BURSA TEKNİK ÜNİVERSİTESİ ❖ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

BİR OTOMOTİV FİRMASINDA KONU MODELLEME YAKLAŞIMI KULLANILARAK ÇALIŞAN ÖNERİLERİNİN DEĞERLENDİRİLMESİ

YÜKSEK LİSANS TEZİ

Mine BOZAN

Akıllı Sistemler Mühendisliği Anabilim Dalı

Akıllı Sistemler Mühendisliği Programı

OCAK, 2022

BURSA TEKNİK ÜNİVERSİTESİ ❖ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

BİR OTOMOTİV FİRMASINDA KONU MODELLEME YAKLAŞIMI KULLANILARAK ÇALIŞAN ÖNERİLERİNİN DEĞERLENDİRİLMESİ

YÜKSEK LİSANS TEZİ

Mine BOZAN
(181324814001**)**

Akıllı Sistemler Mühendisliği Anabilim Dalı

Tez Danışmanı: Dr. Öğr. Üyesi KORAY ALTUN

OCAK, 2022

BTÜ, Lisansüstü Eğitim Enstitüsü'nün 181324814001 numaralı Yüksek Lisans Öğrencisi Mine BOZAN, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı "BİR OTOMOTİV FİRMASINDA KONU MODELLEME YAKLAŞIMI KULLANILARAK ÇALIŞAN ÖNERİLERİNİN DEĞERLENDİRİLMESİ" başlıklı tezini aşağıda imzaları olan jüri önünde başarı ile sunmustur.

Tez Danışmanı : **Dr. Öğr. Üyesi Koray ALTUN** .....
Bursa Teknik Üniversitesi

Jüri Üyeleri : **Doç. Dr. Serkan ALTUNTAŞ** .....
Yıldız Teknik Üniversitesi

Doç. Dr. Gökay BAYRAK .....
Bursa Teknik Üniversitesi

Teslim Tarihi : 15 Şubat 2022
Savunma Tarihi : 18 Ocak 2022

20.04.2016 tarihli Resmi Gazete'de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, Bursa Teknik Üniversitesi'nin abonesi olduğu intihal yazılım programı kullanılarak Lisansüstü Eğitim Enstitüsü'nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

İNTİHAL BEYANI

Bu tezde görsel, işitsel ve yazılı biçimde sunulan tüm bilgi ve sonuçların akademik ve etik kurallara uyularak tarafımdan elde edildiğini, tez içinde yer alan ancak bu çalışmaya özgü olmayan tüm sonuç ve bilgileri tezde kaynak göstererek belgelediğimi, aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul ettiğimi beyan ederim.

Mine BOZAN

Anneme, babama, kardeşime; tüm aileme,

ÖNSÖZ

Tez danışmanım Dr. Öğr. Üyesi Koray Altun'a desteği ve yönlendirmesiyle tezi hazırlamamda yardımcı olduğu için teşekkürlerimi sunarım.

Aralık 2021

Mine Bozan
(Endüstri Mühendisi)

İÇİNDEKİLER

Sayfa

KISALTMALAR

ADM-LDA **:** Aspect Detection Model-LDA (Yön AlgılamaModeli LDA)

AEP-LDA **:** Appraisal Expression Pattern-LDA (Değerleme İfade Modeli LDA)

Ar-Ge **:** Araştırma ve Geliştime

ASUM **:** Aspect Sentiment Unification (Yön Duyarlılığı Birleştirme)

DTAS **:** Dependency - Topic - Affects - Sentiment – LDA (Bağımlılık - Konu - Etkiler - Duyarlılık – LDA)

FLDA **:** Factorized LDA (Faktörleştirilmiş LDA)

GDT **:** Gizli Dirichlet Tahsisi

HDP-LDA **:** Hierarchical Drichlet Process-Latent Dirichlet Allocation (Hiyerarşik Drichlet Süreci LDA)

JAS **:** Joint Aspect/Sentiment (Ortak Yön/Duygu)

LDA **:** Latent Dirichlet Allocation (Gizli Dirichlet Tahsisi)

LSA **:** Latent Semantic Analysis (Gizli Anlamsal Analiz)

PLSA **:** Probabilistic Latent Semantic Analysis (Olasılıksal Gizli Anlamsal Analiz)

S-LDA **:** Sentence-LDA (Cümle LDA)

UFL-LDA **:** Unified Fine-grained Labeled-LDA (Birleşik Detaylı Etiketli LDA)

ŞEKİL LİSTESİ

Sayfa

BİR OTOMOTİV FİRMASINDA KONU MODELLEME YAKLAŞIMI KULLANILARAK ÇALIŞAN ÖNERİLERİNİN DEĞERLENDİRİLMESİ

ÖZET

Veri madenciliğinin bir alt kolu olan metin madenciliği, birçok dilde birçok farklı alanda yazılmış metinlerde özellik çıkarımı için kullanılmaktadır. Son yıllarda önem kazanan konu modelleme yöntemleri metin madenciliği uygulamalarında sıklıkla tercih edilmeye başlanmıştır. Konu modelleme, veri madenciliği, gizli veri keşfi ve veri ile metin belgeleri arasındaki ilişkileri bulmak için metin madenciliğinde en güçlü tekniklerden biridir. Büyük boyutlu dokümanlardaki gizli yapıyı denetimsiz şekilde ortaya çıkaran konu modelleme, başarılı bir yöntem olarak kabul görmüştür.

Bir doküman kümesi içerisindeki belgelerin konu olarak adlandırılan sözcük grupları dokümanlar içerisinde gizli ve yapılandırılmamış bir biçimde bulunur. Bu konuların özelliği metin içerisinde çoğunlukla birlikte görülmeleri ve genellikle ortak veya birbirine benzeyen temayı paylaşan sözcüklerden oluşmalarıdır. Konu modellemede metinde sıkça birlikte görülen kelimeler kümelenerek soyut konular(abstract topic) üretilir ve ilgili metinler içerdikleri kelimelere göre kendisine en yakın olan bir veya daha fazla kümede konuçlandırılır. Birbirine yakın olabilecek ifadeler anlamsal uzayda soyut konu oluşturulacak şekilde gruplanırlar. Daha sonra bu dokümanlar oluşturulan gruplara ve içerdikleri kelimelere göre kümelenir.

Günümüzün en büyük endüstrilerinden biri olan otomotiv endüstrisinde rekabetin fazla olması işletmelerin iyileştirme çalışmalarına önem vermeleri gerekliliğini doğurmuştur. Bu iyileştirmelerin arasında çalışanların önerileri oldukça önemli yer kaplamaktadır.Öneri sistemlerinin içeriğinin metinlerden oluşması, onları ileri metin madenciliği çalışmaları için uygun veri setleri haline getirmiştir.

Çalışan önerilerinin konu modelleme ile analiz edilmesi, en çok hangi konularda öneriler geldiğini, hangi konulara yoğunlaşılması gerektiğini ve gelecekteki öneri ve iyileştirmelerle ilgili tahminler yapabilmeyi olanaklı hale getirebilecektir.

Bu çalışmada bir otomotiv firmasının öneri sistemlerini analiz etmek için metin madenciliğine ait yöntemlerden Gizli Dirichlet Tahsisi (GDT) kullanılmıştır. Projemizde analiz için Azure Machine Learning aracı kullanılmıştır.

En çok verilen öneri çeşiti “getirisi olmayan olumlu öneri”lerdir. Bu öneriler genellikle iş sağlığı ve güvenliği ile ilgili olmaktadır. 2. Sıradaki en çok verilen öneriler ise “öneri” yani getirisi olan, firmaya kazanç sağlayan önerilerdir. 3. Sırada “öneriden hızlı kaizene” yani çok kısa sürede sonuç alınabilen, getirisi yüksek öneriler bulunmaktadır. 4. Sırada “değerlendirilmek üzere havale” edilen öneriler bulunurken, en az verilen öneri türünün ise “devreye alınmayacak öneriler” olduğu görünmektedir.

Anahtar kelimeler: Konu modelleme, Gizli dirichlet tahsisi, LDA, Öneri sistemleri

EVALUATION OF EMPLOYEE SUGGESTIONS BY USING TOPIC MODELING APPROACH IN AN AUTOMOTIVE COMPANY

SUMMARY

Text mining, a sub-branch of data mining, is used for feature extraction in texts written in many languages and in many different fields. Subject modeling methods, which have gained importance in recent years, have started to be preferred frequently in text mining applications. Topic modeling is one of the most powerful techniques in text mining for data mining, discovery of hidden data, and finding relationships between data and text documents. Topic modeling, which reveals the hidden structure of large documents in an unsupervised manner, has been accepted as a successful method.

Phrases called subjects of documents in a set of documents are found in documents in a hidden and unstructured form. The peculiarity of these topics is that they are often seen together in the text and usually consist of words that share a common or similar theme. In topic modeling, abstract topics are produced by clustering the words that are frequently seen together in the text, and the related texts are positioned in one or more clusters closest to the words they contain. Expressions that may be close to each other are grouped in the semantic space to form an abstract subject. These documents are then clustered according to the groups created and the words they contain.

The high competition in the automotive industry, which is one of the largest industries of today, necessitated the businesses to attach importance to improvement studies. Among these improvements, the suggestions of the employees occupy a very important place. The fact that the content of the recommendation systems consists of texts has made them suitable data sets for advanced text mining studies.

Analyzing employee suggestions with subject modeling will make it possible to make suggestions about which subjects are the most, which subjects should be focused on, and to make predictions about future suggestions and improvements.

In this study, Latent Dirichlet Allocation (LDA), one of the text mining methods, was used to analyze the recommendation systems of an automotive company. Azure Machine Learning tool was used for analysis in our project.

The most common type of suggestion is “positive suggestions with no return”. These recommendations are generally related to occupational health and safety. 2. The most frequently given suggestions are "suggestions", that is, those that have returns and provide profit to the company. In the 3rd rank, there are “fast kaizene suggestions”, that is, high-yielding suggestions that can be achieved in a very short time. 4. Suggestions that are "referred to be evaluated" are in the rank, while "recommendations that will not be put into action" seem to be the least given type of suggestion.

Keywords: Topic modeling, Latent Dirichlet Allocation, LDA, Recommendation systems

1. GİRİŞ

Günümüzün en büyük endüstrilerinden biri olan otomotiv endüstrisinde rekabetin fazla olması işletmelerin iyileştirme çalışmalarına önem vermeleri gerekliliğini doğurmuştur. Bu iyileştirmelerin arasında çalışanların önerileri oldukça önemli yer kaplamaktadır. Öneri sistemleri, işletmelerde yapılan işler için o işleri bizzat yapan çalışanların işi en iyi bilen kişiler olduğu düşüncesinden yola çıkılarak oluşturulmuştur. İşletmenin her kademesinde çalışan mavi ve beyaz yaka personellerin katkılarıyla uygulanmaktadır. Çalışan öneri sistemi yenilikçi fikirlerin ortaya çıkmasını, uygulanmasını ve firmanın sürekli iyileştirilmesini sağlamaktadır.

Japonya'da 452 işletmede bir araştırma yapılmıştır ve bu araştırmaya göre bu şirketlerde 23 milyon 532 bin fikir öne sürülmüş, bu fikirlerin yaklaşık olarak yarısı hayata geçirilmiştir. Önerilerden elde edilen kar ise 225,3 milyon yene ulaşmıştır.

Firmanın çalışanlarını iyileştirme önerileri vermeye teşvik etmesi, çalışanların motivasyonlarının artmasına, çalıştıkları firmaya kendilerini ait hissetmelerine yardımcı olmaktadır. İşletmenin gelişime açık konuları ve öne sürülen yenilikçi fikirler öneri sistemi sayesinde ortaya çıkmaktadır ve bunlar firmanın Ar-Ge projelerine katkı sağlamaktadır.

Öneriler metinlerden oluşmaktadır ve gün geçtikçe sayıları artmaktadır. Bu da veri birikiminin artmasına dolayısıyla depolanan verinin boyutun da her geçen gün artmasına neden olmaktadır. Depolanan veriler yüksek oranda metin verisi olduğundan, metin verilerinin otomatik analiz edilmesi önemli bir araştırma konusuna dönüşmektedir. Bundan dolayı büyük ölçekteki metinlerde aramak, anlamak ve işlemek gibi işlemleri yapabilecek otomatik araçların ihtiyacı doğmuştur. Bu ihtiyaç doğrultusunda konu modelleme yöntemlerinden olan makine öğrenmesi, doğal dil işleme, bilgi çıkarımı fazlasıyla yaygınlaşmış şekilde kullanılmaktadır. Konu modelleme, büyük miktardaki metin kaynaklarından anlamlı bilgilere erişebilmek için uygulanan denetimsiz bir makine öğrenmesi yöntemidir.

Konu modelleme, veri madenciliği, gizli veri keşfi ve veri ile metin belgeleri arasındaki ilişkileri bulmak için metin madenciliğinde en güçlü tekniklerden biri olmakla birlikte makine öğreniminde ve doğal dil işleme süreçlerinde doküman kümesinin içerisindeki soyut konuların araştırılarak analiz edildiği istatiksel bir modeldir. Doküman kümesini en iyi şekilde karakterize eden sözcük grupları ile benzer ifadeleri otomatik olarak kümeleyebilen gözetimsiz (unsupervised) öğrenme olan bir makine öğrenmesi tekniğidir.

Konu modelleme genel olarak bir dokümanın konusunun ne hakkında olduğuyla ilgili fikir verir. Bir doküman kümesi içerisindeki belgelerin konu adı verilen sözcük grupları dokümanlar içerisinde gizli ve yapılandırılmamış bir biçimde bulundurulur. Bu bahsedilen konuların özelliği metinlerin içinde sıklıkla bir arada görülmeleri ve genel olarak ortak veya birbirine benzeyen temayı paylaşan sözcüklerden oluşmuş olmalarıdır. Konu modellemede metinde sıkça birlikte görülen kelimeler kümelenerek soyut konular(abstract topic) üretilir ve ilgili metinler içerdikleri kelimelere göre kendisine en yakın olan bir veya daha fazla kümeye atanır.

Literatürde, araştırmacılar tarafından geliştirilen mevcut pek çok konu modeli bulunmaktadır. Bu konu modelleri arasında en yaygın kullanılan yöntemlerden birisi Gizli Dirichlet Tahsisi (Latent Dirichlet Allocation, LDA) yöntemidir [1-5]. LDA, metin dökümanları gibi ayrık verilerin modellenmesi için geliştirilmiş olan grafiksel bir modeldir ve metin belgesinde bulunan gizli konuların açığa çıkarılmasını amaçlar [6]. LDA'nın ana fikri, konuların sabit bir kelime sözlüğünden olasılık dağılımını içerdiği ve belgelerin gelişigüzel bir şekilde birleşmiş gizli konulardan oluştuğudur. Bu temel fikir, LDA'nin belge koleksiyonunda bulunan konuları, kelimelerin konuların içindeki olasılıklarını, belgeyi oluşturan kelimelerin atandığı konuları ve bu belge içindeki konuların nasıl dağıldığını öğrenerek ortaya çıkardığını belirtir [7].

Çalışan önerilerinin konu modelleme ile analiz edilmesi, en çok hangi konularda öneriler geldiğini, hangi konulara yoğunlaşılması gerektiğini ve gelecekteki öneri ve iyileştirmelerle ilgili tahminler yapabilmeyi olanaklı hale getirebilecektir.

Daha önceki metin madenciliği çalışmalarında öneri sistemlerine hiç değinilmediği için bu çalışma literatürdeki ilk çalışma olma özelliği taşımaktadır.

1.1 Tezin Amacı

Bu çalışmada otomotiv sektöründe üretim yapan bir şirketin 2021 yılına ait çalışan önerileri SAP üzerinden otomatik tablolar halinde alınmıştır ve metin madenciliği yöntemlerinden olan Konu modelleme-LDA yöntemleriyle analiz edilecektir.

Amaç, tesisin ürün üretimindeki genel duyarlılığını sağlamak için üretim sahasındaki önerilerin toplanarak müşteri geri bildirimlerini ve çalışanların önerilerini analiz eden bir işlem hattı oluşturarak üretimi 0 hata ile gerçekleştirmek ve müşteri memnuniyetini en yüksek seviyede tutmaktır.

Projemizde Azure Machine Learning aracı kullanılmıştır. Azure Machine Learning, veri bilimcilerin kolayca bir metin sınıflandırma çözümü oluşturmasına ve dağıtmasına yardımcı olacak bir çok araç ve şablon sağlamaktadır.

Literatürdeki çalışmalarda çalışan memnuniyeti konularına değinilmiştir fakat sürekli iyileştirme çalışmalarına değinilmemiştir, bu bağlamda bu çalışma literatürde ilk olma özelliği taşımaktadır.

1.2 Literatür Araştırması

Son yirmi yılda, LSA (Latent Semantic Analysis) (Gizli Anlamsal Analiz), PLSA (Probabilistic Latent Semantic Analysis) (Olasılıksal Gizli Anlamsal Analiz) ve LDA (Gizli Dirichlet Tahsisi) dahil olmak üzere çeşitli konu modelleme tekniklerinin araştırıldığı makale ve araştırma sonuçları olmuştur. Matematikçiler, istatistikçiler, bilgisayar bilimcileri, biyologlar ve sinirbilimciler gibi çeşitli alanlardan araştırmacılar, konu modellemeyi farklı açılardan araştırmışlardır. Konu modelleme alanındaki başlıca araştırmacılar Michael I. Jordan, David M Blei, Hanna M. Wallach, Thomas L. Griffiths, Jordan Boyd Graber, D. Mimno, Mark Johnson, Jonathan Chang ve Katherine Heller'dir. Jordan Boyd Graber, çeşitli saygın dergilerde ve uluslararası konferanslarda yaklaşık 200 yayını ile 2003'ten 2017'ye kadar sürekli araştırma yayınları yapmıştır. 2007 yılında David M. Blei ile birlikte kelime anlamı belirsizleştirmeleri için konu modellemenin kullanımı üzerine ilk makaleyi yazmıştır. 2008'de çok dilli konu modeli ve sözdizimsel konu modeli kavramını sunmuş ve 2010 yılında konu modellerinin dilsel olarak genişletilmesi üzerine doktorasını tamamlamıştır. Jordan, duygu analizi, çok dilli konu modelleme ve konuşma üzerine

konu modelleme, büyük ölçekli konu modelleme, verimli ağaç tabanlı konu modelleme ve çevrimiçi konu modellemeye büyük ölçüde katkıda bulunmuştur [8].

Konu modellemede ankete odaklanan bazı çalışmalar vardır [9]. Chen T-H. vd. konu modellerinin şimdiye kadar bir veya daha fazla yazılım havuzuna nasıl uygulandığını belirtmek için yazılım mühendisliği alanında konu modelleme üzerine bir anket sunmuştur. Aralık 1999 ile Aralık 2014 arasında yazılan makalelere odaklanmışlar ve yazılım mühendisliği alanında konu modellemeyi kullanan 167 makaleyi incelemişlerdir. Konu modelleriyle, yapılandırılmamış veri ambarlarının metin madenciliğindeki araştırma eğilimlerini belirleyerek göstermişlerdir. Çalışmaların çoğunun yalnızca sınırlı sayıda yazılım mühendisliği görevine odaklandığını ve çoğu çalışmanın yalnızca temel konu modellerini kullandığını bulmuşlardır [10]. Daud A. vd. metin külliyatında yumuşak kümeleme yeteneklerine sahip Konu Modelleri araştırmasına odaklanmış ve temel kavramları ve parametre tahmini (Gibbs Örneklemesi gibi) ve performans değerlendirme ölçütleri ile çeşitli kategorilerde mevcut model sınıflandırmasını araştırmışlardır [11]. Debortoli S. vd. bilgi sistemleri araştırmasında metin madenciliği tekniklerinin zorluklarını tartışmışlar ve örnek bir veri kaynağı olarak çevrimiçi müşteri incelemelerini kullanarak açıklayıcı regresyon analizi ile birlikte konu modellemenin pratik uygulamasını gerçekleştirmişlerdir.

[12]'de yazarlar dört SE dergisinden ve on bir konferans bildirisinden Yazılım Mühendisliği görevlerinde konu modellerinin şimdiye kadar nasıl uygulandığına dair bir anket sunmuşlardır. 2003'ten 2015'e kadar seçilen 38 yayını incelemişler ve sosyal yazılım mühendisliği, geliştirici tavsiyesi vb. gibi çeşitli SE görevlerinde konu modellerinin yaygın olarak kullanıldığını bulmuşlardır.

Korfiatis N., Symitsi E., Stamolampros P., Daskalakis G. ve 2018'de çalışan çevrimiçi incelemelerinin bilgi değeri ile ilgili araştırma gerçekleştirmişlerdir.

[13]'te belgedeki belirgin özelliklerin ortaya çıkarılması için Sentiment LDA ve Dependency Sentiment LDA gibi iki teknik önerilmiştir. Önerilen bu yöntemlerle ürünlerin özellikleri ve duygu ifadeleri birlikte çıkarılmıştır. Bu yöntem ve teknikler; belgedeki duygu belirten ifadelerin temel varlıkla ilişkisinin olduğu düşüncesinden esinlenilerek geliştirilmiş ve duygu ifadeleriyle belirgin özellikler bir bütün gibi ele alınmıştır. Dependency Sentiment LDA'da duygu kutuplarının ortaya çıkarılması amaçlanmıştır. [14]'te yarı-denetimli yöntem olarak bilinen Co-LDA tekniğiyle

ürünlerin özelliklerinin pozitif yada negatif duygusunun ortaya çıkarılması amaçlanmıştır. [15]'te aynı başlık altında bulunan ürünlerin özelliklerinin yorumlarda birbirine yakın oldukları düşüncesiyle bir cümlenin içindeki bütün sözcüklerin bir ürünün özelliğiyle ilgili olduğunu iddia eden S-LDA (Sentence-LDA) (Cümle LDA) yöntemi önerilmiştir ve sonrasında ASUM (Aspect Sentiment Unification) (Yön Duyarlılığı Birleştirme) geliştirilmiştir. [16]'da GDT'yi baz alan ve GDT'nin eksik yönlerini tamamlayan JAS (Joint Aspect/Sentiment) (Ortak Yön/Duygu) modeli önerilmiştir. JAS yöntemiyle, ürünlerin özellikleri kullanıcılar tarafından yapılan yorumlardan çıkarılırken bu özelliklerin duygu ifadelerinin kutuplarıyla aynı anda çıkarılarak özellik bağımlı bir sözlük oluşturmuşlardır. [17]'de Çince yazılan yorumlardan GDT ile genel konular çıkartılmıştır. Yerel konuların ve ilişkili duygu ifadelerinin çıkartılmasındaysa kayan pencere tabanlı yöntem önerilmişir. [18]'de GDT'yi baz alan bileşik bir model olan HDP-LDA (Hierarchical Drichlet Process-Latent Dirichlet Allocation) (Hiyerarşik Drichlet Süreci LDA) önerilmiştir. [19]'da FLDA (factorized LDA) (Faktörleştirilmiş LDA) yöntemi, soğuk başlangıç problemlerinden ürün özelliklerinin çıkarılması amacıyla önerilmiştir. Soğuk başlangıç problemi; öneri sistemlerinde ürünlerin özellikleriyle alakalı yeterince yorumun olmamasından kaynaklıdır. [20]'de yarı denetimli konu modelleme yöntemi olan UFL-LDA (Unified Fine-grained Labeled-LDA) (Birleşik Detaylı Etiketli LDA) ve FL-LDA'yla (Fine-grained Labeled LDA) kullanıcı yorumlarından ürünlerin özelliklerini çıkartmışlardır. [21]'de restoranların, otellerin, vb hizmetlerin yorumlarından AEP-LDA (Appraisal Expression Pattern LDA) (Değerleme İfade Modeli LDA) ile belirgin olan ürün özellikleri kullanıcıların yorumları analiz edilerek çıkartılmıştır. [22]'de ürünler için bulanık ontoloji öğrenmesi yönteminin, belirgin özellikleri ve bu özelliklere dayalı olan duygu ifadelerinin ortaya çıkarılması amacıyla geliştirildiği görülmektedir. [23]'te GDT'yi baz alan DTAS (Dependency Topic Affects Sentiment LDA) (Bağımlılık - Konu - Etkiler - Duyarlılık – LDA) tekniğiyle konular ve aynı zamanda duygu ifadeleri çıkartılmıştır. [24]'te kullanıcıların yorumlarından belirgin olan özellikleri çıkartmak için ADM-LDA (Aspect Detection Model-LDA) (Yön Algılama Modeli LDA) yöntemini önermişlerdir. [25]'te GDT'de bulunan kelimelerin dağılımlarını hesaplamak üzere yeni bir teknik geliştirilmiştir. Bu yöntemle Sentic-LDA söz dizilimine bağlı yaklaşımı kullanılmaktansa anlamsal yaklaşım uygulanarak özellik tabanlı duygu analizi gerçekleştirilmektedir [26].

Literatürde, araştırmacılar tarafından geliştirilen mevcut pek çok konu modelleme yöntemleri bulunmaktadır. Vayansky ve Kumar konu modelleme yöntemleri üzerine kapsamlı ve güncel bir literatür taraması sunmaktadır. Vayansky ve Kumar çalışmalarında bir karar ağacı da sunmaktadır [27]. Bu karar ağacı, hangi konu modelleme yönteminin kullanılması gerektiği hususunda, yol gösterici bir rehber niteliğindedir.

Konu modelleme yöntemleri arasında en yaygın kullanılanı; "Gizli Dirichlet Tahsisi (LDA – Latent Dirichlet Allocation)" yöntemidir. Üzerinde çalışılan belgelerdeki kelime sayılarının 50'den fazla olduğu ve çalışılan alan itibariyle kompleks konu ilişkilerinin öngörülmediği durumlar için LDA kullanımı önerilmektedir. Bu çalışma kapsamında da LDA yönteminin kullanımı tercih edilmiştir.

2. VERİ MADENCİLİĞİ

Bilgisayar sistemleri ile üretilen veriler tek başlarına değersizdir, çünkü çıplak gözle bakıldığında bir anlam ifade etmezler. Bu veriler belli bir amaç doğrultusunda işlendiği zaman bir anlam ifade etmeye başlar [28]. Bilgi bir amaca yönelik işlenmiş veridir. "Ham veri" veya "enformasyon"a bağlı olarak karar alınması mümkün değildir. Önemli olan geçmişe ait olaylara dair gizli bilgileri keşfederek, ileriye yönelik durumsal öngörüler veren modeller ile önceden tedbir almamızı sağlayacak bir yönetim anlayışına geçmek ve olası kayıpları öngörebilmektir [29]. Bundan dolayı çok miktardaki verilerin işlenebilmesini sağlayan tekniklerin kullanılabilmesi oldukça önemli hale gelmektedir. Ham verinin bilgiye ya da anlam ifade eden biçime dönüştürülme işlemlerinin veri madenciliğiyle yapılabileceği görülmektedir [28]. Büyük miktardaki verilerde bulunan örüntüleri ve eğilimleri keşfetmek veri madenciliğinin tanımıdır [30].

Günümüzde işletmeler için veri madenciliğinin oldukça önemli olduğu görülmektedir. Çok büyük ölçekteki verilerin analiz edilmesi ve bu analizin sonrasında anlam ifade eden bilgilerin elde edilmesi ve bu bilgilerin yorumlanması insan yeteneğini ve ilişkisel veri tabanlarının yeteneklerini aşmaktadır. Bu ihtiyaçlarla birlikte akıllı ve otomatik veri tabanları analizlerini gerçekleştirmek üzere yeni yöntemler ortaya çıkmıştır. Bahsedilen yeni yöntemler verinin akıllı ve otomatik olarak yararlı bilgiye dönüşmesini sağlamalıdır. Veri madenciliği bunlara çözüm olarak ortaya atılmıştır ve giderek önemi artan bir araştırma konusu olmuştur [31].

Veri madenciliği, büyük miktardaki verinin içinden geleceğin tahminlenmesine katkıda bulunacak yarar sağlayan ve anlam ifade eden bağlantıların ve kuralların bilgisayar programları vasıtasıyla aranarak analiz edilmesidir. Dahası, veri madenciliği, yüksek miktarlardaki verinin içerdiği ilişkileri inceleyen, aralarında bulunan bağlantının bulunmasına katkıda bulunan ve veri tabanı sistemlerinde henüz açığa çıkmamış olan bilgilerin çıkarılmasında rol oynayan veri analizi tekniğidir [28]. Uygulama alanı oldukça geniş olan bu işlemler veri tabanı sistemi, yapay sinir ağı, veri görselliği, istatistik gibi disiplinlerden oluşmaktadır.

Veri madenciliğine ait araçların kullanılmasıyla, işletmelerin daha etkin ve etkili kararları alması için karar destek sistemlerinin gerektirdiği davranış kalıpları ortaya

çıkarılabilir hale gelmektedir. Eski karar destek sistemleri tarafından kullanılan yöntemlerden ziyade otomatik ve kapsamlı analizlerin yapılabilmesi için farklı özellikler bulunmaktadır [29].

Veri madenciliğinin işletmeler için önerdiği en önemli özelliklerden biri, veri gruplarının arasında bulunan birbirine benzeyen eğilimleri belirlemesidir. Bu özellik, özellikle hedef pazarlara yönelik yapılan pazarlamada sıklıkla kullanılmaktadır [29]. Başka bir özelliğiyse önceden bilinmeyen, veri ambarları içinde bulunmasına rağmen ilk etapta görülememiş olan bilgileri ortaya çıkarabilmesidir. Örnek vermek gerekirse bir şirket satmış olduğu ürünlerin analizin yapılmasıyla, daha sonra yapması gereken kampanyaları belirleyebilir veya sattığı ürünlerin arasında bulunan etkileşimleri öğrenebilir.

Geçmiş yıllarda geliştirilen gelişmiş veri tabanı teknolojisinin geniş kullanımı sayesinde, büyük hacimli verileri bilgisayarlarda verimli bir şekilde depolamak ve gerektiğinde geri almak zor değildir. Depolanan veriler bir kuruluşun değerli bir varlığı olmasına rağmen, çoğu kuruluş er ya da geç veri bakımından zengin ancak bilgiden yoksun olma sorunuyla karşı karşıya kalabilir. Bu durum, son zamanlarda veri madenciliği alanındaki araştırma ilgilerinin artmasına neden olmuştur [32,33,34].

2.1. Veri Madenciliği Metotları

Veri madenciliğinde kullanılan birçok yönteme yeni yöntemler ve algoritmalar eklenmektedir. Bunların büyük bir kısmı istatistiksel yöntemlerdir. Diğer yöntemlerinse çoğunlukla istatistik bilimini temel alan ve makine öğrenimiyle ve yapay zekayla desteklenen yeni türden yöntemler olduğu söylenebilir. Veri madenciliğine ait modeller işlevlerine göre 3 gruba ayrılır.

Bu modeller aşağıdaki 3 ana başlık altında incelenebilmektedir:

1. Sınıflama ve Regresyon (Classification and Regression),

2. Kümeleme (Clustering),

3. Birliktelik Kuralları (Association Rules),

Sınıflama ve regresyon modellerinin tahminleme modelleri, kümeleme ve birliktelik kuralları modellerinin tanımlayıcı modeller olduğu söylenebilir [35].

2.1.1. Sınıflama ve regresyon

Sınıflama ve regresyon, kayda değer öneme sahip olan veri sınıflarının ortaya konduğu ya da geleceğe ait veri eğilimlerinin tahmin edildiği modelleri kurabilen iki veri analizi yöntemidir [36]. Regresyon sürekli değerlerin tahmininde kullanılırken, sınıflamaysa kategorik değerlerin tahmininde kullanılmaktadır [36]. Örnek olarak, sınıflandırma modeli bir üretim hattı uygulamasında üretilen parçaların OK veya NOK olma durumlarını kategorilere ayırmak için kullanılırken, regresyon modeliyse mesleğiyle birlikte geliri bilinen potansiyel müşteri adaylarının telefon ile ilgili ürünlerin alışverişini yaparken yapacakları harcamaların tahmin edilmesi amacıyla kullanılabilir. Sınıflama ve regresyon modellerinin uygulanmasında aşağıdaki yöntemler kullanılmaktadır [37].

1 - Karar Ağaçları (Decision Trees)

2- Yapay Sinir Ağları (Artificial Neural Networks)

3- Genetik Algoritmalar (Genetic Algorithms)

4- K-En Yakın Komşu (K-Nearest Neighbor)

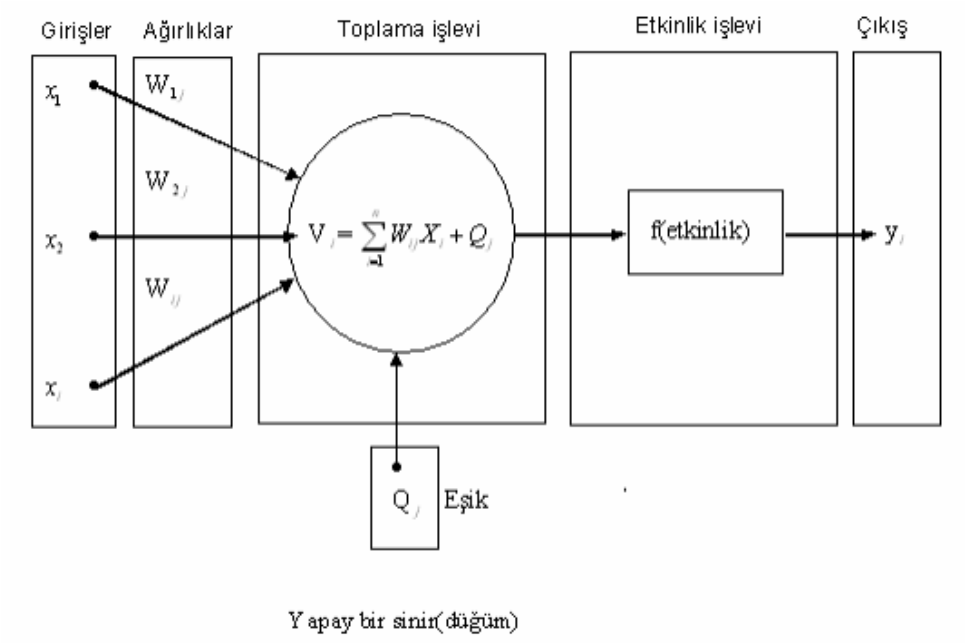
5- Bellek Temelli Nedenleme (Memory Based Reasoning)

6- Naive-Bayes

Karar ağacı, karar düğümlerinden, dallardan ve yapraklardan oluşur [36]. Karar ağaçlarının ilk hücreleri kök olarak adlandırılır. Yapılan her bir gözlem kökte belirtilen koşulun gerçekleşmesine bağlı olarak “Evet” ya da “Hayır”dan birisi seçilerek sınıflandırılır. Kök hücrelerinin altında düğümler bulunur ve her bir gözlem düğümler ile sınıflandırılır. Düğüm sayısı arttıkça modelin karmaşıklığı da artmaktadır. Karar ağacının en altında bulunan yapraklar bize sonucu vermektedir.

Yapay sinir ağları insan beyninden esinlenerek geliştirilen, ağırlıklı bağlantılar vasıtasıyla birbirine bağlanan ve her biri kendi belleğine sahip işlem elemanlarından oluşan paralel ve dağıtılmış bilgi işleme yapılarıdır. Yapay sinir ağlarına bu bilgiler ilgili olaya ait örnekler üzerinde eğitilerek verilir [38]. Şekil 2.1’de bir yapay sinir ağına ait yapı gösterilmektedir.

Yapay sinir ağlarının başlıca uygulama alanlarının sınıflandırma, tahmin ve modelleme olduğu söylenebilir [38].



Şekil 2.1. Yapay Sinir Ağının Yapısı [39].

Genetik algoritmalar, doğal seleksiyon ve canlıların genetiğinin gelişiminin benzetişimli modelini uygulamaktadır. Algoritma, diğer evrimsel algoritmalarda olduğu gibi araştırmanın yapıldığı uzayda bulunan çözümlerden oluşan başlangıç kümesi kullanmaktadır. Başlangıç popülasyonu her jenerasyonda doğal seçme ve tekrar üreme işlemleri aracılığıyla art arda geliştirilir. En son kuşağın en uygun yani en kaliteli bireyi, problem için optimal çözüm olmaktadır [39].

Genetik algoritmaların uygulamaları üç ana gruba ayrılabilir. Bunlardan birincisi deneysel uygulamalardır; mevcuttaki diğer optimizasyon algoritmalarına karşı genetik algoritmalarının daha üstün olduğunu kanıtlamak amacıyla belirli problemlerin çözümlerinin bulunması için genetik algoritmaların kullanıldığı çalışmalardır. İkinci grupsa pratik uygulamalardır; genetik algoritmalar endüstriye ait problemlerinin ve diğer gerçek problemlerin çözümlerinin bulunması için kullanılır. Üçüncü grupsa veri madenciliğinde bilgi çıkarılma amacıyla kullanılan çalışmalardır [39].

K-En Yakın Komşu algoritması, sınıflandırma sürecinde çıkarılmış olan özelliklerin sınıflandırılması istenilen yeni ögenin daha öncekilerden k tanesine olan uzaklığına bakılarak gerçekleştirilir.

Mesafe hesabı yapılırken çoğunlukla öklid mesafesi (euclid distance) kullanılabilmektedir. Yeni bir örnek geldiği zaman mevcuttaki öğrenme verisinde sınıflandırma yapmak, eğitimli öğrenme algoritması olan KNN'in amacıdır. Algoritmada, yeni bir örnek geldiği zaman örneğin en yakın K komşusuna bakılarak örneğin sınıfına karar verilir.

Regresyon analizi, bir bağımlı değişkenle bir ya da birden fazla bağımsız değişkenin arasında bulunan ilişkinin incelenmesi için kullanılmaktadır. Bir tek bağımsız değişkenin kullanıldığı regresyon tek değişkenli regresyon analizi, birden fazla bağımsız değişkenin kullanıldığı regresyon analizi de çok değişkenli regresyon analizi olarak adlandırılır. Tek değişkenli regresyon analizi bir bağımlı değişken ve bir bağımsız değişken arasındaki ilişkiyi incelerken, bağımlı ve bağımsız değişkenler arasında bulunan doğrusal ilişkiyi temsil eden bir doğrunun denklemi oluşturur. Çok Değişkenli Regresyon Analizi, içinde bir adet bağımlı değişken ve birden fazla bağımsız değişkenin bulunduğu regresyon modelleridir.

Naive Bayes algoritması, her bir kriterin sonuca olan etkilerinin olasılık olarak hesaplanmasına dayalı bir VM modelidir. Naive Bayes algoritmasında bir sonucun çıkma olasılığı o sonucu etkileyen tüm faktörlerin o sonucu sağlama olasılıklarının çarpımıdır [40].

2.1.2. Kümeleme

Kümeleme, verinin sınıflara veya kümelere ayrılması işlemidir [41]. Aynı kümeye ait elemanlar birbirlerine benzerken başka kümelere ait elemanlardan farklıdırlar. Kümeleme yöntemi, veri madenciliği, makine öğrenmesi, istatistik, biyoloji gibi birçok alanda kullanılmaktadır. Kümeleme modelinde, sınıflama modelinde bulunan veri sınıfları bulunmamaktadır [42]. Kümeleme modelinde, sınıfları olmayan veriler gruplar halinde kümelere ayrılırlarken, Sınıflama modelinde, verilere ait sınıflar bilinebilmekte ve yeni bir veri geldiğinde bunun hangi sınıfa ait olduğu tahmin edilebilmektedir. Bazı uygulamalarda kümeleme modeli, sınıflama modelinin bir önişlemi durumunda değerlendirilebilmektedir [42]. Marketlerdeki farklı müşteri gruplarının keşfi ve bu grupların alışverişleriyle ilgili örüntülerin ortaya çıkarılması, biyolojide bitkilerin ve hayvanların sınıflandırılmaları ve işlevlerine göre benzer olan genlerin sınıflandırılması, şehir planlanması yapılırken evlerin tiplerine, değerlerine

ve coğrafik konumlarına göre gruplara ayrılması uygulamaları kümeleme uygulamalarıdır. Kümeleme uygulaması aynı zamanda Web üzerindeki bilgilerin keşfi için dokümanların sınıflandırılması amacıyla da kullanılabilmektedir [43].

Kümeleme uygulaması son zamanlarda veri madenciliği uygulamaları arasında oldukça güncel ve yaygın bir konu haline gelmiştir. Literatürde pek çok kümeleme algoritması bulunmaktadır. Kullanılacak olan kümeleme algoritmasının seçimi, verilerin tipine ve amaca bağlıdır. Genel olarak başlıca kümeleme yöntemleri şu şekilde sınıflandınlabilir [36]:

1- Bölme yöntemleri

2- Hiyerarşik yöntemler

3- Yoğunluk tabanlı yöntemler

4- Izgara tabanlı yöntemler

5- Model tabanlı yöntemler

Kümeleme, basitçe benzer veri öğelerinin toplanması ve sunulması anlamına gelir. Bir küme, grup içinde birbirine benzeyen ve diğer kümelere ait nesnelere benzemeyen bir öğe grubudur. İstatistiksel gösterimde "kümeleme en önemli denetimsiz öğrenme algoritmasıdır" [44]. Veri madenciliği sürecinde bir ön işleme algoritması olan kümeleme, veri boyutunu daha fazla veri analizi için kullanılabilecek anlamlı kümelere önemli ölçüde azaltabilir. Bununla birlikte, veri boyutunu küçültürken dikkatli olunmalıdır, çünkü verileri daha az küme şeklinde temsil ederken tipik olarak kayıplı veri sıkıştırmasına benzer bazı ince ayrıntılar kaybolur. Kümeleme algoritmalarının sınıflandırılması, birçoğu birbiriyle örtüştüğü için kesin değildir.

Geleneksel terimlerle, kümeleme teknikleri genel olarak hiyerarşik ve bölümsel olmak üzere iki türe ayrılmıştır. Ancak bu türleri tartışmadan önce, kümeleme (denetimsiz sınıflandırma) ile denetimli sınıflandırma (veya diskriminant analizi) arasındaki ince farkı anlamak önemlidir.

Denetimli sınıflandırmada, bize etiketlenmiş (veya önceden sınıflandırılmış) veri kalıplarının bir koleksiyonu verilir. Amaç, yeni karşılaşılan etiketlenmemiş bir veri kümesi için etiketlemeyi belirlemektir. Oysa kümeleme durumunda sorun,

etiketlenmemiş veri kümesini anlamlı kategorik etiketli desenler veya kümeler halinde gruplamaktır [45].

Kümeleme yöntemlerini sınıflandırırken bir yandan kümeleme yönteminin doğası dikkate alınmalıdır. Bu nedenle, kümeleme çözümünü oluşturan kümelerin yapısı (tek katmanlı veya birkaç küme katmanı) ile ilgili olarak, Bölmeli ve Hiyerarşik yöntemler genellikle ayırt edilir. Ayrıca, veri kümesindeki nesnelerin kümelere nasıl eşlendiğine atıfta bulunulan Sert ve Yumuşak yöntemler arasındaki ayrım (ikili eşleme ve aidiyet derecesi) de çok önemlidir. Kümeleme yöntemleri tipik olarak algoritmayı uygulamak için benimsenen yaklaşıma göre sınıflandırılır.

Bu algoritmalar hakkında daha fazla ayrıntı [46] nolu çalışmada bulunabilir. Kümeleme algoritmaları, büyük veri gibi hacimli veri boyutlarına da uygulanır. Büyük veri kavramı, Sagiroğlu ve Sinanç'a göre, öğrencilerin test veya sınavlarının kayıtları ve muhasebe kayıtları gibi çeşitli depolarda saklanabilen hem dijital hem de fiziksel formatlarda hacimli, çok büyük miktarda veriyi ifade eder [46]. Sayısal boyutu yazılımın işlem sınırını aşan bir veri seti [47]'de önerildiği gibi büyük veri olarak kategorize edilebilir.

Geçmişte, [48]'de yazarlar büyük verilerin katıksız hacimli gücünü evcilleştirmek için kümeleme, tahmin ve ilişkilendirme gibi geleneksel veri madenciliği algoritmalarının uygulanmasına ilişkin ayrıntılı bilgiler sağlayan birkaç çalışma yapmıştır.

Çeşitli bilimsel ve endüstriyel sektörler, veri hacimlerinde üstel bir büyüme yaşamaya devam ediyor ve bu nedenle otomatik kategorizasyon teknikleri, veri kümesi keşfi için standart araçlar haline gelmiştir. Otomatik sınıflandırma teknikleri - tipik olarak kümeleme olarak adlandırılır - bir veri kümesinin yapısını ortaya çıkarmaya yardımcı olur. Örneğin, oluşturulan kümelerin her biri, makul ölçüde benzer ihtiyaçlara ve davranışlara sahip bir müşteri grubuna karşılık gelebilir. Ortaya çıkan kümeler genellikle daha yüksek seviyeli - genellikle özel - tahmine dayalı modeller için yapı taşları olarak kullanıldığından, araştırmacılar sürekli olarak ince ayar yapmış ve yeni kümeleme teknikleri icat etmişlerdir [49].

2.1.3. Birliktelik kuralları

Birliktelik kuralları, büyük veri kümeleri arasındaki birliktelik ilişkilerini bulmaktadırlar [50]. Depolanan ve toplanan verilerin git gide büyümesi sebebiyle

şirketler veritabanlarında bulunan birliktelik kurallarını ortaya çıkarmayı istemektedirler. Şirketlerin karar alma sürecinin daha verimli hale getirilebilmesi için büyük miktardaki mesleki kayıtlarındaki ilginç birliktelik ilişkileri keşfedilmektedir. Birliktelik kurallarının kullanıldığı en tipik örnek market sepeti uygulamasıdır. Bu uygulama, müşterilerin yaptığı alışverişlerindeki ürünlerin arasında bulunan birliktelikleri ortaya çıkararak müşterilerin satın alma alışkanlıklarını analiz eder. Bu tip birlikteliklerin keşfi, müşterilerin hangi ürünleri bir arada aldıkları bilgisini ortaya çıkarır ve market yöneticileri de bu bilgileri kullanarak daha etkili satış stratejileri geliştirebilirler.

Örneğin bir müşterinin süt alışverişini yaparken aynı alışverişte sütün yanında ekmek alma olasılığı hesaplanır ve bu bilgi kullanılmasıyla market rafları düzenlenerek ürünlerdeki satış oranı arttırılabilir. Müşterilerin sütle birlikte ekmek satın alma oranı yüksek ise süt ile ekmek rafları yan yana koyularak ekmek satışları arttırabilir.

2.2. Metin Madenciliği

Metin madenciliği, özel amaçlar için metinden bazı bilgiler çıkarılması için metinin analiz edilmesi işlemidir [51]. Veri madenciliği, büyük veri içindeki gizli örüntülerin ve şablonların çıkarılması, gizli bilginin keşfedilmesidir. Veri madenciliği ile metin madenciliği arasındaki temel fark; metin madenciliği çalışmalarında kullanılan verinin yapısal veri biçimi yerine doğal dil metinlerinden oluşmasıdır [52].

Metin madenciliğinde amaç bilinmeyen bilgileri, henüz kimsenin bilmediği ve henüz yazamadığı bir şeyi keşfetmektir. Metin madenciliği, büyük veri tabanlarından ilginç desenler bulmaya çalışan veri madenciliğinin bir varyasyonudur [53]. Akıllı Metin Analizi, Metin Veri Madenciliği veya Metinde Bilgi Keşfi (KDT) olarak da bilinen metin madenciliği, genellikle yapılandırılmamış metinden ilginç ve önemli bilgiler çıkarma işlemini ifade eder. Çoğu bilgi metin olarak depolandığından, metin madenciliğinin yüksek bir ticari potansiyelinin olduğu düşünülmektedir.

Bilgi birçok bilgi kaynağından keşfedilebilir, ancak yapılandırılmamış metinler kolayca erişilebilen en büyük bilgi kaynağı olmaya devam etmektedir. Metinden Bilgi Keşfi Sorunu (KDT) Doğal Dil İşleme (NLP) tekniklerini kullanarak kavramlar arasındaki açık ve örtük kavramları ve anlamsal ilişkileri çıkarmaktır [54]. Amacı, büyük miktarda metin verisi hakkında fikir edinmektir. KDT, NLP'ye derinden kök

salmış olsa da, keşif süreci için istatistik, makine öğrenimi, akıl yürütme, bilgi çıkarma, bilgi yönetimi gibi yöntemlere dayanır. KDT, Metin Anlama gibi yeni ortaya çıkan uygulamalarda giderek daha önemli bir rol oynamaktadır.

Metin madenciliği veri madenciliğine benzer, fakat veri madenciliği araçları veri tabanlarından yapılandırılmış verilerin işlenmesi için tasarlanmıştır, ancak metin madenciliği e-postalar, tam metin belgeleri ve HTML dosyaları gibi yapılandırılmamış veya yarı yapılandırılmış veri kümeleriyle çalışabilir. Sonuç olarak, metin madenciliği şirketler için çok daha iyi bir çözümdür. Bununla birlikte, bugüne kadar araştırma ve geliştirme çabalarının çoğu yapılandırılmış verileri kullanarak veri madenciliği çabalarına odaklanmıştır.

Metin madenciliğinin getirdiği sorun açıktır: insanların birbirleriyle iletişim kurması ve bilgi kaydetmesi için doğal dil geliştirilmiştir ve bilgisayarlar doğal dili kavramaktan çok uzaktır. İnsanlar metne dilsel kalıpları ayırt etme ve uygulama yeteneğine sahiptir ve insanlar argo, yazım varyasyonları ve bağlamsal anlam gibi bilgisayarların kolayca üstesinden gelemediği engelleri kolayca aşabilirler. Bununla birlikte, dil yeteneklerimiz yapılandırılmamış verileri anlamamıza izin verse de, bilgisayarın metni büyük hacimlerde veya yüksek hızlarda işleme yeteneğinden yoksunuz.

İnsan ve bilgisayar dillerindeki farklılıklar geniş olsa da, bu açığı kapatmaya başlayan teknolojik gelişmeler olmuştur. Doğal dil işleme alanı, bilgisayarlara doğal dili öğreten teknolojiler üretmiştir, böylece metinleri analiz edebilir, anlayabilir ve hatta üretebilirler. Geliştirilen ve metin madenciliği sürecinde kullanılabilecek teknolojilerden bazıları; bilgi çıkarma, konu takibi, özetleme, sınıflandırma, kümeleme, kavram bağlantısı, bilgi görselleştirme ve soru cevaplamadır [55].

Metin tabanlı bilgilerin işlenmesini sağlayan hesaplama sistemlerinin temel stratejilerinden biri sonsuz doğal dil girdilerinin küçük bir kategoriler kümesine indirgenmesidir. Metinlerin kategorilere ayrılması iki açıdan önemli olmuştur. İnternet gibi hızla büyüyen ve gelişen metin tabanlı bilgi kaynaklarının sayesinde bilginin işlenme ihtiyacı da artmıştır. Metin kategorilemesinin bu alandaki kullanımı dokümanların filtrelenmesi, konuya özel işleme mekanizmalarının sağlanmasıyla bilginin edinilmesi şeklindedir.

Bu zamana kadar metin madenciliği analizlerinde genellikle İngilizce ve İngilizce'ye kural ve teknik olarak benzeyen diller üzerinde çalışılmıştır. İngilizce gibi dillerde ek sayısı fazla değildir ve bu analizi kolaylaştırır. Ancak Türkçe gibi sondan eklemeli ve çok fazla ek alabilen, eklerle birlikte kökündeki bazı harfler değişebilen bir dilde analiz yapmak daha zordur, literatürde de çok az örneğe rastlanmaktadır. Türkçe için metin analizleri, metin önişleme tekniklerinin diğer dillerden farklı olması gerekmektedir.

Büyük harflerin küçük harfe dönüştürülmesi, noktalama işaretlerinin kaldırılması, ek ve köklerin ayrılması, analizde anlam ifade etmeyen genel sözcüklerin veri kaynağından çıkarılması gerekmektedir.

Metin madenciliği, metin sınıflandırma, konu çıkarımı, kümeleme, duygusal analiz, metni özetleme vb. amaç ve yöntemlerle yapılmaktadır.

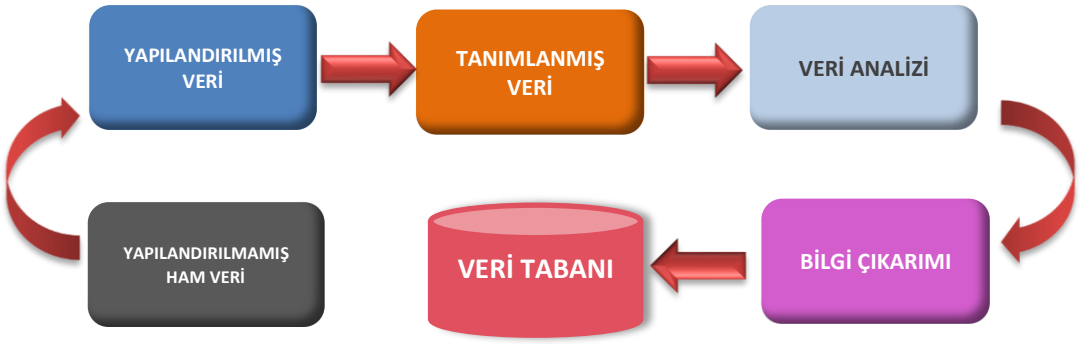
Bu bahsi geçen hedeflere ulaşmak için enformasyon getirimi, hece analizi, örüntü tanıma, kelime frekans dağılımı, enformasyon çıkarımı, etiketleme, veri madenciliği ve görselleştirme gibi yöntemler kullanılmaktadır [56].

2.2.1. Metin madenciliği süreci

Dünya verilerinin yaklaşık %80'i yapılandırılmamış metinde bulunmaktadır [57]. Bu yapılandırılmamış metin, bilgisayar tarafından daha fazla işlem için kolayca kullanılamaz. Bu nedenle, yapılandırılmamış metinden bazı değerli bilgileri çıkarmak için yararlı olan bazı tekniklere ihtiyaç vardır. Bu bilgiler daha sonra yapılandırılmış ve yapılandırılmamış birkaç alan içeren metin veri tabanı biçiminde depolanır. Metinler postalara, sohbetlere, sms'lere, gazete makalelerine, dergilere, ürün incelemelerine ve kuruluş kayıtlarına yerleştirilebilir [58]. Kurumların, kamu sektörlerinin, kuruluşların ve endüstrilerin hemen hemen her birindeki bilgiler elektronik ortamda saklanır. Metin madenciliği için veri madenciliği, metinsel veri tabanlarından bilgi keşfi, akıllı metnin analizi, yapılandırılmamış metinden değerli bilgilerin çıkarılması veya alınması anlamına gelir [59]. Yapılandırılmış veri tabanlarından veri madenciliği veya bilgi keşfinin bir uzantısı olarak görülebilir. Metin madenciliği, daha önce tanımlanamayan veya gizli bilgiler olan metinsel verilerden yeni bilgi parçalarını farklı teknikler kullanarak ayıklayarak keşfeder. Metin madenciliği, bilgi edinme, metin analizi, bilgi çıkarma, sınıflandırma, kümeleme, görselleştirme, veri madenciliği ve makine öğrenimi ile ilgili çok disiplinli bir alandır.

Metin madenciliği sürecinde verinin ham veriden veri tabanına dönüşümü Şekil 2.2'de gösterilmektedir. Aşağıdaki gibi beş temel metin madenciliği adımı vardır:

1) Yapılandırılmamış verilerden bilgi toplama
2) Alınan bu bilgileri yapılandırılmış verilere dönüştürme
3) Yapılandırılmış verilerden kalıbı tanımlama
4) Kalıbı analiz etme
5) Değerli bilgileri ayıklama ve veri tabanında saklama



Şekil 2.2. Metin madenciliği süreci

2.2.2. Temel metin madenciliği teknolojileri

Enformasyon Getirimi (information retrieval): Bu aşama ilgilenilen derlem (corpus) ile ilgili ön bilginin toplandığı aşamadır. Örnek olarak, metin madenciliği web üzerinde bulunan veri kaynaklarının üzerinde yapılacak ise; web sayfaları, adresleri veya dosya sistemi üzerindeyse dosyaların tarihleri, kullanıcı bilgileri, dosya isimleri, dizin bilgileri gibi bilgilerin toplandığı aşamadır. En iyi bilinen bilgi alma (IR) sistemleri, World Wide Web'de belirli bir sözcük kümesiyle ilişkili belgeleri tanıyan Google arama motorlarıdır. Kullanıcı için önemli olan yararlı bilgileri çıkarmak için döndürülen belgelerin işlendiği belge alma işleminin bir uzantısı olarak ölçülür [60]. Böylece belge alma işlemini, kullanıcı tarafından oluşturulan sorguya odaklanan bir metin özetleme aşaması veya bir bilgi çıkarma aşaması izler.

Doğal Dil İşleme: Günlük hayatta sıklıkla sözlü ve yazılı olarak kullanılan Türkçe, İngilizce, Arapça gibi iletişim dillerinin bilgisayar sistemlerinin algılanması ve işlenmesi yoluyla elektronik ortama aktarılmasıdır. Doğal dil işleme, yapay zekanın bir dalı olarak temsil edilir ve onu kullanan birçok sistemde otomatik ve temassız işlemlere izin verir.

Adlandırılmış Varlık Tanıma: Metin belgelerinden önceden tanımlanmış kategorilerin (kişiler, konumlar, kuruluşlar gibi) çıkarılması işlemidir. Makine çevirisinden duygu analizine kadar birçok doğal dil işleme probleminde kullanılan bir bilgi çıkarma dalıdır.

Örüntüsü Tanımlı Varlıkların Bulunması: Bazı durumlarda, metindeki belirli belirli bilgiler metin madenciliğinin konusu olabilir. Örneğin, e-posta adresleri, telefon numaraları, adresler, tarihler gibi belirli bilgileri özel olarak almak isteyebiliriz.

Genellikle, bu durumlarda, düzenli ifadeler veya bağlamdan bağımsız gramerler tanımlanır ve metin üzerinde çalıştırılır [61].

Eş Atıf: Varlıklara (alıntı) gönderme yapan isim tamlamalarını ve diğer terimleri bulmayı/ayırmayı amaçlar.

Duygu Analizi: Bir metinde duyguların nasıl ifade edildiğini ve bu ifadelerdeki olumlu veya olumsuz durumları belirlemeye olanak sağlayan bir analizdir. Duygu analizi veya fikir araştırması, insanların fikirleri, değerlendirmeleri, tutumları ve duygularının hesaplamalı bir çalışmasıdır. Duygu analizi, özellikle çevrimiçi veriler için son yıllarda popüler hale gelmiştir. Duygu analizi programı, kullanılan kelimelerin ve ifadelerin özelliklerine göre metnin duygusal içeriğini tahmin etmeyi amaçlar.

Buna göre bir konu hakkında geçen mesajların veya yazıların olumlu veya olumsuz olmasına göre iki sınıfa ayrılması hedeflenir [62].

Bilgi çıkarma (IE) yöntemlerinin amacı, metinden yararlı bilgilerin çıkarılmasıdır. Varlıkların, olayların ve ilişkilerin yarı yapılandırılmış veya yapılandırılmamış metinden çıkarılmasını tanımlar. Kişinin adı, yeri ve kuruluşu gibi en yararlı bilgiler, metni doğru anlamadan çıkarılabilir [59]. IE, metinlerden alınan seçilmiş ilgili parça bilgilerinin yapılandırılmış bir görüntüsünün oluşturulması olarak tanımlanabilir.

Metin sınıflandırma: Her bir belge için kategorilerin önceden bilindiği ve kesin olarak devam ettiği bir tür “denetimli” öğrenmedir. Sınıflandırma, normal dil belgelerinin içeriğine göre önceden tanımlanmış konular kümesine atanmasıdır. Metin belgelerinin bir koleksiyonudur, her belge için doğru konuyu veya konuları bulma sürecidir.

Kümeleme: Metin madenciliğinde en ilginç ve önemli konulardan biridir. Amacı, bilgideki içsel yapıları bulmak ve bunları daha fazla çalışma ve analiz için önemli alt gruplara ayırmaktır. Nesnelerin kümeler adı verilen gruplara ayrıldığı denetimsiz bir işlemdir. Kümeleme, biyoloji, veri madenciliği, örüntü tanıma, belge alma, görüntü segmentasyonu, örüntü sınıflandırma, güvenlik, iş zekası ve Web araması gibi birçok

uygulama alanında kullanışlıdır. Küme analizi, veri dağıtımını sağlamak için bağımsız bir metin madenciliği aracı olarak veya algılanan kümeler üzerinde çalışan diğer metin madenciliği algoritmaları için bir ön işleme adımı olarak kullanılabilir.

Metin özetleme: Metin madenciliğinde zor bir işlemdir, araştırmacının hesaplama zekası, makine bilgisi ve doğal dil işleme alanlarındaki dikkatine ihtiyaç duymaktadır. Metin özetlemesi, kullanıcıya yararlı bilgiler sağlayan belirli bir metnin sıkıştırılmış bir sürümünü otomatik olarak oluşturma işlemidir. Büyük bir organizasyonda veya şirkette, araştırmacının tüm belgeleri okumak için zamanı yoktur, bu nedenle belgeyi özetler ve özeti ana noktalarla vurgularlar [59]. Özet, bilginin önemli bir bölümünü içeren, uzunluğu azaltılmış ve genel anlamı orijinal metinlerde olduğu gibi tutan bir veya daha fazla metinden üretilen bir metindir. Metin özetlemesi, sinir ağları, karar ağaçları, anlamsal grafikler, regresyon modelleri, bulanık mantık ve sürü zekası gibi metin sınıflandırmasını kullanan çeşitli yöntemleri içerir. Bununla birlikte, tüm bu yöntemlerin ortak bir sorunu vardır, sınıflandırıcıların gelişiminin kalitesi değişkendir ve özetlenen metnin türüne büyük ölçüde bağlıdır [63].

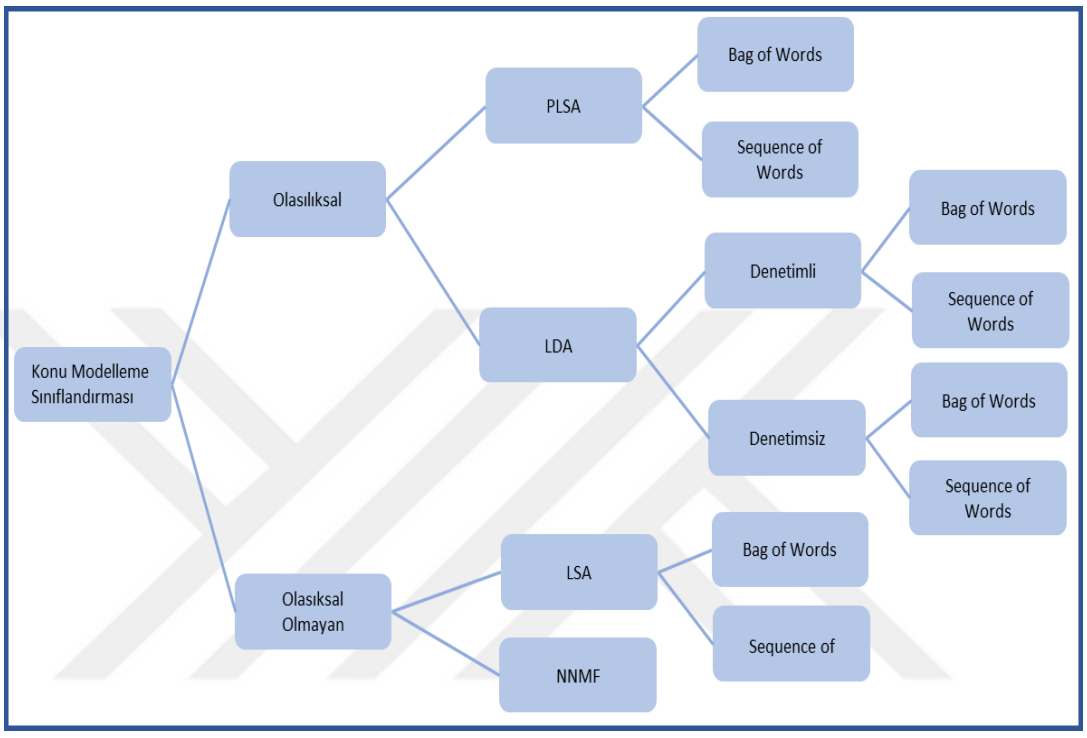
Çok sayıda metinsel veri üreten sosyal medya platformlarının sayısı arttıkça, metinsel verileri analiz etmeye yönelik geleneksel yaklaşım ortadan kalkmış ve yerini metin özetleme, anahtar kelime çıkarma, duygu analizi, konu modelleme vb. dahil olmak üzere metin madenciliği teknikleri almıştır. Konu modelleme, bir dizi belgeyi tarayabilen, içlerindeki sözcük ve kelime öbeği kalıplarını algılayabilen ve bir belge kümesini en iyi karakterize eden sözcük gruplarını ve benzer ifadeleri otomatik olarak kümelendirebilen, denetimsiz bir makine öğrenimi tekniğidir. Konu modelleme, veri madenciliği, gizli veri keşfi ve veri ile metin belgeleri arasındaki ilişkileri bulmak için metin madenciliğinde en güçlü tekniklerden biridir [64].

Güçlü bir yöntem olan konu modelleme büyük boyutlu dokümanlardan denetimsiz bir şekilde gizli yapıyı keşfetmek amacıyla kullanılmaktadır. Konu modelleme, yapılandırılmamış (unstructured) metinlerden oluşan büyük koleksiyonlarını anlamamıza, bilgileri düzenlememize ve sunmamıza yardımcı olur.

Konu modelleme, bir metin belgesinde “topics(konular)” adı verilen kelime gruplarını bulmak için kullanılan denetimsiz(unsupervised) bir yaklaşımdır. Bu konular, sık sık birlikte ortaya çıkan ve çoğunlukla ortak bir temayı paylaşan kelimelerden oluşmaktadır. Bundan dolayı bu konular, önceden tanımlanmış kelime kümesiyle

belgenin tamamını en iyi şekilde tanımlamak için kelime grubu olarak kullanılabilmektedir.

Konu modelleme, bize büyük metin verisi koleksiyonlarını organize etme, anlama ve özetleme yöntemleri sunar. Konu modelleme sınıflandırma hiyerarşisi Şekil 2.3'te gösterilmiştir.



Şekil 2.3. Konu modelleme sınıflandırma hiyerarşisi

Aşağıda üç adımda özetlenen metin madenciliği süreci adımları ile değerli bilginin elde edilmesi sağlanmaktadır [65]:

1. Farklı kaynaklarda bulunan değişik formatlardaki (düz metin, web sayfası, pdf dosyası vb.) verinin toplanması,

2. Veri ön işleme ve temizleme yapılması, aykırı durumların tespit edilmesi ve kaldırılması,

3. Metin madenciliği tekniklerinin kullanılarak örüntü analizinin yapılması.

Metin madenciliği teknolojisinin kullanımı ile bilgi alma, bilgi çıkarma, metin sınıflandırma, kümeleme ve metin özetleme alanlarında araştırmalar yapılmıştır [66, 67]. Metin sınıflandırmada; etiketli örneklerle eğitilmiş bir modele dayalı olarak,

makale veya kitaplardan alınan metin, paragraf veya belgeleri gruplamak ve sınıflandırmak için veri madenciliği sınıflandırma yöntemleri kullanılabilir [68].

Metin sınıflandırmasında da iki temel yöntem vardır [69]:

Birinci yöntem, uzman bir mühendisten sınıf hakkında bilgi alınması veya sınıflandırma kurallarının doğrudan sisteme kodlanması ile elde edilen bilgi mühendisliği yöntemidir.

İkinci yöntem, önceden sınıflandırılmış örneklerden öğrenerek bir sınıflandırıcı oluşturan bir makine öğrenmesi yöntemidir.

Veri setinde, makale adı ve makale özet niteliklerinin önceden işlenmesi gerekir. Literatürde metin madenciliğinde kullanılan aşağıdaki veri ön işleme yöntemlerinin kullanıldığından bahsedilmiştir [70]:

Çıkarım: Bir dosyanın içeriğini tek bir öğeye dönüştürmek için kullanılır.

Etkisiz kelimelerin ayıklanması (stop words): Giriş alanının boyutunu küçültmek için analizde geçersiz sayılan sözcükleri metinden çıkarın.

Veri madenciliği yöntemleri kelime odaklı tarama ve karar verme süreçleri gerçekleştirdiği için “amaç”, “kapsam”, “sonuç”, “bu araştırma”, “bu bağlamda”, “örnek”, “yöntem” vb. sözcük ve sözcük öbeklerinin ve "ve", "veya", "ile", "ama", "veya" gibi edat ve bağlaçların da çıkarılması gerekir. Sayısal ifadeler ve noktalama işaretleri de veri setinden silinmelidir.

Kök bulma (stemming): Sesbilimsel olarak ilişkili sözcüklerin köklerini bulma diğer bir deyişle ortak ekleri çıkarma ve sözcük sayısını doğru köklere uyacak şekilde azaltma işlemidir [71].

Olasılıksal Gizli Semantik Analiz (PLSA) ve Gizli Dirichlet Tahsisi (LDA) gibi geleneksel konu modelleme algoritmaları, belgelerde herhangi bir ön açıklama veya etiketleme gerektirmeden metin korpusundan gizli anlamsal yapıyı keşfetmek için yaygın olarak benimsenmiştir [72].

En belirgin konu modeli, 2003 yılında Blei ve diğerleri tarafından tanıtılan gizli Dirichlet tahsisi (LDA)’dir [6].

Gizli Dirichlet tahsisi (LDA)

Konu modelleme algoritmalarından biri olan Latent Dirichlet Allocation (LDA), her belgenin olasılıksal olarak dağıtılmış konulardan oluştuğunu ve her konunun olasılıksal olarak dağıtılmış kelimelerle temsil edilebileceğini varsayar [73].

Dolayısıyla, bir belge(metin verisi) verildiğinde LDA, belgeyi temel alarak her konu grubunu o grubu en iyi açıklayan bir dizi kelimenin olduğu konu gruplarına kümeler [74].

Latent Dirichlet Allocation (LDA) modeli, son zamanlarda büyük veri kümelerini, özellikle Web'in sosyal çalışmalarında kullanıcı tarafından oluşturulan veri kümelerini anlamak için önemli bir araç olarak ortaya çıkmıştır [75].

İlk olarak 2003 yılında Blei, Ng ve Jordan tarafından tanıtılan Gizli Dirichlet Tahsisi (LDA), konu modellemede en popüler yöntemlerden biridir. LDA, konuları kelimelerin olasılıklarına göre temsil eder. Her konuda en yüksek olasılığa sahip kelimeler, genellikle konunun ne olduğu hakkında fikir verebilir. Bir metni modellemek için denetimsiz bir olasılık yöntemi olan LDA, en yaygın olarak kullanılan konu modelleme yöntemidir. LDA, her belgenin gizli konular üzerinde olasılıklı bir dağılım olarak temsil edilebileceğini varsayar.

Bu durumda LDA her yorumu bir belge olarak görür ve bu belgelere karşılık gelen konuları bulur. Her konunun grubu, konuya yüzde katkısı ile birlikte bir dizi kelime içerir. LDA konuları ve konulara ait kelimeleri yüzdeleriyle birlikte gruplar ve kelimelerin yüzdelik ilişkilerini belirtir [76].

Konular kelime birlikteliği temelinde çıkarılır. Bilgisayara bir metin koleksiyonu yüklendikten sonra, uzman konu modelleme yazılımı, tek tek metinler içinde kelime birlikte bulunma kalıplarını arar. Bir grup kelime aynı metin içinde belirli bir düzenlilikle birlikte bulunma eğiliminde olduğunda, bilgisayar bu kelimeler arasında bir ilişki olduğunu varsayar. Örneğin, istatistiksel bir prosedür, sözcüklerin bir kez ve sonun genellikle bir bütüncede aynı metin içinde yer aldığını belirlerse, bilgisayar tarafından aralarında bir ilişki olduğu varsayılır. Genellikle üç ila beş kelimeden oluşan, nispeten dar birliktelik aralıklarıyla işleme eğiliminde olan bütünce dilsel eşdizim ölçümlerinin aksine konu modelleme genellikle daha geniş bir ölçekte, tipik olarak bir metnin tamamında birlikte meydana gelmeyi açıklar [77].

Bu nedenle, konu modellemede, bilgisayar için, aynı metin içinde sıklıkla birlikte bulunan herhangi bir kelime arasında, tutarlı bir şekilde binlerce kelime ayrı olsa bile, bir ilişki varsayması mümkündür. Bir kez tanımlandıktan sonra, bilgisayar tarafından tematik olarak tutarlı bir konuyu temsil etmek için birlikte oluşan bir kelime grubu alınır. Konu modelleme sürecinin temel çıktısı, analiz edilen tümcedeki metinlerin koleksiyonunda birlikte bulunan bir dizi kelime listesidir. Her kelime listesi prensip olarak ayrı bir konuyu temsil eder ve içerdiği kelimeler o konunun en karakteristik kelimeleridir.

Bir konu modelinden çıkan konular, bütüncedeki metinleri kapsar. Her metin çok sayıda konuyla, az sayıda konuyla veya hiçbir konuyla güçlü bir şekilde ilişkilendirilebilir. Her metin için, bir konu modeli, derlemdeki her metindeki her konunun gücünü bir puan aracılığıyla gösteren bir dizi sonuç verir. O halde, çalışmanın amaçları için, bir konu kelime listesinde görüntülenen kelimeler o konunun 'anahtar kelimeleri' olarak görülebilirken, bir konuyu sergileme olasılığı yüksek olan metinler konunun 'anahtar metinleri' olarak görülebilir [78].

Konular bilgisayar yazılımı tarafından algoritmik olarak oluşturulsa da, bu konulara anlamlar atamak ve bilgisayar tarafından oluşturulduktan sonra tematik tutarlılıklarını çıkarmak kullanıcılara/araştırmacılara kalmıştır.

Yine de bilgisayar verilerdeki konuları keşfetmeden önce, kullanıcının iki ana parametre belirlemesi gerekir: (1) ayıklanacak konu sayısı ve (2) her konuda olması gereken kelime sayısı. Bilgisayar tarafından sağlanan konu sayısı, oluşturulan konuların doğasını etkiler; çok sayıda konu daha ayrıntılı temalar sağlama vaadinde bulunurken, daha az sayıda konunun daha genel temalar sağladığı iddia edilmektedir [78].

Konu modelleme, deterministik olmayan bir yöntemdir; bilgisayar tarafından oluşturulan konular bir dereceye kadar rastgeleliğe tabidir ve veriler ve kullanıcı tarafından belirlenen parametreler sabit kalsa bile prosedür her çalıştırıldığında farklı olabilir.

Konu modellemenin, teorik olarak herhangi bir tür metin koleksiyonuna uygulanabilen esnek bir yaklaşım olduğu iddia edilmektedir. Konu modellemesi, politik basın

bültenleri [79], çevrimiçi forumlar [80] ve akademik araştırma makaleleri [78] gibi sürekli genişleyen metin türlerindeki temaları incelemek için kullanılmıştır.

Derlem destekli herhangi bir söylem çalışmasının başarısının, insan liderliğindeki hesaplamalı ve istatistiksel ölçümlerin etkileşimine dayandığı göz önüne alındığında,

teoriye duyarlı yorumlama, konu modelleme, insan denetimli yorumlama ile otomatik bir analiz yaptığı için harmanlama, prensipte bütünce destekli söylem analizinde faydalı bir rol oynayabilir [81].

Üretilen konuların tematik olarak tutarlı olduğu iddiasını kabul edersek, konu modelleme, belirli bir metin koleksiyonunun ne hakkında olduğunu göstermenin bir yolunu sağlayabilir [73].

Gizli Dirichlet Tahsisi (LDA) ve LSA aynı temel varsayımlara dayanmaktadır: dağılım hipotezi (yani benzer konular benzer kelimeleri kullanır) ve istatistiksel karışım hipotezi (yani belgeler birkaç konu hakkında konuşur) için istatistiksel bir dağılım olabilir belirlenen LDA'nın amacı, külliyatımızdaki her belgeyi, belgedeki kelimelerin büyük bir kısmını kapsayan bir dizi konuya eşlemektir. Bu, belgelerin sözcük düzenlemeleri ile yazıldığı ve bu düzenlemelerin konuları belirlediği varsayımından kaynaklanmaktadır. Yine, tıpkı LSA gibi, LDA da sözdizimsel bilgileri görmezden gelir ve belgeleri kelime paketleri olarak ele alır. Ayrıca, belgedeki tüm kelimelere bir konuya ait olma olasılığı atanabileceğini varsayar. Bununla birlikte, LDA'nın amacı, bir belgenin içerdiği konuların karışımını belirlemektir.

LSA ve LDA arasındaki temel fark, LDA'nın bir belgedeki konuların dağıtımının ve konulardaki kelimelerin dağılımının Dirichlet dağıtımları olduğunu varsaymasıdır. LSA herhangi bir dağıtım varsaymaz ve bu nedenle konuların ve belgelerin daha opak vektör temsillerine yol açar.

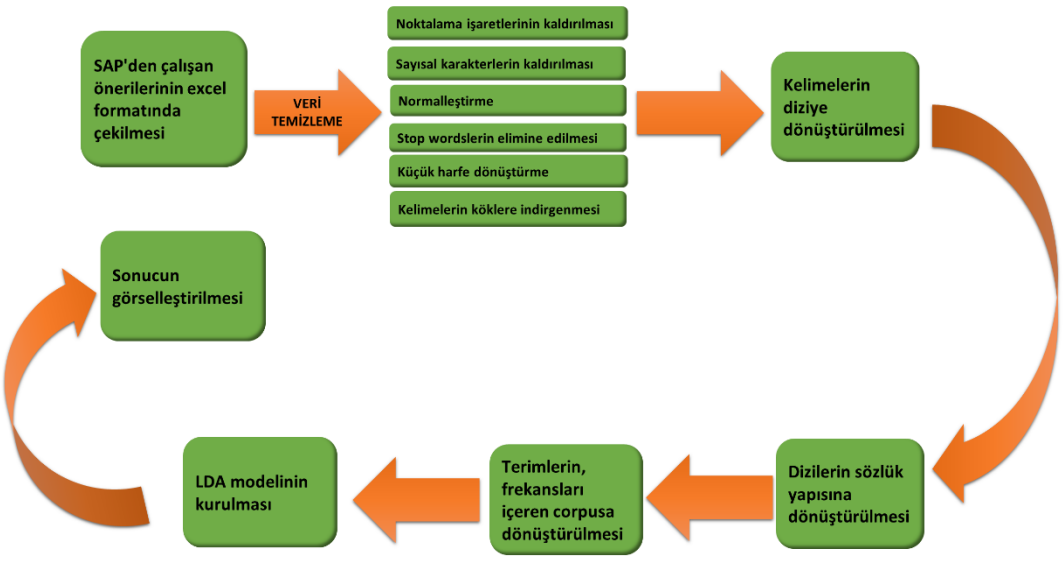
Sırasıyla alfa ve beta olarak bilinen belge ve konu benzerliğini kontrol eden iki hiperparametre vardır. Düşük bir alfa değeri, her belgeye daha az konu atarken, yüksek bir alfa değeri zıt etkiye sahip olacaktır. Düşük bir beta değeri, bir konuyu modellemek için daha az kelime kullanır, oysa yüksek bir değer daha fazla kelime kullanır ve böylece konuları aralarında daha benzer hale getirir. LDA'yı uygularken üçüncü bir hiperparametre, yani, LDA konuların sayısına kendi başına karar veremediği için algoritmanın algılayacağı konu sayısı ayarlanmalıdır.

Algoritmanın çıktısı, modellenen belge için her konunun kapsamını içeren bir vektördür. Uygun bir şekilde karşılaştırılırsa, bu vektörler, külliyatın topikal özellikleri hakkında fikir verebilir.

3. METODOLOJİ

Bu bölümde, bir otomotiv firmasına ait çalışan önerilerinin gizli Dirichlet tahsisi yöntemiyle analiz edilmesi sırasında izlenen yöntem tanıtılmaktadır.

Metin madenciliğinin temeli yapılandırılmamış verilerin yapılandırılmış verilere dönüştürülmesidir. Bunun için veri tabanından alınan veri setinin istenilen özelliklere sahip olması gerekir. İstenilen özelliklerde toplanan veri seti üzerinde tokenizasyon, durdurma sözcükleri filtreleme, kök bulma ön işleme adımları gerçekleştirilir. Daha sonra veri seti doğal dil işleme yöntemleri ile metin madenciliği algoritmalarını kullanmaya uygun hale gelmiş olmaktadır. Metin verilerindeki kelimeler token olarak tanımlanır ve her token için benzersiz bir indeks numarası ile bir sözlük oluşturulur. Bu sözlük yapısı sözcükleri ve onların dizin (id) numaralarını içerir. Sözlük yapısı bir derleme dönüştürülür böylece bağlantılı metinler düzenlenir ve yapısal olarak bir arada bulunur. Derlem metindeki kelimelerin dizin numaralarından ve sıklıklarından oluşur. Oluşturulan derlem bir belgenin terim matrisi ve LDA tematik modeli için girdi matrisidir. LDA konu modelleri bir veri ambarı kullanılarak oluşturulur. Daha sonra oluşturulan konu modelleri karşılaştırma ve görselleştirme yoluyla analiz edilir. Konu bazlı modelde aşağıdaki Şekil 3.1'de gösterilen süreç takip edilir:



Şekil 3.1. Öneri sistemlerinin LDA Konu modelleme ile analizinin adımları

• Toplanacak metinlerin belirlenmesi: Firmada çalışanların son 1 yılda verdiği öneriler SAP'den çekilir.

• Metinleri düzenleme süreci: Metindeki veriler temizlenir. Bu işlemde sınıflandırma adımında anlamı olmayan noktalama işaretleri ham verilerden silinir. Daha sonra sayısal karakterler temizlenir ardından metindeki kelimeler normalleştirilir. Normalleştirme yöntemi kullanılarak yazım hatası olan kelimeler belirlenip uygun kelimelere dönüştürülür. Öznitelikler tanımlanmadan önce anlamsız ve kesin olmayan sözcükler normalize edilerek sınıflandırma adımında öznitelikler en aza indirilir. Daha sonra stop words filtrelemesi yapılır. Böylelikle değerlendirmeye katılmaması gereken sözcükler kapsam dışı bırakılmış olur. Stopwords filtreleme işlemi için bazı veri setleri belirlenerek ortak bir havuz oluşturulur. Gereksiz kelimeler ile bağlaç, edat vb kelimelerin metinlerden çıkarılması için bu havuzdaki veriler kullanılır.

• Temizleme işlemleri tamamlandıktan sonra, programsal anlamda ortak bir kümede işlem yapılabilmesi için metinlerin tamamı küçük harfe dönüştürülür.

• Kelimenin köklere indirgenmesi: Normalize edilmiş olan metinlerde bulunan kelimeler köklerine ayrılır. Aynı köke sahip olup farklı ekleri bulunan kelimelerin ortaklaştırılmasının sağlanması bu işlemin amacıdır. Bağlaçlar, imleçler, kalıplaşmış kısaltmalar gözönünde bulundurularak, kelimelerin en yalın hale getirilmiş olur.

• Modelin oluşturulması: Kelimeler kök haline getirildikten sonra diziye dönüştürülür. Bu diziler gensim kütüphanesinin kullanılmasıyla, token olarak nitelendirilir ve her token için bir dizin numarası üretilir. İlgili diziler sözlük yapısına dönüştürülür. Sözlük haline gelen nesne, sonrasında terimlere ait frekansları içeren corpusa dönüştürülür. Kelime frekans eşleşmesinin ilk indeksinin (0,1) anlamı ilk metin verisindeki kelime indeksi 0 olan sözcüğün tekrarlanma sayısıdır. LDA modeli, frekans ağırlıkları belirlenmiş kelimeler ve belirtilen konu sayısı parametresiyle oluşturulur [74].

3.1.Veri Bilgileri

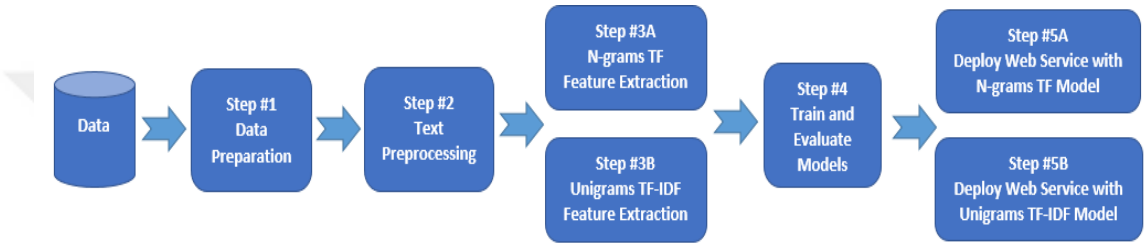
Projede kullanılan veriler SAP'den çekilen önerilerin bulunduğu 2844 satırdan oluşmaktadır. Veri kümesinde 2 kolon bulunmaktadır.

Sentiment_Label=(0=Öneriden Hızlı Kaizene, 1=Getirisi Olmayan Olumlu Öneri, 2=Değerlendirilmek Üzere Havale, 3=Devreye Alınmayacak Öneri, 4=Öneri)

Text=Metin

3.2. İş Akışı

Çalışılan bu yöntemde her kelimenin oluşma sıklığı veya terim-sıklığı (TF), ters belge frekansı ile çarpılır ve bir sınıflandırıcıyı eğitmek için öznitelik değerleri olarak TF-IDF puanları kullanılır. Ayrıca n-gram modelinin başka bir yaygın vektör temsil modeli olduğunu, ancak hangisinin en iyi sonuç verdiğine dair kesin bir cevap olmadığı bilinmektedir, bu verilerin performansına bağlıdır. Metin analizinin aşamaları Şekil 3.1'de gösterilmiştir.



Şekil 3.2. Metin analizi aşamaları

Proje aşamalarının adımları aşağıdaki gibidir:

Adım 1: veri ön işleme

Adım 2: metin ön işleme

Adım 3A: n-gram TF özellik çıkarımı

Adım 3B: unigrams TF-IDF özellik çıkarma

Adım 4: modelleri eğitme ve değerlendirme

Veri Hattı

Azure Machine Learning ile kullanıcılar yerel bir dosyadan bir veri kümesi yükleyebilir veya Azure SQL veritabanı gibi bir çevrimiçi veri kaynağına bağlanabilir.

1-4 adımlar, metin sınıflandırma modeli eğitim aşamasını temsil eder. Bu aşamada, metin örnekleri Azure ML deneyine yüklenir, metin temizlenir ve filtrelenir. Temizlenen metinden farklı türde sayısal öznitelikler çıkarılır ve modeller farklı öznitelik türleri üzerinde eğitilir. Son olarak, eğitilen modellerin performansı,

görünmeyen metin örnekleri üzerinde değerlendirilir ve bir dizi değerlendirme kriterine göre en iyi model belirlenir.

5A ve 5B adımlarında, en doğru model, RRS (Request Response Service) veya BES (Toplu Yürütme Hizmeti) kullanılarak yayınlanmış bir web hizmeti olarak dağıtılır. RRS kullanırken, aynı anda yalnızca bir metin örneği sınıflandırılır. BES kullanırken, aynı anda sınıflandırma için bir grup metin örneği gönderilebilir. Bu web hizmetlerini kullanarak, büyük ölçüde geliştirilmiş verimlilik için harici bir çalışan veya Azure Data Factory kullanarak paralel olarak sınıflandırma gerçekleştirebilir.

Adım 1: Veri hazırlama

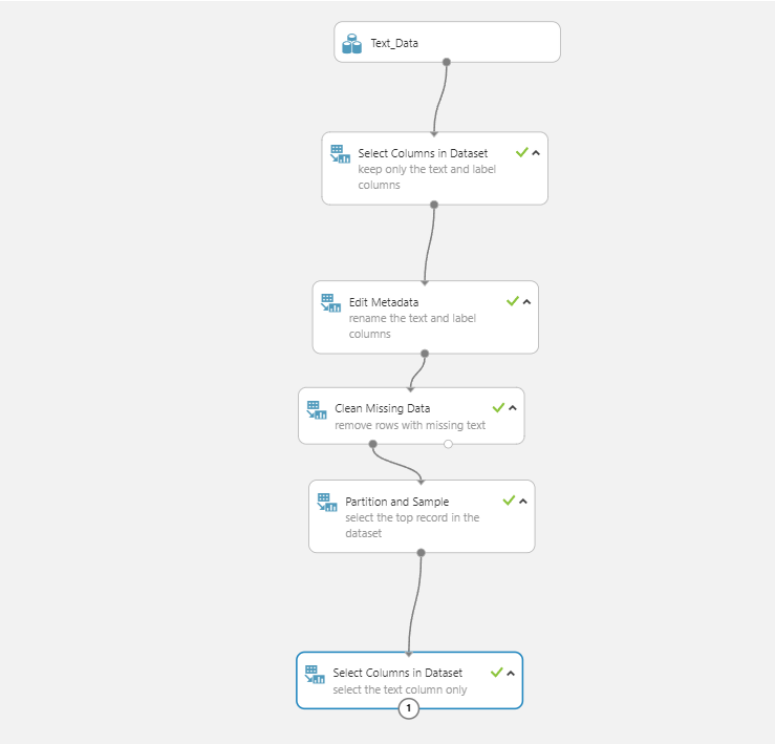
Adım 2: Metin ön işleme

Adım 3: Özellik mühendisliği

Adım 4: Modelleri eğitme ve değerlendirme

Adım 1: Veri hazırlama

Microsoft Azure programı üzerinde veri hazırlama aşamalarının akışı Şekil 3.2'de gösterilmiştir.



Şekil 3.3. Microsoft Azure'da veri hazırlama aşamaları

1.1. Aşağıdaki yöntemlerden biri kullanılarak metin sütununu içeren veri kümesi deneye yüklenir:

Seçenek 1: Yeni 'ye tıklanır, Veri Seti seçilir ve ardından Yerel Dosyadan seçilir.

Seçenek 2: Web, Azure SQL veritabanındaki bir tablo, Windows Azure tablosu, BLOB depolama alanı veya bir Hive tablosu gibi kaynaklardan veri yüklenmesi gerekiyorsa deneye bir Reader modülü eklenir.

1.2. Etiket ve metin sütunlarını seçmek için Proje Sütunları modülü kullanılır.

1.3. Etiket ve metin sütunlarını sırasıyla yeniden adlandırmak için Meta Veri Düzenleyici modülü kullanılarak *label_column ,text_column* adlandırılır.

1.4. Metin değerleri eksik olan kayıtlar, Eksik Değerleri Temizle modülü kullanılarak kaldırılır.

1.5. Veri kümesindeki en üst kaydı seçmek için Bölüm ve Örnek modülü kullanılır.

1.6. Metin sütununu seçmek için Proje Sütunları modülü kullanılır.

1.7. Deney çalıştırılır.

1.8. Deneme başarıyla tamamlandıktan sonra, aşağıdaki seçeneklerden biri kullanılarak veriler kaydedilir:

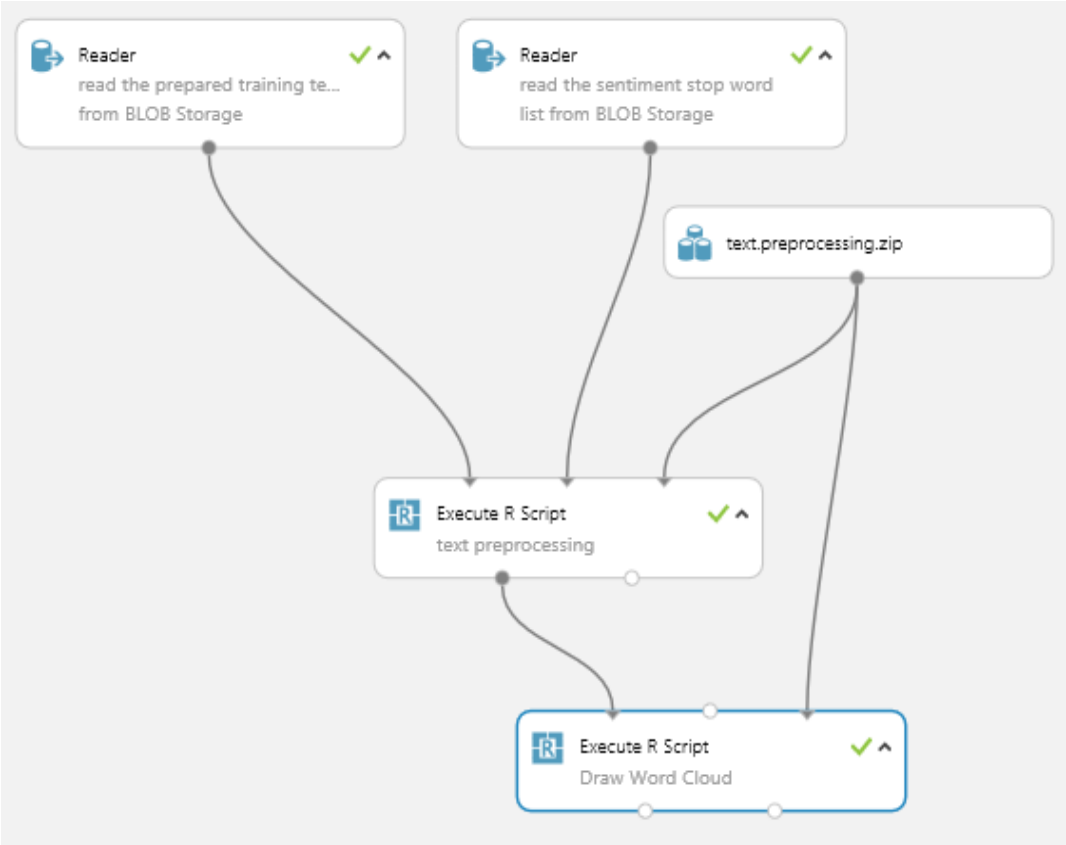
Seçenek 1: Eksik Değerleri Temizle modülünün sol çıkış bağlantı noktasına tıklanarak Veri Kümesi Olarak Kaydet öğesi seçilir ve veri kümesi adlandırılır.

Seçenek 2: Denemeye bir Writer modülü eklenir ve çıktı veri kümesi bir Azure SQL veri tabanındaki, Windows Azure tablosundaki, BLOB depolama alanındaki veya bir Hive tablosundaki bir alana yazılır.

1.9. Proje Sütunları modülünün çıkış bağlantı noktasına tıklanır, Veri Kümesi Olarak Kaydet seçilir ve veri kümesi adlandırılır.

Adım 2: Metin ön işleme

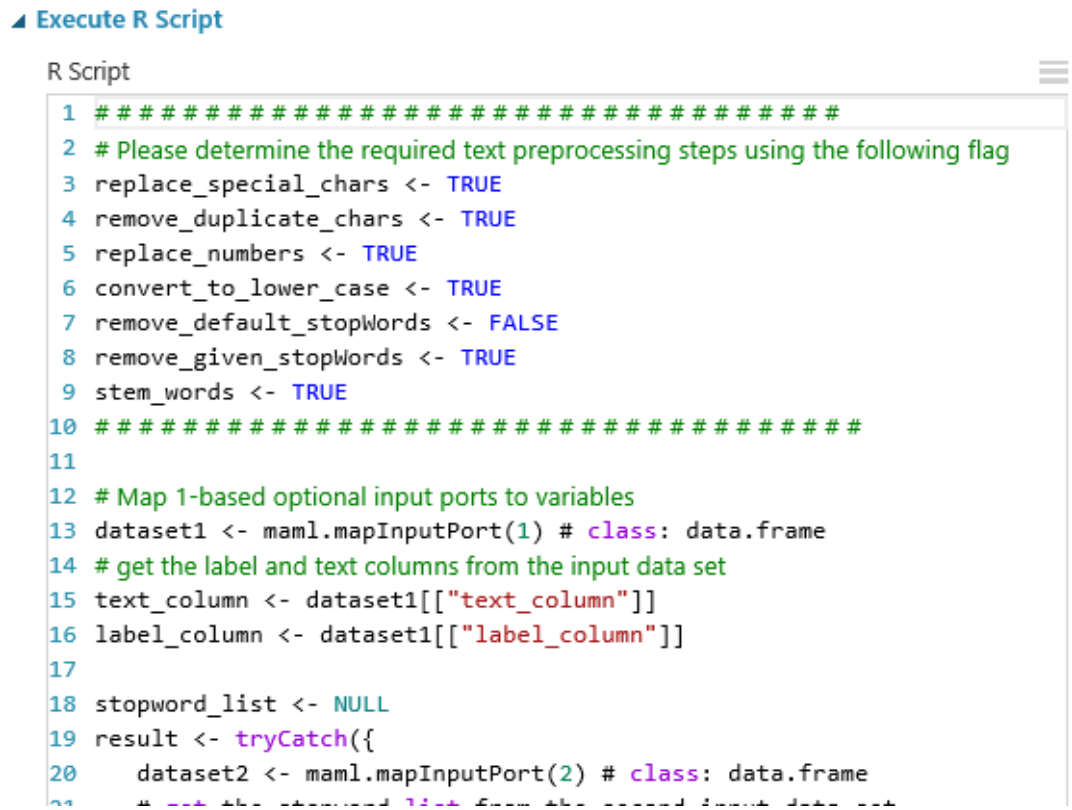
Microsoft Azure programı üzerinde veri ön işleme aşamalarının akışı Şekil 3.3'te gösterilmiştir.



Şekil 3.4. Microsoft Azure'da veri işleme aşamaları

Metinler genellikle analiz edilmeden önce bir miktar ön işleme gerektirir. Bu deney, özel karakterleri ve noktalama işaretlerini boşluklarla değiştirme, büyük/küçük harfe normalleştirme, yinelenen karakterleri kaldırma, kullanıcı tanımlı veya yerleşik durdurma sözcükleri kaldırma ve sözcük köklendirme gibi bir dizi isteğe bağlı metin ön işleme ve metin temizleme adımını içerir. Bu adımlar, R programlama dili kullanılarak uygulanır.

2.1. Verileri daha önce Web'e, bir Azure SQL veri tabanına, Windows Azure tablosuna veya BLOB Depolamaya veya bir Hive tablosuna kaydetmek için Yazıcı modülü kullanıldıysa, verileri yüklemek için Reader modülü kullanılmalıdır.



Şekil 3.5. R kodlamaları

2.2. Gerekli metin ön işleme adımlarını belirtmek için Execute R Script modülü kullanılır. R'da kullanılan kodlamalar Şekil 3.4'te gösterilmiştir.

Seçenek1: Özel karakterleri boşluklarla değiştirmek gerekiyorsa parametre *replace_special chars* TRUE olarak ayarlanır.

Seçenek2: Parametre *remove_duplicate_char*s için TRUE ayarlanır.

Seçenek3: Sayıların boşluklarla değiştirilmesi gerekiyorsa parametre *replace_numbers* TRUE olarak ayarlanır. Bazı metin sınıflandırma görevleri için, eğitim için ayırt edici özellikler olarak sayılar kullanılmalıdır.

Seçenek4: Metni küçük harfe dönüştürmek gerekiyorsa parametre *convert_to_lower_case* TRUE olarak ayarlanır.

Seçenek 5: Önceden tanımlanmış bir ortak sözcük listesi kullanarak metinden durdurma sözcüklerini kaldırmak gerekirse , parametre *remove_default_stopWords* TRUE olarak ayarlanır.

Seçenek 6: Tanımlanmış ortak sözcükler listenizi kullanarak metinden durdurma sözcüklerini kaldırmak gerekirse , parametre *remove_given_stopWords* TRUE olarak

ayarlanır. Durdurma sözcüklerinin uygulamaya bağlı olduğu unutulmamalıdır. Yani bir kelime, bir uygulama için sık görülen, ayrım yapmayan bir özellik olarak kabul edilebilirken, başka bir uygulama için anahtar özellik olarak kabul edilebilir. Örneğin, "iyi , kötü , harika" gibi kelimeler haber makaleleri kategorizasyonu için durak kelimeleri iken duyguyu ifade etmenin temel özellikleridir.

a: Ortak sözcükleri içeren veri kümesi yüklenir (bkz. Adım 1.1).

b: Yüklenen veri kümesi Execute R Script modülünün ikinci portuna bağlanır.

Seçenek 7: Kelimeleri köklendirmek gerekiyorsa parametre *stem_words* TRUE olarak ayarlanır. Köklendirme, çekimli (veya bazen türetilmiş) kelimeleri kelime kökü, taban veya kök biçimlerine indirgeme işlemidir. Örneğin, "bağlı", "bağlanmış", "bağlanan", "bağlanmayan" kelimeleri "bağla" ile eşleştirilir.

2.3. Deney çalıştırılır.

2.4. Deneme başarıyla tamamlandıktan sonra, aşağıdaki seçeneklerden biri kullanılarak sonuçlar kaydedilir:

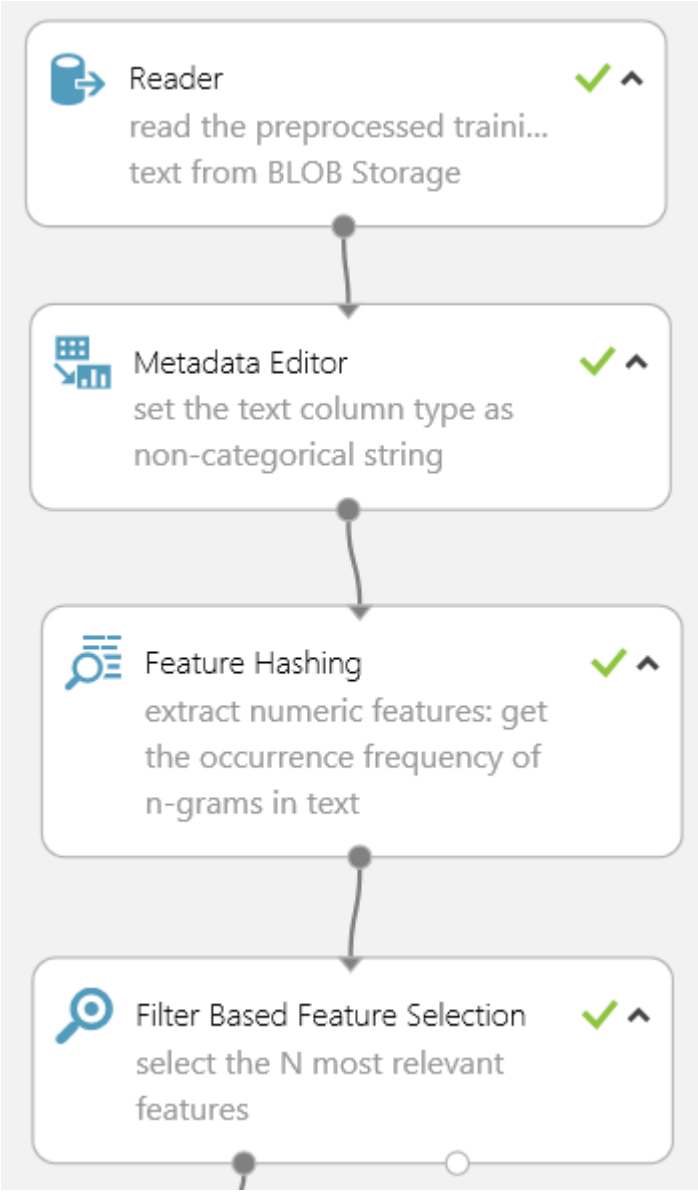
Seçenek 1: Son Satır Ekle modülünün çıkış bağlantı noktasına tıklanır ve Veri Kümesi Olarak Kaydet öğesi seçilir. Veri kümesi adlandırılır (*Text - Preprocessed Input*).

Seçenek 2: Denemeye bir Yazar modülü eklenir ve bu son Satır Ekle modülünün çıkış bağlantı noktasına eklenir ve çıktı önceden işlenmiş metin veri kümesi bir Azure SQL veritabanındaki, Windows Azure tablosundaki, BLOB deposundaki veya bir Hive tablosundaki bir alana yazılır.

2.5. Adlandırılmış son R Script'i Yürüt modülünün ikinci çıkış bağlantı noktasına gidilir ve her sınıf için en sık kullanılan sözcükleri görmek gerekiyorsa Görselleştir seçilir. Kabul edilen kullanım durumunda, ilk kelime bulutu, en iyi pozitif duygu taşıyan kelimeleri temsil eder ve ikinci kelime bulutu, girdi eğitimi korpusunda en sık negatif duygu taşıyan kelimeleri gösterir

Adım 3: Özellik mühendisliği

Microsoft Azure programı üzerinde özellik seçimi aşamalarının akışı Şekil 3.5'te gösterilmiştir.



Şekil 3.6. Özellik seçimi

Bir n-gram, belirli bir metin dizisinden gelen bitişik bir n terim dizisidir. 1 boyutundaki bir n-gram, bir unigram olarak adlandırılır; 2 büyüklüğünde bir n-gram bir bigramdır ; 3 boyutlu bir n-gram bir trigramdır . Daha büyük boyutlu N-gram'lar bazen n değeriyle anılır, örneğin "dört gram", "beş gram" vb.

Özellik karma

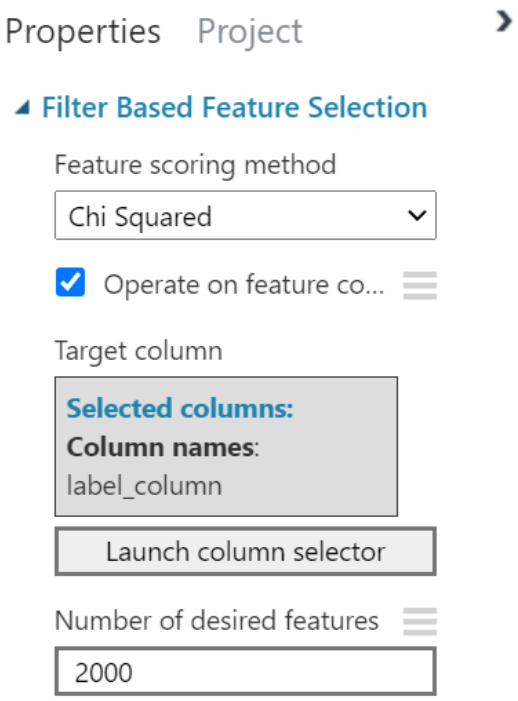
Vowpal Wabbit kitaplığı Özellik Karma modülü tarafından sağlanan bir 32-bit murmurhash v3 karma yöntemi kullanılarak eşit uzunlukta sayısal özellik vektörlerine değişken uzunluklu metin belgesine dönüştürmek için kullanılabilir. Özellik karmasını

kullanmanın amacı, boyutsallığı azaltmaktır; ayrıca özellik karma, dize karşılaştırması yerine karma değer karşılaştırmasını kullandığından, sınıflandırma zamanında özellik ağırlıklarının daha hızlı aranmasını sağlar.

Örnek deneyde, hash bitlerinin sayısı 15'e ve n-gram sayısı2'ye ayarlandı. Bu ayarlarla, hash tablosu 2^15 veya 32.768 giriş tutabilir. Her bir hashing özelliği bir veya more n-gram özellikleri ve değeri, metin örneğinde o n-gram'ın oluşum sıklığını temsil eder. Birçok problem için bu boyuttaki bir hash tablosu fazlasıyla yeterlidir, ancak bazı durumlarda çarpışmaları önlemek için daha fazla alana ihtiyaç duyulabilir.

Özellik seçimi (boyutsallık azaltma)

Eğitimli bir modelin sınıflandırma süresi ve karmaşıklığı, özniteliklerin sayısına (giriş uzayının boyutluluğuna) bağlıdır. Destek vektör makinesi gibi doğrusal bir model için karmaşıklık, özellik sayısına göre doğrusaldır. Metin sınıflandırma görevleri için, kelime dağarcığındaki her bir kelime ve her bir n-gram bir özelliğe eşlendiğinden, özellik çıkarımından kaynaklanan özelliklerin sayısı yüksektir. Ayıklanan karma özelliklerin kapsamlı listesinden daha kompakt bir özellik alt kümesi seçmek için Filtre Tabanlı Özellik Seçimi kullanıldı. Amaç, boyutsallık etkilerinden kaçınmak ve sınıflandırma doğruluğuna zarar vermeden hesaplama karmaşıklığını azaltmaktır. 2^15 ayıklanan özelliklerden duygu etiketine göre en alakalı ilk 2.000 özelliği elde etmek için, hash özelliklerini azalan düzende sıralamak için Ki-kare puanı işlevi kullanıldı. Özellik seçiminin modül girdileri Şekil 3.6'da gösterilmiştir.



Şekil 3.7. Özellik seçimi modül girdileri

Filtre tabanlı özellik seçimi nedir ve neden kullanılır?

Özellik seçimi için bu modül "filtre tabanlı" olarak adlandırılır, çünkü seçili metriği alakasız öznitelikleri belirlemek ve modelden gereksiz sütunları filtrelemek için kullanılır. Verilere uygun tek bir istatistiksel ölçü seçilirse modül her özellik sütunu için bir puan hesaplar. Sütunlar özellik puanlarına göre sıralanmış olarak döndürülür.

Doğru özellikler seçilerek sınıflandırmanın doğruluğu ve verimliliği potansiyel olarak artırabilir.

Tahmine dayalı modelini oluşturmak için genellikle yalnızca en iyi puanlara sahip sütunlar kullanılır. Zayıf özellik seçimi puanlarına sahip sütunlar bir model oluşturulduğunda veri kümesinde bırakılabilir ve yoksayılabilir.

Bir özellik seçimi metriği nasıl seçilir?

Filtre Esaslı özellik seçimi, her sütun içinde bilgi değeri değerlendirmek için bir metrik sunar. Bu bölüm, her bir metriğin genel bir tanımını ve nasıl uygulandığını sunar.

Pearson Korelasyonu

Pearson korelasyon istatistiği veya Pearson korelasyon katsayısı, istatistiksel modellerde r değer olarak da bilinir. Herhangi iki değişken için korelasyonun gücünü gösteren bir değer döndürür.

Pearson korelasyon katsayısı, iki değişkenin kovaryansı alınarak ve standart sapmalarının çarpımına bölünerek hesaplanır. Katsayı, iki değişkendeki ölçek değişikliklerinden etkilenmez.

Karşılıklı bilgi

Karşılıklı bilgi puanı, bir değişkenin başka bir değişkenin değeri hakkındaki belirsizliği azaltmaya yönelik katkısını ölçer. Karşılıklı bilgi puanının birçok varyasyonu farklı dağılımlara uyacak şekilde tasarlanmıştır.

Karşılıklı bilgi puanı, birçok boyutlu veri setlerinde ortak dağılım ve hedef değişkenler arasındaki karşılıklı bilgiyi en üst düzeye çıkardığı için özellik seçiminde özellikle yararlıdır.

Kendall Korelasyonu

Kendall'ın sıra korelasyonu, farklı sıralı değişkenlerin sıralamaları veya aynı değişkenin farklı sıralamaları arasındaki ilişkiyi ölçen birkaç istatistikten biridir. Başka bir deyişle, miktarlara göre sıralandığında siparişlerin benzerliğini ölçer. Hem bu katsayı hem de Spearman's korelasyon katsayısı, parametrik olmayan ve normal dağılıma sahip olmayan verilerle kullanılmak üzere tasarlanmıştır.

Spearman Katsayısı

Spearman katsayısı, iki değişken arasındaki istatistiksel bağımlılığın parametrik olmayan bir ölçüsüdür ve bazen Yunan harfi ρ ile gösterilir. Spearman katsayısı, iki değişkenin monoton olarak ilişkili olma derecesini ifade eder. Sıralı değişkenlerle birlikte kullanılabildiği için Spearman sıra korelasyonu olarak da adlandırılır.

Ki Kare

İki yönlü ki-kare testi, beklenen değerlerin gerçek sonuçlara ne kadar yakın olduğunu ölçen istatistiksel bir yöntemdir. Yöntem, değişkenlerin rastgele olduğunu ve yeterli sayıda bağımsız değişken örneğinden alındığını varsayar. Ortaya çıkan ki-kare istatistiği, sonuçların beklenen (rastgele) sonuçtan ne kadar uzak olduğunu gösterir.

Fisher Skoru

Fisher puanı (Fischer yöntemi veya Fisher birleşik olasılık puanı olarak da adlandırılır) bazen bilgi puanı olarak adlandırılır çünkü bir değişkenin bağlı olduğu bazı bilinmeyen parametreler hakkında sağladığı bilgi miktarını temsil eder.

Skor, bilginin beklenen değeri ile gözlenen değer arasındaki varyans ölçülerek hesaplanır. Varyans minimize edildiğinde bilgi maksimize edilir. Puan beklentisi sıfır olduğu için Fisher bilgisi aynı zamanda puanın varyansıdır.

Sayıma Dayalı

Sayıma dayalı özellik seçimi tahminciler hakkında bilgi bulmanın basit ama nispeten güçlü bir yoludur. Sayıya dayalı özellikleştirmenin altında yatan temel fikir basittir: bir sütundaki tek tek değerlerin sayısı hesaplanarak değerlerin dağılımı ve ağırlığı hakkında bir fikir edinilebilir ve bundan hangi sütunların en önemli bilgileri içerdiği anlaşılabilir.

Sayıya dayalı özellik seçimi denetimsiz bir özellik seçimi yöntemidir yani bir etiket sütununa ihtiyacınız yoktur. Bu yöntem aynı zamanda bilgi kaybı olmadan verilerin boyutsallığını da azaltır.

3A.1. Sonuçları bir Azure SQL veri tabanı, Windows Azure tablosu, BLOB Depolaması veya bir Hive tablosundaki bir alana kaydetmek için daha önce (adım 2'de) Writer modülü kullanılmış ise verileri yüklemek için Reader modülü kullanılır.

3A.2. Özellik Karma modülünde *Hashing bitsize* parametresi kullanılarak karma bitlerin sayısı belirtilir ve parametre kullanılarak n-gram boyutu belirtilir (*N-grams*). Örnek deneyde 2^15 = 32.768 hashing özelliğini çıkarmak için bit boyutu 15

bit olarak belirlendi. Daha iyi sınıflandırma performansı elde etmek için bit boyutu arttırılabilir.

3A.3. Filtre Tabanlı Özellik Seçimi modülü tıklanır ve özellik puanlama yöntemi ve istenen özelliklerin sayısı belirtilir. Örnek deneyde, en alakalı 2.000 özellik seçildi. Daha iyi sınıflandırma performansı elde etmek için istenen özelliklerin sayısı artırılabilir.

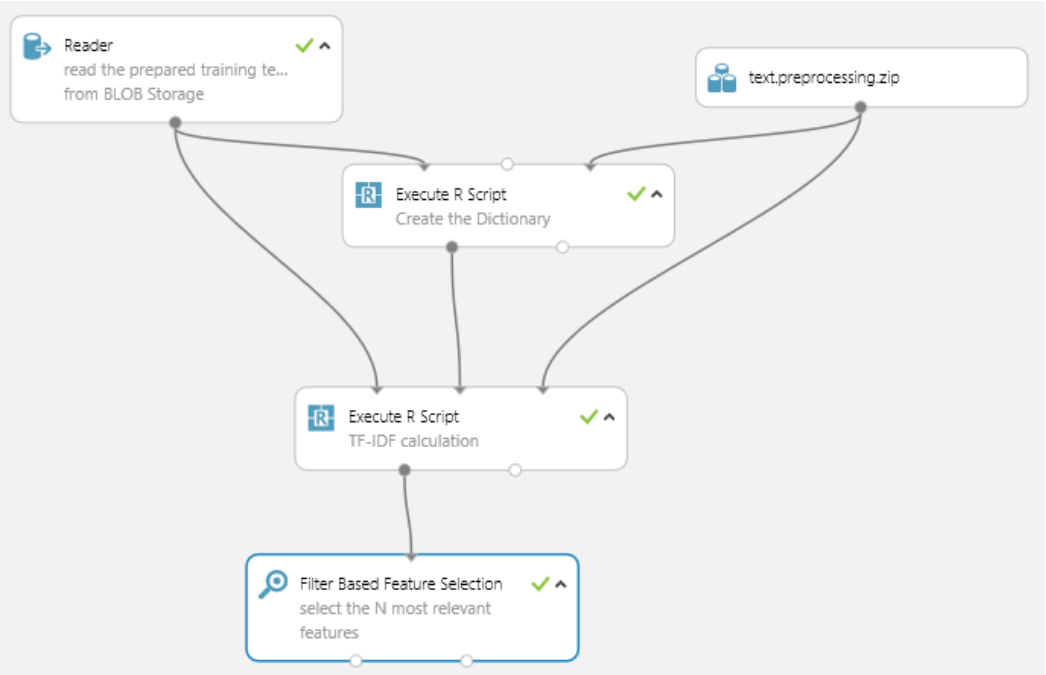
3A.4. Deney çalıştırılır.

3A.5. Deneme başarıyla tamamlandıktan sonra aşağıdaki seçeneklerden biri kullanılarak sonuçlar kaydedilir:

Seçenek 1: Filtre Tabanlı Özellik Seçimi modülünün sol çıkış bağlantı noktasına tıklanır ve Veri Kümesi Olarak Kaydet öğesi seçilir. Veri kümesi adlandırılır (*Text - Extracted N-grams TF*).

Seçenek 2: Denemeye bir Writer modülü eklenir ve bu Filtre Tabanlı Özellik Seçiminin sol çıkış bağlantı noktasına eklenir ve ayıklanan özellikler bir Azure SQL veri tabanındaki, Windows Azure tablosundaki, BLOB deposundaki veya bir Hive tablosundaki bir alana yazılır.

Adım 3B: unigrams TF-IDF özellik çıkarma



Şekil 3.8. Özellik çıkarımı akış

İlk olarak, metin modelini eğitmek için kullanılacak unigram (kelime) seti çıkarılır. Unigramlara ek olarak, metin korpusunda her kelimenin göründüğü belge sayısı sayılır (DF). Sözlüğü metin modelini eğitmek için kullanılan aynı etiketli veriler üzerinde oluşturmak gerekli değildir. Hedef ile açıklama eklenmemiş olsa bile sınıflandırmanın hedef alanındaki kelimelerin sıklığını adil bir şekilde temsil eden herhangi bir büyük bütünlük kullanılabilir. Özellik çıkarımının Microsoft Azure üzerindeki akışı Şekil 3.7'de gösterilmiştir.

TF-IDF Hesaplaması

Bir belgede metrik sözcük oluşum sıklığı (TF) bir özellik değeri olarak kullanıldığında, bir bütüncede (durdurma sözcükleri gibi) sıkça görünen sözcüklere daha yüksek bir ağırlık atanma eğilimi gösterir. Ters belge sıklığı (IDF), sık kullanılan sözcüklere daha düşük bir ağırlık verdiği için daha iyi bir ölçümdür. IDF'yi, eğitim külliyatındaki belge sayısının verilen kelimeyi içeren belge sayısına oranının günlüğü olarak hesaplarsınız. Bu sayıları bir metrikte (TF/IDF) birleştirmek, belgede sık görülen ancak bütüncede nadir bulunan sözcüklere daha fazla önem verir. Bu varsayım sadece unigramlar için değil, bigramlar, trigramlar vb. için de geçerlidir.

Bu, yapılandırılmamış metin verilerini her özelliğin bir metin örneğindeki bir unigramın TF-IDF'sini temsil ettiği eşit uzunluktaki sayısal özellik vektörlerine dönüştürür.

Özellik seçimi (boyutsallık azaltma)

Hash özelliklerini azalan düzende sıralamak için Ki-kare puanı işlevi kullanıldı ve çıkarılan tüm unigramlardan duygu etiketine göre en alakalı ilk 2.000 özellik döndürüldü.

3B.1. Daha önce (2. Adımda) sonuçları kaydetmek için Writer modülünü kullanıldıysa önceden işlenmiş eğitim verilerini yüklemek için Reader modülü bir Azure SQL veri tabanına, Windows Azure tablosuna, BLOB deposuna veya bir Hive tablosuna eklenir.

3B.2. Grafikteki ilk Execute R Script modülünde giriş veri kümesinden oluşturulan sözlüğe dahil edilecek bir kelime için aşağıdaki parametreler belirtilir:

a. bir kelimenin minimum (*minWordLen*) ve maksimum (*maxWordLen*) uzunluğu

b. minimum (*minDF*) ve maksimum (*maxDF*) belge oluşum sıklığı

R komutları Şekil 3.8'de gösterilmiştir.

Execute R Script

R Script

```
# Map 1-based optional input ports to variables
dataset <- maml.mapInputPort(1) # class: data.frame
##################################################
# Determine the following input parameters:-
# minimum length of a word to be included into the dictionary.
# Exclude any word if its length is less than *minWordLen* characters.
minWordLen <- 3

# maximum length of a word to be included into the dictionary.
# Exclude any word if its length is greater than *maxWordLen* characters.
maxWordLen <- 25

# minimum document frequency of a word to be included into the dictionary.
# Exclude any word if it appears in less than *minDF* documents.
minDF <- 9

# maximum document frequency of a word to be included into the dictionary.
# Exclude any word if it appears in greater than *maxDF* documents.
maxDF <- Inf

# we assume that the text is the second column in the input data frame
```

Şekil 3.9. R komutları

3B.3. İkinci R Komut Dosyası Yürütme modülünde (*TF-IDF Calculation*) adım 3B.3'te tanımlanan *minWordLen* ve *maxWordLen* için aynı değerlerin belirtildiğinden emin olunmalıdır . İkinci R komut dosyasının komutları Şekil 3.9'de gösterilmiştir.

Execute R Script

R Script

```
# Map 1-based optional input ports to variables
dataset <- maml.mapInputPort(1) # class: data.frame
input.dictionary <- maml.mapInputPort(2) # class: data.frame
##################################################
# Determine the following input parameters:-
# minimum length of a word to be included into the dictionary.
# Exclude any word if its length is less than *minWordLen* characters.
minWordLen <- 3

# maximum length of a word to be included into the dictionary.
# Exclude any word if its length is greater than *maxWordLen* characters.
maxWordLen <- 25
```

Şekil 3.10. 2. R komutları

3B.4. Filtre Tabanlı Özellik seçimi modülünde özellik puanlama yöntemi ve istenen özelliklerin sayısı belirtilir. Örnek deneyde en alakalı 2.000 özellik seçildi.

3B.5. Deney çalıştırıldı.

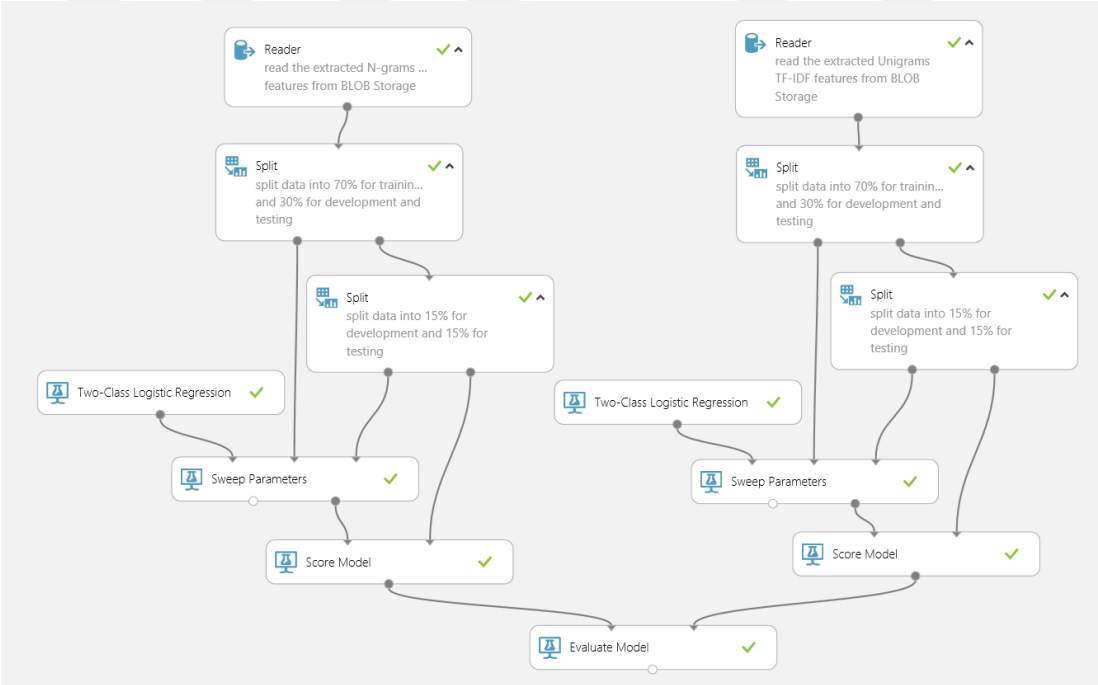
3B.6. Deneme başarıyla tamamlandıktan sonra, Execute R Script modülünün sol çıkış portuna tıklanır ve Save as Dataset öğesi seçilir. Veri kümesi adlandırılır (*Text - Unigrams Dictionary*).

3B.7. Aşağıdaki seçeneklerden biri kullanılarak özellik seçiminin çıktısı kaydedilir:

Seçenek 1: Filtre Tabanlı Özellik Seçimi modülünün sol çıkış bağlantı noktasına tıklanır ve Veri Kümesi Olarak Kaydet öğesi seçilir. Veri kümesi adlandırılır (*Text - Extracted Unigrams TF-IDF*).

Seçenek 2: Denemeye bir Writer modülü eklenir ve bu Filtre Tabanlı Özellik Seçiminin sol çıkış bağlantı noktasına eklenerek çıktı alınan özellikler veri kümesi, bir Azure SQL veri tabanındaki, Windows Azure tablosundaki, BLOB deposundaki veya bir Hive içindeki bir tabloya yazılır.

Adım 4: Modellerin eğitimi ve değerlendirilmesi



Şekil 3.11. Modellerin eğitimi ve değerlendirilmesi

4.1. Sonuçlar bir tabloya kaydetmek için daha önce (adım 2'de) Writer modülü kullanılmışsa, verileri yüklemek için Reader modülü kullanılır.

4.2. Verileri iki alt kümeye bölmek için ilk Bölme modülü kullanılır. İlk alt küme modeli eğitmek için kullanılacak ve ikinci alt küme, bir sonraki adımda geliştirme/doğrulama kümesi ve test kümesine bölünecektir. Örnek deneyde veriler sırasıyla %70 ve %30 olarak ayrıldı.

4.3. Verileri iki alt kümeye bölmek için ikinci Bölme modülü kullanılır. İlk alt küme daha sonra Sweep Parameters modülü tarafından kullanılacaktır. İkinci alt küme eğitilen modelin performansını değerlendirmek için test kümesi olarak kullanılır. Örnek deneyde %30 veri örneği iki yarıya bölündü. Yani geliştirme setinin ve test setinin her biri girdi verilerinin %15'ini temsil etmektedir.

4.4. Sweep Parametreleri altında yatan öğrenme algoritması parametrelerinin optimum değerlerini almak için örnek denemede parametre tarama modu, modülün parametre aralıklarından bir dizi eğitim çalıştırması gerçekleştirmek için Rastgele tarama olarak ayarlanır . İki Sınıflı Lojistik Regresyon modülü gibi öğrenme algoritması modülünde belirtildiği gibi her bir parametre için olası tüm değerleri keşfetmek için bir parametre süpürme modu olarak Tüm ızgara seçeneği kullanılır. Örnek deneyde şu şekilde belirtilir: AUCMetric for measuring performance for classification. Model seçimi için kesinlik, geri çağırma ve F-puanı gibi diğer performans değerlendirme kriterleri kullanılabilir.

4.5. İkili sınıflandırma görevleri için, İki Sınıflı Lojistik Regresyon modülü tutulabilir veya İki Sınıflı Destek Vektör Makinesi, İki Sınıflı Yükseltilmiş Karar Ağacı vb. gibi başka bir ikili sınıf sınıflandırma eğiticisi ile değiştirilebilir.

4.6. Çok sınıflı sınıflandırma görevleri için, İki Sınıflı Lojistik Regresyon modülünün Bire Karşı Tüm Çok Sınıflı, Çok Sınıflı Lojistik Regresyon, Çok Sınıflı Karar Ormanı vb. gibi başka bir çok sınıflı sınıflandırma eğiticisi ile değiştirilmesi gerekir.

4.7. Deney çalıştırılır.

4.8. Deneme başarıyla tamamlandıktan sonra sonuçları aşağıdaki gibi kaydedilir:

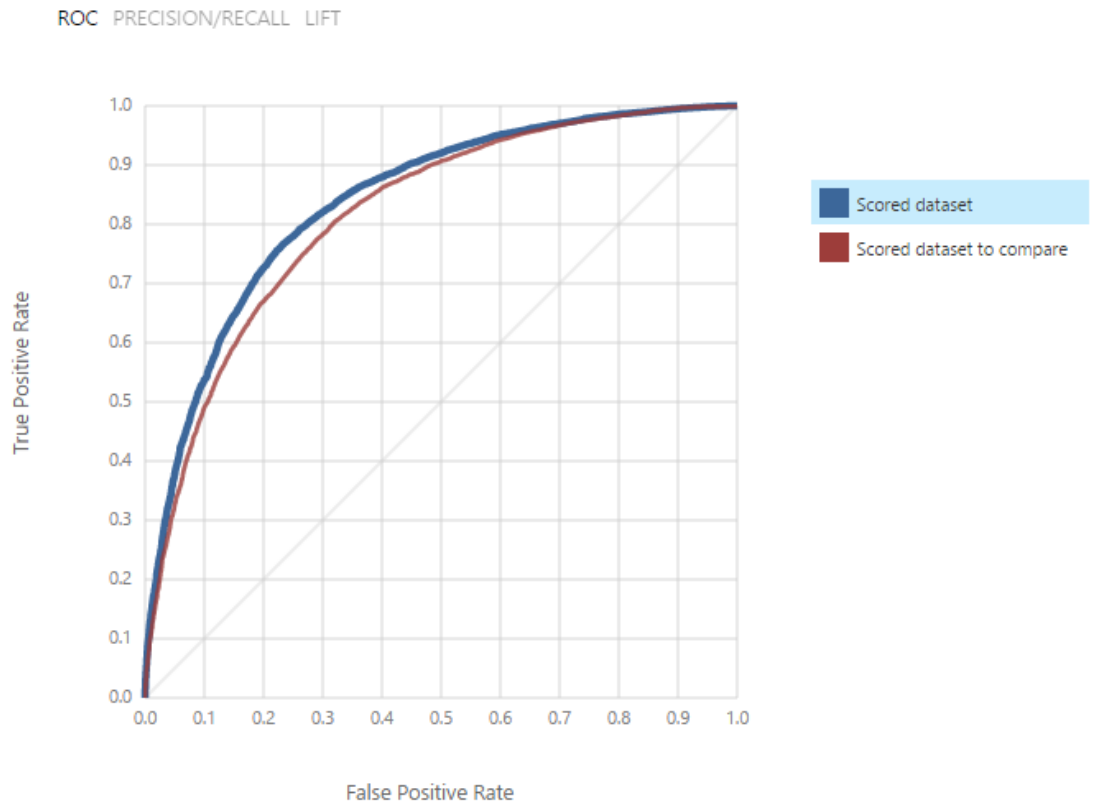
a. Sol Süpürme Parametreleri modülünün sağ çıkış bağlantı noktasına tıklanır ve Eğitilmiş Model Olarak Kaydet'i seçilir. Model adlandırılır (*Text - Trained N-grams model*).

b. Sağ Sweep Parameters modülünün sağ çıkış portuna tıklanır ve Save as Trained Model (Eğitimli Model Olarak Kaydet) öğesi seçilir. Model adlandırılır (*Text -Trained Unigrams model*).

Modellerin eğitimi ve değerlendirilmesinin akışı Şekil 3.10'da gösterilmektedir.

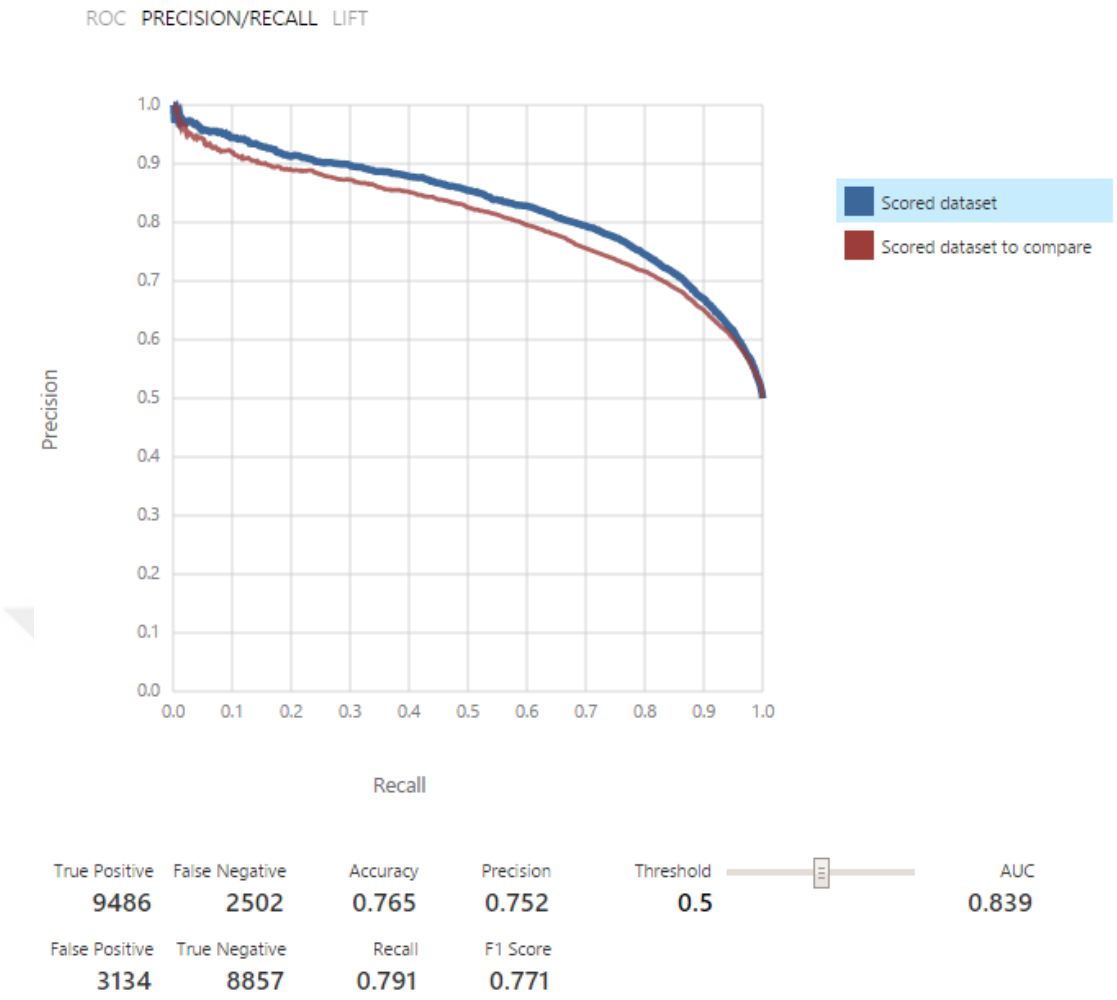
4.9. Evaluate Model modülünün çıkış bağlantı noktasına tıklanır ve iki eğitilmiş model arasındaki karşılaştırmanın sonuçlarını görselleştirilir: N-gram modeli (aşağıdaki grafiklerde mavi renkte) ve unigrams modeli (aşağıdaki grafiklerde kırmızı renkte). Bu değerlendirmeye dayanarak en iyi eğitilmiş model belirtilmiştir.

Şekil 3.11'de ROC Eğrisi (Receiver Operating Characteristic Curve-Alıcı İşletim Karakteristik Eğrisi) grafiği gösterilmiştir. Makine öğrenmesinde bir sınıflandırma probleminin performansının değerlendirilmesinde ROC eğrisinden yararlanılır. Bu eğri modelin tahmininin ne derecede iyi olduğunu açıklar. AUC, "ROC Eğrisi altındaki alan" anlamına gelir. Kapsanan alan ne kadar büyükse, makine öğrenme modelinin verilen sınıfları ayırt etme yeteneğinin de o kadar iyi olduğu söylenebilir.



Şekil 3.12. ROC eğrisi

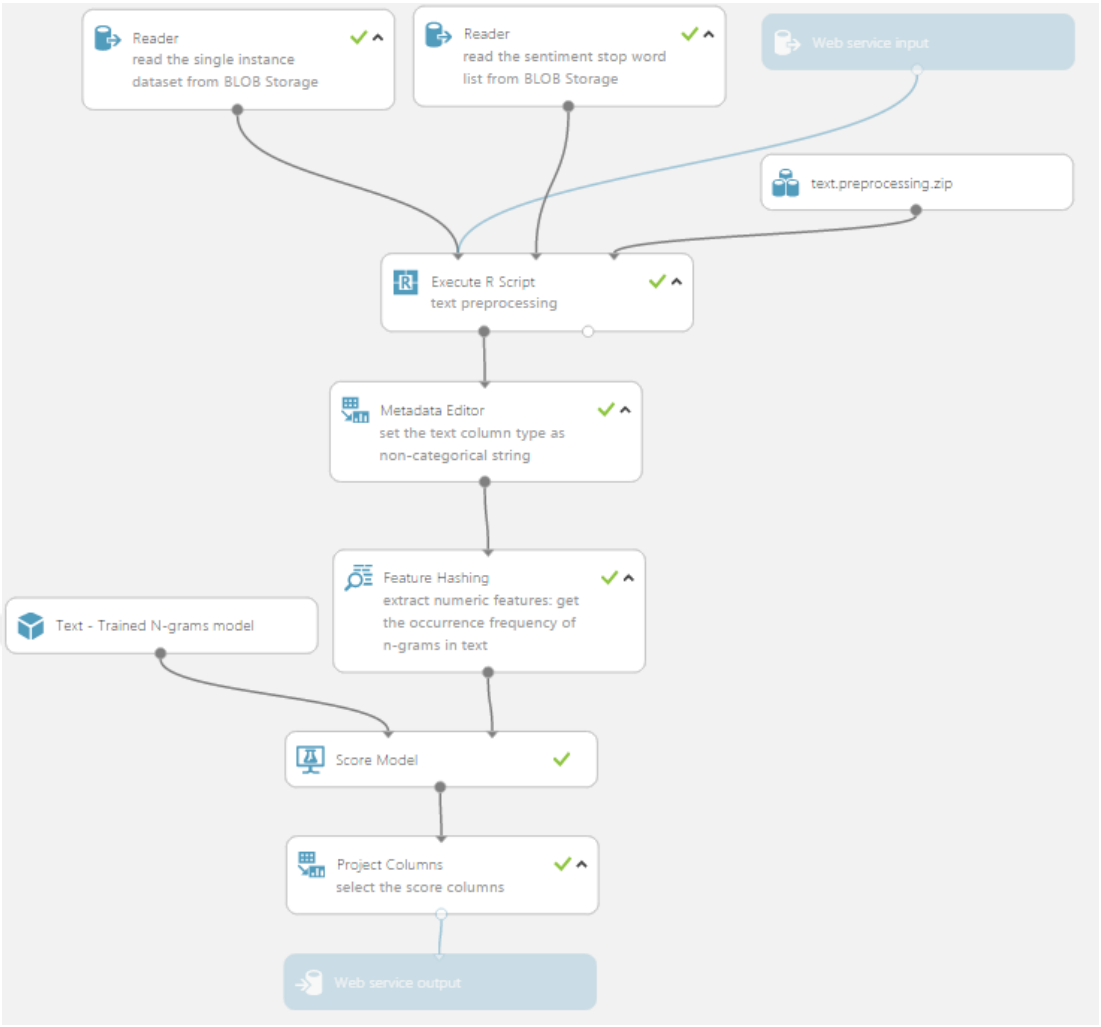
Şekil 3.12'de Precision/Recall (Hassasiyet/Geri çağırma) grafiği gösterilmiştir. Hassasiyet/Geri çağırma, sınıflar çok dengesiz olduğunda tahmin başarısının bir ölçüsüdür. Bilgi alımında hassasiyet, sonuç alaka düzeyinin bir ölçüsüyken, geri çağırma, gerçekten alakalı kaç sonucun döndürüldüğünün bir ölçüsüdür. Eğrinin altındaki yüksek alan, hem yüksek geri çağırmayı hem de yüksek kesinliği (hassasiyeti) temsil eder; burada yüksek hassasiyet, düşük yanlış pozitif oranıyla ve yüksek geri çağırma, düşük yanlış negatif oranıyla ilgilidir. Her ikisi için de yüksek puanlar, sınıflandırıcının doğru sonuçlar verdiğini (yüksek kesinlik) ve ayrıca tüm olumlu sonuçların çoğunluğunu (yüksek hatırlama) gösterir.



Şekil 3.13. Hassasiyet/Geri çağırma eğrisi

Adım 5: Eğitimli modellerin web hizmetleri olarak dağıtımı

Web hizmeti iki modda tüketilebilir: RRS (istek-yanıt hizmeti) ve BES (toplu yürütme hizmeti). Web servislerini aramak için örnek kod (C#/python/R) web servisinde verilmektedir.



Şekil 3.14. Web hizmeti akış

Bu deney, Adım 4'te eğitilmiş N-gram TF metin modelini bir web hizmeti olarak dağıtır.

Web servis giriş ve çıkış noktaları, özel Web Servis modülleri kullanılarak tanımlanır. Web hizmeti giriş modülü, deneydeki giriş verilerinin gireceği düğüme eklenir.

5A.1. Yürütme R Senaryo modülünde, Adım 2.2'de tanımlanan aynı parametreler kullanılarak gerekli metin ön işleme adımları belirtilir.

5A.2. Özellik Hashing modülünde, *Hashing bitsize* parametresi kullanılarak aynı sayıda bit ve Adım 3A.1'de tanımlanan parametre kullanılarak aynı n-gram boyutu belirtilir.

5A.3. Web hizmeti giriş noktası ayarlanır.

5A.4. Web hizmeti çıkış noktası ayarlanır.

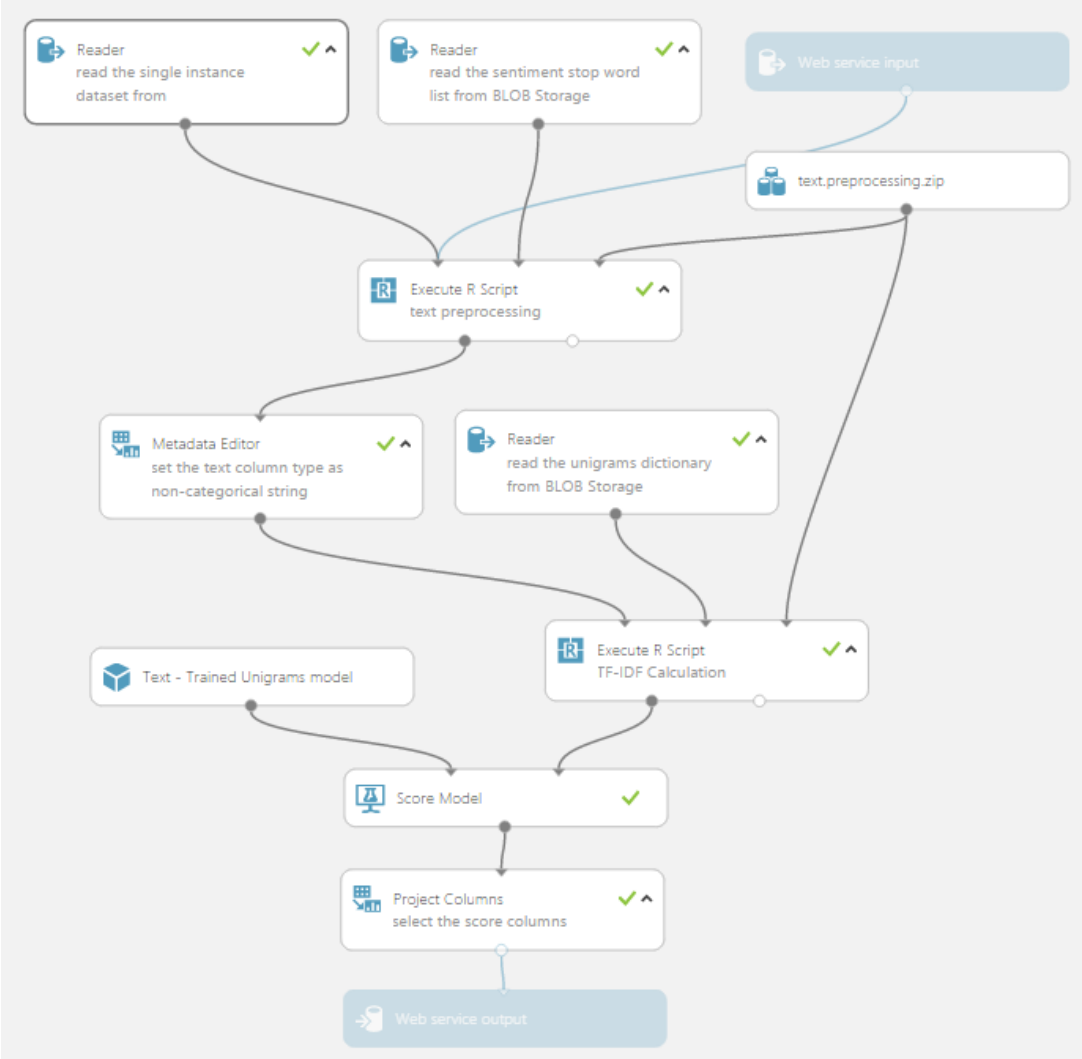
5A.5. Deney çalıştırılır.

5A.6. Deneme başarıyla tamamlandıktan sonra deneme tuvalinin altındaki Web Hizmetini Yayınla seçilir.

5A.7. API yardım sayfasından yayınlanan web hizmetini çağırmak için örnek kodu kullanılarak yazılım entegrasyonu sağlanabilir.

5A.8. Gösterge Tablosu sayfasından API anahtarı kopyalanarak örnek koda yapıştırılır.

Adım 5B: Unigrams TF-IDF tarafından eğitilmiş model bir web hizmeti olarak yayınlanarak kullanıma alma işlemi tamamlanmış olur.



Şekil 3.15. Web hizmeti output

Bu deney, Adım 4'te eğitilmiş unigrams TF-IDF metin modelini bir web hizmeti olarak dağıtır. Web servis giriş ve çıkış noktaları, özel Web Servis modülleri kullanılarak tanımlanır. Web hizmeti giriş modülünün, deneydeki giriş verilerinin gireceği düğüme eklenmiştir.

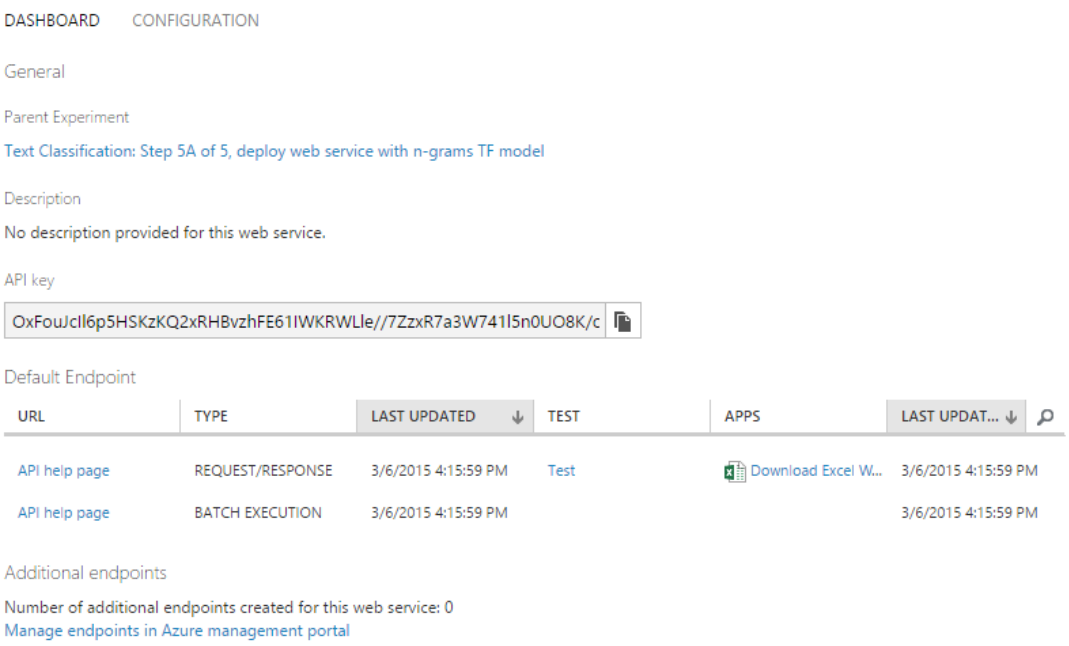
5B.1. Yürütme R Senaryo modülünde, Adım 2.2'de tanımlanan aynı parametreler kullanılarak, gerekli metin önişleme adımları belirtilir.

5B.2. Yürütme R Komut modülünde aşama 3B.3 de tarif edildiği gibi aynı parametreler kullanılır (*TF-IDF Calculation*).

5B.3. Deney çalıştırılır.

5B.4. Deneme başarıyla tamamlandıktan sonra, deneme tuvalinin altındaki Web Hizmetini Yayınla seçilir.

5B.5. API yardım sayfasından, yayınlanan web hizmetini çağırmak için örnek kodu kullanılarak yazılım entegrasyonu sağlanabilir.



Şekil 3.16. Örnek kod ile web hizmeti çağırma

4. SONUÇ VE ÖNERİLER

Bu araştırma, üretim sektöründe çalışanların firmaya bağlılığını arttırmak ve aynı zamanda firmaya kazanç sağlamak amacıyla yürütülen ve her geçen gün sayısı artan çalışan öneri sistemlerinin analizlerinin yapılabilmesi, kolay yorumlanabilir hale getirilmesi ve bu önerilerin firmanın geleceğine dair fikirler vermesinin sağlanması amacıyla yapılmıştır.

Araştırmanın çıktısı olarak, çalışanların verdikleri önerilerin benzer özellikte olanlarının sınıflandırılması ve bunun sonucunda çalışanların hangi konularda daha fazla öneri verdiğinin, hangi konularda öneri vermediğinin, firmanın hangi konulara eğilmesi gerektiğinin belirlenmesi gösterilebilir.

Literatürdeki daha önceki benzer çalışmalarda öneri sistemlerinin konu modellemesi ile analizinin yapılmamış olmasının, bu çalışmayı bilimsel açıdan önemli hale getirdiği söylenebilir.

Yönetsel açıdan bakıldığında ise, uygulama yapılan firmada daha önce öneri sistemleriyle ilgili hiçbir analiz yapılmadığından dolayı bu çalışmanın firmanın yönetsel kararlarını olumlu yönde etkileyeceği söylenebilir.

Çalışmanın uygulama aşamasında, eğitim için gerekli veriler uygulamaya yüklenip analiz tamamlandıktan sonra aşağıda gösterildiği gibi sentiment_label histogram grafiği çıkarılmıştır.

Etiketlerin temsil ettiği kümeler aşağıdaki gibidir;

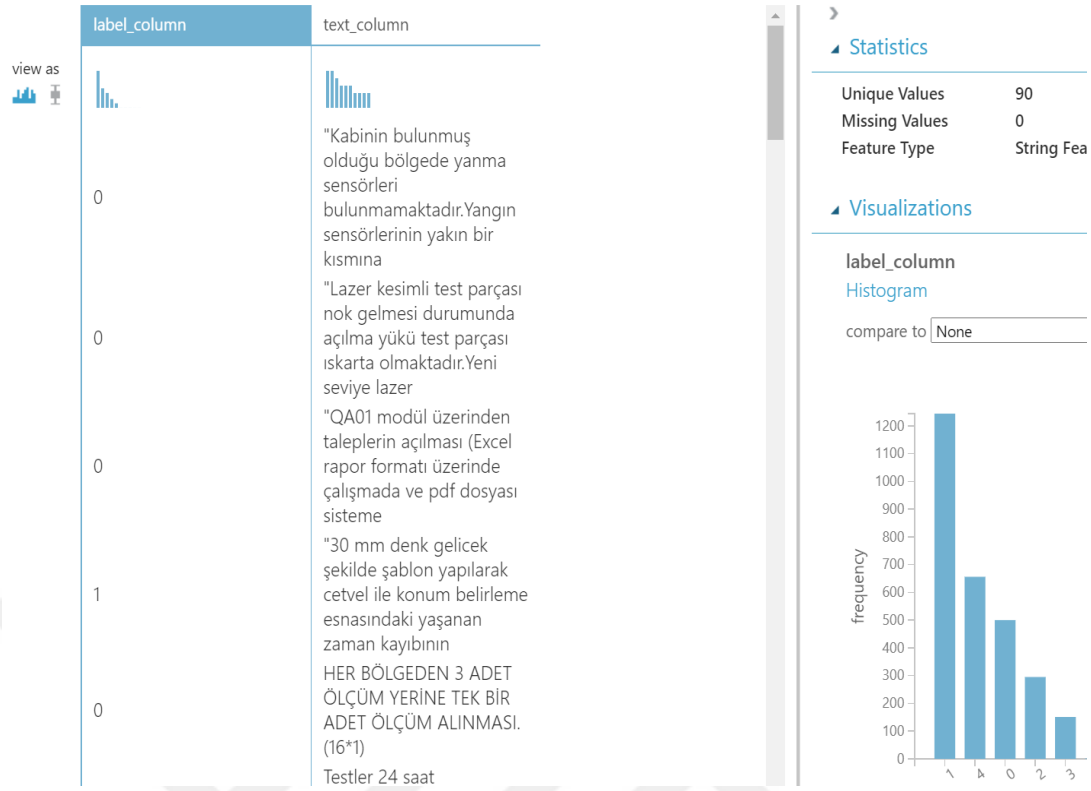
0=Öneriden Hızlı Kaizene,

1=Getirisi Olmayan Olumlu Öneri,

2=Değerlendirilmek Üzere Havale,

3=Devreye Alınmayacak Öneri,

4=Öneri



Şekil 4.1. Sentiment Label Histogram Grafiği

Histogram grafiğinde görüldüğü gibi; en çok verilen öneri çeşiti "getirisi olmayan olumlu öneri"lerdir. Bu öneriler genellikle iş sağlığı ve güvenliği ile ilgili olmaktadır. 2. Sıradaki en çok verilen öneriler ise "öneri" olarak adlandırdığımız yani getirisi olan, firmaya kazanç sağlayan önerilerdir. 3. Sırada "öneriden hızlı kaizene" yani çok kısa sürede sonuç alınabilen, getirisi yüksek öneriler bulunmaktadır. 4. Sırada "değerlendirilmek üzere havale" edilen öneriler bulunurken, en az verilen öneri türünün ise "devreye alınmayacak öneriler" olduğu görülmektedir.

Gelecek çalışmalarda, önerilerin arttığı ve azaldığı dönemler, önerilerin bölümler arası değişkenliği, öneri verilen konuların dağılımı gibi detaylı veri analizleri yapılabilir.

KAYNAKLAR

[1] **Schwarz, C.** (2018). ldagibbs: A command for topic modeling in Stata using latent Dirichlet allocation, *The Stata Journal*, 18(1), 101-117.

[2] **Sun, M., & Zheng, H.** (2018). Topic Detection for Post Bar Based on LDA Model, *International Conference of Pioneering Computer Scientists, Engineers and Educators* (pp. 136-149), Springer, Singapore.

[3] **Shah, A. H.** (2019). How episodic frames gave way to thematic frames over time: A topic modeling study of the Indian media's reporting of rape post the 2012 Delhi gang-rape, *Poetics*, 72, 54-69.

[4] **Karami, A., Ghasemi, M., Sen, S., Moraes, M. F., & Shah, V.** (2019). Exploring diseases and syndromes in neurology case reports from 1955 to 2017 with text mining, *Computers in biology and medicine*, 109, 322-332.

[5] **Onan, A., Bulut, H., & Korukoglu, S.** (2017). An improved ant algorithm with LDA-based representation for text document clustering, *Journal of Information Science*, 43(2), 275-292.

[6] **Blei, D. M., Ng, A. Y., & Jordan, M. I.** (2003). Latent dirichlet allocation, *Journal of machine Learning research*, January, 993- 1022.

[7] **Agrawal, A., Fu, W., & 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.

[8] **Kherwa1, P. ve Bansal, P.** (2019). Topic Modeling: A Comprehensive Review, Maharaja Surajmal Institute of Technology, C-4 Janak Puri. GGSIPU. New Delhi-110058.

[9] **Chen, T-H., Thomas, SW., Hassan, AE.** (2016). A Survey on The Use of Topic Models When Mining Software Repositories, *Empirical Software Engineering*, 21(5):1843–1919.

[10] **Daud, A. et al** (2010). Knowledge Discovery Through Directed Probabilistic Topic Models: A Survey. *Frontiers of Computer Science in China*, 4(2):280–301.

[11] **Debortoli, S. et al** (2016). Text Mining for Information Systems Researchers: An Annotated Topic Modeling Tutorial, CAIS 39:7.

[12] **Sun, X. et al** (2016). Exploring Topic Models in Software Engineering Data Analysis: A Survey, *17th IEEE/ACIS international conference on software engineering, artificial intelligence, networking and parallel/distributed computing (SNPD)*, IEEE.

[13] **Li, F., Huang, M., Zhu, X.** (2010). Sentiment Analysis with Global Topics and Local Dependency, *Proceedings of the 24th AAAI Conference on Artificial Intelligence.* pp. 1371-1376.

[14] **Wang, W.** (2010). Sentiment Analysis of Online Product Reviews with Semi-supervised Topic Sentiment Mixture Model, *Proceedings of 7th International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 2385-2389.

[15] **Jo, Y., Oh, A.** (2011). Aspect and Sentiment Unification Model for Online Review Analysis, *Proceedings of 4th ACM International Conference on Web Search and Data Mining*, pp. 815-824.

[16] **Xueke, X., Xueki, C., Songbo, T., Yue, L., Huawei, S.** (2013). Aspect-Level Opinion Mining of Online Customer Reviews, *China Communications*, 10(3), pp. 25-41.

[17] **Xianghua, F., Guo, L., Yanyan, G., Zhiqiang, W.** (2013). Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and How Net Lexicon, *Knowledge-Based Systems*, 37, pp. 186-195.

[18**] Ding, W., Song, X., Guo, L., Xiong, Z., Hu, X.** (2013). A Novel Hybrid HDP-LDA Model for Sentiment Analysis, *Proceedings of 2013 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, pp. 329-336.

[19] **Moghaddam, S., Ester, M.** (2013). The Flda Model for Aspectbased Opinion Mining: Addressing the Cold Start Problem, *Proceedings of the 22nd International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, pp. 909–918.

[20] **Wang, T., Cai, Y., Leung, H., Lau, R.Y.K., Li, Q., Min, H.** (2014). Product Aspect Extraction Supervised with Online Domain Knowledge, *Knowledge-Based Systems*, 71, pp. 86-100.

[21] **Zheng, X., Lin, Z., Wang, X., Lin, K.J., Song, M.** (2014). Incorporating Appraisal Expression Patterns into Topic Modeling for Aspect and Sentiment Word Identification, *Knowledge-Based Systems, 61*. pp. 29-47.

[22] **Lau, R.Y.K., Li, C., Liao, S.S.Y**. (2014). Social Analytics: Learning Fuzzy Product Ontologies for Aspectoriented Sentiment Analysis, *Decision Support Systems*, 65, pp. 80-94.

[23] **Yin, S., Han, J., Huang, Y., Kumar, K.** (2014). DependencyTopic-Affects-Sentiment-LDA Model for Sentiment Analysis, *Proceedings of 2014 IEEE 26th International conference on Tools with Artificial Intelligence,* pp. 413-418.

[24] **Bagheri, A., Saraee, M., Jong, F.** (2013). Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews, *Knowledge-Based Systems, 52*, pp. 201–213.

[25] **Poria, S., Chaturvedi, I., Cambria, E., Bisio, F.** (2016). Sentic LDA: Improving on LDA with Semantic Similarity for Aspect-Based Sentiment Analysis, IJCNN.

[26] **Ekinci, E., Omurca, S.** (2016). Ürün Özelliklerinin Konu Modelleme Yöntemi ile Çıkarılması Product Aspect Extraction with Topic Model, *Bilişim Vakfi Bilgisayar Bilimleri Ve Mühendisliği Dergisi*, Cilt: 9 - Sayı:1.

[27] **Vayansky, I., Kumar, A.P.S.** (2020). A Review of Topic Modeling Methods, *Information Systems*, Volume 94, December 2020, 101582.

[28] **Kalıkov, A.** (2006). *Veri Madenciliği ve Bir E-Ticaret Uygulaması*. (Yüksek Lisans Tezi). Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

[29] **İnan, O.** (2003). *Veri Madenciliği*. (Yüksek Lisans Tezi). Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Konya.

[30] **Thuarisingham, B.M.** (2003), Web Data Mining and Applications in Business Intelligence and Counter Terrorism, *CRC Press LLC,* Boca Raton, FL, USA.

[31] **Savaş, N., Topaloğlu, N., Yılmaz, M**. (2012). Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri, *Dergipark,* Cilt:11, Sayı:21, Haziran 1-23.

[32] **Agrawal, R., Ghosh, S., Imielinski, T., Iyer, B., and Swami, A. (**1992). An interval classifier for database mining approaches, *Proceedings of the 18th VLDB Conference,* Vancouver, Canada, August 23-27.

[33] **Han, J., Cai, Y. and Cercone, H.** (1992). Knowledge discovery in databases: An attribute-oriented approach, School of Computing Science, Simon Fraser University, Burnaby, British Columbia, Canada WA IS6.

[34] **Agrawal, R., Imielinski, T. and Swami, A**. (1993). Database mining: A performance perspective, *IEEE Tranactions on Knowledge and Data Engineering,* volume:5, issue:6, pp: 914-915.

[35] **Özekes, S.** (2003). Data Mining Models and Application Areas, *İstanbul Commerce University Journal of Science,* No.3, 65-82.

[36] **Han, J., Kamber, M.** (2000). Data Mining Concepts and Techniques, *Morgan Kaufmann Publishers*, 1st Ed., San Francisco, USA.

[37] **Akpınar, H**. (2000). Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği, *İstanbul Üniv. İşletme Fakültesi Dergisi*, C:29, S: 1, Nisan.

[38] **Elmas, Ç.** (2003). Yapay Sinir Ağları (Kuram, Mimari, Eğitim, Uygulama), *Seçkin Yayıncılık,* Ankara.

[39] **Karaboğa, D.** (2004). Yapay Zeka Optimizasyon Algoritmaları, *Atlas Yayın Dağıtım*, İstanbul, s. 78.

[40] **Bilekdemir, G.** (2010). *Veri Madenciliği Tekniklerini Kullanarak Üretim Süresi Tahmini Ve Bir Uygulama.* (Yüksek Lisans Tezi). Dokuz Eylül Üniversitesi, Sosyal Bilimler Enstitüsü, İzmir.

[41] **Karypis, G., Han, E., Kumar, V**. (1999). Chameleon: Hierarchical Clustering Using Dynamic Modeling, *IEEE Computer*, pp: 68-75.

[42] **Ramkumar, G.D., Swami, A.** (1998). Clustering Data Without Distance Functions, *IEEE Bulletin ofthe Technical Committee on Data Engineering,* Vol.21, No.l.9-14.

[43] **Seidman, C.** (2001). Data Mining with Microsoft SQL Server 2000. *Microsoft Press,* 1 st Ed., Washington, USA.

[44] **Madhulatha, T.S.** (2012). *An overview on clustering methods.* Retrieved from https://arxiv.org/abs/1205.1117

[45] **Dutt, A., İsmail, M.A., Herawan, T.** (2017). A Systematic Review on Educational Data Mining, *IEEE*, volume:5. pp. 15991-16005.

[46] **Jain, A.K. and Dubes, R.C.** (1988). Algorithms for Clustering Data, *Prentice Hall*, Englewood Cliffs, NJ, USA.

[47] **Sagiroglu, S. and Sinanc, D.** (2013). Big data: A review. *Proceedings. Int. Conf. Collaboration Technol. Syst. (CTS*). pp. 42–47.

[48] **Zaıane, O.R. and Luo, J.** (2001). Web usage mining for a better Webbased learning environment, *Proceedings. Conf. Adv. Technol. Edu.* pp. 60–64.

[49] **Novikov., A.V**. (2019). Pyclustering: Data Mining Library, *Journal of Open Source Software,* 4(36), 1230.

[50] **Han J., Fu Y.** (1999). Mining Multiple-Level Association Rules in Large Databases, *IEEE Transactions on Knowledge and Data Engineering*, vol 11, no.5.

[51] **Visa, A.** (2001). Technology of Text Mining, *Machine Learning and Data Mining in Pattern Recognition,* pp. 1-11.

[52] **İlhan, S., Duru, N., Karagöz, Ş. ve Sağır, M.** (2008). Metin Madenciliği ile Soru Cevaplama Sistemi, Kocaeli Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Kocaeli.

[53] **Navathe, S.B. and Ramez, E.** (2000). Data Warehousing and Data Mining, *Fundamentals of Database Systems*, Pearson Education pvt Inc, Singapore, 841-872.

[54] **Karanikas, H. and Manchester, B.T.** (2001). Knowledge Discovery in Text and Text Mining Software, *Centre for Research in Information Management*, UK.

[55] **Fan, W., Wallace, L., Rich, S. and Zhang, Z**. (2005). Tapping into the Power of Text Mining, *Communications of the ACM* 49:76-82.

[56] **Seker, S.E., Mert, C., Al-Naami, K., Ozalp, N., Ayan, U.** (2013). Correlation between the Economy News and Stock Market in Turkey, *International Journal of Business Intelligence and Review (IJBIR)*, vol. 4, is. 4, pp. 1-21.

[57] **Ramanathan, V., Meyyappan, T**. (2013). Survey of Text Mining, *International Conference on Technology and Business and Management*, pp. 508-514.

[58] **Vidya K.A., Aghila, G.** (2020). Text Mining Process, Techniques and Tools: An Overview, *International Journal of Information Technology and Knowledge Management*, Vol. 2, No 2, pp.613-622.

[59] **Gupta, V. and Lehal, G.** (2009). A Survey of Text Mining Techniques and Applications, *Journal Of Emerging Technologies in Web Intelligence.* Vol. 1, No. 1.

[60] **Sagayam, R., Srinivasan, S., Roshini, S.** (2012). A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques, *Internaltional Journal of Computational Engineering Research (ijceronline.com),* Vol.2, Issue.5.

[61] **Seker, S.E.** (2012). Turkish Query Engine on Library Ontology, *IKE12, Internet Knowledge Engineering*, ISBN:1-60132-222-4, Pages:26-33.

[62] **Seker, S.E., Al-Naami, K.** (2013). Sentimental Analysis on Turkish Blogs via Ensemble Classifier, *Proceedings of the 2013 International Conference on Data Mining*, ISBN:1- 60132-239-9, DMIN, pp. 10-16.

[63] **Dang, S., Ahmad, P.H.** (2014). Text Mining: Techniques and Its Application, *IJETI International Journal of Engineering & Technology Innovations*, Vol. 1, Issue 4, ISSN (Online): 2348-0866.

[64] **Jelodar, H., Wang Y., Yuan C., Feng X., Jiang X., Li Y., Zhao L.** (2018). Latent Dirichlet Allocation (LDA) And Topic Modeling: Models, Applications, a survey, *Springer Science+Business Media,* LLC, part of Springer Nature 2018.

[65] **Ponweiser, M.** (2012). Latent Dirichlet Allocation in R. Institute for Statistics and Mathematics, Vienna University of Business and Economics.

[66] **Dang, S. ve Ahmad, P. H.** (2014). Text Mining: Techniques & its Application, *International Journal of Enginerring Technologies & Technology Innnovations*, 1(4):22–25.

[67] **Kaşıkçı, T. ve Gökçen, H**. (2014). Metin Madenciliği İle E-Ticaret Sitelerinin Belirlenmesi, *Bilişim Teknolojileri Dergisi,* 7(1):25-32.

[68] Abbott, D. (2013). Introduction to Text Mining, Virtual Data Intensive Summer School, *Abbot Analytics Inc*, Retrieved from http://www.vscse.org

[69] **Feldman, R., Sanger, J**. (2007) The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, *Cambridge University Press.*

[70] **Nogueira, B. M., Moura, M. F., Conrado, M. da S., Rossi, R. G., Marcacini, R. M., Rezende, O. S.** (2008). Winning Some of The Document Preprocessing Challenges in A Text Mining Process, *XXIII Simpósio Brasileiro de Banco de Dados (SBBD) - IV Workshop em Algoritmos e Aplicações de Mineração de Dados (WAAMD).*

[71] **Kartal, E., Emre, İ.E., Özen, Z.** (2018). Metin Madenciliği ile Türkçe Bir Dergi Öneri Sisteminin Geliştirilmesi, *20. Akademik Bilişim Konferansı 2018 Bildiriler Kitabı*, Karabük.

[72] **Hofmann, T.** (1999). Probabilistic Latent Semantic Indexing, *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, *ACM*, pp. 50–57.

[73] **Blei, D.M.** (2012). Probabilistic Topic Models, *Communications of the ACM*, 55 (4) 77–84.

[74] **Onan, A., Yalçın, A., Atik, E.** (2020). Üniversite Bilgi Yönetim Sistemi Servis Destek Taleplerinin Konu Modelleme Tabanlı Analizi. *Avrupa Bilim ve Teknoloji Dergisi Özel Sayı*, S. 389-397.

[75] **Koltcov, S., Koltsova, O., Nikolenka, S. (**2014). Latent Dirichlet Allocation: Stability and Applications to Studies of User-Generated Content, *WebSci '14: Proceedings of the 2014 ACM conference on Web science.* Pages 161–165. https://doi.org/10.1145/2615569.2615680.

[76] **Brookes, G., McEnery, T**. (2018). The utility of topic modelling for discourse studies: A critical evaluation, *Discourse Studies 2019*, Vol. 21(1) 3–21.

[77] **Baker, P.** (2006). *Using Corpora in Discourse Analysis.* London: Continuum.

[78] **Murakami, A., Thompson, P., Hunston, S., Vajn, D**. (2017). 'What is this corpus about?' Using topic modelling to explore a specialised corpus, *Corpora* 12(2): 243–277.

[79] **Grimmer, J**. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases, *Political Analysis* 18: 1–35.

[80] **Törnberg, A., Törnberg, P.** (2016). Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum, *Discourse & Society* 27(4): 401–422.

[81] **Brezina, V.** (2018). Statistical choices in corpus-based discourse analysis, *In: Taylor, C, Marchi, A (eds) Corpus Approaches to Discourse: A Critical Review*. London and New York: Routledge, pp. 259–280.

ÖZGEÇMİŞ

FOTOĞRAF

Ad-Soyad **:Mine Bozan**

Doğum Tarihi ve Yeri **:**

E-posta **:**

ÖĞRENİM DURUMU:

- **Lisans** **:** 2016, Eskişehir Osmangazi Üniversitesi, Mühendislik Mimamrlık Fakültesi, Endüstri Mühendisliği Bölümü
- **Yüksek Lisans** **:** 2022, Bursa Teknik Üniversitesi, Akıllı Sistemler Ana Bilim Dalı, Akıllı Sistemler Mühendisliği

MESLEKİ DENEYİM VE ÖDÜLLER:

- ……..
- ……..

TEZDEN TÜRETİLEN ESERLER, SUNUMLAR VE PATENTLER:

- Bozan, M., Altun, K., Konu modelleme ile çalışan önerileri madenciliği: Bir otomotiv endüstrisi vakası, Ulusal bir dergide hakemlik süreci devam eden makale, 2022.
- Bozan, M., Altun, K., Real-time process control in robotic gas metal arc welding: An automotive industry case, IES'20 International IDU Engineering Symposium, Izmir Democracy University, Izmir, Turkey, December 5-6, 10-13, 2020.
- ……

DİĞER ESERLER, SUNUMLAR VE PATENTLER:

- …….
- ……..