**T.C.**
**İSTANBUL MEDENİYET ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**
**UYGULAMALI MATEMATİK VE HESAPLAMALI BİLİMLER BİLİM DALI**

# Müşteri Yorumları Üzerinde Metin Analitiği Çalışmaları ve Yorumların Makine Öğrenmesi Algoritmaları İle Modellenmesi

**Yüksek Lisans Tezi**

**Muhammed IŞIK**

**Danışman**
**Dr. Öğretim Üyesi Elif CESUR**

**Haziran 2019**

T.C.
İSTANBUL MEDENİYET ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
UYGULAMALI MATEMATİK VE HESAPLAMALI BİLİMLER BİLİM DALI

# Müşteri Yorumları Üzerinde Metin Analitiği Çalışmaları ve Yorumların Makine Öğrenmesi Algoritmaları İle Modellenmesi

Yüksek Lisans Tezi

Muhammed IŞIK

Danışman
Dr. Öğretim Üyesi Elif CESUR

Haziran 2019

## İMZA SAYFASI

Muhammed IŞIK tarafından hazırlanan ' Müşteri Yorumları Üzerinde Metin Analitiği Çalışmaları ve Yorumların Makine Öğrenmesi Algoritmaları İle Modellenmesi ' başlıklı bu yüksek lisans tezi 'Uygulamalı Matematik ve Hesaplamalı Bilimler ' bilim dalında hazırlanmış ve jürilerimiz tarafından kabul edilmiştir.

| **JÜRİ ÜYELERİ** | **İMZA** |
|---|---|
| **Tez Danışmanı:** | |
| [Dr. Öğr. Üyesi Elif CESUR] | ..... |
| Kurumu: İstanbul Medeniyet Üniversitesi | |
| **Üyeler:** | |
| [Doç.Dr. Rahmet SAVAŞ] | ..... |
| Kurumu: İstanbul Medeniyet Üniversitesi | |
| [Doç. Dr. Özlem ŞENVAR] | ..... |
| Kurumu: Marmara Üniversitesi | |

**Tez Savunma Tarihi:** 21 / 06 / 2019

## BİLDİRİM

Hazırladığım tezin tamamen kendi çalışmam olduğunu, akademik ve etik kuralları gözeterek çalıştığımı ve her alıntıya kaynak gösterdiğimi taahhüt ederim.

Muhammed IŞIK

Danışmanlığını yaptığım işbu tezin tamamen öğrencinin çalışması olduğunu, akademik ve etik kuralları gözeterek çalıştığını taahhüt ederim.

Dr. Öğr. Üyesi Elif CESUR

# ÖZET

Metin analitiği teknolojisi; duygu analizi, kategorilendirme ve metinlerden konu çıkarılması gibi birçok alanda kullanıcılarına fayda sağlamaktadır. Kullanıcı bu teknoloji ile müşterinin yaklaşımını, tercihini ve ihtiyacını anlayabilmekte ve stratejisini koşullara göre revize edebilmektedir. Bu çalışma kapsamında öncelikle müşteri yorumlarının analiz yöntemleri ile incelenmesi gerçekleştirilecektir. Yazılı metin içerisinde bulunan içerikler sayesiyle bilgi keşfi adına analiz çalışmaları yapılacaktır. Metin sınıflandırma problemi, belli tekniklerle işlenen metinlerin içeriğine göre sınıflar atanması sürecidir. Metin sınıflandırma uygulamaları, web sayfalarının ve haber metinlerinin kataloglaması, arama motorlarının optimizasyonu, bilgi çıkarımı, e-postaların otomatik olarak işlenmesi gibi çeşitli alanlarda etkin olarak kullanılmaktadır. Bu bağlamda gerçekleştirilen çalışma da sınıflandırma algoritmasına dayanan yordamlarla müşteri yorumların sınıflandırılması amaçlanmaktadır. Bu çalışma kapsamında oluşturulan modellerin müşteri yorumları üzerindeki sınıflandırma başarısı analiz edilmiştir. Naive bayes, lojistik regresyon ve destek vektör makinesi algoritmalarının yorumlar üzerindeki model performansları karşılaştırılmıştır.

**Anahtar Kelimeler:** Müşteri Yorumları, Metin Analitiği, Metin Sınıflandırma, Makine Öğrenmesi, Sınıflandırma Başarı

## ABSTRACT

Text analytics technology is benefit of users in many areas such as emotion analysis, categorization and text extraction. With this technology, the user can understand the customer's approach, preference and need and revise his strategy according to the conditions. Within the scope of this study, firstly customer reviews will be analyzed with analysis methods. Analysis will be carried out on behalf of information discovery through the contents in the written text. The problem of text classification is the process of assigning classes according to the content of the texts processed with certain techniques. Text classification applications are used effectively in various fields such as cataloging web pages and news texts, optimizing search engines, extracting information, automatically processing e-mails. In this context, it is aimed to classify the customer comments with the procedures based on the classification algorithm. The success of the classification of the models created in this study on customer comments was analyzed. Naive bayes, logistic regression and support vector machine algorithms are compared with the model performances on the interpretations.



**Keywords:** Customer Comments, Text Analytics, Text Classification, Machine Learning, Classification Success

## TEŞEKKÜR

Lisansüstü çalışmalarıma her zaman destek olan anneme ve eşime ve hiçbir zaman yardımlarını esirgemeyen çok değerli hocalarım Doç . Dr. Rahmet SAVAŞ ve Dr . Öğretim Üyesi Elif CESUR'a sonsuz teşekkür ederim.

# İÇİNDEKİLER

## TABLOLAR LİSTESİ

## DENKLEMLER LİSTESİ

## ŞEKİLLER LİSTESİ

## KISALTMALAR

NB: Naive Bayes

LR: Lojistik Regresyon

DVM: Destek Vektör Makineleri

CHI2: Ki Kare İstatistiği

RBF: Radyal Tabanlı Fonksiyon

TFIDF: Terim Frekansı – Ters Doküman Frekansı

# 1. GİRİŞ

## 1.1 Tezin Amacı

Bu tez ile ilk aşamada cümlelerin kelime köklerine kadar ayrışımı (tokenization) yapılacaktır. Anlamlı görülen kelime köklerinin birbirleriyle etkileşimi incelenecektir. Son aşamada ise makine öğrenimi algoritmalarını kullanarak ürünler hakkındaki müşteri yorumlarının modellenmesi amaçlanmaktadır. Elde edilen model kalıbı ile (pattern) ürünler hakkındaki yeni yorumların her biri okunmadan yorumun niteliği tespit edilebilecektir.

## 1.2 Tezin Kapsamı

Bu tez kapsamında, Amazon şirketinde yıllar itibari ile müşteri memnuniyeti ölçülmek istenmektedir. Bilgi kaynağında ürünlerin isimleri, ürünler hakkında yorumlar (metin) ve ürün rating (değerlendirme notu) alanı bulunmaktadır. Ürün yorumları ile ürün değerlendirme notu arasında belli analizler yaparak ilişkiler kurulmaya çalışılacaktır. Literatürde sayısal nitelikte olmayan bu verilerin işlendiği alan 'Metin Madenciliği' olarak bilinmektedir. Makine öğrenimi alanında geniş yer tutan sınıflandırma algoritmaları vasıtasıyla ürünler hakkındaki bu yorumların sınıflandırılması da yapılacaktır.

## 1.3 Tezin Özgün Yanı

Tezin konuları literatürde 'Metin Analitiği've 'Metin Sınıflandırma' olarak geçmektedir. Genelde veriler, veri kaynaklarında sayısal veriler olarak tutulmaktadır. Sayısal nitelikli olmayan verilerde veri kaynaklarında çokça bulunmaktadır. Sayısal olmayan verilerden bilgi kazanımı adına 'Metin Analitiği've 'Metin Sınıflandırma' çalışmaları y apılmaktadır. 'Metin Analitiği've 'Metin Sınıflandırma' literatürde 'Metin Madenciliği' alanının alt dallarıdır . Metin madenciliği gelişmekte olan bir alan olan 'Web Madenciliği' alanına metodolojik ve kuramsal açıdan katkı sağlamaktadır.

# 2. LİTERATÜR

## 2.1 Makine Öğrenmesi

Makine öğrenmesi, örnek veri ya da geçmiş deneyimleri kullanarak bir başarım ölçütünü en uygun hale getirmek için bilgisayarları programlar (Alpaydın, 2004). Makine öğrenmesi, mevcut veri ve geçmiş tecrübeleri değerlendirerek olası yeni durumların analiz ve tahmin çalışmalarında kullanıcılara yardımcı olmaktadır. Makine öğrenmesinde kullanıcı eldeki mevcut verileri uygun şekilde bilgisayara tanıtarak öğrenme sürecini başlatmaktadır. Bu süreçte öğrenme sürecini gerçekleştirecek olan algoritma, veri içerisinden açığa çıkarılacak olan örüntüleri belli analiz yaklaşımları ile belirlemektedir. Bu süreçte en önemli husus kullanılan algoritmaların yapısının çok iyi bilinmesi gerektiğidir. Mevcut algoritma yapılarının bilinmemesi durumunda modelleme süreci sonunda farklı sonuçlar alınmakta ve elde edilen çıktılar yanlış yorumlanabilmektedir.

Makine öğrenmesi algoritmalarında öğrenme sürecine başlamadan önce parametre tayininin yapılması önem arz etmektedir. Parametre tayini bir nevi en iyileme çalışmaları olarak görülmektedir. Uygun parametrelerin tespitinden sonra öğrenme verisi üzerinde model kalıbı oluşturulmakta ve bu kalıp sonraki aşamada test verisi üzerinde test edilmektedir. Test etme sürecinde performans ölçütleri kullanılmaktadır. Performans ölçütleri değerlendirilirken karşılaştırmalı bir yaklaşımda bulunmak gerekmektedir. Test etme sonucunda karşılaştırmalı olarak başarılı görülen model kalıbı tahmin veya analiz çalışmalarında kullanılmaktadır.

Makine öğrenmesi, optik karakter algılama, yüz tanıma, spam e-posta filtrelemesi, konuşma dili anlama, tıbbî teşhis, müşteri segmentasyonu, sahtekârlık tespiti ve hava durumu tahmini gibi birçok farklı problemin çözümünde kullanılmaktadır (Schapire, 2008). Kullanım amacına bağlı olarak öğrenme biçimi seçimi yapılmalıdır. Örneğin müşteri gelirlerine göre göre gruplama yaparken öğrenme biçimi olarak kümelemeden, fraud olayını saptamada ise sınıflandırmadan yararlanılmaktadır. Çalışmalara başlamadan önce üzerinde çalışılan konu ile ilgili geçmiş zamanda yapılmış çalışmalara bakmakta fayda vardır.

## 2.2 Öznitelik Seçme

Belirli sayıda ve sınıfta değişkenleri bazı matematiksel algoritmalar ışığında ağırlıklandıran öznitelik seçme metrikleri , uygulanan veriler içerisindeki en değerli veya ayrıştırıcı değişkenlerin tespiti konusunda çok yardımı olan bir uygulamadır . Öznitelik seçme , makine öğrenimi , veri madenciliği ve istatistik bilim alanları içerisinde çok yaygın olarak kullanılan yöntemlerden biridir. Özellikle verileri sınıflama ve eleme işlemleri için yaygın olarak kullanılır . Öznitelik seçme metriklerinin faydası olarak büyük boyutlu veriler üzerinde yapılan verilerin eliminasyonunda harcanan süre ve maliyeti düşürmesi gösterilebilir . Bunu gerçekleştirirken eğitim setinde sıkça kullanılan ve analiz sonucuna katkı sağlamayan kelimelerin elenmesini sağlar. Böylelikle cümlelerde daha az geçen ve sonucu birebir değiştirme etkisi bulunan kelimelerin tespiti gerçekleştirilir.

Öznitelik seçme metrikleri bu işlemi matematiksel denklemler kullanır ve her bir kelime için ayrı ayrı puanlamalar yaparak önceden tanımlanmış kategoriler için en değerli kelimeleri belirler . Analizi yapılan kelimeleri ise sayısal değerlikler vererek hangisinin en değ erli ve değersiz olduklarını ifade eder . Böylelikle cümleler en sağlıklı bir şekilde ve içerisindeki bilgi kirliliği azaltılmış olarak analiz yapılabilir hale getirilir.

Öznitelik seçme yöntemleri farklı denklemler kullanılarak gerçekleştirilebilme ktedir. Bu denklemlerin başarıları genel olarak birbirlerine yakın oranlardadır . Bu denklemlerdeki farklılıklar daha az veya fazla değişkene bağlı hesaplamalar yapılmasından kaynaklanır. Öznitelik seçme metriklerinin etkililik ve tutarlılık oranı geçmişte yapılan akademik literatür de baz alındığında gayet başarılı , ve çok yaygın olarak halen kullanılmakta olduğu söylenebilir. (Akba, 2014)

## 2.3 Chi2 Özellik Çıkarım Yöntemi

Chi2 testi değişkenler arasında olan ilişkiyi istatistiki olarak tespit etmeye çalışan bir metottur. Chi2 oluşturulan hipotezi referans alıp gerçekte olan değerler ile gözlem sonucunda elde edilen değerleri karşılaştırarak değişkenler arasında ilişkinin var olup olmadığını sınamaktadır.

Nitelik seçiminin bir parçası olarak Chi 2 testi kullanıldığında , belirli bir terimin belirli bir sınıf ile ilişkili olup olmadığını ortay a çıkarmak amacıyla kullanılır (Çürük, 2018). Nitelik (özellik) çıkarım aşamasında ilk olarak ilgili metin seçilen ayrışım yöntemi ile terimlerine ayrıştırılmaktadır. Bu ayrışım işleminde metin durak kelimelerden arındırılır. Bu arındırma işleminde tek başına herhangi bir anlamı olmayan terimler analiz sürecinden çıkarılmalıdır. Bu sürecin akabinde elde edilen terimler ile metnin sınıfı arasında Chi2 testi yapılabilir. Terimler ile sınıflar arasında hesaplamalar yapılarak Chi2 testleri gerçekleştirilir. Chi2 testi ile tespit edilen önemli değişkenler ile sınıflandırma ve kümeleme gibi modelleme çalışmaları yapılmaktadır.

$$Chi2 = \sum_{k=1}^{n} \frac{(Ok - Ek)}{Ek}$$ Denklem 1

Chi2 testi metin analitiği çalışmaları kapsamında sıklıkla başvurulan yöntemlerden biridir. Özellik çıkarımı yaparak örüntülerin belirlenmesinde kullanıcılarına fayda sağlamaktadır.

Örneğin x terimi iyi ve kötü yorumların olduğu bir metinde kullanılmış olsun. x teriminin ilişkili olduğu sınıf Chi2 testi ile belirlenerek terimin özelliği tespit edilebilmektedir. Chi2 özellik çıkarım tekniğinde elde edilen önemli terimler bir araya getirilerek model kurulum sürecinde kullanılmaktadır.

## 2.4 Sınıflandırma Algoritmaları

Gözetimli öğrenme algoritmaları olarak bilinmekte olan sınıflandırma algoritmaları eğitim verisi içerisinde örüntüleri tespit eder ve bu örüntüler yardımıyla oluşturduğu model kalıbı ile test verisindeki bir nesnenin hangi sınıfta bulunacağını tahmin etmeye çalışır.

Bu algoritmalardan başlıcaları; lineer diskriminant analizi, karar ağaçları, yapay sinir ağları, destek vektör makineleri, lojistik regresyon, kNN (k-en yakın komşu), genetik algoritmalar, bellek temelli nedenleme ve naive bayes algoritmasıdır. (Atalay & Çelik, 2013)

### 2.4.1 Naive Bayes

Diğer algoritmalarla karşılaştırmak için temel seviye sınıflandırıcı olarak önerilen iyi bilinen bir istatistiksel öğrenme algoritmasıdır . NB belirli bir sınıf için girdilerin birbirinden bağımsız olduğunu varsayarak sınıf koşulu olasılıkları tarafsız tahmin eder. NB, sınıfları koşullu marjinal yoğunlukları bulmak için ayırıcı sınıflar problemini azaltır , bu da belirli bir örneğin olası hedef sınıflardan biri olma olasılığını temsil eder. NB birbirleri ile ilişkili girdiler içermedikçe diğer algoritmalara karşı iyi performans göstermektedir. (Aydın, 2018)

Genel olarak Naive Bayes sınıflandırıcısı her kriterin sonuca olan etkilerinin olasılık değerlerinin hesaplanmasına dayanır (Çalış, Yıldız, & Gazdağı, 2013). Bu hesaplama sürecinde her bir terimin sınıflarla olan koşullu olasılık değerleri ayrı ayrı hesaplanmaktadır. Koşullu olasılık değeri yüksek gelen terimler bir araya getirilerek model kurulumu tamamlanmaktadır. Naive Bayes teoremi Denklem 2'deki eşitlik ile ifade edilmektedir:

$$P(x/y)=\frac{P(y/x)*P(x)}{P(y)} \qquad \text{Denklem 2}$$

P(x) x olayının olasılığı ve P(y) y olayının olasılığı temsil eder. P(y|x) x olayının gerçekleştiği bilindiğinde y olayının şartlı olasılığı ve P(x|y) y olayının olduğu bilindiğinde x olayının şartlı olasılığını göstermektedir.

### 2.4.2 Lojistik Regresyon

Lojistik regresyon bir veya çok ön göstergeli değişkene bağlı kategorik bağımlı değişkenlerin sonuçlarını tahmin etmede kullanılan bir tür regresyon analizidir. Geçmişte, temel lojistik regresyon modelinin genişletilmiş hali olarak farklı türde modeller geliştirilmiştir. Multinominal (çok terimli) lojistik regresyon modeli ikiden fazla ayrı sonuca izin veren lojistik regresyonu genelleştiren bir modeldir. Yani bağımsız değişkenleri verilen , kategorik dağıtılan bağımlı değişkenlerin farklı olası sonuçlarını tahmin etmek için kullanılan modeldir. (Sağbaş & Ballı, 2016)

k bağımsız değişken ve N gözlem olduğunda doğrusal regresyon modelinin genel formu Denklem 3'te belirtilen şekilde yazılır.

$$yi = \beta 0 + \beta 1x1i + \beta 2x2i + \ldots\ldots\ldots + \beta kxki + \varepsilon i \qquad \text{Denklem 3}$$

Bağımlı değişkenin alabileceği değerlerin 0-1 arasında olmasını sağlamak için bağımsız değişken ve bağımlı değişken arasında eğrisel bir ilişkiyi sağlayan modeli kullanmak daha uygundur. $\beta 1$'in işaretine göre S veya ters S şeklinde olan eğrileri sağlayan

$$\mathrm{E(yi)} = \Pi \mathrm{i} = \frac{\exp(\beta 0 + \beta 1\mathrm{x1i} + \beta 2\mathrm{x2i} + \ldots\ldots\ldots + \beta \mathrm{kxki})}{1 + \exp(\beta 0 + \beta 1\mathrm{x1i} + \beta 2\mathrm{x2i} + \ldots\ldots\ldots + \beta \mathrm{kxki})}$$

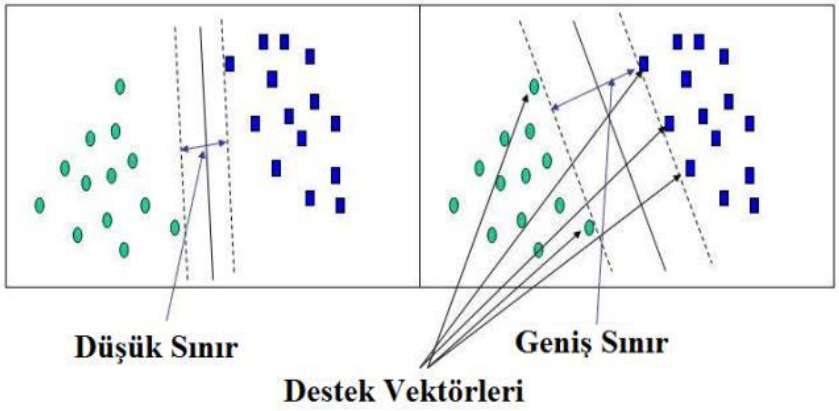
formundaki bu fonksiyona "Lojistik Fonksiyon" adı verilir. Bu lojistik fonksiyonlar genellikle S eklinde fonksiyon olarak isimlendirilir. Bunlar 0 ve 1 asimtotlarına sahiptir ve böylece E(y), 0 ile 1 sınırlar arasında kalır. (Aktaş, 2009)

### 2.4.3 Destek Vektör Makineleri

Destek Vektör Makinaları, Alexey Chervonenkis ve Vladimir Vapnik tarafından geliştirilmiş kullanışlı sınıflandırma algoritmalarından biridir. Destek Vektör Makinaların lineer ve lineer olmayan yaklaşımları mevcuttur. Destek Vektör Makinaları lineer yaklaşımda sınıflar arasına onları en iyi şekilde ayırabilecek şekilde doğrusal sınırlar çizmeye çalışmaktadır. Lineer yaklaşımda tek Cost (C) parametresi mevcuttur.

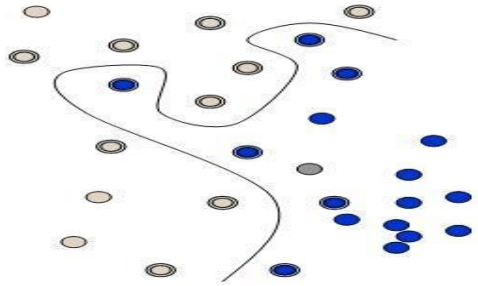
Algoritma lineer olmayan yaklaşımda ise sınıfları en iyi şekilde ayırabilecek eğrisel sınırlar çizmeye çalışmaktadır. Lineer olmayan yaklaşımda algoritma Cost (C) ve Gamma parametrelerine sahiptir.

DVM analizi, olayları ayıran 1-boyutlu düzlemi, hedef kategorilerini temel alarak bulmaya çalışır. Mümkün çizgilerin sınırsız sayısı vardır. Hangi çizginin, daha iyi olduğunu bulmak ve optimal çizgiyi nasıl bulacağımız önemlidir . Noktalı gösterilen çizgiler en yakın vektörler arasında mesafeyi ayıran çizgiye paralel olarak çekilir . Noktalı çizgilerin arasındaki mesafe kenarı çağırır . Kenarın genişliğini zorlayan vektörler, destek vektörleridir. (Karakoyun & Hacıbeyoğlu, 2014)




**Şekil 1.** Düşük Sınır ve Geniş Sınırda Destek Vektörlerin Gösterimi

Sınıflandırılması düşünülen veri Şekil 2'de görüleceği üzere lineer bir sınır ile ayrılmayacak koşullarda olabilmektedir.



**Şekil 2.** Destek Vektör Makineleri Lineer Olmayan Sınıflandırma Gösterimi

## 2.5 Karşılaştırma Ölçütleri

Sınıflandırma başarımını ölçmek için öncelikle hata matrisi kullanılmaktadır . Hata matrisi sınıf sayısı kadar satır ve sütundan oluşan ve her bir hücrede farklı bir değer tutan bir tablodur.

**Tablo 1:** İkili Sınıflandırmaya İlişkin Doğru (1) ve Yanlış (0) Sınıflandırma Sonuç Sayılarını Gösteren Bir Hata Matrisi.

| | | Tahmin Edilen Sınıf | |
|---|---|---|---|
| | | 0 | 1 |
| Gerçek Sınıf | 0 | Doğru Negatif (DN) | Yanlış Negatif (YN) |
| | 1 | Yanlış Pozitif (YP) | Doğru Pozitif (DP) |

Matrisin satırları gerçek sınıflandırma sonuçlarını gösterirken , sütunları tahminleri göstermektedir. Böylece hata matrisinden yararlanarak aşağıdaki ölçütler yöntemleri karşılaştırma amacıyla kullanılmaktadır:

• **Doğruluk ("accuracy"):** Doğru olarak sınıflandırılmış girdilerin toplam girdilere oranını veren ölçüttür. Denklem 4 ile hesaplanmaktadır.

$$Doğruluk \quad Ölçütü = \frac{DP + YP}{DP + YP + DN + YN} \qquad \text{Denklem 4}$$

• **Kesinlik ("precision"):** Doğru olarak sınıflandırılmış pozitif girdilerin toplam pozitif değerlere oranıdır. Denklem 5 ile hesaplanmaktadır.

$$Kesinlik = \frac{DP}{DP + YP} \qquad \text{Denklem 5}$$

• **Geriçağırım ("recall"):** Doğru olarak sınıflandırılmış pozitif girdilerin gerçek pozitif değerlere oranıdır. Denklem 6 ile hesaplanmaktadır.

$$Geriçağırım = \frac{DP}{DP + YN} \qquad \text{Denklem 6}$$

• **F1 ölçütü:** Daha bir karşılaştırma sağlamak için kesinlik ve geriçağırımın harmonik ortalaması şeklinde hesaplan an bir ölçüttür. Denklem 7 ile hesaplanmaktadır. (Kalaycı, 2018)

$$F1 \text{ ölçütü} = \frac{2 * Kesinlik * G Geri çağırım}{Kesinlik + Geri çağırım} \qquad \text{Denklem 7}$$

## 2.6 Metin Madenciliği

Günümüzde metin olarak adlandırılabilecek olan verinin büyük çoğunluğu e – posta, sms, bloglar ve sosyal medya uygulamalarında bulunmaktadır. Bu alanlarda tutulmakta olan metinler genelde düzensiz, gürültülü ve karmaşık yapılarından dolayı yapısal olmayan karakterdedir. Metin madenciliği çalışmalarında bu alanlarda tutulmakta olan metinler bir araya getirilir.

Yapısal olmayan karakteristiklerinden dolayı bu metinler uygun önişleme çalışmaları ile makine öğrenmesi tekniklerinin uygulanabileceği bir veriye dönüştürülür.

### 2.6.1 Metin Madenciliğinin Tanımı

En genel anlamda metin madenciliği, metin içindeki verileri kullanarak bu verilerden anlamlı örüntüler çıkaran yöntemdir. Metin madenciliği çalışmalarında örüntüler genelde makine öğrenmesi algoritmaları ile tespit edilir.

Metin madenciliği çalışmaları ile ilgili araştırma yapıldığında metin analitiği, duygu analizi ve varlık ilişki modellemesi gibi kavramlarla sıklıkla karşılaşılmaktadır. Metin madenciliği çalışmaları bakışları üzerine çeken bir alan olup bu çalışmalar sonucunda elde edilen sonuçlar ve modeller kullanıcılarını memnun etmektedir.

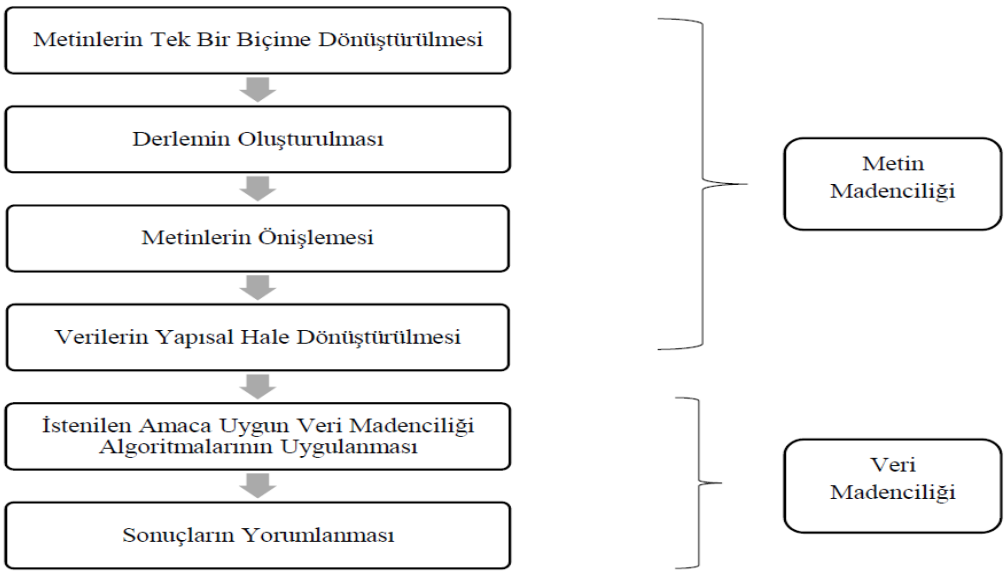
İnsanların ne düşündüğünü, neler hissettiklerini anlamak için geleneksel yöntemler olan anketler, odak görüşmeler ve diğer araştırma yöntemlerinin ötesinde, yayınlanmış yazılı metinleri ele almak ve bu metinleri de tıpkı bir büyük bir veri tabanı gibi analiz etmek ve sayısal değerlere dönüştürmek metin madenciliğinin temel fonksiyonudur. (Yıldız & Ağdeniz, 2018)

Metin madenciliğinde düzensiz karmaşık ve gürültülü yapıdaki veri önişleme çalışmaları ile düzenli bir yapıya kavuşturulur. Bu süreçte verilere makine öğrenmesi algoritmaları uygulanabilir. Önişleme süreçleri sonrası yapılan bu çalışmalar bilinen veri madenciliği çalışmalarıdır.

### 2.6.2 Metin Madenciliği Aşamaları

Metin madenciliği çalışmalarında kullanılan veriler düzensiz ve karmaşık yapılarından dolayı yapısal olmayan verilerdir. Metin madenciliği çalışmaları yapılarak veriler yapısal forma getirilmektedir. Bu süreçte veri karmaşık, düzensiz ve gürültülü yapı giderilmekte ve veriler modellemeye hazır hale getirilmektedir.

Ana hatları ile metin madenciliği çalışmaaları Şekil 4'te verilen aşamalardan oluşmaktadır:

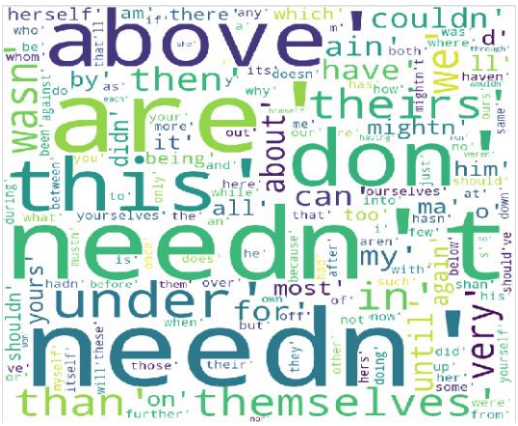



**Şekil 3.** Metin Madenciliği Aşamaları

Şekil 3'te verilen ilk aşamada ham metnin önişleme süreçleri ile yapısal hale dönüştürülmesi görülmektedir. Yapısal dönüşümden sonraki aşamada makine öğrenmesi algoritmaları kullanılarak modelleme süreci tamamlanmaktadır.

### 2.6.3 Durak Kelimeler

Edat, Bağlaç ve zamir gibi çok sık kullanılan ve tek başına anlam ifade etmeyen kelimeler metinden çıkartılır. Örneğin ingilizcedeki "a", "the", "as", "at" gibi kelimeler bu gruba girer. Kelime sayısının azaltılması hem hız hem de performansı iyileştirebilir. (Haltaş, Alkan , & Karabulut, 2015)




**Şekil 4.** Durak Kelimelerin (Stopword) Oluşturduğu Kelime Bulutu

Önişleme safhasında dokümanlardaki boşluk, rakam ve noktalama işareti gibi herhangi bir anlam ifade etmeyen karakterler elenir, büyük harfler küçük harflere dönüştürülerek temizlenmiş doküman haline getirilir. (Yılmaz, 2013)

### 2.6.4 Terim Ağırlığı

Terim ağırlığı , bir belgede terime ait bilgiyi ifade etmektedir. Belgeler içerisinde farklı ağırlıklara sahip birçok terimleri barındırmaktadır. Belgeyi t ile ifade edersek terim ağırlığı denklemi Denklem 8'deki gibi olacaktır.

$$t = (w1, w2, w3, \dots, wT) \qquad \text{Denklem 8}$$

Yukarıdaki formüldeki ağırlık değeri duruma göre iki farklı şekilde ifade edilebilir.

- Bir doküman içindeki ilişkili kelimelerin doküman içindeki görülme sayısı

hesaplanarak.

- Eğer dokümanların çoğunda görülüyorsa, doküman içinde değeri az ise ayırt

edici bir ağırlık değeri belirlenerek.

Terim ağırlığı belirlemede çeşitli yaklaşımlar karşımıza çıkar . Hazırlanan makine öğrenmesi tekniğine uygun olarak aşağıdaki yaklaşımlardan biri veya bir kaçı kullanılabilir:

- **Yanlış/Doğru Ağırlığı**

Temel terim ağırlığı hesaplama yaklaşımı olarak görülmektedir. Terim belgede var ise 1 yok ise 0 değeri terim ağırlığı olarak belirlenir.

$$w_i = \begin{cases} 1, & if\ tf_i > 0 \\ 0, & otherwise \end{cases} \qquad \text{Denklem 9}$$

$tf_i$ : terim frekansı

- **Terim Frekansı × Ters Doküman Frekansı Ağırlığı**

Bu yaklaşım diğer dokümanları da göz önüne alarak bir terim ağırlığı hesaplama yoluna gider . Bu değer hesaplanırken doküman içindeki geçme frekansı terimin görüldüğü tüm dokümanları ters orantısı ile çarpılarak bulunur. (Uzun, 2007)

$$w_i = tf_i \cdot log\frac{N}{N_i} \qquad \text{Denklem 10}$$

n : Tüm Belgelerin Sayısı

$n_i$ : Terimi İçeren Belge Sayısı

$log\frac{n}{n_i}$: Belge Frekansını Tersi

# 3. VERİLERİN TOPLANMASI

## 3.1 Veri Seti Hakkında Bilgiler

Veri seti; müşterilerin 'Amazon' şirketinden almış oldukları tablet bilgisayarlar ile ilgili yorumları ve bu yorumlara istinaden şirket yetkililerinin her bir yorum için vermiş oldukları notları (etiketleri) içermektedir. Veri, 2015 ile 2017 arasındaki yorumları ve notları barındırmaktadır.

Veri seti 33028 satır ve 2 sütundan oluşmaktadır. İlk sütunda müşteri yorumu (metin) yer alırken ikinci sütunda ise ilgili yorumun notu (etiket) yer almaktadır. Veri setinde yorum notu (etiketi) olmayan 27 kayıt vardır. Notları olmayan bu kayıtlar silinince 33001 kayıt kalmıştır. Müşteri yorumunun notu '1' ile '5' arasında değerler almaktadır. '5' notu müşterinin üründen çok memnun kaldığını belirtiyor iken '1' notu ise müşterinin üründen hiç memnun kalmadığını göstermektedir.

| | Metin | Etiket |
|---|---|---|
| 0 | This product so far has not disappointed. My c... | 5.0 |
| 1 | great for beginner or experienced person. Boug... | 5.0 |
| 2 | Inexpensive tablet for him to use and learn on... | 5.0 |
| 3 | I've had my Fire HD 8 two weeks now and I love... | 4.0 |
| 4 | I bought this for my grand daughter when she c... | 5.0 |

**Şekil 5.** Verinin Görünümü

Veri setinde müşterilerin vermiş oldukları notlara bakıldığında '5' etiketli kayıtların sayısının veride çoğunlukta olduğu görülmektedir. '5' etiketli kayıtlardan sonra '4' etiketli kayıtlar çoğunluktadır. Veri setinde en az kaydı olan notlar ise '1' ve '2' notlu kayıtlardır.

```
5.0    22638
4.0     8206
3.0     1426
2.0      370
1.0      362
Name: Etiket, dtype: int64
```

**Şekil 6.** Veride Etiket Alanının Sayı Bazlı Dağılımı

```
5.0    0.685958
4.0    0.248652
3.0    0.043210
2.0    0.011211
1.0    0.010969
Name: Etiket, dtype: float64
```

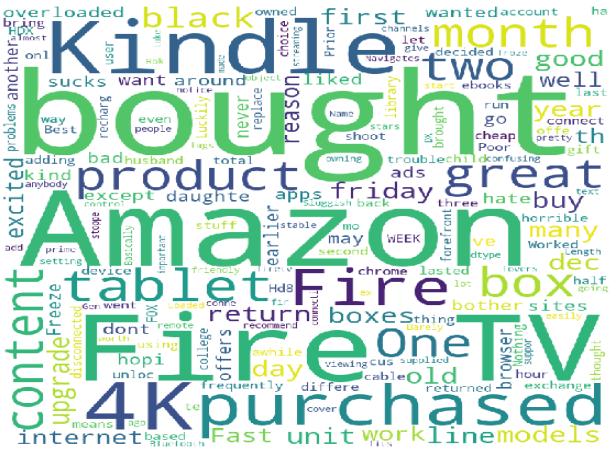
**Şekil 7.** Veride Etiket Alanının Yüzde Bazlı Dağılımı

# 4. VERİLERİN ANALİZİ

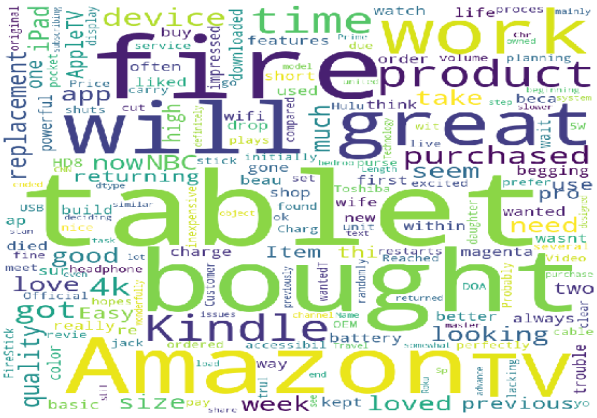
## 4.1 Kelime Bulutları

Kelime bulutu, kelime köklerinin çözümlenmesi vasıtasıyla elde edilen farklı kelime bilgilerini bulut içinde tek temsili terim şeklinde tutarak kullanır. (Heimerl, Lohmann, Lange, & Ertl, 2014)

Bu inceleme sürecinde kelime bulutu oluşturularak notlarla en çok tekrarlanan kelimeler arasındaki ilişki incelenebilir. Kelime bulutu oluşturulurken durak kelimeler bulutlardan çıkarılmıştır. Not (etiket) bazında müşteri yorumlarında en çok tekrarlanan kelimeleri görmek adına kelime bulutlarından faydalanılmıştır. Kelime bulutları ile not bazında en çok tekrarlanan kelimeler (genelde kelime bulutunda en önde ve diğer kelimelere göre daha büyük gösterilen kelimeler) tespit edilebilir. Etiketi '1' ve '2' olan kayıtlarda 'fire' gibi duygu belirten ifadeler kullanılmış iken '3' etiketli kayıtlarda 'Good' ve 'fire' gibi iyi ve kötü adlandırılabilecek duygu belirten ifadeler kullanılmıştır. '4' ve '5' etiketli kayıtlarda 'love' , 'great' ve 'nice' gibi memnuniyet belirten kelimeler daha yoğun olarak kullanılmıştır.

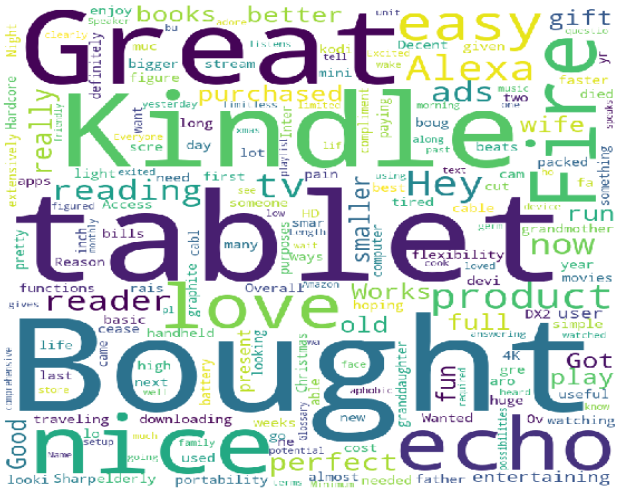



**Şekil 8.** Etiketi '1' Olan Kayıtların Kelime Bulutu




**Şekil 9.** Etiketi ‘2’ Olan Kayıtların Kelime Bulutu


**Şekil 10.** Etiketi ‘3’ Olan Kayıtların Kelime Bulutu




**Şekil 11.** Etiketi ‘4’ Olan Kayıtların Kelime Bulutu


**Şekil 12.** Etiketi ‘5’ Olan Kayıtların Kelime Bulutu

## 4.2 Müşteri Yorumlarının Uzunlukları

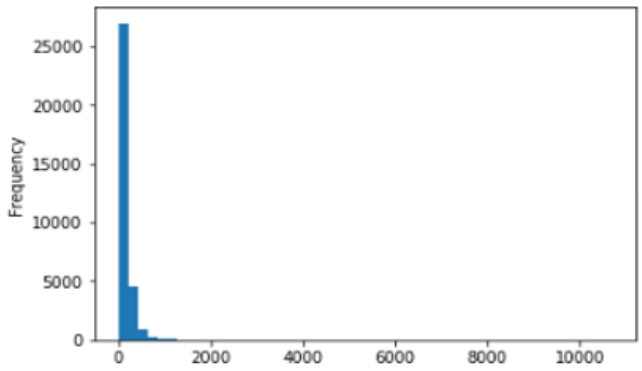
Metin analitiği çalışmaları kapsamında yorumların uzunlukları hesaplanarak etiket bazlı çıkarsamalar yapılabilmektedir. Verideki yorumlar analiz edildiğinde en kısa yorumun 4 karakterden oluştuğu ve en uzun yorumun ise 10670 karakterden oluştuğu tespit edilmiştir. Her bir kayıtta ortalama olarak 155 karakterlik yorum yazmıştır.

| | Etiket | Metin | Uzunluk |
|---|---|---|---|
| 0 | 5.0 | This product so far has not disappointed. My c... | 143 |
| 1 | 5.0 | great for beginner or experienced person. Boug... | 75 |
| 2 | 5.0 | Inexpensive tablet for him to use and learn on... | 131 |
| 3 | 4.0 | I've had my Fire HD 8 two weeks now and I love... | 593 |
| 4 | 5.0 | I bought this for my grand daughter when she c... | 613 |

**Şekil 13.** Müşteri Yorumlarındaki Karakter Uzunluklarının Tespiti

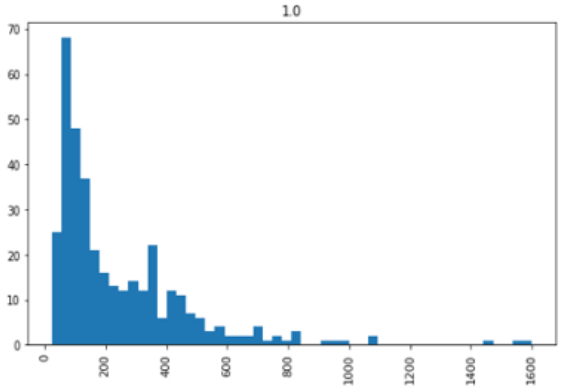
```
count    33001.000000
mean       155.361504
std        176.959427
min          4.000000
25%         69.000000
50%        105.000000
75%        180.000000
max      10670.000000
Name: Uzunluk, dtype: float64
```

**Şekil 14.** Açıklayıcı İstatistik Değerleri

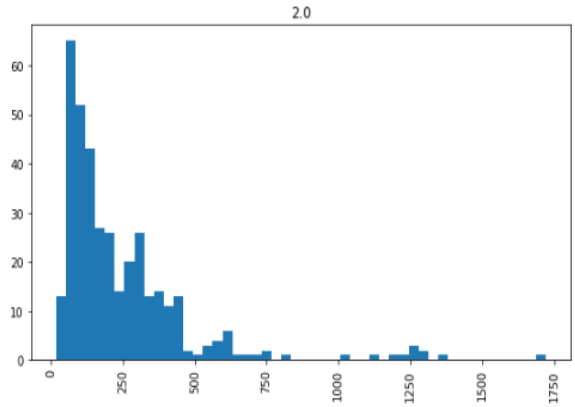



**Şekil 15.** Tüm Kayıtların Karakter Uzunluk Histogramı

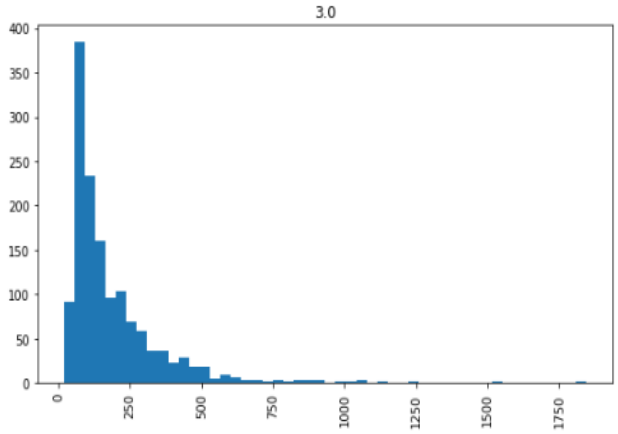
Karakter uzunluk çalışmasını not(etiket) bazında incelediğimizde ise notu ‘1’ ve ‘2’ olan kayıtlardaki yorumların diğer etiketli yorumlara göre daha uzun olduğu görülmüştür. Bu durumdan üründen memnun olmayan müşterilerin uzunca bir şekilde memnuniyetsizliğini belirttiği sonucuna varılabilir. Histogramlarda düşey eksen frekansı gösterirken yatay eksen karakter uzunluğunu göstermektedir.



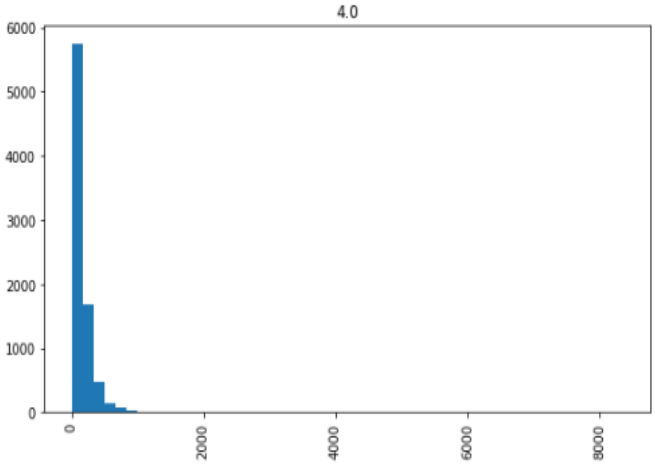

**Şekil 16.** Etiketi ‘1’ Olan Yorumların Karakter Uzunluk Histogramı



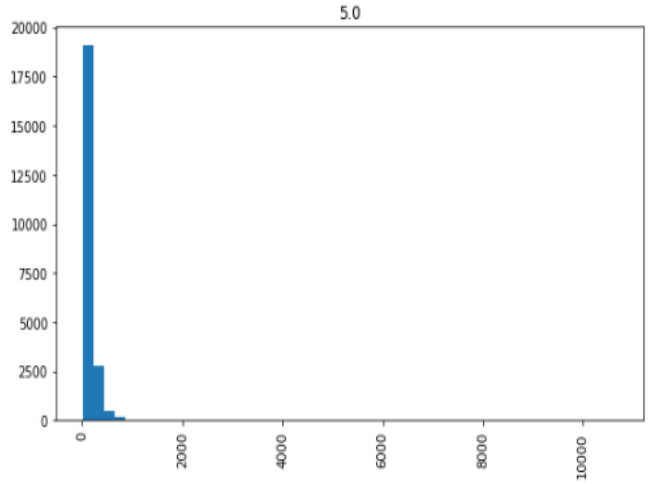

**Şekil 17.** Etiketi ‘2’ Olan Yorumların Karakter Uzunluk Histogramı




**Şekil 18.** Etiketi ‘3’ Olan Yorumların Karakter Uzunluk Histogramı




**Şekil 19.** Etiketi ‘4’ Olan Yorumların Karakter Uzunluk Histogramı




**Şekil 20.** Etiketi ‘5’ Olan Yorumların Karakter Uzunluk Histogramı

## 4.3 Chi2 Bağımsızlık Analizi

Ki-kare en popüler özellik seçme yaklaşımlarından biridir. İstatistikte ki-kare testi iki olayın bağımsızlığını incelenmek için kullanılmaktadır. (Zainuddin & Selamat, 2014)

Metin analitiği çalışmaları kapsamında bir diğer analiz yöntemi ise ki-kare analiz yöntemidir. Bu analiz yönteminde ilk olarak müşteri yorumları, tek kelime ve iki kelime olacak şekilde parçalara ayrılır. Sonraki aşamada ise kelime (tek kelime) ya da kelime gruplarının (iki kelime) notlarla ilişkisine ki-kare analizi ile test edilmektedir.

Analiz sonucunda istatistiki olarak anlamlı görülen kelime (ilk beş) ve kelime grupları (ilk beş) elde edilmiştir. İstatistiki anlamlılığın akabinde çıkan sonuçların not bazında yorumlanması gerekmektedir.

Yapılan analizde ‘1’ ve ‘2’ etiketli kayıtlarda ‘return’ ,’waste’ ve ’worst’ gibi olumsuz nitelikteki kelimelerin olduğu görülmüştür. ‘1’ ve ‘2’ etiketli kayıtlarda bu tarzdaki olumsuz kelimelerin çoğunlukta olduğu tespit edilmiştir.

‘3’ etiketli kayıtlarda ise olumlu ya da olumsuz nitelendirilebilecek kelimelere pek rastlanmadığı görülmüştür. ‘3’ etiketli kayıtlardaki yorumların ‘nötr’ diye adlandırılabilecek nitelikte olduğu söylenebilir.

Son aşamada ise ‘4’ ve ‘5’ etiketli kayıtlarda ‘wish’ ,’love’ ve ’good’ gibi olumlu nitelikteki kelimelerin olduğu görülmüştür. ‘4’ ve ‘5’ etiketli yorumlarda ürünlerin genel itibari ile iyi olduğu dile getirilmiştir.

```
# '1.0':
  . En İlişkili Kelime (Tek Kelime) :
        . return
        . worst
        . waste
        . returning
        . returned
  . En İlişkili Kelime Grubu (İki Kelime):
        . did work
        . turn charge
        . wouldn work
        . wouldn connect
        . waste money
```

**Şekil 21.** Ki Kare Analizi (Etiketi ‘1’ Olan Yorumlar)

```
# '2.0':
  . En İlişkili Kelime (Tek Kelime) :
        . dead
        . closes
        . return
        . slow
        . returned
  . En İlişkili Kelime Grubu (İki Kelime):
        . just ok
        . time trying
        . returned purchased
        . lots ads
        . returned day
```

**Şekil 22.** Ki Kare Analizi (Etiketi ‘2’ Olan Yorumlar)

```
# '3.0':
  . En İlişkili Kelime (Tek Kelime) :
        . decent
        . limited
        . okay
        . slow
        . ok
  . En İlişkili Kelime Grubu (İki Kelime):
        . pretty slow
        . ok price
        . tablet slow
        . ok tablet
        . ok product
```

**Şekil 23.**Ki Kare Analizi (Etiketi ‘3’ Olan Yorumlar)

```
# '4.0':
  . En İlişkili Kelime (Tek Kelime) :
        . wish
        . love
        . overall
        . stars
        . good
  . En İlişkili Kelime Grubu (İki Kelime):
        . good little
        . pretty good
        . little slow
        . overall good
        . good tablet
```

**Şekil 24.**Ki Kare Analizi (Etiketi ‘4’ Olan Yorumlar)

```
# '5.0':
  . En İlişkili Kelime (Tek Kelime) :
       . google
       . ok
       . good
       . love
       . slow
  . En İlişkili Kelime Grubu (İki Kelime):
       . don like
       . play store
       . little slow
       . google play
       . good tablet
```

**Şekil 25.**Ki Kare Analizi (Etiketi ‘5’ Olan Yorumlar)

# 5. SINIFLANDIRMA

## 5.1 Sınıflandırma Öncesi Ön İşlemler

Analiz çalışmaları, verinin yapısının anlaşılması ve sınıflandırma çalışmalarına yardımcı olması adına yapılmıştır. Yapılan analiz çalışmasında etiketi '1' ve '2' olan yorumların 'Kötü Yorum' olarak adlandırılabileceği görülmüştür.Etiketi '4' ve '5' olan yorumların ise yapı itibari ile 'İyi Yorum' olarak görülebileceği sonucuna varılmıştır. '3' etiketli yorumları 'İyi Yorum' ve 'Kötü Yorum' olarak atanabileceği bir durumun olmadığı kanaatine varılmıştır. Çalışma iyi ve kötü yorumların sınıflandırılması adına yağunlaşılacağından etiketi '3' olan kayıtlar verinin içerisinden çıkarılır. Etiketi '3' olan kayıtları veriden çıkardığımızda 33001 olan kayıt sayısı 31575 kayıt sayısına düşmektedir.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 31575 entries, 0 to 33027
Data columns (total 2 columns):
Etiket    31575 non-null float64
Metin     31575 non-null object
dtypes: float64(1), object(1)
memory usage: 740.0+ KB
```

**Şekil 26.**Nötr Nitelikli Kayıtlar Çıkarıldıktan Sonra Veri İle İlgili Özet Bilgi

Veride '3' etiketli olan kayıtlar çıkarıldıktan sonra '1' ve '2' etiketli kayıtlara 'KOTU' olarak nitelik ataması yapılırken '4' ve '5' etiketli kayıtlara ise 'IYI' nitelik ataması yapılır.

| | Etiket | Metin | Yorum_Nitelik |
|---|---|---|---|
| 0 | 5.0 | This product so far has not disappointed. My c... | IYI |
| 1 | 5.0 | great for beginner or experienced person. Boug... | IYI |
| 2 | 5.0 | Inexpensive tablet for him to use and learn on... | IYI |
| 3 | 4.0 | I've had my Fire HD 8 two weeks now and I love... | IYI |
| 4 | 5.0 | I bought this for my grand daughter when she c... | IYI |

**Şekil 27.**Etiketteki Değerlere Göre Nitelik (Duygu) Atamasının Yapılması

Nitelik atamasından sonraki aşamada sınıflandırma için kullanılacak olan metin üzerinde düzenleme çalışması yapılmıştır. Bu çalışma kapsamında ilk olarak metin durak kelimelerden arındırılmıştır. Arındırma işleminden sonra metin karakterlerden arındırılarak daha düzenli ve kullanılabilir bir yapıya dönüştürülmüştür.

| | Etiket | Metin | Duzenlenmis_Metin |
|---|---|---|---|
| 31570 | 5.0 | I purchased this as a gift after using this at... | i purchased this a a gift after using this at ... |
| 31571 | 5.0 | Love my new Echo. Ask it questions...get the a... | love my new echo ask it question get the answe... |
| 31572 | 5.0 | This is a good buy with the recent drop in pri... | this is a good buy with the recent drop in pri... |
| 31573 | 5.0 | Love it! Alexa is Amazing! It is user friendly... | love it alexa is amazing it is user friendly w... |
| 31574 | 5.0 | My wife wanted this for her birthday. She love... | my wife wanted this for her birthday she love ... |

**Şekil 28.**Sınıflandırma Öncesi Metnin Düzenlenmesi

## 5.2 İşaretleme Ve Terim Ağırlıklandırma

Çalışma kapsamında düzenlenmiş olan metindeki her bir yorum cümlenin içinde barındırdığı kelimelerin köklerine kadar ayrışımı (işaretleme) gerçekleştirilmiştir. Bu ayrıştırma işleminde metne Count ( N-Gram) ve TFIDF (N-Gram ) yaklaşımı ayrı ayrı uygulanmıştır. Ayrıştırma işleminden sonra elde edilen kelimeler modellemede kullanılarak ne türde bir ayrıştırma işleminin model başarımını artırdığı da tespit edilecektir.

Bu ayrıştırma (işaretleme) işleminde python içerisinde bulunan sklearn kütüphanesine bağlı CountVectorizer ve TfidfVectorizer prosedürleri kullanılmıştır. Ayrıştırma işlemlerinde metin kelimelere ayrıştırılır ve ayrıştırılan kelimeler sütun başlıkları olacak şekilde konumlandırılır. Count yaklaşımlı ayrıştırma işleminde kelimelerin ilgili satırda varlığı ya da yokluğu 1 ve 0 ile ifade edilir.

TF-IDF, bir kelimenin kitle veya yığındaki önemini yansıtır. TF-IDF değeri, belgedeki ilgili kelimenin sıklığı arttıkça artar. (Tripathy, Agrawal, & Rath, 2016) Count yaklaşımlı ayrıştırma işlemi uygulandıktan sonra müşteri yorumları kelimelere ayrıştırılmıştır. Bu işlem sürecinde elde edilen kelimeler özellik olarak modelleme aşamasında kullanılacaktır. Count yaklaşımın kullanıldığı bu aşamada 25247 adet özellik oluşmuştur.

Aynı şekilde ham veri üzerine count yaklaşımlı ayrıştırma işleminin yanı sıra TF-IDF ayrıştırma işlemi de uygulanmıştır. TF-IDF yaklaşımın kullanıldığı bu aşamada 25247 adet özellik oluşmuştur. Bilgilendirici olmayan kelimeler başlıklar, edatlar, bağlaçlar ve yüksek frekanslı kelimeleri (birçok fiil, zarf ve zamir) içeren yaklaşık 400 kelimeyi içeren bir 'durak kelime listesi' olarak sıklıkla tanımlanmaktadır. (Méndez, Iglesias, Fdez-Riverola, Díaz, & Corchado, 2005)

Bu ayrıştırma işlemleri uygulanırken edat ve bağlaç gibi çok sık kullanılan ve tek başına anlam ifade etmeyen (durak kelimeler) kelimeler metinden çıkarılmıştır.

Metin, python içerisinde bulunan nltk kütüphanesine bağlı 'stopwords' prosedürü sayesiyle 'a' , 'the' ,'as' gibi durak kelimelerden arındırılmıştır. Bu durak kelimelerin azaltılması hem işlem hızında hem de model performansında iyileştirecektir.

## 5.3 Öğrenme Ve Test Etme Verisinin Oluşturulması

Modelleme aşamasında veri setinin %75'lik kısmı (23681 kayıt) öğrenme (training) için %25'lik kısmı (7894 kayıt) ise test etme (testing) için kullanılmıştır. Ayrıştırma işleminde python içerisinde bulunan sklearn kütüphanesine bağlı 'train_test_split' prosedürü kullanılmıştır.

Prosedürde 'stratify' (tabakalı) ibaresi girilerek öğrenme ve test etme verisinde etiket bazında ( Kötü Yorum, İyi Yorum) dengeli bir dağılımın oluşması sağlanmaya çalışılmıştır.

**Tablo 2:** Tüm Verinin Öğrenme Verisi ve Test Etme Verisi Olarak İkiye Ayrılması

| | |
|---|---|
| Öğrenme İçin Kullanılan Kayıt Sayısı | 23681 Kayıt |
| Test Etme İçin Kullanılan Kayıt Sayısı | 7894 Kayıt |
| Toplam Kayıt Sayısı ( Tüm Veri) | 31575 Kayıt |

**Tablo 3:** Test Etme Verisinde Etiket (Label) Dağılımı

| | |
|---|---|
| Test Verisinde IYI Etiketli Kayıt Sayısı | 7711 Kayıt |
| Test Verisinde KOTU Etiketli Kayıt Sayısı | 183 Kayıt |
| Test Verisindeki Toplam Kayıt Sayısı | 7894 Kayıt |

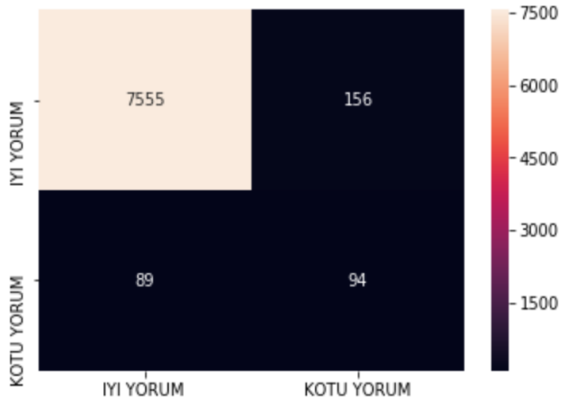
# 6. BULGULAR

Modellemede sınıflandırma algoritması öğrenme verisi üzerinde eğitilmekte ve öğrendikleri örüntüler yardımıyla bir kalıp (pattern) oluşturmaktadır. Sonraki aşamada ise algoritma bu kalıp ile test verisi üzerinde tahminler yapmaktadır.

Sınıflandırma algoritmaları çalıştırılmadan önce algoritmalara class_weight='balanced' ibaresi yazılmıştır. Bu ifadenin yazılması ile sayıca fazla olan iyi yorumların, azınlık olarak adlandırılabilecek kötü yorumları baskılamasının önüne geçilmiş olacaktır.

## 6.1 Count Yaklaşımlı Naive Bayes Modeli

Naive Bayes ( Count Yaklaşımlı ) algoritması ile test verisi üzerinde tahminler yapılmıştır. Test verisinde gerçekteki etiket ile tahmin edilen etiket kıyaslanarak model başarımı incelenmiştir. Model test verisinde 'IYI YORUM' etiketli 7711 kayıtın 7555 adedini doğru tahmin etmiştir. Ayrıca model test verisinde 'KOTU YORUM' etiketli 183 kaydın 94 adedini doğru tahmin etmiştir.




**Şekil 29.**Naive Bayes ( Count Yaklaşımlı ) Modelin Karmaşıklık Matrisi

Modelde doğruluk oranı (accuracy ratio) %96,89 olarak hesaplanmıştır. Veri dengeli değilse ve oluşacak hataların maliyeti önemli derecede ise doğruluk oranı (accuracy ratio) uygun bir yaklaşım olmayacaktır. (Chawla, Data Mining for Imbalanced Datasets: An Overview, 2014)

Dengesiz veri setlerinden öğrenme için ana hedef, precision oranını düşürmeden recall oranını geliştirmektir. Bununla birlikte recall ve precision sıklıkla birbirlerine

aykırı düşen iki metriktir. Azınlık sınıf için ‘DP’ arttığı zaman ‘YP’ ayrıca artacaktır. Bu durum recall metriğini artırırken precision metriğini düşürecektir. F1 ölçütü, precision ve recall metriklerinin birleşimidir ve bu metrik azınlık sınıflarda sınıflandırıcının iyi olup olmadığını göstermektedir. (Chawla, Data Mining for Imbalanced Datasets: An Overview, 2014) Modelde ‘IYI YORUM’ etiketli kayıtlarda F1 değeri 0.98 iken ‘KOTU YORUM’ etiketli kayıtlarda F1 değeri 0.43 hesaplanmıştır. Modelin ağırlıklandırılmış F1 değeri ise 0,97 olarak bulunmuştur.

Count Yaklaşımlı (NB) Model Başarımı :

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| IYI | 0.99 | 0.98 | 0.98 | 7711 |
| KOTU | 0.38 | 0.51 | 0.43 | 183 |
| micro avg | 0.97 | 0.97 | 0.97 | 7894 |
| macro avg | 0.68 | 0.75 | 0.71 | 7894 |
| weighted avg | 0.97 | 0.97 | 0.97 | 7894 |

**Şekil 30.**Naive Bayes ( Count Yaklaşımlı ) Modelin Başarım Metrikleri

## 6.2 TFIDF Yaklaşımlı Naive Bayes Modeli

Naive Bayes ( TFIDF Yaklaşımlı ) algoritması ile test veri seti üzerinde yapılan model tahminleri Şekil 32’de görülmektedir. Modelde ‘IYI YORUM’ etiketli 7711 kayıtın 7370 adetini doğru tahmin etmiştir. Algoritma ‘KOTU YORUM’ etiketli 183 kaydın 126 adetini doğru tahmin etmiştir.


**Şekil 31.**Naive Bayes ( TFIDF Yaklaşımlı ) Modelin Karmaşıklık Matrisi

Modelde doğruluk oranı %94,95 olarak hesaplanmıştır. Modelde 'IYI YORUM' etiketli kayıtlarda F1 değeri 0.97 iken 'KOTU YORUM' etiketli kayıtlarda F1 değeri 0.39 hesaplanmıştır. Modelin ağırlıklandırılmış F1 değeri ise 0,96 olarak bulunmuştur. 'Count yaklaşımlı Naive Bayes' modeli , 'TFIDF yaklaşımlı Naive Bayes' modeline göre kötü yorumları tespit etmede daha iyi başarım göstermiştir.

TFIDF Yaklaşımlı (NB) Model Başarımı :

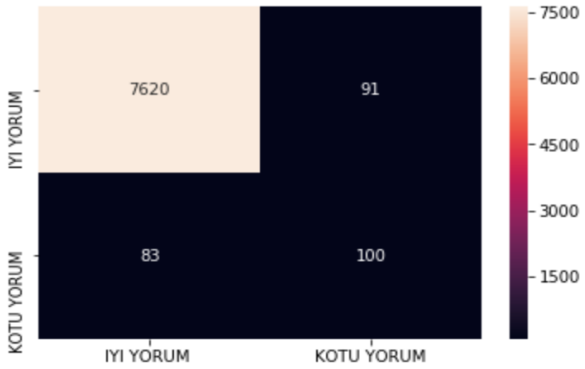
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| IYI | 0.99 | 0.96 | 0.97 | 7711 |
| KOTU | 0.27 | 0.69 | 0.39 | 183 |
| micro avg | 0.95 | 0.95 | 0.95 | 7894 |
| macro avg | 0.63 | 0.82 | 0.68 | 7894 |
| weighted avg | 0.98 | 0.95 | 0.96 | 7894 |

**Şekil 32.**Naive Bayes (TFIDF Yaklaşımlı ) Modelin Başarım Metrikleri

## 6.3 Count Yaklaşımlı Lojistik Regresyon Modeli

Lojistik regresyonda 'saga', 'lbfgs', 'liblinear', 'sag', ve 'newton-cg' gibi farklı çözücüler mevcuttur. (Gupta , Mittal , & Padha, 2017)

Çalışma içerisinde yukarıda belirtilen her bir çözücü için modeller çalıştırılmış ve ortaya çıkan sonuçlar değerlendirilmiştir. Model başarımı açısından karşılaştırmalı iyi sonuç verdiği için 'liblinear' çözücü çeşiti parametre tahmininde kullanılmıştır. Algoritma 'IYI YORUM' etiketli 7711 kaydın 7620 adedini doğru tahmin etmiştir. Algoritma 'KOTU YORUM' etiketli 183 kaydın 100 adedini doğru tahmin etmiştir.




**Şekil 33.**Lojistik Regresyon ( Count Yaklaşımlı ) Modelin Karmaşıklık Matrisi

Modelde doğruluk oranı %97,78 olarak çıkmıştır. Modelde 'IYI YORUM' etiketli kayıtlarda F1 değeri 0.99 iken 'KOTU YORUM' etiketli kayıtlarda F1 değeri 0.53 hesaplanmıştır. Modelin ağırlıklandırılmış F1 değeri ise 0,98 olarak bulunmuştur.

Count Yaklaşımlı (LR) Model Başarımı :

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| IYI | 0.99 | 0.99 | 0.99 | 7711 |
| KOTU | 0.52 | 0.55 | 0.53 | 183 |
| micro avg | 0.98 | 0.98 | 0.98 | 7894 |
| macro avg | 0.76 | 0.77 | 0.76 | 7894 |
| weighted avg | 0.98 | 0.98 | 0.98 | 7894 |

**Şekil 34.**Lojistik Regresyon ( Count Yaklaşımlı ) Modelin Başarım Metrikleri

'Count yaklaşımlı Lojistik Regresyon' çalışmasında model parametreleri incelenmiştir. Toplam skor değeri belli bir eşik değerini aşıyorsa müşteri yorumu 'Kötü Yorum' olarak tahmin edilirken, eşik değerinin altı ise yorum 'İyi Yorum' olarak tahmin edilir. Modelde, değişken katsayısı büyük olan kelimeler olumsuz (returned, slow vb. ) ibareler bulunduruyorken değişken katsayısı küçük olan kelimeler olumlu ibareler (love, excellent vb.) bulundurmaktadır.

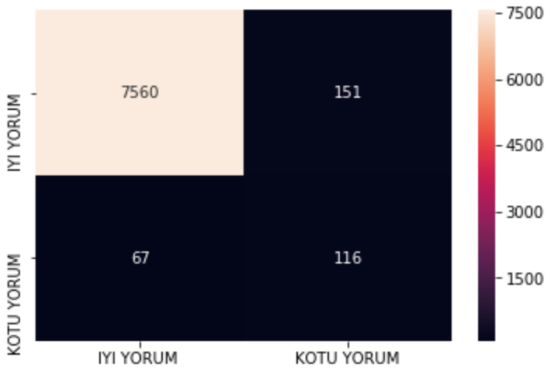
```
Model Parametreleri(Count Yaklaşımlı)


Katsayısı Küçük Olan Kelimeler :
['love' 'excellent' 'great' 'reading' 'overall']

Katsayısı Büyük Olan Kelimeler :
['returned' 'slow' 'picture great' 'price awesome' 'what expected']
```

**Şekil 35.**Lojistik Regresyon ( Count Yaklaşımlı ) Modelinde Önemli Kelimeler

## 6.4 TFIDF Yaklaşımlı Lojistik Regresyon Modeli

Model başarımı açısından karşılaştırmalı iyi sonuç verdiği için 'liblinear' çözücü çeşiti parametre tahmininde kullanılmıştır.

Algoritma 'IYI YORUM' etiketli 7711 kaydın 7560 adedini doğru tahmin etmiştir. Algoritma 'KOTU YORUM' etiketli 183 kaydın 116 adedini doğru tahmin etmiştir.




**Şekil 36.**Lojistik Regresyon ( TFIDF Yaklaşımlı ) Modelin Karmaşıklık Matrisi

Modelde doğruluk oranı %97,36 olarak çıkmıştır. Modelde 'IYI YORUM' etiketli kayıtlarda F1 değeri 0.99 iken 'KOTU YORUM' etiketli kayıtlarda F1 değeri 0.53 hesaplanmıştır. Modelin ağırlıklandırılmış F1 değeri ise 0,98 olarak bulunmuştur. "TFIDF yaklaşımlı LR" modeli ve "Count yaklaşımlı LR" modeli kötü yorumları tespit etmede model başarımları birbirine çok yakın çıkmıştır.

```
TFIDF Yaklaşımlı (LR) Model Başarımı :

              precision    recall  f1-score   support

         IYI       0.99      0.98      0.99      7711
        KOTU       0.45      0.64      0.53       183

   micro avg       0.97      0.97      0.97      7894
   macro avg       0.72      0.81      0.76      7894
weighted avg       0.98      0.97      0.98      7894
```

**Şekil 37.**Lojistik Regresyon ( TFIDF Yaklaşımlı ) Modelin Başarım Metrikleri

'TFIDF yaklaşımlı Lojistik Regresyon' çalışmasında model parametreleri de incelenmiştir. Toplam skor değeri belli bir eşik değerini aşıyorsa müşteri yorumu 'Kötü Yorum' olarak tahmin edilirken, eşik değerinin altı ise yorum 'İyi Yorum' olarak tahmin edilir. Modelde, değişken katsayısı büyük olan kelimeler olumsuz (not, slow vb. ) ibareler bulunduruyorken değişken katsayısı küçük olan kelimeler olumlu ibareler (great, love vb.) bulundurmaktadır.

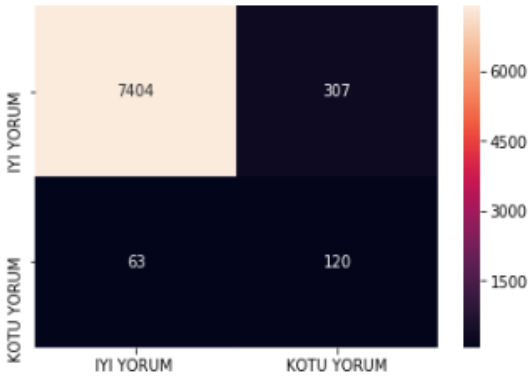
```
Model Parametreleri(TF - IDF Yaklaşımlı)


Katsayısı Küçük Olan Kelimeler :
['great' 'love' 'easy' 'easy to' 'reading']

Katsayısı Büyük Olan Kelimeler :
['not' 'slow' 'returned' 'didn' 'wouldn']
```

**Şekil 38.**Lojistik Regresyon ( TFIDF Yaklaşımlı ) Modelinde Önemli Kelimeler

## 6.5 Count Yaklaşımlı SVM RBF Modeli

SVM RBF ( Count Yaklaşımlı ) algoritması ile test veri seti üzerinde yapılan model tahminleri Şekil 40'da görülmektedir. Python (Jupiter) içerisinde yazılan kod bloğu sayesiyle optimal C (Cost) ve gamma katsayıları tespit edilmiştir. SVM modellerinde C ve gamma katsayılarını tespit edebilmek adına "Grid Search" algoritması önerilmektedir. (Yang, ve diğerleri, 2019) Bu model çalışmasında C (Cost) katsayısı '10' olarak alınırken gamma katsayısı '1e-3' alınmıştır.

Modelde ‘IYI YORUM’ etiketli 7701 kayıtın 7404 adedini doğru tahmin etmiştir. Algoritma ‘KOTU YORUM’ etiketli 183 kaydın 120 adedini doğru tahmin etmiştir.




**Şekil 39.**SVM RBF ( Count Yaklaşımlı ) Modelin Karmaşıklık Matrisi

Modelde doğruluk oranı %95,31 olarak çıkmıştır. Modelde ‘IYI YORUM’ etiketli kayıtlarda F1 değeri 0.98 iken ‘KOTU YORUM’ etiketli kayıtlarda F1 değeri 0.39 hesaplanmıştır. Modelin ağırlıklandırılmış F1 değeri ise 0,96 olarak bulunmuştur.

Count Yaklaşımlı (SVM) Model Başarımı :

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| IYI | 0.99 | 0.96 | 0.98 | 7711 |
| KOTU | 0.28 | 0.66 | 0.39 | 183 |
| micro avg | 0.95 | 0.95 | 0.95 | 7894 |
| macro avg | 0.64 | 0.81 | 0.68 | 7894 |
| weighted avg | 0.98 | 0.95 | 0.96 | 7894 |

**Şekil 40.**SVM RBF ( Count Yaklaşımlı ) Modelin Başarım Metrikleri

## 6.6 TFIDF Yaklaşımlı SVM RBF Modeli

Count yaklaşımlı SVM RBF modelinde olduğu gibi bu çalışmada da optimal C (Cost) ve gamma katsayıları tespit edilmiştir. Bu model çalışmasında C (Cost) katsayısı '10' olarak alınırken gamma katsayısı '1e-3' alınmıştır. Modelde 'IYI YORUM' etiketli 7701 kayıtın 7204 adedini doğru tahmin etmiştir. Model 'KOTU YORUM' etiketli 183 kaydın 137 adedini doğru tahmin etmiştir.


**Şekil 41.**SVM RBF ( TFIDF Yaklaşımlı ) Modelin Karmaşıklık Matrisi

TFIDF Yaklaşımlı (SVM) Model Başarımı :

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| IYI | 0.99 | 0.93 | 0.96 | 7711 |
| KOTU | 0.21 | 0.75 | 0.33 | 183 |
| micro avg | 0.93 | 0.93 | 0.93 | 7894 |
| macro avg | 0.60 | 0.84 | 0.65 | 7894 |
| weighted avg | 0.98 | 0.93 | 0.95 | 7894 |

**Şekil 42.**SVM RBF (TFIDF Yaklaşımlı ) Modelin Başarım Metrikleri

Modelde doğruluk oranı %92,99 olarak çıkmıştır. Modelde 'IYI YORUM' etiketli kayıtlarda F1 değeri 0.96 iken 'KOTU YORUM' etiketli kayıtlarda F1 değeri 0.33 olarak hesaplanmıştır. Modelin ağırlıklandırılmış F1 değeri ise 0,95 olarak bulunmuştur. 'Count yaklaşımlı SVM RBF' modeli , 'TFIDF yaklaşımlı SVM RBF' modeline göre kötü yorumları tespit etmede daha iyi başarım göstermiştir.

# 7. SONUÇLAR

Dengeli veri üzerinde geliştirilen sınıflandırma modellerinde, performans kriteri olarak genellikle accuracy ratio ( doğruluk oranı ) kriterine bakılmaktadır. Ancak dengesiz bir yapıya sahip veride accuracy ratio ( doğruluk oranı ) kriterini referans almak yanlışa düşürebilmektedir.

Dengesiz bir yapıya sahip veri ile oluşturulan sınıflandırma modellerinde model performans kriterlerini belirlemek oldukça önem arz etmektedir. Bu durumdan kaynaklı olarak dengesiz veride accuracy ratio ( doğruluk oranı ) yerine recall, precison ya da F1 kriter değerlerine bakılması daha uygun olacaktır. Çalışma kapsamında recall ve precision metriklerinin harmonik ortalaması olan F1 kriter değerine bakılmıştır. F1 kriter değeri ne kadar yüksek ise model başarımı o kadar iyidir. (Rocca, 2019)

Çoğunluk sınıf olan 'IYI' etiketi kayıtlarda model başarımları yüksek düzeylerdedir. Model başarımlarını yorumlarken azınlık sınıf ile ilgili metrikleri göz önünde bulundurmak gerekmektedir. Modellerdeki azınlık sınıf olan 'KOTU' etiketli kayıtlardaki Precision ve Recall değerleri aşağıdaki tabloda sunulmuştur. Tabloda ayrıca her bir modelin ağırlıklandırılmış ortalama F1 değerleri yer almaktadır.

**Tablo 4:** Modellerin Recall ve Ortalama F1 Değerlerinin Karşılaştırılması

| Geliştirilen Modeller | Metrikler | | |
|---|---|---|---|
| | KOTU Etiketlilerde Precision Değerleri | KOTU Etiketlilerde Recall Değerleri | Ağırlıklandırılmış Ortalama F1 Değerleri |
| Count Yaklaşımlı NB | 0.38 | 0,51 | 0,97 |
| TFIDF Yaklaşımlı NB | 0,27 | 0,69 | 0,96 |
| Count Yaklaşımlı LR | 0,52 | 0,55 | 0,98 |
| TFIDF Yaklaşımlı LR | 0,45 | 0,64 | 0,98 |
| Count Yaklaşımlı SVM | 0,28 | 0,66 | 0,96 |
| TFIDF Yaklaşımlı SVM | 0,21 | 0,75 | 0,95 |

‘KOTU’ etiketli kayıtları diğer modellere göre daha iyi tahmin eden model ‘TFIDF Yaklaşımlı Lojistik Regresyon’ modelidir. Bu modelde recall değeri yükselirken precision değerinde düşüş gerçekleşmemiştir. Bu durum modelin F1 metriğini etkileyerek yükselmesini sağlamıştır.

# 8. ÖNERİLER

Analiz ve sınıflandırma çalışmalarında dengesiz verinin büyük oranda sonuçları etkilediği görülmüştür. Bu olumsuz etkiyi azaltma adına ilk olarak öğrenme ve test etmede tabakalı ( stratify=data['Etiket'] ) örneklem seçme yöntemi kullanılmıştır. Ayrıca sınıflandırma çalışmalarında kullanılan algoritmalara, sınıfların ( etiketlerin ) dengeli dağılması adına kod bloğu ( class_weight='balanced' ) eklenmiştir.

Dengesiz verinin etkileri üzerinde sıklıkla konuşulmakta olup ortaya çıkan sorunların çözümü adına çeşitli fikirler öne sürülmektedir. Dengesiz veri sorunu ile başa çıkılması adına python içerisinde 'imblearn' adı altında belli yaklaşımların olduğu bir kütüphane geliştirilmiştir. (Lemaitre, Nogueira, Oliveira, & Aridas , 2016)

Python içerisinde bulunan bu kütüphane yardımıyla dengesiz olan veriye, klasik makine öğrenmesi yaklaşımından farklı şekilde yaklaşılmaktadır. Öncelikle verinin sınıflandırma öncesinde dengeli bir veri haline gelmesi adına undersampling (örneklem azaltma) veya oversampling (örneklem azaltma) gibi yaklaşımlar uygulanmaktadır.

Undersampling yönteminde sınıf dağılımını dengelemek için çoğunluğu oluşturan örneklemin sayısının azaltılmasını amaçlanmaktadır. Veri setinden gözlemleri kaldırdığı için yararlı bilgiler silinebilmektedir.

Oversampling yönteminde ise azınlığı oluşturan kayıtların sayıları artırılır. Oversampling yönteminin avantajı azınlık ve çoğunluk sınıflarının kayıtları tutulduğu için hiçbir bilginin kaybolmamasıdır. Öte yandan sayıca artırılan azınlık sınıflarda aşırı öğrenme problemi oluşmaktadır. (Minh, 2018)

# KAYNAKÇA


Akba, F. (2014). Duygu Analizinde Öznitelik Seçme Metriklerinin Değerlendirilmesi : Türkçe Film Eleştirileri. Yayımlanmamış Yüksek Lisans Tezi, Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü.

Aktaş, C. (2009). Lojistik Regresyon Analizi : Öğrencilerin Sigara İçme Alışkanlıkları Üzerine Bir Uygulama. *Sosyal Bilimler Enstitüsü Dergisi*, s. 111 - 112.

Alpaydın, E. (2004). *Introduction to Machine Learning* . ABD: The MIT Press .

Atalay, M., & Çelik, E. (2013). Büyük Veri Analizinde Yapay Zekâ Ve Makine Öğrenmesi Uygulamaları. *Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, s. 165 - 166.

Aydın, C. (2018). Makine Öğrenmesi Algoritmaları Kullanılarak İtfaiye İstasyonu İhtiyacının Sınıflandırılması. *Avrupa Bilim ve Teknoloji Dergisi*, s. 172.

Chawla, N. V. (2014). Data Mining for Imbalanced Datasets: An Overview. O. Maimon, & L. Rokach içinde, *Data Mining and Knowledge Discovery Handbook* (s. 877). Springer New York Dordrecht Heidelberg.

Chawla, N. V. (2014). Data Mining for Imbalanced Datasets: An Overview. O. Maimon, & L. Rokach içinde, *Data Mining and Knowledge Discovery Handbook* (s. 878). Springer New York Dordrecht Heidelberg.

Çalış, K., Yıldız, O., & Gazdağı, O. (2013). Reklam İçerikli Epostaların Metin Madenciliği Yöntemleri ile Otomatik Tespiti. *Bilişim Teknolojileri Dergisi*, s.3.

Çürük, E. (2018). Sosyal Ağlardaki Siber Zorbalığın Yapay Zeka Algoritmaları İle Tespiti Ve Sınıflandırılması. Yayımlanmamış Yüksek Lisans Tezi , Mersin Üniversitesi, Fen Bilimleri Enstitüsü.

Gupta , S., Mittal , M., & Padha, A. (2017). Predictive Analytics of Sensor Data Based on Supervised Machine Learning ,International Conference on Next Generation Computing and Information Systems., (s. 173).

Haltaş, A., Alkan , A., & Karabulut, M. (2015). Metin Sınıflandırmada Sezgisel Arama Algoritmalarının Performans Analizi. *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, s.419.

Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). Word Cloud Explorer : Text Analytics based on Word Clouds. *47th Hawaii International Conference on System Science*, (s. 1835).

Kalaycı, T. E. (2018). Kimlik Hırsızı Web Sitelerinin Sınıflandırılması Için Makine Öğrenmesi Yöntemlerinin Karşılaştırılması. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, s. 874-875.

Karakoyun, M., & Hacıbeyoğlu, M. (2014). Biyomedikal Veri Kümeleri İle Makine Öğrenmesi Sınıflandırma Algoritmalarının İstatistiksel Olarak Karşılaştırılması. *DEÜ Mühendislik Fakültesi Mühendislik Bilimleri Dergisi*, s.34-35.

Lemaitre, G., Nogueira, F., Oliveira, D., & Aridas , C. (2016). *Welcome to imbalanced-learn documentation*. Imbalanced Learn: https://imbalanced-learn.readthedocs.io/en/stable/index.html adresinden alındı

Méndez, J. R., Iglesias, E. L., Fdez-Riverola, F., Díaz, F., & Corchado, J. M. (2005). Tokenising, Stemming and Stopword Removal on Anti-spam Filtering Domain. *CAEPIA'05 Proceedings of the 11th Spanish association conference on Current Topics in Artificial Intelligence*, (s. 449).

Minh, H. (2018). *How to Handle Imbalanced Data in Classification Problems*. Medium: https://medium.com/james-blogs/handling-imbalanced-data-in-classification-problems-7de598c1059f adresinden alındı

Rocca, B. (2019). *Handling imbalanced datasets in machine learning*. Towards Data Science: https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e84220f28 adresinden alındı

Sağbaş, E. A., & Ballı, S. (2016). Akıllı Telefon Sensör Verileri ile Eylem Tanımada Lojistik Regresyon ve kNN Yöntemlerinin Karşılaştırılması. *1st International Conference on Engineering Technology and Applied Sciences*, s.895.

Schapire, R. (2008, 02 04). *Theoretical Machine Learning.* 06 14, 2019 tarihinde Theoretical Machine Learning: http://www.cs.princeton.edu/courses/archive/spr08/cos511/scribe_notes/0204.pdf adresinden alındı

Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning. *Expert Systems With Applications*, s.120.

Uzun, E. (2007). İnternet Tabanlı Belge Erişimi Destekli Bir Otomatik Öğrenme Sistemi. Yayımlanmamış Doktora Tezi, Trakya Üniversitesi, Fen Bilimleri Enstitüsü.

Yang, J., Liu, L., Zhang, L., Li, G., Sun, Z., & Song, H. (2019). Prediction of Marine Pycnocline Based on Kernel. *MDPI Sensors*, s.1562.

Yıldız, B., & Ağdeniz, Ş. (2018). Muhasebede Analiz Yöntemi Olarak Metin Madenciliği. *Muhasebe Bilim Dünyası Dergisi*, s. 290 – 296.

Yılmaz, R. (2013). Türkçe Dokümanların Sınıflandırılması. Yayımlanmamış Doktora Tezi, Adnan Menderes Üniversitesi, Fen Bilimleri Enstitüsü.

Zainuddin, N., & Selamat, A. (2014). Sentiment Analysis Using Support Vector Machine. *IEEE International Conference on Computer, Communication, and Control Technology*, (s. 334).

# EKLER

## Ek-1. Kullanılan Parametreler

| Terim Ağırlıklandırıcı | Parametreler |
|---|---|
| CountVectorizer | stop_words="english", ngram_range=(1, 2) |
| TfidfVectorizer | stop_words="english",ngram_range=(1, 2) |

| Sınıflandırıcı | Parametreler |
|---|---|
| MultinomialNB | Default |
| LogisticRegression | class_weight='balanced',solver = 'liblinear', multi_class='auto' |
| SVC (RBF) | class_weight='balanced',C=10, gamma = '1e-3' , kernel='rbf' |

## Ek-2. Özgeçmiş

**Kişisel Bilgiler**

Soyadı, adı: IŞIK, Muhammed

Uyruğu: T.C.

Doğum tarihi ve yeri: 27.04.1988 Erzurum

Medeni hali: Evli

Telefon: 0 (543) 3535875

e-mail: endustri2526@gmail.com

**Lisans Bilgileri**

Mezuniyet Tarihi: 13.06.2013

Okul Adı: Atatürk Üniversitesi

Bölüm Adı: Endüstri Mühendisliği

Mezuniyet Ortalaması: 2,76/4,00

**Yandal Bilgileri**

Mezuniyet Tarihi: 23.07.2013

Okul Adı: Atatürk Üniversitesi

Bölüm Adı: Ekonometri

Mezuniyet Ortalaması: 3,14/4,00

**İlgi Alanları**

Information Management, Machine Learning, Artificial Intelligence, Text Mining, Web Minings