ACTA INFOLOGICA 2021;5(2):359-372
dergipark.org.tr/acin

İSTANBUL UNIVERSITY PRESS

DOI: 10.26650/acin.947747

RESEARCH ARTICLE

LSTM Derin Öğrenme Yaklaşımı ile Covid-19 Pandemi Sürecinde Twitter Verilerinden Duygu Analizi

Sentiment Analysis from Twitter Data during the Covid-19 Pandemic Era with LSTM Deep Learning Approach

Mehmet Can Yılmaz[1], **Zeynep Orman**[2]

[1] (Lisans Öğrencisi), İstanbul Üniversitesi-Cerrahpaşa, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye
[2] (Doç. Dr.), İstanbul Üniversitesi-Cerrahpaşa, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye

ORCID: M.C.Y. 0000-0001-6282-111X; Z.O. 0000-0002-0205-4198

Corresponding author:
Zeynep ORMAN
İstanbul Üniversitesi-Cerrahpaşa, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye
E-mail address: ormanz@istanbul.edu.tr

Submitted: 08.06.2021
Revision Requested: 27.06.2021
Last Revision Received: 01.07.2021
Accepted: 16.07.2021
Published Online: 27.10.2021

Citation: M. C., ve Orman, Z. (2021). LSTM Derin öğrenme yaklaşımı ile Covid-19 pandemi sürecinde Twitter verilerinden duygu analizi. *Acta Infologica, 5*(2), 359-372. https://doi.org/10.26650/acin.947747

ÖZ

Dünyada yaşanan toplumsal olaylar için insanların düşüncelerini anlamak ve bu düşünceleri analiz ederek birtakım çıkarımlar yapmak oldukça önemlidir. Bu analiz ve çıkarımlar sayesinde çeşitli projeler başlatılabilir ve karar verme süreçleri oluşturulabilir. Bu amaçla kullanılan işlemlerden biri de metinlerin çeşitli bilgisayar algoritmaları ile sınıflandırılmasıyla gerçekleştirilen duygu analizi işlemidir. Duygu analizini gerçekleştirmek için kullanılan yöntemler genel olarak sözlük tabanlı yöntemler ve makine öğrenmesi yaklaşımları olarak ikiye ayrılır. Bu makalede, dünyayı etkisi altına alan ve halen devam etmekte olan koronavirüs pandemisi (Covid-19) ile ilgili Twitter sosyal medya platformunda sık konuşulan bir takım terimler gözönüne alınarak duygu analizi çalışması gerçekleştirilmiştir. Bunun için, konu ile ilgili bazı Türkçe başlıklar toplanmış ve bu başlıklar olumlu ve olumsuz düşünceler şeklinde sınıflandırılarak duygu analizi yapılmıştır. Bu analiz için derin öğrenme yöntemlerinden biri olan Uzun Kısa Süreli Hafıza (LSTM) yapısı kullanan bir sistem önerilmiştir. Önerilen bu sistem oluşturulan veri kümelerine uygulanmış ve maksimum %97 doğruluk başarısı elde edilmiştir.
Anahtar kelimeler: Duygu Analizi, Covid-19, LSTM

ABSTRACT

It is very important to understand people's thoughts regarding social events occurring in the world and to make some inferences by analyzing these thoughts. With these analysis and inferences, various projects can be initiated and decision-making processes can be formed. One of the procedures used for these purposes is the sentiment analysis which is performed by classifying text with various computer algorithms. The methods used to perform sentiment analysis are generally categorized as dictionary-based methods and machine learning approaches. In this paper, a sentiment analysis study has been carried out by considering a number of frequently spoken terms on the Twitter social media platform regarding the coronavirus (Covid-19) pandemic, which has affected the world and is still ongoing. For this, some Turkish titles related to the subject were collected and sentiment analysis was conducted by classifying these titles as positive and negative thoughts. For this analysis, a system using a Long Short-Term Memory (LSTM) structure, which is one of the deep learning methods, was proposed. The proposed system was applied on the obtained data sets and a maximum 97% accuracy was achieved.
Keywords: Sentiment Analysis, Covid-19, LSTM

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License

1. GİRİŞ

İletişim teknolojilerindeki ilerlemeler hayatımızı köklü bir şekilde değiştirmiş ve İnternet günümüzün vazgeçilmez bir parçası haline gelmiştir. İnternetin gelişmesiyle beraber birçok sosyal medya platformu yeni iletişim araçları olarak ortaya çıkmıştır. Gazete, televizyon ve radyo gibi geleneksel iletişim araçlarına yeni bir alternatif olan sosyal medya araçlarının popülerliği ve kullanımı her geçen gün artmaktadır. Sosyal medya araçlarıyla birlikte milyonlarca insanın bir konu hakkında düşüncelerine, görüşlerine ve değerlendirmelerine erişebilme imkanı doğmuştur. Bu araçların yoğun bir şekilde kullanımı sonucunda çeşitli konular hakkında üzerinde çalışmaların gerçekleştirilebileceği ve anlamlı çıkarımların yapılabileceği büyük veri ve bilgiler oluşmaktadır. Kuruluşlar, şirketler ve devletler karar mekanizmalarında insanların düşüncelerini önemser ve bu düşüncelere göre çalışma politikalarını belirler. Son zamanlarda, bu kurumlar sosyal medyadaki bilgileri kullanarak çeşitli çıkarımlar yapmaktadır. İnsanların görüşlerini, değerlendirmelerini ve duygularını analiz eden çalışma alanı duygu analizi olarak tanımlanmaktadır. Duygu analizinde genellikle bir kişinin yazdığı metnin olumlu, olumsuz ya da nötr duygu beslediği analiz edilmektedir. Bu analizler birçok şirketin ya da devletin karar verme sürecine doğrudan etki yapabildiği için oldukça değerli bilgilerdir.

Literatürde duygu analizi ile ilgili yapılan çalışmaların sayısı son yıllarda hız kazanmıştır. (Kaynar, Görmez, Yıldız ve Albayrak, 2016) çalışmasında Internet Movie Database (IMDb) haber kaynağında yer alan film yorumlarından makine öğrenmesi algoritmalarını kullanarak duygu analizi yapmışlardır. Bu çalışmada sınıflandırma için kullanılan algoritmalar Naive Bayes (NB), Merkez Tabanlı Sınıflayıcı, Çok Katmanlı Yapay Sinir Ağları (YSA) ve Destek Vektör Makineleri (SVM – Support Vector Machine) algoritmalarıdır. Sınıflandırma performanslarını karşılaştırmak için model başarım ölçütleri kullanılmıştır. Bu ölçütler incelendiğinde en yüksek başarıyı YSA ve SVM algoritmaları elde etmiştir. Eğitim veri kümesinde YSA %89.73 başarı yüzdesiyle %84.07 başarı yüzdesi olan SVM algoritmasından daha yüksek başarı oranı elde etmiştir. (Albayrak, Topal ve Altıntaş, 2017) çalışmasında Twitter üzerinden konuşulan bir konunun duygu analizini gerçekleştirmişlerdir. Sözkonusu çalışmada, Twitter üzerinden "bedelli askerlik" konu başlıklı tweetler Twitter API kullanarak çekilmiştir. Daha sonra çekilen veriler python içerisinde bulunan NLTK kütüphanesi ile veri ön işleme sürecininden geçirilmiştir. Veri ön işleme sürecinden sonra elde edilen veri kümesi ile SentiTurkNet veri kümesi karşılaştırılmış ve her kelimenin polarite skoru belirlenmiştir. Polarite sonuçlarına göre tweetlerin %16'sı pozitif, %5'i negatif ve %79'u nötr olduğu sonucuna varılmıştır. Bedelli askerlikle ilgili genel olarak insanların ne olumlu ne de olumsuz bir düşünceye sahip olduğu belirlenmiştir. (Aytuğ, 2017) çalışmasında Twitter'da paylaşılan Türkçe tweetlerden makine öğrenmesi algoritmalarıyla duygu analizi yapmıştır. Bu çalışmada, Twitter API aracılığıyla bir aylık sürede Türkçe tweetler çekilmiştir. Bu süreçten sonra veri ön işleme süreci gerçekleştirilmiştir. Çalışmada öznitelik kümelerinin oluşturulması için N-gram modeli kullanılmıştır. Sınıflandırma işlemi için NB, SVM ve Lojistik Regresyon (LR – Logistic Regression) gibi makine öğrenmesi algoritmaları kullanılmıştır. Sınıflandırma algoritmalarının ve öznitelik kümelerinin değerlendirilmesinde doğru sınıflandırma oranı, F-ölçütü ve ROC (Receiver Operating Characteristic) eğrisi altında kalan alan performans metrikleri kullanılmıştır. Sınıflandırma başarıları incelendiğinde NB algoritması ile 1-gram ve 2-gram öznitelik kümelerinin birleştirilmesiyle oluşan öznitelik kümesinin en yüksek başarı oranına ulaştığı görülmüştür. (Salur ve Aydın, 2020) çalışmasında GSM operatörlerine yönelik atılan tweetlerden derin öğrenme algoritmalarıyla duygu analizi gerçekleştirilmiştir. Sınıflandırma işlemi için dört farklı derin öğrenme algoritmasını kullanarak sınıflandırma başarıları karşılaştırılmıştır. Evrişimli Sinir Ağları (CNN – Convolutional Neural Network) ve Uzun Kısa Süreli Hafıza (LSTM – Long Short-Term Memory) modeli ile yaklaşık %82'lik başarı oranıyla en yüksek başarı oranı elde edilmiştir. (Ayvaz, Yıldırım ve Salman, 2019) çalışmasında Twitter'da popüler olan konu başlıkları ile ilgili yazılan tweetlerden duygu analizi yapılmıştır. Çalışmada iki farklı veri kümesi kullanılmıştır. Bunlar hava durumu veri kümesi ve survivor televizyon programı veri kümesidir. Bu çalışmada duygu kütüphanesi kullanılarak analiz gerçekleştirilmiştir. Elle veri üzerinde incelemeler yapılmış ve bu incelemelerle duygu kütüphanesi düzenlenmiştir. Böylece her kelimenin duygu polaritesi hesaplanmıştır. Hava durumu veri kümesinin duygu analizi sonuçları incelendiğinde "yaz" etiketi için nötr, "ilkbahar" etiketi için daha çok olumlu ve nötr, "sonbahar" etiketi için olumlu ve "kış" etiketi için daha çok nötr ve olumsuz duygu olduğu sonucuna varılmıştır. Survivor televizyon programı veri kümesinin duygu analizi sonuçları incelendiğinde %27'si olumlu, %39'u nötr ve %34'ünün olumsuz duygu olduğu sonucuna varılmıştır. (Chintalapudi, Battineni ve Amenta, 2021) çalışmasında, Covid-19 sürecinde atılan tweetlerden duygu analizi yapılmıştır. Veriler, 23 Mart 2020 ile

15 Temmuz 2020 tarihleri arasında toplanan tweetleri içermektedir ve duygular korku, üzüntü, öfke ve neşe olarak etiketlenmiştir. Veri analizi ve metin analizi yeni bir derin öğrenme modeli olan Bidirectional Encoder Representations (BERT) modeli ile gerçekleştirilmiş ve önerilen model LR, SVM ve LSTM gibi diğer modellerle karşılaştırılmıştır. Her duygu için doğruluk ayrı ayrı hesaplanmıştır. BERT modeli %89 doğruluk üretmiştir ve diğer üç model sırasıyla %75, %74.75 ve %65 doğruluk üretmiştir. Her duyarlılık sınıflandırması, metin madenciliği algoritmalarında nispeten önemli bir değer olan %79.34 medyan doğrulukla %75.88-87.33 arasında değişen bir doğruluğa sahiptir. (Manguri, Ramadhan ve Amin, 2020) çalışmasında dünya çapındaki Covid-19 salgınlarıyla ilgili twitter verilerinden duygu analizi gerçekleştirmişlerdir. Veriler, koronavirüsün en yaygın haftalarından biri olan 09-04-2020 ile 15-04-2020 arasında, Twitter API ve tweepy python kütüphanesi kullanılarak toplanmıştır. #Koronavirüs ve #COVID-19 anahtar kelimeleri seçilmiştir. 530.232 sayıda tweet toplanmıştır. Elde edilen tweetlerin polaritelerinin belirlenmesi için TextBlob kütüphanesi kullanılmıştır. Sonuçlar incelendiği zaman hem koronavirüs hem de covid-19 anahtar kelimelerinin polarite için önemli ölçüde yüksek olduğu görülmüştür ve kayıtların büyük kısmının yaklaşık %64 oranında objektif olduğu değerlendirilmiştir. (Sarıman ve Mutaf, 2020) çalışmasında Covid-19 sürecinde konuşulan önemli bazı Türkçe konu başlıkları hakkında yazılan tweetlerden duygu analizi yapılmıştır. Bu çalışmada ilk olarak "maske, sokağa çıkma yasağı, kısa çalışma ödeneği, eba" konu başlığı altında atılan tweetler Twitter API aracılığıyla çekilmiştir. Tweetler çekildikten sonra sırayla veri ön işleme, terim ağırlıklandırma, öznitelik oluşturma aşamaları gerçekleştirilmiş. Daha sonar, veriler eğitim ve test veri kümesi olmak üzere ikiye ayrılmıştır. Sınıflandırma için kullanılan algoritma, makine öğrenmesi algoritmalarından lojistik regresyon algoritmasıdır. Eğitim ve test veri kümesi oluşturulurken iki farklı yöntem kullanılmıştır. Birinci yöntemde eğitim kümesinde olumlu olumsuz veriler yarı yarıya işleme alınmıştır. Bu yöntemde en yüksek başarı oranı %82.84 olarak belirlenmiştir. İkinci yöntemde eğitim kümesi olumlu olumsuz kelimelerden çıkarımlarla belirlenmiştir. Bu yöntemde en yüksek başarı oranı %98.13 olarak belirlenmiştir. Sonuçlar incelendiği zaman maske konu başlığı altında atılan tweetlerin genelde olumlu olduğu fakat diğer konu başlığı altında atılan tweetlerin genelde olumsuz olduğu görülmüştür.

Bu çalışmada, Covid-19 sürecinde Twitter sosyal medya aracından atılan Türkçe tweetlerden duygu analizi çalışması yapılmıştır. Literatürde bu konuda yapılmış benzer çalışmalar incelendiğinde Covid-19 ile ilgili Türkçe tweetlerden duygu analizi çalışmalarının az sayıda olduğu ve bu çalışmalarda da genel olarak geleneksel makine öğrenmesi algoritmalarının kullanıldığı görülmüştür. Bu makale çalışmasında ise, derin öğrenme algoritmalarından LSTM algoritması kullanılarak duygu analizi gerçekleştirilmiştir. İlk olarak, 11 Mart 2020 tarihinden ilk normalleşme adımlarının başladığı 1 Haziran 2020'ye kadar olan süreçte Covid-19 hakkında konuşulan önemli konuların analizi yapılmıştır. 5 başlık altında toplanan konular için yapılan yorumlar olumlu ve olumsuz şeklinde sınıflandırılarak genel bakış çıkarılmış ve daha sonra oluşturulan veri kümeleri LSTM algoritmasıyla eğitilmiştir. Türkçe tweetler üzerinde yapılan duygu analizi çalışmalarında genel doğal dil işleme yöntemlerinin kullanılması ile yüksek başarımlı sonuçlar elde edilememesinden dolayı, bu çalışmada derin öğrenme algoritmasıyla sınıflandırma yapılarak sonuç elde edilmiştir. Çalışmanın 2. bölümünde Covid-19, duygu analizi yöntemleri, derin öğrenme, LSTM algoritması ile ilgili genel bilgiler verilmiştir. 3. bölümde ise çalışmada önerilen sistem yapısı ve geliştirilen modelin temel adımları anlatılmıştır. 4. bölüm olan bulgular kısmında veri kümelerindeki tweet sayıları, ön işlemden önce ve sonra oluşturulan veri kümeleri ve modelin doğruluk sonuçları ile ilgili bilgiler verilmiş olup son bölümde ise genel olarak elde edilen sonuçlar tartışılmıştır.

2. MATERYAL VE METOT

2.1. Covid-19

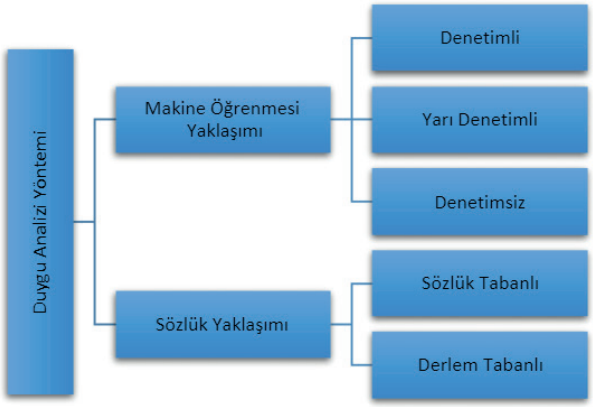
(Aydın ve Doğan, 2020) yapmış oldukları çalışmada belirttikleri gibi, Covid-19 yani tam adıyla koronavirüs hastalığı 2019 yüksek ateş, ökrüsük ve nefes darlığı gibi solunum yolu hastalıklarına neden olan bir virüstür. Virüs ilk olarak Aralık 2019'da Çin'in Wuhan kentinde ortaya çıkmıştır. İlk olarak 31 Aralık 2019'da Dünya Sağlık Örgütü (WHO) tarafından daha önce bilinmeyen bir virüsün ortaya çıktığını ilan etmiştir. Böylece Covid-19 dünya genelinde bilinir bir hale gelmiştir. Covid-19 WHO tarafından "epidemi" olarak belirlenmiştir. Fakat virüs nefes ve hava yolu ile insandan insana hızlı bir şekilde bulaşmasıyla beraber kısa bir süre içinde diğer şehirlere hatta ülkelere sıçramıştır. Bu yayılmalar ile birlikte WHO tarafından 11 Mart 2020 tarihinde korona virüs "pandemi" yani diğer bir deyiş ile coğrafi salgın olarak belirlenmiştir. Covid-19 bulaşıcılık

oranının çok yüksek ve hızlı olduğunu gören devletler salgından ülkelerini koruyabilmek için bir takım tedbirler almışlardır. Bu tedbirler genel olarak sağlık, ekonomi ve eğitim alanlarındadır. Sokağa çıkma yasağı, maske zorunluluğu, ülkeler arası uçuşların durdurulması gibi birçok kısıtlama getirilmiştir. 29 Kasım itibariyle dünya genelinde toplam vaka sayısı 62 milyon kişiyi geçmiş olup toplam vefat edenlerin sayısı ise 1 milyon 450 bini geçmiştir.

Türkiye salgını önleyebilmek için erkenden önlem alan ülkelerden birisi olmuştur. Bu önlemlerden bazıları başka ülkelerden gelen insanların kontrolü ve bazı uçuş seferlerinin durdurulmasıdır. Fakat alınan tedbirler yeterli olmamıştır. Türkiye'de 11 Mart 2020 tarihinde ilk korona virüs vakası tespit edilmiştir. Ülkemizde ilk vakanın görülmesiyle birlikte alınan tedbirler sıkılaştırılmıştır. Bu alınan sıkı tedbirler genel olarak şunlardır; tüm eğitim öğretim faaliyetlerine ara verilmesi, bazı ülkelere uçuşların durdurulması, 65 yaş üstü ve 20 yaş altı sokağa çıkma yasağı, 31 ilin giriş ve çıkışlarının kapatılması, umreden gelen vatandaşların 14 gün karantinaya alınması, bazı illerde hafta sonu sokağa çıkma yasağı, maske takma zorunluluğu ve bazı işletmelerin kapatılmasıdır. Yeni tedbirler ile birlikte düşen vaka ve ölüm sayılarının sonucunda 1 Haziran 2020 tarihinde yeni normalleşme süreci başlamıştır. Günlük vaka sayıları belirli bir sayının altına kadar inmiştir. Fakat şuan Kasım 2020 itibariyle artan vaka sayılarından dolayı tekrar sıkı önlemler alınmaya başlamıştır. 29 Kasım itibariyle Türkiye genelinde toplam vaka sayısı 600 bini geçmiş olup toplam vefat edenlerin sayısı 13 bini geçmiştir.

2.2. Duygu Analizi

İnsanların duygu, düşüncelerinin gerekli bilgisayar algoritmaları ile analiz edilmesine duygu analizi denir. Duygu analizi, insanların duygularını olumlu, olumsuz veya nötr olarak sınıflandırmayı amaçlar. Bu sınıflandırma işlemi sonucunda bir ya da birden çok yazarın o konu hakkındaki düşüncesine karar verilmiş olunur. Duygu analizi çalışmaları günümüzde yaygın olarak kullanılır. Duygu analizi ile elde edilen bilgiler oldukça değerlidir. Bu çerçevede yapılan çalışmalar ile elde edilen sonuçlar birçok şirketin veya devletin bir konu hakkında insanların ne düşündüğünü bilmesini sağlayarak karar verme süreçlerini doğrudan etkilemektedir. Örnek olarak bir medya şirketinin yeni bir dizi çıkardığını varsayalım. Bu dizinin ilk sezonu hakkında yazılan yazıların analizi gerçekleştiğinde, olumlu yorumların çok olması durumunda ikinci sezonunun da çıkarılmasına karar verilebilir. Ama olumsuz yorumlar çok ise dizi çekimine devam edilmeyip boş yatırım yapmaktan da kaçınabilinir. Böylece, şirketler hem maliyet hem zaman açısından büyük bir tasarruf sağlayabilirler. Devletler açısından ise örneğin yatırım yapılacak bir proje hakkından önceden vatandaşlara proje hakkında bilgiler verilir. Daha sonra insanların düşüncelerinin analiz edilmesiyle o proje hakkında insanların ne düşündüğü, bu konuya nasıl yaklaştıkları tespit edilebilir. Böylece devletler bu projeyi yapmaktan vazgeçebilir ya da daha farklı bir şekilde düzenleyebilir. Böylece toplumun huzuru ve refahı daha kuvvetli sağlanmış olur.



Şekil 1. Duygu Analizi Yöntemleri (Sarıman ve Mutaf, 2020)

Duygu analizi yapabilmek için öncelikle bir veri kümesine ihtiyaç vardır. Bu veri kümeleri etiketlenmiş verilerden oluşmalıdır. Yani bir yazının olumlu, olumsuz ya da nötr duygu belirttiği belirtilmelidir. Bu aşamadan sonra veriler gerekli teknik ve algoritmalarla ön işleme sürecine sokulur. Bu şekilde verilerde genel olarak yazım hatası, noktalama işaretleri, gereksiz

kelimeler gibi sorunlar tespit edilir. Ham veriler bu hatalardan arındırılmak için ön işleme sürecine girer. Veri temizleme işlemi bittikten sonra sınıflandırma işlemi geçrekleştirilir. Bu işlemler için daha çok makine öğrenmesi algoritmaları kullanılır. Sınıflandırma işlemi ile eğitilen veriler test verileri ile kıyaslanır ve sistemin ne kadar başarılı olduğu görülür. Duygu analizi için genel kapsamda iki yaklaşım kullanılır. Bunlar Şekil 1'de gösterildiği üzere sözlük tabanlı yaklaşım ve makine öğrenmesine dayalı yaklaşımdır (İlhan ve Sağaltıcı, 2020).

2.3. Derin Öğrenme

Derin öğrenme, bilgisayarların deneyimden ders almasını ve dünyayı kavramların hiyerarşisi açısından anlamasını sağlayan bir makine öğrenmesi yöntemidir (Gündüz ve Cedimoğlu, 2019). Geleneksel makine öğrenmesi algoritmaları doğrusal yapıdadır fakat derin öğrenme algoritmaları yapılacak modelinde pronlemin karmaşıklığına göre değişen bir hiyerarşi modeli bulunmaktadır. Derin öğrenme süreci başarı sonucu belirli bir seviyeye gelene kadar devam eder. Bu süreçte verilerin geçmesi gereken genel adımlar Şekil 2'de gösterilmektedir (Kayaalp ve Süzen).



Şekil 2. Derin Öğrenme Süreçleri (Kayaalp ve Süzen)

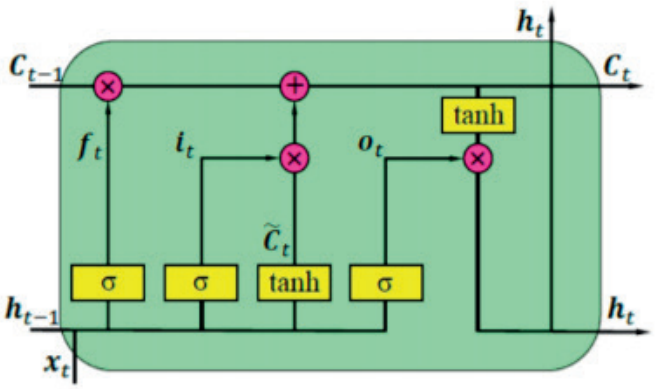
Literatürde derin öğrenme algoritmalarının uygulamaları ile ilgili çok sayıda çalışma yapılmıştır. Nesne tanıma, görüntü ve ses işleme, dil işleme, hastalık tespiti, biyomedikal sinyal ve görüntü işleme, robotik, kimya, reklam, finans gibi birbirinden farklı konularda derin öğrenme uygulamaları geliştirilmektedir (Gündüz ve Cedimoğlu, 2019).

(Kayaalp ve Süzen) yapmış oldukları çalışmada belirttikleri gibi derin öğrenme algoritmaları yaygın olarak kullanılan evrişimsel sinir ağları, tekrarlayan sinir ağları (RNN – Recurrent Neural Network) ve uzun kısa süreli hafıza ağları olmak üzere üçe ayrılır. CNN nesne tanıma ve görüntü sınıflandırma gibi alanlarda kullanılmaktadır. RNN, ardışık bilgileri kullanan bir algoritmadır. Bu algoritma çevirilerde, altyazı oluşturmada, gürültüsüz veri elde edilmesinde, konuşma tanıma gibi alanlarda kullanılır. LSTM modeli, RNN'nin gelişmiş bir versiyonudur. Bu algoritma sessiz videolara ses ekleme, ilişkili metinlerde kelime üretme, düzensiz dillerde öğrenme gibi alanlarda kullanılır.

2.4. Uzun Kısa Süreli Hafıza Algoritması

Derin öğrenme algoritmaları günümüzde birçok sınıflandırma ve tahmin işlemleri için kullanılırlar. Örneğin, bir videoda bir sınıf içerisine sırayla çocuklar girip, sıralara oturup test çözüyor olsunlar. Daha sonra sınıfa girecek çocuğun sıraya oturup test çözeceğini RNN algoritması ile tahmin etmek oldukça kolaydır. Fakat videoda belli zaman aralıklarında başka olaylar gerçekleşir ve bu olaylardan sonra sınıfa bir çocuk girerse, RNN bu çocuğun ne yapacağını tahmin etmesi kolay olmayabilir. Bu tür sorunları çözmek için RNN'nin gelişmiş bir versiyonu olan Uzun Kısa Süreli Hafıza Ağları kullanılır (Kayaalp ve Süzen).

(Görgel ve Kavlak, 2020) yapmış oldukları çalışmada belirttikleri gibi, LSTM uzun vadede gerçekleşen işlemleri öğrenebilen bir algoritmadır. Bu algoritma sıralı verilerin modellenmesinde kullanılan RNN algoritmanın gelişmiş bir türüdür.



Şekil 3. LSTM Mimari Yapısı (Kara, 2019)

(Kara, 2019) yapmış olduğu çalışmada belirttiği gibi, LSTM birbirini takip eden sıralı yapılardan oluşur. Şekil 3'de görüldüğü gibi, LSTM algoritmasının temel olarak üç katmanı vardır. Bu katmanlar unut, girdi ve çıktı katmanlarıdır. Unut katmanı gelen bilginin unutulup unutulmayacağına karar verir. Girdi katmanı hangi bilginin bellekte depolanıp depolanmayacağına karar verir. Çıktı katmanı ise hangi bilginin çıktı olup olmayacağına karar verir.

LSTM algoritmasının ilk adımı girdi olarak X_t ve h_{t-1} girdilerini alarak nelerin silineceğine karar verir. Bu işlemler unut katmanında (f_t) Eşitlik (1) kullanılarak yapılır ve aktivasyon fonksiyonu olarak sigmoid fonksiyonu kullanılır.

$$f_t = (W_{f,} * X_t + W_{f,h} * h_{t-1} + b_f) \quad (1)$$

İkinci adımda yeni bilgilerin belirleneceği girdi katmanı devreye girer ve ilk olarak (i_t) Eşitlik (2) kullanılarak sigmoid fonksiyonu ile bilgiler güncellenir. Ardından Eşitlik (3) ile yeni bilgiyi oluşturacak aday bilgiler tanh fonksiyonu tarafından belirlenir.

$$i_t = (W_{i,} * X_t + W_{i,h} * h_{t-1} + b_i) \quad (2)$$

$$\tilde{C}_t = tanh(W_{c,x} * X_t + W_{c,h} * h_{t-1} + b_c) \quad (3)$$

Eşitlik (4) tarafından yeni bilgiler oluşturulur.

$$C_t = C_{t-1} * f_t + i_t * \tilde{C}_t \quad (4)$$

Son olarak çıktı katmanında Eşitlik (5) ve (6) kullanılarak çıktı değerleri elde edilir.

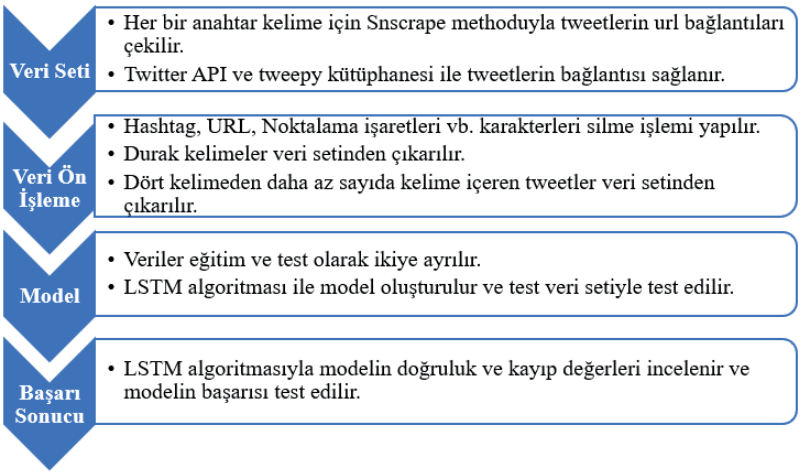
$$o_t = (W_{o,} * X_t + W_{o,h} * h_{t-1} + b_o) \quad (5)$$

$$h_t = ot * tanh(C_t) \quad (6)$$

Yukarıda ifade edilen süreç tekrarlanarak devam eder. Ağırlık parametreleri (W) ve bias parametreleri (b) gerçek eğitim değerleri ile LSTM çıktı değerleri arasındaki farkı minimize edecek şekilde model tarafından öğrenilmektedir.

3. ÖNERİLEN SİSTEM YAPISI

Covid-19 sürecinde Twitter sosyal medya aracından atılan Türkçe tweetlerden duygu analiz çalışmasının yapılacağı sistemin yapısı Şekil 4'te verilmiştir. Bu bölümde çalışmada kullanılacak veri kümesi, veri ön işleme ve sınıflandırma gibi duygu analizi adımları anlatılmaktadır.



Şekil 4. Önerilen Sistemin Uygulama Adımları

3.1. Veri Kümesi

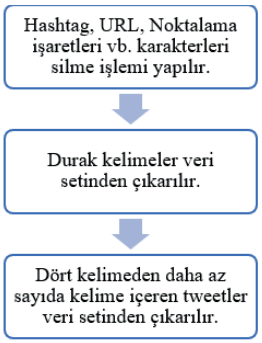
Bu çalışmada Covid-19 sürecinde Twitter üzerinden atılan Türkçe tweetler kullanılmıştır. Bu konuyla ilgili hazır bir veri kümesi olmadığı için Twitter API aracılığıyla, python dilinde yazılmış bir scrape methoduyla tweetler toplanmıştır. 11 Mart 2020 tarihinden normalleşme adımlarının atıldığı 1 Haziran 2020 tarihine kadar 5 başlık altında yaklaşık 485.000 adet Türkçe tweet toplanmıştır. Şekil 5'te veri kümesi içindeki bazı tweetler gösterilmiştir.

```
0    Hadi tüm Ülkemize  geçmiş olsun. Son sokağa çı...
1    2 günlük #SokagaCikmaYasagi sonrası, an itibar...
2    Arizona kertenkeleleri son ses müzik açıp gezd...
3    #SokagaCikmaYasagi ndaki max aktivitem. Son ya...
4    https://t.co/rIo5JUJQo3\nYUZUME İNSTAGRAM POST...
5    @ciicekci 😪😪😪😪\n\nÜzgünüm\n1 pazar arabam bile...
6    https://t.co/rIo5JUJQo3\nYUZUME İNSTAGRAM POST...
7    Son Dakika\n\n15 ili kapsayan sokağa çıkma kıs...
8    2 gün süren sokağa çıkma yasağı sona erdi http...
9    Gece 12 yi bekleyenler yine dökülmüş yollara. ...
```

Şekil 5. Çalışmada Kullanılan Veri Kümesinden Bir Örnek

3.2. Veri Ön İşleme

Çalışmada kullanılacak veri kümesi, gerekli sınıflandırma algoritmalarına girmeden önce daha başarılı bir model oluşturmak amacıyla veri ön işleme aşamalarına sokulur. Şekil 6'da veri ön işleme adımları gösterilmiştir.



Şekil 6. Veri Ön İşleme Adımları

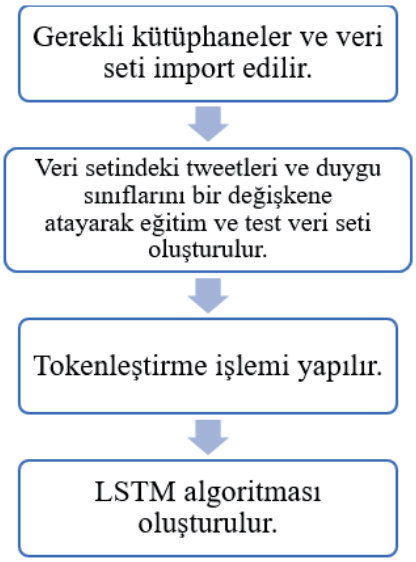
Veri temizleme aşamasında ilk olarak veri kümesinde bulunan http, simge, hashtag, noktalama işaretleri gibi gereksiz ifadeler temizlenmiştir. Ayrıca aynı tweetleri içeren satırlar silinmiştir. Temizlenen veri kümesinde bulunan kelimeler küçük harfe çevrilmiştir. Bu işlemlerden sonra Türkçede sık kıllanılan durak kelimeler veri kümesinden kaldırılmıştır. Son olarak da 4 kelimeden az olan tweetler metin analizinde anlamlı bir sonuç üretemeyeceği için veri kümesinden çıkarılmıştır. Şekil 7'de temizlenmiş veri kümesinden bir örnek gösterilmiştir.

```
40    hafta gün parkta sokakta kalsam anca kendime g...
41    doğum iznindeyken yakalanan hastaneye yatırıla...
42    tarih yaziyo bitmedi bitmiyo corona hirkcilik ...
43    sosyal medya bilgi kartı yakında youtube adres...
44    göğüs hastalıkları hastanesinde çalışıyordu ül...
45    koronavirüs george floyd ölümü arasında bağ ku...
46    koronavirüs yakalanan haftalık bebeği sezeryan...
47    sayın cumhurbaşkanı dileği yerine getirmenizi ...
48    kısıtlamanın bitmesiyle izmir manisa arası sey...
49    ilde seyahat kısıtlaması mayıs saat bitmiştir ...
50    ilde iki gündür devam eden sokağa çıkma yasağı...
51    sporcumuz doruk göktuğ canlandırdığı covid isi...
52    okullar açıldı tarihe notum aydır kapalı olan ...
53    covid salgınına karşı alınan önlemler kapsamın...
54    memleket meselesicanlı yayın haziran pazartesi...
```

Şekil 7. Temizlenmiş Veri Kümesinden Bir Örnek

3.3. Sınıflandırma

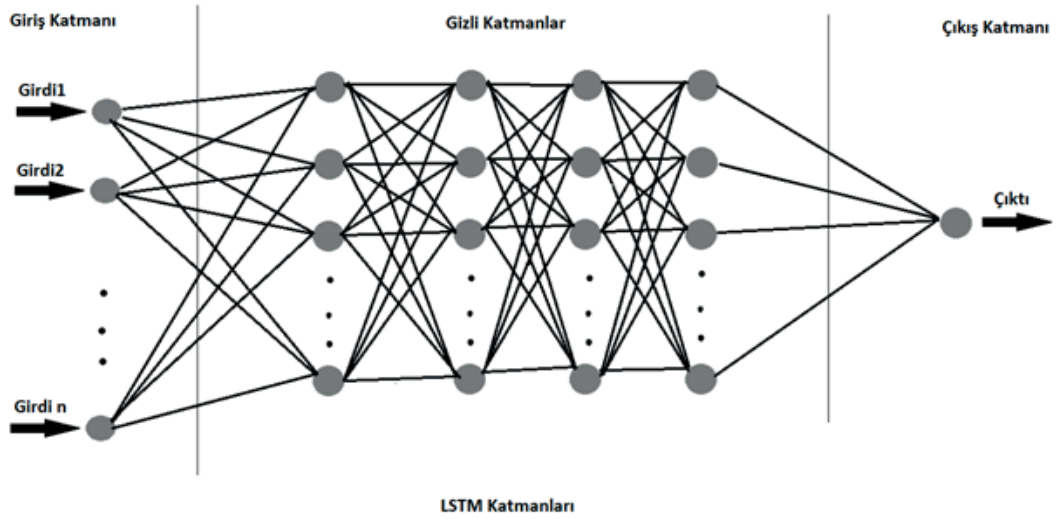
Sınıflandırma, kategorisi bilinmeyen verilere en uygun kategorinin atanmasıdır. Sınıflandırma işlemi için genellikle makine öğrenmesi algoritmaları kullanılır. Bu çalışmada makine öğrenmesinin alt sınıfı olan derin öğrenme algoritmalarından LSTM algoritması kullanılanılmıştır. Şekil 8'de LSTM model adımları gösterilmiştir.



Şekil 8. Uygulama Model Adımları

İlk olarak verideki etiketler ve tweetler birer değişkene atılır. Eğitim ve test veri kümesi oluşturulur. Daha sonra veri kümesinin içinde en çok kullanılan 10.000 kelimeye göre bir sözlük oluşturulur ve kelimelere sayısal değerler atanır. Veri kümesindeki her bir kelime oluşturulan sözlükteki sayısal karşılığı ile değiştirilir. Her bir tweetin içindeki kelime sayıları bulunur. Daha sonra her bir tweetin kelime sayılarının genel ortalamasına bakılır ve bir değer elde edilir. Bu değer verimizdeki aykırı uzunluğa sahip cümleleri ortalamaya indirgememizi sağlar. Bu işlemden sonra veri içindeki tüm

tweetler aynı uzunluğa dönüştürülür. Verimiz LSTM modelinde kullanılacak hale getirilmiştir. Şekil 9'da görüldüğü gibi model içerisinde bir giriş katmanı, 4 adet LSTM katmanı ve bir çıkış katmanı bulunmaktadır. LSTM katmanlarında sırasıyla 32, 16, 8 ve 4 nöron bulunmaktadır. Model katmanları oluşturulduktan sonra model eğitilmiştir.



Şekil 9. Çalışmada Kullanılan LSTM Modeli

4. BULGULAR

Tweetler veri ön işleme aşamalarından geçirilmeden önce 484.002, sonrasında 392.060 adete düşmüştür. Şekil 10'da anahtar kelimelerin veri ön işleme öncesi ve sonrasına ait sayıları verilmiştir.

Anahtar Kelime	Ön işlem öncesi	Ön işlem sonrası
Covid19	176.027	141.711
Eğitim	25.464	18.798
Ekonomi	19.915	13.939
Maske	23.554	20.091
Sokağa çıkma yasağı	484.002	197.521

Şekil 10. Anahtar Kelime Sayıları

Twitter API aracılığıyla çekilen tweetler sırasıyla veri ön işleme ve model oluşturma aşamalarından geçirilmiştir. Veri ön işleme aşamasından sonra verilerin duygu sınıfı her bir tweet için rastgele olarak belirlenmiştir. Fakat rastgele duygu sınıfı atanıp model eğitildiği zaman başarı oranının çok düşük olduğu görülmüştür. Bu yüzden her bir veri kümesinin içinde en çok kullanılan olumlu ve olumsuz kelimeler tespit edilmiştir. Bu kelimelerin geçtiği tweetlere duygu sınıfı atanmıştır. Belirlenen kelimelere göre duygu sınıfı belirlendikten sonra oluşan veri kümesi ve olumlu olumsuz tweet sayıları Şekil 11'de gösterilmiştir.

Veri seti	Toplam	Olumlu	Olumsuz
Covid19	37.454	11.281	26.173
Eğitim	2.932	2.101	831
Ekonomi	2.465	1.051	1.414
Maske	5.862	2.413	3.449
Sokağa çıkma yasağı	57.582	11.790	45.792

Şekil 11. Etiketlenmiş Veri Kümesi

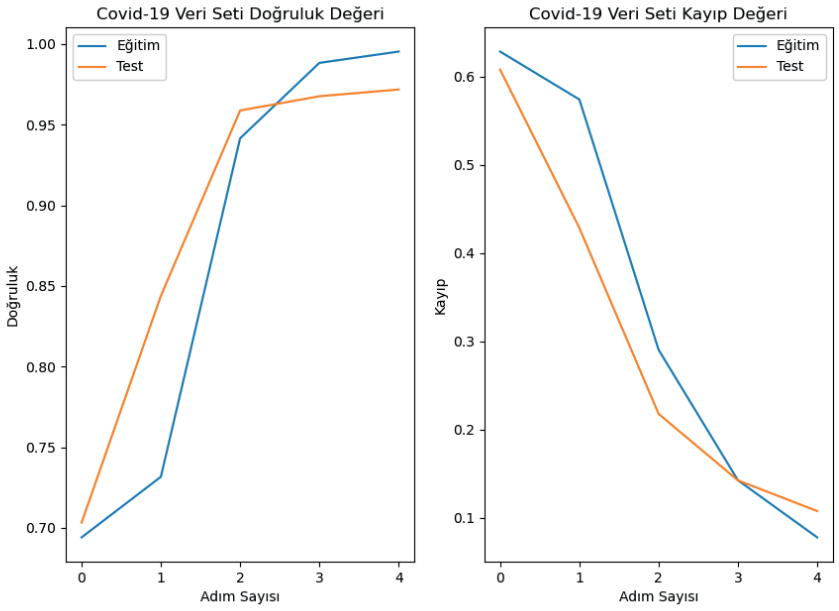
Son aşamada ise oluşturulan LSTM modelinin her bir veri kümesi için doğruluk değeri Şekil 12'de verilmiştir.

Veri Seti	Eğitim Veri Seti	Test Veri Seti	Doğruluk Değeri
Covid-19	29.963	7.491	0.9723
Eğitim	2.345	587	0.8961
Ekonomi	1.972	493	0.8519
Maske	4.689	1.173	0.8875
Sokağa Çıkma Yasağı	46.065	11.517	0.9774

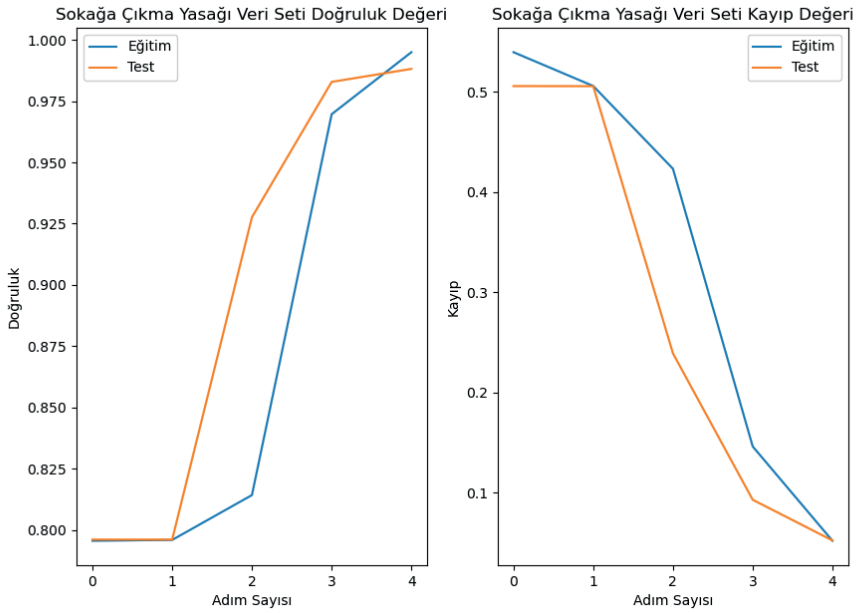
Şekil 12. Veri Kümelerinin Başarı Oranları

Şekil 12'deki sonuçlar incelendiği zaman covid-19 ve sokağa çıkma yasağı ile ilgili veri kümelerinde %97 oranında başarı sağlandığı, diğer veri kümelerinde ise %85 - %90 arasında başarı sağlandığı görülmüştür. Böyle bir durum ortaya çıkmasının en büyük sebebi covid19 ve sokağa çıkma yasağındaki veri kümelerinin boyutlarının diğer veri kümelerine göre çok daha fazla olmasıdır. Veri kümesi boyutunun artması model başarımını da arttırmaktadır.

Çalışmada önerilen modelin doğruluk değerleri incelendiği zaman genel olarak başarılı sonuçlar elde edildiği söylenebilir. Bu tip çalışmalarda ele alınan modelin eğitim için kullanılan veri seti üzerinde gereğinden fazla çalışıp ezber yapmaya başlaması söz konusu olabilir. Dolayısıyla, model eğitim seti için yüksek başarımlar elde ederken test veri seti için düşük başarımlar gösterebilir. Bu durum modelin aşırı öğrenmesine neden olmaktadır. Çalışmada önerilen modelde aşırı öğrenme olup olmadığını analiz etmek için modelin her bir veri kümesinde elde ettiği doğruluk ve kayıp değerlerinin sonucu aşağıdaki grafiklerde incelenmiştir.

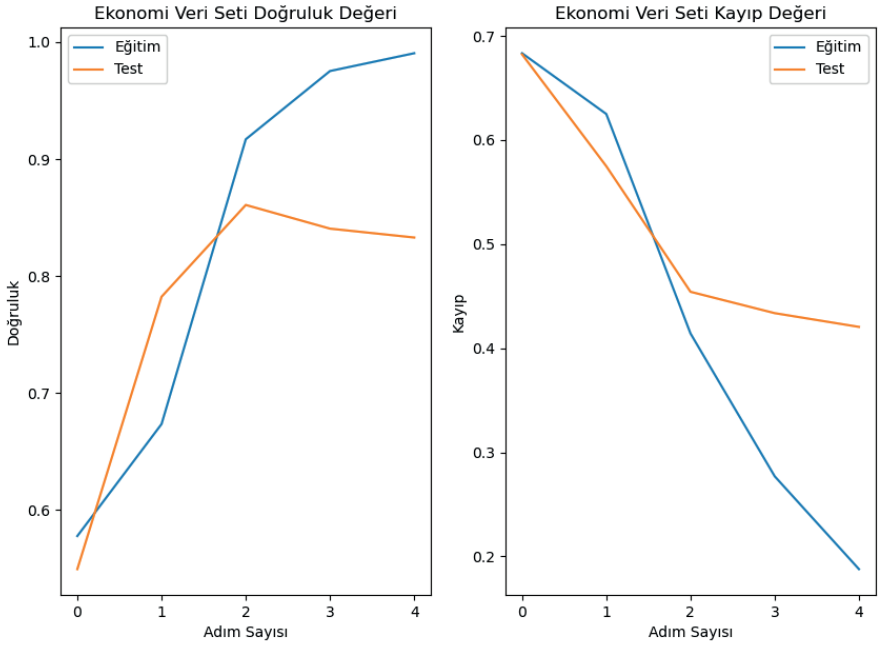


Şekil 13. Önerilen LSTM modelinde Covid19 veri kümesinin doğruluk ve kayıp değer grafiği

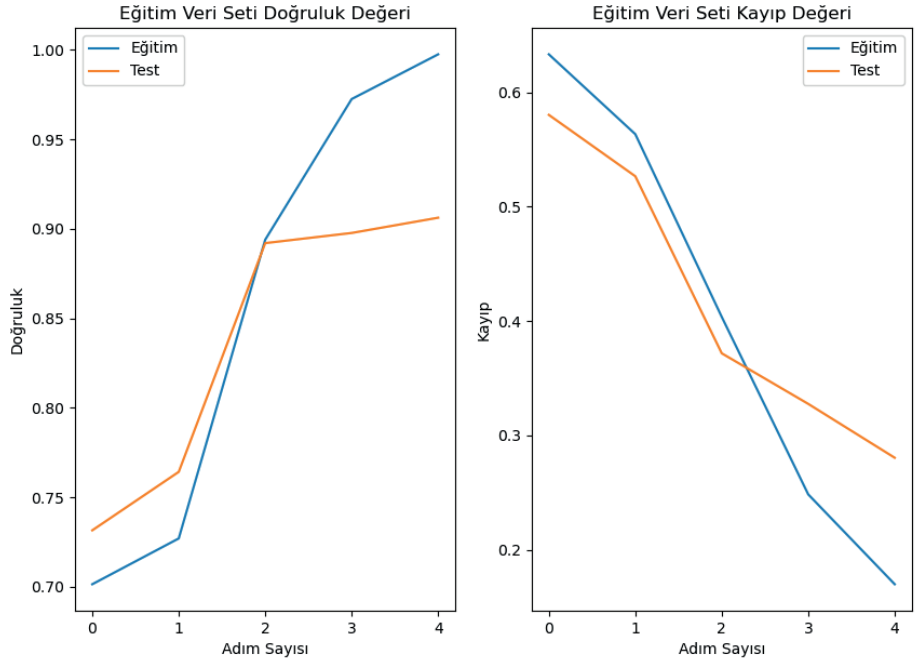


Şekil 14. Önerilen LSTM modelinde Sokağa Çıkma Yasağı veri kümesinin doğruluk ve kayıp değer grafiği

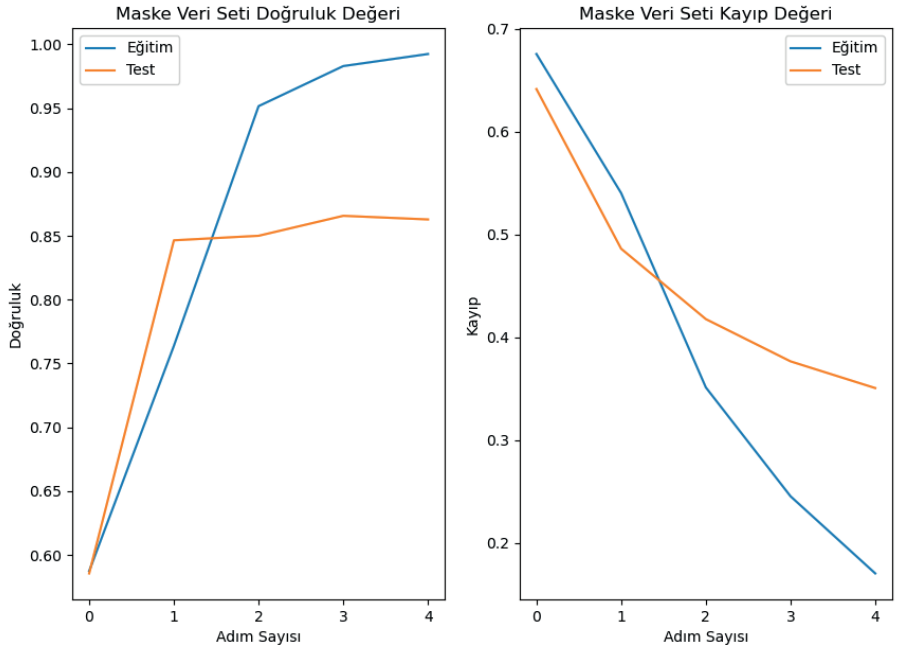
Şekil 13 ve 14 incelendiği zaman eğitim ve test kümelerinin aynı derecede doğruluk değerlerinin arttığı ve aynı derecede kayıp değerlerinin azaldığı görülür. Bu durumda modelde aşırı öğrenmenin olmadığı söylenebilir.



Şekil 15. Önerilen LSTM modelinde Ekonomi veri kümesinin doğruluk ve kayıp değer grafiği



Şekil 16. Önerilen LSTM modelinde Eğitim veri kümesinin doğruluk ve kayıp değer grafiği



Şekil 17. Önerilen LSTM modelinde Maske veri kümesinin doğruluk ve kayıp değer grafiği

Şekil 15, 16 ve 17'deki veri kümelerinin doğruluk ve kayıp değerlerinin her bir adım sayısındaki değerleri grafik üzerinden incelendiği zaman Şekil 13 ve 14'deki değerlerden daha farklı bir görüntüye sahip olduğu görülür. Birinci adımda eğitim ve test verilerin doğruluk değerleri beraber artıp kayıp değerleri beraber düştükten sonra test verilerinin bir süre sonra eğitim

verileri ile aynı yönde ilerlemediği görülür. Yani belirli bir adım değerinden sonra test veri kümesindeki doğruluk ve kayıp değerleri sabit kalır ve eğitim veri kümesi ile arasındaki değer farkı artar. Bu durumda bu veri kümelerinden oluşan modellerimizde aşırı öğrenme durumu olduğu söylenebilir. Bunun birçok nedeni olabilir. Veri kümelerinin boyutlarının az olması bunun en büyük sebepleri arasında gösterilebilir. Aşırı öğrenmeyi düzeltmek için daha fazla sayıda veri toplanabilir, hiperparametre değerleri değiştirilebilir.

5. SONUÇ

Bu çalışmada Covid-19 sürecinde Twitter sosyal medya aracından atılan Türkçe tweetlerden duygu analizi çalışması yapılmıştır. Bu analizler ile elde edilen sonuçlar farklı paydaşlar için önem arz edebilir. Örneğin, hükümetler insanların yeni virüs türlerine nasıl tepki verdiklerini, gıda kıtlığı, panik atak vb. gibi çeşitli zorlukların neler olduğunu bu analizler ile bileceği için bu bilgileri çeşitli alanlarda politikalarını belirlemek için kullanabilir. Firmalar, maske ya da gıda kıtlığı üzerine atılan tweetlerden yapılan duygu analizi ile bu temel öğelerin üretimine başlayabilir ya da mevcut üretimlerini arttırabilir. Çeşitli sosyal toplum kuruluşları, duygu analizi sonuçlarını kullanarak insanları nasıl rehabilite edeceklerine dair stratejilerine karar verebilir. Bu amaçla, Twitter'dan çekilen tweetler gerekli veri temizleme adımlarından geçirilip veri kümesinde en çok kullanılan kelimelere göre bir duygu sözlüğü oluşturulmuştur. Bu sözlükle beraber veri kümesindeki her bir satırın duygu sınıfı belirlenmiştir. Bu işlemlerden sonra LSTM modeli oluşturulmuştur. 5 farklı veri kümesi için model başarı çıktıları incelenmiş ve doğruluk oranları belirlenmiştir.

Sonuçlar incelendiği zaman oluşturduğumuz modelin genel olarak başarı yüzdesinin yüksek olduğu görülmüştür. covid19 ve sokağa çıkma yasağı veri kümeleri diğer üç veri kümesine göre daha iyi bir başarı göstermiştir. Bunun en büyük sebebi bu iki veri kümesinin boyutlarının diğer veri kümelerinin boyutlarına göre daha fazla olmasıdır. Çalışmanın mevcut durumunda yeterince büyük bir duygu sözlüğü kullanılmamıştır. Ayrıca, duygu sözlüğündeki kelimelerden birine sahip olmayan tweetler veri kümesinden çıkarılmıştır. Bu durum veri kaybına sebep olmaktadır. İleriki çalışmalarda, duygu sınıfı etiketleme işlemlerinin daha büyük bir duygu sözlüğü ve daha fazla veri ile gerçekleştirilmesi ile modelin daha yüksek başarı sağlayabilmesi amaçlanmaktadır.

Hakem Değerlendirmesi: Dış bağımsız.
Yazar Katkıları: Çalışma Konsepti/Tasarım-M.C.Y., Z.O.; Veri Toplama-M.C.Y.; Veri Analizi/Yorumlama- M.C.Y., Z.O.; Yazı Taslağı- M.C.Y., Z.O.; İçeriğin Eleştirel İncelemesi-M.C.Y., Z.O.; Son Onay ve Sorumluluk- M.C.Y., Z.O.
Çıkar Çatışması: Yazarlar çıkar çatışması bildirmemiştir.
Finansal Destek: Yazarlar bu çalışma için finansal destek almadığını beyan etmiştir.

Peer-review: Externally peer-reviewed.
Author Contributions: Conception/Design of Study- M.C.Y., Z.O.; Data Acquisition-M.C.Y.; Data Analysis/Interpretation- M.C.Y., Z.O.; Drafting Manuscript- M.C.Y., Z.O.; Critical Revision of Manuscript- M.C.Y., Z.O.; Final Approval and Accountability- M.C.Y., Z.O.
Conflict of Interest: The authors have no conflict of interest to declare.
Grant Support: The authors declared that this study has received no financial support.

Kaynaklar/References

Albayrak, M., Topal, K., & Altıntaş, V. (2017). Sosyal medya üzerinde veri analizi: Twitter. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, *22*(Kayfor 15 Özel Sayısı), 1991-1998.

Aydın, B., & Doğan, M. (2020). Yeni Koronavirüs (Covid-19) pandemisinin turistik tüketici davranışları ve Türkiye turizmi üzerindeki etkilerinin değerlendirilmesi. *Pazarlama Teorisi ve Uygulamaları Dergisi*, *6*(1), 93-115.

Aytuğ, O. (2017). Twitter mesajlari üzerinde makine öğrenmesi yöntemlerine dayali duygu analizi. *Yönetim Bilişim Sistemleri Dergisi*, *3*(2), 1-14.

Ayvaz, S., Yıldırım, S., & Salman, Y. B. (2019). Türkçe duygu kütüphanesi geliştirme: Sosyal medya verileriyle duygu analizi çalışması. *Avrupa Bilim ve Teknoloji Dergisi*, (16), 51-60.

Chintalapudi, N., Battineni, G., & Amenta, F. (2021). Sentimental analysis of Covid-19 tweets using deep learning models. *Infectious Disease Reports*, *13*(2), 329-339.

Görgel, P., & Kavlak, E. (2020). Uzun kısa süreli hafıza ve evrişimsel sinir ağları ile rüzgar enerjisi üretim tahmini. *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi*, *11*(1), 69-80.

Gündüz, G., & Cedimoğlu, İ. H. (2019). Derin öğrenme algoritmalarını kullanarak görüntüden cinsiyet tahmini. *Sakarya University Journal of Computer and Information Sciences, 2*(1), 9-17.

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

Kara, A. (2019). Uzun-kısa süreli bellek ağı kullanarak global güneş ışınımı zaman serileri tahmini. *Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji, 7*(4), 882-892.

Kayaalp, K., & Süzen, A. A. (2018). Derin öğrenme ve Türkiye'deki uygulamaları. *Iksad International Publishing House*, 6-21.

Kaynar, O., Görmez, Y., Yıldız, M., & Albayrak, A. (2016). Makine öğrenmesi yöntemleri ile duygu analizi. In *International Artificial Intelligence and Data Processing Symposium (IDAP'16)*, 17-18.

Manguri, K. H., Ramadhan, R. N., & Amin, P. R. M. (2020). Twitter sentiment analysis on worldwide Covid-19 outbreaks. *Kurdistan Journal of Applied Research, Special Issue on Coronavirus (COVID-19),*54-65.

Salur, M. U., & Aydin, I. (2020). A novel hybrid deep learning model for sentiment classification. *IEEE Access, 8*, 58080-58093.

Sariman, G., & Mutaf, E. Covid-19 sürecinde Twitter mesajlarinin duygu analizi. *Euroasia Journal of Mathematics Engineering Natural and Medical Sciences,*