



Duygu Analizinde Denetimli Makine Öğrenme Algoritmalarının Karşılaştırılmaları, (Kahramanmaraş Depremi Örneği)

Emre BİÇEK, Erol KINA

Van Yüzüncü Yıl Üniversitesi Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Bilgisayar Bilimleri Anabilim Dalı, Van
Van Yüzüncü Yıl Üniversitesi Özalp Meslek Yüksekokulu, Bilgisayar Teknolojisi Bölümü, Bilgisayar Teknolojisi Programı, Van

Doi: 10.55024/buyasambid.1295878

Makale Bilgisi

Özet

Makale geçmişi:

İlk gönderim tarihi: 11.05.2023

Düzeltilme tarihi

Kabul tarihi: 08.06.2023

Yayın tarihi: 14.07.2023

Anahatar Kelimeler:

Duygu Analizi, Makine Öğrenmesi,
Kahramanmaraş Depremi.

* Sorumlu Yazar

E-mail address:

emrebicek@yyu.edu.tr

Orcid:

[0000-0001-6061-9372](https://orcid.org/0000-0001-6061-9372)

Duygu analizi kullanılarak, siyaset, ekonomi, spor, alışveriş gibi farklı alanlarda yapılmış birçok çalışma mevcuttur. Sosyal medya, insanların herhangi bir konu hakkındaki görüşlerini ve duygularını paylaştıkları önemli bir veri merkezidir. İnsanların sosyal medya platformlarında duygularını olumlu, olumsuz veya tarafsız olarak sınıflandırabilmekteyiz. Bu araştırma makalesinde, makine öğrenimi yaklaşımı altındaki denetimli öğrenme algoritmaları kullanıldı. Destek Vektör Makineleri (SVM), Naïve Bayes (NB) ve Oylama Sınıflayıcısı (Lineer Regresyon, Logistik Regresyon, Destek Vektör Makineleri) kullanılarak, algoritmalar karşılaştırıldı. Veri seti Twitter API kullanılarak 2023 yılı içerisinde Türkiye’de Kahramanmaraş merkezli 10 ili etkileyen depremle ilgili paylaşılmış 10.000 tweet’ten oluşturuldu. Veri seti kullanılmadan önce ön işleme adımları uygulandı. Etkisiz Kelimeleri kaldırma (Stop Words), @ kaldırma, Hashtag kaldırma, Metin parçası etiketleme (POS Tagging), duyarlılık puanı hesaplama işlemleri uygulandı. Daha sonra görüş tespitini (olumlu, olumsuz ve tarafsız olarak) sınıflandırmak için SVM, NB, Oylama Sınıflayıcısı (VC) denetimli makine öğrenme algoritmaları kullanıldı. Sonuç olarak Oylama Sınıflayıcısının Destek vektör makinesi ve Naïve Bayes algoritmasından çok daha iyi sonuç verdiğini tespit edildi. Twitter yorumlarının çoğunlukla olumsuz duygu taşıdığı ve önerilen Oylama Sınıflayıcısı modelinin %89,14 başarı oranı ile doğru tespit yaptığı görüldü. Bu çalışmada önerilen model ile, paylaşılan bir içeriğin olumlu ya da olumsuz duygu taşıyıp taşımadığı yüksek bir oranda doğru tahmin edilebilmektedir.

Comparison of Supervised Machine Learning Algorithms in Sentiment Analysis: (Case of Kahramanmaraş Earthquake)

Emre BİÇEK, Erol KINA

Van Yüzüncü Yıl Üniversitesi Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Bilgisayar Bilimleri
Anabilim Dalı, Van

Van Yüzüncü Yıl Üniversitesi Özalp Meslek Yüksekokulu, Bilgisayar Teknolojisi Bölümü, Bilgisayar
Teknolojisi Programı, Van

ARTICLE INFO

ABSTRACT

Article history:

Received: 11.05.2023

Received in revised form

Accepted: 08.06.2023

Available online: 14.07.2023

Key words:

Sentiment Analysis, Machine
Learning, Kahramanmaraş
Earthquake.

* Corresponding author.

E-mail address:

haticeiskender2011@hotmail.com

Many studies have been conducted using sentiment analysis in different fields such as politics, economy, sports, and shopping. Social media is an important data center where people share their opinions and emotions about any subject. We can classify people's emotions on social media platforms as positive, negative, or neutral. In this research paper, supervised learning algorithms under the machine learning approach were used. Support Vector Machines (SVM), Naïve Bayes (NB), and Voting Classifier (Linear Regression, Logistic Regression, Support Vector Machines) were compared using a dataset consisting of 10,000 tweets shared about the earthquake that affected 10 provinces centered in Kahramanmaraş in Turkey in 2023, obtained using the Twitter API. Preprocessing steps were applied to the dataset before use, including removing stop words, @ mentions, hashtags, performing part-of-speech tagging, and calculating sentiment scores. SVM, NB, and Voting Classifier (VC) supervised machine learning algorithms were then used to classify the sentiment (positive, negative, and neutral). As a result, it was found that the Voting Classifier model performed much better than the Support Vector Machines and Naïve Bayes algorithms. It was observed that most of the Twitter comments carried negative sentiments and the proposed Voting Classifier model correctly predicted sentiment with a success rate of 89.14%. With the proposed model in this study, it is possible to predict whether a shared content carries positive or negative emotions with a high accuracy rate.

1. GİRİŞ

Sosyal Medya insanların görüş ve önerilerini paylaştıkları açık ve ücretsiz en büyük veri kaynağıdır. Çevremizde meydana gelen olaylar, insanların yaşadıkları deneyimler, doğal afet, ekonomi, siyaset, spor, sanat gibi birçok alanda paylaşımlar sosyal medya platformlarında yapılmaktadır. İnsanlar bu konularda kendi görüş, öneri ve şikayetlerini kolaylıkla, en kısa sürede ve en fazla kişiye ulaştırabilmektedir. Twitter sosyal medya platformları arasında en popüler olanlarından biridir. Websiterating sitesinin verilerine göre, 17 Mart 2023 itibarıyla 1.3 Milyar Twitter hesabı olup 368.4 Milyon aktif kullanıcı mevcuttur ve her gün 500 milyonun üzerinde tweet paylaşılmaktadır. Dakikada yaklaşık olarak 350.000 tweet paylaşılan bu platform araştırmacılar için eşsiz bir veri kaynağıdır. Her beş Twitter yorumundan birisinde emoji kullanılmaktadır (Websiterating, 2023). Bu durum duygu analizi çalışmalarında karşılaşılan zorluklardan sadece birisidir. İnsanlar görüşlerini farklı ifade yollarıyla ifade etmektedir. Bir kelime kullanım şekliyle bir cümlede olumlu olurken başka bir cümlede olumsuz olarak kullanılabilir. Mesela “Sanki bütün bu olayların sorumlusu

benim.” Cümlesi anlamca olumsuz olmasına karşın biçimce olumsuz olmaktadır. Yine argo dilinin kullanılması, kısaltmalar duygu analizinde zorluklar çıkarmaktadır. İğneleyici sözler (Harika, yardımın gelmesi bir hafta sürdü), çok kutupluluk (Kargo sağlamdı ama içindekiler kırılmıştı.) gibi ifade biçimleri tweetlerin sınıflandırılmasında zorluklar çıkarabilmektedir.

Bu araştırma makalesinde, SVM, Naive Bayes ve VC olmak üzere üç tür makine öğrenimi algoritması kullanıldı. Doğrulukları, kesinlikleri ve geri çağırma değerlerini hesaplandı ve pozitif, negatif, tarafsız, sonuçlar elde edildi.

Kuram

Sosyal medya, insanların görüş ve önerilerini paylaştıkları açık ve ücretsiz bir veri kaynağıdır. Twitter gibi popüler sosyal medya platformları, araştırmacılar için eşsiz bir veri kaynağı sunmaktadır. Ancak, duygu analizi çalışmalarında emoji kullanımı, ifade biçimleri ve argo dilinin kullanımı gibi zorluklarla karşılaşmaktadır. Bu zorluklar, makine öğrenimi algoritmaları kullanılarak çözülebilmekte ve doğruluk, kesinlik ve geri çağırma değerleri hesaplanarak pozitif, negatif ve tarafsız sonuçlar elde edilebilmektedir.

İlgili Çalışmalar

Duygu analizinde insanların birçok konuda farklı fikirleri sınıflandırılmaktadır. Duygu analizi için, metin, ses, resim video gibi veri kaynaklarından yararlanılan çalışmalar literatürde mevcuttur. Bu bölümde Twitter verileri kullanılarak yapılan benzer çalışmalara yer verildi.

Vaghela ve Jadav, Duygu Analizi üzerine çalıştıkları araştırma makalesinde, Twitter’den elde ettikleri Matrix filmine ait veri kümesini kullanmışlardır Yapılandırılmamış verileri, yapılandırılmış forma dönüştürerek yapılandırılmış verinin puanını hesaplamışlardır. Puanı Destek Vektör Makinesi ve Naive Bayes algoritmalarına girdi olarak vererek olumlu ya da olumsuz çıktılar üretmişlerdir. 5331 pozitif, 1802 negatif tweet üzerinde SVM için %74.73, NB için %74.55 doğruluk değerine ulaşmışlardır (Vaghela ve Jadav, 2016).

Duyarlılık analizi üzerine Shah ve arkadaşları bir araştırma makalesi yayınlandı. Veri setlerini twittersentiment.appspot.com üzerinden elde etmişlerdir. İlk olarak verilerin yeniden denenmesi, ardından verilerin ön işlenmesini yaparak trend olan konuyu ve en çok kullanılan tweet ve retweet’i bulmak için hashtag sınıflandırmasını kullanmışlardır. Farklı olarak her bir imojinin kullandıkları karakterleri tanımlayarak imojileride değerlendirmeye almışlardır. Örneğin “ :(“ ifadesini Negatif olarak değerlendirmişlerdir. Naive Bayes algoritmasında pozitif ve negatif olarak ayırdıkları tweetlerde %81 doğruluk oranına ulaşmışlardır (Shah vd., 2016).

Dey ve arkadaşları, Duygu Analizi üzerine bir araştırma makalesi üzerinde çalıştılar. Bu araştırma makalesinde, film incelemesi üzerine bir araştırma yaptılar. Film incelemesi verilerini www.imdb.com üzerinden 5000 negatif ve 5000 pozitif yorum olarak elde ederek Naive Bayes ve K-NN sınıflandırıcılarını kullanmışlardır. Film incelemesinde Naive Bayes için %80.12 doğruluk oranı elde ederken K-NN sınıflandırıcısı için %61.81 doğruluk oranına ulaşmışlardır (Dey vd., 2016).

Onan, Türkçe duygu analizi konusunda yürüttüğü çalışmada, evrişimli sinir ağı mimarilerini kullanarak Twitter API aracılığıyla 10500’er pozitif ve negatif olmak üzere toplamda 21000 Twitter

mesajı topladı. Çalışmada, k-en yakın komşu algoritması, destek vektör makineleri, lojistik regresyon ve Naive Bayes algoritması gibi farklı makine öğrenmesi yöntemleri kullanıldı. Deneysel sonuçlara göre, en yüksek doğruluk oranı %80.94 ile Naive Bayes algoritmasında elde edildi. Destek vektör makineleri %79.83, lojistik regresyon %78.61 ve k-en yakın komşu algoritması %76.63 doğruluk oranlarına sahipti (Onan, A, 2020).

Akba ve arkadaşları, www.beyazperde.com internet sitesinden film inceleme yorumlarından elde ettikleri veri setiyle bir duygu analizi çalışması yapmışlardır. Toplamda 40050 yorum üzerinden yaptıkları çalışmada Destek Vektör Makineleri ve Naive Bayes algoritmalarını kullanmışlardır. Çalışma sonunda pozitif ve negatif kategorileri için %83.9 doğruluk oranı, pozitif, negatif ve tarafsız kategorileri için %63.3 doğruluk oranı elde etmişlerdir (Akba vd., 2014).

Nalçakan ve arkadaşları yaptıkları çalışmada, Twitter üzerinden belirli markalar için yazılan tweetlerin duygu analizini gerçekleştirmeyi hedeflemektedir. Çalışmanın odaklandığı markalar Samsung, Apple ve LG'dir. İlk aşamada 500 tweet toplanmış ve bu tweetler pozitif, negatif veya nötr olarak işaretlenmiştir. Bu çalışmada en iyi sonuç, Naive Bayes (NB) algoritmasının kullanılmasıyla elde edilmiştir (Nalçakan vd., 2015).

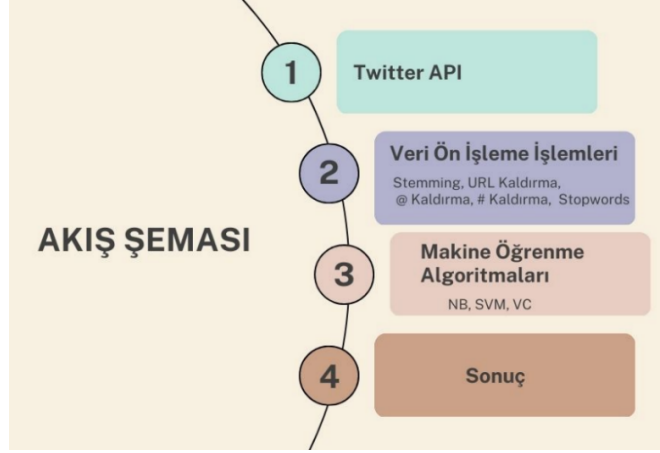
Kaynar ve arkadaşları yaptıkları çalışmada, IMDB'de yer alan film yorumlarını içeren bir veri seti üzerinde çeşitli makine öğrenme tekniklerinin kullanıldığı bir duygu analizi çalışmasını sunmaktadır. Yapay sinir ağları, destek vektör makinaları, Naive Bayes ve merkez tabanlı sınıflayıcı algoritmaları gibi yaygın olarak kullanılan sınıflandırma algoritmaları bu çalışmada değerlendirilmiş ve performansları karşılaştırılmıştır. Elde edilen sonuçlar incelendiğinde, en yüksek başarı değerlerinin yapay sinir ağları ve destek vektör makinalarında elde edildiği görülmüştür. Eğitim veri seti üzerinde yapay sinir ağı, %89.73 doğru sınıflandırma oranıyla destek vektör makinalarını (%84.07) geride bırakarak daha iyi bir performans sergilemiştir. Her iki sınıflandırıcı da test veri seti için yaklaşık %75 doğru sınıflama oranıyla hemen hemen aynı başarıyı göstermiştir (Kaynar vd., 2016).

Dinçer ve arkadaşları siber zorbalığın tespitine yönelik yaptıkları çalışmada, Twitter API kullanarak veri setlerini oluşturmuşlardır. Çalışmada Destek Vektör Makinesi (SVM), Lojistik Regresyon (LR), Naive Bayes (NB) yöntemleri ile analizleri gerçekleştirmişlerdir. Elde ettikleri sonuçlara bakıldığında en iyi sonuca 87.21'lik doğruluk değeriyle Lojistik Regresyon algoritmasıyla ulaşmışlardır. Kullanılan veri setinin örneklem sayısının fazla olmasını tavsiye ederek, karşılaşılabilecek yanlış analiz sonuçlarının önüne geçilebileceği belirtilmiştir. Veri setinde ön işleme adımları uygulanmış ve Türkçe dilinde bazı kelimelerin birden fazla anlam içerimesinin karşılaşılan zorluklardan birisi olduğu üzerinde durulmuştur (Dinçer vd..

Makalenin ikinci bölümü Yöntem, üçüncü bölümü bulgular ve dördüncü bölümü tartışma ve sonuç olarak düzenlenmiştir.

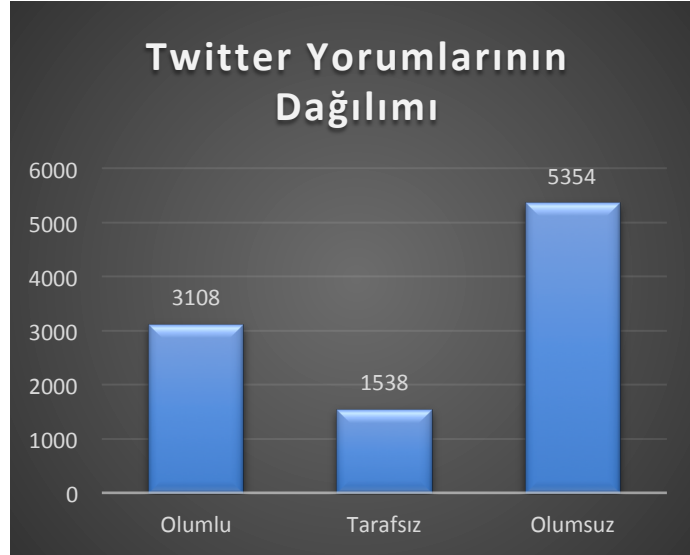
2. YÖNTEM

Bu çalışmanın amacı, denetimli makine öğrenme algoritmalarını önerilen oylama sınıflayıcısıyla karşılaştırarak elde edilen verilerin analizinin yapılmasıdır. ¹Bu amaçla yapılan duygu analizi çalışmasının daha iyi anlaşılabilmesi için iş-akış şeması Şekil 1’de sunulmuştur.



Şekil 1. Akış Şeması

Bu çalışmanın veri seti Twitter sosyal medya platformundan elde edildi. Twitter API kullanarak Python dilinde yazılmış bir uygulamayla Türkiye’de Kahramanmaraş merkezli 10 ili etkileyen depremle ilgili paylaşılmış Şubat 2023 ve Mart 2023 dönemlerini kapsayan 10.000 Twitter yorumu elde edildi. Tweetlerin 140 karakterden, ücretli üyeler için 9 Şubat 2023 itibariyle 4000’e çıkartılması sonucunda çalışmada uzun metinlerin duygu analizinde zorluklar çıkarabileceği düşünülerek sadece 140 karakterle yazılmış tweetler veri setine dahil edildi. 5354 Pozitif, 1538 Nötr, 3108 Pozitif Twitter yorumuyla veri seti oluşturuldu (Şekil 2).



Şekil 2. Veri setindeki Twitter yorumlarının dağılımı

Twitter yorumlarının ön işleme aşamaları gerçekleştirilerek tweetlerin polarizasyonu sağlandı. Aşamaların detayları aşağıda sunulmuştur.

- 1- “@” sembolünün kaldırılması: Bu sembol kişilerin etiketlenmesi için kullanıldığından duygu analizinde bir anlam ifade etmemektedir. Ayrıca etiketlenen kelimedede tamamen çıkartıldı.
- 2- URL kaldırılması: “URL” olarak bilinen (Uniform Resource Loader) linklerin kaldırılması da sağlandı.
- 3- “#” sembolünün çıkarılması: Hashtag olarak isimlendirilen bu sembolle birlikte kullanılan kelimeleri içeren yorumlar elde edilmektedir. Bu çalışmada Sadece “#” sembolü kaldırıldı içerdiği kelime içereceği anlam önem arz edebildiğinden değerlendirmeye alındı.
- 4- Etkisiz kelimelerin kaldırılması: Python’da bulunan Natural Language Toolkit (NLTK) yardımıyla etkisiz kelimeleri kaldırmak olarak Türkçeleştirilen “Stopwords” işlemi uygulandı. Bu işlemde öznitelik çıkarımı yaparken önemli olmayan kelimeler (ben, sen, de, da, ki, ile...) çıkartıldı (Mouthami vd., 2013). NLTK kullanılmasının en önemli nedeni Türkçe dilini desteklemesidir.
- 5- Kelimelerin köklerinin elde edilmesi: “stemming” işlemi uygulanarak kelimelerin köklerine ulaşıldı. Türkçe, sondan eklemeli bir dildir ve dil işleme modellerinde kelime sonlarındaki eklerin doğru bir şekilde yönetilmesi oldukça önemlidir. Bu nedenle, dil işleme uygulamalarında kelimelerin sonundaki eklerin temizlenmesi işlemi, doğru sonuçlar elde etmek için kritik bir adımdır.

Çalışmada NLTK (Natural Language Toolkit) kütüphanesi Türkçe dilini desteklediği için kullanıldı. Bu kütüphane, doğal dil verilerini işlemek, analiz etmek, sınıflandırmak ve çıkarım yapmak için bir dizi araç ve algoritma sunar. Zemberek kütüphanesi yardımıyla da dil kurallarına uymayan yanlış karakterlerin çıkarılması sağlanmıştır. Metinlerdeki yazım hatalarını tespit edip düzeltme işlemin “zemberek.normalization.TurkishSpellChecker” fonksiyonuyla gerçekleştirilmiştir. Bu işlem normalizasyon olarak adlandırılmaktadır. Ayrıca Doğal Dil İşleme alanında yaygın olarak kullanılan Naive Bayes denetimli makine öğrenme algoritması ve Destek Vektör Makineleri denetimli makine öğrenme algoritması, Oylama sınıflayıcısı ile karşılaştırıldı.

Oylama sınıflayıcısı için 3 farklı makine öğrenme algoritması ile hibrit bir yöntem önerildi. Oylama sınıflayıcısı (Voting Classifier), birkaç farklı sınıflandırma algoritmasını kullanarak daha doğru bir sınıflandırma sonucu elde etmek amacıyla kullanılan bir makine öğrenimi yöntemidir. Bu yöntemde, farklı algoritmaların çıktıları bir araya getirilerek bir oylama mekanizması ile son karar verilir.

Örneğin, veri setinde sınıflandırma işlemi yapılacaksa, bir oylama sınıflayıcısı kullanarak, farklı sınıflandırma algoritmalarının çıktuları birleştirilebilir. Her bir algoritmanın sınıflandırma sonuçları belirlenir ve sonunda bu sonuçlar bir oylama mekanizmasıyla birleştirilerek son karar verilir. Oylama sınıflayıcısı, her bir algoritmanın ağırlıklı oylarına dayanarak sonucu tahmin ettiği için, diğer sınıflandırma yöntemlerinden daha yüksek bir doğruluk oranı elde etmek mümkündür.

Bu çalışmada Oylama Sınıflayıcısı için, Lineer Regresyon, Lojistik regresyon ve Destek Vektör Makineleri algoritmaları kullanılarak daha iyi sonuçlar elde edilmek istendi. Oylama sınıflayıcısında tek sayılı algoritma sayısının tercih edilmesinin nedeni her algoritmanın oylayacağı negatif ya da

pozitif oylamanın sonuç olarak bir karara varması ve eşitlik durumunun yaşanmaması içindir. Ayrıca Oylama Sınıflayıcısı geleneksel yöntemleri kullanan makine öğrenme algoritmalarından, tek bir algoritmanın sonucuna göre değil, çoğunluğun kararına göre bir sonuca vararak daha iyi sonuçlar üretmektedir. Bununla birlikte Oylama sınıflayıcısı için seçilen algoritmaların her veri seti için aynı sonuçları vermediği bilinmekle birlikte, birçok deneme ve yanılma yoluyla uygun algoritmaların bir araya getirilmesi önem arz etmektedir.

Bu araştırmada, veri analizi için dört farklı metrik kullanılmıştır. Bu metrikler, doğruluk, keskinlik, duyarlılık ve F1 skoru olarak adlandırılmaktadır. Bu metriklerin hesaplanması için kullanılan formüller, denklem 1, denklem 2, denklem 3 ve denklem 4 olarak sunulmaktadır.

$$\text{Doğruluk (Accuracy)} = \frac{(TP + TN)}{(TP + FN + TN + FP)} \quad (1)$$

$$\text{Keskinlik (Precision)} = \frac{TP}{(TP + FP)} \quad (2)$$

$$\text{Duyarlılık (Recall)} = \frac{TP}{(TP + FN)} \quad (3)$$

$$\text{F1 Skoru (F1 Score)} = \frac{2 TP}{(2TP + FN + FP)} \quad (4)$$

Doğru pozitif (TP), doğru negatif (TN), yanlış pozitif (FP) ve yanlış negatif (FN) terimleri, sırasıyla doğru ve yanlış sınıflandırılmış veri örneklerinin sayısını ifade etmektedir. Bu terimler, çeşitli makine öğrenimi ve veri madenciliği uygulamalarında, performans metriklerinin hesaplanmasında kullanılmaktadır. Bu metrikler, bir modelin doğruluğunun ve performansının ölçülmesinde önemli bir rol oynamaktadır.

Naive Bayes, bayes teoremine bağlı olarak çalışan bir algoritmadır. Veri kümesindeki özniteliklerin birbirinden bağımsız olduğu kabul edilir. Bayes teoremi aşağıda sunulmuştur (Denklem 5).

$$P(A|B) = \frac{(P(B|A) P(A))}{P(B)} \quad (5)$$

$P(A|B)$: B'nin olasılığı verildiğinde A olayının olma olasılığı,

$P(B|A)$: A'nın olasılığı verildiğinde B olayının olma olasılığı,

$P(A)$: A olayının olma olasılığı,

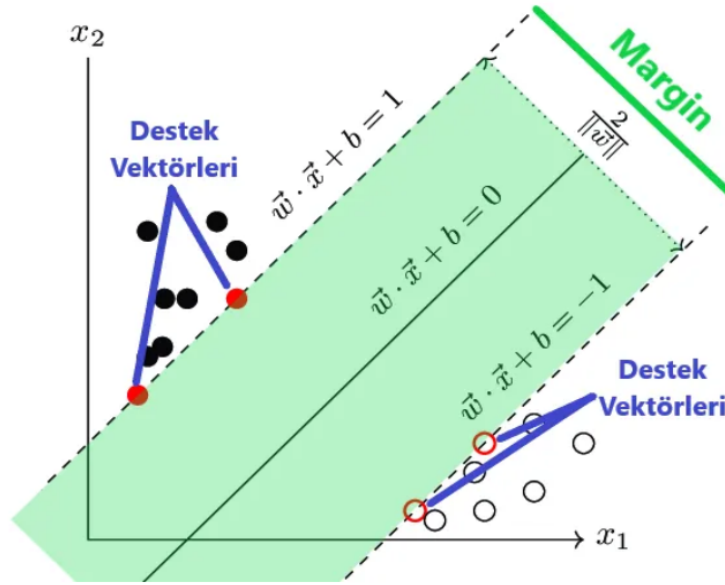
$P(B)$: B olayının olma olasılığı olarak ifade edilir.

Bayes yöntemleri, özellik değerlerine dayalı olarak her bir sınıfın olasılığını hesaplamak için eğitim verilerini kullanarak sınıflandırma yaparlar. Bu yöntem, bir sonucun olasılığını tahmin etmek için birden fazla özellikten gelen bilgileri aynı anda dikkate alır ve özelliklerin zayıf etkileri bile bir araya geldiğinde etkileri oldukça büyük olabilir (Gunawan vd., 2022). Bayes sınıflandırıcıları, verilerdeki hatalı sınıflandırma durumlarını da düzeltmek için mevcut özellikleri kullanırlar ve az

veriyle bile iyi sonuçlar üretebilirler. Bu nedenle, Bayes yöntemleri yüksek boyutlu verilerde de yaygın olarak kullanılan bir algoritmadır (Hofmann ve Klinkenberg, 2016).

Naive Bayes algoritması, sınıf belirleme işlemini, verilen özelliklere göre sınıf olasılıklarını karşılaştırarak gerçekleştirir. Verilen örneğin sınıfı, en yüksek olasılığı veren sınıf olarak belirlenir.

Destek Vektör Makinesi algoritması küçük ve orta ölçekli veri setleri için kullanıma uygun etkili bir denetimli öğrenme algoritmasıdır. Gelecek verinin hangi sınıfa ait olduğuna karar vermektedir. Şekil 3'te beyazlar ve siyahlar iki farklı sınıf olarak ayrılmıştır. Sınıflandırma problemlerinde, sınıfları birbirinden ayırmak için bir doğru veya hiperyüzey belirlenir. Bu doğru, sınıfların olası bölgelerini ayıran ve Margin adı verilen yeşil bir alana sahip olur. Margin, sınıfların ayrımının ne kadar doğru ve geniş olduğunu belirleyen bir ölçüttür. Margin'in genişliği arttıkça, sınıflar daha iyi bir şekilde ayrıştırılabilir. Destek Vektör Makinesi Algoritması Çalışma Mantığı Şekil 3'te sunulmuştur.



Şekil 3. SVM çalışma mantığı

SVM için formül denklem 6'da verilmiştir.

$$\hat{y}^{\text{SVM}} = \begin{cases} 0 & \text{eğer } w^T \cdot x + b < 0, \\ 1 & \text{eğer } w^T \cdot x + b \geq 0. \end{cases} \quad (6)$$

w ; ağırlık vektörü

x ; girdi vektörü,

b ; sapma olarak ifade edilmektedir.

SVM, düzlemdeki noktaların sınıflandırılması için kullanılan bir yöntemdir. Bu yöntemde, bir doğru veya hiper düzlem kullanılarak noktalar farklı sınıflara ayrılır. Eğer çıkan sonuç 0'dan küçükse, nokta beyaz noktalara daha yakın, 0'a eşit veya büyükse siyah noktalara daha yakın kabul edilir. SVM, doğru veya hiper düzlemin seçimi, Margin genişliği ve bazı parametrelerin belirlenmesi gibi faktörlere bağlı olarak iyi sonuçlar verebilir. Bu yöntem, özellikle küçük boyutlu verilerde etkilidir ve geniş Marginler ile iyi bir performans sergiler (Akpınar, 2014).

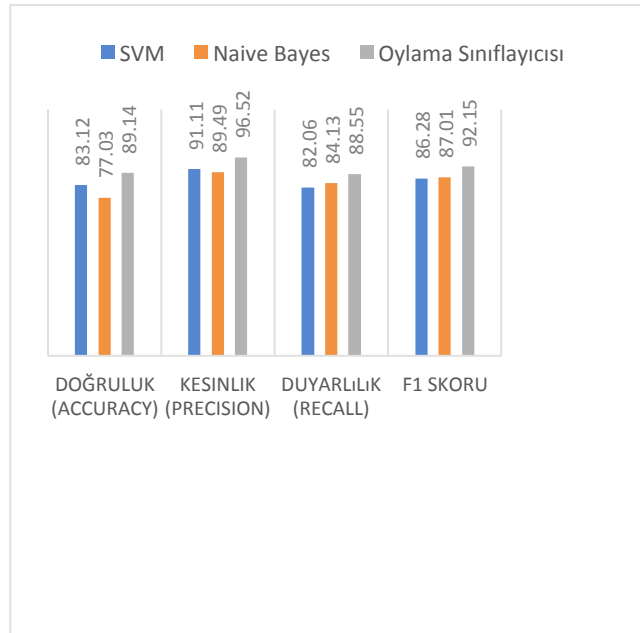
3. BULGULAR

Bu çalışmada toplam veri setinin %80'i eğitim verisi ve %20'si test verisi olarak ayrıldı. Doğruluk (Accuracy) değeri modelde doğru tahmin ettiğimiz alanların toplam veri kümesine oranı ile hesaplanmaktadır. Şekil 4'te Doğruluk değerlerine bakıldığında Lineer Regresyon, Lojistik Regresyon ve Destek Vektör makinesi algoritmalarından oluşan oylama sınıflayıcısı 89,14 ile en başarılı sonucu alırken, SVM algoritması NB algoritmasına göre daha başarılı sonuç üretmiştir.

Kesinlik (Precision), sınıflandırma modelinin pozitif olarak tahmin ettiği değerlerin gerçekten ne kadarının pozitif olduğunu ölçen bir metriktir. Oylama Sınıflayıcısı, %96.52 kesinlik skoru ile en başarılı sonucu elde etmiştir. Bu sonuç, modelin pozitif tahminlerinin büyük bir oranda gerçek pozitif değerlerle eşleştiğini göstermektedir.

Duyarlılık (Recall), sınıflandırıcı tarafından pozitif olarak tahmin edilmesi gereken örneklerin ne kadarının gerçekten pozitif olarak tahmin edildiğini gösteren bir ölçüttür. Oylama Sınıflayıcısı, %88.55 kesinlik skoru ile en başarılı sonucu elde etmiştir.

F1 Skoru, sınıflandırıcının Kesinlik (Precision) ve Duyarlılık (Recall) değerlerinin harmonik ortalamasını gösteren bir ölçüttür. Harmonik ortalamasının kullanılması, uç durumlar da dahil olmak üzere tüm değerleri dikkate almak için tercih edilen bir yöntemdir. Önerilen model olan Oylama Sınıflayıcısı 92.15 F1 Skoru ile en başarılı sonucu üreten algoritma olmuştur. Genel olarak SVM, NB algoritmasına göre daha iyi sonuçlar üretse de Oylama Sınıflayıcısı en başarılı model olmuştur.

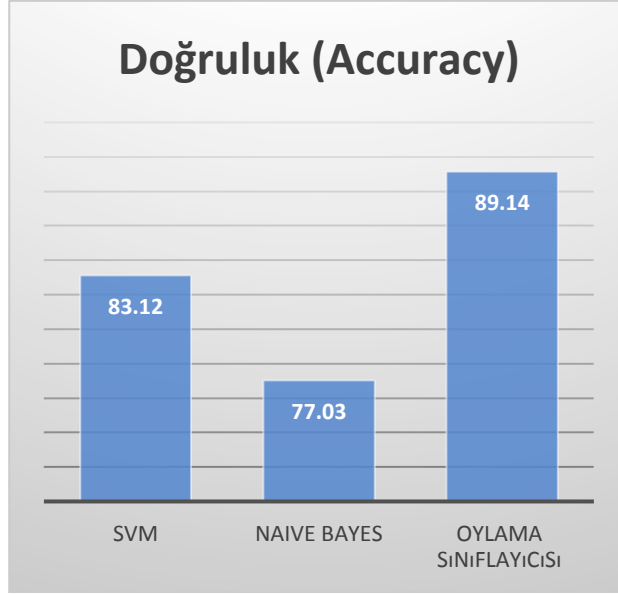


Şekil 4. Algoritmaların metrik değerleri

Tablo 1'de bütün algoritmaların doğruluk (accuracy) değerlerinin karşılaştırılması yapılarak en başarılı algoritmanın %89.14 ile Oylama sınıflayıcısı olduğu görülmektedir.

Algoritma	Doğruluk
SVM	83.12
Naive Bayes	77.03

Şekil 5'te bütün algoritmaların karşılaştırılması yapılarak doğruluk değerine göre en başarılı algoritmanın %89.14 ile Oylama sınıflayıcısı olduğu görülmektedir.



Şekil 5. Algoritmaların doğruluk değerleri karşılaştırmaları

4. TARTIŞMA VE SONUÇ

Bu çalışmada Türkiye’de meydana gelen Kahramanmaraş merkezli 10 ili etkileyen depremle alakalı sosyal medya platformu olan Twitter’da paylaşılan yorumların duygu analizi yapıldı. Twitter yorumları, Twitter API kullanarak Python dilinde yazılmış bir uygulamayla elde edildi. 10.000 Twitter yorumu ön işleme aşamalarından geçirilerek, polariteleri hesaplandı ve olumlu, olumsuz, nötr olarak sınıflandırıldı. Veri setinin %80’i eğitim veri seti ve %20’si test veri seti olarak ayrıldı. Duygu analizinde yaygın olarak kullanılan SVM ve NB denetimli öğrenme algoritmaları önerilen model olan Oylama Sınıflayıcısı ile karşılaştırıldı. Sonuç olarak en başarılı değerler önerilen model olan Oylama sınıflayıcısıyla elde edildi. Depremle alakalı olarak paylaşılan Twitter yorumlarının çoğunlukla olumsuz duygu taşıdığı ve önerilen modelin %89.14 başarı oranı ile doğru tespit yaptığı görüldü. Bu çalışmada önerilen model ile, paylaşılan bir içeriğin olumlu ya da olumsuz duygu taşıyıp taşımadığı yüksek bir oranda doğru tahmin edilebilmektedir. Bu çalışma literatür taramasında belirtilen benzer çalışmalara kıyasla SVM ve NB algoritmalarıyla dahi genel olarak daha iyi sonuçlar elde etmiştir. Bunun temel nedeni, veri ön işleme aşamalarının ayrıntılı olarak ele alınması ve kullanılan kütüphanelerin (NLTK, Zemberek) başarılı olmasıdır. Bhumika M. Jadav ve Vimalkumar B. Vaghela Twitter’den elde ettikleri Matrix filmine ait veri kümesini kullanarak yaptıkları duygu analizi çalışmasında SVM için %74.73, NB için %74.55 doğruluk değerine ulaşmışlarken, bu çalışmada SVM için, %83.12, NB için %77.03 ile daha başarılı sonuçlar elde edildi. Bu sonuçlar, duygu analizi konusunda denetimli öğrenme algoritmalarının etkinliğini ve sosyal medya verilerinin duygu analizi için önemini vurgulamaktadır.

İleride geliştirilecek bir uygulama ile paylaşılan bir Twitter yorumunun taşıdığı duygu gerçek zamanlı olarak tahmin edilebilir.

5. KAYNAKÇA

- Akba, F., Uçan, A., Akcapinar Sezer, E., ve Sever, H. (2014, Temmuz). Assessment of feature selection metrics for sentiment analyses: Turkish movie reviews. In 8th European Conference on Data Mining, Lisbon.
- Akpınar, H. (2014). *Data: Veri Madenciliği Veri Analizi*. Papatya Yayıncılık Eğitim.
- Dey, L., Chakraborty, S., Bose, ve B., Tiwari, S. (2016). Sentiment analysis of review datasets using Naïve Bayes and K-NN classifier. *Information Engineering and Electronic Business*, 4, 54-62. doi: 10.5815/ijieeb.2016.04.07
- Dinçer, E. Ş., Kayaoglu, D., ve Safarlı, S. (2022). Metin Madenciliği ve Duygu Analizi ile Siber Zorbalık Tespiti. *Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi*, 3(2), 38-45.
- Gunawan, P. H., Alhafidh, T. D., ve Wahyudi, B. A. (2022). The Sentiment Analysis of Spider-Man: No Way Home Film Based on IMDb Reviews. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 6(1), 177-182.
- Hofmann, M. & Klinkenberg, R. (2016). *RapidMiner: Data mining use cases and business analytics applications*. United Kingdom: CRC Press.
- Kaynar, O., Görmez, Y., Yıldız, M., ve Albayrak, A. (2016). Makine öğrenmesi yöntemleri ile Duygu Analizi. In International Artificial Intelligence and Data Processing Symposium (IDAP'16) Vol. 17, No. 18, pp. 17-18.
- Mouthami, K., Devi, K. N., ve Bhaskaran, V. M. (2013, Şubat). Sentiment analysis and classification based on textual reviews. In 2013 International Conference on Information Communication and Embedded Systems (ICICES). 21-22 February, Tamilnadu.
- Nalçakan, Y., Bayramoğlu, Ş. S., ve Tuna, S. (2015). Sosyal Medya Verileri Üzerinde Yapay Öğrenme ile Duygu Analizi Çalışması. Technical Report.
- Onan, A. (2020). Evrişimli sinir ağı mimarilerine dayalı Türkçe duygu analizi. *Avrupa Bilim ve Teknoloji Dergisi*, 374-380.
- Ruuska, S., Hämäläinen, W., Kajava, S., Mughal, M., Matilainen, ve P., Mononen, J. (2018). Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle. *Behavioural Processes*, 148, 56-62.
- Shah, S., Kumar, K., ve Saravanaguru, R. K. (2016). Sentiment analysis of Twitter data using classifier algorithm. *International Journal of Electrical & Computer Engineering*, 6(1), 357-366.
- Vaghela, V. B., ve Jadav, B. M. (2016). Sentiment analysis using support vector machine based on feature selection and semantic analysis. *International Journal of Computer Applications*, 146(13), 10-15.
- Websiterating, (2023). Twitter İstatistikleri- 2023- Kullanıcılar, Tweetler ve Markalar. Alınan Yer <https://www.websiterating.com/tr/research/twitter-statistics/> (Erişim Tarihi: 29.04.2023).