METİN ANALİZİ İÇİN MAKİNE ÖĞRENİM HATLARI

Yazılım Mühendisliği Ana

Bilim Dalı

Dönem Projesi

Nurettin Selçuk

ORCID 0009-0003-4910-3887

Proje Danışmanı: Doç. Dr. Sıla Övgü Korkut Uysal

Mayıs 2024

Metin Analizi İçin Makine Öğrenim Hatları

Özet

Bu çalışmanın amacı, farklı makine öğrenmesi algoritmaları ile oluşturulmuş metin analizi hatlarının metin sınıflandırma konusundaki yeteneklerinin ölçülmesi ve karşılaştırılmasıdır.
Proje işletme yorumlarının 1 ila 5 puan arasında sınıflandırılması ve performanslarının karşılaştırılması üzerinedir.

Anahtar Sözcükler: Makine öğrenmesi, Svm, Karar ağaçları, Rastgele Ormanlar, Lojistik Regresyon, KNN

Machine Learning Pipelines for Text Analysis

Abstract

The aim of this study is to measure and compare the capabilities of text analysis pipelines created with different machine learning algorithms in the context of text classification. The project focuses on classifying business reviews with ratings from 1 to 5 and comparing their performance.

Keywords: Machine learning, SVM, Decision trees, Random forests, Logistic regression, KNN.

İçindekiler

Şekiller Listesi

Bölüm 1

Giriş

Metin analizi, günümüzde giderek artan veri hacmi ve bu verinin anlamlandırılması gerekliliği nedeniyle önemli bir araştırma alanı haline gelmiştir. Bu alanda yapılan çalışmalar çeşitli metinlerden anlamlı bilgi çıkarımını, duygu analizini, konu modellemeyi ve daha birçok uygulamayı kapsamaktadır. .Metin analizi pazarlama stratejilerinden müşteri memnuniyetine, akademik araştırmalardan sosyal medya analizlerine kadar geniş bir yelpazede kullanılarak büyük bir değer yaratmaktadır. Yapılan analizlerden bir tanesi saat 16:00 civarında benzin istasyonlarına uğrayan müşterilerin devam eden bir saatlik süre içerisinde restoran ve alışveriş merkezlerini kullandığını göstermiştir (Schonberger ve Cukier, 2013).

Literatürde, metin analizi için kullanılan çeşitli makine öğrenimi yöntemleri ve modelleri bulunmaktadır. Örneğin Destek Vektör Makineleri (SVM) ve Lineer SVC yüksek boyutlu verilerde etkili performans göstermesi ile bilinir. Destek Vektör Makineleri, yapısal risk minimizasyonu prensibine dayanan ve dış bükey optimizasyona dayalı bir makine öğrenmesi algoritmasıdır. Bu algoritma, veriye ait birleşik dağılım fonksiyonu bilgisine ihtiyaç duymadığı için dağılımdan bağımsız olarak çalışabilen öğrenme algoritmalarından biridir (Soman, Loganathan ve Ajay, 2011). Karar Ağaçları ve Rastgele Orman algoritmaları veri içindeki karar noktalarını anlamlandırmada güçlü araçlardır. Lojistik Regresyon ikili sınıflandırma problemlerinde yaygın olarak kullanılırken, K-En Yakın Komşu (KNN) algoritması benzerlik bazlı sınıflandırma yöntemleri için tercih edilir. Sınıflandırma, bir nesnenin sahip olduğu özellikler temelinde hangi kategoriye ait olduğunu belirlemeye yönelik bir işlemdir. Denetimli öğrenme algoritmalarından olan sınıflandırma algoritmaları, mevcut verilerdeki desenleri keşfeder ve yeni eklenen nesnelerin hangi sınıfa ait olacağını tahmin eder. Ayrıca, bu yöntemlerin geliştirilmesinde bulanık mantık da önemli bir rol oynamaktadır (Bayrakçı, 2015, s. 98). Bu

yöntemlerin her biri farklı veri setleri ve problemler üzerinde çeşitli başarı oranlarına sahiptir ve akademik literatürde geniş bir yer tutar.

Bu proje, mevcut çalışmalardan farklı olarak metin analizi için çeşitli makine öğrenimi algoritmalarını karşılaştırmalı olarak incelemeyi hedeflemektedir. Özellikle Destek Vektör Makineleri (SVM) ile Lineer SVC, Karar Ağaçları, Rastgele Orman, Lojistik Regresyon ve K-En Yakın Komşu (KNN) algoritmalarını kullanarak modeller oluşturulacaktır. Bu modellerin performansları belirli metin veri setleri üzerinde değerlendirilecek ve karşılaştırılacaktır.

Bu projenin ilk bölümünde, çalışmanın amacı, metin analizi ve makine öğrenimi algoritmalarına genel bir giriş yapılacaktır. İkinci bölümde projede kullanılan veri setleri ve metodoloji detaylandırılacaktır. Üçüncü bölümde ise, farklı makine öğrenimi algoritmalarının performans sonuçları ve karşılaştırmaları sunulacaktır. Son olarak, proje bulguları ve gelecekte yapılabilecek çalışmalar hakkında kısa bir değerlendirme yapılacaktır.

1.1.Çalışmanın Amacı

Bu çalışmanın amacı farklı algoritmaların metin analizi üzerindeki etkilerini ve performanslarını daha iyi anlamamızı sağlayarak işletmelerin büyük metin verilerini işlemesi için oluşturması gereken hatları oluşturma süreçlerine ışık tutmaktır.

1.2.Çalışmanın Önemi

Günümüz dünyasında ilerleyen teknolojiler ile birlikte ortalama bir insanın bütün işlemleri teknoloji üzerinden yürümektedir bu da sürekli artan bir veri oluşturmaktadır. İşletmelerin kullanıcılarını ve kullanıcılarının davranışlarını daha iyi analiz ederek bu yönde stratejiler belirlemesi işletmelerin geleceği için önemlidir. Büyük verinin analizi bu stratejilerin belirlenmesinde önemli bir rol oynamaktadır, bu çalışma verilerin analizi için önemli bilgiler içermektedir.

1.3. Tanımlar

Büyük veri; Büyük veri, çeşitli kaynaklardan hızla artan devasa veri kümelerini ifade eder. Bu veriler geleneksel veri işleme yöntemleriyle işlenemeyecek kadar karmaşık ve büyüktür (IBM, n.d.).

Veri işleme hattı; Ham verileri makine öğrenmesi algoritmaları ile işleyip analiz ederek işletmelere değerli bilgiler sunan bir dizi adımdan oluşan bir sistemdir (Amazon Web Services, n.d.).

Destek Vektör Makineleri (SVM); Sınıflandırma ve regresyon problemleri için kullanılan, verileri bir hiperplan ile iki sınıfa ayıran ve ayırıcı çizginin mümkün olduğunca geniş olmasını hedefleyen bir modeldir (scikit-learn, n.d.).

Lineer SVC; SVM'in doğrusal bir versiyonudur ve doğrusal olarak ayrılabilen problemler için yoğunlukla kullanılan bir makine öğrenmesi algoritmasıdır, hızlı ve hesaplama açısından ucuzdur (scikit-learn, n.d.).

Karar Ağaçları; Verileri bir dizi kurala göre alt kümelere ayıran bir makine öğrenemesi algoritmasıdır. Her kural bir özelliğe ve bir değere dayalıdır (scikit-learn, n.d.).

Rastgele Orman; Birden fazla karar ağacından oluşan bir makine öğrenmesi algoritmasıdır. Her ağac'ın farklı bir alt küme ile etkileşime geçmesi ile oluşur (scikit-learn, n.d.).

Lojistik Regresyon; Sigmoid fonksiyonu yardımı ile olasılıkların hesaplanması üzerine olan bir makine öğrenmesi algoritmasıdır (scikit-learn, n.d.).

K-En Yakın Komşu (KNN); Verilerin noktalandırılması üzerine bir makine öğrenmesi algoritmasıdır, yeni veriyi noktalandırmak için en yakın K veri noktasına bakar, noktaların sınıflandırılması üzerinedir (scikit-learn, n.d.).

Kesinlik; Modelin pozitif olarak tahmin ettiği örneklerin kaçının gerçekten pozitif olduğunu gösterir. Yani, doğru pozitif tahminlerin toplam pozitif

tahminlere oranıdır. Formülü: TP / (TP + FP), burada TP doğru pozitif, FP ise yanlış pozitif tahminlerdir (Öğündür, n.d.).

Duyarlılık; Gerçek pozitif örneklerin kaç tanesinin model tarafından doğru bir şekilde pozitif olarak tahmin edildiğini gösterir. Yani, doğru pozitif tahminlerin toplam gerçek pozitiflere oranıdır. Formülü: TP / (TP + FN), burada FN yanlış negatif tahminlerdir. (Öğündür, n.d.).

F1-Skoru; Precision ve Recall'un harmonik ortalamasıdır. Precision ve Recall arasında bir denge sağlar. Yüksek Precision ve yüksek Recall değerlerine sahip bir modelin iyi performans gösterdiğini belirtir. Formülü: 2 * (Precision * Recall) / (Precision + Recall) (Öğündür, n.d.).

Bölüm 2

Materyal ve Metod

Bu bölümde, projenin materyalleri ve metodolojisi sunulmaktadır. Bu projenin amacı, belli makine öğrenmesi algoritmalarını kullanarak, Kaggle'dan üzerinden elde edilen Yelp işletme yorumları veri seti üzerinde bir puan ve beş puan arasında sınıflandırma yapmaktır. Bu bölümde veri seti, kullanılan makine öğrenimi modelleri, değerlendirme ölçütleri ve genel çalışma metodolojisi ele alınmaktadır.

2.1 Veri Toplama

Bu projede kullanılan veriler Kaggle platformu üzerinde paylaşılan Yelp işletme yorumları veri setinden alınmıştır. Veri seti içerisinde “review_id”, “used_id”, “bussiness_id”, “stars”, “useful”, “funny”, “cool” ve “text” alanlarından oluşan Json formatındadır.

2.2 Ön İşleme

Veri seti, büyük bir JSON dosyasından satır satır okunarak pandas DataFrame'e dönüştürülmüştür. Kaynak kullanımı sebebi ile veri seti üzerinden1000 adet veri kullanılmıştır. Yorumlar “text” ve yıldız puanları (stars) sütunları bağımsız ve bağımlı değişkenler olarak ayrılmıştır. Veri seti, eğitim ve test setlerine bölünmeden önce metin verisi TfidfVectorizer kullanılarak sayısal vektörlere dönüştürülmüştür.

2.3. Makine Öğrenmesi Modelleri

Verilerin yıldızlarının sınıflandırılması için beş farklı makine öğrenmesi algoritması kullanılarak modeller oluşturuldu. Her bir model, TfidfVectorizer ile birlikte bir hat içerisinde oluşturulmuş ve eğitim seti üzerinde eğitilmiştir.

- Desktek Vektör Makinaları (SVM)
- K- En Yakın Komşu (KNN)
- Lojistik Regresyon
- Karar Ağaçları
- Rastgele Ormanlar
-

2.4. Model Değerlendirme

Modellerin performansı, test seti üzerinde doğruluk (accuracy) ve sınıflandırma raporu (classification report) ile değerlendirilmiştir. Sınıflandırma raporu, her bir sınıf için hassasiyet (precision), duyarlılık (recall) ve F1 skoru gibi metrikleri içermektedir.

Bölüm 3

Sonuçlar ve Tartışma

3.1 Modellerin Sonuçları

Verilerin ön işlemeleri tamamladıktan sonra modeller 80'e 20 oranında eğitim verisi ve test verisi olarak bölünerek modeller eğitildi. TfidfVectorizer ile vektörleştirilen veriler K-En yakın Komuş sınıflandırıcısı, Rastgele Orman Sınıflandırıcısı, Destek Vektör Sınıflandırıcısı ve Lojistik Regrasyon, Destek vektör makineleri algoritmaları ile eğitildi. Her bir model için doğruluk oranı ve sınıflandırma raporu oluşturuldu, bu bölümde her model ve sınıflandırma raporları sonuçları sunularak sonuçlar üzerinde tartışılmıştır.

3.1.1 Destek Vektör Makineleri

	Precision	Recall	F1-score	Support
1.0	0.68	0.75	0.71	115
2.0	0.43	0.16	0.24	73
3.0	0.35	0.25	0.29	116
4.0	0.45	0.39	0.42	263
5.0	0.65	0.81	0.72	435
Accuracy			0.58	1002
Macro avg	0.51	0.47	0.48	1002
Weighted avg	0.55	0.58	0.56	1002

Şekil 3.1.1: Destek Vektör Makineleri Sınıflandırma Raporu

Modelin genel doğruluk oranı 0.578 yani %57,8 çıkmıştır, model için aynı zamanda sınıflandırma raporu incelendiğinde;

1 Yıldız için; 115 destek üzerinden 0,68 kesinlik, 0,75 duyarlılık, 0,71 F1-

skoru elde etmiştir.

2 Yıldız için; 73 destek üzerinden 0,43 kesinlik, 0,16 duyarlılık, 0,24 F1-skoru elde etmiştir.

3 Yıldız için; 116 destek üzerinden 0,35 kesinlik, 0,25 duyarlılık, 0,29 F1-skoru elde etmiştir.

4 Yıldız için; 263 destek üzerinden 0,45 kesinlik, 0,39 duyarlılık, 0,42 F1-skoru elde etmiştir.

5 Yıldız için; 435 destek üzerinden 0,65 kesinlik, 0,81 duyarlılık, 0,72 F1-skoru elde etmiştir.

Modelin genel ortalama doğruluk oranı %58 iken Makro ortalama kesinlik %51, duyarlılık %43 ve F1 skoru 0,48 olmuştur. Ağırlıklı ortalama için ise kesinlik değeri %55, duyarlılık %58, F1 skoru %56 oranındadır.

3.1.2 K-En Yakın Komşu

	Precision	Recall	F1-score	Support
1.0	0.36	0.67	0.47	115
2.0	0.11	0.08	0.09	73
3.0	0.20	0.18	0.19	116
4.0	0.40	0.33	0.36	263
5.0	0.62	0.58	0.60	435
Accuracy			0.44	1002
Macro avg	0.34	0.37	0.34	1002
Weighted avg	0.45	0.44	0.44	1002

Şekil 3.1.2: K-En Yakın Komşu Sınıflandırma Raporu

Modelin genel doğruluk oranı 0.443 yani %44,3 çıkmıştır, model için aynı zamanda sınıflandırma raporu incelendiğinde;

1 Yıldız için; 115 destek üzerinden 0,36 kesinlik, 0,67 duyarlılık, 0,47 F1-skoru elde etmiştir.

2 Yıldız için; 73 destek üzerinden 0,11 kesinlik, 0,08 duyarlılık, 0,09 F1-skoru elde etmiştir.

3 Yıldız için; 116 destek üzerinden 0,20 kesinlik, 0,18 duyarlılık, 0,19 F1-skoru elde etmiştir.

4 Yıldız için; 263 destek üzerinden 0,40 kesinlik, 0,33 duyarlılık, 0,36 F1-skoru elde etmiştir.

5 Yıldız için; 435 destek üzerinden 0,62 kesinlik, 0,58 duyarlılık, 0,60 F1-skoru elde etmiştir.

Modelin genel ortalama doğruluk oranı %44 iken Makro ortalama kesinlik %34, duyarlılık %37 ve F1 skoru 0,34 olmuştur. Ağırlıklı ortalama için ise kesinlik değeri %45, duyarlılık %44, F1 skoru %44 oranındadır.

3.1.3 Lojistik Regresyon

	Precision	Recall	F1-score	Support
1.0	0.73	0.70	0.71	115
2.0	0.38	0.08	0.13	73
3.0	0.38	0.16	0.23	116
4.0	0.42	0.35	0.38	263
5.0	0.62	0.87	0.73	435
Accuracy			0.58	1002
Macro avg	0.51	0.43	0.44	1002
Weighted avg	0.54	0.58	0.53	1002

Şekil 3.1.3: Lojistik Regresyon Sınıflandırma Raporu

Modelin genel doğruluk oranı 0.576 yani %57,6 çıkmıştır, model için aynı zamanda sınıflandırma raporu incelendiğinde;

1 Yıldız için; 115 destek üzerinden 0,73 kesinlik, 0,70 duyarlılık, 0,71 F1-skoru elde etmiştir.

2 Yıldız için; 73 destek üzerinden 0,38 kesinlik, 0,08 duyarlılık, 0,13 F1-skoru elde etmiştir.

3 Yıldız için; 116 destek üzerinden 0,38 kesinlik, 0,16 duyarlılık, 0,23 F1-skoru elde etmiştir.

4 Yıldız için; 263 destek üzerinden 0,42 kesinlik, 0,35 duyarlılık, 0,38 F1-skoru elde etmiştir.

5 Yıldız için; 435 destek üzerinden 0,62 kesinlik, 0,87 duyarlılık, 0,73 F1-skoru elde etmiştir.

Modelin genel ortalama doğruluk oranı %58 iken Makro ortalama kesinlik %51, duyarlılık %43 ve F1 skoru 0,44 olmuştur. Ağırlıklı ortalama için ise kesinlik değeri %54, duyarlılık %58, F1 skoru %53 oranındadır.

3.1.4. Karar Ağaçları

	Precision	Recall	F1-score	Support
1.0	0.40	0.42	0.41	115
2.0	0.19	0.15	0.17	73
3.0	0.27	0.26	0.26	116
4.0	0.33	0.29	0.31	263
5.0	0.59	0.64	0.61	435
Accuracy			0.44	1002
Macro avg	0.35	0.35	0.35	1002
Weighted avg	0.43	0.44	0.44	1002

Şekil 3.1.4: Karar Ağacı Sınıflandırma Raporu

Modelin genel doğruluk oranı 0.443 yani %44,3 çıkmıştır, model için aynı zamanda sınıflandırma raporu incelendiğinde;

1 Yıldız için; 115 destek üzerinden 0,40 kesinlik, 0,42 duyarlılık, 0,41 F1-skoru elde etmiştir.

2 Yıldız için; 73 destek üzerinden 0,19 kesinlik, 0,15 duyarlılık, 0,17 F1-skoru elde etmiştir.

3 Yıldız için; 116 destek üzerinden 0,27 kesinlik, 0,26 duyarlılık, 0,26 F1-skoru elde etmiştir.

4 Yıldız için; 263 destek üzerinden 0,33 kesinlik, 0,29 duyarlılık, 0,31 F1-skoru elde etmiştir.

5 Yıldız için; 435 destek üzerinden 0,59 kesinlik, 0,64 duyarlılık, 0,61 F1-skoru elde etmiştir.

Modelin genel ortalama doğruluk oranı %44 iken Makro ortalama kesinlik %35, duyarlılık %35 ve F1 skoru 0,35 olmuştur. Ağırlıklı ortalama için ise kesinlik değeri %43, duyarlılık %44, F1 skoru %44 oranındadır.

3.1.5. Rastgele Ormanlar

	Precision	Recall	F1-score	Support
1.0	0.79	0.33	0.47	115
2.0	0.00	0.00	0.00	73
3.0	0.43	0.03	0.05	116
4.0	0.38	0.14	0.20	263
5.0	0.50	0.98	0.66	435
Accuracy			0.50	1002
Macro avg	0.42	0.29	0.28	1002
Weighted avg	0.46	0.50	0.40	1002

Şekil 3.1.5: Rastgele Ormanlar Sınıflandırma Raporu

Modelin genel doğruluk oranı 0.502 yani %50,2 çıkmıştır, model için aynı zamanda sınıflandırma raporu incelendiğinde;

1 Yıldız için; 115 destek üzerinden 0,79 kesinlik, 0,33 duyarlılık, 0,47 F1-skoru elde etmiştir.

2 Yıldız için; 73 destek üzerinden 0,00 kesinlik, 0,00 duyarlılık, 0,00 F1-skoru elde etmiştir.

3 Yıldız için; 116 destek üzerinden 0,43 kesinlik, 0,03 duyarlılık, 0,05 F1-skoru elde etmiştir.

4 Yıldız için; 263 destek üzerinden 0,38 kesinlik, 0,14 duyarlılık, 0,20 F1-skoru elde etmiştir.

5 Yıldız için; 435 destek üzerinden 0,50 kesinlik, 0,98 duyarlılık, 0,66 F1-skoru elde etmiştir.

Modelin genel ortalama doğruluk oranı %50 iken Makro ortalama kesinlik %42, duyarlılık %29 ve F1 skoru 0,28 olmuştur. Ağırlıklı ortalama için ise kesinlik değeri %46, duyarlılık %50, F1 skoru %40 oranındadır.

3.2. Modellerin Tartışılması

Model değerlendirmesi sonucunda farklı makine öğrenimi algoritmalarının işletme yorumları üzerindeki performansları incelendi. Destek Karar Makineleri, Karar Ağacı, Lojistik Regresyon, K-En yakın komşu ve Rastgele Ormanlar gibi çeşitli algoritmalar kullanılarak sınıflandırma raporları elde edildi. Bu raporlar doğrultusunda her bir modelin güçlü ve zayıf yönleri ortaya konuldu.

İlk olarak SVM modeli %57,8 doğruluk oranı ile dikkate değer bir performans sergilemiştir. Özellikle 1 ve 5 yıldız sınıflarında yüksek doğruluk oranları elde edilmiştir bununla birlikte, 2, 3 ve 4 yıldız sınıflarında modelin performansı düşüktür. SVM modelinin makro ve ağırlıklı ortalamaları sınıflandırma görevinde belirli sınıflarda iyi performans gösterdiğini ancak genel olarak iyileştirilmesi gerektiğini ortaya koymaktadır.

Karar ağaçları modeli ise %44,3 doğruluk oranı ile daha düşük bir performans göstermiştir. 1 ve 5 yıldız sınıflarında kabul edilebilir doğruluk oranları içinde olsa bile diğer sınıflarda modelin performansı oldukça düşüktür. Özellikle 2 yıldız sınıfında duyarlılık ve F1-skoru çok düşüktür. Bu durum Karar ağacı modelinin belirli sınıflar için uygun olmadığını göstermektedir.

Lojistik Regresyon modeli %57,6 doğruluk oranı ile SVM modeline yakın bir performans sergilemiştir. 1 ve 5 yıldız sınıflarında iyi sonuçlar elde edilmiştir. Ancak, 2, 3 ve 4 yıldız sınıflarında modelin performansı yine düşüktür. Logistic Regression modelinin makro ve ağırlıklı ortalamaları modelin genel olarak dengeli bir performans sergilediğini ancak belirli sınıflarda iyileştirme gerektiğini göstermektedir.

K-en yakın komşu modeli ise %44,3 doğruluk oranı ile Karar Ağacı modeline benzer bir performans sergilemiştir. 1 ve 5 yıldız sınıflarında kabul edilebilir sonuçlar göstersede diğer sınıflarda modelin performansı düşüktür. Özellikle 2 yıldız sınıfında duyarlılık ve F1-skoru oldukça düşüktür. Bu sonuçlar KNN modelinin belirli sınıflar için uygun olmadığını göstermektedir.

Son olarak, Rastgele Ormanlar modeli %50,2 doğruluk oranı ile orta düzeyde bir performans sergilemiştir. 5 yıldız sınıfında oldukça yüksek bir duyarlılık elde edilmiştir. Lakin, diğer sınıflarda modelin performansı düşüktür. Özellikle 2 ve 3 yıldız sınıflarında duyarlılık ve F1-skoru çok düşüktür. Rastgele ormanlar

modelinin makro ve ağırlıklı ortalamaları, modelin belirli sınıflarda iyi performans gösterdiğini ancak genel olarak iyileştirilmesi gerektiğini ortaya koymaktadır.

Genel olarak bu analiz her bir modelin farklı sınıflarda farklı performans sergilediğini göstermektedir. SVM ve Logistic Regression modelleri genel doğruluk oranları ve belirli sınıflardaki performansları ile öne çıkarken Karar Ağacı, K-En yakın komşu ve rastgele ormanlar modelleri belirli sınıflarda düşük performans göstermektedir. Bu sonuçlar model seçimi ve iyileştirme süreçlerinde dikkate alınması gereken önemli faktörler sunmaktadır. Özellikle düşük performans gösteren sınıflar için veri ön işleme ve model optimizasyonu gibi adımların atılması ve veri miktarının daha iyi kaynaklar kullanılarak arttırılması görüşü ortaya çıkmaktadır.

Bölüm 4

Sonuç

Sonuç olarak, Yelp işletme verileri üzerinden hazırlanan ve yorumların 1-5 yıldız üzerinden sınıflandırılmasını hedefleyen bu çalışmada farklı makine öğrenmesi algoritmaları ile makine öğrenmesi hatları oluşturulmuş ve sonuçlar karşılaştırılarak modellerin güçlü ve zayıf yönleri ortaya koyulmuştur.

Modeller 1 ve 5 yıldız sınıflarında yüksek doğruluk oranlarını yakalasada düşük kalan destek sebebi ile 2,3 ve 4 yıldız sınıflarında düşük doğruluk oranlarında kalmıştır.

Bu çalışma Destek Vektör Makineleri ve Lojistik regresyon modellerinin diğer modellere göre daha güçlü sonuçlar elde etmesi ile öne çıkarken, gelecekteki yapılacak çalışmalar için K-En yakın komşu, Karar Ağaçları, Rastgele Ormanlar algoritmaları için daha iyi bir veri seti gerektiğini ortaya koymuştur.

Kaynaklar

Amazon Web Services. (n.d.). What is AWS Data Pipeline? Amazon Web Services.

Bayrakçı, S. (2015). Sosyal Bilimlerdeki Akademik Çalışmalarda Büyük Veri Kullanımı. Yayınlanmamış Yüksek Lisans Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü.

Gülcan Öğündür (n.d.). Doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ya da F1-score. Medium.

https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/what-is-datapipeline.html

https://medium.com/@gulcanogundur/do%C4%9Fruluk-accuracy-kesinlik-precision-duyarl%C4%B1l%C4%B1k-recall-ya-da-f1-score-300c925feb38

IBM. (n.d.). Big data analytics. IBM. https://www.ibm.com/topics/big-data-analytics

Mayer-Schonberger, V. ve Cukier, K. (2013). Big Data: A Revolution That Will Transform How We Live, Work and Think. Boston, Massachusetts: Houghton Mifflin Harcourt

Scikit-learn. (n.d.). User guide. Scikit-learn. https://scikit-learn.org/stable/user_guide.html

Soman, K.P., Loganathan, R. and Ajay, V. (2011). Machine learning with SVM and other kernel methods. PHI Learning Pvt. Ltd., s. 486.