Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Araştırma Makalesi

Farklı Vektörleştirme ve Ön işlem Yöntemleri ile Talep Sınıflandırma

Halil ARSLAN[a], İbrahim Ethem DADAŞ[b], Yunus Emre IŞIK[c*]

[a] *Bilgisayar Mühendisliği Bölümü, Mühendislik Fakültesi, Sivas Cumhuriyet Üniversitesi, Sivas, TÜRKİYE*
[b]*Detay Danışmanlık Bil. Hiz. San. Dış Tic. A.Ş, İstanbul, TÜRKİYE*
[c]*Yönetim Bilişim Sistemleri Bölümü, İktisadi ve İdari Bilimler Fakültesi, Sivas Cumhuriyet Üniversitesi, Sivas, TÜRKİYE*
** Sorumlu yazarın e-posta adresi: yeisik@cumhuriyet.edu.tr*
DOI: 10.29130/dubited.1017422

Öz

Firmalarda, ihtiyaçlara yönelik gelen taleplerin doğru şekilde işlenmesi hem iş sürecini hızlandırır hem de ortaya çıkabilecek sorunları bertaraf eder. Geliştirme, destek, sorun çözme gibi farklı konulardaki taleplerin, verimli ve doğru kişilerce çözülmesi için öncelikle ilgili alt departmana yönlendirilmesi gerekir. Yönlendirmeler belirli kişilerce elle gerçekleştirilebilir. Ancak firma büyüklüğüyle doğru orantılı olarak gelen talep sayısının çok olması süreci zorlaştırıp zaman kaybına yol açmaktadır. Özellikle bilişim sektöründe hizmet veren kurumsal firmalarda taleplerin otomatik olarak alt-departmanlara aktarılabilmesi, işin verimliliğinin ciddi şekilde arttırabilir. Bu ihtiyacın giderilmesi içi metni işleyerek içerisinden kolaylıkla bilgi çıkarımını sağlayabilen metin madenciliği ve makine öğrenmesi yöntemleri kullanılabilir. Çalışmamızda, Detaysoft Danışmanlık firmasına ait gelen taleplerin doğru şekilde alt departmana yönlendirilmesini sağlayan bir sistem önerilmiştir. Sistem performansının ölçülebilmesi amacıyla gerçek müşteri taleplerinden oluşan 2103 veri toplanmış ve işaretlenmiştir. Toplanan verilerin varsayımlardan bağımsız olarak doğru şekilde işaretlenmesi için de veriye göre sınıf etiketlerinin belirlendiği temellendirilmiş teoriden faydalanılmıştır. Ham metinlerin vektörleştirilmesi için kelime çantası ve türevlerinin (TF, TFIDF) yanı sıra GloVe ve Word2Vec gibi kelime gömme yöntemleri de denenmiş ve hangi vektörleştirme yönteminin daha başarılı olduğu irdelenmiştir. Ayrıca gereksiz kelimelerin ve sadece kelime köklerinin kullanılmasının talep sınıflandırmaya etkileri analiz edilmiştir. Yapılan analizler sonucunda SVM algoritmasını kullanan modellerin %79 gibi iyi sayılabilecek bir başarım ile gelen talebi doğru şekilde sınıflandırabildiği gözlemlenmiştir. Elde edilen sonuçların, talep sınıflandırma konularındaki gelecek çalışmalara hem vektörleştirme hem de ön işlem süreçleriyle alakalı ışık tutması beklenmektedir.

Anahtar Kelimeler: *talep sınıflandırma, metin vektörleştirme, makine öğrenmesi, metin madenciliği*

Classification of Support Requests using Different Vectorization and Pre-Processing Methods

Abstract

Processing support requests for needs enables both the acceleration of business processes and the elimination of problems that may arise. To ensure that requests for various topics such as development, support, and problem fixing are handled by the right person, the request should first be routed to the appropriate sub-department. This routing can be done manually by an expert. However, the large number of requests complicates business processes and leads to time loss depending on the size of the company. Therefore, automatic routing of incoming request to the right sub-department can significantly increase work efficiency, especially in companies operating in IT. Text mining and machine learning methods that process information in some form can help meet this need. In our study, a model was proposed to route incoming request belong to Detaysoft to the proper sub-departments. To measure

Geliş: 01/11/2021, Düzeltme: 10/12/2022, Kabul: 26/12/2021

the performance of the system, 2103 data with real customer requests were collected and labeled. Grounded theory was used to correctly label the collected data regardless of the assumptions, where class labels are determined based on the data. For the vectorization of raw texts, Bag of Words such as TF, TF-IDF as well as word embedding methods such as GloVe and Word2Vec were tried to investigate which vectorization method is more successful. Also, the effects of using stop words and word stems on query classification were analyzed. As a result of the analysis, it was observed that the models using SVM algorithm can correctly classify the incoming requests with a performance of 79%, which can be considered reasonable. It is expected that the obtained results will shed light on future studies on request classification in the context of vectorization and preprocessing. It is expected that the results obtained will shed light on future studies on request classification studies related to both vectorization and preprocessing processes.

Keywords: *support ticket classification, text vectorization, machine learning, text mining*

I. GİRİŞ

Gelişen teknoloji ve azalan maliyetler neticesinde birçok işlem elektronik ortamda yürütülmekte ve veri olarak depolanmaktadır. Büyük ölçekte toplanan bu devasa boyuttaki veri yığını arasından anlamlı bilgilerin çıkartılarak kullanılabilir hale getirilmesi, sektörel birçok ihtiyacın çözülmesine de olanak tanır. Firmalar çıkarılan bilgileri kullanarak pazarlama, finans, üretim, süreç yönetimi gibi konularda karar mekanizmasını güçlendirir ve şirket içi verimliliği arttırır. Özellikle bilişim sektöründe faaliyet gösteren firmalar için toplanan verilerin işlenmesi müşteri memnuniyeti, hizmet kalitesinin arttırılması ve dolaylı olarak firmanın sektördeki yerini güçlendirmesini sağlar. Bu şekilde işlendiği taktirde firmaya yüksek katkı sağlayabilecek verilerden birisi de farklı kaynaklardan tamamlanmak üzere gönderilen görev ya da iş talepleridir.

Bilişim sektöründe görev talepleri, genellikle müşterilerin yaşadığı sorunların ya da geliştirme ve destek gibi ihtiyaçların çözülmesine yönelik metinlerden oluşur. Gelen talep yapılacak işin niteliğine göre firma içerisinde yer alan ilgili departmana ya da personele yönlendirilir. Böylece ihtiyacın çözümü sağlanmaktadır. Ancak talepte belirtilen işin hangi personel tarafından ifa edileceğinin belirlenmesi hem personelleri tanıyan hem de talebi anlayabilecek kadar konuya hakim bir uzman tarafından yapılmalıdır. Aksi taktirde talebin çözüm süresi uzayabilir. Bu durum müşteri memnuniyetinin azalmasının yanı sıra firma içi iş akışlarının da kesintiye uğramasına neden olur. Bu tarz problemler müşteri sayısı az olan ve görev talebi trafiğinin yoğun olmadığı ufak ölçekli firmalar için önem arz etmeyebilir. Çünkü ilgili talep belirlenen bir uzman tarafından hızlıca yönlendirilebilir. Ancak eğer firma çok fazla sayıda müşteriye hizmet veriyorsa, gelen taleplerin bir veya birkaç uzman tarafından doğru ve hızlı bir şekilde belirli bir personele atanması pek mümkün gözükmemektedir. Ortaya çıkan bu ihtiyacın otomatik olarak çözümlenebilmesi için gelen talep metnini veri olarak ele alan ve istenilen bilginin çıkartılmasına olanak sağlayan metin madenciliği temelli yaklaşımlardan yararlanılabilir.

Metin madenciliği; istenmeyen e-postaların engellenmesi, sosyal medya paylaşımlarının analizi, toplum fikrinin belirlenmesi, dil tanımlama, personel görev atama ve hatta öz-bildirime göre hastalık tahmini gibi birçok konuda analiz ve yorumlama imkanı sağlamaktadır. Ballı ve Karasoy, telefona gelen kısa mesajların gereksiz olup olmadığına makine öğrenmesi ve metin madenciliği tabanlı bir yaklaşım kullanarak karar vermişlerdir. Mesajın içerdiği kelimeler Word2Vec yöntemi kullanılarak vektörleştirilmiş ve rastgele orman algoritmasıyla test edilmiştir. Elde edilen makine öğrenmesi modeli ile gelen bir mesajın gereksiz olup olmadığı %99'un üzerinde başarıyla tahmin edilmiştir [1].
Benzer bir çalışmada da, elektronik alışveriş sitelerinde yer alan ürün yorumlarının sahte mi yoksa gerçek kullanıcı yorumu mu olduğunu tespit etmek için sistem önerilmiştir. Yorumların vektörleştirilmesi için terim frekansı ve Word2Vec yöntemlerinin ayrı ayrı kullanıldığı çalışmada, evrişimsel ve tekrarlayıcı yapay sinir ağları modelleri ile yorumun durumu tahmin edilmeye çalışılmıştır. İki farklı veri seti ile denenen sistem %94'ün üzerinde doğru tahmin başarısı elde etmiştir [2]. Diğer bir çalışma da ise e-ticaret sitelerinin bulunmasının kolaylaştırılması amacıyla site içerisinde yer alan metinler işlenmiştir. Metnin vektör haline getirilmesinde kelime çantası yönteminden yararlanılırken, karar verme işlemi için k-en yakın komşu ve naive bayes algoritmaları kullanılmıştır.

Yapılan test analizleri sonucunda sistem %85 oranında doğru başarıyla verilen sitenin e-ticaret sitesi olup olmadığını tahmin etmiştir [3].

Metin madenciliği yöntemleri sadece ikili sınıflandırmada değil aynı zamanda çok sınıflı problemlerin çözümünde de kullanılmaktadır. Bouazizi ve Ohtsuki, twitter'da yapılan paylaşımların hangi duyguyu içerdiğini tahmin etmek için metin madenciliği ve makine öğrenmesi yöntemlerinden yararlanmışlardır. Mutluluk, sinirlilik, öfke, umutsuzluk benzeri 7 farklı duygu ile işaretlenen tweetler rastgele orman algoritmasıyla sınıflandırılmıştır. Analizler sonucunda eğlence duygusunun tahmini %40'ta kalırken nefret duygusu içeren metinlerin tahmin başarısı %90ın üzerine çıkmıştır [4]. Arifianto vd., Endonezyada bulunan internet servis sağlayıcısının müşteri yorumlarını kategorize etmek için uzun-kısa süreli bellek bazlı bir sistem önermişlerdir. 4 farklı kategoriden 1107 örneklem ile eğitilen modeller için yorumların vektör haline getirilmesinde Word2Vec yöntemi kullanılmıştır. Yapılan test analizleri sonucunda sistem %88'e yakın bir başarım elde etmiştir [5]. Parmar vd. müşteri ilişkileri yönetimi (CRM) uygulaması üzerinden gelen destek taleplerinin doğru departmana yönlendirilmesi için makine öğrenme temelli bir sistem önermişlerdir. 12 farklı departman için yapılan analizler sonucunda en yüksek başarım %63 ile SVM algoritmasıyla elde edilmiştir [6]. Benzer bir çalışma da üniversite bilgi sistemine gelen hata ve hizmet destek taleplerini içeren metinler sınıflandırılmıştır. Başarımın arttırılması için gereksiz kelime temizleme ve kök alma gibi ön işlemlerinde uygulandığı çalışma da metnin vektörlerle ifade edilebilmesi için TF-IDF yaklaşımı kullanılmıştır. Ayrılan test verilerine yapılan sınıflandırma işlemi sonucunda destek vektör makineleri %92 gibi yüksek bir başarım elde etmiştir [7]. Metin madenciliği yöntemleri sadece özel firmaların değil aynı zamanda kamusal işlemlerin verimliliğini de arttırmaktadır. Boston bölgesindeki vatandaşların ulaşım şikayetlerinin otomatik olarak belirlenmesi için metin madenciliği ve makine öğrenmesi yöntemleri kullanılarak bir model önerilmiştir. Hem denetimli hem de denetimsiz öğrenme yaklaşımlarının kullanıldığı çalışma da talepler %70 oranında başarılı tahmin yapabilmişlerdir [8]. Bilişim sektörüne uygulanan başka bir çalışma da ise kablosuz bağlantı ekipmanlarıyla ilgili gelen teknik destek talepleri otomatik olarak sınıflandırılmıştır. Kullanıcının girdiği destek talebi Word2Vec vektör haline getirildikten sonra evrişimsel ağlar ile model oluşturulmuştur. Yapılan testler sonucunda ekipmanla ilgili sorunun hangi nedenle kaynaklandığı otomatik olarak belirlenmiştir [9].

Metin madenciliği yöntemleri bilişim sektörü dahil farklı alanlarda ortaya çıkan ihtiyaçların çözümlerinde başarıyla uygulanmıştır. Böylece yöntemin yazılım geliştirme veya danışmanlık taleplerinin de doğru şekilde yönlendirilmesinde kullanılabilirliği fikrini doğurmuştur. Bu amaçla ortaya koyduğumuz çalışma, Detaysoft firmasına ait gerçek zamanlı talep verilerinin metin madenciliği ve makine öğrenmesi yöntemleriyle sınıflandırılması süreçlerini içermektedir. Önerilen model ile müşteri tarafından gelen talep metninin madencilik yöntemleriyle işlenmesi, sayısallaştırılması ve doğru alt-departmana yönlendirilebilmesi için sınıflandırılması süreçlerini içerir. Hata payının en aza indirilebilmesi için farklı vektör oluşturma yöntemleri yanı sıra ön işlem yaklaşımları da karşılaştırılmalı olarak sunulmuştur.

II. MATERYAL VE METOD

A. TEMELLENDİRİLMİŞ TEORİ

Nitel araştırma yöntemlerinde, herhangi bir veri grubuyla ilgili ortaya atılacak bir teorinin doğrulanması sürecinde genellikle teoriyi doğrulayacak veriler toplanmaktadır. Ancak bu yaklaşım sadece toplanan veri üzerinden varsayım oluşturması sebebi ile ilgili veri grubunun tamamını doğrulamaz. Veriler hakkında doğru çıkarımların yapılabilmesi için mevcutta olan veriler üzerine teorilerin kurgulanması gereklidir.

Nitel verileri toplayıp analiz ederek sistematik olarak hipotez ve teoriler oluşturmaya dayanan bu yaklaşıma temellendirilmiş teori (TT) denilir. Bu yöntem verilerin özetlenmiş fikir ve kavramlarının ortaya çıkartılmasında da kullanılmaktadır. TT yinelemeli olarak 4 kuralı izlemektedir. Bunlar; veriyi temsil edecek kavramların bulunması, kavramların anahtar kelimelerle ifade edilmesi, ifadelerin

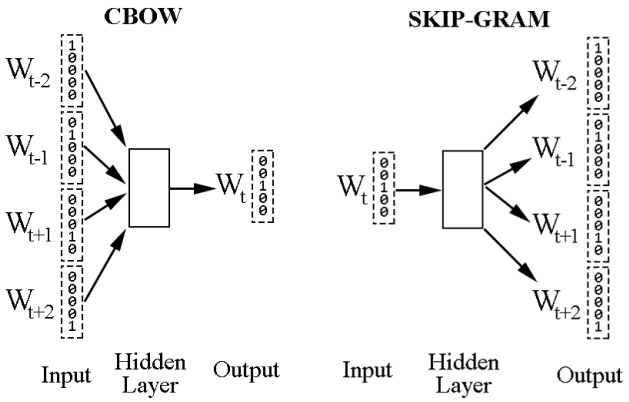
hiyerarşik olarak gruplandırılması ve benzerliklerine göre kategorize edilmesidir. Yinelemeli analiz sonunda kategoriler belirlenmiş olur.

B. TERİM FREKANSI – TERS DOKÜMAN FREKANSI

Metin madenciliği yaklaşımının önemli ön adımlarından birisi metnin algoritmalar tarafından anlayacağı şekilde temsil edilmesidir. Vektörleştirme olarak bilinen yaklaşım ile metinler vektör olarak ifade edilebilir. Böylelikle vektörler arasındaki yakınlık-uzaklık ilişkisine göre birbirlerine olan benzerliğine karar verilebilir.

Literatürde sıklıkla kullanılan vektörleştirme yöntemleri Terim Frekansı (TF) ve Terim Frekansı-Ters doküman Frekansı (TF-IDF) oranıdır. TF yönteminde her bir kelimenin ağırlığı, kelimenin geçme sıklığının dokümanın içerdiği kelime sayısına bölünmesiyle bulunmaktadır [10]. Dolayısıyla eğer kelime doküman içerisinde daha fazla geçiyorsa ayırt edici kabul edilir ve ağırlığı daha yüksek olur. TF-IDF yönteminde ise ağırlık, terim frekansının ters doküman frekansıyla çarpımı sonucunda elde edilir. Ters doküman frekansı (Inverse Document Frequency, IDF) ilgili terimin tüm dokümanlar içerisindeki önemini belirtir [11]. Böylelikle bir kelimenin TF-IDF değeri, kelimenin sadece geçme sıklığına değil aynı zamanda doküman içerisindeki önem derecesine göre hesaplanmış olur.

C. WORD2VEC



Şekil 1. *Word2Vec Öğrenme Modelleri.*

Word2vec, verilen metin içeriğinde anlamsal olarak benzer olan kelimelerin vektör uzayında da birbirlerine yakın olacağı hipotezine dayanan bir kelime temsil yöntemidir [12]. Bazı kelimeler genellikle çiftler halinde kullanılmaktadır. Örneğin hava durumu ile ilgili bir metin içerisinde "hava" kelimesinden sonra "sıcak" ya da "soğuk" kelimelerinin gelmesi muhtemeldir. Bu kelimelerin birlikte geçmesi anlamları hakkında bilgi verebilir. CBOW ve Skip-Gram olarak iki farklı öğrenme algoritmasına sahip olan Word2Vec mimarileri Şekil 1'de gösterilmiştir. Her iki öğrenme yaklaşımı da giriş, projeksiyon ve çıkış katmanlarına sahip yapay sinir ağından oluşur.
CBOW modeli girdi olarak belirli pencerede yer alan kelimeleri alır ve hedef kelimeyi tahmin etmeye çalışır. Böylelikle komşu kelimelerin analizine dayalı olarak kelimenin tahmini sağlanmaktadır. Öte yandan Skip-Gram yaklaşımı ise verilen bir kelimenin öncesi ya da sonrası gelebilecek kelimeleri tahmin etmekte kullanılır. Her iki modelin de projeksiyon katmanı N-boyutlu ve ağırlıkları tutan vektörü içermektedir. Model eğitimine göre vektör içindeki ağırlıklar değişerek optimum kelime tahmini yakalanmaya çalışılır.

D. GLOVE

GloVe (Global Vectors) kelimelerin anlamsal vektörlerini çıkartan bir diğer kelime temsil etme yöntemidir [13]. Word2Vec yöntemi kelimelerin yerel istatistiklerine göre bir vektör oluşturmaktadır. Yani kelimelerin sadece komşularına bakar. GloVe ise farklı olarak kelimelerin tüm metin içeresindeki global istatistiklerini de kullanmaktadır. GloVe yönteminin uygulanması için öncelikle metin içerisinde geçen kelimelerin birlikte geçme matrisi X oluşturulur. X_{ij} j kelimesinin i ile beraber geçme sayısını, X_i i kelimesinin metin içerisinde geçme sayısını, $P_{ij} = \frac{X_{ij}}{X_i}$ j kelimesiyle i kelimesinin beraber geçme olasılığını ve w_i i kelimesini göstermek üzere herhangi farklı bir k kelimesinin i ve j kelimeleriyle beraber geçme olasılığı eşitlik 1 ile hesaplanır.

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}} \quad (1)$$

GloVe yönteminde kelime vektörlerinin öğrenimi için en uygun başlangıç noktasının farklı kelimelerin birlikte meydana gelme olasılıklarına oranlı olması gerektiği öner sürülmektedir. Dolayısıyla fonksiyon içerisinde iki farklı kelimenin uzaklığının, bu iki kelimeyle ilişkili diğer bir kelime olan $\tilde{w}_k$ kullanılmasını öngörmektedir. Bu şekilde kelimeler arasındaki benzerlik hesaplanmaktadır. Ancak binlerce kelimenin geçtiği uzun metinlerde her bir kelime için 3 değişkenli bir hesaplama yapmak pratik olarak mümkün değildir. Bunun yerine matematiksel yöntemler kullanılarak eşitlik 1 eşitlik 3'e evirilmiştir.

$$F\left(\left(w_i - w_j\right)\right)^T * \tilde{w}_k\Big) = F\left({w_i}^T * \tilde{w}_k - {w_j}^T * \tilde{w}_k\right) = \frac{F\left({w_i}^T * \tilde{w}_k\right)}{F\left({w_j}^T * \tilde{w}_k\right)} = \frac{P_{ik}}{P_{jk}} \quad (2)$$

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left({w_i}^T \tilde{w}_j + b_i + \tilde{b}_j - log X_{ij}\right)^2 \quad (3)$$

Burada ise çeşitli matematiksel dönüşümler ile $\tilde{w}_k$ yani ortak kelime devre dışı bırakılarak 2 kelime kullanılarak birbirlerine olan uzaklık hesaplanabilmektedir. V farklı kelime sayısını ifade etmek üzere tüm kelimeler arasında bu işlem gerçekleştirilerek J maliyet fonksiyonunu en aza indirilmektedir.

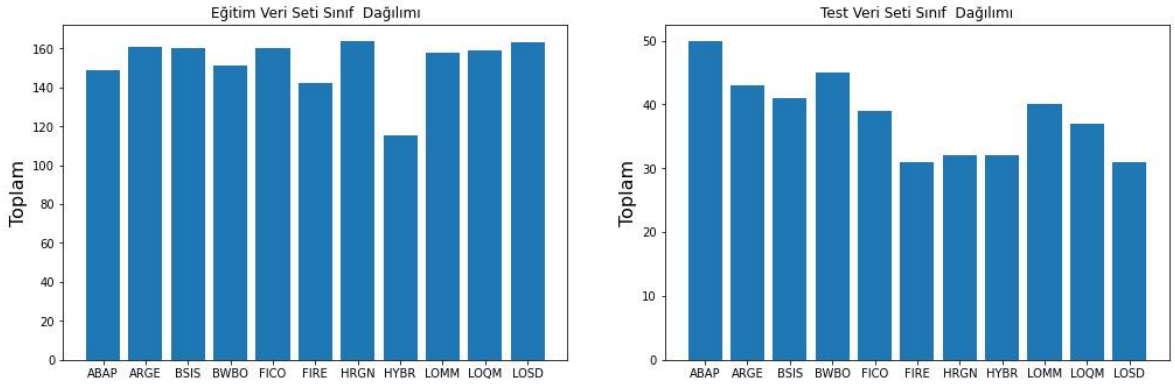
E. VERİSETİ

Çalışma içerisinde Detaysoft Danışmanlık firmasına ait gerçek zamanlı proje/görev isteklerini içeren talepler kullanılmıştır. Veriler için firmaya ait iş taleplerinin ve yapılacakların toplandığı web tabanlı uygulama sistemi kullanılarak elde edilmiştir. Her talep müşteri, yönetici ya da takım lideri tarafından sisteme girilmiş görevlerin başlık ve tanımını içermektedir. 2103 adet görev girdisi bu şekilde toplanarak çalışma içerisinde kullanılacak veri seti belirlenmiştir. Veri setine uygulanan ön işlem adımları "uygulama" bölümünde detaylı olarak anlatılmıştır.

III. UYGULAMA

A. ÖN İŞLEM

Ön işlem süreci metin formatındaki verinin makine öğrenmesi algoritmalarının işleyebileceği vektörlere çevrilmesi sürecini içerir. Veri seti toplanma işlemi bir web uygulaması üzerinden gerçekleştirildiği için metin içerisinde bulunan XML ve HTML etiketleri temizlenmiştir. Metin madenciliği literatüründe, noktalama işaretlerinin kullanımının model başarısı üzerinde olumsuz etkisi olduğu belirtilmektedir [14]. Bu nedenle metinlerde geçen kelime noktalama işaretleri de temizlenmiştir.



Şekil 2. *Eğitim-Test Sınıf dağılımlar.*

Bir sonraki aşama verilerin etiketlenmesi sürecidir. Veri kümesinden bilimsel şekilde en doğru etiket havuzunun çıkartılması için nitel bir araştırma yöntemi olan temellendirilmiş teoriden faydalanılmıştır. Firmanın proje yöneticileri ve takım liderleriyle birlikte temellendirilmiş teori yaklaşımıyla ham veriler gözden geçirilmiş ve geniş kapsamda 11 etiket belirlenmiştir. Etiketlenen veriler rastgele %20'si test geri kalanı eğitim veri seti olacak şekilde ayrılmıştır. Veri setlerinin sınıf dağılımları şekil 2'de gösterilmiştir. Noktalama işaretlerine benzer olarak gereksiz kelimelerin geçmesi, kelime köklerinin kullanılması gibi farklı etmenlerin de metin madenciliğine olumlu veya olumsuz etkileri literatürde raporlanmıştır. Çalışmamızda farklı ön işlem yöntemlerinin talep sınıflandırması üzerine etkileri de incelenmiştir.

Bu amaçla orijinal metinlerden oluşan ham veri seti yanı sıra gereksiz kelimelerin çıkartıldığı, kelime hatalarının giderildiği ve sadece kelime köklerini içeren veri setleri de oluşturulmuştur. Gereksiz kelimelerin temizlenmesi amacıyla Türkçe dilinde var olan anlam taşımayan kelimelerin listesi çıkartılarak verilerden ayıklanmıştır.

Kelime düzgünleştirme ve kök çıkartma süreci için ise bu alanda yüksek doğruluk gösteren ITU NLP Web Servis [15] kullanılmıştır. Düzgünleştirme işlemi ilgili kelimenin bilerek veya bilmeyerek yanlış yazılması durumunda kelimeyi düzeltmeye yarayan bir yaklaşımdır. Örneğin metin içerisinde "dvlet" kelimesi geçiyorsa düzgünleştirme işlemiyle bu kelime "devlet" olarak düzeltilir.Her örneklemde yer alan kelimeler sırayla ilgili servise gönderilerek düzgünleştirilmiş ve daha sonra kökü çıkartılarak saklanmıştır. Son aşama ise metinlerin algoritmaların anlayabileceği şekilde vektörler halinde temsil edilmesidir. TF ve TF-IDF yöntemleri için Python Scikit [16] kütüphanesi kullanılarak her bir talep metni vektöre dönüştürülmüştür. Bunun dışında çalışma içerisinde Word2Vec ve GloVe yöntemleri Python Gensim [17] kütüphanesi kullanılarak denenmiştir.

Kelime temsil yöntemlerinin ana amacı herhangi bir metni vektöre çevirmek değil, metin içeresinde geçen kelimelerin diğer kelimelerle olan uzaklıklarını vektör olarak temsil etmektir. Ancak çalışmamız birden fazla kelimeden oluşan taleplerin sınıflandırılmasını içermektedir. Dolayısıyla her bir talep metninin vektöre dönüştürülmesi için ek işlemlere gerek vardır. Bu aşamada talep vektörlerinin oluşturulması için içerisinde geçen kelimelerin vektör değerleri ortalaması talebin vektör değerini temsil edileceği varsayılmıştır. TF yönteminden farklı olarak kelime temsil yöntemleri arka planda yapay sinir ağı temelli öğrenme algoritması kullanmaktadır. Kelimeler arası vektör uzaklığının optimum şekilde öğrenilebilmesi için öğrenme algoritması içerisinde yer alan parametrelerin optimize edilmesine ihtiyaç vardır. Bu amaçla eğitim veri setinin %20'i doğrulama seti olarak ayrılmış ve parametre optimizasyonunda kullanılmıştır. En uygun parametre uzayı bulunduktan sonra eğitim setinin tamamıyla model tekrar eğitilmiş ve test veri seti içerisinde yer alan kelimelerin vektörleri hesaplanmıştır.

Her bir vektörleştirme yöntemi için herhangi bir ön işleme tabi tutulmamış, gereksiz kelimelerden temizlenmiş, düzgünleştirme işlemi uygulanmış ve metinlerin kökleri alınmış talep verileri ayrı ayrı

saklanmıştır. Dolayısıyla ön işlemler sonucunda analiz edilmeyi bekleyen 16 farklı veri seti oluşturulmuştur. Her veri seti ayrı ayrı makine öğrenmesi modelleriyle eğitilerek test edilmiştir.

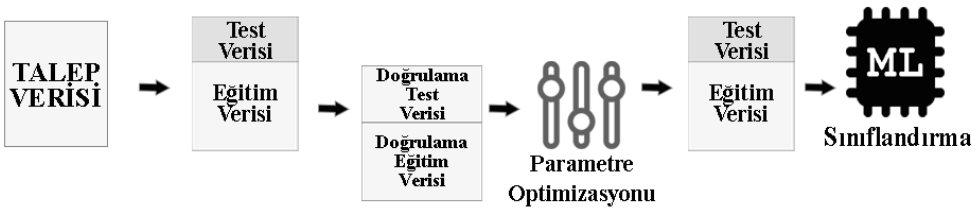
B. SINIFLANDIRMA

İşaretlenen ve vektörleri çıkartılan talep metinlerinin sınıflandırılması için 5 farklı algoritma kullanılmıştır. Bunlar sırasıyla Lojistik Regresyon (LR), Destek Vektör Makineleri (SVM), Rastgele Orman (RF), k-En Yakın Komşu (KNN) ve XGBoost'tur. Öğrenme algoritmalarının parametre optimizasyonu model başarımı üzerinde ciddi bir katkısı sağlamaktadır. Her bir algoritmaya ait en yüksek başarımı veren parametre uzayı Optuna [18] kütüphanesi kullanılarak belirlenmiştir.

Tablo 1. *Parametre Uzayı*

Algoritma	**Parametre Tipi**	**Parametre Uzayı**
LR	Regülürizasyon (C)	$(2^{-12} - 2^{-12})$
	Çözümleyici	Linear, BFGS
SVM	Regülürizasyon (C)	$(2^{-12} - 2^{-12})$
	Çekirdek (Kernel)	Linear, RBF, Polinominal
	Gamma	(0.000001 - 10)
RF	Ağaç Sayısı	(2 - 1000)
	Ölçüt	Gini, Entropy
KNN	Komuş Sayısı	2 - Örnek Sayısı
	Algoritma	KD-Tree, Ball-Tree, Brute
XGB	Ağaç Sayısı	(2 - 1000)
	Öğrenme Oranı	(0.0001 - 10)
	Alpha	(0 - 32)
	Gamma	(0 - 32)

Optuna verilen hiper-parametre uzayı içerisinden algoritmaya uygun parametreleri rastgele arama yöntemi kullanarak denemektedir. Her denenen parametre grubu sonrası ilgili parametreler ile bu parametre grubuyla alınan başarım saklanır. Parametre başarımının hesaplanması için eğitim veri setinin %20'si doğrulama seti olarak ayrılmış ve geri kalan doğrulama-eğitim seti olarak kullanılmıştır. Her rastgele parametre grubuyla algoritmalar 250 defa tekrar edilmiş ve en uygun parametreler belirlenmiştir. Optimizasyon sürecinde her algoritma için kullanılan farklı parametre uzayı tablo 1'deki gibidir.



Şekil 3. *Hiper Parametre Optimizasyon Süreci.*

Daha sonra en uygun parametrelerle oluşturulan modeller eğitim setinin tamamı kullanılarak çalıştırılmıştır. Sınıflandırma sürecinde izlenen adımlar şekil 3'te gösterilmiştir. Nihai model ön işlem sonucunda oluşturulan 4 farklı veri setine de uygulanarak test edilmiştir. Testler sonucunda elde edilen doğruluk değerleri tablo 2'de gösterilmektedir.

Tablo 2. *Doğruluk Değerleri.*

Ham Veri	**Gereksiz Kelimelerden Temizlenmiş Veri**	**Kelime Normalizasyonu Uygulanmış Veri**	**Kök Alınmış Veri**

	TF	*TF-IDF*	*W2V*	*GloVe*	*TF*	*TF-IDF*	*W2V*	*GloVe*	*TF*	*TF-IDF*	*W2V*	*GloVe*	*TF*	*TF-IDF*	*W2V*	*GloVe*
LR	0.76	0.783	0.77	0.745	0.767	0.788	0.769	0.774	0.767	0.784	0.745	0.717	0.741	0.783	0.774	0.774
SVM	0.712	0.786	0.786	0.724	0.724	**0.795**	0.783	0.767	0.7224	0.779	0.7	0.7	0.733	0.786	0.767	0.724
KNN	0.1	0.665	0.695	0. 674	0.1	0.662	0.674	0.665	0.1	0.688	0.684	0.68	0.1	0.7	0.68	0.665
RF	0.731	0.722	0.76	0.738	0.714	0.714	0.776	0.722	0.724	0.717	0.745	0.722	0.762	0.738	0.764	0.738
XGB	0.643	0.745	0.76	0.703	0.703	0.693	0.757	0.733	0.1	0.551	0.736	0.676	0.605	0.676	0.733	0.703

Sonuçlar incelendiğinde gelen olarak %75'in üzerinde bir doğruluk oranıyla gelen taleplerin hangi departmanı ilgilendirdiği tahmin edilmiştir. En yüksek başarım %79,5 ile SVM algoritmasıyla alınırken, TF-IDF yönteminin diğer vektörleştirme yöntemlerine göre daha başarılı olduğu söylenebilir. Kelime temsil yöntemleri olan Word2Vec ve GloVe teknikleri ortalama üzeri sonuç vermelerine rağmen TF-IDF skorlarının altında kalmışlardır. Bu yöntemler temel olarak kelimeler arasındaki anlamsal benzer ilişkisini göstermek için kullanılmaktadır. Kullanılan veri setinin büyüklüğüyle kelimeler arası anlamsal uzaklığın doğru olarak tahmininde doğrusal bir oran vardır. Çalışmamız içerisinde bu yaklaşımların daha düşük sonuç vermesi, kelimeler arasındaki ilişkiyi tam olarak öğrenecek kadar fazla verinin olmadığına işaret olabilir. Çünkü özellikle yapay sinir ağları modellerinde veri sayısı arttıkça tahmin başarısının arttığı bilinmektedir.

Şekil 4. *Sınıf Etiketi Bazlı Başarım Değerleri.*

En başarılı algoritma ve vektörleştirme ikilisi olan SVM/TF-IDF sonuçlarına ait karmaşıklık matrisi şekil 4'te gösterilmektedir. Etiket bazlı incelendiğinde "arge" ve "fire" alanları ile ilgili gelen taleplerin %92'nin üzerinde bir başarıyla doğru tahmin edildiği gözükmektedir. Öte yandan en düşük başarım %58,82 ile "losd" etiketine sahip taleplerin sınıflandırılmasında elde edilmiştir. "abap", "loqm", "losd", "fico", "bwbo" , "hrgn" , "lomm" etiketleri aslında aynı kurumsal kaynak planlama uygulamasına ait alt modülleri tanımlayan etiketlerdir. Dolayısıyla bu etiketlerle gelen talep metinlerinin içerisinde

birbirine benzer kelimelerin geçmesi muhtemeldir. Yaşanan düşük başarımın buna bağlı olduğu düşünülmektedir.

IV. SONUÇ

Bilişim firmalarında müşterilerin ilettiği geliştirme ve sorun çözme gibi talepler genellikle bir uzman personel tarafından incelenerek ilgili departman yönlendirilmektedir. Talep sayısının az olduğu firmalar için bu durum herhangi bir sorun teşkil etmezken, aynı anda çok sayıda taleplerin geldiği danışmanlık firmalarında yönlendirme işleminin bir kişi tarafından yapılması insan hatasına yol açabileceği gibi kaynak ve zaman kaybına da neden olabilir. Bunun yerine gelen talepleri otomatik olarak inceleyerek ilgili departmana aktaracak bir sistem her açıdan firmaya katkı sağlar. Bu çalışmada, bilişim ve yazılım destek konularında gelen gerçek zamanlı taleplerin metin madenciliği ve makine öğrenmesi yöntemleriyle doğru alt-departmana yönlendirilmesini içeren örnek bir uygulama yapılmıştır. Analizler içerisinde metni farklı şekilde vektöre çeviren yöntemlerle beraber ön işlem uygulamalarının da talep sınıflandırılması üzerine etkileri detaylı şekilde incelenmiştir. Sonuçlara bakıldığı zaman literatürde sıklıkla kullanılan TF-IDF vektörleştirme yönteminin neredeyse her algoritma için daha yüksek sonuç aldığı görülmüştür. Metin içerisinde yer alan önemsiz kelimelerin çıkartılması da beklenildiği gibi başarıyı arttırmıştır. Kelime düzgünleştirme ve kök alma ön-işlemleri ise genel olarak başarımda düşüklüğe yol açmıştır. Her ne kadar metnin bu şekilde ön işlemlere sokulması başarımı arttırması beklense de bu sürecin problem bazlı olduğu söylenebilir. Sınıf bazında başarımlara bakıldığında ise aynı departmana ait alt konuları içeren taleplerin tahmin başarımı daha düşük kaldığı gözlemlenmiştir. Bu durum farklı sınıf etiketine sahip olsa da talepler içinde benzer kelimelerin geçmesiyle açıklanabilir. Test işlemleri sonucunda ortalama %79 civarında başarı elde edilmiştir. Bu sonuç ilgili modelin karar destek sistemi olarak firma içlerinde talep yönlendirme sürecinde kullanılabilirliği açısından yeterli görülmüştür.

Çalışmamız firmalara gelen taleplerin ilgili departmanla beraber doğru personele aktarılmasını sağlayacak bir sistemin ön çalışmasıdır. İleriki çalışmalarda personel önerme modeliyle birleştirilerek uçtan uca talep yönlendirme sistemi geliştirilecektir.

TEŞEKKÜR: Bu çalışma, Detaysoft Ar-Ge Merkez bünyesinde yürütülen çalışmaların sonucudur. Desteklerinden dolayı Detaysoft Ar-Ge Merkezine teşekkür ederiz. Bu çalışmada rapor edilen sayısal hesaplamaların bir kısmı TÜBİTAK ULAKBİM Yüksek Başarımlı ve Grid Hesaplama Merkezi (TRUBA)'da gerçekleştirilmiştir.

V. KAYNAKLAR

[1] S. Ballı and O. Karasoy, 'Development of content-based SMS classification application by using Word2Vec-based feature extraction', IET Software, vol. 13, no. 4, pp. 295–304, 2019.

[2] G. M. Shahariar, S. Biswas, F. Omar, F. M. Shah, and S. B. Hassan, 'Spam review detection using deep learning', 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2019, pp. 0027–0033.

[3] T. KAŞIKÇI and H. Gökçen, 'Metin madenciliği ile e-ticaret sitelerinin belirlenmesi', Bilişim Teknolojileri Dergisi, c. 7, s. 1, 2013.

[4] M. Bouazizi and T. Ohtsuki, 'Multi-class sentiment analysis in Twitter: What if classification is not the answer', IEEE Access, vol. 6, pp. 64486–64502, 2018.

[5] A. Arifianto et al., ‘Developing an LSTM-based Classification Model of IndiHome Customer Feedbacks’, 2020 International Conference on Data Science and Its Applications (ICoDSA), 2020, pp. 1–5.

[6] P. S. Parmar, P. K. Biju, M. Shankar, and N. Kadiresan, ‘Multiclass text classification and analytics for improving customer support response through different classifiers’, 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018, pp. 538–542.

[7] A. Onan, E. Atik, and A. Yalçın, ‘Machine learning approach for automatic categorization of service support requests on university information management system’, International Conference on Intelligent and Fuzzy Systems, 2020, pp. 1133–1139.

[8] N. Kim and S. Hong, ‘Automatic classification of citizen requests for transportation using deep learning: Case study from Boston city’, Information Processing & Management, vol. 58, no. 1, p. 102410, 2021.

[9] A. A. Gorbunova, A. S. Trunov, and V. I. Voronov, ‘Intelligent analysis of technical support requests in Service Desk ticketing systems’, 2020 International Conference on Engineering Management of Communication and Technology (EMCTECH), 2020, pp. 1–6.

[10] R. Mitkov, The Oxford handbook of computational linguistics. Oxford University Press, 2004.

[11] C. Manning and H. Schutze, Foundations of statistical natural language processing. MIT press, 1999.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, ‘Efficient estimation of word representations in vector space’, arXiv preprint arXiv:1301.3781, 2013.

[13] J. Pennington, R. Socher, and C. D. Manning, ‘Glove: Global vectors for word representation’, Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[14] W. Etaiwi and G. Naymat, ‘The Impact of applying Different Preprocessing Steps on Review Spam Detection’, Procedia Computer Science, vol. 113, pp. 273–279, Jan. 2017, doi: 10.1016/j.procs.2017.08.368.

[15] G. Eryiğit, ‘ITU Turkish NLP web service’, Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014, pp. 1–4.

[16] F. Pedregosa et al., ‘Scikit-learn: Machine learning in Python’, the Journal of machine Learning research, vol. 12, pp. 2825–2830, 2011.

[17] R. Řehůřek and P. Sojka, ‘Software Framework for Topic Modelling with Large Corpora’, Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, May 2010, pp. 45–50.

[18] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, ‘Optuna: A next-generation hyperparameter optimization framework’, Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 2623–2631.