Amazon Ürün Yorumlarından Olumlu-Olumsuz Duygu Analizi

Arman Kuyucu[1], Recep Aydın[2]

[1,2] *Bilgisayar Mühendisliği, Kocaeli Üniversitesi*
Kabaoğlu Mahallesi, Baki Komsuoğlu Bulvarı No:515, Umuttepe, 41001 İzmit/Kocaeli
[1]190201099@kocaeli.edu.tr
[2]200202093@kocaeli.edu.tr

***Abstract*— Doğal Dil İşleme(Natural Language Processing) insanların birbirleri aralarında konuşmak için kullandıkları dili, bilgisayar-insan etkileşiminde kullanmak için yapılan uygulamaları inceleyen bir yapay zeka alt alanıdır. Duygu analizi (Sentiment Analysis), mesajın duygusal tonunun olumlu, olumsuz veya tarafsız olup olmadığını belirlemek için dijital metnin analiz edilme sürecidir. Amazon ürün yorumlarını barındıran etiketli bir veri seti kullanılmıştır. Çalışmada Rastgele Orman, Karar Ağacı, K-En Yakın Komşu, Naive-Bayes ve Ekstrem Gradyan Artırma makine öğrenme algoritmaları kullanılarak duygu analizi yapılmıştır. Deneysel Sonuçlara bakılacak olursa %84.8006 F1-Skor ile Ekstrem Gradyan Artırma algoritması en iyi sonucu veren algoritma olmuştur.**

***Keywords*— Duygu Analizi, Doğal Dil İşleme, Yapay Zeka, Makine Öğrenmesi.**

I. Giriş

Duygu analizi temel olarak bir metin işleme (text processing) işlemi olup verilen metinin duygusal olarak ifade etmek istediği sınıfı belirlemeyi amaçlar. Duygu analizinin ilk çalışmaları duygusal kutupsallık (sentimental polarity) olarak geçmekte olup verilen metni olumlu (positive), olumsuz (negative) ve nötr olarak sınıflandırmayı amaçlamaktadır[1].

Projede olumlu olumsuz olmak üzere ikili sınıflandırma yapılmıştır. Kullanılan veri setinde[2] Amazon yorumlarının puanı, 1 ve 2 ise negatif, 4 ve 5 ise pozitif olarak etiketlenmiştir. Puanı 3 olanlar görmezden gelinmiştir. Genel veri seti test ve eğitim verileri ayrı olarak verilmiştir. Bu projede 400.000 örnekten oluşan test verileri kullanılmıştır. Veri setinin içeriği şekil 1'de verilmiştir.

	class_index	review_title	review_text
0	2	Great CD	My lovely Pat has one of the GREAT voices of h...
1	2	One of the best game music soundtracks - for a...	Despite the fact that I have only played a sma...
2	1	Batteries died within a year ...	I bought this charger in Jul 2003 and it worke...
3	2	works fine, but Maha Energy is better	Check out Maha Energy's website. Their Powerex...
4	2	Great for the non-audiophile	Reviewed quite a bit of the combo players and ...
...	...	...	...
399995	1	Unbelievable- In a Bad Way	We bought this Thomas for our son who is a hug...
399996	1	Almost Great, Until it Broke...	My son recieved this as a birthday gift 2 mont...
399997	1	Disappointed !!!	I bought this toy for my son who loves the "Th...
399998	2	Classic Jessica Mitford	This is a compilation of a wide range of Mitfo...
399999	1	Comedy Scene, and Not Heard	This DVD will be a disappointment if you get i...

Şekil 1. Veri Setinin İçeriği

Projede Rastgele Ağaç, Karar Ağacı, K-En Yakın Komşu, Naive-Bayes, XGB olmak üzere makine öğrenmesi modelleri kullanılmıştır.

II. Yöntem

A. Kullanılan Makine Öğrenme Algoritmaları

1) *Rastgele Ağaç (Random Forest)(RF) :* Random Forest algoritması, denetimli sınıflandırma algoritmalarından biridir. Hem regresyon hem de sınıflandırma problemlerinde kullanılmaktadır. Algoritma, birden fazla karar ağacı üreterek sınıflandırma işlemi esnasında sınıflandırma değerini yükseltmeyi hedefler. Random forest algoritması birbirinden bağımsız olarak çalışan birçok karar ağacının bir araya gelerek aralarından en yüksek puan alan değerin seçilmesi işlemidir. Ağaç sayısı arttıkça kesin bir sonuç elde etme oranı artmaktadır. Karar ağaçları algoritması ile arasındaki temel fark, Random Forest algoritmasında kök düğümü bulma ve düğümleri bölme işleminin rastgele olmasıdır. Random forest algoritması, elinde yeterli miktarda ağaç varsa aşırı öğrenme sorununu azaltır. Az oranda bir veri hazırlığına ihtiyaç duyar [3].

2) *Karar Ağacı (Decision Tree)(DT) :* Karar ağaçları, belirli bir parametreyi baz alarak verilerin, sonuç alınana kadar bölündüğü bir denetimli makine öğrenmesi çeşididir. Karar ağaçları geçmişte yaşanmış ve sınıfları belirlenmiş verileri baz alarak doğrusal olmayan ilişkilerin, karma veri türlerinin ve aykırı değerlerin olduğu veri setlerinin sınıflandırılmasını sağlar. Karar ağaçları, yapısal olarak kök düğüm, düğüm ve dallardan meydana gelir. Bir karar ağacı oluşturmak istersek öncelikle sorunun sorulacağı ana değişken belirlenmelidir. Bu aşamayla birlikte kök düğüm seçilir, burada kök düğüm bağımlı değişkeni temsil eder. Sonraki aşamalarda ise bir sonuca varana kadar sorular sorulur. Sorular düğüm noktalarını, cevaplar dalları temsil eder. Ana mantıkta C programlama dilindeki if-else yapısına benzer bir sistem vardır[4].

3) *K-En Yakın Komşu (K-Nearest Neighbour)(KNN) :* KNN algoritması, makine öğrenmesinde kullanılan en yaygın sınıflandırma algoritma türlerinden biridir. Tembel bir öğrenme türüdür. Eğitim veri setini öğrenmek yerine ezberler ve tahmin yapılması istendiğinde veri setindeki en yakın komşuları arar. Bilinmeyen bir noktaya en yakın olan komşuların sayısı K olarak tanımlanır. Eğitim veri seti içermediği için geniş bir veri seti kullanılan algoritmalar için iyi bir tercih değildir. Çalışma mantığına gelinecek olursa verinin (hedef noktanın) diğer verilerle (komşularla) olan benzerlik miktarı hesaplanır. Komşuların içinden en yakın K adeti belirlenir(Bu sayı genellikle tek sayıdır). K adet komşunun her biri için sınıflandırma yapılır. Hedef noktaya en yakın olan noktalar aynı sınıfa yazılır[5].

4) *Naive Bayes (NB) :* Naive Bayes sınıflandırıcısı olasılık tabanlı bir sınıflandırıcıdır. Sınıflandırıcı, özniteliklerin birbirinden bağımsız olduğunu varsayar. Bu, bir özniteliğinin var olmasının veya olmamasının, başka bir özniteliğin var olması, olmaması veya değerini etkilemeyeceği anlamına gelir. Naive Bayes sınıflandırıcısı Bayes teoremine dayanmaktadır. Bayes teoremi formülü şekil 2'de verilmiştir[6].

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

Şekil 2. Bayes Teoremi Formülü

Naive Bayes sınıflandırma algoritması şu şekilde açıklanabilir:

a) D'nin veri seti kümesini temsil ettiği ve D'deki her elemanın sınıfının belli olduğu varsayılırsa, X ise n tane nitelikten oluşan bir vektör olarak kabul edilirse X=(x1 , x2 , … , xn) olarak temsil edilmektedir.

b) C ise sınıfları temsil ederken m tane sınıf olduğu varsayılsın. Naive Bayes sınıflandırma algoritması bir X vektörünün C'de yer alan herhangi bir sınıfa ait olup olmadığını belirlemek amacıyla, bütün sınıflar içinde en yüksek olasılığa sahip değeri bulmaya çalışır.

c) Sonuçta, sınıflandırıcı en büyük olasılık değerine sahip olan C_i sınıfını, X vektörünün sınıfı olarak seçer[6].

5) XGBoost (Extreme Gradient Boosting)(XGB) : Gradyan Artırma (Gradient Boosting) algoritmasının çeşitli düzenlemeler ile optimize edilmiş yüksek performanslı halidir. Algoritmanın en önemli özellikleri yüksek tahmin gücü elde edebilmesi, aşırı öğrenmenin önüne geçebilmesi, boş verileri yönetebilmesi ve bunları hızlı yapabilmesidir. Daha az kaynak kullanarak üstün sonuçlar elde etmek için yazılım ve donanım optimizasyon tekniklerini uygulanmıştır. Karar ağacı tabanlı algoritmaların en iyisi olarak gösterilir. Çalışma mantığı Gradyan Artırma (Gradient Boosting) ile oldukça benzerdir[7].

B. Veri Önişleme ve Kelime-Vektör Dönüşümü

Etkisiz kelimeleri elde etmek için NLTK kütüphanesi kullanılmıştır.

TfidfVectorizer fonksiyonuna parametre verilerek kelimeleri küçük harfe dönüştürme, vurgu işaretlerini ve etkisiz kelimeleri çıkarma işlemleri yapılmıştır.

TF-IDF (Term Frequency — Inverse Document Frequency); terim frekansı ve ters belge frekansı anlamına gelir. Metni anlamlı sayılar olarak göstermenin bir yoludur, vektör gösterimi olarak da bilinir. Bir bilgi erişim problemini çözmek için kullanılan TF-IDF, belge sınıflandırması, konu modelleme ve durdurma kelimesi filtreleme (stop-word filtering) dahil olmak üzere çeşitli durumlarda kullanılan doğal dil işleme (NLP) algoritmalarında rol almıştır.

Terim sıklığı(TF), metinde ya da veri kümesinde bulunan her kelimenin kaç kez geçtiğini yakalar.

Ters belge frekansı(IDF), derlemedeki belge sayısının, incelenen anahtar kelimeyi içeren topluluktaki belge sayısına bölünmesiyle elde edilen logaritmadır.

IDF aslında bize kelimenin belge için ne kadar önemli olduğunu söyler. Bu, o kelimenin tüm belge setinde ne kadar yaygın veya nadir olduğu anlamına gelir. 0'a ne kadar yakınsa, kelime o kadar yaygındır. Yani kelime çok yaygınsa ve birçok belgede yer alıyorsa bu sayı 0'a yaklaşacaktır. Aksi takdirde 1'e yaklaşacaktır.[8].

C. K-Fold Çapraz Doğrulama

K-Fold Çapraz Doğrulama, sınıflandırma modellerinin değerlendirilmesi ve modelin eğitilmesi için veri setini parçalara ayırma yöntemlerinden biridir. Veriyi belirlenen bir K sayısına göre eşit parçalara böler, her bir parçanın hem eğitim hem de test için kullanılmasını sağlar, böylelikle dağılım ve parçalanmadan kaynaklanan sapma ve hataları asgariye indirir. Ancak modeli k kadar eğitmek ve test etmek gibi ilave bir veri işleme yük ve zamanı ister. Bu durum eğitim ve testi kısa süren küçük ve orta hacimli veriler için sorun olmasa da büyük hacimli veri setlerinde hesaplama ve zaman yönünden maliyetli olabilir [9].

Bu projede XGB hariç her modelin çapraz doğrulaması hesaplanmıştır. XGB modeli çapraz doğrulama sonucunda nan (not a number) döndürmüştür.

D. Hiper Parametre Ayarlama

Makine öğrenmesinde hiper parametre ayarlaması çok önemlidir. Daha efektif ve başarılı sonucu alabilmek için hiper parametre ayarlaması yapılmıştır.

Ayarlama için GridSearchCV fonksiyonu kullanılmıştır. Bu fonksiyon, parametre olarak verilen hiper parametre

değerlerinden oluşabilecek bütün olasılıkları deneyerek en yüksek çapraz doğrulama başarısını elde eden parametreleri bulmaktadır. Bu projede KNN hariç her modelin hiper parametre ayarlaması yapılmıştır. KNN modelinin çalıştırılması çok zaman aldığından hiper parametre ayarlaması yapılamamıştır.

III. Deneysel Sonuçlar

Çalışma kapsamında ikili sınıflandırma kullanılmış olup olumsuz 0 ve olumlu 1 olmak üzere 2 sınıf için modeller eğitilmiştir ve test edilmiştir.

Ağaç tabanlı algoritmalarda (Rastgele Ağaç, Karar Ağacı) parametreler olduğu gibi alındığında çok uzun sürmektedir. Bunu engellemek için bu algoritmalara "max_depth" parametresi eklenmiştir.

Tablo 1'de kullanılan "V(Varsayılan)" gösterimi Rastgele Ağaç, Karar Ağacı algoritmaları hariç (bu algoritmalarda max_depth=3 alınmıştır.) modellerin parametresiz halini ifade etmektedir.

Tablo 1'de kullanılan "A(Ayarlama)" gösterimi ise hiper parametre ayarlaması işlemini ifade etmektedir.

Metinlerin olumlu ve olumsuz olarak sınıflandırılması için RF, NB, KNN, DT ve XGB algoritmaları test edilmiştir.

Yapılan deneyler sonucunda elde edilen en iyi f-skor(F1), kesinlik, hassasiyet ve doğruluk(Accuracy) skorları Tablo 1'de verilmiştir. Tablodaki veriler, veri setindeki sınıfların örnekleri eşit sayıda olduğundan makro değerler verilmiştir. Tabloya göre en iyi F1 skoru %84.8006 ve %84.8011 doğruluk ile XGB algoritmasında elde edilmiştir. Bu sonuç hiper parametre ayarlaması yapılarak elde edilmiştir.

TABLO I
Algoritmaların Sonuçları

Algoritma	Parametreler	V/A	F-Skor (F1)	Kesinlik (Precision)	Hassasiyet (Recall)	Doğruluk (Accuracy)	ÇD
RF	max_depth=3	V	%74.1217	%74.2313	%74.1426	%74.1464	%73.3065
	criterion='entropy', max_depth=15, n_estimators=1000	A	%81.6808	%81.9802	%81.7142	%81.7200	%81.7050
DT	max_depth=3	V	%58.6516	%65.3366	%61.3546	%61.3221	%61.4190
	max_depth=50, random_state=42	A	%74.1365	%74.3588	%74.1863	%74.1809	%74.1755
KNN	-	V	%67.0182	%67.1148	%67.0478	%67.0518	%67.3135
NB	-	V	%82.7929	%82.8086	%82.7937	%82.7950	%82.9315
	alpha=5.0, binarize=0.01	A	%82.8869	%82.9072	%82.8882	%82.8896	%82.9727
XGB	-	V	%82.7723	%82.7722	%82.7733	%82.7725	-
	learning_rate=0.5, max_depth=15	A	%84.8006	%84.8035	%84.8005	%84.8011	-

V/A: Varsayılan(V), Ayarlama(A). **ÇD:** Ortalama Çapraz Doğrulama Skoru

IV. Sonuçlar

Bu çalışmada, etiketli Amazon ürün yorumlarını barındıran bir veri seti üzerinde makine öğrenmesi algoritmaları kullanılarak duygu analizi gerçekleştirilmiştir. Duygu analizi RF, DT, KNN, NB, XGB algoritmaları kullanılarak yapılmıştır. Deneysel sonuçlara bakılacak olursa hiper parametre ayarlaması yapıldığında tüm algoritmaların sonuçlarında iyileşme görülmüştür.

Kaynaklar

[1] https://www.ybsansiklopedi.com/wp-content/uploads/2016/09/duygu_analizi.pdf (Erişim Zamanı : 25 Kasım 2022)

[2] https://www.kaggle.com/datasets/yacharki/amazon-reviews-for-sa-binary-negative-positive-csv (Erişim Zamanı : 25 Kasım 2022)

[3] https://ece-akdagli.medium.com/makine-%C3%B6%C4%9Frenmesinde-random-forest-algoritmas%C4%B1-a79b044bbb31 (Erişim Zamanı : 28 Kasım 2022)

[4] https://ece-akdagli.medium.com/makine-%C3%B6%C4%9Frenmesinde-decision-tree-42a86502ee75 (Erişim Zamanı : 28 Kasım 2022)

[5] https://ece-akdagli.medium.com/makine-%C3%B6%C4%9Frenmesinde-knn-algoritmas%C4%B1-eaafef16d765 (Erişim Zamanı : 28 Kasım 2022)

[6] İlhan, N. & Sağaltıcı, D. (2020). Twitter'da Duygu Analizi . Harran Üniversitesi Mühendislik Dergisi , 5 (2) , 146-156 . DOI: 10.46578/humder.772929

[7] https://www.veribilimiokulu.com/xgboost-nasil-calisir/ (Erişim Zamanı : 28 Kasım 2022)

[8] https://blog.turhost.com/tf-idf-nedir/ (Erişim Zamanı : 25 Kasım 2022)

[9] https://www.veribilimiokulu.com/bir-bakista-k-fold-cross-validation/ (Erişim Zamanı : 25 Kasım 2022)