Türkçe Dokümanlar İçin N-gram Tabanlı Yeni Bir Sınıflandırma(Ng-ind): Yazar, Tür ve Cinsiyet

Sibel Doğan[1], Banu Diri[2]

[1,2]Yıldız Teknik Üniversitesi, Bilgisayar Mühendisliği, 34349 İstanbul-Türkiye

[1]sibel.dogan@akbank.com, [2]banu@ce.yildiz.edu.tr

Özet *Bu çalışmada Türkçe bir dokümanın türü, yazarı ve doküman yazarının cinsiyeti Türkçe'nin n-gram modeli kullanılarak belirlenmeye çalışılmıştır. N-gram modelinde 2-, 3-, 4-gram'lar kullanılmış ve üç farklı veri seti üzerinde toplam altı adet özellik vektörü oluşturulmuştur. Naive Bayes (NB), Destek Vektör Makinesi (DVM), Rastgele Orman (RO), K-En Yakın Komşuluk (K-EYK) gibi sınıflandırıcıların yanında geliştirdiğimiz* ***Ng-ind*** *yöntemi kullanılarak testler yapılmış ve başarı performansları birbirleri ile karşılaştırılmıştır.* ***Ng-ind*** *yöntemi cinsiyet ve tür belirlemede diğer yöntemlere göre daha iyi sonuç vermiştir. Bununla birlikte* ***Ng-ind****, tür belirlemede birleştirilmiş sınıflandırıcılardan da daha iyi performans göstermiştir.*

Abstract. *In this study, it is tried to find out a Turkish document's genre, author and document author's gender with using the Turkish n-gram model. In N-gram model, 2-, 3-, 4-grams were used, and total 6 feature vectors were produced on 3 different data set. Some tests were made with the Ng-ind method that we produced near the other classifiers such as Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (K-NN) and the success performances were compared with each other. In spite of the Ng-ind method gave better results than the other ones in gender and genre determination, it showed better performance than* the compounded classifiers in genre determination.

1 Giriş

Dünya bilgi çağına girdiğinden beri, gelişen ülkelerde insanlar tarafından kullanılan bilginin miktarı çok hızlı bir şekilde artmaktadır. Bu bilgileri sağlıklı bir şekilde kullanabilmek ve kısa sürede erişebilmek için birbirleri ile ilişkili olan bilgileri bulup aynı bilgi topluluğu içinde toplamak gerekir. Bu da dokümanları sınıflandırmayı gerektirir. Doküman sınıflandırmadaki amaç, bir dokümanın özelliklerine bakılarak önceden belirlenmiş belli sayıdaki kategorilerden hangisine dahil olacağını belirlemektir [1]. Doküman sınıflandırmadaki problemler, dokümanın yazarının kim olduğu, dokümanın türünün ve doküman yazarının cinsiyetinin belirlenmesidir. Bu çalışma, Cavnar ve Trenkle tarafından yazılmış "N-gram Tabanlı Doküman Sınıflandırma" makalesinden esinlenilerek hazırlanmış [2]ve Türk dili için uygulanmıştır.

Yazar tanıma ile ilgili birçok çalışma mevcuttur. Peng ve arkadaşları, n-gram yöntemini kullanarak yazar tanıma yaparken, kullandıkları veri setinin 1'den 10'a kadar olan n-gram'larını çıkarmışlar ve Yunanca, İngilizce, Çince veriler üzerinde denemeler yapmışlardır [3]. Stamatatos ve Kokkinakis'in yaptığı çalışmada ise 10 yazara ait Yunanca iki veri seti kullanılmış ve sırasıyla %85 ve %97'lik doğru sınıflandırma oranı ile başarı sağlanmıştır [4]. Peng ve arkadaşları, aynı veri setini kullanarak yumuşatma tekniğini (absolute smoothing) uygulamış ve en iyi sonucu 3-gram'ları kullandıklarında sırasıyla %74 ve %90 olarak almışlardır [5]. Diri ve Amasyalı n-gram yöntemini kullanarak yaptıkları yazar tanıma çalışmasında (18 yazar için) %83,3'lük bir başarı elde etmişlerdir [6].

Cavnar, tür belirleme ile ilgili n-gram tabanlı çalışmasında %80'lik bir başarı elde etmiştir [2]. Peng ve arkadaşları, karakter seviyeli n-

gram modeline dayalı Naive Bayes sınıflandırma yöntemini kullanılarak altı sınıf için %81'lik bir başarı elde etmiştir [7]. Stamatatos, Kokkinakis ve arkadaşları, yaptıkları tür belirleme çalışmasında farklı özellikler kullanmışlar ve 10 farklı sınıf için aldıkları en yüksek başarı %81 olmuştur [4].

Peng ve arkadaşları da, karakter seviyeli n-gram modelinde n-gram'ların dilden bağımsızlığı ve etkisini kanıtlamak için Yunanca, İngilizce, Japonca ve Çince veriler üzerinde denemeler yapmışlardır [12]. Diri ve Amasyalı, 6 sınıf için yaptıkları tür belirleme çalışmasında %93,6'lık bir başarı elde etmişlerdir [6].

N-gram'ları kullanarak yazarın cinsiyetinin belirlenmesinde Doyle ve Keselj %81'lik bir başarı almışken [13], Nowson ve Oberlander %93'lük bir başarı elde etmiştir [8]. Dupont ise %85,76'lık [9], Diri ve Amasyalı ise %96,3'lük bir başarı elde etmiştir [6].

Makalenin ikinci bölümünde yazarlık özelliklerinden, üçüncü bölümünde de sınıflandırma algoritmalarından bahsedilmiştir. Dördüncü ve beşinci bölümlerde sırasıyla deneysel sonuçlar ve sonuç bölümü yer almaktadır.

2 Yazar Tanıma, Tür ve Cinsiyet Belirleme

Belirli bir kişinin kaleminden dökülen yazılardaki karakter sıklıkları, diğer yazarların yazılarındaki karakter sıklıklarından farklıdır, kısaca her yazarın kendisine özgü bir yazım stili vardır ve bu da yazarlık özelliği olarak adlandırılır. Yazar Tanıma, konudan bağımsız olarak elimizde bulunan dokümanların hangi yazar tarafından yazıldığını tahmin etme işlemi olarak da bilinir.

Belirli bir konuda yazılmış dokümanların karakter tabanlı n-gram sıklıkları birbirine ne kadar yakınsa, farklı konularda yazılmış dokümanların n-gram sıklıkları da birbirinden o kadar uzaktır. Bu düşünceden yola çıkılarak n-gram'ları kullanarak dokümanların konularına göre sınıflandırılması gerçekleştirilmiştir.

Bay ve bayan yazarların yazım üslupları da birbirlerine göre farklılıklar gösterebilmektedir. Bu farklılık karakter tabanlı n-gram özellikleri kullanılarak cinsiyet belirleme çalışmasının da yapılmasına olanak sağlamıştır.

2.1 Derlem (corpus)

Bu çalışmada, dokümanların yazarlarını, türlerini ve yazarların cinsiyetini belirlemede kullanılmak amacıyla üç farklı veri seti oluşturulmuştur. Veri setleri İnternet'te belirli gazete sayfalarından Hürriyet (www.hurriyet.com.tr), Milliyet (www.milliyet.com.tr), Akşam (www.aksam.com.tr) ve Sabah'tan (www.sabah.com.tr) farklı yazarların yazıları toplanarak oluşturulmuştur. Birinci veri seti (Veri Seti – I) cinsiyet belirleme, ikinci veri seti (Veri Seti – II) yazar tanıma ve üçüncü veri seti (Veri Seti – III) ise tür belirleme çalışmalarında kullanılmıştır. Tüm veri setlerinde bulunan her bir yazarın 40 adet yazısı alınmıştır. Kullanılan yazıların boyutları 3 Kb ile 9 Kb arasında değişmektedir.

Veri Seti – I, 10'u bayan ve 10'u bay yazardan oluşan ve her birine ait 40 yazının yer aldığı 800 adet dokümandan oluşmaktadır. Veri Seti-I cinsiyet belirme amacıyla oluşturulduğundan bay ve bayan olmak üzere iki sınıf ile temsil edilmektedir.

Veri Seti – II, yazar tanıma amaçlı kullanılmaktadır. Veri Seti-I' deki dokümanlardan oluşmakta, ancak sınıf sayısı 20 olup, farklı yazar sayısını göstermektedir.

Veri Seti – III, 12 yazardan oluşmakta ve her bir yazarın 40 adet yazısı alınarak toplam 480 adet dokümanı içermektedir. Veri Seti-III dokümanın türünü belirlemek amacıyla kullanılmış olup, sınıf sayısı 6'dır (spor,

magazin, güncel, ekonomi, sağlık ve politika). Her kategoride 2 yazarın yazısı bulunmaktadır.

2.2 N-gram Model

N-gram, bir karakter katarının n adet karakter dilimidir. N-gram tabanlı sınıflandırma yöntemi, doküman içerisindeki karakter tabanlı n-gram'ların kullanım sıklığına dayalı bir işlemdir. Bu çalışmada, n-gram'ın farklı birkaç uzunluğu alınarak 2-, 3- ve 4-gram'lar kullanılmıştır. N-gram'ların elde edilmesinde izlenen yolu bir örnek ile açıklayacak olursak: Örnekte boşluk karakterini göstermek için "_" altçizgi karakteri kullanılmıştır.

Cümlemiz "Yazar Tanıma" ise, bu cümlenin n-gram'ları;

2-gram'lar: "Ya", "az", "za", "ar", "r_", "_T", "Ta", "an", "nı", "ım", "ma"
3-gram'lar: "Yaz", "aza", "zar", "ar_", "r_T", "_Ta", "Tan", "anı", "nım", "ıma"
4-gram'lar: "Yaza", "azar", "zar_", "ar_T", "r_Ta_", "_Tan", "Tanı", "anım", "nıma"
şeklinde çıkarılır.

Tür belirleme, yazar tanıma ve cinsiyet belirleme çalışmalarında en fazla sıklıkta kullanılan n-gram'ların, dokümanları sınıflandırmada önemli etkisi olduğu düşünülerek frekans değeri (kullanım sıklığı) 100'den büyük olan n-gram'lar alınmıştır. Belirlenen eşik değeri parametrik olup, kullanıcının seçimine bırakılmıştır. Yapılan denemelerde ideal eşik değerinin 100 olduğuna karar verilmiş ve test sonuçları bu değer üzerinden alınmıştır.

2.3 Özellik Vektörleri

Sınıflandırma yöntemleri kullanılarak bir dokümanın yazarını, türünü belirlemede ve yazarın cinsiyetini tahmin etme çalışmalarında veri setinde yer alan dokümanlardan belirli özelliklerin çıkarılarak, her bir dokümanın özel bir şekilde ifade edilmesi gerekmektedir. Her doküman, seçtiğimiz eşik değerine göre sayısını bizim belirlediğimiz ***k*** adet özelliğe sahiptir ve ***k*** boyutlu bir vektör ile ifade edilir.

2-gram özellik vektörü

Veri Seti – I, II ve III içerisindeki her bir 2-gram özellik vektörünün *k* değeri sırasıyla 257, 257 ve 217'dir. Yani özellik vektörleri 257/257/217 adet frekans değerine sahiptir ve bu değerler 2-gram özellik vektörünün birer özelliğini temsil etmektedir. Çalışmada bu özellikler kullanılarak oluşturulan vektörden $\mathbf{OV_2}$ olarak bahsedilecektir.

3-gram özellik vektörü

Veri Seti – I, II ve III içerisindeki her bir 3-gram özellik vektörünün *k* değeri sırasıyla 324, 324 ve 208'dir. Çalışmada bu özellikler kullanılarak oluşturulan vektörden $\mathbf{OV_3}$ olarak bahsedilecektir.

4-gram özellik vektörü

Veri Seti – I, II ve III içerisindeki her bir 4-gram özellik vektörünün *k* değeri ise sırasıyla 142, 142 ve 75'tir. Çalışmada bu özellikler kullanılarak oluşturulan vektörden $\mathbf{OV_4}$ olarak bahsedilecektir.

3 Sınıflandırma Yöntemleri

Bu çalışmada, dokümanları yazarına, türüne ve yazarının cinsiyetine göre sınıflandırmak için "Weka" [14] sınıflandırma aracı içerisinde yer alan Naive Bayes, Destek Vektör Makinesi, K-En Yakın Komşuluk, Rastgele Orman yöntemleri ile birlikte bizim geliştirdiğimiz **Ng-ind** yöntemi kullanılmıştır.

Naïve Bayes (NB)

Klasik Naïve Bayes algoritması, genelde kelimelerin ve sınıfların birleşik olasılıkları ile bir dokümanın sınıfının belirlenmesinde kullanılır. Bizim çalışmamızda ise özellikler kelimelerin frekansları değildir ve sürekli dağılımlara sahip olduklarından klasik Naïve Bayes yerine George'un [10] çalışmasında önerilen Naïve Bayes versiyonu kullanılmıştır.

Destek Vektör Makinesi (DVM)

DVM, günümüzde performansı sayesinde oldukça popüler olmuş bir metottur. Sınıfları birbirinden ayıran marjini en büyük, doğrusal bir ayırt edici fonksiyon bulunmasını amaçlar. Doğrusal olarak ayrılamayan örnekler için, örnekler doğrusal olarak ayrılabildikleri daha yüksek boyutlu başka bir uzaya taşınır ve sınıflandırma o uzayda yapılır.

K-En Yakın Komşuluk (K-EYK)

Sınıflandırılmak istenen örneğin sınıfı belirlenirken eğitim kümesinde o örneğe en yakın olan k adet örnek seçilir. Seçilen örnekler en çok hangi sınıfa ait ise ilgili test örneği de o sınıfa dahil edilir.

Rastgele Orman (RO)

Breiman [11] tek bir karar ağacı üretmek yerine, her biri farklı eğitim kümeleriyle eğitilen çok sayıda, çok değişkenli ağacın kararlarının birleştirilmesini önermiştir. Farklı eğitim kümeleri önyükleme (bootstrap) ve rasgele özellik seçimi ile orijinal eğitim setinden oluşturulur. Çok değişkenli karar ağaçları CART [10] algoritmasıyla elde edilir. Önce her karar ağacı kendi kararını verir ve karar ormanı içerisinde maksimum oyu olan sınıf, son karar olarak kabul edilir ve gelen test verisi o sınıfa dahil edilir.

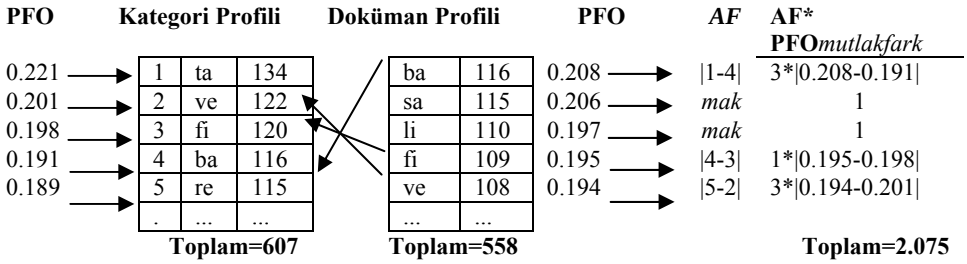
Ng-ind Yöntemi

Ng-ind adını verdiğimiz bu yöntem, Cavnar ve Trenkle tarafından yazılmış "N-gram Tabanlı Doküman Sınıflandırma" makalesindeki "Measure profile distance" adını verdikleri profiller arasındaki benzerlik ölçümünden esinlenilerek hazırlanmıştır [2].

Profiller arasındaki benzerlik ölçümü işleminde, her bir kategorinin ve sınıflanacak olan her bir dokümanın n-gram profili hesaplanır. Profiller, bir dokümanda bulunan n-gram'ların en yüksek frekanstan en düşük frekansa sıralanmış halinden oluşmakta ve ayırt edici olduğu düşünüldüğünden yüksek frekanslı n-gram'lar ($\geq$100) kullanılmıştır. Ölçüm aşamasında kategori ve sınıflandırılacak dokümanın profili olmak üzere iki n-gram profil alınır ve "adres-fark" (AF) olarak isimlendirilen basit bir dizi-sıra istatistik hesabı yapılır. Bu ölçü, doküman profil içerisindeki bir n-gram'ın yerinin, kategori profil içerisindeki yerinden ne kadar uzakta olduğu bilgisini verir. Doküman profilindeki her bir n-gram için, kategori profilindeki onun benzeri bulunur ve aralarındaki uzaklığın mutlak değerce farkı hesaplanır.

Örneğin Şekil 1'de, "ve" n-gram'ının dokümandaki sırası 5, kategorideki sırası 2 olup, "ve" n-gram'ı için "adres-fark" (5 fark 2) 3 olarak hesaplanır. Eğer doküman profilindeki herhangi bir n-gram, kategori profilinde bulunmuyorsa maksimum uzaklık verilir. Daha sonra hem doküman hem de kategori profilinde yer alan n-gram'ların "profil frekans oranı" (PFO) hesaplanır. Profildeki her bir n-gram'ın frekans değerinin, belirli profil uzunluğu içerisinde bulunan tüm n-gram'ların frekans değerlerinin toplamına oranı n-gram'ın "profil frekans oranını" olarak bulunur. Profil uzunluğunun 5 olduğu farz edilen bu örnekte, "ta" n-gram'ının frekans değeri 134 olup, ilk 5 n-gram'ın frekans değerlerinin toplamı 607'ye oranı, ilgili n-gram'ın PFO'nını 0.221 olarak verir.

Ng-ind yönteminin son adımında doküman profilindeki her bir n-gram için, kategori profilindeki benzeri bulunur ve aralarındaki mutlak adres-fark (AF) değeri, doküman ve kategori profilinde hesaplanan profil frekans oranlarının mutlak değerce farkları alınarak çarpılır. AF değeri *mak* olarak belirlenen n-gram'larda $AF*PFO_{mutlakfark}$ değeri, maksimum değer olan ***1*** alınarak işleme dahil edilir.



Şekil 1 Ng-ind yöntemiyle benzerlik değerinin hesaplanması

Bir n-gram'ın profil frekans değerlerinin mutlak farkı ne kadar küçülürse profillerdeki benzerlik o kadar artmaktadır. Dokümanın profilinde bulunan tüm n-gram'lar için bu hesaplanan değerlerin toplamı alınarak, her bir kategori profili için bir benzerlik ölçümü elde edilir. Bu benzerlik ölçüm değerlerinden en küçük değere sahip olan kategori, test edilen dokümanın atanacağı sınıf olarak kabul edilir. Şekil 1'deki örnekte profil uzunluğu 5 olarak alınmış ve test dokümanının karşılaştırıldığı kategori profiline yakınlığı 2.075 olarak bulunmuştur.

Oylama (Vote)

Oylama, doğru sınıflandırma başarısını arttırmak için, başarılı olduğu düşünülen sınıflandırıcıların seçilip birlikte kullanılması işlemidir. Sınıflandırıcıları birlikte kullanma işleminin performansını gözlemlemek amacıyla, Naive Bayes, Rastgele Orman, Destek Vektör Makinesi, K-En Yakın Komşuluk gibi sınıflandırma yöntemleri farklı kombinasyonlar ile birlikte kullanılmıştır. Birlikte kullanım işlemi, tek bir özellik vektörü kullanıldığında, başarısı en yüksek olan sınıflandırıcıların bir araya getirilerek oylama sonucunda dokümanın hangi sınıftan olabileceğine karar veren bir işlemdir.

4 Deneysel Sonuçlar

Bu çalışmanın amacı, dokümanların yazarlarını, türlerini ve yazarların cinsiyetini n-gram modelini kullanarak oluşturduğumuz üç farklı özellik vektörü yardımıyla geliştirdiğimiz Ng-ind yönteminin başarısını test etmek ve alınan sonuçları, Destek Vektör Makinesi, Naive Bayes, Rastgele Orman ve K-En Yakın Komşuluk makine öğrenmesi teknikleri ile karşılaştırmaktır. Ayrıca sınıflandırmadaki başarı oranını arttırmak için başarılı olan sınıflandırıcılar birlikte kullanılmıştır. Birlikte kullanma işlemi, tek bir özellik vektörü kullanıldığında, başarısı en yüksek olan sınıflandırıcıları bir araya getirerek oylama sonucunda dokümanın hangi sınıftan olabileceğine karar veren bir işlemdir. Özellik vektörlerinin yazar belirlemedeki başarılarını test ederken 10'lu çapraz geçerlilik (10-fold cross validation) kullanılmıştır. Weka paketi içerisinde yer alan sınıflandırma yöntemlerini kullanırken ön tanımlı (default) parametreler tercih edilmiştir [14].

Çizelge 1'de Ng-ind yöntemi kullanılarak yazarın cinsiyetini belirmek amacıyla hazırlanmış Veri Seti-I, yazının yazarını tahmin etmek için hazırlanmış Veri Seti-II ve yazının türünü bulmak amacıyla hazırlanmış Veri Seti-III'ün farklı özellik sayıları ile alınmış başarı oranları gösterilmektedir. Özellik sayıları, her veri seti ve n-gram modelinde farklılıklar göstermektedir. Ng-ind yönteminin başarısını hem bu özellikler ile hem de farklı özellik sayılarında görebilmek için özellik sayıları sırasıyla 100, 200, 300, 400 ve 500 alınarak test edilmiştir.

Veri Seti-I kullanılarak yazarın cinsiyetini tahmin ederken, Ng-ind yöntemi OV_2

vektörünün özellik sayısına eşit iken (257) %78,8'lik başarı verirken, aynı yöntem kullanılan özellik vektörü eleman sayısı 400 olduğunda doğru sınıflandırma başarısını %82,5'a yükseltmiştir. 324 elemanlı OV_3 özellik vektörü kullanıldığında %80 olan başarı, 400 elemanlı özellik vektöründe %82,5'a yükselmektedir. OV_4 özellik vektöründe de (142) %85 olarak alınan doğru sınıflandırma başarısı, özellik vektörü eleman sayısı 100 olduğunda %91,3 yükselerek bu veri seti için en başarılı sınıflandırıcı olmuştur.

Çizelge 1 Üç farklı model ve veri seti'nin Ng-ind yöntemi ile farklı özellik sayılarındaki başarı oranları

Ng-ind		**Özellik Sayısı**							
Veri Seti-I		**100**	**142**	**200**	**257**	**300**	**324**	**400**	**500**
Başarı Oranı %	**2-gram**	71,3	-	78,6	**78,8**	80	-	**82,5**	78,8
	3-gram	75	-	82,5	-	78,8	**80**	**82,5**	78,8
	4-gram	**91,3**	**85**	88,8	-	88,8	-	88,8	85
Veri Seti-II		**100**	**142**	**200**	**257**	**300**	**324**	**400**	**500**
Başarı Oranı %	**2-gram**	70	-	71,3	**70**	72,5	-	**75**	71,3
	3-gram	75	-	73,8	-	73,8	**70**	**75**	73,8
	4-gram	75	**70**	73,8	-	75	-	**77,5**	73,8
Veri Seti-III		**75**	**100**	**200**	**208**	**217**	**300**	**400**	**500**
Başarı Oranı %	**2-gram**	-	75	75	-	**81,3**	79,2	**81,3**	75
	3-gram	-	70,8	85,4	**83,3**	-	87,5	**93,8**	93,8
	4-gram	**83,3**	85,4	85,4	-	-	87,5	**93,8**	87,5

Veri Seti-II'yi kullanarak, yazının yazarını tahmin ederken, Ng-ind yöntemi 257 elemanlı OV_2, 324 elemanlı OV_3 ve 142 elemanlı OV_4 vektörleri ile Ng-ind yöntemi kullanılarak sınıflandırılma yapıldığında doğru sınıflandırma başarısı hepsinde %70 olmuştur. Bu veri setinde her üç özellik vektörü sayısı 400 olarak alındığında en yüksek başarı oranları elde edilmiştir. Yazar tanımada %77,5 ile en başarılı sonuç 400 elemanlı OV_4 vektörünün kullanılması ile alınmıştır.

Veri Seti-III kullanılarak yazının türünü tahmin ederken, OV_2, OV_3 ve OV_4 özellik vektörlerinin özellik sayıları 217, 208 ve 75 iken Ng-ind yöntemi ile alınan doğru sınıflandırma başarıları %81,3 - %83,3 ve %83,3 olmuştur. Aynı veri seti için özellik sayısı 400 olarak alındığında başarı oranları sırasıyla %81,3 - %93,8 ve %93,8 yükselmiştir. Bu başarı diğer sınıflandırıcılara göre oldukça yüksektir.

Çizelge 2'de hangi yöntemin hangi özellik sayısı ile ne kadar başarılı olduğu gösterilmektedir. Ng-ind'in en başarılı sonucu verdiği özellik sayısı genelde 400'dür. Sadece yazarın cinsiyetini belirlerken 4-gram modelinde özellik sayısı 100 olarak alınmış ve diğer yöntemlere göre en başarılı sonucu %91,3 ile vermiştir. Ng-ind yöntemi yazar tanımada başarılı olamamıştır, sadece RO ve K-EYK yöntemlerine göre daha iyi performans göstermiştir. Bu veri setinde Ng-ind

yönteminin başarısının düşük olması sınıf sayısının fazla olmasından kaynaklanmaktadır. Yazının türünü belirlemede 3- ve 4-gram modelde Ng-ind yöntemi diğer sınıflandırıcılara göre %93,8 ile çok daha yüksek bir performans sergilemiştir.

Çizelge 2 Ng-ind yönteminin NB, RO, K-EYK ve DVM ile karşılaştırılması

Veri Seti-I		**NB**	**RO**	**K-EYK**	**DVM**	**Ng-ind (400-400-100 öz.)**
Başarı Oranı %	**OV_2(257 öz.)**	67,8	85,6	83,4	91,1	82,5
	OV_3(324 öz.)	73	83	83,3	89,8	82,5
	OV_4(142 öz.)	74,9	84,1	82	85,6	**91,3**
Veri Seti-II		**NB**	**RO**	**K-EYK**	**DVM**	**Ng-ind (400 öz.)**
Başarı Oranı %	**OV_2(257 öz.)**	81,9	67	68	89,5	75
	OV_3(324 öz.)	82	61,5	62,1	88,3	75
	OV_4(142 öz.)	80,8	63,3	57,1	80	77,5
Veri Seti-III		**NB**	**RO**	**K-EYK**	**DVM**	**Ng-ind (400 öz.)**
Başarı Oranı %	**OV_2(217 öz.)**	73,5	74,6	81,5	91,9	81,3
	OV_3(208 öz.)	80,8	72,5	72,3	92,1	**93,8**
	OV_4(75 öz.)	76,7	79,6	60,6	82,5	**93,8**

Sınıflandırma başarısını arttırabilmek için bazı sınıflandırıcıların oylama yöntemi ile birlikte kullanılması düşünülür. Sınıflandırıcıların birlikte kullanılması işleminin performansını gözlemlemek amacıyla sınıflandırıcılar farklı kombinasyonlarda bir araya getirilir. Birlikte kullanma işlemi, daha önceki denemelerde başarısı en yüksek olan sınıflandırıcıların bir araya getirilerek oylama sonucunda dokümanın hangi sınıftan olabileceğine karar verme işlemdir.

Veri Seti-I'de DVM, RO ve K-EYK yöntemleri birlikte kullanılmış ve %92,6'lık bir başarı alınmıştır. Bu başarı Ng-ind yönteminin 4-gram ile aldığı %91,3'lük başarıdan yüksektir. Veri Seti-II'de DVM, NB ve K-EYK yöntemleri birlikte kullanıldığında en yüksek başarı olarak %87,4 elde edilmiştir. Veri Seti-III'de ise birleştirilmiş sınıflandırıcıların hiçbiri Ng-ind yönteminin başarısını geçememiştir. En yüksek başarı DVM, NB ve K-EYK ile %88,2 alınırken, Ng-ind'in başarısı %93,8'dir. Çizelge 3'de birleştirilmiş sınıflandırıcıların doğru sınıflandırma başarıları gösterilmektedir.

Çizelge 3 Ng-ind yönteminin birleştirilmiş sınıflandırıcılar ile karşılaştırılması

Veri Seti-I				**Ng-ind (400-400-100 öz.)**
Başarı Oranı %	**OV_2(257 öz.)**	DVM-RO-K EYK	***92,6***	82,5
	OV_3(324 öz.)	DVM-RO-K EYK	91,8	82,5
	OV_4(142 öz.)	DVM-RO-K EYK	88,4	91,3
Veri Seti-II				**Ng-ind (400 öz.)**
Başarı Oranı %	**OV_2(257 öz.)**	DVM-NB-K EYK	***87,4***	75
	OV_3(324 öz.)	DVM-NB-K EYK	86,3	75
	OV_4(142 öz.)	RO-NB-K EYK	81,1	77,5
Veri Seti-III				**Ng-ind (400 öz.)**
Başarı Oranı %	**OV_2(217 öz.)**	DVM-RO	83,5	81,3
	OV_3(208 öz.)	DVM-NB-K EYK	***88,2***	**93,8**
	OV_4(75 öz.)	DVM-RO	85,4	**93,8**

Çizelge 4 Cavnar'ın yöntemi ve Ng-ind yönteminin karşılaştırılması

		Cavnar'ın Yöntemi			Ng-ind		
		2-gram	3-gram	4-gram	2-gram	3-gram	4-gram
Özellik sayısı	100	70,8	68,8	85,42	75	70,83	85,42
	200	75	85,42	85,42	75	85,42	85,42
	300	75	85,42	87,50	79,17	87,50	87,50
	400	81,25	91,67	87,50	81,25	93,75	93,75
	500	75	91,67	87,50	75	93,75	87,50
Ort. Başarı %		75,41	85	86,67	77	86	88

Cavnar'ın [2] yöntemi ile Ng-ind'in başarılarını karşılaştırabilmek için Cavnar'ın yöntemini de ayrıca kodladık. Üç farklı n-gram modeli ve 5 farklı özellik vektörü kullanılarak Veri Seti-III (tür belirleme) için alınmış olan sonuçlar karşılaştırmalı olarak Çizelge 4'de verilmiştir. Üç farklı n-gram modeli için ortalama sonuçlara bakıldığında Ng-ind yönteminin daha başarılı olduğu görülmüştür. Aynı zamanda Ng-ind yönteminde en yüksek başarı %93,75 iken Cavnar'ın yönteminde %91,67'dir. Sonuçlara göre Cavnar'ın yöntemi üzerinden geliştirdiğimiz Ng-ind yöntemi daha başarılı sonuçlar vermektedir.

5 Sonuç

Bu çalışmada, yazarı bilinmeyen bir dokümanın önceden belirlenmiş 20 farklı yazar içerisinden hangisine ait olabileceği, türü bilinmeyen bir dokümanın 6 farklı kategoriden hangisine dahil olabileceği ve dokümanın yazarının cinsiyetini belirlemek üzere Ng-ind adı verilen bir yöntem geliştirilmiştir. Bu yöntemin başarısını test etmek için Türkçe'nin 2-, 3- ve 4-gram'ları kullanılmış ve yöntemin başarısı Naive Bayes, Destek Vektör Makinesi, Rastgele Orman ve K-En Yakın Komşuluk ile karşılaştırılması önce ayrı ayrı daha sonra da birleştirilmiş sınıflandırıcılar ile yapılmıştır. Ng-ind yöntemi çalışmada adı geçen diğer sınıflandırıcılara göre yazar tanımada başarılı olamamıştır. Bunun sebebi ise, 20 farklı sınıf içerisinden tanıma yapıyor olmasıdır. Cinsiyet ve tür belirlemede yazar tanımaya göre çok daha iyi başarı alınmıştır. En iyi başarı 6 sınıf ile dokümanın türünü bulmada %93,8 ile elde edilmiştir.

Kaynakça

1. Doğan, S., 2006, "*Türkçe Dokümanlar için N-gram Tabanlı Sınıflandırma: Yazar, Tür ve Cinsiyet*", Yıldız Teknik Üniv., Master Tezi
2. Cavnar, W. B. ve Trenkle, J. M., 1994, "*N-gram-based text categorization*", Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. Information Systems Project Management, Jolyon E. Hallows, AMACOM Pres
3. Peng F., Keselj V., Cerconey N., Thomasy C., 2003, "*N-Gram-Based Author Profiles For Authorship Attribution*", Faculty of Computing Science, Dalhousie University, Canada
4. Stamatatos E., Fakotakis N., Kokkinakis G., 2000, "*Automatic Text Categorization in Terms of Genre and Author",* Computational Linguistics, pp.471-495
5. Peng F., Schuurmans D., 2003, "*Combining Naive Bayes and N-gram Language Models for Test Classification*", School of Computer Science, University of Waterloo.
6. Amasyalı M.F., Diri B., 2006, "*Automatic Written Turkish Text Categorization in Terms of Author, Genre and Gender*", 11th International Conference on Applications of Natural Language to Information Systems, Austria
7. Peng F., Wang S., Schuurmans D., 2003, "*Language and Task Independent Text*

Categorization with Simple Language Models", School of Computer Science, University of Waterloo
8. Nowson S., Oberlander J., 2006, "*Openness and gender in personal weblogs*", School of Informatics, University of Edinburgh, 2 Buccleuch Place, Edinburg, EH89LW
9. Dupont P., 2006, "Noisy *Sequence Classification with Smoothed Markov Chains*", Department of Computing Science and Engineering (INGI), Université catholique de Louvain Place Sainte Barbe, 2 B-1348 Louvain-la-Neuve – Belgium
10. George H., 1995, "*Estimating Continuous Distributions in Bayesian Classifiers*", Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, pp. 338-345. Morgan Kaufmann, San Mateo
11. Breiman L., 1999, "*Random forests–random features*", Technical Report 567, Department of Statistics, University of California, Berkeley
12. Peng F., Schuurmans D., 2003, "*Combining Naive Bayes and N-gram Language Models for Text Classification*", School of Computer Science, University of Waterloo
13. Doyle J., Keselj V., 2005, "*Automatic Categorization of Author Gender via N-Gram Analysis*", In The 6th Symposium on Natural Language Processing, SNLP'2005, Chiang Rai, Thailand, December
14. http://sourceforge.net/projects/weka/