

T.C.
İSTANBUL BEYKENT ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ BİLİM DALI

**E-TİCARET SİTELERİNDE YER ALAN YORUMLARIN
BERT VE ALBERT DİL MODELLERİ İLE ANALİZİ**

Yüksek Lisans Tezi

Tezi Hazırlayan

Suat ERKAN

İstanbul, 2023

T.C.
İSTANBUL BEYKENT ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ BİLİM DALI

**E-TİCARET SİTELERİNDE YER ALAN YORUMLARIN
BERT VE ALBERT DİL MODELLERİ İLE ANALİZİ**

Yüksek Lisans Tezi

Tezi Hazırlayan

Suat ERKAN

2020003033

Orcid: 0000-0002-8216-787X

Danışman

Dr. Öğr. Üyesi Ediz ŞAYKOL

İstanbul, 2023

YEMİN METNİ

Yüksek Lisans Tezi olarak sunduğum “E-Ticaret Sitelerinde Yer Alan Yorumların BERT ve ALBERT Dil Modelleri ile Analizi” başlıklı bu çalışmanın, bilimsel ahlak ve geleneklere uygun şekilde tarafımdan yazıldığını, yararlandığım eserlerin tamamının kaynaklarda gösterildiğini ve çalışmamın içinde kullanıldıkları her yerde bunlara atıf yapıldığını belirtir ve bunu onurumla doğrularım. 15.06.2023

Suat ERKAN

T.C.
İSTANBUL BEYKENT ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ MÜDÜRLÜĞÜ
TEZLİ YÜKSEK LİSANS SINAV TUTANAĞI

15/06/2023

Enstitümüz *Bilgisayar Mühendisliği* Anabilim Dalı *Bilgisayar Mühendisliği* Programı yüksek lisans öğrencilerinden **2020003033** numaralı **Suat ERKAN**'ın "*İstanbul Beykent Üniversitesi Lisansüstü Eğitim – Öğretim Yönetmeliği*"nin ilgili maddesine göre hazırlayarak, Enstitümüze teslim ettiği "**E-Ticaret Sitelerinde Yer Alan Yorumların Bert ve Albert Dil Modelleri ile Analizi**" konulu tezini, Yönetim Kurulumuzun 30/05/2023 tarih ve 2023/22 sayılı toplantısında seçilen ve On-Line toplanan biz jüri üyeleri huzurunda, İstanbul Beykent Üniversitesi Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 29. maddesinin 3. fıkrası gereğince 45 dakika süre ile Zoom programı aracılığıyla on-line olarak aday tarafından savunulmuş ve sonuçta adayın tezi hakkında "**OYBİRLİĞİ**" ile "**KABUL**" kararı verilmiştir.

İşbu tutanak, 2 nüsha olarak hazırlanmış ve Enstitü Müdürlüğü'ne sunulmak üzere tarafımızdan düzenlenmiştir.

DANIŞMAN
Dr. Öğr. Üyesi Ed*** ŞA***
(İstanbul Beykent Üniversitesi)

ÜYE
Doç. Dr. At*** YI***
(İstanbul Beykent Üniversitesi)

ÜYE
Dr. Öğr. Üyesi So*** SE***
(Kütahya Dumlupınar Üniversitesi)

Adı Soyadı : Suat ERKAN
Danışmanı : Dr. Öğr. Üyesi Ediz ŞAYKOL
Türü ve Tarihi : Yüksek Lisans Tezi, 2023
Alanı : Bilgisayar Mühendisliği
Anahtar Kelimeler : ALBERT, BERT, Derin Öğrenme, Duygu Analizi

ÖZ

E-TİCARET SİTELERİNDE YER ALAN YORUMLARIN BERT VE ALBERT DİL MODELLERİ İLE ANALİZİ

Günümüzde popüler hale gelen çevrimiçi alışveriş (e-ticaret) platformları hızlı, kolay, güvenli ve temassız bir alışveriş deneyimi sunmaları nedeniyle pek çok insan tarafından sıklıkla tercih edilmektedir. Müşterilerin çevrimiçi alışveriş sırasında satın alma işlemini gerçekleştirmeden önce doğru karar verebilmelerini sağlamak amacıyla başvurdukları ürün ve hizmet ile ilgili çevrimiçi platformlar üzerinden yapılan olumlu veya olumsuz yorumlar alışveriş tercihlerinde önemli bir rol oynamaktadır. E-ticaret platformlarında yer alan ürünlere ait müşteri yorumları, müşterilerin ürün/hizmet hakkında gerçekçi bir fikir edinmelerine ve işletmelerin bu yorumları değerlendirerek performanslarını arttırmasına katkı sağlamaktadır. Ancak bu yorumların müşteriler tarafından tek tek okunması ve analiz edilmesi oldukça zaman alıcı ve zor bir süreçtir. Bu süreç aynı zamanda dikkat dağınıklığı gibi etkiler nedeniyle ürün yorumlarının analizinde insan hatası riskini de beraberinde getirmektedir. Müşterilerin ilgilendikleri ürün ve hizmet hakkında hızlı ve doğru karar verebilmelerini sağlamak için makine öğrenimi algoritmalarına dayalı dil analiz modellerinin geliştirilmesi üzerine çeşitli çalışmalar yapılmaktadır. Ele alınan bu tez çalışmasında BERT ve ALBERT dil modelleri kullanılarak, e-ticaret siteleri üzerinde yer alan ürünlere/hizmetlere ait müşteri yorumları kullanılarak ürün/hizmet hakkında duygu analizi yapılması planlanmıştır. Çalışmada elde edilen analiz sonuçlarının e-ticaret platformları üzerinde müşteri memnuniyetini artırması ve buna bağlı olarak pazarlama stratejileri ve ürün geliştirme gibi konulara katkı sağlaması öngörülmektedir.

Name and Surname : Suat ERKAN
Supervisor : Asst. Prof. Dr. Ediz ŞAYKOL
Degree and Date : Master's Thesis, 2023
Major : Computer Engineering
Key Words : ALBERT, BERT, Deep Learning, Sentiment Analysis

ABSTRACT

ANALYSIS OF COMMENTS ON E-COMMERCE SITES WITH BERT AND ALBERT LANGUAGE MODELS

Online shopping (e-commerce) platforms, which have become popular today, are often preferred by many people because they offer a fast, easy, secure and contactless shopping experience. Positive or negative comments made on online platforms about the product and service that customers apply for, in order to enable them to make the right decision before making the purchase during online shopping, play an important role in their shopping preferences. Customer reviews of products on e-commerce platforms help customers get a realistic idea about the product/service and help businesses improve their performance by evaluating these comments. However, reading and analyzing these comments one by one by customers is a time-consuming and difficult process. This process also brings the risk of human error in the analysis of product reviews due to effects such as distraction. Various studies are carried out on the development of language analysis models based on machine learning algorithms in order to enable customers to make fast and accurate decisions about the product and service they are interested in. In this study, it is planned to conduct a sentiment analysis about the product/service by using customer comments on products/services on e-commerce sites using BERT and ALBERT language models. It is predicted that the results of the analysis obtained in the study will increase customer satisfaction on e-commerce sites and contribute to issues such as marketing strategies and product development.

İÇİNDEKİLER

Sayfa No.

ÖZ

ABSTRACT

ŞEKİLLER LİSTESİ iii

KISALTMALAR iv

SÖZLÜK..... v

GİRİŞ..... 1

BİRİNCİ BÖLÜM

1. LİTERATÜR ÇALIŞMALARI..... 3

İKİNCİ BÖLÜM

2. MATERYAL VE YÖNTEM 8

2.1. Yapay Zekâ 8

2.2. Makine Öğrenmesi..... 9

2.3. Derin Öğrenme 11

2.4. Duygu Analizi..... 12

2.4.1. BERT 13

2.4.2. ALBERT 14

2.4.3. Naive Bayes 15

2.4.4. Rastgele Orman..... 17

2.4.5. Destek Vektör Makinesi..... 18

ÜÇÜNCÜ BÖLÜM

3. DENEYSEL BULGULAR..... 20

3.1. Veri Setinin Elde Edilmesi 20

3.1.1. Zemberek Kütüphanesi 21

3.2. Deneysel Çalışma Ortamı..... 23

3.3. Analiz Yöntemleri 24

3.4. Veri Seti Üzerinde Python Programlama Dili ile Duygu Analizi 25

3.4.1. BERT ve ALBERT Dil Modelleri ile Duygu Analizi	25
3.4.2. Makine Öğrenmesi Yöntemleri ile Duygu Analizi.....	28
SONUÇ VE ÖNERİLER	31
KAYNAKÇA.....	33



ŞEKİLLER LİSTESİ

Sayfa No.

Şekil 1. Yapay Zekâ, Makine Öğrenmesi ve Derin Öğrenme Arasındaki İlişki	9
Şekil 2. Makine Öğrenmesi Çeşitleri.....	10
Şekil 3. Derin Öğrenme Katman Yapısı.....	11
Şekil 4. BERT Model Yapısı.....	13
Şekil 5. ALBERT Model Mimarisi	15
Şekil 6. Naive Bayes Sınıflandırma Algoritma Yapısı.....	16
Şekil 7. Rastgele Orman Algoritması	18
Şekil 8. Destek Vektör Makine Algoritması	19
Şekil 9. Zemberek Kütüphanesi Çalışma Yapısı.....	21
Şekil 10. Zemberek Kütüphanesi ile Türkçe Metnin İşlenmesi.....	22
Şekil 11. Veri Setinin İlk 28 Verisi.....	23
Şekil 12. BERT ve ALBERT Dil Modelinin Çalışma Yapısı	25
Şekil 13. BERTurk Dil Modelinin Çağırılması.....	26
Şekil 14. ALBERT-Tr Dil Modelinin Çağırılması.....	26
Şekil 15. BERTurk Dil Modeli ile Veri Setinin Eğitilmesi.....	27
Şekil 16. ALBERT-Tr Dil Modeli ile Veri Setinin Eğitilmesi.....	27
Şekil 17. BERTurk Dil Modeli Analiz Sonuçları.....	28
Şekil 18. ALBERT-Tr Dil Modeli Analiz Sonuçları.....	28
Şekil 19. Naive Bayes Analiz Sonuçları.....	29
Şekil 20. Rastgele Ağaç Analiz Sonuçları.....	30
Şekil 21. Destek Vektör Makinesi Analiz Sonuçları.....	30

KISALTMALAR

ALBERT	: Representational State Transfer
BERT	: Platform As A Service
NB	: Naive Bayes
NLP	: Natural Language Processing (Doğal Dil İşleme, DDİ)
RF	: Random Forest (Rastgele Orman, RO)
SVM	: Support Vector Machine (Destek Vektör Makinesi, SVM)
YZ	: Yapay Zekâ



SÖZLÜK

Doğal Dil İşleme: İnsan dilini işlemek, anlamak, üretmek ve manipüle etmek için bilgisayar algoritmalarına dayalı çeşitli analiz yöntemlerinin kullanıldığı bir disiplindir.

Dil Modeli: Bir doğal dil işleme tekniği olarak belirli bir dildeki metnin olasılıklı yapılarını ve ilişkilerini öğrenen ve gelecekteki metin tahminleri yapmak için kullanılan bir matematiksel modeldir.

Yapay Zekâ: İnsanların düşünme, öğrenme, problem çözme ve karar verme gibi zekâsal faaliyetlerini taklit etmek için tasarlanan teknolojilerdir.

GİRİŞ

Günümüzde internetin hızlı yayılmasıyla birlikte, e-ticaret siteleri giderek daha popüler hale gelmektedir. Çevrimiçi satın alma işlemi yapanların sayısı arttıkça, e-ticaret sitelerindeki ürün yorumları da daha fazla önem kazanmaktadır. Yapılan araştırmalar, müşterilerin yaklaşık %95'inin ürünleri satın almadan önce daha önceden ürünü satın almış müşterilerin deneyimlerini ve düşüncelerini içeren ürün yorumlarını okuduğunu göstermektedir. Bu nedenle, e-ticaret siteleri için, ürün yorumlarından alınacak geri bildirimlerin takibi ve analizi oldukça önemlidir. E-ticaret siteleri bu geri bildirimler ışığında ürün ve hizmetlerini iyileştirmek için hızlı ve hatasız bir çalışma yapması gerekmektedir. Ancak, e-ticaret sitelerinde yer alan müşteri yorumlarının eski yöntemlerle analiz edilmesi büyük zaman kaybına sebep olduğu gibi zahmetli de bir işlemdir. Bu zorlukların üstesinden gelmek için, Doğal Dil İşleme (DDİ) teknolojileri ve makine öğrenmesi algoritmaları kullanılarak müşteri yorumlarının hızlı ve otomatik olarak analiz edilmesi önerilmiştir.

Yapay zekâ (YZ) teknolojileri, DDİ ve makine öğrenmesi algoritmaları, doğal dilleri anlamak ve yorumlamak için öğrenme yeteneğine sahiptirler. YZ teknolojileri, günümüzde e-ticaret sektöründe daha fazla kullanılmaktadır. Bu teknolojiler, ürün incelemelerini, müşteri hizmetlerini ve müşteri davranışlarını analiz etmek için kullanılabilir. Ayrıca, e-ticaret sitelerindeki ürün yorumlarını analiz etmek ve yorumların türünü, duygusal tonunu ve memnuniyet seviyesini sınıflandırmak gibi işlemler için de kullanılabilirler.

DDİ teknolojileri, doğal dili işlemek, anlamak ve yorumlamak için kullanılır. Bu teknolojiler, e-ticaret sitelerindeki müşteri yorumlarını analiz etmek ve ürün kalitesi, hizmet memnuniyeti, müşteri davranışları ve daha fazlası gibi konularda geri bildirimler sağlamak için kullanılabilir. DDİ teknolojilerinin en yaygın kullanım alanlarından biri, dil modelleri kullanarak otomatik metin üretme ve yorum analizi yapma işlemleridir.

Doğal dil işleme teknolojileri arasında BERT ve ALBERT gibi dil modelleri de yer almaktadır. Google 2018 yılında BERT (Bidirectional Encoder Representations from Transformers) dil modelini literatüre kazandırmıştır. BERT, geniş bir metin veri kümesinde ön eğitilerek, doğal dil anlama işlemleri için kullanılan en gelişmiş

teknolojilerden biridir. ALBERT (A Lite BERT), Google ve Toyota Technological Institute at Chicago (TTIC) tarafından geliştirilmiş bir dil modelidir ve BERT modeline göre daha az parametre kullanarak daha yüksek performans gösterir. ALBERT dil modeli BERT dil modelinin optimize edilmiş bir versiyonudur.

Bu tez e-ticaret sitelerinde yer alan yorumların BERT ve ALBERT dil modelleri ile analizini ele almaktadır. Bu tez çalışmasında BERT ve ALBERT dil modelleri kullanılarak e-ticaret sitelerindeki ürün yorumları analiz edilecek ve yorumların duygu durumu, memnuniyet seviyesi ve diğer özellikleri sınıflandırılacaktır. Bu çalışmanın amacı e-ticaret sitelerinin daha iyi bir kullanıcı deneyimi sağlamasına yardımcı olmak ve müşterilerin geri bildirimlerini daha etkili bir şekilde analiz etmelerine yardımcı olmaktır.

Bu çalışma BERT ve ALBERT dil modellerinin e-ticaret sitelerindeki ürün yorumlarının analizindeki etkinliğini ve performansını karşılaştıracak ve hangi modelin daha iyi sonuçlar verdiğini belirlemeye çalışacaktır. Bu çalışmanın sonuçları e-ticaret sitelerinin ürün yorumlarını analiz etmek için hangi dil modelinin daha iyi performans gösterdiği konusunda bilgi sağlayabilir ve e-ticaret sitelerin işlerinde iyileştirmeler yapmalarına yardımcı olacağı düşünülmüştür.

Sonuç olarak bu tez e-ticaret sitelerindeki ürün yorumlarının analizi için BERT ve ALBERT dil modellerinin kullanımını ele almaktadır. Bu çalışmanın sonuçları, doğal dil işleme teknolojileri alanındaki gelişmeler hakkında daha fazla bilgi sağlayabilir ve gelecekteki araştırmalar için bir temel oluşturabilir. Makine öğrenmesi yöntemleri ile dil modellerinin duygu analizine etkisini incelemek için karşılaştırma yapılmıştır.

Bu tez çalışmasının amacı BERT ve ALBERT dil modellerinin çalışma mantığının incelenmesi ve aralarındaki benzerlik ve farklılıkları ortaya koyarak nasıl etkilerinin olduğu ortaya konmuştur. Böyle bir kıyaslama sayesinde daha az parametreye sahip ALBERT dil modeli ile büyük veri setlerinde BERT dil modeline kıyaslan daha hızlı sonuçlar alabildiği ama daha az parametrelere sahip olduğu için veri setinin büyümesi durumunda doğruluk oranında azalma olduğu gibi bilgiler elde edilebilmektedir. Çalışmanın ilerleyen bölümlerinde literatür çalışması yapılarak dil modelleri ve makine öğrenmesi yöntemlerinde yapılmış çalışmalar incelenmiştir. Sonraki bölümlerde materyal ve yöntem ve deneysel bulgular sunulmuştur.

BİRİNCİ BÖLÜM

1. LİTERATÜR ÇALIŞMALARI

Son yıllarda teknolojiye gerçekleşen gelişmeler ışığında insan duygularının tespitiye yönelik yapılan çalışmalar ve analizler hız kazanmıştır. Makine öğrenmesi yöntemlerinin çeşitli platformlara entegre edilmesiyle birlikte bu analizler ve çalışmalar giderek artmaktadır. İnsanların internet aracılığı ile veri aktarımı, araştırma alanlarındaki çeşitliliği arttırmaktadır.

Günümüzde yapay zekanın alt dallarından olan doğal dil işleme dil modelleri kullanarak insan dilini anlayan ve işleyen bilgisayar sistemleri oluşturmayı hedefler. Dil modelleri büyük veri setleri üzerinde eğitilerek dilin yapısal özelliklerini, kelime anlamlarını ve cümlelerin bağlamsal ilişkilerini öğrenir. Bu sayede dil modelleri doğal dilde yazılmış veya konuşulmuş verileri işleyerek çeşitli uygulamalar için kullanılabilir (örneğin; otomatik metin çevirisi, duygu analizi, konuşma tanıma, makine öğrenmesi ve daha birçok alanda kullanılabilir).

Doğal dil işleme (DDİ) alanında birden fazla model tasarımı ve yöntemlerinin ortaya çıkması ile Young ve vd., (2018) derin öğrenme yöntemlerinin DDİ’de yaygınlaştığı ve başarılı olduğunu incelemiştir. İlgili çalışmada farklı dil modellerinin açıklaması ve karşılaştırılması yer almaktadır. Derin öğrenmenin DDİ’deki geçmişini ve geleceği hakkında durumu detaylı şekilde ele alınmıştır.

Zhang ve vd., (2018) çalışmalarında derin öğrenme içerisinde duygu analizinin uygulanmasını incelemiştir. Farklı derin öğrenme yöntemlerinin duygu analizi performansı üzerindeki etkisini ve mevcut araştırmaların yapısal bir analizini göstermektedir. Bu çalışma ile derin öğrenme modellerinin duygu analizi alanında önemli bir potansiyele sahip olduğunu göstermiştir.

Goldberg, (2017) yaptığı araştırmalarda doğal dil işleme alanında kullanılan yaygın derin öğrenme yöntemlerine odaklanmıştır. Goldberg yapay zekâ içerisinde yer alan dil modellerinin metin sınıflandırma, duygu analizi, dil çevirisi, kelime sıralama ve diyalog sistemleri gibi doğal dil işleme alanındaki uygulama yöntemleri hakkında temel kavramları ele almaktadır.

Liang ve Wang (2019) ele aldıkları çalışmada kullanıcıların arzu ettikleri ürünleri çevrimiçi olarak keşfetmelerine yardımcı olabilecek bir bütünleşik karar

destek modelini literatüre sunmuştur. Geliştirdikleri model ürün özelliklerini ve müşteri yorumlarını içeren bir küme oluşturarak, ürün seçiminde müşterilerin bulanık karar verme sürecini modellemiştir. Bu modelleme ile müşterilerin çevrimiçi ortamda hızlıca alışveriş yapabilmeleri kolaylaştırılmıştır. Sanal ortamda tüketiciyle buluşan ürünlerin gerçek ürünler ile karşılaştırıldığında farklılıkların düşük seviyede olabilmesi için duygu analizinin yapılması önem göstermiştir. Yang ve vd. (2020) literatüre kazandırdıkları yeni bir dil modeli ile Çince dilinde duygu analizini amaçlamışlardır. İlgili çalışmalarında evrişimli sinir ağlarını temel alarak Çince e-platform üzerinde ürün yorumlarını incelemişlerdir. Bu çalışma ile derin öğrenmenin duygu analizi alanında etkili bir araç olduğu ve duygu sözcüklerinin derin öğrenme yöntemleri ile birleştirilerek daha faydalı sonuçlar alınabileceği gösterilmiştir.

Zeng ve vd. (2019) çalışmalarında insanların yaptıkları yorumlarda belirli konuların duygusal yönelimlerini tespit edebilmek için evrişimli sinir ağları mimarisi kullanarak bir dil modeli geliştirmiştir. Bu yöntem ile yüksek doğruluk elde edildiği için hem dilbilimsel hem de makine öğrenimi açısından birleştirilmiş bir yaklaşım sunarak, ürün yorumlarının analizlerini kolaylaştırmıştır.

Gürçan, (2018) çalışmasında Türkçe metinlerin sınıflandırılması için Multinomial Naive Bayes, Bernoulli Naive Bayes, Karar Ağaçları gibi farklı makine öğrenimi algoritmalarını kullanmıştır. Çalışmada veri seti olarak kullanılan Türkçe haber metinleri ilgili algoritmalar ile sınıflandırılmıştır. Elde edilen sınıflandırma sonuçları üzerinden Türkçe doğal dil işleme araçları ve klasik makine öğrenmesi algoritmalarının performansı yorumlanmıştır.

Demir (2021) ele aldığı tez çalışmasında Türkçe ses kayıtlarının olduğu bir veri seti üzerinde BERT dil modeli, TextBlob kütüphanesi ve Naive Bayes, Rastgele Orman, Destek Vektör Makinesi, Karar Ağacı ve K-En Yakın Komşu gibi popüler makine öğrenmesi yöntemlerini uygulamıştır. Common Voice platformunda veri setini metin hale getirerek duygu analizi çalışmasını gerçekleştirmiştir.

Güven, Z.A. tarafından 2021'de ele alınan çalışmada Türkçe ürün yorumlarını içeren veri seti üzerinde BERT, ALBERT ve Electra dil modellerinin yanı sıra Rastgele Orman, Naive Bayes ve Lojistik Regresyon makine öğrenmesi yöntemleri uygulanmıştır. Bu çalışmada Electra ve ALBERT dil modellerinin makine öğrenmesi yöntemlerinden daha başarılı olduğunu tespit etmiştir (Güven, 2021).

Raffel ve vd. (2019) alışmalarında transformatör tabanlı modellerin transfer öğrenme kabiliyetlerini derinlemesine incelemiştir. Ele aldıkları bu alışmada metin tabanlı dil modellerini Text-to-Text modeli ile “metinden metine” formatına dönüştürerek NLP için transfer öğrenme tekniklerini kullanmışlardır. Bu alışmada temel fikir, her metin işleme problemini bir “metinden metine” problemi olarak ele almak, yani metni girdi olarak alıp yeni metni çıktı olarak üretmektir.

Vaswani ve vd. (2017) dil modellemesi ve DDİ alanında büyük bir etki yaratan “attention” yapısını ortaya ıkarmışlardır. Bu yapı ile modellerin belirli bir girdi üzerinde yoğunlaşması ve bu şekilde önemli bilgileri daha iyi yakalamasını sağlamıştır. İlgili alışma kapsamında transformatör tabanlı modellerin temelini oluşturmuş ve ilerlemesinde büyük katkı sağlamıştır. (Vaswani vd., 2017)

Bu tez alışmasının temelini oluşturan BERT dil modeli Devlin ve vd. (2019) tarafından literatüre kazandırılmıştır. Devlin ve arkadaşları ilgili alışmada cümlelerin sol ve sağ bağlamlarını kullanarak kelime veya cümleler arasındaki ilişkileri anlamaya alışan bir dil modeli geliştirmiştir. BERT dil modelinin duygu analizi, soru cevaplama metin sınıflandırma gibi işlemlerde yüksek başarılar elde etmiştir.

ALBERT dil modeli Lan ve vd. (2020) tarafından BERT dil modeli optimize edilerek kendi kendine denetimli öğrenme modeli olarak geliştirilmiştir. ALBERT dil modeli düşük seviyede kaynak tüketimi sağlaması ve eğitim süresinin BERT dil modeline göre daha kısa olması nedeniyle popüler hale getirmiştir. Bu alışmanın DDİ alanında daha az parametrelere sahip ancak daha iyi sonuçlar veren dil modellerinin gelişmesine ışık tutacağı düşünülmüştür. (Lan vd., 2020)

Açık kaynak kütüphanelerinde yabancı diller için bir kütüphane olmasına rağmen Türkçe dil ailesini destekleyen bir kütüphanenin olmaması Türkçe metinlerin DDİ alanında analizini olanaksız kılmıştır. Türkçe dil ailesini içeren kütüphane modeli ilk olarak Kemal Oflazer tarafından 1993 yılında literatüre sunulmuştur (Oflazer, 1993).

Literatürde Zemberek adı ile bilinen açık kaynak kütüphanesi üzerinde alışan Akın ve Akın 2007’de Türkçe dil ailesine ait kütüphaneyi farklı platformlarda alışması amacıyla java tabanlı hale getirilmiştir. Kelime köklerinin çözümlenmesi, yazım denetiminin yapılması, dil tanıma gibi yetenekler ortaya ıkarılmıştır (Akın ve Akın, 2007).

Tunalı ve Bilgin (2012) çalışmalarında Zemberek, Ek Çıkaran ve Sabit Önek gibi farklı açık kaynak kütüphanelerini kullanarak Türkçe metinler üzerindeki etkilerini araştırmıştır. Elde edilen sonuçlar incelendiğinde Zemberek ve Sabit Önek kütüphaneleri ile oluşturulan veri setlerinin diğer yöntemlere kıyasla daha verimli sonuçlar ürettiği gözlenmiştir.

Literatürde yer alan çalışmalar incelendiğinde duygu analizi çalışmalarında BERT ve ALBERT dil modelleri haricinde makine öğrenmesi yöntemlerinden Naive Bayes, Random Forest ve Support Vector Machine yöntemleri de kullanılmıştır. Yu ve vd. (2021) ilgili çalışmalarında örnek ağırlıklı gizli Naive Bayes (Instance Weighted Hidden Naive Bayes-IWHNB) adı verilen yeni hibrit bir yaklaşım ile Naive Bayes algoritmasının yeni bir uygulamasını incelemiştir. Bu çalışma ile Naive Bayes algoritmasının daha fazla hassasiyet ve farklı alanlarda uygulanabilirliği için optimize edilebilmesi hakkında araştırma gerçekleştirilmiştir.

Probst ve Boulesteix (2017) araştırmalarında Random Forest algoritmasının önemli parametrelerinden biri olan ağaç sayısının ayarlanması konusunu ele almıştır. Ağaç sayısının algoritma üzerindeki etkisini inceleyen ve uygun ağaç sayısının belirlenmesini hedefleyen incelemeler gerçekleştirilmiştir.

Cortes ve Vapnik tarafından ele alınan “Support Vector Networks” başlıklı makalede, Destek Vektör Makinelerinin (Support Vector Machine-DVM) temel ilkeleri ve bu teknolojiye ilişkin yeni teorik sonuçlar sunulmuştur (Cortes ve Vapnik, 1995). Bu çalışmada sınıflandırma problemlerinin çözümünde kullanılması amaçlanan DVM algoritmalarının özellikle öğrenme problemlerinin analizinde kolaylık sağladığı gösterilmiştir. Literatürde yer alan pek çok çalışmanın temelini oluşturan DVM algoritmaları bir veriyi çok boyutlu uzayında doğrusal olarak ayrılabilir hale getirebilmektedir. Ek olarak düşük genelleştirme hatası, hesaplama karmaşıklığının düşük olması ve modelin parametrelerinin belirlenmesi DVM algoritmalarını analizlerde tercih edilir kılmaktadır (Al Amrani vd., 2018; Bourequeatve Mourad, 2021; Gupta vd., 2019; Somvanshi ve Chavan, 2016; Zheng ve Ye, 2009)

Choi ve vd. (2021) ele aldıkları çalışmada BERT ve ALBERT dil modellerinin, farklı DDİ problemleri için performans kıyaslamasını incelemiştir. Yapılan çalışmada Morfolojik Analiz, Morfolojik Üretim, Adlandırılmış Varlık Tanıma, Kelime Kökü Bulma ve Ek Ayırma, Dil Tanıma ve Heceleme gibi özellikleri incelenmiş ve

ALBERT dil modelinin BERT dil modeline göre daha hızlı ve verimi yüksek sonuçlar elde ettiğini tespit etmiştir. Bu ve benzeri çalışmalar literatüre yeni kazandırılmış BERT ve ALBERT dil modellerini bilim insanları tarafından tercih edilir kılmaktadır (Alrowili ve Shanker, 2021; Chen vd., 2021; Choi vd., 2021a; Choi vd., 2021b; Koto vd., 2021; Lan vd., 2020; Lee vd., 2021; Zhao vd., 2020).

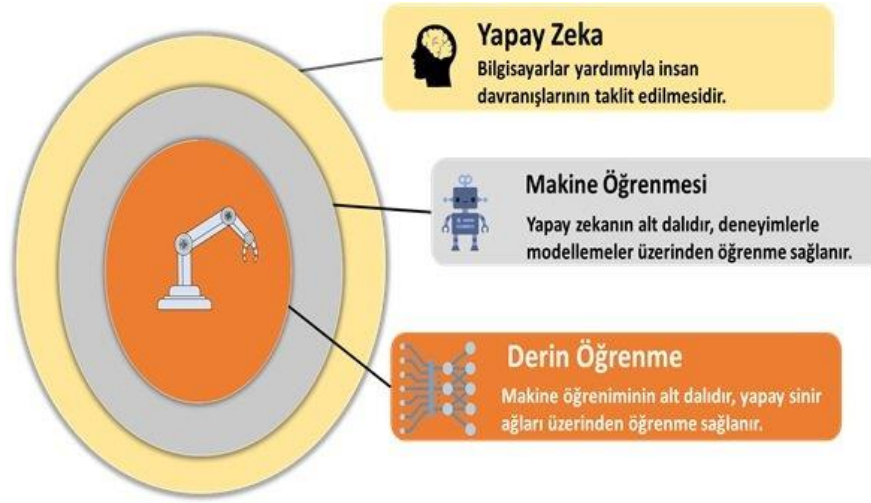


İKİNCİ BÖLÜM

2. MATERYAL VE YÖNTEM

2.1. Yapay Zekâ

Yapay zekâ, bilgisayar sistemlerinin insan zekâsını taklit ederek öğrenme, anlama, çıkarım yapma, planlama ve problem çözme yeteneklerine sahip olmasını sağlayan bir bilim dalıdır (Russell ve Norvig, 2020). Yapay zekâ üzerine 1950 yılında Alan Turing “Makine ve Zekâ” çalışmasında “Makineler Düşünebilir mi?” sorusu ile zekâ ve bilgisayar biliminin sınırlarını zorlayarak makinelerin programlanabileceği fikrini ortaya atmıştır (Turing, 1950). Yapay zekâ geniş uygulama yelpazesi ve potansiyeli ile bilim, teknoloji, tıp, iş dünyası ve daha pek çok alanda büyük atılım sağlamıştır. Yapay zekâ teknolojileri, büyük verinin önem arz ettiği bu zamanlarda karar verme süreçlerimizi, üretim yöntemlerimizi ve yaşam tarzlarımızı dahi değiştirmektedir. Makine öğrenmesi ve derin öğrenme teknikleri, karmaşık veri kümelerinin analizinde ve değerlendirilmesinde önem arz etmektedir. Bu yöntemlerle beraber tıbbi teşhisler, finansal tahminler, hava durumu tahminleri ve özelleştirilmiş reklamlar gibi uygulamalarda kullanılmaktadır. İlerleyen yıllarda yapay zekânın insan zekâsını taklit etme ve aşma potansiyeli olduğu öngörülmektedir. Aynı zamanda insanların öğrenme, karar verme ve problem çözme yeteneklerini simüle edebilen derin öğrenme sistemleri gibi teknolojilerin geliştirilmesini içermektedir (LeCun vd., 2015).



Şekil 1. Yapay Zekâ, Makine Öğrenmesi ve Derin Öğrenme Arasındaki İlişki
Kaynak: Ateş, E. C. (2021). *Yapay Zekâ (Artificial Intelligence)*, pp. 644-647.

2.2. Makine Öğrenmesi

Bilgisayarların veriden öğrenme ve karar verme yetenekleri üzerine geliştirme sağlayan bir yapay zekâ alt dalıdır (Samuel, 1959). Yapay zekânın bu dalı, özellikle büyük veri döneminde çeşitli sektörlerde karmaşık sorunları çözme yeteneği sayesinde önem kazanmıştır. Algoritmalar veriye maruz bırakıldıkça ve yeni verilerle beslendikçe sorumluluk sahasında performanslarını arttırmaktadır. Makine öğrenmesi, Supervised (Gözetimli), Unsupervised (Gözetimsiz), Semi-supervised (Yarı Gözetimli) ve Reinforcement (Destekli) öğrenme teknikleri olarak dört ana gruba ayrılmıştır (Hoş, 2021).

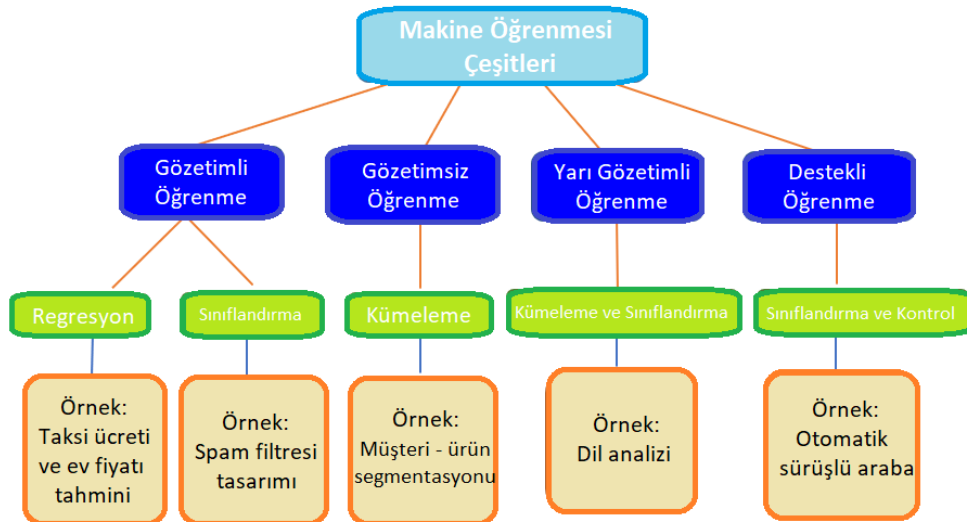
Gözetimli Öğrenme: Modelin öğrenme süreci, genellikle bir etiketli veri seti kullanılarak gerçekleştirilir. Etiketli veri seti hem girdi hem de bu verilerin hangi kategorilere ait olduğunu (örneğin, bir resmin bir köpeği veya bir kediye temsil etmesi veya bir metin parçasının olumlu veya olumsuz bir duyguyu ifade etmesi vb.) belirten etiketleri içerir. Model bu etiketli veri seti üzerinde eğitilir ve böylece girdi verilerinden çıktı etiketlerini tahmin etmek için kullanılacak bir fonksiyon veya harita oluşturur

Gözetimsiz Öğrenme: Verinin doğal yapısını keşfetmek için etiketlenmemiş verilerle çalışmaktır. Model belirgin kalıpları veya yapıları bulmaya çalışır ve genellikle

kümeleme veya yoğunluk tahmini gibi teknikler kullanılmaktadır. Gözetimsiz öğrenme algoritmaları genellikle büyük ve karmaşık veri kümelerinde kullanılır, çünkü bu tür veri kümeleri genellikle el ile etiketlenemez veya insan gözlemciler için anlamlı bir şekilde analiz edilemez. Bu nedenle, gözetimsiz öğrenme, veri madenciliği ve büyük veri analitiği alanlarında özellikle yaygın olarak kullanılmaktadır.

Yarı Gözetimli Öğrenme: Gözetimli ve gözetimsiz öğrenmenin birleşimini kullanmaktadır. Etiketlemenin maliyetli veya zaman alıcı olduğu durumlarda kullanılmaktadır. Çoğu gerçek dünya verisi etiketsizdir. Etiketlemenin hem zaman alıcı hem de maliyetli bir süreç olması nedeniyle, tüm verilerin etiketlenmesi genellikle pratik olmaz. Yarı-gözetimli öğrenme bu tür durumlar için bir çözüm sunmaktadır.

Destekli Öğrenme: Bu metodoloji bir modeli eğitmek için hem girdi (input) hem de çıktı (output) verilerinin kullanıldığı bir yaklaşımdır. Gözetimli öğrenmenin amacı girdi verilerinden çıktıyı tahmin etmek için bir işlev öğrenmektir. Bu türden bir öğrenme bir dizi örnekleme veya örnek veri seti üzerinde gerçekleştirilmektedir. Her örnek bir girdi ve beklenen çıktıyı içermektedir. Bir öğrenme algoritması, girdi ve çıktı arasındaki ilişkiyi temsil eden bir model oluşturur. Model öğrenme süreci sırasında yeterince iyi performans gösterene kadar iteratif olarak ayarlanır. Makine öğrenmesi çeşitleri **Şekil 2**'de yer verilmiştir.



Şekil 2. Makine Öğrenmesi Çeşitleri

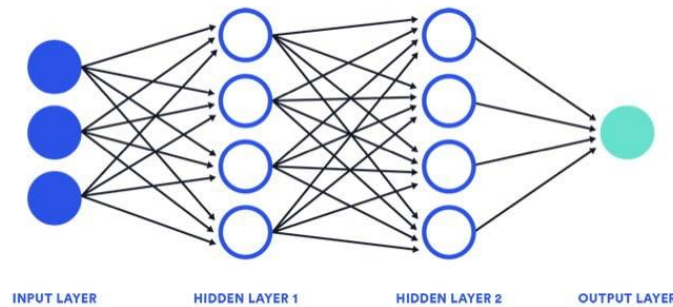
Kaynak: <https://technogezgin.com/makine-ogrenmesi-nedir-ne-ise-yarar/#>, E.T: 15.04.2023.

2.3. Derin Öğrenme

Derin öğrenme, insan beyninin işleme yapısını taklit eden ve büyük miktarda veriden karmaşık özellikler öğrenme yeteneğine sahip olan yapay sinir ağları üzerine kurulu gelişmiş bir makine öğrenme alt dalıdır (LeCun vd., 2015). Çok katmanlı yapay sinir ağları kullanarak derin öğrenme algoritmaları, görüntü ve ses tanıma, sınıflandırma, regresyon, karmaşık desenleri tanıma ve doğal dil işleme gibi bir dizi görevi yerine getirebilmek için geniş bir uygulamaya sahiptir. Derin öğrenmenin popülaritesi ve başarısı büyük ölçüde artan veri miktarları ve hızlı donanım gelişimine dayanmaktadır. Modelin karmaşıklığı katman sayısının artmasıyla doğru orantılıdır.

Bu teknik özellikle büyük veri setleri ve karmaşık problemlerle başa çıkmak için güçlü bir araç olarak görülmektedir. Derin öğrenme süreçlerinde bir sistemdeki hataları düzeltmek için basit yönergelerle gerçekleştirilebilir ve böylece öğrenme sürecini düzenlenebilir. Zaman içerisinde yapay sinir ağları teknolojisi ilerlemiş ve karmaşık problemlerin çözümü için derin öğrenme tekniklerine dönüşmüştür (Kaya vd., 2019).

Böylece, derin öğrenme teknikleri geniş ve karmaşık veri setlerini analiz etme yetenekleri ile öne çıkmaktadır. Bu yetenek aynı zamanda, derin öğrenmeyi büyük veri analizi ve karmaşık problem çözümü için ideal bir yöntem haline getirmektedir. Derin öğrenme, yapay zekâ ve makine öğrenmesi alanındaki en heyecan verici gelişmelerden biri olarak kabul edilir ve birçok uygulama alanında potansiyel olarak dönüştürücü bir etkiye sahip olmaktadır. Derin öğrenmenin çalışma mantığını gösteren görsel **Şekil 3**'te verilmiştir.



Şekil 3. Derin Öğrenme Katman Yapısı

Kaynak: Zaloğlu O. (2023), Dünyayı değiştirmekte olan yapay sinir ağları nedir? <https://bilimfili.com/dunyayi-degistirmekte-olan-yapay-sinir-aglari-nedir>. E.T:15.04.2023

2.4. Duygu Analizi

Duygu analizi, bir veri parçasının pozitif, negatif veya nötr duygusal tonunu değerlendiren bir yaklaşımdır ve son yıllarda veri biliminin kilit bir odak noktası olmuştur. Çeşitli araştırmacıların bu alana olan ilgisinin artması, duygu analizinin kendi içerisinde bir dizi alt kategoriye ayrılmasına yol açmıştır. Bu alt kategoriler; belge düzeyinde, cümle düzeyinde, özellik tabanlı ve karşılaştırmalı duygu analizini içerir. Bu dallar, belirli bir metinden çıkarılan duygusal bilgilerin çeşitli yönlerini vurgular. Bu nedenle duygu analizi genellikle bir metnin genel tonunu anlamak için karmaşık bir yaklaşım gerektirir (Yurt, 2015).

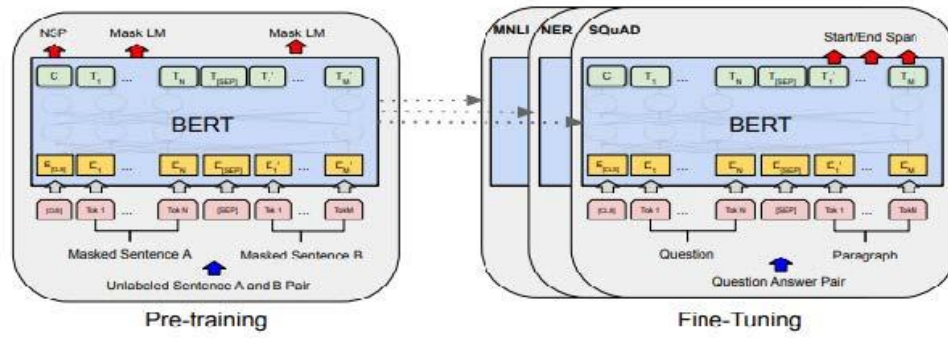
Duygu analizi, verinin türüne veya boyutuna bakılmaksızın uygulanabilir bir araçtır. Metin veya ses kaydından oluşan veri setinde duygu analizi için veriye dahil olan duygusal tonları belirlemek için kullanılabilir. Verilerin ön işleme tabi tutulduktan sonra analiz edilmesi, çeşitli hipotezlerin formülasyonuna izin verir. Örneğin; bir kişi tarafından bir e-ticaret platformunda yazılan bir ürün değerlendirmesi hem ürün hakkında bilgi sağlayabilir hem de potansiyel alıcıların satın alma kararlarını şekillendirebilir. Bu durum, duygu analizinin pratik ve teorik önemini vurgular. Verinin duygusal boyutunu anlamak, genellikle belirli bir metnin veya konuşmanın daha geniş bağlamını anlama anahtarını sağlar.

Duygu analizi genellikle doğal dil işleme (DDİ), derin öğrenme, metin madenciliği ve makine öğrenmesi tekniklerini kullanılmaktadır. Özellikle derin öğrenme duygu analizinde önemli bir rol oynamaktadır. Örneğin, yinelemeli sinir ağları (Recurrent Neural Network-RNN) ve dikkat mekanizmaları gibi derin öğrenme teknikleri metnin semantik ve bağlam bilgisini yakalama yeteneği sayesinde duygu analizinde etkili olmuştur. Bu alandaki en son gelişmelerden biri BERT gibi dönüşümlü dil modellerin kullanılmasıdır. Bu modeller duygu analizi için özellikle etkilidir çünkü dilin çift yönlü doğasını modellemek için tasarlanmışlardır ve bu da duygusal tonun belirlenmesinde büyük öneme sahiptir.

Bu tez çalışmasında duygu analizi için BERT ve ALBERT dil modelleri ile makine öğrenmesi yöntemlerinden Naive Bayes, Rastgele Orman ve Destek Vektör Makinesi tercih edilmiştir. Sonraki bölümlerde söz konusu yöntemlerin tanımları ve açıklamaları ele alınarak kullanılan tekniklerin temelleri ve önemi anlatılmıştır.

2.4.1. BERT

Google firmasında çalışan Jacob Devlin ve ekip arkadaşları tarafından 2019'da oluşturulan bu model, cümleleri hem sol hem de sağdan okuma yeteneği sayesinde kelime ilişkilerini daha etkin bir şekilde çözümlemektedir (Devlin vd., 2019). Bu çözümleme yöntemi sayesinde kelimelerin kendi aralarında oluşturduğu ilişkileri daha gerçekçi yansıtmaktadır. BookCorpus ve Wikipedia gibi geniş kapsamlı veri kümesi kaynakları modelin temelini oluşturmuştur. BERT dil modelinin çalışma yapısı Şekil 4'te verilmiştir.



Şekil 4. BERT Model Yapısı

Kaynak: Devlin, J., Chang, M.-W., Lee, K., ve Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. İçinde J. Burstein, C. Doran, ve T. Solorio (Ed.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)* (ss. 4171-4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/n19-1423>.

BERT modeli, Masked Language Modeling (MLM) ve Next Sentence Prediction (NSP) adı verilen iki temel yöntemle eğitilerek ortaya çıkmıştır. MLM yaklaşımında model maskelenmiş bir kelimenin ne olabileceğini ve çevresindeki kelimelerden elde ettiği bağlam bilgisine dayanarak tahmin etmeye hedefler. MLM esas olarak kelimeler arasındaki ilişkilere yoğunlaşır. NSP yaklaşımında model bir sonraki cümlelerin öncekinin devamı olup olmadığının tahminine odaklanır ve böylece cümleler arası ilişkilerin önemini vurgular. Optimizasyon süreci bu iki yöntemin kombinasyonunu kullanarak modelin eğitim süresince ortaya çıkan hataların en aza indirilmesini amaçlar (Uçar, 2020). BERT modelinin çift yönlü metin analizi yeteneği ve karmaşık eğitim teknikleri daha derin ve daha etkili bir öğrenme süreci sağlamaktadır.

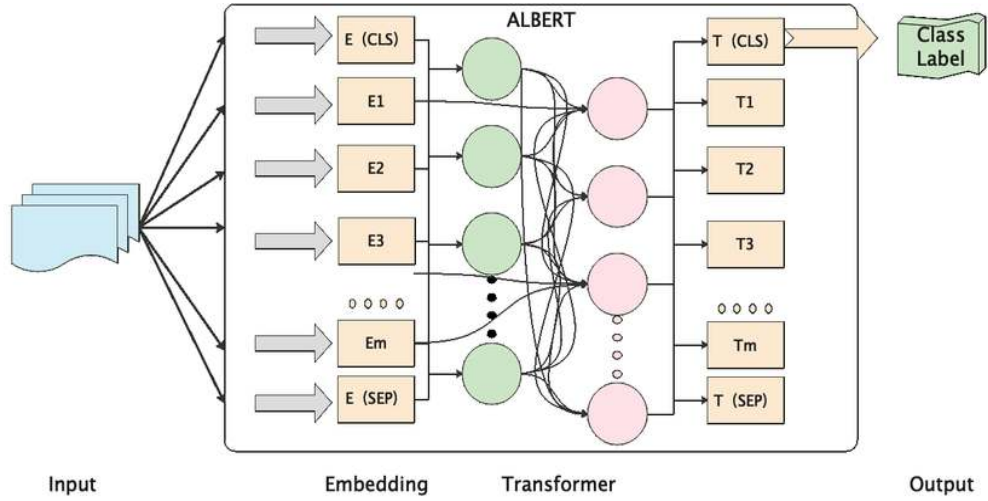
BERT dil modelinin kullanılmaya başlandığı ilk yıllarda dünya üzerinde en çok kullanılan dilleri kapsayan yapıyı temel almıştır. Ancak, zamanla çeşitli dil eklemeleriyle, BERT dil modeli çok dilli bir çerçeve geliştirmeye devam etmiştir. Türkçe diline yönelik geliştirilen BERT modeli, *BERTurk* olarak adlandırılmıştır. BERTurk'ün gelişiminde Prof.Dr.Kemal Oflazer, Reyhan Yeniterzi ve Stefan Schweter gibi bilim insanlarından oluşturduğu ekip ile Türkçe dilinde doğal dil işleme çalışmaları BERT dil modeli aracılığı ile gerçekleştirilmeye başlanmıştır. Türkçe BERT, varlık etiketleme, soru yanıtlama, belge sınıflandırma ve duygu analizi gibi çeşitli doğal dil işleme işlemlerinde kullanılmaktadır (Demir, 2021).

2.4.2. ALBERT

ALBERT dil modeli BERT dil modelinin gelişmiş bir sürümü olup çeşitli doğal dil işleme görevlerinde üst düzey performans sağlamaktadır. Aslında BERT dil modelinin daha hafif ve verimli bir sürümü olarak düşünülmektedir. Google Research ve Toyota Technological Institute at Chicago tarafından oluşturulan ALBERT, BERT dil modelinin bellek tüketimini ve eğitim hızını iyileştiren birkaç optimizasyon içerir. BERT dil modeline kıyasla daha az parametre kullanır böylece model eğitim süresini ve kaynak kullanımını önemli ölçüde azaltır. ALBERT, BERT'ten farklı olarak, parametre paylaşımı yaklaşımını ve çapraz katman normalleştirme adı verilen bir teknik kullanır (Lan vd., 2020).

ALBERT, iki önemli tasarım değişikliğiyle BERT dil modelinin boyutunu ve hesaplama gereksinimlerini azaltmıştır. Parametre paylaşımı ve çapraz katman normalizasyonu modelin büyük değişim geçirmesini sağlamıştır. Parametre paylaşımı, model boyutunu ve eğitim süresini azaltırken, çapraz katman normalizasyonu modelin daha derin öğrenme yeteneğini artırır. Bu özellikler, ALBERT dil modelinin BERT dil modeline kıyasla performansını korurken daha hızlı ve verimli olmasını sağlamaktadır.

Bu modelin hız ve verimlilik avantajları nedeniyle metin sınıflandırma, adlandırılmış varlık tanıma gibi birçok DDİ görevinde kullanılmıştır. Ayrıca, dil çeşitliliği ve transfer öğrenme kabiliyeti nedeniyle birçok farklı dilde doğal dil işleme görevlerine uygulanabilmektedir. ALBERT dil modelinin mimarisi **Şekil 5**'te verilmiştir.



Şekil 5. ALBERT Model Mimarisi

Kaynak: Li, J., Wang, B., ve Ding, H. (2020, December). Lijunyi at SemEval-2020 Task 4: An ALBERT Model Based Maximum Ensemble with Different Training Sizes and Depths for Commonsense Validation and Explanation. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 556-561.

2.4.3. Naive Bayes

Naive Bayes sınıflandırıcısı olasılık teorisine dayalı basit ve etkili bir makine öğrenmesi algoritmasıdır. İsmi olasılık teorisindeki Bayes Teoremi'nden alır ve naive (saf) kelimesi, özellikler arasındaki bağımsızlık varsayımına atıfta bulunmuştur (Zhang, 2004). Naive Bayes, özellikle büyük veri setlerinin işlenmesinde ve metin sınıflandırma gibi alanlarda etkilidir.

Naive Bayes algoritmasının çalışma prensibi her bir sınıfın olasılığını Bayes teoremini kullanarak belirlenmiş bir veri noktasıyla hesaplamaktır. Özellikler arasındaki bağımsızlık varsayımı hesaplamaları basitleştirir ve bu varsayımı sayesinde gerçek dünya verileri için tamamen doğru olmasa da Naive Bayes sınıflandırıcısı genellikle iyi sonuçlar vermiştir.

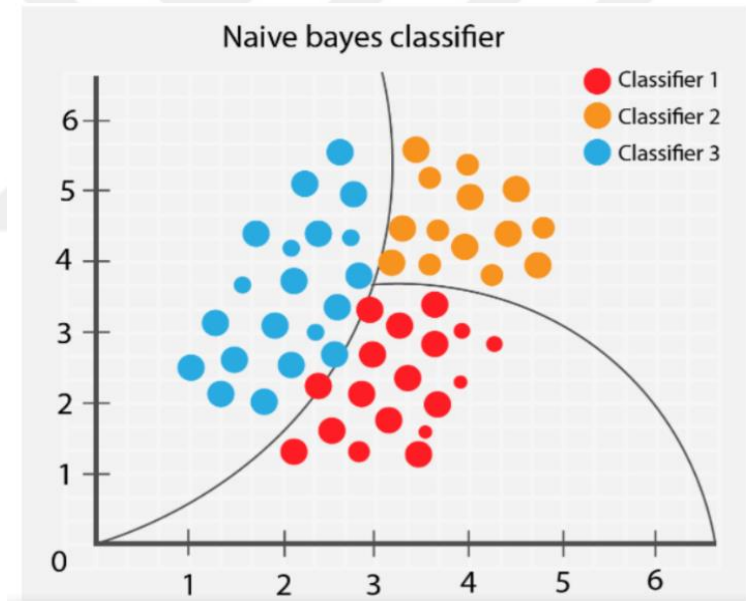
Naive Bayes'in avantajlı yanlarından birisi hesaplamamanın hızlı ve verimli olmasıdır. Bu büyük veri setlerinde ve özellikle de metin sınıflandırma gibi alanlarda, sınıflandırma problemlerinin çözümünde önemlidir. Ayrıca, Naive Bayes algoritmasının parametre ayarlama gerektirmemesi, modelin uygulanmasını ve kullanılmasını kolaylaştırmıştır. Naive Bayes teoremi denklem (2.1) ile gösterilmiştir.

$$P(A|B) = \frac{P(A|B)P(A)}{P(A)} \quad (2.1)$$

Naive Bayes sınıflandırma algoritması denklem (2.2)'de verilmiştir.

$$P(Class_j|x) = \frac{P(x|Class_j)P(Class_j)}{P(x)} \quad (2.2)$$

Yukarıda yer alan denklemde x tahmin olasılığı, j sınıfına ait olasılık ve eğitim veri setinde bulunan bu sınıfın tahmin edici olasılığıyla x tahmin edicisinin bu sınıfa ait olasılığı belirlenebilir. **Şekil 6**'da Naive Bayes sınıflandırılmasına ait algoritma gösterilmiştir.



Şekil 6. Naive Bayes Sınıflandırma Algoritma Yapısı

Kaynakça: Kaya, U., Yılmaz, A., ve Dikmen, Y. (2019). Sağlık Alanında Kullanılan Derin Öğrenme Yöntemleri. *European Journal of Science and Technology*, 792-808
<https://doi.org/10.31590/ejosat.573248>.

2.4.4. Rastgele Orman

Rastgele Orman (RO) makine öğrenmesinde sınıflandırma ve regresyon için kullanılan güçlü bir algoritmadır. Bu yöntem birçok karar ağacını bir araya getirerek bir model oluşturur. Her bir ağaç veri kümesinin alt kümeleriyle eğitilir ve ardından tahminlerini birleştirerek nihai tahmini üretmektedir (Breiman, 2001). RO algoritmasının matematiksel temsili denklem (2.3)'te verilmiştir.

$$f_i = \frac{\sum_{j:i \text{ özelliğine ayrılan } j \text{ dugumu}} n_{ij}}{\sum_{k \in \text{butun dugumler}} n_{ik}} \quad (2.3)$$

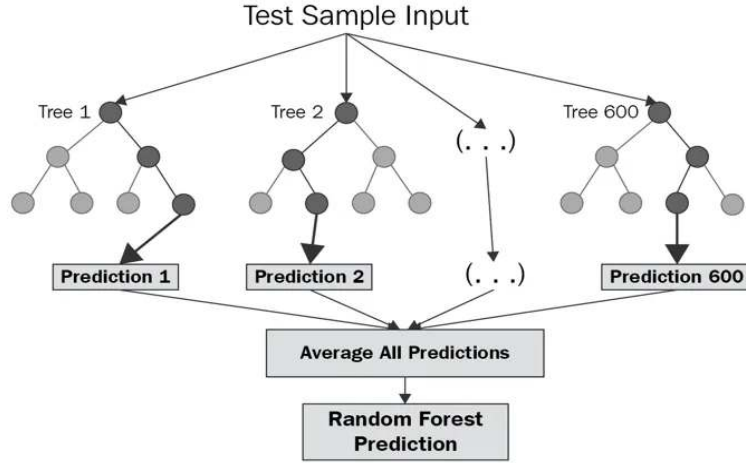
Özellik önem değerlerinin toplamına bölünerek 1 ve 0 arasında bir değere getirdikten sonra denklem (2.4) ile gösterilmiştir.

$$\text{norm}f_i = \frac{f_i}{\sum_{j \in \text{butun özellikler}} f_j} \quad (2.4)$$

Ağaçtaki özelliğin önem değerlerinin toplamı bulunduktan sonra toplam ağaç sayısına bölünür. Denklem (2.5) ile gösterilmiştir.

$$RFf_i = \frac{\sum_{j \in \text{butun ağaclar}} \text{norm}f_{ij}}{T} \quad (2.5)$$

Şekil 7 ile tanımlanan RO yüksek esneklik ve genel performansı nedeniyle çeşitli uygulamalarda başarılı sonuçlar vermiştir. Özellikle sınıflandırma ve regresyon problemlerinde etkili bir şekilde kullanılmaktadır. Sınıflandırma problemlerinde çoğunluk oylaması yöntemiyle sınıf tahminlerini birleştirerek sonuçları belirlemektedir. Regresyon problemlerinde ise ağaçların tahminlerinin ortalaması kullanılarak bir tahmin yapılmaktadır. Her bir ağaç veri setindeki özelliklerin önem sırasını hesaplamaktadır. Böylece veri setinde hangi özelliklerin tahmin performansı üzerinde daha büyük etkiye sahip olduğunu anlamak mümkün olmaktadır. RO algoritması aşırı uyum gibi problemlerin giderilmesini de sağlamaktadır.



Şekil 7. Rastgele Orman Algoritması

Kaynak: CFI Team, Random Forest, <https://corporatefinanceinstitute.com/resources/data-science/random-forest/>. E.T: 10.05.2023.

2.4.5. Destek Vektör Makinesi

Destek Vektör Makinesi (DVM), doğrusal ve doğrusal olmayan sınıflandırma problemlerini çözmek için kullanılan güçlü ve esnek bir makine öğrenmesi algoritmasıdır. DVM, veriyi en iyi şekilde ayırmak için bir hiper düzlem aramaktadır. Bu hiper düzlem iki sınıfı en geniş marjla ayırmak üzere tasarlanmıştır ve bu nedenle maksimum marjın sınıflandırıcısı olarak adlandırılmaktadır (Cortes ve Vapnik, 1995). DVM'nin matematiksel ifadesi denklem (2.6)'da verilmiştir.

$$\vec{x}_i \cdot \vec{w} + b = 1; \text{ (pozitif sınırlar)}$$

$$\vec{x}_i \cdot \vec{w} + b = -1; \text{ (negatif sınırlar)}$$

(2.6)

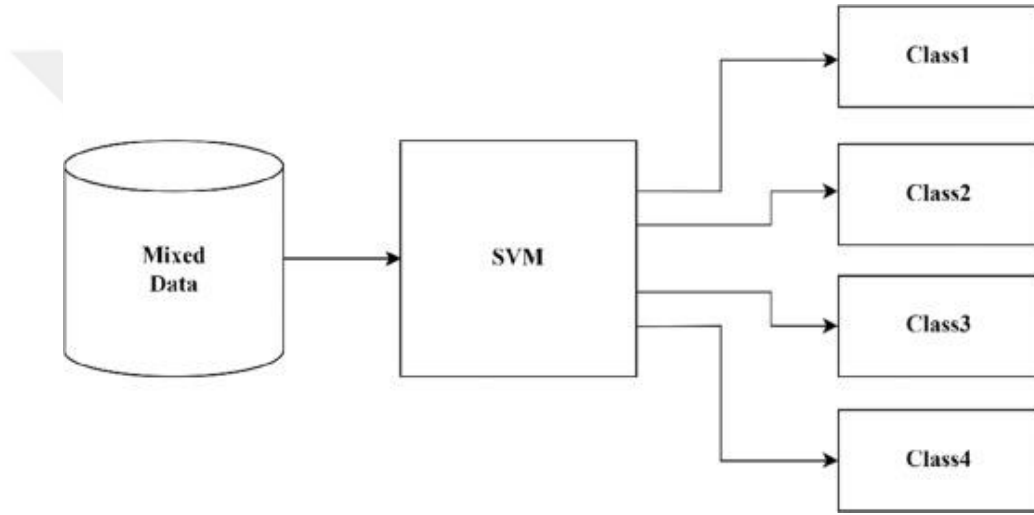
Sabit marjlı Destek Vektör Makineleri (DVM), destek vektörlerini kullanarak hiper düzlemi belirlerken oldukça kesin sonuçlar verir. Hiper düzlemin marjının en büyük hale getirilmesi hedeflendiğinde, sabit marjlı DVM bir optimizasyon sorunu haline gelir.

$$\min_{\vec{w}, b} \frac{1}{2} \|\vec{w}\|^2$$

(2.7)

$$y_i(\vec{x}_i \cdot \vec{w} + b) \geq 1; i = 1, \dots, n$$

Şekil 8 ile tanımlanan DVM'lerin temel amacı veri özelliklerinin doğrusal olarak ayrıştırılamayacağı durumlarda kullanılmak üzere tasarlanmış çekirdek fonksiyonları kullanmaktır. Bu çekirdekler daha yüksek boyutlu bir özellik uzayında doğrusal sınıflandırma yapmayı sağlamaktadır. DVM'lerin daha karmaşık sınıflandırma problemlerine uyum sağlamasını sağlamaktadır.



Şekil 8. Destek Vektör Makine Algoritması

Kaynak: Marneni, D., ve Vemula, S. (2023). Analysis of Covid-19 Using Machine learning Techniques. Statistical Modeling in Machine Learning, 37-53. <https://doi.org/10.1016/B978-0-323-91776-6.00006-3>

DVM genellikle yüksek doğruluk oranlarına sahiptirler ve aşırı uyuma karşı oldukça dirençlidirler. Ayrıca, doğrusal ve doğrusal olmayan sınıflandırma problemlerinde çalışabilme yetenekleri sahiptir ve çekirdek fonksiyonları sayesinde daha yüksek boyutlu veri setleri ile başa çıkabilme yetenekleri vardır. DVM'ler yüksek doğruluk oranları, aşırı uyuma karşı direnç ve çeşitli sınıflandırma problemlerini çözme yeteneği nedeniyle geniş çapta kullanılmaktadır.

ÜÇÜNCÜ BÖLÜM

3. DENEYSEL BULGULAR

3.1. Veri Setinin Elde Edilmesi

Bu tez kapsamında e-ticaret sitesinden elde edilen veri seti üzerinde BERT ve ALBERT dil modelleri ile Naive Bayes, Rastgele Orman ve Destek Vector Makinesi makine öğrenme algoritmaları kullanılarak ortaya çıkan performans kıyaslaması yapılmıştır. Kullanılan veri seti üzerinde birtakım işlemler yapılmıştır. Ortaya çıkan performans sonuçları kıyaslanmış ve özgün bir çalışma yapılmıştır.

3.2. Veri Setinin Elde Edilmesi

Veri bilgi ve içgörülerin bilgisayar sistemlerinde işlenmesi ve analiz edilmesi için uygun bir biçime dönüştürülmüş bilgilerdir (Kitchin, 2014). Teknolojik ilerlemelerin hızlanması beraberinde muazzam bir veri hacmini de getirmiştir. İnternet üzerinde gerçekleşen çeşitli etkinlikler ve sosyal medya kullanımı bu büyük veri yığınlarının oluşmasını sağlamaktadır. İnternet üzerindeki bu etkinlikler alışverişten bilgi araştırmaya ve sosyal etkileşimlere kadar çok çeşitli konuları kapsar. Bu süreçte oluşturulan her türlü dijital iz analiz edilebilir verilerin oluşumuna katkıda bulunmaktadır.

Bu büyük veri kümelerinin analizi işletmeler, hükümetler ve diğer kuruluşlar için önemli fırsatlar sunmaktadır. Bu verileri analiz etmek ve anlamlı içgörüler elde etmek daha iyi kararlar almayı, hedefli pazarlama stratejileri oluşturmayı ve daha etkili politikalar geliştirmeyi sağlamaktadır. Bu yüzden veri modern dünyanın en değerli kaynaklarından biri haline gelmiştir.

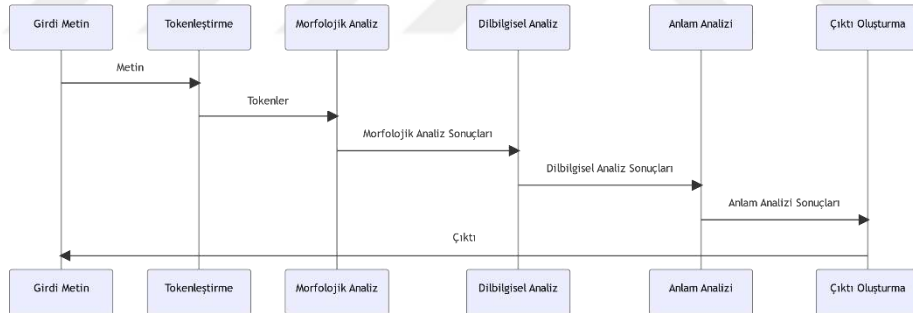
Veri setinin seçimi herhangi bir yapay zekâ alanında uygulanmasında kritik bir adımdır. Uygulamanın başarısı doğru verinin doğru bir şekilde toplanması ve işlenmesine büyük ölçüde bağlıdır. Veri seti sorunun doğru bir şekilde çözülmesi ve modelin etkin bir şekilde eğitilmesi için büyük önem taşımaktadır.

Bu tez çalışması kapsamında, BERT ve ALBERT dil modelleri ile makine öğrenmesi yöntemlerinde Naive Bayes, Rastgele Orman ve Destek Vektör Makinesi uygulamaları ile veri seti üzerinde elde edilen sonuçların kıyaslaması yapılmıştır.

Tez çalışmasında Hepsiburada e-ticaret sitesinde ürünler hakkında müşterilerin yaptıkları yorumları içeren bir veri seti ele alınmıştır. Bu veri seti Kaggle veri tabanında yer almaktadır. Veri setinde toplam 243477 adet pozitif ve negatif müşteri yorumları yer almaktadır. Pozitif yorumlar için 1 ve negatif yorum için 0 etiketlerine sahiptir. Veri seti dağınık bir yapıya sahip olmasından dolayı pozitif ve negatif etiketli yorumlar rastgele seçilmiş ve hızlı sonuç alabilmek için 27351 adet müşteri yorumu kullanılmıştır. Özelleştirilmiş bu veri setinde 13676 adet 0 etiketli negatif yorum ve 13675 adet 1 etiketli pozitif yorum yer almaktadır.

3.2.1. Zemberek Kütüphanesi

Zemberek Türkçe dil işleme ve doğal dil işleme görevleri için kullanılan açık kaynaklı bir Java kütüphanesidir. Adını bir tür Türk müzik aleti olan ve birçok farklı tonu aynı anda çalabilen Zemberek isminden almaktadır. Bu isim kütüphanenin Türkçe'nin zengin ve karmaşık yapısını analiz etme ve işleme yeteneğini yansıtmaktadır.



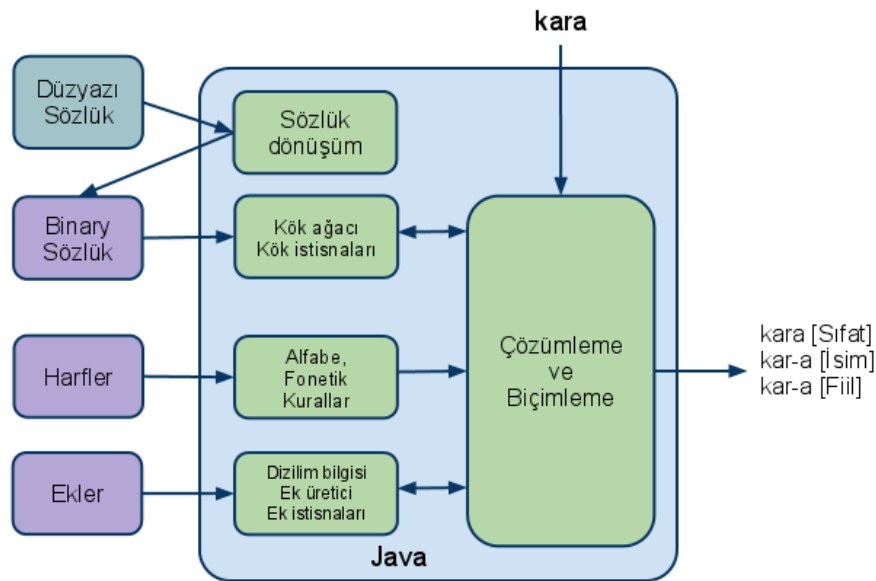
Şekil 9. Zemberek Kütüphanesi Çalışma Yapısı

Şekil 9 üzerinde yer alana tokenizasyon girdi metni kelime ve cümlelere ayırma işlemini ifade etmektedir. Tokenizasyon işlemi sonrası kendi içerisinde ayrılan kelimelerin köklerini ve eklerini morfolojik analiz yöntemi ile analiz etmektedir. Dilbilgisel analiz ile metin içerisinde yer alan her kelimenin dilbilgisel kategorisi belirlemektedir. Anlam analizi kelimelerin kişi isimleri, organizasyonlar, yerler vb. kategorilere ayrılmasını sağlamaktadır.

Zemberek kütüphanesi bir dizi farklı bileşene sahip geniş kapsamlı bir araçtır. Bu bileşenler arasında dilbilgisel ve morfolojik analiz, kelime önerme, kelime kökü bulma, yazım denetimi, adlandırılmış varlık tanıma, dil tespiti ve benzeri yetenekler bulunmaktadır (Akın ve Akın, 2007).

Zemberek Türkçe'nin özelliklerini ve karmaşıklığını anlama konusunda önemli bir rol oynamaktadır. Türkçe'nin ağızları, dile özgü sesbilim ve dilbilgisel özellikleri gibi bir dizi zorluğa rağmen Zemberek bu zorlukların üstesinden gelmeye çalışmıştır. Zemberek kütüphanesi Türkçe'nin dilbilgisel özelliklerini, sesbilimsel özelliklerini ve ağızlarını modellemeye çalışmaktadır.

Zemberek Türkçe dil işleme ve DDİ projelerinde genellikle önemli bir rol oynamıştır. Türkçe metinleri analiz etmek, Türkçe konuşmayı anlamak veya Türkçe metinleri oluşturmak gibi görevlerde kullanılmaktadır. Özellikle metin sınıflandırma, duygu analizi, makine çevirisi ve otomatik özetleme gibi alanlarda kullanılmaktadır. Zemberek kütüphanesi birçok farklı projede kullanılmıştır ve hala aktif olarak geliştirilmektedir. Bu projeler dil teknolojisi ve yapay zekâ araştırmalarından endüstriyel uygulamalara kadar geniş bir yelpazeyi kapsamaktadır. Bu nedenle Zemberek kütüphanesi Türkçe dil işleme ve DDİ çalışmalarında değerli bir araç olarak kabul edilmektedir. Örnek bir Türkçe metnin zemberek kütüphanesi üzerinde işlenmesi **Şekil 10**'da verilmiştir.



Şekil 10. Zemberek Kütüphanesi ile Türkçe Metnin İşlenmesi

Kaynakça: <http://zembereknlp.blogspot.com/>. E.T: 02.05.2023

İlerleyen bölümlerde BERT ve ALBERT dil modelleri ve makine öğrenmesi algoritmaları olan Naive Bayes, Rastgele Orman ve Destek Vektör Makinesi yöntemleri uygulanmış ve sonuçları verilmiştir.

Kaggle üzerinden elde edilen veri seti üzerinde zemberek kütüphanesi ile işlemler yapılmıştır. Zemberek kütüphanesi veri seti üzerinde yazım yanlışların düzeltilmesi, kelimelerin köklerinin bulunması, kelimelerin Türkçe olup olmadığının kontrolü ve hatalı kodlama temizleme gibi yeteneklere sahiptir. Zemberek kütüphanesi ile veri setimizde yer alan müşteri yorumları üzerinde yazım denetimi, kelime kökü bulma, heceleme ve tokenleştirme gibi bir dizi işlemler uygulanmıştır. Zemberek kütüphanesi uygulanan veri setimizin ilk 28 satırı **Şekil 11**'da yer almaktadır.

1	labels,text
2	1,3 yıl kullan ve memnun
3	1,3 yıl kullan müthiş
4	1,ürün bugün el geç çok fazla incele fırsat ol ancak göz çarp ilk şey usb bağlantı için pil kapak alt çok güzel yer yap ol daha önce o aparat kaybet biri ol bu özellik çok sev ayrıca ol çok i
5	1,al karar ver hemencecik gel keyif kullan
6	1,günlük kullanım çok çok iyi kısaç mükemmel
7	1,gayet güzel
8	1,çok kalite bir ürün ve fiyat da uygun uzun araştır sonuç al ve çok memnun
9	1,yaklaşık 5 sene kullan defalarca düş rağmen de 5 yıl daha git diye düşün
10	1,ürün günlük kullanım için çok uygun ürün birlikte 1 adet pil de çık iş ve günlük kullanım için kablo mouse iste düşün al hafta 5 gün yaklaşık 8 saat kullanım 2 ay civar pil ömür var
11	1,ürün bir arkadaş 2009 yıl hediye ol al halen kullan ki çok dikkat kullan bir kullan da değil gayet sağlam dayanıklı ve aynı zaman da ergonomik bir ürün
12	1,1 gün gibi kısa bir süre el geç ve bu süper bir ürün tavsiye et lakin el büyük ol kişi daha büyük model tercih et ürün teşekkür
13	1,fiyat fayda oran göre çok iyi bir ürün
14	1,güzel kullan bir ürün herhangi bir problem yok
15	1,oyun oyna için gayet güzel bir mouse
16	1,ürün yıl kullan fiyat kalite ve performans çok iyi
17	1,hoş ve kalite bir mouse kullanım için mükemmel boyut um uzun süre boz
18	1,ur
19	1,ürün yorum bak al iyi ki de al uygun fiyat ve yüksek kalite bir ürün pişman ol tavsiye et
20	1,ürün süper kalite değil malzeme ağı ama fiyat belli iş gör birçok yüzey rahat kullan
21	1,sağlam ve dayanıklı bir ürün herkes tavsiye et ilk ürün 2013 yıl al hala kullan memnun kal için yeni sipariş ver .
22	0,fiyat göre ortalama oyun için al dene standart kullanım için yeterli ama ben orta teker boz o dışı iyi .
23	1,çok iyi
24	1,ürün gayet güzel kısa süre el geç teşekkür
25	1,ürün al önce yorum oku çok kısa süre gel yaz fiyat da uygun ol sipariş ver ürün gayet şık ve kullan özellikle el küçük ol için gayet pratik tek eksi yön dört gün gel düşün tavsiye et
26	1,ürün bir gün el geç hepsiburada teşekkür et çıkar sonra hemen kullan başla çok memnun tavsiye et
27	1,çok güzel bir mouse hem el otur hem çok kibar
28	1,birçok mouse kullan an logitech bu konu ben en iyi fiyat da süper ben günlük kullanım için al en güzel mouse şarj da çok git

Şekil 11. Veri Setinin İlk 28 Verisi

3.3. Deneysel Çalışma Ortamı

Python kodları Google Colaboratory gibi web tabanlı platformlar üzerinde çalıştırılabilmektedir. Google Colaboratory geliştiricilerin Python kodlarını bir tarayıcı üzerinde çalıştırmasına olanak sağlar ve ayrıca kod bloklarının hızlı şekilde yürütülmesine, çeşitli eklentilerin ve kütüphanelerin kolaylıkla yüklenmesine ve

paylaşılmasına imkân vermektedir. Google Colaboratory Pro paketi içinde yer alan güçlü ekran kartı ve yüksek ram kaynaklarına erişim sağlanabilmektedir. Bu kaynaklar sayesinde daha hızlı sonuçlar elde edilebilmektedir.

Bu çalışmada duygu analizinin gerçekleştirilmesi için Google Colaboratory Pro platformu ve Python programlama dili kullanılmıştır. Colaboratory Pro, duygu analizi için gerekli kütüphanelerin kolayca yüklenmesini, kod bloklarının hızlıca çalıştırılmasını, uygulamanın kolayca paylaşılmasını ve platform bağımsız her yerden erişim imkanını sağlamıştır.

3.4. Analiz Yöntemleri

Bu tez çalışmasında sonuçların başarısını ölçmek için doğruluk (accuracy), duyarlılık (recall), kesinlik (precision) ve F1-Skor (F1-score) gibi performans metrikleri kullanılmıştır. Terimlere ait açıklamalar aşağıda yer almaktadır.

Doğruluk: Bir sınıflandırma modelinin doğru şekilde sınıflandırdığı örneklerin oranını ifade eden bir performans metriğidir. Tüm örneklerin ne kadarının doğru olarak sınıflandırıldığını gösterir. Doğru pozitifler, doğru negatifler, yanlış pozitifler ve yanlış negatiflerin hepsini dikkate almaktadır.

Duyarlılık: Bir sınıflandırma modelinin gerçek pozitifleri ne kadar başarılı bir şekilde tespit ettiğini ölçen bir performans metriğidir. Tüm gerçek pozitiflerin ne kadarının doğru şekilde sınıflandırıldığını ifade eder. Değeri ne kadar yüksekse modelin gerçek pozitifleri tespit etme yeteneği o kadar iyi olacaktır.

Kesinlik: Bir sınıflandırma modelinin pozitif olarak varsaydığı örneklerin ne kadarının gerçekten pozitif olduğunu ölçen bir performans metriğidir. Tüm pozitif tahminlerin ne kadarının doğru olduğunu ifade eder. Precision değeri ne kadar yüksekse, modelin yanlış pozitif tahmin yapma oranı o kadar düşük olacaktır.

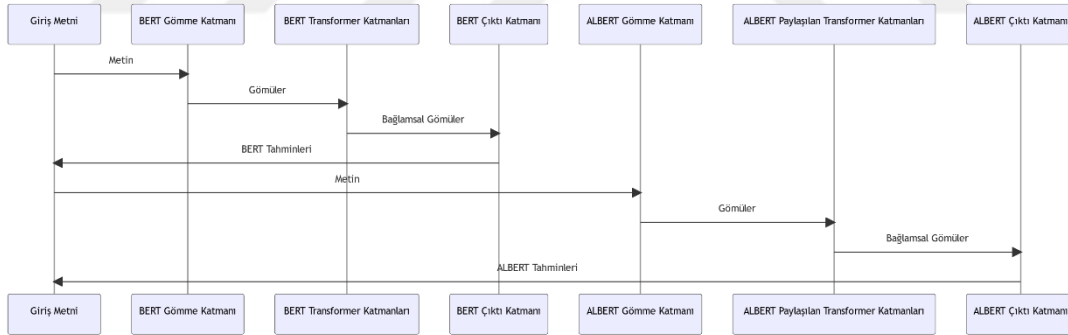
F1-Skoru: Bir sınıflandırma modelinin kesinlik ve duyarlılık metriklerini birleştiren bir performans metriğidir. F1-skor kesinlik ve duyarlılık metriklerinin harmonik ortalamasını alarak hesaplanır. F1-skor bir sınıflandırma modelinin hem doğru pozitifleri ne kadar iyi tespit ettiğini hem de pozitif tahminlerinin ne kadar doğru olduğunu bir arada değerlendirir. F1-skor bir sınıflandırma modelinin dengelemiş bir performansa sahip olup olmadığını göstermektedir.

3.5. Veri Seti Üzerinde Python Programlama Dili ile Duygu Analizi

Duygu analizi veri setinde yer alan yorumların olumlu ya da olumsuz olmadığını belirler ve buna göre sonuç üretir. Tez çalışmasında veri seti üzerinde BERT ve ALBERT dil modelleri ile Naive Bayes, Rastgele Orman ve Destek Vektör Makinesi gibi makine öğrenmesi yöntemleri uygulanmıştır ve sonuçları verilmiştir. BERT ve ALBERT dil modeli Python programlama dilini kullanarak geliştirilmiştir. Python, Guido Van Rossom tarafından oluşturulan bir nesne yönelimli ve yüksek seviye programlama dilidir. Kullanıcı dostu sözdizimi ve çok sayıda işletim sistemi ile uyumluluğu sayesinde Python günümüzde birçok kişi ve kurum tarafından tercih edilen bir dil haline gelmiştir (Doğan, 2021).

3.4.1. BERT ve ALBERT Dil Modelleri ile Duygu Analizi

Bu tezde Türkçe duygu analizi için BERTurk ve Albert-Tr dil modeli uygulanmıştır. Bu dil modelleri ifadeleri olumlu ve olumsuz duygusal anlamlara göre sınıflandırmaktadır.



Şekil 12. BERT ve ALBERT Dil Modelinin Çalışma Yapısı

Araştırmanın ilk aşamasında gerekli kütüphaneleri ve veri setinin yüklenmesi sağlanmıştır. Daha sonra metinler tokenler yani birimlerine ayrılıp modele sunulmuştur. Bu aşamada her bir cümlenin olumlu ya da olumsuz durumları ve onları belirleyen etiketler tanımlanmıştır. Bu süreç modelin duygusal eğilimleri öğrenmesini ve yeni ifadeleri doğru bir şekilde sınıflandırabilmesini sağlamaktadır.

BERTurk ve ALBERT-Tr dil modellerinin çağırılması işlemlerine ait kodlar Şekil 13 ve Şekil 14’te verilmiştir.

```
✓ 19 sn. 1 from simpletransformers.classification import ClassificationModel
2
3 args = {
4     'reprocess_input_data': True,
5     'overwrite_output_dir': True
6 }
7
8 model = ClassificationModel('bert', 'dbmdz/bert-base-turkish-cased',
```

Şekil 13. BERTurk Dil Modelinin Çağırılması

```
✓ 21 sn. [9] 1 from simpletransformers.classification import ClassificationModel
2
3 args = {
4     'reprocess_input_data': True,
5     'overwrite_output_dir': True
6 }
7
8 model = ClassificationModel('albert', 'lodos/albert-base-turkish-uncased',
```

Şekil 14. ALBERT-Tr Dil Modelinin Çağırılması

Aşağıda tanımlanan kod parçacıklarında ana amaç modellerin belirli bir eğitim veri seti üzerinde eğitmek ve modelin eğitim sürecinde genelleme yeteneğini izlemektir. Modelin eğitim süreci belirtilen eğitim ve doğrulama veri setleri üzerinde gerçekleşir. Eğitim tamamlandığında model metin sınıflandırma görevini gerçekleştirmek üzere kullanılabilir hale gelmektedir. Döngü sayısı 3 olarak belirlenmiştir. Test veri setinin BERTurk modeli ile eğitilmesi toplam 17 dakika, Albert-Tr modeli ile eğitilmesi toplam 9 dakika sürmüştür. Bu işlemlere ait kodlar Şekil 15 ve Şekil 16’da yer verilmiştir.

```
[12] 1 print('Training on {:,} samples...'.format(len(train_df)))
      2
      3 # Train the model, testing against the validation set periodically.
      4 out = model.train_model(train_df, eval_df=validation_df)
```

Training on 21,880 samples...
0% | 44/21880 [00:09<1:04:30, 5.64it/s]

Epoch 3 of 3: 100% | 3/3 [17:13<00:00, 345.66s/it]

Epochs 0/3. Running Loss: 0.0115: 100% | 2735/2735 [05:29<00:00, 9.62it/s]

/usr/local/lib/python3.10/dist-packages/torch/optim/lr_scheduler.py:139: UserWarning: Detected call of `lr\_scheduler.step()` before `optimizer.step()`. "

Epochs 1/3. Running Loss: 0.0072: 100% | 2735/2735 [05:17<00:00, 9.47it/s]

Epochs 2/3. Running Loss: 0.0107: 100% | 2735/2735 [05:33<00:00, 9.63it/s]

Şekil 15. BERTurk Dil Modeli ile Veri Setinin Eğitilmesi

```
[17] 1 print('Training on {:,} samples...'.format(len(train_df)))
      2
      3 # Train the model, testing against the validation set periodically.
      4 out = model.train_model(train_df, eval_df=validation_df)
```

Training on 21,880 samples...
0% | 44/21880 [00:07<44:52, 8.11it/s]

Epoch 3 of 3: 100% | 3/3 [09:42<00:00, 199.45s/it]

Epochs 0/3. Running Loss: 0.0140: 100% | 2735/2735 [02:47<00:00, 17.60it/s]

Epochs 1/3. Running Loss: 0.4644: 100% | 2735/2735 [03:02<00:00, 17.66it/s]

Epochs 2/3. Running Loss: 0.0064: 100% | 2735/2735 [03:20<00:00, 17.01it/s]

Şekil 16. ALBERT-Tr Dil Modeli ile Veri Setinin Eğitilmesi

BERTurk ve ALBERT-Tr modelleri ile eğitimi tamamlanan ve test verisine uygulanan modellerim duygu analiz sonuçları aşağıda yer almaktadır. BERTurk modeli ile doğruluk değeri %94 çıkmıştır. ALBERT-Tr modeli ile doğruluk değeri %92 çıkmıştır. Her iki modelin doğruluk (accuracy) değeri haricinde duyarlılık (recall), kesinlik (precision) ve F1-skor (f1-score) metriklerin sonuçları da verilmiştir. BERTurk ve ALBERT-Tr modellerinin test sonuçlarının bütün hali **Şekil 17** ve **Şekil 18**'de yer almaktadır.

```
[16] 1 print(classification_report(test_df.labels,np.argmax(test_outputs,axis=1)))
```

	precision	recall	f1-score	support
0	0.94	0.93	0.94	1379
1	0.93	0.94	0.94	1357
accuracy			0.94	2736
macro avg	0.94	0.94	0.94	2736
weighted avg	0.94	0.94	0.94	2736

Şekil 17. BERTurk Dil Modeli Analiz Sonuçları

```
[18] 1 print(classification_report(test_df.labels,np.argmax(test_outputs,axis=1)))
```

	precision	recall	f1-score	support
0	0.93	0.91	0.92	1379
1	0.91	0.93	0.92	1357
accuracy			0.92	2736
macro avg	0.92	0.92	0.92	2736
weighted avg	0.92	0.92	0.92	2736

Şekil 18. ALBERT-Tr Dil Modeli Analiz Sonuçları

3.4.2. Makine Öğrenmesi Yöntemleri ile Duygu Analizi

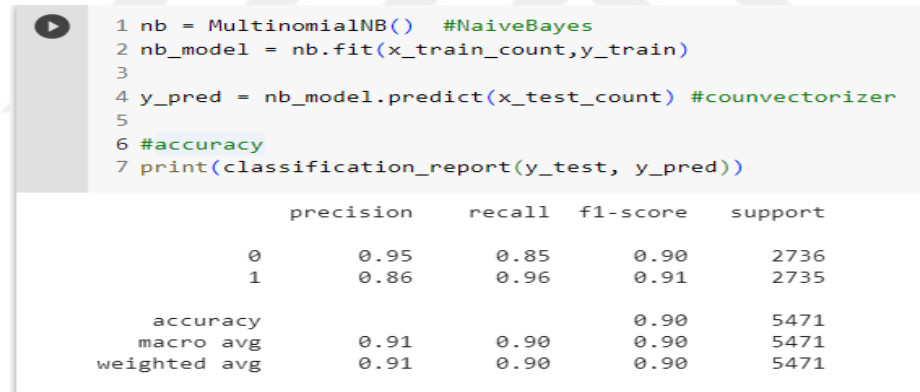
Bu tezde dil modellerine kıyasla makine öğrenme yöntemlerinin aynı veri setinde nasıl bir sonuç alınacağı izlenmiştir. Popüler makine öğrenme yöntemlerinden Naive Bayes, Rastgele Orman ve Destek Vektör Makinesi yöntemleri kullanılmıştır. Bu yöntemler ile olumlu ve olumsuz duygusal anlamlara göre sınıflandırılmaktadır. Çalışmanın ilk aşamasında gerekli kütüphanelerin ve daha önceden zemberek kütüphanesi ile ön işleme sokulmuş veri setinin yüklenmesi gerçekleşmiştir. Metinler birimlerine ayrılıp tez çalışmasında kullanılan makine öğrenmesi yöntemlerine sunulmuştur. Bu aşamada her bir cümlemin olumlu ya da olumsuz durumları ve onları belirleyen etiketler tanımlanmıştır. Bu süreç modelin duygusal eğilimleri öğrenmesini ve yeni ifadeleri doğru bir şekilde sınıflandırabilmesini sağlamaktadır.

Bu tez çalışmasında metin verilerinin sayısal formata dönüştürülmesi için CountVectorizer ve TF-IDFVectorizer teknikleri uygulanmıştır. Bu iki yönteminin detaylı tanımlamaları aşağıda sunulmuştur.

TF-IDFVectorizer: Bir kelimenin belge içerisindeki önemini istatistiksel bir teknikle belirleyen bir ağırlık faktörüdür. Belgenin içinde sıkça geçen terimler diğer terimlere kıyasla belirgin bir şekilde öne çıkarlar. Bu durum bazen diğer önemli terimlerin göz ardı edilmesine yol açtığından bu tür durumları dengelemek için geliştirilmiştir.

CountVectorizer: Bir dizi metin belgesini belirteç (token) frekanslarına dayanan bir matrise dönüştürme işlemini gerçekleştirir. Bu vektörizasyon tekniği, metin belgelerinin ön işlemlerini de içererek metin analitiği yeteneklerini genişletir. Özellikle metin sınıflandırma ve analiz uygulamalarında etkili bir araç olarak kabul edilmektedir.

Analiz sürecinde elde edilen sonuçların değerlendirilmesi için doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F1-skoru (F1-score) metrikleri kullanılmıştır. Naive Bayes, Rastgele Orman ve Destek Vektör Makinesi yöntemlerinin veri seti ile eğitildikten sonra test aşamasındaki sonuçların çağrılması işlemlerine ait kodlar **Şekil 19**, **Şekil 20** ve **Şekil 21**'de verilmiştir.



```
1 nb = MultinomialNB() #NaiveBayes
2 nb_model = nb.fit(x_train_count,y_train)
3
4 y_pred = nb_model.predict(x_test_count) #counvectorizer
5
6 #accuracy
7 print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.95	0.85	0.90	2736
1	0.86	0.96	0.91	2735
accuracy			0.90	5471
macro avg	0.91	0.90	0.90	5471
weighted avg	0.91	0.90	0.90	5471

Şekil 19. Naive Bayes Analiz Sonuçları

```
[117] 1 rf = RandomForestClassifier() #RandomForest-countvectorizer
2 rf_model = rf.fit(x_train_count,y_train)
3
4 y_pred = rf_model.predict(x_test_count)
5
6 #accuracy
7 print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.92	0.85	0.88	2736
1	0.86	0.93	0.89	2735
accuracy			0.89	5471
macro avg	0.89	0.89	0.89	5471
weighted avg	0.89	0.89	0.89	5471

Şekil 20. Rastgele Ağaç Analiz Sonuçları

```
[119] 1 svm = LinearSVC() #Support Vector Machine
2 svm_model = svm.fit(x_train_count,y_train)
3
4 y_pred = svm_model.predict(x_test_count) #countvectorizer
5
6 #accuracy
7 print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.91	0.88	0.90	2736
1	0.89	0.91	0.90	2735
accuracy			0.90	5471
macro avg	0.90	0.90	0.90	5471
weighted avg	0.90	0.90	0.90	5471

Şekil 21. Destek Vektör Makinesi Analiz Sonuçları

Elde edilen sonuçlara göre Naive Bayes ve Destek Vektör Makinesi algoritması %90 oranında bir başarı göstermiştir. Rastgele Orman algoritması %89 oranında bir başarı elde etmiştir.

SONUÇ VE ÖNERİLER

Günümüz teknolojisi büyük veri ve yapay zekâ alanlarının hızla geliştiği bir çağda, duygu analizi ve metin sınıflandırma konusunda önemli adımlar atılmaktadır. Bu tez çalışmasında e-ticaret sitelerinde yer alan yorumların BERT ve ALBERT dil modelleri ile analizi üzerine yoğunlaşmıştır. BERT ve ALBERT dil modelleri ile makine öğrenmesi yöntemleri olan Naive Bayes, Rastgele Orman ve Destek Vektör Makinesi ile karşılaştırılmıştır.

Kullanılan veri setinin BERTürk dil modeli ile eğitilmesi sonucu 17 dakikalık süre zarfında %94 oranında doğruluk performansı göstermiştir. Veri seti ALBERT-Tr ile eğitildiği takdirde 9 dakikalık süre zarfında %92 oranında doğruluk performansı gösterilmiştir. Makine öğrenme yöntemlerinden Naive Bayes Algoritması ve Destek Vektör Makinesi algoritmaları %90 oranında, Rastgele Orman algoritması ise %89 oranında bir doğruluk performansı göstermiştir.

Bu analiz sonuçları Naive Bayes ve Destek Vektör Makinesi algoritmalarının uygulandığı veri seti üzerinde etkili bir doğruluk performansı gösterdi gözlenmektedir. Doğruluk başarı oranının %90 olması bu algoritmaların metin sınıflandırma ve duygu analizi gibi konularda güçlü bir potansiyele sahip olduğunu göstermiştir. Bu sonuçlara ek olarak Rastgele Orman algoritması ile %89 oranında kabul edilebilir bir sonuç elde edilmektedir. Bu sonuçlar kıyaslandığında Rastgele Orman algoritması %1 oranında farkla Naive Bayes ve Destek Vektör Makinesi algoritmalarının gerisinde kalmıştır.

BERT dil modelinin, ALBERT dil modeline göre daha yüksek bir performans gösterdiğini ortaya koymaktadır. BERT dil modelinin daha fazla parametre içermesi ve bu parametrelerin verilerin özelliklerini daha geniş bir ölçekte yakalayabilmesi ile ilişkilendirilebilmektedir. Bu durum BERT dil modelinin dil anlamını ve bağlamı daha iyi bir şekilde anlamasını ve böylece daha doğru sonuçlar vermesini sağlamaktadır. Parametre sayılarına bağlı olarak artan doğruluk performansı hesaplama süresinde ve kaynak tüketiminde önemli bir artışa yol açmaktadır. BERT dil modelini kullanmanın faydaları ve maliyetleri arasında bir denge bulunması gerekmektedir.

ALBERT modeli BERT modelinin daha az parametre içeren bir versiyonu olarak geliştirilmiş olmasına rağmen analiz sonuçları ALBERT'in doğruluk performansının BERT'in gerisinde olduğunu göstermektedir. BERT dil modelinin aksine ALBERT dil modeli parametre sayısını azaltarak ve model boyutunu küçülterek hesaplama maliyetlerini azaltmayı hedeflemiştir fakat bu yaklaşım modelin karmaşıklığını ve dolayısıyla dil anlama yeteneğini azaltmıştır.

Sonuç olarak Naive Bayes, Rastgele Orman ve Destek Vektör Makinesi gibi geleneksel makine öğrenme algoritmaları ile karşılaştırıldığında BERT ve ALBERT dil modellerinin daha etkili performans gösterdiği gözlemlenmiştir. Bu tez çalışması kapsamında elde edilen sonuçların dil modelleri ve makine öğrenmesi algoritmaları tabanlı duygu analizi tekniklerinin önemini vurgulaması hedeflenmiş olup gelecekte yapılacak çalışmalara katkı sağlayacağı öngörülmektedir.

KAYNAKÇA

- Akın, A. A., ve Akın, M. D. (2007). Zemberek, an open source NLP framework for Turkish languages. *Structure*, 10, 1-5.
- Al Amrani, Y., Lazaar, M., ve El Kadiri, K. E. (2018). Random forest and support vector machine based hybrid approach to sentiment analysis. *Procedia Computer Science*, 127, 511-520.
<https://doi.org/https://doi.org/10.1016/j.procs.2018.01.150>.
- Alrowili, S., ve Shanker, V. (2021). *Biom-transformers: building large biomedical language models with BERT, ALBERT and ELECTRA*.
<https://doi.org/10.18653/v1/2021.bionlp-1.24>.
- Bourequat, W., ve Mourad, H. (2021). Sentiment analysis approach for analyzing iphone release using support vector machine. *international journal of advances in data and information systems*, 2(1), 36-44.
<https://doi.org/10.25008/ijadis.v2i1.1216>.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
<https://doi.org/10.1023/A:1010933404324>.
- Chen, B., Chen, B., Gao, D., Chen, Q., Huo, C., Meng, X., Ren, W., ve Zhou, Y. (2021). Transformer-based language model fine-tuning methods for COVID-19 fake news detection. T. Chakraborty, K. Shu, H. R. Bernard, H. Liu, ve M. S. Akhtar (Ed.), *Combating Online Hostile Posts in Regional Languages during Emergency Situation* (ss. 83-92). Springer International Publishing.
- Choi, H., Kim, J., Joe, S., ve Gwon, Y. (2021a). Evaluation of BERT and ALBERT sentence embedding performance on downstream NLP tasks. *2020 25th International Conference on Pattern Recognition (ICPR)*, 5482-5487.
<https://doi.org/10.1109/ICPR48806.2021.9412102>.
- Choi, H., Kim, J., Joe, S., ve Gwon, Y. (2021b). Evaluation of BERT and ALBERT sentence embedding performance on downstream NLP Tasks. *2020 25th International Conference on Pattern Recognition (ICPR)*, 5482-5487.
<https://doi.org/10.1109/ICPR48806.2021.9412102>.
- Cortes, C., ve Vapnik, V. N. (1995). Support-vector networks. *Machine Learning*, 20, 273-297.

- Demir E. (2021). *Türkçe ses kayıtlarındaki konuşmalardan elde edilen verilerden duygu tespiti ve analizi*. (Yüksek Lisans Tezi). Yalova Üniversitesi Lisansüstü Eğitim Enstitüsü.
- Devlin, J., Chang, M.-W., Lee, K., ve Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. J. Burstein, C. Doran, ve T. Solorio (Ed.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)* (ss. 4171-4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/n19-1423>.
- Doğan, Ö. (2021, Aralık 26). *Python nedir? En popüler programlama dili*. <https://teknoloji.org/python-nedir/>. Erişim Tarihi : 10.04.2023
- Goldberg, Y. (2017). Neural network methods for natural language processing. *synthesis lectures on human language technologies*, 10, 1-309. <https://doi.org/10.2200/S00762ED1V01Y201703HLT037>.
- Gupta, A., Tyagi, P., Choudhury, T., ve Shamoon, M. (2019). Sentiment analysis using support vector machine. *2019 International Conference on contemporary Computing and Informatics (IC3I)*, 49-53. <https://doi.org/10.1109/IC3I46837.2019.9055645>.
- Gurcan, F. (2018). *Multi-Class classification of turkish texts with machine learning algorithms*. <https://doi.org/10.1109/ISMSIT.2018.8567307>.
- Güven, Z. A. (2021). The effect of BERT, ELECTRA and ALBERT language models on sentiment analysis for Turkish product reviews. *2021 6th International Conference on Computer Science and Engineering (UBMK)*, 629-632. <https://doi.org/10.1109/UBMK52708.2021.9559007>.
- Hoş, S. (2021). *Makine öğrenimi (machine learning) nedir? Derin öğrenme ile arasındaki farklar*. <https://www.hosting.com.tr/blog/makine-ogrenimi-machine-learning/>.
- Kaya, U., Yılmaz, A., ve Dikmen, Y. (2019). Sağlık alanında kullanılan derin öğrenme yöntemleri. *European Journal of Science and Technology*, 792-808. <https://doi.org/10.31590/ejosat.573248>.

- Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1), 205395171452848. <https://doi.org/10.1177/2053951714528481>.
- Koto, F., Lau, J. H., ve Baldwin, T. (2021). Discourse probing of pretrained language models. *North American Chapter of the Association for Computational Linguistics..*
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., ve Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. <https://openreview.net/forum?id=H1eA7AEtvS>.
- LeCun, Y., Bengio, Y., ve Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444. <https://doi.org/10.1038/nature14539>.
- Lee, H., Yoon, J., Hwang, B., Joe, S., Min, S., ve Gwon, Y. (2021). Kore ALBERT: pretraining a lite bert model for Korean language understanding. *2020 25th International Conference on Pattern Recognition (ICPR)*, 5551-5557. <https://doi.org/10.1109/ICPR48806.2021.9412023>.
- Liang, R., ve Wang, J. (2019). A linguistic intuitionistic cloud decision support model with sentiment analysis for product selection in e-commerce. *International Journal of Fuzzy Systems*, 21(3), 963-977. <https://doi.org/10.1007/s40815-019-00606-0>.
- Oflazer, K. (1993). *Two-level Description of Turkish Morphology*. http://repository.bilkent.edu.tr/bitstream/handle/11693/25941/Two_level_description_of_Turkish_Morphology.pdf;jsessionid=AF414D66B52E0E598D58A833D8C38BFB?sequence=1.
- Probst, P., ve Boulesteix, A.-L. (2017). To tune or not to tune the number of trees in random forest? *Journal of Machine Learning Research*, 18.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., ve Liu, P. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21.
- Russell, S., ve Norvig, P. (2020). What is AI? *Artificial Intelligence: A Modern Approach*, 4th Edition (4th Edition, ss. 1-2). http://lib.ysu.am/disciplines_bk/b7707dde83ee24b2b23999b4df5fd988.pdf.

- Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3), 210-229. <https://doi.org/10.1147/rd.33.0210>.
- Somvanshi, M., ve Chavan, P. (2016). A review of machine learning techniques using decision tree and support vector machine. *2016 International Conference on Computing Communication Control and automation (ICCUBE)*, 1-7.
- Tunalı, V., ve Bilgin, T. (2012). Türkçe metinlerin kümelenmesinde farklı kök bulma yöntemlerinin etkisinin araştırılması. *Elektrik, Elektronik ve Bilgisayar Mühendisliği Sempozyumu (ELECO 2012)*.
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, LIX(236), 433-460. <https://doi.org/10.1093/mind/LIX.236.433>.
- Uçar, K. T. (2020). *BERT modeli ile Türkçe metinlerde sınıflandırma yapmak*. <https://medium.com/@toprakucar/bert-modeli-ile-t%C3%BCrk%C3%A7e-metinlerde-s%C4%B1n%C4%B1fland%C4%B1rma-yapmak-260f15a65611>.
- Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., ve Polosukhin, I. (2017). Attention is all you need. *NIPS*.
- Yang, L., Li, Y., Wang, J., ve Sherratt, R. (2020). Sentiment analysis for e-commerce product reviews in chinese based on sentiment lexicon and deep learning. *IEEE Access*, 8, 1. <https://doi.org/10.1109/ACCESS.2020.2969854>.
- Young, T., Hazarika, D., Poria, S., ve Cambria, E. (2018). Recent trends in deep learning based natural language processing [Review article]. *IEEE Computational Intelligence Magazine*, 13, 55-75. <https://doi.org/10.1109/MCI.2018.2840738>.
- Yu, L., Gan, S., Chen, Y., ve Luo, D. (2021). A novel hybrid approach: instance weighted hidden naive bayes. *Mathematics*, 9(22). <https://doi.org/10.3390/math9222982>
- Yurt, E. A. (2015). *Türkçe metinlerde duygu analizi*. (Yüksek Lisans Tezi). Maltepe Üniversitesi Fen Bilimleri Enstitüsü.
- Zeng, D., Dai, Y., Li, F., Wang, J., ve Kumar, A. (2019). Aspect based sentiment analysis by a linguistically regularized cnn with gated mechanism. *Journal of Intelligent & Fuzzy Systems*, 36, 1-10. <https://doi.org/10.3233/JIFS-169958>.
- Zhang, H. (2004). The optimality of naive bayes. *The Florida AI Research Society*.

- Zhang, L., Wang, S., ve Liu, B. (2018). Deep learning for sentiment analysis: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(4), e1253. <https://doi.org/https://doi.org/10.1002/widm.1253>.
- Zhao, Z., Liu, Y., Chen, L., Liu, Q., Ma, R., ve Yu, K. (2020). An investigation on different underlying quantization schemes for pre-trained language models. Zhu, M. Zhang, Y. Hong, ve R. He (Ed.), *Natural Language Processing and Chinese Computing* (ss. 359-371). Springer International Publishing.
- Zheng, W., ve Ye, Q. (2009). Sentiment classification of chinese traveler reviews by support vector machine algorithm. *2009 Third International Symposium on Intelligent Information Technology Application*, 3, 335-338. <https://doi.org/10.1109/IITA.2009.457>.

