T.C.

HASAN KALYONCU ÜNİVERSİTESİ

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

ELEKTRONİK BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

KLASİK MAKİNE ÖĞRENME ALGORİTMALARI VE TRANSFORMER MODELİ İLE TÜRKÇE TWEET DUYGU ANALİZİ

Aslı GÜRSOY

YÜKSEK LİSANS TEZİ

GAZİANTEP-2024

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
YÜKSEK LİSANS TEZ KABUL VE ONAY FORMU

Elektronik Bilgisayar Mühendisliği Anabilim Dalı **Elektronik Bilgisayar Mühendisliği** Tezli Yüksek Lisans Programı öğrencisi **Aslı GÜRSOY** tarafından hazırlanan **"Klasik Makine Öğrenme Algoritmaları ve Transformer Modeli ile Türkçe Tweet Duygu Analizi"** başlıklı tez, 19/07/2024 tarihinde yapılan savunma sınavı sonucu **başarılı** bulunarak jürimiz tarafından tarafından **Yüksek Lisans** tezi olarak kabul edilmiştir.

Görevi	**Unvanı, Adı ve Soyadı**	**Kurumu/Üniversitesi**	**İmzası:**
Tez Danışmanı	Doç. Dr. Abdul Hafız ABDULHAFIZ	Hasan Kalyoncu Üniversitesi	
Jüri Üyesi	Doç. Dr. Bülent HAZNEDAR	Gaziantep Üniversitesi	
Jüri Üyesi	Dr. Öğr. Üyesi Saed Abde Wahhab Reshid AL QARALEH	Hasan Kalyoncu Üniversitesi	

Bu tez Enstitü Yönetim Kurulunca belirlenen yukarıdaki jüri üyeleri tarafından uygun görülmüş ve Enstitü Yönetim Kurulu kararı ile onaylanmıştır.

Doç. Dr. Ufuk AKBAŞ
Enstitü Müdürü

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Aslı GÜRSOY

Tarih: 19/07/2024

HASAN KALYONCU ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
ELEKTRONİK BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

KLASİK MAKİNE ÖĞRENME ALGORİTMALARI VE TRANSFORMER MODELİ İLE TÜRKÇE TWEET DUYGU ANALİZİ

Aslı GÜRSOY

YÜKSEK LİSANS TEZİ

Danışman
Doç. Dr. Abdul Hafız ABDULHAFIZ

ÖZET

Teknolojik çağın başlamasıyla birlikte birçok uygulama günümüzde popüler hale gelmiştir. Bunlardan biri de Twitter'dır (Yeni adıyla X). Bu platform sayesinde birçok kullanıcı kişisel düşünce ve fikirlerini yazıyla, resimle ya da video ile paylaşabilmektedir. Paylaşılan bu veriler birçok insanın ilgisini çekmektedir. Bu verilerden kişisel çıkarım yapmak isteyen bilim insanınları, akademisyen ya da başka bir meslek grubundan insanlar çalışmalarını bu yönde yapmaktadır. Örneğin suçlu profili oluşturmak isteyen bir polis twitlerden bunu yapma amacıyla çalışma yürütebilir ya da bir reklam ajansı kişinin yazdıklarına ya da paylaştığı resimlere bakarak kişiye özel reklam üretebilir ve bu sayede ürün satıcıları ürünlerini hızlı bir şekilde satabilir. Bu tezde ise yüksek oranda doğruluk oranı aldığımız bir duygu analizi çalışması yapılmıştır.

Bu çalışmada hazır bir Türkçe Tweet verisi üzerinden duygu sınıflandırılması amaçlandı. Bu veri seti 5 farklı etiketle ayrılmış 4000 veriyi içermektedir. Bu ham veri seti üzerinde ön işleme uygulandı. Ön işlemden geçen veri seti eğitim ve test olarak ayrıldı. Klasik makine öğrenmesi algoritmalarının performansları ölçüldü. Bu performanslar doğruluk, kesinlik, duyarlılık ve F1 skorunun yanı sıra makro ve ağırlıklı ortalama açısından ölçülmüştür. Ayrıca, her algoritma için hesaplanan karışıklık matrisi verilmiştir. En yüksek doğruluk oranına %96,88 ile Yığın algoritması ulaşılmıştır.

Bu çalışmada ayrıca derin öğrenme alanı içerisinde olan önceden eğitimiş bir Transformer modeli kullanılmıştır. Veri seti eğitim, doğrulama ve test olarak ayrılmıştır. Aynı şekilde performansı doğruluk, kesinlik, geri çağırma ve F1 skoru açısından ölçülmüştür ve her algoritma için hesaplanan karışıklık matrisi verilmiştir. Bu model ile %93 doğruluk oranına ulaşılmıştır.

Anahtar Kelimeler: Twitter, Duygu Analizi, Makine Öğrenmesi, Derin Öğrenme, Doğal Dil İşleme

HASAN KALYONCU UNIVERSITY
GRADUATE EDUCATION INSTITUTE
DEPARTMENT of ELECTRONIC COMPUTER ENGINEERING

TURKISH TWEET SENTIMENT ANALYSIS WITH CLASSICAL MACHINE LEARNING ALGORITHMS AND TRANSFORMER MODEL

Aslı GÜRSOY

MASTER THESIS

Advisor
Assoc. Prof. Dr. Abdul Hafız ABDULHAFIZ

ABSTRACT

With the onset of the technological age, many applications have become popular today. One of these is Twitter (Newly known as X). Thanks to this platform, many users can share their personal thoughts and ideas with text, pictures or videos. These shared data attract the attention of many people. Scientists, academicians or other professionals who want to make personal inferences from these data do their work in this direction. For example, a police officer who wants to create a criminal profile can use tweets to do this, or an advertising agency can create personalized advertisements by looking at what a person writes or the pictures he/she shares, and thus product sellers can sell their products quickly. In this thesis, a sentiment analysis study was conducted in which we received a high accuracy rate.

In this study, it was aimed to classify emotions using a ready-made Turkish Tweet data. This data set includes 4000 data which is separated by 5 different labels. Pre-processing was applied on this raw dataset. The pre-processed dataset was divided into training and testing. Performances were measured with classical machine learning models. These performances were measured in terms of accuracy, precision, recall, and F1 score as well as macro and weighted averages. Additionally, the confusion matrix calculated for each algorithm is given. The Stack algorithm achieved the highest accuracy rate of 96.88%.

Also in this study, a pre-trained Transformer model, which is in the field of deep learning, was also used. The data set is divided into training, validation and testing. Likewise, its performance was measured in terms of accuracy, precision, recall and F1 score, and the confusion matrix calculated for each algorithm is given. 93% accuracy rate was achieved with this model.

Keywords: Twitter, Sentiment Analysis, Machine Learning, Deep Learning, Natural Language Processing

İÇİNDEKİLER

TABLO LİSTESİ

ŞEKİL LİSTESİ

KISALTMALAR LİSTESİ

ACC :Accuracy

AI :Artificial intelligence

ANN :Artificial neural networks

BERT :Bidirectional encoder representations from transformers

CBOW :Continuous bag of words

CCE :Categorical cross-entropy

CNN :Convolutional neural networks

CRFs :Conditional random fields

DL :Deep learning

DT :Decision tree

ELMO :Embeddings from language model

ET :Extra Tree

FN :False negative

FP :False positive

FT :Fourier transform

GLOVE :Global vectors

HOG :Histogram of oriented gradients

KDE :Kernel density estimation

KNN :K-nearest neighbors

LSTM :Long term short memory

ML :Machine learning

MLM :Masked language modeling

NER :Named-entity recognition

NLP :Natural language processing

NSP :Next sentence prediction

P :Precision

PCA :Principal component analysis

R :Recall

RELU :Rectified linear unit

RF :Random forest

RNN :Recurrent neural networks

ROBERTA :Robustly optimized bert pretraining approach

SIFT :Scale-invariant feature transform

SVC :Support vector classification

SVM :Support vector machine

T-SNE :T-distributed stochastic neighbor embedding

TF-IDF :Term frequency-inverse document frequency

TN :True negative

TP :True positive

TSA :Twitter-based sentiment analysis

WORD2VEC :Word to vector

1.GİRİŞ

Günümüz dünyasında teknolojik gelişmeler bireylerin düşüncelerini çeşitli platformlarda ifade etme biçimlerini kökten değiştirmiştir. Artık pek çok uygulama, kullanıcıların gerçek kimliklerini, kurumsal adlarını kullanarak veya anonim olarak görüşlerini paylaşmalarına olanak tanıyor. Bu uygulamalar arasında en popüler olanlardan biri de eski adıyla Twitter olan X platformudur. Bu platform sayesinde her türlü bilgi, fikir ve haber kullanıcılara bir tık uzak hale gelmiştir. Bu durum ilk bakışta avantajlı görünse de önemli dezavantajları da beraberinde getirmektedir. Örneğin sosyal medya aracılığıyla bir haberin hızla yayılması, bilginin doğruluğu teyit edilmeden spekülasyonlara yol açabilmekte, bu durum habere konu olan kişi veya gruplar üzerinde olumsuz etkiler yaratabilmektedir. Ayrıca toplumun belirli kesimlerini provokatif söylemlerle harekete geçiren anonim hesaplar, amacına ulaştığı takdirde olumsuz sonuçlar doğurabilmektedir. Bu tür olumsuzlukların önüne geçebilmek ve sosyal medya dinamiklerini daha iyi anlayabilmek için bilim insanları, dilbilimciler ve veri madencileri metin analizi alanında yoğun araştırmalar yürütmektedir.

Birjali vd. (2021), sadece araştırmacıların değil aynı zamanda şirketlerin ve hükümetlerin de ilgisini çektiğini ve yakın zamanda yayınlanan duygu analizi makalelerinin sayısının da gösterdiği gibi hızla büyüyen bir alan haline geldiğini belirtmişlerdir (Mäntylä vd., 2018). Örneğin bir ticaret şirketi, kullanıcıların sosyal medya paylaşımlarını analiz ederek, bireylerin ihtiyaçlarına uygun ürünlerin reklamını yapabilmektedirler ve bu sayede ürünlerini daha hızlı ve daha karlı bir şekilde satabilirler. Benzer şekilde bir suç analisti sosyal medya verilerini inceleyerek suç profilleri oluşturabilir ve olası suçların önlenmesine katkı sağlayabilir. Bu nedenle duygu analizi günümüzde birçok sektör için giderek daha önemli hale gelmiştir.

Bu tez çalışmasında, duygu analizi üzerine odaklanılarak, Gaussian Naive Bayes (GNB), K-En Yakın Komşu (KNN), Lojistik Regresyon, Karar Ağacı, Extra Tree, Gradient Boosting Classifier (GBC) ve Random Forest (RF), Lineer Destek Vektör Sınıflandırma (SVC) ve Lojistik Regresyonu birleştiren Stacking sınıflandırıcıları gibi klasik makine öğrenmesi algoritmalarının yanı sıra, Türkçe tweetler üzerinde BERT (Bidirectional Encoder Representations from Transformers) adlı derin öğrenme modeli kullanılarak doğruluk oranları elde edilmiştir.

1.1. Arka Plan

X platformu günümüzde sosyal medya uygulamaları arasında popülerliği hızla artan bir uygulamadır. İnsanlar tweet aracılığıyla görüntüleme metinlerini, videoları, bağlantıları vb. paylaşırlar ve mesajları iletirler (Memiş, 2024). Bu mesajlar toplumun her kesiminden insanın ilgisini çekebilmektedir. Popülerliği ve yaygın kullanımı nedeniyle kullanıcı tarafından gönderilen tweetlerin analizi yararlı hale gelmiştir (Memiş, 2024).

Kullanıcıların bu platformu kullanma amaçları belirli bir konu hakkında bilgi edinmek, doğru bilgiye ulaşmak, görüşlerine delil bulmak veya başka sebepler olabilir. Bu tezin amacı Türkçe tweetlerden duygu analizi yapmaktır. X platformu kullanıcılarının yazdıkları metinlerin hangi duygularla yazıldığını analiz etmek ve sınıflandırmaktır. Bu analiz sürecinde klasik makine öğrenmesi algoritmalarının yanı sıra derin öğrenme modeli de kullanılarak başarılı sonuçlar elde edilmiştir.

1.2. Sorun Açıklaması

Son zamanlarda sosyal ağ hizmetlerinin (SNS) hızla büyümesi nedeniyle, yorum ve değerlendirme gibi büyük miktardaki bilgiler kullanıcılar tarafından sürekli olarak sağlanmaktadır (Zimbra vd., 2018). Bu bilgi, çoğu zaman tek bir ilginç gerçeğe dayanan, insanların düşüncelerini ve duygularını yansıtır (Wang vd., 2022). Rui vd. (2013), bu verilerin büyük veri haline geldiğini, özellikle ürün tahminlerinde insan davranışını analiz etmek için birçok fırsat sunduğunu belirtmişler ve Wang vd. (2012) siyasi seçimlerin sonuçlarını tahmin etmenin oldukça faydalı olduğunu belirtmişlerdir. Twitter Kullanıcı İstatistikleri 2024'e (2022) göre 368 milyon aktif Twitter kullanıcısı bulunmaktadır, dolayısıyla Twitter en popüler blog hizmetlerinden biri haline gelmiştir (Reyna vd., 2022). İnsanların düşünce ve duygularını anlamada önemli bir rol oynayan Twitter tabanlı duygu analizi (TSA) oldukça ilgi çekici olmuştur (Khan vd., 2021; Meng vd., 2022).

Birçok araştırmacı bu alanda çeşitli çalışmalar yürütmekte olup, özellikle İngilizce tweetler üzerine yapılan duygu analizi çalışmaları yaygınlık göstermektedir. Ancak, Türkçe tweetler üzerine gerçekleştirilen araştırmaların sayısı nispeten sınırlıdır. Bu çalışmanın, Türkçe tweetlerden elde edilen verilerle gerçekleştirilecek duygu analizi üzerine önemli bir katkı sunacağı öngörülmektedir.

1.3. Araştırmanın Önemi

Bu tezin önemi, makine öğrenmesi (ML) algoritmaları ve derin öğrenme (DL) modeli olan BERT modeli ile Türkçe tweetler kullanılarak duygu analizinin gerçekleştirilmesinin amaçlanmış olmasıdır. Bu hedefe ulaşmak için atılan adımlar şu şekildedir:

- Türkçe tweetleri içeren bir veri seti bulma.
- Temiz bir veri kümesi elde etmek için verilerin ön işlenmesi.
- GNB, Lojistik Regresyon, KNN, GBC, Yığınlama (temel sınıflandırıcılar olarak RF ve Doğrusal SVC ve son sınıflandırıcı olarak Lojistik Regresyonun kullanılması), Karar Ağacı ve Ekstra Ağaç gibi klasik makine öğrenimi modellerinin uygulanması.
- Bir transformatör modeli olan BERT'in kullanılması.
- Değerlendirme metriklerini kullanarak performansı hesaplamak.

Bu süreç, Türkçe tweetlerden elde edilen verilerle duygu analizi yapmayı ve farklı algoritmaların performanslarını karşılaştırarak en etkili yöntemleri belirlemeyi amaçlamaktadır.

1.4. Tezin Düzeni

Tezin ikinci bölümünde arka plan bilgileri ve literatür taraması yer almaktadır. Üçüncü bölümde metodoloji ayrıntılarıyla anlatılmaktadır. Dördüncü bölümde deneysel çalışmalar ve sonuçlar sunulmaktadır. Beşinci ve son bölümde ise sonuçların değerlendirilmesi, genel çıkarımlar ve gelecek araştırmalara yönelik öneriler yer almaktadır.

2.ARKA PLAN VE LİTERATÜR TARAMASI

2.1 Doğal Dil İşleme (NLP)

Doğal Dil İşleme (NLP), bilgisayarların yazılı veya sözlü metinleri doğal dilde anlaması ve işleyebilmesi için kullanılan bir çalışma ve uygulama alanıdır. Fanni vd. (2023), bu alanın, bilgisayar ve insan arasındaki etkileşimi daha kolay ve verimli hale getirdiğini belirtmiştir. NLP, birçok alanda kullanılabilmektedir. Aşağıda bu alanlar verilmiş ve açıklanmıştır.

2.1.1 Metin sınıflandırma

Metin sınıflandırma, yazıları veya belgeleri belirli etiketlerle ayırma yöntemidir. Bu tezde yapılan duygu analizi çalışması, bu yönteme örnek olarak verilebilir. Çoğu metin sınıflandırma ve sınıflandırma sistemleri dört aşamaya ayrılabilir: öznitelik çıkarma, parametre azaltma, sınıflandırma ve değerlendirme.

2.1.2 Makine çevirisi

Makine çevirisi, bir metnin dilini belirleyip, onu istenilen dile çevirme yöntemidir. Dil modülleri, çeviri hafızaları ve gelişmiş sistemler kullanarak çeviri kalitesini artırır. Google Translate buna bir örnek olarak verilebilir.

2.1.3 Konuşma tanıma

Konuşma tanıma, konuşulan bir metni veya metin parçasını metne dönüştürme yöntemidir. Android işletim sistemlerindeki Google Asistan, IOS işletim sisteminin sesli asistanı Siri ve telefon görüşmeleri, bu teknolojinin örnekleridir. Makine öğrenmesi modelleri ve doğal dil işleme, bu yöntemin doğruluğunu artırmak için kullanılır.

2.1.4 Soru-cevap sistemi

Soru-cevap sistemi, herhangi bir soruya otomatik olarak yanıt üretme yöntemidir. Bu sistem, büyük veri tabanları üzerinde çalışır ve metni anlayarak, sorulan sorulara daha kapsamlı ve doğru yanıtlar sağlayabilir.

2.1.5 Bilgi çıkarma

Bilgi çıkarma, belirli bir metinden bilgi elde etme stratejisidir. Bu alanın bir örneği, bir metin veya içerikten (genellikle yapılandırılmamış) bir olayın tarihi veya bahsedilen ürünün özelliklerinin otomatik olarak tanımlanması ve sınıflandırılmasıdır.

2.1.6 Tavsiye sistemleri

Tavsiye sistemleri, müşterilerin bireysel ihtiyaçlarına veya taleplerine uygun öneriler oluşturma stratejisidir. Özellikle e-ticaret sistemlerinde görülebilir. Bu sistemler, kullanıcı deneyimini iyileştirmek ve kullanıcı sadakatini sağlamak için önemli bir rol oynar.

2.2. Özellik Çıkarma

Özellik çıkarma, ham verilerden bilgi açısından zengin ve daha önemli vurguları çıkarmak için kullanılan yöntemdir. Avinash ve Sivasankar (2019), makine öğrenme modelleri açısından temel ve bilgilendirici özelliklerin çıkarılmasının performansı artırmada ve hesaplama karmaşıklığını azaltmada avantaj sağladığını belirtmişlerdir. Özellik çıkarma, ham bilgileri daha kompakt ve analiz dostu bir forma dönüştürerek bilgi hazırlama formlarını teşvik eder. Özellik çıkarma yöntemleri, bilginin türüne ve analizin nedenine bağlı olarak değişmektedir. Aşağıda birkaç yaygın özellik çıkarma stratejisi verilmiş ve açıklanmıştır.

2.2.1. Metin verilerinden özellik çıkarma

Bu bölümde metinden özellik çıkarma teknikleri verilmektedir.

2.2.1.1. Terim frekansı-ters belge sıklığı (TF-IDF)

Terim frekansı-ters belge frekansı (TF-IDF), kelimelerin önemine karar vermek için kullanılan bir stratejidir. Belirli bir kelimenin bir arşivde ne sıklıkta geçtiğini ve o kelimenin tüm belgeler arasında ne kadar nadir bulunduğunu hesaplar. Metin sınıflandırma, bilgi alma ve duygu analizi için kullanılabilir.

2.2.1.2. Kelime yerleştirme

Peng vd. (2020), kelime yerleştirmenin, duygu analizi gerçekleştirirken doğruluğu ve performansı artırmak için kelimeleri dağıtılmış özellikler olarak temsil etmek için kullanılan bir teknik olduğunu belirtmişlerdir. Bu teknikle, farklı tekniklerle büyük miktarda bağlamsal bilgi kullanılarak kelimelerin dağıtılmış vektör temsilleri öğrenilmeye çalışılır (Kasri vd., 2022). Bu tekniğin kullanılmasının nedeni cümledeki kelime bağlamlarını daha iyi anlamaktır.

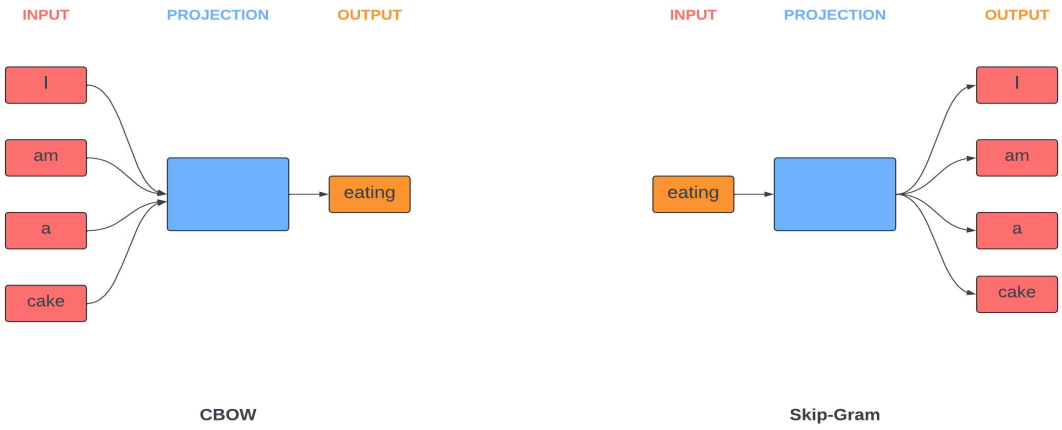
Birçok kelime yerleştirme tekniği vardır. Aşağıda bunlar verilmiş ve açıklanmıştır.

2.2.1.2.1. Kelimeden vektöre (Word2Vec)

Word2Vec, Google ekibinden Tomas Mikolov ve ekip arkadaşları tarafından geliştirilen ve tahmin esasına göre kelimeleri vektör uzayında ifade etmeye çalışan bir model olarak 2013 yılında ortaya çıkan bir modeldir. Bu modelin Sürekli Kelime Çantası (CBOW) ve Skip-gram adı verilen 2 alt yöntemi vardır. Şekil 2.1 CBOW ve Skip-gram'ın çalışma prensibini göstermektedir.

- ***Sürekli kelime çantası (CBOW):*** CBOW modeli, hedef kelimeyi çevreleyen bağlam kelimelerinden tahmin eder. Yani etrafındaki kelimeleri kullanarak ortadaki kelimeyi tahmin edebilir. Bu model tüm bağlam sözcüklerini alır, onları bir araya getirir ve ortaya çıkan vektörü kullanarak hedef sözcüğü tahmin eder.

- ***Skip-gram:*** Skip-gram modeli, hedef kelimeye dayalı olarak çevredeki bağlam kelimelerini tahmin eder. Başka bir deyişle, bir kelimeyi etrafındaki bağlamı tahmin etmek için kullanmak üzere tasarlanmıştır.



Şekil 2. 1. CBOW and Skip-gram modellerinin çalışma prensibi

2.2.1.2.2. Kelime temsili için küresel vektörler (GloVe)

Stanford Üniversitesi'ndeki bir grup araştırmacı tarafından 2014 yılında geliştirilen GloVe, benzer veya yakın anlamlı kelimeleri, kelime uzayındaki temsillerine bakarak birbirine yakın konumlandıran bir tekniktir. Bu teknikle kelimeler arasındaki benzerlik ilişkisi matematiksel olarak modellenebilmektedir.

GloVe ile büyük veri kümelerinde kelime çiftlerinin bir arada görünme olasılığına bakılarak bir frekans matrisi oluşturuluyor. Bu matris ile bu kelime çiftlerinin bir bağlamda kaç kez tekrarlandığı hesaplanır. Oranın dengeli ve düzenli olmasını sağlamak için kelime çiftlerinin birlikte görülme sıklıkları arasındaki oranların logaritması alınır ve bu şekilde modellenir. Logaritmanın kullanılmasının nedeni frekans farklılıklarını azaltmak ve daha yönetilebilir hale getirmektir.

Global istatistiklerden faydalanılması, boyutların küçültülmesi, yüksek performans ve taşınabilirlik GloVe tekniğinin en önemli özellikleri arasındadır. GloVe ile kelime çiftlerinin tümce boyunca birlikte ortaya çıkma olasılıklarını kullanmak, kelime vektörlerini anlamlı hale getirir. Bu teknik büyük veri kümelerinde iyi çalışır. Ayrıca sözcük yerleştirmelerin anlamlı şekilde öğrenilmesini sağlar. Bu teknikle daha küçük boyutlu vektör uzaylarında çalışmak hem daha düşük maliyetlere hem de kullanılan modelden daha fazla verim alınmasına olanak sağlar.

GloVe, metin sınıflandırma, makine çevirileri ve duygu analizi gibi birçok alanda kullanılabilmektedir. Kısaca bu yöntemin kelime gömme yöntemleri alanında önemli bir yere sahip olduğu söylenebilir.

2.2.1.2.3. Transformatörlerden çift yönlü kodlayıcı gösterimleri (BERT)

Transformer Çift Yönlü Kodlayıcı Gösterimleri (BERT) modeliyle, transformerların çift yönlü mimarisi kullanılarak belirli bir kelimenin bağlamsal vektör temsilleri çıkarılır. Ayrıca bu model ile verilen cümlenin hem başından hem de sonundaki bağlam bilgileri kullanılarak kelime anlamı daha doğru bir şekilde çıkarılmaktadır. Metodoloji bölümünde daha detaylı anlatılmıştır.

2.2.1.3. N-gram

N-gram, metinden n ardışık kelime veya karakterden oluşan grupları çıkarır. Tek kelime (unigram), çift kelime (bigram), üçlü kelime (trigram) olabilir. Metin sınıflandırma ve dil modelleme için kullanılabilir.

2.2.2. Görüntü verilerinden özellik çıkarma

Bu bölümde görüntüden özellik çıkarma teknikleri verilmektedir.

2.2.2.1. Yönlendirilmiş degradelerin histogramı (HOG)

Yönlendirilmiş degradelerin histogramı (HOG), görüntülerdeki nesne kenarlarını ve açılarını yakalar. Görüntüyü küçük hücrelere böler ve her hücredeki gradyan yönlerinin histogramını hesaplar. Nesne tanıma ve insan tespiti için kullanılabilir.

2.2.2.2. Ölçekle değişmeyen özellik dönüşümü (SIFT)

Ölçekle değişmeyen özellik dönüşümü (SIFT), görüntüdeki ölçek ve dönüş açısından bağımsız özellik noktalarını (anahtar noktalar) çıkarır. Görüntü eşleştirme ve nesne tanıma için kullanılabilir.

2.2.2.3. Evrişimli sinir ağları (CNN)

Evrişimli Sinir Ağları (CNN), görüntü ve video gibi iki boyutlu veriler üzerinde üstün performans gösteren derin öğrenme modelleridir. Temel bileşenleri evrişimli katmanları, havuzlama katmanlarını, tamamen bağlı katmanları ve aktivasyon fonksiyonlarını içerir. Evrimsel katmanlar, görüntülerdeki belirli desenleri tanımak için filtreler kullanarak özellik haritaları oluşturur. Havuzlama katmanları, özellik haritalarının boyutunu azaltır, hesaplama yükünü azaltır ve modelin genelleme yeteneğini artırır. Tamamen bağlı katmanlar modelin son tahminlerini yapar. Aktivasyon fonksiyonları, doğrusal olmayan dönüşümler yoluyla modelin karmaşıklığını artırır. CNN'ler ayrıca n-gramlar, dikkat mekanizmaları, adım ve dolgu teknikleri gibi özellikleri kullanarak verileri işler ve onlardan öğrenir. Modelin genel başarısını artırmak için geriye yayılım yöntemiyle ağırlıklar güncellenir ve ağırlık paylaşımıyla parametre sayısı optimize edilir. CNN'ler görüntü sınıflandırma, nesne tespiti, yüz tanıma ve tıbbi görüntü analizi gibi alanlarda yaygın olarak kullanılmakta ve yüksek doğruluk oranlarıyla etkili sonuçlar sağlamaktadır.

2.2.3. Sayısal ve kategorik verilerden özellik çıkarma

Bu bölümde sayısal ve kategorik verilerden özellik çıkarma teknikleri verilmektedir.

2.2.3.1. Temel bileşen analizi (PCA)

Temel bileşen analizi (PCA), ana bileşenlerini çıkararak verinin boyutunu azaltır. Ana bileşenler, veri varyansının çoğunu açıklayan doğrusal kombinasyonlardır. Boyut küçültme, veri görselleştirme amacıyla kullanılabilir.

2.2.3.2. T-dağıtılmış stokastik komşu yerleştirme (T-SNE)

T-dağıtılmış stokastik komşu yerleştirme (T-SNE), yüksek boyutlu verileri düşük boyutlu bir alana yansıtarak veri kümelerini görselleştirir. Özellikle küme yapısını korur. Veri görselleştirme ve küme analizi için kullanılabilir.

2.2.3.3. One-hot kodlama

Tek geçişli kodlamada kategorik veriler, her kategori için ayrı bir sütun oluşturarak ikiliyi (0 ve 1) temsil eder. Makine öğrenimi modelleri için kategorik verileri sayısal forma dönüştürmek için kullanılabilir.

2.2.4. Zaman serisi verilerinden özellik çıkarma

Bu bölümde zaman serilerinden özellik çıkarma teknikleri verilmektedir.

2.2.4.1. Fourier dönüşümü (FT)

Fourier dönüşümü (FT), zaman serisi verilerini frekans bileşenlerine ayırır. Frekans analizi yaparak periyodik özellikleri çıkarır. Sinyal işleme ve frekans analizi için kullanılabilir.

2.2.4.2. Dalgacık dönüşümü

Dalgacık dönüşümü, zaman serisi verilerini hem zaman hem de frekans alanlarında analiz eder. Ölçeklenebilirlik sağlar. Zaman-frekans analizi ve sinyal işleme için kullanılabilir.

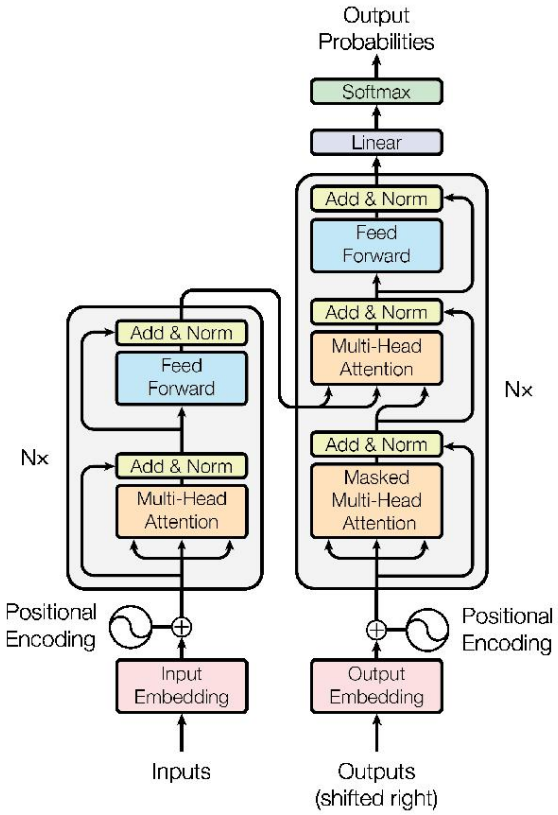
2.2.4.3. Otomatik korelasyon

Otomatik korelasyon, zaman serisi verilerinin kendisiyle olan korelasyonunu hesaplar. Verilerdeki periyodik kalıpları ve ilişkileri tanımlar. Zaman serisi analizi ve sinyal işleme için kullanılabilir.

2.3. Transformer

Transformer, Vaswani ve arkadaşlarının bir makalesinde tanıtılan, doğal dil işleme ve veri işleme alanlarında kullanılan bir makine öğrenme modelidir. Tekrarlayan Sinir Ağları (RNN) ve Uzun Süreli Kısa Bellek (LSTM) gibi modelleri birçok açıdan geride bırakarak, yüksek doğruluk elde edebileceğiniz bir model olarak öne çıkmaktadır.

Transformer mimarisi özellikle dikkat mekanizmasını temel alır ve iki temel bileşenden oluşur: Kodlayıcı ve kod çözücü. Şekil 2.2 Transformer modelinin yapısını göstermektedir (Vaswani vd., 2017). Aşağıda bunlar ayrıntılı olarak açıklanmaktadır.



Şekil 2. 2. Transformer mimarisi (Vaswani vd., 2017)

2.3.1. Kodlayıcı

Transformer kodlayıcı, dikkat mekanizmasına dayalı modern bir sinir ağı bileşenidir. Encoder bölümünde öncelikle giriş işlemleri yapılır. Veriler bir dizi sayısal veriye dönüştürülür. Model daha sonra veri sırasını belirlemek için giriş verilerine konumsal kodlama bilgisi ekler. Daha sonra çok başlı dikkat mekanizması dediğimiz kodlayıcı kısmında belli bir giriş verisinin diğer verilerle ilişkisini öğrenir. Son olarak, önceki mekanizmayı takip ederek veriler tamamen bağlı katman aracılığıyla işlenir.

Transformatör kodlayıcı, özellik çıkarımına benzer ancak bazı farklılıkları vardır. Bu farklar şu şekilde sıralanabilir:

- Kodlayıcı genellikle model içindeki bir bileşendir, özellik çıkarımı ön işlemede kullanılan bir tekniktir.

- Kodlayıcı dikkat mekanizması ve öz-dikkat ile çalışır ancak özellik çıkarımı manuel veya otomatik olarak yapılır.
- Kodlayıcı özellikle metin ve konuşma gibi sıralı veriler üzerinde etkilidir; özellik çıkarımı metin, resim veya sayısal veri gibi her türlü veriye uygulanabilir.

2.3.2. Kod çözücü

Kod çözücü, kodlayıcı tarafından üretilen ara çıktıları alır ve bunları hedef diziyi üretmek için işler. Kod çözücü hem kendi girişleri hem de kodlayıcı çıkışları üzerinde bir dikkat mekanizması kullanır. Aşağıda bu mekanizmalar açıklanmaktadır.

- ***Maskelenmiş çok başlı dikkat:*** Kod çözücü, girdi dizisinin yalnızca belirli bölümlerini dikkate alarak çıktı üretir. Bu özellikle dil modellemede kullanılır.
- ***Çıktı üretimi:*** Kod çözücü, verileri tamamen bağlı katmanlardan geçirerek nihai çıktıyı üretir.

2.3.3. Dikkat mekanizması

Dikkat mekanizması bir girdinin diğer tüm girdilerle ilişkisini öğrenme yeteneğine sahiptir. Bu mekanizma, her bir girdi öğesinin diğer tüm girdiler üzerindeki etkisini hesaplar ve bu hesaplamaları kullanarak çıktıyı oluşturur. Aşağıda dikkat mekanizmasının nasıl çalıştığı ve temel bileşenleri anlatılmaktadır.

2.3.3.1. Puan hesaplama

Puan hesaplamasında her girdinin diğer girdilerle benzerlik puanı hesaplanır.

2.3.3.2. Softmax uygulaması

Hesaplanan puanlar softmax işlevi kullanılarak normalleştirilir, böylece her girdinin diğer girdilerle karşılaştırıldığında göreceli önemi belirlenir.

2.3.3.3. Ağırlıklı toplama

Girdi verileri, elde edilen önem düzeylerinin çarpılmasıyla ağırlıklı olarak toplanır ve bu işlem her bir girdi öğesi için tekrarlanır.

2.4. Metin Madenciliği ve Ön İşleme

Metin madenciliği, verilerden yararlı bilgiler bulma veya çıkarma sürecidir (Vijayarani vd., 2015). Kısa vd. (2010), son otuz yılda araştırmacıların, yazılı bilgiyi özetlemek için anlamla ilgili kelimeleri ve cümleleri sayan yapıları (örneğin, girişimci yönelim) puanlamak için kapalı kelimeler kullanan metin madenciliği yöntemlerini kullandıklarını belirtmişlerdir. Bu teknikler yapılandırılmış veya yapılandırılmamış verilerden anlamlı bilgiler çıkarmayı amaçlamaktadır. Bu şekilde büyük hacimli belgeler işlenir ve analiz edilerek önemli bilgiler ortaya çıkarılır. Bu görevi gerçekleştirmek için birçok farklı algoritma kullanılabilir.

Ön işleme aşamasında veri setindeki temel veriler temizlenir ve analizler için ileri işlemlere hazırlanır. Bu bölüm çok önemlidir çünkü birincil veriler çoğu zaman analitik yöntemlere uygun değildir. Aşağıda ön işleme yöntemleri anlatılmaktadır.

2.4.1. Tokenizasyon

Tokenizasyon, metin verilerini daha küçük parçalara ayırma işlemidir. “Ben bir akademisyenim.” cümlesi, tokenizasyondan sonra [“Ben”, “bir”, “akademisyenim”] olacaktır.

2.4.2. Filtreleme

Filtreleme işlemi ile metne anlam katmayan, sık kullanılan ve önemsiz kelimeler filtrelenir. Filtreleme sonrasında “bu” ve “bu nedenle” gibi kelimeler cümleden silinir. Bu tür kelimelerin silinmesinin amacı analizin daha anlamlı hale getirilmesidir.

2.4.3. Lemmatizasyon ve kök çıkarma

Lemmatizasyon ile bir kelime kök haline getirilir. Örnek olarak “koşuyor” kelimesi lemmatizasyondan sonra “koşmak” olarak, “çiçekler” kelimesi ise “çiçek” olarak köklere dönüştürülür. Kök çıkarma ise kelimelerin basit kök haline getirilmesi işlemidir. Genel olarak önek ve son eklerinin kesilerek kelime alınması işlemi de denebilir. Örnek olarak “koşuyor” kelimesi kök çıkarmadan sonra “koş”, “çiçekler” kelimesi ise “çiçek” haline getirilir.

2.4.4. Normalleştirme

Normalleştirme kısmında birkaç temel adım gerçekleştirilir. Genellikle cümlenin tamamı büyük veya küçük harfe dönüştürülebilir, “@” gibi özel karakterler cümleden çıkarılabilir veya normalizasyon sonrası sayısal veriler çıkarılabilir.

2.5. Literatür İncelemesi

NLP alanlarından biri olan duygu analizi önemli bir araştırma konusudur. Özellikle son yıllarda Türkçe çalışmalar büyük önem kazanmıştır. Çeşitli makine öğrenmesi modellerinin yanı sıra Türkçeye özgü duygu analizi yöntemleri için geliştirilen yöntemlerle de bu yönde yapılan çalışmaların doğruluğu ve etkinliği artmaktadır. Böylece farklı duyguların analizinde ve sınıflandırılmasında bir artış söz konusudur. Literatürdeki bazı çalışmalar Türkçe metinlere özel olarak geliştirilmiş olup, bunların analizinde daha yetkin sonuçlar vermektedir.

Literatür çalışmasının amacı daha önce yapılmış Türkçe metinlerden duygu analizi çalışmalarının yöntem ve yaklaşımlarını tartışmaktır. Bu doğrultuda yapılan çalışmalarda elde edilen yöntem ve sonuçlar bu bölümde ele alınmaktadır. Ayrıca bu konuda karşılaşılan zorluklara ve bu zorluklara yönelik çözüm önerilerine de yer verilmiştir.

Bu çalışmada Türkçe metinden duygu analizine ilişkin mevcut bilgilere ek bilgi sağlanması ve bu konuda gelecekte yapılacak çalışmalara referans olması amaçlanmaktadır. Dile özgü etkili yöntemlerle duygu analizi çalışmalarının daha hızlı, daha verimli ve daha etkili olması muhtemeldir.

Çoban vd. (2015), blog, Facebook ve Twitter gibi sosyal medya platformlarından elde edilen metinler üzerinde bir çalışma yürütmüştür. Çalışmanın amacı firmanın müşterilerinin memnuniyetini arttırmak ve maliyetlerini azaltmaktır. 14777 mesajdan oluşan veri seti, kelime çantası ve karakter tabanlı n-gram modeli kullanılarak analiz edilmiştir. Ayrıca sınıflandırma algoritmalarından Destek Vektör Makinesi, Naive Bayes, Multinomial Naive Bayes ve K-En Yakın Komşular algoritmaları kullanılmıştır. En iyi performans n-gram modelinin 3 gram karakter seviyesinde elde edilmiştir. Multinomial Naive Bayes algoritması her iki model için de en iyi performansı gösterdi.

Doğan & Kaya (2019), sosyal ağlar üzerinde duygu analizi çalışması gerçekleştirmişlerdir. Bu çalışmanın amacı belirli markaların sosyal medya platformlarındaki yorumlarından veri setleri oluşturmak ve bu verileri analiz ederek duygu analizi konusunda bilgi vermektir. Çalışmada %70 doğruluk oranı elde edilmiş olup bu durum makine öğrenmesi modelleri ile duygu analizi yapılabileceğini göstermektedir. Ancak sosyal medya kullanıcılarının yazım kurallarına dikkat etmemesi gibi faktörler doğruluk oranlarının geniş bir aralıkta değişmesine neden olmuştur. Rastgele Orman, Lojistik Regresyon, Multinomial Naive Bayes ve Destek Vektör Makinesi algoritmaları kullanıldı. Yorumlar sayısal verilere dönüştürülerek CountVectorizer ve TF-IDF yöntemleri kullanıldı. Veri seti sosyal medya platformlarından toplanmış olup 1004 olumlu, 1000 olumsuz ve 1000 nötr yorumdan oluşmaktadır. Veri setinin %75'i eğitim, %25'i ise test için kullanıldı. Türkiye'de yapılan çalışmalarda genel olarak anketler, tweet'ler, film eleştirileri, e-ticaret sitelerindeki yorumlar, Facebook yorumları ve çağrı merkezi yorumları kullanıldı. Bu çalışmalarda başarı oranları %42 ile %91 arasında değişmektedir. Lojistik Regresyon algoritması %70 doğruluk oranıyla en iyi sonucu verirken, KNN algoritması %56 doğruluk oranıyla en düşük sonucu verdi. TF-IDF ve CountVectorizer + TF-IDF yöntemleri arasında anlamlı bir fark gözlenmedi. Duygu analizi için oluşturulan modelde kelime gösterimleri değiştirilerek bu değişikliklerin sonuçlara etkisi gözlemlenmiştir. FastText, sınıflandırma problemlerinde Word2Vec modeline göre daha yüksek bir başarı oranı sağlamıştır. Sosyal medya verileri yazım kurallarına uymadığı için başarı oranları yazılı metinlere göre daha düşüktü.

Shehu vd. (2019), polarite sözlüğünü ve yapay zekayı kullanarak Türkçe tweetlerden duygu analizi gerçekleştirmiştir. Polarite sözlüğü yöntemi ile kelimeleri bir duyarlılık sözlüğüyle karşılaştırarak tweet'leri olumlu, olumsuz veya nötr olarak sınıflandırdı.

Yapay zeka yöntemi olarak ise SVM ve RF algoritmaları kullanıldı. Deneysel sonuçlar, SVM'nin köklü verilerde %76 doğrulukla daha iyi performans gösterdiğini, RF'nin ise ham verilerde %88 doğrulukla daha iyi performans gösterdiğini ortaya çıkardı.

Aydın & Güngör (2021), denetimli, yarı denetimli ve denetimsiz öğrenme yöntemlerini karşılaştırmış ve bu tekniklerin etkililiğini ve kullanım alanlarını incelemiştir. Araştırma, Türkçede duygu analizi yapılırken denetimli yöntemlerin yüksek doğruluk oranlarına ulaştığını ancak etiketli veri eksikliği durumunda yarı denetimli ve denetimsiz yöntemlerin de başarılı sonuçlar verdiğini ortaya çıkardı.

Demircan vd. (2021), sosyal medyada ifade edilen metinlere dayanarak duygu analizi tahmini gerçekleştirdi. 5 makine öğrenmesi algoritması kullanılarak metinlerin olumlu, olumsuz ve nötr olarak sınıflandırılması amaçlandı. SVM ve RF algoritmaları diğer algoritmalara göre daha iyi performans göstermiştir. Ancak nötr kelimelerin kesinlik değerlerinde Lojistik Regresyon modeli SVM modelinden, pozitif metinlerin kesinlik değerlerinde ise DT ve KNN modelleri RF modelinden daha iyi sonuçlar vermiştir. Ayrıca negatif etiketli metinlerde DT, LR ve KNN algoritmalarının hatırlama değerleri RF algoritmasına göre daha yüksek çıkmıştır. Sonuç olarak çalışma, Türkçe metinlerin duygu analizi yapılırken SVM ve RF algoritmalarının ihmal edilebilir hata oranlarıyla kullanılabileceği sonucuna varmıştır.

Yıldırım vd. (2015), yaptıkları çalışmada NLP tekniklerinin duygu analizine katkısını gözlemlemişlerdir. Araştırmada TF-IDF ve n-gram gibi metin ön işleme teknikleri, kök bulma ve özellik çıkarma teknikleri kullanılmıştır. Kullandıkları teknikler sayesinde; başarı oranı neredeyse %6 artmıştır.

Çıplak & Yıldız (2024), sosyal ağlardaki paylaşımlar aracılığıyla kişilerin meslek gruplarını tahmin etmeyi amaçlayan bir çalışma yürütmüştür. Özellikle Twitter verileri kullanılarak makine öğrenmesi yöntemleriyle Türkçe paylaşım yapan kullanıcıların meslek gruplarının tahmin edilmesi amaçlandı. Çalışma kapsamında meslek grupları belirlenerek bu meslek gruplarına ait Twitter hesaplarından paylaşılan tweetler toplanmıştır. Toplamda 500.000'den fazla tweet içeren çeşitli veri setleri oluşturulmuş ve Zemberek kütüphanesi kullanılarak bu veri setleri üzerinde ön işlemler yapılmıştır. Tekli ve ikili yedekli kelime listeleri manuel olarak oluşturulmuş, metin verileri sayısal verilere dönüştürülmüş ve özellik çıkarımı ile yeni değişkenler oluşturulmuştur. Araştırmada farklı özellikler kullanmak yerine, en değerli özelliklerin belirlendiği "en

uygun sayıda özelliğin belirlenmesi" yöntemi kullanıldı. Çoklu yaklaşımın kullanıldığı makine öğrenimi deneylerinde %97,3 başarı oranı elde edildi. Bu oran daha önce yapılan benzer çalışmalara göre daha yüksektir.

Alqaraleh (2020), Türkçe metinlerden duyguları analiz ederken topluluk öğrenme yaklaşımını kullanmıştır. 'Beyazperde' internet sitesinden alınan 'Türk film eleştirileri' veri setinde olumlu ve olumsuz olarak sınıflandırılan 34990 metin üzerinden gerçekleştirilen bu çalışmada veri seti 4 veri setine bölünmüştür. Bu veri setleriyle ayrı ayrı 4 deney yapıldı. İlk deneyde RF, AdaBoost ve GBC topluluk öğrenme tekniklerinin performansları gözlemlendi. RF modeli ile %70 doğruluk oranına ulaşılmıştır. İkinci deneyle veri ön işlemenin performansa katkısı gözlemlendi. Ön işleme verilerinin performansa yaklaşık %30 oranında etki ettiği görülmüştür. 3. deneyde özellik çıkarma yöntemleri (Terim Frekansı (TF), Terim Frekansı-Ters Belge Frekansı (TF-IDF) ve Word2Vec) kullanıldığında elde edilen sonuçlara göre Word2Vec yöntemiyle bir veri setinden %87 doğruluk oranı elde edildi. Bu durum bu yöntemin Türkçe veri setleri için uygun bir yöntem olduğunu göstermektedir. Son deneyde, geliştirilen yaklaşımın genel performansı, sağlamlığı ve ölçeklenebilirliği, tüm veri seti kullanılarak performans metrik puanları ile incelenmiştir. Bu deneyle %87 ila %85 arasında doğruluk oranları elde edilmiştir.

Demir & Bilgin (2023), Türkçe haber metinlerine dayalı veri setlerini kullanarak yaptıkları çalışmada BERT tabanlı RoBERTa, DISTILBERT, ALBERTA modellerinin yanı sıra SVM ve Naive Bayes gibi makine öğrenmesi algoritmalarını kullanarak performansları değerlendirmişlerdir. Makine öğrenimi modellerinin %68 ile %71 arasında doğruluk oranlarına ulaştığını, BERT modellerinin ise %80 gibi daha yüksek doğruluk oranlarına ulaştığını gözlemlenmiştir. Buradan yola çıkarak derin öğrenme modellerinin duygu analizi alanında klasik makine öğrenmesi modellerine göre daha etkili olduğu sonucuna varmışlardır.

Kavi (2020), Türkçe metin sınıflandırmada sözlük analizi, SVM ve Extreme GBC gibi geleneksel yöntemlerin yanı sıra BERT gibi derin öğrenme modellerinin başarısını değerlendiren çalışmayı gerçekleştirmiştir. BERT, Türkçe metin sınıflandırmada önceki yöntemlere göre önemli ölçüde daha iyi performans göstermiştir. Özellikle BERT modeliyle elde edilen doğruluk oranı %92,5 olmuştur. Ancak BERT'in eğitim süresi geleneksel makine öğrenimi yöntemlerine göre daha uzun olduğu görülmüştür.

3.METODOLOJİ

Türkçe tweetlerden duygu analizi yapmanın birçok adımı vardır. Bunlar arasında veri toplanması, toplanan ham verilerin temizlenmesi ve ön işleme tabi tutulması, veri setinin eğitilmesi, eğitilen veri setinin performansının klasik makine öğrenmesi algoritmaları ve BERT olan derin öğrenme modeli ile değerlendirilmesi yer almaktadır. Hazır veri setleri kullanılabilir. Performansın artırılabilmesi için bu veri setlerinin tüm gereksiz verilerden temizlenmesi gerekmektedir. Kelime yerleştirmeyle, tüm veriler makine tarafından okunabilir bir forma dönüştürülür ve sınıflandırıcılara aktarılarak eğitilmeleri ve öğrenilmeleri sağlanır. Eğitim sonrasında duygu sınıflandırması için modele test verileri verilerek veri setindeki tüm veriler 5 duygu etiketi ile sınıflandırılarak model performansı değerlendirilmiştir. Bu tez çalışmasında GNB, Logistic Regresion, KNN, GBC, Linear SVC, Extra Trees, Decision Trees ve Staking makine öğrenmesi algoritmaları ile Transformer modeli olan BERT modeli kullanılmıştır.

3.1. Veri Toplama ve Ön İşleme

Bu bölümde tez çalışmasında toplanan veri seti ve bu veri setindeki verilere uygulanan ön işleme teknikleri hakkında bilgi verilmektedir.

3.1.1 Veri toplama

Veri toplama adımı en önemli adımlardan biridir. Bu adımda tek veya farklı kaynaklardan toplanan veriler gerekli işlemler için kullanılır. Doğal dil işlemede, öğrenme sürecini iyileştirmek ve yüksek performansa ulaşmak için büyük veri kümelerine ihtiyaç vardır. Veri toplama aşaması, verilerin toplanması, etiketlenmesi veya iyileştirilmesi süreçlerini içerir. Toplanan veriler, kullanılacak makine öğrenmesi ve derin öğrenme algoritmaları için bilgi kaynağıdır. Bu nedenle verilerin toplanması ve doğru şekilde etiketlenmesi çok önemlidir. Veriler başka kaynaklardan alınabildiği gibi sentetik olarak da oluşturulabilmektedir. Aynı şekilde oluşturulan veri setinin üzerinde çalıştığınız çalışma için yeterli olması ve doğru şekilde etiketlenmesi de önemli ve gereklidir.

3.1.2. Veri ön işleme

Veri ön işleme, doğru ve geçerli veri analizlerinin temel adımlarından biridir (Fan ve ark, 2021). Bu süreç, ham verilerin makine öğrenmesi ve derin öğrenme algoritmalarına uygun hale getirilmesi için temizlenmesini, azaltılmasını, dönüştürülmesini ve zenginleştirilmesini içerir. Amaç, verileri en verimli şekilde kullanarak yüksek performans elde etmektir. Bu aşamada cümlelerdeki tüm kelimelerin küçük harfe dönüştürülmesi, URL'lerin, hashtag'lerin ve emojilerin kaldırılması, sayıların ve noktalama işaretlerinin kaldırılması, tekrar eden kelimelerin silinmesi, anlamsız kelimelerin kaldırılması ve kök kelime çıkarma işlemleri gerçekleştirilebilir. Şekil 3.1 veri ön işleme adımlarını göstermektedir.

Şekil 3. 1. Veri ön işleme adımları

3.2. Makine Öğrenimi Algoritmaları

Makine Öğrenimi (ML); bilim, bilgisayar bilimi, matematik ve diğer birçok alan ve disiplinden gelen geliştirme fikirlerine dayanan disiplinlerarası bir alandır (Cervantes vd., 2020). Makine öğrenmesi yöntemleri denetimli, yarı denetimli ve denetimsiz olmak üzere üç ana kategoriye ayrılmaktadır (Khan vd., 2016). Bu çalışmada denetimli makine öğrenmesi ve topluluk öğrenme algoritmalarını kullandık.

Denetimli öğrenme, algoritmaya bilinen girdilerin ve ilgili sonuçların sağlanmasıyla gerçekleştirilir. Algoritmanın amacı bu ilişkiyi öğrenmek ve yeni girdi hakkında doğru tahminlerde bulunmaktır. Bu yöntem, model oluşturma sürecine birçok kişinin katılımını gerektirir ancak sonuçta yinelemeli süreci hızlı bir şekilde tamamlayabilir (Bansal vd., 2022).

Topluluk öğrenimi, birden fazla makine öğrenimi algoritmasını entegre etme genel yeteneğine sahip, sağlam, yüksek performanslı bir model oluşturmayı amaçlamaktadır.

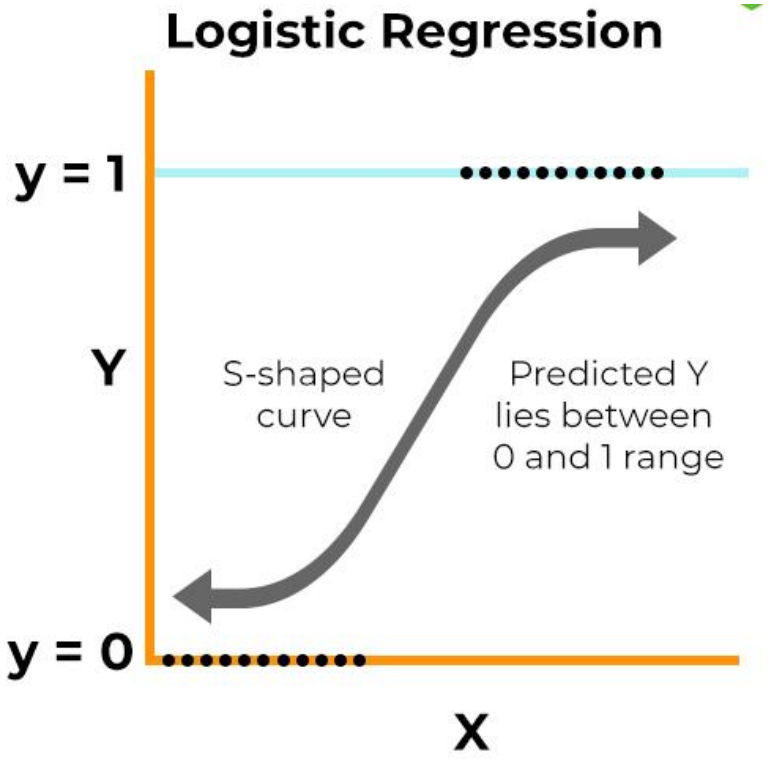
Kanakaraj vd., (2015), makine öğrenimindeki topluluk yöntemlerinin, tek başına çalışan bir algoritmaya göre daha fazla tahmin gücü elde etmek için birden fazla algoritmanın belirli bir sorun üzerindeki etkilerini birleştirdiğini belirtmişlerdir. Bu yöntem genel olarak güvenirlik ve geçerlilik açısından tek bir modele ihtiyaç duymaktadır. Bu öğrenme üç yöntemle uygulanır: Torbalama, artırma ve istifleme.

Aşağıda bu çalışmada kullanılan makine öğrenmesi algoritmaları ve özellikleri verilmektedir.

3.2.1. Regresyon

Regresyon, makine öğrenmesinde en yaygın kullanılan ve en iyi anlaşılan istatistiksel yöntemlerden biridir. İstatistiksel açıdan bakıldığında regresyon, bağımlı değişkenler ile bağımsız değişkenler arasındaki ilişkiyi göstermek ve analiz etmek için kullanılan bir tekniktir (Tyagi vd., 2022). Bu teknikte veri noktalarının, verinin dağılımına bağlı olarak bir eğri veya bir çizgi uygulanır. Amaç buradaki veri noktalarının ve çizilecek eğri/çizgi arasındaki farkın minimum düzeyde olmasıdır.

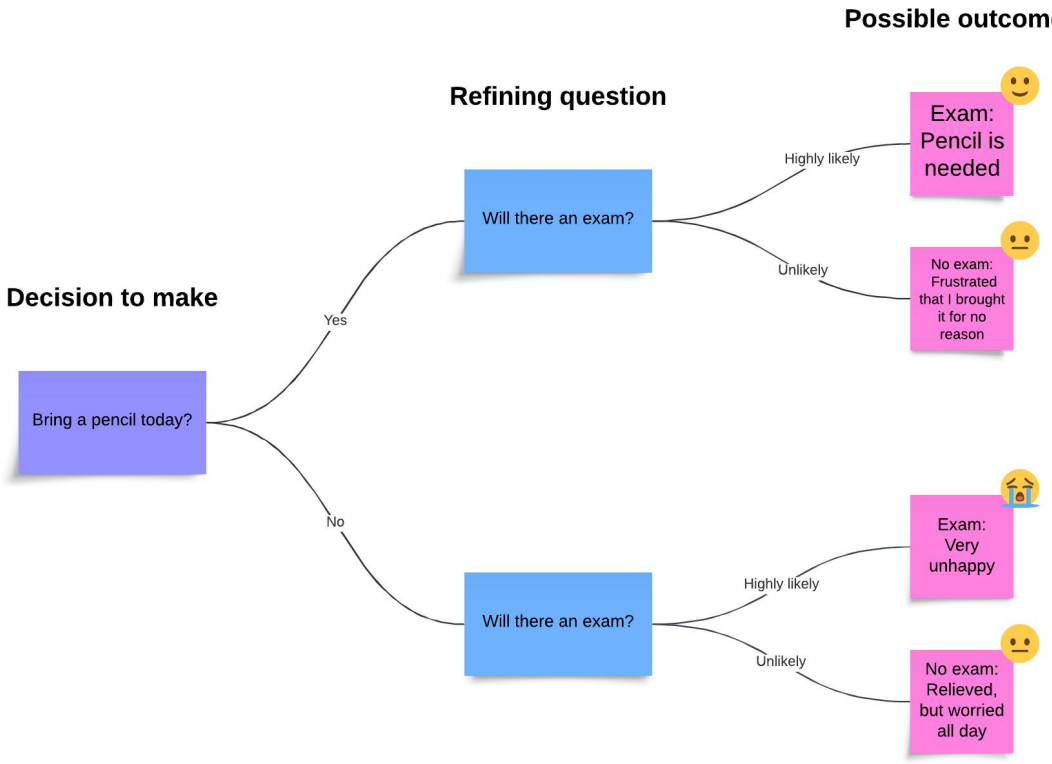
Lojistik regresyon, regresyon tekniklerinden biridir. Mevcut verilere dayanarak bir olayın meydana gelme olasılığının tahmin edilmesidir. Y değişkeninin iki kategori (örneğin 0 ve 1) arasında nasıl dağıldığını tahmin etmek için kullanılır; burada X bağımsız bir değişkendir ve Y bağımlı değişkendir. Şekil 3.2, lojistik regresyonun uygulanmasına yönelik temel varsayımları göstermektedir (“Everything You Need to Know About Logistic Regression - Spiceworks,” tarihsiz).



Şekil 3. 2. Lojistik regresyonun uygulanmasına ilişkin temel varsayımlar ("Everything You Need to Know About Logistic Regression - Spiceworks," tarihsiz)

3.2.2. Karar ağacı (DT)

Bir ağaç yapısında yapraklar sınıf etiketlerini gösterirken, yapraklara ulaşan dallar ilgili özelliklerin değerlendirilmesini temsil eder. Karar ağaçları tepeden aşağıya doğru bir yapıya sahipler. Bu yapının oluşturulmasındaki temel amaç, büyük veri setlerini çeşitli karar kuralları bölümlendirmesine tabi tutarak daha küçük kümelere tabi tutmaktır. Karar ağacı modelleri çok büyük veri setlerine uygulanabilmesi, güvenilirliği birçok istatistiksel testle mümkündür. Şekil 3.3 karar ağacı algoritmasının mantığını göstermektedir.



Şekil 3. 3. Karar ağacı örneği

3.2.3. K-en yakın komşular (KNN)

KNN, nesneleri sınıflandırmak için kullanılan bir yöntemdir, bu yöntemde nesne, etrafındaki en yakın öğrenme verilerine göre sınıflandırılır (Lubis vd., 2020). Hem sınıflandırma hem de regresyon problemlerinde kullanılan basit ama çok güçlü bir yöntemdir. Belirtildiği gibi bu model, bir veri noktası ile diğer tüm eğitim veri noktaları arasındaki mesafenin hesaplanması, KNN'sinin bulunması, komşuların sınıflarının incelenmesi ve yeni veri noktasının sınıfı olarak çoğunluk sınıfının belirlenmesi prensibiyle çalışmaktadır.

KNN sınıflandırıcısı uygun bir metrik olmayan bir yakınlık ölçüsü içerebilir; Ancak temel KNN algoritmasında performans optimizasyonları için uygun bir metriğin kullanılması gerekmektedir (Beygelzimer vd., 2006; Silpa-Anan & Hartley, 2008). Bu ölçümler Öklid, Manhattan veya Minkowski gibi seçenekleri içerir. KNN basit anlaşılırlığıyla öne çıkar ancak aykırı değerlere duyarlı olduğundan modelin performansı zayıf olabilir.

3.2.4. Naive bayes

Naive Bayes makine öğrenmesi algoritması özellikle NLP ve metin sınıflandırmada kullanılan, adını Bayes teoreminden ve saf varsayımından alan, oldukça etkili sonuçlar veren olasılık temelli bir algoritmadır. Bayes teoremi, bir olayın olasılığının, ilgili diğer olayların olasılıklarından yararlanılarak hesaplanmasıyla ilgilidir. Bu algoritmanın temel varsayımı, öznel bir değerin diğer nitelik değerlerine bağlı olmadığı, yani tüm konuların birbirinden bağımsız olduğudur. Naive Bayes, basitliği ve etkinliği ile öne çıkması nedeniyle belge sınıflandırma ve test süreçlerinde yaygın olarak tercih edilen bir yöntemdir (Kim vd., 2002).

GNB en basit sınıflandırma algoritmalarından biridir (Bishop, 2006). Bu algoritma, voksel katkılarının bağımsız olduğu ve Gauss dağılımını takip ettiği varsayımı altında her bir örneğin sınıf etiketi atamasını sağlar (Ontivero - Ortega vd., 2017). Özellikle özelliklerin sürekli olduğu ve Gauss (normal) dağılımı takip ettiğine inanılan sınıflandırma görevlerinde yaygın olarak kullanılan özel bir Naive Bayes algoritmasıdır.

3.2.5. Destek vektör makinesi (SVM)

SVM teorisinde veriler en fazla iki sınıfa aittir. Birbirinden doğrusal/doğrusal olmayan şekilde düzgün bir şekilde ayrılabilen sonsuz sayıda çizginin olduğunu varsayar.

SVC, SVM algoritmasının özel bir türüdür ve ikili sınıflandırma problemlerinde kullanılır. Algoritma, özellik uzayındaki farklı sınıfları etkili bir şekilde ayırabilen en verimli doğrusal karar sınırını (veya hiperdüzlemi) tanımlamak için özel olarak tasarlanmıştır. 'Doğrusal' kısım, sınıflandırıcının verileri düz bir çizgi (iki boyutta) veya düz bir hiperdüzlem (daha yüksek boyutlarda) kullanarak ayırmayı amaçlayan doğrusal bir çekirdek kullandığını belirtir.

3.2.6. Rastgele orman (RF)

RF, sınıflandırma ve regresyon gibi görevler için kullanılan bir topluluk öğrenme yöntemidir. Çoklu karar ağaçları, her ağacın veri setinin farklı bir alt kümesi üzerinde eğitilmesi prensibine göre çalışır. Bölünmüş veri setleri arasından en iyi karar noktasını

seçer. İsmin 'rastgele' kısmı, modelin her ağacı oluştururken veri alt kümelerini ve özellik alt kümelerini rastgele seçme şeklini ifade eder. Her düğümde bölünme noktası öznel olarak en iyi olana göre belirlenir.

Bu yöntem, veri setinin bölünmesi (alt kümelerin oluşturulması), bölünen alt kümelerle karar ağaçlarının eğitilmesi ve tahminlerin oluşturulması şeklinde bir çalışma prensibidir. Sınıflandırmada bu tahminler bağımsız ağaçların tahminleri arasından çoğunluk oyuna göre yapılır. Rastgele orman algoritması, karar ağaçlarıyla karşılaştırıldığında hata oranını daha doğru tahmin eder. Daha spesifik olarak, ağaç sayısı arttıkça hata oranının her zaman yakınsadığı matematiksel olarak kanıtlanmıştır (Breiman, 2001). Bu yöntemle yüksek doğruluk oranına ulaşmak mümkündür.

3.2.7. Ekstra ağaç (ET)

Ekstra ağaçlar, veri kümesinin farklı alt örneklerine uymaya çalışan ve süreçteki veri uyumunu ve doğruluğunu iyileştirmek için ortalama alma kavramını kullanan bir tahmin aracıdır (Chauhan vd., 2023). Karar ağaçlarına dayalı bir tür topluluk öğrenme yöntemidir. Bu algoritma RF algoritmasına benzer ancak daha fazla rastgeleliğe sahip ek yöntemler kullanır. Bu bakımdan bu algoritma problemler ve veri setleri için daha kullanışlı olabilir. Bu model birden fazla karar ağacından oluşmaktadır. Bu ağaçların her biri veri seti üzerinde farklı bir model oluşturur ve bu set bölündüğünde bazı koşullara göre bölünür. Rastgelelik özelliği sayesinde karar ağaçları çeşitliliğe sahiptir. Ayrıca bu rastgelelik sayesinde aşırı öğrenmenin önüne geçilir.

Her karar ağacı tüm veri seti kullanılarak eğitilir. Her düğümde bir niteliğin bir alt kümesi rastgele seçilir. Ayırma noktası aynı şekilde seçilen bir nitelik üzerinde seçilir. En iyisini aramak yerine rastgele seçilir. En iyi bölünme noktası, rastgele seçilen bölünme noktalarına belirli kriterlerin uygulanmasıyla seçilir. Bu kriterler Gini safsızlığı veya entropisi olarak verilebilir. Son olarak eğitilmiş birçok karar ağacı bir araya getirilerek sınıflandırma problemine yönelik oylar sayılır ve en çok oyu alan sınıf seçilir.

Ekstra ağaç büyük veya küçük veri setleri fark etmeksizin tüm veri setleri üzerinde çalışabilen, yüksek hızlı bir algoritmadır.

3.2.8. Gradyan artırıcı sınıflandırıcı (GBC)

GBC güçlü bir topluluk öğrenme yöntemidir. Bu algoritmalar zayıf öğrenenleri (yani rastgele öğrenenlerden biraz daha iyi olanları) sürekli olarak güçlü öğrenenlere dönüştürür (Freund vd., 1999). Özellikle regresyon problemleri için benzer bir güçlendirme algoritmasıdır (Friedman, 2001). Her yeni model, önceki modellerin hatalarını düzeltmeye çalışır. Bu işlem gradyan iniş yöntemiyle optimizasyon yapılarak gerçekleştirilir. GB'nin amacı, doğru tahminler yapmak için bir grup zayıf öğrenciyi sırayla birleştirerek güçlü bir öğrenci yaratmaktır. Bu algoritmanın birkaç çalışma prensibi vardır. Aşağıda bu ilkeler verilmektedir.

- ***İlk model:*** Öncelikle basit bir modelle (genellikle bir ortalama değer) veriyi tahmin etmekle başlar.
- ***Artımlı modelleme:*** Her ardışık model, bir önceki modelin hatalarını tahmin etmeye çalışır. Bu hatalar (artıklar) bir sonraki modelin hedefi haline gelir.
- ***Ağırlıklandırma:*** Her modelin katkısı, gradyan iniş yöntemiyle belirlenen bir ağırlıkla çarpılarak toplam model oluşturulur.
- ***Toplam tahmin:*** Sonunda, tüm zayıf öğrencilerin tahminlerinin ağırlıklı toplamı nihai tahmin olarak kullanılır.

Model, kayıp fonksiyonunu en aza indirecek şekilde sırayla oluşturulmuştur. Kayıp fonksiyonu genellikle hatanın karesi, lojistik kayıp veya başka bir uygun kayıp fonksiyonu olabilir. Her yeni model, bir önceki modelin hatalarını düzeltmek için gradyan yönünde bir güncelleme yapar.

3.2.9. Yığınlama

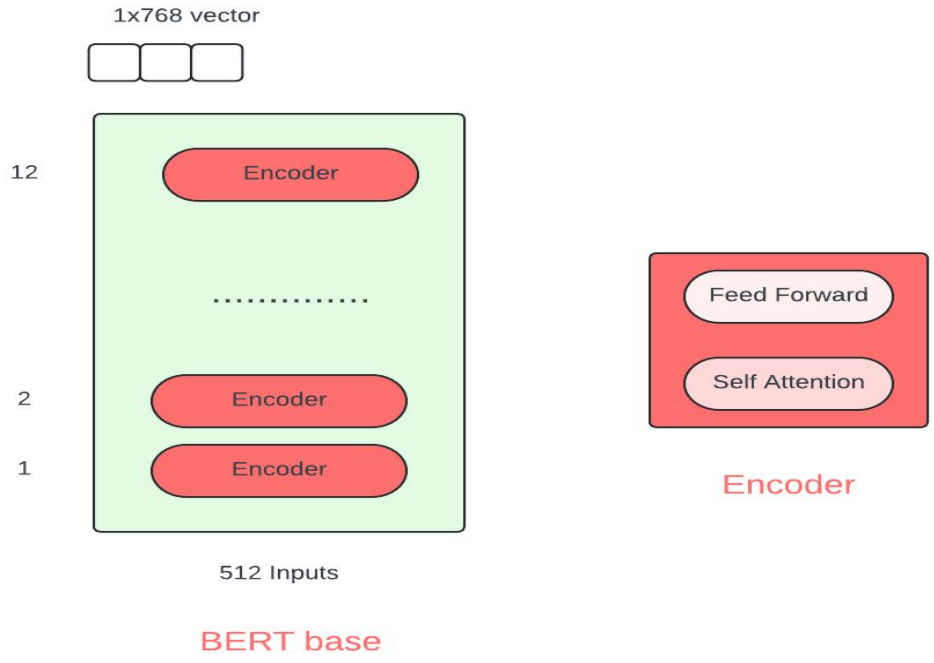
Yığınlama, birden fazla farklı modelin çıktısını birleştirir ve nihai tahmini yapmak için bu çıktıları ikinci bir modele (meta-öğrenici) besler. Birinci aşama modeller çeşitli algoritmalardan oluşur ve ikinci aşamada bu modellerin çıktıları bir model ile birleştirilir. Çalışmamızda ilk aşamada temel modeller RF ve Doğrusal SVC'dir. Daha sonra bu modellerin tahminleri Lojistik Regresyon modeliyle birleştirildi.

3.3. Transformatörlerden Çift Yönlü Kodlayıcı Gösterimleri (BERT)

Devlin vd. (2019), BERT'in Vaswani vd. (2017) tarafından sunulan transformatör mimarisine sahip önceden eğitilmiş bir dil modeli olduğunu belirtmişlerdir. Bu model, ince ayarlı bir şekilde alt NLP görevleriyle kolayca uygulanabilecek şekilde tasarlanmıştır (Zahera vd., 2019). Çoğu geleneksel dil modeli, metni yalnızca soldan sağa veya sağdan sola okur. Bu, modelin bağlamı tam olarak anlamasını zorlaştırır. Örneğin "yaz" kelimesi hem mevsim hem de yazmak fiili anlamına gelebilir. Ancak "yaz" kelimesinin etrafındaki diğer kelimelere bakarak doğru anlamı çıkarmak daha etkilidir. Bu modelin en önemli özelliği metinleri çift yönlü öğrenmesidir. Çift yönlü öğrenme şu şekilde açıklanabilir:

- Sözcüğün önündeki sözcüğü alarak bağlamını öğrenmek
- Sözcüğün bağlamını kendisinden sonraki sözcüğü dikkate alarak öğrenmek

BERT, Transformer mimarisini temel alır. Transformer, dikkat mekanizmasını kullanarak metin içindeki bağımlılıkları öğrenir ve bu sayede metni anlamada yüksek performans sağlar. Transformatörün daha önce de belirtildiği ve açıklandığı gibi iki ana bileşeni vardır: kodlayıcı ve kod çözücü. BERT yalnızca kodlayıcı kısmını kullanır. Şekil 3.4 BERT modelinin yapısını göstermektedir.



Şekil 3. 4. BERT modelinin yapısı

BERT, ön eğitim ve ince ayar olmak üzere iki aşamalı bir eğitim sürecinden geçer. Aşağıda bu süreçler anlatılmaktadır.

3.3.1. Ön eğitim

Ön eğitim aşamasında model büyük miktarda metin verisi üzerinde eğitilir. Bu aşamada iki ana görev kullanılır:

- ***Maskeli dil modelleme (MLM):*** Rastgele seçilen bazı kelimeler maskelenir ve modelin bu kelimeleri tahmin etmesi beklenir. Örneğin, "[MASK] kitaplarını okumaktan keyif alıyorum" cümlesinde model, "[MASK]" ifadesinin yerine hangi kelimenin geçeceğini tahmin eder. Çalışmamızda MLM tekniği ile oluşturulan ve birçok dil desteği sunan "bert-based-multilingual-uncased" model ile çalıştık. Bu model İngilizce, Türkçe, Almanca, Fransızca, İspanyolca, Çince, Japonca gibi 100'den fazla dilde büyük metin verileri üzerinde eğitilmektedir. Modelin en önemli özelliklerinden biri de bir dilde öğrenilenlerin diğer dile aktarılabilmesi, böylece bilgi paylaşımını mümkün kılmasıdır. Model küçük harflere karşı duyarsızdır; bu, modeli eğitirken bir ön hazırlık olarak tüm metnin küçük harflere dönüştürüldüğü anlamına gelir. Modelde 12 transformatör katmanı vardır ve her katmanda 768 gizli birim bulunur; bu, modelin her kelime için 768 boyutlu bir vektör temsili oluşturduğu anlamına gelir. Ayrıca güçlü ve zengin dil temsilleri oluşturmasına olanak tanıyan 110 milyon parametreye sahiptir.
- ***Sonraki cümle tahmini (NSP):*** Bu görevde modele iki cümle verilir ve bu cümlelerin birbirini takip edip etmediğini tahmin etmesi istenir.

3.3.2. İnce ayar

İnce ayar, BERT modeli eğitiminin ikinci aşamasıdır. Önceden eğitilmiş bir BERT modelinin NLP görevlerinden birini gerçekleştirebilmesi için daha az miktarda veri içeren bir veri seti üzerinde bazı ayarlamalar yapılır. Bunu yapmanın amacı modeli belirli bir görevi gerçekleştirebilecek şekilde hazırlamaktır. Şekil 3.5 ince ayar sürecini göstermektedir.

Şekil 3. 5. İnce ayar süreci

3.4. Değerlendirme Metrikleri

Değerlendirme metrikleri, makine öğrenimi modelinin kalitesini ve doğruluğunu hesaplamak için kullanılır. Bu modelleri değerlendirmek ve bunu birden fazla metrikle yapmak çok önemlidir çünkü herhangi bir model herhangi bir metrik değerlendirme ölçüsüyle iyi performans gösterirken, tam tersi başka bir metrik kullanıldığında kötü performans gösterecektir. Bir modelin en iyi şekilde doğru çalışması istiyorsa, bu değerlendirme kriterlerinin kullanmanısı çok önemlidir. Tablo 3.1 ikili sınıflandırma için karışıklık matrisini göstermektedir.

Tablo 3. 1. İkili sınıflandırma için karışıklık matrisi

Asıl/Tahmin	**Negatif**	**Pozitif**
Negatif	Doğru Negatif (TN)	Yanlış Pozitif (FP)
Pozitif	Yanlış Negatif (FN)	Doğru Pozitif (TP)

3.4.1. Doğruluk (Accuracy)

Doğruluk, doğru sınıflandırmanın toplamıdır. Bu, fiili sayının toplama bölünmesine eşit olduğu anlamına gelir. Doğru pozitif ve negatiflerin toplama oranı da denebilir. Doğruluk değerinin hesaplamasında kullanılan denklem 3.1'de gösterilmektedir (Sokolova & Lapalme, 2009).

$$Accuracy\ (acc) = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.1}$$

3.4.2. Kesinlik (Precision)

Kesinlik, tüm sınıflardan ne kadarının doğru tahmin edildiğinin bir ölçüsüdür. Bu, ne kadar yüksek olursa o kadar iyi olduğu anlamına gelir. Pozitif tahmin değeri olarak da bilinir. Kesinlik değerinin hesaplamasında kullanılan denklem 3.2'de gösterilmektedir (Bello vd., 2023).

$$Precision\ (p) = \frac{TP}{TP + FP} \tag{3.2}$$

3.4.3. Duyarlılık (Recall)

Duyarlılık, sınıflandırıcının gerçek pozitif değeri ne kadar doğru tahmin ettiğinin bir ölçüsüdür. Geri çağırma olarak da bilinir. Mümkün olduğu kadar yüksek olmalıdır. Duyarlılık değerinin hesaplamasında kullanılan denklem 3.3'te gösterilmektedir (Bello vd., 2023).

$$Recall\ (r) = \frac{TP}{TP + FN} \tag{3.3}$$

3.4.4. F1 puanı (F1 Score)

F1 puanı, geri çağırma ve kesinlik değerlerinin harmonik ortalamasıdır. Sınıflandırıcı performansını karşılaştırmak için kullanılır. F1 puanı değerinin hesaplamasında kullanılan denklem 3.4'te gösterilmektedir (Bello vd., 2023).

$$F1 - score = 2 * \frac{p * r}{p + r} \tag{3.4}$$

3.4.5. Ortalama yöntemleri

Ortalama alma yöntemi, sınıflandırma veya nesne algılama gibi birçok sınıfı kapsayan iş akışlarının, tüm sınıfların model performansını kesinlik, duyarlılık ve F1 puanı gibi sınıf başına vermek yerine tek bir değer olarak vermesine olanak tanır. Bunu yaparken sınıf başına puanların toplanması ve ortalamasının alınması gerekir. Birkaç ortalama alma yöntemi vardır. Aşağıda bu yöntemler verilmiş ve açıklanmıştır.

3.4.5.1. Makro ortalama

Makro ortalama, tüm sınıflardan elde edilen sonuçların ortalamasına eşittir. Kesinlik, hatırlama ve f1 puanının makro ortalaması sırasıyla 3.5 (Grandini vd., 2020), 3.6 (Grandini vd., 2020) ve 3.7 (Grandini vd., 2020) denklemleriyle verilmiştir.

$$Macro\ Average\ Presicion = \frac{1}{K}\sum_{k=1}^{K} Precision\ (k) \tag{3.5}$$

$$Macro\ Average\ Recall = \frac{1}{K}\sum_{k=1}^{K} Recall\ (k) \tag{3.6}$$

$$Macro\ Average\ F1 - score = 2 * \frac{Macro\ Average\ Precision * Macro\ Average\ Recall}{Macro\ Average\ Precision^{-1} + Macro\ Average\ Recall^{-1}} \tag{3.7}$$

k sınıftır ve değerlerin 1'den K'ya kadar olduğu ve her sayının farklı bir sınıfı temsil ettiği varsayılır.

3.4.5.2. Mikro ortalama

Mikro ortalama almanın amacı, sınıflar arasındaki olası farklılıkları dikkate almadan tüm birimleri bir arada ele almaktır (Grandini vd., 2020). Kesinlik, hatırlama ve f1 puanının mikro ortalaması sırasıyla 3.8 (Grandini vd., 2020), 3.9 (Grandini vd., 2020) ve 3.10 (Grandini vd., 2020) denklemleriyle verilmiştir.

$$Micro\ Average\ Presicion = \frac{\sum_{k=1}^{K} TP\ (k)}{\sum_{k=1}^{K} TotalColumn_k} = \frac{\sum_{k=1}^{K} TP\ (k)}{GrandTotal} \tag{3.8}$$

$$Micro\ Average\ Recall = \frac{\sum_{k=1}^{K} TP\ (k)}{\sum_{k=1}^{K} TotalRow_k} = \frac{\sum_{k=1}^{K} TP\ (k)}{GrandTotal} \quad (3.9)$$

$$Micro\ Average\ F1 - score = \frac{\sum_{k=1}^{K} TP\ (k)}{GrandTotal} \quad (3.10)$$

k sınıftır ve değerlerin 1'den K'ya kadar olduğu ve her sayının farklı bir sınıfı temsil ettiği varsayılır.

3.4.5.3. Ağırlıklı ortalama

Ağırlıklı ortalama, sınıf sayılarının farklı olduğu durumlarda sınıfların büyüklüklerine göre ağırlıklandırılarak ortalamanın alınmasıdır. Yani her sınıfın performansı o sınıfa ait örnek sayısıyla çarpılıp tüm sınıflara ait örnek sayısının toplamına bölünerek hesaplanır.

3.5. Aşırı Öğrenme

Makine öğrenimi ve istatistikte aşırı öğrenme, bir modelin eğitim verilerine gereğinden fazla uyum sağladığı ve dolayısıyla yeni, görünmeyen verilerde düşük performans gösterdiği bir durumu ifade eder. Başka bir deyişle, model eğitim verilerini o kadar iyi öğrenir ki, bu verilere gürültü ve hatta rastgele desenler katarak genelleme yeteneğini azaltır. Şekil 3.6 modeller için ideal ve aşırı öğrenme grafikleri göstermektedir.

Şekil 3. 6. İdeal ve aşırı öğrenme grafikleri (What Is Overfitting in Machine Learning?, 2023)

Aşırı öğrenmenin bazı göstergeleri vardır. Bunlardan biri eğitim performansının yüksek ancak test performansının zayıf olmasıdır. Bu, modelin eğitim verileri üzerinde çok iyi performans gösterdiği ancak test (veya doğrulama) sırasında veriler üzerinde performansın daha düşük olacağı anlamına gelir. Diğeri karmaşıklıktır. Modelin çok fazla parametresi olabilir veya yapısı çok karmaşık olabilir.

Aşırı öğrenmenin önlenmesi için bazı yöntemler vardır. Bu çalışmada çapraz doğrulama yöntemi kullanılmıştır. Aşağıda bu yöntem verilmiş ve türleri açıklanmıştır.

3.5.1. Çapraz doğrulama

Çapraz doğrulama, bir modelin genelleme yeteneğini değerlendirmek için kullanılan bir yöntemdir. Bu yöntem, özellikle veri kümesinin boyutu küçük olduğunda, modelin aşırı öğrnmesini önlemeye yardımcı olur.

Bu, modelin eğitim verilerinin her detayını öğrenmesine olanak tanır. Bu çalışmada, bu durumu önlemek için bu yöntemin uygulandığı bazı deneyler yapılmıştır. Çapraz doğrulamanın bazı temel adımları vardır. Aşağıda bu adımlar verilmiştir.

- ***Veri setini parçalara bölme:*** Veri seti genellikle k eşit parçaya (katlama) bölünür. Bu parçalar arasındaki ilişkiyi kırmak için veriler rastgele karıştırılabilir.
- ***Modelin eğitimi ve doğrulanması:*** Her yinelemede modeli eğitmek için k-1 parça kullanılır. Geriye kalan 1 parça ise modelin performansını test etmek için kullanılır. Bu işlem k kez tekrarlanır ve her parça bir kez test verisi olarak kullanılır.
- ***Performansın değerlendirilmesi:*** Tüm yinelemelerin sonuçları birleştirilir ve ortalama alınarak modelin genel performansı değerlendirilir.

3.5.1.1. K-katlı çapraz doğrulama

Bu yöntemle veri seti k eşit parçaya (katlama) bölünür. Ok yinelemesi boyunca, test seti olarak her seferinde bir kat kullanılırken, geri kalan k-1 kat, modeli eğitmek için kullanılır. Tüm yinelemeler tamamlandığında modelin genel performansı bu test sonuçlarının ortalaması alınarak belirlenir. Ok değeri genellikle 5 veya 10 olarak seçilir. Bu değerler genellikle dengeyi sağlamak için kullanılır. Daha yüksek k değeri, modelin daha genel performansını test eder ancak daha uzun hesaplama süresi gerektirir.

3.5.1.2. Birini dışarıda bırakma çapraz doğrulaması (LOOCV)

Bu yöntemle veri setindeki her örnek sırayla test seti olarak ayrılır ve geri kalan tüm örnekler modelin eğitimi için kullanılır. Veri setinde n adet örnek varsa bu işlem n defa tekrarlanır. LOOCV, her numune test edildiğinden veri seti küçük olduğunda kullanılır. Hesaplama maliyeti yüksektir çünkü veri setindeki örnekler kadar çok sayıda model eğitilip test edilmektedir.

3.5.1.3. Katmanlı k-katlı çapraz doğrulama

Bu yöntem k-katlı çapraz doğrulamaya benzer ancak her katlamanın sınıf dağılımını koruyacak şekilde bölünmesini sağlar. Bu özellikle dengesiz veri kümelerinde kullanışlıdır.

Bu yöntem, sınıf dengesizliği olan veri setlerinde modelin daha adil değerlendirilmesini sağlar. Örneğin, bir sınıfın çok az örneği varsa, bu yöntem bu örneklerin her katta yeterince temsil edilmesini sağlar.

3.5.1.4. Zaman serisi çapraz doğrulaması

Zaman serisi verilerinde sıralama önemli olduğundan standart k-katlama yöntemi uygun değildir. Bu yöntem, zaman dizisini koruyarak modeli geçmiş verilerle eğitme ve gelecekteki verilerle test etme yaklaşımını benimser.

Bu yöntem özellikle zaman serisi tahmin modellerinde kullanılmaktadır. Geçmiş verilere dayanarak tahminler yaparak modelin performansını değerlendirmeye olanak sağlar.

3.5.1.5. Tekrarlanan k-katlı çapraz doğrulama

Bu yöntemle k-katlı çapraz doğrulama birden çok kez tekrarlanır. Her yinelemede veri seti yeniden karıştırılır ve katlamalar yeniden oluşturulur.

Bu yöntem, modelin rastgele değişimlere karşı ne kadar tutarlı olduğunu test etmek için kullanılır. Performans ortalaması daha istikrarlı ve güvenilir hale gelir.

3.6. Aktivasyon Fonksiyonu

Aktivasyon fonksiyonu yapay sinir ağlarında (YSA) kullanılan temel bileşenlerden biridir. Aktivasyon fonksiyonları sinir ağındaki her nöronda kullanılır. Nöronların girdilerini işleyerek belirli bir çıktı üretmesini sağlayan matematiksel işlevlerdir. Nöronların doğrusal olmayan dönüşümler gerçekleştirmesini sağlayarak sinir ağlarının karmaşık veri modellerini öğrenmesini mümkün kılar. BERT modelinin özellikle gizli katmanlarında ve çıktı katmanında yer almaktadır.

Aktivasyon fonksiyonları, sinir ağlarına doğrusal olmama özelliği sağlayarak karmaşık veri yapılarının ve ilişkilerin öğrenilmesine olanak tanır. Sinir ağı yalnızca doğrusal fonksiyonlar kullanılarak çalışsaydı, bu ağın öğrenebileceği ilişkiler son derece sınırlı olurdu ve derin yapısal özellikleri çıkaramazdı. Aktivasyon fonksiyonları sayesinde nöronlar girdileri farklı şekillerde işleyebilir ve çok katmanlı ağlar daha karmaşık fonksiyonları öğrenebilir.

Çalışma prensibi, nöronun aldığı girdileri belirli ağırlıklarla çarparak bu değerlerin toplamını hesaplamasıdır. Bu toplam aktivasyon fonksiyonuna verilir ve nöronun çıkışı elde edilir. Aktivasyon fonksiyonları girdiyi işler ve daha sonra diğer nöronlara veya nihai sonuçlara katkıda bulunan bir çıktı üretir.

Farklı görevler ve veri yapıları için çeşitli aktivasyon fonksiyonları kullanılır. Aşağıda bu tez çalışmasında kullanılan aktivasyon fonksiyonları verilmiştir.

3.6.1. Düzeltilmiş doğrusal birim işlevi (ReLU)

ReLU, yapay sinir ağlarında ve derin öğrenmede yaygın olarak kullanılan bir fonksiyondur. Bu fonksiyon ile girişler doğrusal hale getirilir, pozitifler olduğu gibi bırakılır, negatifler ise sıfıra eşitlenir. Bu işlev hesaplama açısından hızlı ve verimlidir ve nöronların bir kısmını sıfıra eşitleyerek seyreklik sağlar. Ancak negatif girdilerin sıfıra eşitlenmesi nedeniyle nöronların ölmesine neden olabilmesi gibi bir dezavantaja sahiptir. ReLU özellikle NLP, ses işleme ve görüntü işlemede kullanılmaktadır. ReLU denklemi 3.11'de gösterilmiştir

$$ReLU\,(x) = \begin{cases} x & if\ x > 0 \\ 0 & if\ x < 0 \end{cases} \tag{3.11}$$

Burada x girdidir.

3.6.2. Softmax

Softmax, çok sınıflı problemlerde kullanılan, bir vektörün her elemanını 1'e toplayacak şekilde normalleştiren ve bunu bir olasılık dağılımına dönüştüren bir fonksiyondur. Bu fonksiyon genellikle derin öğrenme modellerinin son katmanında kullanılır. Her sınıf için ayrı ayrı tahmini olasılıkları verir. Bu sayede hangi sınıfa ait olduğu yüksek olasılıkla belirlenir. Bir z vektörünün zi elemanının softmax değeri denklem 3.12'de verilmiştir. (Softmax: RK⟶ RK)

$$Softmax\ (zi)\ = \frac{e^{zi}}{\sum_{j=1}^{K} e^{zj}} \quad \text{for i = 1, . . . , K and z = } (z_1, \ldots, z_K) \in R^K \quad (3.12)$$

3.7. Kayıp Fonksiyonu

Kayıp fonksiyonu, makine öğrenimi modellerinin doğruluğunu ölçen bir performans ölçümüdür. Model tahminlerinin gerçek değerlerle ne kadar tutarsız olduğunu tespit etmek için kullanılır. Kayıp fonksiyonu, model sonuçları ile hedef değerler arasındaki hatayı hesaplar ve bu hatayı en aza indirmek model öğrenmenin temel amacıdır. Bu fonksiyon BERT modelinin çıktı katmanında kullanılır.

Kayıp fonksiyonu birkaç temel nedenden dolayı kritiktir:

- **Öğrenme sürecine rehberlik etmek:** Modelin ağırlıkları, kayıp fonksiyonunu en aza indirecek şekilde optimize edilir. Bu işlem genellikle gradyan iniş gibi optimizasyon algoritmaları kullanılarak gerçekleştirilir.
- **Öğrenme kontrolü:** Model ağırlıkları, kayıp fonksiyonunu en aza indirecek şekilde optimize edilmiştir. Bu işlem genellikle gradyan iniş gibi optimizasyon algoritmaları kullanılarak gerçekleştirilir.
- **Modellerin performansının değerlendirilmesi:** Kayıp fonksiyonunun değeri, modelin tahminleri ne kadar doğru veya yanlış yaptığını gösterir.

Farklı problem türleri için çeşitli kayıp fonksiyonları kullanılır. Bu çalışmada kategorik çapraz entropi kaybı fonksiyonları kullanılmıştır.

Kategorik çapraz entropi, sınıflandırma problemlerinde, model tarafından tahmin edilen olasılıklar ile gerçek sınıf etiketleri arasındaki farkı hesaplamak için kullanılır. Bu kayıp fonksiyonu, modelin tahminlerinin doğruluğunu ölçer ve eğitim sırasında bu kaybı en aza indirmeye çalışır. Ayrıca one-hot kodlama ile kullanılır.

3.8. Optimizasyon Algoritması

Optimizasyon algoritması, bir amaç fonksiyonu için bulunabilecek en iyi sonucu elde etmeye yönelik bir yöntemdir. BERT modelinde eğitim sürecinin her adımında (yinelemesinde) optimizasyon algoritmaları kullanılır. Kayıp fonksiyonunun çıktısını en aza indirmek için modelin parametrelerini (ağırlıklar ve sapma) günceller. Derin öğrenmede bu algoritma parametreleri eğitmek için kullanılır. Bu algoritmanın kullanılmasındaki amaç, gerçek sonuçlar ile tahmin edilen sonuçlar arasındaki boşluğu kapatmak, yani hata oranını azaltmak ve en aza indirmektir.

Pek çok optimizasyon algoritması vardır. Bu çalışmada yalnızca Uyarlanabilir Moment Tahmini (ADAM) algoritması kullanılmıştır.

ADAM algoritması, gradyan iniş algoritmasının bir çeşididir. ADAM sayesinde bu algoritma geliştirilmiş ve hızlandırılmıştır. Bu şekilde optimizasyon daha kararlı hale gelir. Öğrenme hızını her algoritma için ayrı ayrı ayarlayan bu algoritma sayesinde öğrenme süreci hızlı ve stabil hale gelir.

Bu algoritma, ortalama gradyanlar ve gradyanların karesi olan 2 momenti kullanır. Bu şekilde parametreler güncellenir. Başlangıç momentumu ile parametrelerin yöneldiği yön tahmin edilir. İkinci momentum trendlerin büyüklüğüdür. Bu iki yönlü süreç ile süreç daha kısa ve verimli hale gelir ve çalışmanın eğitimi sırasında öğrenme süresi ayarlanarak performans çok daha fazla iyileştirilir.

4.BULGULAR VE SONUÇLAR

Bu bölümde makine öğrenmesi algoritmaları ve BERT modeli ile gerçekleştirilen Türkçe tweetlerden duygu analizi performans sonuçlarını değerlendirilmiştir. Veri setimiz halihazırda bir kullanıcı tarafından paylaşılmış ve etiketlenmiş hazır Türkçe tweetlerden oluşmaktadır.

4.1. Duygu Analizi Çalışması İçin Oluşturulan Ortam

Türkçe tweetlerden duygu analizi gibi çalışmaların yapılabilmesi için güçlü makinelere, fonksiyonel kütüphanelere ve GPU desteğine ihtiyaç duyulmaktadır. Önceden eğitilmiş modeller kullanılarak çalışmalarımız daha verimli hale getirilmiştir. Çalışmamız Python dili kullanılarak yapılmıştır. Son yıllarda NLP çalışmalarında etkili olan Keras ve Tensorflow gibi kütüphanelerden faydalanılmıştır. Duygu analizi gibi makine öğrenimi uygulamaları için Google Colab özellikle uygundur çünkü ücretsiz GPU ve TPU erişimi sağlar (Gupta ve diğerleri, 2017). Tablo 4.1, bu çalışmadaki modeller için kullanılan makinenin, çalışma ortamının ve kullanılan kütüphanelerin özelliklerini göstermektedir.

Tablo 4. 1. Çalışma ortamı, makine özellikleri ve kullanılan kütüphaneler

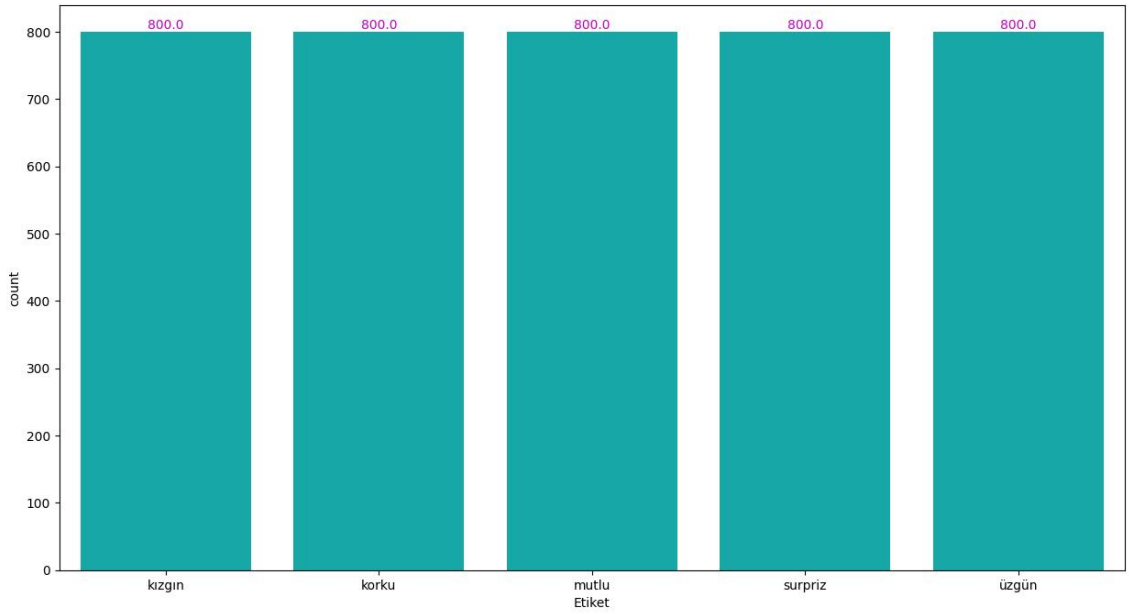
SİSTEM TİPİ	Windows 11 x64
İŞLEMCİ	13th Gen Intel(R) Core(TM) i7-13650HX, 2600 MHz, 14 Core
HAFIZA	16.0 GB
GRAFİK İŞLEM BİRİMİ (GPU)	NVIDIA GEFORCE RTX
PROGRAMLAMA DİLİ	Python
EDITÖR	Google Colab
KÜTÜPHANELER	Keras, Transformers, Numpy, Pandas, Matplotlib, Saeborn, Tensorflow, Sklearn

4.2. Veri Seti

Kullanılan veri seti, https://www.kaggle.com/ internet sitesi adresinden Anıl Güven'e ait TurkishTweets isimli veri setidir. Veri setinin tamamı Türkçe tweetlerden oluşmaktadır. Tweetler 5 etiketle ayrılmıştır. Bunlar sırasıyla 'kızgın', 'korku', 'mutlu', 'sürpriz' ve 'üzgün'dür. Tablo 4.2 veri setindeki örnek olarak seçilen 5 cümleyi ve bu 5 cümlenin etiketlerini gösterilmektedir. Veri setinde her etiket için 800 Tweet, toplamda 4000 veri (tweet) bulunmaktadır. Şekil 4.1 veri setindeki tweetlerin etiketlere göre dağılımını göstermektedir.

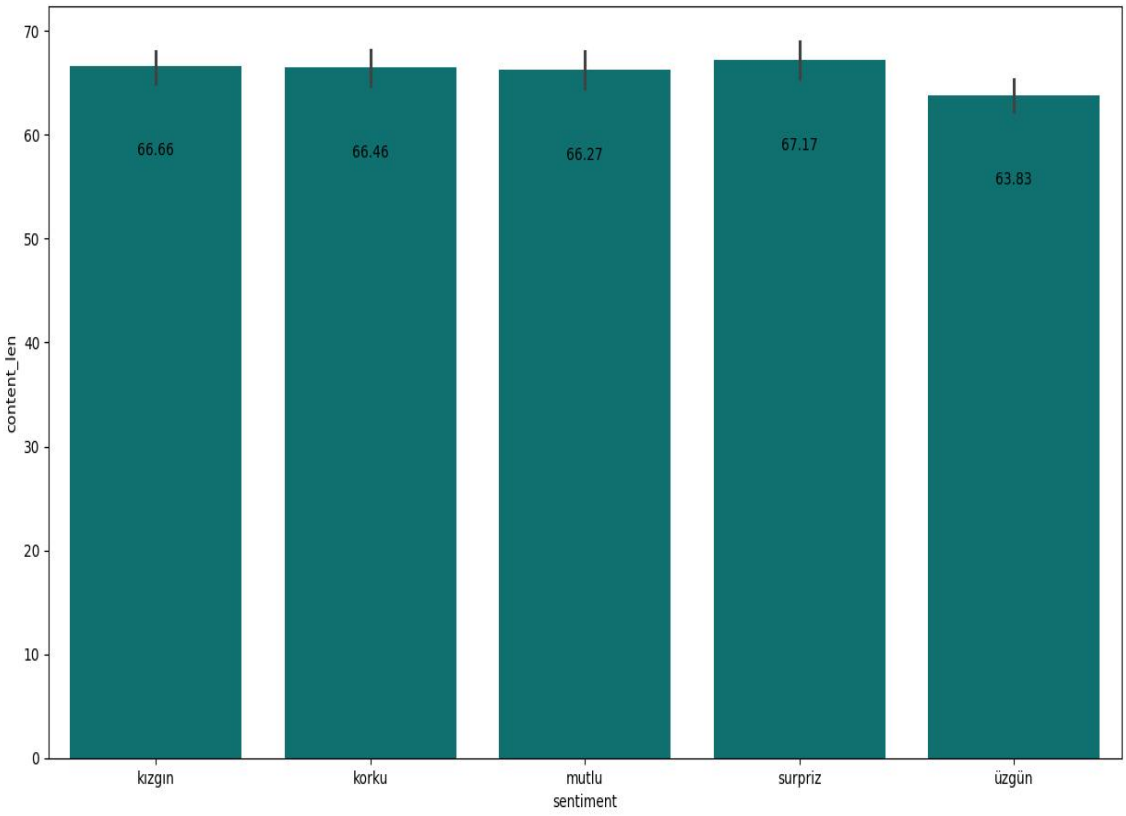
Tablo 4. 2. Türkçe tweet ve etiket örnekleri

Tweet	Etiket
Sebebi neydi ki diye bağıracağım şimdi az kaldı	kızgın
Çok korktum :((hayatım gözümün önünden film şeridi gibi geçti	korku
Şu andan itibaren doğum günüme 2 gün var. Niye bu kadar heyecanla bekliyorsam :D:D	mutlu
Pek bir sürpriz oldu hocam beklemiyorduk bu quizi	surpriz
Uykuyla dinlenemeyecek kadar yorgunum artık...	üzgün



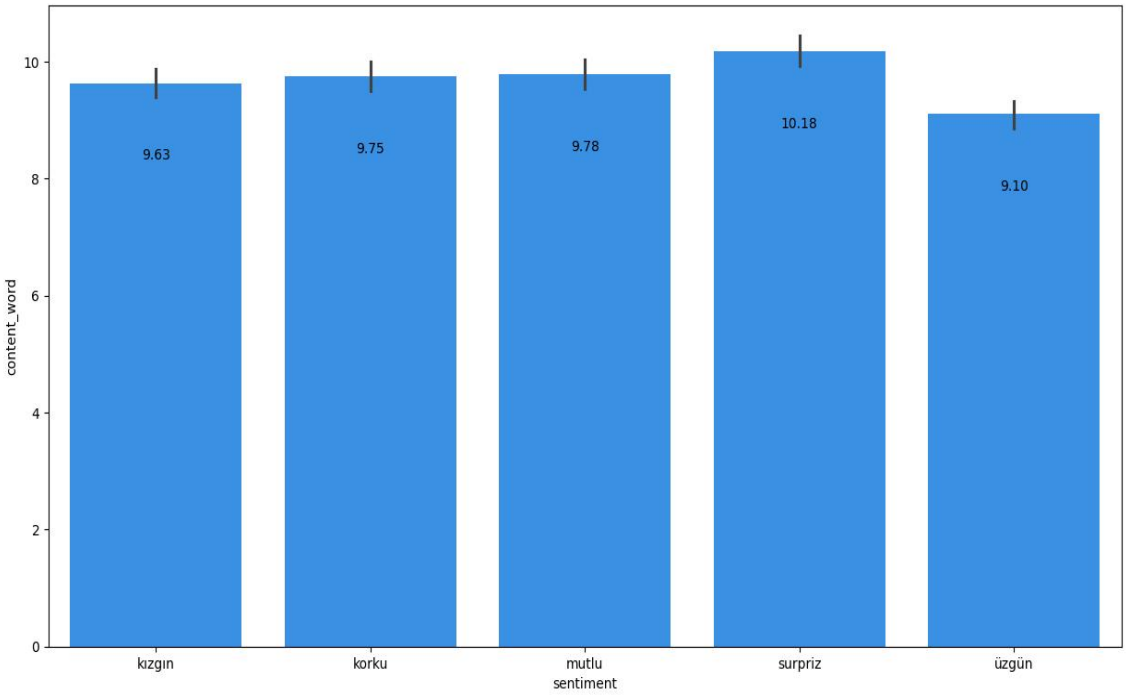
Şekil 4. 1. Veri setindeki tweetlerin etiketlere göre dağılımı

Veri setindeki sütunlarda yer alan her bir tweet metninin uzunluğu hesaplanmış ve bu uzunlukların ortalaması duygu etiketlerine göre analiz edilmiştir. Şekil 4.2 her bir duyguya ilişkin metin uzunluğunun ortalamasını bir çubuk grafik olarak göstermektedir.



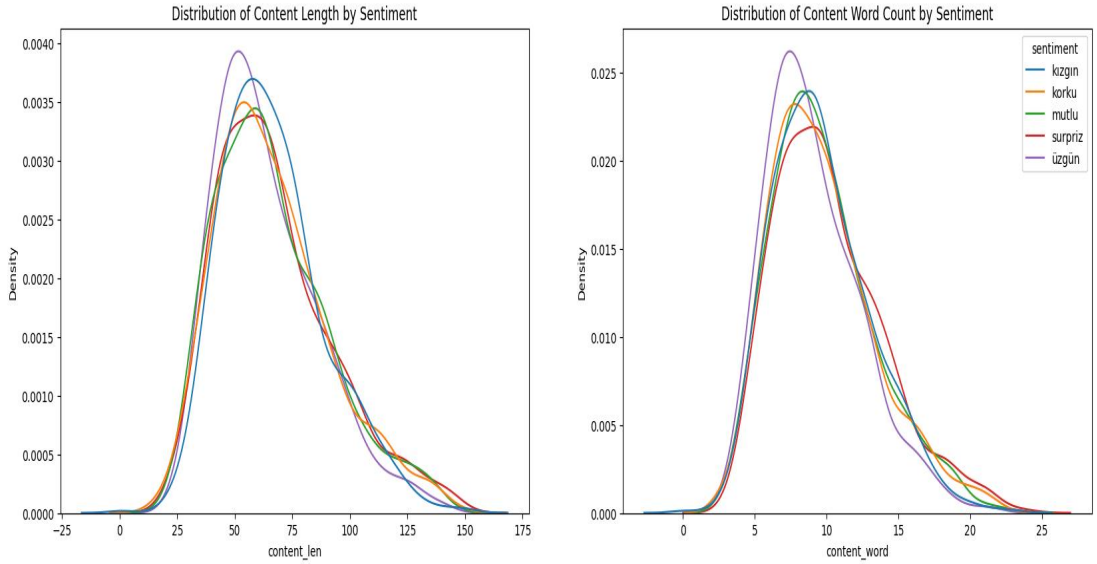
Şekil 4. 2. Her bir duyguya ilişkin metin uzunluğunun ortalamaları

Ayrıca metinlerin kelime sayıları hesaplanmış, bu kelime sayılarının duyguya göre ortalaması alınmıştır ve sonuçlar Şekil 4.3 gösterilmiştir. Bu sayede farklı duygu durumlarına sahip metinlerin ortalama kelime sayıları kolaylıkla karşılaştırılabilir.



Şekil 4. 3. Her duygu için kelime sayısı ortalamaları

Veri setindeki iki farklı özellik (içerik uzunluğu ve içerik kelime sayısı), duygusal duruma göre iki adet Çekirdek Yoğunluğu Tahmini (KDE) grafiği kullanılarak görselleştirilmiştir. Şekil 4.4'e göre, ilk grafik içerik uzunluğunun dağılımını, ikincisi ise içerik kelime sayısının duygu durumuna göre dağılımını göstermektedir. Bu grafikler, farklı duygusal durumların belirli veri özellikleri üzerindeki etkilerini analiz etmek için kullanılabilir.



Şekil 4. 4. İçerik uzunluğunun ve içerik kelimesine göre duygu dağılımı

4.3. Veri Ön Hazırlığı

Veri ön hazırlama, makine öğrenimi algoritmalarını uygulamadan önce atılması gereken çok önemli bir adımdır. Birincil amaç, veri kümesini temizlemek ve algoritmik uygulamalarda doğrudan kullanıma hazırlamaktır. Bu aşamada çeşitli ön işleme teknikleri uygulanabilir. Veri kümemde yürüttüğüm spesifik ön işlemler aşağıdaki gibidir:

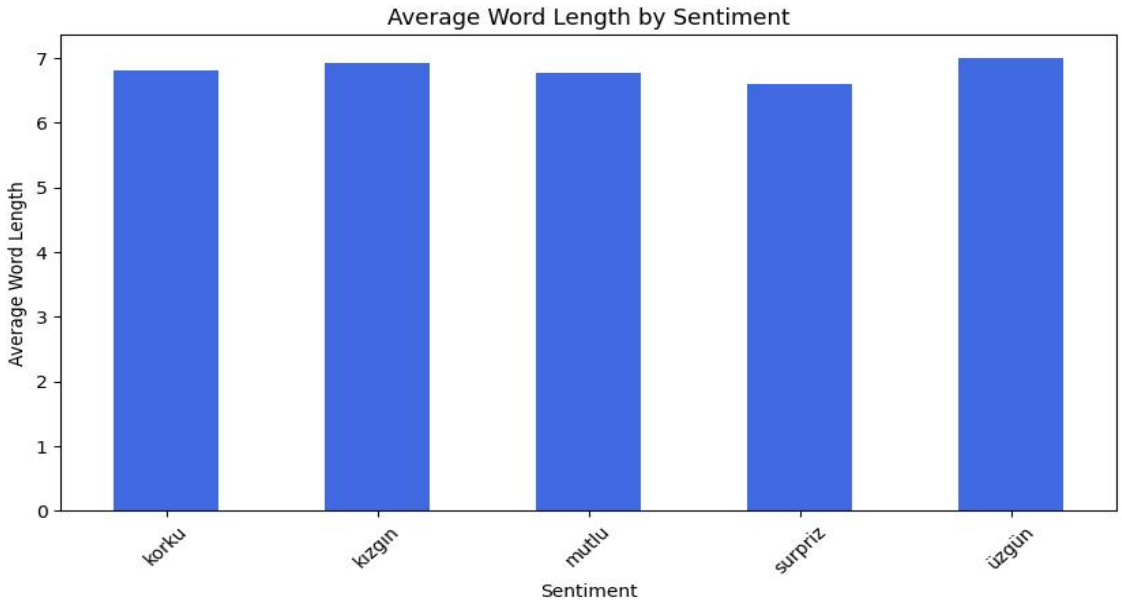
- Büyük harfle başlayan veya büyük harf içeren tüm kelimeleri küçük harfe dönüştürme.
- Tüm '@' ve köprü örneklerini kaldırma.
- Noktalama işaretlerini, sayıları ve emojileri kaldırma.
- 'NULL' içeren satırlar kaldırılıyor.
- 'Tweet' sütununun 'içerik', 'Etiket' sütununun ise 'duygu' olarak yeniden adlandırılması.

Tablo 4.3'te görüldüğü gibi ön işleme sonrasında veri işleme daha kolay hale gelmiştir. Veri setinde '@' veya 'http' yoktur. Veri setinde sadece 1 boş satır bulunmaktadır ve ön işleme tabi tutulduktan sonra bu satır silinmiştir.

Tablo 4. 3. Ön hazırlıktan önce ve sonraki metinler

Orijinal Metin	Ön Hazırlık Sonrası Metin	Duygu
Allah'ım çıldıracağım. Yemin ederim çıldıracağım sinirimden.	allahım çıldıracağım yemin ederim çıldıracağım sinirimden	kızgın
Çok korktum :((hayatım gözümün önünden film şeridi gibi geçti	çok korktum hayatım gözümün önünden film şeridi gibi geçti	korku
11 yıllık eğitim hayatımın bir 15 tatilini daha boşa geçirmiş bulunmaktayım kendime çok teşekkür ederim	yıllık eğitim hayatımın bir tatilini daha boşa geçirmiş bulunmaktayım kendime çok teşekkür ederim	mutlu
2 ay sonra yeniden evimdeyim. Sürpriz geldim bizimkiler nasıl şok oldu.	ay sonra yeniden evimdeyim sürpriz geldim bizimkiler nasıl şok oldu	surpriz
Bazı insanlar espiri yapmasın gün boyunca yoruldum ya..	bazı insanlar espiri yapmasın gün boyunca yoruldum ya	üzgün

Verileri duygu sütununa göre gruplandırarak her duygu kategorisi için toplam içerik uzunluğu ve toplam kelime sayısını hesaplanmıştır. Daha sonra bu değerler kullanılarak her bir duygu kategorisi için ortalama kelime uzunluğu hesaplanmıştır. Şekil 4.5 bu değerler gösterilmektedir.



Şekil 4. 5. Duygulara göre ortalama kelime uzunlukları

Farklı duygu kategorilerine ait metinler için kelime bulutları oluşturularak 2x4 grid düzeninde görselleştirilme yapılmıştır. Her alt grafik belirli bir duygu kategorisine karşılık gelir ve bu kategorinin metin içeriğine göre oluşturulan kelime bulutunu gösterir. Şekil 4.6 her duygu kategorisi için öne çıkan kelimeleri göstermektedir.

Şekil 4. 6. Her duygu için öne çıkan kelimeler

Ön işleme adımlarından biri, makine öğrenimi modellerini metin verileriyle eğitmek ve test etmektir. Bu nedenle veri seti ayrılmıştır. Daha sonra duygu etiketlerini sayısal verilere dönüştürülmüştür. Daha sonra veriler eğitim ve test setlerine ayrılmıştır. Aynı şekilde metin verileri de TF-IDF yöntemi kullanılarak sayısal vektörlere dönüştürülmüştür. Eğitim ve test veri seti oranları sırasıyla %80 ve %20'dir. Tablo 4.4'te her set için örnek sayısı ve bunların oranı gösterilmektedir.

Tablo 4. 4. Her set için örnek sayısı ve oranları

	Eğitim	Test	Toplam
Örnek Sayısı	3199	800	3999
Yüzdelik Oran (%)	80	20	100

Derin öğrenme modellerinden biri olan BERT'i kullanırken veri setini eğitim, doğrulama ve test setlerine ayırdık. Oranlar sırasıyla %65, %9 ve %26 civarındadır. Tablo 4.5 her set için örnek sayısını ve bunların oranını göstermektedir.

Tablo 4. 5. BERT modelinde her set için örnek sayısı ve oranları

	Eğitim	Doğrulama	Test	Toplam
Örnek Sayısı	2599	350	1050	3999
Yüzdelik Oran (%)	65	9	26	100

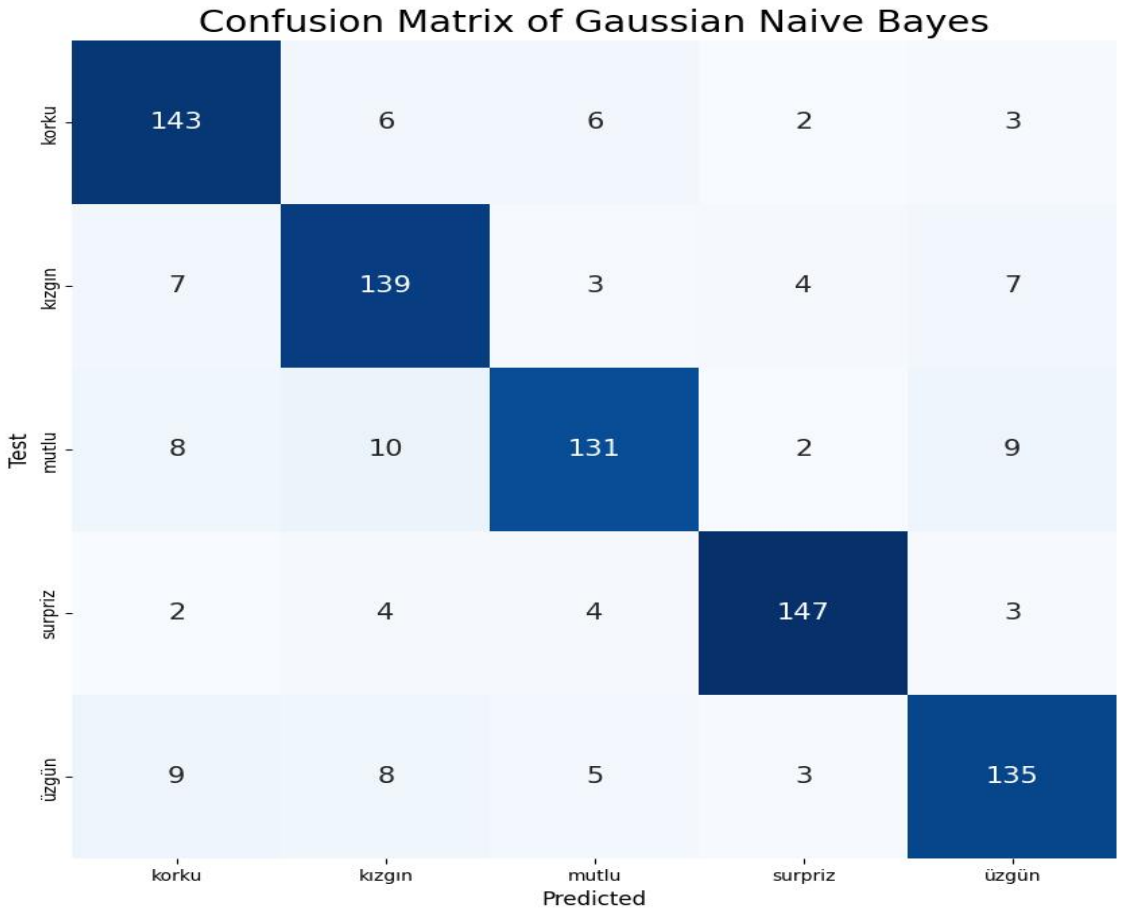
4.4. Çapraz Doğrulama Olmadan Makine Öğrenimi Algoritmalarının Sonuçları

Bu bölümde ön işleme tabi tutulup temizlenen ve daha sonra eğitim ve teste ayrılan veri seti üzerinde klasik makine öğrenmesi modellerinin çapraz doğrulama yapılmadan uygulanması sonucu elde edilen sonuçlar verilmektedir. Modellerin performansları sınıflandırma raporu ve karışıklık matrisi ile ortaya çıkarılmıştır. Sınıflandırma raporunda doğruluk, kesinlik, geri çağırma ve F1 puanı gibi değerlendirme ölçümlerinin yanı sıra makro ve ağırlıklı ortalama değerler de gösterilir. Bu süreç, modelin gerçek dünya verilerine ne kadar iyi uyduğunu anlamak açısından kritik öneme sahiptir. Tüm makine öğrenimi modellerinin sınıflandırma raporları Tablo 4.6'den 4.12'ye kadar ayrıntılı olarak gösterilmektedir. Şekil 4.7'den 4.20'ye kadar, modellerin karışıklık matrislerini ve kategorik değişkenleri görselleştirmek için kullanılan sütun grafiklerini göstermektedir. Karışıklık matrisi, modelin hangi sınıfları doğru veya yanlış tahmin ettiğini gösterir.

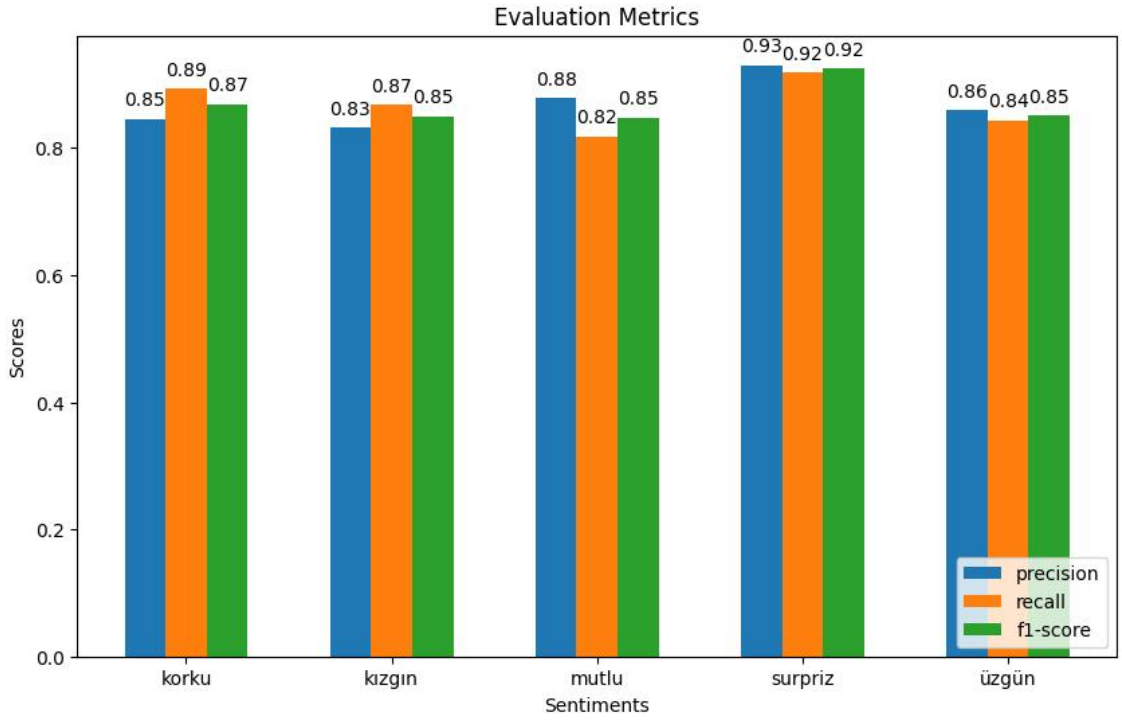
Tablo 4. 6. GNB algoritmasının performansı

	kesinlik	**duyarlılık**	**F1 puanı**
korku	0.85	0.89	0.87
kızgın	0.83	0.87	0.85
mutlu	0.88	0.82	0.85
surpriz	0.93	0.92	0.92
üzgün	0.86	0.84	0.85
doğruluk			0.87
makro ortalama	0.87	0.87	0.87
ağırlıklı ortalama	0.87	0.87	0.87

İlk olarak GNB algoritmasının performansı hesaplanmıştır. GNB algoritması ile elde edilen sonuçlara göre, hassasiyette en yüksek değere (%93) 'sürpriz' etiketi, geri çağırmada ise en düşük değere (%82) 'mutlu' etiketi ulaşmıştır. Bu, bu model için en kötü sonuçtur. Sonuç olarak bu modelin doğruluğu %87'dir. Bu modelle iyi bir sonuç elde ettiğimizi söyleyebiliriz.



Şekil 4. 7. GNB algoritmasının karışıklık matrisi

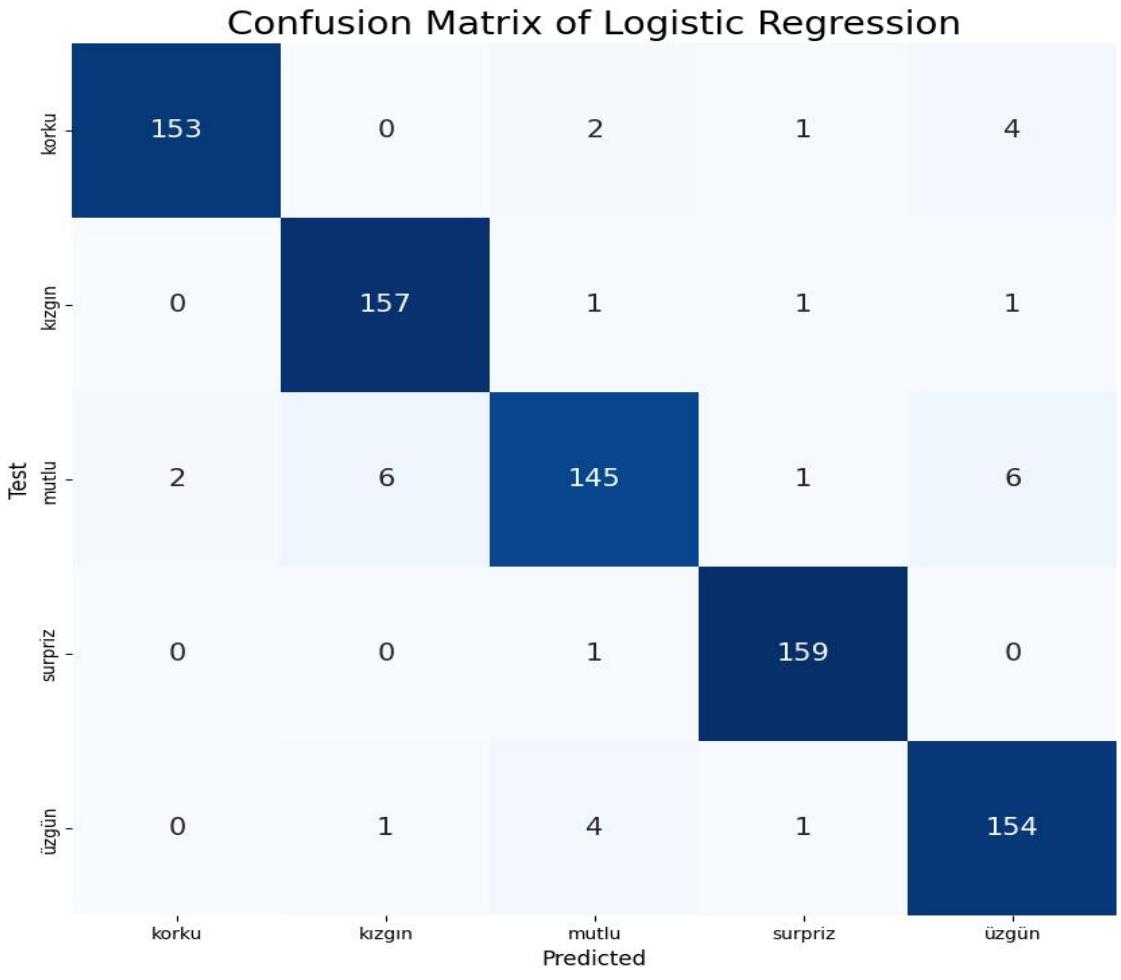


Şekil 4. 8. GNB algoritmasının performans metrikleri için grafik

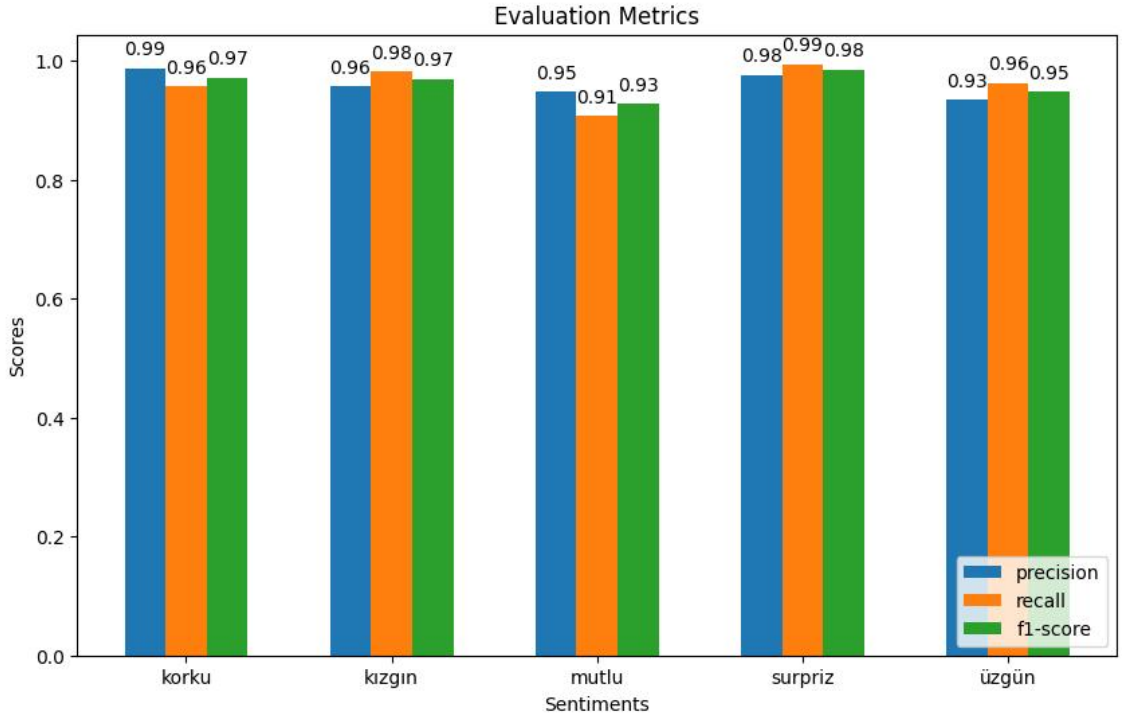
İkinci olarak Lojistik regresyon algoritmasının performansı hesaplanmıştır. Bu algoritmayla daha iyi sonuçlar elde edilmiştir. Duyarlılıkta 'sürpriz' etiketiyle %99'a varan zor bir doğruluk, 'korku' etiketiyle ise kesinlik elde edildi. Bu değeri %98 ile 'sürpriz' etiketinin kesinlik ve F1 skoru değerleri, 'kızgın' etiketinin duyarlılık metriği takip etmektedir. Modelin doğruluğu %96'dır. Bu sonuç çalışmanın, performans açısından üstünlüğünü ortaya koymaktadır.

Tablo 4. 7. Lojistik regresyon algoritmasının performansı

	kesinlik	**duyarlılık**	**F1 puanı**
korku	0.99	0.96	0.97
kızgın	0.96	0.98	0.97
mutlu	0.95	0.91	0.93
surpriz	0.98	0.99	0.98
üzgün	0.93	0.96	0.95
doğruluk			0.96
makro ortalama	0.96	0.96	0.96
ağırlıklı ortalama	0.96	0.96	0.96



Şekil 4. 9. Lojistik regresyon algoritmasının karışıklık matrisi

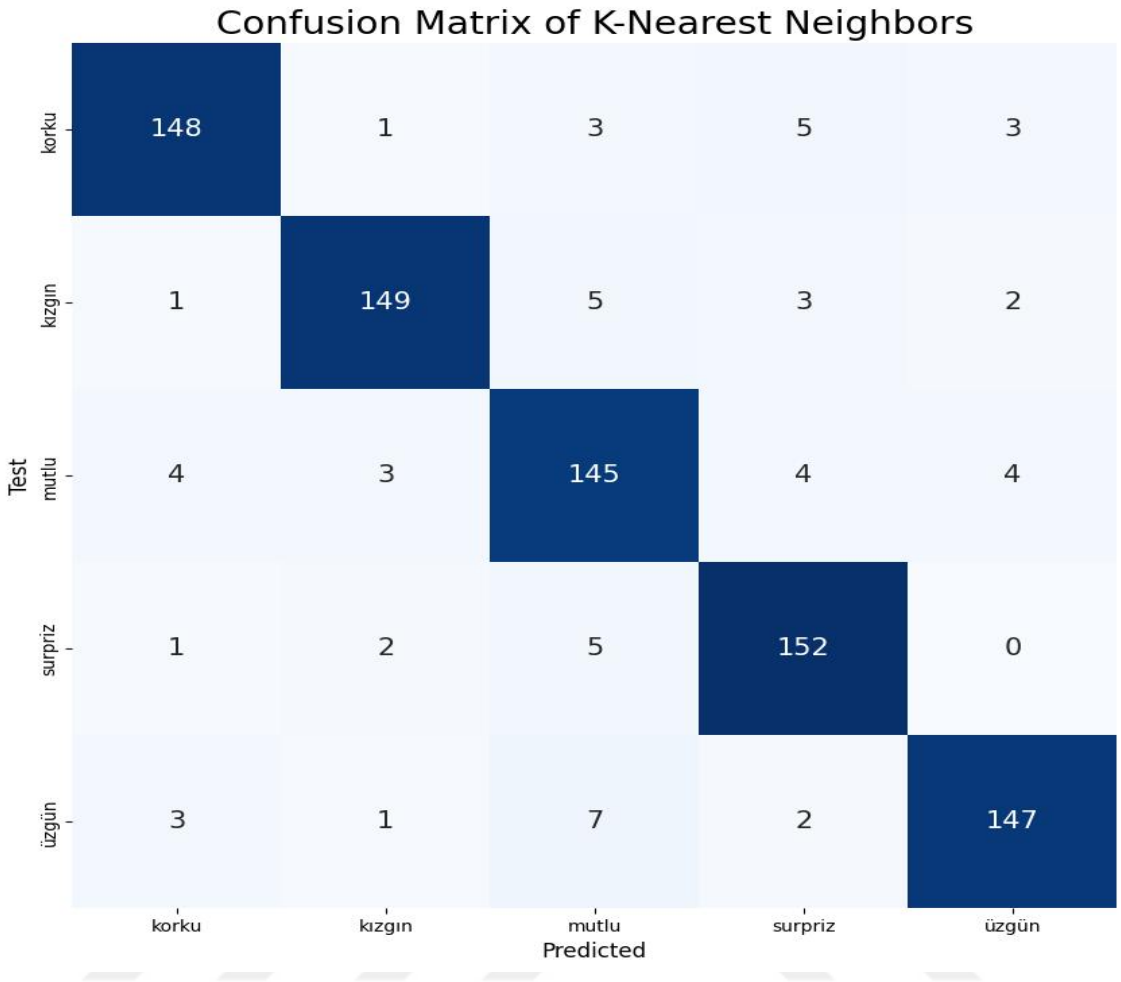


Şekil 4. 10. Lojistik regresyon algoritmasının performans metrikleri için grafik

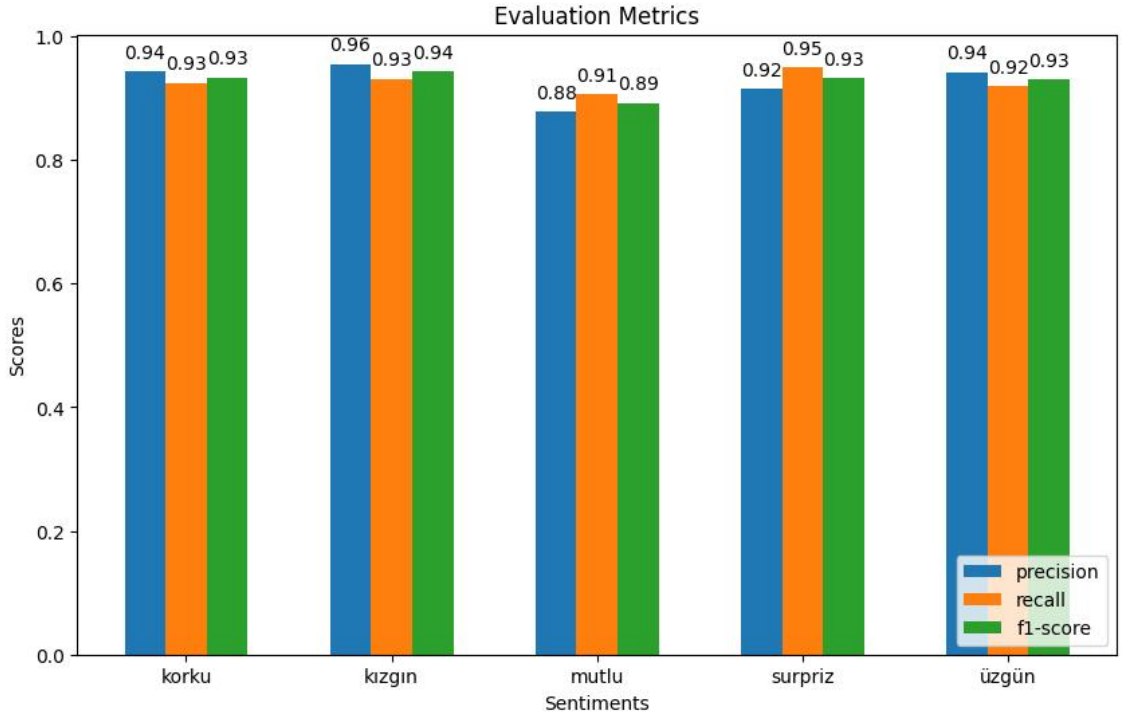
Daha sonra KNN algoritmasının performansı hesaplanmıştır. Bu algoritma aynı zamanda %93'lük bir doğruluk vererek iyi performans göstermiştir. En iyi sonuç "kızgın" etiketinin %96 olan kesinlik metriğinden elde edilmiştir.

Tablo 4. 8. KNN algoritmasının performansı

	kesinlik	**duyarlılık**	**F1 puanı**
korku	0.94	0.93	0.93
kızgın	0.96	0.93	0.94
mutlu	0.88	0.91	0.89
surpriz	0.92	0.95	0.93
üzgün	0.94	0.92	0.93
doğruluk			0.93
makro ortalama	0.93	0.93	0.93
ağırlıklı ortalama	0.93	0.93	0.93



Şekil 4. 11. KNN algoritmasının karışıklık matrisi

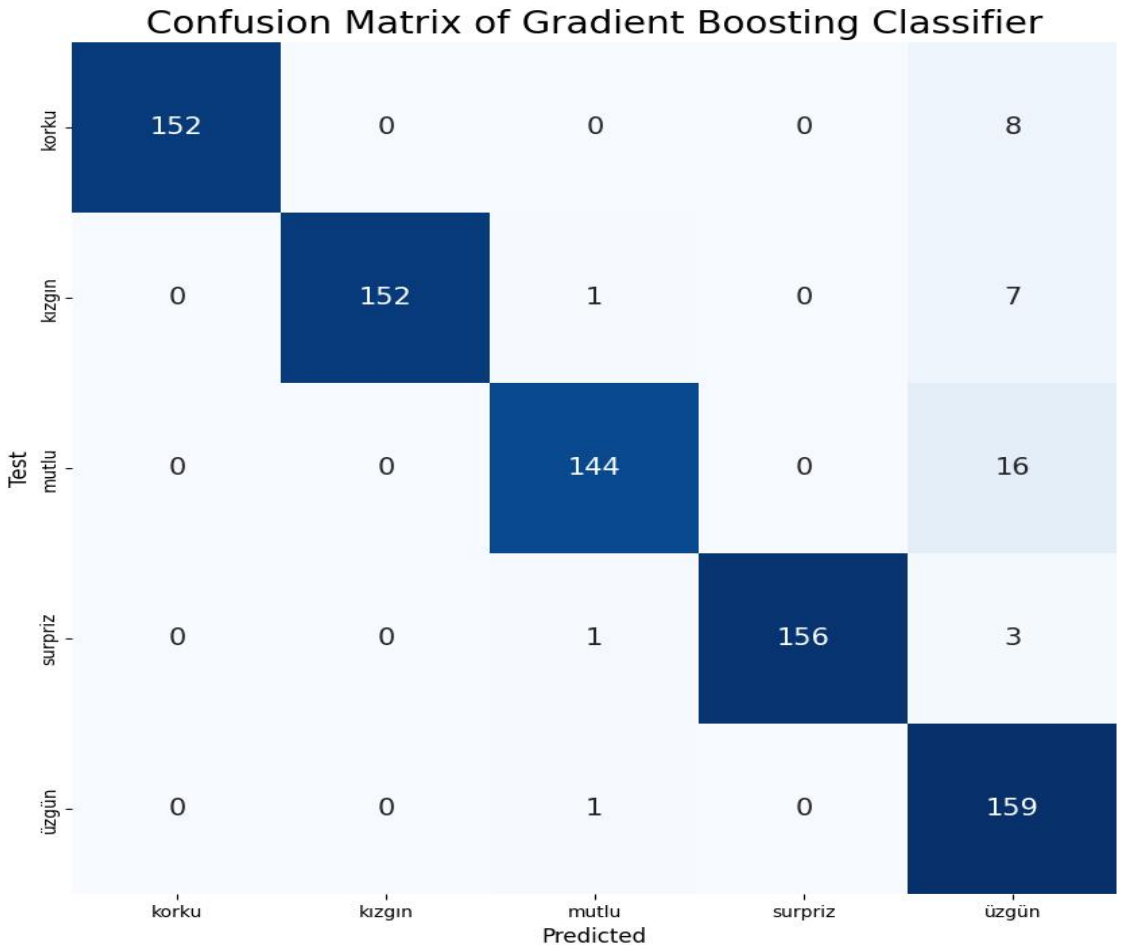


Şekil 4. 12. KNN algoritmasının performans metrikleri için grafik

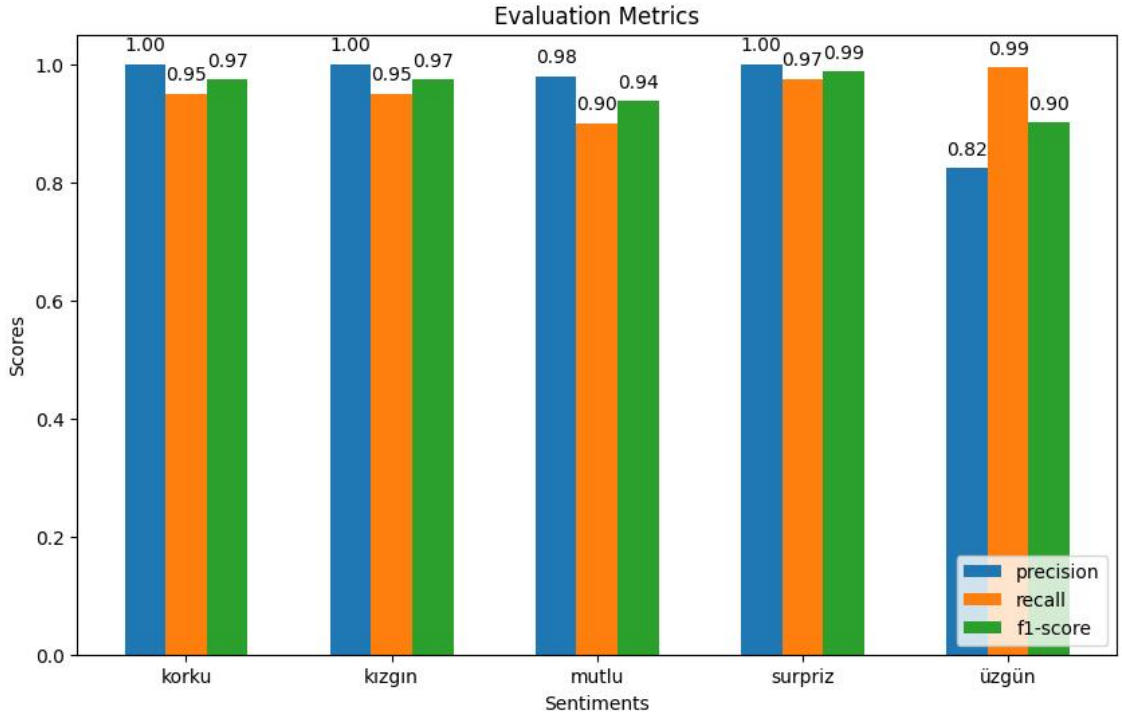
Daha sonra sırasıyla GBC, DT ve ET algoritmalarının sınıflandırma raporları ve modellerin karışıklık matrisleri verilmiştir. GBC algoritması ile %95 doğruluk elde edilMİŞTİR. Tıpkı Lojistik Regresyon modelinde olduğu gibi ulaşmak istediğimiz çok yüksek performansa ulaşılmıştır. 'Korku', 'kızgın' ve 'sürpriz' etiketlerinin kesinlik metriğinde %100 doğruluğa ulaşılmıştır. Bu, elde edilebilecek en iyi sonuçtur. DT algoritması ile %94 doğruluk elde edilmiştir. Diğer birçok makine algoritması gibi bu model de çok iyi performans göstermiştir. Birçok metrikte %90'ın üzerinde doğruluğa ulaşılmıştır. ET algoritması, Lojistik Regresyon gibi %96 ile makine algoritmaları arasında en yüksek doğruluktan birine ulaşmıştır. 'Sürpriz' etiketi her değerlendirme metriğinde %100 doğruluğa ulaşmıştır. Son olarak Yığın modeli uygulanmıştır. Bu yaklaşım, makine öğrenimi çalışmalarında en iyi sonuçları elde etmek için sıklıkla kullanılır ve daha önce de belirtildiği gibi modelin gerçek dünya verilerine daha iyi uymasını sağlar. Bu sayede ET ve Lojistik Regresyondan bile daha iyi sonuç veren bu modelle %96,88 doğruluk oranına ulaşılmıştır.

Tablo 4. 9. GBC algoritmasının performansı

	kesinlik	**duyarlılık**	**F1 puanı**
korku	1.00	0.95	0.97
kızgın	1.00	0.95	0.97
mutlu	0.98	0.90	0.94
surpriz	1.00	0.97	0.99
üzgün	0.82	0.99	0.90
doğruluk			0.95
makro ortalama	0.96	0.95	0.96
ağırlıklı ortalama	0.96	0.95	0.96



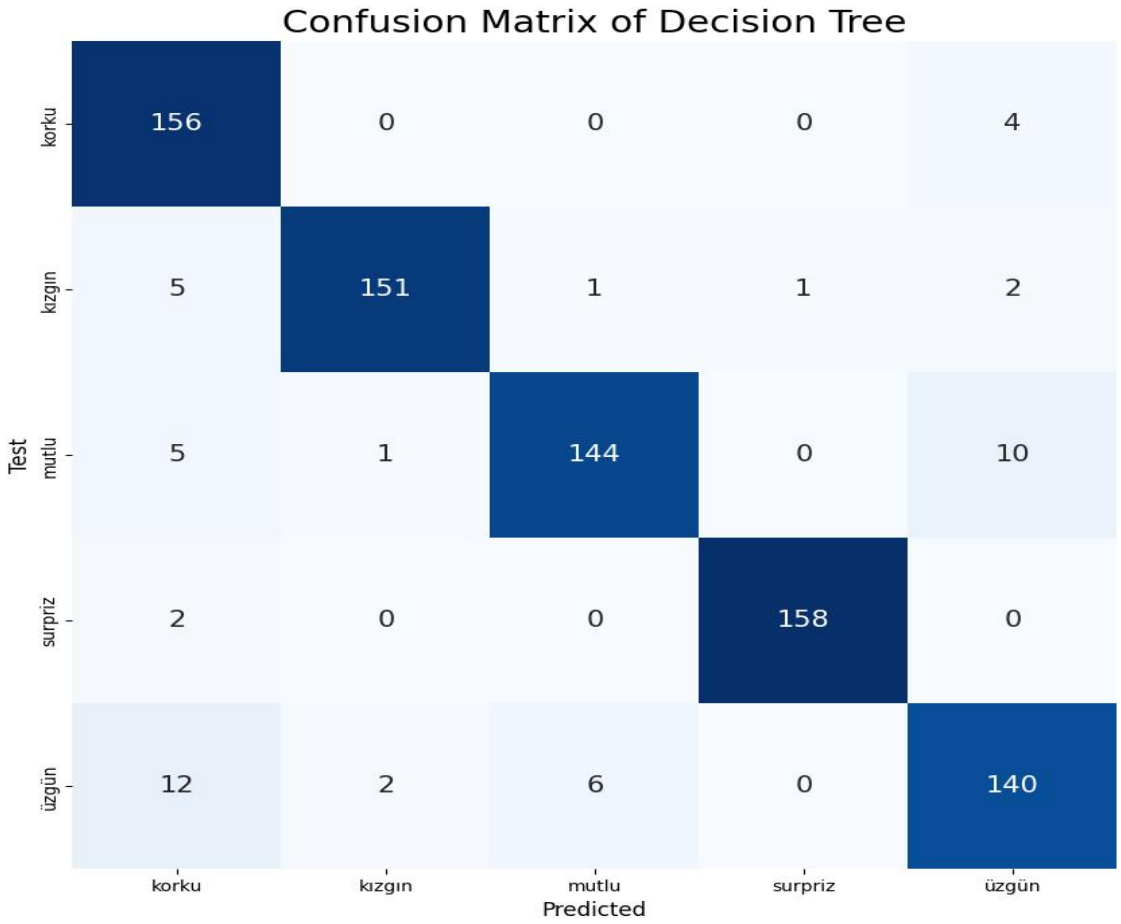
Şekil 4. 13. GBC algoritmasının karışıklık matrisi



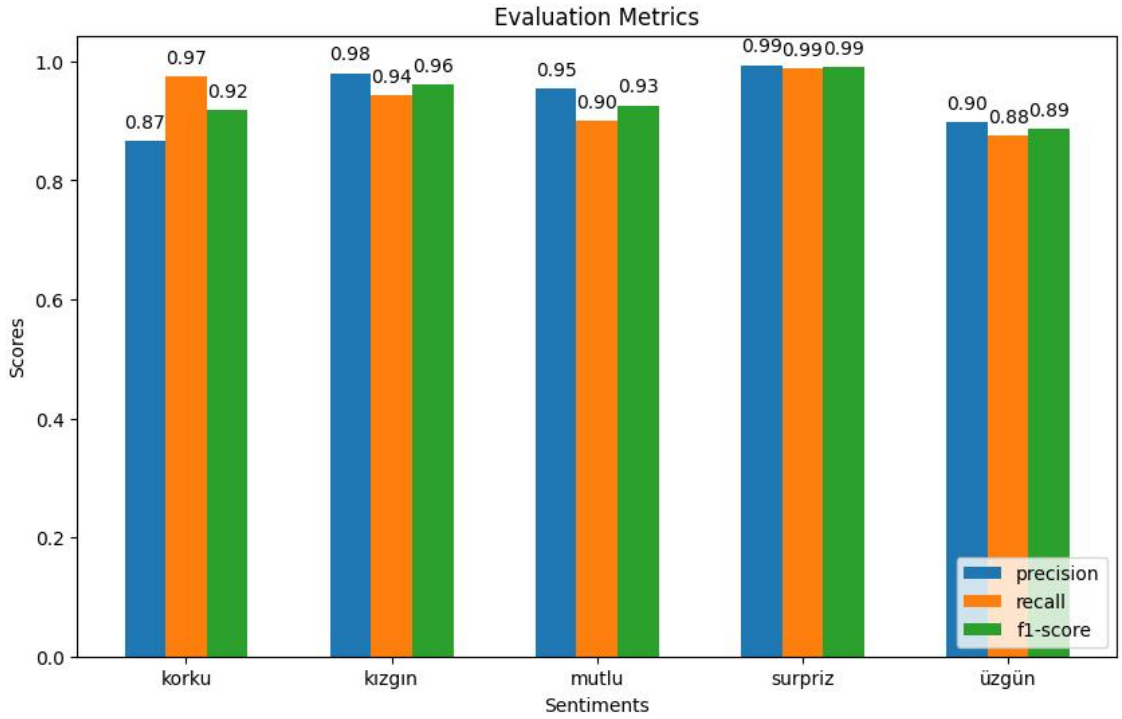
Şekil 4. 14. GBC algoritmasının performans metrikleri için grafik

Tablo 4. 10. DT algoritmasının performansı

	kesinlik	**duyarlılık**	**F1 puanı**
korku	0.87	0.97	0.92
kızgın	0.98	0.94	0.96
mutlu	0.95	0.90	0.93
surpriz	0.99	0.99	0.99
üzgün	0.90	0.88	0.89
doğruluk			0.94
makro ortalama	0.94	0.94	0.94
ağırlıklı ortalama	0.94	0.94	0.94



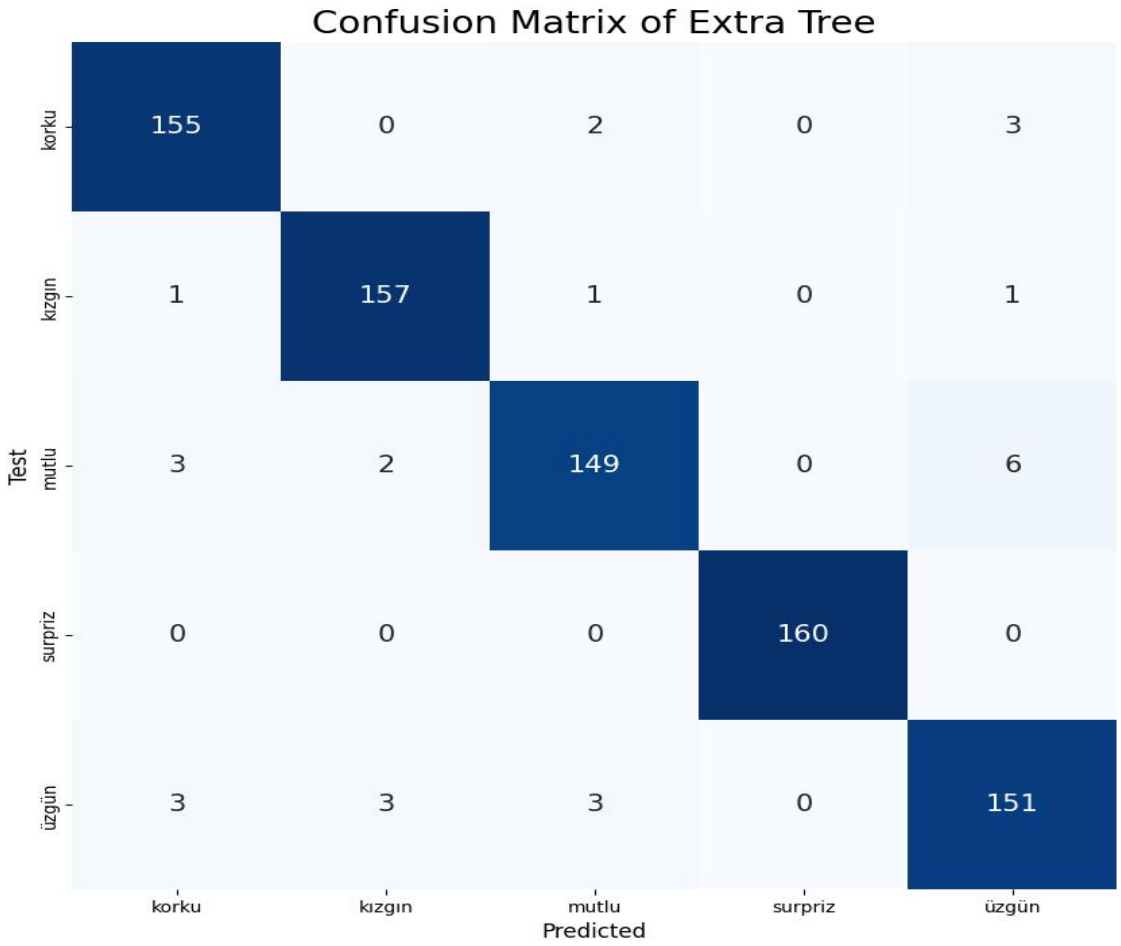
Şekil 4. 15. DT algoritmasının karışıklık matrisi



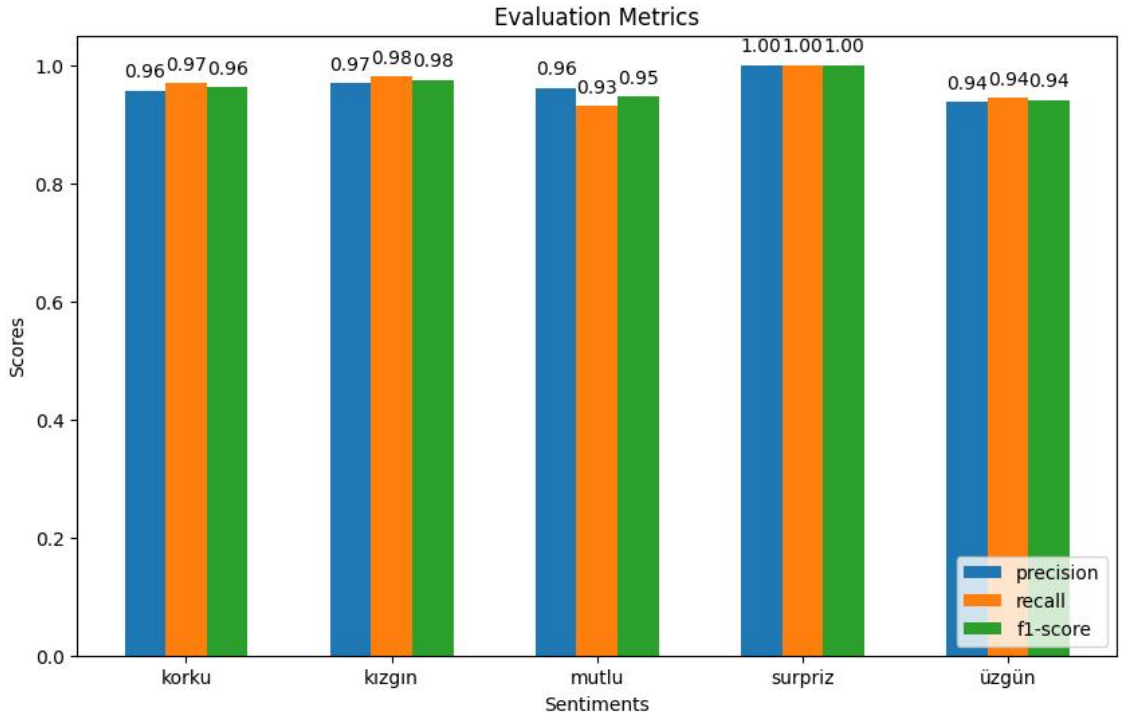
Şekil 4. 16. DT algoritmasının performans metrikleri için grafik

Tablo 4. 11. ET algoritmasının performansı

	kesinlik	**duyarlılık**	**F1 puanı**
korku	0.96	0.97	0.96
kızgın	0.97	0.98	0.98
mutlu	0.96	0.93	0.95
surpriz	1.00	1.00	1.00
üzgün	0.94	0.94	0.94
doğruluk			0.96
makro ortalama	0.97	0.97	0.96
ağırlıklı ortalama	0.97	0.96	0.96



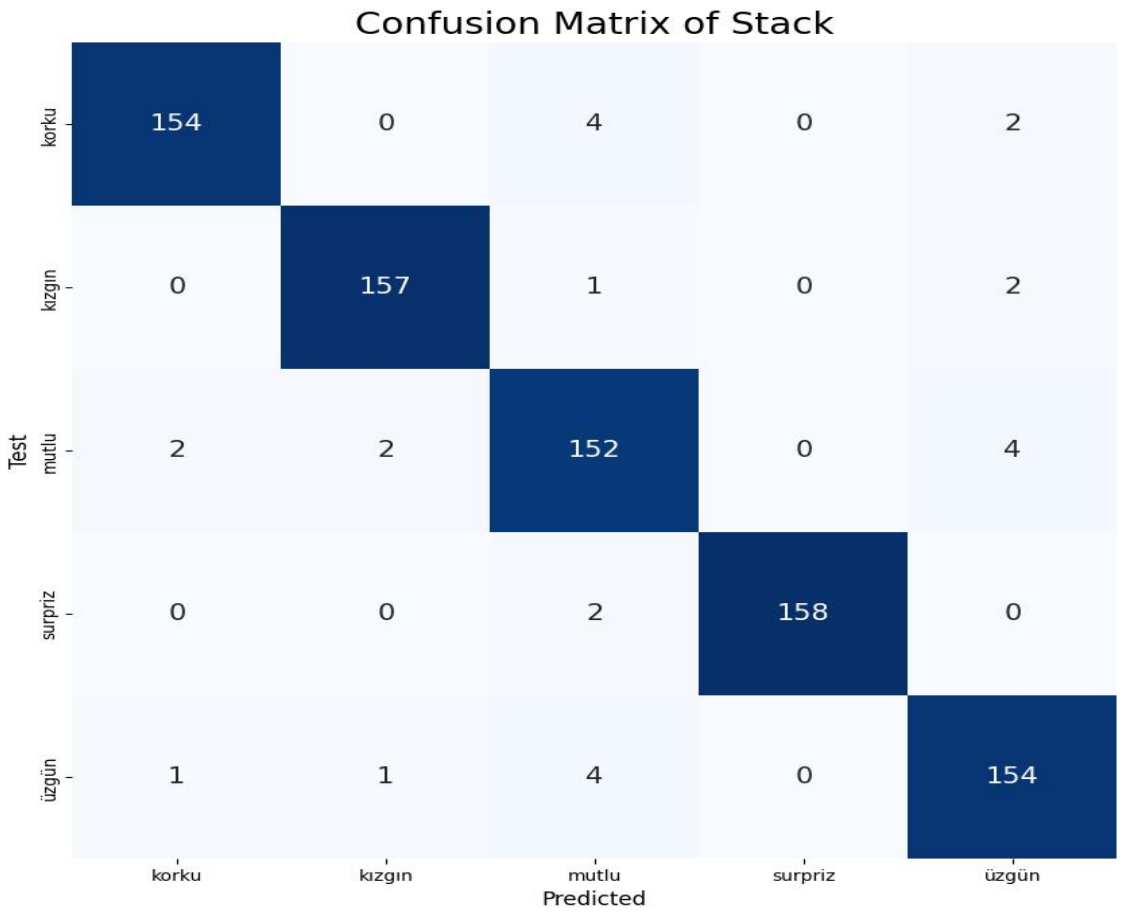
Şekil 4. 17. ET algoritmasının karışıklık matrisi



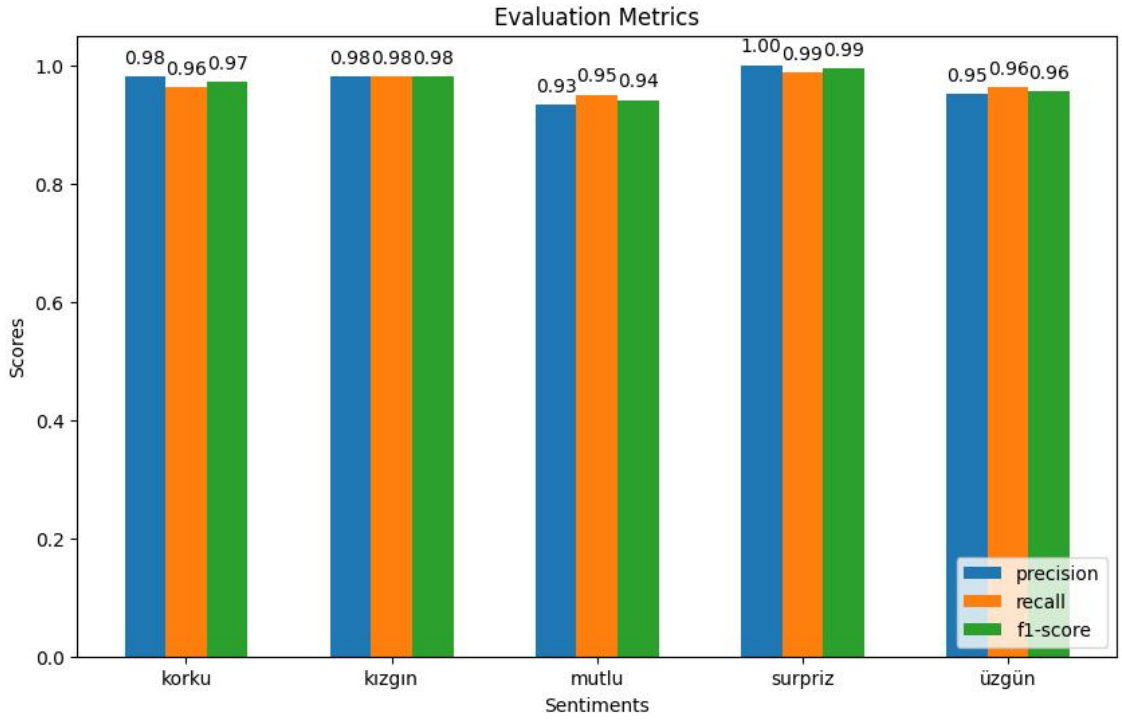
Şekil 4. 18. ET algoritmasının performans metrikleri için grafik

Tablo 4. 12. Yığın algoritmasının performansı

	kesinlik	**duyarlılık**	**F1 puanı**
korku	0.98	0.96	0.97
kızgın	0.98	0.98	0.98
mutlu	0.93	0.95	0.94
surpriz	1.00	0.99	0.99
üzgün	0.95	0.96	0.96
doğruluk			0.97
makro ortalama	0.97	0.97	0.97
ağırlıklı ortalama	0.97	0.97	0.97

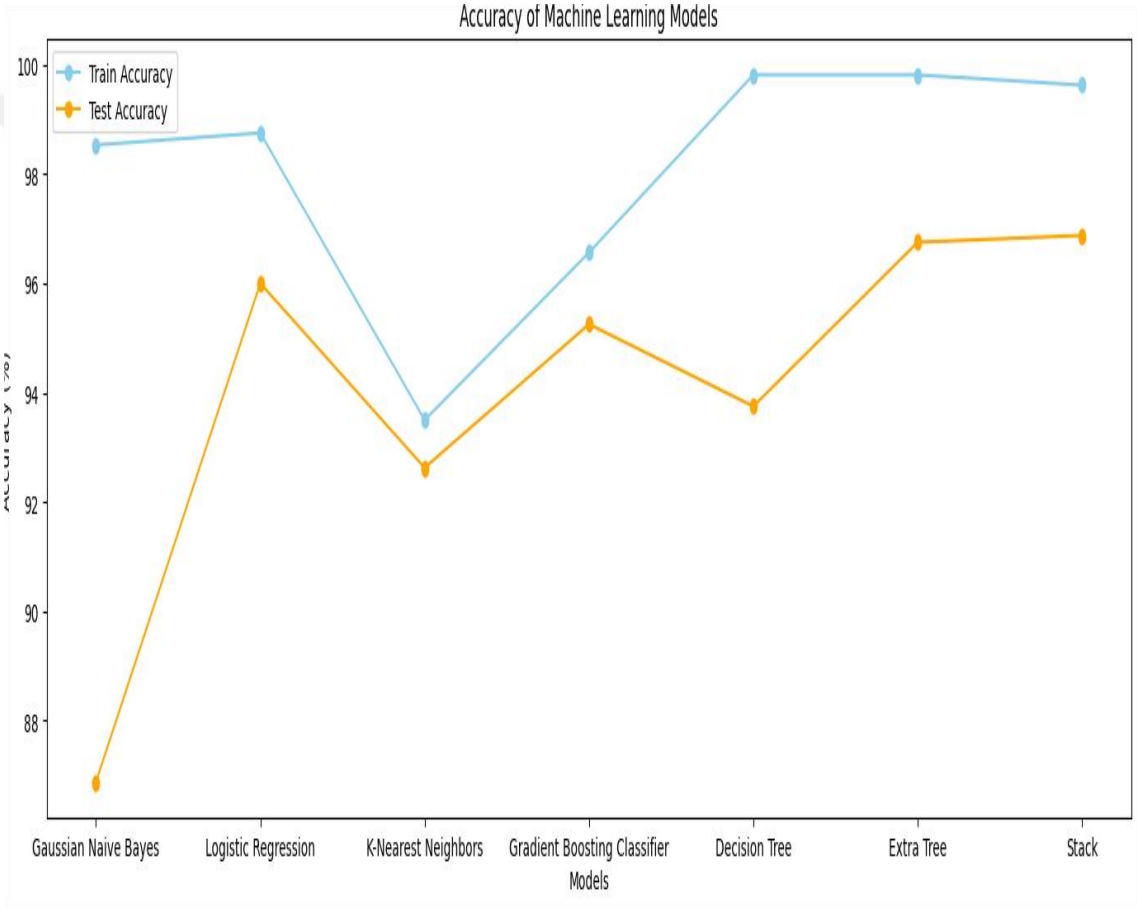


Şekil 4. 19. Yığın algoritmasının karışıklık matrisi



Şekil 4. 20. Yığın algoritmasının performans metrikleri için grafik

Sonuç olarak önceden işlenmiş veri setlerini makine öğrenmesi modelleri ile eğittiğimizde birçok modelin başarılı sonuçlar verdiği görüldü. Buradan verilerin yeterli olduğu ve ön işlemenin doğru yapıldığı sonucunu çıkarabiliriz. Modeller arasında %96,88 test doğruluğuna ulaşan Stack modeli en başarılı model olarak belirlendi. Bunu sırasıyla %96 ve %96,50 doğrulukla Lojistik Regresyon ve Ekstra Ağaçlar algoritmaları takip etmektedir. Bu çalışmada en kötü algoritma %86 doğruluk oranıyla GNB oldu. Şekil 4.21. tüm makine öğrenimi modellerinin doğruluk sonuçlarını grafik olarak gösterir.



Şekil 4. 21. Tüm makine öğrenimi modellerinin eğitim ve test doğruluk değerleri

4.5. Çapraz Doğrulamalı Makine Öğrenimi Algoritmalarının Sonuçları

Bu bölümde, klasik makine öğrenmesi modelleri ile ön işleme tabi tutulup temizlenen ve daha sonra eğitim ve teste ayrılan veri seti üzerinde çapraz doğrulama uygulanan deneyden elde edilen sonuçlar verilmektedir. Modellerin performansları sınıflandırma raporuyla ortaya çıkıyor. Sınıflandırma raporu, Tablo 4.13'te gösterilen her model için doğruluk oranını gösterir.

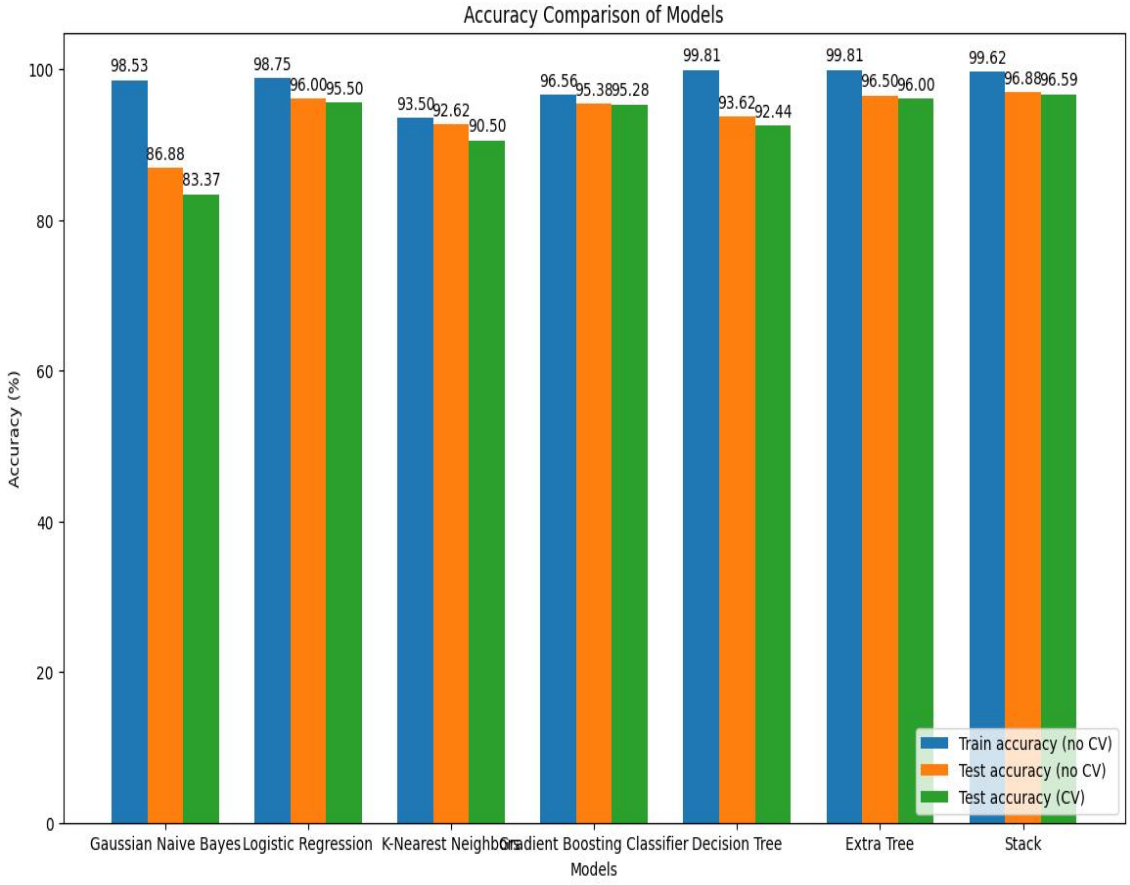
Tablo 4. 13. Çapraz doğrulama uygulandıktan sonra tüm modellerin doğruluk değeri

Model	Doğruluk
GNB	%83,37
Lojistik Regresyon	%95,50
KNN	%90,50
GBC	%95,28
DT	%92,44
ET	%96
Yığın	%96,59

Performanslar incelendiğinde çapraz doğrulama yöntemini uygulayan modellerden biri olan Yığın modelinin bir önceki denemeye benzer şekilde %96,59 doğruluk oranıyla tüm modeller arasında en iyi performansa sahip olduğu belirlendi. Sonrasında %96 ile ET, %95,50 ile Lojistik Regresyon ve %95,28 ile GBC modelleri bu çalışmada en yüksek doğruluk oranlarına ulaşan sınıflandırma algoritmalarıdır.

Bu deneyle aşırı öğrenme sorunu çözülmüştür. Çapraz doğrulama uygulandığında tüm modellerin doğruluk değerleri düşse de bu düşüşe rağmen tüm modeller (GNB hariç) %90'ın üzerinde başarı elde etmişlerdir. Bu bakımdan çalışma amacına ulaşmıştır.

Şekil 4.22, çalışmada uygulanan makine öğrenimi modellerinin çapraz doğrulama uygulanmadan eğitim ve test doğruluklarını ve çapraz doğrulama uygulandığında test doğruluğunu karşılaştırmalı olarak göstermektedir.



Şekil 4. 22. Tüm modellerin doğruluk karşılaştırma grafiği

4.6. BERT Modelinin Sonuçları

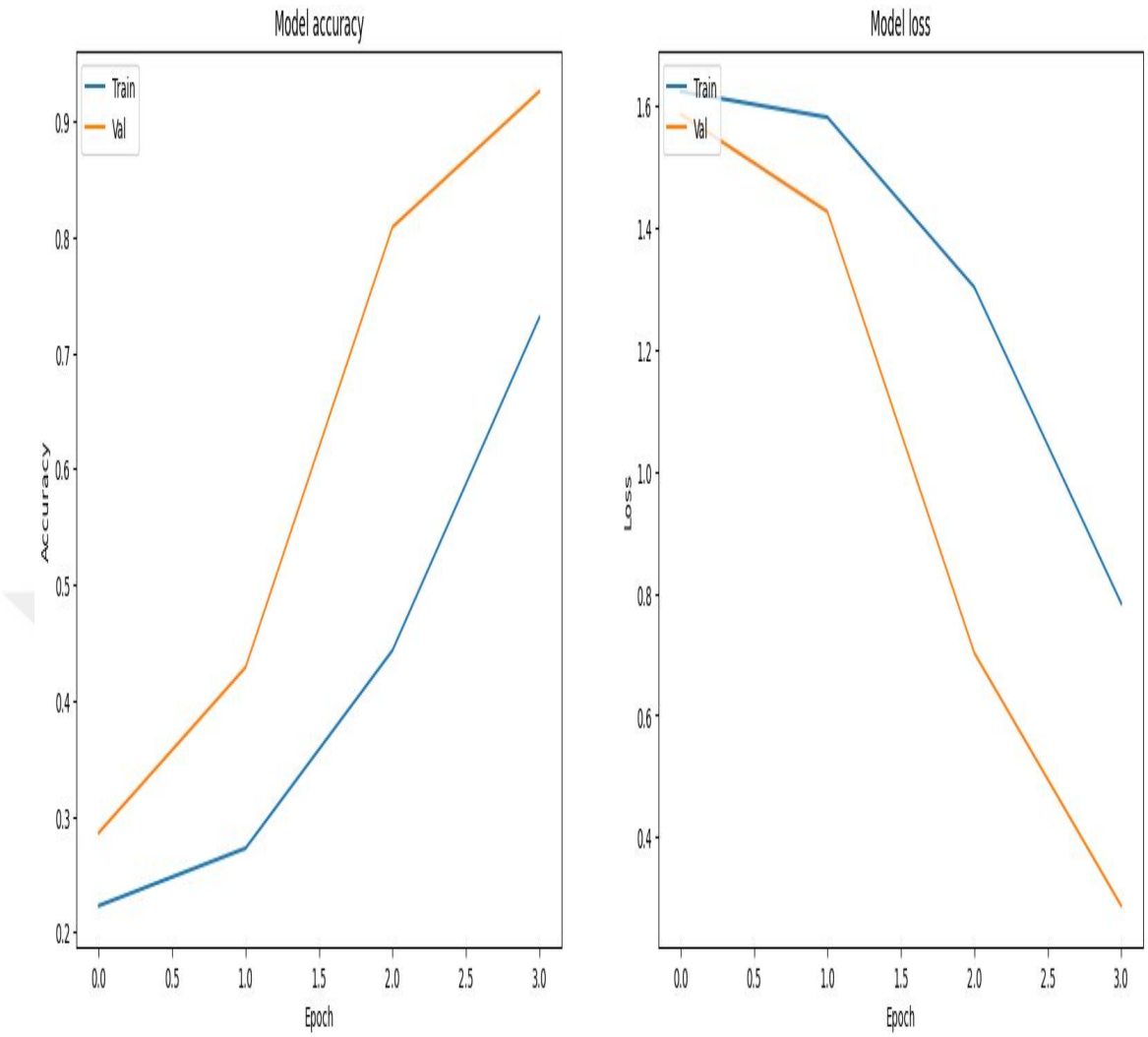
Bu bölümde ön işleme tabi tutulup temizlenen ve daha sonra eğitim, doğrulama ve test olarak ayrılan veri seti üzerinde BERT modeli uygulandıktan sonra elde edilen sonuçlar verilmektedir. Modellerin performansları sınıflandırma raporu ve karışıklık matrisi ile ortaya çıkarılmıştır. Sınıflandırma raporunda doğruluk, kesinlik, geri çağırma ve F1 puanı gibi değerlendirme metriklerinin yanı sıra mikro, makro ve ağırlıklı ortalamalar da gösterilmiştir. Tablo 4.14'te her bir ölçüm için bu modelin sonucu gösterilmektedir.

BERT modeli kullanıldığında 'sürpriz' ve 'üzgün' etiketlerinin hatırlama doğruluğu dışındaki tüm değerlerde %90'ın üzerinde başarı elde edildi. Özellikle ‘korku’ etiketi ile tüm performans değerlerinde %99 gibi yüksek bir doğruluğa ulaşıldı. Sonuç olarak bu model kullanılarak eğitilen verilerle elde edilen doğruluk %93 olmuştur. Araştırmaya göre Türkçe tweet duyarlılık analizinde bu başarıyı BERT modelini kullanarak yakalayan çok az çalışma vardır.

Tablo 4. 14. BERT modelinin performansı

	kesinlik	**duyarlılık**	**F1 puanı**
korku	0.99	0.99	0.99
kızgın	0.95	0.98	0.96
mutlu	0.90	0.94	0.93
surpriz	0.91	0.89	0.90
üzgün	0.94	0.86	0.90
mikro ortalama	0.93	0.93	0.93
makro ortalama	0.94	0.93	0.93
ağırlıklı ortalama	0.94	0.93	0.93
örnek ortalaması	0.93	0.93	0.93

BERT modelinin eğitim süreci sırasında elde edilen doğruluk ve kayıp metrikleri hesaplanmış ve Şekil 4.23'te gösterilmiştir. İki alt grafikte görselleştirilen şekilde görüldüğü üzere, ilk grafik eğitim ve doğrulama doğruluğunu, ikinci grafik ise eğitim ve doğrulama kaybını göstermektedir. Bu tür grafikler modelin eğitim süreci hakkında bilgi sağlar ve modelin performansının anlaşılmasına yardımcı olur.



Şekil 4. 23. BERT modelinin doğruluk ve kayıp grafiği

5.SONUÇ VE ÖNERİLER

Teknoloji çağının da etkisiyle birçok sosyal medya platformu günlük hayatta yerini almıştır. Bunlardan en önemlisi X'tir (Eski adıyla Twitter). Bu platform sayesinde insanlar duygu ve düşüncelerini metin, resim veya video aracılığıyla paylaşabilmekte ve seslerini geniş kitlelere duyurabilmektedir. Buna göre bilim insanları, reklamcılar, pazarlamacılar, polisler ve daha birçok meslek grubundan kişiler bu platformda paylaşılan verileri kendilerine göre kullanmayı hedeflemektedir. Yapay zeka yöntemleri bu amaçlarla araştırmacılara büyük kolaylık sağlar. Özellikle NLP teknolojisi ile birçok veri işlenebilmekte ve X kullanıcıları hakkında bilgi çıkarımı yapılabilmektedir. Bu teknolojinin kullanıldığı alanlardan biri de duygu analizidir. Büyük miktarda veri toplanıp analiz edilerek kullanıcıların tweet atarken duygusal durumları hakkında tahminlerde bulunulabilmektedir.

Bu çalışma, hazır bir Türkçe veri seti üzerinde duygu analizi yapmayı amaçlamıştır. İngilizce tweetlerle duygu analizi çalışmaları yaygın iken Türkçe tweetlerle yapılan çalışmalar nispeten azdır. Bu sebeple bu tez çalışmasında, kendi ana dilim olan Türkçe dilinde yazılmış tweetlerden oluşan bir veri seti kullanmaktı. Kullanılan veri setinde 4000 adet Türkçe tweet bulunmaktadır. Bu tweetler 5 duyguyla (korku, öfke, üzüntü, mutluluk ve şaşkınlık) etiketlendi. Veri seti öncelikle hiperlinklerin silinmesi, noktalama işaretlerinin, sayıların ve emojilerin kaldırılması, boş satırların silinmesi gibi birçok ön işleme tabi tutuldu. Eğitim ve test setlerine ayrıldıktan sonra klasik makine öğrenmesi algoritmaları uygulandı. Bu algoritmalar GNB, Lojistik Regresyon, KNN, GBC, Yığınlama (RF ve Doğrusal SVC temel sınıflandırıcıları ve Lojistik Regresyon son sınıflandırıcısını kullanan katmanlama modelleme yöntemi), DT ve ET algoritmalarıdır. Bu çalışma çapraz doğrulama yöntemiyle ve çapraz doğrulama yöntemi olmadan iki kez tekrarlandı. Daha sonra ana veri seti eğitim, doğrulama ve test setlerine bölünerek DL alanında kullanılan Transformer modellerinden biri olan BERT modeli ile çalışma tekrarlanmıştır.

Sonuç olarak çapraz doğrulama yapılmadan tüm makine öğrenimi algoritmalarında %85'in üzerinde başarı elde edildi. Özellikle Lojistik Regresyon, ET ve Yığın algoritmaları bu çalışmada %96'nın üzerinde doğruluk oranıyla en iyi performansı gösterdi. Birçok değerlendirme metriğinde %100 sonuç veren GBC algoritması OLDU ancak Yığın algoritması %96,88 ile en iyi doğruluk sonucunu verdiğinden bu

algoritmanın diğer algoritmalar arasında en başarılı algoritma olduğu söylenebilir. GNB'nin %86'lık doğruluk oranı nedeniyle bu çalışma için zayıf bir algoritma olduğu belirlendi. Çapraz doğrulama yöntemi ile Lojistik Regresyon, ET ve Yığın algoritmaları bu çalışmada da %95'in üzerinde doğruluk oranıyla en iyi performansı göstermişlerdir. GNB'nin %83'lük doğruluk oranı nedeniyle bu çalışma için zayıf bir algoritma olduğu görüldü. Derin öğrenme modellerinden biri olan BERT modeli kullanıldığında %93 gibi oldukça iyi bir doğruluk elde edilmiştir. Bu tez çalışması birçok makine öğrenmesi algoritmasının %96 ile %94 arasında doğruluk elde etmesi ve ayrıca BERT modeliyle %90'ın üzerinde performans elde etmesi nedeniyle diğer birçok çalışmaya göre üstünlüğünü ortaya koymuştur.

Daha önce çok fazla veri ile eğitilmiş bir modelin kullanılması çalışmamızda faydalı olmuştur.

Bu çalışmanın bulguları özellikle Türkçede duygu analizi yapmak isteyen araştırmacılar için faydalı olabilir. Bu çalışmada birçok makine öğrenmesi algoritması ve derin öğrenme modeli kullanılmış ve iyi sonuçlar elde edilmiştir. Ancak veri seti nispeten küçüktür. Daha büyük veri setleri üzerinde yapılacak çalışmalarda daha iyi sonuçlar elde edilebilir. Ayrıca RoBERTa, ELMO vb. birçok derin öğrenme modeli kullanılarak çalışmalar tekrarlanabilir.

KAYNAKÇA

Alqaraleh, S. (2020). Turkish Sentiment Analysis System via Ensemble Learning. *Avrupa Bilim ve Teknoloji Dergisi*, 122–129.

Anandarajan, M., Hill, C., & Nolan, T. (2019). Text Preprocessing. In M. Anandarajan, C. Hill, & T. Nolan, Practical Text Analytics (Vol. 2, pp. 45–59). Springer International Publishing. https://doi.org/10.1007/978-3-319-95663-3_4

Avinash, M., & Sivasankar, E. (2019). A Study of Feature Extraction Techniques for Sentiment Analysis. In A. Abraham, P. Dutta, J. K. Mandal, A. Bhattacharya, & S. Dutta (Eds.), *Emerging Technologies in Data Mining and Information Security* (Vol. 814, pp. 475–486). Springer Singapore. https://doi.org/10.1007/978-981-13-1501-5_41

Aydın, C. R., & Güngör, T. (2021). Sentiment analysis in Turkish: Supervised, semi-supervised, and unsupervised techniques. Natural Language Engineering, 27(4), 455–483.

Bai, Y. (2022). RELU-function and derived function review. *SHS Web of Conferences*, *144*, 02006. https://www.shs-conferences.org/articles/shsconf/abs/2022/14/shsconf_stehf2022_02006/shsconf_stehf2022_02006.html

Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. *Decision Analytics Journal*, *3*, 100071.

Bello, A., Ng, S.-C., & Leung, M.-F. (2023). A BERT framework to sentiment analysis of tweets. *Sensors*, *23*(1), 506.

Beygelzimer, A., Kakade, S., & Langford, J. (2006). Cover trees for nearest neighbor. *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*, 97–104. https://doi.org/10.1145/1143844.1143857

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, 107134.

Bishop, C. M. (2006). Pattern recognition and machine learning. *Springer Google Schola*, *2*, 1122–1128.

Breiman, L. (2001). [No title found]. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189–215.

Chauhan, R., Savani, J. R., & Sheta, J. M. (2023). *Classification of Gujarati Articles using Bernoulli Naïve Bayes Classifier and Extra-trees Classifier*. https://www.academia.edu/download/103679736/11211.pdf

Ciplak, Z., & Yildiz, K. (2024). Occupational groups prediction in Turkish Twitter data by using machine learning algorithms with multinomial approach. *Expert Systems with Applications*, *252*, 124175.

Çoban, Ö., Özyer, B., & Özyer, G. T. (2015). Sentiment analysis for Turkish Twitter feeds. 2015 23nd Signal Processing and Communications Applications Conference (SIU), 2388–2391. https://ieeexplore.ieee.org/abstract/document/7130362/?casa_token=27Z5qYpdoyUAAAAA:tuKyCaAmrIUaEzc5DRzB2gKS8g4IaPtmwi0DVhRI5y5bNbgY4V_xw5ObhsPXKBCeLZ0P0NjWMq05

Demir, E., & Bilgin, M. (2023). Sentiment Analysis from Turkish News Texts with BERT-Based Language Models and Machine Learning Algorithms. 2023 8th International Conference on Computer Science and Engineering (UBMK), 01–04. https://ieeexplore.ieee.org/abstract/document/10286719/?casa_token=hHrwLl8fXLUAAAAA:ltLnOztd8kz44eIn4KDyi5QKFWXmu0763NN61c1iJ5AykrZDBEkuCm_X3dNUvYs79G3sJinH7yc

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

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. http://arxiv.org/abs/1810.04805

Doğan, E., & Kaya, B. (2019). Deep learning based sentiment analysis and text summarization in social networks. *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 1–6. https://ieeexplore.ieee.org/abstract/document/8875879/?casa_token=EJioF5k9aVcAAAAA:zZvMn7XTgaqXHTcPQiHpUwpt5JdQ2ofhC7x3K2Iy8ZXzXcQCTxHJrsyRB_7wq_n-6o-8OJTIWuLv

Everything You Need to Know About Logistic Regression—Spiceworks. (n.d.). *Spiceworks Inc*. Retrieved June 30, 2024, from https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/

Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Frontiers in Energy Research*, *9*, 652801.

Fanni, S. C., Febi, M., Aghakhanyan, G., & Neri, E. (2023). Natural Language Processing. In M. E. Klontzas, S. C. Fanni, & E. Neri (Eds.), *Introduction to Artificial Intelligence* (pp. 87–99). Springer International Publishing. https://doi.org/10.1007/978-3-031-25928-9_5

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, *14*(771–780), 1612.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: An Overview* (arXiv:2008.05756). arXiv. http://arxiv.org/abs/2008.05756

Gupta, B., Negi, M., Vishwakarma, K., Rawat, G., Badhani, P., & Tech, B. (2017). Study of Twitter sentiment analysis using machine learning algorithms on Python. *International Journal of Computer Applications*, *165*(9), 29–34.

Jurek, A., Mulvenna, M. D., & Bi, Y. (2015). Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics*, *4*(1), 9. https://doi.org/10.1186/s13388-015-0024-x

Kanakaraj, M., & Guddeti, R. M. R. (2015). Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques. *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, 169–170. https://ieeexplore.ieee.org/abstract/document/7050801/?casa_token=qMqUfxU_tA8AAAAA:zIEI1yFpv5JS1QJsKxjQaWz9Ta__72IbFnt9W4OT8mQWRo4E2BBF7WygDpbZBP5wfu0FjfddNg4

Kasri, M., Birjali, M., Nabil, M., Beni-Hssane, A., El-Ansari, A., & El Fissaoui, M. (2022). Refining word embeddings with sentiment information for sentiment analysis. *Journal of ICT Standardization*, *10*(3), 353–382.

Kavi, D. (2020). *Turkish Text Classification: From Lexicon Analysis to Bidirectional Transformer* (arXiv:2104.11642). arXiv. http://arxiv.org/abs/2104.11642

Khan, I. U., Khan, A., Khan, W., Su'ud, M. M., Alam, M. M., Subhan, F., & Asghar, M. Z. (2021). A review of Urdu sentiment analysis with multilingual perspective: A case of Urdu and roman Urdu language. *Computers*, *11*(1), 3.

Khan, M. T., Durrani, M., Ali, A., Inayat, I., Khalid, S., & Khan, K. H. (2016). Sentiment analysis and the complex natural language. *Complex Adaptive Systems Modeling*, *4*(1), 2. https://doi.org/10.1186/s40294-016-0016-9

Kim, S.-B., Rim, H.-C., Yook, D., & Lim, H.-S. (2002). Effective Methods for Improving Naive Bayes Text Classifiers. In M. Ishizuka & A. Sattar (Eds.), *PRICAI 2002: Trends in Artificial Intelligence* (Vol. 2417, pp. 414–423). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45683-X_45

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), 150.

Lubis, A. R., & Lubis, M. (2020). Optimization of distance formula in K-Nearest Neighbor method. *Bulletin of Electrical Engineering and Informatics*, *9*(1), 326–338.

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*, *27*, 16–32.

Memiş, E., Akarkamçı, H., Yeniad, M., Rahebi, J., & Lopez-Guede, J. M. (2024). Comparative Study for Sentiment Analysis of Financial Tweets with Deep Learning Methods. *Applied Sciences*, *14*(2), 588.

Meng, F., Xiao, X., & Wang, J. (2022). Rating the Crisis of Online Public Opinion Using a Multi-Level Index System. *The International Arab Journal of Information Technology*, *19*(4). https://doi.org/10.34028/iajit/19/4/4

Ontivero-Ortega, M., Lage-Castellanos, A., Valente, G., Goebel, R., & Valdes-Sosa, M. (2017). Fast Gaussian Naïve Bayes for searchlight classification analysis. *Neuroimage*, *163*, 471–479.

Peng, B., Wang, J., & Zhang, X. (2020). Adversarial learning of sentiment word representations for sentiment analysis. *Information Sciences*, *541*, 426–441.

Reyna, N. S., Pruett, C., Morrison, M., Fowler, J., Pandey, S., & Hensley, L. (2022). Twitter: More than Tweets for Undergraduate Student Researchers. *Journal of Microbiology & Biology Education*, *23*(1), e00326-21. https://doi.org/10.1128/jmbe.00326-21

Rui, H., Liu, Y., & Whinston, A. (2013). Whose and what chatter matters? The effect of tweets on movie sales. *Decision Support Systems*, *55*(4), 863–870.

Shehu, H. A., Tokat, S., Sharif, M. H., & Uyaver, S. (2019). Sentiment analysis of Turkish Twitter data. *AIP Conference Proceedings*, *2183*(1). https://pubs.aip.org/aip/acp/article-abstract/2183/1/080004/1018823

Short, J. C., Broberg, J. C., Cogliser, C. C., & Brigham, K. H. (2010). Construct Validation Using Computer-Aided Text Analysis (CATA): An Illustration Using Entrepreneurial Orientation. *Organizational Research Methods*, *13*(2), 320–347. https://doi.org/10.1177/1094428109335949

Silpa-Anan, C., & Hartley, R. (2008). Optimised KD-trees for fast image descriptor matching. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. https://ieeexplore.ieee.org/abstract/document/4587638/

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437.

Twitter User Statistics 2024: What Happened After "X" Rebranding? (2022, July 7). https://www.searchlogistics.com/learn/statistics/twitter-user-statistics/

Tyagi, K., Rane, C., & Manry, M. (2022). Regression analysis. In *Artificial intelligence and machine learning for EDGE computing* (pp. 53–63). Elsevier. https://www.sciencedirect.com/science/article/pii/B9780128240540000071

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012). A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. *Proceedings of the ACL 2012 System Demonstrations*, 115–120. https://aclanthology.org/P12-3020.pdf

Wang, Y., Guo, J., Yuan, C., & Li, B. (2022). Sentiment analysis of twitter data. *Applied Sciences*, *12*(22), 11775.

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7–16.

Yıldırım, E., Çetin, F. S., Eryiğit, G., & Temel, T. (2015). The impact of NLP on Turkish sentiment analysis. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, *7*(1), 43–51.

Zabor, E. C., Reddy, C. A., Tendulkar, R. D., & Patil, S. (2022). Logistic regression in clinical studies. *International Journal of Radiation Oncology* Biology* Physics*, *112*(2), 271–277.

Zahera, H. M., Elgendy, I. A., Jalota, R., Sherif, M. A., & Voorhees, E. (2019). Fine-tuned BERT Model for Multi-Label Tweets Classification. *TREC*, 1–7. https://trec.nist.gov/pubs/trec28/papers/DICE_UPB.IS.pdf

Zimbra, D., Abbasi, A., Zeng, D., & Chen, H. (2018). The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation. *ACM Transactions on Management Information Systems*, *9*(2), 1–29. https://doi.org/10.1145/3185045

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Aslı GÜRSOY
Uyruğu : Türk

EĞİTİM

Derece	**Adı**	**Bitirme Yılı**
Üniversite :	Hasan Kalyoncu Üniversitesi	2019
Yüksek Lisans :	Hasan Kalyoncu Üniversitesi	-

İŞ DENEYİMLERİ

Yıl	**Kurum**	**Görevi**
2022-Günümüz	Hasan Kalyoncu Üniversitesi	Araştırma Görevlisi

YABANCI DİLLER

İngilizce