Journal of Computer Science
https://dergipark.org.tr/en/pub/bbd
Anatolian Science

ISSN, e-ISSN: 2548-1304
Volume: 9, Issue:2, pp:133-141, 2024
https://doi.org/ 10.53070/bbd.1437223
Research Paper

Makinelerin Öğrenmesi Yöntemlerinin Duygu Okuma Yeteneği: E-Ticaret Yorumlarini Anlamak

Emotion Reading Ability of Machine Learning Methods: Understanding E-Commerce Comments

Mert Halil DURAK*[1], Cengiz HARK*[2]

[1]Bilgisayar Mühendisliği Bölümü, İnönü Üniversitesi, Malatya, Türkiye

[2]Bilgisayar Mühendisliği Bölümü, İnönü Üniversitesi, Malatya, Türkiye

([1]durakmerthalil@gmail.com, [2]cengiz.hark@inonu.edu.tr)

Received:May 07, 2024 Accepted:Jul.10, 2024 Published:Dec.25, 2024

Özetçe— Günümüzde, forumlar, bloglar ve sosyal medyanın geniş insan kesimleri tarafından yoğun bir şekilde kullanılması nedeniyle bireyler, düşüncelerini, fikirlerini ve duygularını bu platformlar üzerinden paylaşmaya başlamıştır. Sosyal medya kullanımındaki artışla birlikte, araştırmacılar da duygu analizi alanında yaptıkları çalışmaları artırmışlardır. Veri hacmindeki hızlı artışla birlikte, bu verilerin etkili bir şekilde yönetilmesi ve içinden anlamlı bilgilerin çıkarılması gerekliliği ortaya çıkmıştır. Bu bağlamda, akıllı hesaplama yöntemlerinin bu verileri analiz etmesi son derece kritik bir öneme sahiptir. Duygu analizi, verilere uygulanan bir dizi süreçle, anlamlı bilgi elde etmek amacıyla gerçekleştirilen bir yöntemdir. Bu araştırma çerçevesinde, bir metin içinde bulunan düşünce içeren verinin sistemli bir şekilde incelenmesi ve metnin duygu kategorisi ile duygu polaritesinin belirlenmesi amaçlanmıştır. Belirli e-ticaret sitelerinde yer alan ürünler ile ilgili yapılan yorumların yer aldığı bir veri seti kullanılmıştır. Toplanan bu veri setinde CountVectorizer ve TF-IDF Vectorizer yöntemleri ile özellik çıkarımları gerçekleştirilmiş ve 2 farklı etiketlemeyi (Pozitif Cümle-Negatif Cümle) doğru bir şekilde yapması amaçlanmıştır. Bu eğitim sırasında Lojistik Regresyon, Naive Bayes, Random Forests ve XGBoost makine öğrenmesi yöntemleri kullanılarak karşılaştırılıp sonuçlar rapor edilmiştir. Raporlanan sonuçların başarım değerleri tablolar üzerinde gösterilip karşılaştırmalar yapılmıştır. Yapılan karşılaştırmalar sonucunda Naive Bayes algoritmasının Count Vektör yöntemi kullanılarak elde edilen %84,9'luk doğruluk değeri diğer algoritmalarda kullandığımız Count Vektör ve TF-IDF yöntemleri sonucunda elde ettiğimiz doğruluk değerlerine göre daha yüksek bir değer olduğu gözlemlenmiştir.

Anahtar Kelimeler : *Duygu Analizi, Doğal Dil İşleme, Derin Öğrenme, İkili Sınıflandırma, Özellik Çıkarımı, Metin Sınıflandırma Modelleri*

Abstract— Today, due to the widespread use of forums, blogs, and social media by large segments of society, individuals have started sharing their thoughts, ideas, and emotions on these platforms. Along with the increase in social media usage, researchers have also intensified their studies in the field of sentiment analysis. With the rapid growth in data volume, the necessity to manage this data effectively and extract meaningful information has emerged. In this context, it is critical for intelligent computing methods to analyze these data effectively. Sentiment analysis is a method applied to data to obtain meaningful insights through a series of processes. Within this research, the goal is to systematically analyze data containing opinions within a text and determine the text's sentiment category and sentiment polarity. A dataset containing reviews of products on specific e-commerce sites was used. In the collected dataset, feature extractions were performed using CountVectorizer and TF-IDF Vectorizer methods, aiming to correctly label two categories (Positive Sentence - Negative Sentence). During this training, machine learning methods including Logistic Regression, Naive Bayes, Random Forests, and XGBoost were used and compared, with the results reported. The reported performance metrics were displayed in tables for comparison. As a result of these comparisons, it was observed that the Naive Bayes algorithm achieved an accuracy of 84.9% using the Count Vector method, which was higher than the accuracy obtained with Count Vector and TF-IDF methods applied in other algorithms.

Keywords : *Sentiment Analysis, Natural Language Processing, Deep Learning, Binary Classification, Feature Extraction, Text Classification Models.*

1. Giriş

E-ticaret siteleri, çevrimiçi depolama sağlayıcıları ve sosyal medya kullanıcılarının her gün hızla çoğalması mevcut veri miktarını da arttırmaktadır. Bu verilere genellikle herhangi bir yapılandırılma uygulanmamaktadır ve bu verilerin yönetilmesi ile beraber düzenlenmesi içinde doğru bir duygu analizi ile sınıflandırılması gerekmektedir. Uygulanacak duygu analizi yöntemleri ile en doğru sonucu verebilecek metin sınıflandırma modelini bulmak önemlidir. Bundan dolayı duygu analizi yöntemleri ve metin sınıflandırma modelleri her zaman önemli bir araştırma konusu olmuştur.

Duygu analizi, internette veya genel ortamlarda ki birçok insanın günlük yaşamında çok önemli oranda yer alması ve bunun sonucunda kişilerin siyasi ve sosyal durumlar, aldıkları hizmetler ve ticari ürünler gibi konular hakkındaki düşüncelerini oluşturan yazılımlar ile rapor edilmesi ve anlamlandırılması işlemidir.

Şekil 1. Duygu Analizinde Kullanılan Teknikler [1] (Amanet, 2020)

Duygu analizi işleminde metinlerin içeriğinin pozitif, negatif veya nötr olma durumlarını belirleyebilmek için analizler yapılmaktadır. Bu analizler sonucunda insanların veya belirli grupların ilgili konu hakkındaki görüşleri, düşünceleri veya fikirleri belirlenmiş olur. Bunun için duygu analizi şirketler açısından yeni çıkacak bir ürün için ön değerlendirme, alınması planlanan kararlara gelecek pozitif veya negatif tepkiler, film tercihi yapacak kişilerin önceki yorumlardan yola çıkarak film tercihi yapması gibi konularda belirleyici bir görüş olanağı sunabilir. Ancak duygu analizi yapılacak verilerin çok büyük hacimlere sahip olması bu analizin ayrı ayrı kontrol edilerek yapılmasını imkânsız hale getirmektedir. Bundan dolayı duygu analizi, metin madenciliği ile makine öğrenmesi alanları söz konusu amaç için büyük öneme sahiptirler.

Bu çalışmada belirli e-ticaret sitelerinden alınan yorumlardan oluşan veriler, Random Forests, Naive Bayes, XGBoost ve Lojistik Regresyon sınıflandırıcıları kullanılarak yapılan duygu analizi sonucundan 2 farklı kategoriye sınıflandırılmıştır (Pozitif Cümle-Negatif Cümle). Uygulanan Duygu Analizi yöntemleri ile metin sınıflandırma modellerinin doğruluk değerleri tablolarla karşılaştırılmıştır.

2. İlişkili Çalışmalar

Yapılan araştırmalarda; S. Tuzcu, çevrimiçi bir e-ticaret sitesinin kullanıcı yorumlarını alarak bunlar üzerinde duygu analizi yapmak için öncelikle Pyhton programlama dili ile Çok katmanlı algılayıcı (Multi-Layer Perceptron, MLP) algoritması kullanmayı hedeflemiştir. Daha sonra çalışmasında, Lojistik Regresyon, Naive Bayes ve Destek Vektör Makinesi algoritmalarını kullanılmıştır ve başarım oranlarını karşılaştırılmıştır. Sonuçlara bakıldığında; Destek Vektör Makinesi, olumlu yorumları sınıflandırmada diğer algoritmalar içinde en iyi sonucu vermiş ancak olumsuz yorumları sınıflandırmada diğer algoritmalardan oldukça düşük performans rapor etmiştir. Naive Bayes ise bu üç algoritmadan farklı olarak olumsuz yorumları olumlu yorumlara göre daha başarılı olmuştur ancak genel sınıflandırma başarısı bu algoritmaların gerisinde kalmıştır [2].

Aliwy ile beraber çalışma arkadaşları, dönüşüm temelli yöntemler kullanarak Arapçada birden fazla etiketli metinlerin kategorize edilmesi üzerine bir çalışma gerçekleştirmişlerdir. Değerlendirme kriteri olarak beş kategoriye ('bilim', 'ekonomi', 'politika', 'spor' ve 'sanat') dağıtılmış 10.000 veriden oluşan bir data set kullanılmıştır. Kullanılan veri seti BBC Arabic haberlerinin olduğu portaldan toplanmış ve her haber konusunun başlıkları kategori olarak belirlenmiştir. Metin Sınıflandırma için Random Forest, Destek Vektör Makinesi (Support vector machine, SVM) ve K-En yakın komşu (K-Nearest Neighbours, KNN) gibi yöntemler kullanılmıştır [3].

Sevindi çalışmasında Türkçe dilinde ki filmler için yapılan yorumların duygu analizini farklı makine öğrenmesi yöntemleri ile tespit etmeye çalışmış ve elde ettiği sonuçları karşılaştırmıştır. Karar Ağacı (Decision Tree, DT), K-En Yakın Komşu (K-Nearest Neighbours, KNN), Naive Bayes ve Destek Vektör Makineleri (Support vector machine, SVM) yöntemlerini kullanmış ve en yüksek doğruluk derecesine sahip sonucu SVM metin sınıflandırma yöntemi ile bulmuştur [4].

J. Sadhasivam yaptığı çalışmada ensemble yaklaşımı, destek vektör makineleri ve Naive Bayes yöntemleri ile Amazon sitesinde ürünlere yapılan yorumlar üzerinde duygu analizi için yöntemlerin başarım oranlarını değerlendirmiş ve ensemble yönteminin %78 oranında bir doğruluk sonucu elde ettiğini gözlemlemiştir [5].

Alshalabi ile beraber çalışma arkadaşları, Dünyada ki Malayca dilinde olan metinlerin makine öğrenmesi algoritmaları ile otomatik şekilde kategorize edilmesine yönelik bir çalışma yürütmüşlerdir. Bu çalışmada iki özellik çıkarım yöntemi (Ki-kare ve Bilgi kazancı) ile üç makine öğrenmesi algoritması (Naive Bayes, N-gram ve KNN) kullanılmıştır [6].

Yıldırım ve Amasyalı gerçekleştirdikleri projede farklı gazete markalarının web sitelerinde ki haber içeriklerini otomatik olarak metin sınıflandırma işlemi ile sınıflandırma yapmaya çalışmışlar, haber içerikleri 5 farklı kategoriye ayrılmış ve en iyi sonuç %76 doğruluk oranı ile Naive Bayes algoritması sonucunda elde edilmiştir [7].

Tüfekçi ve çalışma arkadaşları, haber sitelerinden elde ettikleri iki farklı haber dataset üzerinde gerçekleştirdikleri bu çalışmada haberleri 5 kategoriye ayırmış ve özellik çıkarımı yaparak Naive Bayes, Random Forest, Destek Vektör Makinesi (Support Vector Machine, SVM) ve C4.5 sınıflandırma yöntemleri ile alınan sonuçları karşılaştırmışlardır. Yaptıkları bu çalışma sonucunda Naive Bayes algoritması %92,73'lük doğruluk oranı ile en yüksek doğruluk değerine ulaşmıştır [8].

[9], çalışmasında derin öğrenme teknikleri ile duygu analizini derinlemesine bir şekilde ele almaktadır. E-ticaret sitelerinde bulunan kullanıcı yorumlarının analizi için kullanılan modelleri ve yöntemleri incelemiştir. [10], çalışmasında tüketici yorumlarının tüketici davranışları üzerindeki etkisini sistematik bir biçimde incelemektedir. E-ticaret platformlarındaki yorumların analizi ve bu yorumlarına karşılık tüketici davranışlarının eğilimini ele almıştır. Türkçe metin verileri üzerinde transformer tabanlı derin öğrenme modellerini kullanarak duygu analizinin gerçekleştirildiği bir başka çalışmada [12], E-ticaret platformlarındaki Türkçe kullanıcı yorumlarının analizi için önerilen yöntemler detaylı bir biçimde incelenmiştir. [13]'de ise, metin tabanlı duygu analizi ve duygu tespitine yönelik yöntemler kapsamlı bir biçimde araştırılmaktadır. Özellikle e-ticaret platformlarındaki kullanıcı geri dönüşleri üzerinde durulmaktadır.

Güran ve Erdinç Türkçe dilinde 11 farklı kategoriye sınıflanmış 5 milyon haber içeren bir dataset ile gerçekleştirdikleri projede bir yarı öğreticili teknik ile FastText, Word2Vec ve Doc2Vec vektör model yöntemlerinin metin sınıflandırma çalışması üzerindeki doğruluk değerlerini karşılaştırmıştır. Metin sınıflandırma işlemi için Naive Bayes, Logistik Regresyon, Karar Ağaçları, Destek Vektör Makineleri ve Yapay Sinir Ağları sınıflandırma algoritmaları çalışmaya dahil edilmiş olup, karakter n-gram kullanımı ile test verisi genelleştirilmiş ve en yüksek doğruluk değeri olan %78 sonucuna Naive Bayes algoritması ile FastText Vektör Model Yöntemi ulaşmıştır [14].

3. Materyal ve Metotlar

Yapılan bu çalışma Pyhton Programlama dili kullanılarak Spyder IDE kodlama ortamında gerçekleştirilmiştir. Veri setinde yer alan kelimelerin analiz edilebilmesi ve daha düzgün bir şekilde sınıflandırma yapılabilmesi için kelimeler farklı ön işlemlerden geçirilmiştir. Bu ön işlemlerin içinde, kelimelerde ki tüm harflerin küçük harfe dönüştürülmesi, noktalama işaretlerini düzenlenmesi, kelimelerin köklerinin belirlenmesi ve kelimelerin etiketlenmesi gibi işlemler yapılmıştır. Veri ön işleme aşamalarından sonra veri setinde ki kelimeleri sayısal verilere dönüştürmek için CountVectorizer ve TF-IDF Vectorizer özellik çıkarım yöntemleri ile vektör uzay modelleri kullanılmıştır. Özellik çıkarım işlemleri gerçekleştirildikten sonra metin sınıflandırma aşamasında ise Lojistik Regresyon, Naive Bayes, Random Forests ve XGBoost makine öğrenmesi yöntemleri kullanılmıştır ve elde edilen doğruluk değerlerine göre bu modeller karşılaştırılmıştır.

3.1. Veri Seti

Çalışmada belirli e-ticaret sitelerinden alınan yorumlardan oluşturulmuş bir veri seti kullanılmıştır. Kullanılan bu veri setinin içerisinde toplamda 76478 veri bulunmaktadır. Bu veri setimiz İngilizce cümlelerden oluşmaktadır. Veri setinin içerisinde Tablo 1'de gösterildiği gibi 42133 tane Pozitif Cümle ve 34345 tane Negatif cümle bulunmaktadır. Oluşturduğumuz bu veri setini çalışmamıza uygun olarak kullanmak için belirli ön işleme aşamalarından geçirip çalışmamıza uygun hale getirilmiştir [22].

Tablo 1. Veri seti yorum sayıları

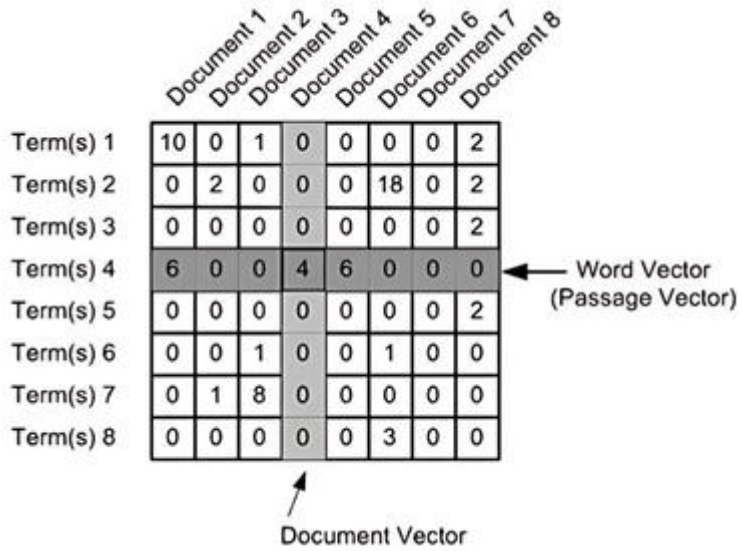
Cümle Etiketi	Yorum Sayısı
Pozitif etiketli cümle sayısı	42133
Negatif etiketli cümle sayısı	34345
Toplam cümle sayısı	76478

3.2. Özellik Çıkarım Yöntemleri

Çalışma kapsamında kullanılan veri setinde yer alan kelimeler üzerinde işlem yapılabilmesi için cümledeki kelimelerin tokenlarına ayrıştırılması gerekmektedir. Sonrasında makine öğrenme algoritmasına girişte kullanılacak kelimelerin, özellik çıkarım işlemi ile tamsayılar ya da kayan noktalı sayılar olarak kodlanması gerekmektedir. Genel olarak 3 özellik çıkarım yöntemi kullanılmaktadır. Bu yöntemler; metni kelime frekans vektörlerine çeviren TF-IDFVectorizer, kelime sayısı vektörüne çeviren CountVectorizer ve benzersiz tamsayılar haline çeviren HashingVectorizer yöntemleridir. Bu çalışmamızda TF-IDFVectorizer ve CountVectorizer yöntemleri kullanılmıştır.

3.2.1. Count Vectorizer

Metinlerden oluşan bir veri setini makinenin anlayacağı şekilde sayısal verilere çevirmek için birden fazla yöntem bulunmaktadır. Bu çalışmamızda bunlardan 2 tanesi ele alınarak karşılaştırma yapılmıştır. Bunlarda ilki CountVectorizer yöntemidir. CountVectorizer yönteminde veri setinde geçen her bir satır verideki benzersiz kelimelerden bir matris oluşturup bu matriste veri setinde ki ilgili kelimelerin geçme sıklığı yer almaktadır.



Şekil 2. Count Vector (DAYIBAŞI, 2018)

3.2.2. TF-IDF Vectorizer

Veri setimizi sayısal değerlere dönüştürmek için çalışmamızda kullandığımız ikinci yöntem ise TF-IDF Vectorizer yöntemidir. TF-IDF, metinlerin belgelerde ki matematiksel önemini belirlemek için kullanılan istatistiksel bir yöntemdir [15]. TF-IDF Vectorizer işleminde sözel ifadeye denk gelen değere 1 yerine TF-IDF değeri gelmektedir. TF-IDF değerini elde etmek için TF değeri ve IDF değerleri çarpılmaktadır. TF (Term

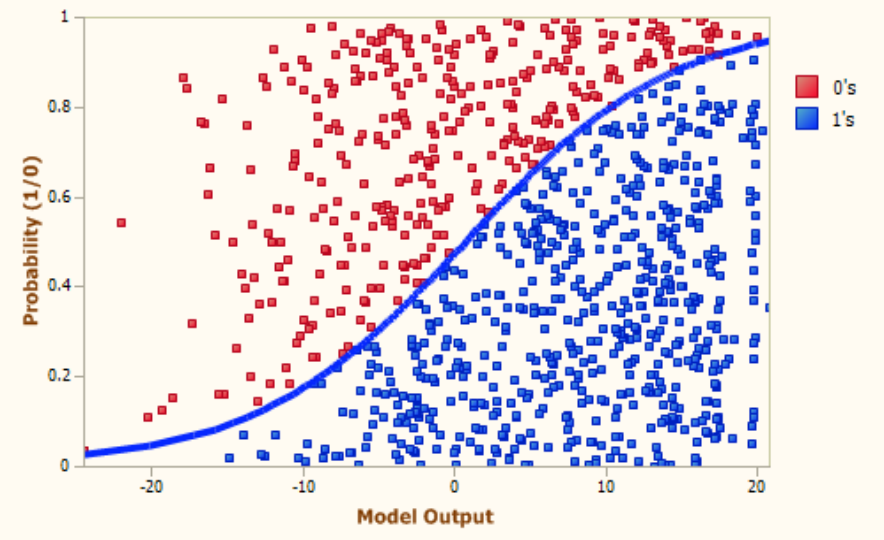
Frequency), En yalın ifadeyle terim sıklığı, belgede ki hedef eleman sayısının belgede ki toplam eleman sayısına oranıdır. IDF değeri ise, toplam belge sayısının hedef elemanın geçtiği belge sayısına oranının logaritması şeklinde ifade edilmektedir. Bu işlemde belge içinde kelimelerin kaç defa yer aldığının önemi yoktur. Kelimenin olup olmadığının bulunması yeterlidir. Elde edilen değerler ile vektörizasyon işlemi gerçekleştirilecek olursa ilk olarak her bir belge için bütün belgelerdeki benzersiz kelime sayısı kadar elemanlı bir vektör oluşturulur. Bu işlemler sırasında oluşabilecek karışıklığı engelleyebilmek için vektörleştirme işleminde TF-IDF değerlerine smoothing işlemi yapılmaktadır. En çok kullanılan yöntem elde edilen sonuç değerlerine 1 eklenmesi yöntemidir. Çözümlenmek istenen probleme göre sonrasında bu değerlere normalizasyon işlemi de yapılabilmektedir.

3.3. Sınıflandırma Modelleri

Çalışmamızda duygu analizi için kullanıp karşılaştırma yaptığımız 4 farklı makine öğrenme modeli kullanılmıştır. Bunlar Lojistik Regresyon, Naive Bayes, Random Forest ve XGBoost modelleridir.

3.3.1. Lojistik Regresyon Modeli

Lojistik Regresyon, İkili sınıfları tanımlamak için kullanılan istatistiksel bir modeldir. Lojistik Regresyon, sadece iki değer verebilen bir problemin sonuçlarının olasılığını tahmin edebilir. Yapılan bu tahmin, bir ya da birden fazla öngörücünün kategorik ve sayısal olarak kullanımına dayanmaktadır. Doğrusal Regresyon true/false, var/yok, evet/hayır gibi ikili sistem üzerinde ifade edilebilecek veriler için uygun değildir. Çünkü, Doğrusal Regresyon değer tahmininde 0 ve 1 aralığı dışında değerlerde tahmin edebilir. Lojistik Regresyon, 0 ile 1 arasında yer alan değerler ile sınırlı bir lojistik eğrisi oluşturur [15].



Şekil 3. Lojistik Regresyon (NAZLI, 2021)

3.3.2. Naive Bayes Modeli

Naive Bayes metin sınıflandırma modeli Bayes teoremine dayanır ve belirlenen sınıf için iletilen değerlerin gerçekleşme olasılığını bildirir [16]. Bu çalışmada kullandığımız Naive Bayes metin sınıflandırma modeli; sınıfı belirlenecek bir metnin diğer tüm sınıflara ait olma olasılığını, eğitim veri setinde bulunan metinlerden belirlenen olasılıklar içerisinde hesaplama yapmak için kullanılan gözetimli bir makine öğrenmesi yöntemidir. Naive Bayes algoritmasına ait formül

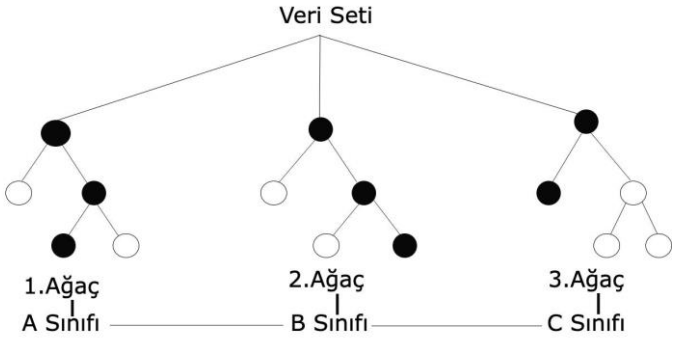
$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$$

Şeklindedir. Formülde yer alan değişkenler ise;
P (A | B) = B durumu olduğunda A durumunun olma ihtimali
P (B | A) = A durumu olduğunda B durumunun olma ihtimali

P (A) = A durumunun olma ihtimali
P (B) = B durumunun olma ihtimali şeklinde ifade edilmektedir [17].

3.3.3. Random Forest Modeli

Random Forest metin sınıflandırma modeli bir denetimli öğrenme algoritmasıdır. Random Forest hem regresyon için kullanılabilir hem de sınıflandırma yapmak için kullanılabilir. Random Forest algoritmasının kullanımı kolay ve esnek bir algoritma yapısı vardır. Rastgele ormanlar yine rastgele seçilen örnek verilerinde (D) karar ağaçları yapabilir, oluşturulan her ağaç için tahmin yapar ve aralarından oylama sonucu ile en iyi çözümü bulur [18].



Şekil 4. Random Forest Yöntemi (TAFRALI, 2022)

Dört aşama ile çalışır;

1. Veri kümesinden rastgele örnekler alır. (D_1, D_2, D_3, D_k.)
2. Aldığı her bir örnek için D_i karar ağacı oluşturur ve oluşturulan her karar ağacından tahmin sonucu gelir.
3. Gelen tahmin sonuçları için oylama gerçekleştirilir.

$$y = mode\{h_1(x), h_2(x), h_3(x), h_4(x)\}$$

$h_i(x)$ i. karar ağacının tahmin ettiği sınıftır.

$$y = \frac{1}{k} + \sum_{i=1}^{k} (h_i(x)\ )$$

$h_i(x)$, i. karar ağacı sınıftır.

4. Seçilecek olan tahmin, oylama sonucu en çok oy alarak belirlenir.

3.3.4. XGBoost Modeli

XGBoost Modeli, bir denetimli öğrenme probleminde kullanılması adına verileri eğitim ve test verisi olacak şekilde ayırarak kullanan bir algoritmadır. XGBoost modeli temel olarak gradyan arttırma ile ilgilidir, ölçeklenebilir ve etkilidir [18]. XGBoost gibi tabanı karar ağacı olan algoritmalar optimizasyon yapılarak geliştirilmiştir. XGBoost modelinin temel amacı loss (kayıp) fonksiyonunun değerini mümkün olduğunca azaltıp oluşturulacak karar ağaçlarını daha iyi bir hale getirmektir. Bu şekilde diğer modellere göre daha etkilidir ve çözüme ulaşma süresi daha kısadır [20].

Şekil 5. Karar Ağacı Algoritmalarının XGBoost'a Evrimi [21]

4. Deneysel Sonuçlar

Günümüz bilgi çağının en büyük kaynağı olarak kabul edilen internette, istenilen veriye ulaşmanın hızlı ve doğru olması ve bilginin erişebilirliğinin kolay olması gerekmektedir. Bu konuda özellik çıkarımı ve metin sınıflandırma yöntemleri kullanılmaktadır. Bu çalışmada belirli e-ticaret sitelerinden alınan yorumlar literatürden farklı olarak, yüksek doğruluk derecelerine sahip, farklı metin sınıflandırma yöntemleri ile karşılaştırılmış 4 yöntem birlikte incelenmiştir. Lojistik Regresyon, Naive Bayes, Random Forest ve XGBoost yöntemleri ile metin sınıflandırma işleminde tüm yöntemler başarıya ulaşmış ve seçilen tüm yöntemlerin doğruluk değerleri Tablo2'de karşılaştırılmıştır. Naive Bayes algoritmasının Count Vektör yöntemiyle elde edilen %84,9'luk doğruluk değeri diğer algoritmaların kullandığımız Count Vektör ve TF-IDF yöntemleri ile elde ettiğimiz doğruluk değerlerine oran ile daha yüksek bir değer olduğu gözlemlenmiştir. Count Vektör yönteminden sonra en iyi sonucu Lojistik Regresyon algoritması vermiştir. Çalışmamızda kullanıp karşılaştırmasını yaptığımız 4 yöntem arasında en düşük doğruluk değerine sahip algoritma ise Xgboost algoritması olmuştur. Çalışmamızda kullandığımız İngilizce metinlerden oluşan veri setimizi sayısal verilere dönüştürmek ve vektör uzay modeli oluşturmak için kullandığımız Count Vektör ve TF-IDF Vektörün 3 farklı modülünü de incelediğimizde, bunların arasında Count Vektör ile oluşturduğumuz model üzerinde uyguladığımız Makine Öğrenmesi yöntemlerinin daha yüksek doğruluk değeri verdiği görülmüştür. TF-IDF Vektör yönteminde kullandığımız 3 farklı modül olan kelime bazında, n-gram ve karakter bazında modülleri arasında ise TF-IDF Vektörünün kelime bazında verdiği doğruluk değerlerinin kullandığımız bu veri setinde diğer TF-IDF modüllerine göre daha yüksek bir doğruluk değeri verdiği gözlemlenmiştir.

Tablo 2. Yöntemlerin Doğruluk Değerleri

	Count Vektör	Tf-idf (kelime bazında)	Tf-idf (n-gram)	Tf-idf (karakter bazında)
Lojistik Regresyon	0,824	0,816	0,522	0,518
Naive Bayes	0,849	0,831	0,547	0,540
Random Forest	0,811	0,796	0,516	0,511
Xgboost	0,638	0,625	0,510	0,508

5. Tartışma ve Öneri

Bu araştırma, metin sınıflandırma modellerinin duygu analizi alanında etkinliğini incelemek üzere gerçekleştirilmiştir. Deneysel süreçler, belirli e-ticaret sitelerinden elde edilen ve yorumlardan oluşan veriler kullanılarak farklı özellik çıkarım ve makine öğrenmesi yöntemlerinin performanslarını karşılaştırmayı amaçlamaktadır. Raporlanan bulgular, çalışmanın önemli sonuçlarını ortaya koymaktadır.

Veri setinin ön işleme aşamasının, sonuçları üzerinde belirgin bir etkisi olduğu görülmektedir. Özellikle, CountVectorizer yönteminin kullanımı, Naive Bayes algoritmasıyla elde edilen yüksek doğruluk değerleriyle

dikkat çekmektedir. Bu sonuçlar, veri setinin önceden işlenmesinin ve özellik çıkarımının sınıflandırma performansı üzerinde kritik bir rol oynadığını göstermektedir.

Ek olarak, farklı özellik çıkarım yöntemlerinin ve makine öğrenmesi algoritmalarının kombinasyonlarının doğruluk üzerindeki etkisi detaylı bir biçimde incelenmiştir. Naive Bayes algoritması, CountVectorizer ile birlikte kullanıldığında en yüksek doğruluk değerlerini vermiştir. Söz konusu bulgular, belirli veri setleri için en uygun özellik çıkarımı ve sınıflandırma algoritmalarının seçilmesinin önemini vurgulamaktadır.

Çalışmanın raporlanan sonuçlarına göre, araştırmacılar için bazı öneriler sunulabilir. Örneğin, farklı endüstri veya dil verileri üzerinde benzer analizler gerçekleştirmek, bu çalışmanın bulgularını genelleştirmeye yardımcı olabilecektir. Ayrıca, daha karmaşık başka makine öğrenmesi yöntemlerinin ve derin öğrenme tekniklerinin bu alandaki etkinliğini araştırmak da ilginç olabilir.

Özetle, bu araştırma, metin sınıflandırma ve duygu analizi alanındaki mevcut yöntemlerin etkinliğini değerlendirerek önemli bir katkı sağlamıştır. Gelecekteki çalışmalar, bu alanın daha da gelişmesine ve endüstriyel uygulamalarda da kullanılmasına ışık tutabilir.

6. Sonuç

Çalışmamızda yüksek doğruluk derecelerine sahip 4 farklı metin sınıflandırma yöntemi kullanışmış olup tüm yöntemlerin uygulanması başarılı şekilde sağlanmıştır. Kullanılan yöntemlerin doğruluk değerleri Tablo2'de belirtilmiştir. Çalışmada kullanılan yöntemler arasında Naive Bayes algoritması ile elde edilen doğruluk değerlerinin diğer metin sınıflandırma algoritmalarına göre daha yüksek olduğu görülmüştür. Vektör uzay modeli oluşturmak için kullandığımız Count Vektör ve TF-IDF Vektörün 3 farklı modülü olan kelime bazında, n-gram ve karakter bazında modülleri de incelediğimizde Naive Bayes algoritmasının Count Vektör yöntemiyle elde edilen doğruluk değerinin diğer vektör uzay modeli oluşturmak için kullanılan yöntemlere göre daha yüksek olduğu gözlemlenmiştir. Çalışmamızın sonucu olarak kullandığımız veri seti üzerinde yaptığımız bu metin sınıflandırma çalışması, Naive Bayes algoritmasının Count Vektör yöntemiyle elde edilen %84,9'luk doğruluk değeri ile çalışmada ki en başarılı sonuç olduğu gözlemlenmiştir.

Kaynaklar

[1] B. Liu, Sentiment analysis and opinion mining. Springer Nature, 2022.

[2] TUZCU S. Çevrimiçi Kullanıcı Yorumlarının Duygu Analizi ile Sınıflandırılması. Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi. 2020;1(2):1-5.

[3] A. H. Aliwy ve E. H. Abdul Ameer, "Comparative Study of Five Text Classification Algorithms with their Improvements", International Journal of Applied Engineering Research, 2017.

[4] Sevindi, B. İbrahim. "Türkçe Metinlerde Denetimli ve Sözlük Tabanlı Duygu Analizi Yaklaşımlarının Karşılaştırılması," Yüksek Lisans Tezi, 2013.

[5] Sadhasivam, Jayakumar & Babu, Ramesh. (2019). Sentiment Analysis of Amazon Products Using Ensemble Machine Learning Algorithm. International Journal of Mathematical, Engineering and Management Sciences. 4. 508-520. 10.33889/IJMEMS.2019.4.2-041.

[6] H. Alshalabi, S. Tiun, N. Omar, M. Albared, "Experiments on the Use of Feature Selection and Machine Learning Methods in Automatic Malay Text Categorization", Science Direct, Procedia Technology, Elsevier, 2013.

[7] Amasyali, M. F., & Yildirim, T. (2004, April). Automatic text categorization of news articles. In Proceedings of the IEEE 12th Signal Processing and Communications Applications Conference, 2004. (Pp. 224-226). IEEE.

[8] Tüfekci, P., Uzun, E., & Sevinç, B. (2012, April). Text classification of web based news articles by using Turkish grammatical features. In 2012 20th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

[9] C. Sahoo, M. Wankhade, and B. K. Singh, "Sentiment analysis using deep learning techniques: a comprehensive review,"Int. J. Multimed. Inf. Retr., vol. 12, no. 2, p. 41, Dec. 2023, doi: 10.1007/s13735-023-00308-2.

[10] S. Kutabish, A. M. Soares, and B. Casais, "The Influence of Online Ratings and Reviews in Consumer Buying Behavior: A Systematic Literature Review," 2023, pp. 113–136. doi: 10.1007/978-3-031-42788-6_8.

[11] O. Ozturk and A. Ozcan, “Sentiment Analysis in Turkish Using Transformer-Based Deep Learning Models,” 2023, pp. 1–15. doi: 10.1007/978-3-031-31956-3_1.

[12]P. Nandwani and R. Verma, “A review on sentiment analysis and emotion detection from text,” Soc. Netw. Anal. Min., vol. 11, no. 1, p. 81, Dec. 2021, doi: 10.1007/s13278-021-00776-6.

[13] Erdinç, H. Y., & Güran, A. (2019, April). Semisupervised Turkish Text Categorization with Word2Vec, Doc2Vec and FastText Algorithms. In 2019 27th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

[14] Aizawa Akiko, (2003). “An information-theoretic perspective of tf–idf measures”. Information Processing and Management. 39 (1), 45–65. doi:10.1016/S0306–4573(02)00021–3

[15] D. G. Kleinbaum ve M. Klein, “Logistic Regression: A Self-Learning Text (Statistics for Biology and Health)”, Third Edition. New York: Springer 2010.

[16] H. Deng, Y. Sun, Y. Chang, J. Han, “Probabilistic Models for Classification” C.C. Aggarwal (Eds.), Data Classification Algorithms and Applications (pp. 67-70), CRC Press, New York, USA, 2015.

[17] Akdağlı, E. (2021). “Makine öğrenmesinde naive bayes yöntemi”. Medium.

[18] G. Louppe, “Understanding Random Forest”, doktora tezi, University of Liege, 2015.

[19] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794), 2016.

[20] Demolli H, Dokuz AS, Ecemis A, Gokcek M. Wind power forecasting based on daily wind speed data using machine learning algorithms. Energy Conversion and Management, 198, 111823, 2019.

[21] A. C. KELLE and H. YÜCE, “MQTT Trafiğinde DoS Saldırılarının Makine Öğrenmesi ile Sınıflandırılması ve Modelin SHAP ile Yorumlanması,” J. Mater. Mechatronics A, vol. 3, no. 1, pp. 50–62, Jun. 2022, doi: 10.55546/jmm.995091.

[22]willzjc, “Word Cloud Dataset.” 2024. [Online]. Available: https://gist.github.com/willzjc/3523f6ecc0a618efaecd4e2183b7efcd