DOI:10.25092/baunfbed.843909 *J. BAUN Inst. Sci. Technol., 23(1), 151-172, (2021)*

Twitter üzerinde Türkçe sahte haber tespiti

Süleyman Gökhan TAŞKIN[1,*], Ecir Uğur KÜÇÜKSİLLE[2], Kamil TOPAL[3]

[1]Bandırma Onyedi Eylül Üniversitesi Bilgi İşlem Daire Bşk., Bandırma, Balıkesir
[2]Süleyman Demirel Üniversitesi Müh. Fak. Bilgisayar Müh. Böl., Batı kampüsü, Isparta
[3]Balıkesir Üniversitesi Müh. Fak. Bilgisayar Müh. Böl., Çağış kampüsü, Balıkesir

Geliş Tarihi (Received Date): 25.04.2020
Kabul Tarihi (Accepted Date): 23.10.2020

Öz

Son yıllarda internet kullanımının artmasıyla insanların bilgi ve haber alma kaynakları da değişmiştir. Radyo, televizyon, gazete ve dergi gibi geleneksel medya platformları yerine sosyal medya platformlarının kullanımı giderek artmaktadır. Geleneksel medyada haberler belirli bir kaynak tarafından gönderilirken, sosyal medyada her kullanıcı bir haber kaynağı olabilmektedir. Bu da sosyal medyadaki haberlerin bir süzgeçten geçirilmeden paylaşılmasına ve sahte haberlerin büyük bir hızla yayılmasına neden olmaktadır. Sahte haber; propaganda, provokasyon veya insanları yanıltma amacıyla sahte veya provokatif kullanıcılar tarafından yayılan haberlerdir. Dikkat çekici özellikte oldukları için sosyal medya aracılığı ile çok kısa sürede yayılabilmektedirler. Bu nedenle sahte haberlerin en kısa sürede tespit edilmesi büyük öneme sahiptir. Çoğu sosyal medya platformunda sahte haber tespiti uzmanlar tarafından yapılmaktadır. Çok yoğun paylaşım trafiği bulunan sosyal medya platformlarında uzmanlar tarafından kısa sürede sahte haber tespiti mümkün olmamaktadır. Bu da sahte haberin kısa sürede çok kişi tarafından paylaşılmasına neden olmaktadır. Bu nedenle yarı otomatik ve otomatik sahte haber tespiti sistemleri, uzmanlara göre daha kısa sürede sahte haber tespitini sağlayabilmektedir. Sahte haberleri kısa sürede tespit edebilmek için otomatik tespit sistemlerinin geliştirilmesi gerekmektedir. Bu çalışmada Türkçe dilinde, denetimli ve denetimsiz makine öğrenmesi algoritmaları kullanılarak Twitter üzerinde sahte haber tespiti yapılmış ve sonuçları incelenmiştir. Denetimsiz öğrenme algoritmalarından, K-ortalamalar (K-means), Negatif Olmayan Matris Çarpımı (Non-Negative Matrix Factorization-NMF) ve Doğrusal Diskriminant Analizi (Linear Discriminant Analysis-LDA); denetimli öğrenme algoritmalarından, K En Yakın Komşu (K Nearest Neighbor-KNN), Destek Vektör Makinaları (Support Vector Machines-SVM) ve Rassal Orman (Random Forest-RF) algoritmaları ile tahmin yapılmıştır. Her bir algoritma 100 defa çalıştırılarak ortalama F1 metrik değerleri incelenmiştir. Denetimli öğrenme algoritmalarında 0.86 F1-metrik değeriyle başarılı sonuçlar alınmıştır. Denetimsiz öğrenme algoritmalarının F1-metrik değeri ise 0.72'de kalmıştır.

Anahtar kelimeler: *Sahte haber tespiti, makine öğrenmesi, yapay zeka.*

[*] Süleyman Gökhan TAŞKIN, taskinster@gmail.com, https://orcid.org/0000-0002-1535-7462
Ecir Uğur KÜÇÜKSİLLE, ecirkucuksille@sdu.edu.tr, https://orcid.org/0000-0002-3293-9878
Kamil TOPAL, kamiltopal@balikesir.edu.tr, https://orcid.org/0000-0002-0266-7365

Turkish fake news detection on twitter

Abstract

In recent years, news and their sources have transformed with the increasing use of the internet. Instead of traditional media platforms such as radio, television, newspaper and magazine, the use of social media platforms is also growing. While certain sources share the news in traditional media, every user can be a news source in social media. Fake news is news produced by fake or provocative users for the purpose of propaganda, provocation or misleading users. Since an ordinary social media user may share any news without any filter and they are usually interesting, a fake news can spread rapidly. For this reason, it is very important to detect fake news as soon as possible. Sometimes, fake news is detected by expert systems. It is not possible to detect fake news in a short time with such expert systems on social media platforms with very dense sharing traffic. This causes fake news to be shared by many people in a short time. Therefore, semi-automatic and automatic fake news detection systems can provide fake news detection in a shorter time than non-autonomous expert systems. Automatic detection systems are needed to be developed in order to overcome this shortcoming. In this study, we collect data from Twitter, annotate them whether they are fake or real news. Then, we use supervised (K-Nearest Neighbor-KNN, Support Vector Machines-SVM, and Random Forest) and unsupervised (K-means, Non-Negative Matrix Factorization-NMF, and Linear Discriminant Analysis-LDA) machine learning algorithms to detect fake news automatically. We run each algorithm 100 times and the average F1-score values were examined. The best results were obtained with 0.86 F1-score value in supervised learning algorithms. The F1- score value of unsupervised learning algorithms remained at 0.72.

Keywords: *Fake news detection, machine learning, artificial intelligence.*

1. Giriş

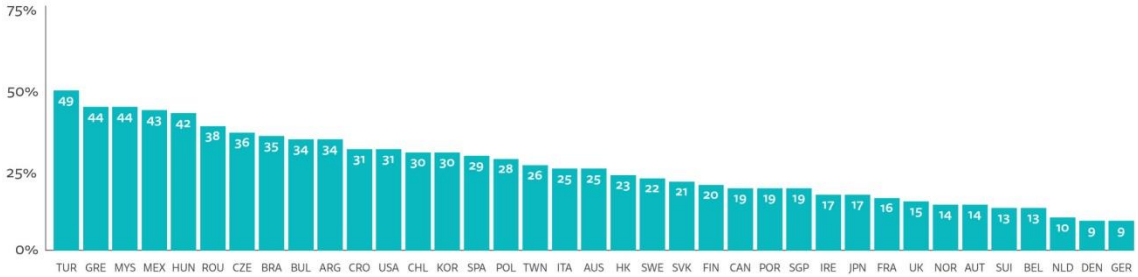
Geleneksel haber paylaşımlarında, iletişim tek yönlü olduğu için haber kaynaklarının sayıları sınırlıdır. Dolayısıyla, bu kaynakları denetlemek ve haberlerin içeriğini önceden kontrol etmek daha kolay olduğu için toplumun sahte haberlere maruz kalmasını engellemektedir. Dijital çağa geçişle birlikte, insanların habere erişim kaynakları çok çeşitlilik göstermeye başlamıştır. İletişim tek yönlü olmaktan çıkıp, haber alıcılarının habere karşı gösterebilecekleri tepkiler de dijitalleşmiştir. Bu tepkiler, paylaşmak, beğenmek/beğenmemek, yorum yapmak gibi farklı türlerde olabilmektedir. Kişiler, haberleri paylaşarak aynı zamanda haber kaynağı olduğundan, tüm paylaşımların denetleyici kurumlar tarafından gerçek kişiler (denetçiler) ile doğruluğunun kontrol edilmesi mümkün değildir. Sahte haber paylaşımlarının ilk 2 saat ile 20. saate kadar hızla yayılması göz önünde bulundurulduğunda sahte haber tespitinin otomatik sistemler ile anlık olarak tespit edilmesi büyük önem taşımaktadır [1].

Kurumsal hesapları veya siteleri taklit ederek birçok dolandırıcılık yöntemleri kullanılmaktadır. Örneğin; T.C. İletişim Başkanlığı tarafından 04.01.2020 tarihinde atılan tweet mesajında, kurumun logo ve ismini kullanarak elektrik ve doğal gaz fatura iadesi bildirimi yaparak dolandırıcılık yapılmaya çalışıldığı bildirilmiştir [2]. Ayrıca 07.01.2020 tarihinde sahte bir Twitter hesabından Eskişehir'de okulların yoğun kar yağışı dolayısıyla

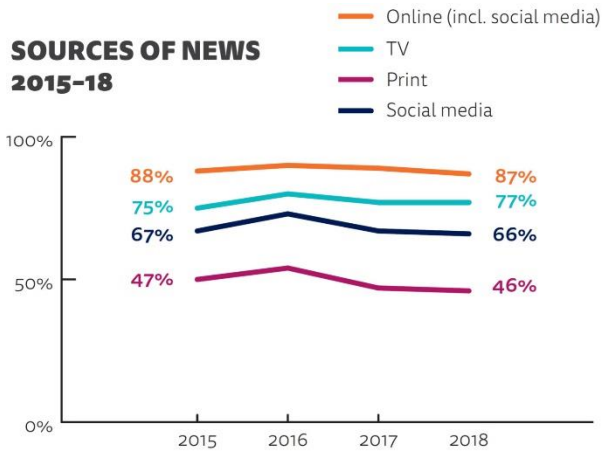
tatil olduğu sahte mesajı çok hızlı bir şekilde yayılmıştır [3]. İhlas Haber Ajansı tarafından bu haberin yetkililer tarafından yalanlandığı bildirilmiştir [4]. Yarı otomatik ve otomatik sahte haber tespiti sistemleri ile bunun gibi çok hızlı yayılan sahte ve dolandırma amaçlı haberlerin hızlıca önlenmesi gerekmektedir. Bu sayede birçok insanın dolandırılması veya yanlış bilgilendirilmesi önlenebilir.

Sahte haberin en genel tanımını, Shu vd., kasıtlı olarak yapılan ve kesinlikle yanlış olan bir haber makalesi olarak tanımlamıştır. Ayrıca geleneksel medya ve sosyal medyayı sahte haber bakımından karşılaştırmıştır. Sahte haberler, genellikle bot veya trol hesaplardan yapıldığını bildirmişler, trol hesapları, insan tarafından kontrol edilen, propaganda amaçlı hesaplar olarak ve bot hesapları ise propaganda amaçlı oluşturulmuş bilgisayar tarafından kontrol edilen hesaplar olarak tanımlamışlar ve bu hesapların kısa sürede ve çok sayıda oluşturulduğunu belirtmişlerdir. Aynı anda birden fazla bot veya trol hesaptan paylaşılan bir haber normal kullanıcılar tarafından gerçekmiş gibi algılanabildiğini belirtmişlerdir [5].

Türkçe dilinde sahte haberle mücadele çalışmaları, İngilizce dilinde sahte haberle mücadele çalışmalarına göre yetersizdir. Literatürde Türkçe dilinde sahte haberle mücadele için yapılmış çalışmalar oldukça az sayıdadır. Bu da Türkiye'de yaşayan insanların sahte habere maruz kalma oranını yükseltmiştir. Newman vd., 2018 yılındaki raporunda ülkelere göre sahte habere maruz kaldığını belirten kişilerin oranının %49 ile Türkiye'de en fazla olduğu belirtilmiştir [6] (Şekil 1). Aynı raporda Türkiye'deki kişilerin %87'si haberleri sosyal medya ve çevrimiçi kaynaklardan takip ettiklerini belirtmişlerdir (Şekil 2). Yine Newman vd. 2019 raporunda 38 ülkede yapılan araştırmaya göre bu ülkelerde yaşayan kişilerin ortalama %55'i internette gerçek ve sahte haberi ayırma yeteneklerinden endişe duymaktadır. Türkiye'de yaşayanlarda ise bu oran %63'tür [7] (Şekil 3). Bu durum Türkiye'de sahte bir haberin çok hızlı bir şekilde yayılmasına neden olmaktadır.



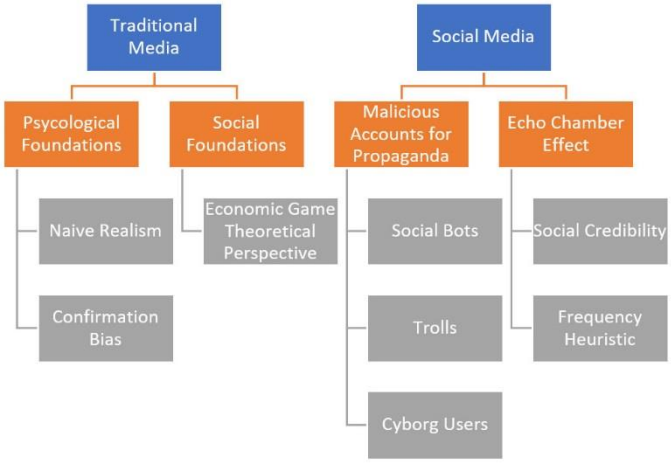
Şekil 1. Ülkelere göre sahte habere maruz kaldığını belirten kişilerin oranı [5].



Şekil 2. Türkiye'de yaşayan kişilerin haberlere erişim kaynakları [5].

Şekil 3. Ülkelere göre sahte habere maruz kaldığını belirten kişilerin oranı [5].

Sahte haber tespiti yapılmadan önce sahte haberin iyi incelenmesi gerekmektedir. İyi kategorize edilmiş bir sahte haberin tespit edilmesi de daha kolay olacaktır. Shu vd. çalışmasında, Şekil 4'de görüldüğü gibi geleneksel medyadaki sahte haberleri, psikolojik ve sosyal bulgular bakımından; sosyal medyadaki sahte haberleri ise iki farklı kategoride incelemişlerdir. Bu kategorilerden ilki, propaganda için açılmış sahte hesaplar ve ikincisi ise kullanıcıların aynı fikirde oldukları kullanıcıları takip ettiklerinden ve bu hesaplardan gelen haberlere güvendiklerinden dolayı sahte haber olsa bile kendi ilgisine yakın haberleri alıp, paylaşmaları olarak tanımladıkları yankı odası etkisidir (echo chamber effect). Sosyal medyada propaganda için açılmış sahte hesaplar tarafından yayılan sahte haberlerin sosyal botlar, trol hesaplar ve yarı robot hesaplar tarafından yayıldığını belirtmişlerdir. Trol veya bot hesaplardan birbirine yakın zamanlarda çok sayıda paylaşılan haber gerçek kullanıcılar için doğru gibi görülerek kısa sürede çok sayıda gerçek hesap tarafından da paylaşılmaktadır. Gerçek hesaplardan paylaşılan sahte haberlerin inandırıcılığı bu sayede artmaktadır [5].



Şekil 4. Sahte haber türleri [1].

Sahte haber tespiti üzerine çalışmalar 2011 yılında Zhao ve Jiang'ın çalışmasıyla başlamış ve 2016 yılındaki Amerika Başkanlık Seçimlerindeki komplo teorileri ile bu çalışmalar artmıştır [8]. İngilizce dilinde birçok çalışma olmasına rağmen Türkçe dilinde otomatik sahte haber tespiti üzerine çalışmalar oldukça sınırlı sayıda kalmıştır.

Türkiye'deki insanların çoğunun sahte haberi ayırt etme yeteneklerinden şüphe etmesi ve sahte haberlerin kısa sürede çok fazla hesap tarafından paylaşılması sahte haber tespitinin önemini göstermektedir. Uzmanlar, kullanıcıları sahte haberi bilgilendirmesi konusunda otomatik sahte haber tespiti sistemlerine göre yavaş kalmaktadır. 2 saatlik zaman diliminde büyük oranda paylaşım yapılan bu sahte haberlerin uzmanlar yerine otomatik tespit sistemleri ile tespit edilmesi önemlidir. Bu çalışmada Twitter [9] platformunda paylaşılan Türkçe dilindeki tweet mesajlarında sahte haber tespiti yapan otomatik bir sistem üzerine çalışılmıştır.

2. Kaynak özetleri

Reuters Gazeteciliği Araştırma Enstitüsü tarafından 2012 yılından itibaren her yıl Oxford Üniversitesinde, "Digital News Report" ismiyle rapor yayınlanmaktadır. Bu rapor dijital haberlerin durumunu, sosyal medyanın haber alma açısından yeri ve önemi ve birçok ülkenin sosyal medya haberleri de dahil olmak üzere dijital haberleri hakkında bilgi sunmaktadır. Ayrıca bu raporda sahte haberler ile ilgili istatistiksel bilgiler paylaşılan bölümlerde bulunmaktadır [6], [7].

Shu vd. çalışmalarında, sahte haberi tanımlamışlar, sahte haber türlerinden bahsetmişlerdir. Geleneksel medyadaki sahte haber türleri ve sosyal medyadaki sahte haber türlerini ayrı ayrı listelemişlerdir. Ayrıca hem geleneksel medyadaki hem de sosyal medyadaki sahte haberleri tespit etmek için çıkarılabilecek özellikleri de açıklamışlardır [1].

Alpaydin [10], makine öğrenmesini, sensörler ve algılayıcılardan toplanan veya veritabanında biriken çok sayıda verinin sınıflandırılması, kümelenmesi veya tahminini olanaklı kılan algoritmaların tasarım ve geliştirme süreçlerini konu alan bir bilim dalı

olarak tanımlamıştır. Makine öğrenmesi; görüntü işleme ([11], [12]), konuşma tanıma ([13], [14]) ve metin işleme ([15], [16]) gibi birçok görevde başarılı bir şekilde kullanılmaktadır.

Doğal Dil İşleme (DDİ), insanların iletişimde kullandıkları doğal dilleri, bilgisayarların anlaması için işleyen bir bilim dalıdır. İlk örnekleri, doğal dillerin birbiri arasında çevrilmesini konu alan makine çevirisi ile başlamıştır. DDİ ilk zamanlarda geleneksel yöntemlerle yapılmış ve ilerleyen zamanlarda makine öğrenmesi yöntemleri kullanılarak daha etkili sonuçlar vermeye başlamıştır. Farklı dillerde sahte haberin kısa sürede tespitinin sağlanması amacıyla da makine öğrenmesi yöntemleri kullanılmıştır.

2015 yılında, Carnegie Mellon Üniversitesi'nde öğretim üyesi olan Dean Pomerleau ve "Joostware AI Research Corporation"ın kurucusu Delip Rao tarafından makine öğrenimi ve doğal dil işlemenin sahte haberlerle savaşmakta ne kadar başarılı olabileceğini keşfetmek amacıyla Fake News Challange-1 (FNC-1) yarışması düzenlendi. Bu yarışmada amaç başlıktaki bilginin makale içeriğiyle ilgisinin olup olmamasını tahmin etmekti. Toplam 50 yarışmacı bu etkinliğe katıldı. Başlık ve haber çiftleri ilişkili olduğunu doğru tahmin eden takım 0.25 ağırlık puanı, ilişkili çiftleri kabul etme (agree), kabul etmeme (disagree) ve tartışma (discussion) olarak doğru etiketleyen takım ise 0.75 ağırlık puanı almaktadır. "SOLAT in the SWEN" takımı 82.02 puan, "Athene (UKP Lab)" takımı 81.97 puan ve "UCL Machine Reading" takımı ise 81.72 puan ile sıralamadaki ilk 3 takım olmuştur [17].

Yine 2015 yılında "Massachusetts Institute of Technology"de Vosoughi tarafından yapılan doktora tezinde 7000 adet etiketlenmiş İngilizce tweet verilerini, Lojistik regresyon kullanarak iddia, ifade, soru, öneri, istek ve diğer etiketleri ile sınıflandırmıştır. Sınıflandırmak için, denetimli öğrenme algoritmalarından, Naive Bayes (NB), Decision Tree (DT), Logistic regresyon (Logistic Regression-LR) ve Linear çekirdek ile Support Vector Machine (L-SVM) algoritmalarını kullanmıştır. LR algoritmasının, 0.70 f1-metik değeri ile diğer algoritmalardan daha yüksek f1-metrik değerine sahip olduğu sonucuna varmıştır [18].

Chen ve Chen çalışmalarında, telefonlar hakkında Çince bir forum sitesi olan "mobile01.com" üzerindeki forum mesajlarını spam veya spam değil olarak etiketlemişlerdir. Forum mesajlarını zamana bağlı olarak incelemişler ve spam mesajlarının daha çok çalışma günleri ve saatlerinde, spam olmayan mesajların ise çalışma saatleri dışında paylaşıldığını belirtmişlerdir. Scikit-Learn kütüphanesini kullanarak; LR, L-SVM, RBF Çekirdek ile SVM ve SVMperf algoritmalarını kullanmış ve F1 skorlarını karşılaştırmışlardır. Faklı özellikler kullanarak, 6 tane model oluşturmuşlar ve en yüksek başarıyı, 5. modellerinde 0.61 f1-metrik değeri ile elde etmişlerdir [19].

Ahmed yüksek lisans tez çalışmasında, semantik benzerlik ve n-gram analizi ile sahte haber tespitini araştırmıştır. TF (Term Frequency) ve TF-IDF (Term Frequency- Inverse Document Fequency) yöntemleriyle özellik çıkarma işlemi yapmış ve Olasılıklı Dereceli Azalma (Stochastic Gradient Descent-SGD), SVM, L-SVM, K En Yakın Komşu (KNN), Lojistik Regresyon ve Karar Ağaçları (Decision Tree-DT) olmak üzere 6 farklı makine öğrenme algoritması ile sınıflandırma işlemini yaparak bu algoritmaların sonuçlarını karşılaştırmıştır. TF-IDF kelime temsili yöntemi ile kullandığı LSVM algoritması 0.90 doğruluk değeri ile en başarılı algoritma olmuştur [20].

Bajaj, derin öğrenme kullanarak sahte haberlerin tespiti üzerine yaptığı çalışmada, Lojistik regresyon, ileri beslemeli ağlar, RNN (Recurrent Neural Network), GRU (Gated Recurrent Unit), LSTM (Long-Short Term Memory), BiLSTM (Bidirectional LSTM), Maxpooling ile CNN (Convolutional Neural Network), maxpooling ve attention ile CNN metotlarını kullanarak, 63 bin veri içerisinde gerçek ve sahte haberlerin sırasıyla 0 ve 1 şeklinde ikili sınıflandırma ile sınıflandırmış ve bu algoritmaların sonuçlarının f1-metrik değerlerini karşılaştırmıştır. GRU algoritması ile 0.84 f1-metrik değerine ulaşmıştır [21].

Granik ve Mesyura çalışmalarında, naif bayes sınıflandırıcı kullanarak sahte haber tespiti için bir yöntem uyguladıklarını belirtmişlerdir. Facebook sayfalarından toplanan 2282 gönderiyi "çoğunlukla doğru (mostly true)", "çoğunlukla yanlış (mostly false)", "doğru ve yanlış karışık (mixture of true and false)" ve "tam bilgi yok (no factual content)" etiketi ile kelime torbası kullanarak etiketlenmişlerdir. Yöntemleri ile 0.75 doğruluk değeri elde etmişlerdir [22].

Patel yüksek lisans tez çalışmasında, 2016 yılındaki Amerika Başkanlık seçimlerinde yayınlanan sahte haberleri tespit etmeye çalışmıştır. Bunun için; KNN, SVM ve LSTM olmak üzere 3 farklı makine öğrenme algoritması kullanmış ve sonuçlarını incelemiştir. LSTM algoritmasının ortalama F1-metrik değerinin 0.90 olduğunu belirtmiştir [23].

Tacchini vd. çalışmalarında, Facebook üzerinde paylaşılan sahte haberlerin tespiti için kullanıcıları incelemişler ve bu kullanıcıların beğendiği gönderi ve sayfalara göre paylaştıkları haberlerin sahte olup olmama durumuna karar vermişlerdir. Sahte haber ile gerçek haberleri sınıflandırmak için lojistik regresyon ve Harmonik Kitle-Kaynak Boole Etiketi (Boolean Label Crowd-sourcing) olmak üzere 2 yöntem kullanmışlardır. LR algoritması ile 0.79, HBLCS ile 0.99 doğruluk değeri elde etmişlerdir. [24].

Ågren ve Ågren yüksek lisans tez çalışmasında, Yenilemeli sinir ağlarını kullanarak sahte haber tespiti üzerine çalışmıştır. Haber makalesi verilerini başlığıyla ilgili durumunu "ilişkisiz (unrelated), onaylama (agree), onaylamama (disagree) ve tartışma (discuss)" etiketleri ile Paralel kodlayıcı ve koşullu kodlayıcılar ile LSTM ve GRU metotları ile sınıflandırmıştır. Algoritma başarılarını FNC yarışmasındaki değerlendirme ölçütleri ile değerlendirmişler ve algoritmaları 0.69 puana erişmiştir [25].

Rajendran vd. çalışmalarında başlık ve içerik çiftlerinin alakalı veya alakasız olarak etiketlenmesi üzerine çalışmışlardır. Alakalı olan başlık-içerik çiftlerini de onaylama, onaylamama ve tartışma olarak 3 farklı etikete bölmüşlerdir. Sınıflandırma işlemi için; BiLSTM, LSTM, BiGRU, GRU, RNN, BiGRU ve BiLSTM algoritmalarının birlikte kullanımı ve MLP (Multi-layered Perceptron) algoritmalarını kullanmış ve karşılaştırmışlardır. En yüksek başarıyı, BiLSTM algoritması kullanarak 0.84 f1-metrik değeri ile elde etmişlerdir. [26].

Türkçe dilinde makine öğrenmesi yöntemleri kullanarak sahte haber tespiti çalışmaları çok sınırlı sayıda kalmıştır. Yapılan çalışmalara bakıldığında aynı kişiler tarafından haber metinlerinde sınıflandırma ve konu bazlı etiketleme olmak üzere 2 adet çalışmaya rastlanılmıştır.

Mertoğlu vd., 2018 yılındaki bildiride TF-IDF kullanarak sahte ve gerçek haber makaleleri üzerine bir prototip sunmuşlardır. Bu prototipte otomatik, yarı otomatik ve manuel veri toplayan bir kullanıcı arayüzü geliştirmişler ve spor, politika, şehir efsanesi,

sağlık ve diğer konulardan oluşan 3 farklı veri kümesi oluşturmuşlardır. Kullanıcı arayüzü ile toplam 250 haber toplamışlardır. SVM ve Naive Bayes (NB) sınıflandırıcıları ile terim sıklığı ve n-gram özellikleri ile elde edilen sonuçları, TF, n-gram, argo kullanımı ve noktalama işareti kullanımı özellikleri ile elde edilen sonuçlarla karşılaştırmışlardır. 4 özellik ile kullanılan SVM kullanarak 0.79 f1-metrik değeri ile diğer sonuçlara göre daha yüksek başarı elde etmişlerdir [27].

Mertoğlu vd., 2019 yılındaki bildirilerinde Türkçe sahte haber tespitinde otomatik etiketleme modeli sunmuşlardır. Çalışmalarında evrensel bir veritabanı projesi olan GDELT (Global Data on Events, Languages and Tone) projesi tarafından ve kendi geliştirdikleri kullanıcı arayüzü tarafından haber makaleleri toplamışlardır. Toplanan verilerin 1305 tanesi etiketli veriden oluşmuştur. Kalan 250 etiketsiz veri için ise 7 tane lisansüstü öğrencisi tarafından en az 2 en fazla 4 konu etiketi ile etiketlenmesini sağlamışlardır. TF-IDF kelime torbası yöntemi ile kelimeler vektör olarak temsil edilmiş ve toplam 1550 verinin 1150 tanesi eğitim setinde ve eğitim setinde kullanılmamış olan 400 tanesi ise test setinde kullanılmıştır. Etiket olma olasılığını hesaplamak için NB kullanmışlardır [28].

Yapılan çalışmalar incelendiğinde, literatürde farklı dillerde sahte haber tespiti üzerine çalışmalar bulunmakta fakat Türkçe dilinde yapılan çalışmalar ise SVM ve NB sınıflandırıcısı ile sınırlı kalmıştır. Bu çalışmada farklı olarak denetimli ve denetimsiz öğrenme algoritmaları ile Türkçe dilinde sahte haber tespiti yapılmıştır. Ayrıca veri seti, Twitter platformunda belirlenen konulardan toplanmış özgün bir veri setidir. Sosyal medya platformlarından yayınlanan Türkçe mesajlar üzerinde sahte haber tespiti yapan ilk çalışmadır. Çalışmada kullanılan yöntem, Türkçe dilinde yapılmış makine öğrenmesi ile yapılan sahte haber tespiti çalışmasından daha iyi sonuçlar vermiştir.

3. Veri seti

Çalışmamızda Twitter internet sitesinde paylaşılan sahte haber tweet mesajlarının tahmin edilmesi amaçlanmıştır. Bu nedenle belirlenen 6 konuda tweet mesajları toplanmış ve manuel olarak etiketlenmişlerdir. Daha sonra bu konuların içinden sahte haber sayısı ve tweet mesajı sayısı fazla olan konular seçilmiştir. Sonuç olarak 3 konu veri setine eklenmiştir.

3.1. Tweet mesajlarının toplanması

Twitter sitesinde paylaşılan tweet mesajlarının toplanması için Twitter tarafından "Twitter Search API" uygulama programlama arayüzü sunulmaktadır. Bu arayüz Standart, Premium ve Enterprise olmak üzere 3 farklı sürümde kullanıcılara sunulmaktadır. Ücretsiz olan standart sürümde son 7 gün içinde paylaşılmış olan tweet mesajları çekilebilmekte, ücretli olan diğer premium ve enterprise sürümlerinde ise herhangi bir tarih sınırlaması bulunmamaktadır [29]. Standart API kullanılması denenmiş fakat ücretsiz sunulan bu uygulama programlama arayüzünün (API) sadece son 7 gün içerisindeki tweet mesajları çekebilmesinden dolayı bu uygulama programlama arayüzünün etiketsiz tweet mesajlarını toplamak için kullanılmasından vazgeçilmiştir.

TweetScraper [30] uygulaması ile Twitter platformundaki aranan konudaki tüm mesajlar ".txt" uzantılı dosyalar halinde çekilip bilgisayarda bir klasöre kaydedilebilmektedir. Bu ".txt" uzantılı dosyaların içinde tweet mesajları JSON formatında tutulmaktadır. Bu

uygulama ile Twitter API uygulama programlama arayüzü ile çekilen bilgiler kadar kapsamlı olmasa da herhangi bir zaman aralığı olmadan tweet mesajları çekilebilmektedir. Uygulama verileri toplarken kullanıcının gördüğü Twitter sayfasındaki içerikleri tarayarak oluşturduğu için Twitter API kadar kapsamlı bilgi verememektedir.

3.2. Konu seçimi

Haber makalelerinin doğruluğunu kontrol eden birçok siteler bulunmaktadır. Teyit.org, organizasyonel ve editoryal süreçlerinin Uluslararası Doğruluk Kontrolü Ağı'nın (IFCN - International Fact-checking Network) prensipleriyle uyuştuğunu gösterir IFCN sertifikası bulunan haber doğrulama sitesidir [31]. Tweet mesajlarının çekilmesinde konu seçimi için teyit.org internet sitesinde bulunan sahte haberlerden sayıca fazla tweet bulunan konular seçilmiştir.

Tüm hastanelere 4440911 telefon numarasından ulaşılabileceği iddiası, Bayburt Havalimanı'nın yolcu garantili olarak yaptırılması iddiası, hamile kadına saldıran baklavacıların dağıttığı ücretsiz baklavaları almak isteyenlerin kuyruk oluşturduğu iddiası, Falcao'nun Galatasaray takımına geldiği iddiası, İstanbul Havalimanı'nda hava taksi yolunun çöktüğü iddiası ve Kuleli Askerî Lisesi'nin satıldığı iddiasından oluşan 76738 tane tweet mesajı toplanmıştır.

3.3. Veritabanının oluşturulması ve tweet mesajlarının kaydedilmesi

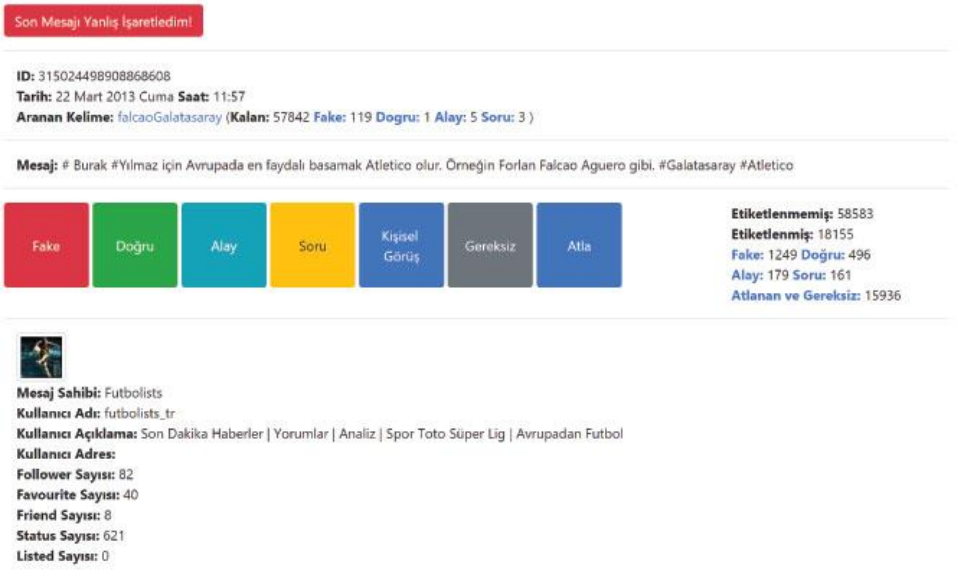
Çekilen tweet mesajlarını saklamak için ve Bölüm 3.4'de anlatılacak olan manuel etiketleme kullanıcı arayüzünün oluşturulması için bir ilişkisel veritabanı oluşturulmuştur. Bu veritabanında Tweet mesajları "Tweet" tablosunda ve bu tweet mesajlarını gönderen kullanıcıların bilgileri ise "User" tablosunda tutulmaktadır.

TweetScraper ile belirlenen 6 konuda toplam 76738 tweet mesajı toplanmıştır. Daha sonra JSON formatında ".txt" uzantılı dosyalar içinde bulunan bu tweet mesajlarını ilişkisel veritabanına aktarmak için C# konsol uygulaması oluşturulmuş ve tüm bu tweet mesajları ilişkisel veritabanında saklanmıştır.

3.4. Verilerin etiketlenmesi

Denetimli öğrenme algoritmaları sınıflandırma işlemini yapabilmeleri için etiketli verilere ihtiyaç duymaktadırlar. Bu nedenle Twitter platformu üzerinden toplanan veriler denetimli öğrenme algoritmalarında kullanılmak üzere etiketli olmaları gerekmektedir.

Manuel etiketlemenin yapılabilmesi için, ASP.NET MVC çatısı kullanılarak C# programlama dili ile bir kullanıcı arayüzü oluşturulmuştur. Toplanan tweet mesaları etiketleyiciye gösterilmiş ve etiketleyiciden "Sahte, Doğru, Alay, Soru, Kişisel Görüş, Gereksiz ve Atla" etiketlerinden birini tercih etmesi istenmiştir (Şekil 5). Aynı zamanda kullanıcı arayüzünde bilgilendirme amaçlı konu bazlı etiketli mesaj sayıları ve tüm etiket sayıları da gösterilmektedir. "Aranan Kelime" kısmında sayısal bilgilerin önündeki Sahte, Doğru, vb. başlıklar tıklandığında o konu ile ilgili etiketlenmiş mesajlar görülebilmektedir. Sağ tarafta bulunan etiketler tıklandığında ise tüm konular için Sahte, Doğru, vb. ile etiketlenen mesajlar görülebilmektedir.



Şekil 5. Manuel etiketleme yapılabilmesi için oluşturulan kullanıcı arayüzü.

Belirlenen konular teyit.org üzerinden ve çeşitli kaynaklardan teyit edildikten sonra, tweet mesajları oluşturulan kullanıcı arayüzü aracılığıyla etiketlenmiştir. Eğer bir tweet mesajı yanlış etiketlenirse veya yanlışlıkla farklı bir etiket seçilirse kullanıcı arayüzünün en üstünde bulunan "Son Mesajı Yanlış İşaretledim" butonu ile son etiketlenen mesajın etiketi iptal edilebilmektedir.

Veritabanında bulunan 76738 adet tweet mesajının 18021 tanesi etiketlenmiştir. Bu etiketlenen tweet mesajlarından 1249 tanesi sahte, 496 tanesi doğru, 179 tanesi alay, 161 tanesi soru, 40 tanesi kişisel görüş ve 15896 tanesi ise gereksiz olarak etiketlenmiştir.

3.5. Etiketlenen verilerin seçilmesi

Tweet mesajları etiketlendikten sonra, etiket bazında en az tweet mesajı sayısı 180 olarak görülmüştür. Her konuda eşit sayıda ve eşit etiketlerde tweet mesajları alınmıştır. Bu nedenle, 180 doğru ve 180 sahte tweet mesajına sahip olan konular seçilmiştir. Daha sonra iki metin arasındaki eklenen silinen karakterleri de dikkate alarak benzerlik puanı oluşturan, Levenshtein Mesafe Algoritması kullanılarak tweet mesajlarının her birinin diğerleri ile benzerlikleri ölçülmüş ve 0.5 değerinin altında benzerlik oranı çıkan tweet mesajları Denetimli öğrenme ve denetimsiz öğrenme modellerinde kullanılmak üzere seçilmiştir.

$$lev_{\mathrm{a,b}}(\mathrm{i,j}) = \begin{cases} max(i,j), & \min(\mathrm{i,j}) = 0 \text{ ise,} \\ min\begin{cases} lev_{\mathrm{a,b}}(\mathrm{i}-1,\mathrm{j}) + 1 \\ lev_{\mathrm{a,b}}(\mathrm{i},\mathrm{j}-1) + 1 \\ lev_{\mathrm{a,b}}(\mathrm{i}-1,\mathrm{j}-1) + 1_{(a_i \neq b_i)} \end{cases}, & \text{diğer durumlarda.} \end{cases} \quad (1)$$

Denklem 1'de a ve b metinlerinin Levenshtein Uzaklığı denklemi verilmiştir. $a_i = b_i$ olduğu zaman, $1_{(a_i \neq b_i)}$ gösterge fonksiyonu 0, diğer durumlarda 1'dir. $lev_{\mathrm{a,b}}(\mathrm{i,j})$, a metninin ilk i karakteri ile b metnin ilk j karakteri arasındaki uzaklığı hesaplar [32].

Levenshtein Uzaklığı uygulanan, benzer tweet mesajları çıkarıldıktan sonra her konuda 90 doğru, 90 sahte tweet mesajı alınmıştır. Bu 3 konudaki tweet mesajları aşağıda açıklanmıştır.

- *Konu-1:* Bayburt Havalimanı'nın T.C. Ulaştırma ve Altyapı Bakanlığı tarafından yıllık 2 milyon yolcu garantili olarak yaptırıldığı iddiasına (Konu-1) ait doğru tweet mesajına örnek olarak "*Bayburt 'un yılda 2 milyon yolcu kapasiteli havalimanı olacakmış ahxnbabxbabd*" tweet mesajı verilebilir. Yanlış tweet mesajına örnek olarak ise "*Bayburt 'a 2 milyon yolcu garantili havalimanı. Kendine güvenen bir ekonomi profesörü bunu bize anlatsın.*" tweet mesajı verilebilir. Doğru ve yanlış tweet mesajı örneklerine bakıldığında "*garantili*" ve "*kapasiteli*" kelimeleri tweet mesajının gerçek veya sahte olduğunu göstermektedir. Bu konudaki tweet mesajları "Bayburt havalimanı" arama cümlesi kullanıldığında Temmuz 2009 ile Eylül 2019 tarihleri arasındaki tweet mesajlarını kapsamaktadır.

- *Konu-2:* Galatasaray futbol kulübünün Radamel Falcao isimli futbolcuyu transfer etti iddiasına (Konu-2) ait doğru tweet mesajına örnek olarak "*Kafayı Falcao ile bozduk doğrudur gelecekse gelsin artık #Galatasaray*" veya "*Bu gece de bekledik gelmedin @FALCAO #ultrAslan #Galatasaray*" tweet mesajları verilebilir. Yanlış tweet mesajına örnek olarak ise "*@FALCAO Galatasaray camiamıza hayırlı olsun.*" veya "*Ve Galatasaray Falcao ile anlaştığını Borsaya bildirdi!*" gibi tweet mesajları verilebilir. İlk konudan farklı olarak bu konudaki tweet mesajlarında sahte tweet mesajları ile gerçek tweet mesajları tamamen farklı kelimelerden oluşabilmektedir. Bu konudaki tweet mesajları "*Falcao Galatasaray*" arama cümlesi kullanıldığında Nisan 2011 ile Ağustos 2019 tarihleri arasındaki tweet mesajlarını kapsamaktadır. Bu tarihlerde Galatasaray futbol takımı tarafından Radamel Falcao isimli futbolcunun transferi gerçekleşmemiş fakat daha sonraki tarihlerde bu transfer gerçekleşmiştir.

- *Konu-3:* Kuleli Askerî Lisesi'nin T.C. Kültür ve Turizm Bakanlığı tarafından satıldığı iddiasına (Konu-3) ait doğru tweet mesajlarına örnek olarak "*Kültür ve Turizm Bakanı Mehmet Nuri Ersoy, Kuleli Askeri Lisesi'nin satıldığına ilişkin sosyal medya ve basına yansıyan haberlerin tamamen asılsız olduğunu belirtti.*" tweet mesajı verilebilir. Yanlış tweet mesajlarına örnek olarak ise "*Kuleli askeri lisesi araplara satılmış"* tweet mesajı verilebilir. Bu konudaki tweet mesajları "*kuleli askeri lisesi*" arama cümlesi kullanıldığında Mayıs 2009 ile Ağustos 2019 tarihleri arasındaki tweet mesajlarından oluşmaktadır. Bu konudaki tweet mesajlarında da sahte tweet mesajlarında "satılmış" kelimesi geçerken, gerçek tweet mesajları ise farklı kelimelerden oluşmaktadır.

4. Materyal- metod

4.1. Doğal dil işleme

Doğal Dil İşleme, insanların iletişimde kullandıkları doğal dillerin makinenin anlamasını sağlayan bir bilim dalıdır. İlk örnekleri, doğal dillerin birbiri arasında çevrilmesini konu alan makine çevirisi ile başlamıştır. Doğal Dil İşleme ilk zamanlarda geleneksel yöntemlerle yapılmakta fakat ilerleyen zamanlarda makine öğrenmesi kullanarak etkili sonuçlar vermeye başlamıştır. Makine öğrenmesinin kullanılması ile Doğal Dil İşleme tanınır hale gelmiş ve yapılan çalışmalar artmıştır. Bu çalışmada da makine öğrenmesi kullanarak doğal dil işleme yöntemleri ile sahte ve gerçek haber tespiti yapılmıştır.

Türkiye'de doğal dil işleme çalışmaları 1973'de kural tabanlı çalışmalarla başlamış olup, 1990 yılından itibaren ise istatistiksel çalışmalarla birlikte makine öğrenmesi çalışmaları başlamıştır. Türkçe de sahte haber tespiti üzerine sadece 2 çalışmaya rastlanmıştır.

4.2. Özellik çıkarımı

4.2.1. TF-IDF

Makine öğrenmesi ile yapılan Doğal Dil İşleme çalışmalarında, matematiksel işlemlerin yapılabilmesi için metinsel ifadelerin vektörel ifadelere dönüştürülmesi gerekmektedir. Bu yöntemlere vektör uzayı modeli adı verilir. Bu yöntemlerden biri olan TF-IDF doğal dil işlemede sıkça kullanılmaktadır. TF-IDF, makine öğrenmesine girdi olarak verilecek olan verilerde tahmin yapmak için bir kelimenin ne kadar önemli olduğunun değerini bulan istatistiksel yöntemdir. Bu yöntemde önce terim sıklığı değeri hesaplanır. Terim sıklığında, her kelimenin her bir dokümanda kaç kere geçtiği hesaplanır. Daha sonrasında ise Ters doküman sıklığı hesaplanarak, terim sıklığı ile ters doküman sıklığı değeri çarpılarak her kelime için TF-IDF skoru belirlenir.

Bu çalışmada vektör uzay modeli, tweet mesajlarını temsil etmek için kullanılmıştır. Tweet mesajları küçük harfe dönüştürülerek python dilinin NLTK kütüphanesinin "*Word_tokenize*" metodu ile kelimelere ayrılmıştır [33]. Yine NLTK kütüphanesinin "*stopwords*" metodunda bulunan Türkçe dilindeki durak kelimeler kaldırılmış ve harflerden ve alfa karakterlerden oluşan kelimeler kullanılmıştır. Her bir kelimeye benzersiz bir numara verilmiştir. Daha sonra "sklearn" kütüphanesinin "*TfidfVectorizer*" metodu kullanılarak her bir kelimenin TF-IDF skoru hesaplanmıştır. Her bir d dokümanında (bu çalışmada, her bir tweet mesajında) geçen t teriminin TF skorunu hesaplamak için Denklem 2 kullanılır.

$f(t,d) = d$ dokümanındaki t teriminin sıklığı

$$\mathrm{TF(t,d)} = \frac{\mathrm{f(t,d)}}{\mathrm{n_d}} \tag{2}$$

Burada $\mathrm{n_d}$, d doküman dizisinin boyutunu (bu çalışmada tweet mesajlarının sayısını) temsil etmektedir. TF skorunun hesaplanmasından sonra ise ters doküman sıklığı değeri hesaplanır. Ters doküman sıklığında ise, bir kelimenin kaç farklı dokümanda geçtiği kontrol edilmektedir. Tersine doküman sıklığını hesaplamak için Denklem 3 kullanılır.

$$\mathrm{IDF(t)} = \log\left(\frac{1+\mathrm{D}}{\mathrm{D_t}}\right) \tag{3}$$

Denklemde D doküman dizisini (bu çalışmada, tweet mesajlarının dizisini), D_t ise t teriminin geçtiği doküman dizisini temsil etmektedir. Denklemde bulunan log fonksiyonu ise sönümleme fonksiyonudur (dampening function). d dokümanındaki her bir t teriminin TF-IDF skoru Denklem 4 ile hesaplanır;

$$\mathrm{TF-IDF(t,d)} = \mathrm{TF(t,d)} \times \mathrm{IDF(t)} \tag{4}$$

4.2.2. Word2Vec

Word2vec kelime temsili modeli bir kütüphanedeki kelimeleri girdi olarak alarak sinir ağı ile eğitip, kelimeler arasındaki benzerlikleri istenen boyutta bir vektör ile temsil eden

yöntemdir. Skip-gram ve Cbow olmak üzere iki yöntem ile sinir ağı eğitilmektedir. Skip gram yönteminde bir kelime modele girdi olarak alınır ve girdi olarak modele giren kelimenin öncesi ve sonrasındaki n adet kelime tahmin edilir. Cbow yönteminde ise girdi olarak bir kelimenin öncesi ve sonrasındaki kelimeler alınır ve model tarafından o kelime tahmin edilir [34].

Word2Vec yöntemi tarafından her kelime için belirlenen boyutta bir vektör oluşturulmaktadır. Bir tweet mesajı, mesajda bulunan her kelimeye karşılık gelen kelime vektörü ile eşleştirilerek, tweet mesajı ve kelime vektörlerinden oluşan bir matrise dönüştürülür. Algoritmalara verilen matris boyutlarının eşit olması gerekmektedir. Bu nedenle bir tweet mesajındaki kelime sayısı 50 olarak belirlenmiştir. 50 kelimeden az olan tweet mesajlarını 50 kelimeye tamamlamak için sonuna vektör boyutuyla aynı boyutta olan sıfır vektörleri eklenmiştir. 50 kelimeden fazla olan tweet mesajlarının ise ilk 50 kelimesi dikkate alınmıştır.

4.2.3. Doc2Vec

Doc2Vec modeli ise Word2Vec modeline benzer şekilde kelime temsili için bir vektör hesaplamaktadır. Word2Vec modelinde her bir kelime için belirlenen boyutta vektör oluşturulurken, Doc2vec yönteminde her bir doküman (bu çalışmada her bir tweet mesajı) için belirlenen boyutta bir vektör oluşturulur [35]. Word2vec modelinin Cbow yöntemine benzer şekilde dokümanda bulunan kelimeleri girdi olarak alır ve her doküman için belirlenen boyutta dokümanı temsil edecek bir vektörü çıktı olarak vermektedir.

4.3. Makine Öğrenmesi

Makine öğrenmesi, sensörler ve algılayıcılardan toplanan veya veritabanında biriken çok sayıda verinin sınıflandırılması, kümelenmesi veya tahminini olanaklı kılan algoritmaların tasarım ve geliştirme süreçlerini konu alan bir bilim dalıdır. Makine öğrenimi ile bilgisayarlara karmaşık örüntüleri tanıma ve karar verme becerisi kazandırılması amaçlanmaktadır. İstatistik, olasılık, veri madenciliği, örüntü tanıma ve yapay zeka gibi alanlarla yakından ilişkilidir [10].

4.3.1. Denetimli Makine Öğrenmesi

Denetimli öğrenmede veriler giriş olarak ve bu verilerin etiketleri de çıktı olarak makine öğrenmesi algoritmasına verilir. Algoritma tarafından eldeki verilerden bir fonksiyon öğrenilerek yeni verilerin hangi çıktıya sahip olacağı tahmin edilir [36].

Bu çalışmada K En Yakın Komşu, Destek Vektör Makinaları ve Rassal Orman denetimli öğrenme algoritmaları kullanılmıştır.

- *K En Yakın Komşu algoritması:* KNN algoritması, dışarıdan bir k hiper parametresi almaktadır. Bu k parametresi sınıflandırılacak olan bir veriye en yakın k adet komşuyu belirtir. Sınıflandırılacak olan veriye en yakın olan k adet elemanla uzaklık ölçülür. Bu uzaklıklar öklid, manhattan, minkowski, hamming ve kosinüs uzaklık metotları gibi farklı metotlar ile ölçülebilmektedir. Uzaklık hesaplama metotlarının sonucuna göre, sınıflandırılacak olan yeni veri hangi sınıfa daha yakın ise o sınıfa dahil edilir [37].

  Bu çalışmada sklearn kütüphanesinin "*KNeighborsClassifier*" metodu kullanılmıştır. Bu metoda k hiper parametresi $k=2$ olarak verilmiştir. Uzaklık hesaplama metodu algoritmaya parametre olarak girilmemiştir. Bu değer girilmediğinde

"*KNeighborsClassifier*" metodu varsayılan olarak "*Minkowski*" metodunu kullanmaktadır.

- *Destek Vektör Makineleri algoritması:* SVM algoritması, eğitim verisinde bulunan vektörleri farklı sınıflara ayırmak için en optimal olan hiper düzlemleri belirler. Lineer olmayan düzlemleri belirlemek için 3 farklı öğrenme makinesi çekirdek fonksiyonu belirlenmiştir. Bunlar; Polinom Öğrenme Makineleri, Radyal Temelli Fonksiyon ve İki Katmanlı Sinir Ağlarıdır [38].

  Bu çalışmada, "*sklearn*" kütüphanesinin "*SVC*" metodu kullanılmıştır. Lineer olmayan düzlem belirlemek için çekirdek fonksiyonu olarak, bu metotta varsayılan olarak kullanılan Radyal Temelli Fonksiyon (Radial Basis Function-RBF) ayarlanmıştır.

- *Rassal Orman algoritması:* Bir karar ağacı algoritması olan Rassal Orman algoritmasında, algoritma farklı alt setler seçerek farklı karar ağaçları oluşturur. Her oluşan karar ağacı tahminde bulunur. Daha sonra sınıflandırmada bu tahminler için en yüksek değer olan seçilir. Birden çok alt set oluşturulmasından dolayı karar ağaçlarının aşırı öğrenme problemi bu algoritmada azalmaktadır [39].

  Bu çalışmada, "*sklearn*" kütüphanesinin "*RandomForestRegressor*" metodu kullanılmıştır. Ormandaki ağaç sayısını belirten hiperparametre olan "*n_estimators*" parametresi 1000 olarak ayarlanmıştır.

4.3.2. Denetimsiz makine öğrenmesi

Denetimsiz öğrenme algoritmalarında veriler algoritmaya girdi olarak verilir. Ayrıca algoritmanın bu verileri kaç kümeye ayıracağı da bildirilir. Algoritma tarafından verilerin ilişkileri ve yapıları öğrenilerek en anlamlı şekilde istenen sayıda kümeye ayrılır [30].

Bu çalışmada K-ortalamalar, Negatif Olmayan Matris Çarpımı ve Doğrusal Diskriminant Analizi denetimsiz öğrenme algoritmaları kullanılmıştır.

- *K-ortalamalar algoritması:* K-means algoritması, dışarıdan hiper parametre olarak küme sayısını almaktadır. Verile küme sayısı kadar merkez nokta rasgele olarak atanarak k adet küme oluşturmaya çalışır. Rasgele atanan noktalar kümelemede zayıflıklar oluşturabilmektedir. Bu problemin önüne geçebilmek ve daha iyi merkez nokta seçilebilmeyi sağlamak için K-means++ algoritması geliştirilmiştir [40].

  Bu çalışmada, "*sklearn*" kütüphanesinin "*KMeans*" metodu kullanılmıştır. Küme sayısını belirten "*n_clusters*" parametresi sahte ve sahte olmayan etiketlerini temsil etmek üzere 2 olarak ayarlanmıştır. Metot varsayılan olarak "*k-means++*" yöntemini kullanmaktadır.

- *Negatif Olmayan Matris Çarpımı algoritması:* Negatif Olmayan Matris Çarpımı, k-ortalamalar kümeleme algoritması ve temel bileşenler analizi yöntemlerine alternatif olarak önerilmiştir [41]. Bu yöntemde amaç, verilen negatif olmayan bir matrise iki negatif olmayan matris çarpımı cinsinden yaklaşmaktır. Birçok çalışmada boyut azaltma algoritması olarak kullanılmış olsa da doküman kümeleme için kullanıldığı çalışmalarda mevcuttur [42], [43].

Bu çalışmada, "*sklearn*" kütüphanesinin NMF metodu kullanılmıştır. Küme sayısını belirten "*n_components*" parametresi Sahte ve Sahte olmayan etiketlerini temsil etmek üzere 2 olarak ayarlanmıştır. TF-IDF, Word2vec ve Doc2vec kelime temsili yöntemleri ile algoritma eğitilmiştir.

Word2Vec ve Doc2Vec yöntemlerinde elde edilen vektörler pozitif ve negatif sayılardan oluşabilmektedir. NMF yöntemi ise sadece negatif olmayan sayıları girdi olarak kabul eder. Bu nedenle Word2Vec ve Doc2Vec tarafından oluşturulmuş vektörlerdeki en küçük negatif sayılar bulunup bu sayıların mutlak değeri, vektördeki tüm sayılara eklenmiştir. Bu sayede kelime vektörü, negatif olmayan sayılardan oluşmuştur.

- *Doğrusal Diskriminant Analizi algoritması:* Doğrusal Diskriminant Analizi, verileri benzerliklerine göre kümelemek için bir hiperdüzlem belirler [44].

  Bu çalışmada, "*sklearn*" kütüphanesinin "*LinearDiscriminantAnalysis*" metodu kullanılmıştır. Küme sayısını belirten "*n_components*" parametresi Sahte ve Sahte olmayan etiketlerini temsil etmek üzere 2 olarak ayarlanmıştır. TF-IDF, Word2vec ve Doc2vec kelime temsili yöntemleri ile algoritma eğitilmiştir.

4.3.3. Makine öğrenmesi algoritmalarını değerlendirme

Makine öğrenmesi algoritmalarının başarısının ölçülmesi için Hata matrisi (confussion matrix) kullanılmaktadır. Tablo 1'de hata matrisinin yapısı gösterilmiştir. Bu matriste kaç adet doğru tahmin yapıldığı, kaç adet yanlış tahmin yapıldığı görülebilmektedir. Çizelgede bu çalışma dikkate alındığında; *S* sahte haberleri, *G* gerçek haberleri, *S'* sahte haber tahminini, *G'* ise gerçek haber tahmini göstermektedir. *P* sahte haberlerin toplam sayısını, *N* gerçek haberlerin toplam sayısını, *P'* tahmin edilen sahte haberlerin toplam sayısını ve *N'* ise tahmin edilen gerçek haberlerin toplam sayısını göstermektedir. *TP*, doğru tahmin edilmiş sahte haber sayısını temsil etmektedir. *FP* ise, sahte haber olarak tahmin edilmiş fakat gerçek haber olması gereken verilerin sayısını temsil etmektedir. *TN*, doğru tahmin edilmiş gerçek haber sayısını temsil etmektedir. Son olarak *FN* ise, gerçek olarak tahmin edilmiş fakat aslında sahte haber olan verilerin sayısını temsil etmektedir.

Tablo 1. Hata Matrisi

	S'	*G'*	
S	*TP*	*FN*	*P*
G	*FP*	*TN*	*N*
	P'	*N'*	

Kesinlik (precision) değeri, algoritmanın tahmin ettiği verilerden ne kadarının ilk kümede doğru tahmin edildiğini gösterir. Sahte ve gerçek haber tespiti problemi dikkate alındığında, veri setinde bulunan sahte haberlerin kaç tanesinin sahte haber olarak seçildiğinin oranını vermektedir. Kesinlik değeri ne kadar yüksek ise ilk kümede doğru etiketlediği eleman sayısı o kadar fazladır. Denklem 5'de Kesinlik metrik değerinin denklemi verilmiştir.

$$Kesinlik = \frac{TP}{TP + FP} = \frac{TP}{P'} \quad (5)$$

Duyarlılık (recall) değeri ise algoritmanın 1. kümede etiketlediği verilerin ne kadarının doğru etiketlendiği oranını verir. Yine sahte haber tespiti problemi düşünüldüğünde; algoritmanın sahte haber olarak işaretlediği verilerin tüm sahte haberlere oranını vermektedir. Denklem 6'da Duyarlılık metrik değerinin denklemi verilmiştir.

$$Duyarlılık = \frac{TP}{TP + FN} = \frac{TP}{P} \quad (6)$$

Kesinlik ve Duyarlılık değerlerinin harmonik ortalaması F1 metrik değeri vermektedir. Denklem 7'de f1-metrik değerinin denklemi verilmiştir.

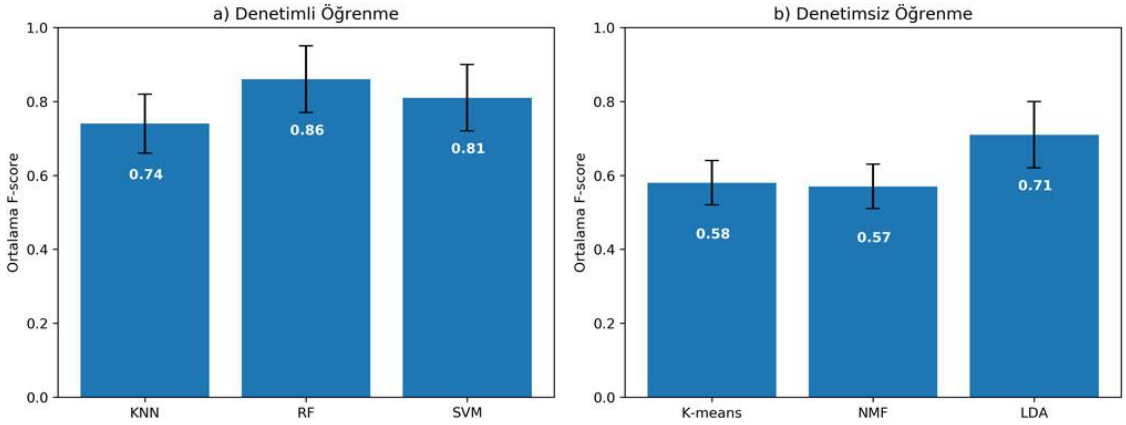
$$F1 - metrik = 2x \frac{Kesinlik\ x\ Duyarlılık}{Kesinlik + Duyarlılık} \quad (7)$$

5. Bulgular

Çalışmada Twitter üzerinden toplanan ve etiketlenmiş 3 konu ile ilgili tweet mesajları, TF-IDF özellik çıkarımı kullanılarak denetimli öğrenme ve denetimsiz öğrenme algoritmaları ile sahte ve gerçek olarak tahmin edilmeye çalışılmıştır. Algoritmaların ön yargısını kaldırmak için her bir algoritma 100 defa çalıştırılmış, her adımda konulardaki tweet mesajları rasgele karıştırılmış ve tüm mesajların hem eğitim setinde hem de test setinde bulunabilmesi sağlanmış ve F1 metrik değerlerinin ortalamaları ile F1 metrik değerlerinin standart sapmaları incelenmiştir.

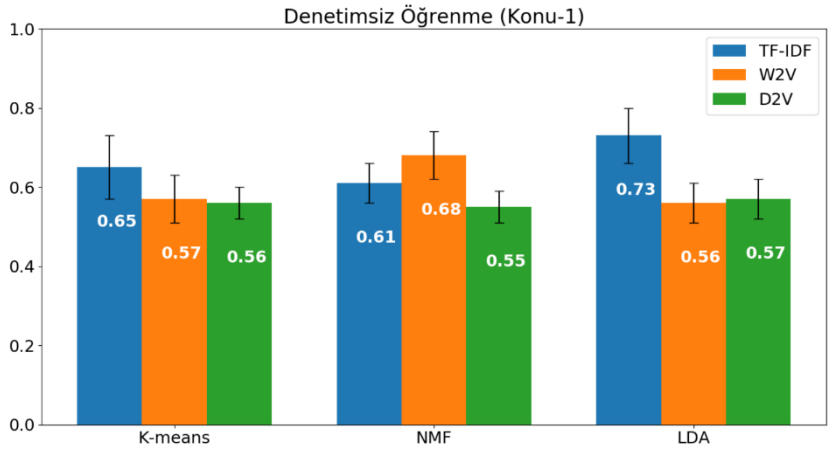
Yapılan tahminler sonucunda denetimli öğrenme algoritmalarının F1 metrik değerlerinin, denetimsiz öğrenme algoritmalarının F1 metrik değerlerine göre daha yüksek olduğu görülmüştür. Denetimsiz öğrenme algoritmalarında sahte haber tespiti neredeyse şans eseri tahmin edilebilirken, denetimli öğrenme algoritmalarında tahmin etme oranı oldukça yüksek çıkmıştır.

Şekil 6.a'da denetimli öğrenme algoritmalarının F1 metrik ortalamaları ve standart sapmaları hata çubukları grafiğinde gösterilmiştir. Şekil 6.b'de ise denetimsiz öğrenme algoritmalarının F1 metrik ortalamaları ve standart sapmaları hata çubukları grafiğinde gösterilmiştir. Denetimsiz öğrenme algoritmaları etiketli veri setine ihtiyaç duymadan var olan verileri benzerliklerine göre kümeler. Bu nedenle denetimsiz öğrenme algoritmalarının, 3 farklı konuyu kümeleme yaparken tüm tweet mesajları verildiğinde, bu tweet mesajlarını Konu-1 ve Konu-3 aynı kümede, Konu-2 ise diğer kümede olacak şekilde ayırdığı görülmüştür. Denetimsiz öğrenme algoritmaları 3 farklı konuyu 2 farklı kümede temsil etmeye çalışmıştır.

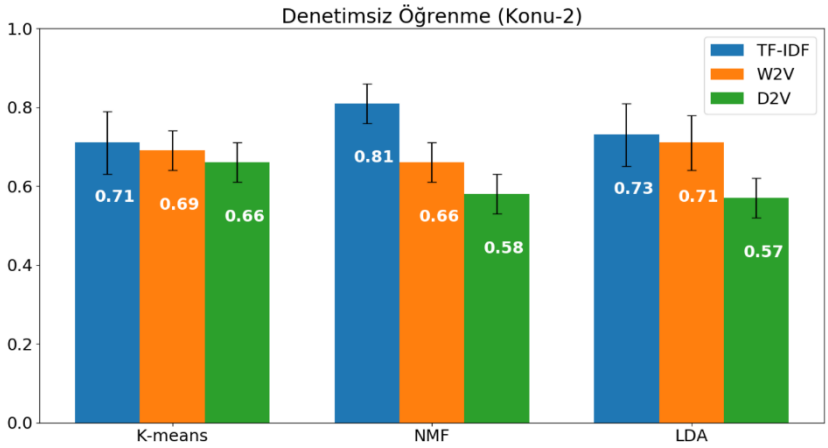


Şekil 6. a) Denetimli öğrenme algoritmalarının F1 metrik ortalamalarını ve standart sapmalarını gösteren hata çubukları grafiği. b) Denetimsiz öğrenme algoritmalarının F1 metrik ortalamalarını ve standart sapmalarını gösteren hata çubukları grafiği.

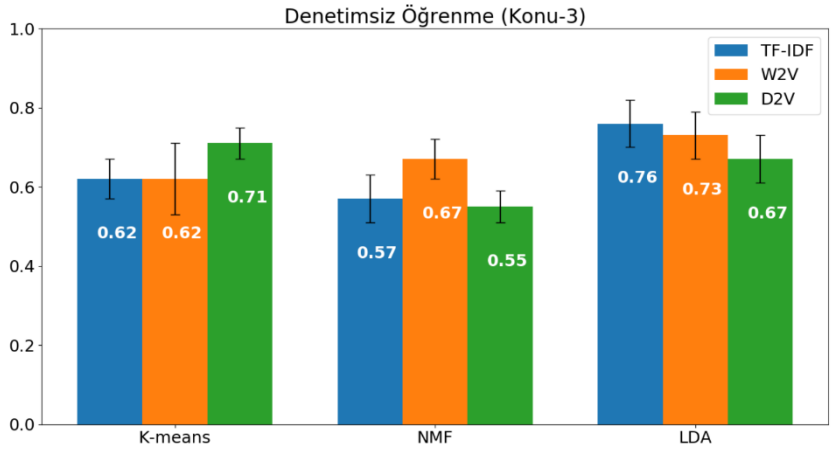
Denetimsiz öğrenme algoritmalarının F1 metrik değerlerini yükseltebilmek için her konu algoritmaya ayrı ayrı verilmiştir. TF-IDF kelime torbası yöntemine ek olarak, Word2Vec ve Doc2Vec kelime gömmeleri yöntemleri de kelime temsili için kullanılmıştır. Algoritmaların ön yargısını kaldırmak için her algoritma 100 defa çalıştırılmıştır. Her adımda konulardaki tweet mesajları rasgele karıştırılmış ve tüm mesajların hem eğitim setinde hem de test setinde bulunabilmesi sağlanmıştır. Bu her adımda karıştırılan veri setinin, %70'i modeli eğitmek için eğitim seti, %30'u ise bu algoritmaların başarısını ölçmek için test seti olarak kullanılmıştır. Her algoritma için 100 adet F1 metrik değeri hesaplanmış ve bu F1 metrik değerlerinin ortalaması alınmıştır (Şekil 7, 8, 9).



Şekil 7. Konu-1 ortalama F1 metrik değerleri.



Şekil 8. Konu-2 ortalama F1 metrik değerleri.

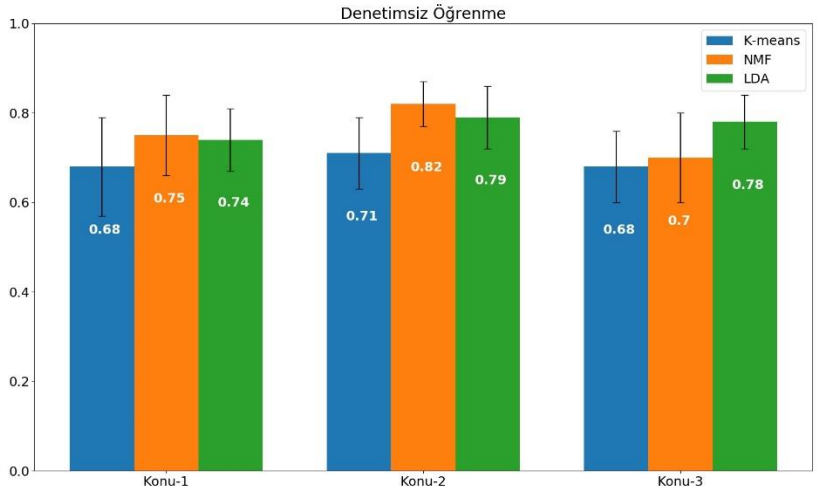


Şekil 9. Konu-3 ortalama F1 metrik değerleri.

Denetimsiz öğrenme algoritmalarından K-means algoritmasının kümelediği verilere bakıldığında; Konu-1 için, "*80 bin nüfusu ile Türkiye'nin en az nüfuslu şehri olan Bayburt 'a, 285 milyon liraya, senelik 2 milyon yolcu kapasiteli havaalanı yapılacakmış. 80 bin nüfuslu şehre, 2 milyon yolcu kapasiteli havaalanı, 285 milyon liraya!"* gerçek haber tweet mesajıdır ve algoritma tarafından da gerçek olarak etiketlenmiştir. "*Bayburt 80 bin nüfusu ile Türkiye'nin en az nüfuslu ili. Bu ile 2 milyon kapasiteli havaalanı yapılacak ve havalanı için yurt dışından borç alınacak. Malum müteahhite 20 yıl kar etme garantisi verilecek. Sonra ekonomi çökünce ABD batırdı, oyunlar oynanıyor.*" tweet mesajı ise sahte haber tweet mesajı olduğu halde gerçek olarak kümelenmiştir. Bu örnekten görüleceği gibi pek çok ortak kelime bulunan iki farklı kümede bulunan tweet mesajları algoritma tarafından aynı kümeye dahil edilmiştir. Ayrıca 54 adet test verisinden 5 tanesi sahte 49 tanesi gerçek olarak tahmin edilmiş, algoritma verilerin çoğunu aynı kümeye dahil etmiştir.

Denetimsiz öğrenme algoritmalarının başarılarını arttırmak için hem sahte haber hem de gerçek haber tweet mesajlarında geçen kelimelerden birkaç tanesi çıkarılarak sonuçlar incelenmiş ve ortalama F1 metrik değerleri biraz daha yükseltilebilmiştir. Word2Vec ve Doc2Vec kelime temsili modellerinde her kelimenin önceki ve sonrasındaki kelimeler de dikkate alındığı için kelime çıkarımı işlemi sadece TF-IDF kelime torbası modeli için yapılmıştır. Kelime çıkarımının ardından TF-IDF kelime torbası modeli ile Denetimsiz

öğrenme algoritmaları 100 defa çalıştırılarak ortalama F1 metrik değerleri hesaplanmıştır. Konu-1 tweet mesajlarından "havaalanı", "havalimanı" ve "nüfus" kelimeleri, Konu-2 tweet mesajlarından "*Falcao*" ve "*Galatasaray*" kelimeleri, Konu-3 tweet mesajlarından ise "*Haydarpaşa*", "*Sirkeci*", "*Suudi*" ve "*araplara*" kelimeleri çıkarılmış ve ortalama F1 metrik değerleri Şekil 10'da gösterilmiştir.



Şekil 10. Konulara göre kelimeler çıkartıldıktan sonra denetimsiz öğrenme algoritmalarının hesapladığı F1 metrik değerleri.

Kelimeler çıkarıldıktan sonra, Konu-2 tweet mesajları ile oluşan veri setinde, 0.82 ortalama f1-metrik değeri ile en başarılı sonucu veren NMF algoritmasının sonuçlarına bakıldığında, veri setinin %30'unu temsil eden 54 test verisinden, 19 tanesini doğru haber kümesine, 35 tanesini ise sahte haber kümesine dahil ettiği görülmüştür. Tahmin edilen doğru haber kümesine bakıldığında ise sadece 2 tweet mesajını doğru kümeleyebilmiştir.

6. Sonuçlar ve tartışma

Yapılan çalışmalara bakıldığında, Türkiye'de yaşayan insanların sahte habere maruz kalma oranı oldukça yüksek ve sahte haberi ayırt edenlerin oranı da düşük olduğu görülmektedir. Ayrıca sahte haberlerin ilk 2 saatlik dilimde çok fazla paylaşıldığı düşünüldüğünde sahte haberlerin tespitinde otomatik tespit sistemlerinin kullanılması büyük bir öneme sahiptir.

Denetimsiz öğrenme algoritmaları etiketli veri setine ihtiyaç duymadan kümeleme yapar. Bu çalışmada 3 farklı konu kullanılmış ve Denetimsiz öğrenme algoritmaları ile her bir konuda iki küme (gerçek, sahte) bulması istenmiştir. Her bir konudaki veriler %70 eğitim seti, %30 test seti olarak ayrılmıştır. Gerçek sahte haber kümeleri eğitim seti kullanılarak 3 farklı Denetimsiz öğrenme algoritması ile bulunmuştur. Bulunan kümeler kümeleme saflığı yöntemi kullanılarak gerçek veya sahte haber oldukları belirlenmiştir.

Buna göre ise, LDA algoritması Konu-1 ve Konu-3 tweet mesajlarında TF-IDF özellik seçimi yöntemi ile, K-ortalamalar ve Negatif Olmayan Matris Çarpımına göre daha yüksek F1 metrik değeri vermiştir. LDA algoritması Konu-2 tweet mesajlarında ise sahte tweet mesajları ile gerçek tweet mesajları farklı kelimelerden oluştuğundan dolayı Konu-1 ve Konu-3 tweet mesajlarında eğitilen modellere göre daha yüksek F1 metrik değeri

vermiştir. Kelimelerin tamamen farklı olduğu Konu-2 tweet mesajlarında NMF algoritması TF-IDF kelime torbası yöntemi ile diğer modellerden daha yüksek F1 metrik değeri verdiği görülmüştür.

Denetimsiz öğrenme algoritmalarının sonuçları incelendiğinde; kelime çıkarımı yapılmadan önce tahmin edilen verilerin bir küme içinde yoğunlaştığı görülmüştür. İki grup içerisinde ortak kelimelerin çıkarımı yapıldıktan sonra ise bir küme içerisindeki yoğunlaşma azalmıştır. Tweet mesajları iki kümeye daha orantılı yayılmıştır. Ancak buna rağmen doğru kümeleme oranında istenilen seviyede F1 metrik değerinde bir artış olmamıştır.

Denetimli öğrenme algoritmalarında KNN algoritması; Rassal Orman ve SVM algoritmasına göre daha düşük F1 metrik değerine ulaşmıştır. SVM algoritması; Rassal Orman algoritmasına göre daha yüksek F1 metrik değerine sahip olsa da aralarındaki fark oldukça azdır.

Sonuç olarak etiketli veriler ile sahte haber tespitinde denetimli öğrenme algoritmalarının, denetimsiz öğrenme algoritmalarına göre daha başarılı olduğu görülmüştür. Bunun nedeni; denetimsiz öğrenme algoritmalarının verileri kümelemek için bir etikete ihtiyaç duymamasından kaynaklanmaktadır. Denetimsiz öğrenme algoritmaları veriler arasındaki benzerliklere göre verileri kümelemektedir. Denetimli öğrenme algoritmalarında ise modele etiketli veri verildiği için aynı etiketler arasındaki benzerlikleri çözmekte daha başarılı olmuşlardır.

Yapılan bu çalışmada sahte haberlerin tespiti otomatik bir sistem tarafından yaptırılmış ve çok kısa sürede sahte haberin tespit edilmesine olanak sağlamıştır. Bu sayede kısa sürede çok fazla paylaşım yapılmadan sahte haberin tespiti ve önlenmesi sağlanabilecektir.

Bu çalışmada kelime temsili yöntemi olarak TF-IDF kelime torbası yöntemi, Word2Vec ve Doc2vec yöntemleri kullanılmıştır. TF-IDF kelime torbası modeli, Word2vec ve Doc2vec kelime temsili yöntemlerine göre kümeleme F1 metrik değerleri daha yüksektir. Ayrıca Word2vec ve Doc2vec yöntemlerine göre algoritmanın eğitim süresi daha kısadır.

Algoritmaların F1 metrik değerlerini yükseltebilmek için sosyal medya platformlarında kişileri takip edenler ve kişinin takip ettiği kişilerde incelenerek, hesapların bot hesap, sahte hesap veya propaganda için açılmış hesap olup olmaması ayrımları dikkate alınabilir. Bu nedenle gelecek çalışmalarda arkadaşlık grafı kullanılarak otomatik tespit sistemine bu graf, girdi olarak verilecek ve algoritmanın başarısı arttırılmaya çalışılacaktır.

Kaynaklar

[1] Del Vicario, M. *vd.*, The spreading of misinformation online, **Proceedings of the National Academy of Sciences**, 113, 3, 554–559, (2016).

[2] Twitter, "KAMUOYUNA DUYURU İletişim Başkanlığı, vatandaşlardan hiçbir şekilde kredi kartı bilgilerini talep etmez. Kurumumuzun adı ve logosu ile yayılan “elektrik ve doğal gaz fatura iadesi” bildirimi, dolandırıcıların milletimizin devletimize olan güvenini kötüye, [Tweet]" (2020). https://twitter.com/iletisim/status/1213530046733979649, (04.1.2020).

[3] Twitter, "Yoğun kar yağışı,buzlanma ve soğuk nedeniyle, 07 Ocak 2020 Salı günü, il merkezi dışında kalan resmi ve özel tüm okul ve kurumlarımızda (okul öncesi, ilkokul, ortaokul, lise ve yaygın eğitim kurumları) eğitim öğretime bir gün ara verilmiştir. [Tweet]", (2020). https://twitter.com/eskvalilik/status/1214309576939573248, (07.1.2020).

[4] Ihlas Haber Ajansi, Eskişehir'de sahte hesaptan kar tatili mesajı atıldı, 2020. https://www.iha.com.tr/haber-eskisehirde-sahte-hesaptan-kar-tatili-mesaji-atildi-821170/, (06.1.2020).

[5] Shu, K., Sliva, A., Wang, S., Tang, J. ve Liu, H., Fake News Detection on Social Media, **ACM SIGKDD Explorations Newsletter**, 19, 1, 22–36, (2017).

[6] Newman, N., Fletcher, R., Kalogeropoulos, A. ve Nielsen, R., Reuters Institute Digital News Report 2018, Teknik Rapor, Reuters Institute for the Study of Journalism, Oxford, (2018).

[7] Newman, N., Fletcher, R., Kalogeropoulos, A. ve Nielsen, R., Reuters Institute Digital News Report 2019, Teknik Rapor, Reuters Institute for the Study of Journalism, Oxford, (2019).

[8] Zhao, X. ve Jiang, J., An empirical comparison of topics in twitter and traditional media, **Singapore Management University School of Information Systems Technical paper series**, (2011).

[9] Twitter, Twitter Inc., 2006. https://twitter.com/, (10.10.2018).

[10] Alpaydin, E., **Machine Learning: The New AI**. Cambridge, MA: The MIT Press, (2016).

[11] Rosten, E. ve Drummond, T., Machine Learning for High-Speed Corner Detection, **European Conference on Computer Vision**, Lecture Notes in Computer Science, 430–443, Graz- Austria, (2006).

[12] Arganda-Carreras, I. *vd.*, Trainable Weka Segmentation: a machine learning tool for microscopy pixel classification, **Bioinformatics**, 33, 15, 2424–2426, (2017).

[13] Amodei, D. *vd.*, Deep Speech 2: End-to-End Speech Recognition in English and Mandarin, **Computing Research Repository**,(2015).

[14] Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K. ve Bengio, Y., Attention-Based Models for Speech Recognition, **Computing Research Repository**, (2015).

[15] Pang, B. ve Lee, L., Opinion Mining and Sentiment Analysis, **Foundations and Trends® in Information Retrieval**, 2, 1–2, 1–135, (2008).

[16] Pang, B. ve Lee, L., A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, (2004), doi: 0409058.

[17] Pomerleau, D. ve Rao, D., Fake News Challenge, (2015). http://www.fakenewschallenge.org/, (11.07.2018).

[18] Vosoughi, S., Automatic Detection and Verification of Rumors on Twitter, Yüksek Lisans Tezi, Massachusetts Institute of Technology**,** Cambridge, (2015).

[19] Chen, Y. ve Chen, H., Opinion Spam Detection in Web Forum : A Real Case Study, **Www 2015**, 1, 173–183, (2015).

[20] Ahmed, H., Detecting Opinion Spam and Fake News Using N-gram Analysis and Semantic Similarity, Yüksek Lisans Tezi, University of Ahram Canadian, Kahire, (2017).

[21] Bajaj, S., "The Pope Has a New Baby !" Fake News Detection Using Deep Learning, 1–8, (2017).

[22] Granik, M. ve Mesyura, V., Fake news detection using naive Bayes classifier, **Electrical and Computer Engineering (UKRCON), 2017 IEEE First Ukraine Conference on**, 900–903, (2017).

[23] Patel, M., Detection of Maliciously Authored News Articles, Yüksek Lisans Tezi,

The Cooper Union For The Advancement of Science and Art, New York, (2017).
[24] Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S. ve de Alfaro, L., Some like it Hoax: Automated fake news detection in social networks, **SoGood 2017 - Second Workshop on Data Science for Social Good**, Skopje-Macedonia,(2017).
[25] Ågren, A. ve Ågren, C., Combating Fake News with Stance Detection using Recurrent Neural Networks, Yüksek Lisans Tezi, University of Gothenburg, Gothenburg, (2018).
[26] Rajendran, G., Chitturi, B. ve Poornachandran, P., Stance-In-Depth Deep Neural Approach to Stance Classification, **International Conference on Computational Intelligence and Data Science (ICCIDS 2018***)*, 132, 1646–1653, (2018).
[27] Mertoğlu, U., Sever, H., ve Genç, B., Savunmada Yenilikçi bir Dijital Dönüşüm Alanı: Sahte Haber Tespit Modeli, **SAVTEK 2018 - 9. Savunma Teknolojileri Kongresi**, 771–778, (2018).
[28] Mertoğlu, U., Genç, B., Sever, H. ve Sağlam, F., Auto-Tagging Model For Turkish News, içinde **International Ankara Conference on Scientific Researches**, 615–623, (2019).
[29] Twitter Search API, Twitter Search API, Twitter, (2018). https://developer.twitter.com/en/docs/basics/getting-started, (10.06.2018).
[30] Github, TweetScraper, (2019). https://github.com/jonbakerfish/TweetScraper, (10.06.2018).
[31] Teyit.org, teyit.org, (2016). https://teyit.org/, (01.08.2018).
[32] Levenshtein, V. I., Двоичные коды с исправлением выпадений, вставок и замещений символов (Binary Codes Capable of Correcting Deletions, Insertions, and Reversals), **Доклады Академий Наук СССР**, 163, 4, 845–848, (1965).
[33] Bird, S., Klein, E. ve Loper, E., **Natural language processing with Python: analyzing text with the natural language toolkit**. O'Reilly Media, Inc., (2009).
[34] Mikolov, T., Chen, K., Corrado, G. ve Dean, J., Efficient Estimation of Word Representations in Vector Space, (2013).
[35] Le, Q. V. ve Mikolov, T., Distributed Representations of Sentences and Documents, (2014). http://arxiv.org/abs/1405.4053, (16.03.2019).
[36] Goodfellow, I., Bengio, Y. ve Courville, A., **Deep learning**. Cambridge, MA: The MIT Press, (2017).
[37] Cunningham, P. ve Delany, S. J., k-Nearest Neighbour Classifiers -- 2nd Edition, (2020). http://arxiv.org/abs/2004.04523, (10.11.2019).
[38] Vapnik, V., **The Nature of Statistical Learning Theory**. Springer, (1995).
[39] Breiman, L., Random Forests, **Machine Learning**, Springer, 5–32, (2001).
[40] Arthur, D. ve Vassilvitskii, S., k-means++: The Advantages of Careful Seeding, (2006). http://ilpubs.stanford.edu:8090/778/, (08.11.2019).
[41] Lee, D. D. ve Seung, H. S., Learning the parts of objects by non-negative matrix factorization, **Nature**, 401, 6755, 788–791, (1999).
[42] Shahnaz, F., Berry, M.W., Pauca, V. P. ve Plemmons, R. J., Document clustering using nonnegative matrix factorization, **Information Processing & Management**, 42, 2, 373–386, (2006).
[43] Xu, W., Liu, X. ve Gong, Y., Document clustering based on non-negative matrix factorization, **Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval - SIGIR '03**, 267, (2003).
[44] Fisher, R.A., The Use of Multiple Measurements in Taxonomic Problems, **Annals of Eugenics**, 7, 2, 179–188, (1936).