Int. J. Pure Appl. Sci. 9(1);20-28 (2023)

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1119552

Konu Modelleme ile Çalışan Önerileri Madenciliği: Bir Otomotiv Endüstrisi Vakası

Mine Bozan[1], Koray Altun[2]*

[1]Bursa Teknik Üniversitesi, Lisansüstü Eğitim Enstitüsü, Akıllı Sistemler Mühendisliği, Bursa, Türkiye
[2]Bursa Teknik Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Endüstri Mühendisliği Bölümü, Bursa, Türkiye
*mineebozan@gmail.com, *koray.altun@btu.edu.tr*
Makale gönderme tarihi:21.05.2022, Makale kabul tarihi: 16.11.2022

Öz

Otomotiv endüstrisindeki yoğun rekabet, sürekli iyileştirme kültürünü zorunlu hale getirmektedir. Çalışan önerileri ve öneri sistemleri bu kültürün önemli bileşenleridir. 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 iyileştirmelerle ilgili tahminler yapabilmeyi olanaklı hale getirebilecektir. Bu çalışmada, bir otomotiv firmasının çalışan önerilerinin analizi için, konu modellemeye ait yöntemlerden, "Gizli Dirichlet Ayrımı (GDA)" kullanılmıştır. En çok verilen öneri çeşidi, "getirisi olmayan olumlu" önerilerdir. Bu öneriler, genellikle iş sağlığı ve güvenliği ile ilgilidir. İkinci sıradaki en çok verilen öneriler ise "öneri", firmaya kazanç sağlayan önerilerdir. Üçüncü sırada, "öneriden hızlı kaizene" yani kısa sürede sonuç alınabilen, getirisi yüksek öneriler bulunmaktadır. Dördüncü sırada, "değerlendirilmek üzere havale" edilen öneriler bulunurken, en az verilen öneri türünün ise "devreye alınmayacak öneriler" olduğu değerlendirilmiştir.

Anahtar Kelimeler: Çalışan önerileri, gizli dirichlet ayrımı, konu modelleme

Mining the Employee Suggestions through Topic Modeling: An Automotive Industry Case

Abstract

Intense competition in the automotive industry necessitates the continuous improvement culture. Employee suggestions and suggestion systems are important components of this culture. The fact that content of the suggestion systems consists of texts has made them suitable data sets for advanced text mining studies. Analyzing employee suggestions with topic modeling will make it possible to assess which subjects are most often received, which subjects should be focused on, and to make predictions about the future improvements. In this study, Latent Dirichlet Allocation (LDA), one of the topic modeling methods, was used for the analysis of the employee suggestions of an automotive company. The most common type of suggestion is "positive suggestions with no return". These recommendations are generally related to occupational health and safety. The second most frequently given suggestions are "suggestions", those are providing profit to the company. In the 3rd rank, there are "fast kaizen from suggestion", which are high-yielding suggestions that can be achieved in a short time. While the fourth rank most common suggestions are "referred to be evaluated", "recommendations that will not be put into action" suggestions are the least given type.

Keywords: Employee suggestions, latent dirichlet allocation, topic modeling

GİRİŞ

Öneri sistemleri, işletmenin her kademesinde çalışan mavi ve beyaz yaka personellerin katkılarıyla uygulanmaktadır. Çalışan öneri sistemi yenilikçi fikirlerin ortaya çıkmasını, uygulanmasını ve firmanın sürekli iyileştirilmesini sağlamaktadır. Öneriler, genellikle metinlerden oluşmaktadır ve gün geçtikçe sayıları artmaktadır. Bu nedenle, arama, anlama ve işleme gibi süreçler için ileri veri analitiği araçlarının kullanımına ihtiyaç duyulur aşamaya erişilmiştir.

Konu modelleme, büyük miktardaki metin kaynaklarından anlamlı bilgilere erişebilmek için

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1119552

uygulanan bir makine öğrenmesi yöntemidir. Doküman kümesini en iyi şekilde karakterize eden sözcük grupları ile benzer ifadeleri otomatik olarak kümeleyebilmektedir. Konu modellemede metinde sıkça birlikte görülen kelimeler kümelenerek soyut konular üretilir ve ilgili metinler içerdikleri kelimelere göre kendisine en yakın olan bir veya daha fazla kümeye atanır.

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, otomotiv sektöründe faaliyet gösteren bir şirketin, 2844 satır metinden oluşan 2021 yılına ait çalışan önerileri, firmanın öneri sistemi üzerinden temin edilmiş ve konu modelleme yöntemiyle analiz edilmiştir. Bu çalışma, bir kurumun gerçek çalışan önerileri veri setini, konu modelleme yöntemi kullanarak analiz eden, literatürdeki ilk çalışmadır.

LİTERATÜR TARAMASI

Konu modelleme, bir metin belgesinde "konu" (topic) 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.

Literatürde, araştırmacılar tarafından geliştirilen mevcut pek çok konu modelleme yöntemi bulunmaktadır.

Vayansky ve Kumar (2020) konu modelleme yöntemleri üzerine kapsamlı ve güncel bir literatür taraması sunmaktadır. Vayansky ve Kumar (2020) çalışmalarında bir karar ağacı da sunmaktadır. Bu karar ağacı, hangi konu modelleme yönteminin kullanılması gerektiği hususunda, yol gösterici bir rehber niteliğindedir.

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

GDA'nın ana fikri, konuların sabit bir kelime sözlüğünden olasılık dağılımını içerdiği ve belgelerin gelişigüzel bir şekilde birleşmiş gizli konulardan oluştuğudur. Bu temel fikir, GDA'nın belge koleksiyonunda bulunan konuları, kelimelerin konuların içindeki olasılıklarını, belgeyi oluşturan kelimelerin atandığı konuları ve bu belge içindeki konuların nasıl dağıldığını öğrenerek ortaya çıkardığını belirtir (Agrawal vd., 2018). Algoritmanın çıktısı, modellenen belge için her konunun kapsamını içeren bir vektördür. Uygun bir şekilde karşılaştırılırsa, bu vektörler, külliyatın "topikal" özellikleri hakkında fikir verebilir.

Konu modelleme ve GDA çalışmaları literatürde yoğun ilgi görmektedir. Bu nedenle, konu özelinde kapsamlı literatür çalışmalarına rastlamak mümkündür. Kapsamlı bir GDA uygulamaları literatür taraması Jelodar vd. (2019) tarafından gerçekleştirilmiştir. Konu modelleme ile ilgili kapsamlı bir diğer çalışma ise Kherwa ve Bansal (2018) tarafından gerçekleştirilmiştir.

Literatürde "konu modelleme" ve "çalışan (employee)" anahtar kelimeleri ile ulaşılabilen sınırlı sayıda çalışma bulunmaktadır. Symitsi vd. (2021), "Glassdoor" isimli kariyer sitesindeki çalışan yorumlarını (employee online reviews), konu modelleme yöntemi ile analiz etmiştir. Çalışanların firmaları ile ilgili pozitif ve negatif geri beslemelerinin modellenmesini konu edinmektedir. Bir diğer çalışmada (Schmiedel vd., 2021), Fortune-500 şirketleri çalışanlarının "Glassdoor" yorumları analiz edilmiştir. Konu modelleme yöntemi ile çalışanların kurum kültürü algısı değerlendirilmiştir. Karkhanis vd. (2022), benzer şekilde, "Glassdoor" verilerini konu modelleme yöntemi ile analiz etmiştir. Firma çalışanlarının paylaşımları kullanılarak "işveren markası" ile ilgili çıkarımda bulunulmuştur. Glassdoor verilerinin konu modelleme yöntemi ile analiz edildiği bir diğer çalışma ise Wang vd. (2022). İlgili çalışmada, firmaların çeşitlilik, eşitlik ve kapsayıcılık açısından bir değerlendirmesi sunulmuştur.

Kurum içerisindeki yönetsel faaliyetleri destekleme amacıyla da konu modelleme yöntemini kullanan çalışmalar literatürde yer almaktadır. Pröllochs ve Feuerriegel (2020), kurumsal stratejik planlama faaliyetlerini destekleme amacıyla konu modelleme yöntemini kullanmıştır. İlgili çalışma, konu modelleme yöntemi ile bir risk iyimserliği analizini bilgisayarlı bir prosedüre dönüştürmüştür.

Konu modelleme yöntemlerinin otomotiv endüstrisindeki uygulamaları da oldukça sınırlıdır. Rhoden vd. (2022), çalışmalarında, otomotiv

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1119552

endüstrisinin sürdürülebilirliği ile ilgili raporların bir değerlendirmesini sunmuştur.

Bu çalışmada, bir otomotiv firmasının "çalışan önerileri" konu modelleme yöntemi ile analiz edilmektedir. Yazarların en iyi bilgisine göre, literatürde ilk defa, "çalışan önerileri" bir konu modelleme yöntemi ile analiz edilmektedir. Bu yönüyle, bu çalışma, literatüre yeni bir konu modelleme, GDA kullanımı vakası sunmaktadır. Benzer çalışmaların endüstri kullanımlarının yaygınlaşması amacıyla, özellikle, bir rehber formatında hazırlanmaya çalışılmıştır.

MATERYAL VE METOT

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

Düz yazı içeren metin verileri yapısal olmayan doğal dil formundaki verilerdir. Kurumsal öneri sistemleri vasıtasıyla toplanan çalışan önerileri de bu formasyondaki verilerdir. Metin verilerinden anlamlı değerlendirmelerin otomatik olarak elde edilmesi metin analizi yöntemleri ile mümkündür. Konu modelleme de bu yöntemlerden biridir. Konu modelleme yöntemlerinden en yaygın kullanılanı GDA'dır. Bu çalışmada çalışan önerilerinin analizi için GDA yöntemi tercih edilmiştir.

GDA, Blei vd. (2003) tarafından önerilmiş bir konu modelleme yaklaşımıdır. Metinlerden oluşan dokümanlarda bulunan gizli anlamsal yapıları olasılık temelli modelleme yaklaşımı ile ortaya çıkarmaya çalışmaktadır. Kelime torbası yaklaşımına dayalı olarak çalışan GDA'da, kelimelerin doküman içerisindeki yerleşimi dikkate alınmazken, kelimelerin birlikte bulunma durumu dikkate alınmaktadır (Altıntaş vd., 2021).

Bir GDA analizi için öncelikle, metinlerden oluşan veri setinin analize uygun hale getirilmesi gerekmektedir. Bu ön işlemler; tokenizasyon, durdurma sözcükleri (stop words) filtreleme ve kök bulma işlemlerinden oluşmaktadır. Bu ön işlemler sonrasında veri seti doğal dil işleme yöntemleri ile metin madenciliği algoritmalarını kullanmaya uygun hale getirilmiş olur.

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 GDA tematik modeli için girdi matrisidir. GDA konu modelleri bir veri ambarı kullanılarak oluşturulur. Daha sonra oluşturulan konu modelleri karşılaştırma ve görselleştirme yoluyla analiz edilir. Konu bazlı modelde aşağıdaki süreç takip edilmiştir:

• Toplanacak metinlerin belirlenmesi: Bu çalışmada, firma çalışanlarının son 1 yılda verdiği öneriler kurumun öneri sisteminden çekilmiştir.

• 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. Stop-word 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 göz önünde 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 "Korpus"a (külliyat) dönüştürülür. Kelime frekans eşleşmesinin ilk indeksinin (0,1) anlamı ilk metin verisindeki kelime indeksi 0 olan sözcüğün tekrarlanma sayısıdır. GDA modeli, frekans ağırlıkları belirlenmiş kelimeler ve

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1119552

belirtilen konu sayısı parametresiyle oluşturulur (Onan vd., 2020).

Veri Seti

Projede kullanılan veriler 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)

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-1'de gösterilmiştir.

"Azure Machine Learning" kullanımı bu aşamaları gerçekleştirebilmek için tercih edilmiştir. Literatürde de benzer farklı uygulamaları yer almaktadır (bkz. Alrumayyan vd., 2018; Khaleg ve Ra, 2019; Balasubramanian vd., 2021).

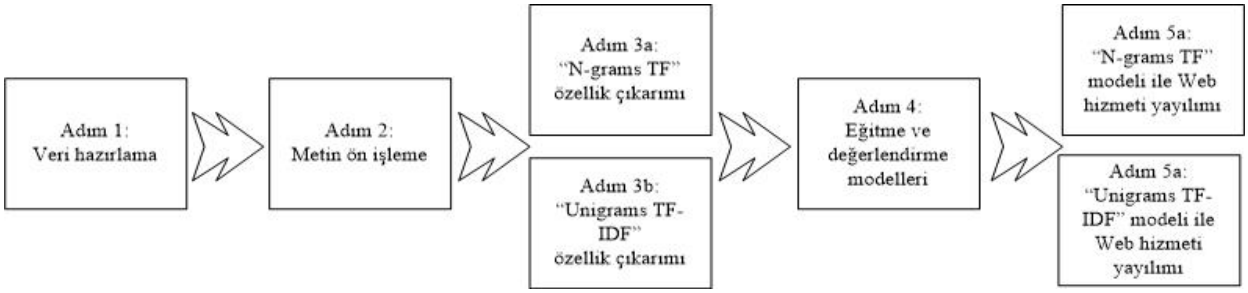
1-4 adımlar, metin sınıflandırma modeli eğitim aşamasını temsil eder. Bu aşamada, metin örnekleri Azure ML deneyine yüklenir, metin temizlenir ve filtrelenir. Temizlenen metinden farklı türde sayısal öznitelikler çıkarılır ve modeller farklı öznitelik türleri üzerinde eğitilir. Son olarak, eğitilen modellerin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

Uygulama

"Microsoft Azure ML" programı üzerinde veri ön işleme aşamalarının akışı Şekil-2'de gösterilmiştir. Metinler genellikle analiz edilmeden önce bir miktar ön işleme gerektirir. Bu aşama, 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.

Özel karakterleri boşluklarla değiştirmek gerekiyorsa parametre "replace_special chars" TRUE olarak ayarlanır. Parametre "remove_duplicate_chars" için TRUE ayarlanır. Sayıların boşluklarla değiştirilmesi gerekiyorsa parametre "replace_numbers" TRUE olarak ayarlanır. Bazı metin sınıflandırma görevleri için, eğitim için ayırt edici özellikler olarak sayılar kullanılmalıdır. Metni küçük harfe dönüştürmek gerekiyorsa parametre "convert_to_lower_case" TRUE olarak ayarlanır. Önceden tanımlanmış bir ortak sözcük listesi kullanarak metinden durdurma sözcüklerini kaldırmak gerekirse, parametre "remove_default_stopWords", TRUE olarak ayarlanır.



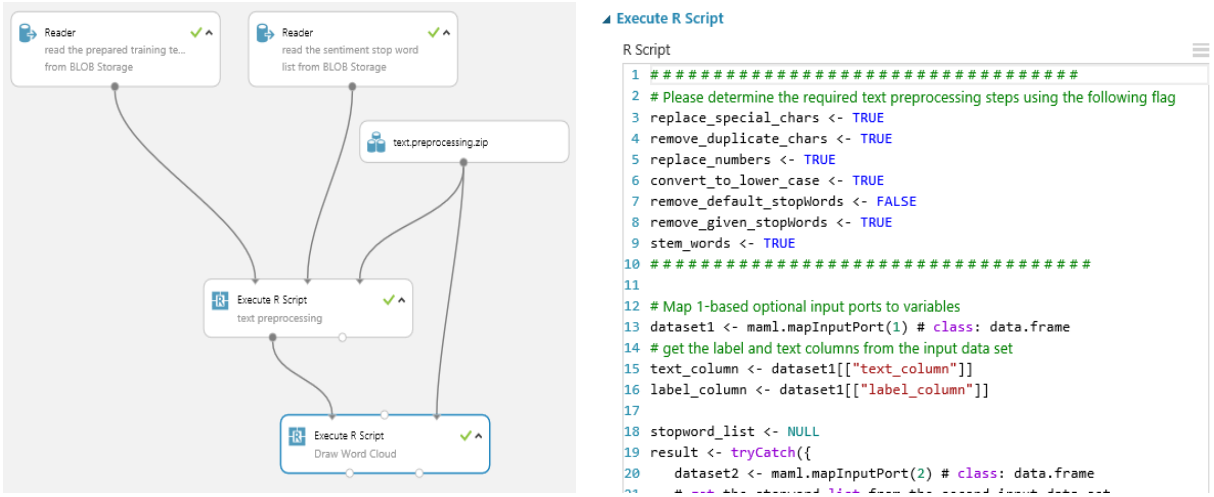
Şekil 1. Metin analizi aşamaları

1

Int. J. Pure Appl. Sci. 9(1);20-28 (2023)

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1119552



(a) **(b)**

Şekil 2. Microsoft Azure'da veri işleme aşamaları (a) ve R kodlamaları (b)

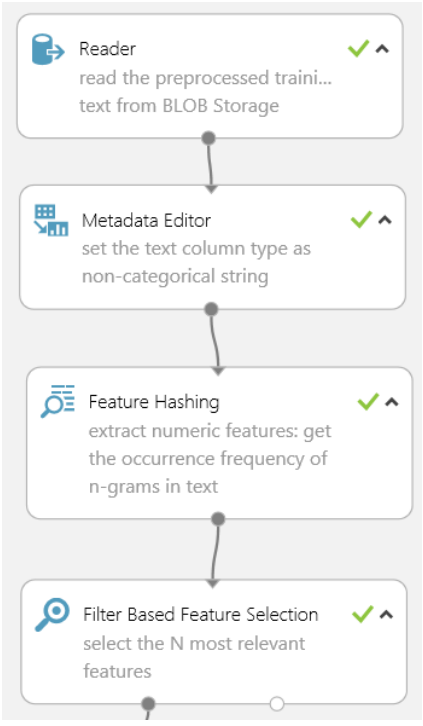
Tanımlanmış ortak sözcükler listesini kullanarak metinden durdurma sözcüklerini kaldırmak gerekirse, parametre "remove_given_stopWords" TRUE olarak ayarlanır. Durdurma sözcüklerinin uygulamaya bağlı olduğu unutulmamalıdır. Yani bir kelime, bir uygulama için sık görülen, ayrım yapmayan bir özellik olarak 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. 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.

Microsoft Azure programı üzerinde özellik seçimi aşamalarının akışı Şekil-3'te gösterilmiştir.

Eğitimli bir modelin sınıflandırma süresi ve karmaşıklığı, özniteliklerin sayısına 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.

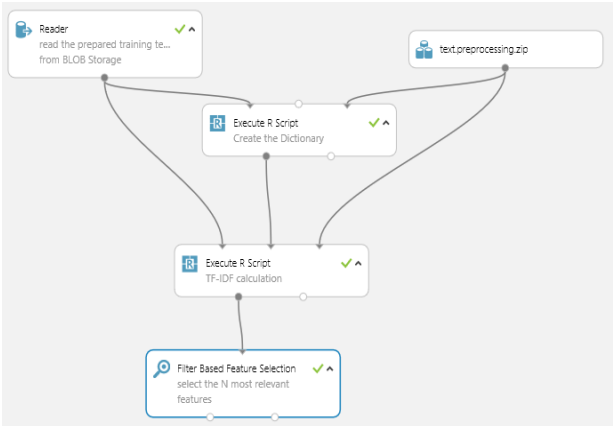
Ayıklanan karma özelliklerin kapsamlı listesinden daha kompakt bir özellik alt kümesi seçmek için "Filtre Tabanlı Özellik Seçimi" kullanıldı. Amaç, "boyutsallık" etkilerinden kaçınmak ve sınıflandırma doğruluğuna zarar vermeden hesaplama karmaşıklığını azaltmaktır. 2^15 ayıklanan özelliklerden duygu etiketine göre en alakalı ilk 2.000 özelliği elde etmek için, "hash" özelliklerini azalan düzende sıralamak için Ki-kare puanı işlevi kullanıldı.



Şekil 3. Özellik seçimi

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1119552



Şekil 4. Özellik çıkarımı akışı

Unigrams TF-IDF özellik çıkarma işlemi (Adım 3b) için ilk olarak, metin modelini eğitmek için kullanılacak unigram (kelime) seti çıkarılır. Unigramlara ek olarak, metin külliyatında 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-4'te gösterilmiştir.

Bir belgede metrik sözcük oluşum sıklığı (TF) bir özellik değeri olarak kullanıldığında, bir korpusda (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, eğitim külliyatındaki belge sayısının verilen kelimeyi içeren belge sayısına oranının günlüğü olarak değerlendirilir. Bu sayıları bir metrikte (TF/IDF) birleştirmek, belgede sık görülen ancak bütüncede nadir bulunan sözcüklere daha fazla önem verir.

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

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

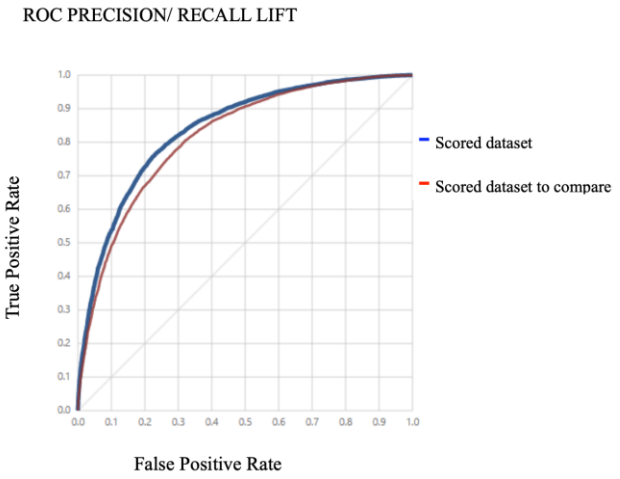
Modellerin eğitimi ve değerlendirilmesi işlemi şu şekilde ilerlemiştir; Verileri iki alt kümeye bölmek için ilk Bölme modülü kullanılmıştır. İlk alt küme modeli eğitmek için kullanılmıştır ve ikinci alt küme, bir sonraki adımda geliştirme/doğrulama kümesi ve test kümesine bölünmektedir.

Veriler sırasıyla %70 ve %30 olarak ayrılmıştır. İkinci alt küme eğitilen modelin performansını değerlendirmek için test kümesi olarak kullanılmıştır. %30 veri örneği ikiye bölünmüştür. Geliştirme setinin ve test setinin her biri girdi verilerinin %15'ini temsil etmektedir.

"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 ayarlanmıştır.

Her bir parametrenin olası tüm değerlerini keşfetmek için bir parametre süpürme modu olarak "tüm ızgara" seçeneği kullanılmıştır.

Şekil-5'te "Alıcı İşletim Karakteristik Eğrisi" (ROC - Receiver Operating Characteristic Curve) verilmiştir. 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.

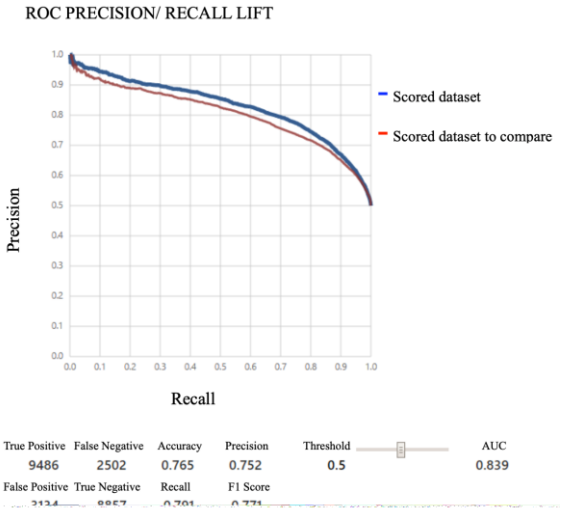


Şekil 5. ROC eğrisi

Int. J. Pure Appl. Sci. 9(1);20-28 (2023)

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

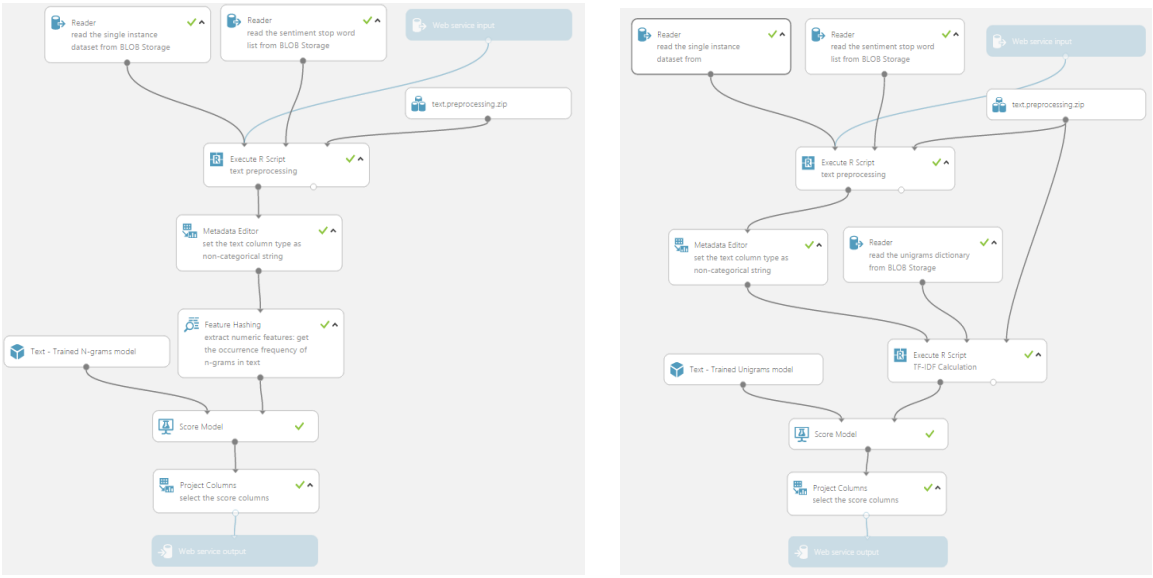
Research article/Araştırma makalesi
DOI:10.29132/ijpas.1119552



Şekil 6. Hassasiyet/Geri çağırma eğrisi

Şekil-6'da Hassasiyet / Geri çağırma (Precision/Recall) grafiği verilmiştir. 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.

Eğitimli modellerin web hizmetleri olarak dağıtım akışı ve çıktısı Şekil-7'de gösterilmektedir. Web hizmeti iki mod ile türetilebilir: RRS (istek-yanıt hizmeti) ve BES (toplu yürütme hizmeti). Adım 4'te eğitilmiş N-gram TF metin modelini bir web hizmeti olarak dağıtılır. Web servis giriş ve çıkış noktaları, özel Web Servis modülleri kullanılarak tanımlanır. Web hizmeti giriş modülü, deneydeki giriş verilerinin gireceği düğüme eklenir. Yürütme R Senaryo modülünde, Adım 2.2'de tanımlanan aynı parametreler kullanılarak gerekli metin ön işleme adımları belirtilir. Özellik "hashing" modülünde, "hashing bitsize" parametresi kullanılarak aynı sayıda bit ve Adım 3a'da tanımlanan parametre kullanılarak aynı n-gram boyutu belirtilir. Web hizmeti giriş noktası ayarlanır. Web hizmeti çıkış noktası ayarlanır. Deneme tuvalinin altındaki "Web Hizmetini Yayınla" seçilir. "Unigrams TF-IDF" tarafından eğitilmiş model bir web hizmeti olarak yayınlanarak kullanıma alma işlemi tamamlanmış olur.



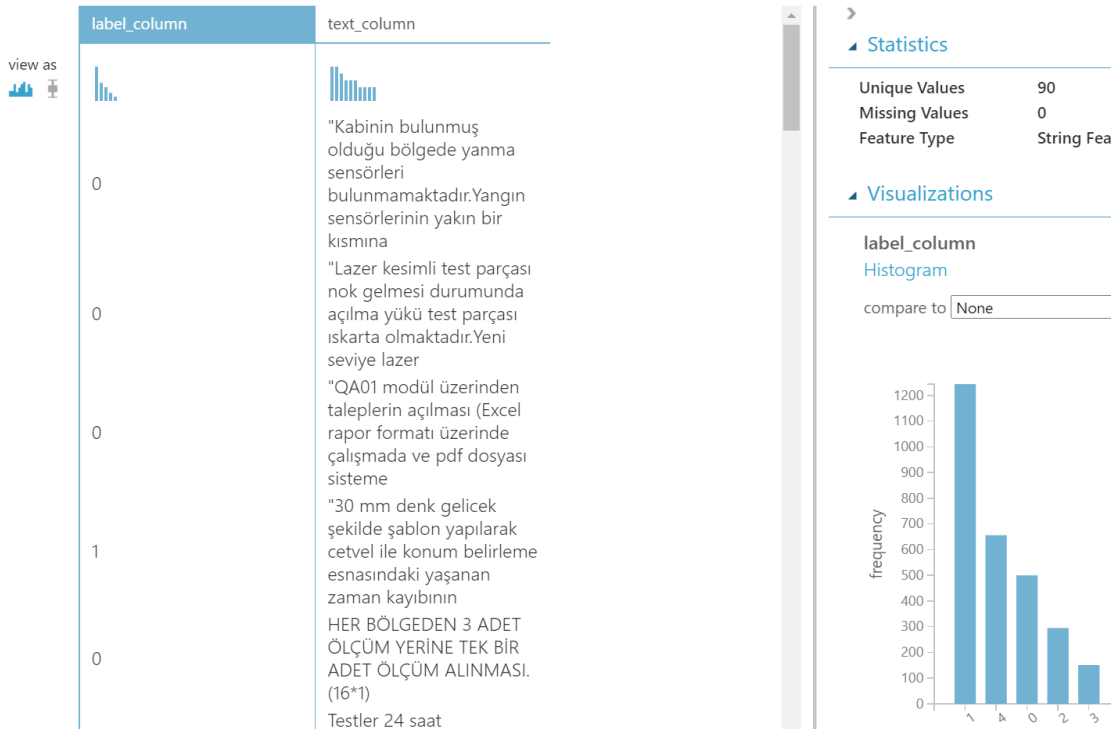
(a) (b)

Şekil 7. Web hizmeti akışı (a) ve çıktısı (b)

Int. J. Pure Appl. Sci. 9(1);20-28 (2023)

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1119552



Şekil 8. "Duygu Etiketi" Histogram Grafiği

TARTIŞMA VE SONUÇ

Eğitim için gerekli veriler yüklenip analiz tamamlandıktan sonra Şekil-8'de 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

Histogram grafiğinde görüldüğü gibi; en çok verilen öneri çeşidi "getirisi olmayan olumlu öneri" dir. Bu öneriler genellikle iş sağlığı ve güvenliği ile ilgili olmaktadır. İkinci sıradaki en çok verilen öneriler ise "öneri" yani getirisi olan, firmaya kazanç sağlayan önerilerdir. Üçüncü sırada "öneriden hızlı kaizene" yani çok kısa sürede sonuç alınabilen, getirisi yüksek öneriler bulunmaktadır. Dördüncü sırada "değerlendirilmek üzere havale" edilen öneriler bulunurken, en az verilen öneri türünün ise "devreye alınmayacak öneriler" olduğu görünmektedir.

Günümüzde, çalışan önerileri, inovasyon sistemlerinin kritik bileşenlerinden biri olmuştur. Çalışanlar, önerileri ile kurumsal inovasyon sistemlerine önemli katkılar sunabilmektedir. Yüksek katılım oranları ise genellikle performans göstergesi olarak takip edilmektedir. Zamanla miktarı artan çalışan önerilerinin değerlendirilmesi ve seçimi de zorlaşmaktadır.

Bu aşamada, makine öğrenmesi uygulamaları önemli katkılar sunma potansiyeline sahiptir. Yapısı ve kurumlar açısından önemi itibariyle, "çalışan önerileri", makine öğrenmesi uygulamaları için oldukça uygun veri setidir ve yüksek katma değer sağlayacak niteliktedir.

Çalışan önerileri, maliyet avantajı ile rekabet ortamında oldukça önem arz etmektedir. Bu çalışmada otomotiv endüstrisinden bir uygulama sunulmuştur.

Yıl içerisinde sisteme gelen çalışan önerilerinin sayıları binleri bulmaktadır. Çalışan önerilerinin konu modelleme yöntemleri ile değerlendirilmesi bir zorunluluk haline gelmektedir. Farklı endüstri uygulamaları, gelecek çalışmaların konusu olabilir.

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1119552

ÇIKAR ÇATIŞMASI BEYANI

Yazarlar bu makale ile ilgili herhangi bir çıkar çatışması bildirmemektedir.

ARAŞTIRMA VE YAYIN ETİĞİ BEYANI

Yazarlar bu çalışmanın araştırma ve yayın etiğine uygun olduğunu beyan eder.

KAYNAKLAR

Agrawal, A., Fu, W. ve 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.

Alrumayyan N., Bawazeer S., AlJurayyad R. ve Al-Razgan M. (2018). Analyzing User Behaviors: A Study of Tips in Foursquare. In: Alenezi M., Qureshi B. (eds) 5th International Symposium on Data Mining Applications. Advances in Intelligent Systems and Computing, vol 753. Springer, Cham.

Altıntaş, V., Albayrak, M. ve Topal, K. (2021). Kanser hastalığı paylaşımları için Dirichlet ayrımı ile gizli konu modelleme. Journal of the Faculty of Engineering and Architecture of Gazi University, 36 (4), 2183-2196.

Balasubramanian S., Kaitheri S., Nanath K., Sreejith S. ve Paris C.M. (2021). Examining Post COVID-19 Tourist Concerns Using Sentiment Analysis and Topic Modeling. In: Wörndl W., Koo C., Stienmetz J.L. (eds) Information and Communication Technologies in Tourism 2021. Springer, Cham.

Blei, D.M., Ng, A.Y. ve Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. ve Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed Tools Appl, 78, 15169–15211.

Karkhanis, G. V., Chandnani, S. U., ve Chakraborti, S. (2022). Analysis of employee perception of employer brand: A comparative study across business cycles using structural topic modelling. Journal of Business Analytics, 1-17.

Khaleq A. A. ve Ra I. (2019). Twitter Analytics for Disaster Relevance and Disaster Phase Discovery. In: Arai K., Bhatia R., Kapoor S. (eds) Proceedings of the Future Technologies Conference (FTC) 2018. FTC 2018. Advances in Intelligent Systems and Computing, vol 880. Springer, Cham.

Kherwa, P. ve Bansal, P. (2018). Topic Modeling: A Comprehensive Review. ICST Transactions on Scalable Information Systems, 159623.

Onan, A., Yalçın, A. ve Atik, E. (2020). Üniversite bilgi yönetim sistemi servis destek taleplerinin konu modelleme tabanlı analizi. Avrupa Bilim ve Teknoloji Dergisi Özel Sayı, 389-397.

Pröllochs, N. ve Feuerriegel, S. (2020). Business analytics for strategic management: Identifying and assessing corporate challenges via topic modeling. Information & Management, 57, 103070.

Rhoden, I., Ball, C. S., Vögele, S. Ve Kuckshinrichs, W. (2022). Minding the gap-relating disclosure to contexts of sustainability reporting in the automotive industry. Corporate Social Responsibility and Environmental Management, 1-12.

Schimiedel, T., Müller, O. ve Brocke, J. V. (2019). Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture. Organizational Research Methods, 22 (4), 941-968.

Symitsi, E., Stamolampros, P., Daskalakis, G. ve Korfiatis, N. (2021). The informational value of employee online reviews. European Journal of Operational Research, 288, 605–619.

Vayansky, I. ve Kumar, S. A. P. (2020). A review of topic modeling methods. Information Systems, 94, 101582.

Wang, W., Dinh, J., Jones, K. S., Upadhyay, S. ve Yang, J. (2022). Machine learning text analysis of corporate diversity statements predicts employees' online ratings. Academy of Management Proceedings, 15107.