MİKROBLOG KULLANICILARININ KATEGORİZASYONU
CLASSIFICATION OF MICROBLOGGING USERS

M.Özgür Cingiz

Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
mozgur@ce.yildiz.edu.tr

Banu Diri

Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
banu@ce.yildiz.edu.tr

ÖZETÇE

Web 2.0 ile birlikte kullanıcılar Twitter ve diğer sosyal ağlarda ilgi alanlarını, düşüncelerini, başka kullanıcıların yazılarını veya gündelik aktivitelerini takipçileriye paylaşabilmektedir. Bu makalede wefollow.com sitesinden kategori bilgisi alınan spor, teknoloji, eğlence ve ekonomi kategorilerindeki Twitter kullanıcılarının takip edildikleri kategorileri ne ölçüde temsil ettiği Twitter kullanıcı içeriklerinin test verisi olarak eğitim modeline verilmesiyle ölçülmüştür. Eğitim verisi olarak bu dört alandan 2015 haber içeriği haber sitelerinin RSS beslemelerinden alınmıştır. Benzer çalışmalardan farklı olarak Twitter kullanıcılarının içeriklerine ait özellikler RSS haber beslemelerinde yer alıyorsa içerik özelliği olarak alınmış, aksi taktirde özellikler değerlendirmeye katılmamıştır. RSS haber beslemelerinden alınan verilerin kategorik içerik olarak Twitter kullanıcılarından alınan verilerden daha değerli olmasından dolayı eğitim modeli RSS haber beslemeleri kullanılarak oluşturulmuş ve özellik uzayı sadece RSS haber beslemelerinden elde edilmiştir. Performans değerlendirme sonuçlarına göre haber içeriği giren kullanıcıların(botlar) içerikleri normal kullanıcı içeriklerine oranla daha başarılı kategorize edildiği gözlemlenmiştir.

ABSTRACT

Recent advancements in Web 2.0, people can't be regarded as simple content reader, they can also contribute content as writers. This work consists of microblogging and text categorization. Text categorization steps were used in microblogs to find out users whose contributions are more valuable for its related category. 2015 RSS news feeds were taken for training and users' tweets were used as test data. This study also differs from other related projects in selection of features. Selected test feature must be also in training data. If it doesn't, test feature can't be taken as feature in test data. In conclusion, contents of news bots in Twitter have more categorical content than ordinary microbloggers.

1. GİRİŞ

Web 2.0 ile birlikte kullanıcılar sadece okuyucu olarak değil aynı zamanda katılımcı olarak da internette yer almaktadır. Bu gelişim sosyal ağ, blog ve mikroblog gibi kavramları kullanıcılarla tanıştırmıştır. Twitter en yaygın olarak kullanılan mikroblog olmakla birlikte kullanıcılar ilgi alanlarını, fotoğraflarını veya gündelik aktivitelerini takipçileriyle paylaşmaktadır. Twitter, kullanıcılara 140 karakter ile sınırlı içerik paylaşım olanağı sunmaktadır. Bu paylaşılan içeriklere tweet adı verilmektedir.

Twitter'da ilgi duyulan alanlara ait içerik giren kullanıcıların keşfi için wefollow.com, twitterholic.com ve Twitter içi gibi arama uygulamaları kullanıcılara yardımcı olmaktadır. Kullanıcıların karakter sayısı kısıtından dolayı yazdığı içeriklerde yaptığı kısaltmalar ve sosyal ağlara özgü yazı dili kullanımı kullanıcıların bir alana kategorize edilmesini zorlaştıran faktörlerdir.

Bu alanda yapılan çalışmaları mikroblogların kategorizasyonu ve mikrobloglar ile veri madenciliği olarak iki gruba ayırabiliriz. Bu alanda Değirmencioğlu [1] içeriklerdeki kelime-etiket, kelime-kullanıcı ve etiket-kullanıcı ilişkisini inceleyerek kullanıcıların ortak ilgi alanlarına göre bir sosyal ağ önermiştir. Yurtsever [2] semantik kaynak kullanarak kullanıcıların girdiği içeriklere göre kullanıcıları kategorize etmiştir. Akman [3] 150 kullanıcının girdiği içeriği alarak seçilen grupların genel karakteristik özelliklerini çıkarmıştır. Aslan [4] ise, haber örüntüsü kullanarak Twitter'da yer alan kullanıcıların içerik olarak haber örüntüsüne benzerliğine bakarak Twitter üzerinden haber içeriği yayınlayan kullanıcıları bulmaya çalışmıştır.

Metin kategorizasyonuyla ilgili çalışmalar uzun yıllardır yapılmaktadır. Pilavcılar [5] metin madenciliği teknikleriyle metin sınıflandırma yapmıştır. Hem mikrobloglar hem de metin kategorizasyonunu birlikte kullanan Güç [6], Twitter kullanıcılarından oluşan sınıf etiketi belli listelerin, sınıflarına uygunluğunu metin madenciliği teknikleriyle araştırmıştır.

Makalenin ikinci bölümünde kullanılan veri seti, üçüncü bölümde tasarlanan model ve modelin adımları anlatılmıştır. Dördüncü bölümde ise, test sonuçları ve sonuçların yorumları verilmiştir.

2. VERİ SETİ

Bu çalışmada BBC, CNN, SKYNews gibi haberlerini RSS haber beslemesi olarak sunan sitelerden RSS4j Java kütüphanesiyle haberler elde edilerek haberlerden oluşan bir eğitim kümesi oluşturulmuştur. Haberlerin çekiminde ilgili sitelerin sunduğu spor, ekonomi, teknoloji ve eğlence alanındaki haber geri beslemeleri alınmıştır. 544'ü eğlence, 470'i teknoloji, 548'i ekonomi ve 543'ü spor haberi olan toplamda 2015 haber, RSS4j kütüphanesiyle ilgili web sayfalarından çekilmiştir.

Test verilerinin oluşturulması içinde wefollow.com kullanılarak kategori bilgisi bilinen 26 haber içeriği yayınlayan kullanıcı(bot) ve 27 tane de normal kullanıcı olmak üzere toplam 53 kullanıcının farklı sayılardaki içeriği Twitter4j java kütüphanesi kullanılarak alınmıştır. Test aşamasında spor, teknoloji, ekonomi ve eğlence kategorilerinden her kategori için 4 kullanıcı seçilmiştir. Bu işlem hem haber içeriği giren

978-1-4673-0056-8/12/$26.00 ©2012 IEEE

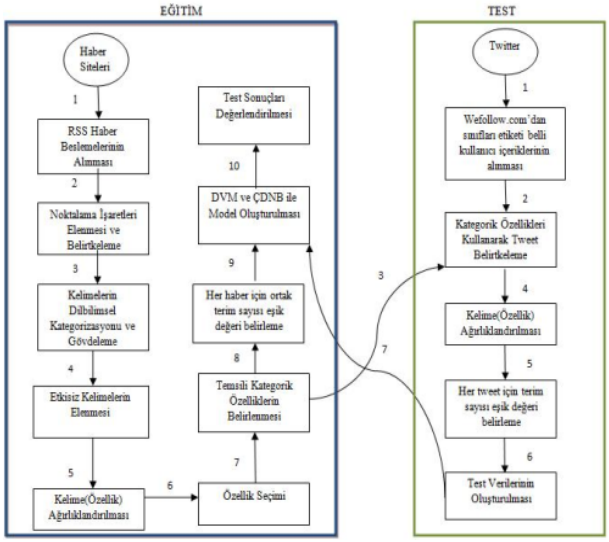
kullanıcılar, hem de normal kullanıcıların ayrı ayrı seçilmesiyle gerçeklenmiştir.

3. GELİŞTİRİLEN SİSTEM

Şekil 1'de test ve eğitim aşamalarında geliştirilen sisteme ait adımlar gösterilmiştir. Bu adımlar eğitim aşamasında sırasıyla eğitim modelinin oluşturulması için eğitim verilerinin toplanması, ön işleme adımları, özelliklerin ağırlıklandırılması, boyut azaltmak için özellik seçimi, tüm haberler için bir terim sayısı eşik değeri belirlenmesi ve çıkarılan özelliklerden oluşan haberlerin Destek Vektör Makinaları (DVM) ve Çok Değişkenli Naive Bayes (ÇDNB) sınıflandırıcıları ile model oluşturulmasıdır.

Test verilerinin oluşturulması için wefollow.com'dan kategori bilgisi alınan 26 haber içeriği yayınlayan kullanıcı(bot), 27'de normal kullanıcı olmak üzere 53 kullanıcının farklı sayılardaki içeriği alınmıştır. Test aşamasında spor, teknoloji, ekonomi ve eğlence kategorilerinden her kategori için 4 kullanıcı seçilmiştir. Bu işlem hem haber içeriği giren kullanıcıların hem de normal kullanıcıların ayrı ayrı seçilmesiyle gerçeklenmiştir. Her kategoriden 4'er tane alınan kullanıcıların oluşturduğu 16 kullanıcıdan alınan içerik belirtkelemeden (tokenization) geçerek kelimeler çıkarılmıştır. Çıkarılan kelimeler eğitim kümesinde yer almamışsa test özelliği olarak değerlendirilmiştir. Test özelliği olarak alınan veriler ise kendi içerisinde ağırlıklandırıldıktan sonra her bir kullanıcı içeriği için 3 ayrı terim sayısı eşik değeri belirlenmiş ve testler bu eşik değerlerine göre yapılmıştır.

Elde edilen kullanıcı girdileri Destek Vektör Makinaları ve Çok Değişkenli Naive Bayes ile oluşturulan eğitim modeline verilerek test sonuçları elde edilmiştir. Bu bölümde geliştirilen sistemde hangi yaklaşımların kullanıldığı detaylı olarak anlatılmıştır.



Şekil 1: Geliştirilen Sistemin Adımları

3.1. Ön İşlemler

Ön işlemler noktalama işaretlerinin atılması, belirtkeleme, özelliklerin dilbilimsel seçimi, özelliklerin gövdelenmesi ve etkisiz kelimelerin elenmesi aşamalarından oluşmaktadır.

3.1.1. Belirtkeleme ve Noktalama İşaretleri Elenmesi

Metin tipindeki verilerde ilk yapılan işlem metinlerdeki özelliklerin çıkarılması işlemidir. Atomik hale gelen metin tipindeki veriler n-gram'lar, söz öbekleri veya kelimeler olabilir. RSS haber beslemeleri alındıktan sonra haberlerdeki noktalama işaretleri atılır ve sonra belirtkeleme ile haberler kelimelerine ayrılır.

3.1.2. Özelliklerim Dilbilimsel Seçimi ve Gövdeleme

Daha önceki çalışmalarda [7][8] metin kategorizasyonunda özellik olarak alınan kelimelerin dilbilimsel etiketleri ayrı ayrı incelenmiş ve bu çalışmalara göre isim ve fiiller kategorik değer olarak metinde geçen sıfat, zamir, edat sözcük türlerindeki kelimelerden daha yüksek çıkmıştır. Bu çalışmada da kelimelerin sözcük türleri Stanford Üniversitesi tarafından geliştirilen POS etiketleyicisi [1] kullanılarak çıkarılmıştır. Sözcük türlerine göre özellik olarak bu çalışmada sadece isim ve fiiller seçilmiştir. Ayrıca kelimelerin gövdeleme işlemi de aynı Stanford Üniversitesi'nin gövdeleme kütüphanesiyle gerçekleştirilerek kelimeler eklerinden ayrılmıştır. Böylece farklı zaman çekimlerindeki, farklı formatta gösterilen fakat gövde olarak aynı olan kelimeler belirlenmiştir.

3.1.3. Diğer Ön- İşlemler

Kategoride yer alan ve sıkça geçtiği için kategorik belirleyiciliği olmayan etkisiz kelimelerde (stop words) bu aşamada elenir.

3.2. Özelliklerin Ağırlıklandırılması

Terim özellik uzayında her metin dosyası bir vektöre, her kelime ise ilgili metin dosyasına ait bir boyuta karşılık gelmektedir. Metinde geçen her kelime sıfırdan farklı bir ağırlık değeri alacaktır. Bu çalışmada metin madenciliğinde en çok tercih edilen tf-idf ağırlıklandırma yöntemi kullanılmıştır. Önceki ağırlıklandırma çalışmalarına [9][10]bakıldığında terimlerin ağırlık seçimlerinin Destek Vektör Makinaları'nın çekirdek fonksiyonunu seçmekten daha önemli olduğu belirtilmiştir. Yine aynı çalışmalarda tf-idf ağırlıklandırmasıyla kullanılan özelliklerin sınıflandırma başarıları ikili ağırlıklandırmadan daha yüksek çıkmıştır.

Tf-idf ağırlıklandırmasında "tf" terim frekansını (1) yani terimlerin bulunduğu metin dosyasında kaç kez geçtiğini, idf (2) ise aynı terimin tüm metin dosyalarında kaç kez geçtiğini gösteren ters doküman frekansıdır. Denklem 3'te görüldüğü üzere bu iki değerin çarpımıyla terimlerin vektör uzayındaki tf-idf ağırlıkları elde edilir. Burada metinde bulunan terim frekansı frekans(d,t), terimin ters doküman değeri idf(t) şeklinde gösterilmiştir.

Bu makalede eğitim ve test verilerine ait özellikler tf-idf ağırlıklandırma metodu kullanılarak hesaplanmıştır.

$$tf(d,f)=\begin{cases} 0, \text{eğer frekans } (d,t)=0 \\ 1+\log(1+\log(\text{frekans}(d,t))) \end{cases} \quad (1)$$

$$idf(t)=\log \frac{n}{|df t|} \quad (2)$$

$$tfidf(d,t)= tf(d,f)*idf(t) \quad (3)$$

[1] http://nlp.stanford.edu/software/tagger.shtml

3.3. Özellik Seçimi ve Terim Sayısı Eşik Değeri

Vektör uzay modelinde vektörler çok sayıda özellikten oluşmakla birlikte, RSS haber beslemeleri kısa haberlerden oluştuğu için her bir haber az sayıda özellik içerir. Tüm RSS haber beslemeleriyle oluşan özellik uzayında özellik sayısı çok olmasına rağmen, vektör modelinde haberlerde geçen terim ağırlığı sıfırdan farklı kelime sayısı çok azdır. Çok sayıda özellikten oluşan eğitim veri seti ve seyrek özellik değerlerine sahip haberler üzerinde çalışmak sınıflandırıcıların performansını düşürür. Bu yüzden eğitim setinde özellik seçimi yöntemleri kullanılır, az sayıda kelimeden oluşan haberler içinde terim eşik değerleri kullanılır.

Özellik seçimi, kategorileri en iyi temsil eden alt özelliklerin belirlenmesi işlemidir. Özellik seçimiyle ilgili pek çok yaklaşım bulunmaktadır. Bu çalışmada en çok kullanılan ve önceki çalışmalarda [11][12] iyi sonuçlar veren özellik seçim algoritmaları olan Ki-kare İstatistiği ve Bilgi Kazanımı (Information Gain) yöntemleri kullanılmıştır. RSS haberlerinden elde edilen özelliklerden Ki-kare ile çıkarılan özelliklerin sınıflandırma başarısı, Bilgi Kazanımı yöntemiyle elde edilen özelliklerin başarı oranından daha yüksek olduğu gözlemlenmiştir. "*t*" özelliğinin *c* sınıfındaki Ki-kare istatistik değeri denklem 4'te verilmiştir. Ki-kare istatistiği özelliğin sadece sınıfta yer almasına göre değil, aynı zamanda özelliğin olmadığı durum olasılığını da değerlendirdiğinden seyrek özellik uzayında daha iyi sonuç vermiştir.

$$X^2(t.c)=N.\frac{[P(t,c).P(\bar{t},\bar{c})-P(t,\bar{c}).P(\bar{t},c)}{P(t).P(\bar{t}).P(c).P(\bar{c})}]^2 \quad (4)$$

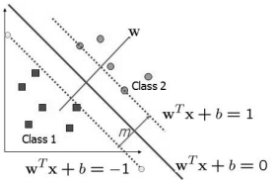
Çalışmada 2105 RSS haber beslemesinden elde edilen 7210 özellik Ki-kare yöntemiyle, 1277 özelliğe indirilmiştir. Ayrıca özellik olarak sadece isim ve fiiller alınmış, etkisiz kelimelerin elenmesiyle de bazı haber verileri özelliksiz kalacağı için her haberin en az 1 özellik taşıması için bir eşik değeri konulmuştur.

3.4. Sınıflandırma

Modelin kurulması ve test aşamasında Çok Değişkenli Naive Bayes ve Destek Vektör Makinaları kullanılmıştır. Özellikle bu iki sınıflandırıcının seçilmesinin nedeni pek çok sınıflandırma algoritmasıyla metin madenciliği üzerinde yapılan sınıflandırma çalışmalarında [13][14] DVM sınıflandırıcısının k-En Yakın Komşuluk, En Küçük Kareler, Yapay Sinir Ağları, Naive Bayes gibi sınıflandırıcılardan daha iyi sonuç verdiği gözlemlenmiştir. Çok Değişkenli Naive Bayes ise özellikle metin kategorizasyonunda başarılı sonuçlar verdiği için bu çalışmada kullanılmıştır.

3.4.1. Destek Vektör Makinaları

Destek Vektör Makinaları karar düzlemlerinin ideal şekilde belirlenmesi prensibine dayanır. İdeal karar sınırı ayırdığı sınıflara ait verilere mümkün olduğu kadar karar düzleminden uzak olmalıdır.



Şekil 2: Destek Vektörler ve Karar Sınırı

Kesikli doğrular her sınıfa ait sınırları göstermektedir. Sınırlar üzerindeki örneklere destek vektör denilmekle birlikte karar sınırları sadece bu vektörler tarafından belirlenmektedir. İdeal bir karar düzlemi belirlemek için şekil 2'de gösterilen *m* değerinin maksimize edilmesi gerekmektedir.

$$m=\frac{2}{||w||} \quad (5)$$

Bir başka deyişle ||w|| değerinin minimize edilerek karar düzleminin ayırdığı sınıflara ait sınır değerlerine uzaklık maksimum yapılmaktadır. Sınıf etiketleri $y_i \in \{1,-1\}$ olmak üzere tüm veri setindeki değerlerin $\{x_1, x_2, x_3,..,x_n\}$ denklem 6'daki gibi belirlenmiş karar sınırı tarafından sınıflandırılması gerekir.

$$yi(w^t x+b)\geq 1 \quad (6)$$

Makale kapsamında Destek Vektör Makinası çekirdek fonksiyonu olarak doğrusal çekirdek fonksiyonu kullanılmıştır.

3.4.2. Çok Değişkenli Naive Bayes

Klasik Naive Bayes kelimelerin yani özelliklerin aynı sınıfta yer alma olasılıklarını birbirinden bağımsız olarak düşünür. Çok değişkenli Naive Bayes klasik Naive Bayes'ten farklı olarak kelimelerin metinlerde geçme sıklığının terimlerin ilgili kategoriye ait olma olasılığını hesaplamada kullanır. Bu sınıflandırıcıda her terimin metinde bulunma frekansı, diğer terimlerin aynı metinde bulunma frekansından bağımsız olduğu düşünülmektedir. Denklem 7'de metinde yer alan tüm terimlerin sınıflara ait olma olasılıkları birbirleriyle çarpılarak en yüksek olasılık değerini veren sınıf, sınıf etiketi olarak belirlenirken denklem 8'de ÇDNB sınıflandırıcısının Laplace yumuşatmasıyla "c" sınıfındaki "t" teriminin olasılık değeri verilmiştir.

$$c = \arg\max_{c \in C} \; P(c|d) = \arg\max_{c \in C} \; P(c) \prod_{1\leq k\leq V} P(t_k|c) \quad (7)$$

$$P(t|c)=\frac{T_{ct}+1}{\sum_{t' \in V} T_{ct}+B'} \quad (8)$$

4. TEST SONUÇLARI VE DEĞERLENDİRME

RSS haber beslemelerinden alınan 2015 haber ile eğitim modeli oluşturulmuştur. Wefollow.com'dan kategori bilgisi alınan 26 haber içeriği yayınlayan kullanıcı ve 27 normal kullanıcı olmak üzere toplam 53 kullanıcının farklı sayılardaki içeriği alınarak eğitim modeline test verisi olarak verilmiştir. Bu işlem gerçekleşirken haber içeriği yayınlayan 26 kullanıcının 5'i ekonomi,5'i teknoloji, 5'i eğlence, 11'i spor kategorisinde içerik giren kullanıcılardır. Spor kullanıcıların fazla sayıda olma neden spora ait alt kategorilerin fazla olmasından kaynaklanmaktadır. Normal kullanıcıların ise 7'si ekonomi, 7'si eğlence, 7'si spor ve 6'sı teknoloji sınıfında yer almaktadır. Test aşaması 3 farklı terim sayısı eşik değeri seçilerek yapılmıştır. Her kategoriye ait 4 kullanıcı aynı kategorideki kullanıcılar arasından seçiler test veri kümeleri oluşturulmuştur. Üç farklı terim sayısı eşik değeri için her kategoriden dörder kullanıcı seçimiyle oluşturulan beş veri seti hem haber içeriği giren kullanıcılar için hem de normal kullanıcılar için ayrı ayrı oluşturulmuş ve 30 farklı test veri seti hazırlanmıştır.

Tablo 1'de 30 farklı test veri kümesi için eğitim modeline verilen testlerin F-ölçüm değerleri verilmiştir. F-ölçüm değeri

kesinlik ve geri çağırma oranlarının harmonik ortalaması olarak kullanılan bir performans ölçü değeridir.

Eğitim verilerinin modellemesi ve test işlemi için ÇDNB ve DVM sınıflandırıcıları kullanılmıştır. F-ölçüm değerleri verilirken ilk değer ÇDNB'den elde edilen, ikinci değer ise DVM'den elde edilen F-ölçüm değerleridir.

Tablo 1: Test Sonuçları, F-Ölçüm Değerleri

Haber Tweetleri (botlar)	**Terim Sayısı Eşik Değeri**					
ÇDNB | DVM F-ölçüm	**>2**		**>3**		**>4**	
1.Veri Kümesi	86,0	70,2	90,3	76,4	92,3	80,7
2.Veri Kümesi	86,9	70,9	91,7	76,3	95,6	78,8
3.Veri Kümesi	85,7	68,8	90,6	72,9	92,2	73,5
4.Veri Kümesi	87,3	71,7	91,1	73,6	96,3	78,5
5.Veri Kümesi	86,3	68,6	92,6	75,2	96,2	79,9
Normal Kullanıcı Tweetleri	**Terim Sayısı Eşik Değeri**					
ÇDNB | DVM F-ölçüm	**>2**		**>3**		**>4**	
1.Veri Kümesi	78,0	64,9	81,0	67,4	85,5	74,7
2.Veri Kümesi	79,0	64,4	74,1	60,4	89,1	81,4
3.Veri Kümesi	82,5	69,3	78,0	64,8	87,8	78,9
4.Veri Kümesi	76,5	61,9	81,9	69,7	85,3	74,0
5.Veri Kümesi	73,7	63,8	79,5	68,6	86,3	66,8

Tablo 1'de yer alan sonuçlara bakıldığında "Terim Sayısı Eşik Değeri" arttıkça verilerin daha iyi şekilde sınıflandırıldığı gözükmektedir. Eşik değeri artışıyla seyrek özellik uzayının yarattığı sınıflandırma problemi azaltılmıştır. Aynı veri setleri üzerinde ÇDNB sınıflandırıcısıyla oluşturulan model ve modelden elden edilen test ölçüm sonuçlarının DVM sınıflandırıcısından alınan test ölçüm sonuçlarından daha iyi sonuç verdiği görülmüştür.

Tablo 1'den elde edilen sonuçlara göre haber içeriği giren kullanıcılar normal kullanıcılara oranla daha iyi sınıflandırılabilmektedir. Bu da haber içeriği yayınlayan kullanıcıların girdikleri içeriğin sınıfa ait özellikleri daha iyi yansıttığını gösterir. Wefollow.com'dan kategori bilgisi alınan haber içeriği yayınlayan kullanıcılar normal kullanıcılardan daha fazla kategorik içerik girmektedir. F-ölçüm sonuçlarına değerlendirildiğinde en yüksek F-ölçüm değerlerinin hem normal kullanıcılar hem de haber içeriği yayınlayan kullanıcılarda terim eşik sayısının 5 ve 5'ten büyük belirlendiği değerlerde alındığında gözlemlenmiştir. En yüksek F-ölçüm değerleri ÇDNB sınıflandırıcısı kullanarak her haber için terim eşik sayısı 5 ve 5'ten büyük olan verilerin haber içeriği yayınlayan kullanıcılardan seçilmesiyle elde edilmiştir.

Çeşitli eşik değerleri kullanılarak normal kullanıcılar için elde edilen F değerlerinde en yüksek başarı oranı spor kategorisine ait kullanıcıların içeriklerinden elde edilmiş. Normal kullanıcılar içerisinde ekonomi ve teknoloji kategorilerinde içerik giren kullanıcıların başarı oranları daha düşüktür. Haber içeriği giren kullanıcıların sınıflandırma başarıları incelendiğinde ise farklı eşik değerlerinden en iyi başarı oranları ekonomi kategorisinde yazan kullanıcıların içeriklerinden elde edilmiştir. Haber içeriği giren kullanıcılar arasında en düşük başarı oranları ise teknoloji kategorisinde yazan kullanıcı içeriklerinden elde edilmiştir.

5. SONUÇ

Bu makalede Twitter kullanıcılarının içerikleri metin madenciliği teknikleriyle kategorize edilmiştir. Bu işlemi yaparken RSS haber beslemelerinden ilgili 4 alana ait haberler alınarak eğitim modeli oluşturulmuştur. Test verileri olarak kullanıcı içerikleri alınarak eğitim modeline verilmiştir ve kullanıcılar yayınladıkları içeriğe göre sınıflandırmıştır.

Sonuç olarak haber içeriği giren kullanıcılar daha değerli kategorik içerikler girmekte ve sınıflandırıcılar tarafından daha iyi sınıflandırılmaktadır. 30 farklı veri setinden elde edilen F-ölçüm değerleri arasında en yüksek elde edilen F-ölçüm değeri, haber içeriği giren kullanıcılar tarafından girilen ve her içerikte RSS haber beslemelerinden elde edilen en az 5 özellik bulunduran içeriklerin oluşturduğu veri setleridir. Bu veri setlerinde ÇDNB sınıflandırıcısı kullanarak elde edilen F-ölçüm değeri %96,3 gibi yüksek bir değer almıştır.

6. KAYNAKÇA

[1] Degirmencioglu, E. A., *Exploring Area-Specific Microblogger Social Networks*, Master's thesis, Bogazici University, 2010

[2] Yurtsever, E., *Sweettweet:A Semantic Analysis For Microblogging Environments*, Master's thesis, Bogazici University, 2010

[3] Akman, D. S., *Revealing Microblogger Interests By Analyzing Contributions*, Master's thesis, Bogazici University, 2010

[4] Aslan, O., *An Analysis Of News On Microblogging Systems*, Master's thesis, Bogazici University, 2010

[5] Pilavcılar, I. F., *Metin Madenciliğiyle Metin Sınıflandırma*, Master Tezi, Yıldız Teknik Universitesi, 2007

[6] Güç, B., *Information Filtering On Micro-Blogging Services*, Master's Thesis, Swiss Federal Institute of Technology, Zurich, 2010

[7] Chua, S. "The Role of Parts-of-Speech in Feature Selection", *The International Conference on Data Mining and Applications-IAENG*, 2008

[8] Masuyama, T. and Nakagawa, H., "Applying Cascaded Feature Selection to SVM Text Categorization", *The DEXA Workshops*, pp. 241-245, 2002

[9] Lan, M., Sung, S. and Low, H., "A comparative study on term weighting schemes for text categorization", *Neural Networks-IJCCN'05*,.IEEE International Conference, 2005

[10] Leopold, E. and Kindermann, J., *Text categorization with support vector machines. How to represent texts in input space?*, Machine Learning, 2002

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

[12] Zheng, Z., and Srihai, R., "Optimally Combining Positive and Negative Features for Text Categorization", *ICML Workshop*, 2003

[13] Yang, Y. and Liu, X., "A re-examination of text categorization methods", *The Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval* (Berkeley, US), pp. 42–49, 1999

[14] Joachims, T., "Text Categorization with Support Vector Machines: Learning with Many Relevant Features", *The European Conference on Machine Learning (ECML*), 1998