek sonuca ulaşan algoritmanın GRU algoritması olduğu görülmektedir. F1-Puanı sonuçlarına bakıldığında da 0,9598 ile yine GRU algoritmasının en yüksek sonucu verdiği görülmektedir.

Tablo 4.6. Manuel Veri Seti Kullanılarak Derin Öğrenme Algoritmaları Tarafından Elde Edilen Doğruluk Sonuçları

Gömme Boyutu	Katman Birimi	RNN	LSTM	GRU	CNN
100	100	0,9519	0,9566	0,9571	0,9574
200	100	0,9510	0,9557	0,9581	0,9549
300	100	**0,9525**	**0,9572**	0,9563	0,9562
100	200	0,9488	0,9533	0,9521	0,9544
200	200	0,9498	0,9542	0,9590	0,9573
300	200	0,9473	0,9550	0,9598	**0,9578**
100	300	0,9486	0,9556	0,9573	0,9559
200	300	0,9490	0,9564	**0,9600**	0,9559
300	300	0,9521	0,9550	0,9588	0,9574

Tablo 4.7. Manuel Veri Seti Kullanılarak Derin Öğrenme Algoritmaları Tarafından Elde Edilen Geri Çağırma Sonuçları

Gömme Boyutu	Katman Birimi	RNN	LSTM	GRU	CNN
100	100	0,9538	0,9561	0,9583	**0,9636**
200	100	0,9542	0,9545	0,9600	0,9578
300	100	0,9569	**0,9622**	0,9573	0,9567
100	200	0,9375	0,9545	0,9564	0,9563
200	200	0,9539	0,9595	**0,9655**	0,9603
300	200	**0,9601**	0,9570	0,9633	0,9603
100	300	0,9517	0,9586	0,9572	0,9581
200	300	0,9547	0,9594	0,9629	0,9597
300	300	0,9550	0,9620	0,9641	0,9623

Tablo 4.8. Manuel Veri Seti Kullanılarak Derin Öğrenme Algoritmaları Tarafından Elde Edilen Kesinlik Sonuçları

Gömme Boyutu	Katman Birimi	RNN	LSTM	GRU	CNN
100	100	0,9501	**0,9566**	0,9556	0,9512
200	100	0,9479	0,9565	0,9560	0,9517
300	100	0,9484	0,9524	0,9547	**0,9553**
100	200	**0,9590**	0,9517	0,9475	0,9522
200	200	0,9458	0,9490	0,9526	0,9541
300	200	0,9359	0,9527	0,9560	0,9550
100	300	0,9456	0,9525	**0,9568**	0,9534
200	300	0,9439	0,9532	**0,9568**	0,9519
300	300	0,9492	0,9483	0,9536	0,9524

Tablo 4.9. Manuel Veri Seti Kullanılarak Derin Öğrenme Algoritmaları Tarafından Elde Edilen F1-Puanı Sonuçları

Gömme Boyutu	Katman Birimi	RNN	LSTM	GRU	CNN
100	100	0,9516	0,9563	0,9568	0,9573
200	100	0,9508	0,9553	0,9578	0,9547
300	100	**0,9524**	**0,9571**	0,9560	0,9559
100	200	0,9478	0,9530	0,9519	0,9541
200	200	0,9496	0,9541	0,9589	0,9571
300	200	0,9476	0,9547	0,9596	**0,9576**
100	300	0,9484	0,9554	0,9569	0,9557
200	300	0,9490	0,9562	**0,9598**	0,9557
300	300	0,9519	0,9550	0,9587	0,9573

Tablo 4.10, Tablo 4.11, Tablo 4.12 ve Tablo 4.13 ise otomatik etiketlenmiş veri seti kullanılarak derin öğrenme algoritmaları ile elde edilen sırasıyla doğruluk, geri çağırma, kesinlik ve F1-Puanı sonuçlarını göstermektedir. **Bu sonuçlarda doğruluk sonuçları incelendiğinde 0,9555 ile en yüksek sonuca ulaşan algoritmanın CNN olduğu görülmektedir.** F1-Puanı sonuçlarına bakıldığında da 0,9505 ile yine CNN algoritmasının en yüksek sonucu verdiği görülmektedir.

Tablo 4.10. Otomatik Veri Seti Kullanılarak Derin Öğrenme Algoritmaları Tarafından Elde Edilen Doğruluk Sonuçları

Gömme Boyutu	Katman Birimi	RNN	LSTM	GRU	CNN
100	100	0,9421	0,9528	0,9509	0,9539
200	100	0,9451	**0,9531**	0,9515	0,9521
300	100	**0,9460**	0,9521	0,9514	0,9538
100	200	0,9447	0,9499	0,9520	0,9552
200	200	0,9452	0,9509	0,9517	0,9550
300	200	0,9411	0,9513	0,9513	**0,9555**
100	300	0,9373	0,9513	**0,9525**	0,9538
200	300	0,9390	0,9521	0,9524	0,9554
300	300	0,9367	0,9507	0,9507	0,9552

Tablo 4.11. Otomatik Veri Seti Kullanılarak Derin Öğrenme Algoritmaları Tarafından Elde Edilen Geri Çağırma Sonuçları

Gömme Boyutu	Katman Birimi	RNN	LSTM	GRU	CNN
100	100	0,9410	0,9392	**0,9502**	0,9469
200	100	0,9415	0,9459	0,9428	0,9347
300	100	0,9350	**0,9464**	0,9488	0,9495
100	200	**0,9451**	0,9346	0,9370	**0,9527**
200	200	0,9383	0,9405	0,9372	0,9431
300	200	0,9291	0,9412	0,9414	0,9481
100	300	0,9271	0,9402	0,9457	0,9467
200	300	0,9274	0,9446	0,9491	0,9492
300	300	0,9112	0,9439	0,9464	0,9493

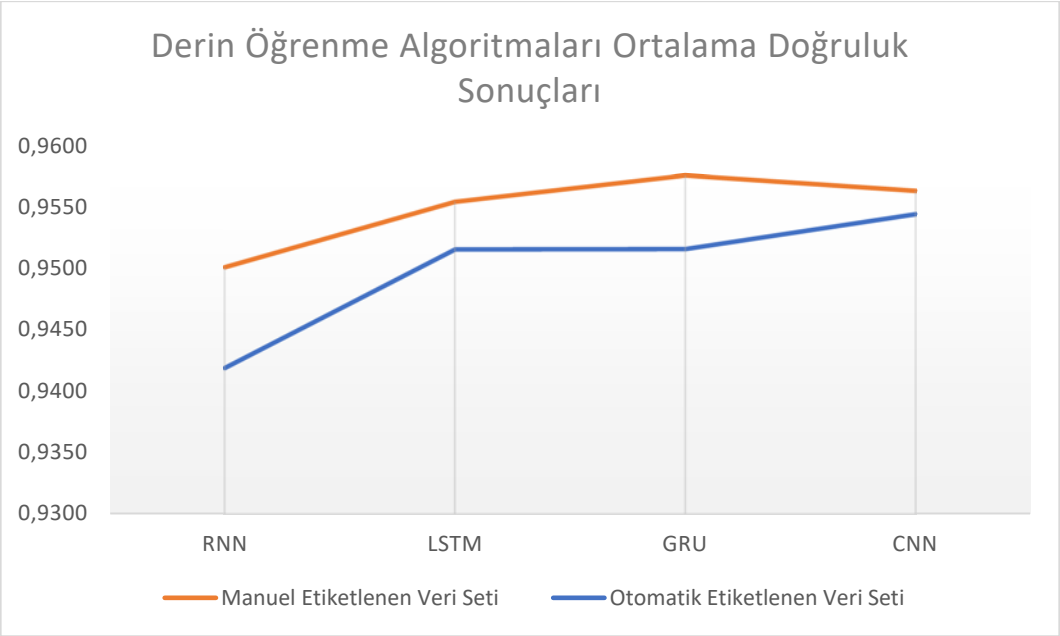
Tablo 4.12. Otomatik Veri Seti Kullanılarak Derin Öğrenme Algoritmaları Tarafından Elde Edilen Kesinlik Sonuçları

Gömme Boyutu	Katman Birimi	RNN	LSTM	GRU	CNN
100	100	0,9318	**0,9554**	0,9418	0,9507
200	100	0,9378	0,9498	0,9495	**0,9584**
300	100	0,9449	0,9476	0,9439	0,9483
100	200	0,9335	0,9535	**0,9560**	0,9483
200	200	0,9406	0,9502	0,9549	0,9568
300	200	0,9404	0,9506	0,9504	0,9531
100	300	0,9338	0,9512	0,9488	0,9511
200	300	0,9377	0,9491	0,9456	0,9518
300	300	**0,9477**	0,9471	0,9448	0,9514

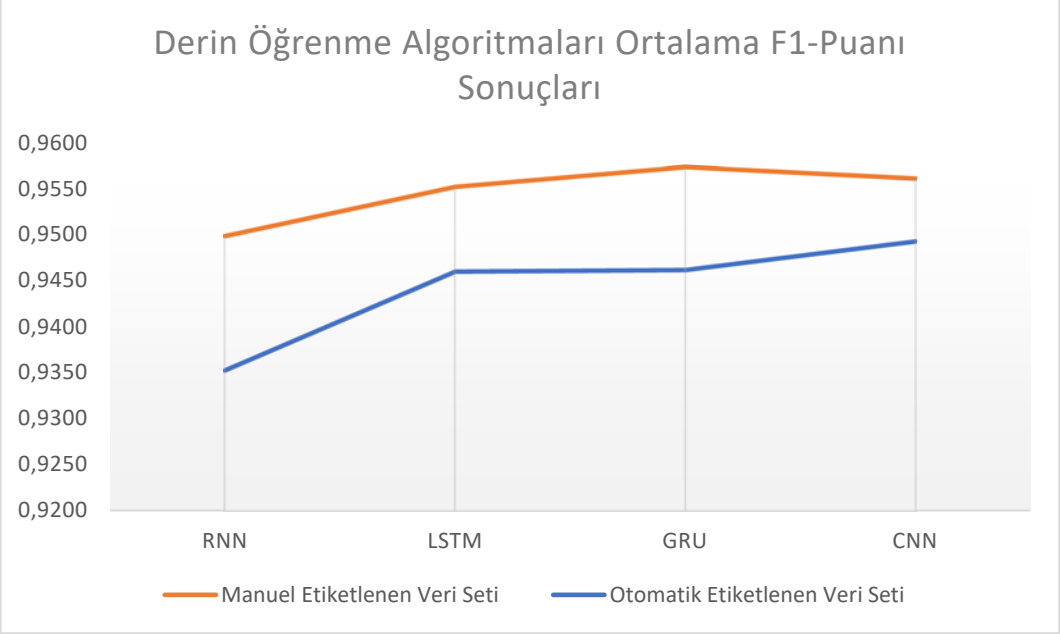
Tablo 4.13. Otomatik Veri Seti Kullanılarak Derin Öğrenme Algoritmaları Tarafından Elde Edilen F1-Puanı Sonuçları

Gömme Boyutu	Katman Birimi	RNN	LSTM	GRU	CNN
100	100	0,9360	0,9471	0,9458	0,9487
200	100	0,9392	**0,9478**	0,9460	0,9462
300	100	**0,9397**	0,9469	0,9463	0,9488
100	200	0,9390	0,9438	0,9462	0,9504
200	200	0,9391	0,9453	0,9459	0,9497
300	200	0,9343	0,9458	0,9457	**0,9505**
100	300	0,9302	0,9456	0,9472	0,9487
200	300	0,9320	0,9467	**0,9473**	0,9504
300	300	0,9279	0,9453	0,9454	0,9503

Manuel ve otomatik etiketlenmiş veri setleri kullanılarak derin öğrenme algoritmaları ile elde edilen ortalama doğruluk ve F1-Puanı sonuçları Şekil 4.4 ve Şekil 4.5'te gösterilmektedir. Bu grafikler incelendiğinde CNN algoritması ile alınan sonuçların manuel ve otomatik etiketlenmiş veri setleri için birbirine yakın olduğu görülmektedir. CNN algoritmasının sonuçları daha detaylı incelendiğinde, manuel etiketlenen veri seti ile elde edilen ortalama doğruluk 0,9563 ve otomatik etiketlenen veri seti ile elde edilen ortalama doğruluk ise 0,9544 olduğu ve aradaki farkın 0,0019 olarak hesaplandığı görülmüştür. Ortalama F1-Puanı sonuçlarına bakıldığında da yine CNN algoritmasının verdiği sonuçların yüksek olduğu görülmektedir. CNN algoritmasında manuel etiketlenen veri seti ile elde edilen ortalama F1-Puanı 0,9562 ve otomatik etiketlenen veri seti ile elde edilen F1-Puanı 0,9493 olduğu görülmüştür. Bu sonuçlarda da aradaki farkın 0,0069 olarak hesaplandığı görülmüştür.



Şekil 4.4. Derin Öğrenme Algoritmaları Ortalama Doğruluk Sonuçları



Şekil 4.5. Derin Öğrenme Algoritmaları Ortalama F1-Puanı Sonuçları

Manuel ve otomatik etiketlenmiş veri setleriyle elde edilen sonuçlar karşılaştırıldığında aradaki farkların çok düşük olduğu görüldüğünden oluşturulan otomatik etiketleme modelinin kullanılabilirliği saptanmıştır. Bu sayede milyonlarca verinin olduğu büyük veri setleri bu model sayesinde dakikalar içerisinde otomatik olarak etiketlenebilecektir. Sonraki aşamada otomatik etiketlenmiş veri setinin geleneksel makine öğrenmesi ile elde edilen sonuçlar gösterilmektedir.

Geleneksel makine öğrenmesi algoritmaları deneylerinde N-gram ve ağırlıklandırma metotları kullanılmıştır. N-gram metodu bir dil modelleme tekniğidir ve metindeki ardışık kelime gruplarını temsil etmektedir. Bu çalışmada 3 farklı N-gram tekniği kullanılmıştır. Bunlar Unigram, Bigram ve Trigram'dır. Unigram, bir metindeki her bir kelimeyi ayrı ayrı temsil eden bir N-gram tekniğidir ve metindeki kelime dağılımını modellemek için kullanılır. Bigram, metindeki ardışık iki kelimeyi temsil eden bir N-gram tekniğidir ve bir kelimenin ardından gelen kelimeyi tahmin etmek için kullanılır. Trigram, metindeki ardışık üç kelimeyi temsil eden bir N-gram tekniğidir ve bir kelimenin iki önceki kelimesine ve bir sonraki kelimesine dayalı olarak bir kelimenin tahmin edilmesi için kullanılır. Ağırlıklandırma metodu, metinlerin temsil edilmesi ve özelliklerin çıkarılması için kullanılan yöntemdir. Bu çalışmada 3 farklı ağırlıklandırma yöntemi kullanılmıştır. Bunlar Terim Varlığı (Term Presence – TP), Terim Frekansı (Term Frequency – TF) ve Terim Frekansı – Ters Doküman Frekansı (Term Frequency – Inverse Document Frequency – TF-IDF)'dır. TP, bir belgede bir terimin var olup olmadığını ifade eder ve belgede terim varsa 1, yoksa değeri 0 olur. TF, bir belgede bir terimin ne sıklıkla geçtiğini ifade eder. TF-IDF ise bir terimin belge içindeki sıklığını ve tüm belgelerdeki yaygınlığını birleştiren bir yöntemdir. TP ve TF basit bir metin temsil yöntemi sağlarken, TF-IDF daha gelişmiş bir özellik çıkarım yöntemidir ve daha ayrıntılı bir metin temsili sunar.

Tablo 4.14, Tablo 4.15, Tablo 4.16 ve Tablo 4.17 otomatik etiketlenmiş veri seti kullanılarak geleneksel makine öğrenmesi algoritmaları ile elde edilen sırasıyla doğruluk, geri çağırma, kesinlik ve F1-Puanı sonuçlarını göstermektedir. Bu sonuçlarda doğruluk sonuçları incelendiğinde 0,9090 ile en yüksek sonuca ulaşan algoritmanın SVM olduğu görülmektedir. F1-Puanı sonuçlarına bakıldığında da 0,8986 ortalama ile yine SVM algoritmasının en yüksek sonucu verdiği görülmektedir.

Tablo 4.14. Otomatik Veri Seti Kullanılarak Geleneksel Makine Öğrenmesi Algoritmaları Tarafından Elde Edilen Doğruluk Sonuçları

N-Gram ve Ağırlıklandırma Metotları	LR	SVM	RF	NB
Unigram, TF	0,9074	0,9065	0,8183	0,7983
Unigram, TP	0,8999	0,8988	0,8168	0,7963

Unigram, TF-IDF	**0,9081**	**0,9090**	**0,8202**	**0,7998**
Bigram, TF	0,8958	0,8952	0,8131	0,7873
Bigram, TP	0,8898	0,8889	0,8121	0,7863
Bigram, TF-IDF	0,8961	0,8971	0,8113	0,7960
Trigram, TF	0,8936	0,8927	0,8124	0,7841
Trigram, TP	0,8878	0,8867	0,8114	0,7832
Trigram, TF-IDF	0,8935	0,8867	0,8106	0,7910

Tablo 4.15. Otomatik Veri Seti Kullanılarak Geleneksel Makine Öğrenmesi Algoritmaları Tarafından Elde Edilen Geri Çağırma Sonuçları

N-Gram ve Ağırlıklandırma Metotları	**LR**	**SVM**	**RF**	**NB**
Unigram, TF	**0,8990**	**0,8994**	**0,7723**	**0,7864**
Unigram, TP	0,8895	0,8897	0,7722	0,7807
Unigram, TF-IDF	0,8916	0,8946	0,7473	0,6952
Bigram, TF	0,8842	0,8844	0,7635	0,7470
Bigram, TP	0,8759	0,8759	0,7625	0,7415
Bigram, TF-IDF	0,8798	0,8830	0,7441	0,6989
Trigram, TF	0,8820	0,8823	0,7614	0,7371
Trigram, TP	0,8731	0,8731	0,7606	0,7315
Trigram, TF-IDF	0,8778	0,8731	0,7439	0,6909

Tablo 4.16. Otomatik Veri Seti Kullanılarak Geleneksel Makine Öğrenmesi Algoritmaları Tarafından Elde Edilen Kesinlik Sonuçları

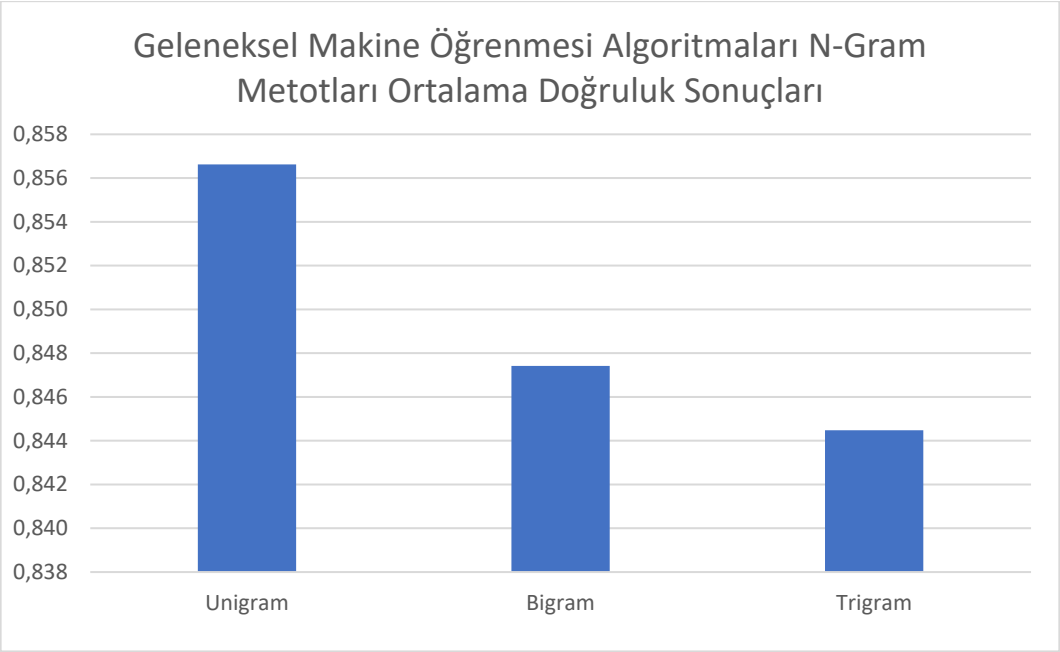
N-Gram ve Ağırlıklandırma Metotları	**LR**	**SVM**	**RF**	**NB**
Unigram, TF	0,8959	0,8939	0,8151	0,7708
Unigram, TP	0,8886	0,8862	0,8122	0,7705
Unigram, TF-IDF	**0,9033**	**0,9027**	**0,8365**	**0,8330**
Bigram, TF	0,8847	0,8832	0,8109	0,7734
Bigram, TP	0,8792	0,8774	0,8097	0,7748
Bigram, TF-IDF	0,8887	0,8882	0,8206	0,8219
Trigram, TF	0,8819	0,8800	0,8109	0,7733

Trigram, TP	0,8775	0,8753	0,8095	0,7749
Trigram, TF-IDF	0,8851	0,8753	0,8193	0,8172

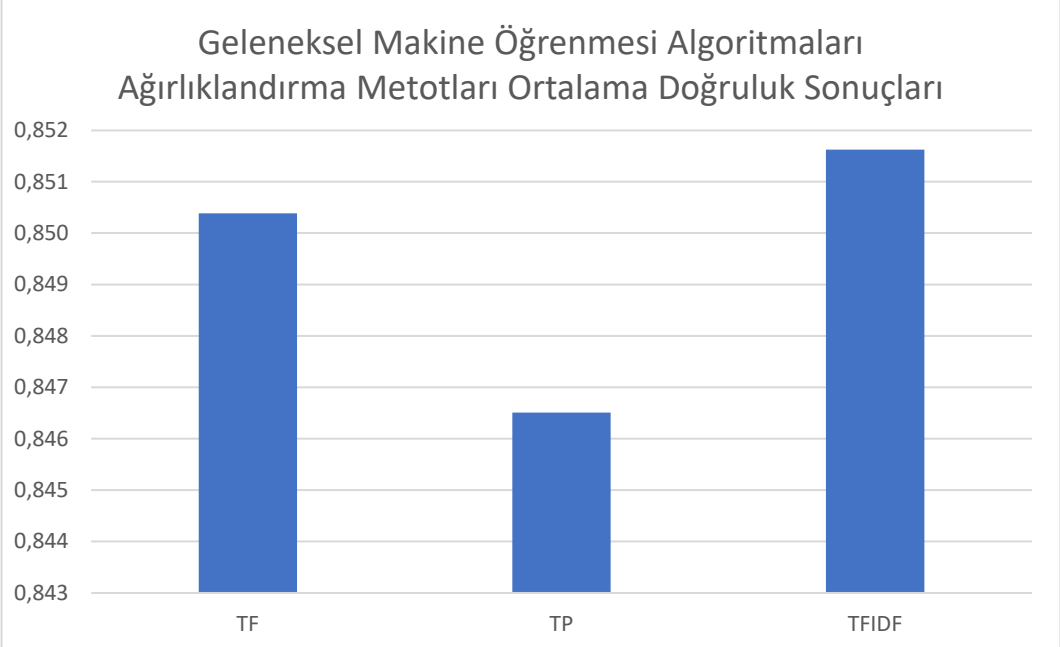
Tablo 4.17. Otomatik Veri Seti Kullanılarak Geleneksel Makine Öğrenmesi Algoritmaları Tarafından Elde Edilen F1-Puanı Sonuçları

N-Gram ve Ağırlıklandırma Metotları	LR	SVM	RF	NB
Unigram, TF	**0,8975**	0,8967	**0,7931**	**0,7785**
Unigram, TP	0,8890	0,8880	0,7917	0,7756
Unigram, TF-IDF	0,8974	**0,8986**	0,7894	0,7579
Bigram, TF	0,8844	0,8838	0,7865	0,7600
Bigram, TP	0,8775	0,8766	0,7854	0,7578
Bigram, TF-IDF	0,8842	0,8856	0,7804	0,7554
Trigram, TF	0,8820	0,8811	0,7854	0,7548
Trigram, TP	0,8753	0,8742	0,7843	0,7526
Trigram, TF-IDF	0,8814	0,8742	0,7798	0,7488

Şekil 4.6 ve Şekil 4.7'de geleneksel makine öğrenmesi algoritmaları ile yapılan deneylerde kullanılan N-Gram ve ağırlıklandırma metotlarının karşılaştırılması gösterilmektedir. Burada görüldüğü gibi N-gram metotlarından Unigram 0,8566 ortalama ile diğer metotlara göre yüksek ortalamaya sahiptir. Ağırlıklandırma metotlarından ise TF-IDF metodu 0,8516 ortalama ile diğer metotlara göre yüksek ortalamaya sahiptir. Tersine en düşük ortalama ise 0,8445 ile Trigram ve 0,8465 ile TP yöntemlerinin olduğu görülmüştür. Yukarıdaki tablolara da bu metotlar özelinde bakıldığında en yüksek sonuçlar Unigram ve TF-IDF yöntemlerinin birlikte kullanıldığı deneylerde elde edildiği net bir şekilde görülmektedir.

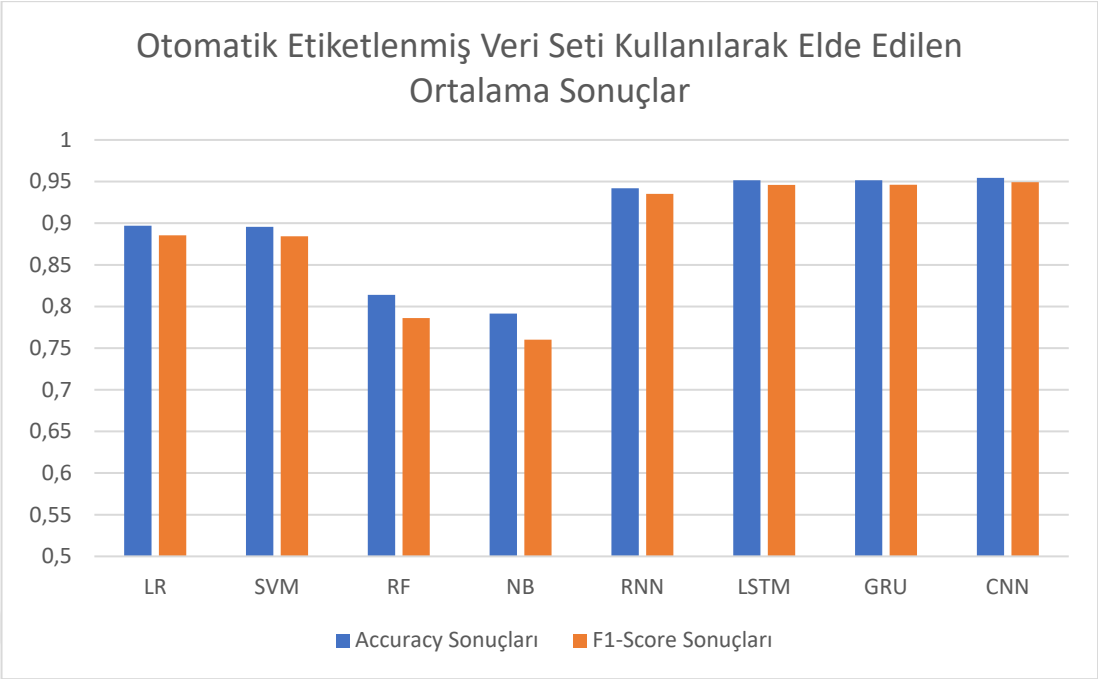


Şekil 4.6. Geleneksel Makine Öğrenmesi Algoritmaları N-Gram Metotları Ortalama Doğruluk Sonuçları



Şekil 4.7. Geleneksel Makine Öğrenmesi Algoritmaları Ağırlıklandırma Metotları Ortalama Doğruluk Sonuçları

Şekil 4.8'de otomatik etiketlenmiş veri seti kullanılarak tüm algoritmalar ile elde edilen sonuçlar gösterilmektedir. Burada derin öğrenme algoritmalarının makine öğrenmesi algoritmalarına karşı başarısı net bir şekilde görülmektedir.



Şekil 4.8. Otomatik Etiketlenmiş Veri Seti Kullanılarak Elde Edilen Ortalama Sonuçlar

5. UYGULAMA

Bu çalışma kapsamında geliştirilen sınıflandırma modelinin kullanıldığı bir web tabanlı uygulama geliştirilmiştir. Bu uygulama sayesinde veri çekme, veri ön işleme ve verinin otomatik etiketlenmesi işlemleri yapılabilmektedir. Bu çalışma kapsamında kullanılan veri setleri bu uygulama kullanılarak hazırlanmıştır.

ASP.NET MVC [43] teknolojisi kullanılarak geliştirilen bu uygulamanın arayüzünde Bootstrap mimarisi kullanılarak kullanıcı dostu bir arayüz geliştirilmiş ve mobil cihazlarda da kolaylıkla kullanılabilmesi sağlanmıştır. Selenium WebDriver [44] kütüphanesi kullanılarak Web Scraping yöntemi ile e-ticaret platformlarından ürün yorumlarının çekilebilmesi sağlanmıştır. Dosya yükleme ve indirme özellikleri sayesinde veri setlerinin CSV formatında oluşturulması sağlanmıştır. Bu çalışma kapsamında test edilen sınıflandırma modelleri uygulamaya entegre edilerek otomatik veri etiketleme işlevi sağlanmıştır.

Bu uygulama e-ticaret platformlarında satış yapan satıcılar için ürünlerdeki müşteri yorumlarının analiz edilebilmesini sağlamaktadır. Bunun yanı sıra akademik çalışmalarda oldukça fayda sağlaması ön görülmektedir.

5.1. Uygulama Modülleri

Geliştirilen uygulamada veri çekme, veri ön işleme ve veri etiketleme olmak üzere 3 modül yer almaktadır. Veri çekme modülü sayesinde e-ticaret platformlarında yer alan ürün yorumlarından etiketlenmemiş bir veri seti oluşturulabilmektedir. Veri ön işleme modülü sayesinde oluşturulan veri setinin ön işlemlerden geçirilerek temizlenmesi sağlanmaktadır. Veri etiketleme modülü sayesinde ise ön işlemlerden geçirilmiş veri setinin sınıflandırma tiplerine göre otomatik etiketlenmesi sağlanmaktadır.

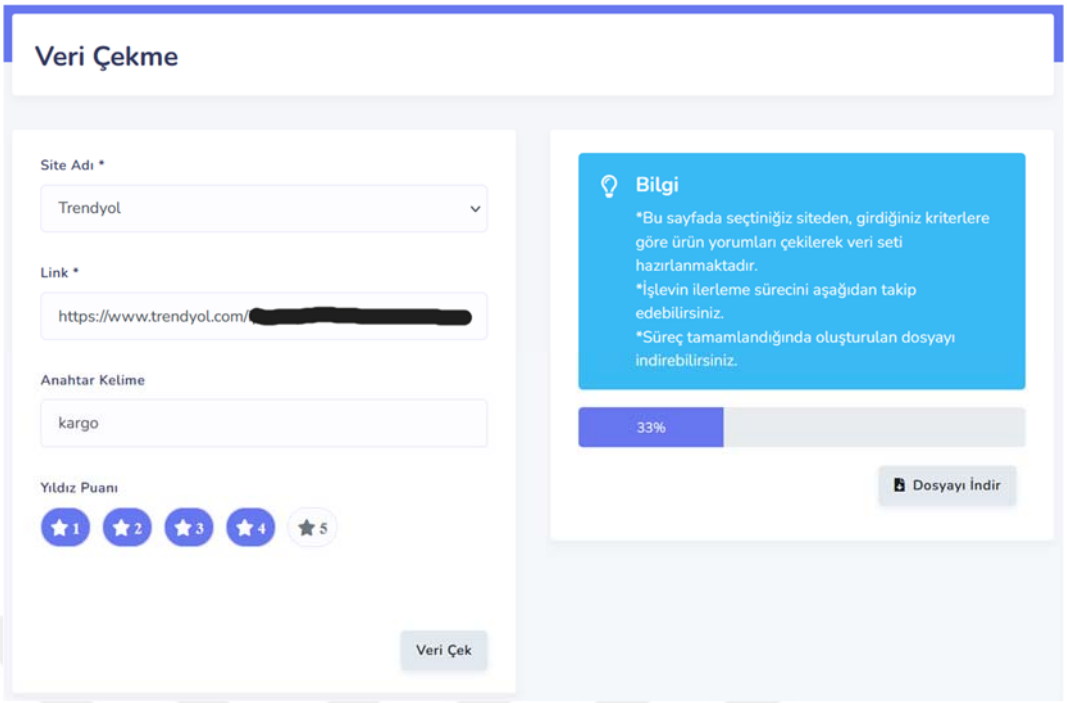
Uygulamada her modül birbirinden bağımsız çalışabilmektedir. Modüller bir arada kullanılabileceği gibi birbirinden bağımsız bir şekilde de kullanılabilmektedir. Önceden hazırlanmış bir veri seti olduğu düşünüldüğünde ve bu veri setinin bir makine öğrenmesi modeliyle otomatik etiketlenmesi istendiğinde sadece veri etiketleme modülü kullanılarak bu işlem yapılabilmektedir. Eğer bir veri seti oluşturup bunun

manuel etiketlenmesi isteniyorsa sadece veri çekme modülü kullanılarak veri seti dosyası indirilip kullanılabilmektedir. Eğer ürün yorumlarını içeren bir veri seti hazırlanıp bu veri setinin ön işlemlerden geçirildikten sonra otomatik etiketlenmesi isteniyorsa uygulamada yer alan 3 modül sırasıyla kullanılabilmektedir.

5.1.1. Veri Çekme Modülü

Veri çekme modülü e-ticaret platformlarından ürün yorumlarını çekip veri seti oluşturan modüldür. Ürün linki verilerek ve anahtar kelime, yıldız puanı gibi kriterler seçilerek istenilen yorumlar çekilebilmektedir. Bu yorumlar CSV formatında dosya olarak indirilebilmektedir.

Şekil 5.1'de veri çekme modülünün ekran görüntüsü yer almaktadır. Site adı seçeneğinde e-ticaret platformlarının isimleri yer almaktadır. Burada istenilen e-ticaret platformu seçilebilmektedir. Link alanına seçilen e-ticaret platformunda yer alan herhangi bir ürünün linki yazılabilmektedir. Anahtar kelime ve yıldız puanı alanları ürün yorumlarını filtrelemek için kullanılmaktadır. Anahtar kelime alanına bir kelime veya kelime öbeği yazarak sadece o kelime veya kelime öbeğini içeren yorumlar çekilebilmektedir. Yıldız puanı seçeneğini de kullanarak sadece seçilen yıldız puanına sahip yorumlar çekilebilmektedir. Tüm bu kriterler girilerek veri çek butonu sayesinde veri çekme işlemi yapılmaktadır. Ardından dosyayı indir butonu sayesinde çekilen verilerin CSV formatında indirilebilmektedir.



Şekil 5.1. Uygulama Veri Çekme Modülü

Şekil 5.2'de veri çekme modülü kullanılarak çekilen ürün yorumlarının bir kısmı yer almaktadır.

	A
1	Metin
2	Hiç rahat değil
3	Oturma yeri çöktü gibi oldu altı batıyo Arada, minderle kullanmakta fayda var
4	Urun minderi aslinda rahat olmasina ragmen, sirt kismi cok arkada duruyor ne yazik ki, ben rahat edemedim :(
5	yapilan yorumlara gore aldim pek rahat değil
6	çocuklar için idare eder kurulumu kolay sırt kısmı üst vida kısa geldi evden bulmak zorunda kaldım
7	ürünü yorumlara göre aldım çok kötü bir plastiği var kumaşı dersen daha kalitesiz kesinlikle tavsiye etmiyorum
8	orta sınıf çok konforlu değil ama fiyata göre idare edilebilir.
9	Her kırılacak gibi.
10	rahat ama sırt kısmı dik değil hafif geriye doğru eğik. ve malzemesi aşırı dandik. çoğu yeri ezik ve baya pis geldi.
11	Burdan yazın dediniz yazdım
12	gayet guzel basariliii
13	fiyat performans ürünü çok bir şey beklemeyin
14	Ürün güzeldi. Ama çok fazla yükselmiyor koltuk
15	Güzel ürün
16	ürün güzel ama kurulumu ile ilgili hiç bir şey yoktu

Şekil 5.2. Veri Çekme Modülü ile Oluşturulan Dosyadan Bir Kesit

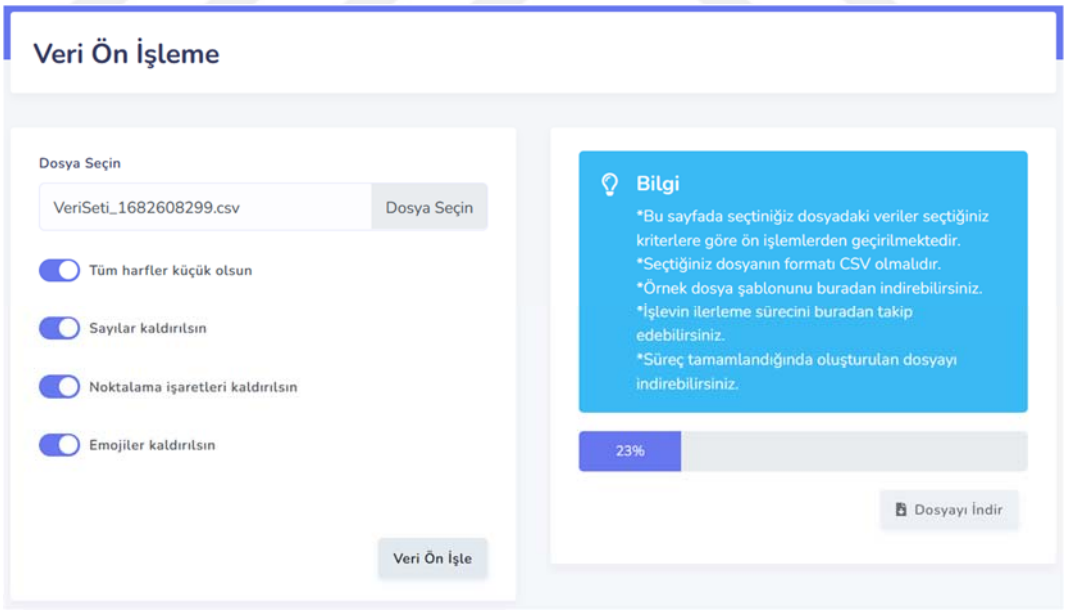
Bu modülde Selenium WebDriver kütüphanesi ile Web Scraping yöntemi kullanılmıştır. E-ticaret platformlarının arayüzleri incelenerek her platforma özel veri çekme yöntemleri geliştirilmiştir.

Veri çekme modülü sayesinde veri seti hazırlama işlemi dakikalar içerisinde yapılabilmektedir. Çalışma yapılacak ürünlere ait yorumlar ve istenilen anahtar kelime ile özgün veri setleri hazırlanabilmektedir. Yıldız puanı seçeneği sayesinde sadece olumsuz yorumlar çekilebilmektedir.

5.1.2. Veri Ön İşleme Modülü

Veri ön işleme modülünde CSV formatında dosya yükleyerek o dosyadaki verilerin seçilen kriterlere göre ön işlemlerden geçirilmesi sağlanmaktadır. Bu modülde kullanıcı istediği ön işleme adımını seçebilmektedir. Ön işleme yapıldıktan sonra yine CSV formatında dosyayı indirebilmektedir.

Şekil 5.3'te veri ön işleme modülünün ekran görüntüsü gösterilmektedir. Literatürde sıklıkla kullanılan ön işleme adımları burada seçenek olarak verilmiştir. Kullanıcı yüklediği veri setini istediği ön işleme adımlarını seçerek veri ön işle butonu ile ön işleme adımını gerçekleştirmektedir. Veri ön işleme işleminin ardından hazırlanan veri setini dosyayı indir butonu ile CSV formatı halinde indirebilmektedir.



Şekil 5.3. Uygulama Veri Ön İşleme Modülü

Şekil 5.4'te veri ön işleme modülü kullanılarak ön işlemlerden geçirilmiş veri setinin bir kısmı yer almaktadır. Tüm harflerin küçük harfe döndüğü ve noktalama işaretlerinin kaldırıldığı görülmektedir.

	A	B	C	D	E	F	G	H	I	J	K
1	Metin										
2	hiç rahat değil										
3	oturma yeri çöktü gibi oldu altı batıyo arada minderle kullanmakta fayda var										
4	urun minderi aslinda rahat olmasina ragmen sirt kismi cok arkada duruyor ne yazik ki ben rahat edemedim										
5	yapilan yorumlara gore aldim pek rahat değil										
6	çocuklar için idare eder kurulumu kolay sırt kısmı üst vida kısa geldi evden bulmak zorunda kaldım										
7	ürünü yorumlara göre aldım çok kötü bir plastiği var kumaşı dersen daha kalitesiz kesinlikle tavsiye etmiyorum										
8	orta sınıf çok konforlu değil ama fiyata göre idare edilebilir										
9	her kırılacak gibi										
10	rahat ama sırt kısmı dik değil hafif geriye doğru eğik ve malzemesi aşırı dandik çoğu yeri ezik ve baya pis geldi										
11	burdan yazın dediniz yazdım										
12	gayet guzel basariliii										
13	fiyat performans ürünü çok bir şey beklemeyin										
14	ürün güzeldi ama çok fazla yükselmiyor koltuk										
15	güzel ürün										
16	ürün güzel ama kurulumu ile ilgili hiç bir şey yoktu										

Şekil 5.4. Veri Ön İşleme Modülü ile Oluşturulan Dosyadan Bir Kesit

Bu modülde ön işleme adımları için Microsoft .NET Framework'ü kullanılmıştır. Tüm harflerin küçük harfe dönüştürülmesi için ToLower fonksiyonu, diğer ön işleme adımları için Regex yapısında yer alan Replace fonksiyonu kullanılmıştır.

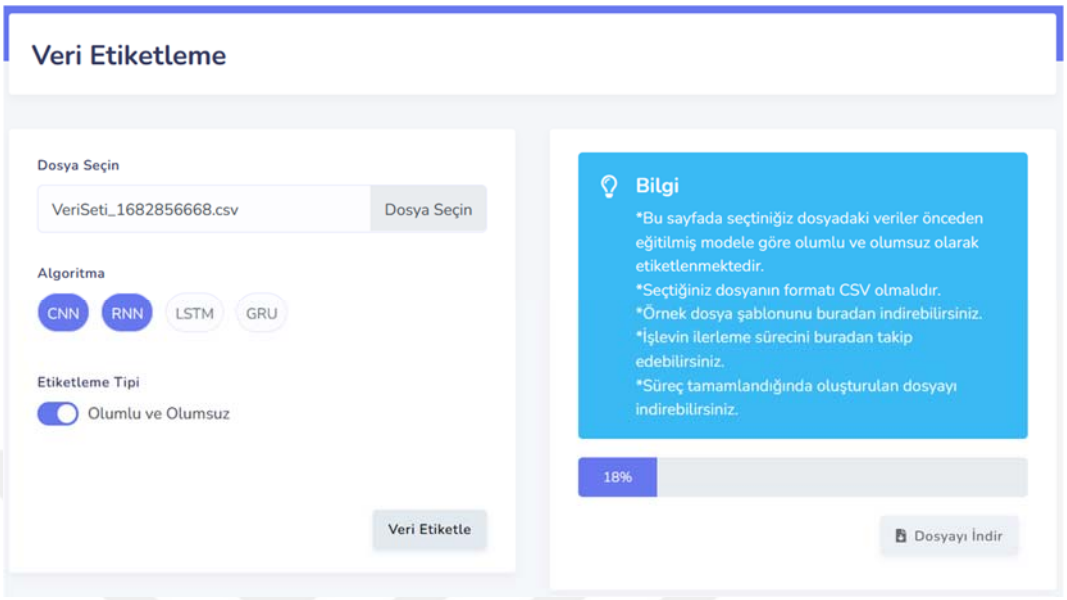
Veri ön işleme modülü sayesinde büyük çaplı veri setleri saniyeler içerisinde tüm ön işlemlerden geçirilerek test aşamasına hazır hale getirilebilmektedir. Yapılan çalışmalara göre veri setleri istenilen ön işleme adımlarından geçirilerek sonuçların karşılaştırılması yapılabilmekte ve ön işleme adımlarının sonuçlarda faydası görülebilmektedir.

5.1.3. Veri Etiketleme Modülü

Veri etiketleme modülünde veri setinin belirli bir sınıflandırma tipine göre otomatik etiketlenmesi sağlanmaktadır. Bu etiketleme işlemi bu çalışma oluşturduğumuz manuel etiketlenen veri seti ile eğitilen LSTM modeli kullanılarak yapılmaktadır. Ayrıca kullanıcı ekrandan birden fazla algoritma seçerek onların karşılaştırılmasının yapıldığı bir çıktı da alabilmektedir.

Şekil 5.5'te veri etiketleme modülünün ekran görüntüsü yer almaktadır. Kullanıcı etiketlenmemiş veri seti dosyasını yükleyerek, algoritma ve sınıflandırma tipini seçerek veri etiketle butonu sayesinde otomatik etiketleme işlemini

yapabilmektedir. Etiketlenen veri seti dosyayı indir butonu sayesinde CSV formatında indirilebilmektedir.



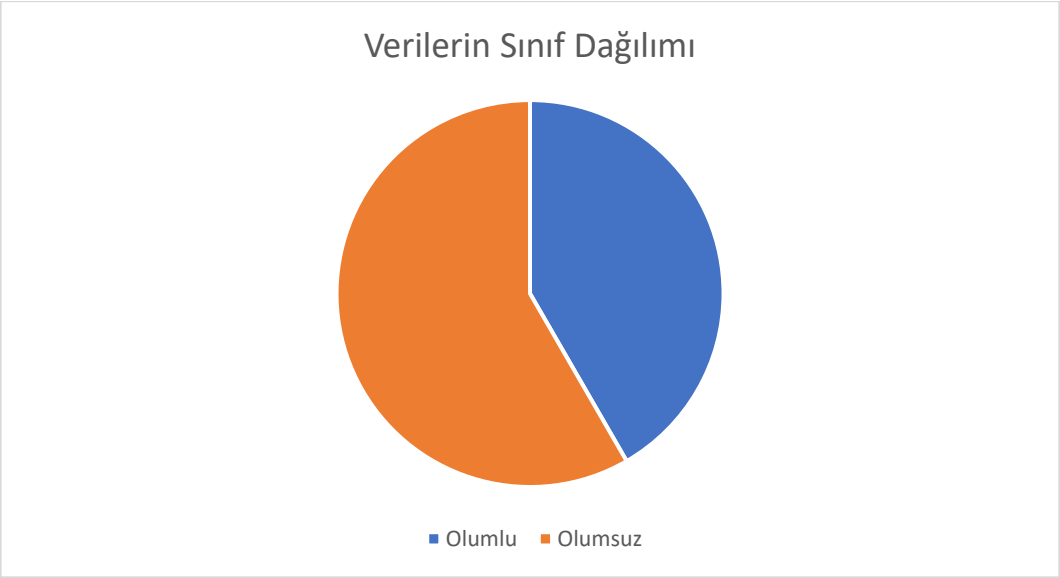
Şekil 5.5. Uygulama Veri Etiketleme Modülü

Şekil 5.6'da veri etiketleme modülü ile oluşturulan otomatik etiketlenmiş veri setinden bir bölüm gösterilmektedir. Buradan da görüldüğü gibi bu çalışma kapsamında oluşturulan sınıflandırma modelimiz oldukça başarılı çalışmaktadır.

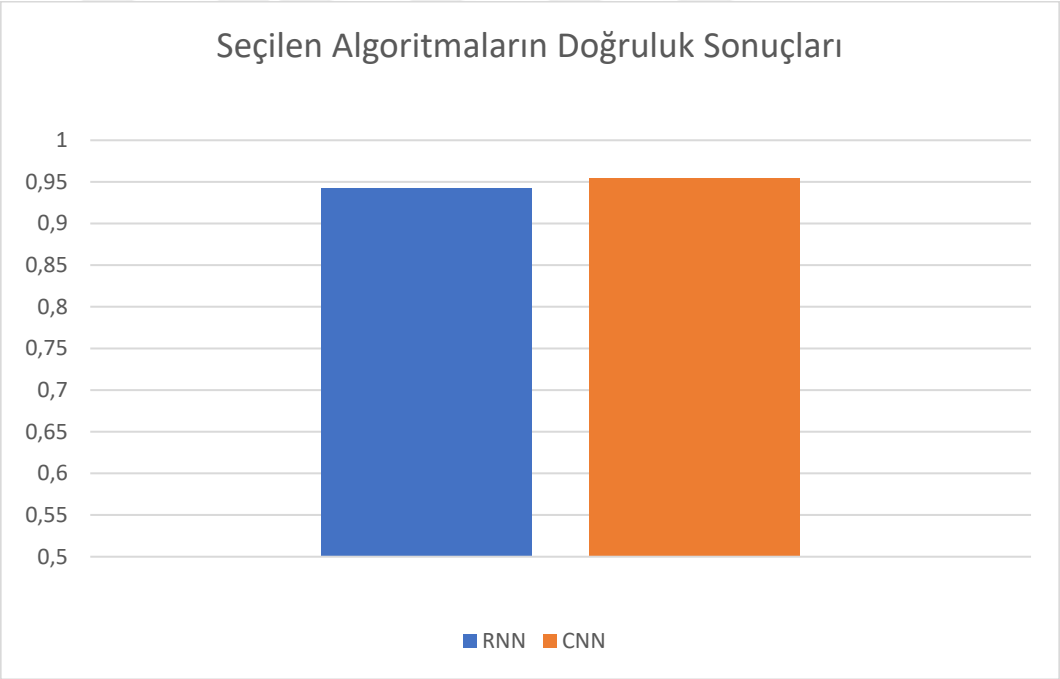
	A	B
1	Metin	Durum
2	hiç rahat değil	0
3	oturma yeri çöktü gibi oldu altı batıyo arada minderle kullanmakta fayda var	0
4	urun minderi aslinda rahat olmasina ragmen sirt kismi cok arkada duruyor ne yazik ki ben rahat edemedim	1
5	yapilan yorumlara gore aldim pek rahat değil	0
6	çocuklar için idare eder kurulumu kolay sırt kısmı üst vida kısa geldi evden bulmak zorunda kaldım	1
7	ürünü yorumlara göre aldım çok kötü bir plastiği var kumaşı dersen daha kalitesiz kesinlikle tavsiye etmiyorum	0
8	orta sınıf çok konforlu değil ama fiyata göre idare edilebilir	0
9	her kırılacak gibi	0
10	rahat ama sırt kısmı dik değil hafif geriye doğru eğik ve malzemesi aşırı dandik çoğu yeri ezik ve baya pis geldi	0
11	burdan yazın dediniz yazdım	0
12	gayet guzel basariliii	1
13	fiyat performans ürünü çok bir şey beklemeyin	1
14	ürün güzeldi ama çok fazla yükselmiyor koltuk	0
15	güzel ürün	1
16	ürün güzel ama kurulumu ile ilgili hiç bir şey yoktu	1

Şekil 5.6. Veri Etiketleme Modülü ile Etiketlenmiş Veri Seti

Şekil 5.7 ve Şekil 5.8'de veri etiketleme modülü kullanılarak elde edilen istatiksel çıktılar gösterilmektedir. Şekil 5.7'deki grafik bu modül sayesinde etiketlenen verilerin sınıf dağılımını gösteren bir grafiktir. Şekil 5.8'deki grafik ise seçilen algoritmaların karşılaştırılmasının yapıldığı bir grafiktir.



Şekil 5.7. Veri Etiketleme Modülü Verilerin Sınıf Dağılımı Çıktısı



Şekil 5.8. Veri Etiketleme Modülü Seçilen Algoritmaların Doğruluk Sonuçları Çıktısı

Bu modülde bu çalışma kapsamında hazırlanan sınıflandırma modeli kullanılmıştır. Modelin çalıştırılması için Microsoft .NET Framework'ünde yer alan Process yapısı kullanılmıştır. Process yapısı sayesinde Python uygulaması çalıştırılmakta ve bu uygulama ile modelimiz çalıştırılarak etiketleme sonuçları alınmaktadır.

Veri etiketleme modülü sayesinde binlerce veri içeren veri setlerinin otomatik etiketlenmesi sağlanmıştır. İnsan eliyle etiketlenmesi belki de aylar sürecek olan veri setlerini dakikalar içerisinde otomatik olarak etiketleyen bu modül oldukça başarılı çalıştığı gözlemlenmiştir. Bu modülde kullanılan sınıflandırma modelinin %96'lık başarı oranı sağladığı gözlemlenmiştir. Manuel etiketlenen veri setlerinde bile etiketleme esnasında hatalar yapıldığı gözlemlenmiş ve hatta iki kişinin dahi kararsız kaldığı verilerle karşılaşıldığı görülmüştür. Bu yüzden bu modülde çalışan sınıflandırma modelimiz oldukça önemli bir işe imza atmaktadır.

6. SONUÇLAR

E-ticaret sektörünün gelişmesiyle birlikte ürünlere yapılan yorumların da muazzam bir şekilde artmasının ardından tüm yorumların manuel analiz edilmesi hem e-ticaret satıcıları hem de müşteriler için önemli bir problem haline gelmiştir. Bu problem karşısında ürün yorumlarının otomatik analizini gerçekleştiren bir çalışma sunulmuştur. Sunulan bu çalışmada ürün yorumlarının otomatik sınıflandırılması sağlanarak sektöre fayda sağlaması amaçlanmıştır. Çalışma sonucunda yüksek başarı sonuçları elde edilmiş ve gerçek hayatta kullanılabilecek bir çalışma gerçekleştirmiştir. Bu kapsamda geliştirilen web tabanlı uygulama sayesinde kullanıcılar istedikleri ürünün yorumlarını analiz ederek olumlu ve olumsuz yorumların sayılarını çıkarabilecektir.

Çalışma kapsamında iki veri seti hazırlanmış ve bu veri setlerinden biri manuel biri de otomatik etiketlenmiştir. Manuel olarak etiketlenen veri farklı zamanlarda iki defa etiketlenmiş ve etiketleme sonuçları karşılaştırılarak doğruluk oranı arttırılmaya çalışılmıştır. Otomatik etiketlenen veri seti ise bu çalışma kapsamında oluşturduğumuz LSTM modeli ile otomatik etiketlenmiştir. Bu LSTM modeli manuel etiketlenen veri seti ile eğitilmiştir. Manuel ve otomatik etiketlenen veri setleri ile 4 derin öğrenme algoritması kullanılarak sonuçlar alınıp karşılaştırılmıştır. Derin öğrenme algoritmalarında çeşitli parametreler kullanılıp en yüksek sonuçlar elde edilmeye çalışılmıştır. Sonuçlar incelendiğinde otomatik etiketlenen veri setinin kullanıldığı deneylerde başarılı sonuçlar elde edilmiştir. Sonrasında otomatik etiketlenmiş veri seti ile makine öğrenmesi algoritmaları test edilmiştir. Makine öğrenmesi algoritmaları ile yapılan deneylerde çeşitli ağırlıklandırma yöntemleri kullanılmış ve en yüksek sonuçlar alınmaya çalışılmıştır. Elde edilen sonuçlar karşılaştırıldığında derin öğrenme algoritmalarının makine öğrenmesi algoritmalarına karşı başarısı gözlemlenmiştir. Derin öğrenme algoritmalarından CNN algoritması kullanılarak manuel etiketlenmiş veri seti ve otomatik etiketlenmiş veri seti ile elde edilen sonuçlar incelendiğinde birbirine yakın sonuçlar olduğu gözlemlenmiştir. Bu otomatik etiketleme modelinin başarısını göstermektedir.

Akademik çalışmalarda veri seti hazırlama sürecinin zorlu ve uzun bir süreç olduğu bilinmektedir. Çalışmanın sonucunda geliştirilen uygulama, veri seti hazırlama

sürecinin kolaylıkla yapılabilmesini de sağlamaktadır. Duygu analizi konulu akademik çalışmalarda bu geliştirilen uygulama kullanılarak otomatik etiketlenmiş veri seti oluşturma işlemi çok kısa bir sürede kolaylıkla yapılabilecektir.

Gelecek çalışmalarda, hazırlanan manuel ve otomatik etiketlenmiş veri setleri kullanılarak farklı sınıflandırma modelleri üzerine yoğunlaşılabilir. Geliştirilen uygulama sayesinde birçok veri seti hazırlanarak Türkçe dili tabanlı akademik çalışmalarda artış görülebilir.

KAYNAKLAR

[1] Çevrimiçi: WikiPedia: https://tr.wikipedia.org/wiki/Elektronik_ticaret (Erişim Tarihi: 13.04.2023).

[2] Çevrimiçi: Ticimax: https://www.ticimax.com/blog/e-ticaretin-gelisim-sureci-hakkinda-kapsamli-rehber (Erişim Tarihi: 13.04.2023).

[3] Çevrimiçi: E-Ticaret Bilgi Platformu: https://www.eticaret.gov.tr/istatistikler (Erişim Tarihi: 13.04.2023).

[4] Gedik, Y. E-Ticaret: Teorik Bir Çerçeve. Ankara Üniversitesi Sosyal Bilimler Dergisi. 2021, 12(1), 184-198.

[5] Rumelli, M., Akkuş, D., Kart, Ö., Isik, Z. Türkçe Metinlerde Makine Öğrenmesi Algoritmaları ile Duygu Analizi. 2019 Innovations in Intelligent Systems and Applications Conference (ASYU). 31 Ekim – 02 Kasım, 2019, İzmir (IEEE, 1-5).

[6] Topçu, A. E., Erkaya, A. E. Recent Research and Assessments for Computer Engineering. Ed.: Ahmet Ercan Topçu, Livre de Lyon, Lyon, Fransa, 2020, s. 1-14.

[7] Tuzcu, S. Çevrimiçi Kullanıcı Yorumlarının Duygu Analizi ile Sınıflandırılması. ESTUDAM Bilişim Dergisi. 2020, 1(2), 1-5.

[8] Mengutayci, Ü., Temurtas, H. Yapay Sinir Ağları ile Türkçe Otel Yorumlarının Sınıflandırılması. International Black Sea Coastline Countries Scientific Research Symposium - VI. Nisan, 2021, 683-687.

[9] Mayda, I., Korkmaz, M. Sentiment Analysis with Turkish Adjective Dictionary. 2018 Innovations in Intelligent Systems and Applications Conference (ASYU). 4-6 Ekim, 2018, Adana (IEEE, 1-6).

[10] Gezici, G., Yanıkoğlu, B. Sentiment Analysis in Turkish. Theory and Applications of Natural Language Processing. 2018, 255–271.

[11] Pervan, N., Keleş, H. Y. Sentiment Analysis Using A Random Forest Classifier On Turkish Web Comments. Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering. 2017, 59(2), 69-79.

[12] Onan, A. Ensemble of Classifiers and Term Weighting Schemes for Sentiment Analysis in Turkish. Scientific Research Communications. 2021, 1(1), 1-12.

[13] Demircan, M., Seller, A., Abut, F., Akay, M. F. Developing Turkish sentiment analysis models using machine learning and e-commerce data. International Journal of Cognitive Computing in Engineering. 2021, 2, 202-207.

[14] Toçoğlu, M. A., Öztürkmenoğlu, O., Alpkoçak, A. Emotion Analysis from Turkish Tweets using Deep Neural Networks. IEEE Access. 2019, 7, 183061-183069.

[15] Çataltaş, M., Doğramacı, S., Yumuşak, S., Öztoprak, K. Extraction of Product Defects and Opinions from Customer Reviews by Using Text Clustering and Sentiment Analysis. 2020 IEEE International Conference on Big Data (Big Data), Atlanta, ABD, 2020, 4529-4534.

[16] Atalay, M., Çelik, E. Büyük Veri Analizinde Yapay Zeka ve Makine Öğrenmesi Uygulamaları. Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü. 2017, 9(22), 155-72.

[17] Kotsiantis, S. B. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering. 2007, 160(1), 3-24.

[18] Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., Aljaaf, A.J. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. Ed: Berry, M., Mohamed, A., Yap, B. Supervised and Unsupervised Learning for Data Science. Springer, Cham. 2020, 3-21 s.

[19] Akanksha, E., Jyoti, Sharma N., Gulati, K. Review on Reinforcement Learning, Research Evolution and Scope of Application. 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, 1416-1423

[20] Maalouf, M. Logistic regression in data analysis: an overview. International Journal of Data Analysis Techniques and Strategies. 2011, 3(3), 281-299.

[21] Mahesh, B. Machine Learning Algorithms - A Review. International Journal of Science and Research (IJSR). 2020, 9(1), 381-386.

[22] Drăguţ, L., Belgiu, M. Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing. 2016, 114, 24-31.

[23] LeCun, Y., Bengio, Y., Hinton, G. Deep learning. Nature. 2015, 521, 436–444.

[24] Yang, S., Yu X., Zhou, Y. LSTM and GRU Neural Network Performance Comparison Study: Taking Yelp Review Dataset as an Example. 2020

International Workshop on Electronic Communication and Artificial Intelligence (IWECAI). Shanghai, China, 2020, 98-101.

[25] Şeker, S. E. Metin Madenciliği (Text Mining). YBS Ansiklopedi. 2015, 2(3), 30-32.

[26] Ain, Q. T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B., Rehman, A. Sentiment Analysis Using Deep Learning Techniques: A Review. International Journal of Advanced Computer Science and Applications (IJACSA). 2017, 8(6), 424-433.

[27] Adalı E. Doğal Dil İşleme. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi. 2012, 5(2).

[28] Süzen, A. A., Yıldız, Z., Yılmaz, T. LSTM tabanlı Derin Sinir Ağı ile Ayak Taban Basınç Verilerinden VKİ Durumlarının Sınıflandırılması. BEÜ Fen Bilimleri Dergisi. 2019, 8(4), 1392-1398.

[29] Yılmaz, M. C., Orman, Z. LSTM Derin Öğrenme Yaklaşımı ile Covid-19 Pandemi Sürecinde Twitter Verilerinden Duygu Analizi. Acta Infologica. 2021, 5(2), 359-372.

[30] Çetiner, H. GRU ve CNN Tabanlı Hibrit Derin Öğrenme Modeli ile Çok Etiketli Metin Analizi. 3rd International Conference on Applied Engineering and Natural Sciences. 20-23 Temmuz, 2022, Konya.

[31] Karga, K. Covid-19 Pandemi Sürecinin Eğitim Üzerindeki Etkilerinin Makine Öğrenme Teknikleriyle Tespit Edilmesi. Manisa Celal Bayar Üniversitesi, Fen Bilimleri Enstitüsü, Yazılım Mühendisliği Ana Bilim Dalı, Manisa, 2022, 54 s. (Yüksek Lisans Tezi)

[32] Habek, G. C. Makine Öğrenmesi Teknikleriyle Kripto Para Duygu Analizi. Manisa Celal Bayar Üniversitesi, Fen Bilimleri Enstitüsü, Yazılım Mühendisliği Ana Bilim Dalı, Manisa, 2022, 64 s. (Yüksek Lisans Tezi).

[33] Toçoğlu, M. A., Çelikten, A., Aygün, İ., Alpkoçak, Adil. Türkçe Metinlerde Duygu Analizi İçin Farklı Makine Öğrenmesi Yöntemlerinin Karşılaştırılması. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi. 2019, 21(63), 719-725.

[34] Göçgün, Ö. F., Onan, A. Amazon Ürün Değerlendirmeleri Üzerinde Derin Öğrenme/Makine Öğrenmesi Tabanlı Duygu Analizi Yapılması. Avrupa Bilim ve Teknoloji Dergisi. 2021, 24, 445-448.

[35] Onan, A. Twitter mesajları üzerinde makine öğrenmesi yöntemlerine dayalı duygu analizi. Yönetim Bilişim Sistemleri Dergisi, 2017, 3(2), 1-14.

[36] Yilmaz, Ş. Ş., Özer, İ., Gökçen, H. Türkçe Metinlerde Derin Öğrenme Yöntemleri Kullanılarak Duygu Analizi. International Symposium of Scientific Research and Innovative Studies. 22-25 Şubat, 2021, Balıkesir.

[37] Çevrimiçi: Hepsiburada: https://www.hepsiburada.com (Erişim Tarihi: 2021).

[38] Çevrimiçi: Trendyol: https://www.trendyol.com (Erişim Tarihi: 2021).

[39] Çevrimiçi: N11: https://www.n11.com (Erişim Tarihi: 2021).

[40] Çevrimiçi: Paperspace: https://blog.paperspace.com/deep-learning-metrics-precision-recall-accuracy (Erişim Tarihi: 13.05.2023).

[41] Kaygusuz, Y. Türkçe Metinlerin Sınıflandırılmasında Kullanılmak İçin Geliştirilen Melez Bir Öz Nitelik Seçim Yöntemi. Manisa Celal Bayar Üniversitesi, Fen Bilimleri Enstitüsü, Yazılım Mühendisliği Ana Bilim Dalı, Manisa, 2021, 58 s. (Yüksek Lisans Tezi).

[42] Bozuyla, M., Özçift, A. Developing a fake news identification model with advanced deep language transformers for Turkish COVID-19 misinformation data. 2022, 30, 908-926.

[43] Çevrimiçi: Microsoft: https://learn.microsoft.com/en-us/dotnet (2021).

[44] Çevrimiçi: Selenium: https://www.selenium.dev/documentation (Erişim Tarihi: 04.08.2021).

[45] Altınel, A. B. Türkçe Metinlerde Makine Öğrenmesi Algoritmalarının Duygu Analizi Problemi Üzerindeki Performansının Kıyaslanması. Avrupa Bilim ve Teknoloji Dergisi. 2021, 28, 1056-1061.

[46] Polat, H., Ağca, Y. Tripadvisor Kullanıcılarının Türkçe ve İngilizce Yorumları Kapsamında Duygu Analizi Yöntemlerinin Karşılaştırmalı Analizi. Abant Sosyal Bilimler Dergisi. 2022, 22(2), 901-916.

[47] Babüroğlu, B., Tekerek, A., Tekerek, M. Development Of Deep Learning Based Natural Language Processing Model For Turkish. 13th International Computer and Instructional Technology Symposium. Mayıs, 2022, Kırşehir.

[48] Özçift, A., Akarsu, K., Yumuk, F., Söylemez, C. Advancing natural language processing (NLP) applications of morphologically rich languages with bidirectional encoder representations from transformers (BERT): an empirical case study for Turkish. Automatika, 2021, 62(2), 226-238.