
GAZİ ÜNİVERSİTESİ
Gazi
1926

MAKİNE ÖĞRENMESİ İLE ETKİLEŞİMLİ YARDIM MASASI SİSTEMİ TASARIMI

Buğra Kaan TÜRKMENOĞLU

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

OCAK 2023

ETİK BEYAN

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

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

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

Buğra Kaan
TÜRKMENOĞLU
12/01/2023

MAKİNE ÖĞRENMESİ İLE ETKİLEŞİMLİ YARDIM MASASI SİSTEMİ TASARIMI

(Yüksek Lisans Tezi)

Buğra Kaan TÜRKMENOĞLU

GAZİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Ocak 2023

ÖZET

Yardım masası sistemi kullanıcılar ile müşteri hizmetleri arasındaki ana iletişim aracıdır. Yardım masası bilet alanının manuel seçilmesi sonucunda bilet alanı yanlış atanabilir, biletin çözüm süresi uzayabilir ve iş gücü kaybedilebilir. Biletin önceliğinin yanlış seçilmesi durumunda kritik önemi olmayan biletlere yüksek öncelik verilebilir. Birçok konu hakkında bilgi sahibi olma zorunluluğu, süreç aktarımının zorluğu ve tekrarlayan biletler sebebiyle müşteri hizmetleri temsilcileri üzerinde çok sayıda bilet birikir. Bu sorunların üstesinden gelmek için biletleri üst ile alt alanlara ve önceliklerine göre sınıflandıracak, müşteri hizmetleri temsilcileri tarafından kapatılan biletlerle aynı anlama gelen biletleri önererek bir an önce kapatılmasını sağlayacak modellerin tasarlanması amaçlanmıştır. Çok sınıflı bilet alanı sınıflandırmanın zorluğunu aşmak için hiyerarşik yaklaşım kullanılmıştır. Bilet sınıflandırma modellerinin eğitimi için bir şirketin yardım masası sisteminden Türkçe bilet açıklamasıyla üst ile alt alan ve öncelik etiketlerini içeren 4 adet veri kümesi toplanmıştır. Benzerlik modelini eğitmek için bir bilet üst alanına ait biletler kartezyen olarak eşleştirilmiş ve manuel olarak etiketlenmiştir. Veri ön işleme adımları, yeniden örnekleme ve boyut küçültme sayesinde, dengesiz bir veri kümesinde geleneksel makine öğrenmesi algoritmalarını kullanarak çok sınıflı metin sınıflandırma probleminin zorluğu aşılmıştır. SVM yöntemi kullanılarak biletlerin %94,4 doğrulukla üst alana, %88,22 ile %98,92 arasında doğrulukla alt alanlara sınıflandırılması gerçekleştirilmiştir. K-NN yöntemi kullanılarak biletlerin %97,35 doğrulukla öncelik sınıflandırması gerçekleştirilmiştir. MaLSTM modeli kullanılarak bilet benzerliği için 0,235 ortalama kare hata, 0,4848 kök ortalama kare hata, 0,4603 ortalama mutlak hata, 0,2528 Pearson korelasyon katsayısı, 0,2487 Spearman'ın sıralama korelasyon katsayısı elde edilmiştir. Bu tez çalışması, TUSAŞ Bilimsel Araştırmalar Programı (TUSAŞ BAP) kapsamında desteklenmiştir.

Bilim Kodu	:	92432
Anahtar Kelimeler	:	Yardım masası, bilet sınıflandırma, anlamsal metin benzerliği, makine öğrenmesi, doğal dil işleme, Word2vec, Doc2vec, LSTM
Sayfa Adedi	:	64
Danışman	:	Doç. Dr. Oktay YILDIZ

INTERACTIVE HELPDESK SYSTEM DESIGN WITH MACHINE LEARNING

(M. Sc. Thesis)

Buğra Kaan TÜRKMENOĞLU

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

January 2023

ABSTRACT

The helpdesk system is the main communication tool between users and customer service representatives. Because of the manual selection of the helpdesk ticket field, the ticket field may be assigned incorrectly, the resolution time of the ticket may be extended, and the workforce may be lost. In case of wrong selection ticket priority, non-critical tickets may be given higher priority. Customer service representatives accumulate a large number of tickets due to the necessity of having knowledge about many subjects, the difficulty of transferring the process, and repetitive tickets. To overcome these problems, it is aimed to design models that will classify tickets according to upper and lower fields and their priorities, and that will suggest the closing of tickets that have the same meaning as tickets closed by customer service representatives. A hierarchical approach was used to overcome the difficulty of multi-class ticket field classification. For the training of ticket classification models, four datasets were collected from a company's help desk system, with the upper field, lower field and priority classes, and ticket description. To train the similarity model, tickets belonging to a ticket upper field were mapped as cartesian and manually labeled. Thanks to data preprocessing steps, resampling and dimension reduction, the challenge of multiclass text classification problem was overcome by using traditional machine learning algorithms on an unbalanced dataset. The classification of the tickets to the upper field with 94.4% accuracy and to lower fields with an accuracy between 88.22% and 98.92% was carried out using the SVM method. Priority classification of tickets with an accuracy of 97.35% was performed using the K-NN method. Using the MaLSTM model, 0.235 mean squared error, 0.4848 root mean square error, 0.4603 mean absolute error, 0.2528 Pearson correlation coefficient, and 0.2487 Spearman's rank correlation coefficient were obtained for ticket similarity. This thesis study was supported within the scope of TUSAŞ Scientific Research Program (TUSAŞ BAP).

Science Code	:	92432
Key Words	:	Help desk, ticket classification, semantic text similarity, machine learning, natural language processing, Word2vec, Doc2vec, LSTM
Page Number	:	64
Supervisor	:	Assoc. Prof. Dr. Oktay YILDIZ

TEŞEKKÜR

Yüksek lisans tez çalışmam süresince kıymetli görüşlerini esirgemeyen, takıldığım noktalarda yol gösteren ve zorluklar karşısında destek olan değerli hocam Doç. Dr. Oktay YILDIZ ‘a teşekkürlerimi ve saygılarımı sunarım.

Çalışmam boyunca takıldığım zorluklar karşısında bana destek olan, yeni fikirler aşılayan ve yapıcı eleştiriler sunan Emrah ÖZTÜRK ‘e yardımları için teşekkürler ederim.

Eğitim hayatımda ve hayatımın her anında desteğini ve varlığını esirgemeyen tüm aileme teşekkürü bir borç bilirim.

İÇİNDEKİLER

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

Şekil **Sayfa**

SİMGELER VE KISALTMALAR

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

Simgeler	**Açıklamalar**
%	Yüzde

Kısaltmalar	**Açıklamalar**
Bagging-DT	Bagging Decision Tree
Bagging-SVM	Bagging Support Vector Machine
BAP	Bilimsel Araştırma Programı
BT	Bilgi Teknolojileri
CBOW	Continuous Bag of Words
D-LSTM	Dependency-based Long Short Term Memory
DBOW	Distributed Bag of Words
DMPV	Distributed Memory Paragraph Vector
HTML	Hypertext Markup Language
IBM	International Business Machines
İTÜ	İstanbul Teknik Üniversitesi
K-NN	K-Nearest Neighbors
KNIME	Konstanz Information Miner
LSTM	Long Short Term Memory
MAE	Mean Absolute Error
MAIS	Multi-Agent Indexing System
MALSTM	Manhattan Long Short Term Memory
MSE	Mean Squared Error
PCA	Principal Component Analysis
PCC	Pearson Correlation Coefficient
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network

Kısaltmalar	Açıklamalar
SG	Skip Gram
SMO	Sequential Minimal Optimization
SRCC	Spearman Rank Correlation Coefficient
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
URL	Uniform Resource Locator

1. GİRİŞ

Kurumlarda finans, akademi ve satın alma gibi çeşitli modülleri içeren ERP yazılımları kullanılmaktadır. Kurumların içerisindeki bölümlerin bağlantılı olmasına benzer olarak ERP modülleri de birbirleriyle ilişkilidir. Dolayısıyla bir modülde meydana gelen problem diğer modülleri etkiliyorsa kurumun işleyişini aksatabilir. Kullanıcılar ile müşteri hizmetleri temsilcileri arasındaki ana iletişim aracı olan yardım masası sistemi üzerinden istek veya sorunlar bildirilir. Yardım masası sistemi kullanıcı memnuniyeti açısından kurumlar için önemli olduğundan üzerinde birçok çalışma yapılmaktadır.

İstek veya sorunların bildirilmesi için yardım masası sisteminde bilet kullanılmaktadır. Bilet, yardım masası sisteminin temel bileşenidir. Bilet açıklaması, öncelik, kategori, oluşturan kişi, oluşturma zamanı ve atanan kişi gibi alanlardan oluşmaktadır. Kullanıcılar tarafından açılan biletlerin statüsü müşteri hizmetleri temsilcileri tarafından değiştirilmektedir. Müşteri hizmetleri temsilcilerine atanan biletler çözümlendikten sonra harcanan süre ve çözüm açıklaması girilerek kapatılmaktadır. Müşteri hizmetleri temsilcileri bilet ile ilgili ek bilgiye ihtiyaç duyarlarsa bileti; kullanıcıdan bilgi talebi durumuna, gereksiz bulurlarsa iptal durumuna veya başka bir duruma getirebilirler.

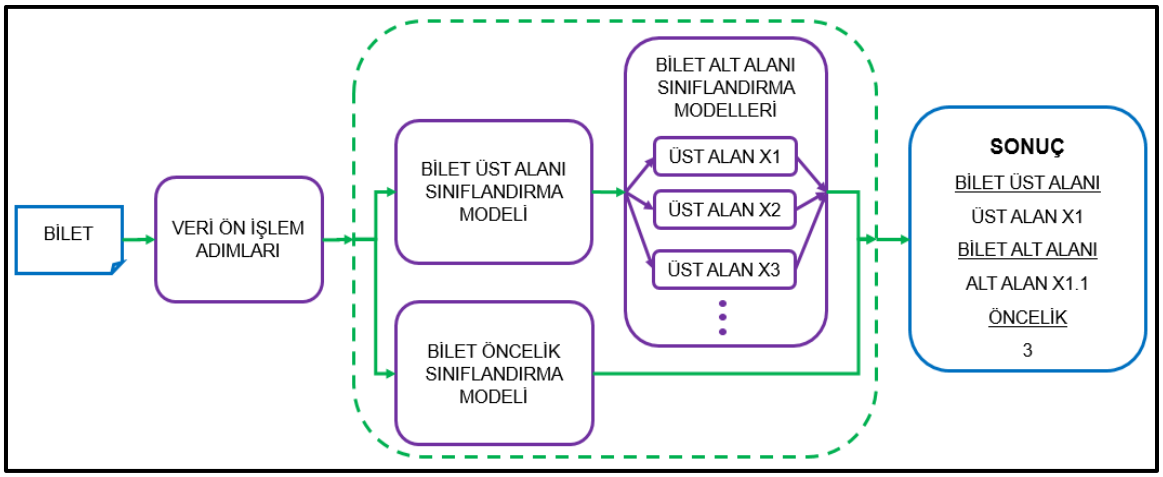
Kullanıcılar tarafından açılan biletler ilgili oldukları kategorilere ve önceliklere göre sınıflandırılmaktadır. Biletlerin sınıflandırılması kullanıcılar tarafından manuel veya yardım masası sistemi tarafından otomatik olmak üzere iki yol ile gerçekleştirilebilmektedir. Biletin kategorisi belirlendikten sonra biletin kategorinin uzmanı olan müşteri hizmetleri temsilcilerine biletler atanmaktadır.

Yardım masası sistemlerinin birçok bilet alanı içermesi durumu bilet atamayı zorlaştırmaktadır. BT kaynaklarının eksik olması problem çözme sürecinde gecikmelere neden olur. Şirketlerin biletleri alanlarına göre sınıflandırmak için havuz personeli kullanması ise iş gücü kaybına sebebiyet verir. Bilet alanının manuel olarak seçilmesi; yanlış çözüm grubuna gönderilmesine, yeniden atanmasına ve çözüm süresinin uzamasına neden olur. Bilet alanının yanlış seçilmesi ise kurumlardaki süreçlerin aksamasına ve kurumlardaki normal işleyişinin bozulmasına yol açar. Yanlış yönlendirmeler sonucunda olumsuz mali sonuçlar ortaya çıkabilir [1]. Yanlış yönlendirmeler sonucunda

gerçekleştirilen çoklu atamalar, genellikle bildirimde veya biletin atanmasında yapılan hataların sonucudur. Açık kaynaklı bir projedeki geliştiricilerin yaptığı atamaların %37 - %44'ü bu tür hatalardır [2]. Gerçek bir üretim destek sistemi üzerinde yapılan bir çalışmada, bir biletin ortalama geri dönüş süresinin, transfer sayısı 2'den 4'e arttığında %100'e kadar yükseldiği gözlemlenmiştir [1]. Biletlere doğru bir şekilde öncelik verilmesi kritik bir görevdir. Kritik durumda olmayan biletlere yüksek öncelik verilmesi hatalı planlamaya ve müşteri memnuniyetsizliğine sebep olur.

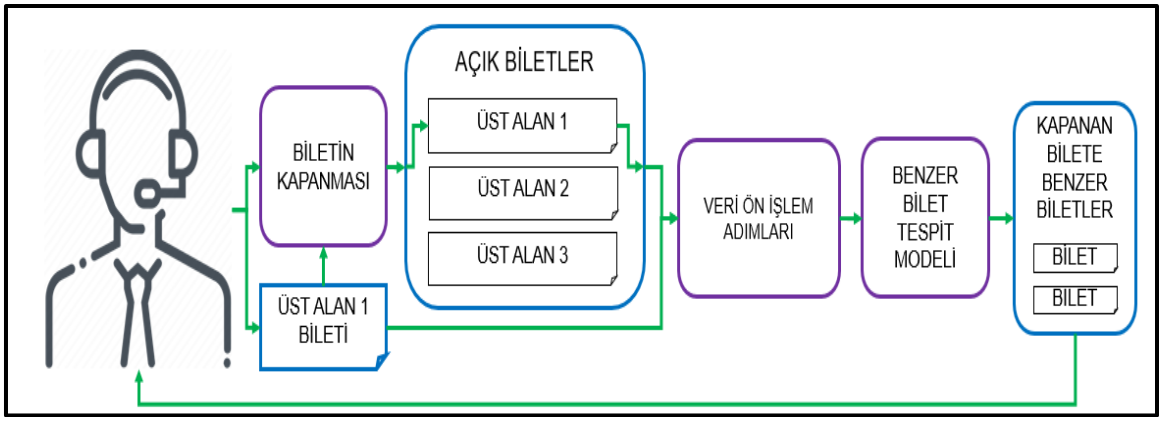
Yardım masası sistemi müşteri hizmetleri temsilcilerinin hâkim olması gereken birçok süreç içerir. Bu süreçler ise müşteri hizmetleri temsilcilerine bağlı olmaktadır. Süreçlerin karmaşıklığı nedeniyle ise bilgi aktarımı zordur. Bunların neticesinde yardım masası sistemlerinde yinelenen biletler meydana gelir. Bu tekrarlayan biletler, müşteri hizmetleri temsilcisi üzerinde birçok biletin birikmesine neden olur. İşlemlerin otomatikleştirilmemesi sebebiyle yardım masası sistemlerinde kullanıcılar uzun bilet çözüm sürelerinden rahatsız olabilirler. Bazı yardım masası sistemleri, yeni açılan bilete benzer kapalı biletlerin çözümlerini öneri olarak listeler. Kullanıcılar ise listeyi inceler. Aynı şekilde, müşteri hizmetleri temsilcileri de üzerinde çok sayıda bilet bulundurabilirler. Bu biletlerden birini çözdüklerinde ise o biletle aynı anlama gelen biletleri gözden kaçırabilirler. Bu nedenle müşteri hizmetleri temsilcileri biletin kapatıldığı anda benzer biletlerden haberdarlarsa, o anda sahip oldukları bilgi birikimi ile o biletleri de hızlıca kapatabilirler.

Akıllı yardım masası sistemine sahip olmak kaliteli müşteri hizmeti sunmak için önemlidir. Müşteri hizmetlerinin kalitesi ise kurumların başarısını doğrudan etkiler. Dolayısıyla akıllı yardım masası sistemleri her geçen gün daha yaygın hale gelmektedir. Çok sayıda kullanıcı ve ürün olduğunda akıllı sistemlere ihtiyaç duyulur. Böylece, kullanıcıların aradıkları seçenekleri bulmaları kolaylaşır. Bu araştırma çalışmasının amacı Şekil 1.1‘de görüldüğü üzere yardım masası biletlerini üst ile alt alanlara ve önceliklerine göre sınıflandıracak, Şekil 1.2’de görüldüğü üzere müşteri hizmetleri temsilcileri tarafından kapatılan biletler ile aynı anlama gelen biletleri müşteri hizmetleri temsilcilerine kapatmaları için önerecek modeller elde etmektir.



Şekil 1.1. Biletlerin üst ile alt alanlara ve önceliklerine göre sınıflandırılması

Bu çalışmanın diğer çalışmalardan ilk farkı, kullanıcılara açık biletler için çözümler önermek yerine, müşteri hizmetleri temsilcilerine o anda sahip oldukları bilgi birikimi ile hızlıca kapatılabilecek biletleri önermesidir. İkinci fark ise Türkçe yardım masası biletlerinin anlamsal benzerliğini hesaplayan ilk çalışma olmasıdır.



Şekil 1.2. Kapatılan bilete benzer biletlerin tespit edilmesi

Bu tez çalışmasında altı adet amacımız bulunmaktadır. Çalışmanın hedefleri şu şekilde özetlenebilir:

- Makine öğrenmesi yöntemlerini kullanarak biletleri üst ile alt alanlara ve önceliklerine göre sınıflandırmak
- Yardım masası sistemindeki çok sayıda alan için çok sınıflı metin sınıflandırma probleminin zorluğunu hiyerarşik yaklaşımı kullanarak aşmak
- Dengesiz bir veri kümesinde bulunan yardım masası bilet açıklamalarına özel veri ön

işleme adımlarını kullanarak geleneksel makine öğrenme algoritmaları ile metin sınıflandırma probleminin zorluğunu aşmak

- Müşteri hizmetleri temsilcileri tarafından kapatılan biletlerle aynı anlama gelen açık biletlerin müşteri hizmetleri temsilcilerine bildirilerek hızlı bir şekilde kapatılmasını sağlamak
- Benzer biletlerin hızlı bir şekilde kapatılmasını sağlayarak müşteri hizmetleri temsilcilerine daha karmaşık ve zor sorunlarla başa çıkmaları için uzun bir süre tanımak
- İş süreçlerinde aksamaların önlenebileceğini, işçilikten tasarruf edilebileceğini, kullanıcı memnuniyeti sağlanabileceğini ve yardım masası sisteminin etkinliğinin sağlanabileceğini göstermek

Bu araştırma çalışmasının ana katkıları aşağıda özetlenmiştir.

- Yardım masası sisteminin bilet alan sınıflandırma ve öncelik verme süreçlerini otomatikleştirilmektedir.
- Yardım masası biletlerine alan ve öncelik atanmasını otomatik hale getirilerek çözüm süreleri kısaltılmaktadır.
- Müşteri hizmetleri temsilcileri tarafından kapatılan biletlerle aynı anlama gelen biletleri müşteri hizmetleri temsilcilerine tavsiyede bulunarak kısa sürede kapatılmasını sağlanmaktır.
- Benzer biletlerin tespit ederek ve bunların hızlı kapanmasını sağlayarak müşteri hizmetleri temsilcilerinin daha zor sorunlarla uğraşmasına olanak tanımaktadır [3].
- Yardım masası bilet alanı ve önceliği atamasını otomatikleştirerek iş gücünden tasarruf sağlanmaktadır.
- Yardım masasının etkili ve etkileşimli çalışmasını sağlayarak müşteri memnuniyetini artırmaktadır.

2. LİTERATÜR ARAŞTIRMASI

Yapay zekâ ve dilbilim alanlarında yapılan çalışmalarda doğal dil işleme yöntemleri, metinlerin işlenmesi için yaygın olarak kullanılmaktadır. Sosyal ağlar, forumlar, web siteleri, e-postalar ve araştırma makaleleri yayınlayan çevrimiçi kütüphaneler gibi çok büyük miktarda veri üreten kaynakların artması nedeniyle doğal dil işleme yöntemlerinin kullanıldığı metin madenciliği çalışmaları giderek önem kazanmaktadır [4].

Metin madenciliği üzerine çeşitli çalışmalar gerçekleştirilmiştir. HaCohen-Kerner, Dilmon, Hone ve Ben-Basan [5] şirketlere gönderilen İbranice yazılmış şikâyet mektuplarının otomatik metin sınıflandırmasını araştırmıştır. Hark et al. [6] doğal dil işleme yöntemlerini kullanarak tamamen bağımsız belge yığınları arasındaki benzerlikleri hesaplamaya çalışmıştır. Boufrida ve Zizette [7] doğal dil işleme ve metin madenciliği yöntemlerini kullanarak metinlerden karmaşık ilişkileri çıkarmak ve bunları kurallar çerçevesinde kodlamak için bir öneri sunmuştur. Ayata, Saraçlar ve Özgür [8] destek vektör makinesi (Support Vector Machine – SVM) yöntemini kullanarak Türkçe tweetlerin sektör bazlı duygu analizini gerçekleştirmiştir. Moukrim, Abderrahim, Tarik ve Benlahmer [9] doğal dil işleme yöntemlerini kullanarak Arapça cümlelerdeki yazım hatalarının giderilmesi için bir yazım hatası düzeltme sistemi geliştirmiştir. Zouaoui ve Rezeg [10] intihal tespiti için doğal dil işleme, indeksleme ve değerlendirme aşamalarından oluşan Multi-Agent Indexing System (MAIS) 'i sunmuştur. Khan ve Shamsi [11] doğal dil işleme yöntemlerini uygulayarak bir hastanın elektronik sağlık kaydı aracılığıyla hastalıkları teşhis etmeye yardımcı olan derin sinir ağı tabanlı bir klinik karar destek sistemi önermiştir. Anand ve Wagh [12] kelime-cümle temsillerini ve sinir ağını kullanarak Hint yasal yargı belgelerini özetlemek için bir yöntem önermiştir. Mohasseb, Bader-El-Den ve Cocea [13] soruların yapısından yararlanan dilbilgisine dayalı bir yaklaşımı kullanarak soruların sınıflandırılması için bir çerçeve önermiştir. Soruların dil bilgisel kalıplara dönüştürülmesi için resmi bir dilbilgisi yaklaşımı kullanılmış ve otomatik sınıflandırma için makine öğrenme algoritmaları uygulanmıştır.

Metin sınıflandırma, bir metin belgesini en alakalı etiketler veya kategorilerle otomatik olarak etiketleme sürecidir [14]. Genel olarak söylemek gerekirse, yapılandırılmamış metin formatında var olan devasa miktardaki bilgiyi organize etmek ve kullanmak için en önemli

yöntemlerden biri de metin sınıflandırmadır [4]. Metin sınıflandırma, dünya çapında giderek artan sayıda elektronik belgeyi organize etme gerekliliği nedeniyle zorlu araştırma konularından biri olmuştur [15]. Doğal dil işleme yöntemleri kullanılarak yardım masası biletlerinin alanlarına ve önceliklerine göre sınıflandırılması konusunda son yıllarda çeşitli araştırmalar yapılmıştır. Al-Hawari ve Barham [16] Alman Ürdün Üniversitesi'nin BT ihtiyaçlarını karşılamak için çevrimiçi bir web tabanlı yardım masası sistemini tanıtmıştır. Yardım masası biletlerinin doğru hizmetle ilişkilendirilmesi için 1254 bilete veri ön işleme adımlarını uygulamıştır. Terim Frekans - Ters Belge Frekansı (Term Frequency-Inverse Document Frequency - TF-IDF) yöntemi kullanılarak belge gösterimleri çıkarılmış ve Sıralı Minimum Optimizasyon (Sequential Minimal Optimization - SMO) yöntemi kullanılarak biletlerin sınıflandırılması gerçekleştirilmiştir. Paramesh ve Shreedhara [17] doğal dil işleme yöntemlerini kullanarak akıllı bir otomatik bilet sınıflandırma sistemi geliştirmiştir. 8468 örnek ve 18 kategoriden oluşan bir kurumsal BT altyapı hizmet yönetimi bilet veri kümesine, veri ön işleme adımları uygulanmıştır. Rastgele Aşırı Örnekleme ile Rastgele Yetersiz Örnekleme yöntemleri kullanılarak veri kümesi dengeli hale getirilmiştir. TF-IDF yöntemi kullanılarak doküman gösterimleri elde edilmiştir. Ki Kare yöntemi kullanılarak verilerin boyutu küçültülmüştür. Random Forest yöntemi kullanılarak bilet sınıflandırması gerçekleştirilmiştir. Agarwal, Sindhgatta ve Sengupta [1] yüksek doğruluk seviyelerini korurken bilet gönderme sürecini otomatikleştirmeyi amaçlayan öğrenme tabanlı bir araç olan SmartDispatch'i sunmuştur. SmartDispatch iki sınıflandırma yaklaşımına sahiptir. Birincisi SVM yöntemi, ikincisi ise ayrımcı terim tabanlı bir yaklaşımdır. International Business Machines (IBM)'in hizmet görevlerinden elde edilen büyük ölçekli veriler üzerinde uygulanan metine dayalı sınıflandırma teknikleri kullanılarak, bu sürecin ne ölçüde otomatikleştirilebileceği incelenmiştir. Paramesh ve Shreedhara [18] bir kuruluşun 10742 bilet ve 18 kategori içeren BT altyapısı bilet verilerine veri ön işleme adımları uygulamış ve örnekleme teknikleri gerçekleştirmiştir. TF-IDF yöntemi kullanılarak belge gösterimleri elde edilmiştir. Ki Kare yöntemi kullanılarak verilerin boyutu küçültülmüştür. Torbalı Destek Vektör Makinesi (Bagging Support Vector Machine - Bagging-SVM) yöntemi kullanılarak bilet sınıflandırması gerçekleştirilmiştir. Pereira, Ribeiro ve Silva [19] makine öğrenmesi tekniklerini kullanarak biletlerin otomatik olarak sınıflandırılması için bir modül önermiştir. Bir firma tarafından sağlanan 10.000 yardım masası biletine veri ön işleme adımları uygulanmıştır. TF-IDF yöntemi kullanılarak belge gösterimleri elde edilmiştir. SVM yöntemi kullanılarak biletlerin sınıflandırılması gerçekleştirilmiştir. Altıntaş ve Tantuğ [20] müşteri hizmetleri temsilcilerine biletleri otomatik olarak atamak için sorun

takip sistemine bir eklenti önermiştir. Veri ön işleme kapsamında İstanbul Teknik Üniversitesi (İTÜ) Durum Takip Sistemi'nden toplanan yaklaşık 10000 Türkçe bilete, İTÜ Türkçe Doğal Dil İşleme Yazılım Zinciri uygulanmıştır. TF-IDF yöntemi kullanılarak belge gösterimleri çıkarılmıştır. Poly Kernel SVM yöntemi kullanılarak bilet sınıflandırması gerçekleştirilmiştir. Paramesh ve Shreedhara [21] yardım masası bilet sınıflandırıcı sisteminin doğruluğunu iyileştirmek için en popüler üç gruplandırma yöntemi olan torbalama, artırma ve oylama grubunu kullanmıştır. 10742 bilet ve 18 sınıf içeren veri kümesine veri ön işlem adımları uygulanmıştır. Sınıf dengesizliği nedeniyle Rastgele Aşırı Örnekleme ve Rastgele Yetersiz Örnekleme yöntemleri uygulanmıştır. TF-IDF yöntemi kullanılarak belge gösterimleri elde edilmiştir. Torbalı Karar Ağacı kullanılarak bilet sınıflandırması gerçekleştirilmiştir. Miliano, Steven, Kosim, Jayadi ve Mauritsius [22] 40351 yardım masası biletine veri ön işleme adımlarını uygulamıştır. Konstanz Information Miner (KNIME) yöntemini kullanarak kelime temsillerini elde etmiştir. Random Forest yöntemini kullanarak biletlerin sınıflandırılmasını gerçekleştirmiştir. Tekin ve Tunalı [23] metin madenciliği yöntemlerini kullanarak müşteri taleplerinin önem derecesini tahmin etmeye çalışmıştır. Kurumsal bir şirketin talep yönetim sisteminden alınan kayıtlara veri ön işleme adımları uygulanmıştır. Kelime temsilleri, TF-IDF yöntemi ile elde edilmiştir. Biletler, Random Forest yöntemi kullanılarak önceliklerine göre sınıflandırılmıştır. Walkek [24] her biletin önemini değerlendirmek, biletleri önemlerine göre sıralamak ve sıralı bilet listesini müşteri hizmetleri temsilcilerine göstermek için yardım masası sistemine akıllı bir sistem önermiştir. Önerilen sistem, biletin seçilen özelliklerine ve bir bilgi tabanının eğer-öyleyse kurallarına dayalı olarak genel önemini değerlendiren bir uzman sisteme bağlanmıştır. Değerlendirilen biletler daha sonra önemlerine göre sıralanmış ve müşteri hizmetleri temsilcilerine görsel olarak sunulmuştur. Önerilen sistem belirli bir örnek üzerinde doğrulanmıştır.

Doğal dil işleme yöntemleri kullanılarak metinlerin benzerliği konusunda son yıllarda birçok araştırma yapılmıştır. Nguyen, Duong ve Cambria [25] anlamsal benzerliği ölçmek için metinlerin birbirine bağlı temsillerini kullanan yeni bir yöntem önermiştir. Yöntem, önceden elde edilmiş kelime vektörlerine ve dış bilgi kaynaklarına dayalı olarak ikiye ayrılmıştır. Anlamsal benzerliği hesaplamak için SVM yöntemi kullanılmıştır. Zhang ve Zhou [26] yasal hükümler arasındaki benzerliği hesaplamayı amaçlamıştır. Temsilleri elde etmek için Doc2Vec yöntemi kullanılmıştır. Doc2vec'in çok iyi sonuçlar vermemesine rağmen metin benzerliği konusunda yine de iyi bir sınıflandırma performansı sunduğu belirtilmiştir.

Qurashi, Holmes ve Johnson [27] demiryolu güvenlik belgelerinin anlamsal benzerliğini hesaplamayı amaçlamıştır. Gösterimler Word2vec yöntemi kullanılarak elde edilmiştir. Benzerliği ölçmek için jaccard ve kosinüs uzaklık türleri tercih edilmiştir. Eşik değeri yüzde 80 olarak belirlenmiştir. Kosinüs, üstün performans göstermiştir. Zahrotun [28] benzerlik hesabında uzaklık türlerinin karşılaştırmasını sunmuştur. Jaccard benzerliği, kosinüs benzerliği ve jaccard ile kosinüs benzerliğinin bir kombinasyonu kullanılmıştır. Kullanılan yöntem, Shared Nearest Neighbor (SNN) ile belge kümelemedir. Kosinüs benzerliği diğer alternatiflerden daha iyi performans göstermiştir. Kiros, Zhu, Salakhutdinov, Zemel, Torralba, Urtasun ve Fidler [29] Skip-thoughts adı verilen bir model önermiştir. Önerilen model, cümleyi çevreleyen cümleleri tahmin etmek için kullanılmıştır. Öğrenilen vektörlerin anlamsal benzerlik performansı gösterilmiştir. Kenter ve Rijke [30] cümlelerin benzerliğini hesaplamak için evrişimli sinir ağlarına dayalı bir model önermiştir. Önerilen model, birden fazla ayrıntı türü, pencere boyutu ve havuzlamaya sahip evrişim filtreleri içermektedir. Cümle benzerliği birkaç ayrıntı düzeyinde hesaplanmıştır. Imtiaz et al. [31] yinelenen soruları tespit etmek için çalışmıştır. Google haber vektörü temsili, FastText tarama temsili ve FastText tarama alt kelime temsili içeren üç tür kelime temsili kullanılmıştır. Daha sonra bu üç temsilin ortalaması alınmıştır. Yinelenen soruları tespit etmek için Manhattan Uzun Kısa Süreli Bellek (Manhattan Long Short Term Memory - MaLSTM) modeli kullanılmıştır. Eşik değeri 0.5 olarak ayarlanmıştır. Srinidhi et al. [32] duyguları sınıflandırmak için MaLSTM ve Support Vector Machine (SVM) yöntemlerini birleştiren bir hibrit model önermiştir. Uzun Kısa Süreli Bellek (Long Short Term Memory - LSTM) modeli kullanılarak gizli gösterimler elde edilmiştir. Ardından SVM yöntemiyle duygular sınıflandırılmıştır. Guo, Zeng, Duan, Ni ve Liu [33] Çin acil müdahale planlarının benzerliğini hesaplamıştır. Planların vektörlerini elde etmek için Ts2vec, Word2vec ve VSM yöntemleri kullanılmıştır. Planlar arasındaki benzerliği ölçmek için kosinüs benzerliği kullanılmıştır. Ts2vec, Word2vec ile VSM yöntemlerine kıyasla daha iyi performans göstermiştir. Vine, Zuccon, Koopman ve Sitbon [34] tıbbi kavramlar arasındaki anlamsal benzerliği hesaplamayı amaçlamıştır. Tıbbi kavramların anlamsal benzerliği, Skip-gram modeli kullanılarak hesaplanmıştır. Pencere boyutunun artırılması model performansının artmasına katkıda bulunmuştur. Thyagarajan ve Mueller [35] bir Siyam Long Short Term Memory (LSTM) ağı sunmuştur. Değişken uzunlukta dizi çiftlerinden oluşan cümlelerin anlamsal benzerliğini hesaplamayı amaçlamışlardır. LSTM'lere kelime temsilleri sağlanmıştır. Model tarafından öğrenilen cümle temsillerinin benzerliği, manhattan uzaklığı kullanılarak hesaplanmıştır. Sonuçlara göre, LSTM'ler karmaşık anlayış gerektiren görevleri

yerine getirebilmiştir. Manhattan uzaklığı, kosinüs benzerliği gibi diğer uzaklık türlerinden daha iyi performans göstermiştir. Zhu, Yao, Ni, Wei ve Lu [36] cümlelerin benzerliğini hesaplamak için Bağımlılığa Dayalı Uzun Kısa Süreli Bellek (Dependency-based Long Short Term Memory - D-LSTM) 'i önermiştir. D-LSTM cümle bağımlılığını kullanmaktadır. Önerilen model, temel bileşen ve destekleyici bileşen içermektedir. Bunlar cümle temsillerinin bir parçasıdır. Destekleyici bileşenin eklenmesi, cümle temsilinin performansını geliştirilebileceği gösterilmiştir. Tai, Socher ve Manning [37] film incelemelerinin anlamsal benzerliğini hesaplamak için Tree-LSTM'i önermiştir. Önerilen model, LSTM'lerin ağaç yapılı ağ topolojilerine genelleştirilmesidir. Ghasemi ve Momtazi [38] benzer kullanıcıları bulmak için bir model önermiştir. Tavsiye sistemlerini iyileştirmeyi amaçlamışlardır. Kullanıcıların incelemelerine ve puanlarına dikkat etmişlerdir. Benzerliği ölçmek için kosinüs benzerliği kullanılmıştır. LSTM yöntemi ile en iyi sonuçlar elde edilmiştir.

Bilet benzerliğinin hesaplanması konusunda son yıllarda birçok çalışma yapılmıştır. Lekshmy, Anusree ve Varunika [39] akıllı bir yardım masası sistemi sunmuştur. Yeni bilete anlamsal olarak benzer olan biletlerin bulunması amaçlanmıştır. Benzer biletleri tespit etmek için Wordnet yöntemi seçilmiştir. Anlamsal olarak en benzer biletleri belirlemek için genetik algoritma kullanılmıştır. Her kümeden en benzer biletler belirlenmiştir. Apriyanto, Rahmawati, Setiawan, Budi ve Haryono [40] kullanıcılara iyi bir hizmet sunacak ve müşteri hizmetleri temsilcileri ile kullanıcılar arasındaki iletişimi kolaylaştıracak bir yardım masası sistemi önermiştir. Her birine en yakın mesafede olan biletleri belirlemek için Nearest Neighbor Retrieval algoritması kullanılmıştır. Niteliklere göre biletlerin benzerliği hesaplanmıştır. Samarakoon, Kumarawadu ve Pulasinghe [3] bir soru cevaplama sistemi önermiştir. Yardım masası sistemini otomatikleştirmeyi amaçlamışlardır. Önerilen sistem, bağlamı ve kelime sırasını dikkate almıştır. Parametreleri güncellemek için bir genetik algoritma kullanılmıştır. Değerlendirme için mean reciprocal rank kullanılmıştır. Liu et al. [41] bakım ekipmanlarının kritik bilgileri ile tarihi sahneler arasındaki benzerliği hesaplamayı amaçlamıştır. Çok değişkenli benzerliğin hesaplanması için bir yöntem önerilmiştir. Cümle temsilleri, TF-IDF kullanılarak elde edilmiştir. Benzerliği ölçmek için kosinüs benzerliği kullanılmıştır. Wang, Li, Zhu ve Gong [42] akıllı bir yardım masası sistemi önermiştir. Bu akıllı sistem, yeni biletle anlamsal benzerliğine göre geçmiş biletleri arar ve sıralar. Benzer biletler farklı kümeler halinde gruplandırılır. Her vaka grubundan bir öneride bulunulur.

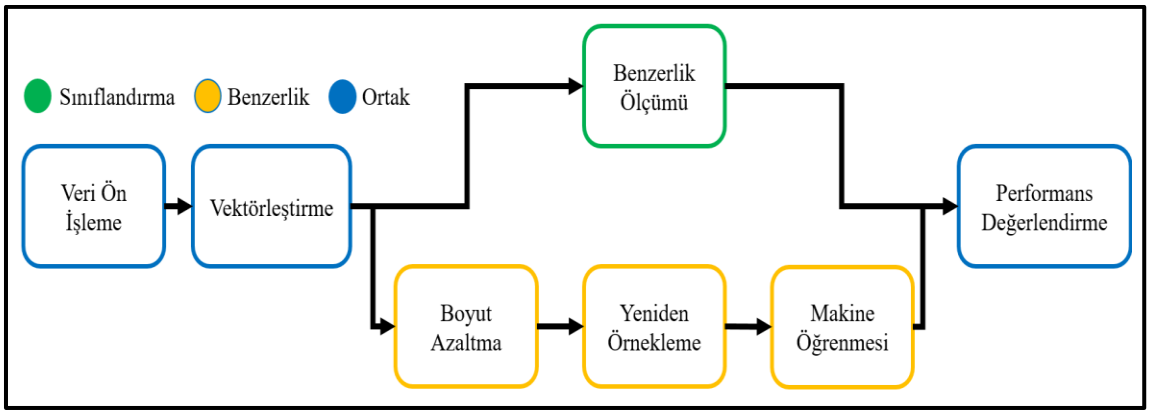
Literatürde doğal Türkçe dili içeren yardım masası bileti üzerinde yapılan çalışmaların sayısının az olması dikkat çekicidir. Son yıllarda yapılan araştırmalarda ağırlıklı olarak yardım masası bileti kategorisinin sınıflandırılmasına odaklanıldığı belirlenmiştir. Türkçe biletlerin benzerliği üzerine herhangi bir çalışma yapılmadığı, Türkçe biletlerin önceliğinin tahmini gibi alternatif konularda az sayıda çalışmanın olduğu gözlemlenmiştir [2]. Bilet kategorisi sınıflandırmasının yapıldığı çalışmalarda, doküman gösterimlerinin sıklıkla TF-IDF yöntemi kullanılarak elde edildiği ve geleneksel makine öğrenmesi yöntemleri kullanılarak sınıflandırıldığı tespit edilmiştir. Araştırma çalışmalarında, benzer türdeki biletleri gruplamak için kümeleme yöntemlerinin kullanılması [18], bilet yorumlarının kullanılmasının sınıflandırma doğruluğunu iyileştirdiği [16] ve geleneksel sınıflandırıcılara kıyasla sınıflandırıcılar topluluğunun iyi performans gösterdiği [21] öne çıkmıştır. Metin benzerliği konusunda birçok çalışmada LSTM yönteminin ve onun türevlerinin tercih edildiği görülmüştür.

3. TEORİK ALTYAPI

Biletleri sınıflandırmak ve benzerliğini ölçmek için Şekil 3.1'de görüldüğü gibi çeşitli görevleri yerine getirmek gerekir. Öncelikle bilet sınıflandırma ve benzerlik performansını doğrudan etkileyen veri ön işlem adımları uygulanır. Böylece veri kümesi sayısal vektörler elde etmek için hazır hale gelir. Bilet açıklamalarının doğal dil içermesi sebebiyle sınıflandırma ve benzerlik görevleri için sayısal vektörlere dönüştürülmesi gerekir. Bu kapsamda TF-IDF, Word2vec, Doc2vec ve LSTM gibi vektörleştirme yöntemleri uygulanır.

Sayısal vektörlerin yüksek boyut içermesi başarıyı etkilemektedir. Dolayısıyla Temel Bileşen Analizi gibi boyut azaltma yöntemleri kullanılarak sayısal vektörlerin boyutu azaltılır. Veri kümesinin dengesiz olması modellerin genelleme yapma performansını olumsuz yönde etkiler. Yeniden örnekleme yöntemi başta olmak üzere örnekleme yöntemleri kullanılarak yukarı ve aşağı örnekleme gerçekleştirilir, sınıflar dengeli hale getirilir. Makine öğrenmesi yöntemleri sınıflandırma görevi için sıklıkla kullanılmaktadır. SVM ve K-NN, metinlerin sınıflandırılmasında yaygın olarak kullanılan makine öğrenmesi algoritmalarıdır. Bilet benzerliği görevinde sayısal vektörler elde edildikten sonra biletlerin anlamını temsil eden vektörlerin benzerliğini hesaplamak için aradaki uzaklık ölçülür. Uzaklık ölçümünde kosinüs, öklid ve manhattan uzaklık türleri yaygın olarak kullanılmaktadır.

Sınıflandırma ve benzerlik görevlerinde performans ölçümünün doğruluğunu sağlamak için genellikle çapraz doğrulama uygulanır. Çapraz doğrulama sayesinde veri setindeki her parça; eğitim ve test olarak kullanılarak daha doğru ölçüm sağlanmış olur. Bilet sınıflandırma görevinde genellikle başarı, kesinlik, duyarlılık ve F1-Puanı metrikleri hesaplanmaktadır. Bilet benzerlik görevinde ise MSE, RMSE, MAE, PCC ve SRCC metrikleri hesaplanmaktadır.



Şekil 3.1. Sınıflandırma ve benzerlik görevleri için kullanılan yöntemlerin blok şeması

3.1. Veri Ön İşlemleri

Metin sınıflandırma ve benzerliği gibi çeşitli görevler için uygulanması gereken temel bileşenlerden biri veri ön işlemedir. Kapsamlı deneysel analizler alana ve dile bağlı olarak veri ön işleme görevlerinin sınıflandırma doğruluğunu önemli ölçüde artırabileceğini ortaya koymuştur [15]. Sınıflandırma doğruluğuna katkısının yanında performans artışı, veri boyutunun azalması ve eğitim süresinin kısalması gibi birçok faydası da bulunmaktadır.

Yardım masası bilet açıklamalarında doğal dil, URL ifadeleri, yazım hataları, noktalama işaretleri ve HTML ifadeleri bulunabilir. Dolayısıyla bilet açıklamalarını makine öğrenmesi modellerinin, sinir ağlarının ve vektörleştirme yöntemlerinin işleyebilmesi için veri ön işleme adımları uygulanır.

Bilet açıklamalarına uygulanan veri ön işleme adımları şu şekildedir:

- Bilet açıklamaları HTML ifadeleri ile birlikte veri tabanlarında tutulabilir. Bu ifadelerin cümleye anlamsal bir katkısı yoktur ve modelin performansını etkiler. Bu sebeple HTML ifadeleri kaldırılmalıdır.
- Hakkında sorun bildirilen veya talep yapılan sistemlerin URL ifadeleri biletlerde bulunabilir. Ancak, URL ifadeleri bilet açıklamasının anlaşılmasına fazla katkı sağlamaz. Bu, müşteri hizmetleri temsilcilerine sistem hakkında bir ipucu verilmesi olarak düşünülebilir. Bu nedenle, URL ifadelerini kaldırılmalıdır.
- Vektörleştirme yöntemlerin doğal dil içeren metinleri işleyebilmesi için kelimelere ayrılması gerekir. Bu kapsamda metinler kelimelere ayrılır.

- Noktalama işaretlerini kaldırmak, doğal dil metinleriyle çalışırken veri ön işlemenin bir parçasıdır. Modelin performansına katkıda bulunur. Bu sebeple kaldırılır.
- Bilet açıklaması yazım hataları ve alfabetik olmayan karakterler içerebilir. Bunlardan bazıları tek karakterli kelimelerdir. Bu nedenle, bu ifadeler kaldırılır.
- Etkisiz kelimeler, bir cümlenin anlamına fazla katkı sağlamaz. Bu nedenle, veri ön işlemenin bir parçası olarak kaldırılır.
- Kelimelerin köklerine indirgenmesi vektörleştirme yöntemine göre değişmekle birlikte belge temsil vektörünün boyutunun küçülmesine katkı sağlamaktadır.

3.2. Vektörleştirme

Doğal dil işleme çalışmalarından biri metni bir temsili dönüştürmektir. Yapılandırılmamış metin belgelerinin makine öğrenimi algoritmalarının işleyebileceği sayısal vektörler gibi uyumlu bir biçimde temsil edilmesi gerekir [14]. Bu nedenle, vektörleştirme yöntemleri kullanılarak belge gösterimlerinin elde edilmesi gerekir. Anlamsal ilişkileri çıkarmak ve benzerliği daha kesin olarak araştırmak için kelime temsilleri kullanılır [43]. Metin temsilinin metnin anlamını göstermesi ise son derece önemlidir [26].

En küçük metin parçası kelimelerdir. Bu sebeple kelimeler metnin temel öğesi olarak kabul edilir [33]. Kelimelerin temsillerini oluşturmak için sinir ağı tabanlı yöntemler etkilidir. Sinir ağı tabanlı dil modellerindeki gelişmeler sayesinde kelimelerin öğrenilmiş temsilleri, anlamları temsil edebilir [28]. Bu modeller bir fonksiyonun optimizasyonuna dayanır. Kelimenin anlamı, yakındaki kelime oluşumları tekrar tekrar gözlemlenerek öğrenilir [34]. Kelimelerin anlamsal uzayını oluşturmak için sinir ağı tabanlı modeller tarafından etiketlenmemiş metin verilerine ihtiyaç duyulur. Bu anlamsal uzayda benzer anlama sahip kelimeler birbirine yakın olur [30]. Kelime temsilleri yaklaşımları genellikle belge temsilleri olarak kabul edilen belge düzeyine genelleştirilir [25].

3.2.1. TF-IDF

TF-IDF, belgedeki kelimelerin önemini hesaplamak için kullanılır. Eş. 3.1'de görüldüğü gibi terim frekansının ters belge frekansı ile çarpılmasıyla hesaplanır. Bir terimin bir doküman içerisinde bulunma olasılığının $tf(i,j)$, toplam doküman sayısının (N) terimin geçtiği

doküman sayısına $df(i)$ bölümünün logaritmik değeri ile çarpımı sonucunda elde edilir. Terim ne kadar az dokümanda geçiyor ise geçtiği dokümanlar için o kadar önemli olur.

$$w(i,j) = tf(i,j) * \log(\frac{N}{df(i)}) \quad (3.1)$$

Aşağıda iki ayrı sistem için açılmış yardım masası bileti açıklamaları görülmektedir. Yardım masası bileti açıklamalarının veri ön işlem öncesi ve sonrası durumları gösterilmiştir;

Stok Sistemi Yardım Masası İsteği Açıklaması (Ön İşlem Uygulanmamış)

- Merhaba Rezervasyon Ekranları Çalışmıyor, Yardımcı Olun Stok.

Stok Sistemi Yardım Masası İsteği Açıklaması (Ön İşlem Uygulanmış)

- Merhaba Rezerv Ekran Çalış Yardım Ol Stok

Tedarik Sistemi Yardım Masası İsteği Açıklaması (Ön İşlem Uygulanmamış)

- Merhaba İrsaliye Ekranları Çalışmıyor, Yardımcı Olun Tedarik.

Tedarik Sistemi Yardım Masası İsteği Açıklaması (Ön İşlem Uygulanmış)

- Merhaba İrsaliye Ekran Çalış Yardım Ol Tedarik

Çizelge 3.1'de görüldüğü üzere stok sistemi ile ilgili önemli sözcükler rezerv ve stok olmuştur. Tedarik sistemi ile ilgili önemli sözcükler ise irsaliye ve tedarik olmuştur. Dolayısıyla makine öğrenmesi modeli bu yardım masası biletlerinin açıklamalarına bakarak başarılı bir şekilde sınıflandırma gerçekleştirebilir.

Çizelge 3.1. TF-IDF yöntemi ile vektör hesaplama örneği

Kelime	TF		IDF	TF × IDF	
	A	B		A	B
Merhaba	1/7	1/7	log(2/2) = 0	0	0
Rezerv	1/7	0	log(2/1) = 0,3	0,043	0
İrsaliye	0	1/7	log(2/1) = 0,3	0	0,043
Ekran	1/7	1/7	log(2/2) = 0	0	0
Çalış	1/7	1/7	log(2/2) = 0	0	0
Yardım	1/7	1/7	log(2/2) = 0	0	0
Ol	1/7	1/7	log(2/2) = 0	0	0
Stok	1/7	0	log(2/1) = 0,3	0,043	0
Tedarik	0	1/7	log(2/1) = 0,3	0	0,043

3.2.2. Word2vec

Dağılım hipotezine göre benzer bağlamdaki kelimeler anlamsal olarak benzerdir [27]. Bu nedenle, Word2vec dağılım hipotezi aracılığıyla kelimelerin anlamsal bilgilerini tespit eder [26]. Word2vec iki yönteme sahiptir: Atlama Gram (Skip Gram - SG) ve Sürekli Kelime Torbası (Continuous Bag of Words - CBOW). SG yönteminde hedef sözcük verilerek bağlam sözcükleri tahmin edilirken, CBOW yönteminde bağlam sözcükleri verilerek hedef sözcük tahmin edilir. Pencere büyüklüğü hiper parametresi, bağlam sözcüklerinin sayısını belirler. Word2vec modelinin giriş ve çıkış katmanında veri kümesindeki benzersiz kelime sayısı kadar nöron bulunur. Giriş nöronu dışındaki diğer nöronlar sıfır değeri ile pasif hale getirilir. Gizli katmandaki nöron sayısı deneysel olarak belirlenir. Bu katmandaki nöron sayısı, kelimeleri temsil eden kelime vektörlerinin boyutunu belirler. N kelime ve B kelime vektör boyutundan oluşan gizli katman bağlantılarının girişi olan NXB boyutlu matristeki i. vektör, i. kelimenin temsilidir.

3.2.3. Doc2vec

Belge temsillerini elde etmek için yaygın bir yaklaşım sözcük temsillerinin ortalamasını almaktadır. Ancak bu yaklaşım, kelime sırasının anlama olan katkısını kaybeder. Metindeki kelimelerin sırası önemli bilgiler içermektedir [26]. Kelime temsillerinin ortalaması alındığında cümlenin anlamı kaybolur [38]. Metinlerin anlamını yakalamada Doc2vec yöntemi Word2vec'ten daha iyidir [26]. Doc2vec yöntemi her belge için bir vektör oluşturur. Word2vec'in aksine kelime sırası kaybedilmez. Böylece kelimelerin diziliminden yararlanarak metnin anlamı daha iyi yakalanır. Doc2vec, Dağıtılmış Kelime Çantası (Distributed Bag of Words - DBOW) ve Dağıtılmış Bellek Paragraf Vektörü (Distributed Memory Paragraph Vector - DMPV) olarak ikiye ayrılır [38]. DMPV modeli kelimelerin hangi bağlamda bulunduğunu öğrenir. Kelime vektörü ve belge kimliği modele girdi olur. Belge kimliği ve kelime vektörlerinin vektör uzayları farklıdır. Bir belgenin eğitim sürecinde belge kimliği değişmez. Tahmin sürecinde belgeye yeni bir belge kimliği atanır. Bu işlemde kelime vektörü ve çıktı katmanı softmax parametrelerinin değişmesine izin verilmez. Hata yakınsanır ve belge vektörü elde edilir. Vektör modelinin amacı, Eş. 3.2'de gösterildiği üzere ortalama logaritmik olabilirliği maksimum yapmaktır. W_t, eğitim kelimeleri dizesinin t indeksindeki kelimeyi tanımlar. P, olasılığı gösterir. Pencereye boyutu ise k ile ifade edilir.

$$\text{Ortalama Logaritmik Olabilirlik} = \frac{1}{T}\sum_{t=k}^{T-k} \log P(W_t \mid W_{t-k}, \dots, W_{t+k}) \tag{3.2}$$

Çoklu sınıflandırıcı kullanılarak tahmin görevleri Eş. 3.3'de görüldüğü gibi gerçekleştirilir [26].

$$p(W_t \mid W_{t-k}, \dots, W_{t+k}) = (e^{y_{w_t}}) / \sum_i e^{y_i} \tag{3.3}$$

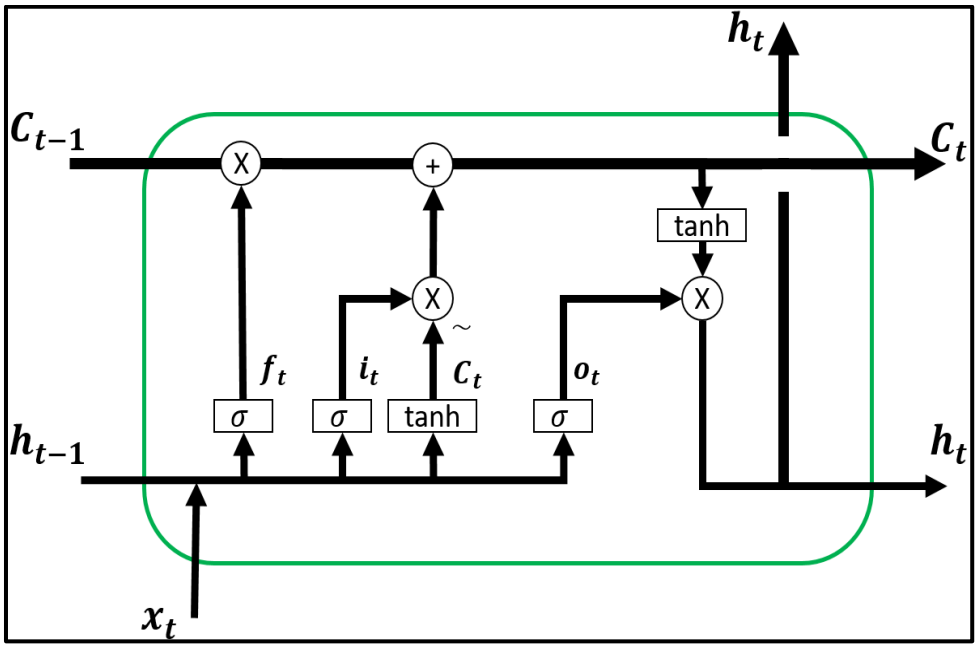
Eş. 3.4'de gösterildiği üzere y_i, her çıktı sözcüğü i için normalleştirilmemiş log olasılığıdır. U ve b, softmax parametreleridir ve h, kelime temsillerinin ortalamasıdır [26]. yi'nin her biri, şu şekilde hesaplanan her çıktı sözcüğü i için normalleştirilmemiş log-olasılığıdır:

$$y = b + Uh(W_{t-k}, \dots W_{t+k}; W) \tag{3.4}$$

3.2.4. LSTM

İleri beslemeli ağların giriş ve çıkışlarının bağımsız olması metinlerin sıralı davranışını yakalayamamasına neden olur. Yinelemeli Sinir Ağı (Recurrent Neural Network – RNN) ise bu sorunun üstesinden gelen bir mantık sunar. Ancak, RNN uzun vadeli bağımlılıkları öğrenemez [38]. LSTM ise bu sorunun üstesinden gelir. LSTM girdi olarak aldığı dizilerdeki bilgileri depolayabilir ve bunlara erişebilir. Bu sayede uzun vadeli bağımlılıkları öğrenir. Ancak, cümle temsillerinin uzayını öğrenmek için LSTM 'in yeterli miktarda veriye ihtiyacı vardır [35]. LSTM girdi olarak cümlelerin vektör temsillerini kullanır ve çıktı olarak cümlelerin anlamsal anlamını kodlayan gizli bir durum üretir.

LSTM ünitesi Şekil 3.2'de gösterilmiştir. LSTM ünitesinin üç çeşit kapısı vardır. Bunlar giriş, unutma ve çıkış kapılarıdır. Bu üç kapı ve hücre belleği hangi bilgilerin kullanılacağına ve modelden çıkarılacağına karar verir [31]. LSTM'in giriş verileri x_t, çıkış verileri h_t, bellek birimi ise c_t ile gösterilir. W, b ve σ sırasıyla ağırlık matrisi, kapı ünitesinin yanlılığı ve aktivasyon fonksiyonudur [44]. Sigmoid işlevi hangi değerin etkileneceğine karar verir.



Şekil 3.2. LSTM Hafıza Hücresi

Giriş kapısı, c_t bellek birimine ne kadar giriş verisinin (x_t) aktarılacağını belirler [44]. Giriş kapısının aktivasyon vektörü (i_t) Eş. 3.5' de ifade edilmiştir.

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i) \tag{3.5}$$

Bazı özellikleri göz önünde bulundurmak yararlı olmayabilir. Bu nedenle, bazı özellikleri yok saymak için unutma kapısı kullanılır [44]. Unutma kapısının aktivasyon vektörü (f_t) Eş. 3.6'da gösterilmiştir. Unutma kapısı, özellikleri göz ardı eder ve Eş. 3.7'de görüldüğü gibi bellek birimini (c_t) değiştirir.

$$f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f) \tag{3.6}$$

$$c_t = ft * c_{t-1} + i_t * \tilde{C}_t \tag{3.7}$$

Çıkış kapısı, bellek birimini (c_t) ve giriş verilerini (x_t) kullanarak çıkış verilerini tanımlar. Çıkış kapısının aktivasyon vektörü (o_t) Eş. 3.8'de ifade edilmiştir. Çıkış verileri (h_t), giriş verileri (x_t) ile önceki ağın bellek birimini (c_{t-1}) birleştirir [44]. Gizli durum vektörü olarak da bilinen çıkış verileri (h_t) Eş. 3.9'da ifade edilmiştir.

$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o) \tag{3.8}$$

$$h_t = o_t \times \tanh(C_t) \tag{3.9}$$

3.3. Temel Bileşen Analizi

Metinlerin vektör uzayı gösterimi genellikle yüksek boyutluluk ile sonuçlanır. Bu durum özellikle çok sayıda sınıf etiketi olduğunda ve her biri için yetersiz eğitim verileri varken bu büyük bir zorluktur [4]. Uzay problemini hafifletmeye yönelik bir yaklaşım boyut azaltma tekniklerini kullanmaktır [45]. Boyut azaltma, özellikleri daha düşük boyutta bir temsile dönüştürerek veya gereksiz özellikleri kaldırarak sınıflandırma görevini basitleştirme sürecidir [46]. Modelin eğitim süresini kısaltmak ve performansını artırmak için boyut azaltma yöntemi kullanılır. Temel Bileşen Analizi (Principal Component Analysis - PCA) metin sınıflandırmasında yaygın olarak kullanılmaktadır [45] ve iyi bilinen bir boyut azaltma yöntemidir [46]. PCA, girdi özellik alanını bileşenin ilişkisiz olacağı yeni bir alana yansıtır. Ayrıca bilgiyi orijinal uzayda mevcut olan bilgilerin çoğunu içeren küçük çıktı özelliklerine sıkıştırır [46]. PCA'da verilerin kovaryans matrisini bulmak ve gizli kalıpları ortaya

çıkarmak için gürültülü ve gereksiz bilgileri filtreleyerek özdeğerleri ve özvektörleri çıkarmak gerekir [43]. Kovaryans Eş. 3.10'da görüldüğü gibi iki değişken arasındaki ilişkiyi gösterir. Eş. 3.10'da görülen X', X değişkenlerinin ortalamasını temsil eder. Y', Y değişkenlerinin ortalamasını temsil eder. N ise veri sayısıdır. Kovaryans matrisi, Eş. 3.11 ile hesaplanır.

$$\text{Kovaryans} = \frac{\sum_{i=1}^{N} (Xi - X')(Yi - Y')}{N - 1} \tag{3.10}$$

$$\text{Kovaryans Matris} = \begin{bmatrix} \text{Var}(x) & \text{Kov}(x,y) & \text{Kov}(x,z) \\ \text{Kov}(x,y) & \text{Var}(y) & \text{Kov}(y,z) \\ \text{Kov}(x,z) & \text{Kov}(y,z) & \text{Var}(z) \end{bmatrix} \tag{3.11}$$

Eş. 3.12 'deki A, λ, ve I sırasıyla kovaryans matrisi, özdeğeri ve birim matrisi temsil eder. Özdeğerler kullanılarak özvektörler Eş. 3.13 ile hesaplanır ve büyüklüklerine göre sıralanır. Eş. 3.13'deki v, özvektörü temsil eder. Eş. 3.14 ile veri boyutu k boyutuna indirgenir. Eş. 3.14'de görülen X^i, veri kümesindeki örneği temsil eder. Vi^T, i'inci özvektörün devriğini temsil eder. N'inci özdeğer n'inci ana bileşeni tanımlar. Birinci temel bileşen en yüksek varyans yüzdesine sahiptir [43].

$$\det(A, \lambda I) = 0 \tag{3.12}$$

$$Av = \lambda v \tag{3.13}$$

$$\text{Xyeni}^i = \begin{bmatrix} V1^T X^i \\ V2^T X^i \\ \dots \\ VN^T X^i \end{bmatrix} \tag{3.14}$$

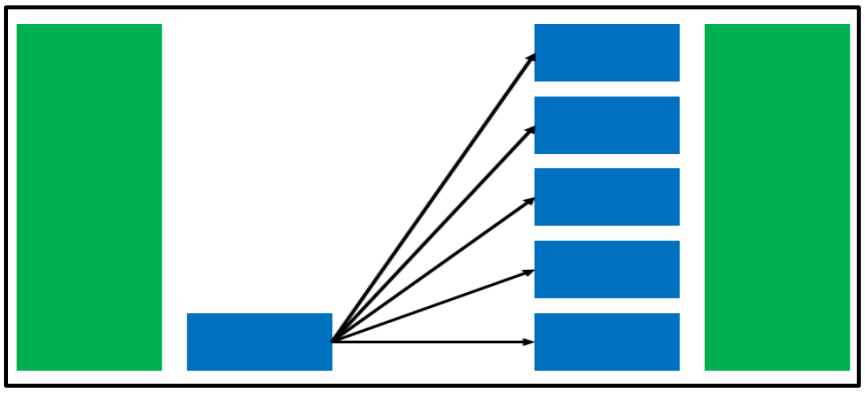
3.4. Yeniden Örnekleme

Yeniden örnekleme, sınıflar arasında eşit olmayan örnek dağılımı olan veri kümelerini dengelemek için kullanılır. Eşit oranlarda sınıfların örneklerinden oluşan bir veri kümesi ile eğitilen model kullanılarak yapılan sınıflandırma tahmini güvenilirdir. Ancak, örneklerin oranında bir sapma varsa dengesiz sınıflandırmanın oluşması nedeniyle sonuç yanlış

olacaktır [47]. Dengesiz veri kümeleriyle uğraşırken sınıflandırıcı yüksek global doğruluk elde etse bile azınlık sınıfına ait örneklerin tanımlanması yaygın bir durum değildir.

Verilerdeki dengesizliği gidermek için çoğunluk ve azınlık örneklerin sayısını ayarlayan Yeniden Örnekleme kullanılır. Yeniden örnekleme yöntemi ikiye ayrılır: Yetersiz Örnekleme ve Aşırı Örnekleme. Yetersiz Örneklemede çoğunluk sınıfa ait örnekler kaldırılırken, aşırı örneklemede azınlık sınıfından bazı örnekler çoğaltılır veya sentetik olarak oluşturulur [48]. Çeşitli çalışmalarda bilet sınıflandırmasının doğruluğunu artırmak için Aşırı Örnekleme ve Yetersiz Örnekleme kullanılmıştır [49].

Rastgele Aşırı Örnekleme Şekil 3.3'de görüldüğü gibi azınlık sınıfına ait örneklerin rastgele seçimi ile veri kümesine eklenmesidir. Rastgele Aşırı Örneklemede kullanılan örneklerin sayısı küçükse aynı azınlık örneklerinden birden fazla örnek çoğaltıldığında oluşturulan örnekler fazla uyum ile sonuçlanabilir. Ayrıca Rastgele Aşırı Örnekleme, eğitim verisi sınıflarında gürültüyü veya çakışmayı artırabilir [43, 66].



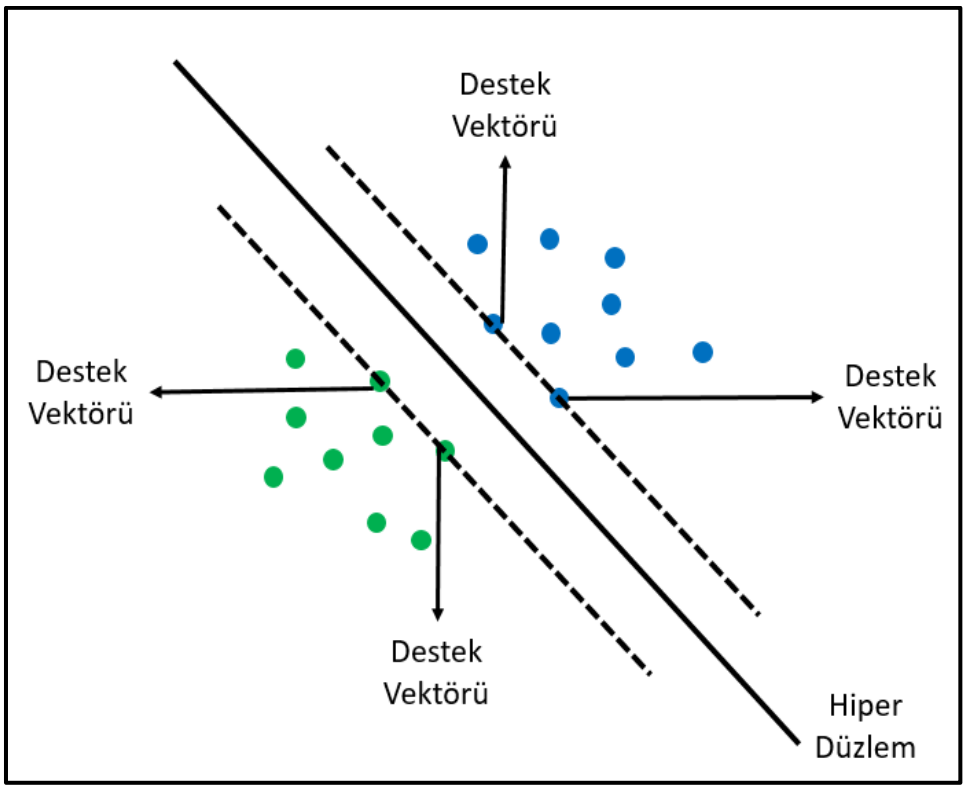
Şekil 3.3. Yukarı Örnekleme yöntemi

3.5. Makine Öğrenmesi

Makine öğrenmesi yöntemleri; sınıflandırma, regresyon ve kümeleme amaçları için sıklıkla kullanılmaktadır. SVM ve K-En Yakın Komşu (K-Nearest Neighbors - K-NN), en yaygın kullanılan makine öğrenmesi algoritmalarındandır.

SVM, sınıflandırma ve regresyon görevlerinde hem doğrusal hem de doğrusal olmayan problemleri çözmek için kullanılan denetimli bir yöntemdir [50]. Yardım masası bilet alanı

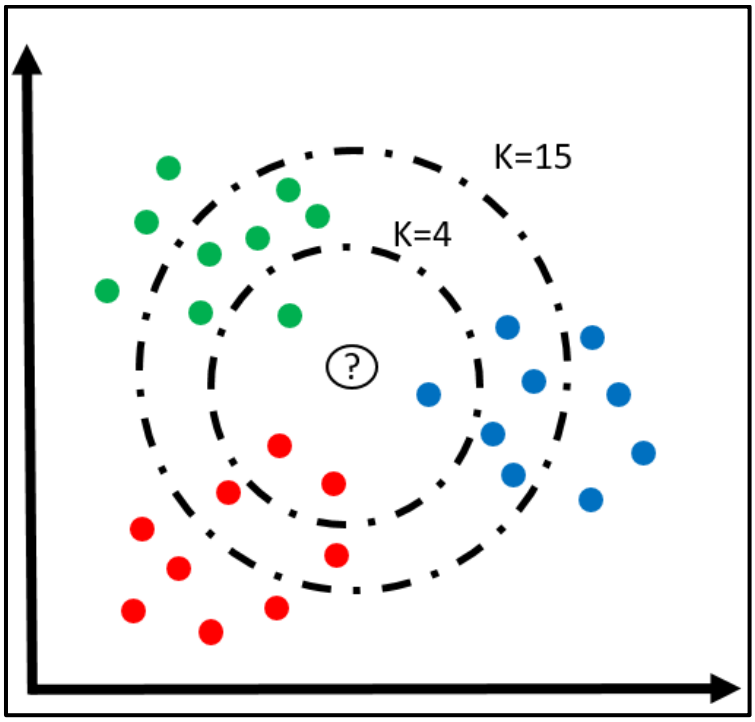
sınıflandırması için çeşitli çalışmalarda SVM yöntemi kullanılmıştır [25, 26, 27] ve metin sınıflandırmada en iyi performansı göstermektedir [49]. SVM yöntemi Şekil 3.4'de görüldüğü gibi diğer öğrenen yöntemlerden farklı olarak sınıflar arasındaki farklılıkları öğrenmenin aksine sınıflar arası benzerlikleri öğrenerek ve onları destek vektörleri olarak kullanarak iki farklı sınıfı ayıran optimum hiper düzlemi bulmaktadır [51]. Optimum düzlemler karar düzlemi olarak da adlandırılabilir. Alt kümelerdeki her bir öğenin benzer özelliklere sahip olacağı şekilde verileri alt kümelere böler [52]. SVM, yüksek boyutlu bir özellik uzayında olağanüstü genelleme yeteneği ile bilinir [51, 65].



Şekil 3.4. SVM yöntemi

K-NN, gelecekteki değerleri tahmin etmek için geçmiş bir veri tabanındaki en benzer K tane özniteliği arayan parametrik olmayan bir tahmin algoritmasıdır [53]. Basit ve etkili bir metin sınıflandırma yöntemi olan K-NN yönteminin, iyi bir doğruluk ve duyarlılık oranına sahip olması onu metin sınıflandırmada yaygın olarak kullanılan sınıflandırıcılardan biri haline getirmektedir [54]. Şekil 3.5'de görüldüğü gibi K-NN yönteminde K değeri belirlendikten sonra örneğin en yakın k komşusu tespit edilir. Ey yakın k tane komşunun çoğunluğunu oluşturduğu sınıf örneğin sınıfı olur. Dolayısıyla K-NN bir tür örnek tabanlı öğrenme uygular [55]. K-NN yönteminde örnekler arası mesafe hesaplanırken Öklid mesafesi gibi bir

ayırma metriği kullanılır [56]. K parametresi için en uygun değer ise deneysel olarak belirlenebilir. Minimum hata oranını veren k değeri seçilir [56].



Şekil 3.5. K-NN yöntemi

3.6. Benzerlik Ölçüleri

Metinlerin benzerliğini hesaplamak üzerine gerçekleştirilen çalışmalarda yaygın olarak kosinüs [27, 28, 33, 35, 36, 41], manhattan [35] ve öklid uzaklıkları kullanılmıştır. Bu sebeple, bu çalışmada bilet çiftlerini temsil eden vektörler arasındaki benzerliği hesaplamak için kosinüs, manhattan ve öklid uzaklıkları kullanılmıştır.

Kosinüs uzaklığı, Eş. 3.15'de görüldüğü gibi vektörler arasındaki açının kosinüsünü hesaplar [35]. Ölçü vektör uzunluğundan bağımsız olduğu için yüksek boyutlu veriler üzerinde sıklıkla kullanılmaktadır [28]. Kosinüs benzerliği, Eş. 3.16'da görüldüğü gibi kosinüs uzaklığının bir eksisidir.

$$\text{Kosinüs Uzaklığı}(S1, S2) = \frac{(S1 * S2)}{||S1||||S2||} = \frac{\sum_{i=1}^{n} S1iS2i}{\sqrt{\sum_{i=1}^{n}(S1i)^2}\sqrt{\sum_{i=1}^{n}(S2i)^2}} \quad (3.15)$$

$$\text{Kosinüs Benzerliği}(S1, S2) = 1 - \text{Kosinüs Uzaklığı}(S1, S2) \tag{3.16}$$

Manhattan uzaklığı, Eş. 3.17'de görüldüğü üzere iki vektör arasındaki mutlak farkların toplamı olarak hesaplanır. Koordinat ekseninden örnek vermek gerekirse segmentlerin uzaklıkları toplanarak iki nokta arasındaki uzaklık hesaplanır.

$$\text{Manhattan Uzaklığı}(S1, S2) = \sum_{i=1}^{d} |S1_i - S2_i| \tag{3.17}$$

Öklid uzaklığı, Eş. 3.18'de görüldüğü gibi bir nokta ile başka bir nokta arasındaki tüm karesi alınmış farkların toplanması ve karekökünün alınmasıyla hesaplanır [57]. Öklid uzaklığını hesaplamak için koordinat noktaları ve pisagor teoremi kullanılır. Öklid uzaklığı formülü Eş. 3.19'da görüldüğü gibi kayıp fonksiyonların nan değeri almasını önlemek için f(euc_dist) olarak değiştirilmiştir.

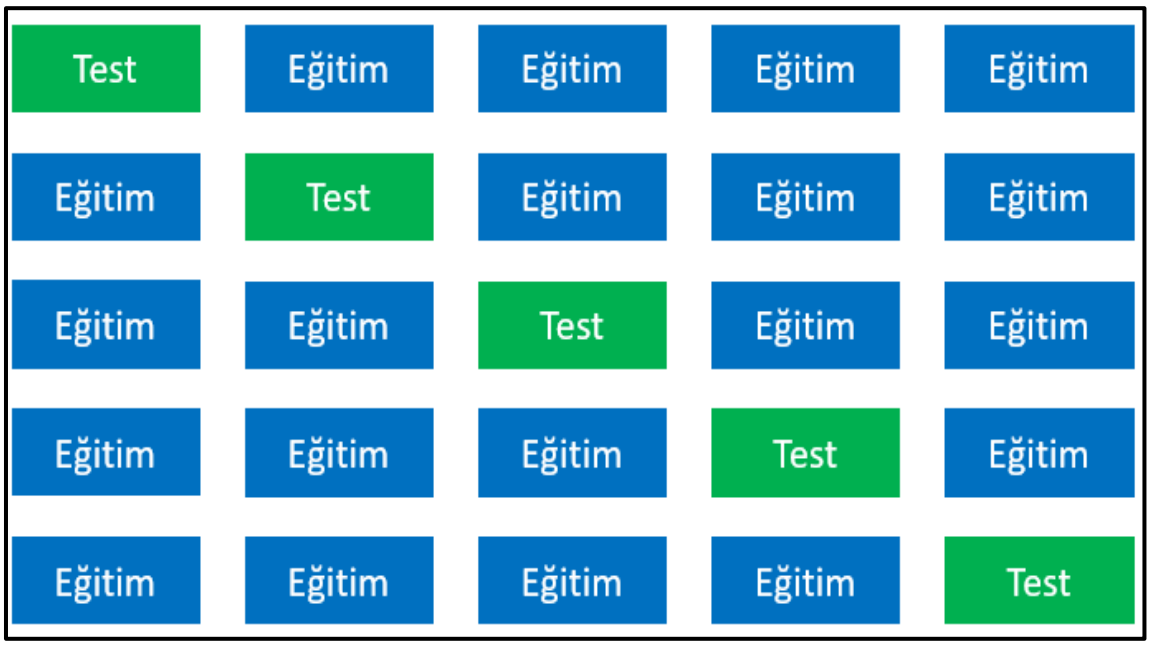
$$\text{Öklid Uzaklığı}(S1, S2) = \sqrt{\sum_{i=1}^{d} (S1_i - S2_i)^2} \tag{3.18}$$

$$f_{euc\_dist}(S1, S2) = \sqrt{\text{maksimum}(\sum_{i=1}^{d} (S1_i - S2_i)^2, 1 * 10^{-9})} \tag{3.19}$$

3.7. Performans Değerlendirme Metrikleri

Sınıflandırma modellerinin performansı, literatürde metin sınıflandırma için en sık kullanılan performans değerlendirme ölçütleri olan doğruluk, kesinlik, duyarlılık ve F1 puanı kullanılarak değerlendirilmiştir [12, 45]. Benzerlik tespit modellerinin performansı ise ortalama kare hata, kök ortalama kare hata, ortalama mutlak hata, Pearson korelasyon katsayısı ve Spearman'ın sıralama korelasyon katsayısı metrikleri kullanılarak değerlendirilmiştir. SVM ve K-NN için her metrik eğitim ve test ayrımın 10 kat çapraz doğrulamasının sonucunda hesaplanmıştır. LSTM için ise her metrik eğitim ve doğrulama ayrımın 5 kat çapraz doğrulamasının sonucunda hesaplanmıştır. Word2vec ve Doc2vec için her metrik eğitim ve test ayrımının 5 kat çapraz doğrulamasının sonunda hesaplanmıştır.

Deneysel çalışmalarda sınıflandırıcıların performansını değerlendirmek için çapraz doğrulama kullanılır. Şekil 3.6'da görüldüğü gibi çapraz doğrulama tekniğinde k değeri belirlenir ve veri kümesi k parçaya bölünür. Sırayla farklı parçalar test parçası olarak seçilir ve diğer parçalar eğitim parçası haline gelir. İşlem k kez tekrarlanır. K adet eğitimin değerlendirme metrik ortalamaları hesaplanır.



Şekil 3.6. Çapraz doğrulama tekniği

Doğruluk, sınıflandırıcı tarafından doğru tahmin edilen örnek sayısının toplam örnek sayısına oranıdır. Doğruluk oranı, gerçek pozitif ve gerçek negatif oranlarının toplamının, Eş. 3.20'de görüldüğü gibi gerçek pozitif, gerçek negatif, yanlış pozitif ve yanlış negatif oranlarının toplamına bölünmesiyle elde edilir.

$$\text{Doğruluk} = \frac{\text{GP} + \text{GN}}{\text{GP} + \text{GN} + \text{YP} + \text{YN}} \tag{3.20}$$

Kesinlik, doğru tahmin edilen örnekler arasındaki başarı oranını gösterir. Kesinlik, Eş. 3.21'de görüldüğü gibi gerçek pozitif oranın, gerçek pozitif ve yanlış pozitif oranlarının toplamına bölünmesiyle elde edilir.

$$\text{Kesinlik} = \frac{\text{GP}}{\text{GP} + \text{YP}} \tag{3.21}$$

Duyarlılık, doğru örnekleri tahmin etmedeki başarı oranını gösterir. Duyarlılık, Eş. 3.22'de görüldüğü gibi gerçek pozitif oranın, gerçek pozitif ve yanlış negatif oranlarının toplamına bölünmesiyle elde edilir.

$$\text{Duyarlılık} = \frac{\text{GP}}{\text{GP} + \text{YN}} \tag{3.22}$$

F1 puanı, kesinlik ve duyarlılık değerlerinin harmonik ortalamasıdır. F1 puanı, Eş. 3.23'de görüldüğü gibi kesinlik ve duyarlılık değerlerinin çarpımının kesinlik ve duyarlılık değerlerinin toplamına bölümünün iki katıdır.

$$\text{F1 Puanı} = 2\left(\frac{\text{Kesinlik} * \text{Duyarlılık}}{\text{Kesinlik} + \text{Duyarlılık}}\right) \tag{3.23}$$

Ortalama kare hata (Mean Squared Error - MSE), Eş. 3.24'de görüldüğü gibi hataların karelerinin ortalamasıdır. Hata, tahmini değer ile gerçek değer arasındaki farktır. MSE, karesi alınmış hata kaybının beklenen değerine karşılık gelen bir risk fonksiyonudur [36].

$$\text{MSE} = \left(\frac{1}{\text{n}}\right)\sum_{\text{i}=1}^{\text{n}}(\text{Y}_\text{i} - \text{Y'}_\text{i})^2 \tag{3.24}$$

Kök ortalama kare hata (Root Mean Squared Error - RMSE), Eş. 3.25'de görüldüğü gibi tahmin hatalarının standart sapmasıdır. RMSE, karesi alınmış hataların ortalamasının karekökünü ifade eder. RMSE, bu tahmin hatalarının ne kadar yayıldığının bir ölçüsüdür.

$$\text{RMSE} = \sqrt{\left(\frac{1}{\text{n}}\right)\sum_{\text{i}=1}^{\text{n}}(\text{Y}_\text{i} - \text{Y'}_\text{i})^2} \tag{3.25}$$

Ortalama mutlak hata (Mean Absolute Error - MAE), Eş. 3.26'da görüldüğü gibi mutlak hataların aritmetik ortalamasıdır. MAE, ölçümdeki hata miktarını ifade eder.

$$\mathrm{MAE} = \left(\frac{1}{n}\right) \sum_{i=1}^{n} (Y_i - Y'_i) \tag{3.26}$$

Pearson korelasyon katsayısı (Pearson Correlation Coefficient - PCC), Eş. 3.27'de görüldüğü gibi, iki değişken arasındaki doğrusal bir ilişkiyi gösterir. -1 ile +1 arasında değişir. Pozitif korelasyon bir değişkendeki değişimin diğer değişkende aynı yönde gerçekleşeceğini gösterir [58].

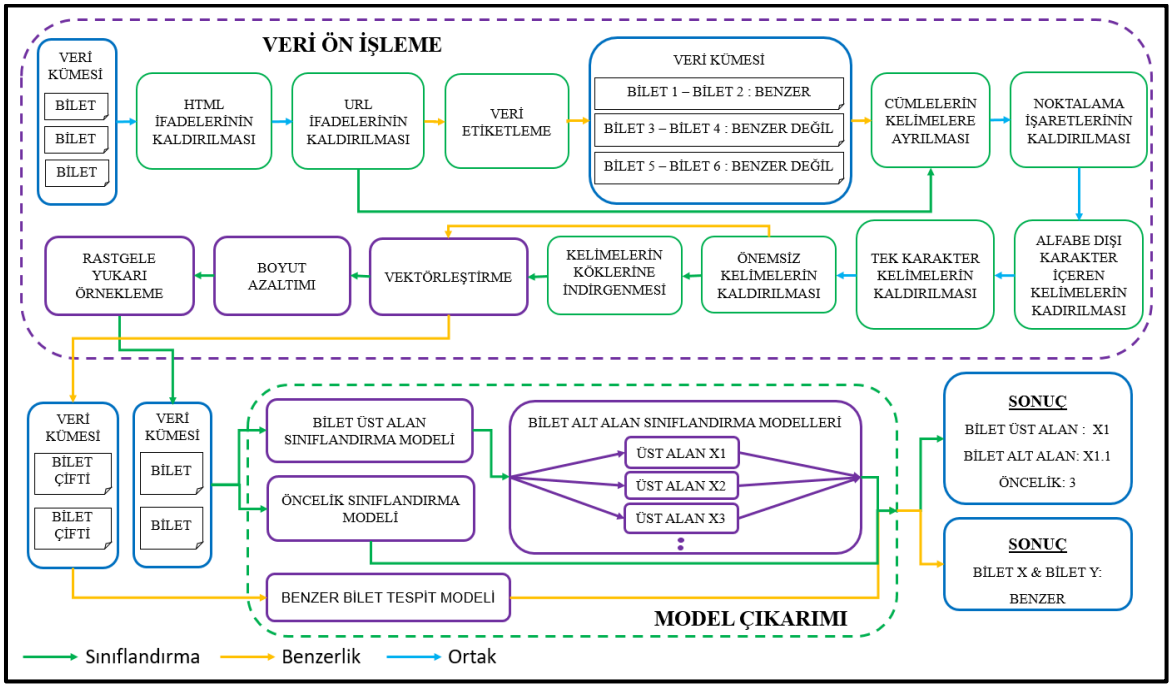
$$\mathrm{PCC} = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[(n\sum x^2) - (\sum x)^2][(n\sum y^2) - (y)^2]}} \tag{3.27}$$

Spearman'ın sıralama korelasyon katsayısı (Spearman's Rank Correlation Coefficient - SRCC), veriler iki değişken arasındaki normal dağılıma uymadığında uygulanır. Eş. 3.28'de görüldüğü gibi sıralanmış iki değişken arasındaki korelasyonu ölçer [59]. Pearson korelasyon katsayısından farklı olarak rastgele değişken yerine Spearman'ın sıralama korelasyon katsayısı kullanılmaktadır [60]. -1 ile +1 arasında değişir. Pozitif korelasyon bir değişkendeki değişimin diğer değişkende aynı yönde gerçekleşeceğini gösterir [61].

$$\mathrm{SRCC} = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{3.28}$$

4. DENEYSEL ÇALIŞMA

Etkili ve etkileşimli yardım masası sistemi tasarımı kapsamında bu araştırma çalışmasının genel yapısı Şekil 4.1'de gösterilmiştir. Bu çalışma kapsamında biletler üst ile alt alanlara ve önceliklerine göre sınıflandırılmış, müşteri hizmetleri temsilcileri tarafından kapatılan biletlerle aynı anlama gelen biletler tespit edilmiştir. Yardım masası biletini üst ve alt alanlara sınıflandırmak için hiyerarşik yaklaşım kullanılmıştır. Hiyerarşik yaklaşım, biletin ilk olarak sınıflandırılacağı üst alanı tahmin etmeye ardından öngörülen üst alanın kapsadığı alt alanı tahmin etmeye dayanır. Böylece çok sayıda alan içeren yardım masasında çok sınıflı sınıflandırmanın zorluğu aşılır.



Şekil 4.1. Önerilen mimari

4.1. Veri Kümesi

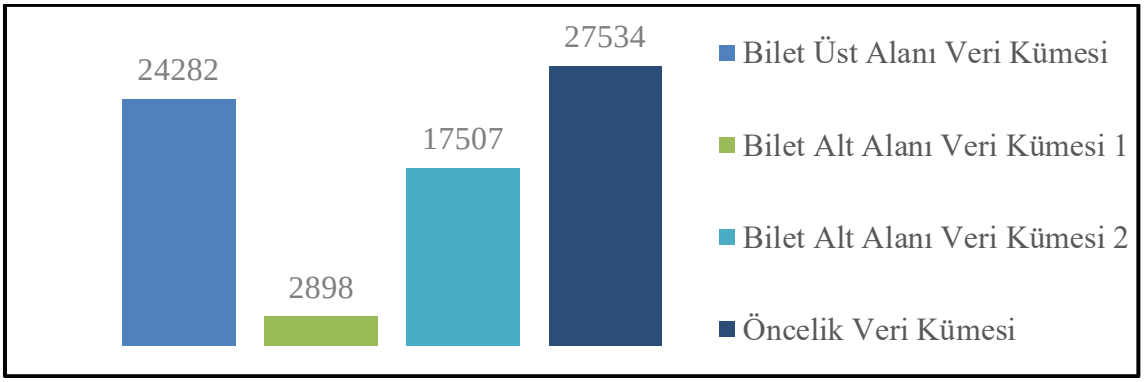
2021 yılının ikinci çeyreğinde bir şirketin yardım masası sisteminden Türkçe bilet açıklaması ile birlikte üst alan, alt alan ve öncelik sınıflarını içeren 4 adet veri kümesi toplanmıştır. Sadece kapalı durumdaki biletler veri kümesine dâhil edilerek bilet sınıflarının doğruluğundan emin olunmuştur. Veri kümesine ait örnek bir bilet açıklamasının doğal dil ve HTML ifadeleri içeren hali (B) ile HTML ifadeleri kaldırılmış ham hali (A) Şekil 4.2'de

gösterilmiştir. Veri kümesinin alındığı şirketin gizliliği nedeniyle örnek bilet açıklamasında URL ve isim değiştirme gerçekleştirilmiştir.

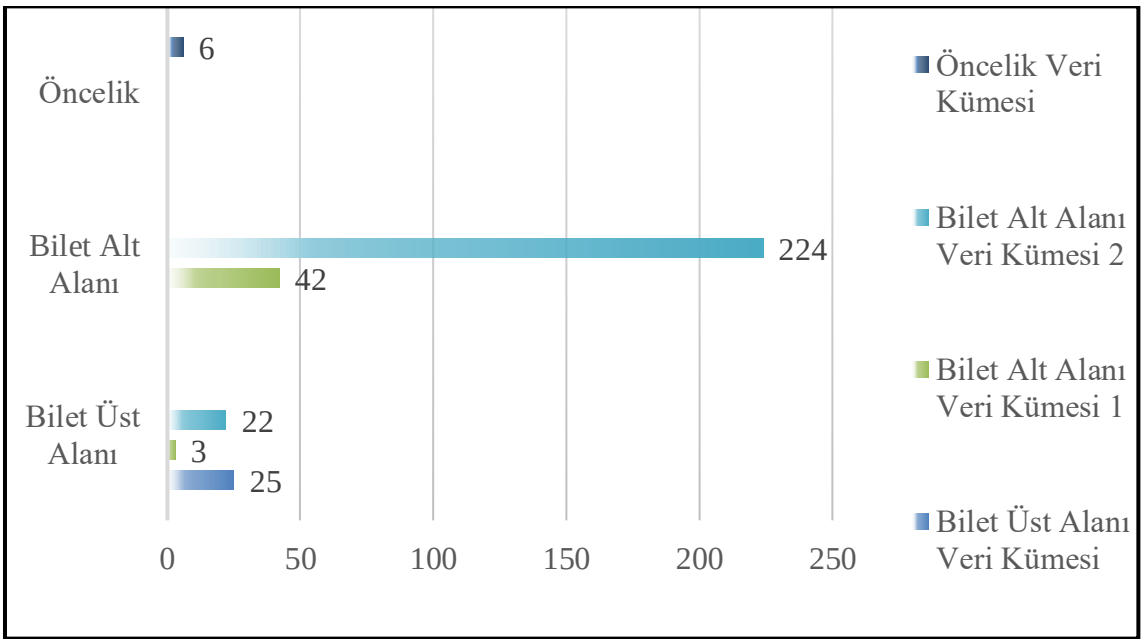
Bilet Açıklaması	
A	**HTML İfadeleri Kaldırılmış** https://url.com.tr/display/Bordro ve altındaki sayfalara Ali, Veli ve Mehmet Beylerin erişmesi gerekiyor, yetkiyi veremedim, aslında "No restriction" gözüküyor ama yine de erişemiyorlar.
B	**HTML İfadeleri Kaldırılmamış** https://url.com.tr/display/Bordro ve altındaki sayfalara Ali, Veli ve Mehmet Beylerin erişmesi gerekiyor, yetkiyi veremedim, aslında "No restriction" gözüküyor ama yine de erişemiyorlar.

Şekil 4.2. Örnek Bilet Açıklaması

Her bir veri kümesinin örnek sayısı Şekil 4.3'de, sınıf sayısı ise Şekil 4.4'de gösterilmiştir. Bilet üst alanını belirleyecek makine öğrenmesi modelinin eğitiminde bilet üst alanı veri kümesi kullanılmıştır. Veri kümesi, 24282 bilet ve 25 bilet üst alanı içerir. Bilet alt alanı veri kümesi 1 ile bilet alt alanı veri kümesi 2, biletin alt alanını belirleyecek makine öğrenmesi modellerinin eğitiminde kullanılmıştır. Bilet alt alanının sistem veya modül olarak farklılaşması nedeniyle iki ayrı veri kümesi kullanılmıştır. Bilet alt alanı veri kümesi 1, 2898 bilet, 3 bilet üst alanı ve 42 bilet alt alanı içermektedir. Bilet alt alanı veri kümesi 2, 17507 bilet, 22 bilet üst alanı ve 224 bilet alt alanı içermektedir. Öncelikli veri kümesi, biletin önceliğini belirleyecek makine öğrenmesi modelinin eğitiminde kullanılmıştır. Öncelik veri kümesi, 27534 bilet ve 6 öncelik içerir.



Şekil 4.3. Veri kümeleri örnek sayıları



Şekil 4.4. Veri kümeleri sınıf sayıları

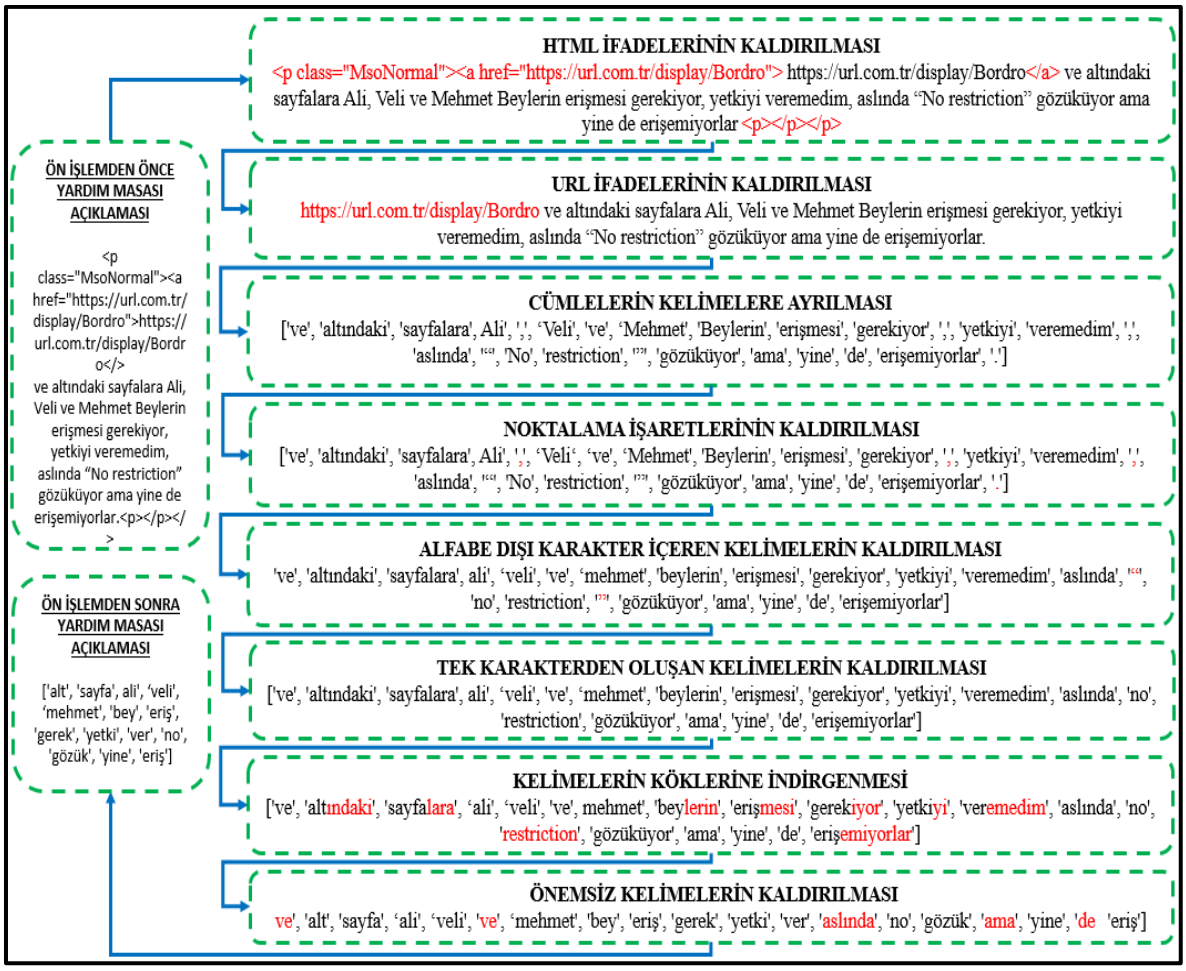
Bilet üst alanı veri kümesinden bir bilet üst alanına ait olan 267 bilet ayrılmış, kartezyen olarak eşleştirilmiş ve her bir bilet çiftinin benzer olup olmadığı manuel olarak etiketlenmiştir. Biletlerin eşleştirilmesi ve etiketlenmesi ile oluşturulan veri kümesi, 48206 bilet çifti içermektedir. Biletler, Word2vec, Doc2vec ve LSTM modellerini eğitmek ve değerlendirmek için kullanılmıştır. Veri kümesi, Word2vec ve Doc2vec yöntemleri için eğitim ve test olarak yüzde 80, 20 oranında bölünmüştür. Veri kümesi, LSTM yöntemi için ise eğitim, doğrulama ve test olarak yüzde 70, 10, 20 oranında bölünmüştür. Veri kümesine ait örnek bilet çiftleri Çizelge 4.1'de görülmektedir. Veri kümesinin elde edildiği firmanın gizliliği nedeniyle Çizelge 4.1'deki örnek bilet açıklamasında isim değişikliği yapılmıştır.

Çizelge 4.1. Benzerlik veri kümesindeki örnek bilet çiftleri

Bilet 1	Bilet 2	Benzerlik
Merhaba, Üretim Kaynak Planlama Sistemi'nde XYZ102 ekranına yetki almak istiyorum. Yetki Talep sisteminde XYZ1 ya da XYZ2 Sistemi olarak seçenek göremedim. Yardımcı olabilir misiniz? Teşekkür ederim,	Merhaba, XYZ1 Projesi'nde XYZ2 Mühendisi olarak çalışıyorum. İşim gereği XYZ3 Sistemi'ni kullanmam gerekiyor, bu sisteme yetki talep ediyorum. Sistem adını yardım masasında da yetki yönetimi sisteminde de bulamadım. Yardımcı olabilir misiniz? Teşekkürler, iyi çalışmalar.	1
Merhabalar Depolarımızdan etiket rezervasyonu ve etiket çıkışı yapamıyoruz. Konunun çözülmesi için acilen destek olunulmasını rica ederim. İyi çalışmalar	Merhabalar, XYZ123 RR'ının depo girişi yapılmış olmasına rağmen XYZ1.XYZ2 veritabanında Statu AC görünmektedir. Statu: KP olarak güncellenmesi konusunda yardımcı olabilir misiniz? Teşekkürler, İyi çalışmalar.	0

4.2. Veri Ön İşlemleri

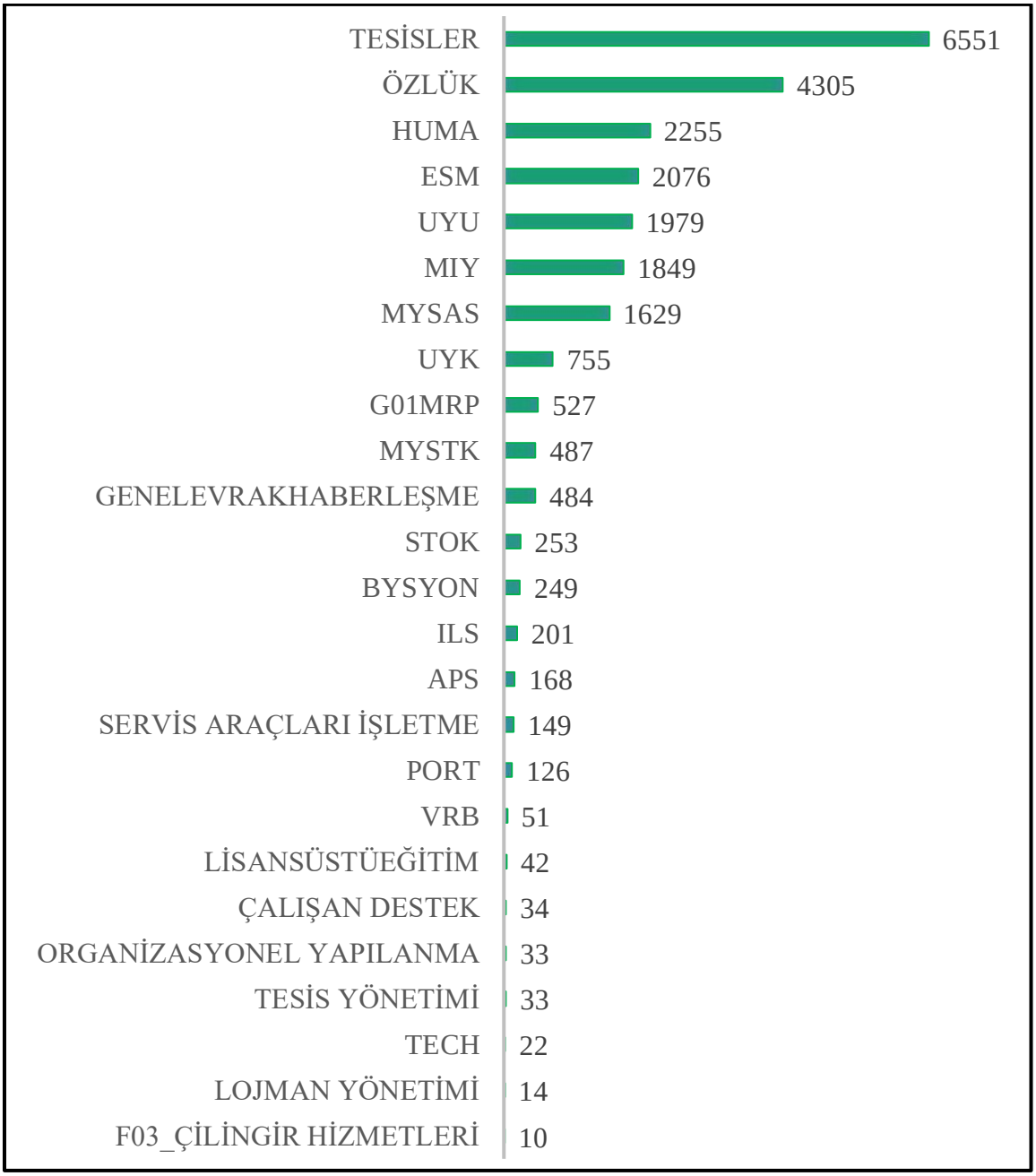
Veri kümesine ait örnek bir bilet açıklaması üzerinde veri ön işleme adımlarının uygulaması Şekil 4.5'de gösterilmiştir. Bilet sınıflandırma ve benzerliği görevleri için bilet açıklamalarındaki HTML ifadeleri kaldırılmış, URL ifadeleri kaldırılmış, cümleler kelimelerine ayrılmış, noktalama işaretleri kaldırılmış, alfabetik olmayan karakterler içeren kelimeler kaldırılmış, tek karakterden oluşan kelimeler kaldırılmış ve etkisiz kelimeler kaldırılmıştır. Bilet sınıflandırma için ek olarak kelimeler köklerine indirgenmiştir.



Şekil 4.5. Örnek bilet açıklaması üzerinde veri ön işlem adımlarının uygulanması

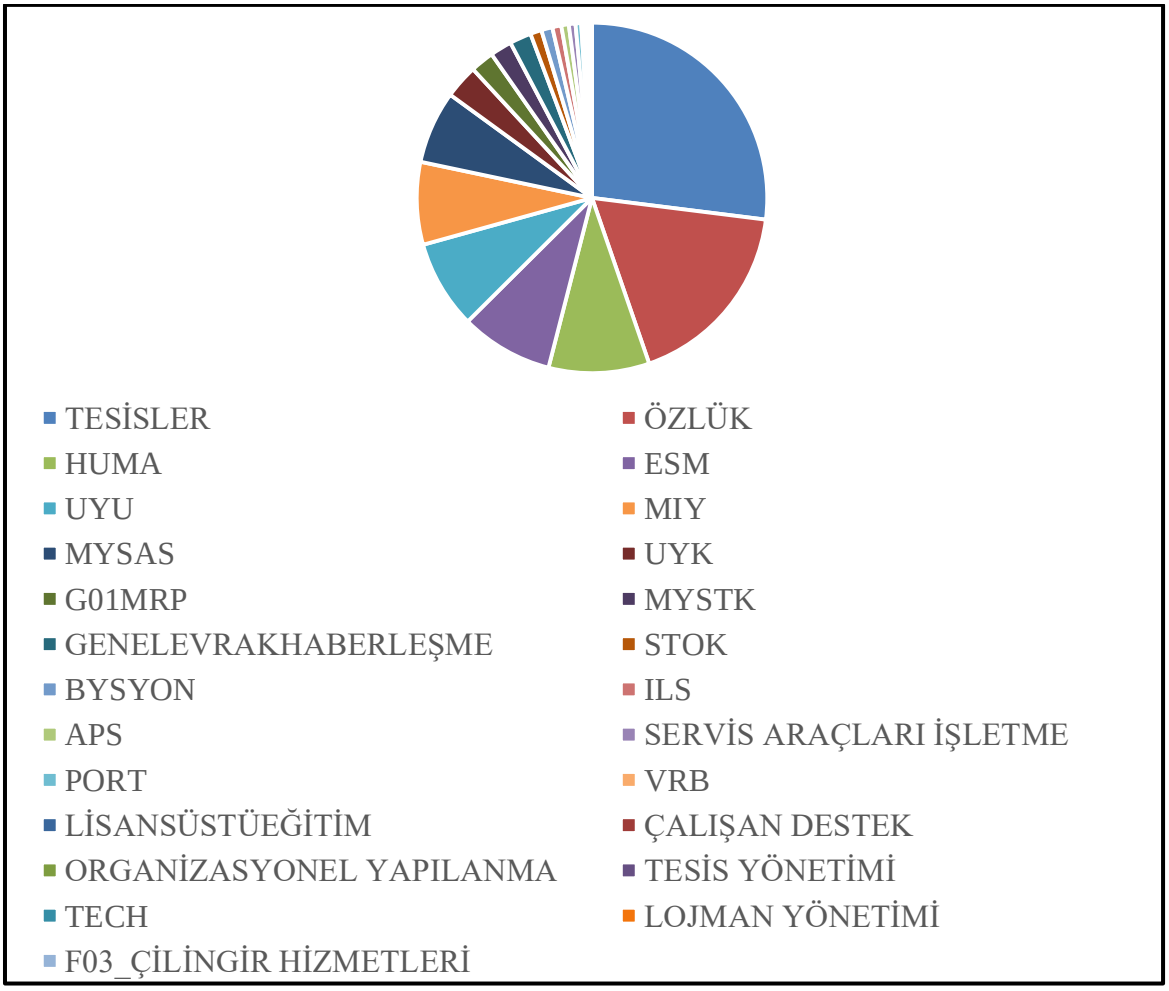
4.3. Bilet Üst Alan Sınıflandırması

Bilet üst alan sınıflandırması için kullanılacak veri kümesine veri ön işlem adımları uygulandıktan ve belge temsilleri elde edildikten sonra hiyerarşik yaklaşımın ilk adımı olarak bilet üst alan sınıflandırması gerçekleştirilmiştir. Veri ön işleme adımlarından sonra 24282 bilet ve 25 bilet üst alanından oluşan veri kümesi kullanılmıştır. Bilet üst alanlarının örnek sayısı Şekil 4.6'da gösterilmiştir.



Şekil 4.6. Bilet üst alanı veri kümesindeki bilet üst alanlarının örneklerinin sayısı

Öznitelik sayısı, PCA yöntemi kullanılarak kademeli olarak artırılmış ve doğruluk oranındaki artışın minimum düzeye indiği 2000 belirlenmiştir. Şekil 4.7'de görüldüğü üzere bilet üst alanı veri kümesinin dengesiz dağılımı nedeniyle, Rastgele Aşırı Örnekleme kullanılarak bilet üst alanlarının örnek sayısı, Tesisler üst alanı örnek sayısı olan 6551'e yükseltilmiştir. Rastgele Aşırı Örnekleme ile toplam bilet üst alanı örnek sayısı 24282'den 163775'e yükselmiştir. Makine öğrenmesi algoritması olarak SVM yöntemi deneysel olarak belirlenmiştir.



Şekil 4.7. Bilet üst alan veri kümesinin örnek dağılımı

4.4. Bilet Alt Alan Sınıflandırması

Bilet alt alan sınıflandırması için kullanılacak veri setlerine veri ön işlem adımları uygulandıktan ve belge gösterimleri elde edildikten sonra hiyerarşik yaklaşımın ikinci adımı olan bilet alt alan sınıflandırması gerçekleştirilmiştir. Bilet alt alanları modüllere veya sistemlere karşılık gelir. Bilet alt alan sınıflandırması için bilet üst alan sayısı kadar sınıflandırıcı model vardır. Bilet alt alanı veri kümeleri 1 ve 2'nin her biri bir bilet üst alanının verilerini içerecek şekilde bölümlere ayrılmış ve modeller eğitilmiştir. Bilet üst alanlarından bir bilet alt alanına sahip olanlar yok sayılmıştır. Bilet alt alanı veri kümesi 1, veri ön işlemden sonra 2866 bilet, 2 bilet üst alanı ve 41 bilet alt alanından oluşur. Bilet alt alanı veri kümesi 2, veri ön işlemden sonra 16564 bilet, 12 bilet üst alanı ve 214 bilet alt alanından oluşur. Bilet alt alan veri kümelerinin istatistiksel bilgisi Çizelge 4.2'de gösterilmiştir. Hüma ve Özlük bilet üst alanları ile bunların bilet alt alanları, bilet alt alan veri kümesi 1'de, diğer bilet üst alanları ile bilet alt alanları, bilet alt alan veri kümesi 2'de yer almaktadır.

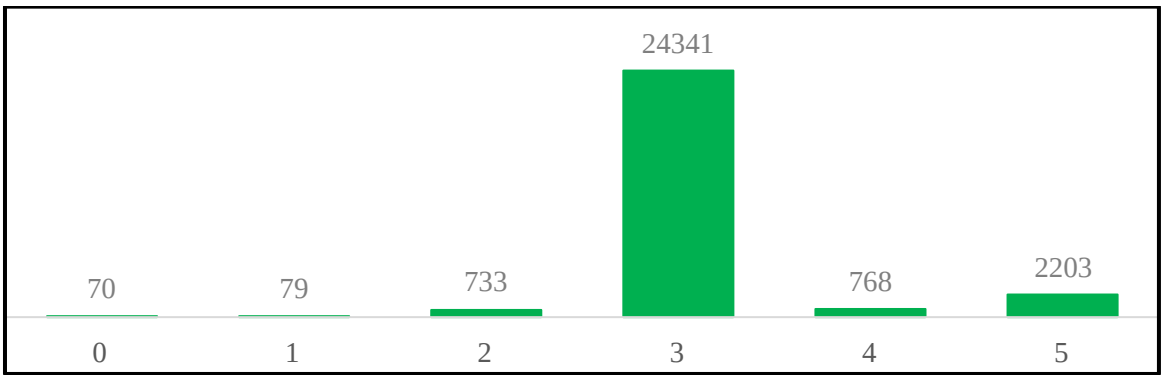
Öznitelik sayıları, PCA yöntemi kullanılarak kademeli olarak artırılmış ve doğruluk oranındaki artışın minimum düzeye indiği 50 ile 1000 arasında değerler belirlenmiştir. Bilet alt alanlarının Rastgele Aşırı Örnekleme ile dengelenmeden öncesi ve sonrasına ait bilet üst alanlarının sahip olduğu örnek sayısı ve özellik sayısı Çizelge 4.2'de gösterilmiştir. Makine öğrenmesi algoritması olarak SVM yöntemi deneysel olarak belirlenmiştir.

Çizelge 4.2. Bilet alt alan veri kümesi 1 ve 2 istatistik bilgileri

Sınıf	Örnekleme Öncesi Örnek Sayısı	Örnekleme Sonrası Örnek Sayısı	Özellik Sayısı	Alt Sınıf Sayısı
TESISLER	6520	39501	250	63
HUMA	2773	36408	1000	37
ESM	2065	8382	250	11
UYU	1960	13664	250	16
MIY	1824	9369	250	27
MYSAS	1612	18657	250	27
UYK	750	4150	250	10
G01MRP	516	1680	250	14
MYSTK	469	1072	250	16
BYSYON	248	574	200	7
STOK	246	1215	200	9
ILS	187	378	150	7
APS	167	525	150	7
ÖZLÜK	93	160	50	4

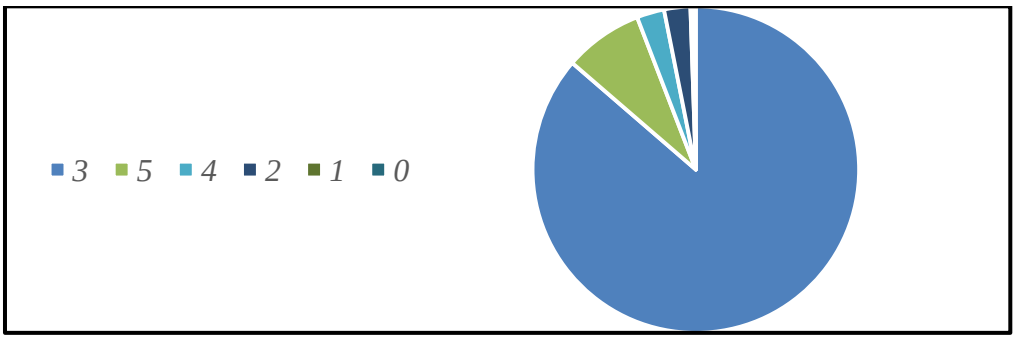
4.5. Bilet Öncelik Sınıflandırması

Bilet öncelik sınıflandırması için kullanılacak veri kümesine, veri ön işlem adımlarının uygulanmasının ve belge temsillerinin alınmasının ardından bilet öncelik sınıflandırması gerçekleştirilmiştir. Biletleri önceliklerine göre sınıflandırmak için veri ön işleme adımlarından sonra 27534 bilet ve 6 öncelikten oluşan bir veri kümesi kullanılmıştır. Veri kümesindeki önceliklerin örnek sayısı Şekil 4.8'de gösterilmiştir.



Şekil 4.8. Bilet öncelik veri kümesindeki önceliklerin örnek sayıları

Öznitelik sayısı, PCA yöntemi kullanılarak kademeli olarak artırılmış ve başarı oranındaki artışın minimum düzeye düştüğü 1000 belirlenmiştir. Şekil 4.9'da görüldüğü gibi öncelik veri kümesinin dengesiz dağılımı nedeniyle, Rastgele Aşırı Örnekleme kullanılarak önceliklerin örnek sayısı, en yüksek örnek sayısına sahip olan öncelik 3 sınıfı ile aynı sayıya yükseltilmiştir. Rastgele Aşırı Örnekleme kullanılarak örnek sayısı 27534'den 146046'ya yükseltilmiştir. Makine öğrenmesi algoritması olarak K-NN yöntemi deneysel olarak belirlenmiştir. K parametresi deneysel olarak 4 ayarlanmıştır.



Şekil 4.9. Bilet öncelik veri kümesinin örnek dağılımı

4.6. Bilet Benzerliği

Bilet çiftlerinin benzerliğinin hesaplanması için öncelikle bilet açıklamalarına veri ön işlem adımlarının uygulanması, bilet açıklamaların sayısal vektörlere dönüştürülmesi ve elde edilen sayısal vektörlerin benzerliğinin ölçülmesi gerekir. Bilet açıklamalarına veri ön işlem adımlarının uygulanması, bilet benzerlik performansının artmasına katkıda bulunmuştur. Bu kapsamda, 48206 bilet çiftine veri ön işlem adımları uygulanmıştır. Bilet açıklamalarının doğal dil içermesi sebebiyle benzerliğin ölçülmesi için sayısal vektörler elde edilmesi gerekir. Dolayısıyla Word2vec, Doc2vec ve LSTM yöntemleri uygulanarak bilet açıklamalarının temsilleri elde edilmiştir. Sayısal vektörler arasındaki benzerliği hesaplamak için uzaklık türleri kullanılmaktadır. Bu çalışmada kosinüs, öklid ve manhattan uzaklık türleri kullanılarak bilet çiftlerinin benzerliği ölçülmüştür.

4.6.1. Word2vec

Word2vec modeli, önceden işlenmiş bilet açıklamaları kullanılarak eğitilmiştir. Eğitim algoritması olarak CBOW seçilmiştir. Pencere büyüklüğü deneysel olarak 10 belirlenmiştir. Word2vec modeli, kelime temsillerini ve Siyam LSTM modelinin gömme matrisini oluşturmak için uygulanmıştır. Kelime temsilleri olarak 300 boyutlu vektörler elde edilmiştir. Doküman temsilleri, kelime temsillerinin ortalaması alınarak hesaplanmıştır.

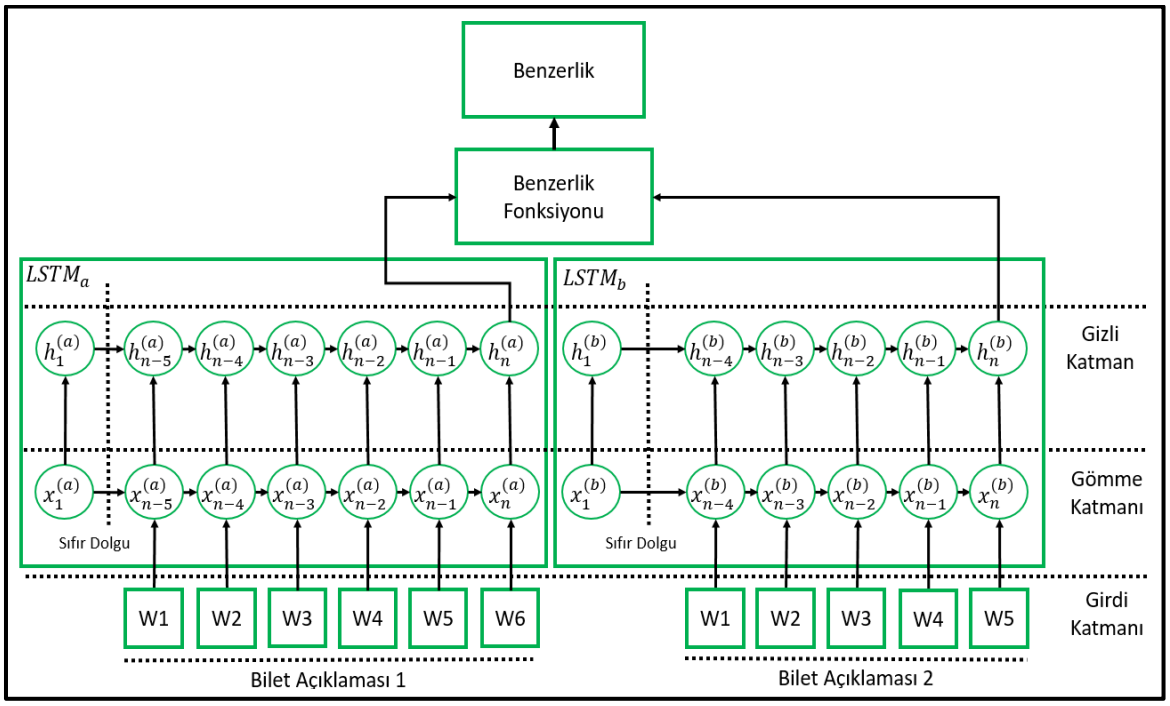
4.6.2. Doc2vec

Doc2vec modeli, önceden işlenmiş bilet açıklamaları kullanılarak eğitilmiştir. Eğitim algoritması olarak DMPV seçilmiştir. Pencere büyüklüğü deneysel olarak 5 belirlenmiştir. Doc2vec, doküman temsillerini elde etmek için uygulanmıştır. Doküman temsilleri olarak 300 boyutlu vektörler elde edilmiştir.

4.6.3. LSTM

Bir Siyam sinir ağı girdi olarak iki veya daha fazla vektör alır. Hesaplamalar aynı sinir ağları ile yapılır. Daha sonra çıktı vektörü üretilir [31]. Her biri belirli bir çiftteki cümlelerden birini işleyen eşit iki ağ olan LSTMa ve LSTMb ‘yi içeren Siyam LSTM modeli Şekil 4.10'da

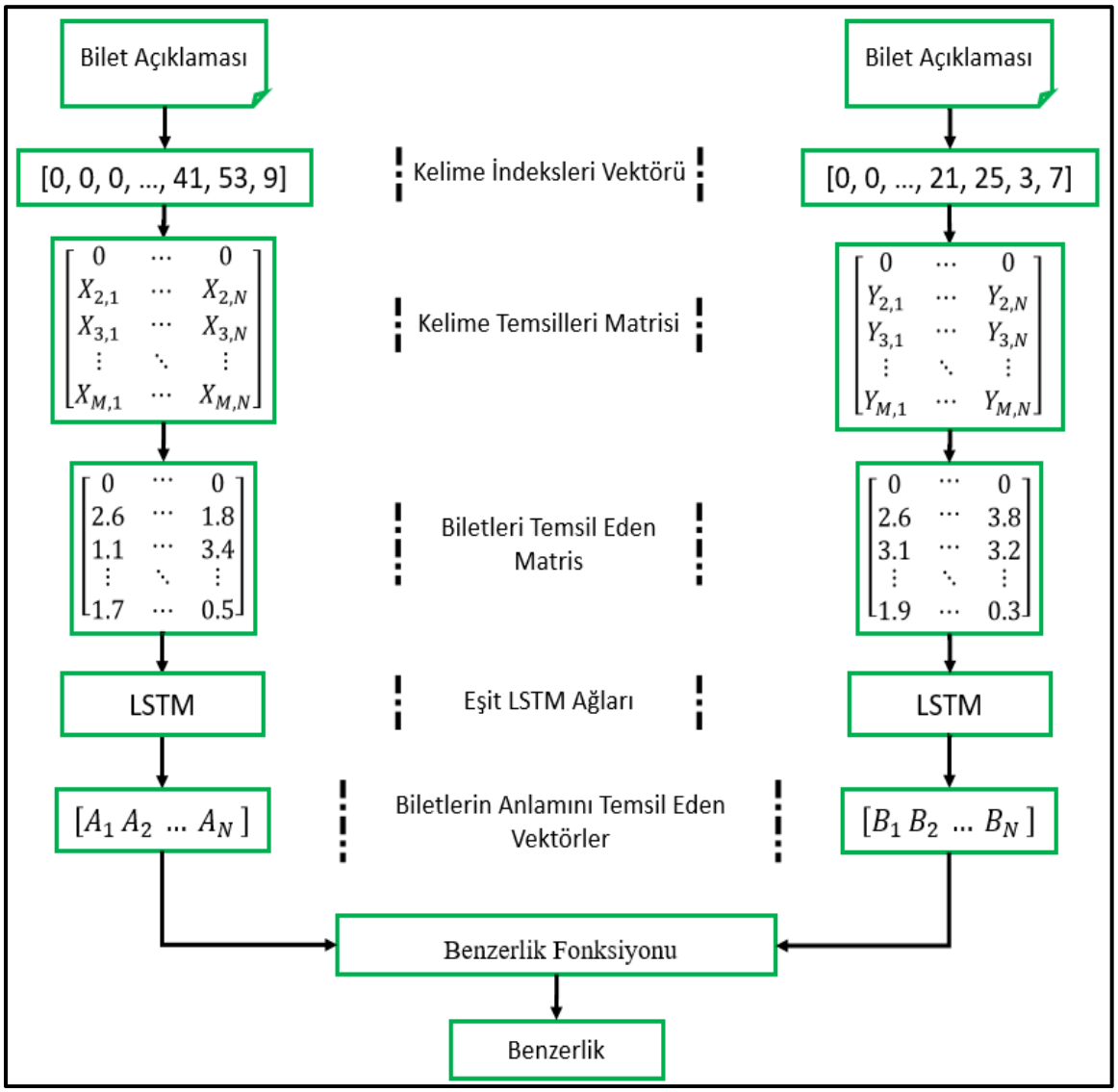
gösterilmiştir. Siyam LSTM modeli, bilet açıklamalarının benzerliğini hesaplamak için eğitilmiştir.



Şekil 4.10. Siyam LSTM ağı

Gömme katmanlarının girişi, sıfır dolgulu dizilerdir. Şekil 4.11'de görüldüğü gibi bu diziler kelime indekslerini temsil eder. Sıfır doldurma ile sabit uzunlukta vektörler elde edilir. Bu katman, Word2vec modeli tarafından oluşturulan kelime temsilleri matrisini kullanır. Sözcüklere karşılık gelen bir dizi temsil çıktısı verir. Gömme katmanlarının çıktısı olan iki matris, LSTM ağlarının girdisi haline gelir. LSTM ise her belgeyi 32 boyutlu bir vektör olarak temsil eder. Bu vektörler belgenin anlamını taşır. Böylece Siyam ağı, sabit uzunluklu vektörler ile değişken uzunluktaki belgeleri temsil eder [35]. Benzerlik fonksiyonu kullanılarak her iki LSTM'in çıktısı olan temsillerin benzerliği hesaplanır. Benzerlik değeri daha sonra gerçek etiketle karşılaştırılır ve MSE hesaplanır. Çünkü kayıp fonksiyonu olarak ortalama kare hata seçilmiştir. Gradyan daha sonra LSTM'lerden birinin ağırlıklarına göre alınır. Ardından, optimize edici ile geri yayılım yapılır ve Eş. 4.1'de görüldüğü gibi geri yayılım için gradyan hesaplanır. Optimize edici olarak Adam seçilmiştir. Gradyan (LSTMa) ve Gradyan (LSTMb), MaLSTM modelinde her birim tarafından hesaplanan gradyanları temsil eder. Bu, LSTM'lerin simetrik geri yayılımına yardımcı olur, böylece LSTMa=LSTMb kriterinin korunması sağlanır [32].

$$\text{Gradyanlar} = \frac{\text{Gradyan}(\text{LSTM}_a) + \text{Gradyan}(\text{LSTM}_b)}{2} \quad (4.1)$$



Şekil 4.11. Siyam LSTM metin benzerliği modeli

4.6.4. Bilet benzerlik ölçümü

Benzerlik fonksiyonu olarak fcos, fmanh ve feuc kullanılarak her bir bilet çiftinin benzerliği hesaplanmıştır. Eş. 4.2'de görüldüğü gibi fcos, kosinüs benzerliğini 0'dan 1'e ölçeklendirir. Eş. 4.3'de görüldüğü gibi fmanh, manhattan uzaklığını 0'dan 1'e ölçeklendirir. Eş. 4.4'de görüldüğü gibi feuc, öklid uzaklığını 0'dan 1'e ölçeklendirir.

$$f_{cos} = \frac{\text{Kosinüs Benzerliği}(A, B) + 1}{2} \tag{4.2}$$

$$f_{manh} = \exp(-\text{Manhattan Uzaklığı}) \tag{4.3}$$

$$f_{euc} = \frac{1}{e^{f_{euc\_dist}}} \tag{4.4}$$

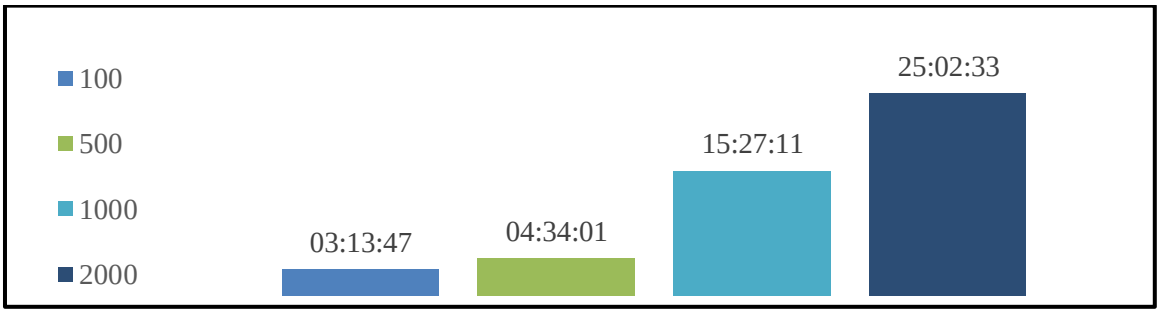
5. DENEYSEL SONUÇLAR

Biletleri alanlara ve önceliklere göre sınıflandıracak modellerin performans değerlendirme metrikleri, çapraz doğrulama tekniği kullanılarak elde edilmiştir. Çapraz doğrulama tekniğinin k parametresi için 10 değeri deneysel olarak belirlenmiştir. Bilet sınıflandırma modellerinin her biri için doğruluk, kesinlik, duyarlılık ve F1 puan değerleri hesaplanmıştır. Örnekleme öncesi alınan kesinlik, duyarlılık ve F1 puan değerlerinin sıfır olmasının nedeni paydadaki değerin sıfır olması sonucunda işlem sonucunun sıfır olarak belirlenmesindendir.

Biletlerin benzerliğini hesaplayacak modellerin performans değerlendirme metrikleri çapraz doğrulama tekniği kullanılarak elde edilmiştir. Çapraz doğrulama tekniğinin k parametresi için 10 değeri deneysel olarak belirlenmiştir. Uzaklık türleri ve vektörleştirme yöntemlerinin her biri için MSE, RMSE, MAE, PCC ve SRCC değerleri hesaplanmıştır.

5.1. Bilet Üst Alan Sınıflandırması Sonuçları

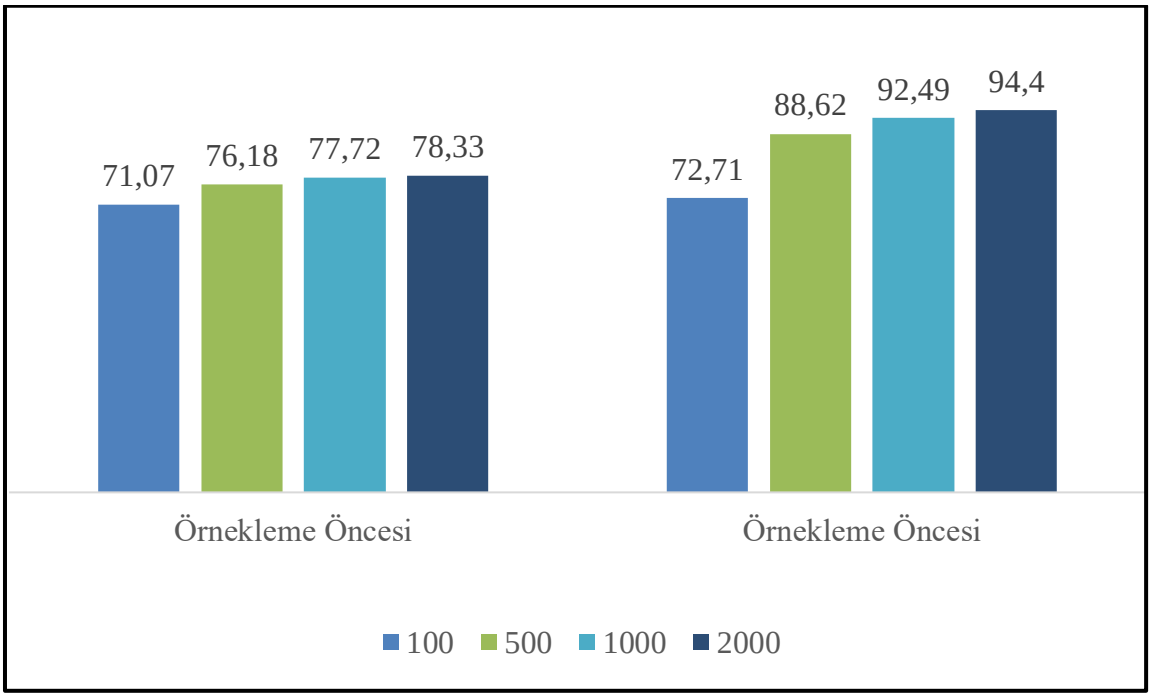
Yardım masası bileti üst alan sınıflandırması için SVM modeli eğitilmiş ve sonuçlar elde edilmiştir. Örnekleme uygulandıktan sonra farklı sayıda özellik için SVM sınıflandırıcının eğitim süreleri Şekil 5.1'de gösterilmiştir. Özellik sayısının artmasıyla birlikte sınıflandırıcının eğitimin süresi uzamış ve sürenin artış hızı yükselmiştir.



Şekil 5.1. Bilet üst alan sınıflandırmasında SVM yönteminin eğitim süresi

Örnekleme öncesine ve sonrasına ait farklı sayıda özellik için SVM sınıflandırıcısının bilet üst alan sınıflandırma doğruluk oranları Şekil 5.2'de gösterilmiştir. Özellik sayısı arttıkça sınıflandırıcının doğruluk oranı artmış ve oranın artış hızı düşmüştür. Şekil 5.2'de görüldüğü gibi %94,40 doğruluk oranı ile biletlerin üst alana sınıflandırılması gerçekleştirilmiştir. Rastgele Aşırı Örnekleme tekniğinin kullanılması modelin 2 saat 19 dakika 9 saniye süren

eğitim süresini 25 saat 2 dakika 33 saniyeye çıkarırken doğruluk oranını %78,33'den %94,40'a yükseltmiştir. Böylece daha uzun bir eğitim süresine karşın daha yüksek bir doğruluk oranı elde edilmiştir.



Şekil 5.2. Bilet üst alan sınıflandırmasında SVM yönteminin doğruluk oranı

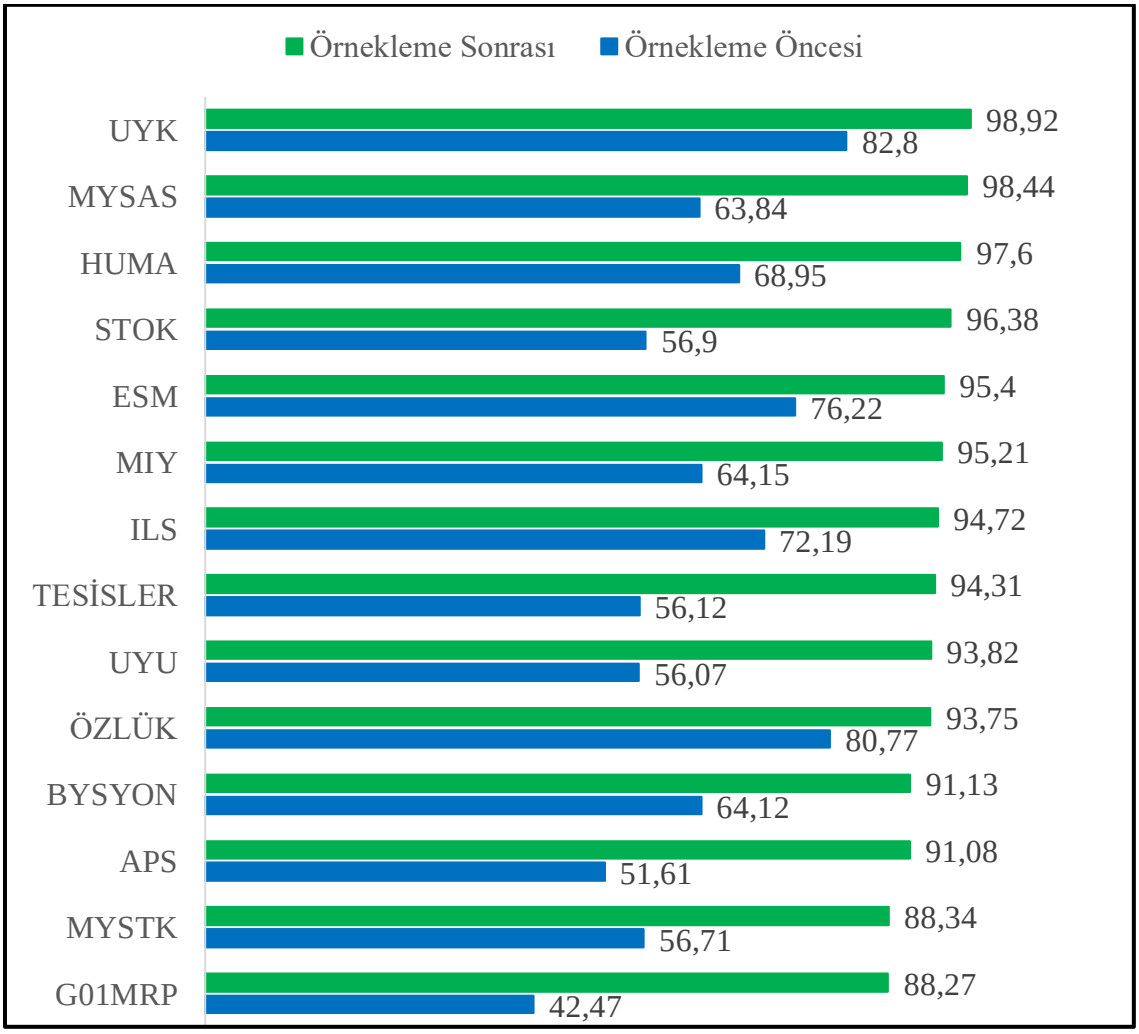
Rastgele Aşırı Örnekleme yönteminin uygulanmasından öncesine ile sonrasına ait bilet üst alan sınıflandırması kesinlik, duyarlılık ve F1 puan değerleri Çizelge 5.1'de gösterilmiştir.

Çizelge 5.1. Bilet üst alan sınıflandırmasında SVM yönteminin kesinlik, duyarlılık ve F1 puan değerleri

Sınıf	Kesinlik (%)	Duyarlılık (%)	F1 Puanı (%)	Kesinlik (%)	Duyarlılık (%)	F1 Puanı (%)
	Örnekleme Öncesi			Örnekleme Sonrası		
APS	8	1	1	94	98	96
BYSYON	74	33	46	95	98	97
ESM	70	84	76	90	90	90
F03_ÇİLİNGİ HİZMETLERİ	0	0	0	89	100	94
G01MRP	39	23	29	84	91	87
GENEL EVRAK HABERLESME	92	93	92	99	100	99
HUMA	54	67	60	82	79	80
ILS	85	33	48	98	99	99
LOJMANYÖNETİMİ	0	0	0	100	100	100
LİSANSÜSTÜ EGİTİM	88	50	64	100	95	98
MIY	71	72	71	85	85	85
TESİS YÖNETİMİ	0	0	0	100	100	100
MYSAS	66	64	65	89	81	85
MYSTK	65	23	34	91	95	93
ORGANİZASYONEL YAPILANMA	0	0	0	100	100	100
SERVİS ARAÇLARI İŞLETME	88	83	85	100	98	99
STOK	59	49	54	97	98	98
TECH	0	0	0	96	95	96
TESISLER	94	98	96	97	96	96
PORT	62	29	40	99	99	99
UYK	70	46	55	90	91	91
UYU	69	77	73	92	78	85
VRB	100	25	41	100	100	100
ÇALIŞAN DESTEK	0	0	0	100	100	100
ÖZLÜK	90	92	1	95	91	93

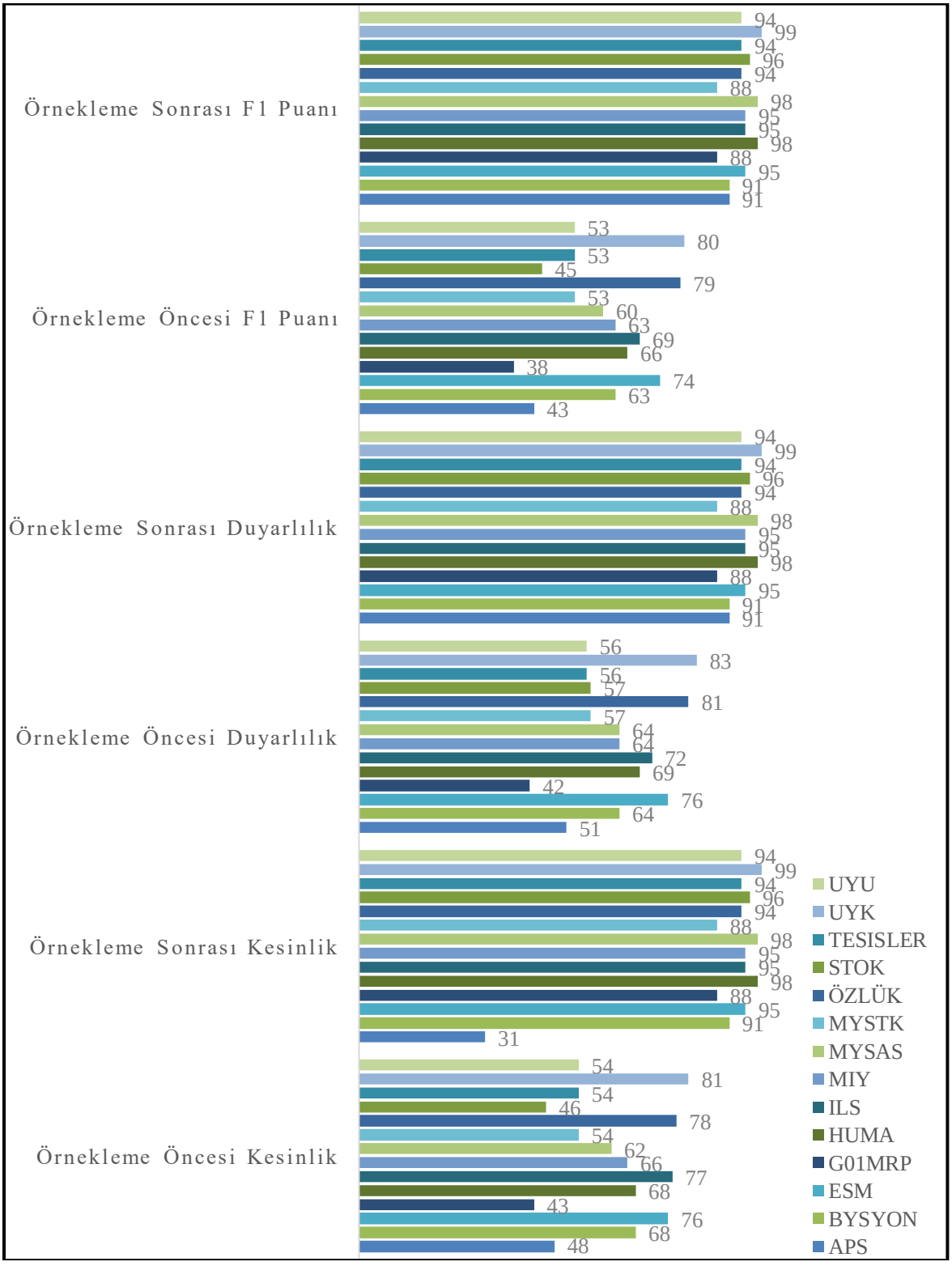
5.2. Bilet Alt Alan Sınıflandırması Sonuçları

Yardım masası bilet alt alan sınıflandırması için bilet üst alan sayısı kadar SVM modeli eğitilmiş ve sonuçlar elde edilmiştir. Bilet alt alanı sınıflandırmasında en yüksek doğruluk oranı %98,92 ile 10 bilet alt alanı içeren UYK bilet üst alanı için elde edilirken en düşük doğruluk oranı 14 bilet alt alanı içeren G01MRP bilet üst alanı için %88,27 ile elde edilmiştir. Bilet alt alan sınıflandırıcılarının doğruluk oranları Şekil 5.3'de gösterilmiştir. Rastgele Aşırı Örnekleme yönteminin kullanılması ile doğruluk oranındaki en büyük artış 14 sistem içeren G01MRP bilet üst alanında %45,80 ile gerçekleşmiştir. Doğruluk oranı %42,47'den %88,27'ye yükselmiştir. En küçük artış ise %16,12 ile 10 sistem içeren UYK bilet üst alanında gerçekleşmiştir. Doğruluk oranı %82,80'den %98,92'ye yükselmiştir. Rastgele Aşırı Örnekleme yönteminin performansa katkısı istatistiksel olarak gösterilmiştir.



Şekil 5.3. Bilet alt alan sınıflandırmasında SVM yönteminin doğruluk oranları

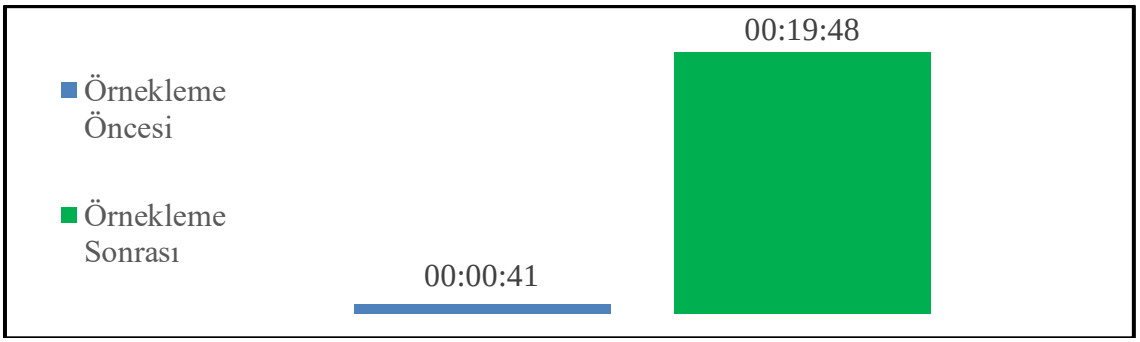
Yardım masası bilet alt alan sınıflandırması için her bir bilet üst alan modelinin ağırlıklı kesinlik, duyarlılık ve F1 puan değerleri Şekil 5.4'de gösterilmiştir.



Şekil 5.4. Bilet alt alan sınıflandırmasında SVM yönteminin ağırlıklı kesinlik, duyarlılık ve F1 puan değerleri

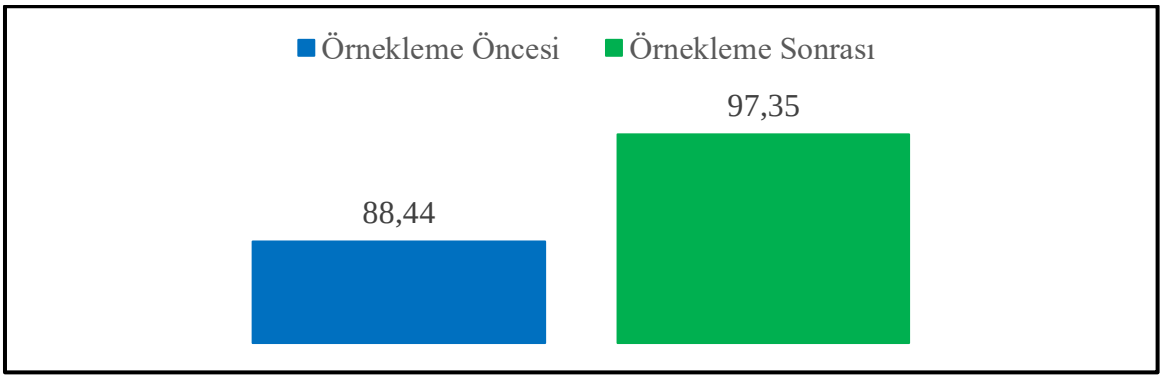
5.3. Bilet Öncelik Sınıflandırması Sonuçları

Yardım masası bilet öncelik sınıflandırması için K-NN modeli eğitilmiş ve sonuçlar elde edilmiştir. Şekil 5.5'de görüldüğü gibi Rastgele Aşırı Örnekleme yönteminin kullanılması ile 41 saniye süren sınıflandırıcının eğitim süresi 19 dakika 48 saniyeye yükselmiştir.



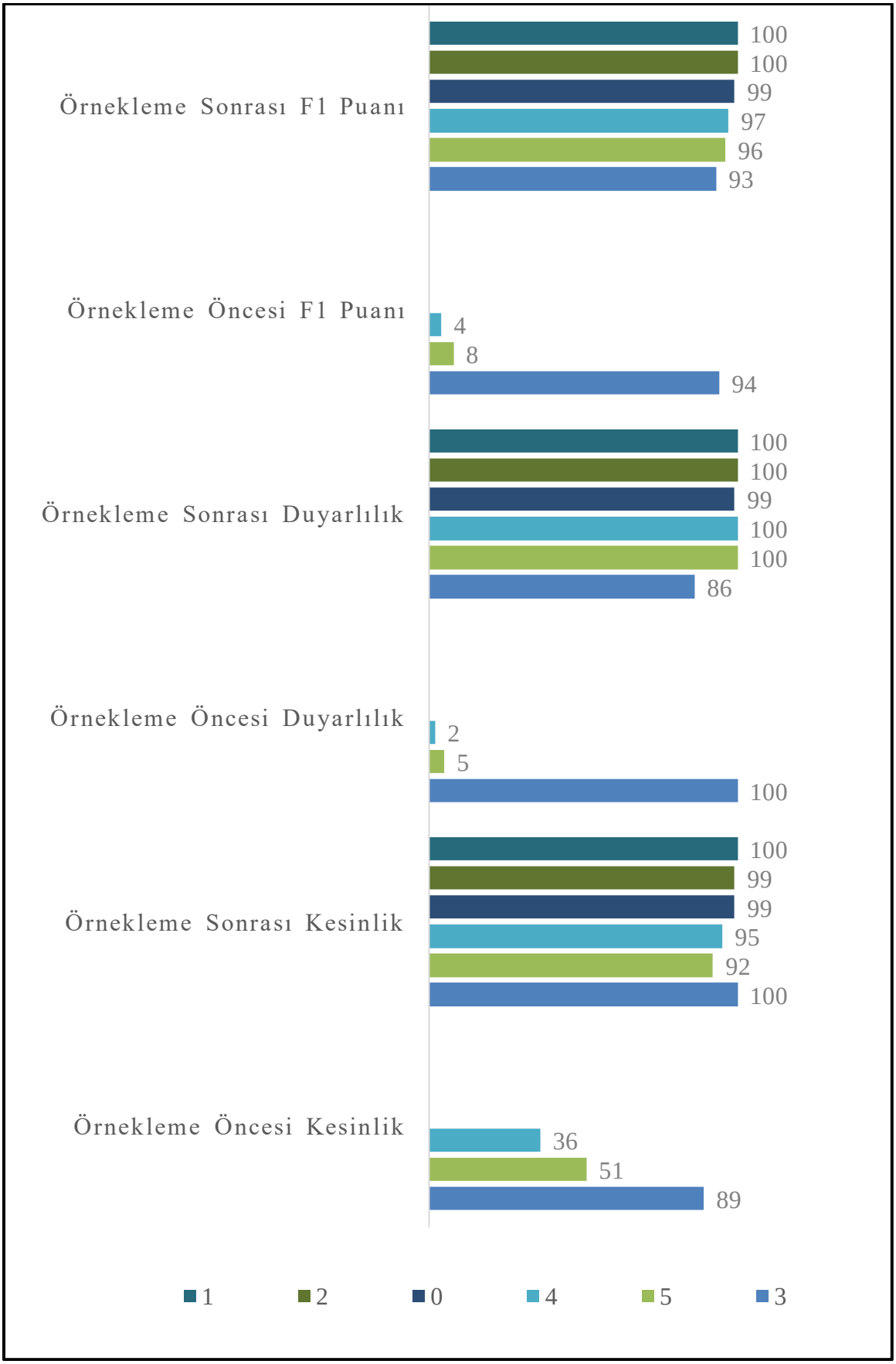
Şekil 5.5. Bilet öncelik sınıflandırmasında K-NN yönteminin örnekleme öncesi ve sonrası eğitim süresi

Yardım masası bilet öncelik sınıflandırması Şekil 5.6'da görüldüğü gibi %97,35 doğruluk oranı ile gerçekleştirilmiştir. Şekil 5.6'da görüldüğü gibi Rastgele Aşırı Örnekleme tekniğinin kullanılması ile sınıflandırıcının doğruluk oranı %88,44'den %97,35'e yükselmiştir. Böylece daha uzun bir eğitim süresine karşın daha yüksek bir başarı oranı elde edilmiştir.



Şekil 5.6. Bilet öncelik sınıflandırmasında örnekleme öncesi ve sonrası için K-NN yöntemi doğruluk oranları

Yardım masası bileti öncelik sınıflandırması için her bir önceliğin örnekleme öncesi ve sonrası kesinlik, duyarlılık ve F1 puan değerleri Şekil 5.7'de gösterilmiştir.

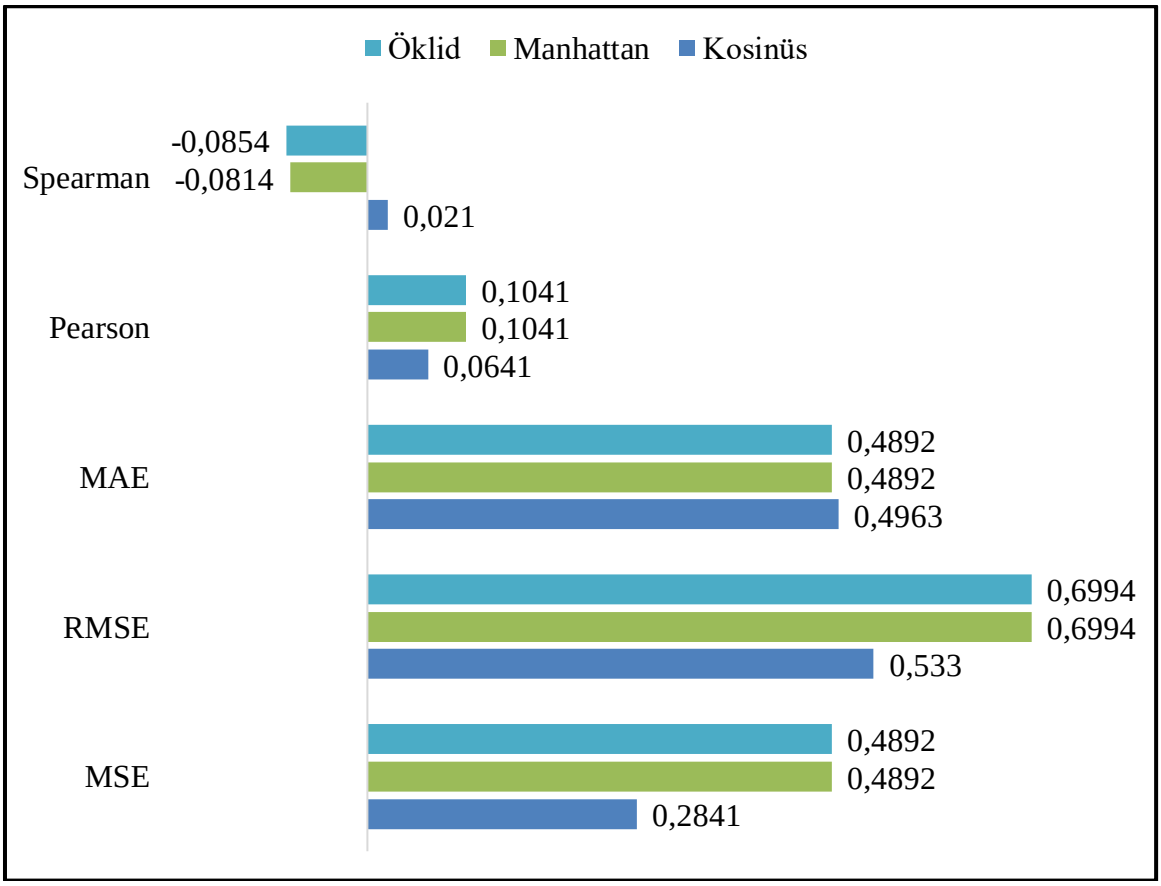


Şekil 5.7. Bilet öncelik sınıflandırmasında K-NN yönteminin örnekleme öncesi ve sonrası için kesinlik, duyarlılık ve F1 puan değerleri

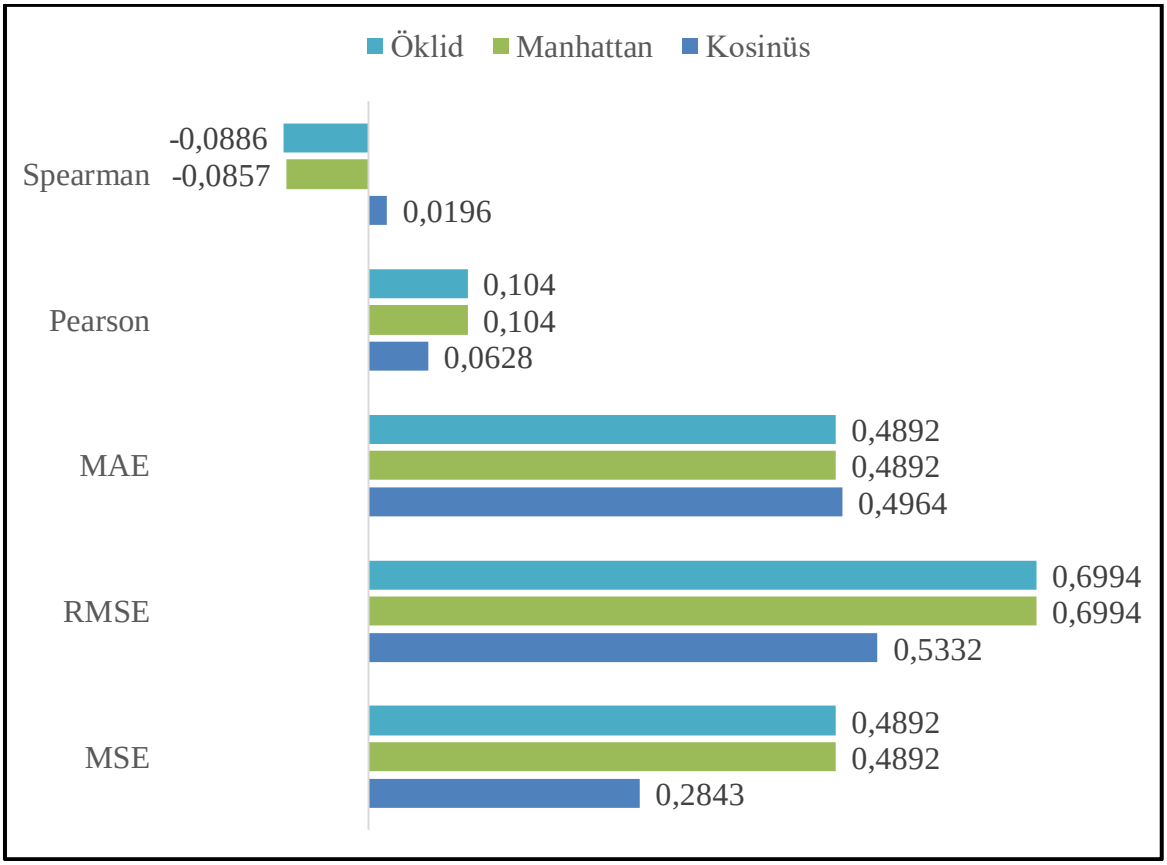
5.4. Bilet Benzerlik Sonuçları

5.4.1. Word2vec

Word2vec yöntemi için en yüksek bilet benzerlik performansı kosinüs uzaklığı kullanılarak elde edilmiştir. Kosinüs uzaklığı ve Word2vec yöntemi kullanılarak eğitim ve test veri kümesi için sırasıyla 0,2841 ve 0,2843 MSE, 0,533 ve 0,5332 RMSE, 0,4963 ve 0,4964 MAE, 0,0641 ve 0,0628 PCC, 0,021 ve 0,0196 SRCC elde edilmiştir. Word2vec'in bilet benzerlik performansı Şekil 5.8'de eğitim veri kümesi ve Şekil 5.9'da test veri kümesi için gösterilmiştir.



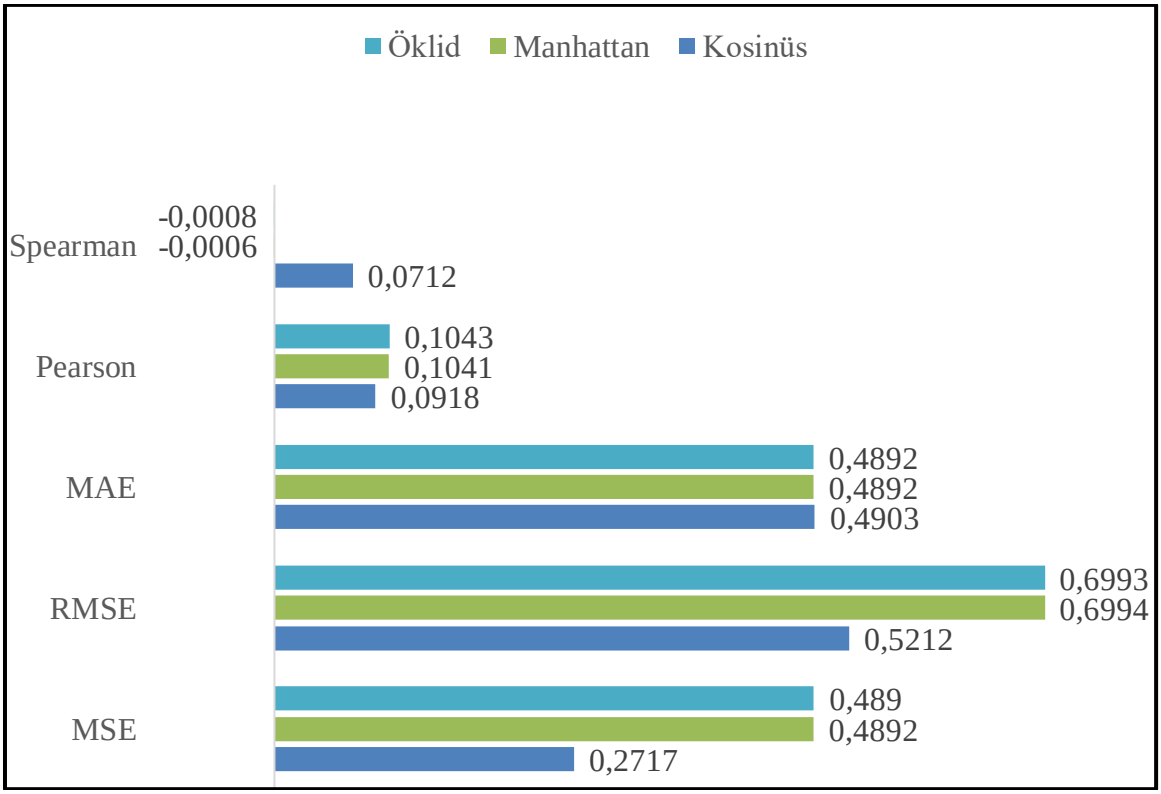
Şekil 5.8. Word2vec modelinin eğitim veri kümesi bilet benzerlik performansı



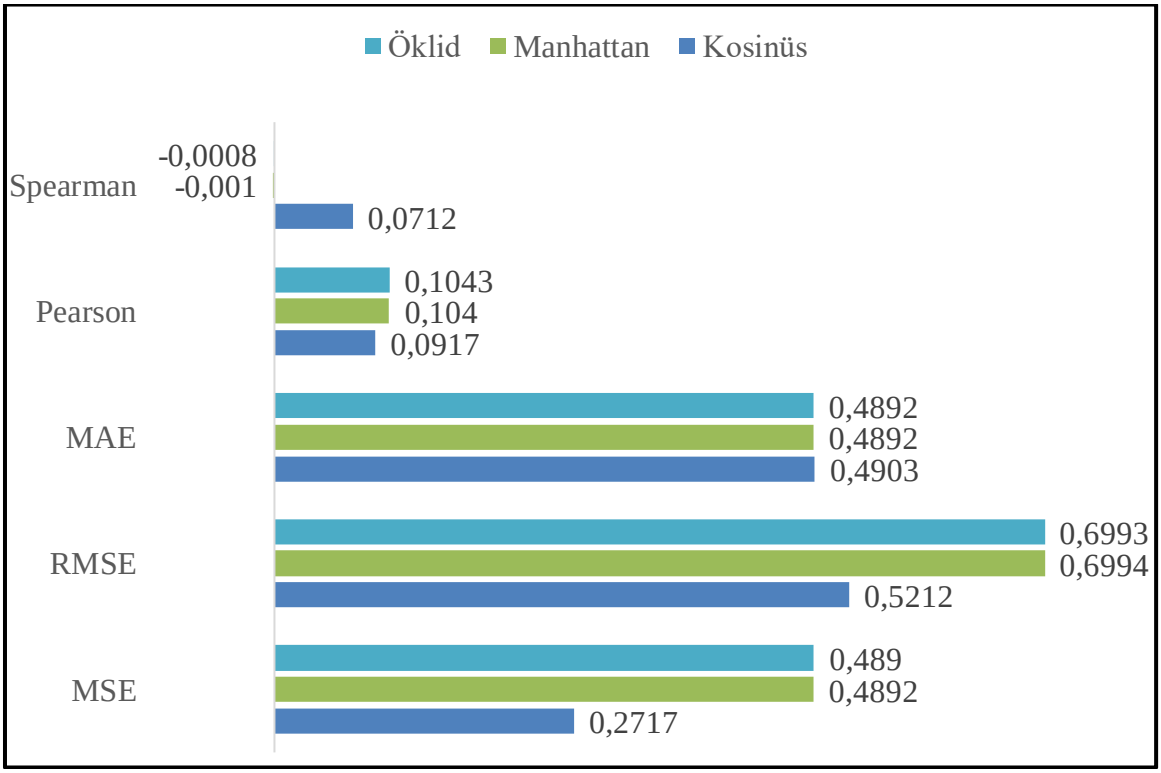
Şekil 5.9. Word2vec modelinin test veri kümesi bilet benzerlik performansı

5.4.2. Doc2vec

Doc2vec yöntemi için en yüksek bilet benzerlik performansı kosinüs uzaklığı kullanılarak elde edilmiştir. Kosinüs uzaklığı ve Doc2vec yöntemi kullanılarak eğitim ve test veri kümesi için sırasıyla 0,2717 ve 0,2717 MSE, 0,5212 ve 0,5212 RMSE, 0,4903 ve 0,4903 MAE, 0,0918 ve 0,0917 PCC, 0,0712 ve 0,0712 SRCC elde edilmiştir. Doc2vec'in bilet benzerlik performansı Şekil 5.10'da eğitim veri kümesi ve Şekil 5.11'de test veri kümesi için gösterilmiştir.



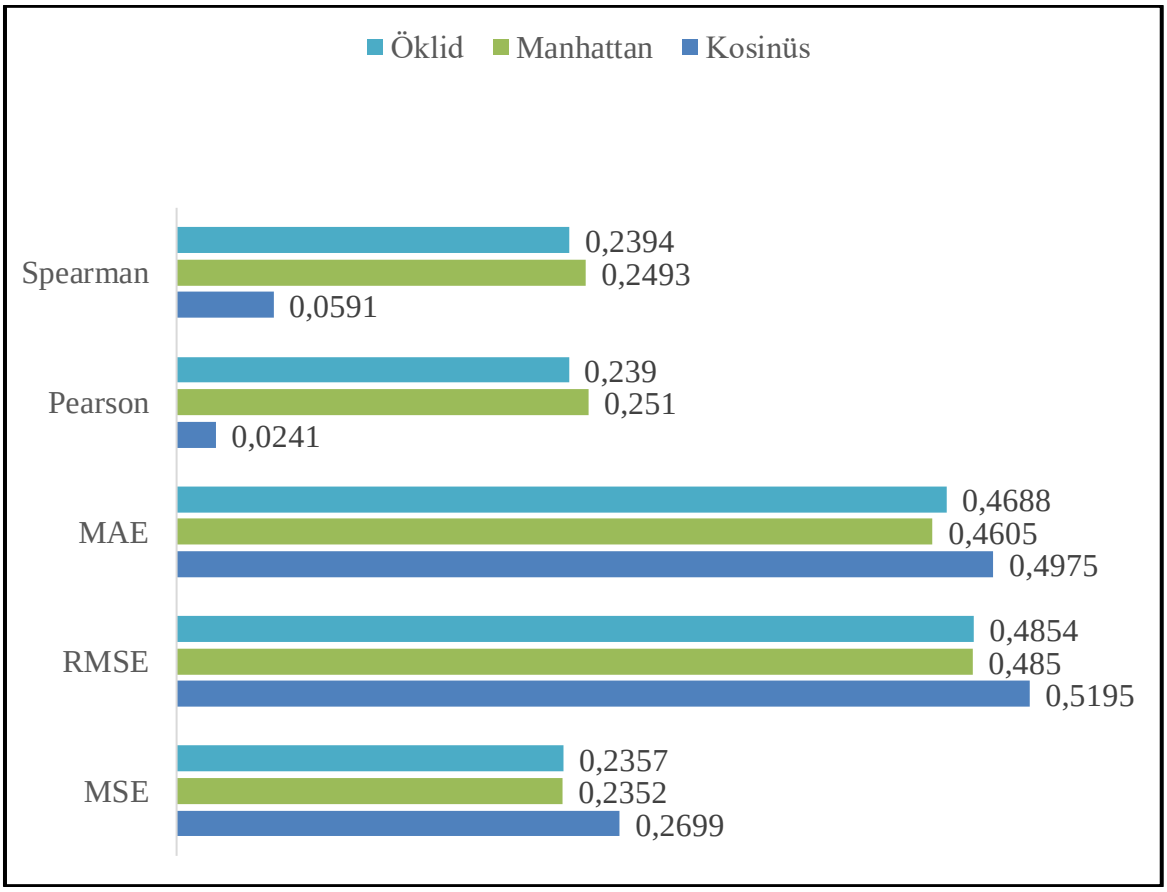
Şekil 5.10. Doc2vec modelinin eğitim veri kümesi bilet benzerlik performansı



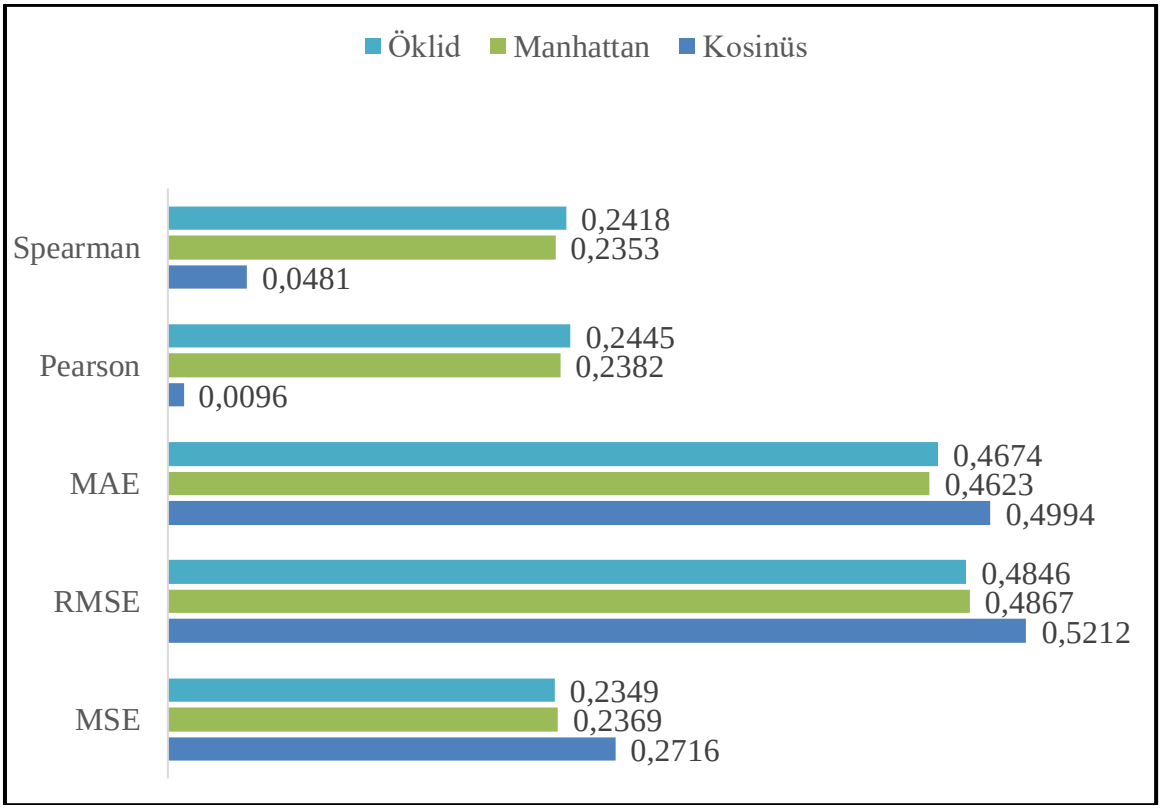
Şekil 5.11. Doc2vec modelinin test veri kümesi bilet benzerlik performansı

5.4.3. LSTM

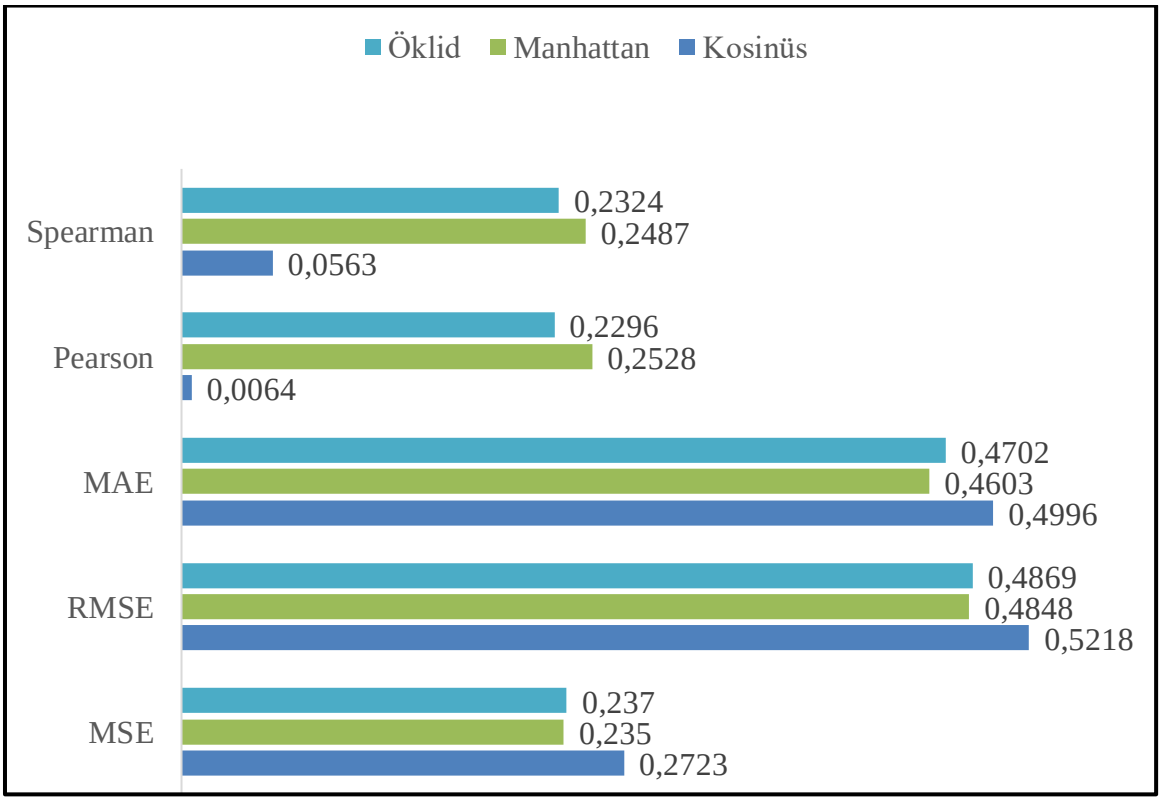
Siyam LSTM modeli için en yüksek bilet benzerlik performansı manhattan uzaklığı kullanılarak elde edilmiştir. MaLSTM modeli kullanılarak eğitim, doğrulama ve test veri kümesi için sırasıyla 0,2352, 0,2369 ve 0,235 MSE, 0,485, 0,4867 ve 0,4848 RMSE, 0,4605, 0,4623 ve 0,4603 MAE, 0,251, 0,2382 ve 0,2528 PCC, 0,2493, 0,2353 ve 0,2487 SRCC elde edilmiştir. LSTM'nin bilet benzerlik performansı Şekil 5.12'de eğitim veri kümesi için, Şekil 5.13'de doğrulama veri kümesi için ve Şekil 5.14'de test veri kümesi için gösterilmiştir.



Şekil 5.12. MaLSTM modelinin eğitim veri kümesi bilet benzerlik performansı



Şekil 5.13. MaLSTM modelinin doğrulama veri kümesi bilet benzerlik performansı



Şekil 5.14. MaLSTM modelinin test veri kümesi bilet benzerlik performansı

6. SONUÇ VE DEĞERLENDİRME

Kullanıcılar yardım masası sistemini kullanarak sorunlarını ve isteklerini müşteri hizmetleri temsilcilerine iletmek için bilet açar. Biletlerin yanlış alanlara açılması, uzun bekleme süreleri, biletlerin yanlış öncelik düzeyinde girilmesi ve bilet dağıtımında iş gücü kaybedilmesi yardım masası sisteminin etkin kullanımında sorunlar yaratmaktadır. Yinelenen durumlar, BT kaynaklarının eksikliği, kullanıcıya ve yöneticiye bağımlılık, müşteri hizmetleri temsilcileri üzerinde biletlerin birikmesine ve çözüm sürelerinin uzamasına neden olur. Dolayısıyla bu sorunların üstesinden gelmek için bu çalışmada biletleri üst ile alt alanlara ve önceliklerine göre sınıflandıracak, müşteri hizmetleri temsilcileri tarafından kapatılan biletlerle aynı anlama gelen biletlerin müşteri hizmetleri temsilcilerine bildirilerek kısa süre içerisinde kapatılmasını sağlayacak modellerin tasarlanması amaçlanmıştır.

Yardım masası bilet alan sınıflandırması üzerine gerçekleştirilen çalışmaların çoğu tek adımlı bir yaklaşım kullanırken [16, 17, 18, 19, 21, 22], hiyerarşik yaklaşımı kullanan çalışmalar da bulunmaktadır [20]. Bilet alan sınıflandırması için literatürdeki çalışmalarda elde edilen daha yüksek performans göz önünde bulundurularak bu çalışmada hiyerarşik yaklaşım tercih edilmiştir. Hiyerarşik yaklaşım önce biletlerin sınıflandırılacağı üst alanı tahmin etmeye ve ardından öngörülen üst alanın kapsadığı alt alanı tahmin etmeye dayanır. Böylece, yardım masasındaki çok sayıda alan için çok sınıflı metin sınıflandırma zorluğunun üstesinden gelinir. Anlamsal metin benzerliği üzerine literatürde birçok çalışma bulunmaktadır; ancak bilet benzerliği ile ilgili az sayıda çalışma yapılmasının yanı sıra Türkçe bilet benzerliği ile ilgili herhangi bir çalışma bulunmamaktadır. Bu tez çalışması ile birlikte literatürde ilk olarak müşteri hizmetleri temsilcileri tarafından kapatılan biletlerle aynı anlama gelen biletlerin müşteri hizmetleri temsilcilerine bildirilerek kısa süre içerisinde kapatılması önerilmiştir. Böylece müşteri hizmetleri temsilcilerinin daha karmaşık sorunlara daha uzun süre ayırmasına olanak sağlanması ve kullanıcı memnuniyetinin artırılması amaçlanmıştır. Dengesiz veri kümelerinde çok sınıflı metin sınıflandırma probleminin üstesinden gelmek için yardım masası bilet açıklamaları için seçilen veri ön işleme adımları kullanılmıştır. Yardım masası sisteminden toplanan üst alan, alt alan ve öncelik sınıfları ile bilet açıklamasını içeren dört veri kümesi kullanılarak bilet alan ve öncelik sınıflandırma modelleri eğitilmiş ve performansı test edilmiştir. Kartezyen olarak eşleştirilmiş ve her bir

bilet çiftinin benzer olup olmadığı manuel olarak etiketlenmiş veri kümesi kullanılarak bilet benzerlik modeli eğitilmiş ve performansı test edilmiştir.

Bilet üst alan sınıflandırması için SVM yöntemi kullanılarak %94,4 doğruluk oranı elde edilmiştir. Bilet alt alan sınıflandırmasında SVM yöntemi kullanılarak %98,92 ile en yüksek doğruluk oranı ve %88,27 ile en düşük doğruluk oranı elde edilmiştir. Yardım masası bilet öncelik sınıflandırması için K-NN kullanılarak %97,35 doğruluk oranı elde edilmiştir. MaLSTM modeli, bilet benzerlik görevinde Word2vec ve Doc2vec modellerine kıyasla daha yüksek bir performans elde etmiştir. MaLSTM modeli 0,235 MSE, 0,4848 RMSE, 0,4603 MAE, 0,2528 PCC ve 0,2487 SRCC elde etmiştir.

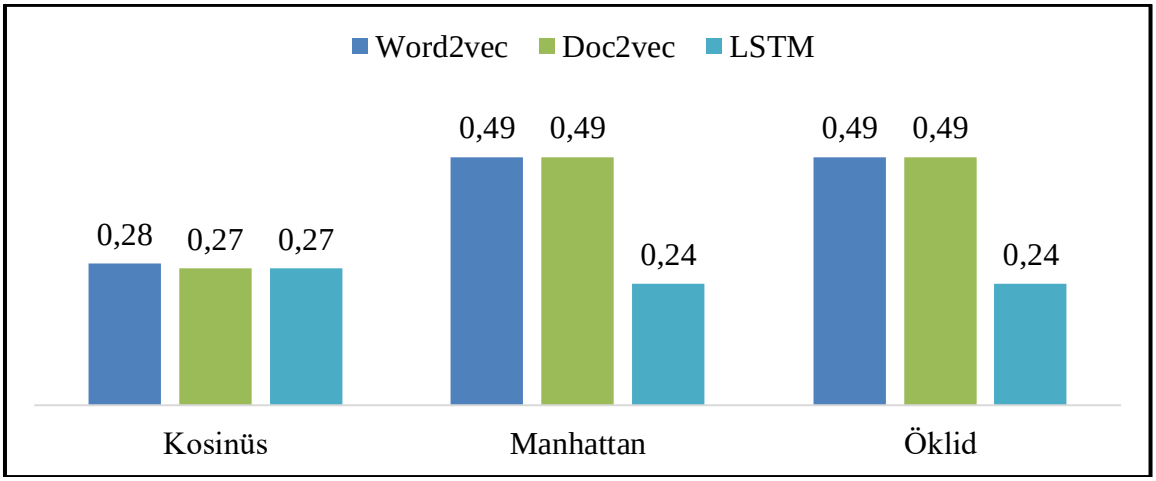
Bu çalışmada Çizelge 6.1'de görüleceği üzere literatürdeki diğer çalışmalara kıyasla daha yüksek bir bilet üst ve alt alan sınıflandırma doğruluk oranı elde edilmiştir. Bilet öncelik sınıflandırmasında ise Tekin ve Tunalı [23] Random Forest yöntemini kullanarak %74,5 F1 puanı elde ederken bu çalışmada K-NN yöntemi kullanılarak %97,5 F1 puanı ile daha iyi bir sonuç elde edilmiştir. Böylece, yardım masası verileri için seçilen veri ön işleme adımlarının kullanılmasıyla literatürdeki çalışmalara kıyasla daha başarılı bilet alan ve öncelik sınıflandırması gerçekleştirilmiştir. Çünkü bilet açıklamaları doğal dil, HTML ifadeleri, URL ifadeleri ve yanlış yazılmış ifadeler içerebilir; dengesiz bir veri kümesinde bulunabilir ve yüksek boyutlu vektörler sunabilir.

Rastgele Aşırı Örnekleme yönteminin kullanılması bilet üst alan sınıflandırma modelinin 2 saat 19 dakika 9 saniye süren eğitim süresini 25 saat 2 dakika 33 saniyeye çıkarırken doğruluk oranını ise %78,33'den %94,40'a yükseltmiştir. Bilet alt alan sınıflandırması için doğruluk oranında Rastgele Aşırı Örnekleme yönteminin sağladığı en büyük artış %42,47'den %88,27'ye %45,80, en küçük artış ise %82,80'den %98,92'ye %16,12 olmuştur. Rastgele Aşırı Örnekleme yönteminin kullanılması bilet öncelik sınıflandırma modeli eğitim süresini 41 saniyeden 19 dakika 48 saniyeye yükseltirken doğruluk oranını ise %88,44'den %97,35'e yükseltmiştir. Öznitelik sayısındaki artış, bilet sınıflandırma başarı oranı doyuma ulaşana kadar yeniden örnekleme yönteminin sağladığı katkıyı da artırmıştır. Bilet açıklamalarından elde edilen vektörlerin boyutunun belirli bir düzeye indirilmesinin eğitim süresini kısalttığı ancak öznitelik sayısının belirli bir düzeyin altına indirilmesi başarı oranını düşürdüğü gözlemlenmiştir. Dolayısıyla başarı oranı ve eğitim süresi açısından öznitelik sayısı deneysel olarak belirlenmelidir.

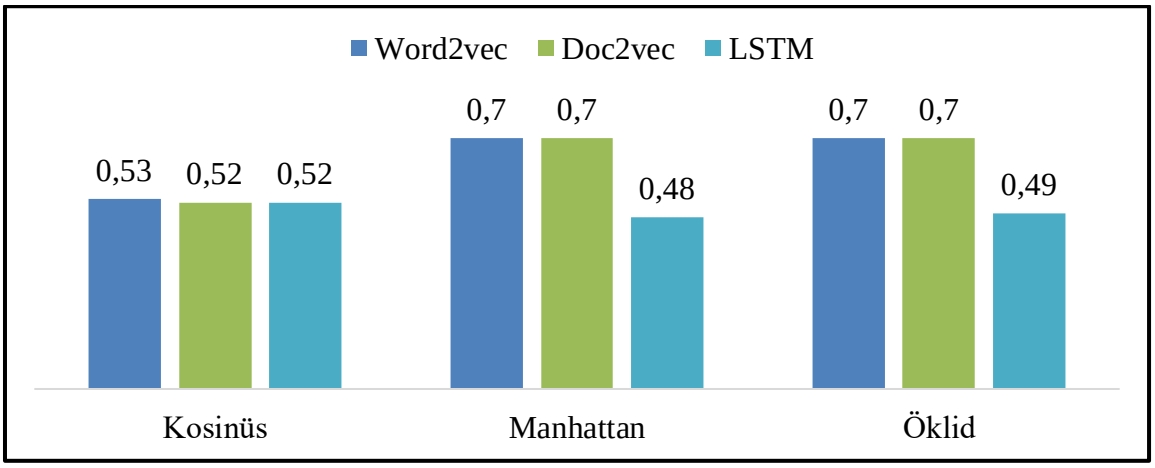
Çizelge 6.1. Bilet alan sınıflandırmasında bu çalışma ile literatürdeki incelenen çalışmaların doğruluk oranlarının karşılaştırılması

Çalışma	Bilet Üst Alanı (%)	Bilet Alt Alanı (%)
SMO [16]	-	81,4
Random Forest [17]	89,37	-
Bagging-SVM [18]	87,78	-
SVM [19]	89	-
SVM [20]	~91	~ [55-95]
Bagging-DT [21]	92,04	-
Random Forest [22]	82	50
Bu çalışma (SVM)	94,4	[88,27-98,92]

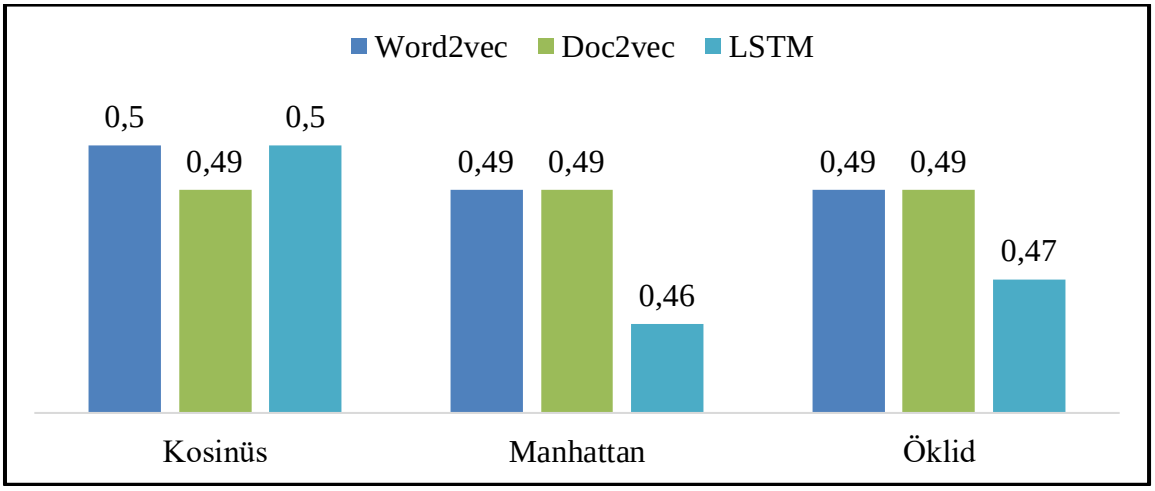
Bilet benzerliğini hesaplamak için Word2vec, Doc2vec ve Siyam LSTM modelleri kullanılmıştır. Siyam LSTM modeli, kosinüs ve öklid uzaklıklarına kıyasla manhattan uzaklığı ile daha yüksek bir bilet benzerlik performansı göstermiştir. Benzer şekilde MaLSTM modeli, Word2vec ve Doc2vec modellerinden daha yüksek bir performans göstermiştir. Test veri kümesi için MaLSTM modeli kullanılarak Şekil 6.1'de görüldüğü gibi 0,24 MSE, Şekil 6.2'de görüldüğü gibi 0,48 RMSE, Şekil 6.3'de görüldüğü gibi 0,46 MAE, Şekil 6.4'de görüldüğü gibi 0,25 PCC, Şekil 6.5'de görüldüğü gibi 0,25 SRCC elde edilmiştir.



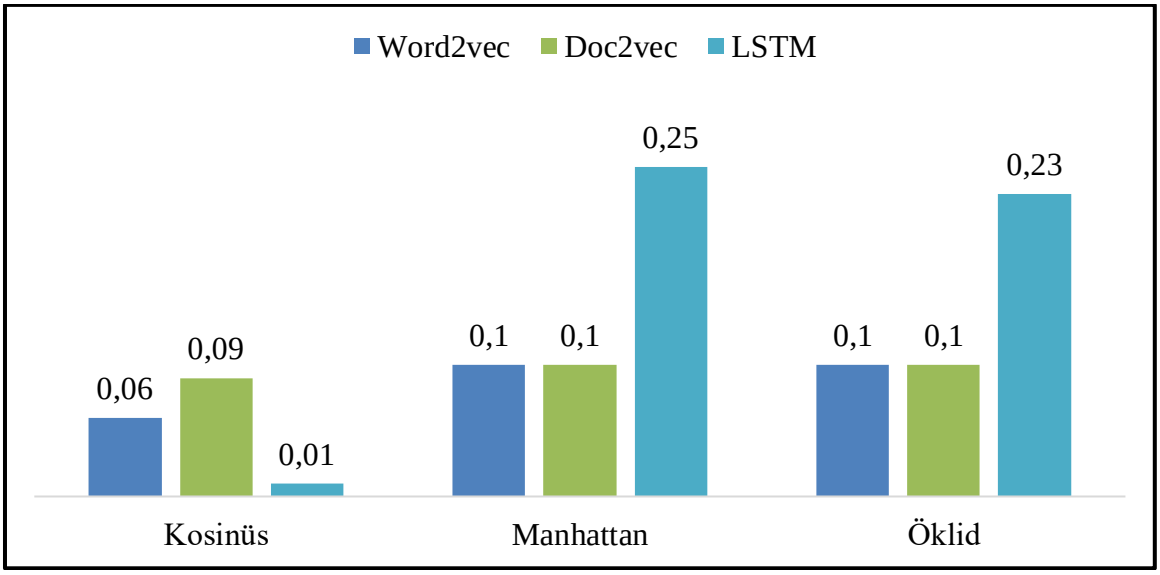
Şekil 6.1. Vektörleştirme yöntemleri ve uzaklık türleri arasında test veri kümesi ortalama kare hata karşılaştırması



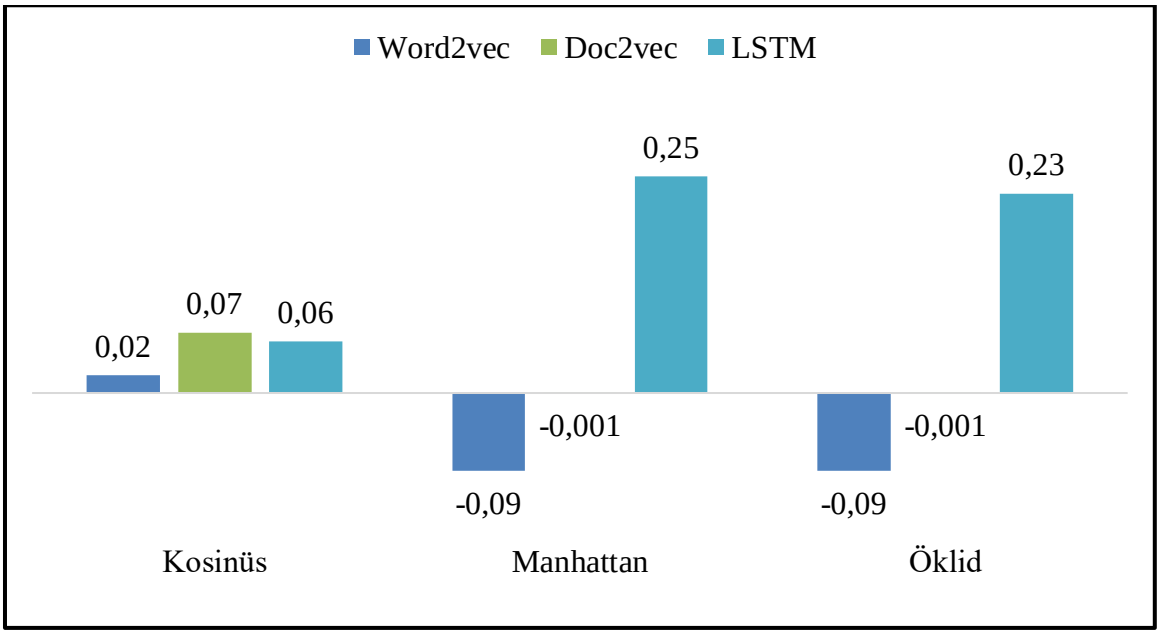
Şekil 6.2. Vektörleştirme yöntemleri ve uzaklık türleri arasında test veri kümesi kök ortalama kare hata karşılaştırması



Şekil 6.3. Vektörleştirme yöntemleri ve uzaklık türleri arasında test veri kümesi ortalama mutlak hata karşılaştırması



Şekil 6.4. Vektörleştirme yöntemleri ve uzaklık türleri arasında test veri kümesi Pearson korelasyon katsayısı karşılaştırması



Şekil 6.5. Vektörleştirme yöntemleri ve uzaklık türleri arasında test veri kümesi Spearman'ın sıralama korelasyon katsayısı karşılaştırması

Bu çalışmanın literatürdeki anlamsal metin benzerliği üzerine yapılan çalışmalarla karşılaştırılması Çizelge 6.2'de gösterilmiştir. Anlamsal benzerlik üzerine yapılan çalışmalara yakın bir sonucun elde edildiği görülmüştür. Manuel etiketlenen veri miktarının artmasıyla elde edilen sonuçların daha yüksek bir noktaya çıkacağı ise aşikârdır.

Çizelge 6.2. Anlamsal benzerlikte bu çalışma ile literatürdeki incelenen çalışmaların benzerlik performanslarının karşılaştırılması

Çalışmalar	MSE
MaLSTM [35]	0,2286
Dependency Tree-LSTM [37]	0,2532
Skip-thought+COCO [29]	0,2561
ConvNet [62]	0,2606
D-LSTM [36]	0,3442
MaLSTM [36]	0,3693
Bu çalışma (MaLSTM)	0,2350

Sonuçlar, yardım masası sisteminin bilet alan sınıflandırma ve öncelik verme süreçlerinin makine öğrenmesi ve doğal dil işleme yöntemleri kullanılarak otomatikleştirilebildiğini, taleplerin müşteri hizmetleri temsilcilerine hızlı bir şekilde ulaştırılabildiğini ve kısa sürede çözümlenebildiğini; etkin kullanım, iş süreçlerinin sürekliliği, emek tasarrufu ve kullanıcı memnuniyeti sağlanabileceğini göstermiştir.

KAYNAKLAR

1. Agarwal, S., Sindhgatta, R. and Sengupta, B. (2012, 12-16 August). *SmartDispatch: enabling efficient ticket dispatch in an IT service environment.* International Conference on Knowledge Discovery and Data Mining, Beijing, 1393–1401.

2. Schad, J., Sambasivan, R. and Woodward, C. (2022). Predicting help desk ticket reassignments with graph convolutional. *Machine Learning with Applications*, 7 (100237), 1-9.

3. Samarakoon, L., Kumarawadu, S. and Pulasinghe, K. (2011, 16-19 August). *Automated question answering for customer helpdesk applications*. 6th International Conference on Industrial and Information Systems, Sri Lanka, 328-333.

4. Altınel, B. and Ganiz, M. C. (2018). Semantic text classification: A survey of past and recent advances. *Information Processing and Management*, 54 (6), 1129-1153.

5. HaCohen-Kerner, Y., Dilmon, R., Hone, M. and Ben-Basan, M. A. (2019). Automatic classification of complaint letters according to service. *Information Processing and Management*, 56 (102102), 1-20.

6. Hark, C. et al. (2017, 16-17 September). *Doğal dil işleme yaklaşımları ile yapısal olmayan dokümanların benzerliği*. International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, 1-6.

7. Boufrida, A. and Boufaida, Z. (2022). Rule extraction from scientific texts: Evaluation in the specialty of gynecology. *Journal of King Saud University - Computer and Information Sciences*, 34 (4), 1150-1160.

8. Ayata, D., Saraçlar, M. and Özgür, A. (2017, 15-18 May). *Turkish tweet sentiment analysis with word embedding and machine learning*. 25th Signal Processing and Communications Applications Conference (SIU), Antalya, 1-4.

9. Moukrim, C., Abderrahim, T., Tarik, A. and Benlahmer, E. H. (2021). An innovative approach to autocorrecting grammatical errors in Arabic texts. *Journal of King Saud University - Computer and Information Sciences*, 33 (4), 476-488.

10. Zouaoui, S. and Rezeg, K. (2022). Multi-Agents Indexing System (MAIS) for plagiarism detection. *Journal of King Saud University - Computer and Information Sciences*, 34 (5), 2131-2140.

11. Khan, S. and Shamsi, J. A. (2021). Health Quest: A generalized clinical decision support system with multi-label classification. *Journal of King Saud University - Computer and Information Sciences*, 33 (1), 45-53.

12. Anand, D. and Wagh, R. (2022). Effective deep learning approaches for summarization of legal texts. *Journal of King Saud University - Computer and Information Sciences*, 34 (5), 2141-2150.

13. Mohasseb, A., Bader-El-Den, M. and Cocea, M. (2018). Question categorization and classification using grammar based. *Information Processing and Management*, 54 (6), 1228-1243.

14. Elnagar, A., Al-Debsi, R. and Einea, O. (2020). Arabic text classification using deep learning models. *Information Processing and Management*, 57 (102121), 1-17.

15. Uysal, A. K. and Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing and Management*, 50 (1), 104-112.

16. Al-Hawari, F. and Barham, H. (2021). A machine learning based help desk system for IT service management. *Journal of King Saud University - Computer and Information Sciences*, 33 (6), 702-718.

17. Paramesh, S. P. and Shreedhara, K. S. (2019). Building intelligent service desk systems using AI. *International Journal of Science, Technology, Engineering and Management*, 1 (2), 1-8.

18. Paramesh, S. P. and Shreedhara, K. S. (2019). IT help desk incident classification using classifier ensembles. *Journal on Soft Computing (ICTACT)*, 9 (4), 1980-1987.

19. Pereira, R., Ribeiro, R. and Silva, S. (2018, 13-16 June). *Machine learning in incident categorization automation*. 13th Iberian Conference on Information Systems and Technologies (CISTI), Caceres, 1-6.

20. Altıntaş, M. and Tantuğ, A. C. (2014, 15-16 September). *Machine learning based ticket classification in issue tracking systems*. 2nd International Conference on Artificial Intelligence and Computer Science (AICS), Astanaanyar, 33-44.

21. Paramesh, S. P. and Shreedhara, K. S. (2018, 20-22 December). *Classifying the unstructured IT service desk tickets using ensemble of classifiers*. 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, 221-227.

22. Miliano, A., Steven, I., Kosim, P. K., Jayadi, R. and Mauritsius, T. (2020, 18-21 December). *Machine learning-based automated problem categorization in a helpdesk ticketing application*. 8th International Conference on Orange Technology (ICOT), Daegu, 1-6.

23. Tekin, M. C. and Tunalı, V. (2019). Prioritization of software development demands with text mining technique. *Journal of Engineering Sciences*, 25 (5), 615-620.

24. Walek, B. (2017, 15-18 November). *Intelligent system for ordering incidents in helpdesk system*. 21st International Computer Science and Engineering Conference (ICSEC), Bangkok, 1-5.

25. Nguyen, H. T., Duong, P. H. and Cambria, E. (2019). Learning short-text semantic similarity with word embeddings and external knowledge sources. *Knowledge-Based Systems*, 182 (104842), 1-9.

26. Zhang, H. and Zhou, L. (2019, 19-21 October). *Similarity judgment of civil aviation regulations based on Doc2Vec deep learning algorithm*. 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Huaqiao, 1-8.

27. Qurashi, A. W., Holmes, V. and Johnson, A. P. (2020, 8-10 August). *Document processing: Methods for semantic text similarity analysis*. International Conference on Innovations in Intelligent Systems and Applications (INISTA), Biarritz, 1-6.

28. Zahrotun, L. (2016). Comparison jaccard similarity, cosine similarity and combined both of the data clustering with Shared Nearest Neighbor method. *Computer Engineering and Applications*, 5 (1), 11-18.

29. Kiros, R. et al. (2015). Skip-Thought vectors. *Advances in Neural Information Processing Systems*, 28, 1-11.

30. Kenter, T. and Rijke, M. D. (2015, 18-23 October). *Short text similarity with word embeddings*. 24th ACM International on Conference on Information and Knowledge Management (CIKM), Melbourne, 1411-1420.

31. Imtiaz, Z., Umer, M., Ahmad, M., Ullah, S., Choi, G. S. and Mehmood, A. (2019). Duplicate questions pair detection using siamese MaLSTM. *IEEE Access*, 8, 21932 - 21942.

32. Srinidhi, H., Siddesh, G. and Srinivasa, K. (2020). A hybrid model using MaLSTM based on recurrent neural networks with support vector machines for sentiment analysis. *Engineering and Applied Science Research*, 47 (3), 232–240.

33. Guo, W., Zeng, Q., Duan, H., Ni, W. and Liu, C. (2021). Process-extraction-based text similarity measure for emergency response plans. *Expert Systems with Applications*, 183 (115301), 1-14.

34. Vine, L. D., Zuccon, G., Koopman, B. and Sitbon, L. (2014, 3-7 November). *Medical semantic similarity with a neural language model*. 23rd ACM International Conference on Conference on Information and Knowledge Management, Shanghai, 1819-1822.

35. Thyagarajan, A. and Mueller, J. (2016, 12-16 February). *Siamese recurrent architectures for learning sentence similarity*. Thirtieth AAAI Conference on Artificial Intelligence (AAAI), Arizona, 2786-2792.

36. Zhu, W., Yao, T., Ni, J., Wei, B. and Lu, Z. (2018). Dependency-based siamese long short-term memory network for learning sentence representations. *PLoS ONE*, 13 (3), 1-14.

37. Tai, K. S., Socher, R. and Manning, C. D. (2015, 26-31 July). *Improved semantic representations from Tree-Structured Long Short-Term Memory Networks*. 7th International Joint Conference on Natural Language Processing, Beijing, 1556-1566.

38. Ghasemi, N. and Momtazi, S. (2021). Neural text similarity of user reviews for improving collaborative filtering recommender systems. *Electronic Commerce Research and Applications*, 45 (101019), 1-9.

39. Lekshmy, V. G., Anusree, P. K. and Varunika, V. S. (2018, 19-22 September). *An implementation of genetic algorithm for clustering help desk data for service automation*. International Conference on Advances in Computing, Communications and Informatics, Karnataka, 952-956.

40. Apriyanto, R., Rahmawati, I. I., Setiawan, H., Budi and Haryono (2019, 16-17 October). *Application of case-based reasoning in helpdesk systems*. International Conference on Informatics and Computing (ICIC), Semarang, 1-7.

41. Liu, T., Li, S., Gu, X., Wang, T., Lu, P., Cao, X., Yang, X., Wang, W., Lv, H. and Feng, C. (2019, 11-12 October). *Historical similar ticket matching and extraction used for power grid maintenance work ticket decision making*. 3rd International Conference on Data Science and Business Analytics (ICDSBA), Istanbul, 302-307.

42. Wang, D., Li, T., Zhu, S. and Gong, Y. (2011). iHelp: an intelligent online helpdesk system. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41 (1), 173-182.

43. Ramos-Pérez, I., Arnaiz-González, A., Rodríguez, J. J. and García-Osorio, C. (2022). When is resampling beneficial for feature selection with imbalanced wide data?. *Expert Systems with Applications*, 188 (116015), 1-12.

44. Zhang, Q., Chen, Z., Zeng, Y., Gao, H., Wei, Q., Luo, T. and Wang, Z. (2021). Data-driven approaches for time series prediction of daily production in the Sulige tight gas field, China. *Artificial Intelligence in Geosciences*, 2, 165-170.

45. Al-Anzi, F. S. and AbuZeina, D. (2018). Beyond vector space model for hierarchical Arabic text. *Information Processing and Management*, 54 (1), 105-115.

46. Wang, Y., Li, J., Pei, Y., Ma, Z., Jia, Y. and Wei, Y. C. (2022). An adaptive high-voltage direct current detection algorithm using cognitive wavelet transform. *Information Processing & Management*, 59 (102867), 1-24.

47. Sasada, T., Liu, Z., Baba, T., Hatano, K. and Kimura, Y. (2020). A resampling method for imbalanced datasets considering noise and overlap. *Procedia Computer Science*, 176, 420-429.

48. Pereira, R. M., Costa, Y. M. and Silla, C. N. (2021). Toward hierarchical classification of imbalanced data using. *Information Sciences*, 578, 344-363.

49. Xu, J., Tang, L. and Li, T. (2016). System situation ticket identification using SVMs ensemble. *Expert Systems with Applications*, 60, 130-140.

50. Katoch, S., Singh, V. and Tiwary, U. S. (2022). Indian sign language recognition system using SURF with SVM and CNN. *Array*, 14 (100141), 1-9.

51. Otchere, D. A., Ganat, T. O. A., Gholami, R. and Ridha, S. (2021). Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models. *Journal of Petroleum Science and Engineering*, 200 (108182), 1-20.

52. Do, D. T. and Khanh, N. Q. (2019). A sequence-based approach for identifying recombination spots in Saccharomyces cerevisiae by using hyper-parameter optimization in FastText and support vector machine. *Chemometrics and Intelligent Laboratory Systems*, 194 (103855), 1-7.

53. Clercq, D. D., Diop, N. F., Jain, D., Tan, B. and Wen, Z. (2019). Multi-label classification and interactive NLP-based visualization of electric vehicle patent data. *World Patent Information*, 58 (101903), 1-12.

54. Chen, Z., Zhou, L. J., Li, X. D., Zhang, J. N. and Huo, W. J. (2020). The Lao text classification method based on KNN. *Procedia Computer Science*, 166, 523-528.

55. Bhuvaneswari, P. and Therese, A. B. (2015). Detection of Cancer in Lung with K-NN Classification Using Genetic Algorithm. *Procedia Materials Science*, 10, 433-440.

56. Maleki, N., Zeinali, Y. and Niaki, S. T. A. (2021). A K-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection. *Expert Systems with Applications*, 164 (113981), 1-7.

57. Debnath, S., Sinha, N. and Bhowmik, B. B. (2022). ML based modulation format identifier using K-NN algorithm. *Materials Today: Proceedings*, 65 (5), 2626-2630.

58. Djordjević, B., Mane, A. S. and Krmac, E. (2021). Analysis of dependency and importance of key indicators for railway sustainability monitoring: A new integrated approach with DEA and Pearson correlation. *Research in Transportation Business & Management*, 41 (100650), 1-12.

59. Jayaraman, A. K. and Murugappan, A. (2018). Aspect-based opinion ranking framework for product reviews using a Spearman's rank correlation coefficient method. *Information Sciences*, 460, 23-41.

60. Zhang, W. Y., Wei, Z. W., Wang, B. H. and Han, X. P. (2016). Measuring mixing patterns in complex networks by Spearman rank correlation coefficient. *Physica A: Statistical Mechanics and its Applications*, 451, 440-450.

61. Prion, S. and Haerling, K. A. (2014). Making Sense of Methods and Measurement: Spearman-Rho Ranked-Order correlation Coefficient. *Clinical Simulation in Nursing*, 10 (10), 535-536.

62. He, H. and Gimpel, K. (2015, 17-21 September). *Multi-Perspective Sentence Similarity Modeling with Convolutional Neural Network*. Conference on Empirical Methods in Natural Language Processing, Lisbon, 1576-1586.

63. Gomaa, W. and Fahmy, A. (2013). A Survey of Text Similarity Approache. *International Journal of Computer Applications*, 68 (13), 13-18.

64. Beltrán-Pérez, G., Castillo-Mixcóatl, J., Muñoz-Aguirre, S. and Rodríguez-Garciapiña, J. L. (2021). Application of the principal components analysis technique to optical fiber sensors for acetone detection. *Optics & Laser Technology*, 143 (107314), 1-7.

65. Raja, K. and Natarajan, J. (2018). Mining protein phosphorylation information from biomedical literature using NLP parsing and Support Vector Machines. *Comput Methods Programs Biomed*, 160, 57-64.

66. Hongbo, S., Ying, Z., Yuwen, C., Suqin, J. and Yuanxiang, D. (2022). Resampling algorithms based on sample concatenation for imbalance learning. *Knowledge-Based Systems*, 245 (108592), 1-19.

Gazili olmak ayrıcalıktır