Twitter Üzerinde Duygu Analizi
Sentiment Analysis on Twitter

Meriç Meral
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
İstanbul, Türkiye
mericmeral@gmail.com

Banu Diri
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
İstanbul, Türkiye
banu@ce.yildiz.edu.tr

***Özetçe*—İnternetin genişlemesi ve insanların kişisel fikirlerini paylaşmaları ile sosyal medya önemli bir bilgi kaynağı haline gelmiştir. Sosyal medya verileri ham haliyle çok kirli olduğu için üzerinde işlem yaparak gerçeğe uygun sonuçlar çıkarmak mümkün olamaz. Bu çalışmada, Twitter'dan toplanan veriler üzerinde algı analizi yapılmıştır. Bu analiz gerçekleştirilirken doğal dil işleme ve Naïve Bayes, Rastgele Orman ve Destek Vektör Makinesi gibi makine öğrenmesi yöntemleri kullanılarak, akıllı bir sistem oluşturulmuş ve elde edilen sonuçlar karşılaştırmalı olarak verilmiştir.**

Anahtar Kelimeler — Twitter, duygu analizi, makine öğrenmesi.

Abstract — **Social media has become an important information source with the expanding internet and ideas shared by the people. The social media raw data which is quite disordered and messy can not be processed as it is and obtain adequate results. In this study, sentiment analysis has been performed by collecting data from the Twitter. To perform this analysis an intelligent system has been created by using machine learning methods such as Naïve Bayes, Random Forest, Support Vector Machine and the compared results has been given.**

Keywords—Twitter, sentiment analysis, machine learning

I. GİRİŞ

Günümüz dünyasında rekabet halindeki firmalar, geleneksel pazarlama yöntemleri ile müşteriye ulaşmanın yeterli olmadığını görmüşler ve sahip olduğu potansiyel nedeni ile gündemlerine sosyal medyayı almışlardır. İnternetin ve sosyal medyanın büyümesi ile her geçen gün bunları günlük rutinleri arasına daha fazla yerleştirmekte olan insanlar, firmaların da sosyal medya ortamlarında daha fazla yer almasına neden olmuştur. Bu açıdan sosyal medya çok önemli bir veri kaynağıdır ve buradan elde edilen bilgi birçok sektöre yön vermektedir. İnsanların bir ürün ya da hizmet hakkındaki olumlu ve olumsuz görüşleri, sosyal medyadaki ağ aracılığı ile bütün dünyaya yayılmaktadır. Sosyal medyada yayılan fikirler ürün ve hizmetlerin ne kadar başarılı olduğunu yansıtmakta ve dolayısı ile satışları ve ekonomiyi etkilemektedir. Oluşan ekonomik etki zamanla firmaların sosyal medyadaki algıyı keşfetmeleri ihtiyacını doğurmuştur ve sosyal medyada algı analizi çalışmalarına ön ayak olmuştur. Milyonlarca insan hergün sosyal medyada satın aldığı ürünler, kullandığı hizmetler ve markalar, kurumlar, ekonomi, siyaset, spor vb. konular hakkında olumlu ve olumsuz fikirlerini paylaşmaktadır. Sosyal medya siteleride (Facebook, Twitter, Linkedin ve Google+, vb.) insanların yayınladığı bütün bu içeriği sağladıkları servisler aracılığı ile paylaşmaktadır. Ancak, paylaşılan bu ham veride anlamsal bir etiket bilgisi (konu, olumluluk, olumsuzluk, vb.) bulunmamaktadır. Dolayısı ile bir konu hakkındaki algının ne olduğu bilgisi sosyal medyada yığın halindeki verinin içerisinde gizlidir. Bu çalışmada Twitter'ın sağladığı API ile verilere erişimin kolay, iletilerin içeriğinin kısa ve açık ifadeler bulundurması nedeniyle Twitter verileri üzerinde yapılmıştır. Sosyal medya verilerinin işlenmemiş haliyle üzerinde çalışılması oldukça zordur ve incelendiğinde çoğunluğunun hatalı yazılmış kelimeler, kısaltmalar ve günlük konuşma dilinde kullanılmayan sosyal medyaya özgü jargon sözcüklerden oluştuğu gözlemlenmiştir. Bu büyük miktardaki veri işlenmemiş ham yapısıyla tek tek incelenmeye çalışıldığında insan algısıyla bile anlaşılması güçtür. Bu nedenle verilerin doğal dil işleme yöntemleri ile süzülmesi ve işlenmesi gerekir. Twitter'dan toplanan veriler ile doğal dil işleme ve veri madenciliği ile pek çok araştırma yapılmakta ve endüstride projeler gerçekleştirilmektedir. Bunlara örnek olarak salgınları önceden tahminleyen [1], ilaçların bilinmeyen yan etkilerini keşfeden [2], algının zamanla değişimini tahminleyen [3], turistik bir beldeye gelen turistlerin tivitleri üzerinde algı analizi yapan [4] çalışmalar verilebilir. Makalenin ikinci bölümünde Twitter'dan verilerin çekilmesi ve temizlenmesinden, üçüncü bölümde geliştirilen sistemden ve dördüncü bölümde de deneysel sonuçlardan bahsedilmektedir.

978-1-4799-4874-1/14/$31.00 ©2014 IEEE

II. TWITTER VERİLERİ

Twitter, kullanıcıların "tivit" adı verilen en fazla 140 karakterden oluşan mesajlar göndermesine ve okumasına izin veren sosyal ağ ve mikroblog sitesidir. 2006 yılında Jack Dorsey tarafından kurulan Twitter, günümüzde en çok ziyaret edilen 10 internet sitesinden biridir.

A. Twitter Jargonu

Twitter'ın 140 karakter sınırlamasından dolayı, kullanıcılar yazdıkları tivitlere mümkün olabildiğince çok bilgi sığdırma stratejileri geliştirmişlerdir [5]. Bu stratejilerin başında gelen diyez (#) karakteri ile yazılan belirli bir başlıkla tivitler etiketlenebilir. Karakter sınırlamasının Twitter jargonuna bir etkisi de çok fazla kısaltmanın varlığıdır ('slm'-selam, 'hrksn'-harikasın, lol-kahkaha).

B. Verilerin Toplanması

Bu çalışma kapsamında veriler, geliştirilen sosyal medya arama katmanı yardımıyla Twitter API'sine erişerek dokuz farklı alan için belirlenen anahtar sözcükler ile 07.06.2013–13.07.2013 tarihleri arasında sorgulanarak toplanmıştır.

C. Verilerin Saklanması ve Etiketlenmesi

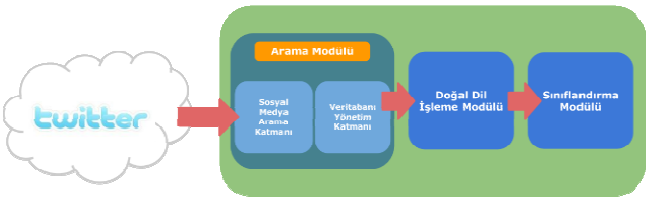
Büyük miktarda verilerin saklanıp üzerinde hızlı arama yapabilmesi için çalışmada NoSQL veritabanları içerisinden Solr seçilmiştir ve dokuz farklı alandan toplanan 8.321 adet tivit işlenmeden ham halleriyle gönüllü kişilerce etiketlenmiştir. Etiketlenen veriler alanları ile birlikte Tablo I'de verilmiştir. Etiketleme işlemi kişisel algıya göre farklılık gösterebileceği için aynı tivitin birden fazla kişi tarafından etiketlenmesi sağlanmıştır. Verilerin farklı alanlardan toplanmasındaki amaç, alana bağımlılığı azaltarak genel anlamda sosyal medya verileri üzerinde başarılı ve alan değişimine dayanıklı bir model oluşturmaktır.

No	Alan Adı	Olumlu	Olumsuz	Nötr
1	Telekom.	161	685	172
2	Sağlık	92	314	299
3	Finans	115	498	376
4	Spor	196	417	566
5	Sigorta	468	686	690
6	Gıda	30	102	299
7	Otomotiv	35	139	185
8	Gayrimenkul	39	83	211
9	Siyaset	208	794	461

Tablo I: Etiketli sosyal medya verileri

III. GELİŞTİRİLEN SİSTEM

Bu çalışmada sosyal medya erişim katmanı, veritabanı yönetim katmanı, doğal dil işleme modülü ve sınıflandırma modülü olmak üzere dört uygulama bileşeninden oluşan bir sistem geliştirilmiştir (Şekil 1).



Şekil 1: Uygulama bileşen modülü

A. Sosyal Medya Erişim Katmanı

Sosyal medya erişim katmanı çalışmanın içerisinde kütüphane olarak uygulamayla entegre çalışmaktadır. Bu modülün işlevi sosyal medya kanallarıyla veri alışverişini sağlamaktır. Bu iş için sosyal medya kanallarının özel uygulama parçaları kullanılmaktadır. Bu modülde belirlenmiş anahtar sözcükler Twitter4J kütüphanesinin sağladığı arayüz ile aratılarak Twitter verileri toplanmıştır [6]. Bu arama işlemi yazılan bir zamanlayıcı uygulama parçası ile periyodik hale getirilmiş ve uygulama kullanıcısının belirlediği periyotlara göre sürekli arama yapmaktadır.

B. Veri Tabanı Yönetim Modülü

Veri tabanı yönetim katmanı Solr veritabanı üzerindeki verilerin eklenmesi, güncellenmesi, silinmesi ve arama işlemlerini yöneten modüldür. Solr, ilişkisel olmayan (NoSQL) verilerin saklandığı bir veritabanıdır. Hızlı sorgulama yapma ve büyük miktarda verileri yüksek performansla yazma ve okuma özelliğinden dolayı tercih edilmiştir. Solr, verileri kendi özel formatıyla dosya sistemi üzerinde tutmaktadır.

C. Doğal Dil İşleme Modülü

Twitter verilerinin işlenmesinde kelime ve n-gram tabanlı iki farklı yöntem kullanılmıştır.

Kelime Tabanlı Yöntem. Türkçe bir tivit incelendiğinde içerdiği kelimeler açısından oldukça çeşitliliğe sahiptir. Kullanılan bir kelimenin değişik çekimlere sahip halleri, farklı yapım ekleri ile farklı anlamlara gelmesi, sosyal medyada kullanılan jargonun Türkçe'deki kelimelerin yerini alması bu çeşitliliğe verilebilecek örneklerden birkaçıdır.

Türkçe tivitlerin doğru sınıflandırılabilmesi için kelimeler kök ve eklerine ayrıştırılmıştır. Kelimelerin ayrıştırılması için Türkçe'de oldukça başarılı olan açık kaynak kodlu Zemberek [7] kütüphenesi kullanılmıştır. Doğru yazılmışTürkçe kelimelerin büyük bir çoğunluğu çözümlenerek gövdeleri elde edilmiş ve olumsuzluk eki dışındaki çoğul, zaman, hal, iyelik, ilgi gibi çekim ekleri temizlenmiştir. Olumsuzluk ekleri kelimenin anlamını değiştirdiği için kelimelerin olumsuz halleri de özellik olarak kabul edilmiştir ve bir dönüşüm uygulanarak

978-1-4799-4874-1/14/$31.00 ©2014 IEEE

olumsuzluk ifadesi kelimenin başına "not_" ön eki eklenerek sağlanmıştır.

Hatalı yazılan kelimeler için, Zemberek kütüphanesinin önerme fonksiyonu kullanılmıştır. Bu öneri fonksiyonuna ek olarak Apache Lucene'in [8] içerisinde yer alan uzaklık algoritması da kullanılmaktadır. Bu algoritma yaptığı tekil öneriler ile sosyal medya verilerinde daha başarılı olmuştur ve başarısından dolayı kelimelerin düzeltilmesinde bu yöntem tercih edilmiştir.

Twitter'ın karakter sayısındaki sınırlama insanların duygu ve düşüncelerini ifade etmek için yetersiz kalmış ve bu durum kullanılan kelimelerin farklılaşmasına neden olmuştur. Jargon kelimelerin bir durumu ifade etmeyi kolaylaştırdığı için çok sıklıkla kullanıldığı gözlemlenmiştir. Veriler üzerinde yapılan gözlemler ile jargon sözlüğü oluşturulmuş ve sözlükteki ifadelerin herbiri sınıflandırıcı için bir özellik olarak kabul edilmiştir.

Verilerin tekrar tekrar işlenmemesi ve sınıflandırıcıya verilen eğitim verisinde tekrar eden verilerin engellenmesi için "RT" ile başlayan tivitler, birinden bahsetme, söz etme durumları için "@kullanıcıadı" şeklinde kullanılan 'mention'lar ve durak sözcükler (stop words) bir olumluluk ya da olumsuzluk ifade etmediği için tivitlerin içerisinden çıkarılmıştır.

Tivitler üzerinde yapılan incelemeler sonucunda içinde jargon, deyim ve kelime kalıplarının, olumlu ve olumsuz ifadelerin bulunduğu 589 adet farklı özellik çıkarılmıştır.

N-gram Model. N-gram'lar bir dizi verinin ardışık sıralı n elemanlı alt kümelerinin bulunduğu yöntemdir [9]. Bu çalışmada n-gram birimleri olarak karakterler kullanılmıştır. Birim olarak karakter seçilmesi durumunda Twitter jargonunda kullanılan çeşitli kısaltma ve yansıma ifadelerinin yakalanması mümkün görülmektedir. Ayrıca, sık kullanılan olumlu ve olumsuz ifadelerin içinde tekrarlanan karakter dizilerinin keşfedilmesinde de faydalı olacağı düşünülmüştür.

Tivitler, 2-gram, 3-gram olmak üzere iki farklı şekilde sıralı karakter dizileri haline getirilmiştir. Böylece elde ettiğimiz karakter ikilileri ve üçlüleri, modellenecek sınıflandırıcı için birer özellik haline getirilmiş, etiketlenen 8.321 tivitin ayrıştırılmasıyla 1.604 adet 2-gram, 10.067 adet 3-gram elde edilmiştir.

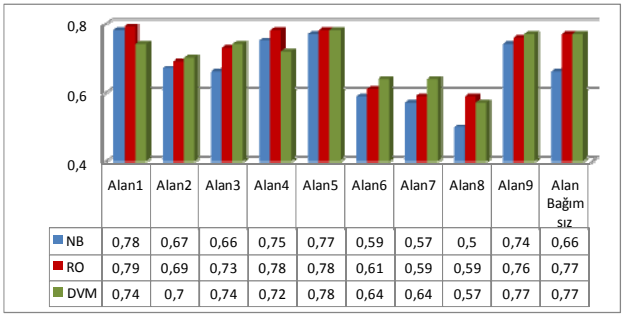
D. Sınıflandırma Modülü

Sınıflandırma modülü belirli yöntemler izlenerek işlenmiş veriler ile sınıflandırma algoritmalarının 10 katlı çapraz doğrulama ile eğitildiği ve test edildiği modüldür. Bu modülde sistemin başarısını ölçmek için Rastgele Orman, Naïve Bayes ve Destek Vektör Makinesi sınıflandırıcıları kullanılmıştır. Deneysel çalışmalarda kullandığımız Rastgele Orman tüm değişkenler arasından en iyi dalı kullanarak her bir düğümü dallara ayırmak yerine, her bir düğümde rastgele olarak seçilen değişkenler arasından en iyisini kullanarak her bir düğümü dallara ayırır. Her bir veri seti orijinal veri setinden yer değiştirmeli olarak üretilir. Sonra rastgele özellik seçimi kullanılarak ağaçlar geliştirilir [10]. Naïve Bayes yöntemi ise, istatistiksel bir yöntem olup, Bayes teoreminin bağımsızlık önermesiyle basitleştirilmiş halidir [10]. Bu önerme örüntü tanımada kullanılacak her özelliğin istatistiksel açıdan bağımsız olması gerekliliğini ortaya atar. İstatistiksel öğrenme teorisine ve yapısal risk minimizasyonuna dayanan Destek Vektör Makineleri, sınıfları birbirinden ayıran marjini en yüksek yapacak doğrusal fonksiyonu seçmeyi hedeflemektedir [11]. Veriler doğrusal olarak ayrılamadığında, veriyi doğrusal olmayan haritalama ile orijinal girdi uzayından daha yüksek boyuttaki bir uzaya aktararak sınıflandırma problemini çözer.

IV. DENEYSEL SONUÇLAR

Sınıflandırma işleminde bazı durumlarda özellik uzayının boyutunun artması sınıflandırıcının başarısını azaltıcı etki yapar. Bu nedenlerden dolayı çalışmamızda özellik seçimi yapılmıştır ve yöntem olarak korelasyon tabanlı özellik seçimi (CFS) kullanılmıştır. Kelime tabanlı yöntem ile eğitilen alan bağımsız veri setinde başarının düşmesine neden olan korelasyon tabanlı özellik seçimi, alan bağımlı veri setinde n-gram yöntemi ile eğitildiğinde başarının artmasını sağlamıştır. CFS kullanılarak seçilen özellikler kullanıldığında başarının %8'e kadar arttığı gözlemlenmiştir. Deneysel sonuçların elde edilmesinde Weka veri madenciliği aracı içerisinde yer alan yöntemler varsayılan parametre değerleri ile kullanılmıştır [12].

2-gram, 3-gram ve kelime tabanlı yöntemler ile eğitilen sistemlerde dokuz farklı alan ve bütün alanları içeren veri setleri kullanılmıştır. Destek Vektör Makinesinin alanların çoğunda ve alan bağımsız veri setinde en iyi f-ölçüm değerlerini verdiği görülmüştür. Bazı alanlarda Naive Bayes başarılı olsa da Rasgele Orman ve diğer yöntemlerden daha az başarılıdır. Şekil 2'de kelime tabanlı yöntem ile tüm veri setlerinde elde edilen sonuçlar yer almaktadır.



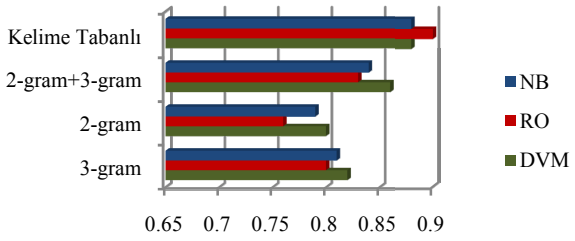
	Alan1	Alan2	Alan3	Alan4	Alan5	Alan6	Alan7	Alan8	Alan9	Alan Bağımsız
NB	0,78	0,67	0,66	0,75	0,77	0,59	0,57	0,5	0,74	0,66
RO	0,79	0,69	0,73	0,78	0,78	0,61	0,59	0,59	0,76	0,77
DVM	0,74	0,7	0,74	0,72	0,78	0,64	0,64	0,57	0,77	0,77

Şekil 2: Kelime tabanlı yöntem başarı grafiği

Şekil 2'deki karşılaştırmaların tümünde korelasyon tabanlı özellik seçimi kullanılmış ve bir istisnai durum hariç Destek Vektör Makinesinin en başarılı sınıflandırıcı

978-1-4799-4874-1/14/$31.00 ©2014 IEEE

olduğu görülmüştür. Bu istisnai durum özellik seçimi kullanılmadan kelime tabanlı yöntemler çalıştırıldığında Rasgele Orman sınıflandırıcısının alan bağımsız veriler ile eğitildiğinde 0.90'lık f-ölçüm değeri ile elde edilen en iyi sonucu vermesidir. Alan 6, 7 ve 8'de başarının düşük çıkması olumlu ve olumsuz etikete sahip tivit sayılarının diğerlerine göre düşük sayıda olmasıdır.

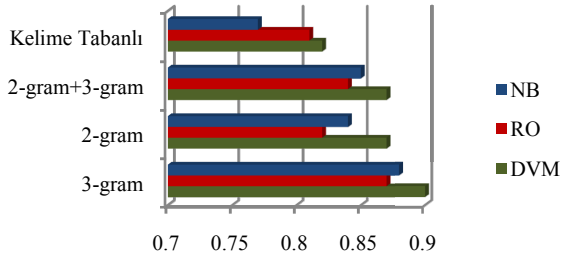


Şekil 3: Alan bağımsız sınıflandırma başarı grafiği

Şekil 3 incelendiğinde alan bağımsız verilerde 2-gram ve 3-gram yöntemlerinin tivitleri sınıflandırmada seçici özelliğinin kelime tabanlı yönteme göre daha az olduğu görülmektedir. 2-gram ve 3-gram özellik vektörlerini birleştirdiğimizde alan bağımsız verilerde diğer n-gram yöntemlerine göre daha fazla başarılı olduğu görülmektedir. Diğer n-gram yöntemlerindeki gibi burada da en iyi sonucu 0.86 f-ölçüm değeri ile Destek Vektör Makinesi vermektedir.

N-gram yöntemine özellik seçimi uygulandığında başarı artmaktadır ancak, kelime tabanlıda aynı durum söz konusu değildir. Korelasyon tabanlı özellik seçiminin kullanılmadığı kelime tabanlı yöntemde 589 adet özellik ile alan bağımsız verilerle eğitilen sınıflandırıcılar en yüksek başarıyı kelime tabanlı yöntemde vermiştir. Bu yöntemde Rasgele Orman sınıflandırıcısı 0.90 f-ölçüm değeri ile diğer yöntemlerden daha başarılı olmuştur.

Alan bağımlı veriler içerisinden siyaset konulu Alan 9 incelendiğinde alan bağımsız veri kümesi ile elde edilen sonuçların aksine, kelime tabanlı yöntemin n-gram yönteminden daha başarısız olduğu görülmüştür. En yüksek başarının 3-gram yöntemi ile elde edildiği denemede Destek Vektör Makinesi 0.90 f-ölçüm değeri elde etmiştir (Şekil 4).



Şekil 4: Alan 9 için başarı grafiği

3-gram yönteminde DVM için çıkarılan hata matrisi incelendiğinde (Tablo II) olumsuz ve nötr sınıfların yüksek başarı ile sınıflandırıldığı, olumlu sınıfının ise başarıyı düşürdüğü görülmektedir. Sınıflandırmada kullanılan özelliklerin olumlular için diğer sınıflar kadar seçici olmadığı görülmektedir. Diğer alanlar ve alan bağımsız veri seti için de elde edilen hata matrisleri incelendiğinde olumlu veri seti için diğer sınıflar kadar özellik olmadığı gözlenmiştir.

a	b	c	Sınıf
251	59	8	a = olumlu
16	297	4	b = olumsuz
4	6	308	c = nötr

Tablo II: Alan 9-Siyaset için 3-gram hata matrisi

V. SONUÇ

Dokuz farklı alan ve birleşiminden oluşturulan veri setleri ile sistem eğitildiğinde kelime tabanlı yöntem kullanıldığında %89.5'lik başarı alınmıştır. N-gram yöntemi, alan bağımlı verilerde %90 gibi yüksek başarı vermesine rağmen, alan bağımsız verilerde aynı başarıyı gösterememiştir. Bu çalışmada kelime tabanlı bir doğal dil işleme sürecinden geçirilen Twitter verileri ile sınıflandırıcılar eğitilmiş ve %90'a yakın başarılı sınıflandırma performansı gösteren bir sistem geliştirilmiştir.

KAYNAKÇA

[1] Szomszor M.,Kostkova P., De Quincey E. Swineflu, "Twitter predicts swine flu outbreak in 2009", 3rd International ICST Conference on Electronic Healthcare for the 21st century, 2009

[2] Bian, J., Topaloglu, U., Yu, F., "Towards Large-scale Twitter Mining for Drug-related Adverse Events", SHB'12, Hawaii, 2012

[3] Nguyen, L.E., Wu, P., Chan, W., Peng, W., "Predicting Collective Sentiment Dynamics from Time-series Social Media", WISDOM '12, Beijing, China, 2012

[4] Claster, W.B., Dinh, H., Cooper, M., "Naive Bayes and Unsupervised Artificial Neural Nets for Caneun Tourism - Social Media Data Analysis", In 2010 Second World Congress on Nature and Biologically Inspired Computing, Fukuoka, Japan, 2010

[5] Horn, C., "Analysis and Classification of Twitter Messages", Master's Thesis, Graz University of Technology, 2010

[6] "Twitter4J: Twitter API erişimi için geliştirilmiş java kütüphanesi", http://twitter4j.org/

[7] https://code.google.com/p/zemberek/downloads/list

[8] http://lucene.apache.org/

[9] Doğan, S., Diri, B., "Türkçe Dokümanlar İçin N-gram Tabanlı Yeni Bir Sınıflandırma(Ng-ind): Yazar, Tür ve Cinsiyet", TBV Bilgisayar Bilimleri ve Mühendisliği Dergisi, sayı:3, 2010

[10] Witten, I. H., Frank, E., Data Mining, Morgan Kaufmann Publishers, 2005

[11] Alpaydın, E., Yapay Öğrenme, Boğaziçi Üniv. Yayınları, 2011

[12] http://www.cs.waikato.ac.nz/ml/weka/

978-1-4799-4874-1/14/$31.00 ©2014 IEEE