

Otel İncelemesi (Hotel Review)

Didem ŞENUSTA¹

1. Yaşar Üniversitesi, Bilgisayar Mühendisliği Bölümü

Özet

Sürekli gelişen teknoloji ve internetin ortaklaşa dünyasında sosyal ağlar, seyahat bilgilerini paylaşmak için en popüler platform haline gelmiştir. Tripadvisor, booking gibi otel rezervasyonu web siteleri, müşterinin karar verme sürecine yardımcı olmak için otel yorumlarını ve puanlamaları kullanır. Bu makale de otel hakkında yapılan yorumlar ve puanlamalar incelenmiştir. Ayrıca metin madenciliği (text mining) kullanılarak, olumlu ve olumsuz kelimeler ayrıştırılıp bu kelimelerin temelini incelemeyi amaçlamaktadır. Bir metin madenciliği yaklaşımı takip edilerek, 2 kategoriye ayrılmış müşteriler tarafından yapılan yorumlar karşılaştırılacak ve buna bağlı olarak her otelin puan ortalamaları alınacaktır. Duygu analizi yöntemiyle, text madenciliğinde ayrıştırılmış kelimeler kullanılarak negatif ve pozitif yorumlar ayrılacak ve diğer işlemlerle birlikte tüketiciye en uygun otel sunulacaktır. 1.000 otel misafirinin online yorumları Amerika'nın farklı şehirleri için kaggle.com'dan toplanmıştır.

Anahtar Kelimeler: text mining, machine learning, data mining, data science

Abstract

Social networks have become the most popular platform for sharing travel information in the world of ever-evolving technology and the Internet. Tripadvisor uses hotel booking websites such as booking websites, hotel reviews and ratings to help the client's decision-making process. In this article, reviews and ratings about the hotel are examined. In addition, by using text mining, positive and negative words are searched and the basis of these words is examined. By following a text mining approach, comments made by 2 categorized customers will be compared and the average score of each hotel will be taken accordingly. By means of emotion analysis, negative and positive interpretations will be made using the separated words in text mining and the most suitable hotel will be presented with the other operations. Online comments of 1,000 hotel guests are collected from Kaggle.com for different cities in the United States.

Keywords: text mining, machine learning, data mining, data science, hotel review

1. Giriş

Günümüzde teknolojinin ilerlemesiyle birlikte içeriğinden anlamlı sonuçlar çıkarmak ve gerektiğinde kullanmak üzere pek çok veri toplanmaktadır. Bu verileri şekiller, grafikler, metinler gibi malzemeler oluşturmaktadır ve bilgisayarlarla elektronik ortamlara taşınmış bulunmaktadır.

Elektronik ortamlara taşınmış dijital verilerin hacmindeki artış beraberinde yeni sorun alanları da yaratmıştır. Bunların başlıcaları; karmaşık verileri, yeni tür verileri, dağılmış verileri işlemek ve yeni yöntem geliştirmek; verilerin kullanımı ve güvenliğiyle ilgili modeller geliştirmek olarak sıralanabilir. Teknolojinin her geçen gün artan bir ivmeyle ilerlemesiyle birlikte, bu verilerin depolanması, işlenmesi, kullanıcının ihtiyaçlarını karşılıyor olması ve sorunsuz bir şekilde çalışması beklenmektedir. Bilimsel yöntemlerle mevcut durum analizi yaparak sistemlerin güçlü-zayıf, olumlu-olumsuz, negatif-pozitif yönlerinin belirlenmesi toplanan veriler sayesinde sağlanabilmektedir. Böylelikle elde edilen veriler sayesinde sistemin durumu incelenebilir, iyileştirilmesi veya güncellenmesi gereken kısımlar netleştirilebilir.

Büyük veri denilince akla ilk olarak Veri Madenciliği gelmektedir. Veri Madenciliği, büyük verilerden örüntü ve ilişki keşfederek, istenilen işe yardımcı olacak anlamlı ve yararlı bağlantı ve kuralların bilgisayar programlarının aracılığıyla aranması ve analizidir. [1]

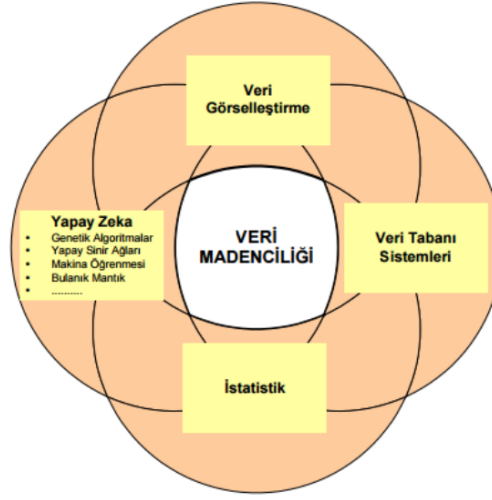
Veri madenciliğinin yoğun şekilde kullanıldığı alanlardan biriside turizm sektörü olmuştur. Günümüzde seyahat planlama ve rezervasyon en önemli ticari kullanımlarından biri haline gelmiştir. Bir seyahatin planlanmasında seyahat deneyimleriyle ilgili yorum, rapor ve verilen puanlamalar bilgi kaynağı olarak artan bir rol oynamaktadır. Bu aşamada bilgi türü ile ilgili bir sorun ortaya çıkmaktadır. Çevrimiçi otel rezervasyonları için kullanılan web sitelerinde müşterinin bir otel hakkında bilgi edinebilmesi için yorumlar ve puanlamalar büyük önem taşımaktadır. Fakat bu yorumlar pozitif veya negatif olarak ayrıştırılmış, puanlamalar ise ortalaması alınmış halde müşteriye sunulmamaktadır.

Bu çalışmada, bu eksikliğin giderilmesi için kullanıcıların yorumları üzerinde metin madenciliği (text mining) ve duygu analizi (sentimental analysis) yöntemleri ile otellere verdikleri puanların ortalamalarına bakılarak oteller arasında kıyaslama yapılması ve tüketiciye en uygun oteli pazarlama amacı güdülmüştür.

2. Veri, metin madenciliği ve duygu analizi

Veri madenciliği çok geniş bir kavramdır ve içinde birden fazla metotları barındırmaktadır. Veri madenciliği, veriler arasındaki ilişkiyi ortaya çıkarırken birçok dalı birlikte kullanır ve daha çok görsel olarak bir programlama ara yüzü üzerinden yapılan bir süreçtir. [2] İstatiksel, matematiksel algoritmalar gibi algoritmalar veri madenciliği algoritmaları içindedir.

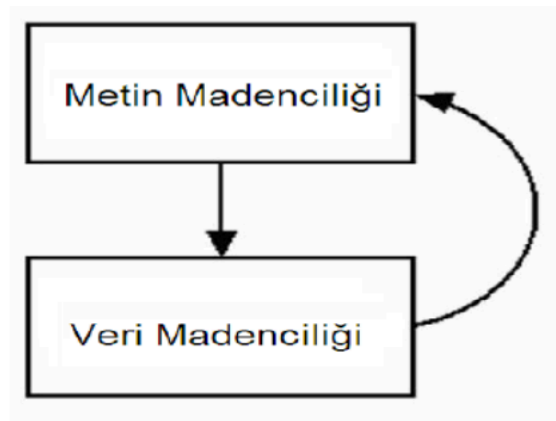
Veri madenciliği yapısal bir veriyi analiz edebilirken, metin madenciliği veri madenciliğinde kullanılmak üzere verileri yapısal hale dönüştürmektedir. Tek bir veri madenciliği modeli yoktur, birbirleriyle bağlantı halinde olan modellerin birleşimi vardır. Şekil 1’de gösterildiği gibi veri madenciliği ve kullanım alanları birbirleri ile bağlantılıdır. [3]

Şekil 1. Veri madenciliği ilgili alanları

2.1. Metin Madenciliği

Veriler farklı şekillerde bulunabilir. Bunların kullanılmasında da farklı metotlar uygulanabilir. Bir verideki anlam ortaya çıkarılmak isteniyorsa metin madenciliği yöntemi uygulanmaktadır.

Metin madenciliği, metinlerin sınıflandırılması ve konu çıkarılması, sınıf taneciklerinin üretilmesi, duygusal analiz, metin özetleme, varlık-ilişki modellemesi gibi çalışmaları hedefler. [4] Natural Language Processing (NLP) yani dilbilimsel yollar olarak ifade edilen terim aslında metin madenciliğinin temelidir. NLP ile bir metin içerisinde istenilen ifade, koşullara uygun çıkarılabilmekte, metin içerisindeki terimler sınıflandırılıp, bir sınıfa atanmaktadır. Metin ve veri madenciliği arasındaki ilişki Şekil 2’de tanımlanmıştır;

**Şekil 2.** Süreçler arasındaki ilişki

Metin madenciliği uygulamaları iki ana sınıfta incelenebilir:

- Metnin anlaşılması/özetlenmesi: İşlenecek metinden uygulanabilir ve anlamlı bilginin çıkarılmasıdır. Böylelikle metnin ifade etmek istediği anahtar içerik anlaşılacaktır.
- Metin ile modelleme: Anahtar içerik anlaşıldıktan sonra elde edilen içerik ile diğer bilgiler kullanılarak bir modelleme oluşturulur.

2.2. Duygu Analizi

Duygu analizi metodu da bir metin işleme işlemi olup verilen metnin duygusal olarak ifade etmek istediği sınıfı belirlemeyi amaçlar. Duygu analizinin ilk çalışmaları olumlu (positive), olumsuz (negative) ve nötr olarak sınıflandırılmıştır ve bu duygusal kutupsallık (sentimental polarity) olarak geçmektedir. [5] Metinlerdeki ruh hali, kanaat ve daha karmaşık duyguların çıkarılması üzerinde de çalışılmaktadır.

Şekil 3'de bir duygu analizi yönteminin bir ürün üzerindeki basamakları sınıflandırılmıştır.



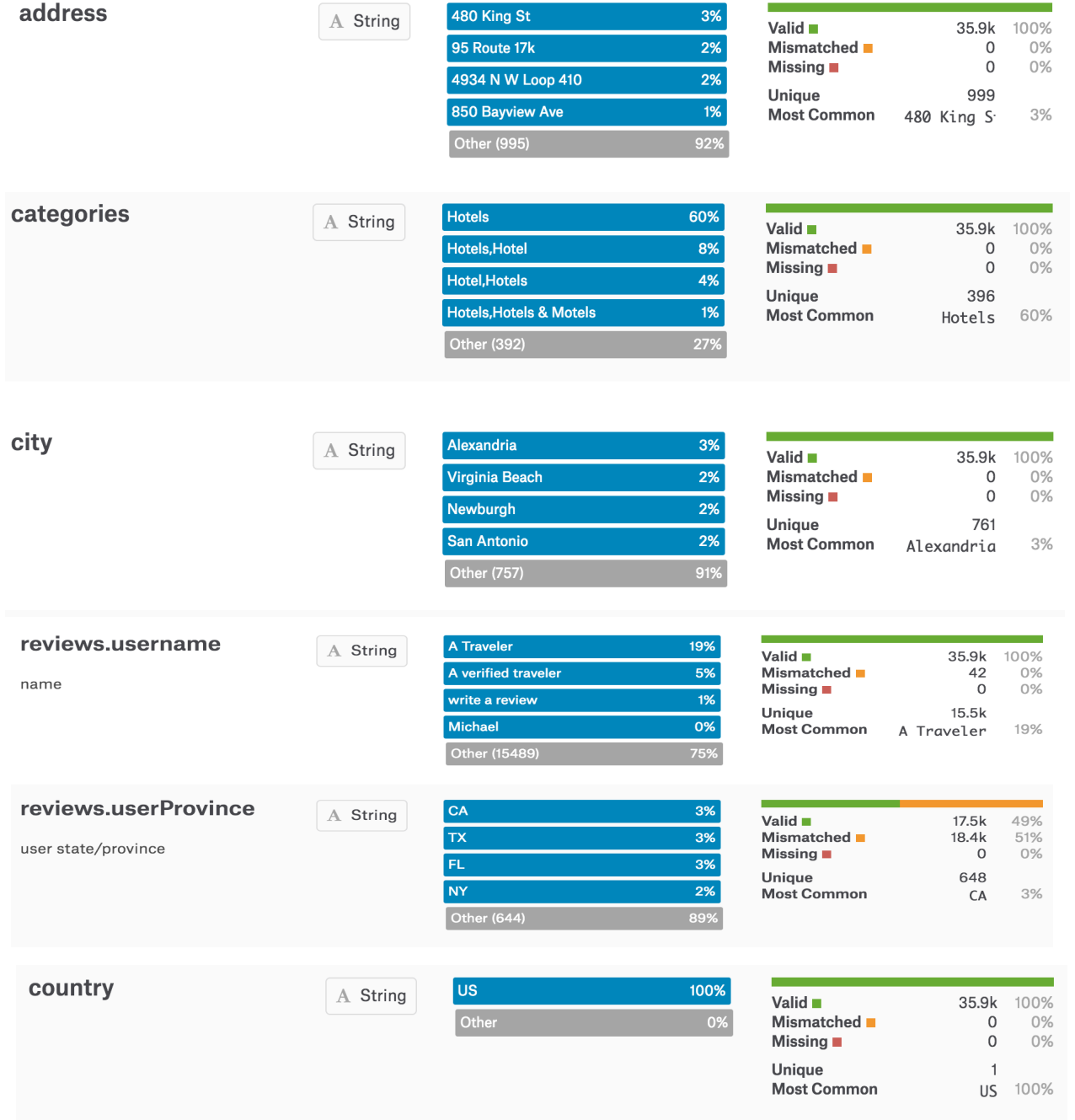
Şekil 3. Ürün yorumlarındaki DA süreci

3. Veriler

Bu çalışmada, kaggle.com üzerinden 1,000 farklı otelin veri kümesi alınmıştır. Bu veri kümeleri otel konumu, ad, derecelendirme, yorumlar, kullanıcı adı gibi öğeler içeren 20 farklı kolondan oluşmaktadır. Bu veriler ile otel değerlendirmelerinin karşılaştırılması ve duygu puanlaması yapılabilmektedir. Veri setinde, otel hakkında yapılan yorumlardaki anahtar kelimeler metin madenciliği yöntemi kullanılarak derecelendirilmeyle ilişki kurulmasına olanak sağlamaktadır. Knime üzerinde farklı şehirlerdeki oteller hakkında yapılan yorumlamalar alınmış, duygu analizine yardımcı olacak şekilde olumlu ve olumsuz kelimeler çıkartılmıştır. Metin madenciliği yönteminden faydalanarak, veriler eğitim ve test olarak ikiye bölünmüş ve makine öğrenmesi algoritması olan sınıflandırma yöntemleri denenerek istatistiksel veriler tablolaştırılmıştır. Veriler Şekil 4'te gösterilmiştir.

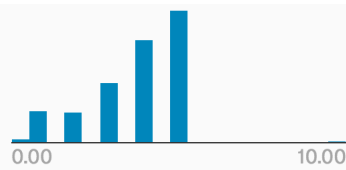
5.2 Analiz ve Karşılaştırma

Knime’de yer alan istatistiksel veriler bu çalışma kapsamında kullanılmıştır. Kolonlara ait kullanım sıklığı, unutulmuş değerlerin kontrolü, integer değere sahip olan kolonların standart sapmaları ve ortalamaları projenin ilerlemesi açısından büyük önem taşımaktadır. Bu verilerin yardımıyla hangi kolonlar üzerinde çalışılması gerektiği kararlaştırılmıştır. Şekil 6’da kategorilendirilmiş istatistikler mevcuttur.



reviews.rating

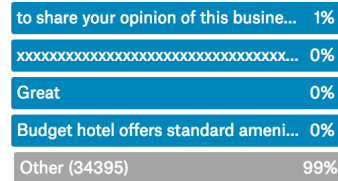
Numeric



Valid	35.0k	98%
Mismatched	862	2%
Missing	0	0%
Mean	3.69	
Std. Deviation	1.42	
Quantiles	0.00	Min
	3.00	25%
	4.00	50%
	5.00	75%
	10.00	Max

reviews.text

A String

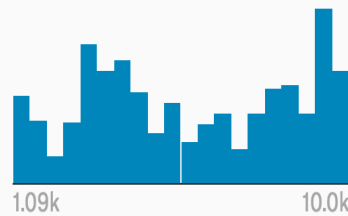


Valid	35.9k	100%
Mismatched	20	0%
Missing	0	0%
Unique	34.4k	
Most Common	to share y	1%

postalCode

Numeric

postalcode



Valid	33.4k	93%
Mismatched	2545	7%
Missing	0	0%
Mean	5.0k	
Std. Deviation	3.0k	
Quantiles	1.09k	Min
	2.9k	25%
	5.0k	50%

province

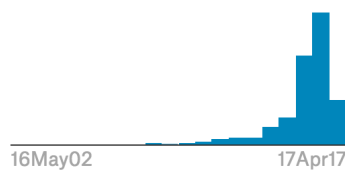
A String



Valid	35.9k	100%
Mismatched	0	0%
Missing	0	0%
Unique	287	
Most Common	CA	11%

reviews.date

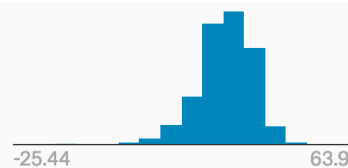
Date



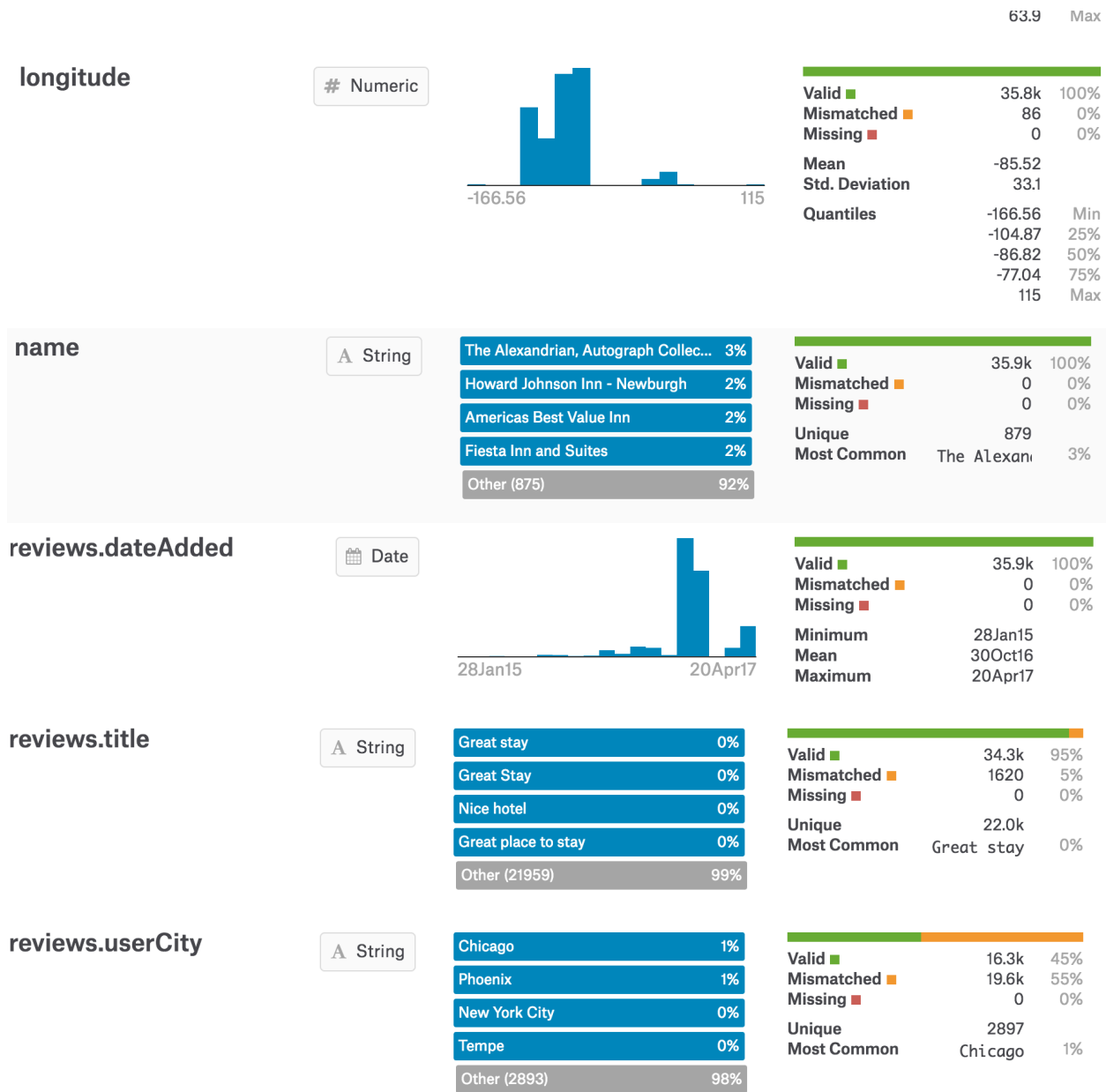
Valid	35.7k	99%
Mismatched	259	1%
Missing	0	0%
Minimum	16May02	
Mean	13May15	
Maximum	17Apr17	

latitude

Numeric



Valid	35.8k	100%
Mismatched	86	0%
Missing	0	0%
Mean	37.2	
Std. Deviation	7.00	
Quantiles	-25.44	Min
	33.8	25%
	37.9	50%
	41.7	75%
	63.9	Max



Şekil 6. Kategorilendirilmiş istatistikler

5.3 Python

Python, nesne yönelimli, yorumlamalı, birimsel ve etkileşimli yüksek seviyeli bir programlama dilidir. Knime üzerinden modellemesi yapılan projenin her bir adımı Python’da kodlanmıştır. Pandas, numpy ve matplotlib kütüphaneleri kurularak kodlanmaya başlanmıştır. Ardından hangi kolon kullanılacaksa kolon filtreleme kodu yazılmış ve ardından yorumlar ele alınmış ve küçük-büyük harf, varsa Türkçe karakterler silinmiştir. Kelimeler olumlu-olumsuz ayrıştırılmıştır. Sınıflandırma yapmak için eğitim ve test olarak 2’ye ayrılmıştır. Birkaç sınıflandırma yöntemi (Logistic Regression, Decision Tree) denenerek en iyi sonuç veren yöntem ele alınarak kodlama yapılmıştır.

5. Sonuç

Bu çalışma, çevrimiçi web sitesinde belirtilen otellerin kullanıcı yorumlarının ve puanlamalarının analizi ile özetlenmesine odaklanmaktadır. İnceleme özeti, kullanıcılara kararsız kaldıkları veya seçtikleri oteller arasından en iyi seçimin hangisi olacağı hakkında daha iyi bir seçim sağlar. Çalışma ayrıca müşteriye otellerin puanlamalarını yorumlarla birlikte göstererek karşılaştırma olanağı sağladığından kararsızlığı ortadan kaldırmayı ve seçilen bölgedeki en iyi otelin pazarlamasının yapılmasını hedeflemiştir.

Teknolojinin günümüzde giderek önem kazanması ile beraber popüler bir konu haline gelmiş veri madenciliği genel olarak anlatılmış ve teknolojinin ilerlemesiyle birlikte gelişen turizm sektörü ele alınmıştır. 1,000 adet farklı otelin kullanıcılarına ait yorumları Knime platformu üzerinde metin madenciliği algoritması kullanılarak sentezlenmiş ve yine veri madenciliğinde çok popüler bir algoritma olan duygu analizi algoritması ile birlikte otellere ait kullanıcılar tarafından yapılan yorumlar ‘pozitif’ ve ‘negatif’ olarak gruplandırılmıştır. Otellere ait puanlamaların ortalaması alınarak birleştirilmiş ve tüketiciye en uygun otelin reklamının yapılabilmesi için Python ortamı kullanılarak kodlanmıştır.

Metin madenciliği ve duygu analizi yönteminin çok daha başarılı sonuçlar verdiği görülmüştür.

Kaynakça

- [1] SAVAŞ, S., TOPALOĞLU, N., YILMAZ, M.,(2011), Veri Madenciliği ve Türkiye’deki Uygulama Örnekleri, *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 11(21), 1-23.
- [2] DOLGUN, M. Ö., ÖZDEMİR, G. T., OĞUZ, D.,(2009), Veri madenciliği’nde yapısal olmayan verinin analizi: Metin ve web madenciliği, *İstatistikçiler Dergisi*, 2(2009), 48-58.
- [3] Albayrak, M. (2017). The use of data mining in scientific research. *International Journal of Social Sciences and Education Research*, 3(2), 751-760.
- [4] ŞEKER, E., Ş.,(2014), *Metin Madenciliği(Text Mining)*, 25 Temmuz 2018 tarihinde Bilgisayar Kavramları: <http://bilgisayarkavramlari.sadiyevrenseker.com/2014/06/15/metin-madenciligi-text-mining/> adresinden alındı.
- [5] ŞEKER, E., Ş.,(2016), Duygu Analizi(Sentimental Analysis), *YBS Ansiklopedi*, 3(3), 22-36.