

**YÖNETİM BİLİŞİM SİSTEMLERİ DERGİSİ**<http://dergipark.ulakbim.gov.tr/ybs>

Yayın Geliş Tarihi: 16.11.2017
Yayına Kabul Tarihi: 23.11.2017
Online Yayın Tarihi: 20.12.2017

Cilt:3, Sayı:2, Yıl:2017, Sayfa: 1-14
ISSN: 2148-3752

**SENTIMENT ANALYSIS ON TWITTER MESSAGES BASED ON
MACHINE LEARNING METHODS**

Aytuğ ONAN*

Manisa Celal Bayar University, Turkey

Abstract: Twitter is an important social platform, in which people can share their opinions about current issues. The opinions and ideas shared on Twitter can serve as an important source of information for researchers and practitioners. The data available on Twitter can be used to identify current events, to collect information about epidemic diseases and to support crisis management. Sentiment analysis is a recent research direction, which utilizes tools and techniques from several fields, such as natural language processing, statistics and computer science, to identify the subjective information of opinion holders. Machine learning classifiers have been successfully employed in several different application fields of text and web mining, including sentiment analysis. The representation schemes utilized to represent raw text documents are essential for the predictive performance of text classifiers. In this regard, three well-known machine learning classifiers (Naïve Bayes algorithm, support vector machines and logistic regression) on Turkish Twitter messages. In order to represent text documents, different feature representation schemes (1-gram, 2-gram and 3-gram) and their subsets are evaluated.

Keywords: Sentiment Analysis, N-Gram, Machine Learning, Twitter.

**TWITTER MESAJLARI ÜZERİNDE MAKİNE ÖĞRENME Sİ
YÖNTEMLERİNE DAYALI DUYGU ANALİZİ**

Özet: Twitter, insanların güncel konular hakkında görüş bildirdikleri önemli bir sosyal mecra. Twitter kullanıcılarının yaptıkları paylaşım ve görüş bildirimleri, araştırmacı ve uygulayıcılar için önemli bir bilgi kaynağı olarak işlev görmektedir. Twitter verileri, güncel olayları belirleme, yaygın hastalıklar hakkında bilgi toplama, kriz yönetimi gibi birçok farklı alanda kullanılabilir. Duygu analizi, doğal dil işleme, istatistik, bilgisayar bilimleri gibi alanlardan yöntem ve tekniklerin kullanılması ile görüş sahibinin metin içerisinde belirttiği, duygu, görüş, tutum gibi öznel bilgilerin belirlenmesini amaçlayan güncel bir araştırma alanıdır. Makine öğrenmesi sınıflandırıcıları, aralarında duygu

* Contact Author: aytug.onan@cbu.edu.tr, Manisa Celal Bayar Üniversitesi, Manisa, Türkiye

analizinin de yer aldığı, metin madenciliği ve web madenciliğine ilişkin birçok alanda başarıyla uygulanmaktadır. Metin sınıflandırıcılarının başarımlarında, ham metin belgelerinin temsil edilmesinde kullanılan öznitelikler büyük önem taşımaktadır. Bu doğrultuda, bu çalışma kapsamında Türkçe Twitter mesajlarının sınıflandırılmasında, üç temel makine öğrenmesi sınıflandırıcısı (Naive Bayes algoritması, destek vektör makineleri, lojistik regresyon) kullanılmıştır. Metin temsiliinde, farklı öznitelik temsili (1-gram, 2-gram ve 3-gram) ve bu öznitelik temsilleri ile elde edilen farklı öznitelik setleri değerlendirilmiştir.

Anahtar Kelimeler: Duygu Analizi, N-Gram, Makine Öğrenmesi, Twitter.

GİRİŞ

Duygu analizi (Görüş madenciliği), doğal dil işleme, istatistik, bilgisayar bilimleri gibi alanlardan yöntem ve tekniklerin kullanılması ile görüş sahibinin metin içerisinde belirttiği duygu, görüş, tutum gibi öznel bilgilerin belirlenmesini amaçlayan güncel bir araştırma alanıdır. Görüş ve duygular, karar verme sürecini etkileyen önemli bilgilerdir. Bilgi ve iletişim teknolojilerindeki ilerlemeler ile birlikte Webin hızla artan içeriği, bireysel karar vericiler ve karar destek sistemleri için en önemli bilgi kaynaklarından biri olarak işlev görmektedir. Duygu analizi, siyasi politikalara ilişkin kamusal görüşün saptanması, iş zekâsı ve pazar analizi, müşteri memnuniyetinin ölçülmesi, film satışlarının tahmin edilmesi gibi birçok farklı alanda uygulanabilmektedir (Ravi ve Ravi, 2015).

Duygu analizi, metin sınıflandırma problemi olarak modellenilebilir (Onan ve Korukoğlu, 2017). Duygu analizi, gerçekleştirilen sınıflandırma işleminin ayrıntı seviyesine dayalı olarak belge seviyesi duygu analizi, tümce seviyesi duygu analizi ve varlık/özellik seviyesi duygu analizi olmak üzere üç temel sınıf altında incelenmektedir (Medhat vd., 2014). Belge seviyesi duygu analizinde, metin belgesinin, tümce seviyesi duygu analizinde metin belgesinde yer alan belirli bir tümcenin duygu yönü belirlenir. Varlık/özellik seviyesi duygu analizinde ise belirli bir varlığın belirli özelliklerine ilişkin duygu yönünün belirlenmesi amaçlanır.

Duygu analizinde kullanılan yöntemler ise temel olarak makine öğrenmesine dayalı yöntemler ve sözlüğe dayalı yöntemler olmak üzere iki temel sınıf gruba ayrılmaktadır (Medhat vd., 2014). Makine öğrenmesine dayalı duygu analizi yöntemlerinde, makine öğrenmesi sınıflandırıcıları, görüş kutbu etiketlenmiş veri seti ile eğitilerek, sınıflandırma modeli oluşturulmakta, ardından, bu model yeni örneklerle ilişkin görüş kutbunun belirlenmesinde kullanılmaktadır. Sözlüğe dayalı duygu analizi yöntemlerinde ise görüş sözcükleri içeren bir sözlük oluşturularak duygu analizi gerçekleştirilmektedir (Onan ve Korukoğlu, 2016). Duygu analizinde, makine öğrenmesine dayalı yöntemler sıklıkla kullanılmaktadır (Onan vd., 2016). Metin madenciliğinde, karar ağacı algoritmaları, istatistik tabanlı sınıflandırıcılar (Naive Bayes algoritması gibi), doğrusal sınıflandırıcılar (destek vektör makineleri gibi) ve yapay sinir ağları, başarıyla uygulanmaktadır (Aggarwal ve Zhai, 2012). Metin sınıflandırmada, makine öğrenme sınıflandırıcılarının uygulanabilmesi için, öncelikle uygun bir yöntem kullanılarak, metin özniteliklerinin çıkarılması ve seçilmesi gerekmektedir. Terim sıklığı, terim varlığı, tümcenin ögeleri, görüş sözcükleri gibi temsil yöntemleri sıklıkla kullanılan temsil yapılarıdır (Medhat vd., 2014).

Bilgi ve iletişim teknolojilerindeki ilerlemeler ile birlikte İnternet, en önemli bilgi kaynaklarından biri haline gelmiştir. Twitter, kullanıcıların "tweet" adı verilen 140 karakter limitine sahip, kısa mesajlar gönderebildikleri popüler bir sosyal

platformdur. Twitter, Haziran 2016 itibariyle 310 milyonun üzerindeki aktif kullanıcı sayısı ile, hızla gelişen bir sosyal ağdır (Twitter, 2017). Twitter, kullanıcıların etkin bir şekilde iletişim kurabildikleri bir platformdur. Twitter ortamında paylaşılan metin mesajları, hem araştırmacılar hem de uygulayıcılar için önemli bir işleve sahiptir. Twitter verileri, güncel olayların belirlenmesinde, salgın hastalıkların önceden tahmin edilmesinde ve kriz yönetimi gibi birçok farklı alanda uygulama alanı bulabilmektedir (Mahmud vd., 2014; Cheng vd., 2010; Hecht vd., 2011). Twitter mesajları üzerinde duygu analizi gerçekleştirilerek pazarlama kararlarına ilişkin stratejik kararlar alınabilmesi, finansal oranların tahmin edilmesi, müşteri kayıp analizi, sektördeki fırsat ve tehdit faktörlerinin belirlenmesi, rakiplerin etkinliklerinin önceden belirlenmesi, pazarlama karar verme sürecinin etkin bir biçimde ele alınabilmesi gibi birçok farklı pazarlama ve iş zekası uygulaması olanaklı hale gelmektedir (Ravi ve Ravi, 2015).

Bu çalışma kapsamında, Türkçe Twitter mesajlarının, makine öğrenmesi sınıflandırıcıları aracılığıyla sınıflandırılmasında, farklı öğrenme algoritmalarının (Naive Bayes algoritması, destek vektör makineleri, lojistik regresyon) ve farklı öznitelik temsili (1-gram, 2-gram ve 3-gram) yöntemlerinin etkinlikleri değerlendirilmektedir. Metin temsili, etkin bir veri temsil yapısı elde edilmesi, makine öğrenmesi sınıflandırıcılarının doğru tahmin etme başarımı için oldukça önemlidir (Aggarwal ve Zhai, 2012). Bu doğrultuda, temel metin temsil yöntemleri ve bu temsil yöntemlerinin bir araya getirilmeleri ile beş farklı set elde edilmiştir. Çalışma kapsamında, 1-gram, 2-gram ve 3-gram temsil yöntemleri ve bunların bir araya getirilmeleri ile elde edilen değişik öznitelik altkümelerinin başarımları deneysel olarak incelenerek, görüş sınıflandırma için etkin bir öznitelik seti belirlenmesi amaçlanmaktadır.

Çalışmanın ikinci bölümünde, Twitter verileri üzerinde gerçekleştirilen duygu analizi yöntemlerine ilişkin başlıca akademik çalışmalar sunulmaktadır. Üçüncü bölümde, çalışmada kullanılan veri seti, öznitelik setleri ve sınıflandırma algoritmalarına yer verilmektedir. Dördüncü bölümde, sınıflandırma algoritmaları ile elde edilen deneysel sonuçlar, beşinci bölümde ise çalışmanın genel değerlendirmesine yer verilmektedir.

İLGİLİ ÇALIŞMALAR

Twitter verileri üzerinde makine öğrenmesi sınıflandırıcıları kullanılarak duygu analizi gerçekleştirimine yönelik birçok çalışma bulunmaktadır.

Go vd. (2009) çalışmalarında, Naive Bayes, maksimum entropi ve destek vektör makineleri sınıflandırıcılarının Twitter mesajları üzerinde duygu analizi için tahmin etme başarımlarını değerlendirmiştir. Bu çalışmada, metin belgeleri, 1-gram, 2-gram ve tümceyi ögeleri gibi farklı yapılar kullanılarak temsil edilmiştir. Geliştirilen yöntem ile %80 civarında doğru sınıflandırma başarımı elde edildiği gözlemlenmiştir. Birmingham ve Smeaton (2010) çalışmalarında, mikroblog ve blog ortamlarında gerçekleştirilen duygu analizi çalışmalarının etkinliklerini

değerlendirmiş ve mikroblog ortamlarında duygu analizinin, daha kolay olduğu sonucuna varmıştır. Bir diğer çalışmada, Agarwal vd. (2011), Twitter verileri üzerinde duygu analizinde, 1-gram veri temsili, öznitelik mühendisliği temelli veri temsili ve ağaç temelli veri temsili öznitelik setlerinin etkinliklerini değerlendirmiştir. 1-gram temsilde veri seti 10000 civarında öznitelik kullanılarak temsil edilirken, öznitelik mühendisliğine dayalı temsilde öznitelik sayısı 100'e düşürülmesine karşın doğru sınıflandırma başarımı daha yüksek tutulmuştur. Buna karşın, en yüksek doğru sınıflandırma başarımlarının, Twitter mesajlarının ağaç yapısı kullanılarak temsil edildiğinde elde edildiği gözlemlenmiştir. Benzer şekilde, Kouloumpis vd. (2011) çalışmalarında, Twitter mesajları üzerinde duygu analizi için, N-gram öznitelikleri, sözlük tabanlı öznitelikler, tümcenin öğelerine dayalı öznitelikler ve mikroblog platformlarına dayalı özniteliklerin etkinliklerini değerlendirmiştir. Deneysel sonuçlar kapsamında, en yüksek başarımın, N-gram, sözlük tabanlı öznitelik ve mikroblog platformlarına dayalı özniteliklerin birleştirilmesi ile elde edilen öznitelik kümesinde alındığı gözlenmiştir. Bir diğer çalışmada, Pak ve Paroubek (2010) negatif ve pozitif görüş kutbu içeren bir duygu analizi Twitter derlemi oluşturmak için yöntem önerisinde bulunmuştur. Oluşturulan derlem üzerinde istatistiksel dil analizi gerçekleştirilerek, yöntemin etkinliği değerlendirilmiştir. Bir diğer çalışmada, Montejo-Raez vd. (2012) Twitter mesajları üzerinde görüş kutbu belirlemek için, SentiWordNet görüş sözlüğü ve rastgele yürüyüş algoritmasına dayalı bir öğreticisiz yöntem geliştirmiştir. Deneysel çalışmalarda, geliştirilen yöntem ile destek vektör makinelerine yakın doğru sınıflandırma başarımı değerleri elde edildiği gözlenmiştir.

Türkçe Twitter mesajları üzerinde duygu analizi yapılmasına yönelik çalışmalar bulunmaktadır. Örneğin, Nizam ve Akın (2014) çalışmalarında, gıda sektöründeki firmaların ürünlerine yönelik Twitter mesajlarını toplayarak, dengeli ve dengesiz iki farklı veri seti oluşturmuştur. Bu veri setlerinin etkinlikleri, Naive Bayes, rastgele orman, destek vektör makineleri ve C4.5 karar ağacı sınıflandırıcıları kullanılarak değerlendirilmiştir. Meral ve Diri (2014) tarafından gerçekleştirilen çalışmada, Twitter mesajlarının duygu analizi ile incelenmesinde, kelime tabanlı, 2-gram ve 3-gram tabanlı temsil yöntemlerinin, Naive Bayes, rastgele orman ve destek vektör makineleri sınıflandırıcıları üzerindeki etkinlikleri irdelenmiştir. Deneysel sonuçlar incelendiğinde, kelime tabanlı yöntem kullanılarak, %89,5 doğru sınıflandırma başarımı elde edildiği gözlenmiştir. Çoban vd. (2015) tarafından gerçekleştirilen çalışmada, Türkçe Twitter mesajları üzerinde duygu analizi gerçekleştirilmesinde, sözcük torbası ve N-gram temsillerinin etkinlikleri, Naive Bayes ve K-en yakın komşu algoritmaları kullanılarak değerlendirilmiştir. Türkmen ve Cemgil çalışmalarında (2014) mikroblog verileri üzerinde politik eğitimlerin belirlenmesine yönelik olarak, Gezi Parkı gösterileri sırasında atılan Twitter mesajlarının bir alt kümesini ele almıştır. Bu çalışmada, en önemli belirtkeler Ki-kare istatistiği kullanılarak saptanmış, sınıflandırma aşamasında destek vektör makineleri ve rastgele orman algoritmalarının etkinlikleri değerlendirilmiştir. Kasaba ve Yıldıztepe (2016) tarafından gerçekleştirilen çalışmada, Türk filmlerine

ilişkin Twitter mesajlarının duygu analizi, destek vektör makineleri kullanılarak yapılmıştır. Bir diğer çalışmada, Akgül vd. (2016) Twitter verisini ayrıştıran bir program geliştirmiştir. Geliştirilen programda, metin mesajları, sözlük tabanlı, 2-gram, 3-gram ve 4-gram temsil yapıları kullanılarak temsil edilmiştir. Deneysel analizlerde, en yüksek doğru sınıflandırma başarımının sözlük tabanlı yaklaşım ile elde edildiği gözlenmiştir. N-gram tabanlı yöntemler arasında, en iyi sonuçlar 3-gram ile alınmıştır.

YÖNTEM

Bu bölümde, çalışma kapsamında kullanılan veri seti, öznitelik setleri ve sınıflandırma algoritmaları sunulmaktadır.

Veri Seti

Türkçe Twitter mesajları üzerinde duygu analizi gerçekleştirilebilmesi için, Twitter API kullanılarak, Python dilinde yazılmış bir uygulama aracılığıyla bir aylık bir süreçte Twitter mesajları toplanmıştır. Oluşturulan duygu analizi veri seti, 5300 pozitif ve 5300 negatif olmak üzere toplam 10600 Twitter mesajı içermektedir. Twitter mesajlarının elde edilmesinde, Go vd. (2009) tarafından uygulanan yöntem benimsenerek, sorgulama ve mesajların etiketlenmesi sırasında, yalnızca duygu belirten simgeler dikkate alınmıştır. Her bir mesaj, içerdiği duygu simgesinin sınıfına göre, negatif ya da pozitif olarak etiketlenmiştir. Örneğin, :) gibi bir ifade içeren mesaj, pozitif olarak etiketlenirken, :(gibi bir duygu simgesi içeren mesajlar negatif olarak etiketlenmiştir. Twitter mesajlarında yer alan, '@' gibi bazı özel karakterler, makine öğrenmesi sınıflandırıcısı için herhangi bir bilgi vericiliğe sahip değildir. Veri setinin, makine öğrenmesi sınıflandırıcıları ile eğitilerek uygun bir sınıflandırma modeli oluşturulabilmesi için, veri setinin ön işleme aşamasından geçirilmesi gerekmektedir. Ön işleme aşamasında, Çoban vd. (2015) tarafından gerçekleştirilen çalışmada kullanılan veri seti ön işleme adımları benimsenmiş, veri seti üzerinde, dizgi parçalama, durak kelimeleri çıkarma ve kök bulma gibi işlemler uygulanmıştır. Dizgi parçalama aşamasında, hem pozitif hem de negatif içeriğe sahip Twitter mesajları elenmiştir. Twitter mesajlarının tamamı küçük harfe dönüştürülmüş, noktalama işaretleri ve rakamlar ise elenmiştir. Bunun yanı sıra, Twitter mesajları içerisinde yer alan '@', '#' gibi karakterler temizlenmiştir. Metin mesajları, terim sayısı ve karakter uzunluğuna göre filtrelenmiştir. Bunun yanı sıra, metin mesajları içerisinde yinelenen harfler çıkarılarak, sözcükler ön işleme tabii tutulmuş biçimde temsil edilmiştir. Durak kelimelerinin çıkarılması aşamasında, Lucene uygulama geliştirme ara yüzü, kök bulma aşamasında ise Zemberek kütüphanesi kullanılmıştır.

Öznitelik Setleri

Çalışma kapsamında, öznitelik setlerinin oluşturulması amacıyla N-gram modeli kullanılmıştır. N-gram temsil modeli, metin sınıflandırmada en sık kullanılan temsil modellerinden biridir. N-gram, bir karakter katarının n adet karakter dilimidir. N-gram temsili, temel hesaplamalı dilbilim yöntemlerinden biridir. Bu temsilde,

belirli bir dile ilişkin dilbilgisi özellikleri, sözcük yapısı gibi dilsel özellikler dikkate alınmaz. N-gram temsiline dayalı metin sınıflandırmada, farklı uzunluk değerleri alınabilmektedir. Çalışmalarda en sıklıkla kullanılan N-gram modelleri, 1-gram (unigram), 2-gram (bigram) ve 3-gram (trigram) temsilleridir. Örneğin, “Makine Öğrenmesi” tümcesinin, bazı N-gramları şu şekilde elde edilmektedir:

- 1-gramlar: “M”, “a”, “k”, “i”, “n”, “e”, “_”, “Ö”, “ğ”, “r”, “e”, “n”, “m”, “e”, “s”, “i”.
- 2-gramlar: “Ma”, “ak”, “ki”, “in”, “ne”, “e_”, “_Ö”, “Öğ”, “ğr”, “re”, “en”, “nm”, “me”, “es”, “si”.
- 3-gramlar: “Mak”, “aki”, “kin”, “ine”, “ne_”, “e_Ö”, “_Öğ”, “Öğr”, “ğre”, “ren”, “enm”, “nme”, “mes”, “esi”

Burada, alt çizgi karakteri (“_”) kullanılarak temsil edilmektedir. 1-gram temsilinde, metin belgesinde yer alan belirli bir sözcüğün varlığı ya da yokluğu modellenmektedir. 2-gram temsilinde, metin belgesinde birbiri ardına gelen iki sözcüğün varlığı ya da yokluğu, 3-gram temsilinde ise ardışık üç sözcüğe ilişkin modelleme yapılabilmektedir. Bu çalışma kapsamında, veri seti, 1-gram, 2-gram ve 3-gram temsilleri kullanılarak modellenmiştir. Bunun yanı sıra, 1-gram ve 2-gram, 1-gram ve 3-gram, 2-gram ve 3-gram ile 1-gram, 2-gram ve 3-gram öznelik setleri birlikte kullanılarak toplam yedi farklı öznelik seti oluşturulmuştur.

Sınıflandırma Algoritmaları

Çalışma kapsamında, Twitter mesajlarının görüş kutbunun belirlenmesinde üç farklı sınıflandırma algoritması kullanılmıştır. Bunlar, Naive Bayes sınıflandırıcısı (NB), destek vektör makineleri (SVM) ve lojistik regresyon (LR) algoritmalarıdır. Naive Bayes algoritması, Bayes teoremine dayalı istatistiksel bir sınıflandırma algoritmasıdır. Naive Bayes algoritmasında, belirli bir sınıfa ait öznelik değerinin etkisinin, diğer öznelik değerlerinden bağımsız olduğu varsayılır. Bu varsayım, sınıf koşul bağımsızlığı olarak adlandırılır. Sınıf koşul bağımsızlığı sayesinde, gerekli hesaplama işlemleri kolay bir biçimde gerçekleştirilebilmektedir. Naive Bayes algoritması, basit yapısı, hesaplama etkinliği ve yüksek doğru sınıflandırma başarımı nedeniyle metin madenciliğinde sıklıkla uygulanmaktadır. Karar ağacı ve yapay sinir ağları yöntemleri ile karşılaştırılabilir sonuçlar vermektedir (Han vd., 2011).

Destek vektör makineleri, hem doğrusal hem de doğrusal olmayan verilerin sınıflandırılmasında kullanılan temel sınıflandırma algoritmalarından biridir. Destek vektör makinelerinde, doğrusal olmayan bir eşleme yöntemi kullanılarak, orijinal veri daha üst bir boyuta dönüştürülür. Bu yeni boyutta, veriyi en uygun şekilde ayırabilecek bir üst düzlem bulunması amaçlanır. Bu üst düzlem, bir sınıfa ait örneklerin diğer sınıfa ait örneklerden ayrılmasını sağlayan karar sınırını (Vapnik, 1995). Destek vektör makineleri, genelleştirme yetenekleri yüksek, gürültülü ve aykırı değerler içeren verilere karşı dayanıklı sınıflandırma algoritmalarıdır.

Lojistik regresyon, istatistiksel bir sınıflandırma algoritmasıdır. Lojistik regresyon algoritmasında, eğitim setinde yer alan örneklere dayalı olarak bir sınıflandırma modeli oluşturulmakta ve yeni örnekler en yüksek olasılık değerine sahip sınıfa atanmaktadır. Naive Bayes algoritmasında, olasılık değerlerinin hesaplanması Bayes teoremine dayalı olarak gerçekleştirilirken, lojistik regresyon sınıflandırıcısında olasılık değeri, doğrudan parametreler üzerinden hesaplanır (Shatkay ve Craven, 2012).

DENEYSSEL SÜREÇ VE BULGULAR

Bu bölümde, deneysel çalışmalarda kullanılan değerlendirme ölçütleri, deneysel süreç ve sınıflandırma algoritmaları ile elde edilen deneysel sonuçlar sunulmaktadır.

Değerlendirme Ölçütleri

Çalışma kapsamında, sınıflandırma algoritmalarının ve farklı öznelik setlerinin başarımlarının değerlendirilebilmesi için doğru sınıflandırma oranı, F-ölçütü ve ROC eğrisi altında kalan alan kullanılmıştır.

Doğru sınıflandırma oranı, sınıflandırma algoritmalarının başarımını değerlendirmek için kullanılan en temel ölçütlerden biridir. Doğru sınıflandırma oranı (ACC), doğru pozitifler ve doğru negatifler toplamının, doğru pozitif, yanlış pozitif, yanlış negatif ve doğru negatiflerin toplamına oranlanması ile Eşitlik 1'e göre hesaplanır:

$$ACC = \frac{TN+TP}{TP+FP+FN+TN} \quad (1)$$

Burada, TN , TP , FP ve FN sırası ile doğru negatif, doğru pozitif, yanlış pozitif ve yanlış negatif sayılarını temsil etmektedir.

F-ölçütü ise geri çağırma (REC) ve kesinliğe (PRE) dayalı bir ölçüttür. F-ölçütü, [0-1] aralığında değer alır ve Eşitlik 2'ye göre hesaplanır:

$$F - ölçütü = \frac{2*PRE*REC}{PRE+REC} \quad (2)$$

$$PRE = \frac{TP}{TP+FP} \quad (3)$$

$$REC = \frac{TP}{TP+FN} \quad (4)$$

ROC eğrisi altında kalan alan (AUC), sınıflandırma algoritmalarının başarımlarının değerlendirilmesinde kullanılan bir diğer ölçüttür. ROC eğrisi altında kalan alan, [0-1] aralığında değer alır ve yüksek değerler alması, sınıflandırma algoritmasının tahmin etme başarımının daha yüksek olduğunu gösterir.

Deneysel Süreç

Deneysel çalışmalarda kullanılan sınıflandırma algoritmaları, WEKA 3.9 kullanılarak gerçekleştirilmiştir. Sınıflandırma algoritmalarında WEKA yazılımında

yer alan varsayılan parametre değerleri kullanılmıştır. Deneysel çalışmalarda, 10-kat çapraz geçirme yöntemi kullanılarak, veri setleri 10 eşit parçaya ayrılmış ve her bir yinelemede parçalarda biri sınaama amacıyla diğer parçalar eğitim amacıyla kullanılmıştır. Çalışma kapsamında, Twitter mesajlarına ilişkin veri setinin temsilinde 1-gram, 2-gram ve 3-gram temsil modelleri ile bu modeller ile elde edilen öznitelikler bir arada kullanıldığında elde edilen toplam yedi farklı öznitelik seti değerlendirilmiştir. 1-gram temsilinde elde edilen toplam öznitelik sayısı 4486, 2-gram temsilinde elde edilen toplam öznitelik sayısı 5810 ve 3-gram temsilinde elde edilen toplam öznitelik sayısı 2327 dir.

Deneysel Sonuçlar

Tablo 1’de farklı öznitelik setleri ile Twitter mesajları üzerinde, Naive Bayes (NB), destek vektör makineleri (SVM) ve lojistik regresyon (LR) sınıflandırıcıları ile elde edilen doğru sınıflandırma oranları sunulmaktadır. Tablo 2’de farklı öznitelik setleri ve sınıflandırıcılar ile elde edilen F-ölçütü değerleri, Tablo 3’te ise ROC eğrisi altında kalan alan (AUC) değerleri sunulmaktadır.

Tablo 1: Doğru sınıflandırma oranları

Öznitelik Temsili	Sınıflandırma Algoritmaları		
	NB	SVM	LR
1-gram	76.33	72.48	76.12
2-gram	70.05	65.90	69.56
3-gram	61.58	59.76	60.84
1-gram+2-gram	77.78	73.68	77.23
1-gram+3-gram	76.70	73.03	76.29
2-gram+3-gram	69.42	65.55	69.28
1-gram+2-gram+3-gram	76.20	72.59	76.20

Bunun yanı sıra, Şekil 1, Şekil 2 ve Şekil 3’te farklı öznitelik temsili yöntemleri ve sınıflandırma algoritmaları ile elde edilen doğru sınıflandırma oranı, F-ölçütü ve AUC değerlerine ilişkin değişimler özetlenmiştir.

Türkçe Twitter mesajları üzerinde yapılan deneysel çalışmalarda, Naive Bayes ve lojistik regresyon sınıflandırıcıları ile elde edilen doğru sınıflandırma oranları birbirlerine yakın olmak ile birlikte, en yüksek doğru sınıflandırma başarımları, Naive Bayes algoritması ile elde edilmiştir. Farklı öznitelik setleri ve sınıflandırma algoritmaları ile elde edilen konfigürasyonlar arasında en yüksek doğru sınıflandırma oranı (%77,78), en yüksek F-ölçütü değeri (0,79) ve en yüksek AUC değeri (0.85), veri seti 1-gram ve 2-gram öznitelik setlerinin birleştirilmesi ile oluşturulan öznitelik seti ile temsil edildiğinde ve sınıflandırma algoritması olarak Naive Bayes algoritması kullanıldığında elde edilmektedir.

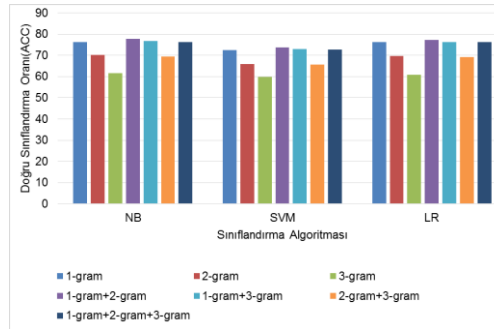
Tablo 2: F-ölçütü değerleri

Öznitelik Temsili	Sınıflandırma Algoritmaları		
	NB	SVM	LR
1-gram	0.77	0.73	0.76
2-gram	0.70	0.66	0.70
3-gram	0.58	0.61	0.60
1-gram+2-gram	0.79	0.75	0.78
1-gram+3-gram	0.77	0.73	0.77
2-gram+3-gram	0.69	0.66	0.69
1-gram+2-gram+3-gram	0.76	0.73	0.76

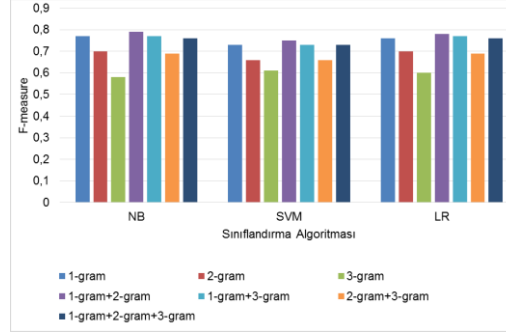
Tablo 3: AUC değerleri

Öznitelik Temsili	Sınıflandırma Algoritmaları		
	NB	SVM	LR
1-gram	0.84	0.72	0.76
2-gram	0.77	0.66	0.70
3-gram	0.67	0.60	0.61
1-gram+2-gram	0.85	0.75	0.78
1-gram+3-gram	0.84	0.73	0.77
2-gram+3-gram	0.76	0.66	0.69
1-gram+2-gram+3-gram	0.84	0.73	0.76

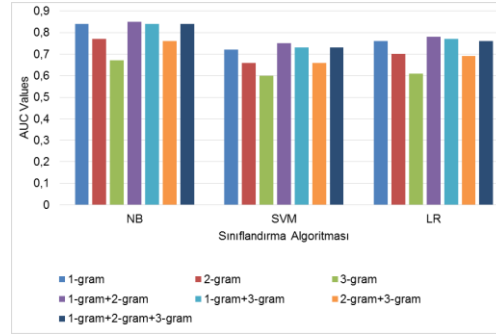
Şekil 1: Doğru sınıflandırma oranları



Şekil 2: F-ölçütü değerleri



Şekil 3: AUC değerleri



Çalışma kapsamında kullanılan farklı N-gram temsil modelleri arasında en yüksek başarımlar, her üç değerlendirme ölçütü için, 1-gram temsil yöntemi ile elde edilmiştir. N-gram temsilindeki N parametre değeri arttıkça, sınıflandırma başarımının düştüğü görülmektedir.

Metin temsilinde, etkin bir veri temsil yapısı elde edilmesi, makine öğrenmesi sınıflandırıcısının doğru tahmin etme başarımı için oldukça önemlidir. Bu doğrultuda, çalışma kapsamında, temel N-gram yöntemlerinin (1-gram, 2-gram ve 3-gram) yanı sıra, bu yöntemler ile elde edilen öznitelik setleri birleştirilerek, veri seti temsil edildiğinde başarımın nasıl etkilendiği de incelenmiştir. Tablo 1-3'te belirtilen değerlendirme ölçütü değerleri ve Şekil 1-3'te sunulan değişim örüntüleri incelendiğinde, karşılaştırılan tüm öznitelik setleri içerisinde, en yüksek başarımın 1-gram temsil modeli ve 2-gram temsil modeli ile elde edilen özniteliklerin birleştirilmesi ile elde edilen öznitelik seti ile elde edildiği görülmektedir. Veri setinin, 3-gram modeli ile temsil edilmesi, 1-gram ve 2-gram temsillerine kıyasla önemli ölçüde performans düşüşüne neden olmaktadır. Buna karşın, 3-gram temsili ile elde edilen öznitelik seti, 1-gram temsili ile elde edilen öznitelik seti ile birleştirildiğinde, en yüksek başarımın elde edildiği öznitelik temsiline (1-gram+2-gram) yakın doğru sınıflandırma, F-ölçütü ve AUC değerleri alınmaktadır. Bunun

yanı sıra, her üç N-gram temsil modeli ile elde edilen öznelik setleri birleştirildiğinde (1-gram+2-gram+3-gram), 1-gram ve (1-gram+2-gram) temsillerine yakın başarımlar elde edilmektedir.

TARTIŞMA ve SONUÇ

Twitter, insanların güncel konular ile ilgili görüş bildirdikleri güncel ve popüler bir sosyal ortamdır. Twitter, araştırmacı ve uygulayıcılar için bilgi sunan önemli bir kaynaktır. Twitter mesajları üzerinden elde edilen veriler, yaygın hastalıkların belirlenmesi, kriz yönetimi gibi birçok farklı alanda uygulanabilmektedir.

Bu çalışma kapsamında, Türkçe Twitter mesajları üzerinde, makine öğrenmesi sınıflandırıcıları kullanılarak, duygu analizi gerçekleştirilmiştir. Metin sınıflandırıcılarının başarımlarında, metin temsilde kullanılan veri temsil modeli büyük önem taşımaktadır. Bu bağlamda, bu çalışma kapsamında Türkçe Twitter mesajlarının sınıflandırılmasında, üç temel makine öğrenmesi sınıflandırıcısı (Naive Bayes algoritması, destek vektör makineleri ve lojistik regresyon) ve üç temel temsil modeli (1-gram, 2-gram ve 3-gram) ile bu temsil modellerinin farklı bileşenleri değerlendirilmiştir. Deneysel çalışmalarda, en yüksek başarımın (%77,78), veri seti 1-gram ve 2-gram öznelik setlerinin birleştirilmesi ile oluşturulan öznelik seti ile temsil edildiğinde ve sınıflandırma algoritması olarak Naive Bayes algoritması kullanıldığında elde edilmektedir. Sınıflandırma algoritmalarının başarımlarında, kullanılan temsil yapısı büyük önem taşımaktadır. Duygu analizinde, N-gram temsil modelinin yanı sıra, sözcük türü bilgisi, POS N-gram gibi farklı öznelik setleri de kullanılabilir. Türkçe’de sözcük türü etiketleme araçlarının yetersizliği nedeniyle, çalışma kapsamında yalnızca N-gram modelleri ile elde edilen öznelik setleri değerlendirilmiştir. Sonraki çalışmalarda, N-gram temsiline yanı sıra, başka temsil modelleri ile elde edilen öznelik setlerinin de değerlendirilmesi yararlı olacaktır.

KAYNAKLAR

- Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media* (pp. 30-38). Association for Computational Linguistics.
- Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In *Mining text data* (pp. 163-222). Springer US.
- Akgül, E.S., Ertano, C., & Diri, B. (2016). Twitter verileri ile duygu analizi. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 22(2), 106-110.
- Birmingham, A., & Smeaton, A. F. (2010, October). Classifying sentiment in microblogs: is brevity an advantage?. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 1833-1836). ACM.
- Cheng, Z., Caverlee, J., & Lee, K. (2010, October). You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the*

19th ACM international conference on Information and knowledge management (pp. 759-768). ACM.

Çoban, Ö., Özyer, B., & Özyer, G. T. (2015, May). Sentiment analysis for Turkish Twitter feeds. In *Signal Processing and Communications Applications Conference (SIU), 2015 23th* (pp. 2388-2391). IEEE.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford, 1*(12).

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011, May). Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 237-246). ACM.

Internet: About.twitter.com, <http://about.twitter.com/company>, 03.05.2017.

Kasaba, E., & Yıldıztepe, E. (2016). Destek vektör makinesi yöntemi ile bir duygu çözümlemesi. *Akademik Bilişim 2016*.

Kouloumpis, E., Wilson, T., & Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg!. *Icwsn, 11*(538-541), 164.

Mahmud, J., Nichols, J., & Drews, C. (2014). Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST), 5*(3), 47.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal, 5*(4), 1093-1113.

Meral, M., & Diri, B. (2014). Twitter üzerinde duygu analizi. *IEEE 22. Sinyal İşleme ve İletişim Uygulamaları Kurultayı, 23-25*.

Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M. T., & Urena-Lopez, L. A. (2012, July). Random walk weighting over sentiwordnet for sentiment polarity detection on twitter. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis* (pp. 3-10). Association for Computational Linguistics.

Nizam, H., & Akın, S. S. (2014). Sosyal Medyada Makine Öğrenmesi ile Duygu Analizinde Dengeli ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması. *XIX. Türkiye'de İnternet Konferansı*.

Onan, A., & Korukoğlu, S. (2016). Makine öğrenmesi yöntemlerinin görüş madenciliğinde kullanılması üzerine bir literatür araştırması. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 22*(2), 111-122.

Onan, A., & Korukoğlu, S. (2017). A feature selection model based on genetic rank aggregation for text sentiment classification. *Journal of Information Science, 43*(1), 25-38.

Onan, A., Korukoğlu, S., & Bulut, H. (2016). A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. *Expert Systems with Applications, 62*, 1-16.

Pak, A., & Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *LREc* (Vol. 10, No. 2010).

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, 89, 14-46.

Shatkay, H., & Craven, M. (2012). *Mining the biomedical literature*. MIT Press.

Turkmen, A. C., & Cemgil, A. T. (2014, April). Political interest and tendency prediction from microblog data. In *Signal Processing and Communications Applications Conference (SIU), 2014 22nd* (pp. 1327-1330). IEEE.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer.