T.C.
KONYA TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ YÖNTEMLERİ KULLANILARAK UZAKTAN EĞİTİM KONULU TÜRKÇE TWEETLERİN DUYGU ANALİZİ

Ali Can AKDENİZ

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Nisan-2022
KONYA

TEZ KABUL VE ONAYI

Ali Can AKDENİZ tarafından hazırlanan "Makine Öğrenmesi Yöntemleri Kullanılarak Uzaktan Eğitim Konulu Türkçe Tweetlerin Duygu Analizi" adlı tez çalışması 14/04/2022 tarihinde aşağıdaki jüri tarafından oy birliği ile Konya Teknik Üniversitesi Lisansüstü Eğitim Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda YÜKSEK LİSANS olarak kabul edilmiştir.

Jüri Üyeleri **İmza**

Başkan
Doç. Dr. Mehmet HACIBEYOĞLU

Danışman
Doç. Dr. İsmail BABAOĞLU

Üye
Dr. Öğr. Üyesi Sait Ali UYMAZ

Yukarıdaki sonucu onaylarım.

Prof. Dr. Saadettin Erhan KESEN
Enstitü Müdürü

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Ali Can AKDENİZ

Tarih: 14.04.2022

ÖZET

YÜKSEK LİSANS

MAKİNE ÖĞRENMESİ YÖNTEMLERİ KULLANILARAK UZAKTAN EĞİTİM KONULU TÜRKÇE TWEETLERİN DUYGU ANALİZİ

Ali Can AKDENİZ

Konya Teknik Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Doç. Dr. İsmail BABAOĞLU

2022, 90 Sayfa

Jüri
Doç. Dr. İsmail BABAOĞLU
Doç. Dr. Mehmet HACIBEYOĞLU
Dr. Öğr. Üyesi Sait Ali UYMAZ

Teknolojinin gelişmesi beraberinde sosyal medya platformlarının da gelişerek büyük kullanıcı kitlelerine ulaşmasına yol açmıştır. Kişiler sosyal medya platformları kullanarak diğer kişilerle iletişim kurabildiği gibi, meydana gelen toplumsal olaylar karşısında, bir ürün ya da bir konu hakkında ortak bir başlıkta bu platformlarda bir araya gelerek duygu ve düşüncelerini paylaşabilmektedir. Bu paylaşımlar duygu analizi çalışmaları için birçok alanda kullanılabilir büyük bir veri kaynağı oluşturmaktadır. Duygu analizi çalışmaları ile bu veriler işlenip analiz edilerek, ilgili konu hakkında olumlu, olumsuz veya tarafsız duygu ifadeleri belirlenebilmektedir. 2020 yılının ocak ayında başlayan ve tüm dünyayı etkisi altına alan korona virüs salgını ile ülke genelinde birtakım tedbirler alınmaya başlanmış, bu tedbirler kapsamında da Mart 2020'den itibaren uzaktan eğitim sürecine geçilmiştir. Bu çalışmada sosyal medya platformu Twitter'da paylaşılan uzaktan eğitim konulu Türkçe tweetler elde edilerek veri ön işleme tabi tutulmuş ve ayrıca zemberek kütüphanesi ile normalleştirilerek işlenebilir bir hale getirilmiştir. Sınıflandırma aşamasında girdi olarak kullanılacak veri seti için manuel etiketleme işleminin yanı sıra farklı bir yaklaşımla dil çeviri işlemi yapılarak İngilizce dilinde doğrudan duygu çıktıları üreten TextBlob, Vader ve Bert ile modeller oluşturulmuştur. Bu modeller farklı sayısallaştırma yöntemleri (BoW, TF-IDF, Word2Vec,) ve farklı makine öğrenmesi algoritmaları (LR, SGD, SVM, RF, NB) ile kullanılarak en iyi performansı gösteren sınıflandırma modeli üzerinden yapılan paylaşımların duygu analizi gerçekleştirilmiştir. Türkçe metinlerin manuel etikete sahip olduğu yapıda en iyi TF-IDF – LR ikilisi ile 0.79'luk bir sınıflandırma başarısı elde edilmiştir. Manuel yöntemle etiketlendirilen tarafsız olarak işaretlenmiş metinler veri setinden çıkarıldığında başarı oranının arttığı ve BoW – LR ikilisinin 0.84'lük oranla en iyi sonucu verdiği görülmüştür. Dil çeviri işlemi ile hazır modeller tarafından etiketlenerek oluşturulan modellerde Türkçe metinler için istenilen seviyede bir başarı elde edilememiştir.

Anahtar Kelimeler: Doğal Dil İşleme, Duygu Analizi, Makine Öğrenmesi, Uzaktan Eğitim, Twitter

ABSTRACT

MS THESIS

TURKISH TWEETS ON DISTANCE EDUCATION USING MACHINE LEARNING METHODS SENTIMENT ANALYSIS

Ali Can AKDENİZ

Konya Technical University
Institute of Graduate Studies
Department of Computer Engineering

Advisor: Assoc. Prof. Dr. İsmail BABAOĞLU

2022, 90 Pages

Jury
Assoc. Prof. Dr. İsmail BABAOĞLU
Assoc. Prof. Dr. Mehmet HACIBEYOĞLU
Asst. Prof. Dr. Sait Ali UYMAZ

The development of technology has led to the development of social media platforms and reaching large user masses. People can communicate with other people by using social media platforms, and they can come together on these platforms to share their feelings and thoughts on a common topic about a product or a topic, in the face of social events that occur. These shares constitute a large data source that can be used in many fields for sentiment analysis studies. With sentiment analysis studies, these data can be processed and analyzed, and positive, negative or neutral emotional expressions about the relevant subject can be determined. With the corona virus epidemic, which started in January 2020 and affected the whole world, some measures were taken across the country, and within the scope of these measures, the distance education process started in March 2020. In this study, Turkish tweets on distance education shared on the social media platform Twitter were obtained and the data were pre-processed and also normalized with the zemberek library and made processable. For the data set to be used as input in the classification phase, besides manual labeling, models were created with TextBlob, Vader and Bert, which directly produce English emotion outputs by making language translation with a different approach. For the data set to be used as input in the classification phase, besides the manual labeling process, a different approach was brought with language translation and models were created with TextBlob, Vader and Bert, which directly produce English emotion outputs. These models were used with different digitization methods (BoW, TF-IDF, Word2Vec,) and different machine learning algorithms (LR, SGD, SVM, RF, NB) and sentiment analysis of the shares made on the best performing classification model was performed. In the structure where Turkish texts have manual tags, 0.79 classification success was achieved with the best TF-IDF – LR pair. When the texts labeled with the manual method and marked as neutral were removed from the data set, it was seen that the success rate increased and the BoW – LR pair gave the best result with a ratio of 0.84. In the models created by labeling ready-made models with the language translation process, the desired level of success for Turkish texts was not achieved.

Keywords: Distance Education, Machine learning, Natural Language Processing, Sentiment Analysis, Twitter

ÖNSÖZ

Bu çalışma sürecinde bilgi ve deneyimlerinden faydalandığım, beni destekleyip ve yönlendiren danışmanım Doç. Dr. İsmail BABAOĞLU'na, çalışmalarım sırasında yardım ve desteklerini esirgemeyen iş arkadaşlarıma ve her koşulda yanımda olan eşim Gülsüm AKDENİZ'e teşekkürlerimi sunarım.

Ali Can AKDENİZ
KONYA-2022

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

Simgeler

@: Twitter kullanıcı adı öneki
#: Hashtag

Kısaltmalar

API: Application Programming Interface
Bert: Bidirectional Encoder Representations from Transformers
BoW: Bag of Words
CBoW: Continous Bag of Words
DDİ: Doğal Dil İşleme
GPU: Graphic Processing Unit
LR: Lojistik Regresyon
NB: Naive Bayes
NLTK: Natural Language Toolkit
RF: Random Forest
SGD: Stochastic Gradient Descent
SVM: Support Vector Machine
TF-IDF: Term Frequency-Inverse Document Frequency
TPU: Tensor Processing Unit
Vader: Valence Aware Dictionary and Sentiment Reasoner

1. GİRİŞ

Teknolojinin her geçen gün gelişmesiyle, internet ortamı ve buna bağlı olarak sosyal medya platformları da çeşitlilik ve gelişme göstermiştir. Bu anlamda büyük bir kullanıcı kitlesine sahip Twitter sadece bir iletişim aracı olmanın yanı sıra sağlık, siyaset, spor, teknoloji, eğitim, e-ticaret, sinema gibi alanlarda da belirli bir ürün ya da konu ile ilgili insanların görüşlerini paylaştığı popüler bir sosyal medya platformu ve önemli bir veri kaynağı haline gelmiştir. Ortaya çıkan bu büyük veri beraberinde birçok firmanın satış ve pazarlama stratejisi, siyasi partilerin seçmenler hakkında fikir edinmesi, sinema sektöründe bir filmin beğeni düzeyi gibi kullanıcı yorum ve görüşlerinden bilgi edinilmesinde duygu analizi çalışmalarını önemli bir noktaya getirmiştir. Bahsedilen duygu analizi çalışmalarında temel amaç bir ürün veya konu hakkında yapılmış olan paylaşımın olumlu, olumsuz ya da tarafsız olup olmadığı sonucuna varabilmektir.

Dünya veya ülke genelinde meydana gelen toplumsal olaylar karşısında kişiler sosyal medya platformlarında bir araya gelerek ortak bir başlık altında duygu ve düşüncelerini paylaşabilmektedir. Covid-19 yani tam adıyla koronavirüs 2019 yılında, solunum yolu hastalıklarına neden olan, öksürük, ateş ve nefes darlığı gibi belirtiler gösteren bir virüstür. Bu virüs ilk olarak 2019 yılının aralık ayında Çin'in Wuhan kentinde görülmüştür. Virüsün nefes ve hava yoluyla bulaşıyor olması insandan insana geçişini hızlandırmış, diğer şehirlere ve ülkelere sıçramasına sebep olmuştur. Bu yayılımlar sonrasında Dünya Sağlık Örgütü tarafından korona virüs coğrafi salgın (pandemi) olarak belirlenmiştir. Bu salgına karşı korunabilmek için ülkeler birtakım tedbirler almıştır. Bunların başında sağlık, ekonomi ve eğitim alanları gelmektedir (Sariman ve Mutaf, 2020). Türkiye de Mart 2020 tarihinde bu tedbirleri alan ülkeler arasındadır. Tedbirler kapsamında tüm eğitim öğretim faaliyetleri durdurularak zamandan ve mekândan tamamen bağımsız bir şekilde öğrencinin ve öğretim üyesinin okula gelme zorunluluğu olmaksızın internet aracılığı ile sanal ortamda canlı, görüntülü, sesli olarak derslerin işlendiği, katılımcının istediği zaman bunları tekrar izleyebileceği bir eğitim sistemi olan uzaktan eğitim sürecine geçilmiştir.

Dünyayı etkisi altına alan ve 2022 yılı itibarıyla hala devam etmekte olan Covid-19 salgını sürecinde sosyal medya platformlarında özellikle Twitter'da, uzaktan eğitim konu başlığı altında birçok paylaşım yer almaktadır. Bu doğrultuda Twitter'da yer alan

uzaktan eğitim konulu Türkçe paylaşımlar tez kapsamında ele alınarak makine öğrenmesi algoritmaları ile duygu analizi gerçekleştirilmiştir.

Duygu analizi uzun zamandır çalışmalara konu olan bir alan olmuştur. Bu anlamda çalışmanın ikinci bölümünde daha önce yapılan ve literatürde yer almış Türkçe ve İngilizce çalışmalardan bahsedilmiştir.

Üçüncü bölümde çalışma kapsamında kullanılan yöntem ve analiz sürecinin adımları anlatılmış ayrıca Twitter ve Twitter verilerine erişimde kullanılan Twitter uygulama programlama arayüzü (Application Programming Interface-API) hakkında bilgilere yer verilmiştir.

Dördüncü bölümde deneysel çalışmalar ile sistem mimarisi üzerinden uygulanan adımlar tek tek ele alınmıştır. Twitter'dan elde edilen veriler literatürde yer alan birçok veri ön işleme adımlarına tabi tutularak işlenebilir bir hale getirilmiştir. Ayrıca veri ön işleme aşamasında java programla dili ile geliştirilmiş zemberek kütüphanesinin python programlama diline uyarlanmış versiyonu ile veri setinde yer alan Türkçe kelimeler için yazım denetimi ve kelime tahmini ile normalleştirme işlemine bu bölümde yer verilmiştir.

Verilerin etiketlenmesi için manuel etiketleme, sözlük oluşturarak etiketleme gibi literatürde farklı yöntemler bulunmaktadır. Verilerin etiketleme aşaması veri setinin büyük olduğu durumlarda hız ve zaman açısından zahmetli olabilmektedir. Bu anlamda İngilizce dili için geliştirilmiş ve literatürde sıkça kullanılan birçok hazır model bulunmaktadır. Çalışma kapsamında sınıflandırma aşamasında girdi olarak kullanılacak veri seti için manuel etiketleme işlemine ek olarak farklı bir yaklaşımla Türkçeden İngilizceye dil çeviri işlemi yapılmış ardından İngilizce metinler üzerinden hazır modeller ile etiketleme işlemi gerçekleştirilmiştir. Manuel etiketleme işlemine kıyasla hazır modellerin etiketleme başarısına yine bu bölümde yer verilmiştir. İngilizce içerikli metinlerde duygu çıktıları veren hazır modeller olduğu gibi Türkçe metinler için geliştirilmiş ve doğrudan duygu çıktıları üreten Bidirectional Encoder Representations from Transformers (Bert) tabanlı modele de çalışma içinde yer verilmiştir.

Veri seti içerisinde metin olarak yer alan verilerin sayısal olarak ifade edilmesi gerekmektedir. Bu doğrultuda kullanılan farklı sayısallaştırma yöntemlerinin sınıflandırma performansına etkisi yine dördüncü bölüm içerisinde ele alınmıştır.

Dördüncü bölümün devamında farklı etiketleme yöntemlerine sahip veri setleri ile oluşturulan modeller farklı sayısallaştırma yöntemleri ile temsil edilerek makine öğrenmesi algoritmalarına girdi olarak kullanılmış ve en iyi performans gösteren

sınıflandırma modeline karar verilmiştir. Duygu analizi kapsamında kelime bulutu ve kelime ağacı diyagramları ile uzaktan eğitim konusunda daha çok nelerin konuşulduğu gibi durumlara dördüncü bölümün sonunda yer verilmiştir.

2. KAYNAK ARAŞTIRMASI

Yerli ve yabancı literatür incelendiğinde sosyal medya üzerindeki verilerden yararlanılarak yapılan birçok çalışmaya rastlanmıştır. Bu kısımda makine öğrenmesi ile Twitter verilerinin duygu analizi konusu ile ilgili bazı literatür çalışmalarına yer verilmiştir.

2021 yılında Kumaş, çalışmasında Türkçe Twitter verilerine metin madenciliği yöntemleri uygulayarak duygu analizi yapmıştır. Naive Bayes (NB), KNN, Destek Vektör Makinesi, Lojistik Regresyon (LR) ve Karar Ağacı sınıflandırma algoritmalarını kullanarak pozitif ve negatif olarak sınıflandırmıştır. Sınıflandırma sonuçlarını f1 skoru ile değerlendirmiş ve sırasıyla NB, KNN, Destek Vektör Makineleri, LR ve Karar Ağacı sınıflandırıcılarıyla elde ettiği f1 skorları %70, %65, %73, %71 ve %69 olarak bulmuştur. Köksal ve ark. (2021) Twitter kullanıcılarının Bitcoin ile ilgili paylaşımlarını derleyerek duygu analizi çalışması gerçekleştirmişlerdir. NB ve LR algoritmalarını kullanarak bu algoritmaların başarı oranlarını karşılaştırmış, NB'nin duyguları tahmin etme üzerindeki başarı oranını %72,19, LR'nin %75,53 olarak yakalamışlardır. Çalışmanın ikinci kısmında "Bitcoin" anahtar kelimesi ile yapılan paylaşımlardan pozitif tweet oranına bakılarak Bitcoin günlük açılış değeri ile birlikte kapanış değerini tahmin etmeye çalışmışlardır. Doğrusal Regresyon ve Rastgele Orman Regresyon yöntemleri ile sırasıyla r^2 değeri %88,97 ve %94,16 ulaşmışlardır. Aynı yıl Özyurt ve Kısa pandemi sürecinde uzaktan eğitim ile ilgili Twitter paylaşımlarında duygu analizi gerçekleştirmişlerdir. Kaggle veri paylaşım platformundan elde ettikleri verilerden rastgele 999 kaydı manuel olarak pozitif ve negatif olarak etiketlendirmişlerdir. Veri seti KNIME üzerinde kurulan model ile uygun düğümler kurarak önce ön işleme ardından duygu analizi aşamalarından geçirilerek başarı hesaplaması yapmışlardır. Temelde sözlük tabanlı yaklaşımı esas aldıkları çalışmada %88,4 oranında başarıya ulaşmışlardır. Yine aynı yıl Yılmaz ve Orman yaptığı çalışmada korona virüs pandemisi ile ilgili Twitter'da paylaşılan birtakım terimler göz önüne alınmış ve duygu analizi çalışmasını gerçekleştirmişlerdir. Konu ile ilgili bir takım Türkçe paylaşımlar üzerinden bu paylaşımların olumlu ya da olumsuz duygu analizini gerçekleştirmişlerdir. Geliştirdikleri Uzun Kısa Süreli Hafıza derin öğrenme yöntemi ile maksimum %97 doğruluk başarısı elde etmişlerdir.

2020 yılında Sarıman ve Mutaf, çalışmalarında, 11 Mart 2020 tarihinden itibaren Türkiye'de kullanıcıların Twitter aracılığı ile yayınladığı korona virüs için konuşulan

önemli başlıkları veri olarak kullanarak duygu analizi yapmışlardır. Duygu analizi yöntemine göre beş başlık altında topladıkları konular için olumlu ve olumsuz yorumları sınıflandırarak genel bakış çıkarmış ve sonrasında haftalık süreçte bu konular hakkında gözle görülebilir bir değişiklik olup olmadığını analiz etmişlerdir. Çalışmanın sonucunda insanların maske uygulamasının genelde olumlu olarak değerlendirildiğini fakat diğer uygulamaların ise genelde olumsuz olarak değerlendirildiğini tespit etmişlerdir. Emekli ve Selvi (2020), GSM Operatörlerine Yönelik Atılan Türkçe tweetlerin Derin Öğrenme Yöntemleriyle Duygu Analizi isimli çalışmalarında, Türkiye'nin önde gelen GSM operatörlerine yönelik paylaşılan Türkçe tweetlerin olumlu ve olumsuz olarak sınıflandırılmasını incelemişlerdir. Çalışma kapsamında tweetlerin sınıflandırılması için derin öğrenme yöntemlerini önererek kullandıkları Twitter verilerini, ilk olarak Doğal Dil İşleme (DDİ) yöntemleri ile hazırladıktan sonra derin öğrenme yöntemlerinden Evrişimli Sinir Ağları, Tekrarlayan Sinir Ağları ve Uzun Kısa Süreli Bellek modelleri ile sınıflandırarak karşılaştırmışlardır. Yine aynı yıl İlhan ve Sağaltıcı ise, Twitter verileri üzerinde çeşitli makine öğrenme teknikleri kullanarak bir duygu analizi çalışması yapmışlardır. N-gram metodu ile olumlu ve olumsuz duyguların analizini yaparak NB ve Destek Vektör Makineleri yöntemleri kullanmış ve ilgili sınıflandırıcıların performans karşılaştırmalarını yapmışlardır. Çalışma sonucunda, en yüksek değerin Destek Vektör Makineleri sınıflandırıcısına ait olduğunu tespit etmişlerdir.

2019 yılında, Bilgin ve Şentürk, Türkçe ve İngilizce tweetler üzerinde duygusal sınıflandırmayı araştırmışlardır. Çalışmada doküman vektörlerini (Doc2Vec) kullanarak hem DBoW ve DM gibi iki farklı doküman vektörü yönteminin çalışması hem de Yarı Danışmanlı ve Danışmanlı öğrenmenin etkileri incelemişlerdir. Çalışma sonuçlarını doğruluk, kesinlik, anma, özgünlük ve F-ölçütü metrikleri ile raporlamış, çalışma sonucunda da Yarı Danışmanlı öğrenme yöntemi Türkçe ve İngilizce veri kümelerinde Danışmanlı öğrenmeye göre daha başarılı sonuçlara ulaşmışlardır. Aynı yıl Ayan ve arkadaşları, Twitter üzerindeki tweetlerin İslamofobik olup olmadığını duygu analizi ile tespit etmeye çalışmışlardır. Çalışmalarında Lineer ridge regresyonu ve NB Sınıflandırıcı ile eğitilen modeller üzerinden precision, Recall, F1 ölçütlerinde hesaplamalarda bulunmuşlardır. Çalışmanın sonucunda pozitif tweetler için Ridge modelinde NB sınıflandırıcıya göre daha iyi sonuçlar elde edilirken Ridge Regresyonunda %96,3, NB Sınıflandırıcıda %95,3 oranında doğru sonuca ulaşmışlardır. Alharbi ve Doncker (2019) tarafından yapılan çalışma kapsamında da derin öğrenme

modellerine değinilerek sadece tweetlere değil, kullanıcı davranışlarını da dikkate alan bir Evrişimli Sinir Ağı kullanılmış ve başarımı gözlemlenmiştir. Çalışma sonucu incelendiğinde yaklaşık %88'lik başarım elde edildiği görülmüştür. Shehu ve arkadaşları (2019) ise çalışmalarında Türkçe tweetler üzerinde duygu analizi için Destek Vektör Makinesi ve Rastgele Orman algoritmalarını kullanmışlardır. Çalışmalarında veri seti üzerinde Zemberek kütüphanesini kullanarak veri ön işleme aşamasından geçirdikleri ve çalışma sonucunda da %88,5'lik bir başarım oranı ile Rasgele Orman algoritması ile daha başarılı bir sonuca ulaşmışlardır.

2018 yılında Kshirsagar ve arkadaşları çalışmalarında, Twitter üzerindeki nefret konuşmalarını makine öğrenmesi yöntemiyle tahmin etmeye çalışmışlardır. Bu çalışma kapsamında altmış bine yakın tweet kullanarak cinsiyetçi söylem için %76, ırkçı söylem için %78 ve diğer nefret söylemlerinin dahil olduğu ana küme için %86 oranıyla F1 skoru elde etmişlerdir. Yine aynı yıl Yelmen ve arkadaşları çalışmalarında üç farklı özellik seçim ve makine öğrenmesi algoritmasını hibrit şekilde kullanarak GSM operatörlerine yönelik atılan Türkçe tweetlerin sınıflandırılmasındaki başarımlarının ölçülmesini hedeflemişlerdir. Çalışma kapsamında veri seti oluşturma aşamasında DDİ yöntemlerinden faydalanarak tweet içeriğini temizlemişler, kelime düzeltme işleminden sonra kelime köklerini bularak tekrarlanan kelimeleri ve cümleleri veri setinden temizlemişlerdir. Çalışmalarında makine öğrenmesi temelli bir model gerçekleştirdikleri için özellik seçiminde farklı algoritmalar kullanarak 200 ayırt edici özellik seçimi yapmışlar ve özellik seçiminde Genetik Algoritma ve sınıflandırmada Destek Vektör Makinelerinin hibrit şekilde kullanımında yüksek başarım elde etmişlerdir.

2017 yılında Ayata ve arkadaşları dört farklı sektör ve konulardaki Türkçe tweetler üzerinde bir duygu analizi çalışması gerçekleştirmişlerdir. Çalışmalarında Karar Destek Makinesi ve Rastgele Orman sınıflandırıcı makine öğrenmesi algoritmalarını kullanmışlar ve çalışma sonunda Karar Destek Makinesinin daha başarılı sonuç verdiğini ortaya koymuşlardır. Aynı yıl Onan (2017) çalışmasında Türkçe Twitter mesajlarının sınıflandırılmasında, NB algoritması, destek vektör makineleri, LR makine öğrenmesi sınıflandırıcılarını kullanmıştır. Metin temsilinde, 1-gram, 2-gram ve 3-gram olmak üzere farklı öznitelik temsili ve bu öznitelik temsilleri ile elde edilen farklı öznitelik setlerini incelemiştir. Çalışma sonucunda, %77,78 oranla en yüksek başarımın, veri seti 1-gram ve 2-gram öznitelik setlerinin birleştirilmesi ile oluşturulan öznitelik seti ile temsil edildiğini ve sınıflandırma algoritması olarak da NB algoritması kullanıldığını ortaya koymuştur. Bu çalışmada yalnızca N-gram modelleri ile elde

edilmiş olan öznitelik setleri değerlendirmiştir. Wehrmann ve arkadaşları (2017) ise yaptıkları çalışmada daha önceki çalışmalardan farklı bir yöntem geliştirerek önceki çalışmaların makine öğrenmesi temelli ve tek dil bazlı olmasını eleştirmişlerdir. Bu eleştiriye istinaden çalışmalarında derin öğrenme yöntemlerinden faydalanmış ve dört farklı dilde model eğitimi gerçekleştirmişlerdir. Kullandıkları Conv-Char-S derin öğrenme yöntemi ile %72 doğruluk oranı, %75 F-ölçümüne ulaşmışlardır.

2016 yılında Yelmen, günlük konuşma dili ile yazılan Türkçe metinlerden öznitelik seçimine odaklandığı bir çalışma gerçekleştirmiştir. Bu çalışmada detaylı ön işlemeden geçen veri üzerinde destek vektör makineleri, yapay sinir ağları ve centroid tabanlı sınıflandırma algoritmalarını kullanmıştır. 3 ayrı GSM operatörünün takipçilerine ait tweetler üzerinde Gini İndeks, Bilgi Kazancı ve Genetik Algoritma olarak 3 farklı sınıflandırma algoritmasını hibrit olarak kullanmıştır. Çalışma sonucunda genetik algoritma ile destek vektör makineleri hibrit olarak kullanıldığında 3 farklı GSM operatörü için de %100 başarı elde edildiğini ortaya koymuştur. Aynı yıl Kaynar ve arkadaşları tarafından yapılan çalışmada, film yorumları üzerinden duygu analizi incelemesi gerçekleştirilmiştir. Kaynar ve arkadaşları çalışmalarında kullandıkları makine öğrenmesi yöntemleri ile başarımı ölçmeyi hedefleyerek Yapay Sinir Ağları ve Destek Vektör Makinelerinde yüksek başarıma ulaşmışlardır.

2015 yılında Akgül ve arkadaşları gerçekleştirdikleri çalışmada, Türkçe tweetler üzerinde duygu analizi incelemişlerdir. Bu anlamda belirli etiketlerle topladıkları tweetler üzerinde ön işlemler uygulayarak ön işlem aşamasında tweetler içerisinde, karakter ve kelimelerin temizlenmesini gerçekleştirerek sonuçların başarımını ölçmek için F-Ölçüm metriğini kullanmışlardır. Uyguladıkları n-gram yöntemleri ile %70 civarı bir başarım elde etmişlerdir. Çoban ve arkadaşları (2015) ise, çalışmalarında Türkçe Twitter mesajlardan oluşturulan veri setini metin sınıflandırma yöntemleri ile analiz ederek olumlu veya olumsuz olup olmadığını incelenmişlerdir. Çalışma sonuçlarını SVM, NB, Multinominal NB ve KNN algoritmalarıyla elde etmişlerdir. Çalışmada Vector Space model ile temsil edilen özniteliklere, kelime torbası (Bag of Words- BoW) ve N-Gram model olmak üzere iki farklı şekilde ulaşılmış ve bu durumun sınıflandırma sonuçlarına olan etkisi araştırılmıştır. Elde edilen sonuçlara göre duygu analizi çalışmaları, bir metin sınıflandırma problemi olarak ele alınabilir. Ayrıca başarı oranının yükseltilebilmesi için harici yöntemler uygulanması gerekmektedir. Bu çalışma ile Twitter mesajlarının makine öğrenmesi yöntemleri ile sınıflandırılabileceği tezi doğrulanmıştır.

3. MATERYAL VE YÖNTEM

3.1. Doğal Dil İşleme

Teknolojinin gelişmesi ile birlikte birtakım problemler ve bu problemlerin çözümüne yönelik birtakım ihtiyaçlarla birlikte de yeni araştırma alanları ve bilgisayar bilimleri ortaya çıkmıştır. Bilgisayarlara, iletişim dilinin aktarılması ve dilin doğru bir şekilde çözümlenmesi ihtiyacı da haliyle beraberinde gelmiştir. Bunun sonucunda da bilgisayar bilimi, yapay zekâ ve bilişimsel dil biliminin birleşiminin bir araya gelmesiyle "Doğal Dil İşleme (Natural Language Processing)" doğmuş olup bu alanın asıl kaynağını bilgisayar kullanarak iki farklı dili birbirine çevirme işlemi oluşturmaktadır (Çoban, 2016; Gordin, 2015; Yıldırım, 2018).

Son yıllarda oldukça önem kazanmış bir çalışma alanı ve bilgisayar bilimi olan DDİ, esasen 20. yüzyılın ikinci yarısının başlarında yapay zekânın küçük bir alt dalı olarak meydana gelmiştir (Oflazer, 2006). 1954 yılında Georgetown deneyi olarak adlandırılan Georgetown Üniversitesi ve IBM'in birlikte geliştirdiği bilgisayarlı çeviri deneyinde 60'dan fazla Rusça cümlenin başarılı bir şekilde İngilizceye çevrilebildiği görülmüştür. 1980'li yıllarda ise, hesaplama gücünün hızla artması ile bu alandaki gelişmeler hızlanmıştır (Gordin, 2015). Zamanla yapılan araştırmalar ve gerçekleştirilen uygulamalar sonucunda elde edilen başarılar ile de DDİ nihayetinde bilgisayar bilimlerinin temel bir disiplini olarak kabul edilmeye başlanmıştır.

DDİ kısaca; sosyal medya platformları, web sayfaları, e-postalar, farklı dillerde yayımlanan gazete makaleleri ve başka birçok kaynaktan elde edilen doğal dilde yazılmış metinler ile ilgilenen bir alan olarak tanımlanabilir (Korkusuz, 2018). DDİ, doğal dillerin kurallı olan yapısını detaylıca incelerken aynı zamanda çözümleyerek işlemekte ve anlaşılması yahut yeniden üretilmesi amacını taşıyarak otomatik çeviri, konuşma, ses tanıma, üretme ve duygu analizi gibi pek çok konuyu içeren çalışmada kullanılmaktadır (Yelmen, 2016). Genellikle metin tabanlı çalışmalarda kullanılmakta olup istatistiksel olarak metnin üzerinden sonuçlar üretmeyi kapsarken çoğunlukla yapay zekâ altındaki dil bilim bilgisine dayalı çalışmaları içermektedir (Şeker, 2014).

Genel olarak DDİ'nin kullanım alanları şu şekilde ifade edilebilir (Kızılırmak, 2020):

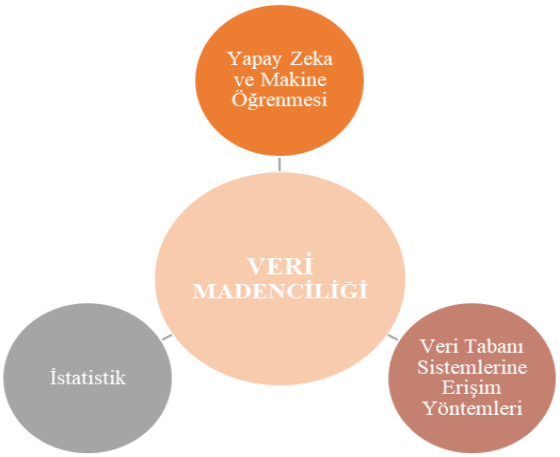
- Soru Cevaplama (Question Answering, QA): Bilgisayara sorulan bir soruya cevap vermeyi ifade eder. Örneğin, kimi web sayfalarında yer alan yardım botlarında kullanılan yöntemler.
- Bilginin Çıkartılması (Information Extraction, IE): Bir metnin analiz edilerek içinden bilgi çıkarılmasını ifade eder. Örneğin, bir davetiye metini içerisinden davetin ne zaman ve nerede gerçekleşeceği gibi bilgilerin analiz edilmesi.
- Duygu Analizi (Sentement Analysis, SA): Kısaca bir cümlenin olumlu ya da olumsuz olup olmadığının analiz edilmesi şeklinde ifade edilebilir. Çalışmanın devamında bu başlıktan detaylıca bahsedilecektir. Örneğin, Torku markasına ait süt ürünlerinin fiyatının çok uygun olduğunu (olumlu) fakat paketlenmesinin kötü olduğunun (olumsuz) analiz edilmesi.
- Makine Diline Çeviri (Machine Translation, MT): Bir dilden diğer dile DDİ işlemiyle çeviriyi ifade eder. Örneğin, pek çok kişi tarafından sıklıkla kullanılan Google Translate.
- Sözcük Anlamı Açıklaştırma (Word Sense Disambiguation, WSD): Bir kelimenin sözlük anlamı olduğunu eğer sözlük anlamı yoksa sözlükte bulunmayan bir kelime olmasının analiz edilmesini ifade eder. Örneğin, ülkemizin şehirlerinden biri olan “Uşak” kelimesinin şehir mi yoksa erkek hizmetçi anlamında olan kelime mi olduğunun tespit edilmesi.
- Özetleme (Summarization): Bir metin ya da paragraf içerisinden konu ile ilgili özet çıkarma işlemini ifade eder. Örneğin, bir blog yazısı içerisinden, yazının özetinin çıkarılması.

Özetle yukarıda bahsedilen soru cevaplama, bilginin çıkartılması, duygu analizi, makine diline çeviri, sözcük anlamı açıklaştırma ve özetleme alanlarında kullanımı olan DDİ yöntemleri, insanların el ile yapmak zorunda olduğu bazı metin sınıflandırmaların otomatik olarak yapılabilmesine olanak vermekte ve birçok spesifik uygulamada ciddi kolaylıklar sağlamaktadır. Bu yönüyle DD’nin metinsel verinin yoğun olduğu tüm alanlarda kullanılabilecek tekniklerden oluştuğu söylenebilir.

3.2. Veri Madenciliği

Son yıllarda teknolojinin hızla gelişmesi ve veri miktarındaki hızlı artışla birlikte ilerleme kaydeden ve farklı türlerine erişilebilen bir alanı oluşturan veri madenciliği, literatürde veri keşfi olarak da adlandırılmaktadır. Artan veri boyutlarından anlamlı bilgi elde edebilmek için, bilgisayar temelli yeni yöntemlere ihtiyaç duyulmaya başlanmasıyla birlikte veri tabanlarında bilgi keşfi ve veri madenciliği konuları araştırmacılar için ilgi çekici olmaya başlamıştır (Fayyad ve ark., 1996; Emre ve Selçukcan Erol, 2017). Bu sebeple araştırmacılar, büyük verinin analizi ya da yorumlanması için veri madenciliği kullanmaktadır.

Veri keşfi bir diğer ifade ile veri madenciliği, Şekil 3.1.'de belirtildiği üzere büyük ölçekli verilerin işlenerek bilgiye ulaşması işlemi için yapay zekâ, makine öğrenmesi, istatistik ve veri tabanı sistemlerine erişim yöntemleri disiplinlerinin bir araya gelmesiyle oluşmuş çok disiplinli bir araştırma alanıdır (Chakrabarti ve ark., 2006). Veri madenciliği, büyük miktarda verinin bilgisayarlar aracılığı ile işlenmesi sayesinde bir sonuca etki eden nedenlerin kolaylıkla ayrıştırılarak analizinin yapılabilmesine ve geleceğe yönelik durum olay tahminine olanak sağlar (Vikipedi, 2021).



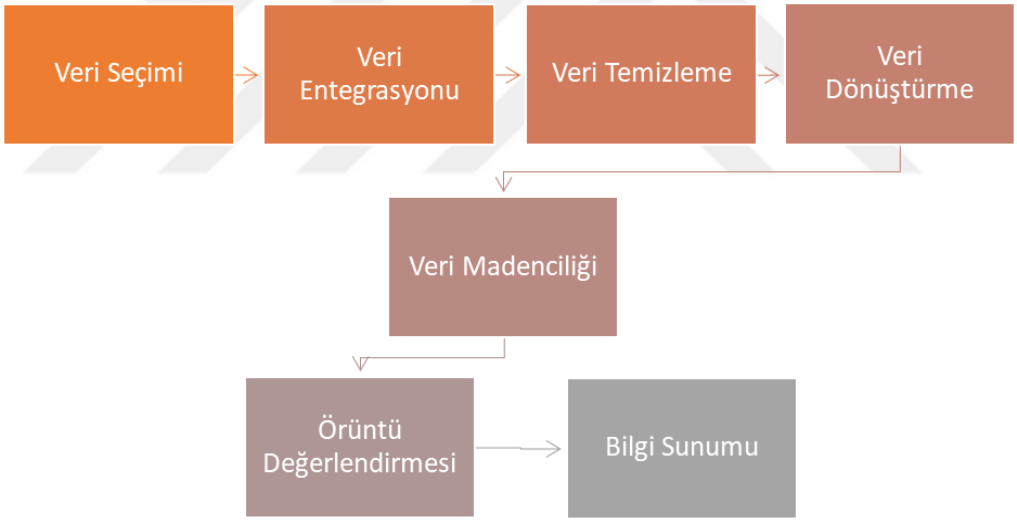
Şekil 3.1. Veri Madenciliğini Oluşturan Disiplinler

Veri madenciliği genel olarak, verileri araştıran, matematiksel modeller geliştiren ve yararlı bilgilerin özünü keşfeden, çıkarım algoritmalarını içeren veri tabanından bilgi bulma sürecinin matematiksel özü şeklinde ifade edilebilir (Maimon ve ark., 2009). Temel sorusunu "Veriden anlamlı bilgi elde edebilmek için neler nasıl yapılmalıdır?" sorusu oluşturur. Bu sorunun cevaplanabilmesi için öncelikle mevcut

problemin iyi tanımlanmış olması gerekmektedir. Bunun için de veri madenciliği modelinin uygulandıktan sonra nasıl bir çıktı almanın hedeflendiği problemin tanımlanması aşamasında belirlenmelidir.

Veri madenciliği modellerinde çıktı genellikle bir tahmin sonucu oluşan ya da modellenmiş bir veri setidir. Kısaca veri üzerinde madencilik yapmak olarak tanımlanabilen veri madenciliğinde veri setleri içerisinde bulunan ve genellikle hemen görülemeyen bilgilerin ortaya çıkarılarak veriye dayalı karar verme ve tahmin çalışmaları için geliştirilmiş birçok teknik olduğu, bu tekniklerin kullanılarak geliştirilen modellerin veri madenciliğinin çıktısı olduğu ifade edilebilir (Erden, 2021).

Veri madenciliği temelde üç adımdan oluşur: Veri ön işleme süreci, gerçek elde işlemi ve veri analizi. Veri ön işleme sürecinde verilerin seçimi, entegrasyonu, temizlenmesi ve dönüştürülmesi aşamaları yer alır. Gerçek elde işlemi veri madenciliği aşamasında elde edilir. Üçüncü ve son adım olan veri analizi adımı ise örüntü değerlendirmesi ve bilgi sunumu adımlarını kapsar (Şekil 3.2).



Şekil 3.2. Veri Madenciliği Adımları

Karmaşık veri yığınından anlamlı bilgiye ulaşılabilmesi için izlenmesi gereken veri madenciliği adımları kısaca şu şekilde ifade edilebilir (Han ve ark., 2011):

1. Veri Seçimi: Üzerinde çalışılacak verinin veri tabanından yahut herhangi bir kaynaktan alınmasını ifade eder.
2. Veri Entegrasyonu: Verinin birden çok kaynaktan alınması durumunda bu verilerin birbirleriyle olan birleştirilme işlemini ifade eder.

3. Veri Temizleme: Veri içindeki tutarsızlıkların ve gürültünün giderilmesini ifade eder.
4. Veri Dönüştürme: Verinin özetleme ya da derleme işlemlerine tabi tutularak kullanıma uygun hale getirilmesini ifade eder.
5. Veri Madenciliği: Veri örüntülerini ortaya çıkarmak için akıllı yöntemlerin uygulandığı süreci ifade eder.
6. Örüntü Değerlendirmesi: Bilgiyi temsil eden ilginç örüntülerin özel ölçümlere dayanarak belirlenmesi işlemini ifade eder.
7. Bilgi Sunumu: Ortaya çıkarılan bilginin görselleştirme ve bilgi sunum yöntemleri kullanılması ile kullanıcıya gösterilmesi aşamasını ifade eder.

Veri seçimi, veri entegrasyonu, veri temizleme ve veri dönüştürme adımları literatürde veri ön işleme süreci olarak adlandırılmaktadır. Veri ön işleme süreci, üzerinde madencilik yapılacak verinin temizlenmesi ve işleme hazır hale getirilmesi için gereken adımları içermekteyken; veri madenciliği adımı ise, kullanıcı ile etkileşimli bir şekilde gerçekleştirilebilen bir adımdır. Bu adımdan sonraki örüntülerin değerlendirilmesi adımında, ilgi alanı dışındaki örüntüler belirli ölçümlerle ayıklanarak önemli olanlar bir sonraki aşamada kullanıcıya değişik yollarla gösterilmektedir. Genellikle küçük veri kümeleri ile ilgilenmeyen veri madenciliği bu noktada çok büyük veri kümelerinde standart yöntemlerle görülemeyecek bilgi ve örüntülerin ortaya çıkarılmasında önem taşır (Aytekin, 2012; Yurt, 2015).

3.3. Metin Madenciliği

Son yıllarda donanım ve yazılım teknolojisindeki büyük gelişmelere bağlı olarak hızlı ilerlemeler gösteren veri madenciliği, farklı veri türlerine uygulanılabilen bir alan haline gelmiştir. Donanım ve yazılım platformlarının geliştirilmesiyle bu durum, özellikle web ve sosyal ağlar için büyük miktardaki metin türündeki veriler için de gerçekleşmiştir. Farklı uygulamalar sonucu elde edilen metin verilerinin miktarının artmasıyla verilerin dinamik ve ölçülebilir bir şekilde öğrenilebilmesini sağlayan algoritmik tasarımların geliştirilmesine ihtiyaç duyulmaya başlanmıştır (Bender ve ark., 2003). İşte bu noktada metin madenciliği devreye girmektedir.

Metin madenciliği; yapılanmamış ya da yarı yapılı metinlerden anlamlı sonuçlar çıkarabilmek amacıyla metinlerin belirli süreçlerden geçirilip yapılandırılmış hale

getirilerek çıktıların analiz edilmesini ifade etmektedir (Karamanlı, 2019). Metin madenciliği uygulaması temelde insan beynindeki en karmaşık analitik işleme sistemine sahip anlayışları, yazı dilinde analiz etmeyi amaçlamaktadır. Metin türündeki verilerde kelime ve cümlelerin her zaman doğal sıralanışında olmamasından dolayı metin madenciliğinde kullanılacak istatistiksel yöntemlerin doğru sonuçlar verebilmesi, DDİ teknikleriyle kelime ve cümlelerin ön işlemesinin dikkatli bir şekilde yapılması ile mümkündür (Berger ve ark., 1996).

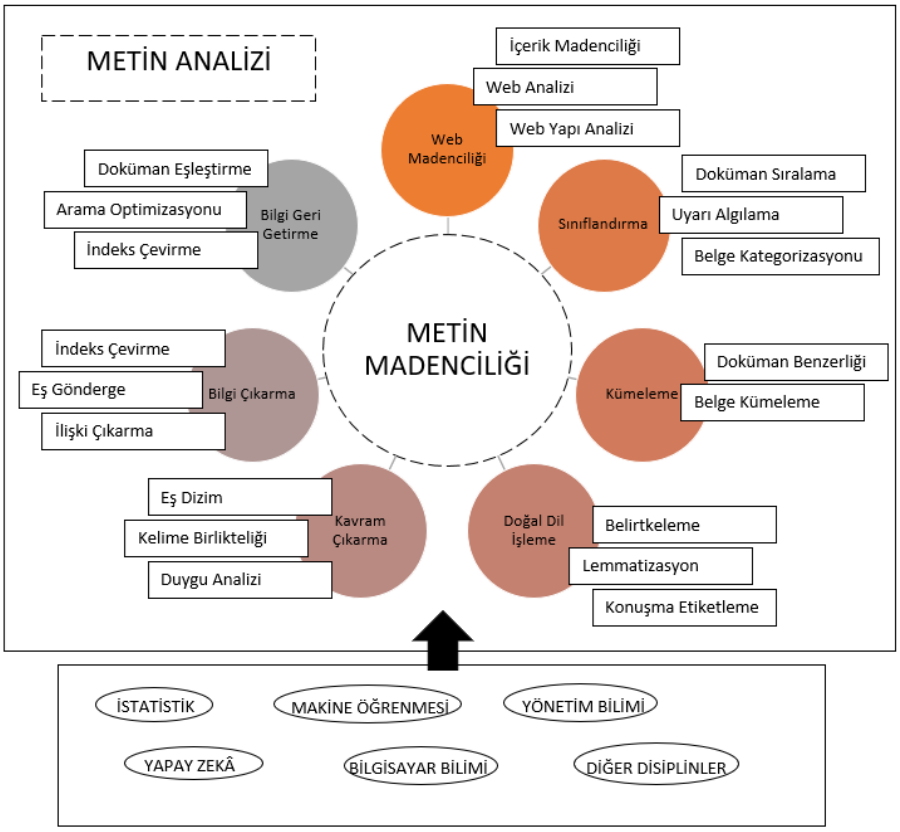
Tablo 3.1. Metin Madenciliği ile Veri Madenciliğinin Karşılaştırılması

Metin Madenciliği	Veri Madenciliği
Dil işleme/ DDİ	Doğrudan işlem
Önceden bilinmeyen bilgilerin keşfi	Nedensel ilişkiyi tanımlama
Yarı yapılandırılmış ve yapılandırılmamış veri seti	Yapılandırılmış veri seti
Birden fazla tipte bulunan belli bir alanda bulunmayan veri seti	Veri tabanında bulunan, formatlı veri

Kaynak: Slideshare (2019).

Tablo 3.1.'de metin madenciliği ile veri madenciliğinin karşılaştırılması yapılmıştır. Mevcut verinin türüne göre metin madenciliği ya da veri madenciliği tercih edilebilir. Genel olarak, veri madenciliğinde veri hazır bir veri tabanından alınır ve formatı bellidir. Yapılandırılmamış veri seti kullanılarak nedensel ilişkiyi tanımlama amacı güdülür. DDİ çalışmaları ile birlikte yürütülen metin madenciliğinde ise, metin kaynaklı literatürdeki diğer bir çalışma alanı olan birçok veri türü aynı anda daha dağınık veri kaynaklarında bulunmaktadır. Ayrıca yarı yapılandırılmış ve yapılandırılmamış veri seti kullanılarak önceden bilinmeyen bilgilerin keşfedilmesi amaçlanır.

Metin madenciliği tekniklerinin kullanılmasıyla gerçekleştirilen işlemler ve ilişkili disiplinler Şekil 3.3.'te verilmiştir (Amanet, 2017). Buna göre metin madenciliğinin bilgisayar bilimi, istatistik, yapay zekâ, makine öğrenmesi, yönetim bilimi ve diğer bilimler ile iç içe olduğu söylenebilir. Esasında metin madenciliği kavram çıkarma, bilgi çıkarma, bilgi geri getirme, web madenciliği, sınıflandırma, kümeleme ve DDİ alanlarında kullanılır.



Şekil 3.3. Metin Madenciliği Tekniklerinin Kullanılmasıyla Gerçekleştirilen İşlemler ve İlişkili Disiplinler

Metin madenciliği teknikleri, temelde dört kategoriye ayrılmaktadır (Tunalı, 2009):

1. Sınıflandırma (Classification): Nesnelerin önceden bilinen sınıflara yahut kategorilere dahil edilmesi işlemini ifade eder.
2. Birliktelik Analizi (Association Analysis): Sıklıkla birlikte yer alan yahut gelişen sözcük veya kavramların belirlenerek doküman içeriğinin veya doküman kümelerinin anlaşılmasını sağlar.
3. Bilgi Çıkarımı (Information Extraction): Dokümanların içerisindeki yararlı veri ya da ifadelerin bulunması işlemini ifade eder.
4. Kümeleme (Clustering): Doküman kümelerinin temelini oluşturan yapıların keşfedilmesi işlemini ifade eder.

Veri kaynağı olarak metinlerin ele alındığı metin madenciliğinde, ilk olarak metin kaynaklarından amaca uygun olan veri seçimi yapılırken sonrasında tekrar eden

ve etkisiz kelimeler filtrelenmekte ve metinler istenilen özelliklere göre parçalara ayrılmaktadır. Bunu yaparken de aynı zamanda metinler noktalama işaretlerinden, sayısal verilerden arındırılarak metinlerdeki büyük harfler küçük harflere de dönüştürülmektedir. Böylece metin analizi için bir ön işleme yapılmaktadır. Metin verilerinin ön işleme aşamasını geçmesinin ardından gösterim aşamasına geçilir ve bu aşamada bir kelimenin metin içerisinde ne kadar önemli olduğunun istatistiki olarak değerlendirilmesi için metnin sayısal olarak ifade edilmesi işlemi gerçekleştirilmektedir (Aninditya ve ark., 2019; Beşkirli ve ark., 2021). Son olarak da pek çok aşamadan geçerek metin madenciliği için kullanıma hazır hale getirilen veriler metin kaynaklarından bilgi keşfi için sınıflandırma, birliktelik analizi, bilgi çıkarımı ve kümeleme teknikleri uygulanarak analiz edilmektedir.

3.3.1. Veri Ön İşleme

Veri seti oluşturma aşamasında elde edilen metinleri kullanabilmek için öncelikle metinlerin düzenlenmesi gerekmektedir. Bu düzenleme ile metinlerde bulunan yazım yanlışlarını, kullanılmayacak olan internet sayfası linkleri, kullanıcı adı gibi verileri kısaltma olarak kullanılan kelimelerde düzenlemeler yaptıktan sonra kullanıma hazır hale getirilir (Kuzucu, 2015). Yapılandırılmamış metin formatındaki verilerin, metin madenciliği süreçlerinde işlenerek yapılandırılmış hale getirilebilmesi için gereken ilk adım ön işleme aşamasıdır. Bu adımda, metinlerdeki gereksiz içerikleri çıkarma, kısaltma olan kelimeleri doğru kelimeler ile değiştirme, kelimeleri sözcük parçacıklarına bölme (tokenizasyon) işlemleri uygulanarak veri analiz edilebilir hale getirilmektedir (Çınar, 2020; Sar, 2021).

Sınıflandırmada kullanılacak veriler, bazen eksik ya da tutarsız olabilmektedir. Veri setinde bulunan eksik ya da hatalı verilere gürültü adı verilmektedir. Veri setinde gürültülü verilerin bulunması halinde, bu sorunun giderilmesi beklenirken bu gibi durumlarda aşağıdaki yöntemler kullanılmaktadır (Özkan, 2008):

- Gürültülü verilerinin ilgili veri setinden silinmesi yahut yerine yenisinin eklenmesi gerekmektedir.
- Gürültülü verinin yerine sabit bir değer kullanılabilir.
- Tüm verilerin ya da bir kısım verilerin ortalaması hesaplanıp gürültülü verilerin yerine bu değer kullanılabilir.

- Gürültülü verilerin yerine, veri setinde yer alan verilerin tamamı ya da belli bir kısmı kullanılarak gürültülü veriler tahmin edilir ve elde edilen bu veriler gürültülü verilerin yerine kullanılabilir.

3.3.2. Veri Etiketleme

Veri etiketleme temel olarak görsel, metin veya işitsel verileri insanların algıladığı özelliklerle işaretlemek ve anlamlandırmaktır. Veriler üzerinde birçok farklı öğe etiketlenebilse de genellikle belirtilen alt kümeye odaklanması beklenmektedir. Makine öğrenmesi çalışmaları 3 temel aşamadan oluşmaktadır, bunlar; veri hazırlama, modelin eğitimi ve modelin uygulanmasıdır. Makine öğrenmesi çalışmalarında veri etiketleme, verilerin hazırlanmasında büyük rol oynamaktadır. Bu aşamada, modelin eğitilmesi için etiketlenmiş veriler kullanılmaktadır (Andırın, 2020). Etiketleme işlemi kişiler tarafından manuel olarak değerlendirilip yapılabildiği gibi geliştirilen hazır modeller ile etiketlendirme işlemi yapmak mümkündür. Bahsedilen hazır modeller ve özellikleri şöyledir:

- **TextBlob**

TextBlob, Natural Language Toolkit (NLTK) üzerine inşa edilmiş bir python kütüphanesidir. Duygu analizi, isim öbeği çıkarma, konuşma bölümü etiketleme ve daha birçok DDİ işlemleri için kolaylık sağlamaktadır. Textblob iki tür çıktı sağlamaktadır. Birincisi her kelime için ne kadar olumlu ya da ne kadar olumsuz olduğunu veren duygu puanı (polarity), ikincisi de bir öznellik puanıdır (subjectivity) (Subramanian, 2019).

Şekil 3.4. TextBlob Çıktıları

Polarite bir kelimenin ne kadar olumlu ya da olumsuz olduğunu verir. Polarite -1 ile +1 aralığındadır. -1 çok olumsuz, +1 çok olumlu olarak ifade edilmektedir. Öznellik ise bir kelimenin ne kadar nesnel ya da öznel olduğunu vermektedir. Öznellik değeri 0 ile +1 aralığındadır. Bu değer 0'a yakınsa nesnel, +1'e yakınsa özneldir.

- **Valence Aware Dictionary and Sentiment Reasoner**

Sözlük ve kural tabanlı duygu analizi aracı olan Valence Aware Dictionary and Sentiment Reasoner (Vader) metin duyarlılığını hesaplamak için anlamsal yönelimlerine göre olumlu ya da olumsuz olarak etkilenen sözcüklerin bir listesini kullanır. Bir kelimenin ne kadar olumlu ya da ne kadar olumsuz (+, - polarite) olduğunun aksine yoğunluğunu dikkate almaktadır. Özellikle sosyal medya türündeki metinlerde oldukça başarılı sonuçlar vermektedir. Eğitilmesine gerek olmayan Vader MIT lisansı altında tamamen açık kaynaklıdır (Pandey, 2018).

```
print(analyser.polarity_scores("I like distance education"))

{'neg': 0.0, 'neu': 0.545, 'pos': 0.455, 'compound': 0.3612}
```

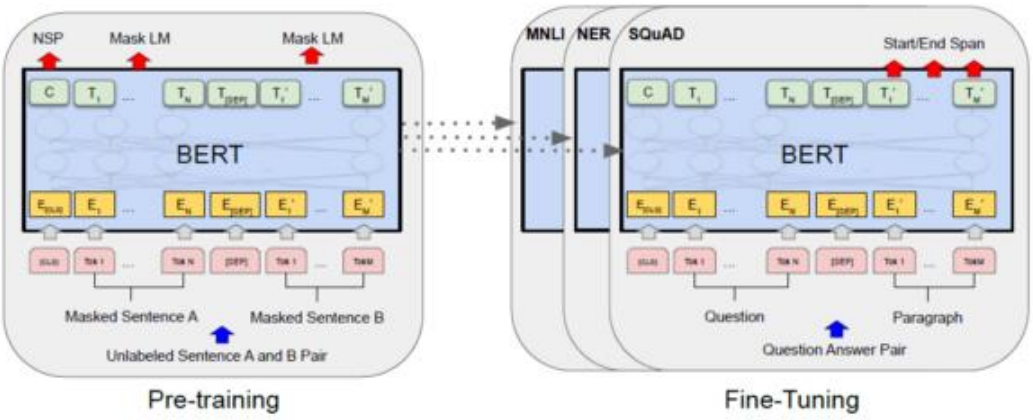
Şekil 3.5. Örnek Vader Çıktısı

Vader negatif, nötr, pozitif ve bileşik değer olarak 4 çıktı vermektedir. Şekil 3.5'teki negatif, nötr ve pozitif değerler bu kategoriye giren kelimelerin oranını temsil etmektedir. Bu oranlar %45.5 pozitif, %54.5 nötr ve %0 negatif olarak metni derecelendirdiği anlamına gelmektedir. Buradaki değerlerin toplamı 1 olmalıdır. Bileşik değer ise, [-1, +1] (-1 çok olumsuz, +1 çok olumlu) arasında normalleştirilmiş tüm sözlük derecelendirmelerinin toplamını hesaplayan bir ölçümdür (Pandey, 2018).

- **Bidirectional Encoder Representations from Transformers**

2018 yılında, Google tarafından duyurulan Bert modelinin diğer modellerden en büyük farklı cümleyi hem sağdan sola hem de soldan sağa tarayarak değerlendirmesidir. Böylelikle kelimelerin birbiriyle olan ilişkilerini ve anlamlarını daha iyi çıkarmayı hedeflemektedir. Bert, 800 milyon İngilizce kelime (BookCorpus) ve 2,5 milyar İngilizce kelime (Wikipedia) toplamda 3,3 milyar kelime hazinesine sahip korpus üzerinde eğitilmiş, bert_large ve bert_base adı verilen 2 temel model ile sunulmuştur.

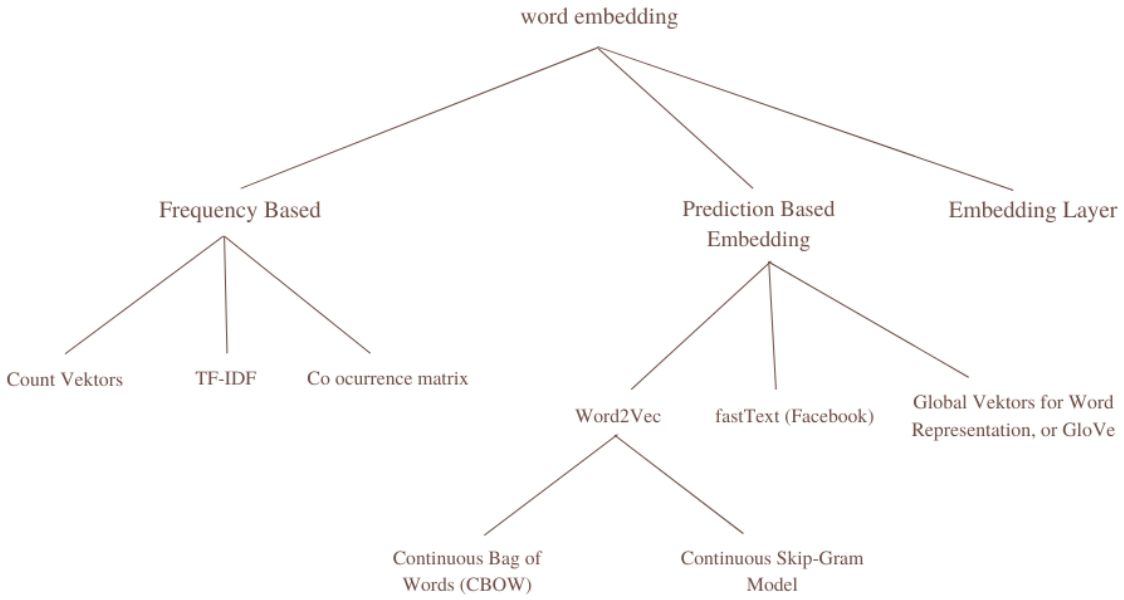
Bu modellerin 4 günlük eğitim sürecinde bert_large için 16 adet TPU, bert_base için 4 adet TPU ya ihtiyaç duyulmuştur. Şekil 3.6'da yer alan fine-tuning adı verilen teknikle hazır eğitilmiş modeller kullanılarak yeni problemlere çözüm üretilebilmektedir (Uçar, 2020).



Şekil 3.6. Bert (Devlin ve ark., 2018)

3.3.3. Veri Sayısallaştırma (Kelime Gömme)

DDİ çalışmalarının tümü metinsel ifadeler ile yapılmaktadır, ancak bilgisayar sistemleri metinsel ifadeleri algılayamadığından DDİ'nin ana işlevi olan dil algılama özelliği metinsel verilerle sonuçlanamaz. Bu sebeple metinlerin sayısal olarak ifade edilmesi gerekmektedir (Sar, 2021). Bir derlemde metin işleme çalışmalarının en kritik yönlerinden biri, metinlerin nasıl temsil edileceğidir. En basit tanımıyla kelime gömme (Word embedding), metinlerin sayısal ifadelere dönüştürülmesi olarak ifade edilebilir (Aydoğan ve Karcı, 2019). Kelime gömmeleri frekans tabanlı ve tahmin tabanlı olarak ikiye ayrılabilir (Şekil 3.7).



Şekil 3.7. Veri Sayısallaştırma Yöntemleri

3.3.3.1. Kelime Çantası

Kelime çantası (Bag of Words-BoW) frekans temelli kelime temsili yöntemlerinde en çok tercih edilen modeldir. Bu modelde, belgedeki her cümle benzersiz kelimelere bölünür ve benzersiz kelime boyutunda bir matris oluşturulur. Matrisin sütunlarını belgedeki kelimeler (N), satırlarını ise doküman sayısı (D) oluşturur. Sonuç olarak, tüm derlem bir DxN matrisi olarak temsil edilmektedir (Aydoğan ve Karcı, 2019).

Model sadece kelimelerin belgede olup olmadığı ile ilgilenir. Başka bir deyişle, bu modelde her kelimenin sayısı bir özellik olarak kabul edilir. Bu kelime sayıları, belgelerin karşılaştırılmasına ve belge sınıflandırma ile konu modelleme gibi uygulamalar için kelime benzerliğinin ölçülmesine olanak tanır. BoW modeli aşağıdaki özellik fonksiyonu ile elde edilir (Albayrak, 2018).

d_i, $w \in d_i$ kelimelerinden oluşan bir derlemde:

$$fi(X) = \begin{cases} 1, & d_i, w_i \ kelimesini \ içeriyorsa \\ 0, & diğer \end{cases} \quad (1)$$

BoW modeli, bir cümleyi vektör olarak temsil etmek için kelimelerin her birini bir özellik olarak kullanmanın yeterli olduğunu varsaymaktadır (denklem 2).

$$d_i = (w_{i1}, w_{i2}, \ldots w_{ij}, \ldots w_{in}) \quad (2)$$

w_{ij}, d cümlesindeki w_i. terimi. n, veri setindeki $|D|$ tüm kelimelerin sayısını ifade etmektedir.

3.3.3.2. Terim Frekansı-Ters Doküman Frekansı

Frekans bazlı temsil yöntemlerinden bir diğeri olan Terim Frekansı-Ters Doküman Frekansı (Term Frequency-Inverse Document Frequency-TF-IDF), bir belgedeki her kelimenin değerlerini, belirli bir belgedeki kelimenin frekansının ve kelimenin göründüğü belgelerin yüzdesinin tersi olarak hesaplamaktadır. Bu hesaplama belirli bir kelimenin belirli bir belgeyle ne kadar alakalı olduğunu sezgisel olarak belirlemektedir. Temel olarak, TF-IDF, belirli bir belgedeki kelimelerin tüm veri kümesi üzerindeki ters oranına dayalı olarak kelimelerin göreli frekansını belirleyerek çalışır. Tek ya da küçük bir belge grubu için ortak olan kelimeler, genel kelimelerden daha yüksek TF-IDF numaralarına sahip olma eğilimindedir (Ramos, 2003; Çelik ve Koç., 2021). TF-IDF modeli hesaplama fonksiyonu denklem 3'deki gibidir.

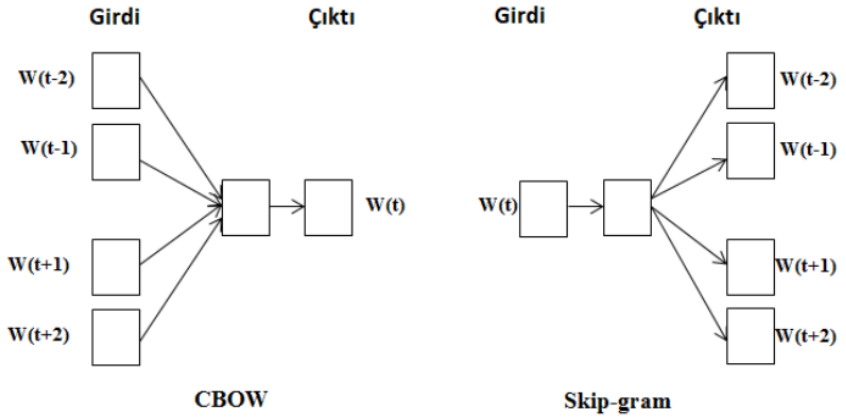
$$w_{i,j} = tf_{i,j} \times \log \frac{N}{df_i} \quad (3)$$

$tf_{i,j}$ = i kelimesinin j belgesindeki frekansı
df_i = i kelimesini barındıran doküman sayısı
N = toplam doküman sayısı

3.3.3.3. Kelime Vektörü

Veri sayısallaştırma yöntemlerinden tahmin bazlı sayısallaştırma yöntemi olan Kelime vektörü (Word2Vec), Mikolov ve ark. (2013) tarafından önerilmiştir. Word2Vec, kelimeleri gömmek için sinir ağı tabanlı bir yaklaşımdır. Büyük bir metin kümesi ile eğitilen bu model kelimeleri n boyutlu bir uzayda benzersiz birer vektör olarak temsil etmektedir ve oluşturulan bu benzersiz vektörlerin özelliği anlamsal benzerliğe sahip kelimelerin birbirine yakın vektörler oluşturmasıdır (Bilgin, 2019).

Word2Vec, anlam ilişkileri kurmak amacıyla iki farklı öğrenme mimarisi ile çalışır. Bunlardan ilki olan Continous Bag of Words (CBoW), pencere merkezindeki kelimeyi tahmin etmek için kelimenin pencere boyutu kadar yakın komşularına bakar. CBoW'a benzer şekilde çalışan ikinci yöntem Skip-Gram ise kelimenin komşularını pencere merkezinde konumlandırılan kelimeden tahmin eder. Bu modelin avantajı, farklı anlamlara sahip kelimelerin birden fazla anlamını yakalayabilmesidir (Çelik ve Koç, 2021).



Şekil 3.8. Word2Vec CBoW ve Skip-gram (Çelik ve Koç., 2021)

3.3.3.4. N-Gram

N-gram, bir karakter katarının n adet karakter dilimi olarak ifade edilmektedir. N-gram tabanlı sınıflandırma yöntemi, doküman içerisindeki karakter tabanlı N-gram'ların kullanım sıklığına dayalı bir işlemdir (Doğan ve Diri., 2010).

Önceki n elemanlı sıralamanın olma olasılığı bilindiğinde sıradaki olayın olma olasılığını tahmin etmeye çalışan N-gram modelinde, DDİ kullanıldığı zaman n-1. sıradan daha önceki kelimeler ile bağımsızlık varsayımı uygulanır. Kelimenin olma olasılığı sadece kendinden önceki n-1 kelimeye bağlıdır (Amanet, 2017).

N-gram modelinde N 1 ise unigram, N 2 ise bigram N 3 ise trigram olarak nitelendirilmektedir. Tablo 3.2'de N-gram modellerinin örnek bir metin üzerinde çıktıları gösterilmiştir.

Tablo 3.2. Örnek Bir Metinde N-Gram Çıktıları

Metin: “Uzaktan eğitimi sevdim. Böyle devam edebilir”	
	N-gram Çıktı
Unigram (N=1)	“uzaktan”, “eğitimi”, “sevdim”, “böyle”, “devam”, “edebilir”
Bigram (N=2)	“uzaktan eğitimi”, “eğitimi sevdim”, “sevdim böyle”, “böyle devam”, “devam edebilir”
Trigram (N=3)	“uzaktan eğitimi sevdim”, “eğitimi sevdim böyle”, “sevdim böyle devam”, “böyle devam edebilir”

N-gramlar, duygu analizi için makine öğrenmesi yaklaşımında oldukça sık kullanılmaktadır. N-gramlar bileşik kelimeler ve deyimler şeklinde beraber kullanıldıklarında daha anlamlı ve yüksek seviyede bilgi barındırırken, kelimelerin birçoğu tek başına kullanıldığında yeterince bilgi içermezler. Bu doğrultuda n-gramlar duygu analizi çalışmalarında duygu içeren kelime sıraları elde etmemizi sağlarlar (Rosenfield ve ark.; Stolcke, 2002; Türkmenoğlu, 2020).

3.4. Makine Öğrenmesi

Bir sorunun bilgisayar ortamında çözülebilmesi için algoritmaya ihtiyaç duyulmaktadır. Algoritma; belli bir problemi çözmek yahut belirli bir amaca ulaşmak için tasarlanan yol olarak tanımlanmaktadır. Algoritmalar, bilgisayar biliminde bir işi yapmak amacıyla tanımlanan, bir başlangıç durumundan başladığında, açıkça belirlenmiş bir son durumunda sonlanan, sonlu işlemler kümesidir (Vikipedi, 2021). Bilgisayar sayesinde saniyede milyonlarca işlem yapabilen bir makine öğrenimi algoritması ile karmaşık ve uzun süren işlemler için otomatik algoritmalar geliştirilebilmektedir (Alpaydın, 2010; Aykul, 2019).

1959 yılında Arthur Samuel tarafından bir araya getirilen “makine öğrenmesi” terimi, kısaca bilgisayarın açıkça programlanmadan öğrenme yeteneği olarak ifade edilmektedir. Otomatik olarak öğrenme işlemini deneyimlerden yola çıkarak geliştiren ve gerçekleştiren bilgisayar sistemlerinin geliştirilmesi olarak tanımlanabilir (Ayodele, 2010). Makine Öğrenmesi, gözlenmiş bir örneklem kümesinden doğru tahminler yapabilmeyi öğrenebilmek amacıyla otomatik teknikler geliştirerek matematiksel ve

istatistiksel yöntemler aracılığıyla örneklem kümesini oluşturan mevcut verilerden çıkarımlar yaparak elde ettiği çıkarımlarla bilinmeyene dair tahminlerde bulunan sistem şeklinde açıklanabilir (Elmas, 2019; Schapire 2003).

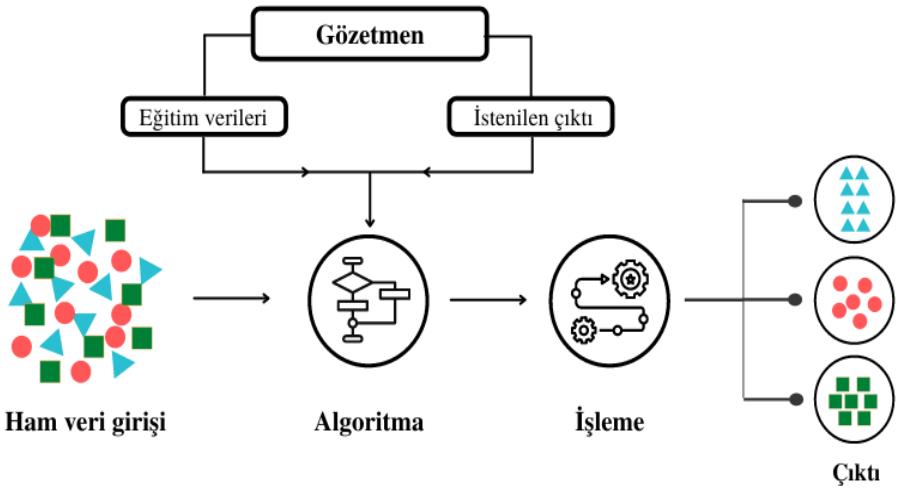
Günümüzde çok sayıda büyük verinin bulunması, bu verilerin çözümlenmesinde bilgisayar teknolojilerine ihtiyaç duyulmasını arttırmıştır. Bu noktada makine öğrenmesi üzerine yapılan araştırmalar da artmıştır. Bu artışla birlikte hesaplamalı öğrenme teorisi, yapay sinir ağları, istatistik ve örüntü tanıma gibi araştırma alanları arasında bağlantı kurularak bu alanlarla birlikte çalışılmaya başlanmıştır. Böylece makine öğrenmesi teknikleri yüz tanıma gibi daha geleneksel sorunların yanı sıra veri tabanlarında bilgi keşfi, dil işleme ve robot kontrolü gibi problemlere uygulanmaya başlanmıştır (Dietterich 1997; Çoban, 2016).

Literatürde sıklıkla kullanılan iki tür makine öğrenmesi makine öğrenmesi yöntemi vardır:

1. Denetimli Makine Öğrenmesi (Supervised Machine Learning)
2. Denetimsiz Makine Öğrenmesi (Unsupervised Machine Learning)

3.4.1. Denetimli Makine Öğrenmesi

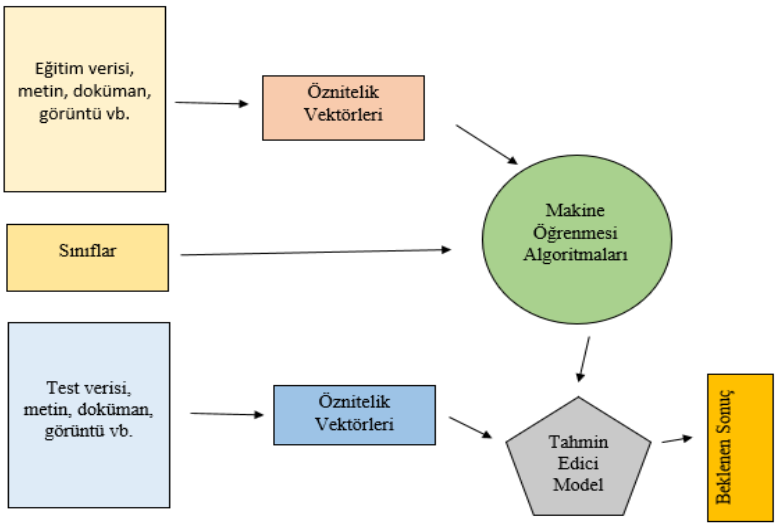
Denetimli Makine Öğrenmesi, sistemin etiketli veriler kullanılarak eğitilmesi ile öğrenmenin sağlanmasını ifade eder. Sistem eğitilirken veri setinde bulunan her bir örneğe ait giriş ve çıkışlar verilmektedir. Giriş, metnin içeriğini; çıkış metnin kategorisini temsil etmektedir. Sistemin doğrulanması amacıyla ise test veri seti kullanılır. Bu aşamada öğrenme algoritması kategorisi bilinmeyen bir test verisine, eğitim verisinde bulunan çıkışlardan herhangi biri atanır (Kotsiantis ve ark., 2007; Bilgin, 2017).



Şekil 3.9. Denetimli Makine Öğrenmesi (Turhost, 2021)

Yukarıda verilen denetimli makine öğrenmesi modelinde de görüleceği üzere önceden belirlenmiş etiketlenmiş verilere dayanarak istenilen çıktılara ulaşılabilmesi için makine öğrenmesi algoritması ham veri girişini işleyerek çıktılara ulaşmaktadır. Böylece girdiler ve çıktılar arasında eşleşme yapan bir fonksiyon elde edilmektedir. Örneğin, son 50 yılda, ülkelerin nüfuslarını içeren bir veri setinden beş yıl sonra Türkiye'nin nüfusunun ne olacağı öğrenilmek istendiğinde nüfus, şehir ve yıl etiketleri kullanılarak girdi ve çıktı arasında eşleşme yapılabilir.

Denetimli makine öğrenmesi modeli süreci Şekil 3.10'da belirtilmiştir (Afrin ve Nahar, 2015; Çoban, 2016).



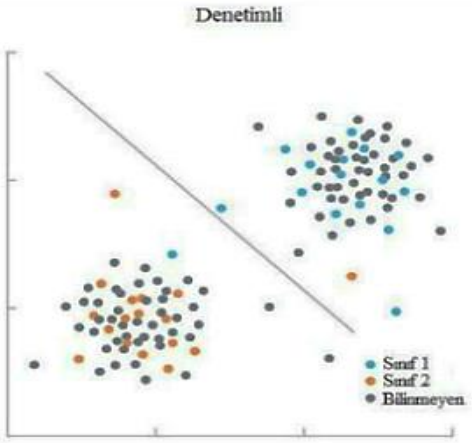
Şekil 3.10. Denetimli Makine Öğrenmesi Modeli Süreci

Denetimli makine öğrenmesi yöntemlerinde veri kümesi ön işlemden geçirilip kullanılabilir hale getirildikten sonra, eğitim ve test işlemlerinde kullanılmak üzere birden fazla alt kümeye bölünmektedir. Eğitim sürecinde algoritmalar değişik parametrelerle test edilip birbirleri ile kıyaslanarak en iyi sonucu sağlayan nihai sınıflandırıcı belirlenmektedir (Topaçan, 2016).

Denetimli makine öğrenmesinin temel amacı, girdileri ve girdilerden elde edilen sonuçları algoritmaya aktararak, algoritmanın bu verilere göre bir fonksiyon oluşturmasıdır. Fonksiyonun, girdi (x) ve çıktısı (y) daha önceden belirlenir. Burada asıl amaç, mevcut girdilerin kullanılmasıyla çıktının nasıl oluştuğunun bulunması, bir başka ifade ile tahmin edici modelin oluşturularak beklenen sonuca ulaşılmasıdır.

$$y = g(x|\theta) \tag{4}$$

Denklem (4) de g oluşturulan modeli, Θ parametreleri, y ise sonucu ifade etmektedir. Girdi ve çıktı arasındaki bağlantının tümevarım yaklaşımıyla bulunma süreci öğrenme olarak isimlendirilir. Öğrenme evresinde ortaya çıkan model ise gelecekteki sonuçların tahmin edilmesi için kullanılır (Alpaydın, 2010). Şekil 3.11'de denetimli makine öğrenmesi örneği bulunmaktadır:



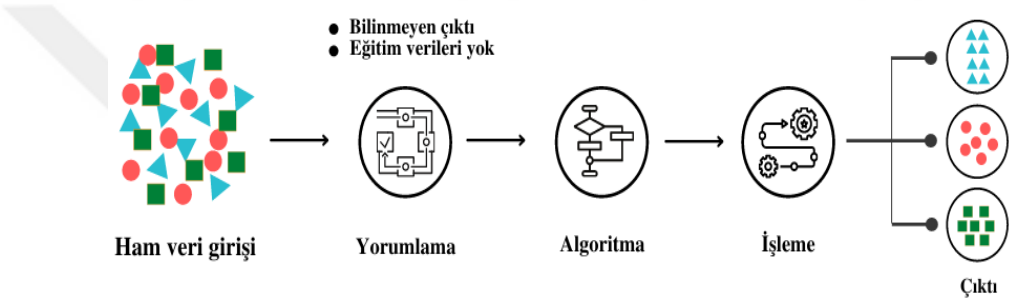
Şekil 3.11. Denetimli Makine Öğrenmesi Örneği (Deveci, 2012)

Denetimli öğrenme yöntemleri sınıflandırma araştırmalarının ana temelini oluşturmaktadır. Eğitim veri kümesi ile eğitilen algoritma test kümesi ile karşılaştırılarak ihtiyacı ne kadar karşıladığı bulunmaktadır. Sonuçların tahminlerin

altında kalması durumunda algoritma parametreleri değiştirilerek performansın arttırılması (parametre tuning) hedeflenmektedir (Aykul, 2019).

3.4.2. Denetimsiz Makine Öğrenmesi

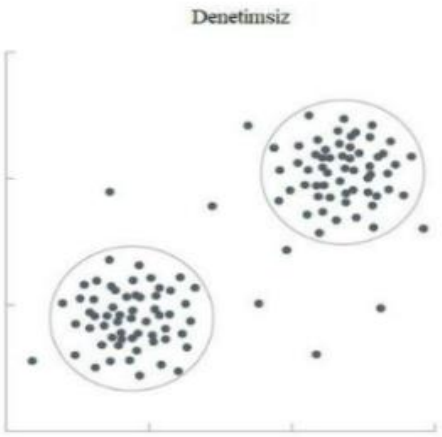
Denetimsiz makine öğrenmesi yöntemi; etiket yapılmamış asıl verideki gözle görülmeyen bağlantıların ortaya çıkarılması işlemlerinden oluşmaktadır. Denetimli makine öğrenmesi yönteminde olduğu gibi verilerin etiketleri ya da sınıfları belli değildir. Veriler benzer özelliklerine ve birbirlerine yakınlık durumuna göre gruplandırılmaktadır (Aykul, 2019).



Şekil 3.12. Denetimsiz Makine Öğrenmesi (Turhost, 2021)

Denetimsiz makine öğrenmesi modelinde sistem eğitilirken etiketsiz veri kullanılarak öğrenmesi sağlanır. Veri setindeki örneklerin çıkışları bilinmediğinden bu modelde amaç tanıma yahut sınıflandırma değil; genellikle kümeleme, olasılık yoğunluk tahmini, öznitelikler arasındaki ilişkilerin bulunması ve boyut indirgemedir. Ayrıca denetimsiz makine öğrenmesi algoritması ile elde edilen sonuçlar denetimli makine öğrenmesi için de kullanılabilmektedir (Chao, 2011). Parçalayıcı ve hiyerarşik kümeleme algoritmaları ise genellikle denetimsiz makine öğrenmesi modeli oluşturulurken kullanılan algoritmalardır (Özgür, 2004).

Denetimsiz makine öğrenmesinde en yaygın kullanılan yöntem öbekleme yöntemidir. Öbekleme yönteminde yakın özniteliklere sahip veriler aynı kümelere gelecek şekilde gruplara ayrılmaktadır. Örnek olarak bir firma müşterilerinin demografik bilgilerini elde tutarak müşterilerin profillerini görmek isteyebilir. Böyle bir durumda, öbekleme yöntemi birbirine benzeyen müşterileri aynı öbeklere aktararak müşteri kümeleri oluşturabilir. Şekil 3.13'te denetimsiz makine öğrenmesi örneği bulunmaktadır:



Şekil 3.13. Denetimsiz Makine Öğrenmesi Örneği (Deveci, 2012)

Denetimli ve denetimsiz makine öğrenmesi modellerinin kullanım amaçları arasındaki fark Tablo 3.3.'te gösterilmiştir:

Tablo 3.3. Denetimli Makine Öğrenmesi ile Denetimsiz Makine Öğrenmesi Arasındaki Farklar

	Denetimli Makine Öğrenmesi	**Denetimsiz Makine Öğrenmesi**
Temel	Etiketli veri ile ilgilenir.	Etiketlenmemiş veri ile ilgilenir.
Hesaplamalı Karmaşıklık	Yüksek	Düşük
Analyzation	Çevrimdışı	Gerçek zaman
Doğruluk	Doğru sonuçlar üretir.	Orta derecede sonuç üretir.
Alt Etki	Sınıflandırma ve regresyon	Kümeleme ve dernek kural madenciliği

Denetimli makine öğrenmesi, etiketli verilerle ilgilenirken; denetimsiz makine öğrenmesi, etiketsiz verilerle ilgilenir. Karmaşıklık söz konusu olduğunda, denetimsiz makine öğrenmesi yöntemi denetimli makine öğrenmesine göre daha karmaşıktır. Denetimsiz makine öğrenmesi, gerçek zamanlı analiz kullanırken; denetimli makine öğrenmesi, çevrimdışı analiz de yapabilmektedir. Denetimli makine öğrenmesi tekniğinin sonucu daha doğru ve güvenilirken; denetimsiz makine öğrenmesi, güvenilir fakat orta derecede güvenilir sonuçlar vermektedir. Denetimli makine öğrenmesi yöntemi ile sınıflandırma ve regresyon problem türleri çözülürken; denetimsiz makine öğrenmesi kümeleme ve ilişkisel kural madenciliği sorunlarını da içermektedir. Sonuç olarak denetimli makine öğrenmesi, sistemlere eğitim, girdi ve çıktı kalıpları sağlayarak bir görevi yerine getirme tekniğiyken; denetimsiz makine öğrenmesi, sistemin girdi

grubunun özelliklerini kendi başına keşfetmesi gereken ve önceden bir kategori grubu bulunmadığı kendi kendine öğrenme tekniğidir (Anonim, 2019).

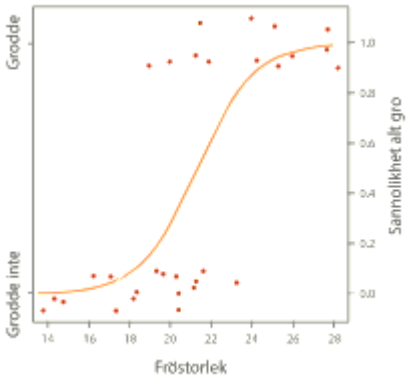
3.4.3. Makine Öğrenmesi Algoritmaları

Makine öğrenmesi algoritmaları, karmaşık veri kümelerinin keşfedilerek analiz edilmesine ve anlamlandırılmasına yardımcı olan kod parçacıkları olarak tanımlanabilir. Her algoritma bir makinenin belirli bir hedefi gerçekleştirmek için izleyebileceği sınırlı ve belirli adım adım ilerleyen yönerge kümesidir. Makine öğrenmesi modelinin hedefi, tahmin yapmak ya da bilgileri kategorilere ayırmak amacıyla kullanılabilecek desenler ortaya çıkarmaktır (Anonim, 2021).

Literatürde pek çok farklı alanda pek çok farklı makine öğrenmesi algoritması bulunmaktadır. Bu bölümde bu çalışmada kullanılan makine öğrenmesi algoritmaları açıklanmaya çalışılmıştır.

3.4.3.1. Lojistik Regresyon

(Logistic Regression-LR), 1900'lü yılların başlarında biyolojik bilimlerde sonrasında ise pek çok sosyal bilim uygulamasında kullanılan algoritma türüdür. Doğrusal sınıflandırmada kullanılan LR, çıkış verisi bir kategoriyi ifade ettiğinde tercih edilmektedir. Bir e-postanın spam olup olmayışının kategorilerinden birini sonuç olarak vermesi örnek olarak verilebilir (Hatipoğlu, 2018).



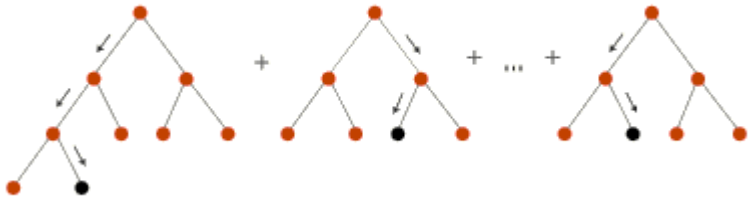
Şekil 3.14. LR

Veri işaretleme işlemlerinin ya da veri ön işlemlerinin LR tiplerine göre yapılması gerekmektedir. Buna göre üç çeşit LR tipi vardır (Swaminathan, 2018; Ballı, 2021):

1. İkili LR: Kategorik yanıtın yalnızca iki olası sonucu bulunmaktadır. Bir e-postanın spam olması ya da olmaması örnek olarak verilebilir. Bu LR tipinin kullanılması durumunda yalnızca iki kategorili (pozitif ve negatif gibi) tweetlerin verilmesi gerekmektedir.
2. Çok Terimli LR: Sıralama olmadan üç veya daha fazla kategoride sonuç verebilmektedir. Hangi tür yiyeceklerin daha çok tercih edildiğinin tahmin edilmesi (sebzeli, siyah etli, beyaz etli, vegan gibi) örnek olarak verilebilir. Bu veri tipinde duygu analizi verileri üç kategoride (pozitif, negatif, nötr) olacak şekilde verilebilmektedir.
3. Sıralı LR: Sıralama ile üç veya daha fazla kategoride sonuç vermektedir. Bir filmin beğeni olarak 1'den 5'e kadar derecelendirilmesi örnek olarak verilebilir.

3.4.3.2. Rastgele Orman Algoritması

Rastgele orman algoritması, karar ağaçlarını fazladan rastgelelik ekleyerek birleştirmektedir. Düğümleri alt düğümlere ayırma işlemini gerçekleştirir ve bunu yaparken en önemli özelliği aramaz. Bunun yerine rastgele bir özellik alt kümesi (subset of features) arasında en iyi özelliği arar. Bu, genellikle daha iyi bir modelle ve geniş bir çeşitlilikle sonuçlanmaktadır. Random Forest (RF) algoritması ile çalışılırken karar ağaçlarının (desicion tree) yaptığı gibi olası en iyi eşikleri aramak yerine her özellik için ek olarak rastgele eşikler kullanılarak ağaçlar daha da rastgele hale getirilebilmektedir. RF algoritması ile tahmin üzerindeki her bir özelliğin diğerlerine göre önemini ölçmek oldukça kolaydır (Ballı, 2021).



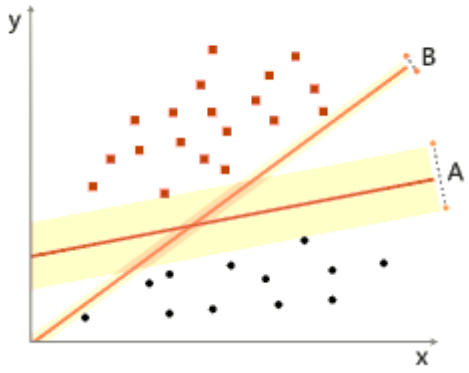
Şekil 3.15. Rastgele Orman Algoritması

RF algoritmaları, karar ağaçlarını temel almaktadır. Önce her karar ağacı kendi kararını vermekte, karar ormanı içerisinde maksimum oyu olan sınıf son karar olarak kabul edilmekte ve gelen test verisi o sınıfa dahil edilmektedir (Kaban ve Diri, 2008). RF algoritmalarında, bir ağaç oluşturmak yerine bir orman oluşturup bu ormandaki ağaçları rastgele düzenlemektedir. Ardından, test nesnesinin son sınıfını belirlemek için farklı rastgele karar ağacı biçimlerinden alınan oyları toplamaktadır.

3.4.3.3. Naive Bayes Sınıflandırması

NB sınıflandırma algoritması, Bayes teoremine dayalı, olayların gerçekleşme sıklığının hesaplandığı koşullu olasılık yöntemini kullan istatistiksel bir sınıflandırma yöntemidir. NB teoreminde, önerme sınıflandırmada kullanılacak olan her özniteliğin istatistiksel olarak bağımsız olduğu üzerine kuruludur ve her bir nitelik verilen sınıf içinde diğer niteliklerden bağımsız olarak kabul edilmektedir. Bir başka ifade ile tüm özellikler sonuç olasılığını birbirinden bağımsız olarak etkilemektedir. Yani bir özniteliğin sınıfta yarattığı etki diğer özniteliklerin var olup olmamasına bağlı değildir (Elmas, 2019; Karaöz, 2018). NB sınıflandırma algoritması basitçe, önceki bilgiler ışığında hipotez olasılığının hesaplanması için iyi bir yoldur

NB sınıflandırma algoritmasının işleyişinde öncelikle sınıfların olasılıkları ve özniteliklerin tüm durumları için sınıf değerlerine bağlı koşullu olasılıkları bulunur. Yeni gelen veriler ise ilgili sınıflar için tüm olasılıklar çarpılması ile elde edilen değerlerden hangisi daha yüksekse o sınıfa atanmaktadır (Dean, 2014).



Şekil 3.16. NB Sınıflandırması

Bayes teoremi, denklem (5) de yer alan formülle ifade edilmektedir. Formülde görülen P(h|d) sonraki olasılığı (posterior probability), d veriyi, h ise hipotezi ifade etmektedir. Sonuç ise, d' ye göre h'nin yani veriye göre hipotezin olasılığıdır. Hipotezin doğru olduğu varsayıldığında verinin olasılığı P(d|h) ile ifade edilir. P(d) ve P(h), birbirilerinden bağımsız olarak sırayla verinin olasılığı ve hipotezin doğru olma olasılığıdır. Buna önceki olasılık (prior probability) adı verilmektedir (Brownlee, 2020).

$$P(h|d) = \frac{\mathrm{P(d|h)}\ x\ P(h)}{\mathrm{P(d)}} \qquad (5)$$

Gandhi (2018) NB türlerini şöyle özetlemiştir:

- Gaussian NB Tahmincilerin değeri sürekli olduğunda ve ayrıklaştırılamadığında, bu değerlerin bir gauss dağılımından veya diğer bir ifadeyle normal dağılımdan örneklendiğini varsayılmaktadır.
- Multinomial NB Çoğunlukla belge sınıflandırma problemleri için kullanılan Multinominal NB bir belgenin spor, sağlık, teknoloji, siyaset gibi çoklu sınıf tahminlerinde kullanılmaktadır. Belgede bulunan kelimelerin sıklığı, sınıflandırıcın kullandığı özellik ya da tahminlere dayanmaktadır.
- Bernoulli NB, Multinominal NB'e benzer, ancak tahmin ediciler boolean değişkenlerdir. Sınıf değişkenini tahminlemede kullanılan parametreler yalnızca evet/hayır, iyi/kötü gibi değerler almaktadır.

NB algoritmaları öneri sistemleri, duygu analizi, spam filtreleme gibi alanlarda oldukça sık kullanılmaktadır. NB hızlı ve kolay uygulanabilir olmasına rağmen, tahmin edicilerin bağımsız olması gerekliliği en büyük dezavantajıdır. Çoğu gerçek yaşam durumlarında, tahmin ediciler bağımlıdır ve bu da sınıflandırıcının performansını olumsuz etkilemektedir (Gandhi, 2018).

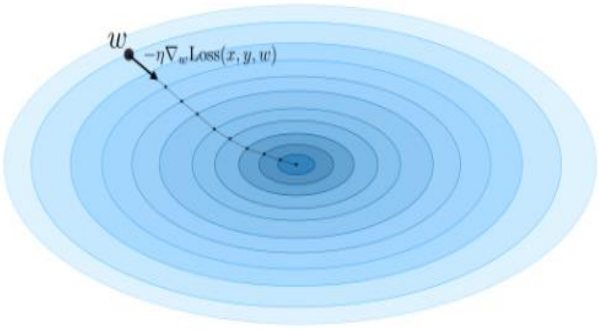
3.4.3.4. Stokastik Gradyan İniş Algoritması

Stokastik, kelime anlamı olarak rastgele bir olasılığa sahip süreç veya sistemi ifade etmektedir. Bir fonksiyonun eğimi olarak tanımlanan gradyan, bir değişkenin geçirdiği değişiklikler karşısında bir başka değişkenin geçirdiği miktarını hesaplamaktadır. Başka bir değişkenin değişikliklerine yanıt olarak bir değişkenin değişim derecesini ölçmektedir. Gradyan inişi maliyet fonksiyonunun minimum değerini bulmak için yinelemeli olarak çalışmaktadır. Stochastic Gradient Descent (SGD) algoritması, makine öğrenimi ve derin öğrenmede sıklıkla kullanılan bir algoritma türüdür. Öğrenme algoritmalarının tamamında olmamakla birlikte çoğunda kullanılabildiği söylenebilir (Ballı, 2021).

$\eta \in R$ öğrenme oranını (aynı zamanda adım boyutu olarak da bilinir) dikkate alınarak, gradyan (bayır) inişine ilişkin güncelleme kuralı, öğrenme oranı ve Loss(x,y,w) kayıp fonksiyonu ile aşağıdaki denklem (6) ile ifade edilebilir (Amidi ve Amidi, 2021) :

$$w \longleftarrow w - \eta \nabla_w \mathrm{Loss}(x, y, w) \tag{6}$$

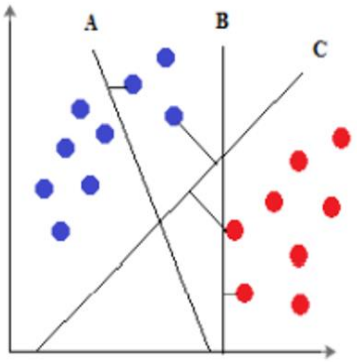
SGD algoritmasında, her bir yineleme için tüm veri seti yerine rastgele birkaç örnek seçilmektedir. Toplu gradyan iniş gibi tipik gradyan iniş algoritmalarında her iterasyonda veri setinden alınan örnek sayısı olarak tüm veri seti alınmaktadır (Ballı, 2021).



Şekil 3.17. Stokastik Gradyan İniş Algoritması

3.4.3.5. Destek Vektör Makineleri

İlk olarak 1995 yılında, Cartes ve Vapnik tarafından sınıflandırma problemlerinin çözümü amacıyla ortaya atılan destek vektör makineleri (Support Vector Machines-SVM) temelde, sınıfları birbirinden ayıran en uygun hiper düzlemin bulunması esasına dayanmakta olan makine öğrenmesi yöntemidir. Bu yöntemde, hiper düzlemin sınıfların üyelerini birbirinden ayıran en uzak mesafeyi içermesi gerekmektedir. İşte bu hiper düzlem üzerindeki sınıflara ait noktalara destek vektörleri adı verilir. Başlangıçta iki sınıflı doğrusal verilerin sınıflandırılması amaçlı tasarlanan destek vektör makineleri yöntemi, daha sonra çok sınıflı ve doğrusal olmayan verilerin sınıflandırılması amacıyla genişletilmiştir (Elmas, 2019).



Şekil 3.18. Destek Vektör Makineleri

İki sınıflı problem üzerinde destek vektör makineleri algoritmasının çalışma prensibi Şekil 3.18.'de gösterilmiştir. Şekildeki iki sınıfı A, B, C ile gösterilen üç farklı düzlem ayırmaktadır. Buna göre A düzleminin sınıfları başarılı bir şekilde ayıramadığı, B düzleminin en yüksek marjini sağlayamadığı ve C düzleminin ise hem sınıfları

başarılı bir şekilde ayırabilmekte olduğu hem de en yüksek marjini sağlamakta olduğu görülmektedir (Elmas, 2019).

3.4.4. Performans Ölçütleri

Doğruluk, duyarlılık, kesinlik, F-ölçütü gibi kavramlar sınıflandırma algoritmalarının performans ölçütünde kullanılmaktadır. Bir sınıflandırma algoritması tarafından gerçekte olan ve tahmin edilen sınıflandırmalar hakkında bilgi edinmek ve bu algoritmanın performansını özetlemek için karışıklık matrisi kullanılır. Matriste yer alan veriler ile bu algoritmaların karşılaştırılması ve performans ölçüsü değerlendirilir. Aşağıdaki Tablo 3.4'te iki sınıflı bir sınıflandırıcı için karışıklık matrisi gösterilmektedir (Albayrak, 2018).

Tablo 3.4. Karışıklık Matrisi

		Tahmin Edilen	
		Pozitif	Negatif
Gerçek	Pozitif	TP	FP
	Negatif	FN	TN

Tablo 3.4'teki TP, FP, FN, TN girdilerinin ifadesi şöyledir:

- TP (True Pozitif): Gerçekte pozitif olan ve tahmin edilenin de pozitif olduğu örnek sayısı.
- FP (False Pozitif): Gerçekte pozitif olan ama tahmin edilenin pozitif olmadığı örnek sayısı.
- FN (False Negatif): Gerçekte negatif olan ama tahmin edilenin negatif olmadığı örnek sayısı.
- TN (True Pozitif): Gerçekte negatif olan ve tahmin edilenin de negatif olduğu örnek sayısı.

Sınıflandırma performansını karşılaştırmak için kullanılan metrikler ve bunlara ait formüller aşağıdaki gibidir (Nalçakan ve ark., 2015).

Doğruluk (Accuracy): Doğru olarak sınıflandırılmış örnek sayısının (TP +TN), toplam örnek sayısına (TP+TN+FP+FN) oranıdır. Modele ait doğruluk oranı, performans ölçümünde kullanılan en basit ve popüler yöntemdir.

$$Doğruluk = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

Duyarlılık (Recall): Pozitif olarak etiketlenmiş (TP) örnek sayısının, toplam pozitif (TP+FN) örnek sayısına oranıdır.

$$Duyarlılık = \frac{TP}{TP+FN} \quad (8)$$

Kesinlik (Precision): Pozitif olarak etiketlenmiş (TP) örnek sayısının, pozitif olarak ön görülen tüm örnek sayısına (TP+FP) oranıdır.

$$Kesinlik = \frac{TP}{TP+FP} \quad (9)$$

F-Ölçütü (F-Measure): Duyarlılık ve kesinlik metrikleri tek başına anlamlı bir performans ölçütü çıkarmakta yeterli değildir. F-Ölçütü, duyarlılık ve kesinlik metrikleri kullanılarak hesaplanmaktadır.

$$Kesinlik = \frac{2 \text{ x Duyarlılık x Kesinlik}}{Duyarlılık+Kesinlik} \quad (10)$$

3.5. Duygu Analizi

Duygu Analizi (DA), makine öğrenmesi çalışmalarına sıklıkla konu olan bir kavramdır. DA insanların, bir varlık üzerindeki tutum, düşünce ve duygularının bilgisayar bilimleri kullanılarak ortaya çıkarılmasını amaçlayan bir araştırma alanıdır (Medhat ve ark., 2014). Temel olarak bir metin işleme işlemidir. Metindeki ifade edilmek istenen duyguyu belirlemeyi, kelimelerin ve ifadelerin özelliklerine dayalı

olarak metinlerin duygu içeriğini tahmin etmeyi amaçlamaktadır (Kouloumpis ve ark. 2011). Mesaj, haber, paylaşım gibi sosyal medya verisinin taşımış olduğu duygu ve düşünceyi anlam bilimsel olarak ortaya çıkarmak amacıyla kullanılır. Bunu yaparken metin duygusal kutuplara ayrılır ve duygu kategorilerinin kullanılması yerine pozitif ve negatif duygular olarak gruplandırılır. Duyguların metinlerde hangi yollarla anlatıldığının ve bu anlatımlarda olumlu ya da olumsuz durumların tespit edilmesini sağlayan bir analizdir (Nasukawa ve ark., 2003). Metinlerin ifade ettiği duyguyu sınıflandırabilmek amacıyla her metnin tek bir duygu ile etiketlenmesi veya metinlerde birden fazla duygunun etiketlenmesi gerekmektedir (Elmas, 2019).

Şekil 3.19. Duygu Sınıflandırma Teknikleri (Amanet, 2017).

DA Şekil 3.19'da belirtildiği üzere makine öğrenmesi yaklaşımı ve sözcük tabanlı yaklaşımla ya da bunların birleştirilmesi ile yapılabilmektedir.

DA'nde temel veri kaynağı daha önce ifade edildiği üzere sosyal medyadır. Bununla birlikte tartışma forumları, inceleme siteleri, bloglar, müşteri geri bildirim siteleri ve metin formatında bilgi talep eden bütün siteler duygu içeren sübjektif dokumanlar içerdiğinden DA için veri kaynağı olarak ele alınabilirler (Poria vd., 2018). Analize konu olan metinlerin bazı özellikleri taşımaları beklenmektedir. Aşağıda bu özellikler kısaca tanımlanmıştır (Cebeci, 2020):

- Duygu: Bir fikir ile doğrudan ilişkili olan his, davranış, değerlendirme gibi sübjektif unsurlardır. Duygular genel kabul görecek şekilde rasyonel olabildiği gibi kişiye özel göreceli de olabilir. Pratikte DA'nin konu alanının kişiye özel duyguların olduğu söylenebilir.

- Duygu Hedefi: Metin içerisinde ifade edilen bir duygunun yöneldiği nesneye karşılık gelir. Metin içerisindeki her bir duygu ifadesinin bir nesneye bağlanması analiz sürecinin başarısı için hayati önem taşımaktadır.
- Varlık: Duygunun ifade edildiği bir ürün, hizmet, konu, organizasyon veya insan olabilir.
- Polarite: Bir duygu ifadesinin yönünü vermektedir. DA sürecinde negatif, nötr ve pozitif olarak üçlü polarite kullanılabileceği gibi pozitiflik ve negatiflik derecesi de (5 yıldız gibi) tercih edilebilir.

Örneğin “MacBook Pro aşırı uzun batarya ömrü ile beni çok etkiledi.” şeklinde bir müşteri yorumu olduğunu varsayalım. Burada duygu ifade eden kelimeler “aşırı uzun” ve “etkiledi” ifadeleridir. Söz konusu duygu için duygu hedefi: batarya ömrüdür. MacBookPro ifadesi de varlık olarak ele alınır. DA tarafından bu cümle pozitif polariteli ya da yüksek pozitif olarak değerlendirilebilir. Eğer varlıklar ve duygu hedefleri doğru şekilde belirlenebilirse analiz başarımı yükselerek hatta topik modelleme yaklaşımları gibi özetleme yöntemleri ile farklı bakış açıları elde edilebilir.

3.6. Twitter ve Twitter API

Kullanıcıların 280 karakter sınırı ile "tweet" adı verilen gönderiler yazabildiği, güncel ve trend konular hakkında duygu ve düşüncelerini paylaşabildiği popüler bir sosyal medya platformu olan Twitter, ilk kez 2006 yılında kullanılmaya başlanmış, 2011 yılında da Türkçe desteği ile kullanıcılara sunulmuştur.

Her geçen gün kullanıcı sayısı giderek artan Twitter markaların, kamu kuruluşlarının, devlet adamlarının, sporcu ve birçok ünlü kişinin de içinde olduğu 1.3 milyar kullanıcıya ulaşarak büyük bir bilgi üretim, dağıtım ve tüketim platformu haline gelmiştir. Platform üzerinde aylık ortalama 330 milyon aktif kullanıcı yer almakta ve her gün 500 milyon tweet atılmaktadır (Ahlgren ve ark. 2021). Bu yapısı ile önemli bir haber kaynağı ve dağıtım kanalı haline gelmiştir. Birçok şirket ürün veya hizmetleri hakkında fikir sahibi olmak adına Twitteri dijital pazar olarak kullanmakta ve her gün Twitter’da ortalama 164 milyon reklam gösterilmektedir. Yine politikacılar ya da ünlü isimler platform sayesinde geniş kitlelere ulaşma imkânı bulmaktadır (Akyul, 2019).

Günümüzde artık vazgeçilmez hale gelen Twitter insanların bilgiye ulaşma, eğlence, sosyalleşme, bir durum veya ürün hakkında fikir beyan etme, toplumda

meydana gelen olaylar karşısında duygu ve düşüncelerini diğer insanlarla paylaşma gibi ihtiyaçlarını karşılamaktadır. Tweet Yapısı:

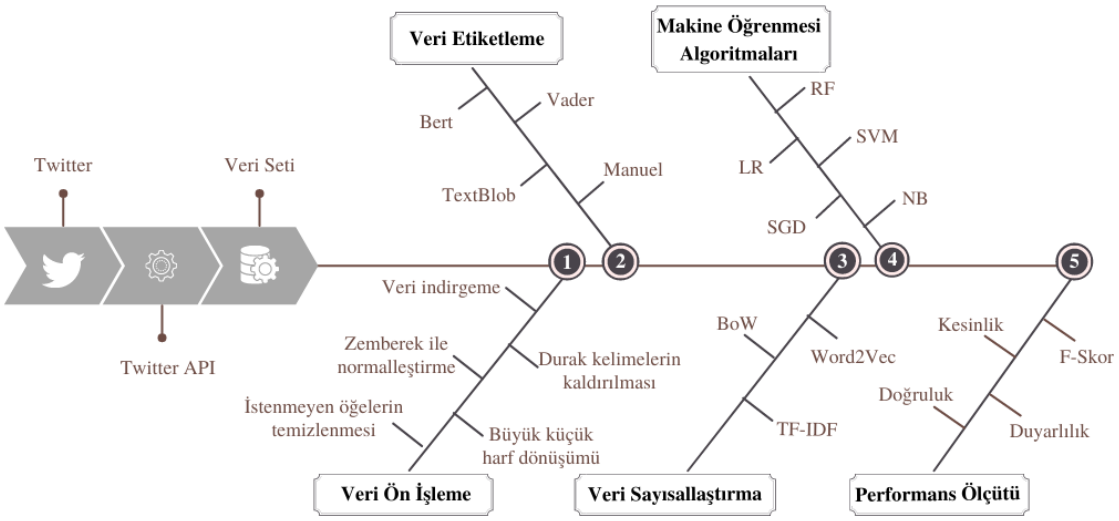
- **Tweet:** Twitter üzerinde yapılan paylaşımlara tweet adı verilmektedir. Bu paylaşımlar metin, ses, video ve görsel içerikler olarak yapılabilmektedir.
- **Retweet (RT):** RT Diğer kullanıcılar tarafından yapılan paylaşımların alıntı yapılarak kişinin kendi zaman tünelinde paylaşmasıdır.
- **Trend Topic (TT):** Twitterda gündem olan listedeki ilk 10 başlığa trend topic adı verilmektedir. Trendler ülkeden ülkeye şehirden şehire hatta bulunduğu konum itibari ile değişiklik gösterebilmektedir.
- **Hashtag (#):** Twitterde "#" işareti ile yazılan kelimelere hashtag Türkçe adıyla başlık etiketi adı verilmektedir. Herhangi bir konu veya başlığı etiketlemek için kullanılmaktadır. Twitter'da etiket ile yapılan aramalar o etikete sahip bütün tweetleri listelemektedir.
- **Mention (@):** Türkçe karşılığı bahsetmek olan mention, başka bir kişinin kullanıcı adının önüne @ işareti koyarak kullanılmaktadır. Tweet içerisinde kullanılan @kullanıcı adı ile paylaşılan tweet bahsedilen kullanıcı ve o kullanıcının takipçileri tarafından görülmektedir.
- **Direct Message (DM):** Twitter'da sıklıkla kullanılan bir diğer kısaltma ise DM ifadesidir. DM kullanıcılar arasında özel mesajlaşma sistemidir ve diğer kullanıcılar tarafından görülmemektedir.
- **Favorite (FAV):** Paylaşılan bir tweette kalp simgesine tıklanarak o tweetin beğenilmesi anlamına gelmektedir.

Twitter API Search, Rest ve Stream kütüphanelerini içeren farklı özelliklere sahiptir. Bu özellikler dışında program geliştiriciler tarafından kullanılan başka yazılımlarda bulunmaktadır. Twitter bünyesindeki verilerin yani tweetlerin elde edilmesi bu tür yazılımlarla oldukça kolay hale gelmiştir. Bunun yanı sıra farklı parametreler (tarih, dil, konum vb.) ile sorgu gönderme, Twitter mesaj paylaşımı, diğer kullanıcıların gönderdiği mesajları liste olarak görme, gibi birçok adımı sıralamak mümkündür. Yapılan bu sorgular Twitter jargonunda var olan hastag(#), mention(@), kelime ve sözcük içerebilmektedir. Fakat gizlilik ve güvenlik politikası gereği Twitter yalnızca herkesin erişimine açık olan tweetlere erişim izni vermektedir (Çoban, 2016).

4. DENEYSEL ÇALIŞMALAR VE TARTIŞMA

4.1. Çalışmanın Mimari Yapısı

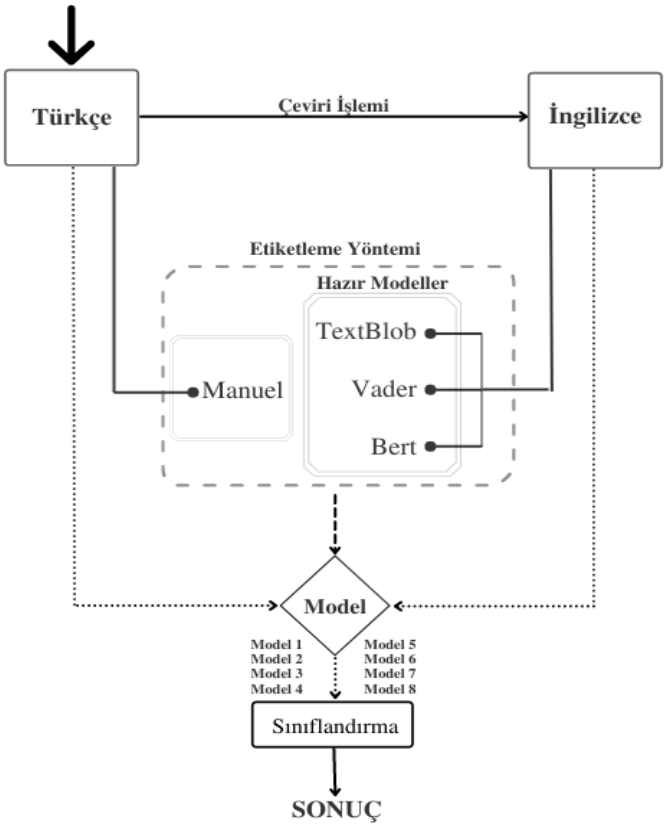
Pandemi ve beraberinde gelen uzaktan eğitim süreci, insanların bu konularda sosyal medya platformlarını kullanarak duygularını ifade etmesine ve metin madenciliği çalışmaları için büyük bir veri kaynağı oluşturmasına sebep olmuştur. Bu çalışmada uzaktan eğitim konu başlığı altında atılan Türkçe tweetler elde edilerek, yapılan paylaşımlar olumlu, olumsuz veya tarafsız olarak manuel ve önceden eğitilmiş hazır modeller (TextBlob, Vader, Bert) ile etiketlendirilmiştir. Etiketleme işlemleriyle çeşitli modeller oluşturulmuş, bu modeller ile farklı veri sayısallaştırma (BoW, TF-IDF, Word2Vec) yöntemleri ve farklı makine öğrenmesi algoritmaları (LR, SGD, SVM, RF, NB) kullanılarak en iyi performansı gösteren sınıflandırma modeli üzerinden duygu analizi amaçlanmıştır. Yapılan çalışmaya ait sistem mimarisi Şekil 4.1'deki gibidir.



Şekil 4.1. Çalışmaya Ait Sistemin Mimarisi

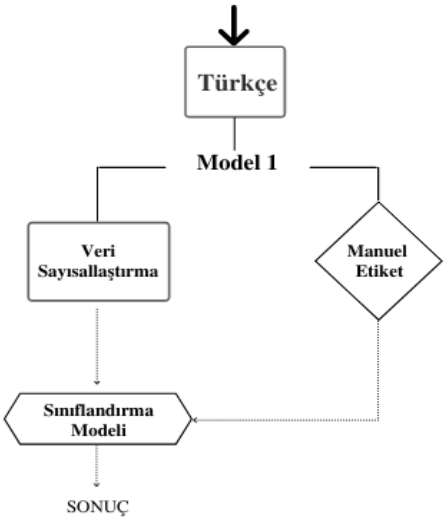
Geliştirilen sisteminin ilk aşamasında elde edilen veri seti çeşitli ön işlem adımlarına tabi tutularak daha anlamlı ve ölçülebilir bir hale gelmesi amaçlanmaktadır. Bu doğrultuda ham hali ile elde edilmiş tweetler ilk olarak küçük harfe dönüştürülmüş ardından özel karakter, emoji, rakam gibi istenmeyen öğeler temizlenmiştir. Yine bu adımda yazım yanlışları, kısaltılarak kullanılan kelimeler zemberek ile normalleştirilmiştir. Veri ön işlemenin son adımı olarak tekrar eden satırlar ve 40 karakterden küçük tweetler veri setinden çıkartılmıştır.

Makine öğrenmesi algoritmalarına girdi olarak kullanılacak veri seti için sistemin ikinci aşamasında atılan tweetin hangi duyguyu ifade ettiği belirlenerek etiketleme işlemi gerçekleştirilmektedir. Bahsedilen etiketleme işleminde Türkçe tweetlerin manuel (el ile) etiketleme işleminin yanı sıra aynı tweetlerin dil çeviri işlemi yapılarak literatürde sıkça geçen ve birçok çalışma gerçekleştirilen TextBlob, Vader ve Bert gibi hazır modeller ile etiketleme işlemi yapılmıştır. Böylece farklı bir yaklaşımla manuel etiketleme işlemine kıyasla kullanılan hazır modellerin etiketleme işlemindeki başarısı, ayrıca dil çeviri işleminin duygu analizi gibi çalışmalara etkisini ölçmek amaçlanmıştır. Bu aşamada 5043 adet tweet birinci aşamada uygulanan veri ön işleme adımlarının ardından İngilizceye çevirisi yapılarak bahsedilen farklı etiketleme yöntemleri ile modeller oluşturulmuştur. Oluşturulan bu modellerin her biri ayrı ayrı makine öğrenmesi algoritmalarına girdi olarak kullanılarak sınıflandırma performansları karşılaştırılmıştır. Veri setini oluşturan metinler ve bu metinlerin etiketleme yöntemi ile kurgulanan modelleme işlemine ait görsel şekil 4.2'deki gibidir.



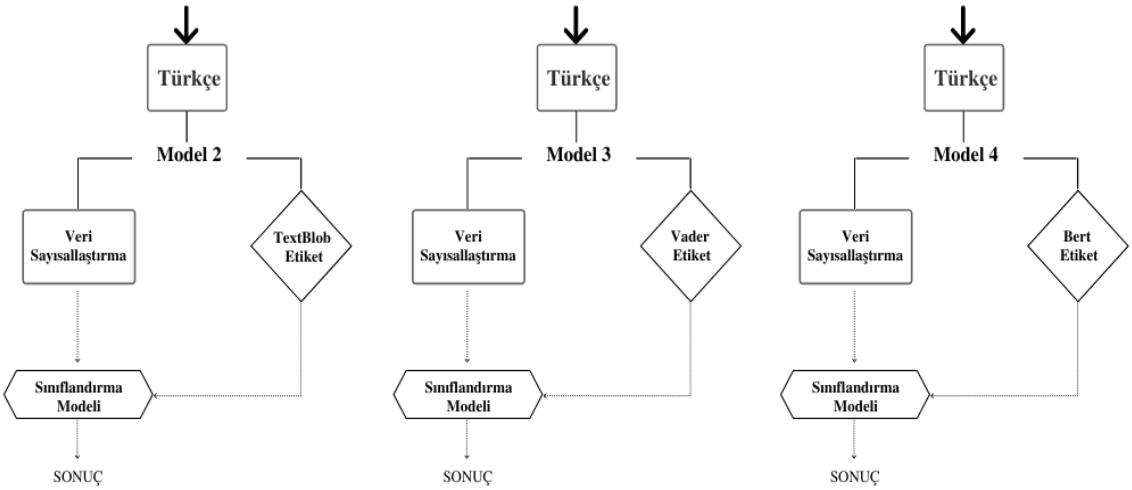
Şekil 4.2. Modelleme İşlemi

Model 1'de veri seti Türkçe metinlerden, etiketler Türkçe metinler üzerinden işaretlenmiş manuel etiketlerden oluşmaktadır. Model 1'in kullanımına ait görsel 4.3'teki gibidir.



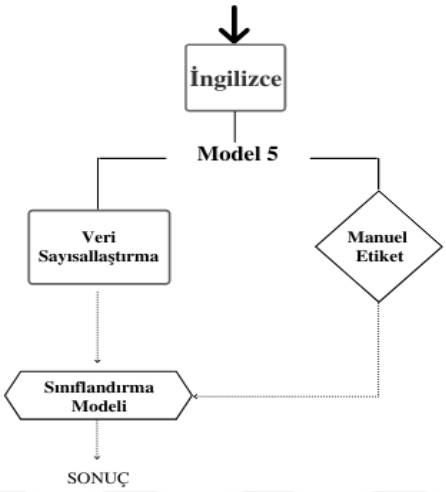
Şekil 4.3. Model 1

Model 2, Model 3 ve Model 4'te veri seti Türkçe metinlerden, bu metinlerin etiketleri ise İngilizceye çevirisi yapılmış metinlerin TextBlob, Vader ve Bert çıktılarından oluşmaktadır. Model 2, Model 3 ve Model 4'ün kullanımına ait görsel 4.4'teki gibidir.



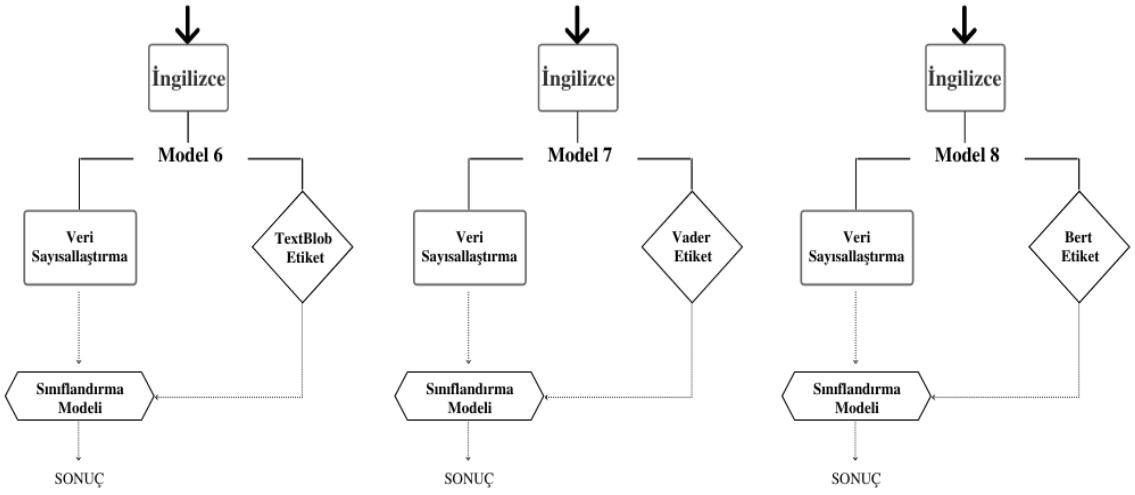
Şekil 4.4. Model 2, Model 3, Model 4

Model 5'te veri seti İngilizceye çevrilmiş metinlerden, etiketleri ise Türkçe metinler üzerinden işaretlenmiş manuel etiketlerden oluşmaktadır. Model 5'in kullanımına ait görsel 4.5'teki gibidir.



Şekil 4.5. Model 5

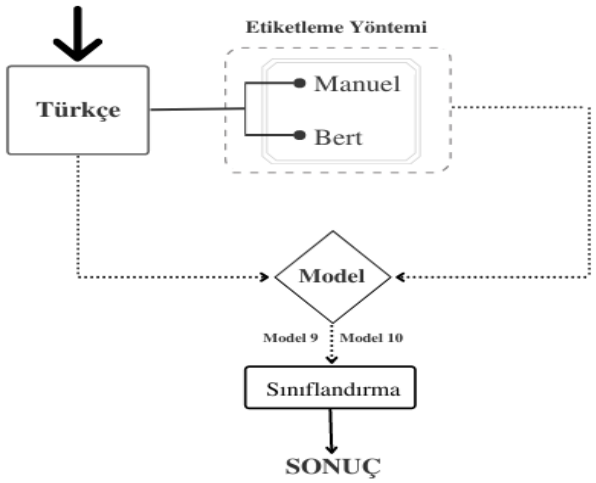
İngilizce metinlerden oluşan Model 6, Model 7 ve Model 8'de etiketler İngilizce çevrisi yapılmış metinlerin TextBlob, Vader ve Bert çıktılarından oluşmaktadır. Model 6, Model 7 ve Model 8'in kullanımına ait görsel 4.6'daki gibidir.



Şekil 4.6. Model 6, Model 7, Model 8

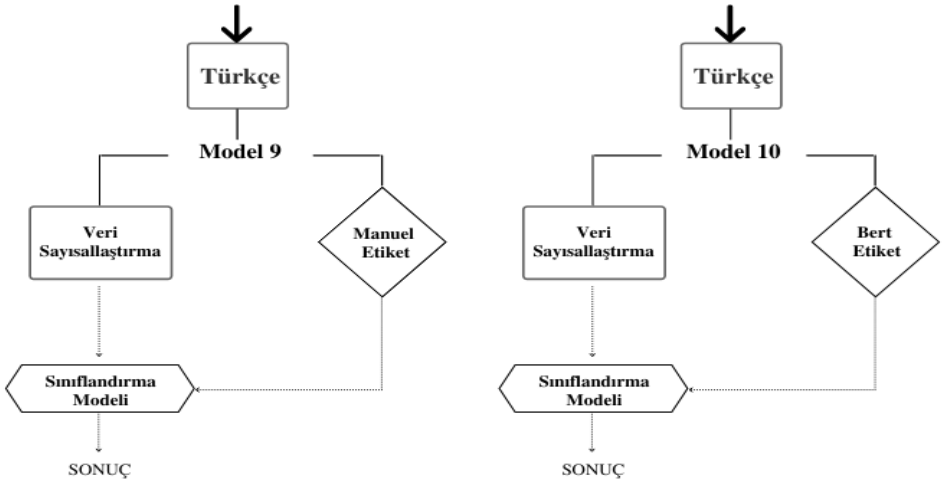
Oluşturulan ilk sekiz modelde dil çeviri işlemi ve beraberinde farklı etiketleme yöntemleri ile metinlerin negatif, pozitif veya nötr duygu etiketleri kullanılmıştır. Oluşturulan sonraki iki modelde ise manuel olarak işaretlenen nötr etikete sahip tweetler veri setinden çıkartılarak sadece negatif ve pozitif etiketler kullanılmıştır. Nötr etikete

sahip tweetler veri setinden çıkarıldıktan sonra 4514 adet tweet Bert'in Türkçe Modeli ile tekrar etiketlendirilmiştir. Böylece Türkçe metinlerde duygu çıktıları üreten bu modelin çalışma kapsamında manuel etiketlemeye kıyasla sınıflandırma performansını ölçmek amaçlanmıştır. Manuel ve Bert'in Türkçe modeli ile modelleme işlemi şekil 4.7'deki gibidir.



Şekil 4.7. Modelleme İşlemi

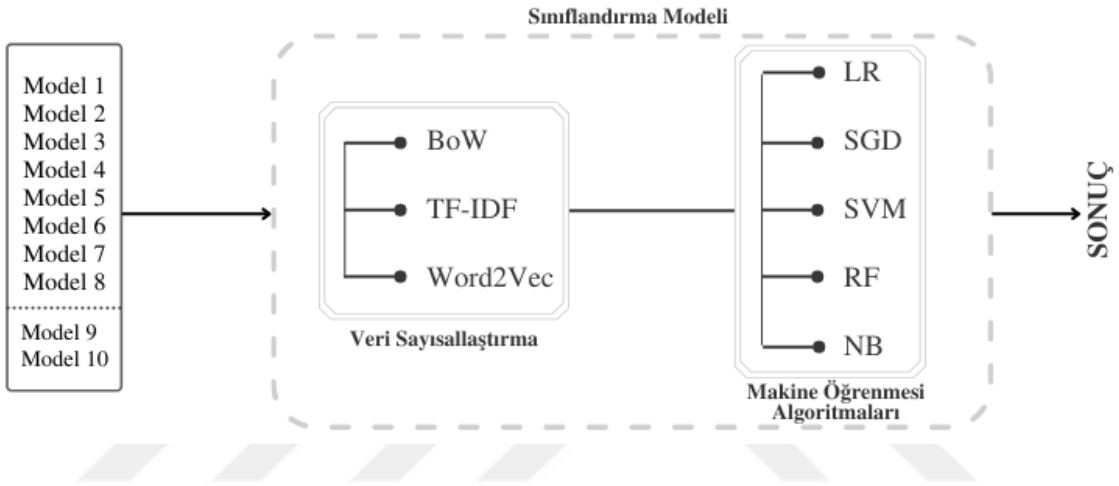
Model 9 Türkçe metinlerden ve manuel olarak işaretlenmiş negatif ve pozitif etiketlerden oluşmaktadır. Model 10 ise Türkçe metinler ve bu metinler için negatif ve pozitif duygu çıktıları üreten Bert'in Türkçe modeline ait etiketlerden oluşmaktadır. Model 9 ve Model 10'un kullanımına ait görsel şekil 4.8'deki gibidir.



Şekil 4.8. Model 9, Model 10

Üçüncü aşamada BoW, TF-IDF ve Word2Vec ile metinler sayısal ifadelere dönüştürülerek temsil edilmektedir. Bu aşamada farklı sayısallaştırma yöntemlerinin makine öğrenmesi algoritmaları ile kullanımında sınıflandırmaya etkisi amaçlanmıştır.

Sistemin dördüncü adımında, oluşturulan 10 farklı modelin farklı sayısallaştırma yöntemleri ve farklı makine öğrenmesi algoritmaları (LR, SGD, SVM, RF, NB) ile birlikle kullanılarak karşılaştırılması ve beşinci adımda da bu karşılaştırma sonucunda performans ölçütlerine bakılarak en iyi sınıflandırma modelinin tespit edilmesi amaçlanmıştır. Sınıflandırma modeline ait görsel Şekil 4.9'daki gibidir.



Şekil 4.9. Sınıflandırma Modeli

4.2. Çalışmada Kullanılan Veri

Covid-19 pandemisi ile birlikte eğitim öğretim alanında alınan tedbirler kapsamında Mart 2020'de ülke genelinde uzaktan eğitim sürecine geçilmiştir. Bu durum beraberinde öğrenci ve velilerin sosyal medya platformlarında uzaktan eğitim konu başlığı altında duygu ve düşüncelerini ifade etmelerine sebep olmuştur. Veri setinin elde edilmesinde zengin metin içerikleri ve popülerliği ile Twitter tercih edilmiş, uzaktan eğitim konu başlığı altında atılan Türkçe tweetler elde edilerek duygu analizinin ilk adımı veri seti oluşturulmuştur.

4.3. Çalışma Ortamı ve Paketler

Son yıllarda makine öğrenimi ve veri biliminde oldukça popüler olan R, python, java gibi birçok programlama dili kullanılmaktadır. Bu çalışmada dinamik, hızlı,

yerleşik kütüphaneleri ve veri madenciliği kabiliyetleri ile programlama dili olarak python tercih edilmiştir. Genel amaçlı bir programlama dili olan python, 1991 yılında Guido van Rossum tarafından ilk sürümü ortaya çıkarılmıştır. Kullanım kolaylığı ve geniş kütüphane desteği ile günümüzde oldukça popülerleşmiş ve geniş bir kullanıcı kitlesine sahip olmuştur (Malkoç, 2012).

Python kodlaması için localde (PyCharm, SPYDER, PYDEV vb. ideler) veya bulut sistemlerde (Jupyter Notebook, Google Colab vb.) kullanabilecek birçok geliştirme ortamı mevcuttur. Geliştirme ortamı olarak da Google Colaboratory (Colab), tercih edilmiştir. Ortam kurulumu gerektirmeden önceden yapılandırılmış hazır paketler ile Graphic Processing Unit (GPU) ve Tensor Processing Unit (TPU) kullanımına imkan sağlayan Colab, makine öğrenimi gibi çalışmalarda oldukça yaygın olarak kullanılmaktadır. Aynı zamanda ücretsiz bir bulut hizmeti olan Colab drive ve github gibi ortamlara kolay paylaşım imkânı da sunmaktadır.

Python içerisinde kolaylıkla kullanılabilen ve geliştiricilerin veri madenciliği, makine öğrenmesi gibi çalışmalarda sıklıkla kullandığı Tensorflow, Keras, Pandas, Numpy, Snscrape, SciKit-Learn, NLTK, Matplotlib gibi açık kaynaklı birçok kütüphane mevcuttur. Bu tez çalışmasında kullanılan bazı python kütüphaneleri aşağıdaki gibidir:

- **NLTK:** DDİ ve metin madenciliğinde popüler olarak kullanılan bir python kütüphanesidir. Kütüphane içerisinde barındırdığı paketler sayesinde köke indirgeme, durak kelimeler, metni kelimelere ayırma gibi işlemler kolaylıkla uygulanabilmektedir. Duygu analizi, sınıflandırma, kaynak oluşturma, ayrıştırma, etiketleme, anlamsal akıl yürütme ve otomatik özet oluşturma gibi insan diliyle ilgili diğer birçok işlem için ayrıca NLTK kütüphanesinden faydalanılabilmektedir (Mehta, 2021).
- **Numpy:** Matrisler ve çok boyutlu dizi nesneleri üzerinde sıralama, şekil işleme, matematiksel ve mantıksal işlemlerde çeşitli kolaylıklar sağlayan, bilimsel hesaplamalar için temel bir python kütüphanesidir. Kütüphane ayrıca dizi formatındaki matematiksel işlemlerin yürütme süresini hızlandırır, vektörlere dönüştürülme performansını artırır (Anonymous, 2021).
- **Pandas:** İlişkisel veya etiketlenmiş verilerle kolay ve sezgisel olarak çalışabilmeyi sağlamak için geliştirilen pandas, veri yapılarının dataframe nesnelerine dönüştürülmesine, dataframe üzerinde sütunların oluşturulması

ve kaldırılmasına, eksik verilerin işlenmesine olanak tanır. Veri ön işleme, veri yönetimi ve analizi için hızlı ve kolay bir araçtır (Anonymous, 2021).

- **Matplotlib:** Veri görselleştirmelerinde kullanılan standart bir python kütüphanesidir. Gelişmiş görselleştirmeler için yetersiz kalan matplotlib sağladığı nesne yönelimli API ile iki boyutlu diyagram ve grafikleri uygulamalara gömmeyi sağlar (Custer, 2020).
- **SciKit-Learn:** Makine öğrenimi algoritmalarını (doğrusal regresyon, LR, karar ağaçları, rastgele orman gibi) içeren bir python kütüphanesidir. Bu kütüphanenin bu kadar popüler olmasının nedeni ihtiyaç duyulan temel yöntemlerin büyük çoğunluğunu içeriyor olmasıdır. Scikit-learn ile veride yer alan eksik değerleri doldurmak, özniteliklere karar vermek, çapraz doğrulama yapmak, sonuç çıktılarını değerlendirmek gibi veri analitiği çalışmalarının baştan sona yürütülmesini mümkün kılmaktadır (Yüceoğlu, 2017).
- **Snscrape:** Twitter geliştirici hesabına ihtiyaç duyulmadan, herhangi bir istek sınırı ya da kısıtlama olmaksızın tweetleri kazıyabileceğimiz bir kütüphanedir. Bu kütüphaneyi kullanabilmek için Python 3.8 veya üzeri sürüm gerekmektedir (Beck, 2020).
- **Seaborn:** İstatistiksel modelleri görselleştirmek, verileri özetlemek ve genel dağılımlarını göstermek için makine öğrenimi aracı olarak hizmet eden bir python kütüphanesidir. Matplotlib'i temel alan bu kütüphane geniş bir görselleştirme galerisi sunmaktadır (Custer, 2020).
- **Zemberek-Python:** 2005 yılında Özgür yazılım ödülünü alan zemberek, açık kaynak Türkçe DDİ kütüphanesidir. Java ortamında geliştirilen bu kütüphane, yazım denetimi, yanlış kelime önerme, heceleme, kök bulma gibi özellikleri içerisinde barındırmaktadır (Anonim, 2021). Python içerisinde yer alan zemberek tabanlı zemberek-python kütüphanesi, java ortamı gerektirmeden tamamen python ile geliştirilmiştir (Anonim, 2021).

4.4. Twitter Verilerine Erişim

Twitter üzerinden verilerin çekilebilmesi için Twitter geliştirici hesabı ile Twitter API'ye erişim sağlamak gerekmektedir. 2018 yılına kadar API'ye erişip tweetleri çekmek mümkündü ancak 2018 yılından sonra geliştirici hesabı için Twittere

başvuruda bulunup, başvurunun onaylanması için de belirli bir süre beklemek gerekmektedir. Onay işleminden sonra API'ye erişimde python dili için geliştirilmiş "tweepy" kütüphanesi kullanılarak uygulamaya özel Consumer Key, Consumer Screet, Access Token ve Acces Token Secret değerleri ile Twitter bağlantısı sağlanarak veriler çekilebilmektedir. Twitter API anlık olarak son yedi günün verilerine erişim izni vermekle birlikte her 15 dakikada 180 istek gibi bazı kısıtlamalarda koymaktadır.

Bu tez çalışmasında veri setinin büyüklüğü, çekilmek istenen verinin zaman aralığı göz önüne alındığında bahsedilen kısıtlamalardan bağımsız dilediğimiz kadar tweetin çekilebildiği snscrape kütüphanesi tercih edilerek, istenilen tarihler arasında anahtar kelimeler girilip, alıntı olmayan Türkçe tweetler kullanılmıştır.

Veri setinin oluşturulmasında snscrape kütüphanesi ile Twitter'a erişim sağlanarak, uzaktan eğitim konu başlığı altında 'uzaktanegitim', 'onlineeğitim', "Uzaktan Eğitim", "öğrencine güven", "online istiyor", "ödev istiyor", "uzaktan öğretim", anahtar kelimeleri girilerek alıntı yapılmamış tweetler elde edilmiştir. Veri seti içerisinde konu ile alakalı olmayan örneğin -bedelli, -ReisBedelli, -BedelliÇözümü, -GaziENT, -papara gibi gündem oluşturmak için atılan uzaktaneğitim etiketli tweetler filtrelenerek veri setinin dışında tutulmuştur. 2020-01-01 ve 2021-11-29 tarihleri arasında atılan toplamda 162,227 tweet KVKK/GDPR kapsamında sadece id, tweets, date başlıkları altında toplanarak veri seti oluşturulmuştur. Veri setine ait özet bilgi şekil 4.10'daki gibidir.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 162227 entries, 0 to 162226
Data columns (total 3 columns):
 #   Column  Non-Null Count   Dtype 
---  ------  --------------   ----- 
 0   id      162227 non-null  int64 
 1   tweets  162227 non-null  object
 2   date    162227 non-null  object
dtypes: int64(1), object(2)
memory usage: 3.7+ MB
```

Şekil 4.10. Veri Setine Ait Özet Bilgi

Veri setinin içeriğine bakıldığı zaman elde edilen tweetlere ait id, tweet ve zaman damgası Şekil 4.11'deki gibidir.

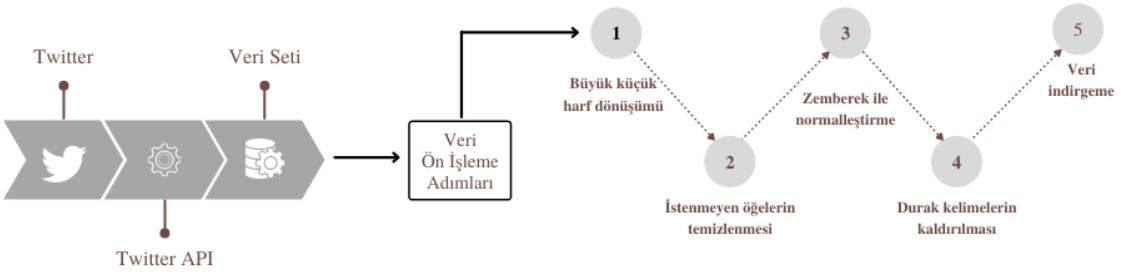
	id	tweets	date
0	1	Uzaktan eğitim gelmez, ekonomi kötü diyenler v...	2021-11-29 09:25:54
1	2	Yetoo ya yetooo artık uzaktan eğitim gelsin\n#...	2021-11-29 09:23:54
2	3	Uzaktan eğitim kararı hiç beklemediğimiz bir a...	2021-11-29 09:08:06
3	4	Yurtta okuyan arkadaşım halsiz ve boğazı ağrıy...	2021-11-29 08:27:03
4	5	Uzaktan eğitim sonrası gireceğim ilk sınav...\...	2021-11-29 08:24:05
...	...	...	...
162222	162223	Sosyal bilgiler öğretmenliği okuyoruz inkılap ...	2020-01-01 21:49:01
162223	162224	Uzaktan eğitim sınavlarına çalışmak istemiyoru...	2020-01-01 21:04:52
162224	162225	okul olmayınca hayat o kkkkadar güzel ki :((❤...	2020-01-01 14:32:44
162225	162226	şu uzaktan eğitim hocalarına bi tablet verin a...	2020-01-01 13:38:06
162226	162227	Kardeşim evde sürekli ödev yapılması söylendiğ...	2020-01-01 08:38:00

162227 rows × 3 columns

Şekil 4.11. Veri Setine Ait Örnek Tweetler

4.5. Veri Ön İşleme

Veri setinde yer alan ham hali ile elde edilmiş tweetlerde birçok anlamsız ifade, gürültülü veri yer almaktadır. Veri setinde yapısal olamayan verilerle analitik işlemlerin yapılabilmesi için işlenebilir, ölçülebilir ve sayısallaştırılabilir hale getirilmesi gerekmektedir. Bunun için de kelimelerin mümkün olduğunca yalın hale getirilmesi frekanslarının ölçülebilmesi için önemlidir. Çalışmada uygulanan veri ön işleme adımları şekil 4.12'deki gibidir.



Şekil 4.12. Veri Ön İşleme Adımları

Bu aşamada elde etmiş olduğumuz tweetler özel karakterler, noktalama işaretleri, emojiler, büyük küçük harf dönüşümü, etkisiz kelimelerin (durak kelimeler-stopwrods) elenmesi gibi veri ön işleme adımlarına tabi tutulmaktadır. Bu işlemlere ek olarak veri setinin daha anlamlı ve yalın hale gelmesi için zemberek ile normalleştirme işlemi yapılmış, tekrar eden satırlar tespit edilerek teke düşürülmüş, hastag ve mention ifadeler (#uzaktanegitim #meb @alican @ktun vb.) ile birlikte veri seti içerisinde en

fazla geçen üç kelime, boş satırlar ve 40 karakterden küçük tweetler veri setinden tamamen çıkarılmıştır.

4.5.1. Büyük Küçük Harf Dönüşümü

Metin içerisindeki büyük harflerin küçük harflere örneğin "UZAKTAN" kelimesinin anlamsal bir değişikliğe uğramadan "uzaktan" olarak dönüştürülmesi işlemidir. Yapılan büyük küçük harf dönüşümü işlemi Şekil 4.13'te gösterilmiştir.

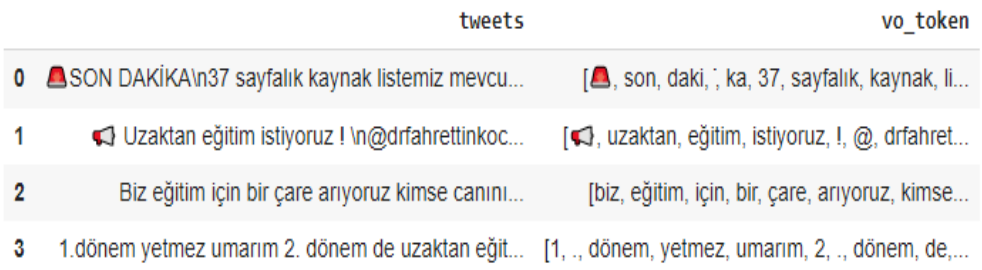
	tweets	vo_kucuk_harf
0	🚨SON DAKİKA\n37 sayfalık kaynak listemiz mevcu...	🚨son dakika\n37 sayfalık kaynak listemiz mevc...
1	📢 Uzaktan eğitim istiyoruz ! \n@drfahrettinkoc...	📢 uzaktan eğitim istiyoruz ! \n@drfahrettinkoc...
2	Biz eğitim için bir çare arıyoruz kimse canını...	biz eğitim için bir çare arıyoruz kimse canını...
3	1.dönem yetmez umarım 2. dönem de uzaktan eğit...	1.dönem yetmez umarım 2. dönem de uzaktan eğit...

Şekil 4.13. Örnek Büyük Küçük Harf Dönüşümü

4.5.2. İstenmeyen Öğelerin Temizlenmesi

- **Metni İstenilen Biçimde Ayırma (Tokenization)**

Metinleri istenilen biçimde parçalanıp dizilere kaydetme işleminin yapıldığı bu adımda, NLTK kütüphanesi içerisinde yer alan "WordPunctTokenizer" fonksiyonu ile tweetler kelime kelime, tüm özel karakter ve noktalama işaretleri ile birlikte ayırılmıştır. Örnek ayrıştırma işlemi şekil 4.14'teki gibidir.



	tweets	vo_token
0	🚨SON DAKİKA\n37 sayfalık kaynak listemiz mevcu...	[🚨, son, daki, ̇, ka, 37, sayfalık, kaynak, li...
1	📢 Uzaktan eğitim istiyoruz ! \n@drfahrettinkoc...	[📢, uzaktan, eğitim, istiyoruz, !, @, drfahret...
2	Biz eğitim için bir çare arıyoruz kimse canını...	[biz, eğitim, için, bir, çare, arıyoruz, kimse...
3	1.dönem yetmez umarım 2. dönem de uzaktan eğit...	[1, ., dönem, yetmez, umarım, 2, ., dönem, de,...

Şekil 4.14. Örnek Metin Ayrıştırma İşlemi

- **Noktalama İşaretleri, Etiket ve Kişilerin Temizlenmesi**

Bu adımda noktalama işaretleri ile birlikte tweet içerisinde etiketlenen konu başlıkları ve bahsedilen kişiler (#uzaktanegitim #meb @alican @ktun gibi) metinlerden tamamen kaldırılmıştır. Böylece metinlerin daha sade ve anlamlı hale gelmesi sağlanmıştır. Yapılan işlemden sonra ortaya çıkan örnek veri şekil 4.15'deki gibidir.

	tweets	vo_hastag_mention_kaldir
0	🚨SON DAKİKA\n37 sayfalık kaynak listemiz mevcu...	🚨son dakika 37 sayfalık kaynak listemiz mevcu...
1	📢 Uzaktan eğitim istiyoruz ! \n@drfahrettinkoc...	📢 uzaktan eğitim istiyoruz
2	Biz eğitim için bir çare arıyoruz kimse canını...	biz eğitim için bir çare arıyoruz kimse canını...
3	1.dönem yetmez umarım 2. dönem de uzaktan eğit...	1 dönem yetmez umarım 2 dönem de uzaktan eğiti...

Şekil 4.15. Örnek Noktalama İşaretleri, Etiket ve Kişilerin Temizlenmiş Hali

- **Özel Karakter, Emoji, Url ve Rakamların Temizlenmesi**

Noktalama işareti, kişi ve etiketlerin ön işleminden sonra tweet içerisinde yer alan rakam, özel karakter, emoji, link gibi ifadelerin temizlenmesi gerekmektedir. Temizleme işleminin ardından veri şekil 4.16'daki gibidir.

	tweets	vo_ozel_karakter_rakam_kaldir
0	🚨SON DAKİKA\n37 sayfalık kaynak listemiz mevcu...	son dakika sayfalık kaynak listemiz mevcut uz...
1	📢 Uzaktan eğitim istiyoruz ! \n@drfahrettinkoc...	uzaktan eğitim istiyoruz
2	Biz eğitim için bir çare arıyoruz kimse canını...	biz eğitim için bir çare arıyoruz kimse canını...
3	1.dönem yetmez umarım 2. dönem de uzaktan eğit...	dönem yetmez umarım dönem de uzaktan eğitim ...

Şekil 4.16. Örnek Özel Karakter, Emoji, Url ve Rakamların Temizlenmesi

4.5.3. Zemberek ile Normalleştirme

Bu aşamada daha sade ve anlamlı bir hale gelen veri seti için "zemberek-python" kütüphanesi kullanılarak yazım hatalarının giderilmesi ve kelime tahmini ile

tweetlerde normalleştirme işlemi gerçekleştirilmiştir. Normalleştirme işlemine ait örnek çıktı şekil 4.17'deki gibidir.

	tweets	normallestirilmis_hali
0	yrn okular acılacakmış hayirlisi bakalım	yarın okular açılacakmış hayırlısı bakalım
1	cok güzel dersten kaldık bu dönemde	çok güzel dersten kaldık bu dönemde
2	arkadaslar savasi yenicez sadece biraz daha sabir	arkadaşlar savaşı yeneceğiz sadece biraz daha ...
3	uzakdan egitm gelmeli bence okullar açılmamai	uzaktan eğitim gelmeli bence okullar açılmaması

Şekil 4.17. Zemberek ile Normalleştirme Örneği

4.5.4. Durak Kelimelerin Kaldırılması

Ön işlemenin bu adımında metin içinde sıkça geçen, önemli anlamlara sahip olmayan durak kelimeler çıkartılmıştır. NLTK kütüphanesi içerisinde farklı diller için hazırlanmış durak kelimeler yer almaktadır. Türkçe kelimeler için hazırlanmış "stopwords.words('turkish')" metodu ile tweetlerdeki durak kelimeler çıkarılmıştır. Şekil 4.18'de Türkçe için hazırlanmış durak kelimelerin listesi, Şekil 4.19'da da örnek veri setinden durak kelimelerin çıkartılmış hali gösterilmiştir.

```
['acaba', 'ama', 'aslında', 'az', 'bazı', 'belki', 'biri', 'birkaç', 'birşey',
 'biz', 'bu', 'çok', 'çünkü', 'da', 'daha', 'de', 'defa', 'diye', 'eğer', 'en',
 'gibi', 'hem', 'hep', 'hepsi', 'her', 'hiç', 'için', 'ile', 'ise', 'kez', 'ki',
 'kim', 'mı', 'mu', 'mü', 'nasıl', 'ne', 'neden', 'nerde', 'nerede', 'nereye',
 'niçin', 'niye', 'o', 'sanki', 'şey', 'siz', 'şu', 'tüm', 've', 'veya', 'ya', 'yani']
```

Şekil 4.18. Türkçe Durak Kelimeler

	tweets	vo_dk_cikarilmis
0	🚨SON DAKİKA\n37 sayfalık kaynak listemiz mevcu...	dakika sayfalık kaynak listemiz mevcut uzaktan...
1	📢 Uzaktan eğitim istiyoruz ! \n@drfahrettinkoc...	uzaktan eğitim istiyoruz
2	Biz eğitim için bir çare arıyoruz kimse canını...	eğitim çare arıyoruz kimse canını pazarda bulm...
3	1.dönem yetmez umarım 2. dönem de uzaktan eğit...	dönem yetmez umarım dönem uzaktan eğitim olur ...

Şekil 4.19. Örnek Metinlerden Durak kelimeler çıkarılmış hali

4.5.6. Veri İndirgeme

- **Tekrar Eden Satırların Silinmesi**

Veri seti içinde farklı kişiler tarafından atılmış fakat aynı içeriği sahip tweetler sadece bir adet kalacak şekilde veri setinden çıkarılmıştır. Satır silme işleminin ardından tekrar eden 26,946 tweet veri setinden çıkartılarak 135,281 tweet elde edilmiştir. Tekrar eden satırların çıkarılmasının ardından veri setine ait özet bilgi şekil 4.20'deki gibidir.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 135281 entries, 0 to 162226
Data columns (total 2 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   tweets        135281 non-null  object
 1   clean_tweets  135281 non-null  object
dtypes: object(2)
memory usage: 3.1+ MB
```

Şekil 4.20. Veri Setine Ait Özet Bilgi

- **İstenmeyen Kelimelerin Kaldırılması**

Bu aşamada veri seti içerisinde yer alan ve çok fazla geçen kelimeler tespit edilmiştir. Bir kelimenin veri seti içinde çok fazla geçiyor olması duygu analizi gibi çalışmalarda analiz sonuçlarını etkileyebilmektedir. Bu nedenle veri seti içinde en çok geçen "uzaktan, 134.651", "eğitim, 133.559", "kadar, 12564" kelimeleri metinlerden kaldırılmıştır.

Veri ön işleme adımları sırasıyla uygulanarak ham hali ile elde edilmiş 162,227 tweet işleme tabi tutulmuştur. Uygulanan bu işlemlerin ardından veri seti içerisinde boş satırların ve kısa cümlelerin yer aldığı gözlemlenmiştir. Son olarak oluşan bu boş satırlar ve 40 karakterden kısa olan tweetler veri setinden çıkartılarak 115,134 tweet kullanılabilir hale getirilmiştir. Veri setinin son haline ait örnek gösterim şekil 4.21'deki gibidir.

	tweets	clean_tweets	result_tweets
0	Uzaktan eğitim gelmez, ekonomi kötü diyenler v...	uzaktan eğitim gelmez ekonomi kötü diyenler dü...	gelmez ekonomi kötü diyenler düşünüyorsunuz
3	Yurtta okuyan arkadaşım halsiz ve boğazı ağrıy...	yurtta okuyan arkadaşım halsiz boğazı ağrıyor ...	yurtta okuyan arkadaşım halsiz boğazı ağrıyor ...
4	Uzaktan eğitim sonrası gireceğim ilk sınav...\...	uzaktan eğitim sonrası gireceğim sınav heyecan...	sonrası gireceğim sınav heyecan kağıda dokunuş...
9	uzaktan tanıdığım birinin kapısı önündeki aya...	uzaktan tanıdığım birinin kapısı önündeki ayak...	tanıdığım birinin kapısı önündeki ayakkabıları...
14	Uzaktan eğitim istiyoruz bukadar net#MebYökKur...	uzaktan eğitim istiyoruz bukadar netmebyökkurt...	istiyoruz bukadar netmebyökkurtuluşuzaktaneğitim
...	...	...	...
162222	Sosyal bilgiler öğretmenliği okuyoruz inkılap ...	sosyal bilgiler öğretmenliği okuyoruz inkılap ...	sosyal bilgiler öğretmenliği okuyoruz inkılap ...
162223	Uzaktan eğitim sınavlarına çalışmak istemiyoru...	uzaktan eğitim sınavlarına çalışmak istemiyoru...	sınavlarına çalışmak istemiyorum skyrim oynama...
162224	okul olmayınca hayat o kkkkadar güzel ki :((💔...	okul olmayınca hayat kkkkadar güzel uzaktan eğ...	okul olmayınca hayat kkkkadar güzel alabilme h...
162225	şu uzaktan eğitim hocalarına bi tablet verin a...	uzaktan eğitim hocalarına tablet verin saat fa...	hocalarına tablet verin saat fareyle yazmaya ç...
162226	Kardeşim evde sürekli ödev yapılması söylendiğ...	kardeşim evde sürekli ödev yapılması söylendiğ...	kardeşim evde sürekli ödev yapılması söylendiğ...

115134 rows × 3 columns

Şekil 4.21. Veri Setinin Son Hali

4.6. Veri Etiketleme

Twitter'dan elde edilen veriler ön işlemden geçirilmiş ve 115,134 tweet duygu analizi için kullanılabilir hale getirilmiştir. Bu aşamada atılan tweetlerden 5134 adet rast gele seçilerek makine öğrenmesi modellerinde kullanılabilmesi için etiketleme işlemi ile hangi duyguyu ifade ettiği belirlenmiştir. Etiketleme işlemi Tablo 4.1'de gösterildiği gibi atılan tweet olumluysa pozitif, olumsuzsa negatif, tarafsızsa nötr olarak yapılmıştır.

Tablo 4.1. Etiketleme Örneği

Tweet	**Etiket**
Fazla ödev sunum yapmak istemiyorum.	Negatif
Uzaktan eğitim olayını baya sevdim.	Pozitif
Online dersler yarın başlıyor.	Nötr

Etiketleme işlemleri genellikle yazar tarafından manuel olarak gerçekleştirilmektedir. Fakat bu çalışmada manuel etiketleme işleminin yanı sıra veri setinin büyüklüğü ve yapılan işlemlerin tamamen kişi yorumlamasından bağımsız gerçekleşmesi adına TextBlob, Vader ve Bert modelleri gibi duygu sınıfı çıktıları veren önceden eğitilmiş hazır modellerde etiketlendirme yapılarak, bahsedilen hazır modellerin manuel etiketlendirme işlemine kıyasla başarıları karşılaştırılmıştır. Aynı zamanda bahsedilen bu hazır modellerin kullanımında gerçekleştirilen dil çeviri işleminin (Türkçe'den İngilizce'ye) etiketleme işlemlerinde paylaşılan tweette verilmek istenen duygudaki değişimi de ölçülmüş olmaktadır.

Bahsedilen hazır modeller ile etiketleme yapılabilmesi için tweetlerin İngilizce içeriğe sahip olması gerekmektedir. Bu sebeple hazır modeller ile etiketleme yapmadan önce Türkçe'den İngilizceye çeviri işlemi gerçekleştirilmiştir. Python ile dil çeviri işlemlerinde Yandex, Google, Microsoft, Bing gibi firmaların çeviri API'lerinden faydalanılabilmektedir. Fakat bu API'lerin ücretsiz kullanımında belirli istek süresi ve çeviri için karakter sayısı kısıtları bulunmaktadır.

Bu çalışmada Yandex çeviri API'si kullanılarak çeviri işlemi gerçekleştirilmiştir. Bu çeviri API'i kullanabilmek için öncelikle bir yandex hesabına sahip olmak ve çeviri API anahtarı edinmek gerekmektedir. Yandex bu API'nin kullanımında 75 dolar hibe ile 5 milyon karakter çevirisine izin vermektedir. Çalışma da çeviri yapılacak veri seti için Yandex'in sağlamış olduğu bu hizmet yeterli olmuştur.

Etiketleme işlemi yapılacak veri setinde Türkçe tweetler ve yapılan çeviri işlemi sonrası Türkçe tweetlerin İngilizce karşılığı şekil 4.22'deki gibidir.

	tweets	eng_tweets
0	kişilik sınıflarda kişi akışkanlar almak istem...	just because you don't want to get people flui...
1	öğrenciler öğretmensiz kalmasın öğrenciler öğr...	students should not be left without a teacher ...
2	milli bakanı selçuk eğitimi önümüzdeki yıllard...	minister of national education selçuk we are i...
3	şurada kalmış sürü sınav online eğitimden anla...	a lot of exams left here are not understood fr...
4	ağustos eylül zaten telafi eğitimi olacaktı ka...	august september was already going to be compe...
...	...	...
5129	özel okullarda yaklaşık öğretmen çalışıyor dev...	as teachers continue to work in private school...
5130	kıymetli öğrenciler üniversitemiz ilgili kurul...	dear students, as a result of the evaluations ...
5131	lisansta sınıf dönem ortalama yüksek lisansta ...	i am sure that I would have pushed the bottom ...
5132	sabah öğretim dersine canlı sistem bakacaktım ...	i was going to look at the live system in the ...
5133	riskin bariz yüksek olduğu dönemde lütfen öğre...	at a time when the risk is obviously high, ple...

5134 rows × 2 columns

Şekil 4.22. Etiketleme Yapılacak Veri Setine Ait Örnek

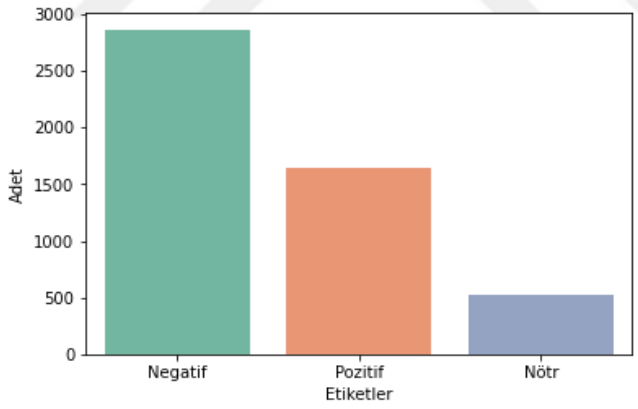
Dil çeviri işleminin ardından İngilizce metinlerde de veri ön işleme yapılarak çeviri işleminde oluşan noktalama işaretleri, İngilizce durak kelimeler veri setinden çıkarılmıştır. Veri etiketleme işleminde ilk olarak rast gele seçilen 5134 adet tweet 4 farklı kişi tarafından Türkçe metinler üzerinden manuel olarak etiketlenmiştir. Manuel etiketleme işlemi aşamasında konu ile ilgili olmayan 91 tweet tespit edilerek etiketlendirilen veri setinden çıkarılmıştır. 5043 adet tweetin manuel etiketleme işlemi

sonrasında 2866 tweet negatif, 1648 tweet pozitif, 529 tweet nötr olarak işaretlenmiştir. Etiketlenen tweetlerin duygu çıktıları Şekil 4.23 ve duygu dağılımları çubuk grafiği Şekil 4.24'teki gibidir.

	sentiment	tweets	eng_tweets
0	Negatif	kişilik sınıflarda kişi akışkanlar almak istem...	just because you don't want to get people flui...
1	Pozitif	öğrenciler öğretmensiz kalmasın öğrenciler öğr...	students should not be left without a teacher ...
2	Nötr	milli bakanı selçuk eğitimi önümüzdeki yıllard...	minister of national education selçuk we are i...
3	Pozitif	şurada kalmış sürü sınav online eğitimden anla...	a lot of exams left here are not understood fr...
4	Negatif	ağustos eylül zaten telafi eğitimi olacaktı ka...	august september was already going to be compe...
...	...	...	...
5038	Negatif	özel okullarda yaklaşık öğretmen çalışıyor dev...	as teachers continue to work in private school...
5039	Nötr	kıymetli öğrenciler üniversitemiz ilgili kurul...	dear students, as a result of the evaluations ...
5040	Pozitif	lisansta sınıf dönem ortalama yüksek lisansta ...	i am sure that I would have pushed the bottom ...
5041	Nötr	sabah öğretim dersine canlı sistem bakacaktım ...	i was going to look at the live system in the ...
5042	Pozitif	riskin bariz yüksek olduğu dönemde lütfen öğre...	at a time when the risk is obviously high, ple...

5043 rows × 3 columns

Şekil 4.23. Manuel Etiketli Tweetlerin Örnek Duygu Çıktıları.



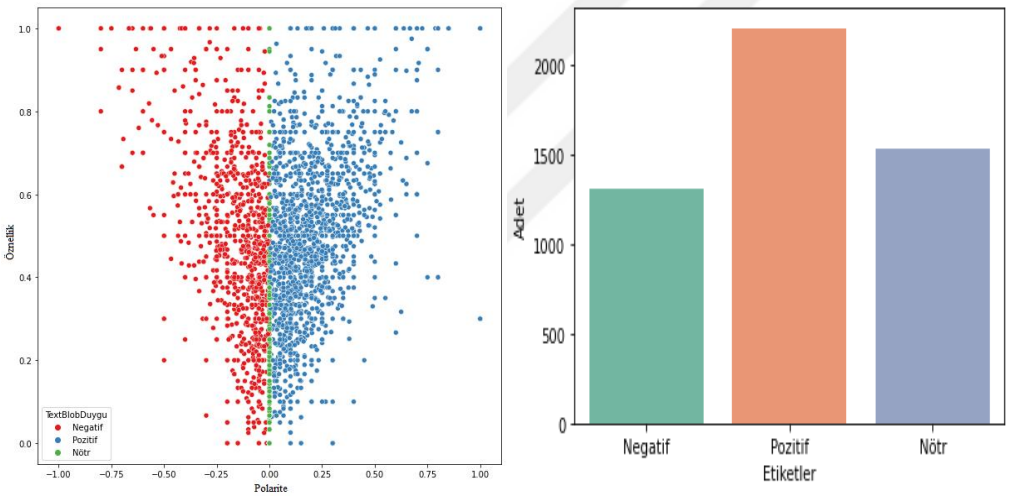
Şekil 4.24. Manuel Etiketleme Duygu Dağılımları.

Önceden eğitilmiş hazır modellerden ilki olan TextBlob -1 ile +1 arasında verdiği polarite değeri ile metnin ne kadar olumlu ya da ne kadar olumsuz olduğu bilgisini vermektedir. Çalışmada TextBlob polarite değeri sıfırdan küçükler negatif, sıfıra eşit olanlar nötr, sıfırdan büyük olanlar ise pozitif olacak şekilde model uygulanmıştır. TextBlob'a ait polarite, öznellik ve duygu çıktıları şekil 4.25'te ve duygu dağılımlarının görselleştirilmiş hali de şekil 4.26'daki gibidir.

	sentiment	tweets	eng_tweets	TextBlobSubjectivity	TextBlobPolarity	TextBlobDuygu
0	Negatif	kişilik sınıflarda kişi akışkanlar almak istem...	want people fluid personality classes make mis...	0.550000	-0.250000	Negatif
1	Pozitif	öğrenciler öğretmensiz kalmasın öğrenciler öğr...	students left without teacher students miss te...	0.250000	0.250000	Pozitif
2	Nötr	milli bakanı selçuk eğitimi önümüzdeki yıllard...	minister national education selçuk process wor...	0.477273	-0.056818	Negatif
3	Pozitif	şurada kalmış sürü sınav online eğitimden anla...	exams left understood online education perform...	0.500000	0.250000	Pozitif
4	Negatif	ağustos eylül zaten telafi eğitimi olacaktı ka...	august september already going compensatory tr...	0.311111	0.101010	Pozitif
...	...	...	...	...	...	...
5038	Negatif	özel okullarda yaklaşık öğretmen çalışıyor dev...	teachers continue work private schools employe...	0.375000	0.000000	Nötr
5039	Nötr	kıymetli öğrenciler üniversitemiz ilgili kurul...	dear students result evaluations made relevant...	0.900000	0.400000	Pozitif
5040	Pozitif	lisansta sınıf dönem ortalama yüksek lisansta ...	sure would pushed bottom reason average rise c...	0.522222	0.075000	Pozitif
5041	Nötr	sabah öğretim dersine canlı sistem bakacaktım ...	going look live system morning teaching class ...	0.262963	-0.006397	Negatif
5042	Pozitif	riskin bariz yüksek olduğu dönemde lütfen öğre...	time risk obviously high please ignore health ...	0.540000	0.160000	Pozitif

5043 rows × 6 columns

Şekil 4.25. Textblob Polarite, Öznellik ve Duygu Çıktıları.



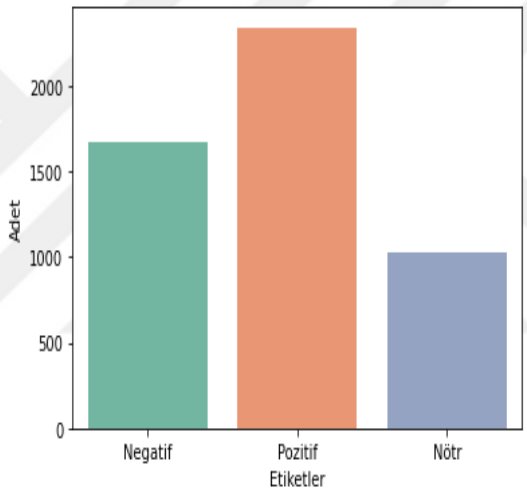
Şekil 4.26. Textblob Polarite ve Öznellik, Duygu Dağılımları.

Vader negatif, nötr ve pozitif etiketlerin yanı sıra bileşik değer çıktısı da üretmektedir. Bu bileşik değere bağlı olarak metnin hangi duyguyu ifade ettiğini belirlemek için ideal ölçekler kullanılmaktadır. Çalışmada +0.05'ten büyük olanlar pozitif, -0.05 ile +0.05 arasında kalanlar nötr, -0.05'ten küçük olanlarda negatif olacak şekilde ideal ölçek belirlenerek model uygulanmıştır. Vader bileşik değer ve duygu çıktıları şekil 4.27'de ve duygu dağılımlarının çubuk grafiği de şekil 4.28'deki gibidir.

	sentiment	tweets	eng_tweets	VaderBilesikDeger	VaderDuygu
0	Negatif	kişilik sınıflarda kişi akışkanlar almak istem...	want people fluid personality classes make mis...	-0.2211	Negatif
1	Pozitif	öğrenciler öğretmensiz kalmasın öğrenciler öğr...	students left without teacher students miss te...	0.4843	Pozitif
2	Nötr	milli bakanı selçuk eğitimi önümüzdeki yıllard...	minister national education selçuk process wor...	0.0000	Nötr
3	Pozitif	şurada kalmış sürü sınav online eğitimden anla...	exams left understood online education perform...	0.6124	Pozitif
4	Negatif	ağustos eylül zaten telafi eğitimi olacaktı ka...	august september already going compensatory tr...	-0.3863	Negatif
...	...	...	...	...	...
5038	Negatif	özel okullarda yaklaşık öğretmen çalışıyor dev...	teachers continue work private schools employe...	-0.3182	Negatif
5039	Nötr	kıymetli öğrenciler üniversitemiz ilgili kurul...	dear students result evaluations made relevant...	0.3818	Pozitif
5040	Pozitif	lisansta sınıf dönem ortalama yüksek lisansta ...	sure would pushed bottom reason average rise c...	0.3182	Pozitif
5041	Nötr	sabah öğretim dersine canlı sistem bakacaktım ...	going look live system morning teaching class ...	0.0000	Nötr
5042	Pozitif	riskin bariz yüksek olduğu dönemde lütfen öğre...	time risk obviously high please ignore health ...	0.3204	Pozitif

5043 rows × 5 columns

Şekil 4.27. Vader Bileşik Değer ve Duygu Çıktıları



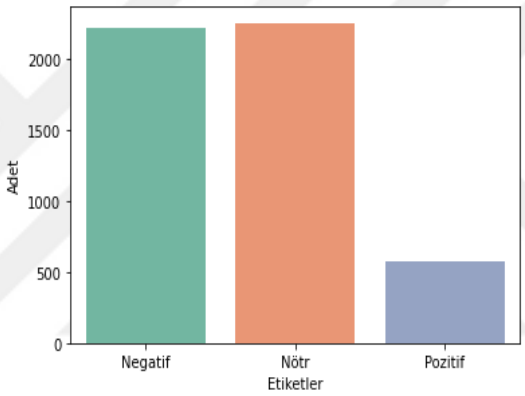
Şekil 4.28. Vader Duygu Dağılımları

Bert modeli için Hugging Face üzerinde faklı büyüklükte ve farklı diller için geliştirilmiş ücretsiz birçok hazır model bulunmaktadır. Bu çalışmada İngilizce metinlerde etiketleme işlemi için Hugging Face'te ücretsiz olarak yer alan yaklaşık 58 milyon tweet ile eğitilmiş, duygu analizi için ince ayar yapılmış Bert tabanlı "cardiffnlp/Twitter-roberta-base-sentiment" (Anonymous, 2020; Barbieri ve ark., 2020) modeli kullanılmıştır. Model çıktı olarak duygu skorları ile birlikte pozitif, negatif ve nötr sonuçlar vermektedir. Bert duygu skoru ve duygu çıktıları şekil 4.29'da ve duygu dağılımlarının çubuk grafiği de şekil 4.30'daki gibidir.

	sentiment	tweets	eng_tweets	BertSentimentscore	BertDuygu
0	Negatif	kişilik sınıflarda kişi akışkanlar almak istem...	want people fluid personality classes make mis...	0.659100	Negatif
1	Pozitif	öğrenciler öğretmensiz kalmasın öğrenciler öğr...	students left without teacher students miss te...	0.572479	Nötr
2	Nötr	milli bakanı selçuk eğitimi önümüzdeki yıllard...	minister national education selçuk process wor...	0.798004	Nötr
3	Pozitif	şurada kalmış sürü sınav online eğitimden anla...	exams left understood online education perform...	0.648503	Negatif
4	Negatif	ağustos eylül zaten telafi eğitimi olacaktı ka...	august september already going compensatory tr...	0.769911	Nötr
...	...	...	...	...	...
5038	Negatif	özel okullarda yaklaşık öğretmen çalışıyor dev...	teachers continue work private schools employe...	0.629711	Negatif
5039	Nötr	kıymetli öğrenciler üniversitemiz ilgili kurul...	dear students result evaluations made relevant...	0.900299	Nötr
5040	Pozitif	lisansta sınıf dönem ortalama yüksek lisansta ...	sure would pushed bottom reason average rise c...	0.591537	Nötr
5041	Nötr	sabah öğretim dersine canlı sistem bakacaktım ...	going look live system morning teaching class ...	0.528989	Nötr
5042	Pozitif	riskin bariz yüksek olduğu dönemde lütfen öğre...	time risk obviously high please ignore health ...	0.487867	Nötr

5043 rows × 5 columns

Şekil 4.29. Bert Duygu Skoru ve Duygu Çıktıları



Şekil 4.30. Bert Duygu Dağılımları.

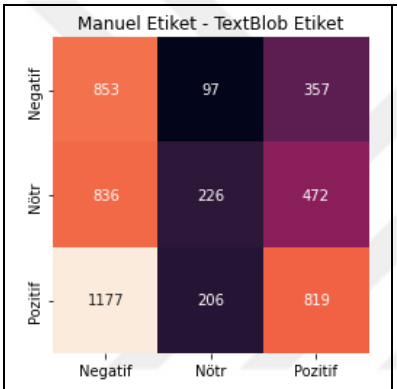
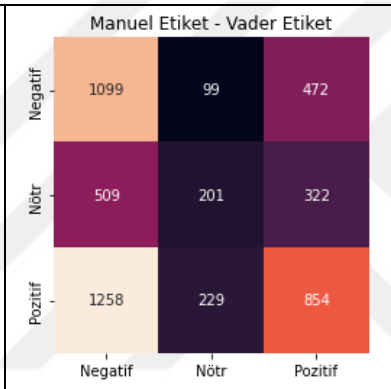
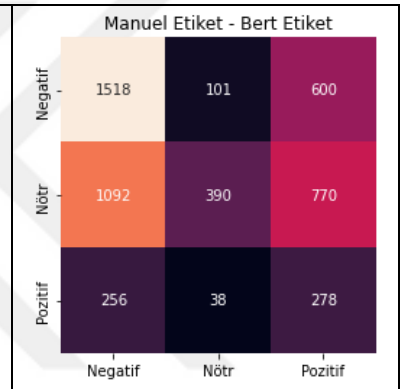
Şekil 4.31'de çeviri işlemi yapılan Türkçe tweetlerin hazır modeller kullanılarak etiketlendirilmesine ait duygu çıktıları birlikte gösterilmiştir. Manuel etiketlemeye kıyasla hazır modellerin performanslarını ölçmek için etiketlenen 5043 adet tweet için doğruluk oranlarına bakılmıştır. Manuel ve hazır modellerin etiketlendirme işlemine ait karışıklık matrisleri ve doğruluk oranları Tablo 4.2'deki gibidir.

	sentiment	tweets	eng_tweets	TextBlobDuygu	VaderDuygu	BertDuygu
0	Negatif	kişilik sınıflarda kişi akışkanlar almak istem...	just because you don't want to get people flui...	Negatif	Negatif	Negatif
1	Pozitif	öğrenciler öğretmensiz kalmasın öğrenciler öğr...	students should not be left without a teacher ...	Pozitif	Pozitif	Nötr
2	Nötr	milli bakanı selçuk eğitimi önümüzdeki yıllard...	minister of national education selçuk we are i...	Negatif	Nötr	Nötr
3	Pozitif	şurada kalmış sürü sınav online eğitimden anla...	a lot of exams left here are not understood fr...	Pozitif	Pozitif	Negatif
4	Negatif	ağustos eylül zaten telafi eğitimi olacaktı ka...	august september was already going to be compe...	Pozitif	Negatif	Nötr
...	...	...	...	...	...	...
5038	Negatif	özel okullarda yaklaşık öğretmen çalışıyor dev...	as teachers continue to work in private school...	Nötr	Negatif	Negatif
5039	Nötr	kıymetli öğrenciler üniversitemiz ilgili kurul...	dear students, as a result of the evaluations ...	Pozitif	Pozitif	Nötr
5040	Pozitif	lisansta sınıf dönem ortalama yüksek lisansta ...	i am sure that I would have pushed the bottom ...	Pozitif	Pozitif	Nötr
5041	Nötr	sabah öğretim dersine canlı sistem bakacaktım ...	i was going to look at the live system in the ...	Negatif	Nötr	Nötr
5042	Pozitif	riskin bariz yüksek olduğu dönemde lütfen öğre...	at a time when the risk is obviously high, ple...	Pozitif	Pozitif	Nötr

5043 rows × 6 columns

Şekil 4.3. Textblob, Vader, Bert Örnek Duygu Çıktıları.

Tablo 4.2. Manuel ve Hazır Modellerin Karşılaştırılmasına Ait Karışıklık Matrisleri

		
Doğruluk Oranı: 0.3764	Doğruluk Oranı: 0.4271	Doğruluk Oranı: 0.4334

Karışıklık matrisine bakıldığında manuel ve TextBlob ile etiketlenen tweetlerden aynı etikete sahip 853 negatif, 226 nötr ve 819 pozitif tweet olmuştur. TextBlob'un manuel etiketlemeye kıyasla doğruluk oranı ise 0.3764 olmuştur. Manuel ve Vader ile etiketlenen tweetlerden aynı etikete sahip 1099 negatif, 201 nötr ve 854 pozitif tweet yer alırken, Vader'in manuel etiketlemeye kıyasla doğruluk oranı ise 0.4271 olmuştur. Manuel ve Bert ile etiketlenen tweetlerden aynı etikete sahip 1518 negatif, 390 nötr ve 278 pozitif tweet vardır. Bert'in manuel etiketlemeye kıyasla doğruluk oranı ise 0.4334 olmuştur.

Farklı etiketleme yöntemleri olan hazır modellerden TextBlob'un performansı diğer iki modele göre düşük kalmıştır. Vader ve Bert modelleri birbirine yakın sonuçlar vermiş olsa da Bert'in doğruluk oranı diğer iki modele göre daha yüksek olmuştur. Makul kabul oranında bir benzerlik seviyesine ulaşılamamış olsa da sistemin analiz edilmesi amacıyla test edilmiştir. Ancak gerçek başarı seviyesi konuşulurken Türkçe

metinlerin manuel etiketlendirilmesi daha anlamlı kabul edilmiştir. Bu doğrultuda yapılan etiketlemeler ile farklı modeller oluşturularak makine öğrenmesi algoritmalarına girdi olarak kullanılmıştır.

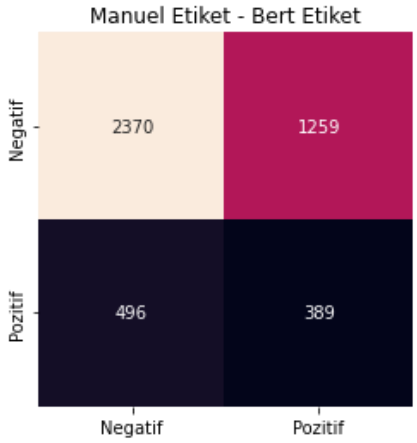
Yapılan etiketleme işlemlerine ek olarak, yukarıda da bahsedilen Bert'in farklı diller için geliştirildiği bir diğer modeli olan Türkçe metinlerde negatif ve pozitif duygu çıktıları üreten Bert tabanlı "savasy/bert-base-turkish-sentiment-cased" (Yıldırım, 2021) modeli kullanılmıştır. Modelin kullanılabilmesi için manuel olarak işaretlenen nötr etiketler veri setinden çıkartılarak 4514 adet tweet bu hazır model ile tekrar etiketlendirilmiştir. Manuel ve Bert'in Türkçe modeli ile etiketlendirilmiş tweetlere ait duygu çıktıları Şekil 4.32'deki gibidir.

	sentiment	tweets	bert_sentimentscore	BertDuygu
0	Negatif	kişilik sınıflarda kişi akışkanlar almak istem...	0.999397	Negatif
1	Pozitif	öğrenciler öğretmensiz kalmasın öğrenciler öğr...	0.785479	Negatif
2	Pozitif	şurada kalmış sürü sınav online eğitimden anla...	0.992008	Negatif
3	Negatif	ağustos eylül zaten telafi eğitimi olacaktı ka...	0.769605	Pozitif
4	Negatif	dışarısı gece ayaza çekecek belli yollar yolla...	0.999169	Negatif
...	...	...	...	...
4509	Negatif	yenilik çağdaşlık köyden uzak diyarlardan eğit...	0.843504	Pozitif
4510	Negatif	yunus aydın önceden güveniyordum artık anladım...	0.937090	Negatif
4511	Negatif	özel okullarda yaklaşık öğretmen çalışıyor dev...	0.942686	Negatif
4512	Pozitif	lisansta sınıf dönem ortalama yüksek lisansta ...	0.637277	Negatif
4513	Pozitif	riskin bariz yüksek olduğu dönemde lütfen öğre...	0.897164	Negatif

4514 rows × 4 columns

Şekil 4.32. Manuel ve Bert'in Türkçe Modeli ile Etiketlenen Örnek Tweetler

Nötr etikete sahip tweetlerin çıkarılmasının ardından veri setinde yer alan 4514 adet tweete ait manuel olarak işaretlenmiş 2866 negatif, 1648 pozitif, Bert modeli ile işaretlenmiş 3629 negatif, 885 pozitif duygu çıktısı yer almaktadır. Şekil 4.33'te yer alan karışıklık matrisi incelendiğinde manuel etiketleme işlemine kıyasla Bert'in Türkçe modeli ile etiketleme işlemine ait doğruluk oranı 0.6112 olmuştur.



Şekil 4.33. Manuel ve Türkçe Bert Modeline Ait Karışıklık Matrisi

Sonuç olarak veri etiketleme işlemlerinde manuel etiketleme yöntemi, veri setinin büyük olduğu durumlarda hız ve zaman açısından zahmetli olabilmektedir. Tez kapsamında bu gibi durumlara alternatif olabileceği düşünülerek literatürde oldukça sık kullanılan İngilizce metinlerde doğrudan duygu çıktıları üreten TextBlob, Vader, Bert gibi hazır modeller kullanılarak test edilmiştir. Bu tür hazır modellerin kullanılabilmesi için çeviri işlemine ihtiyaç duyulmuştur. Türkçe dilinin zengin yapısı, kelimelerin kısaltılarak kullanılması ve iki dilin kökenindeki yapısal farklar çeviri işlemini etkiliyor olsa da metinde verilmek istenen duygu birbirine yakın olabilmektedir. Örneğin Türkçe "bütün çocukların gençlerin hayatı tehlikede uzaktan eğitim şart" negatif olarak etiketlenen bir cümlenin İngilizce çevirisi "The lives of all children and young people are in danger, distance education is a must" olan cümle de negatif duyguyu yansıtmaktadır. Ancak söz konusu modellerin manuel etiketleme işlemine kıyasla değerlendirildiği noktada yüksek derecede pozitif veya negatif duyguya sahip kelimelerin ironi yapılarak kullanılması ya da bir cümlenin pozitif başlayıp negatif, negatif başlayıp pozitif tamamlanması gibi yukarıda da bahsedilen Türkçe'nin zengin biçimsel yapısı çeviri yapılarak kullanılan hazır modellerin kullanımında başarı oranlarına da bakıldığında duygu çıktılarını etkilemektedir. Bu doğrultuda tez kapsamında kullanılan veri seti için manuel etiketleme işlemi ile çeviri yapılarak kullanılan hazır modeller karşılaştırıldığında makul oranda bir benzerlik seviyesi yakalanamamıştır.

4.7. Veri Sayısallaştırma

Makine öğrenmesi modellerine geçmeden önce veri ön işleme tabi tutulan ve etiketlenen tweetlerin sayısal olarak ifade edilmesi gerekmektedir. Bu çalışmada BoW, TF-IDF, Word2Vec yöntemleri kullanılarak sayısallaştırma işlemi gerçekleştirilmiş.

Veri setindeki terimlerin oluşum şeklini (terim sayısını) belirten BoW, terimlerin pozisyonlarını ve sıralamasını dikkate almadan metinleri temsil etmektedir. Örnek ile açıklayacak olursak, kelime torbası bir D dokümanı {d1, d2, …, dD} için bir S sözlüğü ve bu sözlüğün içerdiği N adet benzersiz kelime ile D x N boyutlu bir M matrisi oluşturacaktır. M matrisindeki her bir satır D(i). cümledeki kelimelerin sıklığını vermektedir.

D1: Vaka sayıları böyle devam ederse uzaktan eğitim devam edecek.

D2: Dersler uzaktan eğitim yoluyla devam edecek.

S={‘Vaka’,’sayıları’,’böyle’,’devam’,’ederse’,’uzaktan’,’eğitim’,’edecek’,’dersler’,’yoluyla’}

D= 2, N=10

2 x 10 boyutunda M matrisi şöyle oluşacaktır:

	Vaka	sayıları	böyle	devam	ederse	uzaktan	eğitim	edecek	Dersler	yoluyla
D1	1	1	1	2	1	1	1	1	0	0
D2	0	0	0	1	0	1	1	1	1	1

BoW’ün bu çalışmada python ile kullanımı aşağıdaki gibidir:

```
bow_vektor = CountVectorizer(max_df=0.9, min_df=5, max_features=1000)
bow = bow_vektor.fit_transform(df[‘tweets’])
```

“Sklearn” kütüphanesine bağlı “CountVectorizer” ile veri seti terim sayısı matrisine dönüştürülmüştür. “CountVectorizer” metinlerdeki tüm kelimeler ile bir sözlük oluşturarak her bir tweet için kelimelerin sıklığını saymaktadır. Bu işlemlerin ardından kelime çantası bir sınıflandırıcı için girdi olarak kullanılabilir hale gelmektedir.

TF-IDF, verideki bir kelimenin frekansını, o kelimeyi içeren toplam doküman sayısı ve tüm dokümanların sayısına dayalı olarak ağırlıklandırmaktadır. Böylece verideki önemsiz değerler elenerek önemli öznitelikler tespit edilmektedir ve

sınıflandırma işlemi aşamasında performans artışı sağlanmaktadır. TF-IDF hesaplama adımları aşağıdaki gibidir:

TF = (t teriminin bir belgede görünme sayısı) / (belgedeki terim sayısı)

IDF = log(N/n), N belge sayısı ve n bir t teriminin göründüğü belge sayısı.

TF-IDF = TF x IDF

TF-IDF'ın bu çalışmada python ile kullanımı aşağıdaki gibidir.

```
Tfidf_vektor = TfidfVectorizer(max_df=0.90, min_df=5, max_features=1000)
tfidf = tfidf_vektor.fit_transform(df['tweets'])
```

Veri setinin TF-IDF özellik matrisine dönüştürülmesi için "TfidfVectorizer" kullanılmıştır. "TfidfVectorizer", en sık kullanılan kelimeleri özellik indekslerine eşlemek ve dolayısıyla bir kelime oluşum sıklığı matrisi hesaplamak için bellek içi kelime dağarcığı kullanmaktadır. TF-IDF tüm veri setinde nadir ancak birkaç belgede iyi sayılarda yer alan kelimelere önem verirken, çok sık geçen kelimelere düşük ağırlıklar vermektedir.

Kelime dağarcığı oluşturulurken kelimelerin metinlerde geçen sıklığına göre ayarlama yapılabilmektedir. Bu çalışmada "CountVectorizer" ve "TfidfVectorizer" için kullanılan parametre ve değerleri açıklamaları ile birlikte tablo 4.3'teki gibidir.

Tablo 4.3. "Countvectorizer" ve "Tfidfvectorizer" Parametreleri

Parametre	Değer	Açıklama
max_df	0.9	Veri setinin %90'ından fazlasında geçen terimleri yok say
min_df	5	5'ten daha az metinde görünen terimleri yok say
max_features	1000	Maksimum özellik sayısı

Belirlenen bu parametre değerleri ile makine öğrenmesi modelleri için, modelin fazla öğrenmesini önlemek, eğitim verimliliğini artırmak için modeli besleyecek vektörlerin boyutu kontrol altına alınmış olmaktadır. Ayrıca TF-IDF N-gram (bigram ve trigram) yaklaşımıyla kelimelerin tekli ve ikili, ikili ve üçlü olarak metinlerde geçme sıklıklarına bakılarak makine öğrenmesi modellerine uygulanmıştır.

BoW ve TF-IDF kelime temsili için popüler uygulamalar olsa da kelimelerin anlamsal yakınlıklarını anlamak gerektiğinde eksik kalmaktadırlar. Bu noktada Kelime Vektörü (Word2Vec) devreye girmektedir.

Sayısallaştırma işleminin bir başka yöntemi olan Word2Vec, her kelimeyi çok boyutlu bir uzayda eşler. Bu işlemi kelimenin metinlerde bağlanıma dikkat ederek yapmaktadır. Böylece çok boyutlu bu uzayda benzer anlama sahip kelimeler birlikte kümelenir ve iki kelime arasındaki mesafe de aynı olmaktadır. Word2Vec "Gensim" kütüphanesinin bir parçasıdır. "Gensim" ham dijital metinleri (düz metin) denetimsiz makine öğrenimi algoritmaları kullanarak işlemek için tasarlanmıştır (Rehurek, 2021). Word2Vec'in bu çalışmada python ile kullanımı aşağıdaki gibidir.

```
W2v_model=gensim.models.Word2Vec(tweet_token,size=200,window=5,
min_count=2,sg = 1,hs = 0,negative = 10,workers= 2)
```

Word2Vec birer adet girdi, çıktı ve gizli katmandan oluşan bir yapay sinir ağından oluşmaktadır. Kelime vektörleri oluşturulurken size, window, min_count gibi parametreler kullanılmaktadır. Bu çalışmada kullanılan Word2Vec parametreleri, değer ve açıklamaları ile birlikte Tablo 4.4'teki gibidir.

Tablo 4.4. Word2Vec Parametreleri

Parametre	Değer	Açıklama
size	200	Kelime vektörlerinin boyutu
window	5	Hedef kelimenin sağında ve solunda yer alacak kelime sayısı
min_count	2	min_count değerinden düşük frekanstaki kelimelerin yok sayılması
sg	1	1: Skip-Gram, 0: CBOW modeli
workers	2	Thread sayısı
hs	0	hs=0 negatif örnekleme, hs=1 hiyerarşik softmax
negative	10	Bağlam dışında seçilecek rastgele kelime sayısı

Pencere genişliği hedef kelimenin sağında ve solunda kaç kelime olması gerektiğini belirtirken, kelime vektör boyutu her bir kelimenin kaç boyutlu vektör olarak ifade edileceğini belirtmektedir. Bu da sinir ağındaki gizli katman nöron sayısına karşılık gelmektedir. "hs" parametresi "negative" parametresi ile birlikte kullanılmaktadır ve hs=0 ifadesi negatif örnekleme yapılacağı anlamına gelmektedir. Negatif örnekleme bağlamdaki kelimeler ile bağlam dışından seçilen rastgele kelimelerin karşılaştırılmasıyla yapılmaktadır. Hiyerarşik softmax seyrek kelimeler için iyi çalışırken, negatif örnekleme sık kullanılan kelimeler için iyi çalışmaktadır (Gheorghe, 2018). Word2Vec kelime temsili için Skip-Gram algoritması kullanılmıştır.

Skip-Gram CBOW'a göre büyük veri setlerinde daha iyi çalışmaktadır ve iki veya daha çok anlamlı kelimeleri anlamakta daha başarılı sonuçlar vermektedir ancak hesaplama gücü açısından daha fazla maliyet gerektirmektedir (Keskin, 2018).

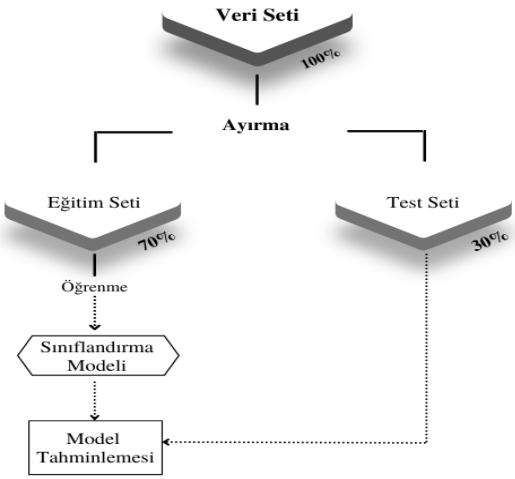
Son olarak tweet başına tüm kelime vektörlerinin ortalaması alınarak, farklı sayıda oluşacak vektöre sahip tweetlerin önüne geçilmiştir. Böylece veri seti için kelime vektörlerinin boyutu kadar (size=200) Word2Vec özelliği oluşmuştur. Word2Vec de en iyi sonuç, pencere boyutu 5, vektör boyutu 200, epoch sayısı 20 olarak uygulandığı durumda elde edilmiştir. Veri setinde yer alan "yurt" kelimesi için örnek Word2Vec çıktısı şekil 4.34'teki gibidir.

```
[('apart', 0.837084174156189),
 ('kiraları', 0.7780308723449707),
 ('odasında', 0.7740583419799805),
 ('isteyenlere', 0.7689255475997925),
 ('şehirlerdeki', 0.7649985551834106),
 ('raporlu', 0.7612568140029907),
 ('ücretlerin', 0.7545772194862366),
 ('ünilerde', 0.7543065547943115),
 ('cesaretiniz', 0.7527318000793457),
 ('isteyene', 0.7483856081962585)]
```

Şekil 4.34. "Yurt" Kelimesi İçin Örnek Word2Vec Çıktısı

4.8. Verilerin Ayrılması ve Modelleme

Tezin önceki bölümlerinde veriler uygun forma sokularak gereken ön modelleme aşamaları tamamlanmıştır. Bu bölümde de en iyi sınıflandırma modelini bulabilmek için etiketlendirilen veri setleri Bow, TF-IDF ve Word2Vec ile ayrı ayrı sayısal olarak temsil edilerek python içerisinde bulunan "sklearn" kütüphanesinde yer alan "train_test_split" metodu ile iki parçaya ayrılmıştır. Birinci parça ile veriler eğitilirken ikinci parça ile doğrulu test edilmiştir. Çalışmada "random_state" değeri 42, "test_size" değeri 0.30 alınarak veri setinin %70'i eğitim, %30'u test için uygulanmıştır. Eğitim ve test verilerinin sınıflandırma modellerinde kullanımına ait görsel şekil 4.35'teki gibidir.



Şekil 4.35. Eğitim ve Test Verilerinin Ayrılması

Tez kapsamında sınıflandırma işlemi için etiketlendirilen metinlerin negatif, pozitif veya nötr duygu dağılımları ile birlikte ayrıca manuel olarak işaretlenmiş nötr etikete sahip metinlerin veri setinden çıkartılması ile sadece negatif ve pozitifliliğe bakılarak makine öğrenmesi algoritmalarının performansları karşılaştırılmıştır. Negatif, pozitif veya nötr gibi çoklu sınıflandırma problemlerinde ikili sınıflandırma algoritmalarını kullanmak için elimizdeki veri setini çoklu ikili sınıflandırma veri setine bölmek ve her birini bir ikili sınıflandırma modeline sığdırmak gerekmektedir. Bire karşı hepsi (OneVsRest) bu yaklaşım için kullanılan bir modeldir. Model çok sınıflı bir sınıflandırmayı sınıf başına bir ikili sınıflandırma problemine böler daha sonra her bir ikili sınıflandırma problemi için bir ikili sınıflandırıcı eğitilir, en güvenilir model

kullanılarak tahminler yapılır. Çalışmada yer alan duygu sınıfları negatif, pozitif, nötr için bire karşı hepsi yaklaşımı aşağıdaki gibidir (Brownlee, 2020).

Negatif / [Nötr, Pozitif]
Nötr / [Negatif, Pozitif]
Pozitif / [Negatif, Nötr]

BoW, TF-IDF, Word2Vec veri sayısallaştırma yöntemleri ile LR, SGD, SVM OneVsRest çoklu sınıflandırma çatısı altında, bunlara ek olarak RF ve Multinominal NB makine öğrenmesi algoritmaları kullanılarak sınıflandırma modelleri oluşturulmuştur. Manuel ve hazır modeller ile etiketlendirilerek oluşturulan veri setleri bu aşamada girdi olarak kullanılmış, her birinin sınıflandırma işlemlerindeki performansları değerlendirilmiştir. Sınıflandırma modellerinden hangisinin en yüksek başarıya sahip olduğuna karar verebilmek için doğruluk oranları karşılaştırılmıştır.

4.9.1. Modellerin Doğruluklarının Karşılaştırılması

Türkçe metinlerden oluşan ve manuel olarak etiketlendirilen Model 1, Türkçe metinlerden oluşan ve çeviri işlemi yapılarak sırasıyla TextBlob, Vader ve Bert ile etiketlendirilen Model 2, Model 3, ve Model 4, İngilizceye çevrilmiş metinlerden oluşan ve manuel etiketlere sahip Model 5, İngilizce metinlerin Textblob, Vader ve Bert ile etiketlendirilmesi ile oluşturulan sırasıyla Model 6, Model 7 ve Model 8 BoW, TF-IDF, Word2Vec ile sayısallaştırılarak LR, SVM, SGD, RF ve Multinominal NM makine öğrenmesi algoritmaları ile her bir model sınıflandırma işlemine tabi tutulmuştur. Sınıflandırma işlemi sonrasında modellere ait doğruluk oranlarına tablo 4.5'te yer verilmiştir. Tablo üzerinde ondalık sayılar sıfırdan sonra dört basamağa yuvarlanmış ve modellere ait en yüksek doğruluk oranları koyu renk ile işaretlenmiştir. Word2Vec ile sayısallaştırılan verilerin "-" (negatif) vektör değerlere sahip olması Multinominal NB algoritmasına girdi olarak kullanılamadığından tablolarda bu ikiliye yer verilmemiştir.

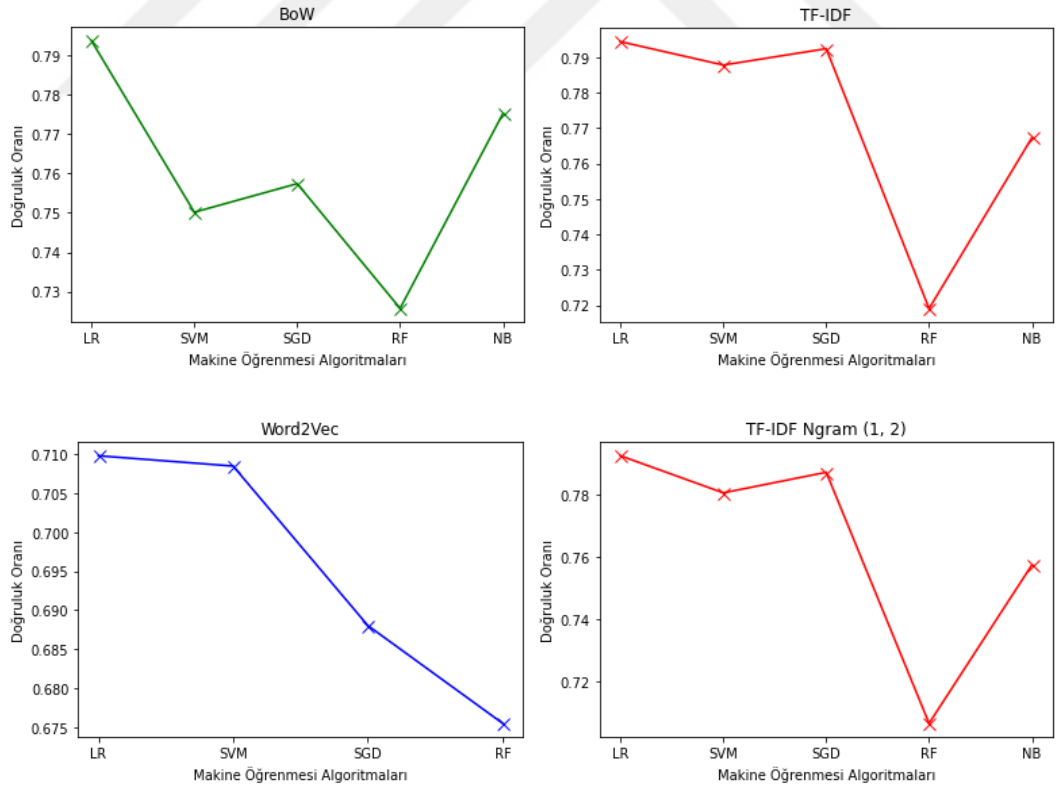
Tablo 4.5. Modellerin Doğruluk Oranları

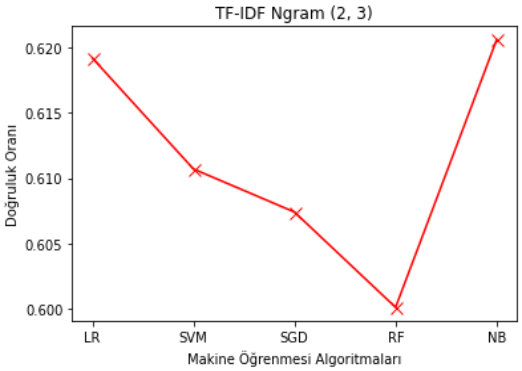
		Türkçe Metin				İngilizce Metin			
	Kelime Gömme	**Manuel Etiket (Model1)**	**TextBlob Etiket (Model2)**	**Vader Etiket (Model3)**	**Bert Etiket (Model4)**	**Manuel Etiket (Model5)**	**TextBlob Etiket (Model6)**	**Vader Etiket (Model7)**	**Bert Etiket (Model8)**
LR	**BoW**	0.7937	0.5042	0.5175	0.5902	0.6470	0.6615	0.6133	0.6080
	TF-IDF	**0.7944**	0.5194	0.5294	0.6034	0.6615	0.6292	0.6087	**0.6140**
	Word2Vec	0.7105	0.4738	0.5135	0.6126	0.6391	0.5062	0.5465	0.6060
	TF-IDF N-gram (1,2)	0.7924	0.5135	0.5333	0.6007	**0.6688**	0.6292	0.6093	0.6034
	TF-IDF N-gram (2,3)	0.6192	0.4342	0.4560	0.5241	0.6014	0.4672	0.4917	0.5254
SVM	**BoW**	0.7501	0.4857	0.4976	0.5651	0.6093	0.6556	0.5935	0.5842
	TF-IDF	0.7878	0.5056	0.5115	0.5888	0.6463	0.6589	0.6232	0.5941
	Word2Vec	0.7144	0.4831	0.5082	**0.6159**	0.6397	0.5346	0.5499	0.6073
	TF-IDF N-gram (1,2)	0.7805	0.5089	0.5188	0.5769	0.6159	0.6536	0.6153	0.5922
	TF-IDF N-gram (2,3)	0.6107	0.4230	0.4514	0.5214	0.5842	0.3952	0.4732	0.5016
SGD	**BoW**	0.7580	0.4930	0.5109	0.5617	0.6080	0.6715	0.6001	0.5968
	TF-IDF	0.7918	0.5102	0.5142	0.5869	0.6444	0.6794	**0.6305**	0.6014
	Word2Vec	0.7111	0.4646	0.4970	0.6140	0.6351	0.5128	0.5472	0.6120
	TF-IDF N-gram (1,2)	0.7871	0.5056	0.5228	0.5703	0.6153	0.6417	0.6239	0.6074
	TF-IDF N-gram (2,3)	0.6074	0.3430	0.4527	0.5254	0.5875	0.4560	0.4771	0.5115
RF	**BoW**	0.7257	**0.5677**	**0.5505**	0.6107	0.6457	**0.7448**	0.6239	0.6100
	TF-IDF	0.7191	0.5512	0.5181	0.5955	0.6358	0.7118	0.6021	0.6066
	Word2Vec	0.6781	0.4699	0.5069	0.5875	0.6338	0.4877	0.5346	0.6060
	TF-IDF N-gram (1,2)	0.7065	0.5452	0.5267	0.5895	0.6483	0.7138	0.6040	0.5875
	TF-IDF N-gram (2,3)	0.6001	0.4269	0.4553	0.5168	0.5953	0.4481	0.4838	0.4864
NB	**BoW**	0.7752	0.4970	0.5261	0.5842	0.6450	0.6159	0.5346	0.6080
	TF-IDF	0.7673	0.5029	0.5247	0.5816	0.6497	0.5875	0.5789	0.6040
	TF-IDF N-gram (1,2)	0.7574	0.4963	0.5287	0.5836	0.6510	0.5789	0.5948	0.5769
	TF-IDF N-gram (2,3)	0.6206	0.4335	0.4580	0.4785	0.5968	0.4619	0.4712	0.5148

Tablo 4.5'te yer alan doğruluk oranları incelendiğinde Türkçe metinler ile kullanılan Model 1'de TF-IDF – LR ile 0.7944'lük en yüksek doğruluk oranı elde edilmiştir. Model 1 için veri sayısallaştırma işlemleri karşılaştırıldığında TF-IDF ile sayısallaştırma işleminde makine öğrenmesi algoritmalarının başarı oranının diğer yöntemlere göre kısmen daha iyi sonuçlar verdiği gözlemlenmiştir. Model 2 için 0.5677 oran ile BoW – RF, Model 3 için 0.5505 oran ile yine BoW – RF, Model 4 için ise 0.6159 oran ile Word2Vec – SVM en yüksek doğruluk oranlarını vermiştir. Bu noktada Türkçe metinlerin manuel etiketler ile birlikte kullanımında elde edilen sınıflandırma başarısı, çeviri yapılarak hazır modeller ile etiketlendirilen modellere göre daha yüksek olmuştur. Çeviri yapılarak hazır modeller ile etiketlendirilen Türkçe metinler için

anlamlı ve makul bir oranda sınıflandırma başarısı yakalanamamıştır. İngilizce metinlerden oluşan modellerden TextBlob ile etiketlendirilen Model 6'da BoW – RF 0.7448'lik doğruluk oranı ile diğer modellere göre en yüksek başarıyı göstermiştir. Manuel etiketleme yöntemine kıyasla Textblob ile etiketleme işleminin başarısı her ne kadar düşük çıkmış olsa da İngilizce içeriğe sahip metinlerde bu modelin makine öğrenmesi algoritmaları ile sınıflandırma başarısı diğer modellere göre daha yüksek çıkmıştır. Bu doğrultuda kullanılan metnin dili ve bu metnin aynı dil yapısıyla etiketlendirilmesi sınıflandırma performanslarını değerlendirmede önemli bir rol oynağı yorumu yapılabilir. Model 5 için TF-IDF Ngram (1, 2) – LR 0.6688 oran ile Model 7 için TF-IDF – SGD 0.6305 oran ile Model 8 için ise TF-IDF – LR 0.6140 oran ile en yüksek başarıyı göstermiştir.

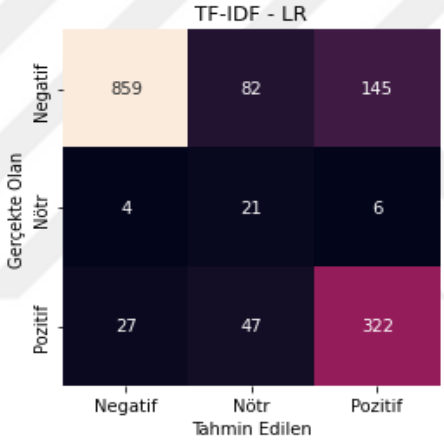
Türkçe metinlerden oluşan ve manuel etiketlere sahip Model 1'in BoW, TF-IDF, Word2Vec, TF-IDF Ngram (1, 2) ve TF-IDF Ngram (2, 3) farklı sayısallaştırma yöntemleri ile kullanılan makine öğrenmesi algoritmalarının doğruluk oranları şekil 4.36'da grafiksel olarak gösterilmiştir.





Şekil 4.36. Model 1'e Ait Doğruluk Oranlarının Grafiksel Gösterimi

Model 1 için en iyi performansı gösteren TF-IDF – LR sınıflandırma işlemine ait karışıklık matrisi şekil 4.37'deki gibidir.



Şekil 4.37. Model 1 TF-IDF – LR Karışıklık Matrisi

Manuel olarak işaretlenen ve nötr etikete sahip tweetler veri setinden çıkartılarak (Model 9) Bert'in Türkçe metinler için duygu çıktıları üreten modeli ile tekrar etiketlendirilmiştir (Model 10). Oluşturulan modellerin her biri BoW, TF-IDF, Word2Vec ile sayısallaştırılarak LR, SVM, SGD, RF ve Multinominal NM makine öğrenmesi algoritmaları ile her bir model sınıflandırma işlemine tabi tutulmuştur. Tablo 4.6'da bu iki model için sınıflandırma performanslarına ait doğruluk oranlarına yer verilmiştir.

Tablo 4.6. Modellere Ait Doğruluk Oranları

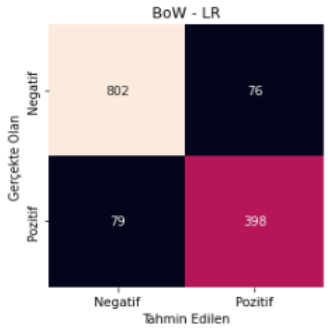
		Türkçe Metin	
	Kelime Gömme	**Manuel Etiket (Model 9)**	**Bert Etiket (Model 10)**
LR	**BoW**	**0.8856**	0.7889
	TF-IDF	0.8811	0.8162
	Word2Vec	0.7579	0.**8214**
	TF-IDF N-gram (1,2)	0.8797	0.8162
	TF-IDF N-gram (2,3)	0.6715	0.8132
SVM	**BoW**	0.8619	0.7402
	TF-IDF	0.8789	0.7815
	Word2Vec	0.7675	0.8228
	TF-IDF N-gram (1,2)	0.8671	0.7815
	TF-IDF N-gram (2,3)	0.6693	0.7867
SGD	**BoW**	0.8678	0.7461
	TF-IDF	0.8693	0.7852
	Word2Vec	0.7232	0.8199
	TF-IDF N-gram (1,2)	0.8590	0.7881
	TF-IDF N-gram (2,3)	0.6583	0.7926
RF	**BoW**	0.8014	0.8029
	TF-IDF	0.7911	0.7933
	Word2Vec	0.7409	0.8114
	TF-IDF N-gram (1,2)	0.7756	0.7904
	TF-IDF N-gram (2,3)	0.6612	0.7785
NB	**BoW**	0.8848	0.7859
	TF-IDF	0.8597	0.8184
	TF-IDF N-gram (1,2)	0.8605	0.8191
	TF-IDF N-gram (2,3)	0.6789	0.8132

Türkçe metinlerin sayısallaştırılarak kullanıldığı pozitif ve negatif etikete sahip Model 9'da 0.8856 doğruluk oranı ile BoW – LR diğer sayısallaştırma ve makine öğrenmesi algoritmalarının birlikte kullanımına göre en iyi performansı göstermiştir. Model 10'da her ne kadar doğruluk oranları kabul edilebilir seviyede olsa da negatif ve pozitif etiket sayılarının (3629 negatif, 885 pozitif) dengesiz dağılımı bu oranlarda etkili olmuştur. Bu sebeple bu iki modelin karşılaştırılmasında sadece doğruluk oranlarına bakmak doğru ve yeterli olmayacaktır. Tablo 4.7'de Model 9 ve Model 10 için karışıklık matrisleri ile birlikte performans ölçütleri gösterilmiştir.

Tablo 4.7. Model 9, Model 10 Karışıklık Matrisleri ve Performans Ölçütleri

BoW - LR

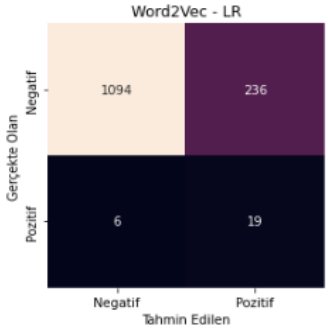
	Kesinlik	Duyarlılık	f1 skor	
Negatif	0.91	0.91	0.91	881
Pozitif	0.83	0.84	0.84	474
Doğruluk			0.89	1355
Makro Ortalama	0.87	0.87	0.87	1355
Ağırlıklı Ortalama	0.89	0.89	0.89	1355



Model 9, BoW – LR

Word2Vec - LR

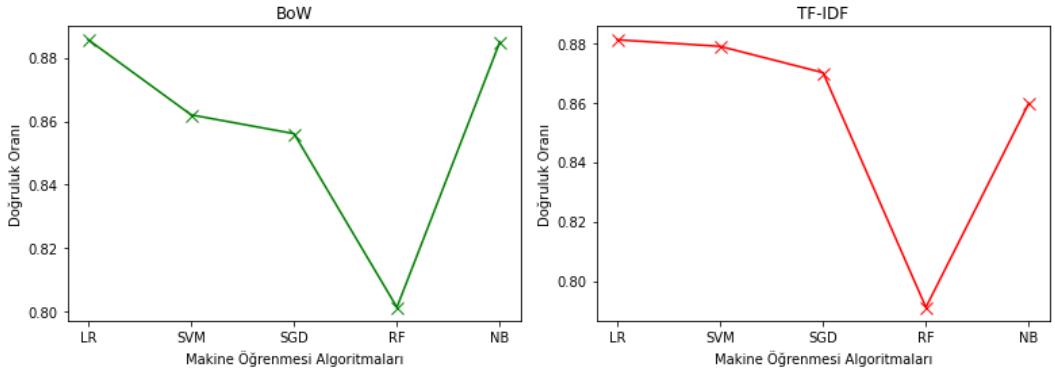
	Kesinlik	Duyarlılık	f1 skor	
Negatif	0.82	0.99	0.90	1100
Pozitif	0.76	0.07	0.14	255
Doğruluk			0.82	1355
Makro Ortalama	0.79	0.53	0.52	1355
Ağırlıklı Ortalama	0.81	0.82	0.76	1355

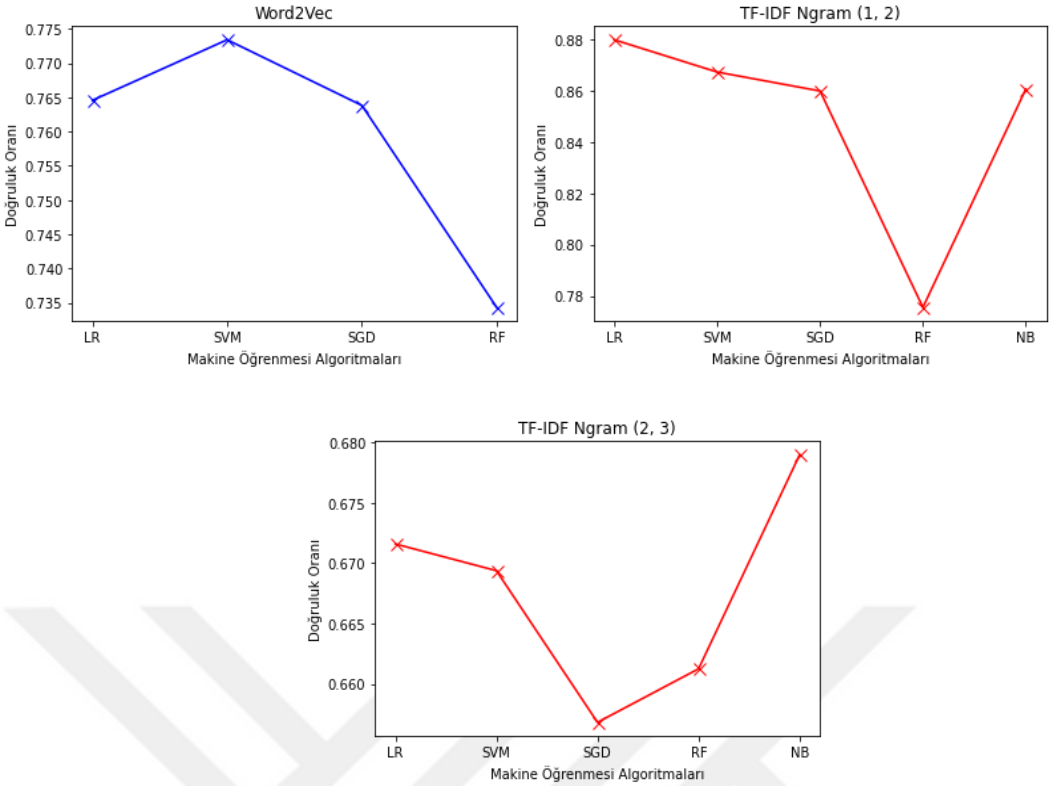


Model 10, Word2Vec – LR

Tablo 4.7'de yer alan Model 9 ve Model 10 için en yüksek doğruluk oranlarını veren sayısallaştırma ve makine öğrenmesi algoritmasına ait doğruluk ve f1 skor değerlerine bakıldığında, Model 9'a ait sonuçların büyük oranda örtüştüğü, Model 10 için ise bu değerlerin tutarsız olduğu dolayısıyla veri dağılımının düzensiz olduğu sonucuna varılmaktadır.

Model 9 için makine öğrenmesi algoritmalarının farklı sayısallaştırma yöntemleri ile kullanımına ait doğruluk oranlarının grafiksel gösterimi şekil 4.38'deki gibidir.

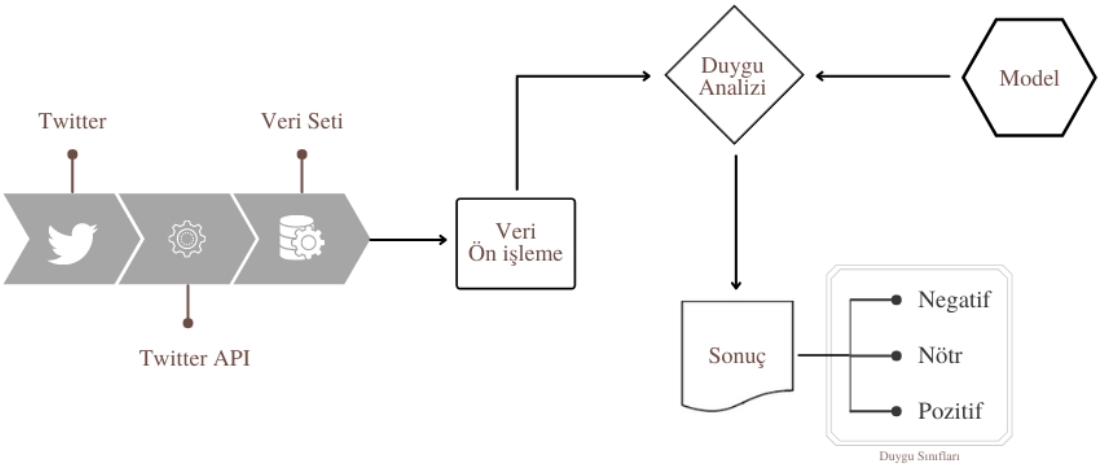




Şekil 4.38. Model 9'a Ait Doğruluk Oranlarının Grafiksel Gösterimi

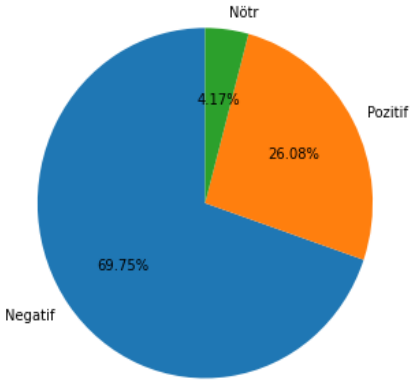
4.10. Duygu Analizi ve Görselleştirme

Türkçe metinlerden oluşan ve manuel olarak etiketlendirilen veri seti için TF-IDF – LR sınıflandırma modelinin performansı en yüksek olmuştur. Bu model kullanılarak 110.000 tweet için tahminleme yapılmıştır. Paylaşılan tweetlerin negatif, pozitif veya nötr olarak hangi duyguyu ifade ettiği belirlenerek analiz gerçekleştirilmiştir. Çalışmada kullanılan duygu analizi mimarisi şekil 4.39'teki gibidir.



Şekil 4.39. Duygu Analizi Mimarisi

TF-IDF – LR ile tüm veri setinde tahminleme yapıldığında 76.723 adet negatif, 28.691 adet pozitif ve 4586 adet nötr tweet atıldığı sonucuna varılmıştır. Duygu sınıflarının dağılım oranı şekil 4.40'ta gösterilmiştir.



Şekil 4.40. Duygu Dağılım Oranları

Covid-19 pandemisi ile hemen hemen her alanda enfeksiyon ve ölüm riskine karşı duyulan kaygı beraberinde uzaktan eğitime geçiş süreci ve sonrasında bu sürecin giderek uzaması kişilerin yapmış olduğu paylaşımlara yoğun olarak olumsuz yansıdığı görülmüştür.

4.10.1. Kelime Bulutu

Duygu analizi yapılarak negatif, nötr veya pozitif sınıflara atanan tweetler kelime bulutu ile görselleştirilmiştir. Kelime bulutu veri setinde en çok geçen kelimeleri büyük, seyrek geçen kelimeleri de küçük olarak görselleştirmektedir. Kelime bulutu ile görselleştirme işleminde her bir duygu sınıfına ait kelimeler ile tüm veri setinde yer alan kelimeler 150 kelime sınırı belirlenerek görselleştirilmiştir.

Etiketlendirilen ve sınıflandırma modellerinde kullanılan 5043 adet tweet için etiket sınıflarına göre ve tüm kelimelere ait kelime bulutları tablo 4.9'daki gibidir.

Tablo 4.9. 5043 Adet Tweet İçin Kelime Bulutları

<table>
<tr><td>  </td><td>  </td></tr>
<tr><td>Tüm Kelimeler Kelime Bulutu</td><td>Negatif Etiketli Kelime Bulutu</td></tr>
<tr><td>  </td><td>  </td></tr>
<tr><td>Nötr Etiketli Kelime Bulutu</td><td>Pozitif Etiketli Kelime Bulutu</td></tr>
</table>

Tahminleme yapılan 110.000 tweet için etiket sınıflarına ait kelimeler ve tüm kelimelere ait kelime bulutları tablo 4.10'da yer almaktadır.

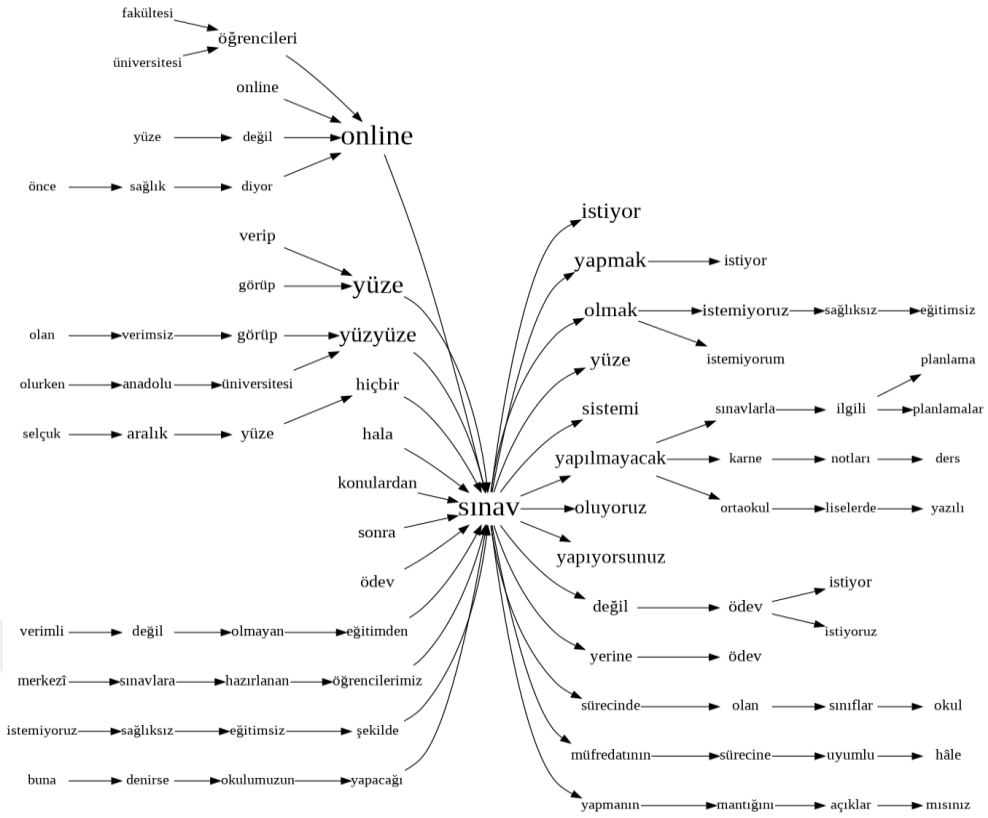
Tablo 4.10. 110.000 Adet Tweet İçin Kelime Bulutları

	
Tüm Kelimeler Kelime Bulutu	Negatif Etiketli Kelime Bulutu
	
Nötr Etiketli Kelime Bulutu	Pozitif Etiketli Kelime Bulutu

4.10.2. Kelime Ağacı

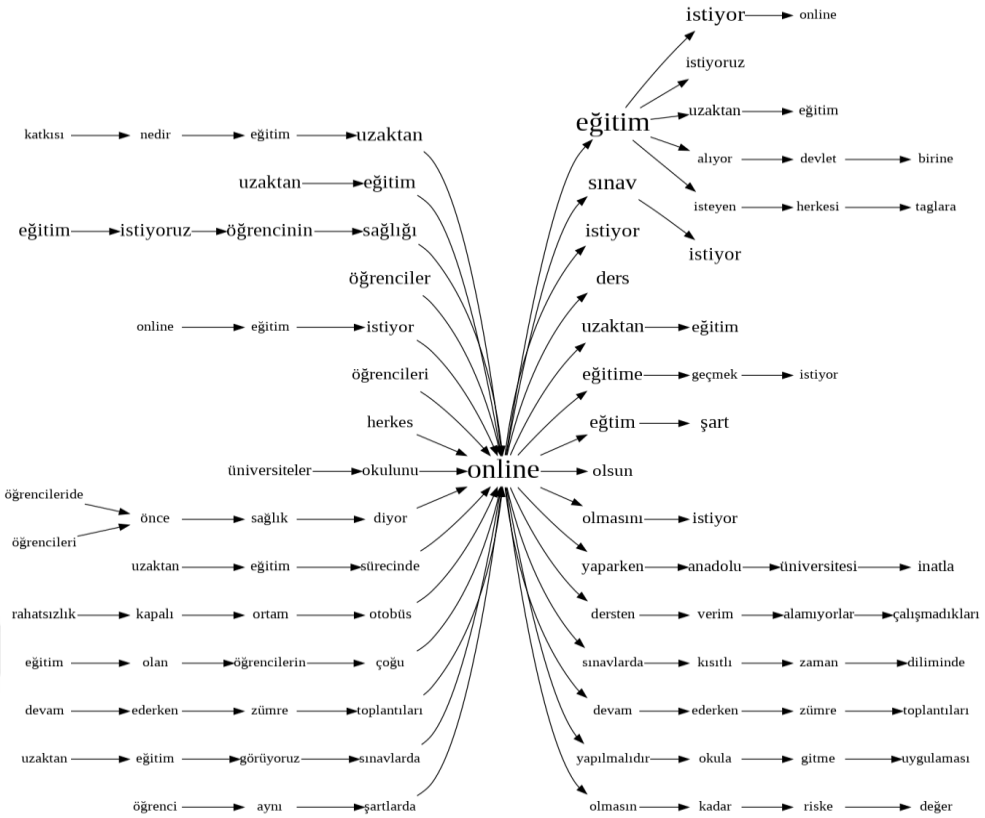
Kelime ağacı (WordTree) diyagramları, anahtar bir kelimenin bir bütünde farklı ifadelerde ne sıklıkta kullanıldığını keşfetmemizi sağlar. Bu kütüphanenin iki ana işlevi vardır: arama işlevi bir yapıdaki kelime (N-gram) frekansını sayarken, çizim ise N-gram frekanslarından bir kelime ağacı diyagramı oluşturmaktadır (Crichton, 2020).

Kelime ağacı diyagramları kullanarak duygu analizi gibi çalışmalarda anahtar bir kelime ile nelerin konuşulduğu, girilen bu anahtar kelime ile en sık kullanılan kelimelerin neler olduğu gibi sonuçlara varılabilmektedir. Çalışmada kullanılan veri setinde yer alan “sınav”, “online” ve “istemiyor” anahtar kelimeleri ile oluşturulan kelime ağacı diyagramları sırayla şekil 4.41, şekil 4.42 ve şekil 4.43’teki gibidir.



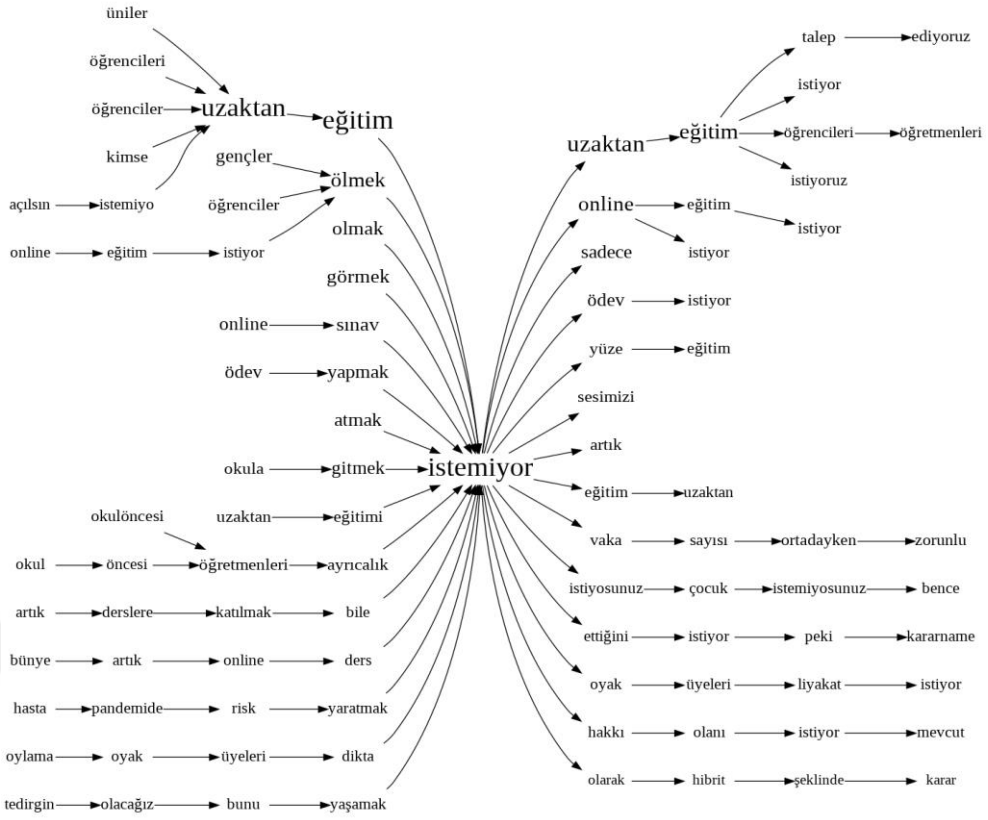
Şekil 4.41. "Sınav" Kelimesine Ait Kelime Ağacı Diyagramı

"Sınav" kelimesine ait diyagram incelendiğinde "online", "istiyor", "ödev", "yüzyüze" kelimelerinin yoğun olarak kullanıldığı, uzaktan eğitim sürecinde kişilerin sınav uygulamalarına yönelik tutumları görülebilmektedir.



Şekil 4.42. "Online" Kelimesine Ait Kelime Ağacı Diyagramı

"Online" kelimesine ait diyagram incelendiğinde yapılan paylaşımların daha çok mevcut uzaktan eğitim sistemini destekler nitelikte, yaşanan pandemi sürecinin kaygılarının taşındığı, "eğitim", "istiyor", "uzaktan" kelimelerinin ağırlıklı olarak paylaşımlarda yer aldığı görülebilmektedir.



Şekil 4.43. "İstemiyor" Kelimesine Ait Kelime Ağacı Diyagramı

"İstemiyor" kelimesine ait diyagram incelendiğinde paylaşılan tweetlerin yoğun olarak iki farklı görüşü ifade ettiği, bir grubun uzaktan eğitimi, ödev, sınav gibi uygulamaları istemediği, diğer bir grubun ise sağlık kaygısı yaşamak istemediği, okula gitmek yerine uzaktan, online eğitimi, sınav yerine ödev uygulamalarını istediği sonucuna varılabilmektedir.

5. SONUÇLAR VE ÖNERİLER

5.1. Sonuçlar

İnsanlar tarafından büyük ilgi gören ve oldukça aktif bir şekilde kullanılan sosyal medya platformları birçok alanda kullanılabilir büyük bir veri kaynağı haline gelmiştir. Ortaya çıkan bu büyük veri duygu analizi gibi çalışmaların temel veri kaynağı olmuştur. Duygu analizi uzun süredir çalışmalara konu olan bir alan olsa da Türkçe verilerle yapılmış çok fazla uygulamasının bulunmadığı, İngilizce dili ile yapılan çalışmaların literatürde oldukça büyük bir yere sahip olduğu gözlemlenmiştir.

Bu tez kapsamında sosyal medya platformu Twitter'da paylaşılan uzaktan eğitim konulu Türkçe tweetler çalışmanın veri setini oluşturmuştur. Bu veri seti çeşitli ön işlemlere tabi tutulmuş ve zemberek kütüphanesi ile normalleştirme işlemi gerçekleştirilmiştir. Verilerin ön işleme sonrasında anlamlı ve işlenebilir olması, zemberek ile imla hataları ve anlam denetiminin yapılması sınıflandırma başarısında önemli bir rol oynadığı görülmüştür.

Makine öğrenmesi algoritmalarına girdi olarak kullanılacak veri seti için manuel etiketleme işleminin yanı sıra, bir dilden başka bir dile çevrilen metinlerde duygu ifadesinde bir değişim olup olmadığını görebilmek adına İngilizce metinlerde duygu çıktıları veren ve literatürde sıkça kullanılan TextBlob, Vader ve Bert gibi hazır modeller ile etiketleme işlemi yapılmıştır. Manuel olarak etiketlenen verilere kıyasla kullanılan hazır modellerin başarısı karşılaştırılmıştır. Türkçe 'den İngilizceye çeviri işlemi sonrasında benzer duygu çıktılarının olduğu fakat Türkçenin zengin biçimsel yapısı, ironi içeren cümleler, çeviri işlemi sonrasında kullanılan hazır modellerin duygu çıktılarını etkilediği görülmüştür.

Çalışma kapsamında BoW, TF-IDF ve Word2Vec yöntemleri kullanılarak veri sayısallaştırma işlemi gerçekleştirilmiştir. Ayrıca kelimelerin Ngram yaklaşımı ile tekli ve ikili, ikili ve üçlü kullanımları TF-IDF ile sayısallaştırma işleminde uygulanmıştır. Verilerin farklı sayısallaştırma yöntemleri ile temsil edilmesi kullanılan makine öğrenmesi algoritmalarına bağlı olarak sınıflandırma performanslarında etkili olmuştur.

Sayısal olarak ifade edilen ve sınıf etiketine sahip olan veri seti eğitim ve test için ayrılarak LR, SGD, SVM, RF ve NB makine öğrenmesi algoritmaları ile sınıflandırma performansları karşılaştırılmıştır. Kullanılan makine öğrenmesi

algoritmaları ve bu algoritmaların sahip olduğu parametrelerin doğru değerlerde seçimi sınıflandırma performanslarını konuşurken doğrudan etkili olduğu görülmüştür.

Sonuç olarak bu tez çalışmasında en iyi sınıflandırma modelinin tespit edilerek devamında bu model üzerinden duygu analizi gerçekleştirmek için önceki bölümlerde bahsedilen manuel ve hazır modeller ile veri etiketleme işlemi gerçekleştirilmiş, her bir model için farklı sayısallaştırma yöntemi ve farklı makine öğrenmesi algoritmaları kullanılarak performansları karşılaştırılmıştır. Karşılaştırma işlemi sonucunda Türkçe metinlerde manuel etikete sahip modelin performansı diğer modellere göre yüksek olmuştur. Çeviri yapılarak hazır modeller ile etiketlendirilen ve Türkçe metinler ile makine öğrenmesi algoritmalarına girdi olarak kullanılan modellerin istenilen seviyede başarıya ulaşamadığı görülmüştür. Manuel olarak işaretlenmiş nötr etiketlerin veri setinden çıkartılarak oluşturulduğu modelde ikili sınıflandırma işleminde başarının arttığı gözlemlenmiştir. Aynı şekilde Bert'in Türkçe modeli ile tekrar etiketlendirilen modelde kabul edilebilir bir başarı seviyesine ulaşılmış olsa da f1 skor göz önüne alındığında sınıfların dengesiz dağıldı dolayısıyla sınıf dağılımlarının performans ölçütlerinde önemli bir rol oynadığı gözlemlenmiştir.

Türkçe metinler ve manuel etiketlere sahip veri seti için en iyi sınıflandırma performansını gösteren sayısallaştırma ve makine öğrenmesi algoritması (TF-IDF – LR) ile tüm veri seti için tahminleme yapılarak duygu analizi gerçekleştirilmiş, analiz sonucunda ise kişilerin uzaktan eğitim konusunda yapmış olduğu paylaşımlarda daha çok olumsuz duyguların yer aldığı görülmüştür.

5.2. Öneriler

Bu tez kapsamında olumlu, olumsuz ve tarafsız duygu ifadeleri kullanılarak çalışılmıştır. Gelecek çalışmalarda üzgün, mutlu, kızgın, şaşkın gibi daha ayrıntılı duygu ifadeleri belirten çalışmalar yapılabilir.

İngilizce metinler için geliştirilmiş ve yüksek başarı oranlarına sahip duygu çıktıları veren birçok hazır model bulunmaktadır. Türkçe çalışmaların çok az olduğu bu alanda bahsedilen hazır modeller Türkçe metinler için de geliştirilebilir.

Veri setinin kalitesi ve verinin doğru etikete sahip olması makine öğrenmesi algoritmalarının sınıflandırma performanslarını değerlendirmede önemli rol oynamaktadır. Türkçe metinlerde ön işleme adımlarının geliştirilip, normalleştirilmesi, aynı zamanda etiketli veri örneklerinin sayısının çoğaltılması başarı oranları artırılabilir.

KAYNAKLAR

Afrin, F. ve Nahar, I., 2015," Incremental learning based intelligent job search system", Doktora Tezi, *BRAC University*, Dakka, Bangladeş, 6-14.

Ahlgren M. ve Team WSHR, 2021, Statistics and 50+ Facts For 2020 [online], https://www.websiterating.com/research/Twitter-statistics/ [Ziyaret Tarihi: 5 Kasım 2021].

Akgül, E. S., Ertano, C. ve Diri, B., 2015, Twitter verileri ile duygu analizi, *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi,* 22 (2), 106-110.

Akyul, F., 2019, "Veri madenciliği teknikleri ile hava yolu firmalarının tweetleri üzerinden duygu analizi", Yüksek Lisans Tezi, *Burdur Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü,* Burdur, 16-21.

Albayrak, A. 2018, "Duygu analizinde farklı vektör temsil yöntemleri ve sınıflayıcıların karşılaştırılması", Yüksek Lisans Tezi, *Sivas Cumhuriyet Üniversitesi Sosyal Bilimler Enstitüsü,* Sivas, 13-43.

Alharbi, A. S. M. ve Doncker, E., 2019, Twitter sentiment analysis with a deep neural network: an enhanced approach using user behavioral information, *Cognitive Systems Research*, 54, 50-61.

Alpaydın, E., 2010, Introduction to machine learning, *the MIT Press Cambridge*, Massachusetts Londra, İngiltere.

Amanet H., 2017, "Türkçe sosyal medya metinlerinde duygu analizi", Yüksek Lisans Tezi, *Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsü,* Trabzon.

Amidi ve Amidi, 2021, Makine öğrenmesi ile refleks-temelli modeller [online], https://stanford.edu/~shervine/l/tr/teaching/cs-221/cheatsheet-reflex-models [Ziyaret Tarihi: 2 Aralık 2021].

Andırın, M. Y., 2020, Veri etiketleme nedir? Neden bu kadar önemli? [online], https://medium.com/startupmef/veri-etiketleme-nedir-neden-bu-kadar-%C3%B6nemli-e503b3c7c96f [Ziyaret Tarihi: 11 Aralık 2021].

Aninditya, A., Hasibuan, M. A., Sutoyo, E. 2019. Text mining approach using tf-ıdf and naive bayes for classification of exam questions based on cognitive level of bloom's taxonomy. *Paper presented at the 2019 IEEE International Conference on Internet of Things and Intelligence System*, IoTaIS.

Anonim, 2019, Denetimli ve denetimsiz öğrenme arasındaki fark [online], https://tr.gadget-info.com/difference-between-supervised [Ziyaret Tarihi: 2 Aralık 2021].

Anonim, 2021, Makine öğrenmesi algoritmaları [online], https://azure.microsoft.com/tr-tr/overview/machine-learning-algorithms/#overview [Ziyaret Tarihi: 2 Aralık 2021].

Anonim, 2021, Zemberek (yazılım) [online], https://tr.wikipedia-on-ipfs.org/wiki/Zemberek_(yaz%C4%B1l%C4%B1m) [Ziyaret Tarihi: 5 Aralık 2021].

Anonim, 2021, Zemberek-Python [online], https://github.com/Loodos/zemberek-python [Ziyaret Tarihi: 5 Aralık 2021].

Anonymous, 2020, Twitter-roBERTa-base for sentiment analysis [online], https://huggingface.co/cardiffnlp/Twitter-roberta-base-sentiment#Twitter-roberta-base-for-sentiment-analysis [Ziyaret Tarihi: 15 Aralık 2021].

Anonymous, 2021, Pandas package overview [online], https://pandas.pydata.org/docs/getting_started/overview.html [Ziyaret Tarihi: 15 Aralık 2021].

Anonymous, 2021, What is numpy [online], https://numpy.org/doc/stable/user/whatisnumpy.html [Ziyaret Tarihi: 15 Aralık 2021].

Ayan, B., Kuyumcu, B. ve Ciylan, B., 2019, Detection of islamophobic tweets on Twitter using sentiment analysis, *Gazi Üniversitesi Fen Bilimleri Dergisi*, 7 (2), 495-502.

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

Aydoğan, M. ve Karcı, A., 2019, Kelime temsil yöntemleri ile kelime benzerliklerinin incelenmesi. *Çukurova Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, 34 (2), 181-196.

Aykul, F., 2019, "Veri madenciliği teknikleri ile hava yolu firmalarının tweetleri üzerinden duygu analizi", Yüksek Lisans Tezi, *Burdur Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü*, Burdur, 16-19.

Ayodele, T. O., 2010. Machine learning overview. *INTECH Open Access Publisher*.

Aytekin, M. K., 2012, "Vekil sunucu verisi üzerinde ile kullanıcı sorguları kümelemesi", Yüksek Lisans Tezi, *Maltepe Üniversitesi Fen Bilimleri Enstitüsü*, İstanbul.

Ballı, Ç., 2021, "Doğal dil işleme ile Türkçe içerikli paylaşımlardan sosyal medya kullanıcılarının duygu analizi", Yüksek Lisans Tezi, *Ankara Üniversitesi Fen Bilimleri Enstitüsü*, Ankara, 21-28.

Barbieri, F., Camacho-Collados, J., Neves, L., ve Espinosa-Anke, L., 2020, Tweeteval: Unified benchmark and comparative evaluation for tweet classification [online], https://arxiv.org/pdf/2010.12421.pdf [Ziyaret Tarihi: 5 Aralık 2021].

Beck M., 2020, How to scrape tweets with snscrape [online], https://betterprogramming.pub/how-to-scrape-tweets-with-snscrape-90124ed006af [Ziyaret Tarihi: 14 Ekim 2021].

Bender, O., Och, F.J. ve Ney, H., 2003, Maximum entropy models for named entity recognition, *Proceedings Of The Seventh Conference On Natural Language Learning at HLT-NAACL*, 27 May- June 1, Edmonton, Canada, 148-151.

Berger, A.L., Pietra, V.J. Della ve Pietra, S.A. Della, 1996, A maximum entropy approach to natural language processing, *Computational Linguistics*, 22, 1, 39–71.

Beşkirli, A., Gülbandılar, E. ve Dağ, İ., 2021, Metin madenciliği yöntemleri ile Twitter verilerinden bilgi keşfi, *Journal of ESTUDAM Information*, 2 (1), 21-25.

Bilgin, M. ve Şentürk, İ. F., 2019, Danışmanlı ve yarı danışmanlı öğrenme kullanarak doküman vektörleri tabanlı tweetlerin duygu analizi, *BAUN Fen Bilimleri Enstitüsü Dergisi*, 21 (2), 822-839.

Bilgin, M., 2017, Gerçek veri setlerinde klasik makine öğrenmesi yöntemlerinin performans analizi, *Breast*, 2 (9), 683-689.

Bilgin, M., 2019, Kelime vektörü yöntemlerinin model oluşturma sürelerinin karşılaştırılması, *Bilişim Teknolojileri Dergisi*, 12 (2).

Brownlee J., 2020, One-vs-rest and one-vs-one for multi-class classification [online], https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification/ [Ziyaret Tarihi: 14 Aralık 2021].

Cebeci, H. İ., 2020, Sosyal medya verileri ile duygu analizi, *Sakarya Üniversitesi Mezunlar Derneği*, 191-211.

Chakrabarti, B., Bullmore, E., ve Baron-Cohen, S., 2006, Empathizing with basic emotions: common and discrete neural substrates. *Social Neuroscience*, 1 (3-4), 364-384.

Chao, W. L., 2011, Machine Learning Tutorial, *Digital Image and Signal Processing*.

Crichton W., 2020, Wordtree [online], https://github.com/willcrichton/wordtree [Ziyaret Tarihi: 9 Ekim 2021].

Custer C., 2020, 15 Python libraries for data science you should know [online], https://www.dataquest.io/blog/15-python-libraries-for-data-science/ [Ziyaret Tarihi: 15 Aralık 2021].

Çelik, Ö. ve Koç, B. C., 2021, TF-IDF, Word2vec ve Fasttext vektör model yöntemleri ile Türkçe haber metinlerinin sınıflandırılması, *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, 23 (67), 121-127.

Çınar, A., 2020, Sınıflandırma algoritmaları ile bir metin madenciliği uygulaması: veri madenciliği ve makine öğrenmesi, temel kavramlar, algoritmalar, uygulamalar, *Çağlayan Kitapevi ve Eğitim Çözümleri Tic. A.Ş.*, İstanbul, 105-140.

Çoban, Ö., 2016, "Metin sınıflandırma teknikleri ile türkçe Twitter duygu analizi", Yüksek Lisans Tezi, *Atatürk Üniversitesi Fen Bilimleri Enstitüsü*, Erzurum.

Çoban, Ö., Özyer, B. ve Özyer, G. T., 2015, Türkçe Twitter mesajlarının duygu analizi, *IEEE 23rd Signal Processing and Communications Applications Conference,* 16-19 Mayıs 2015, Malatya.

Dean, 2014, DEAN, Jared.: Big data, data mining, and machine learning: value creation for business leaders and practitioners. *John Wiley & Sons.*

Deveci, 2012, Denetimli ve denetimsiz makine öğrenmesi nedir? [online], https://www.elektrikport.com/haber-roportaj/denetimli-ve-denetimsiz-makine-ogrenmesi-nedir/22487#ad-image-0 [Ziyaret Tarihi: 11 Aralık 2021].

Devlin, J., Chang, M. W., Lee, K., ve Toutanova, K., 2018, Bert: Pre-training of deep bidirectional transformers for language understanding [online], https://arxiv.org/abs/1810.04805 [Ziyaret Tarihi: 5 Aralık 2021].

Dietterich, T. G., 1997. Machine-learning research. *Aı Magazine*, 18 (4), 97.

Doğan, S. ve Diri, B., 2010, Türkçe dokümanlar için n-gram tabanlı yeni bir sınıflandırma (ng ınd): yazar, tür ve cinsiyet, *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi,* 3,1 (Basılı 3).

Elmas, Ş. Ş., (2019), "Sosyal medya mesajlarının veri madenciliği yöntemi ile duygu analizi (Sivas ili örneği)", Yüksek Lisans Tezi, *Sivas Cumhuriyet Üniversitesi Sosyal Bilimler Enstitüsü,* Sivas, 5-29.

Emekli, B. ve Selvi, İ. H., 2020, GSM operatörlerine yönelik atılan Türkçe tweetlerin derin öğrenme yöntemleriyle duygu analizi, *4. Uluslararası Marmara Fen Bilimleri Kongresi,* 19-20 Haziran 2020, Online.

Emre, İ. E. ve Selçukcan Erol, Ç., 2017, Veri analizinde istatistik mi veri madenciliği mi? *Bilişim Teknolojileri Dergisi*, 10 (2). 161-167.

Erden, C. 2021, Pyton ileveri madenciliği, 1. Baskı, *İnkılap Kitabevi Yayın San. Tic. A.Ş.*, İstanbul, 1-4.

Fayyad P. S., ve Smyth, P., 1996, The KDD Process for extracting useful knowledge from volumes of data, *Communıcatıons Of The ACM*, 39 (11), 27–34.

Gandhi, R., 2018, Naive bayes classifier [online], https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c [Ziyaret Tarihi: 02 Aralık 2021].

Gheorghe R., 2018, Generating word embeddings with gensim's word2vec [online], https://sematext.com/blog/word-embeddings-gensim-word2vec-tutorial/ [Ziyaret Tarihi: 20 Aralık 2020].

Gordin, D. M., 2015, Scientific babel: how science was done before and after global English. Chicago, *Illinois: University of Chicago Press*.

Han, J., Kamber, M., ve Pei, J., 2011, Data mining: concepts and techniques: concepts and techniques. *Elsevier*.

Hatipoğlu, E., 2018, Machine Learning — Classification — Logistic Regression — Part 8 [online], https://medium.com/@ekrem.hatipoglu/machine-learningclassification-logistic-regression-part-8-b77d2a61aae [Ziyaret Tarihi: 15 Mayıs 2021].

İlhan, N. ve Sağaltıcı, D., 2020, Twitter'da duygu analizi, *Harran Üniversitesi Mühendislik Dergisi*, 5 (2), 146-156.

Kaban, Z., ve Diri, B., 2008, Genre and author detection in Turkish texts using artificial immune recognition systems. *In 2008 IEEE 16th Signal Processing, Communication and Applications Conference*, 1-4).

Karamanlı, E., 2019, "Makine öğrenmesi algoritmaları kullanarak, metin madenciliği ve duygu analizi ile müşteri deneyiminin geliştirilmesi", Yüksek Lisans Tezi, *İstanbul Üniversitesi Sosyal Bilimler Enstitüsü,* İstanbul, 3-8.

Karaöz, B., 2018, "Büyük veri ve işletme analitiği: sosyal medya ve duygu analizi ile bir öngörü modeli", *İstanbul Üniversitesi Sosyal Bilimler Enstitüsü*, İstanbul, 47-50.

Kaynar, O., Yıldız, M., Görmez, Y. ve Albayrak, A., 2016, Makine öğrenmesi yöntemleri ile duygu analizi. *International Artificial Intelligence and Data Processing Symposium*. 17-18 September, Malatya.

Keskin M., 2018, Word2Vec, FastText, GloVe [online], https://medium.com/codable/word2vec-fasttext-glove-d4402fa8cce0 [Ziyaret Tarihi: 20 Aralık 2020].

Kızılırmak, E.,2020, "İngilizce–Türkçe çeviri metinlerde levenshteın uzaklığı ile desteklenmiş çapa tabanlı cümle eşleme", Yüksek Lisans Tezi, *Maltepe Üniversitesi Lisansüstü Eğitim Enstitüsü*, İstanbul, 1-8.

Korkusuz, R., 2018, "Futbola ilişkin Twitter paylaşımlarının duygu analizi", Yüksek Lisans Tezi, *Trakya Üniversitesi Fen Bilimleri Enstitüsü*, Edirne, 13-21.

Kotsiantis, S. B., Zaharakis, I. ve Pintelas, P., 2007, Supervised machine learning: A review of classification techniques, *Informatica*, 31, 249-268.

Kouloumpis, E., Wilson, T. ve Moore, J. D., 2011, Twitter sentiment analysis: the good the bad and the omg, *Fifth International AAAI Conference on Weblogs and Social Media, ICWSM*, Barcelona, 538-541.

Köksal, B., Erdem, G., Türkeli, C., ve Öztürk, Z. K., 2021, Twitter'da duygu analizi yöntemi kullanılarak bitcoin değer tahminlemesi. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, 9 (3), 280-297.

Kshirsagar, R., Cukuvac, T., McKeown, K. ve McGregor, S., 2018, Predictive embeddings for hate speech detection on Twitter [online]. https://arxiv.org/pdf/1809.10644.pdf [Ziyaret Tarihi: 24 Ekim 2021].

Kumaş, E., 2021, Türkçe Twitter verilerinden duygu analizi yapılırken sınıflandırıcıların karşılaştırılması, *ESTUDAM Bilişim Dergisi*, 2 (2), 1-5.

Kuzucu, K., 2015, "Müşteri memnuniyeti belirlemek için metin madenciliği tabanlı bir yazılım aracı", Yüksek Lisans Tezi, *Maltepe Üniversitesi Fen Bilimleri Enstitüsü*, İstanbul.

Malkoç, B., 2012, Temel bilimler ve mühendislik eğitiminde programlama dili olarak python. XIV. *Akademik Bilişim Konferansı Bildirileri*, 201.

Medhat, W., Hassan, A., Korashy, H. M., 2014, Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5.4: 1093-1113.

Mikolov, T., Chen, K., Corrado, G., ve Dean, J., 2013, Efficient estimation of word representations in vector space, *International Conference on Learning Representations*, Arizona, USA, 1-12, 2-4.

Nalçakan, Y., Bayramoğlu, Ş. S. ve Tuna, S., 2015, Sosyal medya verileri üzerinde yapay öğrenme ile duygu analizi çalışması. *Trakya Üniversitesi*.

Nasukawa, T. ve Yi, J., 2003, Sentiment analysis: capturing favorability using natural language processing, *Proceedings of the 2nd International Conference on Knowledge Capture*, 23-25 Ekim, Sanibel Island, FL, USA,70–77.

Oflazer, K., Bozşahin H.C., 2006, Türkçe doğal dil işleme [online]. Ç.Ü. Türkoloji-Makale Bilgi Sistemi [online], http://turkoloji.cu.edu.tr/DILBILIM/turkce_dogal_dil_isleme.pdf [Ziyaret Tarihi: 9 Ekim 2021].

Onan, A., 2017, Twitter mesajları üzerinde makine öğrenmesi yöntemlerine dayalı duygu analizi, *Yönetim Bilişim Sistemleri Dergisi*, 3 (2), 1-14.

Özgür, A., 2004, "Supervised and unsupervised machine learning techbiques for text document categorization", Yüksek Lisans Tezi, *Boğaziçi Üniversitesi Fen Bilimleri Enstitüsü*, İstanbul.

Özkan, Y., 2008, Veri madenciliği yöntemleri, 1.baskı, *Papatya Yayıncılık Eğitim*, İstanbul.

Özyurt, Ö. ve Kısa N., 2021, Covid-19 salgını sürecinde uzaktan eğitime ilişkin Tweetlerin duygusal analizi. *Journal of Computer and Education Research,* 9 (18), 853-868.

Pandey, P., 2018, simplifying sentiment analysis using VADER in python [online], https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f [Ziyaret Tarihi:15 Ekim 2021].

Ramos, J., 2003, Using tf-idf to determine word relevance in document queries. *In Proceedings Of The First İnstructional Conference On Machine Learning,* 242 (1), 29-48.

Rehurek R., 2021, What is gensim? [online], https://radimrehurek.com/gensim/intro.html [Ziyaret Tarihi: 19 Aralık 2021].

Rosenfield, R., ve Clarkson, P., 1997, Statistical language modeling using the cmucambridge toolkit. *5th European Conference on Speech Communication and Technology.* In Eurospeech.

Sar, K. T., 2021, "Yapay sinir ağları ve bert dil modeli kullanılarak zaman bazlı duygu analizi: Whatsapp yeni gizlilik sözleşmesine yönelik yorumların araştırılması", Yüksek Lisans Tezi, *Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü*, İzmir.

Sariman G ve Mutaf E., 2020, Covıd-19 Sürecinde Twitter mesajlarının duygu analizi, *Euroasia Journal of Mathematics Engineering Natural and Medical Sciences,* 7 (10), 137-148.

Schapire, R. E., 2003. The boosting approach to machine learning: An overview. *In Nonlinear Estimation And Classification*, New York, 149-171.

Shehu, H. A., Tokat, S., Sharif, Md. H. ve Uyaver, S., 2019, Sentiment analysis of Turkish Twitter data, *Third International Conference of Mathematical Sciences,* 9, 1-5.

Stolcke, A., 2002, SRILM-an extensible language modeling toolkit. *In Seventh İnternational Conference On Spoken Language Processing*.

Subramanian, D., 2019, Sentiment analysis on Ellen's DeGeneres tweets using TextBlob [online], https://medium.com/analytics-vidhya/sentiment-analysis-on-ellens-degeneres-tweets-using-textblob-ff525ea7c30f [Ziyaret Tarihi:15 Ekim 2021].

Swaminathan, S., 2018, Logistic regression — detailed overview [online] https://towardsdatascience.com/logistic-regressiondetailed-overview-46c4da4303bc [Ziyaret Tarihi: 15 Mayıs 2021].

Şeker, Ş. E., 2014, Metin madenciliği: bilgisayar kavramları [online], http://bilgisayarkavramlari.sadievrenseker.com [Ziyaret Tarihi: 10 Ekim 2021].

Topaçan, Ü., 2016, Sosyal medya paylaşımlarında duygu analizi: makine öğrenimi yaklaşımı üzerine bir araştırma, Doktora Tezi, *Marmara Üniversitesi Sosyal Bilimler Enstitüsü*, İstanbul.

Tunalı, V., 2009, Metin madenciliği [online], http://www.vtunali.com/tr/index.php/2009/10/metin-madenciligi-text-miningnedir/ [Ziyaret Tarihi: 25 Ağustos 2021].

Turhost, 2021, Makine öğrenmesi nedir? [online] https://www.turhost.com/blog/makine-ogrenmesi-machine-learning-nedir/ [Ziyaret Tarihi: 15 Ekim 2021].

Türkmenoğlu, C., 2015, "Türkçe metinlerde duygu analizi", Yüksek Lisans Tezi, *İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü*, İstanbul.

Uçar, K.T., 2020, BERT modeli ile Türkçe metinlerde sınıflandırma yapmak [online], https://medium.com/@toprakucar/bert-modeli-ile-t%C3%BCrk%C3%A7e-metinlerde-s%C4%B1n%C4%B1fland%C4%B1rma-yapmak-260f15a65611 [Ziyaret Tarihi: 15 Ekim 2021].

Vikipedi, 2021, Algoritma [online], https://tr.wikipedia.org/wiki/Algoritma#:~:text=Algoritma%2C%20belli%20bir%20problemi%20%C3%A7%C3%B6zmek,durumunda%20sonlanan%2C%20sonlu%20i%C5%9Flemler%20k%C3%BCmesidir [Ziyaret Tarihi: 15 Ağustos 2021].

Wehrmann, J., Becker, W., Cagnini, H. E. L. ve Barros, R. C., 2017, A character-based convolutional neural network for language-agnostic Twitter sentiment analysis [online], https://ieeexplore.ieee.org/document/7966145 [Ziyaret Tarihi: 24 Mart 2021].

Yelmen, İ., 2016, "Doğal dil işleme yöntemleriyle Türkçe sosyal medya verileri üzerinde duygu analizi", Yüksek Lisans Tezi, *İstanbul Aydın üniversitesi Fen Bilimleri Enstitüsü*, İstanbul, 21-33.

Yelmen, İ., Zontul, M., Kaynar, O. ve Sönmez, F., 2018, A novel hybrid approach for sentiment classification of Turkish tweets for GSM operators, *International Journal of Circuits, Systems and Signal Processing*, 12, 637-645.

Yıldırım, S., 2018, "Twitter verileriyle duygu analizi ve Türkçe duygu kütüphanesi, Yüksek Lisans Tezi, *Bahçeşehir Üniversitesi Fen Bilimleri Enstitüsü*, İstanbul, 5-17.

Yıldırım, S., 2021, "Bert-base Turkish Sentiment Model" [online] https://huggingface.co/savasy/bert-base-turkish-sentiment-cased [Ziyaret Tarihi: 14 Kasım 2021].

Yılmaz, M. C., ve Orman, Z., 2021, LSTM derin öğrenme yaklaşımı ile covid-19 pandemi sürecinde Twitter verilerinden duygu analizi. *Acta Infologica*, 5 (2).

Yurt, E. A., 2015, “Türkçe metinlerde duygu Analizi”, Yüksek Lisans Tezi, *Maltepe Üniversitesi Fen Bilimleri Enstitüsü,* İstanbul.

Yüceoğlu B., 2017, Scikit-leearn ile veri analitiğine giriş [online], http://www.veridefteri.com/2017/11/23/scikit-learn-ile-veri-analitigine-giris/ [Ziyaret Tarihi: 16 Aralık 2021].