T.C.
KIRIKKALE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

METİN MADENCİLİĞİ YÖNTEMLERİ İLE HOLLAND MESLEK KİŞİLİĞİ TİPİ ANALİZİ

Ömer DAĞISTANLI
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DOKTORA TEZİ

DANIŞMAN
PROF. DR. HASAN ERBAY

KIRIKKALE-2023

Etik Beyan Sayfası

Kırıkkale Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Ömer DAĞISTANLI

ÖZET

METİN MADENCİLİĞİ YÖNTEMLERİ İLE HOLLAND MESLEK KİŞİLİĞİ TİPİ ANALİZİ

Kırıkkale Üniversitesi

Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı, Doktora Tezi

Danışman: Prof. Dr. Hasan ERBAY

Ortak Danışman: Dr. Öğretim Üyesi Hakan KÖR

Ocak 2023, 78 sayfa

John Holland'ınMeslek Kişiliği Yaklaşımı insanların gerçekçi, sanatsal, sosyal, geleneksel ve araştırmacı olmak üzere altı kişilik tipinden biri olduğunu ve aynı şekilde altı iş çevresi olduğunu iddia etmektedir. Ayrıca kişiliğin kariyer başarısında önemli bir faktör olduğunu da iddia etmektedir. Bu yaklaşım kişilik ile iş çevreleri arasında doğrusal bir ilişki olduğunu, örneğin sosyal mesleki kişiliği baskın olan birinin sosyal iş çevresinde daha başarılı olacağını iddia etmektedir. Holland insanların üç mesleki kişilik tipine sahip olabileceğini ve her bir mesleki kişiliğin birbiri ile yakınlık ve uzaklık ilişkisi olduğunu savunur. Holland mesleki kişilikler arasındaki ilişkiyi ifade etmek için bir altıgen tasarlamıştır. Ayrıca teori insanların sadece beceri ve yeteneklerini kullanmalarına değil, aynı zamanda tutum ve değerlerini ifade etmelerine izin veren ortamlar aradıklarını da belirtir.

Diğer taraftan sosyal ağlar insanlar tarafından çokça ilgi gösterilen ve kayda değer seviyede vakit geçirilen bir alan olmuştur. Özel ve önemli günlerdeki paylaşımlar ile ya da düşünce ve fikir paylaşımları ile sıkça kullanılır olmuştur. Ayrıca sosyal ağlar benzer ilgi alanlarına sahip bireyleri birbirine bağlar ve onların düşüncelerini, hislerini, içgörülerini ve duygularını paylaşmalarına izin verir.

Bu tür paylaşımlar insanların ilgilerinin bir ifadesi olduğu da yadsınamaz. Mesleki ilgi bunlardan birisidir. Mesleki ilgi de bireylerin uyum sağlayacakları eğitim ve çalışma ortamlarını yani kariyerlerini belirlemede önemli bir unsurdur.

Bu çalışmada kişilerin mesleğinin sosyal ağlardaki yansıması incelenmiştir. John Holland'ın kariyer seçimi teorisinden, kişiliğin ve işin tutarlılığından esinlenilmiştir.

ABSTRACT

HOLLAND PROFESSIONAL PERSONALITY TYPE ANALYSIS WITH TEXT MINING METHODS

Kırıkkale University

Graduate School of Natural and Applied Sciences

Department of Computer Engineering, Ph.D. Thesis

Supervisor: Prof. Dr. Hasan ERBAY

Co-Supervisor: Assist. Prof.Dr. Hakan KÖR

January 2023, 78 pages

With the Professional Interest Approach, John Holland claims that people are one of six personality types: realistic, artistic, social, traditional and investigative, and likewise there are six work environments. He also claims that personality is an important factor in career success. This approach claims that there is a linear relationship between personality and work environments, for example, someone with a dominant social professional personality will be more successful in the social work environment. Holland argues that people can have three occupational personality types and that each occupational personality has a close and distant relationship with each other. Holland designed a hexagon to express the relationship between professional personalities. The theory also states that people seek environments that allow them not only to use their skills and abilities, but also to express their attitudes and values.

On the other hand, social networks have become an area where people are very interested and spend a significant amount of time.

It has been used frequently with sharing on special and important days or sharing thoughts and ideas. In addition, social networks connect individuals with similar interests and allow them to share their thoughts, feelings, insights and emotions.

It is undeniable that such posts are an expression of people's interests. Professional interest is one of them. Professional interest is also an important element in determining the education and working environments that individuals will adapt to, namely their careers.

In this study, the reflection of people's profession on social networks was examined. Inspired by John Holland's theory of career choice, consistency of personality and work.

İÇİNDEKİLER

Sayfa

ŞEKİLLER DİZİNİ

ŞEKİL **Sayfa**

TABLOLAR DİZİNİ

TABLO **Sayfa**

KISALTMALAR DİZİNİ

GRU	Geçitli Tekrarlayan Birim Sinir Ağı
LSTM	Uzun Kısa Süreli Bellek Sinir Ağı
STEM	Bilim Teknoloji Matematik Mühendislik Alanları
RIASEC	Gerçekçi Araştırmacı Sanatsal Sosyal Girişimci Resmi Kişilik Tipleri
SDT	Kendi Kaderini Tayin Etme Teorisi
K-NN	K-En Yakın Komşu Algoritması
UCI	Kaliforniya Üniversitesi, IrvineMakine Öğrenimi Deposu
BOW	Kelime Torbası Modeli
TF-IDF	Terim Frekansı-Ters Doküman Frekansı
CNN	Evrişimli Sinir Ağı
CAMEO	Çatışma ve Arabuluculuk Olay Gözlemleri
BERT	Transformatörlerden Çift Yönlü Enkoder Gösterimleri Modeli
DNN	Derin Sinir Ağı
WSN	Kablosuz Sensör Ağı
OSPF	Önce En Kısa Yolu Aç
RNN	Tekrarlayan Sinir Ağı
CBOW	Sürekli Kelime Torbası Modeli
IoT	Nesnelerin İnterneti
SMS	Kısa Mesaj Servisi
CPU	Merkezi İşlem Birimi
YSA	Yapay Sinir Ağı
ROC	Alıcı İşletim Karakteristiği
AUC	Eğri Altında Kalan Alan
ReLU	Doğrultulmuş Doğrusal Birimler

1. GİRİŞ

Sosyal Ağlar günümüzde en çok rağbet gören mecralar olmuştur. İnsanlar anılarını, heyecanlarını, haberlerini, bir konu hakkındaki yorumlarını veya fikirlerini sosyal ağları kullanarak bazen bir fotoğraf bazen bir metin ileçevreleriyle paylaşmaktadırlar. Bu paylaşımlar insanlar hakkında birçok fikir vermektedir.

Sosyal ağlarda yer alan bu tür paylaşımların insanların ilgilerinin bir ifadesi olduğu yadsınamaz. Mesleki ilgi bunlardan biridir ve bireylerin uyum sağlayacakları eğitim ve çalışma ortamlarını yani kariyerlerini belirlemede önemli bir unsurdur [1].

Öte yandan, benlik çocukluktan itibaren şekillenirken tercihler de oluşmaya başlar. Bu tercihler zamanla kalıcı ilgiye dönüşür. Bu ilgi alanları kariyerin ilk kısmıdır [1].Kariyer ise toplumda önemli bir olgudur. İyi bir kariyere sahip olmak yaşamı ve başarıyı etkileyen bir unsurdur. Bu iyi bir planlama gerektirir ve bu planlama gençlik döneminden itibaren yapılmalıdır [2].Ayrıca kariyer hazırlığı konusu kapsamında Gysbers bazı davranış ve becerileri gerekli görmüştür. Bunlar: (a) sosyal yeterlilik, (b) çeşitlilik becerileri, (c) olumlu çalışma alışkanlıkları, (d) kişisel nitelikler, (e) kişilik ve duygusal durumlar ve (f) girişimcilik [3].Öğrencilerin iyi bir kariyer planlaması için bu bileşenlere özellikle dikkat etmeleri, beceri ve kişiliklerini geliştirmeleri gerekmektedir [2].

Bu çalışmada da sosyal ağ paylaşımlarından mesleki kişiliğin yansıması incelenmiştir. Çalışmamız John Holland'ın “Meslek Kişiliği Yaklaşımı” teorisine dayanmaktadır. Bu yaklaşıma göre 6 tip mesleki kişilik vardır. Bunlar gerçekçi, sosyal, araştırmacı, girişimci, geleneksel ve sanatçı kişilik tipleridir. Mesleki kişiliklerle aynı adı taşıyan 6 tip de iş çevresi vardır. Bu yaklaşıma göre meslekler üç kişilik tipini içerir. Örneğin “öğretmenlik” mesleği sosyal, araştırmacı ve kişilik tiplerini içinde barındırır. İnsanlar da üç kişilik tipine sahip olabilir. Holland’a göre mesleki kişilik, kariyer seçiminde, kariyer başarısında ve kariyer tatmininde önemli bir faktördür [4].

Holland, mesleki kişilik ve çalışma ortamlarının uyumluluğunu araştırırkenbaşta mesleki tercih envanteri uygulamıştır. Bu envanterlikert tipi ölçeklerden oluşmaktadır. Anket, her iş çevresine uygun eşit sayıda soru içerir. Daha

sonra ise Holland “Kendine Yönelik Arama” ilgi envateri geliştirmiştir. Bu envanter 1300 civarı mesleğin listesinin yer aldığı bir kitapçık şeklindedir [4].

Çalışmamızda ise mesleki kişilik ve çalışma ortamının uyumunu belirlemek için yaklaşık 200 milyon aktif kullanıcısı olan, günlük 500 milyon gönderi yapılan Twitter platformu kullanılmıştır [5].Meslek sahibi ünlü kişilerin Twitter paylaşımlarından oluşan 10454 adet tweet çalışmanın veri setini oluşturmaktadır. Twitter, veri toplamak isteyen kullanıcılar için dört anahtar vermektedir. Bunlar “consumer key”, “consumer secret”, “access token”, ve “access secret” anahtarlarıdır. Bu anahtarlar kullanılarak twitter verisi toplanabilmektedir.

Çalışmamızın metodolojisi metin madenciliği yöntemleri kullanılarak elde edilmiştir. Metin madenciliği, yapılandırılmamış metin verilerinin yapısallaştırılmasıyla ilgilenir. Büyük metin verilerinden bilgiye erişen ve bunları çıkaran, veritabanlarından bilgi keşfeden, organizasyonlarda bilgi yönetimi ile veri ve bilginin görselleştirilmesini birleştiren bir çalışma alanıdır [6].

Twitter paylaşımlarından oluşan veri önce önişlem aşamasında gereksiz kelime, karakter, noktalama işaretlerinden temizlenmiş, yazım hataları düzeltilmiş, köklere ayırma işlemi gerçekleşmiştir. Bu işlemler Python programlama dili ile gerçekleşmiştir. Gereksiz kelime, noktalama işaretleri ve sayıların temizlenmesi işlemleri metin fonksiyonlarıyla yazım hatalarının düzeltilmesi Python Textblob kütüphanesi fonksiyonlarıyla, kök alma işlemi de yine Python programlama dili ile lematizasyon adı verilen yöntem ile gerçekleşmiştir.

Önişlem aşamasından sonra özellik çıkarımı aşamasında Python Keras kütüphanesikullanılmıştır. Özellik belirleme kullanılmak üzere 8871 kelimeden oluşan kelime indeksi oluşturulmuştur. Her bir kelimenin temsili için kelime gömme katmanı ile 100 sütunlu vektörler oluşturulmuştur. Sınıflama aşamasının daha başarılı olmasıiçin özellik çıkarımı aşamasında önceden eğitilmiş GloVe vektörleride kullanılmıştır.

Konu modelleme yöntemiyle de her bir kişinin tweetlerinin hangi konuları içerdiği tespit edilmiştir ve toplanantweetlerin, hesap sahiplerinin meslekleri ile ilgili olduğu belirlenmiştir.

Sınıflama aşamasında kullanılacak olan derin öğrenme algoritmaları sabit uzunlukta veri girişine izin verdiğinden istatistiki yöntemler kullanılarak her bir

satırın sabit bir sütun değerine sahip olması sağlanmıştır. Ayrıca derin öğrenme algoritmalarına sunulacak eğitim verisi ve test verisi belirleme süreci çapraz doğrulama yöntemi ile gerçekleşmiştir. Veri 5 parçaya ayrılmıştır. Her bir parça 1 kez test verisi olarak kalan 4 parça da eğitim verisi olarak sınıflama aşamasına sunulmuştur.

Sınıflama aşamasında da GRU ve LSTM derin öğrenme sinir ağları ile model oluşturulmuştur. Aşırı uymayı önlemek için seyreltme katmanı kullanılmıştır. Eğitim 50 iterasyon sürmüştür. Bu aşamada gerçekleştirilen tahmin sonuçlarına göre GRU model %94.1 oranında başarı gösterirken LSTM model %93.2 oranında başarı göstermiştir.

Bu tez çalışmasının amacı Twitter gönderileri ile Holland mesleki kişiliğin yansımasının sonuçlarının bulunmasıdır. John Holland likert tipi ölçekle mesleki kişiliği tahmin etmeye çalışmışken bu çalışma Twitter gönderileriile mesleki kişiliği tahmin etmeye çalışmıştır.

Bu araştırma ile aşağıdaki soruların cevapları aranmaktadır.

(a) Sosyal ağlar bireyin mesleki kişiliğini yansıtır mı?

(b) Sosyal ağlar bireyin mesleki kişiliği ile çalışma ortamının tutarlılığını yansıtır mı?

(c) Bireylerin paylaştıkları Twitter gönderileri meslekleri ile ilgili midir?

2. LİTERATÜR ÇALIŞMASI

2.1. Holland Teorisi ile İlgili Çalışmalar

Holland Teorisi ile ilişkili birçok çalışma yapılmıştır. Halen de bu tür çalışmalar devam etmektedir. Örneğin, Hoff ve diğerleri beş kariyer sonucu için ergenlik dönemindeki kişilerin mesleki ilgisini incelemişlerdir. Bu sonuçlar derece elde etme, prestij, gelir ve kariyer ve iş tatminidir. Çalışmada kullanılan veri, ergenliğin son dönemlerinden genç yetişkinliğe kadarki dönemde olan kişilerden 12 yıllık bir süreçte toplanmıştır. Çalışma ergenlerin mesleki ilgilerinin kariyer başarılarında önemli bir etkisi olduğunu belirlemiştir [7].

Babarovic ve diğerleri yaptıkları çalışmada ortaokul çocuklarının STEM(bilim, teknoloji, matematik ve mühendislik)'e olan ilgilerini ölçmüşlerdir. Çalışmaya 13 yaşındaki 727 öğrenci katılmıştır. Çalışma sonucunda erkeklerin Mühendislik ve teknolojiye olan ilgileri kızlara oranla daha yüksek çıkmıştır. Fen ve Matematikte ise bu fark daha az çıkmış, bilime olan ilgide ise cinsiyet farkı oluşmamıştır. Çalışma kızların STEM'e olan ilgilerinin arttırılmasına yönelik çalışmalar yapılmasını önermektedir [8].

Usslepp ve diğerleri çalışmalarında öğrencilerin genel eğitim yolu ve kariyer yolu tercihlerini büyük beş kişilik modeli ve Holland teorisi kapsamında incelemişlerdir. TOSCA ve TOSCA 10 adlı iki veri setinin kullanıldığı çalışmada Lojistik regresyon algoritması işe koşulmuştur. Çalışmanın sonucunda Holland teorisinde bahsedilen girişimci ve araştırmacı kişiliklerin eğitim yolunu, Sosyal ve Geleneksel kişilik tiplerinin de kariyer yolunu tercih ettikleri belirlenmiştir [9].

Oliveira ve diğerleri "Children's Career Expectations and Parents' Jobs: Intergenerational (Dis)continuities" adlı çalışmalarında çocukların mesleki ilgilerinin anne babalarının yaşantısındaki meslekleri ile ilgili olup olmadığını araştırmışlardır. Çalışmada iki ebeveyni de hayatta olan 108 Portekizli çocuk katılımcı kullanılmıştır. Çalışmanın sonuçları çocukların kariyer beklentilerinin ve ebeveynlerinin işlerinin, RIASEC tipolojisi ile pozitif ilişkili ve tutarlı olduğunu göstermiştir [10].

Kennon ve diğerleri "Comparing Holland and Self-Determination Theory Measures of Career Preference as Predictors of Career Choice" adındaki iki farklı

araştırmayı içeren çalışmalarında Holland RIASEC modelini Self Determination Theory (Kendi Kaderini Tayin etme Teorisi-SDT) ile karşılaştırmışlardır. Araştırma iki çalışma şeklinde yürütülmüştür. Birinci çalışmaya Missouri Üniversitesi lisans psikolojisi dersinden 246 öğrenci, ikinci çalışmaya ise yine aynı üniversitedeki kariyer merkezindeki kariyer keşif sınıflarından 92 öğrenci katılım sağlamıştır. İlk çalışmada Holland puanları ile SDT puanları arasında altıda dört oranında anlamlı ilişki bulunmuştur. Her iki çalışmada da içsel motivasyonun alt ölçeği kariyer seçimini tahmin edebilmiştir. Çalışmanın sonucunda öz-uyum değerlendirme metodolojisi kariyer seçiminde kullanılabilecek bir alternatif olduğu ortaya çıkmıştır [11].

Mintram ve diğerleri "An investigation of gender differences in Holland's circumplex model of vocational personality types in South Africa" adındaki araştırma çalışmalarında Güney Afrika Kariyer İlgi Envanteri kullanarak Güney Afrika'daki mesleki ilginin cinsiyete göre değişimini incelemişlerdir. 138 erkek, 268 kadın katılımcının yer aldığı çalışmada likert tipi 143 maddeden oluşan ölçek kullanılmıştır. Çalışmada gerçekçi, araştırmacı ve geleneksel ölçeklerde erkekler, sosyal ölçekte ise kadınlar yüksek puan almıştır [12].

Anggraini ve diğerleri çalışmalarında Holland RIASEC ölçeğinin Endonezya versiyonunun geçerlilik ve güvenilirliğini test etmişlerdir. Çalışmada katılımcılar 15 ile 27 yaşları arasındadır ve bir eğitim kurumunda çalışıyor ya da kariyer seçimi aşamasındadırlar. Ancak kariyerlerinde doyuma ulaşmamışlardır. Kullanılan ölçeğin 43 maddesinin geçerli olduğu 5 maddesinin de geçersiz olduğu belirlenmiş ve güvenirlik katsayısı 0.906 olarak bulunmuştur. Çalışmanın sonucunda Holland RIASEC ölçeğinin Endonezya versiyonu geçerli ve güvenilir bulunmuştur [13].

Rose ve diğerleri Kişisel Küresel İlgi Envanteri'nin Vietnam modelinin yapısal geçerliliği ile ilgilenmişlerdir. Mesleki ilgi ile Mesleki prestijin birleştirildiği çalışmaya 3125 üniversite öğrencisi katılmıştır. Sonuçlar beğeni ve yeterlilik için Holland Ölçeği ile veriler arasında önemli derecede uyum olduğu yönündedir. Çalışma Vietnam'da küresel ilgi envanterinin geçerliliğini doğrulamaktadır [14].

Jones ve diğerleri bir inceleme çalışması olan "Black-White differences in vocational interests: Meta-analysis and boundary conditions" adlı çalışmalarında Amerika'da ırk farklılıklarının mesleki ilgiye etkisini araştırmışlardır. Çalışmada 54

çalışmanın meta analizi yapılmıştır. Sonuçlara bakıldığında siyah amerikalıların sosyal, girişimci ve geleneksel yanlarının, beyaz amerikalıların da gerçekçi ve araştırmacı yanlarının daha baskın olduğu görülmüştür. Sanatsal açıdan ise aralarında anlamlı bir farklılık bulunamamıştır [15].

2.2. Sınıflama Algoritmaları ile ilgili Çalışmalar

Veri madenciliği çalışmalarının sınıflandırma aşamalarında önceleri K-En Yakın Komşu (K-NN), NaïveBayes, Destek Vektör Makinesi gibi algoritmaların kullanımı yoğunluktaydı. Örneğin Fanny ve diğerleri K-NN, NaïveBayes ve Destek Vektör Makinesi algoritmalarını karşılaştırmışlardır. Haber verilerinin oluşturduğu veri setinin kullanıldığı çalışmada üç tip kök bulma yöntemi kullanılmıştır. Bu yöntemlerden en iyisi wordnet olarak bulunan çalışmada sınıflama başarısının en yüksek olduğu algoritma %93,3 ile K-NN olmuştur [16].

Bir diğer çalışmada Pratama ve Sarno Twitter verilerini kullanarak kişilik tespiti yapmışlardır. Çalışmanın birinci aşamasında Twitter paylaşımları ile sınıflama yapılmıştır. NaïveBayes algoritması%60.63 oranında başarı ile diğer algoritmalar arasında en başarılısı olmuştur. Çalışmanın ikinci kısmında ise bir anket kullanılmıştır. Bu anketten alınan veriler ile yapılan kişilik belirleme sınıflandırmasında ise %65 oranında başarı elde edilmiştir [17].

Devika ve diğerleri tıp alanındaki çalışmalarında insanlarda kronik böbrek hastalığının varlığını tahmin etmişlerdir. Çalışmada K-NN, Rastgele Orman ve NaïveBayes algoritmaları kullanılmıştır. Kesinlik, geri çağırma, F1 skoru ve başarı skorunun değerlendirme ölçütleri olarak kullanıldığı çalışmada sınıflama başarı oranı en yüksek algoritma %99.84 ile Rastgele Orman Algoritması olmuştur [18].

Tıp alanındaki bir diğer çalışmada Maheshwar ve Kumar, K-NN, NaïveBayes ve Karar Ağaçları algoritmaları kullanarak memekanseri teşhisi üzerine çalışmışlardır. UCI kütüphanesinden alınan İyi huylu 458, kötü huylu 241 örneğin yer aldığı veri setinde10 özellik bulunmaktadır. Kesinlik, F1 skor, başarı ve geri çağırma metriklerinin yer aldığı çalışmada Karar Ağacı Algoritması%100 sınıflama başarısı elde etmiştir [19].

NaïveBayes, K-NN ve Karar Ağacı algoritmalarının karşılaştırıldığı "Application of NaïveBayes, Decision Tree, and K-Nearest Neighbors for

Automated Text Classification" adlı çalışmasında Ababneh, Suudi basın ajansından topladığı Arapça makaleleri sınıflandırmıştır. Çalışmada kültürel haberler, spor haberleri, sosyal haberler, ekonomi haberleri, siyaset haberleri ve genel haberler konularında veriler toplanmıştır. Sonuç olarak çalışmada %88 F1 skoru ile NaïveBayes Algoritması en yüksek değeri elde etmiştir [20].

2.3. Özellik Çıkarımı Yöntemleri ile ilgili Çalışmalar

Metin madenciliği çalışmalarında yapılan çalışmalarda birçok özellik çıkarım yöntemi ve sınıflandırma ve kümeleme algoritmaları kullanılmaktadır. Sınıflandırma başarısının yüksek olması için özellik çıkarımının iyi yapılması gerekmektedir. Özellik çıkarımı yöntemlerinin bazıları kelime torbası (BoW) modeli, Sözlük Yöntemi ve Terim frekansı-ters döküman frekansı modeli (TF-IDF)'dir. Öte yandanWord2Vec, GloVe, ve FastText özellik çıkarım aşamasında kullanılan kelime gömme metotlarıdır.

Dharma ve diğerleri yaptıkları çalışmada kelime gömme metodlarındanWord2Vec, GloVe ve FastText yöntemlerini karşılaştırmışlardır. Sınıflama için Evrişimli Sinir Ağı (CNN) kullanılan çalışmada veri seti olarak 20 haber başlığı altında 19.977 haberin yer aldığı UCI KDD arşivi kullanılmıştır. Üç kelime gömme metodu ile birbirine yakın sınıflama başarısı elde edilen çalışmadaFastText %97.2 ile en yüksek sonucu vermiştir [21].

Parolinve diğerleri Word2Vec ve GloVe kelime gömme metodlarını kullandıkları çalışmalarında siyasi haber makalelerini incelemişlerdir. Çalışmada CAMEO(Çatışma ve Arabuluculuk Olay Gözlemleri) ontolojisinden yararlanılmıştır.250 bin makalenin incelendiği çalışmada önişlem aşamasından sonra 107.475 cümle ile çalışılmıştır.Word2Vec ile %75.1 GloVe ile %65.8 oranında başarı elde edilmiştir [22].

Son zamanlarda BERT yöntemi de özellik çıkarımı için kullanılan yeni bir yöntem olarak karşımıza çıkmaktadır. Dil modelleme ve sonraki cümle tahmininde kullanılan bu yöntem kelimeler için bağlamsal yöntemleri de öğrenebilmektedir. Bir kelimenin birden fazla anlamı olduğu durumda kelime gömme metotları tek bir vektör üretirken BERT cümleye göre değişen birden fazla vektör üretebilmektedir. BERT yönteminin kullanıldığı bir çalışmada Uday ve diğerleri BoW, Word2Vec,

TF-IDF özellik çıkarımı yöntemleri ile BERT yöntemini karşılaştırmışlardır. Covid 19 pandemisi ile ilgili 3 ayrı veri setinin bulunduğu çalışmada en yüksek sınıflama başarısı %90 oranında BERT ile elde edilmiştir [23].

BERT modelinin çalışıldığı bir diğer çalışmada Mozafari ve diğerleri Twitter verileri kullanarak nefret söylemi tespiti modülü ve önyargı azaltmamodülü geliştirmişlerdir. BERT tabanlı çalışmada Bi-LSTM ve CNN sınıflandırma için kullanılmıştır. Nefret söylemi tespiti çalışmasında CNN sinir ağı ile %92'ye varan F1 skoru elde edilmiştir. Önyargı azaltma modülü çalışmasında ise %91 oranında F1 skor elde edilmiştir [24].

2.4. Derin Öğrenme Algoritmaları ile ilgili Çalışmalar

Sınıflandırma aşamasındason zamanlarda daha çok derin öğrenme algoritmaları kullanılmıştır. Mohanty ve diğerleri kablosuz sensör ağının (WSN) füzyon merkezinde enerji verimliliği ve optimum yük dengesini sağlamak için derin öğrenme tabanlı bir model önermişlerdir. Çalışmadaki model RNN-LSTM uzun kısa süreli bellek içermektedir. Karşılaştırma için OSPF ve DNN modeller kullanılmıştır. Sinyalleme yüküne ve ortalama gecikmeye göre sonuçlar LSTM'nin diğer modellerden daha başarılı olduğunu ortaya koymaktadır [25].

Yang ve diğerleri bilgisayar oyunlarının pazarlamasında kullanılmak üzere bir derin öğrenme modeli önermişlerdir. Oyuncu davranışlarının incelendiği çalışmada GRU modeli kullanılmıştır. Sonuçlar derin öğrenme modeli sayesinde otomatikleştirilmiş bir pazarlama hizmetinin sağlanabilirliğini, maliyetin düşürülmesini ve pazarlama hizmetinin en doğrusuna ulaşabilir olduğunu belirlemiştir [26].

Çalışmalarında Bangla dilindeki farklı kategorilerdeki çevrimiçi gazetelerde yer alan makaleleri inceleyen Banik ve Rahman doğal dil işlemenin alt alanlarından biri olan varlık tanıma sistemini çalışmalarında kullanmışlardır. GRU derin öğrenme ağının kullanıldığı çalışmada %69 F1 puanı elde edilmiştir. Araştırmacılar veri setinin artırılması ve birlikte kullanılacak kelime gömme metodları ile daha başarılı sonuçlar elde edilebileceğini öngörmüşlerdir [27].

Guo ve diğerleriderin öğrenme modelini kullandıkları çalışmalarında köprü inşaatları sürecinde ve bakım sürecinde sensörlerden elde edilen veriler ile çalışıp bu

verilerden çıkarımlar elde etmişlerdir. Çalışmada Kohonen sinir ağı ile erken uyarı ve aykırı tespitinde LSTM sinir ağını ise sapmaların tahmin edilmesinde kullanılmıştır. LSTM model, sapma değeri tahmininde, %99'a kadar başarı göstermiştir. Çalışmada LSTM'nin Köprü sağlığı izleme sistemlerinde izleme parametrelerinin tahmininde kullanılabileceği öngörülmüştür [28].

Massaro ve diğerleriLSTM ve GRU sinir ağları ile küresel dağıtım sistemi için bir karar destek modeli oluşturmuşlardır. Temel performans göstergelerinin tahmin edildiği çalışmakarar destek sistemi, iş zekası stratejileri ile ilgili endüstri faaliyetlerini içeren bilgi tabanının artırılmasını önermektedir [29].

Sun ve diğerleri "Short-Term Building Load Forecast Based on a Data-Mining Feature Selection and LSTM-RNN Method" adlı çalışmalarında bina yük tahmini yapmaya çalışmışlardır. LSTM sinir ağının kullanıldığı çalışmada tanıtılan yöntemin uygun girdi seçimi ve büyük tahmin doğruluğu ile yüksek başarı göstereceği önerilmiştir [30].

Zhang ve diğerleri "Improved Dota2 Lineup Recommendation Model Based on a Bidirectional LSTM" adlı çalışmalarında dünyaca ünlü e-spor oyunu Dota2'nin Bidirectional-LSTM ile kahraman karakterini tahmin etmektedirler. Çalışmada özellik çıkarımı için kelime gömme metodu olan Word2Vec CBOW kullanılmıştır. Deney sonuçları önerilen beş kahramanın ortalama doğruluk oranını %67,74 olarak belirlemiştir [31].

Ryaz ve Ganapathy çalışmalarında saldırı tespit sistemi modellemişlerdir. Çalışmada izinsiz ağa girişi tespit etmek amacıyla özellik çıkarımı aşamasında koşullu rastgele alan, doğrusal korelasyon katsayısı ve evrişimli sinir ağı(CNN) algoritması kullanılmıştır. Önerilen model %98,8 oranında tespit doğruluğu elde etmiştir [32].

Liu ve diğerleri "Spatio-Temporal GRU forTrajectoryClassification" adlı çalışmalarında derin öğrenme algoritması kullanarak Mekansal-zamansal yörünge sınıflandırması yapmışlardır. Bu bağlamda mekansal-zamansal GRU adında bir sınıflandırıcı önerilmiştir. Sonuç olarak çalışmada literatürün aksine hem mekansal hem de zamansal yörünge sınıflandırmasının başarılı olabileceği gösterilmiştir [33].

Smys ve diğerleri göl ve baraj gibi su depolanan yerlerin sensörlerle yönetilmesi üzerine bir model önermişlerdir. Çalışma alanı alıcı göl veya baraj ile

gönderici göl veya barajın olduğu sistemlerdir. Önerilen modelde evrişimli sinir ağı CNN kullanılmış ve alıcı göl veya barajın su seviyesinin tahmini ile taşkınların önüne geçilmesi hedeflenmiştir. Sonuçta önerilen yöntem gerçek zamanlı baraj ve göllerde kullanılabilirliği kanıtlanmıştır [34].

LSTM ve CNN'nin hibrid modelinin önerildiği bir diğer çalışmada da Zhang ve diğerleri iki farklı web sitesinden 10 bin film yorumunu veriseti olarak kullanmışlardır. Sınıflama için LSTM ve CNN algoritmalarının ayrı ayrı da kullanıldığı çalışmada en yüksek başarı skorunu %91.17 ile hibrid model elde etmiştir [35].

3. GENEL BİLGİLER

3.1. Kariyer Seçimi

Kariyer seçimi insan hayatındaki belki de en önemli seçimlerden biri olmaktadır. Birçok zorluğu içinde barındıran kariyer seçimi, kişinin duygu durumundan, hayata bakışından, yaşam stilinden ve ekonomik durumundan etkilenen, kariyer seçimi sonucu isekişinin topluma katkısını ve toplumdaki statüsünü belirleyen bir öneme sahiptir [36]. Ayrıca kariyer seçiminde, kişisel yetenekler, kültür, alınan eğitim, ailenin geçmişi ve kişinin yaşıtları arasındaki başarı durumu gibi birçok faktör etkilidir [37,38,39].

Kariyer seçiminin bu denli önemli olması insanların hayatlarının farklı dönemlerinde bu konuyla meşgul olmalarına [36,40] ve kararlarını düşünerek, planlayarak, strateji yaparak, sistematik bir şekilde vermelerine sebep olmaktadır [37].

Öte yandan kişiler kariyer seçimlerini her zaman düşünerek ve bilinçli bir şekilde yapmazlar. Brimrose çalışmasında toplumun sadece %25'inin kariyer seçiminde stratejikve akla uygun kararlar verdiğini belirtmiştir [41].Gladwell de kişinin planlamadan yaptığı kariyer seçiminin daha etkili olduğu görüşündedir [42]. Miltchellde bu konuda sezginin ön planda olduğunu, yoldaki işaretlere göre karar vermenin etkili olacağını öne sürmektedir [37,43].

Kariyer seçiminin önemi ve zorluğu, bu alanda uygulamaların, kariyer gelişim teorilerinin ve hatta kariyer danışmanlığı vb. gibi mesleklerin doğmasına neden olmuştur.

Araştırmacılar geçmişi açıklamak, yetenekleri dikkate alabilmek ve bu gibi kriterlerle gelecekteki kariyer seçimini doğru tahminine yardımcı olması için bir teori bilgisinin gerekli olduğunu tartışmaktadırlar [37,44].

Korkut-Owen ve diğerleri, kariyer teorilerinin karmaşık davranışları anlaşılabilir durumlara, mantıklı örüntülere ve doğru tahminlere dönüştürebileceğini öne sürmektedir [37,45].

Krumboltz'a göre ise kariyer teorisi gereksiz ayrıntıları atlamamıza ve büyük resmi görmemize olanak sağlar [37,46].

3.2. Holland Meslek Kişiliği Yaklaşımı

Günümüzde kariyer seçimi büyük bir öneme sahiptir. Bu kapsamda kariyer üzerine yapılan araştırmalar ile teoriler üretilmektedir. Teoriler bir kariyer danışmanı görevini üstlenmekte ve kişisel ilgi üzerine odaklanmaktadırlar. Bu teorilerin başlıcalarından bir tanesi de John Holland tarafından geliştirilmiş olan "Holland Meslek Kişiliği Yaklaşımı"dır.

John Holland, geliştirdiğiMeslek Kişiliği Yaklaşımıile kişinin ilgi alanlarının kariyer seçiminde etkili olduğu görüşünü benimsemiştir. Bu yaklaşım kişilerin ilgi duyduğu alanlara ve uygun çalışma ortamına yönelmesinin başarı, istikrar ve iş tatmininde olumlu sonuçlar ortaya çıkardığı ve ilgilerin meslek seçiminin belirleyicisi olduğu esasınadayanır [8,9].

Bu yaklaşıma göre kişiler 6 tip meslek kişiliğine sahiptir. Bunlar gerçekçi, sosyal, araştırmacı, girişimci, geleneksel ve sanatsal kişiliklerdir. Holland Teorisi olarak adlandırılan yaklaşım kişilik tiplerinin baş harflerinin yan yana yazılması ile oluşan RIASEC modeli diye de ifade edilebilmektedir.

Teoriye göre iş çevreleri yani meslekler de bu kişiliklere uyumlu olacak şekilde kategorize edilmiş ve her bir mesleki kişilik ile kategorize edilmiş meslekler arasında bir eşleşme söz konusudur.

Gerçekçi kariyerler arasında aşçı, sporcu, tamirci itfaiyeci, elektrikçi, boyacı gibi meslekler yer alırken, sanatsal kariyerler arasında ressam, oyuncu, tasarımcı gibi meslekler, araştırmacı kariyerler arasında akademisyen, matematikçi, fizikçi gibi meslekler, sosyal kariyerler arasında öğretmen, avukat, politikacı gibi meslekler, girişimci kariyerler arasında insan kaynakları temsilcisi, işadamı gibi meslekler ve geleneksel kariyerler arasında devlet memuru, istatistikçi, muhasebeci gibi meslekler yer almaktadır [11].

Holland Teorisi üzerine yapılan çalışmalar gösteriyor ki sosyal, gerçekçi ve geleneksel kişilik tiplerine sahip insanların eğitim istekleri düşük, araştırmacı, sanatsal ve girişimci kişilik tiplerine sahip insanların da eğitim istekleri yüksektir [9].

Bu yaklaşıma göre meslekler birden fazla meslek kişiliğini içinde barındırabilirler. Örneğin öğretmenlik mesleği hem sosyal hem araştırmacı hem de

geleneksel kişilik tiplerini kapsar. Ayrıca teori bireylerin biri daha baskın olacak şekilde üç tip meslek kişiliğini yansıtabileceğini savunur [4].

3.2.1. Gerçekçi Kişilik Tipi

Gerçekçi Kişilik tipleri teknik ve pratik yetenek gerektiren aktivitelere ilgi duyan uygulamalı problemler çözen, ahşap malzemeler, makine ve aletler gibi somut işlerle uğraşmayı seven kişilik tipleridir. Aşçı, oto tamircisi, boyacı, teknisyen, tesisatçı, sporcu gibimeslekler bu kişilik tipinin başarılı olabileceği meslekler arasında yer almaktadır [47].

3.2.2. Araştırmacı Kişilik Tipi

Araştırmacı kişilik tipleri, araştırma yapmayı, bir konu üzerinde düşünmeyi, fikir üretmeyi, deneyler yapmayı seven ve zihinsel aktivitelere ilgi duyan kişilik tipleridir. Bilimselliğe yakın kişiliklerdir. Bilim insanı olmak isteyen kişiler bu yaklaşıma göre bu kişiliğe sahip bireylerdir [47].

3.2.3. Sanatsal Kişilik Tipi

Sanatsal kişilik tipleri orijinal fikirler üretmeyi iyi becerirler. Hayal güçleri yüksek olan bu kişilikler edebiyat, sinema, tiyatro, müzik, resim, heykel gibi dallarda daha başarılı olabilmektedirler. Yaratıcı bakış açısına sahip olan sanatsal kişilikler için ressam, heykeltıraş, dansçı, oyuncu, müzisyen, fotoğrafçı, tasarımcı gibi meslekler uygun mesleklerdir [47].

3.2.4. Sosyal Kişilik Tipi

Sosyal kişilik tipleri, iletişim kurmayı iyi becerirler. Yardımsever ve fedakârdırlar. Birlikte çalışma konusunda en doğru kişilik tipidir. Sosyal sorumluluk projelerinde yer almayı severler. Ayrıca insanlara bir şeyler öğretme konusunda da başarılıdırlar. Dışa dönük bir yapıya sahip bu kişilikler için öğretmen, avukat ve hemşire gibi meslekler uygun mesleklerdir [47].

3.2.5. Girişimci Kişilik Tipi

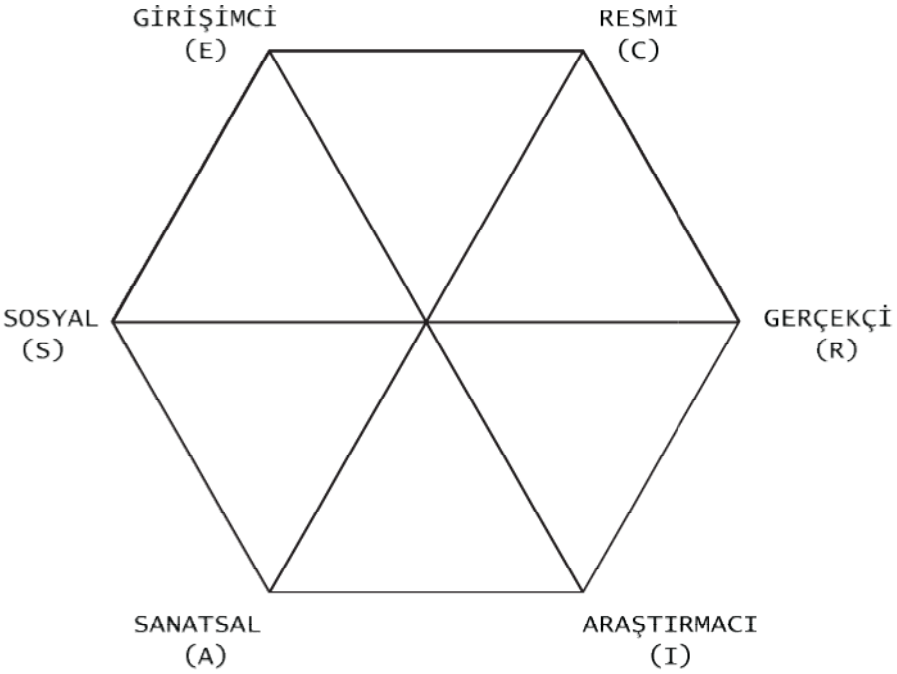
Girişimci kişilikler lider kişilerdir. Enerjik ve hırslı yapıları ile dikkat çekerler. Risk alabilirler. İnsanları yönlendirme ve ikna etme konusunda başarılıdırlar. Dışa dönüktürler. Girişimci olmayı gerektiren mesleklerde başarılı olabilirler. Teoriye göre bu kişilik tipleri için satış pazarlama, siyaset, iş adamı gibi meslekler ve gazetecilik mesleği uygundur [47].

3.2.6. Geleneksel Kişilik Tipi

Geleneksel kişilikler genellikle rutinleri takip eden kişilerdir. "Görev adamı" tabiri ile anılan insanlardır. Titiz, plan yapan ve ayrıntıya önem veren bir yapıya sahiptirler. Devlet dairelerinde memurluk gibi meslekler bu kişilikler için en uygun mesleklerdir. Örneğin muhasebeci, bankacılık, sekreterlik gibi meslekler bu kişilik tipleri için uygundur [47].

3.2.7. Holland Altıgeni

Holland geliştirdiği meslek kişiliği yaklaşımıyla kişilik tiplerinin ve iş çevrelerinin birbiri ile yakınlık ve uzaklık ilişkisini gösteren dairesel bir altıgen tasarlamıştır [8].



Şekil 1 Holland Altıgeni

Şekil 1'de gösterilen Holland Altıgeni olarak bilinen bu tasarıma göre meslek kişiliklerinden yakın olanlar birbiri ile daha yüksek uzak olanlar ise daha düşük seviyede ilişkiye sahiptir. Örneğin bitişik olan girişimci ile sosyal birbirine daha benzerken girişimci ile sanatsal daha az benzerdir. Girişimci ile en az benzer olan ise araştırmacıdır [8].

3.3. Twitter

Sosyal ağların en popülerlerinden biri de Twitter'dır. Twitter aktif 192 milyon kullanıcıya sahiptir. Bu sayının %63'ü 35 ile 65 yaş aralığındadır ve cinsiyet olarak %66'ya %34 erkek kullanıcı üstünlüğü bulunmaktadır. Günlük 500 milyon tweet atılmakta ve Twitter internetin SMS'si olarak kabul edilmektedir. Twitter haberlerin çıkış noktası ve son dakika haberlerinin iletildiği en önemli mecra olarak da kabul görmüştür. Bu da Facebook'daortalama geçirilen 4,96 dakikaya nazaran Twitter'daki sürenin 3,39 dakikada kalmasının açıklaması olarak görülebilir [48].

Twitter paylaşımları genellikle herkese açıktır. Herkes fikirlerini ve düşüncelerini paylaştığı gibi başka kullanıcıların paylaşımlarını da retweet olarak paylaşabilmektedir. Bazı konular günlük "trend topic" yani gündemdeki konu olabilmektedir.

Twitter özellikle politika, ekonomik ve sosyal konuların tartışıldığı bir platform halini de almıştır. Gündelik haberlerin paylaşılması bir haber sitesi gibi görülmesinisağlamıştır. Ayrıca veri toplanmasına izin veren yapısı olması dolayısıyla araştırmacıların ilgisini çeken bir platform olmaktadır. Twitter veri toplamak isteyen kullanıcının başvurusunu onayladıktan sonra ona dört anahtar vermektedir. Bunlar "consumer key", "consumer secret", "access token", ve "access secret" anahtarlarıdır. Kullanıcıya özel olan bu anahtarlar kullanılarak Twitter ile kullanıcı hesaplarının ya da belirli bir konudaki tweet gönderileri toplanabilmektedir.

3.4. Veri Madenciliği

Veri madenciliği, kalıpları belirlemek, bilgi keşfetmek, özellikleri belirleme, eğilimleri belirleme, fikir edinme gibi işlemleri büyük veri içinde uygulamaya verilen genel ifadedir [49].

Birçok endüstride veri madenciliği önemli bir role sahiptir. Bu alanda bir şirketin sattıkları ürünlerin gelecekteki satış miktarının tahmini, kaynak miktarı tahmini, risk analizi, müşterilerin firmaya ve ürünlere bakışı, dolandırıcılığı tespit etmek gibi birçok işlem yapılabilir [49].

Veri madenciliğinde uygulanan yukarıda bahsedilen işlemler ile şirketin rekabet gücü artar, firmanın gelişmesi ve büyümesi sağlanır, ekonomik kayıplar en aza indirgenir ve kurumların müşteri portföyü genişler [49].

Veri Madenciliği ile uğraşan kişilere "Veri Bilimcisi" denmektedir. Bir Veri bilimcisinin bilgi çıkarımında bulunmak, veriler arasındaki kalıpları ve ilişkileri belirlemek gibi sorumlulukları vardır. Veri bilimcisi bu gibi görevleri gerçekleştirebilmek için Hadoop, Pig, Hive, Spark, MapReduce gibi teknolojileri ya da Python, R dili gibi programlama dillerinikullanarak tahmine dayalı makine öğrenmesi modellemesi yapar [50].

Veri madenciliği, Örüntü Madenciliği, Metin madenciliği, Web Madenciliği, Fikir Madenciliği gibi kategorilerde uygulanmaktadır. Pazarlama, bankacılık, sigortacılık, mühendislik, sağlık, ulaştırma, güvenlik vb. birçok alanda veri madenciliği çalışmalarında yararlanılmaktadır.

Metinsel verilerin kullanıldığı çalışmalarda birden fazla alana ayrılmaktadır. Bunlara metin madenciliği, doğal dil işleme ve duygu analizi örnek verilebilir. Bu alanlar birbirleri ile hibrit olarak da kullanılabilmektedirler. Metin madenciliği metinsel veriler ile ilgili çıkarımlar elde etmeyi hedeflerken doğal dil işleme insan dili ile bilgisayar dilini yakınlaştırmayı hedeflemektedir. Duygu analizi ise metinlerdeki duyguyu ortaya çıkarmayı amaçlamaktadır.

3.5. Doğal Dil İşleme

Doğal dil işleme yapay zekânın bir alt dalı olarak insan dilindeki anlamın bilgisayarlar tarafından ortaya çıkarılmasıyla ilgilenmektedir. Doğal dilin anlaşılmasını sağlayan teori ve yöntemleri kapsar. Doğal dil işlemenin ilk önemli gelişimi 1957 ile 1970 arasında gerçekleşmiştir. Bu tarihler arasında doğal dil işleme yapay zekâ ile birleştirilmiştir. Fakat 1970'e yaklaşırken doğal dil işlemede istatistikî yaklaşım ile korpusun oluşturulması konusunda yaşanılan sorunlar bu alana güveni

sarsmıştır. Bu alandaki uygulamaların istenilen seviyeye gelemeyeceği kanısına sebep olmuştur [51].

İkinci önemli gelişim 1971 ile 1993 yılları arasındadır. Araştırmacıların doğal dil işleme alanlarından konuşma tanıma alanında Gizli Markov Modeline dayalı başarıları ve 1980'lerin başı ile söylem analizinin gelişmesi doğal dil işlemenin ivme kazanmasını sağladı [51].

1990'lı yılların ortalarına doğru bilgisayar hızındaki ve depolamadaki hızın artışı doğal dil işlemenin ticari gelişimini sağlamıştır. Ayrıca internetin ticarileşmesi de doğal dil işlemedeki talebi artırmıştır. YoshuaBengio 2001 yılında doğal dil işleme alanında ilk ileri beslemeli sinir ağı modelini önermiştir [51].

Günümüzde özellik çıkarımı yöntemleri olarak karşımıza çıkan BOW yöntemi, TF-IDF modeli, N-gram yöntemleri,Word2Vec, GloVe, FastText gibi kelime gömme metotları, konu modelleme olarak bilinen Gizli Dirichlet Tahsisi, dil modelleme ve sonraki cümle tahmininde kullanılan BERT yöntemi doğal dil işleme alanında kullanılan yöntemlerdir.

3.6. Duygu Analizi

Duygusal durum ve yargı çalışmalarında ortaya çıkan duygu analizi kelime, kelime öbekleri, cümleler ve bazen de tüm belgenin duygu ifade ettiği durumlarla ilgilidir. Kelime tabanlı daha çok tercih edilir. Duygudoğrudan veya örtükolabilir [52].

Duygu analiziproblemin karmaşıklığı sebebiyle üç görevi kapsar. İlki metnin öznel veya nesnel olması anlamındaki fikir tespitidir. Fikir tespiti metindeki sıfatlar ile belirlenebilir. İkincisi ise duygu polaritesinin sınıflanmasıdır. Duygu analizinde polarite genellikle pozitif ve negatif kutupluluk olarak sonuçlanır [52].

Yukarıdaki iki görev için farklı düzeylerde uygulama yapılabilir. Bunlar kelime düzeyi, ifade düzeyi veya cümle düzeyi olabilir. Sözlükler, kelime torbaları, n-gram kelime düzeyinde iken konuşma bölümü etiketleme (part-of-speech-tagging) ise tamlamama cümle tabanlı olmaktadır [52].

Duygu analizinin üçüncü görevi ise görüşün hedefinin belirlenmesidir. Ürün analizlerinde bu çok önemli bir etmendir ve bazen birden fazla hedef vardır. Bu durumda nesneler değerlerini en iyi temsil eden bir sıraya konulabilir [52].

3.7. Metin Madenciliği

Çalışmada Metin madenciliği yöntemleri kullanılmıştır. Metin madenciliği yapısal olmayan metin verilerinin yapısal hale getirilmesi ile ilgilenir. Büyük metin verilerinden bilgiye erişen ve bilgi çıkaran, veri tabanlarından bilgi keşfeden, organizasyonlarda bilgi yönetimini ve veri ile bilginin görselleştirilmesi aşamalarını birleştiren bir mimaridir [6].

Metin madenciliği temel adımları;

a) Veri toplama

b) Önişleme

c) Özellik Çıkarımı

d) Sınıflama

e) Değerlendirme ve Yorumlamadır.

3.7.1.Metin Madenciliğinde Önişlem Aşaması Yöntemleri

3.7.1.1. Tokenizasyon İşlemi

Tokenizasyon cümleleri daha küçük birimlere yani ‘token’lere bölme işlemidir. Doğal dil işleme ve metin madenciliğinde verileri simgeleştirme tam anlamıyla metni bölmeyi gerektirir. Bu işlem metinde eşleşen kalıpları bulmamızı sağlar. Python “nltk” kütüphanesi ile kelime tokenizasyonuve cümle tokenizasyonuyapılabilmektedir [53].

Örnek olarak:

Cümle: “Ayşe hastalandığı için, bugün okula gitmedi.”

Tokenizasyon işlemi sonrası: [“Ayşe”, “hastalandığı”, “için”, “,”, “bugün”, “okula”, “gitmedi”, “.”]

3.7.1.2. Durak kelimelerinin Kaldırılması

Durak kelimeleri her metinde bulunma ihtimali yüksek kelimelerdir. Bunlar bağlaçlar, edatlar, gibi kelimelerdir. Bu kelimelerin metindeki varlığı sınıflama için bir fayda sağlamaz. Yani sınıflama aşamasında tahmini kolaylaştırma açısından bir rolü olmaz. Bu yüzden kaldırılması gerekir. Bunlar gibi sayıların ve gereksiz kelimelerin de metinden kaldırılması uygun olur [55].

3.7.1.3. Kök Alma(Stemming) İşlemi

Kök alma işlemi kelimenin ön ek ve son eklerinin çıkarılma işlemidir. Bu kelime bir isim ise çoğul eki almış olabilir ya da bir fiil ise çekim ekli almış olabilir. Kök alma işlemi ile bu çoğul eklerini ya da çekim eklerini kelimeden çıkarmak içindir. Bu işlem İngilizce kelimeler için Python dilinde bulunan birçok eklenti ile yapılabilmektedir. Türkçe kelimeler için kök alma işleminde ise Zemberek [54] kütüphanesi yaygın olarak kullanılmaktadır.

Bu işlemde kök morfolojik olmak zorunda değildir. Bu işlemin amacı özellik çıkarımı aşamasındaki özellik sayısını azaltmak, bir kökten türemiş kelimelerin tek bir özellikte birleşmesini sağlamaktır. Bu şekilde sınıflama performansı da artmış olacaktır [55].

3.7.1.4. Lemmatizasyon İşlemi

Lemmatizasyon işlemi kök alma işlemi gibi köke indirme işini yapmaktadır. Ama kök alma(stemming) işleminden farklı çalışma yöntemi vardır. Özellikle İngilizce kelimelerde kök alma(stemming) işlemi sadece sondaki eki almaya yöneliktir. Bu işlem bu yüzden kaldırılmaması gereken harfleri silebilir [56].

Lemmatizasyon işlemi ise kelimeleri morfolojik olarak inceler. Hangi türde bir kelime ise o türe göre ekleri kaldırır. Kelime birfiil ise ve çekimli bir kelime ile işlem yapılıyorsa lemmatizasyon çekim eklerini kaldırır. Kelime bir isim ise ve çoğul eki almış ise lemmatizasyon işlemi çoğul ekini kaldırır. Bu işlemin de amacı özellik çıkarımı aşamasındaki özellik sayısını azaltmak, bir kökten türemiş kelimelerin tek bir özellikte birleşmesini sağlamaktır [56].

3.7.2. Metin Madenciliğinde Özellik Çıkarımı Yöntemleri

Metin madenciliğinde özellik çıkarımı, önişleme süreci ile gereksiz tüm öğelerinden temizlenen metinsel verilerin belirleyici bir özellik doğrultusunda sayısal hale getirilmesi işlemidir. Literatürde birçok özellik çıkarımı yöntemi bulunmaktadır.

3.7.2.1. Kelime Torbası Yöntemi (BoW)

Kelime Torbası Model metinde yer alan kelimelerin dilsel ve anlamsal yönüne bakmadan sadece sıklığı ile ilgilenir. Eşsiz kelimeler için oluşturulan kelime torbasındaki her bir kelimenin metin içinde kullanılma sıklığına göre vektör uzayı modeli elde edilir [57].

Daha detaylı anlatmak gerekirse öncelikle verilerin bulunduğu metinsel ifadeler kelime kelime ayrılır. Bu kelimelerden eşsiz bir sözlük yani kelime torbası oluşturulur. Oluşturulan kelime torbasındaki kelimeler ile her veri satırı için matris oluşturulur. Örnek aşağıda gösterilmiştir.

Cümle 1: “Bugün çok ders çalışmam gerekiyor”

Cümle 2: “Eğer bugün hava güzel olursa biraz gezeceğim”

Cümleler önişlemden geçtikten sonra aşağıdaki gibi olacaktır.

Cümle 1: “bugün ders çalışma gerek”, Cümle 2: “bugün hava güzel olur gez”

Kelime Torbası: “bugün,ders,hava,çalışma,gerek,güzel,olur,gez”

	bugün	Ders	hava	çalışma	gerek	güzel	olur	gez
Cümle1	1	1	0	1	1	0	0	0
Cümle2	1	0	1	0	0	1	1	1

Cümle 1= {1,1,0,1,1,0,0,0}

Cümle 2= {1,0,1,0,0,1,1,1}

3.7.2.2. Terim Frekansı – Ters Doküman Frekansı (TF-IDF)

TF-IDF yöntemi, bir belgedeki kelimenin sayısının diğer belgelerdeki sıklığının ters orantılı olarak ifade edilmesidir. Bu yöntemde iki ayrı terim kullanılır. Birincisi terim frekansı (TF), ikincisi ise ters doküman frekansı(IDF)'dır.

Terim frekansı(TF), bir terimin bir belgede bulunma sayısı olarak ifade edilir. Örneğin, 10.000 kelime bulunan bir belgede "**yönetim"** kelimesi 50 kere geçmiş olsun. Ve aynı kelime diğer 200 belgeden 10'unda da geçiyor olsun. Bilinir ki çok kelime sayısına sahip belgelerde bir kelimenin bulunma ihtimali az kelime sayısına sahip belgelerdekine göre daha fazladır. Büyük ve küçük belgelerin bir arada olduğu çalışmalarda bu durum bir sorun teşkil etmektedir. Terim Frekansı hesaplanırken bu sorun göz önüne alınır ve belgedeki tüm kelimelerin sayısı istenen kelimenin sayısına bölünür. Bu şekilde TF değeri hesaplanır. Örneğimize göre;

TF=50/10000=0.005 olarak hesaplanır.

Ters Doküman Frekansı da (IDF) bir kelimenin diğer belgelerin kaçında olduğu ile ilgilidir. Bu kelime "ve" gibi çok kullanılan bir kelime ise bu kelimeyi önemsiz az kullanılan bir kelimeyi de önemli yani belirleyici kelime durumuna getirir. Ağırlığı daha fazla olur. Örneğimize devam edecek olursak,

IDF= log(200/10)=1.301 olarak hesaplanır.

Yukarıdan anlaşılacağı üzere TF ile bir terimin belgedeki fazlalığı ile sıklığının daha çok olacağı, IDF ile bir terimin diğer belgelerde bulunma sayısı ile öneminin artması ifade edilmektedir. TF-IDF değeri de TF değeri ile IDF değerinin çarpılması ile elde edilmektedir [58].

TF-IDF= 0.005*1.301=0,006505 olarak hesaplanır.

3.7.2.3. Word2Vec

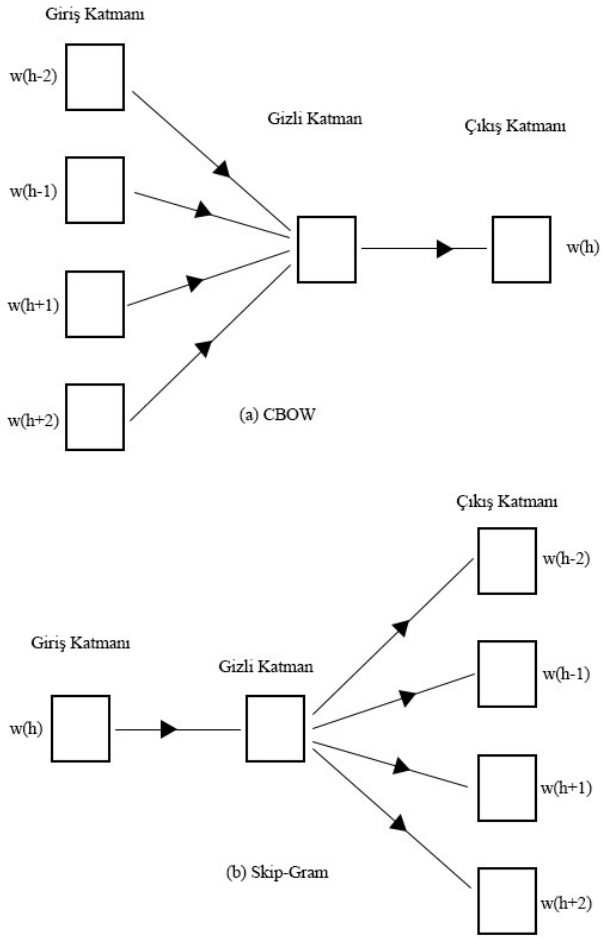
Word2Vec Google tarafından oluşturulmuş olan bir kelime gömme metodudur. Metin verileri vektörel olarak ifade edilir. Word2Vec kullandığı yapay sinir ağı sayesinde kelimeler arasındaki anlamsal ilişki seviyesini de belirleyebilmektedir [59].

Word2Vec kelime gömme metodunun iki öğrenme yöntemi bulunmaktadır. Bunlar CBOW ve Skip-Gram yöntemleridir. Her iki yöntemde kelime tahmininde bulunur. Girdi olarak metin alır çıktı olarak kelime vektörü üretir. Ürettiği kelime

vektörü kelimeyi temsil eder ve bu öğrenme yoluyla gerçekleşir. Kelime vektörünün öğrenildiğini ise yakın kelimeleri tahmin etmesiyle anlayabiliriz. Yöntem yakın kelimeleri bulmak için mesafe aracını kullanır [60].

Word2Vec kelime gömme metodunda birbirine yakın kelimeler birbirine yakın değerdeki vektörlerle ifade edilir [61].

Her iki model de kelimelerdeki bağlamlara dikkat etmektedir. CBOW yöntemi bağlamlardan hedef kelimeyi tahmin etmeye çalışır. Skip-Gram yöntemi ise hedef kelimenin etrafındaki bağlam kelimeleri tahmin etmeye çalışır [62].



Şekil 2 Word2Vec CBOW ve Skip-Gram Metodları

3.7.2.4. FastText

FastText, Facebook yapay zeka laboratuvarlarında oluşturulan kelime gömme metodudur [63]. Yapay sinir ağı kullanır. En önemli özelliği hızlı ve verimli olmasıdır. Facebook standart çok çekirdekli bir CPU ile 10 dakikada 1 milyardan fazla kelime üzerinde eğitilebileceğini savunmaktadır [64] .

FastText verimliliği ve hızı arttırmak için skip-gram yöntemini kullanır [65].Mikolav ve arkadaşlarına göre de Skip-Gram yöntemi CBOW yönteminden daha başarılıdır [66].N-gramların eşleşmesinin verimli ve hızlı olması için karma işlevi kullanılır [67]. Ayrıca çok sınıflı problemlere uygularken Huffman kodlama ağacına dayalı hiyerarşik softmax fonksiyonu kullanılabilir. Bu, hesaplamanın karmaşıklığını azaltmak için yapılır [68].

3.7.2.5. GloVe

Son dönemde kullanılan vektör uzay modelleri kelimeleri temsil eden vektörlerin belirlenmesinde önemli başarı elde etmişler ve çalışmalar kelime vektörleri arasındaki mesafeye dayalı yaklaşımı kullanmışlardır. Bu çalışmalar iki ana modele ayrılmışlardır. Biri Gizli Semantik Analiz gibi matris çarpanlarına ayırma yöntemleri ve diğeri Skip-Gram gibi bağlama dayalı yöntemlerdir [71].

Mikolov ve arkadaşları da önerdikleri yöntemde sözdizimsel ve kelime analojisine dayalı kelime temsillerini elde etmişlerdir. Vektörlerin öğrenilmesi aşamasının RNN algoritması ile yapıldığı çalışmada kosinüs mesafesi ile vektör offset yöntemi çalışılmıştır. Vektör offset kelimenin vektör karşılığı anlamına gelmektedir. Tutarlı bir vektör offset bir kelime çiftleri arasında öğrenilen vektörlerin yakın olması anlamına gelmektedir [72].

GloVeda Stanford Üniversitesi tarafından oluşturulan metinsel verileri vektörel olarak gösterilmesini sağlayan kelime gömme metotlarından bir diğeridir [69].

GloVe metindeki kelimelerin vektörel temsillerini oluşturma için birlikte olma istatistiklerini kullanır. Global düzeyde gerçekleşen yöntem TF-IDF gibi sayıma dayalı denetimsiz bir öğrenme yöntemidir. Başka bir ifadeyle GloVe hem TF-IDF modeli gibi sayıma dayalı modellerin sezgisini kullanırken hem de Word2Vec gibi doğrusal yapıyı da yakalayabilen bir yöntemdir. Bir başka ifadeyle matrise dayalı yöntem ile bağlama dayalı yöntemi birlikte kullanır. “Global Log-Bilinear Regresyon” modeli de denilen bir yönteme dayanan GloVe’un temelinde yatan fikir, iki kelimenin birlikte ortaya çıkma olasılığını bir kelime temsili vektörü olarak ifade edilmesidir [70].

Örneğin GloVe vektörlerinin kullanıldığı bir çalışmada Pennington ve arkadaşları “buz” ve “buhar” kelimeleri incelemiş, “buz” kelimesinin “katı” ile buhar

kelimesinin de “gaz” ile daha çok birlikte ortaya çıktığını bulmuştur. Ayrıca çalışmada her iki kelimenin de su ile sıklıkla birlikte bulunduğu da belirlenmiştir [71].

3.7.2.6. BERT

Devlin ve arkadaşları tarafından 2018 yılında sunulan, açılımı “Bidirectional Encoder RepresentationsfromTransformers” olan BERT yeni bir dil temsili modeli ve makine öğrenimi çerçevesidir. Bu çerçeve önceden eğitilmiş bir çerçevedir [73,74].

Dönüştürücüden (Transformer) Çift Yönlü Kodlayıcı anlamına gelen BERT giriş verisini aynı anda hem soldan hem de sağdan,yani her iki yönden de okuma yeteneğine sahiptir. BERT bu yönüyle diğer yöntemlerden ayrılır. Bu şekilde bir kelimenin bağlamını her yönüyle öğrenir [75].BERT’in iki stratejisi vardır. Bunlar maskeli dil modelleme ve sonraki cümle tahminidir.

Maskeli Dil Modeli

BERT girdi değerlerinin bir kısmını rastgele maskeler. Bunu “Mask” belirteci ile değiştirerek yapar. Daha sonra maskelenmemiş olan değerlerin bağlamını dikkate alarak maskeli değerleri tahmin eder [75].

Sonraki Cümle Tahmini

BERT sonraki cümle tahmininde cümle çiftlerini kullanır.Sistem art arda gelen ikili cümleler ile ikinci cümlesi rastgele belirlenen cümle çiftleri kullanır. Girdi değerlerinin %50’si art arda gelen, %50’si ise ikinci cümlenin rastgele olduğu ikili cümlelerdir. BERT eğitilirken ikinci cümlenin ilk gerçek olan mı yoksa rastgele eklenen cümle mi olduğunu anlamaya çalışır. Sonraki cümle tahmini bu şekilde gerçekleşir [74,75].

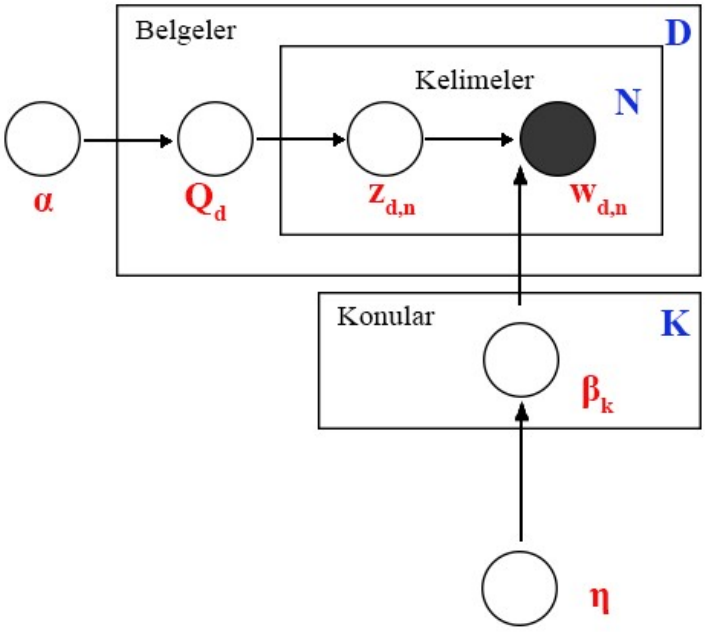
3.7.2.7. Gizli Dirichlet Tahsisi

Gizli Dirichlet Tahsisi ilk olarak 2003 yılında Blei ve arkadaşlarıtarafındangeliştirilmiş olan, metinsel belgelerdeki gizli anlamsal yapıları ortaya çıkaran denetimsiz bir kümeleme modelidir [76,77].

Gizli Dirichlet Tahsisi’nde temel olan, her konuyu anlatan farklı kelimelerin dağıtımı ile belgenin gizli konularının olduğunu ve bu konuların bir karışımı olduğu

fikridir. Ayrıca belge birden fazla konunun parçası olabilir. Amaç bir belgenin ait olduğu konuları bulabilmektir [77].

Gizli Dirichlet Tahsisinin, belgedeki gizli konuları keşfetmek, keşfettiği konulara göre belgeleri kümelemek, kümeleme sonrası verileri vektörel hale getirip sınıflandırma algoritmasına sunmak gibi faydaları bulunmaktadır.



Şekil 3 Gizli Dirichlet Tahsisi

Şekil 3'te Gizli DirichletTahsisi'nin belgenin konularının keşfedilmesi süreci gösterilmektedir. Siyah dairedeki $\mathbf{W_d}$ gözlemlenebilir değişkenleri, $\mathbf{Z_d}$ ise kelime başına atanan konuyu, $\boldsymbol{\beta_k}$ kelimeler üzerinden konu dağılımlarını, $\boldsymbol{\theta_d}$ ise belgelerin konu dağılımlarını gösterir. $\boldsymbol{\eta}$,$\boldsymbol{\beta_k}$'nın, $\boldsymbol{\alpha}$da$\boldsymbol{\theta_d}$'nin önceki dağılımları için hiperparametrelerdir [77].

3.8. Makine Öğrenimi

İnsanlar tarımdan sanayiye, hizmet sektöründen ev işlerine kadar uzanan birçok konuda işlerinin yapılması için makineleri kullanmaktadırlar. Kullanılan makineler ihtiyaca göre güncellenmektegerekirse yenileri icat edilmektedir. İşlemci sektöründeki gelişmeler de kişisel bilgisayarların icadının önünü açmıştır.

Günümüze gelindiğinde artık mobil cihazlar hayatımızı kolaylaştırmış ve birçok makinenin yerini akıllı cihazlar almaya başlamıştır.

Makine öğrenimi ise bilgisayarların açıkça programlanmadan verilere bakarak kendi iç görüleriyle öğrenme yeteneği kazanması olarak ifade edilir. Denetimli öğrenme ve denetimsiz öğrenme olarak ikiye ayrılır [78].

3.8.1. Denetimli Öğrenme

Denetimli Öğrenme giriş verisi ve çıkış verisinin eşlenmesi üzerine oluşturulan bir öğrenme sistemidir. Tüm verinin çıktı etiketi bulunmalıdır. Veriler eğitim verisi ve test verisi olarak ayrılır. Eğitim verisi üzerinden algoritma bir kalıp çıkarımı yapar. Test verisine de bu kalıbı uygular. Denetimli Öğrenme üzerine birçok algoritma bulunur. Bu algoritmalara Doğrusal Regresyon, Lojistik Regresyon, Karar Ağacı, K-en Yakın Komşu, Naïve Bayes, Rastgele Orman ve Destek Vektör Makineleri örnek olarak verilebilir [78].

Doğrusal Regresyon, iki değişken altında toplanan verilerin arasındaki ilişkiyi modelleyen algoritmadır. Değişkenlerden biri bağımlı diğer ise açıklayıcı değişkendir [79,80].

Lojistik regresyon bir regresyon algoritması da ikili sınıflama problemlerinde kullanılır ve değişkenler arasında doğrusal bir ilişki aranmaz. Bağımlı değişken 0 ile 1 arasında değerler alır [81].

Karar ağacı algoritması da regresyon ve sınıflama problemleri için kullanılan denetimli öğrenme algoritmasıdır. Karar ağacı algoritması eğitim verisi ile çıkarımda bulunduğu karar kurallarını test verisine uygulayıp tahminde bulunması yada sınıflandırma yapmasıdır [82].

K-en yakın komşu algoritması sınıflandırma veya tahmin için yakınlığı kullanan denetimli öğrenme algoritmasıdır. Önceden etiketli olan eğitim veri setindeki gruplara test veri setindeki her bir veriyi yakınlık değerine göre grup atamasını yapar. Bunu yaparken en yakın k komşunun ortalaması alınır [86].

Naïve Bayes algoritması koşullu olasılık olarak bilinen Bayes Teorimini kullanan denetimli öğrenme algoritmasıdır. Eğitim verisinde çalışıp her bir veri için olasılık hesabı yapar. Test veri setinde de verilerin olasılığa göre hangi gruba ait olduğunu tahmin eder [87].

Rastgele orman algoritması topluluk öğrenme algoritmalarına bir örnektir. Denetimli öğrenme algoritması olan Rastgele orman algoritması tahmin işlemini birden fazla karar ağacının bir araya gelmesi ve bu karar ağaçlarının ortalamasının alınması ile yapmaktadır. Bu tahmin doğruluğunu artırır. Bu şekilde algoritma sadece bir karar ağacına güvenmemiş olur. Ormanda ne kadar fazla ağaç var ise bu doğruluğun yükselmesini sağlar ve aşırı uymayı önler [89].

Destek vektör makineleri algoritması verileri en iyi şekilde sınıflayabilen hiperdüzlemi bulma üzerine kuruludur. Aslında veriler arasında birçok hiperdüzlem elde edilebilir. Fakat veri grupları arasındaki en uzak mesafeyi elde ettiğimiz hiperdüzlem aradığımızdır. Bu şekilde yeni eklenecek verinin de hangi gruba dahil olacağı daha net belirlenebilecektir [90].

3.8.2. Denetimsiz Öğrenme

Denetimsiz öğrenmede verinin çıkış değeri yoktur. Bilgisayar bir "doğru cevaba" başvurmaz. Verideki ilişkileri ve grupları belirlemeye çalışır. Algoritma kendisi veri içinde gruplar oluşturur. K-means kümeleme, hiyerarşik kümeleme, Apriori algoritmaları denetimsiz öğrenmeye örnek algoritmalardır [91].

K-meanskümeleme algoritması sadece giriş verilerini inceleyerek veriyi kümelere ayıran denetimsiz öğrenme algoritmasıdır. Önceden belirlenmiş k değeri kadar veriyi gruplara ayırır. Bunu yaparken her bir grup için bir merkez noktası belirler. Yeni veri merkezlerden hangisine daha yakın ise o gruba dahil olur [92].

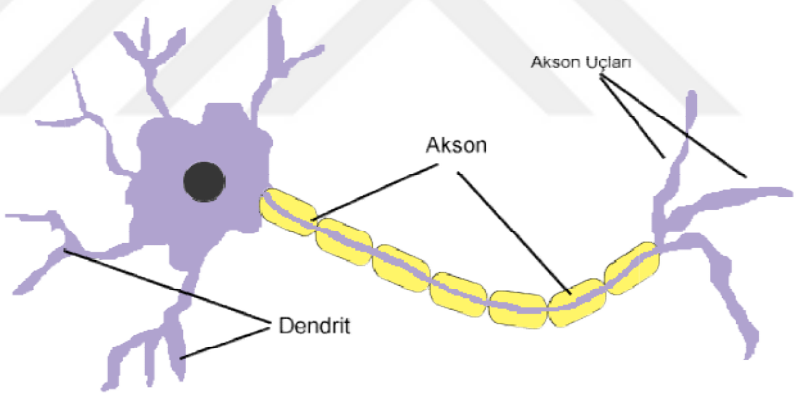
Hiyerarşik kümeleme algoritmasında ise iki yöntem kullanılmaktadır. Biri aglomeratif diğer ise bölücü hiyerarşik kümelemedir. Aglomeratif yöntemde her bir veri bir küme olarak başlar ve daha sonra birbirine yakın kümeler birleşir en az küme sayısına ulaşıncaya kadar bu durum devam eder. Bölücü yöntemde ise başta tüm veri tek bir kümedir. Verilerin arasındaki mesafeye bakarak küme bölünür. İdeal küme sayısına ulaşıncaya kadar bölünme devam eder [93].

Birliktelik kuralı algoritmalarından olan Apriori algoritması da denetimsiz öğrenme modellerine girmektedir. Bu algoritma kullanılan verilerin arasında en sık olanı bulma, hangi veri hangisiyle daha sık bir arada bunu bulmak için kullanılmaktadır. Algoritma istediği en sık olanları bulmak için birleştirme ve budama tekniklerini kullanır [94].

3.9. Yapay Sinir Ağları (YSA)

İnsan vücudunda sinir sistemi önemli bir role sahiptir. Sinir ağı milyonlarca sayıda nöron dediğimiz özelleşmiş hücrelerden meydana gelmektedir. Bu nöronların hepsi birbirine bağlı yapıdadır. Nöronlar arasındaki iletişim birtakım elektriksel ve kimyasal olaylar ile gerçekleşir [95]. Dışarıdan alınan duyusal bilgiyi almak, kaslarımıza motor komutlar göndermek, elektriksel sinyalleri dönüştürmek ve iletmek nöronların görevidir [96].

Bir nöronu bir ağaca benzetebiliriz. Bu ağacın üç bölümü vardır. Bunlar dallar, kökler ve gövdedir. Dallar Dendrit, kökler Akson, gövde ise Hücre Gövdesi veya Soma olarak adlandırılmaktadır. Akson nöron hücresinin verici kısmıdır. Sinir hücresinin gövdesindeki elektriksel sinyalin diğer nörona iletilmesini sağlar. Dendrit ise nöronun alıcı kısmıdır. Başka bir nöronun akson ucundan gelen elektriksel sinyali alıp kendi somasına (hücre gövdesi) ulaştıran yapıdır. Soma (ağaç gövdesi), içinde çekirdeği barındıran dendrit ve akson boyunca iletiminin yapıldığı yerdir [96].



Şekil 4 İnsan Sinir Hücresi

İnsanlardaki bir sinir hücresi Şekil 4'te gösterilmiştir.

Yapay Sinir Ağı Kavramı insan vücudundaki bu sinir hücrelerinden ilham alan bir sistem olarak tanıtılmıştır [96].

Biyolojik nöronun modellemesinde üç temel bileşen yer alır. Birincisi, sinapsler. Ağırlıklar olarak ifade edilir ve nöronlar arasındaki bağın gücü ağırlığın değeridir. İkinci önemli bileşen toplayıcıdır ve girdileri ağırlıklarına göre toplar. Üçüncü önemli bileşen ise aktivasyon fonksiyonudur ve çıkış değerinin genliğini düzenler. Genellikle bu genlik değeri 0 ile 1 arasında ya da 1 ile -1 arasındadır [97].

Bir YSA modeli ağırlıklı bağlantılar ile birbirine bağlı birçok düğümden oluşur. Her bir düğüm bir girdi verisi alır diğer bir düğüme aktarmak üzere bir çıkış verisi üretir. Çıkış verisi aldığı bilgiye bağlıdır [97].

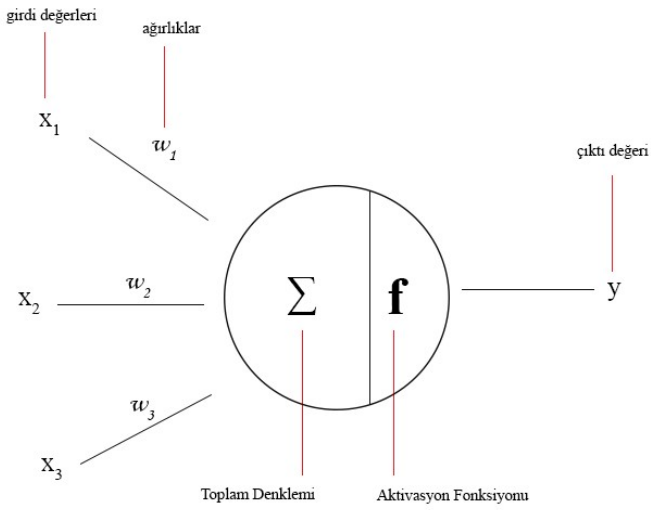
Düğümler tek başlarına çok güçlü sonuçlar elde edemezler fakat düğümlerin çokluğu ve birbiri ile bağlantılı olması sistemin gücünü ortaya çıkarır. Sistemin oluşturduğu sonuçlar ile gerçek sonuçlar arasında bir hata ortaya çıkar. Ortaya çıkan hatalar geri besleme ile sisteme verilir. Bu şekilde ağırlıklar tekrar uyarlanır. Ağırlıkların bu şekilde ağın gömülü olduğu ortamda sürekli uyarlanması sürecine sinir ağının öğrenme süreci denir. Uyarlanma hızınada öğrenme hızı denmektedir. Sistemin performansı istenilen seviyelere gelinceye kadar geri besleme işlemi devam eder [97].

Yapay sinir ağları kesin net bir sonuç vermek yerine yaklaşık sonuç üretir. Yeteri kadar verinin olmadığı veya gürültünün çok olduğu verilerde kullanılması önerilmez. Çok verinin olduğu karmaşık, model oluşturmanın zor olduğu durumlarda yapay sinir ağları iyi bir çözüm olabilir [97].

Yapay sinir ağlarının yapılandırılmış genel bir metodolojisi olmadığından genel amaçlı bir çözüm üretemez. YSAalgoritmalarının çıktı kalitesi de tahmin edilememektedir. Ayrıca ysa algoritmaları aşırı uymaya eğilimlidirler [95].

Yapay sinir ağları haritalama, örüntü tanıma, görüntü işleme, sınıflama ve kümeleme problemlerinde kullanılabilmektedir [97].

Bir Yapay Sinir Ağı NöronuŞekil 5'te gösterilmektedir.



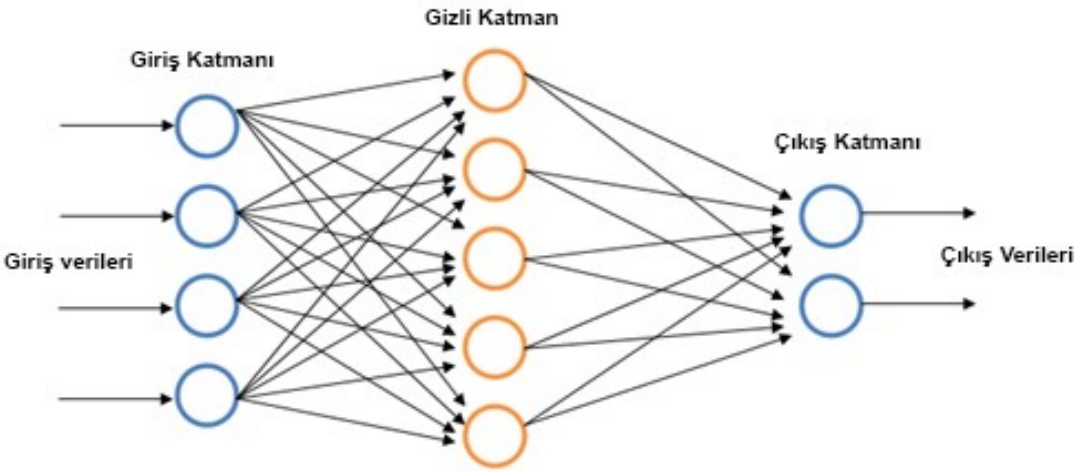
Şekil 5 Yapay Sinir Ağı Nöronu Modeli

3.9.1. İleri Beslemeli Sinir Ağları

İleri Beslemeli sinir ağında bilginin tek yöne doğru işlendiği sinir ağı modelidir. Veriler birden fazla düğümden geçerler ama tek yönde ilerlerler geri dönüş yapmazlar. Bu modelde herhangi bir döngüde oluşmaz [98].

İleri beslemeli sinir ağları çok katmanlı sinir ağları (MLN) olarak da bilinir. Bu modelde sigmoid nöronlar kullanılır. Sigmoid nöronlar doğrusal olmayan verileri işlemek için daha performanslıdır. Sınıflama problemleri ileri beslemeli ağlarda geleneksel sinir ağlarına göre daha kolay çözümlenir [99].

İleri beslemeli sinir ağının işleyişi şekil 6'da gösterilmektedir.



Şekil 6 İleri Beslemeli Sinir Ağı Modeli

Şekilde de görüldüğü gibi ileri beslemeli sinir ağlarında giriş katmanı, gizli katman ve çıkış katmanı bulunmaktadır.

Girdi Katmanı

Verilerin sisteme giriş yaptığı katmandır. Bu katmandaki nöron sayısı giriş verisinin özellik sayısı kadardır. Bu katman verileri bir sonraki katmana aktarır [99].

Gizli Katman

Giriş ve çıkış katmanları arasında yer almaktadır. Gizli katman sayısı modelin türüne bağlıdır. Çok büyük ve anlaşılması zor giriş verisi olan problemlerde gizli

katman sayısının çok olması gerekir. Bu şekilde bir gizli katman sonuçlarını bir sonraki gizli katmana, son gizli katman ise çıktı katmanına iletir [99].

Çıktı Katmanı

En son tahmin edilen değerlerin elde edildiği katmandır. Bu katman çıkış verisi olarak istenen değer kadar nörona sahip olmalıdır [99].

Aktivasyon Fonksiyonu

İleri beslemeli sinir ağlarında nöronların çıkışlarındaki karar verme mekanizmasıdır. Hangi düğümlerin işe koşulacağını da belirler. İleri beslemeli sinir ağlarında en çok kullanılan aktivasyon fonksiyonları ReLU, Sigmoid ve Tanh fonksiyonlarıdır [99].

Kayıp Gradyan Sorunu

Sinir ağları eğitim sırasında kaybolan gradyan problemi ile karşı karşıya kalabilmektedir. Eğitim sırasında sinir ağındaki ağırlıklar hataların tekrar geri besleme ile sisteme girilmesi ile aza indirgenmeye çalışılır. Yani ağırlıklar güncellenir. Bazı durumlarda gradyan o kadar azalır ki ağırlıklara etkisi olmaz. Bu duruma kayıp gradyan sorunu denir. Bu sorun ileri beslemeli sinir ağlarında da, tekrarlayan sinir ağlarında da gerçekleşebilir. Bu Soruna çözüm olarak ReLU fonksiyonu gösterilmektedir. Ayrıca LSTM ve GRU sinir ağlarında da kayıp gradyan sorunu oluşmamaktadır. Yani RNN'nin aksine LSTM ve GRU ağları uzun giriş değerlerini okuyabilir ve öğrenebilir [112].

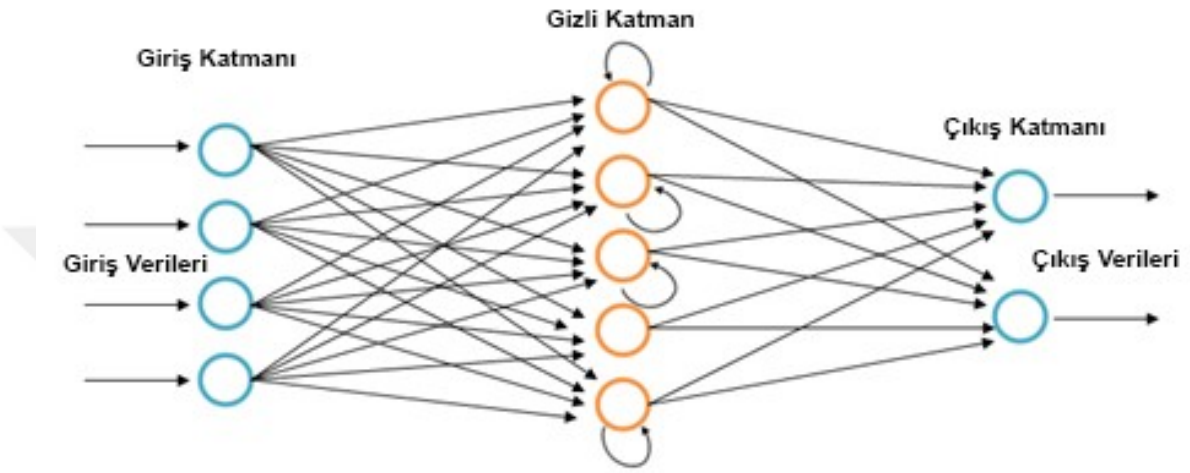
Patlayan Gradyan Sorunu

Sinir ağında eğitim sırasında ağırlıkların her iterasyonda güncellenmesi gerekir. Bu güncelleme hata gradyanlarının geri yayılımla sisteme dahil edilmesiyle gerçekleşir. Amaç güncelleme ile daha iyi tahmin sonucu elde etmektir. Derin sinir ağlarında bu hata graydyanları sistemde birikebilir ve büyük sonuçlar üretebilir. Bu da ağın kararsız bir yapıya bürünmesine yol açar ve bu durumda öğrenme gerçekleşmez, ya da eğitim sonucunda **NaN** değerleriyle karşı karşıya kalınabilir bu durumda da değerler güncellenemez [113].

İşte yukarıda bahsedilen soruna patlayan gradyan sorunu adı verilir. Bu sorunun çözümlerinden biri de LSTM sinir ağını kullanmaktır [113].

3.9.2. Tekrarlayan (RNN) Sinir Ağları

RNN yani tekrarlayan sinir ağları, ileri beslemeli sinir ağlarındaki her adımın bir sonraki adıma girdi olarak kullanılmasıyla oluşturulur. Böylece gizli katmanda bir döngü oluşur kısa hafıza elde edilir ve tekrar eden birimler belirir. Aşağıdaki şekilde bir RNN mimarisi gösterilmektedir [100].



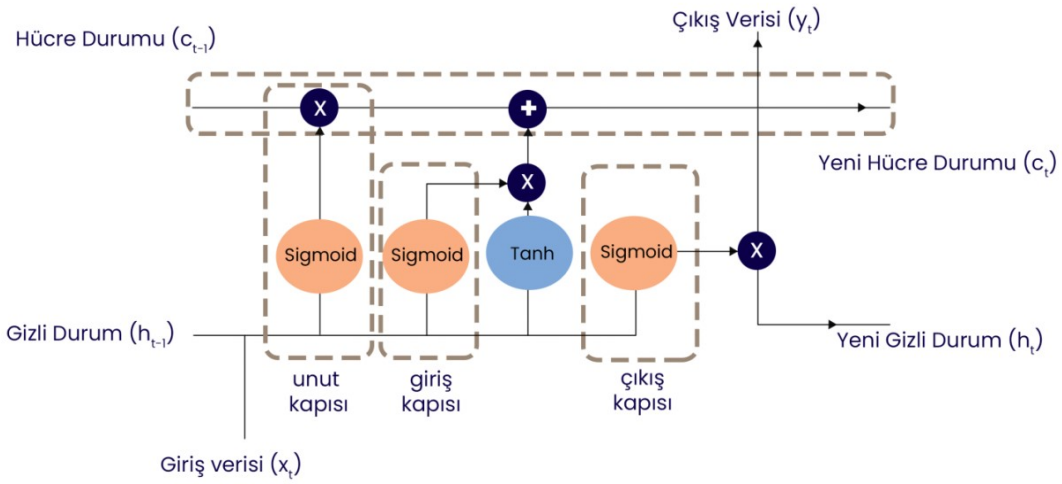
Şekil 7 Tekrarlayan Sinir Ağı Modeli

RNN ile gizli katmanda işlemler belirlenen sayıda zaman adımı için ve bir aktivasyon fonksiyonu ile gerçekleşir. RNN'de, gizli katmandaki ağırlıklar kullanılan gradyan algoritması ile kendilerini güncelleyebilirler. RNN sinir ağının ileri beslemeli bir sinir ağına kıyasla iyi çalıştığı düşünülse de bu çok yeterli bir seviyede değildir. Çok uzun verilerde, RNN kısa hafıza ile başarılı olmakta zorlanmaktadır. Bu durumda kayıp gradyan sorunu oluşur [100].

3.9.2.1. Uzun Kısa Süreli Bellek (LSTM) Ağı

LSTM bir RNN türünde derin öğrenme ağıdır. RNN ile yaşanabilecek olan kayıp gradyan sorunu ya da başka bir ifade ile uzun vadeli bağımlılıklar sorununun üstesinden gelebilen bir model sunmaktadır. Uzun süre hatırlayabilme ve öğrenebilme yeteneğine sahiptir.

LSTM katmanları birkaç özel yapıya, hücre durumuna, gizli duruma ve kapılara sahiptir. Hücre Durumu yapının hafızasıdır. Kapılar, giriş verisinin hangi kısmının unutulacağını ve hatırlanacağını belirler ve yapının sonuna giden hücre durumundan bilgi ekler veya çıkarır.



Şekil 8 LSTM Sinir Ağı Modeli

LSTM ağının mimarisi şekil 8'degösterilmektedir [101].

Unut Kapısı

Bir önceki gizli katmandan gelen veriler ve girdi olarak gelen veriler burada sigmoid fonksiyonuna tabi tutulur ve 0 ve 1 sonuçlarına göre unutulup unutulmayacağı belirlenir. Sonuç 0 ise unutulacak, 1 ise tutulacaktır [101].

Giriş Kapısı

Giriş kapısı, hücre durumunu güncellemek için kullanılır. Gizli katmandan girdi olarak gelen veriler burada hem sigmoid hem de tanh fonksiyonlarına tabi tutulur. Hangi bilgilerin hücre durumunun güncelleneceği ve saklama işlemi bu kapılarda gerçekleştirilir.

Unut kapısından ve Giriş kapısından gelen veriler toplanır ve bu hücre durumunun güncellenmiş değeri olur [101].

Çıkış kapısı

Bu kapı hangi gizli verilerin bir sonraki katmana gideceğini belirler. Sigmoid fonksiyonunda girdiden ve bir önceki gizli katmandan veriyi geçiriyoruz ve bu değer

tanhfonksiyonundan geçirilen hücre durumunda güncellenen verinin sonucu ile çarpılıyor. Sonuç olarak hangi gizli verinin bir sonraki katmana geçeceği belirlenir [101].

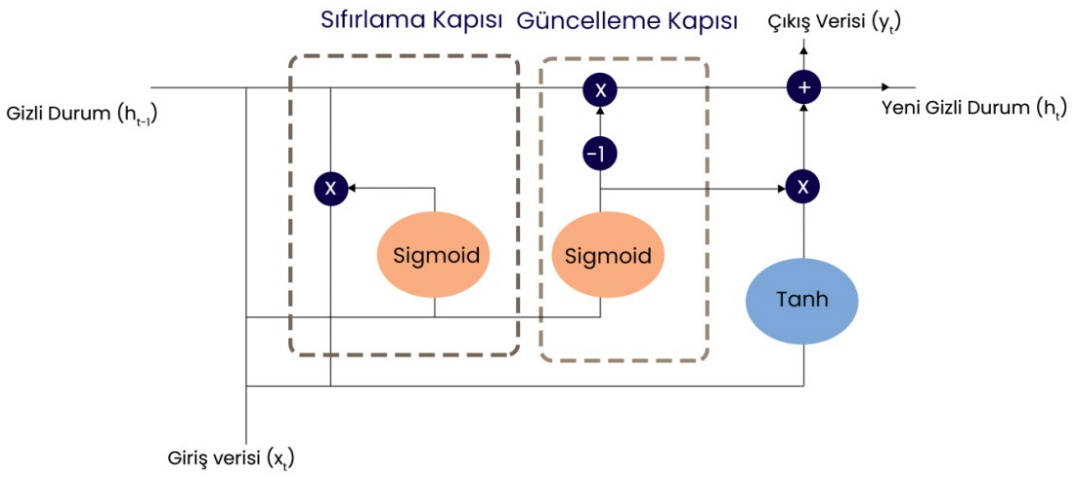
3.9.2.2. Kapılı Tekrarlayan Birim (GRU) Ağı

GRU katmanları da LSTM katmanlarına benzer özel yapılara sahiptir. LSTM'den farklı olarak GRU katmanında bir gizli durum ve iki kapı bulunmaktadır ve bu yapıları ile aynı LSTM gibi uzun vadeli bağımlılıkları koruyabilmektedir. GRU katmanında bulunan kapılar sıfırlama kapısı ve güncelleme kapısıdır. Bu kapılar faydalı olan gerekli bilgileri korurken gereksiz olanları filtrelemektedir. Bunun için gizli durumdaki veri ile ya da giriş verisi ile 0 ile 1 arasındaki vektörleri çarpar [102].

Sıfırlama kapısı ile önceki gizli durumdan gelen verilerle mevcut giriş verilerini kullanır. Ağırlık değerleri ile çarpılan bu iki veriyi toplar ve sigmoid fonksiyonundan geçirir. Sigmoid fonksiyonundan geçen veriye sıfırlama vektörü denilebilir. Sıfırlama vektörü daha sonra matris düzeyinde önceki gizli durumla çarpılır. Bu işlemle sıfırlama kapısı, önceki zaman adımlarında hangi bilgilerin saklanacağına karar verir. Elde edilen değerler girdinin ağırlıkla çarpımı ile toplanır. Daha sonra tanh fonksiyonu ile -1 ile 1 arasındaki değerlere düşürülür. Elde edilen değere "s" diyelim.

Güncelleme kapısında ise önceki gizli durumdan gelen verilerin toplamının ağırlıkları ile çarpılması ve mevcut giriş verileri sigmoid fonksiyonundan geçirilmesi gerçekleşir. Bu değer güncelleme vektörü olarak adlandırılabilir. Güncelleme vektörü, önceki gizli durumdan gelen verilerle matris düzeyinde çarpılır. Bu değere de "g" değerini verelim.

Sıfırlama kapısında oluşturduğumuz "s" değeri, güncelleme vektörünün tersi ile matris düzeyinde çarpılır. Bu şekilde gizli tutulacak veriler belirlenir. Elde ettiğimiz veriler ve "g" değeri toplanır.Böylelikle bir sonraki gizli duruma verilecek değer oluşturulur.



Şekil 9 GRU Sinir Ağı Modeli

Şekil 9'da bir GRU mimarisi gösterilmektedir [102].

3.9.3.Hiperparametreler

3.9.3.1. Öğrenme Oranı

Derin Öğrenmede ve sinir ağlarında öğrenme gerçekleşirken ağırlıklar optimizasyon algoritmaları ile güncellenmektedir. Ağırlıkların güncellendiği miktara öğrenme oranı denmektedir. Geri yayılımla elde edilen hatalar sisteme tekrar geri verilir. Ama bu hatanın tamamı ile yapılmaz öğrenme oranı ile çarpılarak sadece belirli bir oranı ile yapılır. Örneğin öğrenme oranı 0.1 ise hata ile çarpımı sonucu sisteme hatanın %10'u verilmiş olur. Yani böylelikle ağırlıkların %10'u güncellenmiş olur. Öğrenme oranı hiperparametresi modelin öğrenme hızını kontrol etmektedir. Ağırlıkların her güncellemedeki hata miktarı kontrol edilir. İyi belirlenmiş bir öğrenme oranı fonksiyonu en iyi yaklaşmayı sağlayacaktır. Gerekenden büyük bir öğrenme oranı, öğrenmenin hızlı olmasını sağlar ama salınıma sebep olur ve yakınsama gerçekleşmeyebilir. Gerekenden düşük öğrenme oranında ise yakınsama olur fakat öğrenme uzun sürede gerçekleşir [103].

3.9.3.2.Parti Boyutu (Batch Size)

Parti boyutu verinin kaçarlı gruplar halinde eğitime tabii tutulacağının sayısıdır. 100 satırlık veriniz olduğunu varsayalım. Oluşturduğumuz modelde parti boyutunu 5 yaparsak, beşer satırlık 20 parça şeklinde verinin eğitime katılacağını belirtmiş oluruz. Modelin ağırlıkları da her 5 parçanın sonunda güncellenecektir [104].

3.9.3.3. Dönem Sayısı (Number of Epoch)

Dönem sayısı eğitim sırasında veri setinin algoritma tarafından çalışılma sayısıdır. Her Dönemde tekrar algoritma çalıştırılır. Bu da eğitimde öğrenmenin geliştirilmesi anlamına gelir. Parti boyutunun da belirlendiği bir çalışmada dönem sayısı her bir partinin çalışma sayısı anlamına gelir. Dönem sayısı 1 ile sonsuz arasında bir tamsayı olabilir. Çalışmanın en iyi dönem sayısı denenerek bulunabilir. Eğitim sırasında elde edilen hata değeri değişmediği yerde dönem sayısı durdurulabilir. Daha sonra dönem sayısı bu değerle güncellenebilir [104].

3.9.3.4. Optimize Ediciler

Optimize edici görevi ağırlıkları güncelleyip kayıp değerini en aza indirmektir, kayıp fonksiyonunu sıfıra yaklaştırmaktır. Bu da gerçek sonuç ile tahmin edilen sonucun birbirine yakınlaşması anlamına gelmektedir [105].

Optimize edici olarak kullanılan algoritmalar aşağıda gösterilmiştir.

Gradyan İniş

Gradyan iniş bir modelin amaç fonksiyonu minimize etmenin bir yoludur. Sinir ağlarını optimize etmek en çok kullanılan optimizasyon algoritmalarındandır [106].

Gradyan iniş algoritmasının varyantları aşağıda açıklanmıştır.

Toplu Gradyan İniş

Toplu gradyan iniş algoritması eğitim sırasında gradyanları her hesaplamada tüm veriyi kullanır. Öğrenme oranı sabittir. Büyük veri kullanımında hızı yavaştır [106].

Stokastik Gradyan İniş (SGD)

Amaç fonksiyonları genellikle stokastiktir, yani olasılıksaldır. Stokastik Gradyan iniş algoritması toplu gradyan iniş algoritmasına göre daha hızlıdır. Sebebi ise her eğitim sırasında rastgele bir veri kullanıyor olmasıdır [106].

Adagrad

Adagrad öğrenme hızını parametreye uyarlayan bir gradyan iniş algoritmasıdır. Seyrek veriler için uygun bir algoritmadır. Dean ve arkadaşları

Adagrad algoritmasının Stokastik Gradyan İniş Algoritmasını sağlamlaştırdığını iddia etmektedirler [107].

Adagrad öğrenme oranını varsayılan olarak 0.01 değerini kullanır. Manuel ayarlama zorunluluğunu da bu şekilde ortadan kaldırmıştır. Adagrad formülü paydada gradyanların karesinin birikmesine sebep olur. Bu da öğrenme hızının küçülmesine sebep olur [106].

RMSprop

RMSprop algoritması Adagrad algoritmasının sınırlılıklarının düzelten bir algoritmadır. Bu algoritma GeoffHinton tarafından geliştirilmiştir. RMSprop öğrenme oranı değerini Adagrad algoritmasında da katlanarak azalan gradyanlarının kare değerlerinin ortalamasına böler. Hinton momentum faktörünü 0,9'a ayarlanmasını önerirken, öğrenme oranı için varsayılan değer 0,001'dir [106].

Adam

Adam momemtumlu stokastik gradyan iniş algoritması gibi davranır. Momentum eğimlerin yatay olarak yönlenmesini sağlar. Küresel minimuma ulaşmayı hızlandırır. Ayrıca Momentum ile RMsprop algoritmasının birleşimi olarak da görülebilir [106].

3.9.3.5. Kayıp Fonksiyonları

Sınıflandırma problemlerinde algoritmalar sınıfları tahmin etmeye çalışmaktadır. Doğruluk değerinin de artırılması gerekmektedir. Kayıp fonksiyonu algoritmanın çıktısı ile gerçek çıktı arasındaki farkı hesaplayan fonksiyondur [108].Optimizasyon problemlerinde amaç fonksiyonu kayıp fonksiyonuna dönüşür. Kayıp fonksiyonu eğitim sırasında elde edilir. Ne kadar düşürülebilirse model o kadar başarılı demektir.

Kayıp fonksiyonları regresyon için ve sınıflandırma için olmak üzere iki gruba ayrılmaktadır. Sınıflandırma için kullanılan bazı kayıp fonksiyonları İkili Çapraz Entropi, Kategorik Çapraz Entropi ve Seyrek Kategorik Çapraz Entropi fonksiyonlarıdır.

İkili Çapraz Entropi

İkili Çapraz entropi iki sınıflı problemler için kayıp fonksiyonu olarak kullanılmaktadır. İkili çapraz entropinin kullanıldığı problemlerde aktivasyon fonksiyonu olarak sigmoid fonksiyonu kullanılır.

Kategorik Çapraz Entropi

Kategorik Çapraz Entropi kayıp fonksiyonu ile kullanılan aktivasyon fonksiyonu softmax fonksiyonudur. Bu aktivasyon fonksiyonunun çıktısını en aza indirmek kategorik çapraz entropi kayıp fonksiyonunun görevidir. Softmax 0 ve 1 gibi tam sayılar değilde 0 ile 1 arasında ondalık sayı üretir. Kategorik çapraz entropi bu üretilen sayıların 0 ve 1 gibi istenilen değerlere yakınlaşmasını sağlar. Çok sınıflı problemlerde "one-hot encoded" tipte çıktılar için kullanılır.

Seyrek Kategorik Çapraz Entropi

Seyrek Kategorik çapraz entropi, kategorik çapraz entropi ile aynı işlevi yapar ve aynı aktivasyon fonksiyonu ile birliktekullanılır. Ondan tek farkı çıktı değerleri tamsayı tipinde olmasıdır. Kategorik çapraz entropi "one-hot encoded" tipte çıktılar için, seyrek kategorikçapraz entropi tamsayı tipinde çıktılar için kullanılır.

3.9.3.6. Aktivasyon Fonksiyonları

Aktivasyon fonksiyonu bir nöronun ağa girip girmeyeceğinin belirlenmesini sağlayan fonksiyondur. Beyinde de durum benzerdir. Nöronun aktive edilip edilmemesi girdi sinyallerinin yoğunluğuna bağlıdır ve karar buna göre verilir.

Aktivasyon fonksiyonunun en önemli rolü gelen ağırlıklı girdiyi bir sonraki gizli katmanın kullanacağı girdiye dönüştürmektir. Aktivasyon fonksiyonu aslında sinir ağına doğrusal-olmayanlık özelliği kazandırmaktadır.

Denklem 6'daki toplam denklemine göre;

$$z = \sum_i x_i w_i + b \qquad (6)$$

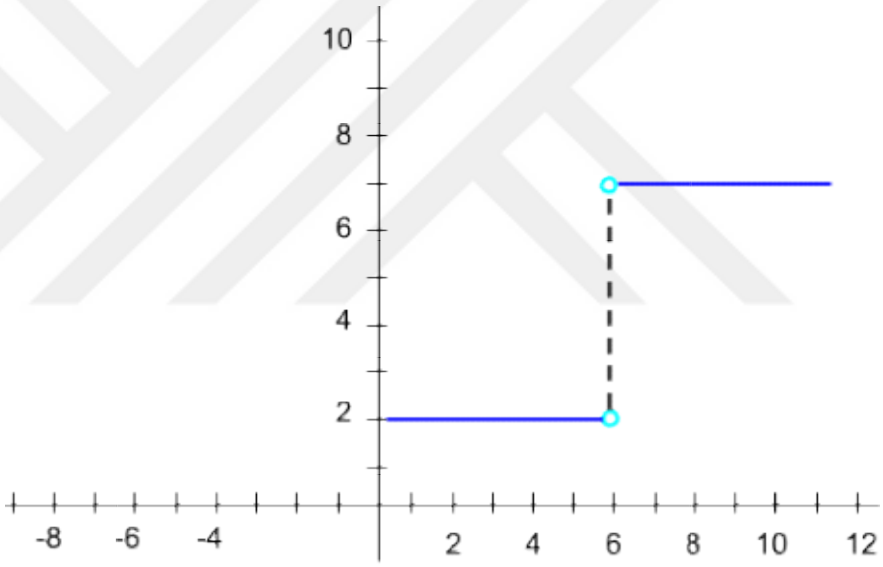
tüm girdi değerlerinin ağırlıklarla çarpıldıktan sonra toplanması ve en son bias değeri ile yani tahmin edilen ile gerçek arasındaki hata ile toplanır ve aktivasyon fonksiyonuna gönderilir. Aktivasyon fonksiyonu da bu toplam ile elde edilen z değerini kendisine girdi olarak kullanır.

Aktivasyon fonksiyonu olmayan bir sinir ağı sadece doğrusal bir sonuç elde edecektir. Ağdaki gizli katman sayısının bir önemi olmayacak model bir doğrusal regresyon sonucu üretebilecektir. Aktivasyon fonksiyonu sinir ağını işte bu doğrusallıktan kurtarmaktadır [109].

Üç çeşit aktivasyon fonksiyonu bulunmaktadır. Bunlar ikili adım aktivasyon fonksiyonu, doğrusal aktivasyon fonksiyonu ve doğrusal olmayan aktivasyon fonksiyonlarıdır.

İkili Adım(Binary Step) Aktivasyon Fonksiyonu

İkili adım aktivasyon fonksiyonunda eğer girdi değeri belirli bir eşik değerinin üstüne çıkarsa aktive edilir ve diğer gizli katmana iletilir, yoksa aktive edilmez yani bir sonraki gizli katmana iletilmez [109].



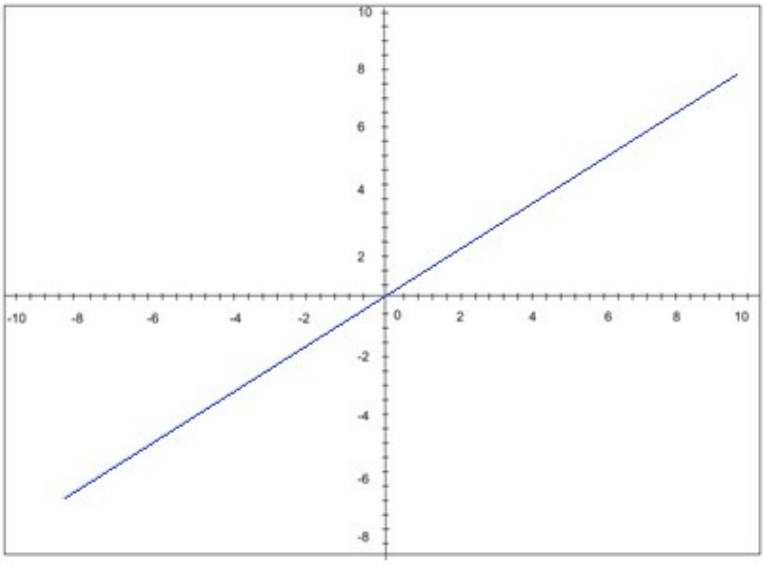
Şekil 10 İkili Adım Aktivasyon Fonksiyonu

İkili Adım aktivasyon fonksiyonu şekil 10'da grafiği gösterilmektedir. Bu aktivasyon fonksiyonu çok değere sahip problemlerde kullanılmaz. Yani çok sınıflı problemler için aktivasyon fonksiyonu olarak kullanılamaz. Adım fonksiyonunun gradyanı sıfıra eşittir. Bu durum da geri yayılımı engeller. İkili Adım Aktivasyon fonksiyonunun denklemi aşağıda gösterilmektedir [109].

$$f(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \qquad (7)$$

Doğrusal Aktivasyon Fonksiyonu

Doğrusal Aktivasyon fonksiyonu ağırlıkları ile çarpımı yapılıp toplanan girdi değerlerine bir etki yapmaz. Sadece diğer gizli katmana bu değeri iletir. Geri yayılıma izin vermez. Sinir ağındaki katman sayısının bir önemi yoktur. Doğrusal tek bir katmanın sonucunu üretir [109].



Şekil 11 Doğrusal Aktivasyon Fonksiyonu

Doğrusal Aktivasyon Fonksiyonunun grafiği şekil 11'de gösterilmiştir. Fonksiyonun denklemi de aşağıda gösterilmektedir [109].

$$f(x) = x \qquad (8)$$

DoğrusalOlmayan Aktivasyon Fonksiyonları

Lineer aktivasyon fonksiyonu geri yayılıma izin vermediğinden giriş katmanı ile çıkış katmanı arasındaki karmaşık eşleşmelerin oluşmasına sağlayamaz. Doğrusal olmayan Aktivasyon Fonksiyonları ise aşağıdaki problemlere çözüm bulur.

- Geri yayılıma izin verir, ve hataların tekrar sisteme dahil edilmesi ile ağırlıklar da güncellenir. Bu sayede daha başarılı bir tahmin mümkün olmaktadır.
- Gizli katmanda karmaşık eşleşmelerin yapılabilmesi ile çoklu sınıfa sahip çıktıların da tahmin edilebilmesi mümkün olabilmektedir.

Doğrusal olmayan bazı aktivasyon fonksiyonları aşağıda açıklanmaktadır.

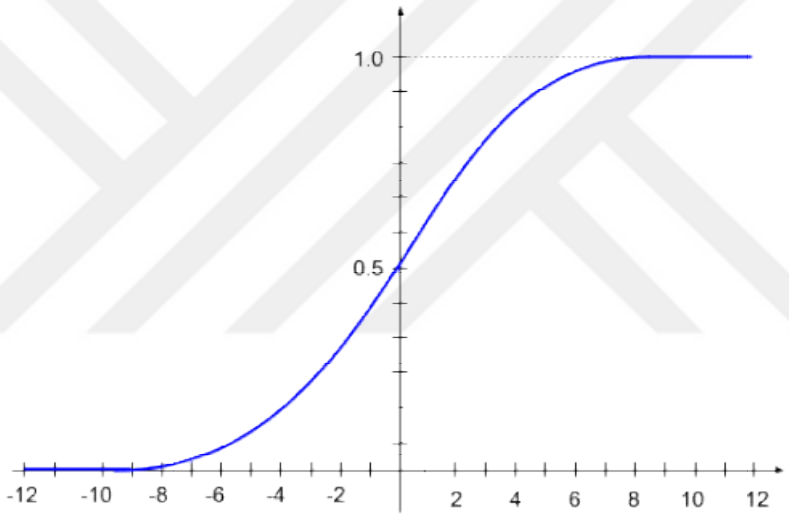
Sigmoid Aktivasyon Fonksiyonu

Sigmoid fonksiyonu girdi olarak verilen değerleri 0 ile 1 arasındaki çıktı değerleri olarak verir. Girdi değeri büyüdükçe çıktı 1 e yaklaşır, girdi değeri küçüldükçe çıktı değeri 0' a yaklaşır. Pozitif negatif ya da olumlu olumsuz gibi çıktı değerlerine sahip problemler için çok kullanılan bir aktivasyon fonksiyonudur. Sigmoid fonksiyonu bazı problemlerde kayıp gradyan sorununa sebep olabilir [109].

Sigomid Fonksiyonunun formülü denklem 9'da gösterilmektedir.

$$f(x) = \frac{1}{1 + e^{-x}} \qquad (9)$$

Fonksiyonun grafiği de şekil 12'de gösterilmektedir.



Şekil 12 Sigmoid Fonksiyonu

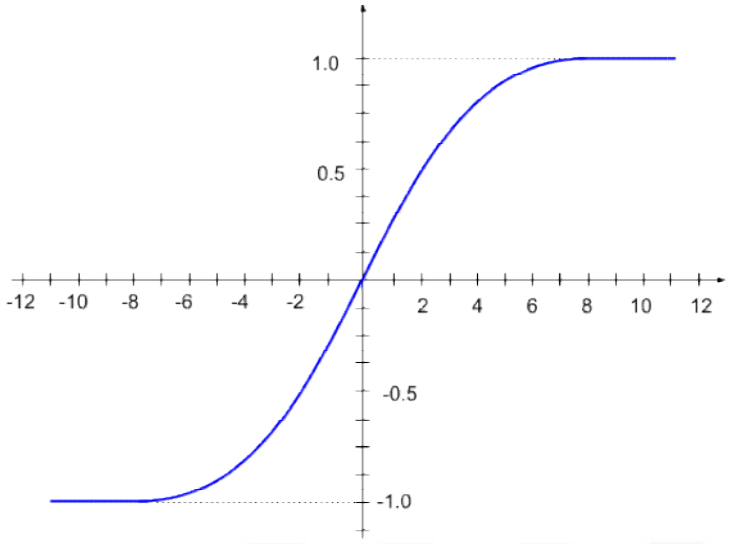
Tanh Aktivasyon Fonksiyonu

Tanh fonksiyonu girdi olarak verilen değerleri -1 ile 1 arasında döndürür. Girdi değeri büyüdükçe çıktı 1 e yakınsar, girdi değeri küçüldükçe çıktı değeri -1' e yakınsar. Tanh fonksiyonunda üç farklı çıktı değeri oluşturulabilir. Bunlar pozitif için +1, nötr için 0, negatif için -1'dir. Tanh aktivasyon fonksiyonu sigmoid fonksiyonuna göre daha dik gradyanlara sahip olur [109].

Tanh Fonksiyonunun formülü denklem 10'da gösterilmektedir.

$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \qquad (10)$$

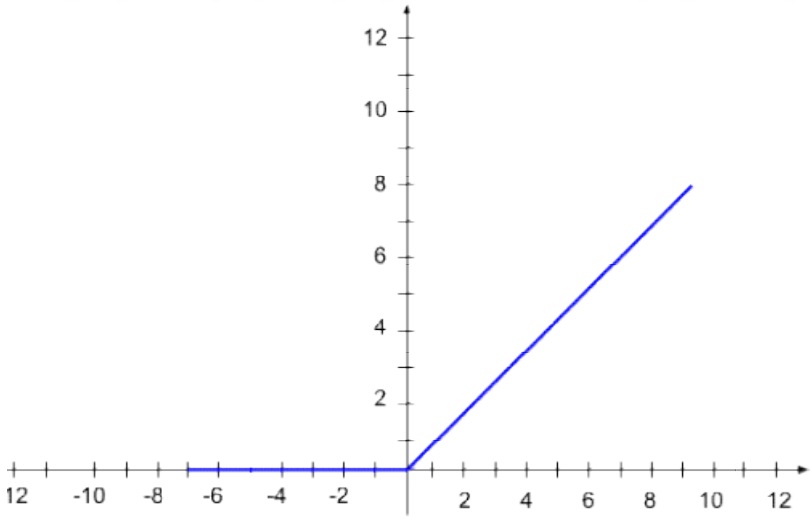
Fonksiyonun grafiği de şekil 13'te gösterilmektedir.



Şekil 13Tanh Fonksiyonu

ReLU Aktivasyon Fonksiyonu

Şekil 14'teki grafikte ReLU fonksiyonu Lineer gibi görünse de, öyle değildir. ReLU türevlenebilir bir fonksiyondur ve geri yayılıma da izin verir.



Şekil 14 ReLU Fonksiyonu

ReLU fonksiyonun en önemli özelliği tüm nöronları aynı anda aktive etmemesidir [109].

ReLU'da fonksiyona giriş yapan değerler negatif ise sonuç 0 döndürülür pozitif ise sonuç olduğu gibi döner [110].

ReLU fonksiyonu lineer fonksiyona göre çok daha verimlidir veglobal minimum bulmada hızlıdır [109].

$$f(x) = \max(0, x) \qquad (11)$$

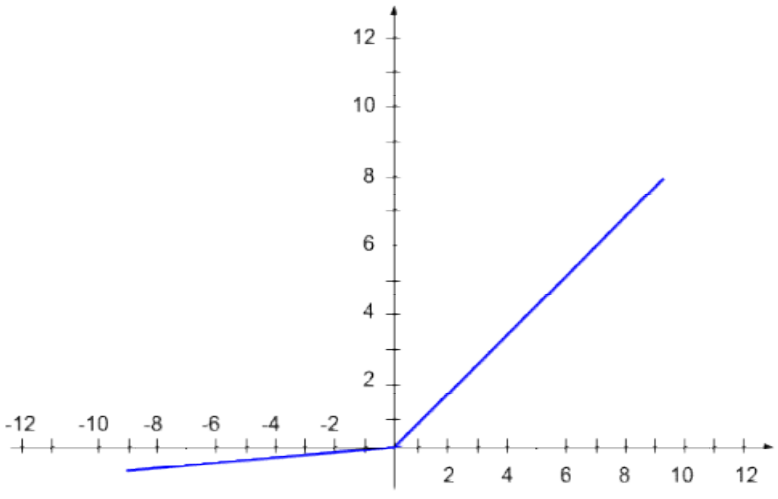
Fonksiyonun denklemi denklem 11'de gösterilmektedir.

ReLU fonksiyonunda grafiğin negatif kısmı hep sıfır döndürdüğü için negatif değerlerin çok olduğu verilerde ağırlıklar güncellenmemektedir. Bu duruma **ölü nöronlar sorunu** denir. Bu sorunu önlemek için ReLU fonksiyonuna farklı teknikler uygulanmıştır [111].

Sızdıran ReLU Aktivasyon Fonksiyonu

Sızdıran ReLU, ReLU fonksiyonunda ölü nöronlar problemi çözüm olarak üretilen fonksiyondur [109].

$$f(x) = \max(0.1x, x) \qquad (12)$$



Şekil 15 Sızdıran ReLU Aktivasyon Fonksiyonu

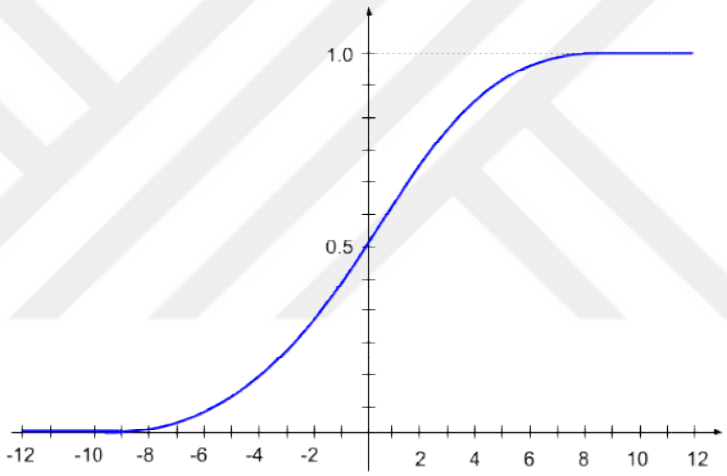
Sızdıran ReLU aktivasyon fonksiyonunun şekil 15'te grafiği bulunmaktadır. Denklem 12 de sızdıran ReLU aktivasyon fonksiyonunun denklemidir [109].

Bu fonksiyonun en önemli avantajı negatif giriş değerleri için geri yayılımı aktif edebilmesidir. Bu şekilde negatif giriş değerleri için gradyan değeri sıfırdan farklı olur [109].

Softmax Aktivasyon Fonksiyonu

Softmax Aktivasyon fonksiyonu sigmoid fonksiyonu gibidir. Grafiği de çok benzerdir. Fakat Softmax fonksiyonu sigmoid fonksiyonu gibi her girdiye sadece 0 ya da 1 değerini geri döndürmez. En önemli özelliği 0 ve 1 gibi iki grup değil de 2 den fazla çıktı tahmini yapılması gerektiğinde bunu gerçekleştirebiliyor olmasıdır. Softmax aktivasyon fonksiyonu, çoklu sigmoid fonksiyonu olarak da düşünülebilir. Her bir sınıf için olasılıksal bir değer döndürür [109].

Softmax aktivasyon fonksiyonunun grafiği ve denklemi aşağıdaki gibidir.



Şekil 16 Softmax Aktivasyon Fonksiyonu

$$softmax(z_i) = \frac{\exp(z_i)}{\sum_j exp(z_j)} \qquad (13)$$

Softmax fonksiyonunda 4 tane vektör girdi değeri olsun. [0.3, 0.6, 1.2, 1.8] değerleri softmax fonksiyonunda işledik ve sonuç [0.1, 0.2, 0.3, 0.6]olduğunu varsayalım.Softmax fonksiyonunun elde ettiği sonuçların toplamı 1 olmaktadır en yüksek değerin grubu çıktı tahmin olarak değerlendirilebilir. Bu şekilde çoklu çıktılarda tahminin gerçekleştirmesi sağlanmış olur [109].

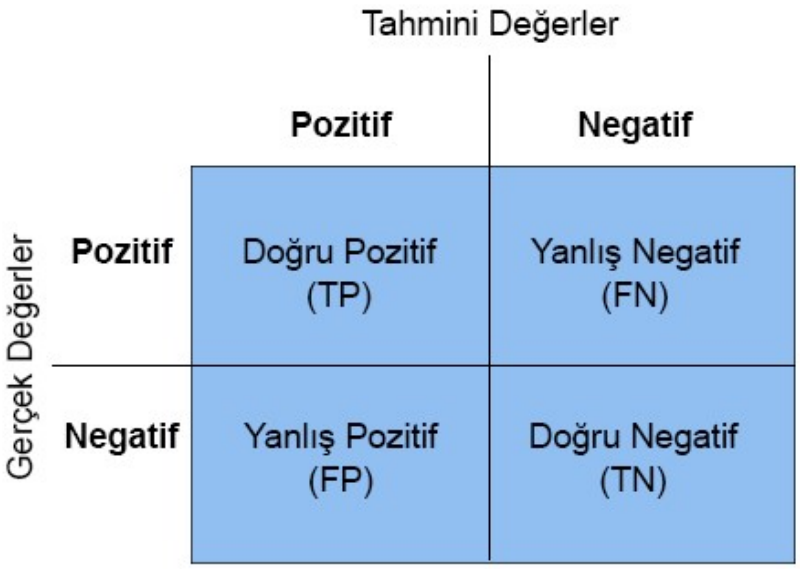
3.9.4. Değerlendirme Metrikleri

Regresyon veya sınıflama problemleri için yapılan çalışmalarda tahmin için uygulanan algoritmanın değerlendirilmesi makine öğreniminin önemli bir parçasıdır. Modelin tahmin başarısından emin olmak için değerlendirme esastır. Regresyon problemlerindeki sürekli değişkenlerin tahmininde kullanılan değerlendirme metrikleri ile sınıflama problemlerindeki nominal değişkenlerin tahmininde kullanılan değerlendirme metrikleri birbirinden farklıdır.

Değerlendirme metrikleri kayıp fonksiyonları gibi değildir. Kayıp fonksiyonları makine öğrenimi modelinin eğitilmesinde kullanılır. Değerlendirme metrikleri ise makine öğrenimi modelinin eğitim ve test sırasında performansını izler. Kayıp fonksiyonları makine öğrenimi modelinin parametrelerinde türevi alınabilir, ancak değerlendirme metriklerinin türevlenebilir olması gerekmemektedir [114].

En çok kullanılan değerlendirme metrikleri doğruluk, kesinlik(precision), duyarlılık(recall), ve F1 skordur. Değerlendirme metriklerinden önce tahmini değerler ve gerçek değerler arasındaki ilişkiyi gösteren karışıklık matrisi elde edilir.

Karışıklık matrisi yapılan çalışmada elde edilen tahminlerin doğruluğu ve yanlışlığını anlatan, tahmin değerleri ile gerçek değerleri karşılaştıran iki boyutlu bir özettir ve her tahmin sınıfına göre ayrılır [115].



Şekil 17 İki Sınıflı Karışıklık Matrisi Örneği

Şekil 17'de iki sınıfa sahip bir karışıklık matrisinin bir örneği gösterilmektedir. Matris ile TP(Doğru Pozitif), FP(Yanlış Pozitif), TN(Doğru Negatif) ve FN(Yanlış Negatif) değerleri belirlenir.

Doğru Pozitif (TP), gerçek pozitif değerlerin doğrutahmin edilenlerinin sayısı,

Yanlış Pozitif (FP), gerçek negatif değerlerin pozitifolduğutahmin edilenlerinin sayısı,

Doğru Negatif (TN), gerçek negatif değerlerindoğru tahmin edilenlerinin sayısı,

Yanlış Negatif (FN), gerçek pozitif değerlerin negatif olduğu tahmin edilenlerinin sayısıdır.

Karışıklık matrisinde yukarıda elde edilen, doğru pozitif, yanlış pozitif, doğru negatif ve yanlış negatif değerleri ile değerlendirme metrikleri elde edilir.

Sınıf sayısının n olduğu bir çalışmada karışıklık matrisi nxn düzeyinde bir matristir.

Şekil 18'de çok sınıflı bir karışıklık matrisinin bir örneği gösterilmektedir.

		Gerçek sınıflar			
		1	2	3	n
Tahmin edilen sınıflar	1	TP_1			
	2		TP_2		
	3			TP_3	
	n				TP_n

Şekil 18 Çok Sınıflı Karışıklık Matrisi Örneği

Modelin doğruluğu doğru tahmin edilen değerlerin sayısının tüm veri sayısına bölünmesi ile elde edilir. Doğruluk denklem 13'te ifade edilmektedir.

$$Doğruluk = \frac{\sum_{i=1}^{n}(tp_i + tn_i)}{\sum_{i=1}^{n}(tp_i + fn_i + fp_i + tn_i)} \quad (14)$$

3.9.4.1. Kesinlik (Precision)

Kesinlik metriği gerçek pozitif değer sayısının tahmin edilen tüm pozitif değer sayısına oranıdır. Burada pozitif olarak tahmin edilenlerin ne kadarının doğru olduğu anlaşılmaktadır. Yanlış pozitif değerine odaklanır. Kesinlik değerinin 0 ile 1 arasında bir değerdir ve 1 olması hiçbir pozitif değeri kaçırmadığı anlamına gelir. Kesinlik formülü denklem 15'te gösterilmektedir.

$$Kesinlik_{(i)} = \frac{\sum_{i=1}^{n} tp_{(i)}}{\sum_{i=1}^{n}(tp_{(i)} + fp_{(i)})} \quad (15)$$

3.9.4.2. Hassasiyet-Geri Çağırma(Recall)

Hassasiyet veya geri çağırma, pozitif olarak tahmin edilenlerin tüm gerçek pozitiflere oranıdır. Metrik yanlış negatif değerine odaklanır.Burada da hassasiyetin 1'e doğru yaklaşması hiçbir pozitif değeri kaçırmadığı anlamına gelir.Hassasiyet formülü denklem 16'da gösterilmektedir.

$$Hassasiyet_{(i)} = \frac{\sum_{i=1}^{n} tp_{(i)}}{\sum_{i=1}^{n} tp_{(i)} + fn_{(i)}} \quad (16)$$

3.9.4.3. F1 Skoru

F1 skoru kesinlik ve hassasiyetin harmonik ortalamasıdır. Kesinlik ve geri çağırma arasındaki dengeyi ifade eder. F1 Skor formülü denklem 17'da gösterilmektedir.

$$F1_{\ (i)} = \frac{2\ x\ Kesinlik_{(i)}\ x\ Hassasiyet_{(i)}}{Kesinlik_{(i)} + Hassasiyet_{(i)}} \quad (17)$$

3.9.4.4. Makro Ortalama Kesinlik

Çok sınıflı sınıflama problemlerinde makro ortalama kesinlik tüm sınıfların ortalama kesinliği olarak hesaplanmaktadır.

Makro ortalama kesinlik formülü denklem 18'de sunulmaktadır.

$$Makro\ ortalama\ kesinlik = \frac{\sum_{i=1}^{n} Kesinlik(i)}{n} \qquad (18)$$

3.9.4.5. Makro Ortalama Hassasiyet

Çok sınıflı sınıflama problemlerinde makro ortalama kesinlik benzeri olarak makro ortalama hassasiyet de tüm sınıfların ortalama hassasiyeti olarak karşımıza çıkar.

Makro ortalama hassasiyet formülü denklem 19'da sunulmaktadır.

$$Makro\ ortalama\ hassasiyet = \frac{\sum_{i=1}^{n} hassasiyet(i)}{n} \qquad (19)$$

3.9.4.6. Makro Ortalama F1 Skoru

Makro ortalama F1 skor makro ortalama kesinlik ile makro ortalama hassasiyet değerlerinin harmonik ortalamasıdır.

Makro ortalama hassasiyet formülü denklem 20'de sunulmaktadır.

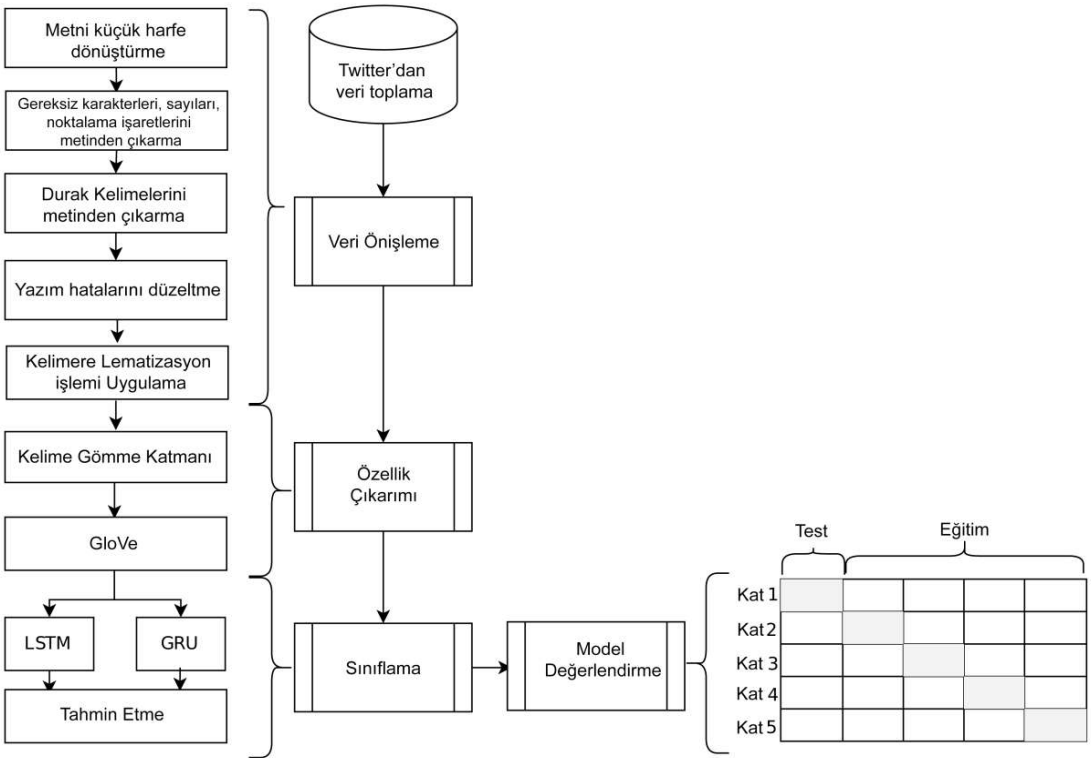
$$Makro\ Ortalama\ F1 = \frac{2\ x\ Makro\ Ortalama\ Kesinlik\ x\ Makro\ Ortalama\ Hassasiyet}{Makro\ Ortalama\ Kesinlik + Makro\ Ortalama\ Hassasiyet} \qquad (20)$$

Son olarak ROC eğrisi adında doğru pozitif değerler ile yanlış pozitif değerlerin oranının grafiği araştırmalarda kullanılmaktadır. Doğru pozitif değerler gerçekte pozitif olup doğru tahmin edilenler, yanlış pozitif değerler ise gerçekte negatif olup yanlış tahmin edilenlerdir.

AUC (Area Under the ROC Curve) ROC eğrisinin altında kalan alanı ifade etmektedir. Bu eğrinin altında kalan alan ne kadar büyükse model o seviyede başarılı olmuş demektir. AUC değeri 1.0 ise tüm tahminlerin doğru olduğu anlamına gelir.0.5 ise başarılı bir model olmadığı modelde değişiklik yapılması gerektiği anlamına gelir. AUC değeri 0 ise tahminlerin hiç tutmadığı anlamına gelir.

4. MATERYAL METOD

4.1. Tez Çalışmasının Mimarisi



Şekil 19 Tez Çalışmasının Aşamaları

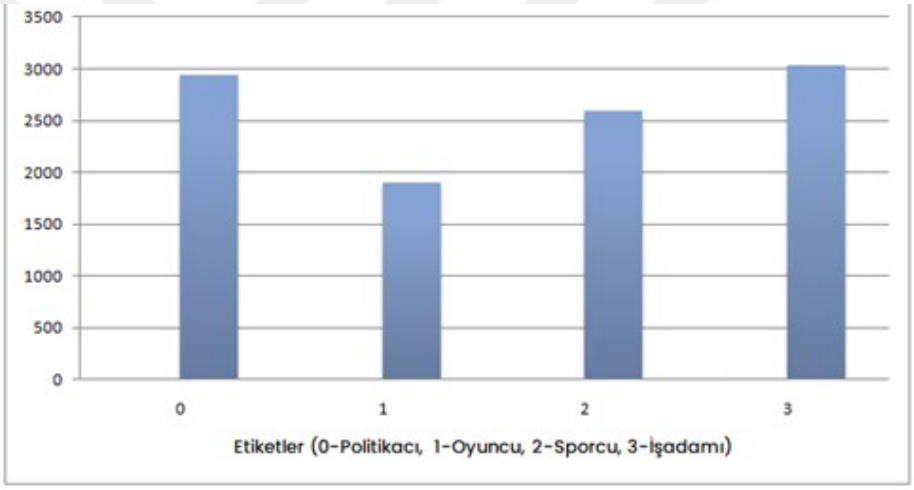
Tez çalışmasının tüm aşamaları Şekil 19'da gösterilmektedir. Twitter'dan sağlanan APIile veri toplama işleminden sonra gereksiz tüm ifadeleri çıkarmak, yazım hatalarını düzeltmek ve morfolojik olarak kök alma işlemi için önişlem aşaması gerçekleştirilmiştir. Özellik çıkarımı aşamasında ise kelime gömme katmanı ve GloVe vektörleri kullanılmıştır. Bu aşamada verideki öznitelikler belirleyici özelliklerine uygun olarak vektörlere dönüştürülmüştür. Önişlem ve özellik çıkarımı aşamalarının başarısı sınıflama aşamasındaki başarı oranını etkileyen en önemli unsurdur.

Sınıflama aşamasında derin öğrenme algoritmalarının kullanıldığı iki model oluşturulmuştur. Oluşturulan modellerin eğitim ve test verisi belirlemesi çapraz doğrulama yöntemiyle gerçekleştirilmiştir. Son olarak algoritmaların tahmin değerleriuygun görülen metrikler ile değerlendirilmiştir. Çalışmadaki tüm aşamalar ayrıntılarıyla birlikte aşağıda sunulmuştur.

4.2. Veri Seti

Araştırmacılar Twitter hesaplarına sahip olmak koşulu ile kendi hesaplarına ekleyebilecekleri bir API aracılığı ile belirledikleri hesapların Twitter paylaşımlarını veya belirledikleri bir konu ile alakalı olan Twitter paylaşımlarını toplayabilmektedir. Twitter bu APIaracılığı ile veri toplamaya izin vermektedir.

Yapılan tez çalışmasında yine Twitter API aracılığı ile 10454adet ingilizce tweet toplanmıştır.Toplanan tweetler 4 ünlü kişinin hesaplarındaki Twitter paylaşımlarıdır. Bu kişilerin meslekleri, Politikacı, Sporcu, Oyuncu ve İşadamı'dır. Bu meslekler Holland Teorisine göre sosyal, geleneksel, sanatsal, girişimci ve gerçekçi iş çevrelerini yansıtır. Toplanan veriler kişinin kendi yazısı olması için retweet içermemektedir.



Şekil 20 Tweetlerin Mesleklere Göre Dağılımı

Tablo 1*Tweetlerin Mesleklere Göre Dağılımı*

Meslek	**Tweet Sayısı**
İşadamı	3034
Politikacı	2933
Sporcu	2586
Oyuncu	1901

Toplanan veri ile ilgili istatistiksel bilgiler Şekil 20'de,veTablo 1'de gösterilmektedir.

4.3. Önişlem Aşaması

Veri önişleme olarak da ifade edilen bu aşama veri toplama aşamasından sonraki ilk adımdır. Gerçek metinsel veya görsel veriler dağınık bir yapıya sahiptir. Veriler uygulamalar veya iş süreçleri tarafından üretilir. Verilerin eksik değerlere, gereksiz karakterlere veya yazım hatalarına sahip olması olası bir durumdur. İşte dağınık olan bu verilerin belirli bir yapıda ve düzende olması gerekir. Verinin düzenli hale getirilmesi de önişlem aşamasında gerçekleştirilir. Doğru yapılmış bir önişlem aşaması sınıflama aşamasında oluşturulacak modelin daha yüksek başarı elde etmesine olumlu bir katkı sağlayacaktır [116].

Tweet paylaşımları da çoğunlukla metin verisi içerdiğinden yapısal olmayan bir veri olarak karşımıza çıkmaktadır. Bir tweet içerisinde hashtagler, bağlantılar, sayılar, noktalama işaretleri ve gereksiz kelimeler bulunabilir. Bu gibi ifadeler verinin anlamlandırılmasını ve veriden bilgi çıkarımını zorlaştıran ifadelerdir. İşte bu sebeple gereksiz görülen tüm ifade ve karakterlerin veriden temizlenmesi gerekmektedir.

Önişlem aşamasında elimizdeki veri önce küçük harflere dönüştürülmüştür. Bu işlemin başta yapılmış olması önişlemin diğer aşamalarında işimizi kolaylaştırmıştır. Küçük harflere dönüştürülen veriden daha sonragereksiz olduğu düşünülen hashtagler ve bağlantılar kaldırılmıştır. Daha sonra kelime olmayan veriler olan sayılar ve noktalama işaretleri kaldırılmıştır. Sayı değerleri ve hashtagler farklı çalışmalarda anlamlı bilgi içerir veya anlamlı görülebilir, fakat kişilerin mesleklerine bakışlarını incelediğimiz bu çalışmada anlamsız görülmüştür. Gereksiz ifadeler, sayılar ve noktalama işaretleri kaldırıldığında veride oluşan birden fazla boşluklar ve satırlar silinmiştir. Tüm bu işlemler için Python re (regularexpressionoperations) kütüphanesi kullanılmıştır.

Sırada durak kelimelerin kaldırılması işlemi vardır. Durak kelimeler dilin içinde yer alan bağlaçlar, edatlar gibi belirleyici özelliği olmayan kelimelerdir. Daha belirleyici olan kelimelerin ortaya çıkması için bu türden durak kelimelerin kaldırılması işlemi gerçekleşmiştir.

Durak kelimelerin kaldırılmasından sonra metinde kalan kelimelerin yazım ve dilbilgisi hatalarıPython TextblobKütüphanesi ile düzeltilmiştir. Örneğin, “tennnnis” kelimesi “tenis” olarak düzeltilmesi bu aşamada gerçekleşmiştir.

En son olarak lemmatizasyon olarak bilinen morfolojik olarak kök alma aşaması gerçekleşmiştir. Lemmatizasyon işlemi tüm kelimeleri “lemma” yani kök haline getirmiştir. Bunu yaparken kelimelerin fiil ise çekimleri düşürülmüş, isim ise çoğul eki kaldırılmıştır. Lemmatizasyon işleminin yerine “stemming” işlemi de kullanılabilirdi. Ama çalışmada lemmatizasyon işleminin “stemming” işleminden daha başarılı olduğu tespit edilmiştir. Lemmatizasyon işlemi ile verinin önişlem aşaması tamamlanmıştır veönişlem aşamasının tüm aşamaları Şekil 21’de gösterilmektedir.

Şekil 21 Önişlem Aşaması

Önişlem aşamasından sonra veri setinin ilk 5 ve son 5 satırı Şekil 22’de gösterilmektedir. Veri seti karışık bir şekilde özellik çıkarımı aşamasına gönderilmektedir.

	tweet	group
0	honor support hero fight folio	3
1	agree	0
2	deangelohall willcompton redskins hata gun mon...	2
3	billgates new book exceptionally clear accessi...	0
4	recommend factfulness highly han family member...	3
...	...	...
10449	second volunteer make call text voter use judg...	0
10450	suedhellmann division parent confidence child ...	3
10451	margin nearly americans believe senate hold he...	0
10452	wonder ll win team medium dmitri water war tom...	1
10453	learn maize wheat carlo slim commit great happen	3

Şekil 22 Önişlem Aşamasından sonra Veri Seti

4.4. Özellik Çıkarımı Yöntemi

Çalışmanın özellik çıkarımı aşamasında **Python Keras** kütüphanesi kullanılmıştır. Önişlem aşaması sonrasında elde edilen metin verisinden bir kelime indeksi oluşturulmuştur. Oluşturulan kelime indeksinde bulunan kelime sayısı **8871**'dir. Bu eşsiz olan kelimelerin sayısıdır ve aynı zamanda öznitelik sayısıdır. Ardından kelime indeksindeki her bir kelimeye belgedeki sıklığı ile ters orantılı olarak bir tamsayı değeri atanmıştır. Örneğin belgede en sık geçen kelimenin karşılığı 1'dir.

Tablo 2*Kelime İndeksinden Örnekler*

Kelimeler	**Atanan Tamsayı Değerleri**
health	27
Work	4
energy	41
family	56
Story	63
student	80
inspire	111
drama	3
investment	213
security	391
Team	1
president	2
Watch	19
malaria	167
Cost	471
Vote	34
Fun	202
economy	95
Court	103
Win	85

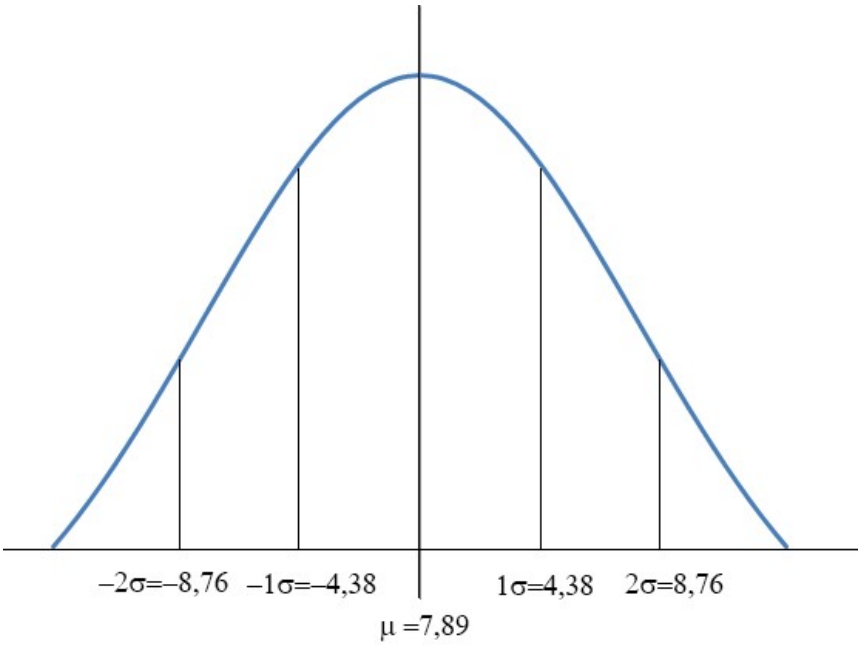
Çalışmanın sınıflandırma aşamasında LSTM ve GRU derin öğrenme ağlarıkullanılmıştır. LSTM ve GRU sinir ağlarına verilecek giriş verisinin ilk aşamasından yani verilerin tamsayı değerlerine çevrilmiş halinden bir kesit Tablo 2'de gösterilmiştir.

Tablo 3*Verinin İstatistiksel Bilgileri*

Kelime indeksinde elde edilen kelime sayısı	8871
Tweetler arsındaki en yüksek kelime sayısı	30
Tweetlerin kelime sayılarının Aritmetik Ortalaması	7,89
Tweetlerin kelime sayılarının Standart Sapması	4,38

Derin öğrenme algoritmalarından olan LSTM ve GRU sinir ağlarına gönderilecek olan giriş verisinin sabit uzunlukta olması gerekmektedir. Bu sabit değeri belirlemek için bir dizi ölçme ve değerlendirme yöntemi kullanılmıştır. Bu sabit uzunluğa göre bazı tweetler kesilecek, bazılarının ise içini sıfır ile dolduracağız. Bu işleme doldurma(Padding) işlemi denilmektedir.

İstatistikte veriyi temsil edebilecek seviyede bir örnekleminin alınması sıkça kullanılan bir yöntemdir. Bu popülasyonun tamamının toplanamaması durumunda özellikle uygulanmaktadır ve maliyetin düşmesini sağlamaktadır [120]. Tez çalışmasında da verinin %95' ini temsil edecek seviyedebir örneklemi kullanılmıştır. Bunun sebebi çok fazla kırpma işlemi yapmamak veya veriyi gereğinden fazla sıfır ile doldurmamaktır.



Şekil 23 Verinin Normal Dağılım Grafiği

Şekil 23'te gösterilen aritmetik ortalamaya eklenecek 1 standart sapma verinin %68'ini 2 standart sapma ise %95'ini ifade etmektedir. Bu istatistikte z tablosu olarak bilinen bir tablo yardımıyla bulunur. Z tablosu istatistikte dağılımın yüzde değerini bulmamızı sağlar. Kalan %5 değeri ondalık olarak gösterdiğimizde alfa değerini elde ederiz ki o da 0.05'tir. Bu alfa değerinin z tablosundaki karşılığı 1.96 değeridir ve bu z-skorudur. Aritemetik ortalamaya standart sapmanın 1.96 katı yani yaklaşık 2 katı eklenince verinin %95 lik kısmı temsil edilmiş olacaktır [117].

LSTm ve GRU sinir ağlarının istediği sabit uzunluktaki değer bu şekilde bulunmuştur ve kullanılan denklem aşağıda yer almaktadır.

$$\textbf{\textit{giriş değeri}} = \textbf{\textit{aritmetik ortalama}}(\textit{token sayısı}) + 1{,}96 \times \textbf{\textit{standart sapma}}(\textit{token sayısı}) \quad (20)$$

Denklemi veri setine uyguladığımızda sabit uzunluk değeri 16 olarak hesaplanmaktadır. Sütun sayısı 16'nın üstünde olan satırlar kesme işlemi ile 16'ya indirilecek, 16'dan az olan satırlara ise sıfır ile doldurma işlemi uygulanacaktır. Yapılan hesaplamalara göre doldurma işleminden sonra veri %93,7 oranında temsil edilmektedir. Değerler ile ilgili bilgiler aşağıda Tablo 4'te sunulmuştur.

Tablo 4*Doldurma İşlemi sonrası İstatistiki Bilgiler*

Tweetlerin kelime sayılarının Aritmetik Ortalaması	7,89
Tweetlerin kelime sayılarının Standart Sapması	4,38
Hesaplanan Sabit Uzunluk Değeri	16
Sabit uzunluk değerinin veri setini temsil etme oranı	%93,7

Doldurma işlemi sonrasında verilerin her satırı sabit bir değer olması için belirlediğimiz 16 sütunlu hale gelmiştir. Sıra gömme katmanı (EmbeddingLayer) oluşturmaya gelmiştir. Gömme katmanı ile Python Keras, verideki bütün tamsayı değerlerini 100 sütunlu vektörlere dönüştürmektedir. Oluşturulan vektörler sınıflama algoritmalarının eğitimi süresince kendini güncelleyecektir.

Özellik çıkarımı aşamasında gömme katmanının oluşturduğu vektörler ile yetinilmeyip önceden eğitilmiş GloVe vektörleri de kullanılmıştır [118].GloVe vektörleri 400 bin adet kelimeye ait 100 sütunlu önceden eğitilmiş vektörlerden oluşmaktadır ve bir txt dosyası içinde yer almaktadır. Daha önce çalışmanın veri setindeki kelimeler gömme katmanı ile 100 sütunlu vektörlere dönüştürülmüştü. Kelime indeksinde yer alan 8871 kelimedenGloVe vektörlerinin bulunduğu dosyadaki kelimelerle eşleşenlerin gömme katmanı vektör değerleriGloVe vektör değerleri ile yer değiştirmektedir. Bu şekilde daha iyi sınıflama başarısı hedeflenmiştir.

Ayrıca çalışmada toplanan tweet paylaşımlarının konularını belirlemek için ***Gizli Dirichlet Tahsisi*** kullanılmıştır. Toplanan tweetlerin konularını belirten tablo aşağıda gösterilmektedir.

Tablo 5*Toplanan Tweetlerin Konuları*

Meslek Türü	Mesleki Kişilik Tipi	Twitter Paylaşımları Konuları
Sporcu	Gerçekçi	Basketbol, Spor, Müzik
Politikacı	Girişimci, Sosyal	Ekonomi, Politika, İletişim, Sosyal Sorumluluk
Oyuncu	Sanatçı	Sinema, Film, Savaş
İşadamı	Girişimci	Çalışma, Eğitim, İklim Değişikliği, Sosyal Sorumluluk, Sağlık, Pandemi, Yenilik

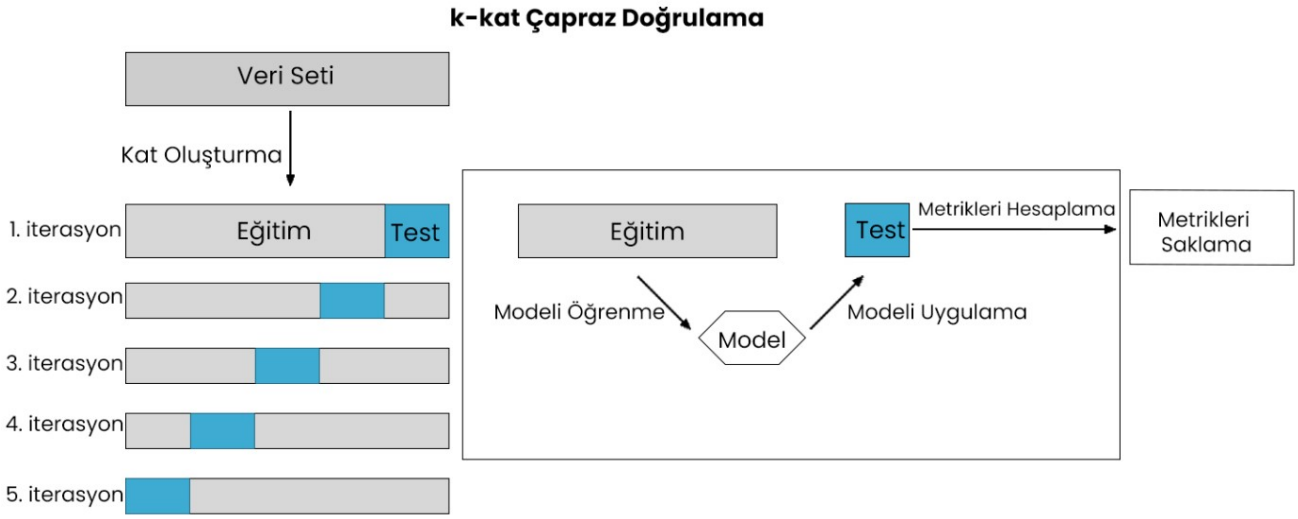
Tablo 5 incelendiğinde, Twitter paylaşımları toplanan ünlü kişilerin mesleklerine uygun tweet attıkları *gizli dirichlet tahsisi* ile tespit edilmiştir. Konuların mesleklerle ilgili olması sınıflama aşamasında mesleki ilgiyi tahmin ettiğimizi ortaya koymaktadır.

4.5. Çapraz Doğrulama ile Eğitim ve Test Verisi Oluşturma

Özellik çıkarımı sonrasında veri artık sınıflandırma aşamasına geçebilecek durumdadır. Yani verinin tamamı vektörel olarak ifade edilmiştir. Denetimli öğrenmede vektörel haldeki verilerin belirlenen bir yüzdesi eğitim verisi kalan veri de test verisi olarak ayrılır. Bu ayrım hep aynı yerden yapılır ve rastgelelik oluşturulmazsa, eğitim verisi her çalışıldığında daha önceki gördüğü veriye

rastlayacak, başarı yüzdesi yüksek çıkacaktır. Ama bu başarı test verisine aynı oranda yansımayacaktır. Bu sorunun oluşmaması için çalışmada 5 kat çapraz doğrulama yöntemi kullanılmıştır. Bu yöntemin kullanılmasındaki asıl amaç sınıflama algoritmasının test verisi üzerinde yapacağı tahminlerin geçerliliğini artırmaktır.

Çapraz doğrulama yönteminin uygulanışı Şekil 24'te gösterilmektedir.



Şekil 24 Çalışmanın Eğitim Verisi ve Test Verisi Ayrımı

Görüldüğü gibi veri 5 kez eğitime sokulmaktadır. Her defasında veri 5 parçaya ayrılmaktadır.1.iterasyonda 5 parçanın ilk 4'ü eğitim verisiyken 5. parça test verisi olarak sınıflama algoritmasına sunulur. Sonraki 4 iterasyonda farklı bir parça test verisi olarak ve kalan 4 parça veri eğitim verisi olarak kullanılır.

Eğitim verisi ile kullanılan sınıflama algoritmaları denetimli öğrenme ile yani sonuçların önceden bilinmesi yöntemi ile eğitilir. Eğitim belirlenen değer kadar iterasyon boyunca devam eder. Daha sonra Test verisi algoritmaya sunulur ve çıktı değerleri tahmin edilmeye çalışılır.

Çapraz doğrulamada 5 kez eğitim ve test verisi algoritmaya sunulur. Her defasında algoritmaya sunulan veri "kat" olarak ifade edilir. Başarı sonucu her bir parça için ayrıca elde edilir. Gerekirse başarı sonuçlarının ortalaması alınır.

4.6.Sınıflama Modeli Oluşturma

Çalışma için farklı GRU ve LSTM modelleri oluşturulup test edilmiştir. Hibrid kullanımların bile söz konusu olduğu çalışmada Şekil 25 ve Şekil 26'da görülen GRU ve LSTM için iki model oluşturulmuştur. Her iki model de gömme katmanı ile başlamaktadır. Gömme katmanı ile her bir eşsiz kelime için 100 adet vektör elde edilmiştir. Bu da 887.200 parametre etmektedir. Her iki modelde de bir dense (yoğun) katmanı ve bir dropout(bırakma) katmanı ile bitmektedir. Birinci modelde 3 adet GRU katmanı ikinci modelde ise 3 adet LSTM katmanı bulunmaktadır. Ayrıca GRU modelinin 893.592 eğitilebilir parametresi bulunurken LSTM modelinin 895.716 parametresi bulunmaktadır.

```
Layer (type)                 Output Shape              Param #
=================================================================
embedding-layer (Embedding)  (None, None, 100)         887200
_________________________
gru_18 (GRU)                 (None, None, 16)          5616
_________________________
gru_19 (GRU)                 (None, None, 8)           600
_________________________
gru_20 (GRU)                 (None, 4)                 156
_________________________
dropout_6 (Dropout)          (None, 4)                 0
_________________________
dense_6 (Dense)              (None, 4)                 20
=================================================================
Total params: 893,592
Trainable params: 893,592
Non-trainable params: 0
```

Şekil 25 Gated Recurrent Unit Network Modeli Yapısı

```
Layer (type)                 Output Shape              Param #
=================================================================
embedding-layer (Embedding)  (None, None, 100)         887200
_________________________
lstm_19 (LSTM)               (None, None, 16)          7488
_________________________
lstm_20 (LSTM)               (None, None, 8)           800
_________________________
lstm_21 (LSTM)               (None, 4)                 208
_________________________
dropout_6 (Dropout)          (None, 4)                 0
_________________________
dense_6 (Dense)              (None, 4)                 20
=================================================================
Total params: 895,716
Trainable params: 895,716
Non-trainable params: 0
```

Şekil 26 Long Short Term Memory Modeli Yapısı

Çıkış değerleri için de her deneme için 4 birimden oluşan "Yoğun katmanı" modele eklenmiştir. Sonuçlarda "aşırı uyma" 'yı engellemek için yoğun katmanından önce 0.43 değerinde "Dropout katmanı" modele eklenmiştir. Deneysel sonuçlar 0.43 değerinin en iyi değer olduğunu göstermiştir. Yoğun katmanında çıkış değerleri 2'den fazla yani 4 olduğu için kullanılan aktivasyon fonksiyonu "Softmax" aktivasyon fonksiyonudur.

Kayıp fonksiyonu"Seyrek Kategorik Çapraz Entropi"olarak belirlenmiştir. Optimize edici olarak "adam" seçilmiştir. Tablo 6'da hiperparametrelerin listesi verilmektedir.

Tablo 6*Hiperparametreler*

Hiperparametreler	**Değerler**
Optimize Edici	Adam
Kayıp Fonksiyonu	Seyrek Kategorik Çapraz Entropi
Aktivasyon Fonksiyonu	Softmax
Öğrenme Oranı	0.001
Dönem Sayısı	50
Parti Boyutu	256
Seyreltme Değeri	0.43

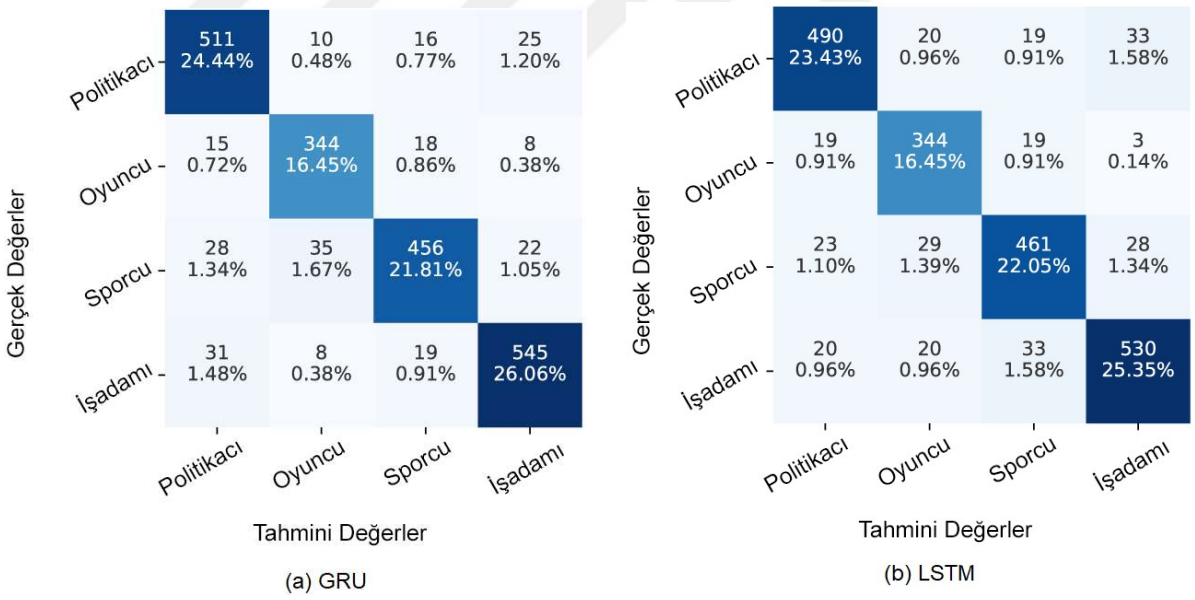
5.BULGULAR

Yapılan tez çalışmasının tamamında Python programlama dili ile çalışılmıştır. Veri toplama, önişlem ve özellik çıkarımı aşamalarından sonra verinin sinir ağı modellerine sunumuna geçilmiştir.

Çalışmada iki adet derin öğrenme sinir ağı modeli önerilmiştir. Biri GRU ikincisi ise LSTM sinir ağıdır. Oluşturulan modeller ROC eğrisi, Karışıklık Matrisi gibi grafikler ile desteklenmiştir. Ayrıca Kesinlik, Geri çağırma ve F1 Skor gibi değerlendirme metrikleri ile değerlendirme yapılmıştır. Eğitim ve Test verisi için başarı grafiği de gösterilmiştir.

5.1. Karışıklık Matrisleri

Şekil 27'de solda GRU modelinin, sağda iseLSTM modelinin karışıklık matrislerinin grafiği yer almaktadır.



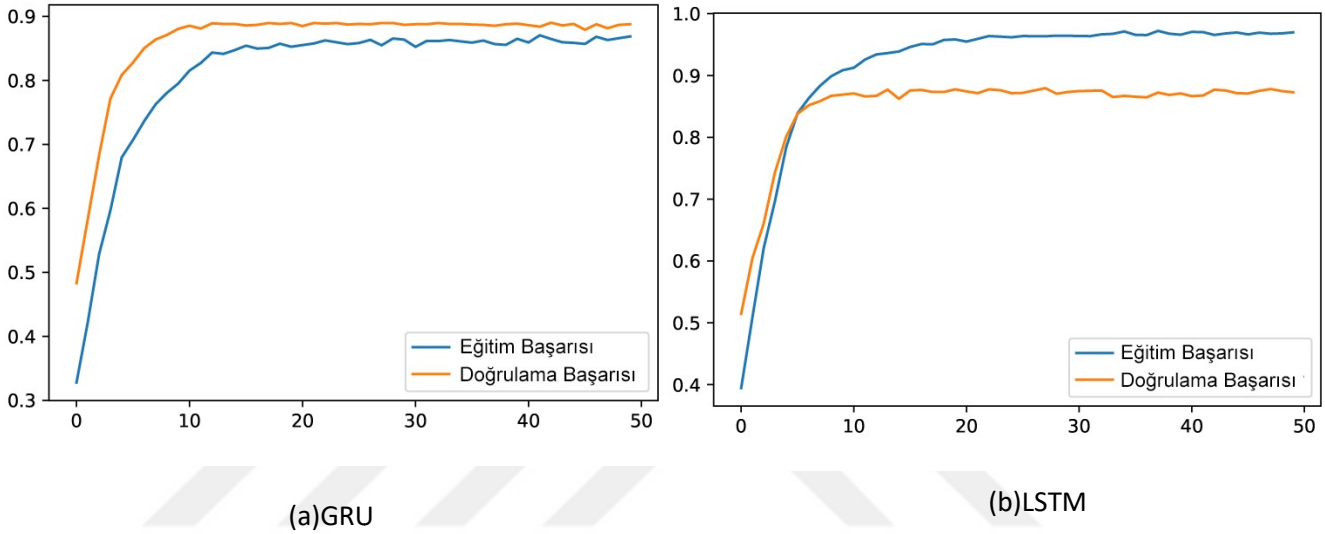
Şekil 27 GRU ve LSTM Model Karışıklık Matrisleri

Grafiğe göre GRU modelinin tahmin performansı LSTM modelininkinden daha başarılı çıkmıştır. GRU model Politikacı ve İşadamı sınıflarında %1 oranında daha başarılı, oyuncu sınıfında her iki model de eşit tahmin oranına sahip, sporcu sınıfında ise LSTM bir miktar yaklaşık %0,24 oranında daha başarılıdır. Her iki

modelde de yanlış tahmin değerleri %1.58'in üstüne çıkmamıştır. Dolayısıyla şekil 27'de gösterilen karışıklık matirisine göre genel olarak her iki model de yüksek oranda sınıflama tahmini başarısına sahiptir.

5.2. Modellerin Eğitim ve Doğrulama Başarıları

Şekil 28'de ise solda GRU modelinin, sağda ise LSTM modelinin eğitim başarısı ve doğrulama başarısı grafikleri yer almaktadır.



Şekil 28GRU ve LSTM eğitim sırasındaki eğitim ve doğrulama başarıları

Grafiği inceleyecek olursak, GRU modeli LSTM modelinden daha başarılı olduğu anlaşılmaktadır. GRU modelde eğitim verisi ile doğrulama verisi arasında bir uyum yakalanmış gözükmektedir. Eğitim başarısı grafiği bir istikrar noktasına çıkmış doğrulama başarısı grafiği de bir karar noktasına çıkmıştır. Bu iki grup verinin grafiği arasında çok fazla boşluk bulunmaması da bir uyuma işaret etmektedir.

LSTM modelinin grafiğinde ise doğrulama başarısı erken bir noktada kalmıştır, ve eğitim başarısı arasında büyüyen bir boşluk vardır. Doğrulama başarısı grafiğinin durumu eğitimin30. iterasyonundan sonra bitirilebileceğini göstermektedir. Ayrıca bu durum bu noktadan sonra aşırı uymanın gerçekleştiğinin göstergesidir.

Aşırı uyma yada ezberleme olarak ifade edilen durum modelineğitimi sırasında meydana gelmektedir.Modeleğitim sırasında tahmin yaparken yüksek değerde sonuçlar üretir ve görünmeyen verilere karşı doğru tahminde bulunma yeteneğini kaybeder. Çalışmada kullanılan LSTM modelinde de model30. iterasyondan sonra aşırı uyma eğilimi göstermiştir. Bu durumda 30. iterasyondan sonra eğitimin sonlandırımasında bir problem olmayacaktır.

5.3. Model Başarı Oranları

Tablo 7'de GRU ve LSTM sinir ağı modellerinin eğitim sırasındaki her bir katın eğitim ve test verisi başarı oranları gösterilmektedir.

Tablo 7*GRU ve LSTM Modelleri Başarı Oranları*

Modeller	**GRU**		**LSTM**	
	Eğitim Verisi Başarı Oranı	**Test Verisi Başarı Oranı**	**Eğitim Verisi Başarı Oranı**	**Test Verisi Başarı Oranı**
Kat 1	%86.7	%88.6	%83.7	%86.4
Kat 2	%84.3	%87.5	%83.4	%83.8
Kat 3	%84.5	%86.7	%83.9	%86.2
Kat 4	%87.1	%88.4	%83.9	%85.9
Kat 5	%84.1	%85.6	%80.8	%85.9

Eğitim 50 iterasyon olarak gerçekleşmiştir. Her bir kat bir kere test verisi olarak kullanılırken diğer katlar eğitim verisi olarak kullanılmıştır. Tablo 7'yi inceleyecek olursak,

Kat 1'de hem eğitim verisinde hem de test verisinde GRU modeli,

Kat 2'de hem eğitim verisinde hem de test verisinde GRU modeli,

Kat 3'te hem eğitim verisinde hem de test verisinde GRU modeli,

Kat 4'te hem eğitim verisinde hem de test verisinde GRU modeli,

Kat 5'te eğitim verisinde GRU modeli test verisinde LSTM modeli, daha başarılı olmuştur.

Eğitim verileri ile test verilerin başarı oranlarının birbirine yakın olması veri setlerinin düzgün dağıldığının bir göstergesidir.

5.4. Değerlendirme

Tablo 8'de GRU ve LSTM sinir ağı modellerinin değerlendirme sonuçları ortaya konmaktadır. Tüm grupların (Politikacı, Oyuncu, Sporcu, İşadamı) ayrı ayrı değerlendirmeleri gösterilmiştir.

Tablo 8 *GRU ve LSTM Modelleri Değerlendirme Sonuçları*

GRU	**Politikacı**	**Oyuncu**	**Sporcu**	**İşadamı**	**Makro Ortalama**
Kesinlik	0.873	0.866	0.895	0.908	0.886
Duyarlılık	0.909	0.893	0.842	0.903	0.887
F1 Skoru	0.891	0.879	0.868	0.906	0.886
Doğruluk	0.937	0.952	0.931	0.943	0.941

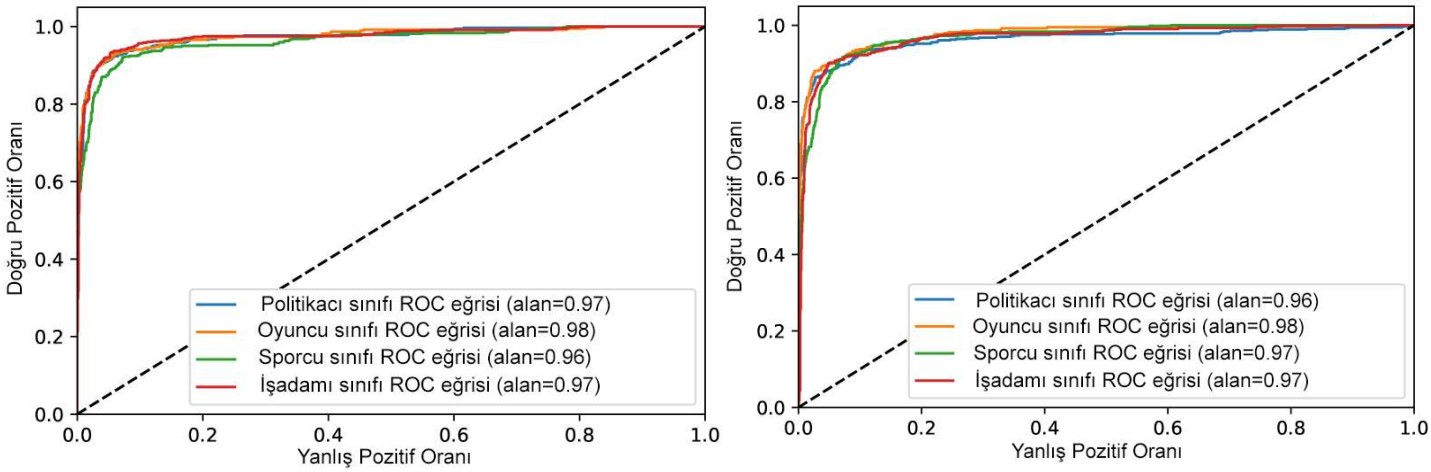
LSTM	**Politikacı**	**Oyuncu**	**Sporcu**	**İşadamı**	**Makro Ortalama**
Kesinlik	0.887	0.832	0.866	0.892	0.869
Duyarlılık	0.871	0.893	0.852	0.878	0.874
F1 Skoru	0.879	0.862	0.859	0.885	0.871
Doğruluk	0.932	0.943	0.924	0.930	0.932

Doğruluğun tek başına yeterli olmadığı durumlarda özellikle kesinlik, duyarlılık ve F1 skoru metrikleri kullanılmaktadır. Örneğin 100 kişinin olduğu bir grupta bir kişinin hasta olması %99 oranında doğruluk değeri verirken hasta kişinin bulaşıcı bir hastalığa sahip olması durumunda sonuç çok kötü olabilir. Bu durumda kesinlik, duyarlılık ve F1 skoru önem kazanmaktadır. Kesinlik doğru diye tahmin edilen grupların kaçının gerçekten doğru olduğunun bir ölçüsüdür. Duyarlılık ise gerçek doğru değerlerin kaç tanesinin doğru tahmin edildiği ile ilgilidir.F1 skoru ise bu her iki değerin ortalamasıdır. Bu ortalama aritmetik ortalama değil harmonik ortalamadır. Harmonik ortalama daha doğru sonuçlar alınmasını sağlamaktadır.

GRU sinir ağı modelinde makro ortalama özelinde doğruluk, kesinlik, duyarlılık ve kesinlik ve duyarlılığın harmonik ortalaması olan F1 skoru sırasıyla, 0.941, 0.886, 0.887, 0.886 olarak hesaplanmıştır. Aynı şekilde LSTM sinir ağında ise değerler sırasıyla, 0.932, 0.869, 0.874 ve 0.871 bulunmuştur. Bu sonuçlara göre mesleki ilgiyi GRU model %94 oranında tahmin ederken LSTM'de bu değer %93 olarak hesaplanmıştır.Ayrıca Kesinlik, duyarlılık ve F1 skoruna göre de GRU sinir ağı modeli LSTM sinir ağ modeline göre daha başarılı sonuçlar elde etmiştir.

5.5. ROC Eğrileri

Şekil 29'da solda GRU modelinin, sağda ise LSTM modelininROC Eğrisi grafikleri yer almaktadır.



(a)GRU (b)LSTM

Şekil 29 GRU ve LSTM Modelleri ROC Eğrisi

ROC eğrisi pozitif sınıfları negatif sınıflardan ayırma oranını gösterir. ROC eğrisinin altında kalan alan AUC olarak ifade edilir. Eğer hiçbir pozitif değer negatif değerlerden ayrılmamış olsaydı eğrimiz grafikteki lineer regresyon çizgisi gibi olacaktı. Burada AUC değeri 0.0 olacaktır. Eğer tüm negatifler ve pozitifler %100 oranında ayrılabilseydi bu durumda ise AUC değeri 1.0 olacaktır [119].

Yapılan tez çalışmasında elde edilen ROC eğrilerine bakacak olursaksolda GRU sağda ise LSTM modeli grafikleri gösterilmiştir. Eğriler tüm sınıflar için ayrı ayrı çizilmiştir. Eğrilerin sol üste doğru yaklaşması tahminde başarılı olunduğunu gösterir. Politikacı sınıfında GRU modelinin, Sporcu sınıfında da LSTM modelinin daha başarılı olduğu, diğer sınıflarda ise eşit başarıya sahip olduğu görülmektedir. Her iki modelde sınıflama tahmininde yüksek başarıya sahip olmaktadır. Bu da gönderilen tweetlerin mesleklerle ilgili olduğunun belirlenmesinde çok iyi bir performans olduğunun göstergesidir.

6.TARTIŞMA ve SONUÇ

John Holland'ın Meslek Kişiliği Yaklaşımına göre insanların ilgileri ile meslekler yani iş çevreleri arasında bir ilişki, bir etkileşim bulunmaktadır. Holland, kişilerin mesleki faaliyetlere olan ilgisini, "mesleki ilgi" olarak ifade etmektedir ve mesleki ilgiyi altı kategoriye ayırmaktadır. Bunlar gerçekçi, sanatsal, geleneksel, girişimci, sosyal ve araştırmacıdır. Bu yaklaşıma göre iş çevreleri de aynı isimlerle altı kategoriye ayrılmaktadır.

Holland, bu yaklaşımın merkezindeki mesleki ilgi kavramının kişilerin kariyerlerinde önemli bir yer tuttuğunu savunmaktadır. Bireylerin mesleki ilgilerini ölçerken de likert tipi anketler uygulamıştır.

Yapılan tez çalımasında da Holland'ın teorisinden esinlenilmiştir. Mesleki ilgiyi ölçerken ise likert tipi anket yerine sosyal ağların en önemlilerinden biri olan Twitter paylaşımları kullanılmıştır.Kişilerin tweet paylaşımları ile meslekleri arasındaki ilişkiyi keşfederek sosyal ağların insanların kariyerlerine karar vermelerinde yardımcı olup olmadığı araştırılmıştır.

Çalışma boyunca dört farklı meslekten kişilerin tweetleri kullanılmış ve tweetlerin meslekleriyle ilgili olup olmadığı incelenmiştir. Bu kapsamda iş adamı, politikacı, sporcu ve oyuncu mesleklerinden dört kişinin tweetleri dikkate alınmıştır.Toplam 10454 tweet incelenmiş olup bunların 3034'ü işadamının, 2933'ü politikacının, 2586'sı sporcunun ve 1901'i oyuncunun paylaşımlarıdır.

Önişlem aşamasında toplanan veri küçük harflere dönüştürülmüş ve hashtag'ler, bağlantılar veriden temizlenmiştir. Kişilerin mesleklerine bakışlarının incelendiği çalışmada veriden sayılar ve noktalama işaretleri de kaldırılmıştır.Daha sonra önişlem aşamasının önemli aşamalarından biri olan dilin yapısında bulunan edat, bağlaç gibi öğelerin genel ifadesi olan durak kelimelerin kaldırılması işlemi gerçekleşmiştir. Yazım hatalarının düzeltilmesinden sonra önişlemin en son işlemi olarak lemmatizayon işlemi gerçekleşmiştir. Lemmatizasyon işlemi ile metnin morfolojik olarak köklerine dönüştürülmesi amaçlanmıştır.

Özellik çıkarımı aşamasında python keras kütüphanesi ile kelime indeksi oluşturulmuştur. Oluşturulan kelime indeksi 8871 eşsiz kelimeden oluşmaktadır. Bu öznitelik sayısıdır. Bu kelimelerin herbiri için tamsayı değerleri atanmıştr. Sınıflama

aşamasında kullanılan LSTM ve GRU modellerine sabit uzunlukta veri girişi yapılması gerektiğinden istatistik alanından yararlanılarak sabit bir giriş verisi değeri elde edilmiştir. Bu değer 16 olarak belirlenmiştir. Bu sayının altında kelime sayısına sahip satırlar 0 ile doldurulmuş, bu sayının üstünde kelime sayısına sahip satırlar kesilip 16'ya eşitlenmiştir. Elde edilen 16 sütunlu satırlardan oluşan verideki her tamsayı değeri python keras gömme katmanı ile 100 sütunlu 0 ile 1 arasındaki vektörlere dönüştürülmüştür. Bu vektörler eğitim sırasında kendisini güncellemektedir. Çalışmada gömme katmanının oluşturduğu vektörlerden ziyade önceden eğitilmiş GloVe vektörleri de kullanılmıştır. Veri setindeki kelimelerden GloVe vektörlerindeki ile eşleşenler bu vektörlerle yer değiştirmiştir. Yapılan çalışmadaki denemeler neticesinde GloVe vektörlerinin kullanımı sınıflama eğitim ve test verisi başarısını artırdığını ortaya koymuştur.Ayrıca yapılan gizli dirichlet tahsisi uygulaması ile verilerin konusu belirlenmiş kişilerin attığı tweetlerin mesleklerini yansıtan konuları içerdiği belirlenmiştir.

Sınıflama aşamasında eğitim verisi ve test verisi çapraz doğrulama yöntemiyle belirlenmiştir. Veri 5 parçaya bölünmüş her bir parça 1 kere test verisi olmuş, kalan veriler eğitim verisi olmuştur.

Derin öğrenme algoritmalarının kullanıldığı çalışmada biri LSTM'ye, diğeri GRU'ya dayalı iki model geliştirilmiştir. Karışıklık matrisine göre modellerin her ikisi de %1.58 oranının üstünde yanlış tahminde bulunmamıştır.Test verisi başarı oranları da GRU için %88.6, LSTM için ise %84.6 olmuştur.

Bu tez çalışması Twitter gönderileri ile Holland mesleki kişiliğin yansımasının sonuçlarını araştırmıştır. Çalışmada her iki model de bireylerin mesleklerinin ve tweetlerinin tutarlı olup olmadığını belirlemede etkili performans sergilemektedir. Bununla birlikte, GRU'nun performansı biraz daha iyidir. GRU %94.025 makro ortalama doğruluk elde ederken, LSTM ise %93.025 makro ortalama doğruluk elde etmiştir.

Kaynakça

[1]Hoff, K. A., Chu, C., Einarsdóttir, S., Briley, D. A., Hanna, A., & Rounds, J. (2022). Adolescent vocational interests predict early career success: Two 12-year longitudinal studies. Applied Psychology, 71(1), 49-75.

[2]Kumara, A. R., Bhakti, C. P., Astuti, B., Ghiffari, M. A. N., & Ammattulloh, F. I. (2019, June). Development of Android Application based on Holland's Theory of Individual Student Planning. In International Conference on Social Science and Character Educations (IcoSSCE 2018) and International Conference on Social Studies, Moral, and Character Education (ICSMC 2018) (pp. 32-36). Atlantis Press.

[3]Gysbers, N. C. (2013). Career-ready students: A goal of comprehensive school counseling programs. The Career Development Quarterly, 61(3), 283-288.

[4] Holland, J.L.: Making vocational choices: A theory of vocational personalities and work environments. Psychological Assessment Resources (1997)

[5]https://www.omnicoreagency.com/twitter-statistics/

[6] Losiewicz, P., Oard, D.W. and Kostoff, R.N., "Textual data mining to support science and technology management", Journal of Intelligent Information Systems 2000, 15 (2): 99-119 (2000)

[7] Hoff, Kevin & Chu, Chu & Einarsdóttir, Sif & Briley, Daniel & Hanna, Alexis & Rounds, James. (2021). Adolescent Vocational Interests Predict Early Career Success: Two 12-Year Longitudinal Studies. Applied Psychology. 10.1111/apps.12311.

[8] Babarović, T., Dević, I. & Burušić, J. Fitting the STEM interests of middle school children into the RIASEC structural space. Int J Educ Vocat Guidance 19, 111–128 (2019). https://doi.org/10.1007/s10775-018-9371-8

[9]Usslepp, Nele & Hübner, Nicolas & Stoll, Gundula & Spengler, Marion & Trautwein, Ulrich & Nagengast, Benjamin. (2020). RIASEC Interests and the Big Five Personality Traits Matter for Life Success—But Do They Already Matter for Educational Track Choices?. Journal of Personality. 88. 10.1111/jopy.12547.

[10] Oliveira, Íris & Porfeli, Erik & Taveira, Maria do céu & Lee, Bora. (2020). Children's Career Expectations and Parents' Jobs: Intergenerational (Dis)continuities. The Career Development Quarterly. 68. 63-77. 10.1002/cdq.12213.

[11]Sheldon, K. M., Holliday, G., Titova, L., & Benson, C. (2020). Comparing Holland and Self-Determination Theory measures of career preference as predictors of career choice. Journal of Career Assessment, 28(1), 28-42.

[12] Mintram, K., Morgan, B., & de Bruin, G. P. (2020). An investigation of gender differences in Holland's circumplex model of vocational personality types in South Africa. International Journal for Educational and Vocational Guidance, 20(2), 293-310.

[13] Anggraini, W., Kurniawan, F., Susilawati, S., & Hasna, A. (2020). Validitas dan realibilitas instrumen teori pilihan karir Holland di Indonesia. Bulletin of Counseling and Psychotherapy, 2(2), 68-73.

[14] Rose, P., Nguyen, N. P., Kim, J. K., & Nguyen, D. (2022). The Personal Globe Inventory: The structure of vocational interest in Vietnam. Journal of Employment Counseling, 59(1), 27-36.

[15] Jones, K. S., Newman, D. A., Su, R., & Rounds, J. (2021). Black-White differences in vocational interests: Meta-analysis and boundary conditions. Journal of Business and Psychology, 36(4), 589-607.

[16] Fanny, F., Muliono, Y., & Tanzil, F. (2018). A comparison of text classification methods k-NN, Naïve Bayes, and support vector machine for news classification. Jurnal Informatika: Jurnal Pengembangan IT, 3(2), 157-160

[17] Pratama, B. Y., & Sarno, R. (2015, November). Personality classification based on Twitter text using Naïve Bayes, KNN and SVM. In 2015 International Conference on Data and Software Engineering (ICoDSE) (pp. 170-174). IEEE.

[18] Devika, R., Avilala, S. V., & Subramaniyaswamy, V. (2019, March). Comparative study of classifier for chronic kidney disease prediction using naïve bayes, KNN and random forest. In 2019 3rd International conference on computing methodologies and communication (ICCMC) (pp. 679-684). IEEE.

[19] Kumar, G. (2019, November). Breast cancer detection using decision tree, naïve bayes, KNN and SVM classifiers: a comparative study. In 2019 International

Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 683-686). IEEE.

[20] Ababneh, J. (2019). Application of Naïve Bayes, decision tree, and K-nearest neighbors for automated text classification. Modern Applied Science, 13(11), 31.

[21] Dharma, E. M., Gaol, F. L., Leslie, H., Warnars, H. S., & Soewito, B. (2022). The accuracy comparison among word2vec, glove, and fasttext towards convolution neural network (cnn) text classification. J Theor Appl Inf Technol, 100(2), 31.

[22] Parolin, E. S., Salam, S., Khan, L., Brandt, P., & Holmes, J. (2019, May). Automated verbal-pattern extraction from political news articles using cameo event coding ontology. In 2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS) (pp. 258-266). IEEE.

[23] Uday, S. S., Pavani, S. T., Lakshmi, T. J., & Chivukula, R. (2022). COVID-19 Literature Mining and Retrieval using Text Mining Approaches. arXiv preprint arXiv:2205.14781.

[24] Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on BERT model. PloS one, 15(8), e0237861.

[25] Sachi Nandan Mohanty a, E. Laxmi Lydia b , Mohamed Elhoseny c , Majid M. Gethami Al Otaibi d , K. Shankar e,(2020), Deep learning with LSTM based distributed data mining model for energy efficient wireless sensor Networks, Physical Communication, 40, 101097.

[26] Xuelian Yang , Jin Bai, and Xiaolin Wang,2021, Game User Preference Data Analysis and Market Guidance Based on Dynamic Attention GRU, Discrete Dynamics in Nature and Society, https://doi.org/10.1155/2021/5666405.

[27] Nayan Banik, Md. Hasan Hafizur Rahman, GRU based Named Entity Recognition System for Bangla Online Newspapers, International Conference on Innovation in Engineering and Technology (ICIET) 27-29 December, 2018.

[28] Aiping Guo, Ajuan Jiang, Jie Lin, Xiaoxiao Li, Data mining algorithms for bridge health monitoring: Kohonen clustering and LSTM prediction

approaches,(2020), The Journal of Supercomputing, 76:932–947 https://doi.org/10.1007/s11227-019-03045-8.

[29] Massaro, A., Dipierro, G., Saponaro, A., Galiano, A., Data Mining Applied In Food Trade Network, International Journal of Artificial Intelligence and Applications (IJAIA), Vol.11, No.2, March 2020.

[30] Gaiping Sun, Chuanwen Jiang, Xu Wang, Xiu Yang, Short-Term Building Load Forecast Based on a Data-Mining Feature Selection and LSTM-RNN Method, IEEJ Transactions On Electrical And Electronic Engineering,(2020), . DOI:10.1002/tee.23144.

[31] Lei Zhang, Chenbo Xu, Yihua Gao, Yi Han , Xiaojiang Du, and Zhihong Tian, Improved Dota2 Lineup Recommendation Model Based on a Bidirectional LSTM,(2020), Tsinghua Science And Technology, DOI: 10.26599/TST.2019.9010065.

[32] B. Riyaz, Sannasi Ganapathy, A deep learning approach for effective intrusion detection in wireless networks using CNN, Soft Computing(2020),24:17265–17278, https: // doi.org/ 10.1007/ s00500-020-05017-0.

[33] Hong-Bin Liu, Hao Wu, Weiwei Sun, Ickjai Lee, Spatio-Temporal GRU for Trajectory Classification,2019, IEEE International Conference on Data Mining (ICDM).

[34] Dr. S. Smys, Dr. Abul Basar, Dr. Haoxiang Wang, CNN based Flood Management System with IoT Sensors and Cloud Data,2020, Journal of Artificial Intelligence and Capsule Networks ,Vol.02/ No.04, 194-200.

[35] Zhang, J., Li, Y., Tian, J., & Li, T. (2018, October). LSTM-CNN hybrid model for text classification. In 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC) (pp. 1675-1680). IEEE.

[36]Gati, I., Tal, S. (2008). Decision-Making Models and Career Guidance. In: Athanasou, J.A., Van Esbroeck, R. (eds) International Handbook of Career Guidance. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-6230-8_8

[37]Matthews, R. J. (2017). A theory for everything? Is a knowledge of career development theory necessary to understand career decision making?. European Scientific Journal, 13(7), 320-334.

[38]Owen, F. K., & Owen, D. W. (2008). School counselor's role and functions: School administrators' and counselors' opinions. Ankara University, Journal of Faculty of Educational Sciences, 41(1), 207-221.

[39]Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (2001). Self-efficacy beliefs as shapers of children's aspirations and career trajectories. Child development, 72(1), 187-206.

[40]Campbell, R. E., & Cellini, J. V. (1981). A diagnostic taxonomy of adult career problems. Journal of Vocational Behavior, 19(2), 175-190.

[41]Bimrose, J. (2006) 'The changing context of career practice: Guidance, counselling or coaching?', CeGS Occasional Paper, .

[42]Gladwell, M. (2005) Blink: The power of thinking without thinking. London: Allen Lane.

[43]Mitchell, L., & Cubey, P. (2003). Characteristics of professional development linked to enhanced pedagogy and children's learning in early childhood settings: Best evidence synthesis. Wellington: Ministry of Education.

[44]Kileen, J. (1996) 'Career Theory - Chapter Two', in Watts, A.G., Law, B., and Killeen, J. (eds.) Rethinking careers education and guidance: Theory, policy and practice. London: Taylor & Francis

[45] Korkut-Owen, F., Demirtas–Zorbas, S., & Mutlu-Sural, T. (2015). Career sailboat model as a tool for the guidance counsellor. School Guidance Handbook.

[46] Krumboltz, J.. (1996) 'A learning theory of career counseling', in Savickas, M. and Walsh, W. (eds.) Handbook of career counseling theory and practice. Palo Alto, CA: Davies-Black.: , pp. 55–81.

[47] Holland's Six Personality Types https://www.cte.nd.gov/sites/www/files/ documents/ CRN/ Docs/ HollandTypes

[48] Oberlo. https://www.oberlo.com/blog/twitter-statistics. Accessed: 2022-02-11

[49] https://bootcamp.pe.gatech.edu/blog/what-is-data-mining/

[50] https://www.techtarget.com/searchenterpriseai/definition/data-scientist

[51] Kang, Y., Cai, Z., Tan, C. W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. Journal of Management Analytics, 7(2), 139-172.

[52] Mejova, Y. (2009). Sentiment analysis: An overview. University of Iowa, Computer Science Department.

[53] Prasad, G. N. R. (2021). Identification of Bloom's Taxonomy level for the given Question paper using NLP Tokenization technique. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(13), 1872-1875.

[54] Akın, A. A., & Akın, M. D. (2007). Zemberek, an open source NLP framework for Turkic languages. Structure, 10(2007), 1-5.

[55] Kadhim, A. I. (2018). An evaluation of preprocessing techniques for text classification. International Journal of Computer Science and Information Security (IJCSIS), 16(6), 22-32.

[56]https://en.wikipedia.org/wiki/Lemmatisation

[57]BoW, https://en.wikipedia.org/wiki/Bag-of-words_model

[58] Qaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. International Journal of Computer Applications, 181(1), 25-29.

[59] Word2Vec , https://code.google.com/archive/p/word2vec/

[60]Xue, B., Fu, C., & Shaobin, Z. (2014, June). A study on sentiment computing and classification of sina weibo with word2vec. In 2014 IEEE International Congress on Big Data (pp. 358-363). IEEE.

[61]Nawangsari, R. P., Kusumaningrum, R., & Wibowo, A. (2019). Word2vec for Indonesian sentiment analysis towards hotel reviews: An evaluation study. Procedia Computer Science, 157, 360-366.

[62]https://www.baeldung.com/cs/word-embeddings-cbow-vs-skip-gram

[63] Fast Text, https://en.wikipedia.org/wiki/FastText

[64]https://www.engadget.com/2016-08-18-facebook-open-sourcing-fasttext.html

[65] Shumaly, S., Yazdinejad, M., & Guo, Y. (2021). Persian sentiment analysis of an online store independent of pre-processing using convolutional neural network with fastText embeddings. PeerJ Computer Science, 7, e422.

[66]Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2019). Efficient estimation of word representations in vector space. arXiv preprint (2013). arXiv preprint arXiv:1301.3781.

[67]https://towardsdatascience.com/fasttext-bag-of-tricks-for-efficient-text-classification-513ba9e302e7

[68] Alessa, A., Faezipour, M., & Alhassan, Z. (2018, June). Text classification of flu-related tweets using fasttext with sentiment and keyword features. In 2018 IEEE International Conference on Healthcare Informatics (ICHI) (pp. 366-367). IEEE.

[69] GloVe, https://nlp.stanford.edu/projects/glove/

[70]https://becominghuman.ai/mathematical-introduction-to-glove-word-embedding-60f24154e54c

[71]https://nlp.stanford.edu/pubs/glove.pdf

[72] Mikolov, T., Yih, W. T., & Zweig, G. (2013, June). Linguistic regularities in continuous space word representations. In Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies (pp. 746-751).

[73]https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model

[74] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[75]https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

[76] Blei, David M.; NG, Andrew Y.; Jordan, Michael I. Latent dirichlet allocation. Journal of machine Learning research, 2003, 3.Jan: 993-1022.

[77] Bastani, K., Namavari, H., & Shaffer, J. (2019). Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. Expert Systems with Applications, 127, 256-271.

[78] Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9, 381-386.

[79]https://www.vedantu.com/formula/linear-regression-formula

[80]http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

[81] https://www.ibm.com/topics/ logistic-regression

[82] https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

[83] https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

[84] https://www.geeksforgeeks.org/decision-tree-introduction-example/

[85] https://addepto.com/blog/what-is-entropy-in-machine-learning/

[86] https://www.ibm.com/topics/knn

[87] Webb, G. I., Keogh, E., & Miikkulainen, R. (2010). Naïve Bayes. Encyclopedia of machine learning, 15, 713-714.

[88]https://www.kdnuggets.com/2020/06/naïve-bayes-algorithm-everything.html

[89]https://www.javatpoint.com/machine-learning-random-forest-algorithm

[90]https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[91] Qifang Bi, Katherine E Goodman, Joshua Kaminsky, Justin Lessler, What is Machine Learning? A Primer for the Epidemiologist, American Journal of Epidemiology, Volume 188, Issue 12, December 2019, Pages 2222–2239, https://doi.org/10.1093/aje/kwz189

[92] https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

[93]https://www.analyticsvidhya.com/blog/2019/05/beginners-guide-hierarchical-clustering/

[94]https://www.softwaretestinghelp.com/apriori-algorithm/

[95] M. Mishra and M. Srivastava, "A view of Artificial Neural Network," 2014 International Conference on Advances in Engineering & Technology Research (ICAETR - 2014), 2014, pp. 1-3, doi: 10.1109/ICAETR.2014.7012785.

[96]https://qbi.uq.edu.au/brain/brain-anatomy/what-neuron

[97]Dongare, A. D., Kharde, R. R., & Kachare, A. D. (2012). Introduction to artificial neural network. International Journal of Engineering and Innovative Technology (IJEIT), 2(1), 189-194.

[98] https://deepai.org/machine-learning-glossary-and-terms/feed-forward-neural-network

[99]https://www.datasciencecentral.com/feedforward-neural-networks/

[100] S.,Dobilas, "RNN: Recurrent Neural Networks—How to Successfully Model Sequential Data in Python",https://towardsdatascience.com/rnn-recurrent-neural-networks-how-to-successfully-model-sequential-data-in-python-5a0b9e494f92

[101]https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

[102]https://blog.floydhub.com/gru-with-pytorch/

[103]https://machinelearningmastery.com/learning-rate-for-deep-learning-neural-networks

[104]https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/#:~:text=The%20batch%20size%20is%20a%20number%20of%20samples%20processed%20before,samples%20in%20the%20training%20dataset.

[105]https://medium.datadriveninvestor.com/overview-of-different-optimizers-for-neural-networks-e0ed119440c3

[106]Ruder, S. (2016). An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747.

[107]Dean, J., Corrado, G. S., Monga, R., Chen, K., Devin, M., Le, Q. V, Ng, A. Y. (2012). Large Scale Distributed Deep Networks. NIPS 2012: Neural Information Processing Systems, 1–11. http://papers.nips.cc/paper/4687-large-scale-distributed-deep-networks.pdf

[108] https://towardsdatascience.com/what-is-loss-function-1e2605aeb904#:~:text=The%20loss%20function%20is%20the,be%20categorized%20into%20two%20groups.

[109]https://www.v7labs.com/blog/neural-networks-activation-functions

[110]https://medium.com/kodcular/

[111]https://www.educative.io/answers/what-is-the-dying-relu-problem

[112]https://towardsdatascience.com/the-vanishing-gradient-problem-69bf08b15484

[113]https://machinelearningmastery.com/exploding-gradients-in-neural-networks/

[114]https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide

[115]https://machinelearningmastery.com/confusion-matrix-machine-learning/

[116]https://www.techtarget.com/searchdatamanagement/definition/data-preprocessing

[117] Hajian-Tilaki, K. (2011). Sample size estimation in epidemiologic studies. Caspian journal of internal medicine, 2(4), 289.

[118] Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., Qin, B., et al.: Learning sentimentspecific word embedding for twitter sentiment classification. In: ACL (1), pp. 1555–1565. Citeseer (2014)

[119]https://developers.google.com/machine-learning/glossary#ROC

[120] https://en.wikipedia.org/wiki/Sampling_(statistics)

ÖZGEÇMİŞ

Adı Soyadı: Ömer DAĞISTANLI

Yabancı Dil: İngilizce

Eğitim Durumu:

Lisans: Balıkesir Üniversitesi, 2003-2007

Yüksek Lisans: Balıkesir Üniversitesi, 2009-2013

Çalıştığı Kurumlar:

Yozgat Bozok Üniversitesi- 2011-Devam ediyor.

Yayınları(SCI):

Dağıstanlı, Ö., Erbay, H., Yurttakal, A., H., Kör, H., "Solar Irradiaton Forecast by Deep Learning Architectures" ,Thermal Science., 2022, Vo:26, No:4A, pp2895,2906

2. Dağıstanlı, O., Erbay, H., Kör, H., Yurttakal, A., H., "Reflection of people's professions on social media platforms" ,Neural Computing and Applications., 2022, https://doi.org/10.1007/s00521-022-7987-8.