**YILDIZ TEKNİK ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

# TÜRKÇE DOKÜMANLAR İÇİN N-GRAM TABANLI SINIFLANDIRMA: YAZAR, TÜR ve CİNSİYET

Bilgisayar Mühendisi Sibel DOĞAN

**FBE Bilgisayar Mühendisliği Anabilim Dalında Hazırlanan**

**YÜKSEK LİSANS TEZİ**

**Tez Danışmanı :** Yrd. Doç. Dr. Banu DİRİ (YTÜ)

İSTANBUL, 2006

**İÇİNDEKİLER**

Sayfa

## KISALTMA LİSTESİ

| | |
|---|---|
| AADTT | Automatic Author Detection for Turkish Text |
| BT | Bilişim Teknolojileri |
| BNN | British National Corpus |
| CWF | Common Word Frequency (en fazla sıklıkla kullanılan) |
| DVM | Destek Vektör Makineleri |
| K-NN | K-Nearest Neighbor |
| K-EK | K-En Yakın Komşuluk |
| KTÖS | Korelasyon Tabanlı Özellik Seçme |
| NB | Naive Bayes |
| RF | Ramdom Forest |
| RO | Rasgele Orman |
| SVM | Support Vector Machine |
| VR | Vocabulary Richness |

## ŞEKİL LİSTESİ

## ÇİZELGE LİSTESİ

## ÖNSÖZ

Bu çalışma, Yapay Zeka'nın bir dalı olan Doğal Dil İşleme alanında yapılmıştır. Çalışmanın amacı Türkçe'nin n-gram modelini çıkarmak ve daha sonra bu modeli kullanarak Türkçe dokümanların önceden belirlenmiş sınıflara atanabilmesidir. N-gram modelinden yararlanarak farklı özellik vektörleri çıkarılmış ve seçilen sınıflandırma yöntemleriyle dokümanın ait olduğu sınıf, dokümanın yazarının kim olduğu ve doküman yazarının cinsiyetinin belirlenmesi sağlanmıştır. Bu çalışma, n-gram modeli kullanılarak Türkçe dokümanlar için tür belirleme, yazar tanıma ve cinsiyet saptama alanında yapılmış en kapsamlı çalışmadır.

Çalışmalarımın her aşamasında bilgi ve deneyimleri ile bana yardımcı olan ve her türlü olanağı sağlayan danışmanım Sayın Yrd.Doç.Dr.Banu Diri'ye içtenlikle teşekkür ederim.

Desteğini esirgemeyip yol almamdaki yardımları için Filiz Türkoğlu'na teşekkür ederim.

Çalışmalarım süresince güvenini ve desteğini esirgemeyen Muharrem Dermenci'ye teşekkür ederim.

Çalışmalarımda bana en büyük moral, destek ve anlayışı gösteren aileme ve arkadaşlarıma sonsuz teşekkür ederim.

## ÖZET

Bizler bilgilerin değerli olduğu bir dünyada yaşıyoruz. Son yıllarda bilgi miktarının artması, ihtiyaç duyduğumuz bilgilere kısa sürede erişebilmeyi zor hale getirmiştir ve bu problem nedeniyle bu işlerin elle yapılabilmesi neredeyse imkansızdır. Problemin çözümü için doküman sınıflandırma sistemlerine ihtiyaç duyulmaktadır. Diğer dillerin aksine bu konuda Türkçe üzerinde çok az çalışma mevcuttur. Sınıflandırma işlemi doküman işleme için önemli bir konu olup, elektronik ortamdaki dokümanların otomatik olarak sınıflandırılmasına izin verir. Bu çalışmada; Türk dilinin 2, 3 ve 4'lü gramları çıkarılarak farklı boyutlarda özellik vektörleri oluşturulmuştur. Daha sonra bu özellik vektörlerinin boyutları korelasyon tabanlı özellik seçiciler kullanılarak azaltılmış ve farklı boyutlarda özellik vektörleri elde edilmiştir. N-gram modeline dayalı bu özellik vektörleri, seçilmiş (sınıflandırma başarısı yüksek) sınıflandırma yöntemleri yardımıyla Türkçe bir dokümanın türünü, yazarını ve doküman yazarının cinsiyetini belirlemek amacıyla kullanılmışlardır.

Kullanılan veri seti spor, magazin, güncel, ekonomi, sağlık ve politika gibi farklı konularda yazan 20 yazara ait, 40 adet doküman alınarak 800 metinden oluşmaktadır. Dokümanın türünü, yazarını ve yazarın cinsiyetini belirlemek için eldeki veri seti üç ayrı formatta düzenlenmiştir. Ayrıca sınıflandırma başarısının tesadüfi olmadığını göstermek için tüm deneylerde 10-kat çapraz geçerlilik uygulanmıştır.

Tür, yazar ve cinsiyet belirlemede hangi n-gram özelliklerin daha başarılı olduğunu analiz etmek amacıyla beş farklı sınıflandırma metodu kullanılarak performansları birbirleri ile karşılaştırılmıştır. Bu metotlardan dördünü Naive Bayes, Destek Vektör Makinesi, Rastgele Orman, K-En Yakın Komşuluk gibi bilinen sınıflandırma yöntemleri, birini de bizim geliştirdiğimiz ng_ind yöntemi oluşturmaktadır. Sınıflandırıcıları birlikte kullanma işleminin başarısını gözlemlemek amacıyla, Naive Bayes, Destek Vektör Makinesi, Rastgele Orman ve K-En Yakın Komşuluk yöntemleri birlikte kullanılmıştır.

Yapılan denemelere göre, cinsiyet belirleme için bayan yazarların yazıları, tür belirleme için spor ve güncel alanlarda yazılmış yazılar, yazar tanımada da bayan yazarların yazıları daha başarılı sonuçlar vermiştir. Özelliklerin azaltılması ile elde edilen özellik vektörleri, diğer özellik vektörlerine göre daha iyi performans göstermiştir ve en yüksek başarı oranını, yazar tanımada DVM, tür ve cinsiyet belirlemede Ng-ind yöntemi vermiştir. Birlikte kullanılan sınıflandırıcılar ile bireysel sınıflandırıcılara göre daha yüksek başarı sonuçları alınmıştır.


**Anahtar kelimeler:** Yazar tanıma, tür belirleme, cinsiyet belirleme, sınıflandırıcıların birlikte kullanılması, Naive Bayes, Destek Vektör Makinesi, Rastgele Orman, K-Enyakın Komşuluk, ng_ind

**ABSTRACT**

We live in a world where the information has an important value. It has been diffucult to access the data we need in a reasonable time because of the increasing amount of data, and this has led a new problem of that doing this by hand has been almost impossible. Thus, document classification systems are needed in order to solve the problem. But, there are only a few studies about this topic in Turkish in spite of the other languages. Classification operation is an important subject for document processing, and it allows digital documents to be processed automatically. In this thesis study, property vectors of Turkish language in different dimensions have been constructed by finding out 2nd, 3rd, and 4th order grams. Then, the dimensions of these property vectors have been reduced by using correlation based propoerty choosers, and property vectors in different dimensions have been obtained. These property vectors have been used in order to determine the type, author name and author sex of a Turkish document by the help of classification methods.

The dataset used in the study contains 800 articles of 40 documents which are belong to 20 authors writing about sports, magazine, health and politics. The used dataset is arranged in three different formats in order to find out the the type, author name and author sex of the documents. Also, 10-times dioganal validity has been applied in order to demonstrate that classification success is not Random.

Performances have been compared by using five different types of classification methods in order to analyze which properties are more successfull for determining type, author name and author sex of documents. The four known ones of these classification methods are Naive Bayes, Support Vector Machine, Random Forest, and K-Nearest Neighbor methods. The last of these methods is ng_ind method that we proposed and developed in this study. Naive Bayes, Support Vector Machine, Random Forest, and K-Nearest Neighbor methods have been used together in order to observe the performance of the operation of using classificators together.

The experiments have shown that the articles about sports and daily news are more successfull for determining document types while the articles of women authors are more successfull for determining the author name and author sex of the documents. The property vector obtained by reducing the propoerties has had a better performance compared to others. DVM method has given the most successful result for author recognition while Ng-ind has given the most successful result for genre and gender determination. Classificators used together have given more successfull results compared with individual classificators.

**Keywords**: Author Recognition, Genre Determination, Gender Determination, Using Classificatiors Together, Naive Bayes, Support Vector Machines, Random Forest, K-Nearest Neighbor, ng_ind.

# 1. GİRİŞ

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. Doküman sınıflandırma bilgi alma (information retrieval), bilgi çıkarma (information extraction), doküman indeksleme, doküman filtreleme, otomatik olarak metaveri elde etme ve web sayfalarını hiyerarşik olarak düzenleme gibi pek çok alanda önemli bir rol oynamaktadır. Eskiden kategoriler insanlar tarafından oluşturulur ve doğru kategoriye atama yapabilmek için dokümanların okunma zorunluluğu vardı. Oysaki doküman sınıflandırma işlemi elle değil de elektronik ortamda, önerilen bazı yöntemler kullanılarak yapılırsa sınıflandırma için harcanan zamanda çok kısalmış olur.

Bu çalışma, Cavnar ve Trenkle tarafından yazılmış "N-gram Tabanlı Doküman Sınıflandırma" makalesinden esinlenilerek hazırlanmış (Cavnar,1994), ve Türk dili için uygulanmıştır. N-gram tabanlı sınıflandırma işlemi, doküman içerisindeki karakterlerin kullanım sıklığına dayalı bir yöntemdir.

Doküman sınıflandırmadaki problemlerden biri, kime ait olduğu bilinmeyen veya yazarının kimliğinden şüphelenilen dokümanların yazarının tahmin edilmesi, bir diğer problem de dokümanın türünün veya yazarının cinsiyetinin belirlenmesidir.

Her yazarın kendisine özgü bir yazım üslubu vardır. Örneğin, yazılarında her zaman okuyucuya soru soran, bazı cümlelerini vurgulamak için noktalama işaretlerini sık kullanan ya da aynı kelimeleri tekrarlayan yazarlarımız olabilir. "Belirli bir kişinin kaleminden dökülen yazılardaki karakter sıklıkları, diğer yazarların yazılarındaki karakter sıklıklarından farklıdır" düşüncesiyle yazar tanıma çalışması yapılmıştır. Yazar Tanıma, konudan bağımsız olarak elimizde bulunan dokümanların hangi yazar tarafından yazıldığını tahmin etme işlemidir. Yazar tanıma çalışmasındaki aynı mantıkla yola çıkılarak bay ve bayan yazarların birbirlerinden farklı bir yazım tarzına sahip olduğunu farz ederek cinsiyet belirleme çalışması da yapılmıştır. Karakter kullanım sıklık değerleri dil içinde birbirinden farklı değerler almaktadır. Belirli bir konuda yazılmış dokümanların karakter sıklıkları birbirine daha yakın iken farklı konularda yazılmış dokümanların karakter sıklıkları ise birbirinden daha uzaktır. Bu düşünceden yola çıkılarak farklı konularda yazılmış dokümanların konusuna göre

sınıflandırılabildiğini göstermek için yapılan tür tanıma çalışması da bu tezin kapsamına dahil edilmiştir.

Bu çalışmada, yukarıda bahsedilen sınıflara ayırmayı gerçekleştirebilmek için gerekli olan veri setleri İnternet'te belirli gazete sayfalarından, farklı yazarların yazıları toplanarak oluşturulmuştur. Kullanılan yazıların boyutları 3 KB ile 9 KB arasında değişmektedir. Tür tanımada kullanılan yazıların konuları spor, magazin, güncel, ekonomi, sağlık ve politika olarak belirlenmiştir ve yazılar bu kategorilere uygun olarak sınıflandırılmıştır. Tüm bu sınıflandırma işlemlerinde N-gram yöntemi kullanılmıştır. N–gram, uzun bir karakter katarının n-karakter dilimidir. Bu çalışma kapsamında, dokümanların yazarlarını, türlerini ve yazarların cinsiyetini belirlemede kullanmak amacıyla üç farklı veri seti oluşturulmuştur. Her veri seti için 2-gram, 3-gram ve 4-gram'lardan oluşan özellik vektörleri çıkarılmıştır. Her bir veri seti için üç özellik vektörü ve daha sonra bu özellik vektörlerinden türetilmiş özellik sayısı daha az olan vektörler kullanılarak sınıflandırma işlemi gerçekleştirilmiştir. Kısacası her veri seti için 6 adet özellik vektörü, tüm veri setleri için de 18 adet özellik vektörü çıkarılmıştır. Bu çalışmada önce Naive Bayes, Rastgele Orman, Destek Vektör Makinesi (DVM), K-En Yakın Komşuluk (K-EK) ve ng_ind gibi sınıflandırma yöntemleri tek başlarına kullanılmış daha sonra da bazı sınıflandırıcılar birlikte kullanılarak yazar, tür belirleme ve cinsiyet tahmin etme çalışması yapılmıştır.

Tezin ikinci bölümde, yazar tanıma, tür ve cinsiyet belirleme alanında daha önce yapılmış olan çalışmalara yer verilmiştir. Üçüncü bölümde n-gram yöntemi kullanılarak dokümanlardan elde edilen özellik vektörlerinin çıkarılması anlatılmış, dördüncü bölümde de bu çalışmada kullanılmış olan sınıflandırıcılardan ve bunların birlikte kullanılmasından bahsedilmiştir. Beşinci bölümde ise kullanılan veri setlerine göre deneysel sonuçlara ve yorumlara yer verilmiştir.

## 2. YAZAR TANIMA, TÜR ve CİNSİYET BELİRLEME ALANINDA YAPILMIŞ ÖNCEKİ ÇALIŞMALAR

Bu bölümde bu konuda yapılan ve literatüre girmiş çalışmalara yer verilecektir.

### 2.1 Yazar Tanıma ile İlgili Daha Önce Yapılan Çalışmalar

"Belirli bir kişinin kaleminden dökülen yazılardaki karakter sıklıkları, diğer yazarların yazılarındaki karakter sıklıklarından farklıdır" düşüncesiyle yazar tanıma çalışması yapılmıştır.Yazar Tanıma, konudan bağımsız olarak elimizde bulunan dokümanların hangi yazar tarafından yazıldığını tahmin etme işlemi olarak da bilinir. Yazarların birbirlerinden farklı bir yazım stiline sahip olduğu düşünülerek birçok yazar tanıma çalışması yapılmıştır.

Keselj 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 deneyler yapmışlardır (Peng ve Keselj, 2003). Bennett tarafından yapılan, n-gram yöntemi ile oluşturulan profil baz alınarak, iki doküman arasındaki benzerliği ölçme yöntemi revize edilerek bu çalışmada kullanılmıştır (Bennett, 1976). Bennett, 2-gram oluşturmak için alfabede bulunan 26 harfi kullanarak iki doküman arasındaki benzerliği (1)'deki formül ile elde etmiştir.

$$\sum_{I,J}\left[M(I,J)-N(I,J)\right]^2 \qquad (1)$$

I ve J, 2-gram'ı oluşturulan alfabenin 26 harfinden her birini, M(I,J) birinci dokümanda bulunan 2-gram'ın frekansını, N(I,J) ikinci dokümanda bulunan 2-gram'ın frekansını temsil etmektedir. Dokümandan elde edilen her *x* 2-gram'ı ve bu *x* 2-gram'ının frekansı, ilgili dokümanın L uzunluğundaki profilini oluşturmaktadır, ($f(x_1, f_1)$, $(x_2, f_2)$,$(x_L, f_L)$). Bennett tarafından kullanılan formül (2)'deki gibi revize edilmiştir.

$$\sum_{n\in profil}\left(\frac{f_1(n)-f_2(n)}{\frac{f_1(n)+f_2(n)}{2}}\right)^2=\sum_{n\in profil}\left(\frac{2*\left(f_1(n)-f_2(n)\right)}{f_1(n)+f_2(n)}\right)^2 \qquad (2)$$

$f_1(n)$ sınıflandırılacak dokümanın profilinde bulunan n-gram'ın frekansı, $f_2(n)$ kategori sınıfının profilinde bulunan n-gram'ın frekansıdır. İki profil arasındaki ölçüm sonucu elde

edilen farklılık (dissimilarity) değeri ne kadar az ise, dokümanın o sınıfa ait olma olasılığı o kadar fazladır. Yani farklılık değeri 0'a yaklaştıkça benzerlik artmaktadır.

Bu yöntem, Yunanca, İngilizce ve Çince'den oluşan 3 veri seti üzerinde uygulanmıştır. İngilizce veri setinde 8 yazarın yazıları kullanılmış ve 1'den 10'a kadar olan n-gram'lar çıkarılmış ve profil uzunluğu da 20 ile 5000 arası olarak alınmıştır. Profil uzunluğu 100 ile 3000, n-gram'lar 3 ile 8 olarak alındığında İngilizce veri seti kullanılarak %100'lük doğru sınıflandırma başarısı sağlanmıştır.

Yunanca veri seti olarak Stamatatos ve Kokkinakis'in yaptığı çalışmada kullanılan iki veri seti alınmıştır. Her veri setinde, 10 yazar ve yazarların 20 yazısı bulunmaktadır (Stamatatos ve Kokkinakis, 2000). Yazarların 20 yazısından 10'u eğitim seti, 10 tanesi de test seti olarak kullanılmıştır. Birinci veri seti gazete yazarlarının değişik konularda yazılmış yazılarından oluşmakta, ikinci veri seti de bilim, tarih ve sanat gibi dallarda yazılmış konuları içeren yazılardan oluşmaktadır. 1'den 10'a kadar olan n-gram'lar çıkarılmış ve profilin uzunluğu 20 ile 5000 arası olarak alınmıştır. Profil uzunluğu 1000 ve n-gram değeri de 3'ten büyük alındığında birinci veri seti kullanılarak en fazla %85'lik doğru sınıflandırma oranı, ikinci veri seti kullanıldığında ise en fazla %97'lik doğru sınıflandırma oranı ile başarı sağlanmıştır.

Çok sayıdaki Çince karakterlerden dolayı, Çince dokümanlardan değişik sayılarda n-gram'lar üretilmiş ve n-gram'ların sayısı 200,000 ile sınırlandırılmıştır. 1'den 10'a kadar olan n-gram'lar çıkarılmış ve profil uzunluğu da 20 ile 5000 arası olarak alınmıştır. Çince veri seti kullanılarak %89'luk doğru sınıflandırma oranı ile başarı elde edilmiştir.

Fuchun ve arkadaşları, karakter seviyeli n-gram modeline dayalı bilgisayar destekli yazar tanıma çalışması yapmıştır (Schuurmans, Peng, 2003). Bu çalışmanın dilden bağımsızlığını ve etkisini kanıtlamak için Yunanca, İngilizce ve Çince veriler üzerinde deneyler yapılmıştır. Her deneyde başarılı sonuçlar elde edilmiştir. Bu yaklaşımda yazarların yazıları n-gram modeli ile yapılandırılmış ve 3-gram, 4-gram, 5-gram ve 6-gram'lar kullanılmıştır. N-gram karakterlerinin hepsi muhtemel özellik olarak alınmıştır. Bu modelde tüm özellikler kullanılarak özellik seçiminden kaçınılmıştır. Ancak frekans değerlerini azaltmak için bazı yumuşatma teknikleri kullanılmıştır. Bayesian sınıflandırma yöntemi kullanılarak yazar tanıma işlemi yapılmıştır.

Fuchun ve arkadaşları tarafından, Yunanca veri seti olarak Stamatatos ve Kokkinakis'in yaptığı çalışmada kullanılan aynı iki veri seti kullanılmıştır (Stamatatos ve Kokkinakis, 2000). Yumuşatma tekniği (absolute smoothing) uygulanarak en iyi sonuç 3-gram'lar kullanılarak

elde edilmiştir. Birinci veri setinde en fazla %74'lük doğru sınıflandırma oranı, ikinci veri seti kullanılarak da en fazla %90'lık doğru sınıflandırma oranı ile başarı sağlanmıştır.

İngilizce veri seti olarak elektronik dokümanların Alex kataloğundan İngilizce dokümanlar alınmıştır ve bu katalogdan 8 tane yazar seçilmiştir. Denemelerde bütün büyük harf karakterleri küçük harfe çevrilmiş ve frekans değeri 30 dan büyük olan değerler kullanılmıştır. Yumuşatma tekniği uygulanarak en iyi başarı 6-gram'lar kullanılarak %98 olarak elde edilmiştir.

Çince veri seti İnternet'ten 8 dövüş ustası yazarının yazıları toplanarak oluşturulmuştur. Her bir yazarın 1 veya 2 yazısı eğitim seti ve 20 yazıda test seti olarak alınmıştır. Wittin Bell Yumuşatma tekniği uygulanarak en iyi sonuç 6-gram'lar kullanılarak elde edilmiştir. Çalışmada %94'lük doğru sınıflandırma ile başarı sağlanmıştır.

Stamatatos, Kokkinakis ve arkadaşları tarafından, Yunanca veri seti, Modern Greek weekly newspaper TO BHMA ismindeki gazetenin İnternet sitesinden dokümanlar toplanarak oluşturulmuştur. Yaptıkları çalışmada kullanılan dokümanlar 2 farklı veri setine dağıtılmıştır (Stamatatos ve Kokkinakis, 2000). Birinci veri seti gazete yazarlarının değişik konularda yazılmış yazılarından oluşurken, ikinci veri seti de bilim, tarih ve sanat gibi dallarda yazılmış konuları içeren yazılardan oluşmuştur. Her veri setinde, 10 yazar ve yazarların 20 yazısı bulunmaktadır. Yazarların 20 yazısından 10'u eğitim seti, 10 tanesi de test seti olarak kullanılmıştır. Bu çalışmada özellik vektörü oluşturabilmek için 4 yöntem kullanılmıştır.

- SCBD (Sentence and Chunk Boundaries), doküman analizi için kullanılan Doğal Dil İşleme Arac'ında (NLP Tool) var olan doküman işleme yöntemidir. Cümle sınır belirleme (Sentence boundary detection) ve yığın sınır belirleme (Chunk boundary detection) ismindeki iki yaklaşımdan oluşur. Cümle sınır belirleme yaklaşımında dokümandaki cümleleri belirlemek için nokta, ünlem, soru işareti ve üç nokta gibi noktalama işaretleri dikkate alınır. Yığın sınır belirleme yaklaşımında, dokümandaki cümleleri belirlemek için (Türkçe'de bulanan "ve, için" kelimeleri gibi) edat vb. anahtar kelimelerden 450 tanesi ve en çok kullanılan son eklerden (Türkçe'de bulanan -ler,-lar ekleri gibi) 300 tanesi alınmıştır. Bu çalışmada her doküman için 22 özelliğe sahip bir vektör kullanılmıştır.
- VR (Vocabulary Richness) sözcüklerin ağırlıklarını ölçmek için kullanılan bir yöntemdir (Holmes 1992).

- CWF-30 eğitim setini, en sık görülen kelimelerden oluşan frekans değeri 30 olan kelimeler oluşturmuştur.
- CWF-50 eğitim setini, en sık görülen kelimelerden oluşan frekans değeri 50 olan kelimeler oluşturmuştur.

Sınıflama için Ayrıştırma Çözümlenmesi (Discriminant Analysis) ve Çoklu Regresyon (Multiple Regression) yöntemleri kullanılmıştır.

Birinci veri setinde Ayrıştırma Çözümlenmesi için en fazla doğru olarak sınıflandırma oranı; SCBD yöntemi ile %72, VR yöntemi ile %41, CWF-50 yöntemi ile %54 ve CWF-30 yöntemi ile %54 olarak elde edilmiştir.

Birinci veri setinde Çoklu Regresyon için en fazla doğru olarak sınıflandırma oranı; SCBD yöntemi ile %66, VR yöntemi ile %37, CWF-50 yöntemi ile %56 ve CWF-30 yöntemi ile %55 olarak elde edilmiştir.

İkinci veri setinde Ayrıştırma Çözümlenmesi için en fazla doğru olarak sınıflandırma oranı; SCBD yöntemi ile %70, VR yöntemi ile %49, CWF-50 yöntemi ile %65 ve CWF-30 yöntemi ile %64 olarak elde edilmiştir.

İkinci veri setinde Çoklu Regresyon için en fazla doğru olarak sınıflandırma oranı; SCBD yöntemi ile %69, VR yöntemi ile %51, CWF-50 yöntemi ile %65 ve CWF-30 yöntemi ile %65 olarak elde edilmiştir.

Diri ve Amasyalı, Türkçe dokümanların yazarlarının belirlenmesi için dokümanın içeriğini ve belirlenen 22 farklı stil özelliğini kullanan 2 sınıflandırma yöntemi ile çalışmışlardır (Diri ve Amasyalı, 2003). Dokümanı içeriğine bağlı sınıflandırmada Naive Bayes, stil özelliklerine göre sınıflandırma da ise kendi geliştirdikleri *Automatic Author Detection for Turkish Text* metodunu kullanmışlardır. Denemeler iki farklı veri seti üzerinde yapılmıştır. Veri setleri İnternet'te belirli gazete sayfasından farklı yazarların yazıları toplanarak oluşturulmuştur. Külliyat-I, 18 yazardan oluşmaktadır. Bu yazarlara ait yazıların konuları birbirlerinden farklıdır. Bu veri setinde her bir yazarın 20 adet yazısı olmak üzere toplam 360 yazı vardır. 20 yazının 15 tanesi eğitim, 5 tanesi de test için kullanılmıştır. Külliyat-II, 18 yazardan oluşmaktadır. Bu veri setinde her bir yazarın 35 adet yazısı olmak üzere toplam 630 yazı vardır. 35 yazının 28 tanesi eğitim, 7 tanesi de test için kullanılmıştır. Külliyat-I kullanılarak elde edilen Naive Bayes yönteminin başarısı %43, AADTT yönteminin başarısı %84 olarak elde edilmiştir. Külliyat-II kullanılarak elde edilen Naive Bayes yönteminin başarısı %53, AADTT yönteminin başarısı ise %70 olarak alınmıştır. Her iki yöntem birlikte kullanıldığında

Külliyat-I için %89, Külliyat-II için %82'lik başarı oranı elde edilmektedir. Elde ettikleri sonuçlara göre stile dayalı özelliklerin daha başarılı olduğunu vurgulamışlardır. Naive Bayes ve AADTT yöntemi birlikte kullanıldığında başarı yükselmiştir.

Diri ve Amasyalı tarafından, farklı özellik vektörleri kullanarak Türkçe dokümanların yazarlarını belirleme çalışması yapılmıştır (Diri, Amasyalı ve Türkoğlu, 2006). Bu çalışmanın amacı, yazarı bilinmeyen bir dokümanın önceden belirlenmiş ve yazarlık özellikleri çıkarılmış 18 farklı yazardan hangisine ait olduğunu tespit etmektir. Veri setleri İnternet'te belirli gazete sayfasından (www.hurriyet.com.tr, www.vatanim.com.tr, www.sabah.com.tr) farklı yazarların politika, güncel, spor ve magazin gibi konularda yazıları toplanarak oluşturulmuştur. Eldeki mevcut veri seti 18 farklı yazarın her birine ait 35 farklı yazısından oluşan, 630 adet dokümandan meydana gelmektedir. Yazarı bulunması istenen dokümanlardan istatistiksel veriler, kelime sözlüğünün zenginliği, sık geçen Türkçe kelimeler, dilbilgisi özellikleri, n-gram'lar ve bunların çeşitli birleşimleri kullanılarak on farklı özellik vektörü elde edilmiştir. Daha sonra NB, DVM, C 4.5 ve RF sınıflandırma yöntemleri kullanılarak her bir özellik vektörünün sınıflandırmadaki başarıları karşılaştırılmıştır. En başarılı sonuç DVM ile sınıflandırılmasında %92,5 ile elde edilmiştir. Kullanılan sınıflandırıcıların en başarılısı, orijinal vektörlerle çalışıldığında açık farkla DVM, özellik azaltılmış vektörlerle çalışıldığında ise Naive Bayes ve DVM'dir. Bu çalışmada dokümanların sınıflandırılmasında, n-gram'ların kullanılmasının yazarlık özelliklerine göre daha başarılı olduğu gözlemlenmiştir.

Diri ve Amasyalı tarafından, n-gram yöntemi kullanılarak Türkçe dokümanların yazarlarını, türlerini ve cinsiyetini belirleme çalışması yapılmıştır (Diri ve Amasyalı, 2006). Veri seti İnternet'te belirli gazete sayfasından farklı yazarların politika, güncel, spor ve magazin gibi konularda yazıları toplanarak oluşturulmuştur. Eldeki mevcut veri seti 18 farklı yazarın 35'şer yazısı alınarak 630 doküman elde edilmiş, bunlardan 35 yazının 28 tanesi eğitim, 7 tanesi de test için kullanılmıştır. Veri seti için 2-gram ve 3-gram'lardan oluşan özellik vektörleri ve bu özellik vektörlerinden türetilmiş özellik sayısı daha az olan vektörler kullanılarak sınıflandırma işlemi gerçekleştirilmiştir. Kısacası 4 adet özellik vektörü çıkarılmıştır. Daha sonra NB, DVM, C 4.5 ve RF sınıflandırma yöntemleri kullanılarak her bir özellik vektörünün sınıflandırmadaki başarıları karşılaştırılmıştır. En başarılı sonuç NB ile sınıflandırılmasında %83,3 ile elde edilmiştir. Özellik azaltılarak yapılan sınıflandırma işleminin, sınıflandırma yöntemlerinin bireysel olarak uygulanmasından daha iyi sonuçlar verdiği gözlemlenmiştir.

Çizelge 2.1'de farklı dillerde n-gram modeli kullanılarak daha önce yapılmış olan yazar tanıma çalışmaları gösterilmektedir.

Çizelge 2.1 N-gram modeli kullanılarak yazar tanıma

| | **Dil** | **Sınıf Sayısı** | **N-gram** | **Başarı Oranı %** |
|---|---|---|---|---|
| Peng ve Keselj, 2003 | Yunanca | 10 | 3-gram | (Veri Seti-I) %85 |
| | | 10 | 4-gram | (Veri Seti-II) %97 |
| | Çince | 8 | 5-gram | %89 |
| | İngilizce | 8 | 6-gram | %100 |
| Fuchun, 2003 | Yunanca | 10 | 3-gram | (Veri Seti-I) %74 |
| | | 10 | 3-gram | (Veri Seti-II) %90 |
| | İngilizce | 8 | 6-gram | %98 |
| | Çince | 8 | 6-gram | %94 |
| Diri ve Amasyalı, 2006 | Türkçe | 18 | 2-gram | %83,3 |
| Ng_ind | Türkçe | 20 | 4-gram | %78 |

## 2.2 Tür Belirleme ile İlgili Daha Önce Yapılan Çalışmalar

Belirli bir konuda yazılmış dokümanların karakter sıklıkları birbirine ne kadar yakınsa, farklı konularda yazılmış dokümanların karakter sıklıkları da birbirinden o kadar uzaktır. Bu düşünceden yola çıkılarak farklı konularda yazılmış dokümanların konusuna göre sınıflandırılabildiğini göstermek için bir çok tür belirleme çalışması gerçekleştirilmiştir.

N-gram tabanlı daha önce yapılmış olan çalışmalarda başarılı sonuçlar elde edilmiştir. Bu yöntem dil sınıflama için denenmiş ve çok iyi sonuç alınmıştır. Sistem aynı zamanda konusuna göre farklı makaleleri sınıflandırmada da iyi bir şekilde çalışmıştır. %80'lik doğru sınıflama oranı gibi yüksek bir başarıya sahiptir (Cavnar,1994).

Dale ve Fuchun tarafından da, karakter seviyeli n-gram modeline dayalı tür belirleme çalışması yapılmıştır. Bu yöntem İngilizce ve Çince'den oluşan 3 veri seti üzerinde uygulanmıştır (Peng, Schuurmans, 2003). Bu yaklaşımda dokümanlar n-gram modeli ile yapılandırılmış ve 1'den 4'e kadar n-gram'lar çıkarılmıştır. Ancak frekans değerlerini azaltmak için bazı yumuşatma teknikleri kullanılmıştır. Naive Bayes sınıflandırma yöntemi kullanılarak tür belirme işlemi yapılmıştır.

İngilizce veri seti oluşturmak için 20 haber grubundan İngilizce dokümanlar seçilmiştir. Dokümanların rastgele %80'i seçilerek eğitim seti, %20'si seçilerek de test seti olarak kullanılmıştır. Yumuşatma tekniği uygulanarak en fazla %87'lik doğru sınıflandırma ile başarı elde edilmiştir. Çin veri seti, TREC-5 (Text Retrieval Conference Proceedings) veri setinden Çince dokümanlar alınarak oluşturulmuştur. TREC-5 veri setinde 101 grup içinden 6 grup seçilmiştir. 6 grubun her birinden 500 doküman eğitim seti, 100 doküman da test seti için alınmıştır. Çalışmada en fazla %81'lik doğru sınıflandırma ile başarı sağlanmıştır.

Stamatatos, Kokkinakis ve arkadaşları, yaptıkları tür belirleme çalışmasında, modern Greek weekly newspaper TO BHMA ismindeki gazetenin İnternet sitesinden çeşitli türlerde dokümanlar toplayarak Yunanca veri seti oluşturmuşlardır. Veri seti içinde 10 farklı konu ve her konu için 25 yazı bulunmaktadır (Stamatatos ve Kokkinakis, 2000). Bu çalışmada özellik vektörü oluşturabilmek için 4 yöntem kullanılmıştır.

- SCBD
- VR
- CWF-30
- CWF-50

Sınıflama için Ayrıştırma Çözümlenmesi ve Çoklu Regresyon yöntemleri kullanılmıştır. Veri setinde Ayrıştırma Çözümlenmesi için en fazla doğru olarak sınıflandırma oranı; SCBD yöntemi ile %82, VR yöntemi ile %59, CWF-50 yöntemi ile %79ve CWF-30 yöntemi ile %78olarak elde edilmiştir.

Veri setinde Çoklu Regresyon için en fazla doğru sınıflandırma başarısı; SCBD yöntemi ile %82, VR yöntemi ile %56, CWF-50 yöntemi ile %78 ve CWF-30 yöntemi ile de %78olarak elde edilmiştir.

Fuchun ve arkadaşları da, karakter seviyeli n-gram modeline dayalı tür belirleme çalışması yapmışlardır (Peng, Wang, 2003). Bu çalışmanın dilden bağımsızlığı ve etkisini kanıtlamak için Yunanca, İngilizce, Japonca ve Çince veriler üzerinde denemeler yapılmıştır. Her denemede başarılı sonuçlar elde edilmiştir. Bu yaklaşımda yazılar n-gram modeli ile yapılandırılmıştır ve 1'den 9'a kadar n-gram'ları çıkarılmıştır. N-gram karakterlerinin hepsi muhtemel özellik olarak alınmıştır. Bu modelde tüm özellikler kullanılarak özellik seçiminden kaçınılmıştır. Ancak frekans değerlerini azaltmak için bazı yumuşatma teknikleri kullanılmıştır. Bayesian sınıflandırma yöntemi kullanılarak tür belirme işlemi yapılmıştır.

Fuchun ve arkadaşları tarafından, Yunanca veri seti olarak Stamatatos ve Kokkinakis'in yaptığı çalışmada kullanılan aynı veri seti kullanılmıştır (Stamatatos ve Kokkinakis, 2000). Yunanca veri seti içinde 10 farklı tür ve her tür için 20 yazı bulunmaktadır. Her türe ait olan 20 yazının 10'u eğitim, 10 tanesi de test seti olarak kullanılmıştır. En fazla %86'lık doğru sınıflandırma ile başarı sağlanmıştır.

İngilizce veri seti olarak 20 haber grubu verisinden İngilizce dokümanlar seçilmiştir. Dokümanların %80'ni eğitim seti, %20'si de test seti olarak kullanılmıştır. Yumuşatma tekniği (Witten-Bell smoothing) uygulanarak ve 6-gram kullanılarak %88'lik, 3-gram kullanılarak %89'luk doğru sınıflandırma gerçekleştirilmiştir.

Çince veri seti, Çince için TREC-5 veri setinden dokümanlar alınarak oluşturulmuştur. Bu veri setinde 101 grup içinden 6 grup seçilmiştir. 6 grubun her birinden 500 doküman eğitim seti, 100 doküman da test seti için ayrılmıştır. Çalışmada 4-gram kullanılarak %81'lik doğru sınıflandırma ile başarı sağlanmıştır.

Japonca olan veri seti de, Japonca için NCIR-JI veri setinden dokümanlar alınarak oluşturulmuştur. Çalışmada %84'lük doğru sınıflandırma ile başarı sağlanmıştır.

Diri ve Amasyalı tarafından, n-gram yöntemi kullanılarak Türkçe dokümanların yazarlarını, türlerini ve cinsiyetini belirleme çalışması yapılmıştır (Diri ve Amasyalı, 2006). Veri seti İnternet'te belirli gazete sayfasından farklı yazarların politika, güncel, spor ve magazin gibi konularda yazıları toplanarak oluşturulmuştur. Eldeki mevcut veri seti 18 farklı yazarın 35'şer yazısı alınarak 630 doküman elde edilmiştir. Her tür için 210 yazının 168 tanesi eğitim, 42 tanesi de test için kullanılmıştır. Veri seti için 2-gram ve 3-gram'lardan oluşan özellik vektörleri ve bu özellik vektörlerinden türetilmiş özellik sayısı daha az olan vektörler kullanılarak sınıflandırma işlemi gerçekleştirilmiştir. Daha sonra NB, DVM, C 4.5 ve RF sınıflandırma yöntemleri kullanılarak her bir özellik vektörünün sınıflandırmadaki başarıları karşılaştırılmıştır. En başarılı sonuç DVM ile sınıflandırılmasında %93,6 ile elde edilmiştir. Özellik azaltılarak yapılan sınıflandırma işleminin, sınıflandırma yöntemlerinin bireysel olarak uygulanmasından daha iyi sonuçlar verdiği gözlemlenmiştir.

Çizelge 2.2'de farklı dillerde n-gram modeli kullanılarak daha önce yapılmış olan tür belirleme çalışmaları gösterilmektedir.

Çizelge 2.2 N-gram modeli kullanılarak tür belirleme

| | Dil | Sınıf Sayısı | N-gram | Başarı Oranı % |
|---|---|---|---|---|
| | Yunanca | 10 | 2-gram | %86 |
| | | 20 | 6-gram | |

1997). Başlangıçta ağırlık vektörü $w$, $w^+ = \{1,1,\ldots,1\}$ ve $w^- = \{-1, -1,\ldots, -1\}$ olarak tanımlanan 2 bileşenden oluşur. Her eğitilen x örneği için, eğer x bayan yazarsa c(x), 1 değerini alır, eğer x bay yazarsa c(x), 0 değerini alır. "$w$.x" skaler çarpım değeri 0'dan büyük ise s($w$,x), 1 değerini, aksi takdirde 0 değerini alır. Her bir örnek için (3)'deki formül kullandıktan sonra elde edilen ağırlıklar güncellenir ve bu işlem tüm eğitim örnekleri doğru olarak sınıflanana kadar devam ettirilir.

$$w_i^+ \leftarrow w_i^+ (1 + \beta X^i)^{(c(x)-s(w,x))}$$

$$w_i^- \leftarrow w_i^- (1 + \beta X^i)^{(s(w,x)-c(x))}$$

$$w = w^+ + w^- \qquad (3)$$

Bazı eşik değerlerinin aşağısında kalarak iptal edilen $w^+ + w^-$ elemanlarına 0 değeri atanır. Güncelleme kuralında değeri olmayan özelliklerin ağırlıkları 0 değerine eğilimlidir. Öğrenme algoritması özellik azaltıcı olarak kullanılmakta ve öğrenme eylemi tekrarlanarak düşük ağırlıktaki özellikler atılmış ve daha az sayıda özellik ile çalışılmıştır.

Kurgu türünde olan veri seti üzerinde sınıflama işlemi yapıldığında, %74'lük bir başarı elde edilmiştir. Kurgu türünde olmayan veri seti üzerinde sınıflama işlemi yapıldığında ise, %79.7'lik bir başarı sağlanmıştır.

Naive Bayes sınıflandırma yöntemi kullanılarak cinsiyet belirme işleminde kurgu türü olan veri seti üzerinde sınıflama işlemi yapıldığında, %64.4'lük bir başarı alınmış iken kurgu türünde olmayan veri seti üzerinde sınıflama yapıldığında, %60.9'luk bir başarı sağlanmıştır.

Ripper sınıflandırma yöntemi kullanılarak cinsiyet belirme işleminde kurgu türü olan veri seti üzerinde sınıflama işlemi yapıldığında, %64.9'luk bir başarı alınmış iken, kurgu türünde olmayan veri seti üzerinde sınıflama işlemi yapıldığında, %73.7'lik bir başarı sağlanmıştır.

Jonathan ve Keselj tarafından, yazarın cinsiyetini belirlemek için n-gram analizi yoluyla sınıflandırılma yapılmıştır [1]. İngiliz öğrencilerinin kompozisyonlarından oluşan bir veri seti kullanılmıştır. Denemelerde, karakter seviyeleri, kelime seviyeleri ve cümlenin öğelerinin n-gram'ları kullanılmıştır. Tüm yöntemler için aşağı yukarı aynı sonuçlar elde edilmiş ve en fazla %81'lik bir başarı sağlanmıştır.

Nowson ve Oberlander, n-gram yöntemini kullanarak cinsiyet belirleme yaparken, kullandıkları veri setinin n-gram'larını çıkarmışlar ve İngilizce veriler üzerinde deneyler yapmışlardır (Nowson ve Oberlander, 2006). Eldeki mevcut veri setleri 71 farklı yazarın yazıları kullanılarak oluşturulmuştur. 71 yazardan, 47 yazarın cinsiyeti bayan, 24 yazarın

cinsiyeti baydır. Her veri seti 8 bay ve 15 bayan yazardan oluşmuştur. Çalışmada dokümanları sınıflandırmak için "Weka" hazır sınıflandırma aracı içerisinde yer alan Destek Vektör Makinesi yöntemi kullanılmıştır. DVM ile sınıflandırılmasında %93'lük başarı elde edilmiştir.

Dupont , n-gram yöntemini kullanarak cinsiyet belirleme yaparken, kullandıkları veri setinin 2-gram'larını çıkarmış ve İngilizce veriler üzerinde deneyler yapmıştır (Dupont, 2006). Veri setinin rastgele %80'i seçilerek eğitim seti, %20'si seçilerek de test seti olarak kullanılmıştır. Çalışmada dokümanları sınıflandırmak için Naive Bayes yöntemi kullanılmıştır ve %85,76'lık başarı elde edilmiştir.

Diri ve Amasyalı tarafından, n-gram yöntemi kullanılarak Türkçe dokümanların yazarlarını, türlerini ve cinsiyetini belirleme çalışması yapılmıştır (Diri ve Amasyalı, 2006). Veri seti İnternet'te belirli gazete sayfasından farklı yazarların yazıları toplanarak oluşturulmuştur. Eldeki mevcut veri seti 18 farklı yazarın 35'şer yazısı alınarak 630 doküman elde edilmiştir. 18 yazardan, 4 yazarın cinsiyeti bayan, 14 yazarın cinsiyeti baydır. Bay yazarların 490 yazısının 392 tanesi eğitim, 98 tanesi de test için kullanılmıştır. Bayan yazarların 140 yazısının 112 tanesi eğitim, 28 tanesi de test için kullanılmıştır. Amacımız cinsiyet belirleme olduğu için bu çalışmada yazılar iki grupta sınıflandırılmıştır. Veri seti için 2-gram ve 3-gram'lardan oluşan özellik vektörleri ve bu özellik vektörlerinden türetilmiş özellik sayısı daha az olan vektörler kullanılarak sınıflandırma işlemi gerçekleştirilmiştir. Daha sonra NB, DVM, C 4.5 ve RF sınıflandırma yöntemleri kullanılarak her bir özellik vektörünün sınıflandırmadaki başarıları karşılaştırılmıştır. En başarılı sonuç DVM ile sınıflandırılmasında %96,3 ile elde edilmiştir. Özellik azaltılarak yapılan sınıflandırma işleminin, sınıflandırma yöntemlerinin bireysel olarak uygulanmasından daha iyi sonuçlar verdiği gözlemlenmiştir.

Çizelge 2.3'te farklı dillerde n-gram modeli kullanılarak daha önce yapılmış olan cinsiyet belirleme çalışmaları gösterilmektedir.

Çizelge 6.7 N-gram modeli kullanılarak cinsiyet belirleme

| | **Dil** | **N-gram** | **Başarı Oranı %** |
|---|---|---|---|
| (Dupont, 2006) | İngilizce | 2-gram | %85,76 |
| Diri ve Amasyalı, 2006 | Türkçe | 3-gram | %96,03 |

# 3. ÖZELLİK VEKTÖRLERİ

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, sayısını bizim belirlediğimiz *k* adet özelliğe sahiptir ve *k* boyutlu bir vektör ile ifade edilir. Dokümanların birer özellik vektörü şeklinde gösterimi en çok kullanılan metodlardan biridir (Salton, 1975).

## 3.1 N-Gram Model

N-gram, bir karakter katarının *n* adet karakter dilimidir. N-gram tabanlı sınıflandırma yöntemi, doküman içerisindeki n-gram karakterlerin kullanım sıklığına dayalı bir işlemdir. Bu çalışmada, n-gram'ın farklı birkaç uzunluğu olarak 2-gram, 3-gram 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 "Yüksek Lisans" ise, bu cümlenin n-gram'ları;

2-gram'lar: "Yü", "ük", "ks", "se", "ek", "k_", "_L", "Li", "is", "sa", "an", "ns"

3-gram'lar: "Yük", "üks", "kse", "sek", "ek_", "k_L", "_Li", "Lis", "isa", "san", "ans"

4-gram'lar: "Yüks", "ükse", "ksek", "sek ", "sek_", "ek_L", "k_Li", "_Lis", "Lisa", "isan", "sans"

şeklinde çıkarılır.

N-gram yöntemi, dokümanları sınıflandırmak için kullanılan basit ve güvenilebilir bir yöntemdir. Temel düşünce, bir doküman içerisindeki n-gram oluşumlarının tanımlanmasıdır. N-gram frekans yaklaşımı dilden bağımsız çalışır. Yani belirli bir dil hakkında detaylı bir dilbilgisine veya bir sözlük yapısına ihtiyaç yoktur.

Tüm harflerin veya hecelerin istatistiklerini kullanarak benzer sonuçlara ulaşmak mümkündür. Bununla beraber, bu düşünce ile birlikte çok kısa paragraflardan oluşan dokümanlar, konu tabanlı kelime istatistiğinde çok yetersiz kalmaktadır. Sonuç olarak eşleme için önemli olabilecek N-gram'ların yeterli şekilde toplanabilmesi için daha uzun paragraflara ihtiyaç vardır.

Sınıflandırma işlemi yazıdaki karakterlerin kullanım sıklığından yararlanılarak yapıldığından, konu türü spor olan bir yazıda, bir kelimenin ilgili formları için ("futbol", "futbolcu",

"futbolu", "futboldan") elde ettiğimiz n-gram'ların sıklığı ile sınıflandırma işlemini kolayca yapabiliriz.

Doğal diller, kullanım sıklıkları diğerlerinden daha fazla olan değişmez bazı kelimelere sahiptir. Bu varsayım tüm diller için geçerlidir. Bu düşünceyi ifade etmenin en genel yollarından biri Zipf'in kanunları olarak bilinir. "The $n^{th}$ most common word in a human language text occurs with a frequency inversely proportional to n." (George, 1949). Bu kanunun anlamı şudur, bir dilde bazı kelimeler diğerlerinden daha fazla kullanılır. Belirli bir konuya özel olan kelimeler için bu doğru bir yaklaşımdır. Bu yüzden önemli frekans değerlerine sahip n-gram'ları diğerlerinden ayırmak için bir eşik değeri (threshold ) kullanırız.

Aynı kategoriden gelen dokümanları karşılaştırdığımızda dokümanların benzer n-gram frekanslarına sahip olduğunu görürüz. Bu çalışmada veri setlerinde bulunan her bir dokümanın 2-gram, 3-gram ve 4-gram'ları elde edilmiştir. 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 değerler alınmıştır.

## 3.2 Özellik Vektörünün Çıkarılması

Dokümana özgü olan özelliklerin çıkarılmasında *Genel n-gram Tablosu*, *Ana n-gram Tablosu* ve *Sonuç n-gram Tablosu* olarak isimlendirilen üç tablo kullanılmıştır.

Her bir veri setinde bulunan dokümanların 2-gram'ları, 3-gram'ları ve 4-gram'larından *Genel n-gram Tabloları* elde edilmektedir. Kısaca 3 farklı n-gram modeli kullanıldığından *n* adet doküman için *nx3* adet *Genel n-gram Tablosu* oluşturulmuştur.Bu tablolarda n-gram'lar kullanım frekans değerleri büyükten küçüğe doğru sıralanarak tutulmuştur.

Her veri seti için, veri setindeki dokümanların tümünü ifade eden ortak bir n-gram tablosu oluşturulmalıdır. Bu ortak n-gram tablosunu *Ana n-gram Tablosu* olarak isimlendiririz. Ana n-gram tabloları, her veri setimize özgü olarak ortak 2-gram, 3-gram ve 4 gram'ların çıkarılması ile oluşturulmuştur. Bu tablonun oluşumunda veri setinde bulunan her yazarın 5 yazısı alınmıştır. Örneğin cinsiyet belirleme işlemini ele alırsak, kullanacağımız veri setinde 20 yazar bulunmaktadır. Her bir yazarın 5 yazısını kullanacağımızdan toplam 100 yazının hepsi için tek ortak 2-gram, 3-gram ve 4-gram tabloları oluşturulmuştur. Cinsiyet belirmede olduğu gibi Ana n-gram tablosunu oluşturmak için yazar tanıma işleminde 100 doküman, tür tanıma işleminde de 60 doküman kullanılmıştır. Çalışmalarımızda her veri setimiz için 3 ana tablo oluşturulmuştur. Ana tablolarda n-gram'lar kullanım frekansı büyükten küçüğe doğru

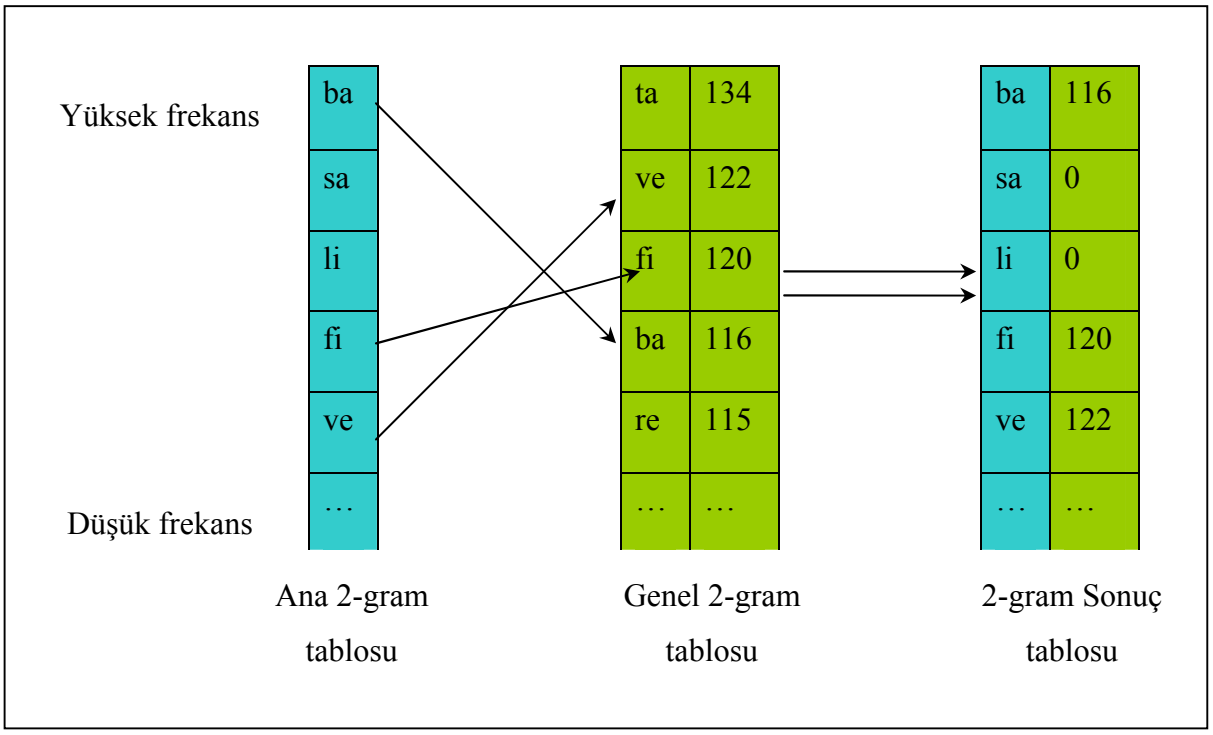
sıralanmıştır. En fazla sıklıkla kullanılan n-gram'ların dokümanları sınıflandırmada büyük etkisi olduğu düşünülerek, ana tablolarda kullanım sıklığı 100'den büyük olan değerler alınmıştır. Çizelge 3.1'de bu çalışmada kullanılan dokümanların, frekans değeri 100'den büyük olan Ana 2-gram, 3-gram ve 4-gram tablolarında bulunan n-gram'ların sayısı gösterilmektedir.

Çizelge 3.1 Özellik Frekans değeri 100'den büyük olan n-gram sayısı

| | **Ana 2-gram tablosu** | **Ana 3-gram tablosu** | **Ana 4-gram tablosu** |
|---|---|---|---|
| **Cinsiyet Belirleme** | 257 | 324 | 142 |
| **Yazar Tanıma** | 257 | 324 | 142 |
| **Tür Tanıma** | 217 | 208 | 75 |

Özellik vektörünün oluşturulması için n-gram sonuç tablolarındaki değerler gereklidir. *N-gram sonuç tablosu*, Genel n-gram tablosu ile Ana n-gram tablosunun karşılaştırılmasıyla elde edilmiştir. Ana tablo ile Sonuç tablolarının boyutları ve bu tablolarda bulunan n-gram'lar ile sıraları birebir aynıdır. Fakat sonuç tablosundaki n-gram'ların frekans değerleri genel tablolardan alınmıştır.

Şekil 3.1'de veri setlerindeki her bir dokümanın Genel 2-gram tablosu ile o veri setinin Ana 2-gram tablosunun karşılaştırılması ile, o yazı için oluşturulan 2-gram sonuç tablolarının elde edilişi gösterilmektedir. Frekans değerleri büyükten küçüğe sıralanmış olan bu tablolar karşılaştırılırken Ana 2-gram tablosundaki n-gram'ların sırası önem taşımaktadır. Ana tabloda bulunan n-gram'ların sırasıyla Genel n-gram tablosunda var olup olmadığı kontrol edilir. Eğer n-gram'lar var ise sonuç tablosuna, genel tablodaki frekans değerleri yazılır. Eğer n-gram'lar yok ise sonuç tablosuna "0" değeri yazılır. Örneğin, Ana 2-gram tablosunun ilk elemanı "ba" 2-gram'ı Genel tablosunda mevcut ve frekans değeri 116 ise, sonuç tablosunun ilk elemanı "ba" 2-gram'ının frekans değerine de 116 yazılır. Ana 2-gram tablosunun ikinci elemanı "sa" 2-gram'ı genel tablosunda mevcut değilse, o zaman sonuç tablosunun ikinci elemanı "sa" 2-gram'ının frekans değerine "0" yazılır. Bu şekilde ana tablo ve genel tablolar karşılaştırılarak sonuç tablosu elde edilir.




Şekil 3.1 2-gram sonuç tablosunun oluşturulması

Bu işlemler çalışmalarımızda kullanılan üç veri setindeki tüm dokümanlar için gerçekleştirilir ve tüm dokümanların 2-gram, 3-gram ve 4-gram sonuç tabloları elde edilir. Çizelge 3.2'de bu çalışmada elde edilen n-gram sonuç tablolarındaki eleman sayısı verilmektedir. Sonuç tablosunun boyutu ana tablonun boyutu ile aynıdır ve bu tabloda bulunan her bir frekans değeri özellik vektörünün bir özelliğini oluşturur. Özellik vektöründe bulunan özelliklerin sayısı sonuç tablosunun büyüklüğü ile aynıdır.

Çizelge 3.2 Elde edilen sonuç tablo sayısı

| | Sonuç tablosundaki n-gram sayısı | | |
|---|---|---|---|
| | **2-gram** | **3-gram** | **4-gram** |
| **Cinsiyet Belirleme (Veri Seti - I)** | 800 | 800 | 800 |
| **Yazar Tanıma (Veri Seti - II)** | 800 | 800 | 800 |
| **Tür Tanıma (Veri Seti - III)** | 480 | 480 | 480 |

Sınıflandırma işlemi özellik vektörleri kullanılarak gerçekleştirilmiştir. Özellik vektörü oluşturulurken şu adımlar izlenmiştir;

- Her veri setinde bulunan yazıların genel 2-gram, 3-gram ve 4-gram tabloları çıkarılmıştır. Bu tablolarda n-gram'ların frekans değerleri büyükten küçüğe doğru sıralanmıştır.
- Her veri seti için Ana 2-gram, 3-gram ve 4-gram tabloları oluşturulmuştur. Bu tablolarda da n-gram'lar frekans değerleri büyükten küçüğe doğru sıralanmıştır.
- Bir veri setinde yer alan her bir dokümanın genel tablosu ile, o veri setinin ana tablosu karşılaştırılıp sonuç tabloları elde edilmiştir. Her veri seti için bir 2-gram, 3-gram ve 4-gram sonuç tablosu oluşturulmuştur. Çizelge 3.2'de bu çalışmada elde edilen sonuç tablosundaki n-gram sayısı verilmektedir.
- Her bir sonuç tablosunda bulunan değerler, özellik vektörlerini oluşturan özellikleri temsil ederler. Özellik vektöründe bulunan özelliklerin sayısı sonuç tablosunun uzunluğu ile aynıdır.

Bu çalışmada her veri seti için 2-gram, 3-gram ve 4-gram'lardan oluşan özellik vektörleri çıkarılmıştır.

### 3.2.1 2-gram özellik vektörü

Veri Seti - I içindeki her bir 2-gram özellik vektörünün uzunluğu 257'dir. Yani özellik vektörleri 257 adet frekans değerine sahiptir ve bu değerler 2-gram özellik vektörünün birer özelliğini temsil etmektedir. Veri Seti - II içindeki her bir 2-gram özellik vektörünün uzunluğu 257 ve Veri Seti - III içindeki her bir 2-gram özellik vektörünün uzunluğu da 217'dir. Çalışmada bu özellikler kullanılarak oluşturulan vektörden $\boldsymbol{OV_2}$ olarak bahsedilecektir.

### 3.2.2 3-gram özellik vektörü

Veri Seti - I içindeki her bir 3-gram özellik vektörünün uzunluğu 324, Veri Seti - II içindeki her bir 3-gram özellik vektörünün uzunluğu 324 ve Veri Seti - III içindeki her bir 3-gram özellik vektörünün uzunluğu da 208'dir. Çalışmada bu özellikler kullanılarak oluşturulan vektörden $\boldsymbol{OV_3}$ olarak bahsedilecektir.

### 3.2.3 4-gram özellik vektörü

Veri Seti - I içindeki her bir 4-gram özellik vektörünün uzunluğu 142, Veri Seti - II içindeki her bir 4-gram özellik vektörünün uzunluğu 142 ve Veri Seti - III içindeki her bir 4-gram

özellik vektörünün uzunluğu da 75'tir. Çalışmada bu özellikler kullanılarak oluşturulan vektörden $OV_4$ olarak bahsedilecektir.

### 3.3 Özellik Azaltma Yoluyla Oluşturulan Vektörler

Her bir dokümanın özel bir şekilde ifade edilmesi amacıyla çıkarılmış özeliklerden bazıları dokümanları ayırt edici nitelik taşımamaktadır. Ayırt edici nitelikte olmayan özelliklere sahip özellik vektörleri kullanıldığı durumda sınıflandırma başarısı düşebilir. Gereksiz özelliklerin çıkarılıp sistemin başarısını tekrar ölçmek amacıyla özellik azaltma işlemi yapılmıştır. Bu çalışmada Weka Tool'u [2] içerisinde bulunan CfsSubsetEval fonksiyonu özellik azaltma amacıyla kullanılmıştır.

Çizelge 3.3'te bu çalışmada kullanılan özellik vektörlerinin özellik sayıları gösterilmektedir. Her bir veri seti için elde edilen 2-gram, 3-gram ve 4-gram özellik vektörlerinin özellik sayıları azaltılarak yeni özellik vektörleri elde edilmiştir. Çalışmada 2-gram özellik vektörü azaltılarak elde edilen vektörden $OV_{2A}$, 3-gram özellik vektörü azaltılarak elde edilen vektörden $OV_{3A}$, 4-gram özellik vektörü azaltılarak elde edilen vektörden $OV_{4A}$ olarak bahsedilecektir.

Çizelge 3.3 Özellik vektörlerinde bulunan özelliklerin sayısı

| | **Özellik Vektöründe bulunan özellik sayısı** | | | | | |
|---|---|---|---|---|---|---|
| | | | | Özelliği azaltılmış | | |
| | $OV_2$ | $OV_3$ | $OV_4$ | $OV_{2A}$ | $OV_{3A}$ | $OV_{4A}$ |
| **Cinsiyet Belirleme** (Veri Seti - I) | 257 | 324 | 142 | 10 | 27 | 12 |
| **Yazar Tanıma** (Veri Seti - II) | 257 | 324 | 142 | 12 | 14 | 6 |
| **Tür Tanıma** (Veri Seti - III) | 217 | 208 | 75 | 15 | 21 | 15 |

Her bir veri seti için, üç özellik vektörü ve bunlardan türetilmiş özellik sayısı azaltılmış vektörler kullanılarak sınıflandırma işlemi gerçekleştirilmiştir. Bu çalışmada Naive Bayes (NB), Rasgele Orman (RO), Destek Vektör Makinesi (DVM), K-En Yakın Komşuluk (K-EK)ve ng_ind gibi sınıflandırma yöntemleri kullanılarak yazar tanıma, tür belirleme ve cinsiyet tahmin etme çalışması yapılmıştır.

## 3.4 Veri Setleri

Bu çalışma kapsamında, dokümanların yazarlarını, türlerini ve yazarların cinsiyetini belirlemede kullanılmak amacıyla üç farklı veri seti oluşturulmuştur. Veri setleri İnternet'te belirli gazete sayfalarından Hürriyet [3], Milliyet [4], Akşam [5] ve Sabah'tan [6] 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.

### 3.4.1 Veri Seti - I

Veri Seti – I, 20 yazardan oluşmaktadır. 20 yazardan, 10 yazarın cinsiyeti bayan, 10 yazarın cinsiyeti baydır. Bay ve bayan yazarların 400'er yazısı olmak üzere toplam 800 doküman vardır. Amacımız cinsiyet belirleme olduğu için bu çalışmada yazılar iki grupta sınıflandırılmıştır. Bu sınıflardan biri bayan diğeri bay yazar sınıfını temsil etmektedir.

### 3.4.2 Veri Seti - II

Veri Seti – II, 20 yazardan oluşmaktadır. Bu yazarlara ait yazıların konuları birbirlerine yakın ya da farklı olabilir. Bu veri setinde her bir yazarın 40 adet yazısı olmak üzere toplam 800 dokümanı vardır. Amacımız yazar tanıma olduğu için dokümanlar yirmi farklı grupta sınıflandırılmıştır. Bu sınıflardan her biri bir yazarı temsil etmektedir. Çizelge 3.4'te Veri Seti II'yi oluşturan yazarlar verilmiştir.

Çizelge 3.4 Veri Seti – II 'de dokümanları kullanılan bay ve bayan yazarlar

| Kullanılan Doküman Sayısı : 40 (her yazar için) | | | |
|---|---|---|---|
| **Yazarlar** (politika, magazin, güncel, ekonomi**)** | | | |
| 1 | Ayşe Arman (magazin**)** | 11 | Can Dündar (güncel) |
| 2 | Ebru Çapa (güncel) | 12 | Emin Çölaşan (güncel) |
| 3 | Pakize Suda (magazin) | 13 | Hadi Uluengin (güncel) |
| 4 | Banu Tuna (güncel) | 14 | Taha Akyol (politika) |
| 5 | Tuğba Akyol (güncel) | 15 | Çetin Altan (güncel) |
| 6 | Meral Tamer (güncel) | 16 | Melih Aşık (güncel) |
| 7 | Serpil Yılmaz (ekonomi) | 17 | Abbas Güçlü (güncel) |
| 8 | Ferai Tinç (politika) | 18 | Fikret Bila (politika) |
| 9 | Gila Benmayor (güncel) | 19 | Hasan Cemal (politika) |
| 10 | Figen Batur (güncel) | 20 | Güneri Civaoğlu (politika) |

### 3.4.3 Veri Seti - III

Veri Seti – III, 12 yazardan oluşmaktadır. Bu veri setinde her bir yazarın 40 adet yazısı olmak üzere toplam 480 yazı vardır. Tür belirleme çalışmasında spor, magazin, güncel, ekonomi, sağlık ve politika olarak toplam 6 tane kategori belirlenmiştir. Her kategoride 2 yazarın yazısı bulunmaktadır. Üçüncü veri setini tür belirleme için kullandığımızdan dolayı bu çalışmada altı sınıf mevcuttur. Bu sınıflardan her biri bir konuyu temsil etmektedir.

# 4. SINIFLANDIRMA YÖNTEMLERİ

Bu çalışmada dokümanları sınıflandırmak için "Weka" hazır sınıflandırma aracı içerisinde yer alan Naive Bayes, Destek Vektör Makinesi, K-En Yakın Komşuluk, Rastgele Orman yöntemleri ve bizim geliştirdiğimiz ng_ind yöntemi kullanılmıştır.

## 4.1 Naive Bayes

Naive Bayes yöntemi doküman sınıflandırma işlemlerinde en sık kullanılan, pratik, olasılığa dayanan bir sınıflayıcıdır. Diğer bütün sınıflandırıcılarla karşılaştırıldıklarında en düşük hata oranına sahiptirler. Elimizde n adet sınıf olduğunu farz edelim, $S_1,S_2,\ldots,S_n$. Herhangi bir sınıfa ait olmayan bir veri örneği *X*'in, hangi sınıfa ait olduğu Naive Bayes sınıflandırıcı tarafından belirlenir. Veri örneği *X*, verilen sınıflara ait olma olasılığı en yüksek değere sahip sınıfa atanır. Sonuç olarak, Naive Bayes sınıflandırıcı bilinmeyen örnek *X*'i, $S_i$ sınıfına atar. Her veri örneği, m boyutlu özellik vektörleri ile gösterilir, $X = (X_1,X_2,\ldots,X_m)$. Özelliklerin hepsi aynı derecede önemlidir ve birbirinden bağımsızdır. Bir özelliğin değeri başka bir özellik değeri hakkında bilgi içermez. $X = (X_1,X_2,\ldots,X_m)$ örneğinin $S_i$ sınıfında olma olasılığı (3)'teki gibidir.

$$P(S_i \mid X) = \frac{P(X \mid S_i)P(S_i)}{P(X)} \tag{3}$$

P(X) bütün sınıflar için sabit ise, *X* örneğinin $S_i$ sınıfında olma olasılığına, $P(X|S_i)\ P(S_i)$ ifadesi ile ulaşabiliriz. $P(S_i)$, her bir sınıfın olasılığı olup (4)'teki gibidir.

$$P(S_i) = \frac{O_i}{O} \tag{4}$$

Burada $O_i$, $S_i$ sınıfına ait eğitilen örnek sayısı, ve *O* ise toplam eğitilen örnek sayısıdır. Eğer sınıf öncelik olasılığı bilinmiyorsa, o zaman genel olarak sınıflar eşit kabul edilir, $P(S_1) = P(S_2) = \ldots\ldots = P(S_n)$, ve bu sebeple $P(X|S_i)$ ifadesi, *X* örneğinin $S_i$ sınıfında olma olasılığını bulmak için kullanılır. Aksi taktirde, $P(X|S_i)\ P(S_i)$ ifadesi bizim için en anlamlı ifadedir. Olasılıklar $P(X_1|S_i)$, $P(X_2|S_i),\ldots,$ $P(X_m|S_i)$ eğitim örneklerinde tahmin edilebilirler,

$$P(X_k \mid S_i) = \frac{O_{ik}}{O_i} \tag{5}$$

burada $O_{ik}$, $x_k$ değerine sahip olan $S_i$ sınıfına ait eğitim seti sayısı, ve $O_i$'*de* $S_i$'ye ait olan eğitim seti sayısıdır.

Bilinmeyen örnek *X*'i sınıflandırmak için, her $S_i$ sınıfı $P(X|S_i)\ P(S_i)$ ifadesi hesaplanır. Örnek X'i en yüksek değere sahip $S_i$ sınıfına atanır (6).

$$P(X \mid S_i) = \prod_{k=1}^{m} P(X_k \mid S_i) \qquad (6)$$

Çizelge 4.1'deki eğitim verisinden yararlanarak bilinmeyen bir X örneğinin hangi sınıfa ait olduğunu Naive Bayes sınıflandırma kullanılarak tahmin etmek isteyelim. Eğitim verisi çizelge 4.1'deki veri olup, örnekleri *vites, renk, yakıt,* ve *kapı* özellikleri ile tanımlanır. Elimizde A ve B olmak üzere 2 sınıf mevcuttur. Sınıflandırmak istediğimiz bilinmeyen örnek, X = (vites = "otomatik", renk = "gri", yakıt = "benzinli", kapı = "4") olsun.

Çizelge 4.1 A ve B eğitim seti

| **Sınıf** | | **Vites** | **Renk** | **Yakıt** | **Kapı** |
|---|---|---|---|---|---|
| 1 | A | Otomatik | Gri | Dizel | 2 |
| 2 | B | Normal | Gri | Benzinli | 4 |
| 3 | A | Otomatik | Kırmızı | Benzinli | 2 |
| 4 | A | Otomatik | Gri | Benzinli | 4 |
| 5 | A | Otomatik | Beyaz | Dizel | 2 |
| 6 | B | Otomatik | Kırmızı | Benzinli | 4 |
| 7 | A | Normal | Gri | Dizel | 4 |
| 8 | B | Otomatik | Gri | Benzinli | 4 |
| 9 | A | Normal | Beyaz | Benzinli | 4 |
| 10 | A | Otomatik | Kırmızı | Benzinli | 4 |
| 11 | B | Normal | Kırmızı | Dizel | 4 |
| 12 | A | Normal | Beyaz | Dizel | 4 |
| 13 | A | Normal | Gri | Benzinli | 2 |
| 14 | B | Otomatik | Beyaz | Benzinli | 4 |
| 15 | B | Otomatik | Beyaz | Benzinli | 4 |

*X* bilinmeyen verisinin hangi sınıfa ait olduğunu bulabilmek için $P(X|S_i)P(S_i)$ değerini maksimize etmemiz gerekmektedir. A sınıfı 9 elemandan, B sınıfı da 6 elemandan oluşmuş olsun. $P(S_i)$ için, her sınıf için olasılık değerleri eğitim örneklerinden hesaplanabilir:

P(A) = 9/15 =0.600

P(B) = 6/15 =0.400

$P(X|S_i)$, i =1,2 değerlerinin koşullu olasılıkları hesaplarsak,

P(vites = "otomatik" | *A)* = 5/9 = 0.555

P(renk = "gri" | *A)* = 4/9 = 0.444

P(yakıt = “benzinli” | *A)* = 5/9 = 0.555

P(kapı = “4” | *A)* = 5/9 = 0.555

P(vites = “otomatik” | *B)* = 4/6 = 0.667

P(renk = “Gri” | *B)* = 4/6 = 0.667

P(yakıt = “benzinli” | *B)* = 5/6 = 0.833

P(kapı = “4” | *B)* = 6/6 = 1

elde ederiz. Hesaplanan bu olasılık değerleri kullanılarak

P(X|A) = 0.555*0.444*0.555*0.555 = 0.076

P(X|B) = 0. 667*0. 667*0. 833*1 = 0.370

P(A) P(X|A) = 0.600*0.076= 0.046

P(B) P(X|B) = 0.400*0.370= 0.148

sonuçları elde edilir. Elde edilen sonuca göre *X* arabası en yüksek olasılık değerine sahip olan A sınıfına aittir.

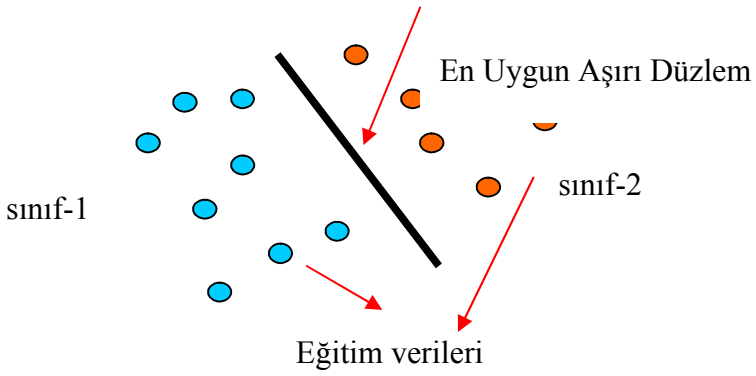
Özgür tarafından yapılan çalışmada, otomatik belge sınıflandırma için gözetimli ve gözetimsiz temel yöntemler ele alınmıştır (Özgür, 2002). Bu temel yöntemler, beş standart veri seti üzerindeki başarıları farklı kıstaslara dayanarak incelenmiş ve Destek Vektör Makinesi, Naive Bayes, K-En Yakın Komşuluk, K-Means ve Açıortay k-Means’in (bisecting k-means) öğrenme yaklaşımları birbiriyle kıyaslanmıştır. Bu çalışma sonucunda bu yöntemler içinde K-Means ve Açıortay K-Means’in belge sınıflanması için daha elverişli olduğunu görülmüştür. Üzerinde çalışılan veri setleri Classic3 3,891 dokümandan ve 3 sınıftan, Hitech 1,530 doküman ve 6 sınıftan, LA1 2,134 dokümandan ve 6 sınıftan, Reuters-21578 12,902 dokümandan ve 90 sınıftan, Wap 1,560 dokümandan ve 20 sınıftan oluşmaktadır. Naive Bayes ile yapılan sınıflandırma işleminde Classic3 veri seti için %98.7’lik, Hitech veri seti için %53.3’lük, LA1 veri seti için %57.4’lük, Reuters-21578 veri seti için %62.5’lik, Wap veri seti için %62.3’lük doğru sınıflandırma başarısı sağlanmıştır.

Ko ve Seo tarafından, özelik izdüşüm tekniğine (Text Categorization using Feature Projections (TCFP)) dayalı farklı bir doküman sınıflandırma yöntemi geliştirilmiştir (Ko, Seo, 2002). Önce, eğitim dokümanlarının izdüşümü olarak özellikler belirlenmiş, daha sonra bireysel özellik izdüşümleri temelinde, sınıflama için oylama işlemi gerçekleştirilmiştir. Test dokümanın asıl sınıfı, her özelliğin bireysel sınıflamasından elde edilen oy çoğunluğuna göre

belirlenmiştir. Sınıflandırma için K-NN, Rocchio ve Naive Bayes yöntemleri kullanılmıştır. Eğitim setinde16,000 doküman ve 10,000 özelik kullanılmıştır. Naive Bayes ile yapılan sınıflandırma işleminde %82.5'lik doğru sınıflandırma başarısı sağlanmıştır.

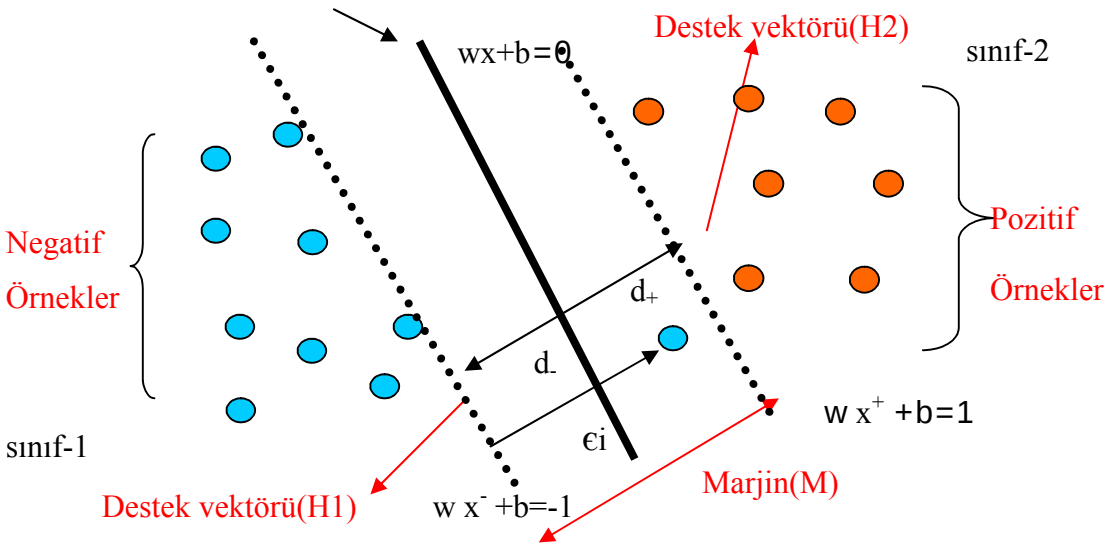
## 4.2 Destek Vektör Makinesi (DVM)

Destek Vektör Makinesi, makine öğrenmesi yöntemlerinden biri olup V.Vapnik tarafından ortaya atılmıştır. Çekirdek tabanlı doğrusal olmayan sınıflandırıcıların sinyal işleme, yapay öğrenme ve veri madenciliği alanındaki pratik problemlerde iyi sonuçlar verdiği kanıtlanmıştır (Vapnik, 1999), (Weston, 1999). Temelde DVM iki sınıflı problemlerle ilgilenir. Girişte alınan veriler destek vektörler ile tanımlanabilen bir aşırıdüzlem (hyperplane) tarafından şekil 4.1'de gösterildiği gibi ikiye ayrılır. Amaç 2 sınıfı birbirinden ayırabilecek en uygun aşırıdüzlemi bulabilmektir.




Şekil 4.1 Gelen verileri ayıran aşırıdüzlem

En uygun aşırıdüzlemi bulabilmek için, her iki sınıfın en uygun aşırıdüzlemine en yakın veri noktalarından geçen aşırıdüzlemler çizilir ve bu iki düzlem birbirine paraleldir. Bu düzlemler arasındaki mesafe en uygun aşırıdüzleminin kalitesini belirler. DVM iki sınıf arasındaki sınırı ayırt etme yüzeyini belirlemekte, yani eğitim kümesi ile ayırt etme yüzeyine en yakın noktaların arasındaki mesafeyi maksimumlaştırmaktadır. Destek vektörleri, aşırıdüzlemler ve marjin kavramları şekil 4.2'de gösterilmiştir.




Şekil 4.2 Doğrusal ayrılabilir veriler üzerinde destek vektörleri

İki sınıfın örneklerini birbirinden ayıran bir aşırıdüzlem vardır, bu düzlem üzerindeki noktalar w.x+b=0 eşitliğini sağlayacaktır, burada w aşırıdüzleme olan normal ve |b|/||w|| aşırıdüzlemden orijine olan dik uzaklıktır. Destek vektör yöntemi aşırıdüzleme en yakın pozitif ve negatif örnekler arasındaki mesafenin (marjin genişliğinin) en yüksek olduğu bir aşırıdüzlem bulmaya çalışır. Marjin (M) genişliği (7)'deki gibidir.

$$\left.\begin{array}{l} w \,.\, x^{+} + b = +1 \\ w \,.\, x^{-} + b = -1 \end{array}\right\} \quad w.\,(x^{+}-x^{-}) = 2$$

$$M = \frac{(x^{+} - x^{-}) \cdot w}{|w|} = \frac{2}{|w|} \tag{7}$$

"w" değeri ne kadar küçülürse marjin genişliği o kadar artar. Her nokta $x_i$ olarak gösterilir. İki sınıfa ait eğitim verisi $x = (x_1,x_2,\ldots,x_n)$ ve etiket değeri $y = (-1, +1)$ ise en uygun aşırıdüzlem (8)'deki formülle ifade edilir;

$$\left.\begin{array}{l} \text{eğer } y = +1 \text{ ise } w \,.\, x_i + b >= 1 \\ \text{eğer } y = -1 \text{ ise } w \,.\, x_i + b < 1 \end{array}\right\} \quad \text{tüm i'ler için } y_i(w \,.\, x_i + b) >= 1\text{'dir} \tag{8}$$

Verilen *X* örneğini sınıflandırmak için öncelikle en uygun aşırıdüzlem bulunur. Bu aşırıdüzlem taraflarından biri negatif sınıfı, diğeri ise pozitif sınıfı temsil eder. *X* örneği DVM

yöntemi ile formüle edilir ve eğer f(x) fonksiyonu sıfırdan büyük çıkarsa pozitif sınıfa, negatif çıkarsa negatif sınıfa atanır.

f(x) = sign(w x + b)

f(x) >= 0 pozitif sınıfı temsil etmektedir.

f(x) < 0 negatif sınıfı temsil etmektedir.

Çoklu-sınıf verilerinin DVM ile sınıflandırılması için Kneer ve arkadaşları bire-karşı-bir (one-against-one) yöntemini önermişlerdir (Kneer, 1990). 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 iki sınıf problemine çevrilmiş olur. Ayrıca eğitim için her sınıflandırıcıda sadece iki sınıfa ait verilerin kullanılması toplam eğitim zamanını azaltacaktır.

Çoklu-sınıf verilerinin DVM ile sınıflandırılması için bir diğer yöntem ise bire-karşı-hepsi (one-against-all) adı verilen yöntemdir (Platt, 2000). 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. Dikkat edilirse eğitim zamanı, eğitim için fazla sayıda örüntünün kullanıldığı sınıflandırma problemleri için, bire-karşı-bir (Müller, Mika, 2001) yöntemine göre oldukça büyük olacaktır.

DVM yöntemi cinsiyet belirleme, yüz tanıma, karakter tanıma vb. çalışmalarda kullanılmıştır. Kepenekçi ve Akar, Gabor dalgacık dönüşümü kullanılarak bulunan yüz öznitelik vektörlerinin destekçi vektör makinesi ile sınıflandırılmasına dayalı bir yüz tanıma çalışması yapmışlardır (Kepenekçi, Akar, 2005). Her yüz için bire karşı geri kalan olmak üzere iki sınıflı destekçi vektör makinesi eğitilmiştir. Gabor öznitelikleri kullanılarak doğrudan görüntü bilgisi kullanımına oranla destekçi vektör sayısı azaltılmıştır, ayrıca daha iyi genelleme performansı sağlanmıştır. Önerilen yöntemin performansı ORL yüz veritabanı kullanılarak sınanmıştır [7]. Önerilen yöntem ile ORL yüz veritabanında %98.75 doğru tanıma başarısı elde edilmiştir.

Basu ve arkadaşları, DVM yöntemini kullanarak dokümanları sınıflandırma işlemini yapmışlardır. Bu çalışmada amaç, yeni bir dokümanı otomatik olarak sınıflandırmaktır. Otomatik sınıflandırma, özellik setlerini azaltmak için bir kelime filtresi kullanılarak geliştirilmiştir. Yapay sinir ağları ve DVM algoritması doküman sınıflandırıcı olarak kullanılması için karşılaştırılmıştır. Yüksek sonuca ulaşılmak için özellik azaltıcı

tanımlanmıştır. Bu çalışmada DVM ile sınıflandırılması için bire-karşı-bir (one-against-one) yöntemini önermişlerdir. Bu yöntemde 64 adet sınıf için 64(64−1)/2 = 2016 sınıflandırıcı oluşturulmuştur. Önerilen yöntem ile Reuters-21578 veritabanı kullanılarak %82'lik doğru tanıma başarısı elde edilmiştir (Basu, Watters, 2002).

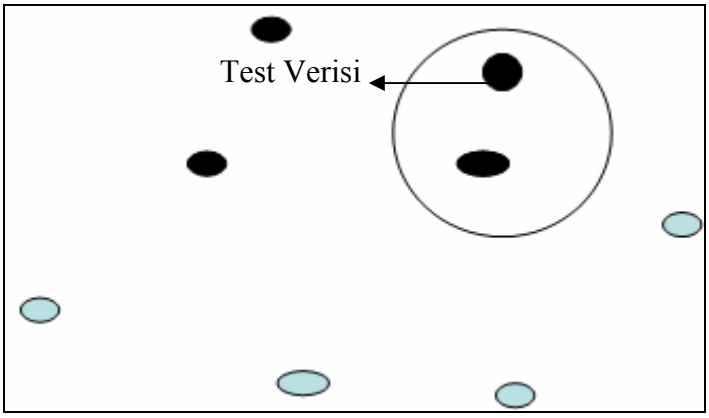
Kısaca DVM, doğrusal olmayan bir şekilde ayrılabilen öbekler için optimal aşırıdüzlem bulmaya çalışır. Bu yüzden DVM'nin VM'deki uygulamaları özellikle sınıflama tekniğinde ortaya çıkmıştır. Elde edilen sonuçlar bu yöntemin sınıflama tekniğinde oldukça başarılı olduğunu göstermiştir (Fung ve Mangasarian, 2002).

### 4.3 K-En Yakın Komşuluk (K-EK)

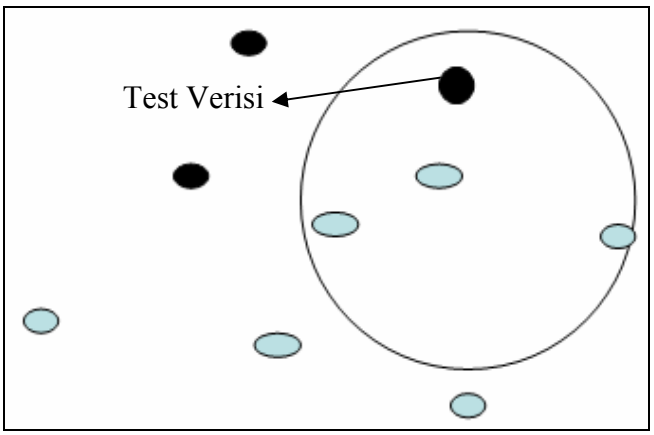
Sınıflandırılmak istenen örneğe en yakın örnekleri bulan eğiticili bir yöntemdir. En yakın komşu sınıflandırma yöntemi, sınıflandırılmak istenen örnek ile sınıflarda bulunan örnekler arasındaki benzerliğe dayalıdır. Eğitim örnekleri n-boyutlu sayısal niteliklerle tanımlanır. Her örnek n-boyutlu uzayda bir noktayı gösterir. Bu yolla, bütün eğitim örnekleri n-boyutlu örnek uzayda depolanır. Bu k-eğitim örnekleri, bilinmeyen örnek A'ya k en yakın komşudur. Yakınlık, Öklid (Euclidean Distance) uzaklığı hesaplanarak bulunur. X=( $X_1$,$X_2$,…,$X_n$) ve Y=( $Y_1$,$Y_2$,…,$Y_n$) arasındaki Öklid uzaklığı (9) ile hesaplanır.

$$D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (9)$$

Bilinmeyen örnek A, en yakın k-komşularının arasında en sık bulunan sınıfa atanır. Şekil 4.3'te, k değerinin 1, şekil 4.4'te k değerinin 2 olduğu duruma örnek gösterilmektedir. k değerinin 1 olduğu durumda, bilinmeyen örnek, örnek uzayında en yakın eğitim örneğinin sınıfına aittir.



Şekil 4.3 1-En Yakın Komşuluk




Şekil 4.4 3-En Yakın Komşuluk

Çizelge 4.2'deki eğitim verisinden yararlanarak bilinmeyen bir *X* örneğinin hangi sınıfa ait olduğunu k-En Yakın Komşuluk sınıflandırma yöntemi kullanılarak tahmin edilmiştir. Veri örnekleri, *araba sayısı, ev sayısı* ve *mağaza sayısı* özellikleri tarafından tanımlanır. Elimizde A, O ve Z olmak üzere 3 sınıf mevcuttur. X = (araba sayısı= "2", ev sayısı = "3", mağaza sayısı = "0") örneği, sınıflandırmak istediğimiz bilinmeyen örnek olsun.

Bilinmeyen örnek *X*'in, tüm sınıflardaki her elemanla arasındaki Öklid uzaklığı hesaplanır ve k değeri 5 seçildiği için en küçük ilk 5 değere sahip komşularının arasında en sık bulunan sınıfa atanır.

Çizelge 4.2 A, O ve Z sınıfları ve *X* bilinmeyen örneğin bu sınıflara uzaklıkları

| | **Sınıf** | **Araba Sayısı** | **Ev Sayısı** | **Mağaza Sayısı** | **Öklid Uzaklığı** |
|---|---|---|---|---|---|
| 1 | Zengin Yaşam(Z) | 5 | 10 | 4 | $((5-2)^2+(10-3)^2+(4-0)^2)^{1/2} = 12.08$ |
| 2 | Zengin Yaşam(Z) | 3 | 5 | 6 | $((3-2)^2+(5-3)^2+(6-0)^2)^{1/2} = 6.40$ |
| 3 | Zengin Yaşam(Z) | 4 | 8 | 7 | $((4-2)^2+(8-3)^2+(7-0)^2)^{1/2} = 8.83$ |
| 4 | Orta Halli Yaşam (O) | 2 | 2 | 2 | $((2-2)^2+(2-3)^2+(2-0)^2)^{1/2} = 2.24$ |
| 5 | Orta Halli Yaşam (O) | 2 | 3 | 2 | $((2-2)^2+(3-3)^2+(2-0)^2)^{1/2} = 2$ |
| 6 | Orta Halli Yaşam (O) | 1 | 2 | 2 | $((1-2)^2+(2-3)^2+(2-0)^2)^{1/2} = 2.45$ |
| 7 | Alt Seviye Yaşam(A) | 1 | 1 | 0 | $((1-2)^2+(1-3)^2+(0-0)^2)^{1/2} = 2.24$ |
| 8 | Alt Seviye Yaşam(A) | 0 | 1 | 0 | $((0-2)^2+(1-3)^2+(0-0)^2)^{1/2} = 2.83$ |
| 9 | Alt Seviye Yaşam(A) | 0 | 1 | 1 | $((0-2)^2+(1-3)^2+(1-0)^2)^{1/2} = 3$ |

Elde edilen ilk 5 değer sırasıyla; 2(O), 2.24(A), 2.24(O), 2.45(O), 2.83(A) değerleridir. Bu değerler arasında en çok o*rta Halli* sınıfa ait elemanlar bulunduğundan *X* bilinmeyen örneği *orta halli* sınıfa atanır.

K-EK yöntemi doku tanıma (pattern recognition), doküman sınıflandırma vb. çalışmalarda kullanılmıştır.

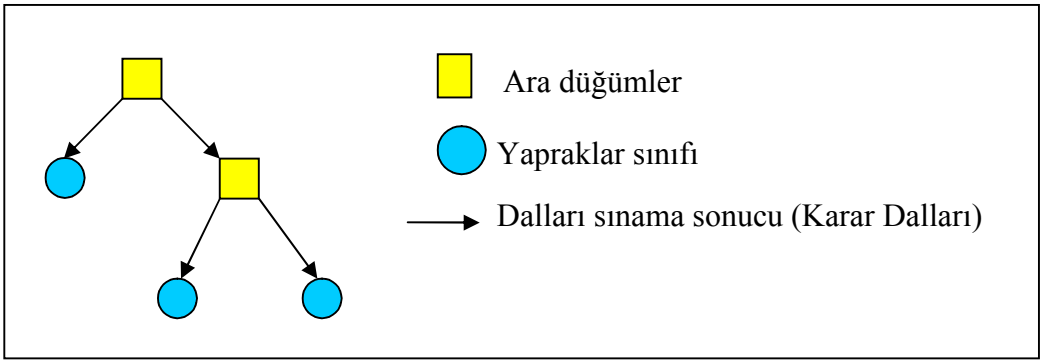
Özgür tarafından yapılan ve bölüm 4.1'deki veri setlerinin kullanıldığı çalışmada (Özgür, 2002), K-En Yakın Komşuluk yöntemi ile yapılan sınıflandırma işleminde Classic3 veri seti için k değeri 35 iken %98.7'lik, Hitech veri seti için k değeri 25 iken %66.8'lik, LA1 veri seti için k değeri 20 iken %83.5'lik, Reuters-21578 veri seti için k değeri 30 iken %75.7'lik, Wap veri seti için k değeri 35 iken %76.8'lik doğru sınıflandırma başarısı sağlanmıştır.

Ko ve Seo tarafından bölüm 4.1'de anlatılan veri setleri kullanılarak K-EK yöntemi ile doküman sınıflandırma için vektör uzayı modeli tanımlanmış ve her doküman için ağırlık vektörü oluşturulmuştur. Uygun vektörler arasındaki açının kosinüs değeri ile iki doküman arasındaki benzerlik ölçülmüştür. K-EK ile yapılan sınıflandırma işleminde %82.5'lik doğru sınıflandırma başarısı sağlanmıştır.

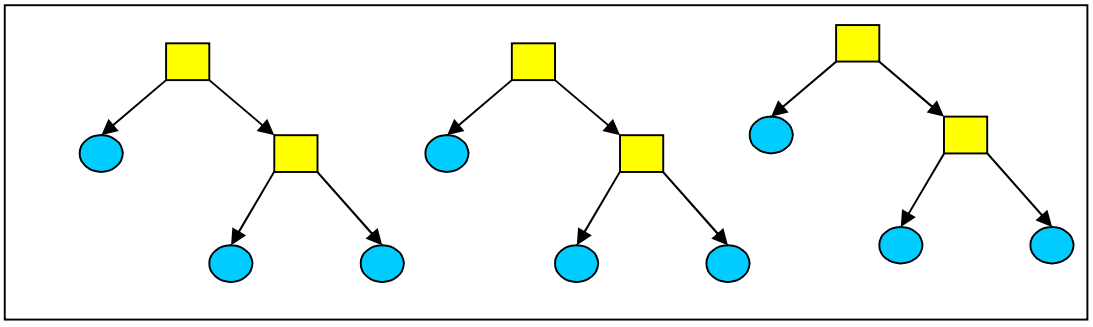
### 4.4 Rastgele Orman (RO)

Rastgele Orman, Breiman ve Cutler tarafından sunulmuş, durum ya da değişkenlerin rastgele örneklerinden üretilmiş çoklu ağaçlardan oluşmuş bir sınıflayıcıdır [8]. Rastgele Orman, akış şemasına benzer bir çok ağaç yapısından oluşur. Şekil 4.5'te Rastgele Orman Yapısı gösterilmektedir.

Ağaç yapısında olup, dallar ve yapraklar ağacın elemanlarıdır. Her bir nitelik bir düğüm tarafından temsil edilir [9]. En son yapı yaprak, en üst yapı kök ve bunların arasında kalan yapılar ise dal olarak adlandırılır. Şekil 4.6'da akış diyagramı şeklinde ağaç yapısı verilmiştir. Rastgele Orman yönteminde, ağaç bütün verinin oluşturduğu tek bir düğümle başlamakta, eğer örneklerin hepsi aynı sınıfa aitse düğüm yaprak olarak sonlanmakta ve sınıf etiketi verilmektedir. Eğer örnekler aynı sınıfa dahil değilse, örnekleri sınıflara en iyi bölecek olan nitelik seçilmektedir.




Şekil 4.5 Akış diyagramı şeklinde ağaç yapısı



Şekil 4.6 Rastgele Orman yapısı

Ağaç yapısı yukarıdan aşağıya, yinelemeli olarak böl ve kazan yöntemine göre inşa edilirken şu adımlar izlenmiştir (Demiriz, 2006);

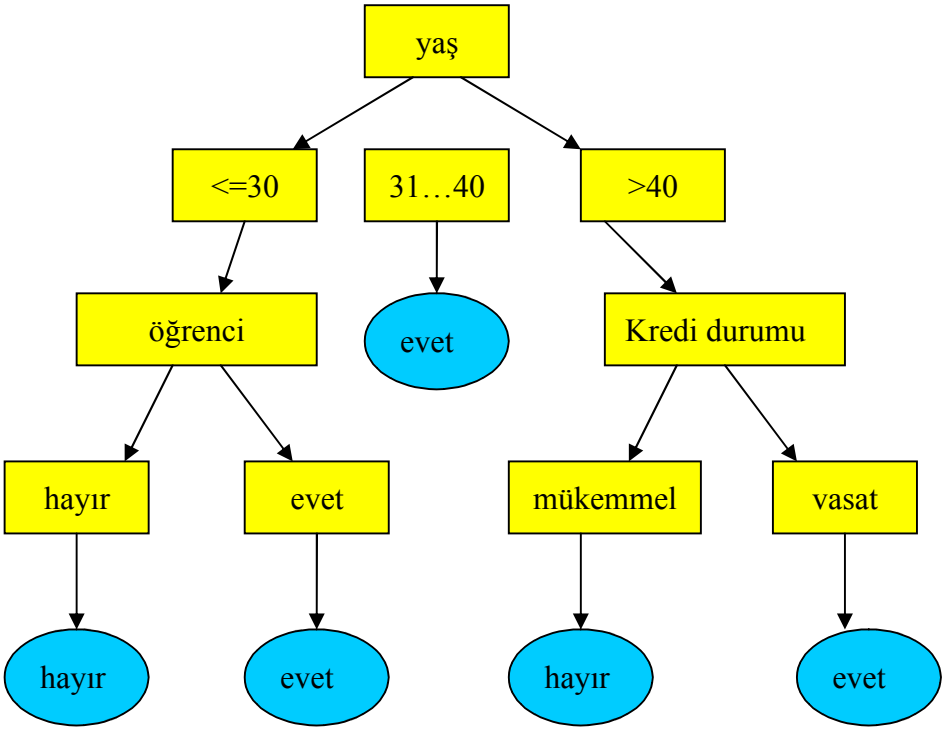
- Başlangıçta bütün noktalar ağacın kökünde toplanmaktadır.
    - Kategorik veriler kullanılır, sürekli değişkenlerin önceden kesikli hale getirilmesi gerekir.
    - Örnekler, seçilen değişkenlere göre yinelemeli olarak bölümlenir.
    - Değişkenlerin seçimi sezgisel veya belli bir istatistiksel ölçüye dayanır.
- Bölümlemenin durması için şartlar:
    - Bir düğümde bulunan bütün örnekler aynı sınıfa aittir.
    - Bölümlemenin yapılacağı değişken kalmamıştır. Yani o düğüme (yaprak) gelene kadar bütün değişkenler kullanılmıştır.
    - Başka örnek kalmamıştır.

Yeni bir Rastgele Orman oluşturulduğu zaman veriler üzerindeki gürültüye bağlı olarak çok çeşitli ve sayıda yaprak ve dal oluşabilir. Budama işlemleri bu gereksiz ve ağaçta karmaşıklığa yol açan kayıtların silinmesi anlamına gelir. Gürültülü verilerden oluşan ve sınama kümesinde hataya neden olan dalların silinmesi sınıflandırma başarımını artırır. Çizelge 4.3'te Rastgele Orman oluşturulacak veri seti verilmiştir. Şekil 4.7'de, çizelge 4.3'te verilen örnek setinin akış diyagram şeklindeki Rastgele Orman yapısı verilmiştir.

Rastgele Orman, sınıflandırma algoritmalarını uygulayabilmek için uygun bir yapıdır. IF-THEN yapısı kullanılarak Rastgele Orman içerisinde sınıflandırma kuralları sorgulanabilir. Ağacın kökünden bir yaprağa giden yol için bir if-then kuralı tanımlanabilir. Bir nesneyi sınıflamak için o nesnenin giriş vektörü, ormandaki her bir ağaca konulur ve yukarıdan aşağı doğru test edildikten sonra ağaçlar bu nesnenin sınıfını verir. Nesne tüm ağaçların verdiği sınıf bilgisinden en çok olan sınıfa atanır.

Çizelge 4.3 Rastgele orman oluşturulacak örnek veri seti

| **yaş** | **öğrenci** | **kredi durumu** | **Bilgisayar Alır?** |
|---|---|---|---|
| ≤30 | hayır | vasat | hayır |
| ≤30 | hayır | mükemmel | hayır |
| 31…40 | hayır | vasat | evet |
| >40 | hayır | vasat | evet |
| >40 | evet | vasat | evet |
| >40 | evet | mükemmel | hayır |
| 31…40 | evet | mükemmel | evet |
| ≤30 | hayır | vasat | hayır |
| ≤30 | evet | vasat | evet |
| >40 | evet | vasat | evet |
| ≤30 | evet | mükemmel | evet |
| 31…40 | hayır | mükemmel | evet |
| 31…40 | evet | vasat | evet |
| >40 | hayır | mükemmel | hayır |




Şekil 4.7 Örnek ağaç yapısı

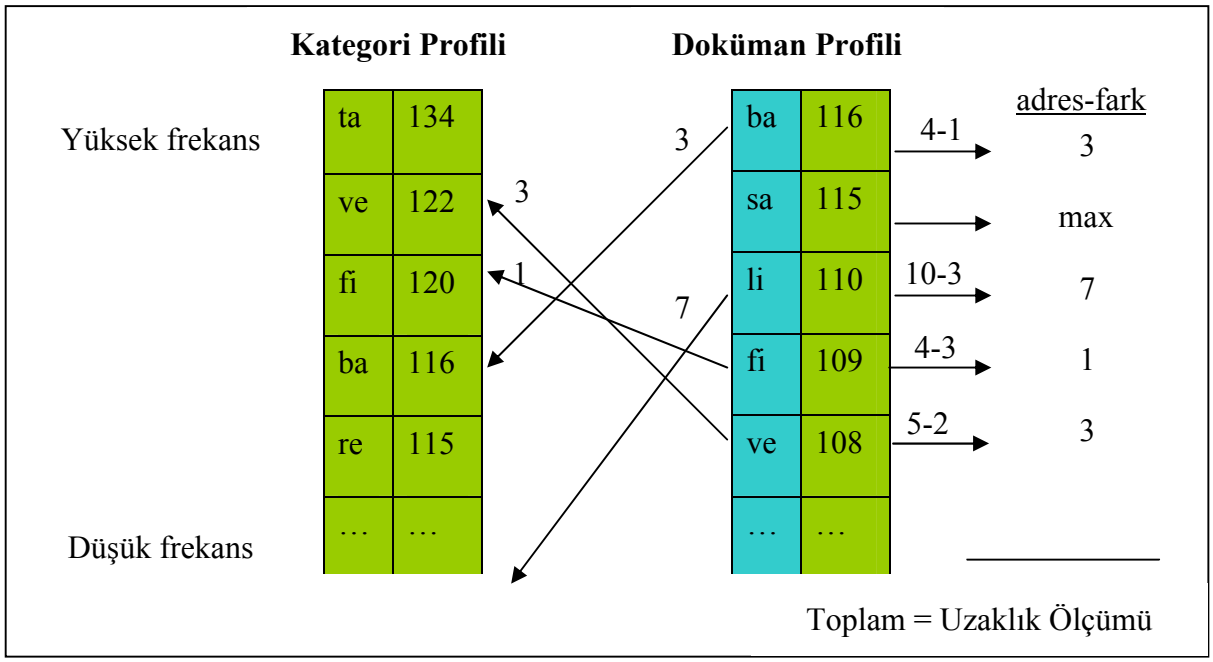
Rastgele Orman bir çok alanda kullanılmaktadır. Bunların başlıcaları sağlık, oyun ve iş sektörüdür.

Kramer ve arkadaşları tarafından, SIPPER veri setinden (Shadow Image Particle Profiling Evaluation Recorder) su altı botanik görüntüsünü tanıma işlemi yapılmıştır (Kramer, 2004). 6,000 gri seviyesinde görüntü kullanılmıştır. SIPPER veri setindeki bir çok görüntünün çizgileri net değildir ve gürültü kirliği çok fazladır. Çizgilerin net olmamasına çok bağlı olamayan şeklin özellikleri oluşturulmuştur. Özellikler, özellik seçim işlemi yapılarak azaltılmıştır. Özellik vektörleri kullanılarak DVM ve Rastgele Orman sınıflandırma yöntemleri kullanılmıştır. Rastgele Orman yönteminde 100 ağaç kullanılmış ve %74.5'lik doğru tanıma başarısı elde edilmiştir.

## 4.5 Ng_ind

*ng_ind* adını verdiğimiz bu yöntem, Cavnar, W.B. ve J.M. Trenkle tarafından yazılmış "N-gram Tabanlı Doküman Sınıflandırma" makalesindeki "Measure profile distance" adını verdikleri profiller arasındaki benzerlik ölçümünden esinlenilerek hazırlanmıştır (Cavnar,1994).

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 oluşmaktadır. Ayırt edici özelliği olduğu düşünülerek profildeki yüksek frekanslı n-gram'lar kullanılır. Ölçüm aşamasında kategori profili ve sınıflandırılacak dokümanın profili olmak üzere iki n-gram profil alınır ve "adres-fark" olarak isimlendirilen basit bir dizi-sıra istatistik hesabı yapılır. Bu ölçü, bir profil içerisindeki bir n-gram'ın yerinin diğer profil içerisindeki yerinden ne kadar uzakta olduğu bilgisini verir. Şekil 4.8'de, n-gram kullanılarak yapılan hesaplama örneği verilmiştir. Doküman profilindeki her bir n-gram için, kategori profilindeki onun benzeri bulunur ve daha sonra kendi yeri ile buradaki yeri arasındaki uzaklık farkı hesaplanır. Örneğin Şekil 4.8'de, "ve" n-gram'ı için dokümandaki sırası 2, fakat kategorideki sırası 5'tir. O zaman "ve" n-gram için "adres-fark" (5 fark 2) 3 tür. Eğer doküman profilindeki herhangi bir n-gram, kategori profilinde bulunmuyorsa maksimum uzaklık değeri verilir. Sınıflandırılacak dokümanın profilinde bulunan tüm n-gram'lar için uzaklık değerleri toplanır. Böylece doküman, en küçük değere sahip olan kategoriye atanır.

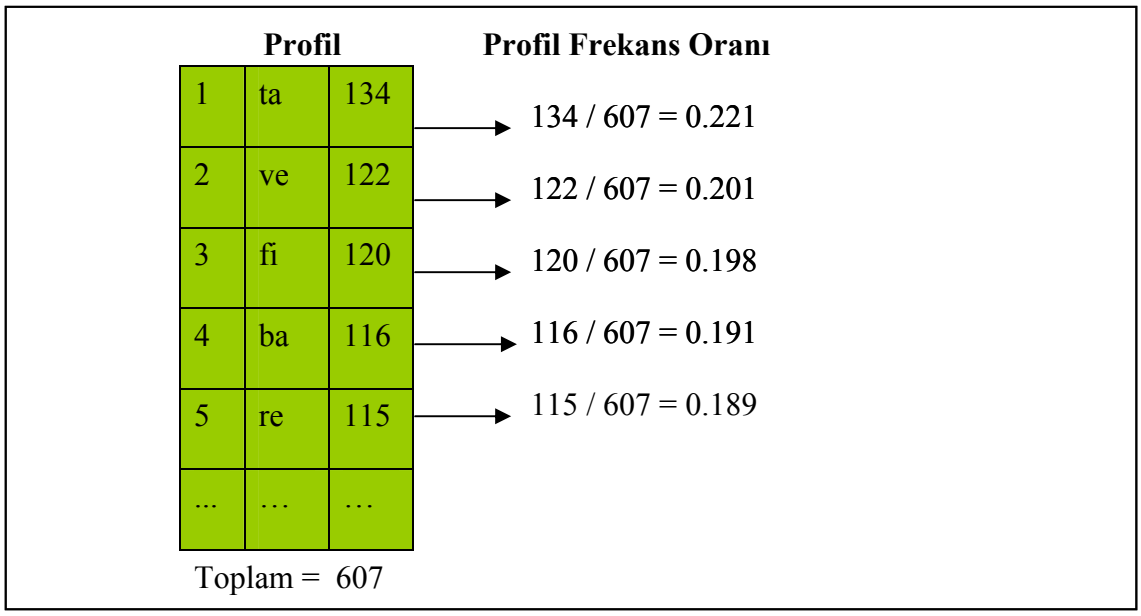



Şekil 4.8 Adres-fark değerinin hesaplanması

Kategori ve doküman profilleri belirli uzunluklarda alınarak sınıflandırma işlemi yapılır. Örneğin kategori profilinin 2000, doküman profilinin 500 adet n-gram'a sahip olduğunu farz edelim. Ölçüm için profiller karşılaştırılırken, profil uzunluğu ilk 100 veya 200 gibi değerlerini alabilir. *Ng_ind* adını verdiğimiz yöntemin ölçüm aşamasında "adres-fark" olarak

isimlendirilen dizi-sıra istatistik hesap işlemine n-gram'ların *"profil frekans oranı"* katılmıştır.

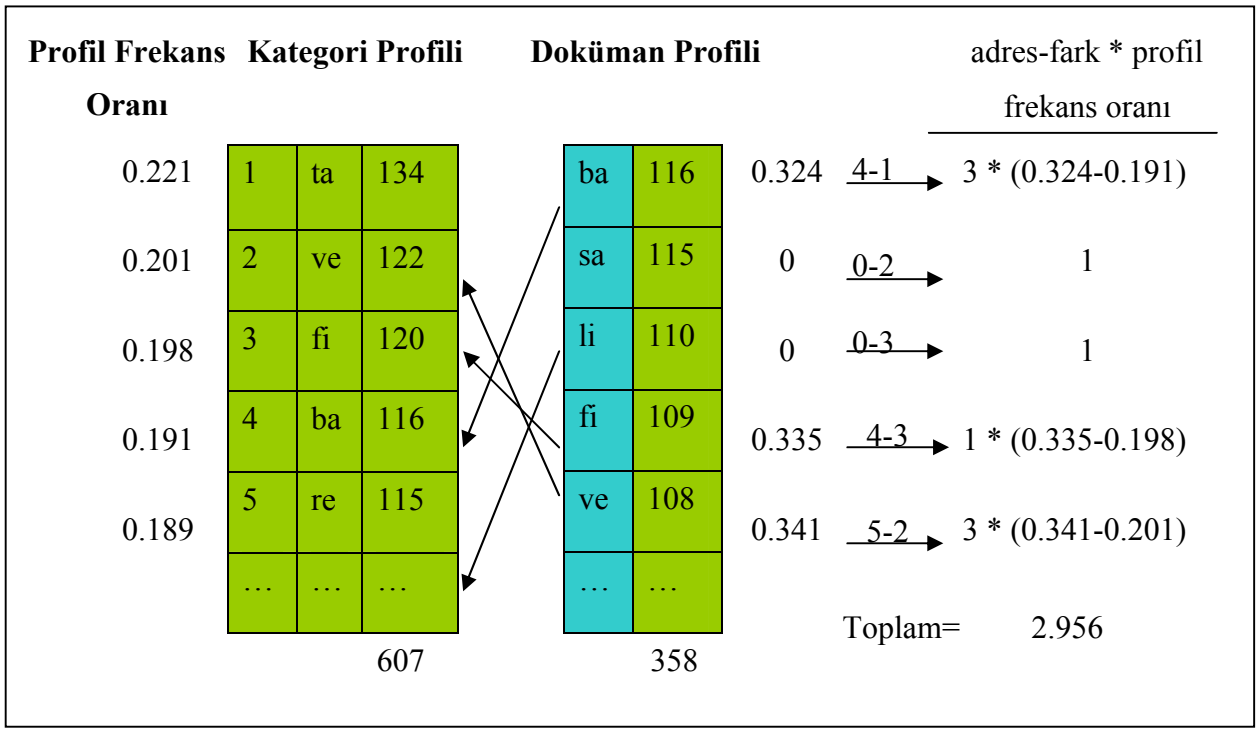
Profildeki her bir n-gram'ın frekans değerinin, belirli profil uzunluğu içerisinde bulunan tüm n-gram'ların frekans değerlerinin toplamına oranı n-gram'ın *"profil frekans oranını"* verir. Şekil 4.9'da, profildeki n-gram'ların her biri için profil frekans oranı hesaplama örneği verilmektedir. Örnekte profil uzunluğunun 5 olduğu farz edilmiştir. Toplam frekans değeri, ilk 5 n-gram'ın frekans değerlerinin toplanmasıyla 607 olarak bulunur. Örneğin "ta" n-gram'ının frekans değeri 134'ün, toplam frekans değeri 607'ye oranı, "ta" n-gram'ının profil frekans oranını verir ve bu değer 0.221'dir.




Şekil 4.9 Profil frekans oranı hesaplanması

*Ng_ind* yönteminde doküman profilindeki her bir n-gram için, kategori profilindeki onun benzeri bulunur ve daha sonra kendi yeri ile buradaki yeri arasındaki adres-fark değeri hesaplanır. N-gram'ın uzaklık farkı ile doküman ve kategori profilinde hesaplanan profil frekans oranlarının farkının mutlak değeri çarpılır. Bir n-gram'ın profil frekans değerlerinin mutlak farkı ne kadar küçülürse profillerdeki benzerlik o kadar artmaktadır. Eğer doküman profilindeki herhangi bir n-gram, kategori profilinde bulunmuyorsa maksimum uzaklık değeri 1 olarak alınır. Sınıflandırılacak dokümanın profilinde bulunan her bir n-gram'ın adres-fark değeri ile profil frekans oranı çarpımı, o n-gram'ın uzaklık değerini verir. Dokümanın profilinde bulunan tüm n-gram'lar için uzaklık değerleri toplanır. Test edilecek doküman profili tüm kategori profilleri ile karşılaştırılır ve elde edilen sonuçlara göre doküman, en

küçük değere sahip olan kategoriye atanır. Şekil 4.10'da, *ng_ind* yöntemiyle kategori ve sınıflandırılacak doküman arasındaki benzerliğin hesaplanması örneği verilmiştir. Profil uzunluğu 5 olarak alınmış ve toplam frekans değeri, kategori profili için 607, sınıflandırılacak doküman için 358 olarak bulunmuştur. Örneğin "ba" n-gram'ının kategori profilindeki profil frekans oranı 0.191, doküman profilindeki profil frekans oranı 0.324'tür. "ba" n-gram'ı için profil frekans oranlarının farkının mutlak değeri (0.324 mutlak fark 0.191) 0.133, adres-fark değeri (4 fark 1) 3 ve uzaklık değeri 0.399'dur. Belirlenmiş profil uzunluğunda bulunan tüm n-gram'lar için uzaklık değerleri toplamı 2.956'dir ve bu değer dokümanın kategoriye olan benzerliğini verir.




Şekil 4.10 *Ng_ind* yöntemiyle benzerlik değerinin hesaplanması

## 4.6 Korelasyon Tabanlı Özellik Seçme

Her bir dokümanın özel bir şekilde ifade edilmesi amacıyla çıkarılmış özeliklerden bazıları dokümanları ayırt edici nitelik taşımamaktadır. Ayırt edici nitelikte olmayan özelliklere sahip özellik vektörleri kullanıldığı durumda doğru sınıflandırma başarısı düşebilir. Gereksiz özelliklerin çıkarılıp sistemin başarısını tekrar ölçmek amacıyla özellik azaltma işlemi yapılmıştır.

Korelasyon Tabanlı Özellik Seçme (KTÖS), korelasyona dayanan bir özellik alt küme seçme metodudur. KTÖS, konu ile ilgisiz, gereksiz ve gürültü niteliğine sahip özellikleri çok çabuk teşhis eden ve onları eleyen bir yöntemdir. KTÖS genellikle özelliklerin yarısından fazlasını

eler ve böylece öğrenme yöntemlerinin başarı performansının arttırılmasını sağlar. (Hall ve Smith, 1999).

Bu çalışmadaki özellik alt kümeleri, WEKA'nın [2] KTÖS uygulaması ile oluşturulmuştur.

### 4.7 Oylama (Vote)

Sınıflandırma başarısını arttırabilmek için bazı sınıflandırıcıların birlikte kullanılması yoluna gidilmiştir. Sınıflandırıcıları birlikte kullanma işleminin performansını gözlemlemek amacıyla, Naive Bayes, Rasgele 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.

Oylama, doğru sınıflama başarısını arttırmak için başarılı olduğu düşünülen sınıfları tahmin yoluyla seçip birlikte kullanma işlemidir.

# 5. DENEYSEL SONUÇLAR

Bu çalışmanın amacı, dokümanların yazarlarını, türlerini ve yazarların cinsiyetini belirlemektir. 3 adet veri seti için 2-gram, 3-gram ve 4-gram'lardan oluşan özellik vektörleri çıkarılmış, daha sonra bu üç özellik vektöründen türetilmiş, özellik sayısı daha az olan vektörleri de kullanılarak, toplam 6 özellik vektörü ile sınıflandırma işlemi gerçekleştirilmiştir. Ayrıca sınıflandırmadaki başarı oranını arttırmak için başarılı olan sınıflandırıcılar birlikte kullanılmıştır. Birlikte kullanma işlemi, tek bir özellik vektörü kullanıldığında, başarısı en yüksek olan sınıflandırıcıların bir araya getirlerek oylama sonucunda dokümanın hangi sınıftan olabileceğine karar veren bir işlemdir. Dokümanları sınıflandırmak için verilerin n gruba ayrıldığı n katlı çapraz geçerlilik *(N-Fold Cross Validation)* kullanılmıştır ve çapraz geçerlilik değeri 10 (*10-fold cross validation*) olarak alınarak, Weka hazır sınıflandırma aracı içerisinde yer alan Naive Bayes, Destek Vektör Makinesi, K-En Yakın Komşuluk, Rastgele Orman yöntemleri ve bizim geliştirdiğimiz ng_ind yöntemi kullanılmıştır.

Çapraz geçerlilik yönteminde veriler n gruba ayrılırken, ilk aşamada birinci grup test, diğer gruplar eğitim için kullanılır. Bu süreç her defasında bir grubun test, diğer grupların eğitim amaçlı kullanılması ile sürdürülür. Sonuçta elde edilen 10 hata oranının ortalaması, kurulan modelin tahmini hata oranı olacaktır [10].

Ayrıca Weka Tool'u kullanılarak Naive Bayes, Destek Vektör Makinesi, K-En Yakın Komşuluk, Rastgele Orman sınıflandırıcıları birlikte kullanarak her bir veri seti için denemeler yapılmıştır. Cinsiyet belirleme işleminde ele edilen en iyi doğru sınıflandırma oranı %92,63 iken, yazar tanıma işleminde alınan en iyi sonuç %89,63 ve tür belirleme işleminde ele edilen en iyi sonuç ise %92,08 olmuştur. Cinsiyet belirleme, yazar tanıma ve tür belirleme için özellik sayısı azaltılarak yapılan sınıflandırma işlemlerinde, doğru sınıflandırma başarısının arttığı görülmüştür.

## 5.1 Veri Setine Göre Deneysel Sonuçlar

Bu bölümde üç farklı veri seti üzerinde yapılan deneysel sonuçlar sırasıyla verilmiştir.

### 5.1.1 Veri Seti-I için Deneysel Sonuçlar

Cinsiyet belirleme çalışmasında kullanılan veri seti 10 tanesi bay, 10 tanesi de bayan olmak üzere 20 yazardan oluşmaktadır. 20 yazara ait 40 doküman alınarak oluşturulan Veri Seti-I üzerinde, 6 farklı özellik vektörü kullanılarak 5 farklı sınıflandırıcıyla deneyler yapılmıştır.

Çizelge 5.1'de ng_ind yönteminin farklı özellik sayılarındaki doğru sınıflandırma başarıları gösterilmektedir. Çizelge 5.2'de de 20 farklı yazarın cinsiyetlerine göre sınıflandırılmasından alınan başarı sonuçları verilmiştir.

Çizelge 5.1 ng_ind yönteminin cinsiyet belirlemedeki başarısı

| **n-gram model** | **Özellik sayısı ve Başarı Oranı %** | | | | | |
|---|---|---|---|---|---|---|
| | **100** | **200** | **257** | **300** | **400** | **500** |
| **2-gram** | %71,25 | %78,75 | **%78,75** | %80 | %82,5 | %78,75 |
| | **100** | **200** | **300** | **324** | **400** | **500** |
| **3-gram** | %75 | %82,5 | %78,75 | **%80** | %82,5 | %78,75 |
| | **100** | **142** | **200** | **300** | **400** | **500** |
| **4-gram** | %91,25 | **%85** | %88,75 | %88,75 | %88,75 | %85 |

Ng_ind yöntemi $OV_2$'nin özellik sayısına eşit iken (257) %78,75'lik başarı ile doğru sınıflandırma yapıyor iken kullanılan özellik vektörü eleman sayısı 400 iken %82,5'lik başarı elde edilmiştir. $OV_3$'ün özellik sayısına eşit iken (324) %80'lik başarı alınıyor iken özellik vektörü eleman sayısı 400 alındığında bu başarı %82,5 yükselmektedir ve sadece NB sınıflandırıcısına göre daha başarılı olmuştur. $OV_4$'ün özellik sayısına eşit iken (142) %85'lik başarı vermiş iken özellik vektörü eleman sayısı 400 alındığında %88,75 ile en başarılı sınıflandırıcı olmuştur.

Çizelge 5.2'de NB, RO, K-EK ve DVM yöntemleri kullanılarak her özellik vektörü için alınmış başarı ortalamaları verilirken son kolonda da ng_ind yönteminin başarısı verilmiştir. Özellik azaltılmadan önce cinsiyet belirlemede en başarılı sınıflandırıcı DVM olup en başarılı özellik vektörü $OV_3$ olmuştur. Özellik azaltıldıktan sonra en başarılı sınıflandırıcı RO, en başarılı özellik vektörü de $OV_{3A}$ olmuştur.

Genel ortalama başarıya bakılırsa RO ve DVM sınıflandırıcıları cinsiyet belirlemede en iyi sonuçları vermiş iken, özellik vektörlerindeki en yüksek başarı özellik sayısı azaltıldığında alınmıştır.

Çizelge 5.2 Veri Seti-I üzerinde elde edilen sınıflandırma başarıları

| Özellik Vektörü | %NB | %RO | %K-EK | %DVM | %Ort Başarı | %NG_IND |
|---|---|---|---|---|---|---|
| **OV2 (257 öz.)** | 67,75 | 85,62 | 83,37 | 91,12 | 81,96 | 400 öz.<br>82.5 |
| **$OV_3$** ( 324 öz.) | 73 | 83 | 83,25 | 89,75 | 82,25 | 400 öz.<br>82,5 |
| **$OV_3$** ( 324 öz.) | 74,87 | 84,12 | 82 | 85,62 | 81,65 | 100 öz.<br>91,25 |
| **Ortalama Başarı %** | 71,88 | 84,25 | 82,88 | 88,83 | 81,95 | - |
| **$OV_{2A}$**(10 öz.) | 77,62 | 86,75 | 85,75 | 79,75 | 82,47 | - |
| **$OV_{3A}$**(27 öz.) | 85,12 | 89,87 | 86,87 | 89,37 | 87,80 | - |
| **$OV_{4A}$**(12 öz.) | 76,50 | 86,75 | 81,63 | 80,62 | 81,38 | - |
| **Ortalama Başarı %** | 79,75 | 87,79 | 84,75 | 83,25 | 83,88 | - |
| **Genel Ortalama Başarı %** | 75,81 | 86,02 | 83,81 | 86,04 | 82,91 | - |

### 5.1.2 Veri Seti-II için Deneysel Sonuçlar

Politika, magazin, güncel, spor, sağlık ve ekonomi gibi farklı konularda yazan 20 yazara ait 40 doküman alınarak oluşturulan Veri Seti-II üzerinde, 6 farklı özellik vektörü kullanılarak 5 farklı sınıflandırıcıyla deneyler yapılmıştır. Çizelge 5.3'de ng_ind yönteminin farklı özellik sayılarındaki doğru sınıflandırma başarıları gösterilmektedir. Çizelge 5.4'de de 20 farklı yazarın sınıflandırılmasından alınan başarı sonuçları verilmiştir.

Çizelge 5.3 ng_ind yönteminin yazar tanımadaki başarısı

| **n-gram model** | **Özellik sayısı ve Başarı Oranı %** | | | | | |
|---|---|---|---|---|---|---|
| | **100** | **200** | **257** | **300** | **400** | **500** |
| **2-gram** | %70 | %71,25 | **%70** | %72,5 | %75 | %71,25 |
| | **100** | **200** | **300** | **324** | **400** | **500** |
| **3-gram** | %75 | %73,75 | %73,75 | **%70** | %75 | %73,75 |
| | **100** | **142** | **200** | **300** | **400** | **500** |
| **4-gram** | %75 | **%70** | %73,75 | %75 | %77,5 | %73,75 |

Ng_ind yöntemi $OV_2$'nin özellik sayısına eşit iken (257) %70'lik başarı ile doğru sınıflandırma yapıyor iken, kullanılan özellik vektörü eleman sayısı 400 iken %75'lik başarı elde edilmiştir. $OV_3$'ün özellik sayısına eşit iken (324) %70'lik başarı alınıyor iken, özellik vektörü eleman sayısı 400 alındığında bu başarı %75'e yükselmektedir. $OV_4$'ün özellik sayısına eşit iken (142) %70'lik başarı vermiş iken özellik vektörü eleman sayısı 400 alındığında %77,5 ile sadece RO ile DVM sınıflandırıcısına göre daha başarılı olmuştur. Çizelge 5.4'de NB, RO, K-EK ve DVM yöntemleri kullanılarak her özellik vektörü için alınmış başarı ortalamaları verilirken son kolonda da ng_ind yönteminin başarısı verilmiştir. Özellik azaltılmadan önce yazar tanımada en başarılı sınıflandırıcı DVM olup en başarılı özellik vektörü $OV_2$ olmuştur. Özellik azaltıldıktan sonra en başarılı sınıflandırıcı RO, en başarılı özellik vektörü de $OV_{2A}$ olmuştur. Genel ortalama başarıya bakılırsa NB sınıflandırıcısı yazar tanımada en iyi sonucu vermiş iken, özellik vektörlerindeki en yüksek başarı özellik sayısı azaltıldığında alınmıştır. Cinsiyet belirlemeye göre yazar tanımanın başarısının düşük olmasının sebebi sınıf sayısının çok gazla olmasıdır.

Çizelge 5.4 Veri Seti-II üzerinde elde edilen sınıflandırma başarıları

| Özellik Vektörü | %NB | %RO | %K-EK | %DVM | %Ort Başarı | %NG_IND |
|---|---|---|---|---|---|---|
| **OV**$_2$(257 öz.) | 81,88 | 67 | 68 | 89,50 | 76,59 | 400 öz.<br>75 |
| **OV**3 ( 324 öz.) | 82 | 61,50 | 62,13 | 88,25 | 73,47 | 400 öz.<br>75 |
| **OV**4 (142 öz.) | 80,75 | 63,25 | 57,13 | 80 | 70,28 | 400 öz.<br>77,5 |
| **Ortalama Başarı %** | 81,54 | 63,92 | 62,42 | 85,92 | 73,45 | - |
| **OV**$_{2A}$ (12 öz.) | 86,63 | 87,75 | 84,13 | 75,50 | 83,50 | - |
| **OV**$_{3A}$ (14 öz.) | 83 | 83,13 | 76,63 | 78,75 | 80,40 | - |
| **OV**$_{4A}$ (6 öz.) | 64,63 | 67,88 | 61,63 | 49,50 | 60,91 | - |
| **Ortalama Başarı %** | 78,08 | 79,58 | 74,13 | 67,92 | 74,94 | - |
| **Genel Ortalama Başarı %** | 79,81 | 71,75 | 68,27 | 76,92 | 74,20 | - |

### 5.1.3 Veri Seti-III için Deneysel Sonuçlar

Politika magazin, güncel, spor, sağlık ve ekonomi gibi farklı konularda yazan 20 yazara ait 40 doküman alınarak oluşturulan Veri Seti-III üzerinde, 6 farklı özellik vektörü kullanılarak 5 farklı sınıflandırıcıyla deneyler yapılmıştır. Çizelge 5.5'de ng_ind yönteminin farklı özellik sayılarındaki doğru sınıflandırma başarıları gösterilmektedir. Çizelge 5.6'da da 6 farklı doküman türünün sınıflandırılmasından alınan başarı sonuçları verilmiştir.

Çizelge 5.5 ng_ind yönteminin tür belirlemedeki başarısı

| **n-gram model** | **Özellik sayısı ve Başarı Oranı %** | | | | | |
|---|---|---|---|---|---|---|
| | **100** | **200** | **217** | **300** | **400** | **500** |
| **2-gram** | %75 | %75 | **%81,25** | %79,17 | %81,25 | %75 |
| | **100** | **200** | **208** | **300** | **400** | **500** |
| **3-gram** | %70,83 | %85,42 | **%83,33** | %87,5 | %93,75 | %93,75 |
| | **75** | **100** | **200** | **300** | **400** | **500** |
| **4-gram** | **%83,33** | **%85,42** | %85,42 | %87,5 | %93,75 | %87,5 |

Ng_ind yöntemi $OV_2$'nin özellik sayısına eşit iken (217) %81,25'lik başarı ile doğru sınıflandırma yapıyor iken aynı başarıyı kullanılan özellik vektörü eleman sayısı 400 iken de elde edilmiştir. $OV_3$'ün özellik sayısına eşit iken (208) %83,33'lük başarı alınıyor iken, özellik vektörü eleman sayısı 400 alındığında bu başarı %93,75'e yükselmektedir. $OV_4$'ün özellik sayısına eşit iken (75) %83,33'lük başarı vermiş iken özellik vektörü eleman sayısı 400 alındığında %93,75 başarı elde edilmiştir. Bu başarı diğer sınıflandırıcılara göre oldukça yüksektir. Çizelge 5.6'da NB, RO, K-EK ve DVM yöntemleri kullanılarak her özellik vektörü için alınmış başarı ortalamaları verilirken son kolonda da ng_ind yönteminin başarısı verilmiştir. Özellik azaltılmadan önce tür belirlemede en başarılı sınıflandırıcı DVM olup en başarılı özellik vektörü $OV_3$ olmuştur. Özellik azaltıldıktan sonra en başarılı sınıflandırıcı RO, en başarılı özellik vektörü de $OV_{3A}$ olmuştur. Genel ortalama başarıya bakılırsa DVM sınıflandırıcısı dokümanın türünün belirlenmesinde en iyi sonucu vermiş iken, özellik vektörlerindeki en yüksek başarı özellik sayısı azaltıldığında alınmıştır. Dokümanın türünü belirleme, cinsiyet belirleme ve yazar tanımaya göre en başaralı sonucu vermiştir. Yazar tanımaya göre sınıf sayısının daha az olması bu başarıyı vermişken, cinsiyet belirlemeye göre sınıf sayısının daha fazla olması n-gram modelinin tür tanımada cinsiyet belirlemeye göre daha başarılı olduğunu göstermektedir.

Çizelge 5.6 Veri Seti-III üzerinde elde edilen sınıflandırma başarıları

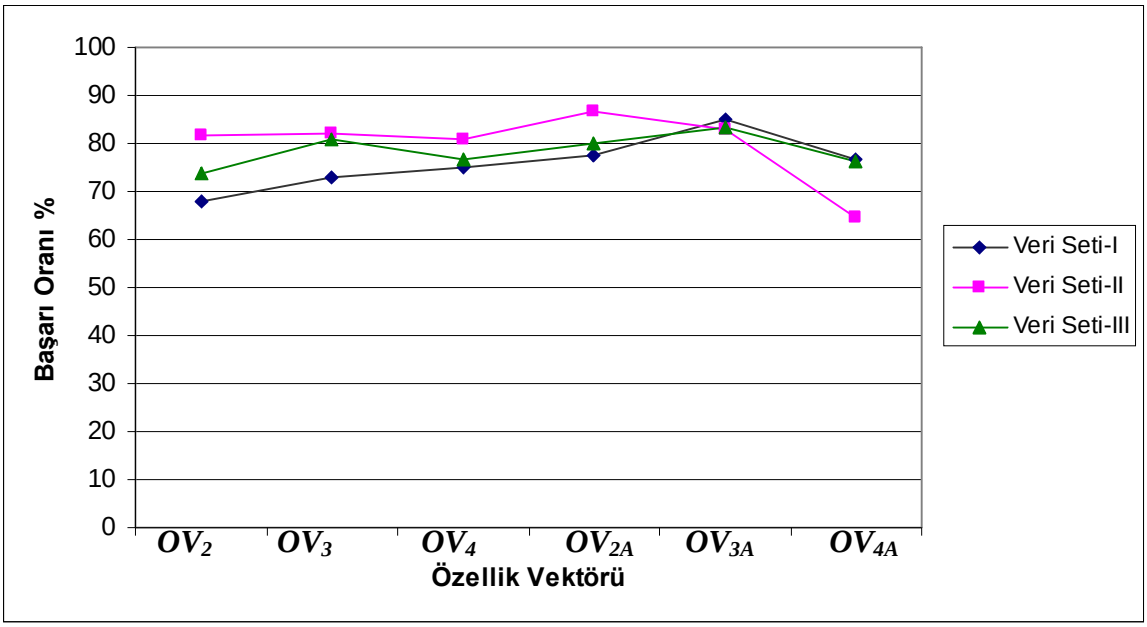
| Özellik Vektörü | %NB | %RO | %K-EK | %DVM | %Ort Başarı | %NG_IND |
|---|---|---|---|---|---|---|
| $\mathbf{OV_2}$( 217 öz.) | 73,54 | 74,58 | 81,46 | 91,88 | 80,54 | 400 öz.<br>81,25 |
| $\mathbf{OV_3}$ ( 208 öz.) | 80,83 | 72,50 | 72,29 | 92,08 | 82,29 | 400 öz.<br>93,75 |
| $\mathbf{OV_4}$ ( 75 öz.) | 76,67 | 79,58 | 60,63 | 82,50 | 78,63 | 400 öz.<br>93,75 |
| **Ortalama Başarı %** | 77,01 | 75,55 | 71,46 | 88,82 | 80,49 | - |
| $\mathbf{OV_{2A}}$ ( 15 öz.) | 80 | 86,25 | 80 | 83,75 | 82,5 | - |
| $\mathbf{OV_{3A}}$ ( 21 öz.) | 83,33 | 87,29 | 83,54 | 87,92 | 85,52 | - |
| $\mathbf{OV_{4A}}$ ( 15 öz.) | 76,25 | 84,58 | 69,58 | 74,58 | 76,25 | - |
| **Ortalama Başarı %** | 79,86 | 86,04 | 77,71 | 82,08 | 81,42 | - |
| **Genel Ortalama Başarı %** | 78,43 | 80,79 | 74,58 | 85,45 | 80,95 | - |

## 5.2 Yöntemlere Göre Deneysel Sonuçlar

Bu bölümde kullandığımız beş yöntemin yazar tanıma, tür ve cinsiyet belirlemedeki başarıları ayrı ayrı verilmektedir.

### 5.2.1 Naive Bayes

Şekil 5.1'deki grafikte Naive Bayes yöntemi kullanılarak her üç veri setinden elde edilen başarı sonuçları gösterilmiştir.




Şekil 5.1 Naive Bayes yöntemi kullanılarak her üç veri setinden elde edilen başarı oranları

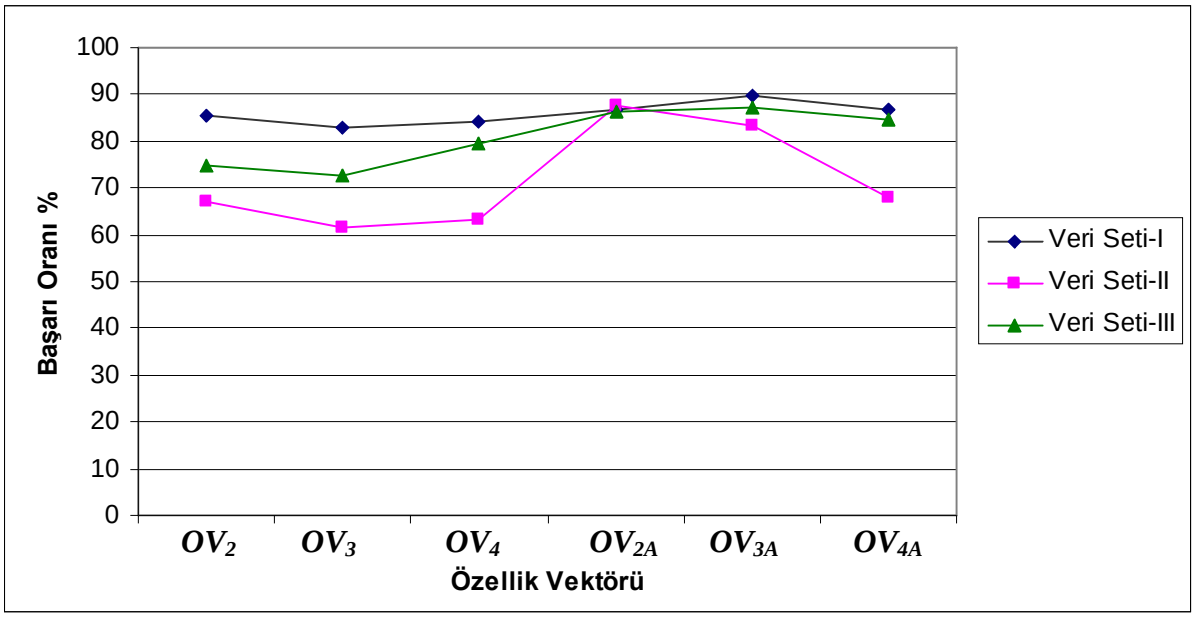
Naive Bayes yönteminin Veri Seti-I ile yapılan denemelerinde bu yöntemin en iyi sonucu %85,12 olup, $OV_{3A}$ özellik vektöründen elde edilmiştir. En kötü sonuç ise $OV_2$ özellik vektörü ile alınan %67,75'dir. Veri Seti-I için başarı ortalamaları hesaplandığında % 75,81'lik bir sonuç edilmiştir. Veri Seti-II ile yapılan deneylerde bu yöntemin en iyi sonuç %86,62 olup, $OV_{2A}$ özellik vektöründen elde edilmiştir. En kötü sonuç ise $OV_{4A}$ özellik vektörü ile alınan %64,62'tir. Veri Seti-II için başarı ortalamaları hesaplandığında % 79,81'lik bir sonuç edilmiştir. Veri Seti-III üzerindeki en yüksek başarı %83,33 olup, $OV_{3A}$ özellik vektöründen elde edilmiştir. En kötü sonuç yine $OV_2$ özellik vektöründen elde edilen %73,54'tür. Veri Seti-III için başarı ortalamaları hesaplandığında % 78,44'lük bir sonuç edilmiştir.

Her üç veri setindeki sonuçlara göre ortalama alındığında Naive Bayes'in en başarılı olduğu özellik vektörü $OV_{3A}$'dır. Sınıflandırma için en elverişsiz özellik vektörü ise ile $OV_2$'dir.

Her veri seti için başarı ortalamaları hesaplandığında % 79,81'lik başarı oranıyla Veri Seti-II, Naive Bayes'in en başarılı sonuç verdiği veri setidir.

### 5.2.2 Rastgele Orman

Şekil 5.2'deki grafikte Rastgele Orman yöntemi kullanılarak her üç veri setinden elde edilen başarı sonuçları gösterilmiştir.




Şekil 5.2 Rastgele Orman yöntemi kullanılarak her üç veri setinden elde edilen başarı oranları

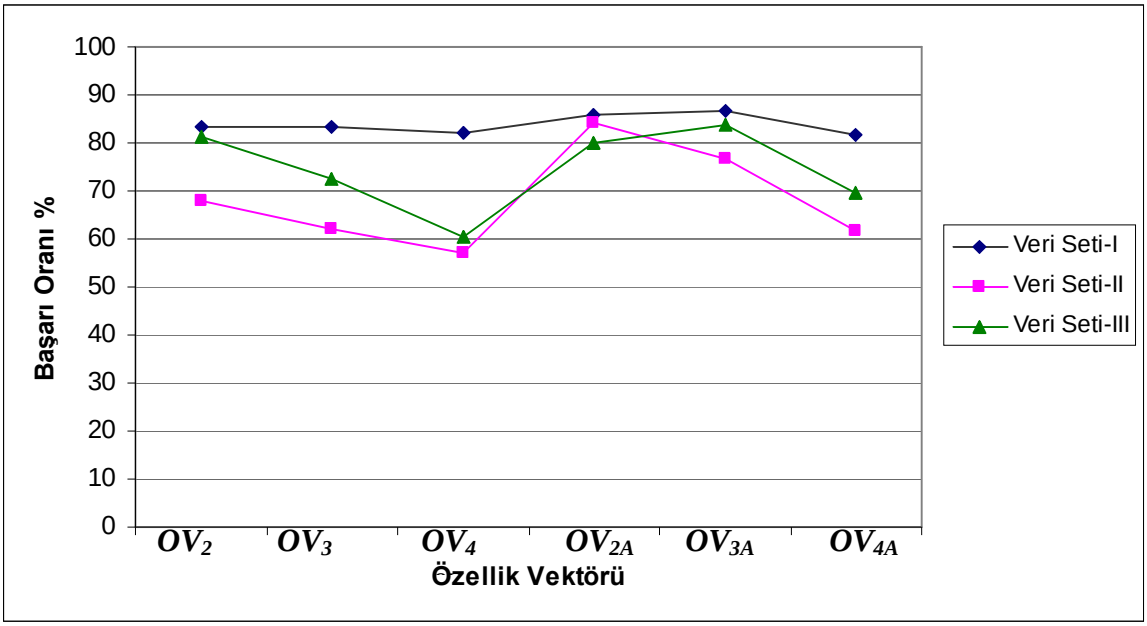
Rastgele Orman yönteminin Veri Seti-I ile yapılan deneylerinde bu yöntemin en iyi sonucu %89,87 olup, $OV_{3A}$ özellik vektöründen elde edilmiştir. En kötü sonuç ise $OV_3$ özellik vektörü ile alınan %83'tür. Veri Seti-I için başarı ortalamaları hesaplandığında % 86,02'lik bir sonuç edilmiştir. Veri Seti-II ile yapılan deneylerde bu yöntemin en iyi sonuç %87,75 olup, $OV_{2A}$ özellik vektöründen elde edilmiştir. En kötü sonuç ise $OV_3$ özellik vektörü ile alınan %61,50'dir. Veri Seti-II için başarı ortalamaları hesaplandığında % 71,75'lik bir sonuç edilmiştir. Veri Seti-III üzerindeki en yüksek başarı %87,29 olup, $OV_{3A}$ özellik vektöründen elde edilmiştir. En kötü sonuç yine $OV_3$ özellik vektöründen elde edilen %72,50'dir. Veri Seti-III için başarı ortalamaları hesaplandığında % 80,80 bir sonuç edilmiştir.

Her üç veri setindeki sonuçlara göre ortalama alındığında Rastgele Orman yönteminin en başarılı olduğu özellik vektörü $OV_{3A}$'dır. Sınıflandırma için en elverişsiz özellik vektörü ise ile $OV_3$'tür.

Her veri seti için başarı ortalamaları hesaplandığında %89,87'lik başarı oranıyla Veri Seti-I, Rastgele Orman yönteminin en başarılı sonuç verdiği veri setidir.

### 5.2.3 K-En yakın Komşuluk

Şekil 5.3'teki grafikte K-En yakın Komşuluk yöntemi kullanılarak her üç veri setinden elde edilen başarı sonuçları gösterilmiştir.




Şekil 5.3 K-En yakın Komşuluk yöntemi kullanılarak her üç veri setinden elde edilen başarı oranları

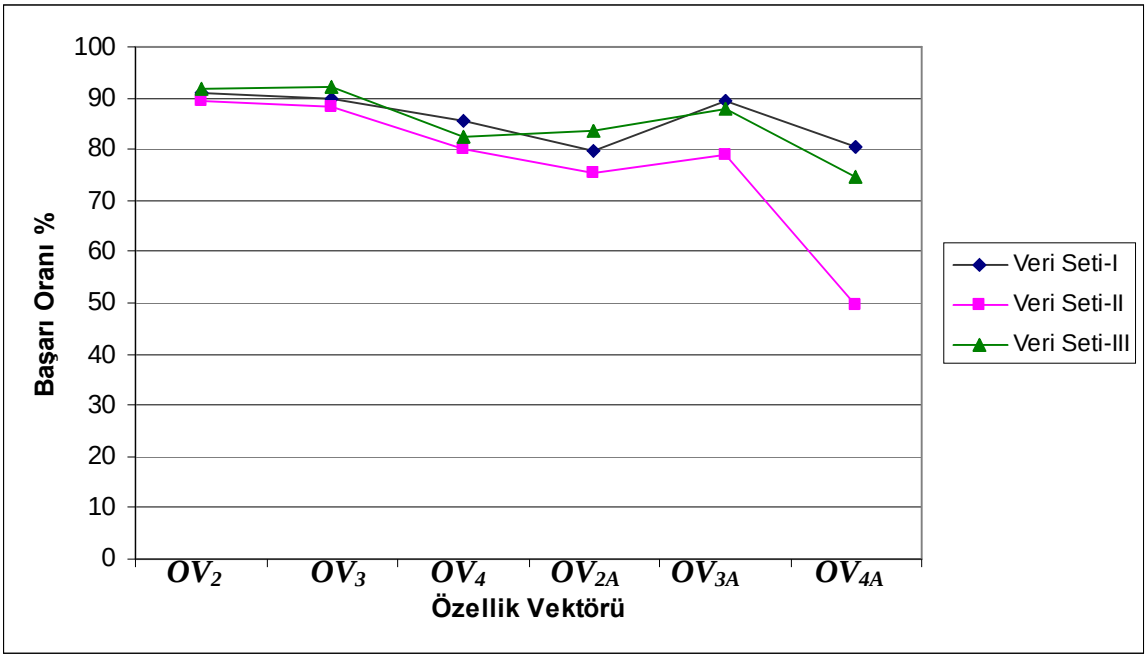
K-En yakın Komşuluk yönteminin Veri Seti-I ile yapılan denemelerinde bu yöntemin en iyi sonuç %86,87 olup, $OV_{3A}$ özellik vektöründen elde edilmiştir. En kötü sonuç ise $OV_{4A}$ özellik vektörü ile alınan %81,62'dir. Veri Seti-I için başarı ortalamaları hesaplandığında % 83,81'lik bir sonuç edilmiştir. Veri Seti-II ile yapılan deneylerde bu yöntemin en iyi sonuç %84,12 olup, $OV_{2A}$ özellik vektöründen elde edilmiştir. En kötü sonuç ise $OV_4$ özellik vektörü ile alınan %57,12'tir. Veri Seti-II için başarı ortalamaları hesaplandığında % 68,27'lik bir sonuç edilmiştir. Veri Seti-III üzerindeki en yüksek başarı %83,54 olup, $OV_{3A}$ özellik vektöründen elde edilmiştir. En kötü sonuç yine $OV_4$ özellik vektöründen elde edilen %60,63'tür. Veri Seti-III için başarı ortalamaları hesaplandığında % 74,58'lik bir sonuç edilmiştir.

Her üç veri seti için K-En yakın Komşuluk yönteminin en başarılı olduğu özellik vektörü $OV_{3A}$'dır. Sınıflandırma için en elverişsiz özellik vektörü ise ile $OV_4$'tür.

Her veri seti için başarı ortalamaları hesaplandığında % 83,81'lik başarı oranıyla Veri Seti-I, K-En yakın Komşuluk yönteminin en başarılı sonuç verdiği veri setidir.

### 5.2.4 Destek Vektör Makinesi

Şekil 5.4'deki grafikte Destek Vektör Makinesi yöntemi kullanılarak her üç veri setinden elde edilen başarı sonuçları gösterilmiştir.




Şekil 5.4 Destek Vektör Makinesi yöntemi kullanılarak her üç veri setinden elde edilen başarı oranları

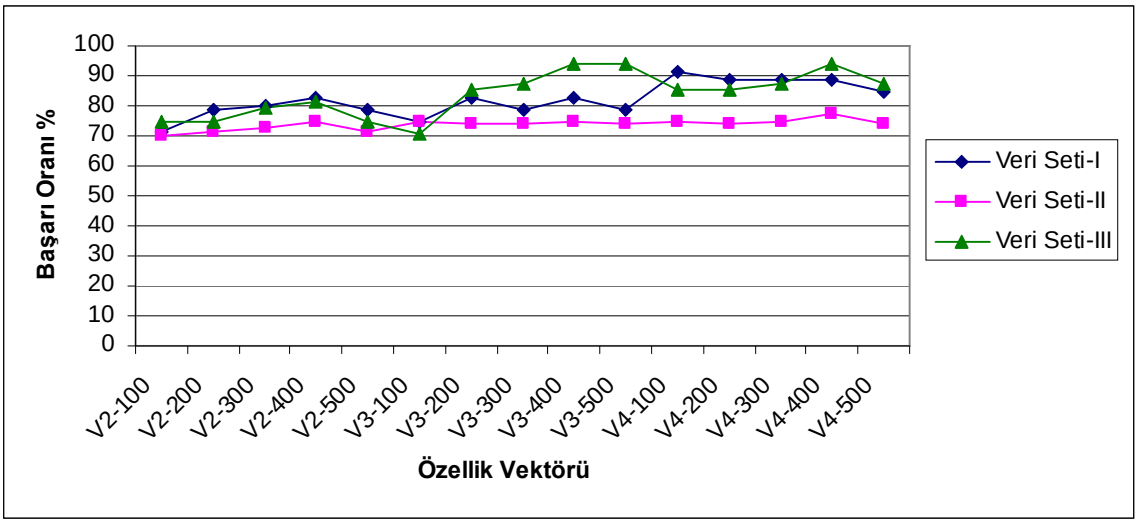
Destek Vektör Makinesi yönteminin Veri Seti-I ile yapılan denemelerinde bu yöntemin en iyi sonuç %91,12 olup, $OV_2$ özellik vektöründen elde edilmiştir. En kötü sonuç ise $OV_{2A}$ özellik vektörü ile alınan %79,75'dir. Veri Seti-I için başarı ortalamaları hesaplandığında % 86,04'lük bir sonuç edilmiştir. Veri Seti-II ile yapılan deneylerde bu yöntemin en iyi sonuç %89,50 olup, $OV_2$ özellik vektöründen elde edilmiştir. En kötü sonuç ise $OV_{4A}$ özellik vektörü ile alınan %49,50'dir. Veri Seti-II için başarı ortalamaları hesaplandığında % 76,92'lik bir sonuç edilmiştir. Veri Seti-III üzerindeki en yüksek başarı %92,08 olup, $OV_3$ özellik vektöründen elde edilmiştir. En kötü sonuç yine $OV_{4A}$ özellik vektöründen elde edilen %74,58'tir. Veri Seti-III için başarı ortalamaları hesaplandığında % 85,45'lik bir sonuç edilmiştir.

Her üç veri setindeki sonuçlara göre ortalama alındığında Destek Vektör Makinesi yönteminin en başarılı olduğu özellik vektörü $OV_3$'dür. Sınıflandırma için en elverişsiz özellik vektörü ise ile $OV_{4A}$'dir.

Her veri seti için başarı ortalamaları hesaplandığında % 86,04'lük başarı oranıyla Veri Seti-I, Destek Vektör Makinesi yönteminin en başarılı sonuç verdiği veri setidir.

### 5.2.5 Ng_ind

Şekil 5.5'teki ng-ind yöntemi kullanılarak her üç veri setinden elde edilen başarı sonuçları gösterilmiştir. Veri Seti-I, Veri Seti-II ve Veri Seti-III içindeki her bir 2-gram, 3-gram ve 4-gram özellik vektörlerinin uzunluğu 100, 200, 300, 400 ve 500 olarak alınmıştır. Çalışmada uzunluğu 100 olan 2-gram özellik vektöründen *V2-100* olarak bahsedilecektir. Aynı şekilde ng_ind yönteminde kullanılan özellik vektörlerinden V2-100, V2-200, V2-300, V2-400, V2-500, V3-100, V3-200, V3-300, V3-400, V3-500, V4-100, V4-200, V4-300, V4-400, V4-500 olarak bahsedilecektir.




Şekil 5.5 Ng_ing yöntemi kullanılarak her üç veri setinden elde edilen başarı oranları

Ng-ind yönteminin Veri Seti-I ile yapılan denemelerinde en iyi sonuç %91,25 ile V4-100 özellik vektöründen elde edilmiştir. En başarısız sonuç ise V2-100 özellik vektörü ile alınan %71,25'dir. Veri Seti-I için başarı ortalamaları hesaplandığında % 82,83'lik bir sonuç edilmiştir. Veri Seti-II ile yapılan denemelerde en iyi sonuç %77,5 olup, V4-400 özellik vektöründen elde edilmiştir. En başarısız sonuç ise V2-100 özellik vektörü ile alınan %70'tir. Veri Seti-II için başarı ortalamaları hesaplandığında % 73,75'lik bir sonuç edilmiştir. Veri Seti-III üzerindeki en yüksek başarı %93,75 olup, V3-400, V3-500 ve V4-400, özellik vektörden elde edilmiştir. En başarısız sonuç yine V3-100 özellik vektöründen elde edilen %70,83'tür. Veri Seti-III için başarı ortalamaları hesaplandığında % 83,75'lik bir sonuç edilmiştir.

Her üç veri seti için Ng-ind yönteminin en başarılı olduğu özellik vektörü V4-400'dür. Sınıflandırma için en elverişsiz özellik vektörü ise ile V2-100'dür.

Her veri seti için başarı ortalamaları hesaplandığında % 83,75'lik başarı oranıyla Veri Seti-III, Ng_ind yönteminin en başarılı sonuç verdiği veri setidir. Kısaca ng-ind yöntemi tür belirlemede cinsiyet ve yazar tanımaya göre daha başarılı olmuştur.

Çizelge 5.7 Cavnar'ın ortaya attığı yöntem 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 |

Çizelge 5.7'de Cavnar'ın ortaya attığı (Cavnar,1994) yöntem ve bizim bu yöntem üzerinde yaptığımız değişiklikler ile oluşturup ng_ind adını verdiğimiz yöntem karşılaştırmalı olarak verilmiştir. Üç farklı n-gram modeli ve 5 farklı özellik vektörü kullanılarak bu sonuçlar alınmıştır. 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 başarılıdır.

## 5.3 Sınıflandırıcıların Birlikte Kullanılmasına Göre Deneysel Sonuçlar

Sınıflandırma başarısını arttırabilmek için bazı sınıflandırıcıların oylama yöntemi ile birlikte kullanılması yoluna gidilmiştir. Sınıflandırıcıların birlikte kullanılması işleminin performansını gözlemlemek amacıyla, Naive Bayes, Rasgele Orman, Destek Vektör Makinesi, K-En Yakın Komşuluk gibi sınıflandırma yöntemleri farklı kombinasyonlarda bir araya getirilmiştir. Birlikte kullanma işlemi, daha önceki deneylerde başarısı en yüksek olan sınıflandırıcıların bir araya getirilerek oylama sonucunda dokümanın hangi sınıftan olabileceğine karar veren bir işlemdir. Denemelerde $OV_2$, $OV_3$, $OV_4$, $OV_{2A}$, $OV_{3A}$, $OV_{4A}$ özellik vektörleri kullanılmıştır.

Veri Seti-I üzerinde sınıflandırıcıların birlikte kullanılması ile elde edilen sonuçlar Çizelge 5.8'de verilmiştir.

Çizelge 5.8 Veri Seti-I için sınıflandırıcılar birlikte kullanılarak alınan sonuçlar

| | **Oylama ile Birlikte Kullanılan Sınıflandırıcılar** | | | |
|---|---|---|---|---|
| **Özellik Vektörü** | **DVM-RO %** | **DVM-KEK %** | **DVM-RO-KEK %** | Ort. Başarı |
| $OV_2$ | 91,50 | 85,87 | 92,62 | 90 |
| $OV_3$ | 90 | 83,12 | 91,75 | 88,29 |
| $OV_4$ | 85,75 | 83,50 | 88,37 | 85,87 |
| $OV_{2A}$ | 80,87 | 81,37 | 86,87 | 83,04 |
| $OV_{3A}$ | 89,50 | 87,12 | 91,62 | 89,41 |
| $OV_{4A}$ | 81 | 79,75 | 86 | 82,25 |
| Ort. Başarı | 86,44 | 83,45 | 89,54 | 86,48 |

Veri Seti-I için her bir sınıflandırıcı ile elde edilen sonuçların başarısı sınıflandırıcıların birlikte kullanılması ile arttığı görülmüştür. DVM sınıflandırıcı yöntemi tek başına kullanıldığında en yüksek başarı %91,12 olarak elde edilmişken, DVM, RO ve K-EK yöntemleri birlikte kullanıldığında elde edilen sonucun %92,62 olarak arttığı görülmüştür. Bireysel olarak sınıflandırma yapıldığında en yüksek %91,12'lik başarı değerine sahip olan DVM kendinden daha düşük sonuca sahip KNN sınıflandırıcısıyla birlikte kullanıldığında başarısının genel olarak düştüğü görülmüştür. Sınıflandırma işlemi, oylama mantığına dayandığından bu da başarının düşmesine sebep olmuştur. DVM ile RO sınıflandırıcıları birlikte kullanıldığında elde edilen sonucun, her iki yönteminde bireysel olarak sahip olduğu sonuçlardan daha başarı olduğu görülmüştür.

Veri Seti-I için sınıflandırıcıların birlikte kullanımı ile elde edilen en yüksek başarı $OV_2$ vektörü ile DVM, RO ve K-EK yöntemlerinin birlikte kullanılmasından elde edilen %92,62'dir. En düşük sonuç ise $OV_{4A}$ vektöründeki özelliklerle DVM ve K-EK yöntemleri birlikte kullanılarak yapılan sınıflandırmada alınan %79,75'dir.

Veri Seti-I için birlikte kullanılmış sınıflandırıcıların başarı ortalamasına bakıldığında %89,54 ile DVM, RO ve K_EK yöntemlerinin en başarılı sınıflandırıcı kombinasyonu olduğu görülmüştür. En başarılı özellik vektörü ise $OV_2$ vektörünün başarı ortalaması %90'dır. Veri Seti-I'in sınıflandırıcıların birlikte kullanımındaki genel başarısı %86,48'dir.

Veri Seti-II üzerinde sınıflandırıcıların birlikte kullanımı ile elde edilen sonuçlar Çizelge 5.9'da verilmiştir. Birlikte kullanım işleminde, her özellik vektörüne göre en yüksek başarıya sahip sınıflandırıcıların çeşitli kombinasyonları oluşturulmuştur. $OV_2$ ve $OV_3$ vektörleri için DVM, NB ve K-EK sınıflandırıcılarının, $OV_4$ ve $OV_{3A}$ vektörleri için DVM, NB ve RO sınıflandırıcılarının, $OV_{2A}$ ve $OV_{3A}$ vektörleri için K-EK, NB ve RO sınıflandırıcılarının kombinasyonları kullanılmıştır.

Çizelge 5.9 Veri Seti-II için sınıflandırıcılar birlikte kullanılarak alınan sonuçlar

| | **Oylama ile Birlikte Kullanılan Sınıflandırıcılar** | | | |
|---|---|---|---|---|
| **Öz.Vektörü** | **DVM-NB %** | **DVM-KEK %** | **DVM-NB-KEK %** | **Ort. Başarı** |
| $OV_2$ | 81,87 | 68 | 87,37 | 79,08 |
| $OV_3$ | 82 | 62,12 | 86,12 | 76,75 |
| **Ort. Başarı** | 81,93 | 65,06 | 86,75 | 77,91 |
| **Öz.Vektörü** | **DVM-NB %** | **RO-DVM %** | **NB-RO-DVM %** | **Ort. Başarı** |
| $OV_4$ | 80,75 | 67,50 | 81,12 | 76,46 |
| $OV_{3A}$ | 83 | 84,12 | 85,50 | 84,21 |
| **Ort. Başarı** | 81,87 | 75,81 | 83,31 | 80,33 |
| **Öz.Vektörü** | **NB-RO %** | **RO-KEK %** | **NB-RO-KEK %** | **Ort. Başarı** |
| $OV_{2A}$ | 87,62 | 84,12 | 89,62 | **87,12** |
| $OV_{4A}$ | 68,37 | 62 | 62,25 | 64,21 |
| **Ort. Başarı** | 78 | 73,06 | 75,94 | 75,66 |

Veri Seti-II için, her bir sınıflandırıcıdan alınan sonuçların başarısının sınıflandırıcıların birlikte kullanım işlemi ile arttığı görülmüştür. Ancak, sınıflandırıcıların bireysel olarak sahip oldukları başarıların, K-EK sınıflandırıcısıyla birlikte kullanıldıklarında genel olarak düştüğü görülmüştür. Bunun nedeni K-EK'un bireysel sınıflandırma yapıldığında tüm özellik vektörleri için en düşük sonuca sahip olmasıdır. NB ile DVM sınıflandırıcıları birlikte kullanıldıkları zaman elde edilen sonucun, NB'nin bireysel olarak sahip olduğu sonuçlarla aynı olduğu görülmüştür. DVM ile bireysel olarak sınıflandırma yapıldığında en yüksek %89,5'lik başarı değerine ulaşılırken diğer sınıflandırıcılar ile birlikte kullanıldığında en yüksek sonuç olarak %87,37 alınmıştır. Diğer sınıflandırıcılar RO sınıflandırıcısı ile birlikte

kullanıldıkları zaman elde edilen sonucun, sınıflandırıcıların bireysel olarak sahip olduğu sonuçlardan daha başarılı olduğu gözlemlenmiştir.

Veri Seti-II için sınıflandırıcıların birlikte kullanımı ile elde edilen en yüksek başarı $OV_{2A}$ vektörü ile K-EK, NB ve RO yöntemlerinin birlikte kullanılması ile elde edilen %89,62'dir. En düşük sonuç ise $OV_{4A}$ vektöründeki özelliklerle RO ve K_EK yöntemlerinin birlikte kullanılması ile alınan başarı %62'dir.

Veri Seti-II için birlikte kullanılmış sınıflandırıcıların başarı ortalamasına bakıldığında %86,75 ile DVM, NB ve K-EK yöntemlerinin en başarılı sınıflandırıcı kombinasyonu olduğu görülmüştür. En başarılı özellik vektörü ise $OV_{2A}$ vektörü olup başarı ortalaması %87,12'dir.Veri Seti-II'in sınıflandırıcıların birlikte kullanım işlemindeki genel başarısı %77,97'dir.

Veri Seti-III üzerinde sınıflandırıcıların birlikte kullanılması ile elde edilen sonuçlar Çizelge 5.10'da verilmiştir. Birlikte kullanım işleminde, her özellik vektörüne göre en yüksek başarıya sahip sınıflandırıcıların çeşitli kombinasyonları oluşturulmuştur. $OV_2$ vektörü için DVM, RO ve K-EK sınıflandırıcılarının, $OV_3$ vektörü için DVM, NB, RO ve K-EK sınıflandırıcılarının, $OV_4$, $OV_{2A}$, $OV_{3A}$ ve $OV_{4A}$ vektörleri için DVM, NB ve RO sınıflandırıcılarının kombinasyonları kullanılmıştır.

Çizelge 5.10 Veri Seti-III için sınıflandırıcılar birlikte kullanılarak alınan sonuçlar

| | **Oylama ile Birlikte Kullanılan Sınıflandırıcılar** | | | |
|---|---|---|---|---|
| **Öz. Vektörü** | **DVM-RO %** | **DVM-KEK %** | **DVM-KEK-RO %** | **Ort. Başarı** |
| $OV_2$ | 83,54 | 81,46 | 81,67 | 82,22 |
| **Öz. Vektörü** | **DVM-NB %** | **DVM-RO %** | **DVM- KEK-NB %** | **Ort. Başarı** |
| $OV_3$ | 81,04 | 83,37 | **88,12** | 84,18 |
| **Öz. Vektörü** | **DVM-NB %** | **DVM-RO %** | **DVM-NB-RO %** | **Ort. Başarı** |
| $OV_4$ | 77,71 | 85,42 | 80,62 | 81,25 |
| **Öz. Vektörü** | **NB-RO %** | **DVM-RO %** | **DVM-NB-RO %** | **Ort. Başarı** |
| $OV_{2A}$ | 82,92 | 86,46 | 84,17 | 84,52 |
| $OV_{3A}$ | 85,42 | 84,37 | 86,25 | 85,35 |
| $OV_{4A}$ | 81,67 | 86,25 | 81,25 | 83,06 |
| **Ort. Başarı** | 83,34 | 85,69 | 83,89 | 84,31 |

Veri Seti-III için her bir sınıflandırıcı ile elde edilen sonuçların başarısı sınıflandırıcıların birlikte kullanılması ile işlemi ile arttığı görülmüştür. Ancak, DVM bireysel olarak sınıflandırma yapıldığında en yüksek %92,08'lik başarı değerine sahip iken, bulunduğu tüm sınıflandırıcıların birlikte kullanılmasında en yüksek sonuç olarak %88,12 alınmıştır. Diğer sınıflandırıcılar RO sınıflandırıcısı ile birlikte kullanıldığı zaman elde edilen sonucun, sınıflandırıcıların bireysel olarak sahip olduğu sonuçlardan daha başarı olduğu görülmüştür.

Veri Seti-III için sınıflandırıcıların birlikte kullanım işlemi ile elde edilen en yüksek başarı $OV_3$ vektörü ile DVM, K-EK ve NB yöntemlerinin birlikte kullanılması ile elde edilen %88,12'dir. En düşük sonuç ise $OV_4$ vektöründeki özelliklerle DVM ve NB yöntemleri birlikte kullanılarak yapılan sınıflandırmada alınan %77,71'dir.

Veri Seti-III için birlikte kullanılan sınıflandırıcıların başarı ortalamasına bakıldığında en başarılı özellik vektörü $OV_{3A}$ vektörünün başarı ortalaması %85,35'dir.

Veri Seti-III'ün sınıflandırıcıların birlikte kullanımındaki genel başarısı %83,43'tür.

# 6. SONUÇ

Bu çalışmanın amacı, dokümanların yazarlarını, türlerini ve yazarların cinsiyetini belirlemektir. Dokümanların yazarlarını, türlerini ve yazarların cinsiyetini belirlemede kullanılmak amacıyla üç farklı veri seti oluşturulmuştur. 3 adet veri seti için 2-gram, 3-gram ve 4-gram'lardan oluşan özellik vektörleri çıkarılmış, daha sonra bu üç özellik vektöründen türetilmiş, özellik sayısı daha az olan vektörlerde kullanılarak, toplam 6 özellik vektörü ile sınıflandırma işlemi gerçekleştirilmiştir. Ayrıca sınıflandırmadaki başarı oranını arttırmak için başarılı sonuç veren sınıflandırıcılar birlikte kullanılmıştır. Birlikte kullanılma işlemi, başarısı en yüksek olan sınıflandırıcıların bir araya getirilerek oylama sonucunda dokümanın hangi sınıftan olabileceğine karar veren bir işlemdir.

Sınıflandırma yapılırken Naive Bayes, Destek Vektör Makinesi, K-En Yakın Komşuluk, Rastgele Orman yöntemleri ve bizim geliştirdiğimiz ng_ind yöntemi kullanılmıştır.

Veri Seti-I için bu yöntemlerden alınan en başarılı sonuçlar ve bu sonuçları veren özellik vektörleri Çizelge 6.1'de gösterilmektedir. Ortalama başarı ile kastedilen ilgili sınıflandırıcının 6 farklı özellik vektöründen alınmış olan sonuçlarının ortalamasıdır. Ng_id yönteminde 3 n-gram modeli ve 5 farklı özelik vektörü ile toplamda 15 özellik vektörünün ortalaması alınarak elde edilmiştir. Ng_ind yöntemi genelde NB'ye göre daha başarılı, buna karşılık özellik sayısı 100 alındığında ve 4-gram modeli kullanıldığında en başarılı sınıflandırıcı olmuştur.

Çizelge 6.1 Veri Seti-I için yöntemlerden alınan en başarılı sonuçlar ve bu sonuçları veren özellik vektörleri

| **Sınıflandırıcı** | **Özellik Vektörü** | **Özellik Vektörü Başarısı** | **Ortalama Başarı** |
|---|---|---|---|
| **DVM** | $OV_2$ | % 91,12 | % 86,04 |
| **K-EK** | $OV_{3A}$ | % 86,87 | % 83,81 |
| **NB** | $OV_{3A}$ | %85,12 | % 75,81 |
| **RO** | $OV_{3A}$ | % 89,875 | % 86,02 |
| **NG_IND** | V4-100 | % 91,25 | % 82,08 |

Veri Seti-II için ng-ind yönteminden alınan en başarılı sonuç ve bu sonuçları veren özellik vektörleri Çizelge 6.2'de gösterilmektedir.

Çizelge 6.2 Veri Seti-II için yöntemlerden alınan en başarılı sonuçlar ve bu sonuçları veren özellik vektörleri

| **Sınıflandırıcı** | **Özellik Vektörü** | **Özellik Vektörü Başarısı** | **Ortalama Başarı** |
|---|---|---|---|
| **DVM** | $OV_2$ | % 89,50 | % 76,92 |
| **K-EK** | $OV_{2A}$ | % 84,13 | % 68,27 |
| **NB** | $OV_{2A}$ | % 86,63 | % 79,81 |
| **RO** | $OV_{2A}$ | % 87,75 | % 71,75 |
| **NG_IND** | V4-400 | %77,5 | % 73,75 |

Ng-ind yöntemi, özellik sayısı 400 alındığında ve 4-gram modeli seçildiğinde en başarılı sonucu %77,5 olarak vermesine rağmen, cinsiyet belirlemeye göre yazar tanımada başarısız olmuştur. Yazar tanımada sınıflandırıcıya bireysel olarak baktığımızda diğer sınıflandırıcılara göre başarısız olmasına karşılık genel başarı ortalamasına bakıldığında K-EK ve RO'ya göre daha başarılı sonuç vermiştir.

Çizelge 6.3'de görüldüğü gibi ng_ind yöntemi dokümanın türünü belirlemede yazar tanıma ve cinsiyet belirlemeye göre daha yüksek bir başarı göstermiştir. Aynı zamanda diğer sınıflandırıcılar ile karşılaştırıldığında da başarısı yüksektir. Özellik sayısı 400 ve 500 alındığında 3 veya 4 gram modeli kullanıldığında %93,75'lik doğru sınıflandırma başarısı vermiştir. Ortalama genel başarıya bakıldığında DVM'den sonra 2. başarılı sınıflandırıcı yöntemidir.

Çizelge 6.3 Veri Seti-III için yöntemlerden alınan en başarılı sonuçlar ve bu sonuçları veren özellik vektörleri

| **Sınıflandırıcı** | **Özellik Vektörü** | **Özellik Vektörü Başarısı** | **Ortalama Başarı** |
|---|---|---|---|
| **DVM** | $OV_3$ | % 92,08 | % 85,45 |
| **K-EK** | $OV_{3A}$ | % 83,54 | % 74,58 |
| **NB** | $OV_{3A}$ | % 87,29 | % 78,43 |
| **RO** | $OV_{3A}$ | % 83,33 | % 80,79 |
| **NG_IND** | V3-400 ,V3-500, V4-400 | %93,75 | %83,75 |

Veri Seti-I için başarısı en yüksek olan DVM, RO ve K-EK sınıflandırıcıları birlikte kullanarak elde edilen sonuçlardan en başarılı olanlar Çizelge 6.4'de verilmiştir. DVM, RO ve K-EK sınıflandırıcıları birlikte kullanıldığında $OV_2$ vektörü ile %92,62'lik başarı elde

edilmiştir. Oysaki bu sınıflandırıcılar bireysel olarak kullanıldıklarında sırasıyla en başarılı sınıflandırma sonucu olarak %91,12, %89,37 ve %86,87 vermişlerdir.

Çizelge 6.4 Veri Seti -I için sınıflandırıcılar birlikte kullanarak elde edilen en iyi sonuçlar ve özellik vektörleri

| Oylama ile Birlikte Kullanılan Sınıflandırıcılar | Özellik Vektörü | Özellik Vektörü Başarısı | Ortalama Başarı |
|---|---|---|---|
| **DVM-RO** | $OV_2$ | % 91,5 | % 86,44 |
| **DVM-KEK** | $OV_{2A}$ | % 81,37 | % 83,45 |
| **DVM-RO-KEK** | $OV_2$ | % 92,62 | % 89,54 |

Veri Seti-II için başarısı en yüksek olan DVM, RO, NB ve K-EK sınıflandırıcıları birlikte kullanılarak elde edilen sonuçlardan en başarılı olanlar Çizelge 6.5'de gösterilmektedir.

Çizelge 6.5 Veri Seti -II için sınıflandırıcılar birlikte kullanarak elde edilen en iyi sonuçlar ve özellik vektörleri

| Oylama ile Birlikte Kullanılan Sınıflandırıcılar | Özellik Vektörü | Özellik Vektörü Başarısı | Ortalama Başarı |
|---|---|---|---|
| **DVM-NB** | $OV_3$ | % 82 | % 81,93 |
| **DVM-KEK** | $OV_2$ | % 68 | % 65,06 |
| **RO-DVM** | $OV_{3A}$ | % 84,12 | % 75,81 |
| **NB-RO** | $OV_{2A}$ | % 87,62 | % 78 |
| **RO-KEK** | $OV_{2A}$ | % 84,12 | % 73,06 |
| **DVM-NB-KEK** | $OV_2$ | % 87,37 | % 86,75 |
| **NB-RO-KEK** | $OV_{2A}$ | % 89,62 | % 75,94 |
| **NB-RO-DVM** | $OV_{3A}$ | % 85,50 | % 83,31 |

Veri Seti-II için oluşturulan sınıflandırıcıların birlikte kullanım işleminde %89,62'lik en iyi sonuç NB, RO ve K-EK yöntemlerinin kombinasyonu tarafından alınmıştır ve $OV_{2A}$ en başarılı özellik vektördür.

Veri Seti-III için başarısı en yüksek olan DVM, RO, NB ve KNN sınıflandırıcıları birlikte kullanılarak elde edilen sonuçlardan en başarılı olanlar Çizelge 6.6'de gösterilmektedir.

Çizelge 6.6 Veri Seti -III için sınıflandırıcılar birlikte kullanarak elde edilen en iyi sonuçlar ve özellik vektörleri

| Oylama ile Birlikte Kullanılan Sınıflandırıcılar | Özellik Vektörü | Özellik Vektörü Başarısı | Ortalama Başarı |
|---|---|---|---|
| **DVM-NB** | $OV_3$ | % 81,04 | % 79,37 |
| **DVM-KEK** | $OV_2$ | % 81,46 | % 81,46 |
| **RO-DVM** | $OV_{2A}$ | % 86,46 | % 84,50 |
| **NB-RO** | $OV_{3A}$ | % 85,42 | % 83,34 |
| **DVM-NB-KEK** | $OV_3$ | % 88,12 | % 88,12 |
| **DVM-RO-KEK** | $OV_2$ | % 81,67 | % 81,67 |
| **NB-RO-DVM** | $OV_{3A}$ | % 86,25 | % 83,07 |

Veri Seti-III %88,12'lik en iyi sonucunu DVM, NB ve K-EK sınıflandırıcılarının birlikte kullanıldığında ve özellik vektörü $OV_3$ kullanıldığında vermiştir.

Sınıflandırıcılar birlikte kullanıldığında en yüksek başarıları sırasıyla Veri Seti-I'de yani cinsiyet belirlemede, Veri Seti-II yazar tanımada ve Veri Seti-III cinsiyet belirleme de vermiştir.

Yapılan deneyler sonucunda, sınıflandırma yöntemlerinin birlikte kullanılmasının ve özellik azaltılarak yapılan sınıflandırma işleminin, yöntemlerin bireysel olarak uygulandığından daha iyi sonuçlar verdiği gözlemlenmiştir. Sınıflandırıcıların birlikte kullanılması işleminin de özellik azaltılarak yapılan sınıflandırma işleminden daha başarılı olduğu görülmüştür. Bu Veri Setleri üzerinde sınıflandırıcıların birlikte kullanılmasının çok başarılı bir yöntem olduğu söylenebilir.

Geliştirilen ng_ind yöntemi cinsiyet belirlemede %91,25, yazar tanımada %78'lik başarıyı 4-gram modeli ile verirken, tür belirlemede %93,75'lik başarıyı 3-gram modelini kullanarak elde etmiştir.

Ng_ind yöntemi Türkçe'de daha önce yapılmış çalışmaya göre daha düşük sonuç vermesine rağmen sınıf sayısı daha fazla olduğundan daha başarılı olarak görülmektedir.

Kısaca bu çalışma, n-gram yöntemi kullanılarak bugüne kadar yapılan Türkçe dokümanların yazarlarının, türlerinin ve yazarların cinsiyetlerinin belirlenmesindeki en kapsamlı çalışmadır.

**KAYNAKLAR**

Argamon S., Fine J., Koppel M., Shimoni A. R., (2003), "Automatically Categorizing Written Texts by Author Gender", 1Dept. of Computer Science, Bar-Ilan University Ramat Gan 52900, Israel.

Basu A., Watters C., and Shepherd M., (2002), "Support Vector Machines for Text Categorization", Faculty of Computer Science Dalhousie University Halifax, Nova Scotia, Canada.

Bennett W. R., (1976), "Scientific and engineering computer", Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

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.

Demiriz A., (2006), Karar Ağaçları İle Sınıflandırma.

Diri, B., Amasyalı, M.F., (2003), "Automatic Author Detection for Turkish Text", 13th International Conference on Neural Information Processing, Turkey.

Diri, B., Amasyalı, M.F., Türkoğlu, F., (2006), "Farklı Özellik Vektörleri ile Türkçe Dokümanların Yazarlarının Belirlenmesi", Yıldız Teknik Üniversitesi, Turkey.

Diri, B., Amasyalı, M.F., (2006), "Automatic Turkish Text Categorization in Terms of Author, Genre and Gender", Yıldız Teknik Üniversitesi, Turkey.

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

Fung, G. ve Mangasarian, O. L., (2002), Incremental Support Vector Machine Classification Second SIAM International Conference on Data Mining.

Fakotakis N., Stamatatos E., Kokkinakis G., (2000), "Automatic text categorization in terms of genre and author", Computational Linguistics, (pp. 471-495).

Hall, M.A., (1999), Correlation-based feature selection machine learning. PhD thesis, Department of Computer Science, University of Waikato, New Zealand.

Kahraman F., Çapar A., Ayvacı A. , Demirel H., Gökmen M., (2004), Elyazısı Verileri Üzerinde YSA ve DVM'nin Sınıflandırma Başarımlarının Karşılaştırılması, İTÜ Bilişim Enstitüsü, Bilgisayar Bilimleri Bölümü, İstanbul.

Kepenekçi B., Akar G. B., (2005),"Destekçi Vektör Makinesi İle Yüz Sınıflandırma", Orta Doğu Teknik Üniv., Elektrik ve Elektronik Mühendisliği Bölümü, Ankara.

Kivinen, J., M. Warmuth, (1997), "Additive versus exponentiated gradient updates for linear prediction", Information and Computation, 132, 1, (pp. 1-64).

Kneer, S., Personnaz, L. and Dreyfus, G., (1990), Single-layer learning revisited: "A stepwise procedure for building and training a neural network Neurocomputing", Algorithms, Architectures and Applications NATO ASI Series.

Koppel M., Argamon S., Shimoni A. R., (2003), "Automatically Categorizing Written Texts by Author Gender", 1Dept. of Computer Science, Bar-Ilan University Ramat Gan 52900, Israel.

Ko Y., Seo J., (2002), "Text Categorization using Feature Projections", Department of Computer Science, Sogang University.

Kramer K., Luo T., Goldgof D. B., Hall L. O., (2004), "Recognizing Plankton Images from the Shadow Image Particle Profiling Evaluation Recorder", Dept. of Computer Science & Engineering,University of South Florida.

Müller K. R., Mika S., Ratsch G., Tsuda K., Schölkopf B., (2001), "An Introduction to Kernel-Based Learning Algorithms", IEEE Trans. On Neural Networks, vol. 12, no. 2.

Nowson S., Oberlander J., (2006), "Openness and gender in personal weblogs", School of Informatics, University of Edinburgh,2 Buccleuch Place, Edinburgh, EH8 9LW

Özgür A., (2002), "Supervised And Unsupervised Machine Learning Techniques For Text Document Categorization" ,Boğaziçi Üniversitesi, İstanbul.

Peng F., Keselj V., Cerconey N., Thomasy C., (2003), "N-Gram-Based Author Profiles For Authorship Attribution", yFaculty of Computing Science, Dalhousie University, Canada.

Peng F., Wang S., Schuurmans D., (2003), "Language and Task Independent Text Categorization with Simple Language Models", School of Computer Science, University of Waterloo.

Peng F., Schuurmans D., (2003), "Combining Naive Bayes and N-gram Language Models for Test Classification", School of Computer Science, University of Waterloo.

Platt, J., Cristianini, N., and Shawe, J., (2000), "Large magrin DAGs for multiclass classification", In Advance in Neural Information Processing System volume 12. MIT Pres.

Salton, G., Yang, C. ve Wong, A., (1975), "A vector-space model for automatic indexing", Communications of the ACM, Vol. 18, No. 11, pp. 613–620

Schuurmans D., Peng F., Keselj V., Wang S., (2003), "Language Independent Authorship Attribution using Character Level Language Models", Dalhousie University, Canada.

Vapnik, V. N., ( "An overview of statistical learning theory", IEEE Transactions on Neural Networks.

Weston, J., Watkins, C., 1999), "Support vector machines for multiclass", Proceedings of ESANN.99, Belgium.

Zipf, George K., (1949), "Human Behavior and the Principle of Least Effort", an Introduction to Human Ecology, Addison-Wesley, Reading, Mass.

## İNTERNET KAYNAKLARI

**[1]** http://torch.cs.dal.ca/~nlp/abstracts/2005-12-05.html

**[2]** http://www.cs.waikato.ac.nz/~ml/weka/

**[3]** http://www.hurriyet.com.tr

**[4]** http://www.milliyet.com.tr

**[5]** http://www.aksam.com.tr

**[6]** http://www.sabah.com.tr

**[7]** http://www.uk.research.att.com/facedatabase.html.(A. L. Cambridge. The olivetti research ltd yüz veritabanı.)

**[8]** http://www.cs.itu.edu.tr/~gunduz/courses/verimaden/

**[9]** http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm#overview

**[10]** http://www.isletme.istanbul.edu.tr/surekli_yayinlar/dergiler/nisan2000/1.htm

## ÖZGEÇMİŞ

| | | |
|---|---|---|
| Doğum tarihi | 01.01.1980 | |
| Doğum yeri | Malatya | |
| Lise | 1995-1998 | Üsküdar Fen Lisesi |
| Lisans | 1998-2002 | İstanbul Üniversitesi Mühendislik Fak.<br>Bilgisayar Mühendisliği Bölümü |
| Yüksek Lisans | 2003-2006 | Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü<br>Bilgisayar Müh. Anabilim Dalı,<br>Bilgisayar Müh. Programı |

**Çalıştığı Kurumlar**

| | |
|---|---|
| 2003-2004 | IAS Yazılım A.Ş |
| 2004-2004 | Bilge Adam Bilgi Teknolojileri |
| 2004-Devam ediyor | Akbank Bilgi Teknolojileri |