
GAZİ ÜNİVERSİTESİ
Gazi
1926

ÇİN RESTORANINA YAPILAN MÜŞTERİ YORUMLARININ METİN MADENCİLİĞİ TEKNİKLERİ VE MAKİNE ÖĞRENME YÖNTEMLERİYLE ANALİZ EDİLMESİ

Elif BOZTÜRK KILINÇ

YÜKSEK LİSANS TEZİ
İSTATİSTİK ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

NİSAN 2024

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Elif BOZTÜRK KILINÇ

16/04/2024

ÇİN RESTORANINA YAPILAN MÜŞTERİ YORUMLARININ METİN MADENCİLİĞİ TEKNİKLERİ VE MAKİNE ÖĞRENME YÖNTEMLERİYLE ANALİZ EDİLMESİ

(Yüksek Lisans Tezi)

Elif BOZTÜRK KILINÇ

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Nisan 2024

ÖZET

Metin madenciliği, büyük metin veri setlerini keşfetme, anlama ve özetleme konusunda bir dizi teknik ve yöntemi içerir. Bilgi çıkarma ve keşif süreçlerine rehberlik eder. Büyük metin veri setlerinde gizlenmiş olan temaları ve önemli bilgileri belirleyerek, kullanıcıların odaklanmalarını sağlar. Aynı zamanda duygu analizi yaparak, metinlerdeki duyguları─anlamak da metin madenciliğinin önemli bir işlevidir. Müşteri yorumları veya sosyal medya paylaşımları üzerinden duygu analizi yapmak, işletmelerin ürün veya hizmet memnuniyet seviyelerini değerlendirmelerine yardımcı olur. Bu çalışma, Çin restoranlarının online platformlardaki müşteri yorumlarını metin madenciliği yöntemleri olan Latent Dirichlet Allocation (LDA) ve Bags of Words (BoW) kullanarak analiz etmeyi ve bu analizler aracılığıyla restoranların hizmet kalitesi ve müşteri memnuniyeti hakkında bilgi elde etmeyi amaçlamaktadır. Metin madenciliği yöntemleri kullanılarak belirlenen konular, restoranın çeşitli yönlerini temsil etmekte ve bu konuların içeriği, müşteri yorumlarındaki belirgin kelimeler aracılığıyla anlaşılmaktadır. Makine öğrenimi uygulamasıyla yapılan analizler ile müşteri yorumlarına dayanarak restoranın genel kalitesini ve müşteri memnuniyeti tahmin edilmiştir. Bu analizde kullanılan Çok Katmanlı Algılayıcı (MLP), Karar Ağacı (DecisionTree), Stokastik gradyan inişi (SGD), AdaBoost, GradientBoosting, LogisticRegression, RandomForest, Destek Vektör Makinesi (SVC), LightGBM, Kneighbors ve XGBoost algoritmalar arasında en iyi performansı gösteren algoritmanın Multilayer Perceptron (MLP) olduğu belirlenmiştir. MLP'nin, müşteri yorumlarından elde edilen verilerle restoranın kalitesini değerlendirme konusunda yüksek doğruluk oranları sağladığı görülmüştür. Bu çalışmada yapılan analizler, restoranın müşteri profilini daha iyi anlamak, pazarlama stratejilerini optimize etmek ve rekabet avantajı sağlamak için kullanılabilir. Bu yöntemler bilgi çağında rekabet avantajı elde etmek isteyen işletmeler için vazgeçilmez bir öneme sahiptir.

Bilim Kodu : 20513

Anahtar Kelimeler : Metin Madenciliği, Müşteri, Makine Öğrenmesi Python

Sayfa Adedi : 93

Danışman : Doç. Dr. Filiz KARDİYEN

ANALYZING CUSTOMER COMMENTS OF A CHINESE RESTAURANT USING TEXT MINING TECHNIQUES AND MACHINE LEARNING METHODS

(M. Sc. Thesis)

Elif BOZTÜRK KILINÇ

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

April 2024

ABSTRACT

Text mining involves a set of techniques and methods for exploring, understanding and summarizing large text data sets. It guides knowledge extraction and discovery processes. It allows users to focus by identifying themes and important information hidden in large text data sets. At the same time, understanding emotions in texts by performing sentiment analysis is also an important function of text mining. Conducting sentiment analysis on customer reviews or social media posts helps businesses evaluate product or service satisfaction levels. This study aims to analyze customer comments of Chinese restaurants on online platforms using text mining methods Latent Dirichlet Allocation (LDA) and Bags of Words (BoW) and to obtain information about the service quality and customer satisfaction of restaurants through these analyses. Topics identified using text mining methods represent various aspects of the restaurant, and the content of these topics is understood through distinct words in customer comments. The overall quality of the restaurant and customer satisfaction were estimated based on customer comments and analyzes made with the machine learning application. Among the Multilayer Perceptron (MLP), Decision Tree (DecisionTree), Stochastic gradient descent (SGD), AdaBoost, GradientBoosting, LogisticRegression, RandomForest, Support Vector Machine (SVC), LightGBM, Kneighbors and XGBoost algorithms used in this analysis, the algorithm with the best performance was determined. It was determined to be Multilayer Perceptron (MLP). MLP has been shown to provide high accuracy rates in assessing restaurant quality with data obtained from customer reviews. The analyzes made in this study can be used to better understand the restaurant's customer profile, optimize marketing strategies and gain competitive advantage. These methods are of indispensable importance for businesses that want to gain competitive advantage in the information age.

Science Code : 20513

Key Words : Text Mining, Client, Machine Learning Python

Page Number : 93

Supervisor : Assoc. Prof. Dr. Filiz KARDİYEN

TEŞEKKÜR

Yüksek lisans eğitimim boyunca bana sabır ve sevgiyle, engin bilgileri ve tecrübeleri ile ışık olan, tez çalışmam dışında da her danıştığım konuda desteğini gördüğüm sevgili tez danışmanım Doç. Dr. Filiz KARDİYEN'e ve değerli katkılarından dolayı sevgili jüri hocalarım Prof. Dr. Hacı Hasan ÖRKCÜ ve Prof. Dr. Furkan EMİRMAHMUTOĞLU'na teşekkür ederim.

Tez çalışmamın her anında desteğini ve emeğini esirgemeyen, aramızdaki mesafelere ve zorlu görevlerine rağmen sabırla yardım eden, çok değerli eşim İsmail KILINÇ'a teşekkür ederim.

Hayatımın her anında sonsuz destekleri ve sevgileri ile yanımda olan annem Zafer BOZTÜRK, babam Şahin BOZTÜRK ve kıymetli kardeşlerime teşekkür ederim. Tez çalışmamın iş hayatımla beraber ilerlemesinde sabır ve anlayış gösteren Onur DUYGUN ve Hakan YILMAZ başta olmak üzere tüm çalışma arkadaşlarıma desteklerinden dolayı teşekkür ederim.

İÇİNDEKİLER

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	**Açıklamalar**
$\prod$	Çarpım Sembolü
$\sum$	Toplam Sembolü
$\boldsymbol{\alpha}$	Alfa
$\boldsymbol{\theta}$	Teta

Kısaltmalar	**Açıklamalar**
BoW	Bags of Words (Kelime Torbaları)
LDA	Latent Dirichlet Allocation (Gizli Dirichlet Ayrımı)
ML	Makine Öğrenmesi
MLP	Multilayer Perceptron (Çok Katmanlı Algılayıcı)
NLP	Natural Language Processing (Doğal Dil İşleme)
PCA	Principal Component Analizi (Temel Bileşen Analizi)
SGD	Stochastic Gradient Descent (Stokastik gradyan inişi)
SVM	Support Vector Machine (Destek Vektör Makinesi)

1. GİRİŞ

Veri madenciliği, günümüzde büyük miktarda veri üreten ve bu verilerden anlamlı bilgiler çıkarmak amacıyla kullanılan bir disiplindir. İstatistik, matematik ve bilgisayar bilimleri alanlarını içine alan veri madenciliği, bu verileri analiz ederek öngörülerde bulunmaya, bilgi çıkarmaya ve kararlar almaya yönelik bilgiler elde etmeyi hedefler. Veri madenciliği, bilgisayarların ve teknolojinin gelişmesiyle birlikte daha da önem kazanmıştır. Metin madenciliği ise 20. yüzyılın sonlarına doğru yapay zekâ teknolojisinden farklı bir dal olarak meydana gelmiştir. Metin madenciliği veri madenciliğinin parçası olarak görülse bile geleneksel veri madenciliğinden farklı denilebilir.

Takan (2022)'a göre temel fark olarak, metin madenciliğinin olaya dayalı bir veri tabanı yerine doğal dildeki metinlerden desenler çıkarmasıdır. Olesen ve Kisjes (2018)'e göre ise metin madenciliği, elektronik ortamda bulunan yapılandırılmamış ve kaotik metin yığınlarından önceden bilinmeyen, kullanışlı, yapılandırılmış ve organize edilmiş verilerin elde edilmesi süreci olarak tanımlanabilir. Teknoloji ve dijital dünyanın hızla genişlemesiyle birlikte, büyük miktarda bulunan metin verileri önümüze çeşitli şekillerde gelmektedir. Bu metinler, sosyal medya paylaşımları, müşteri yorumları, haber makaleleri ve daha birçok kaynaktan oluşmaktadır. Bu büyük veri denizi içinde kaybolmadan, bu verilerden anlamlı bilgiler çıkarmak ve içerdikleri değeri ortaya çıkarmak ise metin madenciliğinin önemli bir rol oynadığı alandır. Metin madenciliği, doğal dil işleme (NLP) ve makine öğrenimi tekniklerini kullanarak metin verilerini analiz etme sürecini ifade eder. Doğal dil işleme, metin madenciliği için temel olan dilbilim ve bilgisayar bilimini birleştirirken, makine öğrenimi ve veri madenciliği metinlerdeki desenleri ve yapılardaki ilişkileri bulmak için kullanılır. Temel amacı, metinlerin içinde gizli kalmış bilgileri çıkararak, anlamlı desenleri ve ilişkileri keşfetmektir.

Kapukaya (2023)'ya göre insanların duygu ve düşüncelerini anlamanın geleneksel yöntemlerden olan anket ve diğer araştırma yöntemlerinin yanı sıra, metin madenciliğinin temel görevlerinden biri de yazılı metinleri işleyerek bu metinleri geniş bir veri tabanı olarak analiz edip sayısal değerlere dönüştürmektir. Hassani, Beneki, Unger, Mazinani ve Yeganegi, (2020) çalışmalarında her türlü iş modeli, pazar araştırması, pazarlama planları, siyasi kampanyalar veya stratejik karar verme, rekabeti ele almak için metin madenciliği

tekniklerine artan bir ihtiyaçla karşı karşıya olunduğundan bahseder. İş dünyasında müşteri ilişkileri yönetimi, pazarlama stratejileri ve risk analizi gibi alanlarda önemli bir rol oynamaktadır.

Gelecekte yapay zeka ve otomatik metin anlama sistemlerinin geliştirilmesiyle daha da büyük bir öneme sahip olacağı öngörülmektedir. Günümüzde, daha karmaşık algoritmalar ve derin öğrenme teknikleri, metin madenciliğinin daha hassas ve etkili olmasını sağlamıştır. Ayrıca, büyük veri setlerinin artması ve bilgiye erişimdeki gelişmeler, metin madenciliğinin daha yaygın bir şekilde kullanılmasını sağlamıştır. Bu sayede, örneğin, müşteri yorumlarından pozitif ve negatif geri bildirimleri ayırt etmek, bir belge koleksiyonundan belirli konulara odaklanmak ile mümkün olabilir. Veri madenciliği ve metin madenciliği, günümüzde bilgiye erişimdeki artışla birlikte önemli hale gelmiş iki disiplindir. Günümüzdeki veri bolluğu içinde değerli bilgiler elde etmek için önemli araçlar haline gelmişlerdir. Bu teknikler, iş dünyasında rekabet avantajı sağlamanın yanı sıra, sağlık, güvenlik, eğitim gibi birçok alanda da önemli katkılar sunmaktadır. Veri ve metin madenciliğinin doğru bir şekilde uygulanması, daha bilinçli ve bilgi odaklı kararlar alınmasına olanak tanır. Bu teknoloji, metin verilerini analiz ederek bilgi çıkarma süreçlerini hızlandırır ve derinlemesine anlam sağlar.

Nilashi ve diğerleri (2021), çalışmalarında metin madenciliğinin, birçok farklı uygulama alanında mevcut olduğunu ve örneğin pazarlama alanında ise bir şirketin hedef kitlesini, tercihlerini ve ihtiyaçlarını karşılamak için müşteri yorumlarını analiz etmek için kullanılır. Bu analiz, şirketin müşteri portföyünü artırmayı ürün ve hizmetleri daha iyi bir şekilde pazarlamak için faydalı bilgiler sunduğunu tespit etmiştir. Şeker (2015a) ise metin madenciliğinde kullanılan ve veri setinde yer alan içeriklerin, veri tabanlarında olduğu gibi tam olarak düzenli olmadığından, ön işleme süreci yapılarak derlenmesi gerektiğinden ve bu sebeple doğal dil işleme yöntemleri aracıyla eldeki karmaşık ifadelerin anlamlı bir şekilde sınıflandırılması gerektiğini saptamıştır. Kılıç, Ateş, N ve Karakaya (2015) bu tür durumlar için metin madenciliği, veri kaynakları üzerinden alınan ve bu verilerin derlenerek analiz edilmesiyle birlikte çıkarımlarda bulunulması, hatta geliştirilebilecek algoritmalar ve yapay zekâ aracılığıyla bunlara yönelik çalışmalar oluşturulmasını bile mümkün kılabilir. Örneğin metin madenciliğinde, büyük miktarda metin verisi içeren belgeler, makaleler, e-postalar, sosyal medya gönderileri ve daha fazlasını inceleyerek, bilgi çıkarımı ve özetleme gibi işlemler gerçekleştirilir. Geniş veri kümelerinden anlam

çıkarılmasına ve bilgi keşfine katkıda bulunarak, bilgi çağında büyük bir öneme sahiptir. Bu alanın gelişimi, daha fazla veriyle başa çıkma ve bu verilerden değerli bilgiler elde etme ihtiyacının artmasıyla paralel olarak ilerlemiştir. Bunların yanında sınıflandırma, kümeleme, duygusal analiz gibi metin madenciliği görevleri, makine öğrenimi algoritmalarıyla gerçekleştirilir.

Isaeva ve Aldarova (2021)'e göre Büyük miktarda metinsel veriden anlamlı bilgiler çıkarmak için gizli anlamsal analiz veya ortalama bağlantı gibi algoritmalar kullanılır. Metin madenciliği, doğal dil işleme, makine öğrenimi ve veri madenciliği gibi alanlar arasında kesişen bir araştırma sahasıdır. Makine öğrenmesi ise bilgisayarların belirli görevleri, deneyimler ve veriler aracılığıyla öğrenmelerini sağlayan yapay zekânın önemli bir parçasıdır ve alt kümesidir. Bilgisayarlar, karmaşık sorunları çözme yetenekleriyle her geçen gün daha da gelişiyor. Bu gelişimlerin arkasındaki temel konulardan biri de "Makine Öğrenmesi"dir. Makine öğrenmesi, bilgisayar sistemlerinin deneyimlerle öğrenmelerine ve veriler üzerindeki işlemleri tanıyarak kendilerini geliştirmelerine olanak tanıyan bir disiplindir. Geleneksel programlamada olduğu gibi açıkça belirtilmiş bir dizi kuralla çalışmak yerine, makine öğrenmesi algoritmaları veri analizi ve süreçlerinde kendilerini eğitirler.

Diez-Olivan, Del Ser, Galar ve Sierra (2019), çalışmalarında Makine öğrenmesi araçlarının amacının, büyük veri analizine dayalı olarak öğrenme ve uyum sağlama yeteneği sağlayarak algoritmaların verimliliğini artırmak olduğundan bahsediyor. Bu hedeflere ulaşmak için uygulanan metin madenciliği çalışmaları genellikle metin yoluyla istatistiksel bilgi elde etmeyi amaçlamaktadır. Bu nedenlerden dolayı metin madenciliği, hızla büyüyen metin verilerinin analiz edilmesini ve anlamlı bilgilere dönüştürülmesini mümkün kılmaktadır.

Kapukaya (2023), çalışması kapsamında internet kullanıcıları tarafından internet üzerindeki IMDB listesindeki ve Twitter(X) kullanıcıları tarafından paylaşılan ve beğenilerin benzerlikleri duygu analizi teknikleri ile analiz edilmiştir. IMDB puanlamasında beğeni olarak yüksek puan ile derecelendirilen üç film için Twitter platformundaki kullanıcı yorumları ele alınarak analiz edilmiştir. Bahse konu filmler hakkında tweet olarak paylaşılan yorumların duygu puanlaması ile pozitif duygu kelimelerinin ağırlıklı olduğu tespit edilmiştir.

Şen (2022), çalışmasında Eğitim Yönetimi, Eğitim Yönetimi ve Planlaması, Eğitim Yönetimi ve Politikası gibi ana bilim dallarında yazılan lisansüstü tezlerinin metin madenciliği ile incelenmesi sonucunda eğitim yönetimi ana bilim dalı ile ilgili değerlendirme yapmıştır. Eğitim yönetimi programlarının genellikle lisansüstü düzeyde ilerlemesine rağmen doktora programı düzeyinde yapılan çalışmaların çok az olduğu görülmektedir. Üniversitelere göre dağılımına bakıldığında ise eğitim yönetimi tezlerinin çoğunluğu vakıf üniversiteleri tarafından üretilmiş olması dikkat çektiğini vurgulamıştır.

Akis (2019), çalışmasında omurilik cerrahisi hastaları tarafından elde edilen hasta hikâyeleri metin madenciliği yöntemleriyle incelenmiş ve karar sistemi oluşturulmuştur. Geliştirilen sistem, hasta hikâyeleri ile teşhis tahmininde yardımcı olmaktadır. Geliştirilen karar sistemi, yeni gelen verilerle kendini geliştirerek zamanla daha iyi kararlar vereceği ön görülmektedir. Geliştirilen karar sistemi ile bu konunun uzmanları için yanlış teşhisi önlemek amaçlanmıştır. Sistem ve tedavi detayları açıklanmış ve sonuçlar paylaşılmıştır. Çalışmanın sonuçları, sistemle başarılı önerilerde ve tecrübe aktarımında bulunulabileceğini göstermektedir. Sistemin hastaların tedavi kalitesini artırmada önemli katkılar sunabileceğini göstermektedir.

Daşgın ve Adem (2023), çalışmasında çevrimiçi öğrenme platformlarından biri olan Udemy akademi sitesinde bulunan kurslara yapılan öğrenci yorumlardan veri elde edilmiştir. Öğrenci yorumlarına verilen puanlamalar dahil olmak üzere kurs hakkında olumlu veya olumsuz değerlendirmeleri metin madenciliği yöntemleri, klasik makine öğrenmesi ve derin öğrenme açısından analiz edildiğinde Udemy (2022)'de bulunan kurs yorumlarının olumlu veya olumsuz olarak ayrılmasında başarılı sonuçlar elde edildiği gözlemlenmiştir.

Okatan (2023), yakın zamanda yaptığı çalışmasında internet ortamındaki kaynaklardan elde ettiği müşteri yorumlarının metin madenciliği yöntemleri ile müşterilerin internet bankacılığına bakış açıları analiz edilmek istenmiştir. Bunun yanında internet bankacılığının günümüz literatüründeki yeri ayrıntılı olarak incelenmek istenmiştir. Pandemi döneminin, internet bankacılığı üzerinde yenilikçi olarak pozitif etkiye sahip olduğu saptanmıştır. Bu kanıya pandemi dönemiyle birlikte yorum yapılma oranlarının yüksekliğiyle ilgilidir. Tespit edilen yorumların tarih veriler göstermektedir ki 2021-2022 aralığında veriler daha fazladır. Diyebiliriz ki pandemi dönemi ile internet bankacılığı kullanımı büyük ölçüde artmıştır.

Çakmak (2023), çalışmasında sosyal medya uygulamalarından olan Twitter(X) veri seti olarak kullanılmış ve metin madenciliği tekniği ile elde ettiği analiz dayanan modeli çalışılmıştır. Çalışmada, ilk olarak araştırma konusuna uygun konu etiketleri belirlenmiştir. Twitter(X)'dan konu ile ilgili veriler analiz edildiğinde, ülkemizde ve dünyada adını duyurmuş şirketlerden bahsedildiği ve çoğu şirketlerin neredeyse hiçbir başlıkta isimlerinin geçmediği tespit edilmiştir. Veri setimizi baz alarak yaptığımız analizler sonucunda, bu durumun bu tür şirketlerin ekonomide tekelleşmeye ulaştığı konusunda tespitte bulunmuştur. Rekabete katılmak için yeni şirketlerin ülkelerdeki rekabet politikaları tarafından desteklenmesi, denetleme yapılması ve takip edilmesi gerektiği belirtilmiştir. Bu tür önlemler yeni firmaların gelişebilmesi için tekelleşme durumunun önlenmesini sağlayacaktır.

Gürbüz (2022), çalışmasında DergiPark-Akademik dijital platformdaki araştırma makalelerini metin madenciliği yöntemi uygulayarak konularına göre sınıflandırmıştır. Tez çalışmaları için yapılan ön araştırma ve çalışma kapsamında, makalelerin öz kısmına göre yapılan sınıflandırmalara bakıldığında çalışma performanslarının yüksek olduğu tespit edilmiştir.

Özdemir (2022), çalışmasında yolcuların sosyal medyada bulanan değerlendirmelerini metin madenciliği yöntemleri ile araştırmanın veri seti olarak almıştır. İnceleme konuları olarak hizmet türlerine yönelik duygularının belirlenmesi ve fiyat-değer algısı, memnuniyet ve başkasına tavsiye davranışı etkilerinin incelenmesi yapılmıştır. Sonuçlara bakıldığında kabin bölgesindeki personelin müşteri ilişkileri, hizmet ve sunumu gibi konularda iletişim becerileri yüksek kişilerin işe alım süreçleri gibi durumlarda olumlu değerlendirilmesi uygulanmalarının havayolları şirketleri arasında rekabet sağlayacağı tespit edilmiştir.

Başer (2022), çalışmasında Japonya'da bulunan Henn-na Hotel çalışanları olan robotlarla ilgili müşterilerin yaptıkları yorumlara metin madenciliği ile duygu analizi gerçekleştirilmiştir. Bu kapsamda faaliyet gösteren Henn-na Hotel çalışanları olan robotlarla iletişim kuran bireylerin deneyimlerini paylaştıkları online sitelerde bulunan kullanıcı yorumları kullanılarak yapılan duygu analizi sonuçlarına bakıldığında olumlu, olumsuz ve nötr yorumlar içerisinden olumlu yorumların oranının daha fazla olduğu gözlenmektedir.

Atabay (2020), çalışmasında Antalya, Mayorka ve Şarm El Şeyh bölgelerindeki seçili otellere ait toplanan veriler metin madenciliği yöntemleri ile analiz edilmiştir. Bu bölgelerin seçiminde turizm ürünleri ve hizmetleri konusunda rakip olmaları ve Akdeniz içinde her şey dahil hizmeti olan otelleri barındırmaları önemli kriter olmuştur. Antalya bölgesinde bulunan otellerden bahsedilen yorumlarda odaların temiz ve düzenli olduğu, havalandırma ve gider yerlerinin kirli olduğundan bahsedilmektedir. Antalya'da ve Şarm El Şeyh'te yemeğin genellikle yorumlandığı görülmüştür. Özellikle çalışan servislerinde Antalya'nın çok yardımsever olduğu vurgulanmıştır. Mayorka bölgesinde ise otellerin berbat olduğu ama çalışanlar hakkında daha olumlu düşünceler olduğu görülmektedir.

Yüksel ve Tan (2018), yaptıkları çalışmada dijital medya üzerinde bulunan mekânlar için yapılan yorumlar ile metin madenciliği analizi gerçekleştirmiştir. Uygulanan analiz ile bahsedilen mekân için yapılan yorumların olumlu veya olumsuz olma durumu incelenmiştir. Dijital medya üzerinde yapılan yorumların analizleri sonucu internet kullanımının artması ile ticareti etkileyen veri miktarının da artması demek olduğu saptanmıştır.

Çalışmaların ışığında, metin madenciliği ve makine öğrenmesi yöntemlerinin kapsamlı bilgi çıkarımında birçok sektöre fayda sağladığı tespit edilmiştir.

Bu çalışma Çin restoranına ait müşteri yorumlarının internet siteleri ve yemek sipariş platformları aracılığıyla toplanmasının ardından metin madenciliği ve makine öğrenmesi yöntemleriyle analiz etme ve kullanıcıların restoranlara ilişkin görüşlerini ve deneyimlerini anlamlandırmayı amaçlıyor. Metin madenciliği teknikleri kullanarak yemek sitelerindeki müşteri yorumları üzerinde duygu analizi, kelime bulutu oluşturma ve trend analizi gibi işlemler gerçekleştirilebilir. Bu sayede, işletmeler müşteri geri bildirimlerini anlamak ve işletme kararları almak için veriye dayalı bilgiler edinebilirler.

Çalışmada bahsedilen Çin restoranları dünya genelinde büyük bir popülerliğe sahiptir. Aynı zamanda popülerliğini sürdüren, benzersiz bir yemek deneyimi ve lezzetler sunan mekânlar olarak bilinir. Çin mutfağı, binlerce yıllık bir tarihe sahip ve farklı bölgelerden etkilerle zenginleşmiş bir mutfaktır. Her ne kadar geleneksel Çin mutfağına ait yemekler sunsalar da her restoranın kendine özgü lezzetleri ve hizmet kalitesi vardır. Ancak, bu restoranların başarısı sadece lezzetli yemeklerle sınırlı değildir. Müşteri yorumları, restoran

işletmecileri ve yöneticileri için bu memnuniyeti ölçmek ve iyileştirmek adına önemli bir geri bildirim kaynağıdır. Bu geri bildirimler restoranların başarısı için hayati öneme sahiptir. Günümüzde, müşterilerin yemek deneyimleri hakkında görüşlerini ifade etmek için internet üzerinde birçok platform bulunmaktadır. Yemek siteleri ve online seyahat acenteleri gibi siteler kullanıcıların yemek yeme deneyimleri hakkında geri bildirimler vermek için sıklıkla kullandıkları platformlardan biridir. Bu geri bildirimler, işletmelerin hizmetleri hakkında önemli bilgiler sağlar ve müşteri memnuniyetini artırmak için kullanılabilir. Buradaki veriler, restoran işletmecilerinin hizmet kalitesini artırmak ve müşteri memnuniyetini sağlamak için önemli bir bilgi kaynağı olduğu gibi müşterilerin mekâna gitmeden önce alacakları hizmet ve ürünler hakkında bilgi sahibi olmasını sağlamaktadır. İnternetin hayatımızın her alanına yayılması insanlara fikir, duygu ve düşüncelerini paylaşabilecekleri sosyal medya, forum, blog ve e-ticaret siteleri benzeri sanal ortamlar sunmaktadır.

Özyurt ve Akçayol (2017)'a göre insanların görüşlerinin ifade edildiği bu sanal ortamlarda çok büyük miktarlarda kıymetli veri bulunmaktadır. Bu bilgiler, ürün ve hizmet üreten veya satan firmalar başta olmak üzere çok fazla kişinin ve araştırmacının ilgisini üzerine çekmektedir. Müşterilerin çeşitli restoran, otel vs. seyahat edebilecekleri her türden mekanlara gelen yorumları bünyesinde bulunduran internet siteleri ve yemek sipariş platformları insanların değerlendirmelerini ve deneyimlerini paylaşmaları için zengin bir veri kaynağı sağlamaktadır. İnsanlar bu tür mecralara aktardıkları deneyimlerini diğer müşterileri de büyük ölçüde yönlendirmektedir.

Kotler, Armstrong, Saunders ve Wong (1999), Namkung ve Jang (2007), tarafından yapılan çalışmada sunulan hizmetlerin kalitesi, hizmet sektörü tarafından faaliyet gösteren tüm rekabetçi firmalarda olduğu gibi restoran endüstrisinde de rekabet gücü açısından yapı taşıdır. Çünkü bir hizmet işletmesinde sunulan hizmetlerin kalitesini, müşteri memnuniyetinde ve müşterilerin o işletmeyi tekrar tercih edip etmeyeceklerinde önemli bir faktör olduğu bilinir. Restoran işletmecileri ve gıda sektörü profesyonelleri için müşteri memnuniyeti ve hizmet kalitesini artırmak, rekabetçi bir avantaj elde etmek ve sürdürülebilir başarı sağlamak için kritik bir öneme sahiptir. Müşterilerin restoran deneyimlerini değerlendirmek ve geri bildirimlerini anlamak, işletme performansını iyileştirmenin temel taşlarından biridir. Restoran deneyimi birçok insan için sadece yemek yemekten öte bir anlam taşıyor. Müşteriler, yemeğin kalitesi, servis hızı, atmosfer ve daha

birçok faktör değerlendirirler.

Bilgin ve Şentürk (2017), yaptığı araştırmada müşterilerin tüm diğer hizmetler gibi yiyecek ve içecekleri de satın almadan önce deneme ve satın aldıktan sonra geri dönme fırsatından yoksun olduğunu tespit etmiştir. Bu durumun müşterilerin hizmet sunan firmalara karşı daha temkinli davranmasına neden olduğu, hem kendi deneyimlerinden hem de daha önce hizmet satın almış olan müşterilerin deneyimlerinden daha fazla yararlanmalarına neden olduğu belirlendi. Ancak, online sitelerdeki müşteri yorumları sayısız olabildiği gibi manuel olarak her bir yorumu okumak ve yorumlamak zor olabilir. Bu sebeple, müşteriler tarafından yapılan geri bildirimler üzerinden metin madenciliği tekniklerini kullanarak bu yorumları analiz etmek, işletmelerin müşterilerinin görüşlerini anlamalarına ve işletme stratejilerini geliştirmelerine yardımcı olabilir. Bu analizler, belirli bir restoranın müşteri memnuniyeti düzeyi, yemeklerin kalitesi veya fiyatlandırma stratejileri hakkında fikir edinmenize yardımcı olabilir.

Hussaın, Wang, Jafar, Ilyas, Mustafa ve Jianzhou (2018), çalışmalarında müşterinin ürünlerle ilgili geri bildirimi, bu ürünün müşterinin ihtiyaçlarını ne kadar karşıladığını belirlediğini tespit etmiştir. Bu tez çalışmasının bütününü bir Çin restoranının online sitelerdeki müşteri yorumları oluşturmaktadır. Çin restoranına yapılan müşteri yorumlarının analizi için ilk adım, bu yorumları toplamak ve düzenlemektir. Sosyal medya platformları, yemek değerlendirme siteleri ve rezervasyon uygulamaları gibi kaynaklardan elde edilen yorumlar, bir veri kümesi oluşturmak için kullanıldı. Tüm bunların ışığında çalışmanın uygulama kısmında toplanan veriler veri madenciliği tekniklerinden metin madenciliğinin yöntemi olan LDA(Latent Dirichlet Allocation) ve Bags of Words (BoW) kullanarak müşterilerin yemekleri, hizmeti ve diğer faktörler hakkındaki görüşler analiz edildi. Bu analizlerin yardımıyla, restoranların müşterileri nezdindeki algısını, hizmet kalitesini ve memnuniyet düzeyini anlamayı hedeflendi. Yorumları ana temalara veya konulara göre sınıflandırarak restoranın hangi yönlerinin en çok öne çıktığı belirlendi. Örneğin, yemek kalitesi, hızlı servis veya fiyatlandırma gibi temaları belirlenebilir.

LDA yöntemi, müşteri yorumlarını tematik olarak gruplandırmak ve restoranın hizmet, yemek kalitesi, fiyatlandırma vb. gibi farklı yönlerini anlamak için kullanılacaktır. LDA, konulardaki belirgin kelimeleri ve bu kelimelerin ağırlıklarını belirleyerek konuların içeriğini anlamamıza yardımcı olan tekniktir. Belirlenen konular, metin verilerinin içeriğini anlamak ve yorumları farklı konu gruplarına göre analiz etmek için kullanılır. Bu sayede

müşteri yorumlarının içerdiği temaları belirleyerek, restoranın hizmet kalitesi, müşteri memnuniyeti ve diğer önemli faktörler hakkında bilgi edinilebilir. Metin madenciliği alanında önemli bir role sahip olan Bags Of Words (BoW) yöntemi kullanılarak metin verilerinin sayısal bir forma dönüştürülüp analiz edilmesi hedeflenmiştir. Metin belgelerinin analizinde ve sınıflandırılmasında yaygın olarak kullanılan bir tekniktir. Temel olarak bir belgedeki kelime sıklıklarını ve dağılımlarını temsil etmek için kullanılabilir. Analizler, metin verilerinden anlamlı bilgiler elde etmeyi sağlar. Bu yöntemle, her yorumdaki kelimelerin ve frekanslarının bir matris şeklinde temsil edilmesi sağlanır. Bu matris makine öğrenimi algoritmalarına girdi olarak verilir. LDA ve Bags of Words (BoW) yöntemlerinin analizlerinin tümü ve verilerin görselleştirmelerinde Python programlama dili kullanılarak elde edilmiştir. Çalışmamızın en önemli detaylarından biri ise yorumlardan yola çıkarak puanlamanın tahmin edilmesine dayanmasıdır. Bu çalışmanın sonuçları, restoran işletmecilerine müşteri geri bildirimlerini daha etkili bir şekilde analiz etmeleri ve işletme performansını artırmak için stratejik kararlar almaları konusunda rehberlik edebilir. Ayrıca, makine öğrenimi yöntemlerinin yemek yorumları gibi metin verilerinin analizindeki uygulanabilirliğini de göstererek, akademik araştırmalara ve literatüre katkı sağlayabilir.

Çalışmada Çin restoranını tercih ederek ülkemizin farklı lezzetlere bakış açısını merak ettiğimiz gibi Çin restoranının popülaritesi anlamaya çalışıldı. Çin mutfağı, dünya genelinde büyük bir popülerlik kazanan ve her damak zevkine hitap eden lezzetler sunan bir mutfaktır. Bu büyüleyici mutfağın bir parçası olarak Çin restoranları, farklı tatlar arayan insanlar için önemli bir tercih haline gelmiştir.

Bu çalışmanın sonuçları, restoran işletmecilerine ve online site kullanıcılarına önemli bilgiler sunabilir ve yemek endüstrisinde hizmet kalitesi üzerinde olumlu etkiler yaratabilir. Hangi alanlarda iyileştirmeler yapılması gerektiği ve müşteri beklentilerine nasıl daha iyi cevap verilebileceği konusunda değerli bilgiler sağlar.

Metin madenciliği ve makine öğrenmesi, Çin restoranları ve diğer işletmeler için müşteri geri bildirimlerini anlamak ve işletme performansını iyileştirmek için güçlü araçlar sunar. Bu analizler, müşteri memnuniyetini artırarak restoranların rekabet avantajını sürdürmelerine yardımcı olabilir.

2. VERİ MADENCİLİĞİ

Dijital çağın hızlı gelişimiyle birlikte, veri miktarı ve çeşitliliği de büyük bir hızla artmaktadır. İnternet, sosyal medya, akıllı cihazlar ve sensörler gibi kaynaklar her geçen gün sayısız veri noktasını üretmektedir. Bu veri denizi içinde kaybolmak yerine, bu verileri anlamak ve kullanmak, istenilen sonuçları elde etmek için güçlü bir araç olan veri madenciliği (Data Mining) önemli bir rol oynamaktadır. Veri madenciliğinin amacı, tahminlerde bulunmak veya karar vermek için kullanılabilecek verilerdeki kalıpları ve ilişkileri ortaya çıkarmaktır (Sohrabi, Raeesi Vanani, Nikaein ve Kakavand, 2019). Veri madenciliğini bir süreç olarak düşündüğünüzde, gizli bilgileri keşfetmek ve büyük veri kümelerini anlamlandırmak için veri setlerini istatistiksel ve matematiksel teknikler kullanarak analiz etme ve yorumlama süreci olarak tanımlanabilir. Veri madenciliği, büyük miktarda bilgi depolayan ve geleceği tahmin eden veri tabanlarından, hedeflere ulaşmak için gerekli olan bilgilerin elde edilmesini sağlayan, ilgili verilerin toplanması ve kullanılması sürecidir (Savaş, Topaloğlu ve Yılmaz, 2012). Küreselleşen dünyada toplanan veriye ulaşmanın yanı sıra bu verinin kullanılması ile anlamlı aynı zamanda daha üst seviyede bilgiye erişmek çok daha önemli olmuştur (Can, 2017; Koç, 2023).

Günümüz dünyasında veri dünyası işletmelerin, araştırmacıların ve kurumların en değerli varlıklarından biri haline gelmiştir. Dijitalleşme ve teknolojik ilerlemeler her gün büyük miktarda veri üretilmesine ve depolanmasına olanak tanımaktadır. Veri madenciliği, veri ambarlarından veya tabanlarında elde edilen, farklı şekillerde depolanan, kullanıma hazır bir konumda bulunan çok büyük ölçekteki verilerin içerisindeki gizli kalmış, daha önceden bilinmeyen, kullanışlı ve faydalı olduğu düşünülen anlamlı bilgilerin bulunup keşfedilmesi sürecidir (Öztürk, 2021). Veri setlerini keşfetme, temizleme ve modelleme üzerine kurulu veri madenciliği büyük ve karmaşık veri setlerini incelemek ve bu veriler içindeki örüntüleri, ilişkileri ve anlamlı bilgileri ortaya çıkarmak için kullanılan bir analiz disiplinidir. Bu analizler, işletmelerin, bilim insanlarının, kamu kuruluşlarının ve birçok farklı alandaki profesyonellerin verileri daha iyi anlamalarına, kararlar almalarına ve süreçleri optimize etmelerine yardımcı olabilir. Bu süreç, istatistiksel teknikler, makine öğrenimi algoritmaları ve matematiksel modelleme gibi bir dizi analitik yöntemi içerir.

Veri madenciliği, büyük veri kümelerinden değerli bilgileri çıkarmak için çeşitli alanlarda kullanılan temel bir tekniktir. İş zekası, makine öğrenimi ve yapay zeka alanlarında birçok uygulamaya sahiptir (Çakmak, 2023).

Veri madenciliği, birçok endüstri ve alan için büyük bir değere sahiptir. Avantajlarından bazısından bahsetmek gerekirse bunların en başında verilerdeki gizli örüntüleri keşfederek rekabet avantajı elde etmek, müşteri davranışlarını daha iyi anlayarak pazarlama stratejilerini optimize etmek, verilere dayalı doğru ve hızlı kararlar almak gibi birçok özellik sayabiliriz. Özellikle finans, perakende, sağlık, pazarlama ve üretim gibi birçok sektörde uygulama bulmuştur. Örneğin, müşteri satın alma alışkanlıklarını analiz etmek veya üretim verimliliğini artırmak için kullanılabilir. Buradan yola çıkarak şirketler müşteri davranışlarını anlamak, pazarlama stratejilerini optimize etmek, maliyetleri azaltmak ve gelirleri artırmak için veri madenciliği tekniklerini kullanırlar. Sağlık sektöründe, hastalıkların teşhisinde ve tedavilerin geliştirilmesinde kullanılmaktadır. Bilim dünyasında, büyük veri setlerinden yeni keşifler yapmak için kullanılır. Veri madenciliği, veri üretilebilen pazarlama, alışveriş alışkınlarının tespiti, müşteri kazanımı, reklamların etkisini arttırma, tıp, kredi işlemleri, eğitim, askeriye, Kredi skorlama ulaştırma, telekomünikasyon, tarım, şehircilik, bilgi güvenliği, sosyal ağ analizi gibi birçok alanda sıklıkla kullanılmaktadır (Yalçın, 2019).

Veri madenciliğinde yaygın olarak kullanılan bir teknik ise, küme analizidir (Yan, Yang, Peng ve Ren, 2020). Bu teknik, benzer nesneleri bir araya getirmek için kullanılır. Örneğin, benzer satın alma alışkanlıklarına sahip müşterileri bir arada gruplandırmak için küme analizi kullanılabilmektedir. Elde edilen kümeleme sonuçları doğrultusunda bilgiler, pazarlama kampanyalarını hedeflemek veya müşteri deneyimini kişiselleştirmek için kullanılabilmektedir (Çakmak, 2023). Veri madenciliği güçlü bir araçtır, ancak sınırlamaları da vardır. Veri madenciliğindeki en büyük zorluklardan biri mevcut olan büyük hacimli verilerle uğraşmaktır (Romero ve Ventura, 2020; Çakmak, 2023). Çok büyük veya karmaşık veri kümelerinden anlamlı içgörüler çıkarmak zor olabilir (Hariri, Fredericks, Bowers, 2019).

Günümüzde restoran endüstrisi, müşterilerin geniş bir yelpazede yiyecek seçenekleri arasından seçim yapabildiği oldukça rekabetçi bir alandır. Bu rekabet ortamında, restoranların müşteri memnuniyetini artırmak ve hizmet kalitesini iyileştirmek kritik bir

öneme sahiptir. Veri Madenciliği, büyük veri setlerinden anlamlı bilgileri çıkarmak için güçlü bir araçtır ve müşteri yorumlarının analizi gibi birçok alanda kullanılabilir. Müşterilerin restoran deneyimlerini değerlendirmeleri için kullandıkları müşteri yorumları, restoran işletmecileri için altın niteliğindedir. Ancak bu yorumlar sayısızdır ve her birini manuel olarak incelemek zor olabilir. İşte bu noktada Veri Madenciliği (Data Mining) devreye girer. Veri madenciliğinin uygulanabileceği veri araştırmasının amacı ve kapsamı belirlendikten sonra faydalı bilgiler elde etmek için ilgili veri madenciliği teknikleri izlenmelidir (Gürbüz, 2023). Metin belgelerinden edinilen verilerin analiz edilmesi için veri madenciliği uygulamalarından olan makine öğrenme algoritmaları kullanılabilmektedir. Veri madenciliğinde olduğu gibi metin madenciliği de, bilgi keşfi ve anlam çıkarma süreçlerini içeren benzer disiplinlerdir ve birbirleriyle sıkça entegre edilebilirler. Hem veri madenciliği hem de metin madenciliği, verileri kategori ve sınıflandırma görevlerini içerebilir. Veri madenciliği genellikle sayısal veri setlerinde kategorizasyon yaparken, metin madenciliği metin belgelerini belirli kategorilereve konulara göre sınıflandırabilir. Bu kategoriler daha sonra veri madenciliği analizlerinde kullanılabilir. Veri madenciliği ve metin madenciliği, öznitelik mühendisliği (Veri setlerinden anlamlı bilgi çıkarmak ve bu bilgiyi kullanarak makine öğrenimi modellerini geliştirmek amacıyla veri özniteliklerini tasarlamak, seçmek ve çıkarmakla ilgilenen bir süreçtir. Bu süreç, makine öğrenimi modellerinin eğitimi ve performansını artırmak için veri setlerindeki öznitelikleri düzenleme, dönüştürme veya oluşturma işlemlerini içerir.

Veri madenciliği, sayısal özniteliklerin (Bir nesnenin belirli ve ayırt edici bir özellik ya da karakteristik anlamı) çıkarılması ve işlenmesi ile ilgiliyken metin madenciliği, kelime frekansları, duygu analizi veya konu modelleme gibi metin özniteliklerini çıkarabilir. Tahmin açısından bakacak olunursa hem veri madenciliği hem de metin madenciliği, gelecekteki olayları tahmin etme ve mevcut bilgilerden anlam çıkarma amacını taşır. Bu bağlamda, veri madenciliği modeli sayısal veriler üzerinde tahminler yapabilirken, metin madenciliği ise metin verilerinden yararlanarak gelecekteki eğilimleri veya olayları anlamaya çalışabilir. Veri madenciliği ve metin madenciliği, birbirini tamamlayan disiplinlerdir. Özellikle, büyük şirketlerin ve kuruluşların sahip olduğu geniş veri setleri içindeki metin verilerinin analizi, bu iki alanın entegrasyonu bakımından gerekli kılar. Metin madenciliği, veri madenciliğine daha fazla anlam katarken, veri madenciliği de metin verilerini daha etkili bir şekilde değerlendirmeye yardımcı olabilir. Veri madenciliği ve metin madenciliği, günümüzde bilgi çağında büyük bir rol oynamaktadır. Bu iki

alandaki çalışmalar, işletmelerin rekabet avantajı elde etmelerine, hizmet kalitesini artırmalarına ve karar alma süreçlerini iyileştirmelerine yardımcı olmaktadır. Ancak, bu alanlardaki hızla ilerleyen teknolojik gelişmelerle birlikte, daha derinlemesine ve karmaşık analiz yöntemlerine olan ihtiyaç da artmaktadır. Veri madenciliği yöntemlerini kullanarak verilerin içindeki anlam ve ilişkiyi kurabilmesi, bu verilerin sınıflandırılabilmesi ve aynı zamanda tahmin yapılabilmesi gibi işlevsel amaçlar mümkün olabilmektedir. Veri madenciliği uygulama alanında işlevselliğinin yanında veri bilimi için birçok sektörde kolaylık sağlamaktadır bunlardan bazılarından bahsetmek gerekirse;

Bankacılık alanında müşterilerin gereksinimlerini veri madenciliği yöntemleriyle anlayarak müşterilerine özel profil oluşturabilir. Bu oluşturulan müşteri profili ile doğru hizmet ve ürünü hem maliyet açısından hem de müşterileri ihtiyaçları olmayan konularda rahatsızlık vermemek adına banka prestiji için önemlidir. Veri madenciliği sayesinde dolandırıcılık tespiti, kredi riski değerlendirmeleri ve müşteri ilişkileri yönetimi gibi önemli konularda bilgi elde ederler. Mevcut müşteri davranışları incelenerek çapraz satış yapılabilir ve müşteri sadakati sağlanabilir (Paftalı, 2021).

Sağlık sektörü alanında sağlık kuruluşları, veri madenciliği ile önceki yıllardaki hastaların teşhisleri ve tedavilerinden oluşan büyük veri setlerini analiz ederek hastalık teşhislerini hızlandırabilir, tedavi yöntemlerini iyileştirebilir ve analizler yapabilirler. Bu, sağlık hizmetlerini optimize etme ve hastaların yaşam kalitesini artırma konularında katkı sağlar.

E-Ticaret sektöründe veri madenciliği çok önemli etkiye sahiptir. E-ticaret platformları, kullanıcı davranışlarına dayalı kişiselleştirilmiş öneriler sunmak, web sitesi deneyimini iyileştirmek ve satışları artırmak için veri madenciliğini kullanabilirler. Bu, müşteri sadakatini artırma ve rekabet avantajı elde etme konularında önemlidir. Web sayfalarına yapılan ziyaretlerin kaydedilmesinde kullanılabilir. Örneğin internet bankacılığı veya online alışveriş sitelerinin ekran tasarımları ve müşterilerin beklentilerine ve kullanım alışkanlıklarına göre analiz edildikten sonra tasarlanması sağlanabilir. Müşterilerin almış oldukları ürün çeşitleri üzerinden analiz yapılarak yan ürün satın alma alışkanlıkları sonucunda müşteri için en uygun ürün yelpazesi sunulabilir.

Telekomünikasyon sektörüne baktığımızda şirketlerin, müşteri şikâyetlerini analiz etme, ağ performansını izleme ve abone taleplerini anlama adına veri madenciliği kullanabilirler.

Bu, hizmet kalitesini artırma ve müşteri memnuniyetini sağlama konusunda önemlidir. Ayrıca veri madenciliği sayesinde güvenlik ağı oluşturularak müşterilerin dolandırıcılıktan korunması sağlanarak şirket prestiji sağlanabilir.

Veri madenciliği günümüzde tüm sektörlerin işleyişinde önemli rol oynamaktadır. Son olarak satış ve pazarlama alanından bahsedecek olursak stok tedarik yönetiminde mevcut veriler üzerinden geleceğin tahmin edilmesi kullanılabilir (Paftalı, 2021).

Pazarlama ve satışa alanında müşteri verilerini analiz ederek benzer özelliklere sahip müşteri gruplarını belirlemeyi sağlar. Bu sayede şirketler, hedef kitlelerine daha özgü ve etkili pazarlama stratejileri oluşturabilirler. Müşteri davranışlarından elde edilen verilerle, kişiselleştirilmiş pazarlama stratejileri geliştirilebilir. Bu, müşterilere daha önce ilgilendikleri ürün veya hizmetleri önerme ve özel teklifler sunma konusunda şirketlere avantaj sağlar. Veri madenciliği, geçmiş satış verilerini analiz ederek gelecekteki talepleri tahmin etmeyi sağlar. Bu, stok yönetimini optimize etme, talep karşılamada esneklik sağlama ve müşteri memnuniyetini artırma konularında önemli rol oynar. Pazarlama stratejilerini ve fiyatlandırmayı belirlemede veri madenciliği, rakip firmaların faaliyetlerini analiz etmeye yardımcı olur. Bu, rekabet avantajı elde etme ve sektördeki değişikliklere daha hızlı adapte olma konularında şirketlere yardımcı olabilir. Sosyal medya ve müşteri yorumları gibi metin verilerini analiz eden duygu analizi, müşteri memnuniyetini değerlendirmek ve ürün/hizmet algısını anlamak için kullanılabilir. Bu sayede şirketler, müşteri geri bildirimlerine daha etkili bir şekilde yanıt verebilirler. Veri madenciliği, fiyatlandırma stratejilerini belirlemede kullanılabilir. Müşterilerin demografik bilgileri üzerinden profilleri oluşturularak pazarlama aktivitelerindeki içerik ve yöntem farklılaştırılabilir (Paftalı, 2021).

Müşteri talepleri, rekabet durumu ve pazar koşulları analiz edilerek optimâl fiyatlandırma stratejileri oluşturulabilir. Şirketler, müşteri sadakatini artırmak ve müşteri kayıplarını önlemek adına etkili stratejiler geliştirebilirler. Müşteri iletişimi ve etkileşimleri analiz edilerek, hangi pazarlama kanallarının ve stratejilerinin daha etkili olduğu belirlenebilir. Bu, bütçenin doğru kanallara yönlendirilmesine yardımcı olabilir. Müşteri geri bildirimleri ve talepleri üzerinden yapılan veri madenciliği analizleri, yeni ürün veya hizmetlerin geliştirilmesinde şirketlere rehberlik edebilir. Veri madenciliği, risk analizi yaparak, özellikle finansal anlamda, belirli pazar koşullarına veya ekonomik değişikliklere karşı

şirketleri bilgilendirir. Satış ve pazarlama sektörlerinde veri madenciliği kullanımı, rekabet avantajı elde etme, müşteri memnuniyetini artırma ve daha etkili stratejiler geliştirme konularında önemli bir rol oynar. Şirketler, bu teknikleri kullanarak büyük veri setlerinden anlam çıkarma ve bu anlamı stratejik avantajlara dönüştürme konusunda önemli kazanımlar elde edebilirler. Online satış siteleri, restoranlara gibi platformlara gelen müşterilerle ilgili veri madenciliği uygulamaları, satış ve pazarlama stratejilerini iyileştirmek, müşteri memnuniyetini artırmak ve işletme performansını optimize etmek için önemli bir rol oynayabilir. Sosyal medya platformlarındaki etkileşimleri analiz ederek, müşteri algısını anlayabilir ve sosyal medya pazarlama stratejileri geliştirebilir.

2.1. Metin Madenciliği

Metin madenciliğinin temelleri, bilgisayarların dil işleme yeteneklerinin gelişmeye başladığı 1950'li yıllara kadar gitmektedir. Ancak, bu dönemde dil işleme ve metin madenciliği konuları hala gelişme aşamasındaydı. 2010'lardan günümüze kadar ise metin madenciliğindeki en büyük gelişmelerden biri, makine öğrenmesi ve derin öğrenme tekniklerinin kullanılabilir hale gelmesidir. Bu teknikler, büyük veri setlerindeki karmaşık desenleri daha iyi anlamak ve daha etkili tahminlerde bulunmak için kullanılır. Gelişen teknoloji ve artan veri hacmi, metin madenciliği alanındaki araştırmaları ve uygulamaları şekillendirmeye devam etmektedir. Tanım olarak bahsedilirse metin madenciliği, doğal dil işleme teknikleriyle birlikte çeşitli yöntem ve teknikler kullanarak büyük hacimli verilerden bilgi çıkarma işlemidir denilebilir. Aynı zamanda doğal dil işleme ve veri madenciliği tekniklerini birleştirerek büyük metin veri kitlelerinden anlamlı bilgiler çıkarmak amacıyla yapılan çalışmaları kapsar. Bu yöntem, metinleri analiz ederek desenleri, eğilimleri ve ilişkileri keşfetmeyi mümkün kılar.

Dijital çağın hızlı gelişimi, internet kullanımının yaygınlaşması ve sosyal medyanın etkisiyle birlikte günlük hayatta milyonlarca insan tarafından üretilen büyük miktardaki metin verisi, bilgi çağının en değerli varlıklarından biri haline gelmesiyle birlikte büyük veri denizinde kaybolmak yerine, bu metin verilerini anlamak, analiz etmek ve değerli bilgiler çıkarmak için kullanılan birçok yöntem ve araç geliştirilmiştir. Bu noktada karşımıza çıkan önemli bir disiplin, metin madenciliği ya da bilgi çıkarma olarak da adlandırılan bu alan, metin verilerindeki desenleri, ilişkileri ve anlamlı bilgileri ortaya çıkarmak amacıyla kullanılan güçlü bir analiz aracıdır. Metin verilerini işlemek ve

anlamlandırmak zor iken bu görevi metin madenciliği üstlenmiştir. Başka bir deyişle doğal dil işleme metin madenciliği araştırmalarında sıkça kullanılan, insana özgü doğal dillerin bilgisayara tanıtılması ve işlenerek doğru bir şekilde analiz edilmesini sağlayan bir mühendislik alanıdır (Başkaya ve Aydın, 2017).

Verilerden anlamlı bilgi çıkarmak için kullanılan veri madenciliği dalıdır. Pazarlama, sağlık, eğitim gibi birçok alanda araştırmalarda rol oynar. Bu teknik, metinlerdeki kalıpları, duygu ve düşünceleri ortaya çıkarmaya yardımcı olur. Metin madenciliği, geniş metin veri setlerinde anlam çıkarma işlemiyle uğraşır. Bu veriler, sosyal medya yazıları, müşteri yorumları, makaleler, haberler ve daha birçok kaynaktan gelir. Temel amacı, metin verilerinde gizli kalmış olan bilgileri keşfetmek, örüntüleri analiz etmek ve bu verilerden anlam çıkarmaktır. Bu hedeflere ulaşmak için veri madenciliği ve görselleştirme gibi teknikler, metin madenciliği araştırmaları, bilgi erişimi, sözcüksel analiz, örüntü tanıma, kelime sıklığı dağıtımı, etiketleme ve bilgi çıkarma için kullanılır (Sönmez, 2017). Günümüzde çoğu kurum ve kuruluş büyük miktarlarda veriyi veri ambarlarında ve bulut platformlarında toplamakta ve saklamaktadır. Bu veriler, birden fazla kaynaktan yeni veriler geldikçe büyümeye devam etmektedir. İşletmelerin ve kuruluşların büyük miktarlardaki metin verilerini geleneksel araçları kullanarak depolaması, işlemesi ve analiz etmesi oldukça zor olduğundan günümüzde yazılımların gelişmesiyle bu sorunlar çözülmüştür (Çelik, 2020).

Farklı araştırmacı görüşlerine göre metin madenciliği veri madenciliğinin bir uzantısı olarak görülmektedir (Irfan ve diğerleri, 2015). Bu görüşün temelinde yapılandırılmamış olan metin dosyalarının sayısallaştırıldıktan sonra analiz sürecinde veri madenciliği ile aynı aşamaları izlemesi yer almaktadır. Veri madenciliği ve metin madenciliği arasındaki güçlü ilişki nedeniyle gelişim süreçleri de birbirlerine paralellik göstermektedir. Veri madenciliği pek çok yöntem ve tekniği ham veri üzerine uygulayarak anlamlı bilgiler edinilmesini sağlamaktadır. Veri madenciliğinde tercih edilen veri tabanları düzgün ve yapısal verileri içermektedir (Öztürk, 2021). Metin madenciliğinin ne yaptığına bakılırsa, en basit düzey olarak, yapılandırılmamış metin belgelerini sayısallaştırıp, veri madenciliği araç ve tekniklerini kullanarak onlardan anlamlı bilgiler çıkardığı görülür. Başka bir deyişle, metin madenciliği en genel haliyle doğal dilde yazılmış metinler arasında aynı konudaki belgeleri ve birbiriyle ilgili olan belgeleri arar ve bulunan belgeleri sıralar.

Metin madenciliği, işletme belgeleri, müşteri incelemeleri, web sayfaları ve XML dosyaları gibi yapılandırılmamış verilerden önceden bilinmeyen ve potansiyel olarak yararlı bilgilerin keşfedilmesi sürecidir. Elde edilen bilgilere dayanarak, analiz edilen metin kaynağında açıkça görülemeyen ilişkiler, hipotezler veya eğilimlerin olduğu varsayılır (Mecca, Raunıch ve Pappalardo, 2007).

Metin madenciliği, birçok farklı uygulama alanında kullanılabilir. Örneğin, pazarlama ve reklamcılık alanında müşteri geri bildirimlerinin analizi için, sosyal medya verilerindeki eğilimlerin ve kullanıcı görüşlerinin belirlenmesi için, sağlık sektöründe hastaların semptomlarını ve tedavi sonuçlarını değerlendirmek için, finansal sektörde piyasa trendlerini belirlemek için ve daha birçok alanda kullanılabilir. Daha ileri düzeyde olarak düşünülürse belgeleri özetlemek ve bilgi çıkarmak için metin madenciliği teknikleri kullanılır. Metin madenciliği, büyük miktarda metin verisinden anlamlı bilgilerin çıkarılması için kullanılan bir veri madenciliği yöntemidir.

Bu yöntem pazarlama, sosyal medya analizi, tıp ve diğer birçok alanda kullanılabilir ve bu alanlarda bilgi çıkarmak ve karar vermek için faydalı bir araçtır. Hastalık teşhisi veya tedavi seçeneklerinin belirlenmesi gibi, büyük miktarda tedavi literatürünün analiz edilmesi yoluyla değerli bilgiler elde edilebilir (Nilashi ve diğerleri, 2021). Metin madenciliğinin, metin verilerinin çoğunluğunun işletmelerden elde edildiği alanlarda açıkça görülmektedir. Örneğin müşteri şikayetleri ve memnuniyet metinleri aracılığıyla sağlanan anlamlı bilgiler, şirketlere ürün geliştirme, kusur takibi, garanti süresi gibi konular hakkında bilgi sağlayabilir (Delen ve Crossland, 2008). Metin madenciliği aynı zamanda bilgi çekme ile birleştirilerek belirli kategorilerdeki bilgilerin çıkartılmasına odaklanır. Örneğin, bir haber makalesinden belirli konuların veya anahtar kelimelerin çıkartılmasını sağlar. Bu, büyük veri setlerindeki önemli bilgilerin belirlenmesini sağlar.

Metin madenciliği teknikleri arasında doğal dil işleme, makine öğrenmesi ve istatistiksel analiz yer alır. Bu teknikler sayesinde, büyük veri kümelerindeki metinlerdeki önemli bilgileri belirlemek ve anlamak mümkün hale gelir (Irfan ve diğerleri 2015). Yorumları ve metin verilerini analiz etmek, daha bilinçli ve veri odaklı kararlar almanıza yardımcı olabilir. İnsanların düşünce ve duygularını anlamanın geleneksel yolları olan anket ve diğer araştırma yöntemlerinin yanı sıra, metin madenciliğinin temel görevi basılı yazılı metinleri işleyerek bu metinleri geniş bir veri tabanı olarak analiz edip sayısal değerlere

dönüştürmektir (Atan, 2016). Metin madenciliği, başlıca yemek sektöründe olduğu gibi birçok sektörde müşteri deneyimini ve işletmenin performansını iyileştirmek için güçlü bir araçtır. Bu araştırma, Çin restoranına yapılan online müşteri yorumlarını metin madenciliği analizi ile değerlendirerek restoranın performansını daha iyi anlamayı amaçlamaktadır. Metin madenciliği tekniklerinden olan Gizli Dirichlet Ayırımı (LDA) (Latent Dirichlet Allocation) makine öğrenimi ve doğal dil işleme alanlarında kullanılan bir istatistiksel modeldir. Bu çalışmamızda LDA yöntemini kullanarak metin belgeleri arasındaki benzerlikleri ve ilişkileri ortaya çıkararak ve bu bilgiyi kullanıp belgeleri gruplamak ve sınıflandırmak amacıyla müşteri yorumlarını daha iyi anlamamızı ve yorumlamamızı sağlayacaktır. Metin madenciliği, bir işletmenin veya ürünün rakipleriyle karşılaştırılması için kullanılabilir. Bu, pazar payını artırmak ve rekabet avantajı elde etmek için stratejik bilgiler sağlar.

Büyük metin veri setleri üzerinde yapılan analizlerle gelecekteki trendleri ve gelişmeleri öngörmekte kullanılabilir. Bu alan, ilerleyen süreçlerde daha fazla işletme ve sektör tarafından benimsenmektedir, çünkü veri tabanlı kararlar almak ve değerli bilgilerden yararlanmak, rekabet avantajı sağlamak adına için kritik öneme sahiptir.

2.1.1. Metin madenciliği çalışma alanları

Metin madenciliği, geniş bir uygulama yelpazesi bulunan çok disiplinli bir alandır.

Bilgi çıkarma (Information Extraction)

Bilgi çıkarımı, yapılandırılmış bilgilerin (varlıklar, varlıklar arasındaki ilişkiler ve varlıklar arasındaki nitelikler gibi) yapılandırılmamış kaynaklardan otomatik olarak çıkarılmasını ifade eder (İçöz, 2021). Bu teknik, metin belgelerinden veya verilerinden belirli yapısal veya yapısal olmayan bilgileri çıkarmayı amaçlar. Bilgi çıkarma, metin verilerindeki önemli bilgileri tanımlamak ve düzenlemek için kullanılır. Araştırmacılar, literatürde var olan metinsel kalıpları tespit ederek, metinleri karşılaştırarak ve metinleri birbiriyle ilişkilendirerek kolaylıkla analiz edebilirler (Richards, Tudhope ve Vlachidis, 2015).

Metin belgelerinden veya metin verilerinden özel bilgileri belirli bir yapı içinde çıkarmak için kullanılan bu teknikte belirli konular, tarihler, yerler, kişiler, organizasyonlar, ürünler

veya olaylar gibi özel bilgileri tanımlamak için kullanılabilir. Bilgi çıkarma, birçok farklı uygulama alanında kullanılır. Örneğin, finansal raporlardan şirket performansı bilgisi çıkarma, medya izleme ve analizi hakkında bilgi çıkarma gibi alanlarda kullanılır. Bilgi çıkarma, büyük metin veri kümelerini otomatik olarak işlemek ve içerdikleri bilgilere ulaşmak için güçlü bir araçtır. Özellikle büyük kurumsal veriler veya web'den toplanan veriler gibi geniş veri kümeleri üzerinde kullanılarak önemli bilgilerin hızlı ve etkili bir şekilde elde edilmesini sağlar. Örneğin, restoranın adı, konumu, fiyat aralığı, yemek çeşitleri, servis kalitesi gibi belirli bilgileri hedeflenebilir. Bilgi çıkarma için kurallar veya şablonlar oluşturulur. Örneğin, restoranın adı genellikle "Restoran Adı: ABC" biçiminde bir cümle içinde bulunabilir. Bu cümleyi tanımlayan bir kural oluşturulur. Oluşturduğunuz kuralları veya şablonları kullanarak bilgi çıkarma işlemine başlanır. Metin verilerini işleyerek, bu tür kurallara uyan bilgileri çıkarılabilir. Örneğin internetin kaynaklarından olan veri tabanları, internetteki belgeler ve taranmış metinler verilerin kaynağını oluşturabilir (İçöz, 2021).

Doğal dil işleme (natural language processing)

Yaygın olarak NLP (Doğal Dil İşleme) olarak bilinen yapay zekâ ve dilbilimin bir alt kategorisi olan Türkçe ve İngilizce gibi doğal dillerin işlenmesini ve kullanımını inceleyen bir disiplindir (İçöz, 2021). Doğal dil işleme (NLP), metin madenciliğinin temelini oluşturan ve metin verilerini anlama, işleme ve yorumlama amacı taşıyan uygulama alanıdır. NLP, insanların doğal dilde iletişim kurduğu metin veya konuşma gibi metin tabanlı verileri incelemek ve analiz etmek için kullanılır. Başka bir deyişle doğal dil işleme metin madenciliği araştırmalarında sıkça kullanılan, insana özgü doğal dillerin bilgisayara tanıtılması ve işlenerek doğru bir şekilde analiz edilmesini sağlayan bir mühendislik alanıdır (Başkaya ve Aydın 2017). Karmaşık bir süreç olan doğal dil işlemede sadece metin içerisindeki sözcükleri bulmak ve sınıflandırmak yeterli değildir. Sözcüklerin aldığı eklerin ve cümle içerisindeki taşıdığı anlamlarında bilinmesi gerekmektedir. Özellikle doğal dil işlemede sözcüklerin hangi anlam bağlamında kullanıldığını anlamada yapay zekâ sıkça tercih edilmektedir (Ergün, 2016).

Doğal dil işlemede iki farklı yaklaşım bulunmaktadır. Bunlardan birincisi metnin yazıldığı dilin dilbilgisi özelliklerine dikkat edilerek, metin bir bütün olarak analiz edilmektedir. İkincisi ise kelime çantası yaklaşımıdır. Bu yaklaşımda metinler parçalanarak sözcükler

bazında incelenmektedir ve dilbilgisi yapısı dikkate alınmamaktadır. Metinler cümlelere, cümleler sözcüklere, sözcükler ise köklerine ayrılarak ve metin içerisindeki köklerin frekansı dikkate alınarak analiz yapılmaktadır (Çeliksu, 2017).

Metinleri küçük parçalara, yani "belirteçlere" ayırma işlemine tokenizasyon (tokenization) denilmektedir. İşlemin amacı ise kelime veya cümle seviyesinde işlemler yapabilmek için metni temel birimlere bölmektir. Örneğin "Günaydın, nasılsınız?" cümlesi belirteçlere ayrıldığında: ["Günaydın", ",", "nasılsınız", "?"] şeklinde yapılmaktadır. Kelimelerin temel yapıları ve anlamlarını anlamak için köklerine ve eklerine ayrıştırılması işlemi ise biçimbilimsel analiz (Morphological Analysis) ile yapılmaktadır. Örneğin, "Geliyorduk" kelimesi ayrıştırıldığında: ["gel" (kök), "-iyor" (ek), "-duk" (ek)] olarak yapılır. Kelime Türü Etiketleme ile kelimelerin dil bilgisel türlerine (örneğin, isim, sıfat, zarf) etiket atanması için cümledeki kelimelerin gramatik rollerini belirliyor. Örneğin "İsot Pencereden dışarı atladı" cümlesi etiketlendiğinde: [("İsot", "İsim"), ("Pencereden", "İsim"), ("dışarı", "Zarf"), ("atl", "Fiil"), ("dı", "Zaman")] şeklinde ayrıntılı inceleme yapar. Sözdizimi ile cümlelerin gramatik yapısını anlama süreci başlar. Kelimeler arasındaki bağlantıları ve cümle yapısını çözümleme yapılır. Örneğin "Ben deftere yazdım" cümlesi ayrıştırıldığında: [("Ben", "Altöz"), ("deftere", "Nesne"), ("yazdım", "Fiil")] olarak yapılır. Anlambilimi oluşturularak kelimelerin ve cümlelerin anlamını çıkarma süreci gerçekleştirilir. Kelimeler arasındaki ilişkiler ve anlam belirlenir. Örneğin, "Elime sıcak su dökülürse yanar" cümlesi anlambilim açısından analiz edildiğinde, eline sıcak su döküldüğünde yanacağı anlaşılabilir.

Duygu Analizi ise metinlerdeki duygusal tonu anlama süreci aşamasında işlevseldir. Bir metnin olumlu, olumsuz veya tarafsız duygular içerip içermediğini belirler. Örneğin, Bir müşteri yorumunun olumlu veya olumsuz olup olmadığını belirlemek için önemlidir. Konu Modelleme metin içindeki ana temaları belirlerken konulara veya kategorilere ait metin gruplarını oluşturur. Örneğin, haber veya bilimsel makalelerinin içeriğini analiz ederek belirli konulardaki grupları tanımlamayabilir. Metin sınıflandırma ve tahmin işlemi metinleri belirli kategorilere sınıflandırma veya belirli bir sonuca ulaşma sürecinde metni belirli bir kategoriye atama veya gelecekteki bir durumu tahmin etmek için gereklidir. Örneğin, bir müşteri yorumunu "olumlu" veya "olumsuz" olarak sınıflandırma yapabilir.

NLP'nin uygulama amacında bazı işlevler vardır. Bunlarında başında gelen NLP'nin metin verilerini anlama işlevi metinlerdeki dil bilgisel yapısı, dil öğeleri (kelimeler, cümleler, paragraflar) ve anlamı çıkarma işlemini içerir. Örneğin, bir metindeki anahtar kelimeleri veya cümlenin yapısını anlama işlemi NLP'nin bir parçasıdır. Anlamsal düzeyde, cümlenin anlamı ile ilgilidir. Makinelerin doğru işlemleri yapabilmesi için cümlelerin anlamlarını doğru tanıması gerekmektedir (Şeker, 2015b).

Dil modelleme işleminde NLP, dilin yapısını ve özelliklerini modelleme amacı taşır. Bu, dilin kuralları, kelime öğeleri ve gramatik yapıları hakkında bilgi içerir. Bu modele dayalı olarak, metinlerin düzgün bir şekilde işlenebilmesi için analizler yapılabilir. Kelimelerde işleme yapılarak metin verilerindeki kelimeleri tanıma, ayrıştırma ve işleme yeteneği sunar. Kelimelerin köklerini (stemming), lemmatization, sıklığını ve anlamını çıkarma gibi işlemleri içerir. Kelimelerin bu farklı özelliklerinin keşfi de sözlükbilim düzeyinde gerçekleşir. Kelime bilgisi düzeyindeki bir diğer örnek ise yanlış yazılan kelimeleri tanımak ve düzeltmektir (Değer, 2017). Kelimelerin işleme aşamasının ardından cümlenin ayrıştırılması NLP işleminde bir cümlenin öznenin, nesnenin ve fiilin tanımlanması gibi dilbilgisi analizlerini içerir. Kelimeler arasındaki ilişkiler incelenir ve cümle yapısının kurallara uygun olarak oluşup oluşmadığı kontrol edilir (Değer, 2017).

Genel bir yapı tamamlandıktan sonra duygu analizi işlemi yapılarak metinlerin içeriğinden pozitif, negatif veya nötr duygu tonunu belirlenir. Bu, metinlerin hissi analizi veya duygu analizi olarak adlandırılır ve özellikle sosyal medya verilerinin analizinde yaygın olarak kullanılır. Duygu analizi sonrası konu modellemesinde metinlerin içeriğinden konuları veya temaları belirlemeye yardımcı olan konu modelleme tekniklerini içerir. Örneğin, Latent Dirichlet Allocation (LDA) gibi algoritmalarla metinlerdeki gizli temaları keşfetmek mümkündür. Konu modellemesi ardından metin sınıflandırmada metin verilerini belirli kategorilere veya sınıflara atama işlevi sunar. Bu spam filtreleri, duygu analizi, konu kategorizasyonu ve daha birçok uygulamada kullanılır. Bu, uzun metinleri daha kısa bir özetleme dönüştürme veya anahtar bilgileri çıkarma işlemini içerir. Bu işlem aşamaları NLP'nin metin madenciliği uygulamalarında metin verilerini anlamak, analiz etmek ve bilgi çıkarmak için temel araçtır. Bu yöntemler, metindeki kelimelerin frekansını ve istatistiksel özelliklerini analiz eder. Kelime bulutları ve kelime dağılımı grafikleri oluşturulabilir. NLP'nin uygulama alanları oldukça geniştir. Bu teknoloji, metin tabanlı verilerin analizi, dil çevirisi, otomatik yanıt sistemleri, duygu analizi, reklamcılık, e-posta

sınıflandırma, sağlık analizi, eğitim ve daha birçok alanda kullanılır. İnsan-makine iletişimi, veri analizi ve karar verme süreçlerini geliştirmek için büyük bir potansiyele sahiptir. Her gün e-postalardan, kısa mesajlardan, tweet'lerden, geri bildirimlerden, sosyal medyadan, ürün/hizmet incelemelerinden, bloglardan, makalelerden, belgelerden vb. milyonlarca yapılandırılmamış gömülü metin ve dil verisi üretiliyor (Liu, Singh ve Srinivasan, 2016).

Uygulama alanlarının en başında metin madenciliği (Text Mining) gelmektedir. Metin verilerinde anlam çıkarmak için NLP teknikleri kullanılır. Duygu analizi, konu modelleme ve bilgi çıkarma gibi konularda önemlidir. Konuşma Tanıma (Speech Recognition) alanında konuşma verilerini yazılı metne dönüştürme sürecinde başarılıdır. Sesli asistanlar, konuşma tabanlı arama ve komut algılama gibi uygulamalarda kullanılır. Sıkça kullandığımız uygulamalarda dil çevirisi (Language Translation)'de kullanılabilir. İki farklı dil arasında metin çevirisi yapma işlemidir. Çevrim içi çeviri araçları bu alanda kullanılan örneklerdir. Doğal Dil İşleme, günümüzde pek çok sektörde kullanılan ve sürekli olarak gelişen bir alan olarak karşımıza çıkmaktadır. Bu teknoloji, bilgisayar sistemlerini dil anlama ve kullanma konusunda daha yetenekli hale getirerek, insanlarla daha etkili bir şekilde iletişim kurmalarına olanak tanır.

Adlandırılmış varlık tanıma (named entity recognition)

Named Entity Recognition (NER), bir metin içindeki belirli varlıkları (named entities) tanıma ve sınıflandırmayı amaçlayan bir doğal dil işleme (NLP) görevidir. Bu varlıklar genellikle belirli bir anlam taşıyan ve özel bir bilgi içeren kelimeler veya kelime gruplarıdır. Varlık adı tanıma, doğal dil işlemenin çalışma alanlarından biri olan ve bilgi çıkarmanın bir alt dalı olarak metinlerde bulunan varlık adlarının tanımlanması ve sınıflandırılması ile ilgilidir (Nadeau ve Sekine, 2007). Kural tabanlı yaklaşımlar, dilin gramer özelliklerine dayalı kuralları kullanarak varlık adı tanımayı gerçekleştirirken istatistiksel yaklaşımlar, makine öğrenimi teknikleri kullanılarak eğitilen istatistiksel modelleri kullanır (Eken, 2015). NER, metin içindeki bu tür varlıkları tanımlayarak, metni daha anlamlı ve yapısallaştırılmış hale getirir. Varlık tanımanın ne olduğunu bilmek için, varlık tespit modeli adı verilen bir modelin, birimi oluşturan kelimeyi veya kelime dizisini (örneğin "İzmir şehri") tanımlayabilmesi ve hangi varlık sınıfına ait olduğunu bilmesi gerekir (Nasiboğlu ve Gencer, 2023). Bir metinde geçen isim, yer, tarih gibi bilgileri

belirleyerek bu bilgileri daha geniş bir bağlam içinde anlamamıza yardımcı olan temel bir tekniktir. Müşteri geri bildirimleri, ürünler hakkındaki düşünceleri düzenleme ve çok fazla tekrarlayan durumları saptamak amacıyla kullanılabilir. Örneğin, adlandırılmış varlık tanımlayıcı kullanılarak olumsuz müşteri geri bildirimlerinde en sık bahsedilen lokasyon ve illerin belirlenmesi ile müşterinin belirli bir ofis veya şubeye yönlendirilmesi sağlanabilir (Han, Sun, Cong, Zhao, Ji ve Phan, 2017). Başka bir örnek ise Netflix'te çok fazla komedi izleniyorsa komedi varlıkları olarak kategorize edilmiş daha fazla öneri görülebilir veya YouTube, izlenen video türlerine göre benzer türde videolar önerir ve bu durumda video türü, adlandırılmış varlık tanıma olarak kabul edilir (Bowden, Wu, Oraby, Misra ve Walker, 2018). NER'in metin içindeki önemli bilgileri tanıma ve sınıflandırma becerisi nedeniyle oldukça değerlidir. NER, bilgisayarların büyük metin verilerini daha etkili bir şekilde işlemelerine ve anlamalarına yardımcı olan temel bir NLP bileşenidir. Sosyal medya platformlarındaki büyük veri setlerini analiz etmek için NER kullanılabilir. Kullanıcı adları, etiketler, marka adları gibi varlıkların tanınması, sosyal medya kampanyalarının etkisini değerlendirmede önemlidir.

Duygu analizi (sentiment analysis)

Genellikle doğal dil işleme ve metin madenciliği tekniklerini kullanarak bir metin belgesinin duygusal tonunu anlama ve sınıflandırma işlemidir. Bu analiz, metinlerin içerdiği duygusal öğeleri belirleyerek genellikle pozitif, negatif veya nötr gibi duygu kategorilerine sınıflandırmayı amaçlar. Duygu analizi, otomatik araçlar kullanarak metinde ifade edilen görüşler, tutumlar ve duygular gibi öznel bilgileri tanımlamayı amaçlamaktadır ve son yıllarda doğal dil işlemeye olan ilginin hızlı bir şekilde artmasına yol açmıştır (Fadel, 2020). Duygu analizi, doğal dil işleme (NLP), hesaplamalı dilbilim ve metin madenciliğindeki problemlerle ilgilidir (Pang, Lee ve Vathyanathan, 2002). Duygu analizi, metindeki kelime ve kelime gruplarını analiz ederek metnin içerdiği duyguyu ortaya çıkarır (Başkaya, 2017). Duygu sınıflandırma ise bir konu hakkında yazılmış olan ifadelerin analiz edilmesiyle birlikte yazarın sahip olduğu duyguyu genellikle olumlu, olumsuz veya tarafsız gibi kategorilere sınıflandırmayı amaçlayan ve duygu analizinin alt dalı olarak ele alınan bir çalışma alanıdır (Onan, Korukoğlu ve Bulut, 2017). Duygu analizi, metin verilerinin anlaşılmasını ve duygusal içeriğin etkili bir şekilde yönetilmesini sağlayarak birçok sektörde önemli bir araç haline gelmiştir. İnsan duygularını anlamak, daha etkili iletişim, hızlı tepki ve daha iyi kararlar almak için temel bir unsurdur.

Günümüzde duygu analizi, ticari işletmeler, politikacılar, medya kuruluşları, kamu kurumları ve güvenlik kuruluşları gibi birçok alanda olduğu gibi piyasa analizi, bilimsel, tıbbi araştırmalarda, suç tespitinde, siyasi görüş belirlemede ve sosyolojik araştırmalarda da yaygın olarak kullanılmaktadır (Nasukawa ve Yi, 2003). Duygu analizi, metin tabanlı büyük veri setlerini anlama ve yönetme konusunda önemli bir araç olup, birçok sektörde karar alma süreçlerine katkı sağlar. İnsanların duygusal eğilimlerini anlamak, daha etkili iletişim, müşteri memnuniyeti ve marka itibarı yönetimi için kritik bir öneme sahiptir.

2.1.2. Metin madenciliği aşamaları

Önceden belirlenen hedeflere uygun analiz ve modelleme yöntemlerini kullanarak verileri analiz eder ve anlam çıkarılır. Metinleri anlamak, kategorize etmek ve ilişkileri belirlemek için NLP (Doğal Dil İşleme) algoritmaları ve teknikleri kullanılır. Büyük veri kitlelerinde saklı olan anlamlı bilgileri bulmak için veri madenciliği yöntemleri ve modelleri kullanılır**.** Metin madenciliği aşamalarında verileri ön işleme, özetleme, kategorizasyon, kümeleme ve bilgi görselleştirme gibi teknolojiler kullanılır.

Veri ön işleme (data preprocessing)

Metin verilerini toplamak için farklı kaynakları kullanırız. Örneğin web sayfaları, akademik çalışmalar, e-postalar, raporlar, sosyal medya platformları veya kurumsal veri tabanları gibi herhangi bir metin içeriği olabilir. Verilerin ön işleme aşaması metin madenciliği sürecindeki en uğraştırıcı ve zaman alan aşama denilebilir. Bu aşamada ilk olarak veri temizleme ve ayıklama işlemleri yapılır. Toplanan verileri temizler, düzenler ve özelleştirilir. Bu adımda, gereksiz karakterleri kaldırır, metni küçük harfe dönüştürür ve metinleri parçalara ayırır. Veri temizleme işlemi yoluyla verilerdeki gürültü, düzensizlikler ve tutarsızlıklar giderilerek veri kalitesi artırılır (İnan, 2015). Veriler eksikse veri kümesinden kaldırmak, sabit bir değer ayarlamak veya ortalama atama gibi bir süreç uygulanabilir (Değer, 2017). Analizde anlamlı olmayabilecek düşük frekansta görülen kelimeler çıkarılır. Sıkça kullanılan ama anlamsal olarak pek bir bilgi taşımayan kelimelerin (örneğin, "ve," "veya," "bir") çıkartılır. Metin madenciliği uygulamalarında genel olarak köke indirgeme (lemmatization) ve etkisi olmayan kelimeleri çıkarma (stop word removal) işlemleri yapmak mümkündür (Kurt, Güldal ve Batmaz, 2022).

Etkisiz Kelimeleri (Stop Words) Kaldırma: "Stop words" veya etkisiz kelimeler, dilin yapı taşları olmasına rağmen genellikle analiz sırasında çok fazla bilgi içermeyen yaygın kelimelerdir. Etkisiz kelimelere, "ve", "veya", "ama", "bu", "şu", "gibi" kelimeler örnek verilebilir. Bu kelimeler genellikle analiz sırasında göz ardı edilir veya kaldırılır çünkü anahtar bilgi içerme olasılıkları düşüktür.

Normalizasyon: Metin verilerindeki kelime biçimleri genellikle farklı olabilir. Normalizasyon adımı, farklı biçimlere sahip kelimeleri benzer hale getirerek, örneğin küçük harfe dönüştürme gibi işlemleri içerir. Bu, aynı anlama gelen kelimelerin aynı şekilde temsil edilmesine yardımcı olur. Sayılar genellikle metin madenciliği için anlam ifade etmeyen ögelerdir. Bu nedenle, sayılar genellikle belirli bir sembolle değiştirilir. Tarih ve saat bilgileri benzer şekilde sembolik temsil ile değiştirilebilir.

Kök Bulma (Stemming): Kelimelerin köklerini (stem) bulma sürecidir. Bu adım, benzer kelimelerin aynı kökü paylaşmasını sağlar. Örneğin, "koşuyorum" ve "koşarken" gibi kelimelerin kökü "koş-u" dur. Bu, analizin tutarlı ve etkili olmasına yardımcı olur. Benzer anlam taşıyan kelimelerin aynı temsil ile ifade edilmesine yardımcı olur.

Bu ön işleme adımları, metin verilerinin daha homojen, temiz ve analiz için uygun hale getirilmesine yardımcı olur. Hangi adımların kullanılacağı, analiz yapılacak veri setine ve hedeflenen sonuca bağlı olarak değişebilir.

Özetleme (summarization)

Metin özetleme, uzun metinlerin veya belgelerin daha kısa bir özetini oluşturmayı amaçlayan bir tekniktir. Metin özetleme, kullanıcıya faydalı bilgiler sağlayan belirli bir metnin yoğunlaştırılmış bir formunun otomatik olarak oluşturulması sürecini içermektedir (Öztürk, 2021). Büyük kuruluşlarda veya organizasyonlarda, araştırmacılar eldeki tüm dokümanları okumak için gerekli zamana sahip olmayabilirler. Bu nedenle temel ve önemli noktaları vurgulanmış özet dokümanlar oluşturulmaktadır (Dang ve Ahmad, 2014). Metindeki cümleler veya paragraflar, özgün metin ile aynı anlamı taşıyan daha kısa bir şekilde ifade edilir.

Kategorizasyon (categorization)

Metin kategorizasyonu metin verilerini belirli kategorilere veya sınıflara atama işlemidir. Bu sınıflar genellikle önceden tanımlanmış kategorilerdir. Metin kategorizasyonu, kategorilerin (sınıfların) önceden belli olduğu ve her eğitim belgesi için kesin ve net olduğu bir tür denetimli öğrenme yöntemidir (Dang ve Ahmad, 2014). Kategorizasyon tekniği, en yalın haliyle ifade edilirse bir veri seti üzerinde tanımlanmış olan belirli sayıdaki kategoriler arasında eldeki veriyi dağıtma işlemidir. Sınıflandırma yöntemi olarak da adı geçen kategorize etme yöntemi, belirlenen eğitim verisinden sınıfların dağılım şeklini öğrenerek, sınıfları belirsiz olan test verileri tanımlandığında en doğru şekilde sınıflandırmaya çalışmaktadır (Öztürk, 2021). Metin sınıflandırma görevi için makine öğrenimi algoritmaları kullanılır. Önceden etiketlenmiş eğitim verileri kullanılarak bir model eğitilir ve daha sonra yeni metinler bu modele göre sınıflandırılır. Metinlerin içeriğini temsil etmek için belirli özellikler veya terimler çıkarılır. Ardından, bu özellikler kullanılarak metinler sınıflandırılır. Örneğin, spam e-postaları tespit etmek veya haber makalelerini kategorilere ayırmak için kullanılabilir.

Kümeleme (clustering)

Metin kümeleme, benzer özelliklere sahip metinleri gruplandırma işlemidir. Bu yöntem, metin verilerindeki benzerlikleri ve ilişkileri bulma ve daha iyi anlama amacı taşır. Kümeleme bir araştırmada incelenen nesnelerin benzerliklerine göre gruplandırılarak sınıflandırılması, varlıkların ortak özelliklerinin ve bu kategorilerin genel tanımlarının ortaya konulmasıdır (İçöz, 2021).

Metin verileri belirli bir sayıda küme veya grup içinde sınıflandırılır. K-Means algoritması, benzer özelliklere sahip metinleri aynı kümelere atar. Hiyerarşik kümeleme yönteminde metinler bir hiyerarşi ağacında gruplandırılır. Benzer metinler alt düzeydeki kümelere, daha genel benzerliklere sahip metinler ise üst düzeydeki kümelere atanır. Temel Bileşen Analizi (PCA) gibi boyut azaltma teknikleri, metin verilerini daha düşük boyutlu temsilcilerine dönüştürerek benzer metinlerin gruplanmasına yardımcı olabilir. Kümeleme yöntemi, analiz edilecek metin belgelerini farklı kümeleme algoritmaları kullanarak gruplar şeklinde sınıflandırmak için tercih edilen denetimsiz bir işlem sürecini kapsamaktadır (Öztürk, 2021). Kümeleme, birçok metin madenciliği ve bilgi çıkarımı

sistemlerinin etkili bir parçası olarak gerçekleştirilen önemli bir yöntemdir (Berry ve Kogan, 2010).

Bilgi görselleştirme (ınformation visualization)

Metin verilerini grafikler, görsel öğeler ve interaktif arayüzler aracılığıyla görsel olarak temsil etme ve anlama işlemidir. Bu teknik, büyük ve karmaşık metin verilerini daha anlaşılır, görsel bir formata dönüştürerek veri keşfi, analiz ve karar verme süreçlerini kolaylaştırmak için kullanılabilir. Metin madenciliğinde bilgi görselleştirme yöntemleri, ilgili bilgilerin keşfedilmesini geliştirebilmekte ve anlaşılmasını basitleştirebilmektedir. İnsan beyni görsel materyali metinsel materyalden daha iyi işler, bu nedenle verileri çizelgeler, grafikler ve tasarım ögeleri kullanarak görselleştirmek trendleri, istatistikleri ve ilgili sonuçları daha kolay iletmemize yardımcı olur (Öztürk, 2021). Normal şartlarda metin formatındaki verilerin kafa karıştırıcı olmasına rağmen, görsel bir şekilde sunulan veriler, kullanıcıların bu verilerden çok daha hızlı ve etkili bir şekilde anlam çıkarmasına yardımcı olmaktadır (Ergün, 2017).

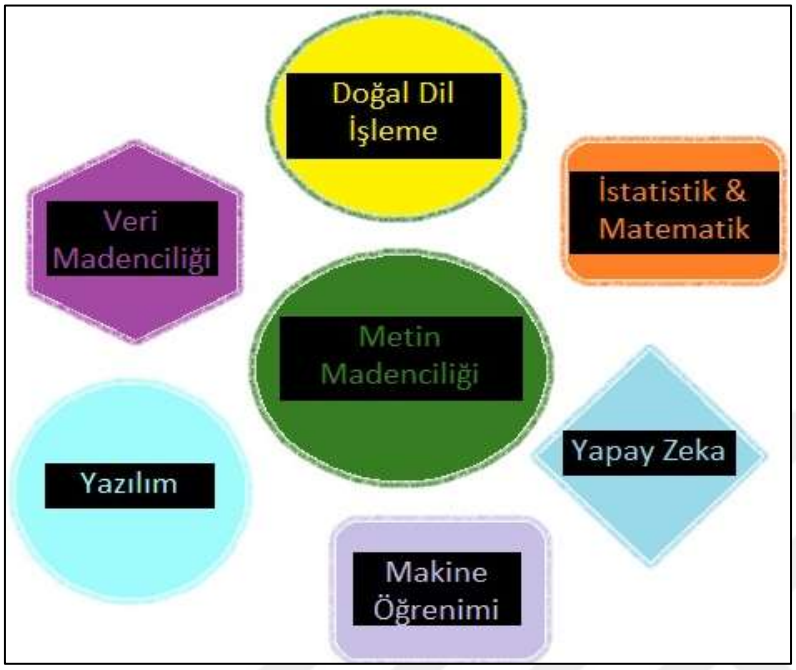
Bilgi görselleştirme, metin madenciliği projelerinde metinlerin anlamını ve içeriğini daha iyi anlamak, keşfetmek ve paylaşmak için güçlü bir araçtır. Bu teknik, metin verilerinin daha erişilebilir ve etkili bir şekilde kullanılmasına katkıda bulunur. Elde edilen sonuçlar incelenir ve yorumlanır. Metin madenciliği bilgi çıkarmak ve anlayış geliştirmek için kullanılır, bu nedenle sonuçlar genellikle raporlar veya görselleştirmelerle sunulur. Metin madenciliği, bilgiye erişimi artırmak, trendleri belirlemek, müşteri geri bildirimlerini anlamak için kullanılabilir. Sonuçların incelenip ve yorumlanarak bilgiye ulaşmamızda hem kullanıcılara hem de bilime büyük ışık tutar.

2.1.3. Metin madenciliği uygulama alanları

Metin madenciliği, birçok farklı alanda uygulanabilir. Örneğin, müşteri geri bildirimlerini analiz ederek pazar eğilimlerini belirlemek, sosyal medya mesajlarını kategorize etmek ve anlam çıkarmak, haberleri otomatik olarak sınıflandırmak gibi uygulamaları vardır. Geleneksel veri madenciliği araçlarını kullanarak veri toplamak çok fazla zaman ve çaba gerektirdiğinden, bu araçlar metinsel verilerin işlenmesinde yeterli değildir (Talib, Hanif, Ayesha ve Fatima, 2016). Bugünlerde gelişmekte olan R, Python vb. gibi yazılımlar ve

yapay zekâ uygulamaları sayesinde bu tür sorunlar büyük ölçüde ortadan kalkmıştır. Veri üretimindeki hızlı büyüme ile birlikte oluşan zorluk veri üretiminin hızlı büyümesiyle birleştiğinde, yalnızca büyük miktarda metinsel veriyi işlemekle kalmayıp aynı zamanda daha iyi kararlar alınmasına da yardımcı olan analitik araçların geliştirilmesine yol açmaktadır. Metin analizindeki yazılım, farklı kaynakları kullanarak elde ettikleri büyük miktarda veri kümelerinden anlam çıkarmayı sağlamaktadır. Bu sayede hızlı kararlar almak, müşteri ihtiyaçlarını anlamak ve doğru bilgilere dayanan stratejiler oluşturmak mümkün olur. Son yıllarda önem kazanan konu modelleme yöntemleri metin madenciliği uygulamalarında sıklıkla tercih edilmeye başlanmıştır. Konu modellemede veri madenciliği, gizli bilgilerin aranması ve bilgi ile metin belgeleri arasındaki ilişkilerin bulunması için metin madenciliğinin en etkili tekniklerden biridir (Bozan, 2022). Metin madenciliği birçok farklı alan ve disiplinle birlikte çalışabilen bir tekniktir. İşbirliği yaptığı alanlardan başlıca olarak doğal dil işleme (NLP) gelmektedir. Metin madenciliği genellikle doğal dil işleme ile birlikte kullanılır. NLP, bilgisayarların insan dilini anlaması ve işlemesi için tasarlanmış bir alanı kapsar. Metin madenciliği, NLP tekniklerini kullanarak metin verilerini analiz eder, kelime dağarcığı, cümle yapıları, duygusal tonlar gibi unsurları anlar. Makine Öğrenimi (ML) algoritmalarını kullanarak büyük metin veri setlerini analiz edebilir. Örneğin, duygu analizi için metin madenciliği, bir metin belgesindeki duygusal tonları belirleyerek olumlu, olumsuz veya tarafsız olduğunu tahmin edebilir. Veri Madenciliği ve Analitiği geniş metin veri setlerinde gizli bilgileri ortaya çıkarmak için veri madenciliği ve analitik yöntemleri ile birleştirilebilir (Phan ve diğerleri 2011). Bu, pazar trendlerini, müşteri davranışlarını ve diğer önemli örüntüleri keşfetmek için kullanılabilir. Bilgi Çıkarma alanı metin madenciliğinde belirli bilgileri çıkarmak ve ilişkileri anlamak için bilgi çıkarma ve ontoloji oluşturma gibi yöntemlerle birleştirilebilir. Bu sayede, metinlerdeki konseptler ve bağlantılar daha iyi anlaşılabilir. Grafik veri analizi ile birleştirilerek metin verileri arasındaki ilişkileri ve ağları görselleştirebilir. Bu, karmaşık metin veri setlerinden anlam çıkarmak için kullanışlı bir yaklaşımdır. Metin madenciliği, tıbbi literatür, hasta kayıtları ve sağlık anketleri üzerinde çalışarak sağlık bilgi sistemlerinde kullanılabilir. Hastalık teşhisi ve tedavi yöntemleri gibi konularda bilgi çıkarmak için kullanılır. Büyük veri analitiği ile birleştirilerek geniş metin veri setlerinden hızlı ve etkili bir şekilde bilgi çıkarmak için kullanılabilir. Bu, büyük ölçekli metin verilerindeki önemli örüntüleri belirlemek açısından önemlidir.

Bu alanlarla birlikte çalışarak büyük veri setlerinden anlam çıkarmak, önemli bilgileri ortaya çıkarmak ve bilgiye dayalı kararlar almak için güçlü bir araç haline gelir.



Şekil 2.1. Metin madenciliğinin çalıştığı alanlar

Metin madenciliğinin yedi farklı uygulama alanı bulunmaktadır ve bunlar, bilgiye erişim, bilgi çıkarımı, doğal dil işleme, belge sınıflandırma, belge kümeleme, web madenciliği ve kavram çıkarımıdır (Miner, Delen, Elder, Fast, Hill ve Nisbet, 2012). Veri madenciliği, büyük miktarda metinden bilgi çıkarmanın bir yoludur. Araştırmacılar edebiyattaki metin kalıplarını tespit edebilir, metinleri karşılaştırabilir ve birbirleriyle ilişkilendirerek kolayca analiz edebilirler (Richards ve diğerleri, 2015). Metin madenciliği çok fazla alanda uygulanabilmesi nedeniyle çok fazla sektörün kapısını açabilmektedir. Metin madenciliği (text mining), sektörlere çeşitli avantajlar sağlayabilen güçlü bir analitik araçtır.

2.1.4. Metin madenciliğinin yararları ve işlevleri

Metin madenciliği, büyük metin veri kümelerindeki desenleri ve ilişkileri tespit ederek yeni bilgiler keşfetmemizi sağlar. Anlamlı verilere dayanarak doğru stratejik kararlar almayı destekler ve rekabet avantajı sağlar. Metinlerdeki eğilimleri ve trendleri belirleyerek piyasa koşulları, müşteri talep ve istekleri hakkında önemli bilgiler sunar. Metin madenciliği teknikleriyle, iki farklı belgenin birbirleriyle benzer olup olmadığına,

karar verilebilir. Bu işlem frekans analizi ile yapılabilir. Belgelerin içerisindeki kelimelerin tekrar sayıları birbirleriyle karşılaştırılarak, iki belgenin birbirleriyle benzerliğine karar verilebilir (Alpaydın, 2000).

Metin madenciliği, verilerdeki kavramları, kalıpları, temaları, anahtar kelimeleri ve diğer birçok özelliği tanımlayabilen yazılıma dayalı olarak büyük miktarlarda yapılandırılmamış metin verilerinin madenciliği süreci ve analizidir (Çelik, 2020). Metin madenciliği, büyük metin veri setlerini analiz ederek anlamlı bilgiler çıkarmak amacıyla kullanılan güçlü bir veri analizi yöntemi olmasıyla birlikte teknik olarak doğal dil işleme, istatistik ve makine öğrenimi gibi alanları içerir ve birçok sektörde önemli avantajlar sağlar. Metin verileri içerisinde anlam çıkarma ve bilgi keşfi sürecinde metin madenciliği, metin verilerinde gizli olan anlamlı bilgileri keşfetmek için kullanılır. Büyük metin veri setleri üzerinde yapılan analizler, belirli kelimelerin, konuların veya duygusal tonların öne çıkmasını sağlar. Bu sayede, işletmeler, araştırmacılar ve kurumlar verilerindeki önemli desenleri ve ilişkileri daha iyi anlayabilir. Metin verilerindeki duygu analizi ve müşteri memnuniyetini metin madenciliği aracılığıyla müşteri yorumlarından veya sosyal medya paylaşımlarından duygusal tonları çıkarma yeteneği ile müşteri memnuniyetini değerlendirmek için kullanılır. Bu sayede, işletmeler müşteri geri bildirimlerine hızlı bir şekilde tepki verebilir ve ürün veya hizmetlerini iyileştirebilir. Metin araçları içerisinde pazarlama stratejilerini geliştirme süreçlerinde metin madenciliği, pazarlama stratejilerini oluşturmak veya iyileştirmek için kullanılır. Tüketicilerin beklentilerini, tercihlerini ve tepkilerini anlamak, hedef kitleye daha etkili bir şekilde ulaşmak ve ürünleri pazarlamak için önemlidir.

Büyük metin koleksiyonlarındaki ana temaları belirlemek için kullanılır (Pasin, 2018). Bu, araştırmacılara, haber analistlerine veya içerik oluşturuculara hangi konuların popüler olduğunu ve hangi trendlere odaklanmaları gerektiğini gösterir. Güvenlikle ilgili metin verilerini analiz ederek tehditleri veya anormallikleri tespit etme yeteneği sağlar. Bu, siber güvenlik alanında önleyici önlemler almak için önemlidir. Metin madenciliği, çalışan geri bildirimleri ve anket verileri üzerinde analiz yaparak şirket içindeki çalışan memnuniyetini değerlendirebilir. Bu, insan kaynakları departmanlarında işyeri kültürünü geliştirmek ve çalışanların ihtiyaçlarına daha iyi yanıt vermek için veri odaklı bir perspektif sunar. **S**ağlık alanında hastane kayıtları, hasta notları ve tıbbi literatür üzerinde analizler yapmak için kullanılır. Bu, hastalıkların yayılması, tedavi yöntemlerinin etkinliği ve sağlık eğilimleri gibi konularda önemli öngörüler sağlar. Metin madenciliğinin başlıca kullanım alanları

kaynak kod incelemesi yaparak yazılım hatalarını tespit etmek, tıp ve sağlık analizi, müşteri incelemelerini analiz ederek ürün veya hizmet performansını değerlendirmek, sosyal medya mesajlarını izleyerek marka itibarını takip etmek, tıklatmalar, görüntülemeler ve dönüşümler gibi dijital pazarlama verilerini analiz etmek gibi birçok alan sayabiliriz. Metin madenciliği, birçok sektöre katkı sağlar. Finans sektöründe kullanılarak, piyasalar hakkında doğru tahminlerde bulunulabilir; sağlık sektöründe kullanılarak, hastalık teşhisi ve tedavi planlama süreci hızlandırılabilir; enerji sektöründe kullanılarak, üretim maliyetleri azaltılabilir. Metin madenciliği, verilerden anlamlı bilgiler çıkararak işletmelerin rekabet avantajı elde etmelerine, karar verme süreçlerini iyileştirmelerine ve müşteri memnuniyetini artırmalarına yardımcı olur. Metin madenciliği doğru tekniklerle kullanılırsa, işletmeler piyasada oluşması muhtemel trendleri öngörebilirler. Metin madenciliği, insanların manuel olarak işleyemeyeceği miktarda veri toplayıp işleyebilir. Bu sayede daha anlamlı bilgiye ulaşılabilir. Metin madenciliği, müşterilerin taleplerini daha iyi anlamak için kullanılır ve müşteri memnuniyeti artırılabilir.

2.1.5. Gelecekte metin madenciliğinin rolü ve gelişimi

Metin madenciliği, hızla büyüyen bir alandır ve gelecekte daha da önemli bir rol oynayacaktır. Yapay zekâ ve derin öğrenme gibi yeni teknolojilerin gelişimi ile birlikte, metin madenciliği daha karmaşık ve etkili hale gelecektir. Bu sayede, daha fazla veriye erişim sağlamak, daha doğru sonuçlar elde etmek ve daha verimli iş süreçleri oluşturmak mümkün olacaktır. Metin madenciliği, teknoloji dünyasının ve veri biliminin hızla evrildiği bir dönemde önemli bir yere sahip olduğu gibi her gün milyonlarca metin belgesi, e-posta, sosyal medya gönderisi ve haber makalesi gibi birçok veri üretiliyor. Bu büyük metin verisi bilgi çıkarma, duygu analizi, trend keşfi ve daha birçok uygulama için potansiyel barındırır. Sosyal medya kullanımı arttıkça tüketicilerin sanal ortamda oluşturduğu veriler, tüketici tercihlerinin belirlenmesinde kullanılabilecek kişilerin dijital ayak izlerini oluşturduğundan bu bilgiler geleceği tahmin etmek ve analiz etmek için kullanılabilir (Taş ve Bülbül, 2021).

Gelecekte, metin madenciliğinin rolü ve gelişimi büyük ölçüde etkili olacak ve iş dünyasını ve toplumu derinlemesine dönüştürecektir. Metin madenciliğinin gelecekteki rolünü anlamak için öncelikle teknolojik gelişmeler incelenebilir. Yapay zekâ ve derin öğrenme teknikleri, metin madenciliğini daha da güçlendirir. Doğal Dil İşleme (NLP) alanında yapılan ilerlemeler, metin verisinin anlaşılmasını, sınıflandırılmasını ve

özetlenmesini daha hassas hale getirir. Bu, metin madenciliğinin daha geniş ve karmaşık veri kümelerini ele almasına olanak tanır. Büyük veri ve bulut bilişim de metin madenciliğinin gelişimini destekliyor. Veri depolama ve işleme kapasitesi arttıkça, daha büyük veri kümesi analizleri yapmak mümkün olur. Bu, özellikle işletmeler için pazarlama stratejilerini geliştirmek, müşteri ilişkilerini yönetmek ve rekabet avantajı elde etmek için büyük önem taşır. Gelecekte metin madenciliği aynı zamanda çoklu dil ve çoklu kültürlü analizde de büyük rol oynayabilir. Küresel bağlantılar arttıkça, metin verisinin farklı dillerde ve kültürlerde anlaşılması ve yorumlanması daha fazla önem kazanır. Bu, uluslararası işletmeler için küresel pazarlama stratejileri oluştururken ve kültürel etkileşimleri anlarken büyük bir avantaj sağlar.

Metin madenciliğinin gelecekteki rolü ayrıca sağlık hizmetleri, eğitim, haber medyası ve kamu politikası gibi birçok sektörü etkileyecektir. Sağlık hizmetlerinde hastaların tıbbi verilerini analiz etmek, hastalıkları daha erken teşhis etmekte ve tedavi planlarını iyileştirmekte kullanılabilir. Eğitimde, öğrenci performansını izlemek ve öğrencilere daha kişiselleştirilmiş eğitim sağlamak için kullanılabilir.

Sonuç olarak, metin madenciliği gelecekte iş dünyasının ve toplumun birçok yönünü derinlemesine etkileyen güçlü bir araç olarak kalmaya devam edebilir. Teknolojik ilerlemeler ve veri bolluğu, metin madenciliğinin rolünü daha da önemli hale getirebileceği gibi daha fazla uygulama alanı sunacaktır. Bu, bilgi çağında verilerden anlam çıkarmak ve daha bilinçli kararlar almak isteyen herkes için büyük bir fırsat sunabilir.

2.2. Python

Python, genel amaçlı ve yüksek seviyesi olan bir programlama dili olmaktadır. Guido van Rossum sayesinde geliştirilmeye başlanmış ve sürekli olarak geliştirilmeye devam edilmektedir. Python'ın en belirgin özelliklerinden biri, kodunun okunabilir ve anlaşılır olmasıdır. Python yazılım dili olarak diğer dillere kıyasla yazım biçimi açısından daha kolaydır ve herhangi bir düzenleyiciye ihtiyaç duyulmamaktadır (Özgül, 2013). Bu, hem yeni başlayanlar hem de deneyimli geliştiriciler için büyük bir avantajdır. Python, büyük bir geliştirici topluluğuna sahiptir ve bu topluluk, kullanıcılar arasında bilgi paylaşımını destekleyen zengin bir kaynak oluşturur. Ayrıca, Python'un geniş bir dokümantasyonu bulunmaktadır, bu da yeni başlayanlar için öğrenme sürecini kolaylaştırır. Python,

programlamaya başlamak veya bu dili öğrenmenin başlıca bazı nedenlerinden biri okunabilir olmasıdır. Birçok geniş kapsamlı programlama dilinden biri olan Python, okunması ve anlaşılması kolay, güçlü bir programlama dilidir.

Yazılım dili olan Python'ın kod yazarken düzenli ve anlaşılır olunmasına yardımcı olur. Girinti kullanarak kod bloklarını tanımlar, bu da kodun düzenini sağlar. Python dili geniş kütüphane desteği vermektedir. Birçok kullanışlı standart kütüphane içerir ve daha fazla işlevselliği kütüphanelerle genişletilebilir. Bu kütüphaneler, veri analizi, web geliştirme, yapay zekâ, veri tabanı yönetimi ve daha birçok alanda kullanılır. Python'ın çoklu platform desteği ile Windows, macOS ve Linux gibi birçok işletim sistemi üzerinde çalışabilir. Bu, farklı platformlarda uygulama geliştirmenizi sağlar. Topluluk ve kaynaklar sayesinde Python, büyük ve aktif bir geliştirici topluluğuna sahiptir. Projelerin geliştirebileceği birçok çevrimiçi kaynak ve forum bulunmaktadır. Web geliştirme, veri bilimi, yapay zekâ, oyun geliştirme, otomasyon, bilimsel hesaplamalar ve daha birçok alanda kullanılabilir. Bu nedenle sınırları olmayan bir programlama dilidir. Hem veri hem de metin madenciliğini destekleyen uygulama, makine öğrenimi algoritmaları oluşturmak için de kullanılabilen açık kaynaklı bir dildir (Kavak, 2022).

Çalışmada ön işlemler, veri setlerini okuma, makine öğrenmesi yöntem ve algoritmaları gibi tüm işlemler Python programlama dilinde görselleştirme, yazım ve analiz işlemleri yapılmıştır. Tüm yöntemler için gerekli kütüphaneler Python ortamına eklenmiştir. Python ücretsiz olmasının avantajının yanında açık kaynak kodlu anlaşılır bir programlama dilidir. Python için çok fazla kütüphane olmasının avantajı Python dilini diğerlerinden farklı olmasını sağlayan özelliklerin başında gelmesini sağlar. Python, birçok kütüphane ve modülle birlikte gelir. Bu kütüphaneler, çeşitli görevleri gerçekleştirmek için kullanılabilir. Veri analizi, makine öğrenmesi, web geliştirme ve doğal dil işleme gibi birçok alanda güçlü kütüphanelere sahiptir.

Bu kütüphanelerin başlarında Gensim, Optuna, Nltk, SpaCy ve Scikit learn gibi kütüphaneler gelir. Bu tür kütüphanelere ek veri bilimi içinde makine öğrenmesi alanında sıkça kullanılan pandas, numpy, missingno ve matloplit gibi veri analizinde görselleştirme yapan kütüphaneler vardır (Arslan, 2019). Pandas işlemleri daha hızlı yapabilmek amacıyla geliştirilmiş işlev olarak zaman serileri gibi sayısal dizin işlemleri ve bunlarla birlikte veri yapıları kolayca oluşturulup işleme sokulabilmektedir (Chen, 2017). Çok boyutlu dizinlerde matematiksel dizin işlemleri hızlıca ve rahatça yapılabilmesi adına geliştirilen

Python dilinin bir kütüphanesidir. Matplotlib ve Seaborn kütüphanesi görselleştirilmesi istenilen verileri belli bazı formatlarda, grafik çeşitleriyle etkin bir şekilde görselleştirme için kullanılan bir kütüphanedir. Python'ın geniş kütüphane desteği, okunabilirliği, çok yönlülüğü ve topluluk desteği gibi avantajları, hem bireysel geliştiriciler hem de işletmeler için çeşitli projelerde etkili bir araç olmasını sağlar.

2.3. Metin Madenciliğinde Kullanılan Makine Öğrenmesi Algoritmaları

2.3.1. Gizli dirichlet ayrımı yöntemi (latent dirichlet allocation)

LDA (Latent Dirichlet Allocation), metin madenciliği ve doğal dil işleme alanlarında belgelerin gizli temalarını çıkarmak için kullanılan bir olasılık modelleme tekniğidir. Bu yöntem, metin madenciliği, doğal dil işleme ve bilgi çıkarma gibi alanlarda sıklıkla kullanılmaktadır. Aynı zamanda büyük metin verilerinin analizi için kullanılan bir makine öğrenme yöntemi ve makine öğrenimi algoritması diyebiliriz. Bu algoritma, bir metnin içindeki gizli (latent) konuları (topics) belirleyen bir temel modeldir. Bir belgenin her bir kelimesini, belgenin hangi gizli konuları içerdiğine göre sınıflandırır. LDA'nın temel fikri, konuların sabit bir sözlüğün olasılık dağılımını içermesi ve belgelerin rastgele birleştirilmiş gizli konulardan oluşmasıdır. Bu temel fikir, LDA'nın bir dizi belgenin konularını, konulardaki kelimelerin olasılıklarını, belgeyi oluşturan kelimelerin hangi konulara atandığını ve konuların o belgede nasıl dağıldığını öğrenerek keşfetmesini ifade eder (Blei, Ng ve Jordan, 2003). LDA algoritması kullanılarak belgeler hem tüm arşivin metinsel verilerinin bir özeti hem de önemli tanımlayıcı anahtar kelimeler içeren konuların bir özeti olarak ifade edilebilir. LDA, arşivin tamamındaki konuların bir belgede bulunma olasılığını hesaplayarak tespit edilmesini sağlar (Kartal, 2017).

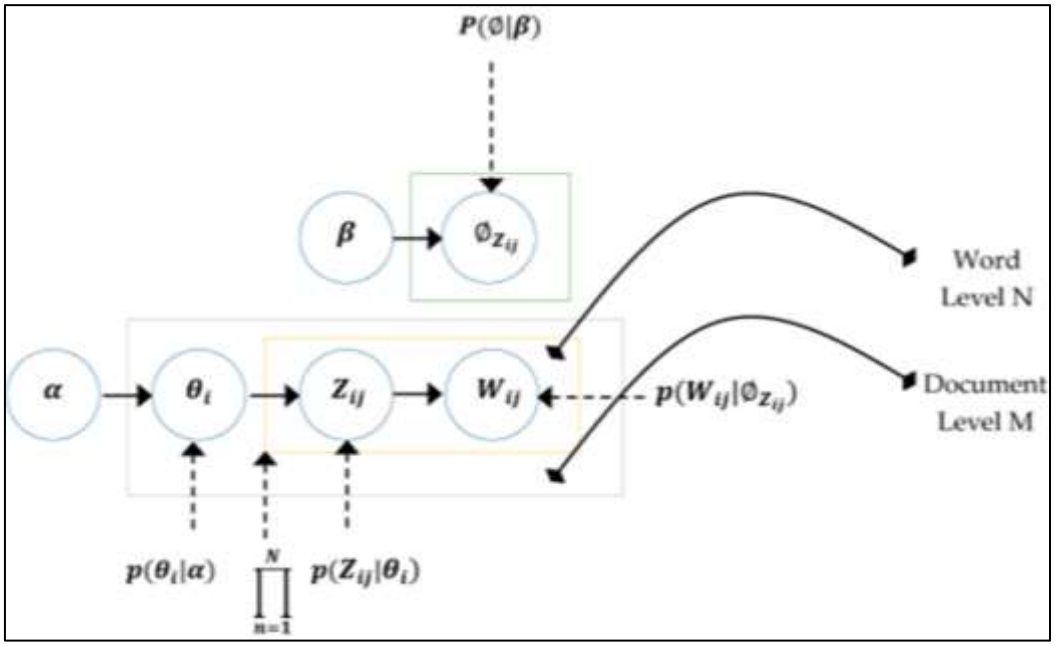
Bir çalışma olarak, her belgedeki kelimelerin konuların bir karışımından oluştuğunu, konunun tamamının sabit bir kelime dağarcığı yerine bir karışımdan olduğunu varsayar. Koleksiyondaki tüm belgeler ortak konuları paylaşır ancak konu oranları, Dirichlet dağılımından rastgele seçildikleri için belgeler arasında stokastik olarak farklılık gösterilebilir (Blei ve Lafferty, 2007).

LDA algoritmasına göre her kelime ve kelime kökü bir konuyu temsil eder. Bütün belgeler belli ölçülerde konu içermektedir (Chen ve diğerleri 2015). Genel olarak her belge, belgeye

dahil edilme olasılığını hesaplayan ve arşivler arasında konu tespitine olanak tanıyan konu ve algoritmaların bir karışımı olarak temsil edilebilir. LDA'nın temel amacı, bir belge koleksiyonundaki temaları ve her belgenin bu temalar üzerindeki dağılımını modellemektir. Bu, belgelerin içerdiği gizli temaların keşfedilmesine ve benzer temalara sahip belgelerin gruplandırılmasına olanak tanır (Alrayashi, 2023) . İşleyiş açısından, LDA, her belgenin farklı temalardan oluştuğunu ve her bir kelimenin bu temalar arasında bir ilişkiye sahip olduğunu varsayar.

Model, belgelerin içerdiği temaların belirli bir dağılımını ve her tema içinde geçen kelimelerin olasılık dağılımını belirlemeye çalışır (Budak ve Sökmen, 2022). Bu, belgelerin temalar üzerindeki içsel yapısını anlamak için kullanılır. LDA'nın uygulama alanları arasında konu modelleme, belge sınıflandırma, öneri sistemleri ve içerik analizi gibi çeşitli görevler bulunmaktadır. Bu yöntem, büyük belge koleksiyonları üzerinde desenleri ve temaları çıkarmak için güçlü bir araç olabilir.

LDA algoritmasını kullanarak bir belge, arşivin tamamındaki metinsel verilerin bir özeti haline gelir ve önemli tanımlayıcı anahtar sözcükleri içeren konularla ifade edilebilir (Öztürk, Sankur, Güngör ve Yılmaz, 2014). Gizli Dirichlet Ayrımı (LDA), metin verilerini analiz etmek, belgeleri temsil ve gruplama yapmak için kullanılan bir olasılık modelidir. Model örneği ise;



Şekil 2.2. LDA grafik modeli (Nilashi ve diğerleri, 2021)

Modelin formülü ise;

$p(D|\alpha,\beta)=\prod_{d=1}^{M}\int p(\theta_d,\alpha)(\prod_{n=1}^{N_d}\sum_{Z_{d_n}} p(Z_{d_n}|\theta_d)p(W_{d_n}|Z_{d_n},\beta))d\theta_d$

Şekil 2.2'de verilen modelde;

m parametresi, üretilen belge sayısını;
α, konu dağılımını;
z, bir konuyu;
β, belirli doküman için konu dağılımını;
w, kelimeleri temsil etmektedir

θ parametresi ise her bir doküman için ayrı örneklenmektedir (Nilashi ve diğerleri, 2021)

p(D|α,β): Tüm belgelerin (D) belirli α ve β parametreleri altında belirli bir olasılığını ifade eder.

$\prod_{d=1}^{M}$:M adet belgeyi temsil eden döngüyü ifade eder. Yani, belge sayısı kadar işlem yapılır.
$\int p(\theta_d,\alpha)$, d belgesinin dağılımını (topic distribution) ifade eder. Bu belgenin içeriğindeki temaların dağılımını temsil eder ve bu dağılım α parametresine göre tanımlanır.

$\prod_{n=1}^{N_d}\sum_{Z_{d_n}} p(Z_{d_n})$: Her bir belge içindeki terimlere (kelimelere) ve bu terimlerin gizli konularına (temalarına) dair bir döngüyü ifade eder.

$p(Z_{d_n})$, d belgesinin n terimine karşılık gelen gizli konuyu (temayı) ifade eder. Bu, belge içindeki terimlerin hangi temalara ait olduğunu belirler ve θ_d'ye bağlı olarak belirlenir.

$p(W_{d_n}|Z_{d_n},\beta)$, d belgesinin n terimini ifade eder. Bu, belgenin içeriğindeki terimin hangi temaya ait olduğunu ve bu terimin belirli bir temadan gelme olasılığını β parametresine göre belirler.

$d\theta_d$, Bu, θ_d'nin dağılımını ifade eder ve θ_d'nin α parametresine göre bir olasılık dağılımını temsil eder.

Ayrıntılı şekilde incelenecek olursa, LDA belgelerin şu şekilde üretildiğini varsayılır;

- Belgelerin içerisinde bulunan kelime sayısı(N) belirlenir.
- İlk olarak olasılık dağılımlarına bakılarak bir konu seçimi yapılır.
- Dirichlet dağılımı içerisindeki k adet konudan amaca göre bir kaç tanesi için seçim işlemi yapılır.
- Belge içerisindeki bir konuya ait "wi" kelimesi eklenir. Eklenecek olan kelime seçilen konunun multinomial dağılımına uygun seçilir. Bu şekilde bir sürü belgenin elde edildiğini varsayım yaparsak k tane konuyu ve konuların içerisindeki kelimeleri LDA algoritmasıyla belirlemek için şu adımlara bakarsak;

1. Her belge içerisindeki kelimeler konu kapsamına göre rasgele olarak seçilerek belli bir konu içerisine atanır.
2. Rastgele atama, kelimelerin tüm belgelerdeki ve tüm konulardaki dağılımını ve oranını belirler. Tüm belgeler için yapılan atamalar iyileştirilir. Belgedeki w kelimelerinin her biri ele alınır. i. Her konu için iki tür oran hesaplanır. Belgede ise o süreç içinde konu içeriğine tanımlanmış kelimelerin oranı w. kelime tarafından gelen konunun bütün belgeler içerisindeki oranı ve w. kelimenin olasılık dağılımı hesaplanıp yeni konunun içine atanır.
3. Bu adımda sürekli tekrarlandığı için konu atamalarında istikrar olduğu belirlenen bir duruma ulaşılması sağlanır. Atamalar sayesinde belgelerin her birinin içerdiği konu olasılıklarının yanında her konunun kelimeleri de, tekrarlanma sayılarıyla birlikte elde edilebilir.

LDA algoritması tarafından tüm belgelere uygulanan bu işlemlerin adımları daha açık şekilde örnekleyecek olursak;

- İnek eti ve balık eti yerim
- İnekler kümes hayvanıdır
- İnek balık yemez

Burada LDA, altı çizili kelimeleri Y konusu altında toplar. Y konusunu Yiyecek olarak etiketleyebiliriz. LDA, aynı şekilde italik kelimeleri de H konusu altında toplar. H konusunu da Hayvanlar olarak etiketleyebiliriz. LDA'nın kelime seviyesinde işlem yapması iki açıdan önemlidir. Her bir cümlenin içeriğine bakarsak kelime sayısı ile çıkarılabilir. Birinci cümle: %100 Y konusu İkinci cümle: %100 H konusu Üçüncü cümle: %40 H konusu ve %60 Y konusu 2. Her kelimenin ilgili konudaki oranı çıkarılabilir. Örneğin Y konusu %40 balık, %40 et, %20 yemek içermektedir.

Bu şekilde tüm kelimeler için bu işlem adımları her seferinde sürekli tekrarlanır. Bu süreç bitiminde nihai sonuca ulaşılır. Belge kümelemeye benzer şekilde LDA yöntemi, belgeleri belirli sayıda k sayısı kadar konuyla rastgele eşleştirerek konu modellemesi gerçekleştirilir. Her belgeyi en iyi anlatan konuların belirlenmesinde, her konuyu anlatan en iyi kelimelerin belirlenmesiyle de öğrenme sağlanır (Aydın ve Hallaç, 2021). Mesela bir sağlık makalesinde hem hasta hem de ilaç konularını içerebilir. Her kelimenin belgedeki belirli bir konuya ait bir olasılığa sahip olduğunu varsayar. Örneğin, "hasta" kelimesi teşhis konusuyla daha fazla ilişkilendirilebilirken, "ilaç" kelimesini ise tedavi konusuyla daha fazla ilişkilendirilebilir.

LDA, bir belgenin bir konunun birleşiminden oluştuğunu varsayar ve her bir belgenin kaydedildiği yere göre gruplandırılır. Bu gruplama, belgelerin içerdiği kelimeleri temel alır. LDA'nın çalışma programı şu şekildedir: öncelikle belgelerdeki kelimelerin frekanslarını hesaplar ve bu frekanslara dayanarak bir kelime konumu görünür olarak oluşturulur. Daha sonra, içerdikleri sözcüklerin temellerini alarak, konu dağıtımlarını tahmin eder. Bu tahmin, bir başlangıçtan başlayarak yinelenen bir şekilde gerçekleşir. Son olarak, konu dağılımlarını ve sözcüklerin konu dağılımlarını birleştirerek, her yerde hangi konuya ait olduğunu tahmin eder (Nilashi ve diğerleri, 2021).

LDA, her belgenin gizli temalara olan katkısını istatistiksel olarak hesaplar. Bu sayede, bir belgenin hangi temalarla ilişkilendirildiği ve hangi temaların bu belgede ne kadar etkili olduğu anlaşılabilir (Agrawal ve diğerleri, 2018). Büyük metin koleksiyonlarını daha anlaşılır hale getirerek öngörü elde etmeyi amaçlar. Bu öngörüler, metin verilerini analiz eden araştırmacılara veya işletmecilere yardımcı olabilir. LDA, metin belgelerini tematik gruplara ayırarak belge sınıflandırma işleminde kullanılabilir. Kullanıcıların ilgi alanlarına dayalı olarak içerik önerileri yapmak için kullanılabilir. Kullanıcının geçmiş tercihlerine

dayalı olarak hangi temalara ilgi duyabileceğini tahmin edebilir. LDA, belgelerdeki duygusal tonlamaları ve temaları analiz ederek duygu analizi yapmak için kullanılabilir.

LDA, metin belgeleri üzerinde gizli temaları ve bu temaların belgeler arasındaki dağılımları için kullanılan güçlü bir olasılık modellemesi yöntemi olması nedeniyle Çin restoranına yapılan müşteri yorumlarının LDA ile analizi, bu yorumların altında yatan gizli konuları ve ana temaları ortaya çıkarmak için güçlü bir araç sağlar.

LDA (Latent Dirichlet Allocation) üzerine yapılan çalışmaları incelediğimizde

Afaq, Gaur ve Singh (2022), Güney Asya'daki zincir otellerin bazılarının çevrimiçi yorumları ile konu modellemesi yapmışlardır. Çalışmalarında veriler TripAdvisor'dan alınmıştır. Konu modellemesi olarak müşteriler tarafından yorumlarda belirtilen başlıca konuları analiz etmek için veri seti kullanılarak LDA analizi yapılmıştır. Analiz çalışmasına göre hijyen, yemek lezzeti, personel davranışı ve hizmetlerin otel misafirleri tarafından temel sorunlar olduğu belirlenmiştir. Ayrıca, oteldeki misafirleri etkileyen temel sorunların servis süresi olduğu tespit edilmiştir.

Xiang, Du, Ma ve Fan (2017), çalışmalarında TripAdvisor ve Yelp seyahat sitelerinde bulunan otellere yapılan yorumlar LDA yöntemi kullanılarak analiz edilmiştir. Yapılan LDA analizi sonucunda yapılan yorumların temel konular (hizmet, ambiyans, gezilecek yerler, yemek, ürün çeşidi) altında değerlendirilmiştir. Sonuçlara bakıldığında, otel endüstrisinin bu tür platformlarda temsil edilmesinin farklılıklar meydana getirdiğini göstermiştir.

Nilashi ve diğerleri (2021), çalışmalarında vejetaryen restoranlara yapılan müşteri yorumlarını TripAdvisor ve sosyal ağ sitelerinden toplanan veri setleri üzerinde değerlendirilmiştir. Müşterileri değerlendirme kriterlerinin temelinde dört müşteri segmenti sunulmaktadır. Çok düşük ve düşük memnuniyet seviyeleri için, müşterilerin çoğunluğunun Segment 1'de atmosfer, hizmet, yemek ve değer konularında memnun olmadığı görülmüştür. Ayrıca, Segment 2, Segment 3 ve Segment 4'te müşteriler, tüm yönlerden restoran hizmetlerinden daha memnun görünmektedir. Sonuçlar, Segment 3 ve Segment 4'ün çok yüksek memnuniyet seviyesine sahip müşterilerin çoğunluğunu içerdiğini göstermektedir. Bu, dört segmentte müşteriler için hangi kriterlerin önemli

olduğunu gösterilmiştir. Ayrıca, bu sonuçlar müşterilerin tercihlerini ve restoran hizmetlerinden nasıl memnun olduklarını göstermiştir.

Ekinci, İlhan, Kırık ve Taşçı (2020), yaptıkları çalışmada 2007- 2017 yılları aralığında tıp alanında Türkiye'de bulunan araştırmacılar tarafından oluşturulan literatür çalışmaları elde edilmiştir. Bu elde edilen literatür ile LDA kullanılarak konu modellemesi yapılmış ve sonucunda hangi yıllarda daha çok hangi konuların ağırlıklı olarak çalışıldığı tespit edilmiştir. Model değerlendirilmiş ve LDA'nın anlamsal analiz olarak başarılı olduğu gözlemlenmiştir.

Kartal (2017), yaptığı araştırmanın kapsamında LDA algoritmasını kullanarak konu modellemeye ilişkin yıl bilgisi ile TOJDE dergilerinden makaleler çıkartılarak, makale arşivinde çalışılan konuların çeşitliliğinin yanı sıra araştırmacıların hangi konuları incelediği veya nelerden vazgeçtiği hakkında bilgi edinilmiştir. Yıllar geçtikçe hangi konuların daha çok ilgi gördüğü konuları ile ilgili bilgiye ulaşılmıştır.

Güven, Diri ve Çakaloğlu (2018), yapılan çalışmada sosyal medyada atılan tweetlerden yola çıkarak belirli duygu türüne ait olan konu modelleme algoritması LDA ile tespit edilmiştir.

Karaosmanoğlu (2022), koronavirüs konu kapsamında medya haber içeriklerini hem insanların hem ülke yöneticilerinin takip edebilecekleri bir analiz yapmak için geniş veri kümesi vardır. Konu Modelleme yöntemleri ile Koronavirüs Salgını süresi boyunca Türkçe yayınlanan medya haberlerinin konu temaları belirlenmiştir. Araştırma kapsamında LDA konu modelleme yapılmıştır. Uygulanan analiz sonucunda Covid-19 Salgını süresince yayınlanan medya haberlerinin konu başlıkları toplanmış. En fazla geçen konular 'maske', 'virüs', 'hijyen', 'dezenfektan' ve 'aşı' gibi gündem süreside en fazla olan kelimeler, en çok geçen konular içinde yer almaktadır.

LDA'nın çalışma aşamaları kısaca özetlemek gerekirse;

Veri toplama ve hazırlık

İlk olarak analiz yapmak istenilen metin belgelerini toplamak ve bu belgeleri uygun bir formatta hazırlamaktır. Bu adımda, metinler temizlenir, gereksiz karakterler ve stop kelimeleri kaldırılır. Metin verileri daha sonra belirli bir formatla düzenlenir. Örneğin her bir doküman bir dizi kelime veya terim olarak temsil edilir. Bu belgeler örneğin makaleler, haberler veya müşteri yorumları olabilir.

Belge-terim matrisi oluşturma

LDA'nın temelinde bir kelime-terim matrisi bulunur. Verileri işlemeye başlamadan önce, her belgenin içerdiği terimleri ve terimlerin frekanslarını temsil eden bir belge-terim matrisi oluşturulur. Bu matris, belgelerin ve terimlerin sayısal bir temsilini sağlar. Belge koleksiyonundaki her kelimenin veya terimin kaç kez geçtiğini gösterir. Bu matris, metinlerin içeriğini sayısal bir formata dönüştürmek için kullanılır.

Model parametrelerinin belirlenmesi

LDA'nın başarılı bir şekilde uygulanabilmesi için bazı parametrelerin belirlenmesi gerekir. Bu parametreler arasında toplam konu sayısı (k), her belgedeki konu dağılımı için kullanılan Dirichlet dağılım parametreleri ve her konu için kelime dağılımı için kullanılan Dirichlet dağılım parametreleri bulunur. Bu parametreler, analizin doğruluğunu ve sonuçların kalitesini etkiler.

LDA modelinin eğitimi

LDA modeli, veri hazırlığı ve parametre belirleme adımlarının ardından eğitilir. Model, metin koleksiyonundaki gizli konuları keşfetmeye çalışır. LDA modelini eğitmek için örnekleme veya değişken belirleme gibi özel bir algoritma kullanılır. Bu algoritma, belge-terim matrisini ve model parametrelerini kullanarak belgelerdeki temaları ve temaların dağılımlarında tahminde bulunur. LDA modelinin oluşturulmasında, belirli bir sayıda konu (topic) belirlemek önemlidir. Bu sayı, elde edilmek istenen detay seviyesine bağlı olarak ayarlanır. LDA modeli, metin verilerini ve doküman terim matrisini kullanarak gizli konuları ve bu konuların dağılımlarını çıkarmak için eğitilir.

Temaların ve terimlerin analizi

Eğitilen model tamamlandığında, LDA modeli her bir belgenin konu dağılımını ve her bir konunun kelime dağılımını verir. Her belge, farklı konuların bir karışımı olarak temsil edilir. Bu karışım, LDA modeli tarafından tahmin edilir. Örnek olarak, bir konu "yemeklerin çeşitliliği" ile ilişkilendirilebilirken, başka bir konu "hizmet kalitesi" ile ilişkilendirilebilir. Bu sonuçlar, metin koleksiyonundaki gizli temaları veya konuları keşfetmemize yardımcı olur. Belge konu dağılımları, müşteri yorumlarının altında yatan ana temaları ve konuları anlamamıza yardımcı olur. Örneğin, bir haber makalesinin konusu veya bir müşteri incelemesinin ana teması belirlenebilir.

Sonuçların incelenmesi ve değerlendirilmesi

LDA sonuçları, belgeler arasındaki temaları, önemli terimleri ve temaların belgelerdeki katkılarını içerir. Bu sonuçlar, metin verileri üzerinde yapılan analizlerin yorumlanması ve işletme kararları için kullanılması için incelenir ve değerlendirilir. Her belge, farklı konuların bir karışımı olarak temsil edilir. Bu karışım, LDA modeli tarafından tahmin edilir. Belge konu dağılımları, her yorumun altında yatan ana temaları ve konuları anlamamıza yardımcı olur. Elde edilen sonuçlar, restoran işletmecileri ve yöneticileri tarafından incelenir. Hangi konuların müşteriler için en önemli olduğu ve iyileştirilmesi gereken alanlar belirlenir.

Sonuçların görselleştirilmesi ve kullanılması

Elde edilen sonuçlar, makine öğrenmesindeki görselleştirme kütüphaneleri kullanılarak daha anlaşılır hale getirilebilir. Örneğin, temaların dağılımları veya belgeler arasındaki benzerlikler görsel grafiklerle temsil edilebilir. LDA analizi sonuçları, işletme kararlarını desteklemek veya metin verileri üzerindeki öngörülerini paylaşmak için kullanılır. Örneğin, müşteri yorumlarının analizi sonucunda restoranın hizmet kalitesini artırmak için önlemler alınabilir. LDA, büyük metin veri setlerini analiz etmek ve içerik analizi yapmak için güçlü bir araçtır. Bu aşamalar, LDA modelini uygulamak ve gizli temaları ortaya çıkarmak için izlenen temel adımları temsil eder.

Bir ürünün online yorumlarını analiz ederek, kullanıcıların ürünle ilgili ne düşündüğünü anlamak mümkün olabilir. Gizli Dirichlet Ayrımı (LDA) makine öğrenmesi ve metin madenciliği uygulamalarında çok büyük öneme sahip olan, en temel ve en popüler konu modelleme tekniklerinden biri haline gelen, belge türü gibi ayrık verileri modellemek ve bir belgeyi oluşturan konuları ortaya çıkarmak için kullanılan üretken bir grafik modeldir (Mei, Shen ve Zhai, 2007). Sonuç olarak, LDA, metin madenciliği ve doğal dil işleme alanında gizli temaların keşfi ve metin verilerinin daha iyi anlaşılması için önemli bir araçtır. Bu araç, araştırmacılar ve işletmeciler için metin verilerini daha verimli bir şekilde kullanmalarına yardımcı olabilir.

2.3.2. Kelime çantası (Bags of Words - BOW)

Metin madenciliği ve doğal dil işleme alanında önemli bir role sahip olan Bags Of Words (BoW) yöntemi metin belgelerinin analizinde ve sınıflandırılmasında yaygın olarak kullanılan bir tekniktir. Temel olarak bir belgedeki kelime sıklıklarını ve dağılımlarını temsil etmek için kullanılabilir. Bu temsil, metin verilerini sayısal bir formata dönüştürerek bilgisayar ortamında işlenebilir hale getirilir. Belge sınıflandırması gibi sistemler genellikle Kelime Çantası (BOW) yaklaşımı gibi kelime temsil yöntemlerini kullanır. Bu yöntemler kelimelerin anlamsal ilişkilerini ve kelimeler arasındaki tekrarları değerlendirir ve sınıflandırılan tüm dokümanların içerisindeki kelimeler üzerinde işlemler gerçekleştirilir (Çelenli, 2020).

Geleneksel kelime temsil yöntemlerinin aksine, modern kelime temsil yöntemleri son zamanlarda kelimelerin anlamsal ve sözdizimsel benzerliklerini daha net bir şekilde ortaya çıkarmıştır (Çelenli, 2020). Kelime anlamları ve kelime sırası gibi benzerliklerin yanı sıra kelimelerin etrafındaki bağlantıları arayıp çıkarmak ve bunları düşük boyutlu vektörler olarak ifade edilebilmektedir. Kelime çantası (BoW) yöntemi sınıflandırma çalışması yapılan metinleri terimlere ayırmaktadır. Oluşturulan terimler birer öznitelik olarak ele alınır. Sonraki aşamada ise her bir terimin tüm metin içinde geçme veya tekrarlanma sıklığı özniteliğin bir değeri olarak atanır. Böylece kategorik bir veri sayısal hale dönüştürülmektedir (Aksu ve Karaman, 2020). Kelime Çantası (BOW) modeli, doğal dil işlemede ve makine öğreniminde, bir belge veya cümle gibi bir metin gövdesini kelimelerden oluşan bir koleksiyon ve "torba" olarak temsil etmek için kullanılan bir tekniktir. BOW modelinde metnin yapısı ve grameri göz ardı edilir ve kelimelerin sıklığına vurgu yapılır. Bir kelime torbası oluşturma süreci, metni

simgeleştirmeyi, yani onu ayrı ayrı sözcüklere veya simgelere bölmeyi içerir. Bu kelimeler daha sonra toplanır ve bir matriste temsil edilir; burada her satır bir belgeye veya cümleye karşılık gelir ve her sütun, benzersiz bir kelimeyi temsil eder (Kowsari, Jafari Meimandi, Haydarşafa, Mendu, Barnes ve Brown, 2019).

Bags Of Words (BoW), metin belgelerini temsil etmek ve aynı zamanda analiz etmek üzere kullanılan bir yöntemdir. Bu yöntemde, metin belgesinde geçen her kelime bir vektörün bir bileşeni olarak kabul edilir ve bu kelimelerin sayılarına dayalı bir vektör oluşturulur. Bu vektör, metin belgesinin temsilini sağlar. Metin madenciliği uygulamaları, haber analizi, duygu analizi, belge sınıflandırma ve daha birçok alanda kullanılmaktadır. Örneğin, "lezzetli" kelimesi metinde 3 kez geçiyorsa ve "hızlı" kelimesi 2 kez geçiyorsa, ilgili öznitelik vektörü [3, 2, ...] şeklinde olabilir.

TF-IDF, bir kelimenin belirli bir belgedeki sıklığını (TF) ve tüm belgeler içinde o kelimenin görülme sıklığını (IDF) birleştirerek bir kelimenin önemini belirler. Metin belgelerindeki kelimelerin önemini belirlemek için kullanılan bir istatistiksel ölçüdür. Temel olarak, bir kelimenin bir belgedeki göreceli önemini değerlendirir ve bu kelimenin diğer belgelerde ne kadar yaygın olduğunu dikkate alır.

TF-IDF iki bileşenden oluşur:

Terim frekansı (TF): Bir belgedeki bir kelimenin sıklığını ifade eder. Yani, bir kelimenin bir belgedeki tekrar sayısının belge içindeki toplam kelime sayısına oranıdır. TF, bir kelimenin belge içinde ne kadar sık kullanıldığını gösterir. Örneğin, "lezzet" kelimesinin bir belgede 20 kez geçtiği ve belgedeki toplam kelime sayısının 200 olduğunu varsayalım. Bu durumda "lezzet" kelimesinin TF değeri 20/200 = 0.1 olacaktır.

Ters belge frekansı (IDF): Bir kelimenin belgedeki yaygınlığını ölçer. Yani, bir kelimenin tüm belgelerde ne kadar yaygın olduğunu ifade eder. Bu, bir kelimenin belge koleksiyonunda ne kadar benzersiz veya nadir olduğunu gösterir. IDF, belirli bir kelimenin belge koleksiyonundaki belge sayısının toplam belge sayısına oranının logaritması şeklinde hesaplanır. Bu sayede nadir kelimeler daha büyük bir ağırlığa sahip olur. Örneğin, "lezzet" kelimesinin 1000 belgeden oluşan bir koleksiyonda sadece 10 belgede bulunduğunu varsayalım. Bu durumda "lezzet" kelimesinin IDF değeri log(1000/10) = 2 olacaktır.

TF-IDF değeri, bir kelimenin belgedeki sıklığını (TF) ve belge koleksiyonundaki yaygınlığını (IDF) birleştirir. Bir kelimenin TF-IDF değeri, hem belge içinde sıklığına hem de genel koleksiyondaki yaygınlığına dayanarak o kelimenin önemini belirler. Yani, çok sık kullanılan kelimelerin (örneğin "ve", "veya") önemi düşük olurken, nadir kullanılan ve belirli bir belgeyi daha iyi tanımlayan kelimelerin önemi yüksek olur. TF-IDF, metin sınıflandırma, bilgi geri getirme ve metin madenciliği gibi alanlarda sıklıkla kullanılan bir öznitelik çıkarma yöntemidir.

Bu uygulamalar, büyük miktarda metin verisinin etkili bir şekilde işlenmesini gerektirir. İşte bu noktada BoW yöntemi devreye girer. Bags Of Words (BoW), metin belgelerini temsil etmek ve analiz etmek için kullanılan bir yöntemdir. BoW metin sınıflandırma, kelime yoğunluğu hesaplama ve dil modellemede kullanılır. Avantajlarından bahsetmek gerekirse yüksek hassasiyetli sonuçlar verir. Kullanımı hızlıdır ve metinler arası karşılaştırma yapmak kolaydır diyebiliriz.

Bilgin ve Şentürk (2017), yaptıkları çalışmada Türkçe ve İngilizce Twitter(X) üzerindeki veriler ele alındığında Doc2vec yöntemi ile analiz yapılmıştır. Dağıtık Bellek yöntemi ve Kelime Çantası (BoW) yöntemi etkili sonuç verme kapsamında karşılaştırılmıştır. Kelime Çantası yöntem olarak çok daha iyi sonuçlar vermiştir.

Çelenli (2020), çalışmasında Kelime Çantası yönteminin vektörlerindeki temsil etme yöntemlerini karşılaştırmıştır. Kelime kalıplama yöntemlerinde Doc2vec ile başarı oranlarının yüksek olduğu belirlenmiştir. TopWord2vec'in Türkçe ve İngilizce olarak sunulan argüman koleksiyonlarında Doc2vec'ten çok daha fazla hızlı olduğu belirlenmiştir.

Esen ve Özkan (2017), çalışmalarında TBMM tutanaklarına bakıldığında kelime çantası temsil yöntemleriyle parti ile özdeşleşme bakımından bir analiz çalışması yapılmıştır. Bunun için siyasi içerikli metinler kullanılarak analiz edilmiş ve buna bakılarak parti özdeşleşme durumu ölçülmüştür.

Çoban ve Özyer (2016), çalışmalarında Twitter (X)'dan elde ettikleri veri setini ön işleme adımlarından sonra kelime torbası yöntemi ile vektör oluşturmuşlardır. SVM, k-NN, NB, MNB, ME algoritma yöntemi ile sınıflandırma yapıp sonuçları değerlendirmişlerdir.

Sonuçlar değerlendirildiğinde BoW yöntemi için en başarılı algoritma %92 başarı oranında sınıflandırma yapan MNB olmuştur.

Kaçdıoğlu (2020), yaptığı çalışmasında OSB şüphesi olan çocukların en belirgin hareketlerini anlamak amacı ile ev ortamında kayda alınan video görüntülerini güncel derin öğrenme algoritmaları ve görsel kelime çantası yaklaşımı ile incelemiştir. Çalışmalarında, OSB'nin önemli olan tanı ölçütlerinden kendisini uyarıcı hareketlerin tanıması için Görsel Kelime Çantası yaklaşımı yapılmıştır. Farklı görsel kelimelerin sayısı (100, 200, 400, 600, 800, 1000) ile 5 (beş) öznitelik tanımlayıcı ve 3 (üç) sınıflandırıcı yöntemi kullanarak deneyler yapılmış ve en iyi sonuç için 200 görsel kelime, HOF tanımlayıcı ve MLP sınıflandırıcısında %80 doğruluk ile alınmıştır. Sonuç olarak videolardaki hareketler %80 oranda doğru sınıflandırılmıştır.

2.4. Sınıflandırma Amaçlı Makine Öğrenme Algoritmaları

Bu algoritmalar verileri belirli kategorilere veya sınıflara atamak için kullanılır. Bir modelin eğitildiği ve daha sonra yeni veriler için tahminlerin yapıldığı süreçte çeşitli makine öğrenimi algoritmaları kullanılır.
Çok Katmanlı Algılayıcılar (Multi-Layer Perceptron), (MLP), yapay sinir ağlarında bulunan yaygın türlerinden biridir. Genellikle sınıflandırma ve regresyon problemlerini çözmek için kullanılır. İnsan beynine benzer bir şekilde modellenmiş olan bu yapılar bilgisayarlara öğrenme yeteneği kazandırma amacıyla kullanılmaktadırlar (Öztemel, 2003). MLP, çok katmanlı bir yapıya sahip olduğu için öğrenme modellerinin temeli denilebilir. Bu yapı sayesinde veri üzerinde çok yüksek seviyede özelliklerin öğrenilmesini sağlamaktadır. MLP, çok fazla sınıfı içeren karmaşık sınıflandırma problemleri ve regresyon problemlerini çözebilir. Örneğin, görüntü ve ses tanıma yanı sıra doğal dil işleme gibi alanlarda sınıflandırma problemi, finansal verilerin analiz edilmesi gibi alanlarda regresyon problemleri kullanılabilir.

Destek vektör makineleri (SVM), makine öğrenimi alanında sınıflandırma ve regresyon problemlerini çözmek amacıyla kullanılan güçlü bir algoritmadır. Sınıflandırma probleminde temel amaç, sınıfları birbirinden ayırmayı sağlayan optimal ayırma hiper düzleminin elde edilmesi ve bu sayede, farklı sınıflarda bulunan destek vektörleri ile arasındaki uzaklık maksimum edilmektedir (Ayhan ve Erdoğmuş, 2014). Sınıflandırma

arasında bir hiperdüzlem bulma SVM'nin temel prensibi denilebilir. Hiperdüzlem, veri setindeki sınıflar arasında olan ayrımı belirleyen matematiksel konsepttir. SVC nesne tanıma, görüntü sınıflandırma problemi, bir hisse senedinin fiyatını tahmin etme gibi sürekli değerlerin tahmin edilmesinde regresyon problemi kullanılabilir.

Lojistik Regresyon (LogisticReg), sınıflandırma problemlerini çözmek ve olayın gerçekleşme olasılığını tahmin etmek amacıyla kullanılan bir makine öğrenimi algoritmasıdır. İki sınıflı (binary) sınıflandırma problemlerinde yaygın olarak kullanılır. Lojistik Regresyon bir e-postanın reklam içeriği olup olmadığını sınıflandırma problemi, bir hastanın bir hastalığa sahip olup olmadığını tahmin etme gibi regresyon problemlerinde kullanılabilir.

Random Forest (RF), sınıflandırma ve regresyon problemlerini çözmek için kullanılan bir makine öğrenimi algoritmasıdır. Yüksek doğruluk ve tahmin performansı sağlamaktadır. Hem sınıflandırma hem de regresyon problemlerinde etkili bir şekilde kullanılabilir. Ensemble learning (ensemble öğrenimi) kategorisine dahildir ve birden çok karar ağacının (decision tree) bir araya getirilmesiyle oluşturulur. Random Forest, bir e-ticaret şirketi, müşterilerin "reklamı tıkladı" veya "reklamı tıklamadı" durumu için sınıflandırma problemi, bir araç üretim firması, araç satış fiyatlarını tahmin etme problemlerinde regresyon kullanılabilir.

Karar Ağacı (Decision Tree), basit ve anlaşılması kolay bir yapıya sahiptir. Karar ağaçları, ağaç yapısında karar düğümleri ve sonuç düğümleri içerir. Her bir karar düğümü, bir özellik veya değişken üzerinde bir koşulu temsil eder ve her bir sonuç düğümü, sınıflandırma veya regresyon içerir. Tek başına çok karmaşık ilişkileri modellemekte zorlanabilir ve uyum riski taşıyabilir. Bu nedenle, genellikle ensemble yöntemlerine (örneğin, Random Forest veya Gradient Boosting) entegre edilerek daha güçlü tahmin modelleri oluşturmak için kullanılır. Örnek olarak e-ticaret şirketi, bir müşterinin ürünü satın alıp almama olasılığını tahmin etmek için müşteri verileri ve satın alma geçmişi kullanılabilir. Verileri müşterinin yaş, cinsiyet, geçmiş alışveriş sayısı, indirim kullanma alışkanlığı gibi bilgileri sınıflandırabilir.

Stokastik gradyan inişi (SGD), makine öğrenimi alanlarında yaygın olarak kullanılan bir optimizasyon algoritmasıdır. Amacı, bir kayıp fonksiyonunu (loss function) en aza indiren model parametrelerini bulmaktır. Gradyan, bir modelin parametrelerine göre kayıp fonksiyonunun türevi olarak düşünülebilir. SGD'nin adı, her bir adımda bir örnek veya bir

alt küme kullanarak gradyan tahmini yapmasından gelir. Bu gradyan tahmini, modelin parametrelerini güncellemek ve kayıp fonksiyonunu en aza indirmek için kullanılır. SGD'nin temel avantajı, büyük veri setleri üzerinde hızlı ve etkili bir şekilde çalışabilmesidir. Ayrıca, SGD'nin esnek bir algoritma olması ve çeşitli makine öğrenimi modellerinde kullanılabilmesi de avantajları arasındadır.

Boosting yöntemleri; Sınıflandırma, regresyon ve sıralama gibi farklı makine öğrenimi görevlerinde kullanılır.

Bu yöntemler GradientBoost (Gradient Boosting), AdaBoost (Adaptive Boosting), XGBOOST (Extreme Gradient Boosting) ve LightGBM'den oluşmaktadır. Algoritmaların her biri, zayıf öğrenicileri (genellikle karar ağaçları olarak adlandırılan basit modeller) bir araya getirerek güçlü bir tahmin modeli oluşturmayı hedefler. Aralarında bazı farklılıklar vardır.

GradientBoost (Gradient Boosting): Bu algoritma, önceki ağaçların tahmin hatalarını azaltarak yeni ağaçlar ekler. Genellikle sınıflandırma ve regresyon problemleri için kullanılır.

AdaBoost (Adaptive Boosting), değişen koşullara uyum sağlayacak bir şekilde zayıf öğrenicileri bir araya getirerek güçlü bir tahmin modeli oluşturan bir Boosting algoritmasıdır. Her iterasyonda yanlış sınıflandırılan örneklerin ağırlıklarını artırır ve bu ağırlıklarla yeni bir öğrenici oluşturur. Tahmin, tüm öğrenicilerin ağırlıklı ortalamasını alarak yapılır. Genellikle sınıflandırma problemleri için kullanılır.

XGBoost (Extreme Gradient Boosting), Gradient Boosting algoritmasının hız, performans ve doğruluk açısından iyileştirilmiş bir versiyonudur. Büyük veri kümeleri ve yüksek boyutlu uzaylarıyla çalışırken etkilidir. Sınıflandırma, regresyon ve sıralama gibi çeşitli problemleri çözmek için kullanılır.

LightGBM (Light Gradient Boosting Machine): LightGBM, Microsoft tarafından geliştirilen hızlı ve ölçeklenebilir bir Gradient Boosting kütüphanesidir. Büyük veri kümeleri üzerinde hızlı bir şekilde çalışabilir ve yüksek doğruluk sağlar. Sınıflandırma, regresyon ve sıralama gibi çeşitli makine öğrenimi problemlerini çözmek için kullanılır.

3. UYGULAMA ÇALIŞMASI

3.1. Araştırmanın Amacı

Bu araştırmada bir Çin restoranı hakkında yapılan müşteri yorumları kullanılarak, makine öğrenmesi yöntemleriyle restoranın hizmet kalitesini artırmak, müşteri memnuniyetini sağlamak ve rekabet ortamında daha etkili stratejiler geliştirmek amaçlanmıştır. Bu amaçla, online yemek inceleme sitelerinden ve internet tavsiye sitelerinden toplanan müşteri yorumları kullanılarak metin madenciliği yöntemleri uygulanmıştır. Bu yöntemlerden Gizli Dirichlet Ayrımı (LDA) modelinin kullanılması öne çıkmaktadır. LDA modeli, yorumları farklı konulara ayırarak belge terim matrisine dönüştürmüş ve her yorumu kelime vektörleriyle temsil etmiştir. Elde edilen temalar ve terim sıklıkları incelenerek restoranın hizmet kalitesi, müşteri memnuniyeti ve stratejileri hakkında önemli öngörüler elde edilmiştir. Bu öngörüler, restoran yönetimi tarafından stratejik kararlar alınmasına ve işletmenin gelişimine katkı sağlamaya yönelik olabilecektir.

3.2. Örneklem

Restoran çeşitliliğinin çok olması ve yemek kültürünün çok geniş olması sebebiyle online sitelerde veri büyüklüğü çok fazla olmaktadır. Araştırmamızda daha spesifik olması açısından bir Çin restoranı tercih edilmiştir. Veriler, internet siteleri ve yemek sipariş platformlarından tarama yolu ile elde edilen 303 adet yorumdan oluşmaktadır.

3.3. Araştırmanın Yöntem ve Kapsamı

Restoranın hizmet kalitesini iyileştirmek, müşteri memnuniyetini ve rekabet ortamında restoranın stratejilerini gelişmek amacıyla Çin restoranına ait müşteri yorumlarını online yemek inceleme sitelerinden ve internet tavsiye siteleri (Getir, TripAdvisor vs.) aracılığıyla toplanmıştır. Toplanan veriler temizlenip ve ön işleme yapılmıştır. Metin madenciliği için uygun bir formata dönüştürülerek, gereksiz karakterler ve noktalama işaretleri kaldırılmış, büyük/küçük harf dönüşümü yapılıp ve etkisiz kelimeler kaldırılmıştır.

Metin verilerini konulara ayırmak ve temsil etmek için LDA modeli kullanılmıştır. Python'da "gensim" veya "scikit-learn" gibi kütüphanelerle LDA modeli eğitilmiştir. Kelime çantası (Bags of Words) yöntemiyle yorumlar bir belge terim matrisine (document-term matrix) dönüştürülerek, her belge kelime vektörleriyle temsil edilmiştir. LDA'nın çıkardığı temalar ve Bags of Words'un sunduğu terim sıklıkları incelenerek bir takım öngörüler elde edilmiştir.

3.4. Verilerin Analizi ve Yorumlanması

Çalışmada ön işlemler, veri setlerini okuma, makine öğrenmesi yöntem ve algoritmalarından oluşan tüm işlemler Python programlama dili kullanılarak yapılmıştır. Tüm yöntemler için gerekli kütüphaneler Python ortamına eklenmiştir.

Bu kütüphaneler ;

- Pandas
- Numpy
- Matplotlib.pyplot
- Seaborn
- Collections, Counter
- WordCloud
- Gensim
- pyLDAvis
- Corpora, models, similarities
- Locale
- Time
- Warnings
- Imissingno

Python'da veri analizi ve görselleştirme işlemleri için kullanılan temel kütüphaneleri içe aktarmaktadır.

Pandas, Python'daki veri analizi için yaygın olarak kullanılan pandas kütüphanesini içe aktarmaktadır. pd kısaltması, genellikle bu kütüphaneye yapılan çağrılarda kullanılır. pandas, veri çerçeveleri (dataframes) gibi yüksek performanslı veri yapıları sunar.

Numpy, Bilimsel hesaplamalar ve çok boyutlu dizilerle çalışmak için kullanılan numpy kütüphanesini içe aktarır. np kısaltması, genellikle bu kütüphaneye yapılan çağrılarda kullanılır. numpy, vektör ve matris operasyonları gibi temel matematiksel işlemleri hızlı bir şekilde gerçekleştirmek için kullanılır.

Matplotlib.pyplot, veri görselleştirmesi için kullanılan matplotlib kütüphanesinin pyplot modülünü içe aktarır. plt kısaltması, genellikle bu kütüphaneye yapılan çağrılarda kullanılır. matplotlib ile çizimler, grafikler ve görselleştirmeler oluşturulabilir.

Missingno, eksik verileri daha görsel ve anlamlı bir şekilde analiz etmek ve göstermek için kullanılan bir Python kütüphanesidir. Bu kütüphane, eksik verileri incelemek ve veri setindeki eksik değerlerin desenlerini anlamak için kullanıcıya yönelik grafikler oluşturulur.

Gensim kütüphanesi, Python programlama dilinde metin madenciliği ve doğal dil işleme (NLP) uygulamaları adına kullanılan bir kütüphanedir. Özellikle büyük metin koleksiyonları üzerinde çalışan ve kelime gömme (word embedding) gibi gelişmiş metin madenciliği tekniklerini uygulayan bir kütüphanedir.

Locale, bilgisayar ortamındaki yerel ayarları kontrol etmek ve yönetmek için kullanılan bir Python modülüdür. Bu modül, dil, para birimi, tarih ve saat formatları gibi yerel ayarları programınızda kontrol edilmesini sağlar. Python'da locale modülü, bilgisayarınızdaki yerel ayarları çekmek ve değiştirmek için bir araç olarak hizmet verir.

PyLDAvis.gensim, kütüphanesi, Latent Dirichlet Allocation (LDA) modelinin görselleştirilmesi için kullanılan bir Python kütüphanesidir. LDA, bir belge koleksiyonundaki temaları ve bu temaların belgeler arasındaki dağılımını modelleyen bir olasılık temelli bir konu modelleme tekniğidir.

Makine öğrenme algoritmaları için kütüphaneler;

- Optuna
- Sklearn.ensemble (Ada Boost Classifier, Extra Trees Regressor, Random Forest Classifier, Gradient Boosting Classifier)
- Sklearn.feature_extraction.text (CountVectorizer, Tfidf Transformer, Tfidf Vectorizer)
- Sklearn.gaussian_process (Gaussian Process Classifier)
- Sklearn.gaussian_process.kernels (RBF)
- Sklearn.linear_model (Logistic Regression, Ridge, SGD Classifier)
- Ssklearn.metrics (accuracy_score, classification_report, confusion_matrix, Confusion Matrix Display, precision_recall_fscore_support, RocCurveDisplay, roc_curve, roc_auc_score)
- Sklearn.model_selection (train_test_split)
- Sklearn.naive_bayes (BernoulliNB, GaussianNB, MultinomialNB)
- Sklearn.neighbors (Kneighbors Classifier)
- Sklearn.neural_network (MLPClassifier)
- Sklearn.pipeline (Pipeline)
- sklearn.tree (DecisionTreeClassifier)
- Sklearn.svm (SVC)
- Xgboost (XGBClassifier)
- Lightgbm (LGBMClassifier)
- Catboost (CatBoostClassifier)
- Tqdm (tqdm)

Optuna, açık kaynaklı bir hiperparametre optimizasyon çerçevesidir. Hiperparametre optimizasyonu, makine öğrenimi algoritmalarının veya model mimarilerinin performansını artırmak için kullanılan parametrelerin (hiperparametrelerin) en iyi değerlerini bulma sürecidir. Optuna, bu optimizasyon sürecini sistemli bir şekilde en iyi hiperparametre setini bulmak için tasarlanmış bir Python kütüphanesidir. Bir dizi farklı makine öğrenimi çerçevesi ve kütüphanesi ile kullanılabilir. Bu, farklı modelleri ve algoritmaları kullanmak isteyen kullanıcılar için esnek bir çözüm sunabilir. Hiperparametre setinin performansını değerlendirmek için minimum sayıda denemeyi gerçekleştirmek üzere tasarlanmıştır. Bu, optimizasyon sürecini daha verimli hale getirir. Hiperparametre optimizasyonunu hızlandırmak için birden çok bilgisayar veya işlemci kullanma yeteneğine sahiptir.

Optuna, optimizasyon sürecinin sonuçlarını ve hiperparametre setlerini görselleştirmek için kullanışlı araçlar içerir. Özellikle derin öğrenme gibi hiperparametre hassas modellerle çalışırken en iyi parametre ayarlarını bulmak için yaygın olarak kullanılan bir araçtır.

Bu kütüphaneler ve modüller, makine öğrenimi ve veri bilimi projelerinde yaygın olarak kullanılan araçlardır. Optuna, bu araçları daha etkili ve performanslı hale getirmek için hiperparametre optimizasyonu sağlar.

Doğal dil işleme ve Makine Öğrenmesi Kütüphaneleri;

- Nltk (tokenize)
- nltk.corpus (stopwords)
- nltk.probability (FreqDist)
- nltk.tokenize (word_tokenize, sent_tokenize)
- nltk.stem (WordNetLemmatizer, PorterStemmer)

NLTK, doğal dil işleme görevlerini gerçekleştirmek için kullanılan geniş bir Python kütüphanesidir.

Tokenize, cümle veya kelimeleri ayırmak için tokenize modülünü içe aktarır. Metni anlamak için genellikle cümle veya kelime seviyesine bölme işlemi yapılır.

Stopwords, NLTK içindeki stopwords modülünü içe aktarır. Stopwords, genellikle metin madenciliği veya NLP uygulamalarında analiz dışı bırakılan yaygın kelimelerdir.

FreqDist, kelime frekanslarını hesaplamak için kullanılır. Bir belgedeki kelimelerin görülme sıklıklarını sayar.

PorterStemmer, kelimeleri köklerine indirgemek için kullanılır. Stemming, kelimenin kökünü bulma işlemidir.

Bu modüller, metin madenciliği, metin analizi, duygu analizi ve benzeri NLP görevlerinde kullanılır. Çalışmamızdaki verilerdeki kelimelerin frekansını analiz etmek, stop words'leri

temizlemek, kelime köklerine indirgemek veya lemmatize etmek gibi işlemlerde bu modüller kullanılabilir. Kütüphaneler tanımlanarak konu modelleme yöntemleri gerçekleştirilerek Visual Studio ortamında çalıştırılmıştır.

Veriler Çizelge 3.4'de dosyadan Visual Studio ortamında bulunan Python'a aktarılmıştır.

Çizelge 3.1. Visual Studio ortamına aktarılan verileri

	Müşteri Adı	Müşteri Yorumu	VERİLEN PUAN
0	Müşteri 1	Güler yüzlü personel mekan harika yemeklerde b...	5
1	Müşteri 2	Japonya'da yaşamış arkadaşımızla gittik. Şu an...	5
2	Müşteri 3	Bu restoranı normalde severdim ancak geçtiğimi...	1
3	Müşteri 4	Eskiden Ankara'daki en sevdiğim mekanlardan bi...	1
4	Müşteri 5	Arkadaşımla yoğun bir iş günü sonrasında gidip...	4
5	Müşteri 6	Çarşamba günü akşam 17:30 da arkadaşımla Armad...	1
6	Müşteri 7	Dün gop taki Quick China ya gittik tek kelime ...	1
7	Müşteri 8	Kapı önünde biri var rezervasyon alıyor sözde,...	4
8	Müşteri 9	Daha önce de birçok kez gittiğim restauranta s...	1
9	Müşteri 10	Personellerin kabalık ve ukalağından mı bahset...	1
…	…	…	…

İlk aşamada, yorumların metin verileri çeşitli online sitelerden toplanmış ve ön işleme yapılmıştır. Veriler dosyadan okutulup, bu adımda kelimeler parçalanmıştır. Gereksiz kelimeler çıkarılarak, özel karakterler ve sayılar kaldırılmış, tüm harfler küçültülmüş ve kelime kökleri çıkartılmıştır. Daha sonra veriler incelenerek, Çizelge 3.5'de olduğu gibi yorum uzunluğu, kelime sayısı ölçüldükten sonra veriler düzenlendi.

Çizelge 3.2. Verilerin yorum uzunluğu ve kelime sayısı

S.NO	Müşteri Adı	Müşteri Yorumu	Verilen Puan	Yorum Uzunluğu	Kelime Sayısı
0	Müşteri 1	Güler yüzlü personel mekan harika yemeklerde b...	5	142	20
1	Müşteri 2	Japonya'da yaşamış arkadaşımızla gittik. Şu an...	5	236	33
2	Müşteri 3	Bu restoranı normalde severdim ancak geçtiğimi...	1	488	68
3	Müşteri 4	Eskiden Ankara'daki en sevdiğim mekanlardan bi...	1	255	30
4	Müşteri 5	Arkadaşımla yoğun bir iş günü sonrasında gidip...	4	218	32
….	….	….	….	….	….

Çizelge 3.3. Yorumların kelime sayısına ilişkin betimsel istatistikler

Kelime sayısına ilişkin betimsel istatistikler	
Minimum	1
Maksimum	212
Ortalama	38,42
Standart Sapma	30,49
Çarpıklık Değeri	2,36
Basıklık Değeri	7,20

Bu istatistikler, müşteri yorumları veya metin verileri gibi bir veri setinin kelime sayısı dağılımını anlamak ve analiz etmek için kullanılır. Bu sayede, veri setinin genel özellikleri hakkında bilgi edinilir ve olası eğilimler veya anormallikler belirlenir. Özellikle bir metnin uzunluğunun önemli olduğu durumlarda (örneğin, makaleler, yorumlar, metin analizi vb.), bu istatistikler metnin yapısal özellikleri hakkında fikir verir ve metinler arasında karşılaştırma yapmayı sağlar. Yorumların ortalama kelime sayısı 38,42'dir. Standart sapmanın 30,49 olması kelime sayılarının geniş bir dağılıma sahip olduğunu gösterir. Pozitif bir çarpıklık değeri, dağılımın sağa doğru çarpık olduğunu gösterir, yani ortalamadan sağa doğru uzanan uzun kuyruğa sahiptir. Yüksek bir basıklık değeri, dağılımın normal dağılıma göre daha sivri olduğunu ve uç değerlerin daha sık görüldüğünü gösterir.

Çizelge 3.4. Kelime sayılarının memnuniyet düzeylerine göre dağılımı

MEMNUNİYET	YORUM SAYISI	KELİME SAYISI	KELİME SAYISI YÜZDESİ	ORTALAMA KELİME SAYISI	YORUM UZUNLUĞU
OLUMSUZ	96	4489	38,56%	46,76	31879
OLUMLU	207	7154	61,44%	34,56	51164

Tabloya göre, olumsuz ve olumlu memnuniyet düzeylerine sahip yorumların kelime sayıları ve dağılımları arasında farklılıklar olduğu görülmektedir. Olumlu memnuniyet düzeyine sahip yorumların sayısı ve kelime sayısı, olumsuz memnuniyet düzeyine sahip yorumlardan ve kelime sayılarından daha fazladır. Ancak olumsuz memnuniyet düzeyine sahip yorumların ortalama kelime sayısı, olumlu memnuniyet düzeyine sahip yorumlardan daha yüksektir. Bu durumda, olumsuz yorumların daha uzun ve detaylı olduğu söylenebilir.

Verilerin analiz edilmesi ile müşteri memnuniyetini anlamak ve işletmenin ürün veya hizmetler hakkındaki algısını değerlendirmek için kullanılabilir. Örneğin, "OLUMLU"

memnuniyet kategorisinde daha yüksek kelime sayısı ve yorum uzunluğu, işletmenin müşteriler üzerinde olumlu bir etkiye sahip olduğunu gösterebilir.

Memnuniyet düzeyleri ile ortalama kelime sayısı bakımından anlamlı fark olup olmadığı test edilmiştir. Bu test öncesi memnuniyet düzeylerinin ortalama kelime sayıları bakımından normal dağılıma uygunluğunu Shapiro-Wilk ile test ettiğimizde,

OLUMSUZ Memnuniyet Düzeyleri İçin Shapiro-Wilk Testi:

Test İstatistiği: 0,8168
p-değeri: 1,4513e-09 Bu durumda, p-değeri 0.05'ten küçük olduğu için normal dağılıma uymaz.

OLUMLU Memnuniyet Düzeyleri İçin Shapiro-Wilk Testi:

Test İstatistiği: 0,7650
p-değeri: 6,8458e-17 . Bu durumda, p-değeri 0.05'ten küçük olduğu için normal dağılıma uymamaktadır.

Bu sonuçlara dayanarak, her iki memnuniyet düzeyi için de kelime sayılarının normal dağılıma uymadığı sonucuna varırız. Memnuniyet düzeyleri ile ortalama kelime sayısı arasında anlamlı bir farkın olup olmadığını test etmek için veriler normal dağılıma uymuyorsa, non-parametrik testlerden biri olan Mann-Whitney U Testi uygulanabilir. Bu test, iki bağımsız grup arasında medyan değerlerinin eşit olup olmadığını test etmek için kullanılır ve veriler normal dağılıma uymadığında güvenilir sonuçlar verir.

Mann-Whitney U Testi;

Mann-Whitney U Testi İstatistiği: 11487,5
p-değeri: 0.0287

p-değeri (0.0287) 0.05 anlamlılık düzeyinden küçük olduğu için iki grup arasında istatistiksel olarak anlamlı bir fark olduğunu gösterir. Bu farkın rastgele oluşmadığı ve istatistiksel olarak anlamlı olduğu söylenebilir. Bu durumda, gruplar arasındaki farkın

gerçek ve anlamlı olduğu kabul edilir. Memnuniyet düzeyleri ile ortalama kelime sayısı bakımından farklılıklar, müşterilerin deneyimlerini ifade etme şekilleri ve işletme ile olan etkileşimleriyle ilgili olduğu söylenebilir.

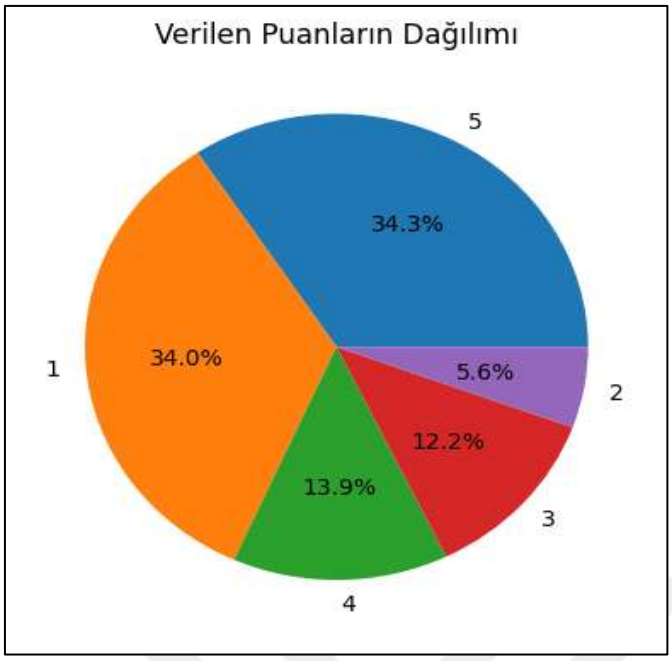
Olumsuz Yorumlar: Olumsuz memnuniyet düzeyine sahip yorumların ortalama kelime sayısının, olumlu memnuniyet düzeyine sahip yorumlardan daha fazla olması, olumsuz yorumların daha uzun ve detaylı olduğunu gösterebilir. Müşterilerin olumsuz deneyimlerini daha ayrıntılı bir şekilde ifade etme eğiliminde olmaları bu farklılığın bir nedeni olabilir. Olumsuz deneyimlerin ifade edilmesi, müşterilerin dikkatini çekmek ve sorunları çözmek için işletmenin dikkatini çekmek için bir yol olabilir.

Olumlu Yorumlar: Olumlu memnuniyet düzeyine sahip yorum sayısının daha yüksek olması, işletmenin genel olarak olumlu bir algıya sahip olduğunu gösterebilir. Müşterilerin olumlu deneyimlerini paylaşma eğiliminde olmaları, işletmenin sunduğu ürün veya hizmetlerden memnun kaldıklarını ve bunu başkalarıyla paylaşmak istediklerini gösterebilir. Ayrıca, olumlu yorumların sayısının fazla olması, işletmenin müşteri memnuniyetini artırmak için başarılı bir strateji uyguladığını da gösterebilir. Ortalama kelime sayısının olumsuz yorumlara göre daha az olması ise yorumların kısa ve öz bir şekilde ifade edildiği söylenebilir.

Çizelge 3.5. Verilen puanların betimsel istatistikler

BETİMSEL İSTATİSTİKLER	
VERİLEN PUAN	
Frekans	303
Ortalama	3,69
Standart Sapma	1,35
Minimum Değer	1
İlk Çeyrek (25%)	3
Medyan (50%)	4
Üçüncü Çeyrek (75%)	5
Maksimum Değer	5

Her bir yorum sahibinin verdiği puanların ortalaması 3.69, standart sapması ise 1.35’dir.



Şekil 3.1. Yorumlara verilen puanların dağılımı

Müşterilerin %34'ü 1 puan, % 5,6'sı 2 puan, %12,2'si 3 puan, %13,9'u 4 puan ve %34,3'ü de 5 puan vermiştir. Grafik çizme işlemleri yapılarak, "VERİLEN PUAN" sütunundaki değerlerin dağılımını ve bu puanların yüzde dağılımı görsel olarak gösterildi.

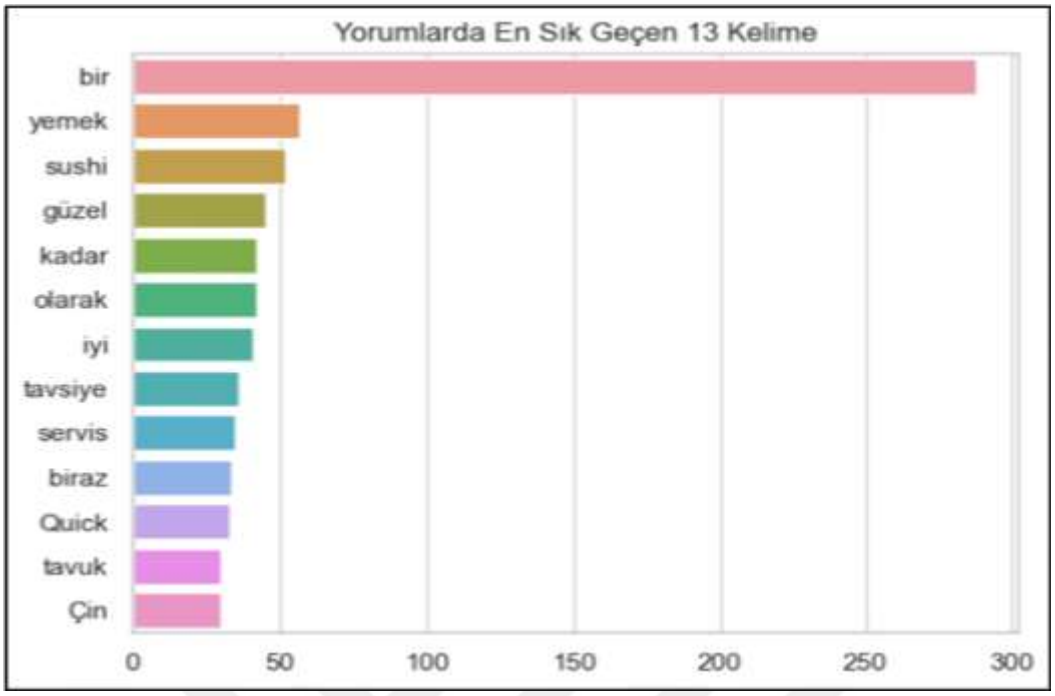
Çizelge 3.6. Memnuniyet düzeylerine göre puanların dağılımı

VERİLEN PUAN	1	2	3	4	5
OLUMLU	-	-	-	104	103
OLUMSUZ	42	17	37	-	-

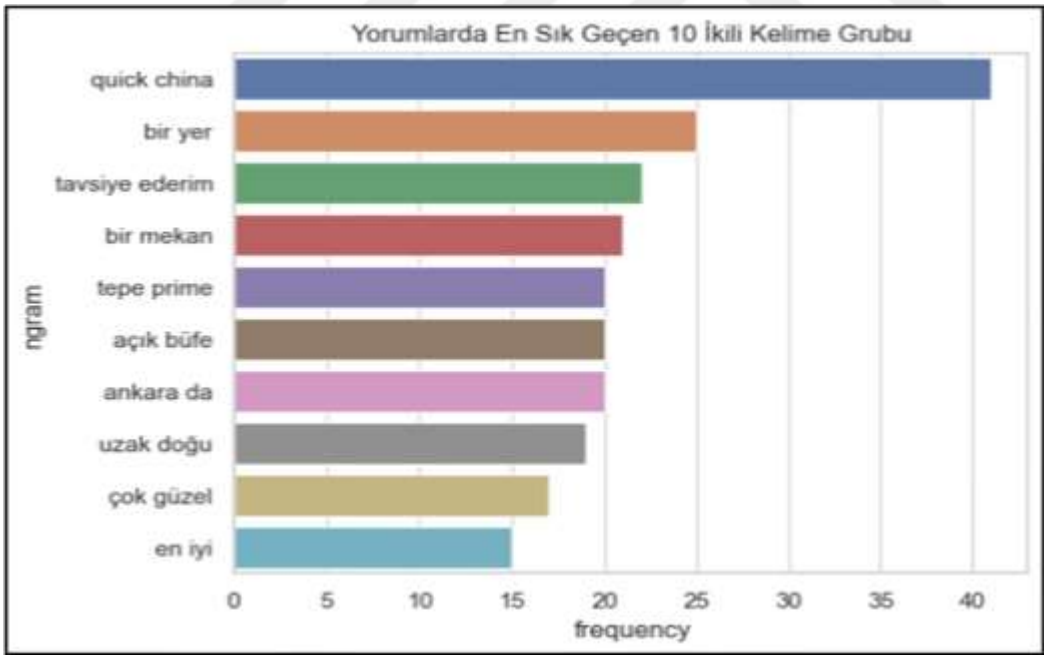
Olumlu grup için, En yüksek memnuniyet seviyesi olan 4 ve 5 puanı 207 katılımcının verdiği görülüyor.

Olumsuz grup için, 1, 2 ve 3 puanı 96 kişinin verildiği görülüyor.

Çalışma kapsamında verilen yorumların incelenebilmesi yorumlar ilk aşamada ön incelemeye tutulmuşlardır. Ön inceleme aşamasında verilen yorumların kelime uzunlukları ve bunlarla birlikte verilen puanlar arasındaki ilişki izleyen şekillerde görülebilmektedir. Yorumlarda kullanılan ve Türkçe'de çok sıklıkla kullanılan kelimeler haricindeki en sık kullanılan kelimeler aşağıdaki şekildedir:



Şekil 3.2. Yorumlarda en sık bulunan kelimeler



Şekil 3.3. Yorumlarda en sık bulunan ikili kelime grubu

Yorumlarda en sık geçen kelimeler “ve”, “de”, “için” gibi Türkçe'de çok yaygın kullanılan kelimelerdir. Bunlar haricinde olanların frekanslarına baktığımızda Şekil 3.8’de görebiliriz.

Şekil 3.4. Müşteri yorumlarından oluşturulan kelime bulutu

Kelime bulutları, karmaşık veri setlerini görsel olarak özetleyerek, bir ürün veya hizmetin müşteri deneyimini daha anlaşılır ve erişilebilir hale getirir. Genel ayarlamalar yapıldıktan sonra temel incelemeler yapılır ve incelemeleri yapmak için eklenen sütunlar silinir. Kelimeler ve işaretler ayarlanır. Büyük/küçük harf ve noktalama işaretleri gibi temizlemeler yapılır. Türkçe'de en sık kullanılan kelimeler 've', 'ile','de' gibi kelimeler çıkartılır.

3.5. LDA Yöntemi Analizi

Analiz kısmında öncelikle Latenet Dirichlet Allocation (LDA) yöntemi ile verilen yorumların oluşturulacak başlıklar altında gruplanması sağlanmaya çalışılmıştır. LDA algoritması temel olarak Bayesyen olasılık modeli üzerine kuruludur. Yorumlardaki kelimelerin konu başlıklarıyla ilişkilendirilmesini sağlar. Her yorumun baskın konusunu, katkı yüzdesini ve konuya ait anahtar kelimeleri belirlemek amacıyla kullanılmıştır. Algoritmada her bir yorumda geçen kelimenin, oluşturulacak olan bir konu başlığına ait olma olasılığının belirlenmesini amaçlar. Algoritmanın uygulanabilmesi amacıyla yapılan yorumlar kelimelere bölünmüş, küçük/büyük harf uyumu sağlanmış ve Türkçe'de sıklıkla kullanılan kelimeler çıkartılarak gerekli uygulamalar yapılmıştır. Sonrasında her bir kelimenin ait olabileceği konu başlığı olasılıkları elde edilerek baskın olan konu başlığı grubu yorumlara atanmıştır.

LDA, Çin restoranına yapılan müşteri yorumlarını analiz etmek için güçlü bir araç olabilir. Örneğin, "yemek kalitesi," "hizmet," "ambiyans" gibi temaları keşfedebilir ve bu temaların müşteri deneyimine nasıl katkıda bulunduğunu anlayabiliriz

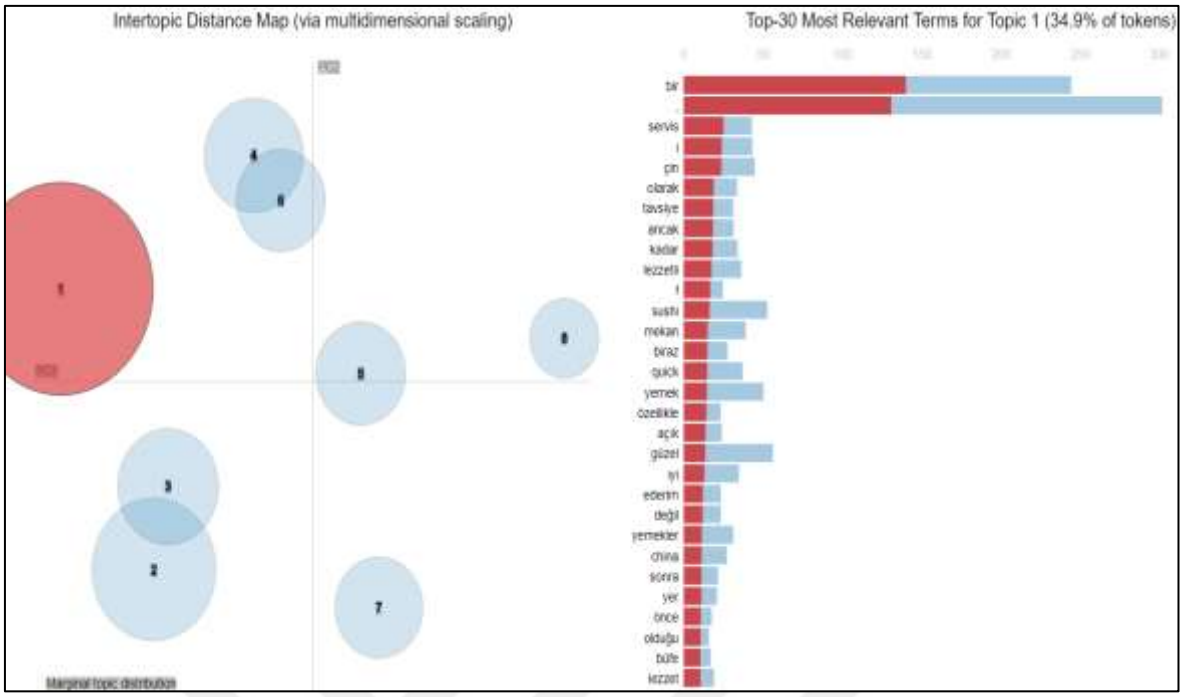
LDA sonuçlarına dayalı olarak, Çin restoranının müşteri memnuniyetini etkileyen faktörleri daha iyi anlayabiliriz. Özellikle hangi temaların daha fazla vurgulandığını ve müşteri yorumlarının hangi konularda odaklandığını gözlemleyebiliriz. Ayrıca müşterilerin yaptıkları yorumların yanında restorana verdikleri puanlarda düşünceleri hakkında bilgi sahibi olmamızı sağlamaktadır.

Çizelge 3.7. Yorumların baskın konuya göre sınıflandırılması

Baskın Konu Numarası	Baskın Konu Adı	Frekans	Yüzde	Birikimli Yüzde
1	Genel restoran deneyimi	52	17,16%	17,16%
2	Memnuniyet ve değişim	25	8,25%	25,41%
3	Sushi deneyimi	35	11,55%	36,96%
4	Yemekler ve mekan atmosferi	70	23,10%	60,07%
5	Lezzet ve Mekan	18	5,94%	66,01%
6	Qick Chinadaki yemek deneyimi	49	16,17%	82,18%
7	Genel mekan ve hızlı servis durumu	26	8,58%	90,76%
8	Uzak doğu mutfağı ve menü çeşitliliği	28	9,24%	100,00%

Öncelik sırasını belirlemek için, genellikle bir baskın konunun frekansı ve birikimli yüzdesi gibi faktörler göz önünde bulundurulur. Örneğin, daha yüksek bir frekansa veya birikimli yüzdeye sahip olan baskın konular genellikle daha önemli kabul edilir. Buna göre, hakkında en çok yorum yapılan baskın konu restoranın yemekleri ve mekân atmosferidir. İkinci sırada ise genel restoran deneyiminden bahsedilmektedir. Üçüncü olarak ise bahse konu restoranın yemek deneyiminden bahsedilmektedir. Diğer konular da önemlidir ancak daha düşük bir frekans veya birikimli yüzdeye sahip oldukları için öncelik sıralamasında daha aşağıda yer alabilirler. Öncelik sırası, belirli bir durum veya gereksinime göre değişebilir. Bu nedenle, bir restoranın belirli bir odak noktası veya stratejik hedeflerine göre öncelikler belirlenebilir. Baskın Konu (Dominant_Topic) sütunundaki sayılar her bir müşteri yorumu için belirlenen en baskın konu numaralarını temsil eder. Baskın konu içeriğindeki kelimeleri ve anahtar kelime özeti şu şekildedir.

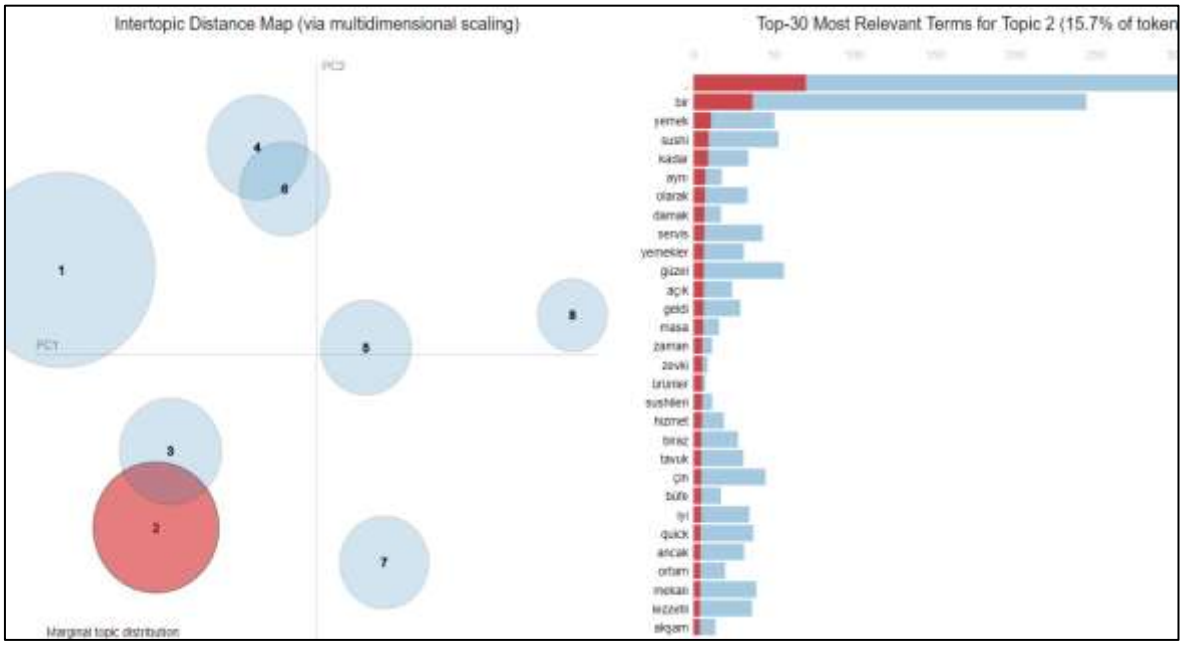
Konu 1: Anahtar Kelimeler: sushi, çin, servis, tavsiye, lezzet, güzel
Örnek yorum: "Japonya'da yaşamış arkadaşımızla gittik. Şu anda..."



Şekil 3.5. LDA modelinde birinci grup baskın konu

Konu 2: Anahtar Kelimeler: güzel, yemek, sushi, yemeği, mekan, sos, fiyat

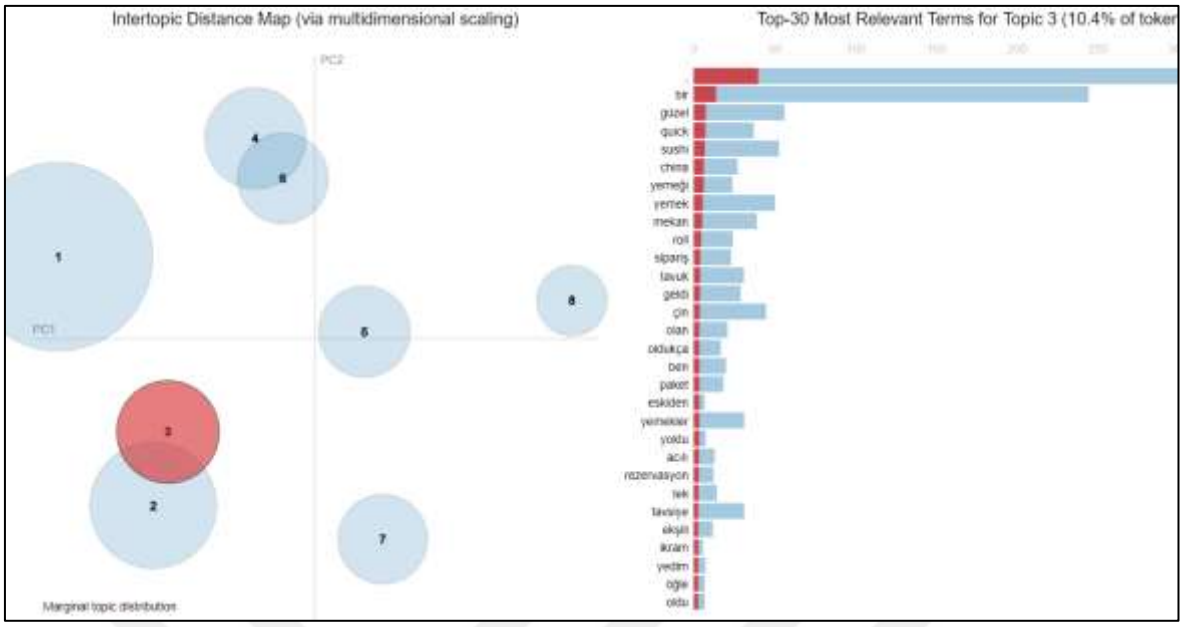
Örnek Yorum: "Arkadaşımla yoğun bir iş günü sonrasında gidip..."



Şekil 3.6. LDA modelinde ikinci grup baskın konu

Konu 3: Anahtar Kelimeler: harika, doğu, sushi, uzak, lezzet, yemek, güzel

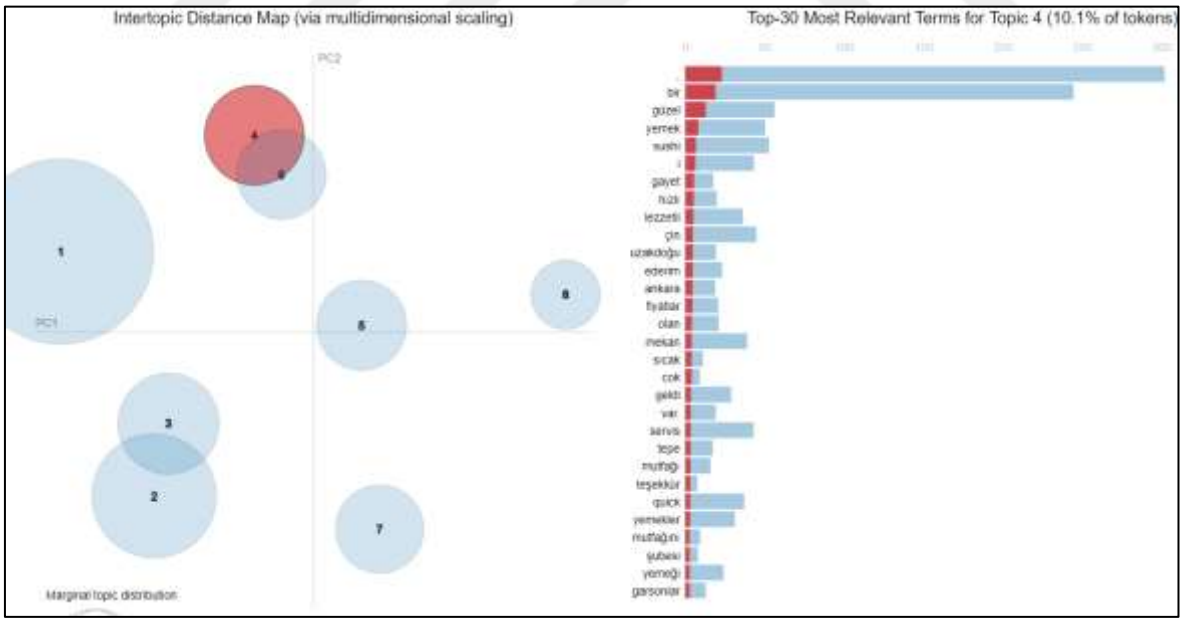
Örnek Yorum: "Çarşamba günü akşam 17:30 da arkadaşımla Armad..."



Şekil 3.7. LDA modelinde üçüncü grup baskın konu

Konu 4: Anahtar Kelimeler: biraz, sushi, fiyat, güzel, tavsiye, sıcak, çin

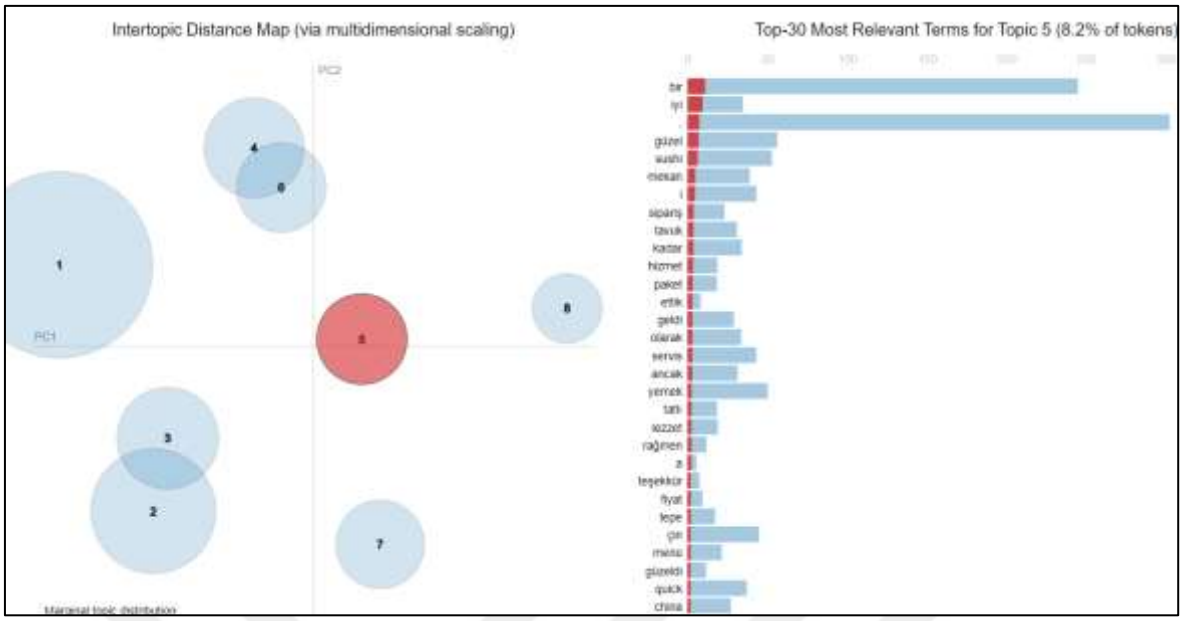
Örnek Yorum: "Dün gop taki Quick China ya gittik tek kelime..."



Şekil 3.8. LDA modelinde dördüncü grup baskın konu

Konu 5: Anahtar Kelimeler: yemek, lezzetli, güzel, gittik, quick, china, kadar

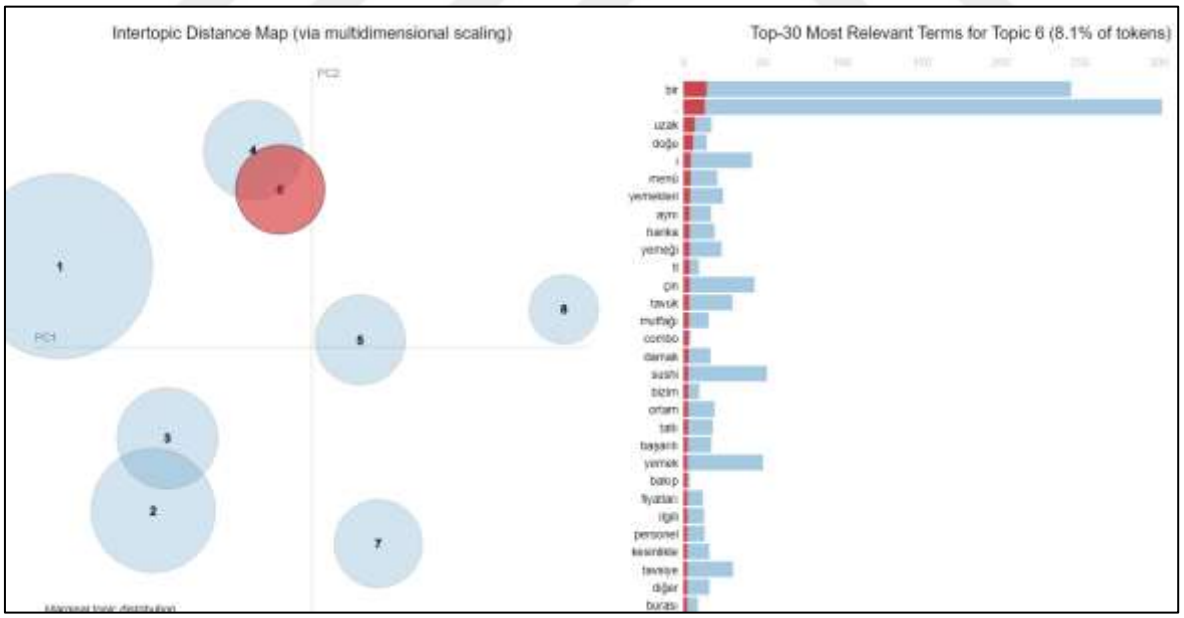
Örnek Yorum: "Güler yüzlü personel mekan harika yemeklerde b..."



Şekil 3.9. LDA modelinde beşinci grup baskın konu

Konu 6: Anahtar Kelimeler: yemekler, tavsiye, mekan, güzel, oldu, hizmet, fiyat

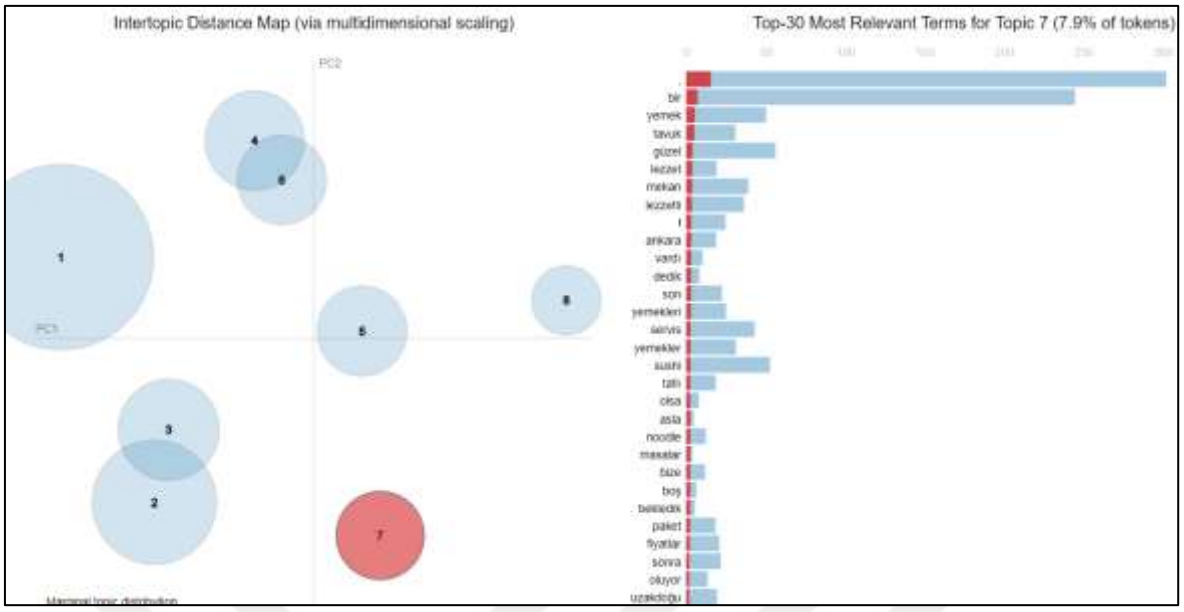
Örnek Yorum: "Personellerin kabalık ve ukalağından mı bahset..."



Şekil 3.10. LDA modelinde altıncı grup baskın konu

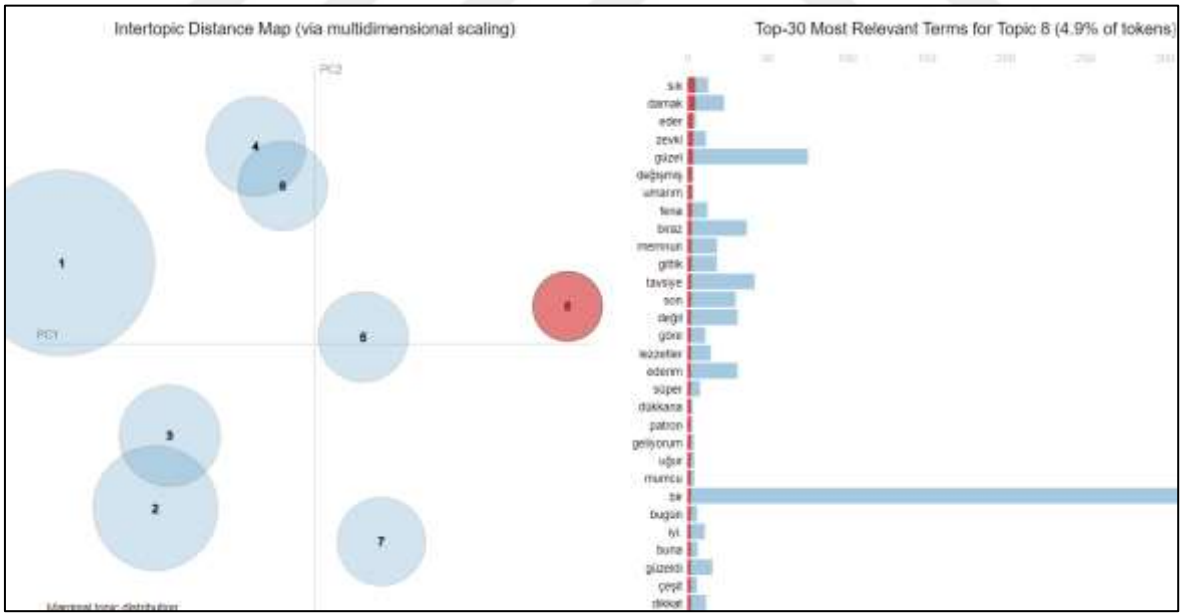
Konu 7: Anahtar Kelimeler: bir, güzel, sushi, ben, servis, iyi, fiyatları

Örnek Yorum: "Eskiden Ankara'daki en sevdiğim mekanlardan bi..."



Şekil 3.11. LDA modelinde yedinci grup baskın konu

Konu 8: Anahtar Kelimeler: bir, kadar, sushi, çin, servis, i, ancak, güzel
Örnek Yorum: "Kapı önünde biri var rezervasyon alıyor sözde..."



Şekil 3.12. LDA modelinde sekizinci grup baskın konu

LDA, baskın konulardaki belirgin kelimeleri ve bu kelimelerin katkı yüzdesini belirleyerek konuların içeriğini anlamamıza yardımcı olan bir topic modeling tekniğidir. Her bir baskın konu, belirli anahtar kelimelerle temsil edilir. Örneğin, Konu 1 sushi, Çin, servis gibi kelimelerle ilişkilidir ve bu konu, müşteri yorumlarındaki belirli öğelerle ilgili

görünmektedir. Verilerdeki bazı müşteri yorumlarının hangi baskın konu içerisinde olduğu, bu yorumların baskın konuya olan katkı yüzdeleri ve anahtar kelimeler Çizelge 3.8.'de gösterilmiştir.

Çizelge 3.8. Verilerin baskın konu ve katkı yüzdesi

	Baskın Konu Numarası (Dominant_Topic)	Katkı Yüzdesi (Perc_Contribution)	Anahtar Kelimeler (Topic_Keywords)	Müşteri Yorumu
0	4	0,9539	yemek, lezzetli, ,, güzel, gittik, kadar, quic...	Güler yüzlü personel mekan harika yemeklerde b...
1	2	0,9675	,, bir, kadar, sushi, çin, servis, i, ancak, g...	Japonya'da yaşamış arkadaşımızla gittik. Şu an...
2	2	0,9852	,, bir, kadar, sushi, çin, servis, i, ancak, g...	Bu restoranı normalde severdim ancak geçtiğimi...
3	6	0,9735	,, bir, güzel, sushi, ben, servis, iyi, fiyatl...	Eskiden Ankara'daki en sevdiğim mekanlardan bi...
4	1	0,9687	,, bir, güzel, yemek, sushi, yemeği, mekan, so...	Arkadaşımla yoğun bir iş günü sonrasında...
5	3	0,9849	bir, ,, harika, doğu, sushi, uzak, ederim, yem...	Çarşamba günü akşam 17:30 da arkadaşımla Armad...
6	2	0,9928	,, bir, kadar, sushi, çin, servis, i, ancak, g...	Dün gop taki Quick China ya gittik tek kelime ...
7	1	0,9888	,, bir, güzel, yemek, sushi, yemeği, mekan, so...	Daha önce de birçok kez gittiğim restauranta s...
8	5	0,965	bir, ,, yemekler, i, tavsiye, mekan, güzel, ol...	Personellerin kabalık ve ukalağından mı bahset...
…	…	…	…	…

Çıktılara göre, bazı yorumların baskın konusu ve bu konunun katkı yüzdesi şu şekildedir:

- Baskın Konu (Dominant_Topic):

Anahtar Kelimeler: "bir, ,, servis, i, çin, olarak, tavsiye, ancak, kadar, lezzetli, sushi, mekan, biraz, quick, yemek, güzel, iyi, ederim, değişmiş, memnun, süper, damak, fena, tavsiye, lezzetler, uğur, çeşit, dikkat"

Katkı Yüzdesi: %96.75
Örnek Yorum: "Japonya'da yaşamış arkadaşımızla gittik. Şu anda..."

- Baskın Konu (Dominant_Topic):

Anahtar Kelimeler: "bir, kadar, sushi, çin, servis, i, ancak, güzel, yemek, lezzet, mekan, iyi, sipariş, tavuk, hizmet, paket, etik, geldi, olarak, tatlı, rağmen, a, teşekkür, fiyat, tepe, çin, menü, güzeldi, quick, china"

Katkı Yüzdesi: %98.52

Örnek Yorumlar: "Bu restoranı normalde severdim ancak geçtiğimi...", "Dün G.O.P'taki Quick China'ya gittik, tek kelime..."

- Baskın Konu (Dominant_Topic):

Anahtar Kelimeler: "bir, güzel, sushi, ben, servis, iyi, fiyatlı, ankara, fiyatlar, olan, mekan, sıcak, çok, geldi, var, servis, tepe, teşekkür, quick, yemekler, mutfağı, mutfağını, şubesi, yemeği, garsonlar"

Katkı Yüzdesi: %97.35
Örnek Yorum: "Eskiden Ankara’daki en sevdiğim mekanlardan biri..."

- Baskın Konu (Dominant_Topic):

Anahtar Kelimeler: "bir, güzel, yemek, sushi, yemeği, mekan, sos, sipariş, aldık, lezzet, güzel, izmir, özellikle, işletme, çin, tencere, porsiyon, baya, bir, geldi, çok, o, olarak, ertesi, teşekkür, açık, fiyat, tavsiye, bugün"

Katkı Yüzdesi: %96.87
Örnek Yorum: "Arkadaşımla yoğun bir iş günü sonrasında gidip..."

- Baskın Konu (Dominant_Topic):

Anahtar Kelimeler: "bir, harika, doğu, sushi, uzak, ederim, yemek, china, roll, tavuk, sipariş, geldi, çin, olan, oldukça, ben, paket, eskiden, yemekler, yoktu, acılı, rezervasyon, tek, tavsiye, ekşili, ikram, yedim, öğle, oldu"

Katkı Yüzdesi: %98.49
Örnek Yorum: "Çarşamba günü akşam 17:30'da arkadaşımla Armada'ya gittik..."

- Baskın Konu (Dominant_Topic):

Anahtar Kelimeler: "bir, güzel, quick, sushi, china, yemeği, yemek, mekan, roll, tavuk, sipariş, geldi, çin, olan, oldukça, ben, paket, eskiden, yemekler, yoktu, acılı, rezervasyon, tek, tavsiye, ekşili, ikram, yedim, öğle, oldu"
Katkı Yüzdesi: %99.28
Örnek Yorum: "Dün G.O.P'taki Quick China'ya gittik, tek kelime..."

- Baskın Konu (Dominant_Topic):

Anahtar Kelimeler: "bir, kadar, sushi, çin, servis, i, ancak, güzel, yemek, lezzet, mekan, iyi, sipariş, tavuk, hizmet, paket, etik, geldi, olarak, tatlı, rağmen, a, teşekkür, fiyat, tepe, çin, menü, güzeldi, quick, china"

Katkı Yüzdesi: %97.5
Örnek Yorum: "Kapı önünde biri var rezervasyon alıyor sözde..."

- Baskın Konu (Dominant_Topic):

Anahtar Kelimeler: "bir, kadar, sushi, çin, servis, i, ancak, güzel, yemek, lezzet, mekan, iyi, sipariş, tavuk, hizmet, paket, etik, geldi, olarak, tatlı, rağmen, a, teşekkür, fiyat, tepe, çin, menü, güzeldi, quick, china"

Katkı Yüzdesi: %97.88
Örnek Yorum: "Daha önce de birçok kez gittiğim restoranta s..."

- Baskın Konu (Dominant_Topic):

Anahtar Kelimeler: "bir, yemekler, i, tavsiye, mekan, güzel, oldukça, özellikle, damak, güzel, quick, sushi, china, yemeği, yemek, mekan, roll, tavuk, sipariş, geldi, çin, olan, oldukça, ben, paket, eskiden, yemekler, yoktu, acılı"

Katkı Yüzdesi: %96.5
Örnek Yorum: "Personellerin kabalık ve ukalalığından mı bahset..."

Bu çıktılar, her bir yorumun hangi konuyla ilişkilendirildiğini, baskın konuyu ve bu konunun yoruma olan katkısını göstermektedir. LDA modelinin belirlediği konulara dair özet bilgileri incelendiğinde, her bir konunun anahtar kelimelerine dayanarak temsil edildiğini görebiliriz.

İlgili konu içeriğini anlamak üzere anahtar kelimelerle genel olarak bakılırsa:

Konu 1: Konu geneline bakıldığında genel restoran deneyiminden bahsedilmektedir.

Anahtar Kelimeler: bir, servis, çin, tavsiye, lezzetli, sushi, mekan, quick, yemek

Ortalama Puan: 3.55
Standart Sapma: 1.45

Konu 2: Konu geneline bakıldığında müşteri memnuniyeti ve değişimden söz edilmektedir.

Anahtar Kelimeler: sık, damak, güzel, değişmiş, umarım, memnun, tavsiye, lezzetler, süper, güzeldi

Ortalama Puan: 3.92
Standart Sapma: 1.26

Konu 2: Konu geneline bakıldığında iyi Sushi deneyiminden bahsedilmiştir.

Anahtar Kelimeler: bir, iyi, güzel, sushi, mekan, hizmet, lezzet, fiyat, teşekkür, çin

Ortalama Puan: 4.05
Standart Sapma: 0.91

Konu 3: Konu geneline bakıldığında yemekler ve mekan Atmosferinden

Anahtar Kelimeler: bir, yemek, sushi, lezzetli, açık, hizmet, çin, büfe, güzel, quick

Ortalama Puan: 3.58
Standart Sapma: 1.47
Konu 4: Konu geneline bakıldığında lezzet ve mekan İncelemesi yapılmaktadır.

Anahtar Kelimeler: yemek, tavuk, güzel, lezzet, mekan, lezzetli, sushi, fiyatlar, uzakdoğu,

Ortalama Puan: 3.27
Standart Sapma: 1.67

Konu 5: Konu geneline bakıldığında Quick China'da yemek deneyiminden söz edilmektedir.

Anahtar Kelimeler: güzel, quick, sushi, china, yemek, mekan, roll, tavuk, çin, ekşili

Ortalama Puan: 4.05
Standart Sapma: 1.05

Konu 6: Konu geneline bakıldığında genel mekan ve hızlı servis durumundan bahsedilmiştir.
Anahtar Kelimeler: bir, güzel, yemek, sushi, çin, uzakdoğu, hızlı, lezzetli, fiyatlar, garsonlar

Ortalama Puan: 3.15
Standart Sapma: 1.66

Konu 7: Konu geneline bakıldığında uzak doğu mutfağı ve menü Çeşitliliğinden söz edilmektedir.

Anahtar Kelimeler: bir, uzak, doğu, yemekleri, harika, çin, tavuk, mutfağı, sushi, tatlı, combo

Ortalama Puan: 3.567
Standart Sapma: 1.16

Baskın konuyu belirlemek için genellikle belirlenen her konunun ağırlığını değerlendirmek gerekir. En yüksek ağırlığa sahip olan konu baskın olarak kabul edilir.

Her bir yorum için baskın konu, katkı yüzdesi ve anahtar kelimeleri makine öğrenmesi teknikleriyle analiz edilmiştir. İlk adım, müşteri yorumları veri seti üzerinde LDA analizi yapılmıştır. Bu analizi gerçekleştirmek için bir metin madenciliği ve LDA uygulaması kullanıldı. Her belge için LDA analizi sonucunda baskın konu belirlenmiştir. Baskın konu, belgenin hangi temanın ağırlıklı olarak bulunduğunu gösterir. Bu temanın ne hakkında olduğu ise "yiyecek kalitesi" veya "hizmet deneyimi" gibi belli başlı konular olmuştur. Her belge için LDA analizi sonucunda katkı yüzdesi hesaplandı. Bu yüzde, belgenin baskın konuya ne kadar katkı sağladığını gösterir. Örneğin, bir belgenin baskın konusu %70 ise, bu belge bu konu hakkında yazılmıştır. Katkı yüzdesini inceleyerek, belgelerin hangi konulara daha fazla odaklandığını belirlenir.

LDA analizi sonucunda her bir konu için ilişkilendirilen anahtar kelimeleri incelenmiştir. Bu anahtar kelimeler, belgelerin içeriği hakkında daha fazla öngörü sağlar. Hangi kelimelerin hangi konularla ilişkilendirildiğini gözlemleyerek, belirli konuların hangi kelimelerle öne çıktığı belirlenmiştir. #sent_topics_df adlı veri çerçevesi, her belgenin en baskın konusunu, bu konunun katkısını, konu anahtar kelimelerini ve metin içeriğini içerir. Bu bilgiler, metin madenciliği veya doğal dil işleme projelerinde belgelerin içeriğini ve önemli konularını anlamak için kullanılır. LDA modelinin sonuçlarını daha anlaşılır bir şekilde görselleştirmek için kullanılır. PyLDAvis kullanıcı dostu bir ara yüz sağlar ve metin verilerindeki konuları, kelime dağarcığını ve konular arasındaki ilişkileri daha iyi anlamanıza yardımcı olur. Bir metin veya belge koleksiyonundan çıkarılan bir konu modellemesi sonucu elde edilen bir konu için en önemli 30 terimi ifade eder. Bu terimler,

belirli bir konunun içeriğini temsil ettiği düşünülen kelimeleri içerir. "34.9% of tokens" kısmı, bu terimlerin belirli bir konunun toplam içeriğinde ne kadar oranda bulunduğunu gösterir. Yani, bu terimlerin toplam belge içindeki token'ların %34.9' unu oluşturduğunu ifade eder. Sonuç olarak, bu analiz sayesinde her bir müşteri yorumunun hangi konu veya tema üzerine odaklandığı belirlenmiş, baskın konuların anahtar kelimeleri ortaya çıkarılmış ve bu bilgiler görselleştirilmiştir. Bu tür analizler, büyük metin verilerini anlamak ve özetlemek için kullanılır ve müşteri geri bildirimlerinden anlamlı öngörüler elde etmeye yardımcı olabilir. LDA (Latent Dirichlet Allocation) modeli tarafından belirlenen her bir konunun içeriğini değerlendirmek, yorumlamak ve her konu için seçilen anahtar kelimeleri ve bu kelimelerin sıklıkları belirlenir.

Çizelge 3.9. Baskın konuların memnuniyet düzeylerine dağılımı

BASKIN KONU	1	2	3	4	5	6	7	8	TOPLAM
OLUMLU	46	16	39	28	13	42	11	12	207
OLUMSUZ	15	13	11	13	5	23	7	9	96
TOPLAM	61	29	50	41	18	65	18	21	303

Verilen çapraz tabloya dayalı olarak, her baskın konu için olumlu ve olumsuz memnuniyet düzeylerine sahip yorum sayılarını görebiliriz. Ayrıca baskın konu 1 için olumlu memnuniyet düzeyine sahip 46 yorum ve olumsuz memnuniyet düzeyine sahip 15 yorum vardır. Toplamda, bu konuda 61 yorum bulunmaktadır. Bu çapraz tablo, farklı baskın konulara göre olumlu ve olumsuz memnuniyet düzeylerinin dağılımını açıkça göstermektedir.

Ki-kare istatistiği, baskın konularla memnuniyet düzeyi arasında bir ilişki olup olmadığını belirlemek için kullanılan bir ölçüdür. Bu istatistik, gözlenen ve beklenen frekanslar arasındaki farkı değerlendirerek bu ilişkiyi ölçer. p değeri, bu ilişkinin istatistiksel olarak anlamlı olup olmadığını belirler.

Verilen sonuçlara göre:

Ki-kare istatistiği 8.0801 ve p değeri 0.3255'dir.

p değeri 0.05'ten büyük olduğu için, baskın konularla memnuniyet düzeyi arasında istatistiksel olarak anlamlı bir ilişki bulunmamaktadır. Bu sonuç, baskın konuların memnuniyet düzeyini etkilemediğini veya bu değişkenler arasında zayıf bir ilişki olduğunu gösterebilir.

3.6. Kelime Çantası (Bag of Words - BOW) Yöntemi

3.6.1. Veri toplama

İlk olarak müşteri yorumlarını içeren bir veri kümesi oluşturmaktır. Bu yorumlar, restoranın hizmeti, yemekleri, ambiyansı ve diğer ilgili faktörler hakkında kullanıcıların görüşlerini ve sayısal olarak puanlama içermelidir. Çin restoranına yapılan yorumlar internet siteleri ve yemek sipariş platformlarından seçilerek toplanmıştır.

3.6.2. Metin ön işleme

Veri kümesindeki metinleri ön işleme yapmak ve düzenlemek önemlidir. Bu adımlar ise;

- Metinler küçük harfe çevrilir.
- Özel karakterleri ve noktalama işaretlerini kaldırılır.
- Etkisiz kelimeler (örneğin, "ve", "veya", "bu", vb.) çıkarılır
- Kelimeler kökleri veya temel formlarına indirgenir (kök bulma (stemming) veya lemmatization).
- Metin belgeleri belirli bir format veya yapıya getirilir.

BOW, metin verilerini sayısal özelliklere dönüştürmek için kullanılan bir yöntemdir. Bu yöntemle, her yorumdaki kelimelerin ve frekanslarının bir matris şeklinde temsil edilmesi sağlanır.

Kelime: güzel, Sütun İndeksi: 165
Değerler: [1 0 1 0 1 0 2 0 1 1 0 1 0 1 0 0 2 0 0 0 0 0 1 1 0]
Kelime: servis, Sütun İndeksi: 337
Değerler: [0 1 3 1 0 0 0 0 0 1 1 1 0 0 1 0 0 0 0 0 0 0 1 1 1]
Kelime: mekan, Sütun İndeksi: 244
Değerler: [1 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 1 1 0]

Çalışma mantığı olarak oluşturulan matris "servis" kelimesi için sütun indeksi 337'dir ve bu kelimenin ilk 25 satırdaki değerleri gösterilir. Bu değerler, ilgili metinlerde "servis" kelimesinin geçtiği frekansları temsil eder.

Bu matris makine öğrenimi algoritmalarına girdi olarak verilir. Makine öğrenmesi algoritmalarını uygulamadan önce yorumlara ilişkin veri temizliği yapılmıştır. Veriler temizlendikten sonra Bags of Words (BoW) yöntemi ile yorumlara karşılık gelen matrisler oluşturulmuştur. Sonra veriler 0,2 oranında test verisi olacak şekilde eğitim ve test verisi olarak ikiye ayrılmıştır.

3.7. Sınıflandırma

Her bir müşteri yorumu, verilen memnuniyet puanına göre iyi veya kötü olarak etiketlenmişti. Bu adımda, MLP, DecisionTree, SGD, AdaBoost, GradientBoost, LogisticReg, RandomForest, SVC, LGBM ve XGB algortimaları ile yorumlar, metin verilerinin sayısallaştırılması ile dönüştürüldükleri vektörlerin modelin bağımsız değişkenleri olarak kullanarak sınıflandırılmıştır.

Optuna ile makine öğrenmesi algoritma seçimi ve hyperparameter tunning için gerekli olan fonksiyon tanımlanıyor. Makine öğrenmesi algoritmaları olarak AdaBoost, DecisionTree, GradientBoosting, Kneighors, LGBM, LogisticRegression, MLP, RandomForest, SGDC, SVC, XGB algoritmaları kullanılmıştır. Oluşturulan BoW matrisi, makine öğrenmesi algoritmalarında X değerleri olarak atanıyor. Yorumlara karşılık gelen puanlar ise y değerleri olarak atanıyor. Model, eğitim verileri (X_train ve y_train) kullanılarak eğitilir ve test verileri (X_test) üzerinde tahminler yapılır. Tahminler ve gerçek etiketler kullanılarak doğruluk (accuracy) skoru hesaplanır ve bu skor accuracy değişkenine atanır. Makine öğrenmesi algoritmaların (XGB için özellikle) kategorilerin 0'dan başlaması için puanlardan 1 çıkartılıyor. Algoritmaların belirlenmesi ve algoritma parametrelerinin iyileştirilmesi için Optuna kütüphanesinden faydalanılmıştır. Optuna kütüphanesi kullanılarak, makine öğrenmesi algoritmaları için en iyi hyperparametre ayarlarını bulma işlemi gerçekleştirildi. "Optuna" denilen hiperparametre (bir makine öğrenimi modelinin veya algoritmasının davranışını ve performansını kontrol eden parametrelerdir. Bu parametreler, modelin nasıl öğrenme, yapacağını ne kadar karmaşık, olacağını ne kadar hızlı öğreneceğini ve benzeri özellikleri belirler) optimizasyon kütüphanesi kullanarak, farklı sınıflandırma (classification) algoritmalarının farklı hiperparametre kombinasyonlarını değerlendiren bir "objective" fonksiyonunu tanımlar. Bu "objective" fonksiyonu, belirli bir algoritma ve hiperparametrelerle bir modeli eğitir, bu modeli test verileri üzerinde değerlendirir ve sonuç olarak doğruluk (accuracy) skoru döndürür. Bu,

algoritmaların performansını artırmak ve modelin doğruluğunu optimize etmek için önemlidir. Optuna ile 11 algoritma ve 47 hyperparametre üzerinden yapılan iyilenme çalışmasında 200 deneme yapılmış ve bu denemeler toplamda 8,68 dakika sürmüştür. Kullanılan metin madenciliği ve makine öğrenimi teknikleri ile müşteri yorumlarını sınıflandırılmıştır. Bag of Words (BoW) yöntemi kullanılarak yapılan makine öğrenmesi analizi sonuçlarına göre en iyi test doğruluk oranına sahip algoritmanın MLP (Multilayer Perceptron) olduğu görülmüştür.

Optuna'nın hiperparametre optimizasyon süreci, deneme sırasında belirli bir sınıflandırıcı ve hiperparametre kombinasyonunun performansını değerlendirir. En iyi performansı gösteren sınıflandırıcı ve hiperparametrelerin seçilmesinin nedeni, bu kombinasyonun veri kümesi ve problem bağlamında en iyi sonuçları vermesidir. Bu nedenle, MLPClassifier'ın seçilmesi, veri kümesi ve probleme en uygun modelin seçildiğini gösterir.

Modelin Karmaşıklığı ve Esnekliği: MLP, çok katmanlı bir yapay sinir ağı modelidir ve oldukça karmaşık yapıları dahi öğrenme yeteneğine sahiptir. Bu, veri kümesindeki daha karmaşık ilişkileri ve desenleri daha iyi anlayabilecek kapasiteye sahiptir.

Optuna İyileştirmesi: Optuna kullanılarak uygulanan hiperparametre optimizasyonu ile MLP'nin hiperparametre kombinasyonunda en iyi performansı gösterdiğini belirlemiştir. Bu, MLP'nin diğer algoritmaları aşan bir şekilde ayarlandığı anlamına gelebilir.

Veri Kümesi Özellikleri: Veri kümesinin yapısı, belirli bir modelin performansını etkileyebilir. MLP'nin esnekliği, farklı türlerdeki veri yapılarına daha iyi uyarlanabilirken, diğer algoritmaların bu yapıları anlamakta veya öğrenmekte zorlanabileceği durumlar olabilir.

Veri Ön işleme: Metin verilerinin uygun bir şekilde işlenmesi, modelin performansını büyük ölçüde etkileyebilir. MLP'nin daha iyi sonuçlar vermesinin nedenlerinden biri, metin verilerinin daha etkili bir şekilde ön işleme durumundan kaynaklı olabilir.

Diğer algoritmaların iyi sonuç vermemesinin nedenleri ise genellikle modelin karmaşıklığı, hiperparametrelerin ayarlanması ve veri setinin özelliklerinden kaynaklanabilir. Bu nedenlerin bir kombinasyonu, sonuçların farklılık göstermesine yol açabilir.

Diğer algoritmalarla karşılaştırma yapmak için doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F1-score gibi ölçümleri kullanabiliriz. Bu ölçümler, bir sınıflandırma modelinin performansını değerlendirmede yaygın olarak kullanılan ölçütlerdir.

Doğruluk (Accuracy): Modelin doğru tahmin ettiği toplam örnek sayısının, tüm örneklerin sayısına oranıdır. Yani doğru tahmin edilenlerin toplam tahminlere oranıdır.

Kesinlik (Precision): Pozitif olarak tahmin edilen örneklerin gerçekten pozitif olduğu oranıdır.

Duyarlılık (Recall): Gerçekten pozitif olan tüm örneklerin doğru bir şekilde pozitif olarak tahmin edilme oranıdır. Duyarlılık, gerçek pozitiflerin ne kadarının yakalandığını gösterir.

F1-Score: Kesinlik ve duyarlılığın harmonik ortalamasıdır. Dengeli bir performans ölçüsüdür, çünkü her iki ölçümde dikkate alır ve aynı zamanda düşük değerlere sahip olan bir metriğin sonucunu etkiler. F1 puanı yüksek olduğunda, modelin hem kesinlik hem de duyarlılık açısından iyi bir performans gösterdiği söylenebilir.

Çizelge 3.10. Algoritmaların performans ölçütleri

Algoritma	Doğruluk (accuracy)	Kesinlik (precision)	Duyarlılık(recall)	F1-Score
MLP	0.90	0.89	0.92	0.90
DecisionTree	0.64	0.65	0.63	0.64
SGD	0.38	0.40	0.35	0.37
AdaBoost	0.82	0.81	0.84	0.82
GradientBoost	0.79	0.78	0.80	0.79
LogisticReg	0.75	0.74	0.76	0.75
RandomForest	0.74	0.73	0.75	0.74
SVC	0.70	0.71	0.69	0.70
LGBM	0.67	0.66	0.68	0.67
XGB	0.67	0.67	0.68	0.67

MLP (Yapay Sinir Ağları): Diğer algoritmalar arasında en yüksek doğruluk (0.90), kesinlik (0.89), duyarlılık (0.92) ve F1-Skoru (0.90) değerlerine sahiptir. Bu, veri kümesinde iyi bir sınıflandırma performansı gösterdiğini gösterir.

AdaBoost: Doğruluk, kesinlik, duyarlılık ve F1-Skoru açısından ikinci en yüksek performansı sergiler (0.82). MLP'nin hemen arkasında yer alır.

GradientBoost ve LogisticReg: Bu iki algoritma, orta düzeyde performans sergiler. MLP ve AdaBoost'un gerisinde, ancak diğer algoritmaların önünde yer alır.

SGD (Stokastik Gradyan İniş): En düşük performansa sahip algoritma olarak öne çıkar. Doğruluk, kesinlik, duyarlılık ve F1-Skoru değerleri diğer algoritmaların altındadır.

Genel olarak, MLP ve AdaBoost, veri kümesinde en iyi performansı gösteren algoritmalardır. SGD ise diğer algoritmalara göre daha düşük performans göstermektedir. Bu karşılaştırma, farklı algoritmaların bir sınıflandırma görevindeki performansını değerlendirmek için kullanılabilir.

Modelin performansını iyileştirmek için doğru algoritmayı seçmek ve uygun hiperparametreleri belirlemek önemlidir.

Bununla birlikte yapılan incelemede puanlar 3'ten büyük ve küçük olarak ikiye ayrılmıştır. İki kategorili y değerleri üzerinden en iyi test doğruluk seviyesi MLP (Multilayer Perceptron) ile elde edilmiştir. Bu nedenle puanlar 3'ün üstü ve altı olarak (sırasıyla iyi ve kötü) şeklinde ikili sınıflandırmaya tabi tutulmuştur. Yine aynı algoritmalar kullanılarak yapılan incelemede en iyi test doğruluk oranının 0,90 olduğu görülmüştür. Bu, modelin verilen yorumlara dayanarak restoranın kalitesini ve müşteri memnuniyetini tahmin etmede daha başarılı olduğunu göstermektedir. Analiz, müşteri deneyimlerini değerlendirme ve restoranın güçlü ve zayıf yönlerini belirleme amacı taşımaktadır. Bu bilgiler, işletmenin gelecekteki stratejilerini şekillendirmek ve müşteri memnuniyetini artırmak için kullanılabilir.

4. SONUÇ VE ÖNERİLER

Çalışmada bir Çin restoran müşterilerinin restoran hakkındaki yorumları metin madenciliği teknikleri ile incelenmiştir. Elde edilen 303 yorumun, puanlamalarına göre %68,32 olumlu, %31,68 olumsuz olmuştur. Müşteri yorumlarında ortalama 11643 kelime kullanılmış, kullanılan ortalama kelime sayısı bakımından olumlu ve olumsuz yorum grupları arasında Mann-Whitney U Testi (11487,5) sonucuna göre gruplar arasındaki farkın olduğu, p-değeri (0,02877) sonucuna göre bu farkın rastgele oluşmadığı ve istatistiksel olarak anlamlı olduğu söylenebilir. Olumlu müşteriler ile olumsuz müşteriler arasında yorumlarında kullandıkları ortalama kelime sayısı bakımından anlamlı fark bulunmuştur. Bu sonuca dayanarak, olumlu müşterilerin olumsuz müşterilere göre ortalama kelime sayısı olarak daha az olduğu söylenebilir. Olumsuz yorum yapan müşterilerin daha az olmasına rağmen deneyimlerini daha detaylı bir şekilde ifade etmeye meyilli oldukları belirlenmiştir. Olumsuz deneyimlerin ayrıntılı ve uzun ifade edilmesi, müşterilerin sorunları çözmek ve işletmenin dikkatini çekmek için bir yol olabilir. Yapılan yorumlarda tespit edilen baskın konular ise LDA analizi ile belirlenmiştir.

Uygulanan LDA algoritması ile yorumlarda tespit edilen 8 baskın konu:

1. Genel restoran deneyimi: 46 olumlu, 15 olumsuz yorum yapılmıştır. Genel olarak restoranın güzel yemekleri, temiz atmosferi ve samimi personeli övülmüştür. Bu tür bir yorum, diğer potansiyel müşterilere restoran hakkında olumlu bir izlenim vermek için kullanılabilir.
2. Memnuniyet ve değişim: 16 olumlu, 13 olumsuz yorum yapılmıştır. Hizmet ve yemek çeşidi konusunda olumlu anlamda değişimden bahsedilmiştir. Lezzet ve hizmetin genel olarak beğenildiği ancak bazılarının iyileştirilebileceği belirtilmiştir.
3. Sushi deneyimi: 39 olumlu, 11 olumsuz yorum yapılmıştır. Özellikle Roll sushi ve diğer sushi lezzetleri övülmüştür.
4. Yemekler ve mekân atmosferi: 28 olumlu, 13 olumsuz yorum yapılmıştır. Bazı şubelerin menü çeşitliliğinin fazla olmasına rağmren ürünlerin hızlı tükendiğinden istenilen yemek çeşitlerin olmamasından bahsedilmiştir.

5. Lezzet ve Mekân: 13 olumlu, 5 olumsuz yorum yapılmıştır. Çin mutfağına özgü yemeklerin lezzeti övülmüştür. Bazı yemeklerin acılığı dışında genel olarak memnun kalındığı vurgulanmıştır.
6. Yemek deneyimi: 42 olumlu, 23 olumsuz yorum yapılmıştır. Eskiden olmayan yeni çeşitlerin deneyimlenmesi olumlu bir şekilde değerlendirilmiş ve restoranın atmosferi olumlu bir şekilde vurgulanmıştır.
7. Genel mekân ve hızlı servis durumu: 11 olumlu, 7 olumsuz yorum yapılmıştır. Restoranın bulunduğu lokasyonun, iç mekânın ve servisin kalitesi vurgulanmıştır. Ayrıca, ilgili garsonlar ve servis hızı gibi olumsuz deneyimler belirtilmiştir. Rezervasyon yapmanın önemi üzerinde durulmuştur.
8. Uzak doğu mutfağı ve menü çeşitliliği olmuştur. 12 olumlu, 9 olumsuz yorum yapılmıştır. Özellikle sushi ve Çin mutfağından memnun kalınmıştır.

Bu konular ile yorumların olumlu ya da olumsuz oluşu arasında ise anlamlı ilişki bulunamamış, dolayısı ile müşterilerin memnuniyet durumları ile restorana öne çıkan ilişkin konular arasında bir ilintilendirme yapılamamıştır.

Çalışmada, ayrıca yorumlar metin madenciliği teknikleri ile sayısallaştırılırak bu verinin girdi olarak kullanılması ile müşteri yorumları sınıflandırılmıştır. Bu amaçla kullanılan MLP, DecisionTree, SGD, AdaBoost, GradientBoost, LogisticReg, RandomForest, SVC, LGBM ve XGB algoritmaları arasından MLP yöntemi en yüksek doğruluk ve 0.90 değerleri ile en iyi sınıflandırma performansı gösteren algoritma olmuştur. MLP, çok katmanlı yapısı sayesinde daha karmaşık ilişkileri öğrenme yeteneğine sahip olduğundan geniş bir öğrenme yeteneği sunar ve modelin veri özelliklerini daha iyi öğrenmesini sağlamıştır. Bu, modelin verilen müşteri yorumlarına dayanarak restoranın kalitesini ve müşteri memnuniyetini başarılı bir şekilde tahmin edebildiğini göstermektedir.

Analiz sonuçlarına bakılırsa, belirli konu başlıkları altında gruplanan yorumlarda öne çıkan temalar şunlar görülmektedir: yemek kalitesi, sushi deneyimi, hizmet kalitesi, mekân atmosferi gibi sıralanabilmektedir.

Bütün bu bulgulara dayanarak, restoran için geliştirilebilir alanlar rezervasyon ve bekleme süreleri, sosyal medya ve iletişim, personel eğitiminin tamamlanması, hizmet kalitesi ve servis hızı olarak belirlenebilir.

Bu, restoran işletmecilerine müşteri geri bildirimlerini daha etkili bir şekilde analiz etme ve işletme performansını artırma konusunda rehberlik etmektedir. Ayrıca, makine öğrenimi yöntemlerinin yemek yorumları gibi metin verilerinin analizindeki uygulanabilirliğini göstererek, literatüre katkı sağlamayı amaçlamaktadır.

Çalışmanın sonuçları, restoran işletmecilerine ve online site kullanıcılarına önemli bilgiler sunar. İyileştirmelerin hangi alanlarda yapılması gerektiği, müşteri beklentilerine nasıl daha iyi cevap verilebileceği konusunda değerli bilgiler sağlayarak, işletmelerin rekabet avantajını sürdürmelerine yardımcı olabilir. Elde edilen sonuçlar, restoranın müşteri profilinn daha iyi anlamak, hizmet kalitesini iyileştirmek ve pazarlama stratejilerini optimize etmek için kullanılabilir. Restoran yönetimi, müşteri taleplerine daha iyi yanıt verebilmek için bu analizleri kullanarak stratejilerini geliştirebilir. Metin madenciliği ve makine öğrenmesi, Çin restoranları ve benzeri işletmeler için müşteri geri bildirimlerini anlamak ve işletme performansını iyileştirmek için güçlü araçlar sunmaktadır. Bu analizler, müşteri memnuniyetini artırarak restoranların sürdürülebilir başarı elde etmelerine katkıda bulunabilir. İşte bu bağlamda Çin restoranına yönelik öneri olarak şu yorumlar yapılabilir.

Rezervasyon ve bekleme sürelerinde iyileştirme: Müşterilerin belirttiği bir konu olarak, rezervasyon ve bekleme sürelerinin optimize edilmesi önemlidir. Restoran, rezervasyon süreçlerini ve bekleme zamanlarını yönetmek için online rezervasyon sistemleri gibi teknolojik çözümleri kullanabilir.

Hizmet kalitesi ve servis hızı iyileştirmeleri: Personel eğitimine ve servis süreçlerinin iyileştirilmesine odaklanmak önemlidir. Personelin müşteriye daha dikkatli ve hızlı hizmet sunması için eğitimler düzenlenebilir.

Sushi deneyimi ve Çin mutfağına odaklanma: Müşteri yorumlarında sıkça övülen sushi deneyimi ve Çin mutfağına özgü yemekler üzerinde odaklanılabilir. Restoran, bu alanlarda kaliteyi koruyarak ve müşteri taleplerine yanıt vererek rekabet avantajı elde edebilir.

Sosyal medya ve iletişim stratejileri: Restoran, müşteri geri bildirimlerini almak ve müşteri iletişimini güçlendirmek için aktif bir şekilde sosyal medya platformlarını kullanabilir. Müşterilerin restoran hakkında olumlu deneyimlerini paylaşmaları teşvik edilebilir. Çin restoran sektöründeki rekabeti anlamak için rakip restoranların faaliyetlerini ve müşteri

geri bildirimlerini analiz edebilir. Bu, fiyatlandırma stratejilerini ve benzersiz satış noktalarını belirlemede yardımcı olabilir.

Müşteri Segmentasyonu: Restorana gelen müşteri verilerini analiz ederek benzer özelliklere sahip müşteri gruplarını belirleyebilir. Bu, farklı müşteri segmentlerine özel pazarlama kampanyası oluşturulmasını mümkün kılar.

Kişiselleştirilmiş Pazarlama: Müşteri verileri, alışkanlıklar ve tercihler açısından analiz edilerek, kişiselleştirilmiş pazarlama stratejileri oluşturulabilir. Müşterilere özel teklifler, indirimler veya promosyonlar sunarak müşteri bağlılığını artırabilir.

Menü Optimizasyonu: Müşteri sipariş veri setlerini analiz ederek, en popüler yemekleri ve içecekleri belirleyebilir ve menü buna göre optimize edilebilir.

Çin restoranına gelen müşteri verileri etkili bir şekilde analiz edilerek, bu verilerden çıkarılan anlamlı bilgileri işletme stratejilerini optimize etmek için kullanabilir. Bu, müşteri memnuniyetini artırmanın yanı sıra satışları ve işletme performansını genel olarak iyileştirilmesine yardımcı olabilir. Modeli sürekli olarak güncelleyerek ve daha fazla veri toplayarak, gelecekteki müşteri deneyimlerini daha iyi anlamak ve yönetmek için değerli öngörüler elde edilebilir.

KAYNAKLAR

Afaq, A., Gaur, L. and Singh, G. (2022). *Küresel Zincir Otellerin Çevrimiçi İncelemelerinin Görüş Madenciliği için Gizli Dirichlet Tahsis Tekniği.* 3. Uluslararası Akıllı Mühendislik ve Yönetim Konferansı (ICIEM) , Londra.

Agrawal, A., Fu, W. and Menzies, T. (2018). What is wrong with topic modeling and how to fix it using search-based software engineering. *Information and Software Technology, 98,* 74-88.

Akis, M. S. (2019). *Using Text Mining Algorithms of Health Data for Analysis Purposes*. Master's thesis, Dokuz Eylül University, Graduate School of Natural and Applied Sciences, İzmir.

Aksu, M. Ç. ve Karaman, E. (2020). FastText ve Kelime Çantası: Kelime temsil yöntemlerinin turistik mekanlar için yapılan Türkçe incelemeler kullanılarak karşılaştırılması. *Avrupa Bilim ve Teknoloji Dergisi,* 20, 311-320.

Alpaydın, E. (2000). *Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri*. Bilişim 2000 Eğitim Semineri, İstanbul.

Alrayashi, A. (2023). *Analyzing Higher Education Research Trends With Latent Dirichlet Allocation: A Text Mining Approach.* Master's Thesis, Atılım University, Graduate School of Natural and Applıed Sciences, Ankara.

Arslan, S. (2019). *Makine Öğrenmesi Uygulamalarında Yapay Arı Koloni Programlama Temelli Yeni Yöntemlerin Geliştirilmesi.* Doktora Tezi, Erciyes Üniversitesi, Fen Bilimleri Enstitüsü, Kayseri.

Atabay, L. (2020). *Otel Yorumlarının Metin Madenciliği Yöntemleri ile Karşılaştırmalı Analizi: Mayorka, Antalya, Şarm-El Seyh Örneği*. Yüksek Lisans Tezi, Akdeniz Üniversitesi, Sosyal Bilimler Enstitüsü, Antalya.

Atan, S. (2016). *Metin Madenciligi ile Sentiment Analizi ve Borsa İstanbul Uygulaması*. Doktora Tezi, Ankara Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Aydın, G. ve Hallaç, İ. R. (2021). Türkçe metinlerde otomatik konu tespiti. *Fırat Üniversitesi, Mühendislik Bilimleri Dergisi, 33*(2), 599-606.

Ayhan, S. ve Erdoğmuş, Ş. (2014). Destek vektör makineleriyle sınıflandırma problemlerinin çözümü için çekirdek fonksiyonu seçimi. *Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi, 9*(1), 175-201.

Başer, S. H. (2022). *Konaklama Sektöründe İnsansı Robota İlişkin Bakış Açısının Metin Madenciliğiyle Belirlenmesi*. Yüksek Lisans Tezi, Aksaray Üniversitesi, Sosyal Bilimler Enstitüsü, Aksaray.

Başkaya, F. (2017). *Kısa Metinlerden Sosyal Duygu Sınıflandırma İçin Makine Öğrenmesi Tabanlı Yöntemlerin Geliştirilmesi*. Yüksek Lisans Tezi, Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Elazığ.

Başkaya, F. ve Aydın, İ. (2017). *Haber metinlerinin farklı metin madenciliği yöntemleriyle sınıflandırılması*. International Artificial Intelligence and Data Processing Symposium (IDAP), 16-17 Eylül, Malatya, Türkiye.

Berry, M. W. and Kogan, J. (2010). *Text Mining: Applications and Theory*. New York: John Wiley & Sons Ltd.

Bilgin, M. and Şentürk, İ. F. (2017). *Sentiment analysis on Twitter data with semisupervised Doc2Vec*. 2rd International Conference on Computer Science and Engineering (UBMK), 5-8 October, Antalya, Türkiye.

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of science. *Annals of Applied Statistics, 1*(1), 17-35.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research,* 3, 993-1022.

Bowden, K. K., Wu, J., Oraby, S., Misra, A. and Walker, M. (2018). *SlugNERDS: A named entity recognition tool for open domain dialogue systems*. Language Resources and Evaluation Conference (LREC), Miyazaki, Japan.

Bozan, N. (2022). *Extendıng The Boundrıes In Descrıptıve Translatıon Analysıs: A Text Mınıng Approach*. Yüksek Lisans Tezi, İstanbul 29 Mayıs Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul.

Budak, İ. ve Sökmen, A. (2022). Otel hizmetlerinin değerlendirilmesinde Gizli Dirichlet ayrımı ile analiz: Kastamonu ili örneği. *Journal of Tourism and Gastronomy Studies, 10*(4), 2942-2954.

Can, Ş. (2017). *Veri Madenciliği ve Eğitim Sektöründe Bir Uygulama.* Yüksek lisans tezi, Celal Bayar Üniversitesi, Sosyal Bilimler Enstitüsü, Manisa.

Chen, D. Y. (2017). *Pandas for Everyone: Python Data Analysis*. Boston: Addison-Wesley Professional.

Chen, Y. S., Chen, L. H. and Takama, Y. (2015). *Proposal of LDA-based sentiment visualization of hotel reviews.* International Conference on Data Mining Workshop (ICDMW), 14-17 November, Atlantic City, USA.

Çakmak, C. (2023). *Ekonomilerinin Gelecek Perspektifi ve Kullanıcı Yorumlarının Metin Madenciliği ile Analizi*. Yüksek Lisans Tezi, Cumhuriyet Üniversitesi, Sosyal Bilimler Enstitüsü, Sivas.

Çelenli, H. İ. (2020). *Gizli Dirichlet Ayrımı ve Word2vec Yöntemlerinin Birleşimi ile Özgün Bir Metin Temsil Modeli Geliştirilmesi*. Yüksek Lisans Tezi, Kocaeli Üniversitesi, Fen Bilimleri Enstitüsü, Kocaeli.

Çelik, S. (2020). Metin madenciliği ile Shakespeare külliyatının incelenmesi. *MANAS Sosyal Araştırmalar Dergisi, 9*(3), 1343-1357.

Çeliksu, Z. (2017). *Yabancı Dizilerin Altyazı ve Twitter Yorumlarının Metin Madenciliği ile İncelenmesi*. Yüksek Lisans Tezi, Mimar Sinan Güzel Sanatlar Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Çoban, Ö. and Özyer, G. T. (2016). *Sentiment classification for Turkish Twitter feeds using LDA.* 24th Signal Processing and Communication Application Conference (SIU), 16-19 May, Zonguldak, Türkiye.

Dang S. and Ahmad P.H. (2014). Text mining: Techniques and its application. *International Journal of Engineering & Technology Innovations, 1*(4), 22-25.

Daşgın, R. ve Adem, K. (2023). Kitlesel çevrimiçi ders platformlarında yapılan yorumların metin madenciliği kullanılarak duygu analizinin yapılması. *Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi, 15*(2), 636-646.

Değer, N. S. (2017). *Sosyal Medya Mesajlarında Veri Madenciliği ile Duygu Analizi*. Doktora Tezi, İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul.

Delen, D. and Crossland, M., D. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications, 34*(3), 1707-1720.

Diez-Olivan, A., Del Ser, J., Galar, D. and Sierra, B. (2019). Data fusion and machine learning for industrial prognosis: Trends and perspectives towards Industry 4.0. *Information Fusion, 50*, 92-111.

Eken, B. (2015). *Kısa Metinlerde Varlık İsmi Tanıma*. Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Ekinci, E., İlhan, S., O., Kırık, E. ve Taşçı, Ş. (2020). Tıp veri kümesi için Gizli Dirichlet ayrımı. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, 22*(64), 67-80.

Ergün, M. (2016). Using the techniques of data mining and text mining in educational research. *Elektronik Eğitim Bilimleri Dergisi, 6*(12), 180-189.

Ergün, M. (2017). Eğitim araştırmalarında data mining ve text mining tekniklerinden yararlanma. *Electronic Journal of Education Sciences, 6*(12), 180-189.

Esen, E. and Özkan, S. (2017). *Analysis of Turkish Parliament records in terms of party coherence*. 25th Signal Processing and Communications Applications Conference (SIU), 15-18 May, Antalya, Türkiye.

Fadel, A. İ. (2020). *Terörizm İçerikli Online Sosyal Yorumların Duygu Analizi ve Fikir Madenciliği*. Doktora Tezi, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Sakarya.

Gürbüz, E. (2023). *Endüstriyel Sistemlerde Veri Madenciliği Yaklaşımlarının Kullanımı ve Bir Uygulama*. Yüksek Lisans Tezi, Bursa Uludağ Üniversitesi, Fen Bilimleri Enstitüsü, Bursa.

Gürbüz, I. (2022). *Metin Madenciliği Yöntemi ile Araştırma Makalesi Sınıflandırması.* Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Güven, Z. A., Diri, B. ve Cakaloglu, T. (2018, 18-19 Nisan). *n-seviyeli Gizli Dirichlet ayırımı ile Türkçe tivit duygularının sınıflandırılması*. Elektrik-Elektronik, Bilgisayar, Biyomedikal Mühendislikleri Bilimsel Toplantısı, İstanbul, Türkiye.

Han, J., Sun, A., Cong, G., Zhao, W. X., Ji, Z. and Phan, M. C.,(2017). Linking fine-grained locations in user comments. *IEEE Transactions on Knowledge and Data Engineering*, *30*(1), 59-72.

Hariri, R. H., Fredericks, E. M. and Bowers, K. M. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data*, *6*(1), 1-16.

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T. and Yeganegi, M. R. (2020). Text mining in big data analytics. *Big Data and Cognitive Computing*, *4*(1), 1.

Hussain, S., Wang, G., Jafar, R. M. S., Ilyas, Z., Mustafa, G. and Jianzhou, Y. (2018). Consumers online information adoption behavior: motives and antecedents of electronic word of mouth communications. *Computers in Human Behavior*, 80, 22-32.

Irfan, R., King, C. K., Grages, D., Ewen, S., Khan, S. U., Madani, S. A., Kolodziej, J., Wang, L., Chen, D. and Rayes, A. (2015). A survey on text mining in social networks. *The Knowledge Engineering Review*, 30(2), 157-170.

Irfan, R., King, C. K., Grages, D., Ewen, S., Khan, S. U., Madani, S. A., Kolodziej, J., Wang, L., Chen, D., Rayes, A. Tziritas, N., Xu, C.Z., Zomaya, A. Y., Alzahrani, A.S. and Li, H. (2015). A survey on text mining in social networks. *The Knowledge Engineering Review*, *30*(2), 157-170.

Isaeva, E. and Aldarova, D. (2021). *Text-mining in terms of methodology and development.* IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), 26-29 January, St. Petersburg and Moscow, Russia.

İçöz, E. (2021). *Covid-19 Pandemi Sürecinde Milli Eğitim Bakanı' Nın Twitter Mesajlarının Metin Madenciliği Yöntemiyle İncelenmesi.* Yüksek Lisans Tezi, Akdeniz Üniversitesi, Eğitim Bilimleri Enstitüsü, Antalya.

İnan, O. (2015). *Veri Madenciliği Uygulamaları için Veri İndirgeme Algoritmalarının Geliştirilmesi ve Resim Madenciliğine Uygulanması*. Doktora Tezi, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Konya.

İnternet: Udemy. (2022). About Udemy. Web: https://about.udemy.com/, Son Erişim Tarihi: 10.09.2023.

Kaçdıoğlu, S. (2020). *Otizm Spektrum Bozukluklarında Kendini Uyarıcı Hareketlerin Görsel Kelime Çantası Yaklaşımı ve Derin Öğrenme Yöntemleri ile Tanınması*. Yüksek Lisans Tezi, Atatürk Üniversitesi, Fen Bilimleri Enstitüsü, Erzurum.

Kapukaya, N. (2023). *Metin Madenciliği ve Duygu Analizi: IMDB En İyi Üç Filmin Twitter Yorumlarının Analizi.* Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Karaosmanoğlu, A. (2022). *Gizli Dirichlet Ayırımı Kullanılarak Covid-19 Salgını Döneminde Yayınlanan İnternet Haberlerinin Konu Modellemesi*. Yüksek Lisans Tezi, Doğuş Üniversitesi, Lisansüstü Eğitim Enstitüsü, İstanbul.

Kartal, Y. (2017). *TOJDE Dergisi Üzerinde LDA ile Konu Modelleme.* Yüksek Lisans Tezi, Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, Eskişehir.

Kavak, D. F. (2022). *Sosyal Medyada Gıda Sektörünün Netnografi Ve Metin Madenciliği Yöntemi İle İncelenmesi: Torku Markasının Twitter Analizi.* Yüksek Lisans Tezi, Aksaray Üniversitesi, Sosyal Bilimler Enstitüsü, Aksaray.

Kiliç E., Ateş, N. and Karakaya, A., (2015, 16-19 May). *Two new feature extraction methods for text classification: TESDF and SADF*. 23th Signal Processing and Communications Applications Conference (SIU), Malatya, Türkiye.

Koç, A. (2023). *Ortaokul Mezunlarının Ortaöğretimi Tamamlama Durumlarının Veri Madenciliği İle Tahmini: Trabzon İli Örneği.* Yüksek Lisans Tezi, Afyon Kocatepe Üniversitesi, Fen Bilimleri Enstitüsü, Afyon.

Kotler, P., Armstrong, G., Saunders, J. and Wong, V. (1999). *Principles of Marketing.* Prentice Hall, Harlow, England.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L. and Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), 150.

Kurt, L., Güldal, O. ve Batmaz, İ. (2022). Bilgi yönetimi bağlamında metin madenciliği teknikleri ile dijital içerik analizi. *Türk Kütüphaneciliği*, *36*(4), 472-494.

Liu, X., Singh, P. V. and Srinivasan, K. (2016). A structured analysis of unstructured big data. *Leveraging Cloud Computing. Marketing Science*, *35*(3), 363-388.

Mecca, G., Raunıch, S. and Pappalardo, A. (2007). A new algorithm for clustering search results. *Data & Knowledge Engineering*, *62*(3), 504-522.

Mei, Q., Shen, X. and Zhai, C. (2007). *Automatic labeling of multinomial topic models*. 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 12-15 August, California, USA.

Miner, G., Delen, D., Elder, J., Fast, A., Hill, T. and Nisbet, R. (2012). *Practical Text Mining and Statistical analysis for Non-Structured Text Data Applications*. Waltham, USA: Elsevier.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, *30*(1), 3-26.

Namkung, Y. and Jang, S. S. (2007). Does food qualıty really matter in restaurants? Its ımpact on customer satısfactıon and behavıoral ıntentions. *Journal of Hospitality & Tourism Research, 31*(3), 387-410.

Nasiboğlu, R. ve Gencer, M. (2023). Adlandırılmış varlık tanıma modelleri ile Türkçe sosyal medya metinlerinde küfürlü sözlerin sansürlenmesi. *Afyon Kocatepe Üniversitesi Fen Ve Mühendislik Bilimleri Dergisi*, *23*(1), 72-88.

Nasukawa T. and Yi, J. (2003, 23-25 October). *Sentiment analysis: Capturing favorability using natural language processing*. 2nd International Conference on Knowledge Capture, Florida, USA.

Nilashi, M., Ahmadi, N., Arji, G., Alsalem, K. O., Samad, S., Ghabban, F., Alzahrani, A. O., Ahani, A. and Alarood, A. A. (2021). Big social data and customer decision making in vegetarian restaurants: A combined machine learning method. *Journal of Retailing and Consumer Services, 62*, 3-7.

Okatan, S. B. (2023). *Dijital Bankacılık Uygulamalarına İlişkin Müşteri Yorumlarının Metin Madenciliği Yöntemleri ile İncelenmesi.* Yüksek Lisans Tezi, Gümüşhane Üniversitesi, Lisansüstü Eğitim Enstitüsü, Gümüşhane.

Olesen, C. G. and Kisjes, I. (2018). From text mining to visual classification: Rethinking computational new cinema history with Jean Desmet's digitised business archive. *TMG Journal for Media History*, *21*(2), 127-145.

Onan, A., Korukoğlu, S. and Bulut, H. (2017). A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. *Information Processing and Management*, 53, 814-833.

Özdemir, E. (2022). *Havayolu Yolcularının Çevrimiçi Değerlendirmelerinin Metin Madenciliği ile Analizi.* Doktora Tezi, Anadolu Üniversitesi, Sosyal Bilimler Enstitüsü, Eskişehir.

Özgül, F. (2013). *Her Yönüyle Python* (2. Baskı). İstanbul: Kodlab Yayıncılık.

Öztemel, E. (2003). *Yapay Sinir Ağları*. İstanbul: Papatya Yayıncılık.

Öztürk, S., Sankur, B., Güngör, T. and Yılmaz, M., B. (2014, 23-25 April). *Turkish labeled text corpus*. 22nd Signal Processing and Communications Applications Conference (SIU), Trabzon, Türkiye.

Öztürk, Y. (2021). *Etkileşimli Tahta Kullanımına Yönelik Sosyal Medya Yorumlarının Metin Madenciliği Yöntemleri İle Analizi*. Yüksek Lisans Tezi, Sivas Cumhuriyet Üniversitesi, Sosyal Bilimler Enstitüsü, Sivas.

Özyurt, B. ve Akçayol, M. A. (2017). Fikir madenciliği ve duygu analizi, yaklaşımlar, yöntemler üzerine bir araştırma. *Selçuk Üniversitesi Mühendislik, Bilim ve Teknoloji Dergisi*. *6*(4), 668-693.

Paftalı, S. (2021). *Bireysel Kredilerini Erken Kapatan Müşterilerin Veri Madenciliği Yaklaşımı İle Analizi.* Yüksek Lisans Tezi, Bahçeşehir Üniversitesi, Lisansüstü Eğitim Enstitüsü, İstanbul.

Pang, B., Lee, L. and Vathyanathan, S. (2002, 7-11 December). *Thumbs up? Sentiment classification using machine learning techniques*. The Conference on Empirical Methods in Natural Language Processing (EMNLP), Abu Dhabi, United Arab Emirates.

Pasin, E. (2018). *Investigation of Text Mining Methods on Turkish Text*. Yüksek Lisans Tezi, Dokuz Eylül Üniversitesi, Fen Bilimleri Enstitüsü, İzmir.

Phan, X-H., Nguyen, C.T., Le, D.T., Nguyen, L.M., Horiguchi, S. and Ha, Q.T. (2011). A hidden topic-based framework toward building applications with short web documents. *IEEE Transactions on Knowledge and Data Engineering*, *23*(7), 961-976.

Richards, J. D., Tudhope, D. and Vlachidis, A. (2015). Text mining in archaeology: Extracting information from archaeological reports. In J. Barcelo, and I. Bogdanovic (Eds.), *Mathematics and Archaeology*. Boca Raton, FL: CRC Press.

Romero, C. and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1355.

Savaş, S., Topaloğlu, N. ve Yılmaz, M., (2012). Veri madenciliği ve Türkiye'deki uygulama örnekleri. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, *11*(21), 1-23.

Sohrabi, B., Raeesi Vanani, I., Nikaein, N. and Kakavand, S. (2019). A predictive analytics of physicians prescription and pharmacies sales correlation using data mining. *International Journal of Pharmaceutical and Healthcare Marketing*, *13*(3), 346-363.

Sönmez, N. (2017). *Çevrimiçi Yorumların Metin Madenciliği ile Analizi: İstanbul'daki Alışveriş Merkezleri Üzerine Bir Çalışma*. Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Şeker, S. E. (2015a). Metin madenciliği (text mining). *YBS Ansiklopedi*, *2*(3), 32-33.

Şeker, Ş. E. (2015b). Doğal dil işleme. *YBS Ansiklopedi*, *2*(4), 14-22.

Şen, F. (2022). *Examination of The Postgraduate Theses Written on Education Management in Turkey by Text Mining Method.* Yüksek lisans tezi, Bahçeşehir Üniversitesi, Lisansüstü Eğitim Enstitüsü, İstanbul.

Takan (2022). *Sosyoenformatik Yaklaşımdan Yaşlılık Temalı Filmlerde Duygu Yapısı: Metin Madenciliği Yöntemi Üzerinden Bir Analiz.* Doktora Tezi, Ege Üniversitesi, Sosyal Bilimler Enstitüsü, İzmir.

Takan E., D. (2022). *Sosyoenformatik Yaklaşımdan Yaşlılık Temalı Filmlerde Duygu Yapısı: Metin Madenciliği Yöntemi Üzerinden Bir Analiz.* Doktora Tezi, Ege Üniversitesi, Sosyal Bilimler Enstitüsü, İzmir.

Talib, R., Hanif, M. K., Ayesha, S. ve Fatima, F. (2016). Text mining: techniques, applications and issues. *International Journal of Advanced Computer Science and Applications*, *7*(11), 414-418.

Taş, A. ve Bülbül, H. İ. (2021). Sosyal medya kullanıcılarının dijital ayak izi farkındalığı. *Sosyal, Beşeri ve İdari Bilimler Dergisi*, *4*(3), 205-216.

Xiang, Z., Du, Q., Ma, Y. and Fan, W. (2017). A comparative analysis of major online review platforms: Implications of social media analytics in hospitality and tourism. *Tourism Management*, 58, 51-65.

Yalçın, L. (2019). *Sağlık Sektöründe Veri Madenciliği.* Yüksek Lisans Tezi, Milli Savunma Üniversitesi, Hezarfen Havacılık ve Uzay Teknolojileri Enstitüsü, İstanbul.

Yan, H., Yang, N., Peng, Y. and Ren, Y. (2020). Data mining in the construction industry: Present status, opportunities, and future trends. *Automation in Construction*, 119, 103331.

Yüksel, A. S. ve Tan, G. F. (2018). Metin madenciliği teknikleri ile sosyal ağlarda bilgi keşfi. *Mühendislik Bilimleri ve Tasarım Dergisi*, *6*(2), 324-333.

Gazili olmak ayrıcalıktır