Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 26, S. 295-300, Temmuz 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 26, pp. 295-300, July 2021

Research Article

Haber Metinlerinden Sosyo-ekonomik ve Epidemiyolojik Konuların Metin Madenciliğine Dayalı Belirlenmesi

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.957004)

ATIF/REFERENCE: Onan, A. (2021). Haber metinlerinden sosyo-ekonomik ve epidemiyolojik konuların metin madenciliğine dayalı belirlenmesi. *Avrupa Bilim ve Teknoloji Dergisi*, (26), 295-300.

Öz

Bilgi teknolojilerindeki ilerlemeler ile, Web'te aralarında sosyo-ekonomik ve epidemiyolojik konuların da yer aldığı birçok konuda önemli boyutta metin belgeleri paylaşılmaktadır. Internetteki çeşitli platformlarda paylaşılan haber makaleleri, hastalık raporları ve haber bültenleri gibi metin-tabanlı paylaşımlar, ortaya çıkan bulaşıcı hastalık salgınlarının erken tespiti için de önemli bir bilgi kaynağı niteliğine sahiptir. Bu bilgi, web tabanlı biyo-gözetim sistemleri geliştirilmesi için de son derece kritik önem taşımaktadır. Webte yayınlanan haber makalelerinin sayısının sürekli olarak artması, bu kaynaklarının hastalık, salgın ve sosyo-ekonomik faktörleri önceden belirlemede kullanılmasını zorlaştırmaktadır. Bu nedenle, etkin bir web tabanlı biyogözetim sistemi geliştirilmesi için, haber metinlerini uygun konulara hızlı ve yüksek başarım ile atayan metin madenciliği ve makine öğrenmesi tabanlı sistemlere gereksinim duyulmaktadır. Bu çalışmada, hayvanlar üzerinde viral bir hastalık olan ASF ve sosyo-ekonomik konularda haber metinleri içeren bir derlem üzerinde temel makine öğrenmesi sınıflandırma algoritmalarının, sınıflandırıcı topluluğu mimarilerinin ve temel metin temsil yöntemlerinin başarımları karşılaştırmalı olarak değerlendirilmiştir. Haber metinlerinin temsil edilmesinde üç temel n-gram modeli olan (1-gram, 2-gram ve 3-gram) temsilleri, terim sıklığı, terim varlığı ve TF-IDF terim ağırlıklandırma yaklaşımları ile birarada kullanılarak toplam dokuz farklı metin temsili elde edilmiştir. Elde edilen metin temsilleri, dört temel sınıflandırma algoritması olan Naive Bayes algoritması, destek vektör makineleri, k-en yakın komşu algoritması ve lojistik regresyon algoritmaları ile değerlendirilmiştir. Bunun yanı sıra, torbalama yöntemi, yükseltme yöntemi, rastgele alt-uzay yöntemi ve çoğunluk oylaması algoritması kullanılarak, haber metinlerinden sosyo-ekonomik ve epidemiyolojik konuların saptanmasında, topluluk öğrenme yöntemlerinin etkinlikleri de analiz edilmiştir. Deneysel analizlerde kullanılan temel sınıflandırıcılar arasında en yüksek başarım Naive Bayes algoritması ile topluluk öğrenmesi mimarileri arasında en yüksek başarım ise rastgele alt-orman algoritmasının Naive Bayes ile kullanılmasıyla elde edilmiştir. Deneysel sonuçlar, metin madenciliği ve makine öğrenmesi yöntemlerinin salgın hastalıkların erken belirlenmesi için kullanılmasının uygun olduğunu göstermektedir.

Anahtar Kelimeler: Metin madenciliği, Makine öğrenmesi, Topluluk öğrenmesi.

Identification of Socio-economic and Epidemiological Issues from News Texts Based on Text Mining

Abstract

With the advances in information technologies, important text documents are shared on the Web on many topics, including socio-economic and epidemiological issues. Text-based posts, such as, news articles, disease reports and news bulletins shared on various platforms on the Internet are also important sources of information for early detection of emerging infectious disease outbreaks. This information is also critical for the development of web-based bio-surveillance systems. The continuous increase in the number of news articles published on the web makes it difficult to use these sources to predict disease, epidemic and socio-economic factors. Therefore, in order to develop an effective web-based bio-surveillance system, text mining and machine learning-based systems are required that assign news texts to appropriate topics with high predictive performance and speed. In this study, the performance of

[*] 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

conventional machine learning classifiers, ensemble learning architectures and conventional text representation methods were evaluated comparatively on a collection of ASF, a viral disease on animals, and news texts on socio-economic issues. A total of nine different text representations were obtained by using three basic *n*-gram model (1-gram, 2-gram and 3-gram) representations, term frequency, term existence and TF-IDF term weighting approaches to represent news texts. The text representations obtained were evaluated using five basic classification algorithms, namely, Naive Bayes algorithm, support vector machines, k-nearest neighbor algorithm, and logistic regression algorithms. In addition, the predictive performances of ensemble learning methods (namely, Bagging method, Boosting method, random subspace method and majority voting algorithm) have been evaluated on the identification of socio-economic and epidemiological issues from news texts. Among the basic classifiers used in experimental analysis, the highest performance was obtained with Naive Bayes algorithm and community learning architectures, while the highest performance was obtained by using the random sub-forest algorithm with Naive Bayes. Experimental results show that it is appropriate to use text mining and machine learning methods for early detection of epidemics.

Keywords: Text mining, Machine learning, Ensemble learning.

1. Giriş

Bilgi teknolojilerindeki ilerlemeler ile Web'te aralarında sosyo-ekonomik ve epidemiyolojik konuların da yer aldığı birçok konuda önemli boyutta metin belgeleri paylaşılmaktadır. Internetteki çeşitli platformlarda paylaşılan haber makaleleri, hastalık raporları ve haber bültenleri gibi metin-tabanlı paylaşımlar, ortaya çıkan bulaşıcı hastalık salgınlarının erken tespiti için de önemli bir bilgi kaynağı niteliğine sahiptir. Bulaşıcı hastalık salgınlarıyla ilgili dünyadaki ilk haberlerin önemli bir bölümün Internetteki çeşitli mecralarda oluşturulan gayri resmi paylaşımlardan geldiği gözlenmektedir. Dünya Sağlık Örgütü (WHO) tarafından incelenen neredeyse tüm büyük salgınların ilk olarak Internetteki gayri resmi kaynaklar üzerinden paylaşıldığı tespit edilmiştir [1]. Bu bilgi, web tabanlı biyo-gözetim sistemleri geliştirilmesi için de son derece kritik önem taşımaktadır. Web tabanlı, olay güdümlü biyo-gözetim ya da dijital hastalık tespit sistemleri, insan, hayvan ve bitkiler üzerindeki bulaşıcı hastalıklara ek olarak insanlığı etkileyen kimyasal, radyolojik ve nükleer tehditlere yönelik erken uyarı ve farkındalık sağlamak amacıyla, çeşitli web tabanlı kaynaklardan erişilen yapılandırılmamış verileri kullanır [2]. Web tabanlı biyogözetim sistemleri, geleneksel gösterge temelli halk sağlığı sürveyans yöntemlerine, gerçek zamanlı tamamlayıcı sistemler olma özelliği taşımaktadır. Bu açıdan hem hükümet organizasyonları hem de halk sağlığı uzmanları, sağlık çalışanları ve kamu ile özel sektör için salgın yönetiminin etkin ve uygun bir şekilde yönetimini olanaklı kılar [3]. Web tabanlı, olay güdümlü biyo-gözetim sistemleri, Internet teknolojilerindeki ilerlemeler ile, epidemiyolojik verilerin izlenmesi ve modellenmesi için oldukça önemli bir karar destek sistemi haline gelmiştir. Salgın gözetiminin web tabanlı biyo-gözetim sistemleri aracılığıyla yapılması ile aralarında sosyal medya platformlarının da yer aldığı birçok Internet mecrasında kullanıcılar bireysel olarak halk sağlığı olaylarına yönelik bildirimlerde bulunabilmektedir. Bu, halk sağlığı bilgilerinin tek elden kontrolünü olanaksız hale getirmekte, böylelikle, salgın hastalıklara yönelik paylaşımların, raporlamaların saklanması ya da geciktirilmesi önemli ölçüde zorlaşmaktadır [4]. Buna ilaveten, web tabanlı biyo-gözetim sistemleri, salgın bildirimi öncesi, protokol ya da doğrulayıcı herhangi bir teste gereksinim duymadığından, geleneksel gözetim sistemlerine kıyasla, salgına ilişkin çok daha hızlı bildirim sağlama niteliğine sahiptir [5].

Etkin bir web tabanlı biyogözetim sistemi geliştirilmesi için, haber metinlerini uygun konulara hızlı ve yüksek başarım ile atayan metin madenciliği ve makine öğrenmesi tabanlı sistemlere gereksinim duyulmaktadır. Bu çalışmada, hayvanlar üzerinde viral bir hastalık olan ASF ve sosyo-ekonomik konularda haber metinleri içeren bir derlem üzerinde temel makine öğrenmesi sınıflandırma algoritmalarının, sınıflandırıcı topluluğu mimarilerinin ve temel metin temsil yöntemlerinin başarımları karşılaştırmalı olarak değerlendirilmiştir. Haber metinlerinin temsil edilmesinde üç temel n-gram modeli olan (1-gram, 2-gram ve 3-gram) temsilleri, terim sıklığı, terim varlığı ve TF-IDF terim ağırlıklandırma yaklaşımları ile birarada kullanılarak toplam dokuz farklı metin temsili elde edilmiştir. Elde edilen metin temsilleri, dört temel sınıflandırma algoritması olan Naive Bayes algoritması, destek vektör makineleri, k-en yakın komşu algoritması ve lojistik regresyon ile değerlendirilmiştir. Bunun yanı sıra, torbalama yöntemi, yükseltme yöntemi, rastgele alt-uzay yöntemi ve çoğunluk oylaması algoritması kullanılarak, haber metinlerinden sosyo-ekonomik ve epidemiyolojik konuların saptanmasında, topluluk öğrenme yöntemlerinin etkinlikleri de analiz edilmiştir. Deneysel analizlerde kullanılan temel sınıflandırıcılar arasında en yüksek başarım Naive Bayes algoritması ile topluluk öğrenmesi mimarileri arasında en yüksek başarım ise rastgele alt-orman algoritmasının Naive Bayes ile kullanılmasıyla elde edilmiştir. Deneysel sonuçlar, metin madenciliği ve makine öğrenmesi yöntemlerinin salgın hastalıkların erken belirlenmesi için kullanılmasının uygun olduğunu göstermektedir. Çalışmanın geri kalanı şu şekildedir: İkinci bölümde ilgili çalışmalar, üçüncü bölümde çalışmanın metodolojisi, dördüncü bölümde deneysel sonuçlar ve tartışma, beşinci bölümde ise temel sonuçlar tartışılmaktadır.

2. İlgili Çalışmalar

Web tabanlı biyogözetim sistemleri ile ilgili temel çalışmalardan biri, 1997 yılında Kanada'da Dünya Sağlık Örgütü iş birliği ile hayata geçirilen küresel halk sağlığı istihbarat ağı (global public health intelligence network) projesidir [6]. Bu proje, dünya çapında olası hastalıkları ve diğer sağlık tehditlerini tespit etmeye ve bildirmeye yardımcı olmak için geliştirilmiştir. Sistem, Internette aralarında haber paylaşım sistemlerinin de yer aldığı yaklaşık 32,000 civarındaki kaynaklardan, dokuz farklı dilde veri toplayarak, salgın hastalıkların izlenmesi ve gözetimi sürecini yürütmektedir [7]. Web tabanlı biyogözetime yönelik olarak gerçekleştirilen bir diğer proje, Avrupa Komisyonu'nun ortak araştırma merkezi tarafından 2014 yılında hayata geçirilen tıbbi bilgi sistemi projesidir. Bu proje, dünya çapında olası salgın hastalıklara yönelik gözetim gerçekleştirmek amacıyla, yüzün üzerinde tıbbi web sitesi ve bin civarında haber paylaşım sitesinden elde edilen kırk farklı dilde verileri toplamaktadır [8, 9]. Benzer şekilde, ABD'deki Georgetown Üniversitesi Tıp Fakültesi bünyesinde gerçekleştirimi yapılan Argus projesinde, Dünya genelinde gözlemlenen insan, bitki ya da hayvan sağlığını tehdit etme potansiyeli bulunan salgın hastalıklara yönelik biyogözetim gerçekleştirilmektedir. Bu proje kapsamında, aralarında Dünya Sağlık Örgütü gibi resmi kaynakların da bulunduğu kaynaklardan bulaşıcı hastalıklara yönelik olarak haber

verilerinin toplanması, toplanan verilerin, uygun kavramlara, arama terimlerine ve sorgulara uygunluklarına dayalı olarak yorumlanması yapılmaktadır. Argus sistemi, kırkın üzerinde farklı dilde toplanan makale ve diğer kaynaklardan verileri işlemektedir [10]. Web tabanlı biyogözetim alanında gerçekleştirilen bir diğer proje, 2006 yılında Boston Çocuk Hastanesi tarafından gerçekleştirimi yapılan HealthMap projesidir [6]. Bu proje, sürekli olarak veri eklenen bir tıbbı sözlükten insan, bitki ve hayvan hastalıkları ile coğrafi isimlere yönelik veri almaktadır. Sistemin veri kaynağını, Google haberler, biyogözetim raporları ve onaylanmış resmi uyarı ve duyurular gibi birçok farklı kaynaktan elde edilmiş veriler oluşturmaktadır. Proje kapsamında, makine öğrenmesi sınıflandırma algoritmaları kullanılarak, makaleler hastalık ve konuma göre uygun konulara atanmakta, otomatik olarak etiketlenmekte ve coğrafi bir harita üzerinde görsel olarak sunulmaktadır [11]. Benzer biçimde, 2013 yılında Melbourne Üniversitesi tarafından uluslararası biyolojik gözetim sistemi adı verilen bir biyogözetim sistemi gerçekleştirimi yapılmıştır. Bu proje kapsamında, arama motorlarından elde edilen genel arama örüntüleri, haber metinleri ve bloglar üzerinde yapılan paylaşımlar veri kaynağı olarak kullanılmıştır. Böylelikle, bitki ve hayvanlara yönelik bulaşıcı hastalıkların erken tespiti ve gözetimi yapılmaktadır [12].

Metin sınıflandırma, bir metin belgesini bir veya daha fazla önceden tanımlanmış sınıfa veya kategoriye atayan, metin madenciliğinin önemli bir alt alanıdır [13]. Metin madenciliği, web sayfası sınıflandırma [14], duygu analizi [15-20], istenmeyen belgelerin filtrelenmesi [21] ve metin türü belirleme [22] gibi birçok farklı alanda başarıyla uygulanmaktadır. Etkin bir web tabanlı biyogözetim sistemi geliştirilmesi için, metin belgelerinin uygun konulara yüksek başarım ile atanması önem taşımaktadır. Bu nedenle, çalışma kapsamında ilgili problem için temel metin temsil yöntemleri ve sınıflandırma algoritmalarının etkinlikleri değerlendirilmektedir.

3. Metodoloji

Bu bölümde, çalışmada kullanılan temel metin temsil yöntemleri, temel sınıflandırma algoritmaları ve topluluk öğrenmesi algoritmalarına değinilmektedir.

3.1. Metin Belgelerinin Temsili

Metin madenciliği alanında yapılan çalışmalarda, metin belgelerinin temsilinde en yaygın kullanıma sahip modellerden biri *n*-gram metin temsil yöntemidir. Bir *n*-gram, verilen bir metin belgesi örneğinden, *n* öğeden oluşan bitişik bir dizidir. Hece, harf, kelime ve karakter gibi ögeler bu şekilde modellenebilir. Kelime tabanlı *n*-gram modelleri ve karakter tabanlı *n*-gramlar, doğal dil işlemede sıklıkla kullanılmaktadır [23]. N-gram boyutu bir olarak alındığında "unigram" olarak, *n*-gram boyutu iki olarak alındığında "bigram" ve *n*-gram boyutu üç olarak alındığında "trigram" olarak adlandırılır. Bu çalışmada, kelime tabanlı *n*-gram modelleme kullanılarak, unigram, bigram ve trigram temsilleri elde edilmiştir. Metin belgelerini sınıflandırmak için, vektör uzay modelinde, terim varlığı (TP), terim sıklığı (TF) ve TF-IDF ağırlıklandırma ölçütü dikkate alınarak toplam dokuz farklı metin temsil konfigürasyonu ele alınmıştır.

3.2. Temel Sınıflandırma Algoritmaları

Çalışma kapsamında, dört temel sınıflandırma algoritması (Naive Bayes algoritması, destek vektör makineleri, k-en yakın komşu algoritması ve lojistik regresyon algoritması) değerlendirilmiştir. Bu bölümün geri kalanında, ilgili öğreticili öğrenme algoritmaları kısaca tanıtılmaktadır.

3.2.1. Naive Bayes algoritması

Naive Bayes algoritması (NB), Bayes teoremine dayalı olasılık tabanlı temel bir öğrenme algoritmasıdır. NB algoritması, öğrenme modelini kurgularken, veri setinde yer alan özniteliklerin sınıf etiketini belirlemede birbirlerinden bağımsız olduğu varsayımı olan koşullu bağımsızlık varsayımına dayanır. Algoritmanın, koşullu bağımsızlık varsayımına dayanması ve basit yapısına karşın, metin madenciliği, istenmeyen e-posta filtreleme ve web madenciliği gibi birçok alanda daha karmaşık yapıya sahip öğrenme algoritmalarıyla rekabet edebilir performanslar gösterdiği gözlenmektedir [13, 14].

3.2.2. Destek vektör makineleri algoritması

Destek vektör makineleri (SVM), sınıflandırma ve regresyon uygulamalarında kullanılabilen öğreticili bir öğrenme yöntemidir. SVM hem doğrusal hem de doğrusal olmayan sınıflandırma problemlerinde başarıyla uygulanabilmektedir [24]. Destek vektör makineleri, sınıflandırma sürecini temelde veri setini daha yüksek boyutlu bir hiper düzlem oluşturacak şekilde bölümleyerek gerçekleştirir. Burada hiper düzlemin temel amacı, sınıfların en yakın eğitim örneklerine maksimum mesafeye ulaşarak iyi bir ayrım elde etmektir, buna işlevsel sınır adı verilir.

3.2.3. K-en yakın komşu algoritması

K-en yakın komşu algoritması (KNN), örnek tabanlı bir sınıflandırma algoritmasıdır. KNN algoritmasında, sınıf etiketi belirlenmek istenen örnek için, öncelikle veri setinden *k* tane en yakın (en benzeyen) örnek belirlenir. Ardından, sınıflandırılmak istenen örnek, en yakın komşularının sınıf etiketlerinin çoğunluk oylamasına dayalı olarak sınıf etiketine atanır.

3.2.4. 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 [23]. Lojistik regresyon algoritmasında, sınıf etiketlerinin belirlenmesinde kullanılan olasılık değeri, doğrudan parametreler üzerinde doğrusal fonksiyona dayalı olarak hesaplanır.

3.3. Topluluk Öğrenmesi Algoritmaları

Topluluk öğrenme algoritmaları, sınıflandırılacak olan örneklere sınıf etiketini, tek bir sınıflandırma algoritması yerine birden fazla öğrenme algoritmasının çıktısına göre atamaya yönelik bir makine öğrenmesi çalışma alanıdır. Topluluk öğrenme algoritmalarının, temel sınıflandırıcı algoritmalarına kıyasla, genelleştirme yeteneklerinin daha iyi olması ve aşırı uygunluk gösterme risklerinin daha düşük olması beklenmektedir [25]. Bu bölümün geri kalanında, çalışmada kullanılan topluluk öğrenmesi algoritmaları kısaca tanıtılmaktadır.

3.3.1. Torbalama algoritması

Torbalama algoritması (Bagging algoritması), eğitim setini öncelikle, rastgele yeniden yerine koyma ile altkümelere ayırır. Ardından, altkümeler üzerinde temel öğrenme algoritmaları eğitilerek öğrenme modelleri oluşturulur. Temel öğrenme algoritmalarının çıktıları ise, çoğunluk oylamasına tabi tutularak, sınıflandırılmak istenen örneğin sınıf etiketi belirlenmiş olur [26].

3.3.2. Destekleme algoritması

Destekleme algoritması, temel öğrenme algoritmalarının, farklı dağılışlara sahip eğitim setleri üzerinde yinelemeli olarak eğitilmesi ve ardından temel öğrenme algoritmaları sonucu elde edilen öğrenme modellerinin birleştirilmesi ile tek bir güçlü sınıflandırıcı elde edilmesi amacını taşır [27]. AdaBoost (adaptif destekleme) algoritmasında, sınıflandırılması zor olan örneklere daha fazla odaklanmak amacıyla, ilgili örneklerin ağırlık değerleri yinelemeli olarak artırılır. Deneysel analizlerde, AdaBoost algoritması kullanılmıştır.

3.3.3. Rastgele alt-uzay algoritması

Rastgele alt-uzay algoritması (RS), torbalama algoritmasında olduğu gibi eğitim setinden örneklemler alınarak temel öğrenme algoritmalarının eğitildiği bir topluluk öğrenmesi algoritmasıdır. Ancak, burada, eğitim setinden farklı altkümelerin elde edilmesinde, örnek tabanlı değil, öznitelik uzayı tabanlı bölümleme gerçekleştirilir [28].

3.3.4. Çoğunluk oylaması algoritması

Sınıflandırma algoritmalarının birleştirilmesinde yaygın kullanıma sahip birleştirme kurallarından biri çoğunluk oylamasıdır. Burada, sınıflandırıcı topluluğunu oluşturan temel öğrenme algoritmalarının çıktıları çoğunluk oyuna tabi tutularak en yüksek oyu alan sınıf etiketinin, topluluğun sınıf etiketi olarak alınması söz konusudur [15].

4. Deneysel Süreç ve Sonuçlar

Bu bölümde, deneysel analizlerde kullanılan veri seti, gerçekleştirimde kullanılan yöntemler ve deneysel sonuçlara ilişkin bilgiler sunulmaktadır. Deneysel analizlerde, ASF derlemi kullanılmıştır. ASF derlemi, hayvanlar üzerinde viral bir hastalık olan ASF ve sosyo-ekonomik konularda haber metinleri içeren bir derlemdir [29]. ASF derlemi, epidemiyolojik konular ile ilgili 69 haber içermektedir. Bu haberler, ASF şüphesi ile ilgili, çeşitli hayvanlarda bilinmeyen hastalık ya da açıklanamayan klinik bulgulara ilişkin temel verileri içeren metinlerden oluşmaktadır. Benzer şekilde veri setinde, ASF salgınının bir ülke ya da bölge üzerindeki sosyo-ekonomik etkisi hakknda 69 haber içeren ikinci bir kategori bulunmaktadır. Deneysel analizlerde kullanılan temel öğrenme algoritmalarının ve topluluk öğrenme algoritmalarının 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 temel sınıflandırma algoritmaları ile ve çoğunluk oylaması ile elde edilen deneysel sonuçlar, Tablo 2'de ise topluluk öğrenmesi algoritmaları ile elde edilen sonuçlar sunulmaktadır. Deneysel analizlerde ele alınan dokuz temel metin temsiline ilişkin sonuçlar incelendiğinde, en yüksek başarımın TF-IDF ağırlıklandırma temsili ve bigram n-gram modeli ile elde edildiği görülmektedir. TF-IDF ağırlıklandırma yönteminin, terim sıklığı ya da terim varlığına dayalı temsile kıyasla daha yüksek sonuçlar verdiği görülmektedir. Temel öğrenme algoritmaları arasında, en yüksek başarım Naive Bayes algoritması ile elde edilmektedir. Naive Bayes algoritmasını, lojistik regresyon sınıflandırıcısının izlediği görülmektedir. Üçüncü en yüksek başarım destek vektör makineleri ile en kötü doğru sınıflandırma oranları ise k-en yakın komşu algoritması ile elde edilmektedir. Çoğunluk oylamalası yöntemi ile dört temel sınıflandırma yöntemi birleştirildiğinde performansta iyileşme elde edilemediği gözlenmektedir.

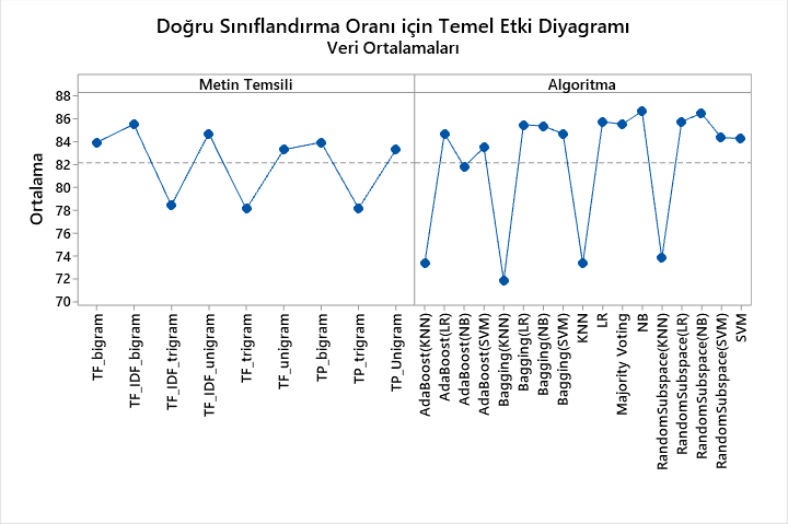
Tablo 1. Temel Sınıflandırma Yöntemlerine İlişkin Deneysel Sonuçlar

Metin Temsili	NB	SVM	LR	KNN	Çoğunluk Oylaması
TF_IDF_bigram	92,62	87,12	88,71	75,31	88,86
TF_IDF_trigram	80,96	82,47	85,17	68,52	84,74
TF_IDF_unigram	89,86	86,70	87,85	76,21	87,99
TF_bigram	91,33	83,84	85,93	75,31	85,57
TF_trigram	77,80	83,21	83,99	68,52	82,92
TF_unigram	88,13	84,01	85,10	76,21	85,25
TP_bigram	91,47	83,77	85,86	75,31	85,93
TP_trigram	79,25	83,21	84,00	68,52	83,14
TP_unigram	88,42	84,01	85,03	76,21	85,25

Tablo 2. Topluluk Öğrenmesi Algoritmalarına İlişkin Deneysel Sonuçlar

	Adaptif destekleme				Torbalama				Rastgele alt-uzay			
Metin Temsili	**NB**	**SVM**	**LR**	**KNN**	**NB**	**SVM**	**LR**	**KNN**	**NB**	**SVM**	**LR**	**KNN**
TF_IDF_bigram	83,90	86,70	88,30	75,30	92,60	87,60	88,60	74,50	92,80	87,00	88,60	75,10
TF_IDF_trigram	70,40	80,90	82,80	68,50	77,60	82,90	85,20	66,00	80,70	83,60	85,50	67,50
TF_IDF_unigram	84,80	84,90	86,20	76,20	89,60	87,00	87,80	75,10	89,70	85,50	86,80	78,00
TF_bigram	89,40	82,90	84,30	75,30	90,60	84,40	85,40	74,50	91,10	84,10	85,40	77,30
TF_trigram	73,00	82,20	83,00	68,50	75,10	83,40	84,20	66,00	78,00	83,60	85,30	68,90
TF_unigram	86,00	84,20	85,40	76,20	87,60	84,50	84,50	75,10	87,50	83,90	85,50	76,80
TP_bigram	89,90	83,50	84,10	75,30	91,50	84,50	85,00	74,50	91,10	84,50	85,70	75,10
TP_trigram	72,60	81,70	83,00	68,50	76,10	83,40	84,20	66,00	79,20	83,40	84,40	68,00
TP_Unigram	86,20	84,40	85,40	76,20	87,60	84,50	84,50	75,10	88,20	83,60	84,30	77,60

Tablo 2'de topluluk öğrenmesi algoritmaları ile elde edilen sonuçlar listelenmektedir. Tablo 2'deki doğru sınıflandırma oranlarına bakıldığında, rastgele alt-uzay algoritmasının diğer topluluk öğrenmesi yöntemlerine kıyasla daha yüksek başarım elde ettiği görülmektedir. Şekil 1'de ise karşılaştırılan yöntemlere ve faktörlere ilişkin temel etki diyagramı aracılığıyla temel sonuçlar özetlenmektedir.



Şekil. 1. Karşılaştırılan yöntemlere ve faktörlere ilişkin temel etki diyagramı

5. Sonuç

Etkin bir web tabanlı biyogözetim sistemi geliştirilmesi için, haber metinlerini uygun konulara hızlı ve yüksek başarım ile atayan metin madenciliği ve makine öğrenmesi tabanlı sistemlere gereksinim duyulmaktadır. Bu çalışmada, hayvanlar üzerinde viral bir hastalık olan ASF ve sosyo-ekonomik konularda haber metinleri içeren bir derlem üzerinde temel makine öğrenmesi sınıflandırma algoritmalarının, sınıflandırıcı topluluğu mimarilerinin ve temel metin temsil yöntemlerinin başarımları karşılaştırmalı olarak değerlendirilmiştir. Haber metinlerinin temsil edilmesinde üç temel n-gram modeli olan (1-gram, 2-gram ve 3-gram) temsilleri, terim sıklığı, terim varlığı ve TF-IDF terim ağırlıklandırma yaklaşımları ile birarada kullanılarak toplam dokuz farklı metin temsili elde edilmiştir. Elde edilen metin temsilleri, dört temel sınıflandırma algoritması olan Naive Bayes algoritması, destek vektör makineleri, k-en yakın komşu algoritması ve lojistik regresyon ile değerlendirilmiştir. Bunun yanı sıra, torbalama yöntemi, yükseltme yöntemi, rastgele alt-uzay yöntemi ve çoğunluk oylaması algoritması kullanılarak, haber metinlerinden sosyo-ekonomik ve epidemiyolojik konuların saptanmasında, topluluk öğrenme yöntemlerinin etkinlikleri de analiz edilmiştir. Deneysel sonuçlar, metin madenciliği ve makine öğrenmesi yöntemlerinin salgın hastalıkların erken belirlenmesi için kullanılmasının uygun olduğunu göstermektedir. En yüksek başarım Naive Bayes rastgele alt-uzay algoritması ile birlikte kullanıldığında elde edilmiştir.

Kaynakça

1. 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.
2. Walters, R. A., Harlan, P. A., Nelson, N. P., & Hartley, D. M. (2008). Data sources for biosurveillance. *Wiley handbook of science and technology for Homeland Security*, 1-17.
3. Hartley, D. M., Nelson, N. P., Arthur, R. R., Barboza, P., Collier, N., Lightfoot, N., ... & Brownstein, J. S. (2013). An overview of internet biosurveillance. *Clinical Microbiology and Infection*, *19*(11), 1006-1013.
4. Tsai, F. J., Tseng, E., Chan, C. C., Tamashiro, H., Motamed, S., & Rougemont, A. C. (2013). Is the reporting timeliness gap for avian flu and H1N1 outbreaks in global health surveillance systems associated with country transparency?. *Globalization and health*, *9*(1), 1-7.
5. Hartley, D., Nelson, N., Walters, R., Arthur, R., Yangarber, R., Madoff, L., ... & Lightfoot, N. (2013). Landscape of

international event-based biosurveillance. Emerg Health Threats J. 2010; 3: e3.
6. Keller, M., Blench, M., Tolentino, H., Freifeld, C. C., Mandl, K. D., Mawudeku, A., ... & Brownstein, J. S. (2009). Use of unstructured event-based reports for global infectious disease surveillance. *Emerging infectious diseases*, *15*(5), 689.
7. Mykhalovskiy, E., & Weir, L. (2006). The global public health intelligence network and early warning outbreak detection. *Canadian journal of public health*, *97*(1), 42-44.
8. Mantero, J., Belyaeva, J., & Linge, J. P. (2011). How to maximise event-based surveillance web-systems the example of ECDC/JRC collaboration to improve the performance of MedISys. *Luxembourg: Publications Office of the European Union*.
9. Steinberger, R., Fuart, F., van der Goot, E., Best, C., von Etter, P., & Yangarber, R. (2008). Text mining from the web for medical intelligence. In *Mining massive data sets for security* (pp. 295-310). IOS Press.
10. Nelson, N. P., Brownstein, J. S., & Hartley, D. M. (2010). Event-based biosurveillance of respiratory disease in Mexico, 2007–2009: connection to the 2009 influenza A (H1N1) pandemic?. *Eurosurveillance*, *15*(30), 19626.
11. Freifeld, C. C., Mandl, K. D., Reis, B. Y., & Brownstein, J. S. (2008). HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *Journal of the American Medical Informatics Association*, *15*(2), 150-157.
12. Lyon, A., Grossel, G., Burgman, M., & Nunn, M. (2013). Using internet intelligence to manage biosecurity risks: a case study for aquatic animal health. *Diversity and Distributions*, *19*(5-6), 640-650.
13. Onan, A., Korukoğlu, S., & Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*, 57, 232-247.
14. Onan, A. (2016). Classifier and feature set ensembles for web page classification. *Journal of Information Science*, *42*(2), 150-165.
15. 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.
16. 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.
17. Onan, A. (2017). Hybrid supervised clustering based ensemble scheme for text classification. *Kybernetes*.
18. 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.
19. 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.
20. 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.
21. 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.
22. 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.
23. Onan, A. (2017). Twitter mesajları üzerinde makine öğrenmesi yöntemlerine dayalı duygu analizi. *Yönetim Bilişim Sistemleri, 3*(2), 1-14.
24. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.
25. Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1-15). Springer, Berlin, Heidelberg.
26. Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.
27. Schapire, R. E. (2013). Explaining adaboost. In *Empirical inference* (pp. 37-52). Springer, Berlin, Heidelberg.
28. Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, *20*(8), 832-844.
29. Arsevska, E., Roche, M., Hendrikx, P., Chavernac, D., Falala, S., Lancelot, R., & Dufour, B. (2016). Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web. *Computers and Electronics in Agriculture*, *123*, 104-115.
30.