SOSYAL MEDYADA TÜRKÇE NEFRET SÖYLEMLERİNİN VE COVID-19 YORUMLARININ MAKİNE ÖĞRENMESİ, DERİN ÖĞRENME VE BERT TEKNİKLERİ İLE ANALİZİ

DOKTORA TEZİ

HABİBE KARAYİĞİT

MERSİN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ
ANABİLİM DALI

MERSİN
TEMMUZ - 2022

SOSYAL MEDYADA TÜRKÇE NEFRET SÖYLEMLERİNİN VE COVID-19 YORUMLARININ MAKİNE ÖĞRENMESİ, DERİN ÖĞRENME VE BERT TEKNİKLERİ İLE ANALİZİ

DOKTORA TEZİ

HABİBE KARAYİĞİT
ORCID ID: 0000-0002-7518-8699

MERSİN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ
ANA BİLİM DALI

Danışman
Prof. Dr. ALİ AKDAĞLI
ORCID ID: 0000-0003-3312-992X

İkinci Danışman
Doç. Dr. ÇİĞDEM ACI
ORCID ID: 0000-0002-0028-9890

MERSİN
TEMMUZ - 2022

ÖZET

SOSYAL MEDYADA TÜRKÇE NEFRET SÖYLEMLERİNİN VE COVID-19 YORUMLARININ MAKİNE ÖĞRENMESİ, DERİN ÖĞRENME VE BERT TEKNİKLERİ İLE ANALİZİ

Instagram, her kullanıcının bir profilinin olduğu, takipçilerin görüntüleme, beğeni ve yorum yapması için fotoğraf ya da videolar yükleyebildiği ücretsiz bir paylaşım platformudur. Paylaşılan görsellere yönelik küfürlü veya homofobik yorumlar içeren nefret söylemleri küçük düşürücü ve incitici olabilmektedir. Bu tür iletişime maruz kalan kişilerde ciddi psikolojik travmalar meydana gelebilmektedir. Sosyal medya içeriğini incelemek, nefret söylemleri türlerini ayırt etmek için dil modellerine dayalı sınıflandırma sistemlerin geliştirilmesi önemlidir. İngilizce dışındaki dillerde nefret yorumları algılama filtresi geliştirmek daha zor ve zaman alıcıdır.

Bu tez çalışmasında, Türkçe diline yönelik küfürlü ve homofobik söylemleri tespit edebilen yapay zekâ algoritmaları geliştirilmiştir. Bu kapsamda, sırasıyla, Küfürlü Türkçe Yorumlar (Abusive Turkish Comments - ATC) ve Homofobik Küfürlü Türkçe Yorumlar (Homofobik Abusive Turkish Comments - HATC) şeklinde isimlendirilen Türkçe küfür ve homofobi içeren veri kümeleri elde edilmiştir. Yorumlar Türkçe dil kurallarına uygun olarak düzeltilmiş, nefret türlerine göre etiketlenmiş ve eksik yorumlar silinmiştir. Veri kümeleri yorumların dağılımına göre dengelenerek verilerin sınıflandırma başarımları hem orijinal hem de dengelenmiş sürümleri ile elde edilmiştir. Küfürlü söylemlerin sınıflandırma sonuçları baz alındığında Derin Öğrenme (Deep Learning – DL) modellerinden Konvolüsyonel Sinir Ağı (Convolutional Neural Network - CNN) modeli diğer sınıflandırma modellerinden daha iyi bir performansa sahip olduğu gösterilmiştir.

Homofobik söylemlerin tespitinde ise 104 dilde önceden eğitilmiş Çok dilli Dönüştürücü Temelli Çift Yönlü Kodlayıcı Temsilleri (Multilingual Bidirectional Encoder Representations from Transformers - M-BERT) modeli, bu tez çalışmasında da duygu temelli metin sınıflandırma amacıyla kullanılmış ve başarılı sonuçların elde edildiği görülmüştür.

Doktora tez araştırmasının son çalışması olarak Instagram platformunda pandemi döneminde yapılan COVID-19 ile ilgili yorumlardan veri setleri oluşturulmuş ve kişilerin duygu analizine yönelik sınıflandırma modelleri geliştirilmiştir. Bu amaçla, sosyal medyada kullanıcıların pandemi ile ilgili söylemleri olumlu/olumsuz/nötr olarak etiketlenmiş ve bu söylemlerde bir etkileşim olup olmadığı analiz edilmiştir. Bu analiz sonuçlarına göre COVID-19 pandemisinde sosyal medyada Türkçe yorumlar arasında anlamlı bir etkileşim bulunamamıştır. Ayrıca, Türkçe metin verileri kısıtlı olduğu için çalışma kapsamında oluşturulan ATC, HATC ve COVID-19 veri kümeleri araştırma yapacak kişiler için literatüre sunulmuştur.

Sonuç olarak, Türkçe diline yönelik yapay zekâ algoritmaları kullanarak nefret söylemlerinin tespiti ile duygu analinizi yapılmış ve başarılı sonuçlar elde edilmiştir.

Anahtar Kelimeler: Nefret söylemleri, Sosyal medya, Instagram, Küfürlü söylem, Homofobik söylem, Çok dilli BERT, Metin sınıflandırma, Makine öğrenmesi, Duygu analizi, Derin öğrenme, Transfer öğrenme.

Danışman: Prof. Dr. Ali AKDAĞLI, Mersin Üniversitesi, Elektrik-Elektronik Mühendisliği Anabilim Dalı, Mersin.

ABSTRACT

ANALYSIS OF TURKISH HATEFUL DISCOURSES AND COVID-19 COMMENTS IN SOCIAL MEDIA WITH MACHINE LEARNING, DEEP LEARNING AND BERT TECHNIQUES

Instagram is a free sharing platform where each user has a profile and followers can upload photos or videos to view, like and comment. Hate speech containing abusive or homophobic comments towards shared images can be humiliating and hurtful. People who are exposed to this type of communication can experience serious psychological trauma. It is important to develop classification systems based on language models in order to examine social media content and distinguish types of hate speech. Developing a hate comments detection filter is more difficult and time consuming in terms of other languages than English.

In this thesis, the artificial intelligence algorithms that can detect abusive and homophobic discourses for the Turkish language have been developed. In this context, datasets containing Turkish abusive and homophobia, named as Abusive Turkish Comments (ATC) and Homophobic Abusive Turkish Comments (HATC), respectively, were obtained. The comments have been corrected in accordance with Turkish language rules, labeled according to hate types, and missing comments have been deleted. The datasets were labeled, balanced according to the distribution of the comments, and classification performances of the data were obtained with both the original and the balanced versions. Based on the classification results of abusive discourses, it has been shown that the Convolutional Neural Network (CNN) model, one of the Deep Learning (DL) models, has a better performance than other classification models.

In the detection of homophobic discourses, the Multilingual Bidirectional Encoder Representations from Transformers (M-BERT) model, which was pre-trained in 104 languages, was used for the purpose of sentiment-based text classification in this thesis, and it was seen that successful results were obtained.

As the last study of the doctoral thesis research, data sets were created from the comments about COVID-19 made during the pandemic period on the Instagram platform and classification models were developed for sentiment analysis of the users. For this purpose, the discourses of users about the pandemic on social media were labeled as positive/negative/neutral and it was analyzed whether there was an interaction in these discourses. According to the results of this analysis, no significant interaction was found between Turkish comments on social media during the COVID-19. In addition, since the Turkish text data is limited, the ATC, HATC and COVID-19 datasets created within the scope of the study were presented to the literature for those who will conduct research.

As a result, for the Turkish language, the sentiment analysis and the detection of hate speech by using the artificial intelligence algorithms were made and successful results were obtained.

Keywords: Types of hate speech, Social media, Instagram, Abusive speech, Homophobic speech, Multilingual BERT, Text classification, Machine learning, Sentiment analysis, Deep learning, Transfer learning.

Advisor: Prof. Dr. Ali AKDAĞLI, Department of Electrical and Electronics Engineering, University of Mersin, Mersin.

TEŞEKKÜR

Doktora çalışmam boyunca desteğini esirgemeyen, değerli bilgilerinden, akademik tecrübesinden yararlandığım, tez aşamasının planlanmasından yürütülmesine bilimsel anlamda bana yardımcı olup beni yönlendiren, birlikte çalışmaktan mutluluk duyduğum çok değerli danışmanlarım sayın Prof. Dr. Ali Akdağlı ve Doç. Dr. Çiğdem Acı'ya teşekkürlerimi ve saygılarımı sunarım. Doktoramı bitirmem için her zaman yanımda olan, desteğini hiçbir zaman esirgemeyen eşim Mehmet Karayiğit ve kardeşim Doç. Dr. Şule Yücelbaş'a sonsuz teşekkür ederim.

Artık yanımda olmayan ebediyete uğurladığım "canım annem", emeklerinden dolayı sana her daim müteşekkirim ve her daim kalbimde olacaksın.

İÇİNDEKİLER

TABLOLAR DİZİNİ

ŞEKİLLER DİZİNİ

SİMGELER VE KISALTMALAR

Kısaltma/Simge	Tanım
AdaBoost	Adaptive Boosting - Uyarlanabilir Artırma
Adam	Ayarlanabilir Öğrenme Oranı Tahmini - Adaptive Moment Estimation)
API	Application Programming Interface - Uygulama Programlama Arayüzü
ATC	Abusive Turkish Comments - Küfürlü Türkçe Yorumlar
Bagging	Torbalama
Batch size	Küme boyutu
BERT	Bidirectional Encoder Representations from Transformers - Dönüştürücü Temelli Çift Yönlü Kodlayıcı Temsilleri
BiLSTM	Bidirectional LSTM - Çift Yönlü LSTM
Boosting	Artırma
Bootstrap	Yeniden örnekleme
Bot	Sahte makine hesabı
BoW	Bag of Word - Kelime Çantası
BPE	Byte Pair Encoding - Bayt Çifti Kodlaması
CBoW	Continuous Bag of Word - Sürekli Kelime Çantası
CNN	Convolutional Neural Network - Konvolüsyonel Sinir Ağı
Dataset1	COVID-19 Veri Kümesi 1
Dataset2	COVID-19 Veri Kümesi 2
DC	Data Clean - Veri Temizleme
DL	Deep Learning - Derin Öğrenme
Dropout	Seyreltme
DT	Decision Tree - Karar Ağacı
Embedding	Kelime yerleştirme
Entropi	Toplam belirsizlik
Epoch	Eğitim tur sayısı
F1 score	F1 skor
FN	False Negative - Yanlış Negatif
FP	False Positive - Yanlış Pozitif
GCR-NN	Gated Convolutional Recurrent-Neural Networks - Geçitli Konvolüsyonel Tekrarlayan Sinir Ağları
Gradient Boosting	Gradyan Artırma
GloVe	Global Vectors for Word Representation - Global Vektörler
GRU	Gated Recurrent Unit - Geçitli Tekrarlayan Birim
HATC	Homofobik Abusive Turkish Comments - Homofobik Küfürlü Türkçe Yorumlar
IDF	Inverse Document Frequency - Ters Döküman Sıklığı
κ	Cohen kappa katsayısı
kNN	K-Nearest Neighbor - K En Yakın Komşulu
LDA	Latent Dirichlet Allocation - Gizli Dirichlet Tahsisi
LR	Logistic Regression - Lojistik Regresyon
LSTM	Long Short-Term Memory - Uzun Kısa Süreli Bellek
M-BERT	Multilingual BERT - Çok Dilli BERT
Makro F1	Makro ortalamalı F1
Mikro F1	Mikro ortalamalı F1
ML	Machine Learning-Makine Öğrenmesi
NB	Naive Bayes
NDC	Non Data Clean - Veri Temizleme Yok
NLP	Natural Language Processing - Doğal Dil İşleme
One-hot kodlama	Tek sıcak kodlama
Overfitting	Aşırı öğrenme
Position embedding	Pozisyon yerleştirme

Kısaltma/Simge	Tanım
Precision	Kesinlik
Random subspace	Rastgele alt uzay
Recall	Hassasiyet
Recurrent dropout	Tekrarlayan seyreltme
RF	Random Forest - Rasgele Orman
RNN	Recurrent Neural Network - Yinelemeli Sinir Ağı
ROS	Random Over Sampling - Rasgele Aşırı Örnekleme
RUS	Random Under Sampling - Rasgele Alt Örnekleme
SBTC	Sentiment-Based Text Categorization - Duygu Temelli Metin Kategorizasyonu
Segment embedding	Bölüm yerleştirme
Sub-sampling	Alt örnekleme
SVM	Support Vector Machine - Destek Vektör Makinesi
TN	True Negative - Doğru Negatif
TP	True Positive - Doğru Pozitif
Word2Vec	Word to Vector - Kelimeden Vektörlere
TF-IDF	Term Frequency-Inverse Document Frequency - Terim Frekansı - Ters Doküman Frekansı
TF	Term Frequency - Terim Sıklığı
TM	Text Mining - Metin Madenciliği
TML	Traditional Machine Learning - Geleneksel Makine Öğrenmesi
Token embedding	Belirteç yerleştirme
XGBoost	eXtreme Gradient Boosting - Aşırı Gradyan Artırma
YSA	Yapay Sinir Ağları

1. GİRİŞ

Sosyal medya, kullanıcıların metin, bağlantı, fotoğraf veya video gibi dijital içerikleri paylaşmalarını sağlayan mikrobloglardır. Mobil cihazların sıklıkla kullanılmasıyla birlikte sosyal medya kullanımı da hızla artmaya başlamıştır. 2019 istatistiklerine göre Türkiye'de 59.36 milyon internet kullanıcısı bulunmakta ve bu oran nüfusun yaklaşık %72'sini oluşturmaktadır. Türkiye sosyal medya kullanımında dünyada ilk beşte yer almaktadır. İnternet kullanıcılarının büyük bir kısmı (52 milyon) sosyal medya platformunda aktif bir rol oynamaktadır (Statista, 2020). Günümüzde popüler sosyal medya ağları Instagram (Instagram, 2018), Twitter (Twitter, 2022), Facebook (Facebook, 2022), YouTube (YouTube, 2022), WhatsApp (WhatsApp, 2022), Snapchat (Snapchat, 2022), TikTok (TikTok, 2022), LinkedIn (LinkedIn, 2022)'dir.

Ücretsiz fotoğraf ve video paylaşma hizmeti sunan Instagram, yeni ve popüler bir mecra olarak ortaya çıkmıştır. Bu sosyal ağ kullanıcılara bir dizi (filtre ayarlı) fotoğraf ve video aracılığıyla hayatlarındaki önemli anları yakalayabilmeleri ve yakınlarıyla paylaşabilmeleri için anlık hizmet sunar. 6 Ekim 2010'da kurulan Instagram, dünya çapında aylık bir milyardan fazla aktif kullanıcıya sahiptir. Dünya genelinde, 18 ile 29 yaşları arasındaki her üç kişiden ikisi bu sosyal ağı kullanmaktadır (Pewresearch, 2021). Günlük 95 milyon paylaşıma ev sahipliği yapmakta ve paylaşılan içeriklere yorum yapılabilmektedir (Thejas vd., 2019). Ekim 2010 tarihindeki lansmanından bu yana Instagram, 1 milyardan fazla aktif kullanıcıya sahiptir ve şimdiye kadar paylaşılan 40 milyardan fazla fotoğrafı bünyesinde barındırmaktadır (Hu & Kambhampati, 2014). Mart 2020'de Türkiye'de tüm nüfusun %46.7'sini oluşturan 38.870.000 Instagram kullanıcısı bulunmaktadır ve bu ağın kullanımında Türkiye, dünyada 6. sırada yer almaktadır (Statista, 2020).

Twitter, kullanıcılarının 140 karakterle sınırlı kısa mesajlar oluşturmasına ve göndermesine olanak tanıyan başka bir popüler sosyal ağdır. Dünyada olan olaylara karşı kişilerin düşüncelerini ifade etmek için sıklıkla kullanılması, araştırmacıların özellikle Metin Madenciliği (Text Mining - TM) alanında çalışmalarını bu ağda yapmasını sağlamıştır. Erişebilirliğinin artmasıyla birlikte kullanım sıklığı artmış ve günde yaklaşık 10 milyon tiviti tivitleyen 3.7 milyondan fazla aktif kullanıcısı ile en popüler sosyal ağlardan biri olmuştur (Hu & Kambhampati, 2014). Twitter seçim sonuçlarını tahmin etme, siyasi olayları ve tartışmaları paylaşma, suç tahmini, terör faaliyetlerini izleme ve nefret söylemlerini tespit etme gibi önemli bir analitik araç haline gelmiştir (Wang vd., 2012).

Facebook, Ocak 2021'de 2.8 milyar aktif kullanıcıya ve 1.8 milyar tekil ziyaretçiye sahip popüler bir başka sosyal ağ sitesidir (Kemp, 2021). Bu platform farklı coğrafi bölgelerden ve farklı kültürlerden insanlara arkadaşlarıyla iletişim kurmak için mesaj, durum güncellemesi, fotoğraf, video vb. öğeleri kullanma fırsatı sunar. 2009 yılında tüm sosyal ağlar için bir başlangıç kabul edilen kullanıcıları "like" özelliği ile tanıştırmış olan Facebook bünyesinde birçok teknoloji şirketinin faaliyet

gösterdiği büyük bir yapıdır. Bu yapının en temel amacı diğer sosyal ağlar gibi insanlar arasında iletişimi sağlamaktır.

Her ne kadar sosyal medya platformları sosyalleşme ve iletişim için geliştirilen internet fenomenleri olsalar da şu anda veri elde etmek için sıklıkla kullanılan araçlara dönüşmüştür. Sosyal medyada araştırmacıların farklı alanlarda kullanabilecekleri işlenmemiş veri yığınları ortaya çıkmıştır. Sosyal medya verileri yeni sorunlara çözüm bulmak için yaygın kullanılan kaynaklar haline gelmiştir. Sosyal medya verilerini toplama ve verileri saklama işlemi her sosyal ağın kendi API'si (Application Programming Interface - Uygulama Programlama Arayüzü) ile elde edilmektedir (Kaya, 2021).

Duygu analizi, sosyal medyada kişilerin duygu ve düşüncelerini belirlemek için sıklıkla kullanılmaktadır. Duygu analizi, kontrollü bir NLP (Natural Language Processing - Doğal Dil İşleme) problemidir ve en basit tanımıyla, gönderilen yorumların pozitif, negatif veya nötr duyguları ifade edip etmediğini belirleyen bir metin sınıflandırma işlemidir (Mossie & Wang, 2020).

TM'nin bir alt dalı olan duygu analizi, öznel bilgileri çıkarmak için insanların görüş ve tutumlarını araştıran bir dizi tekniktir. Sosyal medyanın duygu analizi alanında kullanımı, ilk başlarda sosyal medya içeriklerinden elde edilen verilerle kullanıcının belirli bir ürün ya da hizmete ilişkin algısını belirlemek içindi. Sosyal medyada kullanıcı içeriğine çevrimiçi ücretsiz olarak erişebilir olması duygu analizi çalışmalarının çok sayıda işletme tarafından benimsenmesini kolaylaştırmıştır (He vd., 2013). Günümüzde duygu analizi ile nefret söylemi tespiti, fikir çıkarımları ve piyasa tahminleri gibi daha birçok alanda sosyal medya analizleri yapılmaktadır (Mäntylä vd., 2018).

Sosyal medyada bir görsele ya da videoya ait çoğunlukla güncel olan meselelerle ilgili tartışmalar yaygınlaşmıştır. Bundan dolayı da bu içeriklerin duygu analizinin çıkarılması kişilerin bakış açılarını ve o toplum ya da topluluğa ait fikirlerin öğrenilmesi için önemlidir. Sosyal medyada tartışmalarda yapılan yorumlar her zaman olumlu değildir. Yorumlar içerisinde siber şiddet içeren farklı nefret söylemleri de yaygın olarak kullanılmaktadır.

Sosyal medya, insanlara duygularını özgürce ifade etmeleri için özgür bir platform sunar. Kullanıcılar, sosyal medyadaki gönderilere düşüncelerini paylaşabilmekte, yayabilmekte ve yorum yazabilmektedir (Jenkins, 2002). Sosyal medyada insanlara yapılan yapıcı yorumlar olduğu kadar rahatsız edici nefret söylemleri de vardır. Sosyal medyada her gün çok sayıda paylaşım veya etkileşim yaşanması ve sosyal medyanın merkezi olmayan yapısı nefret söyleminin artmasının en önemli nedenleri arasında yer almaktadır (Banks, 2014; Bilge, 2016; Çomu & Binark, 2012). Sosyal medyanın sık kullanımı ile toplumda karşılaşılan ötekileştirme söylemleri bu platformlara taşınmıştır (Cammaerts, 2010). Nefret söylemi sosyal medyada önemli olayların (başkanlık seçimleri, terör saldırıları, savaş durumları, hastalık veya salgınlar gibi) yaşanması durumunda daha da artmaktadır (Cruz vd., 2022). Değişen sosyal koşullar karşısında kültür, ırk, din ve cinsiyet farklılığına saygı gösterme ile ifade özgürlüğü arasında ki dengenin sağlanması çok güç olmaktadır. Nefret söylemi kamu güvenliğini tehlikeye sokabilecek kadar hızlı bir şekilde kişiler ya da gruplar üzerinde şiddeti tetikleyebilmektedir.

Sosyal ağlarda yayınlanan içeriklerin fazla yani çok boyutlu olması ve bağlama duyarlı doğası nedeniyle izlenmesi ve kontrol edilmesi zordur. Denetimin yeterince etkin olmamasından dolayı kullanıcılar nefret söylemine veya siber zorbalığa uğraması olasıdır. İnternet bağlantısına ulaşmanın kolaylaştığı çağımızda nefret söylemi hızlıca yaygınlaşmaktadır. Nefret söyleminde bulunan kullanıcılardan bir bölümü ceza almayacakları düşündükleri için sahte profillerle bu fiili gerçekleştirmektedirler. İnsanlar gibi davranan bot hesaplar (sahte makine hesapları) her ne kadar yazılma amaçları farklı olsa da belli bir zaman sonra nefret söylemini yayan bir algoritmaya dönüşebilmektedir (Cruz vd., 2022). Örneğin; Twitter sosyal ağında kullanıcıların yaklaşık %8.5'nin bot hesap olduğu iddia edilmektedir (Subrahmanian vd., 2016).

Tablo 1.1. Dereceye dayalı nefret söylemi sınıfları (Sharma & Shrivastava, 2018).

1. Sınıf	2. Sınıf	3. Sınıf
Irkçılık	Gözdağı	Alay
Cinsiyetçilik	Tehdit/Korku	Trolleme
Homofobik	İhlâl	İroni
Küfür (dışkısal küfür, cinsel içerikli küfür vb.)	Düşmanlık	Aşağılama
Ofansif	Suçlamak	
Saldırgan	Rahatsızlık verme	
Kadın düşmanlığı (misogyny)		
Toksik		
Taciz		

Nefret söylemleri tür olarak çok geniş bir alana sahiptir. Bu bağlamda yukarıda verilen Tablo 1.1'e göre üç sınıfa ayrılabilir (Sharma & Shrivastava, 2018). 1. Sınıftan 3. Sınıfa doğru gidildikçe nefretin şiddeti azalmaktadır. 1. Sınıf nefret söylemleri yazışmanın ötesinde şiddet içeren eylemleri ve ağır aşağılamayı teşvik eder. 2. Sınıfta nefret söylemleri, duyguları incitir ancak 1. Sınıftaki gibi şiddetli bir tepki uyandıracak derecede değildir. Bir ideoloji veya topluluktan (gruptan) ziyade bir bireye hitap ederken özel alana izinsiz girme durumu vardır ve yorumlar kışkırtıcıdır. 3. Sınıf nefret söyleminde ise çoğunlukla tekil bir aşağılama olayı vardır ve hafif kışkırtıcıdır. Yazan kişiden alıcıya bağlamı olmayan saygısız ve kötü kelimeler kullanılır. 3. Sınıf yorumların bağlamı trolleme, ironi ve aşağılama etrafında döner.

Son yıllarda nefret tespiti ile alakalı farklı nefret söylemi tespit sınıflarının irdelendiği artan sayıda araştırma görülmüştür. Bu tez çalışmasında ise 1. Sınıf nefret söylemleri ile ilgili küfür ve homofobi türlerini ele alan Türkçe NLP araştırması yapılmıştır. 1. Sınıf nefret söylemlerinden küfürlü ifadeler, insanların öfke, tutku, korku gibi en temel duygularını ifade etmek için kullanılan ve duygusal acıya neden olma, şiddet ve çatışmayı kışkırtma gibi kapasitesi yüksek olan kelimelerdir. Türkçe küfürlerin içerikleri genellikle cinsellik, cinsel yönelim ve dışkılama ile ilgilidir. Küfürlü kelimenin

etkisi, küfre uğrayan kişi için ne kadar tabu olduğu ile ilgilidir. Kullanımı istenmeyen bir iletişim şekli olarak görülse de özellikle sosyal medyada sıklıkla kullanılmaktadır.

Homofobi, kelime olarak, sosyal ve tıbbi nedenlerle farklı cinsel yönelimlere sahip kişilere yönelik bir küçümseme ve önyargı durumudur (Oxford, 2022). Farklı cinsel yönelimlerin ötekileştirildiği cinsel kimlik temelli nefret söylemidir diye de tanımlanabilir (Dondurucu, 2018). Sosyal medyada bu tür ifadelere maruz kalan kişiler her zaman cinsel yönelimleri veya davranışlarından dolayı hakarete uğramazlar. Örneğin; futbolcular maç kaybettikten sonra taraftarları tarafından homofobik söylemlere maruz bırakılabilmektedir (Karayigit vd., 2021). Aldatan, ahlaksız, güvenilmez, hain, kaba, sahtekâr, karaktersiz ve geveze anlamlarında da kullanılan homofobik nefret söylemleri ayrımcılık, değersizleştirme ve düşmanlaştırma davranışıdır.

Nefret söylemiyle ötekileştirmenin diğer biçimleri ırk, etnik köken, din, cinsiyet, engellilik veya hastalık açısından olabilmektedir (Mossie & Wang, 2020). Cinsiyete yönelik nefret söylemi genel olarak kadınlara yapılan cinsel içerikli küfürlü nefret söylemlerine dönüşmektedir. Geçmişten günümüze kadar gelen ataerkil toplumun baskıcı söylemlerinden dolayı kadınlar toplumun bazı kesimlerinde arka plana itilmiştir. Topluma hâkim olan insanların hafızasında yer eden kanıksanmış bazı sözlerde bile kadınlara olan olumsuz bakış açısının etkileri görülebilmektedir. Bu nedenle kadınlara yönelik nefret ve küçümseme gibi nefret söylemlerine sosyal medyada sıklıkla karşılaşılabilmektedir.

Irkçılıkta temel düşünce bir ırkın diğerinden üstün olduğunu düşünülmesidir. Irka ya da etnik kökene yapılan nefret söylemleri kişilerin bulundukları ortama sonradan geldikleri için istenmeme durumundan kaynaklanabilir ve kişilere ya da gruplara belli bir ırka ait olmalarından dolayı duyulan nefreti ifade eder. Sosyal medyada etnik kökenlere önyargı içeren hakaret ifadelerinin artması nefret söylemlerinin yaygınlaşmasına katkıda bulunmaktadır. Etnik kökeni hedef alan aşağılayıcı dil, literatürde nefret söylemi kavramı içerisinde yer alır ve farklı tanımları yapılabilmektedir (Pronoza vd., 2021). Sosyal hayatın olumsuz dinamiklerinden olan ırkçılık dijitalleşmeyle birlikte sosyal medyaya da taşınarak araştırmacılar tarafından siber ırkçılık şeklinde kuramsallaştırılmıştır.

İnanca yönelik nefret söylemi de sosyal medyada sıklıkla karşılaşılan bir durumdur. Dini değerlere küfretme, hakaret ve dinleri nedeniyle insanlar hakkında kötü ifadeler kullanma inanca yönelik nefret söyleminin içeriğini oluşturmaktadır. Sosyal medyada belirli bir dini inanca mensup olanların ayrımcılığa tabi tutulması yönünde kışkırtıcılık söylemleri yaygın bir şekilde görülmektedir. Dini görüş veya inançlar bağlamında başkalarına haksız yere saldırıda bulunan ve herhangi bir kamuoyu tartışmasına neden olan nefret söylemleri, bazı ülkelerin yasalarında farklı biçimlerde cezalandırılmaktadır.

Sosyal medyada nefret söylemlerine maruz kalan kişilerde görülen zararlı psikolojik etkiler son birkaç yıldır daha belirginleşmeye başlamıştır. Bu konu çeşitli kuruluşlar ve araştırmacıların çalışma alanı haline gelmiştir. Sosyal medyadaki nefret söylemini engellemek için yapılan tüm çabalara rağmen

kullanıcılar hala bu söylemlerden etkilenmektedir. Nefret söylemine maruz kalan kişilerde duygusal durum bozukluğundan, psikolojik ve fiziksel olumsuz etkilere kadar birçok olumsuz sonuç görülmektedir (Van Royen vd., 2015).

Sosyal medyadaki yorumlar kullanıcılar için her zaman olumlu değildir; zarar verme potansiyeli sıklıkla görülmektedir (Tang & Dalzell, 2019). Ayrım gözetmeyen argo, küfürlü dil ve aşağılama gibi sosyal medyadaki taciz edici mesajlar sadece bir mesaj olarak görülmemelidir. Nefret söylemleri Tablo 1.2'de görüldüğü gibi siber şiddete açılan şiddetli ve acımasız bir araçtır (Lee vd., 2018). Twitter, Facebook ve YouTube gibi sosyal medya şirketleri yıllardır bu sorunla mücadele etmektedir. İnsan kaynakları da dâhil olmak üzere siber şiddete karşı alınan önlemlere her yıl yüz milyonlarca avronun harcandığı tahmin edilmektedir (Zhang vd., 2018). Siber zorbalığa uğrayan öğrencilerin başarısızlıkları, devamsızlıkları ve okulu bırakma davranışları arasında bazı bağlantılar olduğu bildirilmiştir. Siber zorbalık, depresyon, düşük benlik saygısı, intihar düşüncesi ve hatta intihar gibi potansiyel olarak yıkıcı psikolojik etkilere sahiptir (Van Royen vd., 2015). Türkiye'de her beş kişiden biri siber zorbalığa maruz kalmaktadır ve bu oran dünya ortalaması ile benzerdir. Siber zorbalık için en yaygın kullanılan yöntemler mesaj göndermek ve yorum yazmaktır (BBC_News, 2020). Bu riskleri mümkün olduğunca önceden öngörerek tespit etmek ve önlemek gerekir.

Tablo 1.2. Nefret söylemleri kategorileri için örnek yorumlar listesi (Sharma & Shrivastava, 2018).

Sınıf	Örnek Yorumlar
1. Sınıf	1. Feministlerden nefret ediyorum ve hepsi tecavüze uğramalı ve yakılarak öldürülmeli. 2. Bütün Müslümanlar ibnedir ve domuz gibi katledilmelidir. 3. Artık Trump başkan olduğuna göre seni ve bulabildiğim tüm siyahları vuracağım.
2. Sınıf	1. Sen sadece kendine saygısı olmayan mal bir ilgi tenekesisin. 2. Seni pislik seni mahvedeceğim. 3. Senin fikrinin ne önemi var vasat herif, fikrini bir yerine sok ve bir maymun gibi dans et.
3. Sınıf	1. Umarım hepiniz harika bir hafta sonu geçirirsiniz. Sen hariç, Lisa Kudrow. 2. Benedict Cumberbatch'i çekici buluyorsanız, doğrudan kakaya bakmaktan da zevk alacağınızı tahmin ediyorum. 3. Afrika'ya gitmek. Umarım AIDS kapmam. Ben beyazım.

Nefret söyleminin etkileri farklı travmatik olayların etkileriyle karşılaştırılabilir ve sosyal kimliği etkilediğinden büyük topluluklara ulaşabilir ve daha yıkıcıdır (Quandt vd., 2022). Bu söylemler toplumdaki kolektif refahı bozabilir, duyarsızlaştırma yoluyla kişileri değersizleştirir ve önyargıyı

artırır. Çoğunluk tarafından nefret dili geliştirildiğinde kolektif esenliğe zarar vermekte ve nesiller arası ilişkiler olumsuz etkilenmektedir.

Nefret söylemleri, nefrete maruz kalan kişi veya grupların taciz edilmesine, küçük düşürülmesine, sindirilmesine ve gaddarlığa yol açmaktadır. Yine nefrete maruz kalan kişi veya gruplarda susma ve kendini ifade etmeyi reddetme durumları görülmektedir. Depresyon ve intihar eğilimi, bu duruma maruz kalan bireylerde tanımlanan diğer davranışlardır (Kumar vd., 2021). Agresif ve saldırgan nefret söylemleri, toplumsal şiddeti tetikler, akıl sağlığı sorunlarını artırmakta ve hatta intiharı teşvik edebilecek kadar psikolojik etkiye sebep olmaktadır. Böylesine tehlikeli durumlarla başa çıkmak için sosyal medyada nefret söyleminin erken tespiti ve engellenmesi önemlidir. Sosyal medya üzerinden yapılsa bile söylemlerin eyleme dönüşmeden kontrol edilmesi gerekmektedir. Bu nedenle, insanları rahatsız eden uygunsuz içeriği tespit etmek ve önlemek için otomatik dil modelleri geliştirilmelidir (Kumar vd., 2021).

Sosyal ağlarda küfürlü konuşmaların manuel analizi zor ve zaman alıcıdır. Duygu analizi, sosyal medyadaki yorumlardan kullanıcıların düşüncelerini, duygularını çıkarmakta ve kutuplarını ayırt etmektedir (Hemmatian & Sohrabi, 2019). Duygu Temelli Metin Kategorizasyonu (Sentiment-Based Text Categorization - SBTC), önceden etiketlenmiş veri kümesini kullanarak bir model oluşturmakta ve etiketlenmemiş verileri doğru kategoriye atamaya çalışmaktadır. Etiketlere bakarak siber zorbalığı olumlu ve olumsuz olarak sınıflandırmaya yardımcı olmaktadır (Sebastiani, 2002). Nefret söyleminin otomatik olarak algılanması ve önlenmesi sosyal medyadan elde edilen verinin doğru bir şekilde analiz edilmesini sağlayan gürbüz, verimli ve akıllı sistemler gerektirmektedir.

Instagram ve Facebook gibi sosyal ağlar, nefret söylemleriyle mücadele etmek için veri tabanlarındaki nefret söylemine benzeyen yorumları silmektedir. Yorumları silmek veya engellemek yapılan olayın suç olmadığı anlamına gelmemektedir. Hakaretin boyutu cezalandırılabilmekte ve güvenlik güçleri tarafından takip edilebilmektedir. Örneğin; saldırganlık, tehdit ve ırkçılık gibi toplum huzurunu etkileyen söylemler yasal açıdan cezai takip altındadır. Ancak manuel izleme pahalı ve zaman alıcıdır. Olumsuz dili otomatik olarak algılayan ve analiz eden bir sistem geliştirmek çok önemlidir (Kasakowskij vd., 2020).

Sosyal medya aracılığıyla nefret söylemlerinin fikir özgürlüğü sınırlarını aşması ve hızlı bir şekilde toplumu etkileyecek şekilde yayılması hükümetlerin ve farklı kuruluşların dikkatini çekmektedir. Bu bağlamda Avrupa Komisyonu, Facebook, Microsoft, Twitter ve YouTube gibi sosyal ağların ilgili yöneticileriyle nefret söyleminin yayılmasını erken tespit ve önlemek için belirli maddeler üzerinde anlaşmıştır. Uzlaşılan maddeler her ne kadar nefret söylemi olaylarını azaltsa da toksik söylemlerin tamamen ortadan kaldırılması için daha fazla çalışma yapılması gerekmektedir (Charitidis vd., 2020).

Facebook ve Twitter gibi bazı sosyal ağlarda kullanıcıların herhangi olumsuz söylemi şikâyet edebileceği mekanizmalar vardır ve kullanıcıların beyanı esas alınarak kullanıcı hesapları askıya

alınmaktadır (Fortuna vd., 2021). Ama bu durum bazen zararsız hesapların yanlış değerlendirilebilmesine ve rahatsız edici bazı hesapların ise bildirilmediği durumlarda hesabın devam etmesine neden olmaktadır. Nefret söylemlerinin içeriğinin yanlış değerlendirilmesi ve otomatik tespitinde yanılsamalar olabilmekte ve bu durum güvenlik mekanizmalarının sorgulanmasına neden olmaktadır (Arango vd., 2019). Nefret söylemlerine maruz kalma durumu her anlamda ciddi olumsuz sonuçlar doğurabileceğinden, dünyada büyük çapta kullanıma sahip sosyal medyadaki olumsuz söylemlerin tespit ve analizini yapmak için güvenilir modeller gerekmektedir.

Nefret söylemlerinin tespiti çalışmalarında güvenilir ve gürbüz modellerin oluşturulabilmesi için veri kümelerine özellik seçimi ve çıkarımı uygulanması işlemi en önemli adımlardan biridir. Nitelik seçimi ve çıkarımı veri kümelerinin boyutunu azaltmak ve verileri vektörleştirmek için kullanılmaktadır. Yüksek boyutlu verilerin özellik uzayında bulunması önemli bir sorundur. Yüksek boyutluluğun azaltılması en iyi özellik alt kümelerinin bulunmasına bağlıdır. Farklı diller ile ifade edilen kelimeler yapısı itibariyle bilgiyle doludur. Küçük miktarlardaki metni bile temsil etmek için oldukça büyük miktarda veriye ihtiyaç duyulmaktadır. Vektörler bu format için kullanılan matematiksel modellerdir. Nitelik seçimi ve nitelik çıkarımı işleminde en iyi vektör seçimi yapabilmek seyrek vektör gösterimi ve yoğun vektör gösterimi sıklıkla kullanılmaktadır.

Duygu analizi alanında Makine Öğrenmesi (Machine Learning - ML) tekniklerini kullanarak sosyal medya yorumları gibi öznel bir yapıdan bilgi çıkarmaya odaklanan birçok çalışma yapılmıştır (Al-Garadi vd., 2016; Balakrishnan vd., 2020; Burnap & Matthew, 2016; Chatzakou vd., 2019; Hosseinmardi vd., 2015; Ibrohim & Budi, 2018). Geleneksel Makine Öğrenmesi (Traditional Machine Learning - TML) tabanlı teknikler yaygın olarak kullanılmasına ve oldukça başarılı performans göstermesine rağmen nitelik tanımının çok çaba gerektirmesi ve manuel olarak tanımlanan niteliklere güçlü bir şekilde bağlı olması bu modellerin zorluklarındandır.

ML modellerinde sıklıkla kullanılan topluluk metotları, zayıf öğrenen öğrenicilerin tahminlerini birleştirerek güçlü bir öğrenici oluşturmaya çalışmaktadır. Birçok topluluk modellerinin temel amacı zayıf tahmin edicileri kümülatif olarak eğitmektir. Bu modellerin ağaç yapısında önceki ağaçtan gelen hata oranı bir sonraki ağaçta azaltılmaya çalışılmaktadır (Kilincer vd., 2022).

Öznitelik seçimi için harcanan çabayı azaltabileceği ve nispeten yüksek performans elde edebileceği için Derin Öğrenme (Deep Learning - DL) teknikleri son zamanlarda büyük dikkat çekmektedir (Kim & Jeong, 2019a). DL modeli, verileri farklı soyutlamalarla temsil etmek için önemli sayıda gizli katman ve nöron kullanmaktadır. DL modelleri büyük veri kümeleriyle verimli ve etkili bir şekilde çalışmaktadır. Birçok gizli katmana sahip Konvolüsyonel Sinir Ağı (Convolutional Neural Network - CNN) metodu buna bir örnektir (Alayba vd., 2017). CNN, son zamanlarda sosyal medya yorumlarındaki duyguları sınıflandırmak için de kullanılmıştır (Ayata vd., 2017; Kim & Jeong, 2019b; Mozafari vd., 2020; Park & Fung, 2017; Shushkevich & Cardiff, 2019; Zhang & Luo, 2018).

Uzun Kısa Süreli Bellek (Long Short-Term Memory - LSTM), Geçitli Tekrarlayan Birim (Gated Recurrent Unit - GRU), Çift Yönlü LSTM (Bidirectional LSTM - BiLSTM) gibi Yinelemeli Sinir Ağı (Recurrent Neural Network - RNN) modelleri, kelime dizilerini sınıflandırma, etiketleme ve oluşturma gibi çeşitli görevlerde yaygın olarak kullanılmaktadır (Liu & Qi, 2021). Belirtilen RNN modelleri, önceki adımın çıktısının mevcut adıma giriş verisi olarak beslendiği bir sinir ağıdır. Girdi (input) verileri zaman serisine göre işlenmekte ve elde edilen çıktı (output) bir sonraki durum için girdi olarak kullanılmaktadır (Liu & Qi, 2021). Geçitli Konvolüsyonel Tekrarlayan Sinir Ağları (Gated Convolutional Recurrent-Neural Networks - GCR-NN) hibrit DL modeli de metin sınıflandırma ve duygu analizi gibi görevlerde başarılı sonuçlar vermektedir (Lee, Rustam, Washington, vd., 2022).

Dönüştürücü Temelli Çift Yönlü Kodlayıcı Temsilleri (Bidirectional Encoder Representations from Transformers - BERT) modeli, çift yönlü transformatör mimarisini uygulayan denetimsiz derin çift yönlü bir sinir ağıdır. BERT tabanlı aktarım öğrenme yaklaşımı, sınıflandırma performansının artmasına ve eğitim süresinin azalmasına yol açtığı için nefret sınıflandırma çalışmalarında sıklıkla kullanılmaya başlanmıştır (Sharma vd., 2022). Transfer öğrenme yaklaşımı ayrıca önceden eğitilmiş bir modelle sınırlı etiketli verilerden etkili öğrenme sağlar. Önceden eğitilmiş bir dil modeli, az etiketli veri kaynaklarında bile mevcut dili anlamayı kolaylaştırır.

Bu tez çalışmasının amacı, Türkçe küfürlü, homofobik ve diğer nefret söylemi türlerini tespit edip önlemek için başarılı sınıflandırma modelleri geliştirmektir. Bu amaçla, Instagram'dan toplanan veri kümeleri (Küfürlü Türkçe Yorumlar (Abusive Turkish Comments - ATC) ve Homofobik Küfürlü Türkçe Yorumlar (Homofobik Abusive Turkish Comments - HATC)) farklı ML sınıflandırıcıları ile analiz edilmiştir.

Ayrıca, COVID-19 pandemisinin Türkiye'de yayıldığı ilk günden salgının hızla arttığı ilk ay boyunca Instagram yorumları toplanarak kullanıcıların salgınla ilgili yorumları arasında faydalı bir etkileşim olup olmadığı, kişilerin duygu ve düşüncelerindeki değişiklikler ile olumlu-olumsuz etkilerinin duygu analizi çalışmaları yapılmıştır.

Tez çalışmasında, Instagram platformundan elde edilen veri kümeleri (ATC, HATC ve COVID-19) Türkçe kaynaklar için bir ilktir ve Türkçe dil kurallarına uygunluğu ve etiketleme işleminin yapılma süreçleri ayrıntılarıyla açıklanmıştır. Veri kümelerindeki etiket dağılımı farklı olduğu için çoğu sınıflandırma çalışması için (topluluk modelleri hariç) veri kümeleri yeniden örnekleme işlemine (aşırı örnekleme ve alt örnekleme) tabii tutulmuş ve sınıflandırma sonuçları bu açıdan da değerlendirilmiştir. Ayrıca, tez çalışması kapsamında oluşturulan bu veri kümeleri, başka bilim insanlarının farklı NLP ve TM çalışmaları için kullanıma açıktır.

Tablo 1.3'te gösterildiği gibi nefret söylemlerinin sınıflandırmasında TML yöntemlerinin yanı sıra topluluk modelleri, popüler olan DL modelleri ve BERT gibi anlam dönüştürücü yöntemleri kullanılmıştır.

Tablo 1.3. Tez çalışmasında kullanılan modeller.

Modeller	Yöntemler
TML	Lojistik Regresyon (Logistic Regression - LR)
	Naive Bayes (NB)
	Destek Vektör Makinesi (Support Vector Machine - SVM)
	Karar Ağacı (Decision Tree - DT)
	Rasgele Orman (Random Forest - RF)
Topluluk Modelleri	Uyarlanabilir Artırma (Adaptive Boosting - AdaBoost)
	Aşırı Gradyan Artırma (eXtreme Gradient Boosting - XGBoost)
	Gradyan Artırma (Gradient Boosting)
DL	CNN
	LSTM
	GRU
	BiLSTM
	GCR-NN
BERT	BERT metin sınıflama

Ayrıca küfür, homofobik ve diğer nefret söylemi türlerinin sınıflandırma doğruluğunu geliştirmek için seyrek vektör gösterimleri, yoğun vektör gösterimleri ve transfer öğrenme yöntemi ile önceden eğitilmiş güçlü nitelik çıkarım modelleri de tez çalışmasına dâhil edilmiştir (Tablo 1.4).

Tablo 1.4. Tez çalışmasında kullanılan nitelik çıkarım yöntemleri.

Kullanılan Nitelik Çıkarımları	Yöntemler
Seyrek Vektör Gösterimleri	Kelime Çantası (Bag of Word - BoW)
	ngramlar (unigram, bigram ve trigram)
	Terim Frekansı - Ters Doküman Frekansı (Term Frequency - Inverse Document Frequency - TF-IDF)
	One-hot (Tek sıcak) kodlama
Yoğun Vektör Gösterimleri	Kelimeden Vektörlere (Word to Vector - Word2Vec) (Sürekli Kelime Çantası (Continuous Bag of Word - CBoW) ve SkipGram)
	Bayt Çifti Kodlaması (Byte Pair Encoding - BPE)
	Global Vektörler (Global Vectors for Word Representation - GloVe)
BERT	BERT embedding (kelime yerleştirme)

Bu doktora tez çalışmasında, Türkçe dil kurallarına uygun olarak veri kümeleri toplanıp TM çalışmaları için uygun hale getirilmiş, küfürlü ve homofobik nefret söylemi türlerinin tespiti (ayrıca COVID-19 duygu analizi) için popüler DL, BERT gibi sınıflandırma modellerinin kullanılarak sınıflandırma başarısının geliştirilmesi amaçlanmıştır. Bu kapsamda;

1. Türkçe küfürlü yorumların tespitini mümkün kılan ATC adlı yeni bir veri seti ve HATC adında yeni bir Türkçe homofobik veri seti sunulmuştur. Bildiğimiz kadarıyla, sosyal medya platformlarında Türkçe küfürlü yorumlarla ve homofobi ilgili herhangi bir çalışma yapılmamıştır.

2. Veri setlerine etiketleme yapılırken homofobik ve küfürlü ifadelerin yanı sıra homofobi ve küfür ile ilgili emojiler de dikkate alınmıştır.
3. COVID-19'a ilişkin Türkçe duygu analizini farklı tarihler arasında tespit edilmesini mümkün kılan yeni veri kümeleri sunulmuştur (Dataset1 ve Dataset2).
4. SBTC çalışmalarının büyük çoğunluğu Twitter için yapılmıştır. Son zamanlarda Instagram için nefret söylemi türleri için yapılan çalışmaları artmış olsa da Twitter veri setlerini kullanan çalışmaların sayısı nispeten daha fazladır. Ancak Instagram'da paylaşılan fotoğraf ve videolar, Twitter'dan daha çok küfürlü yorumlar için uygun bir ortam oluşturmaktadır.
5. Elde edilen veri kümelerinin yeniden örneklenmiş dengeli versiyonlarının sınıflandırma sonuçları da sunulmuştur (Aşırı örneklenmiş ATC, resHATC, resDataset1 ve resDataset2).

Bu tezin geri kalanı şöyle organize edilmiştir: Bölüm 2'de nefret söylemi veri kümeleri ve Türkçe küfür, homofobik sınıflandırmaları ile ilgili literatür özetlenirken çalışmada kullanılan materyal ve yöntemler Bölüm 3'te açıklanmıştır. Çalışma kapsamında elde edilen iki veri kümesi ile ilgili (ATC ve HATC) nefret tespit sistemlerinin test verileri üzerinden elde edilen değerlendirme sonuçları ve COVID-19 veri kümesinin duygu analizi sonuçları Bölüm 4'te sunulmuştur. Bölüm 5'te ise tez çalışmasının sonuçları, tartışmalar ve öneriler verilmiştir.

2. KAYNAK ARAŞTIRMALARI

Nefret söylemleri ile ilgili birçok çalışma yapılmış ve çeşitli yaklaşımlar geliştirilmiştir. İngilizce dili en fazla nefret türü analizi çalışmasına sahipken, bu araştırma alanı Türkçe dâhil diğer diller için sınırlıdır. Tezin bu bölümünde nefret söylemleri için İngilizce veya diğer dilleri kullanan veri kümeleri, nefret söylemlerinin tespiti ve önlenmesi için kullanılan farklı nitelik seçimleri ve sınıflandırma metotları ile ilgili çalışmalar sunulmaktadır.

2.1. Nefret Söylemi Veri Kümeleri

Sosyal medya ağları gibi mikroblogların yaygınlaşmasıyla özellikle son on yılda nefret türlerinin tespiti üzerine çalışmalar artmıştır. Nefret söylemlerinin tespiti için en önemli ve ilk adım yüksek kapasiteli veri kümelerinin elde edilmesidir. Sosyal medyadan veri kümesi elde edilmesini hedefleyen çalışmalar azınlıktadır. Davidson ve ark., Twitter'dan ırkçı, cinsiyetçi ve homofobik içeren verileri toplamıştır (Davidson vd., 2017). Twitter'dan dinlerin, cinsel yönelimlerin, cinsiyetin ve etnik azınlıkların karalandığı 16 bin tivitten oluşan bir veri kümesi elde edilmiş ve manuel olarak etiketlenmiştir (Waseem, 2016; Waseem & Hovy, 2016). Elde edilen bu veri kümesine daha sonra cinsiyetçi ve ırkçı tivitler eklenmiştir. Sharma ve ark., art niyetli konuşma içeren 9 bin tivit toplamış ve nefret derecelerine göre bu tivitleri üç sınıf olacak şekilde manuel olarak etiketlemişlerdir (Sharma & Shrivastava, 2018). Irkçı Stormfront adlı web forum sayfasından nefret söylemi içeren 10 binden fazla cümle taranmış ve benzer bir çalışma olarak sunulmuştur (Gibert vd., 2018). Homofobik, cinsiyetçi ve ırkçı gibi çeşitli nefret türlerini içeren 28 binden fazla tivitten oluşan bir veri kümesi elde edilip veri kümesi üzerinde işlem adımlarının fazla olduğu bir sınıflandırma mimarisi tanımlanmıştır (ElSherief vd., 2018).

Nefret söylemi ile ilgili çalışmaların çoğu İngilizce dili üzerine odaklanmaktadır. Fakat farklı dilde yapılan veri kümesi çalışmaları da vardır. Örneğin; Twitter'da mültecilere yönelik Almanca nefret verileri toplanmıştır (Ross vd., 2017; Wiegand vd., 2018). Facebook sosyal ağından toplanan İtalyanca nefret verileri farklı nefret kategorilerine ayrılmıştır (Vigna vd., 2017). Saldırgan ve toksik ifadeler içeren birçok nefret veri kümesi Kaggle adlı web sayfasında kullanılabilmektedir ve Kaggle'ın Toksik Yorum Sınıflandırma Yarışması'nda (Kaggle's Toxic Comment Classification Challenge) kullanılan bu veri kümesi toksik etiketli 150 bin yorumdan oluşmaktadır. Tablo 2.1, literatürde mevcut olan ilgili veri kümelerinin bir özetini sunmaktadır.

Tablo 2.1. Nefret söylemi ile ilgili literatürde bazı veri kümeleri.

İlgili Çalışma	Platform	Dil	Etiketler
(Kwok & Wang, 2013)	Twitter (24.000)	İngilizce	Irkçılık
(Waseem, 2016)	Twitter (16.000)	İngilizce	Cinsiyetçilik/Irkçılık
(Wulczyn vd., 2016)	Vikipedia (100.000)	İngilizce	Saldırganlık
(Ross vd., 2017)	Twitter (541)	Almanca	Irkçılık
(Davidson vd., 2017)	Twitter (25.000)	İngilizce	Saldırgan/Diğer Nefret Söylemi Türleri
(Vigna vd., 2017)	Facebook (17.500)	İtalyanca	Nefret Söylemi
(ElSherief vd., 2018)	Twitter (28.000)	İngilizce	Taciz/Siber Zorbalık/Diğer Nefret Söylemi Türleri
(Wiegand vd., 2018)	Twitter (9.500)	Almanca	Aşağılama/Küfür
(Gibert vd., 2018)	Stormfront (10.000)	İngilizce	Nefret Söylemi
(Founta vd., 2018)	Twitter (80.000)	İngilizce	Saldırganlık/Küfür/Diğer Nefret Söylemi Türleri
(Sharma & Shrivastava, 2018)	Twitter (9.000)	İngilizce	Nefret Söylemi

2.2. Küfür, Homofobik ve Nefret Söylemi Türleri Tespit Yöntemleri

Olumsuzluklardan (art niyetli, toksik ve saldırgan etkileşimler) beslenen siber şiddet, geleneksel şiddetten farklıdır. Çünkü sosyal medyada sunulan küçük düşürücü metinler veya görsel materyaller herkes tarafından kolay erişilebilirdir (Heirman & Walrave, 2008). Siber şiddetin yaygın biçimi, birinin gönderisine küfürlü bir dil kullanarak yorum yapma eylemidir (Jones vd., 2013). Nefret söylemi, bir kişi veya gruba karşı nefret oluşturmak için sosyal medyada metinsel paylaşımlar olarak tanımlanabilir (Charitidis vd., 2020). Son yıllarda nefret söylemi analizi ile ilgili birçok çalışma yapılmıştır (küfür (Mossie & Wang, 2020) (Jones vd., 2013), (Waseem vd., 2018), taciz tespiti (Huang & Raisi, 2018), siber zorbalık (Balakrishnan vd., 2020), saldırganlık tespiti (Chatzakou vd., 2019), kadın düşmanlığının tespiti (Shushkevich & Cardiff, 2019) ve saldırgan dil analizi (Burnap & Williams, 2015)).

Küfürlü dil tespiti, nefret söylemi kapsamına dâhil edilmiştir ve SBTC için önemli bir araştırma alanı haline gelmiştir (Chatzakou vd., 2019; Chen vd., 2016; Huang & Raisi, 2018; Ibrohim & Budi, 2018; Johnson, 2018). Küfürlü konuşma, bireylerin ırkını, etnik kökenini, dinini, cinsiyetini, cinsel yönelimini, engellerini veya hastalığını hedef alır (Mossie & Wang, 2020). Küfürlü konuşma, doğrudan ağır bir aşağılama unsuru içermektedir.

Son araştırmalar, nefret söyleminin tespiti için sosyal ağ kaynaklarından (Facebook, Twitter, Instagram ve YouTube) elde edilen veri kümelerine SBTC metotlarının uygulanmasına odaklanmıştır (Charitidis vd., 2020). Küfürlü dil kullanımı nefret söylemi çatısı altında değerlendirilmektedir. Bu terminoloji aynı zamanda küfrü (uygun olmayan kelimelerin kullanımı) de kapsar. Bununla birlikte, birçok araştırmacı küfürlü dilden saldırgan dil olarak bahsetmektedir (Al-Hassan & Al-Dossari, 2019). Saldırgan/küfür içerikli mesajların tespitine yönelik yapılan çalışmalar kronolojik olarak şu şekilde

özetlenebilir: Kaynak (Park & Fung, 2017)'te yazarlar, taciz edici dilin tespit edilmesi için önce iki aşamalı bir yaklaşım önermişlerdir. Daha sonra bunu belirli türler halinde sınıflandırmışlar ve cinsiyetçi - ırkçı dil üzerinde birçok etiketli sınıflandırma ile tek aşamalı yaklaşımı karşılaştırmışlardır. Temel sinir ağı tekniklerini kullanarak küfür söylemi tespiti üzerine değerlendirmelerin gerçekleştirildiği bir çalışmada hazır sinir ağı tabanlı sınıflandırıcının (FastText) etkisi araştırılmıştır (Chen vd., 2017). Endonezya dilinde Instagram yorumlarından nefret söylemleri tespiti yapmak için yine FastText algoritması tercih edilmiştir (Pratiwi vd., 2019). Küfürlü kelimeler sözlüğü otomatik olarak oluşturulan bir başka çalışmada, Twitter, Wikipedia ve UseNet platformlarında küfürlü kelimeleri tespit etmek için hem kurumsal hem de sözlüksel kaynaklar kullanılmıştır (Wiegand vd., 2018). Endonezya sosyal medyasında küfürlü dili tespit etmek için ML tabanlı yöntemler kullanılarak bir çalışma yürütülmüştür (Ibrohim & Budi, 2018). Arapça çevrimiçi iletişimde anti-sosyal davranışların tespiti için müstehcen veya rahatsız edici kelimeler ve ifadeler içeren yorumların tahmine dayalı bir model sunulmuştur (Alakrot vd., 2018). Bu model için YouTube yorumlarından oluşan büyük bir veri kümesi toplamış, etiketlemiş ve kelime düzeyinde ngram özellikleri ve çeşitli ön işleme tekniklerinin kombinasyonlarıyla bir SVM sınıflandırıcıyı eğitmişlerdir. Bengalce dilinde küfürlü dil tespiti için ML tabanlı bir otomatik sistem oluşturulmuştur (Chakraborty vd., 2020). Verileri, tehdit edici ve küfürlü olarak sınıflandırmak için NB, SVM ve CNN metotları uygulanmıştır. Endonezya dilinde k En Yakın Komşulu (k-Nearest Neighbor - kNN) sınıflandırma yöntemiyle Instagram yorumlarının nefret söylemini tespit edip etmediğini belirleyen bir sistem oluşturulmuştur (Briliani vd., 2019). Facebook, Twitter, Instagram ve YouTube'dan veri kümeleri toplanmış ve 3 yorumcu tarafından manuel olarak iki dengeli sınıf olarak nefret ve nefret içermeyen biçiminde etiketlenmiştir (Omar vd., 2020). Bu veri kümesi, 12 ML tabanlı sınıflandırıcı ve 2 DL metodu ile sınıflandırma başarım sonuçları elde etmek için kullanılmıştır. Sonuç olarak RNN metodu, %98.7 doğrulukla diğer sınıflandırıcılardan daha iyi performans göstermiştir. Verilerin elde edildiği sosyal medya platformu dikkate alındığında Instagram'da siber zorbalık mesajlarını tespit etmeye yönelik çalışmalarda yapılmıştır (Bimantara vd., 2019; Hosseinmardi vd., 2015; Özel vd., 2017; Priyoko & Yaqin, 2019). Tablo 2.2, sosyal medya platformlarında küfürlü davranışların tespiti ile ilgili çalışmaları özetlemektedir.

Nefret söylemi, ofansif ve saldırganlık başlıkları altında yürütülen çalışmalarda genellikle homofobik dil analizi diğer nefret sınıflarıyla ile birlikte alt etiket olarak sınıflandırılmıştır. Bir nefret analizi çalışmasında, Twitter ve Whisper'dan (Whisper, 2021) elde edilen nefret ifadeleri altı nefret kategorisine (etnik köken, davranış, fiziksel özellikler, homofobi, sınıf ve cinsiyet) göre sınıflandırılmıştır (Silva vd., 2016). Kategorilerin her iki sosyal medya platformunda benzer olduğu sonucu analiz edilmiştir. Küfürlü dil tespiti için başka bir çalışmada, tivitler homofobik ve ırkçı olarak etiketlenmiştir (Davidson vd., 2017). Cinsiyetçi ifadeler ise saldırgan olarak etiketlenmiştir. Portekizcede saldırgan dil tespiti için yapılan bir çalışmada saldırgan veriler ırkçılık, cinsiyetçilik,

Tablo 2.2. Saldırgan/küfürlü mesajları tespit etmeye ilişkin çalışmalar ve en iyi sınıflandırma sonuçları.

İlgili çalışma	Platform	Dil	Nitelik gösterimi	En iyi sınıflandırıcı	Performans		
					Precsion	Hassasiyet	F1 score
(Park & Fung, 2017)	Twitter	İngilizce-2017	Karakter ve Word2vec	Hibrit CNN	0.71	0.75	0.73
(Chen vd., 2017)	YouTube, Myspace, SlashDot	İngilizce-2017	Kelime yerleştirme	FastText	-	0.76	-
(Wiegand vd., 2018)	Twitter, Wikipedia, UseNet	İngilizce-2018	Sözcüksel, dilbilim ve kelime yerleştirme	SVM	0.82	0.80	0.81
(Alakrot vd., 2018)	Youtube	Arapça-2018	ngram	SVM	0.88	0.80	0.82
(Ibrohim & Budi, 2018)	Twitter	Endonezyaca-2018	ngram	NB	-	-	0.86
(Briliani vd., 2019)	Instagram	Endonezyaca-2019	TF-IDF	kNN	0.98	0.98	0.98
(Pratiwi vd., 2019)	Instagram	Endonezya dili-2018	ngram	FastText	-	-	0.68
(Chakraborty vd., 2020)	Facebook	Bengalce-2019	Unicode, Bengalce kelimeler ve duygular	SVM	0.78		
(Omar vd., 2020)	Facebook, Twitter, Instagram, YouTube	Arapça-2020	Kelime yerleştirme	RNN	0.98	0.98	0.98
*(Karayigit vd., 2021)	Instagram	Türkçe	BoW, ngramlar, Word2Vec, BERT, BPE	CNN			0.973

homofobi, yabancı düşmanlığı, dini hoşgörüsüzlük ve küfür olarak sınıflandırılmıştır (Pelle & Moreira, 2017). İtalyancada yapılan bir nefret söylemi çalışmasında cinsiyetçilik, ırkçılık ve homofobik ifadeleri içeren bir veri seti homofobik veya homofobik olmayan olarak sınıflandırılmıştır (Akhtar vd., 2019). Nefret söyleminin etnik köken, din, cinsiyet veya cinsel yönelim olarak kategorize edildiği bir çalışmada, nitelik şablonları kullanılarak nefret söylemi tespit edilmiştir (Warner & Hirschberg, 2012). Ayrıca sözlük tabanlı ve duygusal yaklaşımlar kullanılarak çevrimiçi nefret söylemi adı altında ırkçılık, cinsiyetçilik ve homofobi kategorileri tespit edilmiştir. İngilizce tivitlerden oluşan bir veri setinde ırkçılık, cinsiyetçilik, homofobi kategorileri altında nefret söylemini tahmin etmek için sözlük tabanlı algoritmalar ve makine öğrenmesi yaklaşımlarının bir yöntem kombinasyonu sunulmuştur (Martins vd., 2018). Siyasi aşağılama, cinsiyetçilik, homofobi ve ayrımcılık kategorisinin saldırgan olarak tanımlandığı başka bir çalışmada diğer kategori saldırgan olmayan olarak etiketlenmiş ve Meksika

İspanyolcası tivitler için yazar ve saldırgan analizi yapılmıştır (Carmona vd., 2019). Homofobi ile ilgili daha önce yapılan çalışmalarda kullanılan verilerin kaynağı incelendiğinde, verilerin çoğunun Twitter'dan (Chatzakou vd., 2019; Sadiq vd., 2021; Waseem & Hovy, 2016) elde edildiği görülmektedir. Facebook (Aroyehun & Gelbukh, 2018), Instagram (Karayigit vd., 2021), YouTube (Salminen vd., 2018) ve diğer web platformlarından (Almerekhi vd., 2019; Gibert vd., 2018; Wulczyn vd., 2016) elde edilen veri kümeleri de mevcuttur.

Kullanılan yöntemler açısından önceki çalışmalar incelendiğinde, BoW, ngram, DL tabanlı yöntemler (CNN, RNN, LSTM, GRU ve BiLSTM), TML metotları (LR, NB, DT, RF ve SVM) homofobinin saptanmasında sıklıkla kullanılmıştır (Akhtar vd., 2019; Davidson vd., 2017; Kwok & Wang, 2013; Mehdad & Tetreault, 2016; Pelle & Moreira, 2017; Warner & Hirschberg, 2012; Waseem & Hovy, 2016). Yüksek sınıflandırma başarısı nedeniyle, homofobi tespiti için çoğunlukla DL tabanlı metotlar tercih edilmektedir (Badjatiya vd., 2017; Gambäck & Sikdar, 2017; Gao & Huang, 2017; Zhang & Luo, 2018). Ayrıca, transformatör mekanizmalarına dayalı önceden eğitilmiş modeller, nefret söyleminin analizinde önemli sınıflandırma başarıları elde etmiştir (Benballa vd., 2019; Gertner vd., 2019; Zhang vd., 2019). Nefret söylemi tespiti için genellikle TML ve DL sınıflandırma yöntemlerini kullanan çok dilli çalışmalar, birden çok dilde önerilen modellerin gürbüzlüğünü, diller arası bir ortamda gerçekleme yapmadan aynı anda değerlendirmiştir (Corazza vd., 2020; Ousidhoum vd., 2019; Stappen vd., 2020). Nefret söyleminde kullanılan bulanık mantık yöntemi 0 ile 1 arasındaki değerleri kategorize eden mantıktan oluşmaktadır. Çoğu dil probleminde, belirsizliği ortadan kaldırmak ve kesin sınıflandırma sonuçları elde etmek için bulanık mantık algoritmaları kullanılmaktadır. Fuzzy Rule-Based (Haque, 2014; Tashtoush & Al Aziz Orabi, 2019), Fuzzy Multi-Task Learning (Liu vd., 2019) ve birliktelik kuralı türlerini (Wadhwa & Bhatia, 2014) kullanan nefret söylemi çalışmaları vardır. Tablo 2.3, sosyal medya platformlarında homofobi ve cinsel yönelimle ilgili nefret söylemine ilişkin son çalışmaları kronolojik olarak özetlemektedir.

Tablo 2.3. Homofobik kategorileri kullanarak nefret söylemini tespit etmeye yönelik önceki çalışmalar.

İlgili Çalışma	Platform	Dil	Etiketler	Performans
(Warner & Hirschberg, 2012)	Yahoo! ve Amerikan Yahudi Kongresi (1.000 paragraf)	İngilizce - 2012	Irk, Etnik, Cinsiyet, Cinsel yönelim, Uyruk, Din ve Diğer özellikler	0.63 F1 score
(Silva vd., 2016)	Twitter ve Whisper (20.305 tivit ve 7.604 whispers)	İngilizce - 2016	Etnik, Davranış, Fiziksel Özellikler, Cinsel yönelim, Sınıf veya Cinsiyet	Tanımlanmamış
(Davidson vd., 2017)	Twitter (24.802 tivit)	İngilizce - 2017	Irkçılık, Cinsiyetçilik, Homofobi	0.90 F1 score
(Pelle & Moreira, 2017)	News web sitesi (115 haber 10.336 yorum)	Brezilya Portekizcesi - 2017	Irkçılık, Cinsiyetçilik, Homofobi, Yabancı Düşmanlığı, Dini Hoşgörüsüzlük, Küfür	0.70 F1 score
(Martins vd., 2018)	Twitter (975 tivit)	İngilizce - 2018	Irkçılık, Cinsiyetçilik, Homofobi	0.806 Kesinlik
(Akhtar vd., 2019)	Twitter (1.859 tivit)	İtalyanca - 2019	Homofobik ve Homofobik olmayan	0.80 F1 score
(Ousidhoum vd., 2019)	Twitter (13.014 tivit)	İngilizce, Fransızca ve Arapça - 2019	Köken, Cinsiyet, Cinsel Yönelim, Din, Engellilik, Diğer	0.86 Makro F1 (Makro ortalamalı F1)
(Carmona vd., 2019)	Twitter (10.85 tivit)	İspanyolca - 2019	Politika, Cinsiyetçilik, Homofobi, Ayrımcılık	0.65 Makro F1
*(Karayigit vd., 2022)	Instagram (31.290 yorum)	Türkçe 2022	Homofobi Diğer Nefret Söylemleri Nötr	0.90 Makro F1

3. MATERYAL VE YÖNTEM

Bu bölümde, elde edilen veri kümelerinin (ATC, HATC ve COVID-19) ayrıntıları, nefret söylemi türlerinin (küfürlü ve homofobik) tespiti ve COVID-19 duygu analizi için kullanılan TM mimarileri sunulmaktadır.

3.1. Materyaller

3.1.1. Türkçe Dilinin Yapısı

Türkçe dili, Ural - Altay dil grubunun Altay alt grubuna aittir (Kilinç vd., 2017). Kırk küsür dilden oluşan Türk dilleri, dünya da yaklaşık 165 - 200 milyon kişi tarafından ana dil olarak konuşulmaktadır. Latin alfabesini kullanan Türkçe, 21 ünsüz ve 8 sesli harften oluşur (Mujahid vd., 2021). Türkçe, Fince ve Macarca gibi içinde tek bir fiilin daha uzun kelimelere çevrilebildiği sondan eklemeli bir dildir (Çakıcı vd., 2018). Bu nedenle isimler ve fiiller yanlış cümle biçimlerine neden olabilmektedir. Türkçe kelimelerin sonuna eklenen eklerle kelime sayısı hızla artabilmekte ve bu durum kelime formlarının sayısını önemli ölçüde artırabilmektedir (Eryiğit vd., 2008). Türkçe sondan eklemeli bir dil olduğu için cümleleri daha basit parçalara bölmek zordur (El-Kahlout & Akin, 2013).

Türkçede bir kök sözcüğe "ipteki boncuk" gibi biçim birimler eklenerek farklı anlamlara sahip sözcükler elde edilir (Oflazer & Saraçlar, 2018). Türkçe kelimeler bir cümlede birçok çekim ve yapım eki alabilir. Türkçede çekim eki alarak değişen bir ifade İngilizce bir cümleye karşılık gelebilir (gör+ebil+ecek+se+k → if we will be able to see gibi). Türkçe "anahtar" kelimesinin beş veya daha fazla türevden kök alabileceğini ve beş türevden sonra farklı bir kelime haline gelebileceğini gösteren kelime "anahtar+lan+dır+ıl+ama+ma+sı+nda+ki" kelimesidir.

Tablo 3.1, büyük bir Türkçe külliyatta en sık kullanılan on sekiz sözcüğü, sözcükteki biçim birim sayısı ve her birinin biçim birimsel belirsizliği ile birlikte göstermektedir. Çoğu yüksek frekanslı kelimeler, bir morfemli (dilin anlamlı en küçük birimi) kelimeler için farklı konuşma köklerine sahip olmasına karşılık nispeten yüksek morfolojik belirsizliğe sahiptir. Bu çalışmada, yüksek morfolojik belirsizliğe neden olacak 201 kelimelik bir Türkçe durak listesi oluşturulmuş ve veri kümelerinden (ATC ve HATC) çıkarılmıştır.

Tablo 3.1. Sık kullanılan bir biçim birimli on sekiz Türkçe kelime ile ilgili istatistikler (Oflazer & Saraçlar, 2018).

Kelime	bir	bu	da	için	de	çok	ile	en	daha	kadar	ama	gibi	var	ne	sonra	ise	o	ilk
Belirsizlik	4	2	1	4	2	1	2	1	1	2	3	1	2	2	2	2	2	1

3.1.2. Küfürlü Türkçe Yorumlar (Abusive Turkish Comments - ATC) Veri Kümesi

Veri kümesi edinme temel bir adımdır ve metin sınıflandırma sürecinin en çok zaman alan kısmıdır (Omar & Al-Tashi, 2018). ATC veri kümesini elde etmek için güncel tartışmalı konuları ve görüşleri içeren Türkçe magazin sayfasının (ör. @2.sayfaofficial), futbol takımlarının (ör. @fenerbahce, @galatasaray) ve futbolcuların (ör. @ardaturan, @1volkandemirel) açık Instagram hesapları seçilmiştir. Bu tür açık sosyal hesaplarda küfürlü ve nefret dolu yorumlar bulmak daha olasıdır. Türkçe argo sözlüğüne göre (TDK, 2020), veri setindeki tüm yorumlar küfürlü veya küfürsüz olarak manuel etiketlenmiştir. ATC veri seti, 2017 ve 2019 yılları arasında yayınlanan 10.528 küfürlü ve 19.826 küfürsüz yorum içermektedir. Python programlama dili ile Instagram'dan metin verilerini toplamak için bir Instagram API edinilerek kodlama yapılmıştır. ATC veri seti, araştırmacıların ticari olmayan kullanımı için herkese açıktır (Karayigit vd., 2020).

Küfürlü ve küfürsüz yorumların sayısından da anlaşılacağı gibi, ATC veri seti etiketlerin dağılımı açısından dengesizdir. Dengesiz veri setlerinde belgelerin sınıflara dağılımı oldukça değişkendir; bir sınıf birkaç terim içerebilirken diğer sınıf çok sayıda terim içerebilir (Agnihotri vd., 2017). Bir veri kümesinde bir dengesizlik sınıfı mevcut olduğunda, DT ve SVM gibi sınıflandırma yöntemleri çoğunluk sınıfını dikkate alır. Sınıflandırıcılar genellikle azınlık sınıfındaki etiketleri yanlış etiketlenmiş olarak ele alır (Le vd., 2018). Bu nedenle, problemi çözmek için alt örnekleme ve aşırı örnekleme gibi dengeleme teknikleri kullanılmaktadır. Alt örnekleme tekniği, çoğunluk sınıfının sayısını azaltarak verileri dengelemektedir. Alt örneklemenin aksine, aşırı örnekleme ile azınlık verilerine yeterli sayıda rastgele örnek eklenmektedir (Gazzah & Amara, 2008). ATC veri seti büyük olmadığı için, alt örnekleme yerine aşırı örnekleme yöntemini kullanmak daha mantıklıdır (Tablo 3.2).

Tablo 3.2. ATC veri setinin iki versiyonu.

Veri kümesi	Yorumlar	
	Küfürlü	Küfürsüz
Orijinal ATC	10.258	19.826
Aşırı örneklenmiş ATC	19.826	19.826

3.1.3. Homofobik Küfürlü Türkçe Yorumlar (Homophobic Abusive Turkish Comments - HATC) Veri Kümesi

NLP çalışmalarında kullanılan veri kümeleri, sınıflandırma performansını iyileştirmek için çok önemlidir. HATC veri seti, homofobik söylem içerme potansiyeline sahip bazı hesaplardan elde edilen Instagram yorumlarından (@utandiran_paylasimlar, @kerimcandurmaz, @sametlicina vb.) ve ATC veri setindeki küfürlü Instagram yorumlarından oluşmaktadır (Karayigit vd., 2020). ATC veri setinde yer alan küfürlü yorumlar cinsel içerikli, homofobik, kişilik aşağılama ve dışkı ifadelerinden

oluşmaktadır. Tablo 3.3, ATC veri setinde yüksek frekansa sahip nefret dolu Türkçe kelimeleri göstermektedir.

Tablo 3.3. ATC veri setinde yüksek frekanslı nefret içeren bazı Türkçe kelimeler.

Kelime	Nefret türü	Frekansı
siktir	Cinsel içerikli	1182
amk	Cinsel içerikli	917
orospu	Cinsel içerikli	791
bok	Dışkı ifadesi	523
şerefsiz	Kişilik aşağılama	451
amına	Cinsel içerikli	405
amik	Cinsel içerikli	310
sikeyim	Cinsel içerikli	298
aq	Cinsel içerikli	288
piç	Kişilik aşağılama	260
sapık	Kişilik aşağılama	240
ananı	Cinsel içerikli	207
top	Homofobik	185
mk	Cinsel içerikli	159
yavşak	Kişilik aşağılama	156
mal	Kişilik aşağılama	146
kodumun	Cinsel içerikli	131
ibne	Homofobik	108

Homofobik yorumlar, ATC veri setindeki küfür etiketli yorumlardan çıkartılmış, Instagram'dan elde edilen homofobik yorumlarla birleştirilmiş ve manuel olarak homofobik (2 etiketinde) kategoriye çevrilmiştir. ATC veri setindeki cinsel içerikli, şiddetli aşağılama ve dışkılama ifadeleri ise nefret dolu yorumlar etiketinde "1" olarak etiketlenmiştir. Kalan yorumlar nötr yani "0" olarak işaretlenmiştir. Buna göre 1.226'sı homofobik, 10.237'si nefret dolu ve 19.827 nötr olan 31.290 Instagram yorumu toplanarak HATC veri seti oluşturulmuştur (Tablo 3.4). Instagram'dan homofobik verileri toplamak için ATC veri kümesinde olduğu gibi Instagram API ve Python programlama dili kullanılmıştır. Instagram, bir kullanıcı hesabıyla erişilebilen hesaplara ait verilerin uygun şekilde çıkarılmasına ve düzenlenmesine olanak tanıyan açık kaynaklı yapılandırılmamış veriler sağlamaktadır.

Tablo 3.4. HATC veri kümesi kategorileri.

Yorum Sayıları		
Homofobik	Diğer nefret söylem türleri (cinsel içerikli, ağır aşağılama ve dışkılama ifadeleri)	Nötr
1.226 (%3.9)	10.237 (%32.7)	19.827 (%63.4)

Homofobik veri setinin etiketlenmesi Türk Dil Kurumu (TDK, 2020) ve Büyük argo sözlüğüne (Aktunç, 2000) göre yapılmıştır. Türkçe'de en sık kullanılan homofobik ifade "top" kelimesi ve türevleridir (topitoş, topitop, totoş, toplar vb.). "Puşt, ibne", "lavuk" gibi kelimeler homofobiktir ve güvenilmez, aldatıcı kişiler için de kullanılmaktadır. Ayrıca homofobik ifadeler içermeyen ancak homofobiyi bazı emojilerle ifade eden yorumların analiz edilmesine özen gösterilmiştir. Emoji olarak ⚽,🏐, 👌 gibi şekillere sahip yorumlar homofobik olarak etiketlenmiş ve diğer homofobik anlam içermeyen emojiler HATC veri setinden çıkarılmıştır.

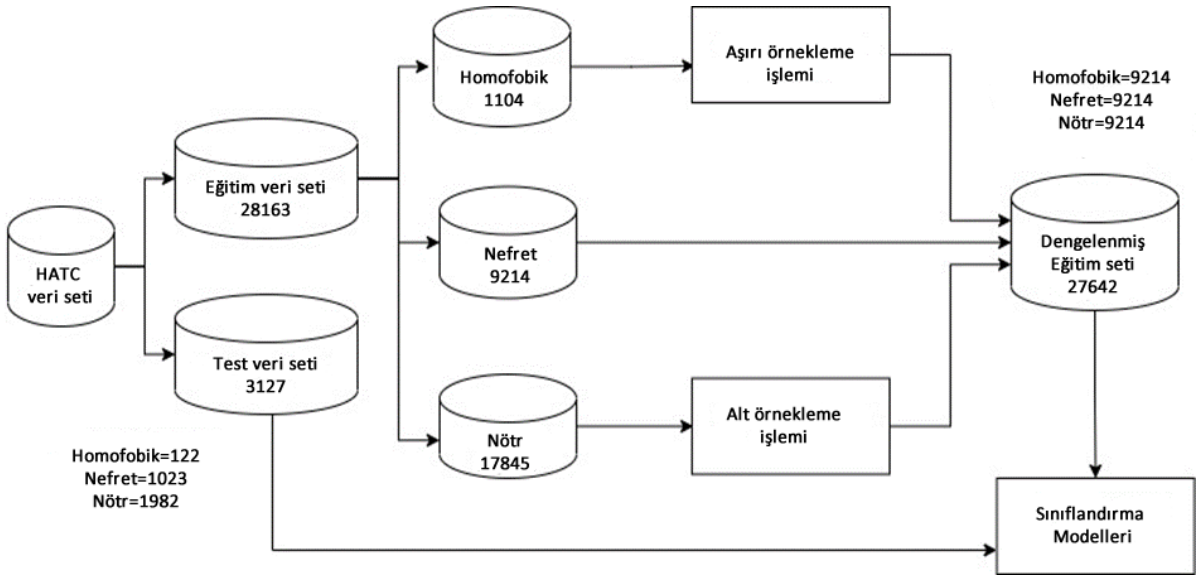
HATC veri kümesi, Tablo 3.4'te görüldüğü gibi dengesizdir. HATC veri setini dengelemek için rasgele aşırı örnekleme ve rasgele alt örnekleme işlemleri kullanılmıştır. Sınıflandırma performansı, DL tabanlı modeller kullanılarak yeniden örneklenen dengesiz veri kümelerinde hem iyileştirebilir hem de fazla uydurma ile azaltılabilir. Aşırı örneklenmiş veri kümelerinde, DL tabanlı yöntemler daha iyi performans göstermektedir, daha seçicidir, daha hızlı öğrenmekte ve daha az uyum sağlamaktadır (Simpson, 2015). Değerlendirme adımlarında HATC veri seti ilk olarak 10 katlı çapraz doğrulama kullanılarak bir eğitim - test setine bölünmüştür (Şekil 3.1). Her eğitim veri setinde, homofobik yorumların sayısı, nefret dolu yorumların sayısına eşit olana kadar rastgele artırılmıştır (aşırı örnekleme). Nötr veri sayıları ise nefret dolu yorumların sayısına eşit olana kadar rastgele azaltılmıştır (alt örnekleme). Böylece HATC veri setine yeniden örnekleme işlemi yapılarak dengelenmiş ve elde edilen yeni veri seti resHATC veri seti olarak adlandırılmıştır. Değerlendirmelerde HATC ve resHATC veri setlerinin sınıflandırma sonuçları ayrı ayrı karşılaştırılmıştır.

Şekil 3.1'de HATC veri seti 10 kat çapraz doğrulama için 10 bölüme ayrıldıktan sonra eğitim setindeki homofobik yorum sayısı 1.104, nefret dolu yorum sayısı 9.214 ve nötr yorum sayısı 17.845'tir. Yeniden örnekleme tekniği uygulandıktan sonra homofobik yorum sayısı 9.214, nefret dolu yorum sayısı 9.214 ve nötr yorum sayısı 9.214'tür. Test setinde homofobik yorum sayısı 122, nefret dolu yorum sayısı 1.023 ve nötr yorum sayısı 1.982'dir.

3.1.4. COVID-19 Veri Kümesi

Bu çalışmada COVID-19 ile ilgili ilk veri seti (Dataset1), Instagram'da Türkçe magazin hesabı "2.SayfaOfficial"in 11 Mart 2020 tarihinde COVID-19 paylaşımlarına yapılan yorumların toplanmasından elde edilmiştir. 12 Mart 2020 ile 10 Nisan 2020 tarihleri arasında aynı Instagram hesabında paylaşılan COVID-19 gönderilerine yapılan yorumlar Instagram API üzerinden alınmış ve Dataset2 veri kümesi elde edilmiştir.

Tablo 3.5, COVID-19 veri kümelerinin her bir kategorisi için yorum sıklığını göstermektedir. Toplanan toplam yorum sayısı 22.878'di. Etiketlenip temizlendikten sonra 9.708 yorum çalışma için uygun görülmüştür.



Şekil 3.1. HATC veri sayılarına göre dengelenme aşaması.

Bunlardan 2.745 pozitif yorum, 3.037 negatif yorum ve 3.926 nötr yorum vardır. Veri setlerinde toplanan yorumlar, duyguyu cümle düzeyinde ifade etmede manuel olarak pozitif, negatif ve nötr olarak etiketlenmiştir (Sağlam, 2019). Anahtar kelimeye dayalı ek açıklama kullanılmamıştır. COVID-19 pandemisi ile ilgili öfke, karamsarlık, alay, korku, nefret, küfür, şiddet gibi duygu örnekleri "1" yani "negatif" olarak etiketlenmiştir. Hastalıkla ilgili geçeceğine dair iyimser ve güven dolu duygular içeren yorumlar "2", yani "pozitif" olarak etiketlenmiştir. Son olarak, herhangi bir duygu veya beklenti belirtisi içermeyen yorumlar "nötr", yani "0" olarak etiketlenmiştir.

COVID-19 veri kümeleri için farklı tarihlerde veri setleri elde etmenin amacı insanların COVID-19 hakkındaki duygu ve düşüncelerinde zaman içinde meydana gelen değişiklikleri daha iyi gözlemlemektir. Veri kümelerinin Gizli Dirichlet Tahsisi (Latent Dirichlet Allocation - LDA) konu çıkarımları açık kaynak kodlu Python programlama dili kullanılarak elde edilmiştir. LDA hesaplanmadan önce kullanılacak veri kümeleri içerisindeki derlemin Türkçe dil bilgisi kurallarına göre tekrar imla kontrolleri yapılmıştır.

Tablo 3.5. COVID-19 veri kümelerinin karşılaştırmalı değerleri.

Veri Kümelerine Ait Değerler	**Yorum Sayısı**	
	Dataset1	**Dataset2**
Pozitif	671	2.074
Negatif	1.473	1.564
Nötr	743	3.183
Etiketlenen ve kullanılan	2.887	6.821
Toplam yorum	4.875	18.003

Değerlendirme çalışmaları öncesinde COVID-19 veri kümelerinde ki (Dataset1 ve Dataset2) sınıf dağılımı dengesiz olduğu için aşırı örneklenmiştir. Pozitif ve nötr örnekler, Dataset1'deki eğitim verilerinin sınıf dağılımını dengelemek için negatif örneklerin sayısına eşit olacak şekilde rastgele artırılmıştır. Negatif ve pozitif örnekler, Dataset2'deki eğitim verilerinin sınıf dağılımını dengelemek için nötr örneklerin sayısına eşit olacak şekilde rastgele artırılmıştır.

Dataset1 veri kümesinde "nötr" olarak etiketlenen yorumların çoğu ilk bildirilen COVID-19 vakasıyla ilgili kişinin bulunduğu yer ile ilgili sahte haberler ve söylentiler içermektedir. Nötr yorumlardaki bir diğer önemli nokta ise özellikle çocukları hastalıktan korumak için okulların kapatılmasının tekrar tekrar talep edilmesidir. Nötr yorumlarda el yıkamanın ve alkol bazlı dezenfektan kullanmanın önemine değinen yorumlar da sıklıkta yer almaktadır. İlk vakanın bildirilmesinden sonraki ay boyunca (12 Mart 2020 - 10 Nisan 2020) toplanan ve Dataset2'yi oluşturan yorumlar, Dataset1'den daha nötr ve pozitif yorumlar içermektedir. Bu durum insanların bu süre zarfında korku, endişe ve güvensizlikten kısa bir süre olsa da kurtulduğunu göstermektedir. Ayrıca pandemi sırasında sağlık görevlilerine olan güvenin arttığı da gözlemlenmiştir (Bhat vd., 2020).

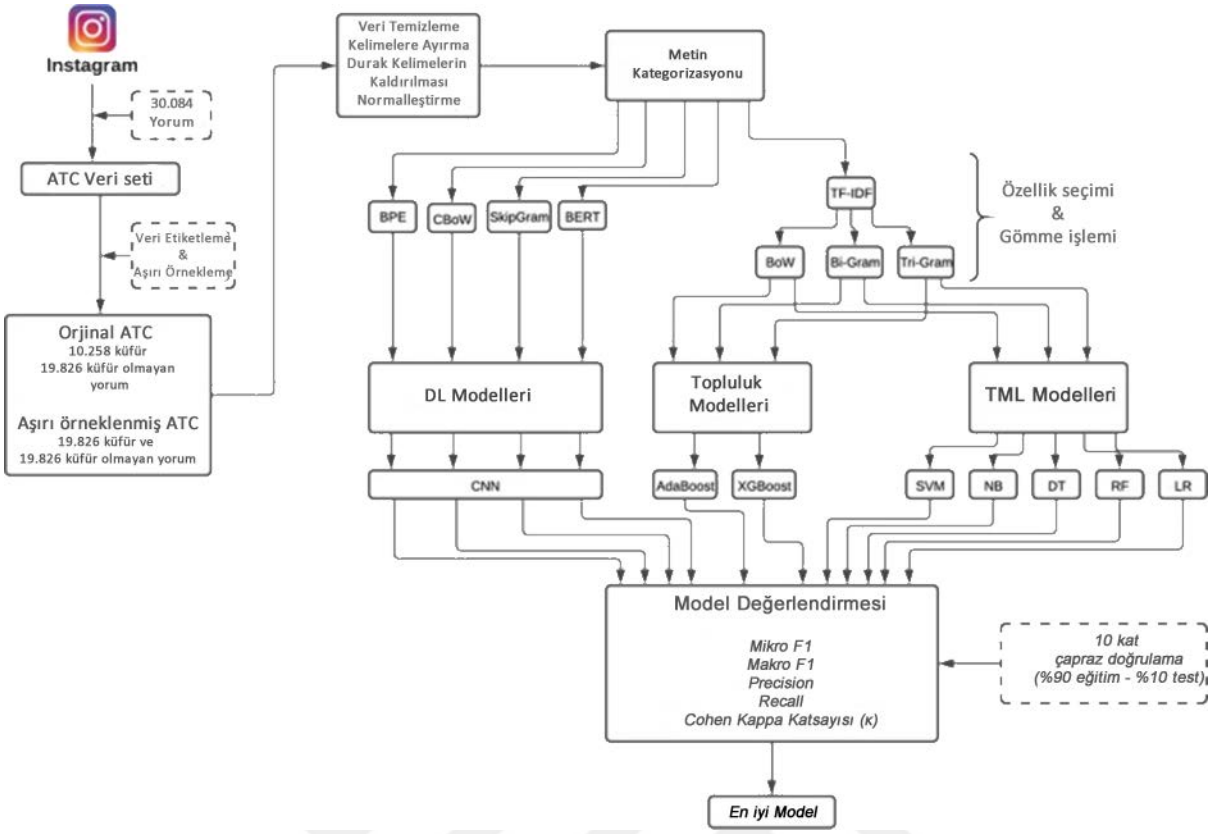
3.2. Yöntemler

3.2.1. Küfür Söylemi Tespit Mimarisi

Bu çalışmanın amacı, küfürlü metinleri tespit eden bir SBTC sistemi oluşturmaktır. Farklı nitelik çıkarma ve metin sınıflandırma yöntemleri arasından en iyi küfürlü yorum algılama modeli seçilmiştir. Bu bölümde kullanılan metotlar hakkında genel bilgiler verilmektedir. Yöntemlerin tümü iyi bilinen teknikler olduğu için kısaca açıklanmış ve teorik detaylar aşağıda diğer bölüm başlıkları altında verilmiştir. Genel model tasarımı Şekil 3.2'de sunulmaktadır.

Küfürlü yorum algılama sisteminin birkaç veri işleme adımı vardır: etiketleme, ön işleme, nitelik seçimi ve sınıflandırma. İlk olarak Instagram verileri elde edilmiş ve ATC veri seti oluşturulmuştur. Bundan sonra, veri kümesi manuel olarak iki sınıfa (küfürlü ve küfürsüz) etiketlenmiş ve bir tane daha veri kümesi oluşturmak için aşırı örnekleme uygulaması gerçekleştirmiştir. Yorumların kelime temsili olarak BPE, Word2Vec ve BERT kelime yerleştirme modelleri kullanılmıştır. ML tabanlı sınıflandırıcılar kullanılarak kategorizasyon için TF-IDF ile BoW, bigram ve trigram öznitelik seçim yöntemleri kullanılmıştır.

ATC veri seti, 10 kat çapraz doğrulama kullanılarak test edilmiştir. 10 katlı çapraz doğrulamada, orijinal numune rastgele on eşit boyutlu alt örneğe bölünür. On alt örnekten, modeli test etmek için doğrulama verisi olarak tek bir alt örnek tutulur ve kalan dokuz alt örnek eğitim verisi olarak kullanılmaktadır.



Şekil 3.2. Küfürlü içerik algılama sisteminin genel mimarisi.

Çapraz doğrulama işlemi daha sonra on kez tekrarlanmakta on alt örneğin her biri doğrulama verisi olarak bir kez kullanılmaktadır. Her çapraz bölümden elde edilen on sonucun daha sonra tek bir tahmin üretmek için ortalaması alınmakta veya başka bir şekilde birleştirilebilmektedir. Bu yöntemin avantajı, elde edilen tüm gözlem sonuçlarının her iki set (eğitim-test) içinde kullanılması ve elde edilen her bir gözlem sonucunun test edilmesi için bir kez kullanılmasıdır.

Tablo 3.6, sınıflandırma aşamasında kullanılan yöntemlere genel bir bakışı göstermektedir. En iyi sınıflandırma modelini bulmak için bu algoritmaların farklı kombinasyonları ile değerlendirmeler yapılmıştır.

Tablo 3.6. Türkçe küfürlü içerik algılama sistemi değerlendirmelerinde kullanılan metotlar.

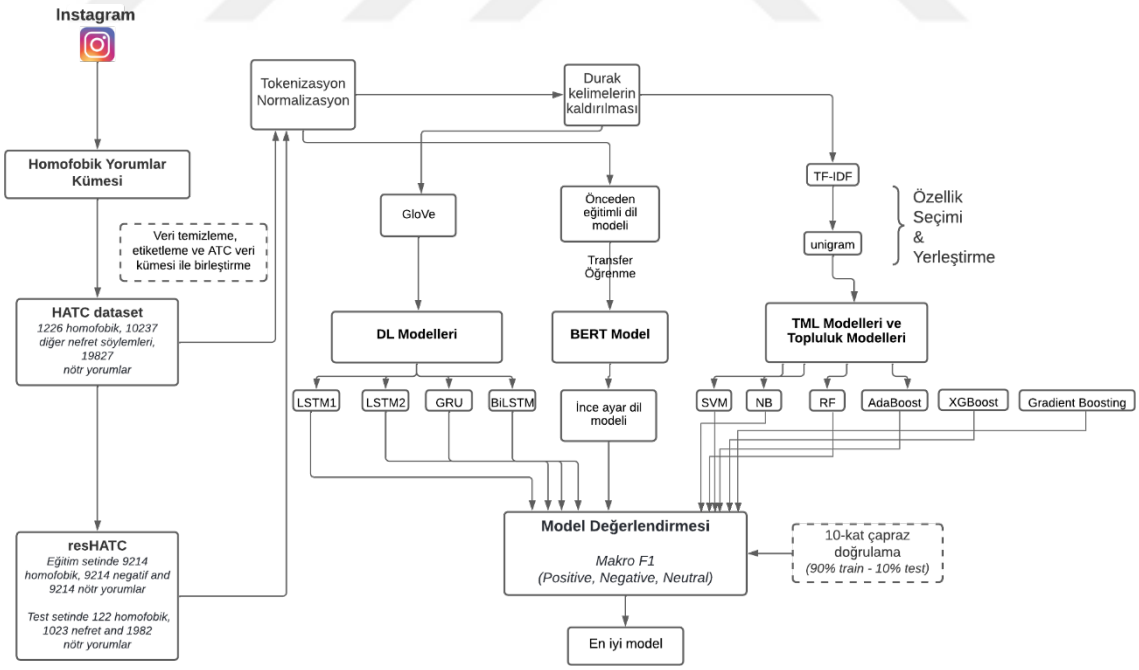
Veri Kümesi	Nitelik Çıkarma ve Yerleştirme	Sınıflandırıcılar	Eğitim Seti (%90)	Test Seti (%10)
Orijinal ATC	BoW + TF-IDF bigram + TF-IDF trigram + TF-IDF BPE, CBoW Skip-Gram BERT	CNN, NB, SVM, DT, RF, LR, Adaboost, XGBoost	27.076 yorum	3.008 yorum
Aşırı Örneklenmiş ATC			35.687 yorum	3.965 yorum

3.2.2. Homofobik Söylem Tespit Mimarisi

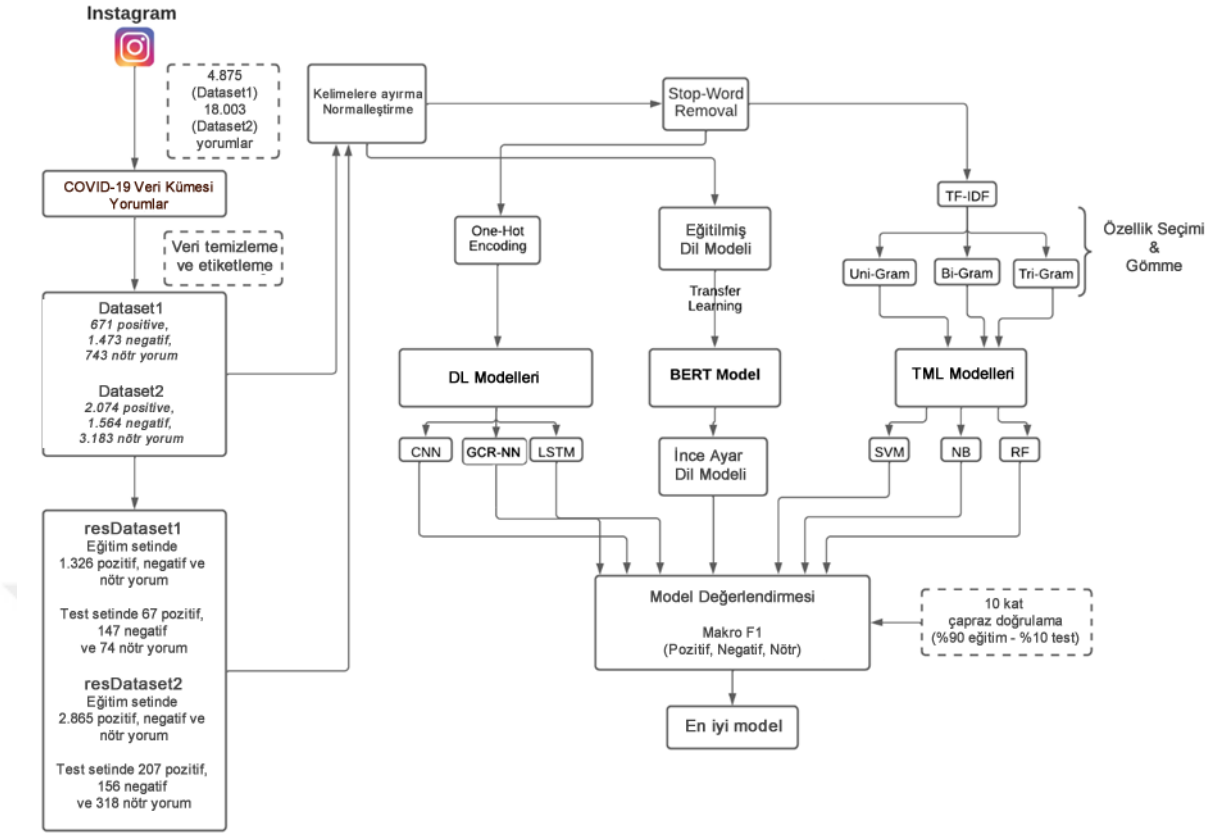
Bu bölümde homofobik ve ilgili nefret yorumlarının tespiti için kullanılan sınıflandırıcılar kısaca sunulmuştur. Vektörleştirilmiş nitelik çıkarımı için ngram, TF-IDF ve GloVe yöntemleri benimsenmiştir. TML yöntemleri olarak SVM, NB ve RF metotları kullanılmıştır. AdaBoost, XGBoost ve Gradient Boosting metotları topluluk modelleri olarak kullanılmıştır. DL tabanlı sınıflandırma için LSTM, BiLSTM ve GRU yöntemleri geliştirilmiştir. BERT adında önceden eğitilmiş verileri kullanabilen çok dilli bir model türü kullanılmıştır. Böylece, homofobik yorumları sınıflandırmak için toplam 22 yeniden örnekleme, nitelik alt kümesi seçimi ve sınıflandırma modeli kombinasyonu eğitilmiş ve doğrulanmıştır. Bu çalışmada kullanılan yöntemlerin şematik gösterimi Şekil 3.3'te verilmiştir.

3.2.3. COVID-19 Duygu Analizi Mimarisi

Şekil 3.4'te görüldüğü gibi COVID-19 duyarlılık analizi için üç model kullanılmıştır: TML modelleri (SVM, RF ve NB), DL modelleri (CNN, LSTM ve GCR-NN) ve BERT tabanlı transfer öğrenme sınıflandırma modeli. TML modellerinde, nitelik çıkarma için kelime ngramları (unigram, bigram ve trigram) ve özellik ağırlıklandırma için TF-IDF kullanılmaktadır. TML modellerinde kullanılan sınıflandırıcılar SVM, RF ve NB'dir.



Şekil 3.3. Homofobik ve nefret söylemi tespit mimarisi.



Şekil 3.4. COVID-19 duygu analizi için önerilen mimari.

Önerilen DL modellerinde nitelik çıkarımı için One-hot kodlama metodu kullanılmıştır. DL modellerinde kullanılan sınıflandırıcılar CNN, LSTM ve GCR-NN'dir. Önerilen üçüncü model olan BERT tabanlı transfer öğrenme sınıflandırma modeli, 104 dilde Wikipedia verileriyle önceden eğitilmiş çok dilli bir modelini kullanmıştır. COVID-19 duygu analizi için yapılan tüm değerlendirme çalışmalarında programlama dili olarak Python kullanılmıştır.

3.2.4. Ön işleme

Ön işleme adımı gereksiz verileri ortadan kaldırarak istatistiksel adımların gerçekleştirilmesine olanak tanıyan bir yapıya dönüştürür (Omar & Al-Tashi, 2018). Bu tezde yapılan çalışmalarda kullanılan ön işleme adımları aşağıda verilmiştir;

3.2.4.1. Veri Temizleme

Web bağlantı adresleri, hashtagler, sayısal karakterler ve noktalama işaretleri genellikle NLP uygulamaları için çok önemli değildir. Ancak hashtagler ve noktalama işaretleri gibi eklentileri kaldırmak, metin verilerini temizlemenin en iyi yolu olmayabilir. Örneğin; hashtagler, küfürlü, homofobik ve ya diğer nefret söylemi türlerinin tespiti için önemli olabilecek zengin anlamsal bilgiler

içerebilir. Hatta noktalama işaretleri, kullanıcıların seçeneklerini ifade etmek için alternatif noktasal emoji anlamı içerebilir. Bu nedenle, sonuçlara etkisini izlemek için küfürlü söylemler tespiti için ön işleme aşaması veri temizleme adımıyla birlikte ve olmadan gerçekleştirilmiştir.

3.2.4.2. Tokenizasyon

Tokenizasyon, metnin anlamlı parçalara ayrılmasına sağlar. ATC, HATC ve COVID-19 veri kümelerindeki yorumlar, boşluklar ve noktalama işaretleri temel alınarak kelimelerine ayrıştırılmıştır.

3.2.4.3. Durak Kelimeleri Kaldırma

Durak kelimeler veri kümesi içerisinde sıklıkla tekrar eden ve tek başına bir anlamı olmayan kelimelerdir. Durak kelimeleri (rakamlar, edatlar, zamirler ve bağlaçlar) ATC, HATC ve COVID-19 veri kümelerinden kaldırılmıştır.

3.2.4.4. Normalizasyon

Tüm veri kümelerinde metinler küçük harfe dönüştürülmüştür. Türkçe dili sondan eklemelidir, eklenen her son ek kelimenin anlamını değiştirir (Alparslan vd., 2013). Bu nedenle kelimelerin kök anlamına ulaşmak zorlaşır. Türkçede nefret söylemleri genellikle kök şeklindedir. Çekim eki olan nefret sözlerinde köklendirme yapıldığında anlam değişmektedir. Örneğin; "şerefsiz" kelimesi, "siz" eki aldığı için nefret uyandırmaktadır. Kelime olarak "şerefsiz" kelimesinin kökü "şeref" kelimesidir. Kelime olarak "şeref" kelimesinin anlamı "şerefsiz" kelimesinden farklıdır ve nefret içermez. Bu nedenle ATC, HATC veri setine kök bulma işlemi uygulanmamıştır. COVID-19 veri kümelerinde ise frekansı yüksek ilk 30 kelimeye kök çıkarma işlemi uygulanmıştır.

3.2.4.5. Sınıflandırma Performans Metrikleri

Önerilen sınıflandırma modellerinin performansını değerlendirmek için dört değerlendirme ölçütü kullanılmıştır: F1 (mikro ve makro ortalama), Cohen kappa katsayısı (κ), Kesinlik (Precision) ve Hassasiyet (Recall). Kullanılan performans metriklerinin tanımları aşağıda verilmiştir:

Karışıklık matrisi, dört bileşeni olan Doğru Pozitif (True Positive - TP), Yanlış Negatif (False Negative - FN), Yanlış Pozitif (False Positive - FP) ve Doğru Negatif (True Negative - TN) yardımıyla herhangi bir sınıflandırıcının performansını değerlendirmek için temel sağlar.

Örneğin; ATC veri kümesi için karışıklık matrisini yorumlarsak; TP, küfürlü olarak etiketlenen küfürlü yorumların sayısını temsil ederken, TN küfürlü olmayan olarak etiketlenen küfürsüz yorumların

sayısıdır. Ayrıca FP küfürlü olmayan olarak etiketlenen küfürlü yorumların sayısını temsil etmektedir ve FN küfürlü olarak etiketlenen küfürlü olmayan yorumların sayısıdır.

Belirli bir veri kümesindeki bir sınıflandırıcının sınıflandırma doğruluğu, sınıflandırıcı tarafından doğru olarak sınıflandırılan test kümesi gruplarının yüzdesini ifade etmektedir. Sınıflandırıcı çeşitli sınıfların kategorilerini ne kadar iyi tanıdığını yansıtır (Patro & Patra, 2015). Kesinlik değeri herhangi bir istatistiksel analizde sınıflandırıcı performansında önemli bir role sahiptir. Etiketli sınıfların sayısı, hem doğru hem de yanlış sınıflandırmalar dâhil olmak üzere bir sınıfın etiketli öğelerinin sayılarının toplamına bölünür. Hassasiyet, veri kümesine ait TP sonuçlarının TP+FN toplamına bölündükten sonra elde edilen değerdir. Düşük Hassasiyet değeri sınıfta bilinen kaç değerin eksik olduğunu göstermektedir. F1 score, Kesinlik ve Hassasiyet değerleri ile hesaplanmaktadır (Dwivedi vd., 2019). F1 score değeri, Kesinlik değeri ile Hassasiyet değerinin harmonik ortalaması olarak ifade edilir.

$$\text{Kesinlik} = \frac{TP}{TP + FP} \tag{3.1.}$$

$$\text{Hassasiyet} = \frac{TP}{TP + FN} \tag{3.2.}$$

$$\text{F1 skor} = 2.\frac{\text{Kesinlik} \,.\, \text{Hassasiyet}}{\text{Kesinlik} + \text{Hassasiyet}} \tag{3.3.}$$

Makro F1 her bir etiket için ölçüyü belirler ve ardından veri kümesindeki etiket sayısı üzerinden ortalamayı hesaplar.

$$\text{Makro F1} = \frac{1}{|\text{Classes}|}.\sum_{i \in \text{Classes}} F1(i) \tag{3.4.}$$

Mikro F1 (Mikro ortalamalı F1) önce her bir etiket için sırasıyla TP, FP, FN ve TN değerlerinin toplamını belirlemekte ve ardından bu sonuçlar üzerinden ölçüleri bulmaktadır.

$$\text{Mikro F1} = \frac{1}{D}.\sum_{i \in \text{Classes}} |i|.F1(i) \tag{3.5.}$$

Burada D, test setindeki örneklerin sayısıdır ve |i| i sınıfının önem derecesidir (Terragni vd., 2020). Makro F1 metriği, dengesiz veri setlerinde Mikro F1 metriğine göre adil değerlendirme için daha uygundur çünkü az örnekleme sahip kategorilerde bile sınıflandırıcının farklı yeteneklerini daha etkin bir şekilde göstermesini sağlamaktadır (Dogan & Uysal, 2019). Dengesiz veri kümelerinde sınıflandırma sonuçlarını Mikro F1 puanındansa Makro F1 puanı ile almak daha iyi olur. Makro F1

puanı, TP ve FN değerlerinin aritmetik ortalamasını sağladığından dengesiz veri setlerinde bile gerçek sınıflandırma doğruluk değerini verebilmektedir (Saraç & Özel, 2016).

Veri kümelerinin sağlamlığını ve kullanışlılığını test etmek için bir kıyaslama olarak sınıflandırıcı performanslarını kullanan κ değeri, tüm kategoriler (0 ve 1) için tahmin edilen ve gerçek örnekler arasındaki uyum derecesini göstermektedir. 1'e yaklaşan κ değeri iki gözlem sonucu arasındaki mükemmel uyumu gösterirken, κ değerinin − 1'e yaklaşması ise uyumsuzluğu göstermektedir (Kiliç, 2015).

3.2.5. Nitelik Seçimi ve Çıkarımı

Öznitelik seçimi ve çıkarımı adımı sınıflandırıcılar kullanılmadan önce, istenilen sonuçla ilgili olmayan ilgisiz özellikleri veri setlerinden çıkaran bir yöntemdir (Sebastiani, 2002) . Nitelik seçimi yöntemleri, veri kümesi içindeki veri örneklerinin sınıflandırıcılar tarafından işlenebilmesi için sayılara dönüştürülmesidir. Metinler üzerinde ML sınıflandırma yöntemlerini çalıştırmak için veri kümelerindeki örneklemlerin vektör temsillerine dönüştürülmesi gereklidir. Bu sürece nitelik çıkarma veya daha basitçe vektörleştirme denir. Dile duyarlı analize yönelik ilk önemli adım nitelik seçimi ve çıkarımı adımıdır.

NLP'de nitelik seçimi için kullanılan kelime temsil yöntemleri seyrek vektör gösterimleri (frekans temelli) ve yoğun vektör gösterimleri (tahmin temelli) olmak üzere ikiye ayrılmaktadır. En sık kullanılan seyrek vektör gösterimi olan BoW gibi algoritmalar metinler içerisinde ki kelimelerin sayılarına yoğunlaşmaktadır. Bu temsillerin en büyük dezavantajı satır ve sütunlardaki benzersiz kelimelerin fazlalığı büyük boyutlarda sıfırlardan oluşan matrisler oluşturmakta ve bu durum seyrek matris yapılarını ortaya çıkarmaktadır. Ayrıca seyrek vektör gösterimleri kelimeler arasındaki benzerliğe odaklanmaz. Diğer kelime temsil yöntemi olan yoğun vektör gösterimleri derlem içindeki sözcüklerin anlamsal ilişkilerini de hesaba katarak özelliklerini çıkarmaktadır.

3.2.5.1. Seyrek Vektör Gösterimleri

Her cümle için kelime sıklığı ile vektörlerin oluşturulması işlemi seyrek matris olarak adlandırılmaktadır. Vektör uzay modeli, metin verilerini sunan sayısal bir modeldir. Bu modelde metin ifadeleri ve sorgular vektör olarak ifade edilir. Vektör uzay modelinde kelime dağarcığı büyüdükçe vektörler son derece seyrek hale gelebilir. BoW, bigram, trigram gibi nitelik seçim yöntemleri temsil ettikleri seyreklik nedeniyle bu kategoride değerlendirilmektedir.

Metin verileri doğrudan bir temsil sağlamamaktadır. Bilgisayarlar sadece sayıları anladığı için metin verilerinin matematiksel bir biçimde temsil edilmesi gerekmektedir. Vektörleştirme işlemi ile derlemlerin matematiksel olarak temsil edilmesi sağlanır. Basit bir vektörleştime işlemi ile bir metin

belgesi içerisindeki kelimelerin sayısı âdetince bir dizi yapay değişken belge terim matrisi oluşturulmaktadır. Vektörleştirici, veri kümesinde tek tek kelimelerin belirli bir belgede terim frekansı olarak da bilinen kelime sıklığını gösteren ve sütunların her bir kelimeye tahsis edildiği bir belge terim matrisi oluşturacaktır.

N benzersiz kelimenin çıkarıldığı $(d_1, d_2, \ldots, d_N)$ D cümlelerini içeren bir veri kümesinde, sözlük N kelimeden oluşmakta ve elde edilen M vektör matrisinin boyutu ise D x N ile gösterilmektedir. M matrisindeki her satır Di veri kümesinde bulunan kelimelerin sıklığını tanımlamaktadır. Örnek olarak, ATC veri kümesinden iki cümle ile bir derlem oluşturulursa;

DERLEM:

Yorum 1 (D1) : Yazık olmuş yakıştıramadım.

Yorum 2 (D2) : Millet bu herifin avukatı olmuş.

Yorum 3 (D3) : Gevşek ya yemin ederim bu kadar mı karaktersiz olunur (ön işleme adımında altı çizili kelimeler silinir.)

İlk adımda benzersiz kelimeler belirlenerek kelimelerin (belirteçlerin) listesi oluşturulmuştur.

Benzersiz kelimeler: [‘yazık’, ‘olmuş’, ‘yakıştıramadım’, ‘millet’, ‘herifin’, ‘avukatı’, ‘olmuş’, ‘gevşek’, ‘yemin’, ‘ederim’, ‘kadar’, ‘karaktersiz’, ‘olunur’]

D = 3 satır

N = 13 benzersiz kelimedir.

Böylece, 3 x 13 boyutundaki M matrisi aşağıdaki gibi temsil edilir (Tablo 3.7).

Tablo 3.7. Seyrek vektör gösterimi.

Derlem	Derlem İçinde Geçen Kelimelerin Vektörel Gösterimi										
	yazık	olmuş	yakıştıramadım	millet	herifin	avukatı	gevşek	yemin	ederim	karaktersiz	olunur
D1	1	1	1	0	0	0	0	0	0	0	0
D2	0	1	0	1	1	1	0	0	0	0	0
D2	0	0	0	0	0	0	1	1	1	1	1

Her bir sütun, M matrisinde kendine karşılık gelen kelime için bir kelime vektörüdür. 3 x 13 boyutundaki M matrisi için oluşturulan kelime vektörleri; “yazık” vektörü için [1, 0, 0] ve “karaktersiz” kelimesi için vektör [0, 0, 1]’dir. Satırlar, veri kümesindeki yorumları, sütunlar ise veri kümesindeki kelimeleri göstermektedir. Bu varyasyonun ardındaki fikir, gerçek dünya uygulamalarında milyonlarca belgeye sahip bir veri kümesinin varlığı ve bu milyonlarca belgenin yardımıyla çok büyük bir sayı olan yüz milyonlarca benzersiz kelimeyi çıkartabilecek olmasıdır.

3.2.5.1.1. Kelime Çantası (Bag of Word - BoW)

BoW metodu, bir kelimenin derlem içerisinde kaç kez göründüğünü sayan bir algoritmadır. Bu kelime sayıları, arama, belge sınıflandırması, konuları modelleme gibi uygulamalar için belgelerin karşılaştırılmasına ve benzerliklerin ölçülmesine olanak tanımaktadır. Bu tür bir yaklaşımın arkasındaki sezgi, benzer belgelerin benzer kelimeler içermesidir. BoW vektörleştirme tekniği, metinleri sayısal nitelik vektörlerine dönüştürerek temsil etmektedir. BoW, veri kümesinden cümleyi alır ve her cümle içeriğindeki kelimeleri ML modellerinin algılayacağı nitelik vektörüne eşleyerek sayısal bir vektörlere dönüştürmektedir.

BoW metin vektörleştirme işlemi basit vektörleştirme işleminde anlatıldığı gibi kelimeleri ayırma (tokenizasyon) ve vektöre dönüştürme olmak üzere iki adımda yapılmaktadır. Veri kümesi içindeki cümlelere ya da ifadelere tokenizasyon işlemi uygulandıktan sonra veri kümesinin vektör boyutu benzersiz sözcüklerin sayısına eşittir. Her cümle için, bir vektörün tüm girişleri belgeye karşılık gelen kelime sıklığıyla doldurulmaktadır. ATC veri kümesinden 2 tane cümle örneği kullanarak bir derlem oluşturulursa;

DERLEM:

Yorum 1 (D1): Senlik suç yok takım oynayamıyor.

Yorum 2 (D2): Kulakları küpeleri benleri her şeyi benziyor:))

Cümleler kelimelere bölündükten (tokenizasyon) ve alfabetik sırayla benzersiz kelime listesi oluşturulduktan sonra derlem içindeki kelimelerin gösterimi aşağıdaki gibidir.

Benzersiz kelimeler: [“benziyor”, “kulakları”, “küpeleri”, “benleri”, “oynayamıyor”, “senlik”, “takım”, “suç”, “yok”]

Tablo 3.8 derlemin BoW yöntemi ile sunulan seyrek matris gösterimini vermektedir. Her cümle toplam nitelik sayısının uzunluğu ile aynı boyuta sahip bir matris dizisi olarak temsil edilir.

Tablo 3.8. BoW ile bir derlemin örnek gösterimi.

Derlem	Derlem İçinde Geçen Kelimelerin Vektörel Gösterimi								
	benleri	benziyor	kulakları	küpeleri	oynayamıyor	senlik	takım	suç	yok
D1	0	0	0	0	1	1	1	1	1
D2	1	1	1	1	0	0	0	0	0

Örnek derlem dizisinin tüm değerleri için eğer sözcükler konum ve nitelik vektörü içindeyse bir ile eğer vektör dışında kelimeler varsa bu durum sıfır ile temsil edilir. Bir derlemin BoW gösterimi tüm vektörlerin toplamı olur.

BoW modeli, sayıları kullanarak metin verilerini temsil etmek için bir mekanizma sağlar. Ancak, bunun belirli sınırlamaları vardır. BoW yalnızca bir derlemdeki terimlerin sayısına dayanır. Bu durum belirli görevler veya sınırlı kelime hazinesi olan durumlar için işe yarayabilir ancak büyük kelime dağarcığına sahip veri kümelerinde verimli bir şekilde ölçeklenemez. Bir belgedeki bir belirteç veya cümlelerle ilişkili semantikleri (anlambilim) dikkate almaz. Bir kelimenin bağlamsal ipucu verebilecek komşuluğundan nitelikleri yakalama olasılığını yok saymaktadır. BoW modeli geniş bir metin bütünü için kelime hazinesi açısından son derece büyük olabilmektedir. Bu büyüklük her belgeyi temsil eden çok büyük boyutlu vektörlere yol açabilmekte ve bu da model performansının düşmesine neden olabilmektedir.

3.2.5.1.2. ngramlar

ngram metodu cümle içindeki kelimelerin istenen kelime grubu âdetince alınarak sayısallaştırılmasıdır. Nitelik çıkarımı için kullanılan ngram algoritmaları, unigram, bigram, trigram gibi kullanımlarıyla aslında verilerin ardışık "n" sözcük dizisini belirtmektedir. Seyrek ngram algoritmasında "n" belirteci kelime tekrarının kontrol edildiği değeri temsil etmektedir, gram ise dizideki kelime tekrarının ağırlığını temsil etmektedir.

Sayı vektörleştirme tekniğine benzer olarak ngram yönteminde de bir derlemde terim matrisi oluşturulmakta ve bu matris kelime sayısını temsil etmektedir. Sütunlar, n uzunluğundaki bitişik sözcükleri temsil etmektedir. BoW'dan farklı olarak, ngramlarda kelime sırası korunmaktadır. Tablo 3.7'deki D derlemine göre bigram (ngram_range = (1,2)) için kelimelerin vektör temsili Tablo 3.9'daki gibidir.

Tablo 3.9. Nitelik seçimi bigram ile vektörel gösterimi.

Derlem	**Derlem İçinde Geçen Kelimelerin Vektörel Gösterimi**								
	[yazık, olmuş]	[olmuş, yakıştıramadım]	[millet, herifin]	[herifin, avukatı]	[avukatı, olmuş]	[gevşek, yemin]	[yemin, ederim]	[ederim, karaktersiz]	[karaktersiz, olunur]
D1	1	1	0	0	0	0	0	0	0
D2	0	0	1	1	1	0	0	0	0
D3	0	0	0	0	0	1	1	1	1

Otomatik küfürlü konuşma algılama ve ilgili görevlerde (nefret dolu, saldırganlık, toksik, ırkçılık ve cinsiyetçilik) sık kullanılan nitelik seçme teknikleri arasında olan ngram modelleri, kelimeleri analiz etmeden önce n-1 kelimenin varlığını tahmin etmektedir (Mossie & Wang, 2020). Bu tez çalışmasında, n uzunluğunun 1 ile 3 arasında değiştiği kelime tabanlı ngram modeller kullanılmıştır.

Küçük bir sayı olarak n değeri seçilirse veri kümesinde ki yararlı bilginin elde edilmesi için yeterli olmayabilir. Yüksek bir değer olarak n değeri seçildiğinde ise çok sayıda özelliği olan büyük bir matris elde edilebilmektedir. Çok fazla özellik nedeniyle, nitelik setinin seyrek hale gelmesi hesaplama

açısından zahmetli olmaktadır. Bu nedenle ngram algoritmasında n değerinin optimal değerini seçmek zordur.

3.2.5.1.3. Terim Sıklığı - Ters Doküman Sıklığı (Term Frequency - Inverse Document Frequency - TF-IDF)

Seyrek vektörleştirme tekniklerinin kullanımı basittir ancak tüm kelimeler eşit olarak ele alınmaktadır. Sık kullanılan kelimelerin veya önemli kelimeleri ayırt etmek için TF-IDF metodu geliştirilmiştir. Bu metot belirli kelimelerin sıklığına bakarak bir veri setindeki kelimelerin önemini tanımlayan istatistiksel bir değerlendirme ölçüsüdür. Kelimenin bir cümlede ve bir veri kümesinde ne sıklıkta geçtiğine bağlı olarak o kelimeye ait bir değer verir ve terimlerin ağırlığını hesaplamak için kullanılan popüler bir yöntemdir (Wang vd., 2010).

TF-IDF değeri, bir kelime hedef metinde sık kullanılıyorsa, diğer metinlerde daha az kullanılıyorsa frekansı yüksektir (Hernández-Castañeda vd., 2022). Bu değer, TF ve IDF olmak üzere iki terim olarak ayrı ayrı hesaplanmaktadır. TF, bir cümlede herhangi bir kelimenin kullanım sıklığını belirtmektedir ve belirli bir kelime için, cümlede kelimenin görünme sayısının cümledeki toplam kelime sayısına oranı olarak tanımlanır.

$$\text{TF(terim)} = \frac{\text{Cümlede terimin görülme sıklığı}}{\text{Cümlede toplam terim sayısı}} \tag{3.6.}$$

IDF ağırlıklandırma ise bir metindeki bir terimin önemini belirtmekte ve en yüksek ağırlığa sahip en iyi terimleri seçmektedir (Al-Radaideh & Al-Abrat, 2019). IDF sözcüğün derlem içinde ne kadar yaygın kullanıldığını belirleyerek nadir kullanılan sözcüklerin çıkarılmasına yardımcı olmaktadır. IDF, toplam belge sayısının içinde "hedef kelime" olan belge sayısına bölümünün logaritmasıdır. Bir kelime birçok cümlede birden çok kez görülüyorsa payda artmakta ve IDF'nin değeri düşmektedir. Örneğin; herhangi bir derlemde "ve", "bir" gibi kelimeler yaygındır ve hemen hemen her cümlede bulunmaktadırlar. 1000 cümlelik bir derlemin tamamında en az bir defa kullanıldığını varsayarsak IDF değeri log(1) = 0 çıkmaktadır. IDF bize sık kullanılan ortak kelimelerin daha az öneme sahip olduğu ana fikrini sunmaktadır.

$$\text{IDF(terim)} = \log\left(\frac{\text{Derlemde bulunan toplam cümle sayısı}}{\text{Derlemde terimin geçtiği toplam cümle sayısı}}\right) \tag{3.7.}$$

TF-IDF bulma denklemi Eşitlik (3.8)'teki gibi elde edilmektedir.

$$\text{TF-IDF(terim)} = \text{TF(terim)}.\text{IDF(terim)} \quad (3.8.)$$

Bu tez çalışmasında TF-IDF ağırlıklandırması ile nitelik seçim yöntemlerinin çeşitli kombinasyonları veri kümelerinde ki (ATC, HATC ve COVID-19) seyrek vektör gösterimlerinin belirledikleri frekanslara göre hesaplanmış ve veri kümelerine ait vektör uzayları oluşturulmuştur.

3.2.5.1.4. One-hot (Tek sıcak) Kodlama

İkili kodlama olarak da bilinen One-hot kodlama yöntemi, metin örneklerini seyrek gösterim olarak temsil etmekte ve örneklem kümesini sınıflandırmaya hazır duruma getirmektedir. Derlem kümesinde öznitelikler tamsayı değerlerine eşlenmektedir. Kelimelere denk gelen her tamsayı değerine 1 değeri verilmektedir. Tamsayı dışındaki tüm sayılara sıfır verilmekte ve ikili vektör olarak gösterilmektedir (Wei vd., 2020).

Veri kümesinde örnek kelime boyutuna N denilirse her kelime 0'dan farklı bir tam sayıya karşılık gelmektedir. One-hot kodlama vektör gösterimini i indeksli herhangi derlemde temsil etmek için sıfırlardan oluşan bir N uzunluk vektörü oluşturulur ve i konumundaki kelimeye 1 değeri verilir. Örneğin, D(i) kelime dağarcığının boyutu, kelime boyutu n = 5 benzersiz kelime için i = 2 olan kelime kimliğinin One-hot kodlama vektörü D(2) = [0, 1, 0, 0, 0] vektörüdür.

$$\left|\left|D(i)+D\left(i^{'}\right)\right|\right|^{2} = 0 \text{ eğer } i = i' \quad (3.9.)$$

$$\left|\left|D(i)+D\left(i^{'}\right)\right|\right|^{2} = 1 \text{ eğer } i \neq i^{'} \quad (3.10.)$$

One-hot kodlama algoritması seyrek vektör gösterimi olduğundan veri kümesindeki örneklerin anlamsal benzerliğini dikkate almaz ve kelimeler vektör uzayında eşit uzaklıktadır. Bundan dolayı da bu algoritmanın en büyük sorunu elde edilen temsil vektörlerinin yüksek boyutlu olmasıdır. Yüksek boyutluluk sınıflandırma sonuçlarında fazla uydurmaya yol açabilir ve çok fazla girdi örnekleri üzerinde bir sinir ağını eğitmek hesaplama açısından pahalılık oluşturabilmektedir.

3.2.5.2. Yoğun Vektör Gösterimi

3.2.5.2.1. Kelimeden Vektörlere (Word to Vector - Word2Vec)

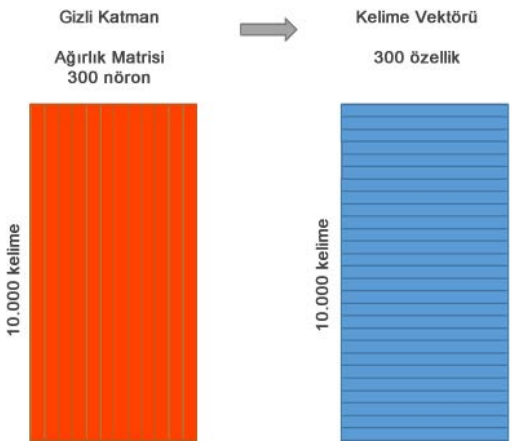
Word2Vec metodu NLP'de kelimeleri anlamlandırmak ve temsil etmek için kullanılmaktadır. Word2Vec, kelimeleri diğer kelimelerle birlikte nasıl kullanıldıklarına bağlı olduğu vektör uzayında temsil etmektedir. Eğitilmiş modeller, bağlamdan bağımsız veya bağlamsal olabilmektedir. Bağlamdan bağımsız modellerde, her kelimenin bir belgede çevresinde bulunan diğer kelimelerle değişmeyen tek bir benzersiz vektörü vardır (Babaeianjelodar vd., 2020). Kelime yerleştirme modellerinde, kelimeler

eşlenmekte veya düşük boyutlu gerçek değerli vektörlere gömülmektedir. Doğal olarak, sık kullanılan sözcükler dağılım niteliklerinin daha iyi bir temsilini sağlar; bu nedenle, kelime yerleştirmelerinin kalitesi, kelimelerin sıklığı ile doğrudan ilişkilidir.

Mikolov ve arkadaşları (Mikolov vd., 2013) tarafından geliştirilen Word2Vec, girdi katmanından cümleleri alan ve çıktı olarak kelimelerin vektör temsillerini üreten bir gizli katmana sahip iki katmanlı bir yapay sinir ağıdır. Her bağlamdaki sözcük sayısı sınırı, "pencere boyutu" adı verilen bir parametre ile belirlenmektedir. İki Word2Vec modeli vardır; Skip-Gram ve CBoW.

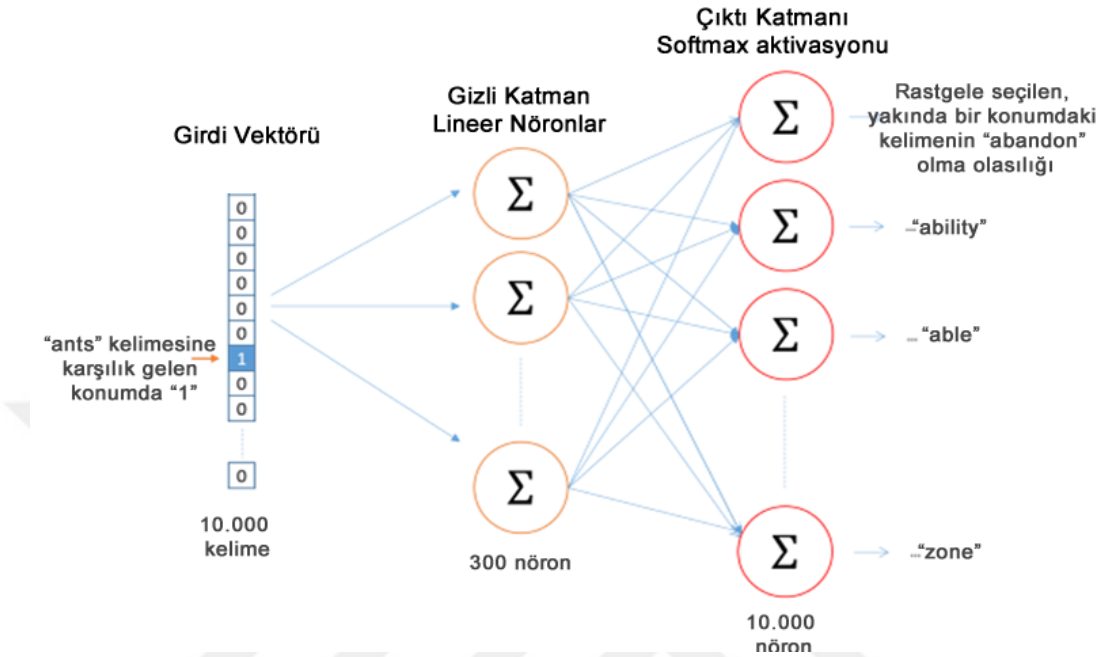
Skip-Gram modeli bağlamı tahmin etmek için girdi kelimesinin dağıtılmış temsilini kullanmaktadır. Kategorik çapraz entropisine dayanan bir kayıp işlevine sahip belirteç dizileri üzerinde denetimsiz olarak eğitilirler (Chamberlain vd., 2020). Skip-Gram modelinde merkezdeki hedef kelime girdi olarak alınır ve çevresindeki kelimeler çıktı olarak tahmin edilmeye çalışılır. Bu işlemler cümlenin sonuna kadar devam eder. Bir cümleye uygulanan bu işlemler tüm derleme uygulandığında başlangıçta etiketli olmayan veri kümesine haritalama (mapping) işlemi uygulanmış olmakta ve eğitime hazır hale gelmektedir. Örnek olarak; tek pencere seçimi için "Yazık olmuş yakıştıramadım" cümlesi için "Yazık" kelimesi girdi olarak verilirse "olmuş" ve "yakıştıramadım" kelimeleri çıktı olarak elde edilmektedir.

One-hot vektör olarak "yazık" girdi kelimesi temsil edilmektedir. Bu vektör örneğin 10.000 bileşene sahip olacak olsa (kelime dağarcığındaki her kelime için bir tane) ve "yazık" kelimesine karşılık gelen konuma 1 ve diğer tüm pozisyonlara 0 yerleştirilmiştir. Ağın çıktısı, derlemdeki her kelime için, rastgele yakınındaki bir kelimenin o kelime olma olasılığını içeren tek bir vektördür. Gizli katman nöronlarında aktivasyon fonksiyonu yoktur, ancak çıkış nöronları Softmax kullanmaktadır. Örneğin; bir derlemde 300 nitelikli bir kelime vektörü ve 10.000 satır var ise her gizli nöron için bir tane kullanılmak üzere 300 sütundan oluşan bir ağırlık matrisi ile temsil edilmektedir. 300 nitelik, kelime yerleştirme katmanı için kullanılan önceden eğitilmiş bir veridir. Nitelik sayısı, uygulamaya göre değişen ve ayarlanması gerekli olan hiperparametre değerleridir. Ağırlık matrisinin satırları kelime vektörleridir (Şekil 3.5).



Şekil 3.5. Ağırlık matrisleri ile kelime vektörlerinin ilişkisi (Chablani, 2017).

Skip-Gram modelinin gerçek hedefi gizli katmanda örnek özniteliklerin ağırlık matrisini öğrenmektir (Şekil 3.6). Herhangi bir Skip-Gram işleminin sonucunda "yakıştıramadım" girdi kelimesi temsili 1 x 300 vektörüdür ve bu vektör Softmax çıktı katmanına gönderilmektedir.

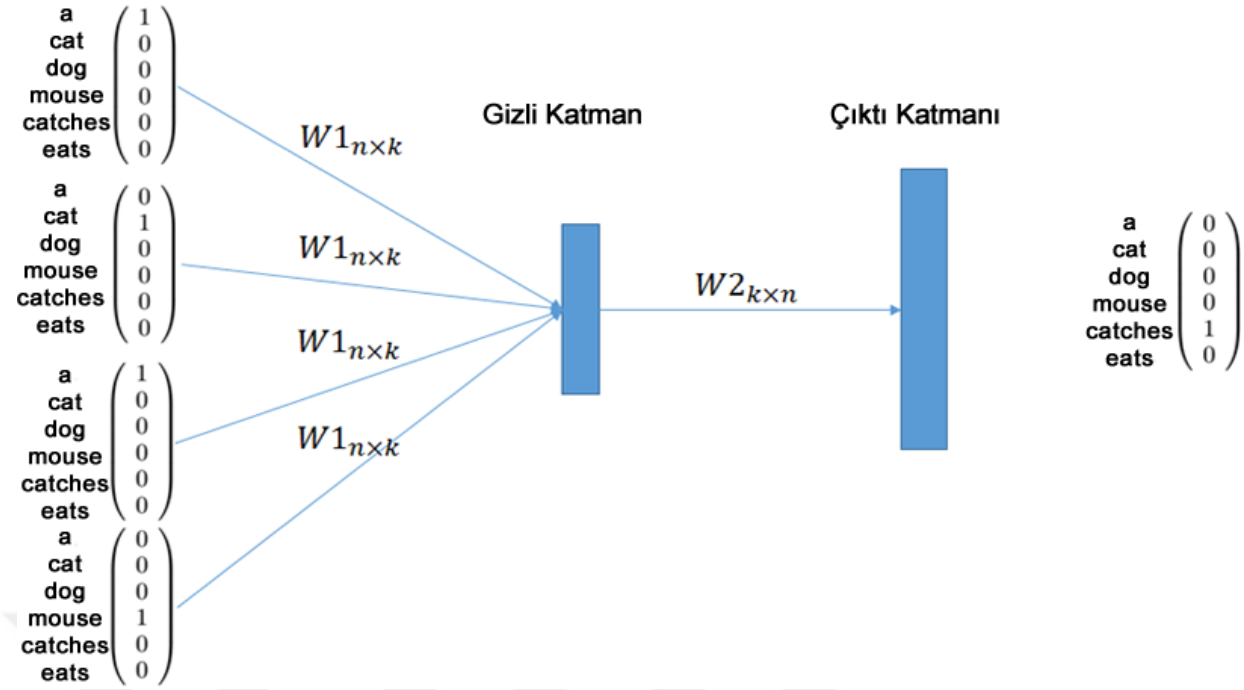


Şekil 3.6. Skip-Gram modeli (Chablani, 2017).

CBoW modelinde, ortadaki kelimeyi tahmin etmek için bağlamın veya çevresindeki kelimelerin dağıtılmış temsilleri birleştirilmiştir. CBoW modelinde merkezde olmayan kelimeler girdi olarak alınıp, bağlama bağlı olarak çıktı ya da hedef kelime tahmin edilmeye çalışılmıştır (Şekil 3.7). Örnek olarak "yazık olmuş yakıştıramadım" cümlesi için CBoW modeli kullanılırsa "yazık" ve "olmuş" kelimeleri girdi olarak verildiğinde hedef kelime yüksek olasılıkla "yakıştıramadım" olarak tahmin edilmiştir (Çelenli, 2020).

Skip-Gram ve CBoW algoritmaları gizli katmanlarında lineer aktivasyon işlemini kullandıkları için büyük veri kümelerinde bile hedef kelime ya da kelime gruplarını hızlı bir biçimde tespit edebilmektedir. Log-lineer modeller olarak da adlandırılan Skip-Gram ve CBoW algoritmalarında hedef kelimenin hesaplanması işleminde hata oranı negatif log olabilirliği ile elde edilir.

Skip-Gram algoritmasında derlem içindeki girdi kelimesi w_i verildiğinden çıktı olarak (w_{i-n}, …, w_{i-2}, w_{i-1}, w_{i+1},w_{i+2}, …, w_{i+n}) çıktı kelimeleri tahmin edilecek biçimde eğitim aşaması gerçekleştirilir. CBoW algoritmasında ise derlem içindeki girdi kelimeleri w_{i-n}, …, w_{i-2}, w_{i-1}, w_{i+1},w_{i+2}, …, w_{i+n}) verildiğinde çıktı olarak $\boldsymbol{w_i}$ hedef kelimesi tahmin edilir. CBoW modelinde giriş katmanına verilen girdi kelimelerinin ağırlık değerleri toplanır ve ortalaması çıktı katmanına gönderilir. Gönderilen çıktı değeri ile One-hot vektörü elde edilmiş hedef kelimenin değerleri karşılaştırıldığında bir hata değeri ortaya çıkar.



Şekil 3.7. CBoW modeli (Ferrone & Zanzotto, 2020).

Skip-Gram modelinde negatif log olabilirlik değerlerinin ortalaması minimize edilerek çıktı olarak istenen hedef kelimeleri bulma olasılığı artırılmış olur. Kelime vektörüne ait parametre θ, pencere boyutu n ile ifade edilirse N tane kelimeden oluşan bir derlemde her kelime w_i için aynı pencerede yer alan diğer kelimelerin olasılığı (maksimum olabilirlik fonksiyonu) L ve kayıp fonksiyonu J için ;

$$L(\theta)=\prod_{i=1}^{N}\prod_{\substack{-n\le j\le n\\ j\neq 0}} P(w_{i+j}\,|w_i\,;\theta) \tag{3.11.}$$

$$J(\theta)=-\frac{1}{N}\log L(\theta)=-\frac{1}{N}\cdot\sum_{i=1}^{N}\sum_{\substack{-n\le j\le n\\ j\neq 0}}\log P(w_{i+j}\,|w_i\,;\theta) \tag{3.12.}$$

Skip-Gram modelinde giriş katmanı ile gizli katman arasındaki kelimelerin vektör temsilleri v_{wi}, gizli katman ile çıkış katmanı arasındaki temsilleri v'_{wi} şeklinde gösterilmek üzere çıkış katmanında $P(w_{i+j}|w_i)$ olasılığı Softmax aktivasyon fonksiyonu ile hesaplanmaktadır.

$$P\left(w_{i+j}\middle|w_i\right) = P(w_{i-n}, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{i+n}\,|w_i) \tag{3.13.}$$

$$=\frac{\exp(v'^{\,T}_{w_{i+j}} v_{w_i})}{\sum_{k=1}^{v}\exp(v'^{\,T}_{w_k} v_{w_i})}$$

Softmax fonksiyonu ile derlemde bulunan her kelime için işlem yapılması büyük eğitim verisinden dolayı oldukça zahmetlidir. Bu durumun önlenmesi için negatif örnekleme ve alt örnekleme (sub-sampling) yöntemleri kullanılmıştır. Sık kullanılan sözcüklere sub-sampling ve negatif örnekleme yapma işlemleri yalnızca eğitim sürecinin hesaplama yükünü azaltmaz aynı zamanda elde edilen kelime vektörlerinin de kalitesini de iyileştirmektedir.

CBoW modelinde Skip-Gram modelinden farklı olarak girdi - çıktı katmanları farklılaşmaktadır ve J kayıp fonksiyonu;

$$J(\theta) = -\frac{1}{N}.\sum_{i=1}^{N} \log P(w_i \,|w_{i-n}, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{i+n}) \tag{3.14.}$$

3.2.5.2.2. Global Vektörler (Global Vectors for Word Representation - GloVe)

GloVe yoğun vektör gösterimi yalnızca kelimelerin yerel bağlam bilgisini dikkate almaz aynı zamanda kelime vektörlerini elde etmek için tüm derlem içindeki global istatistiklerden de (kelimelerin birlikte bulunma verileri) yararlanmaktadır. GloVe basit ağırlıklı en küçük kareler yöntemini kullanan bir log-bilinear regresyon modelini kullanmaktadır. GloVe, Word2vec gibi tahmin modeli yerine sayım tabanlı denetimsiz öğrenme modeli sunmaktadır. Bu modelde kelimelerin bir arada bulunması, modelin kelime temsilini öğrenmesi için mevcut olan en önemli istatistiksel bilgidir. Modelde daha düşük boyutlu bir matris elde etmek için yeniden kelime oluşturma ile ortaya çıkan kayıp olabildiğince azaltılarak daha iyi bir kelime yerleştirme modeli elde edilmeye çalışılmaktadır.

w_i ve w_j derlemdeki kelimeler olmak üzere koşullu olasılıkların oranı, kelimenin anlamlarını temsil etmektedir. Birliktelik matris değeri hesaplanarak derlem içindeki tüm kelimeler arasındaki birliktelik olasılığı elde edilmektedir. w_i kelime kontekstinde w_j'nin bulunma oranı x_{ij} olmak üzere derlemde bulunan sözcükler için w_i kontekstinde bulunma sayıları $\sum_i x_{ik} = x_i$'dir. w_i kelime kontekstinde w_j'nin bulunma olasılığı P_{ij}'nin denklemi;

$$P_{ij} = P(w_j|w_i)=x_{ij}/x_i \tag{3.15.}$$

Üçüncü bir kelime w_k bağlamı ile w_i ve w_j'nin temsili;

$$F\left((w_i-w_j)^T \widetilde{wk}\right) \approx P_{ik}/P_{jk} \tag{3.16.}$$

F fonksiyonu w_i ve w_j kelime vektörlerine uygulanan fark işleminin transpozunun w_k bağlam kelimesi vektörüyle iç çarpımını bulur ve w_k bağlam kelimesi için ayrı ayrı hesaplanan birliktelik

olasılıkları oranı üzerinden bir öğrenme gerçekleştirir. Semantik benzerlik bakımından w_k kelimesi w_i ile benzer w_j ile benzemezdir bundan dolayı P_{ik}/P_{jk} oranı büyür. Eşitlik (12)'deki olasılık değerleri bulunup bias terimleri eklendiğinde denklem aşağıdaki gibi sadeleştirilmiştir.

$$w_i^T.\widetilde{w}_k+b_i+\tilde{b}_k=\log(x_{ik}) \tag{3.17.}$$

GloVe modelinde w_i ve w_j kelime vektörlerine ait kayıp fonksiyonu (w_i, $w_j \in R^D$);

$$J=\sum_{i,j=1}^{v} f(x_{ij})(w_i^T.\widetilde{w}_j+b_i+\tilde{b}_j-\log x_{ij})^2 \tag{3.18.}$$

x_{ij} değeri ile f fonksiyonunda bir ağırlıklandırma değerinin elde edilmesi derlemdeki kelimelerin frekansına göre kelimelerin kapasitesinin artmasını engellemektedir. Frekansı yüksek olan kelimeler hata fonksiyonu değerini fazla etkilememekte ve yönlendirmemektedir.

$$f(x)=\min((x/x_{max})^a,1) \tag{3.19.}$$

Bu tez çalışmasında homofobik söylemlerin sınıflandırılmasında GloVe kelime yerleştirme yöntemi, DL tabanlı sınıflandırıcılarla birlikte kullanılmıştır. Kullanılan GloVe algoritması, Common Crawl (Crawl, 2022) üzerinde eğitilmiştir. Kelime hazinesinde 253 bin kelime vardır ve boyutu 300'dür. Eğitim verileri 2.736 kelimeli web taramalı çok dilli metindir. Veri kümesi boyutu 21 GB'tır.

3.2.5.3. Bayt Çifti Kodlama (Byte Pair Encoding - BPE)

Alt kelime seviyesindeki bilgiler, morfolojiyi yakalamak ve kelime dağarcığı dışındaki girdiler kompozisyon temsillerini geliştirmek için çok önemlidir (Li vd., 2018). BPE algoritması, metni bir sembol dizisi olarak gören ve en sık görülen sembol çiftini tekrarlayarak yeni bir sembolde birleştiren değişken uzunluklu bir kodlamadır (Heinzerling & Strube, 2017; Sennrich vd., 2015).

BPE, sözcük içindeki ortak sembolleri keşfetmek için rastgele uzunluktaki ardışık karakterler gibi eğitim veri kümesinin istatistiksel analizini gerçekleştirmektedir. Daha uzun yeni sembolleştirilmiş örneklemler üretmek için bir uzunluğunda sembolleştirilmiş örneklemlerden başlayarak sık görülen ardışık sembolleri yinelemeli olarak birleştirmektedir. Kelime sınırlarını aşan çiftler modelin verimliliği açısından dikkate alınmamaktadır. Derlem içindeki kelimelerin segmentlerine ayrılması için alt kelimeler semboller gibi kullanabilmektedir.

BPE, karakter ve sözcük düzeyinde hibrit temsiller arasında iyi bir denge sağlar ve bu da onu büyük veri kümelerini yönetme yeteneğine sahip kılar. Bu model aynı zamanda, derlem içinde olmayan

herhangi bir bilinmeyen kelimeyi eklemeden, sözcük dağarcığındaki nadir sözcükleri uygun alt sözcük belirteçleriyle kodlayabilir.

3.2.5.4. Dönüştürücü Temelli Çift Yönlü Kodlayıcı Temsilleri (Bidirectional Encoder Representations from Transformers - BERT)

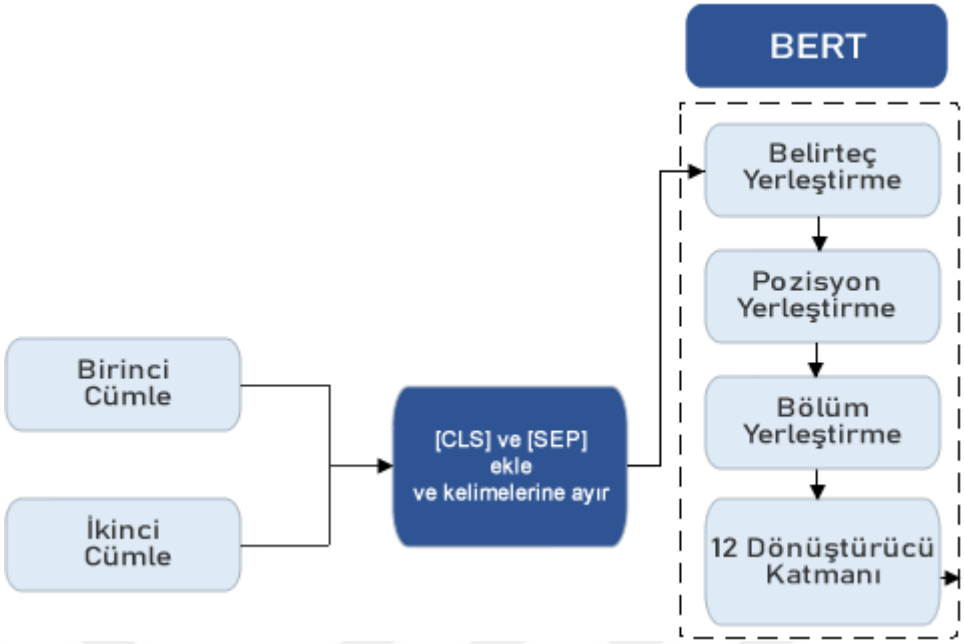
Yakın bir zaman da literatüre giren BERT yöntemi, bir dil modelini bir bütünlük üzerinde önceden eğitmek için çift yönlü dönüştürücü ağı kullanan ve önceden eğitilmiş modele diğer görevler için ince ayar yapan yeni bir dil temsil modelidir (Devlin vd., 2018). Göreve özel BERT tasarımı, tek bir cümleyi veya bir dizi cümleyi ardışık belirteçler dizisi olarak temsil edebilmektedir. Girdi temsili, belirli bir kelimeye karşılık gelen belirteç (token), bölüm (segment) ve pozisyon (position) kelime yerleştirmelerinin toplanmasıyla oluşturulur.

BERT modelde kelimenin tahmin edilmesi hem soldan sağa hem de sağdan sola olmaktadır. Cümle etiketleme görevi için soldan sağa yaklaşımda kelimeler teker teker tahmin edilir. Derlem içinde bir cümlede eksik kelimenin tahmin edilmesi görevinde ise tüm cümledeki bilgi dikkate alınarak rastgele bir kelime maskelenmekte ve tahmin edilmeye çalışılmaktadır.

BERT yöntemi derin çift yönlü bir model olarak tasarlanmıştır. Yapısındaki ağ ile ilk katmandan son katmana kadar belirtecin hem sağ hem de sol bağlamından bilgileri etkin bir şekilde yakalar. Bir sınıflandırma görevi için dizinin ilk sözcüğü benzersiz bir belirteçle tanımlanır ve son kodlayıcı katmanının konumuna tam bağlı bir katman bağlanır. Son olarak Softmax katmanı cümle veya cümle çifti sınıflandırma işlemini sonuçlandırır (Gao vd., 2019).

Modelin girişinde ham derlem verileri vardır ve her kelime bir belirteç için yerleştirilmektedir. BERT kelime yerleştirme modeli üç bölümden oluşmaktadır: 1. Belirteç yerleştirme, 2. Pozisyon yerleştirme, 3. Bölüm yerleştirme'tir ve üç yerleştirme bölümünün de boyutları da aynıdır (Ghorbanali vd., 2022). Modelin tüm girdilerinden elde edilen sonuç üç yerleştirme modelinin toplamında elde ettiği çıktının sonucudur (Şekil 3.8). Cümleler her zaman [CLS] ile başlamakta ve [SEP] ile bitmektedir. Bu çalışmada, dizi kodlayıcı birincil parametreleri kısmında bert-base-multilingual-cased ve bert-base-turkish-cased BERT önceden eğitilmiş veriler kullanılmıştır.

BERT kelime yerleştirme modeli önceden eğitimli veriler kullandığı için küçük çaplı veri kümelerine de uygulanabilmekte ve başarılı sonuçlar elde edebilmektedir. Model her kelime için o kelimeden farklı kelime vektörlerini de dikkate almakta ve farklı kelime yerleştirmeler döndürmektedir. Kelimelerin farklı kullanımları ve anlamları ayırma yeteneğinden dolayı bağlama duyarlı bir modeldir. Çok duyarlı yapısı bu modelin veri kümesinden daha çok bilgi edinmesine ve sınıflandırma performansını artırmasını sağlamaktadır.



Şekil 3.8. BERT kelime yerleştirme mimarisi.

3.2.6. Makine Öğrenmesi Modelleri

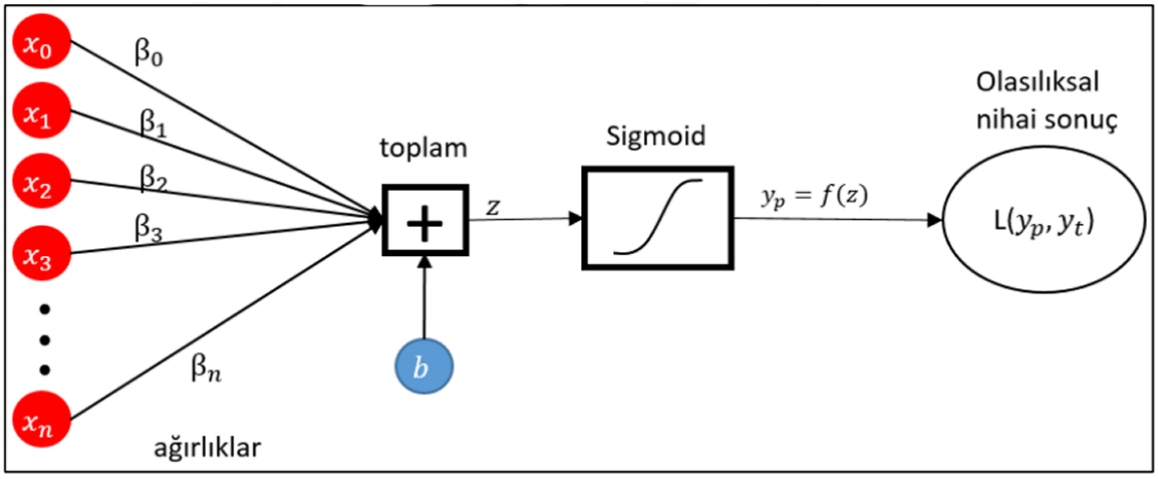
3.2.6.1. Lojistik Regresyon (Logistic Regression - LR)

LR metodu, nefret söylemi türleri sınıflandırma problemleri için yaygın olarak kullanılan istatiksel bir yöntemdir. LR yöntemi 1 veya 0 durumları için (evet - hayır, küfürlü - küfürsüz, pozitif - negatif vb.) sonuç vermektedir. Şekil 3.9'da 1 ve 0 durumları için LR modelinde aktivasyon fonksiyonu olarak Sigmoid fonksiyonunun kullanıldığı görülmektedir. Şekildeki x_0, x_1,x_2,x_3,…, x_n değerleri girdi değerlerini β_0, β_1,β_2, β_3, …, β_n değerleri ağırlık katsayısını, b değeri ise bias değerini, z değeri LR modelini oluşturmaktadır (Kaplan, 2020).

$$z = \sum_{i=0}^{n} \beta_i . x_i + b \tag{3.20.}$$

Çıktı değeri y_p değeridir ve bu değer z değerinin Sigmoid fonksiyonundan geçirilerek elde edilen değerdir.

$$y_p = \frac{1}{1+e^z} \tag{3.21.}$$



Şekil 3.9. LR model mimarisi (Kaplan, 2020).

Elde edilen çıktı değeri ile hedeflenen çıktı değeri arasındaki kayıp fonksiyonu ise $L(y_p, y_t)$ sonucu elde edilen değerdir.

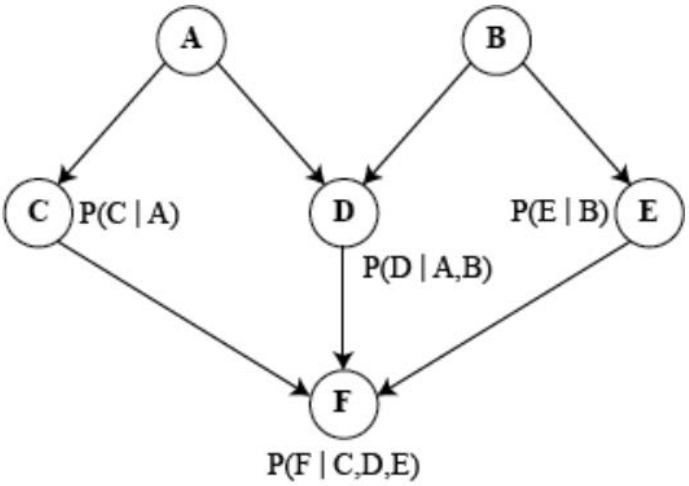
$$L(y_p,y_t) = -(y_t \times \log\left(y_p\right) + \left(1 - y_t\right) \times \log\left(1 - y_p\right)) \quad (3.22.)$$

Bu değer elde edildikten sonra β ağırlık katsayısı en düşük hata değeri biçiminde maksimumum olabilirlik yöntemi kullanılarak tahmin edilir. LR modelinin veri kümesindeki eğitim aşaması böylelikle tamamlanır. Test verileri ile LR modelinin eğitim aşamasında elde ettiği ağırlık ve bias değerleri kullanılarak sınıflandırma sonucu ortaya çıkmış olur.

LR'nin avantajı, eğitim setinde öğrenilen metin kategorilerini diğer sınıflandırıcılara göre daha kolay yorumlanmasıdır. LR'nin dezavantajı, daha fazla sınıflandırma sonucu için farklı sınıflandırıcılara göre daha fazla veriye ihtiyaç duyulmasıdır (Segura-Bedmar vd., 2018).

3.2.6.2. Naive Bayes (NB)

NB sınıflandırıcı, nefret söylemi gibi duygu analizi problemlerinde yaygın olarak kullanılan kısa sınıflandırma süresine sahip ve başarılı sonuçlar verebilen basit bir sınıflandırıcıdır. NB sınıflandırıcının mantığı koşullu olasılığa dayanmaktadır (Abooraig vd., 2018) ve yüksek yanlılığa, düşük varyansa sahiptir. NB sınıflandırıcısı, koşullu olarak bağımsız nitelikleri varsayan Bayes teoremini kullanmaktadır.



Şekil 3.10. NB model mimarisi (Kinalioğlu, 2019).

Şekil 3.10'da gösterildiği gibi NB sınıflandırıcı Bayes ağlarından ağaç nodülleri kullanarak rasgele alınan girdi değişkenlerinin koşullu olasılıkları ile sınıflandırma sonucu elde etmektedir. Olasılıksal akıl yürütme ile yapılan işlemler sayesinde diğer tüm girdi değerlerinin sonsal olasılığı tahmin edilerek sınıflandırma sonucu gösterilmektedir. Bayes denklemi;

$$P(A/B) = \frac{P(B/A).P(A)}{P(B)} \tag{3.23.}$$

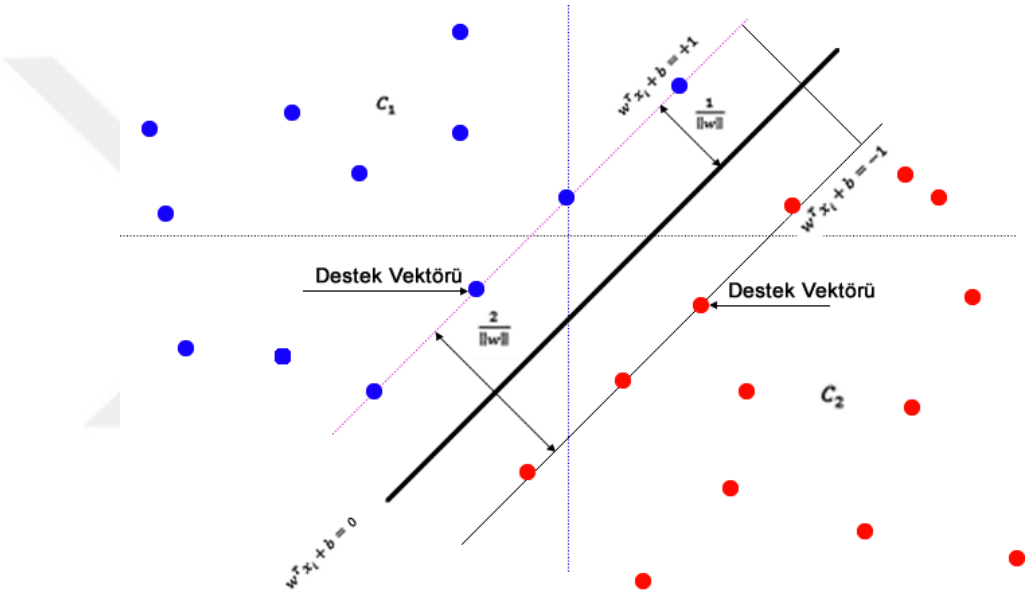
NB sınıflandırıcısına uyarlandığında; x_1, x_2, …, x_n girdi değerleri (öznitelikler), y değerinin ise sınıf etiketi olduğunu varsayarsak; i=1, …, n, ve j=1, …, k ise x_i girdi değerinin bilinmesi durumunda y_j sınıf etiketinin değeri (Kinalioğlu, 2019);

$$P(y_j/x_i) = \frac{P(x_i/y_j).P(y_j)}{P(x_i)} \tag{3.24.}$$

NB yönteminin avantajları kullanımındaki basitlik, gürbüzlük ve elde ettiği sonuçlar açısından yorumlanabilirliktir (Hmeidi vd., 2015). Tüm niteliklerin bağımsız olduğunun varsayılması BoW notasyonu gibi nitelik seçimlerini kullanmayı kolaylaştırır. NB sınıflandırıcının test etme ve tahmin etme becerisi son derece seridir (Xiang vd., 2014).

3.2.6.3. Destek Vektör Makinesi (Support Vector Machine - SVM)

SVM sınıflandırıcısı, doğrusal ve doğrusal olmayan verileri sınıflandırmak için kullanılır (Abooraig vd., 2018) ve nefret duygu durumlarını (homofobik, küfürlü, aşağılama vb.) belirlemek için mükemmel bir sınıflandırıcıdır. SVM yöntemi, sınıflandırma sonuçları üretmek için büyük miktarda veriye ihtiyacı yoktur. Bu metodun amacı, sınıfları ayırmak için optimal bir hiperdüzlem bulmaktır ve gürbüz teorik temelleri olan bir sınıflandırıcıdır (Saric vd., 2015). SVM, veri setini iki sınıfa ayıran n-boyutlu bir hiperdüzlem oluşturmaktadır (Han vd., 2012). Her iki hiperdüzlem sınıfından en yakın eğitim noktasına etkili bir ayrım yaparak genelleme hatasını azaltmaktadır (Hemmatian & Sohrabi, 2019). SVM'nin temel amacı marjin uzaklığını maksimize ederek marjin değerine bağlı olarak elde edilen destek vektörlerinin tam ortasına bir hiperdüzlem yerleştirmektir (Şekil 3.11).



Şekil 3.11. Doğrusal olarak ayrılabilen veri setleri için hiperdüzlemin belirlenmesi (Cortes & Vapnik, 1995).

Hiperdüzlemde ağırlık vektörü w ve bias değeri b, girdi değerleri x_i, destek vektör çizgisinin ifadesi (hiperdüzlemin fonksiyonu) ve iki sınıfı ayıran optimal sınıflandırma çizgi eşitliği (destek vektörlerinin fonksiyonları) ;

$$w^T.x_i + b = 0 \quad (3.25.)$$

$$w^T.x_i + b = 1 \quad (3.26.)$$

$$w^T.x_i + b = -1 \quad (3.27.)$$

$C=(C_1,C_2, \ldots, C_i)$ kümesi i adet sınıftan oluşan bir veri kümesi, örneklem kümesindeki girdiler $x_i \in C_i$ ise marj iki paralel hiperdüzlem ile tanımlanmaktadır.

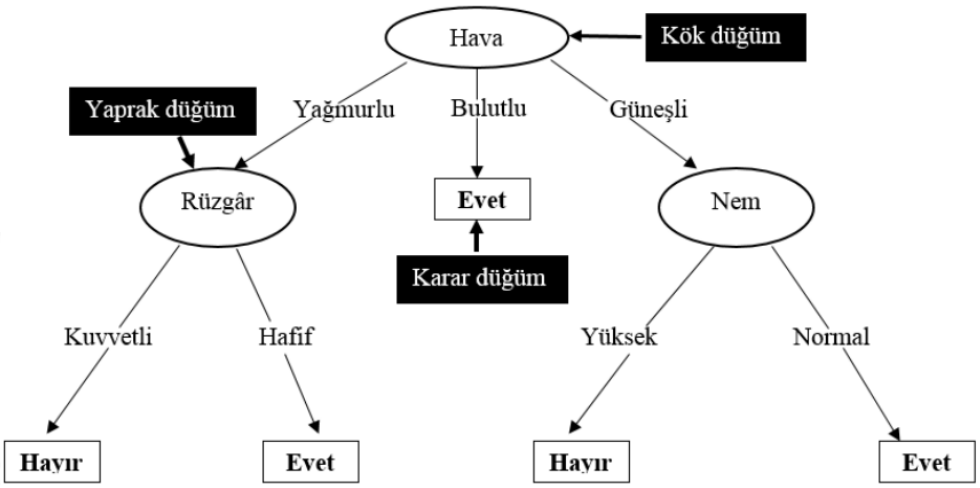
$$w^T . x_i + b \geq 0 \qquad y_i=1 \text{ için} \tag{3.28.}$$

$$w^T . x_i + b < 0 \qquad y_i = -1 \text{ için} \tag{3.29.}$$

SVM sınıflandırıcının avantajları, gürültüye karşı gürbüz olması fazla uyum riskinin düşük olması ve çekirdek (kernel) fonksiyonları nedeniyle karmaşık sınıflandırma problemlerini rahatça çözmesidir (Fatima & Pasha, 2017). Büyük metin içeren veri kümelerinde sıklıkla kullanılmaktadır.

3.2.6.4. Karar Ağacı (Decision Tree - DT)

DT metodu herhangi bir veri kümesinde çok sayıda nitelik içeren verileri, belli kurallar çerçevesinde daha küçük gruplara bölerek sınıflandırmak için kullanılır. Bu algoritmada en üstteki eleman kök, kökün altındaki elmanlar dal ve en son karar kısmında ki elemanlar ise yapraklardır. Bu yapısı nedeniyle metodun veri kümesinden öğrenmesi ağaca benzetilmektedir. DT sınıflandırıcılar, basit kullanımı ve gürültülü veriler üzerinde bile etkin sonuçları nedeniyle yaygın olarak kullanılmaktadır (Rokach & Maimon, 2014).



Şekil 3.12. DT metodunda düğümlerin belirlenmesi (Sarıbacak, 2021).

DT metodunda ağaç oluşturmak için kök değişken niteliği belirlenmektedir. Seçilen değişken niteliğine göre eğitim kümesi yinelenerek sınıflara en iyi şekilde ayıracak olan öncelikli değişken özelliği belirlenmektedir (Şekil 3.12). DT metodunda dallanma olayının hangi öncelikli niteliğe göre

yapılacağını belirlemek içim bilgi kazancı (information gain) ve toplam belirsizlik (entropi) hesaplanmaktadır. Sınıflandırma sonucunda veri kümesindeki verilerin tamamı aynı etikete ait çıkmış ise entropi = 0, etiketler kendi arasında eşit dağılmış ise entropi = 1, etiketler kendi arasında rastgele dağılmış ise 0 < entropi < 1 olmaktadır.

Z, sınıf etiketleri bulunan eğitim kümesini D_i, sınıf sayısını (i = 1, …, n), entropisi hesaplanacak durum sayısını n, P_i i durumun olasılığını göstermek üzere entropi değeri:

$$\text{Entropi}(Z) = -\sum_{i=1}^{n} P_i \log_2 P_i \tag{3.30.}$$

Z sınıf etiketleri bulunan eğitim kümesindeki nitelikler $\boldsymbol{A}\ \{a_1,a_2,a_3, \ldots, a_k\}$ gibi k'ya kadar farklı değer aldığı düşünüldüğünde eğitim kümesi $Z\ \{z_1,z_2,z_3, \ldots, z_k\}$ gibi k değerde parçaya ayrılabilmektedir. Sınıflandırma için başlanacak kök ile ağacın oluşturulması ve seçilen değişkenlere bağlı olarak örnek kümenin yinelenerek bölünmesi için örnekleri sınıflara en iyi şekilde bölecek olan nitelikler belirlenmektedir. Her nitelik parçası için ayrılan parçanın bilgisinin hesaplanması Eşitlik (3.31)'de gösterilmiştir.

$$\text{Entropi}_A(Z) = -\sum_{j=1}^{k} \frac{|Z_j|}{|Z|} . \text{Entropi}(Z_j) \tag{3.31.}$$

Bilgi kazancı, veri kümesinde elde edilen entropi değerinin veri kümesinde her nitelik için elde edilen entropi değerlerinden çıkarılmasıyla elde edilen değerdir.

$$\text{Bilgi}(A) = \text{Entropi}(Z) - \text{Entropi}_A(Z) \tag{3.32.}$$

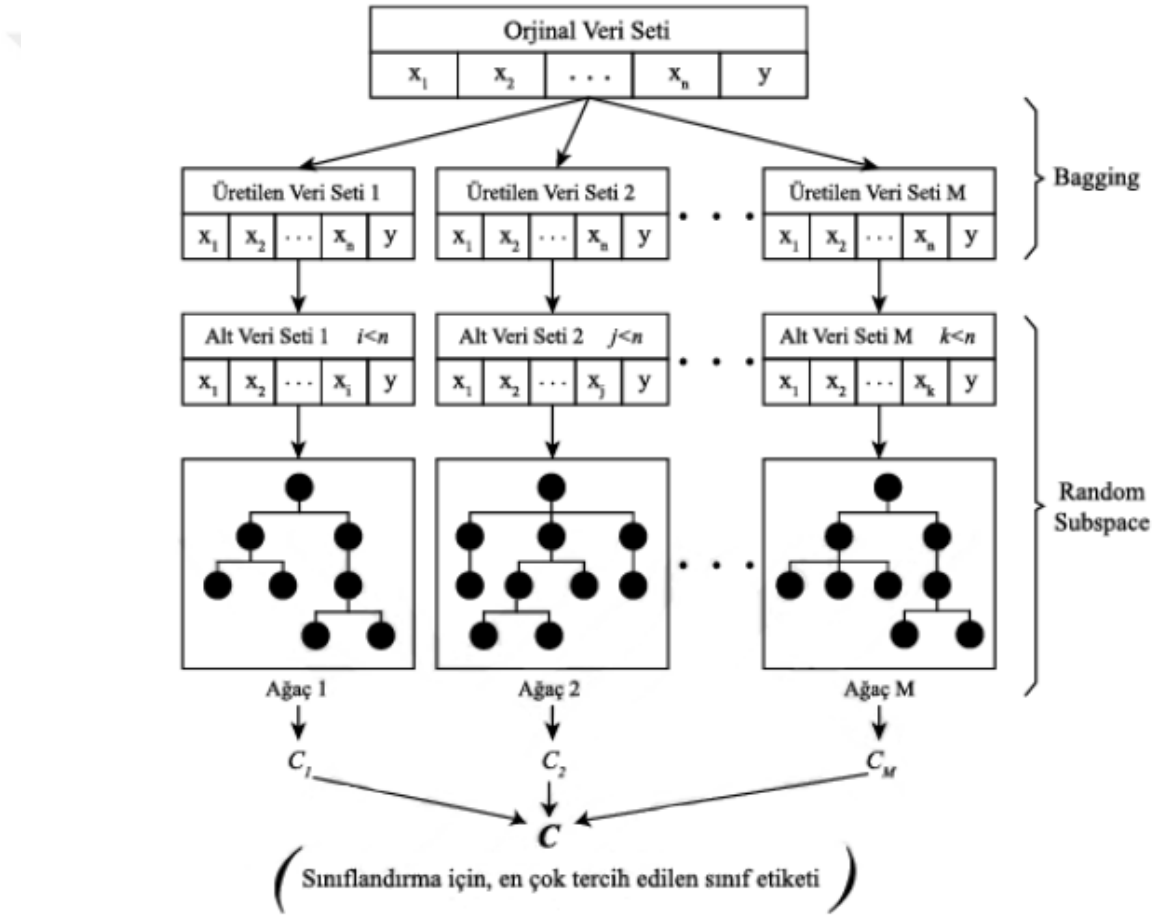
DT'nin avantajları, yorumlanabilirlik ve veri niteliklerini başarılı bir biçimde işleyebilme yeteneğidir (Hmeidi vd., 2015). Sezgisel bir algoritma olması münasebetiyle veri kümesinde eksik değerler olsa bile karar ağaçları her türlü veriye kolaylıkla uygulanabilir. Model eğitiminin farklı ML metotlarına göre daha uzun sürmesi, daha iyi tahmin için rasgele orman gibi karmaşık daha büyük ağaç yapılarına ihtiyaç duyması dezavantajları olarak nitelendirilebilmektedir (Sarıbacak, 2021).

3.2.6.5. Rasgele Orman (Random Forest - RF)

RF sınıflandırıcısı aslında torbalama (bagging) temeline bağlı bir topluluk öğrenme yaklaşımıdır. Bu metot NLP çalışmalarında sıklıkla kullanılan gelişmiş bir DT yöntemidir. DT

algoritması, yüksek varyans nedeniyle kararsızlık sorununu çözmek için RF sınıflandırıcısı ile kullanılmaktadır. RF, birçok farklı karar ağacı oluşturur ve karar ağaçları tarafından elde edilen sınıflandırma puanlarının ortalamasını alır. Böylelikle fazla uydurma ve yanlılık sorunları azaltılmış olur (Breiman, 2001).

RF metodunda karar ağaçlarından oluşan bir yapı ile veri kümesi verilerinden yeniden örnekleme (bootstrap) yoluyla elde edilen vektörleri sınıflandırarak bir karar çıktısı oluşturmaktadır. Veri kümesinin var olan orijinal halinden yeniden örneklem yoluyla yeni veriler üretip karar ağaçlarının kullanılmasına bagging yöntemi adı verilmiştir. Yine bu algoritmada kullanılan rastgele alt uzay yöntemi (random subspace) tekniği ile de rasgele belirlenen özniteliklerle karar ağaçları oluşturulmakta ve karar ağaçlarının uygun dal seçimine sahip olması için gerekli özniteliklere karar verilmektedir.



Şekil 3.13. RF metodunun mimarisi (Kinalioğlu, 2019).

RF metodu kısaca bagging ve random subspace tekniklerinin kullanıldığı birden fazla karar ağacından en çok oy alan sınıflandırma sonucunu kabul eden bir sınıflandırma yöntemidir (Şekil 3.13). Bagging ile hâlihazırda var olan veri kümesinden rasgele yeni veri kümeler ile istenen sayıda karar

ağacı yapısı oluşturularak random subspace yöntemi ile her karar ağacı yapısı için rasgele farklı değerlerde değişkenler kullanılmaktadır (Kinalioğlu, 2019).

3.2.6.6. Uyarlanabilir Artırma (Adaptive Boosting - Adaboost)

Artırma (boosting), dengesiz veri kümelerinin sınıflandırma performansını iyileştirmek için tasarlanmış bir meta öğrenme tekniğidir. Son yıllarda geleneksel sınıflandırıcıların dengesiz veri kümeleri üzerindeki performansını iyileştirmeye yönelik veri düzeyi ve algoritmik düzey açısından önemli çabalar sarf edilmiştir. Sınıf dağılımları doğrudan manipüle edilerek, pozitif örnekler veri kümesinde daha iyi temsil edebilmektedir (aşırı örnekleme). Algoritmik düzeyde azınlık ve çoğunluk sınıflarına farklı maliyetler atanabilmekte ve bu durum amaç fonksiyonunu etkin bir şekilde güncellemektedir. Ayrıca azınlık ve çoğunluk sınıflarına farklı ağırlıklar verilebilmekte ve böylece bir sorgu noktasının sınıf etiketi belirlenirken azınlık sınıfı örnekleri daha fazla önem kazanmaktadır (Seiffert vd., 2008).

AdaBoost (Freund & Schapire, 1996), eğitim setindeki örnekler üzerinde ağırlıkları uyarlanabilir bir şekilde ayarlayarak sıralı olarak bir dizi temel sınıflandırıcı oluşturmaktadır. AdaBoost'ta, mevcut sınıflandırıcı tarafından yanlış sınıflandırılan örneklere daha yüksek ağırlıklar verilir ve bunun tersi de geçerlidir. Bunu yaparak, sınıflandırıcılar önceki sınıflandırıcılar tarafından düzgün bir şekilde tanımlanmayan örneklere yoğunlaşmaya zorlanmakta ve çeşitli temel sınıflandırıcılar oluşturulabilmektedir. Teoride, AdaBoost'un eğitim hatasının üst sınırı, temel sınıflandırıcı rastgele bir tahminden biraz daha doğru olduğu sürece monoton olarak azalmaktadır. Her bir temel sınıflandırıcı, nihai karar fonksiyonunu oluşturmak için birleştirildiğinde de ağırlıklandırılmaktadır (Yuan & Ma, 2012).

Boosting yöntemlerinin bagging yöntemlerinden farkı eğitilen tahmin edicilerin sırayla eğitilmesi ve her birinin bir önceki eğitim sonucunu düzeltmeye çalışmasıdır. Adaboost metodu karar ağaçlarının tahminlerini kullanması bakımından RF metoduna benzer. AdaBoost'ta karar ağaçlarının derinliği bir düğüm iki yaprak şeklindedir.

AdaBoost'un ilk adımında derlemde ki her örnek, sınıflandırma açısından ne kadar önemli olduğunu gösteren bir ağırlıkla ilişkilendirilir. Başlangıçta tüm örneklere aynı ağırlıklar atanır. Her nitelik için bir derinliğinde bir karar ağacı oluşturulur ve verileri sınıflandırma içinde bu karar ağaçları kullanılır. Her bir ağaç tarafından yapılan tahminler eğitim setindeki gerçek etiketlerle karşılaştırılır. Eğitim örneklerini sınıflandırmada en iyi işi yapan özellik sonucuna karşılık gelen ağaç, ormandaki bir sonraki ağaç olur.

$$x_i \in R^n,\ y_i \in (-1,\ 1) \qquad (3.33.)$$

Veri kümesinde N toplam veri noktası sayısı olduğu varsayılırsa n gerçek sayıların boyutu veya veri kümesindeki özniteliklerin sayısıdır. Veri noktaları kümesi x ve y birinci veya ikinci sınıfı ifade eden bir ikili sınıflandırma problemi olduğu için -1 veya 1 olan hedef değişkendir (örneğin, küfürlü ve küfürsüz).

Her veri noktası için ağırlıklı örnekler hesaplanır. AdaBoost, eğitim veri kümesinin önemini belirlemek için her eğitim örneğine ağırlık atanır. Atanan ağırlıklar yüksek değerler içerdiğinde veri noktaları eğitim setinde daha etkilidir. Benzer şekilde, atanan ağırlıklar düşükse eğitim veri setinde etki düzeyleri de düşüktür. Başlangıçta, tüm veri noktaları aynı ağırlığa (w) sahiptir.

$$w = \frac{1}{N} \in [0, 1] \tag{3.34.}$$

Her bir ağırlığın değeri her zaman 0 ile 1 arasında olur ve veri noktalarını sınıflandırmanın bu sınıflandırıcıya gerçek etkisi hesaplanır.

$$a_t = \frac{1}{2} \ln \frac{(1\text{-ToplamHata})}{\text{ToplamHata}} \tag{3.35.}$$

a_t, ağacın son sınıflandırmada ne kadar etkisi olacağını bulur. ToplamHata, eğitim seti için toplam yanlış sınıflandırma sayısının eğitim seti boyutuna bölünmesidir. Karar ağacı iyi sonuç verdiğinde veya herhangi bir yanlış sınıflandırmaya sahip olmadığında sıfır hata oranıyla ve büyük pozitif bir a_t değeriyle sonuçlanır.

Karar ağacı yarısı doğru, yarısı yanlış sınıflandırma yaparsa hata oranı 0.5 ve a_t değeri 0 olur. Karar ağacı yanlış sınıflandırılmış sonuçlar verdiğinde a_t değeri negatif değer alır. Her ağaç için ToplamHata değeri gerçek değerlerini aldıktan sonra, her veri noktası için başlangıçta 1/N olarak alınan örnek ağırlıkları Eşitlik (36)'daki denklem ile güncellenir.

$$w_i = w_{i-1} e^{\mp a} \tag{3.36.}$$

Örneklerin yeni ağırlığı, eski örnek ağırlığının Euler sayısı ile çarpımı sonucunda artı veya eksi yükseltilmiş a_t değerine eşit olur. a_t'nin pozitif değeri için tahmin edilen ve gerçek çıktı aynıdır, örnek doğru sınıflandırılmıştır. a_t'nin negatif değeri için tahmin edilen değer ile gerçek çıktı değeri uyuşmaz. Bundan dolayı da "örnek yanlış sınıflandırılmıştır" sonucu elde edilir.

3.2.6.7. Gradyan Artırma (Gradient Boosting)

Gradient Boosting metodu, verileri eksik değerlere ihtiyaç duymadan esnek bir şekilde işleyen sınıflandırma ve regresyon problemlerini etkili bir biçimde çözen etkili sınıflandırıcılardandır. Karar ağaçlarındaki fazla uyum ve bias bir grup ağaç kullanılarak gradyan artırma ile önemli ölçüde azaltılır (Morris & Yang, 2021).

Gradient Boosting sınıflandırıcılarının amacı, eğitim örneklerinin gerçek sınıf değeri ile tahmin edilen sınıf değeri arasındaki kaybı en aza indirmektir. Sınıflandırıcının kaybını azaltma süreci bir sinir ağındaki gradyan inişine benzer şekilde çalışır. Gradient Boosting sınıflandırıcısı türevlenebilir bir kayıp fonksiyonuna bağlıdır. Gradient Boosting zayıf öğrenicileri alarak regresyon karar ağaçlarını kullanmakta ve bu regresyon ağaçları gerçek değerleri vermektedir. Çıktılar gerçek değerler olduğundan, modele yeni öğrenenler eklendikçe, tahminlerdeki hataları düzeltmek için regresyon ağaçlarının çıktıları ile birlikte düzenlenmektedir.

Veri kümesindeki girdi değerleri $\{x_i,y_i\}_{i=1}^{N}$ ve kayıp fonksiyonu $L(y_i, F(x))$ olmak üzere, Gradient Boosting modelin regresyon fonksiyonu olan $f_0(x)$ sabit başlangıç değeri için açılımı aşağıdaki gibi ifade edilir.

$$f_0(x) = \text{argmin}_\gamma \sum_{i=1}^{N} L(y_i,\gamma) \tag{3.37.}$$

Karar ağacı sayısı m, i = 1, 2, …, N her bir örnek değer için x değişkenine göre kayıp fonksiyonunun türevi alınır ve hata değerini belirlemek için -1 ile çarpılır.

$$\gamma_{im} = -\left[\frac{\partial L(y_i,\ f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}} \tag{3.38.}$$

Veri kümesindeki her örnek için x değerine göre kayıp fonksiyonun türevi alınır ve gerçek değer tahmini için -1 ile çarpılır. Temel öğrenici $h_m(x)$, ile uyumlu olması i için eğitim seti eğitim sürecinden geçirilir. Yapılan optimizasyon sonucunda hedef değerin azalmasının sağlandığı gradyan inişi ile γ_m değeri elde edilir.

$$\gamma_m = \text{argmin}_\gamma \sum_{i=1}^{N} L(y_i,F_{m-1}(x_i) + \gamma h_m(x_i)) \tag{3.39.}$$

Gradient Boosting modelinde, elde edilen değerler güncellenerek $F_m(x)$ değeri elde edilir.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{3.40.}$$

3.2.6.8. Aşırı Gradyan Artırma (eXtreme Gradient Boosting - XGBoost)

XGBoost metodu Gradient Boosting modelinin eğitim aşamasını kısaltmak ve aşırı uyum problemini kaldırmak için ortaya çıkmış bir gradyan yükseltme temelli boosting yöntemidir (Friedman, 2001). Gradyan artırmaya benzer şekilde bu metot zayıf bir temel sınıflandırıcıyı daha güçlü bir sınıflandırıcıyla birleştirmektedir. Eğitim sürecinin her yinelemesinde, bir temel sınıflandırıcının ağırlığı bir sonraki sınıflandırıcıda amaç fonksiyonunu optimize etmek için kullanılmaktadır. XGBoost, veri bilimi problemlerini daha hızlı ve doğru bir biçimde çözmeye çalışan paralel bir ağaç güçlendirme işlemi sağlamaktadır (Shi vd., 2019).

Metodun istenen amaç fonksiyonu L^t, eğitimin kayıp fonksiyonu ile düzenlileştirme teriminin ($\Omega(f_t)$) toplamı sonucu elde edilmektedir.

$$L^t = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) \tag{3.41.}$$

i örnek niteliğinin doğru sınıf etiketini y_i, i özniteliğinin tahmini sınıf etiketini $\hat{y}_i$, temel öğrenme karar ağacı fonksiyon değerlerini f_t, eğitim veri kümesindeki örnek yorumların sayısını n temsil etmektedir.

$$\Omega(f_t) = \gamma T_t + \frac{1}{2}\lambda \sum_{J=1}^{T} w_j^2 \tag{3.42.}$$

Düzenlileştirme fonksiyonu $\Omega(f_t)$ değerini elde etmek için γ ve λ hiperparametre değerleridir, ağaçtaki yaprak sayısını T ve yaprakların ağırlık değerlerini ise w göstermektedir.

$$L^t \cong \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{t-1}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{3.43.}$$

İkinci dereceden Taylor açılımı kayıp fonksiyonun sonucunu tahmin etmek için kullanılmakta ve g_i birinci dereceden h_i ikinci dereceden kayıp fonksiyonun gradyan değerini vermektedir.

3.2.6.9. Derin Öğrenme (Deep Learning – DL) Modelleri

3.2.6.9.1. Konvolüsyonel Sinir Ağı (Convolutional Neural Network - CNN)

CNN derin öğrenme modeli konvolüsyon, havuzlama (pooling) ve tam bağlı katmanların birleşiminden oluşan bilgisayarlı görüş ve görüntü işleme gibi alanlar için önerilen birçok katmanlı yapay sinir ağı modelidir. CNN derin öğrenme modelinde ilk adım veri kümesindeki örneklemlerin ilişkili yerel özniteliklerin çıkartılması ikinci adım ise çok katmanlı sinir ağı yapısı ile bir sınıflandırma sonucu elde edilmesidir (Şekil 3.14).

Konvolüsyon, havuzlama ve düzleştirme (flatten) işlemi ilk adımı oluşturmaktadır. Konvolüsyon işlemi ilgili pozisyondaki girdi örnekleri üzerinde filtrelerin içindeki değerlerin gezdirilerek çarpılması ve toplanması işlemidir. Konvolüsyon işleminin amacı derlemdeki örneklerin derlem içerisinde nerede olduğunu belirlemeye yarar. Konvolüsyon filtresinden geçirilen örneklem kümesinin yeni temsiline öznitelik haritası (feature map) adı verilmektedir. Çıkış matrisi w_{jm}^{l}, bias değeri b_j^l, aktivasyon fonksiyonu f değeri olmak üzere öznitelik haritaları I_j^l Eşitlik (39) ile elde edilmektedir.

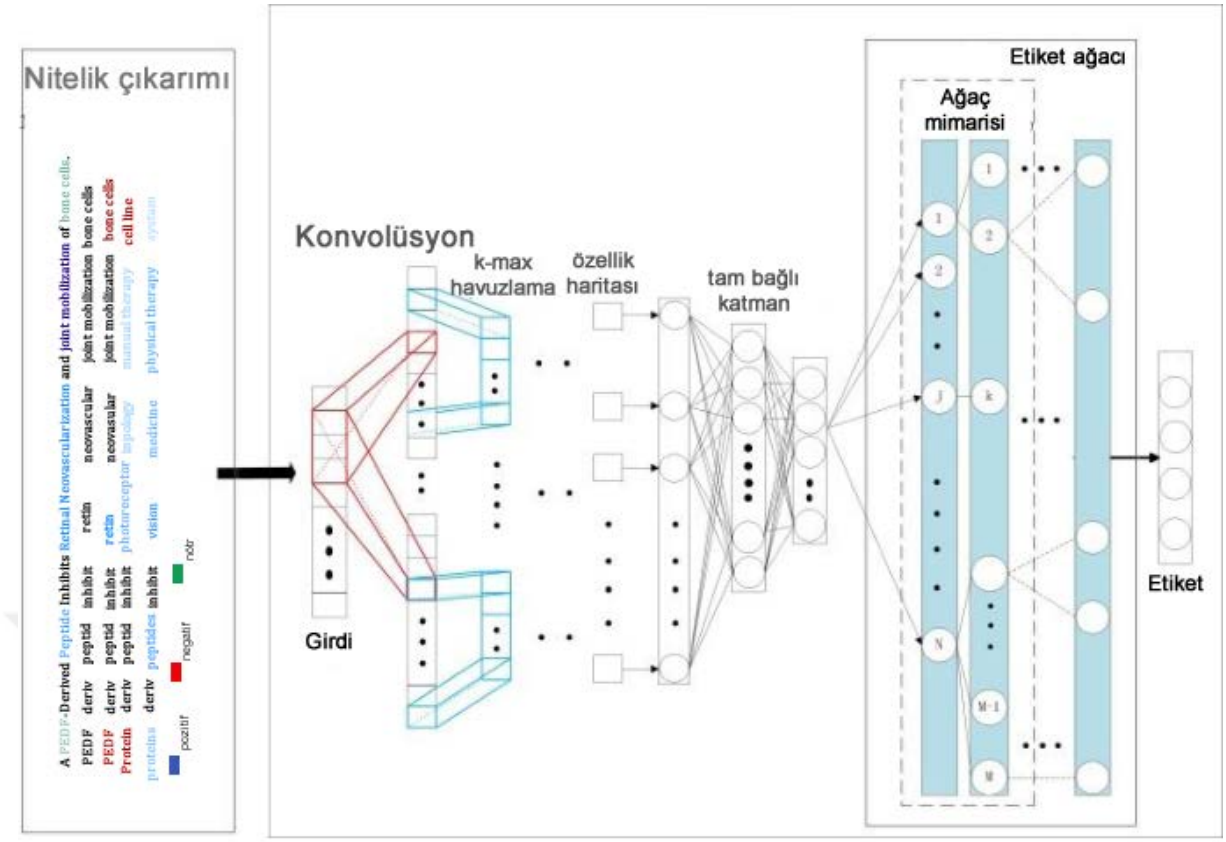
$$I_j^l = f\Big(\sum_m I_j^{l-1} \otimes w_{jm}^l + b_j^l\Big) \tag{3.44.}$$

Derlem içindeki bir örneklere uygulanan filtre sayısı adedince öznitelik haritası elde edileceğinden daha az filtre sayısı ile maliyet azaltılmak istenmektedir. Eğitimde kullanılan rasgele konvolüsyon ağırlık matrisi (filtrenin parametreleri) değerleri eğitim boyunca güncellendiğinden değerler ideal değere yaklaşmaktadır.

G x G boyutunda bir örnekleme D x D boyutunda bir konvolüsyon filtresi 1 adım boyu (s) uygulandığında elde edilen öznitelik haritasının boyutu (G− D)/s + 1 işlemiyle bulunmaktadır. Öznitelik haritasına uygulanan ReLU aktivasyon fonksiyonu negatif değerler üreten nöronların (öznitelik değerleri $x \leq 0$ olan değerler) sıfır değerini almasını sağlamaktadır. Değerlerin $x > 0$ olması durumunda ise x değeri üretmektedir. ReLU fonksiyonunun türevi sıfıra yakınsamadığı için patlayan gradyan olayı görülmemektedir.

$$f(x) = (0, \text{eğer } x < 0 \; x, \text{eğer } x \geq 0 \tag{3.45.}$$

$$x_j^l = \max\Big(0, \sum_{i \epsilon M_J} I_j^l + b_j^l\Big) \tag{3.46.}$$

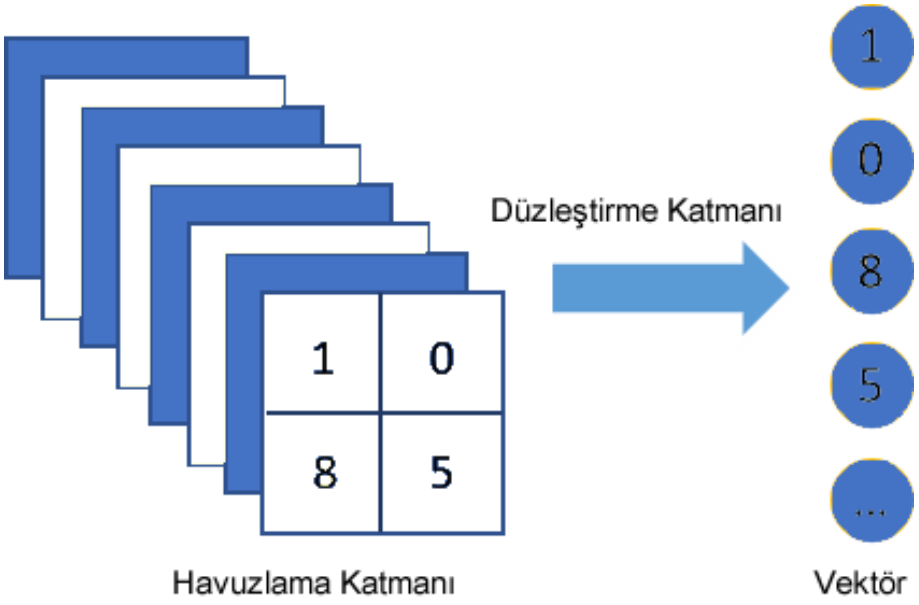


Şekil 3.14. CNN yönteminin mimarisi (Yan, 2016).

Aktivasyon işleminden sonra özniteliklerin çeşitlilikleri azaltmak için havuzlama işlemine tabii tutularak öznitelik haritalarının boyutu azaltılmış olmakta ve boyuttaki bu azalmaya rağmen öznitelik bilgisi değişmemektedir. Birçok havuzlama yöntem vardır ama en sık kullanılanı maksimum havuzlama (max-pooling) işlemidir.

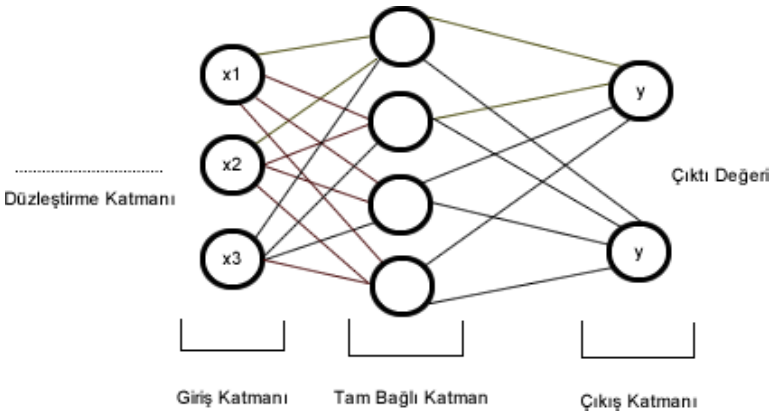
$$x_j^{l+1} = f_p(\beta_j^{l+1}(x_j^l) + b_j^{l+1}) \tag{3.47.}$$

Konvolüsyon ve havuzlama adımları model içinde tekrarlanabilmektedir. Düzleştirme katmanı tam bağlı katmana verileri hazırlamak için konvolüsyon ve havuzlama katmanlarından gelen matrislerden oluşan verileri tek boyutlu dizilere dönüştürmektedir (Şekil 3.15).



Şekil 3.15. Düzleştirme katmanında matrislerin diziye dönüştürülmesi.

Tam bağlı katman ağırlıklar, bias değerleri ve nöronlardan oluştuğu bilinen yapay sinir ağıdır (Şekil 3.16). Tam bağlı katmanda diğer katmanlardan elde edilen örneklemlere ait özellik vektörlerinin sınıflandırılmasını gerçekleştiren katmandır. Tam bağlı katman gelen vektörlerin işlem görmesi ile modelin ileri yayılım işlemi tamamlanmaktadır. Elde edilen sonuç ile bu defa geri yayılım algoritması başlatılmaktadır. Ağırlıkların rasgele atanmasından dolayı hata değeri yüksektir ve geri yayılım ile ağrılıklar tekrar güncellenmekte ve modelin sınıflandırma başarısının artırılması hedeflenmektedir.



Şekil 3.16. CNN metodunda tam bağlı katmanlar.

Modele ait hata fonksiyonu, hedeflenen çıktı değeri ile elde edilen çıktı değerinin farkı n tane örnek için ortalamaları alınarak bulunmaktadır.

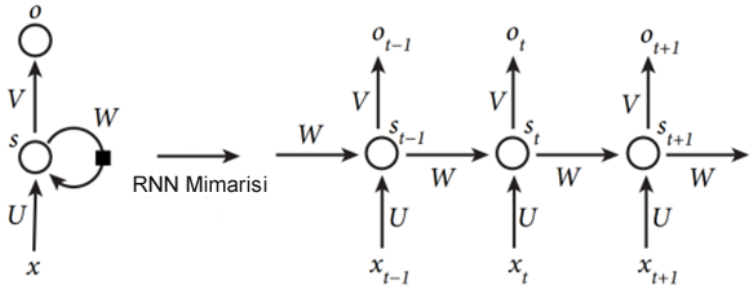
$$H = \frac{1}{n}\sum_{i=1}^{n} (\text{hedef} - \text{eldeedilen})^2 \quad (3.48.)$$

Hata değeri yüksek bir noktadan sıfır noktasına doğru azaltılması için hatanın ağırlığa göre türevinin alınması gerekmektedir. Geri yayılım boyunca ağırlıklar ile hata değeri arasında türev hesabı yapılarak ağırlık güncelleme işlemi gerçekleştirilmektedir. Hata değerinin sıfıra doğru azaltılabilmesi için hatanın ağırlığına göre türevi alınmaktadır.

$$w = w_i - \eta \cdot \frac{dH}{dw} \quad (3.49.)$$

3.2.6.9.2. Uzun Kısa Süreli Bellek (Long Short-Term Memory - LSTM)

Yapay sinir ağları veri kümesindeki girdi değerlerini bağımsız bir biçimde işlemektedir. DL yöntemlerinden biri olan RNN, girdi değerleri arasında ilişki durum bilgilerini tutmaktadır ve t anındaki girdi değeri (t-1) anındaki çıktı değerine bağlıdır. Çıktı değerlerinin yeni girdi değerleri olarak işlem gördüğü RNN, duygu analizi, metin sınıflandırma, konuşmadan metne ve makine çevirisi gibi çeşitli doğal dil işleme görevlerinde yaygın olarak kullanılmaktadır. RNN yöntemleri sıralı modellemede dayanıklı olmasına rağmen, uzun vadede kaybolan ve patlayan gradyanlardan mustariptir. Kaybolan ve patlayan gradyan problemi aktivasyon fonksiyonunun kullanımından sonra geriye yayılım için türevinin değerinin sıfır olmasıdır (Mansoor vd., 2020). RNN modelleri girdi değerleri arasındaki bilgileri sakladığından hafızalı dinamik modeller olarak adlandırılmaktadır (Chollet, 2018; Staudemeyer & Morris, 2019). Durum bilgileri her döngü için bir sonraki girdi değerine bilgi taşır (Şekil 3.17).

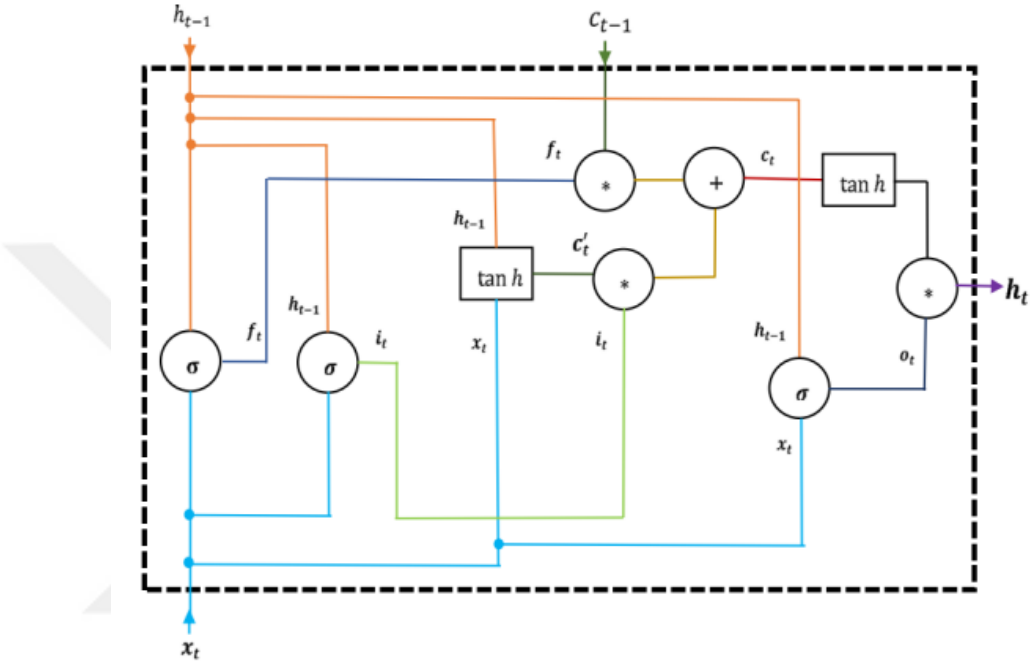


Şekil 3.17. RNN mimarisi ve zaman serisi boyunca açılımı (LeCun vd., 2015).

Fonksiyon f herhangi bir aktivasyon fonksiyonunu, x_t t anındaki girdi değerini, u (girdi değeri ile gizli düğüm arasında), w (gizli düğüm), v (gizli düğüm ile çıktı değeri arasında) ağırlık parametrelerini temsil etmek üzere s_t değeri Eşitlik (45)'teki gibi hesaplanmaktadır.

$$s_t = f(u.x_t + w.s_{t-1}) \quad (3.50.)$$

RNN metotlarındaki gradyan problemlerini çözmek için geliştirilen LSTM metodu, bu sorunu unutma ve çıkış kapıları aracılığıyla çözmeye çalışmıştır. Bu metot, giriş kapısı (i_t), unutma (f_t) ve çıkış kapılarından (o_t) oluşmaktadır. Unutma kapısı mevcut girdi (x_t) ve bir önceki durum bilgisine (h_{t-1}) bakarak unutulması gereken duruma karar verir. Giriş, unutma ve çıkış kapıları gizli durum bilgilerinin (h_t) ve bellek hücrelerinin (c_t) güncellenme durumlarına karar veren birimlerdir (Şekil 3.18).



Şekil 3.18. LSTM metodunun mimarisi (Salur, 2021).

Giriş kapısı ağırlık değeri w_i, unutma kapısı ağrılık değeri w_f ve çıkış kapısı ağırlık değeri w_o, Sigmoid aktivasyon fonksiyonu σ, bias değeri b, o_t ise t anındaki düğüm çıktı değeri olmak üzere LSTM kapıları arasındaki matematiksel denklemler, aşağıdaki gibidir.

$$i_t = \sigma(w_i.[h_{t-1}, x_t] + b_i) \quad (3.51.)$$

$$f_t = \sigma(w_f.[h_{t-1},x_t] + b_f) \quad (3.52.)$$

$$o_t = \sigma(w_o.[h_{t-1},x_t] + b_o) \quad (3.53.)$$

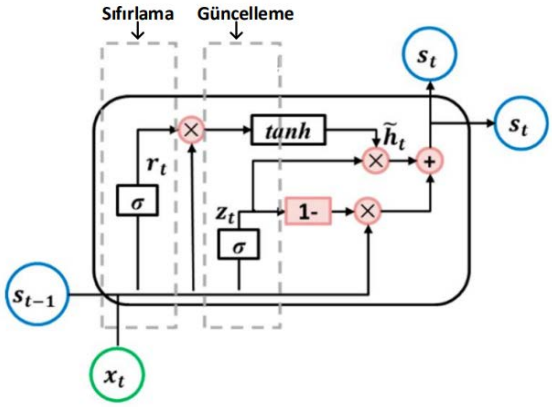
$$c_t^{'} = \tanh(w_c.[h_{t-1},x_t] + c) \quad (3.54.)$$

$$c_t = f_t.c_{t-1} + i_t.c_t^{'} \quad (3.55.)$$

$$h_t = o_t.\tanh(c_t) \quad (3.56.)$$

3.2.6.9.3. Geçitli Tekrarlayan Birim (Gated Recurrent Unit - GRU)

Kaybolan gradyan sorununa karşı geliştirilen bir diğer algoritma olan GRU, sorunu bir sıfırlama kapısı ve bir güncelleme kapısı ile çözmeye çalışmıştır. LSTM modelindeki unutma ve giriş kapılarının görevini güncelleme kapısı yapmaktadır. Güncelleme kapısı bilgilerin saklanması ve ya atılması durumlarında karar mekanizması görevi yapmaktadır. Sıfırlama kapısı ise, güncellenen bilgiler geçmişte elde edilen bilginin hangi kısımlarının göz ardı edilip edilmeyeceğine karar vermektedir (Şekil 3.19). GRU algoritması LSTM yöntemine göre daha az hesaplama yapmakta ve bundan dolayı daha az belleğe ihtiyacı olmaktadır.



Şekil 3.19. GRU metodunun mimarisi (Salur, 2021).

Sıfırlama kapısı çıktı değeri r_t, güncelleme kapısı çıktı değeri z_t, güncelleme kapısı ağırlık değeri w_z, sıfırlama kapısı ağırlık değeri w_r, çıkış kapısı ağırlık değeri w_s, kapıların bias değeri b_z, b_r ve b_s olmak üzere GRU düğümünün denklemsel ilişkileri aşağıda verilmiştir.

$$z_t = \sigma\left(w_z.[s_{t-1},x_t] + b_z\right) \quad (3.57.)$$

$$r_t = \sigma\left(w_r.[s_{t-1},x_t] + b_r\right) \quad (3.58.)$$

$$s_t^{'} = (\tanh\left(w_s.[r_{t-1},x_t\right] + b_s) \quad (3.59.)$$

$$s_t = (1\text{-}z_t).s_{t-1} + z_t + s_t^{'} \quad (3.60.)$$

3.2.6.9.4. Çift Yönlü LSTM (Bidirectional LSTM - BiLSTM)

Günümüzde metin sınıflandırma, ses tanıma, duygu analizi için sıklıkla kullanılan, önceki ve sonraki bağlamları dikkate alan BiLSTM algoritması ileri ve geri katmanları birleştirerek sıralı metin problemlerini daha iyi çözmektedir (Kapočiūtė-Dzikienė & Tesfagergish, 2020). LSTM'nin

geliştirilmiş versiyonu olan çift yönlü BiLSTM veri kümesi girdileri bir baştan sona (ileri - sağ yön) ve sondan başa (geri - sol yön) doğru işlenmektedir. BiLSTM metodunda bulunan iki paralel katmandan geri yayılım yönünde olan sonraki özelliklerin geriye doğru bir LSTM katmanı tarafından yakalanması ve ileri yayılım yönünde olan ise önceki niteliklerin ileri bir LSTM katmanı ile çıkarılması görevini görmektedir.

Girdi veri değerleri (x_{t-1}, x_t , x_{t+1}), ileri LSTM gizli durum bilgisi $\vec{h}_t$, geri LSTM gizli durum bilgisi $\overleftarrow{h}_t$, ileri LSTM ağırlığı $w_{\vec{h}y}$, geri LSTM ağırlığı $w_{\overleftarrow{h}y}$, çıktı katmanındaki bias değeri b_y olmak üzere BiLSTM y_t çıkışı aşağıdaki denklemlerde gösterildiği gibi hesaplanmaktadır.

$$\vec{h}_t = \text{LSTM}(x_t, \vec{h}_{t-1}) \tag{3.61.}$$

$$\overleftarrow{h}_t = \text{LSTM}(x_t, \overleftarrow{h}_{t+1}) \tag{3.62.}$$

$$y_t = w_{\vec{h}y}.\vec{h}_t + w_{\overleftarrow{h}y}.\overleftarrow{h}_t + b_y \tag{3.63.}$$

3.2.7. BERT Sınıflandırma Modeli

Google'ın geliştirdiği BERT tekniği, önceden eğitimli büyük miktarda etiketlenmemiş metinleri kullanarak genel amaçlı dil temsil modellerini eğitmektedir. BERT modeli, çift yönlü transformatör mimarisini uygulayan denetimsiz derin çift yönlü bir sinir ağıdır. BERT tabanlı aktarım öğrenme yaklaşımı, sınıflandırma performansının artmasına ve eğitim süresinin azalmasına yol açtığı için nefret sınıflandırması çalışmalarında sıklıkla kullanılmaya başlanmıştır (Sharma vd., 2022). Transfer öğrenme yaklaşımı ayrıca önceden eğitilmiş bir modelle sınırlı etiketlenmiş verilerden etkili öğrenme sağlamaktadır. Önceden eğitilmiş bir dil modeli, az etiketli veri kaynaklarında bile mevcut dili anlamayı kolaylaştırmaktadır.

Önceden eğitimli dil modeli BERT, önemli miktarda veri ile önceden eğitilmiştir ve bu eğitilmiş verinin diğer daha küçük dil işleme görevlerine aktarılmasına izin verilmiştir. BERT modelin önceden eğitilmesi aşaması çok fazla kaynak gerektirir ancak bu görev önceden eğitilmiş dil modelinde zaten yapılmıştır. Araştırmacılar, çıktı katmanını ekleyerek farklı dil görevleri için sürece ince ayar yapabilmektedir. Bu durum eğitim için önemli ölçüde daha az kaynak gerektirdiğinden eğitim süresini en aza indirmektedir (Devlin vd., 2018).

BERT modeli mantığı, bir metindeki kelimeler arasındaki bağlamsal ilişkileri öğrenen dikkat mekanizmasına, yani dönüştürücü yapısına dayanmaktadır. Temel bir transformatör yapısı, metin girişlerini okuyan bir kodlayıcı ve görev için tahminler üreten bir kod çözücüden oluşmaktadır. BERT modeli, girdi verisi olarak 512 kelimeden daha az bir diziyi alır ve çıktı olarak verinin bir temsilini verir. Tokenizasyon, WordPiece belirteci (Johnson vd., 2017) ile iki adımda (ön metin normalleştirme ve noktalama ayırma) gerçekleştirilmektedir. Belirteçleştirilmiş dizi, her cümlenin başına bir [CLS]

belirteci ve her cümlenin sonuna bir [SEP] belirteci eklenerek elde edilmektedir. BERT modeli, ortaya çıkan belirteç dizilerinin bir temsili olarak ilk belirtecin [CLS] son gizli h durumunu kullanarak metin sınıflandırması gerçekleştirmektedir (Smetanin & Komarov, 2021).

M-BERT modeli, 104 dilden oluşan Wikipedia külliyatında eğitilmiş önceden eğitilmiş bir dil modelidir (Pires vd., 2019). Bu modelin en önemli başarısı, 104 farklı dilden oluşan veriler üzerinden önceden eğitilmiş olması ve düşük kaynaklı dillerde bile oldukça iyi performans gösterebilmektedir. Ayrıca M-BERT modeli tüm dillerin yapılarını dikkate alarak eğitim gerçekleştirmektedir (Guven, 2021). Bir M-BERT modeli birçok ana dilde veriye sahiptir ve ortak dillerde de yeterli veriye sahiptir. Büyük veri kaynaklarına sahip M-BERT, diğer dillerdeki veri kümelerini sınıflandırmak için kolayca kullanılabilmektedir (Yoo & Jeong, 2020).

Yine COVID-19 duygu analizi sınıflandırma çalışmasında Python'da önceden eğitilmiş model ktrain kütüphanesi (Maiya, 2020) ile kullanılmıştır. Çok çeşitli görevleri çok az kod satırıyla çözmeye çalışan ktrain paketi basit bir ara yüze sahiptir. Eğitim modeli oluşturmak için ktrain algoritması model inceleme ve model eğitme gibi adımların kolayca tamamlanmasını sağlamaktadır. Dil ve karakter kodlaması otomatik olarak algılanan ktrain metodunda diğer işlemler buna göre devam etmektedir. Hedeflerin sayısal veya kategorik olup olmadığı otomatik olarak analiz edilmektedir. Model en uygun şekilde yapılandırılmaktadır.

Önceden eğitilmiş M-BERT modeli, rastgele başlatılan bir son yoğun (dense) katmanla yüklenmektedir. Son yoğun katman rastgele başlatılsa da eğitim sürecinde güncelleme yapılmaz. Ağırlıklar, M-BERT algoritmasının önceden eğitilmiş katmanlarından çıkarılan bağlamsal bilgileri analiz etmek için yeni etiketlenmiş bir veri kümesi kullanılarak güncellenmektedir. Hiçbir katman durdurulmadığından ve tüm model katmanları eğitilebilir olduğundan, ağırlıkları geri yayılımda güncellenmektedir. M-BERT modeli, ktrain yapısı içinde "fit_onecycle" yöntemi ile beş terim için eğitilmiştir. Varsayılan ktrain konfigürasyonu, kategorik çapraz entropi kaybı fonksiyonu ve Softmax tahmin katmanı ile 5 periyot için 2×10^{-5} öğrenme hızı ile kullanılmıştır.

4. BULGULAR ve TARTIŞMA

4.1. Küfürlü Söylemler Analizi

Küfürlü söylemler analizi için yapılan değerlendirmeler ATC veri setinin iki versiyonunda ayrı ayrı gerçekleştirilmiş ve sınıflandırma sonuçları 10 kat çapraz doğrulama kullanılarak elde edilmiştir. Pozitif ve negatif sınıfların tespitinin doğruluğunu belirlemek için uygun performans ölçümleri kullanılarak sonuçların çapraz katlar arasında ortalaması alınmıştır.

Uygulanan tüm sınıflandırıcılarda Python tabanlı Scikit-Learn ve Keras kütüphanesi kullanılmış ve test edilmiştir. TML ve topluluk sınıflandırıcılarının uygun parametre değerleri ızgara arama (grid-search) ile belirlenmiştir. Izgara arama, bir ML sınıflandırıcısının varsayılan olarak belirtilen bir alt kümesi aracılığıyla sınıflandırma başarısına hiperparametrelerin etkisini bulmaya çalışan bir arama yöntemidir (Aya vd., 2016).

Küfürlü söylemler analizinde modellerin sınıflandırılması için tüm değerlendirmeler Google'ın Colaboratory (Colaboratory, 2022) hizmetinde gerçekleştirilmiştir. Google Colaboratory kurulum gerektirmeyen ve tamamen bulutta çalışan bir Jupyter notebook ortamıdır. Eğitim süresi, veri kümesinin %90'ı ile belirtilen metot kombinasyonunu kullanarak kategorizasyon modelini eğitmek için gereken süredir. Test süresi, veri setinin %10'u ile aynı metot kombinasyonunu kullanarak sınıflandırma modelini test etmek için gereken süredir.

Veri temizleme aşamasının sonuçlara etkisini gözlemlemek için ön işleme adımı noktalama işaretleri, hashtagler, Türkçe durak sözcükler, emojiler ve web bağlantıları temizlenerek ve temizlenmeden şeklinde gerçekleştirilmiştir. Hem Veri Temizleme (Data Clean - DC) hem de Veri Temizleme Yok (Non Data Clean - NDC) denemelerinde küfürle bağlantılı olduğu düşünülen emoji (Unicode: 1F4A9) kaldırılmamıştır. Bölüm 4.1.1, Bölüm 4.1.2 ve Bölüm 4.1.3. Türkçe küfürlü yorumları algılama modelinin teknik ayrıntılarını ve değerlendirme sonuçlarını sunmaktadır.

4.1.1. Geleneksel Makine Öğrenmesi (Traditional Machine Learning - TML) Modellerinin Değerlendirme Sonuçları

İyi bilinen NLP yaklaşımları genellikle ML tabanlı yöntemleri veya daha genel olarak ML tabanlı istatistiksel metotları benimser. ML tabanlı yöntemler genellikle geçmiş verilerden öğrenebilen akıllı modüllerden oluşmaktadır (Khan vd., 2016). Küfür söylemi tespiti çalışmasında TML tabanlı kategorizasyon modelleri tasarlanmış, DL tabanlı modellerle ve topluluk modelleriyle karşılaştırılmıştır.

Tablo 4.1, TML tabanlı sınıflandırıcılarda kullanılan hiperparametreleri göstermektedir. Tablo 4.2, Tablo 4.3, Tablo 4.4, Tablo 4.5, Tablo 4.6, 10 kat çapraz doğrulama kullanan TML tabanlı

modellerin Makro ve Mikro F1 puanlarını, κ değerlerini ve eğitim - test sürelerini vermektedir. Şekil 4.1, Şekil 4.2, Şekil 4.3, Şekil 4.4, Şekil 4.5, TML tabanlı sınıflandırıcıların Kesinlik ve Hassasiyet sonuçlarını göstermektedir.

Tablo 4.1. TML tabanlı sınıflandırıcılarda kullanılan hiperparametreler.

Sınıflandırıcı	Parametreler ve İfade/Değer
NB	Alpha = 0.1
SVM	C = 10.01, LinearSVC() işlevi
DT	rastgele durum değeri (random state value) = 25
RF	Karar ağacı sayısı = 50
LR	Scikit-Learn'deki varsayılan değerler

Tablo 4.2. SVM modelinin değerlendirme sonuçları (10 kat çapraz doğrulama).

Veri kümesi	Değerlendirmenin SVM Metot Kombinasyonu	Mikro F1	Makro F1	κ	Ort. Süreler (sn.)	
					Eğitim	Test
Orijinal ATC	NDC + BoW + TF-IDF + SVM	0.932	0.923	0.847	**1.577**	**0.041**
	NDC + bigram + TF-IDF + SVM	**0.933**	0.924	**0.849**	2.926	0.062
	NDC + trigram + TF-IDF + SVM	0.927	0.919	0.838	5.635	0.099
	DC + BoW + TF-IDF + SVM	0.932	0.923	0.846	1.710	0.048
	DC + bigram + TF-IDF + SVM	**0.933**	**0.925**	0.850	4.057	0.082
	DC + trigram + TF-IDF + SVM	0.930	0.922	0.845	6.689	0.108
Aşırı Örneklenmiş ATC	NDC + BoW + TF-IDF + SVM	0.972	0.972	**0.945**	**2.286**	**0.057**
	NDC + bigram + TF-IDF + SVM	0.970	0.970	0.940	4.958	0.107
	NDC + trigram + TF-IDF + SVM	0.967	0.964	0.928	7.754	0.149
	DC + BoW + TF-IDF + SVM	**0.973**	**0.973**	**0.945**	2.310	0.061
	DC + bigram + TF-IDF + SVM	0.971	0.971	0.942	5.428	0.110
	DC + trigram + TF-IDF + SVM	0.967	0.967	0.935	8.488	0.152

Türkçe küfürlü söylemler analizi için kullanılan SVM sınıflandırıcı kombinasyonlarının sonuçları incelendiğinde genel olarak sınıflandırma sürelerinin hızlı olduğu ve SVM için BoW nitelik seçiminin ngramlara göre daha iyi bir seçim olduğu gözlenmektedir. Verileri ön işlemeden geçirme ve

verileri artırarak dengelemenin SVM'nin en iyi sınıflandırma sonuç kombinasyonunu oluşturduğu gözlemlenmektedir.

Tablo 4.3. NB modelinin değerlendirme sonuçları (10 kat çapraz doğrulama).

Veri kümesi	Değerlendirmenin NB Metot Kombinasyonu	Mikro F1	Makro F1	κ	Ort. Süreler (sn.)	
					Eğitim	Test
Orijinal ATC	NDC + BoW + TF-IDF	**0.924**	**0.915**	**0.831**	0.997	0.050
	NDC + bigram + TF-IDF	0.921	0.911	0.823	2.566	0.089
	NDC + trigram + TF-IDF	0.919	0.908	0.817	4.058	0.123
	DC + BoW + TF-IDF	0.920	0.910	0.821	**0.914**	**0.044**
	DC + bigram + TF-IDF	0.920	0.910	0.819	2.301	0.075
	DC + trigram + TF-IDF	0.918	0.907	0.814	3.713	0.101
Aşırı Örneklenmiş ATC	NDC + BoW + TF-IDF	0.924	0.924	0.848	**1.273**	0.068
	NDC + bigram + TF-IDF	0.933	0.933	0.866	3.111	0.152
	NDC + trigram + TF-IDF	**0.934**	**0.934**	**0.868**	5.327	0.179
	DC + BoW + TF-IDF	0.921	0.920	0.842	1.395	**0.065**
	DC + bigram + TF-IDF	0.927	0.927	0.855	3.393	0.118
	DC + trigram + TF-IDF	0.929	0.929	0.859	5.230	0.162

Türkçe küfürlü söylemler analizi için kullanılan NB sınıflandırıcı kombinasyonlarının sonuçları incelendiğinde genel olarak sınıflandırma sürelerinin diğer sınıflandırıcılardan daha iyi olduğu ve NB için bigram nitelik seçiminin daha iyi bir seçim olduğu gözlenmektedir. Verileri ön işlemeden geçirmeden ve verileri artırarak dengelemenin NB'nin en iyi sınıflandırma sonuç kombinasyonunu oluşturduğu gözlemlenmektedir.

Türkçe küfürlü söylemler analizi için kullanılan RF sınıflandırıcı kombinasyonlarının sonuçları incelendiğinde genel olarak sınıflandırma sürelerinin TML sınıflandırıcılar içerisinde yüksek olduğu ve bigram nitelik seçiminin RF için iyi bir seçim olduğu gözlenmektedir. Verileri ön işlemeden geçirerek ve verileri artırarak kategorileri dengelemenin RF'nin en iyi sınıflandırma sonuç kombinasyonunu oluşturduğu gözlemlenmektedir.

Türkçe küfürlü söylemler analizi için kullanılan LR sınıflandırıcı kombinasyonlarının sonuçları incelendiğinde genel olarak sınıflandırma sürelerinin kısa olduğu ve BoW nitelik seçiminin LR için iyi bir vektörleştirici olduğu gözlenmektedir. Verileri ön işlemeden geçirerek ve verileri artırarak kategorileri dengeleme işlemleri LR'nin en iyi sınıflandırma sonuç kombinasyonunu oluşturduğu gözlemlenmektedir.

Tablo 4.4. RF modelinin değerlendirme sonuçları (10 kat çapraz doğrulama).

Veri kümesi	Değerlendirmenin RF Metot Kombinasyonu	Mikro F1	Makro F1	κ	Ort. Süreler (sn.)	
					Eğitim	Test
Orijinal ATC	NDC + BoW + TF-IDF	0.921	0.909	0.820	**29.911**	**0.503**
	NDC + bigram + TF-IDF	0.914	0.901	0.803	83.731	0.689
	NDC + trigram + TF-IDF	0.908	0.894	0.788	143.128	0.800
	DC + BoW + TF-IDF	**0.930**	**0.919**	**0.837**	34.908	0.678
	DC + bigram + TF-IDF	0.922	0.914	0.828	101.555	0.881
	DC + trigram + TF-IDF	0.921	0.909	0.816	165.700	0.995
Aşırı Örneklenmiş ATC	NDC + BoW + TF-IDF	0.966	0.966	0.933	**42.183**	**0.541**
	NDC + bigram + TF-IDF	0.967	0.964	0.933	115.373	0.760
	NDC + trigram + TF-IDF	0.965	0.966	0.931	197.081	0.916
	DC + BoW + TF-IDF	**0.970**	0.970	0.940	43.924	0.687
	DC + bigram + TF-IDF	**0.970**	**0.971**	0.939	121.880	0.909
	DC + trigram + TF-IDF	0.969	0.970	**0.941**	200.923	1.045

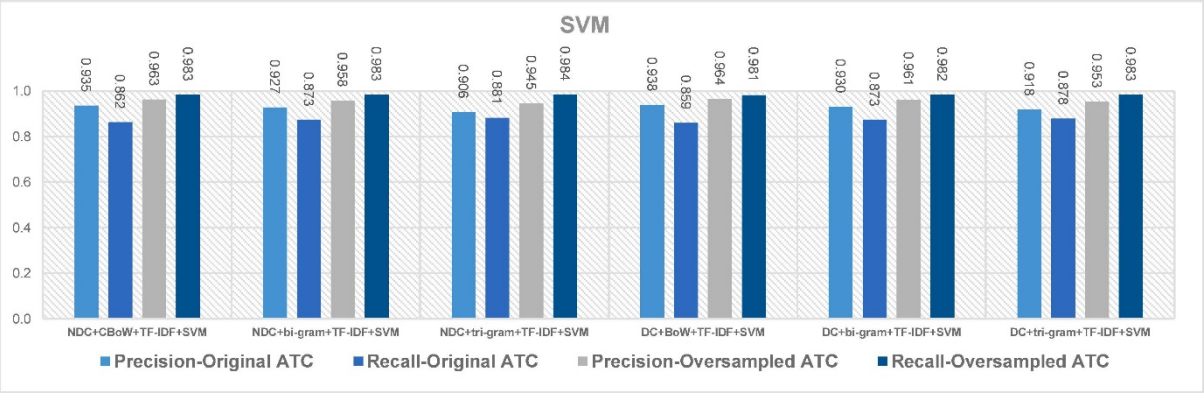
Tablo 4.5. LR modelinin değerlendirme sonuçları (10 kat çapraz doğrulama).

Veri kümesi	Değerlendirmenin LR Metot Kombinasyonu	Mikro F1	Makro F1	κ	Ort. Süreler (sn.)	
					Eğitim	Test
Orijinal ATC	NDC + BoW + TF-IDF	**0.907**	**0.892**	**0.785**	2.575	0.048
	NDC + bigram + TF-IDF	0.904	0.888	0.777	6.380	0.085
	NDC + trigram + TF-IDF	0.901	0.886	0.772	11.026	0.116
	DC + BoW + TF-IDF	0.906	0.890	0.782	**2.363**	**0.044**
	DC + bigram + TF-IDF	0.904	0.888	0.778	6.308	0.130
	DC + trigram + TF-IDF	0.903	0.886	0.774	10.474	0.098
Aşırı Örneklenmiş ATC	NDC + BoW + TF-IDF	0.917	0.917	0.833	2.982	0.064
	NDC + bigram + TF-IDF	0.911	0.911	0.822	7.607	0.119
	NDC + trigram + TF-IDF	0.904	0.904	0.807	13.215	0.168
	DC + BoW + TF-IDF	**0.918**	**0.918**	**0.837**	**2.787**	**0.060**
	DC + bigram + TF-IDF	0.916	0.916	0.833	6.983	0.109
	DC + trigram + TF-IDF	0.916	0.916	0.831	11.233	0.147

Tablo 4.6. DT modelinin değerlendirme sonuçları (10 katlı çapraz doğrulama).

Veri kümesi	Değerlendirmenin DT Metot Kombinasyonu	Mikro F1	Makro F1	κ	Ort. Süreler (sn.)	
					Eğitim	Test
Orijinal ATC	NDC + BoW + TF-IDF	0.904	0.893	0.713	32.926	0.088
	NDC + bigram + TF-IDF	0.899	0.887	0.775	108.646	0.104
	NDC + trigram + TF-IDF	0.888	0.877	0.754	201.349	0.157
	DC + BoW + TF-IDF	**0.914**	**0.904**	**0.810**	**32.383**	**0.058**
	DC + bigram + TF-IDF	0.907	0.897	0.794	102.488	0.089
	DC + trigram + TF-IDF	0.906	0.896	0.791	181.865	0.112
Aşırı Örneklenmiş ATC	NDC + BoW + TF-IDF	0.949	0.949	0.898	**29.937**	0.079
	NDC + bigram + TF-IDF	0.943	0.943	0.887	94.171	0.140
	NDC + trigram + TF-IDF	0.935	0.934	0.869	174.687	0.192
	DC + BoW + TF-IDF	**0.957**	**0.957**	**0.913**	34.623	**0.073**
	DC + bigram + TF-IDF	0.951	0.951	0.901	108.994	0.124
	DC + trigram + TF-IDF	0.947	0.947	0.894	189.594	0.161

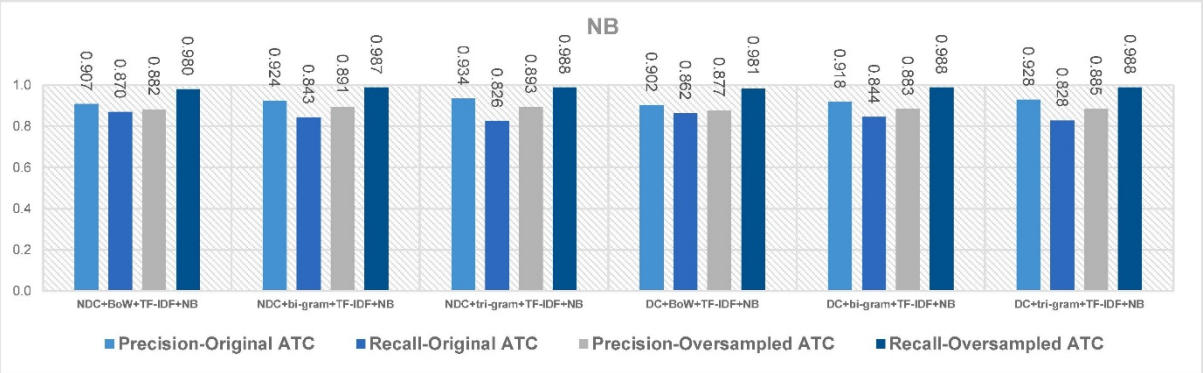
Türkçe küfürlü söylemler analizi için kullanılan DT sınıflandırıcı kombinasyonlarının sonuçları incelendiğinde genel olarak sınıflandırma sürelerinin RF'den daha iyi olduğu ve BoW nitelik seçiminin DT için iyi bir seçim olduğu gözlenmektedir. Verileri ön işlemeden geçirerek ve verileri artırarak dengeleme işlemleri DT'nin en iyi sınıflandırma sonuç kombinasyonunu oluşturduğu gözlemlenmektedir.



Şekil 4.1. Orijinal ATC ve aşırı örneklenmiş ATC üzerinde 10 kat çapraz doğrulama kullanan SVM model kombinasyonlarının ortalama Kesinlik ve Hassasiyet sonuçları.

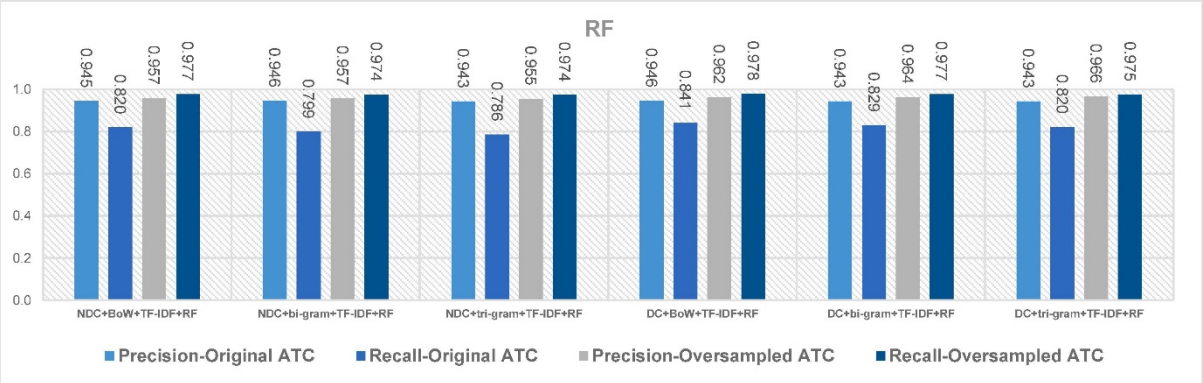
Şekil 4.1'de SVM kombinasyonları için en yüksek Hassasiyet değeri aşırı örneklenmiş ATC veri kümesi ile trigram nitelik seçiminde ve ön işleme yapılmadan elde edilmiştir. En yüksek Kesinlik

değeri ise yine aşırı örneklenmiş ATC veri kümesinde BoW nitelik seçimi ve ön işleme adımı uygulanarak elde edildiği gözlemlenmiştir.



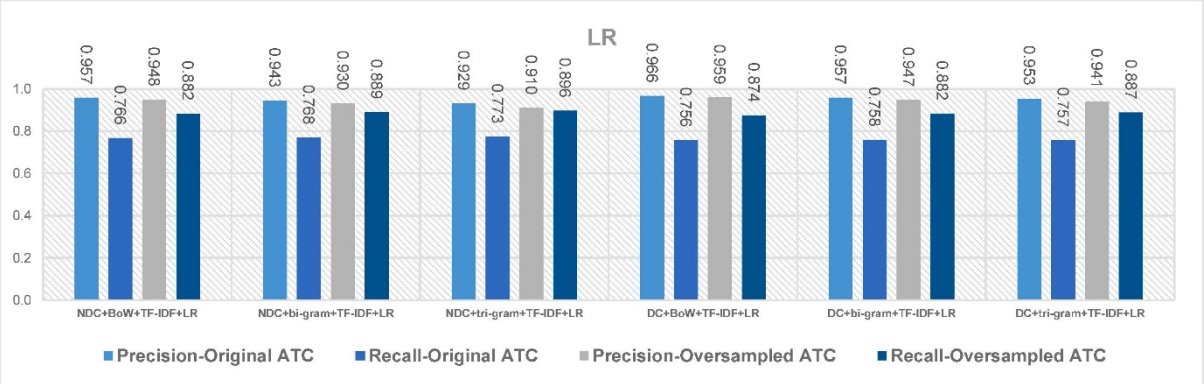
Şekil 4.2. Orijinal ATC ve aşırı örneklenmiş ATC üzerinde 10 kat çapraz doğrulama kullanan NB model kombinasyonlarının ortalama Kesinlik ve Hassasiyet sonuçları.

Şekil 4.2'de NB kombinasyonları için en yüksek Hassasiyet değeri aşırı örneklenmiş ATC veri kümesi ile trigram nitelik seçiminde ve ön işleme yapılmadan/yapılarak elde edilmiştir. En yüksek Kesinlik değeri ise orijinal ATC veri kümesinde trigram nitelik seçimi ve ön işleme adımı uygulanmadan elde edilmiştir.



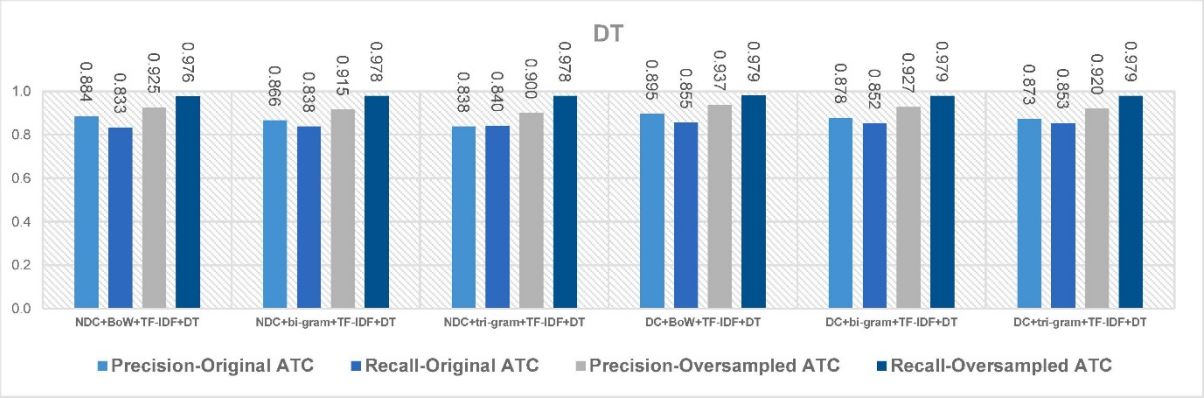
Şekil 4.3. Orijinal ATC ve aşırı örneklenmiş ATC üzerinde 10 kat çapraz doğrulama kullanan RF model kombinasyonlarının ortalama Kesinlik ve Hassasiyet sonuçları.

Şekil 4.3'te RF kombinasyonları için en yüksek Hassasiyet değeri aşırı örneklenmiş ATC veri kümesi ile BoW nitelik seçiminde ve ön işleme yapılarak elde edilmiştir. En yüksek Kesinlik değeri ise aşırı örneklenmiş ATC veri kümesinde trigram nitelik seçimi ve ön işleme adımı uygulanarak elde edildiği görülmüştür.



Şekil 4.4. Orijinal ATC ve aşırı örneklenmiş ATC üzerinde 10 kat çapraz doğrulama kullanan LR model kombinasyonlarının ortalama Kesinlik ve Hassasiyet sonuçları.

Şekil 4.4'te LR kombinasyonları için en yüksek Hassasiyet değeri, aşırı örneklenmiş ATC veri kümesi ile trigram nitelik seçiminde ve ön işleme yapılmadan elde edilmiştir. En yüksek Kesinlik değeri ise orijinal ATC veri kümesinde BoW nitelik seçimi ve ön işleme adımı uygulanarak elde edildiği gözlemlenmiştir.



Şekil 4.5. Orijinal ATC ve aşırı örneklenmiş ATC üzerinde 10 kat çapraz doğrulama kullanan DT model kombinasyonlarının ortalama Kesinlik ve Hassasiyet sonuçları.

Şekil 4.5'te DT kombinasyonları için en yüksek Hassasiyet değeri aşırı örneklenmiş ATC veri kümesi ile BoW/trigram nitelik seçiminde ve ön işleme yapılarak elde edilmiştir. En yüksek Kesinlik değeri ise orijinal ATC veri kümesinde BoW nitelik seçimi ve ön işleme adımı uygulanarak elde edildiği gözlemlenmiştir.

Eklenen düzgünleştirme parametresi (Laplace/Lidstone) olarak kullanılan Alpha değeri (düzleştirme yok için 0), NB sınıflandırıcısı için 0.1 olarak seçilmiştir. SVM metodunda düzenlileştirme için C parametresi kullanılmıştır. Bu metotta kullanılan LinearSVC() fonksiyonu metin sınıflandırmasında yararlı bir tekniktir ve TF-IDF ile kullanıldığında sınıflandırma performansını geliştirmektedir (Zin vd., 2018).

DT metodunda kullanılan rasgele durum değeri, DT'nin farklılaştığı durumlar için en iyi planı oluşturur ve dener. Rasgele durum değeri 25 olarak seçilmiştir. RF sınıflandırıcısında elli ağaç yapısı kullanılmıştır. Bu değer, RF sınıflandırıcı tarafından üretilen DT'lerin sayısıdır. TML tabanlı yöntemlerin parametrelerinin tanımlanmasında genel olarak optimal değerler, problemin fiziksel karmaşıklığına dayalı olarak ampirik olarak seçilmektedir.

4.1.2. Topluluk Modellerinin Değerlendirme Sonuçları

Boosting işleminin verilerin dengesiz olduğu durumlar da dâhil olmak üzere birçok durumda sınıflandırıcıların performansını iyileştirdiği gözlenmiştir. Yeniden ağırlıklandırma ile güçlendirme yapılırken, her bir örneğin sayısal ağırlıkları doğrudan temel öğreniciye iletilir. Temel öğrenici, hipotezini oluştururken ağırlıklandırma bilgilerini kullanır (Seiffert vd., 2008). Küfür söylemleri analizi için topluluk modeli sınıflandırıcıları olarak AdaBoost ve XGBoost uygulanmıştır. DT metodu, artırma işlemi için temel öğrenici olarak kullanılmıştır. AdaBoost ve XGBoost yöntemleri zaten bir artırma işlemine (yeniden ağırlıklandırma) sahip olduğundan, başka bir artırma işlemiyle birleştirilmemiştir. Bu nedenle AdaBoost ve XGBoost yöntemleri yalnızca orijinal ATC veri kümesine uygulanmıştır.

Tablo 4.7, topluluk modeli sınıflandırıcılarında kullanılan hiperparametreleri göstermektedir. Tablo 4.8, AdaBoost ve XGBoost yöntemlerinin performans sonuçlarını açıklamaktadır. Şekil 4.6 ve Şekil 4.7, topluluk modeli sınıflandırıcılarında Kesinlik ve Hassasiyet sonuçlarını göstermektedir.
Tablo 4.8'de Türkçe küfürlü söylemler analizi için kullanılan topluluk sınıflandırıcılarından AdaBoost ve XGBoost sınıflandırıcı kombinasyonlarının sonuçları incelendiğinde XGBoost metodunun en iyi sınıflandırma sonucuna verileri ön işleme aşamasından geçirmeden ve BoW nitelik seçimiyle ulaştığı gözlemlenmiştir.

Şekil 4.6'da AdaBoost kombinasyonları için en yüksek Hassasiyet değeri orijinal ATC veri kümesi ile BoW nitelik seçiminde ve ön işleme yapılarak elde edilmiştir. En yüksek Kesinlik değeri ise orijinal ATC veri kümesinde BoW nitelik seçimi ve ön işleme adımı uygulanarak elde edildiği görülmüştür.

Şekil 4.7'de görüldüğü üzere XGBoost kombinasyonları için en yüksek Hassasiyet değeri orijinal ATC veri kümesi ile BoW nitelik seçiminde ve ön işleme yapılmadan elde edilmiştir. En yüksek Kesinlik değeri ise orijinal ATC veri kümesinde BoW nitelik seçimi ve ön işleme adımı uygulanarak elde edilmiştir.

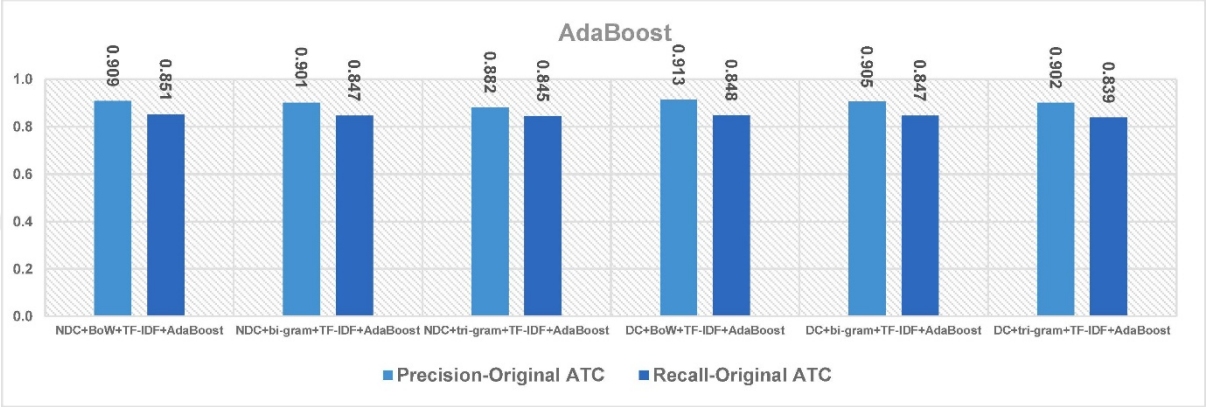
Tablo 4.7. Topluluk modeli sınıflandırıcılarında kullanılan parametreler.

Sınıflandırıcı	Parametre	Değer	Tanım
AdaBoost	Algoritma	SAMME.R	SAMME gerçek artırma algoritması
	Temel tahmin edici (Base_estimator)	DT Sınıflandırıcı	Yükseltilmiş topluluğun inşa edildiği temel tahmin edici.
	n_tahminciler (n_estimators)	3000	Yükseltmenin sonlandırıldığı maksimum tahmin edici sayısı.
	Öğrenme oranı (Learning_rate)	1	Öğrenme oranı, her sınıflandırıcının katkısını azaltır.
XGBoost	n_tahminciler (n_estimators)	3000	Yükseltme için tur sayısı (tahmin ediciler)
	Max_dept	10	Bir ağacın maksimum derinliği
	Öğrenme oranı	0.1	Modele eklendiğinde yeni ağaçların düzeltmeleri için küçülme faktörü
	Alt örnek (Subsample)	1	Fazla takmayı önlemek için eğitim örneğinin alt örnek oranı
	Min_child_weight	1	İhtiyaç duyulan minimum örnek ağırlığı toplamı
	Colsample_bytree	1	Her bir ağacı oluştururken rastgele sütunların (özelliklerin) alt örnek oranı

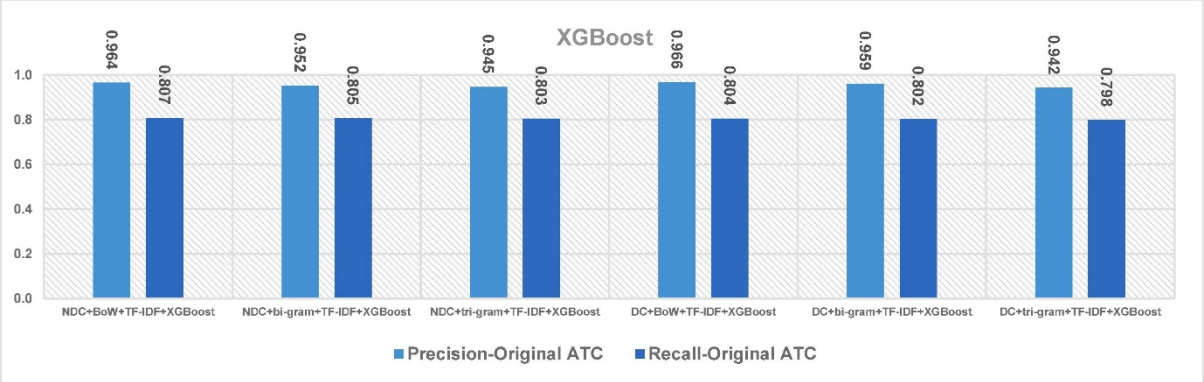
Tablo 4.8. AdaBoost ve XGBoost modellerinin değerlendirme sonuçları (10 kat çapraz doğrulama.

Veri kümesi	Değerlendirmenin Model Kombinasyonu	Mikro F1	Makro F1	κ	Ort. Süreler (sn.)	
					Eğitim	Test
Orijinal ATC	NDC + BoW + TF-IDF + AB	0.919	0.909	0.818	202.654	2.690
	NDC + bigram + TF-IDF + AB	0.914	0.904	0.809	533.504	4.123
	NDC + trigram + TF-IDF + AB	0.907	0.896	0.798	930.810	5.816
	DC + BoW + TF-IDF + AB	0.919	0.909	0.819	**176.067**	2.567
	DC + bigram + TF-IDF + AB	0.916	0.906	0.812	471.727	3.728
	DC + trigram + TF-IDF + AB	0.912	0.902	0.804	794.009	4.818
	NDC + BoW + TF-IDF + XB	**0.923**	**0.911**	**0.823**	282.081	0.429
	NDC + bigram + TF-IDF + XB	0.918	0.906	0.813	909.373	0.478
	NDC + trigram + TF-IDF + XB	0.917	0.903	0.807	1605.153	0.600

Veri kümesi	Değerlendirmenin Model Kombinasyonu	Mikro F1	Makro F1	κ	Ort. Süreler (sn.)	
					Eğitim	Test
	DC + BoW + TF-IDF + XB	0.921	0.910	0.821	265.888	**0.357**
	DC + bigram + TF-IDF + XB	0.920	0.907	0.816	807.654	0.433
	DC + trigram + TF-IDF + XB	0.922	0.906	0.811	1340.781	0.558



Şekil 4.6. Orijinal ATC üzerinde 10 kat çapraz doğrulama kullanan AdaBoost model kombinasyonlarının ortalama Kesinlik ve Hassasiyet sonuçları.



Şekil 4.7. Orijinal ATC üzerinde 10 kat çapraz doğrulama kullanan XGBoost model kombinasyonlarının ortalama Kesinlik ve Hassasiyet sonuçları.

4.1.3. DL (Deep Learning - Derin Öğrenme) Modellerinin Değerlendirme Sonuçları

CNN'ler, grid benzeri bir yapıya sahip verileri işlemek için bilinirliği yüksek bir YSA (Yapay Sinir Ağları) türüdür (Abroyan, 2017). CNN modelleri iki ana katmandan oluşur: konvolüsyon ve tam bağlantılı katman. Konvolüsyon katmanı, konvolüsyon ve havuzlama işlemlerini içermektedir. Konvolüsyon işlemi, seçilen çekirdeğe bağlı olarak metnin özelliklerinin çıkarılmasını temsil

etmektedir. Havuzlama işlemi, elde edilen öznitelik haritasının havuzlama boyutuna ve maksimum havuzlama gibi seçilen yöntemlere göre boyutlarının küçültülmesi işlemidir (Ornek vd., 2019). Çıktı değerlerini ayarlamak için aktivasyon fonksiyonları kullanılmıştır.

Tablo 4.9. CNN modelinin hiperparametreleri ve açıklamaları.

Parametre	İfade/Değer	Tanım
Konvolüsyonel katman (Convolutional layer)	3 tane 1D Konvolüsyonel	Girilen verilerden öğrenme özellikleri
Kernel boyutu	2,3,4	1D Convolution penceresinin uzunluğunu belirtme
Padding	Same	Girişin soluna, sağına veya yukarı, aşağıya eşit şekilde piksel ekleme
Aktivasyon	ReLu ve Sigmoid	Çıktı değerlerinin ayarlanması
Filtreler	100,50,50	Çıktı özellik haritalarının çıkarılması
Havuzlama	GlobalMaxPooling1D	Zaman boyutu üzerinden maksimum değeri alarak girdi gösterimini alt örnekleme
Yoğun katman	İlk yoğun katman 512 nörona, ikinci yoğun katman ise 1 nörona sahiptir.	Çıktıyı döndürme
Optimizasyon	Adam	Adam algoritmasını uygulayan optimize edici
Kayıp fonksiyonu	binary_crossentropy	Gradyanları hesaplama
Eğitim tur sayısı (Epoch)	BPE modeli için 30, diğerleri için 15	Veri kümesinin CNN üzerinden ileri ve geri dönem sayısı
Küme boyutu (Batch size)	128	Dâhili model parametrelerini güncellemeden önce üzerinde çalışılması gereken numune sayısı

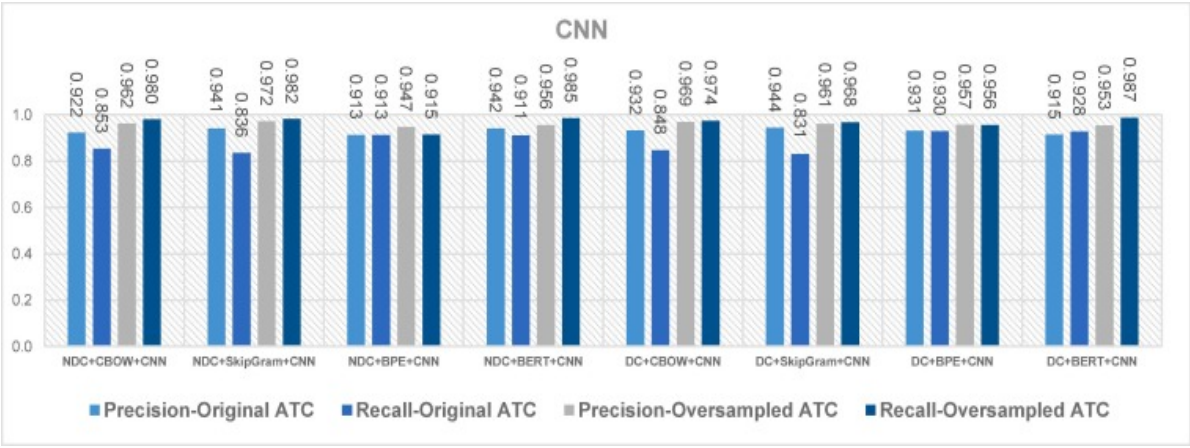
Küfürlü mesaj tespiti için DL tabanlı bir sınıflandırıcı olan CNN modeli kullanılmıştır. CNN modeli, farklı kelime yerleştirme yöntemlerini (CBoW, SkipGram, BPE ve BERT) ve ön işleme adımlarını (DC/NDC) birleştirerek orijinal ve aşırı örneklenmiş ATC veri kümesine uygulanmıştır.

Tablo 4.9, CNN sınıflandırıcısında kullanılan parametreleri göstermektedir. Tablo 4.10, CNN modellerinin 10 kat çapraz doğrulama sonuçları (Makro ve Mikro F1 puanları ve κ değerleri) ve eğitim - test sürelerinin ortalamasını vermektedir. Şekil 4.8, CNN model kombinasyonlarındaki Kesinlik ve Hassasiyet sonuçlarını sunmaktadır.

Tablo 4.10'da DL model kombinasyonlarında en iyi sınıflandırma sonucunun örneklenmiş ATC veri kümesi ile DC + CBoW + CNN kombinasyonu olduğu görülmektedir. Ayrıca NDC + BERT + CNN kombinasyonu en iyi sınıflandırma sonucuna yakın değerler vermiştir. Aşırı örnekleme işleminin DL kombinasyon sonuçlarına olumlu etkisi olmuştur denilebilir. BERT kelime yerleştirme modelinin Word2Vec kelime yerleştirme modellerinden daha düşük sınıflandırma süresine sahip olması kullanılan kelime yerleştirme derlemleriyle ile ilgilidir.

Tablo 4.10. CNN modelinin değerlendirme sonuçları (10 kat çapraz doğrulama).

Veri kümesi	Değerlendirmenin CNN Model Kombinasyonu	Mikro F1	Makro F1	κ	Ort. Süreler (sn.)	
					Eğitim	Test
Orijinal ATC	NDC + CBoW	0.924	0.914	0.829	1552.744	0.769
	NDC + Skip-Gram	0.925	0.914	0.829	1573.601	0.788
	NDC + BPE	0.913	0.906	0.803	2379.531	2.026
	NDC + BERT	**0.947**	**0.942**	**0.883**	280.012	0.039
	DC + CBoW	0.918	0.930	0.834	1482.808	0.748
	DC + Skip-Gram	0.923	0.913	0.827	1462.641	0.579
	DC + BPE	0.931	0.924	0.847	2214.171	2.038
	DC + BERT	0.942	0.939	0.878	**228.979**	**0.008**
Aşırı örneklenmiş ATC	NDC + CBoW	0.971	0.971	0.941	2001.791	0.961
	NDC + Skip-Gram	0.970	0.969	0.938	2051.781	1.009
	NDC + BPE	0.945	0.945	0.893	2771.635	2.586
	NDC + BERT	0.973	**0.974**	0.940	411.745	**0.009**
	DC + CBoW	**0.974**	0.973	**0.946**	1996.982	1.016
	DC + Skip-Gram	0.967	0.968	0.941	2191.014	64.905
	DC + BPE	0.956	0.956	0.913	2805.852	2.468
	DC + BERT	0.970	0.970	0.939	**406.056**	**0.009**



Şekil 4.8. Orijinal ATC ve aşırı örneklenmiş ATC üzerinde 10 kat çapraz doğrulama kullanan CNN model kombinasyonlarının ortalama Kesinlik ve Hassasiyet sonuçları.

Keras (Chollet, 2018) ile içe aktarılan CNN modeli, üç konvolüsyonel, bir global maksimum havuzlama ve iki yoğun katman içermektedir. Konvolüsyon katmanları sırasıyla 100, 50, 50 filtrelerinden oluşur (aktivasyon fonksiyonu = ReLU). Birinci yoğun katman 512 nörondan (aktivasyon fonksiyonu = ReLu) oluşmakta ve çıktı katmanı olarak adlandırılan ikinci katman bir nöron içermektedir (aktivasyon fonksiyonu = Sigmoid). Optimize edici olarak Ayarlanabilir Öğrenme Oranı Tahmini (Adaptive Moment Estimation - Adam) kullanılmıştır. Stokastik amaç fonksiyonlarının birinci mertebeden gradyan tabanlı optimizasyonu için kullanılan bir algoritmadır ve uyarlanabilir tahminlere dayanmaktadır (Huang vd., 2019). Kayıp fonksiyonu olarak binary_crossentropy benimsenmiş ve küme boyutu 128 ve eğitim tur sayısı 15 olarak belirlenmiştir.

CBoW ve Skip-Gram kelime yerleştirmelerini oluşturmak için Türkçe bir derlem kullanılmıştır. Yerleştirme vektör boyutu 400 ve pencere boyutu maksimum uzunluğu 5 olarak seçilmiştir. Türkçe alt kelime yerleştirme için önceden eğitilmiş bir kelime yerleştirme modeli (Heinzerling & Strube, 2017) benimsenmiştir. Önceden eğitilmiş dağıtık modelin kelime boyutu 25.000, kelime yerleştirme boyutu 50 olarak seçilmiştir. CNN modeline BERT kelime yerleştirme uygulamak için BERT tabanlı bir Türkçe Duyarlılık Modeli kullanılmıştır. Model diğer Türkçe NLP çalışmalarında da kullanılmıştır. Kelime yerleştirme boyutu 200, maksimum dizi uzunluk değeri 128 olarak seçilmiştir.

4.1.4. Küfürlü Söylemler Analizi Performans Karşılaştırması

İki veri seti (orijinal ATC ve aşırı örneklenmiş ATC), CNN modeline, geleneksel ML tabanlı modellere (NB, SVM, DT, RF ve LR) ve topluluk modellerine (AdaBoost ve XGBoost) uygulanmıştır. Ön işleme adımı DC/ NDC olarak gerçekleştirilmiş ve elde edilen sonuçlar karşılaştırılmıştır. Sınıflandırıcıların Makro F1, Mikro F1, Precison ve Hassasiyet değerlendirme sonuçları önceki bölümlerde sunulmuştur.

Sınıflandırma performans sonuç tablolarının son iki sütunu, 10 kat çapraz doğrulama için her kategorizasyon modelinin ortalama eğitim ve test sürelerini göstermeye ayrılmıştır. Her kat için eğitim ve test süreleri Kaggle sayfasına yüklenen ayrıntılı deney sonuçları dosyası olarak eklenmiştir. Kategorizasyon modellerinin performans karşılaştırması aşağıda noktadan noktaya verilmiştir:

- Tüm öznitelik seçimi, kelime yerleştirme ve sınıflandırma metotları için aşırı örnekleme yönteminin tüm performans ölçümlerinde daha iyi sonuçlar verdiği şekil ve tablolardan açıkça görülmektedir. Aşırı örnekleme yönteminin sonuçlar üzerinde pozitif bir etkisi olmuştur.
- Değerlendirmemizde kullanılan tüm metotlar için tatmin edici sonuçlar elde edilmiştir. Bununla birlikte, en iyi sınıflandırma performansı (Mikro F1 değeri: 0.974, Makro F1 değeri: 0.973, κ: 0.946), aşırı örneklenmiş ATC'deki CNN modelinden elde edilmiştir. CNN modelinin ve aşırı örneklenmiş ATC üzerindeki SVM modelinin diğer

kombinasyonları çok benzer sonuçlar vermiştir. CNN modeli, küfürlü ve küfürsüz sınıfları ayıran en iyi sınıflandırıcıdır ve yanlış sınıflandırma hatalarını en aza indirmektedir. CNN sınıflandırıcının yüksek Kesinlik ve Hassasiyet değerleri, ATC veri setlerinde pozitif (küfürlü) ve negatif (küfürsüz) yorumları doğru bir oranda ayırabildiğini de kanıtlamaktadır.

- SVM sınıflandırıcı, CNN'den sonra en iyi ikinci performans sonuçlarına sahiptir. Aşırı örnekleme yönteminin etkisi, CNN modelinde Kesinlik açısından daha az etkilidir. SVM modeli, modellerin eğitilmesi için harcanan zaman dikkate alındığında çok hızlı ve doğru sonuçlar vermiştir.
- RF sınıflandırıcı, SVM'den sonra en iyi üçüncü performans sonuçlarına sahiptir. Bazı kombinasyonlarda SVM modeliyle hemen hemen aynı sonuçları üretmiştir. Aşırı örneklemenin etkisi, RF modelinde diğerlerinden daha belirgin hale gelmiştir.
- DT ve NB modelleri benzer performans gösterirken, performans ölçütlerinin çoğunda DT modelinin NB modelinden daha iyi olduğu söylenebilmektedir. LR metodu ve topluluk modelleri (AdaBoost ve XGBoost) diğer metotlardan daha kötü performans göstermiştir.
- "DC" ve "NDC" adımlarının birbirlerine kesin bir üstünlüğü olmadığı gözlemlenmiştir. Bununla birlikte, BoW vektör gösterimi çoğunlukla iyi performans sonuçları üretmiştir.
- Veri setlerinin doğruluğunu, hassasiyetini, sağlamlığını ölçmek için κ değeri kullanılabilir (Talukder & Ahammed, 2020). Elde edilen κ değerlerine göre veri setlerinin ayırt edici kesinliğinin oldukça gürbüz olduğu görülmektedir.

4.2. Homofobik ve Nefret Söylemleri Analizi

Homofobik ve nefret söylemleri analizinde önerilen sınıflandırıcılar, HATC ve resHATC veri kümeleri üzerinde test edilmiştir. Tüm eğitim ve test rutinleri Google'ın Colaboratory hizmetinde gerçekleştirilmiştir.

Bu çalışmada, TML ve topluluk sınıflandırma modellerinin parametre seçimi 10 kat çapraz doğrulama kullanılarak ızgara arama tekniği ile yapılmıştır. SVM sınıflandırıcısının en iyi versiyonunu elde etmek için kullanılan ızgara arama tekniği literatürde iyi bilinen bir yaklaşımdır. SVM'nin parametre değerleri aşağıdaki gibi tanımlanmıştır (Ornek vd., 2019; Wu & Dredze, 2020). C (maliyet parametresi) = (0.01, 0.1, 1, 10, 10.01, 10.1, 100, 100.01, 100.1) ve kernel = (rbf, linearSVC) değerleri test edilmiştir. Izgara arama sürecinin arkasındaki ölçeklendirme motivasyonu, küçük C değeri kenar boşluklarından büyük C değeri marjlarına kadar C parametrelerinin kapsamlı bir değerlendirmesini

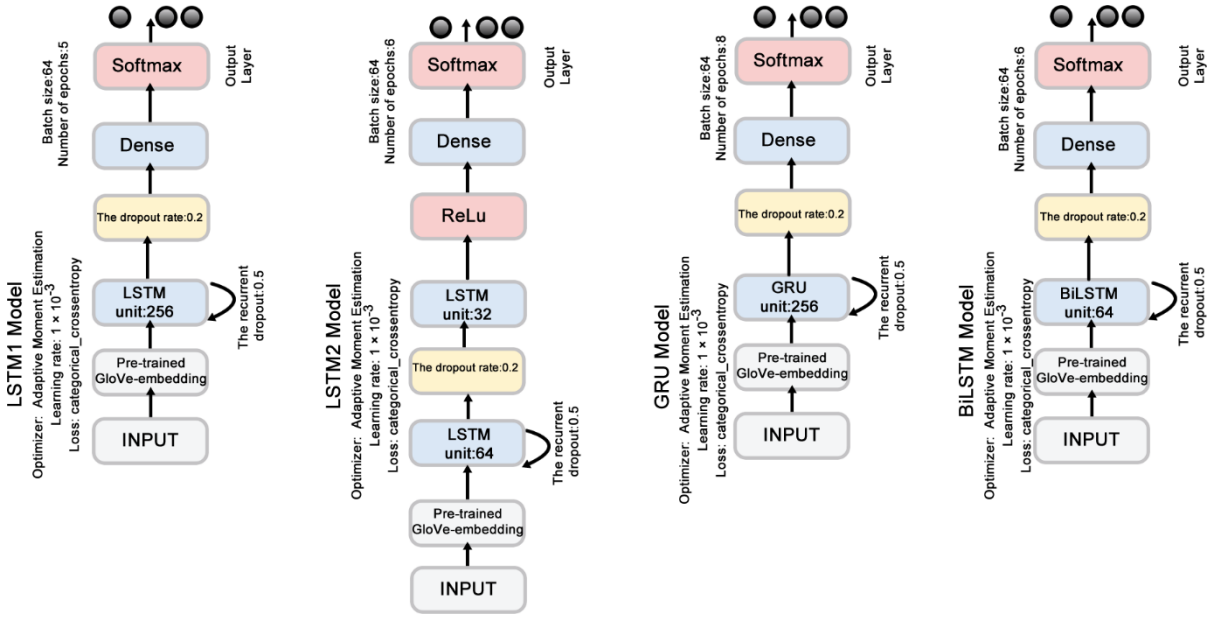
yapmaktır. SVM modeli en iyi sonuçları C = 10.01 ve kernel = linearSVC değerleri ile vermiştir. NB metodunda çok sınıflı kategoriler için kullanılan multinominal NB seçilmiş ve Alpha değeri 0.1 olarak belirlenmiştir. RF metodu için n_estimators değeri 50 olarak belirlenmiştir. Kullanılan tüm TML yöntemleri için hiperparametre değerleri Tablo 4.11'de verilmiştir.

Tablo 4.11. TML sınıflandırıcılarının hiperparametreleri.

Sınıflandırıcı	Parametre ve İfade/Değer
SVM	C = 10.01, kernel = linearSVC
NB	Alpha = 0.1, MultinomialNB
RF	Karar ağacı sayısı (n_estimatators) = 50

Topluluk sınıflandırıcıları için test edilen parametre değerleri: ızgara arama algoritması ile n_estimators = (10, 20, 50, 100, 500,1000, 2000, 3000) değerleri test edilerek optimal tahmin edici sayısı 3000 olarak seçilmiş ve tüm topluluk modellerine uygulanmıştır.

Bu çalışmada, DL tabanlı yöntemler için en uygun parametre değerleri deneme - yanılma (trial - error) yöntemi ile belirlenmiştir. Veri setindeki kelimeleri temsil etmek için 300 boyutlu GloVe vektörleri kullanılmıştır. Tüm DL tabanlı sınıflandırıcılar için ağ mimarilerinin detayları ve hiperparametreler Şekil 4.9'da verilmiştir. DL modellerinde katman sayısı artırılarak denenmiş ve sadece LSTM modelinin iki katmanlı olması durumunda sınıflandırma başarısı azalmamıştır. Bu nedenle, LSTM modelinden bir LSTM katmanlı (LSTM1) ve iki LSTM katmanlı (LSTM2) olmak üzere iki model oluşturulmuştur.



Şekil 4.9. DL tabanlı sınıflandırıcıların hiperparametreleri.

Seyreltme (dropout) değeri katmanlar arasındaki girişleri rastgele kaldırır. Tekrarlayan seyreltme (recurrent dropout) zaman adımları arasındaki girişleri ortadan kaldırır. Dropout ve recurrent dropout model yapısı içerisinde düzenlileştirici bir etkiye sahiptir ve aşırı öğrenmeyi (overfitting) önler. DL tabanlı tüm modeller için farklı dropout değerleri (0.2, 0.3, 0.4 ve 0.5) denenmiş ve optimum dropout değeri 0.2 olarak bulunmuştur. Aynı şekilde optimum recurrent dropout değeri 0.5 olarak kullanılmıştır. Adam optimize edici DL tabanlı modellerde kullanılmıştır. Öğrenme oranı 1×10^{-3} ve kayıp fonksiyonu categorical_crossentropy'dir. Eğitim sırasında küme boyutu 64'tür, eğitim tur sayısı sırasıyla LSTM1 = 5, LSTM2 = 6, GRU = 8 ve BiLSTM = 6'dır.

Homofobik ve nefret söylemleri analizi çalışmasında, Türkçe dâhil 104 dili destekleyen, 12 yığılmış transformatör bloğu, 768 gizli birim, 12 attention head ve toplam 110.000.000 parametreyi destekleyen önceden eğitilmiş bir M-BERT modeli kullanılmıştır. Kullanılan M-BERT modeli, çeşitli dillerden gelen verileri inceleyerek farklı metin dillerinin biçimini dikkate alma yeteneğine sahiptir (Smetanin & Komarov, 2021).

Kullanılan BERT modelde, son katman olarak ReLU aktivasyon fonksiyonlu iki yoğun katman, iki dropout katmanı (0.2) ve Softmax aktivasyon fonksiyonlu yoğun katman bulunmaktadır. BERT model Adam optimizer kullanılarak optimize edilmiş ve küme boyutu 32, eğitim tur sayısı 3 ve öğrenme oranı 1×10^{-5} olan BERT modelinin bir kombinasyonu üzerinde eğitilmiştir.

Tablo 4.12, homofobik ifadeleri tespit etmek için farklı kombinasyonlara sahip sınıflandırma modellerinin performans metriklerini göstermektedir. Tablo 4.12'de, her iki veri setinde de en iyi F1 score değeri M-BERT modelinden elde edilmektedir. Bunun en önemli nedeni, dönüştürücü yapısının ve dikkat mekanizmasının duygu bilgisini daha iyi ve doğru bir şekilde yakalayabilmesidir. Önceden eğitilmiş farklı dillerde büyük veri ve kelime çeşitliliği kullanan M-BERT modeli, tüm yaklaşımlardan daha iyi performans göstermiştir.

BiLSTM resHATC veri setinde en iyi ikinci modeldir. LSTM1 ve LSTM2 modelleri gradyan kaybolma sorunlarını hafifletmesine rağmen, BiLSTM modeli bağlamın anlamsal bilgisini LSTM modellerinden daha etkili bir şekilde yakalayabilmiştir. BiLSTM modeli, ileri - geri zaman yönleri arasındaki çift yönlü uzun vadeli bağımlılıkların öğrenilmesine yardımcı olmuş ve LSTM modellerinden ve GRU modelinden daha iyi nitelikler çıkarmıştır.

SVM modeli, HATC veri setindeki LSTM1, LSTM2 ve GRU modelleriyle çok yakın F1 skor performansı göstermiştir. AdaBoost, XGBoost ve Gradient Boosting modelleri, HATC veri setinde resHATC veri setinden daha iyi F1 skor sonuçları vermiştir. Yeniden örnekleme yönteminin F1 skor değeri açısından TML ve topluluk sınıflandırıcıları üzerinde hiçbir etkisi olmamıştır.

Nefret söylemi kategorisine ilişkin sınıflandırma modellerinin performans sonuçları Tablo 4.13'te verilmiştir. Tablo 4.13'e göre her iki veri setinde de nefret söylemlerinin sınıflandırılması için en iyi model M-BERT modelidir. LSTM1 ve LSTM2 modelleri, HATC veri setinde BiLSTM modeline yakın ikinci en iyi F1 skor değerleri üretmiştir.

Tablo 4.12. Homofobik kategori için sınıflandırma modellerinin performans karşılaştırması.

Model	Homofobik kategori performansı (%)		
	Kesinlik	**Hassasiyet**	**F1 skor**
HATC + unigram + TF-IDF	81.51	61.32	69.99
HATC + unigram + TF-IDF + NB	**96.52**	33.40	49.63
HATC + unigram + TF-IDF + RF	85.31	49.30	62.49
HATC + unigram + TF-IDF + AdaBoost	59.32	47.03	52.46
HATC + unigram + TF-IDF + XGBoost	81.93	53.94	65.05
HATC + unigram + TF-IDF + Gradient Boosting	76.34	61.31	68.00
HATC + GloVe + LSTM1	78.61	61.40	68.95
HATC + GloVe + LSTM2	74.52	62.31	67.87
HATC + GloVe + GRU	72.93	66.72	69.69
HATC + GloVe + BiLSTM	75.62	67.52	71.34
HATC + M-BERT	90.81	**76.29**	**82.64**
resHATC + unigram + TF-IDF	62.31	66.01	64.11
resHATC + unigram + TF-IDF + NB	36.52	63.52	46.38
resHATC + unigram + TF-IDF + RF	58.71	58.20	58.45
resHATC + unigram + TF-IDF + AdaBoost	45.22	54.14	49.28
resHATC + unigram + TF-IDF + XGBoost	50.73	67.83	58.05
resHATC + unigram + TF-IDF + Gradient Boosting	56.22	67.52	61.35
resHATC + GloVe + LSTM1	69.21	72.91	71.01
resHATC + GloVe + LSTM2	69.21	68.51	68.86
resHATC + GloVe + GRU	55.23	76.51	64.15
resHATC + GloVe + BiLSTM	**78.71**	69.50	73.82
resHATC + M-BERT	77.00	**86.37**	**80.88**

Tablo 4.13. Nefret kategorisi için sınıflandırma modellerinin performans karşılaştırması.

Model	Nefret kategori performansı (%)		
	Kesinlik	**Hassasiyet**	**F1 skor**
HATC + unigram + TF-IDF	90.8	84.12	87.33
HATC + unigram + TF-IDF + NB	85.4	86.23	85.81
HATC + unigram + TF-IDF + RF	92.12	76.18	83.40
HATC + unigram + TF-IDF + AdaBoost	86.31	74.61	80.03

Model	Nefret kategori performansı (%)		
	Kesinlik	Hassasiyet	F1 skor
HATC + unigram + TF-IDF + XGBoost	94.9	80.22	86.94
HATC + unigram + TF-IDF + Gradient Boosting	**95.31**	79.21	86.52
HATC + GloVe + LSTM1	90.61	87.32	88.92
HATC + GloVe + LSTM2	91.71	85.61	88.56
HATC + GloVe + GRU	87.5	88.82	88.16
HATC + GloVe + BiLSTM	89.01	88.84	88.92
HATC + M-BERT	94.02	**89.65**	**91.75**
resHATC + unigram + TF-IDF	84.01	84.11	84.06
resHATC + unigram + TF-IDF + NB	76.51	85.32	80.68
resHATC + unigram + TF-IDF + RF	86.42	79.71	82.93
resHATC + unigram + TF-IDF + AdaBoost	82.62	75.51	78.91
resHATC + unigram + TF-IDF + XGBoost	90.22	78.91	84.19
resHATC + unigram + TF-IDF + Gradient Boosting	**90.81**	79.1	84.55
resHATC + GloVe + LSTM1	87.81	87.12	87.46
resHATC + GloVe + LSTM2	84.82	87.11	85.95
resHATC + GloVe + GRU	82.82	87.11	84.91
resHATC + GloVe + BiLSTM	89.35	88.5	88.92
resHATC + M-BERT	88.97	**89.86**	**89.06**

Tablo 4.14, nötr kategori için sınıflandırma modellerinin performans ölçütlerini göstermektedir. Modellerin nötr kategoriyi belirlemede diğer kategorilerin (homofobik ve nefret dolu) tespitine göre daha başarılı F1 skor değerleri ürettiği gözlemlenmiştir. En iyi model, diğer kategorilerdeki sınıflandırma sonuçları gibi HATC veri setinde M-BERT modeli olmuştur. BiLSTM modeli, resHATC veri setinde en iyi ikinci F1 skor sınıflandırma puanını üretmiştir.

Tablo 4.14. Nötr kategori için sınıflandırma modellerinin performans karşılaştırması.

Model	Nötr kategori performansı (%)		
	Kesinlik	Hassasiyet	F1 skor
HATC + unigram + TF-IDF	91.13	95.81	93.41
HATC + unigram + TF-IDF + NB	91.21	95.22	93.17
HATC + unigram + TF-IDF + RF	89.61	96.86	93.09
HATC + unigram + TF-IDF + AdaBoost	87.42	94.68	90.91
HATC + unigram + TF-IDF + XGBoost	88.63	**98.01**	93.08

Model	Nötr kategori performansı (%)		
	Kesinlik	Hassasiyet	F1 skor
HATC + unigram + TF-IDF + Gradient Boosting	85.32	97.68	91.08
HATC + GloVe + LSTM1	90.81	95.08	92.90
HATC + GloVe + LSTM2	92.22	95.59	93.87
HATC + GloVe + GRU	93.02	92.62	92.82
HATC + GloVe + BiLSTM	93.51	94.61	94.06
HATC + M-BERT	**94.56**	97.67	**96.08**
resHATC + unigram + TF-IDF	91.71	91.59	91.65
resHATC + unigram + TF-IDF + NB	93.42	83.89	88.40
resHATC + unigram + TF-IDF + RF	89.44	93.21	91.29
resHATC + unigram + TF-IDF + AdaBoost	88.23	91.42	89.80
resHATC + unigram + TF-IDF + XGBoost	89.52	93.40	91.42
resHATC + unigram + TF-IDF + Gradient Boosting	89.61	94.41	91.95
resHATC + GloVe + LSTM1	93.72	93.59	93.65
resHATC + GloVe + LSTM2	93.71	92.62	93.16
resHATC + GloVe + GRU	94.72	89.61	92.09
resHATC + GloVe + BiLSTM	94.72	**94.61**	**94.66**
resHATC + M-BERT	**95.16**	93.17	93.99

Üç kategori (homofobik, nefret dolu ve nötr) için ortalama performans sonuçları Tablo 4.15'te sunulmuştur.

Tablo 4.15. Veri kümelerinde üç kategorinin (homofobik, nefret ve nötr) ortalama performans karşılaştırması.

Model	Ortalama Performans (%)		
	Kesinlik	Hassasiyet	F1 skor
HATC + unigram + TF-IDF	87.81	80.42	83.95
HATC + unigram + TF-IDF + NB	91.04	71.62	80.17
HATC + unigram + TF-IDF + RF	89.01	74.11	80.88
HATC + unigram + TF-IDF + AdaBoost	77.68	72.11	74.79
HATC + unigram + TF-IDF + XGBoost	88.49	77.39	82.57
HATC + unigram + TF-IDF + Gradient Boosting	85.66	79.40	82.41
HATC + GloVe + LSTM1	86.68	81.27	83.89
HATC + GloVe + LSTM2	86.15	81.17	83.59
HATC + GloVe + GRU	84.48	82.72	83.59

Model	Ortalama Performans (%)		
	Kesinlik	**Hassasiyet**	**F1 skor**
HATC + GloVe + BiLSTM	86.05	83.66	84.84
HATC + M-BERT	**93.13**	**87.87**	**90.15**
resHATC + unigram + TF-IDF	79.34	80.57	79.95
resHATC + unigram + TF-IDF + NB	68.82	77.58	72.94
resHATC + unigram + TF-IDF + RF	78.19	77.04	77.61
resHATC + unigram + TF-IDF + AdaBoost	72.02	73.69	72.85
resHATC + unigram + TF-IDF + XGBoost	76.82	80.05	78.40
resHATC + unigram + TF-IDF + Gradient Boosting	78.88	80.34	79.60
resHATC + GloVe + LSTM1	83.58	84.54	84.04
resHATC + GloVe + LSTM2	82.58	82.75	82.66
resHATC + GloVe + GRU	77.59	84.41	80.86
resHATC + GloVe + BiLSTM	**87.59**	84.20	85.86
resHATC + M-BERT	87.05	**89.80**	**87.98**

Sınıflandırma modellerinin genel performans karşılaştırması aşağıda verilmiştir:

- M-BERT modeli, deneylerde kullanılan tüm modeller arasında en iyi sınıflandırma performansına (homofobik kategori F1 skor puanı: %82.64, nefret dolu kategori F1 skor puanı: %91.75 ve nötr kategori F1 skor puanı: %96.08) sahip olduğundan, ortalama F1 skor performansı (%90.15) diğer modellerden daha iyidir.
- M-BERT modeli, derin katmanlarda farklı diller arasındaki dilsel ve evrimsel ilişkileri daha iyi yansıtmak için var olan alanı bölümlere (segment) ayırır. Diller, sözlükler kullanılarak aralarında hizalanır ve diller arası kelime yerleştirme tamamen denetimsiz yöntemlerle işbirliği içinde öğrenilir. Bu model çok dilli kelime gömmeleri ve çeşitli kontrol seviyeleri ile yüksek kaynaklı diller (%70) ve düşük kaynaklı (%30) diller arasında transfer öğrenimi konusunda eğitilmiştir. Bu sıralamada Türkçe dili yüksek kaynaklı dil grubuna girmektedir. Yüksek kaynak dillere sahip diğer dillerde yapılan transfer sınıflandırma başarısının Türkçe dili M-BERT modelindeki sınıflandırma başarısına yakın olduğu kanıtlanmıştır (Wu & Dredze, 2020). Bu nedenle değerlendirmelerimizde kullanılan M-BERT modeli diğer diller için de kullanılabilir ve tavsiye edilir.
- M-BERT modeli, tüm kategorilerde resHATC veri kümesine kıyasla HATC veri kümesinde daha yüksek F1 skor performans değerleri vermiştir. M-BERT modelinin önceden eğitilmiş yeterli Türkçe veriye sahip bir model olduğu için sınıf dengesizliği sorununu dikkate almadığı düşünülmektedir.

- Üç kategorinin F1 skor sonuçlarının ortalama performansı göz önüne alındığında, BiLSTM modeli resHATC veri setinde en iyi ikinci modeldir. Verileri her iki yönde de işleyen BiLSTM modeli, hem önceki hem de ardışık bağlamlardan bir metin parçasının sıralı bağımlılıklarını modelleme yeteneği nedeniyle daha iyi performans göstermiş olabilir. Üçüncü en iyi sınıflandırma modeli, resHATC veri setindeki LSTM1 modelidir.
- BiLSTM modeli, tüm kategorilerde HATC veri kümesine kıyasla resHATC veri kümesinde daha yüksek F1 skor performans değerleri vermiştir. Veri kümesinin dengelenmesi BiLSTM modeli için sınıflandırma başarısı üzerinde olumlu bir etkiye sahiptir.
- HATC veri setindeki SVM modelinin F1 skor sonuçları, DL tabanlı modellerin sonuçlarına yakındır. ResHATC veri kümesindeki SVM modelinin performansı, HATC veri kümesindeki sonuçlardan daha kötü olmuştur. SVM yönteminde kategoriler arasındaki hiperdüzlemler destek vektörlerine göre hesaplandığından, her kategorideki örnek sayısı sınıf sınırını fazla etkilememektedir. Bu nedenle, SVM'nin sınıf dengesizliği sorununa potansiyel olarak daha az duyarlı olduğu bilinmektedir (Japkowicz & Stephen, 2002; Sun vd., 2009). Ancak, SVM metodunun bazı yeniden örnekleme veri kümelerinde iyi sınıflandırma sonuçları verdiği de kanıtlanmıştır (Karayigit vd., 2021; Moraes vd., 2013). Veri kümesini yeniden örnekleme algoritmalarıyla dengelemek, TML ve topluluk sınıflandırıcılarında değişken sınıflandırma sonuçları (daha iyi veya daha kötü) verebilmektedir. Bu çalışmada HATC veri setinin dengelenmesi, TML ve topluluk sınıflandırıcılarının F1 skor performansını düşürmüştür.
- Topluluk sınıflandırıcılar arasında ortalama F1 skor puanına sahip en iyi sınıflandırıcı, HATC veri setinde %82,57 ile XGBoost metodudur.
- Tüm modeller arasında en düşük ortalama F1 skor puanı, HATC veri setinde %80,17 ile NB sınıflandırıcı olmuştur. NB sınıflandırıcısı, resHATC veri setinde de %72.94 ile en düşük sınıflandırma sonucuna sahiptir.
- Adam optimizer, DL tabanlı modellerin eğitimi için stokastik gradyan inişinin yerini alır. LazyAdam ve AdamW yöntemleri de çalışmamızda değerlendirilmiştir. LazyAdam, seyrek güncellemeleri işlemede daha verimli olacak şekilde tasarlanmış Adam'ın yükseltilmiş bir versiyonudur (Sowinski-Mydlarz vd., 2021). AdamW, Adam'ın bir varyasyonudur, burada ağırlık azaltma sadece adım boyutunu parametre bazında kontrol ettikten sonra gerçekleştirilmektedir (Llugsi vd., 2021). Ancak çalışmamızda LazyAdam ve AdamW optimizasyon yöntemlerinin kullanılması sonuçları etkilememiştir. LazyAdam, DL tabanlı modellerde sınıflandırma sonuçlarını artırmamış ancak M-BERT modelinde Adam optimizasyonuna kıyasla sınıflandırma sonuçlarında düşüşe neden olmuştur. Ayrıca Adam yerine AdamW yöntemi kullanıldığında her iki modelin performansında da herhangi bir iyileşme gözlenmemiştir.

4.3. COVID-19 Duygu Analizi

Bu bölümde COVID-19 duygu analizi için önerilen modellerin (TML modelleri, DL modelleri ve BERT tabanlı transfer öğrenme modeli) sınıflandırma sonuçları ve süreleri kıyaslanmıştır. Bu bölümün son bölümünde, mevcut iki veri kümesi (Dataset1 ve Dataset2) arasındaki sosyal medya ilişkisi, LDA ve p değerleri ile karşılaştırılmıştır.

Bu çalışmada değerlendirme işlemleri için Google Colab makinesi kullanılmıştır. Google Colab makinesinin özellikleri Tesla K80 GPU, 12 GB RAM ve Intel Xeon CPU 2.20 GHz'dir. Önerilen modellerin performansını karşılaştırmak için Makro F1 puanı kullanılmıştır.

Tablo 4.16. ızgara arama ile 10 kat çapraz doğrulama ile belirlenen TML yöntemlerinin (SVM. NB ve RF) hiperparametre değerlerini gösterir. COVID-19 duygu analizi çalışmasında SVM'nin hiper parametre değerleri için maliyet parametresi (C) = (0.1. 1. 2. 3. 4. 5. 10. 20. 30. 40. 50) değerleri için araştırılmıştır. SVM modeli Dataset1'de C = 2 değeri ve Dataset2'de C = 1 değeri ile en iyi sonuçları vermiştir. NB metodunda çok sınıflı kategoriler için kullanılan MultinominalNB seçilmiş. NB'nin Alpha değeri = (0.1. 0.2. 0.3. 0.4. 0.5. 0.6) için denenmiş ve Dataset1'de Alpha = 0.1 değeri ve Dataset2'de Alpha = 0.5 değeri ile en iyi sonuçları vermiştir. RF yönteminin n_estimators parametre değeri yani algoritmadaki DT sayısı tüm veri setlerinde (Dataset1, resDataset1, Dataset2 ve resDataset2) 70'tir.

Tablo 4.16. Veri kümeleri için TML modellerinin en iyi parametreleri.

Veri kümesi	Metot	Parametre	İfade/Değer
Dataset1 ve resDataset1	SVM	C	2
	NB	alpha	0.1
	RF	n_estimators	70
Dataset2 ve resDataset2	SVM	C	1
	NB	alpha	0.5
	RF	n_estimators	70

Derin Öğrenme modellerinin (LSTM ve CNN) hiperparametre değerlerini belirlemek için deneme - yanılma yöntemi uygulanmıştır (Dang vd., 2020). Tablo 4.17, CNN sınıflandırıcının Dataset1 ve Dataset2 veri kümelerinde elde ettiği hiperparametre değerlerini göstermektedir.

CNN katmanları sırasıyla 1, 2, 3 olarak denenmiş ve sınıflandırma başarısına göre CNN katmanı 1 olarak seçilmiştir. CNN katmanında filtre değeri sırasıyla 10, 50, 100, 150 olarak denenmiş

ve sınıflandırma başarısına göre filtre değeri 100 olarak belirlenmiştir. Kernel boyutu değeri sırasıyla 3,4,5 olarak denenmiş ve sınıflandırma başarısına göre değer 3 olarak kararlaştırılmıştır. GlobalMaxPooling1D, zaman boyutu üzerinden maksimum değeri alarak giriş temsilinin alt örneklerini alır. GlobalMaxPooling1D, havuzlama katmanında daha verimli olduğu için kullanılmıştır (Kishore Kumar & Sreenivasa Rao, 2022). Yoğun katmanda 3 kategorik veri tipi (pozitif, negatif ve nötr) olduğu için çıktı değeri 3 ve çıktı fonksiyonu Softmax kullanılır.

Tablo 4.17. Dataset1ve Dataset2 için kullanılan CNN modelinin en iyi parametreleri.

Parametre	İfade/Değer
Yerleştirme katmanındaki çıktı boyutu sayısı	512
Konvolüsyon katmanlarının sayısı	1
Filtre	100
Kernel	3
Aktivasyon fonksiyonu	ReLU
Havuzlama	GlobalMaxPooling1D
Yoğun katman aktivasyon işlevi	ReLU
Çıktı aktivasyon fonksiyonu	Softmax
Kayıp fonksiyonu	Categorical cross-entropy
Eğitim tur sayısı	15
Optimize Edici	Adam
Tam bağlı birimlerin değeri	512
Çıktı sayısı	3
Küme boyutu sayısı	64 (Dataset1 and resDataset1)
	32 (Dataset2 and resDataset2)

LSTM ağları, önceki verilerin hatırlanmasını sağlayan RNN modelin değiştirilmiş bir versiyonudur. RNN'de ortaya çıkan gradyan problemleri LSTM (García-Díaz vd., 2020) ile çözülür. LSTM modeli, verileri geri yayılım algoritması kullanarak eğitir. Tablo 4.18, LSTM sınıflandırıcısının Dataset1 ve Dataset2 veri kümelerindeki hiperparametre değerlerini gösterir.

Tablo 4.18. Veri kümeleri için LSTM modellerinin en iyi parametre değerleri

Parametre	İfade/Değer
Yerleştirme katmanındaki boyutu	512
LSTM katmanlarının sayısı	2
Dropout	0.4 (Dataset1 and resDataset1)
	0.5 (Dataset2 and resDataset2)

Parametre	İfade/Değer
Recurrent dropout	0.4 (Dataset1 and resDataset1)
	0.6 (Dataset2 and resDataset2)
Yoğun katman aktivasyon işlevi	Softmax
Kayıp fonksiyonu	categorical_crossentropy
Eğitim tur sayısı	10
Optimizer	Adam (lr = 0.0001)
Küme boyutu sayısı	64

LSTM modelinde farklı dropout değerleri (0.2, 0.3, 0.4 ve 0.5) denenmiş ve Dataset1 ile resDataset1'de optimum dropout değeri 0.4 olarak bulunmuştur. Dataset2 ve resDataset2'de optimum dropout değeri 0,5 olarak bulunmuştur. Benzer şekilde, Dataset1 ve resDataset1'de optimum recurrent dropout değeri 0,4'tür. Dataset1 ile resDataset1'de optimum recurrent dropout değeri 0,6 olarak kullanılmıştır. LSTM modelinde Adam optimizer kullanılmıştır, öğrenme oranı 1 x 10^{-4}'dir ve kayıp fonksiyonu category_crossentropy'dir. Eğitim sırasında küme boyutu 64'tür ve eğitim tur sayısı 10'dur.

Tablo 4.19, GCR-NN modelinin tüm veri kümelerindeki hiperparametre değerlerini göstermektedir. GCR-NN modeli için ağ yapısı şu şekilde ayarlanmıştır: kelime yerleştirme katmanı (birim = 512), 1. GRU katmanı (birim = 64), 2. CNN katmanı (filtre = 64, kernel =4), 3. MaxPooling1D katmanı (pool_size = 4), 4. dropout katmanı (0.2), 5. SimpleRNN katmanı (birim = 16) - yoğun katman (birim=16), yoğun katman (çıktı sayısı = 3). GCR-NN modelinde Adam optimizer kullanılmıştır ve kayıp fonksiyonu category_crossentropy'dir. Eğitim sırasında küme boyutu 16'dır ve eğitim tur sayısı 100'dür (early stopping ile).

Tablo 4.19. Veri kümeleri için GCR-NN modelinin en iyi parametre değerleri.

Parametre	İfade/Değer
Kelime yerleştirme katmanındaki çıktı boyutu sayısı	512
GRU katmanlarının sayısı	1
GRU birimi	64
Konvolüsyon katmanlarının sayısı	1
Filtre	64
Kernel	4
Aktivasyon konvolüsyon fonksiyonu	ReLU
Havuzlama	MaxPooling1D (pool_size = 4)
Dropout	0.2
SimpleRNN katman sayısı	1

Parametre	İfade/Değer
SimpleRNN birimi	16
Tam bağlı birim	16
Çıktı sayısı	3
Aktivasyon çıktı fonksiyonu	Softmax
Kayıp fonksiyonu	categorical_cross entropy
Optimizer	Adam
Eğitim tur sayısı	100 (early_stopping ile)
Küme boyutu sayısı	16

Verilerdeki gürültünün giderilmesi için ön işleme aşamasında bir çalışma yapılmıştır. URL bağlantıları, HTML etiketleri, hashtagler, emojiler, semboller, noktalama işaretleri ve durak Türkçe kelimeler veri setlerinden kaldırılmıştır. Tüm kelimeler küçük harfe dönüştürülmüştür. Türkçe dilbilgisi kurallarına göre yazım hataları düzeltilmiştir ve tekrarlanan semboller kaldırılmıştır.

Makro F1 puanı, veri kümelerinin eşit olmayan dağılımında bile iyi performans gösterdiği için COVID-19 veri kümelerinin analizinde kullanılmıştır. Dört veri seti (Dataset1, resDataset1, Dataset2 ve resDataset2), eğitim ve test olarak k-kat çapraz doğrulama yöntemiyle on parçaya bölünmüş, dokuz parça eğitim ve bir parça test için kullanılmıştır. Modellerin performansı, 10 kat parçaların test sürelerinin ortalaması alınarak belirlenmiştir.

Türkçe COVID-19 veri kümelerinin sınıflandırılmasında M-BERT dil modeli ile bir dilde öğrenilen sözdizimsel bağımlılık ilişkilerinin diğer dillerde de sürdürüldüğünü doğrulayarak diller arasında sözdizimsel bilginin transferine katkıda bulunmaktadır (Guarasci vd., 2022). COVID-19 veri kümelerinin analizinde kullanılan "bert-base-multilingual-cased", her biri 768 gizli birim ve 12 attention head içeren 12 katmandan oluşmaktaydı. Pytorch Huggingface dönüştürücü (transformatör) kitaplığından "bert-base-multilingual-cased" modeli seçilmiş ve metinsel veriler ayarlanmıştır.

4.3.1. TML, DL ve BERT Modellerinin Değerlendirme Sonuçları

Tablo 4.20 ve Tablo 4.21, Makro F1 sonuçlarını ve farklı öznitelik çıkarımı ile sınıflandırıcılar kullanılarak değerlendirilen toplam eğitim - test sürelerini göstermektedir. TML modelleri (SVM, NB ve RF), CNN, LSTM ve GCR-NN modelleri ile karşılaştırıldığında, ilk veri kümesinin (Dataset1 ve resDataset1) tüm sürümlerinde benzer şekilde performans göstermiştir. Ancak, One-hot kodlama + LSTM model kombinasyonu, TML ve DL modelleri arasında en yüksek Makro F1 puanını (Dataset1 veri kümesinde 0.717) elde etmiştir. Ayrıca, SVM sınıflandırıcı, NB sınıflandırıcıdan sonra ikinci en düşük sınıflandırma süresine sahiptir ve resDataset1'de TML ve DL modelleri arasında en iyi

sınıflandırma Makro F1 sonucuna (0.719) sahiptir. SVM sınıflandırıcı, metin sınıflandırma problemlerinin çözümünde etkili olmuştur (Kwak & Park, 2019).

Dataset1'in tüm sürümleri için yapılan değerlendirmelerde, BERT tabanlı transfer öğrenme modeli en iyi sonuçları vermiştir (Dataset1 veri setinde 0.742 ve resDataset1 veri setinde 0.786 Makro F1 puanı). Ancak BERT modelin oldukça uzun bir eğitim - test süresine sahip olduğu görülmektedir.

Tablo 4.20. Dataset1 veri setinde 10 kat çapraz doğrulama kullanan tüm sınıflar için modellerin Makro F1 puanları.

Modeller	Nitelik Çıkarımı	Sınıflandırıcı	Pozitif	Negatif	Nötr	Makro F1	Süre (saniye)
TML + Dataset1	unigram + TF-IDF	SVM	0.672	0.780	0.650	0.701	0.155
	bigram + TF-IDF		0.686	0.783	0.632	0.700	0.308
	trigram + TF-IDF		0.686	0.779	0.610	0.692	0.470
	unigram + TF-IDF	NB	0.658	0.775	0.615	0.683	**0.104**
	bigram + TF-IDF		**0.689**	0.786	0.613	0.696	0.282
	trigram + TF-IDF		0.680	0.783	0.609	0.691	0.366
	unigram + TF-IDF	RF	0.658	0.764	0.656	0.693	2.949
	bigram + TF-IDF		0.651	0.752	0.661	0.688	6.212
	trigram + TF-IDF		0.645	0.742	0.658	0.682	9.774
DL + Dataset1	One-hot kodlama	LSTM	0.676	**0.809**	0.667	0.717	283.31
	One-hot kodlama	CNN	0.645	0.785	0.639	0.69	90.121
	One-hot kodlama	GCR-NN	0.606	0.743	0.624	0.658	530.21
BERT temelli transfer öğrenme + Dataset1	BERT kelime yerleştirme		0.665	0.756	**0.806**	**0.742**	35998.12
TML + resDataset1	unigram + TF-IDF	SVM	0.671	0.771	0.654	0.699	0.266
	bigram + TF-IDF		0.685	0.787	0.674	0.715	0.414
	trigram + TF-IDF		0.694	0.788	0.675	0.719	0.588
	unigram + TF-IDF	NB	0.662	0.773	0.674	0.703	**0.138**
	bigram + TF-IDF		0.682	0.773	0.684	0.713	0.259
	trigram + TF-IDF		0.682	0.771	0.683	0.712	0.411
	unigram + TF-IDF	RF	0.670	0.737	0.658	0.688	3.090
	bigram + TF-IDF		0.671	0.723	0.650	0.681	6.296
	trigram + TF-IDF		0.663	0.726	0.652	0.680	9.600
DL + resDataset1	One-hot kodlama	LSTM	0.698	0.774	0.651	0.708	459.705
	One-hot kodlama	CNN	0.654	0.762	0.637	0.684	145.799
	One-hot kodlama	GCR-NN	0.617	0.768	0.645	0.677	700.854
BERT temelli transfer öğrenme + resDataset1	BERT kelime yerleştirme		**0.711**	**0.803**	**0.846**	**0.786**	55100.69

Dataset2'nin ayrı sınıflar (pozitif, negatif ve nötr) için değerlendirme sonuçları Tablo 4.21'de verilmiştir. İkinci veri seti sürümlerinin (Dataset2 ve resDataset2) sınıflandırma süreleri ile ilgili olarak, en iyi sınıflandırma süresinin unigram + TF-IDF + NB modeli ile elde edilen 0.244 saniye değeri olduğu görülmektedir. NB sınıflandırıcı, tüm veri setlerinde diğer sınıflandırıcılara kıyasla en iyi sınıflandırma süresine sahiptir.

Tablo 4.21. Dataset2 veri setindeki 10 kat çapraz doğrulama kullanan tüm sınıflar için modellerin Makro F1 puanları.

Modeller	Nitelik Çıkarımı	Sınıflandırıcı	Pozitif	Negatif	Nötr	Makro F1	Süre (saniye)
TML + Dataset2	unigram + TF-IDF	SVM	0.710	0.483	0.651	0.614	0.400
	bigram + TF-IDF		0.713	0.507	0.669	0.63	0.777
	trigram + TF-IDF		0.710	0.521	0.674	0.635	1.239
	unigram + TF-IDF	NB	0.707	0.447	0.669	0.608	**0.244**
	bigram + TF-IDF		0.711	0.463	0.678	0.617	0.615
	trigram + TF-IDF		0.707	0.443	0.670	0.607	1.001
	unigram + TF-IDF	RF	0.695	0.393	0.655	0.581	9.232
	bigram + TF-IDF		0.698	0.387	0.655	0.580	21.766
	trigram + TF-IDF		0.704	0.376	0.642	0.574	43.166
DL + Dataset2	One-hot kodlama	LSTM	0.656	0.496	0.648	0.600	530.307
	One-hot kodlama	CNN	0.662	0.525	0.644	0.610	128.184
	One-hot kodlama	GCR-NN	0.709	0.414	0.574	0.566	790.154
BERT temelli transfer öğrenme + Dataset2	BERT kelime yerleştirme		**0.782**	**0.578**	**0.772**	**0.711**	86098.73
TML + resDataset2	unigram + TF-IDF	SVM	0.641	0.496	0.705	0.614	0.466
	bigram + TF-IDF		0.661	0.515	0.708	0.628	0.991
	trigram + TF-IDF		0.673	0.526	0.709	0.636	1.460
	unigram + TF-IDF	NB	0.654	0.514	0.693	0.620	**0.330**
	bigram + TF-IDF		0.655	0.512	0.691	0.619	0.749
	trigram + TF-IDF		0.666	0.513	0.689	0.623	1.140
	unigram + TF-IDF	RF	0.646	0.483	0.705	0.611	9.197
	bigram + TF-IDF		0.655	0.468	0.721	0.615	22.128
	trigram + TF-IDF		0.647	0.465	0.699	0.604	39.284
DL + resDataset2	One-hot kodlama	LSTM	0.675	0.514	0.604	0.598	783.155
	One-hot kodlama	CNN	0.663	0.479	0.606	0.583	236.022
	One-hot kodlama	GCR-NN	0.697	0.399	0.576	0.557	1033,014

Modeller	Nitelik Çıkarımı	Sınıflandırıcı	Pozitif	Negatif	Nötr	Makro F1	Süre (saniye)
BERT temelli transfer öğrenme + resDataset2	BERT kelime yerleştirme		**0.792**	**0.57**	**0.774**	**0.712**	120039.46

Tüm modeller arasında en iyi sınıflandırma sonucuna sahip olan COVID-19 duyarlılık sınıflandırma modeli, veri kümelerinin iki versiyonundada M-BERT tabanlı transfer öğrenme modeli olmuştur.

Daha önce açıklandığı gibi, Dataset1 ve Dataset2'deki sınıflar (pozitif, negatif ve nötr) eşit olmayan bir dağılıma sahiptir. Sınıflandırma sonuçları, her sınıfın (pozitif, negatif ve nötr) performansının eğitim setindeki örnek sayısına bağlı olduğunu ortaya koymaktadır. Bu nedenle, daha fazla eğitim örneğine sahip sınıflar (resDataset1 ve resDataset2), eğitim kümesinde daha az örnek içeren sınıflardan (Dataset1 ve Dataset2) genellikle daha iyi sonuçlar elde etmiştir. Tablo 4.19 ve Tablo 4.20'de görüldüğü gibi BERT modeli diğer tüm modellerden (TML modelleri ve DL modelleri) daha iyi sonuçlar vermektedir. BERT modelinin sınıflandırma sonuçlarındaki başarısına rağmen tüm veri setlerinde sınıflandırma süreleri oldukça yüksektir.

4.3.2. Gizli Dirichlet Tahsisi (Latent Drichlet Allocation – LDA) Kullanılarak Konu Modelleme ile Duygu Analizi

COVID-19 veri kümelerindeki konu çıkarımı, LDA yöntemi kullanılarak analiz edilmiştir. LDA yöntemi, bir veri setinde verilen metinlerden konuları tanımlayabilen istatistiksel bir denetimsiz öğrenme modelidir (Xie vd., 2021). COVID-19 veri kümelerinin zaman içindeki duyarlılık değişiklikleri araştırılmış ve öne çıkan konular önemleri açısından analiz edilmiştir (Lee, Rustam, Ashraf, vd., 2022). Yorumlarda duygu analizine anlam katmak için konu çıkarımı yapılmıştır. Veri setlerindeki birbirleriyle ilgili konular konu çıkarım süreci ile birlikte sınıflandırılmaktadır.

Tüm etiketlerde en çok tartışılan üç konuyu (pozitif, negatif ve nötr) çıkarmak için n_components parametresi üç olarak ayarlanmıştır. Tablo 4.22, pozitif, nötr ve negatif olarak etiketlenmiş yorumlardan oluşan Dataset1'den elde edilen LDA konu çıkarımlarını ve içeriğini göstermektedir.

Konu içerikleri incelendiğinde Dataset1'de sınıf dağılımına göre (negatif, pozitif ve nötr) Dataset2'ye göre daha fazla "negatif" yorum bulunmaktadır. Dataset1'in toplandığı gün insanların daha fazla korku, öfke, şiddet, iğrenme, üzüntü ve alay duygularını ifade ettikleri söylenebilir. "Negatif" olarak etiketlenen yorumlarda, virüs Allah'tan gelen bir ceza olarak görülmüş ya da Allah'tan virüsü

bulaştıranları cezalandırması istenmiştir. "Pozitif" yorumların çoğunda, salgının sona ermesi için Allah'a duaların edildiği ve iyi dileklerin yazıldığı görülmektedir. Dataset1'de "nötr" olarak etiketlenen yorumların çoğu, Türkiye'de COVID-19'a yakalanan ilk kişi olabilecek kişinin yeri hakkında yalan haberler ve söylentiler içermektedir. Nötr yorumlardaki bir diğer önemli bulgu ise okulların kapatılması ve özellikle çocukların tekrar tekrar hastalıktan korunması için gerekli tedbirlerin alınmasının talep edilmesi olmuştur. Nötr yorumlarda el yıkamanın ve alkol bazlı dezenfektan kullanmanın önemine de değinilmiştir.

İlk vakanın bildirilmesinden sonraki ay boyunca (12 Mart 2020'den 10 Nisan 2020'ye kadar) toplanan ve Dataset2'yi oluşturan yorumlar, Dataset1'den daha nötr ve pozitif yorumlar içermektedir. Bu durum insanların bu süre zarfında korku, endişe ve güvensizlikten kurtulduğunu göstermektedir. Ayrıca bu sonuçlar pandemi sırasında hükümete ve sağlık görevlilerine olan güveni de belirtmektedir (Bhat vd., 2020).

Tablo 4.22. Dataset1 veri setindeki pozitif, negatif ve nötr yorumlardan çıkarılan LDA konu çıkarımları.

Dataset1	Konular	İçerikler
Pozitif	Konu 1	0.015*"olsun" + 0.010*"geç" + 0.010*"kişi" + 0.009*"bol" + 0.009*"virus" + 0.009*"öl" + 0.007*"rabbim" + .007*"önlem" + 0.006*"yok" + 0.006*"insan"
	Konu 2	0.009*"öl" + 0.009*"virus" + 0.009*"kork" + 0.009*"allah" + 0.008*"ülke" + 0.008*"yok" + 0.007*"insan" + 0.005*"panic" + 0.005*"tedbir" + 0.005*"bence"
	Konu 3	0.058*"allah" + 0.053*"okul" + 0.009*"virüs" + 0.007*"ülke" + 0.007*"korusun" + 0.006*"inşallah" + 0.006*"rabbim" + 0.006*"temiz" + 0.005*"il" + 0.005*"çocuk"
Negatif	Konu 1	0.016*"allah" + 0.013*"yurt" + 0.009*"yok" + 0.008*"okul (school)" + 0.008*"ülke" + 0.007*"kork" + 0.007*"dışına" + 0.007*"insan" + 0.006*"diyor" + 0.006*"oldu"
	Konu 2	0.011*"virüs" + 0.010*"giriş" + 0.010*"ülke" + 0.009*"çıkış" + 0.008*"öl" + 0.008*"türkiye" + 0.008*"yurt" + 0.007*"insan" + 0.006*"yurtdışı" + 0.006*"önlem"
	Konu 3	0.012*"virüs" + 0.009*"ülke" + 0.009*"korona" + 0.009*"il" + 0.008*"geç" + 0.008*"olsun" + 0.008*"çin" + 0.007*"kişi" + 0.006*"eyvah" + 0.006*"allah"
Nötr	Konu 1	0.023*"il" + 0.010*"şehir" + 0.009*"acaba" + 0.008*"diyor" + 0.007*"insan" + 0.006*"hastane" + 0.006*"kişi (person)" + 0.005*"virüs" + 0.005*"amin" + 0.005*"bence"
	Konu 2	0.018*"okul" + 0.014*"şehir" + 0.010*"virüs" + 0.008*"yok" + 0.007*"allah" + 0.006*"geç" + 0.005*"il" + 0.005*"vaka" + 0.004*"haber" + 0.004*"peki"
	Konu 3	0.018*"istanbul" + 0.010*"cnn" + 0.009*"gün" + 0.007*"kişi" + 0.007*"kuralı" + 0.007*"bakan" + 0.006*"hasta" + 0.005*"virüs" + 0.005*"evet"

Tablo 4.23, Dataset2'den elde edilen LDA konu çıkarımlarını ve içeriğini göstermektedir. Sonuçlardan da anlaşıldığı üzere pozitif yorumlardan alınan Konu1, Konu2 ve Konu3'te Allah'a dua etmenin önemi, yasakların faydaları ve sokağa çıkma yasağı lehinde görüşler yer almaktadır. Negatif yorumlarda ise Negatif yorumların Konu1, Konu2 ve Konu3 bölümlerinde gösterildiği gibi, negatif yorum içeren konuların içeriğinde çoğunlukla insanların COVID-19 hastalığına yönelik kaygıları ve ölüm korkusundan bahsedilmektedir. Dataset1 ve Dataset2'de pozitif ve negatif yorumlarda "Allah" kelimesi sıklıkla kullanılmıştır.

Tablo 4.23. Dataset2 veri setindeki pozitif, negatif ve nötr yorumlardan çıkarılan LDA konu çıkarımları.

Dataset2	Konular	İçerikler
Pozitif	Konu 1	0.015*"yasak" + 0.013*"inşallah" + 0.012*"çok" + 0.010*"iyi" + 0.009*"öl" + 0.009*"şükür" + 0.008*"gün" + 0.008*"güzel" + 0.008*"rabbim" + 0.008*"oldu"
	Konu 2	0.017*"hafta" + 0.014*"inşallah" + 0.011*"gün" + 0.010*"olsun" + 0.009*"sokağa" + 0.009*"yasak" + 0.008*"karar" + 0.008*"allah" + 0.007*"çıkma" + 0.007*"bakan"
	Konu 3	0.018*"yasak" + 0.015*"çıkma" + 0.015*"allah" + 0.015*"sokağa" + 0.013*"geç" + 0.011*"gelsin" + 0.009*"olsun" + 0.009*"bile" + 0.009*"insan" + 0.009*"sonunda "
Negatif	Konu 1	0.013*"insan" + 0.010*"öl" + 0.010*"diyor" + 0.009*"allah" + 0.007*"millet" + 0.006*"yasak" + 0.005*"kişi" + 0.005*"sokağa" + 0.005*"evde" + 0.005*"çalışan"
	Konu 2	0.021*"yasak" + 0.017*"sokağa" + 0.016*"çıkma" + 0.010*"öl" + 0.007*"gün" + 0.006*"evde" + 0.006*"kişi" + 0.006*"diyor" + 0.006*"insan" + 0.005*"yok"
	Konu 3	0.009*"insan" + 0.009*"virüs" + 0.008*"evde" + 0.007*"geç" + 0.007*"diyor" + 0.005*"yok" + 0.005*"öl" + 0.005*"yasak" + 0.004*"gün" + 0.004*"kork"
Nötr	Konu 1	0.006*"allah" + 0.006*"insan" + 0.006*"yardım" + 0.005*"yasak" + 0.005*"borç" + 0.004*"olsun" + 0.004*"amin" + 0.004*"gün" + 0.004*"aynen" + 0.004*"git"
	Konu 2	0.009*"insan" + 0.007*"öl" + 0.007*"yok" + 0.005*"virüs" + 0.005*"diyor" + 0.005*"çalışan" + 0.004*"olan" + 0.004*"haber" + 0.004*"evde" + 0.004*"kişi"
	Konu 3	0.009*"inşallah" + 0.009*"diyor" + 0.008*"öl" + 0.006*"yasak" + 0.006*"evde" + 0.006*"genç" + 0.005*"virüs" + 0.005*"yok" + 0.004*"yaşlı" + 0.004*"millet"

Dataset1 ve Dataset2 veri kümelerine ait Mann-Whitney parametrik olmayan t testi p değerleri, aynı veri kümelerinde kategoriler (pozitif ve negatif) arasında anlamlı bir fark olup olmadığını belirlemek için hesaplanmıştır (Mancini vd., 2011). Dataset1 ve Dataset2'nin Mann-Whitney parametrik olmayan t-testi p değerleri sırasıyla 0,0012 ve 0,0011'dir. Bu değerler $p < 0,05$ olduğu için negatif yorumlar ile pozitif yorumlar arasında farklılıklar vardır. Sonuçlar, Instagram'da COVID-19 salgını hakkında yapılan pozitif ve negatif yorumlar arasında anlamlı bir benzerlik olmadığını kanıtlamıştır. Türkiye'de salgın sürecini olumlu yönde desteklemek veya salgınları önlemek açısından faydalı bir etkileşimin olmadığı p değerleri ile kanıtlanmıştır.

4.3.3. COVID-19 Duygu Analizi Performans Karşılaştırması

Sınıflandırma modellerinin performans karşılaştırması ve Türk sosyal medyasının salgına etkisine ilişkin sonuçlar aşağıda verilmiştir:

- Dataset1'e uygulanan yeniden örnekleme yönteminin, SVM, NB, GCR-NN ve BERT modelleri için Makro F1 metriklerinde daha iyi sonuçlar verdiği görülmektedir. Dataset2 için yeniden örneklemenin SVM, NB, RF ve BERT modellerinde olumlu sonuç verdiği söylenebilmektedir.
- Önerilen BERT tabanlı transfer öğrenme modeli, tüm modeller arasında en iyi performansı göstermiştir. Modeller arasında en yüksek sınıflandırma sonuçları, resDataset1 veri kümesinde BERT modeli tarafından elde edilen 0.7864 Makro F1 değeri ve resDataset2'deki 0.7120 değeri olmuştur. BERT modeli milyonlarca önceden eğitilmiş veri içerdiğinden, bu model ile dört veri setinde önceden eğitilmiş verilerle sınıflandırma sonuçlarına ulaşılmıştır. Böylece model, bol miktarda veriyi eğitmeden veri kümelerini test etme fırsatı sağlamıştır. BERT modeli ile veri setlerinin sınıflandırma başarısı diğer modellere göre önemli ölçüde artmıştır.
- BERT modeli diğer modellere göre daha iyi sonuçlar vermesine rağmen tüm veri setlerinde sınıflandırma süresi çok uzun olmuştur. NB sınıflandırıcı, tüm veri setlerinde en kısa sınıflandırma süresine sahip sınıflandırıcı olmuştur.
- CNN, LSTM ve GCR-NN modelleri, her iki veri kümesi için de TML modellerinden (SVM, NB ve RF) daha iyi performans göstermemiştir. Veri kümeleri aşırı örneklenmiş olsa da, DL mimarileri için daha fazla veriye ihtiyaç olduğu düşünülmektedir.
- SVM sınıflandırıcısı, BERT tabanlı transfer öğreniminden sonra en iyi ikinci performans sonuçlarına sahiptir. Ayrıca, SVM sınıflandırıcı, tüm veri kümelerinde ikinci en iyi sınıflandırma süresine sahiptir. SVM modeli, sınıflandırma zamanı ile birlikte düşünüldüğünde COVID-19 duygu analizi çalışması için başarılıdır.

- Veri kümelerinde LDA konu çıkarımları yapılması sonucunda Dataset1'de salgının başlangıcındaki negatif yorumların (korku, panik ve öfke) fazlalığının Dataset2'de pozitif ve nötr yorumlara dönüştüğü görülmektedir. Dataset2'de pozitif ve nötr yorumların artmasından, salgının ilk ayında (11 Mart - 10 Nisan 2020) insanların duygularının olumlu ya da kararsız olduğu anlaşılmaktadır. LDA ile elde edilen konulara göre her iki veri setinde de özellikle nötr yorumlarda söylenti ve yalan haber sayısının yüksek olduğu gözlemlenmiştir.
- COVID-19 duygu analizi çalışması kapsamında elde edilen Dataset1 ve Dataset2'de benzer birçok yorum bulunsa da salgınla ilgili yönlendirici ve bilgilendirici açıklamalar tespit edilememiştir. Ayrıca her iki veri setinde (Dataset1 ve Dataset2) parametrik olmayan t testi p değerinin 0,05'ten küçük olması nedeniyle Instagram sosyal ağının pandemi hakkında bilgi vermediği veya farkındalığa neden olmadığı sonucuna varılmıştır.

5. SONUÇLAR ve ÖNERİLER

Bu tez çalışmasında, sosyal medyada Türkçe küfür ve homofobi içeren nefret söylemleri tespitine yönelik DL ve BERT gibi son teknoloji sınıflandırıcılar TML ve topluluk modeli sınıflandırıcıları ile karşılaştırılmıştır. Öncelikle SBTC çalışmalarında kullanılacak veri kümeleri Instagram ağından elde edilmiş, etiketlenmiş, eksik ya da hatalı yorumlar silinmiş ve yorumlarda varolan yazım yanlışları Türkçe dil kurallarına göre düzeltilmiştir. Sosyal ağlarda nefret söylemi analizi birçok dilde yapılmış olsa da çalışmamızda kullanılan veri kümeleri bildiğimiz kadarıyla homofobi ve küfür verilerinin ayrı ayrı sınıflandırılmasından dolayı ilk Türkçe veri setleridir. Veri kümeleri etiket dağılımı bakımından dengesiz olduğundan ATC veri kümesinde rasgele aşırı örnekleme ve HATC veri kümesine ise rasgele yeniden örnekleme işlemi uygulanarak veri kümelerinin dengeli versiyonları ile de sınıflandırma sonuçları alınmıştır. Homofobi ve küfür ile ilgili emojiler silinmemiş ve etiketleme adımları sırasında göz önünde bulundurulmuştur.

Ayrıca COVID-19 hastalığının yayılmaya başladığı bir aylık süre boyunca kişilerin duygu değişimlerini incelemek ve otomatik sınıflandırma yapmak için sosyal medyada yazdıkları yorumlar toplanmıştır. Tüm yapılan çalışmaların değerlendirme adımları ayrıntılı olarak açıklamalarla ilgili bölümlerde verilmiştir. Önerilen yöntemlerin etkinliğini ve geçerliliğini test etmek için veri kümeleri hem orijinal halleriyle hem de dengelenerek veriler hazırlanmıştır. Sınıflandırma sonuçları bu tez çalışmasında önerilen BERT ve DL tekniklerinin başarısını ortaya koymuştur.

Türkçe küfürlü ifadelerin tespitinde CNN metodu, F1 skor, Kesinlik ve Hassasiyet açısından diğer sınıflandırma modellerine göre iyi bir sınıflandırma performansı elde edildiği görülmüştür. Homofobik söylemlerin tespitinde önceden eğitilmiş M-BERT modeli, Türkçe yorum filtrelerinde kullanılacak homofobi tespit modeli için uygun bir aday olma potansiyeline sahiptir. Kullanılan M-BERT modeli, 104 dilde önceden eğitilmiş kaynaklara sahiptir ve farklı metin dillerinin formatını dikkate alabildiğinden, farklı dillerde homofobik ve nefret söylemi çalışmalarında kullanılabilir.

Sosyal medyanın pandemideki rolünü ve katkısını belirlemek için yapılan değerlendirmelerin (LDA, kelime sıklığı, doğrusal olmayan t-testinin p değerleri vb.) sonuçlarına göre, sosyal medyada COVID-19 hakkındaki Türkçe yorumlar arasında entelektüel bir etkileşim bulunamamıştır. COVID-19 Türkçe duygu analizi çalışmasında, M-BERT tabanlı transfer öğrenmesi modeli Makro F1 puanı açısından diğer modellerden daha iyi sınıflandırma performansı elde etmiştir. Türkçe veri setlerinde önceden eğitilmiş çok dilli bir dil modelinin kullanılmasının sınıflandırma performansı açısından diğer modellere göre daha başarılı olduğu kanıtlanmıştır. Nefret söylemi çalışmalarında en çok kullanılan ağ Twitter'dır. Instagram'ın artan popüleritesi ile birlikte nefret söylemi türleri için yapılan çalışmalarda artmıştır. Instagram'ın görseller ve videolar üzerinden kullanıcılarla kurduğu iletişim nefret söylemi türleri içinde istenilen türde yorumları elde etmede kolaylık sağlamaktadır.

Gelecekte ATC veri setinden farklı nefret türleri kullanılarak daha fazla SBTC modeli geliştirmek için çalışmalar yapılabilir. Nefret söylemlerinden 3. Sınıf türe giren ve Türkçe veri kümesi olmayan aşağılamayla ilgili sınıflandırma modelleri geliştirilebilir. Veri kümelerinde yorumlar artırılarak, DL modellerinin parametre değerleri değiştirilerek ve güncellenmiş kelime yerleştirme modelleri kullanılarak sınıflandırma doğruluk değerleri artırılabilir.

KAYNAKLAR

Abooraig, R., Al-Zu'bi, S., Kanan, T., Hawashin, B., Al Ayoub, M., & Hmeidi, I. (2018). Automatic categorization of Arabic articles based on their political orientation. *Digital Investigation*, *25*, 24-41. https://doi.org/10.1016/j.diin.2018.04.003

Abroyan, N. (2017). Convolutional and recurrent neural networks for real-time data classification. *7th International Conference on Innovative Computing Technology, INTECH 2017*, 42-45. https://doi.org/10.1109/INTECH.2017.8102422

Agnihotri, D., Verma, K., & Tripathi, P. (2017). Variable Global Feature Selection Scheme for automatic classification of text documents. *Expert Systems with Applications*, *81*, 268-281. https://doi.org/10.1016/j.eswa.2017.03.057

Akhtar, S., Basile, V., & Patti, V. (2019). A New Measure of Polarization in the Annotation of Hate Speech. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11946 LNAI*, 588-603. https://doi.org/10.1007/978-3-030-35166-3_41

Aktunç, H. (2000). *Big slang dictionary of Turkish: (with witnesses)*. YapıKredi Yayınları.

Al-Garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior*, *63*, 433-443. https://doi.org/10.1016/J.CHB.2016.05.051

Al-Hassan, A., & Al-Dossari, H. (2019). Detection of Hate Speech In Social Networks: A Survey On Multilingual Corpus. 83-100. https://doi.org/10.5121/csit.2019.90208

Al-Radaideh, Q. A., & Al-Abrat, M. A. (2019). An Arabic text categorization approach using term weighting and multiple reducts. *Soft Computing*, *23*(14), 5849-5863. https://doi.org/10.1007/s00500-018-3249-z

Alakrot, A., Murray, L., & Nikolov, N. S. (2018). Dataset Construction for the Detection of Anti-Social Behaviour in Online Communication in Arabic. *Procedia Computer Science*, *142*, 174-181. https://doi.org/10.1016/j.procs.2018.10.473

Alayba, A. M., Palade, V., England, M., & Iqbal, R. (2017). Arabic Language Sentiment Analysis on Health Services. 114-118. https://doi.org/10.1109/ASAR.2017.8067771

Almerekhi, H., Jansen, B. J., Kwak, H., & Salminen, J. (2019). Detecting toxicity triggers in online discussions. *HT 2019 - Proceedings of the 30th ACM Conference on Hypertext and Social Media*, 291-292. https://doi.org/10.1145/3342220.3344933

Alparslan, E., Karahoca, A., & Bahşi, H. (2013). Security-level classification for confidential documents by using adaptive neuro-fuzzy inference systems. *Expert Systems*, *30*(3), 233-242. https://doi.org/10.1111/J.1468-0394.2012.00634.X

Arango, A., Pérez, J., & Poblete, B. (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. *SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 45-53. https://doi.org/10.1145/3331184.3331262

Aroyehun, S. T., & Gelbukh, A. (2018). Aggression Detection in Social Media: Using Deep Neural Networks, Data Augmentation, and Pseudo Labeling. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, 90-97. https://aclanthology.org/W18-4411

Aya, S. A., Ormanci Acar, T., & Tufekci, N. (2016). Modeling of membrane fouling in a submerged membrane reactor using support vector regression. *Desalination and Water Treatment*, *57*(51), 24132-24145. https://doi.org/10.1080/19443994.2016.1140080

Ayata, D., Saraçlar, M., & Özgür, A. (2017). *Political opinion/sentiment prediction via long short term memory recurrent neural networks on Twitter* 25th Signal Processing and Communications Applications Conference, https://ieeexplore.ieee.org/abstract/document/7960733/

Babaeianjelodar, M., Lorenz, S., Gordon, J., Matthews, J., & Freitag, E. (2020). Quantifying Gender Bias in Different Corpora. *The Web Conference 2020 - Companion of the World Wide Web Conference, WWW 2020*, 752-759. https://doi.org/10.1145/3366424.3383559

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets. *26th International World Wide Web Conference 2017, WWW 2017 Companion*, 759-760. https://doi.org/10.1145/3041021.3054223

Balakrishnan, V., Khan, S., & Arabnia, H. R. (2020). Improving cyberbullying detection using Twitter users' psychological features and machine learning. *Computers and Security*, *90*, 101710-101710. https://doi.org/10.1016/j.cose.2019.101710

Banks, J. (2014). European Regulation of Cross-Border Hate Speech in Cyberspace: The Limits of Legislation. *Eur. J. Crime Crim. L. & Crim. Just.*, *4*(13), 33-48. https://dergipark.org.tr/tr/pub/kbchd/905863

BBC_News. (2020). *Siber zorbalık nedir, siber zorbalıkla mücadele için hangi önlemler alınabilir? - BBC News Türkçe*. https://www.bbc.com/turkce/haberler-turkiye-51614553

Benballa, M., Collet, S., & Picot-Clemente, R. (2019). Saagie at Semeval-2019 Task 5: From Universal Text Embeddings and Classical Features to Domain-specific Text Classification. *In Proceedings of the 13th International Workshop on Semantic Evaluation*, 469-475. https://doi.org/10.18653/V1/S19-2083

Bhat, M., Qadri, M., Beg, N. u. A., Kundroo, M., Ahanger, N., & Agarwal, B. (2020). Sentiment analysis of social media response on the Covid19 outbreak. *87*, 136-137. https://doi.org/10.1016/j.bbi.2020.05.006

Bilge, R. (2016). The Construction of Hate Speech on Social Media and Legal Regulations on Hate Crimes. *Yeni Medya*(1), 1-14. https://dergipark.org.tr/tr/pub/yenimedya/760546

Bimantara, A. A., Larasati, A., Risondang, E. M., Naf'an, M. Z., Alim, N., & Nugraha, S. (2019). Sentiment Analysis of Cyberbullying on Instagram User Comments. *Journal of Data Science and Its Applications*, *2*(1), 38-48. https://doi.org/10.21108/JDSA.2019.2.20

Breiman, L. (2001). Random Forests. *Machine Learning 2001 45:1*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Briliani, A., Irawan, B., & Setianingsih, C. (2019). Hate speech detection in indonesian language on instagram comment section using K-nearest neighbor classification method. *Proceedings - 2019 IEEE International Conference on Internet of Things and Intelligence System, IoTaIS 2019*, 98-104. https://doi.org/10.1109/IOTAIS47347.2019.8980398

Burnap, P., & Matthew, L. W. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics | Semantic Scholar. *EPJ Data Science*(5), 1-15.

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy and Internet*, *7*(2), 223-242. https://doi.org/10.1002/poi3.85

Cammaerts, B. (2010). Radical pluralism and free speech in online public spaces: the case of North Belgian extreme right discourses. *International journal of cultural studies*. https://doi.org/10.1177/1367877909342479

Carmona, M. A. Á., Guzmán-Falcón, E., Montes-y-Gómez, M., Escalante, H., Pineda, L. V., Reyes-Meza, V., & Sulayes, A. R. (2019). Overview of MEX-A3T at IberLEF 2019: Authorship and Aggressiveness Analysis in Mexican Spanish Tweets. *IberLEF@ SEPLN*, 478-494.

Chablani, M. (2017). *Word2Vec (skip-gram model): PART 1 - Intuition. | by Manish Chablani | Towards Data Science*. Erişim Tarihi 01 Ocak 2022, https://towardsdatascience.com/word2vec-skip-gram-model-part-1-intuition-78614e4d6e0b

Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R., & Hassanien, A. E. (2020). Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media. *Applied Soft Computing Journal*, *97*, 106754-106754. https://doi.org/10.1016/j.asoc.2020.106754

Chamberlain, B. P., Rossi, E., Shiebler, D., Sedhain, S., & Bronstein, M. M. (2020). Tuning Word2vec for Large Scale Recommendation Systems. *RecSys 2020 - 14th ACM Conference on Recommender Systems*, 732-737. https://doi.org/10.1145/3383313.3418486

Charitidis, P., Doropoulos, S., Vologiannidis, S., Papastergiou, I., & Karakeva, S. (2020). Towards countering hate speech against journalists on social media. *Online Social Networks and Media*, *17*, 100071-100071. https://doi.org/10.1016/J.OSNEM.2020.100071

Chatzakou, D., Leontiadis, I., Blackburn, J., De Cristofaro, E., Stringhini, G., Vakali, A., & Kourtellis, N. (2019). Detecting Cyberbullying and Cyberaggression in Social Media. *ACM Transactions on the Web*, *13*(3). https://arxiv.org/abs/1907.08873v1

Chen, H., McKeever, S., & Delany, S. J. (2017). Harnessing the power of text mining for the detection of abusive content in social media. *513*, 187-205. https://doi.org/10.1007/978-3-319-46562-3_12

Chen, H., McKeever, S., Jane Delany, S., & McKeever, S. (2016). Harnessing the Power of Text Mining for the Detection of Abusive Content in Social Media, 16th UKCI. https://doi.org/10.1007/978-3-319-46562-3_12

Chollet, F. (2018). *Keras: The Python Deep Learning library - NASA/ADS*. Astrophysics source code library. Erişim Tarihi 20 Mart 2022, https://ui.adsabs.harvard.edu/abs/2018ascl.soft06022C/abstract

Colaboratory. (2022). *Colaboratory*. Erişim Tarihi 2 Mayıs 2022, https://colab.research.google.com/

Corazza, M., Menini, S., Cabrio, E., Tonelli, S., & Villata, S. (2020). A Multilingual Evaluation for Online Hate Speech Detection. *ACM Transactions on Internet Technology*, *20*(2), 1-22. https://doi.org/10.1145/3377323

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning 1995 20:3*, *20*(3), 273-297. https://doi.org/10.1007/BF00994018

Crawl, C. (2022). Erişim Tarihi 1 Ocak 2022, https://commoncrawl.org/2021/

Cruz, R. M. O., de Sousa, W. V., & Cavalcanti, G. D. C. (2022). Selecting and combining complementary feature representations and classifiers for hate speech detection. *Online Social Networks and Media*, *28*. https://doi.org/10.48550/arxiv.2201.06721

Çakıcı, R., Steedman, M., & Bozşahin, C. (2018). Wide-Coverage Parsing, Semantics, and Morphology. In (pp. 153-174). https://doi.org/10.1007/978-3-319-90165-7_8

Çelenli, H. İ. (2020). *Gizli dirichlet ayrımı ve Word2vec yöntemlerinin birleşimi ile özgün bir metin temsil modeli geliştirilmesi* [Master, Kocaeli Üniversitesi]. https://acikbilim.yok.gov.tr/handle/20.500.12812/410902

Çomu, T., & Binark, M. (2012). *Video Paylaşım Ağlarında Nefret Söylemi: Youtube Örneği* [Master, Ankara Üniversitesi].

Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment Analysis Based on Deep Learning: A Comparative Study. *Electronics (Switzerland)*, *9*(3). https://doi.org/10.3390/electronics9030483

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). *Automated Hate Speech Detection and the Problem of Offensive Language* Proceedings of the international AAAI conference on web and social media,

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 4171-4186. http://arxiv.org/abs/1810.04805

Dogan, T., & Uysal, A. K. (2019). Improved inverse gravity moment term weighting for text classification. *Expert Systems with Applications*, *130*, 45-59. https://doi.org/10.1016/j.eswa.2019.04.015

Dondurucu, Z. B. (2018). Sexual Identity Based Hate Speech in New Media: Inci Sozluk Sample. *Gumushane University E-Journal of Faculty of Communication*, *6*(2), 1376-1405. https://doi.org/10.19145/E-GIFDER.435744

Dwivedi, R. K., Aggarwal, M., Keshari, S. K., & Kumar, A. (2019). Sentiment analysis and feature extraction using rule-based model (RBM). *Lecture Notes in Networks and Systems*, *56*, 57-63. https://doi.org/10.1007/978-981-13-2354-6_7

El-Kahlout, I. D., & Akin, A. A. (2013). Turkish Constituent Chunking with Morphological and Contextual Features. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7816 LNCS*(PART 1), 270-281. https://doi.org/10.1007/978-3-642-37247-6_22

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., & Belding, E. (2018). Peer to Peer Hate: Hate Speech Instigators and Their Targets. *12th International AAAI Conference on Web and Social Media, ICWSM 2018*, 52-61. https://arxiv.org/abs/1804.04649v1

Eryiğit, G., Nivre, J., & Oflazer, K. (2008). Dependency Parsing of Turkish. *Computational Linguistics*, *34*(3), 357-389. https://doi.org/10.1162/coli.2008.07-017-R1-06-83

Facebook. (2022). Erişim Tarihi 3 Şubat 2022, https://tr-tr.facebook.com/

Fatima, M., & Pasha, M. (2017). Survey of Machine Learning Algorithms for Disease Diagnostic. *Journal of Intelligent Learning Systems and Applications*, *09*(01), 1-16. https://doi.org/10.4236/jilsa.2017.91001

Ferrone, L., & Zanzotto, F. M. (2020). Symbolic, Distributed, and Distributional Representations for Natural Language Processing in the Era of Deep Learning: A Survey. *Frontiers in Robotics and AI*, *6*. https://doi.org/10.3389/FROBT.2019.00153

Fortuna, P., Soler-Company, J., & Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, *58*(3), 102524-102524. https://doi.org/10.1016/J.IPM.2021.102524

Founta, A.-M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. http://arxiv.org/abs/1802.00393

Freund, Y., & Schapire, R. E. (1996). Experiments with a New Boosting Algorithm Draft-Please Do Not Distribute. http://www.research.att.com/orgs/ssr/people/fyoav,schapireg/

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine on JSTOR. *The Annals of Statistics*, *29*(5), 1189-1232. https://www.jstor.org/stable/2699986

Gambäck, B., & Sikdar, U. (2017). Using Convolutional Neural Networks to Classify Hate-Speech. *In Proceedings of the first workshop on abusive language online*, 85-90. https://doi.org/10.18653/v1/W17-3013

Gao, L., & Huang, R. (2017). Detecting Online Hate Speech Using Context Aware Models. *arXiv preprint arXiv:1710.07395.*, 260-266. https://doi.org/10.26615/978-954-452-049-6_036

Gao, Z., Feng, A., Song, X., & Wu, X. (2019). Target-dependent sentiment classification with BERT. *IEEE Access*, *7*, 154290-154299. https://doi.org/10.1109/ACCESS.2019.2946594

García-Díaz, J. A., Cánovas-García, M., & Valencia-García, R. (2020). Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in Latin America. *Future Generation Computer Systems*, *112*, 641-657. https://doi.org/10.1016/j.future.2020.06.019

Gazzah, S., & Amara, N. E. B. (2008). New oversampling approaches based on polynomial fitting for imbalanced data sets. *IAPR International Workshop on Document Analysis Systems*, 677-684. https://doi.org/10.1109/DAS.2008.74

Gertner, A. S., Henderson, J., Merkhofer, E., Marsh, A., Wellner, B., & Zarrella, G. (2019). MITRE at SemEval-2019 Task 5: Transfer Learning for Multilingual Hate Speech Detection. *Proceedings of the 13th International Workshop on Semantic Evaluation*, 453-459. https://doi.org/10.18653/V1/S19-2080

Ghorbanali, A., Sohrabi, M. K., & Yaghmaee, F. (2022). Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks. *Information Processing & Management*, *59*(3), 102929-102929. https://doi.org/10.1016/J.IPM.2022.102929

Gibert, O. d., Pérez, N., Pablos, A. G., & Cuadros, M. (2018). Hate Speech Dataset from a White Supremacy Forum. *arXiv preprint arXiv:1809.04444*, 11-20. https://doi.org/10.18653/V1/W18-5102

Guarasci, R., Silvestri, S., De Pietro, G., Fujita, H., & Esposito, M. (2022). BERT syntactic transfer: A computational experiment on Italian, French and English languages. *Computer Speech & Language*, *71*, 101261-101261. https://doi.org/10.1016/J.CSL.2021.101261

Guven, Z. A. (2021). Comparison of BERT Models and Machine Learning Methods for Sentiment Analysis on Turkish Tweets. 98-101. https://doi.org/10.1109/UBMK52708.2021.9559014

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. Elsevier Inc. https://doi.org/10.1016/C2009-0-61819-5

Haque, M. A. (2014). Sentiment Analysis by Using Fuzzy Logic. *International Journal of Computer Science, Engineering and Information Technology*, *4*(1), 33-48. https://arxiv.org/abs/1403.3185v1

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, *33*(3), 464-472. https://doi.org/10.1016/J.IJINFOMGT.2013.01.001

Heinzerling, B., & Strube, M. (2017). BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 2989-2993. https://doi.org/10.48550/arxiv.1710.02187

Heirman, W., & Walrave, M. (2008). Assessing Concerns and Issues about the Mediation of Technology in Cyberbullying. *Cyberpsychology*, *2*, 1-12.

Hemmatian, F., & Sohrabi, M. K. (2019). A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review*, *52*(3), 1495-1545. https://doi.org/10.1007/s10462-017-9599-6

Hernández-Castañeda, Á., García-Hernández, R. A., Ledeneva, Y., & Millán-Hernández, C. E. (2022). Language-independent extractive automatic text summarization based on automatic keyword extraction. *Computer Speech & Language*, *71*, 101267-101267. https://doi.org/10.1016/J.CSL.2021.101267

Hmeidi, I., Al-Ayyoub, M., Abdulla, N. A., Almodawar, A. A., Abooraig, R., & Mahyoub, N. A. (2015). Automatic Arabic text categorization: A comprehensive comparative study. *Journal of Information Science*, *41*(1), 114-124. https://doi.org/10.1177/0165551514558172

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2015). Detection of Cyberbullying Incidents on the Instagram Social Network. *Association for the Advancement of Artificial Intelligence*.

Hu, Y. M. L., & Kambhampati, S. (2014). What we instagram: A first analysis of instagram photo content and user types. *In Eighth International AAAI conference on weblogs and social media*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewPaper/8118

Huang, B., & Raisi, E. (2018). Weak Supervision and Machine Learning for Online Harassment Detection. *Online Harassment*, 5-28. https://doi.org/10.1007/978-3-319-78583-7_2

Huang, Z., Cao, Y., & Wang, T. (2019). Transfer learning with efficient convolutional neural networks for fruit recognition. *Proceedings of 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2019*, 358-362. https://doi.org/10.1109/ITNEC.2019.8729435

Ibrohim, M. O., & Budi, I. (2018). A Dataset and Preliminaries Study for Abusive Language Detection in Indonesian Social Media. *135*, 222-229. https://doi.org/10.1016/j.procs.2018.08.169

Instagram. (2022). Erişim Tarihi 9 Şubat 2022, https://www.instagram.com/

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *C*, *6*(5), 429-449. https://doi.org/10.3233/IDA-2002-6504

Jenkins, J. (2002). A Sociolinguistically Based, Empirically Researched Pronunciation Syllabus for English as an International Language. *Applied Linguistics*, *23*(1), 83-103. https://doi.org/10.1093/APPLIN/23.1.83

Johnson, B. G. (2018). Tolerating and managing extreme speech on social media. *Internet Research*, *28*(5), 1275-1291. https://doi.org/10.1108/IntR-03-2017-0100

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., & Dean, J. (2017). Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, *5*, 339-351. https://doi.org/10.1162/TACL_A_00065/43400/Google-S-Multilingual-Neural-Machine-Translation

Jones, L. M., Mitchell, K. J., & Finkelhor, D. (2013). Online harassment in context: Trends from three youth internet safety surveys (2000, 2005, 2010). *Psychology of violence*, *3.1*, 53-69. https://doi.org/10.1037/a0030309

Kaplan, K. (2020). *Beyin Tümör Tiplerinin Makine Öğrenmesi Ve Derin Öğrenme Tabanlı Teknikler İle Sınıflandırılması*

Kapočiūtė-Dzikienė, J., & Tesfagergish, S. G. (2020). Part-of-Speech Tagging via Deep Neural Networks for Northern-Ethiopic Languages. *Information Technology and Control*, *49*(4), 482-494. https://doi.org/10.5755/J01.ITC.49.4.26808

Karayigit, H., İnan Acı, Ç., & Akdagli, A. (2020). Abusive Instagram Comments in Turkish | Datasets Novice | Kaggle. In.

Karayigit, H., İnan Acı, Ç., & Akdagli, A. (2021). Detecting abusive Instagram comments in Turkish using convolutional Neural network and machine learning methods. *Expert Systems with Applications*, *174*, 114802-114802. https://doi.org/10.1016/J.ESWA.2021.114802

Kasakowskij, T., Fürst, J., Fischer, J., & Fietkiewicz, K. J. (2020). Network enforcement as denunciation endorsement? A critical study on legal enforcement in social media. *Telematics and Informatics*, *46*, 101317-101317. https://doi.org/10.1016/J.TELE.2019.101317

Kaya, A. (2021). Bir Araştırma Kaynağı Olarak Arşivlenen Sosyal Medya Verilerinin Kullanımı. *Bilgi Ve Belge Araştırmaları*, *0*(16), 49-79. https://doi.org/10.26650/BBA.2021.16.1008002

Kemp, S. (2021). *Digital 2021 October Global Statshot Report. DataReportal—Global Digital Insights*. https://datareportal.com/reports/digital-2021-october-global-statshot.

Khan, W., Daud, A., Nasir, J. A., Science, T. A. K. j. o., & undefined. (2016). A survey on the state-of-the-art machine learning models in the context of NLP. *journalskuwait.org*, *43*(4), 95-113. https://journalskuwait.org/kjs/index.php/KJS/article/view/946

Kiliç, S. (2015). Kappa Test. *Psychiatry and Behavioral Sciences*(5(3)), 142-142. https://doi.org/10.5455/jmood.20150920115439

Kilincer, I. F., Ertam, F., & Sengur, A. (2022). A comprehensive intrusion detection framework using boosting algorithms. *Computers and Electrical Engineering*, *100*, 107869-107869. https://doi.org/10.1016/J.compeleceng.2022.107869

Kilinç, D., Ozcift, A., Bozyigit, F., Yildirim, P., Yücalar, F., & Borandağ, E. (2017). TTC-3600: A new benchmark dataset for Turkish text categorization. *Journal of Information Science*, *43*(2), 174-185. https://doi.org/10.1177/0165551515620551

Kim, H., & Jeong, Y. S. (2019a). Sentiment Classification Using Convolutional Neural Networks. *Applied Sciences 2019, Vol. 9, Page 2347*, *9*(11), 2347-2347. https://doi.org/10.3390/APP9112347

Kim, H., & Jeong, Y. S. (2019b). Sentiment classification using Convolutional Neural Networks. *Applied Sciences (Switzerland)*, *9*(11). https://doi.org/10.3390/APP9112347

Kinalioğlu, İ. H. (2019). *Yapay Zeka Tabanlı Yöntemler Kullanılarak Futbol Müsabakalarının Sonuçlarının Kestirilmesi Ve Hibrit Model Önerileri* [Phd, Selçuk Üniversitesi].

Kishore Kumar, R., & Sreenivasa Rao, K. (2022). A novel approach to unsupervised pattern discovery in speech using Convolutional Neural Network. *Computer Speech & Language*, *71*, 101259-101259. https://doi.org/10.1016/J.CSL.2021.101259

Kumar, A., Abirami, S., Trueman, T. E., & E., C. (2021). Comment toxicity detection via a multichannel convolutional bidirectional gated recurrent unit. *Neurocomputing*, *441*, 272-278. https://doi.org/10.1016/J.NEUCOM.2021.02.023

Kwak, G.-H., & Park, N.-W. (2019). Impact of Texture Information on Crop Classification with Machine Learning and UAV Images. *Applied Sciences*, *9*(4), 643-643. https://doi.org/10.3390/app9040643

Kwok, I., & Wang, Y. (2013). *Locate the Hate: Detecting Tweets against Blacks* Twenty-seventh AAAI conference on artificial intelligence, http://tempest.wellesley.edu/~ywang5/aaai/paper.html.

Le, T., Hoang Son, L., Vo, M., Lee, M., & Baik, S. (2018). A Cluster-Based Boosting Algorithm for Bankruptcy Prediction in a Highly Imbalanced Dataset. *Symmetry*, *10*(7), 250-250. https://doi.org/10.3390/sym10070250

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature.com*. https://doi.org/10.1038/nature14539

Lee, E., Rustam, F., Ashraf, I., Washington, P. B., Narra, M., & Shafique, R. (2022). Inquest of Current Situation in Afghanistan Under Taliban Rule Using Sentiment Analysis and Volume Analysis. *IEEE Access*, *10*, 10333-10348. https://doi.org/10.1109/ACCESS.2022.3144659

Lee, E., Rustam, F., Washington, P. B., Barakaz, F. E., Aljedaani, W., & Ashraf, I. (2022). Racism Detection by Analyzing Differential Opinions Through Sentiment Analysis of Tweets Using Stacked Ensemble GCR-NN Model. *IEEE Access*, *10*, 9717-9728. https://doi.org/10.1109/ACCESS.2022.3144266

Lee, H. S., Lee, H. R., Park, J. U., & Han, Y. S. (2018). An abusive text detection system based on enhanced abusive and non-abusive word lists. *Decision Support Systems*, *113*, 22-31. https://doi.org/10.1016/j.dss.2018.06.009

Li, B., Drozd, A., Liu, T., & Du, X. (2018). Subword-level Composition Functions for Learning Word Embeddings. 38-48. https://doi.org/10.18653/V1/W18-1205

LinkedIn. (2022). Erişim Tarihi 25 Mart 2022, https://www.linkedin.com/home

Liu, H., Alorainy, W., Burnap, P., & Williams, M. L. (2019). Fuzzy multi-task learning for hate speech type identification. *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, 3006-3012. https://doi.org/10.1145/3308558.3313546

Liu, X., & Qi, F. (2021). Research on advertising content recognition based on convolutional neural network and recurrent neural network. *International Journal of Computational Science and Engineering*, *24*(4), 398-404. https://doi.org/10.1504/IJCSE.2021.117022

Llugsi, R., Yacoubi, S. E., Fontaine, A., & Lupera, P. (2021). Comparison between Adam, AdaMax and Adam W optimizers to implement a Weather Forecast based on Neural Networks for the Andean city of Quito. *ETCM 2021 - 5th Ecuador Technical Chapters Meeting*. https://doi.org/10.1109/ETCM53643.2021.9590681

Maiya, A. S. (2020). ktrain: A Low-Code Library for Augmented Machine Learning. *arXiv*. http://arxiv.org/abs/2004.10703

Mancini, F., Sousa, F. S., Teixeira, F. O., Falcão, A. E. J., Hummel, A. D., da Costa, T. M., Calado, P. P., de Araújo, L. V., & Pisa, I. T. (2011). Use of Medical Subject Headings (MeSH) in Portuguese for categorizing web-based healthcare content. *Journal of Biomedical Informatics*, *44*(2), 299-309. https://doi.org/10.1016/j.jbi.2010.12.002

Mansoor, M., Rehman, Z. u., Shaheen, M., Khan, M. A., & Habib, M. (2020). Deep Learning based Semantic Similarity Detection using Text Data. *Information Technology and Control*, *49*(4), 495-510. https://doi.org/10.5755/J01.ITC.49.4.27118

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*, *27*, 16-32. https://doi.org/10.1016/J.COSREV.2017.10.002

Martins, R., Gomes, M., Almeida, J. J., Novais, P., & Henriques, P. (2018). Hate speech classification in social media using emotional analysis. *Proceedings - 2018 Brazilian Conference on Intelligent Systems, BRACIS 2018*, 61-66. https://doi.org/10.1109/BRACIS.2018.00019

Mehdad, Y., & Tetreault, J. (2016). Do Characters Abuse More Than Words? *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 13-15.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/1310.4546v1

Moraes, R., Valiati, J. F., & Gavião Neto, W. P. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*, *40*(2), 621-633. https://doi.org/10.1016/J.ESWA.2012.07.059

Morris, C., & Yang, J. J. (2021). A machine learning model pipeline for detecting wet pavement condition from live scenes of traffic cameras. *Machine Learning with Applications*, *5*, 100070-100070. https://doi.org/10.1016/J.MLWA.2021.100070

Mossie, Z., & Wang, J. H. (2020). Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, *57*(3), 102087-102087. https://doi.org/10.1016/J.IPM.2019.102087

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. *Studies in Computational Intelligence, 881 SCI*, 928-940. https://doi.org/10.1007/978-3-030-36687-2_77

Mujahid, M., Lee, E., Rustam, F., Washington, P. B., Ullah, S., Reshi, A. A., & Ashraf, I. (2021). Sentiment Analysis and Topic Modeling on Tweets about Online Education during COVID-19. *Applied Sciences 2021, Vol. 11, Page 8438*, *11*(18), 8438-8438. https://doi.org/10.3390/APP11188438

Oflazer, K., & Saraçlar, M. (2018). Turkish and Its Challenges for Language and Speech Processing. *Turkish Natural Language Processing Springer*, 1-19. https://doi.org/10.1007/978-3-319-90165-7_1

Omar, A., Mahmoud, T. M., & Abd-El-Hafeez, T. (2020). Comparative Performance of Machine Learning and Deep Learning Algorithms for Arabic Hate Speech Detection in OSNs. *Advances in Intelligent Systems and Computing*, *1153 AISC*, 247-257. https://doi.org/10.1007/978-3-030-44289-7_24

Omar, N., & Al-Tashi, Q. (2018). Arabic nested noun compound extraction based on linguistic features and statistical measures. *GEMA Online Journal of Language Studies*, *18*(2), 93-107. https://doi.org/10.17576/gema-2018-1802-07

Ornek, A. H., Ceylan, M., & Ervural, S. (2019). Health status detection of neonates using infrared thermography and deep convolutional neural networks. *Infrared Physics & Technology*, *103*. https://doi.org/10.1016/j.infrared.2019.103044

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., & Yeung, D.-Y. (2019). Multilingual and Multi-Aspect Hate Speech Analysis. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 4675-4684. https://doi.org/10.18653/V1/D19-1474

Oxford. (2022). Erişim Tarihi 26 Şubat 2022, https://www.oxfordlearnersdictionaries.com/

Özel, S. A., Akdemir, S., Saraç, E., & Aksu, H. (2017). Detection of cyberbullying on social media messages in Turkish. *International Conference on Computer Science and Engineering (UBMK)*, 366-370. https://doi.org/10.1109/UBMK.2017.8093411

Park, J. H., & Fung, P. (2017). One-step and Two-step Classification for Abusive Language Detection on Twitter. 41-45. https://doi.org/10.48550/arxiv.1706.01206

Patro, V. M., & Patra, M. R. (2015). A Novel Approach to Compute Confusion Matrix for Classification of n-Class Attributes with Feature Selection. *Transactions on Machine Learning and Artificial Intelligence*, *3*(2), 52-52. https://doi.org/10.14738/TMLAI.32.1108

Pelle, R. P. d., & Moreira, V. P. (2017). Offensive Comments in the Brazilian Web: a dataset and baseline results. *Anais do Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*. https://doi.org/10.5753/BRASNAM.2017.3260

Pewresearch. (2021). *Social Media Use in 2021 | Pew Research Center*. Erişim Tarihi 27 Eylül 2021, https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is Multilingual BERT? *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 4996-5001. https://doi.org/10.18653/v1/p19-1493

Pratiwi, N. I., Budi, I., & Alfina, I. (2019). Hate speech detection on Indonesian instagram comments using FastText approach. *2018 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2018*, 447-450. https://doi.org/10.1109/ICACSIS.2018.8618182

Priyoko, B., & Yaqin, A. (2019). Implementation of naive bayes algorithm for spam comments classification on Instagram. *2019 International Conference on Information and Communications Technology, ICOIACT 2019*, 508-513. https://doi.org/10.1109/ICOIACT46704.2019.8938575

Pronoza, E., Panicheva, P., Koltsova, O., & Rosso, P. (2021). Detecting ethnicity-targeted hate speech in Russian social media texts. *Information Processing & Management*, *58*(6), 102674-102674. https://doi.org/10.1016/J.IPM.2021.102674

Quandt, T., Klapproth, J., & Frischlich, L. (2022). Dark social media participation and well-being. *Current Opinion in Psychology*, *45*, 101284-101284. https://doi.org/10.1016/J.COPSYC.2021.11.004

Rokach, L., & Maimon, O. (2014). Data Mining with Decision Trees. *81*. https://doi.org/10.1142/9097 (Series in Machine Perception and Artificial Intelligence)

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. https://doi.org/10.17185/duepublico/42132

Sadiq, S., Mehmood, A., Ullah, S., Ahmad, M., Choi, G. S., & On, B. W. (2021). Aggression detection through deep neural model on Twitter. *Future Generation Computer Systems*, *114*, 120-129. https://doi.org/10.1016/J.FUTURE.2020.07.050

Sağlam, F. (2019). Otomatik Duygu Sözlüğü Geliştirilmesi ve Haberlerin Duygu Analizi. http://www.openaccess.hacettepe.edu.tr:8080/xmlui/handle/11655/9305

Salminen, J., Almerekhi, Milenković, M., Jung, S.-G., An, J., Kwak, H., & Jansen, B. J. (2018). Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. *Twelfth International AAAI Conference on Web and Social Media*. www.aaai.org

Salur, M. U. (2021). *Derin Öğrenme Tabanlı Çok Modlu Duygu Analizi Yöntemlerinin Geliştirilmesi* [Phd, Fırat Üniversitesi].

Saraç, E., & Özel, S. A. (2016). Effects of Feature Extraction and Classification Methods on Cyberbully Detection. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *21*(1), 190-190. https://doi.org/10.19113/sdufbed.20964

Sarıbacak, B. (2021). *Makine Öğrenmesi Sınıflandırma Yöntemleri İle Hematoloji Hastalıklarından Demir Eksikliği Anemisinin Erken Teşhis Edilmesi* [Phd, Ondokuz Mayıs Üniversitesi].

Saric, M., Dujmic, H., & Russo, M. (2015). Scene Text Extraction in IHLS Color Space Using Support Vector Machine. *Information Technology and Control*, *44*(1), 20-29. https://doi.org/10.5755/J01.ITC.44.1.5757

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47. www.ira.uka.de/bibliography/Ai/automated.text.

Segura-Bedmar, I., Colón-Ruíz, C., Tejedor-Alonso, M. Á., & Moro-Moro, M. (2018). Predicting of anaphylaxis in big data EMR by exploring machine learning approaches. *Journal of Biomedical Informatics*, *87*, 50-59. https://doi.org/10.1016/j.jbi.2018.09.012

Seiffert, C., Khoshgoftaar, T. M., Hulse, J. V., & Napolitano, A. (2008). Resampling or Reweighting: A Comparison of Boosting Implementations. *2008 20th IEEE International Conference on Tools with Artificial Intelligence*, *1*, 445-451. https://doi.org/10.1109/ICTAI.2008.59

Sennrich, R., Haddow, B., & Birch, A. (2015). Neural Machine Translation of Rare Words with Subword Units. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, *3*, 1715-1725. https://doi.org/10.48550/arxiv.1508.07909

Sharma, A., Kabra, A., & Jain, M. (2022). Ceasing hate with MoH: Hate Speech Detection in Hindi–English code-switched language. *Information Processing & Management*, *59*(1), 102760-102760. https://doi.org/10.1016/J.IPM.2021.102760

Sharma, S., & Shrivastava, M. (2018). Degree based Classification of Harmful Speech using Twitter Data. *arXiv preprint, arXiv:1806.04197*, 106-112.

Shi, H., Wang, H., Huang, Y., Zhao, L., Qin, C., & Liu, C. (2019). A hierarchical method based on weighted extreme gradient boosting in ECG heartbeat classification. *Computer Methods and Programs in Biomedicine*, *171*, 1-10. https://doi.org/10.1016/J.CMPB.2019.02.005

Shushkevich, E., & Cardiff, J. (2019). Automatic misogyny detection in social media: A survey. *Computacion y Sistemas*, *23*(4), 1159-1164. https://doi.org/10.13053/CyS-23-4-3299

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the Targets of Hate in Online Social Media. *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016*, 687-690. https://arxiv.org/abs/1603.07709v1

Simpson, A. J. R. (2015). Over-Sampling in a Deep Neural Network. *arXiv preprint arXiv:1502.03648* https://arxiv.org/abs/1502.03648v1

Smetanin, S., & Komarov, M. (2021). Deep transfer learning baselines for sentiment analysis in Russian. *Information Processing & Management*, *58*(3), 102484-102484. https://doi.org/10.1016/J.IPM.2020.102484

Snapchat. (2022). Erişim Tarihi 15 Şubat 2022, https://accounts.snapchat.com/accounts/login?continue=%2Faccounts%2Fwelcome

Sowinski-Mydlarz, V., Li, J., Ouazzane, K., & Vassilev, V. (2021). Threat intelligence using machine learning packet dissection. *Transactions on Computational Science and Computational Intelligence*. http://repository.londonmet.ac.uk/id/eprint/6919

Stappen, L., Brunn, F., & Schuller, B. (2020). Cross-lingual Zero- and Few-shot Hate Speech Detection Utilising Frozen Transformer Language Models and AXEL. *arXiv preprint arXiv:2004.13850.* https://arxiv.org/abs/2004.13850v1

Statista. (2020). *Turkey: number of Instagram users 2020 | Statista*. Erişim Tarihi 9 Temmuz 2020, https://www.statista.com/statistics/1024714/instagram-users-turkey/

Staudemeyer, R. C., & Morris, E. R. (2019). Understanding LSTM-a tutorial into Long Short-Term Memory Recurrent Neural Networks. *arXiv preprint arXiv:1909.09586*.

Subrahmanian, V. S., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A., Menczer, F., Stevens, A., Dekhtyar, A., Gao, S., Hogg, T., Kooti, F., Liu, Y., Varol, O., Shiralkar, P., Vydiswaran, V., . . . Hwang, T. (2016). The DARPA Twitter Bot Challenge. *Computer*, *49*(6), 38-46. https://doi.org/10.1109/MC.2016.183

Sun, Y. P., Wang, Y. F., & Zheng, X. J. (2009). Analysis the height of water conducted zone of coal seam roof based on GA-SVR. *Journal of China Coal Society*. http://en.cnki.com.cn/Article_en/CJFDTotal-MTXB200912007.htm

Talukder, A., & Ahammed, B. (2020). Machine Learning Algorithms for Predicting Malnutrition among Under-Five Children in Bangladesh. *Nutrition*, *78*, 110861. https://doi.org/10.1016/j.nut.2020.110861

Tang, Y., & Dalzell, N. (2019). Classifying Hate Speech Using a Two-Layer Model. *Statistics and Public Policy*, *6*(1), 80-86. https://doi.org/10.1080/2330443x.2019.1660285

Tashtoush, Y. M., & Al Aziz Orabi, D. A. (2019). Tweets Emotion Prediction by Using Fuzzy Logic System. *2019 6th International Conference on Social Networks Analysis, Management and Security, SNAMS 2019*, 83-90. https://doi.org/10.1109/SNAMS.2019.8931878

TDK. (2020). Erişim Tarihi 9 Temmuz 2020, http://www.tdk.gov.tr/

Terragni, S., Fersini, E., & Messina, E. (2020). Constrained Relational Topic Models. *Information Sciences*, *512*, 581-594. https://doi.org/10.1016/j.ins.2019.09.039

Thejas, G. S., Kumar, K., Iyengar, S. S., Badrinath, P., & Sunitha, N. R. (2019). AI-NLP Analytics: A thorough Comparative Investigation on India-USA Universities Branding on the Trending Social Media Platform 'Instagram'. *CSITSS 2019 - 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution, Proceedings*. https://doi.org/10.1109/CSITSS47250.2019.9031050

TikTok. (2022). Erişim Tarihi 9 Şubat 2022, https://www.tiktok.com/tr-TR/

Twitter. (2022). Erişim Tarihi 3 Ocak 2022, https://twitter.com/home?lang=tr

Van Royen, K., Poels, K., Daelemans, W., & Vandebosch, H. (2015). Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. *Telematics and Informatics*, *32*(1), 89-97. https://doi.org/10.1016/j.tele.2014.04.002

Vigna, F. D., Cimino, A., Dell'orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook. *In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, 86-95. https://curl.haxx.se

Wadhwa, P., & Bhatia, M. P. S. (2014). *Classification of Radical Messages on Twitter Using Security Associations*. Auerbach Publications. https://doi.org/10.1201/B17352-18

Wang, N., Wang, P., & Zhang, B. (2010). An improved TF-IDF weights function based on information theory. *International Conference on Computer and Communication Technologies in Agriculture Engineering*, *3*, 439-441. https://doi.org/10.1109/CCTAE.2010.5544382

Wang, X., Gerber, M. S., & Brown, D. E. (2012). Automatic Crime Prediction Using Events Extracted from Twitter Posts. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7227 LNCS*, 231-238. https://doi.org/10.1007/978-3-642-29047-3_28

Warner, W., & Hirschberg, J. (2012). Detecting Hate Speech on the World Wide Web. *Proceedings of the second workshop on language in social media*, 19-26. https://aclanthology.org/W12-2103

Waseem, Z. (2016). Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. 138-142. www.spacy.io

Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. *In Proceedings of the NAACL student research workshop*, 88-93. https://doi.org/10.18653/V1/N16-2013

Waseem, Z., Thorne, J., & Bingel, J. (2018). Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection. In (pp. 29-55). Springer, Cham. https://doi.org/10.1007/978-3-319-78583-7_3

Wei, J., Liao, J., Yang, Z., Wang, S., & Zhao, Q. (2020). BiLSTM with Multi-Polarity Orthogonal Attention for Implicit Sentiment Analysis. *Neurocomputing*, *383*, 165-173. https://doi.org/10.1016/j.neucom.2019.11.054

WhatsApp. (2022). Erişim Tarihi 9 Şubat 2022, https://www.whatsapp.com/?lang=tr

Whisper. (2021). Erişim Tarihi 30 Eylül 2021, http://whisper.sh/

Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018). Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language.

Wu, S., & Dredze, M. (2020). Are All Languages Created Equal in Multilingual BERT? *arXiv*, 120-130. https://doi.org/10.18653/v1/2020.repl4nlp-1.16

Wulczyn, E., Thain, N., & Dixon, L. (2016). Ex Machina: Personal Attacks Seen at Scale. *26th International World Wide Web Conference, WWW 2017*, 1391-1399. https://arxiv.org/abs/1610.08914v2

Xiang, Z. L., Yu, X. R., Hui, A. W. M., & Kang, D. K. (2014). Novel Naive Bayes based on Attribute Weighting in Kernel Density Estimation. *2014 Joint 7th International Conference on Soft Computing and Intelligent Systems, SCIS 2014 and 15th International Symposium on Advanced Intelligent Systems, ISIS 2014*, 1439-1442. https://doi.org/10.1109/SCIS-ISIS.2014.7044787

Xie, R., Chu, S. K. W., Chiu, D. K. W., & Wang, Y. (2021). Exploring Public Response to COVID-19 on Weibo with LDA Topic Modeling and Sentiment Analysis. *Data and Information Management*, *5*(1), 86-99. https://doi.org/10.2478/DIM-2020-0023

Yan, Y. (2016). Hierarchical Classification with Convolutional Neural Networks for Biomedical Literature. *International Journal of Computer Science*. https://search.proquest.com/openview/e61e7d0d21e0f1973aad42c6d997f9ec/1?pq-origsite=gscholar&cbl=2044552

Yoo, S. Y., & Jeong, O. R. (2020). Automating the expansion of a knowledge graph. *Expert Systems with Applications*, *141*, 112965-112965. https://doi.org/10.1016/J.ESWA.2019.112965

YouTube. (2022). *YouTube* , Erişim Tarihi 3 Ocak 2022, https://www.youtube.com/

Yuan, B., & Ma, X. (2012). Sampling + reweighting: Boosting the performance of AdaBoost on imbalanced datasets. *Proceedings of the International Joint Conference on Neural Networks*. https://doi.org/10.1109/IJCNN.2012.6252738

Zhang, H., Wojatzki, M., Horsmann, T., & Zesch, T. (2019). ltl.uni-due at SemEval-2019 Task 5: Simple but Effective Lexico-Semantic Features for Detecting Hate Speech in Twitter. *Proceedings of the 13th International Workshop on Semantic Evaluation*, 441-446. https://doi.org/10.18653/V1/S19-2078

Zhang, Z., & Luo, L. (2018). Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter. *Semantic Web*, *10*(5), 925-945. https://arxiv.org/abs/1803.03662v2

Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. *European semantic web conference*, *10843 LNCS*, 745-760. https://link.springer.com/chapter/10.1007/978-3-319-93417-4_48

Zin, H. M., Mustapha, N., Murad, M. A. A., & Sharef, N. M. (2018). Term Weighting Scheme Effect in Sentiment Analysis of Online Movie Reviews. *Advanced Science Letters*, *24*(2), 933-937. https://doi.org/10.1166/asl.2018.10661