Mikroblog Verilerinden Politik İlgililik ve Eğilim Tahmini
Political Interest and Tendency Prediction from Microblog Data

Ali Caner Türkmen, Ali Taylan Cemgil
Bilgisayar Mühendisliği Bölümü
Boğaziçi Üniversitesi
34342 Istanbul, Türkiye
Email: {caner.turkmen, taylan.cemgil}@boun.edu.tr

***Özetçe*—Çevirimiçi sosyal ağlar, özellikle de Twitter gibi mikroblog siteleri, son yıllarda kullanıcıların güncel gündem hakkında fikirlerini sıklıkla beyan ettikleri bir platform olarak kabul görmüştür. Bu platformlarda paylaşılan verilerden, kullanıcıların politik eğilimlerini anlamak son yıllarda oldukça ilgi gören bir araştırma konusudur. Bu çalışmada, 2013 yılında Türkiye'de gerçekleşen Gezi Parkı gösterileri sırasında Twitter üzerinden atılan mesajların küçük bir alt kümesi üzerinde çalışılmış, bu mesajlarla politik ilgililik ve eğilimin tahmin edilmesi amacıyla Ki-kare istatistiği kullanılarak seçilen en anlamlı belirtkeler ile Destek Vektör Makineleri ve Rassal Orman sınıflayıcıları kullanılmıştır. Çalışma sonucunda, bu yöntemler kullanılarak ilgililik ve eğilim belirlemede belirli bir verim elde edilebildiği gösterilmiştir.**

***Anahtar Kelimeler*—Twitter, Yapay Öğrenme, Destek Vektör Makineleri**

***Abstract*—Online social networks, and especially the popular microblogging service Twitter have taken to be the epicenter of massive social movements, where users often openly express political tendencies –a trend which has led to making the classification of political tendencies from social shares a research question of interest. In this research, we collect and hand label a small subset of political messages sent during the follow up period of Gezi Park protests that took place in Turkey in 2013. We demonstrate that in order to predict political relevance and tendency, a Chi-square statistic based feature selection approach coupled with Support Vector Machine and Random Forest classifiers yields significant prediction power.**

***Index Terms*—Twitter, Machine Learning, Support Vector Machines**

I. GİRİŞ

Son yıllarda Twitter gibi çevirimiçi sosyal platformlar göstericilerin ve karşıtlarının fikirlerini yoğunlukla beyan ettikleri bir ortam olma özelliği ile öne çıkmıştır. Bu davranış, daha önce özellikle Arap Baharı olarak isimlendirilen gösteriler dalgasında ve Wall Street'i İşgal Et hareketinde de olduğu gibi, Türkiye'de 2013 yılında meydana gelen Gezi Parkı gösterilerinde de gözlemlenmiştir.

Twitter üzerindeki kullanıcılar tarafından oluşturulan içeriği sınıflayarak politik konularla ilgililik ve eğilim tahmin etmek; bir çok uygulama alanında değer ifade etmektedir. Hatta politika haricinde herhangi bir konu ile ilgililiği ve bu konu bağlamında eğilim veya duyguyu tahmin etmek son yıllarda oldukça ilgi gören bir araştırma konusudur.

978-1-4799-4874-1/14/$31.00

Bu çalışmada politika bağlamından elde edilmiş bir veri seti, ilgililik ve iki taraflı bir tartışmada eğilim bakımlarından incelenmekte ve bu iki boyut tahmin edilmeye çalışılmaktadır.

Twitter içeriği üzerinden doğal dil işleme yordamıyla kullanıcıları ve mesajları sınıflandırma uygulamaları bulunmaktadır. Twitter mesajları üzerinden siyasi eğilim tahminlemesi daha önce Rao v.d. [1] tarafından incelenmiş olup, dilbilimsel bir ön incelemeden sonra n-gram öznitelik yöntemi ve Destek Vektör Makineleri (DVM) tabanlı bir sınıflandırma algoritması ile %82 başarım elde etmişlerdir.

Tumasjan v.d. [2], Twitter üzerinden siyasi içerikli mesajları kendi geliştirdikleri yöntemlerle sınıflandırarak, Twitter'ın siyasi seçimlerdeki sonuçları tahminlemeye dönük anlamlı veri barındırdığını göstermişlerdir.

Pennacchiotti ve Popescu [3], Twitter mesajlarına eğiticili ve karar ağacı tabanlı bir yapay öğrenme yaklaşımıyla Amerika Birleşik Devletlerinde Cumhuriyetçi (Republican), ve Demokrat kullanıcıların tespiti için çok yüksek kesinlik (precision) ve çağrı (recall) elde etmişlerdir.

Conover v.d. [4], 2011 Amerika Birleşik Devletleri Kongre seçimleri öncesi toplanan Twitter mesajları ile tüm yazılı içerik yerine Twitter etiketleri (hashtag) üzerinde çalıştırılan bir DVM algoritması ile %91 gibi yüksek başarımlı bir sınıflandırma yapılabildiğini göstermişlerdir.

Stieglitz ve Dang-Xuan [5], Almanya'da iki eyalet meclisi (*Landtag*) seçimi öncesi toplanan Twitter mesajlarının yeniden paylaşılma olasılıklarını kullanıcı davranışları boyutlarıyla mesaj içeriği boyutlarını harmanlayarak tahmin etmişler ve bunun seçim sonucuna etkisini araştırmışlardır.

Türkiye'de gönderilmiş Twitter mesajları üzerinde yapay öğrenme yoluyla sınıflandırma uygulamaları yapmanın önünde engel oluşturan birkaç konudan bahsedilebilir. İlk olarak, Twitter'ın 140 karakterle sınırlandırılmış mesaj uzunluğu kullanıcıları eklemeli bir dil olan Türkçe'yi kısaltmalar ve yazım

bozuklukları yoluyla sınırlandırmaya itmekte ve mesajların yapay olarak okunmasını zorlaştırmaktadır. İkinci olarak da, çevirimiçi ağlarda oluşan kendine has iletişim kültürü ve normal doğal dilden sapan bir jargon kullanımından bahsedilebilir.

Türkçe üzerinde yapılan uygulamalara bakıldığında, ticari uygulama sahasında Çetin ve Amasyalı'nın [6] Twitter üzerinde duygu analizi yaparken eğiticili terim ağırlıklandırma yöntemiyle geleneksel yöntemlere kıyasla daha yüksek verim aldıkları, Türkmen v.d.'nin [7] eğiticili terim seçimi yöntemleriyle Twitter mesajları üzerinden demografik bilgi sınıflandırma uygulamasında başarılı sonuç elde ettikleri görülmektedir.

Cingiz ve Diri [9], çalışmalarında mikroblog kullanıcılarının sınıflandırılması esnasında öznitelik seçiminde Ki-kare yönteminden faydalanmış, ve Destek Vektör Makinelerine kıyasla Çok Değişkenli Naïve Bayes yöntemiyle daha yüksek sınıflandırma başarısı elde ettiklerini bildirmişlerdir.

Türkiye tabanlı Twitter mesajlarına yukarıdaki çalışmalarla karşılaştırmalı olarak bakıldığında birkaç temel fark ortaya koyulabilir. Türkiye, Rao v.d.'nin incelediği Amerika Birleşik Devletleri'nden çok daha yoğun bir siyasi gündem içermekte, bu sebeple daha zengin veri sunmaktadır. Ancak, Türkçe'nin eklemeli dil yapısı ve Amerika Birleşik Devletleri'ndeki çoğunlukla iki partili siyasi ortama kıyasla Twitter kullanıcılarının çokseslliliği sınıflandırmayı zorlaştıran bir etken olarak baş göstermektedir.

Bu çalışmanın ikinci bölümünde kullanılan veri seti ve toplanışı tanıtılmakta, üçüncü bölümünde öznitelik çıkarımı ve sınıflandırma algoritmaları ile yapılan deneyler aktarılmaktadır. Dördüncü ve son bölümde varılan sonuçlar ve muhtemel sonraki adımlar tartışılmıştır.

II. VERİ SETİ

A. Verinin Toplanması

Gezi Parkı gösterileri, Türkiye'de Mayıs 2013 sonlarında başlayıp Temmuz 2013'e kadar süren ve birkaç ilde birden meydana gelmiş geniş toplumsal olaylardır. Taksim Gezi Parkı'nın yıkılmasına tepki olarak başlayan olaylar, bir aylık bir süre içinde Türkiye'nin yakın tarihindeki en kapsamlı toplumsal olaylara dönüşmüştür.

Bu olaylar esnasında, hem göstericilerin hem de gösteri karşıtı ve mevcut yönetim yanlılarının fikirlerini paylaşmak, kitlelere duyurmak hatta örgütlenmek amacıyla çevirimiçi sosyal mecraları yoğunlukla kullandıkları, Türk hükümeti de dahil olmak üzere birçok gözlemci tarafından kaydedilmiştir. Bu sosyal mecralar arasında belki de en ünlüsü, doğası itibariyle genelde tüm kullanıcılara açık bir şekilde kısa mesajlar yayınlanmasını ve bu mesajların tekrar paylaşılarak (örn. "Retweet") kitlelere ulaştırılmasını kolaylaştıran Twitter hizmetidir.

Bu çalışma kapsamında, Gezi Parkı olaylarını izleyen bir dönemden, 19-24 Temmuz 2013 tarihleri Türk kullanıcılar tarafından gönderilen Twitter mesajları, Twitter'ın sunduğu uygulama programlama arayüzü (Streaming API) vasıtasıyla toplanmış ve bunlardan 1315 mesaj, el ile işaretlenmiştir. El ile işaretlenen bu 1315 mesaj, sınıflandırmayı kolaylaştırmak amacıyla "Gösteriler Yanlısı", "Gösterilere Karşı", "Nötral/İlgisiz" mesajlar olarak sınıflandırılmıştır.

Twitter mesajlarını toplamak için özel olarak yazılan Python[1] tabanlı bir istemci kullanılmış, verilerin depolanması için ise ilişkisel olmayan veritabanı MongoDB'den[2] faydalanılmıştır.

B. Ön İşleme

Elde edilen veriler öncelikle kelime dışı noktalama işaretleri, kesme ile ayrılan ekler ayıklanmıştır. Türkçe'ye özgü harfler Latin alfabesindeki eşlerine çevrilerek Türkçe ve İngilizce klavye kullanımıyla oluşan farklılıklar giderilmiştir.

Ek olarak, Twitter'da kullanıcıların diğer kullanıcılara atıfta kullandığı "@" işaretiyle başlayan kelimeler (bkz. "mention"), internet bağlantıları ve rakamlar da ayıklanmış ve tüm kelimeler küçük harflere çevrilmiştir.

Son olarak, her mesaj belirtkelemeden (tokenization) geçirilerek, satırları her bir Twitter mesajını, sütunları ise örneklemde geçen tüm kelimelerden her birini ifade eden ikili bir matris oluşturulmuştur. A matrisinin içeriği Denklem 1'de gösterilmektedir.

$$A_{ij} = \begin{cases} 1 & i \text{ indisli mesajda } j \text{ indisli kelime geçiyorsa,} \\ 0 & \text{diğer halde} \end{cases} \tag{1}$$

III. DENEYSEL ÇALIŞMALAR

Deneysel çalışmalar için geliştirilen sistemde başlıca iki aşamalı bir yöntem izlenmiştir. İlk olarak, her bir belirtkenin (kelime) ifade ettiği öznitelikler arasından eğiticili bir öznitelik seçimi yöntemiyle en anlamlı öznitelikler seçilmiştir. Son olarak ise, her bir Twitter mesajı için amaçta belirtilen konu ile ilgililik (relevance) ve siyasi eğilim (tendency) değişkenlerini tahmin etmek için bir sınıflandırma algoritması kullanılmıştır.

Deneysel çalışmalar esnasında, Python dili üzerinde geliştirilen bir yapay öğrenme kütüphanesi olan scikit-learn [8] kullanılmış olup, aksi belirtilmediği hallerde modelleme parametreleri varsayılan değerlerinde bırakılmıştır.

A. Öznitelik Seçimi

Öznitelik seçimi denemeleri esnasında, öznitelik kümesi içinden eğiticili bir yöntem yoluyla en anlamlı özniteliklerin çıkarılması amaçlanmıştır. Bu amaçla, önceki çalışmalarda [9], [10] iyi sonuç verdiği gözlenmiş Ki-kare istatistiğinden faydalanılmıştır.

Ki-kare istatistiği, özetle t özniteliği ile c sınıfı arasındaki bağımlılığı anlatan bir istatistiktir, ve t özniteliğinin c sınıfı tabanlı Ki-kare (χ^2) istatistiği Denklem 2'de verildiği şekliyle hesaplanmaktadır.

$$\chi^2(t,c) = \frac{N \times [P(t,c) \times P(t^-,c^-) - P(t,c^-) \times P(t^-,c)]^2}{P(t) \times P(t^-) \times P(c) \times P(c^-)} \tag{2}$$

[1] www.python.org
[2] www.mongodb.org

Denklem 2'de verilen değerlerden $P(t,c)$, t ve c'nin beraber pozitif oldukları örnek sayısını, $P(t,c^-)$, t'nin pozitif olup c'nin negatif olduğu durum sayısını, N toplam örnek sayısını, $P(t)$ t'nin pozitif olduğu toplam durum sayısını ifade etmektedir.

Çalışmada elde edilen 6376 öznitelik Ki-kare istatistiğine göre sıralanarak en anlamlı K öznitelik seçilmiştir. K değeri, bir deney parametresi olarak tutulmuş olup, farklı K değerleri için her iki sınıflandırma yöntemiyle elde edilen sonuçlar, aşağıdaki bölümde paylaşılmıştır.

Çalışmada aynı zamanda öznitelik çıkarımı için Negatif Olmayan Matris Ayrışımı ile ayrıştırılan K öznitelik denenmiş ancak bu eğiticisiz yöntemin Ki-kare istatistiği ile elde edilen özniteliklere göre istikrarlı bir şekilde daha düşük başarım verdiği görülmüştür.

B. Sınıflandırma Yaklaşımı

Sınıflandırma araştırması, iki ana sorun etrafında şekillenmiştir. Bunlardan ilki, bir Twitter mesajının eldeki konuya ilgililik ifade edip etmediğini ayıran; böylelikle eğilim ve fikir beyan eden mesajları, yalnız haber niteliğinde olan ya da konu ile ilgisiz mesajlardan ayırt etmek amacıyla bir ilgililik sınıflandırmasıdır. İkinci olarak ise, eğilim ifade ettiği bilinen mesajların iki taraflı bir tartışmada hangi tarafa eğilim gösterdiğinin belirlenmesidir.

Araştırma kapsamında, tüm mesajların el ile "Gezi Parkı Gösterileri Yanlısı", "Gösteriler Karşıtı" ve "Tarafsız/İlgisiz" olarak işaretlendiğinden bahsedilmişti. İlk sınıflandırma uygulamasında Gösteriler Yanlısı/Karşıtı olarak işaretlenen mesajların Tarafsız/İlgisiz mesajlardan ayırt edilmesi araştırılmakta, ikinci uygulamada ise taraf ifade ettiği bilinen ("Tarafsız/İlgisiz" olmayan) mesajların Gösteriler Yanlısı ya da Karşıtı olarak sınıflandırılması çalışılmaktadır.

Her iki uygulamaya girdi oluşturan öznitelik kümesi, Ki-kare yöntemiyle belirlenen K en anlamlı özniteliklerdir. Sınıflandırmanın hedef değişkeni olarak ilgililik (R, relevance) ve eğilim (T, tendency) aşağıdaki şekilde ifade edilebilir:

$$R_i = \begin{cases} 1 & i \text{ indisli mesaj siyasi olarak ilgili ise,} \\ 0 & \text{diğer halde} \end{cases} \quad (3)$$

$$T_i = \begin{cases} 1 & i \text{ indisli mesaj gösterilere karşıt fikirde ise,} \\ 0 & \text{diğer halde} \end{cases} \quad (4)$$

İki uygulamada da, sınıflandırma başarısı aşağıda tanıtılan iki sınıflandırma algoritması yoluyla ölçülmüş ve karşılaştırılmıştır. Bunlardan ilki Rassal Orman (Random Forest), ikincisi ise daha önceki çalışmalarda başarılı bulunan Destek Vektör Makineleri (DVM, Support Vector Machines) algoritmalarıdır.

Rassal Orman, Breiman tarafından tarif edilmiş [11] ve çok büyük sayıda karar ağacını, öznitelik kümesinden rassal altkümeler alarak eğitilmesini içeren bir yapay öğrenme algoritmasıdır.

Son yıllarda sınıflandırma uygulamalarında sıklıkla kullanılan Rassal Orman'da N adet karar ağacı eğitilmektedir. Bu N karar ağacının eğitimi sırasında her dallanmada rassal bir öznitelik altkümesi ve eğitim veri seti alt kümesi seçilmektedir ve bu öznitelik altkümesinden en iyi ayrımı ifade eden değişken ile dallanma gerçekleşmektedir.

Rassal orman ile tahmin yapmak için ise, yeni bir örnek N karar ağacının tümünde sınıflandırılmakta, ve bu ağaçların oylamaları (çoğunluk ağacın kararı) ile sınıflandırma sonucu bildirilmektedir.

Bu çalışma kapsamında, Breiman tarafından geliştirilen bu algoritmanın bir uygulaması kullanılmaktadır. Bölünmeler için Gini ölçütü hedef olarak belirlenmiş, ve Rassal Ormandaki ağaç sayısı 50 ve 100 olarak farklı iki değerle test edilmiştir. Bu modellere dair sonuçlar, ilgili bölümde $RO50$ ve $RO100$ altında bildirilmiştir.

Destek Vektör Makineleri, Cortes ve Vapnik [12] tarafından tanıtılmış olup N-boyutlu uzayda bilinen iki sınıfı en iyi ayıran karar düzlemini bulma ilkesine dayanır.

Diğer bir anlatımla, $D = \{x_i | x_i \in R^p\}_{i=0}^n$ veri seti verildiğinde, p elemanlı her bir x_i vektöründen ilk sınıfa ait olanlar $w.x - b \geq 1$, diğer sınıfa ait olanlar $w.x - b \leq -1$ kısıtını sağladıklarında, iki sınıfa ait olan veri örnekleri birbirlerinden doğrusal olarak ayrılmış olacaklardır. Bunun da ötesinde en iyi karar düzleminin bulunması için, her iki sınıfın düzleme en yakın örneklerini birbirinden en uzak şekilde ayıracak $m = \frac{2}{||w||}$ marjının ençoklanması gerekmektedir.

Bu yolla, problem Denklem 5'da ifade edilen bir optimizasyon problemi şeklinde ifade edilebilmektedir, ve DVM yöntemi ile çözülmektedir.

$$\begin{aligned} &\underset{(w,b)}{\operatorname{argmin}} \frac{1}{2}||w||^2 \\ &\text{s.t.} \\ &wx - b \geq 1, \text{ilk sınıf için} \\ &wx - b \leq -1, \text{ikinci sınıf için} \end{aligned} \quad (5)$$

Destek Vektör Makineleri ile çalışma yapılırken daha önceki çalışmalarda verimli sonuç verdiği bildirilen ve bu araştırmanın ön çalışmalarında daha iyi sonuç verdiği gözlemlenen doğrusal çekirdek fonksiyonu kullanılmıştır.

Deney başarısını etkin bir şekilde ölçmek ve elde edilen modellerin genelleştirilebildiğini göstermek amacıyla tüm sınıflandırma çalışmalarında örneklemin %70'i eğitim, %30'u test veri seti olarak ayrılmıştır. Raporlanan başarı ölçüleri, eğitilmiş modelin test seti üzerindeki başarısını ifade etmektedir.

Ayrıca, testlerde bahsedilen örneklemlerdeki sınıf sayılarının farkları, farklı algoritmalarca yanlılık (bias) doğurabileceğinden her veri seti öncelikle sınıf sayıları eşitlenecek şekilde dengelenmiş, bu esnada sayıca fazla olan sınıftan rastgele seçilen örnekler veri setinin dışında bırakılmıştır.

Sınıflandırma başarısı, başarıma (accuracy) ek olarak bir sınıflandırma başarısını ölçmede sıkça kullanılan ve kesinlik (P, precision) ve çağrı (R, recall) ölçütlerinin bileşimi olan

Tablo I
İLGİLİLİK TAHMİNİ DENEYLERİ BAŞARIM VE F1-DEĞERLERİ

K Öznitelik Sayısı	Sınıflandırma Algoritması		
Başarım | F1-değeri	RO50	RO100	DVM
50	0.71 | 0.66	0.72 | 0.77	0.74 | 0.70
100	0.73 | 0.70	0.73 | 0.70	**0.77 | 0.75**
200	0.71 | 0.67	0.73 | 0.69	0.75 | 0.72

Tablo II
EĞİLİM TAHMİNİ DENEYLERİ BAŞARIM VE F1-DEĞERLERİ

K Öznitelik Sayısı	Sınıflandırma Algoritması		
Başarım | F1-değeri	RO50	RO100	DVM
50	0.76 | 0.75	**0.81** | 0.80	0.73 | 0.70
100	0.77 | 0.76	0.76 | 0.75	0.75 | 0.72
200	0.74 | 0.77	0.80 | **0.81**	0.79 | 0.77

F1-değeri kullanılmıştır. F1 değerinin hesaplanması aşağıdaki gibidir:

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{6}$$

C. Deney ve Sonuçlar

İlgililik tahmini uygulaması için toplanan 1209 mesaj ve 6376 belirtke üzerinden Ki-kare istatistiği kullanılarak en iyi K belirtke belirlenmiş, bu belirtkelerin ifade ettiği öznitelikler ile yukarıda anlatılan 50 ve 100 karar ağaçlı Rassal Orman ($RO50$ ve $RO100$) ve Destek Vektör Makineleri (DVM) yöntemleri kullanılarak R_i değişkeni sınıflandırılmaya çalışılmıştır.

Deneyler esnasında K parametresine sırasıyla 50, 100, 200 değerleri verilmiştir. Farklı deneylerden elde edilen sonuçlar Tablo I'de gösterilmiştir. Tabloda gösterilen F1-değerleri, pozitif sınıf ($R_i = 1$) durumları içindir.

Sonuçlarda da görüldüğü üzere, DVM algoritması istikrarlı şekilde en iyi sonucu verirken, tüm modellerde K değerine 100 verildiğinde en yüksek verim alınmaktadır.

Eğilim belirleme çalışmasında ise, eğitim veri setinden "İlgisiz/Nötr" ($R_i = 0$) olarak işaretlenmiş veriler çıkarılmış ve geriye kalan 728 örnek üzerinden "Gösteri Karşıtı" olarak işaretlenen mesajlar ($T_i = 1$) pozitif sınıf kabul edilerek araştırma yapılmıştır.

İlgi tahmini çalışmasında olduğu gibi, Ki-kare istatistiğine göre en anlamlı 50, 100, 200 belirtke seçilmiş ve $RO50$, $RO100$ ve DVM yöntemleri çalıştırılarak elde edilen bulgular Tablo II'de sunulmuştur.

Çalışmada en verimli sonuç yüksek karar ağacı sayılı Rassal Orman ($RO100$) algoritmasından alınmış ve 0.8 in üstünde başarım elde edilmiştir. Ayrıca K parametresine göre duyarlılığın az olduğu görülmüştür.

İki çalışma karşılaştırıldığında, ilkinde Destek Vektör Makinelerinin istikrarlı şekilde daha başarılı sonuç vermesi; belirtkelerin doğrusal kombinasyonlarının ilgililik tahmin etmede daha fazla tahmin gücü verdiğini göstermektedir. Ayrıca K parametresine göre en verimli sonuçların 100 değeri civarında alınması, ilgililiği tahmin eden 100 civarında belirtkenin varlığını işaret etmektedir.

İkinci çalışmada Rassal Orman yaklaşımının daha başarılı sonuç vermesi ve K parametresinin değerine göre istikrarlı artış / azalma görülmemesi; eğilimi tahmin eden ve beraber kullanılan sayıca az anlamlı belirtkenin varlığını göstermektedir.

IV. VARGILAR

Bu çalışmada, Türk kullanıcıların Twitter mesajları işlenerek bir konuya dair ilgililiği ve eğilimi tahminin mümkün olduğu, politika ile ilgili fikirleri içeren bir veri seti üzerinden gösterilmiştir.

Öznitelik elde edilmesine dair ön çalışmalarda en iyi sonucu verdiği gözlenen Ki-kare yöntemi kullanılarak en anlamlı K belirtkenin bulunması amaçlanmış, ilgililik tahmini için $K = 100$ durumunda en verimli sonuç alındığı görülmüştür. Eğilim tahmini için ise K parametresi ile başarım arasında belirleyici bir bağ görülmemiştir.

Karşılaştırılan iki sınıflama algoritması arasından ilgililik tahmininde doğrusal çekirdek fonksiyonlu Destek Vektör Makinesinin daha verimli sonuç verdiği gözlenirken, eğilim tahmininde Rassal Orman yönteminin yüksek başarım elde ettiği bulgu olarak sunulmuştur.

Bu araştırmanın olası sonraki adımları arasında, tek kelime belirtkelere ek olarak *n-gram* kelime silsilelerinin de birer öznitelik olarak araştırılması, Twitter kullanıcılarının sık olarak yaptığı bilinçli yazım hatalarının tespit edilmesi ve düzeltilmesi amacıyla bir doğal dil işleme aracının kullanılması ve farklı öznitelik seçimi ve sınıflandırma algoritmaları ile başarımın yükseltilmesi sayılabilir.

KAYNAKLAR

[1] D. Rao et al. "Classifying Latent User Attributes in Twitter," *Proceedings of the 2nd international workshop on Search and mining user-generated contents.*, ACM, 2010.

[2] A. Tumasjan et al. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," *ICWSM 10* pp.178-185 2010.

[3] M. Pennacchiotti, ve A.M. Popescu. "A Machine Learning Approach to Twitter User Classification," *ICWSM* 2011.

[4] M.D. Conover et al. "Predicting the political alignment of twitter users," *Privacy, security, risk and trust (passat), 2011 IEEE third international conference on*. IEEE, 2011.

[5] S. Stieglitz, L. Dang-Xuan. "Political Communication and Influence Through Microblogging - An Empirical Analysis of Sentiment in Twitter Messages and Retweet Behavior," *45th Hawaii International Conference on System Sciences*, pp.3500-3509, 2012.

[6] M. Çetin, M.F. Amasyalı "Eğiticili ve Geleneksel Terim Ağırlıklandırma Yöntemleriyle Duygu Analizi," *21. IEEE Sinyal Isleme ve Iletisim Uygulamaları Kurultayı*, 2013.

[7] H.I. Turkmen et al. "Konuşma Dili Kullanılarak Demografik Bilgilerin Sınıflandırılması," *19. IEEE Sinyal Isleme ve Iletisim Uygulamaları Kurultayı*, 2011.

[8] F. Pedregosa. "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research* vol 12, pp.2825–2830, 2011.

[9] M.Ö. Cingiz, B. Diri. "Mikroblog Kullanıcılarının Kategorizasyonu," *20. IEEE Sinyal Isleme ve Iletisim Uygulamaları Kurultayı*, 2012.

[10] Y. Yang, J. Pedersen. "A Comparative Study on Feature Selection in Text Categorization," *The Proceedings of ICML 97*, 1997.

[11] L. Breiman. "Random Forests," *Machine Learning*, 45(1), pp.5-32, 2001.

[12] C. Cortes, V. Vapnik. "Support Vector Networks," *Machine Learning*, 20, pp.273-297, 1995.