DUYGU ANALİZİ PROBLEMİ İÇİN GELENEKSEL VE DERİN TRANSFER MAKİNE ÖĞRENMESİ YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Nuray YILDIZLI

Yüksek Lisans Tezi
Bilgisayar Mühendisliği Ana Bilim Dalı
Dr. Öğr. Üyesi Gülşah TÜMÜKLÜ ÖZYER
2021

T.C.
ATATÜRK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI

DUYGU ANALİZİ PROBLEMİ İÇİN GELENEKSEL VE DERİN TRANSFER MAKİNE ÖĞRENMESİ YÖNTEMLERİNİN KARŞILAŞTIRILMASI

(A Comparison of Traditional and Deep Transfer Machine Learning Methods for the Sentiment Analysis Problem)

YÜKSEK LİSANS TEZİ

Nuray YILDIZLI

Danışman: Dr. Öğr. Üyesi Gülşah TÜMÜKLÜ ÖZYER

Erzurum

Aralık, 2021

KABUL VE ONAY TUTANAĞI

Nuray YILDIZLI tarafından hazırlanan "DUYGU ANALİZİ PROBLEMİ İÇİN GELENEKSEL VE DERİN TRANSFER MAKİNE ÖĞRENMESİ YÖNTEMLERİNİN KARŞILAŞTIRILMASI " başlıklı çalışması 14 / 12 / 2021 tarihinde yapılan tez savunma sınavı sonucunda başarılı bulunarak jürimiz tarafından Bilgisayar Mühendisliği Ana Bilim Dalı, Bilgisayar Mühendisliği Bilim Dalında Yüksek Lisans tezi olarak kabul edilmiştir.

Jüri Başkanı:	Prof.Dr.Köksal ERENTÜRK *Atatürk Üniversitesi*	Aslı Islak İmzalıdır
Danışman:	Dr.Öğr.Üyesi Gülşah TÜMÜKLÜ ÖZYER *Atatürk Üniversitesi*	Aslı Islak İmzalıdır
Jüri Üyesi:	Dr. Öğr. Üyesi ALEV MUTLU *Kocaeli Üniversitesi*	Aslı Islak İmzalıdır

Bu tezin Atatürk Üniversitesi Lisansüstü Eğitim ve Öğretim Yönetmeliği'nin ilgili maddelerinde belirtilen şartları yerine getirdiğini onaylarım.

Prof. Dr. Saltuk Buğrahan CEYHUN
Enstitü Müdürü
Aslı Islak İmzalıdır

Not: Bu tezde kullanılan özgün ve başka kaynaklardan yapılan bildiriş, Tablo, şekil ve fotoğrafların kaynak olarak kullanımı, 5846 sayılı Fikir ve Sanat Eserleri Kanunundaki hükümlere

ETİK BİLDİRİM VE İNTİHAL BEYAN FORMU

Yüksek Lisans Tezi olarak *Dr. Öğr. Üyesi Gülşah Tümüklü Özyer* danışmanlığında sunulan "DUYGU ANALİZİ PROBLEMİ İÇİN GELENEKSEL VE DERİN TRANSFER MAKİNE ÖĞRENMESİ YÖNTEMLERİNİN KARŞILAŞTIRILMASI" başlıklı çalışmanın tarafımızdan bilimsel etik ilkelere uyularak yazıldığını, yararlanılan eserlerin kaynakçada gösterildiğini, Fen Bilimleri Enstitüsü tarafından belirlenmiş olan Turnitin Programı benzerlik oranlarının aşılmadığını ve aşağıdaki oranlarda olduğunu beyan ederiz.

Tez Bölümleri	**Tezin Benzerlik Oranı (%)**	**Maksimum Oran (%)**
Giriş	*4*	**30**
Kuramsal Temeller	*9*	**30**
Materyal ve Yöntem	*7*	**35**
Bulgular	*5*	**20**
Tartışma	*0*	**20**
Tezin Geneli	*9*	**25**

Not: *Yedi kelimeye kadar benzerlikler ile Başlık, Kaynakça, İçindekiler, Teşekkür, Dizin ve Ekler kısımları tarama dışı bırakılabilir. Yukarıdaki azami benzerlik oranları yanında tek bir kaynaktan olan benzerlik oranlarının* ***%5'den*** *büyük olmaması gerekir.*

Beyan edilen bilgilerin doğru olduğunu, aksi halde doğacak hukuki sorumlulukları kabul ve beyan ederiz.

Tez Yazarı (Öğrenci)	**Tez Danışmanı**
NURAY YILDIZLI	Dr. Öğr. Üyesi Gülşah TÜMÜKLÜ ÖZYER
11.11.2021	**11.11.2021**
İmza: Aslı Islak İmzalıdır	İmza: Aslı Islak İmzalıdır

***** Tez ile ilgili YÖKTEZ'de yayınlamasına ilişkin bir engelleme var ise aşağıdaki alanı doldurunuz.

☐ Tezle ilgili patent başvurusu yapılması / patent alma sürecinin devam etmesi sebebiyle Enstitü Yönetim Kurulunun …./…/…. tarih ve ……. sayılı kararı ile teze erişim 2 (iki) yıl süreyle engellenmiştir.

☐ Enstitü Yönetim Kurulunun …./…/…. tarih ve ……. sayılı kararı ile teze erişim 6 (altı) ay süreyle engellenmiştir.

TEŞEKKÜR

Yüksek Lisans tezimin araştırma konusunda ve çalışmalarım sürecinde yol gösteren danışman hocam Sayın Dr. Öğr. Üyesi Gülşah TÜMÜKLÜ ÖZYER' e teşekkür ederim.

Tezimin analizlerinde kullandığımız ve 360 Yorumlar veri seti olarak adlandırdığımız veri seti için Dr. Öğr. Üyesi Barış ÖZYER hocamıza ve Vektora firmasından Betül Bulut ve Semih Çelik'e teşekkür ederim. Bu veri seti TÜBİTAK Kobi Ar-Ge başlangıç destek programı kapsamında "Firmaların insan kaynakları departmanlarında kullanılmak üzere, bulut tabanlı, metin işleme yapabilen 360 derece performans değerlendirme yazılımı geliştirilmesi" isimli ve 7180750 numaralı projeden elde edilmiştir.

Tezimin başından beri bana destek olan Aras Elektrik Dağıtım A.Ş. Bilgi Teknolojileri Koordinatörü Sayın Ali ARSLAN'a ve Bilgi Teknolojileri Müdürü Sayın Sinan KAHRAMAN'a teşekkürlerimi sunuyorum.

Hiçbir zaman benden desteklerini esirgemeyen ve hayatımın her evresinde yanımda olan değerli aileme şükranlarımı sunuyorum, iyi ki varsınız.

Tez çalışmalarım sürecinde inanılmaz derecede destekleyici olan arkadaşlarıma minnettarım.

Nuray YILDIZLI

ÖZET

YÜKSEK LİSANS TEZİ

DUYGU ANALİZİ PROBLEMİ İÇİN GELENEKSEL VE DERİN TRANSFER MAKİNE ÖĞRENMESİ YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Nuray YILDIZLI

Danışman: Dr. Öğr. Üyesi Gülşah TÜMÜKLÜ ÖZYER

Amaç: Elektronik hizmet sektörünün gelişmesi, sosyal medya kullanımının artması ile birlikte kişilerin yorumlarına, düşüncelerine ve duygularına göre hızlı ve doğru analizler yapılarak, kısa zamanda aksiyon planları hazırlamak gerekli bir problem haline gelmiştir. Bu problemin çözümü metin sınıflandırmanın bir alanı olan duygu analizi içerisinde ele alınmıştır. Bu tez çalışmasında e-hizmet sektöründen, sosyal medyadan toplanan veriler için duygu analizi çalışması yapılmıştır. Yapılan çalışmayla amaç Türkçe, İngilizce dillerindeki veri setlerinde geleneksel ve derin transfer makine öğrenme yaklaşımlarının performanslarını incelemektir.

Yöntem: Çalışmamızda doğal dil çıkarım alanlarında kullanılan BERT, RoBERTa ve BART model kullanılmıştır. Geleneksel ve derin transfer makine öğrenme yaklaşımlarının başarısını karşılaştırmadaki doğruluğumuzu artırmak amacıyla her iki yöntem için de farklı modeller kullanılarak sonuçlar elde edilmiştir. Geleneksel makine öğrenmesi için TF-IDF ve Doc2vec modelleri ile LR, SVM, KNN ve NB algoritmaları çalıştırılmıştır. Derin transfer öğrenme yaklaşımları için BERT, RoBERTa ve BART modelleri ile sıfır atış algoritması çalıştırılmıştır.

Bulgular: Geleneksel makine öğrenmesi yaklaşımlarının, derin transfer öğrenme yaklaşımlarına göre zaman karmaşıklığı daha düşük olmakla birlikte daha yüksek doğrulukta sonuçlar ürettiği gözlemlenmiştir. Derin transfer öğrenme modelinde BART ve RoBERTa modellerin Türkçe veri setlerinde yüksek doğrulukta başarı oranları ürettiği gözlemlenmiştir.

Sonuç: Bu tez çalışmasında 3 Türkçe 3 İngilizce olmak üzere toplam 6 farklı büyüklükte veri seti kullanılmıştır. Deneysel sonuçlar incelendiğinde geleneksel makine öğrenmesi yaklaşımlarının derin transfer öğrenme yaklaşımlarına oranla daha başarılı olduğu gözlemlenmiştir. Türkçe veri setlerinde sıfır atış algoritmasıyla BART ve RoBERTa modellerinde yüksek doğrulukta başarı oranları elde edilmiştir. Etiket sayısı arttıkça sıfır atış algoritmasında daha düşük doğruluk sonuçları elde edildiği gözlemlenmiştir.

Anahtar Kelimeler: Makine Öğrenme, Derin Öğrenmesi, LR, SVM, KNN, NB, BERT, BART, RoBERTa, Doğal Dil Çıkarımı

Aralık 2021, 68 Sayfa

ABSTRACT

MASTER THESIS

A COMPARISON OF TRADITIONAL AND DEEP TRANSFER MACHINE LEARNING METHODS FOR THE SENTIMENT ANALYSIS PROBLEM

Nuray YILDIZLI

Supervisor: Dr. Öğr. Üyesi Gülşah TÜMÜKLÜ ÖZYER

Purpose: With the development of the electronic service sector and the increase in the use of social media, it has become a necessary problem to prepare action plans in a short time by making fast and accurate analyzes according to the comments, thoughts and feelings of the people. The solution to this problem is discussed in sentiment analysis, which is an area of text classification. In this thesis, a sentiment analysis study was conducted for the data collected from the e-service sector and social media. The aim of the study is to examine the performances of traditional and deep transfer machine learning approaches in data sets in Turkish and English languages.

Method: BERT, ROBERTa and BART models used in natural language inference were used in our study. In order to increase our accuracy in comparing the success of traditional and deep transfer machine learning approaches, results were obtained by using different models for both methods. For traditional machine learning, TF-IDF and Doc2vec models and LR, SVM, KNN and NB algorithms were run. For deep transfer learning approaches, the zero shot algorithm was run with BERT, ROBERTa and BART models.

Findings: It has been observed that traditional machine learning approaches produce results with higher accuracy, although the time complexity is lower than deep transfer learning approaches. In the deep transfer learning model, it has been observed that BART and RoBERTa models produce high accuracy success rates in Turkish data sets.

Results: In this thesis study, a total of 6 different data sets, 3 in Turkish and 3 in English, were used. When the experimental results were examined, it was observed that traditional machine learning approaches were more successful than deep transfer learning approaches. High accuracy success rates were obtained in BART and RoBERTa models with the zero shot algorithm in Turkish data sets. It has been observed that as the number of tags increases, lower accuracy results are obtained in the zero shot algorithm.

Keywords: Machine Learning, Deep Learning, LR, SVM, KNN, NB, BERT, BART, ROBERTa, Natural Language Inference

December 2021, 68 pages

İÇİNDEKİLER

TABLOLAR DİZİNİ

ŞEKİLLER DİZİNİ

KISALTMALAR VE SİMGELER DİZİNİ

Alg : Algoritma

BERT **:** Bidirectional Encoder Representations from Transformers

D2V **:** Doc2vec

DDİ **:** Doğal Dil İşleme

DistilBERT **:** Distillation BERT

DVM **:** Destek Vektör Makinesi

KNN **:** K En Yakın Komşu

LR **:** Lojistik Regrasyon

LSTM **:** Long short-term memory (Uzun Kısa Vadeli Hafıza Ağları)

MLM **:** Masked Language Model (Maskelenmiş Dil Modeli)

NB **:** Naive Bayes

NLI **:** Natural Language Inference (Doğal Dil Çıkarımı)

NSP **:** Next Sentence Prediction (Sonraki Cümle Tahmini)

RoBERTa **:** Robustly Optimized BERT

SAA **:** Sıfır Atış Algoritması

SNLI **:** Standford Natural Language Inference

SVM **:** Support Vector Machine (Destek Vektör Makinesi)

SVSM **:** Sıfır Vuruş Metin Sınıflandırma

TF-IDF **:** Term Frequency -Inverse Document Frequency(Terim Sıklığı-Ters Doküman Sıklığı)

GİRİŞ

Günlük hayatımızda internetin kullanımı oldukça yaygınlaşmıştır. İnternet kullanımının yaygınlaşması teknolojik anlamda rekabeti ön plana çıkarmıştır. Rekabet ortamında firmaların hizmet verdikleri sektörde müşterilerinin yorumlarını, düşüncelerini öğrenmek için web sitelerinde, çağrı merkezlerinde hemen hemen tüm platformlarında ürünler veya durumlar hakkında insanların düşüncelerini almak için belli bir alan açılmaktadır. İnsanlardan alınan düşünceler kayıt altına alınmaktadır ve gittikçe büyüyen bir veri kümesi oluşmaktadır. Verilecek hizmet kalitesinin artırılması ve rekabet ortamında ön plana çıkabilmek için verilerin analiz edilmesi, verilerin içerdiği olumlu veya olumsuz duyguların çıkarılması gerekmektedir. Tüm bu verilerden insanların düşüncelerinin tespit edilmesi duygu analizi problemi olarak tanımlanmaktadır. Duygu analizi duyguların metinlerde hangi yollarla anlatıldığını ve bu anlatımlarda olumlu, olumsuz veya tarafsız durumları tespit etmeyi sağlayan bir analizdir. Duygu analizi fikir madenciliği olarak da ifade edilmektedir. Duygu analizi insanların görüşlerini, değerlendirmelerini ve duygularını hesaplamaya dayanan bir yaklaşımdır. Herhangi bir metne bakıldığında olumlu, olumsuz ya da tarafsız olarak değerlendirilmesi bir çok problemi de beraberinde getirmektedir. Diller arası farklılıklar, bir kelimenin bir cümle içindeki kullanımına göre farklı anlamlara gelmesi, duyguların insanlara göre değişiklik göstermesi, duyguların kişisel olması, yazı dilindeki kültürel farklar duygu analizi çalışmaları için birer problem haline gelebilmektedir. Bu problemler değerlendirildiğinde duygu analizi çalışmasında veri setlerinin özellikleri de önemlidir. Genellikle üzerinde çalışılan veriler insanların resmi olmayan ortamlarda yaptıkları yorumlar olduğundan, duygu analizi problemlerinde doğal dil işleme yöntemleri incelenmiştir. Özellikle metinlerin içerisine eklenen emojiler, hashtagler ifade biçiminin değişmesine yol açabilmektedir. İnsanlardan alınan verilerin makineler tarafından öğrenilmesi ve makinelerin aldıkları veriyi insan gibi algılayarak kategorize etmesi bilgisayar bilimleri için geniş bir çalışma alanıdır. İnsanlar kavramları dil yoluyla rahatlıkla öğrenebilirler ve çok fazla örneklemeye ihtiyaç duymazlar, ancak makineler öğrenmek için çok fazla veriye ihtiyaç duymaktadırlar. İnsan ve makine öğrenimi arasındaki bu farklılık öğrenme gelişmeleri için önemli bir zemin sağlamaktadır ve makinelere bilişsel öğrenme beceresinin nasıl kazandırılabileceği üzerinde çalışılmaktadır (Srivastava *et al.* 2017). Makinelere bu beceri kazandırılırken kullanılan veri setlerinin kalitesi de incelenmesi gereken bir diğer çalışma alanıdır. Duygu analizi veriler yetersiz olduğunda veya görünmeyen sınıflara uyarlanması gerektiğinde zorlanma eğilimindedir. Gerçek hayat uygulamalarında toplanan verilerin kalitesi o veriler üzerinde çalışılabilmesi için önemli bir etkendir. Bu nedenle duygu

analizi çalışmalarında kullanılacak veri setini işleme adımları da önemli bir problemdir. Ön işleme adımının temel amacı, metinsel verilerden örnek kategorilerin birbirinden ayırt edilmesini sağlayabilecek önemli özniteliklerin ortaya çıkarılması, makine öğrenmesi veya derin öğrenme yaklaşımlarının çalıştırılabileceği formata dönüştürülmesidir. Böylelikle verilerin uyumsuz, tutarsız tarafları çıkarılmış olunacaktır. Tüm bu işlemler veri kalitesini artırmak için uygulanmaktadır. Daha kaliteli veri setinde çalıştırılacak algoritmalar daha doğru değerlendirilir. Özellikle derin öğrenme mimarileri, büyük etiketli veri kümeleri gerektirmesiyle ünlüdür. Bu problem, küçük veri setlerinin doğal dil açıklamalarından elde edilen veriler ile desteklenmesi ile hafifletilmektedir (Srivastava *et al.* 2018). Küçük veri setlerinin desteklenmesine yönelik çalışmalar yapılsa da bir yandan da az sayıda veri içeren veri setlerinde çalıştırılan derin transfer öğrenme yaklaşımları da bulunmaktadır. Bu yaklaşımlardan biri olan sıfır atış öğrenme, yalnızca küçük bir sınıf kümesinden gelen etiketli verileri ve sınıf ilişkileri hakkındaki dış bilgileri kullanarak çok sayıda görünmeyen sınıfı tahmin etmeyi amaçlamaktadır. Özellikle görüntü tanıma çalışmalarında bu yöntem kullanılmaktadır. Nesneler için eğitim verisi mevcut olmasa bile görüntülerdeki nesneleri tanıyabilen modeller üzerinde çalışılmıştır (Socher *et al.* 2013). Ayrıca sıfır atış yaklaşımı metin sınıflandırma problemlerinde de kullanılmıştır. Ancak bu çalışmaların eğitim veri setleri oldukça büyüktür (Pushp and Srivastava 2017). Az sayıda veri içeren veri setlerindeki geleneksel makine öğrenme yaklaşımlarının ve derin transfer öğrenme yaklaşımlarının başarısı da merak konusudur. Bu kapsamda derin öğrenme ile geleneksel makine öğrenme yaklaşımlarının performansları da incelenmiştir (Usherwood and Smit 2019). Çalışmalarda farklı veri setleri üzerinde deneyler yapılarak karşılaştırmalar yapılmıştır ve veriler hazırlanırken özellikle zaman maliyetini düşürmek için doğal dil çıkarımlarının olduğu büyük veri setleri de oluşturulmuştur (Bowman *et al.* 2015). Yapılan literatür araştırması sonucunda derin transfer öğrenme yaklaşımlarının uygulandığı, Türkçe veri setiyle çalışıldığı sınırlı sayıda çalışma olduğu gözlemlenmiştir. Bu çalışmalarda İngilizce olan veri setleri Türkçe'ye çevrilerek kullanılmıştır (Budur vd 2020). Ancak Türkçe dili, yapısı gereğiyle anlaşılması zor bir dil olması sebebiyle çeviri ile oluşturulan veri setleri dışında özgün, daha resmi ortamda gramer olarak daha düzgün veri setleriyle ve resmi olmayan ortamlarda toplanan veri setleriyle çalışmak, incelemek istediğimiz bir çalışma alanıdır. Bu tez kapsamında farklı ortamlardan elde edilen Türkçe ve İngilizce veri setleri ile kapsamlı bir karşılaştırma yapılmıştır. Tez de bir firmaya ait personellerin birbiri hakkındaki yorumları ile oluşturulan özgün bir Türkçe veri seti, sosyal medyadan toplanan verilerden oluşan Twitter veri seti, bir web sitesi aracılığıyla kullanıcıların bir telefon markası ile ilgili yapmış oldukları yorumları içeren bir Türkçe veri seti, AG News adı verilen konu sınıflandırma örneği olan İngilizce veri seti, emotion ve IMDB adında yine

kullanıcı yorumlarından oluşan İngilizce veri setleri kullanılmıştır. Kullanılan veri setleri genellikle etiket dağılımı dengesiz veri setleridir. Çalışmamızda, dengeli ve dengesiz dağılım gösteren Türkçe ve İngilizce veri setlerinde BERT, RoBERTa, BART doğal dil çıkarımı modelleriyle çalıştırılan sıfır atış algoritmasının ve geleneksel makine öğrenme yaklaşımlarının performansları incelenerek veri setlerindeki anlamsal bilginin ortaya çıkarılması ve sonuçlarının karşılaştırılması hedeflenmiştir. Özellikle bir kaç etiketin daha yoğun dağılım gösterdiği veri setlerinde çalıştırılan geleneksel makine ve derin transfer öğrenme algoritmalarının performanslarının karşılaştırılması gelecekteki tutarsız veri setleri için önemli bir gelişme olacaktır.

KURAMSAL TEMELLER

Bu bölümde tez kapsamında araştırılan literatür taraması sonucu incelenen konular ve çalışmalar kısaca açıklanmıştır. Tez kapsamında metin sınıflandırma problemlerinden biri olan duygu analizi çalışmaları incelenmiştir. Metin sınıflandırma çalışmalarında kullanılan geleneksel makine öğrenmesi ve derin transfer öğrenme yöntemlerinden bahsedilmiştir. İncelenen yöntemlerde metin sınıflandırma ile ilgili çalışma alanı sunan sosyal medya verileri kullanılmıştır. Tez kapsamında yapılan çalışmalarda, sosyal medya verileri, belli bir ürün, film gibi insanların yapmış oldukları yorumlardan oluşan fikir verileri, haber verileri, bir firmaya ait insanların birbirleri hakkında yapmış oldukları yorum verileri kullanılmıştır. Genellikle insanların resmi olmayan ortamlarda yapmış oldukları yorumları içeren veriler kullanılması sebebiyle doğal dil işleme yöntemleri de incelenmiştir. Son yıllarda özellikle web üzerinde yapılan sorguların iyileştirilmesi için doğal dil işleme yöntemleri makine öğrenme yöntemleri ile birlikte kullanılarak yeni algoritmalar geliştirilmiştir. Tez kapsamında kullanılan Doğal Dil Çıkarımı (Natural Language Inference, NLI) (Bowman *et al.* 2015), BERT (Devlin *et al.* 2019), RoBERTa (Liu *et al.* 2019) ve BART (Lewis *et al.* 2019) modelleri açıklanmıştır. İnsanların günlük hayatta kullandıkları doğal dil ile yapmış oldukları yorumlardan oluşan metinlerin içerdiği duygu, fikir bilgisini bulmaya yönelik çalışıldığından duygu analizinden bahsedilmiştir. İncelediğimiz duygu analizi bir metin sınıflandırma çalışma alanı olduğundan metin sınıflandırma problemi, metin sınıflandırma alanında kullanılan veri işleme adımları, sınıflandırma yöntemleri, literatürde yapılan çalışmalar açıklanmıştır.

Bölüm akışı; ilk paragrafta açıklandığı gibi doğal dil işleme, metin sınıflandırma ve literatür taraması ana başlıklarından oluşmaktadır.

Doğal Dil İşleme

Günümüzde teknolojinin gelişmesiyle birlikte insanlar, dijital ortamlarda birçok uygulama kullanarak kolay ve hızlı bir şekilde çoğu ihtiyaçlarını karşılayabilmektedir. Alışveriş, eğitim, spor, psikoloji gibi birçok alanda çevrim içi ortamda insanlara hizmet verilmektedir. Özellikle akıllı telefonlarda bulunan kişisel asistan uygulamaları ile plan program hazırlama, farklı dillere çeviri yapabilme, metin seslendirme, sesli olarak konuşup komut verilip arama ve mesaj atma gibi birçok işlem yapılabilmektedir. Bütün bu uygulamalar doğal dil işleme (DDİ) çalışma kapsamını oluşturmaktadır. DDİ, doğal dillerin yapısının çözümlenerek bir bilgisayarın anlayabilmesi ve işleyebilmesi amacını taşır. DDİ çalışmaları daha çok yapay zeka altındaki dil bilim bilgisine dayalı çalışmaları kapsamaktadır. Metin

anlama ve sınıflandırma, bir metnin özetinin çıkarılması gibi metin madenciliği çalışmalarında da DDİ kullanılarak özellik çıkarımı yapılmaktadır (Şeker 2015). DDİ yukarıda da örnekleri verildiği gibi bir çok çalışma alanında kullanılmaktadır. Tez kapsamında DDİ yönelik incelenen NLI, BERT, RoBERTa, BART modellerinden ve sıfır atış algoritmasından bahsedilecektir.

Doğal dil çıkarımı (Natural Language Inference)

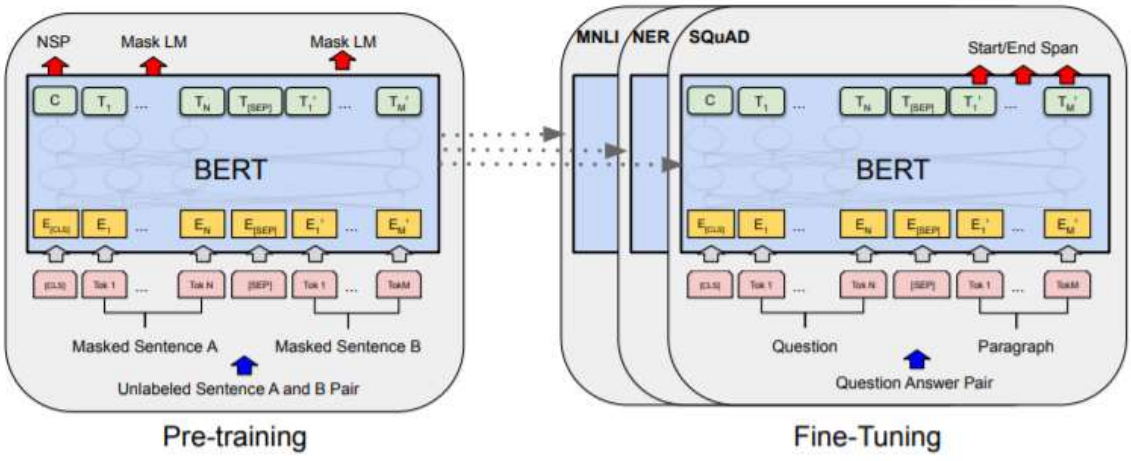
Doğal dil çıkarımı, öncül verilen hipotezin doğru (gerektirme), yanlış (çelişki) veya belirsiz (nötr) olup olmadığını belirleme görevidir. Doğal dili anlamak için dildeki çıkarımı yapabilmek gerekmektedir. Dildeki mantık ve çelişkinin anlamsal kavramları, sözlükten tüm metinlerin içeriğine kadar doğal dil anlamının tüm yönlerinin merkezinde yer alır. Dolayısıyla, doğal dil çıkarımı bu ilişkileri hesaplama sistemlerinde karakterize ederek kullanmaya, bilgi erişiminden anlamsal çözümlemeye ve sağduyu muhakemesine kadar değişen görevlerde gereklidir. Doğal dil çıkarımı, sembolik mantık, bilgi tabanları ve sinir ağlarına dayalı olanlar dahil olmak üzere çeşitli teknikler kullanılarak ele alınmıştır. Son yıllarda, dağıtılmış kelime ve kelime öbeği temsillerini kullanan yaklaşımlar için önemli bir test alanı haline gelmiştir (Bowman *et al.* 2015).

BERT algoritması

"Bidirectional Encoder Representations from Transformers" ifadesinin baş harflerinden oluşan, Transformers'ın Çift Yönlü Kodlayıcı Temsilleri adı verilen BERT algoritması; Google tarafından geliştirilen, DDİ için kullanılan eğitilmiş sinir ağına dayalı bir algoritmadır. BERT modelde cümleler hem sağdan sola hem de soldan sağa doğru değerlendirilmektedir. Cümledeki kelimeleri tek tek değil, sağında ve solunda bulunan diğer tüm kelimelerle ilişkili olarak işleyen bir modeldir. Bu nedenle BERT modelleri, bir kelimenin anlamını, o kelimeden önce ve sonra gelen kelimelere bakarak daha iyi anlayabilmektedir. Ayrıca cümle içerisinde bulunan edat ve bağlaçların kelimelerle ilişkisinin yorumlanması da BERT modelde sağlanmaktadır. 2019 'dan günümüze BERT model 70'den fazla dil için yayınlanmıştır.

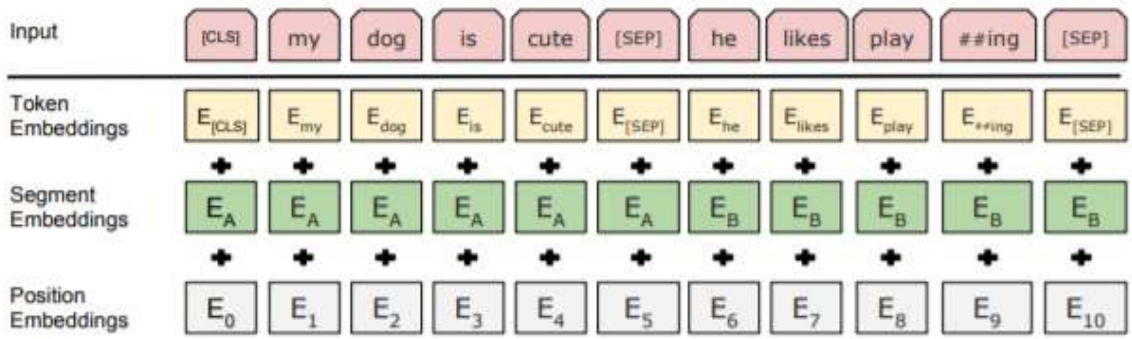
Çalışmada 800 milyon kelime haznesine sahip olan BooksCorpus ve 2500 milyon kelime haznesine sahip Wikipedia veri setleri kullanılarak, BERT_BASE ve BERT_LARGE olarak adlandırılan 2 model sunmuşlardır. BERT_BASE eğitiminde 4 adet TPU, BERT_LARGE eğitiminde 16 adet TPU kullanmışlardır. Her bir ön eğitimin tamamlanması 4 gün sürmüştür. BERT modelini, Masked Language Modeling (MLM) ve Next Sentence Prediction (NSP) adı verilen iki teknikle eğitmişlerdir. Bir cümle modele girdiğinde, cümledeki kelimelerin %15'inde MLM tekniği kullanmışlardır. Bu tekniğin kullanıldığı kelimelerin %80'i "[MASK]" işareti ile, %10'unu rastgele başka bir kelimeyle değiştirmişler, geri kalan %10'

unuda değiştirmeden bırakmışlardır. Çok fazla kelimeyi maskelemenin eğitimi çok zorlaştırdığı, çok az kelimeyi maskelemenin de cümledeki içeriğin çok iyi kavranamama durumuna sebep olduğundan, bu nedenle %15 değerinin seçildiğini belirtmişlerdir (Devlin *et al.* 2019). MLM tekniğinde sadece maskelenen kelimelerin tahmin edilmeye çalışıldığını, açık olan veya üzerinde işlem uygulanmayan kelimelerle ilgili herhangi bir tahminde bulunulmadığını ifade etmişlerdir. Bu sebeple kayıp(loss) değerini sadece işlem uygulanan kelimeler üzerinden değerlendirmişlerdir. MLM de, cümle içerisindeki kelimeler arasındaki ilişki üzerinde durulurken, ikinci teknik olan NSP'de ise cümleler arasında ilişki kurmuşlardır. Eğitim esnasında ikili olarak gelen cümle çiftinde, ikinci cümlenin ilk cümlenin devamı olup olmadığı tahmin edilmiştir. Bu teknikten önce ikinci cümlelerin %50'sini rastgele değiştirmişler, %50'sini ise aynı şekilde bırakmışlardır. Eğitim sırasındaki optimizasyonun, bu iki tekniği kullanırken ortaya çıkan kaybın minimuma indirilmesi olduğunu ifade etmişlerdir. Hem ön eğitimde hem de ince ayarda (fine-tuning) aynı mimarileri kullanmışlardır. Ön eğitim, verilerden bilgi elde edebilmek için kullandıkları aşamadır. İnce ayar ise eğitilmiş modelin farklı problemlerde kullanılması için yapılan ayarlamaları ifade etmektedir. Örneğin ince ayar aşamasında; "[CLS]" sembolünü her giriş örneğinin önüne eklenen özel bir sembol olarak, "[SEP]" sembolünü ise özel bir ayırıcı belirteç olarak kullanmışlardır. BERT algoritmasına ait eğitim ve ince ayar çalışma şekilleri Şekil 1'de gösterilmiştir (Devlin *et al.* 2019).



Şekil 1. BERT ön eğitim ve ince ayar prosedürü (Devlin *et al.* 2019)

Şekil 2'de modelin girdi gösterimi gösterilmiştir. Girdi gösterimini örnek ile açıklayacak olursak örneğin; giriş verisi "Hello I am a [MASK] model." 'dir. Bu cümleyi "[CLS] hello i'm a fashion model. [SEP]" olarak ifade etmişlerdir. Cümledeki "[MASK]" işareti yerine "fashion" ifadesi kullanmışlardır. Böylece cümledeki "[MASK]" tahmin edilmeye çalışılmıştır. Birden fazla cümle içeren veriyi "[CLS] Sentence A [SEP] Sentence B [SEP]" şeklinde ifade etmişlerdir.



Şekil 2. BERT girdi gösterimi (Devlin *et al.* 2019)

Çalışmamızda DistilBERT modeli kullanılmıştır. DistilBERT damıtılmış BERT olarak ifade edilmektedir. BERT' in dil anlama yeteneklerinin % 97' sini korurken, %40 daha küçük ve % 60 daha hızlı olduğu BERT versiyonudur (Sanh *et al.* 2020).

RoBERTa algoritması

"A Robustly Optimized BERT Pretraining Approach" ifadesinin baş harflerinden oluşan, RoBERTa; geniş veri kümesi üzerinde eğitilmiş bir modeldir. BERT modelinden farkları;

- Modelin daha büyük veri setleri üzerinde daha büyük gruplarla daha uzun süre eğitilmesi
- Sonraki cümle tahmin (NSP) adımının kaldırılması
- Daha uzun diziler üzerinde eğitilmesi
- Eğitim verilerine uygulanan maskeleme yönteminin dinamik olması 'dır.

Daha büyük veri kümeleri üzerinde BERT eğitiminin, performansını büyük ölçüde iyileştirdiğini gözlemlemişlerdir. Bu nedenle RoBERTa, 160 GB'ın üzerinde sıkıştırılmamış metin içeren geniş bir veri kümesi üzerinde eğitmişlerdir. BookCorpus ve Wikipedia verileri BERT modelin eğitildiği verilerdi. RoBERTa, CommonCrawl News Data'nın İngilizce bölümünden toplanan Eylül 2016 ile Şubat 2019 arasında taranan 63 milyon İngilizce haber makalesi içeren 76 GB boyutunda CC-News veri seti, OpenAI GPT'yi eğitmek için kullanılan WebText veri kümesinin açık kaynak versiyonu olan 38 GB boyutunda olan OpenWebText veri seti, 16 GB boyutunda BookCorpus ve Wikipedia veri seti, Winograd şemalarının hikaye benzeri stiliyle eşleşmesi için filtrelenmiş CommonCrawl verilerinin bir alt kümesi olan 31 GB boyutundaki hikayeler veri seti üzerinde eğittiklerini ifade etmişlerdir (Liu *et al.* 2019).

BERT modelin ön eğitimindeki maskeli dil modelleme hedefi, her dizide birkaç kelimeyi rastgele maskelemek ve ardından maskelenen kelimeleri tahmin etmekti. Ancak, BERT' in orijinal uygulamasında, diziler ön işlemede yalnızca bir kez maskelenmişti. Bu, tüm

eğitim adımlarında aynı dizi için aynı maskeleme modelinin kullanıldığı anlamına gelmektedir. Bunu önlemek için, RoBERTa model de BERT eğitim verilerini 10 kez çoğaltmışlardır, böylece her dizi 10 farklı desende maskelenmiştir. Bu çalışma da, aynı maskeleme desenleri için her diziyi 4 kez eğitmişler, toplamda 40 tur eğitmişlerdir. Modelde bir dizi beslendiğinde bir maskeleme deseni oluşturulduğundan modelin maskeleme işlemini dinamik maskeleme olarak tanıtmışlardır. Bu değişiklikler ile RoBERTa algoritmasının, BERT algoritması ile hemen hemen aynı sonuçlar verdiğini gözlemlemişlerdir. Ancak dinamik maskeleme kullandıklarında sonuçların biraz daha iyi olduğunu, bu nedenle de RoBERTa'da ön eğitim için dinamik maskeleme yaklaşımını benimsediklerini ifade etmişlerdir (Liu *et al.* 2019). Orijinal BERT makalesinde, modelden en iyi sonuçları elde etmek için sonraki cümle tahmini (NSP) görevinin gerekli olduğunu öne sürmüşlerdi. Son çalışmalar, ön eğitimde bu amacın gerekliliğini sorgulamıştır. Ön eğitimde NSP hedefi olmadan nasıl çalışılacağı aşağıdaki maddelerde açıklanmıştır (Liu *et al.* 2019).

- Segment Çifti + NSP: Bu, NSP kaybıyla birlikte BERT'de (Devlin *et al.* 2019) kullanılan orijinal giriş biçimini izler. Her girdinin, her biri birden fazla doğal cümle içeren metinlerdir, Bu metinlerin toplam uzunluğu 512 kelimeden daha az olmalıdır.
- Cümle Çifti + NSP: Her girdi, bir belgenin bitişik bir bölümünden veya ayrı belgelerden örneklenmiş bir çift doğal cümle içerir. Bu girdiler 512 kelimeden önemli ölçüde daha kısa olduğundan, toplam kelime sayısının Segment çifti + NSP' ye benzer kalması için dizi boyutu artırılıyor. NSP kaybı korunuyor.
- Tam Cümleler: Her girdi, toplam uzunluk en fazla 512 kelime olacak şekilde, bir veya daha fazla belgeden bitişik olarak örneklenen tam cümlelerle paketlenir. Girdiler belge sınırlarını geçebilir. Bir belgenin sonuna gelindiğinde, bir sonraki belgeden cümleler örneklemeye başlanılıyor ve belgeler arasına fazladan bir ayırıcı belirteç ekleniyor. NSP kaybı ortadan kaldırılıyor.
- Belge Cümleler: Girişler, belge sınırlarını geçmemeleri dışında tam cümleler'e benzer şekilde oluşturulur. Bir belgenin sonuna yakın örneklenen girdiler 512 kelimeden daha kısa olabilir, bu nedenle bu durumlarda tam cümleler ile benzer sayıda toplam kelime sayısını elde edebilmek için dizi boyutu dinamik olarak artırılır. NSP kaybı ortadan kaldırılır.

Sonraki cümle tahminini çalışmasının kaldırılmasının sonuçları iyileştirdiğini gözlemlemişlerdir. Tokenleştirme(simgeleştirme) için; RoBERTa da, BERT'in 30K kelime dağarcığına sahip karakter seviyesi Bayt Çifti Kodlama (BPE)'sının aksine, 50K alt kelime birimi içeren bir kelime dağarcığına sahip bayt seviyesinde Bayt Çifti Kodlama (BPE) şeması

kullanmışlardır. Örneğin; “Hello I'm a <mask> model.” cümlesi alınarak “<mask>” ifadesi yerine, “Hello I'm a fashion model.” cümlesinde olduğu gibi kelimeler getirilerek doğruluk değerleri hesaplanmaktadır. RoBERTa'nın, topluluk modelleri de dahil olmak üzere neredeyse tüm GLUE(The General Language Understanding Evaluation benchmark) görevlerinde son teknolojiden daha iyi performans gösterdiğini gözlemlemişlerdir (Liu *et al.* 2019).

BART algoritması

Diziden diziye (sequence to sequence) modellerin ön eğitimi için gürültü giderici bir kodlayıcı olan BART, Lewis *et al.* (2019) tarafından önerilmiştir. Modeli, bir gürültü işleyici ile metni bozma ve orijinal metni yeniden oluşturma üzerine bir model öğrenerek eğitmişlerdir. BART, çift yönlü kodlayıcı (BERT gibi) ve soldan sağa kod çözücü (GPT gibi) içeren standart bir seq2seq/makine çevirisi mimarisini kullanmaktadır. Ön eğitim görevi, orijinal cümlelerin sırasını rastgele karıştırmayı ve metin aralıklarının tek bir maske belirteci ile değiştirildiği yeni bir doldurma şemasını içermektedir. BART model yazarları çeşitli bozulma şemalarını denemişlerdir. Kullanılan bozulma şemaları aşağıda maddeler halinde verilmiştir.

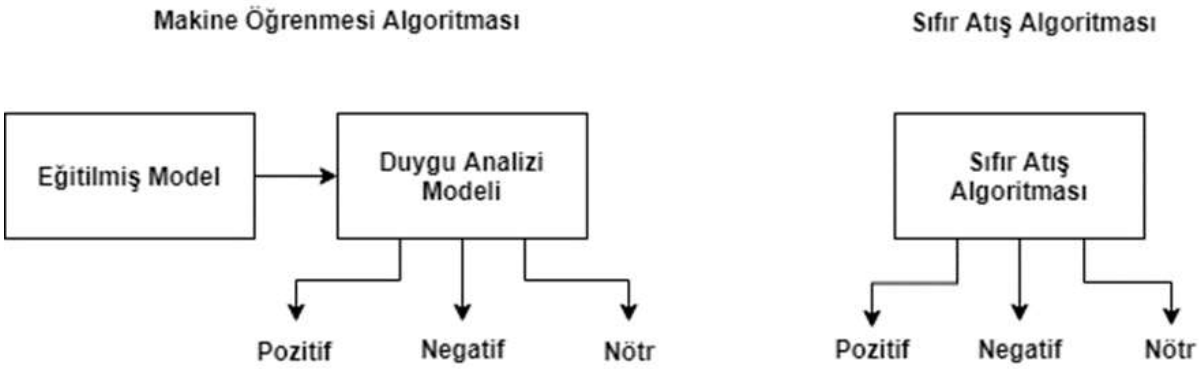
- Token maskeleme: BERT tarafından kullanılan MLM ön eğitim görevidir. Tokenler rastgele bir “<MASK>” simgesi ile değiştirilir ve modelin bu simgeleri tahmin etmesi gerekir.
- Belirteç silme: Belirteçler girdiden rastgele silinir ve model, simgelerin silindiği konumları tahmin etmelidir.
- Metin doldurma: Tek belirteçlerin maskelenmesi yerine, metin aralıkları “<MASK>” simgesiyle değiştirilir. MLM' ye benzer şekilde, model metin aralıklarını tahmin etmelidir.
- Belge döndürme: Rastgele bir belirteç seçilir ve belge bu simgeyle başlaması için döndürülür. Model, belgenin orijinal başlangıcını tahmin etmelidir.
- Cümle karıştırma: Cümleler rastgele sırayla karıştırılır. Model, cümlelerin orijinal sırasını tahmin etmelidir.
- Metin doldurma + Cümle karıştırma birlikte kullanılmıştır.

BART modelin, metin oluşturma için ince ayar yapıldığında özellikle etkili olduğu, ancak anlama görevleri için de iyi çalıştığı, ayırt edici görevlerde RoBERTa'ya benzer bir performans elde ettiğini soyutlayıcı diyalog, soru cevaplama ve özetleme görevlerinde en son teknolojiye sahip yeni sonuçlar elde ettiğini ifade etmişlerdir (Lewis *et al.* 2019).

Sıfır atış algoritması

Derin öğrenme de verilerin birden fazla özellik temsillerinin öğrenilmesine dayanan bir yapı bulunmaktadır. Derin öğrenme sistemi, büyük boyutlu veri setleri ve çoklu katmanları kullanarak, veriyi farklı katmanlarda işleyerek öğrenme işlemini gerçekleştirmektedir. Derin öğrenmede kullanılan modellerde, veri setinin özellikleri makineye belirtilmez, makine verilerin hangi özelliklerine bakacağına kendi karar vererek, veri de hangi fonksiyona ağırlık vereceğini kendi hesaplar. Derin öğrenme uygulamalarında büyük verilerin işlenmesi için güçlü bilgisayar donanımlarına ihtiyaç duyulmaktadır. Son yıllarda maliyeti azaltmak, hızlı analizler yapabilmek için az sayıda etiketli veya etiketsiz veri bulunan görevler için farklı yaklaşımlar kullanılmaktadır. Bu tezde derin transfer öğrenme yöntemleriyle kullanılan sıfır atışlı öğrenme (Zero-shot) algoritmasından bahsedilecektir. Derin öğrenme algoritmaları büyük veri setlerinde çalıştığından, büyük veri setleri oluşturma, veri etiketleme işlemleri özellikle zaman maliyeti düşünüldüğünde zorlu bir süreçtir. Bu nedenle küçük veri setlerinin çeşitli yöntemlerle zenginleştirilmesi üzerine çalışmalar yapılmaktadır. Sıfır atış algoritması (SAA); küçük veri setlerinden, yeni veriler oluşturularak veri setini zenginleştirmek için kullanılan derin öğrenme yöntemlerinden biridir. Özellikle görüntü işleme çalışmalarında nesne tanıma sürecinde yüksek başarı elde edebilmek için nesne sınıflarının belirlenmesi gerekmektedir. Nesne sınıfları belirlenirken bu nesnelere ait farklı ortamlardan alınan görüntü örnekleri bulmak sınıflandırma başarısını artıracak bir işlemdir. Ancak her nesne sınıfına ait yeterli sayıda veri bulunamamakla birlikte bazı durumlarda bazı nesne sınıflarına ait hiç veri bulunamamaktadır. SAA, insanlar gibi karar verme ve muhakeme yeteneğine sahip bilgisayarların daha önce hiç görmediği yeni bir sınıfı tanımalarını sağlar. Örneğin, daha önce hiç zebra görmemiş bir insana: "Zebra, ata benzeyen ve vücudunda siyah ve beyaz çizgiler olan bir hayvandır." diye ifade edilirse hayvanat bahçesinde hangi hayvanın zebra olduğunu insan kolayca bulabilir. Bununla birlikte, geleneksel görüntü tanıma algoritmalarında, makinenin "zebra" yı tanımasını istiyorsak, makineyi yeterli ölçekte "zebra" örnekleriyle beslemek gerekmektedir. Zebra örnekleriyle eğitilen bir model diğer türleri tanımaz ancak SAA bunu yapabilmektedir. SAA modelinin en az bir kere öğrenmesi gerekmez. Özellik tanımlarına göre yeni kavramları öğrenebilmektedir. SAA'nın yeni kavramları öğrenebilme özelliği, insan zekasına yakın bir çalışma prensibi olduğunu göstermektedir. SAA'nın bu özelliği, metin sınıflandırma, doğal dil işleme problemlerinde kullanılmaktadır. Son yıllarda dijitalleşme ile birlikte insanlar akıllı cihazlar üzerinden kendi programlarını oluşturup, bir çok ihtiyacını çevrim içi olarak giderebilmektedir. Bizimle doğal dilde etkileşim kuran bilgisayar sistemleri Siri gibi yaygınlaştıkça kullanıcıların makinelere dilde öğretmesine izin verme olasılığı artmaktadır. Dilden öğrenme yeteneği, sınırlı sayıda eğitim verisi olduğunda ya da hiç veri olmadığında

kullanıcıların kişiselleştirilmiş kavramları öğretmesine olanak tanır. SAA ile herhangi bir görev için eğitim olmadan sınıflandırma işlemleri yapılabilmektedir. Şekil 3'de geleneksel makine öğrenme yöntemleri ile sıfır atış öğrenme çalışma prensibi gösterilmiştir.



Şekil 3. Geleneksel makine öğrenmesi ve sıfır atış algoritması

Sıfır atış tanıma sonuçlarını değerlendirmek için kategori başına ortalama en yüksek değere sahip olan sonuçları kullanarak hesaplama yapar. Matematiksel olarak, bir dizi Y sınıfı için N sınıfları ile, sınıf başına ortalama doğruluk Eşitlik 2.1'de gösterildiği gibi verilir.

$$\alpha_y = \frac{1}{N} \sum_{c=1}^{N} \frac{c'de\ doğru\ yapılan\ tahmin\ sayısı}{c'\ deki\ toplam\ örnek\ sayısı} \quad (1)$$

Her bir sınıf için doğruluğu ayrı ayrı hesaplanarak diğer tüm kategoriler arasında ortalaması alınır. Bu, hem seyrek hem de yoğun sayılı sınıflarda yüksek performansı teşvik etmektedir. Genelleştirilmiş bir sıfır atış ayarı yönteminde amaç, hem görünen sınıflarda hem de görünmeyen bir dizi kategoride yüksek doğruluk sağlamaktır. Böylece SAA performans metriği, görünen sınıflar ve gözlemlenmeyen kategorilerdeki performansın harmonik ortalaması olarak tanımlanır.

Metin Sınıflandırma

Mevcut bir metnin önceden belirlenen sınıflardan hangisine ya da hangilerine dahil edileceğinin belirlendiği metin sınıflandırma işlemi en temelinde, T={t1,t2,…,tn}kümesinde yer alan her bir metin ya da belgenin, önceden tanımlanmış olan C={c1,c2,…,cm} kümesindeki sınıflara ait olup olmadığının belirlenmesi şeklinde gerçekleştirilmektedir (Tantuğ 2012). Bu işleme göre metinlerin otomatik analiz edilmesi ve metinlerin içerdiği duyguların çıkarımı sağlanabilmektedir. Sınıflandırma işlemleri yapılırken uygulanan yöntemler farklılık gösterebilmektedir. Ancak temelde metin sınıflandırma veri toplama, ön işleme, özellik çıkarımı, öğrenme algoritması ve yapılandırmış veri adımlarından oluşmaktadır. Süreç Şekil 4'de gösterilmiştir.

Şekil 4. Metin sınıflandırma adımları

Devam eden alt bölümde; bu tezde kullanılan ve metin sınıflandırma çalışma alanı olan duygu analizinden, metin önişleme adımından ve öğrenme algoritmalarından bahsedilecektir.

Duygu analizi

Metin madenciliği, yapılandırılmamış metinlerden otomatik olarak faydalı bilgilerin çıkarılmasını sağlayan bir çalışma alanıdır. Bu amaçla kullanılan metin madenciliği yöntemleri ile büyük miktardaki metin verileri, kısa zamanda ve yüksek performans ile analiz edilebilmektedir. Metin madenciliği alanlarından biri olan duygu analizi veya fikir madenciliği, görüşler, tutumlar ve duygular gibi öznel bilgilerin algılanmasını otomatikleştirmek için kullanılmaktadır. Teknolojinin gelişmesiyle birlikte dijital ortamda rekabetin artmasıyla insanlardan alınan verilerin olumlu, olumsuz veya tarafsız olarak ayrıştırılması ve analiz sonucunda ihtiyaçların belirlenerek bu yönde iyileştirmeler yapılması duygu analizini önemli bir konu haline getirmiştir. Yapılan duygu analizleri makineler tarafından hızlı, tarafsız ve maliyetsiz olarak yapılmaktadır. Sosyal medyanın aktif olarak kullanılması insanların günlük hayatta kullanmış oldukları dilde farklılıklar oluşturmaktadır. Alınan veriler kelime kısaltmalarını, emojileri, hashtag gibi kelime gruplarını, noktalama işaretlerini içermektedir. Dildeki farklılıklar duygu analizi yöntemleri için kapsamlı çalışma alanı oluşturmaktadır. Özellikle Türkçe gibi gramer yapısı zor olan dillerde analiz yapmak önemli bir problemdir.

İncelediğimiz duygu analizi çalışması metin sınıflandırma problemi olduğundan bir sonraki bölümde; metin sınıflandırmada kullanılan veri işleme adımları, sınıflandırma yöntemleri açıklanmıştır.

Metin ön işleme adımları

Metinler genellikle resmi olmayan ortamlarda insanların dil bilgisi kurallarına dikkat etmeden, metnin içerisinde günümüzde çok kullanılan emojilerin yer aldığı, noktalama işaretlerinin bulunduğu verilerden oluşmaktadır. Bu nedenle metinlerden anlam çıkarılabilmesi, makine veya derin öğrenme algoritmalarının uygulanabilmesi için bazı işlemlerden geçmesi gerekmektedir. Metin ön işleme temel adımları aşağıda kısaca açıklanmıştır.

- Metinlerin cümle, harf, kelime veya n-gramlara parçalanması (Tokenization)
- Metin içerisinde bulunan, anlamda değişiklik yapmayan ile, bazı, da, ve ya gibi durağan kelimelerin çıkarılması (Stopwords)

- Metin içerisindeki noktalama işaretlerinin, emojilerin, özel karakterlerin veya sayıların çıkarılması
- Metindeki büyük küçük harf ayrımının ortadan kaldırılması (Normalization)
- Metinde geçen kelimelerden eklerin kaldırılarak kelime köklerinin kalması (Stemming)

Genellikle öğrenme algoritmaları verilere uygulanmadan önce veriler bahsedilen ön işlemlerden geçirilmektedir.

Özellik çıkarımı

Verilerin bilgisayarlar tarafından anlaşılabilmesi için sayılar ile ifade edilmesi gerekmektedir. Geleneksel yöntemlerde analiz yapılabilmesi için metinlerin vektörlere çevrilmesi temel bir adımdır. Tez kapsamında makine öğrenmesi algoritmasında verinin vektörel olarak gösteriminde 2 farklı yöntem kullanılmıştır. Bu yöntemler TF-IDF ve Doc2vec' dir. TF-IDF, Terim Frekansı- Ters Doküman Frekansı (term frequency-inverse document frequency) baş harflerini içeren, dokümanlardan oluşan bir veri kümesindeki kelimelerin, ilgili dokümanlarla ne kadar ilişkili olduğunu belirten bir ölçüm yöntemidir. Metin sınıflandırma da kullanılan bir yöntem olan TF-IDF ölçümünde bir kelimenin dokümanda geçme sıklığı arttıkça TF değeri artmaya başlar ancak kelimeyi içeren doküman sayısı arttığında IDF değeri bir dengeleme oluşturur. Böylece her dokümanda geçmesi muhtemel olan kelimeler (örnek, neden, için, eğer vb.) ilgili bir dokümanda sıkça geçmelerine rağmen çok sayıda doküman tarafından kapsandığı sebebiyle yüksek TF-IDF değerine sahip olamazlar. Çünkü herhangi bir doküman için özel bir anlam ifade etmezler. Bununla beraber örneğin bir kelime (Ay, Gezegen vb.) bir dokümanda pek çok kez geçiyor ancak veri kümesinde çokça doküman tarafından kapsanmıyorsa, bu kelimenin ilgili doküman için önemli olduğu söylenebilir ve kelime TF-IDF'in yapısından ötürü yüksek TF-IDF değerine sahip olur. TF-IDF, bir dokümandaki her kelimenin dokümanla ne kadar alakalı olduğunu gösteren bir sayısal ölçümdür. Eşitlik 2, 3, 4'de gösterilmiştir.

$$TF = \frac{Seçili\ Terim}{Metin\ İçinde\ Bulunan\ Toplam\ Terim\ Sayısı} \tag{2}$$

$$IDF = \log\left(\frac{Toplam\ Metin\ Adeti}{Terimi\ içeren\ Metin\ Adeti}\right) \tag{3}$$

$$TF - IDF = TF \times IDF \tag{4}$$

Diğer yöntem olan Doc2vec(D2V), Le and Mikolov (2014) tarafından önerilen dokümanların yönetilebilir boyutlu vektörler ile ifade edilmesini sağlayan yapay sinir ağlarına dayalı bir metottur. Bu metot hedef kelimeyi tahmin etmek için sadece kelimeler kullanmak

yerine, belgeye özgü vektör eklemesi yaparak çalışmaktadır. Doc2Vec, hassas söz dizimsel ve anlamsal kelime ilişkilerini yakalayan verimli ve yüksek kaliteli dağıtılmış vektörler üretir. Doc2Vec algoritması içerisinde değiştirilmiş CBOW algoritması, Doküman Vektörlerinin Dağıtılmış Bellek Modeli(Distributed Memory Model of Paragraph Vectors – PV-DM) olarak isimlendirilir. Eğitim süresi boyunca paragraf vektörü yerel içerik çerçevesine eklenir. Bu çerçeve içerisinde herhangi bir kelime tarafından kullanılabilir. Bu işlemin yapılmasının sebebi, zamanla paragraf vektörünün dokümandaki tahminlerin çoğuna katkıda bulunabilecek bir bağlam vektörüne dönüşmesidir. Paragraf vektörü kelimeler arasında sağlanamayan bağlamı tutabilmek için tasarlanmış bir vektördür. Bu vektör mevcut bağlamda neyin eksik olduğunu hatırlayan bir bellek görevi görmektedir. DBOW modelinde ise; PV-DM modelinin aksine buradaki paragraf vektörü dokümandaki tüm kelimeler için bir bağlam oluşturacaktır. Tüm yerel içerik çerçevelerini tahmin edebilmek amacıyla kullanılacaktır (Kınık 2020). Çalışma kapsamımızda PV-DBOW modeli kullanılmıştır.

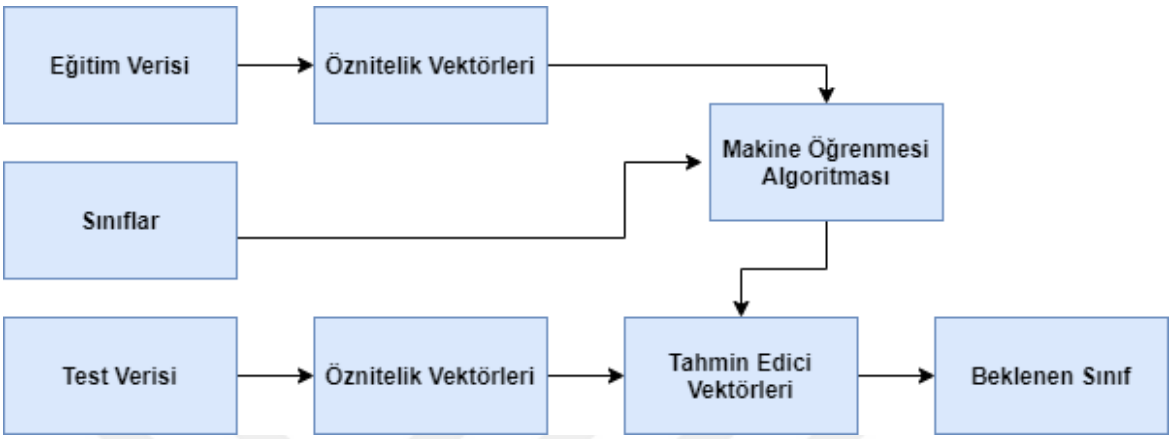
Makine öğrenmesi

Makine öğrenmesi, bir bilgisayarın doğrudan komutlar olmadan insan gibi öğrenme işlemini gerçekleştirmek amacıyla matematiksel modellemeyi kullanan bilgisayar bilimidir. Yapay zekanın bir alt kümesidir. Yapay zeka, insan zekasını taklit eden sistemler anlamına gelen kapsamlı bir terimdir. Makine öğrenimi de yapay zeka çözümlerinden biridir. Makine öğrenmesi pratik yaşamdaki çözülebilir sorunların ele alınmasında, yapay zekanın insan zekasını yakalayacak şekilde gelişmesinde kullanılmaktadır. Makine öğrenmesi algoritmaları metin, sayısal veri veya görüntü verileri üzerinden öğrenme işlemini gerçekleştirebilmektedir. Özellikle tezde de kullandığımız metin verileri üzerinde çalışılan metin sınıflandırma çalışmaları da makine öğrenmesi için önemlidir. Makine öğrenmesi bilinen özelliklerden faydalanılarak öğrenilen verilerden yapılan tahminlere dayanan bir yaklaşımdır. Metin sınıflandırma ise verilerdeki bilinmeyen bilgileri ortaya çıkarmayı, verilerin önceden tanımlanmış sınıflardan hangisine ait olup olmadığının belirlenmesini hedefleyen bir çalışma alanıdır. Bu sınıflandırma işlemi yapılırken makine öğrenmesi algoritmaları kullanılmaktadır. Makine öğrenmesi algoritmaları denetimli ve denetimsiz öğrenme olarak iki ana başlık altında incelenmektedir.

Denetimli öğrenme

Modelin etiketli bir veri kümesi üzerinde eğitilerek öğrenmenin sağlandığı makine öğrenmesi yöntemidir. Sistem eğitilirken kullanılan veri kümesinde bulunan her bir örneğe ait hem giriş hem de çıkış parametreleri bulunmaktadır. Metin Sınıflandırma çalışmalarında giriş

metnin içeriğini, çıkış ise kategorisini temsil eder. Eğitim veri seti sistemi eğitmek, test veri seti ise sistemin doğrulanması amacıyla kullanılır. Sistemin doğrulanması aşamasında öğrenme algoritması kategorisi bilinmeyen bir test verisine, eğitim verisinde bulunan çıkışlardan herhangi birini atar (Kotsiantis *et al.* 2007). Denetimli öğrenme modeli süreci Şekil 5'te gösterilmiştir.

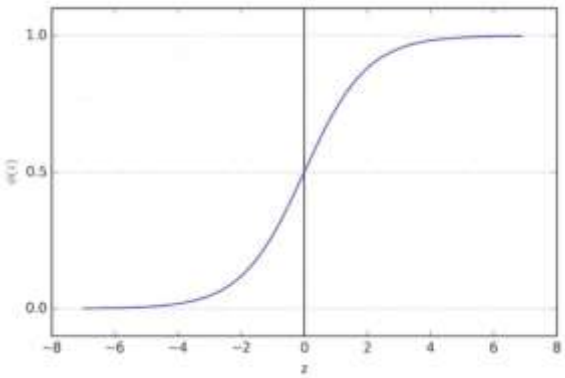


Şekil 5. Denetimli öğrenme modeli

Destek Vektör Makinesi, Yapay Sinir Ağları, Lojistik Regresyon, Naive Bayes, k-En Yakın Komşu, Rastgele Orman ve Karar Ağaçları algoritmaları denetimli öğrenme modeli oluşturulurken yaygın olarak kullanılan yöntemlerdendir. Bu tez de kullanılan 4 denetimli makine öğrenmesi algoritmasından bahsedilecektir. Bu algoritmalar;

- Lojistik Regrasyon
- Destek Vektör Makinesi
- K- en Yakın Komşu
- Naive Bayes 'dır.

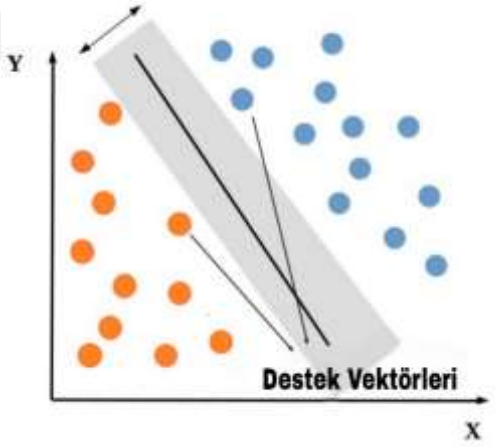
Lojistik Regrasyon (LR): İkili sonuç veren yani sonucun 0 veya 1 olduğu sistemlerde kullanılan modeldir. Örneğin; bir kişinin belirli parametrelere bakılarak hasta olup olmadığı, kredi verilip verilemediği gibi problemlerde kullanılmaktadır. Lojistik modellemede kullanılan fonksiyon Şekil 6 ve Eşitlik 5' de gösterilmiştir. Lojistik fonksiyon (f(x)), -∞/+∞ aralığındaki tüm değerleri girdi olarak kabul edebilir. Çıktı olarak ise 0 ile 1 arasında değerler almaktadır.



Şekil 6. Lojistik regrasyon grafik gösterimi

$$Q(z) = \frac{1}{1 + e^{-z}} \tag{5}$$

Destek vektör makinesi (DVM): Sınıflandırma ve istatistiksel öğrenme alanında ortaya çıkmış veriyi analiz eden denetimli öğrenme modellerindendir. Belirlenen iki kategoriden birine ya da diğerine ait olarak işaretlenmiş bir dizi eğitim örneği verildiğinde, DVM eğitim algoritması, bir olasılık dışı ikili doğrusal sınıflandırıcı haline getirerek bir kategoriye ya da diğerine yeni örnekler atayan bir model oluşturur. Şekil 7'de gösterilmiştir.



Şekil 7. Destek vektör makinesi gösterimi

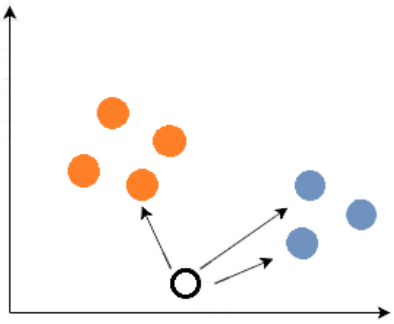
K-en yakın komşu algoritması (KNN): Sınıfları belli olan bir örnek kümedeki verilerden yararlanılarak sınıflandırma ve regrasyon problemlerinde kullanılan algoritmadır. Örnek veri setine katılacak yeni verinin sınıfı belirlenirken, mevcut verilere göre uzaklığı hesaplanıp, k sayıda yakın komşuluğuna bakılarak öznitelik değerlerine göre k komşu veya komşuların sınıfına atanır. Verinin komşularına olan uzaklığı hesaplanırken genellikle Öklit (Euclidean), Manhattan, Minkowski uzaklık formülleri kullanılmaktadır. Herhangi iki nokta P ve Q olsun. P = (x1, x2, x3,….,xn) ve Q = (y1, y2, y3…..yn) olmak üzere eşitlik 6 Öklit, eşitlik 7 Manhattan ve eşitlik 8 Minkowski hesaplama metodlarını göstermektedir. Şekil 8'de KNN algoritması gösterilmiştir. Çalışmamızda KNN algoritması uygulanırken Minkowski mesafe hesabı

kullanılmıştır. Tezde çalıştırılan KNN algoritmasında k değeri 3, 5 ve 7 seçilerek deneyler yapılmıştır. En başarılı sonuç k değeri 7 alındığında elde edildiğinden çalışmamıza k=7 sonuçları eklenmiştir.

$$D = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{6}$$

$$D = \sum_{i=1}^{n} |x_i - y_i| \tag{7}$$

$$D = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p} \tag{8}$$



Şekil 8. K en yakın komşu gösterimi

Naive Bayes algoritması (NB): Naive Bayes sınıflandırma işleminin temeli Bayes algoritmasına dayanmaktadır. Bayes algoritması 1812 yılında Thomas Bayes tarafından bulunan koşullu olasılık hesaplama formülüdür. Formül Eşitlik 2.9'da gösterilmiştir. Formülde P(A|B); B'nin doğru olduğu bilindiğinde A'nın olma olasılığını, P(B|A) ; A'nın doğru olduğu bilindiğinde B'nin olma olasılığını, P(A); A'nın olma olasılığını, P(B); B'nin olma olasılığını ifade etmektedir. Bayes teoremi, olasılık kuramı içinde incelenen önemli bir konudur. Naive Bayes sınıflandırma algoritması bir veri için her durumun olasılığını hesaplayarak, olasılık değeri en yüksek değere sahip olana göre sınıflandırma işlemini yapar.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{9}$$

Denetimsiz öğrenme

Etiketsiz verilerden herhangi bir eğitim olmadan modelin kendi kendine öğrenme sağladığı makine öğrenmesi yöntemleridir. Sistem girdi verilerinden yapısal özellikler çıkartarak çalışır. Denetimsiz öğrenme, denetimli öğrenmeye göre daha karmaşık

algoritmalarla çalışmaktadır. Denetimsiz öğrenme kümeleme ve ilişkisel kuram konularına yönelik çalışmalarda kullanılmaktadır.

Kümeleme, denetimsiz öğrenmede önemli bir alandır. Kategorize edilmemiş verilerden oluşan bir problemde veriler işlenerek doğal gruplar bulunur. Grupların özellikleri belirlenerek ayarlamalar yapılabilmektedir. Örneğin; bir web sitesinde gelen ziyaretçilerin cinsiyet ayrımına göre hangi yazıları okudukları gibi problemlerde kullanılmaktadır. İlişkisel kuram tabanlı denetimsiz öğrenme algoritmalarında ise verilerin birbirleriyle olan birliktelik ilişkileri bulunarak kurallar oluşturup bir sınıflandırma yapılır. Örneğin; bir ürünü alan kullanıcıların yanında aldıkları ürünlerin yoğunluğuna göre diğer kullanıcılara ürün tavsiye edilmesi gibi problemlerde kullanılmaktadır.

Metin Sınıflandırma Probleminde Makine ve Derin Transfer Öğrenme Literatür Taraması

Duygu analizi bir metnin olumlu, olumsuz veya tarafsız olarak sınıflandırılması yaklaşımıdır. Teknolojinin gelişmesiyle birlikte dijitalleşen dünyamızda internet ortamında kullanılan uygulama sayısının artması, insanların kısa sürede çevrim içi ortamda hemen hemen tüm ihtiyaçlarını karşılayabilecek düzeye gelmesini sağlamıştır. Bu durum da beraberinde özellikle başta duygu analizi olmak üzere metin sınıflandırma çalışma alanlarını önemli hale getirmiştir. Bu başlık altında literatür taraması sonucunda incelenen duygu analizi konusunda yapılmış akademik çalışmalara yer verilmiştir.

Geçmiş yıllarda yapılan araştırmalarda çeşitli yöntemler kullanılarak sınıflandırma işlemi yapılmıştır. Çalışmaların birçoğunda etiketsiz ve iyi miktarda veri içeren sosyal medya veri setleri kullanılmıştır. Çoban ve Özyer (2018), Türkçe ve İngilizce Twitter mesajlarından oluşan iki farklı veri seti kullanarak terim ağırlıklandırma yöntemlerini incelemişlerdir. Geleneksel ve yeni yaklaşımları içeren çalışmaları sonucunda PTW yöntemi ile çalışılan DVM algoritmasıyla %97.6 sınıflandırma başarısı elde etmişlerdir. Yeşilyurt ve Şeker(2017), 15 ayrı emoji grubu oluşturularak her grup için Twitter üzerinden veri toplamışlardır. Veri seti üzerinde Random Forest, Random Tree, Gradient Boosted Trees, Naive Bayes, Naive Bayes Kernel algoritmaları çalıştırmışlar ve %52.09 ile en iyi başarıyı Naive Bayes Kernel algoritmasıyla sağlamışlardır. Usherwood and Smit (2019), derin öğrenmeyi temsil eden BERT ve ULMFIT algoritmaları ile DVM, Naive Bayes makine öğrenme algoritmalarını Amazon ve Twitter veri setleri üzerinde çalıştırarak sonuçları karşılaştırmışlardır. 100, 300 ve 1 000 örnekten oluşan veri setleri üzerinden çalışmışlardır. Sonuç olarak BERT algoritmasının 100 örnekte ortalama %9.7, 1000 örnekte %1.8 daha iyi performans gösterdiğini gözlemlemişlerdir. Ancak elde edilen sonuçların hala alt sınırda olduğunu ve bazı uygulamalar için yeterince yüksek kalitede

olmadığını belirtmişlerdir. Geng *et al.* (2019), az sayıda metin sınıflandırılmasını hedefledikleri çalışmalarında birkaç atışlı derin öğrenme algoritması kullanmışlardır. Çalışmada yeni bir sinir modeli yaklaşımı olarak İndüksiyon ağı modelini tanıtmışlardır. Bu modelle yapılan çalışmada ODIC veri setini kullanmışlardır. Sonuç olarak 10 atış uygulamada %88.49 başarı elde etmişlerdir.

Bowman *et al.* (2015), doğal dili anlamak amacıyla, dildeki karmaşıklığın ve çelişkilerin anlaşılmasının önemli olduğunu ve doğal dildeki makine öğrenimi yaklaşımlarının büyük ölçekli veri setlerinin olmaması nedeniyle önemli ölçüde sınırlandırıldığını ifade etmişlerdir. Bu sorununu ele almak için, insanlar tarafından yazılan, etiketli cümle çiftlerinden oluşan ücretsiz Standford Doğal Dil Çıkarımı (Standford Natural Language Inference, SNLI) külliyatını oluşturmuşlardır. Külliyat 570.152 cümle çiftinden oluşmaktadır. Cümle çiftleri "contradiction", "neutral" ve "entailment" kelimeleri ile etiketlenmiştir. Veri seti oluşturulurken Amazon Mechanical Turk uygulamasını kullanmışlar. İnsanlara 160 bin başlıktan oluşan Flickr30k veri setinden yalnızca başlıkları göstererek, başlık hakkında başlığı destekleyen, destekleyebilecek ve başlıkla çelişen 3 farklı cümle yazmaları istenmiştir. Flickr30k başlıkları öncül ve insanların yazdığı cümleler hipotez olarak kullanılmıştır. Cümleleri oluştururlarken herhangi bir kural kısıtlamasına gitmemişlerdir. Sadece birden fazla başlık için aynı cümlelerin kurulması gibi durumlarda insanlar uyarılmıştır. Oluşan veri setinin kalitesini ölçmek için verilerin %10 'una ek bir doğrulama çalışması yapmışlardır. Bu doğrulamada 5 kişiden oluşan gruplara cümle çiftleri sunulmuş ve her bir cümle çifti için 3 etiketten birini seçmeleri istenmiştir. 5 kişiden en az 3'ü tarafından aynı etiket seçilmişse bu etiketi altın etiket olarak işaretlemişlerdir. %2'lik bir oranla fikir birliği olmayan cümle çiftleri çalışmada kullanılmamıştır. %91.2'lik bir oranla seçilen altın etiketin gerçek etiketle eşleştiğini gözlemlemişlerdir. Mevcut veri setlerinden daha geniş bir veri seti olduğu için bu görevde daha önce rekabet etmemiş sinir ağları gibi parametre açısından zengin modelleri eğitmek için uygun olacağını düşünmüşlerdir. SNLI verilerini; kural tabanlı sistemler, basit doğrusal sınıflandırıcılar ve sinir ağı tabanlı modeller dahil olmak üzere çeşitli doğal dil çıkarım modellerinde kullanmışlardır. Bu üç model için SNLI, SICK ve RTE-3 veri setleri kullanılarak performans sonuçlarını karşılaştırmışlardır. SICK; resim ve video alt yazılarından oluşturulan bir veri setidir. RTE-3 ise uzun, karmaşık doğal olarak oluşan metinlerden elde edilen bir veri setidir. Uzaklık tabanlı sınıflandırma performansında SNLI veri seti, SICK ve RTE-3 ten daha iyi sonuç göstermiştir. Sözcüksel sınıflandırıcı için SNLI ve SICK veri setinin hem test hem eğitim sonuçları karşılaştırılmış, önerme ve hipotezlerdeki sözcüklerin örtüşmesine dayalı modelde SNLI'nin daha başarılı olduğunu, önerme ve hipotezdeki kelimelere uygulanan unigram modelde yine SNLI'nin daha başarılı olduğunu, tüm sözcüksel özelliklerden uzak olan

modelde SICK'in daha başarılı olduğunu gözlemlemişlerdir. 3 farklı cümle yerleştirme modelinde de veri setinin performansını ölçmüşlerdir. Her cümledeki kelimelerin gömülmelerini içeren temel bir modelde %75.3, tekrarlayan sinir ağları (Recurrent Neural Network, RNN) modelde %72.2, LSTM RNN modelde %77.6 başarı sağlanmıştır. Ayrıca transfer öğrenme modeli için sadece SNLI kullanılan, sadece SICK kullanılan ve SNLI ile SICK beraber kullanıldığı(transfer öğrenme) veri setlerinde sonuçları incelemişlerdir. SNLI ile SICK kullanıldığı transfer öğrenme modelinde %80 lik bir başarı elde etmişlerdir. RTE-3 üzerinde de çalıştıklarını ancak modelin derlemede kullanılan metin türüne uyum sağlayamadığını ve hiçbir eğitim yapılandırması olmadığını ifade etmişlerdir. Sonuç olarak NLI'nin semantik temsil çalışmaları için oluşturdukları veri setinin ideal bir test ortamı olacağını, zengin genel anlamsal temsiller sağlayabileceğini, hem sözcükleştirilmiş modellerin hem de sinir ağı modellerinin iyi performans gösterdiğini bulmuşlardır. Yin *et al.* (2019), sıfır atışlı metin sınıflandırmasını, veri setinin oluşturulması ve değerlendirilmesi açısından daha detaylı bir çalışma haline getirmişlerdir. Geçmişte yapılan çalışmaların, SVMS probleminde konu sınıflandırılması görevine odaklanarak kısıtlayıcı bir vizyonla çalışmalar modellendiklerini ifade etmişlerdir. İkinci olarak daha önce yapılan çalışmalar da sınıfların bir kısmının görülmesi ve bir modeli eğitmek için etiketli örneklerin mevcut olması gerektiği koşulu olduğunu, yapmış oldukları çalışmada daha geniş bir vizyon ile bakarak metin verisine konu, duygu veya metinde açıklanan durum yönü gibi farklı etiketlerin atanabileceğini ve metnin farklı yönlerinin değerlendirilebileceğini belirtmişlerdir. Bu nedenle gerçek dünya sorunlarını gidermek için görülmeyen etiketlerin olduğu, eğitim verisinin olmadığı durumlar üzerinde de çalışmışlardır. Çalışmanın hedefinin insanlar gibi çalışan bir sistem kurmak olduğunu ifade etmişlerdir. Çeşitli yönlerden etiketlerle başa çıkabilmek için herhangi bir göreve özel etiketli veriye erişmeden bir SVMS sistemi nasıl geliştirilir sorusuna cevap aramışlardır. Genellikle insanların bir metni kategorize ettiklerinde zihinsel olarak "Bu metin ne hakkında?" sorusuna cevap aradıklarını, bir hipotez oluşturduklarını ve metne bakarak bu hipotezin doğru olup olmadığını sorduklarını ifade etmişlerdir. İnsanların herhangi bir yönden etiketlerin gerçek değerine nasıl karar verdiklerini taklit etmek için SVMS'yi metinsel bir problem olarak görmüşlerdir. Konu sınıflandırması için Yahoo, duygu sınıflandırması için Bostan and Klinger (2018), tarafından yayınlanan UnifyEmotion, durum sınıflandırılması için Mayhew *et al.* (2019), tarafından yayınlanan veri kümelerini kullanmışlardır. Bilinmeyen etiketlerinde sınıflandırılmasının sağlanması amacıyla problemin ve etiketlerin model tarafından anlaşıldığından emin olunması gerektiğini belirtmişlerdir. Bu nedenle de etiketleri hipotezlere dönüştürmüşlerdir. Bir yorumu hipoteze dönüştürürken ilk olarak etiket adını kullanmışlar, ikinci olarak WordNet'teki etiketin tanımını kullanmışlardır. MNLI, FEVER ve RTE veri kümeleri ile eğitilen BERT model için

görülen ve görülmeyen sınıfların performansını karşılaştırmışlardır. MNLI veri seti; yazılı ve sözlü kaynaklardan elde edilen yaklaşık 400.000 metin verisinden oluşmaktadır. FEVER ise sahte haberler ve otomatik bilgi kontrolü ile ilgili sorunlardan esinlenerek oluşturulan veri setidir. Konu, duygu ve durum yönünde görülen ve görülmeyen sınıflandırma çalışmasında Majority'nin, Word2Vec ve ESA modelin de yapılan etiket sınıflandırma çalışmasından daha başarılı olduğunu gözlemlemişlerdir. Ayrıca görünmeyen sınıflandırma için Wikipedia'dan toplanan veri kümesinde BERT'e dayalı ikili sınıflandırma eğiterek çalışma yapmışlardır. Bu çalışmada da Wikipedia tabanlı eğitimin konu algılama görevinin duygu ve durum tespitinden daha iyi performans gösterdiğini, ayrıca yine yapılan çalışma da RTE 'nin diğer veri kümelerinden daha iyi performans gösterdiğini belirtmişlerdir.

Srivastava *et al.* (2017), doğal dilden kavram öğrenme problemi üzerinde çalışmışlardır. Sistemin kendi kendine insan gibi dilden öğrenme yetisi kullanılarak az veri içeren veri kümelerinde sınıflandırma başarısı incelenmiştir. Çalışmada 7 kavram (contact, employee, event, humor, meeting, policy, reminder, average) için katılımcılar tarafından yapılan 230 doğal dil ifadeleri ile eşleştirilmiş 1.030 epostadan oluşan verikümesini kullanmışlardır. Çalışmada düşük veri yoğunluğunda rekabetçi sınıflandırma yöntemlerine göre(BoW, BoW tf-idf, Para2vec, Bigram, ESA vb.) F1 puanında %30 'luk bir iyileşme ile sınıflandırma performansı elde etmişlerdir. Srivastava *et al.* (2018), yapmış oldukları bir diğer çalışmada etiketli örnek olmadan açıklamaların kullanılacağı bir model önermişlerdir. Yapılan çalışmada açıklamalarda gizli etiketler ve etiketsiz verilerin gözlemlenen özniteliklerine dayanan olasılıksal ifadelerle eşleştirmek için anlam bilimsel ayrıştırma kullanılmakta ve model eğitimi için dilsel niceleyicilerin farklı anlam biliminden (Örneğin; "genellikle" ve "her zaman") yararlanılmaktadır. Kavramları tanımlayan dilin istatistiksel öğrenmeye yardımcı olabilecek temel özelliklerini kodlamışlardır. Örneğin; Bir epostanın yanıtlanıp yanıtlanmadığı, bir yanıtın bu e-postanın önemli olduğunu ima etmesi gibi. Çalışmada şekillerden, emaillerden ve kuş görsellerinden oluşan 3 farklı veri seti kullanmışlardır. Şekil veri seti için, katılımcılara örnek veriler gösterilerek, 50 tane veri kümesi örnekleri için katılımcılardan bu kavramları tanımlayan ifadeler yazmaları istenmiştir. Toplamda 30 katılımcı kullanılarak veri kümesi başına ortalama 4.3 ifade oluşturulmuştur. Email veri seti, 7 farklı email kategorisini tanımlayan kullanıcılardan gelen dil açıklamalarını içeren 1.030 eposta örneğini içermektedir. Kuş verileri ise boyut, renk, şekil gibi gözlemlenebilir nitelikleri içeren açıklamalardan oluşan veri kümesidir. Veri kümeleri üzerinde yapmış oldukları deneylerde, öğrenilen sınıflandırıcıların, sınırlı verilerle öğrenmeye yönelik önceki yaklaşımlardan daha iyi performans gösterdiğini ve az sayıda etiketli örneklerden eğitilmiş tam denetimli sınıflandırıcılarla karşılaştırılabilir olduğunu göstermişlerdir. Pushp and Srivastava (2017), yapmış oldukları başka bir çalışmada

metin sınıflandırma işlemi için bir sıfır vuruşlu öğrenici yaklaşımı önermişlerdir. Önerilen yöntem, bir cümle ile cümle etiketlerinin yerleştirilmesi arasındaki ilişkiyi öğrenmek için geniş cümle külliyatı üzerinde eğitilen eğitim modeli içermektedir. Bu tür bir ilişkinin öğrenilmesinin modelin görülmeyen cümlelere, etiketlere ve hatta aynı gömme alanına yerleştirilebilmeleri koşuluyla yeni veri kümelerine genellenmesini sağlayacağını düşünmüşlerdir. Model belirli bir cümlenin bir etiketle ilişkili olup olmadığını tahmin etmeyi öğrenir. Yapılan çalışmada 3 farklı sinir ağı modeli geliştirilmiş ve doğruluğu eğitim aşamasında kullanılan veri seti dışında test aşamasında eğitimin yapılmadığı iki farklı veri seti kullanılarak raporlanmıştır. İlk sinir ağı modelinde cümledeki kelime yerleştirmelerinin ortalamasını cümle yerleştirme olarak almışlar ve onu etiket yerleştirme ile birleştirmişlerdir. Bu vektör daha sonra, cümle ve etiketin ilişkili olup olmadığını sınıflandırmak için tam bağlantılı bir katmandan geçirilmiştir. 2. modelde ortalamayı almak yerine, kelime yerleştirmeleri bir LSTM'den geçirilmiş ve ağın son gizli durumu cümle vektörü olarak ele alınmıştır. Etiket kelime vektörü ile birleştirilmiş ve daha sonra cümle ve etiketin ilişkili olup olmadığını sınıflandırmak için tam bağlantılı bir katmandan geçirilmiştir. Bu mimaride, her kelimenin cümleye gömülmesi, etiketin gömülmesiyle birleştirilmiştir. Birleşik yerleştirme bir LSTM'den geçirilmiş ve ağın son gizli durumu alınmıştır. Daha sonra, cümle ve etiketin ilişkili olup olmadığını sınıflandırmak için tam bağlantılı bir katmandan geçirilmiştir. Yazarlar web'den 4,2 milyon haber başlığını taramış ve haber makalesi için SEO etiketlerini etiket olarak kulanımışlardır. Taramadan sonra, etiket olarak toplam 300.000 benzersiz etiket almışlardır. Test veri seti olarak UCI News Aggretor ve tweet sınıflandırma veri kümelerini kullanmışlardır. UCI News Aggretor veri seti teknoloji, işletme, tıp ve eğlence kategorilerine ait toplam 420.000 cümle içermektedir. Tweet sınıflandırma veri seti iş, sağlık, politika, spor, teknoloji ve eğlence olmak üzere 6 kategoriye ait 1.993 cümleden oluşmaktadır. Yazarlar, eğitim sırasında mevcut olmayan etiketleri içeren bir bekleme testi setinde modellerini test etmişlerdir. Mimari 1, 2 ve 3 ile ikili sınıflandırma görevinde sırasıyla %78, %76 ve %81 doğruluk elde etmişlerdir. Bansal *et al.* (2020), farklı sınıf sayılarına sahip görevler arasında optimizasyona dayalı meta öğrenmeyi mümkün kılan yöntemler üzerinde de çalışmışlardır. Meta öğrenme yöntemi olarak sunulan LEOPARD etiket başına 4 örnekle eğitim sırasında hiç görülmeyen görevlere daha iyi bir genelleme yapabildiğini göstermişlerdir. Doğal dil çıkarımı, duygu analizi ve diğer birkaç metin sınıflandırma görevi dahil olmak üzere 17 NLP görevinde test etmişlerdir. Çalışmalarda %14,6 ortalama bağıl kazanç sağladığını göstermişlerdir. Ancak meta öğrenme kullanılarak iyileştirmeler yapılsa da hala insan düzeyindeki performansın gerisinde kaldığını gözlemlemişlerdir. Puri and Catanzaro (2019), yeni görevlere sıfır atışlı model uyarlamasını sağlamak için doğal dilin kullanımını araştırmışlardır. Eğitim görevi için sosyal yorum platformlarından metin verilerini

kullanmışlardır. Dil modeline girdi olarak sınıflandırma görevlerinin doğal dil açıklamalarını ekleyerek doğal dilde doğru cevap üretilmesi için eğitme işlemi yapılmıştır. Yapılan deney sonuçlarında doğal dilin görev uyarlaması için basit ve güçlü tanımlayıcılar olarak kullanılabileceği göstermişlerdir. Yin (2020), NLP alanındaki az atışlı uygulamalara odaklanarak meta öğrenme uygulanmış çalışmaları tanımlamış, bu alanda yapılmış ilerlemeleri özetlemiş ve bu alanda kullanılan veri kümelerini incelemiştir. Çalışmada birkaç atışlık NLP veri setlerinin varlığına rağmen, daha zorlu ve gerçekçi kıyaslamalara ihtiyaç olduğunu ifade etmiştir.

Zhang *et al.* (2019), yapmış oldukları çalışmada yetersiz, örneği olmayan eğitim verileri olduğunda zorlu sınıflandırma işlemine çözüm olarak veri ve özellik büyütme şeklinde iki aşamalı bir model önermiştirler. İlk aşama kaba taneli sınıflandırma olarak adlandırılan bir girdinin görünen veya görünmeyen sınıflardan gelip gelmediğini tahmin eden yaklaşımdır. Ayrıca birinci aşamada sınıflandırıcıların gerçek verilerine erişmeden görünmeyen sınıfların varlığından haberdar olmak için bir veri büyütme tekniği uygulamışlardır. Öğrenme aşamasında, ilk aşamadaki sınıflandırıcılar yalnızca görünen sınıflardan olumsuz örnekler kullandığını, bu nedenle pozitif sınıfı görünmeyen sınıflardan ayırt edemediğini bundan dolayı da çıkarım aşamasında görünmeyen sınıfların adları bilindiğinde, onları Aşama 1'deki sınıflandırıcılara artırılmış veriler aracılığıyla tanıtmaya çalışarak, görünmeyen sınıflardan olası örnekleri reddetmeyi öğrenebilineceğini düşünmüşlerdir ve analoji kullanarak bir belgeyi orijinal görünen sınıfından yeni bir görünmeyen sınıfa çevirerek veri büyütme işlemi yapmışlardır. Bu işleme de konu çevirisi adını vermişlerdir. İkinci aşamanın ise ince taneli olarak adlandırılan girdinin nihai sınıf bilgisinin belirlendiği yaklaşım olarak tanımlamışlardır. Model için yapılan deneylerde 2 veri seti kullanmışlardır. Bunlar 14 örtüşmeyen sınıf ve metinsel veriden oluşan her sınıfta 40.000 eğitim, 5.000 test örneği içeren DBpedia ontoloji veri seti ile her biri yaklaşık 1.000 belge içeren 20 konuya sahip 20newsgroups veri setidir. Çalışmada her sınıfa ait dokümanların %70'i rastgele seçilerek eğitim için, kalan %30'u ise test veri seti olarak kullanmışlardır. Deneylerde %50 ve %25 olmak üzere iki farklı görünmeyen sınıf oranı seçerek inceleme yapmışlardır. Deneylerin sonucunda, konu çevirisi ile veri büyütmenin görünmeyen sınıflardan örnekleri tespit etmede doğruluğu iyileştirdiğini, özellik artırma işleminin sıfır atış öğrenme için görünenden görünmeyen sınıflara bilgi aktarımını mümkün kıldığını ifade etmişlerdir. Ye *et al.* (2020), eğitim sırasında görünmeyen sınıflar için etiketlenmiş veri bulunmadığında sıfır atış öğrenmenin zorlanma eğiliminde olduğunu, bu sorunun üstesinden gelmek için bir yöntem sunduklarını ifade etmişlerdir. Etiketlenmemiş verilerden verimli bir şekilde yararlanmak için kendi kendine eğitim tabanlı bir yöntem önermişlerdir. Yöntemin amacı görünmeyen sınıflardan yüksek kaliteli verileri otomatik olarak

seçmek ve bu verileri temel eşleştirme modelinin performansını artırmak için kullanmaktır. Temel eşleştirme modelini görünen sınıf verileri üzerinde eğitmişler, görünmeyen sınıf verileri üzerinde tahmin yapmışlardır. Sistemi daha verimli hale getirmek için her yinelemede tüm etiketlenmemiş veriler yerine etiketlenmemiş verilerin bir alt kümesinden veri seçmişlerdir. Yapmış oldukları kıyaslamalar ve e-ticaret veri kümesi üzerindeki deneysel sonuçlar ile etiketlenmemiş verilerin entegrasyonunun ve güçlendirilmiş veri seçim politikalarının etkinliğini göstermişlerdir. Chen *et al.* (2021), büyük miktardaki sosyal medya metin verilerinin sınıflandırılması için mevcut bilgi grafiklerinden etkin bir şekilde yararlanan bir sıfır vuruşlu öğrenme yöntemi önermişlerdir. S-BERT modelinde amaç cümle düzeyinde bir gösterimi öğrenmek olduğundan çoğu sınıf etiketi bir veya birkaç kelime içerirken, S-BERT gömme modeli, kelime düzeyinde gömme yöntemleri kadar anlamsal olarak tutarlı olmayacağını düşünerek bu sorunu çözmek için ConseptNet bilgi grafiği ile etiket temsili için bir bilgi grafiği gömme modeli oluşturmuşlardır. Cümle yerleştirme çalışmasını daha sonra en küçük kareler doğrusal projeksiyon yoluyla bilgi grafiğine yansıtmışlardır. Önerdikleri modeli S-BERT-KG olarak tanımlamışlardır. Modeli herhangi bir etiketli verisi olmayan İngilizce COVID-19 tweet veri setinde incelemişlerdir. Deney sonuçlarının umut verici olduğunu ifade etmişlerdir. Birim vd (2021), yılında yapmış oldukları çalışma da 2 farklı Türkçe veri setinde sıfır atış algoritmasıyla analiz yapmışlardır. TTC-3600 veri kümesi 6 haber kategorisinden oluşmaktadır. Diğer kullandıkları veri seti ise; pozitif ve negatif etiketlerden oluşan Twitter mesajlarından topladıkları veri setidir. Sıfır atış metin sınıflandırması için DDİ çalışma alanı olan doğal dil çıkarımından yararlanmışlardır. Sıfır atış yönteminin ne kadar etiketli veri ile yapılan gözetimli öğrenmeye karşılık geldiğini ölçmek için BERT yöntemi uygulamışlardır. Kullandıkları doğal dil çıkarım yönteminde hipotez cümle oluşturulurken kullanılan '{}' yerine konulan kategori ile anlam bütünlüğü sağlanmasına dikkat etmişlerdir. Örneğin; "Bu cümlenin duygusu {}" hipotez cümlesi için "{}" yerine pozitif ve ya negatif etiketleri geldiğinde cümlenin anlamlı olup olmadığına dikkat etmişlerdir. Farklı doğal dil çıkarım şablonları kullanarak sınıflandırma başarımı performansı ölçülmüştür. BERT yönteminde ise ön eğitimli bir model kullanılarak testler gerçekleştirmişler ve ihtiyaç duyulandan az örnek ile başarılı sonuçlar ürettiğini ifade etmişlerdir.

MATERYAL VE YÖNTEM

Bu bölümde bu tez kapsamında gerçekleştirilen çalışmalarda kullanılan materyal ve yöntemler açıklanmıştır. Yapılan çalışmamızda amacımız dengesiz ve resmi olmayan ortamlardan toplanan verilerdeki geleneksel makine ve derin transfer öğrenme yaklaşımlarının performanslarını incelemektir. Bu kapsamda farklı özeliklere sahip veri setleri kullanılarak gerçek hayattaki senaryolara yöntemlerin uygunluğu incelenmiştir. Materyal başlığı altında bu tez kapsamında kullanılan veri setlerinin özellikleri açıklanmıştır. Yöntem başlığı altında geleneksel makine öğrenmesi ve derin transfer öğrenme algoritmalarının duygu analizi çalışmasında kullandığımız yöntemleri açıklanmıştır.

Materyal

Tez kapsamında özellikle dengesiz veri setlerinde geleneksel makine ve derin transfer öğrenme yaklaşımlarının performansları incelenmiştir. Bu kapsamda kullanılan veri setlerinde bulunan cümlelerin maksimum ve minumum uzunluk değerleri ile metin uzunluklarına ait hesaplanan standart sapma değerleri Tablo 1 ve 2'de gösterilmiştir. Tablo 1'de İngilizce veri setleri için değerler verilmiştir.

Tablo 1. İngilizce Veri Setine ait Metin Uzunluk Değerleri

	AG News		**IMDB**		**Emotion**	
	Eğitim	Test	Eğitim	Test	Eğitim	Test
Max Metin U.	729	513	7.756	8.502	204	196
Min Metin U.	2	23	17	31	3	9
Ort Metin U.	128,88	128,10	758,38	769,03	52,97	52,52
Standart Sapma	45,42	43,76	583,09	593,33	30,69	30,16

İngilizce veri setleri incelendiğinde Emotion ve AG News veri setlerinde standart sapma düşük çıkmıştır. Tablo 2'de Türkçe veri setlerine ait değerler verilmiştir. Türkçe veri setlerinde Twitter veri setinde standart sapma diğer Türkçe veri setleri ile karşılaştırıldığında daha düşük çıkmıştır.

Tablo 2. Türkçe Veri Setine ait Metin Uzunluk Değerleri

	Telefon		**Twitter**		**360 Yorumlar**	
	Eğitim	Test	Eğitim	Test	Eğitim	Test
Max Metin U.	1721	1288	211	5447	1006	740
Min Metin U.	3	3	3	4	5	5
Ort Metin U.	185,98	188,66	55,29	56,84	155,5	158,18
Standart Sapma	198,72	174,10	28,62	86,44	111,42	108,95

Tablolar incelendiğinde IMDB, Telefon ve 360 Yorumlar veri setlerinde standart sapmanın yüksek olduğu görülmektedir. Bu da veri setlerinde bulunan cümlelerin uzunluk açısından sabit olmadığını göstermektedir. AG News, emotion ve Twitter veri setlerinin standart sapma değerleri incelendiğinde cümle uzunluğu açısından daha dengeli dağılım gösterdiği görülmüştür. Tez kapsamında kullanılan veri setlerinin özellikleri alt başlıklarda daha detaylı açıklanmıştır.

AG News veri seti

1 milyondan fazla haber makalesinden oluşan koleksiyondur. Haber makaleleri ComeToMyHead tarafından 1 yılı aşkın bir sürede 2000' den fazla haber kaynağından toplanmıştır. ComeToMyHead, Temmuz 2004' ten beri çalışan bir akademik haber arama motorudur. Veri seti, akademik topluluk tarafından veri madenciliği (kümeleme, sınıflandırma, vb.), bilgi alımı (sıralama, arama vb.), xml, veri sıkıştırma, veri akışı ve diğer ticari olmayan faaliyetlerde araştırma amacıyla kullanılmaktadır. Zhang *et al.* (2015), metin sınıflandırması için karakter düzeyinde evrişimli ağların kullanımına ilişkin deneysel bir araştırma yapmışlardır. Karakter seviyesindeki evrişimli ağların en gelişmiş veya rekabetçi sonuçlara ulaşabileceğini göstermek için birkaç büyük ölçekli veri seti oluşturmuşlardır. Bu veri setlerinden biri de AG haber veri setidir. AG'nin haber konusu sınıflandırma veri seti, orijinal derlemden en büyük 4 sınıf seçilerek oluşturulmuştur. Bu sınıflar dünya, spor, iş ve teknoloji etiketlerinden oluşmaktadır. Her sınıf 30.000 eğitim örneği ve 1.900 test örneği içerir. Toplam eğitim örneği sayısı 120.000'dir ve 7.600 veri ile test edilir. Kelime çantası, n-gram ve TF-IDF varyantları gibi geleneksel modellerle ve kelime tabanlı ve tekrarlayan sinir ağları gibi derin öğrenme modelleriyle karşılaştırmalar yapılarak sonuçlar sunulmuştur. Veri seti özellikleri Tablo 3'de gösterilmiştir.

Tablo 3. AG News Veri Seti Sınıf Dağılımı

Sınıf	Eğitim Veri Sayısı	Test Veri Sayısı	Toplam
Dünya (world)	30.000	1900	31.900
Spor (sport)	30.000	1.900	31.900
İş (business)	30.000	1.900	31.900
Teknoloji (ski/tech)	30.000	1.900	31.900
Toplam	120.000	7.600	127.600

IMDB veri seti

IMDB en popüler film web sitesidir ve film konusu açıklamasını, Metastore derecelendirmelerini, eleştirmen ve kullanıcı derecelendirmelerini ve incelemelerini, yayın tarihlerini ve daha birçok yönü birleştirir. Web sitesi, şimdiye kadar vizyona giren (en eskisi 1874 - "Passage de Venus") veya yeni çıkması planlanan (en yeni film 2027 - "Avatar 5") hemen hemen her filmi saklamasıyla ünlüdür. IMDB, 6 milyondan fazla kitapla neredeyse 500.000'i öne çıkan filmlerden oluşmaktadır. İlgili bilgileri depolar ve 1998'den beri Amazon'a aittir. Bu, önceki kıyaslama veri kümelerinden önemli ölçüde daha fazla veri içeren ikili duyarlılık sınıflandırması için bir veri kümesidir. Maas *et al.* (2011), yaptıkları çalışma ile anlamsal terim-belge bilgilerinin yanı sıra zengin duygu içeriği yakalayan kelime vektörlerini öğrenmek için denetimsiz ve denetimli tekniklerin karışımı bir model sunmuşlardır. Modeli geniş bir film incelemesi veri seti oluşturarak sonuçları değerlendirmişlerdir. Oluşturdukları veri seti ile eğitim için 25.000 ve test için 25.000 olmak üzere yüksek düzeyde kutupsal film incelemesi sağlamışlardır. Veriler pozitif ve negatif olarak 2 sınıftan oluşmaktadır. Veri seti özellikleri Tablo 4'de gösterilmiştir.

Tablo 4. IMDB Veri Seti Sınıf Dağılımı

Sınıf	Eğitim Veri Sayısı	Test Veri Sayısı	Toplam
Pozitif (positive)	12.500	12.500	25.000
Negatif (negative)	12.500	12.500	25.000
Toplam	25.000	25.000	50.000

Emotion veri seti

Saravia *et al.* (2018), yaptıkları çalışma ile metinden bağlamsallaştırılmış duygu temsilleri oluşturmak için yapı taşları olarak hizmet eden zengin yapısal tanımlayıcılar üretmek için yarı denetimli, grafik tabanlı bir algoritma önermişlerdir. Önerilen model ile örüntüye dayalı temsiller, sözcük yerleştirmeleri ile daha da zenginleştirilmiş ve çeşitli duygu tanıma görevleri ile değerlendirilmiş, emotion veri seti üzerinde çalışılmış ve sonuçlar sunulmuştur.

Emotion veri seti; altı temel duyguya sahip İngilizce Twitter mesajlarından oluşan bir veri kümesidir. Altı duygu; öfke, korku, sevinç, aşk, üzüntü ve şaşkınlık'tır. Eğitim 16.000, test 2.000 veriden oluşmaktadır. Veri seti özellikleri Tablo 5 'de detaylı gösterilmiştir.

Tablo 5. Emotion Veri Seti Sınıf Dağılımı

Sınıf	Eğitim Veri Sayısı	Test Veri Sayısı	Toplam
Öfke(anger)	2.159	275	2.434
Korku (fear)	1.937	224	2.161
Sevinç (joy)	5.362	695	6.057
Aşk (love)	1.304	159	1.463
Üzüntü (sadness)	4.666	581	5.247
Şaşkınlık (suprise)	572	66	638
Toplam	16.000	2.000	18.000

Twitter veri seti

Çoban ve Özyer (2015), yaptıkları çalışmada Türkçe twitter mesajlarından oluşturdukları veri setiyle, metin sınıflandırma yöntemlerini analiz ederek sonuçları sunmuşlardır. Kullanılan veri seti Türkçe twitter mesajlarından pozitif ve negatif olmak üzere iki sınıf etiketi ile oluşturulmuş veri setidir. Veri seti 6.829 pozitif ve 7.828 negatif veri olmak üzere toplam 14.657 kayıt bulunmaktadır. Bu veri seti içerisindeki kayıtlar yoğun bir şekilde hashtag ve emojiler içermektedir. Resmi dil kullanımından uzak uygulama yorumlarının analiz edilmesi için oldukça güzel bir örnek olmaktadır. Geleneksel makine öğrenme yöntemleri için veri setinin %70'i eğitim %30'u test için kullanılmıştır. Derin transfer öğrenme yöntemleri için de test veri seti kullanılarak sonuçlar sunulmuştur. Tablo 6'da veri setleri özellikleri verilmiştir.

Tablo 6. Twitter Veri Seti Sınıf Dağılımı

Sınıf	Eğitim Veri Sayısı	Test Veri Sayısı	Toplam
Pozitif	4.783	2.046	6.829
Negatif	5.476	2.352	7.828
Toplam	10.259	4.398	14.657

Bir markaya ait mobil telefon hakkında yorumlar içeren veri seti

Yılmaz (2019), veri setinde yapmış olduğu duygu analizi çalışmasının sonuçlarını Kaggle sitesi üzerinden yayınlamıştır. Veri seti web sitesi üzerinden bir telefon markasına ait müşterilerin ürün hakkındaki Türkçe yorumlarını içermektedir. Yorumlar resmi olmayan bir ortamdan oluşturulduğu için emojiler ve noktalama işaretleri içermektedir. Veri seti, 740 tane pozitif, 236 tane negatif olarak etiketlenmiş toplam 976 kayıt içermektedir. Çalışmamızda

geleneksel makine öğrenme yöntemleri için veri setinin %70'i eğitim %30'u test için kullanılmıştır. Derin transfer öğrenme yöntemleri için de test veri seti kullanılarak sonuçlar sunulmuştur. Tablo 7' de veri seti özellikleri detaylı verilmiştir.

Tablo 7. Telefon Yorumları Veri Seti Sınıf Dağılımı

Sınıf	Eğitim Veri Sayısı	Test Veri Sayısı	Toplam
Pozitif	525	215	740
Negatif	158	78	236
Toplam	683	293	936

360 Yorumlar veri seti

Çelik vd (2021), yapmış oldukları çalışma ile oluşturmuş oldukları veri seti üzerinde TF-IDF ve Bow modeli kullanarak geleneksel makine öğrenmesi yöntemleri uygulamışlar ve deney sonuçlarını sunmuşlardır. Kullandıkları veri seti, özel bir şirket olan Vektora'daki Türk çalışanların yorumlarından oluşmaktadır. Yorumlar, Vektora Bilgi Teknolojileri A.Ş. bünyesinde yürütülen HR360 projesi kapsamında toplanmıştır. HR360 projesi, 360 derece değerlendirme adı verilen çalışanları her yönüyle değerlendirmeyi hedefliyor. Proje kapsamı, çalışan yorumlarının değerlendirilmesi, müşteri yorumlarının değerlendirilmesi, katılan görev puanlamasını ve yönetici kararlarını içermektedir. Ortam profesyonel bir çalışma alanı olduğu için yorumlarda kullanılan dil daha resmi ve gramer açısından uygundur. Yorumların çoğu dil bilgisi açısından doğru ve anlamlıdır. İlk yıl faz 1 ve ikinci yıl faz 2 kapsamında toplanan verilerden oluşan veri setidir. Olumlu, olumsuz ve tarafsız olmak üzere 3 sınıftan oluşan toplam 2.442 kayıt içermektedir. Çalışmamızda, geleneksel makine öğrenme yöntemlerinde veri setinin %70'i eğitim %30'u test için kullanılmıştır. Derin transfer öğrenme yöntemleri için de test veri seti kullanılarak sonuçlar sunulmuştur. Farklı yıllara ait faz veri seti özellikleri Tablo 8'de gösterilmiştir. Test ve eğitim veri setlerine ait özellikler Tablo 9'da verilmiştir.

Tablo 8. 360 Yorumlar Faz Veri Seti Sınıf Dağılımı

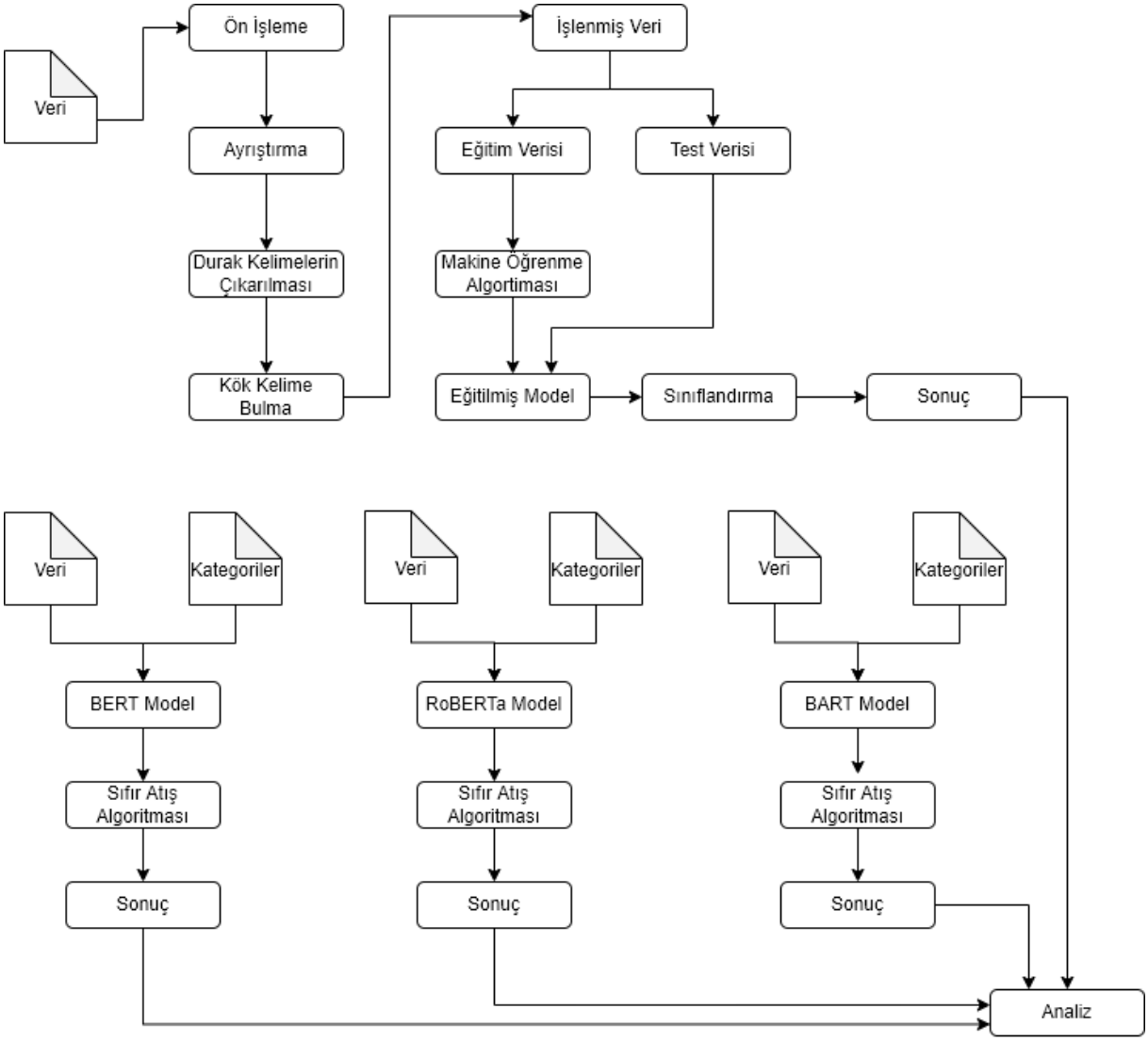
	Olumlu	Olumsuz	Tarafsız	Toplam
Faz 1	619	98	33	750
Faz 2	1.289	243	160	1.692
Toplam	1.908	341	193	2.442

Tablo 9. 360 Yorumlar Veri Seti Sınıf Dağılımı

Sınıf	Eğitim Veri Sayısı	Test Veri Sayısı	Toplam
Olumlu	1.335	573	1.908
Olumsuz	236	105	341
Tarafsız	138	55	193
Toplam	1.709	733	2.442

Yöntem

Çalışmamızda farklı özellikte veri setleri kullanılarak geleneksel makine ve derin transfer öğrenme yöntemleri incelenmiştir. Tez kapsamında çalıştırılan algoritmalar ve analiz süreci Şekil 9'da gösterilmiştir.



Şekil 9. Tez uygulama çalışma diyagramı

Şekil 9'da gösterildiği gibi geleneksel makine öğrenmesi modelimiz için verilerimiz ön işlemlerden geçirilerek sınıflandırma işlemi yapılmış ve sonuçlar elde edilmiştir. Derin transfer

öğrenme modelimizde ise veriler ve kategoriler verilerek BERT, RoBERTa ve BART model ile sıfır atış algoritması çalıştırılmıştır. Veri setlerimiz üzerinde hem geleneksel makine öğrenmesi algoritmaları hem de derin transfer öğrenme yöntemleri ile sıfır atış algoritması çalıştırılarak elde edilen doğruluk değerleri incelenmiştir.

Sistemimiz 2 farklı model üzerinde çalışmaktadır. İlk model geleneksel makine öğrenme modeli, ikinci model derin transfer öğrenme modelidir. Geleneksel makine öğrenme modeli ön işleme, öznitelik çıkarımı ve sınıflandırma aşamalarından oluşmaktadır. Bu model de tez kapsamında veri setleri öncelikle ön işleme adımlarından geçirilmektedir. İngilizce veri setleri için İngilizce durak kelimeleri, Türkçe veri setleri için Türkçe durak kelimeleri verilerden çıkarılmaktadır. Daha sonra verilerde bulunan noktalama işaretleri, birden fazla ard arda bulunan boşluklar, özel karakterler ve sayılar veri setinden çıkarılmaktadır. Kalan kelimeler için kök ayırma işlemi yapılarak kelimeler eklerden temizlenmektedir. Ön işlem adımı tamamlandıktan sonra Türkçe ve İngilizce veri setlerinin makinenin anlayabileceği sayısal formata dönüştürülmesi adımı olan öz nitelik çıkarımında vektörleştirme için TF-IDF ve Doc2Vec modelleri kullanılmıştır. TF-IDF ilgili terimlerin dokümanlarda geçme sıklığına göre hesaplama işlemi yapan modeldir. Yapay Sinir ağlarına dayalı bir metot olan Doc2vec ise hedef kelimeyi tahmin etmek için sadece kelimeler kullanmak yerine, belgeye özgü ek bir özellik vektörü de ekleyerek işlem yapmaktadır. Ek özellik vektörü (paragraf vektörü) eklemek için çalışmamızda Doc2vec de PV-DBOW model kullanılmıştır. Veriler bu yöntemler ile vektörel hale getirildikten sonra sınıflandırma aşamasına geçilmektedir. Bu aşama da veri bir geleneksel makine öğrenme algoritmasına verilmekte ve eğitilmiş bir model oluşturulmaktadır. Yeni gelen veriler eğitilmiş model ile sınıflandırılmaktadır. İkinci model olan derin transfer öğrenme modelinde ise önceden eğitilmiş NLI üzerinde çalıştırıldığı için, herhangi bir ön işlem veya eğitim süreci içermemektedir. Veriler ve kategoriler belirlenerek algoritma çalıştırılmaktadır. Derin transfer öğrenme yönteminde BERT, RoBERTa ve BART modelleriyle sıfır atış algoritması çalıştırılmıştır. Bu modeller özellikle doğal dil anlama üzerine geliştirilmiş modellerdir. Büyük miktarda verilerle eğitilmişlerdir. Çalıştığımız modellerden BERT model cümledeki kelimeleri tek tek değil, sağında ve solunda bulunan diğer tüm kelimelerle ilişkili olarak işleyen bir modeldir. RoBERTa model, BERT model ile benzer mantıkta çalışan, eğitim verisinde kelimelerin maskeleme işleminde farklılık göstererek dinamik maskeleme yöntemi kullanan modeldir. Çalışma da huggingface model merkezinden hem Türkçe hem İngilizce dillerini kapsayan "joeddav/xlm-roberta-large-xnli" RoBERTa (Conneau *et al.* 2020) modeli kullanılmıştır. BART ise, BERT gibi çift yönlü kodlayıcı ve soldan sağa kod çözücü içeren standart bir makine çevirisi mimarisi kullanan modeldir. Çalışma da huggingface model merkezinden hem Türkçe hem İngilizce için "facebook/bart-large-mnli"

BART (Lewis *et al.* 2019) modeli kullanılmıştır. Uygulamamızda BERT modeli için sıfır atış algoritmasında BERT modeli ile aynı korpus üzerinde eğitilmiş BERT'den daha küçük ve hızlı bir tranformatör modeli olan DistilBERT (Sanh *et al.* 2020) kullanılmıştır. Huggingface model merkezinden Türkçeyi kapsayan "dbmdz/distilbert-base-turkish-cased" modeli ve İngilice'yi kapsayan "typeform/distilbert-base-uncased-mnli" modeli kullanılmıştır. Pipeline kütüphanesi kullanılarak DistilBERT, RoBERTa ve BART modelleri ile sıfır atış algoritması çalıştırılmıştır. Çalışmamızda hem geleneksel makine öğrenmesi hem de derin transfer öğrenme çalışmalarında performans ölçütü olarak doğruluk değeri kullanılmıştır. Doğruluk formülü Eşitlik 3.1'de gösterilmiştir. Doğruluk, doğru sınıflandırılmış örnek sayısının tüm örneklere olan oranı ile gösterilmektedir.

$$Doğruluk = \frac{TP+TN}{TP+TN+FP+FN} \tag{10}$$

Doğruluk formülünde yer alan TP, TN, FP, FN değerleri aşağıda açıklanmıştır.

- Gerçek Pozitifler (TP): Bunlar gerçek değeri pozitif ve tahmin ettiğimiz değerin de pozitif olduğu örneklerdir.
- Gerçek Negatifler (TN): Bunlar gerçek değeri negatif ve tahmin ettiğimiz değerin de negatif olduğu örneklerdir.
- Yanlış Pozitifler (FP): Bunlar gerçek değeri negatif ancak tahmin ettiğimiz değerin pozitif olduğu örneklerdir.
- Yanlış Negatifler (FN): Bunlar gerçek değeri pozitif ancak tahmin ettiğimiz değerin negatif olduğu örneklerdir.

Çalışma kapsamında geliştirilen kodlar Python dili kullanılarak kodlanmıştır. Kodlama Anaconda uygulaması üzerinden Jüpiter Notebook aracı ve Google Colab kullanılarak yapılmıştır. Anaconda veri bilim uygulamaları için kullanılan bir python dağıtımıdır. Kısaca Colab olarak ifade edilen Colaboratory ise tarayıcımızda Python yazmamızı ve çalıştırmamızı sağlayan, ücretsiz GPU'lara erişim imkanı sunan Google uygulamasıdır. Python arayüzündeki kütüphanelerin birçoğu makine öğrenimi ve veri bilimi üzerine elverişlidir. Bu alanlardaki kütüphanelerde; yüksek kaliteli komutları, makine öğrenimi kütüphaneleri ve nümerik algoritma kütüphaneleri sürekli gelişmektedir. Python pandas, numpy, scikit learn, pipeline kütüphaneleri aktif olarak kullanılmıştır. Pandas, veri işleme ve temizleme alanında sıklıkla kullanılan python kütüphanesidir. Numpy bilimsel hesaplarda(maskelenmiş diziler, matrisler, şekil işleme,sıralama vb.) kullanılan python kütüphanesidir. Scikit learn, doğrusal regrasyon, lojistik regrasyon, destek vektör makineleri gibi birçok yapay öğrenme yöntemini içeren python kütüphanesidir. Ayrıca Türkçe kelimeleri köklerine ayırmak için Zemberek kütüphanesi ve

İngilizce kelimeleri köklerine ayırma işlemi için nltk kütüphanesi kullanılmıştır. Pipeline, karmaşık kodlardan soyutlayan, modelleri çıkarım için kullanmayı sağlayan api'lerdir. Çalışmamız da sıfır atış algoritması için kullanımı kolay olan ve NLP desteği olan huggingface kullanılmıştır. Huggingface özellikle transformers doğal dil işleme modellerini araştıran ve geliştiren bir şirkettir. Huggingface, herhangi bir etiketli veri olmadan veya modellerin eğitilme sürecine ihtiyaç duymadan metin sınıflandırması, duygu sınıflandırması gibi problemler için çalışan pipeline kütüphanesi sunmaktadır. Sıfır atış algoritması için Huggingface tarafından sunulan pipeline kullanılmıştır. Bu kütüphane kullanılarak sıfır atış algoritması çalıştırılmıştır. Sıfır atış algoritması ile derin transfer öğrenme modelleri yüksek gpu 'lara ücretsiz erişim sağlayan Google Colab platformu ve Anaconda üzerinde çalıştırılmıştır.

Gelecek bölümde Türkçe ve İngilizce veri setlerine ait geleneksel makine ve derin transfer öğrenme algoritmalarının deneysel sonuçları sunulacaktır.

ARAŞTIRMA BULGULARI VE TARTIŞMA

Teknolojinin gelişmesi internet kullanımını artırmaktadır. İnternet kullanımının artması ile birlikte hemen hemen tüm ihtiyaçlar çevrim içi ortamda temin edilebilmektedir. Çevrim içi alışveriş, haber okuma, sesli mesajlar gönderme, kişisel asistanlar kullanma gibi bir çok işlem mobil ve web ortamda yapılabilmektedir. Dijital platformların kullanımı arttıkça rekabet ortamında ürün pazarlama önemli bir konu haline gelmiştir. Özellikle ürünler veya durumlarla ilgili insanlardan geri dönüş almak, alınan dönüşlerin hızlıca analiz edilerek geliştirme planları yapmak rekabet ortamında ön plana çıkmak için önemlidir. İnsanlardan alınan geri dönüş verilerini kısa sürede analiz etmek, verileri olumlu, olumsuz veya tarafsız olarak sınıflandırma yapmak önemli bir araştırma alanı haline gelmiştir. Verilerin pozitif veya negatif ya da olumlu, olumsuz, tarafsız olarak sınıflandırmak duygu analizi olarak ifade edilmektedir. Tez kapsamında amacımız; özellikle az ve dengesiz dağılım gösteren resmi olmayan ortamlardan elde edilen verilerde duygu analizi çalışması yaparak hem geleneksel makine öğrenmesi yöntemlerinin hem de derin transfer öğrenme yaklaşımlarının performanslarını incelemektir. Bu bölümde 3 İngilizce 3 Türkçe veri seti olmak üzere toplam 6 veri seti üzerinde geleneksel makine öğrenmesi ve derin transfer öğrenme yaklaşımları ile ilgili deneyler yapılarak sonuçlar sunulmuştur.

Geleneksel ve Derin Transfer Makine Öğrenme Algoritmalarının Türkçe ve İngilizce Veri Setlerinde Elde Edilen Sonuçları

Bu bölümde 3 İngilizce ve 3 Türkçe veri seti olmak üzere toplam 6 veri seti üzerinde yapmış olduğumuz geleneksel makine ve derin transfer öğrenme yöntemlerine ilişkin deney sonuçlarını inceleyeceğiz. SA sıfır atış modelini ifade etmektedir. D2V doc2vec modeli ifade etmektedir.

AG News veri seti sonuçları

AG News eğitim ve test veri setinde sık kullanılan kelimelere ait kelime bulutları Şekil 10’da gösterilmiştir.

(a) (b)

Şekil 10. AG News eğitim(a) ve test(b) veri setlerinde sık kullanılan kelimeler

Konu sınıflandırma işlemi için örnek bir veri setidir. Metinlerin dünya (world), spor (sport), iş (business), teknoloji (ski/tech) sınıflarından hangi sınıfa ait olduğu tahmin etmek için çalıştırılmıştır. Tablo 10'da geleneksel makine öğrenmesi sonuçları gösterilmiştir.

Tablo 10. AG News Veri Seti Geleneksel Makine Öğrenmesi Sonuçları

Makine Öğrenmesi Alg.	TF-IDF	Doc2vec
Lojistik Regrasyon	**%90,39**	%84,92
K En yakın Komşu	%90,06	%83,72
Destek Vektör Makinesi	%87,02	**%89,57**
Naive Bayes	%89,39	%82,77

AG News veri setinde en yüksek sonuç geleneksel makine öğrenmesinde TF-IDF modelinde LR algoritmasında %90,39 olarak elde edilmiştir. Derin transfer öğrenme modelinde elde edilen sonuçlar Tablo 11'de gösterilmiştir.

Tablo 11. AG News Veri Seti Derin Transfer Öğrenme Sonuçları

Sıfır Atış Modeli	Sonuç
BERT	**%60,00**
ROBERTA	%46,00
BART	%59,00

AG News veri setinde derin transfer öğrenme algoritması olan sıfır atış algoritmasında en iyi sonuç BERT modelde %60,00 olarak elde edilmiştir. Bu veri seti için geleneksel makine öğrenmesi ve derin transfer öğrenme modeller karşılaştırıldığında geleneksel makine öğrenmesi modelleri daha başarılı sonuçlar vermiştir.

IMDB Veri seti sonuçları

IMDB eğitim ve test veri setinde sık kullanılan kelimelere ait kelime bulutları Şekil 11'de gösterilmiştir.

Şekil 11. IMDB eğitim(a) ve test(b) veri setlerinde sık kullanılan kelimeler

Duygu sınıflandırma çalışmaları için kullanılan IMDB pozitif ve negatif olmak üzere 2 farklı sınıftan oluşan bir veri setidir. Tablo 12'de geleneksel makine öğrenmesi ile yapılan duygu sınıflandırma doğruluk yüzdeleri verilmiştir.

Tablo 12. IMDB Veri Seti Geleneksel Makine Öğrenmesi Sonuçları

Makine Öğrenmesi Alg.	TF-IDF	Doc2vec
Lojistik Regrasyon	**%87,50**	%87,17
K En yakın Komşu(k=7)	%66,14	%60,14
Destek Vektör Makinesi	%84,64	**%87,73**
Naive Bayes	%81,94	%86,28

IMDB veri setinde en yüksek sonuç geleneksel makine öğrenmesinde Doc2vec modelinde % 87,73 ile DVM algoritmasında elde edilmiştir. Derin transfer öğrenme modelinde elde edilen sonuçlar Tablo 13 'de gösterilmiştir.

Tablo 13. IMDB Veri Seti Derin Transfer Öğrenme Sonuçları

Sıfır Atış Modeli	Sonuç
BERT	%67,00
ROBERTA	%83,00
BART	**%89,45**

IMDB veri setinde derin transfer öğrenme algoritması olan sıfır atış algoritmasında en iyi sonuç BART modelde %89,45 olarak elde edilmiştir. Bu veri seti için geleneksel makine öğrenmesi ve derin transfer öğrenme modelleri karşılaştırıldığında derin transfer öğrenme modeli daha başarılı sonuç vermiştir.

Emotion veri seti sonuçları

Emotion eğitim ve test veri setinde sık kullanılan kelimelere ait kelime bulutları Şekil 12'de gösterilmiştir.

Şekil 12. Emotion eğitim(a) ve test(b) veri setlerinde sık kullanılan kelimeler

Emotion veri seti öfke, korku, sevinç, aşk, üzüntü ve şaşkınlık sınıflarına sahip İngilizce Twitter mesajlarından oluşan bir veri setidir. Tablo 14'de geleneksel makine öğrenmesi ile elde edilen duygu sınıflandırma doğruluk yüzdeleri verilmiştir.

Tablo 14. Emotion Veri Seti Geleneksel Makine Öğrenmesi Sonuçları

Makine Öğrenmesi Alg.	TF-IDF	Doc2vec
Lojistik Regrasyon	%84,55	%67,95
K En yakın Komşu(k=7)	%76,45	%70,05
Destek Vektör Makinesi	**%84,85**	**%71,15**
Naive Bayes	%78,25	%43,95

Emotion veri setinde en yüksek sonuç geleneksel makine öğrenmesinde TF-IDF modeli DVM algoritması ile %84,85 olarak elde edilmiştir. Derin transfer öğrenme modelinde elde edilen doğruluk yüzdeleri Tablo 15'de verilmiştir.

Tablo 15. Emotion Veri Seti Derin Transfer Öğrenme Sonuçları

Sıfır Atış Modeli	Sonuç
BERT	**%47,00**
ROBERTA	%31,00
BART	%38,00

Emotion veri setinde derin transfer öğrenme yaklaşımında en yüksek sonuç BERT model de %47,00 olarak elde edilmiştir. Genel olarak veri setinde elde edilen sonuçların düşük olmasının, veri setindeki duygu sınıflarının birbirine yakın olmasından ve sınıf sayısının fazla olmasından kaynaklandığını düşünüyoruz. Geleneksel makine ve derin transfer öğrenme

yaklaşımları karşılaştırıldığında en iyi sonuçlar geleneksel makine öğrenmesi algoritmalarıyla sağlanmıştır.

Twitter veri seti sonuçları

Twitter eğitim ve test veri setinde sık kullanılan kelimelere ait kelime bulutları Şekil 13'de gösterilmiştir.

Şekil 13. Twitter eğitim(a) ve test(b) veri setlerinde sık kullanılan kelimeler

Twitter veri seti pozitif ve negatif olmak üzere 2 farklı sınıftan oluşan veri setidir. Tablo 16'da geleneksel makine öğrenmesi ile yapılan duygu sınıflandırma doğruluk yüzdeleri verilmiştir.

Tablo 16. Twitter Veri Seti Geleneksel Makine Öğrenmesi Sonuçları

Makine Öğrenmesi Alg.	TF-IDF	Doc2vec
Lojistik Regrasyon	%65,86	%64,60
K En yakın Komşu(k=7)	%63,30	%61,73
Destek Vektör Makinesi	%64,00	**%65,89**
Naive Bayes	**%66,60**	%60,19

Twitter veri setinde en yüksek sonuç geleneksel makine öğrenmesinde TF-IDF modeli NB algoritması ile %66,60 olarak elde edilmiştir. Derin transfer öğrenme doğruluk yüzdeleri Tablo 17'de verilmiştir.

Tablo 17. Twitter Veri Seti Derin Transfer Öğrenme Sonuçları

Sıfır Atış Modeli	Sonuç
BERT	%53,66
ROBERTA	%85,54
BART	**%88,86**

Twitter veri setinde en yüksek sonuç derin transfer öğrenme BART modelinde %88,86 olarak elde edilmiştir. Bu veri seti için derin transfer öğrenme yaklaşımlarından RoBERTa ve BART model, geleneksel makine öğrenmesi yaklaşımlarından daha iyi performans göstermiştir. Twitter veri seti verilerin içerdiği emojilere göre etiketlendiğinden geleneksel makine öğrenmesi yaklaşımlarının düşük performans gösterdiğini düşünüyoruz.

Telefon markası yorumlarını içeren veri seti sonuçları

Telefon yorumlarını içeren eğitim ve test veri setinde sık kullanılan kelimelere ait kelime bulutları Şekil 14'de gösterilmiştir.

(a) (b)

Şekil 14. Telefon yorumları eğitim(a) ve test(b) veri setlerinde sık kullanılan kelimeler

Telefon yorumları veri seti pozitif ve negatif olmak üzere 2 farklı sınıftan oluşan veri setidir. Tablo 18'de geleneksel makine öğrenmesi ile elde edilen duygu sınıflandırma doğruluk yüzdeleri verilmiştir.

Tablo 18. Telefon Yorumları Veri Seti Geleneksel Makine Öğrenmesi Sonuçları

Makine Öğrenmesi Alg.	TF-IDF	Doc2vec
Lojistik Regrasyon	%78,62	%80,23
K En yakın Komşu(k=7)	%83,89	%83,45
Destek Vektör Makinesi	**%89,19**	**%88,87**
Naive Bayes	%84,91	%81,69

Telefon yorumları veri setinde en yüksek sonuç geleneksel makine öğrenmesinde TF-IDF modeli DVM algoritmasıyla %89,19 olarak elde edilmiştir. Derin transfer öğrenme doğruluk yüzdeleri Tablo 19'da verilmiştir.

Tablo 19. Telefon Yorumları Veri Seti Derin Transfer Öğrenme Sonuçları

Sıfır Atış Modeli	Sonuç
BERT	%31,06
ROBERTA	**%88,74**
BART	%55,63

Telefon yorumları veri setinde derin transfer öğrenme de en yüksek sonuç RoBERTa model ile %88,74 olarak elde edilmiştir. Bu veri seti için geleneksel makine öğrenmesi TF-IDF DVM modeli derin transfer öğrenme yaklaşımlarından daha iyi sonuç vermiştir. Ancak derin transfer öğrenme RoBERTa modeli Doc2vec LR, Doc2vec KNN(k=7), Doc2vec NB ve TF-IDF LR, TF-IDF KNN(k=7), TF-IDF NB algoritmalarından daha iyi sonuç vermiştir.

360 Yorumlar veri seti sonuçları

360 Yorumlar eğitim ve test veri setinde sık kullanılan kelimelere ait kelime bulutları Şekil 15'de gösterilmiştir.

(a) (b)

Şekil 15. 360 Yorumlar eğitim(a) ve test(b) veri setlerinde sık kullanılan kelimeler

360 yorumlar veri seti olumlu, olumsuz ve tarafsız olmak üzere 3 farklı sınıftan oluşan veri setidir. Tablo 20'de geleneksel makine öğrenmesi ile elde edilen duygu sınıflandırma doğruluk yüzdeleri verilmiştir.

Tablo 20. 360 Yorumlar Veri Seti Geleneksel Makine Öğrenmesi Sonuçları

Makine Öğrenmesi Alg.	TF-IDF	Doc2vec
Lojistik Regrasyon	%84,85	%79,80
K En yakın Komşu (k=7)	%84,17	%78,85
Destek Vektör Makinesi	**%85,40**	**%86,90**
Naive Bayes	%85,12	%68,75

360 yorumlar veri setinde geleneksel makine öğrenmesinde en yüksek sonuç Doc2vec modeli ile DVM algoritmasında %86,90 olarak elde edilmiştir. Derin transfer öğrenme doğruluk yüzdeleri Tablo 21'de verilmiştir.

Tablo 21. 360 Yorumlar Veri Seti Derin Transfer Öğrenme Sonuçları

Sıfır Atış Modeli	Sonuç
BERT	%21,83
ROBERTA	**%84,99**
BART	%56,75

360 yorumlar veri setinde derin transfer öğrenme de en yüksek sonuç RoBERTa ile %84,99 olarak elde edilmiştir. Bu veri seti için geleneksel makine öğrenmesi algoritmaları derin transfer öğrenme yaklaşımlarından daha iyi sonuç vermiştir. Ancak derin transfer öğrenme RoBERTa modeli TF-IDF LR, TF-IDF KNN(k=7), TF-IDF NB ve Doc2vec LR, Doc2vec KNN(k=7), Doc2vec NB makine öğrenme yaklaşımlarından daha iyi sonuç vermiştir.

Veri Setleri Deney Sonuçları

Bu bölümde Türkçe ve İngilizce veri setlerine ait sonuçlar detaylı sunulmuştur.

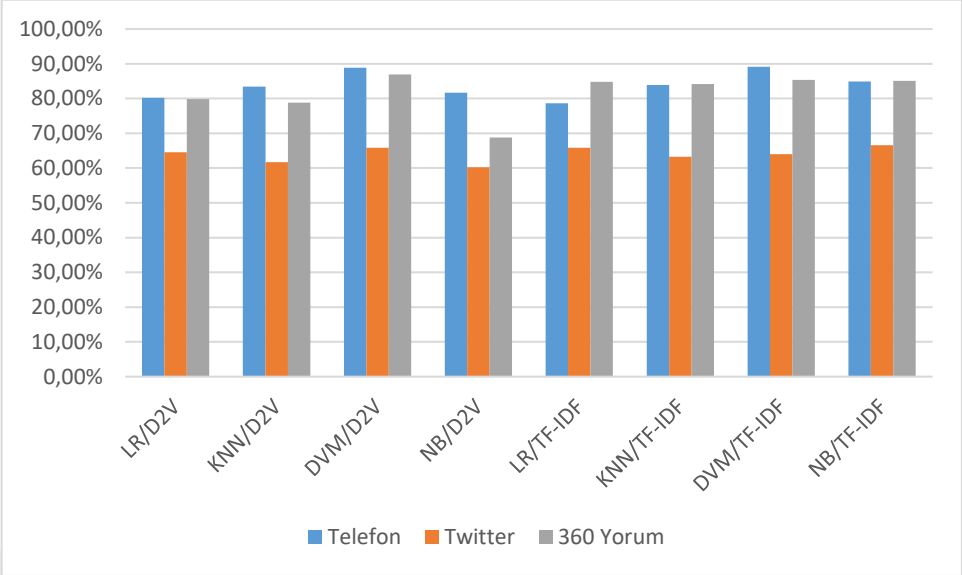
Türkçe veri setleri sonuçları

Bu bölümde Türkçe veri setlerinde yapılan deneylerin sonuçları sunulmuştur. Türkçe veri setlerinde geleneksel makine öğrenmesi sonuçları Tablo 22'de gösterilmiştir. Telefon veri setinde DVM/TF-IDF, Twitter veri setinde NB/TF-IDF, 360 yorum veri setinde DVM-D2V modelleri ile sırasıyla %89,19, %66,60, %86,90 en yüksek sonuçlar elde edilmiştir.

Tablo 22. Türkçe Veri Setleri Geleneksel Makine Öğrenmesi Sonuçları

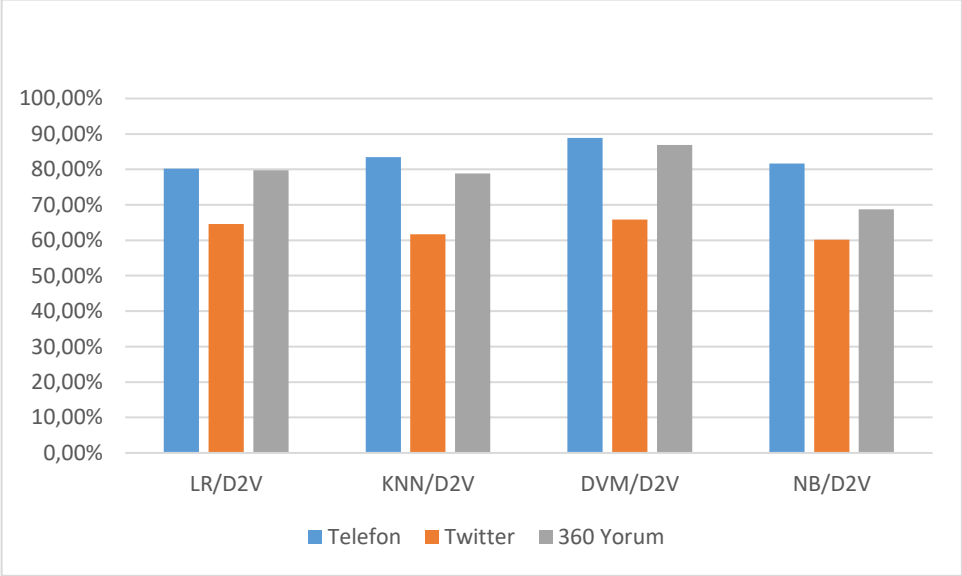
	Telefon	Twitter	360 Yorum
LR/D2V	%80,23	%64,60	%79,80
KNN/D2V	%83,45	%61,73	%78,85
DVM/D2V	%88,87	%65,89	**%86,90**
NB/D2V	%81,69	%60,19	%68,75
LR/TF-IDF	%78,62	%65,86	%84,85
KNN/TF-IDF	%83,89	%63,30	%84,17
DVM/TF-IDF	**%89,19**	%64,00	%85,40
NB/TF-IDF	%84,91	**%66,60**	%85,12

Türkçe veri setleri, geleneksel makine öğrenmesi yaklaşımlarının sonuçları Şekil 16'da grafik olarak gösterilmiştir.



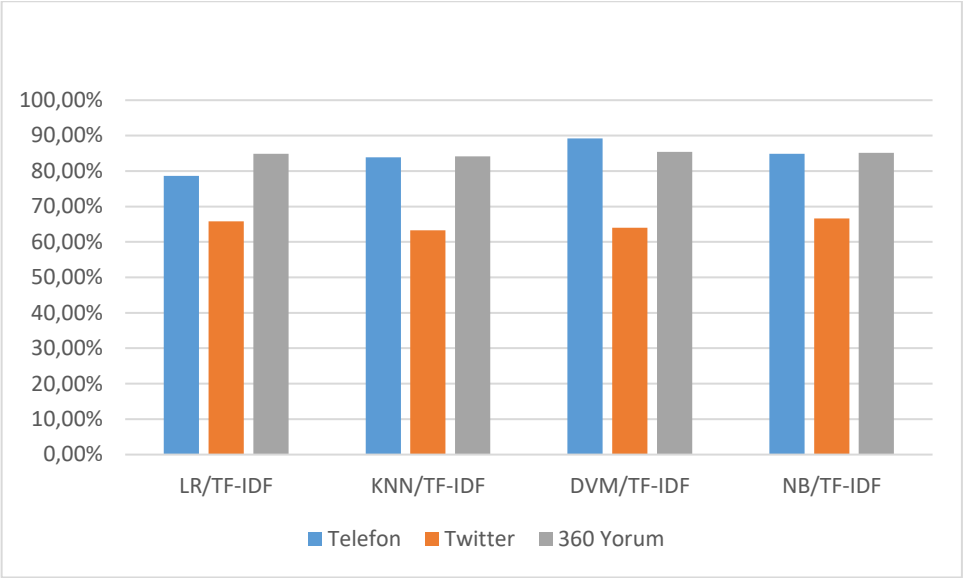
Şekil 16. Türkçe veri setleri geleneksel makine öğrenmesi sonuçları

Türkçe veri setleri, geleneksel makine öğrenmesi Doc2vec model yaklaşımının sonuçları Şekil 17'de grafik olarak gösterilmiştir.



Şekil 17. Doc2vec model Türkçe veri setleri sonuçları

Türkçe veri setleri, geleneksel makine öğrenmesi TF-IDF model yaklaşımının sonuçları Şekil 18'de grafik olarak gösterilmiştir.



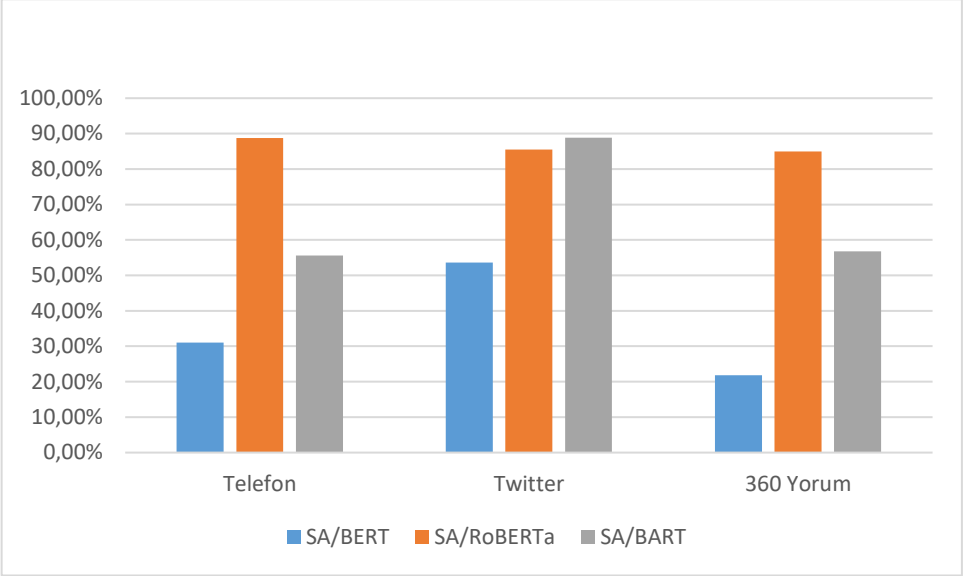
Şekil 18. TF-IDF model Türkçe veri setleri sonuçları

Türkçe veri setlerinde çalıştırılan derin transfer öğrenme algoritmalarından sıfır atış algoritması BERT, RoBERTa ve BART model sonuçları Tablo 23'de gösterilmiştir.

Tablo 23.Türkçe Veri Setleri Derin Transfer Öğrenme Sonuçları

	Telefon	Twitter	360 Yorum
SA/BERT	%31,06	%53,66	%21,83
SA/RoBERTa	**%88,74**	%85,54	**%84,99**
SA/BART	%55,63	**%88,86**	%56,75

Türkçe veri setlerinde; Telefon ve 360 yorum veri setleri için en yüksek doğruluk sonuçları SA/RoBERTa algoritması sırasıyla %88,74, %84,99, Twitter veri seti için SA/BART algoritmasıyla %88,86 olarak elde edilmiştir. Türkçe veri setleri derin transfer öğrenme sonuçları Şekil 19'da grafik olarak gösterilmiştir.



Şekil 19. Türkçe veri setleri derin transfer öğrenme sonuçları

Ayrıca Türkçe veri setlerinde derin transfer öğrenme yöntemleri için farklı etiketler ile testler yapılarak sonuçlar incelenmiştir. 2 sınıftan oluşan Twitter ve telefon yorumları veri setleri için "Pozitif, Negatif" ve "Olumlu, Olumsuz" etiket kümeleriyle, 3 sınıftan oluşan 360 yorumlar veri seti "Pozitif, Negatif, Nötr" ve "Olumlu, Olumsuz, Tarafsız" etiket kümeleriyle BERT, RoBERTa ve BART sıfır atış algoritmaları çalıştırılmıştır. Sonuçlar Tablo 24, Tablo 25 ve Tablo 26'da gösterilmiştir.

Tablo 24. Telefon Yorumları Veri Seti Farklı Etiket Kümeleri İçin Derin Transfer Öğrenme Sonuçları

	Pozitif, Negatif	Olumlu, Olumsuz
SA/BERT	31,06	**32,42**
SA/RoBERTa	88,74	**89,42**
SA/BART	55,63	**58,70**

Tablo 25. Twitter Veri Seti Farklı Etiket Kümeleri İçin Derin Transfer Öğrenme Sonuçları

	Pozitif, Negatif	Olumlu, Olumsuz
SA/BERT	53,66	**55,14**
SA/RoBERTa	85,54	**86,58**
SA/BART	**88,86**	52,84

Tablo 26. 360 yorumlar Veri Seti Farklı Etiket Kümeleri İçin Derin Transfer Öğrenme Sonuçları

	Pozitif, Negatif, Nötr	Olumlu, Olumsuz, Tarafsız
SA/BERT	16,51	**21,83**
SA/RoBERTa	84,72	**84,99**
SA/BART	11,32	**56,75**

Farklı etiket kümeleri için çalıştırılan derin transfer öğrenme yöntemlerinde etiketlerin performansı etkilediği gözlemlenmiştir. Twitter veri seti için SA/BERT ve SA/RoBERTa modellerde "Olumlu, Olumsuz" etiket kümesi daha başarılı performans gösterirken, SA/BART modelde "Pozitif, Negatif" etiket kümesi daha başarılı performans göstermiştir. Telefon veri seti için SA/BERT, SA/RoBERTa ve SA/BART modelde "Olumlu, Olumsuz" etiket kümesi daha başarılı performans göstermiştir. 360 yorum veri seti için SA/BERT, SA/RoBERTa ve SA/BART modelde "Olumlu, Olumsuz, Tarafsız" etiket kümesi daha başarılı performans göstermiştir. Tablo 27'de Türkçe veri setlerinin test veri setlerinde derin transfer öğrenme yöntemlerinin sonuçları verilmiştir.

Tablo 27. Türkçe Test Veri Setleri Derin Transfer Öğrenme Sonuçları

	Pozitif, Negatif / Pozitif, Negatif, Nötr			Olumlu,Olumsuz/ Olumlu,Olumsuz,Tarafsız		
	Telefon	Twitter	360 Yorum	Telefon	Twitter	360 Yorum
SA/BERT	31,06	53,66	16,51	32,42	55,14	21,83
SA/RoBERTa	**88,74**	85,54	**84,72**	**89,42**	**86,58**	**84,99**
SA/BART	55,63	**88,86**	11,32	58,70	52,84	56,75

Tablo 28'de Türkçe veri setlerinde test ve eğitim verileri dahil olmak üzere tüm verilerin bulunduğu veri setlerinde derin transfer öğrenme yöntemleri çalıştırılmıştır.

Tablo 28. Türkçe Veri Setleri Derin Transfer Öğrenme Sonuçları

	Pozitif, Negatif/Pozitif, Negatif, Nötr			Olumlu,Olumsuz/ Olumlu,Olumsuz,Tarafsız		
	Telefon	Twitter	360 Yorum	Telefon	Twitter	360 Yorum
SA/BERT	34,22	56,38	61,92	54,41	56,74	31,98
SA/RoBERTa	**91,39**	85,31	**85,95**	**91,60**	**86,13**	**85,30**
SA/BART	60,86	**88,22**	11,92	59,32	53,69	54,95

Türkçe veri setleri için test veri setleri ve tüm verilerin bulunduğu veri setleri için sıfır atış algoritması çalıştırılmıştır. Doğruluk sonuçları benzer olduğu için derin transfer öğrenmenin veri sayısından bağımsız çalıştığı gözlemlenmiştir.

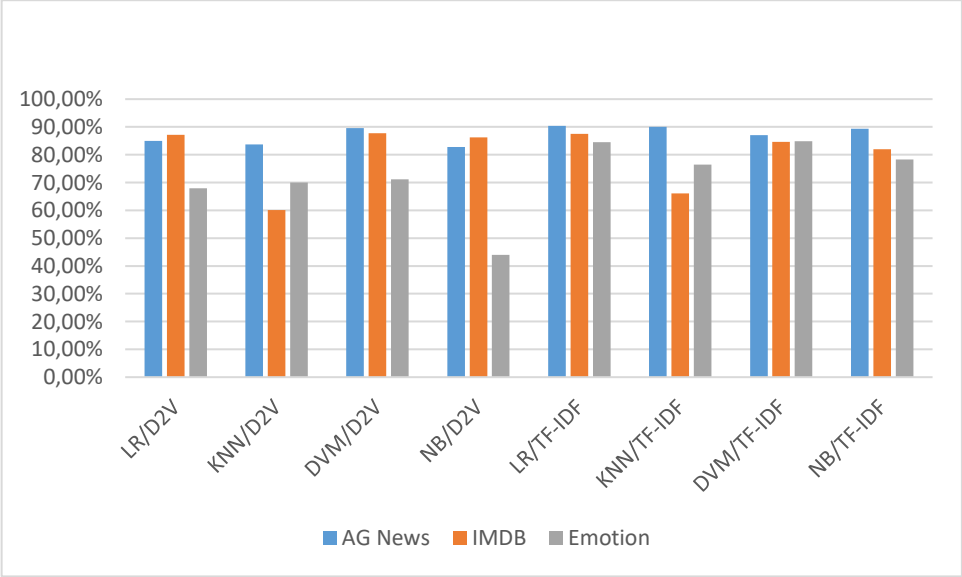
İngilizce veri setleri sonuçları

Bu bölümde İngilizce veri setlerinde yapılan deneylerin sonuçları sunulmuştur. İngilizce veri setlerinde geleneksel makine öğrenmesi sonuçları Tablo 29'da gösterilmiştir. AG News veri setinde LR/TF-IDF, IMDB veri setinde DVM/D2V, emotion veri setinde DVM/TF-IDF modelleri ile sırasıyla %90,39, %87,73, %84,85 en yüksek doğruluk sonuçları elde edilmiştir.

Tablo 29. İngilizce Veri Setleri Geleneksel Makine Öğrenmesi Sonuçları

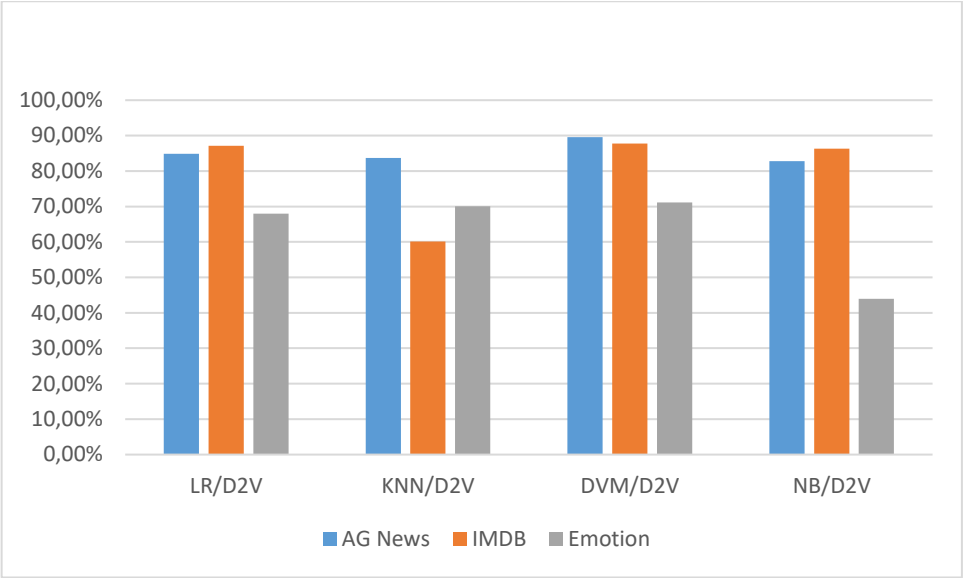
	AG News	IMDB	Emotion
LR/D2V	%84,92	%87,14	%67,95
KNN/D2V	%83,72	%60,14	%70,05
DVM/D2V	%89,57	**%87,73**	%71,15
NB/D2V	%82,77	%86,28	%43,95
LR/TF-IDF	**%90,39**	%87,50	%84,55
KNN/TF-IDF	%90,06	%66,14	%76,45
DVM/TF-IDF	%87,02	%84,64	**%84,85**
NB/TF-IDF	%89,39	%81,94	%78,25

İngilizce veri setleri, geleneksel makine öğrenmesi yaklaşımlarının sonuçları Şekil 20'de grafik olarak gösterilmiştir.



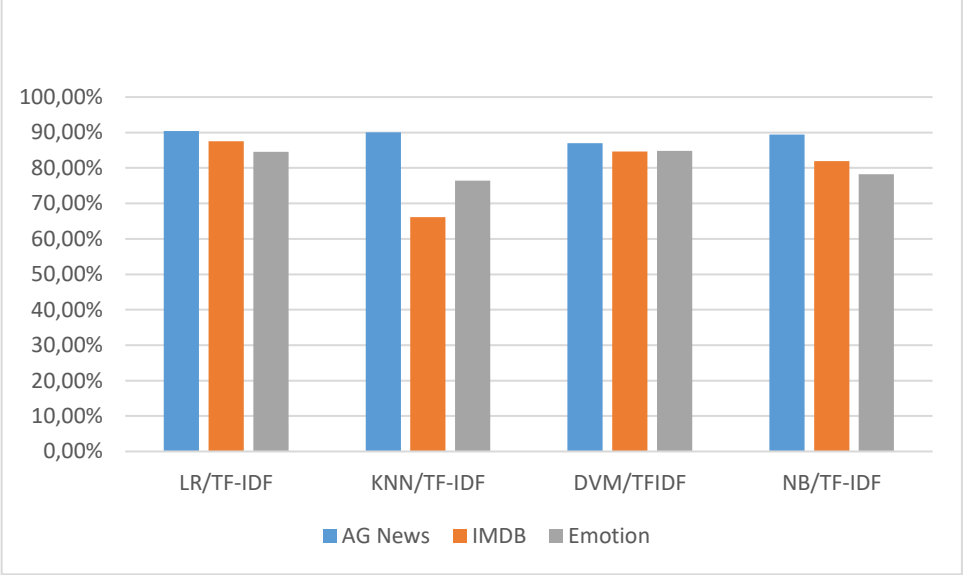
Şekil 20. İngilizce veri setleri geleneksel makine öğrenmesi sonuçları

İngilizce veri setleri, geleneksel makine öğrenmesi Doc2vec model yaklaşımının sonuçları Şekil 21'de grafik olarak gösterilmiştir.



Şekil 21. İngilizce veri setleri Doc2vec model sonuçları

İngilizce veri setleri, geleneksel makine öğrenmesi TF-IDF model yaklaşımının sonuçları Şekil 22'de grafik olarak gösterilmiştir.



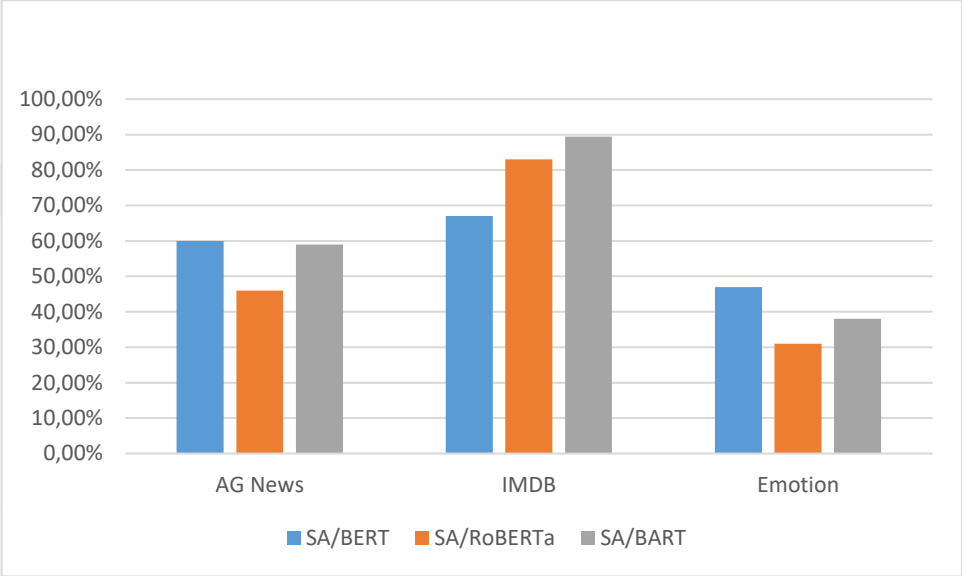
Şekil 22. İngilizce veri setleri TF-IDF Model Sonuçları

İngilizce veri setlerinde çalıştırılan derin transfer öğrenme algoritmalarından sıfır atış algoritması BERT, RoBERTa ve BART model sonuçları Tablo 30'da gösterilmiştir.

Tablo 30. İngilizce Veri Setleri Derin Transfer Öğrenme Sonuçları

	AG News	IMDB	Emotion
SA/BERT	**%60,00**	%67,00	**%47,00**
SA/RoBERTa	%46,00	%83,00	%31,00
SA/BART	%59,00	**%89,45**	%38,00

İngilizce veri setlerinde; AG News ve emotion veri setleri için en yüksek sonuç SA/BERT algoritması sırasıyla %60,00, %47,00, IMDB veri seti için SA/BART algoritmasıyla %89,45 olarak elde edilmiştir. İngilizce veri setleri derin transfer öğrenme sonuçları Şekil 23'de grafik olarak gösterilmiştir. Emotion veri setinde 3 modelinde düşük performans göstermesinin, sınıfların birbirine olan benzerliklerinin yüksek ve sınıf sayısının fazla olmasından kaynaklandığını düşünüyoruz. IMDB veri seti dışında İngilizce veri setleri için derin transfer öğrenme yaklaşımlarında düşük performans elde edilmiştir. AG News ve emotion veri setlerindeki doğruluk oranlarının düşük olmasının, veri setlerindeki sınıf sayısının fazla olmasından kaynaklandığı gözlemlenmiştir.



Şekil 23. İngilizce veri setleri derin transfer öğrenme sonuçları

Sonuçların incelenmesi

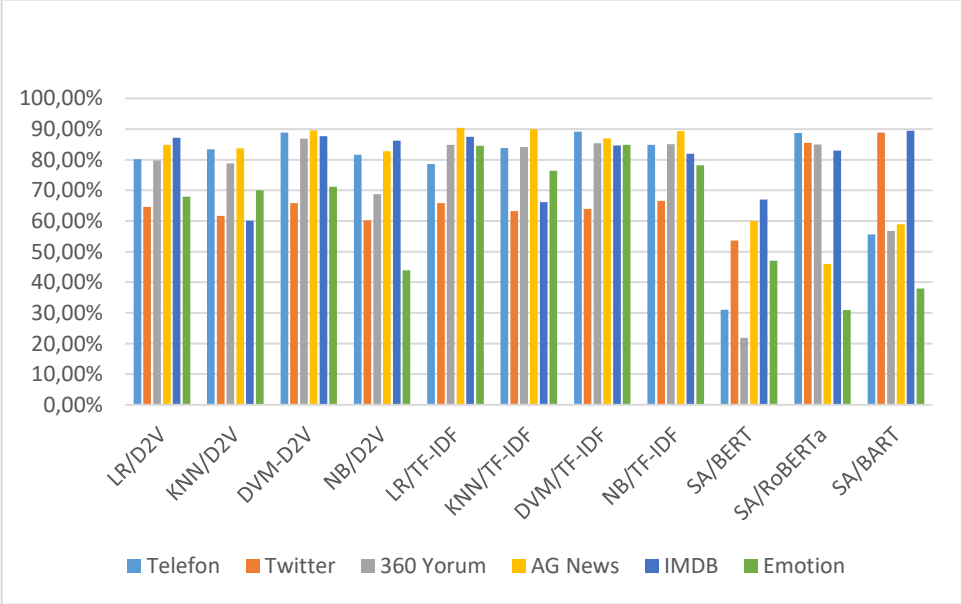
Türkçe ve İngilizce veri setlerinde yapılan deneyler sonucu elde edilen doğruluk değerleri Tablo 31'de gösterilmiştir. Türkçe veri setlerinde; telefon veri setinde en yüksek sonuç DVM/TF-IDF algoritmasında %89,19, Twitter veri setinde en yüksek sonuç SA/BART algoritmasında %88,86, 360 yorum veri setinde en yüksek sonuç DVM/D2V algoritmasında %86,90 olarak elde edilmiştir. İngilizce veri setlerinde; AG News veri setinde en yüksek sonuç LR/TF-IDF algoritmasında %90,39, IMDB veri setinde en yüksek sonuç SA/BART algoritmasında %89,45, emotion veri setinde en yüksek sonuç DVM/TF-IDF algoritmasında %84,85 olarak elde edilmiştir. Twitter ve IMDB veri setleri dışında diğer veri setlerinde geleneksel makine öğrenmesi yöntemleri daha başarılı olurken Twitter ve IMDB veri setlerinde derin transfer öğrenme sıfır atış algoritması BART model daha başarılı performans göstermiştir. Ancak sonuçlar incelendiğinde SA/RoBERTa modelin Türkçe veri setlerinde başarılı

performans gösterdiği gözlemlenmiştir. İngilizce veri setlerinde ise, geleneksel makine öğrenmesi yaklaşımlarının daha başarılı performans gösterdiği gözlemlenmiştir.

Tablo 31. Türkçe ve İngilizce Veri Setleri Algoritma Sonuçları

	Telefon	Twitter	360 Yorum	AG News	IMDB	Emotion
LR/D2V	%80,23	%64,60	%79,80	%84,92	%87,14	%67,95
KNN/D2V	%83,45	%61,73	%78,85	%83,72	%60,14	%70,05
DVM/D2V	%88,87	%65,89	**%86,90**	%89,57	%87,73	%71,15
NB/D2V	%81,69	%60,19	%68,75	%82,77	%86,28	%43,95
LR/TF-IDF	%78,62	%65,86	%84,85	**%90,39**	%87,50	%84,55
KNN/TF-IDF	%83,89	%63,30	%84,17	%90,06	%66,14	%76,45
DVM/TF-IDF	**%89,19**	%64,00	%85,40	%87,02	%84,64	**%84,85**
NB/TF-IDF	%84,91	%66,60	%85,12	%89,39	%81,94	%78,25
SA/BERT	%31,06	%53,66	%21,83	%60,00	%67,00	%47,00
SA/RoBERTa	%88,74	%85,54	%84,99	%46,00	%83,00	%31,00
SA/BART	%55,63	**%88,86**	%56,75	%59,00	**%89,45**	%38,00

Türkçe ve İngilizce veri setlerinde, geleneksel makine öğrenmesi ve derin transfer öğrenme yaklaşımlarının sonuçları Şekil 24'de grafik olarak gösterilmiştir.



Şekil 24. Türkçe ve İngilizce veri setleri sonuçları

Yapılan deneyler sonucu çıkarılan yorumlar aşağıda bulunmaktadır.

- Türkçe veri setlerinde Twitter veri seti dışında geleneksel makine öğrenme yaklaşımları daha başarılı sonuç vermiştir. Twitter veri setinde SA/BART algoritması daha başarılı sonuç vermiştir.

- İngilizce veri setlerinde AG News ve emotion geleneksel makine öğrenmesi yöntemlerinde daha başarılı performans gösterirken, IMDB veri setinde SA/BART algoritması daha başarılı performans göstermiştir.
- Emotion veri setinin derin transfer öğrenme yöntemlerinde düşük performans gösterme sebebinin sınıf sayısının fazla ve etiketlerin birbirlerine ifade olarak yakınlığından kaynaklandığı gözlemlenmiştir.
- SA/RoBERTa algoritmasında Türkçe veri setlerinde %84,99 ve üzeri doğruluk oranı elde edilmiştir. İngilizce veri setleri ile karşılaştırıldığında bu modelin Türkçe veri setlerinde daha başarılı sonuç verdiği gözlemlenmiştir.
- SA/BART algoritmasında IMDB veri setinde %89,45 doğruluk oranı elde edilmiştir. AG News ve emotion veri setleri ile karşılaştırıldığında bu model IMDB veri setinde daha başarılı sonuç vermiştir. Burada diğer veri setlerindeki başarı oranı incelendiğinde IMDB veri setinde sınıf sayısının az olmasının elde edilen başarı oranını etkilediği gözlemlenmiştir.
- Geleneksel makine öğrenmesi yöntemlerinde TF-IDF ve Doc2vec model karşılaştırıldığında 360 Yorum Türkçe veri setinde ve IMDB İngilizce veri setinde Doc2vec model en başarılı sonucu gösterirken diğer Türkçe ve İngilizce veri setlerinde TF-IDF modelin daha başarılı sonuçlar gösterdiği gözlemlenmiştir.
- Veri setleri incelendiğinde Twitter Türkçe veri setinde yapılan deneylerde geleneksel makine öğrenme yaklaşımları daha düşük performans göstermiştir. Bunun sebebinin Twitter veri setinin oluşturulurken metin içerisinde bulunan emojiye göre etiketlenmesinden kaynaklandığını düşünüyoruz.
- Türkçe veri setlerinde farklı etiket kümeleri ile çalıştırılan derin transfer öğrenme yöntemlerinde etiketlerin performansı etkilediği gözlemlenmiştir.
- Türkçe veri setlerinde farklı veri sayısında çalıştırılan derin transfer öğrenme yöntemlerinin veri sayısından etkilenmediği gözlemlenmiştir.
- Derin transfer öğrenme yöntemlerinde veri setlerindeki sınıf sayısı arttığında sonuçların düştüğü gözlemlenmiştir. Derin transfer öğrenme performansını sınıf sayısı etkilemektedir.
- AG News veri seti konu sınıflandırma veri setidir. Diğer veri setleri duygu sınıflandırma veri setleridir. Deney sonuçları incelendiğinde konu sınıflandırma ve duygu sınıflandırma da benzer sonuçlar elde edildiği gözlemlenmiştir.

SONUÇ VE ÖNERİLER

Bu çalışmada elektronik hizmet sektörünün gelişmesi ve dijitalleşmesi ile birlikte insanların yorumlarına, düşüncelerine, duygularına göre hızlı ve doğru analizler yapılarak, kısa zamanda aksiyon planları hazırlamaya yardımcı olmak için geleneksel makine öğrenmesi ve derin transfer öğrenme yaklaşımlarının Türkçe, İngilizce veri setlerindeki performansları incelenmiştir. Yapılan çalışma ile özellikle az veri bulunan durumlarda farklı yaklaşımların performanslarının değerlendirilmesi sağlanmıştır. Doğal dilde etkileşim kuran sistemlerin artması ile birlikte kişisel olarak insanların duygularının analiz edilmesi için geleneksel makine öğrenmesi dışında derin transfer öğrenme yaklaşımlarının performansları değerlendirilmiştir. Geleneksel makine öğrenmesi yöntemleri ile derin transfer öğrenme yaklaşımlarının az sayıda veriye sahip Türkçe ve İngilizce veri setlerindeki başarısı karşılaştırılmıştır.

Yaptığımız çalışmalarda 3 Türkçe 3 İngilizce olmak üzere farklı veri setlerinde geleneksel makine öğrenmesi ve derin transfer öğrenme algoritmaları çalıştırılmıştır. Türkçe veri setlerinden, telefon ve 360 yorum veri setlerinde geleneksel makine öğrenme yaklaşımları daha başarılı olurken, Twitter veri setinde SA/BART derin transfer öğrenme yaklaşımı daha başarılı olmuştur. İngilizce veri setlerinde AG News ve emotion veri setlerinde geleneksel makine öğrenme yöntemleri daha başarılı performans gösterirken, IMDB veri setinde SA/BART model daha başarılı performans göstermiştir. Sınıf sayısı az olan Türkçe veri setlerinde SA-RoBERTa yaklaşımı başarılı performans gösterirken, sınıf sayısı az olan İngilizce veri setinde SA/BART yaklaşımı daha başarılı performans göstermiştir. Kontrollü etiketlenen veri setlerinde geleneksel makine öğrenme yaklaşımları daha başarılı performans göstermiştir. Veri setleri içerisinde Twitter veri setinde yapılan deneylerde geleneksel makine öğrenme yaklaşımları daha düşük performansa sahiptir. Bu durumun Twitter veri setinin emoji içeriğine göre etiketlenmesinden kaynaklandığını düşünüyoruz.

Gelecek çalışmalarımızda doğal dil de toplanan verilerden oluşan veri seti sayısı artırılarak özellikle derin transfer öğrenme yaklaşımlarının performanslarının incelenmesi planlanmaktadır. Veri içeriğinin pozitif ve negatif yönde net ayrıma sahip öznitelikler belirlenerek derin transfer öğrenme yöntemlerinin performanslarının incelenmesi faydalı olacaktır. Ayrıca bir veri setinin çeviri yapılarak farklı dildeki derin transfer öğrenme ve geleneksel makine öğrenme algoritmalarının performansları incelenebilir. Böylelikle çeviri ve orijinal veri setlerindeki derin transfer öğrenme ve geleneksel makine öğrenme yaklaşımlarının farkları gözlemlenebilir. Özellikle bu işlem yapılırken veri sayısı biraz daha fazla olan veri

setleri kullanılabilir. Türkçe veri setlerinde farklı etiket kümeleri ile yapılan çalışma da etiketlerin sonuçları etkilediği gözlemlenmiştir. Etiket kümeleri için yakın anlamlar içeren kelimeler seçilerek testler yapılabilir. Sıfır atış algoritmasında Huggingface modellerinde derin transfer algoritmalarının sunduğu çoklu etiket özelliğiyle birden fazla etiket için doğruluk oranları üzerinde çalışılarak en yüksek doğruluk oranına sahip iki sınıf üzerinden analiz çalışmaları yapılabilir. BERT, RoBERTa veya BART modelde veri etiketleme yapılarak elde edilen veri setinde makine öğrenme yaklaşımlarının performansı incelenebilir. Son yıllarda oluşturulan modeller oldukları için oldukça fazla çalışma alanlarında incelenmesinde fayda olabileceği görüşündeyiz.

KAYNAKLAR

Bowman, S.R., Angeli, G., Potts, C., Manning, C.D., 2015. A large annotated corpus for learning natural language inference,Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 632-642.

Bostan, L.A.M., Klinger, R., 2018. An Analysis of Annotated Corpora for Emotion Classification in Text, Proceedings of the 27th International Conference on Computational Linguistics,2104-2119.

Bansal, T., Jha, R., McCallum, A., 2020. Learning to Few-Shot Learn Across Diverse Natural Language Classification Tasks, Proceedings of the 28th International Conference on Computational Linguistics, 5108-5123.

Birim, A., Erden, M., Arslan, L. M.,2021. Sıfır Atış Türkçe Metin Sınıflandırma. 2021. 29th Signal Processing and Communications Applications Conference(SIU) IEEE.

Budur, E., Özçelik, R., Güngör, T., Potts, C., 2020. Data and Representation for Turkish Natural Language Inference, arXiv preprint arXiv:2004.14963.

Chen, Q., Wang, W., Huang, K., Frans, C.,2021.Zero-shot Text Classification via Knowledge Graph Embedding for Social Media Data.IEEE Internet of Things Journal.1-1

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzman, F., Grave, E., Ott, M., Zettlemoyer, L.,Stoyanov, V.,2020, Unsupervised Cross-lingual Representation Learning at Scale, arXiv.

Çoban, Ö., Özyer, B., Özyer, G.T.,2015. Sentiment analysis for Turkish Twitter feeds. 2015 23nd Signal Processing and Communications Applications Conference (SIU). Y, B., 2019. Turkish Sentiment Analysis.

Çoban, Ö., Özyer, G.T., 2018. Twitter duygu analizinde terim ağırlıklandırma yönteminin etkisi, Pamukkale Üniversitesi Mühendislik Bilişim Dergisi,283-291.

Çelik, S., Özyer, G.T.,2021. Annotated Corpus of Comments and Basic Semantic Analysis.2021 2nd International Conference on Computing and Data Science (CDS).

Devlin ,J., Chang, M.W., Lee, K., Toutanova, K.,2019.BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv.

Geng, R., Li, B., Li, Y., Zhu, X., Jian, P., Sun, J., 2019. Induction Networks for Few-Shot Text Classification, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 3904-3913.

Kotsiantis, S.B., Zaharakis, I.D., Pintelas, P.E., 2007, Machine learning: a review of classification and combining techniques ,Artificial Intelligence Review,159-190.

Kınık, D., 2020, TF-IDF ve Doc2vec Tabanlı Metin Sınıflandırma Sisteminin Başarım Değerinin Ardışık Kelime Grubu Tespiti ile artırılması. Doğuş Üniversitesi,51.

Le, Q., Mikolov, T,2014. Distributed Representations of Sentences and Documents. Proceedings of the 31st International Conference on Machine Learning, PMLR 32(2):1188-1196.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.,2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019.RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv.

Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.,2011.Learning Word Vectors for Sentiment Analysis.Association for Computational Linguistics,142-150.

Pushb, K.P., Srivastava, M.M.,2017. Train Once, Test Anywhere: Zero-Shot Learning For Text Classification, ICLR 2018.

Puri, R., Catanzaro, B., 2019.Zero-shot Text Classification With Generative Language Models, arXiv.

Saravia, E., Liu, T.H., Huang, Y, Wu, J., Chen, Y., 2018.CARER: Contextualized Affect Representations for Emotion Recognition,Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,3687–3697.

Socher,R., Ganjoo,M., Sridhar,H., Bastani,O., Manning,C.D., Ng,A.Y.,2013. Zero-Shot Learning Through Cross-Modal Transfer, NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems (1), 935-943.

Srivastava, S., Labutov, I., Mitchell, T., 2017. Joint Concept Learning and Semantic Parsing from Natural Language Explanations, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 1527-1536.

Srivastava,S.,Labutov,I., Mitchell,T.,2018. Zero-shot Learning of Classifiers from Natural Language Quantification, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), 306-316.

Safali, Y., Nergiz, G., Avaroğlu, E., Doğan, E., 2019. Türkçe Akademik Çalışmaların Doc2Vec Modeli Kullanılarak Derin Öğrenme Tabanlı Sınıflandırılması, 2019 International Artificial Intelligence and Data Processing Symposium (IDAP).

Sanh, V., DEBUT, L., CHAUMOND, J., WOLF, T., 2020. DistilBERT, a distilled version of BERT: smaller,faster, cheaper and lighter. arXiv.

Şeker, S.E., 2015. Metin Madenciliği (Text Mining),YBS Ansiklopedi(2).

Şeker, Ş.E., Yeşilyurt, A., 2017. Metin Madenciliği Yöntemleri ile Twitter Duygu Analizi, YBS Ansiklopedi(4).

Tantuğ, A.C., 2012. Metin Sınıflandırma. Dergipark(5)(2).

Usherwood, P., Smit, S., 2019. Low-Shot Classification: A Comparison of Classical and Deep TransferMachine Learning Approaches, arXiv.

Yin, W., Hay, J., Roth, D., 2019. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 3914-3923.

Yılmaz, B.,2019. Product Comment Dataset, Kaggle.

Yin, W., 2020. Meta-learning for Few-shot Natural Language Processing: A Survey, arXiv.

Ye, Z., Geng, Y., Chen, J., Chen, J., Xu, X., Zheng, S., Wang, F.,Zhang, J., Chen, H.,2020.Zero-shot Text Classification via Reinforced Self-training.Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,3014-3024.

Zhang, J., Lertvittayakumjorn, P., Guo, Y., 2019.Integrating Semantic Knowledge to Tackle Zero-shot Text Classification,Proceedings of NAACL-HLT 2019, 1031-1040.

Zhang, X., Zhao, J., LeCun, Y.,2015,Character-level Convolutional Networks for Text Classification.arXiv.

ÖZGEÇMİŞ

Kişisel Bilgiler	
Adı Soyadı:	Nuray YILDIZLI
Doğum tarihi:	
Doğum Yeri:	
Uyruğu:	
Adres:	
Tel:	
E-mail:	
Eğitim	
Lise:	Şükrüpaşa Lisesi
Lisans:	Ankara Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği
Yüksek lisans:	Atatürk Üniversitesi, FEN Bilimler Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı (2018)
Doktora:	
Yabancı Dil Bilgisi	
İngilizce:	İyi
Üye Olunan Mesleki Kuruluşlar	
Tezden Üretilmiş Yayınlar	