TRANSFER ÖĞRENME TABANLI AKTİF ÖĞRENME METODU İLE DUYGU ANALİZİ

Seher LORT TOSUN

2021
YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ

Tez Danışmanı
Prof. Dr. Oğuz FINDIK

TRANSFER ÖĞRENME TABANLI AKTİF ÖĞRENME METODU İLE DUYGU ANALİZİ

Seher LORT TOSUN

Karabük Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalında
Yüksek Lisans Tezi
Olarak Hazırlanmıştır

KARABÜK
Nisan 2021

Seher LORT TOSUN tarafından hazırlanan "TRANSFER ÖĞRENME TABANLI AKTİF ÖĞRENME METODU İLE DUYGU ANALİZİ" başlıklı bu tezin Yüksek Lisans Tezi olarak uygun olduğunu onaylarım.

Prof. Dr. Oğuz FINDIK .....
Tez Danışmanı, Bilgisayar Mühendisliği Anabilim Dalı

Bu çalışma, jürimiz tarafından oy birliği ile Bilgisayar Mühendisliği Anabilim Dalında Yüksek Lisans tezi olarak kabul edilmiştir. 27/04/2021

Ünvanı, Adı SOYADI (Kurumu)	İmzası
Başkan : Prof. Dr. Oğuz FINDIK (KBÜ)	.....
Üye : Dr. Öğr. Üyesi Rafet DURGUT (KBÜ)	.....
Üye : Dr. Öğr. Üyesi İlker YILDIZ(İBU)	.....

KBÜ Lisansüstü Eğitim Enstitüsü Yönetim Kurulu, bu tez ile, Yüksek Lisans derecesini onamıştır.

Prof. Dr. Hasan SOLMAZ .....
Lisansüstü Eğitim Enstitüsü Müdürü

"Bu tezdeki tüm bilgilerin akademik kurallara ve etik ilkelere uygun olarak elde edildiğini ve sunulduğunu; ayrıca bu kuralların ve ilkelerin gerektirdiği şekilde, bu çalışmadan kaynaklanmayan bütün atıfları yaptığımı beyan ederim."

Seher LORT TOSUN

ÖZET

Yüksek Lisans Tezi

TRANSFER ÖĞRENME TABANLI AKTİF ÖĞRENME METODU ile DUYGU ANALİZİ

Seher LORT TOSUN

Karabük Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı:
Prof. Dr. Oğuz FINDIK
Nisan 2021, 54 sayfa

Günümüzde insanların birbirleri ile etkileşim kurdukları, interaktif paylaşımlar yaptıkları internet tabanlı sosyal medya araçları popülerleşmiştir. Bu araçlar ile video, ses ya da metin içerikli paylaşımlar kolaylıkla yapılabilmektedir. Kullanıcılar bu araçlar sayesinde kişisel görüşlerini, kurum-nesnelere karşı tutumlarını sıklıkla paylaşmaktadır. Bu uygulamaların da yaygınlaşmasıyla Doğal Dil İşleme (DDİ) alt çalışmalarından biri olan Duygu Analizi (DA) önem kazanmıştır.

DA bireylerin ya da toplumun kurum, nesne ya da olgular hakkındaki fikirlerinin çeşitli makine öğrenimi yaklaşımları ile ortaya çıkarılmasıdır. DA çalışmalarında duygu sınıfı belirlenmiş etiketli verilere ihtiyaç vardır. Ancak veri etiketlemek zaman ve maliyet gerektiren bir işlemdir. Bu problemi çözmek için Aktif Öğrenme (AÖ) ve Transfer Öğrenme (TÖ) gibi yaklaşımlar kullanılabilir. AÖ mevcuttaki az sayıdaki etiketli veriden faydalanılarak duygu analizi yapılan çalışmalarda kullanılan bir

yaklaşımdır. TÖ bir kaynak domainden edinilen bilginin başka bir domain olan hedef domaininde kullanılması prensibine dayanır.

Bu tez çalışmasında daha önce yürütülmüş çalışmalardan farklı olarak AÖ ve TÖ yaklaşımları birlikte kullanılarak hibrit bir modelle Türkçe metinlerde farklı domainler arası duygu analizi çalışması yapılmıştır. Böylece farklı domainler arasında temsil yeteneği yüksek az miktardaki etiketli veri transfer edilerek sınıflandırma yapılabilmiştir. Çalışmada film, kitap, mutfak, elektronik ve DVD olmak üzere beş farklı domain kullanılmıştır. Domainler kullanıcıların bu ürün grupları için yaptıkları yorumlardan oluşmaktadır.

Çalışmada AÖ' nin sınıflandırma başarısının başlangıç kümesine oldukça bağlı olduğu görülmüştür. Hedef ve kaynak domainlerin birbirine benzer veri setler olması sınıflandırma başarısını artırır. Çalışmamızda da en yüksek sınıflandırma başarısı film yorumları veri seti kaynak domain iken kitap yorumları veri seti hedef domain olması durumunda görülmüştür. Bu iki domain arası duygu analizinde TÖ tabanlı AÖ yapıldığında ortalama %8 sınıflandırma başarısı artışı olmuştur. Sınıflandırma için Lojistik Regresyon, Destek Vektör Makinesi ve Yapay Sinir Ağları kullanılmış, başarı oranları kıyaslanmıştır. Domainler arası duygu analizi çalışması yapılırken TÖ tabanlı AÖ yaklaşımının ürettiği çıktılar ile başarılı sonuçlar elde edilmiştir.

Anahtar Sözcükler : Duygu Analizi, Transfer Öğrenme, Aktif Öğrenme, Doğal Dil İşleme

Bilim Kodu : 92416

ABSTRACT

M. Sc. Thesis

SENTIMENT ANALYSIS WITH TRANSFER LEARNING-BASED ACTIVE LEARNING METHOD

Seher LORT TOSUN

Karabük University
Institute of Graduate Programs
Department of Computer Engineering

Thesis Advisor:
Prof. Dr. Oğuz FINDIK
April 2021, 54 pages

Today, internet-based social media tools, where people interact with each other and share interactively, have become popular. With these tools, it is possible to share video, audio or text content easily. Thanks to these tools, users often share their personal opinions and attitudes towards institutions-objects. With the spread of these applications, Sentiment Analysis (SA), one of the sub-studies of Natural Language Processing (NLP), has gained importance.

SA is the revealing of the ideas of individuals or society about institutions, objects or phenomena with various machine learning approaches. In SA studies, labeled data with determined emotion class are needed. However, tagging data is a time- and cost-intensive process. Approaches such as Active Learning (AL) and Transfer Learning (TL) can be used to solve this problem. AL is an approach used in studies in which sentiment analysis is made by utilizing the limited number of labeled data available.

TL is based on the principle that the information obtained from one source domain is used in another domain, the target domain.

In this thesis study, different from the previous studies, AL and TL approaches were used together, and a hybrid model of sentiment analysis between different domains was conducted in Turkish texts. Thus, classification can be made by transferring a small amount of labeled data with high representativeness between different domains. Five different domains were used in the study: film, book, kitchen, electronics and DVD. Domains consist of comments made by users for these product groups.

In the study, it was seen that the classification success of AL was highly dependent on the starting set. The fact that the target and source domains are similar data sets increases the success of classification. In our study, the highest classification success was seen when the movie reviews dataset was the source domain, while the book reviews dataset was the target domain. In the sentiment analysis between these two domains, there was an average of 8% increase in classification success when TL-based AL was performed. Logistic Regression, Support Vector Machine and Artificial Neural Networks were used for classification and success rates were compared. While conducting sentiment analysis between domains, successful results were obtained with the outputs produced by the TL-based AL approach.

Key Word **:** Sentiment Analysis, Transfer Learning, Active Learning, Natural Language Processing

Science Code : 92416

TEŞEKKÜR

Bu tez çalışmasının planlanmasında, araştırılmasında, yürütülmesinde ve oluşumunda ilgi ve desteğini esirgemeyen, engin bilgi ve tecrübelerinden yararlandığım, yönlendirme ve bilgilendirmeleriyle çalışmamı bilimsel temeller ışığında şekillendiren sayın hocam Prof. Dr. Oğuz FINDIK'a sonsuz teşekkürlerimi sunarım.

Sevgili aileme manevi hiçbir yardımı esirgemeden yanımda oldukları için tüm kalbimle teşekkür ederim.

İÇİNDEKİLER

Sayfa

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

ÇİZELGELER DİZİNİ

Sayfa

SİMGELER VE KISALTMALAR DİZİNİ

SİMGELER

$\sum$: toplama işlemi

$\prod$: çarpma işlemi

exp : üstel fonksiyon

ln : e tabanında log

$\sqrt{a}$: karekök alma

KISALTMALAR

AÖ : Aktif Öğrenme
DA : Duygu Analizi
DDİ : Doğal Dil İşleme
MÖ : Makine Öğrenmesi
TÖ : Transfer Öğrenme
LR : Logistic Regression (Lojistik Regresyon)
SVM : Support Vector Machine (Destek Vektör Makinası)
RF : Random Forrest (Rastgele Orman)
KNN : K-Nearest Neighbors (K-En Yakın komşuluk)
YSA : Yapay Sinir Ağları

BÖLÜM 1

GİRİŞ

Duygu, belirli nesne, olay veya bireylerin insanın iç dünyasında uyandırdığı izlenim olarak tanımlanmaktadır [1]. Özellikle kişilerin ürünler, olgular, durumlar, siyasi partiler ya da firmalar hakkındaki duygu ve düşüncelerinin analizi pek çok paydaş tarafından önemsenmektedir. Örneğin ürünlerinin son kullanıcıları tarafından memnuniyet derecesinin ölçülebilmesi şirketlerin iyileştirmesi gereken özelliklerini belirlemede son derece önemlidir. Pazarlama, üretim, hizmet ve reklamcılık gibi sektörler için kullanıcıların ya da potansiyel müşterilerinin kendi ürünlerine tutumu öğrenebilmesi değerlidir. Artık kişiler de bir ürün ya da hizmet satın almadan önce yapılan yorum ve değerlendirmeleri inceleyerek karar verme eğilimindedir. Siyasi kuruluşlar seçmen kitlesinin tutumunu ilgili metinlerin analizi ile kolaylıkla ölçebilmektedir. Ayrıca bahsedilen pek çok gereksinimle birlikte duygu analizi, yeni bir akademik pencere olarak başarılı sonuçların elde edildiği ve gelecekte pek çok yaklaşım ve algoritma ile daha da iyileştirilebilecek bir çalışma alanıdır.

Günümüzde yazılı basının yanı sıra kişilerin de artık duygu ve düşüncelerini kolaylıkla paylaşmalarına olanak sağlayan web tabanlı platformların popülerliği, büyük metin veri kümelerinin oluşmasına neden olmuştur [2]. Bu metinlere erişim ise farklı programlama dilleri ve bunların editörleri ile kolaylıkla yapılabilmektedir. Elde edilen büyük miktardaki bu verinin manuel işlenmesi söz konusu değildir. Metinlerin işlenmesi ve analizinde Doğal dil işleme yaklaşımları kullanılmaktadır.

Doğal dil işleme, işlenmemiş metinden bilgi çıkarımı yapabilen bir yapay zeka yaklaşımıdır. Metinler insanların birbirleri ile etkileşimini sağlayan araç olan doğal dilin yazılı formlarıdır. Dolayısıyla doğal dil işleme ile insanlar arasındaki etkileşimi sağlayan dilin yapısı dijital ortama aktarılmış, dijital ortamda incelenebilir ve modellenebilir hale gelir. Bilgisayarın çeşitli matematiksel ve modelleme çıkarma

özelliği ile insan dilinin entegrasyonu bu bilimin temel prensibidir. Doğal dil işleme yaklaşımı ile duygu analizi, metin özetleme, farklı diller arası çeviri yapma, bilgi çıkarma, metin sınıflandırma, konuşmacı tanıma gibi pek çok alanda çalışma yapılmıştır.

Dillerin birbirinden farklı yapılara sahip olması, farklı dil ailelerine ait olması daha çeşitli çalışmaların yürütülmesine olanak sağlamaktadır. Ancak dilin yapısı ya da ait olduğu aileye bakılmaksızın tüm diller için alt dilbilim gruplarından bazıları aşağıdaki gibidir [3].

- Fonoloji: Dilin seslerini inceleyen bu dilbilim dalı sözcüklerin içinde ve arasındaki seslerin yorumlanması prensibine dayanmaktadır. Genel olarak fonoloji kelimedeki seslerin belirlenmesi, telaffuzu ve vurgusu olmak üzere üç temel alana sahiptir [4]. Fonoloji temelli doğal dil işleme yöntemlerinde insan sesi çalışılan bilgisayar modeli için girdi olarak kabul edilmektedir [5]. Bu yöntemlerde ses dalgalarının çeşitli sinyal işleme algoritmaları ile işlenmesi gerekmektedir. Fonoloji bilimi kullanılarak ses tanıma, konuşmacı tanıma, sesten kelimelerin çıkarılması ve çıkarılan kelimelerin anlamlandırması yapılabilmektedir.

- Morfoloji: Kelimelerin bileşenlerine ayrıştırıldığı dilbilim dalıdır. Kelime kökünün ön eklerinin ya da son eklerinin ayrıştırılmasında dilin bağlı olduğu dil ailesine göre sonlu otomatlar kullanılmaktadır [6]. Morfoloji temelli doğal dil işleme yöntemlerinde cümledeki kelimeler ve kelimelerin birbiriyle ilişkileri üzerinde durulmaktadır. Morfoloji bilimi kullanılarak kelimeler ek ve köklerine ayrıştırılır, cümlede özne, fiil ya da nesne gibi ögelere ayrıştırma yapılır, kelimelerin birbiriyle ilişkisi incelenerek metin sınıflandırma işlemleri yapılır [7]. Ayrıca spam mail tespiti işleminde de Morfoloji bilimi kullanılmaktadır.

- Semantik: Kelime ve cümlelerin anlamlarının incelendiği bu dilbilim dalında anlamsal olarak benzer kelimelerin-terimlerin benzer dokümanlarda yer alacağı varsayılmaktadır. Terimlerin ve dokümanların bağlam olarak

benzerliklerini göstermek için matrisler kullanılmaktadır [8]. Kelime ve cümlelerin birbiri ile ilişkileri incelenerek anlam çıkarma işlemleri yapılırken özellikle kelime türü bilgisi önemlidir. Semantik bilimi kullanılarak metinlerin özetlenmesi, soru-cevap otomatları, dünya bilgisinin kullanıldığı sistemler tasarlanmaktadır [9].

- Pragmatik: Kelimelerin gerçek anlamları ya da bağlamsal bilgilerinin dışında insan iletişimi sırasında gerçek anlamlarının ötesine geçen kullanımlarının incelendiği dilbilim dalıdır [10]. Anlamlardaki belirsizliğin giderilmesi ya da analizi ile ilgilenilmektedir. Eş anlamlı kelimelerin ya da söylenenin tam tersini ifade eden sarkastik metinlerin oluşturacağı anlamca belirsizlik konusu bu bilim dalında incelenmektedir.

Her geçen gün doğal dil işleme tabanlı sistemlerin başarı oranı artmakta daha da güvenilir modeller üretilebilmektedir. Başarı oranının artmasında donanımlı bilgisayarların üretilmesi de rol oynamaktadır. Doğal dil işlemede kullanılan karmaşık Makine Öğrenmesi algoritmaları bu donanımlı cihazlarda daha performanslı çalışmaktadırlar. Bahsedilen gereksinimler ve gelecek için ümit vaat eden bir alan olan Doğal Dil İşleme bu çalışmanın motivasyon sebebi olmuştur. Doğal dil işleme ile yazım yanlışlarını düzeltme, özetleme, bilgi çıkarma, diller arası çevrim, metinleri seslendirilme, soru-cevap uygulamaları ya da duygu analizi gibi çalışmalar yapılmaktadır.

Duygu analizi, kişinin kurum, nesne, olgu vb. üzerine paylaştığı metin işlenerek o kurum, nesne ya da olguya karşı tutumu bilgisinin ortaya çıkarıldığı Doğal Dil İşleme çalışma alanıdır [11]. Duygu analizi 2000'li yılların başından beri en popüler konularından biridir [12]. Özellikle Web 2.0 teknolojisi ile kullanıcıların birbirleriyle interaktif bir şekilde iletişime geçip birçok platformda paylaşım yapmaya başlamaları bu alana ilgiyi artırmıştır. Duygunun tanımı için varlık, varlığın özelliği, duygu sınıfı, duygu sahibi ve zaman olmak üzere beş parametre kullanılır. Beş parametrenin tamamı kullanılmak zorunda değildir.

Duygu analizi, doküman seviyesinde, cümle seviyesinde ve özellik seviyesinde olmak üzere üç temel alanda incelenmektedir [13]. Doküman seviyesinde dokümanın tamamını değerlendirilerek duygu sınıfını belirlerken, cümle seviyesinde her bir cümle için duygu sınıfına etiketleme yapılmaktadır [14]. Ancak doküman seviyesinde duygu sınıfı belirlenirken, her bir cümlenin de ayrı ayrı duygu sınıfları belirlenip, dokümanın duygu sınıfı belirlemesinde kullanılabilir. İki seviyenin de birlikte kullanılmasının sınıflandırma başarısını artırdığı görülmüştür [15]. Özellik seviyesinde ise metindeki varlıklar ve varlıkların özellikleri çıkarıldıktan sonra her bir özelliğinden bahsedilen varlık için sınıflandırma yapılmaktadır [16]. Metinlerden varlıkların özelliklerinin çıkarılması aşamasında kelimeler ve kelimeler arasındaki mesafeler ya da ilişkilerinin göz önünde bulundurulmaktadır [17]. Ayrıca Duygu Analizi' de kullanılan yaklaşıma göre 2 gruba göre ayrım yapılmaktadır [18].

Duygu analizinde kullanılan yaklaşıma göre Makine Öğrenme Yaklaşımı ve Veri Sözlüğü Tabanlı Yaklaşım olmak üzere iki grup bulunmaktadır. Makine Öğrenme Yaklaşımında sınıflandırma, kümeleme, regresyon algoritmaları ağırlıklı olarak kullanarak verilerin sayısallaştırılmasında istatistiksel hesaplamalar yapılmaktadır. Veri Sözlüğü Tabanlı Yaklaşımda ise duygu sınıfı belirlenmiş sözlük kümeleri ya da korpus kümeleri kullanılmaktadır [18].

Makine öğrenmesi Denetimli Öğrenme ve Denetimsiz Öğrenme olmak üzere iki yaklaşım içermektedir. Denetimli Öğrenme' de verilerin ait olduğu sınıf etiketli iken, Denetimsiz Öğrenme' de verilerin ait olduğu sınıf etiketli değildir. Denetimli Öğrenme grubu da verilerin sınıfını belirlemek için kullanılan sınıflandırıcıya göre dört alt gruba ayrılmaktadır. Karar Ağacı Sınıflandırıcıları grubu verilerin sınıfının belirlenmesinde kök, yaprak ve dalların olduğu bir graf yapısı kullanmaktadır. Ağacın oluşturulması modeli eğiten veri seti ile sağlanmaktadır [19]. Lineer Sınıflandırıcı grubundaki verilerin sınıfının belirlenmesinde ayrık parametreler ve parametre ağırlıkları kullanılmaktadır [20]. Support Vector Machine (SVM) (Destek Vektör Makinesi), Yapay Sinir Ağları (YSA) algoritmaları bu sınıflandırıcılardandır. Kural Tabanlı Sınıflandırıcılı grupta duygu sınıfı belirlenirken ise sözlük türleri ve sözlük türleri örüntülerine göre kurallar belirlenmektedir. Kurallara uygun koşulları sağlayan durumlar için duygu sınıf belirteci ataması yapılmaktadır [21]. Duygu sınıfı

belirlenmesinde cümledeki kelimelerin çeşitli duygu grubundaki dokümanlarda daha çok daha az bulunma olasılıklarının göz önünde bulundurulduğu grup ise Olasılıksal Sınıflandırıcı grubudur [22-23].

Denetimli Öğrenme ile duygu analizi yapılırken duygu sınıfını belirlemek için bir sınıflandırıcı kullanılabildiği gibi birden çok sınıflandırıcı da kullanılabilmektedir. Birden çok sınıflandırıcı kullanımına tümleşik-çoklu sınıflandırıcılar denir [24]. Bu çalışmalarda bir sonraki sınıflandırıcıyı besleyen değerler için farklı yaklaşımlar kullanılmıştır. Bir sonraki sınıflandırıcıyı besleyen değer veri seti üzerinden sağlanabildiği gibi, ilk sınıflandırıcıdan da bu besleme değeri elde edilebilir [25]. Tümleşik-çoklu sınıflandırıcı olarak Yapay Sinir Ağları da sıklıkla kullanılmaktadır. Yapay Sinir Ağları' nın kullanılmasının olasılıksal sınıflandırıcılara göre daha yüksek başarılı sınıflandırmalar yaptığı görülmüştür. [26-27]

Duygu analizinde Veri Sözlüğü Tabanlı Yaklaşım' ın kullanılması Sözlük Tabanlı ve Korpus Tabanlı olmak üzere 2 grupta incelenmektedir. Sözlük Tabanlı yaklaşımda sözcüklerden oluşan küme kullanılırken Korpus Tabanlı Yaklaşımda korpus kümesi kullanılır, korpusların birbirleri ile ilişkisi göz önünde bulundurulur. Veri sözlüğü tabanlı yaklaşımda ilgili her kelimenin ait olduğu korpus ya da sözlük incelenmektedir [28]. Veri sözlüğü tabanlı yaklaşımlarda 3 temel prensip bulunmaktadır. İlk prensipte, kelimenin geçtiği korpuslara göre kendinden önceki kelimenin grubu ve kendinden sonraki kelimenin grubunun göz önünde bulundurularak olasılıksal hesaplama yapılmaktadır. İkinci prensipte, kelimenin ait olduğu gruba göre ağaç yapısı oluşturulmaktadır. Üçüncü prensipte ise kelimelerin tür bilgileri de kullanılarak kural tabanlı tanımlar yazılmaktadır [29]. Duygu analizinde Veri Sözlüğü Tabanlı Yaklaşım' ın hem de Makine Öğrenmesi Yaklaşımı'nın birlikte kullanımı da söz konusudur [30].

Duygu analizi yapılan doğal dil işleme çalışmalarında başarıyı etkileyen en önemli faktörlerden biri duygu sınıfı etiketlenmiş veri setidir. Günümüzde metinlere erişim daha kolay olsa da metinlerin sınıfını etiketlemek hala büyük zaman ve maliyet gerektiren bir işlemdir. Transfer öğrenme ve aktif öğrenme etiketli veriye olan bu ihtiyacı azaltmak için kullanılabilecek yaklaşımlardır.

Transfer Öğrenme, bir problem çözerken elde edinilen bilginin başka ama ilişkili problemde kullanması prensibine dayanan bir makine öğrenme yaklaşımıdır [31]. Böylece yeni probleme önceki problem çözümünden elde edinilen bilgi aktarımı yapılmış olur. Dolayısıyla hız, performans ve maliyette kazanım sağlanır. Transfer öğrenmede, kaynak adı verilen uzay kümesinden öğrenilen bilgi, hedef adı verilen uzay kümesinde kullanılmaktadır. Doğal Dil İşleme, Görüntü İşleme, Ses Analizi, Bilgisayarla Görme gibi pek çok alanda kullanımı vardır.

Doğal Dil İşlemede kullanılan Transfer Öğrenme yaklaşımı iki farklı metin kümesi arasında bilgi aktarımını sağlamaktadır. Kaynak metin kümesine göre eğitilen model, bazı ayarlamalar yapılarak hedef metin kümesi için kullanılmaktadır. Duygu Analizi' nde kullanılan Transfer Öğrenme yaklaşımı üç farklı grupta incelenmektedir [32]:

- Parametre transferi: Kaynak metin kümesi için kullanılan parametrelerin hedef metin kümesinde de düzenlenerek kullanılmasıdır [33]. Örneğin "Büyük" kelimesinin kaynak domaindeki gösterimi ile hedef domaindeki "Büyükçe", "Büyüklük" gibi "Büyük" kelimesine benzer kelimelerin gösterimi-temsili parametrik olarak aktarılabilir. Bu yaklaşım domainler arasında kelimelerin vektörize gösterimlerinin aktarımı olarak düşünülmektedir [34]. Kelime vektörize gösterimlerinin aktarımı için kullanan yöntemlerden biri de word2vec yaklaşımıdır.

- Kayıt transferi: Kaynak metin kümesindeki bazı örnek kayıtların hedef metin kümesine aktarılmasıdır [35]. Aktarım yapılan örneklere ağırlıklandırma işlemi yapılmaktadır. Hedef ve kaynak metin kümesinin veri dağılımının birbirine yakın olması halinde başarılı sonuçlar elde edilmektedir. İngilizce data setinin Çince karşılıkları için İngilizce için eğitilen modelin metin dili Çince olan data set için kullanılması bu grupta incelenir.

- Özellik temsili transferi: Kaynak metin kümesi temsil eden özelliklerin hedef metin kümesine aktarılmasıdır [36]. Kaynak metin kümesinin sınıflandırılmasında kullanılan ve kaynak veri kümesini en iyi temsil eden özellikler hedef ve kaynak metin kümelerinin birbirine yakın domainler olduğu

varsayımıyla hedef metin kümesini de iyi temsil edecektir yaklaşımı vardır. Her iki domainde de ayırt edicilik yeteneği olan özelliklere pivot özellikler, her iki domain için de ayırt edicilik yeteneği olmayan özelliklere non-pivot özellikler denerek pivot ve non-pivotların korelasyonuna göre aktarım yapılmaktadır [37].

Duygu analizinde kullanılan Transfer Öğrenme yaklaşımı incelendiğinde hedef domain ile kaynak domainin birer adet olabildiği gibi çoklu sayıda da olabildiği görülmüştür. Çoklu sayıda olan domainlerde transfer için domain bilgisi de kullanılmıştır [38]. Örneğin uzun kelimesi, bilgisayar domainde uzun çalışma zamanları gibi ifadeler için kullanıldığında negatif sınıf bilgisi içerirken, güzel kelimesi her domain için domainden bağımsız pozitif sınıf bilgisi içermektedir.

Transfer Öğrenme tabanlı Duygu analizinde domainler arası aktarım kayıt, parametre ya da özellik üzerinde sağlandıktan sonra sınıflandırma yöntemleri kullanılmaktadır. Kullanılan sınıflandırıcı bir ya da birden fazla olabilmektedir [39]. Tümleşik-çoklu sınıflandırıcılı domainler arası duygu analizi modeli ile tek sınıflandırıcı modeller kıyaslanmış, tümleşik-çoklu sınıflandırıcıların göre daha yüksek başarılı sınıflandırmalar yaptığı görülmüştür.

Transfer öğrenme gibi aktif öğrenme de veri etiketleme maliyetini düşürmek için kullanılan iteratif bir makine öğrenme yaklaşımıdır. Aktif Öğrenmede veri setinden başlangıç kümesini oluşturmak için örnekler seçilmektedir ve örneklerin sınıfı etiketlenmektedir. İteratif olarak etiketsiz verinin içinden bir miktar daha veri seçimi yapılmaktadır ve model yeniden eğitilmektedir. Yeni aday seçme işlemi belirlenen adım sayısı, eşik değeri ile kontrol edilmektedir ya da etiketlenecek veri kalmayınca kadar işlem devam etmektedir. Aktif öğrenme, modelin temelini oluşturacak elle ya da otomatik etiketlenecek ilk verinin seçilmesi ve modeli besleyecek yeni aday veri seçme adımları olmak üzere iki temel adımdan oluşmaktadır [40]. Modeli besleyecek ilk verinin, veri seti içindeki sınıfı en belirsiz adaylardan seçilmesi ya da rastgele seçim yapma ile belirlenmesi aktif öğrenmede kullanılmış yaklaşımlardır.

İnsanların duygularını, fikirlerini, tutumlarını paylaşabilecekleri sosyal medya platformlarında Türkçe metinlerin de sıkça paylaşılmasıyla ve bu metinlere erişimin kolaylaşmasıyla Türkçe dili için duygu analizi konusu popülerleşmiştir. Türkçe dili ile duygu analizi alanında pek çok tez, bildiri ve makale bulunmaktadır.

Türkçe, Ural-Altay dil grubuna bağlı sondan eklemeli bir dildir. Bu gruptaki diğer diller gibi yeni kelime türetmek için kelime sonlarına ekler gelmektedir. Türkçe duygu analizi alanında dilin dil grubu özellikleri de göz önünde bulundurularak kelime kökleri ve morfolojik yapıları incelenmiştir. Türkçe kelimelerin ek ve köklerine ayrılması, heceleme, imla hatalarının düzeltilmesi, kelime öbeklerinin çıkarılması gibi amaçlar için Zemberek isimli açık kaynak kodlu kütüphane geliştirilmiştir [41-42]. Kelime köklerinin elde edilmesinde iki yaklaşım kullanılmaktadır. İlk yaklaşımda eklerden arındırılan kelimenin köküne ulaşılmaya çalışılmaktadır. İkinci yaklaşımda ise soldan sağa doğru n sayıdaki harf seçilerek kök olduğu varsayımı vardır [43].

Türkçe duygu analizi alanında tek domain kullanıldığı gibi domainler arası duygu analizi de yapılabilmektedir [44]. Türkçe metinler için domainler arası duygu analizinde Transfer Öğrenme yaklaşımı temeline dayanan parametre transferi, öznitelik transferi alt yaklaşımları kullanılmıştır [45-46].

Günümüzde domainler arası duygu analizinde etiketli veriye olan ihtiyaç hala önemini korumaktadır. Bu problem için önerilmiş transfer öğrenme yaklaşımı ile domainler arası adaptasyonun sağlanmasında parametreler, öznitelik temsilleri ya da domainlerdeki kayıtlar kullanılmıştır. Türkçe dili için domainler arası duygu analizi bilgi aktarımı için parametre ya da öznitelik transferi yapılmıştır. Parametre transferi kullanımında derin öğrenme ağlarından LSTM ağında parametrelerin düzenlenmesi ile aktarımı söz konusudur [45]. Öznitelik transferinde ise kaynak domain ile hedef domain arasındaki adaptasyonu sağlayan özniteliklerin aktarımı yapılmıştır [46]. Bu çalışmada Türkçe metinler için transfer öğrenme ile birlikte aktif öğrenme kullanılarak hibrit bir domainler arası duygu analizi modeli önerilmiştir. Bu çalışmada, ilgili alanda yapılan ilk çalışmalardandır. Kaynak domaini en iyi temsil eden etiketli metinler aktif öğrenmenin ilk adımı olan başlangıç kümesini belirleme adımında kullanılmıştır. İlgili domaini en iyi temsil eden örnekleri belirlemek için K-En Yakın Komşuluk (KYK)

algortiması kullanılmıştır. Aktif öğrenmenin ikinci adımı olan yeni etiketlenecek metinlerin belirlenmesinde duygu sınıfı en belirsiz örnekler seçilerek kullanılmış, yeni etiketlenecek verilerle model yeniden eğitilmiştir. Hedef domaindeki metinlerin etiketlenmesi için yeni etiketlenecek metinlerin belirlenmesi ve modelin eğitilmesi iteratif olarak tekrarlanmıştır. Böylece farklı domainler arası temsil yeteneği yüksek az miktardaki etiketli veri transfer edilerek sınıflandırma yapılabilmiştir. Çalışmada film, kitap, mutfak, elektronik ve DVD olmak üzere beş farklı domain kullanılmıştır. Domainler kullanıcıların bu ürün grupları için yaptıkları yorumlardan oluşmaktadır. Aktif öğrenmede sınıflandırma başarısı başlangıç kümesine oldukça bağlıdır. Hedef ve kaynak dominlerin birbirine benzer datasetler olması sınıflandırma başarısını artırır. Çalışmamızda da en yüksek sınıflandırma başarısı film yorumları dataseti kaynak domain iken kitap yorumları dataseti hedef domain olması durumunda görülmüştür. Bu iki domain arası duygu analizinde ortalama %8 sınıflandırma başarısı artışı olmuştur [47]. Sınıflandırma için Lojistik Regresyon, Destek Vektör Makinesi ve Yapay Sinir Ağları kullanılmıştır.

Tez metni, genel itibarıyla literatür çalışmaları araştırması, materyal ve metot, deneysel sonuçlar ve tartışma ile sonuç olmak üzere dört kısımdan oluşmaktadır. Birinci bölüm “Giriş” olup burada çalışmanın kısa özeti ile birlikte literatür araştırması yapılmıştır. İkinci bölümde bu çalışmada kullanılan metotlar tanıtılmıştır. Üçüncü bölümde deneysel sonuçlar verilmiş, sonuçlar tartışılmıştır.

Deneysel çalışmaların nihai sonuçlarının açıklandığı son bölümde, deneysel çalışmalar sonucu elde edilen bulgular, deneysel çalışmanın amacına uygun bir biçimde yorumlanarak sonuçlandırılmıştır.

BÖLÜM 2

MATERYAL VE METODLAR

2.1 DUYGU ANALİZİ

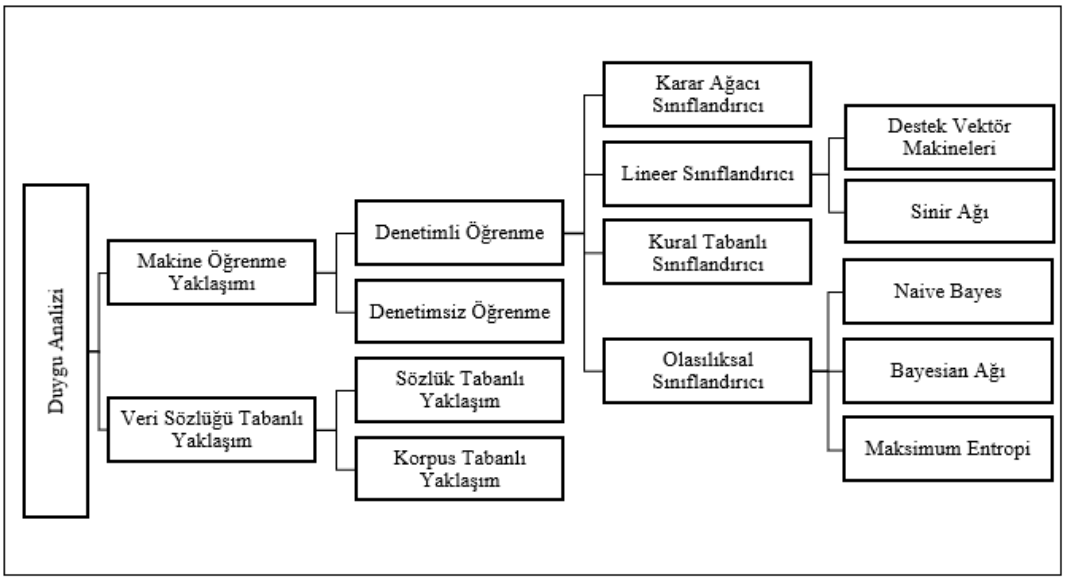
Duygu analizinde metnin işlenerek içerdiği duygusal sınıf ortaya çıkarılmaktadır. [48]. Metindeki duygusal sınıflar olumlu (pozitif), olumsuz (negatif) ya da nötr olabilmektedir.

Eşitlik 2.1'de duygu analizinde metindeki duygu tanımı için kullanılan gösterim yer almaktadır [11].

$$(e_i, a_{ij}, S_{ijk\alpha}, h_k, t_\alpha) \tag{2.1}$$

Metindeki i indeksli varlık e_i ile gösterilir, bu varlığın j indeksli özelliğini ifade etmek için a_{ij} gösterimi kullanılmaktadır. $S_{ijk\alpha}$ ise e_i varlığın a_{ij} özelliğine karşı duygu sınıfını temsil eder. Duygunun sahibi h_k ile duygunun zamanı ise t_α ile gösterilir. Örneğin "Cep telefonunun kamerasını beğendim" cümlesinde varlık cep telefonu, varlığın özelliği telefonun kamerası, duygu sınıfı pozitiftir. Cümlede duygu sahibi cümleyi kurandır. Zaman bilgisi içermemektedir. Duygu Analizi çalışmalarında beş parametrenin tamamı kullanılmak zorunda değildir. Çalışmanın ihtiyacına göre istenen parametrelerle kullanılmaktadır.

Duygu analizinde Makine Öğrenme ve Veri Sözlüğü Tabanlı Yaklaşım olmak üzere 2 kategoride incelenmektedir [18]. Şekil 2.1'de duygu analizi çalışma grupları verilmektedir.



Şekil 2.1. Duygu analizi kategorileri.

2.2 VERİ SETİNİN İNCELENMESİ

Türkçe veri setinden oluşan korpusun oluşturulması çalışmanın ilk adımı olmuştur. Demirtas ve Pechenizkiy paylaştıkları veri setindeki etiketlenmiş verilere HepsiBurada alışveriş sitesi kullanıcı yorumlarından ekleme yapılarak veri seti oluşturulmuştur [49]. Veri setinde film, kitap, DVD, mutfak ve elektronik olmak üzere beş domainden oluşan bu ürünlerin kullanıcı yorumları bulunmaktadır.

Veri setindeki metinlerin dağılımı Çizelge 2 .1'de görülmektedir.

Çizelge 2.1. Veri seti dağılımı.

		Film Domaini	**Kitap Domaini**	**DVD Domaini**	**Mutfak Domaini**	**Elektronik Domaini**
Duygu Sınıfı	Pozitif etiketli yorum adedi	5328	700	1048	1667	1774
	Negatif etiketli yorum adedi	5328	700	823	1522	1324

Domainlerden seçilmiş bazı kullanıcı yorumları Çizelge 2.2'de görülmektedir.

Çizelge 2.2. Domainlerdeki bazı kullanıcı yorumları.

Domain	**Duygu Sınıfı**	**Yorum**
Film	Olumlu	Yönettiği ikinci filmiyle en iyi yönetmen dalında Oscar almış olan Mel Gibson harika bir ise imza atmış. Yönetmenliğinin dışında oyunculuğuna da söylenecek kötü bir laf olduğunu da sanmıyorum, mükemmel bir film.
Film	Olumsuz	Beklentimin altında kalması beni hayal kırıklığına uğrattı. Oysaki sabırsızlıkla beklemiştim, bir fantastik film hastası olarak. İçi doldurulamamış bir şeyler var. Robert De Niro sırtlamış filmi.
Kitap	Olumlu	Kesinlikle okunası kitap. Hatta bu zamana kadar Ahmet Ümit kitabı okumamış olan kesinlikle gidip yazarın tüm kitaplarını almak isteyecektir. Özellikle patasananın anlatımları mükemmel. Şiddetle tavsiye ederim.
Kitap	Olumsuz	Bence Paulo bu sefer çuvallamış. Okurken hiç zevk almadım. Tavsiye etmem. Kitabın kapağı çok çekici ama içi boş. Tabi okurken kapağını görmüyorsunuz.
Mutfak	Olumlu	Özellikle fırından çok memnun kaldım, çok güzel pişiriyor set üstü ocak da çok güzel, davlumbaz da iyi diyebilirim gönül rahatlığı ile alabilirsiniz.
Mutfak	Olumsuz	Ürünü alalı 3 ay kadar oldu bazı eksikleri var: öncelikle demlik çay sızdırıyor. İkincisi derece ayarı yok. Kaynatırken çok ses yapıyor. Yıkaması zor.
Elektronik	Olumlu	Ürün gerçekten fiyatını hak eden bir cihaz. Kesinlikle hem tizlerine hem basslarına hayran kaldım. Ayrıca dışarıdan hiç ses almıyor. Tavsiye ediyorum!
Elektronik	Olumsuz	Kullandığım toplam süre 2 saat izlediğim bir film cızırtı yapmaya başladı çok bir şey beklemeyin ayrıca alırsanız bas vurguları minimal seviye alın.

2.3 METİNLERİN ÖN İŞLENMESİ

Korpusun elde edilmesi sonrası metinlerin duygu analizi çalışması yapılmasından önce ön işleme adımı uygulanmıştır. Böylece metinlerin daha doğru duygu analizi yapılması hedeflenmiştir. Ön işlemede aşağıdaki adımlar uygulanmıştır.

1. Metinlerdeki büyük harflerin küçüklere çevrilmesi
2. Metinlerdeki noktalama işaretleri, parantez, tırnak işareti gibi alfabetik olmayan karakterlerin silinmesi
3. Metinlerdeki rakamların silinmesi
4. Metinlerden Türkçe' deki edat ve bağlaçların silinmesi

Ön işleme adımı için Python kod bloğu ekran görüntüsü Şekil 2.2'de görülmektedir.

```
f = open('Negative_DVD.txt', 'r', encoding='utf8')
text = f.read()
t_list = text.split('\n')

corpus = []
corpus_all = []
listToStr =""

stopWords = set(stopwords.words('turkish'))
punctuation = ['.', '!', ',', ':', ';','...','..','(',')','\\','\?']
numeric = ['0', '1', '2', '3', '4', '5', '6', '7', '8','9']

for cumle in t_list:
    cumle.strip()
    kelimeler = word_tokenize(cumle)
    resultwords  = [word for word in kelimeler if word.lower() not in stopWords]
    resultwords  = [word for word in resultwords if word.lower() not in punctuation]
    resultwords  = [word for word in resultwords if word.lower() not in numeric]
    result = ' '.join(resultwords)
    corpus.append(result)

for cumle in corpus:
    cumle = cumle.replace('.', '')
    cumle = cumle.replace(',', '')
    cumle = cumle.replace('!', '')
    cumle = cumle.replace('..', '')
    cumle = cumle.replace('...', '')
    cumle = cumle.replace('"', "")
    cumle = cumle.replace('?', '')
    cumle = cumle.replace('(', '')
    cumle = cumle.replace(')', '')
    cumle = cumle.replace('0', '')
    cumle = cumle.replace('1', '')
    cumle = cumle.replace('2', '')
    cumle = cumle.replace('3', '')
    cumle = cumle.replace('4', '')
    cumle = cumle.replace('5', '')
    cumle = cumle.replace('6', '')
    cumle = cumle.replace('7', '')
    cumle = cumle.replace('8', '')
    cumle = cumle.replace('9', '')
    corpus_all.append(cumle)
```

Şekil 2.2 Ön işleme adımı için Python kod bloğu.

Ön işleme adımı öncesi ve sonrası korpustan seçilmiş bazı kullanıcı yorumu örnekleri Çizelge 2.3' de görülmektedir.

Çizelge 2.3. Ön işleme öncesi ve sonrası bazı kullanıcı yorumları.

Domain	**Ön İşleme Öncesi Metin**	**Ön İşleme Sonrası Metin**
Film	Filmde aranan her türlü özellik bulunmakta; görsel efekt, senaryo, oyunculuk hepsi mükemmel...	filmde aranan her türlü özellik bulunmakta görsel efekt senaryo oyunculuk hepsi mükemmel
Kitap	Kitabı sırf merakımdan ve bestseller olduğu için okudum. Şunu söylemeliyim ki övüldüğü kadar abartıldığı kadar yok. Kitabı incelemek gerekirse küçük arı ile Sarah ve diğer kahramanlar arasındaki diyaloglar çok iyi. Kitaptaki olaylar akıcı ve sürükleyici dili sade okurken sayfanın nasıl akıp gittiğini anlamıyorsunuz.	kitabı sırf merakımdan bestseller olduğu okudum şunu söylemeliyim övüldüğü abartıldığı yok kitabı incelemek gerekirse küçük arı Sarah diğer kahramanlar arasındaki diyaloglar çok iyi kitaptaki olaylar akıcı sürükleyici dili sade okurken sayfanın nasıl akıp gittiğini anlamıyorsunuz
Mutfak	Eco modu 3 saatte yıkıyor, en kısa programı 80 dakika ama onda da çok iyi yıkıyor. Sessiz çalışıyor. Bu makineler genelde 1 ile 2 saat arasında yıkama yapıyor alınabilir bir ürün.	eco modu saatte yıkıyor en kısa programı dakika onda çok iyi yıkıyor sessiz çalışıyor bu makineler genelde saat arasında yıkama yapıyor alınabilir bir ürün
Elektronik	Ürün fiyat olarak, performans olarak çok iyi bir fiyat/performans oranı veriyor. Şu ana kadar hiçbir sorun çıkmadı. Kafamdaki tek soru işareti ütüde çok plastik kullanılmış olması. onun haricinde iyi bir ütü herkese öneririm.	ürün fiyat olarak performans olarak çok iyi bir fiyat performans oranı veriyor şu ana hiçbir sorun çıkmadı kafamdaki tek soru işareti ütüde çok plastik kullanılmış olması onun haricinde iyi bir ütü herkese öneririm.
DVD	Filmde asıl konu (veya benim beklediğim konu) ikinci plana itilmiş. Yine de bilim kurgu meraklıları çok sıkılmazlar diye düşünüyorum.	filmde asıl konu benim beklediğim konu ikinci plana itilmiş yine bilim kurgu meraklıları çok sıkılmazlar diye düşünüyorum

2.4 METİN TEMSİL YÖNTEMLERİ

Doğal Dil İşleme çalışmalarında ilk adım metinleri sayısallaştırılmak ve makine öğrenme yaklaşımlarında kullanılmak üzere vektörize etmektir. Metinlerin sayısallaştırılması için pek çok yaklaşım bulunmaktadır. Bu bölümde çalışmada

kullanılan Kelime Çantası Modeli (Bag of Word model), Terim Frekans Ters Doküman Frekansı (Term Frequency Inverse Document Frequency) (TF-IDF) Model, N-Gram Model ve Word2Vec yaklaşımı Skip-Gram Modeli teorik altyapısı ve çalışmada kullanım amaçları anlatılmıştır.

2.4.1 Kelime Çantası Modeli

Kelime çantası modeli metinlerin vektörel gösterim yöntemlerindendir. Çalışmada beş domainin her biri farklı vektörel gösterimlerle temsil edilip sınıflandırma algoritmaları ile sınıflandırılarak etkili ve performanslı temsil yöntemleri kıyaslanmıştır. Metinlerin temsil edildiği vektör boyutunun korpustaki birbirinden farklı kelime sayısı kadar olması kelime çantası modelinin dezavantajıdır.

Kelime çantası model yaklaşımında üç adım bulunmaktadır. İlk adım dokümanların toplanması adımıdır. İkinci adım dokümanlardaki farklı kelimelerin tamamını içeren korpus oluşturulması adımıdır. Son adımda ise vektörize edilmek istenen doküman ya da metin bu korpusta bulunan kelimeleri içerip içermediğine göre 1 ya da 0 değeri almaktadır [50].

Aşağıdaki örnekte kelime çantası modeli incelenmiştir. Charles Dickens'ın İki Şehrin Hikayesi kitabından alınmış örnek metin dokümanı şöyledir:

- It was the best of times,
- it was the worst of times,
- it was the age of wisdom,
- it was the age of foolishness

Dokümandan korpusu oluşturacak her bir farklı kelime elde edildiğinde aşağıdaki sözlük oluşturulmaktadır.

- "it"
- "was"
- "the"
- "best"
- "of"
- "times"
- "worst"

- “age”
- “wisdom”
- “foolishness”

Her bir metin vektörize edildiğinde aşağıdaki vektörler elde edilmektedir.

- "it was the worst of times" = [1, 1, 1, 0, 1, 1, 1, 0, 0, 0]
- "it was the age of wisdom" = [1, 1, 1, 0, 1, 0, 0, 1, 1, 0]
- "it was the age of foolishness" = [1, 1, 1, 0, 1, 0, 0, 1, 0, 1]

2.4.2 Terim Frekans Ters Doküman Frekansı Modeli

Terim Frekans Ters Doküman Frekansı (Term Frequency Inverse Document Frequency) (TF-IDF) metinlerin vektörize temsil yöntemlerindendir. Çalışmada film, kitap, DVD, mutfak ve elektronik ürün yorumları domainlerinin her birini en başarılı temsil eden yöntem ve sınıflandırıcının öğrenilmesi için domainler TF-IDF yönetimine göre de temsil edilmiş ve ölçümler yapılmıştır.

TF-IDF yaklaşımında metindeki kelime frekansı ile o kelimenin tüm korpustaki frekans logaritması çarpımı kullanılmaktadır. Yaklaşım; bir kelimenin tüm dokümandaki önemi, ayırt edici vurgu değeri hesaplaması olarak yorumlanmaktadır [51].

Eşitlik 2.2’de terim frekansı (Term Frequency) hesaplanması formülü yer almaktadır. n değeri kelimenin dokümandaki frekansı, total değeri dokümandaki toplam kelime sayısıdır. Yani her kelime için o dokümanda geçme sıklığı hesaplanır.

$$tf = n/total \tag{2.2}$$

Eşitlik 2.3’de ters doküman frekansı (Inverse Document Frequency) hesaplanması formülü yer almaktadır. C toplam döküman sayısı, n_i i. kelimenin içinde geçtiği doküman sayısıdır. Her dokümanda geçen kelimelerde idf değeri küçüktür. Özellikle bağlaç ve edatlar idf değerinin küçük hesaplandığı kelime gruplarıdır.

$$idf = log_2(C/n_i) \tag{2.3}$$

Eşitlik 2.4'de terim frekans ters doküman frekansı hesaplanma formülü yer almaktadır. Her kelime için için hesaplanan Term Frequency ve Inverse Document Frequency değerleri çarpımı ile normalizasyon yapılmış olur.

$$tf\text{-}idf = tf^{*}idf \tag{2.4}$$

2.4.3 N-Gram Modeli

N-Gram Model metinlerin vektörel temsil yöntemlerindendir. Çalışmada kullanılan beş domain için, domainleri kendi içinde en iyi temsil eden yöntemi ve sınıflandırıcıyı belirlemek amacıyla Kelime Çantası Modeli, TF-IDF ve Word2Vec ile birlikte N-Gram yöntemi de kullanılmıştır. Kelime öbeklerine göre temsil vektörlerin hesaplandığı N-Gram Model için komşuluk değeri 3 olarak kullanılmıştır.

N-Gram yaklaşımda kelimelerin birbirine komşulukları göz önünde bulundurularak terim sıklıkları ve olasılıkları hesaplamaları yapılmaktadır [52]. Komşuluk değerleri olarak en sık kullanılan modeller; unigram, bigram ve trigram modellerdir.

Unigram modelde her kelime için kelime sıklığı hesaplanırken, bigram modelde kelime ve bir komşusu için kelime sıklığı hesaplanmaktadır; trigram modelde kelime ve iki komşusu için bu hesaplama yapılmaktadır. N değeri kelime öbeği anlamı taşımaktadır. Kelime ya da kelime öbeklerinin ardışık gelme ihtimalleri Markov hipotezine göre hesaplanmaktadır. Eşitlik 2.5'te hipotezde kullanılan formül görülmektedir.

$$\begin{aligned} P(w_{1:n}) &= P(w_1)P(w_2|w_1)P(w_3|w_{1:2}) \dots P(w_n|w_{1:n-1}) \\ &= \prod_{k=1}^{n} P(w_k|w_{1:k-1}) \end{aligned} \tag{2.5}$$

$P(w_1)$ 1 numaralı kelimenin korpusta bulunma olasılığıdır. $P(w_2|w_1)$ 1 numaralı seçilmiş kelimenin 2 numaralı kelime ile olan komşuluğuna göre korpusta bulunma olasılıklarıdır. $P(w_3|w_{1:2})$ 1 ve 2 numaralı kelimelerin komşu olduğu durumlarda 3 numaralı kelimenin bu kelime grubuna korpusta komşu olma olasılıklarıdır. $P(w_n|w_{1:n-1})$ 1 numaralı kelimeden n-1 numaralı kelimeye kadar olan kelimelerin

komşu oldukları durumlarda n numaralı kelimenin bu kelime gruplarına korpusta komşu olma olasılığıdır. Her ayrık olayın olasılıkları ayrı ayrı hesaplanıp, çarpılır.

Koşullu olasılıkların çarpımı ile kelimelerin ardışık gelme ihtimalleri hesaplanmaktadır. Yukarıdaki eşitliğe benzer şekilde eşitlik 2.6'te bigram model için kullanılan olasılık hesaplama formülü görülmektedir.

$$P(w_{1:n}) = P(w_n|w_{1:n-1})P(w_n|w_{n-1}) \quad (2.6)$$

Bu yaklaşım ile N=2 için örneğin "Merkez Bankası" kelime öbeğinde "Merkez" kelimesinden sonra korpusta "Bankası" kelimesinin gelme ihtimali hesaplanmış olur.

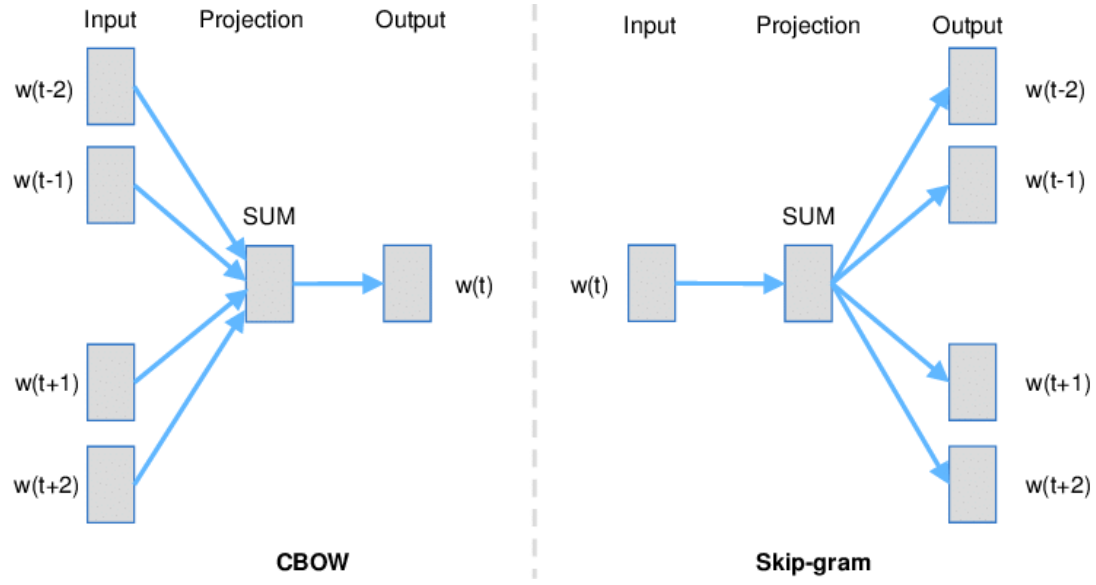
2.4.4 WORD2VEC Modeli

Word2Vec yaklaşımı; Kelime Çantası Modeli, TF-IDF Model ve N-Gram Model gibi metinlerin temsil yöntemlerinden biridir. Yukarıdaki başlıklarda anlatılan temsil yöntemleri ile birlikte Word2Vec temsil ile de gösterilen beş domain için sınıflandırma başarıları ölçülerek domainler arası duygu analizinde hangi temsil yöntemine göre domainlerin temsil edileceğine karar verilmiştir. Word2Vec yaklaşımında kelime ya da metinin temsil edileceği vektör boyutu kullanıcı tarafından belirlenir. Çalışmada 50, 100, 150, 200 ve 250'lik vektör boyutları ile metinler temsil edilmiş ve her boyut için sınıflandırma başarıları ölçülmüştür.

Word2Vec yaklaşımı ile korpustaki her kelime ayrık vektörize temsil edilmektedir. CBOW (Continous Bag of Words) ve Skip-Gram modelleri olmak üzere iki alt yaklaşım bulunmaktadır. Her iki yaklaşımda da de metin üzerinde gezinilen bir pencere vardır. Pencere boyutu N ise N kelime komşuluğundaki kelimeler göz önünde bulundurulur. Tomas Mikolov ve ekibi tarafından 2013 yılında geliştirilmiştir.

CBOW ve Skip-Gram modelleri birbirinin tersi modellerdir. CBOW modelde pencerenin merkezinde olmayan kelimelere göre merkezdeki kelime tahmin edilmeye çalışılırken, Skip-Gram modelde tam tersi olarak pencerenin merkezindeki kelimeye göre merkezinde olmayan kelimeler tahmin edilmeye çalışılmaktadır. Böylece

kelimelerin birbirleri ile ilişkileri, benzerlikleri ortaya çıkmış olur [53]. Şekil 2.3' de bu modeller gösterilmektedir.



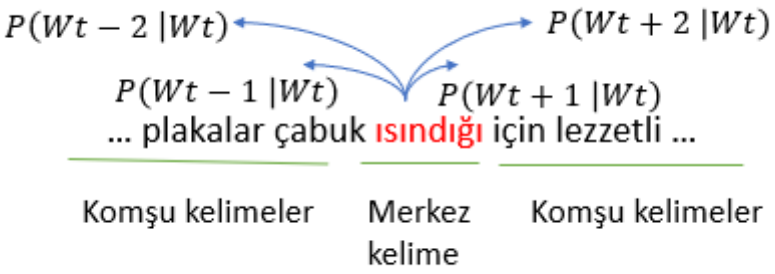
Şekil 2.3. CBOW ve Skip-Gram model yaklaşımları [53].

CBOW ile Skip-Gram modellerinin karşılaştırması Çizelge 2.4' de yer almaktadır. [54]

Çizelge 2.4. CBOW ve Skip-Gram modelleri karşılaştırması.

Numara	**CBOW Modeli**	**Skip-Gram Modeli**
1	Eğitim daha hızlıdır.	Eğitim CBOW'a göre daha yavaştır.
2	Sık kullanılan kelimeleri daha iyi temsil eder.	Seyrek kullanılan kelimeleri daha iyi temsil eder.
3	Seyrek kullanılan kelimelerin temsili için daha büyük korpusa ihtiyaç duyar.	CBOW'a göre daha küçük korpuslarda bile iyi tahminlerde bulunabilir.

Word2Vec yaklaşımında Skip-Gram modelinde pencere boyutu m için j. kelimenin m kadar sağı ve solundaki kelimelerin yüksek ihtimalle tahmin edilmesi hedefi bulunmaktadır. Hem kelime hem komşu kelimeler vektörel olarak sembolize edilmektedir. Şekil 2.4' de örnek bir metin üzerinden 2 komşuluk değeri için yapılan hesaplamalar gösterilmektedir. Örneğin merkez kelimenin "ısındığı" olduğu metinde $P(W_{t+2} \mid W_t)$ değeri, merkez kelimenin 2 kelime kadar sağında olan "lezzetli" kelimesi için korpusta bulunma olasılığının hesaplanması ile elde edilir.



Şekil 2.4 Skip-Gram modelinde merkeze göre komşuluk olasılıkları hesaplanması.

Korpusun tamamı için her kelimede bu hesaplamalar yapılmaktadır. Yapılan hesaplama ile koşullu olasılık dağılımları elde edilmektedir. Olasılık dağılımda j. kelimenin m kadar sağı ve solundaki komşuları en yüksek ihtimalle doğru tahmin edilmesinde parametrik değerler pencere boyutunu gösteren m değeri ve kelimenin vektörel olarak gösterilmesinde kullanılan vektör boyutudur. Eşitlik 2.7'da bahsedilen yaklaşımın formülü görülmektedir.

$$\frac{-1}{T}\sum_{t=1}^{T}\sum_{-m<j<m} logP(Wt+j\,|Wt) \tag{2.7}$$

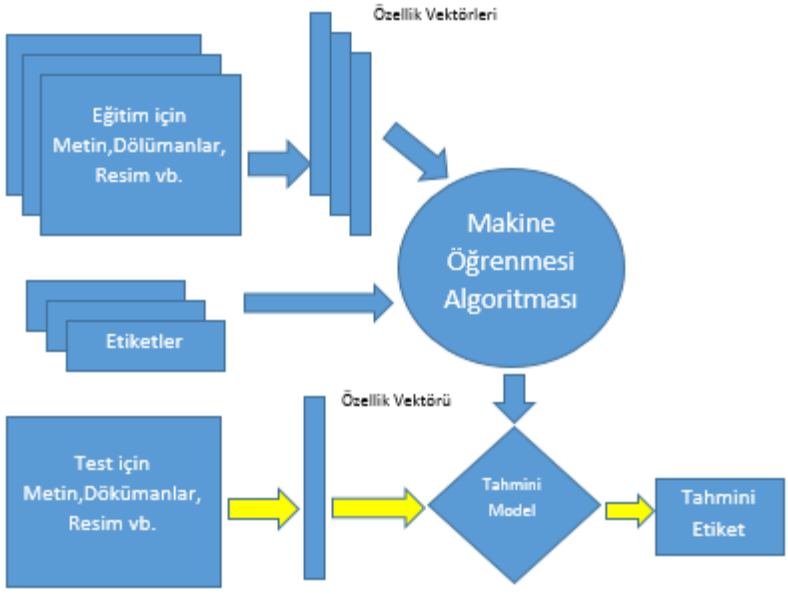
Formüldeki T değeri kelimenin pozisyonunu, $P(Wt+j\,|Wt)$ değeri t pozisyonudaki kelime için j kadar komşulukta bulunan kelimenin korpusta bulunma olasılığını göstermektedir. Eşitliğin en yüksek değerler ile hesaplanabilmesi hedefi vardır. Böylece merkez kelimeye göre en yüksek ihtimalli komşu olabilecek kelimeler hesaplanır.

Kelime vektörleri temsilinde kelimenin cümlenin başında ya da ortasında yer almasına göre farklı değerleri olur. Bu değerlerin optimize edilmesinde kullanılan formül eşitlik 2.8'de görülmektedir. $u_o$vektörü pencere içindeki indeks değerini, c orta merkezli kelime indeksini, v orta kelime vektörünü ifade etmektedir. Olasılık hesabında kelimenin pozisyonu göz önünde bulundurulmaktadır.

$$P(Wt+j\,|Wt) = \frac{\exp(u_o \mathrm{Tvc})}{\sum_{w=1}^{v}\exp(u_o \mathrm{Tvc})} \tag{2.8}$$

2.5 METİN SINIFLANDIRMA YÖNTEMLERİ

Metin sınıflandırma, her bir belgenin önceden belirlenmiş sınıf kümesindeki sınıflardan hangisine dahil olduğunun belirlenmesidir [55]. Lojistik Regresyon, Naive Bayes, Yapay Sinir Ağları gibi pek çok makine öğrenme yaklaşımı metin sınıflandırma için kullanılmaktadır. Metin sınıfını belirleme için kullanılan makine öğrenme yöntemlerinden belge için tahmin edilen sınıf ile belgenin gerçek sınıfı değerlerine göre üretilen modelin doğru sınıflandırabilme metrikleri hesaplanır. Hesaplamalarda belgelerin tamamı eğitim ve test olmak üzere iki gruba ayrılmaktadır. Eğitim kümesindeki belgeler kullanılarak makine öğrenmesi modelindeki parametrelerin optimum değerlerine getirilmesi hedeflenmektedir. Yeteri kadar çok eğitim belgesi ile eğitilen model, test grubu ile sınıflandırma başarısı test edildiğinde daha yüksek başarılar elde edecektir [56]. Şekil 2.5'te makine öğrenmesi ile metin sınıflandırma akışı görülmektedir.



Şekil 2.5 Makine öğrenmesi ile metin sınıflandırma akışı [57].

2.5.1 Lojistik Regresyon Algoritması

Lojistik Regresyon makine öğrenme çalışmalarında kullanılan denetimli öğrenme algoritmalarından biridir. Beş domainin her biri için en başarılı sınıflandırma

yöntemlerini belirlemek adına Naïve Bayes ve Rastgele Orman yaklaşımları ile birlikte duygu analizi için kullanılmış, başarıları ölçülmüştür.

Lojistik Regresyon yaklaşımında bağımsız değişkenler ile bağımlı değişkenler arasında ilişki istatistik olarak hesaplanmaktadır. Bağımlı değişkenlerin sürekli olması halinde doğrusal regresyon modeli kullanılırken bağımlı değişkenler atomik değerlere sahip ise lojistik regresyon modeli kullanılmaktadır. İstatistik modelde bağımsız değişkenlerin ikili sınıfa olan bağımlılığının hesaplanma olasılığından bahsedilmektedir. [58]

Lojistik regresyonda, bağımlı değişkenlerin varlığının olasılığını hesaplama için logit dönüşümünü kullanmaktadır. Eşitlik 2.9'de dönüşüm formülü görülmektedir. X_k bağımsız değişkeninin katsayısı olan b_k değeri, bağımsız değişkenin y bağımlı değişkeni olan etiket sınıfının üzerine yaptığı etki olarak yorumlanmaktadır.

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \cdots b_k X_k \tag{2.9}$$

Eşitlikteki p değeri, karakteristik ya da bağımlı özelliğinin var olma olasılığıdır. Yani bir bağımlı değişkene ait kategorinin olma olasılığı ile olmama olasılığı arasındaki ilişkiye göre katsayılar yorumlanmaktadır. Olasılık hesaplanması için Eşitlik 2.10'daki formül kullanılmaktadır.

$$logit(p) = \ln\left(\frac{p}{1-p}\right) \tag{2.10}$$

2.5.2 Naive Bayes Algoritması

Naïve Bayes, makine öğrenmesi ve yapay zeka çalışmalarında kullanılan olasılıksal hesaplama tabanlı denetimli öğrenme yaklaşımlarından biridir. Çalışmada film, kitap, DVD, mutfak ve elektronik olmak üzere her bir domainin ayrı ayrı sınıflandırma başarılarını farklı temsil yöntemleri ile birlikte ölçmek için kullanılmıştır.

Naïve Bayes algoritması koşullu olasılık hesaplamasına dayanan sınıflandırma yaklaşımıdır. Bu yaklaşıma "sınıf koşullu bağımsızlık" ismi de verilmektedir. Modelin üretilmesi için girdi olarak sağlanan öznitelik vektörlerindeki her bir özniteliğin sınıfın belirlenmesindeki bağımsız olasılıkları göz önünde bulundurulmaktadır [59]. Örneğin golf oyunu oynanıp oynanmamasına karar verilen modelde öznitelikler hava durumu, sıcaklık, nem oranı ve rüzgarın durumu ise hava durumunun iyi olması, sıcaklık ya da nem oranından daha önemli bir etkiye sahip olmayacaktır. Her bir özniteliğin sınıfın belirlenmesindeki önemi bağımsız olarak değerlendirilmektedir. Eşitlik 2.11'da Bayes eşitliği yer almaktadır.

$$P(y\,|X) = \frac{P(X\,|y).P(y)}{P(X)} \tag{2.11}$$

Eşitlikteki y değeri sınıfı, X değerleri öznitelikleri temsil etmektedir. Örneğin hava durumunun güneşli olması durumunda golf oynanma olasılığı hesaplanırken, golf oynandığında hava durumunun güneşli olma olasılığı, havanın güneşli olma olasılığı ve golf oynanma olasılığı göz önünde bulundurulmaktadır. Eşitlik 2.12'de öznitelik vektörleri formülü görülmektedir.

$$X = (\,x_1, x_2, x_3, \dots, x_n) \tag{2.12}$$

Eşitlikteki x_1, x_2 kolonları hava durumu, sıcaklık gibi özniteliklerin değerlerini içermektedir. Eşitlik 2.13 ve 2.14'te her bir öznitelik için sınıf koşullu bağımsızlık değerleri hesaplaması görülmektedir.

$$P(y\,|x_1, \dots, x_n) = \frac{P(x_1\,|y).P(x_2\,|y)\dots P(x_n\,|y)P(y)}{P(x_1).P(x_2)\dots P(x_n)} \tag{2.13}$$

$P(y\,|x_1, \dots, x_n)$ özniteliklerin $x_1, \dots, x_n$ değerleri içermesi durumunda verinin y sınıfına ait olma olasılığıdır. $P(x_1\,|y)$ y sınıfına ait örneklerin korpusta x_1 öznitelik değeri içerme olasılığıdır. Benzer şekilde $P(x_2\,|y)$ y sınıfına ait örneklerin korpusta x_2 öznitelik değeri içerme olasılığıdır. $P(y)$ örneğin y sınıfına ait olma olasılığıdır. Her öznitelik için payda

hesaplaması aynı olduğu için Eşitlik 2.14' deki gibi payda değeri çıkarılıp orantılık kullanılabilir.

$$P(y\,|x_1,\ldots,x_n) \propto P(y)\prod\prod_{i=1}^{n} P(x_i\,|y) \tag{2.14}$$

Sınıflandırıcı eğer ikiden fazla sınıf için kullanılacaksa y sınıfı maksimum olasılık fonksiyonu ile bulunmaktadır. Eşitlik 2.15'te fonksiyon görülmektedir.

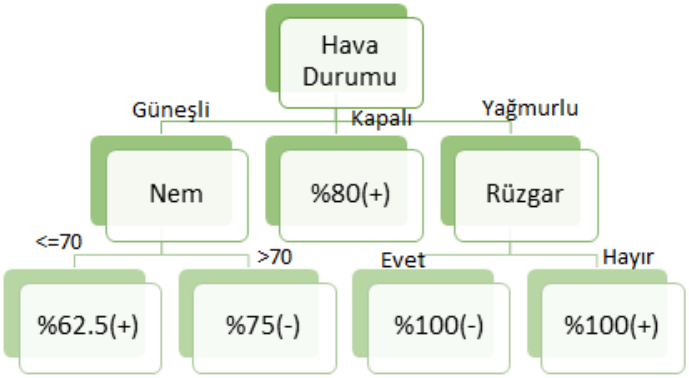
$$y = argmax_y\, P(y)\prod\prod_{i=1}^{n} P(x_i\,|y) \tag{2.15}$$

2.5.3 Rastgele Orman Algoritması

Rastgele Orman algoritması, karar destek ağacı temelli denetimli öğrenme yaklaşımlarındandır. Çalışmada domainler arası duygu analizinde kullanılmak üzere seçilecek algoritmaya karar vermek için Naïve Bayes ve Lojistik Regresyon algoritmaları ile birlikte Rastgele Orman algoritması her bir domaini ayrı ayrı sınıflandırmak için kullanılmıştır.

Rastgele Orman algoritmasında modelin üretilmesi için girdi olarak sağlanan eğitim veri seti rastgele dağılımlara bölünmektedir. Her bir dağılım birbirinden bağımsızdır ve her dağılımdan üretilen karar ağacı bulunmaktadır. Veri setinden sınıfın belirlemesinde tek karar ağacının kullanması yerine birden çok karar ağacının kararlarının birleştirilmesi söz konusudur. [60]

Karar ağaçları kök, dallar, düğümler ve yapraklardan oluşan ağaç yapısındaki sınıflandırıcılardır. Şekil 2.6'da örnek bir karar ağacı çalışma prensibi görülmektedir.



Şekil 2.6 Karar ağacı çalışma prensibi.

Örnek çalışma prensibinde hava durumu, nem ve rüzgar özniteliklerinin olduğu ve bu özniteliklere göre oyun oynanıp oynanmayacağına karar verilen model incelenmiştir. Modele göre hava durumu kapalı ise %80 ihtimalle oyun oynanabileceği, hava yağmurlu ve rüzgar da var ise %100 ihtimalle oyun oynanmayacağı görülmektedir.

Karar Ağacı oluşturulmasında ilk adım kökün belirlenmesidir. Kök öznitelikler arasında sınıfı belirlemede en etkili özniteliktir. Kök öznitelik seçiminde Eşitlik 2.16'de görülen bilgi kazanımı formülüne göre önce eğitim veri setindeki bilgi kazanımı hesaplanır. p_i verilerdeki i. üyenin sınıflardan her birine ait olma olasığıdır.

$$Entropi(D) = -\sum_{i=1}^{m} p_i \, log_2 p_i \tag{2.16}$$

Her bir öznitelik için ayrı ayrı bilgi kazanımları hesabı Eşitlik 2.17'da görülmektedir. $Entropi(D_j)$ her bir özniteliğin değerlerinin j. sınıfa ait olma olasılıklarıdır. Örneğin hava durumu özniteliği için; hava durumunun güneşli olması durumunda oyun oynanma olasılığı, hava durumunun kapalı olması durumunda oyun oynanma olasılığı ya da yağmurlu olması durumundaki oyun oynanma olasılıkları ayrı ayrı hesaplanmaktadır.

$$Entropi_A(D) = \sum_{i=1}^{v} \frac{|D_j|}{|D|} \, Entropi(D_j) \tag{2.17}$$

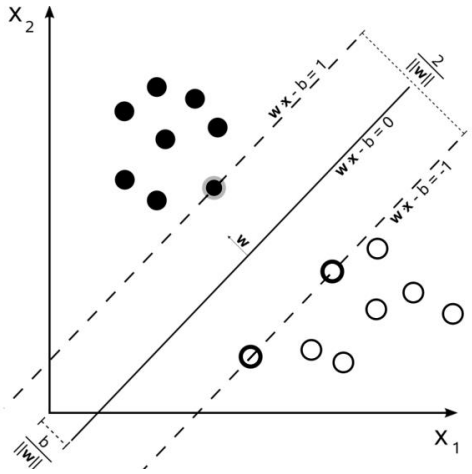
Veri setinin tamamı için hesaplanan bilgi kazanımından her öznitelik için ayrı hesaplanan bilgi kazanımları çıkarılarak en yüksek bilgi kazanımı sağlayacak öznitelik

belirlenmektedir. Ağaç için bu öznitelik kök olarak belirlenmektedir. Kök özniteliğin içerdiği değerlere göre dallar belirlenmektedir. Dalların bağlandığı düğümler yukarıda bahsedilen formüllerle hesaplanmış kökten bir sonraki en yüksek bilgi kazanımı sağlayan özniteliktir.

2.5.4 Destek Vektör Makinesi Algoritması

Destek Vektör Makinesi algoritması yapay zeka çalışmalarında kullanılan lineer sınıflandırıcı grubundaki sınıflandırma yaklaşımlarındandır. Çalışmada domainler arası duygu analizinde Aktif Öğrenme yaklaşımı ile birlikte duygu analizinde kullanılmıştır.

Destek Vektör Makinası algoritması, Vladimir Vapnik ve Alexey Chervonenkis tarafından geliştirilmiş Vapnik-Chervonenkis teorisine dayanan sınıflandırma yaklaşımıdır. Model etiketli verileri girdi olarak alır ve bir hiperdüzleme yerleştirmektedir. Farklı sınıflar arasında bir karar çizgisi belirlemeye çalışılmaktadır. Karar çizgisinin doğrusal olup olmamasına göre Linear SVM ya da NonLinear SVM olmak üzere iki alt kırılma sahiptir. Karar çizgisini istatiksel öğrenme ve yapısal risk minimizasyonuna göre belirlemektedir. Çizgi her iki sınıf örneklerinin en uzağında olacak şekilde hesaplamalar yapılmaktadır. [61] Şekil 2.7' de Linear SVM algoritmasındaki karar çizgisi belirlemeye ait çalışma prensibi görülmektedir.



Şekil 2.7 Linear SVM örnek karar çizgisi.

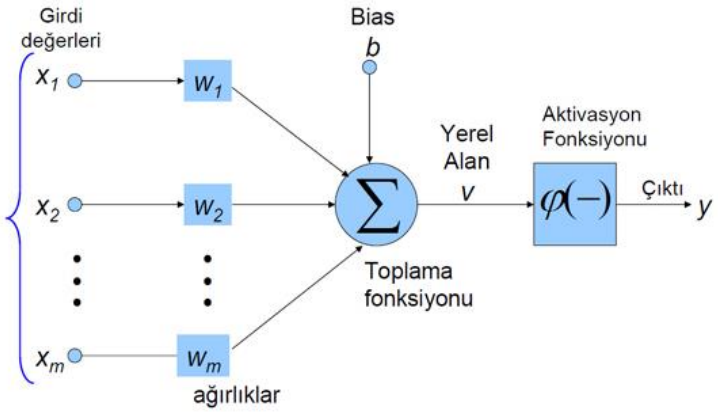
w vektörü hiperdüzlemin ağırlık vektörü, b bias değeridir. X_1 ve X_2 düzlemdeki boyutları temsil etmektedir. Hiperdüzlemdeki her renklendirilmiş ve renklendirilmemiş noktalar modele girdi olarak verilmiş iki sınıfa ait etiketli verilerdir. Etiketli veriler özniteliklerine göre temsil değerlerine sahiptir. Öznitelik vektörlerin farklı sınıflar arası karar sınırı belirlenirken hiperdüzlemdeki bu örneklerin sınıra en uzak değere sahip olması gözetilmektedir.

2.5.5 Yapay Sinir Ağları

Yapay sinir ağları, makine öğrenme çalışmalarında kullanılan lineer sınıflandırıcı temelli denetimli öğrenme yaklaşımıdır. Çalışmada Logistik Regresyon, Destek Vektör Makineleri ile birlikte domainler arası duygu analizinde kullanılmıştır. Aktif Öğrenme yaklaşımı ile birlikte kullanılmıştır.

Yapay sinir ağları, beyindeki sinir hücrelerinin çalışma prensibinden esinlenerek üretilmiş modellerdir. YSA üzerine ilk çalışmalar W.S. McCulloch ve W.A. Pitts tarafından yapılmıştır. YSA tek katmanlı, çok katmanlı olabilirler, denetimsiz ya da denetimli öğrenmede kullanılmışlardır, ileri beslemeli ya da geri beslemeli olmak üzere farklı bağlantı yapılara sahiptirler.

Tek katmanlı YSA tek bir sinir ağı içermektedir. Girdi katmanı ve çıktı katmanından oluşmaktadır. Girdi katmanındaki bileşenlerin ağrılıkları bulunmaktadır. Şekil 2.8'de tek katmanlı YSA çalışma prensibi görülmektedir. YSA'da en uygun ağırlıkların keşfedilmesi prensibine dayanarak, modelin eğitilmesi sağlanmaktadır.



Şekil 2.8 Tek katmanlı YSA çalışma prensibi.

Girdi değerleri ağırlıklarla çarpılıp elde edilen değerler toplanır ve bias değeri eklenmektedir. Eşitlik 2.18'de toplam fonksiyonu hesaplaması görülmektedir. x_i vektörü i. girdi vektörüdür. w_i i. ağırlık vektörünü temsil etmektedir. Vektörel çarpım işlemi sonrası b bias değeri eklemesi yapılır. Ağırlık vektörü ve girdi vektörü boyutu m kadardır.

$$Z = \sum_{i=1}^{m} x_i w_i + b \tag{2.18}$$

Toplam fonksiyonu yukarıdaki formülün dışında Çizelge 2.5'te görülen formüllere göre de hesaplanabilmektedir. Farklı NET hesaplama yaklaşımları vardır [62].

Çizelge 2.5. YSA toplama fonksiyonları.

Numara	**Formül**	**Açıklama**
1	$NET = \sum_{i=1}^{m} x_i w_i$	Ağırlık değerleri girdiler ile çarpılır ve bulunan değerler birbirleriyle toplanarak Net girdi hesaplanır
2	$NET = \prod_{i=1}^{m} x_i w_i$	Ağırlık değerleri girdiler ile çarpılır ve daha sonra bulunan değerler birbirleriyle çarpılarak Net Girdi Hesaplanır

3	$NET = \text{Max}(x_i * w_i)$	m adet girdi içinden ağırlıklar girdilerle çarpıldıktan sonra içlerinden en büyüğü Net girdi olarak kabul edilir
4	$NET = \text{Min}(x_i * w_i)$	m adet girdi içinden ağırlıklar girdilerle çarpıldıktan sonra içlerinden en küçüğü Net girdi olarak kabul edilir
5	$NET = \sum_{i=1}^{m} Sgn(x_i w_i)$	n adet girdi içinden girdilerle ağırlıklar çarpıldıktan sonra pozitif ile negatif olanların sayısı bulunur. Büyük olan sayı hücrenin net girdisi olarak kabul edilir
6	$NET = NET(Eski) + \sum_{i=1}^{m} (x_i w_i)$	Hücreye gelen bilgiler ağırlıklı olarak toplanır. Daha önce hücreye gelen bilgilere yeni hesaplanan girdi değerleri eklenerek hücrenin net girdisi hesaplanır.

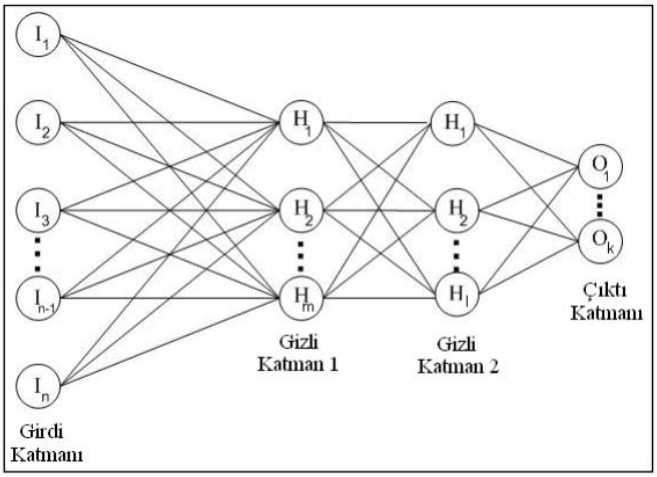
Toplam fonksiyonundan elde edilen değer aktivasyon fonksiyonundan geçirilerek çıktı elde edilmektedir. Eşitlik 2.19'da örnek bir aktivasyon fonksiyonu görülmektedir. Örnekteki fonksiyona göre veri seti iki sınıfa göre çıktı üretmektedir. Eşitlik eğer NET fonksiyonu ilgili i. örnek için pozitif sonuç üretiyorsa örnek 1 numaralı duygu sınıfına aittir ya da sonuç negatif ise örnek 0 numaralı duygu sınıfına aittir şeklinde değerlendirilmektedir.

$$\begin{aligned} &Eğer \sum_{i=1}^{m} x_i w_i + b > 0 \;\; ise, y = 1 \\ &Eğer \sum_{i=1}^{m} x_i w_i + b < 0 \;\; ise, y = 0 \end{aligned} \tag{2.19}$$

Aktivasyon fonksiyonu seçimi probleme, veri setine uygun olarak yapılmalıdır. Genellikle sigmoid fonksiyonu aktivasyon fonksiyon olarak kullanılmaktadır.

Çok katmanlı sinir ağlarında ise girdi katmanı ile çıktı katmanı arasında gizli katman bulunmaktadır. Girdi katmanında veri işlenmeden gizli katmana iletilir. Girdi katmandaki her eleman gizli katmandaki her elemana bağlıdır. Gizli katmandaki eleman sayısı değişkendir ve probleme uygun olarak değişken sayıda olabilmektedir.

Çıktı katmanında ise gizli katmandan gelen bilgiler iletilir. Şekil 2.9'da çok katmanlı YSA çalışma prensibi görülmektedir.



Şekil 2.9 Çok katmanlı YSA çalışma prensibi.

2.6 AKTİF ÖĞRENME YAKLAŞIMI

Aktif Öğrenme, az sayıda etiketli veriye sahip olunan durumlarda etkili sınıflandırma yapabilmek için geliştirilmiş bir makine öğrenme yaklaşımıdır. Çalışmada domainler arası duygu analizinde Aktif Öğrenme ve Transfer Öğrenme temelli Aktif Öğrenme yaklaşımları kullanılmış, sınıflandırma başarıları ölçülmüştür.

Aktif Öğrenme Algoritması, gerektiğinde verileri etiketlemek üzere kullanıcılara danışılan makine öğrenme yaklaşımıdır. Etiketli verilerin az olduğu durumlarda, veri etiketlemenin zaman ve maliyet gerektirmesi, bu yaklaşımın kullanımını pekiştirmektedir.

Aktif Öğrenme temelde üç yaklaşım içermektedir. [63]

- Üyelik Sorgu Sentezi Yaklaşımı: Sorgulanacak verinin kullanıcı tarafından üretilmesinden sonra etiketlenmesidir.
- Akış Tabanlı Seçmeli Örnekleme Yaklaşımı: Etiketlenmemiş verinin bilgi kazanımı değerine göre etiketlenip etiketlenmeyeceğine karar verilen yaklaşımdır.

- Havuz Tabanlı Örnekleme Yaklaşımı: Havuzdaki etiketsiz veriler arasından seçim yapılarak etiketleme yapılması yaklaşımıdır.

Havuz tabanlı örnekleme yaklaşımında kullanılacak verilerin seçimine göre çalışmalar 3 gruba ayrılmaktadır.

- Rastgele Seçim: Havuzdaki etiketsiz verilerden rastgele örneklerin seçilmektedir.
- Entropi Temelli Seçim: Etiketsiz veriler arasında bilgi kazanımı sağlayacak en yüksek entropiye ait örneklerin seçilmektedir.
- Margin Temelli Seçim: Etiketlencek herhangi bir sınıfa ait olma olasılığı diğerlerine göre çok daha yüksek olan örneklerin seçilmektedir.

Aktif Öğrenme yaklaşımı çalışma prensibi aşağıdaki gibidir.

1.) İlk örnek eğitim veri setinin seçilmektedir ve etiketlenmektedir.
2.) Kalan etiketlenmemiş eğitim setinin değerlendirilmesi etiketlenecek örneklerin nasıl seçileceğine karar verilmektedir.
3.) Seçilen örneklerin etiketlenmesi ve modelin yeniden eğitilmesi sağlanmaktadır.
4.) İkinci adıma geri dönülüp hedeflenen eşiğe kadar sürecin devam etmesi sağlanmaktadır.

2.7 K-EN YAKIN KOMŞULUK YAKLAŞIMI

K-En Yakın Komşuluk algoritması makine öğrenme ve yapay zeka çalışmalarında kullanılan kümeleme algoritmalarından biridir. Çalışmada domainler arası duygu analizi çalışması yapılırken kaynak domaini en iyi temsil eden örnekleri belirlemek için kullanılmıştır. Kitap, film, DVD, mutfak ve elektronik ürün yorumlarından oluşan her bir domain pozitif ve negatif olmak üzere iki duygu sınıfı ile etiketlenmiştir. Her domain için iki sınıf merkezi hesaplanarak merkeze en yakın örneklerin kümeyi en iyi temsil eden örnekler olması yaklaşımı ile K-En Yakın Komşuluk algoritması kullanılmıştır.

K-En Yakın Komşuluk algoritması, 1967 yılında T. M. Cover ve P. E. Hart tarafından geliştirilmiştir. Algoritmada, etiketli veri setlerindeki örneklerden yararlanılarak

etiketsiz verilerin sınıflandırılması yapılmaktadır. Etiketlenecek yeni verinin, mevcut verilere göre uzaklığı hesaplanıp, k sayıda en yakın komşuluğu bakılmaktadır. Uzaklık hesapları için farklı fonksiyonlar kullanılmaktadır. Kullanılan bazı fonksiyonlar aşağıda listelenmiştir.

- Euclidean Mesafe
- Manhattan Mesafe
- Minkowski Mesafe

Çalışmada örneklerin uzaklık hesaplaması için Euclidean Mesafe fonksiyonu kullanılmıştır. Eşitlik 2.20'de Euclidean mesafe formülü görülmektedir. Mesafe hesaplanmak istenen ilk verinin kartezyen koordinatlarının $(p_1, p_2, .. p_n)$ ile temsil edildiği, ikinci verinin koordinatlarının $(q_1, q, .. q_n)$ ile temsil edildiği düşünülmektedir. Her iki örneğin ilgili kartezyen koordinat nokta değerleri çıkarılıp kareleri alınarak toplanır, elde edilen toplam değer karekök içine alınır. Yapılan hesaplama ile iki örnek arasındaki uzaklık elde edilmektedir.

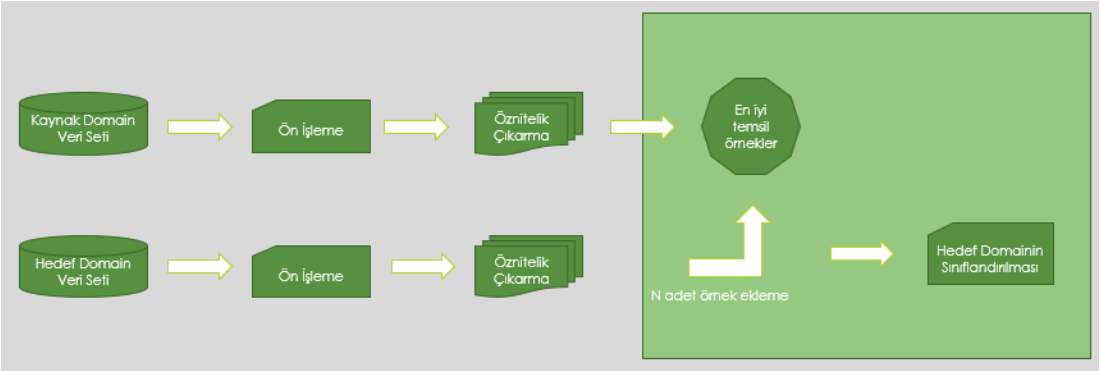
$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} \quad (2.20)$$

2.8 TRANSFER ÖĞRENME TEMELLİ AKTİF ÖĞRENME ADIMI

Domainler arası duygu analizi aşamasında Aktif Öğrenme ve Transfer Öğrenme temelli Aktif Öğrenme kullanılarak Lojistik Regresyon, Destek Vektör Makinesi ve Yapay Sinir Ağları yöntemleri ile sınıflandırma başarıları ölçülmüştür. Sadece Aktif Öğrenmenin kullanıldığı çalışmada hedef domaini en iyi temsil eden örnekler ile model eğitilmeye başlanmıştır. Her iterasyonda yine hedef domainden seçilen modele göre duygu sınıfı en belirsiz örnekler etiketlenmiştir. Eğitim 100 iterasyon sonunda tamamlanmış, hedef domaindeki test verisi ile modelin eğitim başarısı test edilmiştir.

Domainler arası duygu analizinde Transfer Öğrenme temelli Aktif Öğrenme adımında kaynak domaini en iyi temsil eden kayıtlar ile model eğitilmeye başlanmıştır. Her iterasyonda hedef domainden seçilen modele göre duygu sınıfı en belirsiz örnekler

etiketlenmiştir. Eğitim 100 iterasyon sonunda tamamlanmıştır. Şekil 2.10'da ilgili akış görülmektedir.



Şekil 2.10 Domainler arası duygu analizi iş akışı.

Aktif Öğrenme kullanılan duygu analizi çalışması ile Transfer Öğrenme temelli Aktif Öğrenme kullanılan duygu analizi çalışması başarıları kıyaslanmıştır. Şekil 2.11'de Transfer Öğrenme temelli Aktif Öğrenme için Python ile yazılmış kod bloğu görülmektedir.

```
import sklearn.model_selection as model_selection
from sklearn import svm

dataset = pd.read_csv("5-Word2Vec.csv").values[:, ]
dataset_train = pd.read_csv("19-KNNMostRepresentiveWord2Vec100.csv").values[:,]
x_train = dataset_train[:, :-1]
y_train = dataset_train[:, -1]

x_target = dataset[:, :-1]
y_target = dataset[:, -1]


ac1, ac2 = [], []  # arrays to store accuracy of different models
```

```
for i in range(100):
    # split dataset into train(5 %), test(25 %), unlabel(70 %)
    unlabel,x_test,label, y_test = model_selection.train_test_split(x_target,y_target,train_size=0.9, random_state=42)

    # train model by active learning
    for i in range(5):
        #classifier1 = LogisticRegression()
        #classifier1.fit(x_train, y_train)
        classifier1 = svm.SVC(kernel='linear',probability=True)
        classifier1.fit(x_train, y_train)
        if( len(unlabel)!= 0 ):
            y_probab = classifier1.predict_proba(unlabel)[:, 0]
        else:
            break

        #y_probab = classifier1.predict_proba(unlabel)[:, 0]

        p = 0.48 # range of uncertanity 0.47 to 0.53
        uncrt_pt_ind = []
        #print(unlabel.shape[0])
        for i in range(unlabel.shape[0]):
            if(y_probab[i] >= p and y_probab[i] <= 1-p):
                uncrt_pt_ind.append(i)
```

```
    classifier2 = svm.SVC(kernel='linear')
    classifier2.fit(x_train, y_train)
    ac1.append(classifier2.score(x_test, y_test))

        #''' split dataset into train(same as generated by our model),
        #test(25 %), unlabel(rest) '''
    train_size = x_train.shape[0]/dataset_train.shape[0]
    x_train, y_train, x_test, y_test, unlabel, label = split(
        dataset, 0.8, 0.2)

        # train model without active learning
    #classifier3 = LogisticRegression()
    #classifier3.fit(x_train, y_train)
    #ac2.append(classifier3.score(x_test, y_test))
    classifier3 = svm.SVC(kernel='linear')
    classifier3.fit(x_train, y_train)
    ac2.append(classifier3.score(x_test, y_test))


print("Accuracy by active model :", mean(ac1)*100)
print("Accuracy by random sampling :", mean(ac2)*100)
```

Şekil 2.11 Metinlerin vektörize edilmesi için yazılmış kod bloğu.

BÖLÜM 3

DENEYSEL SONUÇLAR VE TARTIŞMA

Bu bölümde veri seti üzerine yapılan çalışmada elde edilen sonuçlar anlatılmıştır. Film yorumları domaini Kelime Çantası, TF-IDF, N-Gram ve Skip-Gram yaklaşımları ile vektörize edilmişlerdir. N = 3 olarak kullanılmış kelimelerin 3 komşuluğu için N-gram hesaplamaları yapılmıştır. Skip-gram yaklaşımında pencere boyutu 50, 100, 150, 200 ve 250 olarak kullanılmıştır. Vektörize edilen metinler Lojistik Regresyon, Naïve Bayes ve Rastgele Orman sınıflandırma metodlarına göre sınıflandırılmışlardır. Çizelge 3.1'de sınıflandırma başarıları görülmektedir.

Çizelge 3.1. Film yorumları domaini sınıflandırma başarıları

	Metinlerin Sayısallaştırılması için Kullanılan Vektörizasyon Metodları							
	Kelime Çantası	**Tf-IDF**	**N-Gram**	**Skip-Gram 50**	**Skip-Gram 100**	**Skip-Gram 150**	**Skip-Gram 200**	**Skip-Gram 250**
Lojistik Regresyon	0.83926	0.84928	0.77735	0.76395	0.76499	0.75310	0.75408	0.77037
Naive Bayes	0.87490	0.86614	0.77860	0.51950	0.57112	0.54265	0.69974	0.70619
Rastgele Orman	0.83927	0.82489	0.77735	0.78076	0.77485	0.77481	0.77433	0.78764

Film yorumları domaininde veri setinin %85'i eğitim %15'i test verisi olarak kullanılmıştır. Domainin kelime çantası ve TF-IDF metin temsilleri ile sınıflandırıldıklarında daha başarılı sonuçlar elde edildiği görülmüştür. Ancak bu temsil yöntemleri için üretilen öznitelik vektör boyutları Skip-Gram yöntemine göre temsil edilen öznitelik vektör boyutlarından çok daha büyüktür. Kelime çantası ve TF-IDF modellerine göre temsil edilen domainde 27944 öznitelikten bahsedilirken, Skip-Gram yönteminde 50, 100, 150, 200 ve 250 boyutlu vektörler kullanılmıştır. Domain

sınıflandırmasında en başarılı sınıflandırıcı temsil yöntemlerine göre değişmektedir. Kelime çantası, TF-IDF ve N-Gram ile temsil edilen domainde en başarılı sınıflandırıcı Naive Bayes iken Skip-Gram ile temsil edilen domain için Logistik Regresyon ve Rastgele Orman sınıflandırma yaklaşımı daha başarılı olmuştur.

Kitap yorumları domaini Kelime Çantası, TF-IDF, N-Gram ve Skip-Gram yaklaşımları ile vektörize edilmişlerdir. N = 3 olarak kullanılmış kelimelerin 3 komşuluğu için N-gram hesaplamaları yapılmıştır. Skip-gram yaklaşımında pencere boyutu 50, 100, 150, 200 ve 250 olarak kullanılmıştır. Vektörize edilen metinler Lojistic Regresyon, Naïve Bayes ve Rastgele Orman sınıflandırma metodlarına göre sınıflandırılmışlardır. Çizelge 3.2'de sınıflandırma başarıları görülmektedir.

Çizelge 3.2. Kitap yorumları domaini sınıflandırma başarıları.

Metinlerin Sayısallaştırılması için Kullanılan Vektörizasyon Metodları								
	Kelime Çantası	**Tf-IDF**	**N-Gram**	**Skip-Gram 50**	**Skip-Gram 100**	**Skip-Gram 150**	**Skip-Gram 200**	**Skip-Gram 250**
Lojistik Regresyon	0.77142	0.78095	0.71904	0.77777	0.76274	0.77008	0.76324	0.75128
Naive Bayes	0.80952	0.78095	0.70952	0.76680	0.76250	0.74017	0.71780	0.75897
Rastgele Orman	0.72380	0.72380	0.71904	0.78917	0.74000	0.80056	0.79330	0.79672

Kitap yorumları domaininde veri setinin %85'i eğitim %15'i test verisi olarak kullanılmıştır. Domainin kelime çantası, TF-IDF, N-Gram ya da Skip-Gram yöntemlerine göre temsil edilmesi sınıflandırma başarısını film yorumları domainindeki kadar etkilememiştir, sınıflandırma başarıları benzer olmuştur. Ancak yine de kelime çantası ve TF-IDF modellerine göre temsil edilen domainde 9526 öznitelikten bahsedilirken, Skip-Gram yönteminde 50, 100, 150, 200 ve 250 boyutlu vektörler kullanılmıştır. Domain sınıflandırmasında en başarılı sınıflandırıcı temsil yöntemlerine göre değişmektedir. Kelime çantası ile temsil edilen veri setinde en başarılı sınıflandırıcı Naive Bayes iken, TF-IDF ile temsil edilen veri setinde Lojistik Regresyon ve Navive Bayes yüksek başarılı sonuçlar elde etmiş, N-Gram ile temsil

edilen domainde en başarılı sınıflandırıcılar Lojistik Regresyon ve Rastgele Orman olmuştur.

DVD yorumları domaini Kelime Çantası, TF-IDF, N-Gram ve Skip-Gram yaklaşımları ile vektörize edilmişlerdir. N = 3 olarak kullanılmış kelimelerin 3 komşuluğu için N-gram hesaplamaları yapılmıştır. Skip-gram yaklaşımında pencere boyutu 50, 100, 150, 200 ve 250 olarak kullanılmıştır. Vektörize edilen metinler Lojistik Regresyon, Naïve Bayes ve Rastgele Orman sınıflandırma metodlarına göre sınıflandırılmışlardır. Çizelge 3.3' de sınıflandırma başarıları görülmektedir.

Çizelge 3.3. DVD yorumları domaini sınıflandırma başarıları.

Metinlerin Sayısallaştırılması için Kullanılan Vektörizasyon Metodları								
	Kelime Çantası	**Tf-IDF**	**N-Gram**	**Skip-Gram 50**	**Skip-Gram 100**	**Skip-Gram 150**	**Skip-Gram 200**	**Skip-Gram 250**
Lojistik Regresyon	0.75670	0.74380	0.71428	0.53658	0.66157	0.53658	0.54158	0.53658
Naive Bayes	0.74285	0.75714	0.70952	0.53658	0.53513	0.53658	0.61113	0.52649
Rastgele Orman	0.67619	0.66666	0.71428	0.67404	0.64263	0.67936	0.66809	0.69944

DVD yorumları domaininde veri setinin %85'i eğitim %15'i test verisi olarak kullanılmıştır. Domainin Rastgele Orman sınıflandırıcı kullanıldığı test senaryolarında, metin temsil yöntemlerine göre sınıflandırma başarı farklarının çok belirgin olmadığı görülürken, Lojistik Regresyon ve Naive Bayes ve sınıflandırıcılar için temsil yöntemlerine göre daha belirgin sınıflandırma başarı farkları oluşturmuştur. Bu test senaryolarında domainin kelime çantası, TF-IDF ve N-Gram metin temsilleri ile sınıflandırıldıklarında daha başarılı sonuçlar elde edildiği görülmüştür. Kelime çantası ve TF-IDF modellerine göre temsil edilen domainde 12432 öznitelikten bahsedilirken, Skip-Gram yönteminde 50, 100, 150, 200 ve 250 boyutlu vektörler kullanılmıştır. Domain sınıflandırmasında en başarılı sınıflandırıcı temsil yöntemlerine göre değişmektedir. Kelime çantası ile temsil edilen veri setinde en başarılı sınıflandırıcı Lojistik Regresyon iken, TF-IDF ile temsil edilen veri setinde

Naive Bayes yüksek başarılı sonuçlar elde etmiş, N-Gram ile temsil edilen domainde en başarılı sınıflandırıcılar Lojistik Regresyon ve Rastgele Orman olmuştur.

Elektronik ürün yorumları domaini Kelime Çantası, TF-IDF, N-Gram ve Skip-Gram yaklaşımları ile vektörize edilmişlerdir. N = 3 olarak kullanılmış kelimelerin 3 komşuluğu için N-gram hesaplamaları yapılmıştır. Skip-gram yaklaşımında pencere boyutu 50, 100, 150, 200 ve 250 olarak kullanılmıştır. Vektörize edilen metinler Lojistik Regresyon, Naïve Bayes ve Rastegele Orman sınıflandırma metodlarına göre sınıflandırılmışlardır. Çizelge 4.4' de sınıflandırma başarıları görülmektedir.

Çizelge 4.4. Elektronik ürün yorumları domaini sınıflandırma başarıları.

Metinlerin Sayısallaştırılması için Kullanılan Vektörizasyon Metodları								
	Kelime Çantası	**Tf-IDF**	**N-Gram**	**Skip-Gram 50**	**Skip-Gram 100**	**Skip-Gram 150**	**Skip-Gram 200**	**Skip-Gram 250**
Lojistik Regresyon	0.84093	0.78487	0.69523	0.7215	0.71278	0.7096	0.71025	0.71472
Naive Bayes	0.85370	0.70476	0.67525	0.5670	0.69452	0.6740	0.66040	0.66890
Rastgele Orman	0.78501	0.76345	0.69523	0.7537	0.71278	0.7555	0.73297	0.74696

Elektronik ürün yorumları domaininde veri setinin %85'i eğitim %15'i test verisi olarak kullanılmıştır. Domainin Rastgele Orman sınıflandırıcı kullanıldığı test senaryolarında, metin temsil yöntemlerine göre sınıflandırma başarı farklarının çok belirgin olmadığı görülürken, özellikle Naive Bayes sınıflandırıcısı için temsil yöntemlerine göre daha belirgin sınıflandırma başarı farkları oluşturmuştur. Diğer domainlerdeki gibi veri setinin kelime çantası ve TF-IDF metin temsilleri ile sınıflandırıldıklarında daha başarılı sonuçlar elde edildiği görülmektedir. Bu domainde kelime çantası ve TF-IDF modellerine göre temsilde 12476 öznitelikten bahsedilirken, Skip-Gram yönteminde 50, 100, 150, 200 ve 250 boyutlu vektörlerden bahsedilir. Domain sınıflandırmasında en başarılı sınıflandırıcı temsil yöntemlerine göre değişmektedir. Kelime çantası ile temsil edilen veri setinde en başarılı sınıflandırıcı Naive Bayes iken, TF-IDF ile temsil edilen veri setinde Lojistik Regresyon yüksek

başarılı sonuçlar elde etmiş, N-Gram ile temsil edilen domainde en başarılı sınıflandırıcılar Lojistik Regresyon ve Rastgele Orman olmuştur.

Mufak yorumları domaini Kelime Çantası, TF-IDF, N-Gram ve Skip-Gram yaklaşımları ile vektörize edilmişlerdir. N = 3 olarak kullanılmış kelimelerin 3 komşuluğu için N-gram hesaplamaları yapılmıştır. Skip-gram yaklaşımında pencere boyutu 50, 100, 150, 200 ve 250 olarak kullanılmıştır. Vektörize edilen metinler Lojistik Regresyon, Naïve Bayes ve Rastgele Orman sınıflandırma metodlarına göre sınıflandırılmışlardır. Çizelge 3.5' de sınıflandırma başarıları görülmektedir.

Çizelge 3.5. Mutfak ürün yorumları domaini sınıflandırma başarıları.

Metinlerin Sayısallaştırılması için Kullanılan Vektörizasyon Metodları								
	Kelime Çantası	**Tf-IDF**	**N-Gram**	**Skip-Gram 50**	**Skip-Gram 100**	**Skip-Gram 150**	**Skip-Gram 200**	**Skip-Gram 250**
Lojistik Regresyon	0.78271	0.79951	0.66666	0.6914	0.68424	0.6832	0.67487	0.67991
Naive Bayes	0.78293	0.70141	0.67007	0.6700	0.68121	0.6667	0.65719	0.67497
Rastgele Orman	0.79738	0.79955	0.66666	0.7177	0.69705	0.6930	0.70801	0.68163

Mutfak ürün yorumları domaininde veri setinin %85'i eğitim %15'i test verisi olarak kullanılmıştır. Domainin kelime çantası ve TF-IDF yöntemlerine göre temsil edilmesi sınıflandırma başarısını etkilemiştir, bu temsil yöntemleri kullanıldığında sınıflandırma başarılarının diğer temsil yöntemlerine göre daha yüksek olduğu görülmüştür. Ancak kelime çantası ve TF-IDF modellerine göre temsil edilen domainde 13138 öznitelikten bahsedilirken, Skip-Gram yönteminde 50, 100, 150, 200 ve 250 boyutlu özniteliklerden bahsedilir. Domain sınıflandırmasında en başarılı sınıflandırıcı temsil yöntemlerine göre değişmektedir.

Film yorumları, DVD ve mutfak ürün yorumları domaininde; metin temsil yöntemi olarak kelime çantası ya da TF-IDF kullanıldığı test senaryolarının, Skip-Gram yöntemine göre temsil edilmiş metinlere göre daha başarılı sınıflandırma yapılabildiği

görülmektedir. Ancak kelime çantası ve TF-IDF temsil yöntemlerinde domaindeki benzersiz kelime sayısı kadar öznitelik vektör boyutundan bahsedilir. Öznitelik vektör boyutunun büyüklüğü sınıflandırma hızını ve performansını etkilemektedir. Ayrıca domainler arası duygu analizinde transfer edilecek kayıtlar her iki domainde de vektör boyutu olarak eşit olmalıdır. Domainler arası duygu analizinde kullanılmak üzere her domainin temsilinde Skip-Gram 100'lük vektör kullanılmasına karar verilmiştir. Böylece domainlerdeki her kayıt 100 öznitelik içeren vektörler ile temsil edilecektir.

Aktif öğrenme transfer öğrenme ile besleyerek modellemek başarı oranını artırmıştır. En çok başarı kitap yorumları ile film yorumları domainleri arasında, DVD yorumları arasında olduğu görülmüştür. Çizelge 3.6'da hedef kaynak domainler arası transfer öğrenme tabanlı aktif öğrenmede Logistik Regresyon yaklaşımı ile sınıflandırma başarıları görülmektedir.

Çizelge 3.6. Logistik regresyon ile domainler arası duygu analizi metrikleri.

Hedef Domain – Kaynak Domain	**Sadece AÖ Yaklaşımı ile Doğruluk Oranları**	**TÖ Tabanlı AÖ ile Doğruluk Oranları**	**TÖ Tabanlı AÖ ile Precision Oranları**	**TÖ Tabanlı AÖ ile Recall Oranları**	**TÖ Tabanlı AÖ ile F-Score Oranları**
Film - DVD	**63.02162**	**65.92592**	**0.632**	**0.646**	**0.639**
Film – Kitap	**71.05555**	**77.84962**	**0.759**	**0.804**	**0.781**
Film - Elektronik	**71.21186**	**75.45423**	**0.628**	**0.647**	**0.638**
Film – Mutfak	70.43442	69.02970	0.666	0.681	0.674
DVD - Kitap	**70.57407**	**77.75187**	**0.690**	**0.704**	**0.697**
DVD - Elektronik	**72.52542**	**75.27796**	**0.628**	**0.647**	**0.638**
DVD - Mutfak	70.72131	69.20462	0.619	0.634	0.627
Kitap - Elektronik	**71.19491**	**75.51864**	**0.711**	**0.727**	**0.719**
Kitap - Mutfak	69.27392	69.79508	0.680	0.667	0.674

DVD domaini kaynak domain iken film yorumları domaini hedef domain olduğunda sadece aktif öğrenme kullanılarak ve transfer öğrenme tabanlı aktif öğrenme yaklaşımı kullanılarak duygu analizi yapılmıştır. Sadece aktif öğrenme kullanıldığında film yorumları domaini bu domainden rastgele seçilmiş örneklerle eğitilmeye başlanır, her

iterasyonda yine film yorumları domaininden duygu sınıfı en belirsiz örnekler etiketlenip modelin tekrar eğitilmesi sağlanır. Eğitim tamamlandıktan sonra test veri seti ile test yapılır. Film yorumları domainin %80'i eğitim %20'si test verisi olarak kullanıldığında ve Lojistik Regresyon ile sınıflandırma yapıldığında %63.02 doğrulukla duygu sınıfı doğru tespit edilebilmektedir. Ancak transfer öğrenme tabanlı aktif öğrenme yaklaşımı kullanıldığında duygu analizi yapacak model üretilirken, ilk eğitimi sağlayacak veri DVD domainini en iyi temsil eden örnekler olmaktadır. Bu modelde her iterasyonda film yorumları domaininden duygu sınıfı en belirsiz örnekler seçilir ve etiketlendikten sonra modelin tekrar eğitimi sağlanır. Eğitim tamamlandıktan sonra film yorumları domainindeki test verisi ile test yapılır. Transfer öğrenme temelli aktif öğrenme yaklaşımında film yorumları domainin %80'i eğitim %20'si test verisi olarak kullanıldığında ve Lojistik Regresyon ile sınıflandırma yapıldığında %65.92 doğrulukla duygu sınıfı doğru tespit edilebilmektedir. Diğer hedef ve kaynak domain çiftleri için de benzer modeller üretilmiştir.

Film yorumları hedef domain iken kitap yorumları kaynak domain olduğunda sadece aktif öğrenme modelinde %71 doğrulukla sınıflandırma yapılırken, transfer öğrenme ile birlikte aktif öğrenme yaklaşımı kullanıldığında %77 doğrulukla sınıflandırma yapılabilmiştir. Film yorumları hedef domain iken elektronik ürün yorumları kaynak domain olduğunda sadece aktif öğrenme modelinde %71 doğrulukla sınıflandırma yapılırken, transfer öğrenme ile birlikte aktif öğrenme yaklaşımı kullanıldığında %75 doğrulukla sınıflandırma yapılabilmiştir. Film yorumları hedef domain iken mutfak ürünleri yorumları kaynak domain olduğunda sadece aktif öğrenme modelinde %70 doğrulukla sınıflandırma yapılırken, transfer öğrenme ile birlikte aktif öğrenme yaklaşımı kullanıldığında %69 doğrulukla sınıflandırma yapılabilmiştir. Bu iki domainde transfer öğrenme yaklaşımı kullanımı negatif etki yaratmıştır. DVD yorumları hedef domain iken kitap yorumları kaynak domain olduğunda sadece aktif öğrenme modelinde %70 doğrulukla sınıflandırma yapılırken, transfer öğrenme ile birlikte aktif öğrenme yaklaşımı kullanıldığında %77 doğrulukla sınıflandırma yapılabilmiştir. DVD yorumları hedef domain iken elektronik ürün yorumları kaynak domain olduğunda sadece aktif öğrenme modelinde %72 doğrulukla sınıflandırma yapılırken, transfer öğrenme ile birlikte aktif öğrenme yaklaşımı kullanıldığında %75 doğrulukla sınıflandırma yapılabilmiştir. DVD yorumları hedef domain iken mutfak

ürünleri yorumları kaynak domain olduğunda sadece aktif öğrenme modelinde %70 doğrulukla sınıflandırma yapılırken, transfer öğrenme ile birlikte aktif öğrenme yaklaşımı kullanıldığında %69 doğrulukla sınıflandırma yapılabilmiştir. DVD ile mutfak ürünleri domainleri arasında transfer öğrenme ile kayıtların transfer edilmesi negatif bir etki yaratmıştır. Kitap yorumları hedef domain iken elektronik ürünleri yorumları kaynak domain olduğunda sadece aktif öğrenme modelinde %71 doğrulukla sınıflandırma yapılırken, transfer öğrenme ile birlikte aktif öğrenme yaklaşımı kullanıldığında %75 doğrulukla sınıflandırma yapılabilmiştir.

Kitap yorumları hedef domain iken mutfak ürünleri yorumları kaynak domain olduğunda sadece aktif öğrenme modelinde %69 doğrulukla sınıflandırma yapılırken, transfer öğrenme ile birlikte aktif öğrenme yaklaşımı kullanıldığında yine %69 doğrulukla sınıflandırma yapılmıştır. Bu iki domain arasında transfer öğrenme kullanımı etkisiz kalmıştır. Çizelge 3.7'de hedef kaynak domainler arası transfer öğrenme tabanlı aktif öğrenmede Destek Vektör Makinesi yaklaşımı ile sınıflandırma başarıları görülmektedir.

Çizelge 3.7. Destek vektör makinesi ile domainler arası duygu analizi metrikleri.

Hedef Domain – Kaynak Domain	**Sadece AÖ Yaklaşımı ile Doğruluk Oranları**	**TÖ Tabanlı AÖ ile Doğruluk Oranları**	**TÖ Tabanlı AÖ ile Precision Oranları**	**TÖ Tabanlı AÖ ile Recall Oranları**	**TÖ Tabanlı AÖ ile F-Score Oranları**
Film - DVD	**68.45390**	**73.21629**	**0.711**	**0.696**	**0.703**
Film – Kitap	**64.62962**	**75.87218**	**0.733**	**0.717**	**0.723**
Film - Elektronik	**73.99152**	**75.75113**	**0.744**	**0.729**	**0.737**
Film – Mutfak	70.93442	69.84158	0.686	0.700	0.693
DVD - Kitap	**64.22222**	**74.95488**	**0.729**	**0.729**	**0.729**
DVD - Elektronik	**73.45762**	**75.61016**	**0.628**	**0.647**	**0.638**
DVD - Mutfak	71.52459	69.98349	0.619	0.634	0.627
Kitap - Elektronik	**72.81355**	**75.44406**	**0.750**	**0.735**	**0.742**
Kitap - Mutfak	70.07590	70.38524	0.720	0.692	0.706

Film yorumları hedef domain iken DVD yorumları kaynak domain olduğunda; destek vektör makinesi yaklaşımının kullanıldığı test senaryolarında sadece aktif öğrenme modelinde ile %68 doğrulukla sınıflandırma yapılırken, transfer öğrenme ile birlikte aktif öğrenme yaklaşımı kullanıldığında %73 doğrulukla sınıflandırma yapılabilmiştir. Ayrıca film yorumları hedef domain iken kitap yorumları kaynak domain olduğunda sadece aktif öğrenme modelinde %64 doğrulukla sınıflandırma yapılırken, transfer öğrenme ile birlikte aktif öğrenme yaklaşımı kullanıldığında %75 doğrulukla sınıflandırma yapılabilmiştir. Diğer domainler arası duygu analizi senaryolarında destek vektör makinesinin kullanım sonuçları da yukarıda görülmektedir. Çizelge 3.8'de hedef kaynak domainler arası transfer öğrenme tabanlı aktif öğrenmede Yapay Sinir Ağları yaklaşımı ile sınıflandırma başarıları görülmektedir.

Çizelge 3.8. Yapay sinir ağları ile domainler arası duygu analizi metrikleri.

Hedef Domain – Kaynak Domain	**Sadece AÖ Yaklaşımı ile Doğruluk Oranları**	**TÖ Tabanlı AÖ ile Doğruluk Oranları**	**TÖ Tabanlı AÖ ile Precision Oranları**	**TÖ Tabanlı AÖ ile Recall Oranları**	**TÖ Tabanlı AÖ ile F-Score Oranları**
Film - DVD	**65.36111**	**70.12359**	**0.666**	**0.723**	**0.694**
Film – Kitap	**72.55555**	**79.64661**	**0.833**	**0.769**	**0.800**
Film - Elektronik	79.65762	79.98305	0.829	0.765	0.796
Film – Mutfak	74.99180	72.04290	0.716	0.745	0.731
DVD - Kitap	**72.11111**	**79.48872**	**0.791**	**0.776**	**0.784**
DVD - Elektronik	**78.67796**	**79.23728**	**0.803**	**0.788**	**0.796**
DVD - Mutfak	74.48360	72.76237	0.717	0.688	0.702
Kitap - Elektronik	**78.16949**	**79.05084**	**0.800**	**0.784**	**0.792**
Kitap - Mutfak	73.16171	73.90163	0.739	0.694	0.716

Film yorumları hedef domain iken DVD yorumları kaynak domain olduğunda; yapay sinir ağları yaklaşımının kullanıldığı test senaryolarında sadece aktif öğrenme modelinde ile %65 doğrulukla sınıflandırma yapılırken, transfer öğrenme ile birlikte aktif öğrenme yaklaşımı kullanıldığında %70 doğrulukla sınıflandırma yapılabilmiştir. Ayrıca film yorumları hedef domain iken kitap yorumları kaynak domain olduğunda sadece aktif öğrenme modelinde %72 doğrulukla sınıflandırma yapılırken, transfer

öğrenme ile birlikte aktif öğrenme yaklaşımı kullanıldığında %79 doğrulukla sınıflandırma yapılabilmiştir. Diğer domainler arası duygu analizi senaryolarında yapay sinir ağları kullanım sonuçları da yukarıda görülmektedir.

Hedef domainin film yorumları olup, kaynak domainin DVD yorumları olduğu duygu analizi modelinde Transfer Öğrenme temelli Aktif Öğrenme yaklaşımı ve sadece Aktif Öğrenme yaklaşımı kullanılarak sınıflandırma başarıları ölçüldüğünde başarının arttığı görülmektedir. Film yorumları ve DVD yorumları domainlerinin birbirine benzer domainler olması bu başarının sebebi olarak görülebilir. LR, DVM ve YSA yaklaşımlarının tamamı için sınıflandırma başarısı Transfer Öğrenme yaklaşımı kullanıldığında artmıştır. DVM ve YSA yaklaşımı kullanılan modelde yaklaşık %5 sınıflandırma başarısı artmıştır.

Hedef domainin film yorumları olup, kaynak domainin kitap yorumları olduğu duygu analizi modelinde bu iki domainin benzer olması yine başarıyı artıran etkendir. Film yorumları domainini en iyi temsil eden örneklerle modelin eğitilmeye başlanması kitap domainini için duygu analizi çalışması yapılırken pozitif etki yaratmıştır. LR, DVM ve YSA yaklaşımlarının tamamı için sınıflandırma başarısı Transfer Öğrenme yaklaşımı kullanıldığında artmıştır. DVM yaklaşımı kullanılan modelde yaklaşık %9 sınıflandırma başarısı artarken, YSA yaklaşımı kullanılan modelde bu başarı yaklaşık %7 olmaktadır.

Hedef domainin film yorumları olup, kaynak domainin elektronik ürün yorumları olduğu duygu analizi modelinde sınıflandırma başarısının LR, DVM ve YSA yaklaşımlarının tamamı için de arttığı söylenemez. YSA yaklaşımı kullanılan modelde başarı hiç artmazken, DVM yaklaşımı kullanılan modelde başarı %2 artmış, LR yaklaşımında yaklaşık %4 artmıştır. Hedef ve kaynak domainlerinin benzer domainler olmaması nedeniyle sınıflandırma başarısı beklendiği gibi çok artmamıştır.

Hedef domainin film yorumları olup, kaynak domainin mutfak ürün yorumları olduğu duygu analizi modelinde LR, DVM ve YSA yaklaşımlarının tamamı için Transfer Öğrenme temelli modelde başarının azaldığı görülmektedir. Bu iki domainin birbirine

benzemeyen domainler olması Transfer Öğrenme temelli yaklaşımda negatif bir etki oluşturmuştur.

Hedef domainin DVD yorumları olup, kaynak domainin kitap ürün yorumları olduğu duygu analizi modelinde sınıflandırma başarıları ölçüldüğünde Transfer Öğrenme temelli yaklaşımın başarıyı artırdığı görülmektedir. DVD yorumları ve kitap yorumları domainlerinin birbirine benzer domainler olması bu başarının sebebi olarak görülebilir. LR yaklaşımı kullanılan modelde yaklaşık %7 sınıflandırma başarısı artarken, DVM yaklaşımı kullanılan modelde bu başarı yaklaşık %10, YSA yaklaşımı kullanılan modelde ise bu başarı yaklaşık %7 olmaktadır.

Hedef domainin DVD yorumları olup, kaynak domainin elektronik ürün yorumları olduğu duygu analizi modelinde sınıflandırma başarıları ölçüldüğünde Transfer Öğrenme yaklaşımı kullanılmasının başarıyı artırdığı görülmüştür. LR yaklaşım kullanılan modelde başarı yaklaşık %3, DVM yaklaşım kullanılan modelde yaklaşık %4 ve YSA yaklaşım kullanılan modelde yaklaşık %1 artış olmuştur.

Hedef domainin DVD yorumları olup, kaynak domainin mutfak ürün yorumları olduğu duygu analizi modelinde LR, DVM ve YSA yaklaşımlarının tamamı için Transfer Öğrenme temelli modelde başarının azaldığı görülmektedir. Transfer öğrenme birbirine benzemeyen bu iki domain için bu iki kullanıldığında negatif bir etki oluşturmuştur.

Hedef domainin kitap yorumları olup, kaynak domainin elektronik ürün yorumları olduğu duygu analizi modelinde Transfer Öğrenme yaklaşımı kullanılmasının başarıyı artırdığı görülmüştür. LR yaklaşım kullanılan modelde başarı yaklaşık %4, DVM yaklaşım kullanılan modelde yaklaşık %3 ve YSA yaklaşım kullanılan modelde yaklaşık %1 artış olmuştur.

Hedef domainin kitap yorumları olup, kaynak domainin mutfak ürün yorumları olduğu duygu analizi modelinde LR, DVM ve YSA yaklaşımlarının tamamı için Transfer Öğrenme temelli modelde başarının artmadığı görülmüştür. Transfer Öğrenme

yaklaşımı kullanılması sonrası başarı artışı olmaması iki domainin birbirine benzememesi dolayısıyladır.

Transfer öğrenme temelli Aktif Öğrenme yaklaşımı kullanılarak domainler arası duygu analizi çalışılan modelde domainlerin birbirine benzer yapıya sahip olması başarıyı artırırken birbirinden farklı olan domainlerde negatif bir etki oluşturmuş ya da kayıt transferinin etkisiz kaldığı görülmüştür.

BÖLÜM 4

SONUÇ

Bu tez çalışmasında domainler arası duygu analizi yapan bir model sunulmuştur. Geliştirilen transfer öğrenme temelli aktif öğrenme yaklaşımı üç temel adımdan oluşmaktadır. İlk adımda kaynak domaini en iyi temsil eden örnekler ile model eğitilmiştir. İkinci adımda duygu analizi çalışılan hedef domainden mevcut modele göre duygu sınıfı en belirsiz örnekler seçilir ve duygu sınıfına göre etiketlenir. Üçüncü adımda yeni etiketlenmiş veriler ile modelin yeniden eğitilmesi sağlanır. İkinci ve üçüncü adımlar iterasyon sayısı kadar tekrarlandıktan sonra hedef domaindeki test verisiyle modelin başarısı ölçülür.

Türkçe metinlerden oluşan beş domainden hedef ve kaynak domainler olarak kullanılmak üzere dokuz adet ikili grup oluşturulmuştur. Çalışma gruplarının her biri LR, DVM ve YSA yaklaşımları ile sınıflandırılarak önerilen modelin sınıflandırma başarıları ölçülmüştür. Çalışma gruplarının birbirine benzer domainler olması durumlarında; domainler arası kayıtların transfer edildiği önerilen modelin başarısı görülebilmiştir. Örneğin film yorumları ile kitap domaini arasında transfer öğrenme temelli aktif öğrenme kullanılarak duygu analizi yapıldığında ortalama %8 daha yüksek doğrulukla duygu sınıfı tespit edilebilmektedir. Benzer şekilde DVD domaini ile kitap yorumları domainleri kullanımında da ortalama %8 daha yüksek doğrulukla sınıflandırma yapılmıştır. Birbirine benzer oldukları düşünülen iki domainde duygu analizi için az sayıda etiketli veri bulunması durumlarında önerilen modelin kullanılabileceği görülmektedir. Böylece veri etiketlemenin zaman ve performans gerektirmesi ihtiyacının önüne geçilebilir.

KAYNAKLAR

1. İnternet: Türkiye Dil Kurumu Sözlükleri, "Güncel Türkçe Sözlük", **https://sozluk.gov.tr/?kelime=duygu** (2020).

2. L. Povoda, R. Burget and M. K. Dutta, "Sentiment analysis based on Support Vector Machine and Big Data", ***39th International Conference on Telecommunications and Signal Processing (TSP)***, 543-545, (2016).

3. Reshamwala, A., and Mishra, D., Pawar, P., "Review on Natural Language Processing", ***Engineering Science and Technology: An International Journal (ESTIJ)***, 3:113 – 116 (2013).

4. Tejedor, J.,Garcia, R., Fernandez, M., Lopez-Colino, F.J. and Perdrix, F., "Ontology-Based Retrieval of Human Speech", ***18th International Workshop on Database and Expert Systems Applications,*** Regensburg,485-489 (2007).

5. Kurdi, M. Z., "Natural Language Processing and Computational Linguistics: Speech, Morphology and Syntax", 1. Ed, John Wiley & Sons, (2016).

6. Beesley, K. R., "Morphological Analysis and Generation: A First-Step in Natural Language Processing", ***Proceedings of the SALTMIL Workshop at LREC 2004,*** Lisbon,1-8, (2004).

7. Mathew, K., and Issac., "Intelligent spam classification for mobile text message", ***2011 International Conference on Computer Science and Network Technology***, USA, 423-428 (2011).

8. Merchant, K, and Pande, Y., "NLP Based Latent Semantic Analysis for Legal Text Summarization", **2018 International Conference on Advances in Computing Communications and Informatics (ICACCI),** Bangalore, India, 1803-1807, (2018).

9. Glökner, I., Hartrumpf, S., Helbig, H., Leveling, J. and Osswald, R., "Automatic Semantic Analysis for NLP Applications", ***Zeitschrift für Sprachwissenschaft***, 26:241- 266 (2007).

10. Mahler, T., Cheung, W., and Elsner, M., "Breaking NLP: Using Morphosyntax, Semantics, Pragmatics and World Knowledge to Fool Sentiment Analysis Systems", ***In Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems***, Copenhagen, Denmark, 33–39, (2017).

11. Liu, B., "Sentiment analysis and opinion mining", ***Synthesis Lectures on Human Language Technologies,*** 5 :1–167 (2012).

12. Hussein, D. M. M., "A Survey of Sentiment Analysis Challenges***", Journal Of King Saud University***, 30(4): 330-338, (2018).

12. Özyurt, B. and Akçayol, M. A., "Fikir Madenciliği ve Duygu Analizi, Yaklaşımlar, Yöntemler Üzerine Bir Araştırma", Selçuk Üniversitesi Mühendislik Bilim ve Teknoloji Dergisi, 6(4): 668-693, (2018).

13. Fan, X., Li X., Du, F., and Wei, M, "Apply Word Vectors for Sentiment Analysis of App Reviews",***3rd International Conference on System and Informaties***, Shanghai,1062-1066, (2016).

14. Mcdonald, R., Hannan, K., Neylon, T., Wells, M., and Reynar, J., "Structured Models for Fine-to Coarse Sentiment Analysis", ***In Proceedings of ACL-07, 45th Annual Meeting of the Association of Computational Linguistics***, Prag, 432-439, (2007).

15. Sobhani, P., Mohammad, S. M. and Kiritchenko, S., "Detecting Stance in Tweets and Analyzing its Interaction with Sentiment", ***Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics***, Berlin, 159-169 (2016).

16. Ding, X., Liu, B., Yu, PS., "A Holistic Lexicon-Based Approach to Opinion Mining", ***In Proceedings of WSDM-2008, Conference on Web Search and Web Data Mining***, Stanford, 231-240, (2008).

17. Can, U., Alatas, B., "Duygu Analizi ve Fikir Madenciliği Algoritmaları İncelenmesi", ***International Journal of Psychology and Educational Studies***, 3(1): 75-111, (2017).

18. Çetin, M., Amasyalı, M.F., "Supervised and Traditional Term Weighting Methods for Sentiment Analysis***", Signal Processing and Communications Applications Conference (SIU)***, Haspolat - KKTC, 1-4, (2013).

19. Tan S, Zhang J. "An empirical study of sentiment analysis for Chinese document". ***Expert Systems with Applications***, 34(4), 2622-2629, (2008).

20. Turney P.D., "Thumbs up or Thumbs down?: Semantic Orientation Applied to Unsupervised Classification of Reviews", ***In Proceedings of ACL'02, 40th Annual Meeting of the Association for Computational Linguistics***, Pennsylvania, ABD, 417-424, (2002).

21. Blei, D.M., Ng, A.Y., Jordan, M.I., "Latent Dirichlet Allocation", ***The Journal of Machine Learning Research***, Ed. 3, 993-1022, (2003).

22. Asur, S. and Huberman, B. A., "Predicting the Future with Social Media", ***2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology,*** Canada, 492-499, (2010).

23. Zafra, S. M., Raez, A., M., Valdivia, M.T. and Lopez, L. A. U., "SINAI at SemEval-2017 Task4: User Based Classification", ***Proceedings of the 11th International Workshop on Semantic Evaluations,*** Canada, 634-639, (2017).

24. Correa, E., A., Marinho, V., Q. and Santos, L., B., "NILL-USP at SemEval-2017 Task 4: A Multi-view Ensemble for Twitter Sentiment Analysis", ***Proceedings of the 11th International Workshop on Semantic Evaluations,*** Canada, 611-615, (2017).

25. Cliche, M., "BB_Twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs", ***Proceedings of the 11th International Workshop on Semantic Evaluations,*** Canada, 573-579, (2017).

26. Rozental, A. and Fleischer, D., "Amobe at SemEval-2014 Task 4: Deep Learning System for Sentiment Detection on Twitter", ***Proceedings of the 11th International Workshop on Semantic Evaluations,*** Canada, 653-667, (2017).

27. Jiao, J. and Zhou, Y., "Sentiment Polarity Analysis based Multi-Dictionary", ***Journal of Physics Procedia***, 22:590-596, (2011).

28. Mulki, H., Haddad, H., Gridach, M. and Babaoglu, I., "Tw-Star at SemEval-2017 Task 4: Sentiment Classification of Arabic Tweets", ***Proceedings of the 11th International Workshop on Semantic Evaluations,*** Canada, 664- 667, (2017).

29. Nguyen, L. T., Chan, W., Peng, W. and Zhang, Y., "Predicting Collective Sentiment Dynamics from Time-Series Media", ***Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining,*** China, 1-8, (2012).

30. Zhuang, F., QI, Z., Duan, K., XI, D., Zhu, Y., Duan, K., Zhu, H., He, Q., and Xiong, H., "A Compresive Survey on Transfer Learning", ***Proceedıngs of the IEEE,*** 109(1), 43-76, (2021).

31. Liu, R., Shi, Y., Ji, C., and Jia, A. M., "A Survey of Sentiment Analysis Based on Transfer Learning", ***Proceedıngs of the in*** **IEEE Access**, Ed. 7, 85401-85412, (2019).

32. Mikolov, T., Chen, K., Corrado, K. and Dean, J., "Efficient Estimation of Word Representations in Vector Space", ***Proceedings of the International Conference on Learning Representations,*** Arizona,647-658, (2013).

33. Gupta, B., Ram, L., Awasthi S., Kumar, P., Agarwal, S., Singh, P. and Prasad, B., R., "Cross Domain Sentiment Analysis Using Transfer Learning", ***Proceedings of 2017 IEEE International Conference on Industrial and Information Systems (ICIIS)***, Peradaniya, 1-5, (2017).

34. Xu, R., Xu, R. and Wang, X., "Instance Level Transfer Learning for Cross Lingual Opinion Analysis", ***Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis,*** Oregon, 182- 188, (2011).

35. Blitzer, J., McDonald, R. and Pereira, R. "Domain Adaptation with Structural Correspondence", ***Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing***, Sdney, 120-128, (2006).

36. Blitzer, J., Drenze, M. and Pereira, F., "Biographies, Bollwood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification", ***Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics***, Prague, 440-447, (2007).

37. Yoshida, Y., Hirao, T., Iwata, T., Nagata, M. and Matsumoto, Y., "Transfer Learning for Multiple-Domain Sentiment Analysis-Identifying Domain Dependent/Independent Word Polarity", ***Proceedings of the 25th AAAI Conference on Artificial Intelligence,*** San Francisco, 1286-1291, (2011).

38. Glorot, X., Bordes, A. and Bengio, Y., "Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach", ***Proceedings of the 28th International Conference on Machine Learning,*** USA, 513-520, (2011).

39. Sattless, B. and Craven, M, "An Analysis of Active Learning Strategies for Sequence Labeling Task", ***Proceedings of the Conference on Empirical Methods in Natural Language Processing,*** Hawaii, 1070-1079, (2008).

40. Türkmenoğlu, C., "Türkçe Metinlerde Duygu Analizi", Yüksek Lisans Tezi, ***İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü***, İstanbul, 21-41 (2015).

41. Ant, K. and Diri, B., "Sosyal Ağımdaki Duygusal Uyum", ***Akıllı Sistemler ve Uygulamaları Dergisi.***, 1(1): 117-121 (2018).

42. İnternet: Zemberek NLP, "Zemberek nasıl çalışır? 1.Sözlük ve Kök ağacı", **http://zembereknlp.blogspot.com/2007/02/zemberek-nasl-alr-1szlk-ve-kk-aac.html** (2021).

43. Kaya, M., "Sentiment Analysis of Turkish Political Columns with Transfer Learning", Yüksek Lisans Tezi, ***Orta Doğu Teknik Üniversitesi Fen Bilimleri Enstitüsü***, İstanbul, 19-67 (2013).

44. Aydoğan, M. and Karcı, A., "Makine Öğrenmesi ve Transfer Öğrenme ile Türkçe Metin Sınıflandırma", ***2019 International Artificial Intelligence and Data Processing Symposium (IDAP),*** Turkey,1-6, (2019).

45. Akın, S. E. and Yıldız, T., "Sentiment Analysis through Transfer Learning for Turkish Language", ***2019 IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA),*** Sofia, Bulgaria, 1-6, (2019).

46. Lort Tosun, S. "Sentiment Analysis Using Active Learning Based Transfer Learning Approach in Turkish Texts" ***International Hazar Scientific Research Congress-II(Baku)***, Baku, Azerbaijan,741-752, (2021).

47. Şeker, S. E., "Duygu Analizi", Yönetim Bilişim Sistemleri Akademisi, 3(3), 21-36 ,(2016).

48. İnternet: Mining Social Media, "Code & Datasets", **https://www.win.tue.nl/~mpechen/projects/smm/#Datasets**, (2020).

49. İnternet: Machine Learning Mastery, "A Gentle Introduction to the Bag-of-Words Model", **https://machinelearningmastery.com/gentle-introduction-bag-words-model/** (2021).

50. Atan, S., "Metin Madenciliği: İmkanlar, Yöntemler ve Kısıtlar***", Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi***, 30(4): 220-239, (2020).

51. Jurafsky, D. and Martin, J. H., "Speech and Language Processing", 2nd ed. New York, 30-58 (2020).

52. Landthaler, J., Waltl, B., Huth, D., Braun, D., Matthes, F., Geiger, T. and Stocker C., "Extending Thesauri Using Word Embeddings and the Intersection Method",***Second Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2017),*** London, UK, (2017).

53. İnternet: Analytics Vidhya, "Word2Vec – CBOW&Skip-gram: Algoritmic Optimizations", **https://medium.com/analytics-vidhya/word2vec-cbow-skip-gram-algorithmic-optimizations-921d6f62d739,** (2020).

54. Tantuğ, A. C., "Metin Sınıflandırma", ***Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi,*** 5(2), 1305-1317, (2016).

55. Sebastiani, F., "Machine learning in automated text categorization", ***ACM Computing Surveys***, 34(1), 1-47, (2002).

56. Bilgin, M., Şentürk, İ. F., "Danışmanlı ve Yarı Danışmanlı Öğrenme Kullanarak Doküman Vektörleri Tabanlı Tweetlerin Duygu Analizi", ***BAUN Fen Bilimleri Enstitüsü Dergisi***, 21(2), 822-839, (2019).

57. Kleinbaum, G. D., "A Self-learning Text Logistic Regression", ***Springer***, Atlanta, (1994).

58. Das, S. ve Chen, M., "Yahoo! for Amazon: Extracting market sentiment from stock message boards", ***Proceedings of the Asia Pacific finance association annual conference (APFA)***, Bangkok, Thailand, (2001).

59. Çelik, U., Akçetin, E., Gök M., "Rapidminer ile Uygulamalı Veri Madenciliği", ***Pusula Yayıncılık***, (2017).

60. Hsu, C.W., Chang, C.C., Lin, C.J., "A practical guide to support vector classification". **www.csie.ntu.edu.tw/∼cjlin/papers/guide /guide.pdf**, (2003).

61. Çayıroğlu İ., "Yapay Sinir Ağları". Karabük Üniversitesi, Karabük, Türkiye, (2015).

62. Y. Guo and D. Schuurmans, "Discriminative batch mode active learning", In Advances in Neural Information Processing Systems (NIPS), 593–600, (2008).

ÖZGEÇMİŞ

Seher LORT TOSUN ilk ve orta öğrenimini aynı şehirde tamamladı. Yıldız Teknik Üniversitesi Elektrik-Elektronik Fakültesi Bilgisayar Mühendisliği Bölümü'nde öğrenime başlayıp 2015 yılında iyi derece ile mezun oldu. 2016 yılında Türkiye Finans Katılım Bankası'nda yazılımcı olarak çalışmaya başladı. 2019 yılında ise Kuveyt Türk Katılım Bankası'nda çalışma başladı ve halen aynı kurumda çalışmaya devam ediyor.