TÜRKÇE YOUTUBE YORUMLARI ÜZERINDE SPAM FILTRELEME

Sevinj SHİRZADOVA

Yüksek Lisans Tezi

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Bilimleri Bilim Dalı

Danışman: Doç. Dr. Alper Kürşat UYSAL

Eskişehir

Eskişehir Teknik Üniversitesi

Lisansüstü Eğitim Enstitüsü

Haziran 2020

ÖZET

TÜRKÇE YOUTUBE YORUMLARI ÜZERINDE SPAM FILTRELEME

Sevinj SHİRZADOVA

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Bilimleri Bilim Dalı

Eskişehir Teknik Üniversitesi, Lisansüstü Eğitim Enstitüsü, Haziran 2020

Danışman: Doç. Dr. Alper Kürşat UYSAL

İnternet kullanımının gün geçtikçe yaygınlaşmasına paralel olarak sosyal medya kullanım oranları da hızla artış göstermektedir. Sosyal medya kullanıcıları tarafından en çok tercih edilen platformlardan biri de YouTube'tur. YouTube kullanımının artması beraberinde bazı problemleri de getirmiştir. Genellikle paylaşılan video içerikleriyle alakası olmayan, reklam amaçlı ve sürekli tekrarlayan istenmeyen (spam) yorumlar boşuna kaynak kullanımına sebep olmaktadır. Bu çalışmada, YouTube yorumları üzerinde istenmeyen yorumların otomatik tespit edilmesi amaçlanmaktadır. Bu amaca ilişkin daha önce yapılan çalışmaların araştırma sonuçları, metin sınıflandırma problemlerinin çözümü için diğer dillerde gerekli sistemler geliştirilse de Türkçe için yapılan çalışmaların oldukça sınırlı olduğunu göstermiştir. Bu tezde Türkçe Youtube yorumlarından oluşan veri setleri oluşturulmuş ve veri setleri üzerinde otomatik metin sınıflandırma algoritmalarının performansları değerlendirilmiştir. Bu tezin önemli bir katkısı da gelecek akademik çalışmalarda kullanılmak üzere erişime açık olacak Türkçe veri setleri oluşturulmuş olmasıdır. Çalışmada, Weka doğal dil işleme aracı kullanılarak doğruluk ve hız açısından iyi sonuçlar veren sınıflandırma algoritmalarının performansları karşılaştırılmıştır. Doğruluk değerleri açısından bakıldığında SMO ve Rastgele Orman makine öğrenimi algoritmaları Türkçe Youtube yorumları sınıflandırma problemi üzerinde diğerlerine göre daha başarılı olarak görünmektedir.

Anahtar Sözcükler: Spam filtreleme, Metin sınıflandırma, YouTube.

ABSTRACT

SPAM FİLTERİNG ON TURKİSH YOUTUBE COMMENTS

Sevinj SHİRZADOVA

Department of Computer Engineering

Programme in Computer Science

Eskişehir Technical University, Institute of Graduate Programs, June 2020

Supervisor: Assoc. Prof. Dr. Alper Kürşat UYSAL

In parallel with the increasing spread of Internet usage, social media usage rates are also increasing rapidly. One of the most preferred platforms by social media users is YouTube. Increased use of YouTube has brought some problems. Repeated spam comments, which are unrelated with shared video content and used for advertising purposes, cause usage of resources unnecessarily in general. This study aims to detect spam comments automatically on YouTube comments. Research results of previous studies performed for this purpose showed that although systems were developed for solving text classification problems in other languages, these kinds of studies for Turkish language were quite limited. In this thesis, datasets consisting of Turkish Youtube comments were created and the performance of automatic text classification algorithms were evaluated on the datasets. An important contribution of this thesis is the creation of Turkish datasets that will be available for use in future academic studies. In the study, the performances of classification algorithms that yield good results in terms of accuracy and speed were compared using the Weka natural language processing tool. In terms of accuracy values, the SMO and Random Forest machine learning algorithms appear to be more successful on the Turkish Youtube comments classification problem than others.

Keywords: Spam filtering, Text classification, YouTube

TEŞEKKÜR

Kıymetli zamanını ayırıp tez danışmanlığımı üstlenen, her zaman bana sabır ve ilgiyle yol gösteren ve çalışmam boyunca önemli katkılarda bulunan çok değerli hocam, Doç. Dr. Alper Kürşat UYSAL'a derin minnettarlığımı bildiririm. Tezin yürütülmesinde değerli fikir ve önerileriyle beni yönlendiren, yardımlarını hiçbir zaman eksik etmeyen saygıdeğer hocalarım Araş. Gör. Gökhan GÖKSEL ve Araş. Gör. Bekir PARLAK'a teşekkür eder, eğitimim sürecinde emeği geçen tüm hocalarıma şükranlarımı sunarım. Ayrıca manevi destekleriyle beni her zaman motive eden ve azimle çalışmaya teşvik eden sevgili arkadaşım Barış DURSUN'a ve canım aileme teşekkürü borç bilirim.

Sevinj SHİRZADOVA

30/06/2020

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Bu tezin bana ait, özgün bir çalışma olduğunu; çalışmamın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamalarında bilimsel etik ve kurallara uygun davrandığımı; bu çalışma kapsamında elde edilen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi; bu çalışmanın Eskişehir Teknik Üniversitesi tarafından kullanılan "bilimsel intihal tespit programı"yla tarandığını ve hiçbir şekilde "intihal içermediğini" beyan ederim. Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçları kabul ettiğimi bildiririm.

Sevinj SHİRZADOVA

İÇİNDEKİLER

Sayfa

TABLOLAR DİZİNİ

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

KISALTMALAR DİZİNİ

APİ : Uygulama programlama arayüzü (App. programming interface)

BOW : Kelime torbası (Bag of Words)

CSV : Virgülle ayrılan değerler dosyası (Comma Separated Values)

DFS : Derinlik öncelikli arama (Depth First Search)

e-posta : Elektron-posta

GainRAE : Kazanım oranı özellik değerlendirici (GainRatioAttributeEval)

Gİ : Gini İndex

İBk : Örnek tabanlı öğrenme algoritması (İnstance Based learning)

ID : Kimlik numarası (Identification number)

IDF : Ters belge frekansı (Inverse Document Frequency)

JSON : JavaScript Nesne Notasyonu (JavaScript Object Notation)

JRip : JRipper- Repeated İncremental Pruning to Produce Error Reduction

kNN : k en yakın komşu (k nearest neighbors)

NBM : NaiveBayesMultinominal sınıflandırıcısı

OneR : Tek sınıf sınıflandırma (One Rule)

REPTree : Reduced Error Pruning Tree

SMO : Sıralı minimal optimizasyon (Sequential Minimal Optimization)

SMS : Kısa mesaj (Short Message)

STK : Sivil Toplum Kuruluşu

TF : Terim frekansı (Term Frequency)

.txt : Metin dosya uzantısı

.arff : Öznitelik dosya biçimi (Attribute file format)

1. GİRİŞ

Web 2.0 teknolojisine geçişle beraber insanların internet kullanım biçimlerinde bazı değişiklikler kendini göstermiştir. Web 1.0 olarak adlandırılan internetin ilk yaygınlaştığı zamanlarda, bilgi web siteleri üzerinden tek taraflı olarak sunulmuş ve kullanıcılar bu bilgiyi sadece tüketmiş ancak kendi fikir ve yorumlarını bildirememişler. Ancak internetin gelişmesiyle, kullanıcıların da sürece katılmasına olanak sağlayan ve çift taraflı bilgi paylaşımını mümkün kılan web siteleri ortaya çıkmaya başlamıştır. Bu durum, Web 2.0 teknolojisine geçişin başlangıcı olmuştur (http-1).

Web 2.0 teknolojisi, sosyal medya platformlarının oluşmasına ve hızla yaygınlaşmasına ortam sağlamıştır. Günümüzde sosyal medya insanların görüşlerini kolayca ifade edebildikleri küresel bir platforma dönüşmüştür. Sosyal medya, internet kullanıcılarının birbirileriyle bilgi, görüş, ilgi alanlarını, görsel ya da işitsel bir şekilde paylaşarak iletişim kurmaları için olanak sağlayan araçlar ve web sitelerini içermektedir (Tunç). Sahip olduğu avantajlar nedeniyle sosyal medya diğer geleneksel medya platformlarından daha fazla tercih edilerek gelişimini sürdürmekte ve kendine yeni özellikler katarak çok hızlı yayılmaktadır.

Sosyal medya platformları kullanım amaçlarına ve hedef kitlesine göre çeşitlilik göstermektedir. Bazıları fotoğraf paylaşımına olanak tanırken, bazıları video içerikleri, bazıları bilgi paylaşımı üzerine kurulmuştur. En çok tercih edilen sosyal medya platformları Facebook, Twitter, Instagram, Youtube, LinkedIn vb. paylaşım ağlarıdır.

"We Are Social 2020" raporuna göre Türkiye'de nüfusun %74'ü internet kullanıcısı, nüfusun %64'ü sosyal medya kullanıcısı ve nüfusun %92'si mobil telefon kullanıcısıdır. Bu da 2019 yılı raporuyla kıyaslandığında internet kullanımında %4, sosyal medya kullanımında %4.2 ve mobil kullanımda %3.4 oranlarında artış olduğunu göstermektedir (http-2).

Sosyal medyanın herkesin söz hakkı olduğu, insanların düşüncelerini özgürce ifade edip topluma etki edebildiği bir platform olması kötü amaçlı kullanıcıların sosyal medyayı olumsuz yönde kullanmalarına neden olmuştur. Bunlara sahte ürün reklamları, kötücül yazılımlar, kötü içerikli linkler vb. örnek olarak gösterilebilir. Genellikle rahatsız edici ve insanlar için tehlike oluşturan bu tür iletiler spam olarak adlandırılır.

Çalışmada, Türkçe Youtube yorumları üzerinde otomatik spam tespiti yapılması amaçlanmıştır. Literatürdeki çalışmalar incelendiğinde konu ile ilgili genelde İngilizce iletiler için makine öğrenme yöntemleri kullanılarak spam tespiti yapıldığı görülmektedir. Bunun yanısıra, Türkçe için yapılan çalışmaların çok daha kısıtlı olduğu söylenebilir. Bunları dikkate alarak Türkçe konusunda bu alandaki literatüre katkıda bulunmak için Türkçe Youtube yorumları içeren veri setleri oluşturulmuş ve bunların üzerinde makine öğrenme yöntemleri ile otomatik spam tespiti yapılmıştır. Bunun yanı sıra veri setleri gelecek akademik çalışmalarda kullanılmak üzere araştırmacıların erişimine açık hale getirilecektir. Veri setleri oluştururken Türkiye'de en çok izlenen 5 Youtube video klibi seçilerek bu kliplere yapılan yorumlar kaydedilmiş ve 5 tane veri seti oluşturulmuştur. Daha sonra her biri üzerinde makine öğrenme algoritmalarının performansları ölçülerek sonuçlar karşılaştırılmış ve böylece daha başarılı sonuç veren algoritmalar tespit edilmiştir.

Tezin ilerleyen bölümleri şu akışta oluşturulmuştur. Tezin ikinci bölümünde literatürdeki konu ile ilgili çalışmalardan bahsedilmektedir. Üçüncü bölümünde ise sosyal medya, Youtube, Türkiye'de Youtube istatistikleri, Youtube'dan verilerin çekilmesi ile ilgili detaylar anlatılmıştır. Dördüncü bölümde çalışmada kullanılan materyal ve yöntemlerden ve uygulama aşamalarından ayrıntılı şekilde bahsedilmiştir. Beşinci bölümde araştırma bulgularından bahsedilmektedir. Altıncı bölümde ise sonuçlar üzerinde tartışmalar, değerlendirmeler ve gelecek çalışmalara dair öneriler bulunmaktadır.

2. KAYNAK ÖZETLERİ

Literatürde spam tepiti ile ilgili çalışmaları incelediğimizde web spam tespiti (Silva, Almeida, & Yamakami, 2012), blog spam tespiti (Romero, Valdez, & Alanis, 2010), e-posta spam tespiti (Ze Li & Shen), sms spam tespiti (Hidalgo, Almeida, & Yamakami, 2012) ile ilgili çok sayıda çalışma olduğunu görmekteyiz. Lakin çalışmalar genelde İngilizce iletilerden ibaret veri seti üzerinde yapılmış olub Türkçe için bu anlamda çalışma sayısı oldukça azdır.

Silva, Almeida ve Yamakami (Silva, Almeida, & Yamakami, 2012) tarafından 2012 yılında yapılan çalışma web yorumları içerisinde spam yorumların tespit edilmesi ile ilgilidir. Bunun için web yorumlarından ibaret bir veri seti kullanılmış ve bu veri setini sınıflandırmak için yapay sinir ağı (Artificial Neural Network) modellerinin performans değerlendirmesi sunulmuştur. Sonuçlar, Levenberg-Marquardt yöntemi ile eğitilmiş Çok Katmanlı Algılayıcı (Multilayer Percepteron) sinir ağının en başarılı model olduğunu göstermektedir. Bu model aynı zamanda Karar Ağaçları (Decision Trees) ve Destek Vektör Makinesi (Support Vektor Machine) gibi yaygın olarak kullanılan ve başarılı olarak bilinen algoritmalardan da daha iyi sonuç vermiştir.

Romero, Valdez ve Alanis (Romero, Valdez, & Alanis, 2010) tarafından 2010 yılında yapılan çalışmada blog yorumları içinde spam yorumları tespit etmek için 4 tane makine öğrenmesi tekniğini karşılaştırılmıştır. Bunlar Sade Bayes (Naive Bayes), k En Yakın Komşu (k Nearest Neighbor), Sinir Ağları (Neural Networks) ve Destek Vektör Makinesi (Support Vector Machine) sınıflandırıcılarıdır. Bu 4 sınıflandırıcıdan Destek Vektör Makinesi yönteminin Sade Bayes ve k En Yakın Komşu yöntemlerine göre daha iyi sonuç verdiği, çok az bir farkla da Sinir Ağları tekniklerinden daha iyi sonuç verdiği tespit edilmiştir.

Li ve Shen (Ze Li & Shen), yapmış oldukları çalışmada istenmeyen elektronik postaların tespiti için yeni bir yöntem önermişler. Bu yönteme kadar en iyi sonuç veren yöntem Bayes filtresi olmuştur ki bu da elektronik postanın spam olup olmadığını bazı anahtar sözcüklere göre ayırt etmektedir. Lakin belli anahtar kelimelerin dışında kelimeler kullanarak bu yöntemin doğruluğunu da kolaylıkla etkilemek mümkün olabilmektedir. Ayrıca Bayes filtresinin her yeni anahtar kelimeyi öğrenip filtreleme yapması belli bir zaman alıyor ki bu da yöntemin eksik yanlarından bir tanesidir. Bu

çalışmada önerilen SOAP ismini verdikleri yöntem Bayes filtresine 3 yeni bileşen entegre edilmesiyle oluşturulmuş ve otomatik filtreleme yapıyor. Bunlar, istenmeyen e-postaları önlemeye, kişiselleştirilmiş spam filtrelemeyi gerçekleştirmeye ve sahte kimliğe bürünme saldırılarını tespit etmeye yardımcı olan bileşenlerdir. Deneysel sonuçlar, SOAP'ın Bayes spam filtrelerinin performansını, spam tespitinin doğruluğu, saldırıya dayanıklılığı ve verimliliği açısından büyük ölçüde geliştirebileceğini göstermektedir.

Hidalgo, Almeida ve Yamakami (Hidalgo, Almeida, & Yamakami, 2012) 2012 yılında yaptıkları çalışmada sms spamların gün geçtikce artmasına ve bunlarla ilgili iyi çalışmalar yapılmadığına dikkat çekilmiştir. Son araştırmalar da SMS spam mesajlarının her geçen yıl önemli ölçüde arttığını ve bunları önlemeye yönelik herhangi çalışma yapılmadığını açıkca gösteriyor. Böyle çalışmaların literatürde bulunmamasının en önemli nedenlerinden biri yeterli verinin olmaması olabilir. Üzerinde çalışmalar yapılabilecek farklı sınıflandırıcılar uygulanarak performansları değerlendirilebilecek ve karşılaştırılabilecek kamuya açık SMS spam veri setlerinin az ve ya yetersizdir. Bu sorunu çözmek için akademik çalışmaların yapılabileceği büyük hacimli, kamuya açık ve gerçek SMS mesajlardan içeren yeni bir veri kümesi oluşturulmuş ve bu veri seti SMS Spam Koleksiyonu olarak adlandırılmıştır. Bu makalede önerilen yeni SMS Spam Koleksiyonunun kapsamlı analizini de yapılmış ve sonuçlar önerilen veri setinin, farklı sınıflandırıcılar tarafından elde edilen performansın değerlendirilmesi ve karşılaştırılmasında kullanılması için güvenilir olduğunu göstermiştir.

Wang, Irani ve Pu (Wang, Irani, & Pu, 2011) tarafından yapılan bu çalışmada Facebook, MySpace ve Twitter gibi sosyal ağların gün geçtikce popülerliğinin artmasına, kullanıcıların milyonlarca insanlara ulaşmak için sosyal ağları tercih etmesine ve aynı zamanda kötü amaçlı kullanıcıların spam gönderileri yaymak için bu ağları kullanmasına değinerek spam filtreleme tekniklerinin önemi vurgulanmıştır. Bu çalışmada tüm sosyal ağ sitelerinde kullanılabilecek bir spam algılama tekniği önerilmiştir. Tekniğin esnekliğini ve uygulanabilirliğini göstermek için de sosyal ağlardan toplanan gerçek veri setleri üzerinde deneysel bir çalışma da yapılmıştır.

Diğer sosyal ağlarda ve e-postalarda yayılan spam'dan farklı olarak YouTube'da yayınlanan spam'ler gerçek kullanıcılar tarafından oluşturularak genelde popüler videolarda kendi kendini tanıtmayı hedeflemektedir. Bu nedenle, bu tür mesajlar meşru

mesajlara benzerliği nedeniyle tespit edilmesi daha zordur. 2013 yılında Chaudhary ve Sureka (Chaudhary & Sureka, 2013) tarafından yapılmış olan bir çalışmada YouTube videolarına yapılan video yorumlar içerisinde spam videoları- kötü içerikli ve ya reklam içerikli videoları tespit etmeye dayalı bir çalışma yapılmıştır. Video spam'leri otomatik olarak tespit edebilecek bir yaklaşım önerilmiştir. Bu yaklaşım tek sınıf sınıflandırıcı – OneR ve ya one class classifier algoritmasına dayanıyor. Video yorumlardan ibaret veri seti oluşturularak önerilen yaklaşım bu veri seti üzerinde denenmiştir. Yapılan deneysel çalışmanın önerilen yaklaşımın %80 oranında doğruluk gösterdiğini ortaya çıkarmıştır.

Khan (Khan, 2019) tarafından 2019 yılında yapılan bu çalışmada YouTube videolarına yapılan video ile ilgisi olmayan, reklam ve takipçi toplama amaçlı alakasız yorumları ve bu yorumları yapan kullanıcıları tespit etmek için bir yaklaşım önerilmiştir. Youtube'un para kazandırma özelliğinin olması onun kullanıcılar tarafından daha çok tercih edilmesine neden olmuştur. Bu aynı zamanda kötü amaçlı kullanıcıları da cezbetmiştir. Bu kullanıcılar otomatik botlar kullanarak videolar altına spam yorumları yapabiliyor ve diğer kanallara abone olarak daha çok kitleye kısa sürede ulaşabiliyor. YouTube farklı farklı yöntemler kullanarak otomatik botların kötü amaçlı yorumlarını tespit edip engellemeye çalışarak bu durumla mücadele ediyor. Ama bu yöntemler yeterince etkili olmamıştır. Bu çalışmada spam ve spam olmayan normal yorumların sınıflandırılması ve böylece sınıflandırma performansını artırmak için farklı teknikler kullanılmıştır.

Alberto, Lochter ve Almeida (Alberto, Lochter, J.V., & Almeida, 2015) 2015 yılında yaptıkları çalışmada YouTube spam yorumları tespit etmek amacıyla YouTube yorumlarından ibaret veri kümeleri oluşturulmuş, bu veri kümeleri üzerinde yaygın sınıflandırma algoritmalarını kullanarak spam tespiti yapılmış ve performanslar karşılaştırılmıştır. Daha sonra elde edilen sonuçlara dayanarak YouTube spam yorumlarını otomatik tespit edebilen TubeSpam ismini verdikleri çevirimiçi bir teknik geliştirilmiştir.

Abd, Altabrawe ve Ajmi (Abd, Altabrawe, & Ajmi, 2018) 2018 yılında yaptığı çalışmada YouTube spam yorumlarını tespit etmek için Yapay Sinir Ağı modelleri kullanmıştır. Burada aynı amaçla daha önce Alberto tarafından yapılan bir çalışmayla karşılaştırma yapılmıştır. O çalışmada TubeSpam adı verdikleri otomatik spam filtreleme yöntemi önerilmiş ve veri seti olarak 5 tane en çok dinlenen şarkının

yorumlarından oluşturulmuş 5 adet veri seti kullanılmıştır. Burada da yine aynı veri setleri kullanılarak Yapay Sinir Ağı modeli kullanılarak sınıflandırma yapılmış ve sonuçlar Alberto'nun yaptığı çalışmadan daha iyi sonuçların elde edildiğini göstermiştir.

Samsudin vd. (Samsudin, Alias, Shamala, Othman, & Din, 2019) 2019 yılında yaptıkları çalışmada Youtube spam yorumların tespiti için Sade Bayes (Naive Bayes) ve Lojistik Regresyon (Logistic regression) sınıflandırma algoritmalarının performanslarını Weka ve Rapid Miner veri madenciliği programlarında karşılaştırmıştır. Weka'da Sade Bayes ve Lojistik Regresyon %87.21 ve %85.29 oranlarında doğru sonuç verirken, Rapid Miner'da Sade Bayes ve Lojistik Regresyon %80.41 ve %80.88 oranlarında doğru sonuç vermiştir. Sonuç olarak Sade Bayes algoritmasının daha iyi sonuç verdiği görülmüştür.

Uysal (Uysal, 2018) tarafından 2018 yılında yapılan çalışmanın amacı YouTube spam filtrelemede 5 adet özellik seçim metodunun 2 farklı sınıflandırma algoritması kullanılarak performanslarının analiz edilmesidir. Sınıflandırıcı olarak Sade Bayes ve Karar Ağaçları algoritmaları kullanılmıştır. Veri seti olarak Alberto'nun sunduğu veri seti kullanılmıştır ki bu da 5 şarkıcının şarkı klibine yapılan İngilizce yorumlardan ibaret 5 veri setidir. Değerlendirme için Macro-F1 başarı ölçütü kullanılmıştır. Performans değerlendirilmesi için 3-katmanlı çapraz doğrulama yöntemi kullanılmıştır. Sonuçlar DFS ve Gİ metotlarının diğer üçüne göre daha iyi sonuç verdiğini göstermiştir. Bununla beraber YouTube spam filtrelemede genellikle Karar Ağaçları algoritmasının Sade Bayes algoritmasından daha iyi sonuç verdiği görülmüştür.

Ali'nin (Ali, 2019) 2019 yılında yaptığı çalışmada YouTube Spam videoların tespit edilmesinin önemi vurgulanarak bu amaçla daha önce yapılan çalışmalar değerlendirilerek kullanılan sınıflandırma tekniklerinin başarı oranları karşılaştırılmıştır. Sonuçların istatistik analizi Çok Katmanlı Algılayıcı (Multilayer Perceptron) ve Destek Vektör Makinesi (Support Vector Machine) algoritmaların en iyi ve en doğru sonuç verdiğini göstermiştir.

Spamla ilgili Türkçe veriler üzerinde yapılan çalışmalara baktığımızda literatürde sınırlı sayıda çalışmaların olduğunu görebiliriz. Yapılan çalışmalar da genelde e-posta spam ((Şahin, 2018), (Altunyaprak, 2006), (Çalış, Gazdağı, & Yıldız, 2013)) SMS

(Örnek, 2019) ve Twitter verileri üzerinde spam tespiti (Karamollaoğlu, 2019) ile ilgilidir. Türkçe YouTube verileri üzerinde hemen hemen hiç çalışma bulunmamaktadır. Bunun nedenlerinden biri de Türkçe YouTube içerikli veri setinin bulunmamasıdır.

Örnek (Örnek, 2019) tarafından 2019 yılında yapılan çalışmada Orange 3 uygulaması kullanarak sms spam mesajları sınıflandırma ve tanımlama için en iyi algoritmayı bulmak amaçlanmıştır. Sms mesajlar haberleşme ve bilgilendirme için hızlı ve etkili bir yol olduğundan bu kötü amaçlar için kullanılması durumları da artarak kullanıcılar için problem oluşturmaktadır. İstenmeyen kandırma amaçlı kötü içerikli ve yanlış bilgi içeren mesajlar gönderilebilmektedir. Bu problemlerin giderilmesi amacıyla bu çalışmada TurkishSMS mesaj ve UCI SMS Spam koleksiyonları kullanılarak Türkçe ve İngilizce içeriklere sahip SMS'ler için spam tespiti yapılmıştır. Çalışmada doğruluk ve hata oranları baz alındığında TurkishSMS veri kümesi için Sinir Ağları, UCI SMS Spam veri kümesi için ise Sade Bayes algoritmasının büyük bir doğruluk ve daha az hata oranına sahip olduğu tespit edilmiştir.

Karamollaoğlu'nun (Karamollaoğlu, 2019) 2019 yılında yaptığı çalışmada sosyal medya platformlarından Twitter üzerinde spam mesajların tespit edilmesi amaçlanmıştır. Makine öğrenmesi modellerinden Vektör Uzay Modeli ile Türkçe ve İngilizce yazılmış tweet'ler üzerinde ayrı ayrı spam tespiti gerçekleştirilmiştir. Bunun için hem Türkçe hem de İngilizce için ayrı ayrı veri setleri oluşturulmuştur. Veri setleri üzerinde Vektör Uzay Modeli ile yapılan spam tespiti sonucunda İngilizce tweet'ler üzerinde %92 oranında, Türkçe tweet'ler üzerinde ise %97 oranında başarı elde edilmiştir. Vektör Uzay Modeli dışında Weka kullanılarak bazı sınıflandırma yöntemlerinin de ilgili problem üzerinde performansları değerlendirilmiştir. Bu yöntemler arasında en başarılı olanları %93 doğruluk oranıyla Sade Bayes (Naive Bayes) ve %94 doğruluk oranıyla Rastgele Orman (Random Forest) yöntemleri olmuştur.

Çıtlak (Çıtlak, 2018) tarafından 2018 yılında yapılan çalışmada sosyal ağların kendi spam hesap politikalarının yetersiz olduğuna dayanarak yeni bir model önerilmiştir. Bu model üç bileşenden oluşuyor ve öğrenmeye dayalıdır. Bu bileşenler link analizi, makine öğrenmesi ve metin analizi bileşenleridir. Link analizinde sosyal ağ kullanıcısının attığı iletilerdeki linkler incelenmekte olup spam link paylaşan sosyal ağ

kullanıcısı spam hesap olarak belirlenir. Link paylaşımı yapmayan spam hesapları tespit etmek için makine öğrenmesi yöntemleri kullanılıyor. Bunun için sosyal ağ üzerindeki paylaşımlardan ibaret bir veri seti oluşturularak, spam hesapların öznitelikleri tespit edilmiş ve bununla makine öğrenme bileşeni modellenmiştir. Metin analizi bileşeninde ise sosyal ağ kullanıcılarının gönderilerindeki metinler incelenerek hassas kelimelerin bulunup bulunmadığına dikkat edilmiştir. Önerilen modelin bir de web tabanlı bir uygulaması gerçekleştirilmiştir. Bu uygulama vasıtasıyla yapılan çalışmalar, önerilen modelin %96,23 oranında başarılı olduğunu göstermiştir.

Şahin (Şahin, 2018) tarafından 2018 yılında yapılan çalışmada e-postaların günümüzde önemine değinerek istenmeyen e-postaların insanlara zarar vermeden önce tespit edilerek önleyebilmenin önemi de ayrıca vurgulanmıştır. İstenmeyen e-postaların tespit edilmesiyle ilgili yapılan önceki çalışmalarla arasındaki fark, bu çalışmada e-posta içerisindeki link metinlerinin analiz edilmesi olmuştur. Link metinlerine göre e-postanın istenmeyen veya normal e-posta olması makine öğrenme yöntemleri ve Kelime Kümesi Tekniği uygulanarak tespit edilmiştir. Tüm deneyler RapidMiner aracı kullanılarak 10 katmanlı çapraz doğrulama ile gerçekleştirilmiştir. Deneyler, Bayes, Destek Vektör Makineleri, Sinir Ağları ve En Yakın Komşu algoritmalarının en iyi sonuç veren, Karar Ağaçları Algoritmalarının düşük başarı gösteren algoritmalar olduğunu göstermiştir. Aynı zamanda %95 üzerinde doğruluk gösteren makine öğrenme tekniklerinin başarısı, Kelime Kümesi Tekniği (BOW) ile elde edilen N gramlar uygulanarak test edilmiş, sonuçlar sınıflandırma performansına en iyi katkı sağlayanın 5 gramlar olduğunu göstermiştir.

Altunyaprak (Altunyaprak, 2006) tarafından 2006 yılında yapılan çalışmada daha önceki e-posta spam filtreme yöntemlerinin çoğunun el ile geliştirilmiş anahtar kelimelere dayanan yöntemler olduğu ve böyle statik filtreleme yöntemlerinin sürekli değişebilen istenmeyen elektronik postaların belirlenmesinde ve filtrelenmesinde yetersiz olduğu vurgulanmıştır. Buna dayanarak sıkça kullanılan ve dinamik bir yaklaşım olan Bayes Filtreleme yöntemi tanıtılmış ve bir uygulama ile sonuçları değerlendirilmiştir.

Çalış, Gazdağı ve Yıldız'ın (Çalış, Gazdağı, & Yıldız, 2013) 2013 yılında yapmış oldukları çalışmada reklam içerikli elektronik posta sorunlarının gün geçtikçe

büyümekle beraber internet trafiğini, aynı zamanda posta sunucularını gereksiz yere meşgul ettiği belirtilmiştir. Bu tür elektronik postaların tespiti için Türkçe yapılan çalışmaların çok az ve yetersiz olması gibi problemler göz önünde bulundurularak metin madenciliği yöntemleri ile Türkçe içerikli reklam e-postaları tespit edilmiştir. Bunun için Destek Vektör Makinesi, k En Yakın Komşu ve Sade Bayes sınıflandırma algoritmalarının başarım oranları değerlendirilmiştir. Deneyler sonucu %96,5 doğruluk oranıyla k En Yakın Komşu (k Nearest Neighbor – kNN) algoritmasının en iyi sonuç verdiği kanaatine varılmıştır.

3. SOSYAL MEDYA VE YOUTUBE

Sosyal medya, Web 2.0 teknolojisine geçişle birlikte oluşan bir medya biçimidir. Sosyal medya, kullanıcılar tarafından oluşturulan bilginin basit, anlık ve çift taraflı olarak paylaşılmasını ve ulaşılmasını sağlayan etkileşimlilik, paylaşımlılık ve katılımcılık olgularını barındıran bir iletişim aracıdır. Aynı zamanda eğlence ve bilgi kaynağıdır (http-3).

Sosyal medyanın geleneksel medyadan aşağıdaki birçok avantajları bulunmaktadır:

- **Maaliyet:** Sosyal medyada daha az maaliyetle hedef kitlelere ulaşmak mümkündür.
- **Etkileşim:** Geleneksel medyada kullanıcılar sadece tüketici rolündeyse sosyal medyada bunun yanı sıra kendi yorumlarını bildirerek ve ya diğerleriyle paylaşarak sürekli etkileşim halindedirler.
- **Zaman:** Sosyal medyada yeni bir haberi anında paylaşmak ve büyük kitlelere duyurmak çok kolaydır.
- **İnteraktiflik:** Geleneksel medyadan farklı olarak sosyal medya çift taraflı iletişimi destekliyor.

Tüm bu pratik yanları sosyal medyanın gün geçtikçe popülerleşmesine ve hızla yayılmasına neden olmuştur. İnsanların sosyal medya platformlarını kullanmalarının birçok amaçları vardır:

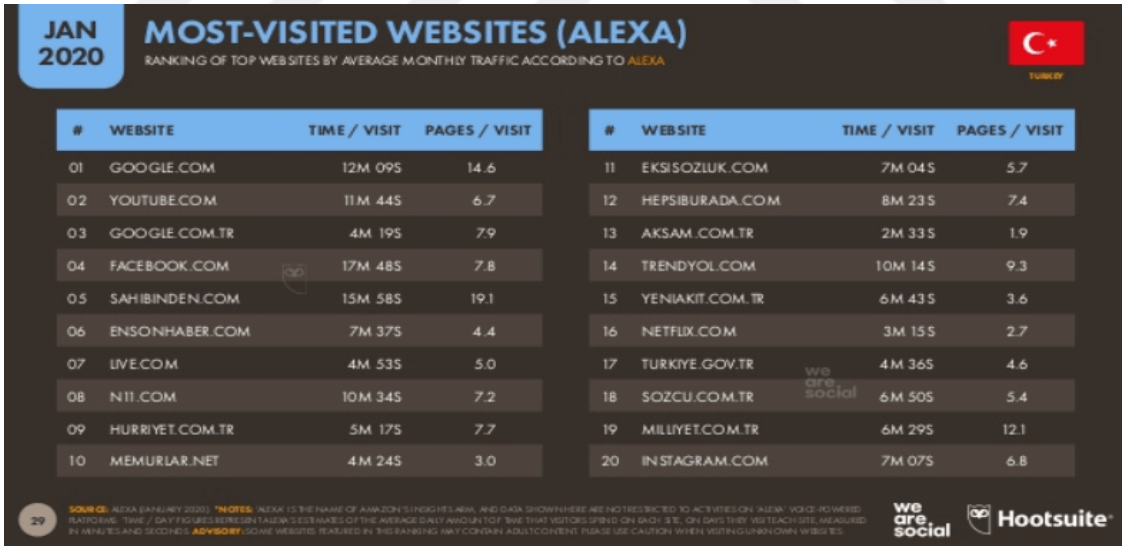
- **Gündemi takip etme:** Sürekli güncel bir akışa sahip olan sosyal medyayı kullanarak dünyada neler olup bittiğini neler değiştiğini öğrenebilirsiniz.
- **Eğlence:** Oyun, sohbet, ilginç yazılar, eğlenceli video ve fotoğraflar ve daha bir çok ilgi alanınıza hitap edebilecek uğraşlarla boş zamanınızı değerlendirebilirsiniz.
- **İletişim**: Normalde bağınızın kopabileceği arkadaş grupları, uzakta yaşayan aile üyeleriniz ve ya ilginizi çeken diğer insanlarla kolayca iletişim kurabilirsiniz.
- **İş olanakları:** Bu amaca hizmet eden platformları da göz önünde bulundurursak sosyal medya iş olanakları sunmakla beraber sosyal medyadaki profiliniz iş görüşmelerinizi etkileyen önemli unsurlardan biridir.

- **Eğitim:** Spordan eğitime, müzikten sanata dair ilgi alanınıza dair hemen her şeye sosyal medya aracılığıyla ulaşabilir, online eğitimlere katılarak bilginizi artırabilirsiniz.

Göründüğü gibi sosyal medya kişiler ve ya kurumlar tarafından doğru kullanıldığında büyük başarılar elde edilebilir. Lakin tüm bu olumlu tarafların yanı sıra kötü amaçlı kullanıcılar sosyal medya aracılığıyla insanlar için tehlike oluşturabilecek, rahatsız edici ve zaman kaybettirici içerikler hazırlayıp paylaşmakla toplumu kötü anlamda etkilemektedirler. Linkler vasıtasıyla yönlendirme yapma, kötücül yazılımları bilgisayara indirmeyi sağlayıp kullanıcı blgilerine erişme vb. örnekler bu tehlikelerden bazılarıdır.

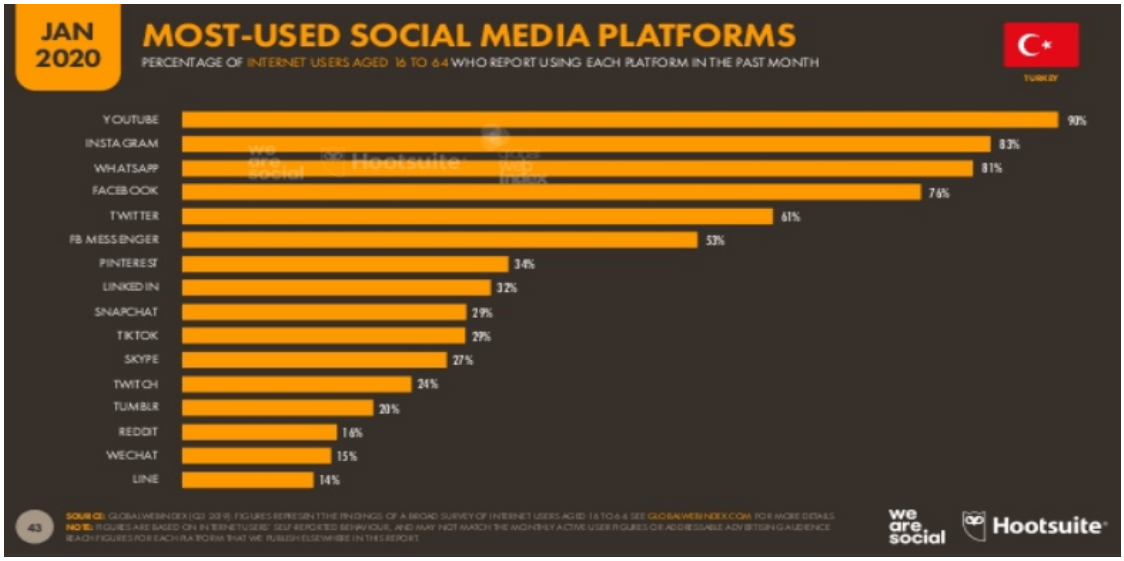
3.1. Sosyal Medya Kullanım Oranları

"We Are Social ve Hootsuite"'in Türkiye için birlikte yayımladığı 2020 yıl raporuna göre en çok ziyaret edilen siteler aşağıdaki gibidir (http-2).



Şekil 3.1. *2019 yılı için Türkiye'de en çok ziyaret edilen siteler*

Şekil 3.1'de görüldüğü gibi Türkiyede en çok ziyaret edilen sitelerin başında Google.com, onun ardından ise YouTube.com en çok ziyaret edilen sitedir.
En çok kullanılan sosyal medya platformları ise Şekil 3.2'de verilmektedir.



Şekil 3.2. *2019 yılı için Türkiye'de en çok kullanılan sosyal medya platformları*

Şekil 3.2'ye baktığımızda son yılda YouTube Türkiye'de en çok kullanılan sosyal medya platformları arasında ilk sırada gelmektedir.

Türkiyede mobil telefon, internet ve sosyal medya kullanım istatistikleri de aşağıdaki gibidir:

Şekil 3.3. *2019 yılı için Türkiye'de internet, sosyal medya ve mobil kullanım istatistikleri*

Şekil 3.3'e bakacak olursak, Türkiye'de 83.88 milyon nüfusun, 77.39 milyonu mobil telefon kullanıcısı, bu da toplam nüfusun %92'si anlamına gelmektedir. Nüfusun %74'üne denk gelen 62.07 milyon internet kullanıcısı ve nüfusun %64'ünü kapsayan 54 milyon aktif sosyal medya kullanıcısı bulunmaktadır.

İnternet, sosyal medya ve mobil kullanım için yıllık büyüme oranları Şekil 3.4'teki gibidir:

Şekil 3.4. *2019 yılı için Türkiye'de internet, sosyal medya ve mobil kullanım istatistiklerinin yıllık değişim oranları*

Şekil 3.4'e bakacak olursak mobil kullanıcı sayısında %3.4, internet kullanıcı sayısında %4 ve sosyal medya kullanıcı sayısında %4.2 artış olduğunu görebiliriz.

Sonuçlar tüm dünyada olduğu gibi Türkiye'de de internet ve sosyal medya kullanım oranlarının gün geçtikçe arttığını, bu artışın avantajlarının yanı sıra oluşabilecek tehlikelerin ve dezavantajların da beraberinde arttığını göstermekle, güvenlik tedbirlerini almanın zorunluluğunu da ifade etmektedir.

3.2. Youtube Ve Spam

YouTube, 2005 yılında PayPal çalışanları tarafından bir video barındırma web sitesi olarak kurulmuş, bir sonraki yıl ise Google tarafından satın alınmıştır. Günümüzde de Google'ın bir yan kuruluşu olarak faaliyetini sürdürmektedir. Google siteyi devraldıktan sonra, eklenen birçok yeni özelliklerle YouTube'in her geçen gün daha fazla ilgi görmesine neden oldu. Google'ın YouTube için seçtiği "broadcast yourself", Türkçe anlamıyla "kendini yayınla" sloganı çok sayıda kullanıcının siteye abone olmasını sağladı. Böylece YouTube, sadece ünlülerin değil, aynı zamanda sıradan insanların da kendine ait videolarını yayınlama imkanı olduğu bir platforma dönüştü.

YouTube'da kullanıcılar video izleyebilir, paylaşabilir ve üyelik alan kullanıcılar kanallarına video yükleyebilirler. Site içi üyelik almayan kullanıcılar ise sadece videoları izleyebilir, izledikleri videoları değerlendirip not verebilir ve yorum

yapabilirler. YouTube'un sunduğu API hizmeti, kullanıcılara YouTube videolarını kendi site ve bloglarında yayınlama imkanı da tanıyor (http-4).

YouTube'un para kazandırma özelliği, kısa sürede kullanıcı sayısının daha çok artmasına neden oldu. Bu avantajlarıyla YouTube kötü niyetli kullanıcıların da hedefi oldu. Bu kullanıcılar daha çok videolar altına video ile ilgisi olmayan, çoğu zaman rahatsız edici yorumlar yapmakla, sahte vaatlerle kullanıcıları yanlış yönlendirme, bilgi çalma gibi amaçlarla kullanıcılar için tehlike oluşturacak biçimde karşımıza çıkmaktadır. YouTube böyle uygunsuz içeriklerin işaretlenmesi ve önlemlerin uygulanması için hem gerçek kişilerden hem de teknolojiden yararlanmaktadır. İşaretlemeler, otomatik işaretleme sistemlerinden, Güvenilir İşaretleyici programının üyelerinden (STK'lar, devlet kurumları ve şahıslar) ve ya genel Youtube topluluğundaki kullanıcılardan gelebilir. Bu işaretlerin incelenmesi için dünya genelinde faaliyet gösteren ekiplerle çalışılmaktadır. Yeterli sayıda kullanıcı bir yorumu spam olarak işaretlerse o yorum gizlenir. Lakin bu yöntemler yorum denetleme için yeterli olmadığından spam hacmi gün geçtikçe artmakta devam ediyor. Şekil 3.5'te YouTube ortamında spam ve normal yorumlara örnekler bulunmaktadır.

Şekil 3.5. *Örnek Youtube yorumları*

YouTube, hem görsel, hem de işitsel temelli olduğundan diğer sosyal medya platformlarına göre etkileşimi daha üst düzeyde tutarak geniş kitlelere hitap etmenizi sağlıyor. YouTube, ilk ortaya çıktığı zamanlarda sadece video yükleme ve paylaşma imkanı sağlasa da, şimdilerde ise haberciler, sinemacılar, müzisyenler, TV kanalları ve hatta özel ve resmi kurumlar tarafından bir televizyon kanalı gibi kullanılan önemli

platforma dönüşmüştür. Canlı yayınlardan interaktif videolara, eğlenceden sanata, hemen her alana dair video içerikler bulmak mümkündür. YouTube'un, video içerikler yayınlamanın yanı sıra bu yayınlarla para kazanma imkanı sağlaması ve bu özelliğinin daha çok kullanıcının siteye abone olmasına neden olması, sitede bulunan spam içeriklerin, sistem tarafından sunulan cazip kazanç imkanıyla doğrudan ilgili olduğunu gösteriyor. E-postalar üzerinden gönderilen spam mesajlar artık daha etkili bir şekilde filtrelenebildiği için sanal dolandırıcılar e-postalar yerine sosyal ağları tercih ederek hem daha büyük kitlelere kolaylıkla ulaşıyor, hem de büyük miktarda para kazanma imkanını da elde etmiş oluyorlar. Sanal dolandırıcıların spam gönderiler için sosyal ağları tercih etmelerinin başlıca nedeni kısa sürede çok sayıda kullanıcıya ulaşması ve para kazandırma imkanı olsa da, diğer bir avantaj bu tür spam içerikleri tespit etmenin zor olmasıdır. Araştırmalar sosyal medyada yayınlanan spam linklerin sadece %15'inin tespit edilebildiğini gösteriyor.

Sanal dolandırıcıların spam gönderileri sosyal ağlarda yaymak için kullandıkları en yaygın yöntemlerden bazıları, "Kazanmak için hemen tıkla" gibi kullanıcıyı üçüncü parti sitelere yönlendiren linkler, sahte hesaplarla ve "Profilinizi kim ziyaret etti, hemen öğrenin" gibi vaatlerle kullanıcıları kandırma amaçlı uygulamalar örnek olabilir. (http-5).

YouTube'da bazı spam yorum örnekleri:

- Ürün tanıtımı yapan anket ve ya hediyelerle ilgili yorumlar
- Yorumlarda " Tıklama başına ödeme" yönlendirme bağlantıları
- Gerçek olmadığı halde tam video içeriğini sunduğunu iddia eden yorumlar
- Kötü amaçlı yazılım ve ya kimlik avı sitesi bağlantıları içeren yorumlar
- Sahte mağazalara bağlantılar içeren yorumlar
- Kanalın ve ya videonun, yorumun gönderildiği videoyla hiçbir ilgisi olmamasına rağmen "Merhaba arkadaşlar, kanalıma/videoma göz atın" gibi yorumlar
- Kanalınızın bağlantısını içeren aynı yorumu tekrar tekrar göndermek

Sosyal medyada spam içeriklerin yol açtığı sorunlar 2010 yılından itibaren ciddi tartışılmaya başlandı. YouTube ile ilgili istenmeyen yorumların hâlâ platformun topluluğuna zarar vermesi, bu tür bir sorunun dikkat ve araştırma gerektirdiğini kanıtlıyor.

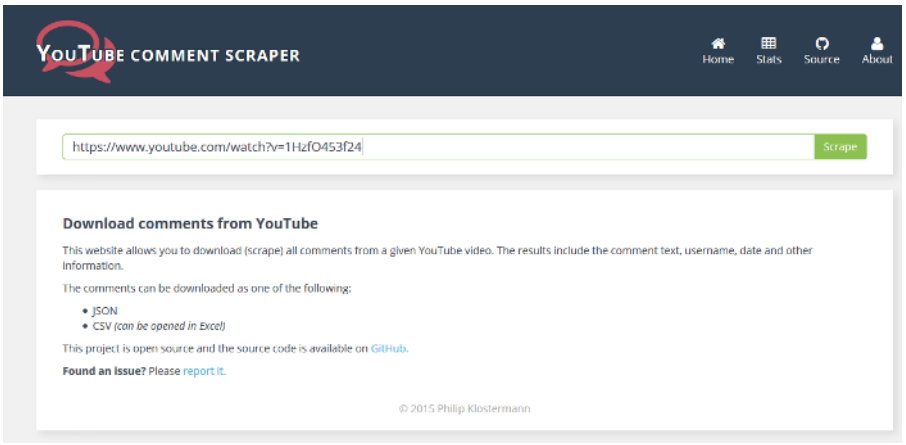
4. MATERYAL VE YÖNTEM

Bu bölümde Türkçe YouTube yorumları spam ve ham olarak sınıflandırma yöntemleri detaylı şekilde ele alınmıştır. Gerekli verinin toplanması, verinin işlenmesi, araç seçimi ve deneylerin yapılması işlemleri anlatılmıştır.

4.1. Veri Kaynağının Toplanması ve Doğal Dil İşleme Aşaması

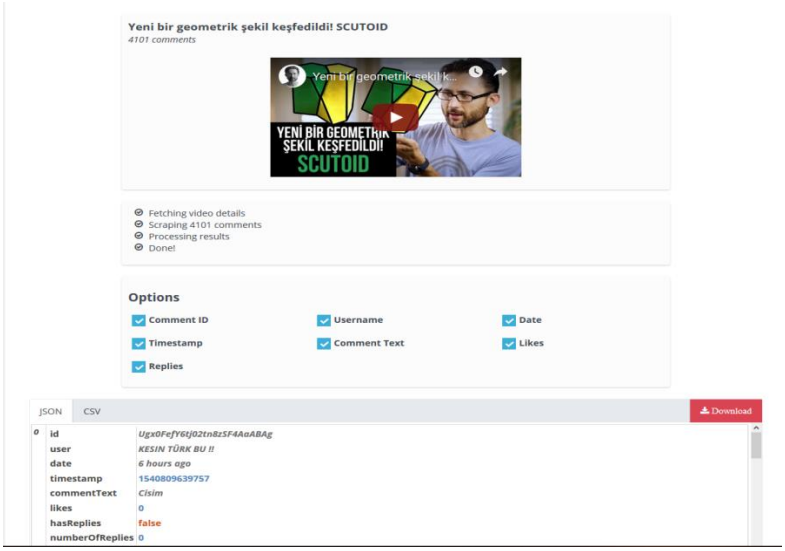
İlgili çalışmalar araştırıldığında Türkçe YouTube yorumları üzerinde herhangi çalışma bulunmamakla beraber Türkçe Youtube yorumlarını içeren veri setinin de olmadığı ortaya çıkmıştır. Bu yüzden çalışmaya önce veri setini oluşturmakla başlanmıştır. Bunun için 2017 yılının en çok izlenen video klipleri içerisinden 5 tanesi seçilerek, bu kliplere yapılan yorumlar indirilerek kaydedilmiştir.

Yorumları kaydetmek için YouTube video yorumlarını CSV ve ya JSON formatında dışarı aktarıp farklı araçlarda analiz etmeye olanak sağlayan Youtube Comment Scraper (YouTube Yorum Ayrıştırıcı) aracı kullanılmıştır. İstenilen videonun linkini kopyalayıp aracın sitesinde ilgili kutucuğa yapıştırıp Scrape tuşuna basmakla yorumların aktarımına başlayabiliriz (http-6). Detayları Şekil 4.1'de görebiliriz.



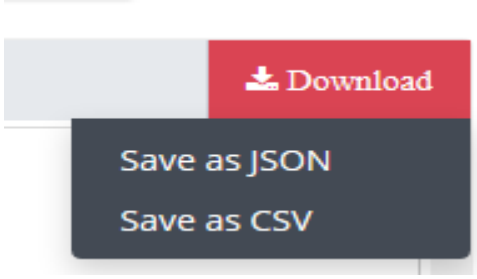
Şekil 4.1. *YouTube Comment Scraper aracı ana penceresi*

Scrape tuşuna bastıktan sonra önümüzde ilgili videoyu, yorum sayısını ve yorumlarla ilgili bazı detayları gösteren pencere açılmaktadır. Şekil 4.2'te yorumları yüklenecek videonun detaylarını gösteren pencerenin arayüzü gösterilmektedir.



Şekil 4.2. *Yorumları yüklenecek videonun detaylarını gösteren pencere*

Options kısmında yorumların detayları ile ilgili bazı seçenekler bulunmaktadır. Burada amacımıza uygun olmayan seçenekleri kaldırarak gereksiz bilgilerin yüklenmesini önleyebilir ve böylece hem hafıza hem de zaman açısından tasarruf edebiliriz. Gerekli seçenekleri de işaretledikten sonra Download tuşuna basarak açılan pencereden indirmek istediğimiz dosya formatını seçiyoruz. Şekil 4.3'te yorumları kaydetmek için sunulan dosya formatı seçenekleri gösterilmektedir.



Şekil 4.3. *Dosya formatı seçim penceresi*

Bu çalışmada yorum metni, yorumcunun ID'si, yayınlanma tarihi, yorumun beğeni sayısı seçenekleri işaretlenerek veri seti CSV dosyası olarak kaydedilmiştir. İndirilen CSV dosyası Excel yazılımı yardımıyla açılmıştır. Bununla yorum ayrıştırma işlemi tamamlanmıştır.

Böylece 5 videoya yapılan yorumlar 5 Excel dosyası olacak şekilde kaydedilmiştir. Daha sonra her bir dosyada yorum metinleri spam ve ham olarak el yordamıyla etiketlenerek .txt dosya formatında ilgili dökümanlara kaydedilmiştir.

Böylelikle spam ve ham yorumlar içeren 5 tane veri seti oluşturulmuştur. Veri seti hakkında detaylar aşağıdaki Tabloda gösterilmiştir.

Tablo 4.1 *Veri kümesi detayları*

Veri seti	Spam	Ham	Toplam
Ece Seçkin	135	269	404
Gülşen	173	194	367
Hadise	145	189	334
Hande Yener	94	179	273
Tarkan	147	184	331

Veri madenciliğinde kullanılan verinin kalitesi sonuçları doğrudan etkiler. Çalışmamızda sınıflandırma algoritmalarının başarımını artırmak adına veri setinin ön işleme tabi tutulması oldukça önemlidir. Çalışmada kullanılan veri ön işleme aşamaları aşağıdaki açıklanmışlardır. Bunlar küçük harflere dönüştürme, noktalama işaretlerin atılması (nokta, virgül, soru vb), durak kelimelerin çıkarılması, dizgeciklere ayırma (tokenization) ve gövdeleme (stemming) işlemleridir.

Küçük harf dönüşümü- metin madenciliği araçları büyük-küçük harf duyarlı olmadığından, büyük-küçük harf farkı olan aynı kelimeler farklı kelimeler gibi algılanıp veri boyutunun gereksiz artmasına ve sınıflandırma algoritmalarının doğruluğunu etkilenmasine neden olabilir. Bu yüzden tüm veriyi küçük harflere dönüştürmekte fayda vardır.

Gereksiz işaretlerin ve durak kelimelerin metinden çıkarılması- Nokta, virgül vb anlam ifade etmeyen işaretler, bunların yanı sıra metinde olması veya olmamasının bilgi çıkarımına herhangi bir etkisi olmayan, veri boyutunu artırıp metin analizini zorlaştıran durak kelimelerin de metinden çıkarılması oldukça önemlidir. Durak kelimeler aşağıda gösterilmiştir:

```
stopwords =("acaba","altmış","altı","ama","ancak","arada","aslında","ayrıca",
            "bana","bazı","belki","ben","benden","beni","benim","beri","beş","bile",
            "bin","bir","birçok","biri","birkaç","birkez","birşey","birşeyi","biz",
            "bize","bizden","bizi","bizim","böyle","böylece","bu","buna","bunda","bundan",
            "bunlar","bunları","bunların","bunu","bunun","burada","çünkü","da","daha",
            "dahi","de","defa","değil","diğer","diye","doksan","dokuz","dolayı","dolayısıyla",
            "dört","edecek","eden","ederek","edilecek","ediliyor","edilmesi","ediyor","eğer",
            "elli","en","etmesi","etti","ettiği","ettiğini","gibi","göre","halen","hala","hangi",
            "hatta","hem","henüz","hep","hepsi","her","herhangi","herkesin","hiç","hiçbir",
            "için","iki","ile","ilgili","ise","işte","itibaren","itibariyle","kadar","karşın",
            "kendi","kendilerine","kendini","kendisi","kendisine","kendisini","kez","ki","kim",
            "kimden","kime","kimi","kimse","kırk","milyar","milyon","mu","mü","mı","mi","nasıl","ne",
            "neden","nedenle","nerde","nerede","nereye","niye","niçin","o","olan","olarak","oldu",
            "olduğu","olduğunu","olduklarını","olmadı","olmadığı","olmak","olması","olmayan","olmaz",
            "olsa","olsun","olup","olur","olursa","oluyor","on","ona","ondan","onlar","onlardan",
            "onları","onların","onu","onun","otuz","oysa","öyle","pek","rağmen","peki","sadece","sanki",
            "sekiz","seksen","sen","senden","seni","senin","siz","sizden","sizi","sizin","şey","şeyden",
            "şeyi","şeyler","şöyle","şu","şuna","şunda","şundan","şunları","şunu","tarafından","trilyon",
            "tüm","üç","üzere","var","vardı","ya","yani","yapacak","yapılan","yapılması","yapıyor","yapmak",
            "yaptı","yaptığı","yaptığını","yaptıkları","yedi","yerine","yetmiş","yine","yirmi","yoksa",
            "yüz","zaten","ve","veya","çok")
```

Şekil 4.4. *Türkçe durak kelimeler listesi*

Dizgeciklere ayırma (Tokenization) - Metnin boşluk, -, vb özelliklere göre parçalara ayrılması işlemidir ki bu parçalara da token denilmektedir (http-7). Bu işlem metin üzerinde bir sıra matematiksel işlemlerin yapılması ve daha kolay analiz edilebilmesi için önemli metin işleme aşamalarından biridir.

Gövdeleme (Stemming) – Metin ön işlemede oldukça önemli bir adımdır. Aynı kökten olup farklı eklerle farklı kelimeler olarak algılanan kelimeler veri boyutunu gereksiz artırmakta ve aslında aynı anlam ifade eden kelimelerin farklı kelimeler gibi değerlendirilerek anlam çıkarımını zorlaştırmakta ve yanlış yönlendirmektedir. Bu açıdan kök bulma işlemi oldukça önemlidir.

Yukarıda sıralanan metin ön işleme aşamalarının gerçekleştirilmesi için İntelliJ İDEA ortamında Java programlama dili kullanılmıştır. Türkçe sondan eklemeli ve çok ek alan bir dil olduğundan doğal dil işleme için diğer dillerin aksine fazla kaynak bulunmamaktadır. Türkçe için en yaygın olan doğal dil işleme aracı Zemberek kütüphanesidir.

Zemberek, Ahmet A. Akın tarafından Java programlama dili kullanılarak geliştirilmiş açık kaynak kodlu Türkçe diller için kullanılabilecek doğal dil işleme kütüphanesidir. Bu kütüphaneyi kullanarak kelime köklerinin bulunması haricinde, heceleme, bir kelimenin Türkçe olub olmadığının kontrol edilmesi, kelime yazım hatalarının düzeltilmesi vb yapılabilir.

Çalışmada bu amaçla Zemberek kütüphanesi (http-8) Java koduna entegre edilerek kullanılmıştır. Veri bahsedilen ön işlemlerden sonra diğer işlemler ve

sınıflandırma için Weka kullanılmışdır. Verinin ön işlemden önceki ve ön işlem uygulandıktan sonraki örneği aşağıdaki şekillerde yer almaktadır:

```
'Küçük kardeşim çok seviyo 10 aylık',1
'Bu Adam harbi adam ',1
'116 bin yorum',1
'Bu arada türkçe yorumdan çok yabancı yorum var bu da tarkanın dünyada ki marka değerini gösteriyor gurur verici',1
'Tebrikler tarkan yorumlar dünya karması gerçekten mega starsın',1
'Samimi söylüyorum hayatımda dinlediğim en guzel sarki',1
'285.368.513 görüntüleme',1
'Yolla hayatı insana anlatan birazda sitem eden şarkı',1
'300 milyon olsun artık',1
'baktım eskiler gelmiş ben de geleyim bari:)',1
'Yorumlara bakindimda hep övgü hep basarisindan bahsedilmis hiç argo laf yok  İşte kalite  budur',1
'Her insan her yükü kaldıramaz kısacası her bünye',1
'bu çok güzel!! Youtube kanalıma izlemeye davetlisiniz!. Teşekkürler',1
'Ders yaparken hep bunu diniyorum',1
'Adam marka dünya tanıyor helal olsun',1
'Yiğenimin sayesine uzun bir aradan sonra tekrar dinledim',1
'300 milyon cok yakisir krala',1
'Kalitenin ozu ve sozude tarkandir varmi itirazi olan',1
'Ablamın kına gecesi giriş müziği',1
'Tarkan yolun acik olsun',1
'turk deyilim ama seviyorum sarkilarini',1
'tarkan abi çok teşekkür ederim çok güzel bir şarkı olumuş ya',1
'Bu şarkıyı izlenmeden çok telefona indiriyorlar',1
'Tarkan Beni çok sev klipsiz 87 milyon benim oldu :D',1
'Dunya yakisiklisi',1
'Adam star megastar. Vessalam',1
```

Şekil 4.5. *Ön işlem uygulanmayan veri kümesi örneği*

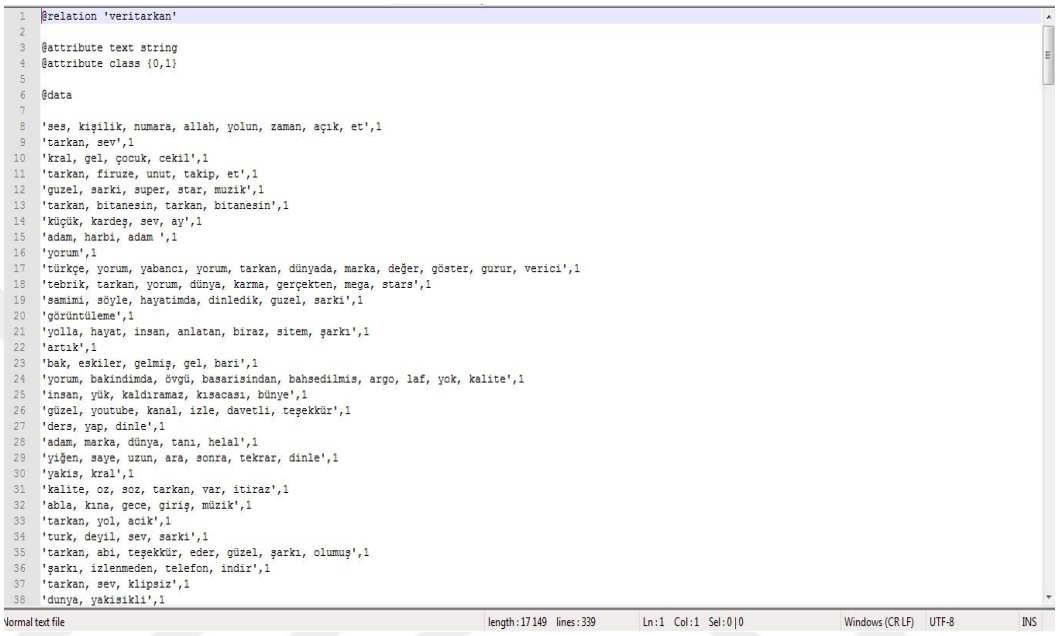
```
'küçük, kardeş, sev, ay',1
'adam, harbi, adam ',1
'yorum',1
'türkçe, yorum, yabancı, yorum, tarkan, dünyada, marka, değer, göster, gurur, verici',1
'tebrik, tarkan, yorum, dünya, karma, gerçekten, mega, stars',1
'samimi, söyle, hayatımda, dinledik, guzel, sarki',1
'görüntüleme',1
'yolla, hayat, insan, anlatan, biraz, sitem, şarkı',1
'artık',1
'bak, eskiler, gelmiş, gel, bari',1
'yorum, bakindimda, övgü, basarisindan, bahsedilmis, argo, laf, yok, kalite',1
'insan, yük, kaldıramaz, kısacası, bünye',1
'güzel, youtube, kanal, izle, davetli, teşekkür',1
'ders, yap, dinle',1
'adam, marka, dünya, tanı, helal',1
'yiğen, saye, uzun, ara, sonra, tekrar, dinle',1
'yakis, kral',1
'kalite, oz, soz, tarkan, var, itiraz',1
'abla, kına, gece, giriş, müzik',1
'tarkan, yol, acik',1
'turk, deyil, sev, sarki',1
'tarkan, abi, teşekkür, eder, güzel, şarkı, olumuş',1
'şarkı, izlenmeden, telefon, indir',1
'tarkan, sev, klipsiz',1
'dunya, yakisikli',1
'adam, star, megastar, vessalam',1
```

Şekil 4.6. *Ön işlem uygulanan veri kümesi örneği*

4.2. Weka

Weka, makine öğrenimi amacıyla Java programlama dili üzerinde geliştirilen ve açık kaynak kodlu olarak kullanıcılara sunulan bir yazılımdır. Waikato Üniversitesinde geliştirilen yazılım ismini de "Waikato Environment for Knowledge Analysis" kelimelerinin baş harflerinden almıştır (http-9). Makine öğreniminde yaygın olarak kullanılan algoritma ve metotların çoğunu içermesi Weka'nın tercih edilme nedenlerindendir. Weka yardımıyla sınıflandırma, bölütleme, ilişkilendirme gibi veri

madenciliği işlemlerinin yanı sıra, veri kümeleri üzerinde veri ön işleme ve görselleştirme de yapılabilmektedir. Weka, ARFF dosya yapısını destekler. Bunun yanı sıra Weka içerisinde CSV dosyalarını da ARFF formatına dönüştürmek mümkün. Çalışmada ön işlemden geçirilen veri kümeleri Wekada kullanabilmek için ARFF dosya formatına çevrilmiştir.



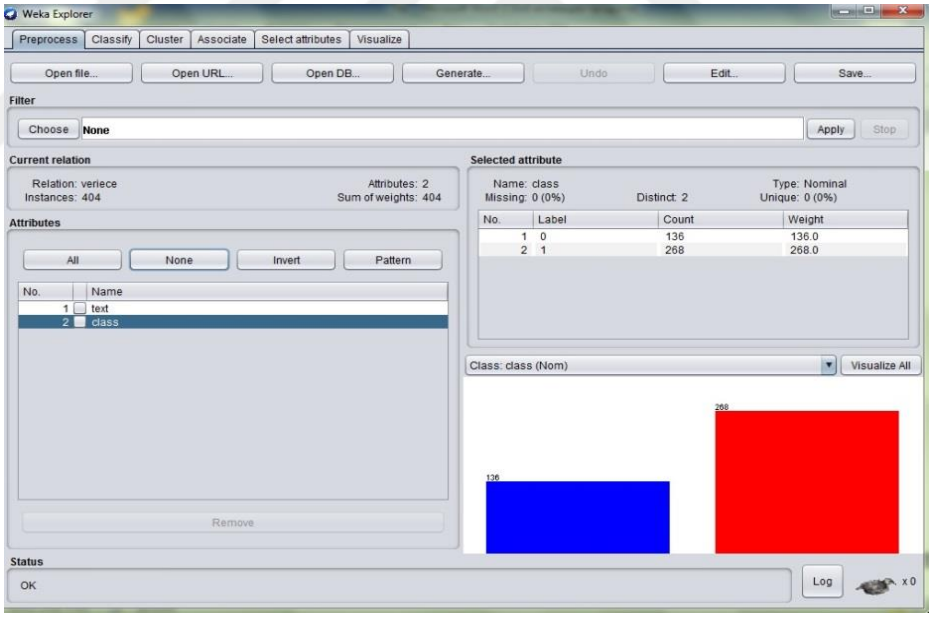
Şekil 4.7. *Örnek .arff dosyasına ait ekran görüntüsü*

ARFF dosya örneğine baktığımızda ilk satırda dosyadaki ilişki tipinin (relation) gösterildiğini, sonraki satırlarda veri kümesindeki özelliklerin (attributes) yazıldığını görebiliriz. Kullandığımız veri kümesinde metin özelliği ve metnin sınıfını gösteren 2 etiket (label) bulunmaktadır. Özelliklerden hemen sonra ise her satır bir örnek olacak şekilde veri kümesi yer alıyor. Her satırın karşısında ise o örneğin etiketi gösterilmekte ve bir birinden virgül ayıracı ile ayrılmıştır. Burada 0 spam, 1 ise ham etiketini ifade etmektedir.

Gerekli formata dönüştürülen dosya artık Weka'da çağrılabilir. Weka çalıştırıldığında aşağıdaki gibi bir ekran açılacaktır.

Şekil 4.8. *Weka ana ekran penceresi*

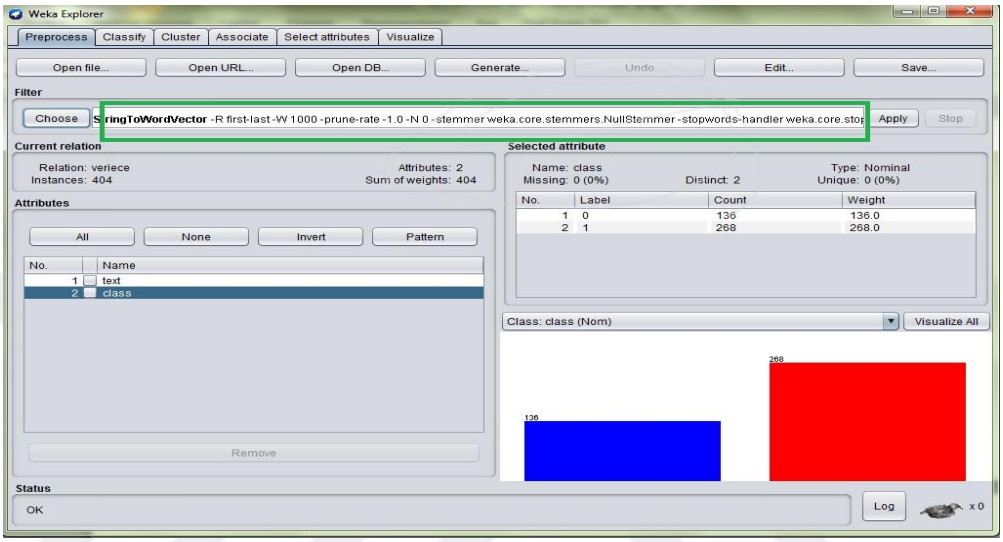
Açılan ekranda veri üzerinde işlemler yapmak için oluşturulan Explorer seçeneğine tıklamakla işleme başlayabiliriz. Açılan ekranda Preprocess sekmesi altındaki Open file düğmesine tıklayıp gerekli dosyayı yüklü olduğu klasörden seçip yüklüyoruz.



Şekil 4.9. *Weka veri dosyası yükleme ve ön işlem penceresi*

Şekilden göründüğü gibi veri dosyasını Weka'ya yükledik. Ekranda veri kümesi hakkında bazı bilgilere ulaşabiliriz. Veri kümesinde 404 tane örnek ve 2 tane sınıf etiketi bulunduğu, sınıfların da 0 ve 1 olarak işaretlendiğini görebiliriz. Daha önce de bahsettiğimiz gibi 0 spam, 1 ise ham sınıfına ait metinleri ifade etmektedir. Böylece mavi renkle gösterilen 136 tane spam içerik ve kırmızı renkle ifade edilen 268 tane ham içerik bulunduğu bilgisini elde edebiliriz.

Sınıflandırmaya geçmeden önce veri kümesi üzerinde bazı hazırlık işlerini tamamlamak gerekiyor. Bu amaçla Filter kısmında bulunan Choose düğmesine tıklayarak açılan pencereden sırayla Filters- unsupervised- attribute- StringToWordVector seçeneklerini seçiyoruz. StringToWordVector filtresi metnin içerisinde bulunan kelimelerin sayısal niteliklere dönüşmesini sağlayan bir fonksiyondur (http-10).



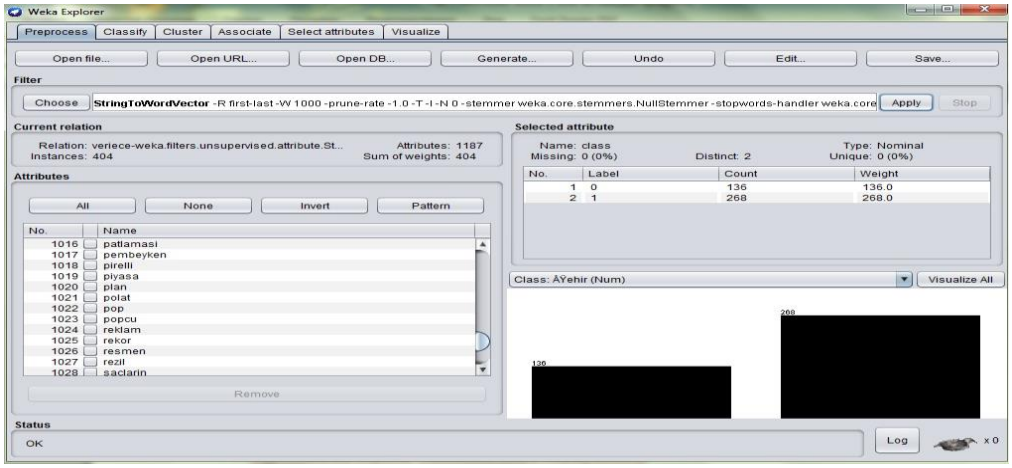
Şekil 4.10. *Weka örnek ön işlem penceresi*

Şekil 4.9'da gösterilen yeşil çerçeve ile işaretlenmiş kutucuğa tıkladığımızda açılan pencereden bazı ayarları değiştirebiliriz. Bu pencerede TFTransform ve IDFTransform sekmeleri karşımıza çıkıyor.

TF- Bir döküman içerisinde geçen terim ağırlıklarını hesaplamak için kullanılan yöntemdir.

IDF-Birden fazla dökümanda kelimenin geçme sayısını bularak bu kelimenin terim olup olmadığını bağlaç, durak kelimesi vb olduğunu anlamaya çalışır (http-11).

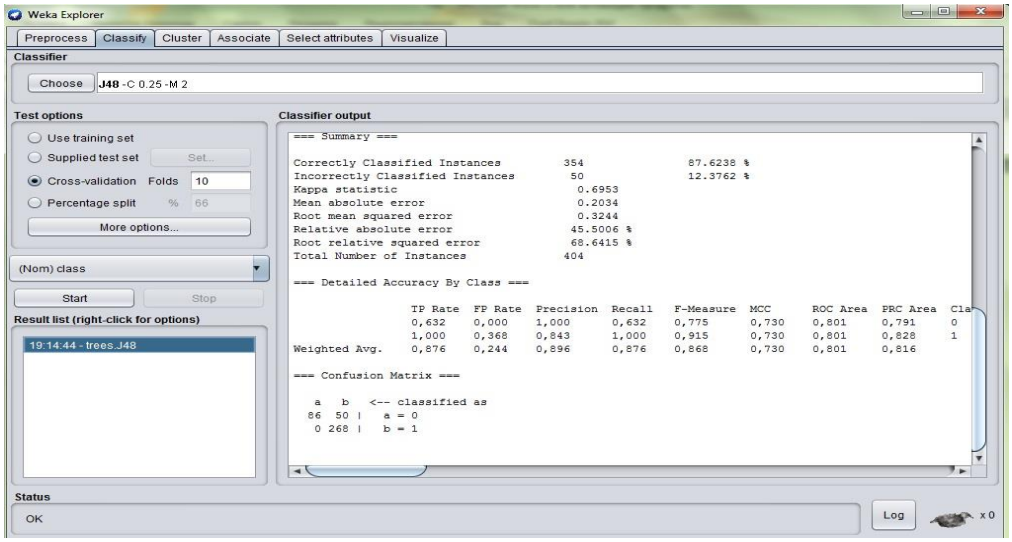
Bu çalışmada TFTransform ve IDFTransform true olarak seçildikten sonra Ok tıklayıp ardından Apply butonuna bastığımızda aşağıdaki gibi bir pencere açılacaktır. Gerekli görmediğiniz kelimeleri seçip Remove butonuna tıklamakla silebilirsiniz.



Şekil 4.11. *Weka'da ön işlem uygulanmış veri kümesini gösteren pencere*

Artık veri kümesi hazır şekle getirildi. Bir sonraki adım olan sınıflandırma aşamasına geçmek için Classify sekmesine tıklayarak uygun sınıflandırma algoritmasını seçerek sınıflandırmaya başlayabiliriz.

Classify sekmesine tıkladığımızda açılan pencerede Classifier kısmında Choose butonuna tıklayarak uygun sınıflandırma algoritmasını seçebiliriz. Test options kısmında test ve eğitim veri kümesini ayarlıyoruz. Çalışmada veri kümesinin %10'luk kısmının eğitim, %90'lık kısmının test için kullanıldığı 10 fold Cross- validation yani 10 katmanlı çapraz doğrulama tekniği seçilmiştir. Seçimler yapıldıktan sonra Start butonuna tıklıyoruz. Sonuçlar sağ tarafda Classifier output kısmında detaylı şekilde görünecektir.



Şekil 4.12. *Weka sınıflandırma sonuç penceresi*

Sınıflama sonuçları penceresinde örneklerden kaç tanesinin doğru kaç tanesinin yanlış sınıflandırıldığı aynı zamanda oranları gösterilmişdir. Bunların yanı sıra sınıflama işlemine ilişkin istatistikler ve karışıklık matrisi de yer almaktadır.

Weka içerisinde 6 kategoride olmak üzere toplam 71 adet algoritmayı barındırır. Bu çalışmada Bayes kategorisinden Bayes ağı (BayesNet), Sade Bayes (NaiveBayes) ve Çok Terimli Sade bayes (NaiveBayesMultinominal) algoritmaları, functions kategorisinden SMO algoritması, lazy kategorisinden IBk algoritması, rules kategorisinden JRip ve Karar Tablosu (DecisionTable) algoritmaları, trees kategorisinden J48, Rastgele Orman (RandomForest) ve RepTree algoritmaları kullanılmıştır.

Bayes Ağı (BayesNet) - Bayes ağı, istatistiksel modelin bir çeşiti olan olasılıksal yönlü dönüşsüz çizge modelidir. Model, birbirileri arasında koşulsal bağımlılıklar bulunan rassal değişkenlerinin yönlü dönüşsüz çizgelerin birleştirilmesi ile ifade edilir. Örneğin bu çalışmada Bayes ağı kullanarak, yorumlar ve içerikleri arasındaki olasılıksal ilişkiler modellenebilir. Bu model kullanılarak, bir yorumun içerisindeki kelimeler verildiğinde bu yorumun spam olub olmama olasılığı hesaplanabilir.

Sade Bayes (Naive Bayes) - Bayes teoremine dayanan olasılıksal sınıflandırma yöntemidir. Daha önce sınıflandırılmış verilerden yaralanarak yeni verinin sınıfını tahmin etmeye çalışır. Örneğin çalışmada daha önceden spam ve ham yorumlar olarak sınıflandırılan veri kümesinden yararlanarak yeni gelen bir verinin spam olup olmadığını tahmin eder.

Çok Terimli Naive Bayes (NaiveBayesMultinominal) - NaïveBayes sınıflandırıcısının bir çeşiti olup genelde belge sınıflandırma problem için kullanılır. Sınıflandırıcının kullandığı öngörücüler belgede bulunan kelimelerin sıklığıdır.

SMO - sıralı minimal optimizasyon algoritması olan SMO algoritması destek vektör makinesi tabanlıdır. Destek vektör makinelerinin eğitim sürecinde optimizasyon problemlerini çözmek için kullanılır. SMO algoritması eksik değerleri değiştirip nominal nitelikleri ikili niteliklere çevirir. Tüm özellikleri varsayılan olarak normalleştirir. Bir iç döngü olarak sayısal kvadratik programlama kullanan standart Destek Vektör Makinesi (Support Vector Machine-SVM) algoritmasından farklı olarak analitik kuadratik programlama tekniği kullanan SMO algoritması ekstra matrise ihtiyaç

duymadan kvadradik programlama problemlerini hızlı bir şekilde çözebiliyor. (Tan, 2018)

IBk - mesafeye dayalı algoritma olub en yakın k-komşu algoritmasıdır. Burada k farklı değerler alabilir. Örneğin k=5 ise, 5 en yakın mesafede olan komşuların sınıfına bakılarak sınıflandırma işlemi yapılmaktadır.

JRip - IREP algoritmasının optimize edilmesiyle oluşturulan bir algoritmadır. Tekrarlayan artımlı budama yöntemiyle hatayı azaltma mantığına dayanan algoritma, kural tabanlı sınıflandırma tekniği kullanır. JRip algoritması önce eğitim kümesini budama ve gelişen kurallar listesi olmak üzere iki alt kümeye ayırır, daha sonra gelişen kurallar listesinde yer alan örneklere göre bir kural oluşturur. Kural modeli oluşturulduktan sonra ise kural listesinin performansını artırmak için budama yapılır. Bu işlemden sonra kurallar listesi artık yeni gelen bir örneğin sınıflandırılması için kullanılabilir. Yeni gelen örnek, listenin en başındaki kuralla karşılaştırılır, eşleşme bulunmadığı takdirde bu karşılaştırma sırayla bir sonraki kurala devredilir ve böylece yeni gelen örneğin ait olduğu sınıf tahmin edilmeye çalışılır. Eğer hiçbir eşleşme bulunmazsa o zaman listenin en altında varsayılan bir kural belirlenir ve sınıflandırılamayan tüm örnekler o sınıfta toplanır. (Özarslan, 2014)

Karar Tablosu (Decision Table) – oluşturduğu karar tablosuna göre sınıflandırma yapan bir algoritmadır. Karar tablosunu ise eğitim kümesindeki verilerin özelliklerine dayanarak oluşturur. Tablolar, üst satırlarda koşullar, alt satırlarda ise onların sonuçları olmak üzere bir matris şeklinde gösteriliyor. Matrisin her sütunu bir kuralı ifade ediyor. JRip algoritması gibi bir kural bulma algoritmasıdır. Yeni gelen örnek, oluşturduğu karar tablosundaki kurallara göre sınıflandırılır. (Turna, 2011)

J48 – bir karar ağacı algoritması olup makine öğreniminde yaygın kullanılan C4.5 algoritmasının Weka uyarlamasıdır. Sınıflandırma mantığı, Divide and Conquer yaklaşımına dayanmaktadır. Veriyi eğitim örneğindeki özellik değerlerine göre aralıklara bölmeyi amaçlar.

J48 algoritması, verileri özyinelemeli olarak sınıflandıran tek değişkenli karar ağacı yapısına sahiptir. Bu yapısıyla çok değişkenli yapıya sahip algoritmalardan daha iyi performans göstermese de tek değişkenli algoritmalar arasında daha iyi performans

gösteren algoritmadır. Ağaç yapısı düğüm ve dallardan oluşup, düğü mler sınıf etiketlerini ifade eder. (Oralhan, 2020)

Rastgele Orman (Random Forest) – J48 algoritmasından farklı olarak tek değişkenli tek bir karar ağacı oluşturmak yerine çok değişkenli çok sayıda karar ağacı oluşturarak sınıflandırma başarısını artırmayı amaçlayan bir algoritmadır. Karar ağaçlarının her biri farklı eğitim verileriyle eğitilerek elde edilen sonuçlar birleştirilir ve yeni gelen örneğin sınıflandırılması bu bireysel sınıflandırma ağaçlarının tahminlerinden elde edilen oylara dayanarak yapılır. RandomForest algoritması büyük hacimli verilerin sınıflandırılmasında da oldukça başarılı bir algoritmadır. Aynı zamanda dengesiz veri kümeleri için hata dengeleme yöntemlerine sahiptir. Kaybolan verilerin çoğunda doğruluğu korumakla beraber bu verilerin tahmin edilmesinde etkili bir yöntemdir.

REPTree – Regresyon ağacı mantığına dayanan, farklı yinelemelerle çok sayıda karar ağacı oluşturup aralarından en iyisini seçmeyi hedefleyen bir karar ağacı algoritmasıdır. REPTree algoritması varyansdan kaynaklanan hatayı en aza indirme ve entropi ile bilgi kazanımı ilkesini benimsemektedir.

Weka'da hem ön işlemden geçirilmemiş veri, hem de ön işlemden geçirilmiş verinin sınıflandırılması ve sınıflandırma sonuçları karşılaştırılmıştır. Daha sonra özellik seçimi metotlarının sınıflandırmaya etkisini araştırmak amacıyla bu metotlar uygulanarak veri üzerinde sınıflandırma yapılmış ve sonuçlar daha öncekilerle karşılaştırılmıştır. Özellik seçiminde amaç, veri setinin tahmin sürecine uygun olmayan özelliklerinin elenerek sınıflandırma performansını düşürmenin önüne geçmektir. Böylece özellik seçim yöntemleri uygulanarak büyük hacimli veriden az hacimli ve tahminde etkili olan veriyi içeren daha güvenilir veri elde edilmiş oluyor. Özetle özellik seçimi sürecini gerekli niteliklerin belirlenip gereksiz niteliklerin veri kümesinden çıkarılması olarak tanımlayabiliriz.

Bu çalışmada Weka'da bulunan özellik seçim metotları kullanılmıştır. Bu metotların performansları detaylı incelendikten sonra bazılarının aynı sonucu verdiği görülmüştür. Aynı sonuç veren metotlar arasından daha hızlı sonuç verenlerin kullanılmasına karar verilmiştir. Böylece özellik değerlendiriciler olarak aşağıdakiler kullanılmıştır:

Korelasyon tabanlı Özellik Altküme Seçim Değerlendirici (Correlation based Feature Subset Selection-CfsSubset Eval) – filtreleme mantığına dayanan bu özellik seçim algoritması, alt özellik kümelerini korelasyon bazlı değerlendirip aralarından en iyi olanı bulmaya çalışır. (Gümüşçü, Aydilek, & Taşaltın, 2016)

Kazanım Oranı Özellik Değerlendirici (GainRatioAttributeEval) – Bir özelliğin değerini, özelliğin sınıfa göre kazanım oranını ölçerek hesaplar (Şık, 2014).

Araştırma yöntemleri olarak aşağıdakiler kullanılmıştır:

Açgözlü Adımsal Araştırma (Greedy Stepwise) – Özellik alt kümesi üzerinde ileriye ve ya geriye doğru açgözlü araştırma yapar. Özellik uzayındaki herhangi bir noktadan başlayabilir. Kalan özelliklerin eklenmesi veya silinmesi değerlendirmede bir düşüşe neden olana kadar devam eder (Şık, 2014).

Sıralayıcı (Ranker) - Her bir özelliği ayrı ayrı değerlendirerek sıralama yapar. GainRatioAttributeEval, ClassifierAttributeEval, OneRAttributeEval gibi özellik değerlendiricileri ile beraber kullanılır.

5. ARAŞTIRMA BULGULARI

Bu bölümde materyal ve yöntem bölümünde anlatıldığı şekilde yapılan deneyler ve sonuçlarından bahsedilmiştir.

Veri madenciliğinde verinin ön işlemden geçirilmesinin sonuçlara etkisini daha açık şekilde gözlemlemek için ön işlem uygulanmayan ve ön işlem uygulanan veri kümesi üzerinde sınıflandırma algoritmaları uygulanmıştır. Weka ile sınıflandırmada değerlendirme metriği olarak doğruluk oranlarının yeterli olacağı düşünülmüştür. 5 veri kümesinin her iki durumu için sonuçlar aşağıdaki Tablolarla ifade edilmiştir.

Tablo 5.1. *Ece ve Ece-S veri kümelerinin sınıflandırma sonuçları*

Veri seti / Sınıflandırma algoritması	Ece				Ece-S			
	D/S	Y/S	Doğruluk yüzdesi	Model oluşum zamanı	D/S	Y/S	Doğruluk yüzdesi	Model oluşum zamanı
J48	335	69	82.92 %	3.68 sn	354	50	87.62 %	1.4 sn
RandomForest	359	45	88.86 %	11.2 sn	378	26	**93.56 %**	6.46 sn
REPTree	364	40	**90.10 %**	3.12 sn	373	31	**92.33 %**	1.73 sn
DecisionTable	360	44	89.11 %	41.3 sn	370	34	91.58 %	18.7 sn
JRip	371	33	**91.83 %**	9.14 sn	376	28	**93.07 %**	5.13 sn
İbk	350	54	86.63 %	0.01 sn	356	48	88.12 %	0.01 sn
SMO	373	31	**92.33 %**	0.37 sn	380	24	**94.06 %**	0.34 sn
BayesNet	362	42	89.60 %	0.9 sn	376	28	**93.07 %**	0.54 sn
NaiveBayes	370	34	**91.58 %**	0.52 sn	378	26	**93.56 %**	0.37 sn
NBM	357	47	88.37 %	0.01 sn	381	23	**94.31 %**	0.01 sn

D/S – Doğru sınıflandırma, Y/S – Yanlış sınıflandırma

Bu Tabloya baktığımız zaman Ece veri kümesinde SMO, JRip, Sade Bayes (Naive Bayes), REPTree algoritmalarının %90 üzerinde doğruluk gösterdiği görülüyor. Bu algoritmalardan SMO algoritması yüksek doğruluk göstermekle beraber, hız açısından da iyi performans göstermiştir. JRip ise Sade Bayes sınıflandırıcısından doğruluk açısından az daha başarılı olsa da hız açısından başarılı olmadığı görünüyor.

Ece-S veri kümesinde de SMO, Çok Terimli Sade Bayes (NaiveBayesMultinominal- NBM), Rastgele Orman (RandomForest), Sade Bayes (Naive Bayes), JRip, Bayes Ağı (BayesNet) algoritmalarının daha yüksek doğruluk gösterdiğini görebiliriz. NBM ve SMO algoritmalarının hem hız, hem de doğruluk açısından başarılı oldukları görülüyor. Sade Bayes ve Rasgele Orman aynı doğruluk oranı gösterse de Sade Bayes daha hızlı sonuç vermiştir. Yine JRip ve Bayes Ağı sınıflandırıcıları aynı doğruluk yüzdesine sahip, hız açısından Bayes Ağı daha başarılıdır.

Genel olarak Tabloya baktığımız zaman sınıflandırıcıların ön işlem uygulanan Ece-S veri kümesi üzerinde, ön işlem uygulanmayan Ece veri kümesine göre daha başarılı olduğu görülüyor.

Tablo 5.2. *Gülşen ve Gülşen-S veri kümelerinin sınıflandırma sonuçları*

Veri seti / Sınıflandırma algoritması	Gülşen				Gülşen-S			
	D/S	Y/S	Doğruluk yüzdesi	Model oluşum zamanı	D/S	Y/S	Doğruluk yüzdesi	Model oluşum zamanı
J48	331	36	90.19%	2.33 sn	353	14	**96.19 %**	1.25 sn
RandomForest	333	34	90.74 %	6.43 sn	351	16	**95.64 %**	4.94 sn
REPTree	335	32	**91.28 %**	1.96 sn	350	17	**95.37 %**	0.91 sn
DecisionTable	337	30	**91.83%**	16.8 sn	352	15	**95.91 %**	9.91 sn
JRip	336	31	**91.55 %**	3.58 sn	352	15	**95.91 %**	2.29 sn
İbk	333	34	90.74 %	0.01 sn	350	17	**95.37 %**	0 sn
SMO	336	31	**91.55 %**	0.49 sn	353	14	**96.19 %**	0.2 sn
BayesNet	329	38	89.65 %	0.61 sn	352	15	**95.91 %**	0.53 sn
NaiveBayes	323	44	88.01%	0.4 sn	348	19	94.82 %	0.29 sn
NBM	327	40	89.10 %	0.01 sn	347	20	94.55 %	0.01 sn

Tablo 5.2'ye baktığımız zaman Gulsen veri kümesi için REPTree, Karar Tablosu (Decision Table), JRip ve SMO algoritmaların doğruluk açısından daha başarılı olduğu görülüyor. SMO algoritması hem doğru, hem de hızlı sonuç vermesiyle diğerlerinden seçiliyor. NBM ve IBk algoritmaları hızlı olsa da doğruluk oranları az daha düşük olmuştur.

Gulsen-S veri kümesinde algoritmaların genelde iyi sonuçlar verdiği görülüyor. SMO, J48, Rasgele Orman, Karar Tablosu, JRip, Bayes Ağı daha yüksek başarı oranı gösteren algoritmalardandır. SMO ve Bayes Ağı daha hızlı sonuç vermesiyle de seçiliyor. İBk, NBM sınıflandırıcıları daha hızlı sonuç vermişler. Ön işlemin uygulandığı veri kümesinde algoritmaların başarı oranlarında büyük farkla artış görünüyor.

Tablo 5.3. *Hadise ve Hadise-S veri kümelerinin sınıflandırma sonuçları*

Veri seti / Sınıflandırma algoritması	Hadise				Hadise-S			
	D/S	Y/S	Doğruluk yüzdesi	Model oluşum zamanı	D/S	Y/S	Doğruluk yüzdesi	Model oluşum zamanı
J48	302	32	90.42 %	1.89 sn	303	31	90.72%	1.16 sn
RandomForest	303	31	**90.72 %**	6.61 sn	314	20	**94.01 %**	4.75 sn
REPTree	302	32	90.42 %	1.67 sn	302	32	90.42%	1.24 sn
DecisionTable	302	32	90.42 %	20.24 sn	306	28	91.62%	13.32 sn
JRip	303	31	**90.72 %**	4.13 sn	310	24	92.81 %	2.65 sn
İbk	299	35	89.52 %	0.04 sn	309	25	92.52 %	0 sn
SMO	306	28	**91.62 %**	0.36 sn	315	19	**94.31 %**	0.21 sn
BayesNet	300	34	89.82 %	0.72 sn	309	25	92.52 %	0.35 sn
NaiveBayes	303	31	**90.72%**	0.38 sn	311	23	**93.11 %**	0.28 sn
NBM	300	34	89.82%	0.01 sn	314	20	**94.012 %**	0 sn

Hadise veri kümesinde algoritmalar yakın sonuçlar göstermişler. SMO, Sade Bayes, JRip, Rastgele Orman sınıflandırıcıları diğerlerine göre nisbeten daha doğru sonuç göstermiş, SMO ve Sade Bayes hızlı sonuç göstermede de başarılı algoritmalardır. NBM, İBk ve Bayes Ağı hızlı sonuç verse de doğruluk açısından biraz düşük sonuç göstermişler.

Hadise-S veri kümesinde SMO, NBM ve Rastgele Orman algoritmaları yüksek doğruluk göstermiş, SMO ve NBM hem de hızlı sonuç göstermişler. IBk, Sade Bayes ve Bayes Ağı hızlı sonuç vermiş, doğruluk oranları nisbeten düşük olmuştur. Genel olarak ön işlemden geçirilen veri üzerinde yapılan sınıflandırma daha başarılı olmuştur.

Tablo 5.4. *Hande ve Hande-S veri kümelerinin sınıflandırma sonuçları*

Veri seti / Sınıflandırma algoritması	Hande				Hande-S			
	D/S	Y/S	Doğruluk yüzdesi	Model oluşum zamanı	D/S	Y/S	Doğruluk yüzdesi	Model oluşum zamanı
J48	240	33	87.91 %	2 sn	257	16	**94.14 %**	1.12 sn
RandomForest	251	22	**91.94 %**	6.45 sn	261	12	**95.60 %**	3.56 sn
REPTree	248	25	90.84 %	1.54 sn	257	16	**94.14 %**	0.74 sn
DecisionTable	246	27	90.11 %	19.83sn	256	17	93.77 %	8.88 sn
JRip	247	26	90.48%	4.19 sn	262	11	**95.97 %**	1.5 sn
İbk	160	113	58.61 %	0.02 sn	225	48	83.52 %	0 sn
SMO	252	21	**92.31%**	0.27 sn	262	11	**95.97%**	0.21 sn
BayesNet	243	30	89.01 %	0.56 sn	258	15	**94.51 %**	0.34 sn
NaiveBayes	250	23	**91.58 %**	0.36 sn	257	16	**94.14%**	0.21 sn
NBM	231	42	84.62 %	0.01 sn	238	35	87.18 %	0.01 sn

Hande veri kümesinde SMO, Rastgele Orman, Sade Bayes diğerlerine göre daha doğru sonuç vermiş, SMO ve Sade Bayes aynı zamanda hızlı sonuç veren algoritmalardır.

Hande-S veri kümesinde Rastgele Orman, JRip, Bayes Ağı daha yüksek doğruluk göstermiş, SMO algoritması yüksek doğruluk göstermekle beraber hızlı da sonuç vermiştir. REPTree, Bayes Ağı, Sade Bayes hızlı sonuç göstermiş, doğruluk oranı da nisbeten iyi olan algoritmalardandır. Diğer veri kümelerinde olduğu gibi burada da ön işlem uygulanan verinin sınıflandırma başarısının daha yüksek olduğu görülüyor.

Tablo 5.5. *Tarkan ve Tarkan-S veri kümelerinin sınıflandırma sonuçları*

Veri seti / Sınıflandırma algoritması	Tarkan				Tarkan-S			
	D/S	Y/S	Doğruluk yüzdesi	Model oluşum zamanı	D/S	Y/S	Doğruluk yüzdesi	Model oluşum zamanı
J48	286	45	86.40 %	1.77 sn	313	18	**94.56 %**	1.62 sn
RandomForest	303	28	91.54 %	6.92 sn	312	19	**94.26 %**	4 sn
REPTree	303	28	91.54%	1.74 sn	313	18	**94.56 %**	0.84 sn
DecisionTable	303	28	91.54%	22.16sn	308	23	93.05 %	12.89sn
JRip	308	23	**93.05 %**	4.67 sn	312	19	**94.26 %**	2.21 sn
İbk	293	38	88.52 %	0 sn	308	23	93.05%	0.02 sn
SMO	307	24	**92.75 %**	0.23 sn	314	17	**94.86 %**	0.34 sn
BayesNet	291	40	87.92 %	0.64 sn	309	22	93.35 %	0.46 sn
NaiveBayes	301	30	90.94%	0.38 sn	317	14	**95.77 %**	0.24 sn
NBM	301	30	90.94 %	0.04 sn	306	25	92.45 %	0.01 sn

Tarkan veri kümesine baktığımız zaman JRip ve SMO algoritmalarının daha yüksek doğruluk oranı gösterdiğini görebiliriz. SMO algoritması doğrulukla beraber hızlı olmasıyla da seçiliyor.

Tarkan-S veri kümesinde ise algoritmaların genellikle yakın performansla yüksek doğruluk gösterdiği görülüyor. SMO ve Sade Bayes hem doğruluk hem de hız açısından daha başarılı sonuç vermiştir. Yine ön işlemden geçirilen veri üzerinde sınıflandırma algoritmaları daha iyi performans göstermişler.

Yapılan sınflandırmaya özellik seçimi metotlarının etkisini araştırmak adına bazı özellikler seçilerek sınıflandırma yapılmıştır. Özellikle veri kümesi çok sayıda özellik içerdiği zaman bazı amaca uygun olmayan özellikler veri kümesinin boyutunu artırarak sınıflandırma performansını düşürebildiği için bu özelliklerin elenmesi gerekli oluyor. Bu amaçla özellik seçim metotları kullanılıyor. Bu çalışmada da Weka'da bulunan metotlar uygulanarak sınıflandırma performansını nasıl etkilediği araştırılmıştır. Bunun için 2 tane özellik değerlendirici ve araştırma yöntemi kombinasyonu kullanılmıştır. Bunlar Korelasyon tabanlı Özellik Altküme Seçim Değerlendiricisi (CfsSubset Eval) ile Açgözlü Adımsal Araştırma (GreedyStepwise) yöntemi ve Kazanım Oranı Özellik

Değerlendirici (GainRatioAttributeEval) ile Sıralayıcı (Ranker) araştırma yöntemi kombinasyonlarıdır. Bu özellik seçimi yöntemleri kullanılarak yapılan sınıflandırma sonuçları aşağıdaki Tablolarde ifade edilmiştir.

CfsSubsetEval+GreedyStepwise kombinasyonu çalıştırıldığında, default olarak her veri kümesi için en etkili özellikleri seçip sınıflandırma yaptığı görülmüştür.

GainRatioAttributeEval+Ranker kombinasyonu çalıştırıldığında default olarak tüm özellikleri sıralayıb seçtiği ve ondan dolayı normal sınıflandırmadan farklı sonuç vermediği görülmüştür. Daha sonra özellik seçimi sayısı değiştirilerek sınıflandırma performansı ölçülmüş ve genelde tüm veri kümeleri için seçilen özellik sayısı 100 olduğunda sınıflandırma algoritmalarının iyi sonuç gösterdiği kannatine varılmıştır.

Tablo 5.6 *Ece veri kümesinin özellik seçim metotları ile sınıflandırma sonuçları*

Veri seti / Sınıflandırma algoritması	Ece					
	CfsSubsetEval + GreedyStepwise (default)	Model Oluşum Zamanı	GainRAE + Ranker (default)	Model Oluşum Zamanı	GainRAE + Ranker (seçilen özellik sayısı: 100)	Model Oluşum Zamanı
J48	81.19 %	2.15 sn	82.92 %	3.3 sn	81.19 %	0.85 sn
RandomForest	91.58%	2.74 sn	88.86 %	12.12 sn	**93.07 %**	2.16 sn
REPTree	89.60 %	2.32 sn	90.10 %	3.88 sn	89.85 %	0.71 sn
DecisionTable	89.60 %	3.19 sn	89.11 %	43.9 sn	**90.35 %**	2.44 sn
JRip	91.34%	1.84 sn	91.83 %	11.71 sn	**92.08 %**	1.11 sn
İbk	91.34 %	2.06 sn	86.63 %	0.81 sn	**91.83 %**	0.73 sn
SMO	91.34 %	1.91 sn	92.33 %	1.01 sn	**92.57 %**	1.09 sn
BayesNet	88.61 %	1.95 sn	89.60 %	1.47 sn	**89.85 %**	0.78 sn
NaiveBayes	89.36 %	1.79 sn	91.58 %	1.25 sn	91.34 %	0.74 sn
NBM	90.59 %	2.01 sn	88.37 %	0.98 sn	**91.09 %**	0.75 sn

Sonuçlara baktığımızda, genelde 100 tane özellik için uygulanan GainRAE+Ranker kombinasyonunun sınıflandırma başarısını artırdığı görülüyor. Rastgele Orman (RandomForest) algoritması 88.86 % başarı oranı, özellik seçimiyle bu oran 93.07 % olmuştur. Onun dışında iyi performans gösteren SMO ve JRip algoritmalarının özellik seçimiyle performansları daha da artmıştır.

Tablo 5.7. *Ece-S veri kümesinin özellik seçim metotları ile sınıflandırma sonuçları*

Veri seti / Sınıflandırma algoritması	Ece-S					
	CfsSubsetEval + GreedyStepwise (default)	Model Oluşum Zamanı	GainRAE + Ranker (default)	Model Oluşum Zamanı	GainRAE + Ranker (seçilen özellik sayısı: 100)	Model Oluşum Zamanı
J48	87.38 %	1.18 sn	87.62 %	2.23 sn	87.38 %	0.61 sn
RandomForest	92.08 %	1.86 sn	93.56 %	8.73 sn	**94.55 %**	2.19 sn
REPTree	90.84 %	1.33 sn	92.33 %	2.1 sn	92.08 %	0.71 sn
DecisionTable	89.11 %	1.16 sn	91.58 %	22.13 sn	**92.33 %**	1.95
JRip	91.09 %	0.95 sn	93.07 %	5.19 sn	92.82 %	1.03 sn
İbk	91.58 %	1.14 sn	88.12 %	0.48 sn	**93.07 %**	0.43 sn
SMO	91.09 %	1.31 sn	94.06 %	0.88 sn	**94.55 %**	0.64 sn
BayesNet	89.36 %	1.12 sn	93.07 %	1.22 sn	**93.32 %**	0.53 sn
NaiveBayes	89.60 %	1.19 sn	93.56 %	0.9sn	92.57 %	0.61 sn
NBM	91.34 %	1.32 sn	94.31 %	0.6 sn	94.31 %	0.44 sn

Bu veri kümesinde de yine genelde GainRAE+Ranker kombinasyonu 100 tane özellik seçiminde sınıflandırma başarısını artırmıştır. En iyi sonuç gösteren Rastgele Orman ve SMO algoritmaları olmuştur.

Tablo 5.8. *Gülşen veri kümesinin özellik seçim metotları ile sınıflandırma sonuçları*

Veri seti / Sınıflandırma algoritması	Gülşen					
	CfsSubsetEval + GreedyStepwise (default)	Model Oluşum Zamanı	GainRAE + Ranker (default)	Model Oluşum Zamanı	GainRAE + Ranker (seçilen özellik sayısı: 100)	Model Oluşum Zamanı
J48	90.19 %	2.03 sn	90.19%	2.83 sn	90.19 %	0.64 sn
RandomForest	91.01 %	2.45 sn	90.74 %	8.9 sn	**91.83 %**	2.25 sn
REPTree	91.28 %	2.39 sn	91.28 %	2.87 sn	91.28 %	0.45 sn
DecisionTable	91.83 %	2.91 sn	91.83%	24.53 sn	91.83%	2.08 sn
JRip	**91.83 %**	2.36 sn	91.55 %	5.45 sn	**91.83 %**	0.82 sn
İbk	90.19 %	2.32 sn	90.74 %	0.58 sn	90.19 %	0.57 sn
SMO	91.01 %	2.29 sn	91.55 %	1 sn	**92.10 %**	0.63 sn
BayesNet	**90.19 %**	2.22 sn	89.65 %	0.92 sn	89.65 %	0.71 sn
NaiveBayes	**89.37 %**	2.02 sn	88.01%	0.89 sn	87.47 %	0.41 sn
NBM	88.83 %	2.5 sn	89.10 %	0.71 sn	88.01 %	0.64 sn

Bu Tabloda bazı durumlarda CfsSubsetEval+GreedyStepwise, bazı durumlarda ise GainRAE+Ranker sınıflandırma başarısını artırmıştır. En iyi sonuç gösteren SMO algoritması olmuştur.

Tablo 5.9. *Gülşen-S veri kümesinin özellik seçim metotları ile sınıflandırma sonuçları*

Veri seti / Sınıflandırma algoritması	Gülşen-S					
	CfsSubsetEval + GreedyStepwise (default)	Model Oluşum Zamanı	GainRAE + Ranker (default)	Model Oluşum Zamanı	GainRAE + Ranker (seçilen özellik sayısı: 100)	Model Oluşum Zamanı
J48	96.19 %	1.11 sn	96.19 %	1.93 sn	96.19 %	0.5 sn
RandomForest	**96.19 %**	1.29 sn	95.64 %	6.57 sn	95.91 %	2.71 sn
REPTree	**95.64 %**	1.27 sn	95.37 %	1.49 sn	**95.64 %**	0.52 sn
DecisionTable	95.91 %	1.29 sn	95.91 %	13.74 sn	95.91 %	1.62 sn
JRip	**96.19 %**	1.22 sn	95.91 %	2.87 sn	95.91 %	0.55 sn
İbk	**95.64 %**	1.06 sn	95.37 %	0.35 sn	94.55 %	0.51 sn
SMO	96.19 %	1.1 sn	96.19 %	0.57 sn	95.91 %	0.6 sn
BayesNet	95.91 %	1.2 sn	95.91 %	0.76 sn	95.91 %	0.69 sn
NaiveBayes	**95.64 %**	1.44 sn	94.82 %	0.62 sn	94.28 %	0.41 sn
NBM	**94.82 %**	1.16 sn	94.55 %	0.45 sn	90.74 %	0.4 sn

Bu veri kümesinde genelde CfsSubsetEval+GreedyStepwise özellik seçim metodunun sınıflandırma başarısını artırdığı görünüyor. En iyi sonuç göstrenler Rastgele Orman ve JRip algoritmaları olmuştur. Lakin aynı başarı oranını SMO ve J48 algoritmaları özellik seçimi uygulanmadan da göstermişler.

Tablo 5.10. *Hadise veri kümesinin özellik seçim metotları ile sınıflandırma sonuçları*

Veri seti / Sınıflandırma algoritması	Hadise					
	CfsSubsetEval + GreedyStepwise (default)	Model Oluşum Zamanı	GainRAE + Ranker (default)	Model Oluşum Zamanı	GainRAE + Ranker (seçilen özellik sayısı: 100)	Model Oluşum Zamanı
J48	89.22 %	1.62 sn	90.42 %	3.15 sn	90.42 %	0.7 sn
RandomForest	89.82 %	2.36 sn	90.72 %	7.87 sn	**91.02 %**	5.2 sn
REPTree	88.92 %	2.02 sn	90.42 %	1.83 sn	90.42 %	1.31 sn
DecisionTable	88.92 %	2.13 sn	90.42 %	27.47 sn	90.42 %	12.73 sn
JRip	89.22 %	2.2 sn	90.72 %	4.77 sn	**91.32 %**	0.86 sn
İbk	88.92 %	1.62 sn	89.52 %	0.59 sn	**91.62 %**	0.5 sn
SMO	89.82 %	2.42 sn	91.62 %	0.75 sn	**91.92 %**	0.62 sn
BayesNet	88.32 %	2.31 sn	89.82 %	1.18 sn	89.82 %	0.63 sn
NaiveBayes	88.92 %	1.79 sn	90.72%	0.67 sn	90.72%	0.65 sn
NBM	88.62 %	2.25 sn	89.82%	0.59 sn	**91.02 %**	0.53 sn

Normal sınıflandırmada en iyi sonuç gösteren SMO algoritması, özellik seçimi uyglandıktan sonra da yine az bir artışla başarı oranı en yüksek olan algoritma olmuştur.

Tablo 5.11. *Hadise-S veri kümesinin özellik seçim metotları ile sınıflandırma sonuçları*

Veri seti / Sınıflandırma algoritması	Hadise-S					
	CfsSubsetEval + GreedyStepwise (default)	Model Oluşum Zamanı	GainRAE + Ranker (default)	Model Oluşum Zamanı	GainRAE + Ranker (seçilen özellik sayısı: 100)	Model Oluşum Zamanı
J48	88.62 %	1.09 sn	90.72%	1.49 sn	90.12 %	0.41 sn
RandomForest	89.82 %	1.88 sn	94.01 %	5.28 sn	94.01 %	1.39 sn
REPTree	89.52 %	1.34 sn	90.42%	1.59 sn	**91.02 %**	0.55 sn
DecisionTable	89.82 %	1.47 sn	91.62%	13.64 sn	91.62%	1.69 sn
JRip	89.82 %	1.49 sn	92.81 %	3.93 sn	**93.41 %**	0.89 sn
İbk	89.22 %	1.24 sn	92.52 %	0.45 sn	**93.41 %**	0.45 sn
SMO	89.82 %	1.31 sn	94.31 %	0.58 sn	93.11 %	0.39 sn
BayesNet	89.52 %	1.3 sn	92.52 %	1.03 sn	92.52 %	0.39 sn
NaiveBayes	89.52 %	1.21 sn	93.11 %	0.63sn	93.11 %	0.46 sn
NBM	88.62 %	1.3 sn	94.01 %	0.29 sn	93.71 %	0.41 sn

Bu veri kümesinde özellik seçim algoritmaları başarı yüzdesini çok fazla etkilememiştir. JRip ve İBk algoritmalarının başarı oranı, özellik seçiminden sonra arttığı görünüyor. Lakin özellik seçimi uygulanmadan SMO algoritması daha yüksek performans göstermiştir.

Tablo 5.12. *Hande veri kümesinin özellik seçim metotları ile sınıflandırma sonuçları*

Veri seti / Sınıflandırma algoritması	Hande					
	CfsSubsetEval + GreedyStepwise (default)	Model Oluşum Zamanı	GainRAE + Ranker (default)	Model Oluşum Zamanı	GainRAE + Ranker (seçilen özellik sayısı: 100)	Model Oluşum Zamanı
J48	84.62 %	3.3 sn	87.91 %	2.4 sn	87.91 %	0.48 sn
RandomForest	90.84 %	3.28 sn	91.94 %	7.17 sn	**93.04 %**	1.76 sn
REPTree	90.48 %	2.83 sn	90.84 %	1.81 sn	**91.58 %**	0.47 sn
DecisionTable	90.48 %	3.11 sn	90.11 %	17.08 sn	**90.84 %**	1.61 sn
JRip	90.48 %	2.74 sn	90.48%	4.79 sn	**91.94 %**	0.84 sn
İbk	89.01 %	2.76 sn	58.61 %	0.44 sn	**90.48 %**	0.44 sn
SMO	90.84 %	3.14 sn	92.31%	0.68 sn	**92.68 %**	0.49 sn
BayesNet	87.91 %	2.57 sn	89.01 %	1.08 sn	89.01 %	0.5 sn
NaiveBayes	89.01 %	2.88 sn	91.58 %	0.65 sn	**91.94 %**	0.43 sn
NBM	89.38 %	3.22 sn	84.62 %	0.52 sn	**90.84 %**	0.46 sn

Bu veri kümesi üzerinde yapılan sınıflandırmada GainRAE+Ranker özellik seçim metodu başarı yüzdesini artırmakta daha etkili olmuştur. En iyi sonucu Rastgele Orman ve SMO algoritmaları göstermiştir. Bu algoritmaların doğruluk yüzdesi, özellik seçim metoları uygulanmadan önceki sınıflandırmaya kıyasla daha başarılı olmuştur.

Tablo 5.13. *Hande-S veri kümesinin özellik seçim metotları ile sınıflandırma sonuçları*

Veri seti / Sınıflandırma algoritması	Hande-S					
	CfsSubsetEval + GreedyStepwise (default)	Model Oluşum Zamanı	GainRAE + Ranker (default)	Model Oluşum Zamanı	GainRAE + Ranker (seçilen özellik sayısı: 100)	Model Oluşum Zamanı
J48	93.77 %	0.49 sn	94.14 %	1.12 sn	94.14 %	0.41 sn
RandomForest	95.60 %	0.93 sn	95.60 %	4.04 sn	**96.70 %**	1.38 sn
REPTree	94.14 %	0.59 sn	94.14 %	0.89 sn	**94.51 %**	0.43 sn
DecisionTable	93.77 %	0.66 sn	93.77 %	9.35 sn	93.77 %	1.76 sn
JRip	95.60 %	0.64 sn	95.97 %	1.62 sn	95.97 %	0.47 sn
İbk	**95.24 %**	0.72 sn	83.52 %	0.3 sn	94.14 %	0.22 sn
SMO	95.60 %	0.87 sn	95.97%	0.32 sn	**96.34 %**	0.54 sn
BayesNet	92.67 %	0.56 sn	94.51 %	0.58 sn	94.51 %	0.29 sn
NaiveBayes	**95.24 %**	0.64 sn	94.14%	0.47sn	94.51 %	0.28 sn
NBM	**94.14 %**	0.68 sn	87.18 %	0.2 sn	91.58 %	0.21 sn

Bu veri kümesinde de yine SMO ve Rastgele Orman algoritmaları hem özellik seçimi uygulanmadan önce, hem de özellik seçimi uygulandıktan sonra başarı yüzdesi en yüksek olan algoritmalar olmuşlar.

Tablo 5.14. *Tarkan veri kümesinin özellik seçim metotları ile sınıflandırma sonuçları*

Veri seti / Sınıflandırma algoritması	Tarkan					
	CfsSubsetEval + GreedyStepwise (default)	Model Oluşum Zamanı	GainRAE + Ranker (default)	Model Oluşum Zamanı	GainRAE + Ranker (seçilen özellik sayısı: 100)	Model Oluşum Zamanı
J48	85.50 %	1.63 sn	86.40 %	1.93 sn	85.80 %	0.81 sn
RandomForest	91.24 %	1.44 sn	91.54 %	6.34 sn	**93.36 %**	1.84 sn
REPTree	90.63 %	1.56 sn	91.54%	1.81 sn	91.54 %	0.7 sn
DecisionTable	89.43 %	1.46 sn	91.54%	21.57 sn	90.63 %	1.81 sn
JRip	91.24 %	1.57 sn	93.05 %	6.34 sn	92.45 %	0.84 sn
İbk	90.63 %	1.52 sn	88.52 %	0.51 sn	**90.94 %**	0.53 sn
SMO	91.24 %	1.34 sn	92.75 %	0.7 sn	**93.35 %**	0.6 sn
BayesNet	85.50 %	1.37 sn	87.92 %	1.14 sn	87.92 %	0.56 sn
NaiveBayes	90.03 %	1.41 sn	90.94%	0.8 sn	90.63 %	0.5 sn
NBM	90.33 %	1.22 sn	90.94 %	0.64 sn	**93.35 %**	0.61 sn

Özellik seçimi yapıldıktan sonra sınıflandırma başarısı artarak en iyi sonuç gösteren algoritmalar SMO, Rasgele Orman ve Çok Terimli Sade Bayes (NBM) algoritmaları olduğu görünüyor.

Tablo 5.15. *Tarkan-S veri kümesinin özellik seçim metotları ile sınıflandırma sonuçları*

Veri seti / Sınıflandırma algoritması	Tarkan-S					
	CfsSubsetEval + GreedyStepwise (default)	Model Oluşum Zamanı	GainRAE + Ranker (default)	Model Oluşum Zamanı	GainRAE + Ranker (seçilen özellik sayısı: 100)	Model Oluşum Zamanı
J48	93.05 %	0.88 sn	94.56 %	1.52 sn	**94.86 %**	0.44 sn
RandomForest	93.96 %	0.94 sn	94.26 %	4.14 sn	**95.17 %**	1.73 sn
REPTree	93.35 %	0.7 sn	94.56 %	1.23 sn	94.56 %	0.48 sn
DecisionTable	93.05 %	1.04 sn	93.05 %	10.14 sn	93.05 %	1.37 sn
JRip	92.75 %	0.77 sn	94.26 %	2.37 sn	94.26 %	0.54 sn
İbk	**93.96 %**	0.66 sn	93.05%	0.37 sn	93.05%	0.4 sn
SMO	94.26 %	0.83 sn	94.86 %	0.5 sn	94.26 %	0.41 sn
BayesNet	90.33 %	0.8 sn	93.35 %	0.81 sn	93.35 %	0.38 sn
NaiveBayes	93.66 %	0.69 sn	95.77 %	0.59 sn	95.47 %	0.39 sn
NBM	93.96 %	0.82 sn	92.45 %	0.41 sn	**94.86 %**	0.35 sn

Özellik seçimi uygulandıktan sonra bazı sınıflandırma algoritmalarının başarı oranlarında artış görülüyor. Bunlar içerisinde nisbeten iyi sonuç gösteren Rasgele Orman algoritması olmuştur. Lakin özellik seçim metotları uygulanmadan daha yüksek başarı gösteren algoritmalar vardır. Bunlar Sade Bayes ve SMO algoritmalarıdır.

Tablolarda, özellik seçimi sonrası özellik seçimi olmadan yapılan sınıflandırmadan daha iyi sonuç verenler kalın olarak gösterilmiştir. Korelasyon tabanlı Özellik Altküme Seçim Değerlendirici + Açgözlü Adımsal Araştırma (CfsSubsetEval+GreedyStepwise) kombinasyonunun bazı durumlarda sınıflandırma performansını artırdığı görülüyor. Kazanım Oranı Özellik Değerlendirici + Sıralayıcı (GainRatioAttributeEval+Ranker) kombinasyonu ise default olarak çalıştırıldığında sınıflandırma başarısını etkilemese de, sıralı özelliklerin ilk 100 tanesi için sınıflandırma başarısını genelde artırdığı görülmüştür. Bir sonraki bölümde sonuçlar daha detaylı ele alınmıştır.

6. TARTIŞMA VE SONUÇ

Araştırma bulguları bölümünde ilk Tablolar veri kümelerinin ham ve ön işlemden geçirilmiş hallerinin sınıflandırma sonuçlarını ifade ediyor. Sonuçlara baktığımızda genel olarak tüm algoritmaların ön işlem uygulanmış veri kümesi üzerinde sınıflandırma performansının ön işlem uygulanmayan veri kümesine göre çok daha iyi olduğu görünmektedir. Tablolarin her birini incelediğimizde tüm veri kümeleri üzerinde Rastgele Orman (Random Forest) ve SMO algoritmalarının başarılı sonuçlar verdiğini görebiliriz. Rastgele Orman algoritması hız açısından biraz düşük performans gösterse de, SMO algoritması hem hız hem de doğruluk yüzdesi açısından başarılı olmuştur. Bu algoritmalar ön işlem uygulanmayan veri kümeleri üzerinde 88.86 %-91.94 % ve 91.55 %-92.75 % arasında doğruluk gösterirken, ön işlemden geçirilmiş veri kümeleri üzerinde 93.56 %-95.64 % ve 94.06 %-96.19 % arasında doğruluk göstermişler. Hemen ardından Jrip ve Sade Bayes algoritmaları da genellikle yüksek başarım gösteren algoritmalardandır. Bu algoritmalar arasında da yine SMO algoritmasının en yüksek başarı oranına sahip olduğunu görebiliriz.

İkinci Tablolar grubu ise aynı veri kümelerinin özellik seçimi metotları uygulanarak sınıflandırılan sonuçlarını göstermektedir. Özellik seçim metotları uygulayarak sınıflandırma yaptığımız zaman toplamda 14 kombinasyonun daha iyi ve hızlı sonuç veren 2 tanesi seçilerek sınıflandırma yapılmıştır. Sonuçları incelediğimiz zaman özellik seçim metotları uygulanmayan veri kümeleri ile uygulanan veri kümelerinin başarım oranları arasında çok büyük fark olmadığını görebiliriz. Bu da veri kümesinin hacminin aşırı büyük olmaması ve gereksiz özellik sayısının fazla olmaması ile ilgili bir durum diyebiliriz. Özellik seçim metoduyla daha iyi sonuç verenler kalın olarak gösterilmiştir. Korelasyon tabanlı Özellik Altküme Seçim Değerlendirici + Açgözlü Adımsal Araştırma (CfsSubsetEval+GreedyStepwise) kombinasyonu bazı algoritmaların başarımını artırsa da bazılarına etki etmemiş ve ya hatta başarım oranını düşürdüğü görülmüştür. Kazanım Oranı Özellik Değerlendirici + Sıralayıcı (GainRatioAttributeEval+Ranker) kombinasyonu ise özellik seçimini default değerlere göre yaptığında veri kümesindeki tüm özellikleri seçtiği için hiçbir algoritmanın başarım oranını ne iyi ne de kötü yönde etkilememişdir. Lakin aynı özellik seçim algoritmasını 100 tane özellik için denediğimizde başarı oranında artış olduğu görülmüştür. Sonuç olarak veri kümeleri üzerinde özellik seçim metotları

uygulanmadan yapılan sınıflandırmada iyi sonuç gösteren Rastgele Orman ve SMO algoritmaları bu metotlar uygulandıktan sonra da başarı oranı biraz daha artarak yine diğerlerine göre daha iyi doğruluk oranı gösteren algoritmalar olmuştur. Algoritmaların başarı oranları ön işlem uygulanmayan veri kümeleri üzerinde 91.02 %-93.36 % ve 91.92 %-93.35 % arasında iken, ön işlemden geçirilmiş veri kümeleri üzerinde 94.01 %-96.70 % ve 93.11 %-96.34 % arasında olmuştur.

YouTube video içeriklerin yayınlandığı bir sosyal medya platformu olarak aynı zamanda kullanıcılarına para kazandırma olanağı sağlaması onun daha fazla tercih edilmesinin nedenlerinden biri olmuştur. YouTube'un popülerliği beraberinde bazı sorunları da getirmiştir. Bunlardan biri de platformun spam içeriklerle dolup taşması olmuştur. Spam içerikler genellikle video ile ilgisi olmayan, reklam amaçlı ve ya kötü içerikli linklerin paylaşıldığı yorumlar şeklinde karşımıza çıkmaktadır. Bu tür yorumlar kullancıların zamanını boşa harcamakla beraber internet traffiği oluşturmakta ve bazen de kullanıcılar için tehlike oluşturabilmektedir.

YouTube'nin yorum denetleme teknikleri yeterli olmadığından platformda spam içerikler nedeniyle oluşan sorunlar hala devam etmektedir. Araştırmalardan yola çıkarak literatürde İngilizce için çok sayıda spam tespiti ile ilgili çalışmalar bulunsa da, Türkçe için bu çalışmalar oldukça sınırlı olmakla beraber YouTube Spamla ilgili çalışma bulunmamaktadır.

Bu sorunlar göz önünde bulundurularak Türkçe YouTube yorumları üzerinde spam tespiti yapılmıştır. Amacımız gelecekte kullanılmak üzere Türkçe YouTube yorumlarını içeren veri kümeleri sunmak ve aynı zamanda o veri kümeleri üzerinde spam tespiti yapmak olmuştur.

Amaç doğrultusunda 2017 senesinde en çok izlenen 5 Türkçe video klipe yapılan yorumlar online bir araçla çekilerek veri seti oluşturulmuş, verinin ön işlemden geçirilmesi için Java programlama dilinden ve Türkçe verilerin işlenmesine olanak sağlayan Zemberek kütüphanesinden yararlanılmıştır. Sınıflandırma aşaması için ise açık kaynak kodlu Weka veri madenciliği aracı kullanılmıştır. Sınıflandırma algoritmaları arasında SMO ve Rastgele Orman sınıflandırıcılarının daha iyi sonuç verdiği görülmüştür.

KAYNAKÇA

Tunç, İ. (Tarihsiz). İnternet ve Sosyal Medya kullanımı. T.C Çalışma ve Sosyal Güvenlik Bakanlığı. 11.

Silva, R.M., Almeida, T.A. and Yamakami, A. (2012). Artificial Neural Networks for Content-based Web Spam Detection. Proceedings of the 14th International Conference on Artificial Intelligence (ICAI'12).

Romero, C., Valdez, M.G. and Alanis, A. (2010). A Comparative Study of Machine Learning Techniques in BlogComments Spam Filtering. WCCI 2010 IEEE World Congress on Computational Intelligence. Barcelona, Spain.

Ze Li and Shen, H. SOAP: A Social Network Aided Personalized and Effective Spam Filter to Clean Your E-mail Box. Clemson, SC 29631.

Hidalgo, J. M. G., Almeida, T.A. and Yamakami, A. (2012). On the Validity of a New SMS Spam Collection. Machine Learning and Applications(ICMLA) 11th International Conference, Volume:2

Wang, D., Irani, D. and Pu, C. (2011). A social-spam detection framework. 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS'11), Perth, Australia

Chaudhary, V. and Sureka, A. (2013). Contextual Feature Based One-Class Classifier Approach for Detecting Video Response Spam on YouTube. IIIT-D-Mtech-CS-IS-13-MT11016.

Khan, R.A. (2019). Spammer Detection: A study of Spam Filter Comments on YouTube Videos. Lahore Garrison University.

Alberto, T.C., Lochter, J.V., Almeida, T.A. (2015). TubeSpam: Comment Spam Filtering on YouTube. IEEE14th International Conference on Machine Learning and Applications.

Abd, T., Altabrawe, H. and Ajmi, S. Q. (2018). YouTube Spam Comments Detection Using Artificial Neural Network . Journal of Engineering and Applied Sciences, 13(22).

Samsudin, N. M., Alias, N., Shamala, P., Othman, N.F. and Din, W.I.S.W. (2019). YouTube spam detection framework using naive bayes and logistic regression. Indonesian Journal of Electrical Engineering and Computer Science, 14(3), pp. 1508-1517.

Uysal, A. K. (2018). Feature Selection for Comment Spam Filtering on YouTube. Data Science and Applications, 1(1).

Ali, A. (2019). Spam Detection in YouTube Comments. Thesis for Bachelor's Degree, University of Engineering and Technology, Lahore, Pakistan

Örnek, Ö. (2019). "Orange 3 ile Türkçe ve İngilizce SMS Mesajlarında Spam Tespiti. ESTUDAM Bilişim Dergisi, 1(1), 1-4.

Karamollaoğlu, H. (2019). Metin verilerinde içerik tabanlı spam tespiti. Yüksek lisans tezi, Gazi Üniversitesi, Bilişim Enstitüsü.

Çıtlak, O. (2018). Sosyal ağlara yönelik öğrenmeye dayalı bir spam hesap tespit modeli ve uygulaması. Yüksek lisans tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü.

Şahin, E. (2018). Makine öğrenme yöntemleri ve kelime kümesi tekniği ile istenmeyen e-posta/e-posta sınıflaması. Yüksek lisans tezi, Hacettepe Üniversitesi, Bilgisayar Mühendisliği ABD.

Altunyaprak, C. (2006). Bayes yöntemi kullanarak istenmeyen elektron postaların filtrelenmesi. Yüksek lisans tezi, Muğla Üniversitesi, Fen Bilimleri Enstitüsü.

Çalış, K., Gazdağı, O. ve Yıldız, O. (2013). Reklam içerikli e-postaların metin madenciliği yöntemleri ile otomatik tespiti**.** Bilişim Teknolojileri Dergisi, 6(1).

Tan, B. (2018). Belediyelerde Sosyal Yardım Hizmetleri Dağıtımına Yönelik Karar Destek Sistemi Geliştirilmesi. Yüksek lisans tezi, Fırat Üniversitesi, Fen Bilimleri Enstitüsü, 22-23.

Özarslan, S. (2014). Öğrenci Performansının Veri Madenciliği ile Belirlenmesi. Yüksek lisans tezi, Kırıkkale Üniversitesi, Fen Bilimleri Enstitüsü.

Turna, F. (2011). Veri Madenciliği Teknikleriyle Tramvay Arıza Kayıtlarından Kural Çıkarımı. Yüksek lisans tezi, Erciyes Üniversitesi, Fen Bilimleri Enstitüsü, 19.

Oralhan, B. (2020). Veri Madenciliği Yaklaşımı ile Telekomunikasyon Sektöründe Arıza Giderme Analizi. BMIJ, 8(1): 1026-1043.

Gümüşçü, A., Aydilek, İ.B. ve Taşaltın, R. (2016). Mikro-dizilim Veri Sınıflandırmasında Öznitelik Seçme Algoritmalarının Karşılaştırılması. Harran Üniversitesi Mühendislik Dergisi, 01(2016), 1-7.

Şık, M. Ş. (2014). Veri madenciliği ve kanser erken teşhisinde kullanımı. Yüksek lisans tezi, İnönü Üniversitesi, Sosyal Bilimler Enstitüsü, 92-99.

http-1: https://www.mediaclick.com.tr/blog/sosyal-medya-nedir-sosyal-medya-siteleri-sosyal-aglar-nelerdir , Son erişim tarihi: 28.05.2020.

http-2: https://dijilopedi.com/2020-turkiye-internet-kullanimi-ve-sosyal-medya-istatistikleri/, Son erişim tarihi: 15.05.2020.

http-3: https://www.brandingturkiye.com/sosyal-medya-ile-geleneksel-medya-karsilastirmasi/, Son erişim tarihi: 15.05.2020.

http-4: https://tr.m.wikipedia.org/wiki/YouTube, Son erişim tarihi: 15.05.2020

http-5: https://www.google.com/amp/s/sosyalmedya.co/sosyal-spamler/amp/, Son erişim tarihi: 28.05.2020

http-6: https://teknojen.net/youtube-video-yorumlarini-excele-aktar/, Son erişim tarihi: 15.05.2020

http-7: https://medium.com/algorithms-data-structures/metin-madencili%C4%9Finde-text-mining-kavramlar-1-e11b87b28847, Son erişim tarihi: 15.05.2020

http-8: https://github.com/ahmetaa/zemberek-nlp/blob/master/README.md , Son erişim tarihi: 30.05.2020

http-9: https://tr.m.wikipedia.org/wiki/Weka, Son erişim tarihi: 28.05.2020

http-10:https://www.google.com/amp/s/uguraltuntass.wordpress.com/2018/03/03/weka-ile-metin-madenciligi/amp/, Son erişim tarihi: 15.05.2020

http-11: https://medium.com/algorithms-data-structures/tf-idf-term-frequency-inverse-document-frequency-53feb22a17c6 , Son erişim tarihi: 15.05.2020

ÖZGEÇMİŞ

ORCID NO: 0000-0002-5819-9599

Adı Soyadı : Sevinj SHİRZADOVA

Yabancı Dil : Türkçe, Rusca, İngilizce

Doğum Yeri ve Yılı : Jambul, Kazakistan / 1991

E-Posta : sevinc.shirzadova@mail.ru

Eğitim ve Mesleki Geçmişi:

- 2009 – 2013, Lenkeran Devlet Üniversitesi / Matematik ve Bilgisayar Öğretimi
- 2015-2016, Matematik öğretmeni, Devlet okulu