Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 26, S. 138-143, Temmuz 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 26, pp. 138-143, July 2021

Research Article

COVID-19 ile İlgili Sosyal Medya Gönderilerinin Metin Madenciliği Yöntemlerine Dayalı Olarak Zaman-Mekansal Analizi

Aytuğ Onan[1*]

[1] İzmir Kâtip Çelebi Üniversitesi, Mühendislik-Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, İzmir, Türkiye (ORCID: 0000-0002-9434-5880)

(3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications June 11-13, 2021)

(**DOI:** 10.31590/ejosat.957020)

ATIF/REFERENCE: Onan, A. (2021). COVID-19 ile ilgili sosyal medya gönderilerinin metin madenciliği yöntemlerine dayalı olarak zaman-mekansal analizi. *Avrupa Bilim ve Teknoloji Dergisi*, (26), 138-143.

Öz

COVID-19, hastalığın ilk bildirildiği dönemden bu yana, şiddetli akut solunum sendromu büyük salgınlara neden olmaktadır ve dünya çapında bir pandemiye dönüşmüştür. Dünyanın birçok ülkesinde, COVID-19 salgınının zaman-mekansal analizine yönelik olarak önemli sayıda gerçek zamanlı, etkileşimli mobil ya da çevrimiçi coğrafi bilgi sistemleri, web siteleri ve uygulamalar geliştirilmiştir. Bilgi ve iletişim teknolojilerindeki ilerlemeler ile pek çok farklı kaynaktan COVID-19 salgınına yönelik olarak elde edilen veriler, salgın durumuna ilişkin bilgilerin etkin ve zamanında elde edilebilmesi için büyük önem taşımaktadır. Internetteki medya ve iletişim platformlarında paylaşılan haber makaleleri, bulaşıcı hastalık salgınlarının izlenmesi ve takip edilmesi için önemli bir veri kaynağı niteliğindedir. Bu çalışmada, İngiltere ve İspanya'da COVID-19 sürecine ilişkin 2020 yılının mart, mayıs ve temmuz aylarında yayınlanan 299'ar tane haber makalesi toplanarak oluşturulan derlem kullanılmaktadır. Metin belgelerinin temsilinde, üç temel n-gram modeli olan (1-gram, 2-gram ve 3-gram) temsilleri, tümce ögeleri 2-gram ve tümce ögeleri 3-gram öznitelikleri, kelime/tümce ögesi çiftleri, karakter n-gram (n=2) ve karakter n-gram (n=3) öznitelikleri ve bu özniteliklerin biraraya getirilmesi ile elde edilen topluluk öznitelik kümelerinin etkinlikleri değerlendirilmektedir. Öznitelik kümelerinin başarımlarının değerlendirilmesinde, altı temel makine öğrenmesi sınıflandırıcısı olan Naive Bayes algoritması, lojistik regresyon algoritması, destek vektör makineleri, C4.5 karar ağacı, k-en yakın komşu algoritması ve rastgele orman algoritması kullanılmaktadır. Deneysel analizlerde kullanılan on yedi farklı metin temsil yöntemi arasında en yüksek başarımın, sözcük tabanlı 1-gram özniteliklerin karakter tabanlı 3-gram modeli ile kullanıldığında elde edildiği görülmektedir. Deneysel analizlerde kullanılan temel sınıflandırma algoritmaları arasında en yüksek başarım rastgele orman algoritmasıyla, ikinci en yüksek başarım ise lojistik regresyon algoritmasıyla alınmaktadır. Deneysel analizler, makine öğrenmesi ve metin madenciliği tekniklerinin, salgın hastalıklara ilişkin sosyal medya gönderilerinin zaman/mekânsal analizi için uygun teknikler olduğunu göstermektedir.

Anahtar Kelimeler: Metin madenciliği, Makine Öğrenmesi, Veri Bilimi.

Spatio-Temporal Analysis of Social Media Posts Related to COVID-19 Based on Text Mining Methods

Abstract

Since COVID-19 was first reported, severe acute respiratory syndrome has been causing massive outbreaks and has turned into a worldwide pandemic. In many countries of the world, a significant number of real-time, interactive mobile or online geographic information systems, websites and applications have been developed for the time-spatial analysis of the COVID-19 outbreak. The advances in information and communication technologies and the data obtained from many different sources regarding the COVID-19 outbreak are of great importance in order to obtain effective and timely information on the epidemic situation. News articles shared on media and communication platforms on the Internet are an important source of data for monitoring and tracking infectious disease outbreaks. In this study, 299 news articles published in March, May and July 2020 on the COVID-19 process in England and Spain are used. In the representation of text documents, the three basic n-gram models (1-gram, 2-gram, and 3-gram), part-of-speech 2-gram

[*] Sorumlu Yazar: İzmir Kâtip Çelebi Üniversitesi, Mühendislik-Mimarlık Fakültesi Fakültesi, Bilgisayar Mühendisliği Bölümü, İzmir, Türkiye, ORCID: 0000-0002-9434-5880, aytug.onan@ikcu.edu.tr

and part-of-speech 3-gram features, word / part-of-speech pairs, character n-gram (for, n = 2) and character n-gram (for, n = 3) features and the efficiency of the ensemble feature sets obtained by combining these features are evaluated. Naive Bayes algorithm, logistic regression algorithm, support vector machines, C4.5 decision tree, k-nearest neighbor algorithm and random forest algorithm are used to evaluate the performance of feature sets. Among the seventeen different text representation methods used in experimental analysis, it is seen that the highest performance is achieved when word-based unigram features are used with a character-based 3-gram model. Among the basic classification algorithms used in experimental analysis, the highest performance is obtained with the random forest algorithm, and the second highest performance is obtained with the logistic regression algorithm. Experimental analysis shows that machine learning and text mining techniques are suitable techniques for the spatio-temporal analysis of social media posts regarding epidemics.

Keywords: Text mining, Machine learning, Data Science.

1. Giriş

COVID-19, hastalığın ilk bildirildiği dönemden bu yana, şiddetli akut solunum sendromu büyük salgınlara neden olmaktadır ve dünya çapında bir pandemiye dönüşmüştür. COVID-19 salgını, dünya çapında birçok insanın yaşamını kaybetmesine neden olmuş, aralarında sağlık, eğitim, gıda ve iş organizasyonlarının da yer aldığı birçok alanda küresel ölçekte önemli değişikliklere yol açmıştır [1]. COVID-19 salgını süreci, hükümet ve araştırmacıların, akademik kurumların ve endüstri kuruluşlarının, salgını önlemeye yönelik olarak ortak hedefler etrafında birleşmesine neden olmuştur. Bu, sağlık kaynakları yönetimi, sosyal politika belirleme, salgın önleme ve tedavisi ile aşı geliştirmeyle ilgili süreçlere ilişkin birçok çıktı elde edilmesini olanaklı hale getirmiştir [2]. Buna paralel olarak, Internetteki medya ve iletişim platformlarında, COVID-19 salgını sürecinde Dünyanın farklı ülkelerinde uygulanan sosyal politikalara, salgın önleme ve tedavi uygulamalarına ve aşı geliştirme süreçlerine yönelik olarak paylaşılan birçok sosyal medya gönderisi ve haber makaleleri bulunmaktadır. Internetteki gayri resmi paylaşım platformlarının, bulaşıcı hastalık ve salgınlara yönelik gönderilerin önemli bir bölümünü oluşturduğu ve Dünyadaki ilk ve zamanlı haberlerin bu tarz platformlardan elde edildiği görülmektedir. Dünya Sağlık Örgütünün (WHO) incelemeye aldığı tüm önemli salgınların ilk olarak Internetteki gayri resmi platformlarda paylaşıldığı görülmektedir [3]. Internetteki medya ve iletişim platformlarında paylaşılan haber makaleleri, bulaşıcı hastalık salgınlarının izlenmesi ve takip edilmesi için önemli bir veri kaynağı niteliğindedir. Dünyanın birçok ülkesinde, COVID-19 salgınının zaman-mekansal analizine yönelik olarak önemli sayıda gerçek zamanlı, etkileşimli mobil ya da çevrimiçi coğrafi bilgi sistemleri, web siteleri ve uygulamalar geliştirilmiştir. Bilgi ve iletişim teknolojilerindeki ilerlemeler ile pek çok farklı kaynaktan COVID-19 salgınına yönelik olarak elde edilen veriler, salgın durumuna ilişkin bilgilerin etkin ve zamanında elde edilebilmesi için büyük önem taşımaktadır.

Bu çalışmada, İngiltere ve İspanya'da COVID-19 sürecine ilişkin 2020 yılının mart, mayıs ve temmuz aylarında yayınlanan 299'ar tane haber makalesi toplanarak oluşturulan derlem kullanılmaktadır. Metin belgelerinin temsilinde, üç temel n-gram modeli olan (1-gram, 2-gram ve 3-gram) temsilleri, tümce ögeleri 2-gram ve tümce ögeleri 3-gram öznitelikleri, kelime/tümce ögesi çiftleri, karakter *n*-gram (*n*=2) ve karakter *n*-gram (*n*=3) öznitelikleri ve bu özniteliklerin biraraya getirilmesi ile elde edilen topluluk öznitelik kümelerinin etkinlikleri değerlendirilmektedir. Öznitelik kümelerinin başarımlarının değerlendirilmesinde, altı temel makine öğrenmesi sınıflandırıcısı olan Naive Bayes algoritması, lojistik regresyon algoritması, destek vektör makineleri, C4.5 karar ağacı, k-en yakın komşu algoritması ve rastgele orman algoritması kullanılmaktadır. Deneysel analizlerde kullanılan on yedi farklı metin temsil yöntemi arasında en yüksek başarımın, sözcük tabanlı 1-gram özniteliklerin karakter tabanlı 3-gram modeli ile kullanıldığında elde edildiği görülmektedir. Deneysel analizlerde kullanılan temel sınıflandırma algoritmaları arasında en yüksek başarım rastgele orman algoritmasıyla, ikinci en yüksek başarım ise lojistik regresyon algoritmasıyla alınmaktadır. Deneysel analizler, makine öğrenmesi ve metin madenciliği tekniklerinin, salgın hastalıklara ilişkin sosyal medya gönderilerinin zaman/mekânsal analizi için uygun teknikler olduğunu göstermektedir. Çalışmanın geri kalanı şu şekilde yapılandırılmıştır: İkinci bölümde, ilgili çalışmalar tanıtılmaktadır. Çalışmanın üçüncü bölümünde, çalışmanın metodolojisi, dördüncü bölümde deneysel sonuçlar ve tartışma, son bölümde ise çalışmanın genel sonuçlarına değinilmektedir.

2. İlgili Çalışmalar

Metin sınıflandırma, metin madenciliğinin, metin belgelerini daha önceden belirlenmiş bir veya daha fazla sınıf etiketine atayan önemli bir uygulama alanıdır [4]. Metin sınıflandırma, web sayfası sınıflandırma [5], duygu analizi [6-11], istenmeyen e-postaların filtrelenmesi [12] ve metin türü belirleme [13] gibi birçok alanda başarıyla uygulanmaktadır. COVID-19 gönderilerinin metin madenciliğiyle analizine yönelik olarak gerçekleştirilmiş birçok bilimsel çalışma bulunmakla birlikte, bu bölümün geri kalanında ilgili alandaki temel çalışmalara değinilmektedir.

Jahanbin ve Rahmanian [14] çalışmalarında, COVID-19 ile ilgili süreci kontrol etmek amacıyla, sosyal medya platformlarında paylaşılan salgınla ilgili haberleri ve sosyal ağ paylaşımlarını izleme ve takip etmeye yönelik olarak, bulanık c-ortalamalar kümeleme algoritmasına dayalı bir yöntem önerisi gerçekleştirmiştir. Ordun vd. [15] tarafından gerçekleştirilen çalışmada, COVID-19 ile ilgili Twitter üzerinde yapılan gönderilerin konu, anahtar terim ve özelliklere dayalı analizi için gizli Dirichlet tahsisi konu modelleme algoritması kullanılmaktadır. Bir başka çalışmada, Peng vd. [16] çalışmalarında, Çin'de yaygın kullanıma sahip olan bir sosyal medya platformu olan Sina Weibo üzerinde COVID-19 pnömoni vakalarının yardım alması amacıyla oluşturulan gönderileri zaman-mekansal olarak analiz etmektedir. Şubat 2020'de on günlük süreçte coğrafi etiketleme kullanılarak analiz edilen veriler ile COVID-19 aktarımının kentsel ve mekânsal özelliklerinin modellenmesi amaçlanmıştır. Li vd. [17] çalışmalarında, Amerika Birleşik Devletleri'nde zaman-mekansal ölçekte COVID-19 ile ilgili stres belirtilerini tespit etmek için korelasyon açıklaması öğrenme algoritmasına ve klinik tabanlı bir sözlüğe dayalı bir algoritma sunmaktadır. Geliştirilen yöntem, gizli Dirichlet tahsisine kıyasla insan müdahalesini daha aza indirgeyerek, başarımı artırmayı amaçlamaktadır. Benzer şekilde, Chen vd. [18] tarafından COVID-19 ile ilgili Twitter mesajlarını zaman/mekânsal analiz etmeye yönelik olarak makine öğrenmesi ve konu modelleme

tabanlı bir yöntem geliştirilmiştir. Gerçekleştirilen çalışma, sosyal medya kullanıcılarının COVID-19 süreciyle ilgili zaman içerisinde nasıl farklı tepkiler verdiklerini anlamayı amaçlamaktadır. Deneysel analizler, analizlerin gerçekleştirildiği ülkelerde Twitter gönderi sayısı ile COVID-19 vakaları arasında korelasyon olduğunu göstermektedir. Bir diğer çalışmada, Boon-Itt ve Skunkan [19] çalışmalarında, Twitter kullanıcılarının COVID-19 salgınına yönelik gönderilerini analiz etmeye yönelik olarak duygu analizi ve gizli Dirichlet tahsisi algoritmasına dayalı konu modelleme yöntemlerine başvurmuştur.

3. Metodoloji

Bu bölümde, metin belgelerinin temsil edilmesinde kullanılan temel metin temsil yöntemleri ve deneysel analizlerde kullanılan temel makine öğrenmesi algoritmaları tanıtılmaktadır.

3.1. Metin Belgelerinin Temsili

Metin belgelerinin temsilinde, üç temel n-gram modeli olan (1-gram, 2-gram ve 3-gram) temsilleri, tümce ögeleri 2-gram ve tümce ögeleri 3-gram öznitelikleri, kelime/tümce ögesi çiftleri, karakter n-gram (n=2) ve karakter n-gram (n=3) öznitelikleri ve bu özniteliklerin biraraya getirilmesi ile elde edilen topluluk öznitelik kümelerinin etkinlikleri değerlendirilmektedir.

3.1.1. N-gram metin temsili

Metin belgelerinin temsil edilmesinde en yaygın kullanıma sahip olan yöntemlerin başında *n*-gram metin temsili yöntemi gelmektedir. Bir *n*-gram, herhangi bir metin belgesinin *n* öğeden oluşan bitişik bir dizidir. Kelime tabanlı ve karakter tabanlı *n*-gramlar, metin madenciliği ve doğal dil işlemede yaygın kullanıma sahiptir [20]. Unigram (1-gram) öznitelik temsili, belirli bir sözcüğün metin belgesi içerisinde var olması ya da olmaması durumunu modeller. Bigram (2-gram) öznitelik temsili, birbiri ardına gelen sözcüklerin varlığını modellemek için kullanılır. Benzer şekilde, trigram (3-gram) öznitelik temsili, birbiri arında gelen üç ardışık kelimenin ilgili metin parçası içerisinde varlığını modellemek için kullanılmaktadır. Kelime tabanlı *n*-gram modellerinin yanı sıra, karakter tabanlı *n*-gram modellerinde karakter n-gram (n=2), birbiri ardına gelen iki karakterin, karakter n-gram (n=3), ardışık üç karakterin ilgili metin parçası içerisinde varlığını modellemek için kullanılır. Metin belgesinden çıkarılan kelime tabanlı bazı unigram öznitelikler, “announced”, “april”, “are”, bazı bigram öznitelikler “accross the”, “as being”, “cases have”, bazı trigram öznitelikler “has said that”, “hold daily press”, “in coronavirus battle” şeklindedir. Metin belgesinden çıkarılan karakter tabanlı n-gram (n=2) öznitelik örnekleri, “ra”, “sp”, “xt”, bazı karakter tabanlı n-gram (n=3) öznitelik örnekleri “bet”, “bla”, “boa” şeklindedir.

3.1.2. Tümce ögesi n-gram metin temsili

Metin madenciliği alanında yapılan çalışmalarda, tümce ögesi etiketleme (POS tagging), kelimenin tanımına ve bağlamına dayalı olarak, bir metindeki kelimelerin belirli bir tümce ögesine atanmasına yönelik bir doğal dil işleme uygulamasıdır. Etkin bir sınıflandırma modeli oluşturmak için, tümce ögesi etiketleme doğal dil işlemede yaygın kullanıma sahiptir. Bu çalışmada, tümce ögeleri 2-gram ve tümce ögeleri 3-gram öznitelikleri değerlendirmeye alınmıştır. Metin belgesinden çıkarılan, tümce ögeleri 2-gram özniteliklerinden bazıları “CC_NN”, “CC_NNS”, “CC_PRP”, tümce ögeleri 3-gram özniteliklerinden bazıları ise “CC_CD_IN”, “CC_CD_IN” ve “CC_JJ_NN” şeklindedir.

3.1.3. Kelime/tümce ögesi çiftleri

Bazı durumlarda kelimelerin anlamları, metin içerisinde kullanıldıkları tümce ögesi çeşidine dayalı olarak farklılık gösterebilmektedir. Bu durumu modellemek amacıyla, kelime/tümce ögesi çiftleri özniteliği, herbir farklı kelime ve sözcük türü eşleşmesi için ayrı bir öznitelik çıkararak metin belgesini temsil etmektedir. Metin belgesinden çıkarılan bazı kelime/tümce ögesi çiftleri özniteliği örnekleri, “buy / VB”, “called / VBD”, “cases / NNS” şeklindedir.

3.2. Sınıflandırma Algoritmaları

Deneysel analizlerde, altı temel makine öğrenmesi sınıflandırıcısı olan Naive Bayes algoritması, lojistik regresyon algoritması, destek vektör makineleri, C4.5 karar ağacı, k-en yakın komşu algoritması ve rastgele orman algoritması kullanılmaktadır.

3.2.1. Naive Bayes algoritması

Naive Bayes algoritması (NB), öğrenme problemini modellerken, problemdeki öznitelikler arasında bağımsızlık olduğu varsayımına dayanan, Bayes teoremine dayalı olasılık tabanlı bir basit öğrenme algoritmasıdır. NB algoritmasının, özniteliklerin sınıf etiketini belirlemede birbirlerinden bağımsız olduğuna dayalı varsayımı, algoritmayı az sayıda parametre gerektiren ve oldukça ölçeklenebilir bir yapıya koymaktadır. NB algoritması, istenmeyen e-postaların filtrelenmesi başta olmak üzere, birçok metin madenciliği probleminde başarıyla kullanılabilmekte, daha karmaşık sınıflandırma algoritmaları ile rekabet edebilir performanslar elde edebildiği görülmektedir [4, 5].

3.2.2. Lojistik regresyon algoritması

Lojistik regresyon (LR) algoritması, meydana gelen herhangi bir olayın olasılığını modellemek için yordayıcı değişkenler arasında doğrusal bir fonksiyon kullanan temel bir öğrenme algoritmasıdır [20]. Lojistik regresyon algoritmasında, sınıf etiketlerinin belirlenmesinde kullanılan olasılık değeri, doğrudan parametreler üzerinde doğrusal fonksiyona dayalı olarak hesaplanır. Doğal dil işleme ve metin madenciliği uygulamalarında, lojistik regresyon algoritmasının, NB algoritması ile ortak birçok tasarım avantajına sahip olduğu, ölçeklenebilir olduğu ve doğru sınıflandırma başarımı bakımından oldukça etkili sonuçlar verdiği görülmektedir.

3.2.3. Destek vektör makineleri algoritması

Destek vektör makineleri algoritması (SVM), sınıflandırma ve regresyon problemlerinde uygulanabilen temel öğrenme algoritmaları arasındadır. SVM hem doğrusal hem de doğrusal olmayan sınıflandırma problemlerinde kullanılabilmektedir. Burada, sınıflandırma süreci veri setinin yüksek boyutlu bir hiper düzlem oluşturacak şekilde bölümlenmesi ile oluşturulur [21].

3.2.4. C4.5 karar ağacı algoritması

Karar ağacı algoritmaları, sınıflandırma ve regresyon problemlerinde başarıyla uygulanabilen, parametrik olmayan öğreticili öğrenme algoritmalarıdır. Burada, temel amaç, veri setinde yer alan özniteliklerden çıkarılan temel karar kurallarını öğrenerek, bir hedef değişkenin değerini tahmin eden bir

öğrenme modeli elde edilmesidir. Karar ağacı algoritmaları, bir özniteliğin önemini bağlamına dayalı olarak değiştirerek öğrenme modelini kurgulamaya çalışır. Karar ağacı algoritmaları sonucu elde edilen öğrenme modelleri ve kurallar, sınıflandırma sürecine ilişkin kararların kolay biçimde anlaşılması ve yorumlanmasını olanaklı hale getirir. Algoritma sonucu elde edilen ağaç yapısı görselleştirilebilir niteliktedir. C4.5 karar ağacı algoritması, ID3'ün halefidir ve sürekli öznitelik değerini ayrı bir aralıklar kümesine bölen ayrı bir özniteliği dinamik olarak tanımlayarak özniteliklerin kategorik olması gerektiği kısıtlamasını kaldırmıştır. Budama, kuralın doğruluğu o olmadan iyileşirse bir kuralın ön koşulu kaldırılarak yapılır [22].

3.2.5. K-en yakın komşu algoritması

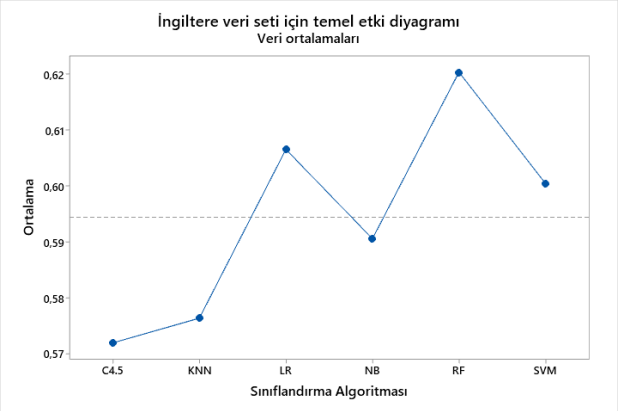
K-en yakın komşu algoritması (KNN), örnek tabanlı bir sınıflandırma algoritmasıdır. K-en yakın komşu algoritmasında, *k* komşu sayısı parametresine dayalı olarak, sınıf etiketi belirlenmek istenen örnek için öncelikle herhangi bir uzaklık/yakınlık ölçütüne dayalı olarak *k* tane en yakın komşu belirlenir. Ardından, sınıflandırılmak istenen örnek, komşularının sınıf etiketlerinin çoğunluk oylamasına dayalı olarak sınıf etiketine atanır.

3.2.6. Rastgele ortman algoritması

Rastgele orman algoritması (RF), veri kümesinin çeşitli alt örneklerine bir dizi karar ağacı sınıflandırma algoritmasını uygulayan ve tahmini doğruluğu ve aşırı uyumu kontrol etmek için ortalamayı kullanan bir meta tahminleyicidir. Alt örnek boyutu, önyükleme parametresi ile kontrol edilir [22].

4. Deneysel Süreç ve Sonuçlar

Deneysel analizlerde, İngiltere ve İspanya'da COVID-19 sürecine ilişkin 2020 yılının mart, mayıs ve temmuz aylarında yayınlanan 299'ar tane haber makalesi toplanarak oluşturulan derlem kullanılmaktadır [23]. İlgili veri seti, zaman tahmini için farklı dönemlerde elde edilen haber metni gönderilerine ilişkin anahtar sözcükleri öznitelik olarak içermektedir. Veri setinde tanımlı olan zaman periyotlarının, öğreticili öğrenme yöntemleri ile tahmin edilmesini gerektirmektedir. Deneysel analizlerde kullanılan temel öğrenme algoritmalarının ve metin temsil yöntemlerinin gerçekleştirimi Python dili kullanılarak scikit-learn kütüphanesi aracılığıyla yapılmıştır. Deneysel analizlerde, öğrenme algoritmaları, 10-kat çapraz geçerleme kullanılarak, doğru sınıflandırma oranına dayalı olarak değerlendirilmiştir. Tablo 1'de İngiltere veri seti üzerinde, Tablo 2'de ise İspanya veri seti üzerinde, on yedi metin temsil yöntemi ile altı farklı makine öğrenmesi aracılığıyla elde edilen sonuçlar sunulmaktadır. Deneysel analizlerde incelenen, kelime tabanlı n-gram modelleri olan, unigram, bigram ve trigram öznitelik temsil yöntemleri incelendiğinde, unigram metin temsili yönteminin diğer n-gram modellerine kıyasla daha yüksek başarım elde ettiği görülmektedir. Tümce ögesi bigram öznitelik temsili, tümce ögesi trigram öznitelik temsiline kıyasla daha yüksek başarım vermektedir. Karakter tabanlı n-gram modelleri arasında trigram modelin diğer birçok metin temsil yöntemine göre daha iyi performans elde ettiği görülmektedir. Deneysel analizlerde, temel metin tabanlı öznitelik temsil yöntemlerinin yanı sıra, bu özniteliklerin biraraya getirilmesi ile elde edilen topluluk öznitelik kümelerinin etkinlikleri değerlendirilmektedir. Deneysel sonuçlar, tüm konfigürasyonlar arasında en yüksek başarımın, kelime tabanlı unigram ile karakter tabanlı trigram özniteliklerin birlikte kullanılmasıyla elde edildiğini göstermektedir. Karşılaştırmalı analizlerde kullanılan sınıflandırma algoritmalarının başarımı Şekil 1'de özetlenmektedir. En yüksek performansın rastgele orman algoritmasıyla, ikinci en yüksek başarımın lojistik regresyon algoritmasıyla elde edildiği görülmektedir. Bunu, destek vektör makineleri algoritması takip etmektedir.



Şekil. 1. Karşılaştırılan yöntemlere ilişkin temel etki diyagramı

Tablo 1. İngiltere Veri Seti Üzerinde Temel Sınıflandırma Yöntemlerine İlişkin Deneysel Sonuçlar

Metin Temsil Yöntemi	NB	LR	SVM	C4.5	KNN	RF
Unigram	0,6388	0,6589	0,6622	0,5552	0,5819	0,6355
Bigram	0,5418	0,6154	0,6154	0,5619	0,5853	0,6488
Trigram	0,4916	0,5719	0,5652	0,5719	0,5351	0,5786
Tümce ögesi bigram	0,5619	0,5518	0,5284	0,5184	0,5786	0,5619
Tümce ögesi trigram	0,5452	0,4883	0,4916	0,4883	0,5518	0,5251
Kelime/tümce ögesi çiftleri	0,6187	0,6455	0,6388	0,5719	0,5786	0,6689
Unigram+Bigram+Trigram	0,5318	0,6321	0,6421	0,6087	0,5485	0,6288
Tümce ögesi bigram+Tümce ögesi trigram	0,5719	0,5151	0,5117	0,4669	0,5619	0,5652
Unigram+Bigram+Trigram+Tümce ögesi bigram+Tümce ögesi trigram+Kelime/tümce ögesi çiftleri	0,5686	0,6421	0,6288	0,5786	0,5485	0,6522
Karakter n-gram (n=2)	0,5987	0,5652	0,5552	0,5853	0,5619	0,5819
Karakter n-gram (n=3)	0,6722	0,6522	0,6522	0,6087	0,6321	0,6488

Unigram+karakter n-gram (n=3)	0,6756	0,6622	0,6555	0,6455	0,6020	0,6488
Bigram+karakter n-gram (n=3)	0,6321	0,6656	0,6455	0,5953	0,6187	0,6555
Trigram+karakter n-gram (n=3)	0,6120	0,6555	0,6622	0,6254	0,6187	0,6455
Unigram+karakter n-gram (n=2)	0,6522	0,6221	0,6221	0,5819	0,5652	0,6321
Bigram+karakter n-gram (n=2)	0,5886	0,5953	0,5920	0,5652	0,5686	0,6288
Trigram+karakter n-gram (n=2)	0,5385	0,5719	0,5385	0,5953	0,5619	0,6388

Tablo 2. İspanya Veri Seti Üzerinde Temel Sınıflandırma Yöntemlerine İlişkin Deneysel Sonuçlar

Metin Temsil Yöntemi	**NB**	**LR**	**SVM**	**C4.5**	**KNN**	**RF**
Unigram	0,6644	0,7013	0,6879	0,5570	0,5940	0,6644
Bigram	0,6141	0,6678	0,6342	0,5906	0,6242	0,6879
Trigram	0,5302	0,5878	0,5570	0,5772	0,5940	0,6275
Tümce ögesi bigram	0,5336	0,5671	0,5436	0,5805	0,6007	0,6174
Tümce ögesi trigram	0,5707	0,6007	0,6040	0,6074	0,5906	0,5940
Kelime/tümce ögesi çiftleri	0,6342	0,6812	0,6510	0,5638	0,5638	0,6980
Unigram+Bigram+Trigram	0,6141	0,6913	0,6812	0,6174	0,6007	0,6846
Tümce ögesi bigram+Tümce ögesi trigram	0,5839	0,5940	0,6007	0,6242	0,5805	0,6040
Unigram+Bigram+Trigram+Tümce ögesi bigram+Tümce ögesi trigram+Kelime/tümce ögesi çiftleri	0,6242	0,7013	0,6913	0,5872	0,6242	0,6242
Karakter n-gram (n=2)	0,6577	0,6309	0,6242	0,5839	0,6141	0,6409
Karakter n-gram (n=3)	0,6846	0,6745	0,6644	0,5906	0,6510	0,6208
Unigram+karakter n-gram (n=3)	0,6846	0,6745	0,6644	0,5906	0,6510	0,6208
Bigram+karakter n-gram (n=3)	0,6745	0,6745	0,6544	0,5738	0,6376	0,6946
Trigram+karakter n-gram (n=3)	0,6477	0,6812	0,6577	0,5872	0,6309	0,6510
Unigram+karakter n-gram (n=2)	0,6745	0,6779	0,6678	0,5235	0,6107	0,6745
Bigram+karakter n-gram (n=2)	0,6577	0,6544	0,6577	0,6007	0,6174	0,6611
Trigram+karakter n-gram (n=2)	0,6107	0,6544	0,6342	0,6040	0,5973	0,6141

5. Sonuç

COVID-19, hastalığın ilk bildirildiği dönemden bu yana, şiddetli akut solunum sendromu büyük salgınlara neden olmaktadır ve dünya çapında bir pandemiye dönüşmüştür. COVID-19 salgını, dünya çapında birçok insanın yaşamını kaybetmesine neden olmuş, aralarında sağlık, eğitim, gıda ve iş organizasyonlarının da yer aldığı birçok alanda küresel ölçekte önemli değişikliklere yol açmıştır. Bilgi ve iletişim teknolojilerindeki ilerlemeler ile pek çok farklı kaynaktan COVID-19 salgınına yönelik olarak elde edilen veriler, salgın durumuna ilişkin bilgilerin etkin ve zamanında elde edilebilmesi için büyük önem taşımaktadır. Bu çalışmada, İngiltere ve İspanya'da COVID-19 sürecine ilişkin 2020 yılının mart, mayıs ve temmuz aylarında yayınlanan 299'ar tane haber makalesi toplanarak oluşturulan derlem kullanılmaktadır. Metin belgelerinin temsilinde, üç temel n-gram modeli olan (1-gram, 2-gram ve 3-gram) temsilleri, tümce ögeleri 2-gram ve tümce ögeleri 3-gram öznitelikleri, kelime/tümce ögesi çiftleri, karakter *n*-gram (*n*=2) ve karakter *n*-gram (*n*=3) öznitelikleri ve bu özniteliklerin biraraya getirilmesi ile elde edilen topluluk öznitelik kümelerinin etkinlikleri değerlendirilmektedir. Öznitelik kümelerinin başarımlarının değerlendirilmesinde, altı temel makine öğrenmesi sınıflandırıcısı olan Naive Bayes algoritması, lojistik regresyon algoritması, destek vektör makineleri, C4.5 karar ağacı, k-en yakın komşu algoritması ve rastgele orman algoritması kullanılmaktadır. Deneysel analizlerde kullanılan on yedi farklı metin temsil yöntemi arasında en yüksek başarımın, sözcük tabanlı 1-gram özniteliklerin karakter tabanlı 3-gram modeli ile kullanıldığında elde edildiği görülmektedir. Deneysel analizlerde kullanılan temel sınıflandırma algoritmaları arasında en yüksek başarım rastgele orman algoritmasıyla, ikinci en yüksek başarım ise lojistik regresyon algoritmasıyla alınmaktadır. Deneysel analizler, makine öğrenmesi ve metin madenciliği tekniklerinin, salgın hastalıklara ilişkin sosyal medya gönderilerinin zaman/mekânsal analizi için uygun teknikler olduğunu göstermektedir.

Kaynakça

1. Chawla, S., Mittal, M., Chawla, M., & Goyal, L. M. (2020). Corona virus-SARS-CoV-2: an insight to another way of natural disaster. *EAI Endorsed Transactions on Pervasive Health and Technology*, *6*(22).
2. Wang, L. L., & Lo, K. (2021). Text mining approaches for dealing with the rapidly expanding literature on COVID-19. *Briefings in Bioinformatics*, *22*(2), 781-799.

3. Gajewski, K. N., Peterson, A. E., Chitale, R. A., Pavlin, J. A., Russell, K. L., & Chretien, J. P. (2014). A review of evaluations of electronic event-based biosurveillance systems. *PloS one*, *9*(10), e111222.
4. Onan, A., Korukoğlu, S., & Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*, 57, 232-247.
5. Onan, A. (2016). Classifier and feature set ensembles for web page classification. *Journal of Information Science*, *42*(2), 150-165.
6. Onan, A., Korukoğlu, S., & Bulut, H. (2016). A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. *Expert Systems with Applications*, *62*, 1-16.
7. Onan, A., & Korukoğlu, S. (2017). A feature selection model based on genetic rank aggregation for text sentiment classification. *Journal of Information Science*, *43*(1), 25-38.
8. Onan, A. (2017). Hybrid supervised clustering based ensemble scheme for text classification. *Kybernetes*.
9. Onan, A., & Toçoğlu, M. A. (2021). A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification. *IEEE Access*, *9*, 7701-7722.
10. Onan, A., Korukoğlu, S., & Bulut, H. (2017). A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. *Information Processing & Management*, *53*(4), 814-833.
11. Toçoğlu, M. A., & Onan, A. (2019, August). Satire detection in Turkish news articles: a machine learning approach. In *International Conference on Big Data Innovations and Applications* (pp. 107-117). Springer, Cham.
12. Onan, A. (2018, May). Review spam detection based on psychological and linguistic features. In *2018 26th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.
13. Onan, A. (2018). An ensemble scheme based on language function analysis and feature engineering for text genre classification. *Journal of Information Science*, *44*(1), 28-47.
14. Jahanbin, K., & Rahmanian, V. (2020). Using Twitter and web news mining to predict COVID-19 outbreak. *Asian Pacific Journal of Tropical Medicine*, *13*(8), 378.
15. Ordun, C., Purushotham, S., & Raff, E. (2020). Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs. *arXiv preprint arXiv:2005.03082*.
16. Peng, Z., Wang, R., Liu, L., & Wu, H. (2020). Exploring urban spatial features of COVID-19 transmission in Wuhan based on social media data. *ISPRS International Journal of Geo-Information*, *9*(6), 402.
17. Li, D., Chaudhary, H., & Zhang, Z. (2020). Modeling spatiotemporal pattern of depressive symptoms caused by COVID-19 using social media data mining. *International Journal of Environmental Research and Public Health*, *17*(14), 4988.
18. Chen, N., Zhong, Z., & Pang, J. (2021). An exploratory study of COVID-19 information on twitter in the greater region. *Big Data and Cognitive Computing*, *5*(1), 5.
19. Boon-Itt, S., & Skunkan, Y. (2020). Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance*, *6*(4), e21978.
20. Onan, A. (2017). Twitter mesajları üzerinde makine öğrenmesi yöntemlerine dayalı duygu analizi. *Yönetim Bilişim Sistemleri, 3*(2), 1-14.
21. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.
22. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.
23. Roche, M. (2020). COVID-19 and Media datasets: Period- and location-specific textual data mining. *Data in brief*, *33*, 106356.