T.C.
İSTANBUL AYDIN ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

ONTOLOJİ BOYUT İNDİRGEMELİ DERİN ÖĞRENME YAKLAŞIMI: YAPISAL OLMAYAN DOKÜMANLARIN SINIFLANDIRILMASI ÜZERİNE BİR UYGULAMA

DOKTORA TEZİ

İLKAY YELMEN

Bilgisayar Mühendisliği Ana Bilim Dalı
Bilgisayar Mühendisliği Programı

HAZİRAN, 2023

T.C.
İSTANBUL AYDIN ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

ONTOLOJİ BOYUT İNDİRGEMELİ DERİN ÖĞRENME YAKLAŞIMI: YAPISAL OLMAYAN DOKÜMANLARIN SINIFLANDIRILMASI ÜZERİNE BİR UYGULAMA

DOKTORA TEZİ

İLKAY YELMEN
(Y1813.610003)

Bilgisayar Mühendisliği Ana Bilim Dalı
Bilgisayar Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Ali GÜNEŞ

HAZİRAN, 2023

ONAY FORMU

ONUR SÖZÜ

Doktora tezi olarak sunduğum "Ontoloji Boyut İndirgemeli Derin Öğrenme Yaklaşımı: Yapısal Olmayan Dokümanların Sınıflandırılması Üzerine Bir Uygulama" adlı çalışmanın, tezin proje safhasından sonuçlanmasına kadarki bütün süreçlerde bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurulmaksızın yazıldığını ve yararlandığım eserlerin Kaynakça'da gösterilenlerden oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve onurumla beyan ederim. (16/06/2023)

İlkay YELMEN

ÖNSÖZ

Doktora tezim boyunca benden yardımlarını esirgemeyen, bilgi ve deneyimlerini benimle paylaşan değerli hocam ve tez danışmanım Prof. Dr. Ali GÜNEŞ'e, çalışmamda destekleri olan Prof. Dr. Metin ZONTUL ve Prof. Dr. Zafer ASLAN'a ve ayrıca 2211-C Öncelikli Alanlara Yönelik Yurt İçi Doktora Burs Programı kapsamında almış olduğum desteklerden dolayı TÜBİTAK'a teşekkürü bir borç bilirim.

Son olarak, hayatımın her döneminde olduğu gibi tez çalışmalarım boyunca da destek olan aileme sonsuz teşekkür eder, saygılarımı sunarım.

Haziran, 2023 İlkay YELMEN

ONTOLOJİ BOYUT İNDİRGEMELİ DERİN ÖĞRENME YAKLAŞIMI: YAPISAL OLMAYAN DOKÜMANLARIN SINIFLANDIRILMASI ÜZERİNE BİR UYGULAMA

ÖZET

Yapısal olmayan veriler önceden tanımlanmış bir veri modeli içermediği için düzensizdir. İnternet ortamında yapısal olmayan metinsel dokümanların artması ile birlikte bu dokümanların yönetilebilirliği de zorlaşmıştır. Sınıf etiketinden yoksun bir şekilde her geçen gün sürekli artan verinin doğru olarak manuel bir şekilde etiketlenmesi oldukça zordur. Bu zorluğu kolaylaştırmak için yapay zeka yöntemlerinin kullanılması gerekmektedir. Araştırmacılar bu zamana kadar bir çok makine öğrenimi ve derin öğrenme modelini farklı türde veriler üzerinde uygulamıştır. Bu modellerin başarılı olarak çalışmasında verinin eğitim için düzgün hale getirilmesi oldukça önemlidir. Bu aşamada veri içerisinden eğitim için anlam ifade etmeyen sözcüklerin çıkarılması ve eğitimin daha iyi yapılmasını sağlamak amacıyla veri üzerinde çeşitli yöntemler uygulanır. Burada yapılacak detaylı çalışmalar modelin başarısına doğrudan etki etmektedir. Bunun yanı sıra öznitelik sayısının fazla olması ve vektör uzayının büyüklüğü hem model başarısını hem de performansı etkilemektedir. Ayrıca sınıf etiketinin fazla olması da eğitimi zorlaştırmaktadır. Yapılan literatür araştırmasında sınıf etiketinin az ve veri sayısının fazla olduğu araştırmalar daha fazla olduğu görülmüştür. Makine öğrenimi ve derin öğrenme modelleri ile veri sayısının fazla olduğu ve sınıf etiketinin az olduğu veriler üzerinde daha kolay öğrenme gerçekleştirilip, daha başarılı sonuçlar alınabilmektedir. Ancak veri sayısının daha az, sınıf etiketinin ise fazla ve dengesiz olduğu durumlarda öğrenme zorlaşmaktadır. Bir de bunlar yapısal olmayan metinsel veriler ise öğrenme daha da zorlaşmaktadır. Bu tez çalışmasında yapısal olmayan ve 7 sınıf içeren haber verisi kullanılarak, sınıflandırma başarısını artırmaya yönelik deneysel çalışmalar yapılmıştır. Çalışmada detaylı veri önişleme yapıldıktan sonra farklı kelime temsil yöntemleri ile makine öğrenimi ve derin öğrenme sınıflandırma yöntemleri ile model başarısı ölçülmüştür. Ayrıca WordNet ontolojisi de kullanılarak kelimeler anlamsal

yönden de değerlendirilip, öznitelik boyut indirgemesi de yapılmıştır. Yapılan çalışmalar sonucunda metin sınıflandırma probleminde çok sayıda, dengesiz sınıf etiketi olan ve az sayıda veri üzerinde yüksek doğrulukta sınıflandırma yapan ontoloji ve derin öğrenme tabanlı hibrit bir yaklaşım önerilmiştir. WordNet ontolojisi ve BERT kullanılarak sağlanan çözüm önerisi özgün olup, yapısal olmayan metinsel dokümanların sınıflandırılmasında bir yol gösterici olmaktadır.

Anahtar Kelimeler: Çok Sınıflı Sınıflandırma, WordNet, Ontoloji, Derin Öğrenme, BERT

DEEP LEARNING APPROACH WITH ONTOLOGY BASED DIMENSION REDUCTION: AN APPLICATION ON CLASSIFICATION OF UNSTRUCTURED DOCUMENTS

ABSTRACT

Unstructured data is disordered because it does not contain a predefined data model. With the increase of unstructured textual documents on the Internet, the manageability of these documents has become more difficult. Accurate manual labeling of ever-increasing data, devoid of class labeling, is extremely difficult. Artificial intelligence methods need to be used to facilitate this challenge. Researchers have so far applied many machine learning and deep learning models on different types of data. It is very important to make the data smooth for training in the successful operation of these models. At this stage, various methods are applied on the data in order to remove the words that do not make sense for training and to ensure that the training is done better. Detailed studies to be made here directly affect the success of the model. In addition, the large number of features and the size of the vector space affect both model success and performance. Besides, too many class labels make training difficult. In the literature research, it has been seen that there are more studies in which the class label is low and the number of data is high. With machine learning and deep learning models, it is easier to learn and more successful results can be obtained on data with a large number of data and a low class label. However, learning becomes difficult in cases where the number of data is less and the class label is large and inbalanced. Also, if these are unstructured textual data, learning becomes even more difficult. In this thesis, experimental studies were carried out to increase the classification success by using unstructured news data containing 7 classes. In the study, after detailed data preprocessing, model success was measured with different word representation methods, machine learning and deep learning classification methods. In addition, by using WordNet ontology, the words were also evaluated in terms of semantics and feature dimension reduction was made. As a result of the studies, a hybrid approach based on ontology and deep learning has been proposed in

the text classification problem, which has a large number of imbalanced class labels and makes high accuracy classification on a small number of data. The solution proposal provided by using WordNet ontology and BERT is unique and guides the classification of unstructured textual documents.

Keywords: Multi Class Classification, WordNet, Ontology, Deep Learning, BERT

İÇİNDEKİLER

KISALTMALAR

BERT : Bidirectional Encoder Representations from Transformers

BoW : Bag of Words

CBOW : Continuous Bag of Words

CNN : Convolutional Neural Network

DDİ : Doğal Dil İşleme

DistilBERT : Distilled BERT

LDA : Linear Discriminant Analysis

LSTM : Long Short-Term Memory

MLP : Multilayer Perceptron

PCA : Principal Component Analysis

PV-DBOW : Distributed Bag Of Words of Paragraph Vector

PV-DM : Distributed Memory Model of Paragraph Vectors

RF : Random Forest

RNN : Reccurrent Neural Network

RoBERTa : A Robustly Optimized BERT Pretraining Approach

SVM : Support Vector Machine

TF-IDF : Term Frequency–Inverse Document Frequency

ÇİZELGELER LİSTESİ

ŞEKİLLER LİSTESİ

I. GİRİŞ

İnternetin hızla gelişmesi ile birlikte büyük hacimli veriler de artmaktadır. Özellikle yapısal olmayan verilerdeki artış verilerin yönetimini zorlaştırmaktadır. Verileri çeşitli amaçlarla kullanabilmek için de sınıflandırmaya ihtiyaç duyulmaktadır. Sürekli artan verinin çeşitli analizler ve değerlendirmeler için manuel olarak sınıflandırılması da zor olduğundan dolayı otomatik sınıflandırma yapan yöntemlere ihtiyaç duyulmaktadır.

Otomatik doküman sınıflandırması, bilgileri keşfetmet, yönetmek, filtrelemek ve işlemek için kullanılmaktadır. Sınıflandırmada dokümanlara önceden atanmış sınıf etiketi buluunmakta ve bu etiketlere göre yapılan eğitim sonrası yeni gelen dokümanın sınıf etiketi tahmin edilmektedir. Dokümanlar, elektronik yayınlar, elektronik kitaplar, e-postalar, haberler, dijital kütüphaneler, akademik makaleler, web sayfaları vb. olabilir. Dokümanların otomatik olarak sınıflandırılmasında çeşitli makine öğrenimi algoritmaları kullanılmaktadır (Çobanoğlu, 2015).

Doküman sınıflandırma, cümle sınıflandırmasından çok farklıdır. Belgeler genellikle birden fazla cümleden olmakta ve çok fazla kelime içermektedir. Cümleler arasında karmaşık ve belirsiz anlamsal ilişkiler bulunmasından dolayı, doküman sınıflandırma cümle sınıflandırmadan daha zordur (Kong et al., 2022).

Derin öğrenme modelleri, çeşitli Doğal Dil İşleme (DDİ) görevlerinde etkileyici ilerlemeler göstermiştir. Doküman sınıflandırmada kullanılan bu modeller, evrişimli sinir ağları (CNN'ler) (Kim, 2014) (Zhang et al., 2015), tekrarlayan sinir ağları (RNN'ler) (Irsoy ve Cardie, 2014) (Yogatama et al., 2017), kapılı tekrarlayan birim (GRU) ağları (Chung et al., 2014) ve uzun kısa süreli bellek (LSTM) ağlarıdır (Tai et al., 2015). CNN'ler bilgisayarlı görüde başarılı olmuş ve belge sınıflandırmasında da kullanılmıştır. Bir belgedeki her tokenın sınıflandırmaya, öz-dikkat (Bahdanau et al., 2014) ve dinamik yönlendirmeye (Sabour et al., 2017) (Gong et al., 2018) eşit şekilde katkıda bulunmadığı varsayılarak, metni otomatik olarak hizalayan ve önemli tokenları vurgulayan bir süreç önerilmiştir. Bu da CNN'lerin, GRU ve LSTM ağlarının

performansını daha da artırmaktadır. Ayrıca, belge sınıflandırması için hiyerarşik dikkat ağı (Yang et al., 2016) önerilmiştir. GRU ile sırasıyla cümle düzeyinde ve belge düzeyinde semantik modelleme yapmıştır. Ancak bu, cümle bağlamında tokenların sözdizimsel bağımlılıklarının kaybolmasına yol açabilir.

Yakın zamanda ortaya çıkan BERT (Devlin et al., 2019), ALBERT (Lan et al., 2019), ve RoBERTa (Liu et al., 2019), gibi önceden eğitilmiş dil modelleri çeşitli NLP işlerinde başarılı olmuştur. Bu modeller sıfırdan bir model oluşturma ihtiyacını ortadan kaldırarak, transfer öğrenme ile metinlerin yüksek kaliteli bağlamsal temsillerini öğrenmek için dönüştürücüleri (Vaswani et al., 2017) kullanır.

Derin öğrenme yöntemlerindeki gelişmelere paralel olarak metin sınıflandırma alanında da gelişmeler olmuştur. BERT modeli ile birlikte sınıflandırma modelinin başarısını artırmak için birçok deneysel çalışma yapılmıştır. (Lu et al., 2020) yapmış oldukları araştırmada, BERT gibi dikkat mekanizmalarını kullanan yöntemlerin belgedeki bağlamsal bilgileri yakalama yeteneğine sahip olduğundan bahsetmiştir. Çalışmada, BERT'in kabiliyetini bir Vocabulary Graph Convolution Network (VGCN) ile birleştiren VGCNBERT modeli önerilmiştir. Yapılan deneylerde en yüksek sonuç F1 91,93 olarak 2 sınıflı SST-2 veri seti ile elde edilmiştir. Kelime grafiği BERT'e yararlı genel bilgiler getirdiğinden, WordNet'in kullanımı gelecekteki çalışmalarda yazarlar tarafından da önerilmektedir. Bu çalışmadan WordNet kullanımının faydalı olabileceği sonucuna varılmıştır.

Son zamanlarda yapılan bazı çalışmalarda, WordNet sözcük ontolojisi ve BERT dil modeli, belge sınıflandırmaları için birlikte kullanılmaktadır; öyle ki, WordNet'in rolü path2vec ve wnet2vec gibi kelime gömmeleri olarak anlamsal bilgi iken, BERT'nin rolü dokümanların yerel öznitelik bilgisini çıkarmak ve sınıflandırmaktır (Lu et al., 2020; Barbouch et al., 2021).

Bu çalışmada ise sınıflandırma modelinin dengesiz veri seti üzerindeki başarısını artırmak için alan ontolojisi yerine boyut indirgeme amaçlı WordNet sözlükbilimi ontolojisi ve BERT sınıflandırma modeli önerilmektedir. Deneylerde makine öğrenimi algoritmaları, BERT ve DistilBERT yöntemlerinden önce WordNet, 7 sınıflı ve dengesiz veri seti üzerinde uygulanmıştır. Yeni hibrit modelin avantajı, öznitelik vektör boyutunu küçültmesi, cümledeki kelimelerin anlamsal benzerliklerini

yakalaması ve yapısal olmayan, dengesiz ve yüksek boyutlu çok sınıflı belgelerde daha yüksek sınıflandırma başarısı sağlamasıdır.

Tez 7 bölümden oluşmakta olup 2. bölümde literatür taraması, 3. bölümde yöntem, 4. bölümde deneysel çalışma, 5. bölümde deneyler ve sonuç, 6. bölümde sonuç ve öneriler ve son bölümde ise tez çalışmasında yararlanılan kaynaklara yer verilmiştir.

A. Tezin Literatüre Katkısı

Bu tezin literatüre katkıları şunlardır:

- WordNet ontolojisinin öznitelik boyut indirgeme için kullanılarak, dengesiz ve çok sınıflı (7 sınıf etiketli) veriler üzerindeki sınıflandırma başarısını değerlendirmek.
- Farklı sayılara sahip çok sınıflı dengesiz veri seti üzerinde makine öğrenmesi ve derin öğrenme sınıflandırma algoritmalarının performansını karşılaştırmak.
- Klasik makine öğrenimi yöntemlerinden önce birlikte kullanılan alan ontolojisi ve bazı kelime gömme yöntemlerinin yerine boyut indirgeme amacıyla kullanılan WordNet sayesinde bazı sınıflandırma modellerinin başarısı artmaktadır.
- En yüksek başarı, boyut küçültme için kullanılan WordNet ve BERT algoritmasının hibrit olarak kullanılmasıyla elde edilmiştir. Yapılan deneysel çalışmalarda sözlükbilime dayalı özellik boyut indirgemesinin sınıflandırma başarısını artırdığı görülmüştür.

II. LİTERATÜR TARAMASI

Veri ön işleme, öznitelik çıkarma aşamasına girdi sağlamak için metinsel verinin dilden bağımsız veya bağımlı bir takım önişleme yöntemleri uygulanması aşamasıdır (Uysal ve Gunal, 2014).

Metin madenciliğinde metin içeriklerinin sınıflandırıcıların kullanacağı sayısal verilere dönüştürme işi öznitelik çıkarımı aşamasında yapılmaktadır. Bu sürecin sonucunda özgün öznitelikler çıkarılarak dokümanlar sayısal değerlerle gösterilen öznitelik vektörleri ile temsil edilirler.

Metinsel dokümanların sınıflandırılmasında, öznitelik çıkarımından elde edilen vektörlerinin boyutları çok yüksektir. Bu nedenle sınıflandırma başarısını artırmak için çeşitli öznitelik seçim metotları ile eğitim süreci için önemli olanlar seçilerek vektör boyutu azaltılır (Schneider, 2005; Uysal ve Gunal, 2012; Agnihotri et al., 2017).

Doküman sınıflanmada özniteliklerin dokümanla olan ilişkisi Kelime Çantası modeli ile temsil edilir (Aggarwal and Zhai, 2012). Modelde ilişkileri gösteren sayısal değerlere ağırlık, bu değerlerin hesaplanıp atanması işlemine de terim ağırlıklandırma denir. Ağırlık değerinin atanması metinsel dokümanların sınıflandırılmasında performansı artırdığı için birçok araştırmacı bu alanda çalışma yapmakta olup, literatürde bu konu ile ilgili önerilen farklı yöntemler bulunmaktadır (Sparck, 2004; Lan et al., 2009; Ren and Sohrab, 2013; Chen et al., 2016; Dogan ve Uysal, 2019).

Metinsel doküman sınıflandırma için kullanılan algoritmaların ile farklı kelime yöntemleri birlikte kullanılabilmektedir. Bu temsiller arasında TF-IDF, Doc2Vec ve LDA vb. yer alır. (Kim et al., 2019) doküman sınıflandırmasında bu kelime temsil yöntemlerini ele almışlardır.

Kullanıcıları sınıflandırmak için TF-IDF kullanılarak sınıflandırma modellerinin eğitildiği bir çalışmada %90 başarı elde edilmiştir (Hong and Davison, 2010).

Tilve ve arkadaşları, 5 sınıflı iki farklı veri setinde metin sınıflandırması için üç sınıflandırma algoritması kullanmış olup, Naïve Bayes'in diğer algoritmalara göre daha iyi olduğu görülmüştür (Tilve and Jain, 2017). Aynı şekilde (Gogoi and Sarma,

2015) belge sınıflandırmasında Naive bayes tekniğini kullanmanın performansını vurgulamıştır.

Tan ve arkadaşları, CNN ve RNN ile birlikte cevap yerleştirmeleri oluşturmak için soru bağlamından etkilenen basit tek yönlü bir dikkat mekanizmasını geliştirdi (Tan et al., 2016). Bu dikkat mekanizması, iç içe geçmiş sorular ve yanıtlar arasındaki karmaşık anlamsal ilişkileri yakalamaktadır. BERT modeli (Devlin et al., 2019) gibi transformatör tabanlı modeller, ardışık bilgileri dikkate almadan hesaplamayı paralel hale getirebilmekte ve büyük boyutlu veri kümeleri için uygun bir şekilde çalışabilmektedir.

Reuters, AAPD ve IMDB veri setleri kullanılarak yapılan belge sınıflandırması çalışmasında ince ayarlamalar ile BERT modeli kullanılmıştır. Ayrıca tek katmanlı bir sistemin etkinliğini artırmak için LSTM modeli ile BERTlarge'a göre 40 kat daha hızlı çıkarım yapabilmişlerdir (Adhikari et al., 2019).

Wu ve arkadaşları, Wikipedia sayfalarında semantik bir yaklaşımla doküman sınıflandırması yapmışlardır (Wu et al., 2017). Zhang ve arkadaşları sınıflandırma yaklaşımlarına ilişkin bir araştırma makalesi sunmuştur (Zhang et al., 2017). Ayrıca yapılan bir çok çalışmada, Random Forest (Pranckevičius and Marcinkevičius, 2017; Kumar and Kaur, 2020) ve SVM (Sotiropoulos et al., 2017; Isa et al., 2008) yöntemleri kullanılarak yapılan doküman sınıflandırmanın popüler olduğu ve başarılı sonuçlar alındığı görülmüştür.

Öğrenme sürecini ve başarısını artıran birçok teknik bulunmakta olup özellikle Transfer Learning ve Bidirectional Encoder Representations from Transformers (BERT) teknikleri son zamanlarda sınıflandırma dahil birçok farklı alanda kullanılmaya başlamıştır. Bunun yanı sıra ontoloji tabanlı semantik anlamlandırma tekniği de başta duygu analizi olmak üzere doküman sınıflandırma alanında da literatürde yerini almıştır. Bilgisayar bilimlerine ait bir terim olan ontoloji, ajan veya ajan toplulukları özelinde var olabilecek kavram ve ilişkilerin özel olarak tanımlanmasıdır. Ontoloji içerisinde bulunan her terimin tanımlanmış olması ontolojinin birinci özelliği olup, bu terimlerin sonlu sayıda olma zorunluluğu bulunmaktadır. Terimlerin ilişkili anlamlar içermesi durumu da ontolojinin ikinci önemli özelliğidir. Üçüncü özellik ise terimlerin sistematik olmasıdır. Kısacası bir sınıfın, alt sınıfına ait örnekler kendisi için de birer örnek olmalıdır. XML dilinde,

veriler sadece sınıflandırılabilmekte ve sınıflandırılmış olan veriler bilgisayarlar için bir anlam ifade etmezken insanlar tarafından anlamlı olmaktadır. Sınıflandırmanın zamanla gelişmesi ile birlikte makinelerin yargılama ve çıkarım yapma yetenekleri geliştirilmiştir. Bu sayede anlamsal web dilleri ortaya çıkmıştır.

Güvenlik gereksinimlerini otomatik olarak belirlemek için yapılan bir çalışmada yazarlar ontoloji tabanlı bir yaklaşım önermişlerdir (Li and Chen, 2020). İlk olarak, kavramsal katmandaki güvenlik gereksinimleri ontolojisini dilsel katmandaki metinsel güvenlik gereksinimlerinin sözdizimsel ve sözlüksel özellikleriyle ilişkilendirecek kapsamlı ve ankete dayalı genişletilmiş bir güvenlik ontolojisi tanımlanmıştır. Daha sonra tipik makine öğrenimi algoritmalarını kullanarak güvenlik gereksinimleri sınıflandırıcılarını eğitmek için kullanılan bir dizi dilsel kural ve güvenlik anahtar sözcüğü tanımlanmıştır. Yapılan deneysel çalışmalar, önerilen yaklaşımı farklı uygulama alanlarındaki gereksinimleri de sınıflandırabilmek için genelleştirilebilir olduğu görülmüştür. Bu açıdan önerilen yöntemin, mevcut yaklaşımlara göre daha performanslı çalıştığı görülmüştür.

Yapılan bir çalışmada eğitim dokümanlarında yer alan Birleşik Tıbbi Dil Sistemi kavramlarını kullanarak oluşturulan alana özgü benzerlik matrisi yardımıyla her bir tıbbi metin dokümanı için otomatik olarak kavram grafiği oluşturup zenginleştirmiştir (Shanavas et al., 2020). Önerilen yöntemde tıbbi metin belgeleri, bir grafik çekirdeği kullanılarak zenginleştirilmiş kavram grafiklerine göre karşılaştırılır. Daha sonra karşılaştırma sonucuna göre sınıflandırma yapılmaktadır. Bu yaklaşımın faydası, alan bilgisinin sınıflandırmaya dahil edilmesine izin vermesidir. Önerilen yöntem, Birleşik Tıbbi Dil Sistemi semantik ağından gelen bilgiyi kolayca metin sınıflandırmaya dahil etmektedir. Sonuç itibariyle zenginleştirilmiş grafik tabanlı benzerlik ölçüsü, tıbbi belge sınıflandırması için yaygın olarak kullanılan diğer benzerlik ölçülerine göre daha iyi sonuç vermiştir.

Rajbabu ve arkadaşları metinsel dokümanlar üzerinde bilgi keşfi için iki aşamalı bir sınıflandırma yaklaşımı önermiştir (Rajbabu et al., 2018). Yaklaşımda, ilk aşamada cümle sınıflandırması, ardından kelime sınıflandırması gerçekleştirilmektedir. Ayrıca çok alanlı katmanlı bir endüstriyel ontoloji kullanılmıştır. Deneysel çalışmalarda, karar ağaçlarının büyük özellik kümelerini işlemede Naïve Bayes, Destek Vektör Makineleri ve Lojistik regresyon sınıflandırıcılarından daha iyi performans gösterdiği görülmüştür.

Ali ve arkadaşları yapmış oldukları çalışmada duygu sınıflandırması için bir Ontoloji ve Latent Dirichlet Allocation (OLDA) tabanlı konu modelleme ve kelime gömme yaklaşımı önermiştir (Ali et al., 2019). Önerilen sistem, ulaşım içeriğini sosyal ağlardan alıp, anlamlı bilgiler elde etmek için alakasız içeriği kaldırır ve OLDA kullanarak çıkarılan verilerden konular ve özellikler üretir. Ayrıca, kelime gömme tekniklerini kullanarak belgeleri temsil eder ve ardından kelime gömme modelinin doğruluğunu artırmak için sözlük tabanlı yaklaşımlar kullanır. Önerilen ontoloji ve akıllı model sırasıyla Web Ontology Language ve Java kullanılarak geliştirilmiştir. Yöntem, önerilen yaklaşımın duygu sınıflandırması için %93 doğruluk ile etkili olduğunu göstermektedir.

Ontoloji yapısı kullanılarak sınıflandırma yapılan bir çalışmada soru cevap kullanım senaryosunu ele alınmıştır. Soru sınıflandırması, soru cevaplama sistemlerinin düşük performansının ana faktörlerinden biri olarak gösterilmekte ve soru sınıflandırma modülü tasarımı önemli hale gelmektedir. Ray ve arkadaşları çalışmada soru sınıflandırması için WordNet ve Wikipedia'nın kullanımına dayalı yeni bir yöntem önermişlerdir (Ray et al., 2010). Önerilen yaklaşım TREC veri kümeleri üzerinde uygulanarak %89,55 sınıflandırma doğruluğu elde edilmiştir.

Abdollahi ve arkadaşları, hasta taburcu dokümanlarını kullanılarak sınıflandırma üzerinde araştırma yapmışlardır (Abdollahi et al., 2020). Tıbbi taburcu notları gerçek hastalardan toplandığından, genellikle dengesizdir. Ayrıca, bu veri setleri nadir hastalıklarla ilgili de çok az sayıda veri barındırır. Bu durumlar sınıflandırma performansını düşürür. Çalışma kapsamında azınlık sınıfı üzerinde örnekleme yaparak sorunları ele almak için yeni bir olasılıksal sözlük tabanlı veri artırma yaklaşımı önerilmiştir. Bu yöntem, WordNet'ten çıkarılan eşanlamlıları kullanarak, eşanlamlıların orijinal sözcükle benzerliklerinin farkında olarak, yüksek çeşitlilikte yeni belgeler oluşturarak çalışmaktadır. Çalışmada CNN, RNN ve HAN sınıflandırma yöntemleri kullanılmış olup, önerilen yaklaşımın dengesiz veri kümesinde iyi sonuç verdiği görülmüştür.

Ledmi ve arkadaşları, XML belgelerinin otomatik olarak önceden tanımlanmış kategoriler halinde sınıflandırılması konusunda belgelerin içeriğini ve yapısını birleştirerek sınıflandıran bir model önermişlerdir (Ledmi et al., 2021). Ayrıca, terimler arasındaki benzerliğe ilişkin bir hesaplama kullanarak anlamsal komşuluk kavramını modellemek özellikle WordNet ontolojiyi kullanmayı önermişlerdir.

Yapılan deneysel çalışmalarda XML belgelerinin sınıflandırılmasında semantik indekslemenin başarıyı artırdığı görülmüştür.

Irfani ve arkadaşları, ilk belgeye terim eklemek için ek öznitelik genişletme ile Naive Bayes yöntemini kullanarak sınıflandırma yapmıştır (Irfani et al., 2018). Terim eklenmesi, sınıflandırmada zorluk yaşanan kısa metinlerden oluşan haber tweetlerinin sınıflandırma sürecini optimize etmektedir. Yapılan eklemeler, WordNet'ten çıkarılan orijinal belgelerden gelen alt ve üst adlardır. Yapılan deneylerde elde edilen doğruluk değeri öznitelik genişletmesiz sınıflandırma için %72, eş anlamlı ve alt kavram ekleme için %65,75, eş anlamlı ekleme için %66,3 ve alt kavram ekleme için %67,5 olarak elde edilmiştir.

Gawade ve arkadaşları, veritabanındaki toplu belgeleri konu etiketine göre sınıflandırmak için otomatik bir metin sınıflandırma sistemi önermiştir (Gawade et al., 2018). Önerilen sınıflandırma metodolojisi, kelimelerin tekrarından ve aynı anlama sahip kelimelerin ortaya çıkmasından kaçınarak boyutluluk problemini azaltmak için öznitelik çıkarmada semantik işlemeyi kullanmıştır. Her kelimenin uygun bağlamda kendi anlamı ile değiştirilmesi ve kelimelerin belirsizliğini ortadan kaldırmak için WordNet ontolojisi ile kelime gömme algoritması kullanılarak Konvolüsyonel Sinir Ağı algoritması ile de sınıflandırma yapmışlardır.

Çok sınıflı veri setleri üzerinde yapılan ilk çalışmalar ağırlıklı olarak TF-IDF vektörlerini kullanırken (Xiao et al., 2018), yeni sistemlerde ise çoğunlukla gömme tabanlı özellikler benimsenmektedir (Jain et al., 2019; Tagami, 2017).

Çok etiketli belge sınıflandırması, haber makalesi konu etiketleme, duygu analizi, tıbbi kod sınıflandırması vb. gibi çeşitli pratik sorunlara geniş bir uygulama alanına sahiptir. Ayrıca, çok etiketli sınıflandırmaya ilişkin mevcut yöntemler tipik olarak çoğunluk sınıflarına odaklanır ve bu da yeterli eğitim örneğine sahip olmayan diğer önemli sınıflar için tatmin edici olmayan bir performansla sonuçlanır. Song ve arkadaşları, çalışmada veri seti olarak 50.000 yasal görüş veri kullanarak alana özgü ön eğitimi benimseyen Etiket Katılımlı Çok Görevli Çok Etiketli Sınıflandırma mimarisi önermişlerdir (Song et al., 2022). Önerilen mimari hem POSTURE50K hem de EUROLEX57K veri kümeleri üzerinde uygulanmış olup, APLC_XLNet ve X-Transformer yöntemlerine göre daha iyi sonuç verdiği görülmüştür.

Doküman sınıflandırmaya yönelik yapılan bir diğer çalışmada yazarlar tahmine dayalı kodlamada farklı derin öğrenme teknolojilerini kullanarak yaptıkları deneysel çalışmada üç açık kaynaklı yasal belge inceleme veri setinde CNN modelinin diğer modellerden daha iyi performans gösterdiğini tespit etmişlerdir (Han and Snaidauf, 2021).

Veri kümesine ait sınıflardaki verilerin dengesiz olduğu durumlarda sınıflandırma zorlaşmaktadır. Özellikle de küçük veri setlerinde bu durum daha da zor olmaktadır. Araştırmacılar da özellikle bu konuya yoğunlaşmakta ve sınıflandırma başarısını artırmak amacıyla çeşitli yöntemler denemektedirler. Sun ve arkadaşları dengesiz metin veri akışları için bir topluluk sınıflandırma algoritmasını önermişlerdir (Sun et al., 2020). İlk olarak, dengeli veri altkümeleri oluşturmak için geliştirilmiş bir yeniden örnekleme yöntemi kullandıktan sonra konu modeli, belge-konu eğitimi alt kümeleri oluşturmak üzere dengeli veri alt kümelerinde konu modellemesi gerçekleştirmiştir. Deneysel sonuçlar, önerilen algoritmanın yalnızca pozitif örnekler için değil, tüm örnekler için iyi bir sınıflandırma performansına sahip olduğunu göstermektedir.

Valarmathi ve arkadaşları dengesiz banka pazarlama veri setini kullanarak, Naive Bayes, J48, KNN ve Bayesnet gibi farklı sınıflandırma algoritmaları üzerinde deneysel çalışma yapmışlardır. CfsSubsetEval yöntemi, kestirim yeteneğine dayalı olarak özniteliklerin alt kümesini değerlendirerek ve seçilen öznitelikler arasındaki tekrarı bularak boyut indirgeme yapılmıştır. Boyut azaltmadan önce, J48 %89 başarı sağlarken, boyut indirgemeden sonra modelin başarısı J48 algoritması ile %91,2'ye yükselmiştir (Valarmathi et al., 2019).

Gün geçtikçe ortaya çıkan farklı zararlı yazılımlar ciddi tehdit oluşturmaktadır. Yeni çıkan zararlı yazılımlar bir küme de çok fazla yer almadığından dolayı da otomatik olarak sınıflandırmakta zordur. Ding ve arkadaşları, yaptıkları çalışmada dengesiz veri seti kullanarak çoklu sınıflandırma için bir model önermişlerdir. Önerilen derin öğrenme tabanlı yöntem ile dengesiz veri seti üzerinde %98,48 doğruluk oranına ulaşılmıştır (Ding et al., 2020).

Metin sınıflandırma performansının iyileştirilmesi için Wikipedia ve WordNet gibi bilgi tabanlarındaki bilgiler uygulanabilmektedir. Bloehdorn ve arkadaşları, Anlamsal düzeltmeler yapmak amacıyla WordNet'ten alınan önsel bilgileri metin

sınıflandırmasına dahil etmişlerdir (Bloehdorn et al., 2006). Kelimeler arasındaki anlamsal benzerliği birleştirmek için, WordNet'ten türetilen bir yumuşatma matrisi kullanmışlardır. Yumuşatma, anlamsal tutarlılığı iyileştirmek için TF-IDF özellik vektörlerine uygulanmıştır. Bu da anlamsal olarak ilişkili terimlerin öznitelik değerlerinin artmasını sağlamıştır. Cristianini ve arkadaşları ise çalışmalarında metin sınıflandırması için bir anlamsal düzeltme çekirdeği tasarlamak amacıyla WordNet kullanmışlardır. Paylaşılan üst kavramlara dayalı olarak kelimeler arasındaki benzerlik hesaplanmıştır (Cristianini et al., 2002).

1.578.627 tweet kullanılarak yapılan başka bir çalışmada ise duygu sınıflandırması yapılmıştır. Sınıflandırma başarısı BoW ve Semantic BoW kullanılarak değerlendirilmiştir. Burada kelimeler arasındaki benzerlik hesabı WordNet kullanılarak yapılmış olup, benzer kelimeler bire indirilerek nitelik boyutu küçültülmüştür. AdaBoost Sınıflandırma ve KNeighbors Sınıflandırma ile Semantic BoW birlikte kullanıldığında sınıflandırma başarısının klasik BoW yöntemine göre %1-4 arasında arttığı gözlemlenmiştir. Ancak Semantic BoW yönteminde doğruluk %69 olduğu için daha fazla başarıya ihtiyaç duyulmaktadır (Bamatraf and Bin-Thalab, 2021).

Sınıflandırma ile ilgili yapılan çalışmalar incelendiğinde ontoloji temelli boyut indirgeme yönteminin sınıflandırma başarısı için önemli olduğu görülmüştür. Gawade ve arkadaşları, WordNet ile kelime tekrarından kaçınılarak öznitelik boyutu küçültme ile ilgili yaptığı çalışmada CNN ile sınıflandıran bir yapı önermişlerdir (Gawade et al., 2018). Bu konuda yapılan başka bir çalışmada ise Naive Bayes, Jrip, J48 ve SVM sınıflandırma yöntemleri PCA ve ontoloji tabanlı özellik azaltma yöntemi ile karşılaştırılmıştır olup, ontoloji tabanlı indirgemenin PCA'ya göre daha iyi sonuçlar verdiği görülmüştür. Çalışmada Reuters-21578 veri setine ait 15 dengesiz kategori kullanılmıştır ve en yüksek başarı %85 ile SVM'de elde edilmiştir (Elhadad et al., 2017).

III. YÖNTEM

Bu bölümde çalışmada kullanılan veri ön işleme yöntemleri, WordNet ontolojisi, kelime temsil yöntemleri ve sınıflandırma yöntemleri anlatılmıştır.

A. Veri Ön İşleme

Veri ön işleme, (García et al., 2015) bilgi keşfi sürecindeki temel faaliyetlerden biridir. Deneysel çalışmalarda kullanılacak veriler genellikle tutarsızlıklar, eksik ve gürültülü değerler gibi bir çok sorunla birlikte gelir. Eğer veriler bu seviyede yani sorunlu bir şekilde deneysel çalışmalarda kullanılırsa öğrenme algoritmalarının performansı zayıflayacaktır. Bu durumu tersine çevirmek için uygun ön işleme adımları uygulanarak otomatik keşiflerin, kararların kalitesi ve güvenilirliği önemli ölçüde etkilenmektedir (Ramírez-Gallego et al., 2017).

Veri ön işleme adımında uygulanan teknikler 4 aşamadan oluşmakta olup bunlar; veri temizleme, birleştirme, dönüştürme ve indirgeme işlemlerdir.

Bu tez çalışmasında veri ön işleme aşamasında uygulanan tekniklerle ilgili detaylar aşağıda açıklanmıştır.

1. Veri Temizleme

Bilim insanları makine öğrenimi görevlerinde daha başarılı sonuçlar alabilmek için veri kümesinin üzerinde bir çok ön işleme yöntemi uygulamaktadır (Shevlyakov and Kan, 2020).

Veri temizleme, kayıp, gürültü, yanlış girilmiş ve aykırı verilerin ortadan kaldırılması işlerini içerir (Silahtaroğlu, 2016)

Ham veri setlerinde en önemli sorunlardan birisi bazı değerlerin boş olmasıdır. Böyle bir durumda kayıtların bazıları veya tamamı veri setinden çıkarılabilir. Bir diğer yöntem de ilgili özelliğin sahip olduğu değerlerin ortalaması alınarak tamamlama yapılmasıdır (Han, 2012).

2. Lemmatizasyon

Lemmatizasyon, metin madenciliğinde uygulanan önemli ön işleme adımlarından biridir. Ayrıca DDİ ve dilbilimi ile ilgili yapılan çalışmalarda sıklıkla kullanılır. Bunun yanı sıra, arama motorları için genel anahtar kelimeler veya kavram haritaları için etiketler oluşturmada verimli bir yol sağlar.

Lemmatizasyon, bir kelimenin çekimli kısımlarını, kelimenin lemması veya kelime dağarcığı olarak adlandırılan tek bir öğe olarak tanınabilecek şekilde bir araya getirme işlemidir (Balakrishnan, 2014). Bir kelimenin bu kök tabanlı sözlük formuna ise lemma adı verilir (Sharma et al., 2022). Bu süreç kök çıkarma ile aynıdır ancak belirli kelimelere anlam katar. Bir örnekle ifade etmek gerekirse, İngilizce kelimeler olan "computes", "computing", "computed" kelimelerinin kökü comput olur. Ancak normalleştirilmiş biçimi fiilin mastarı olan "compute" dur (Plisson et al., 2004).

3. Durak Kelimeleri Kaldırma

Metin içerisinde tek başına anlamı olmayan ve sık geçen kelimeler (fakat, ve, için, gibi, ancak, ile, ben, sen, o) durak kelime (stop words) olarak ifade edilmektedir (Haltaş vd., 2015).

Durak kelimeler ile ilgili standart bir liste bulunmamakla birlikte her dilin kendine özgü durak kelimeleri bulunmaktadır (Çoban et al., 2015). Veri ön işleme aşamasında ilgili dile özgü oluşturulan durak kelime listesi baz alınarak veri kümesi içerisinden bu gereksiz olan kelimeler çıkarılır (Yıldız, 2016).

4. Yazım Düzeltme

Yazım düzeltme sorunları iki kategoriye ayrılır. İlk kategori, amaçlanmamış olsa da geçerli sözcüklerle sonuçlanan yazım düzeltme sorununu (örneğin, "in a peace of cake", "piece" olarak yazılmalı) ve ayrıca belirli sözcük kullanım hatalarını ("among" ve "between" gibi) düzeltme sorununu ele alır. İkinci kategori ise bir sözlükte bulunamayan sözcüklerle sonuçlanan hatalarla ilgilidir (Ruch et al., 2003).

B. WordNet Ontolojisi

WordNet, büyük bir sözlüksel İngilizce veritabanıdır. Psikoloji profesörü George A. Miller'ın yönetiminde Princeton Üniversitesi Bilişsel Bilim Laboratuvarı'nda oluşturulmuş ve sürdürülmektedir (Taieb et al., 2014). İsimler, fiiller,

sıfatlar ve zarfların, her biri farklı bir kavramı ifade eden eşanlamlılar kümeleri halinde gruplandırılır. Eş anlamlılar, kavramsal-anlamsal ve sözlüksel ilişkiler aracılığıyla birbirine bağlanır. Ortaya çıkan anlamlı bir şekilde ilişkili kelime ve kavramlar ağında tarayıcı ile gezinilebilir. WordNet ayrıca ücretsiz olup halka açıktır. Yapı itibariyle de WordNet, DDİ ve hesaplamalı dilbilim alanları için de faydalı bir araç olmuştur (Farkiya et al., 2015).

WordNet'te, eşanlamlı sözcükler bilişsel olarak eşdeğer durumda oldukları için gruplar halinde, "eşküme" (synset) olarak adlandırılmıştır ve her bir eşküme benzersiz bir numara ile diğerlerinden ayrılmaktadır. WordNet ayrıca, bu eşkümelere ilişkin kısa ve genel tanımların yanı sıra, eşkümeler ve sözcükler arasındaki çeşitli ilişkileri de içinde barındırır.

WordNet'te, eşkümeler arasındaki ilişkiler anlam temelli olarak kurulmuştur. Her eşküme, bir kavramı temsil eder ve eşkümenin üyeleri, aynı anlamı taşıyan sözcüklerden oluşur. Her eşküme için yalnızca bir tanımsal ifade mevcuttur. WordNet'te, her eşkümeyle birlikte tanımsal ifadelerin yanı sıra örnek kullanımları da kodlanmıştır. Çizelge 1'de örnek WordNet eşkümesi ve parçaları gösterilmektedir.

Çizelge 1 WordNet Eşkümesinin Parçalarına Ait Örnek

Eşanlamlı sözcükler	Tanımsal ifade	Örnek kullanım
car, auto, automobile, machine, motorcar	A motor vehicle with four wheels, usually propelled by an internal combustion engine.	He needs a car to get to work.

WordNet'te, eşkümeler arasında çift yönlü anlamsal ilişkiler olan üst/alt anlam (hypernymy-hyponymy), parça-bütün (holonymy-meronymy) ve bazı gerektirim (entailment) ilişkileri gibi yapısal bir ağ düzenlemesi bulunmaktadır.

Üstanlam / Altanlam İlişkisi: IS-A ilişkisi, WordNet içindeki kavram hiyerarşisinin temelini oluşturur. Eğer bir X eşkümesinin her bir öğesi aynı zamanda bir Y eşkümesinin de içine giriyorsa, Y eşkümesi X eşkümesinin üst anlamlı (hypernymic) bir kümesini temsil eder.

Parça-bütün İlişkisi: Bir kavram, X eşkümesi ile ifade edildiğinde, bu kavram Y eşkümesinin bütünlüğü içinde bir parça olarak yer alıyorsa, X eşkümesi Y eşkümesinin bir parçasıdır.

Gerektirim İlişkisi: Eğer bir X eyleminin gerçekleşmesi, bir Y eyleminin gerçekleştirilmiş olmasına bağlı ise, X eşkümesinin karşılık geldiği kavramın Y eşkümesini gerektirdiği ifade edilir.

WordNet'te, sözcük formları arasında üst ve alt anlam ilişkileri oluşturulamaz. Üstanlamlılık, WordNet'te sözcükler arasında belirli anlamlar yani eşkümeler arasında kurulan bir ilişkidir. Bu ilişki, geçişli ve asimetrik bir anlamsal ilişkidir ve genellikle IS-A ya da IS-A-KIND-OF ilişkisi olarak kabul edilir. Bu yapı, hiyerarşik bir yapı oluşturur, alt seviyelerdeki özelleşmiş eşkümelerden yukarıdaki daha genel anlamlı eşkümelere doğru uzanır. WordNet'teki {cat, true cat} eşkümesinden başlayan ve tüm isim eşkümeleri için en üst düzeyde bulunan {entity} eşkümesine kadar uzanan bir üstanlam zinciri, Şekil 1'de gösterilmiştir.

{cat, true cat}
=> {feline, felid}
=> {carnivore}
=> {placental, placental mammal, eutherian, eutherian mammal}
=> {mammal, mammalian}
=> {vertebrate, craniate}
=> {chordate}
=> {animal, animate being, beast, brute, creature, fauna}
=> {organism, being}
=> {living thing, animate thing}
=> {object, physical object}
=> {physical entity}
=> {entity}

Şekil 1 Örnek Üstanlam Zinciri (Güner, 2015)

Üstanlamlılık ilişkisinin bu yapılandırılmasıyla sözcüksel bir hiyerarşi oluşur ve genellikle bu tür hiyerarşiler, önceki örneklerde olduğu gibi ağaç yapılarıyla temsil edilir. Bu hiyerarşi aynı zamanda taksonomik bir yapıya sahiptir.

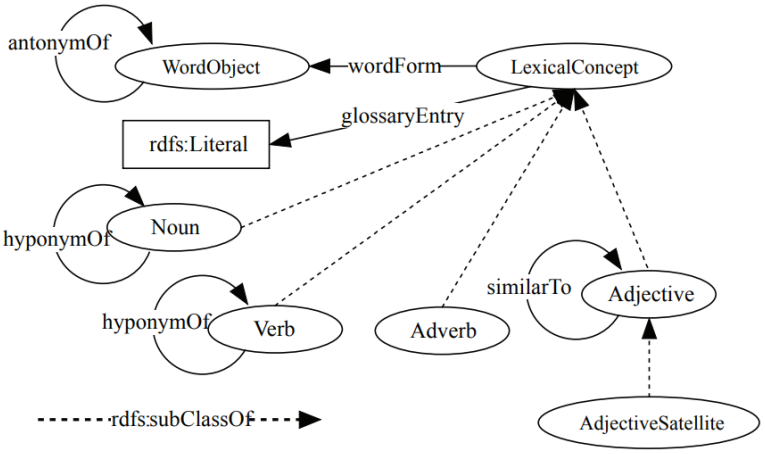
Ancak, WordNet içinde bazı eşkümelerin hiyerarşideki yerleşimleri, taksonomik bir ağaç yapısının gerektirdiği kısıtlamaları ihlal etmektedir. Taksonomik bir ağaç yapısında, bir düğüm yalnızca bir üst düğümle bağlantı kurabilirken birden fazla alt düğümle bağlantı kurabilme kısıtlamasına tabidir. Kavramsal bir taksonomik yapı düşünüldüğünde, bu durum bir kavramın birden fazla da alt kavramı olabileceği ama sadece bir üst kavramı olabileceği ve birden fazla da alt kavrama sahip olabileceği anlamına gelir.

WordNet'te eşküme hiyerarşisi oluşturulurken, bazı eşkümelerin birden fazla eşkümenin altanlamlı eşkümesi olması gerektiği göz önünde bulundurulmuştur. Bu nedenle, WordNet eşküme hiyerarşisi tam olarak taksonomik bir ağaç yapısı şeklinde değildir.

WordNet, temel yapıtaşları olan eşkümelerle bir kavrama karşılık gelen tüm sözcükleri içerdiği için bir eşanlamlılar sözlüğüne (thesaurus) benzer. Bununla birlikte, WordNet'i diğer eşanlamlılar sözlüklerinden ayıran bazı farklar vardır. Öncelikle, WordNet sadece sözcük formlarını değil, aynı zamanda sözcüklerin belirli anlamlarını birbirine bağlar. WordNet'in eşanlamlılar sözlüğünden farklı bir yönü daha vardır: eşanlamlılar sözlüğündeki sözcük grupları, anlamsal yakınlığın ötesinde belirli bir yapıya sahip değilken, WordNet'te sözcükler arasındaki anlamsal ilişkiler açık bir şekilde etiketlerle belirtilir (Güner, 2015).

WordNet 3.0 sürümünün veritabanında 155287 sözcük, 117659 eşküme altında sınıflandırılmıştır. Eşkümeler ad, ön ad, eylem ve belirteçler olmak üzere dört farklı sözdizimsel kategoriyle belirtilmiştir.

Şekil 2’de WordNet sözcük ontolojisinin yapısı gösterilmektedir.



Şekil 2 WordNet Ontoloji Yapısı (Chebotko et al., 2008)

C. Kelime Temsil Yöntemleri

DDİ'de kelime temsil yöntemleri sınıflandırma başarısı için oldukça önemlidir. Başarının artmasındaki en önemli etki metinlerin anlamsal ilişkileri ve bağlılıkları korunarak temsil edilmesidir. Karakter tabanlı temsilde, metindeki her karaktere benzersiz bir indeks değeri atanmaktadır. N boyutlu vektörde karaktere karşılık gelen indeksler 1 diğerleri de 0 olacak şekilde metinden seyrek ve yüksek boyutlu vektörler elde edilmektedir.

Algoritmalar sadece sayısal girdiler üzerinde hesaplama yapabilmektedirler. Bu nedenle metinsel verilerin algoritmalar tarafından işlenmesi metin sayısallaştırma aşamasını gerektirmektedir. Veriler algoritmaya girdi olarak verilmeden önce girdi formatına dönüştürülmelidir. Bu verilerin sayısal olarak temsil edilmesinde genellikle kelime veya karakterler temel alınmaktadır (Chollet, 2017).

Vektör uzayında temsil edilen her kelime bir nokta olarak gösterilmektedir. Cümlelerdeki kelimeler arasındaki anlamsal ilişkinin korunması DDİ'nin önemli araştırma konularından birisidir. Karakter tabanlı temsil (Zhang et al, 2015) ve Word2vec (Mikolov et al, 2013) gibi kelime temsil yöntemleri günümüzde derin öğrenme yöntemleri ile birlikte sıkça kullanılmaktadır.

İki katmanlı bir yapay sinir ağı olan Word2vec kelimeleri vektör olarak ifade etmek için kullanılmaktadır. Bu yöntem girdi olarak metinsel veriyi almakta ve çıkış olarak anlamsal bir vektör kümesi oluşturmaktadır. Bu sayede kelimeler özellik vektörlerine dönüştürülmektedir. Kelime vektörleri üreten gözetimsiz bir öğrenme

algoritması olan Glove ise sondan eklemeli olamayan dillerde yüksek performans sağlamaktadır. Bu yöntem ise bağlamdaki kelimelerin birlikte bulunma istatistiklerine göre kelimelere anlamsal vektörler atamaktadır (Pennington et al, 2014).

1. BoW

Metinlerin özniteliklerini çıkarırken kullanılan ve oldukça yaygın bir yöntem olan Bag of Words dokümandaki kelimelerin frekansından oluşan bir vektör olarak temsil edilir (Zhang et al., 2010). Bu yöntemde metnin hepsi terimlerine ayrılır. Ayrılan terimler de öznitelik olarak ifade edilmektedir. Öznitelik değeri ise terimin tüm metinde geçme sıklığı olarak ifade edilir. Bu şekilde kategorik bir veri sayısal hale dönüştürülmüş olur (Aksu ve Karaman, 2020).

2. TF-IDF

TF-IDF algoritması ilk olarak (Salton and Yu, 1973) tarafından önerilmiş olup, vektör uzayı modelinde en çok tercih edilen öznitelik kelime ağırlığı hesaplama yöntemidir. Temel olarak kelimelerin frekansı ve ters metinlerin frekansı olmak üzere iki bölümden oluşur. Kelime frekansı, dosyada belirli bir kelimenin tekrarlanma sayısını ifade eder. Ters dosya frekansı ise bir kelimenin genel öneminin bir ölçüsünü gösterir. TF-IDF'e ait formüller Denklem 1 ve Denklem 2'deki gibidir.

$$TFIDF(t,d) = TF\ (d,t), IDF(t)$$

Denklem 1

TF ve IDF çarpımı bir metinde çok bulunan fakat diğer metinlerde daha az bulunan bir terimin ağırlığının fazla olduğunu göstermektedir.

$$TFIDF\left(t_k, d_j\right) = \#(t_k, d_j).\log_2 \frac{|T_r|}{|T_r(t_k)|}$$

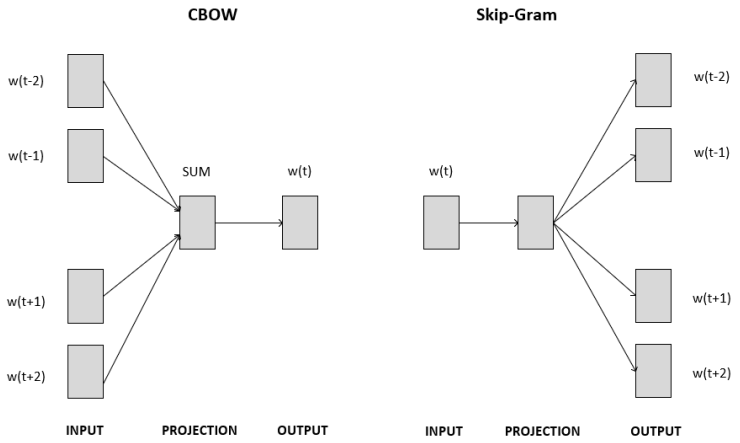
Denklem 2

Burada $\#\left(t_k, d_j\right)$, t_k kelimesinin d_j dokümanı içinde geçme sayısını, |Tr| bütün dokümanları, $|T_r(t_k)|$ içinde en az bir kere t_k kelimesi geçen dokümanları göstermektedir. Metin ağırlıklandırmasında terim dokümanda yoksa dokümanın vektöründe terim sıfır ile ağırlıklandırılmaktadır.

3. Word2Vec

Word2vec, bir kelime gömme yaklaşımı olup, Mikolov tarafından 2013 yılında Google'da bir yöntem olarak önerilmiştir (Kang et al., 2018).

Word2vec yönteminin 2 farklı öğrenme modeli bulunmaktadır. Bunlar, CBOW (Continuous Bag-of-Words Model) ve Skip-gram (Continuous Skipgram Mode) modelleridir (Mikolov et al., 2013). Bu modellere ait mimari Şekil 3'te gösterilmiştir.



Şekil 3 Word2Vec Modelleri: CBOW ve Skip-Gram

CBOW, bağlamda verilen bir kelimenin olasılığını tahmin etmek için kullanılır. Bağlam, kelime grubu veya tek bir kelimeden oluşabilir. Bir cümledeki tüm komşu kelimeler çıkış vektörünü tahmin etmek için girdi olarak alınır.

CBOW temel olarak 3 katmandan oluşur:

Giriş katmanı eğitim setinin bilgi olarak alındığı katmandır. Gizli katman, en önemli katman olup, bu katmanda sadece beklenti gerçekleşir. Çıkış katmanında ise önemli vektör elde edilir.

Skip gram modeli temel olarak CBOW modelinde olanın tersi olan adımları içerir. Hedef kelime, bağlam kelimesi yardımıyla tahmin edilirken, skip gram modelinde ise verilen hedef kelime ile çevresindeki kelimeyi tahmin eder (Choudhary and Beniwal, 2021).

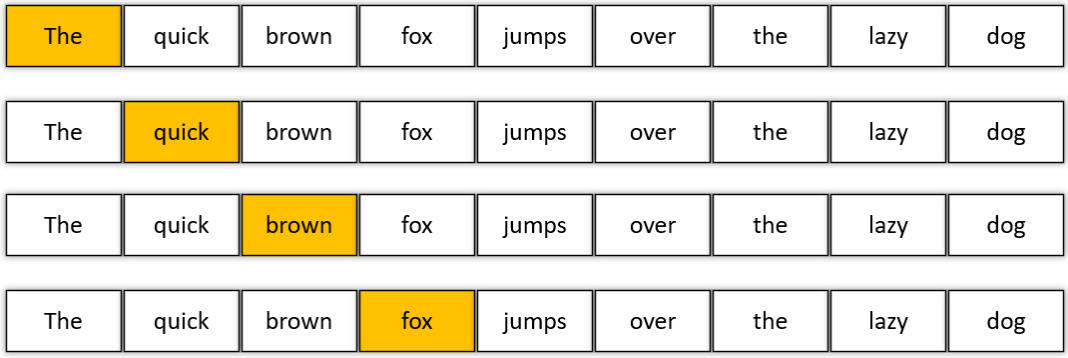
Skip-gram modelinin üstlendiği görev, seçilen bir kelimenin, etrafındaki kelimeler ile arasındaki ilişkiden vektör temsillerini çıkarmasıdır. Formül ile ifade etmek gerekirse, amaçlanan verilen kelimeler w_1, w_2, w_3, …, w_t için logaritmik olasılığı maksimize etmektedir.

Denklem 3'te gösterildiği gibi eğitim içerik boyutu c ile temsil edilmekte olup, modelinin kalitesinin artmasının içeriğe bağlı olduğu hipotezi kurulabilmektedir.

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log_p(w_{t+j} \mid w_t)$$

Denklem 3

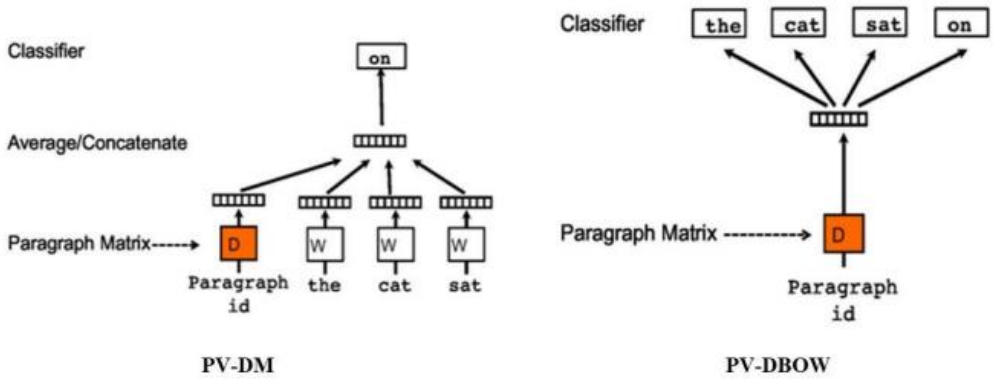
Skip-gram ile CBOW arasındaki ilişki Şekil 4'te gösterilmiş olup, skip-gram modelinde "the" ve "brown" kelimeleri "quick" kelimesi kullanılarak tahmin edilmeye çalışılırken CBOW modelinde ise "the" ve "brown" kelimeleri kullanılarak "quick" kelimesi tahmin edilmektedir.



Şekil 4 Pencere Boyutu Örneği

4. Doc2Vec

Le ve Mikolov tarafından geliştirilen Doc2Vec, hedef kelimeyi tahmin etmek için belgeyi temsil eden bir vektör üretir (Le and Mikolov, 2014). Bunu yaparken, belgenin uzunluğu sayılmaz. Doc2Vec iki farklı yöntemi içerir. Bunlardan birincisi Distributed Memory Model of Paragraph Vectors (PV-DM), ikincisi ise Distributed Bag Of Words of Paragraph Vector (PV-DBOW). Modellerin yapısı Şekil 5'te gösterilmiştir.



Şekil 5 Doc2vec - PV-DM ve PV-DBoW (Le and Mikolov, 2014)

PV-DM yönteminde her paragraf bir kelime olarak kabul edilir ve her paragrafın özel bir kimliği, yani bir vektör gösterimi vardır. PV-DBOW ise hedef kelimeyi tahmin etmek yerine dokümandaki kelimeleri sınıflandırmak amacıyla bir paragraf vektörü kullanır (Dogru vd., 2021).

D. Sınıflandırma Yöntemleri

En sık karşılaşılan karar verme görevlerinden birisi sınıflandırmadır. Bir dizi grup veya sınıfın kümelenmesinden sonra, bir nesnenin, o nesneyle ilgili önceden tanımlanmış bir gruba veya sınıflara atanması gerektiğinde bir sınıflandırma devreye girmektedir. Burada belirli bir test örneğinin bir dizi sınıf arasından hangi sınıfa ait olduğuna karar verilmesi gerekir (Zhang, 2020).

Sınıflandırmadaki hedef fonksiyonların kesikli olması oldukça önemlidir. Genel olarak, sınıf etiketine sayısal veya diğer bazı değerler anlamlı bir şekilde atanamaz. Bu, değeri belirlenmesi gereken sınıf özelliğinin kategorik özellik olduğu anlamına gelir (Novaković et al., 2017).

Bu çalışma kapsamında tüm veriler sınıf etiketine sahip olduğu için denetimli öğrenme uygulanarak 4 farklı sınıflandırma algoritması (Random Forest, Support Vector Machine, Multilayer Perceptron, BERT ve DistilBERT) kullanılmıştır.

1. Random Forest (RF)

(Breiman, 2001) tarafından literatüre kazandırılan RF Algoritması her bir karar ağacını farklı bir gözlem örneği üzerinde eğitip çeşitli modeller üreterek, sınıflandırmayı sağlamaktadır. Hem sınıflandırma hem de regresyon problemlerinde kullanılan algoritma, diğer yöntemlere göre daha hızlı eğitilebilmesi, daha yüksek

tahmin hızı, parametresinin az olması ve çok boyutlu problemlere doğrudan uygulanabilmesi gibi özellikleriyle dikkat çekmektedir (Cutler et al., 2012).

RF algoritmasında, bağımsız olarak örneklenen rastgele bir vektörün değerlerine bağlı olan ve aynı dağılıma sahip ağaçlar bulunmaktadır. Algoritma düğümleri ilgili düğümden rasgele seçilen bir tahmin edici alt kümesi içerisinden en iyisini kullanarak bölme işlemini gerçekleştirir. Yani her düğümü bölmez. RF'in adımları aşağıdaki gibidir (Acet, 2022).

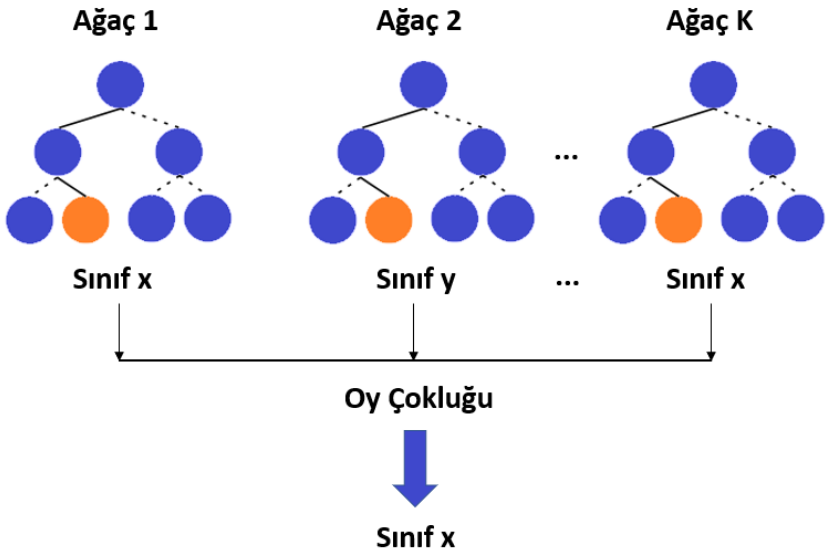
1. Giriş veri setinden rastgele örnek seçimleri yapılır.

2. Algoritma seçilen her örnek için tahmin sonucunu verecek bir karar ağacı oluşturur.

3. Tahmin edilen her bir sonuç için sınıflandırma probleminde mod kullanılır.

4. En çok oy alan tahmin çıkış olacaktır.

RF yönteminin K tane ağaç ile sınıflandırma modeli Şekil 6'da gösterilmiştir.



Şekil 6 RF Sınıflandırma Modeli

Eğitilen k adet karar ağacının RF modeli içerisindeki tanımı Denklem 4'te verilmiştir (Chen et al., 2016). Denklemdeki meta karar ağacı sınıflandırıcı olan $H(X, \theta_j)$ sınıf değerini doğrudan tahmin etmez. Bunun yerine hangi temel seviye sınıflandırıcının tahminde kullanılması gerektiğini belirtir. (Todorovski and Džeroski, 2003)

$$H(X, \theta_j) = \sum_{i=0}^{k} h_i(x, \theta_j), \qquad (j = 1,2,3, \dots, m)$$

Denklem 4

2. Support Vector Machine (SVM)

SVM (Cortes and Vapnik, 1995) güçlü teorik temellere ve mükemmel başarılara sahiptir. Başta sınıflandırma olmak üzere, kümeleme ve tahminleme çalışmalarında kullanılmaktadır. Bu yöntem, farklı örnekler arasındaki maksimum sınırın tespit edilmesi durumuna dayanmakta olup, lineer olmayan örnek uzayını, örneklerin lineer olarak ayrılabileceği yüksek boyuta taşımaktadır (Elmas 2012).

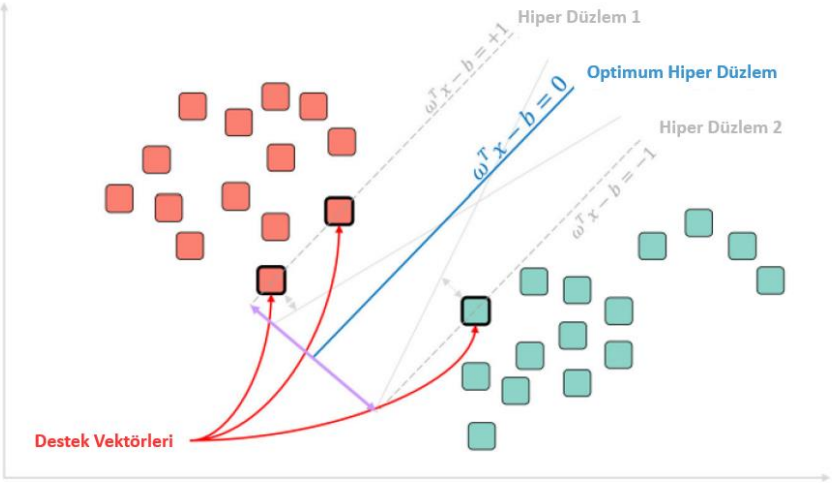
SVM metin kategorizasyonu ve yapay görme alanlarında iyi sonuçlar vermesi ile birlikte dikkate alınmaya başlanmıştır. Bu yöntem bir çok araştırma çalışmasında yapay sinir ağları ve diğer istatistiksel yöntemlere göre çok daha iyi sonuç vermektedir (Elmas 2012).

Parametrik olmayan bir model olan SVM'nin parametrik olmaması, modeldeki tüm parametrelerden yoksun olduğu anlamına gelmemektedir. Burada özellikle "öğrenme" önemli bir konudur. Parametreler, klasik istatiksel çıkarsama modellerinin aksine burada önceden tanımlı olmayıp sayı olarak kullanılan eğitim verilerine bağlıdır. Vapnik, Chervonenkis ve diğer bilim adamları tarafından tanıtılan bu model yapısal risk minimizasyonunun temel paradigmasıdır. İyi bir genelleme özelliğine sahip olması istenen bir model tasarımı yapısal iki temel yaklaşımdan oluşmaktadır. (Özkan, 2008).

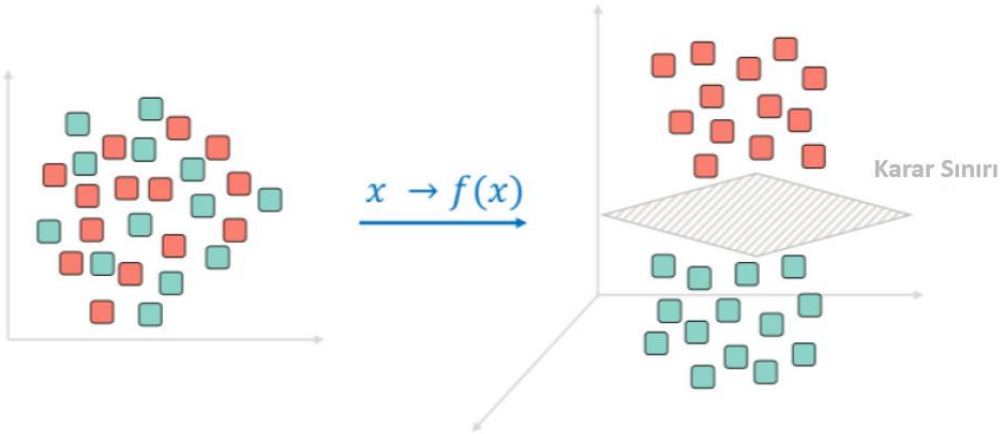
SVM, veriyi birbirinden ayırt edebilmek için kullanılmakta olup, en uygun fonksiyonun tahmin edilmesi temeline dayanmaktadır (Wang, 2005). Eğitme örnekleri arasında ve her iki sınıfın uç noktasında seçilen destek vektörleri sınıflandırmanın temelindeki ana elemanları oluşturmaktadır. Ayrıca SVM'de düzlemin optimum olması genelleme yeteneğinin maksimum olmasına bağlıdır.

SVM, doğrusal ve doğrusal olmayan problemleri çözebilmektedir. Doğrusal SVM, çok boyutlu verilerin girdi uzayında doğrusal olarak ayrılabileceğini varsayar. Verileri düz bir çizgiyle ayıramadığımızda ise Doğrusal Olmayan SVM kullanılır.

Doğrusal olarak ayrılabilen SVM Şekil 7'de, doğrusal olarak ayrılamayan SVM'de Şekil 8'de gösterilmiştir.



Şekil 7 Doğrusal Olarak Ayrılabilen SVM (Sheykhmousa et al., 2020)



Şekil 8 Doğrusal Olarak Ayrılamayan SVM (Sheykhmousa et al., 2020)

İki sınıflı ve doğrusal olarak ayrılabilen bir sınıflandırmada SVM modelinin eğitimi için n sayıda örnekten oluşan eğitim verisinin {xi, yi}, i = 1,.....,n olduğu durumda, optimum hiper düzleme ait eşitsizlikler Denklem 5 ve Denklem 6'daki gibi olur:

$$w \cdot xi + b \geq +1 \text{ her } y = +1 \text{ için}$$

Denklem 5

$$w \cdot xi + b \leq +1 \text{ her } y = -1 \text{ için}$$

Denklem 6

Sonuç olarak, iki sınıfa sahip ve doğrusal olarak ayrılabilen bir problem özelinde karar fonksiyonu Denklem 7'de belirtildiği gibi yazılabilir (Osuna et al.,1997).

$$f(x) = sign\,(\sum_{i=1}^{k} \lambda_i\, y_i(x * x_i) + b)$$

Denklem 7

Sınırın maksimum, yanlış sınıflandırma hatalarının da minimum seviyeye getirilmesi arasındaki denge, pozitif değerler alan ve C ile gösterilen bir düzenleme parametresi ($0 < C < \infty$) tanımlanmasıyla kontrol edilebilir (Cortes and Vapnik, 1995). Düzenleme parametresi ve yapay ile doğrusal olarak ayrımı yapılamayan veriler için optimizasyon problemi Denklem 8'deki şeklini alır.

$$\min[\frac{\left|\left|w\right|\right|^2}{2} + C * \sum_{i=1}^{r} \delta_i$$

Denklem 8

Buna bağlı sınırlamalarda Denklem 9 ve Denklem 10'da gösterilen şekilde ifade edilir.

$$y_i(w * \varphi(x_i) + b) - 1 \geq 1 - \delta_i$$

Denklem 9

$$\delta_i \geq 0 \text{ ve I = 1,.....N}$$

Denklem 10

SVM, matematiksel olarak $K\left(x_i, x_j\right) = \varphi(x) * \varphi(x_j)$ şeklinde ifade edilen bir kernel fonksiyonu yardımıyla doğrusal olmayan dönüşümler yapabilmekte ve verilerin doğrusal olarak yüksek boyutta ayrımına olanak tanımaktadır. Sonuç itibariyle, iki sınıflı ve doğrusal olarak ayrılamayan bir problemin çözümü ile ilgili karar kuralı, kernel fonksiyonu kullanılarak Denklem 11'deki gibi yazılabilir (Osuna et al,1997).

$$f(x) = sign\,(\sum_{i} a_i y_i \varphi(x) * \varphi(x_i) + b$$

Denklem 11

3. Multilayer Perceptron (MLP)

Yapay Sinir Ağlarının en önemli sınıflarından biri MLP'dir. Etiketlenmiş veri örneklerini kullanarak denetimli bir öğrenme gerçekleştiren güçlü bir modelleme

yöntemidir. Bu yöntem, verilen girdilerden çıktıların tahmin edilmesini sağlayan doğrusal olmayan bir fonksiyon modeli oluşturur (Taud and Mas, 2018).

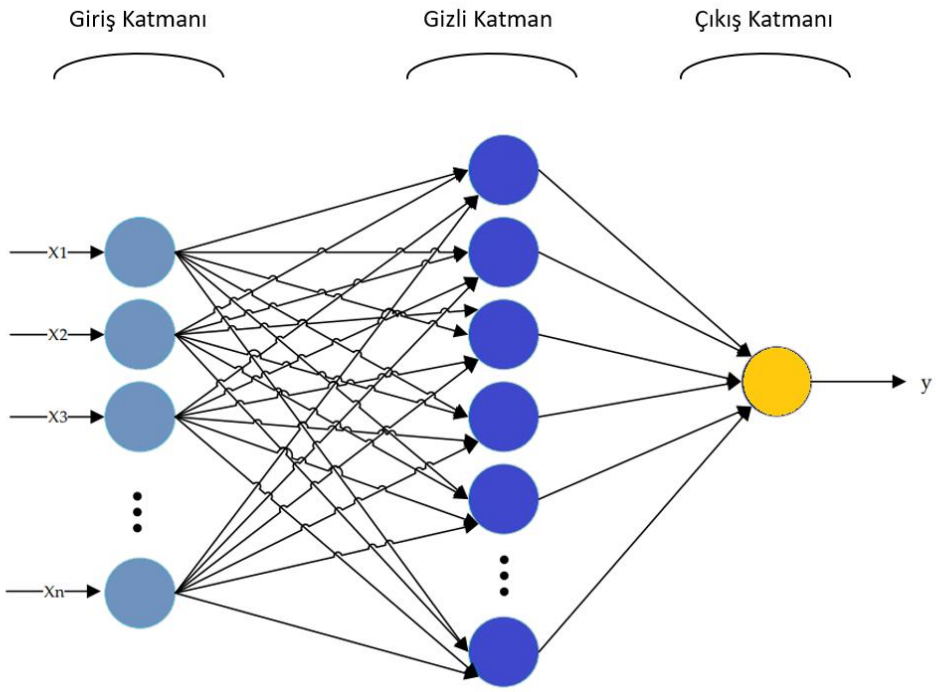
MLP, giriş ve çıkış katmanlarına ve birçok nöronun bir arada bulunduğu bir veya daha fazla gizli katmana sahiptir. MLP'deki nöronlar herhangi bir aktivasyon işlevini kullanabilir. MLP özellikle sınıflandırma ve genel tahmin problemlerinde etkili sonuçlar vermektedir. MLP yapısı şu şekilde tanımlanabilir (Zainal-Mokhtar and Mohamad-Saleh, 2013):

Girdi katmanı: Girdileri bir sonraki katmanlara gönderen, işleme kapasitesi olmayan düğümlerden oluşur.

Gizli katman (bir veya daha fazla): Hesaplama çıktıları sonraki sinir birimlerine girdi olan sinir elemanlarından oluşur.

Çıktı katmanı, hesaplama sonucunda ağın gerçek çıktıları olan düğümlerden oluşur.

Bu katmanlar Şekil 9'da gösterilmiştir.



Şekil 9 MLP Modeli

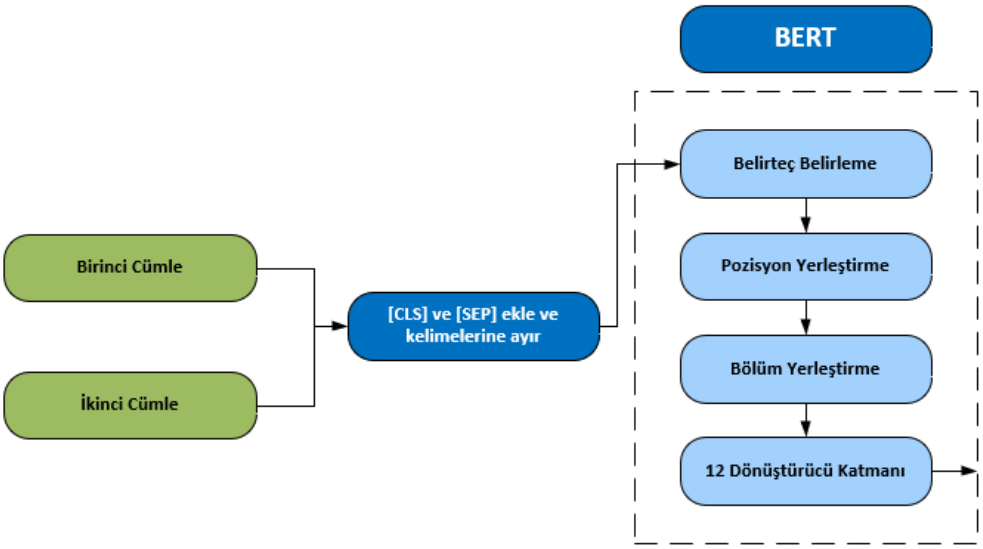
4. Bidirectional Encoder Representations from Transformers (BERT)

Yakın zamanda literatüre giren BERT, Google tarafından Ekim 2018'de önerilen DDİ ön eğitimi için Transformer (derin öğrenme modeli) tabanlı bir makine öğrenme tekniğidir (Devlin et al., 2018).

BERT, kelime temsillerini öğrenmek için Transformer (Vaswani et al., 2017) mimarisini kullanır. Transformer, dizi modelleme için derin ağlara dahil edilebilecek yeni bir mimaridir. Transformer, yalnızca dikkat mekanizmalarını kullanarak girdi ve çıktı arasındaki global bağımlılıkları öğrenir (Nozza et al., 2020).

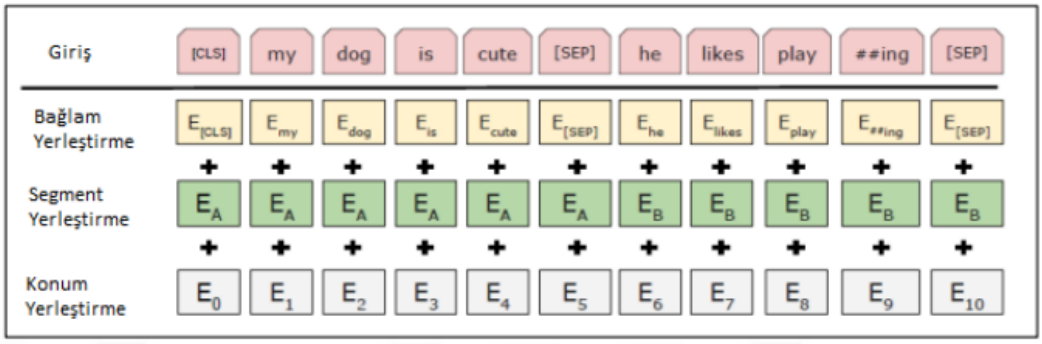
BERT yapısındaki ağ ile ilk katmandan son katmana kadar belirtecin hem sağ hem de sol bağlamından bilgileri etkin bir şekilde yakalar. Sınıflandırmada son kodlayıcı katmanının konumuna tam bağlı bir katman bağlanması için dizinin ilk sözcüğü benzersiz bir belirteçle tanımlanır. Son olarak Softmax katmanı veri sınıflandırma işlemini sonuçlandırır (Gao et al., 2019).

BERT kelime yerleştirme modeli belirteç yerleştirme, bölüm yerleştirme ve pozisyon yerleştirme olmak üzere üç bölümden oluşmaktadır (Ghorbanali et al., 2022). Modelin tüm girdilerinden elde edilen sonuç üç yerleştirme modelinin toplamında elde ettiği çıktının sonucudur. Şekil 10'da BERT modeline ait detaylar gösterilmektedir.



Şekil 10 BERT Modeli

Şekil 11'de belirtiği gibi BERT modelinde her cümlenin başına başlangıcı temsil eden CLS (Classifier Token) ve cümle sonunu temsil eden SEP (Sentences Separator) eklenir. Bu yerleştirme BERT'in belirteç oluşturması ile elde edilmektedir.



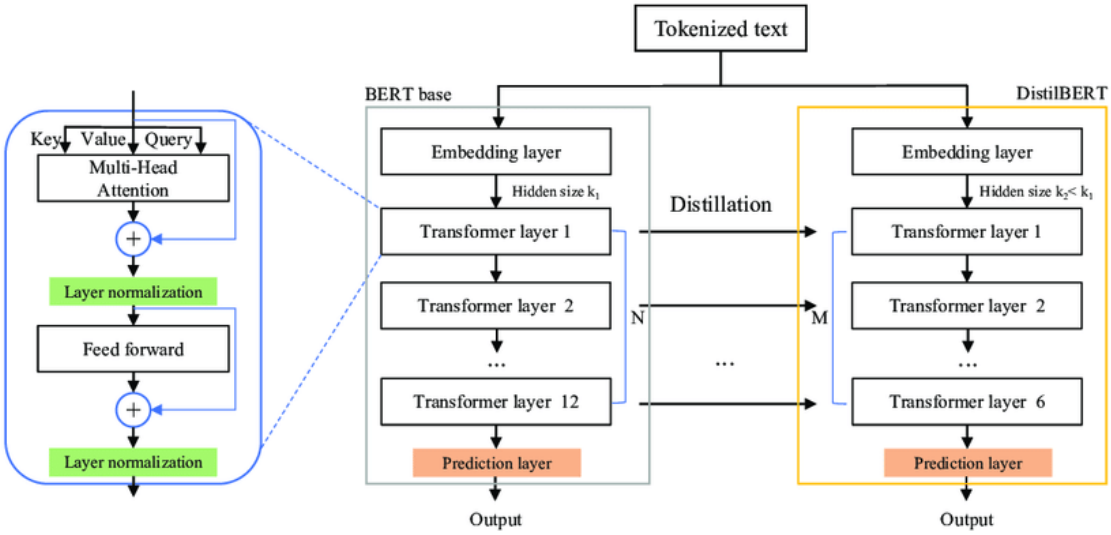
Şekil 11 BERT Örnek Cümle Gösterimi (Devlin et al., 2018)

BERT, Base ve Large olmak üzere 2 farklı varyasyona sahiptir. BERT-Base, 12 katman, 768 gizli boyut ve toplam 110M parametreli 12 dikkat başlığı içerir. BERT-Large ise 24 katman, 16 dikkat başlığı, 1024 gizli boyut ve 340M toplam parametre içermektedir. Base ve Large varyasyonları da cased ve uncased olmak üzere 2 farklı versiyona sahiptir. Uncased versiyon, sözcük tokenizasyon işleminden önce metni küçük harfe dönüştürür. Diğer yandan, cased versiyon büyük/küçük harfe duyarlıdır (Chiorrini et al., 2021).

5. DistilBERT

DistilBERT, BERT modeli ile aynı mimari yapıya sahiptir (Sanh et al., 2019). DistilBERT hızlı, ekonomik, küçük ve çok daha hafif yapılar kullanmaktadır. Ayrıca, %40 daha az parametre kullanarak %60 daha hızlı çalışma özelliğine sahiptir (Hussna et al., 2021). Distilasyonun temel fikri, DistilBERT gibi daha küçük bir model kullanarak BERT modelinin tam çıktı dağılımlarını yaklaştırmaktır (Adel et al., 2022).

DistilBERT-base-uncased modeli, BERT-base-uncased modelinden distile edilmiş olup 6 transformer ve 12 öz-dikkat katmanı, 768 gizli katman ve 66M parametreye sahiptir (Gupta et al., 2021). DistilBERT'e ait mimari Şekil 12'de gösterilmiştir.



Şekil 12 DistilBERT Model Mimarisi (Adel et al., 2022)

BERT ve DistilBERT arasındaki temel farklar şunlardır:

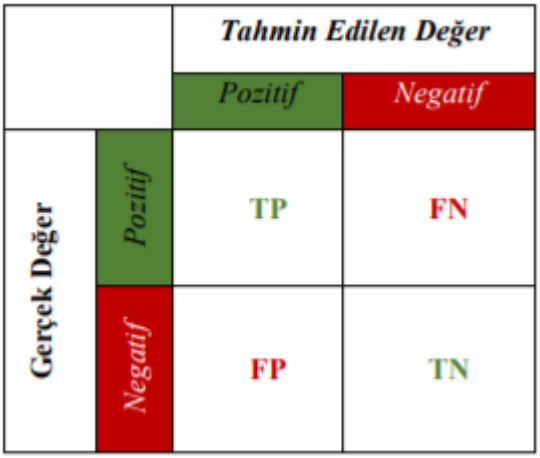
Model boyutu olarak BERT, büyük bir model olup DistilBERT'e göre daha fazla parametre içermektedir. DistilBERT ise daha küçük bir model olup, BERT'in bilgi distilasyonuyla eğitilmiş ve sıkıştırılmış versiyonudur. Hesaplama gücü açısında bakıldığında ise BERT, büyük boyutu nedeniyle daha fazla hesaplama gücü gerektirirken, DistilBERT daha hafif bir modeldir ve daha düşük hesaplama gücü gerektirir. Eğitim süresi açısından ise BERT, daha büyük boyutu ve daha fazla parametreye sahip olduğu için eğitimi daha uzun sürebilir. DistilBERT, daha küçük boyutta olmasından dolayı daha hızlı eğitilebilir.

E. Sınıflandırma Değerlendirme Metrikleri

Sınıflandırma performans ölçümünde Doğruluk (Accuracy), Hassasiyet (Precision), Duyarlılık (Recall) ve F1 ölçütü değerleri kullanılmaktadır. Şekil 13'te gösterilen karmaşıklık matrisindeki TP (True Positive), FP (False Positive), FN (False Negative) ve TN (True Negatif) kısaltmaları aşağıdaki anlamları temsil etmektedir (Salur and Aydin, 2020).

TP, sınıflandırma sonucunda gerçek sınıf etiketinin pozitif ve tahmin edilen sınıf etiketinin de pozitif olduğu örneklerin sayısını, FP, gerçek sınıf etiketinin negatif ve tahmin edilen sınıf etiketinin pozitif olduğu örneklerin sayısını, FN, gerçek sınıf etiketinin pozitif ve tahmin edilen sınıf etiketinin negatif olduğu örneklerin sayısını,

TN ise gerçek sınıf etiketinin negatif ve tahmin edilen sınıf etiketinin negatif olduğu örneklerin sayısını ifade etmektedir.



Şekil 13 İki Sınıflı Bir Problem İçin Karmaşıklık Matrisi

Doğruluk değerinin hesaplanması Denklem 12'ye göre yapılmaktadır.

$$Doğruluk = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Denklem 12

Hassasiyet, her bir sınıf için doğru bir şekilde tahmin edilen sınıf etiketlerinin toplam tahmine oranını ifade etmektedir. Hassasiyet değeri Denklem 13'e göre hesaplamaktadır.

$$Hassasiyet = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Denklem 13

Duyarlılık, her bir sınıf için her bir sınıf için doğru bir şekilde sınıflandırılan doğru etiketlerin ağırlıklı ortalamasıdır. Duyarlılık değeri Denklem 14'e göre hesaplamaktadır.

$$Duyarlılık = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Denklem 14

F1 ölçütü ise hassasiyet ve duyarlılık değerlerini birlikte kullanan bir diğer performans ölçütüdür. F1 ölçütü 0 ile 1 arasında değer almaktadır. F1 ölçütü ne kadar 1'e yakınsa, sınıflandırıcının performansı o kadar iyi demektir. F1 ölçütünün hesaplanması için Denklem 15'teki eşitlik kullanılmaktadır.

$$F_1 = \frac{2 * \text{Hassasiyet} * \text{Duyarlılık}}{\text{Hassasiyet} + \text{Duyarlılık}}$$

Denklem 15

IV. DENEYSEL ÇALIŞMA

Bu bölümde yazılım ve donanım ortamı, veri kümesi, veri ön işleme, ontoloji tabanlı öznitelik boyut indirgeme ve deneysel çalışmalara yer verilmiştir. Sınıflandırma modellerinin başarısı için Doğruluk (Accuracy), Hassasiyet (Precision), Duyarlılık (Recall) ve F1 ölçütü kullanılmış olup, deneylerde kullanılan parametrelere ayrıca yer verilmiştir.

A. Yazılım ve Donanım Ortamı

Tez kapsamında tüm deneyler 2.60 GHz 6 çekirdek Intel Core i7 işlemcili ve 16 GB belleğe sahip bir bilgisayarda ve Windows 10 Enterprise işletim sistemi üzerinde gerçekleştirilmiştir. Çalışmada Python 3.8.2 programlama dili ve IDE olarak Visual Studio Code kullanılmıştır. Veri ön işleme aşamasında nltk.corpus, SymSpell ve WordNetLemmatizer, ontoloji tabanlı boyut indirgeme işlemi için nltk.corpus WordNet, kelime temsil yöntemleri için gensim ve sklearn, sınıflandırma işlemi için ise sklearn, keras ve transformers kütüphaneleri kullanılmıştır.

Deneylerin kolay ve hızlı olarak yapılabilmesi için bir uygulama geliştirilmiştir. Uygulama menüsünde seçilen kriterlere göre veri önişleme, WordNet, kelime gömme ve sınıflandırma yöntemleri çalıştırılabilmektedir.

B. Veri Kümesi

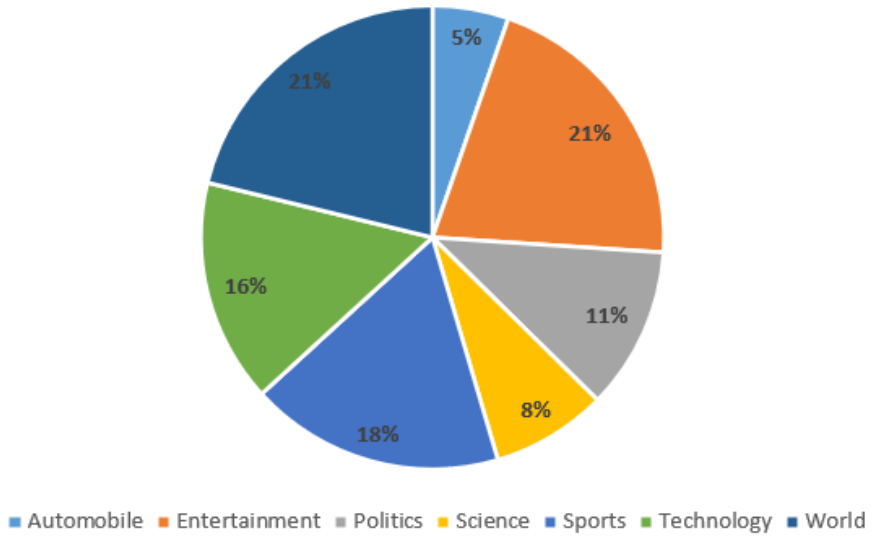
Doküman sınıflandırmada bir çok farklı veri seti ile çalışmalar yapılmıştır. Örneğin duygu analizi çalışmalarında genellikle minimum 2 sınıf etiketi kullanılmıştır. Bunun haricinde Twitter, şikayet, yorum, spam ve haber gibi veri setleri kullanılarakta yapılan çalışmalar mevcuttur. Bu veri setlerinde bulunan sınıf etiketleri de genellikle 2-7 arasında değişmektedir. Sınıf etiketi sayısının az olduğu araştırmalarda yüksek başarıya daha kolay ulaşılırken, fazla olduğu durumlarda da eğitim zorlaştığı için yüksek başarıya ulaşmak çok kolay değildir. Ayrıca veri sayısınını fazla ve denegeli olması da modelin başarısını artırmaktadır. Bu çalışma kapsamında da daha az veri ve

daha çok sınıf etiketinin olduğu bir veri seti seçilmiş olup, makine öğrenimi ve derin öğrenme modellerinin başarı durumları değerlendirilmiştir. Çalışma kapsamında Kaggle'dan alınan İngilizce haberler yayınlayan Inshorts haber sitesine ait 4817 satır haber içerik verisi kullanılmıştır. Veri setinde "Automobile", "Entertainment", "Politics", "Science", "Sports", "Technology", "World" olmak üzere 7 farklı kategoride (sınıf etiketi) haber bulunmaktadır. (Yadav, 2022). Deneysel çalışmalarda kullanılan veri setine ait detaylı bilgiler Çizelge 2'de gösterilmektedir.

Çizelge 2 Veri Seti Detayı

No	Kategori (Sınıf Etiketi) Adı	Veri Sayısı
1	Automobile	256
2	Entertainment	998
3	Politics	546
4	Science	389
5	Sports	856
6	Technology	751
7	World	1021

Kullanılan ham veri seti 1.68 MB metinsel veri içermektedir. Ayrıca Şekil 14'te de tabloda da görüldüğü üzere sınıf etiketleri arasında tam bir denge bulunmamaktadır. Bu durum da modelde yüksek başarı elde etmeyi zorlaştırmaktadır.



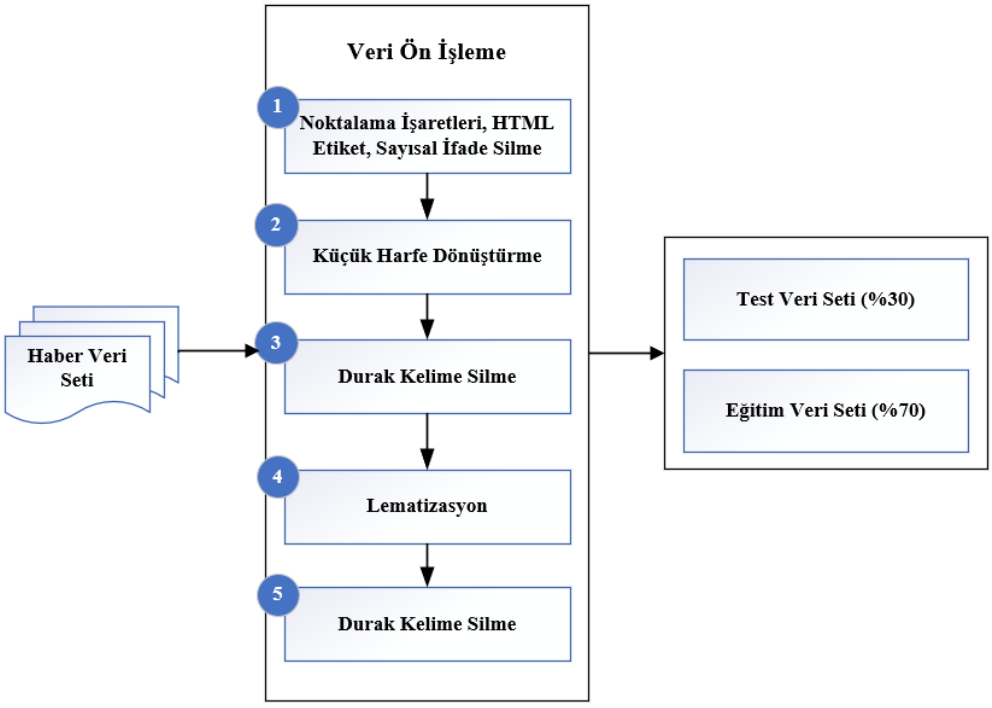
Şekil 14 Sınıf Etiketi Dağılımı

C. Veri Ön İşleme

Veri ön işleme aşamasında sırasıyla aşağıdaki çalışmalar yapılmıştır:

- Noktalama işaretleri kaldırıldı.
- HTML etiketler kaldırıldı.
- Sayısal ifadeler kaldırıldı.
- Tüm kelimeler küçük harfe dönüştürüldü.
- Durak kelimeler kaldırıldı.
- Kelime yazım düzeltimi yapıldı.
- Lematizasyon uygulandı.
- Lematizasyon sonrası oluşan durak kelimeler kaldırıldı.

Ön işleme aşamasında kelime düzeltme için Python SymSpell, durak kelimeleri kaldırmak ve lematizasyon işlem için NLTK kütüphanesi kullanılmıştır. Veri ön işleme sonunda veri seti %70 eğitim ve %30 test olarak 2'ye ayrılmıştır. Tüm veri ön işleme adımları Şekil 15'te gösterilmiştir.



Şekil 15 Veri Ön İşleme Aşamaları

D. Ontoloji Tabanlı Öznitelik Boyut İndirgeme

Bu çalışmanın en önemli ve fark oluşturan aşamalarından birisi öznitelik boyut indirgeme aşamasıdır. Bu aşmada NLTK WordNet kütüphanesi kullanılarak vektör uzayının boyutu daraltılmıştır.

WordNet'in geliştirmesi sırasında sentez kümeleri, sözdizimsel kategoriye ve mantıksal gruplamalara dayalı olarak kırk beş sözlükbilimci dosyası halinde düzenlenmiştir.

Her bir sözlükbilimci dosyasına bir dosya numarası karşılık gelir. Dosya numaraları, bir sözlükbilimci dosya adını belirtmek için etkili bir yol olarak WordNet sisteminin çeşitli bölümlerinde kodlanmıştır. Dosya sözcük adları, dosya adları ve numaraları arasındaki eşlemeyi listeler ve programlar veya son kullanıcılar tarafından ikisini ilişkilendirmek için kullanılabilir.

Sözlükbilimci dosyalarının adları ve bunlara karşılık gelen dosya numaraları, her dosyanın içeriğinin kısa bir açıklamasıyla birlikte Çizelge 3'te listelenmiştir.

Çizelge 3 Sözlükbilimci Dosyaları (wordnet.princeton.edu, 2022)

Dosya Numarası	İsim	Açıklama
0	adj.all	tüm sıfat kümeleri
1	adj.pert	ilişkisel sıfatlar (pertainyms)
2	adv.all	tüm zarflar
3	noun.Tops	isimler için benzersiz başlangıç
4	noun.act	eylemleri veya eylemleri ifade eden isimler
5	noun.animal	hayvanları ifade eden isimler
6	noun.artifact	insan yapımı nesneleri ifade eden isimler
7	noun.attribute	insanların ve nesnelerin niteliklerini gösteren isimler
8	noun.body	vücut kısımlarını gösteren isimler
9	noun.cognition	bilişsel süreçleri ve içerikleri ifade eden isimler
10	noun.communication	iletişimsel süreçleri ve içerikleri ifade eden isimler
11	noun.event	tabiat olaylarını ifade eden isimler
12	noun.feeling	duygu ve duyguları ifade eden isimler
13	noun.food	yiyecek ve içecekleri ifade eden isimler
14	noun.group	insan veya nesne gruplarını ifade eden isimler

15	noun.location	uzamsal konumu ifade eden isimler
16	noun.motive	hedefleri ifade eden isimler
17	noun.object	doğal nesneleri ifade eden isimler (insan yapımı olmayan)
18	noun.person	insanları ifade eden isimler
19	noun.phenomenon	doğal olayları ifade eden isimler
20	noun.plant	bitkileri ifade eden isimler
21	noun.possession	mülkiyeti ve mülkiyeti devretmeyi ifade eden isimler
22	noun.process	doğal süreçleri ifade eden isimler
23	noun.quantity	miktarları ve ölçü birimlerini ifade eden isimler
24	noun.relation	insanlar veya şeyler veya fikirler arasındaki ilişkileri ifade eden isimler
25	noun.shape	iki ve üç boyutlu şekilleri ifade eden isimler
26	noun.state	istikrarlı iş durumlarını ifade eden isimler
27	noun.substance	maddeleri ifade eden isimler
28	noun.time	zaman ve zamansal ilişkileri ifade eden isimler
29	verb.body	bakım, giyinme ve bedensel bakım fiilleri
30	verb.change	büyüklük fiilleri, sıcaklık değişimi, yoğunlaşma vb.
31	verb.cognition	düşünme, yargılama, analiz etme, şüphe duyma fiilleri
32	verb.communication	söyleme, sorma, sipariş verme, şarkı söyleme fiilleri
33	verb.competition	dövüş fiilleri, atletik faaliyetler
34	verb.consumption	yeme içme fiilleri
35	verb.contact	dokunmak, vurmak, bağlamak, kazmak fiilleri
36	verb.creation	dikmek, pişirmek, boyamak, icra etmek fiilleri
37	verb.emotion	duygu fiilleri
38	verb.motion	yürümek, uçmak, yüzmek fiilleri
39	verb.perception	görme, duyma, hissetme fiilleri
40	verb.possession	satın alma, satma, sahip olma fiilleri
41	verb.social	siyasi ve sosyal faaliyet ve olayların fiilleri
42	verb.stative	olmak, sahip olmak, uzamsal ilişkiler fiilleri
43	verb.weather	yağmur yağdırmak, kar yağdırmak, eritmek, gürlemek fiilleri
44	adj.ppl	Katılımcı sıfatlar

Çalışmada 11 adet sözlükbilimci dosya kullanılmış olup, bunlar; “noun.food”, “noun.location”, “'noun.person”, “noun.possession”, “noun.quantity”, “noun.shape”, 'noun.time', “verb.emotion”, “verb.possession”, “verb.social”, “verb.weather”’dır.

Çizelge 4, veri ön işleme adımından sonra ve WordNet sözlüksel ontoloji uygulandıktan sonra veri kümesindeki örnek 3 satırın orijinal durumunu göstermektedir. Kalın olarak işaretlenen sözcükler, WordNet sözlükbilimci dosyasına göre değişiklikleri gösterir.

Çizelge 4 Veri Ön İşleme ve WordNet Sonrası Verilerdeki Değişiklik

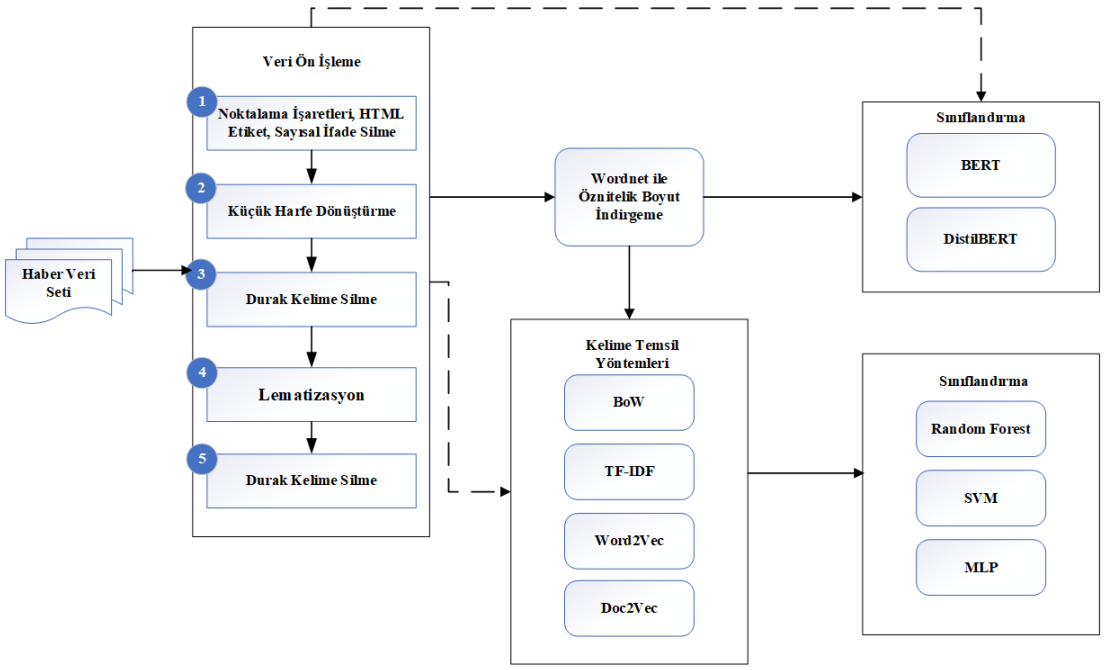
Orjinal Veri	**Veri Ön İşleme Sonrası**	**WordNet Uygulandıktan Sonra**
Iranian authorities on Saturday executed journalist Ruhollah Zam over his online work that helped inspire nationwide economic protests in 2017. A court had sentenced Zam to death in June after he was found guilty of "corruption on earth", one of the country\'s most serious offences. Zam had been living in exile in France but was arrested in October last year.	iranian authority saturday executed journalist roll online work helped inspire nationwide economic protest court sentenced death june found guilty corruption earth one country serious offence living exile france arrested october last year	**person** authority **time social person** roll online work **social emotion** nationwide economic protest court sentenced death **time possession** guilty corruption earth quantity country serious offence living **person location** arrested **time time time**
Tokyo Stock Exchange (TSE) President and CEO Koichiro Miyahara will step down to accept responsibility over a system failure last month that resulted in the first all-day stoppage of trading since the exchange switched to all-electronic trading in 1999. Akira Kiyota, the Group CEO of Japan Exchange Group that runs the TSE, will temporarily take over Miyahara's role.	tokyo stock exchange president co cairo micah ara step accept responsibility system failure last month resulted first day stoppage trading since exchange switched electronic trading akita toyota group co japan exchange group run temporarily take micah ara role	**location possession** exchange **person** co **location person** ara step accept responsibility system failure **time time** resulted first **time** stoppage trading since exchange switched electronic trading akita toyota group co **location** exchange group run temporarily **possession person** ara role
Mick Schumacher, son of seven-time world champion Michael Schumacher, will be racing for Haas in the next Formula One season. The 21-year-old German signed a multi-year agreement and will partner Russian Nikita Mazepin. "The prospect of being on the Formula One	mick schumacher son seven time world champion michael schumacher racing haas next formula one season year old german signed multi year agreement partner russian nikita maze prospect formula one grid next year make incredibly	**person** schumacher **person quantity time** world champion **person** schumacher racing haas next formula **quantity time time time person** signed multi **time** agreement **person person** nikita maze prospect formula **quantity** grid next **time** make incredibly happy

grid next year makes me incredibly happy...I\'m simply speechless," said Mick. He is currently leading the Formula Two championship.	happy simply speechless said mick currently leading formula two championship	simply speechless said person currently leading formula **quantity** championship

V. DENEYLER VE SONUÇ

A. Önerilen Model

Veri ön işleme aşaması tamamlandıktan sonra deneyler WordNet kullanılan ve WordNet kullanılmayan olmak üzere 2 kısma ayrılmıştır. Daha sonra kelime temsil yöntemleri hem WordNet ile hem de doğrudan RF, SVM ve MLP sınıflandırma algoritmaları ile kullanılmıştır. BERT ve DistilBERT algoritmalarında kelime temsil yöntemlerinin kullanılmasına ihtiyaç olmadığı için bu aşamada deneyler hem WordNet ile hem de WordNet olmadan yapılmıştır. Çalışmada WordNet kullanılmasının amacı, WordNet ile öznitelik vektör boyutunun küçültülmesinin sınıflandırma başarısına olan katkısını araştırmaktır. Yapılan çalışmalara ilişkin sistem mimarisi Şekil 16'da gösterilmektedir. Önerilen bu mimari sırasıyla veri seti, veri ön işleme, WordNet ile öznitelik boyut indirgeme, kelime temsil yöntemlerinin uygulanması ve sınıflandırma olmak üzere 5 ana bölümden oluşmaktadır.



Şekil 16 Önerilen Sistemi Mimarisi

B. Makine Öğrenimi Yöntemleri ile Sınıflandırma

Deneylerde kelime temsil yöntemlerinin her biri için farklı parametre değerleri kullanılmış olup, BoW, TF-IDF, Word2Vec, Doc2Vec, SVM, MLP algoritmalarının deneylerde belirlenen optimal parametreler aşağıdaki çizelgelerde gösterilmektedir.

BoW kullanılarak yapılan deneylerde Çizelge 5'te gösterilen optimum parametreler max öznitelik 500, min df 5 ve max df 0.7 olarak bulunmuştur.

Çizelge 5 BoW için Uyarlanmış Optimum Parametreler

Parametre	Parametre Değeri
Max Öznitelik	500
Min df	5
Max df	0.7

TF-IDF kullanılarak yapılan deneylerde Çizelge 6'da gösterilen optimum parametreler max öznitelik 1000, min df 5 ve max df 0.7 olarak bulunmuştur.

Çizelge 6 TF-IDF için Uyarlanmış Optimum Parametreler

Parametre	Parametre Değeri
Max Öznitelik	1000
Min df	5
Max df	0.7

Word2Vec kullanılarak yapılan deneylerde Çizelge 7'de gösterilen optimum parametreler eğitim algoritması skip-gram, pencere boyutu 5, min count 5, vektör boyutu 200, workers 100 ve epoch 100 olarak bulunmuştur.

Çizelge 7 Word2Vec için Uyarlanmış Optimum Parametreler

Parametre	Parametre Değeri
Eğitim Algoritması	skip-gram
Pencere Boyutu	5
Min Count	5
Vektör Boyutu	200

Workers	100
Epoch	100

Doc2Vec kullanılarak yapılan deneylerde Çizelge 8'de gösterilen optimum parametreler eğitim algoritması skip-gram, pencere boyutu 8, vektör boyutu 200, workers 100 ve epoch 25 olarak bulunmuştur.

Çizelge 8 Doc2Vec için Uyarlanmış Optimum Parametreler

Parametre	Parametre Değeri
Eğitim Algoritması	PV-DM
Vektör Boyutu	200
Pencere Boyutu	8
Workers	100
Epoch	25

RF ile yapılan deneylerde Çizelge 9'da gösterilen varsayılan parametre kullanılmıştır.

Çizelge 9 RF için Uyarlanmış Optimum Parametreler

Parametre	Parametre Değeri
Random State	0

SVM kullanılarak yapılan deneylerde Çizelge 10'da gösterilen optimum parametreler max iterasyon 15000, kernel linar ve gamma auto olarak bulunmuştur.

Çizelge 10 SVM için Uyarlanmış Optimum Parametreler

Parametre	Parametre Değeri
Max Iter	15000
Kernel	Linear
Gamma	Auto

Son olarak MLP kullanılarak yapılan deneylerde Çizelge 11'de gösterilen optimum parametreler max iterasyon 50, solver lbfgs, hidden layer boyutu 50, 50, 50 ve aktivasyon relu olarak bulunmuştur.

Çizelge 11 MLP için Uyarlanmış Optimum Parametreler

Parametre	Parametre Değeri
Solver	lbfgs
Max Iter	50
Hidden Layer Boyutu	50, 50, 50
Aktivasyon	Relu

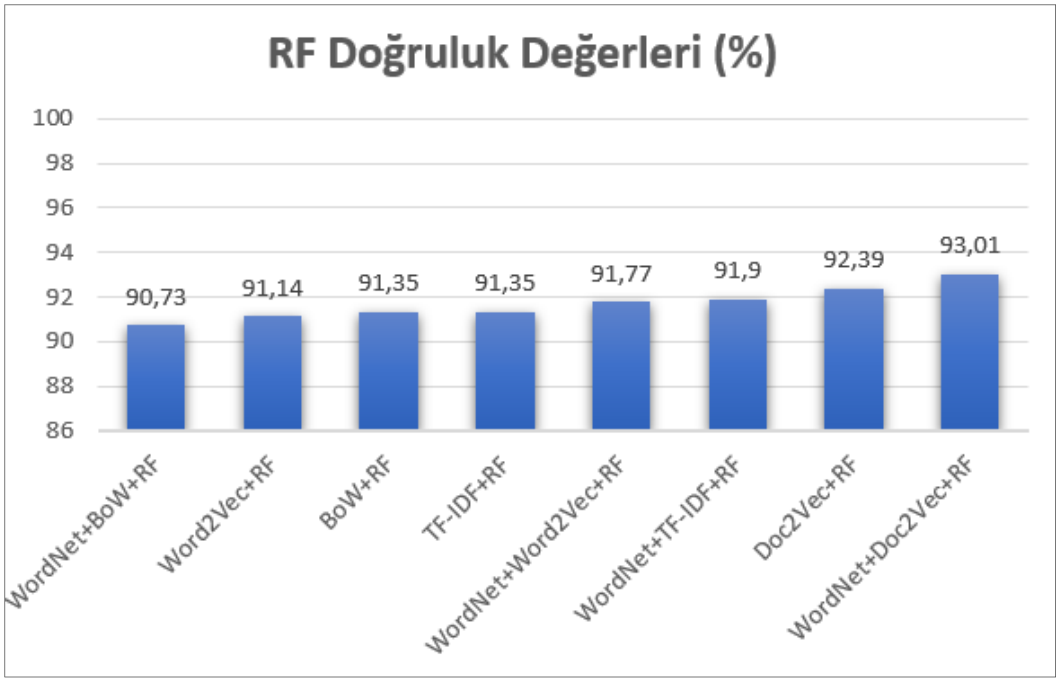
Modellere göre farklı parametrelerle deneyler yapılmış ve sadece RF'te varsayılan parametreler kullanılmıştır. SVM'de doğrusal, polinom, RBF ve sigmoid gibi çekirdek fonksiyonları, 1000 ile 15000 arasında değişen maksimum yineleme ve "auto" olarak ayarlanmış gamma ile deneyler yapılmıştır. Son olarak MLP'de relu ve sigmoid gibi aktivasyon fonksiyonları, 50 ile 100 arasında değişen maksimum iterasyonu ve 30 ile 50 arasında değişen gizli katman boyutları denenmiştir. Deneysel sonuçlar analiz edildikten sonra, her model için en uygun parametreler belirlenmiştir.

Veri seti üzerinde farklı makine öğrenimi sınıflandırıcılarının deney sonuçları Çizelge 12, Çizelge 13 ve Çizelge 14'te gösterilmektedir.

Çizelge 12 RF Sınıflandırmasında Makro Ortalamalı Puanlar

Yöntem	Hassasiyet	Duyarlılık	F1 Ölçütü	Doğruluk
BoW+RF	%91.29	%90.53	%90.87	%91.35
TF-IDF+RF	%90.66	%90.33	%90.43	%91.35
Word2Vec+RF	%90.57	%88.34	%89.32	%91.14
Doc2Vec+RF	%92.67	%91.41	%91.96	%92.39
WordNet+BoW+RF	%90.40	%89.78	%90.04	%90.73
WordNet+TF-IDF+RF	%92.19	%91.70	%91.94	%91.90
WordNet+Word2Vec+RF	%92.60	%90.27	%91.33	%91.77
WordNet+Doc2Vec+RF	%92.55	%92.79	%92.62	%93.01

Şekil 17'de RF ile yapılan deneylerin doğruluk değerlerine ait karşılaştırma grafiği gösterilmektedir.

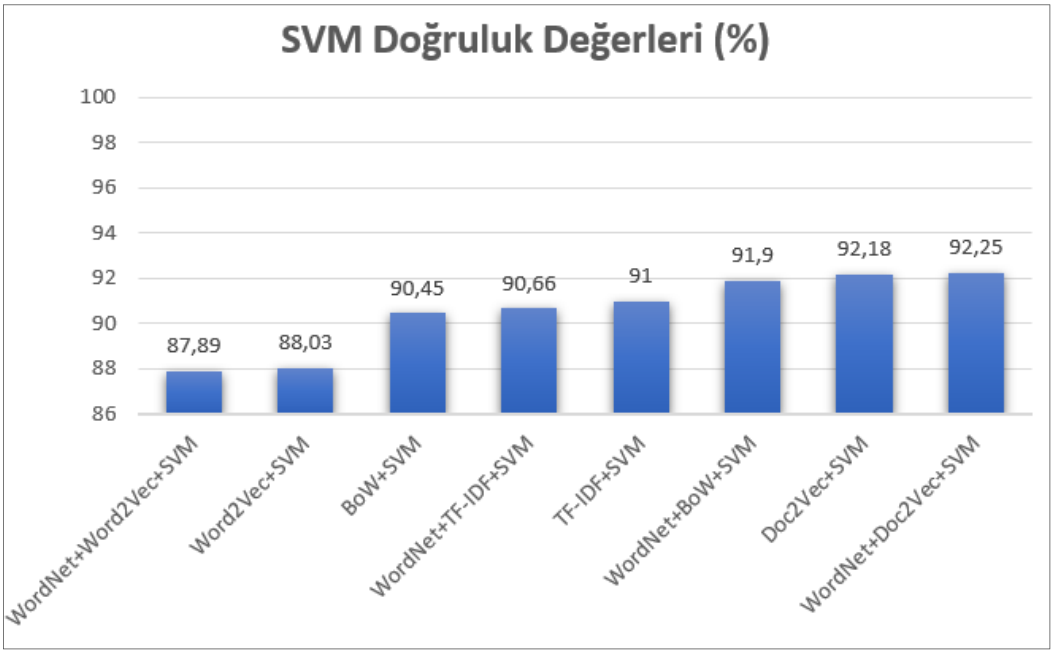


Şekil 17 RF Doğruluk Değerleri

Çizelge 13 SVM Sınıflandırmasında Makro Ortalamalı Puanlar

Yöntem	Hassasiyet	Duyarlılık	F1 Ölçütü	Doğruluk
BoW+SVM	%89.36	%90.42	%89.70	%90.45
TF-IDF+SVM	%89.95	%90.09	%89.91	%91.00
Word2Vec+SVM	%86.96	%87.36	%87.07	%88.0
Doc2Vec+SVM	%91.46	%91.52	%91.41	%92.18
WordNet+BoW+SVM	%91.76	%91.10	%91.38	%91.90
WordNet+TF-IDF+SVM	%90.04	%90.94	%90.37	%90.66
WordNet+Word2Vec+SVM	%87.20	%86.39	%86.68	%87.89
WordNet+Doc2Vec+SVM	%91.70	%92.05	%91.85	%92.25

Şekil 18'de SVM ile yapılan deneylerin doğruluk değerlerine ait karşılaştırma grafiği gösterilmektedir.

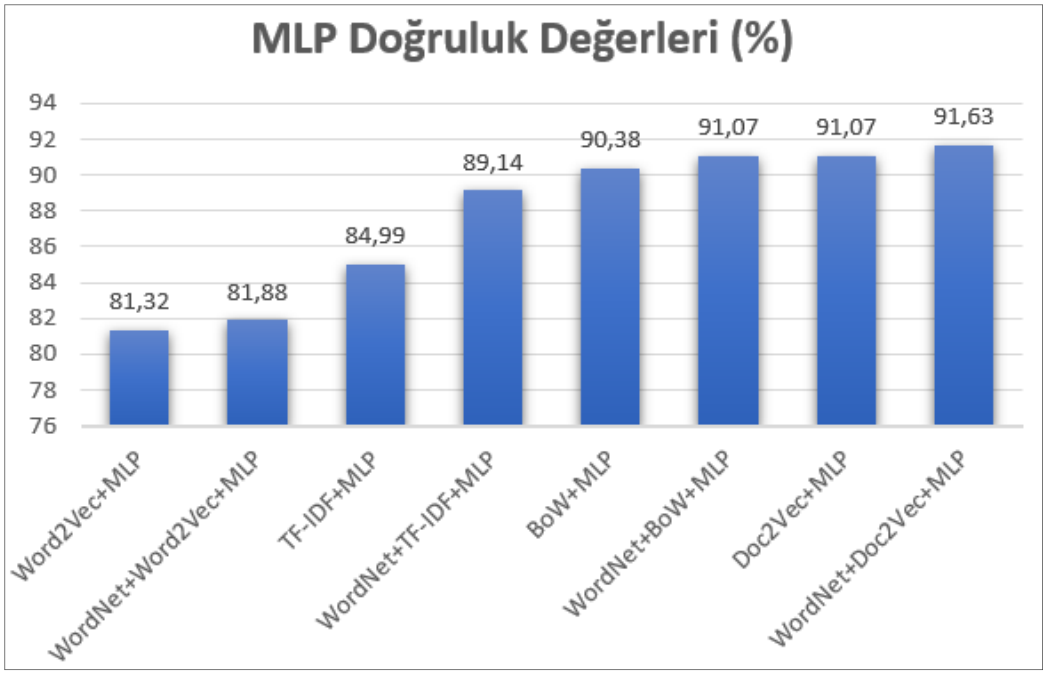


Şekil 18 SVM Doğruluk Değerleri

Çizelge 14 MLP Sınıflandırmasında Makro Ortalamalı Puanlar

Yöntem	Hassasiyet	Duyarlılık	F1 Ölçütü	Doğruluk
BoW+MLP	%89.54	%89.37	%89.43	%90.38
TF-IDF+MLP	%83.72	%82.64	%82.87	%84.99
Word2Vec+MLP	%75.91	%73.89	%74.63	%81.32
Doc2Vec+MLP	%90.69	%89.48	%90.00	%91.07
WordNet+BoW+MLP	%90.35	%90.78	%90.51	%91.07
WordNet+TF-IDF+MLP	%87.60	%87.30	%86.41	%89.14
WordNet+Word2Vec+MLP	%79.41	%75.75	%77.03	%81.88
WordNet+Doc2Vec+MLP	%91.84	%90.73	%91.21	%91.63

Şekil 19'da MLP ile yapılan deneylerin doğruluk değerlerine ait karşılaştırma grafiği gösterilmektedir.



Şekil 19 MLP Doğruluk Değerleri

Çizelge 12, RF algoritması ile sınıflandırma çalışmasının sonuçlarını göstermektedir. Bu kategoride yapılan deneylerde BoW ve RF algoritmasının birlikte kullanılması ile %91.35, TF-IDF ile RF'in birlikte kullanılması ile %91.35, Word2Vec ve RF'in birlikte kullanılması ile %91.14, Doc2Vec ve RF'in birlikte kullanılması ile %92.39, WordNet, BoW ve RF'in birlikte kullanılması ile % 90.73, WordNet, TF-IDF ve RF'in birlikte kullanılması ile %91.90, WordNet, Word2Vec ve RF'in birlikte kullanılması ile %91.77 doğruluk değeri elde edilmiştir. RF kategorisinde en yüksek başarı %93,01 doğrulukla WordNet ontolojisi ve Doc2Vec'in birlikte kullanılması ile elde edilmiştir. Bu başarı, vektör boyutu olarak N = 200, pencere boyutu olarak W = 8, 100 worker ve 25 epoch kullanan Doc2Vec ile elde edilmiştir.

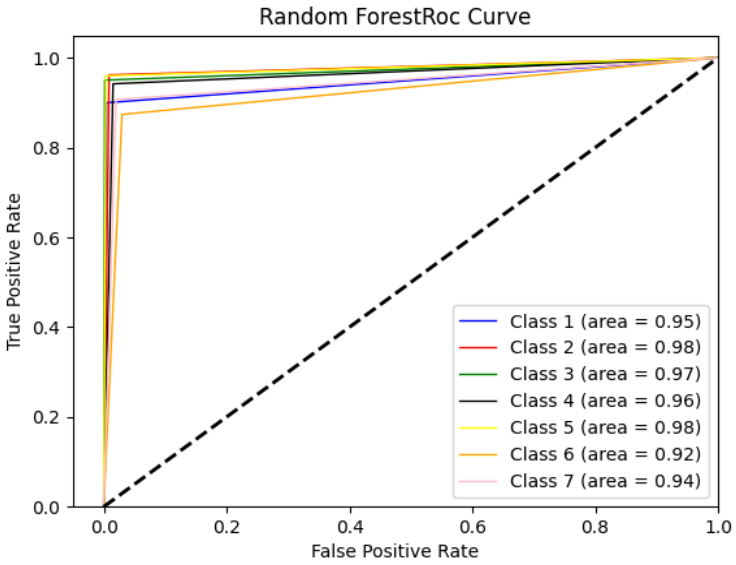
SVM algoritması kullanılarak yapılan deneylerin sonuçları Çizelge 13'te gösterilmiştir. Bu kategoride yapılan deneylerde BoW ve SVM algoritmasının birlikte kullanılması ile %90.45, TF-IDF ile RF'in birlikte kullanılması ile %91.00, Word2Vec ve SVM'in birlikte kullanılması ile %88.03, Doc2Vec ve SVM'in birlikte kullanılması ile %92.18, WordNet, BoW ve SVM'in birlikte kullanılması ile %91.90, WordNet, TF-IDF ve SVM'in birlikte kullanılması ile %90.66, WordNet, Word2Vec ve SVM'in birlikte kullanılması ile %87.89 doğruluk değeri elde edilmiştir. Deneylerde en yüksek başarı değeri WordNet ve Doc2Vec yöntemleri birlikte kullanılarak %92,25 doğrulukla elde edilmiştir. Bu başarı, vektör boyutu olarak N = 200, pencere boyutu

olarak W = 8, 100 worker ve 25 epoch kullanan Doc2Vec ile elde edilmiştir. MLP algoritmasında ise max iter 15000, KernelLinear ve Gamma Auto olarak kullanılmıştır.

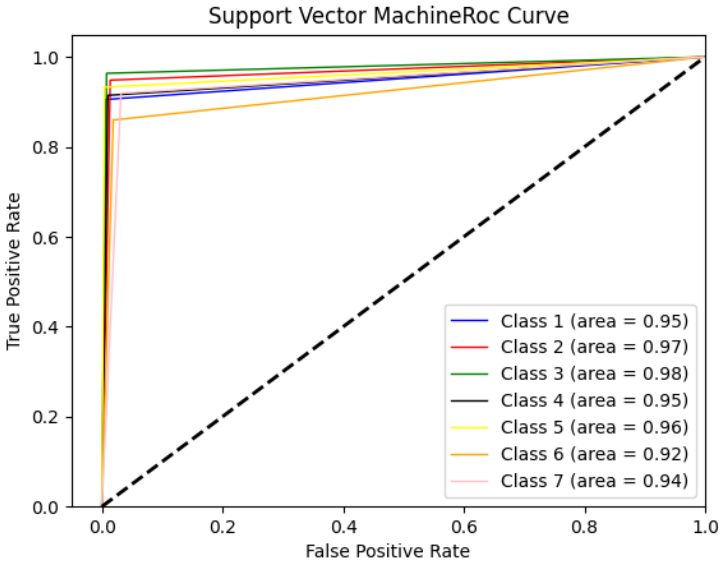
Son olarak Çizelge 14'te gösterilen MLP kategorisinde yapılan deneylerde, BoW ve MLP algoritmasının birlikte kullanılması ile %90.38, TF-IDF ile MLP'nin birlikte kullanılması ile %84.99, Word2Vec ve MLP'nin birlikte kullanılması ile %81.32, Doc2Vec ve MLP'nin birlikte kullanılması ile %91.07, WordNet, BoW ve MLP'nin birlikte kullanılması ile %91.07, WordNet, TF-IDF ve MLP'nin birlikte kullanılması ile %89.14, WordNet, Word2Vec ve MLP'nin birlikte kullanılması ile %81.88 doğruluk değeri elde edilmiştir. WordNet ve Doc2Vec modelleri birlikte kullanılarakta en yüksek doğruluk değeri %91,63 olarak elde edilmiştir. Bu başarı, vektör boyutu olarak N = 200, pencere boyutu olarak W = 8, 100 worker ve 25 epoch kullanan Doc2Vec ile elde edilmiştir. MLP algoritmasında ise solver olarak lbfgs, Max Iter 50, Hidden Layer Sizes 50, 50, 50 ve Activation Relu kullanılmıştır.

Sonuç olarak RF, SVM ve MLP kategorilerinde yapılan deneylerde WordNet'in yanı sıra Doc2Vec modelinin de sınıflandırma başarısını artırmada etkili olduğu görülmüştür. Bunun nedeni, Doc2Vec yönteminde tüm kelime vektörlerinin toplamı ve ortalaması alınarak belgelerin anlamsal olarak korunmasıdır. Diğer bir bulgu ise Doc2Vec modelinde düşük epoch sayılarında yüksek başarı elde edilirken Word2Vec modelinde bunun tam tersi olmasıdır.

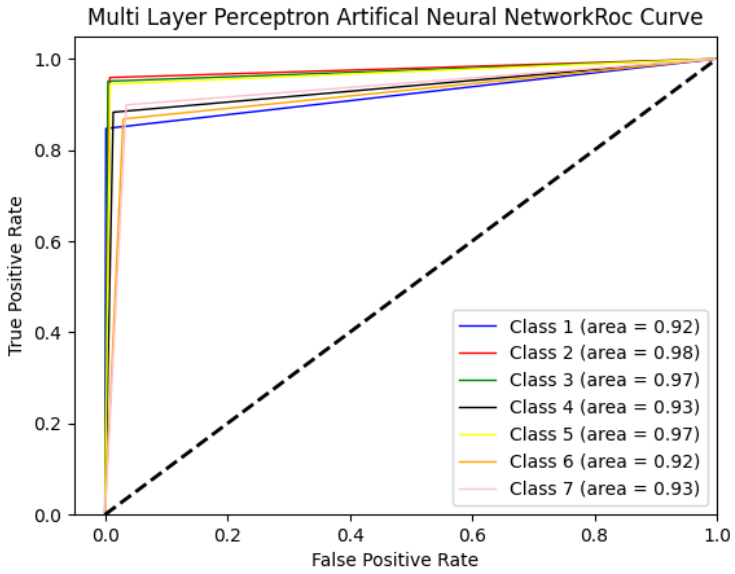
Şekil 20, 21 ve 22'de makine öğrenimi sınıflandırma modelleri arasında en yüksek başarıya sahip yöntemler için ROC eğrileri gösterilmiştir. Her sınıf etiketinin alan değerleri, ROC eğrilerinde ayrı ayrı gösterilir.



Şekil 20 WordNet+Doc2Vec+RF Roc Eğrisi



Şekil 21 WordNet+Doc2Vec+SVM Roc Eğrisi



Şekil 22 WordNet+Doc2Vec+MLP Roc Eğrisi

C. Derin Öğrenme Yöntemleri ile Sınıflandırma

Derin öğrenme tabanlı sınıflandırma olarak önceden eğitilmiş BERT ve DistilBERT kullanılmış ve deneyler sadece BERT, DistilBERT ve BERT+WordNet ve DistilBERT+WordNet kombinasyonu kullanılarak gerçekleştirilmiştir. Test edilen parametreler, 1e-5 ila 4e-5 arasında değişen bir öğrenme oranı, 4 ila 16 arasında değişen parti boyutu, 128 ila 512 arasında değişen maksimum uzunluk ve 1, 3, 5, 7 ve 10 eğitim dönemini içeriyordu. Deney sonuçlarına göre Çizelge 15'te gösterilen BERT ve DistilBERT'in optimal parametreleri belirlenmiştir.

Çizelge 15 BERT ve DistilBERT için Uyarlanmış Optimum Parametreler

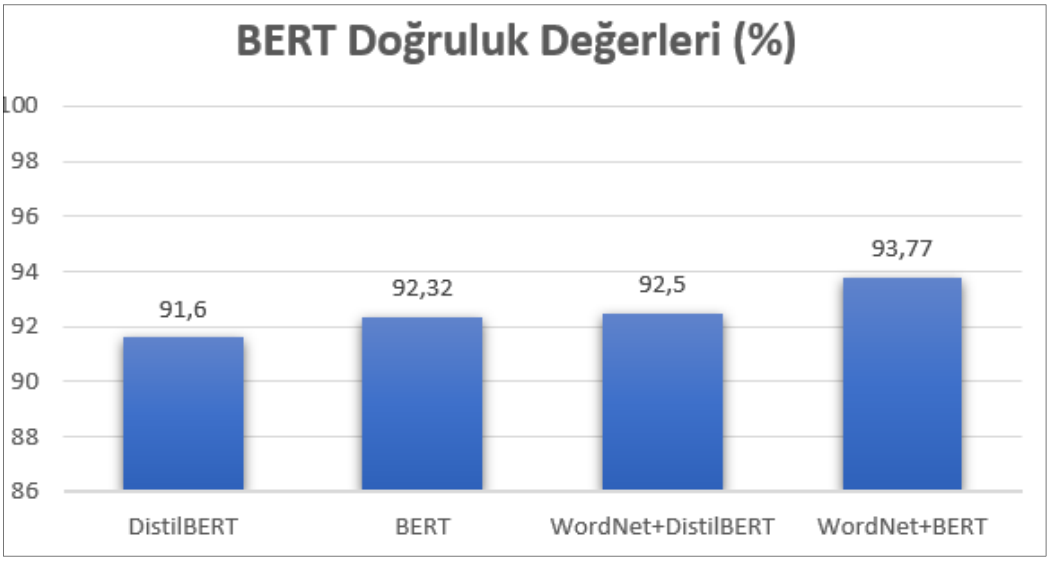
Parametre	Parametre Değeri
Optimizer	Adam
Learning Rate	1e-5
Epsilon	1e-8
Max Length	256
Batch Size	4
Epochs	3 - 5

BERT cased ve uncased versiyonları kullanılarak deneylerin yapıldığı çalışmalarda (Dumitrescu et al., 2020; Jahan et al., 2021; Keya et al., 2022; Yang et al, 2020) uncased versiyonun daha başarılı olduğu görülmüştür. Bu nedenle, deneylerde ince ayar için BERT-Base uncased önceden eğitilmiş modeli seçilmiştir. Transformers kütüphanesi kullanılarak deneyler yapılmıştır. BERT-Base-Uncased modeli Adam optimizer kullanılarak optimize edilmiş olup, en iyi doğruluk değeri 3 ve 5 epoch ile elde edilmiştir. Ayrıca farklı parametre değer kombinasyonları kullanılarakta deneyler gerçekleştirilmiş ve diğer optimum parametre değerleri Çizelge 15'te gösterilmiştir. Batch size hem eğitim hem de doğrulama setinde 4 olarak sabitlenmiştir. 1, 3, 5, 7 ve 10 epoch değerlerinin kullanıldığı deneylerde en yüksek başarı değeri 3 ve 5 epochlarda elde edilmiş, 7 epoch kullanıldığında başarının düştüğü görülmüştür. Ayrıca BERT'in daha küçük, daha hızlı ve daha hafif versiyonu olan DistilBERT üzerinde aynı parametrelerin kullanıldığı deneyler yapılmıştır.

Çizelge 16 BERT ve DistilBERT Sınıflandırmasında Makro Ort. Puanlar

Yöntem	Hassasiyet	Duyarlılık	F1 Ölçütü	Doğruluk
BERT	%92.49	%91.58	%91.94	%92.32
WordNet+BERT	%94.31	%92.99	%93.60	%93.77
DistilBERT	%90.51	%92.57	%91.34	%91.6
WordNet+DistilBERT	%92.71	%92.47	%92.56	%92.5

Şekil 23'te BERT ve DistilBERT ile yapılan deneylerin doğruluk değerlerine ait karşılaştırma grafiği gösterilmektedir.

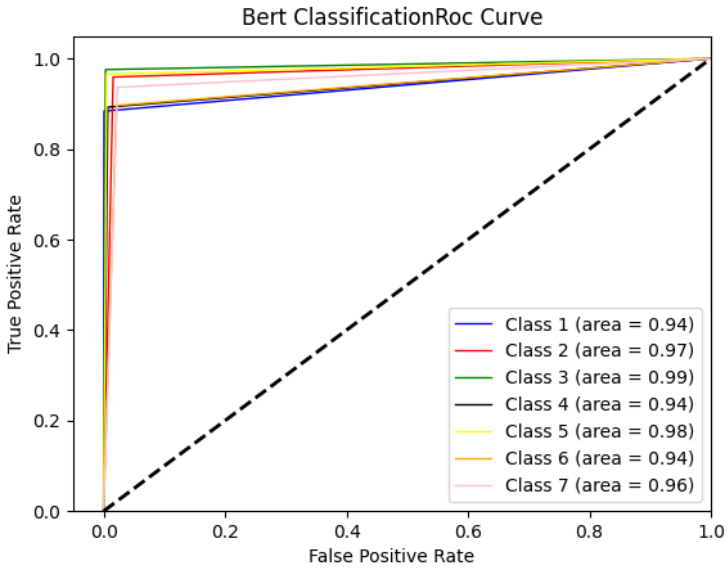


Şekil 23 BERT Doğruluk Değerleri

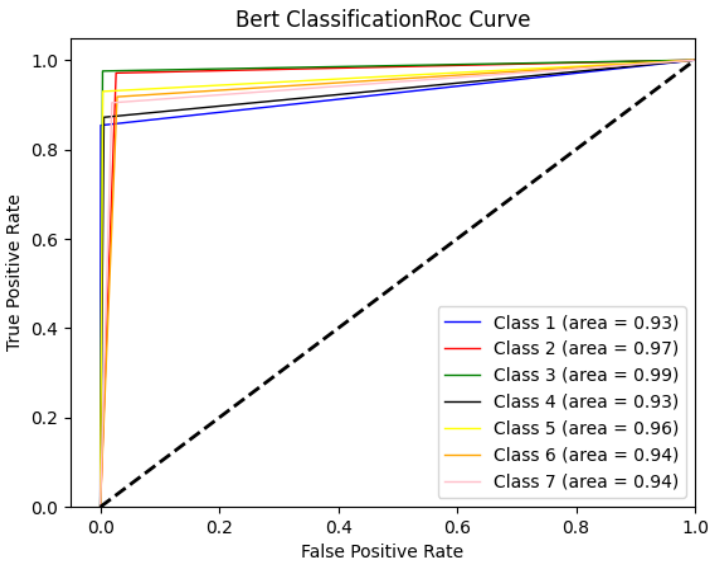
Çizelge 16'da, BERT ve DistilBERT kullanılan deneysel çalışmaların kesinlik, hatırlama, F1 puanı ve doğruluk değerleri gösterilmektedir. Bu kategoride yapılan deneylerde sadece BERT kullanılması ile % 92.32, sadece DistilBERT kullanılması ile % 91.6, WordNet ve DistilBERT'in birlikte kullanılması ile de % 92.5 doğruluk değeri elde edilmiştir. Derin öğrenme kategorisinde yapılan çalışmada en yüksek başarı WordNet ve BERT'in hibrit kullanımı ile elde edilmiş ve doğruluk oranı %93,77 olarak tespit edilmiştir. Bu değer bu çalışmanın en yüksek sonucunu göstermektedir. En yüksek değeri elde etmek için ince ayar yapılmış ve Çizelge 8'deki parametreler kullanılmıştır. Deneyde, parti büyüklüğü hem eğitim hem de doğrulama setinde 4 olarak sabitlenmiştir. Yine 1, 3, 5, 7 ve 10 epoch değerleri kullanılarak en yüksek başarı değeri 3 ve 5 epochlarda elde edilmiş, 7 epoch kullanıldığında başarının düştüğü görülmüştür. Güçlü modeli nedeniyle BERT en yüksek başarıya sahiptir (Devlin et al., 2018). BERT, önceden eğitilmiş derin çift yönlü temsilleri etkinleştirmek için maskelenmiş dil modelleri kullanır. Maskeli dil modeli, girdiden bazı belirteçleri rasgele maskeler ve yalnızca bağlamına dayalı olarak maskelenmiş kelimenin orijinalini tahmin eder. BERT, bağlam ağırlıklı metinleri anlamak için çok güçlüdür. Deneylerde kullanılan veri seti İngilizce ve uzun haber içeriklerinden oluşmaktadır. Bu nedenle, onu anlamsal ve bağlamsal bir bakış açısıyla analiz etmek önemlidir. Bunun için BERT ve WordNet birlikte kullanılarak derin öğrenme yapılmış ve yüksek başarı elde edilmiştir.

BERT modelinin kullanıldığı bir çalışmada, önceden eğitilmiş BERT modelinin doğrudan sınıflandırma görevinde kullanılması performansta istatistiksel olarak anlamlı bir artışa neden olmamıştır. İstatistiksel anlamlılık yerine hiperparametre kullanmanın önemi vurgulanmıştır (Gasmi 2022). Benzer şekilde bu çalışmada da ön eğitim aşamasından sonra farklı parametrelerle ince ayar yapılmıştır.

Şekil 24 ve Şekil 25'te BERT ile WordNet'in ve DistilBERT ile WordNet'in birlikte kullanılması sonucunda elde edilen ROC eğrileri görülmektedir. Her sınıf etiketinin alan değerleri, ROC eğrilerinde ayrı ayrı gösterilmektedir. Derin öğrenme tabanlı bir sınıflandırıcı olan BERT'de en yüksek doğruluk değeri WordNet ve BERT'in hibrit kullanımı ile elde edilmiştir. Bu yöntemi gösteren Şekil 19 incelendiğinde; sınıf etiket alanı değerlerinin birbirine Şekil 20'dekinden daha yakın olduğu görülmektedir.



Şekil 24 BERT+WordNet Roc Eğrisi



Şekil 25 DistilBERT+WordNet Roc Eğrisi

VI. SONUÇ VE ÖNERİLER

Dengesiz ve çok sınıflı metinsel veri ile sınıflandırma yapmak az sayıda ve dengeli sınıf etiketinin bulunduğu veri seti ile sınıflandırma yapmaktan daha zordur. Çünkü öğrenme kolay olmamaktadır. Bu nedenle yüksek sınıflandırma başarısı elde etmek amacıyla yenilikçi yöntemlere başvurulması gerekmektedir.

Yapılan tez çalışmasında dengesiz ve çok sınıflı bir veri seti kullanarak sözlüksel ontolojinin sınıflandırma modellerinin başarısı üzerindeki etkisini araştırmak için makine öğrenmesi ve derin öğrenme modelleri olmak üzere iki farklı kategoride deneyler gerçekleştirilmiştir. Öncelikle İngilizce haber veri seti üzerinde veri ön işleme çalışmaları yapılarak veri seti deneysel çalışmalara hazırlanmıştır. Bu aşamada DDİ tekniklerinden faydalanılarak noktalama işaretleri, HTML etiketler, sayısal ifadeler kaldırılmıştır. Devamında, tüm kelimelerin küçük harfe dönüştürülmesi, durak kelime kaldırma işlemi, kelime yazım düzeltimi, lematizasyon uygulaması ve son olarak lematizasyon sonrası oluşan durak kelimeleri silme işlemi yapılmıştır. Makine öğrenmesi kategorisinde RF, SVM ve MLP modelleri BoW, TF-IDF, Word2Vec ve Doc2Vec kelime gömme yöntemleri ile birlikte eğitilmiştir. Word2Vec yönteminde, eğitim algoritması olarak skip-gram, Doc2Vec yöntemimde ise PV-DM kullanılmıştır. Ayrıca algoritmalara ait diğer parametreler üzerinde de farklı kombinasyonlar denenmiştir. Daha sonra bu modeller WordNet öznitelik boyut indirgeme uygulanarak aynı kelime temsil yöntemleri ile yeniden eğitilmiştir. Model başarısını değerlendirme için Doğruluk, Hassasiyet, Duyarlılık ve F1 ölçütü metrikleri kullanılmıştır. Makine öğrenimi modelleri ile yapılan deneylerde en yüksek başarı, %93,01 doğruluk oranı ile RF modelinin WordNet ve Doc2Vec ile birlikte kullanılmasından elde edilmiştir. Makine öğrenimi sınıflandırma yöntemleri kullanılarak yapılan deneylerde WordNet'in yanı sıra Doc2Vec modelinin de sınıflandırma başarısını artırmada etkili olduğu görülmüştür. Bunun nedeni, Doc2Vec yönteminde tüm kelime vektörlerinin toplamı ve ortalaması alınarak dokümanların anlamsal olarak korunmasıdır. Bir diğer bulgu ise Doc2Vec'te düşük epoch sayılarında yüksek başarı elde edilirken Word2Vec'te tam tersi bir durum söz konusudur. Derin öğrenme kategorisinde ise

BERT ve DistilBERT kullanılarak deneylere devam edilmiştir. Burada veri seti önce BERT ve DistilBERT üzerinde ayrı ayrı eğitilmiş olup, daha sonra WordNet uygulandıktan sonra tekrar eğitilerek deneyler tamamlanmıştır. Çalışmanın en yüksek başarısı %93,77 doğruluk oranı ile WordNet ve önceden eğitilmiş BERT'nin hibrit olarak kullanılmasıyla elde edilmiştir. Deneysel çalışmalar yapılırken modellere ait farklı parametre kombinasyonları denenerek optimal parametreler tespit edilmiştir. Öznitelik boyut indirgeme amacıyla kullanılan WordNet'te, aynı anlama gelen sözcükleri gruplandırmak için veri seti için uygun olan 11 genel sözlükbilimci dosyası seçilmiş ve kullanılmıştır. Veri setiyle ilişkili olmayan sözlükbilimci dosyaları ise göz ardı edilmiştir. Sonuç itibariyle hem geleneksel makine öğrenmesi hem de derin öğrenme tabanlı BERT ve DistilBERT sınıflandırma modellerinde sözlüksel bir ontoloji olan WordNet kullanılarak öznitelik boyut indirgeme işleminin 7 sınıflı ve dengesiz veri seti üzerindeki başarıyı arttırdığı gözlemlenmiştir. Çalışmada kullanılan veri seti İngilizce ve uzun haber içeriklerinden oluştuğu için onu anlamsal ve bağlamsal bir bakış açısıyla analiz etmek önemlidir. Bu nedenle BERT ve WordNet birlikte kullanılarak derin öğrenme yapılmış ve yüksek başarı elde edilmiştir. Ayrıca BERT'in türevi olan DistilBERT BERT'den hızlı çalışmakta olup, WordNet ile birlikte kullanıldığında iyi sonuç vermiştir.

Bu tez çalışması, araştırmacıları sözlüksel ontolojiyi kullanarak belge sınıflandırma araştırması yapmaya motive edebilecek ve önerilen modelin, özellikle yapılandırılmamış verilerin ve sınıflandırılacak birden çok sınıfın olduğu durumlarda, çeşitli metin sınıflandırma görevlerinde uygulanabileceği gösterilmiştir.

Gelecekteki çalışmalarda, deneylerin ALBERT, RoBERTa, XLNet vb. diğer BERT modelleri kullanılarak farklı türde çok sınıflı ve dengesiz veri kümeleri ile WordNet kullanılarak yapılması planlanmaktadır. Ayrıca, çok sınıflı veri setindeki sınıf dengesizliğini gidermek için çeşitli yöntemler kullanarakta deneyler yapılacaktır. Bunun yanı sıra, Graph Convolutional Networks (GCN) son zamanlarda metin sınıflandırmada başarılı sonuçlar verdiği için çok sınıflı ve dengesiz veri seti üzerindeki başarısını ölçmek için kullanılacaktır.

VII. KAYNAKÇA

KİTAPLAR

AGGARWAL, C. C. ve ZHAI, C. (2012). **Mining Text Data**, Springer Science & Business Media.

BREIMAN, L. (2001). **Random Forests**. Machine learning, Springer.

CHOLLET, F. (2017). **Deep Learning with Python, Manning Publications**, Shelter Island.

CORTES, C. ve VAPNIK, V. (1995). **Support-Vector Network,** Machine Learning, Springer.

CUTLER, A., CUTLER, D.R. ve STEVENS, J.R. (2012). **Random Forests,** In: Zhang, C., Ma, Y. (eds) Ensemble Machine Learning. Springer, New York.

GARCÍA, S., LUENGO, J. ve HERRERA, F. (2015). **Data Preprocessing in Data Mining**, Springer.

HAN, J., KAMBER, M. ve PEI, J. (2012). **Data Mining: Concepts and Techniques**, Morgan Kauffman.

OSUNA, E.E., FREUND, R. ve GIROSI, F. (1997). **Support Vector Machines: Training and Applications**, Massachusetts Institute of Technology and Artificial Intelligence Laboratory, Massachusetts.

ÖZKAN, Y. (2008). **Veri Madenciliği Yöntemleri**, Papatya Yayınları, İstanbul.

SİLAHTAROĞLU, G. (2016). **Veri Madenciliği: Kavram ve Algoritmaları**, Papatya Bilim, İstanbul

TAUD, H. ve MAS, J. F. (2018). **Multilayer perceptron (MLP)**, Geomatic Approaches for Modeling Land Change Scenarios.

WANG, L. (2005). **Support Vector Machines: Theory and Applications**, Springer, New York.

ZHANG, X. D. (2020). **A Matrix Algebra Approach to Artificial Intelligence**, Springer.

MAKALELER

ABDOLLAHI, M., GAO, X., MEI, Y., GHOSH, S. ve LI, J. (2020). "A dictionary-based oversampling approach to clinical document classification on small and imbalanced dataset", **In 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)**, ss. 357-364.

ADEL, H., DAHOU, A., MABROUK, A., ABD ELAZİZ, M., KAYED, M., EL-HENAWY, I. M., ALSHATHRI, S. ve ALI, A. A. (2022). "Improving crisis events detection using distilBERT with hunger games search algorithm", **Mathematics**, cilt 10, sayı 3, ss. 447.

ADHIKARI A., RAM A., TANG R. ve LIN J. (2019). "DocBERT: BERT for document classification", **arXiv preprint arXiv:1904.08398**.

AGNIHOTRI, D., VERMA, K. ve TRIPATHI, P. (2017). "Variable Global Feature Selection Scheme for automatic classification of text documents", **Expert Systems with Applications**, cilt 81, ss. 268-281.

AKSU, M. Ç. ve KARAMAN, E. (2020). "FastText ve Kelime Çantası Kelime Temsil Yöntemlerinin Turistik Mekanlar İçin Yapılan Türkçe İncelemeler Kullanılarak Karşılaştırılması", **Avrupa Bilim ve Teknoloji Dergisi**, sayı 20, ss. 311-320.

ALI, F., KWAK, D., KHAN, P., EL-SAPPAGH, S., ALI, A., ULLAH, S., KIM, K. H. ve KWAK, K. S. (2019). "Transportation sentiment analysis using word embedding and ontology-based topic modeling", **Knowledge-Based Systems**, sayı 174, ss. 27-42.

BAHDANAU, D., CHO, K. ve BENGIO, Y. (2014). "Neural machine translation by jointly learning to align and translate", **arXiv preprint arXiv:1409.0473.**

BALAKRISHNAN, V. ve LLOYD-YEMOH E. (2014). "Stemming and Lemmatization: A Comparison of Retrieval Performances", **Lecture Notes on Software Engineering,** cilt 2, sayı 3, ss. 262-267.

BAMATRAF, S. A. ve BIN-THALAB, R. A. (2021). "Semantic Classification Model for Twitter Dataset Using WordNet", **International Research Journal of Innovations in Engineering and Technology**, cilt 5, sayı 2, ss. 5-9.

BARBOUCH, M., VERBERNE, S. ve VERHOEF, T. (2021). "WN-BERT: Integrating WordNet and BERT for Lexical Semantics in Natural Language Understanding", **Computational Linguistics in the Netherlands Journal**, cilt 11, ss. 105-124.

BLOEHDORN, S., BASILI, R., CAMMISA, M. ve MOSCHITTI, A. (2006). "Semantic kernels for text classification based on topological measures of feature similarity", **In Sixth International Conference on Data Mining (ICDM'06)**, ss. 808-812.

CHEBOTKO, A., LU, S., ve ATAY, M. ve Fotouhi, F. (2008). "Efficient processing of RDF queries with nested optional graph patterns in an RDBMS", **International Journal on Semantic Web and Information Systems (IJSWIS)**, cilt 4, ss. 1-30.

CHEN, K., ZHANG, Z., LONG, J. ve ZHANG, H. (2016). "Turning from TF-IDF to TF-IGM for term weighting in text classification", **Expert Systems with Applications**, cilt 66, ss. 245-260.

CHEN, J., LI, K., TANG, Z., BILAL, K., YU, S., WENG, C. ve LI, K. (2016). "A parallel random forest algorithm for big data in a spark cloud computing environment", **IEEE Transactions on Parallel and Distributed Systems**, cilt 28, sayı 4, ss. 919-933.

CHOUDHARY, K. ve BENIWAL, R. (2021). "Xplore Word Embedding Using CBOW Model and Skip-Gram Model", **In 2021 7th International Conference on Signal Processing and Communication (ICSC)** ss. 267-270.

CHIORRINI, A., DIAMANTINI, C., MIRCOLI, A. ve POTENA, D. (2021). "Emotion and sentiment analysis of tweets using BERT", **In EDBT/ICDT Workshops 2021**, cilt 3.

CHUNG, J., GULCEHRE, C., CHO, K. ve BENGIO, Y. (2014). "Empirical evaluation of gated recurrent neural networks on sequence modeling", **arXiv preprint arXiv:1412.3555.**

CRISTIANINI, N., SHAWE-TAYLOR, J. ve LODHI, H. (2002). "Latent semantic kernels", **Journal of Intelligent Information Systems**, cilt 18, ss. 127-152.

ÇOBAN, Ö., ÖZYER, B. ve ÖZYER, G. T. (2015). "Sentiment analysis for Turkish Twitter feeds", **In 2015 23nd Signal Processing and Communications Applications Conference (SIU)**, ss. 2388-2391.

DEVLIN, J., CHANG, M. W., LEE, K. ve TOUTANOVA, K. (2018). "BERT: Pre-training of deep bidirectional transformers for language understanding", **arXiv preprint arXiv:1810.04805.**

DEVLIN, J., CHANG, M. W., LEE, K. ve TOUTANOVA, K. (2019). "BERT: Pre-training of deep bidirectional transformers for language understanding", **Proceedings of NAACL-HLT**, cilt 1, ss. 4171–4186.

DING, Y., WANG, S., XING, J., ZHANG, X., QI, Z., FU, G., QIANG, Q., SUN, H. ve ZHANG, J. (2020). "Malware classification on imbalanced data through self-attention", **In 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)**, ss. 154-161.

DOGAN, T. ve UYSAL, A. K. (2019). "Improved inverse gravity moment term weighting for text classification", **Expert Systems with Applications**, cilt 130, ss. 45-59.

DOGRU, H. B., TILKI, S., JAMIL, A. ve HAMEED, A. A. (2021). "Deep learning-based classification of news texts using doc2vec model", **In 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA)** ss. 91-96.

DUMITRESCU, S. D. AVRAM, A. M. ve PYYSALO, S. (2020). "The birth of Romanian BERT", **arXiv preprint, arXiv:2009.08712.**

ELHADAD, M. K., BADRAN, K. M. ve SALAMA, G. I. (2017). "A novel approach for ontology-based dimensionality reduction for web text document classification", **International Journal of Software Innovation (IJSI)**, cilt 5, sayı 4, ss. 44-58.

FARKIYA, A., SAINI, P., SINHA, S. ve DESAI, S. (2015). "Natural language processing using NLTK and WordNet", **International Journal of Computer Science and Information Technologies**, cilt 6, sayı 6, ss. 5465-5469.

GAO, Z., FENG, A., SONG, X. ve WU, X. (2019). "Target-dependent sentiment classification with BERT", **IEEE Access**, cilt 7, ss. 154290-154299.

GASMI, K. (2022). "Improving BERT-Based Model for Medical Text Classification with an Optimization Algorithm" **In International Conference on Computational Collective Intelligence**, ss. 101-111.

GAWADE, M., MANE, T., GHONE, D., KHADE, P. ve RANJAN, N. (2018). "Text Document Classification by using WordNet Ontology and Neural Network", **International Journal of Computer Applications**, cilt 182, sayı 33, ss. 33-36.

GHORBANALI, A., SOHRABI, M. K. ve YAGHMAEE, F. (2022). "Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks", **Information Processing & Management**, cilt 59, sayı 3, ss. 1-23.

GOGOI, M. ve SHARMA S. K. (2015). "Document Classification of Assamese Text Using Naïve Bayes Approach", **IJCTT**, cilt 30, sayı 4, ss. 182-186.

GONG, J., QIU, X., WANG, S. ve HUANG, X. (2018). "Information aggregation via dynamic routing for sequence encoding", **arXiv preprint arXiv:1806.01501.**

GUPTA, P., GANDHI, S. ve CHAKRAVARTHI, B. R. (2021). "Leveraging transfer learning techniques-BERT, roBERTa, alBERT and distilBERT for fake review

detection", **In Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation**, ss. 75-82.

HALTAŞ, A., ALKAN, A. ve KARABULUT, M. (2015). "Metin Sınıflandırmada Sezgisel Arama Algoritmalarının Performans Analizi", **Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi**, cilt 30, sayı 3, ss. 417-427.

HAN, Q. ve SNAIDAUF D. (2021). "Comparison of Deep Learning Technologies in Legal Document Classification", **2021 IEEE International Conference on Big Data (Big Data)**, ss. 2701-2704.

HONG, L. ve DAVISON, B. (2010). "Empirical Study of Topic Modeling in Twitter", **In Proceedings of the First Workshop on Social Media Analytics**, ss. 80–88.

HUSSNA, A. U., TRISHA, I. I., KARIM, M. S. ve ALAM, M. G. R. (2021). "COVID-19 fake news prediction on social media data", **In 2021 IEEE Region 10 Symposium (TENSYMP)**, ss. 1-5.

IRFANI, F. F., FAUZI, M. A. ve SARI, Y. A. (2018). "News Classification on Twitter Using Naive Bayes and Hypernym-Hyponym Based Feature Expansion", **In 2018 International Conference on Sustainable Information Engineering and Technology (SIET)** ss. 317-321.

IRSOY, O. ve CARDIE, C. (2014). "Opinion mining with deep recurrent neural networks", **In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)**, ss. 720-728.

ISA, D., LEE, L. H., KALLIMANI, V. P. ve RAJKUMAR, R. (2008). "Text document preprocessing with the Bayes formula for classification using the support vector machine", **Int. J. IEEE Transactions on Knowledge and Data Engineering**, cilt 20, sayı 9, ss. 1264-1272.

JAIN, H., BALASUBRAMANIAN, V., CHUNDURI, B. ve VARMA, M. (2019). "Slice: Scalable linear extreme classifiers trained on 100 million labels for related searches", **In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining**, ss. 528-536.

KANG, M., AHN, J. ve LEE, K. (2018). "Opinion mining using ensemble text hidden Markov models for text classification", **Expert Systems with Applications**, cilt 94, ss. 218-227.

KEYA, A. J., WADUD, M. A. H., MRIDHA, M. F., ALATIYYAH, M. ve HAMID, M. A. (2022). "AugFake-BERT: Handling Imbalance through Augmentation of Fake News Using BERT to Enhance the Performance of Fake News Classification". **Applied Sciences**, cilt 12, sayı 17, ss. 1-21.

KIM, D., SEO, D., CHO, S. ve KANG, P. (2019). "Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec", **Information sciences**, cilt 477, ss. 15-29.

KIM, Y. (2014). "Convolutional neural networks for sentence classification", **arXiv preprint arXiv:1408.5882.**

KONG, J., WANG, J. ve ZHANG, X. (2022). "Hierarchical BERT with an adaptive fine-tuning strategy for document classification", **Knowledge-Based Systems**, cilt 238, ss. 1-11.

KUMAR, R. ve KAUR, J. (2020). "Random forest-based sarcastic tweet classification using multiple feature collection", **Multimedia Big Data Computing for IoT Applications: Concepts, Paradigms and Solutions**, ss. 131-160.

LAN, M., TAN, C. L., SU, J. ve LU, Y. (2009). "Supervised and traditional term weighting methods for automatic text categorization", **IEEE Transactions on Pattern Analysis and Machine Intelligence**, cilt 31, sayı 4, ss. 721-735.

LAN, Z., CHEN, M., GOODMAN, S., GIMPEL, K., SHARMA, P. ve SORICUT, R. (2019). "AlBERT: A lite BERT for self-supervised learning of language representations", **arXiv preprint arXiv:1909.11942.**

LE, Q. ve MIKOLOV, T. (2014). "Distributed representations of sentences and documents", **In International conference on machine learning** ss. 1188-1196.

LEDMI, M., LEDMI, A. ve SOUIDI, M. E. H. (2021). "Classification of XML Documents Using Semantic Resources", **In 2021 International Conference on Recent Advances in Mathematics and Informatics (ICRAMI)**, ss. 1-5.

LI, T. ve CHEN, Z. (2020). "An ontology-based learning approach for automatically classifying security requirements", **Journal of Systems and Software**, cilt 165, ss. 1-13.

LIU, Y., OTT, M., GOYAL, N., DU, J., JOSHI, M., CHEN, D., LEVY, O., LEWIS, M., ZETTLEMOYER, L. ve STOYANOV, V. (2019). "RoBERTa: A robustly optimized BERT pretraining approach", **arXiv preprint arXiv:1907.11692.**

LU, Z., DU, P. VE NIE, J. Y. (2020). "VGCN-BERT: augmenting BERT with graph embedding for text classification", **In Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020**, cilt 12035, ss. 369-382.

MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. ve DEAN, J. (2013). "Distributed Representations of Words and Phrases and their Compositionality", **Adv. Neural Inf. Process. Syst**, ss. 3111–3119.

MIKOLOV, T., YIH, W. T. ve ZWEIG, G. (2013). "Linguistic regularities in continuous space word representations", **In Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies**, ss. 746-751.

NOVAKOVIĆ, J. D., VELJOVIĆ, A., ILIĆ, S. S., PAPIĆ, Ž. ve MILICA, T. (2017). "Evaluation of classification models in machine learning", **Theory and Applications of Mathematics & Computer Science**, cilt 7, sayı 1, ss. 39-46.

NOZZA, D., BIANCHI, F. ve HOVY, D. (2020). "What the [mask]? making sense of language-specific BERT models", **arXiv preprint arXiv:2003.02912.**

PENNINGTON, J., SOCHER, R. ve MANNING, C. D. (2014). "Glove: Global vectors for word representation", **In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)**, ss. 1532-1543.

PLISSON, J., LAVRAC, N. ve MLADENIC, D. (2004). "A rule based approach to word lemmatization", **In Proceedings of IS**, cilt 3, ss. 83-86.

PRANCKEVIČIUS, T. ve MARCINKEVIČIUS, V. (2017). "Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification", **Baltic Journal of Modern Computing**, cilt 5, sayı 2, ss. 221-232.

RAJBABU, K., SRINIVAS, H. ve SUDHA, S. (2018). "Industrial information extraction through multi-phase classification using ontology for unstructured documents", **Computers in Industry**, cilt 100, ss. 137-147.

RAMÍREZ-GALLEGO, S., KRAWCZYK, B., GARCÍA, S., WOŹNIAK, M. ve HERRERA, F. (2017). "A survey on data preprocessing for data stream mining: Current status and future directions", **Neurocomputing**, cilt 239, ss. 39-57.

RAY, S. K., SINGH, S. ve JOSHI, B. P. (2010). "A semantic approach for question classification using WordNet and Wikipedia", **Pattern Recognition Letters**, cilt 31, sayı 13, ss. 1935-1943.

REN, F. ve SOHRAB, M. G. (2013). "Class-indexing-based term weighting for automatic text classification", **Information Sciences**, cilt 236, ss. 109-125.

RUCH, P., BAUD, R. ve GEISSBÜHLER, A. (2003). "Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record", **Artificial Intelligence in Medicine**, cilt 29, sayı 1-2, ss.169-184.

SABOUR, S., FROSST, N. ve HINTON, G. E. (2017). "Dynamic routing between capsules", **Proceedings of Advances in Neural Information Processing Systems, NIPS-2017**, ss. 3857–3867.

SALTON, G. ve YU, C. T. (1973). "On the construction of effective vocabularies for information retrieval", **Acm Sigplan Notices**, cilt 10, sayı 1, ss. 48-60.

SALUR, M. U. ve AYDIN, I. (2020). "A novel hybrid deep learning model for sentiment classification", **IEEE Access**, cilt 8, ss. 58080-58093.

SANH, V., DEBUT, L., CHAUMOND, J. ve WOLF, T. (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter", **arXiv preprint arXiv:1910.01108.**

SCHNEIDER, K. M. (2005). "Weighted average pointwise mutual information for feature selection in text categorization", **In European Conference on Principles of Data Mining and Knowledge Discovery** ss. 252-263.

SHANAVAS, N., WANG, H., LIN, Z. ve HAWE, G. (2020). "Ontology-based enriched concept graphs for medical document classification", **Information Sciences**, cilt 525, ss. 172-181.

SHARMA, P., TULSIAN, D., VERMA, C., SHARMA, P. ve NANCY, N. (2022). "Translating speech to Indian Sign Language using natural language processing", **Future Internet**, cilt 14, sayı 9, ss. 1-17.

SHEVLYAKOV, G. ve KAN, M. (2020). "Stream data preprocessing: Outlier detection based on the Chebyshev inequality with applications", **In 2020 26th Conference of Open Innovations Association (FRUCT)**, ss. 402-407.

SHEYKHMOUSA, M., MAHDIANPARI, M., GHANBARI, H., MOHAMMADIMANESH, F., GHAMISI, P. ve HOMAYOUNI, S. (2020). "Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review", **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, cilt 13, ss. 6308-6325.

SPARCK J., K. (2004). "A Statistical Interpretation of Term Specificity and Its Application in Retrieval", **Journal of Documentation**, cilt 28, sayı 1, ss. 11-21.

SOTIROPOULOS, D. N., POURNARAKIS, D. E. ve GIAGLIS, G. M. (2017). "SVM-based sentiment classification: a comparative study against state-of-the-art classifiers", **International Journal of Computational Intelligence Studies**, cilt 6, sayı 1, ss. 52-67.

SONG, D., VOLD, A., MADAN, K. ve SCHILDER, F. (2022). "Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training", **Information Systems**, cilt 106, ss. 1-12.

SUN, G., LIU, J., MENGXUE, W., ZHONGXIN, W., JIA, Z. ve XIAOWEN, G. (2020). "An ensemble classification algorithm for imbalanced text data streams", **In 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)**, ss. 1073-1076.

JAHAN, M. S., BEDDIAR, D. R., OUSSALAH, M., ARHAB, N. ve BOUNAB, Y. (2021). "Hate and Offensive language detection using BERT for English Subtask A", **In Working Notes of FIRE 2021-Forum for Information Retrieval Evaluation**.

TAI, K. S., SOCHER, R. ve MANNING, C. D. (2015). "Improved semantic representations from tree-structured long short-term memory networks", **Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, ACL-IJCNLP**, ss. 1556–1566.

TAIEB, M. A. H., AOUICHA, M. B. ve HAMADOU, A. B. (2014). "Ontology-based approach for measuring semantic similarity", **Engineering Applications of Artificial Intelligence**, cilt 36, ss. 238-261.

TAGAMI, Y. (2017). "Annexml: Approximate nearest neighbor search for extreme multi-label classification", **In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining**, ss. 455-464.

TAN, M., DOS SANTOS, C., XIANG, B. ve ZHOU, B. (2016). "Improved representation learning for question answer matching", **In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics** cilt 1, ss. 464-473.

TILVE, A. K. S. ve JAIN, S. N. (2017). "Text Classification using Naïve Bayes, VSM and Pos Tagger", **International Journal of Ethics in Engineering & Management Education**, cilt 4, sayı 1.

TODOROVSKI, L. ve DŽEROSKI, S. (2003). “Combining classifiers with meta decision trees”, **Machine learning**, cilt 50, sayı 3, ss. 223-249.

UYSAL, A. K. ve GUNAL, S. (2014). “The impact of preprocessing on text classification”, **Information Processing & Management**, cilt 50, sayı 1, ss. 104-112.

UYSAL, A. K. ve GUNAL, S. (2012). “A novel probabilistic feature selection method for text classification”. **Knowledge-Based Systems**, cilt 36, ss. 226-235.

VALARMATHI, B., CHELLATAMILAN T., MITTAL, H., JAGRIT, J. ve SHUBHAM, S. (2019). "Classification of Imbalanced Banking Dataset using Dimensionality Reduction", **2019 International Conference on Intelligent Computing and Control Systems (ICCS)**, ss. 1353-1357.

VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł. ve POLOSUKHIN, I. (2017). “Attention is all you need”, **In Advances in Neural Information Processing Systems**, ss. 5998–6008.

WU, Z., ZHU, H., LI, G., CUI, Z., HUANG, H., LI, J., CHEN E. ve XU, G. (2017). “An efficient Wikipedia semantic matching approach to text document classification”, **Information Sciences**, cilt 393, ss. 15-28.

XIAO, C., ZHONG, H., GUO, Z., TU, C., LIU, Z., SUN, M., FENG, Y., HAN, X., HU, Z., WANG, H. ve XU, J. (2018). “Cail2018: A large-scale legal dataset for judgment prediction”, **arXiv preprint arXiv:1807.02478.**

YANG, Y., UY, M. C. S. ve HUANG, A. (2020). “FinBERT: A pretrained language model for financial communications”, **arXiv preprint arXiv:2006.08097.**

YANG, Z., YANG, D., DYER, C., HE, X., SMOLA, A. ve HOVY, E. (2016). “Hierarchical attention networks for document classification”, **In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies**, ss.1480-1489.

YILDIZ, O. (2016). “Metin madenciliğinde anahtar kelime seçimi bir üniversite örneği”, **Yönetim Bilişim Sistemleri Dergisi**, cilt 2, sayı 1, ss. 29-50.

YOGATAMA, D., DYER, C., LING, W. ve BLUNSOM, P. (2017). "Generative and discriminative text classification with recurrent neural networks", **arXiv preprint arXiv:1703.01898.**

ZAINAL-MOKHTAR, K. ve MOHAMAD-SALEH, J. (2013). "An oil fraction neural sensor developed using electrical capacitance tomography sensor data", **Sensors**, cilt 13, sayı 9, ss. 11385-11406.

ZHANG, C., LIU, C., ZHANG, X. ve ALMPANIDIS, G. (2017). "An up-to-date comparison of state-of-the-art classification algorithms", **Expert Systems with Applications**, cilt 82, ss. 128-150.

ZHANG, X., ZHAO, J. ve LECUN, Y. (2015). "Character-level convolutional networks for text classification", **Advances in Neural Information Processing Systems**, pp. 649–657.

ZHANG, Y., JIN, R. ve ZHOU, Z. H. (2010). "Understanding bag-of-words model: a statistical framework", **International Journal of Machine Learning And Cybernetics**, cilt 1, sayı 1-4, ss. 43-52.

ELEKTRONİK KAYNAKLAR

URL-1 "WordNet A Lexical Database for English", https://WordNet.princeton.edu/documentation/lexnames5wn, (Erişim Tarihi: 23.12.2022).

YADAV, K., "News Classification", Kaggle, https://www.kaggle.com/datasets/kishanyadav/inshort-news?select=inshort_news_data-1.csv, (Erişim Tarihi: 23.12.2022)

TEZLER

ACET, A. (2022). "SVM, NB, KNN, ADABOOST ve Random Forest Sınıflandırma Algoritmaları Kullanılarak Meme Kanserinin Tahmini", (Yüksek Lisans Tezi), Fen Bilimleri Enstitüsü, İnönü Üniversitesi.

ÇOBANOĞLU, Ö. E. (2015). "Comparison of Document Classification Approaches for Turkish Texts", (Yüksek Lisans Tezi), Mühendislik ve Fen Bilimleri Enstitüsü, İzmir Yüksek Teknoloji Enstitüsü.

ELMAS, M. (2012). “Destek Vektör Makineleri ile Fiyat Tahminleri ve Kuyumculuk Sektöründe Bir Uygulama”, (Yüksek Lisans Tezi), Fen Bilimleri Enstitüsü, İstanbul Üniversitesi.

GÜNER, E., S. (2015). “Makine Çevirisinde Yeni Bir Bilgisayımsal Yaklaşım”, (Doktora Tezi), Fen Bilimleri Enstitüsü, Trakya Üniversitesi.

ÖZGEÇMİŞ

Ad-Soyad **:** İlkay YELMEN

ÖĞRENİM DURUMU:

- **Lisans :** 2013, İstanbul Aydın Üniversitesi, Mühendislik Mimarlık Fakültesi, Yazılım Mühendisliği Bölümü
- **Yüksek Lisans :** 2016, İstanbul Aydın Üniversitesi, Bilgisayar Mühendisliği A.B.D., Bilgisayar Mühendisliği Programı

MESLEKİ DENEYİM VE ÖDÜLLER:

11.2022 – Devam : Turkcell Dijital Eğt. Teknolojileri, Eğitim Teknolojileri Direktörü
04.2022 – 11.2022 : Turkcell, Teknik Lider
08.2020 – 04.2022 : Turkcell, Kıdemli Teknik Ürün Yöneticisi
12.2018 – 02.2020 : Atlasglobal, Proje Yönetim Ofisi ve Ar-Ge Merkezi Yöneticisi
01.2015 – 12.2018 : Türk Hava Yolları A.O., İş Analisti, Proje Yöneticisi
09.2013 – 06.2014 : Özyeğin Üniversitesi, Araştırma Görevlisi

TEZDEN TÜRETİLEN YAYINLAR, SUNUMLAR VE PATENTLER:

- Yelmen, I., Gunes, A. ve Zontul, M. (2023). "Multi-Class Document Classification Using Lexical Ontology-Based Deep Learning", Applied Sciences, 13(10), 6139.
- Yelmen, I., Gunes, A., Zontul, M. ve Aslan, Z. (2022). "Multi-class document classification based on deep neural network and Word2Vec", Journal of Aeronautics and Space Technologies, 15(1), 59-65.

DİĞER YAYINLAR, SUNUMLAR, PATENTLER:

- Kiani, F., Randazzo, G., Yelmen, I., Seyyedabbasi, A., Nematzadeh, S., Anka, F. A., Erenel, F., Zontul, M., Lanza, S. ve Muzirafuti, A. (2022). "A smart and mechanized agricultural application: From cultivation to harvest", Applied Sciences, 12(12), 6021.

- Caloglu, Z. V., Zontul, M., Yelmen, I. ve Bagriyanik, S. (2021) “Software Quality Measurement Modelling Using AHP and Promethee Methods”, In 2021 6th International Conference on Computer Science and Engineering (UBMK) pp. 608-612.
- Asarkaya, S., Kaynar, O., Yelmen, İ., Yıldırım, F. ve Zontul, M. (2021). “DDOS Saldırılarının Makine Öğrenimi Algoritmalarıyla Tespiti”, Tasarım Mimarlık ve Mühendislik Dergisi, vol. 1(3), pp. 221-232.
- Üstebay, S. Yelmen, I. ve Zontul, M. (2020) “Customer Segmentation Based on Self-Organizing Map: A Case Study on Airline Passengers”, Journal of Aeronautics and Space Technologies, Vol.13, No.2, pp.227-233.
- Ersan, Z. G., Zontul, M. ve Yelmen, I. (2020) “Map Matching with Kalman Filter and Location Estimation. Cumhuriyet Science Journal”, Vol.41, No.1, pp.43-48.
- Demir, V., Zontul, M. ve Yelmen, I. (2020) “Drug Sales Prediction with ACF and PACF Supported ARIMA Method”, In 2020 5th International Conference on Computer Science and Engineering (UBMK) pp. 243-247.
- Yelmen, I., Zontul M., Kaynar, O. ve Sönmez, F. (2018). “A Novel Hybrid Approach for Sentiment Classification of Turkish Tweets for GSM Operators”, Internatıonal Journal of Circuits, Systems And Signal Processing Vol. 12 pp.637-645.
- Yelmen, I. ve Zontul, M. (2016). “Sentiment Analysis Tool for Daily Speech Turkish Texts”, International Journal of Research in Engineering and Technology (IJRET) Vol. 4, No. 1.
- Yelmen I ve Zontul M. (2016) “Determining The Core Part of Software Development Curriculum Applying Association Rule Mining on Software Job Ads In Turkey”, Fourth International Conference on Data Mining & Knowledge Management Process (DKMP 2016)
- Yelmen, I. ve Tosyalıoğlu, N. (2013). “Evaluation of the University Students’ Opinions on Environmental Awareness Using Data Mining Association Rule”, International Journal of Electronics, Mechanical and Mechatronics Engineering, Vol.2, Num.4, pp.384-390.