

Makine Öğrenme Yöntemleriyle N-Gram Tabanlı Dil Tanıma

N-Gram Based Language Identification with Machine Learning Methods

Şengül Bayrak¹, Hidayet Takçı², Mübariz Eminli¹

¹Bilgisayar Mühendisliği Bölümü, Haliç Üniversitesi

sengulbayrak@halic.edu.tr, mubarizeminli@halic.edu.tr

²Bilgisayar Mühendisliği Bölümü, Cumhuriyet Üniversitesi

htakci@cumhuriyet.edu.tr

Özet

İletişim olanaklarının gelişmesi aynı zamanda birçok dilden dokümanı kullanılabilir hale getirmiştir. Bu dokümanların anlaşılır olabilmesi ve üzerinde doğal dil işleme prosedürlerinin çalışabilmesi için dil tanıma anahtar konumdadır. Artan metinsel dokümanlar ve uluslar arası iletişim ihtiyacı dil tanıma alanında yeni çalışmaları zorunlu hale getirmiştir. Bugüne kadar dil tanıma probleminin çözümünde harfler, kelimeler ve n-gram dizileri makine öğrenimi teknikleriyle kullanılmıştır. Bu çalışmada, n-gram özellik seti kullanılarak; Destek Vektör Makinaları, Yapay Sinir Ağları, Centroid Sınıflayıcı, K-Ortalama ve Bulanık C Ortalamalar algoritması gibi sınıflayıcılar performans bakımından karşılaştırılmış ve dil tanıma problemi için en uygun sınıflayıcılar tespit edilmiştir. Deneyler sırasında çok dilli ECI verisi kullanılmıştır.

Abstract

Documents are available in many languages has made by the development of communication facilities at same time. Language identification is the key for these documents to be understood and to work on language identification and natural language processing procedures. Increased needs for the field of language identification has made the new studies on language identification necessary. Until now, to solve the problem of language identification different techniques have been used as letters, words and n-gram sequences, machine learning algorithms. In this study, n-gram feature set has been used and Support Vector Machines, Artificial Neural Networks, Centroid Classifier, K-Means and Fuzzy C Means algorithms are compared in terms of performance and the most appropriate classifiers for the problem of language identification. ECI multi-lingual data is used for the empirical results.

1. Giriş

Her geçen gün daha fazla globalleşen dünyanın çözmesi gereken en önemli problemlerden biri de dil problemidir. Kullanılabilir dokümanların miktarı artmasına rağmen dili bilinmeyen dokümanlardan faydalanma imkânı maalesef yoktur. Eldeki bilgi kaynaklarının daha faydalı hale gelebilmesi için dil tanıma ve dil anlama önemli işlerdendir. Dil tanıma dil anlamının ilk adımı olması dolayısıyla bu çalışmalarda anahtar görevi görmektedir. Dil tanıma, dili bilinmeyen bir dokümanın içerisinde yer alan özellikler ve algoritmalar yardımıyla dilinin tayin edilmesi işlemidir. Dil tanıma; konuşma ve yazılı metnin tanınması olmak üzere ikiye ayrılır. Konuşma dili tanıma genellikle sinyal işleme teknikleri ile fonem adını verdiğimiz yapılar kullanılarak yerine getirilir. Yazılı dilin tanınması ise harflerle, harf dizileriyle, kelimelerle ve n-gram adını verdiğimiz kelime veya harf dizileri ile yerine getirilir. Çalışmamız yazı dili tanıma üzerinedir. Yazı dili tanıma en sık kullanılan yöntem istatistiksel dil tanımadır. Dilin anlamından çok dile ait özelliklerin dağılımı ile ilgilenir. Dil tanıma kimi zaman da dili bilinen dokümanlardan makine öğrenimi yöntemleriyle modellerin çıkarılması ve bu modeller yardımıyla dili bilinmeyen dokümanların dilinin bulunması şeklinde yerine getirilir. Bu tür dil tanıma makine öğrenimli dil tanıma olarak bilinir. Dil tanıma probleminde en sık kullanılan makine öğrenmesi tekniği sınıflandırma olduğundan dil tanıma konusunda metin sınıflandırma yöntemleri önemli bir yer tutar.

Bu çalışmada, n-gram özellik seti kullanılarak; Destek Vektör Makinaları, Yapay Sinir Ağları, Centroid Sınıflayıcı, K-Ortalama ve Bulanık C Ortalamalar algoritması gibi sınıflayıcılar performans bakımından karşılaştırılmış ve dil tanıma problemi için en uygun sınıflayıcılar tespit edilmiştir.

2. İlişkili Çalışmalar

Metinsel dil tanıma en basit yaklaşımlardan birisi genel kelimelerin kullanılmasıdır [1]. Genel kelimeler; belirleyici, bağlaç, edat gibi bir dilde sık rastlanan ve dilin tespit edilmesinde önemli bilgi sağlayan kelimelerdir. Bir diğeri n-gram dizilerinin kullanılmasıdır. Konuyla ilgili en bilinen çalışma Cavnar ve Trenkle [2] tarafından yapılan çalışmadır.

Bu çalışma metin sınıflandırma yöntemleriyle dil tanıma örneği olması açısından da değerlidir. Cavnar ve Trenkle'in çalışmasında farklı n değerlerine göre birçok deney yapılmış ve dil tanımda en çok 3-gramların başarılı olduğu görülmüştür. Ayrıca karakter kodlamaları ile birlikte n-gramlar Suzuki [3] ve arkadaşları tarafından kullanılmıştır. N-Gram tabanlı bir başka çalışmada Adams ve Resnik [4], dinamik tüm belgeler için web sayfalarına dil etiketleri ekleyen bir sistem önermiştir. Bu amaçla, 220 KB eğitim verisinden 5-gramlar çıkarılmış ve her bir 100-500 byte test verisi için doğruluk değeri %98.68 elde edilmiştir. Bir başka yaklaşım harflerin kullanılmasıdır. Beesley [5] çalışmasında 2, 3, 4 ve daha fazla harf kullanarak İngilizce, İspanyolca, Fransızca ve Portekizce dilleri üzerinde dil tanıma yapmıştır. Dunning [6] İspanyolca ve İngilizce üzerinde yaptığı çalışmada makine öğrenme yaklaşımı kullanılmıştır. Bayesian sınıflayıcı kullanan yazar 20 byte gibi kısa metinlerde bile %92 gibi yüksek sonuçlar elde edilmiştir. Combrink ve Botha [7] tarafından yapılan çalışma metin sınıflandırma yöntemleri ile dil tanıma çalışmasının yapıldığı çalışmalardan bir diğeridir. Bu çalışmada 12 dil kullanılmış ve dilleri sınıflandırmak için Histogram Metodu kullanılmıştır. Grefenstette [1], çalışmasında trigram ve kısa terimler(short terms) metodunu kullanmıştır ve bu iki yöntem için en büyük problem yüksek boyutlu özellik setidir. Trigram yöntemiyle yapılan testlerde 100 byte uzunluğundaki test verisi için %98.96 oranında başarı elde edilmiştir. Kısa terimler metodu ile yapılan testlerde ise 100 byte uzunluğundaki test metinleri için sınıflandırma doğruluğu %98.68'dir. Prager [8], dil tanımda metin sınıflandırma yöntemi kullanan bir başka yazardır. O vektör uzayı modeli kullanan Linguini sistemini önermiştir. Xafopoulous ve arkadaşları [9] dil tanımda, karakter dizileri özellik setinin HMM (Hidden Markov Model) ile kullanımını önermiştir. Önerilen bu sistem; İngilizce, Almanca, Fransızca, İspanyolca ve İtalyancaya ait web dokümanları için otomatik olarak dil tanıma yapmaktadır. 140 byte test datası için %99 doğruluk oranına erişilmiştir. Düşük bir özellik seti ile dil tanımayı öneren Takçı ve Soğukpınar [10] İngilizce, Fransızca, Almanca ve Türkçe dilleri üzerinde yaptığı çalışmada 22 harften oluşan bir özellik seti ve centroid-tabanlı bir algoritma kullanmışlardır. 500 KB eğitim ve 100 byte test verisi için %98 başarı oranına ulaşılmıştır.

3. Yöntem Ve Uygulama

Çalışmamızın amacı, dil tanımda bugüne kadar dil tanıma problemi için n-gram özellik seti kullanılarak makine öğrenme yaklaşımına dayanan birçok sınıflandırma yöntemlerini karşılaştırılması ve böylece en uygun yöntemlerin açığa çıkarılmasıdır.

3.1. Önışlem Aşaması

Dil tanıma metinler üzerinde yapıldığı için metinsel verilerin bir dönüşümden geçirilmesi ihtiyacı vardır. Bu çalışmada n-gram özellik seti kullanılacağı için dönüşüm aşamasında metinler n-gram sıklıklarına dönüştürülecektir. Tüm sınıflandırma yöntemleri için vektör uzayı modeli benimsenmiştir. Vektör uzayı modeline göre dil profilleri kullanarak, doküman profilleri ve doküman örnekleri birer vektör olarak sunulacaktır.

3.1.1. N-Gram Önışlem

Metinsel veri boşluk, noktalama, sayısal karakterlerden arındırıldıktan sonra n-gramlar oluşturulur. N-gram; ardı sıra gelen n adet karakterden meydana gelir. Metinler üzerinde genişliği n olan kayan pencereler yardımıyla o metinde yer alan bütün n-gramlar elde edilir. Elde edilen bu n-gramlardan tekrar edenler sayılarak her bir n-gramdan kaç adet yer aldığı bulunur ve n-gram profili adı verilen bir vektörde tutulur.

3.2.Yöntemlerin Uygulanması

Eğitim ve test metinleri için n-gram profilleri uygun şekilde elde edildikten sonra sıradaki işlem bu verilere makine öğrenmede sınıflandırma algoritmalarının uygulanmasıdır. Karşılaştırması yapılacak makine öğrenimi algoritmalarının sırayla açıklanması konunun anlaşılması açısından elzemdir.

3.2.1. Destek Vektör Makineleri (DVM)

Destek Vektör Makineleri, Vapnik [11] tarafından geliştirilmiştir. Algoritma; sinyal işleme, yapay öğrenme ve veri madenciliği alanındaki birçok problemde iyi sonuçlar vermiştir [11][6]. Temelde DVM iki sınıflı problemlerle ilgilenir. Çok-sınıflı verilerinin DVM ile sınıflandırılması için Knerr ve arkadaşları bire-karşı-bir (one-against-one) yöntemini önermişlerdir [12]. Bu yöntemde n adet sınıf için $n(n-1)/2$ sınıflandırıcı oluşturulur. Her biri sadece iki sınıftan oluşan veriler ile eğitilir. Bu yöntem sayesinde çoklu sınıf problemi ikili sınıf problemine çevrilmiş olur. Çok-sınıflı verilerin DVM ile sınıflandırılması için bir diğer yöntem ise bire-karşı-hepsi (one-against-all) adı verilen yöntemdir [13]. Bu yöntemde n adet sınıf için n adet DVM kurulur ve i.nci DVM, i sınıfındaki verileri kendi sınıf verileri olarak kullanırken, diğer sınıflardan gelen verilerin hepsini sanki 2.sınıfa ait veriymiş gibi kabul eder. Yani kendi verilerine +1 etiketi verirken, diğer sınıflara ait olan tüm verilere -1 etiketini verir ve eğitimi bu şekilde n adet DVM için yapar. Bu çalışmada DVM algoritmasının, ikiden fazla sınıflı uygulamalarda kullanılan C-SVC algoritması üzerinde performans analizi yapılmıştır ve LibSVM kütüphanesi kullanılmıştır.

3.2.2. Centroid Tabanlı Sınıflayıcı

Centroid tabanlı sınıflayıcı vektör uzayı modeli tabanlı, performansı yüksek bir metin sınıflandırma yöntemidir. Doküman sunumunda vektör uzayı modelinin sağladığı imkânlarla sahiptir. Model her bir dokümanı d terim uzayında bir vektör olarak ele alır. Vektörün her bir boyutu dokümanda geçen bir terimin ağırlıklandırılmış frekansını tutar. Centroid tabanlı sınıflandırmada, sınıflar centroid adı verilen vektörlerle sunulur. Centroid, sınıf elemanlarını sunan ortalama bir değerdir ve bu orta değer bütün sınıfı temsil ettiği kabul edilir. Eğitim seti k farklı kategori içeriyorsa bu eğitim verilerinden k adet centroid vektörü elde edilir [10].

$$\bar{C} = \frac{1}{|S|} \sum_{d \in S} \bar{d} \quad (1)$$

Dili bilinmeyen tanınması için doküman ile centroid vektörü arasındaki benzerlik ise aşağıdaki gibi bulunur.

$$\text{sim}(\vec{x}, \vec{C}_{j,j=k}) = \frac{\vec{x} \cdot \vec{C}_j}{\|\vec{x}\|_2 * \|\vec{C}_j\|_2} \quad (2)$$

Benzerlik hesabı ardından doküman en yüksek benzerliği veren centroide göre bir dile atanır. Benzerlik için cosine yöntemi kullanılır. Aşağıdaki denklemde benzerliğin bulunması ve maksimum benzerliğin seçimi sunulmaktadır.

$$\arg \max (\cos_j(d1,d2)) = d1*d2 / \|\vec{d1}\| \|\vec{d2}\| \quad (3)$$

3.2.3.Yapay Sinir Ağları (YSA)

Yapay Sinir Ağları, genel anlamda beynin çalışma prensibini modellemek üzere tasarlanan bir sistem olarak tanımlanabilir. Öğrenme süreci, arzu edilen amaca ulaşmak için YSA ağırlıklarının yenilenmesini sağlayan öğrenme algoritmalarını içerir. YSA'ların hesaplama ve bilgi işlem gücünü paralel yapısından, öğrenebilme ve genelleme yeteneğinden aldığı söylenebilir. Genelleme, eğitim ya da öğrenme sürecinde karşılaşılmayan girişler için de YSA'nın uygun tepkileri üretmesi olarak tanımlanır. Bu üstün özellikleri, YSA'nın karmaşık problemleri çözebilme yeteneğini gösterir [14].

3.2.4.K-Ortalamalar Algoritması

K-ortalamalar (K-means) kümeleme problemini çözen en basit denetimsiz öğrenme yöntemleri arasında yer alır. Algoritmanın genel mantığı n adet veri nesnesinden oluşan bir veri kümesini, giriş parametresi olarak verilen k adet kümeye bölümlenektir. Amaç, gerçekleştirilen bölümlenme işlemi sonunda elde edilen kümelerin, küme içi benzerliklerinin maksimum ve kümeler arası benzerliklerinin minimum olmasını sağlamaktır. Yöntemin performansını k küme sayısı, başlangıç olarak seçilen küme merkezlerinin değerleri ve benzerlik ölçümü kriterleri etkilemektedir [14].

Yöntemin adımları aşağıda gösterilmiştir.

Adım 1: Kümeleme işlemine başlamadan önce k adet küme için başlangıç küme merkezleri belirlenir, $C = \{c_1, c_2, \dots, c_{15}\}$. Bunun için nesneler arasından k adet rastgele nokta seçilebilir ya da tüm nesnelerin ortalaması ile de belirlenebilir.

Adım 2: Eğitim kümesindeki her veri, $X = \{x_1, x_2, \dots, x_{15}\}$ seçilen başlangıç küme merkezlerinden kendine en yakın (veya benzer) kümeye dahil edilir. Kümeye dahil etme işleminde benzerlik kullanılması durumunda aşağıdaki formül kullanılır.

$$\cos(x_i, \text{merkez}(c_j)) = \frac{x_i}{\|x_i\|} * \frac{\text{merkez}(c_j)}{\|\text{merkez}(c_j)\|} \quad (4)$$

(i = {1, 2, 3...n}, j = {1, 2, 3...k})

Adım 3: Oluşan kümelerin merkez noktaları o kümedeki tüm nesnelerin ortalama değerleri ile değiştirilir.

$$\text{merkez}(c_j) = \frac{\sum_{i=1}^{n_j} (x_i)}{n_j} \quad (5)$$

($x_i \in c_j$) ve $n_j = c_j$ kümesindeki veri sayısı

Adım 4: Merkez noktalar değişmeye kadar 3. ve 4. adımları tekrarlanır.

3.2.5.Bulanık C Ortalamalar Algoritması

Bulanık mantık, belirsizliklerin anlatımı ve belirsizliklerle çalışılabilmesi için kurulmuş katı bir matematik düzen olarak tanımlanabilir. Bilindiği gibi istatistikte ve olasılık kuramında, belirsizliklerle değil kesinliklerle çalışılır ama insanın yaşadığı ortam daha çok belirsizliklerle doludur. Bu yüzden insanoğlunun sonuç çıkarabilme yeteneğini anlayabilmek için belirsizliklerle çalışması gereklidir [15]. Kimi zaman nesnelerin bir kümeye kesin olarak atanmadığı durumlar olabilir, bu tip durumlar için Bulanık C Ortalamalar(BCO) uygun bir yöntem olacaktır.

BCO Algoritması için Gerekli Adımlar:

Adım1: Küme sayısı ve bulanıklık indeksi m (bulanıklaştırma(fuzzication) değeri) değeri seçilir. Üyelik dereceleri matrisini gösteren U matrisi için başlangıç değerleri belirlenir.

Adım 2: Centroid (V) küme prototiplerinin rastgele üretildiği varsayılırsa bu değerler kullanılarak üyelik dereceleri matrisini aşağıdaki formüle göre hesaplanır.

$$u_{ik} = \sum_{l=1}^c \left(\frac{\|v^i, x_k\|}{\|v^l, x_k\|} \right)^{-2/m-1} \quad (5)$$

Adım 3: V küme prototip değerlerini 6 numaralı formüle göre güncellenir.

$$v_i = \frac{\sum_{j=1}^n u_{ij}^m x_j}{\sum_{j=1}^n u_{ij}^m} \quad (6)$$

Adım 4: Eğer elde edilen centroid değeri önceki elde edilen centroid değerinden farkı epsilon(0.1) dan küçük değilse 2. adıma git. Bulanık C Ortalamalar(BCO) algoritması uygulandıktan sonra hangi nesnenin hangi kümeye dahil edileceğine karar vermek için üyelik dereceleri kullanılır. Her bir nesne için o nesnenin kümelere üyelik derecelerine bakılır ve en büyük üyelik derecesini veren kümeye nesne dahil edilir. Bir nesnenin birden fazla küme ile ilişkisi olabilir, ilişki derecesi üyelik değeri ile sunulur.

4. Deneyisel Çalışmalar

Deneyisel çalışmalar sırasında, farklı eğitim seti boyutları, farklı algoritmalar ve farklı test seti boyutları için deneyler yapılmıştır. Deneylerden beklenti en uygun kesişimlerin bulunmasıdır.

4.1. Veri Seti

Deneylerde kullanılmak üzere seçilen derlem ECI [16] adıyla bilinen çok dilli bir derlemidir. Bu derlem, dilbilimsel

çalışmalarda sıklıkla tercih edildiğinden ve Avrupa dilleri ile Türkçe için örnek verilere sahip olduğundan tercih edilmiştir. Klasik n-gram yönteminde İngiliz alfabesinde yer alan temel 26 harf kullanılarak n-gram serileri elde edilir ve bunlar üzerinde işlem yapılır. Bu çalışmada ise daha geniş bir alfabe üzerinde çalışma yapılacaktır. Önerdiğimiz özellik setinde aksan karakteri adı verilen özel karakterler de yer almaktadır. 15 dile ait özel karakterlerin de eklenmesiyle elde edilen bu alfabede 58 karakter bulunmaktadır. 58 adet karma özellik seti için olası bütün 3-gramların sayısı 195112 olmakla birlikte özellik seçimi yapılarak bu rakam 300 değerine indirilmiştir. Özellik seçiminde Fisher Filtering algoritması [17] kullanılmıştır.

Karma Özellik Seti:

a,â,á,ä,ã,ä,ä,æ,b,c,ç,d,e,é,ê,ë,f,g,ğ,h,i,ı,ı,ı,ı,j,k,l,m,n,ñ,o,ò,ó,ö,õ,p,q,r,s,ş,t,u,ü,û,ü,ü,v,w,x,y,ÿ,ß,ø,z

4.1.1. Metinsel Verilerin Sayısal Hale Dönüşümü

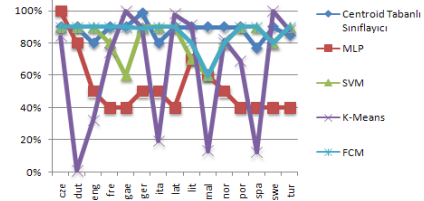
Metinsel veriler üzerinde çalışabilmek için dönüşüm ihtiyacı vardır, dönüşüm için de en uygun yöntem sayısalılaştırmadır. Dönüşüm işleminden önce her dil için 100 KB ve 1 KB eğitim verisi kullanılmış ve bu veriler 1 KB(100 KB için), 100 byte(1 KB için) uzunluklu parçalara ayrılmıştır. Böylece bütün derlem için 1500(100 KB eğitim metni için) ve 150(1 KB eğitim metni için) adet kayıt elde edilmiştir. Ayrıca, derlemde 100 karakter boyutunda 1500 adet kayıt elde edilmiş ve test işleminde kullanılmıştır. Gerek eğitim amaçlı veriler gerekse test amaçlı veriler özellik setinde yer alan 300 adet trigramaya dayalı olarak özetlenmiştir. Deneyler sırasında her bir dilde en önemli 3-gramların neler olduğu da tespit edilmiştir. Bu bilgi de dil tanıma çalışmalarında kullanılan vektör uzayıdır.

Çizelge 1: 100 KB Dosyada Dillere Ait Verilerden En Önemli Trigram Seti.

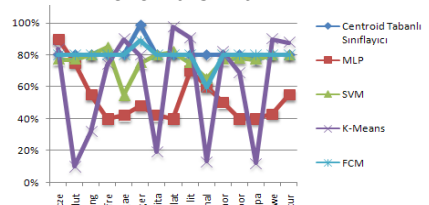
the	1483	ENG
kan	1168	MAL
zio	212	ITA
ikk	417	NOR
die	212	DUT
lar	760	TUR
ada	196	SPA
fro	343	GAE
som	71	SWE
jeg	535	NOR
ndo	302	POR
ost	320	CZE
que	314	LAT
ent	663	FRE
usi	329	LIT

4.3. Deneysel Sonuçlar ve Analiz

ECI CDROM derleminden elde edilen 15 dile (Türkçe-tur, İngilizce-eng, Almanca-ger, Hollandaca-dut, Fransızca-fre, İtalyanca-ita, Cezayirce-cze, İspanyolca-spa, Portekizce-por, Norveççe-nor, Maltaca-mal, Latince-lat, Litvanyaca-lit, İsveççe-swe, Andoa Dili-gae) ait dokümanlar; 300 adet 3-gram ile sunulmuş bu veriler üzerinde; centroid tabanlı, k-ortalamlar, bulanık c ortalamlar, yapay sinir ağları ve destek vektör makinesi yöntemleri çalıştırılmış ve aşağıda sunulan Şekil 1 ve Şekil 2'de sınıflandırma başarıları yüzdelik dilim olarak verilmiştir..



Şekil 1: Yöntemlerin 100 KB'lık Metin Boyutuna Göre Test Performans Analizi



Şekil 2: Yöntemlerin 1KB'lık Metin Boyutuna Göre Test Performans Analizi

Şekil 1 ve Şekil 2'den de görüldüğü üzere centroid tabanlı sınıflayıcı ve BCO(FCM) algoritması dil bazında en iyi tanıma performansı vermiştir. DVM (SVM) performansı onlara yakındır. En düşük performansı ise K-ortalamlar(K-Means) ve MLP(Çok Katmanlı Perceptron) vermiştir. Centroid sınıflayıcı başka çalışmalarda da iyi sonuç veren bir algoritmadır [10]. DVM ise genellikle metin sınıflandırma algoritmalarının hemen hepsinde yüksek sınıflandırma doğruluğuna sahiptir. MLP algoritmasının başarısız olması çok sayıda giriş ile K-Ortalamlar algoritmasının başarısızlığı centroid başlangıç değerlerinin rastgele atanmasıdır.

5. Sonuç

K-ortalamlar algoritması dışında tüm sınıflandırma yöntemleriyle yapılan eğitim verileri üzerindeki deneylerde; dosya boyutunun artmasıyla sınıflandırma algoritmalarının daha başarılı sonuçlar verdiği görülmüştür. Çok katmanlı yapay sinir ağı, çok girişli destek vektör makinesi ve centroid tabanlı sınıflandırma dilin tanınması konusunda başarılı sonuç vermiştir. Fakat k-ortalamlar metin boyutu artsa da başarılı sonuç verememiştir. Buna karşılık; k-ortalamlar algoritması gibi kesin sonuç vermeyen ama genel özellikler yoluyla kendini bir kümeye yakınsama yoluyla atayan BCO algoritması dosya boyutu arttığında iyi sonuç vermiştir. Bunun sebebi olarak da dilin ayırt edici özelliklerinin metin boyutuna bağlı olduğu düşünülmektedir.

Test amacıyla kullanılan bütün yöntemler dil tanımda

başarılı olmakla birlikte en yüksek başarıyı Centroid tabanlı sınıflayıcı ve BCO algoritması vermiştir. Dil gibi birbirine benzer özellikler gösteren sınıflandırma problemlerinde bulanık kümelemeye dayalı yaklaşımına yakın sınıflayıcıların daha yüksek başarı verdiği bir bulgu olarak ortaya çıkarılmıştır.

Bu çalışmada; özel harfleri dışarıda bırakmayan karma bir alfabe kullanılmış, karma alfabeden dolayı artan 3 - gram sayıları ise özellik seçimi ile düşürülmüştür. Dolayısıyla hem önemli bilginin yakalanması hedeflenmiş hem de performans düşmesine engel olunmuştur.

6. Kaynaklar

- [1] Grefenstette, G., "Comparing two language identification schemes", Proceedings of JADT 3rd International Conference on Statistical Analysis of Textual Data, 1995.
- [2] Cavnar, W.B. ve Trenkle, J.M., "N-gram-based text categorization", Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, US 161-175, 1994.
- [3] Suzuki, I., Mikami, Y., Ohsato, A., Chubachi, Y., "A language and character set determination method based on N-gram statistics", ACM Transactions on Asian Language Information Processing (TALIP), Volume 1 Issue 3, Pages 269 - 278, 2002.
- [4] Adams, G., ve Resnik, P., "A language identification application built on the Java client-server platform", In J. Burstein & C. Leacock (Eds), From Research to Commercial Applications: Making {NLP} Work in Practice Somerset, New Jersey: Association for Computational Linguistics, pp. 43 - 47, 1997.
- [5] Beesley, K., "Language Identifier: A computer program for automatic natural language identification of on-line text", Proceedings of the 29th Annual Conference of the American Translators Association, 47 - 54, 1988.
- [6] Dunning, T., "Statistical identification of language", Technical Report, New Mexico State University, CRL MCCC-94-273, 1994.
- [7] Combrinck, H., ve Botha, E., "Automatic language identification: Resisting complexity", South African Computer Journal, 27, 18 - 26, 1995.
- [8] Prager, J. M., Linguini, "Language identification for multilingual documents", Journal of Management Information Systems, 16(3), 71 - 101, 1999.
- [9] Xafopoulos, A., Kotropoulos, C., Almpandis, G., ve Pitas, I., "Language identification in web documents using discrete hidden Markov models", Pattern Recognition, 37(3), 583 - 594, 2004.
- [10] Takçı, H., Soğukpınar, İ., "Centroid-Based Language Identification Using Letter Feature Set", Lecture Notes in Computer Science, (CICLING 2004) Springer-Verlag, Vol.2945/2004, pages 635-645, February 2004.
- [11] Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3) 273-297, 1995.
- [12] Knerr, S., Personnaz, L., Dreyfus, G., "Single-layer learning revisited: A stepwise procedure for building and training a neural network," in Neurocomputing: Algorithms, Architectures and Applications, J. Fogelman, Ed. New York: Springer-Verlag, 1990.
- [13] Mayoraz E., ve Elmas, E., "Support vector machines for multi-class classification," in IWANN, vol. 2,

pp. 833-842, 1999.

- [14] Elmas Ç., *Yapay Sinir Ağları Kuram, Mimari, Uygulama*, Seçkin Yayıncılık, Ankara, 2003.
- [15] Şen Z., *Mühendislikte Bulanık (Fuzzy Mantık) ile Modelleme Prensipleri*, Su Vakfı, İstanbul, 2004.
- [16] European Corpus Initiative Multilingual Corpus (ECI/MCI) (2005), <http://www.elsnet.org/resources/eciCorps.html>, Page last modified 29-03-2005.
- [17] Kira, K. ve Rendell, L.A., "A practical approach to feature selection.", Machine Learning: Proceedings of International Conference (ICML92) 249C256, 1992.