Türkçe Metinlerde Makine Öğrenmesi Algoritmalarının Duygu Analizi Problemi Üzerindeki Performansının İncelenmesi

Investigation of the Performance of Machine Learning Algorithms on Sentiment Analysis Problem in Turkish Texts

Sühenda Hilal ETO1, Süha KÜÇÜKALİ2, Mehmet Kaan EROL3 Emine Büşra SALİHOĞLU4 Ömer Faruk KOÇAL5

1,2,3,4,5*Bilgisayar Mühendisliği Bölümü, Marmara Üniversitesi, İstanbul, Türkiye*

{hilaleto, suhakucukali, kaanerol, busrasalihoglu, omerkocal}@marun.edu.tr

***Özetçe*—***Günümüzde gittikçe yaygınlaşan sosyal medya kullanımı ile duygular ve fikirler bu platformlar üzerinden ifade edilmektedir. Bu platformlarda paylaşılan fikirler ile büyük miktarda veri ortaya çıkmaktadır. Bu verilerin sınıflandırılmasının ve analizinin manuel olarak yapılması büyük bir iş gücü gerektirdiğinden bazı algoritmalar ile duygu analizi yapılması gereksinimi ortaya çıkmıştır. Bu çalışmada çeşitli platformlardan alınan beş farklı veri kümesi ve her bir veri kümesi için dört farklı makine öğrenmesi algoritması(KNN, Naive Bayes, Rassal Orman, DVM) kullanılmıştır. Çalışma sonucunda DVM algoritması ile veri setlerinin genelinde daha doğru sonuçlar,  Rassal Orman ve Naive Bayes algoritmaları ile veri setleri ve eğitim yüzdelerine göre değişken sonuçlar elde edilmiştir. KNN algoritması ile veri setlerinin genelinde doğruluğu en düşük sonuçlar elde edilmiştir.*

***Anahtar Kelimeler—Duygu Analizi, Veri Madenciliği, Makine Öğrenmesi.***

***Abstract*—***Recently, with the use of social media, which is becoming more and more widespread today, emotions and ideas are expressed through these platforms. Huge amounts of data emerge with ideas shared on these platforms. Since the classification and analysis of these data requires a large labor force, the need for sentiment analysis with some algorithms has emerged. In this study, five different datasets from various platforms and four different machine learning algorithms (kNN, Naive Bayes, Random Forest, SVM) were used for each dataset. As a result of the study, more accurate results were obtained in general with the SVM algorithm, and variable results were obtained with the Random Forest and Naive Bayes algorithms according to the data sets and training percentages. With the KNN algorithm, the lowest accuracy results were obtained across the data sets.*

***Keywords—Sentiment Analysis, Data Mining, Machine Learning.***

# GİRİŞ

Duygu analizi, herhangi bir metin yazarının görüşlerini, duygularını ve tutumlarını anlamak için kullanılan bir çalışma alanıdır. Duygu analizinde, metin belirli cümlelere veya kelimelere bölünerek analiz edilir ve metnin içeriği olumlu, olumsuz veya nötr olarak sınıflandırılır. Duygu analizinde, yazarın herhangi bir ürün, olay, kişi veya fikir hakkında ifade ettiği görüşleri analiz etmek amaçlanır.

Sosyal medya araçları insanların farklı sosyal medya kanalları aracılığıyla ürünler, toplumsal ve politik olaylar, farklı fikirler, düşünceler hakkında yorum ve değerlendirme yapabilmesi, bunlara internet erişimi olan herkesin kolayca erişebilmesi, metin ve duygu analizi için işlenmemiş bir veri kaynağıdır.

Bu veriler gerçek toplum hakkında bilgi edinebilmek için üreticiler, politikacılar, sosyal bilimciler ve diğer insan ve toplumu ele alan araştırmalarda çok değerli olabilmektedir. Bu platformlarda ortaya çıkan ve gün geçtikçe artan ham veri yığınları bu amaçlar için kullanılmaya hazır değildir. Duygu analizi ile veriler amaca uygun olarak işlenebilir ve bahsedilen araştırmalar için değerli hale getirilebilir.

Verilerin anlamlı hale getirilebilmesi için bilinen birçok yöntem vardır. Manuel veya bir algoritma yardımı ile sınıflandırma yapılabilir. Manuel sınıflandırma yapmak yüksek iş gücü ve zaman kaybı ortaya çıkarabilmektedir. Makine öğrenmesi algoritmaları, daha kısa sürede, daha az iş gücü ile doğruluğu yüksek işlenmiş veri kümeleri oluşturulabilir.

# İLGİLİ ÇALIŞMALAR

Bu kısımda metin ve duygu analizi ile ilgili daha önce yapılmış çalışmalara yer verilmektedir.

Kaynar O. vd. yapmış oldukları çalışmada duygu analizinde öznitelik düşürme yöntemlerinin oto kodlayıcı derin öğrenme makinaları ile karşılaştırılması yapılmıştır. Metin analizinde frekansı düşük olan kelimelerin cümle vektörlerinde bolca sıfır verisi bulundurması makine öğrenme algoritmalarının verimini düşürdüğü için boyut küçültme uygulanmıştır. Doğrusal ve doğrusal olmayan gözetimsiz makine öğrenmesi tabanlı boyut düşürme algoritmaları kullanılmıştır. IMDB film yorumları veri seti 1000 pozitif ve 1000 negatif yorum olacak şekilde dengeli olarak ayrılmıştır.  Kullanılan kütüphane olarak her bir kelimenin kökleri NLTK kütüphanesi yardımıyla bulunmuştur. Deney sonucunda boyut düşürme teknikleri uygulanan veri setlerinden daha yüksek başarı elde edilmiştir.  Boyut düşürme yöntemleri (TDA, TBA, çTBA, FA) uygulanmadan makine öğrenme algoritmaları (DVM, YSA) kullanılan sonuç ile boyut düşürme yöntemleri uygulanarak yapılan sonuç ile karşılaştırma yapılmıştır [1].

Akgül E. vd. yapmış olduğu çalışmada Twitter verileri ile duygu analizi yapılmıştır. Duygusal Twitter adındaki bu çalışmada bir anahtar kelime belirlenerek her bir tweete ait duygu analizi olumlu, olumsuz veya nötr olarak etiketlenmeye çalışılmış, bu sayede bir sözlük elde edilmesi amaçlanmıştır. Sözlük ve n-gram olmak üzere iki adet farklı yöntem kullanılmıştır. Tarafsız bakış açısı yakalayabilmek adına belirlenen iki kişi tweetleri olumlu, olumsuz veya nötr şeklinde etiketlenmiştir. Daha sonra veri setlerinden duygu analizi için gereksiz karakterler temizlenmiş, kelimelerin tekrar sayısı ile frekansları bulunup, her veri seti için farklı kelime frekansları hesaplanmıştır. N-gram modelinde ise 2, 3 ve 4 gramlar kullanılarak sistem ayrı ayrı test edilmiştir. Deney sonuçlarında en fazla başarı en çok tweet sayısına sahip VS3 veri setinden alınmıştır. 3 farklı veri seti 2 farklı etiketleme yöntemiyle karşılaştırılmıştır [2].

Tuzcu S. yaptığı çalışmada Türkçe çevrimiçi kullanıcı yorumlarının yüksek doğrulukla yönlü sınıflandırılması yapılmak istenmiştir. Python programlama dili ve rapid miner üzerinde veri ön işleme, makine öğrenmesi yöntemlerinin uygulanması gerçekleştirilmiştir. Veri ön işleme işleme için NLTK kütüphanesi kullanılmıştır. Çalışmada MLP, NB, DVM, LR makine öğrenmesi yöntemleri aynı veri seti üzerinde sınıflandırma başarılarına göre karşılaştırılmıştır. Bu karşılaştırma sonucunda MLP pozitif ve ortalama sınıflandırma başarısında ön plana çıkmış, NB ise diğer yöntemlere göre negatif sınıflandırmada pozitife göre daha başarılı sonuç vermiştir [3].

Aytekin Y. E. ve Keskin Ö. yapmış oldukları çalışmada faizsiz finans sistemi ile ilgili türkçe metinler üzerinden duygu analizi yöntemi ile potansiyel müşterilerin faizsiz finans sitemine yönelik pozitif ve negatif algıları anlaşılmak istenmiştir. Veri seti 2019 Ocak ayına ait faizsiz finans sistemi ile ilgili türkçe metinlerden oluşmaktadır. Bu metinler sosyal medya ve internet kaynaklarına aittir. Bu çalışmada teknik herhangi bir ayrıntı paylaşılmamıştır. Çalışmada analiz kısmı [www.mediatoolkit.com](http://www.mediatoolkit.com) adresinde bulunan çevrimiçi program ile yapılmıştır. Yazarların vardığı sonuç şu: “İslami temellere dayanan faizsiz finans sistemindeki kurumlar, sömürü mantığıyla ortaya çıkan konvansiyonel rakiplerinden; üretimi, paylaşımı ve ahlaki değerleri ön planda tutacak şekilde ayrışmalıdır. Bu farklılığın göstergesi olarak en azından “banka” terimi yerine “finans” teriminin kullanılması daha doğru olacaktır.” doğrudan alıntılanmıştır [4].

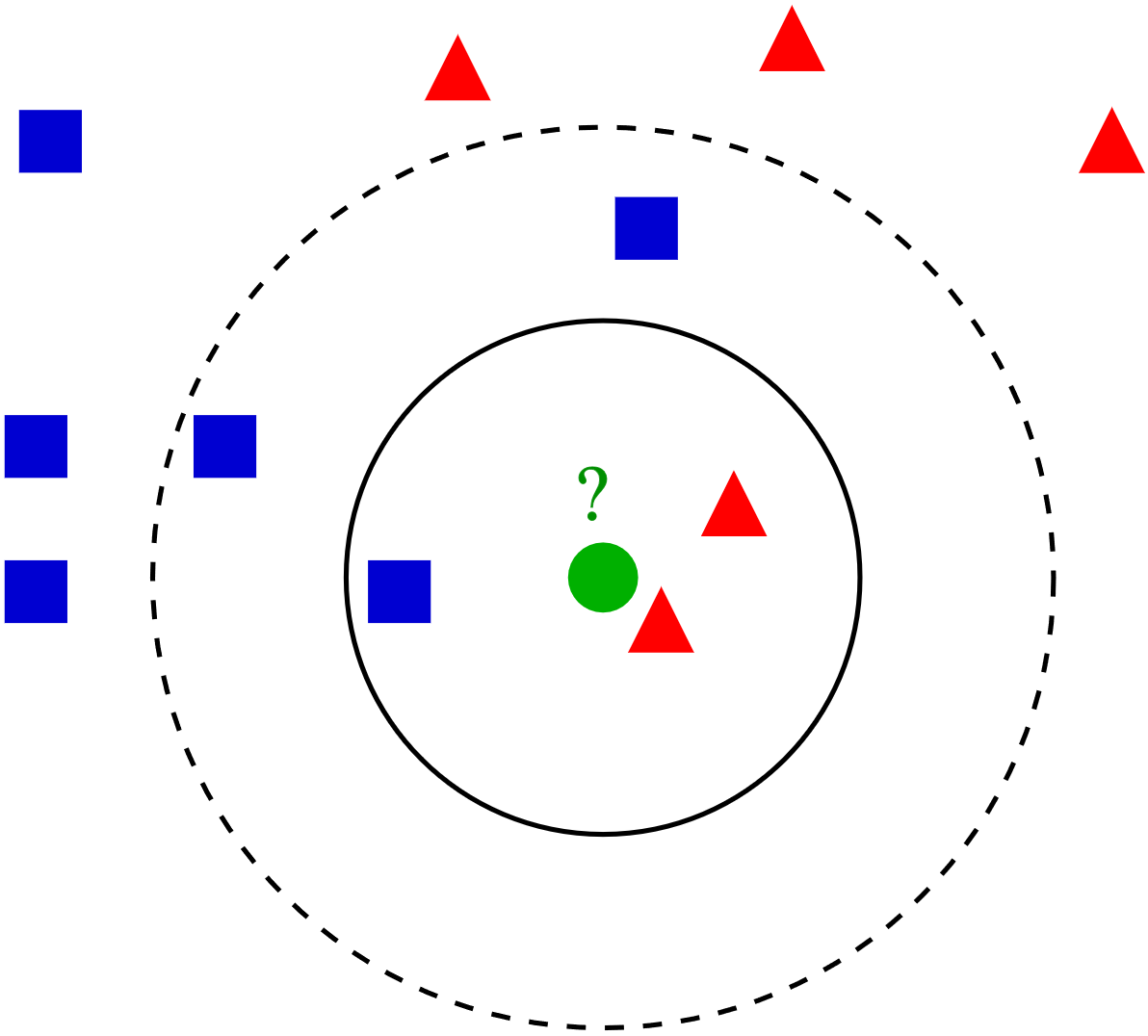
Eryılmaz E. E. vd. yapmış oldukları çalışmada reklam, alakasız içerik barındıran ve kişiye zarar veren yaramaz epostaların makine öğrenmesi yöntemleri ile tespiti amaçlanmıştır.  2 tane Türkçe veri seti kullanılmıştır. Bunlardan biri  "TurkishEmail" veri kümesi diğeri ise yazarlar tarafından hazırlanan "TrHamSpamEmailv1.0" veri setidir.  İlk olarak veri setlerinde öznitelik çıkarımı ve özellik seçimi yapılmıştır. Veri setlerinde çalışılan makine öğrenme algoritmaları RF, C4.5 karar ağacı, SMO, KNN, LR, NB, MLP'dir. Veri kümeleri ön işlemden geçerken işleme tabii tutulan her sözcük Türkçe doğal dil işleme kütüphanesi olan "Zemberek" yazılımından geçirilip köklerine ayrıştırılmıştır. Elde edilen sonuçlara göre %98'in üzerinde bir orana kadar yaramaz e-postaların filtrelenebildiği görülmüştür [5].

# UYGULAMA

Duygu analizi (veya fikir madenciliği), verilerin olumlu, olumsuz veya tarafsız olup olmadığını belirlemek için kullanılan bir doğal dil işleme tekniğidir. Duyarlılık analizi, işletmelerin müşteri geri bildirimlerinde marka ve ürün duyarlılığını izlemelerine ve müşteri ihtiyaçlarını anlamalarına yardımcı olmak için genellikle metin tabanlı veriler üzerinde gerçekleştirilmektedir. Bu amaçla duyarlılık analizi için farklı öğrenme modelleri kullanılmaktadır [6]. Bu çalışmada kullanılan modeller bu bölümde daha ayrıntılı olarak ele alınmıştır.

1. *KNN (K-En Yakın Komşu)*

KNN algoritması temel olarak, sınıflandırılmak istenen bir veriyi daha önceki verilerle olan komşuluk ilişkisine göre sınıflandıran bir algoritmadır. KNN ile yeni noktaya en yakın noktalar aranır. K tane, bilinmeyen noktanın en yakın komşularını temsil eder [7]. Analitik olarak izlenebilir olması, yerel bilgilere uyarlanabilir olması, tutarsız eğitim verilerine karşı dirençli olması KNN algoritmasının avantajlarındandır. KNN algoritmasının dezavantajlarına yüksek miktarda bellek alanlarına ihtiyaç duyması, veri seti boyutu arttıkça işlem yükünün artması örnek gösterilebilir.



Şekil 1. KNN algoritması sınıflandırma örneği [15]

Şekil 1’ de örnek bir KNN sınıflandırması gösterilmiştir. Şekildeki yeşil nokta, mavi karelere veya kırmızı üçgenlere göre sınıflandırılmaktadır. Şekil 1’de K=3 olduğu durumda, daire içerisindeki duruma bakılır. Kırmızı üçgenlerin sayısı mavi karelerden fazla olacağından sınıflandırma kırmızı üçgenlere göre yapılır. Ancak K=5 olduğu durumda, kesikli daire içerisindeki duruma bakılır ve kırmızı üçgenlerin sayısı mavi karelerden daha az olduğundan sınıflandırma mavi karelere göre yapılır [8].

1. *Naive Bayes*

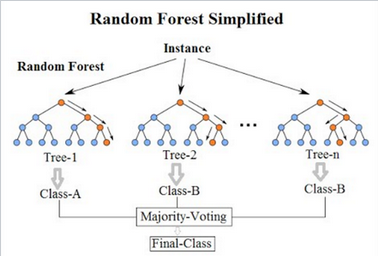
Naive Bayes yöntemi, değişkenlerinin değeri verilen her özellik çifti arasında koşullu bağımsızlık varsayımıyla uygulanmasına yönelik bir dizi denetimli öğrenme algoritmasıdır [9]. Naive Bayes modelinin kullanımı kolaydır ve büyük veri kümeleri için kullanışlı olup karmaşık sınıflandırma yöntemlerine kıyasla daha verimli olduğu bilinmektedir. Şekil 2’de Bayes Teoreminin sonlu olasılık denklemi gösterilmektedir.

(1)

Denklem (1)’ de sonlu olasılığı, verilen sınıfın tahmin edicisinin olasılığı olan olabilirliktir. tahmin edicinin önceki olasılığını ifade etmektedir [10].

1. *Rassal Orman*

Rassal Orman, davranış analizinde tahminleri modellemede kullanılan bir tekniktir ve karar ağaçları üzerine kuruludur. Bir dizi regresyon karar ağacının sonuçlarını birleştirerek çıktı tahminleri yapan makine öğrenmesi modelidir [11]. Her ağaç bağımsızdır ve ormandaki tüm ağaçlar aynı dağılıma sahip olup, girdi verilerinden örneklenen rastgele vektöre bağlıdır [12]. Rassal Orman algoritması her ağacın rastgele vektör değerlerine bağlı olduğu bir ağaç tahmincisi görevi görür. Büyük verileri sınıflandırmak için en iyi algoritmalardan bir tanesidir.

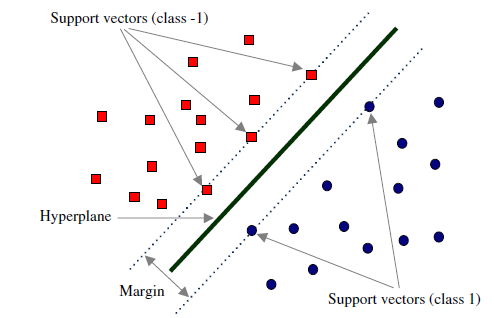


Şekil 2. Rassal Orman modeli için örnek bir karar şeması [16]

Şekil 2’de üç adet karar ağacı gösterilmiştir. Bir karar ağacında, en üstte bir girdi girilmekte ve ağaçta aşağıya inildikçe veriler daha küçük kümeler halinde toplanmaktadır. Tek bir karar ağacına güvenmek yerine her ağaçtan tahmin alınmakta ve tahminlerin çoğunluk oylarına göre sonuç çıktısı tahmin edilmektedir.

1. *DVM (Destek Vektör Makinesi)*

DVM sınıflandırma ve regresyon problemleri için kullanılan bir makine öğrenmesi modelidir. Öncelikli olarak sınıflandırma problemleri için kullanılmaktadır. DVM, yeni veri noktasını kolayca tespit edebilmek için n-boyutlu uzayı sınıflara ayırabilen en iyi çizgi veya karar sınırını oluşturmaktadır. Bu en iyi karar sınırına hiperdüzlem denmektedir. DVM, hiper düzlemi oluşturmaya yardımcı olan vektörleri seçmekte ve “Destek Vektör Makinesi” olarak adlandırılmaktadır [13].



Şekil 3. Destek vektör makinelerinin küme ayrışımını temsil etmekte [17]

Şekil 3’te bir karar sınırı veya hiperdüzlem kullanılarak sınıflandırılan iki farklı kategorinin bulunduğu görülmektedir.

Yüksek boyutlu uzaylarda etkili kümeleme yapması ve belleği daha etkin kullanması DVM’in avantajları arasındadır. Boyutların sayısının örnek sayısından fazla olduğu durumlarda DVM daha etkili olmaktadır. Ancak büyük veri kümeleri için uygun değildir. Veri kümesine oranla gürültü (signal-to-noise ratio) oranı arttıkça DVM sınıflandırma performansı düşmektedir [14].

# SONUÇLAR VE DEĞERLENDİRME

Bu çalışmada makine öğrenmesi modellerinin uygulanması için scikit-learn, veri setinin manipülasyonu, numerik hesapla, değerlendirme süreçlerinde pandas, numpy ve sci-py kütüphaneleri kullanılmıştır. Çalışmada beş farklı veri set üzerinde, dört farklı makine öğrenmesi modeli ile duygu analizi yapılmıştır.

Tablo 1’deki veri seti turizm başlığı altında atılan toplam 4202 tweet’ten oluşmaktadır. Bu veriler olumlu ve olumsuz olmak üzere 2 sınıfa ayrılmıştır. Tablo 2’deki veri seti satın alınan ürünler hakkında yapılan yorumlardan oluşmakta ve olumlu, olumsuz ve tarafsız olarak sınıflara ayrılmıştır. Toplam 8492 adet mağaza yorumuna ait duygu ifadelerini içermektedir. Tablo 3’teki veri seti satın alınan ürünler hakkında yapılan yorumlardan oluşmakta ve olumlu, olumsuz ve tarafsız olarak sınıflara ayrılmıştır. Toplam 11430 adet mağaza yorumlarının duygu ifadesi içermektedir. Tablo 4’teki veri seti çeşitli blog yazarlarının neşeli, sinirli, üzgün, karışık olmak üzere 4 farklı sınıfta incelendiği toplamda 157 blog yazısı içermektedir. Tablo 5’teki veri kümesi kullanıcıların izledikleri film hakkında yapılan yorumlardan oluşmakta ve pozitif, negatif ve tarafsız olarak sınıflara ayrılmıştır. Her sınıfta 35’er adet olmak üzere toplamda 105 adet film yorumu bulunmaktadır.

Deney sonuçlarına göre DVM algoritması tüm veri setlerinde diğer algoritmalara göre daha yüksek, KNN algoritması ise daha düşük doğrulukta sınıflandırma başarısına sahip olmuştur. Rassal Orman ve Naive Bayes algoritmaları ise belirgin olarak daha başarılı sonuç verdikleri veri setleri bulunmasada DVM’ye yakın sonuçlar vermişlerdir. Ek olarak gözetimli öğrenme modelleri olan DVM, Rassal Orman ve Naive Bayes algoritmaları, gözetimsiz öğrenme modeli olan KNN’e göre daha başarılı olmuştur.

İleri çalışmalarda derin öğrenme tabanlı sınıflandırma yöntemleri kullanılabilir. Duygu analizinde sınıflandırma yöntemleri kadar verinin doğru değerlendirilmesi de önemli olduğundan veri setlerine uygun metrikler kullanılabilir ve bunlar karşılaştırılabilir.

TABLO 1. Turizm hakkında tweet’lerden oluşan veri seti. Toplam 4202, 3044 olumlu, 1158 olumsuz

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algoritmanın İsmi** | %30 eğitim-  %70 test | | %40 eğitim-  %60 test | | %50 eğitim-  %50 test | | %70 eğitim-  %30 test | | Accuracy | F1 score |
| **RandomForest** | 0.722 | 0.838 | 0.725 | 0.840 | 0.726 | 0.841 | 0.725 | 0.840 | 0.724 | 0.839 |
| **NaiveBayes** | 0.702 | 0.796 | 0.713 | 0.803 | 0.712 | 0.802 | 0.722 | 0.811 | 0.712 | 0.803 |
| **KNN** | 0.687 | 0.788 | 0.700 | 0.794 | 0.708 | 0.801 | 0.725 | 0.813 | 0.705 | 0.799 |
| **SVM** | **0.738** | **0.845** | **0.741** | **0.845** | **0.740** | **0.844** | **0.742** | **0.844** | **0.740** | **0.844** |

TABLO 2. Mağaza yorumları hakkında tweet’lerden oluşan veri seti. Toplam 8492, 4253 olumlu, 4239 olumsuz

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algoritmanın İsmi** | %30 eğitim-  %70 test | | %40 eğitim-  %60 test | | %50 eğitim-  %50 test | | %70 eğitim-  %30 test | | Accuracy | F1 score |
| **RandomForest** | 0.740 | 0.678 | 0.717 | 0.642 | 0.727 | 0.661 | 0.706 | 0.625 | 0.722 | 0.651 |
| **NaiveBayes** | 0.754 | 0.756 | 0.751 | 0.755 | 0.748 | 0.755 | 0.751 | 0.763 | 0.751 | 0.757 |
| **KNN** | 0.560 | 0.580 | 0.572 | 0.599 | 0.589 | 0.625 | 0.725 | 0.813 | 0.705 | 0.799 |
| **SVM** | **0.874** | **0.869** | **0.878** | **0.875** | **0.883** | **0.881** | **0.890** | **0.890** | **0.879** | **0.787** |

TABLO 3. Mağaza yorumları hakkında tweet’lerden oluşan veri seti. Toplam 11430, 4253 olumlu, 4238 olumsuz, 2939 tarafsiz

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algoritmanın İsmi** | %30 eğitim-  %70 test | | %40 eğitim-  %60 test | | %50 eğitim-  %50 test | | %70 eğitim-  %30 test | | Accuracy | F1 score |
| **RandomForest** | 0.659 | 0.513 | 0.683 | 0.562 | 0.720 | 0.638 | 0.815 | 0.809 | 0.719 | 0.630 |
| **NaiveBayes** | 0.762 | 0.764 | 0.762 | 0.769 | 0.755 | 0.763 | 0.743 | 0.757 | 0.751 | 0.757 |
| **KNN** | 0.521 | 0.678 | 0.528 | 0.680 | 0.534 | 0.684 | 0.533 | 0.684 | 0.529 | 0.681 |
| **SVM** | **0.870** | **0.866** | **0.880** | **0.878** | **0.885** | **0.883** | **0.886** | **0.885** | **0.880** | **0.878** |

TABLO 4. Çeşitli blog yazarların 4 farklı ruh hallerinde yazdıkları toplam 157 blog yazısından oluşan veri seti. 40 Neşeli, 40 sinirli, 39 üzgün, 38 karışık

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algoritmanın İsmi** | %30 eğitim-  %70 test | | %40 eğitim-  %60 test | | %50 eğitim-  %50 test | | %70 eğitim-  %30 test | | Accuracy | F1 score |
| **RandomForest** | **0.4** | **0.47** | 0.33 | 0.43 | 0.329 | 0.42 | 0.35 | 0.43 | 0.352 | 0.437 |
| **NaiveBayes** | 0.35 | 0.41 | 0.34 | **0.46** | **0.36** | **0.44** | 0.5 | 0.52 | **0.387** | 0.457 |
| **KNN** | 0.35 | 0.47 | **0.36** | 0.37 | 0.3 | 0.41 | **0.533** | **0.684** | 0.385 | **0.482** |
| **SVM** | 0.3 | 0.44 | 0.3 | 0.41 | 0.25 | 0.35 | 0.2 | 0.34 | 0.262 | 0385 |

TABLO 5. Veri seti 3 farklı yorum sınıfına ait 35’er toplamda 105 film yorumu içermekte

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algoritmanın İsmi** | %30 eğitim-  %70 test | | %40 eğitim-  %60 test | | %50 eğitim-  %50 test | | %70 eğitim-  %30 test | | Accuracy | F1 score |
| **RandomForest** | 0.29 | 0.42 | 0.26 | 0.4 | 0.35 | 0.45 | 0.37 | 0.51 | 0.317 | 0.445 |
| **NaiveBayes** | **0.31** | **0.46** | **0.3** | **0.41** | **0.39** | **0.44** | **0.46** | **0.57** | **0.365** | **0.47** |
| **KNN** | **0.31** | 0.4 | 0.23 | 0.32 | 0.24 | 0.26 | 0.43 | 0.5 | 0.320 | 0.370 |
| **SVM** | 0.29 | 0.45 | 0.25 | 0.4 | 0.28 | **0.44** | 0.375 | 0.47 | 0.298 | 0.416 |

##### KAYNAKÇA

[1] Kaynar, O., Aydın, Z., Görmez, Y. “Sentiment analizinde öznitelik düşürme yöntemlerinin oto kodlayıcılı derin öğrenme makinaları ile karşılaştırılması,” *Bilişim Teknolojileri Dergisi,* vol. 10, no. 3, pp. 319-326, 2017.

[2] Akgül, E. S., Ertano, C., Diri, B. “Twitter verileri ile duygu analizi,” *Pamukkale University Journal of Engineering Sciences*, vol. 22, no. 2, 2016.

[3] TUZCU, S. “Çevrimiçi Kullanıcı Yorumlarının Duygu Analizi ile Sınıflandırılması,” *Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi*, vol. 1, no. 2, pp. 1-5, 2020.

[4] Aytekin, Y. E., Keskin, Ö. “Türkiye’de Faizsiz Finans Sisteminin Duygu Analizi Bağlamında Değerlendirilmesi,” *Uluslararası İslam Ekonomisi ve Finansı Araştırmaları Dergisi*, vol. 5, no. 3 pp. 87-112, 2019.

[5] ERYILMAZ, E. E., ŞAHİN, D. Ö., KILIÇ, E. “Türkçe İstenmeyen E-postaların Farklı Öznitelik Seçim Yöntemleri Kullanılarak Makine Öğrenmesi Algoritmaları ile Tespit Edilmesi,” *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, vol. 13, no. 2, pp. 57-77.

[6] Liu, Bing. "Sentiment analysis and subjectivity." *Handbook of natural language processing* 2.2010, 627-666,2010

[7] Sutton, Oliver. "Introduction to k nearest neighbour classification and condensed nearest neighbour data reduction." *University lectures, University of Leicester* 1 ,2012.

[8] Arroyo, Javier, and Carlos Maté. "Forecasting histogram time series with k-nearest neighbours methods." *International Journal of Forecasting* 25.1,192-207,2009.

[9] Harry Zhang, “The Optimality of Naive Bayes”, *Conference: Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, 2004

[10] Shah, K., Patel, H., Sanghvi, D., Shah, M. “A comparative analysis of logistic regression, random forest and KNN models for the text classification,” *Augmented Human Research,* vol. 5, no. 1, pp. 1-16, 2020.

[11] [Bianca Williams, Caroline Halloin, Wiebke Löbel,Ferdous Finklea,Elizabeth Lipke,Robert Zweigerdt,Selen Cremaschi](https://www.sciencedirect.com/science/article/pii/B9780128233771502743?via%3Dihub#!),” Data-Driven Model Development for Cardiomyocyte Production Experimental Failure Prediction”, vol. [48](https://www.sciencedirect.com/science/journal/15707946/48/supp/C), pp. 1639-1644, 2020

[12] Breiman, “Random Forests”, *Machine Learning*, no. 1, pp. 5-32, 2001.

[13] Platt, J. “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61-74, 1999.

[14] Lin, C. J., Chang, C. C. LIBSVM: a library for support vector machines, 2001.

[15]<https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/>

[16] <https://en.wikipedia.org/wiki/Random_forest>

[17]https://www.researchgate.net/figure/Classification-of-data-by-support-vector-machine-SVM\_fig8\_304611323