Fırat Üniversitesi Müh. Bil. Dergisi
33(2), 745-754, 2021

Araştırma Makalesi
https://doi.org/10.35234/fumbd.986500

Çevrimiçi Sosyal Ağlarda Nefret Söylemi Tespiti için Yapay Zeka Temelli Algoritmaların Performans Değerlendirmesi

Cem BAYDOĞAN[1], Bilal ALATAŞ[2*]
[1] Yazılım Mühendisliği Bölümü, Teknoloji Fakültesi, Fırat Üniversitesi, Elazığ, Türkiye
[2] Yazılım Mühendisliği Bölümü, Mühendislik Fakültesi, Fırat Üniversitesi, Elazığ, Türkiye
[1] vcbaydogan@firat.edu.tr, [*2] balatas@firat.edu.tr

(Geliş/Received: 24/08/2021; Kabul/Accepted: 31/08/2021)

Öz: Çevrimiçi sosyal medya araçlarının kullanımının artması Nefret Söylemi (NS) başta olmak üzere birçok sosyal ağ problemini beraberinde getirdi. Sosyal ağlarda hızla yayılan NS içeren yazı, resim, kışkırtıcı karikatür, tweet, post vb. iletiler ifade özgürlüğünün ötesine geçmektedir. Dahası bir olayı, rejimi, etnik kökeni, cinsiyet ayrımcılığını, krizi, gündemi vb. durumları hedef alan ve kontrolsüz bir şekilde yayılan bu içerikler insanlar arasında korku ve endişeye sebep olmaktadır. Bu problemlerin çözümü için çalışmada önerilen NS tespit sisteminin geliştirilmesi son derece kritiktir. Önerilen NS tespit sisteminde, sosyal ağlar üzerinde paylaşılan NS tweetlerin otomatik tespiti için yapay sinir ağları ve makine öğrenmesi yöntemlerinden oluşan yapay zeka temelli algoritmalar kullanıldı. Çalışmanın ilk adımında seçilen veri seti üzerinde temel doğal dil işleme teknikleri uygulandı. Ardından, veri setinin temsili için kelime çantası (BoW), terim frekansı (TF) ve terim doküman matris (t-DM) gibi özellik çıkarım teknikleri gerçekleştirildi. Naive Bayes, Destek Vektör Makinesi, iki farklı Karar Ağacı ve Çok Katmanlı Algılayıcı olmak üzere beş farklı yapay zeka temelli algoritma ile NS tespit sistemi tamamlandı. Önerilen sistemin güvenilirliğini kanıtlamak için farklı eğitim ve test teknikleri kullanılarak performans değerlendirme metrikleri hesaplandı. Farklı test teknikleriyle en yüksek doğruluk değeri Karar Ağaçları ve Çok Katmanlı Algılayıcılar tarafından %80 olarak elde edildi. Önerilen NS tespit sistemine ait diğer tüm deney sonuçları tablo ve grafiklerle ayrıntılı bir şekilde Bölüm 4'de sunulmuştur. Ulaşılan umut verici sonuçlar birçok farklı sosyal ağ problemlerinin çözümü için önerilen otomatik tespit sisteminin kullanılabileceğini göstermektedir.

Anahtar kelimeler: Doğal Dil İşleme, Makine Öğrenmesi, Nefret Söylemi Tespiti, Sosyal Medya, Yapay Zekâ.

Performance Assessment of Artificial Intelligence-Based Algorithms for Hate Speech Detection in Online Social Networks

Abstract: The increase in the use of online social media platforms caused many social network problems, especially Hate Speech (NS). Shares such as contexts, pictures, provocative cartoons, tweets, posts, etc. containing NS, which spread rapidly on social networks, have violated freedom of expression. Moreover, these shares, which target an event, regime, ethnic origin, gender discrimination, crisis, agenda, etc. and spread uncontrollably, have given rise to fear and anxiety among people. To solve these problems, it is extremely critical to develop the proposed NS detection system in the study. In the proposed NS detection system, artificial intelligence-based algorithms consisting of artificial neural networks and machine learning methods were used for automatic detection of NS tweets shared on social networks. In the first step of the study, basic natural language processing techniques were applied on the selected data set. Then, feature extraction techniques such as bag of words (BoW), term frequency (TF) and term-document matrix (t-DM) were performed to represent the data set. The NS detection system was completed with five different artificial intelligence-based algorithms: naive bayes, support vector machine, two different decision trees and multi-layer perceptron. Performance evaluation metrics were calculated using different training and testing techniques to prove the reliability of the proposed system. With different test techniques, the highest accuracy value was obtained by Decision Tree and Multilayer Perceptron's as 80%. All other test results of the proposed NS detection system are presented in Chapter 4 with tables and graphics in detail. The promising results show that the proposed automatic detection system can be used to solve many different social network problems.

Key words: Natural Language Processing, Machine Learning, Hate Speech Detection, Social Media, Artificial Intelligence.

1. Giriş

Gelişen dijital çağla birlikte, insanlar kendilerini internet aracılığıyla özellikle de sosyal medya platformları üzerinden ifade etme konusunda uzman hale getirdiler. Sosyal medyanın hızlı büyümesiyle birlikte, kullanıcılar başta yetişkinler olmak üzere, başkalarıyla bağlantı kurmak, bilgi paylaşmak ve ortak ilgi alanlarını sürdürmek için çeşitli çevrimiçi sosyal ağ sitelerinde önemli miktarda zaman harcamaktadırlar. Sosyal ağların bu derece

[*] Sorumlu yazar: balatas@firat.edu.tr. Yazarların ORCID Numarası: [1] 0000-0002-6125-2442, [2] 0000-0002-3513-0329

yoğun kullanımı beraberinde NS gibi birçok sosyal ağ problemlerini ortaya çıkarmıştır [1]. NS'nin genel bir tanımı olmamakla birlikte, sosyal medya kullanıcılarının bilerek veya bilmeyerek bir bireyi ve/veya toplumun bir kesimini aşağılayıcı olabilecek ileti veya yorum yapabileceği senaryoların tamamında görülen suç teşebbüsü sayılabilecek sosyal medya problemidir [2].

NS'nin demokrasinin temel taşlarından biri olarak kabul edilen ve aynı zamanda en çok tartışılan insan haklarından biri olarak kabul edilen ifade özgürlüğü ile karıştırılmaması gerekmektedir. Sosyal medya kullanımı açısından bakıldığı zaman, ifade ve basın özgürlüğünün mutlak olması gerekir. Ancak sosyal ağlarda şiddete teşvik içerikle paylaşımlar yasa dışı olarak değerlendirilir [3]. Aynı zamanda bu paylaşımların yayılmasının engellenmesi gerekir. Sosyal medya iletileri kavga, küfür, hakaret, aşağılama, ayrımcılık vb. düzeyine yükselmediği veya gerçek bir şiddet tehdidi oluşturmadığı sürece ifade özgürlüğü garantisi ile korunmaktadır [4, 5]. Sosyal medyada NS'ni engellemek için birçok ülke tarafından katı yasalar çıkarmıştır. NS paylaşımlarının ardında daha çok taraftar kazanma, dikkat çekme ve para kazanma gibi çeşitli hilelerinde olduğu söylenebilir. Birçok sosyal medya platformu kullanıcılarına kullanım sözleşmesinde ulusal, ırksal, dini, nefreti ve hoşgörüsüzlüğü alevlendirmeyi, şiddete ve savaşa teşviki içeren NS paylaşımlarının durdurulacağını ve hesaplarının askıya alınacağını garanti eden mektuplar imzalatır. Ayrıca sosyal medyada NS paylaşımlarının yayılmasının önlenmesi için her geçen gün yeni tedbirler alınmaktadır [6].

Sosyal medya platformları genel olarak kendileriyle benzer düşünen veya ortak ilgi alanlarına sahip insanlar arasında hızlı iletişim sağlamalarına imkân tanır [7]. Bununla birlikte, kültürel ve kişisel yatkınlığımız nedeniyle, herhangi birisi tarafından normal ya da eğlenceli olarak algılanabilen paylaşımlar, bir başkası için rahatsız edici olabilir. İlerleyen durumlarda bu paylaşımlar yerini sözlü tacizlere, hakaretlere veya küfürlü dil kullanımına bırakabilir. Buda kullanıcılar arasında uyumsuzluk, rahatsızlık, korku ve paniğe neden olur. Ayrıca bu durum sosyal medya kullanıcılarının zihninde olumsuz bir algı oluşturur ve ileride bu etkinliklere katılımını kısıtlayabilir [8]. Çünkü sosyal medyanın genel kullanımından uzaklaşarak, yararlı bilgiler ararken bu aşağılayıcı imaları bulmak, kullanıcının cesaretini kırabilir ve hayal kırıklığına neden olur.

NS problemini çözmek için akla gelen ilk çözüm, sosyal medya paylaşımlarını insan kontrolünde gözlemleyen yöneticiler belirlemektir [9]. Ancak sosyal medya üzerinden paylaşılan verilerin sayısı göz önüne alındığında bu durumun insan gözetiminde yapılması zaman ve maliyet açısından imkânsızdır. Ayrıca insan gözetiminde yapılsa bile hata payı hiçbir zaman sıfır olmayacaktır. Başka bir yöntem de NS sözlüğü oluşturmak olabilir [10]. Bir ileti veya yorum sözlükten kelimeler içeriyorsa, onu hakaret olarak tespit edecektir. Ancak bu yönteminde kelimelerin esnek kullanımından dolayı etkili olmayacağı aşikârdır. Diğer en etkili alternatif ise çalışmada önerilen, yorumları ve paylaşımları otomatik sınıflandıran bir NS tespit sisteminin oluşturulmasıdır [11]. Bu çalışmanın en büyük amacı sosyal medya platformlarından paylaşılan NS içeriklerinin tespitini yapay zeka temelli algoritmalarla hızlı, etkili ve otomatik bir şekilde tanımlayabilmektir. Böylelikle bir kullanıcı hakaret, küfür, aşağılama, taciz vb. içeren bir ileti paylaşırsa, bu paylaşımın yayınlaması durdurulabilir veya bu tür davranışların tekrarlanması engellenebilir. En önemlisi ise otomatik NS tespit sistemi ile öngörülmeyen hata payı azaltılacak, zaman ve maliyet verimliliği sağlanmış olacaktır.

Bu çalışmada, NS otomatik tespit sistemi geliştirmek için 40.623 sentetik metinden oluşturulan veri seti kullanılmıştır. Dinamik olarak üretilen bu metinler 'hate' ve 'nothate' olmak üzere iki etiketle sınıflandırılmıştır. Bu yöntemle sosyal ağ çalışmalarının en önemli problemlerinden olan biri olan veri sayısı yetersizliğinin önüne geçilmiştir. Ardından noktalama işaretlerinin temizlenmesi, sayısal ifadelerin silinmesi, belirli karakterden az sayıda kelimelerin silinmesi, büyük-küçük harf dönüşümü vb. temel metin ön işleme adımları uygulanarak veri seti NS tespit sistemi için uygun formatta hazırlanmıştır. Bir sonraki adımda veri setini temsil edecek özelliklerin çıkarılması için BoW, TF ve t-DM teknikleri uygulanmıştır. Artık NS bir sınıflandırma problemi haline gelmiştir. Seçilen yapay zeka temelli sınıflandırıcılar kullanılarak 'hate' ve 'nothate' tespiti sağlanmıştır. Sistemin güvenilirliğini garanti etmek için çeşitli eğitim ve test teknikleri kullanılmıştır. Sistemin başarımını ölçmek için performans değerlendirme metrikleri hesaplanmıştır. Elde edilen tüm deney sonuçları tablo ve grafiklerle analiz edilmiştir. Geliştirilen sistemde elde edilen tatminkar çıktılar diğer sosyal ağ problemlerinin çözümü içinde önerilen bu yöntemlerin uygulanmasının yolunu açacağı öngörülmektedir.

Çalışmanın devam eden bölümleri şu şekilde düzenlenmiştir. Çalışmanın ikinci bölümünde NS ile ilgili literatür taraması yapılmıştır. NS tespiti için kullanılan yöntemler elde edilen başarımlar ve eksiklikler analiz edilmiştir. Üçüncü bölümünde çalışmada kullanılan veri setinin karakteristiği, doğal dil işleme teknikleri, özellik çıkarım yöntemleri ve yapay zeka algoritmalarının özellikleri sıralanmıştır. Dördüncü bölümde performans değerlendirmesi için yapılan farklı test teknikleri ve elde edilen çıktılar analiz edilmiştir. Algoritmaların farklı değerlendirme kriterleri karşısındaki başarımları değerlendirilmiştir. Son bölümde bu çalışmanın gelecek

çalışmalar açısından önemi vurgulanmıştır. Ayrıca çalışmanın yaygın etkisi, literatüre katkısı, eksileri ve gelecek çalışmalar için gerekli yönlendirmeler yapılmıştır.

2. İlgili Çalışmalar

Sosyal ağların NS'nin yayılmasındaki rolü göz önüne alındığında literatürdeki araştırmaların sayısının her geçen gün arttığı görünmektedir. NS tespit çalışmalarında literatürde birçok farklı yöntem ve veri setleri kullanılmıştır. Bu bölümde NS ile ilgili çalışmalar özetlenmiştir.

Abro ve arkadaşları yürüttükleri çalışmalarında NS tespiti için makine öğrenmesi yöntemlerini önermişlerdir. Çalışmada araştırmacılar üç farklı sınıf etiketine sahip halka açık bir veri setinde performanslarını değerlendirmek için üç özellik mühendisliği tekniğini ve sekiz makine öğrenme algoritmasını karşılaştırmışlardır. Elde ettikleri deney sonucunda, destek vektörü makine sınıflandırıcısı ve bi-gram özellik mühendisliği tekniği ile %79 oranında doğruluk değeriyle en iyi performansın yakalandığı görülmüştür [12]. Yine bu çalışmada %77 hassasiyet değeri bi-gram ve destek vektörü makine sınıflandırıcısı tarafından elde edilmiştir. Daha yüksek başarım elde edebilmeleri için gelecek çalışmalarında farklı özellik mühendisliği teknikleri ve yapay zeka algoritmaları kullanacaklarını ayrıca vurgulamışlardır.

Pathak ve ekibi Malayalam ve Tamilce dillerinde NS tespit çalışması yapmışlardır. Toplam 8 bin veriden oluşan veri seti üzerinde farklı makine öğrenmesi algoritmaları ile çalışmışlardır. Twitter verilerinden oluşan bu veri seti 'saldırgan' ve 'saldırgan olmayan' etiketleriyle iki sınıfa ayrılmıştır. TF-IDF ve n-gram özellik çıkarım yöntemlerini kullanan araştırmacılar Malayalam ve Tamilce veri setleri için sırasıyla %77 ve %87 f-skor değerleri elde etmişlerdir. Araştırmacılar yapılan çalışmanın konuşmadan NS tespit sistemi şeklinde geliştirilmesi için teşebbüslerinin olduklarını vurgulanmışlardır [13]. Literatürde farklı diller için makine öğrenmesi algoritmaları kullanılarak geliştirilen çeşitli NS tespit çalışmaları da bulunmaktadır [14-17].

Yapılan incelemeler sonucunda NS tespiti için derin öğrenme yaklaşımları ile yapılan çalışmalarda literatürde yer almaktadır [18-21]. Tripathy ve arkadaşları Derin Evrişimli Sinir Ağı (DCNN) kullanılarak otomatik bir NS tespit sistemi geliştirilmiştir. Önerilen DCNN modeli sırasıyla %97 hassasiyet, %88 duyarlılık ve %92 f-skoru performansı gösterdi. Araştırmacılar özellik çıkarım aşamasında yaklaşık 32 bin tweetten oluşan veri setinde GLoVe yönteminden faydalanmışlardır [22]. Yalnızca İngilizce dilindeki tweetler üzerinde önerilen sistemin ilerleyen aşamalarında farklı diller içinde hizmet verebilmesini hedeflemişlerdir.

Emmanuel-Ayo ve takımı makine öğrenmesi yöntemleri ile geçmişten günümüze NS tespit sistemi çalışmalarını analiz etmişlerdir [23]. Güncel bir araştırma makalesi yayınlayan ekip tekli ve hibrit makine öğrenimi yöntemlerinin artılarını ve eksilerini sunmuşlardır. Araştırılan makine öğrenimi yöntemleri için performans değerlendirmesinin özeti de ayrıca sunulmuştur. Kendilerinin önerdikleri kural tabanlı sınıflandırma yaklaşımı ile %97'lik AUC değerine sahip NS tespit sistemi geliştirmişlerdir. Elde ettikleri deneysel sonuçların istatistiksel doğrulamasını yapan ekip, geliştirilen sistemin verimliliğine dikkat çekmişlerdir. Son olarak bu sonuçlar, geliştirilen NS sisteminin otomatik konu tespiti ve diğer sınıflandırma problemlerinin çözümü için uygun olduğunu da göstermiştir.

Sosyal ağlarda en çok NS paylaşımlarının yayınlandığı bir başka önemli konu göç ve mültecilere karşı saldırgan içeriklerdir. [24] nolu çalışmada araştırmacılar bu konuya dikkat çekmiş ve yaklaşık 8 bin tweet'e sahip veri seti üzerinde çalışmalarını tamamlamışlardır. Makine öğrenmesi algoritmalarını kullanan araştırmacılar Lojistik regresyon algoritması ve n-gram yöntemiyle %84 f-skor değerine ulaşmışlardır. Araştırmacılar farklı diller için NS tespit sistemleri geliştirmeyi amaçlamaktadırlar. Ayrıca model doğruluğunu artırmak için veri setini genişletmeyi ve derin öğrenme yaklaşımlarını kullanmayı hedeflemektedirler.

Garain ve Basu yürüttükleri NS tespit çalışmasında Twitter'da kadınları ve mültecileri konu alan tweetlerden oluşan veri seti üzerinde bir analiz çalışması yapmışlardır [25]. 10 bin tweetten oluşan bu veri setinde Bi-LSTM derin öğrenme yöntemi kullanmışlardır. Araştırmacılar önerdikleri NS tespit sisteminde %82 doğruluk ve %70 f-skor değerlerine ulaşmışlardır. Sağlanan veri kümesi dışında herhangi bir harici veri kullanmaması önerilen NS tespit sisteminin en büyük dezavantajıdır. Gelecek çalışmalarla ilgili veri sayılarını artırmayı hedefleyen araştırmacılar ayrıca farklı ön işlem adımları uygulayarak model başarımını arttırabileceklerini belirtmişlerdir. Diğer bir dezavantaj olarak önerilen sistemde yalnızca bir sınıflandırıcı kullanılmış olmasıdır. Farklı derin öğrenme mimarileri kullanarak sistem daha güçlü bir yapıya getirilebilir. Ayrıca önerilen Bi-LSTM modelinde parametre optimizasyonu yapılarak iyileştirme sağlanabilir.

Sosyal medya platformlarının çok yoğun kullanımı ve bilgi paylaşımının önüne geçilememesi, NS tespit sistemlerinin sürekli güncellenmesi, iyileştirilmesi ve daha hassas çalışmaları için gerekli adaptasyonların yapılmasına duyulan ihtiyacı da artırmaktadır. Bu çalışmada önerilen otomatik NS tespit sistemi sayesinde duyulan

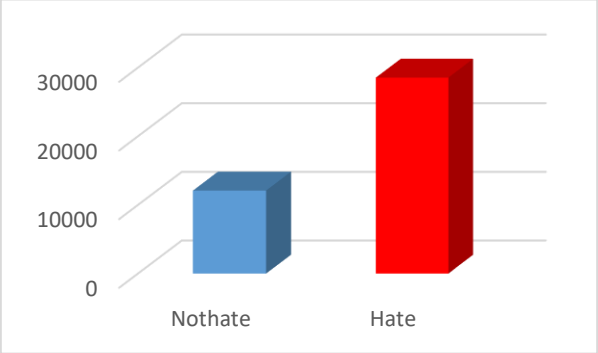
bu ihtiyacı karşılamak için yapay zeka temelli algoritmalar ile tatmin edici sonuçlar elde edilmiştir. Bir sonraki bölümde NS tespit sisteminde kullanılan materyal ve yöntemler detaylıca açıklanmıştır.

3. Metodolojiler

Bu bölümde otomatik NS tespiti için geliştirilen sistemde kullanılan materyal ve yöntemler hakkında detaylı bilgiler verilmiştir. İlk önce üzerinde çalışılan veri seti tanımlanmış, ardından metin önişleme adımları sıralanmıştır. Özellik çıkarımı için veri setine uygulanan NLP (Doğal Dil İşleme - Natural Language Processing) teknikleri sunulmuştur. NS tespit sisteminde tweetlerin doğru sınıflandırılması aşamasında seçilen yapay zeka temelli algoritmalar ve özellikleri açıklanmıştır.

3.1. Veri Setinin Karakteristiği

Çalışmada 40.623 tweetten oluşan Kaggle platformunda paylaşılan açık erişimli bir veri seti kullanılmıştır [26]. Bu veri setinde dinamik bir şekilde üretilen tweetler yer almaktadır. Sentetik veriler olarak adlandırılan bu veriler, genel sosyal ağ problemlerinin çözümünde karşılaşılan engellerden biri olan veri sayısının az olması sorununu ortadan kaldırmıştır. Verilerin sınıflara göre dağılımını Şekil 1'de gösterilmiştir.



Şekil 1. Çalışmada kullanılan verilerin dağılımı.

Bu veri setinde İngilizce dilinde 28.540 adet 'hate - nefret' etiketine sahip veri varken, 12.083 adet 'nothate – nefret olmayan' etikete sahip tweet vardır. Dengesiz bir dağılıma sahip veri setinin olduğu anlaşılmaktadır. Tablo 1'de her iki sınıfa ait veri setinde yer alan tweetlerden bazı örnekler listelenmiştir. Listelenen örneklerden anlaşılacağı üzere tweetler ırkçı, cinsiyetçi ve etnik kökeni hedef alan NS ifadelerinden oluşmaktadır.

Tablo 1. Veri setinde sınıflara ait örnekler.

Orijinal Tweet	Etiketi
"dalits are deserves to be treated as lowlife"	hate
"Women just don't have the brainpower to compete with men"	hate
"Chinese should be banned from entering this country"	hate
"All women are rat snakes"	hate
"Black guys are always looking for trouble"	hate
⋮	⋮
"Short people are smart to some extent"	nothate
"I do not think that all gay people are scum"	nothate
"Women should even try to play football"	nothate
"Muslims don't eat food that contains maggots"	nothate
"discrimination against blacks is wrong and disgusting"	nothate

3.2. Veri Ön İşleme ve NLP Adımları

Veri seti üzerinde NS tespit çalışması yapabilmek ve model oluşturabilmek için ilk adım metin önişlem adımlarının uygulanmasıdır. Bu çalışmada en yaygın kullanılan NLP adımları model eğitimi için kullanılmıştır. Metin önişleme safhasında ilk olarak tweetlerde yer alan noktalama işaretleri temizlenmiştir. Ardından sayısal

ifadeler veri setindeki her bir tweetten silinmiştir. *n*-karakterden az harfe sahip tweetler içerisindeki kelimeler çıkarılmıştır (n>3). Yine kelimelerin ortak bir şekilde tanımlanması için büyük-küçük harf dönüşümü yapılmıştır. Veri setinde yer alan tweetlerin içeriği küçük harfe dönüştürülmüştür. Tüm bu metin önişlem adımları model kurma için en yaygın kullanılan yöntemlerdir.

Özellik çıkarı için t-DM kullanılmıştır. t-DM oluşturulması için öncelikle veri setinde yer tüm kelimeler köklerine ayrılmıştır. Kökleri ayırma adımında İngilizce kök ayrıştırma için Snowball-Stemmer algoritması kullanılmıştır. Ardından her bir kök kelimenin tüm veri setinde kaç adet kullanıldığı terim frekans ağırlığı yardımıyla hesaplanmıştır. Ardından NS için hazırlanan veri setinde en çok geçen kelimelerin yer aldığı kelime çantası oluşturulmuştur. Son olarak metinleri temsil edecek özellik matrisi t-DM oluşturulmuştur. NS tespit sisteminde oluşturulan t-DM Tablo 2 ile tanımlanabilir.

Tablo 2. NS için önerilen t-DM yapısı.

	t_1	t_2	t_3	...	t_{k-2}	t_{k-1}	t_k
C_1	1	0	0		0	1	1
C_2	0	1	0		1	0	1
C_3	0	1	1		0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
C_{n-2}	0	1	0		0	1	0
C_{n-1}	1	0	0		0	1	1
C_n	0	1	0		0	0	1

Tablo 2'de *t* veri seti için bir özelliği, *C* ise veri setinde yer alan her bir tweeti teslim etmektedir. *k* özellik sayısını *n* ise veri setindeki toplam tweet sayısını temsil etmektedir. C_1t_1 çıktısı 1 olduğu için C_1. tweette t_1. özellik var demektir. C_2. tweet için t_2. özellik olmadığı için ise 0 değerini almıştır. Bu şekilde tüm veri seti sayısal olarak temsil edilmiştir.

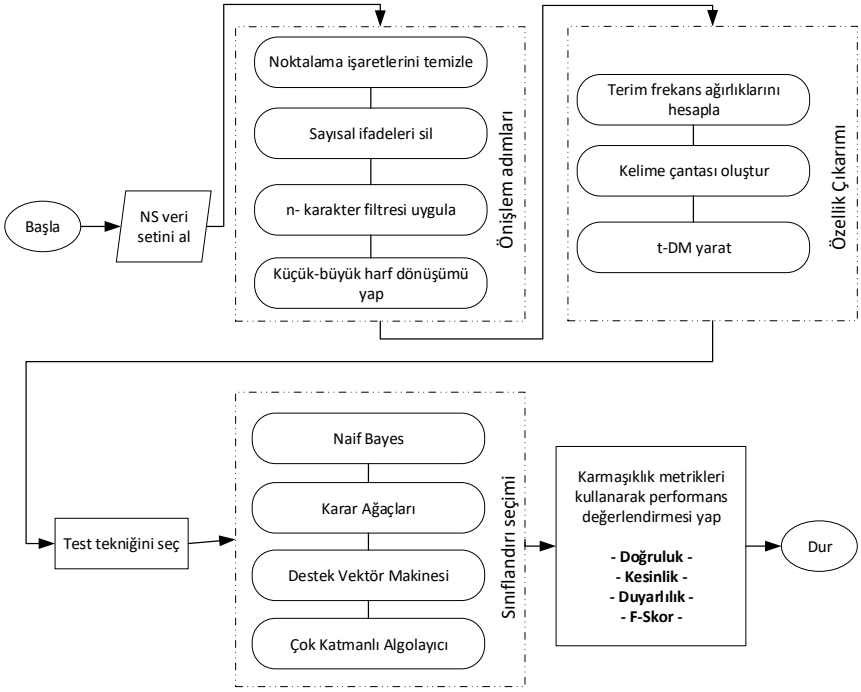
3.3. Yapay Zekaya Dayalı Algoritmaların Seçimi

Yapay zekaya dayalı algoritmalar için NS tespiti ve birçok sosyal medya problemini çözmek tipik görevlerden biridir. NS tespit sistemi için bu çalışmada beş farklı yapay zeka temelli algoritma kullanmıştır. Bu algoritmaların parametrelerini ve özelliklerini şu şeklinde sıralayabiliriz.

- **Multinominal Naive Bayes (MNB):** Metin sınıflandırma, spam filtreleme ve birçok sosyal ağ probleminin çözümünde etkili bir şekilde kullanılmaktadır [27]. MNB araştırmacılar tarafından yaygın olarak kullanılan Naive Bayes sınıflandırıcının başka bir versiyonudur. MNB için kullanılırken parametreler varsayılan olarak seçilmiştir.
- **Lib-Destek Vektör Makinesi (Lib-SVM):** Çok sınıflı sınıflandırmayı destekleyen SVM için geliştirilen bir kütüphanedir [28]. Metin sınıflandırma ve sosyal ağ problemlerinin çözümünde oldukça yaygın kullanılmaktadır. NS problemi için Lib-SVM çalıştırılırken başlangıç parametreleri varsayılan olarak ayarlanmıştır.
- **CVFDT (Concept Adapting Very Fast Decision Tree – Konsepte Uyarlanan Çok Hızlı Karar Ağacı):** Hoeffding Bound karar ağacında çalışan bir algoritmadır [29]. Hoeffding ağacı, karar ağacının oluşturulması ve analizi için örneklerin zaman içinde değişmediği varsayımıyla büyük veri akışlarında yüksek öğrenme yeteneğine sahiptir. CVFDT'nin başlangıç parametreleri varsayılan olarak sabit bırakılmıştır.
- **DT-Part (Decision Tree-Part – Ağaç Ağacı-Kısımlı):** Böl ve yönet yaklaşımıyla çalışan bir diğer karar ağacı algoritmasıdır [30]. Karar verme aşamasında kural tabanlı C4.5 yapısını kullanır. NS tespit sisteminde DT-Part algoritması uygulanırken parametreler varsayılan olarak seçilmiştir.
- **Çok Katmanlı Algılayıcı (MLP):** Girdi, gizli ve çıkış katmanlarından oluşan temel bir sinir ağı ve denetimli öğrenme algoritmasıdır [31]. MLP özellikle NS ve sosyal ağ problemlerinde sınıflandırma ve tespit aşamasında etkin olarak çalışmaktadır. Öğrenme aşamasında geriye doğru yayılım tekniğini

kullanan bu algoritmanın gizli katmanında 20 nöron kullanılmıştır. Bu ağda aktivasyon fonksiyonu olarak sigmoid fonksiyonunu seçilmiştir. İterasyon sayısı 500 olarak ayarlanmıştır. Öğrenme oranı ve momentum katsayıları 0.3 ve 0.2 olarak sabit bırakılmıştır.

Yapay zekaya dayalı algoritmalar ile gerçekleştirilen otomatik NS tespit çalışmasının tüm adımlarını özetleyen akış diyagramı Şekil 2'de belirtilmiştir. Bir sonraki bölümde NS tespit sistemi için yapılan deneyler ve uygulanan algoritmaların performansları değerlendirilmiştir.



Şekil 2. Çalışmanın akış diyagramı.

4. Deneyler ve Sonuçları

Otomatik NS tespit sisteminde kullanılan veri seti üzerinde uygulanan NLP adımlarının ve seçilen yapay zeka temelli algoritmalarının başarısını değerlendirmek için çeşitli eğitim ve test teknikleri kullanılmıştır [32]. Uygulanan eğitim ve test tekniklerinde algoritmaların performansını değerlendirmek için karmaşıklık matrisi kullanılarak çeşitli performans değerlendirme metrikleri hesaplanmıştır. Tablo 3'te performans değerlendirmesi için hesaplanan karmaşıklık metrikleri ve matematiksel denklemleri ile birlikte sunulmuştur.

Tablo 3. Karışıklık metrikleri ve performans kriterleri.

Performans Kriterleri		Denklem	Karmaşıklık Metrikleri
Doğruluk	=	$\frac{(DP + DN)}{(DP + DN + DP + DN)}$	**DP:** Doğru Pozitif, **DN:** Doğru Negatif **YP:** Yanlış Pozitif, **YN:** Yanlış Negatif
Kesinlik	=	$\frac{(DP)}{(DP + YP)}$	
Duyarlılık	=	$\frac{(DP)}{(DP + YN)}$	
F-skor	=	$2\,x\frac{(DP)}{(DP + YP + YN)}$	

NS tespit sisteminin güvenirliliğini garanti etmek için farklı 4 farklı eğitim ve test tekniği kullanılmıştır. Bunlar literatürde en çok kullanılan teknikler olarak bilinmektedir [33]. Böylelikle algoritmaların farklı sınama teknikleri karşısında göstermiş oldukları performansları değerlendirilmiştir. Bu eğitim ve test tekniklerini şu şekilde sıralayabilir.

- Veri setinin %50'sinin eğitim %50'sinin test için ayrılması
- Veri setinin %70'inin eğitim %30'unun test için ayrılması
- 5-kat çapraz doğrulama
- 10-kat çapraz doğrulama

NS tespit sistemi için önerilen modeli eğitmek ve test etmek için ilk iki teknikte veri seti eğitim ve test verileri olarak farklı oranlarda ayrılmıştır. Son iki eğitim ve test tekniğinde ise önce 5-kat sonra 10-kat çapraz doğrulama teknikleri kullanılmıştır. Böylelikle otomatik NS sisteminin güvenilirliği garanti edilmiştir. Tablo 4'te ilk test tekniği için 5 farklı algoritmaların göstermiş olduğu performanslar sıralanmıştır.

Tablo 4. %50 eğitim-%50 test tekniğiyle elde edilen performans sonuçları.

%50 eğitim-%50 test		Doğruluk	Kesinlik	Duyarlılık	F-Skor	Mutlak Hata
Algoritmalar	**MNB**	0.7574	0.739	0.757	0.717	0.3419
	Lib-SVM	0.7728	0.771	0.773	0.73	**0.2272**
	CVFDT	**0.8035**	**0.805**	**0.804**	0.777	0.2839
	DT-Part	0.8033	**0.805**	0.803	0.777	0.2812
	MLP	0.7962	0.785	0.796	**0.785**	0.2966

Tablo 4 incelendiğinde bu deney için NS tespitinde veri setinin yarısı eğitim diğer yarısının ise test için kullanılmıştır. CVFDT, DT-Part ve MLP algoritmaları birbirine çok yakın değerler elde ederek yaklaşık %80 doğruluk oranına sahip performans göstermişlerdir. Sırasıyla bu algoritmaları %77 ve %76 doğruluk değerleriyle Lib-SVM ve MNB takip etmişlerdir. Kesinlik ve duyarlılık performans kriterleri içinde bu sıralama değişmemiştir. %79 en yüksek f-skor değeri MLP sinir ağı tarafından elde edilerek sıralamada birinci sıraya yerleşmiştir. Bu deney için hesaplanan mutlak hata en düşük Lib-SVM algoritması tarafından elde edilmiştir. MNB sınıflandırıcısı ilk deney için en kötü performans gösteren algoritma olmuştur. Tablo 5'te veri setinin %70'inin eğitim, %30'unun test için ayrılması sonucu elde edilen deney sonuçları verilmiştir.

Tablo 5. %70 eğitim-%30 test tekniğiyle elde edilen performans sonuçları.

%70 eğitim-%30 test		Doğruluk	Kesinlik	Duyarlılık	F-Skor	Mutlak Hata
Algoritmalar	**MNB**	0.7553	0.739	0.755	0.714	0.3416
	Lib-SVM	0.7765	0.785	0.776	0.732	**0.2235**
	CVFDT	0.8004	**0.813**	0.8	0.768	0.2849
	DT-Part	0.7991	0.805	0.799	0.769	0.2816
	MLP	**0.8006**	0.803	**0.801**	**0.773**	0.262

İkinci deney için NS tespitinde veri setinin %70'ini eğitim geriye kalanı ise test için kullanılmıştır. CVFDT, DT-Part ve MLP algoritmaları birbirine çok yakın doğruluk değerleri elde etse de en yüksek doğruluk değerine MLP sinir ağı tarafından ulaşılmıştır. Sırasıyla bu algoritmaları %78 ve %76 doğruluk değerleriyle Lib-SVM ve MNB sınıflandırıcıları takip etmiştir. Kesinlik performansında CVFDT algoritması en yüksek değere ulaşmıştır. Ardından DT-Part algoritması ikinci en büyük kesinlik değerini elde etmiştir. MLP bu deney için çok az farkla üçüncü en yüksek kesinlik değerine sahip algoritma olmuştur. Duyarlılık performansında sıralamasıyla MLP, CVFDT, DT-Part, Lib-SVM ve MNB en yüksek değerlere ulaşmıştır. %77 en yüksek f-skor değeri yine MLP sinir ağı tarafından elde edilmiştir. DT-Part ve CVFDT algoritmaları çok az farkla MLP algoritmasını takip etmişlerdir. Dördüncü ve beşinci sırada ise sırasıyla Lib-SVM ve MNB algoritmaları yerini almıştır. İkinci deney için hesaplanan mutlak hata yine en düşük Lib-SVM algoritması tarafından elde edilmiştir. MNB sınıflandırıcısı ikinci deney için en kötü performans gösteren algoritma olmuştur.

NS tespit sisteminde kullanılan diğer bir yöntem olan 5-kat çapraz doğrulama tekniği kullanılarak elde edilen performans sonuçları Tablo 6'da listelenmiştir.

Tablo 6. 5-kat çapraz doğrulama ile elde edilen performans sonuçları.

	5-kat çapraz doğrulama	Doğruluk	Kesinlik	Duyarlılık	F-Skor	Mutlak Hata
Algoritmalar	**MNB**	0.7553	0.737	0.755	0.714	0.342
	Lib-SVM	0.7807	0.784	0.781	0.742	**0.2193**
	CVFDT	**0.8002**	**0.808**	**0.8**	0.769	0.2845
	DT-Part	0.799	**0.808**	0.799	0.767	0.2806
	MLP	0.7947	0.785	0.795	**0.776**	0.2814

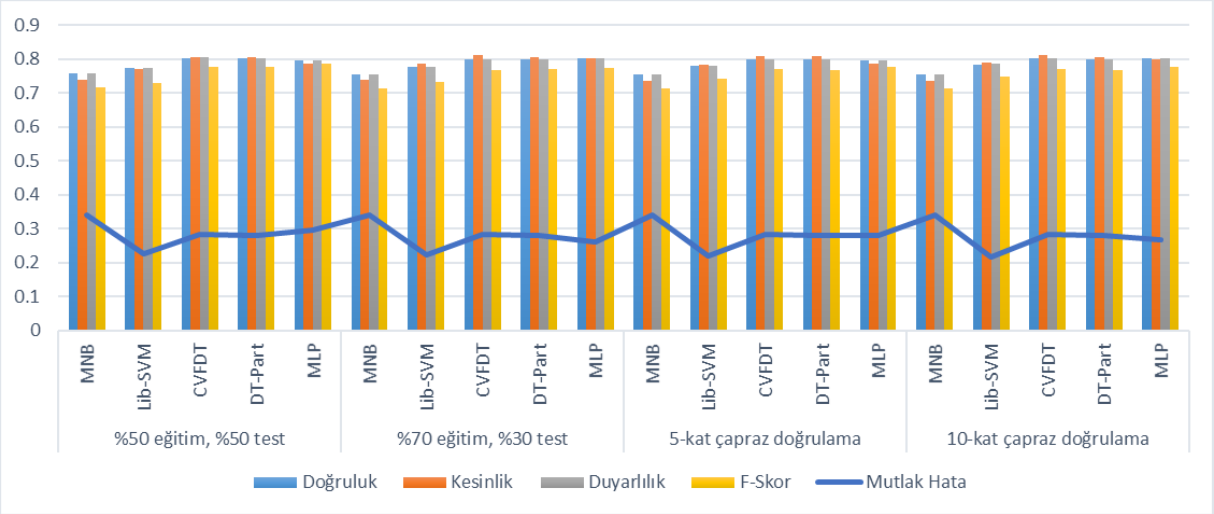
Tablo 6 incelendiğinde üçüncü deney için NS tespitinde 5-kat çapraz doğrulama tekniği kullanılmıştır. Üçüncü deney için CVFDT, DT-Part ve MLP algoritmaları birbirine çok yakın değerler elde ederek yaklaşık %80 doğruluk oranına sahip performans üretmişlerdir. Lib-SVM ve MNB sırasıyla bu algoritmaları %78 ve %76 doğruluk değerleriyle takip etmişlerdir. Kesinlik ve duyarlılık performans kriterleri içinde bu sıralama değişmemiştir. En yüksek %81 kesinlik değeri karar ağaçları ile elde edilmiştir. %78 en yüksek f-skor değeri ile MLP sinir ağı üçüncü deney için sıralamada birinci sıraya yerleşmiştir. Üçüncü deney için hesaplanan mutlak hata en düşük yine Lib-SVM algoritması tarafından elde edilmiştir. MNB sınıflandırıcısı ilk deney için en kötü performans gösteren algoritma olmuştur. MNB algoritması üçüncü deney için tekrar en kötü performans gösteren sınıflandırıcı olmuştur. Tablo 7'de 10-kat çapraz doğrulama tekniği kullanılarak elde edilen performans sonuçları sunulmuştur.

Tablo 7. 10-kat çapraz doğrulama ile elde edilen performans sonuçları.

	10-kat çapraz doğrulama	Doğruluk	Kesinlik	Duyarlılık	F-Skor	Mutlak Hata
Algoritmalar	**MNB**	0.7553	0.737	0.755	0.714	0.342
	Lib-SVM	0.7847	0.789	0.785	0.748	**0.2153**
	CVFDT	**0.8009**	**0.811**	**0.801**	0.769	0.2826
	DT-Part	0.7987	0.805	0.799	0.768	0.2811
	MLP	0.8007	0.799	**0.801**	**0.776**	0.2687

Son deney için otomatik NS tespitinde 10-kat çapraz doğrulama tekniği kullanılmıştır. Sırasıyla CVFDT, MLP ve DT-Part algoritmaları birbirine çok yakın %80'lik doğruluk değerine sahiptirler. Bu algoritmaları son deneyde de %78 ve %76 doğruluk değerleriyle Lib-SVM ve MNB sınıflandırıcıları takip etmiştir. Kesinlik performansında CVFDT ve MLP algoritmaları eşit değere ulaşmıştır. Ardından DT-Part algoritması ikinci en büyük kesinlik değerini elde etmiştir. Kesinlik değerinde de en kötü iki algoritma Lib-SVM ve MNB olmuştur. Duyarlılık performansında algoritmalar sıralamasıyla MLP (%80), CVFDT (%80), DT-Part (%80), Lib-SVM (%79) ve MNB (%76) en yüksek değerlere ulaşmıştır. Son deney için %78 en yüksek f-skor değeri yine MLP sinir ağı tarafından elde edilmiştir. CVFDT ve DT-Part algoritmaları çok az farkla (%77) MLP algoritmasını takip etmişlerdir. Son deneyde dördüncü ve beşinci en yüksek Lib-SVM (%75) ve MNB (%71) f-skor değerleri ile algoritmalar son sıralarda yerlerini almıştır. Son deney için hesaplanan mutlak hata yine en düşük Lib-SVM algoritması tarafından elde edilmiştir. MNB sınıflandırıcısı dördüncü deney için tekrar en kötü performans gösteren algoritma olmuştur.

Tüm deney sonuçlarından elde edilen analizler ışığında %80 en yüksek doğruluk performans değeri CVFDT, DT-Part ve MLP algoritmaları tarafından elde edilmiştir. Tüm deney sonuçlarında %78 en yüksek f-skor performans değerine MLP sinir ağı ulaşmıştır. En yüksek kesinlik değerine CVFDT algoritması %81 değerliyle sahip olmuştur. CVFDT, DT-Part ve MLP algoritmaları %80 en yüksek duyarlılık değerlerine tüm deneylerde ulaşmışlardır. NS sistemi için önerilen modelde en düşük hata oranına Lib-SVM algoritması sabittir. MNB algoritması diğer algoritmaların gerisinde kalarak tüm deneylerde en kötü performans göstererek performans sıralamasında sonuncu olmuştur. Şekil 3'te tüm deney sonuçları çubuk grafiklerle özetlenmiştir.



Şekil 3. Tüm eğitim ve test tekniklerine ait performans değerlendirmesi.

5. Sonuç

Çeşitli çevrimiçi sosyal medya ve paylaşım platformlarının kullanımının artmasıyla son yıllarda NS'nin yayılması önemli ölçüde artmıştır. Bir kişiyi, ulusu, devleti veya etnik kökeni hedef alan ve hız kesmeden yayılan NS, insanlar arasında kaos ve huzursuzluk yarattığı için bu içeriklerin tespiti ve yayılmasının engellenmesi son derece kritik bir görevdir. Bu çalışmada NS içeren metinsel ifadelerin sosyal ağlarda yayılmasını engellemek için yapay zeka temelli otomatik NS tespit sistemi önerilmiştir. NS'nin tespiti ve yayılmaması için kolluk kuvvetlerinin, yasama organlarının ve sosyal medya şirketlerinin milyonlarca yatırımına ve çabalarına rağmen, en etkili yöntemin metin tabanlı otomatik semantik analize dayandığı yaygın olarak kabul edilmektedir. Bu doğrultuda önerilen NS tespit sisteminin gerekliliği bir kez daha anlaşılmaktadır.

Otomatik NS tespit sistemi geliştirmek için, 40.623 sentetik metinden oluşturulan veri seti kullanılmıştır. Kullanılan bu veri setiyle sosyal ağ çalışmalarının en önemli problemlerinden olan biri olan veri sayısı yetersizliğinin önüne geçilmiştir. Kullanılan veri seti üzerinde temel metin önişleme ve NLP teknikleri kullanılarak özellik çıkarımı gerçekleştirilmiştir. 5 farklı yapay zeka tabanlı algoritma seçilerek güçlü bir NS sistemi önerilmiştir. Ardından önerilen NS tespit sisteminin güvenilirliğini garanti etmek için 4 farklı eğitim ve test teknikleri kullanılmıştır. Tüm deney sonuçlarının değerlendirilmesi sonucunda, en yüksek doğruluk performans değeri (%80) CVFDT, DT-Part ve MLP algoritmaları tarafından elde edilmiştir. En yüksek f-skor performans değerine (%78) MLP sinir ağı ulaşmıştır. En yüksek kesinlik değerine (%81) CVFDT algoritması sahip olmuştur. CVFDT, DT-Part ve MLP algoritmaları en yüksek duyarlılık değerine (%80) tüm deneylerde ulaşmışlardır. NS sistemi için önerilen modelde en düşük hata oranına Lib-SVM algoritmasının sahip olduğu görülmüştür. Bu derece büyük bir veri setinde elde edilen başarı oldukça dikkat çekicidir. Başarımı daha da artırmak için parametre optimizasyonu yapılabilir. Ayrıca farklı yapay zeka temelli algoritmalar kullanılarak kıyaslama çalışmaları geliştirilebilir.

Kaynaklar

[1] Baydogan C, Alatas B. Metaheuristic Ant Lion and Moth Flame Optimization-Based Novel Approach for Automatic Detection of Hate Speech in Online Social Networks. IEEE Access, 2021; Vol. 9: 110047-110062.

[2] MacAvaney S, Yao HR, Yang E, Russell K, Goharian N, Frieder O. Hate speech detection: Challenges and solutions. PloS one, 2019; 14(8): e0221152.

[3] Gitari ND, Zuping Z, Damien H, Long J. A lexicon-based approach for hate speech detection. International Journal of Multimedia and Ubiquitous Engineering, 2015; 10(4): 215-230.

[4] Köffer S, Riehle DM, Höhenberger S, Becker J. Discussing the value of automatic hate speech detection in online debates. Multikonferenz Wirtschaftsinformatik (MKWI 2018): Data Driven X-Turning Data in Value, 2018.

[5] Waseem Z, Thorne J, Bingel J. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In Online harassment, Springer, Cham. 2018; 29-55.

[6] Badjatiya P, Gupta M, Varma V. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In The World Wide Web Conference, 2019; 49-59.
[7] Mossie Z, Wang JH. Social network hate speech detection for Amharic language. Computer Science & Information Technology, 2018; 41-55.
[8] Miok K, Škrlj B, Zaharie D, Robnik-Šikonja, M. To ban or not to ban: Bayesian attention networks for reliable hate speech detection. Cognitive Computation, 2021; 1-19.
[9] Robinson D, Zhang Z, Tepper J. Hate speech detection on twitter: Feature engineering vs feature selection. In European Semantic Web Conference Springer, Cham, 2018; 46-49.
[10] Korzeniowski R, Rolczyński R, Sadownik P, Korbak T, Możejko M. Exploiting Unsupervised Pre-training and Automated Feature Engineering for Low-resource Hate Speech Detection in Polish. arXiv preprint, 2019; arXiv:1906.09325.
[11] Ombui E, Muchemi L, Wagacha P. Hate speech detection in code-switched text messages. In 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT - IEEE), 2019; 1-6.
[12] Abro S, Shaikh ZS, Khan S, Mujtaba G, Khand ZH. Automatic Hate Speech Detection using Machine Learning: A Comparative Study. International Journal of Advanced Computer Science and Applications(IJACSA), 2020; 10(6): 484-491.
[13] Pathak V, Joshi M, Joshi P, Mundada M, Joshi T. KBCNMUJAL@ HASOC-Dravidian-CodeMix-FIRE2020: Using Machine Learning for Detection of Hate Speech and Offensive Code-Mixed Social Media text. arXiv preprint arXiv:2102.09866, 2021.
[14] Omar A, Mahmoud TM, Abd-El-Hafeez T. Comparative performance of machine learning and deep learning algorithms for Arabic hate speech detection in osns. In: The International Conference on Artificial Intelligence and Computer Vision. Springer, Cham, 2020; 247-257.
[15] Fauzi MA, Yuniarti A. Ensemble method for indonesian twitter hate speech detection. Indonesian Journal of Electrical Engineering and Computer Science, 2018; 11(1): 294-299.
[16] Plaza-del-Arco FM, Molina-González MD, Ureña-López LA, Martín-Valdivia MT. Comparing pre-trained language models for Spanish hate speech detection. Expert Systems with Applications, 2021; 166: 114120.
[17] Bohra A, Vijay D, Singh V, Akhtar SS, Shrivastava M. A dataset of hindi-english code-mixed social media text for hate speech detection. In Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media, 2018; 36-41.
[18] Alshalan R, Al-Khalıfa H. A deep learning approach for automatic hate speech detection in the saudi twittersphere. Applied Sciences, 2020; 10(23): 8614.
[19] Al-Makhadmeh Z, Tolba A. Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach, Computing, 2020; 102(2): 501-522.
[20] Zhou Y, Yang Y, Liu H, Liu X, Savage N. Deep learning based fusion approach for hate speech detection. IEEE Access, 2020; 8: 128923-128929.
[21] Pitsilis GK, Ramampıaro H, Langseth H, Effective hate-speech detection in Twitter data using recurrent neural networks. Applied Intelligence, 2018; 48(12): 4730-4742.
[22] Roy PK, Tripathy AK, Das TK, Gao XZ. A Framework for Hate Speech Detection Using Deep Convolutional Neural Network. IEEE Access, 2020; 8: 204951-204962.
[23] Ayo FE, Folorunso O, Ibharalu FT, Osinuga IA. Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. Computer Science Review, 2020; 38: 100311.
[24] Pitropakis N, Kokot K, Gkatzia D, Ludwiniak R, Mylonas A, Kandias M. Monitoring Users' Behavior: Anti-Immigration Speech Detection on Twitter. Machine Learning and Knowledge Extraction, 2020; 2(3): 192-215.
[25] Garain A, Basu A. The titans at SemEval-2019 task 5: Detection of hate speech against immigrants and women in twitter. In Proceedings of the 13th International Workshop on Semantic Evaluation, 2019; 494-497.
[26] https://www.kaggle.com/usharengaraju/dynamically-generated-hate-speech-dataset.
[27] Kibriya AM, Frank E, Pfahringer B, Holmes G. Multinomial Naive Bayes for Text Categorization Revisited. Advances in Artificial Intelligence. 2004; 3339: 488-499.
[28] Chih-Chung C, Chih-Jen L. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2011; 2(27):1-27.
[29] Kumar A, Kaur P, Sharma P. A survey on Hoeffding tree stream data classification algorithms. CPUH-Res. J, 2015; 1(2): 28-32.
[30] Frank E, Witten IH. Generating accurate rule sets without global optimization. In fifteenth international conference on machine learning, 1998; 144-151.
[31] Roul RK, Asthana SR, Kumar G. Study on suitability and importance of multilayer extreme learning machine for classification of text data. Soft Computing, 2017; 21(15): 4239-4256.
[32] Baydogan C, Alatas B. Detection of Customer Satisfaction on Unbalanced and Multi-Class Data Using Machine Learning Algorithms. In 2019 1st International Informatics and Software Engineering Conference (UBMYK - IEEE), 2019; 1-5.
[33] Baydogan C, Alatas B. Sentiment analysis using Konstanz Information Miner in social networks. In 6th International Symposium on Digital Forensic and Security (ISDFS - IEEE), 2018; 1-5.