*Avrupa Bilim ve Teknoloji Dergisi*
*Özel Sayı 24, S. 328-334, Nisan 2021*

***Araştırma Makalesi***

www.ejosat.com ISSN:2148-2683

*European Journal of Science and Technology*
*Special Issue 24, pp. 328-334, April 2021*

***Research Article***

# Türkçe Tweetler üzerinde Makine Öğrenmesi ile Nefret Söylemi Tespiti

İslam Mayda[1*], Banu Diri[2], Tuğba Dalyan[3]

[1*] Yıldız Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye, (ORCID: 0000-0001-5584-0259), islam.mayda@stu.khas.edu.tr
[2] Yıldız Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye (ORCID: 0000-0002-6652-4339), diri@yildiz.edu.tr
[3] İstanbul Bilgi Üniversitesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye (ORCID: 0000-0002-5868-5407), tugba.yildiz@bilgi.edu.tr

(2nd International Conference on Access to Recent Advances in Engineering and Digitalization (ARACONF)-10–12 March 2021)

(**DOI:** 10.31590/ejosat.903854)

**ATIF/REFERENCE:** Mayda, İ., Diri, B. & Dalyan, T. (2021). Türkçe Tweetler üzerinde Makine Öğrenmesi ile Nefret Söylemi Tespiti. *Avrupa Bilim ve Teknoloji Dergisi*, (24), 328-334.

**Öz**

Sosyal medya ağlarının sayısının ve kullanımının artması beraberinde nefret söylemi içeriklerinin de daha çok paylaşılması problemini doğurmuştur. Gerek kamu otoriteleri gerekse sosyal medya ağlarının kendileri, artan nefret söylemiyle mücadele kapsamında çeşitli politikalar üretmektedir. Kullanıcılar tarafından üretilen verinin hacminin oldukça büyük olması nedeniyle nefret söylemi tespitinde otomatik sistemlere ihtiyaç duyulmaktadır. Özellikle son yıllarda başta İngilizce olmak üzere birçok dil üzerinde otomatik nefret söylemi çalışması yapılmış olmasına rağmen Türkçe üzerine kapsamlı bir çalışma henüz sunulmamıştır. Bu çalışma bu ihtiyaca karşılık vermek amacıyla yapılmıştır. Farklı hedef gruplara dair anahtar kelimelerin geçtiği 1000 adet Türkçe tweet toplanmış ve iki değerlendirici tarafından üç sınıflı (nefret söylemi, saldırgan ifade, hiçbiri) olarak ayrı ayrı etiketlenmiştir. Oluşturulan Türkçe nefret söylemi veri seti sonraki çalışmalarda kullanılabilmesi için kamuya açık olarak paylaşılmıştır. Bu veri seti üzerinde farklı özellik kümeleri ve farklı makine öğrenmesi algoritmaları kullanılarak çeşitli testler gerçekleştirilmiştir. Üç sınıflı veri seti üzerinde en yüksek performans %79,9 F-ölçüm değeri ile SMO (Sıralı Minimal Optimizasyon) algoritmasının kullanıldığı testte elde edilmiştir. Türkçe nefret söylemi tespitinde daha başarılı sonuçlar almak için veri seti boyutunun artırılması gerekirken, sunulan bu çalışmanın gelecekte yapılacak çalışmalara öncü niteliğinde olması beklenmektedir.

**Anahtar Kelimeler:** Nefret söylemi tespiti, makine öğrenmesi, Türkçe tweetler.

# Hate Speech Detection with Machine Learning on Turkish Tweets

**Abstract**

The increase in the number and use of social media networks has led to the problem of sharing hate speech content more. Both public authorities and social media networks themselves produce various policies within the scope of combating increasing hate speech. Automated systems are needed to detect hate speech due to the very large volume of the data produced by users. Although, in recent years, automatic hate speech studies have been conducted on many languages, especially English, a comprehensive study on Turkish has not been presented yet. This study was carried out in order to meet this need. 1000 tweets in Turkish with keywords for different target groups were collected and labeled separately in three categories (hate speech, offensive expression, none of them) by two evaluators. The Turkish hate speech data set created was shared publicly for use in future studies. Various tests were carried out on this data set using different feature sets and different machine learning algorithms. The highest performance on the three-class data set was obtained in the test using the SMO (Sequential Minimal Optimization) algorithm with 79.9% F-measure value. While the size of the data set needs to be increased in order to achieve more successful results in detecting Turkish hate speech, this study is expected to be a pioneer for future studies.

**Keywords:** Hate speech detection, machine learning, Turkish tweets.

[*] Sorumlu Yazar: islam.mayda@stu.khas.edu.tr

# 1. Giriş

İnternet kullanımının yaygınlaşması, sosyal medya ağlarının sayısının artmasıyla üretilen web içeriği de hızlı bir şekilde artış göstermektedir. Sosyal medya siteleri hem dünyada[†] hem de Türkiye'de[‡] en çok ziyaret edilen web siteleri arasındadır. Bu tür platformlar, kullanıcılarına her türlü düşüncesini paylaşma imkânı sağlamaktadır. Buna bağlı olarak nefret söylemi içeren ifadeler de çoğalmaktadır. Nefret söylemi ve nefret suçu hakkında bir bilgi modeli geliştirme, test etme ve aktarmaya katkı sağlamayı amaçlayan eMORE Projesi raporunda; günümüzde kitle iletişim araçlarında, sosyal medyada ve günlük yaşamda "ötekine" yönelik korkunun gittikçe arttığı belirtilmiştir (RiSSC, 2017). Rapor, son yıllarda artan terör eylemleri, mülteci krizi gibi etkenlerin nefret söylemi artışında temel sebepler olduğunu ortaya koymaktadır. Avrupa'da nefret söyleminin giderek yaygınlaşmasına karşı olarak Avrupa Konseyi tarafından desteklenen *No Hate Speech Movement* hareketi ortaya çıkmıştır[§]. UNESCO da ülkelerin nefret söylemi sorunuyla başa çıkmalarına yardımcı olmak amacıyla *Countering Online Hate Speech* adlı bir çalışma yayınlamıştır (Gagliardone vd., 2015). Birleşmiş Milletler Genel Kurulu tarafından kabul edilen ve 167 üye ülkenin taraf olduğu Kişisel ve Siyasal Haklar Uluslararası Sözleşmesi (International Covenant on Civil and Political Rights - ICCPR) ayrımcılığa, düşmanlığa veya şiddete teşvik oluşturan her türlü ulusal, ırksal veya dini nefretin savunuculuğu kanunla yasaklanacaktır şeklinde bir madde içermektedir (Birleşmiş Milletler, 1976).

Nefret söylemi ve nefret suçlarının dünyayla paralel olarak Türkiye'de de arttığı tespit edilmiştir (Arcan, 2013). ABD, İngiltere ve Almanya gibi pek çok gelişmiş ülkede nefret söylemi ve suçuna yönelik yasal düzenlemeler olmasına rağmen T.C. Anayasası'nda nefret söylemiyle ilgili doğrudan bir düzenleme veya özel bir kanun bulunmamaktadır. Ancak, Türk Ceza Kanunu'nda nefret suçlarıyla ilişkilendirilebilecek bazı maddeler yer almaktadır ve buna bağlı olarak çeşitli olaylarda birtakım cezai yaptırımlar uygulanmaktadır (Kaya, 2018).

Nefret söyleminin hedef kişi veya gruplar üzerindeki fiziksel ve zihinsel zararlarına yönelik çeşitli araştırmalar yayınlanmıştır (Waldron, 2014; Gelber & McNamara, 2016). Bu tür paylaşımların hedef kitle üzerindeki doğrudan zararlarının yanı sıra diğer okuyucuların görüşlerini etkilemeleri ve onları da harekete geçirmesi bakımından dolaylı zararları da vardır. İçerik ne kadar uzun süre paylaşımda kalırsa, mağdurlara o kadar fazla zarar verir (Gagliardone vd., 2015). Bu yüzden nefretin daha fazla yayılmaması için nefret söylemi içeriğinin en kısa süre içerisinde paylaşımdan kaldırılması gerekir. Sosyal medya ağlarındaki toplam kullanıcı sayıları ve paylaşım sayıları[**] düşünüldüğünde, bunun manuel olarak yapılmasının mümkün olmadığı görülmektedir.

Kullanıcılar tarafından içeriği oluşturulan birçok sosyal medya ağı (Facebook[††], Twitter[‡‡], Youtube[§§], vs.) nefret söylemi içeren ifadelerin paylaşımını yasaklamıştır ve bu paylaşımları engellemek için algoritmik çözümler uygulamaya çalışmaktadır. Ancak, üretilen içeriğin büyüklüğü ve problemin çok yönlü olması bu durumu zorlaştırmakta ve nefret söylemleri çevrimiçi platformlarda sorun olmaya devam etmektedir (Djuric vd., 2015). Bu nedenle özellikle son yıllarda, nefret söylemi içeriklerinin otomatik olarak tespit edilmesine yönelik çalışmalara daha çok yoğunlaşılmıştır.

Nefret söyleminin evrensel olarak kabul gören bir tanımı bulunmamaktadır. Nefret söylemi tespiti üzerine yapılan çalışmalarda bu konudaki sınırların net olarak belirlenmiş olmadığı ifade edilmekte ve yapılan tanıma göre veri seti etiketleme sürecinin değiştiğine, dolayısıyla çalışmanın başarısının da etkilediğine dikkat çekilmektedir (Waseem, 2016). Bu yüzden öncelikle bu konuda bir çalışma yapmadan önce nefret söylemi tanımının net olarak yapılması gerekmektedir. Bugüne kadar yapılan bazı nefret söylemi tanımları şöyledir:

Avrupa Konseyi'nin Bakanlar Komitesi 97(20) sayılı Tavsiye Kararı'nda nefret söylemini şu şekilde tanımlamıştır: "nefret söylemi" kavramının ırkçı nefreti, yabancı düşmanlığını, Yahudi aleyhtarlığını veya azınlıklara, göçmenlere veya göçmen kökenli insanlara karşı saldırgan milliyetçilik ve etnik merkezcilik, ayrımcılık ve düşmanlık ifadesi içeren hoşgörüsüzlüğe dayanan diğer nefret biçimlerini yayan, teşvik eden, destekleyen veya meşrulaştıran tüm ifade biçimlerini kapsadığı anlaşılmalıdır (Weber, 2009).

Hukuk profesörü olan Nockleby Amerikan Anayasası Ansiklopedisi'nde nefret söylemi kavramını "bir kişiyi veya grubu ırk, renk, etnik köken, cinsiyet, cinsel yönelim, milliyet, din veya diğer özellikler gibi bazı özelliklere göre ayıran her türlü ifade" şeklinde tanımlamıştır (Nockleby, 2000). UNESCO tarafından yayınlanan Online Nefret Söylemi ile Mücadele kitabında bu tanım "belirli bir sosyal veya demografik grupla tanımlanan hedefe zarar vermeye (özellikle de ayrımcılığa, düşmanlığa veya şiddete) teşvik eden ifadeler" şeklinde yer almıştır (Gagliardone, 2015).

Nefret söyleminin tanımı üzerinde düşünülürken bireylerin ifade özgürlüğü hakkı da göz önünde bulundurulmalıdır. Diğer bir deyişle, bir yandan nefret söylemi içeriklerini yasaklarken diğer yandan da kişilerin ifade özgürlüğü hakkının korunması gerekmektedir. Ancak, nefret söylemi ile ifade özgürlüğü arasındaki çizgi günümüzde çok net çizilememektedir. Bu da diğer bir tartışma konusudur (Çelik, 2013).

Nefret söyleminin otomatik olarak tespit edilmesine yönelik son yıllarda çok sayıda yayın yapılmıştır (Fortuna & Nunes, 2018; Schmidt & Wiegand, 2017). Bu yayınların büyük bir kısmı İngilizce üzerine iken Almanca, İtalyanca, İspanyolca, Arapça, Endonezce dilleri üzerine de bazı çalışmalar literatürde yer almaktadır. Buna karşılık, bugüne kadar Türkçe dili üzerine nefret söyleminin otomatik tespitine yönelik, kadına karşı nefret söylemine dair sunulan bir çalışma (Şahi vd., 2018) haricinde başka bir çalışmaya rastlanmamıştır. Şahi vd. yaptıkları çalışmada öncelikle Twitter'da *#kıyafetimekarisma* etiketiyle atılan tweetleri iki sınıflı olarak etiketlemiş, daha sonra dengeli bir veri seti oluşturmak için bunların içinden 318 tweet (159 adedi nefret söylemi, 159 adedi nefret söylemi değil) seçerek

[†] https://www.alexa.com/topsites

[‡] https://www.alexa.com/topsites/countries/TR

[§] https://www.coe.int/en/web/no-hate-campaign

[**] https://dustinstout.com/social-media-statistics/

[††] https://www.facebook.com/communitystandards/hate_speech

[‡‡] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

[§§] https://support.google.com/youtube/answer/2801939

deneylerini bu küçük veri seti üzerinde gerçekleştirmiştir. Farklı makine öğrenmesi yöntemlerinin denendiği çalışmada en yüksek F-ölçüm değeri %72 ile Rastgele Ağaç sınıflandırıcıyla yapılan deneyde elde edilmiştir. Bu çalışma Türkçe üzerine yapılan ilk çalışma olsa da kullanılan veri setinin gerek boyutunun çok küçük olması gerekse kapsamının çok dar olması nedeniyle yeterli değildir.

Ayrıca, Türkçe'de nefret söylemine dair Sosyal Bilimler alanındaki çeşitli dergilerde yer alan araştırmalar, bu kavramın analizi ve Türkiye'deki örnekleri üzerine tartışmalardan ibarettir (Vardal, 2015; Alp, 2016; Alp, 2018). Türkçe metinler üzerinde nefret söyleminin otomatik tespitine yönelik çalışmalar konusunda ciddi bir açık olduğu görülmektedir.

Bu çalışmanın ilk amacı, söz konusu ihtiyaca karşılık vermektedir. Bu tür bir çalışma yapabilmek için öncelikle Türkçe nefret söylemi veri setine sahip olunması gerekmektedir. Ancak, bu konuda daha önce kamuya açık olarak paylaşılan yapısal bir veri seti oluşturulmamıştır. Bu eksikliği gidermek de çalışmanın amaçlarından biridir. Çalışma kapsamında öncelikle Türkçe nefret söylemi veri seti oluşturulmuş, daha sonra oluşturulan veri seti üzerinde farklı öğrenme yöntemleri ve farklı özellik kümeleriyle deneysel çalışmalar yapılarak en ideal model bulunmaya çalışılmıştır.

# 2. Materyal ve Metot

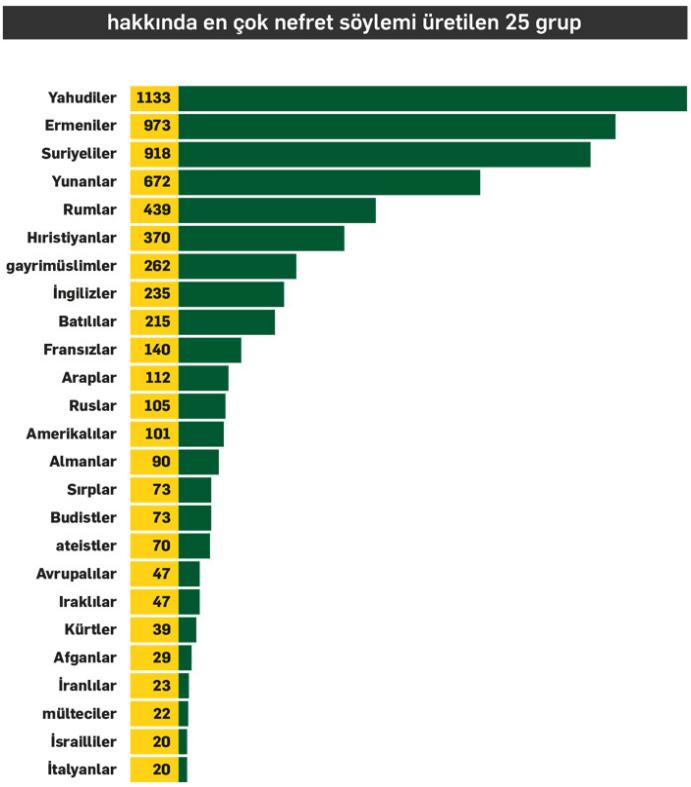
## 2.1. Türkçe nefret söylemi veri seti

Üzerinde sınıflandırma çalışması yapılacak Türkçe tweet veri seti oluşturulmadan önce, tweetleri toplarken aramada kullanılacak olan hedef gruplara dair anahtar kelimelere karar verilmiştir. Bunun için öncelikle, ulusal ve yerel medyadaki nefret söylemi haberlerinin listelendiği NefretSoylemi.org sitesinde yer alan içerikler incelenmiştir. Hrant Dink Vakfı'nın yürütmekte olduğu *Medyada Nefret Söyleminin İzlenmesi Projesi* kapsamında oluşturulan bu sitede, gönüllüler günlük ve haftalık gazeteleri takip ederek tespit ettikleri nefret söylemi haberlerini kamuoyu ile paylaşmaktadırlar.

Hedef gruplara dair anahtar kelimeler belirlenirken, NefretSoylemi.org sitesindeki nefret söylemi içeriklerinde en çok geçen terimlerle birlikte, Şekil 1'de sunulan ve yine Hrant Dink Vakfı tarafından aynı proje kapsamında hazırlanmış olan "Medyada Nefret Söylemi ve Ayrımcı Söylem 2018 Raporu"nun[***] sonuçları göz önünde bulundurulmuştur.

Bu doğrultuda Twitter'da arama yapılırken kullanmak üzere seçilen anahtar kelimeler şöyledir: *suriyeli, ermeni, ingiliz, kürt, türk, yunan, arap, rum, alevi, müslüman, yahudi, ateist, gavur, hıristiyan, kadın, sapık, eşcinsel*.

Şekil 1'deki rapora göre nefret söylemine en çok maruz kalan ilk üç grup olduğu için *yahudi, ermeni* ve *suriyeli* anahtar kelimeleri için 100'er, seçilen diğer anahtar kelimeler için 50'şer olmak üzere toplamda 1000 tweet toplanmıştır. Tweet toplama işlemi için KNIME aracı kullanılmıştır. Belirlenen anahtar kelimeler ile yapılan aramalarda Türkçe dilinde yazılmayan tweetler ile tweet tekrarları veri setine dâhil edilmemiştir.




*Şekil 1. Medyada Nefret Söylemi ve Ayrımcı Söylem 2018 Raporu'na göre hakkında en çok nefret söylemi üretilen 25 grup*

Küfür veya hakaret içeren her ifade nefret söylemi olarak görülemeyeceği gibi içerisinde hiçbir küfür, hakaret veya kötü söz içermeyen her ifade de nefret söylemi değildir diye değerlendirilemez (Malmasi & Zampieri, 2017). Bu noktada, saldırgan bir ifade içeren ancak nefret söylemi olmayan verilerin, nefret söylemi ifadesi içeren verilerden ayırt edilebilmesi önemli bir problemdir. Literatürdeki birçok araştırmada bu iki tür ifade aynıymış gibi değerlendirilmiş ve saldırgan ifade içeren her veri nefret söylemi içeriyor şeklinde etiketlenmiştir (Warner & Hirschberg, 2012; Kwok & Wang, 2013). Söz konusu çalışmalarda veri seti, nefret söylemi içeren ve içermeyen olmak üzere iki sınıftan oluşmaktadır ve sınıflandırma işlemi buna göre yapılmaktadır. Gerçekte ise saldırgan ifadeler ile nefret söylemi ifadeleri birbirinden farklı şeyler olduğu için, bu çalışmada Türkçe nefret söylemi veri seti oluştururken verilerin 3 sınıflı (nefret söylemi, saldırgan ifade, hiçbiri) olarak etiketlenmesi tercih edilmiştir.

Nefret söylemi olarak etiketlenen verilere ayrıca en az bir alt sınıf etiketi de (etnik, dini, cinsiyetçi, siyasi/ideolojik) verilmiştir. Bu çalışmada nefret söylemi alt sınıflarına dair bir deney yapılmamıştır, ancak bu alt sınıf etiketlerinin gelecek çalışmalarda daha kapsamlı deneylerde kullanılması düşünülmektedir.

Bir ifadenin nefret söylemi içerip içermediği farklı değerlendiriciler tarafından farklı şekilde yorumlanabilmektedir. Bir değerlendiricinin "nefret söylemi" olarak etiketlediği bir veri, başka bir değerlendirici tarafından "saldırgan ifade içeren ancak nefret söylemi olmayan" şeklinde etiketlenebilirken, bir başka değerlendirici tarafından "ne nefret söylemi ne de saldırgan bir ifade" olarak sınıflandırılabilir. Bu yargının kesin bir çerçevesinin olmaması, değerlendiricinin bakış açısına göre değişkenlik göstermesi veri seti etiketlemede ciddi bir sorun olarak gözlemlenmektedir (Razavi vd., 2010). Bu nedenle

[***] https://hrantdink.org/tr/asulis/faaliyetler/projeler/medyada-nefret-soylemi/2002-medyada-nefret-soylemi-ve-ayrimci-soylem-2018-raporu-yayimlandi

etiketleme işleminde en az iki değerlendiricinin görev alması önem arz etmektedir.

Bu çalışmada kullanılan veri setini etiketleme sürecinde öncelikle iki farklı değerlendirici birbirinden bağımsız olarak tweetleri etiketlenmiştir. Yapılan bağımsız etiketlemenin sonucunda iki değerlendiricinin vermiş olduğu etiketlerin uyuşum oranının %83.4 olduğu görülmüştür. Daha sonra iki değerlendirici bir araya gelerek farklı etiketlenen veriler üzerinde karşılıklı tartışarak birbirini ikna etme yoluyla uyuşum oranını %96.4'e yükseltmiştir. Üzerinde mutabık kalınamayan tweetler ise üçüncü değerlendiriciye sorulmuş ve en sonunda üçüncü değerlendiricinin verdiği etiketle birlikte oy çoğunluğuna sahip etiket değeri o verinin etiketi olarak atanmıştır. İlk iki değerlendiricinin sınıf etiketleri üzerinde uzlaşamadığı tweetlere ait örnekler Tablo 1'de sunulmuştur.

*Tablo 1. İki değerlendiricinin üzerinde uzlaşamadığı tweetler*

| Tweet |
|---|
| *Adana'da Suriyeli çok birine aşık olacam diye korkuyorum çünkü güzeller* |
| *İzmir'de denize döktüğümüz Yunanlılardan, Venizelos kolundaki kadın kim? bilen varmı? https://t.co/cq3A3LuPSe* |
| *Birader beni karıştırma bu işlere. zaten Yahudi zannediyolar isimden ötürü 😐 https://t.co/pg8oSgvSmB* |
| *Bugün müşterime ilahiyat fakültesi karşısından ev gösterdim. Bu muhitten nefret ediyorum işim olmasa durmam buralarda dedi. Neden dedim. İlahiyatı sevmiyorum dedi. Yahudi misiniz diye sordum. Yok canım Elhamdülillah müslümanım dedi. 3-5 saniye sessizce bakışıp dağıldık.* |
| *1891 ABD de her yıl 6000 kadın Doğumdan sonra ölüyor kafayı buna takan Dr lee bir buluşuyla 1 e düşürüyor ama hıristiyan. Tanrım bu adamı nereye alırsın. ödül mü cezamı verirsin öylesine sordum.* |

İki değerlendiricinin üzerinde uzlaşamadığı tweetler için üçüncü değerlendiricinin kararı alındıktan sonra veri setinin etiketleme süreci tamamlanmıştır. Veri setinin sınıf dağılımı Tablo 2'deki gibi olmuştur.

*Tablo 2. Oluşturulan veri setinin sınıf dağılımı*

| Etiket | Tweet Sayısı |
|---|---|
| nefret söylemi | 276 |
| saldırgan ifade | 60 |
| hiçbiri | 664 |
| **TOPLAM** | **1000** |

## 2.2. Önişleme ve Özellik Çıkarımı

Önişleme adımları olarak veri setindeki tüm metinler küçük harfe çevrilmiş, URL adresleri ve kullanıcı adları temizlenmiştir. Karakter ngram özellikleri çıkarılırken, sınıflandırma başarısına olumlu etki edeceği düşünülen nokta, ünlem, soru işareti, tek tırnak ve çift tırnak haricindeki diğer noktalama işaretleri de silinmiş, fazladan boşluklar tek boşluğa indirilmiştir. Yazım hataları öncelikle Google Dokümanlar servisinin Yazım Denetimi seçeneği ile otomatik olarak düzeltilmiş, daha sonra bütün tweetler gözden geçirilerek otomatik olarak düzeltilemeyen yazım hatalarının elle düzeltimi yapılmıştır. Veri setinin önişleme adımlarından sonraki son hali web tabanlı bir depolama servisi olan Github'da[†††] açık kaynak olarak sunulmuştur.

Kelimelerin morfolojik analizinde açık kaynak kodlu Türkçe doğal dil işleme kütüphanesi olan Zemberek (Akın & Akın, 2007) tercih edilmiştir. Özellik kümeleri olarak karakter bigram ve trigramları, kelime unigram, bigramları ve tweete özgü özellikler çıkarılarak kullanılmıştır. Kelime ngramları için, Zemberek ile kök ve eklerine ayrılan kelimelerin çekim ekleri atılarak kök ve yapım eklerinden oluşan gövdeleri elde edilmiş ve kelime gövdeleri terim olarak kullanılmıştır. Karakter ngramları bulunurken veriler olduğu gibi işlenmiş, kelime ngramları bulunurken ise verilerdeki tüm noktalama işaretleri temizlenerek işleme alınmıştır.

Bu aşamada elde edilen özellik kümelerindeki özellik sayıları Tablo 3'teki gibidir.

*Tablo 3. Özellik kümeleri ve sayıları*

| Özellik Kümesi | Özellik Sayısı |
|---|---|
| Karakter bigram | 1162 |
| Karakter trigram | 7686 |
| Kelime unigram | 4838 |
| Kelime bigram | 18166 |
| Tweete özgü özellikler | 6 |
| **TOPLAM** | **31858** |

Tweete özgü özellik kümesinde o tweetin beğeni sayısı, retweet sayısı, tweeti atan kişinin toplam paylaşım sayısı, takipçi sayısı, takip ettiği kişi sayısı ve toplam beğeni sayısı özellikleri mevcuttur.

## 2.3. Özellik Seçimi ve Sınıflandırma

Sınıflandırma deneyleri için WEKA aracı kullanılmıştır. Hem sınıflandırma başarısını yükseltmek hem de deney sürelerini kısaltarak daha hızlı sonuç elde edebilmek için özellik kümeleri üzerinde özellik seçimi yapılmıştır. Özellik seçimi aşamasında WEKA'nın içerisinde hazır olarak tanımlanmış olan *InfoGainAttributeEval* fonksiyonu tercih edilmiştir. Bu fonksiyon, bilgi kazancı değerlerini hesaplayarak özelliklerin değerini belirlemektedir[‡‡‡].

Sınıflandırma aşamasında ise öncelikle özellik kümelerinin çeşitli kombinasyonları oluşturulmuş, özellik seçimi fonksiyonu ile bu kümelerden ilk 1000 özellik seçilmiş ve daha sonra Naive Bayes (NB), Karar Ağacı (J48), SMO ve Rastgele Orman (RF) gibi farklı makine öğrenmesi algoritmalarıyla deneyler gerçekleştirilmiştir. Deneylerde 10-katlı çapraz doğrulama uygulanmış ve veri setinde sınıf dağılımı eşit olmadığı için başarı ölçütü olarak ise F-ölçüm değeri esas alınmıştır.

[†††] https://github.com/imayda/turkish-hate-speech-dataset-1

[‡‡‡] https://weka.sourceforge.io/doc.dev/weka/attributeSelection/InfoGainAttributeEval.html

## 3. Araştırma Sonuçları ve Tartışma

Çeşitli özellik kümesi kombinasyonları içinden özellik seçimi yapılarak elde edilen 1000'er özellik üzerinde farklı makine öğrenmesi algoritmalarıyla gerçekleştirilen deneylerin sonuçları Tablo 4'te verilmiştir.

*Tablo 4. Seçilen 1000'er özellik ile farklı makine öğrenmesi algoritmalarının sınıflandırma performansı*

| | **Algoritma performansı (F-ölçüm)** | | | |
|---|---|---|---|---|
| **Özellik kümesi** | **NB** | **J48** | **SMO** | **RF** |
| Kelime unigram | %68,5 | %66,7 | %71,1 | %69,1 |
| Kelime bigram | %56,8 | - | %68,2 | %68,0 |
| Karakter bigram | %61,2 | %58,1 | %60,5 | - |
| Karakter trigram | %65,8 | %65,2 | %67,3 | %66,9 |
| Kelime bigram +Karakter trigram | %69,0 | %65,7 | %75,8 | %69,0 |
| Kelime unigram +Karakter trigram | %71,2 | %65,1 | %73,1 | %68,4 |
| Kelime unigram +Kelime bigram +Karakter trigram | %70,7 | %65,5 | **%77,8** | %69,8 |
| Kelime unigram +Kelime bigram +Karakter bigram +Karakter trigram | %70,4 | %65,6 | %73,8 | %68,6 |
| Kelime unigram +Kelime bigram +Karakter trigram +Tweet özellikleri | %71,3 | %66,3 | **%77,8** | %69,6 |

Tablo 4'teki sonuçlara bakıldığında bazı deneylerde kötü sınıflandırmadan dolayı duyarlılık (recall) veya kesinlik (precision) değerleri hesaplanamadığı için F-ölçüm yüzdeleri de hesaplanamayıp, deney sonucu olarak "–" yazılmıştır. Deney sonuçlarında genel olarak SMO sınıflandırıcının diğer makine öğrenmesi yöntemlerinden daha iyi sonuç verdiği açıkça görülmektedir.

Özellik kümelerinin tek başlarına kullanıldığında karakter bigramları genelde diğer özellik kümelerine göre daha kötü sonuç vermiştir. En yüksek F-ölçüm değerinin alındığı, kelime unigram, kelime bigram ve karakter trigramların kullanıldığı deneyden sonra özellik kümesine karakter bigramların dâhil edilmesi genel olarak başarıyı düşürmüştür. Bu sonuçlara göre karakter bigramların kullanılmasının sınıflandırma başarısında olumsuz etkisi olduğu söylenebilir.

Diğer yandan, tweete özgü özelliklerin, en yüksek F-ölçüm değerinin alındığı deneyin başarı oranını artırmaya yönelik bir etkisinin olmadığı gözlemlenmiştir. Tweete özgü özellik kümesinesinde çok az sayıda özellik bulunmasının bu etkinin sınırlı kalmasında temel sebep olduğu düşünülebilir.

Farklı makine öğrenmesi algoritmalarıyla yapılan deneylerde en iyi sonucun SMO sınıflandırıcı ile alınması üzerine, bu sınıflandırıcı ile farklı testler de gerçekleştirilmiştir. Farklı sayılarda özellik seçiminin başarı oranı üzerindeki etkisini görmek amacıyla, çeşitli özellik kombinasyonları üzerinde farklı sayılarda özellik seçilerek SMO sınıflandırıcı ile gerçekleştirilen deneylerin sonuçları Tablo 5'te sunulmuştur.

*Tablo 5. Özellik kümeleri üzerinde farklı sayıda özellik seçimi yapılarak SMO ile gerçekleştirilen deneylerin sonuçları*

| | **SMO performansı (F-ölçüm)** | | | | |
|---|---|---|---|---|---|
| **Özellik kümesi** | **İlk 3000 özellik** | **İlk 2000 özellik** | **İlk 1000 özellik** | **İlk 600 özellik** | **İlk 300 özellik** |
| Kelime unigram | %67,1 | %67,1 | %71,1 | %71,5 | %72,4 |
| Kelime bigram | %66,4 | %66,7 | %68,2 | %68,4 | %67,7 |
| Karakter trigram | %64,8 | %68,2 | %67,3 | %70,2 | %74,7 |
| Kelime bigram +Karakter trigram | %73,4 | %75,6 | %75,8 | %76,1 | %76,5 |
| Kelime unigram +Karakter trigram | %71,2 | %71,6 | %73,1 | %75,0 | %77,5 |
| Kelime unigram +Kelime bigram +Karakter trigram | %72,2 | %75,6 | %77,8 | **%79,7** | %77,5 |
| Kelime unigram +Kelime bigram +Karakter trigram +Tweet özellikleri | %71,8 | %75,5 | %77,8 | **%79,9** | %77,4 |

Farklı sayıda özellik seçilerek yapılan deneylerde en iyi sonuçlar, %80'e yakın F-ölçüm değerleri ile 600 özelliğin seçildiği deneyde alınmıştır. Bu sonuçlar, onbinlerce sayıda özellik barındıran özellik kümelerinde özellik seçimi yapılarak daha az sayıda ve daha anlamlı özelliklerle daha başarılı sonuçlar alındığını bir kez daha göstermektedir. En yüksek F-ölçüm değerinin elde edildiği deneyin sonucunda oluşan karmaşıklık matrisi Tablo 6'da verilmiştir.

*Tablo 6. Üç sınıflı veri setinde en yüksek F-ölçüm değerinin alındığı deneye ait karmaşıklık matrisi*

| | | **TAHMİN** | | |
|---|---|---|---|---|
| | **SINIF** | Nefret | Saldırgan | Hiçbiri |
| **GERÇEK** | Nefret | 175 | 3 | 98 |
| | Saldırgan | 9 | 18 | 33 |
| | Hiçbiri | 33 | 14 | 617 |

Literatürde yer alan birçok çalışmada yapıldığı gibi saldırgan ifadeler de nefret söylemi şeklinde değerlendirdiği durumda iki sınıflı hale gelen veri seti üzerinde yapılacak deneylerin başarı oranlarını görebilmek amacıyla, üç sınıflı veri setindeki *Saldırgan* veriler *Nefret* verisine dâhil edilmiş ve SMO sınıflandırıcı ile tekrar deney yapılmıştır. Bu deneylerin sonuçları Tablo 7'deki gibidir.

*Tablo 7. Nefret ve saldırgan veriler birleştirilerek iki sınıflı veri seti üzerinde SMO ile gerçekleştirilen deneylerin sonuçları*

| | SMO performansı (F-ölçüm) | | | | |
|---|---|---|---|---|---|
| **Özellik kümesi** | **İlk 3000 özellik** | **İlk 2000 özellik** | **İlk 1000 özellik** | **İlk 600 özellik** | **İlk 300 özellik** |
| Kelime unigram +Kelime bigram +Karakter trigram | %78,5 | %80,6 | %82,3 | %81,8 | %79,0 |
| Kelime unigram +Kelime bigram +Karakter trigram +Tweet özellikleri | %78,6 | %80,2 | **%83,1** | %82,4 | %78,5 |

İki sınıflı veri seti üzerinde en yüksek F-ölçüm değeri %83,1 ile kelime unigram, kelime bigram, karakter trigram ve tweete özgü özelliklerin bir araya getirilip, özellik seçimi ile ilk 1000 özelliğin seçildiği deney sonucunda alınmıştır. Bu deneyin karmaşıklık matrisi Tablo 8'de verilmiştir.

*Tablo 8. İki sınıflı veri setinde en yüksek F-ölçüm değerinin alındığı deneye ait karmaşıklık matrisi*

| | | TAHMİN | |
|---|---|---|---|
| | **SINIF** | Nefret + Saldırgan | Hiçbiri |
| **GERÇEK** | Nefret + Saldırgan | 212 | 124 |
| | Hiçbiri | 38 | 626 |

Son olarak, *Saldırgan* etiketli verilerin veri setinden çıkarılması durumunda, sadece *Nefret* ve *Hiçbiri* etiketli veriler üzerinde sınıflandırma başarısını görebilmek için 940 tweetten oluşan (276 *Nefret*, 664 *Hiçbiri*) veri seti üzerinde son testler tekrar yapılmıştır. Bu testlerde elde edilen sonuçlar Tablo 9'da sunulmuştur. .

*Tablo 9. Saldırgan veriler veri setinden çıkartılarak iki sınıflı veri seti üzerinde SMO ile gerçekleştirilen deneylerin sonuçları*

| | SMO performansı (F-ölçüm) | | | | |
|---|---|---|---|---|---|
| **Özellik kümesi** | **İlk 3000 özellik** | **İlk 2000 özellik** | **İlk 1000 özellik** | **İlk 600 özellik** | **İlk 300 özellik** |
| Kelime unigram +Kelime bigram +Karakter trigram | %81,6 | %84,9 | %85,4 | %84,6 | %84,8 |
| Kelime unigram +Kelime bigram +Karakter trigram +Tweet özellikleri | %82,0 | %84,6 | **%86,4** | %85,2 | %84,8 |

Tablo 9'da görüldüğü gibi, Saldırgan etiketli veriler çıkartıldığında iki sınıflı veri setinde en yüksek F-ölçüm değeri %86,4 olarak alınmıştır. En iyi sonuç veren bu testin sonucuna ait karmaşıklık matrisi Tablo 10'daki gibidir.

*Tablo 10. 940 tweetten oluşan iki sınıflı veri setinde en yüksek F-ölçüm değerinin alındığı deneye ait karmaşıklık matrisi*

| | | TAHMİN | |
|---|---|---|---|
| | **SINIF** | Nefret | Hiçbiri |
| **GERÇEK** | Nefret | 183 | 93 |
| | Hiçbiri | 30 | 634 |

# 4. Sonuç

Sosyal medya kullanımının artışına paralel olarak çevrimiçi platformlarda daha fazla görülmeye başlanan nefret söylemiyle mücadelede otomatik tespit sistemlerine ihtiyaç duyulmaktadır. Türkçe üzerine kapsamlı bir nefret söylemi tespiti çalışması ise henüz ortaya konmamıştır. Bu yapılan çalışma bu ihtiyaca yönelik olarak sunulmuştur.

Öncelikle farklı hedef gruplara dair 1000 tweet toplanmış ve iki değerlendirici tarafından ayrı ayrı etiketlenmiştir. Farklı etiketlenen tweetler üzerinde değerlendiriciler kendi aralarında anlaşamadıklarında üçüncü değlendiriciye başvurulmuş ve verilerin son etiketi çoğunluğun oyuna göre atanmıştır. Bu veri seti gelecek çalışmalarda kullanılabilmesi için açık kaynak olarak sunulmuştur. Bu veri seti kamuya açık olarak paylaşılan ilk Türkçe nefret söylemi veri seti olma niteliğindedir.

Kelime ve karakter ngram özellikleriyle tweete özgü özelliklerin farklı kombinasyonları oluşturularak elde edilen veri kümelerinden, bilgi kazancına dayalı özellik seçimi yöntemiyle ilk 1000'er özellik seçilmiş ve farklı makine öğrenmesi algoritmaları kullanılarak testler gerçekleştirilmiştir. Bu testlerde en yüksek F-ölçüm değeri %77,8 ile SMO sınıflandırıcının kullanıldığı testlerde alınmıştır. Bu aşamadaki deneylerde karakter bigram özelliklerinin kullanılmasının genel olarak başarıya olumsuz etki ettiği, tweete özgü özelliklerin etkisinin ise sınırlı kaldığı gözlemlenmiştir. Daha sonra, özellik seçimi işleminde farklı sayılarda özellik seçilmesinin başarıya etkisini görmek amacıyla yapılan testlerde ilk 600 özelliğin kullanıldığı deneylerde %79,9 F-ölçüm değeri ile en yüksek başarı oranına ulaşılmıştır.

SMO sınıflandırıcı ile gerçekleştirilen sonraki testlerde, *Saldırgan* ve *Nefret* sınıfına ait verilerin bir arada değerlendirildiği iki sınıflı veri seti üzerinde en yüksek F-ölçüm değeri %83,1 olarak; *Saldırgan* sınıfına ait verilerin tamamen çıkartıldığı 940 tweetlik iki sınıflı veri seti üzerinde ise en yüksek F-ölçüm değeri %86,4 olarak elde edilmiştir.

Bu çalışma nefret söylemi tespitinde hedef grupların çeşitliliği bakımından Türkçe üzerine yapılan ilk kapsamlı çalışma olsa da, çalışmada kullanılan veri setinin boyutunun düşük olduğu söylenebilir. Ayrıca, veri setinin 1000 örnek gibi az sayıda tweetten oluşmasının, gerçekleştirilen testlerde başarı oranının yüksek çıkmamasında en önemli faktör olduğu düşünülmektedir. Bu yüzden gelecek çalışmalarda veri seti boyutunun 5000-10000 arasına yükseltilmesi hedeflenmektedir.

Veri seti boyutunun artırılmasını yanı sıra, deneylerde farklı sınıflandırma algoritmalarının ve özellik seçim yöntemlerinin de denenmesi, ayrıca Doc2Vec, Glove, BERT gibi kelime yerleştirme (word embedding) algoritmalarından elde edilen vektör değerleri ile farklı özellik kümeleri üzerinde sınıflandırma testlerinin yapılması planlanmaktadır.

## 5. Teşekkür

Çalışmada kullanılan veri setinin etiketlenmesi sürecinde sunduğu katkılardan ötürü Yıldız Teknik Üniversitesi Bilgisayar Mühendisliği Bölümü (çift anadal) öğrencisi İrem ATILGAN'a teşekkür ediyoruz.

## Kaynakça

Akın, M. D., & Akın, A. A. (2007, Ağustos). Türk Dilleri İçin Açık Kaynaklı Doğal Dil İşleme Kütüphanesi: Zemberek. *Elektrik Mühendisliği*, (431), 38-44.

Alp, H. (2016). Çingenelere Yönelik Nefret Söyleminin Ekşi Sözlük'te Yeniden Üretilmesi. *Ankara Üniversitesi İLEF Dergisi*, *3*(2), 143-172. https://doi.org/10.24955/ilef.305520

Alp, H. (2018). Suriyeli Sığınmacılara Yönelik Ayrımcı ve Ötekileştirici Söylemin Yerel Medyada Yeniden Üretilmesi. *Karadeniz Teknik Üniversitesi İletişim Fakültesi Elektronik Dergisi*, *5*(15), 22-37.

Arcan, H. E., (2013). Interrupted Social Peace: Hate Speech in Turkish Media. *The IAFOR Journal of Media, Communication and Film*, *1*(1), 43-56. https://doi.org/10.22492/ijmcf.1.1.04

Çelik, E. (2013). Nefret Söylemi İfade Özgürlüğünün Neresinde?. *İnönü Üniversitesi Hukuk Fakültesi Dergisi*, *4*(2), 205-239. https://doi.org/10.21492/inuhfd.239845

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015, Mayıs). *Hate Speech Detection with Comment Embeddings*. Proceedings of the 24th International Conference on World Wide Web (WWW'15) (pp. 29-30). https://doi.org/10.1145/2740908.2742760

Fortuna, P., & Nunes, S. (2018, Temmuz). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys (CSUR)*, *51*(4). https://doi.org/10.1145/3232676

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering Online Hate Speech*. Paris: UNESCO Publishing.

Gelber, K., & McNamara, L. (2016). Evidencing the harms of hate speech. *Social Identities*, *22*(3), 324-341. https://doi.org/10.1080/13504630.2015.1128810

İnsan Hakları Yüksek Komiserliği Ofisi, Birleşmiş Milletler. (1976, Mart 23). *International Covenant on Civil and Political Rights*. https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx

Kaya, S. (2018). Nefret Söyleminin Üretimi Ve Nefret Suçlarının Dolaşıma Girmesinde Facebook'un Etkisi ve Facebook Kullanım Pratiklerine Bakış. *Journal of Social and Humanities Sciences Research (JSHSR)*, *5*(28), 3263-3275. https://doi.org/10.26450/jshsr.735

Kwok, I., & Wang, Y. (2013, Temmuz). *Locate the Hate: Detecting Tweets against Blacks*, Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence (pp. 1621-1622).

Malmasi, S., & Zampieri, M. (2017, Eylül). *Detecting Hate Speech in Social Media*. Proceedings of the International Conference Recent Advances in Natural Language Processing (pp. 467–472). http://dx.doi.org/10.26615/978-954-452-049-6_062

Nockleby, J. T. (2000). Hate Speech. In *Encyclopedia of the American Constitution* (2nd ed.). Macmillan Reference USA.

Razavi, A. H., Inkpen, D., Uritsky, S., & Matwin, S. (2010). Offensive Language Detection Using Multi-level Classification, *The 23th Canadian Conference on Artificial Intelligence* (pp. 16-27). Springer. https://doi.org/10.1007/978-3-642-13059-5_5

Research Centre on Security and Crime (RiSSC). (2017). *An Overview on Hate Crime and Hate Speech in 9 EU Countries*.

Schmidt, A., & Wiegand, M. (2017). *A Survey on Hate Speech Detection using Natural Language Processing*. Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media (pp. 1-10). http://dx.doi.org/10.18653/v1/W17-1101

Şahi, H., Kılıç, Y., & Sağlam, R. B. (2018). *Automated Detection of Hate Speech towards Woman on Twitter*. Proceedings of the 3rd International Conference on Computer Science and Engineering (UBMK'18) (pp. 533-536). https://doi.org/10.1109/UBMK.2018.8566304

Vardal, Z. B. (2015). Nefret Söylemi ve Yeni Medya. *Maltepe Üniversitesi İletişim Fakültesi Dergisi*, *2*(1), 132-156.

Waldron, J. (2014). *The Harm in Hate Speech*. Harvard University Press.

Warner, W., & Hirschberg, J. (2012, Haziran). *Detecting Hate Speech on the World Wide Web*, Proceedings of the 2012 Workshop on Language in Social Media (pp. 19-26).

Waseem, Z. (2016, Kasım). *Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter*. Proceedings of EMNLP 2016 Workshop on Natural Language Processing and Computational Social Science (pp. 138-142). http://dx.doi.org/10.18653/v1/W16-5618

Weber, A. (2009). *Manual on hate speech*. Council of Europe Publishing.