Doğu Fen Bilimleri Dergisi / Journal of Natural & Applied Sciences of East 6(1): 57-68 (2023)

Araştırma Makalesi / Research Article

Tweetlerin Duygu Analizi İçin Hibrit Bir Yaklaşım

Erol KINA[1]*, Emre BİÇEK[2]

[1]Van Yüzüncü Yıl Üniversitesi, Özalp Meslek Yüksekokulu, Bilgisayar Programcılığı Bölümü, Van
[2]Van Yüzüncü Yıl Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Van
*e-mail: erolkina@yyu.edu.tr

DOI: 10.57244/dfbd.1314901
Geliş tarihi/Received:15/06/2023 Kabul tarihi/Accepted:14/12/2023

Özet

Sosyal medyada ifade edilen görüşler, çeşitli işletmeler için her zaman dikkate alınan ve faydalı bir kaynak olmuştur. Duygu analizi, kullanıcılar tarafından oluşturulan içeriği belirli kutuplara (pozitif, negatif) dağılmış ve etkin bir şekilde sınıflandırmayı ifade eden genel bir terimdir. Duyguların sınıflandırma ve analizini gerçekleştirmek için çeşitli araçlar ve teknikler bulunmaktadır. Bunlar, veri üzerinde ön işleme adımları tamamlandıktan sonra hedef grubu sınıflandıran denetimli makine öğrenimi tekniklerini içermektedir. Hibrit araçlar, makine öğrenimi ve sözlük tabanlı algoritmaların birleşimini kullanarak, işaretlenmiş verilere dayalı olarak sınıflandırma yapar. Bu makalede, duyguların analizinde SVM algoritmasını Weka adında açık kaynaklı bir yazılım ile birlikte kullanılmıştır. İki önceden kategorize edilmiş tweet veri seti kullanıldı. SVM algoritmasının performansı, analitik metrikler yardımıyla değerlendirildi.

Anahtar Kelimeler: Hibrit, Makine öğrenmesi, duygu analizi

A Hybrid Approach for Sentiment Analysis of Tweets

Abstract

The views and inputs expressed by the community have always been a crucial and valuable resource for various enterprises. The advent of widespread community media has provided an exceptional opportunity for studying and assessing diverse fields, replacing the peculiar, laborious, and inaccurate approaches that companies used to rely on. This particular type of analysis falls under the subclass of sentence analysis. Sentiment analysis, a broad term, refers to the process of effectively classifying user-generated content into specific polarities. To perform sentiment identification and analysis, a range of tools and techniques are available, including supervised machine learning techniques that classify the target group after training on data. Hybrid instruments, combining machine learning and lexicon-based algorithms, classify content based on annotated dictionaries. In this study, we employed the Support Vector Machine (SVM) algorithm with Weka, an open-source software, to analyze sentiments. Two pre-categorized datasets of tweets were utilized. The performance of the SVM algorithm was assessed using analytical metrics.

Keywords: Hybrid, Machine Learning, Sentiment Analysis

Giriş

Metin verilerinin büyük hacmi nedeniyle, metin madenciliği araçlarına ve yöntemlerine olan talep hızla artmaktadır. Özellikle Facebook ve Twitter gibi sosyal medya platformları, bu verilerin her geçen gün artmasına yol açmaktadır. Bu büyük miktardaki veri ve değerlendirmelerin duygusal eğilimleri ve polaritesi, şirketlere ve kurumlara sınırsız faydalar sağlayabilir. Şirketler ve kurumlar, ürün veya hizmet

gereksinimlerini belirlemek ve pazarlarının konumunu sürdürmek ve güçlendirmek amacıyla duygu analizini etkili bir şekilde kullanarak bu verileri değerlendirebilirler.

Duygusal eğilimleri belirlemek için, önceden belirlenmiş bir kelime dağarcığından ve ağırlıklı terimlerden ve duygu yönelimlerinden yararlanan sözlük tabanlı bir teknik kullanılır. Bu yöntem, metinleri etkili bir şekilde sınıflandırmak için kendi içinde bulunan bir sözlüğü kullanır. Linguistic Inquiry Word Count, WordNet, SentiStrength 3.0, SentiWordNet, Affective Norms for English Words ve SenticNet gibi sözlük tabanlı uygulamalar popüler yöntemler olarak kabul edilmektedir.

Denetimli makine öğrenimi yaklaşımlarında, gerçek dünya verilerine dayanak sağlamak için bir eğitim veri setinin mevcut olması önemlidir. Hibrit bir platform, iki stratejiyi birleştirir. Verileri hem makine öğrenimi algoritmalarıyla sınıflandırmakta hem de belirli bir sözlük üzerinden kategorize etmektedir.

Bu çalışmada, duygu analizi için SVM (Destek Vektör Makineleri) algoritması, önceden sınıflandırılmış tweet veri setleri üzerinde kullanılmıştır. Her iki veri seti için de SVM'nin hassasiyetini değerlendirmek için Geri Çağrılama (Recall) ve F-Skoru (F-Measure) kullanılmıştır. Araştırmanın bölümleri aşağıda açıklanmaktadır: İkinci bölümde ilgili çalışmalar açıklanmaktadır. Üçüncü bölümde veri toplama süreci daha detaylı bir şekilde ele alınmaktadır. Dördüncü bölüm sınıflandırma yöntemlerine odaklanmaktadır. Beşinci bölüm bulguları özetlerken, altıncı bölümde çalışma bir sonuçla tamamlanmaktadır.

Sonuç olarak, bu çalışmada metinlerin duygu analizi için bir hibrit yaklaşım önerilmektedir. Sözlük tabanlı lexicon yöntemi ve makine öğrenimi algoritmalarının birleşimi kullanılarak duygu analizinin doğruluğunun artırılması hedeflenmektedir. Önerilen hibrit yaklaşımın performansı, SVM algoritması ve önceden sınıflandırılmış tweet veri setleri kullanılarak değerlendirilmiştir. SVM'nin hassasiyeti Geri Çağrı ve F-Skoru metrikleri kullanılarak değerlendirilmiştir.

Birçok araştırmacı, sosyal ağ platformları aracılığıyla üretilen büyük miktardaki veriyi otomatik olarak çıkarmak ve analiz etmek için çalışmaktadır. Literatür incelendiğinde bu alanda yapılmış çeşitli çalışmalar mevcuttur. Bunlardan bazıları aşağıda sunulmuştur.

Liu ve ark. (2013) tarafından yapılan çalışmada, Twitter üzerinde duygu sınıflandırması için adaptif bir eş zamanlı SVM yöntemi geliştirmişlerdir. Bu yöntem, tweetler üzerinde duygu sınıflandırması yapmak için eş zamanlı öğrenme (co-training) yaklaşımını kullanmaktadır. Çalışmanın amacı, sınırlı etiketli veriye sahip olduğumuz durumlarda duygu sınıflandırmasının doğruluğunu artırmaktır. Bunun için, iki farklı SVM sınıflandırıcının birbirini eğitmesi ve güncellemesi gerekmektedir. Bu şekilde, her bir sınıflandırıcının diğerinden farklı örnekleri etiketlemesi ve öğrenmesi sağlanır. Yöntem, Twitter üzerindeki tweetlerden elde edilen verileri kullanarak yapılmıştır. Öncelikle, tweetlerin içerdikleri kelimelerin özellik vektörleri çıkarılır ve bu vektörler üzerinde önceden eğitilmiş SVM sınıflandırıcılar kullanılarak duygu etiketleri tahmin edilir. Ardından, iki farklı sınıflandırıcının karşılıklı olarak birbirini eğitmesi için bir geri besleme döngüsü oluşturulur. Çalışmanın sonuçları, önerilen adaptif eş zamanlı öğrenme yaklaşımı SVM yönteminin diğer yöntemlere kıyasla daha yüksek doğruluk oranları sağladığını (%85) göstermektedir. Özellikle, tweetlerin duygusal özelliklerini daha iyi öğrenebilen ve sınıflandırabilen bir model elde edilmiştir. Bu çalışma, sınırlı etiketli veriye sahip olduğumuz durumlarda duygu sınıflandırması için etkili bir yöntem sunmaktadır. Adaptif co-training SVM yaklaşımı, Twitter gibi sosyal medya

platformlarından elde edilen büyük veri setleri üzerinde duygu analizi yapmak için potansiyel bir çözüm olabilir.

Zainudin ve ark. (2016) yaptıkları çalışmada, Twitter yorumlarını olumlu veya olumsuz olarak sınıflandırmak için yeni bir yöntem önermişlerdir. Bu yöntemde, doğal dil işleme (NLP) teknikleri ve makine öğrenimi algoritmaları kullanılmıştır. Yazarlar, uzak gözetim yöntemini kullanarak, duygu analizi için makine öğrenme algoritmalarının Twitter üzerindeki sonuçlarını sunmuş ve tartışmışlardır. Yazarlar, gürültülü etiketler olarak kullanılan duygulu tweetleri eğitim verisi olarak kullanmışlardır ayrıca farklı metin veri kümesi üzerinde yöntemlerini test edip elde ettikleri sonuçları sunmuşlardır. Sonuçlar, önerilen yöntemin duygu analizi konusunda etkili olduğunu ve iyi bir performans gösterdiğini göstermektedir. Olumlu ve olmuşuz duygu taşıyan tweetlerle eğitildiğinde, Naive Bayes ve SVM gibi ML algoritmalarının %80 doğrulukla çalışabileceğini belirtmişlerdir. Sınıflandırma sırasında kullanılan ön işleme aşamasındaki ölçümler çalışmada vurgulanmıştır.

Mudinas ve ark. (2012), yaptıkları çalışmada çeşitli veri madenciliği teknikleri kullanılarak öğrencilerin spekülatif başarılarının tahmin edilmesi ve incelenmesi amaçlanmıştır. Karar ağacı, Çok Katmanlı Algılama ve Naive Bayes yöntemleri kullanılmıştır. Bu stratejiler, iki dönem boyunca, iki lisans dersindeki öğrenci verilerini analiz etmek için kullanılmıştır. Bulgulara göre, Naive Bayes'in tahmin doğruluğu %86 olarak belirlenmiştir ve bu değer Karar ağacı ve Çok Katmanlı Algılama yöntemlerinden daha yüksek olarak hesaplanmıştır. Bu tahmin yöntemi sayesinde eğitimciler, belirli bir dersin başarısını erken aşamada tespit edebilmektedirler. Sonuç olarak, eğitimcilerin bu öğrencilere ekstra dikkat göstererek akademik performanslarını artırması mümkün olabilmektedir.

Erşahin ve ark. (2019) tarafından yapılan çalışmada, duygu analizi için bir hibrit yöntem önerilmektedir. Makalede, duygu analizi için birleşik bir yaklaşım kullanılmaktadır. Sözlük tabanlı bir teknik olan lexicon yöntemi ve denetimli makine öğrenimi algoritmaları bir araya getirilerek duygu analizinin doğruluğunun artırılması amaçlanmıştır. Lexicon yöntemi, önceden belirlenmiş bir kelime dağarcığı ve duygu yönelimlerini kullanarak metinleri sınıflandırmada kullanılmaktadır. Denetimli makine öğrenimi algoritmaları ise eğitim veri setine dayanarak duygu sınıflandırması yapmaktadır. Makine öğrenimi tarafında, Naive Bayes, destek vektör makineleri ve J48 gibi üç denetimli sınıflandırıcı kullanarak sınıflandırma problemini ele alınmıştır. Makalede, Türkçe tweet veri setleri üzerinde yapılan deneyler ve performans değerlendirmeleri sunulmaktadır. Elde edilen sonuçlar, önerilen hibrit yöntemin Türkçe metinlerde duygu analizinde %7 oranında daha doğru sonuçlar elde ettiklerini göstermektedir. Hibrit bir metodoloji kullanılarak, film veri seti üzerinde NB algoritmasıyla elde edilen doğruluk oranı %88,93, otel veri seti üzerinde SVM algoritmasıyla elde edilen doğruluk oranı %91,96 ve Twitter verileri üzerinde NB algoritmasıyla elde edilen doğruluk oranı ise %83,37 olarak belirlenmiştir.

Rumelli ve ark. (2019) tarafından gerçekleştirilen çalışmada, Hepsiburada.com'daki müşterilerin ürün yorumları ve değerlendirmeleri üzerinde duygu analizi modeli geliştirmek amacıyla makine öğrenimi algoritmaları ve sözlük tabanlı yaklaşımların birleşimini kullanılmışlardır. Çalışmanın ilk aşamasında, her bir kelimenin sözlükteki puan değeriyle birlikte toplama yöntemi kullanılarak bir model oluşturulmuş ve hesaplamalar yapılmıştır. Daha sonra, metinlerin polarite puanlarına dayanarak NB, RF, SVM, KNN gibi makine öğrenimi algoritmaları duygu analizi için

eğitilmiştir. Araştırma sonuçları, insan müdahalesi olmaksızın duygu analizinin %73 doğruluk oranında gerçekleştirilebildiğini ortaya koymaktadır.

Appel ve ark. (2016), yaptıkları çalışmada NLP teknikleri, semantik kurallar ve bulanık kümeleri ile bir hibrit yöntem kullanılarak film incelemeleri üzerinde başarılı sonuçlar elde ettiklerini ve %76 doğruluk oranına ulaştıklarını belirtmektedirler.

Ohana ve Tierney (2009), çalışmalarında SentiWordNet'i kullanarak duygu yönünü hesapladıktan sonra SVM sınıflandırıcısını uygulamışlardır. Film incelemeleri üzerinde SentiWordNet kelime dağarcığının duygu analizi uygulanması sonuçları sunulmuştur. Yaklaşımları, duyguyu belirlemek için pozitif ve negatif terim puanlarını içermektedir. SentiWordNet'i kaynak olarak kullanarak ilgili özelliklerin bir veri seti oluşturarak iyileştirme sağlamışlardır. Ardından makine öğrenimi algoritmalarını uygulayarak SentiWordNet puanlarının özellik olarak kullanıldığı en iyi doğruluk oranını %69,35 olarak elde edilmişlerdir.

Türkmenoğlu'nun (2015) yüksek lisans çalışması, film yorumları ve Twitter sosyal medya verileri üzerinde yapılan duygu analizi araştırmasını içermektedir. Çalışmada, sözlük tabanlı ve makine öğrenmesi yaklaşımları olmak üzere iki farklı yöntem kullanılmış ve elde edilen sonuçlar karşılaştırılmıştır. Sözlük tabanlı yöntem için Sentistrength sözlüğü Türkçe'ye çevrilerek kullanılmıştır. Makine öğrenmesi yaklaşımında ise SVM, NB, DT (Desicion Tree – Karar Ağaçları) algoritmaları değerlendirilmiştir. Twitter veri seti üzerinde yapılan analizde, sözlük tabanlı yöntem %75,2 başarı oranı elde ederken, makine öğrenmesi yaklaşımında SVM algoritması %85 başarı oranıyla öne çıkmıştır. Film yorumlarından oluşturulan veri setinde ise sözlük tabanlı yöntem %79,5 başarı oranına ulaşırken, makine öğrenmesi yaklaşımı SVM algoritmasıyla %89 başarı oranı sağlanmıştır.

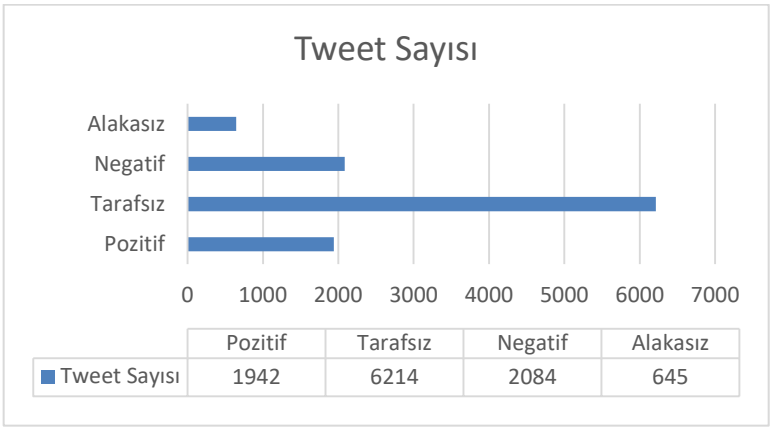
Naz ve arkadaşları (2018) yaptıkları çalışmada, halka açık SemEval 2016 veri seti üzerinde duygu analizi gerçekleştirmişlerdir. Bu çalışmada, duyarlılık puanları oluşturulmuş ve SVM algoritması kullanılmıştır. Elde edilen sonuçlar, duygu puan vektörünün dikkate alınmadığı durumda %79,6 doğruluk elde edildiğini göstermiştir. Ancak duygu puan vektörü dikkate alındığında ise doğruluk oranı %81 olarak belirlenmiştir.

Beleveslis ark. (2019) tarafından gerçekleştirilen çalışma, Yunanca tweetlerden toplanan verilerin analizini içermektedir. Bu araştırmada, önceden hazırlanan sözlükler kullanılmış ve RF, DT ve XGBoost gibi algoritmalar karşılaştırılmıştır. Araştırmanın sonuçları, RF algoritmasının en yüksek doğruluk oranını (%80) elde ettiğini göstermektedir.

Sham ve Mohamed (2022) tarafından gerçekleştirilen çalışmada, sözlük tabanlı, makine öğrenmesi ve hibrit yaklaşımların performansı karşılaştırılmıştır. Çalışmada, Twitter verilerinden oluşan üç farklı hazır veri seti üzerinde analiz yapılmıştır. Sözlük tabanlı yaklaşımda, çeşitli sözlükler arasında karşılaştırma yapılmıştır. Makine öğrenmesi yöntemi için SVM, NB ve LR (Lojistik Regresyon) algoritmaları kullanılmış ve sonuçlar karşılaştırılmıştır. En yüksek doğruluk oranı, sözlük tabanlı yaklaşımda birleştirilmiş veri seti ve VADER sözlüğü ile %57 olarak elde edilirken, makine öğrenmesi yaklaşımında LR algoritmasıyla %70,2 olarak elde edilmiştir. Hibrit yaklaşımda ise en yüksek F-skoru, birleştirilmiş veri seti, TextBlob ve LR algoritması kullanılarak %75,3 olarak belirlenmiştir.

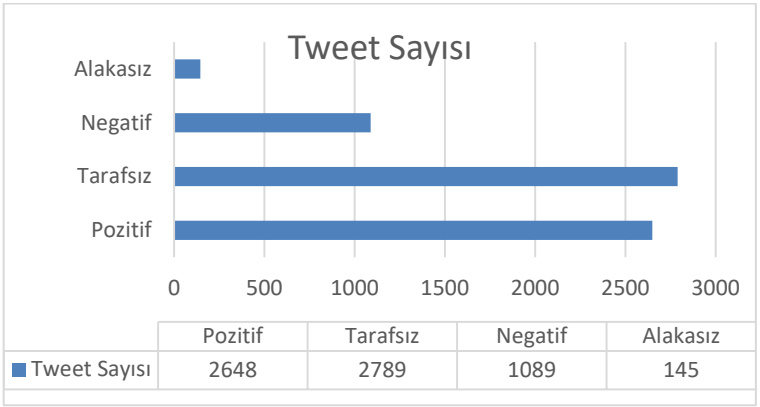
Materyal ve Yöntem

Bu çalışma, SVM'nin metin tabanlı duygu tespitinde tarafsız, olumsuz veya olumlu gibi kategorilere ayrılan ifadeler üzerinde odaklanmaktadır. Twitter platformundan elde edilen iki önceden kategorize edilmiş veri seti kullanılmıştır. Performans ve doğruluk değerlendirmesi için Twitter platformundan elde edilen tweetler test verisi olarak seçilmiştir. SVM algoritmasının çıktı polaritesi, her bir tweet için önceden belirlenmiş sınıf ile karşılaştırılarak fark hesaplanmıştır. Bu çalışmada precision (kesinlik), recall (geri çağırma) ve F-score (F-skoru) gibi metrikler kullanılmıştır. Bulguları analiz etmek ve görselleştirmek için Weka aracı kullanılmıştır. Weka, GNU Genel Kamu Lisansı şartları altında serbestçe kullanılabilen bir yazılımdır. Veriler Twitter API kullanılarak elde edilmiştir. İlk veri seti, film incelemeleriyle ilgili 2023 yılı Ocak – Mayıs ayları arasında yazılmış Twitter yorumlarından oluşmaktadır. 1942 olumsuz, 6214 tarafsız, 2084 olumlu ve 645 alakasız olmak üzere toplamda 10885 Twitter yorumu bulunmaktadır. Birinci veri setinin Şekil olarak gösterimi aşağıda sunulmuştur (Şekil 1).



Şekil 1. Birinci veri seti için elde edilen tweet sayısı ve taraflılık düzeyleri.

İkinci veri seti çevrimiçi eğitim süreciyle ilgili 2023 yılı Ocak – Mayıs ayları arasında yazılmış Twitter yorumlarından oluşmaktadır. Bu veri setinde 2648 olumsuz, 2789 tarafsız, 1089 olumlu ve 145 alakasız olmak üzere toplamda 6671 Twitter yorumu bulunmaktadır. İkinci veri setinin Şekil olarak gösterimi aşağıda sunulmuştur (Şekil 2).

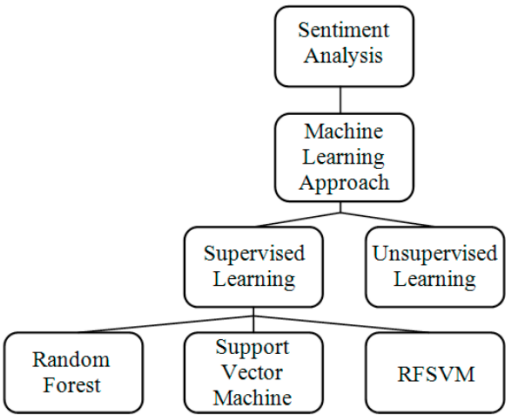


Şekil 2. İkinci veri seti için elde edilen tweet sayısı ve taraflılık düzeyleri.

Sınıflandırma

Bu adımda, hibrit SVM, normalize edilmiş veriler üzerinde sınıflandırma yapmaktadır ve sonuçları göstermektedir. Giriş verilerini işlemeye geçmeden önce, sınıflandırma tekniği çok önemli bir adımdır. Bu noktada, veri seti normal hale getirilmekte ve sınıflandırma algoritması için hazır hale gelmektedir. Böylece sorunsuz bir şekilde çalışmakta ve minimum sürede etkili sonuçlar üretmektedir. Bu çalışmada, Weka ön işleme parametrelerini varsayılan olarak kullanmıştır.

Bu çalışma, Random Forest ve SVM'nin birleştirildiği bir yaklaşım kullanarak denetimli sınıflandırma problemini çözmeyi amaçlamaktadır. Random Forest, bir topluluk öğrenme tekniği olarak bilinen ve test ağacının sınıfını tahmin etmek için rastgele seçilen bir karar ağacı koleksiyonu oluşturan bir algoritmadır. Bu yaklaşım, iki sınıfı ayırmak için bir marj üzerinde odaklanmaktadır. Marjı maksimize ederek, ayırma hiperdüzlemi ile her iki tarafındaki örnekler arasında mümkün olan en geniş alanı yaratmayı hedeflemekte ve tahmin edilen genelleme hatasını azaltmaktadır. Giriş parametreleri Random Forest için hassas olmamakla birlikte her sınıflandırıcı için varsayılan ayarlar kullanıldı. Eğitilmiş değerlendirmeler 0 ile 1 arasında olduğunda, "negatif" veya "pozitif" olarak puan dönüşümü yapılmaktadır. Her bir kombinasyon için öğenin varlığı pozitif (P) veya negatif (N) olarak kabul edilmektedir. Çalışmada kullanılan duygu analizi sınıflandırılması Şekil 3'te sunulmuştur.



Şekil 3. Duygu analizi için kullanılan sınıflandırmanın gösterimi

Önişleme

Veri temizleme, veri birleştirme, veri dönüşümü ve veri azaltma işlemlerinin gerçekleştirildiği bir dizi adımdır. Bu aşama önemlidir çünkü dikkatlice işlenmemiş verilerin analiz edilmesi yanıltıcı bilgiler sağlayabilir. Eğer çok fazla ilgisiz bilgi veya veri gürültüsü varsa, veri kalitesi azalacaktır. Yüksek kaliteli veri üretmek için her vaka çalışması için uygun ön işleme yöntemlerini anlamak ve uygulamak gereklidir. Gerçekleştirilen ön işleme aşamaları veri temizleme, büyük-küçük harf dönüşümü, tokenleştirme, stop word kaldırma, normalleştirme ve kök bulma olarak sıralanabilir. Ön işleme aşaması, sağlanan NLTK kütüphanesi kullanılarak Python'da gerçekleştirilir.

Veri Temizleme

Ön işleme aşamasında ilk adım veri temizlemedir. Twitter'dan elde edilen veriler, DKI Jakarta'daki sel yönetimi konusuyla ilgisi temel alınarak veri temizleme sürecinde seçilir.

Büyük-Küçük Harf Dönüşümü

Bu aşama, yorum metnindeki harflerin yazımdaki düzensiz kullanımından kaynaklanan tutarsız metni düzeltme sürecidir. Bu büyük-küçük harf dönüşümü süreci, temizlenmiş yorum metnindeki harfleri standart bir forma dönüştürmek için kullanılır, yani tüm harfleri küçük harfe dönüştürür.

Kelimelere Ayırma

Bir sonraki aşama tokenize işlemidir. Bu süreçte karakterler, URL'ler, duygusal ifadeler, etiketler ve diğer karakterler kaldırılır. Bu işlem, yazım hatalarını en aza indirmek için duygu analizinde önemli bir işlemdir. Karakterler kaldırıldıktan sonra, cümle birden çok kelimeye bölünür.

Stop Word Kaldırma

Dördüncü aşama stop word kaldırmadır. Bu süreçte cümle içerisinden ayrılmış kelimeler filtrelenir. Anlamı olmayan kelimeler atılır veya silinir. Bu, duygu analizi sürecinde çok yardımcı olur.

Normalleştirme

Bir sonraki aşama kelime normalleştirme sürecidir. Büyük-küçük harf dönüşümü, kelimelere ayıtma ve stop word kaldırma işlemlerinden geçmiş kelimeler tekrar normalleştirilir. Aynı anlama gelen farklı kelimeleri veya argoyu düzeltmek için kullanılır.

Kök Bulma

Bu aşamada, her filtrelenmiş kelimenin kök kelimesini bulmak için kök bulma işlemi gerçekleştirilir. Kök bulma, metin işleme sürecini maksimize etmek ve optimize etmek amacıyla her eklentili kelimenin kök kelimesini bulmak için yapılır.

Model Değerlendirme

Bir sonraki aşama, K-Katlamalı Çapraz Doğrulama tekniğini kullanarak modeli test etmektir. Model performans değerlendirmesi, hata metriklerine dayalı olarak gerçekleştirilir ve model doğruluğunu elde etmek için yapılır. Modelin performansını değerlendirmek için K-katlamalı çapraz doğrulama yöntemi kullanılmıştır. Değerlendirme ve eğitim için kullanılan veri sayısı 10'dur. Veri daha sonra kesin bir değer elde etmek için eğitilir.

Random Forest (Rastgele Orman) Algoritması

Random forest, karar ağacını bir bireysel tahminleyici olarak alan ve Bagging, Randomizing Outputs ve Random Subspace gibi yöntemlere dayanan yöntem seti içinde yer almaktadır. Random forest algoritması, büyük miktarda veriyi doğrulukla sınıflandırabilen sınıflandırma algoritmaları arasında en iyi olanlardan birisi olarak kabul edilmektedir. Sınıflandırma ve regresyon için birçok karar ağacı oluşturan ve eğitim zamanında bu ağaçların çıktısı olan sınıfların modunu veren bir ensemble öğrenme yöntemidir. Ensemble, birden fazla tahminleyiciyi bir araya getirerek daha güçlü ve daha kesin sonuçlar elde etmek için kullanılan bir yöntemdir. Ensable yöntemleri, farklı özelliklere veya örneklemelere dayanan tahminleyicilerin birleştirilmesiyle daha iyi bir genelleme performansı sağlamaktadır (Genuer, 2010).

Random forest sınıflandırma yöntemi, giriş veri setinin alt kümeleme yöntemiyle daha küçük alt kümelerinden birçok sınıflandırıcı oluşturmakta ve daha sonra bu sınıflandırıcıların bireysel sonuçlarını bir oylama mekanizmasıyla birleştirerek giriş veri setinin istenen çıktısını üretmektedir. Bu ensamble öğrenme stratejisi son zamanlarda büyük popülarite kazanmıştır. Random forest'tan önce, Boosting ve Bagging yalnızca iki ensamble öğrenme yöntemi olarak kullanılmaktaydı (Rodríguez-Galiano ve ark., 2011). Random Forest, regresyon ve sınıflandırma problemlerinde kullanılan denetimli bir öğrenme algoritmasıdır. Random Forest, basitçe ağaçların bir koleksiyonu olarak adlandırılmakta ve her bir ağaç birbirinden farklıdır. Çok sayıda karar ağacı oluşturmakta ve sonunda bunları birleştirerek kesin ve istikrarlı bir değer elde etmektedir. Bu değerler genellikle eğitim ve sınıf çıktısının elde edildiği zamanda kullanılmaktadır.

SVM (Destek Vektör Makinesi) Algoritması

SVM, sınıflar arasındaki mesafeyi maksimize ederek en iyi hiperdüzlemi bulmak için kullanılır. Hiperdüzlem, sınıfları ayırmak için kullanılan bir fonksiyondur. İki boyutta öğeleri kategorize etmek için çizgiler, üç boyutta nesneleri kategorize etmek için düzlemler ve yüksek boyutlu sınıf uzaylarında nesneleri kategorize etmek için hiperdüzlemler kullanılır. SVM tarafından bulunan hiperdüzlem, iki sınıfı ayırır, yani en dış sınıftan veri öğelerinden daha uzakta konumlanır. Destek vektör, hiperdüzleme en yakın olan en dış veri öğesidir.

Metriklerin Hesaplanması

Doğru Pozitif (True Positive-TP), Doğru Negatif (True Negative- TN), Yanlış Pozitif (False Positive-FP) ve Yanlış Negatif (False Negative-FN) metrikleri kullanılarak Kesinlik, Geri Çağırma ve F1-Skoru hesaplamaları yapılabilmektedir.

Kesinlik metriği (Precision) modelin pozitif olarak tahmin ettiği örneklerin gerçekten pozitif olma olasılığın ölçen bir metriktir (Denklem 1) (Polat ve Ağca, 2022).

Geri çağırma (Recall), modelin doğru olarak sınıflandırdığı pozitif örnek sayısının tüm pozitif örneklere oranlanması ile bulunabilmektedir (Denklem 2).

F1-Skoru(F1-Measure), Kesinlik ve Geri çağırma metriklerinin harmonik ortalamasıdır (Denklem 3).

Doğruluk (Accuracy) modelin doğru tahminlerinin tüm tahminlere oranıyla elde edilmektedir (Denklem 4) (Çelik ve ark., 2021).

$$Kesinlik = \frac{TP}{TP+FP} \quad (1)$$

$$Geri\ Çağırma = \frac{TP}{TP+FN} \quad (2)$$

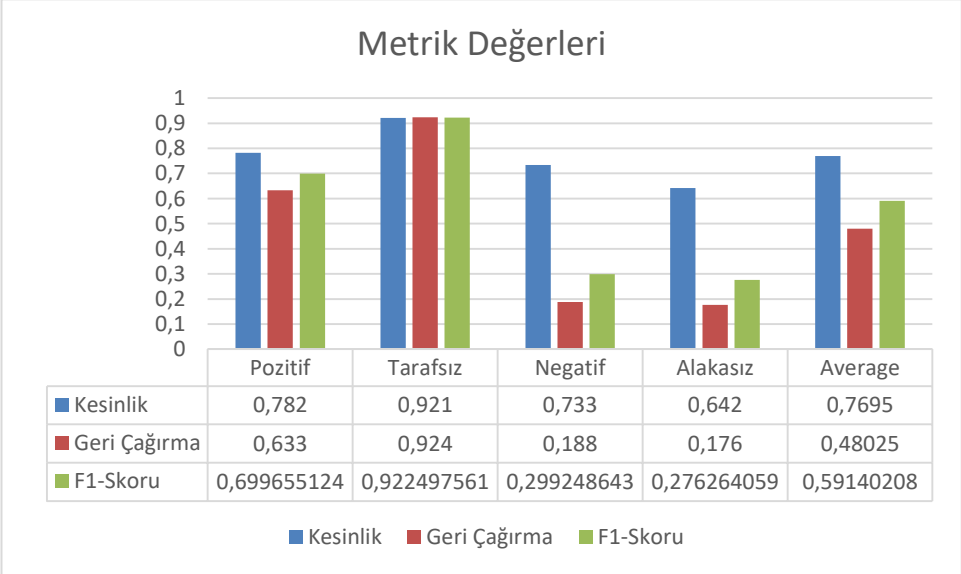
$$F1 - Skoru = 2x\frac{Kesinlik\ x\ Geri\ Çağırma}{Kesinlik+Geri\ Çağırma} \quad (3)$$

$$Doğruluk = \frac{TP+TN}{TP+FP+TN+TF} \quad (4)$$

Bulgular

Bu bölümde her iki veri seti için, hibrit SVM'nin sonuçlarını ve karşılaştırmalı analizi açıklanmaktadır. Bu çalışma karşılaştırma için üç değerlendirme metriği kullanmaktadır. Bu metrikler, Kesinlik (Precision), Geri çağırma (Recall) ve F-skoru (F Measure) olarak isimlendirilmektedir.

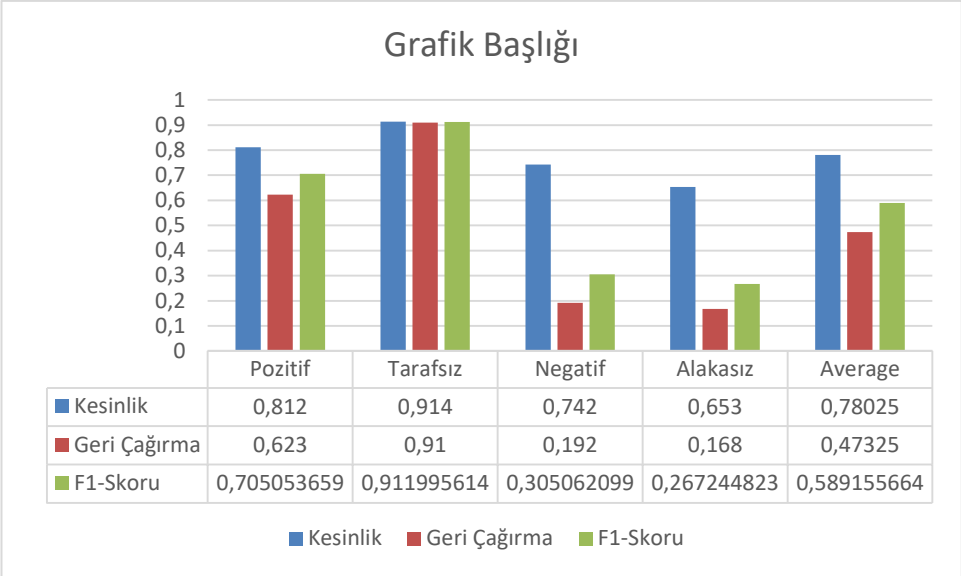
İlk veri seti, film incelemeleriyle ilgili tweetleri içermektedir (VT1). Verilere ve değerlendirmeye göre, ortalama Kesinlik (Precision) %76,95 Geri Çağırma (Recall) %48,02 ve F-Ölçütü (F-Measure) %59,14 olarak belirlenmiştir (Şekil 4).



	Pozitif	Tarafsız	Negatif	Alakasız	Average
Kesinlik	0,782	0,921	0,733	0,642	0,7695
Geri Çağırma	0,633	0,924	0,188	0,176	0,48025
F1-Skoru	0,699655124	0,922497561	0,299248643	0,276264059	0,59140208

Şekil 4. Birinci veri seti için elde edilen metrik değerleri

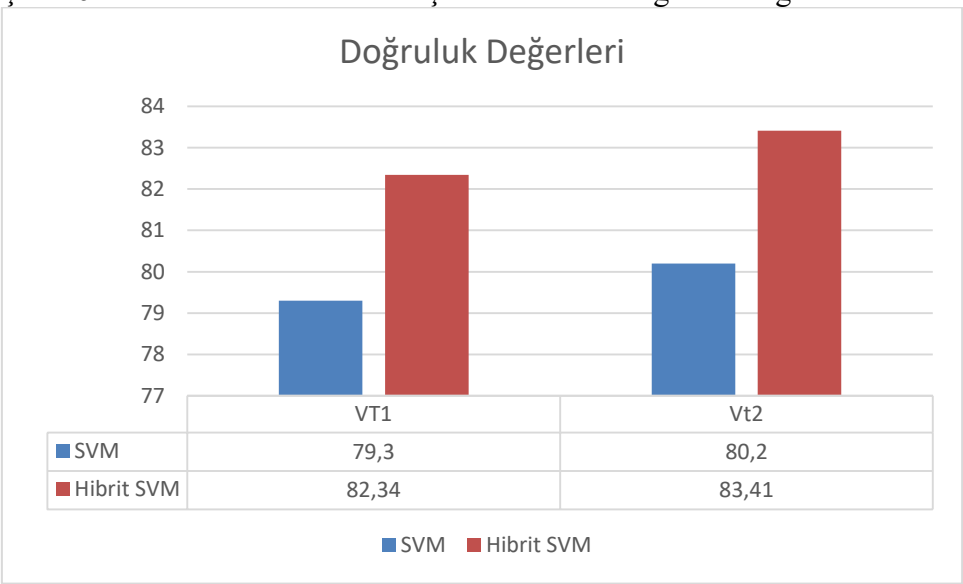
Bir sonraki veri seti, çevrimiçi eğitim süreciyle alakalı Twitter yorumlarını içermektedir (VT2). Ortalama Kesinlik (Precision) %78,02, Geri çağırma (Recall) %47,32 ve F-Ölçütü (F-Measure) %58,91 olarak belirlenmiştir. Şekil 5, her ölçüt için kategorilere göre sonuçları gösteren grafiklerle birlikte tüm verileri içermektedir (Şekil 5).



	Pozitif	Tarafsız	Negatif	Alakasız	Average
Kesinlik	0,812	0,914	0,742	0,653	0,78025
Geri Çağırma	0,623	0,91	0,192	0,168	0,47325
F1-Skoru	0,705053659	0,911995614	0,305062099	0,267244823	0,589155664

Şekil 5. İkinci veri seti için elde edilen metrik değerleri

Şekil 6'da SVM ve Hibrit SVM için elde edilen doğruluk değerleri sunulmuştur.



	VT1	Vt2
SVM	79,3	80,2
Hibrit SVM	82,34	83,41

Şekil 6. SVM ve Hibrit SVM için elde edilen doğruluk (Accuracy) değerleri

Tartışma ve Sonuç

Elde edilen sonuçlara göre VT1 için ve VT2 için en yüksek kesinlik değerleri Tarafsız yorumlardan elde edilmiştir. VT1 için bu değer, %92 olarak, VT2 için %91 olarak ölçülmüştür. En düşük kesinlik değerleri Alakasız yorumlarda elde edilmiştir. VT1 için %64, VT2 için %65 olarak ölçülmüştür.

SVM için VT1 %79,3 doğruluk oranı verirken, VT2 %82,34'lük bir doğruluk oranı vermiştir. Hibrit SVM ile yapılan ölçümlerde VT1 %80,2 ve VT2 %83,41'lik

doğruluk değerine ulaşmıştır. Hibtir SVM modelinin , SVM modeline göre daha doğru sonuçlar üretildiği açıkça görülmektedir.

Bu çalışmada, duygu analizi için hibrit SVM performansı incelenmiştir. Hibrit SVM performans analizi için iki önceden sınıflandırılmış özelleştirilmiş tweet veri seti kullanılmıştır. VT1 film incelemeleri hakkında tweetleri içerirken, VT2 çevrimiçi eğitim süreciyle alakalı tweetleri içermektedir. Karşılaştırma ve analiz için Weka aracı kullanılmıştır. Sonrasında, sonuçları değerlendirmek için kesinlik (Precision), hatırlama (recall) ve F-ölçütü (f-measure) metrikleri kullanılmıştır. Hibrit SVM'nin performansı SVM'den daha iyidir, çünkü sonuçlar açıkça göstermektedir ki hibrit SVM'nin performansı girdi veri setinden etkilenmektedir. Büyük ve çeşitli veri setleri kullanarak, SVM ve diğer makine öğrenimi yaklaşımları performansı iyileştirmek için daha fazla incelenebilir. Bu çalışmanın sonuçları karşılaştırmalı analiz için bir temel olarak kullanılabilir.

Kaynaklar

Appel, O., Chiclana, F., Carter, J., & Fujita, H. (2016). A hybrid approach to sentiment analysis. In IEEE Congress on Evolutionary Computation.

Beleveslis, D., Tjortjis, C., Psaradelis, D., & Nikoglou, D. (2019). A Hybrid Method For Sentiment Analysis Of Election Related Tweets. In 4th SouthEast Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM).

Çelik, E., Dal, D.., & Aydın, T. (2021). Duygu Analizi İçin Veri Madenciliği Sınıflandırma Algoritmalarının Karşılaştırılması. Avrupa Bilim ve Teknoloji Dergisi, (27), 880-889.

Erşahin, B., Aktaş, Ö., Kılınç, D., & Erşahin, M. (2019). A hybrid sentiment analysis method for Turkish. Turkish Journal of Electrical Engineering and Computer Science, 27, 1780–1793.

Genuer, R. (2010). Forêts aléatoires: aspect théoriques, sélection de variables et applications (Thèse de Doctorat Mathématiques, Université de Paris-Sud XI).

Liu, S., Li, F., Li, F., Cheng, X., & Shen, H. (2013). Adaptive co-training SVM for sentiment classification on tweets. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management ACM.

Mudinas, A., Zhang, D., & Levene, M. (2012). Combining lexicon and learning based approaches for concept-level sentiment analysis. In Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining.

Naz, S., Sharan, A., & Malik, N. (2018). Sentiment Classification On Twitter Data Using Support Vector Machine. In IEEE/WIC/ACM International Conference on Web Intelligence (WI).

Ohana, B., & Tierney, B. (2009). Sentiment classification of reviews using SentiWordNet. In 9th IT&T Conference.

Polat, H., & Ağca, Y. (2022). TripAdvisor Kullanıcılarının Türkçe ve İngilizce Yorumları Kapsamında Duygu Analizi Yöntemlerinin Karşılaştırmalı Analizi. Abant Sosyal Bilimler Dergisi, 22(2), 901-916.

Rodríguez-Galiano, V. F., Abarca-Hernández, F., Ghimire, B., Chica-Olmo, M., Akinson, P. M., & Jeganathan, C. (2011). Incorporating Spatial Variability Measures in Land-cover Classification using Random Forest. Procedia Environmental Sciences, 3, 44-49.

Sham, N. M., & Mohamed, A. (2022). Climate Change Sentiment Analysis Using Lexicon, Machine Learning and Hybrid Approaches. Sustainability, 14(8), 4723-4751. DOI: 10.3390/su14084723.

Türkmenoğlu, C. (2015). Türkçe Metinlerde Duygu Analizi (Yüksek Lisans Tezi, Bilgisayar Mühendisliği Anabilim Dalı, İstanbul Teknik Üniversitesi).

Zainudin, S., Jasim, D. S., & Bakar, A. A. (2016). International Journal on Advanced Science, Engineering and Information Technology, 6(6), 1148-1153.