# Yapay Bağışıklık Sistemleri ile Türkçe Metinlerde Tür ve Yazar Tanıma

# Genre and Author Detection in Turkish Texts Using Artificial Immune Recognition Systems

*Zafer Kaban,Banu Diri*

Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi, İstanbul
zafer.kaban@argela.com.tr, banu@yildiz.edu.tr

## Özetçe

Bu çalışma [1]'de bahsedilmiş olan bir dokümanın farklı şekilde ifade edilmesi yöntemi kullanılarak, Yapay Bağışıklık Sistemlerinin bir metnin türünü ve yazarını tanımadaki başarısını araştırmak amacıyla yapılmıştır. Günümüzde yapılan metin sınıflandırma sistemlerinin bir çoğu, kelime kök ya da gövdelerini özellik olarak alan Bag of Words (Kelime Torbası) modeli ile yapılmaktadır. Bu durum hem özellik sayısını hem de sınıflandırma süresini arttırmaktadır. Bu çalışmada, Yıldız ve arkadaşları tarafından geliştirilen ve YTU yöntemi adını verdiğimiz metot kullanılarak, metinlerde geçen kelime gövdelerine bir ağırlıklandırma algoritması uygulanarak özelliklerin sayısı sınıf sayısına indirilmekte ve bu sayede hem sınıflandırma süresi hem de başarısı arttırılmış olmaktadır. Bu yöntem ile test ettiğimiz Yapay Bağışıklık Sistemi algoritmaları olan YBS1, YBS2, YBS2Parallel tür ve yazar tanımada alınan başarı sonuçlarının yükselmesini sağlamıştır.

Makalenin deneysel sonuçlar bölümünde tür ve yazar tanıma çalışmalarında çok sık kullanılan Naive Bayes (NB), Destek Vektör Makinesi (DVM), Rasgele Orman (RO) ve K-En Yakın Komşuluk (K-EK) sınıflandırıcılarından alınan sonuçlar, Yapay Bağışıklık Sistemi algoritmalarından elde edilen sonuçlar ile karşılaştırmalı olarak verilmiştir. Özellikle tür tanımada YBS2Parallel sınıflandırıcısı %99,6'lık başarı oranı ile K-En Yakın Komşuluk ve Rastgele Orman'la birlikte en yüksek başarı oranını vermektedir. Bu da bu yöntem kullanılarak Yapay Bağışıklık Sistemi algoritmalarının tür tanımada kullanılabileceğini göstermektedir.

## Abstract

This study is made for investigating the performance of Artificial Immune Recognition Systems on genre and author detection by using method referenced as [1] based on representation of a document in a different scheme. Most of the studies done nowadays depend on Bag of Words model which takes the roots or the stems of the words as features. This situation both increases the number of features and the classification time. In this study, the method we named as YTU is used which applies a weighting algorithm on word stems and decreases the number of features to the number of classes resulting in lower classification time and better performance. Artificial Immune Recognition algorithms AIRS1, AIRS2, AIRS2Parallel which we tested by this method increased the performance in genre and author detection.

In the experimental results section of the paper the comparison of the classification performance of mostly used classifiers on author and genre detection Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbourhood (K-NN) and the Artificial Immune Systems algorithms are presented. Especially in genre detection AIRS2Parallel classifier gives the highest performance of 99,6% with Random Forest and K-Nearest Neighbourhood. This shows that Artificial Immune Recognition algorithms can be used in genre detection.

## 1. Giriş

Günümüzde bir çok bilgi kaynağı, elektronik ortamda yapılandırılmamış metin belgeleri şeklinde tutulmaktadır. Bu durum kullanıcıların aradıklarını bulabilmeleri için, sayısı hızla artan dokümanlar arasından dokümanları araştırma, analiz etme, gruplandırma uğraşlarını ortaya koymalarını gerektirmektedir. Bu duruma çözüm bulmak amacıyla otomatik olarak dokümanları önceden belirlenmiş sınıflara ayrıştıran sistemler geliştirilmektedir. Son zamanlarda Destek Vektör Makinesi, Naive Bayes, K-En Yakın Komşuluk gibi bir çok makine öğrenmesi yöntemi metin sınıflandırma çalışmalarında kullanılmıştır. Yapay Bağışıklık Sistemleri ise metin sınıflandırma çalışmalarında daha önce denenmemiş bir yöntemdir ve yapılan denemelerde uygun özellik vektörleri kullanıldığında, özellikle tür tanıma sistemlerinde başarılı sonuçlar vereceğini göstermiştir.

Doküman sınıflandırma çalışmalarında bugüne kadar çeşitli özellikler kullanılmıştır. Burrows [2] kelime zenginliğine dayalı özellikleri, Stamatatos ve arkadaşları [3] sözdizimsel stil özelliklerinin çeşitli kombinasyonlarını, Fürnkranz [4] n-gramları, Tan ve arkadaşları [5] 2-gramları kullanmış, Çatal ve arkadaşları [6] n-gramları geliştirerek NECL adını verdikleri bir sistem geliştirmişlerdir. Amasyalı ve Diri [7] n-gramları kullanarak tür, yazar ve cinsiyet belirleme üzerine çalışmışlardır. Morton [8] cümle uzunluklarını, Brainerd [9] ortalama hece sayısını, Holmes [10] kullanılan kelime sayısı ve doküman uzunluğunu, Yıldız ve arkadaşları [1] ise YTU adını verdikleri kelimelerin ağırlıklandırılmasına dayalı kendi geliştirdikleri bir yöntemi tür tanıma uygulamasında kullanmıştır.

Çalışmanın ikinci bölümünde sınıflandırma yöntemleri, üçüncü bölümde uygulanan sistemin yapısı, dördüncü ve

978-1-4244-1999-9/08/$25.00 ©2008 IEEE

beşinci bölümlerde sırasıyla deneysel sonuçlar ve sonuç yer almaktadır.

## 2. Sınıflandırma Yöntemleri

Bu çalışmada, yazarı ve türü bilinmeyen bir dokümanın hangi sınıfa ait olduğunun bulunmasında makine öğrenmesi yöntemlerinden Naïve Bayes, Destek Vektör Makinesi, K-En Yakın Komşuluk ve Rasgele Orman seçilmiştir. Adı geçen bu sınıflandırma algoritmaları ve Yapay Bağışıklık Sistemine (Artificial Immune System) ait algoritmalar WEKA ortamında gerçekleştirilmiştir [11].

**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 [12] çalışmasında önerilen Naïve Bayes versiyonu kullanılmıştır.

**Destek Vektör Makinesi (DVM)** 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.

**Rasgele Orman (RO)** Breiman 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 [13] algoritmasıyla elde edilir. Önce her karar ağacı kendi kararını verir, karar ormanı içerisinde maksimum oyu olan sınıf son karar olarak kabul edilir ve gelen test verisi o sınıfa dahil edilir.

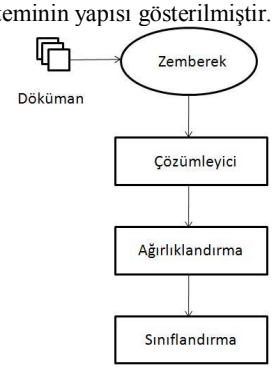
**K-En Yakın Komşuluk (K-EK)** Sınıfı bulunacak olan verinin özellik vektörü değerine en yakın olan *k* adet özellik vektörlerinin bulunması prensibine dayanır. En yakın özellik vektörleri bulunurken Euclidean mesafesi kullanılmıştır. Bulunan bu *k* adet vektör en fazla hangi sınıfa ait ise, o sınıf etiketi sınıflandırılacak olan verinin sınıfı olarak belirlenir.

**Yapay Bağışıklık Sistemleri (YBS)** Her canlı kendisini dış tehditlerden koruyacak bir savunma yeteneğine sahiptir. Omurgalı bağışıklık sistemleri vücutlarına yayılmış molekül, hücre ve organlardan oluşmaktadır. Bağışıklık sistemini kontrol eden merkezi bir organ yoktur. Bağışıklık sisteminin temel görevi hatalı davranış sergileyen hücreler (kanser ve tümör) ile hastalığa sebebiyet veren dış öğeleri (virüs ve bakteriler) tanıyarak vücudu bunlara karşı korumaktır. Bağışıklık sistemi tarafından tanınan her öğeye antijen denmektedir. Vücuda ait ve zararsız olan hücrelere self, hastalığa sebebiyet veren öğelere ise non-self denmektedir. Antijen tanıma, bağışıklık sisteminin karşı bir cevap başlatabilmesi için bir ön gereksinimdir. Tanıma sırasında önce hücre reseptörü belirli bir yakınlık (afinite) ile antijeni tanır. Eğer afinite bir eşik değerinden fazlaysa, bağışıklık sistemi aktive edilir. Antijenin yapısı, tanıma hücresinin türü ve tanıma bölgesi de antijen ve hücre arasındaki etkileşimin sonucunu belirleyen etkenlerdir [14]. Bu özellikleriyle YBS, örüntü tanıma, hesapsal güvenlik, anomali tespiti, optimizasyon, makine öğrenmesi, robotik, hata teşhisi gibi alanlarda ve bunların alt dallarında; ayrıca ekoloji, üretim sistemleri, akıllı evler, açık web sunucu koordinasyonu, protein yapısı tahmini gibi alanlarda başarıyla kullanılmış ve etkili sonuçlar alınmıştır (de Castro ve Timmis, 2002) [15]. Bu çalışmada da Türkçe dokümanlar üzerinde yazar ve tür tanımada da Yapay Bağışıklık sistemlerinin kullanılabileceğini gösterdik. Yapay bağışıklık sistemlerinin genellikle uygulanan algoritmaları negatif ve pozitif seçim algoritmaları, klonal seçim algoritması ve bağışık ağ modelleridir.

**Korelasyon Tabanlı Özellik Seçici (Correlation-based Feature Selection-CFS)** CFS, Weka içerisinde yer alan özellik seçici metotlardan biridir. Diğer özelliklerle düşük korelasyonlu, sınıf değişkeni ile yüksek korelasyonlu olan özellikleri seçer.

## 3. Sınıflandırma Sisteminin Yapısı

Sınıflandırma işlemi için ilk önce metinlerde geçen kelimelerin gövdeleri Zemberek [16] adlı açık kaynak kodlu kütüphane kullanılarak çıkarılmıştır. Çözümleyici modül yardımıyla bulunan gövdelerin her sınıfta kaç kez geçtiği ve sınıf içerisinde kaç farklı dokümanda bulunduğu bilgisi Trie ağaç yapısında tutulur. Bu yapı bilgiye erişimde kolaylık sağlamakta ve aynı zamanda hızı arttırmaktadır. Her dokümana ait çıkarılan özelliklerle birlikte sınıf sayısı kadar özellik içeren bir özellik vektörü elde edilir. Şekil-1'de sınıflandırma sisteminin yapısı gösterilmiştir.




*Şekil 1:* Sınıflandırma Sisteminin Yapısı

### 3.1. Veri Seti

Bu çalışmada veri seti olarak Hürriyet (www.hurriyet.com.tr), Milliyet (www.milliyet.com.tr) ve Sabah (www.sabah.com.tr) gazetelerinin internet sitelerinden alınan 18 farklı köşe yazarına ait makaleler kullanılmıştır. Bu on sekiz yazarın 4'ü kadın, 14'ü erkek olup, her yazara ait 35 doküman alınmıştır. Tür belirlemede spor, ekonomi, politika, sağlık ve popüler olmak üzere beş farklı sınıf seçilmiş ve her sınıf için 50 doküman alınmıştır. Sınıflandırma işleminde on kat çapraz geçerlilik (10-fold cross validation) kullanılarak sonuç alınmıştır.

### 3.2. Kelime Gövdelerinin Bulunması

Türkçe sondan eklemeli bir dil olup, kelimenin köklerine yapım ekleri eklenerek yeni kelime gövdeleri, çekim ekleri eklenerek de anlamı değişmeyen çok farklı kelimeler elde edebiliriz. Bu çalışmada kelimelerin kendileri değil kelime gövdeleri kullanılmıştır. Böylece eklerden doğan kelime farklılıkları ortadan kaldırılarak, frekansı bulunacak kelime sayısı azaltılmıştır. Bu çalışmada Doğal Dil İşleme kütüphanesi olan açık kaynak kodlu Zemberek [15] kullanılarak kelimeler çözümlenmiş ve kökleri bulunmuştur.

### 3.3. Kelime Frekanslarının Bulunması ve Trie Ağacı

Kelime gövdelerinin dokümanlarda bulunma sıklıkları ve kaç farklı dokümanda geçtiği bilgisi Trie ağaç yapısı kullanılarak elde edilmiştir. Trie, İngilizce re*trie*val kelimesinden gelmekte olup, ön ek ağaç olarak da adlandırılmaktadır. Trie ağacı, düğümlerinde kelimelerin harflerini sıralı bir şekilde tutar ve kelimelerin son harfini tutan düğümlerde de o kelimenin kullanım frekansı yer alır. Ağaca yeni bir kelime eklerken ağaçta harf harf gezilerek kelimenin son harfine kadar ilerlenir. Eğer kelimenin son harfi bir düğüme karşılık geliyorsa, bu düğümün bilgisi güncellenir, eğer gelinmiyor ise ağaca yeni bir düğüm eklenir ve başlangıç değeri verilir. Ağaca kelime eklemek ve aramak bu yöntem ile çok hızlı bir şekilde gerçekleştirilir. Kelimenin arama karmaşıklığı, aranan kelimenin uzunluğu ile orantılıdır.

### 3.4. Özellik Vektörünün Oluşturulması

Günümüzde yapılan bir çok metin sınıflandırma sistemi tüm dokümanlarda geçen kelime kök ya da gövdelerini özellik olarak alan Bag of Words (Kelime Torbası) modeli ile gerçekleştirilmektedir. Bu durum hem özellik sayısını hem de sınıflandırma süresini arttırmaktadır. Bazı kelimeler sadece birkaç metinde geçerken bazıları her dokümanda yer almaktadır. Bu kelimelerin ayırt edici bir özelliği olmadığından, metnin sınıfına olan etkisini azaltmak için Eşitlik 1'deki ağırlıklandırma kullanılmaktadır.

$$w_i = \log(tf_i + 0.5) * \log(D/df) \quad (1)$$

$D$ = Toplam metin sayısı
$df$ = Kelimenin geçtiği metin sayısı
$tf_i$ = Kelimenin i. sınıfta geçtiği metin sayısı
$w_i$ = Kelimenin i. sınıfa göre ağırlığı

Önerdiğimiz özellik vektörünün boyutu, sınıf sayısına eşit olup, doküman içerisinde yer alan her kelime gövdesinin bütün sınıflar içerisindeki kullanım frekansları çıkarılıp, sınıf genelinde toplamları alındığında, her metnin ağırlıklandırılması gerçekleştirilmiş olur. Sonuç olarak yazar tanımada her doküman dosyası 18 boyutlu, tür tanımda ise 5 boyutlu özellik vektörü ile ifade edilir.

Elimizde iki sınıfa ait sekiz adet doküman ($d_i$), üç adet kelime ($k_j$) ve kelimelerin, doküman ve sınıflara göre dağılımı Tablo 1'deki gibi verilmiş olsun.

*Tablo 1:* Örnek veri seti

| | ***Sınıf 1*** | | | | ***Sınıf 2*** | | | |
|---|---|---|---|---|---|---|---|---|
| | **d1** | **d2** | **d3** | **d4** | **d5** | **d6** | **d7** | **d8** |
| **k1** | 1 | 0 | 2 | 1 | 1 | 1 | 0 | 0 |
| **k2** | 1 | 0 | 1 | 0 | 2 | 1 | 0 | 1 |
| **k3** | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |

YTU yöntemini kullanarak her dokümanı farklı kelime sayısı yerine, sınıf sayısı kadar bir özellik vektörü ile ifade edebiliriz. Örneğin *d1* dokümanı ifade etmek için kullandığımız özellik vektörünün değerleri sırasıyla 0.6058 ve 0.3941'dir.
***k1* kelimesi:** *sınıf 1* için $w_1$=0.1333, *sınıf 2* için $w_2$=0.0812
***k2* kelimesi:** *sınıf 1* için $w_1$=0.0812, *sınıf 2* için $w_2$=0.1333
***k3* kelimesi:** *sınıf 1* için $w_1$=0.1966, *sınıf 2* için $w_2$=0.053
$w_1$'ler toplamı 0.4111, $w_2$'ler toplamı 0.2675 olup iki sınıf toplamı 0.6786'dır. Buna göre *d1*'in ilk özellik değeri 0.4111/0.6786, ikinci özellik değeri de 0.2675/0.6786'dır.

## 4. Deneysel Sonuçlar

Çalışmanın deneysel sonuçlarını alırken, dokümanlarda üç veya daha az sayıda geçen kelime gövdeleri değerlendirmeye katılmamıştır. Çünkü bu kelimelerin aynı sınıftaki dokümanlarda geçme olasılıklarının çok düşük olacağı düşünülmüştür.

YTU yöntemini kullanarak, dokümanları içerisinde geçen farklı kelime sayısı kadar bir boyutla ifade etmek yerine, sınıf sayısı kadar bir boyut ile göstermek, sınıflandırma başarısını arttırmaktadır. Kısaca kelimelerin dokümanlardaki ağırlıkları yerine, sınıflardaki ağırlıkları kullanılmış ve dokümanda geçen kelimelerin sınıf ağırlıkları toplanarak metnin yeni özellik vektörü oluşturulmuştur.

Tablo 2 tür ve Tablo 3 yazar tanımada eşitlik 1'deki ağırlıklandırma yöntemi kullanılarak sırasıyla tüm kelime gövdelerine (tür tanımada 5873, yazar tanımada 8488 farklı kök) göre, daha sonra özellik azaltarak (tür tanımada 74, yazar tanımada 66 farklı kök) ve son aşamada da YTU yöntemi kullanılarak klasik makine öğrenmesi metotlarından Destek Vektör Makinesi, Naive Bayes, K-En Yakın Komşuluk (k=1 ve k=3 için) ile Yapay Bağışıklık Sistemi algoritmalarının (YBS1, YBS2 ve YBS2Paralel) sınıflandırmadaki başarı sonuçları karşılaştırmalı olarak verilmiştir.

Tablo 2 tür tanımada görüldüğü gibi tüm gövdeler özellik olarak alındığında çoğu sınıflandırıcının başarısı oldukça düşük çıkmaktadır. Aynı özellik vektörüne, özellik azaltma yöntemi uygulanarak elde edilen yeni özellik vektörü ile sınıflandırma yapıldığında başarının yükseldiği izlenmektedir. Metinlerin farklı şekilde ifade edilip, sınıf sayısı kadar özelliğe sahip olunan YTU ile sınıflandırma yapıldığında ise bütün sınıflandırma algoritmalarından yüksek performanslar alınmıştır. Özellikle Yapay Bağışıklık Sistemi algoritmalarının sınıflandırma başarılarındaki artışlar göze çarpmaktadır. En yüksek başarıyı %99.6 ile YBS2, RO ve K-EK sınıflandırıcısı vermiştir.

*Tablo 2:* Tür Tanıma için Başarı Oranları (%)
(w=0 tüm kök değerleri, $w=0_{CFS}$ özellik azaltılmış, w=1 YTU yöntemi)

| **w** | **K-EK (k=1)** | **K-EK (k=3)** | **NB** | **RO** | **DVM** |
|---|---|---|---|---|---|
| **0** | 54.4 | 50.0 | 80.8 | 76.0 | 96.4 |
| **$0_{CFS}$** | 85.2 | 84 | 93.6 | 86.4 | 94 |
| **1** | 99.6 | 99.6 | 98.8 | 99.6 | 98.8 |
| | **YBS1** | **YBS2** | | **YPSP** | |
| **0** | - | 24.4 | | 40.8 | |
| **$0_{CFS}$** | 60.4 | 58.4 | | 61.6 | |
| **1** | 99.2 | 99.6 | | 99.2 | |

Yine Tablo 3' te görüldüğü gibi yazar tanımada da, YTU yöntemi tüm algoritmalarda özellikle de Yapay Bağışıklık Sistemi algoritmalarında sınıflandırma başarısında artış sağlamıştır. Ancak en yüksek başarı %99.20 ile DVM'den alınmıştır. Yapay Bağışıklık Sistemi algoritmalarından YBSP %98.25'ik bir başarı elde etmiştir. DVM ile arasında yaklaşık %1'lik bir fark vardır.

Tablo 4'de tür tanımada 250 adet dokümanın birbirlerine göre sınıflandırma doğrulukları Hata Matrisi (confusion matrix) ile gösterilmiştir. Bu sonuçlar dokümanların Eşitlik 1'le ifade edilip, YBS2P yöntemi ile sınıflandırıldığında alınmış olan değerlerdir.

*Tablo 3* Yazar Tanıma için Başarı Oranları (%)
(w=0 tüm kök değerleri, w=$0_{CFS}$ özellik azaltılmış, w=1 YTU yöntemi)

| w | K-EK (k=1) | K-EK (k=3) | NB | RO | DVM |
|---|---|---|---|---|---|
| **0** | 9.21 | 17.46 | 54.6 | - | 90.48 |
| **$0_{CFS}$** | 53.02 | 53.65 | 69.37 | 53.65 | 72.70 |
| **1** | 98.89 | 98.57 | 92.70 | 93.49 | 99.20 |
| | **YBS1** | **YBS2** | | **YBSP** | |
| **0** | NA | 10.32 | | 16.83 | |
| **$0_{CFS}$** | 26.83 | 34.13 | | 34.13 | |
| **1** | 95.87 | 95.56 | | 98.25 | |

*Tablo 4:* Tür Tanıma için Hata Matrisi

| | **Tahmin Edilen Tür** | | | | | |
|---|---|---|---|---|---|---|
| **Gerçek Tür** | Ekonomi | Popüler | Sağlık | Siyaset | Spor | Hata |
| Ekonomi | **49** | 1 | 0 | 0 | 0 | 0.02 |
| Popüler | 0 | **50** | 0 | 0 | 0 | 0 |
| Sağlık | 0 | 0 | **50** | 0 | 0 | 0 |
| Siyaset | 0 | 1 | 0 | **49** | 0 | 0.02 |
| Spor | 0 | 0 | 0 | 0 | **50** | 0 |

Ortalama: 0.008

Tür tanımada Yapay Bağışıklık Sistemi algoritmalarından YBSP sınıflandırmada en iyi sonucu vermiştir. Ekonomi ve siyaset içerikli dokümanları yüzde iki hata ile sınıflandırmıştır. Bu iki türde hatalı sınıflandırılan doküman popüler olarak etiketlenmiştir. Metnin yazıldığı dönem ekonomi veya siyaset alanında konuşulan bir konunun, herkes tarafından tartışılan bir platforma taşınması durumunda ekonomi ve siyaset sınıfına ait metinlerde geçen kelimelerin, popüler sınıfında da geçiyor olmasına neden olduğundan hatalı sınıflandırma yapılmıştır. Sistemin ortalama hatası Tablo 4'de gösterildiği gibi yüzde 0.8'dir.

## 5. Sonuçlar

Bu çalışmada, yazarı bilinmeyen bir dokümanın önceden belirlenmiş 18 farklı yazar içerisinden hangisine ve türü bilinmeyen bir dokümanın, 5 sınıf içerisinden hangisine ait olabileceği sorusunun cevabı aranmıştır. Kelimelerin metinlerdeki ağırlıklarının yerine, sınıflardaki ağırlıkları kullanılmış ve metinde geçen kelimelerin sınıf ağırlıkları toplamı alınarak doküman için yeni bir özellik vektörü oluşturulmuştur. Bu özellik vektörleri kullanılarak K-En Yakın Komşuluk, Naïve Bayes, Rasgele Orman, Destek Vektör Makinesi gibi makine öğrenmesi yöntemleri ile birlikte Yapay Bağışıklık Sistemleri kullanılarak 10'lu çapraz geçerlilik ile sınıflandırma yaptığımızda oldukça başarılı sonuçlar elde edilmiştir.

Tablo 2 ve Tablo 3'te görüldüğü gibi YTU yöntemi kullanılarak kelime gövdelerinin ağırlıklandırılmasından elde edilen özelliklerin kullanılması, özellikle Yapay Bağışıklık Sistemi sınıflandırıcılarında başarının artmasını sağlamıştır. Özellikle tür tanımada YBS2Parallel sınıflandırıcısından %99,6'lık başarı elde edilmiştir.

Bu çalışma her ne kadar Türkçe için yapılsa da dilden bağımsız olup, farklı diller içinde kullanılabilir.

## 6. Kaynakça


[1] Yıldız, H. K., Gençtav, M., Usta N., Diri, B. ve Amasyalı ,M.F., "Metin Sınıflandırmada Yeni Özellik Çıkarımı", *IEEE SIU 2007 15. Sinyal İşleme, İletişim ve Uygulamaları Kurultayı*, 2007.

[2] Burrows, J. F., "Not unless you ask nicely: the interpretative nexus between analysis and information", *Literary Linguist Compu,t* 7:91-109, 1992.

[3] Stamatatos, E., Fakotakis, N. ve Kokkinakis, G., "Automatic Text Categorization in Terms of Genre and Author", *Computational Linguistics*, pp.471-495, 2000.

[4] Fürnkranz, J., "A Study using n-gram Features for Text Categorization", *Austrian Research Institute for Artifical Intelligence,* 1998.

[5] Tan, C. M., Wang, Y. F. ve Lee C. D., "The Use of Bi-grams to Enhance" *Journal Information Processing and Management*, Vol:30 No:4 pp.529-546, 2002.

[6] Çatal, Ç., Erbakırcı, K. ve Erenler, Y., "Computer-based Authorship Attribution for Turkish Documents", *Turkish Symposium on Artificial Intelligence and Neural Networks,* 2003.

[7] Amasyalı, M. F., Diri, B., "Automatic Written Turkish Text Categorization in Terms of Author, Genre and Gender", *11th International Conf. on App. of Natural Language to Information Systems*, Austria, 2006.

[8] Morton, A. Q., "The Authorship of Greek Prose ", *Journal of the Royal Statistical Society*, Series A, 1965.

[9] Brainerd, B., "Weighting Evidence in Language and Literature: A Statistical Approach", *University of Toronto Pres,* 1974.

[10] Holmes, D. I., "Authorship Attribution", *Comput Humanities*, 28:87-106, 1994.

[11] www.cs.waikato.ac.nz/ml/weka/

[12] George, H., "Estimating Continuous Distributions in Bayesian Classifiers". *Proceedings of the Eleventh Conference on Uncertainty in Artificial In*telligence, pp. 338-345. Morgan Kaufmann, San Mateo, 1995.

[13] Breiman, L., "Random forests–random features", Technical Report 567, *Department of Statistics, University of California, Berkeley*, 1999.

[14] Forrest, S., Perelson, A., Allen, L., "Self-Nonself Discrimination in a Computer ", *Proceedings of the IEEE Symp. on Research in Security and Privacy*, 1994.

[15] Alataş, B., Akın, E.,"Yapay Zekada Yeni Bir Alan: Yapay Bağışıklık Sistemleri", *YA/EM'2004-Yöneylem Araştırması/Endüstri Müh.-XXIV Ulusal Kongresi*

[16] http://code.google.com/p/zemberek/