

**YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**YAPAY BAĞIŞIKLIK SİSTEMLERİNİ KULLANARAK
TÜRKÇE METİNLERDE TÜR, YAZAR ve CİNSİYET
TANIMA**

Bilgisayar Mühendisi Zafer KABAN

**FBE Bilgisayar Mühendisliği Anabilim Dalında
Hazırlanan**

YÜKSEK LİSANS TEZİ

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

İstanbul, 2008

İÇİNDEKİLER

KISALTMA LİSTESİ.....	v
ŞEKİL LİSTESİ.....	vi
ÇİZELGE LİSTESİ	ix
ÖNSÖZ	xi
ÖZET	xii
ABSTRACT	xiii
1. GİRİŞ	1
2. YAZAR TANIMA, TÜR ve CİNSİYET BELİRLEME ALANINDA YAPILMIŞ ÖNCEKİ ÇALIŞMALAR	4
2.1 Yazar Tanıma ile İlgili Daha Önce Yapılan Çalışmalar	4
2.2 Doküman Türünü Belirleme ile İlgili Daha Önce Yapılan Çalışmalar	7
2.3 Doküman Yazarının Cinsiyetini Belirleme ile İlgili Daha Önce Yapılan Çalışmalar	8
3. SINIFLANDIRMA YÖNTEMLERİ	9
3.1 Eğiticili Sınıflandırma Yöntemleri	9
3.1.1 Naive Bayes (NB).....	9
3.1.2 K-En Yakın Komşuluk (K-EYK).....	10
3.1.3 Destek Vektör Makinesi (DVM).....	11
3.1.4 Rastgele Orman (RO)	13
3.2 Yapay Bağışıklık Sistemleri (YBS).....	14
3.2.1 Doğal Bağışıklık Sistemi	14
3.2.1.1 Bağışıklık Sistemi Organları	14
3.2.1.2 Bağışıklık Sistemi Hücreleri.....	15
3.2.1.3 Bağışıklık Sisteminin Çalışması	17
3.2.1.4 Klonal Seçim Prensibi.....	17
3.2.1.5 Kendinden Olan/Olmayan (Self/Nonself) Ayırımı	18
3.2.2 Yapay Bağışıklık Sistemleri.....	19
3.2.2.1 Klonal Seçim Algoritması	19
3.2.2.2 Negatif Seçim Algoritması	20

3.2.2.3	Şekil Uzayı ve ARB Kavramları	21
3.2.2.4	AIRS (Artificial Immune Recognition System) Algoritması.....	22
3.2.2.5	AIRS Algoritması Türevleri (AIRSv1, AIRSv2 ve AISv2Parallel)	24
3.3	Boyut İndirgeme.....	24
3.3.1	YTU Yöntemi.....	25
3.3.2	Korelasyon Tabanlı Özellik Seçici (KTÖS/Correlation-based Feature Selection-CFS)	27
4.	VERİ KÜMELERİ	28
4.1	Cinsiyet Belirleme Çalışmasında Kullanılan Veri Kümesi (VKc).....	28
4.2	Tür Belirleme Çalışmasında Kullanılan Veri Kümesi (VKt).....	28
4.3	Yazar Tanıma Çalışmasında Kullanılan Veri Kümeleri	28
4.3.1	Yazar Tanıma için Karma Türlerde Yazan Yazarlardan Oluşturulan Veri Kümeleri.....	29
4.3.2	Aynı Türde Yazan Yazarlar için Veri Kümeleri	29
5.	ÖZELLİK VEKTÖRLERİ	32
5.1	İstatistiksel Özellikler (V_{IST}).....	32
5.2	Kelime Zenginliğine Dayalı Özellikler (V_{KZ}).....	33
5.3	Dilbilgisi Özellikleri (V_{DB})	34
5.4	İşlevsel Kelimeler ($V_{İK}$)	35
5.5	Karakter N-Gram'ları (V_{nG})	36
5.6	Kelime Kök veya Gövdeleri ($V_{KÖK}$ veya $V_{GÖVDE}$)	36
5.7	Boyut İndirgeme Yoluyla Oluşturulan Vektörler.....	37
6.	DENEYSEL SONUÇLAR.....	39
6.1	Cinsiyet Belirleme İçin Deneysel Sonuçlar (VKc)	39
6.2	Tür Belirleme İçin Deneysel Sonuçlar (VKt).....	50
6.3	Yazar Tanıma İçin Deneysel Sonuçlar	62
6.3.1	Aynı Türde Yazar Tanıma İçin Deneysel Sonuçlar.....	62
6.3.1.1	Aynı Türde Yazar Tanıma İçin Deneysel Sonuçlar ($V_{K_{ya10}}$).....	62
6.3.1.2	Aynı Türde Yazar Tanıma İçin Deneysel Sonuçlar ($V_{K_{ya8}}$)	73

6.3.2	Farklı Türde Yazar Tanıma İçin Deneysel Sonuçlar	84
6.3.2.1	Farklı Türde Yazar Tanıma İçin Deneysel Sonuçlar (VK _{yk20})	84
6.3.2.2	Farklı Türde Yazar Tanıma İçin Deneysel Sonuçlar (VK _{yk18})	90
6.3.2.3	Farklı Türde Yazar Tanıma İçin Deneysel Sonuçlar (VK _{yk16})	96
6.3.2.4	Farklı Türde Yazar Tanıma İçin Deneysel Sonuçlar (VK _{yk14})	102
6.3.2.5	Farklı Türde Yazar Tanıma İçin Deneysel Sonuçlar (VK _{yk12})	108
6.4	YTU Boyut Azaltma Yönteminin Farklı Bir Dil için Kullanımı	115
6.4.1	Reuters Veri Kümesi.....	115
6.4.2	Özellik Vektörü	116
6.4.3	Deneysel Sonuçlar	116
7.	SONUÇ	119

KISALTIMA LİSTESİ

AADTT	Automatic Author Detection for Turkish Text
AIRS	Artificial Immune Recognition Systems
AIRS2P	Artificial Immune Recognition Systems 2 Parallel
AP	Associated Pres
APC	Antigen Presenting Cells
ARB	Artificial Recognition Balls
ASH	Antijen Sunan Hücreler
BNC	British National Corpus
CFS	Correlation-Based Feature Selection
DAA	Doğrusal Ayırtedan Analizi
DVM	Destek Vektör Makinesi
K-EYK	K-En Yakın Komşuluk
KNN	K-Nearest Neighborhood
KTÖS	Korelasyon Tabanlı Özellik Seçme
LDA	Linear Discriminant Analysis
MHC	Major Histocompatibility Complex
NA	Not Available
NB	Naive Bayes
PCA	Principal Component Analysis
RF	Random Forests
RO	Rastgele Orman
SVM	Support Vector Machine
TBA	Temel Bileşen Analizi
TKV	Türkçe Kelime Veritabanı
TREC	Text Retrieval Conference
VK	Veri Kümesi
YBS	Yapay Bağışıklık Sistemleri

ŞEKİL LİSTESİ

Şekil 3.3-1 K-En Yakın Komşuluk.....	10
Şekil 3.2 Destek Vektör Makinesi [7].....	12
Şekil 3.3 Bağışıklık Sistemi Organları (Alataş vd., 2005).....	15
Şekil 3.4 Klonal Seçim Prensibi (Parmaksızoğlu, 2005)	18
Şekil 3.5 Klonal Seçim Algoritması Adımları (Engin ve Döyen, 2004)	20
Şekil 3.6 Negatif Seçim Algoritması (Engin ve Döyen, 2004)	21
Şekil 3.7 Şekil Uzayı (De Castro, 1999).....	22
Şekil 6.1 VKc için İstatistiksel Özellik Vektörleri Başarı Grafiği	46
Şekil 6.2 VKc için İşlevsel KelimelerÖzellik Vektörleri Başarı Grafiği.....	46
Şekil 6.3 VKc için 2-gram Özellik Vektörleri Başarı Grafiği.....	47
Şekil 6.4 VKc için 3-gram Özellik Vektörleri Başarı Grafiği.....	47
Şekil 6.5 VKc için Kök Özellik Vektörleri Başarı Grafiği	48
Şekil 6.6 VKc için Gövde Özellik Vektörleri Başarı Grafiği.....	48
Şekil 6.7 VKc için Boyut İndirgeme Yapılmayan Özellik Vektörleri Başarı Grafiği.....	49
Şekil 6.8 VKc için En Başarılı Özellik Vektörleri Başarı Grafiği.....	49
Şekil 6.9 VK _t için İstatistiksel Özellik Vektörleri Başarı Grafiği	58
Şekil 6.10 VK _t için İşlevsel Kelimeler Özellik Vektörleri Başarı Grafiği.....	58
Şekil 6.11 VK _t için 2-gram Özellik Vektörleri Başarı Grafiği.....	59
Şekil 6.12 VK _t için 3-gram Özellik Vektörleri Başarı Grafiği.....	59
Şekil 6.13 VK _t için Kök Özellik Vektörleri Başarı Grafiği	60
Şekil 6.14 VK _t için Gövde Özellik Vektörleri Başarı Grafiği.....	60
Şekil 6.15 VK _t için Boyut İndirgeme Yapılmayan Özellik Vektörleri Başarı Grafiği.....	61
Şekil 6.17 VK _{ya10} için İstatistiksel Özellik Vektörleri Başarı Grafiği	69
Şekil 6.18 VK _{ya10} için İşlevsel Kelimeler Özellik Vektörleri Başarı Grafiği.....	69
Şekil 6.19 VK _{ya10} için 2-gram Özellik Vektörleri Başarı Grafiği.....	70
Şekil 6.20 VK _{ya10} için 3-gram Özellik Vektörleri Başarı Grafiği.....	70
Şekil 6.21 VK _{ya10} için Kök Özellik Vektörleri Başarı Grafiği	71
Şekil 6.22 VK _{ya10} için Gövde Özellik Vektörleri Başarı Grafiği	71

Şekil 6.23 VK_{ya10} için Boyut İndirgeme Yapılmayan Özellik Vektörleri Başarı Grafiği.....	72
Şekil 6.24 VK_{ya10} için En Başarılı Özellik Vektörleri Başarı Grafiği	72
Şekil 6.25 VK_{ya8} için İstatistiksel Özellik Vektörleri Başarı Grafiği	80
Şekil 6.26 VK_{ya8} için İşlevsel Kelimeler Özellik Vektörleri Başarı Grafiği.....	80
Şekil 6.27 VK_{ya8} için 2-gram Özellik Vektörleri Başarı Grafiği.....	81
Şekil 6.28 VK_{ya8} için 3-gram Özellik Vektörleri Başarı Grafiği.....	81
Şekil 6.29 VK_{ya8} için Kök Özellik Vektörleri Başarı Grafiği	82
Şekil 6.30 VK_{ya8} için Gövde Özellik Vektörleri Başarı Grafiği.....	82
Şekil 6.31 VK_{ya8} için Boyut İndirgeme Uygulanmayan Özellik Vektörleri Başarı Grafiği	83
Şekil 6.32 VK_{ya8} için En Başarılı Özellik Vektörleri Başarı Grafiği.....	83
Şekil 6.33 VK_{yk20} için İstatistiksel Özellik Vektörleri Başarı Grafiği	86
Şekil 6.34 VK_{yk20} için İşlevsel Kelimeler Özellik Vektörleri Başarı Grafiği	86
Şekil 6.35 VK_{yk20} için 2-gram Özellik Vektörleri Başarı Grafiği	87
Şekil 6.36 VK_{yk20} için 3-gram Özellik Vektörleri Başarı Grafiği	87
Şekil 6.37 VK_{yk20} için Kök Özellik Vektörleri Başarı Grafiği.....	88
Şekil 6.38 VK_{yk20} için Gövde Özellik Vektörleri Başarı Grafiği	88
Şekil 6.39 VK_{yk20} için Boyut İndirgeme Yapılmayan Özellik Vektörleri Başarı Grafiği.....	89
Şekil 6.40 VK_{yk20} için En Başarılı Özellik Vektörleri Başarı Grafiği	89
Şekil 6.41 VK_{yk18} için İstatistiksel Özellik Vektörleri Başarı Grafiği	92
Şekil 6.42 VK_{yk18} için İşlevsel Kelimeler Özellik Vektörleri Başarı Grafiği	92
Şekil 6.43 VK_{yk18} için 2-gram Özellik Vektörleri Başarı Grafiği	93
Şekil 6.44 VK_{yk18} için 3-gram Özellik Vektörleri Başarı Grafiği	93
Şekil 6.45 VK_{yk18} için Kök Özellik Vektörleri Başarı Grafiği.....	94
Şekil 6.46 VK_{yk18} için Gövde Özellik Vektörleri Başarı Grafiği	94
Şekil 6.47 VK_{yk18} için Boyut İndirgeme Yapılmayan Özellik Vektörleri Başarı Grafiği.....	95
Şekil 6.48 VK_{yk18} için En Başarılı Özellik Vektörleri Başarı Grafiği	95
Şekil 6.49 VK_{yk16} için İstatistiksel Özellik Vektörleri Başarı Grafiği	98
Şekil 6.50 VK_{yk16} için İşlevsel Kelimeler Özellik Vektörleri Başarı Grafiği	98
Şekil 6.51 VK_{yk16} için 2-gram Kelimeler Özellik Vektörleri Başarı Grafiği.....	99

Şekil 6.52 VK_{yk16} için 3-gram Kelimeler Özellik Vektörleri Başarı Grafiği	99
Şekil 6.53 VK_{yk16} için Kök Özellik Vektörleri Başarı Grafiği	100
Şekil 6.54 VK_{yk16} için Gövde Özellik Vektörleri Başarı Grafiği	100
Şekil 6.55 VK_{yk16} için Boyut İndirgeme Yapılmayan Özellik Vektörleri Başarı Grafiği.....	101
Şekil 6.56 VK_{yk16} için En Başarılı Özellik Vektörleri Başarı Grafiği	101
Şekil 6.57 VK_{yk14} için İstatistiksel Özellik Vektörleri Başarı Grafiği	104
Şekil 6.58 VK_{yk14} için İşlevsel Kelimeler Özellik Vektörleri Başarı Grafiği	104
Şekil 6.59 VK_{yk14} için 2-gram Özellik Vektörleri Başarı Grafiği	105
Şekil 6.60 VK_{yk14} için 3-gram Özellik Vektörleri Başarı Grafiği	105
Şekil 6.61 VK_{yk14} için Kök Özellik Vektörleri Başarı Grafiği	106
Şekil 6.62 VK_{yk14} için Gövde Özellik Vektörleri Başarı Grafiği	106
Şekil 6.63 VK_{yk14} için Boyut İndirgeme Yapılmayan Özellik Vektörleri Başarı Grafiği.....	107
Şekil 6.64 VK_{yk14} için En Başarılı Özellik Vektörleri Başarı Grafiği	107
Şekil 6.65 VK_{yk12} için İstatistiksel Özellik Vektörleri Başarı Grafiği	110
Şekil 6.66 VK_{yk12} için İşlevsel Kelimeler Özellik Vektörleri Başarı Grafiği	110
Şekil 6.67 VK_{yk12} için 2-gram Özellik Vektörleri Başarı Grafiği	111
Şekil 6.68 VK_{yk12} için 3-gram Özellik Vektörleri Başarı Grafiği	111
Şekil 6.69 VK_{yk12} için Kök Özellik Vektörleri Başarı Grafiği.....	112
Şekil 6.70 VK_{yk12} için Gövde Özellik Vektörleri Başarı Grafiği	112
Şekil 6.71 VK_{yk12} için Boyut İndirgeme Yapılmayan Özellik Vektörleri Başarı Grafiği.....	113
Şekil 6.72 VK_{yk12} için En Başarılı Özellik Vektörleri Başarı Grafiği	113
Şekil 6.73 V_{3G_YTU} Özellik Vektörü ile AIRS1 ve KNN Algoritmalarının Başarı Grafiği....	114
Şekil 6.74 V_{3G_YTU} Özellik Vektörü ile YBS Algoritmalarının Başarı Grafiği.....	114
Şekil 6.75 VK_{TEM_YTU} ve $V_{GÖVDE_YTU}$ Özellik Vektörlerinin Tür Tanıma Başarı Grafiği (%)	117

ÇİZELGE LİSTESİ

Çizelge 3.1: Örnek veri kümesi	26
Çizelge 4.1: Cinsiyet Belirleme Çalışmasında Kullanılan Yazarlar	28
Çizelge 4.2: Karma Türde Yazar Tanıma Çalışmasındaki Veri Kümeleri	30
Çizelge 4.3: Aynı Türde Yazar Tanıma Çalışmasında Kullanılan Yazarlar	31
Çizelge 5.1: Boyut İndirgenmiş Özellik Vektörleri.....	38
Çizelge 6.1: VKc için Genel Başarı Oranları (%)	41
Çizelge 6.2: VKc için Boyut İndirgeme Yapılmayan Özellik Vektörleri İçin Başarı Oranları.....	42
Çizelge 6.3: VKc için CFS ile Boyut İndirgenmiş Özellik Vektörleri İçin Başarı Oranları.....	42
Çizelge 6.4: VKc için YTU ile Boyut İndirgenmiş Özellik Vektörleri İçin Başarı Oranları.....	43
Çizelge 6.5: VKc için İstatistiksel Özellik Vektörleri İçin Başarı Oranları.....	43
Çizelge 6.6: VKc için İşlevsel Kelimeler Özellik Vektörleri İçin Başarı Oranları	44
Çizelge 6.7: VKc için 2-gram Özellik Vektörleri İçin Başarı Oranları	44
Çizelge 6.8: VKc için 3-gram Özellik Vektörleri İçin Başarı Oranları	44
Çizelge 6.9: VKc için Kök Özellik Vektörleri İçin Başarı Oranları.....	45
Çizelge 6.10: VKc için Gövde Özellik Vektörleri İçin Başarı Oranları	45
Çizelge 6.11: VK _i için Genel Başarı Oranları (%)	53
Çizelge 6.12: VK _i için Boyut İndirgeme Yapılmayan Özellik Vektörleri İçin Başarı Oranları (%).....	54
Çizelge 6.13: VK _i için CFS ile Boyut İndirgenmiş Özellik Vektörleri İçin Başarı Oranları (%)	54
Çizelge 6.14: VK _i için YTU ile Boyut İndirgenmiş Özellik Vektörleri İçin Başarı Oranları (%)	55
Çizelge 6.15: VK _i için İstatistiksel Özellik Vektörleri İçin Başarı Oranları (%)	55
Çizelge 6.16: VK _i için İşlevsel Kelimeler Özellik Vektörleri İçin Başarı Oranları (%).....	56
Çizelge 6.17: VK _i için 2-Gram Özellik Vektörleri İçin Başarı Oranları (%).....	56
Çizelge 6.18: VK _i için 3-Gram Özellik Vektörleri İçin Başarı Oranları (%).....	56
Çizelge 6.19: VK _i için Kök Özellik Vektörleri İçin Başarı Oranları (%).....	57
Çizelge 6.20: VK _i için Gövde Özellik Vektörleri İçin Başarı Oranları (%)	57
Çizelge 6.21: VK _{ya10} için Genel Başarı Oranları (%)	63

Çizelge 6.22: VK _{ya10} için Boyut İndirgeme Yapılmayan Özellik Vektörleri İçin Başarı Oranları (%).....	65
Çizelge 6.23: VK _{ya10} için CFS ile Boyut İndirgenmiş Özellik Vektörleri İçin Başarı Oranları (%).....	65
Çizelge 6.24: VK _{ya10} için YTU ile Boyut İndirgenmiş Özellik Vektörleri İçin Başarı Oranları (%).....	66
Çizelge 6.25: VK _{ya10} için İstatistiksel Özellik Vektörleri İçin Başarı Oranları (%).....	66
Çizelge 6.26: VK _{ya10} için İşlevsel Kelimeler Özellik Vektörleri İçin Başarı Oranları (%)	67
Çizelge 6.27: VK _{ya10} için 2-gram Özellik Vektörleri İçin Başarı Oranları (%)	67
Çizelge 6.28: VK _{ya10} için 3-gram Özellik Vektörleri İçin Başarı Oranları (%)	67
Çizelge 6.29: VK _{ya10} için Kök Özellik Vektörleri İçin Başarı Oranları (%).....	68
Çizelge 6.30: VK _{ya10} için Gövde Özellik Vektörleri İçin Başarı Oranları (%)	68
Çizelge 6.31: VK _{ya8} için Genel Başarı Oranları (%)	74
Çizelge 6.32: VK _{ya8} için Boyut İndirgeme Yapılmayan Özellik Vektörleri Başarı Oranları (%)	76
Çizelge 6.33: VK _{ya8} için CFS ile Boyut İndirgenmiş Özellik Vektörleri Başarı Oranları (%)	76
Çizelge 6.34: VK _{ya8} için YTU ile Boyut İndirgenmiş Özellik Vektörleri Başarı Oranları (%).....	77
Çizelge 6.35: VK _{ya8} için İstatistiksel Özellik Vektörleri Başarı Oranları (%).....	77
Çizelge 6.36: VK _{ya8} için İşlevsel Kelimeler Özellik Vektörleri Başarı Oranları (%)	78
Çizelge 6.37: VK _{ya8} için 2-Gram Özellik Vektörleri Başarı Oranları (%).....	78
Çizelge 6.38: VK _{ya8} için 3-Gram Özellik Vektörleri Başarı Oranları (%).....	78
Çizelge 6.39: VK _{ya8} için Kök Özellik Vektörleri Başarı Oranları (%).....	79
Çizelge 6.40: VK _{ya8} için Kök Özellik Vektörleri Başarı Oranları (%).....	79
Çizelge 6.41: VK _{yk20} için Genel Başarı Oranları (%)	85
Çizelge 6.42: VK _{yk18} için Genel Başarı Oranları (%)	91
Çizelge 6.43: VK _{yk16} için Genel Başarı Oranları (%)	97
Çizelge 6.44: VK _{yk14} için Genel Başarı Oranları (%)	103
Çizelge 6.45: VK _{yk12} için Genel Başarı Oranları (%)	109
Çizelge 6.46: İngilizce ve Türkçe Gövde Özellikleri ile Tür Belirleme Başarı Oranları (%)	116

ÖNSÖZ

Bu çalışma, Yapay Bağışıklık Sistemleri kullanılarak Türkçe dokümanların türünü, yazarını ve yazarının cinsiyetini belirlemek için hazırlanmıştır. Çalışmada çeşitli özellik vektörleriyle Yapay Bağışıklık Sistemi algoritmalarının sınıflandırmadaki başarısı, literatürde bu alanda daha önce sıkça kullanılmış ve başarıları kanıtlanmış Naive Bayes, K-En Yakın Komşuluk, Rastgele Orman ve Destek Vektör Makineleri gibi diğer sınıflandırma yöntemleriyle karşılaştırmalı olarak verilmiştir. Bu çalışmayla Türkçe dokümanların sınıflandırılmasında bugüne kadar kullanılmamış olan Yapay Bağışıklık Sistemleri kullanılarak bu alanda farklı bir yaklaşım sunulmuştur.

Çalışmam boyunca ilgi ve desteğini hiç esirgemeyen, bilgi ve deneyimlerini benimle paylaşarak büyük katkı sağlayan danışmanım Yrd. Doç. Dr. Banu Diri'ye sonsuz teşekkür ederim.

Çalışma boyunca maddi manevi her konuda destek olan aileme ve tüm arkadaşlarıma teşekkürü borç bilirim.

ÖZET

İnternet kullanımının hızla yaygınlaşmasıyla birlikte her geçen gün farklı kategorilerde bir çok doküman elektronik ortamda yerini almaktadır. Artan doküman sayısı ile birlikte bu dokümanların benzer olanlarının önceden belirlenmiş gruplara ayrılması ihtiyacı ortaya çıkmıştır. Doküman sınıflandırma dediğimiz bu işlem sayesinde dokümanlar önceden belirlenmiş sınıflara ayrıştırılmaktadır. Bu çalışmada dokümanlar, dokümanların türüne, yazarına ve yazarının cinsiyetine göre olmak üzere üç ana başlık altında sınıflandırılmıştır.

Yapay Bağışıklık Sistemleri, doğal bağışıklık sisteminden esinlenerek, bu sistemin mühendislik açısından incelenerek karmaşık problemlerin çözümünde kullanılan yöntemlerden biridir. Daha önce örüntü tanıma, hesapsal güvenlik, anomali tespiti, optimizasyon, makine öğrenmesi, robotik, kontrol, çizelgeleme, hata teşhisi gibi alanlarda ve bunların alt dallarında; ayrıca ekoloji, üretim sistemleri, akıllı evler, adaptif gürültü nötralizasyonu, indüktif problem çözümü, açık web sunucu koordinasyonu, protein yapısı tahmini gibi alanlarda başarıyla kullanılmış ve etkili sonuçlar alınmış olan bu yöntem doküman sınıflandırma alanında ilk kez bu çalışmada kullanılmıştır.

Çalışmada 16 farklı özellik vektörü oluşturularak, Yapay Bağışıklık Sistemi algoritmalarıyla ve literatürde daha önce bu alanda sıkça kullanılan diğer sınıflandırma yöntemleri olan Naive Bayes, K-En Yakın Komşuluk, Destek Vektör Makinesi ve Rastgele Orman gibi sınıflandırıcılarla deneyler gerçekleştirilerek Türkçe dokümanlar üzerinde dokümanın türü, yazarı ve yazarının cinsiyeti belirlenmeye çalışılmıştır.

Özellik vektörleri üzerinde boyut indirgeme işlemleri uygulanarak sınıflandırma yöntemlerinin başarılarının arttığı gözlenmiştir.

Yapılan denemelerde karakter n-gram'ları, kelime kökleri ve kelime gövdeleri gibi uygun özellik vektörlerinden YTU boyut indirgeme algoritmasıyla oluşturulmuş yeni özellik vektörleriyle Yapay Bağışıklık Sistemi algoritmalarının Türkçe dokümanların türünü, yazarını ve yazarının cinsiyeti belirlemede çok başarılı sonuçlar verdiği ve bu alanda geliştirilecek olan sistemlerde kullanılabileceği görülmüştür.

Anahtar kelimeler: Yapay Bağışıklık Sistemleri, YBS, yazar tanıma, tür belirleme, cinsiyet belirleme, doküman sınıflandırma, boyut indirgeme, Naive Bayes, Destek Vektör Makinesi, Rastgele Orman, K-En Yakın Komşuluk

ABSTRACT

With the rapid growth of internet usage many documents arise in electronic form everyday in different categories. The increase in the number of documents arises the need for categorizing the likely documents in predefined groups. Documents are categorized to the predefined classes by a process called document classification. In this study, documents are classified in three main headlines which are according to their genre, author and author's gender.

Artificial Immune Systems are inspired from natural immune systems that are used in engineering applications to solve complex problems. Previously, they are used in pattern recognition, computational security, anomaly detection, optimization, machine learning, robotics control, tabulation, error detection and their branches, ecology, product systems, smart homes, adaptive noise neutralization, inductive problem solving, web server coordination, protein structure guessing and succesful results are gained but they are first used in document classification problem within this thesis study.

In this study, 16 different feature vectors are constructed and tested on document genre detection, document author's gender detection and authorship attribution with Artificial Immune Systems algorithms and mostly used classifiers in literature on document classification which are Naive Bayes, K-Nearest Neighborhood, Support Vector Machines and Random Forests.

It is observed that dimension reduction techniques that are applied to the feature vectors increase the classification performance of classifiers.

In the experiments it is seen that when we apply dimension reduction algorithm called YTU to suitable featurers vectors such as character n-grams, word roots and word stems and create new feature vectors, Artificial Immune Systems algorithms give successful results in document genre detection, document author's gender detection and authorship attribution of Turkish documents and can be used in such systems.

Keywords: Artificial Immune Systems, AIRS, authoship attribution, document classification, genre detection, gender detection dimension reduction, Naive Bayes, Support Vector Machines, Random Forests, K-Nearest Neighborhood

1. GİRİŞ

Günümüzde internet kullanımı hızla yaygınlaşmaktadır. Kullanıcı sayısı arttıkça farklı kategorilerde, farklı kişilerce yazılmış çok sayıda doküman internet üzerinde yerini almaktadır. Artan doküman sayısı, beraberinde bu dokümanların birbiriyle ilişkili olanlarını kümelenendirme ihtiyacını ortaya çıkarmaktadır. Bu işlem, artan dokümanlar içerisinde ihtiyacımız olanı bulmakta bize büyük kolaylık sağlayacaktır.

Doküman sınıflandırma, bir dokümanın önceden belirlenmiş sınıflara ayrıştırılması işlemidir. Bu işlemi yapabilmek için o dokümanı diğer dokümanlardan ayırt edebilecek doğru özelliklerin belirlenmesi ve belirlenen bir sınıflandırma yönteminin uygulanması gerekmektedir. Bir doküman sınıflandırma sisteminin temel bileşenleri şunlardır (Koppel, Argamon, Shimoni, 2002):

1. Doküman Betimleme : Dokümanı temsil edebilecek özelliklerin belirlenerek her bir özelliğin frekans değerleri çıkarılıp bir özellik vektöründe tutulur.
2. Boyut İndirgeme : Oluşturulan özellik vektöründe boyut indirgeyerek ayırt edici olmayan özellikler elenir.
3. Sınıflandırma Yöntemi : Destek Vektör Makinesi, K-En Yakın Komşuluk, vd. gibi makine öğrenmesi yöntemlerinden biri kullanılır.
4. Test Protokolü : Çapraz geçerlilik gibi test protokolünü belirleyerek testler gerçekleştirilir.

Bu çalışmada doküman sınıflandırma işlemi, Türkçe metinler üzerinde ve üç farklı kategoride yapılmıştır. Bunlar:

- Doküman yazarının cinsiyetini belirleme
- Dokümanın türünü belirleme
- Dokümanın yazarını belirleme

Literatürde doküman sınıflandırma için çok çeşitli özellikler kullanılmıştır. Bunlardan bazıları istatistiksel özellikler, kelime zenginliğine dayalı özellikler, işlevsel kelimeler, dilbilgisi özellikleri ve n-gram'lardır. Doküman sınıflandırma konusunda yapılan çalışmalar arttıkça kullanılabilecek yeni özellikler de ortaya çıkarılmıştır.

Bugüne kadar çeşitli özellik vektörleri kullanılarak Destek Vektör Makinesi, Naive Bayes, K-En Yakın Komşuluk, Rasgele Orman gibi bir çok makine öğrenmesi yöntemiyle doküman sınıflandırma çalışmaları yapılmıştır.

Yapay Bağışıklık Sistemleri, doğal bağışıklık sisteminden esinlenerek, bu sistemin mühendislik açısından incelenerek, karmaşık problemlerin çözümünde kullanılan yöntemlerden biridir. Bu konuda yapılan ilk en önemli çalışma Farmer (1986) tarafından, bağışıklık sistemini diğer yapay zeka yaklaşımları ile ilişkilendirme üzerine yapılan çalışma kabul edilebilir. Bu alanda yapılmış diğer önemli çalışmalar arasında Ishida (1996); Hunt & Cooke (1996); Dasgupta (1997) ve Hofmeyr & Forrest (1999) tarafından yapılmış çalışmalar da sayılabilir.

Yapay Bağışıklık Sistemleri, örüntü tanıma, hesapsal güvenlik, anomali tespiti, optimizasyon, makine öğrenmesi, robotik, kontrol, çizelgeleme, hata teşhisi, üretim sistemleri, akıllı evler, adaptif gürültü nötralizasyonu, indüktif problem çözümü, açık web sunucu koordinasyonu, protein yapısı tahmini gibi alanlarda başarıyla kullanılmış ve etkili sonuçlar alınmıştır (de Castro ve Timmis, 2002).

Yapılan literatür çalışmalarında, bu yöntemin doküman sınıflandırmada kullanılmadığı görülmüştür. Bu tez çalışmasında uygun özellik vektörleri kullanıldığı takdirde, özellikle tür tanıma sistemlerinde başarılı sonuçlar alınacağı gösterilmiştir (Kaban ve Diri, 2008).

Bu çalışmada cinsiyete, türe ve yazara göre sınıflandırılacak olan dokümanlar, çeşitli günlük gazetelerin internet sayfalarından derlenerek elde edilmiştir. Cinsiyet, tür ve yazar tanıma çalışmalarının her biri için ayrı derlemeler oluşturulmuştur. Özellik vektörleri olarak istatistiksel özellikler, dil zenginliğine dayalı özellikler, işlevsel kelimeler, kelime kök ve gövdeleri, dilbilgisine dayalı özellikler ve karakter n-gramları kullanılmıştır. Her bir özellik vektörü kullanılarak K-En Yakın Komşuluk, Naive Bayes, Rasgele Orman ve Destek Vektör Makinesi ile Yapay Bağışıklık Sistemlerinin doküman sınıflandırmadaki başarısı diğer yöntemlerle kıyaslanmıştır. Buna ek olarak Yıldız ve arkadaşları (Yıldız vd., 2007) tarafından geliştirilen ve YTU yöntemi adını verdiğimiz yöntem kullanılmış, metinlerde geçen kelime gövdelerine bir ağırlıklandırma algoritması uygulanmış ve özelliklerin sayısı, sınıf sayısına indirilerek bu sayede hem sınıflandırma süresi hem de başarısı arttırılmış olduğu gösterilmiştir.

Tezin ikinci bölümünde, cinsiyet, tür ve yazar tanıma alanında daha önce yapılmış olan çalışmalara, üçüncü bölümünde bu çalışmada kullanılmış olan sınıflandırma yöntemleri ile Yapay Bağışıklık Sistemleri ve YTU yöntemine, dördüncü bölümde çalışmada kullanılan veri kümelerine, beşinci bölümde özellik vektörlerine, altıncı bölümde deneysel sonuçlara, yedinci bölümde ise sonuçlara ve yorumlara yer verilmiştir.

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

Bu bölümde doküman sınıflandırma probleminin yazar tanıma, dokümanın türünü belirleme ve doküman yazarının cinsiyetini belirleme konularındaki, daha önce yapılan ve literatüre girmiş çalışmalarının bazılarına yer verilmiştir.

2.1 Yazar Tanıma ile İlgili Daha Önce Yapılan Çalışmalar

Stamatatos ve arkadaşları (Stamatatos vd., 1999), çeşitli stil özelliklerinden faydalanarak Yunanca metinler üzerinde yazar tanıma çalışması yapmışlardır. Çalışmalarında SCBD (Sentence and Chunk Boundaries) adlı doğal dil işleme kütüphanesini kullanarak kelime ve tamlama ayraçları elde etmişlerdir. Metinlerdeki kelime sayısı, cümle sayısı, noktalama işareti sayıları, isim sayısı, fiil sayısı, zarf sayısı, edat sayısı, bağlaç sayısı ile bunların kelime ve tamlama sayısına oranı gibi hem ayraç seviyesinde hem de dilbilgisel 22 adet özellik içeren bir özellik vektörü kullanmışlardır. Sınıflandırma yöntemi olarak Çoklu Regresyon (Multiple Regression) ve Diskriminant Analizi (Discriminant Analysis) kullanılmıştır. 10 farklı yazarın 20'şer adet makalesi alınarak, bunların 10 tanesi eğitim seti diğer 10 tanesi ise test kümesi olarak kullanılmıştır. Sonuçta %69 gibi bir başarı oranı elde edilmiştir.

Diri ve Amasyalı tarafından, metinlerdeki karakter n-gram'ları özellik olarak alınarak Türkçe dokümanların yazarlarını, türlerini ve cinsiyetini belirleme çalışması yapılmıştır (Amasyalı ve Diri, 2006). Veri kümesi 18 farklı yazarın 35'şer yazısı alınarak, 630 dokümandan oluşturulmuştur. Bunlardan her yazarın 35 yazının 28 tanesi eğitim, 7 tanesi de test için kullanılmıştır. Özellik olarak karakter 2-gram ve 3-gram'lardan oluşan özellik vektörleri ve bu özellik vektörlerinden, özellik indirgeme yoluyla elde edilmiş vektörler ile NB (Naive Bayes), DVM (Destek Vektör Makinesi-SVM Support Vector Machine), C4.5 ve RO (Rasgele Orman- RF Random Forest) sınıflandırma yöntemleri kullanılarak her bir özellik vektörünün sınıflandırmadaki başarıları karşılaştırılmıştır. En başarılı sonuç NB sınıflandırılmasında %83,3 ile elde edilmiştir. Özellik azaltılarak sınıflandırma yapıldığında, daha iyi sonuçlar elde edildiği gözlemlenmiştir.

Diri ve Amasyalı, (Diri ve Amasyalı, 2003) Türkçe dokümanlar üzerinde yaptıkları yazar tanıma çalışmasında dokümanın içeriğini ve belirledikleri 22 farklı stil özelliğini

kullanmışlardır. Doküman içeriğini özellik olarak aldıklarında Naive Bayes, stil özelliklerini kullandıklarında ise kendi geliştirdikleri Automatic Author Detection for Turkish Text (AADTT-Automatic Author Detection for Turkish Text) yöntemini kullanmışlardır. 18 yazara ait politika, güncel ve sağlık konulu dokümanları gazetelerin internet sayfalarından elde etmişlerdir. Testlerini iki farklı veri kümesi üzerinde yapmışlardır. Birinci veri kümesi, 18 farklı yazarın her birinden 20 yazı alınarak toplam 360 dokümandan oluşmaktadır. Bunlardan 20 yazının 15 tanesi eğitim, 5 tanesi de test için kullanılmıştır. İkinci veri kümesi ise, 18 farklı yazarın 35’şer yazısı alınarak toplam 630 dokümandan oluşmaktadır. Bunlardan 35 yazının 28 tanesi eğitim, 7 tanesi de test için kullanılmıştır. Stile dayalı özelliklerin daha başarılı olduğunu görmüşlerdir. Naive Bayes ve stile dayalı özellikleri kullanan yöntem birleştirildiğinde tanıma başarısı yükselmiştir. Birinci veri kümesinde Naive Bayes’in başarısı %43, AADTT’nin başarısı %84’ tür. İkinci veri kümesinde ise Naive Bayes’in başarısı %53 iken, AADTT’nin başarısı ise %70’tir. Her iki yöntem birleştirildiğinde birinci veri kümesi için %89, ikinci veri kümesi için %82’lik başarı oranı elde edilmiştir.

Koppel ve Schler (Koppel ve Schler, 2003), İngilizce metinlerdeki yazım yanlışlarının (eksik harf kullanımı, yanlış yerde büyük harf kullanımı, eksik kelimeler, imla hataları vb.) yazar tanımadaki başarısını araştırmak için bir çalışma yapmışlardır. Çalışmada yazım yanlışlarına ek olarak ikili kelime grupları ve işlevsel kelimeler de özellik olarak alınmış, bu üç özellik birbirleriyle birleştirilerek de denemeler yapılmıştır. Derlem 11 farklı yazarın ortalama 200 kelime uzunluğunda toplam 480 adet e-posta mesajından oluşmaktadır. Deneyler 10 kat çapraz geçerlilik ile DVM ve C4.5 sınıflandırma yöntemleri ile gerçekleştirilmiştir. Yalnızca yazım yanlışları özellik olarak alındığında C4.5 sınıflandırıcısı %67,6, tüm özellikler kullanıldığında ise %72 ile en iyi başarı oranını vermiştir. Yazım yanlışları ilk bakışta yazarı ayırt etmede can alıcı bir özellik olarak görünse de, seyrek olması ve yazarların bunu yazıdan yazıya düzelterek bir alışkanlık haline getirilmesini önleyebilir olması dezavantaj sağlayan durumlardır.

Clark ve Hannon (Clark ve Hannon, 2007) metinlerde kullanılan eş anlamlı sözcük sayılarının yazarları birbirinden ayırt etmede kullanılabileceğini iddia etmişlerdir. Eş anlamlı sözcükleri bulmada WordNet isimli sözcük veritabanından faydalanılmıştır. Klasik edebiyattan Charles Dickens ve Shakespeare’in eserlerinden kelime sayıları birbirine yakın

olacak şekilde dokümanlar oluşturarak eğitim ve test kümelerini hazırlamışlardır. Deneylerde dokümanların %62'sinin yazarı doğru olarak tahmin edilebilmiştir.

Ying Zhao ve Justin Zobel (Zhao ve Zobel, 2005) işlevsel kelimeleri özellik olarak İngilizce metinler üzerinde yazar tanıma çalışması yapmışlardır. İşlevsel kelimelerin yazarları ayırt etmede kullanılabileceği ilk olarak Burrows tarafından (Burrows, 1987) ortaya atılmıştır. Benzer bir çalışmayı Baayen ve Tweedie (Baayen ve Tweedie, 2002) 72 öğrencinin kompozisyonlarını kullanarak en sık kullanılan 42 işlevsel kelime ile gerçekleştirmişlerdir. Lineer Diskriminant Analizi ve Ana Bileşen Analizi kullanarak %81,5 gibi bir başarı yakalamışlardır. İşlevsel kelimeler, edat ve bağlaçlar gibi kendi başlarına anlam ifade etmeseler bile cümle içinde kullanıldıklarında cümlenin anlamını değiştiren kelimelerdir. Türkçedeki “ve” ve “veya” kelimelerini sıkça kullanılan işlevsel kelimelere örnek olarak gösterebiliriz. Deneylerinde kullanılacak dokümanları TREC (Text Retrieval Conference) derleminin Associated Press (AP) alt koleksiyonunu seçmişlerdir. Buna gerekçe olarak zengin yazar ve tür çeşitliliği gösterilmiştir. Sınıflandırma yöntemi olarak Naive Bayes, Bayes Ağları, K-En Yakın Komşuluk ve Karar Ağaçlarını kullanmışlardır. Özellik vektörleri, 365 adet işlevsel kelimedenden oluşmaktadır. Testlerde yazar sayısını 2 ile 5 arasında, doküman sayısını da 50 ile 300 arasında değişken tutarak 10 kat çapraz geçerlilik uygulamışlardır. En iyi sonucu her yazar için 300 doküman kullanıldığında Bayes Ağları ile %76.33 başarı oranı ile almışlardır, en kötü sonucu ise %62.15 ile Karar Ağaçları vermiştir.

Morales ve arkadaşları (Morales vd., 2006) kelime dizilerinin yazar tanımadaki başarısını araştırmışlardır. Çalışmalarında kelime n-gram'ları yerine en sık kullanılan değişken uzunluktaki kelime dizilerini özellik olarak kullanmışlardır. 5 Meksikalı yazarın 353 adet yazısını kullanarak Naive Bayes sınıflandırıcısı ile testlerini yapmışlar, testlerde boyut indirgeme ve 10 kat çapraz geçerlilik uygulamışlardır. Karşılaştırma yapabilmek için aynı derlem üzerinde işlevsel kelimeler, tüm kelimeler, kelime 2-gram'ları ve kelime 3-gram'ları ile de testler yapmışlardır. Kendi geliştirdikleri en sık kullanılan değişken uzunluktaki kelime dizilerini kullanan özellik vektörü ile %83 başarı oranı yakalayarak, karşılaştırdıkları diğer özellik vektörlerinden daha iyi başarı elde etmişlerdir.

Peng ve arkadaşları (Peng vd., 2004) kelime n-gram'larının yazar tanımadaki başarısını araştırmak için çalışmayı Yunanca yazarların çeşitli gazete ve dergilerden elde ettikleri yazılarıyla, Stamatatos ve arkadaşları ile aynı veri kümesini kullanarak

gerçekleştirmişlerdir. Çalışmada $n=1$, $n=2$, $n=3$ ve $n=4$ için kelime n -gram'larını özellik olarak Naive Bayes sınıflandırıcısı ile testlerini yapmışlardır. Sonuç olarak her n değeri için %96'lık bir başarı elde etmişlerdir. Benzer çalışmayı karakter n -gram'ları ile de tekrarlamışlardır. Bu defa $n=1$ 'den $n=6$ 'ya kadar bütün n -gram'ları çıkararak her biri için ayrı ayrı testler gerçekleştirmişlerdir. En yüksek başarı oranını %90 ile $n=3$ değeri için almışlardır.

2.2 Doküman Türünü Belirleme ile İlgili Daha Önce Yapılan Çalışmalar

Peng ve arkadaşları (Peng vd., 2004) kelime n -gram'larının doküman türünü belirlemedeki başarısını araştırmışlardır. Her bir türde 20'şer doküman olmak üzere 10 farklı türden 200 doküman kullanarak Yunanca metinler üzerinde tür tanıma çalışması yapmışlardır. Derlemi çeşitli Yunan dergi ve gazetelerinden elde etmişlerdir. Çalışmada $n=1$, $n=2$ ve $n=3$ için kelime n -gram'larını kullanmışlardır. En iyi başarıyı %81 ile $n=1$ için elde etmişlerdir. Diğer n değerleri için başarı %80 olmuştur. Karşılaştırma yapabilmek için aynı veri kümesi karakter n -gram'ları ile de test edilmiştir. Burada $n=1$ 'den $n=7$ 'ye kadar olan bütün karakter n -gram'lar çıkarılarak her biri için ayrı ayrı testler yapılmıştır. En iyi sonucu %86 ile $n=2$ değeri için elde etmişlerdir.

Diri ve Amasyalı (Amasyalı ve Diri, 2006), karakter n -gram'larını özellik olarak kullanarak Türkçe dokümanların yazarlarını, türlerini ve yazarların cinsiyetini belirlemek için bir çalışma yapmıştır. Derlem olarak çeşitli gazetelerin internet sayfalarından politika, güncel ve spor konularında farklı yazarların yazıları toplanarak oluşturulmuştur. Toplam 18 farklı yazarın 35'şer yazısı alınarak 630 doküman kullanılmıştır. Her tür için 210 yazının 168 tanesi eğitim, 42 tanesi de test için kullanılmıştır. Özellik olarak 2-gram ve 3-gram'lardan oluşan özellik vektörleri ve bu özellik vektörlerinden özellik indirgeme yoluyla oluşturulan vektörler kullanılarak sınıflandırma işlemi gerçekleştirilmiştir. Naive Bayes, Destek Vektör Makinesi, C4.5 ve Rasgele Orman gibi çeşitli 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 daha iyi sonuçlar verdiği gözlemlenmiştir.

2.3 Doküman Yazarının Cinsiyetini Belirleme ile İlgili Daha Önce Yapılan Çalışmalar

Koppel ve arkadaşları (Koppel, Argamon, Shimoni, 2002), British National Corpus (BNC) üzerinden elde ettikleri 566 doküman üzerinde cinsiyet belirleme çalışması yapmışlardır. Çalışmada işlevsel kelimeleri, kelime n-gramlarını ve noktalama işaretlerini özellik olarak almışlardır. Öğrenme algoritması aynı zamanda özellik azaltma için de kullanılmıştır. Her öğrenmede belirli bir değerin altında kalan özellikler atılarak daha az özellikli bir model üretilebilmektedir. Sistemi kurgu türünde olmayan dokümanlarla eğitip, test ettiklerinde %82,6 gibi bir başarı sağlamışlardır. Kurgu türündeki dokümanlarda ise %79,5 gibi bir başarı sağlamışlardır.

Diri ve Amasyalı (Amasyalı ve Diri, 2006), karakter n-gram'larını özellik olarak kullanarak Türkçe dokümanların yazarlarını, türlerini ve yazarların cinsiyetini belirlemek için bir çalışma yapmıştır. Toplam 18 farklı yazarın 35'şer yazısı alınarak 630 doküman kullanılmıştır. Bu yazarların 4 tanesi bayan diğerleri ise 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. Özellik olarak karakter 2-gram ve 3-gram'lardan oluşan özellik vektörleri ve bu özellik vektörlerinden özellik indirgeme yoluyla oluşturulan vektörler kullanılarak sınıflandırma işlemi gerçekleştirilmiştir. Naive Bayes, Destek Vektör Makinesi, C4.5 ve Rasgele Orman gibi çeşitli sınıflandırma yöntemleri kullanılarak her bir özellik vektörünün sınıflandırmadaki başarıları karşılaştırılmıştır. En başarılı sonuç DVM kullanılarak %96,3 ile elde edilmiştir. Özellik azaltılarak yapılan sınıflandırma işleminin daha iyi sonuçlar verdiği gözlemlenmiştir.

Corney ve arkadaşları (Corney vd., 2002), 342 yazara ait 8820 e-posta mesajını kullanarak cinsiyet belirleme çalışması yapmışlardır. Çalışmalarında e-posta mesajlarının başlık, imza gibi kısımlarını çıkararak gövde bölümü en az 50 kelimedenden oluşan e-postaları seçmişlerdir. Çalışmayı 100, 150 ve 200 kelimedenden oluşan e-postaları alarak tekrarlamışlardır. Çeşitli istatistiksel özellikler (boş satır sayısı, ortalama cümle uzunluğu vb.), kelime zenginliğine dayalı özellikler (hapax legomena), işlevsel kelimeler gibi 221 adet özellikten oluşan bir özellik vektörü oluşturmuşlardır. Çalışmada sınıflandırma yöntemi olarak DVM kullanmışlardır. En başarılı sonucu %71.1 ile 100 ve daha fazla kelime içeren 1000 e-postadan oluşan veri kümesi ile elde etmişlerdir.

3. SINIFLANDIRMA YÖNTEMLERİ

Öğrenme yöntemlerini eğitici (supervised), eğitici (unsupervised) ve yarı-eğitici (semi-supervised) olarak üç bölümde ele alabiliriz. Bu çalışmada eğitici yöntemlerden Naive Bayes, K-En Yakın Komşuluk, Destek Vektör Makineleri ve Rastgele Orman, sezgisel yöntemlerden de Yapay Bağışıklık Sistemleri (YBS/Artificial Immune Systems-AIRS) kullanılmıştır.

3.1 Eğitici Sınıflandırma Yöntemleri

Bu bölümde sırasıyla Naive Bayes, K-En Yakın Komşuluk, Destek Vektör Makinesi ve Rastgele Orman hakkında bilgi verilecektir.

3.1.1 Naive Bayes (NB)

Naive Bayes sınıflandırıcısı Bayes Teorisine dayanan ve özelliklerin birbirinden bağımsız olduğunu varsayan bir sınıflandırıcıdır. Yeni örnekler üzerinde sınıflandırma Bayes kuralına göre, sınıflandırılacak örneğe en yüksek olasılıklı benzerlik gösteren sınıf seçilerek yapılır. Bütün özelliklerin birbirinden bağımsız olarak düşünülmesi sınıflandırıcıyı basitleştirse de gerçek hayatta böyle bir durumla karşılaşmak neredeyse imkansızdır. Buna rağmen NB'nin sınıflandırmadaki başarısı çeşitli çalışmalarla kanıtlanmıştır. Özelliklerin birbirinden bağımsız olarak varsayılması, her birinin ayrı ayrı öğrenilmesine imkan vermektedir. Bu da öğrenmeyi çabuklaştırmakta, çok fazla özellik olan sistemlerde bile sınıflandırmanın hızlı bir şekilde gerçekleştirilmesine izin vermektedir. Kelime özelliklerine dayalı doküman sınıflandırma sistemlerinde binlerce özellik olduğu için bu alanda kullanılması uygun olmaktadır.

Naive Bayes sınıflandırıcısı için bir örneğin bir C sınıfına ait olma olasılığı (3.1)'deki gibi hesaplanır:

$$P(C, X_1, \dots, X_n) = P(C) \prod_{i=1}^n P(X_i|C) \quad (3.1)$$

Burada her sınıf için olasılıklar hesaplanarak hangi sınıfa ait değer en yüksek çıkarsa, örnek o sınıfa aittir denilir.

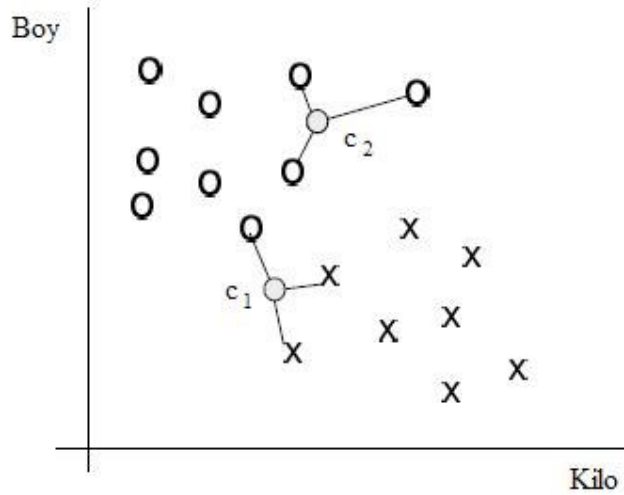
Morales ve arkadaşları (Morales vd., 2006) kelime dizilerinin yazar tanımadaki başarısını araştırmışlardır. Çalışmalarında kelime n-gram'ları yerine en sık kullanılan değişken

uzunluktaki kelime dizilerini özellik olarak kullanmışlardır. 5 Meksikalı yazarın 353 adet yazısını kullanarak Naive Bayes sınıflandırıcısı ile testlerini yapmışlar ve %83'lük bir tanıma başarısı elde etmişlerdir. Bu çalışmada WEKA[1] uygulamasında yer alan Naive Bayes modülü kullanılmıştır.

3.1.2 K-En Yakın Komşuluk (K-EYK)

En Yakın Komşuluk sınıflandırmasının özünde yatan fikir oldukça basittir. Örnekler, kendilerine en yakın olan komşularının sınıf etiketlerine göre sınıflandırılırlar. Genelde birden fazla en yakın komşuyu dikkate almak daha faydalı olmaktadır, dolayısıyla sınıflandırıcı k adet komşu dikkate alındığı için K-En Yakın Komşuluk adıyla bilinir.

Şekil 3.1'de sistemin çalışma mantığı basit olarak gösterilmiştir. Burada $k=3$ alınarak en yakın 3 komşuya göre sınıflandırma yapılmıştır. Şekil 3.1'de 2 boyutlu uzayda 2 sınıfın birbirinden ayırt edilmesi gösterilmektedir. C_1 örneği için en yakın 3 komşudan 2'si X sınıfına, 1'i de O sınıfına aittir. X sınıfına ait yakın komşu sayısı daha fazla olduğu için bu örnek X sınıfına dahil edilir. C_2 örneğinde ise en yakın 3 komşu da O sınıfından olduğu için c_2 O sınıfına dahil edilir.



Şekil 3.3-1 K-En Yakın Komşuluk

Bu çalışmada WEKA uygulamasında yer alan K-En Yakın Komşuluk modülü kullanılmıştır. Bu modül Aha ve Kibler (Aha ve Kibler, 1991) tarafından bahsedilen K-En Yakın Komşuluk yöntemi esas alınarak geliştirilmiştir. Yöntem iki örnek arasındaki

mesafe ölçümünü yaparken Öklid bağıntısını kullanmaktadır. A ve B noktaları arasındaki mesafe (distance) Öklid bağıntısına göre (3.2)'deki gibi hesaplanır:

$$D(A,B) = \sum_{i=1}^n \sqrt{(A_i - B_i)^2} \quad (3.2)$$

Ying Zhao ve Justin Zobel (Zhao ve Zobel, 2005) işlevsel kelimeleri özellik olarak alarak İngilizce metinler üzerinde yazar tanıma çalışması yaparken K-En Yakın Komşuluk sınıflandırıcısını kullanmışlardır. Sınıf sayısı 5 alındığında %70, 2 alındığında ise %85 gibi bir başarı elde edilmiştir.

3.1.3 Destek Vektör Makinesi (DVM)

Destek Vektör Makinesi, Vapnik (Vapnik, 1995) tarafından geliştirilmiş bir makine öğrenmesi yöntemidir. Bugüne değin el yazısı tanıma (Cortes ve Vapnik, 1995), nesne tanıma (Blanz vd., 1996), ses tanıma (Schmidt, 1996), yüz tanıma (Osuna, Freund ve Girosi, 1998) ve metin sınıflandırma (Joachims, 1998) gibi bir çok alanda kullanılmıştır.

DVM, iki sınıf bilgiyi birbirinden ayırmak için, en yakın farklı sınıf örnekleri arasındaki mesafeyi maksimum yapan aşırıdüzlemi bulmayı amaçlar. Bu maksimum mesafeye marjın denir. Şekil 3.2'de açık(negatif) ve koyu renkli(pozitif) iki sınıfı birbirinden ayıran aşırıdüzlem gösterilmiştir. Bu düzlem üzerindeki noktalar $w x + b = 0$ eşitliğini sağlamaktadır. Burada w , düzleme olan normal ve $|b| / \|w\|$ düzlemden orijine dik uzaklıktır. DVM ile sınıflandırma fonksiyonu $f(x)$ (3.3)'teki gibi ifade edilecek olursa:

$$f(x) = \text{sign}(w x + b) \quad (3.3)$$

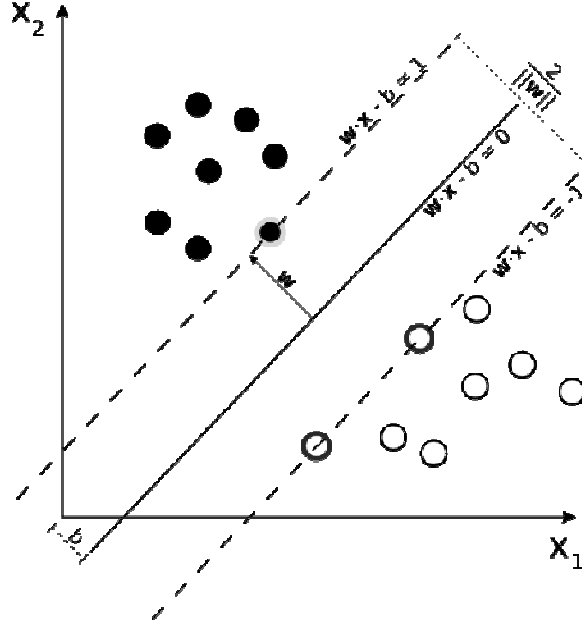
$f(x) \geq 0$ olduğunda pozitif, $f(x) < 0$ olduğunda negatif olarak sınıflandırılır.

DVM iki farklı sınıfı birbirinden ayırmak için tasarlanmıştır. Bu nedenle çok sınıflı sistemleri DVM ile sınıflandırabilmek için çeşitli algoritmalar geliştirilmiştir. Bu algoritmalar şunlardır:

Bire Karşı Heps (One Against Rest): K sınıflı bir sınıflandırma sistemi için K adet sınıflandırıcı oluşturulur. Her sınıflandırıcı kendi verilerini pozitif, diğerleri negatif olacak şekilde tanımlar.

Bire Karşı Bir (Pairwise): Sınıflandırıcı her sınıf çifti için ayrı ayrı eğitilir. K sınıflı bir sistem için $K(K-1)/2$ adet sınıflandırıcı oluşturulur. Oluşturulan çiftler birbiriyle

karşılaştırılarak sonuç elde edilir. Bu sonuçtan geriye kalan çiftler yine birbiriyle karşılaştırılır. En sonunda bir tek sınıflandırıcı kalana kadar işlem sürdürülür.



Şekil 3.2 Destek Vektör Makinesi [7]

Yönlendirilmiş Çevrimsiz Çizge (Directed Acyclic Graph): K sınıflı bir sistem için ilk önce 1 ve 2. sınıflar için olan sınıflandırıcılar kıyaslanır. Devamında her defasında kazanan sınıflandırıcı ile geriye kalan sınıflandırıcılardan biri karşılaştırılarak en sonunda bir sınıf kalıncaya kadar işlem sürdürülür.

Bu çalışmada WEKA uygulamasında yer alan DVM sınıflandırıcı modülü kullanılmıştır. John Platt (Platt, 1998) tarafından geliştirilmiş modül çok sınıflı sınıflandırma işlemleri için bire karşı bir algoritmasını kullanır.

Thorsten Joachims (Joachims, 1998), DVM'nin doküman sınıflandırmadaki başarısını denemek için yaptığı çalışmada, Reuters ve Ohsumed derlemlerinden elde ettiği dokümanları türlerine göre sınıflandırmak için DVM sınıflandırıcısını kullanmıştır. Özellik olarak dokümanlarda 3 ve daha fazla sayıda geçen kelime gövdelerini almıştır. Elde ettiği %86'lık yüksek başarı oranı ile DVM'nin doküman sınıflandırma için uygun bir sınıflandırıcı olduğunu göstermiştir.

Diri ve Amasyalı (Amasyalı ve Diri, 2006), karakter n-gram'larını özellik olarak kullanarak Türkçe dokümanların yazarlarını, türlerini ve yazarların cinsiyetini belirlemek için bir çalışma yapmıştır. Çalışmada Naive Bayes, DVM, C 4.5 ve Rastgele Orman gibi çeşitli sınıflandırma yöntemleri kullanılmış ve sonuçları karşılaştırmalı olarak verilmiştir.. DVM sınıflandırıcısı tür tanımada %93.6, cinsiyet tanımada %96.3'lük başarı oranlarıyla hem tür hem de cinsiyet tanımada en yüksek başarıyı veren sınıflandırıcı olmuştur.

3.1.4 Rastgele Orman (RO)

Rastgele Orman sınıflandırıcısı bir çok karar ağacından oluşan ve ağaçlar arasında oylama sonucuna göre kazanan sınıfı çıktı olarak veren bir makine öğrenmesi yöntemidir. Yöntem Leo Breiman ve Adele Cutler tarafından, Tin Kam Ho'nun Rastgele Karar Ağaçları fikrinden esinlenerek geliştirilmiştir. Rastgele Orman yöntemindeki karar ağaçları şu şekilde oluşturulur:

- M adet özelliği olan, N elemanlı bir eğitim kümesi olduğunu düşünelim,
- Bir ağacın düğümündeki kararı verebilmek için bize gereken girdi değişkenlerinin sayısı olan m verilir, öyle ki m değeri M değerinden çok daha küçüktür,
- Elimizde olan N adet eğitim kümesindeki örneklerden N kez yenisiyle değiştirme (replacement) yaparak rastgele eğitim örnekleri seçilir. Kalan örnekler ağacın hata oranını hesaplamak için kullanılır,
- Ağacın her düğümünde seçilmiş olan m adet değişken özellikten faydalanarak bir karar verilir. Oylamayla verilen karara göre o düğümdeki en iyi ayırım hesaplanır.

Diri ve Amasyalı (Amasyalı ve Diri, 2006), karakter n-gram'larını özellik olarak kullanarak Türkçe dokümanların yazarlarını, türlerini ve yazarların cinsiyetini belirlemek için yaptıkları çalışmada NB, DVM, C4.5 ve RO gibi çeşitli sınıflandırıcıların başarı oranlarını karşılaştırmışlardır. RO yönteminin yazar tanımada %80.2, tür tanımada %92.7 ve cinsiyet tanımada da %92.2 başarı oranlarıyla, doküman sınıflandırma için uygun bir yöntem olduğunu göstermişlerdir.

Bu çalışmada WEKA uygulamasında yer alan RO sınıflandırıcı modülü kullanılmıştır. Modül Leo Breiman'ın (Breiman, 2001) çalışması esas alınarak geliştirilmiştir.

3.2 Yapay Bağışıklık Sistemleri (YBS)

Bu bölümde sırasıyla doğal bağışıklık ve yapay bağışıklık sistemleri hakkında bilgi verilmiştir.

3.2.1 Doğal Bağışıklık Sistemi

Bağışıklık sistemi bütün vücuda yayılmış çeşitli hücre ve organlardan oluşur. Temel görevi canlıyı hatalı davranış sergileyen hücreler (tümör ve kanser hücreleri) ile vücuda zarar verebilecek dış öğelerden (virüs ve bakteriler) korumaktır. Vücuda ait ve zararsız olan hücrelere self, hastalığa sebebiyet veren öğelere ise non-self denmektedir. Bağışıklık sistemi self ve non-self öğeleri birbirinden ayırt etmekle görevlidir.

Bağışıklık sistemi organları lenfoid organlar olarak da bilinir. Lenfoid organlar timus, dalak ve lenf düğümleridir. Bunlar birincil lenfoid organlar (Central Lymphoid Organs) ve ikincil lenfoid organlar (Peripheral Lymphoid Organs) olmak üzere ikiye ayrılırlar. Birincil lenfoid organlar timus ve kemik iliğidir, görevleri lenfositlerin (B ve T hücreleri) üretilmesi, çoğaltılması ve farklılaştırılmasıdır. İkincil organlar, lenfositlerin antijenlerle karşılaştığı yerlerdir.

3.2.1.1 Bağışıklık Sistemi Organları

Tonsil ve Adenoid : Solunum yoluyla vücuda girebilecek yabancı mikroorganizmalara karşı bağışıklık hücreleri içeren özel lenf düğümleridir.

Lenf Damarları : Antijenlerin, lenf sıvılarının ve lenf hücrelerinin kandaki dolaşımını ve bağışıklık sistemi organlarına iletimini sağlayan damarlardır.

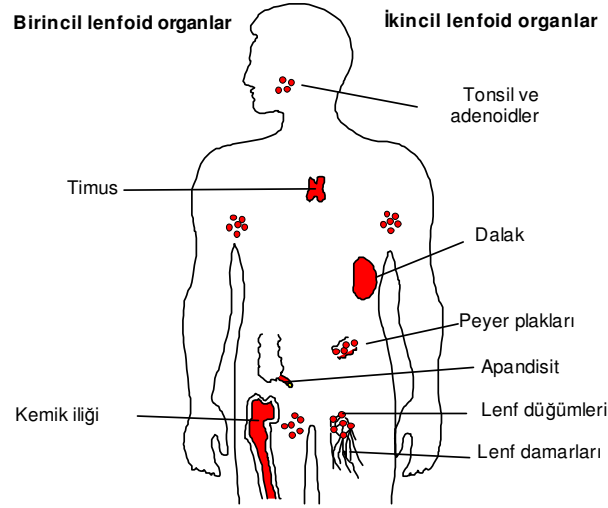
Kemik İliği : Uzun kemiklerin içinde bulunan ve bağışıklık hücrelerini üreten yumuşak dokudur.

Lenf Düğümleri : Lenf damarlarının düğümlendiği yerde bulunan, B ve T lenfositlerinin depolandığı yerdir.

Timus : Kemik iliğinde üretilen hücreler burada çoğalıp farklılaşmalarını tamamlarlar.

Dalak : Lökositlerin kan dolaşımında bulunan yabancı mikroorganizmaları yok ettiği organdır.

Apandisit ve Payer Plakları : Sindirim sistemini koruyan özel lenf düğümleridir.



Şekil 3.3 Bağışıklık Sistemi Organları (Alataş vd., 2005)

3.2.1.2 Bağışıklık Sistemi Hücreleri

Bağışıklık sistemi hücreleri kemik iliğinde üretilirler. Bazıları vücudun genel savunması için görevliken, bazıları da belli patojenler için özelleşmişlerdir.

Lenfositler : Bağışıklık sisteminin sorumluluğunu büyük ölçüde yüklenen hücrelerdir. En önemli iki türü B ve T lenfositlerdir.

B Hücreleri : B hücrelerinin ana görevi yabancı organizmalara karşı antikorların üretimi ve salgılanmasıdır. Antikorlar yabancı maddeleri tanıma ve onlara bağlanma kabiliyetleri olan özel proteinlerdir. Antikorların üretilmesi ve bağlanmaları genellikle diğer hücreler için yok etme sinyali anlamına gelmektedir. Antikorlar, bağışıklık sisteminin yabancı organizmaları tanıma işlevini yerine getiren en önemli elemanıdır.

T Hücreleri ve Lenfokinler : Timus içinde olgunlaştıkları için T hücreleri olarak adlandırılmışlardır. T hücrelerinin görevi, diğer hücrelerin çalışmalarını kontrol etmek ve enfekte olmuş hücreleri yok etmektir. Üç çeşit T hücresi vardır. Bunlar: yardımcı, baskılayıcı ve öldürücü T hücreleri.

Yardımcı T hücreleri B hücrelerinin, diğer T hücrelerinin, makrofajların ve doğal öldürücü hücrelerin etkin hale gelmesini sağlarlar.

Öldürücü T hücreleri kanser yapıcı hücreleri, virüsleri ve istilacı hücreleri kimyasal salgılarla yok eden hücrelerdir.

Baskılayıcı T hücreleri diğer hücreleri kontrol ederek sistemin kontrolden çıkmasını ve kendi hücrelerine zarar vermesini engeller.

T hücreleri bir çeşit kimyasal salgı olan lenfokinleri salgılar. Lenfokinler güçlü kimyasal sinyallerdir. Hücre büyümesine, yenilenmesine ve etkinleşmesine destek olurlar. Bunun yanında, zararlı hücreleri öldürme yetenekleri de vardır.

Doğal Öldürücü Hücreler : Öldürücü T hücreleri gibi güçlü kimyasal salgılar içeren granüllere sahiptirler. T hücrelerinin aksine harekete geçmeden önce antijeni tanımaya ihtiyaç duymazlar. Tümörleri ve enfeksiyon yapan mikropların büyük bölümünü yok ederler.

Fagosit, Monosit ve Makrofajlar : Fagositler literatürde hücre yiyiciler olarak da bilinir. Bazı fagositlerin lenfositlere antijenleri sunma (Antigen Presenting Cells [APC]) özelliği de vardır. En önemli fagositler, monosit ve makrofajlardır. Monositler kanda dolaşarak dokulara yerleşirler ve makrofaj (Büyük Yiyiciler) haline gelirler. Makrofajlar antijenleri parçalayıp sindirdikten sonra T lenfositlerine sunarlar. Bağışıklık reaksiyonun başlamasında önemli rol oynarlar.

Tamamlayıcı (Complement) Sistem : Antikorların işlevlerine yardımcı olan protein kompleksini yaratan sistemdir. Tamamlayıcı sistem yabancı bir organizma ile karşılaştığında bileşenleri ile bir zincir reaksiyon başlatır. İstilacı organizmanın yüzeyine bağlanarak, lezyon oluşturur. Tamamlayıcı sistem vücudun çeşitli bölgelerinde bulunan 25 proteinden oluşmaktadır.

Antikorlar : Antikorlar moleküler örüntüleri tanıma özelliklerinden dolayı bağışıklık sisteminde önemli bir role sahiptirler. Antijenlerin yapısal olarak çok çeşitli olması antikorların repertuarının da aynı derecede çeşitli olmasını gerektirmektedir.

Antikorlar ikişer ağır, ikişer hafif zincirden oluşmuştur. Antikoru değişken ve sabit olmak üzere iki kısma ayırabiliriz. Değişken bölge (V-Bölgesi) antijenleri tanıma görevini yerine getiren bölümdür. Sabit bölge ise diğer fonksiyonları yerine getirir.

MHC (Major Histocompatibility Complex) : İki sınıfı vardır: I. ve II. sınıf. I. sınıf bütün hücrelerde vardır, II. sınıf MHC molekülleri ise Antijen Sunan Hücrelerde (ASH) bulunur. Antijen sunan hücreler, antijenleri peptitlere ayırır, MHC-II molekülleri de bu peptitlerle birleşirler ve ASH'nin yüzeyinde taşınırlar (De Castro, 1999).

3.2.1.3 Başıřıklık Sisteminin Çalışması

Başıřıklık sistemi çok sayıda molekül ve hücreyle vücudu antijenlere karşı korur. Koruma şu aşamalardan oluşur:

- Makrofajlar gibi antijen sunan hücreler, buldukları antijenleri antijenik protein parçacıklarına dönüřtürürler.
- Bu protein parçacıkları MHC molekülleri ile birleşirler. T lenfositler, reseptörleri vasıtasıyla bu protein-MHC birleşimini ayırt edebilirler.
- T lenfositler bu tanıma sonrasında aktive olurlar ve salgıladıkları lenfokinler ile diğerk başıřıklık sistemi hücrelerini harekete geçirirler.
- B hücreleri de reseptörleri yardımıyla aktive olurlar. B hücrelerinin, T hücrelerinin aksine MHC molekülleri olmaksızın serbest antijenleri tanıyabilme özellikleri vardır.
- Aktive olan B hücreleri plazma hücrelerine dönüřürler ve antikorları üretirler.
- Antikorlar buldukları antijenlere bağlanarak onları pasif hale getirirler. Tamamlayıcı sistem veya diğerk enzimler yardımıyla antijeni yok ederler.
- Tekrar karşılaşıldığında daha güçlü bir reaksiyon gösterebilmek için bazı T ve B hücreleri antijenle uyarıldıktan sonra hafıza hücrelerine dönüřürler. (De Castro, 1999)

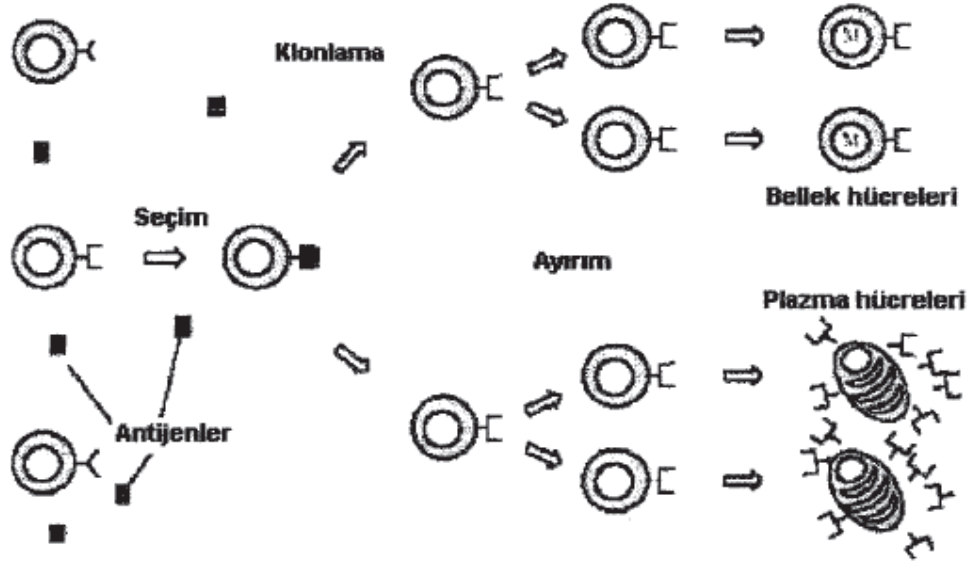
3.2.1.4 Klonal Seçim Prensibi

Antijene karşı başıřıklık sisteminin cevap vermesini sağılayan, T ve B hücreleri tarafından sadece antijene cevap verebilecek hücrelerin seçilmesine dayanan prensiptir. Vücuda giren antijene karşı germinal merkezde başıřıklık cevabı oluşur. Kemik iliğinden köken alan B hücreleri, bu antijenlere bağlanarak ve yardımcı T hücrelerinin sinyali ile çoğalıp olgunlaşarak plazma hücrelerine dönüřürler. Bir kısmı da hafıza hücrelerine dönüřüp kanda ve lenf damarlarında dolaşımda kalırlar. Böylece benzer bir antijenle karşılaştıklarında daha hızlı antikor üretebilirler.

Klonal seçim teorisinin özellikleri (Burnet 1978):

- Yeni hücreler yüksek oranda mutasyona uğramış ebeveynlerinin kopyasıdır (klon).

- Mutasyon, tanıma miktarı ile ters orantılı olarak gerçekleşir. Yani antijene yüksek duyarlılıkta olan hücreler daha az mutasyona uğrarlar.
- Düşük seviyede benzerlik taşıyan, farklılaşmasını yeni tamamlamış lenfositler elenir.
- Antijenleri tanıyan hücreler, tanıdıkları oranda çoğaltılır.



Şekil 3.4 Klonal Seçim Prensibi (Parmaksızoğlu, 2005)

3.2.1.5 Kendinden Olan/Olmayan (Self/Nonsself) Ayrımı

Bağışıklık sistemi, hem vücudun kendi hücrelerini hem de yabancı hücreleri tanıma özelliğine sahiptir. Vücut, kendi hücrelerini yok etmeme özelliğine sahiptir. Buna, kendine karşı toleranslıdır da denir. Bu toleranstan Yardımcı T Hücreleri (Th-Cells) sorumludur. Timusta olgunlaşmakta olan Th hücreleri, olgunlaşmaları esnasında kendinden olan bir hücreye bağlanır ve aktif hale gelirlerse negatif seçim ya da klonal silme denen bir işlemle öldürülürler. Böylece olgunlaşmalarını tamamlayıp timustan vücuda yayılacak olan Th hücreleri, vücudun kendi hücrelerine karşı toleranslı durumda olmuş olurlar.

3.2.2 Yapay Bağışıklık Sistemleri

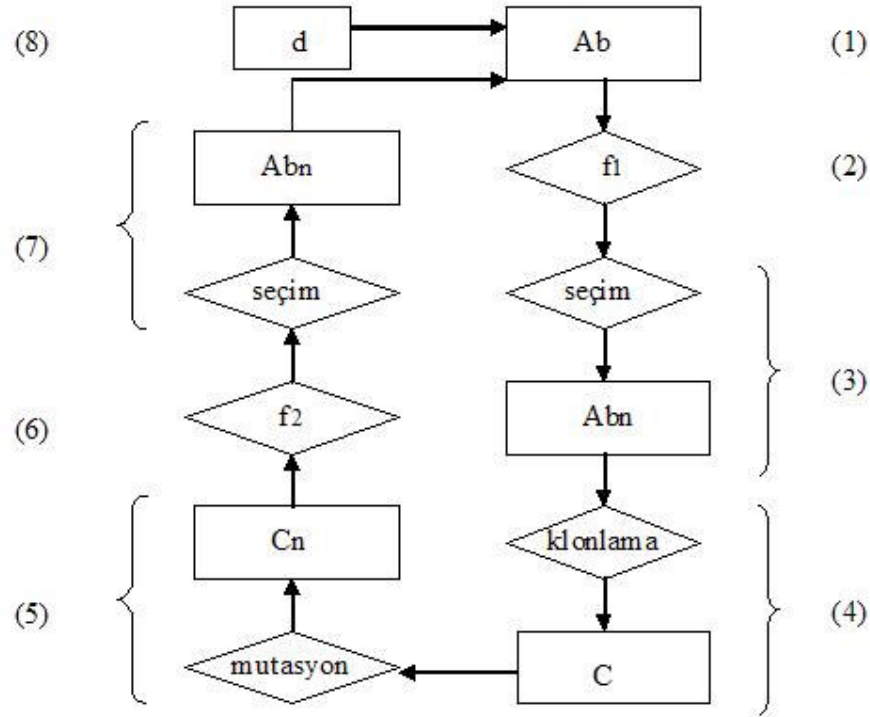
Doğal bağışıklık sistemlerinin çalışma prensiplerinin, çeşitli problemlerin çözümünde kullanılabileceği düşünülmüştür. YBS, örüntü tanıma, hesapsal güvenlik, anomali tespiti, optimizasyon, makine öğrenmesi, robotik, hata teşhisi gibi alanlarda ve bunların alt dallarında, ayrıca ekoloji, üretim sistemleri, akıllı evler, açık web sunucu koordinasyonu, protein yapısı tahmini gibi alanlarda başarıyla kullanılmış ve etkili sonuçlar alınmıştır (de Castro ve Timmis, 2002) Yapay Bağışıklık Sistemleri doğal bağışıklık sistemlerinin en çok Klonal Seçim, Negatif Seçim gibi mekanizmalarından esinlenmiştir. Bunlara bağlı geliştirilen algoritmalar şu şekildedir:

3.2.2.1 Klonal Seçim Algoritması

Hunt ve Cooke (Hunt ve Cooke,1996) yaptıkları çalışmayla klonal seçim prensibini bir algoritma olarak sunmuşlardır:

1. Vücut antikor repertuarını oluşturan antikorlar, başlangıç çözüm kümesini oluştururlar (Ab).
2. Antikorların benzerlik dereceleri hesaplanır (f_1).
3. n adet en yüksek benzerlikli antikor seçilir (Ab_n).
4. Seçilen n adet antikorun benzerlik dereceleri ile doğru orantılı olarak, yüksek benzerlik dereceli antikorun daha fazla olacak şekilde klonlanması (C).
5. Antikorların benzerlik derecesi yüksek olan, daha az olacak şekilde mutasyona uğratılır (C_n).
6. Mutasyona uğramış klonların benzerlik dereceleri belirlenir (f_2).
7. n yüksek benzerlik dereceli antikorların yeniden seçilmesi (Ab_n).
8. d adet en düşük benzerlik derecesindeki antikorların yeni üretilen antikorlarla değiştirilmesi (Ab_n).

Algoritmanın adımları Şekil 3.5’de verilmiştir.



Şekil 3.5 Klonal Seçim Algoritması Adımları (Engin ve Döyen, 2004)

3.2.2.2 Negatif Seçim Algoritması

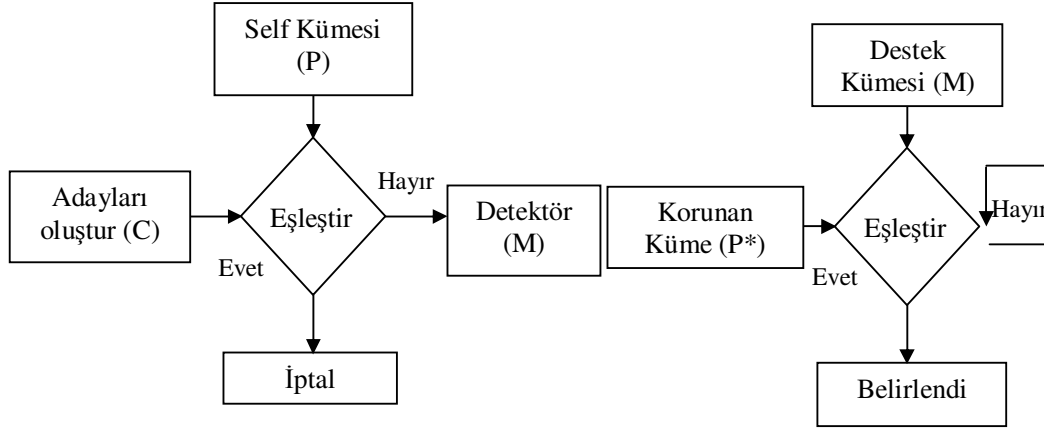
Bir bilgi işleme perspektifi ile olaya bakılırsa, negatif seçim, desen tanımayı gerçekleştirmek için alternatif bir paradigma sunar. Bu paradigma tanınacak desenleri (self) tamamlayıcı küme (nonself) hakkında bilgi depolama mantığı ile çalışır. Literatürde; bilgisayar ve ağ saldırıları, zaman serileri tahmini, şekil denetimi ve kesimlemesi ve donatım hatası toleransı gibi anormallik belirleme problemleri ile ilgili uygulamalarla birlikte bir negatif seçim algoritması (Forrest vd., 1994) önerilmiştir.

Öncelikle, korunacak modellerin kümesi belirlenir ve 'self-set'(P) olarak adlandırılır. Negatif seçim algoritmasına dayanarak, 'self-set' kümesine ait olmayan elemanları tanımakla sorumlu bir algılayıcılar (detektörler) (M) kümesi oluşturulur.

1. Aynı gösterim kabulünü kullanan rastgele aday elemanlarını (C) oluştur.

2. C' deki elemanlarla P' deki elemanları karşılaştır (eşle). Bir eşleşme olursa, örneğin, P' 'nin bir elemanı C' 'nin bir elemanı tarafından tanınırsa, C' 'nin bu elemanını at. Ters durumda, C' 'nin bu elemanını algılayıcı (dedektör) kümesi içinde sakla.

Algılayıcı kümesini (M) oluşturduktan sonra, algoritmanın diğer aşaması nonself modellerin varlığı için sistemi izlemekten oluşur. Bu durumda bir P^* korunacak kümesi oluşturulur. Bu küme, P kümesi ve diğer yeni modellerden veya tamamıyla yeni bir kümeden meydana gelir. Nonself modellerine karşılık gelen, algılayıcı kümesinin tüm elemanları için, P^* kümesinin bir elemanını tanıyıp tanıyamadığı (eşleşme) kontrol edilir. Eşleşme var ise, o zaman bir nonself modeli tanınmıştır ve reaksiyon başlamalıdır. Nonself tespitinde sonuç aksiyonu değerlendirme problemine göre çeşitlilik gösterir, ve negatif seçim algoritmasının model tanıma kapsamı dışındadır. Algoritmanın adımları Şekil 3.6'da verilmiştir.



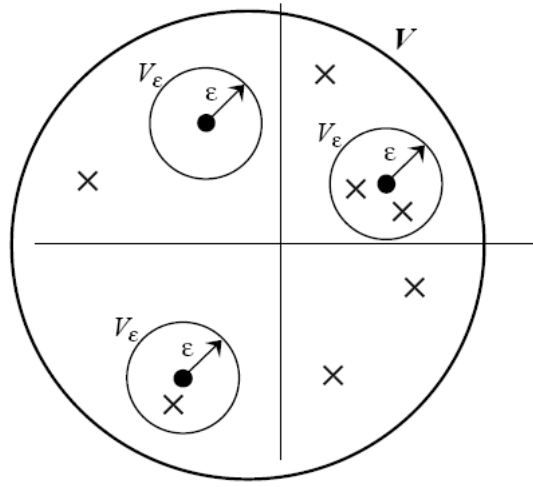
Şekil 3.6 Negatif Seçim Algoritması (Engin ve Döyen, 2004)

3.2.2.3 Şekil Uzayı ve ARB Kavramları

Antijen tanıma için gerekli tüm hücrelerin oluşturduğu kümeye bağışıklık repertuarı denir. İnsan vücudu herhangi bir antijeni tanıyabilecek protein genlerine sahiptir ve bu genlerin çeşitli kombinasyonları ile tüm antijenleri tanıyabilecek kapasitededir. Bağışıklık hücresi molekülleri ile antijenler arasındaki etkileşimi anlatmak için Perelson ve Oster (Perelson ve Oster, 1979) Şekil Uzayı modelini ortaya atmışlardır.

Şekil Uzayında gösterilen her bir self hücrenin (B ve T hücreleri) yarıçapı bir eşik seviyesiyle belirlenen bir tanıma çemberi vardır. Her hücre bu çember içerisinde bulunan antijenleri tanıyabilir.

Artificial Recognition Ball (ARB), Timmis ve Neal (Timmis ve Neal, 2000) tarafından önerilmiş bir kavramdır. ARB aynı antikora sahip çok sayıda B hücresi olarak tanımlanabilir. Yapay bağışıklık sistemi çok sayıda ARB'den oluşmaktadır. Timmis ve Neal sistem kaynakları olarak B hücrelerinin sayısını göstermişlerdir. Kaynaklar sınırlı olduğu için öğrenme rekabete dayalı olmaktadır. Bir antijenle karşılaştığında ARB'ler uyarılırlar. Uyarılma seviyelerine göre kaynak tüketmeye başlarlar. Daha çok uyarılanlar daha çok kaynak tüketeceğinden ve kaynaklar da sınırlı olduğundan bir süre sonra daha az uyarılan ARB'ler sistemden çıkarılırlar. Böylece en güçlü ARB'lerin sistemde kalması sağlanmış olur (Şekil 3.7).



Şekil 3.7 Şekil Uzayı (De Castro, 1999)

3.2.2.4 AIRS (Artificial Immune Recognition System) Algoritması

AIRS algoritması Watkins (Watkins, 2001) tarafından bağışıklık sisteminin bilinen özellikleri ve algoritmaları kullanılarak oluşturulmuş bir sınıflandırma algoritmasıdır. Algoritmada YBS'nin klonal seçim, somatik hipermutasyon, şekil uzayı gibi kavramlarından faydalanılmıştır.

AIRS algoritmasında 2 popülasyon vardır. Bunlar ARB'ler ve hafıza hücreleridir. Bir eğitim hücresi sunulduğunda, antijenle uyuşan ARB'ler seçilir ve onlara daha fazla kaynak ayrılır. Bu uyum, mutasyon ve seçim süreci boyunca yeterli bilgiye sahip olan aday hafıza hücresi, genel hafıza hücresi kümesine yerleştirilir. Bu işlem her eğitim elemanı için tekrarlanır. En sonunda sınıflandırma en yakın hafıza hücresi K-En Yakın Komşuluk algoritmasıyla bulunarak gerçekleştirilir.

Algoritmayı açıklarken elimizde X gibi, $ag_i = \{x_i, t_i\} \in R$. Z olacak şekilde elemanlardan oluşan bir eğitim kümesi olduğunu düşünelim. Burada x_i , d özelliği olan bir girdiyi, t_i ise ($i=1,2,...,n$) olacak şekilde bir sınıfı temsil etmektedir. $MC=\{mc_1, mc_2, ..., mc_n\}$ hafıza hücreleri kümesini, $AB=\{ab_1, ab_2, ..., ab_r\}$ ARB popülasyonu kümesini göstermektedir.

Algoritma temelde 4 aşamadan oluşmaktadır:

1. **Veri Normalizasyonu ve Başlangıç :** Veri kümesindeki bütün antijenler arasındaki öklid uzaklığı $[0,1]$ arasında olacak şekilde normalize edilir. Bu işlemden sonra duyarlılık eşiği (Affinity Threshold) hesaplanır. Duyarlılık eşiği bütün eğitim verilerinin özellik vektörleri üzerindeki ortalama duyarlılıktır. Bu aşamada son işlem olarak ise bellek hücresi ve ARB popülasyonu oluşturulur. Bu rastgele olarak 0 veya daha fazla antijenin bellek hücresi ve ARB kümesine eklenmesiyle olur.
2. **Hafıza Hücresi Belirleme ve ARB Oluşumu :** Burada eğitim antijenine ($ag_i = \{x_i, t_i\}$) en büyük uyarma değerine sahip, aynı sınıftan olan ($t_{match}^{mc} = t_i$) hafıza hücresi ($mc_{match} = x_{match}^{mc}, t_{match}^{mc}$) bulunur. Uyarma, antijen ve hafıza hücresi arasındaki öklid uzaklığıyla bağıntılıdır. Eğer antijen ile aynı sınıftaki hafıza hücresinin kümesi boş küme ise hafıza hücresinin kümesine antijen eklenir ve antijen mc_{match} olarak kabul edilir. Bulunan hafıza hücresi mevcut ARB'lerin popülasyonuna yeni ARB'ler eklemek için kullanılır. Eklenen ARB hücreleri hafıza hücresinin uyarılma değeriyle orantılı olarak mutasyona uğramış hafıza hücrelerinden oluşmaktadır.
3. **Aday Hafıza Hücresinin Gelişimi:** Bu aşamada elimizde bir önceki adımdan gelen ARB'ler için en uygunlarını hayatta bırakacak bir mekanizma uygulanır. AIRS'te, aynı sınıftaki bir antijenle uyarıldığında yüksek bir uyarı değeri veren ve farklı bir sınıftaki antijenle uyarıldığında az uyarı veren hücreler ödüllendirilerek daha fazla kaynak verilir. Kaynak elde etme mücadelesini kaybeden hücreler elenir. Böylece yüksek kaynağa sahip ve etkili uyarılan hücreler sistemde kalmış olurlar.

İşleme ARB kümesinin her bir alt sınıfının ortalama uyarılma değeri, uyarılma eşik değerinden küçük olduğu sürece devam edilir. Daha sonra eğitim antijeniyle aynı sınıftaki en çok uyarılan $mc_{candidate}$ aday hafıza hücresi seçilir.
4. **Hafıza Hücresinin Sunumu:** Eğer $mc_{candidate}$ antijen ile, mc_{match} 'ten daha fazla uyarıldıysa $mc_{candidate}$ hafıza hücresi kümesine yerleştirilir. Bu mekanizmayla daha iyi

hafıza hücreleri elde edilmiş olunur. Bu aşama da tamamlandıktan sonra ağı antijeni için eğitim tamamlanmış olunur ve diğer antijenlere geçilir.

5. **Sınıflandırma** : Eğitim tamamlandıktan sonra sınıflandırma işlemi, K-En Yakın Komşuluk algoritmasına göre test verisine en fazla sayıda yakın elemanı olan hafıza hücresinin sınıfı seçilerek yapılır.

3.2.2.5 AIRS Algoritması Türevleri (AIRSv1, AIRSv2 ve AISv2Parallel)

Watkins (Watkins, 2001) tarafından geliştirilen AIRS algoritması AIRSv1, AIRSv2 ve AIRSv2Parallel gibi 3 farklı türeve sahiptir. Bunlardan AIRSv1 ve AIRSv2, bölüm 3.2.2.4'te anlatılan temel adımları aynen izlemekle birlikte bazı küçük farklılıkları bulunmaktadır.

AIRSv2 algoritması AIRSv1'e göre daha kolay ve karmaşıklığı az olan bir algoritma olmakla birlikte performansında küçük azalmalar olabileceği öngörülmüştür (Çatal, 2007). ARB popülasyonu AIRSv1 için kalıcı olmasına karşın, AIRSv2 için geçici bir kaynaktır. Önceki adımlardan gelen ARB'ler farklı sınıflara da ait bilgiler içereceğinden ve aynı sınıfa ait ARB'lerin antijene daha yakın olması beklendiğinden bu konuda AIRSv2'nin yaklaşımı daha uygundur. AIRSv1'de mutasyon işleminden sonra klonlanan hücrelerin sınıfları değiştirilebilir ancak AIRSv2'de bu yapılmaz. AIRSv1'de mutasyon oranı kullanıcı tarafından verilmekte olup AIRSv2'de somatik hipermutasyonla yani afinite değeriyle orantılı olarak mutasyon uygulanır.

AIRSv2Parallel algoritması (Watkins, 2005) ise AIRSv2 algoritmasının paralel programlama mantığına uyarlanmış halidir. AIRSv2Parallel, paralel işlem yapılabilir olması sebebiyle en hızlı çalışan YBS algoritmasıdır.

3.3 Boyut İndirgeme

Literatürde boyut azaltmak için kullanılan yöntemler arasında en bilinenleri Temel Bileşen Analizi (TBA/Principal Component Analysis-PCA) ve Doğrusal Ayırteden Analizi (DAA/Linear Discriminant Analysis-LDA) dir. Bu çalışmada özellik sayısını sınıf sayısına indirerek boyut azaltmaya gittiğimiz ve YTU adını verdiğimiz bir yöntem kullanılmıştır.

3.3.1 YTU Yöntemi

Günümüzde yapılan metin sınıflandırma sistemlerinin bir çoğu, kelime kök ya da gövdelerini özellik olarak alan Bag of Words (Kelime Torbası) modeli ile yapılmaktadır. Bu durum hem özellik sayısını hem de sınıflandırma süresini arttırmaktadır. Yıldız ve arkadaşları (Yıldız vd., 2007) tarafından geliştirilen ve YTU yöntemi adını verdiğimiz yöntem kullanılarak, metinlerde geçen kelime gövdelerine bir ağırlıklandırma algoritması uygulanarak, özellik sayısı sınıf sayısına indirilmekte ve bu sayede hem sınıflandırma süresi hem de başarısı arttırılmış olmaktadır. Kaban ve Diri (Kaban ve Diri, 2008) bu yöntemi kullanarak yaptıkları çalışmada, Yapay Bağışıklık Sistemi algoritmaları olan AIRS1, AIRS2, AIRS2Parallel sınıflandırıcıları ile tür ve yazar tanımda başarılı sonuçlar almışlardır.

YTU yöntemi ile özellik vektörünün boyutunu indirgeme herhangi bir çok boyutlu özellik vektörüne de uygulanabilir. Bu çalışmada kelime kökleri, kelime gövdeleri gibi özelliklerin yanı sıra karakter n-gram'larına da bu boyut indirgeme yöntemi uygulanmış ve başarılı sınıflandırma sonuçları elde edilmiştir.

Özelliklerin bir kısmı yalnızca bazı sınıflara ait dokümanlarda geçiyorken, bazı özellikler ise tüm sınıflardaki dokümanlarda ortak olarak geçmektedir. Bu özellikler ayırt edici olmadığından sınıfına olan etkisini azaltmak için (3.4)'teki ağırlıklandırma formülü kullanılmıştır.

$$\log(tf_i + 0.5) * \log(D/df) \quad (3.4)$$

D = Toplam metin sayısı

tf_i = Özelliğin i . sınıfta geçtiği metin sayısı

df = Özelliğin geçtiği metin sayısı

$X = \{x_1, x_2, \dots, x_n\}$ başlangıç özellik vektörü olmak üzere, bir d dokümanı için YTU yöntemi ile elde edilen ve sınıf sayısı kadar özellik içeren yeni X' özellik vektörü şu adımlarla elde edilir:

- Her s sınıfı için x_i özelliğinin, (3.4)'teki ağırlıklandırma formülü kullanılarak ağırlıklandırılmış değeri w_{is} elde edilir.
- Her s sınıfı için elde edilen w_{is} değerleri toplanarak w_s değeri elde edilir.

- Tüm sınıflar için elde edilen w_s değerleri toplanarak w toplamı bulunur.
- Sınıflar için elde edilen her w_s değerinin w toplamına bölünerek, sınıf sayısı kadar özellik içeren yeni X' özellik vektörünün değerleri elde edilir.
- Bu adımlar tüm dokümanlar için tekrarlanır.

Bir örnekle açıklamak gerekirse, elimizde iki sınıfa ait sekiz adet doküman (d_i), üç adet kelime (k_j) ve kelimelerin, doküman ve sınıflara göre dağılımı Tablo 3.1'deki gibi verilmiş olsun.

Çizelge 3.1: Örnek veri kümesi

	<i>Sınıf 1</i>				<i>Sınıf 2</i>			
	d1	d2	d3	d4	d5	d6	d7	d8
k1	1	0	2	1	1	1	0	0
k2	1	0	1	0	2	1	0	1
k3	1	0	1	2	1	0	0	0

YTU yöntemini kullanarak her dokümanı farklı kelime sayısı yerine, sınıf sayısı kadar bir özellik vektörü ile ifade edebiliriz. Örneğin $d1$ dokümanını ifade etmek için kullandığımız özellik vektörünün değerleri sırasıyla 0.6058 ve 0.3941'dir. Eşitlik 3.4'deki ağırlıklandırma formülünü kullanarak, her kelimenin sınıflara göre ağırlık değerlerini hesaplayacak olursak:

$$w_{11} = \log(3+0.5) * \log(8/5) = 0.111$$

$$w_{12} = \log(2+0.5) * \log(8/5) = 0.0812$$

$$w_{21} = \log(2+0.5) * \log(8/5) = 0.0812$$

$$w_{22} = \log(3+0.5) * \log(8/5) = 0.111$$

$$w_{31} = \log(3+0.5) * \log(8/4) = 0.164$$

$$w_{32} = \log(1+0.5) * \log(8/4) = 0.053$$

k1 kelimesi: *sınıf 1* için $w_{11}=0.111$, *sınıf 2* için $w_{12}=0.0812$

k2 kelimesi: *sınıf 1* için $w_{21}=0.0812$, *sınıf 2* için $w_{22}=0.111$

k3 kelimesi: *sınıf 1* için $w_{31}=0.164$, *sınıf 2* için $w_{32}=0.053$

$$w_1 = 0.111 + 0.0812 + 0.164 = 0.3562$$

$$w_2 = 0.0812 + 0.111 + 0.053 = 0.2452$$

w_1 toplamı 0.3562, w_2 toplamı 0.2452 olup, iki sınıf toplamı $w=0.6014$ 'dür. Buna göre dI 'in ilk özellik değeri $0.3562/0.6014=0.5923$, ikinci özellik değeri de $0.2452/0.6014=0.4077$ 'dir.

3.3.2 Korelasyon Tabanlı Özellik Seçici (KTÖS/Correlation-based Feature Selection-CFS)

Makine öğrenmesi, büyük ölçekli verilerin otomatik olarak analiz edilmesini sağlayan araçlar sağlamaktadır. Makine öğrenmesinin en önemli öğelerinden biri de özellik seçimidir. Özellik seçimi tüm özellikler arasından öğrenme ve sınıflandırma için en uygun özelliklerin seçilmesi prensibine dayanır. Hall (Hall, 1998) eğitici öğrenme yöntemleri için özellikler arasındaki korelasyonu temel alan bir özellik seçme yöntemi önermiştir. KTÖS bu anlamda basit ve hızlı olarak özellik seçimi sunmaktadır.

Hall(Hall, 1998) bu çalışmasında iyi bir özellik alt kümesinde birbiriyle korelasyonu düşük, sınıfla olan korelasyonu ise yüksek olan özellikler bulunur prensibiyle hareket etmiştir.

KTÖS, gereksiz, sınıflandırmaya katkısı olmayan ve gürültü niteliğine sahip özellikleri eleyen bir yöntemdir. KTÖS genellikle özelliklerin yarısından fazlasını eler ve bir çok makine öğrenmesi yönteminin başarı performansının arttırılmasını sağlar (Hall ve Smith, 1999). Bu çalışmada WEKA uygulaması içinde yer alan CfsSubsetEval modülü kullanılmıştır.

4. VERİ KÜMELERİ

Bu bölümde tez çalışmasında kullanılan veri kümeleri üç başlık altında toplanarak anlatılmaktadır. Dokümanın türünü, yazarını ve yazarının cinsiyetini belirlemek için ayrı ayrı dört farklı veri kümesi oluşturulmuştur. Bu veri kümelerinin her biri Milliyet (www.milliyet.com.tr), Hürriyet (www.hurriyet.com.tr) ve Sabah (www.sabah.com.tr) gibi günlük ulusal gazetelerin yazarlarının makaleleri derlenerek elde edilmiştir. Her bir veri kümesi farklı sayıda sınıfa içerdiği için, farklı doküman sayısına sahiptir.

4.1 Cinsiyet Belirleme Çalışmasında Kullanılan Veri Kümesi (VKc)

Cinsiyet belirleme çalışmasında kullanılan veri kümesi (VKc) iki sınıftan (bay/bayan) ve her sınıfta 5 adet yazarın 35'er dokümanından oluşmaktadır. Bu çalışmada kullanılan veri kümesi içerisindeki yazarların isim ve dokümanlarının toplam boyutu yazarların listesi Çizelge 4.1'de verilmiştir. Bu veri kümesinde (VKc) bulunan toplam doküman sayısı 350'dir.

Çizelge 4.1: Cinsiyet Belirleme Çalışmasında Kullanılan Yazarlar

Bayan Yazarlar	Boyut (KB)	Bay Yazarlar	Boyut (KB)
Ayşe Arman	284	Can Dünder	140
Ebru Çapa	208	Fikret Bila	156
Meliha Okur	140	Güneri Cívaoğlu	140
Pakize Suda	168	Hıncal Uluç	176
Tuğçe Baran	136	Taha Akyol	136

4.2 Tür Belirleme Çalışmasında Kullanılan Veri Kümesi (VKt)

Tür belirleme çalışmasında kullanılan veri kümesi (VKt), 5 farklı türden (Siyaset, Ekonomi, Spor, Popüler Konular ve, Sağlık), gibi 5 farklı türden oluşmaktadır. Farklı yazarlara ait olan ve türdenden 50 adet yazı alınarak, toplamda 250 her birinden 50'şer tane olmak üzere, toplam 250 adet dokümandan oluşmaktadır.

4.3 Yazar Tanıma Çalışmasında Kullanılan Veri Kümeleri

Yazar tanıma çalışmasında iki farklı veri kümesi kullanılmıştır. Birinci veri kümesi kendi içerisinde 5 farklı alt küme oluşturacak şekilde, farklı türlerde yazılmış dokümanlardan

oluşmaktadır. İkinci veri kümesi ise, aynı türde yazan yazarların yazılarından oluşan 2 alt kümeden oluşmaktadır. karma türlerde yazan yazarları tanıma ve aynı türde yazan yazarları tanıma şeklinde iki kısma. ayrılmıştır. Buradaki amaç, yazarları tanıırken farklı türde ve aynı türde yazanların tanınmadaki başarılarının test edilmesidir.

4.3.1 Yazar Tanıma için Karma Türlerde Yazan Yazarlardan Oluşturulan Veri Kümeleri

Karma türlerde yazar tanıma için oluşturulan veri kümeleri; 20,18,16,14 ve 12 yazarlı olmak üzere 5 alt kümeye ayrılarak incelenmiştir. Burada sınıf sayısı değişken tutularak sınıf sayısı ile sınıflandırma başarısının nasıl değiştiği gözlenmiştir. Bu veri kümelerinde 4 farklı türde (Ekonomi, siyaset, popüler ve spor) ve her türde 5 farklı yazarların yazılarına yer verilmiştir.

Bu çalışmada da her yazarın 35'er yazısı alınmıştır. Çizelge 4.2'de değişen sınıf sayılarında hangi türden hangi yazarların dokümanlarına yer verildiği gösterilmiştir. VKy_{k20} , 20 yazarlık, VKy_{k18} , 18 yazarlık, VKy_{k16} , 16 yazarlık, VKy_{k14} 14 yazarlık, VKy_{k12} , 12 yazarlık veri kümesini ifade etmektedir. Her veri kümesi için o veri kümesinde bulunan yazarlar ilgili sütunda x ile işaretlenmiştir.

4.3.2 Aynı Türde Yazan Yazarlar için Veri Kümeleri

Aynı türde yazan yazarları tanıma çalışmasında 10 ve 8 yazardan oluşan 2 alt veri kümesi oluşturulmuştur. içeren veri kümeleri kullanılmıştır. Tür olarak *siyaset* türünde yazan yazarlar seçilmiştir. Sınıf sayısı değişken tutularak sınıf sayısı ile sınıflandırma başarısının nasıl değiştiği gözlenmiştir. Bu çalışmada da her yazarın 35'er yazısına yer verilmiştir. Çizelge 4.3'te değişen sınıf sayılarında hangi yazarların dokümanlarına yer verildiği gösterilmiştir. VKy_{a10} , 10 yazarlık, VKy_{a8} , 8 yazarlık veri kümesini ifade etmektedir. Her veri kümesi için o veri kümesinde bulunan yazarlar ilgili sütunda x ile işaretlenmiştir.

Çizelge 4.2: Karma Türde Yazar Tanıma Çalışmasındaki Veri Kümeleri

Yazar	Tür	VKy _{k20}	VKy _{k18}	VKy _{k16}	VKy _{k14}	VKy _{k12}
A.Savaş Akat	Ekonomi (832 KB)	x	x			
Erdal Sağlam		x	x	x	x	x
Güngör Uras		x	x	x	x	
Meliha Okur		x	x	x	x	x
Vahap Munyar		x	x	x	x	x
Bekir Coşkun	Siyaset (716 KB)	x	x	x	x	x
Can Dünder		x	x	x	x	x
Fikret Bila		x	x	x		
Güneri Civaoglu		x	x	x	x	x
Taha Akyol		x	x			
Ayşe Arman	Popüler (992 KB)	x	x	x	x	x
Ebru Çapa		x	x	x	x	x
Hıncal Uluç		x	x	x	x	x
Pakize Suda		x	x	x	x	
Savaş Ay		x	x			
Cüneyt Tanman	Spor (752 KB)	x	x	x	x	x
Gürcan Bilgiç		x	x	x	x	x
Kazım Kanat		x	x	x	x	x
Levent Tüzmen		x		x		
Talay Erker		x				

Çizelge 4.3: Aynı Türde Yazar Tanıma Çalışmasında Kullanılan Yazarlar

Yazar	Boyut (KB)	VKya₁₀	VKya₈
Bekir Coşkun	140	x	x
Can Dünder	140	x	x
Cüneyt Ülsever	144	x	
Fikret Bila	160	x	x
Güneri Cívaoğlu	140	x	x
Mustafa Sarıkaya	140	x	x
Oktay Ekşi	140	x	x
Taha Akyol	136	x	x
Tayfun Türeñç	152	x	
Yalçın Bayer	300	x	x

5. ÖZELLİK VEKTÖRLERİ

Bu bölümde çalışmada kullanılan özellik vektörlerinden bahsedilmektedir. Doküman sınıflandırma çalışmalarında, sınıfları birbirinden ayırt edebilmek için, önce uygun özellikler seçilerek özel bir şekilde ifade edilir. Özellikleri içeren bu gösterim biçimine özellik vektörü denir. Doküman sınıflandırma çalışmalarında sıkça kullanılan özellikler olan istatistiksel özellikler, kelime zenginliğine dayalı özellikler, işlevsel kelimeler, dilbilgisine dayalı özellikler ve karakter n-gram'ları ile dokümanlarda geçen kelimelerin kök ve gövdelerinden özellik vektörleri oluşturularak bunların sınıflandırmadaki başarıları test edilmiştir. Ayrıca bazı özellik vektörlerine de özellik indirgeme işlemleri uygulanarak boyutları düşürülmüş, bu işlem sonucunda ortaya çıkan yeni özellik vektörlerinin sınıflandırma performansına olan etkileri gözlenmiştir.

5.1 İstatistiksel Özellikler (V_{IST})

Doküman sınıflandırmada en çok kullanılan özelliklerden biri olan istatistiksel özellikler, dokümanlarda geçen bazı sembollerini sayma, kelime ve cümlelerin uzunluk değerlerini istatistiksel olarak ifade etme biçiminde oluşturulmaktadır. Brinegar (Stamatatos, 2000) kelime uzunlukları, Morton (Stamatatos, 2000) cümle uzunlukları, Holmes (Stamatatos, 2000) kullanılan kelime sayısı ve doküman uzunluğu gibi özellikleri kullanmışlardır. Stamatatos (Stamatatos, 2000) Yunanca dokümanlar üzerinde yaptığı çalışmada 22 farklı stil özelliğine yer vermiş, bunların içinde çeşitli noktalama işaretlerinin kelime sayısına oranını da kullanmıştır. Bu çalışmada V_{IST} olarak adlandırılan ve 26 adet istatistiksel özellik içeren vektör şu özelliklerden oluşmaktadır:

- Aç Köşeli Parantez Sayısı
- Aç Küme Sayısı
- Aç Parantez Sayısı
- Alt Çizgi Sayısı
- Ampersant Sayısı
- Çift Tırnak Sayısı
- Çizgi Sayısı

- Diyez Sayısı
- Eşittir Sayısı
- İki Nokta Sayısı
- Kapa Köşeli Parantez Sayısı
- Kapa Küme Sayısı
- Kapa Parantez Sayısı
- Nokta Sayısı
- Noktalı Virgül Sayısı
- Bölü İşareti Sayısı
- Soru İşareti Sayısı
- Ters Bölü İşareti Sayısı
- Ünlem Sayısı
- Virgül Sayısı
- Yıldız İşareti Sayısı
- Yüzde İşareti Sayısı
- Üç Nokta Sayısı
- Cümle Sayısı
- Kelime Sayısı
- Ortalama Cümle Uzunluğu

5.2 Kelime Zenginliğine Dayalı Özellikler (V_{KZ})

Bir yazarın kelime dağarcığının zenginliği onu diğer yazarlardan ayırt etmek için kullanılabilir özellikler sağlamaktadır. Literatürde kelime zenginliğine dayalı özellikler çeşitli çalışmalarda (Burrows 1992, Holmes 1994, Forsyth 1995) kullanılmıştır. Bu çalışmada oluşturulan V_{KZ} özellik vektörü şu özellikleri içermektedir:

- a. Farklı kelime sayısının toplam kelime sayısına oranı (type/token ratio)

- b. Dokümanda yalnızca bir kere geçen kelimelerin sayısı (hapax legomena)
- c. Dokümanda sadece iki defa geçen kelimelerin sayısı (hapax dislegomena)

Birinci özellik V/N olarak gösterilmektedir. V dokümandaki farklı sözcük sayısına, N 'de dokümandaki toplam sözcük sayısına karşılık gelir. Yazarın dokümanda farklı kelimeler kullanabilme yeteneğini ifade eden bir orandır. V_{KZ} vektöründe bu üç özellik kullanılarak, doküman sınıflandırmadaki başarısı test edilmiştir.

5.3 Dilbilgisi Özellikleri (V_{DB})

Türkçe'de sözcükler dilbilgisel açıdan isim, sıfat, fiil, bağlaç, edat, zarf, zamir ve ünlem gibi farklı sözcük türlerinde gruplandırılırlar. V_{DB} özellik vektörü, bu 8 özelliği içeren bir özellik vektörüdür. Bu özellikler çıkarılırken Türk Dil Kurumu'nun Türkçe Kelime Veritabanı'ndan (TKV) (Diri ve Amasyalı, 2003) faydalanılmıştır. TKV bir sözcüğün olası sözcük türlerini vermektedir. Ancak Türkçe'de sözcükler cümlede geçtiği yere göre farklı türde olabilmektedirler. Örneğin "Çok hızlı koşuyorsun." cümlesindeki "hızlı" sözcüğü zarf iken "Hızlı tren seferleri yıl sonunda başlayacak." cümlesindeki "hızlı" sözcüğü sıfattır. Bu gibi tür farklılıklarını belirleyebilmek için aşağıdaki şu kurallar uygulanmıştır:

- Bir kelimenin olası tiplerinden biri sıfat ve kelime ek almamışsa ve bir sonraki kelime isim veya zamir ise bu kelime sıfattır.
- Bir kelimenin olası tiplerinden biri sıfat ise ve kelime ek almamışsa, kelimenin olası tiplerinden sıfat çıkarılır ve tip sayısı azaltılır. Eğer tip sayısı 1'e düşüyorsa kelimenin tipi olarak alınır.
- Bir kelimenin olası tiplerinden biri sıfat ve kelime ek almamışsa son madde uygulanır.
- Bir kelimenin olası tipleri arasında sıfat yok, fakat zarf varsa ve bir sonraki kelime fiil ise veya bir sonraki kelime cümlede sonunda ise bu kelime zarf olarak alınır.
- Bir kelimenin olası tipleri arasında sıfat yoksa, fakat zarf varsa ve bir sonraki kelime isim ise, kelimenin olası tiplerinden zarf çıkarılır ve tip sayısı azaltılır. Eğer tip sayısı 1'e düşmüşse kelimenin tipi olarak alınır. (Diri ve Amasyalı, 2003)
- Bir kelimenin olası tipleri arasında sıfat yoksa, fakat zarf varsa ve kelime cümlede sonundaysa bu kelime fiil olarak alınır.

- Eğer kelime türünün 3 olasılığı varsa ve bunlardan ikisi sıfat ve zarf olmasına rağmen önceki kurallara göre sıfat ya da zarf koşullarını sağlamamışsa, diğer tür kelimenin türü olarak kabul edilir.
- Eğer kelimenin olası türlerinden biri fiilse ve kelime cümlenin sonunda yer alıyorsa kelime fiildir.
- Kelimenin olası yalnızca 1 türü varsa o tür kabul edilir.
- Yukarıdaki kurallarla tipi belirlenemeyen kelime için olası tipler arasında kelimenin en fazla sıklıkla kullanılan tipi tercih edilir.

Diri ve Amasyalı'nın (Diri ve Amasyalı, 2003), Türkçe dokümanlar üzerinde yaptıkları yazar tanıma çalışmasında kullandıkları 22 özelliğin 8 tanesi, bu sözcük türlerinin sayısının cümle sayısına oranından oluşmaktadır.

5.4 İşlevsel Kelimeler (V_{İK})

Tek başlarına kullanıldıklarında anlamı olmasa da cümle içinde kullanıldıklarında diğer kelimelerin anlamını değiştirebilen ve cümleye anlam katan kelimeler işlevsel kelimeler (Function Words) olarak adlandırılırlar. Örneğin “ile” sözcüğü tek başına kullanıldığında anlam ifade etmez, ancak “Ahmet ile Mehmet geldiler.” cümlesindeki “ile” sözcüğü cümleye birliktelik anlamı katar. İşlevsel kelime olmayan ve tek başlarına da anlamları olan kelimelere içerik kelimeleri (Content Words) denir. İsim, fiil, sıfat gibi sözcük türleri tek başına anlam taşıdıkları için içerik kelimeleridir. İşlevsel kelimeler, edat, zamir ve bağlaç sözcük türlerine ait kelimelerden oluşur.

İşlevsel kelimelerin yazarları ayırt etmede kullanılabileceği ilk olarak Burrows tarafından (Burrows, 1987) ortaya atılmıştır. Benzer bir çalışmayı Baayen ve Tweedie (Baayen ve Tweedie, 2002) 72 öğrencinin kompozisyonlarını kullanarak en sık kullanılan 42 işlevsel kelime ile gerçekleştirmişlerdir. Çalışmada %81,5 gibi bir başarı yakalamışlardır. Ying Zhao ve Justin Zobel (Zhao ve Zobel, 2005) işlevsel kelimeleri özellik olarak alarak İngilizce metinler üzerinde yazar tanıma çalışması yapmışlardır. Özellik vektörlerini 365 adet işlevsel kelimeden oluşturmuşlardır.

Türkçe’de işlevsel kelimelerin listesi çıkarılmadığından, bu çalışmada TKV kullanılarak bir işlevsel kelimeler listesi oluşturulmuştur. TKV’de yapılan bir sorgulamayla tipi edat,

zamir ve bağlaç olan kelimeler çıkarılmıştır. Bu kelimelerle oluşturulan V_{IK} adı verilen özellik vektörü 154 özellikten oluşmaktadır.

5.5 Karakter N-Gram'ları (V_{nG})

Karakter n-gram'ları bir karakter katarının n elemanlı karakter dizileridir. N-gram'lar n sayısının değerine göre adlandırılır. Alınan n değeri 2 iken 2-gram, 3 iken 3-gram'lar olarak söz edilir. Bu çalışmada karakter n-gram'ları olarak 2-gram'lar ve 3 gram'lar kullanılmıştır. Bir örmekle açıklayacak olursak “Sinema” kelimesinin 2-gram'ları “Si”, “in”, “ne”, “em”, “ma” şeklindedir. Aynı kelimenin 3-gram'ları ise “Sin”, “ine”, “nem”, “ema” şeklinde olacaktır. Bu şekilde tüm dokümanlar bir karakter katarı olacak şekilde ele alınıp olası tüm 2-gram'lar ve 3-gram'lar çıkarılarak özellik vektörleri oluşturulmuştur. 2-gram özellik vektörü V_{2G} , 3-gram özellik vektörü ise V_{3G} olarak adlandırılmıştır.

Diri ve Amasyalı tarafından, metinlerdeki karakter n-gram'lar özellik olarak alınarak Türkçe dokümanların yazarlarını, türlerini ve cinsiyetini belirleme çalışması yapılmıştır (Amasyalı ve Diri, 2006). 18 yazardan oluşan bir veri kümesiyle yazar tanımda Naive Bayes sınıflandırıcısı ile %83,3'lük başarı elde etmişlerdir. Peng ve arkadaşları (Peng vd., 2004) karakter n-gram'larının yazar tanımdaki başarısını araştırmak için Yunanca metinler üzerinde bir çalışma yapmışlardır. Çalışmalarında 1'den 6'ya kadar bütün n-gram'ları çıkararak her biri için ayrı ayrı testler gerçekleştirmişlerdir. En yüksek başarı oranını %90 ile n değeri 3 iken almışlardır.

5.6 Kelime Kök veya Gövdeleri ($V_{KÖK}$ veya $V_{GÖVDE}$)

Doküman içerisinde geçen kelimelerin birer özellik olarak kabul edilmesine Kelime Torbası (Bag of Words) denmektedir. Bu çalışmada doküman içerisinde yer alan kelimeler yerine, kelimelerin kök veya gövdeleri dokümanın sınıflandırılmasında kullanılmıştır.

Bir kelimenin aldığı yapım ve çekim eklerinin çıkarılmış haline kök denir. Örneğin “kitapçılar” kelimesinin kökü “kitap”tır. Kelimenin çekim eklerinin çıkarılmış hali o kelimenin gövdesini verir. Bu durumda “kitapçılar” kelimesinin gövdesi “kitapçı” olmaktadır.

Kelime kök ve gövdelerinden özellik vektörleri oluşturulmasının nedeni aynı sınıfa ait dokümanlarda aynı kelime kök veya gövdesinden daha fazla olacağı düşüncesidir. Örneğin

“Göz Hastalıkları” ile ilgili bir yazıda “göz”, “gözlük”, “gözlükçü”, “gözbebeği” gibi göz kökünden türemiş bir çok kelime olabilir. Bu sebeple bu doküman için “göz” kökü ayırt edici bir özelliktir. Bu düşünceden hareketle dokümanlarda bulunan kelime kök veya gövdeleri açık kaynak kodlu bir Türkçe Doğal Dil İşleme kütüphanesi olan Zemberek [6] yardımıyla çıkarılmış ve iki farklı özellik vektörü elde edilmiştir. Kelime köklerinin frekanslarını özellik olarak alan vektör $V_{KÖK}$ ve kelime gövdelerinin frekanslarını özellik olarak alan vektör de $V_{GÖVDE}$ olarak adlandırılmıştır.

5.7 Boyut İndirgeme Yoluyla Oluşturulan Vektörler

Bir özellik vektöründe çok fazla özellik olması onun sınıflandırma için iyi bir özellik vektörü olduğu anlamına gelmez. Aksine, özellik seçme işlemiyle daha ayırt edici özelliklerin elde edildiği, boyutu indirgenmiş vektörler daha iyi başarı oranları verebilir. Bu çalışmada WEKA [1] içerisinde yer alan CfsSubsetEval fonksiyonuyla ve Yıldız ve arkadaşları (Yıldız vd., 2007) tarafından geliştirilen YTU yöntemiyle özellik vektörlerinin boyutları indirgenerek yeni özellik vektörleri oluşturulmuştur. Çizelge 5.1’de tüm özellik vektörleri ile boyut indirgenmesiyle oluşturulan yeni özellik vektörlerinin isimleri ve hangi veri kümesi için kaçar özellik içerdikleri verilmiştir.

Çizelge 5.1: Boyut İndirgenmiş Özellik Vektörleri

	VK _c	VK _t	VK _{ya10}	VK _{ya8}	VK _{yk20}	VK _{yk18}	VK _{yk16}	VK _{yk14}	VK _{yk12}
V _{IST}	26	26	26	26	26	26	26	26	26
V _{IST_CFS}	6	9	12	12	12	12	10	12	14
V _{IST_YTU}	2	5	10	8	20	18	16	14	12
V _{KZ}	3	3	3	3	3	3	3	3	3
V _{DB}	8	8	8	8	8	8	8	8	8
V _{IK}	154	154	154	154	154	154	154	154	154
V _{IK_CFS}	22	19	11	12	17	14	21	16	20
V _{IK_YTU}	2	5	10	8	20	18	16	14	12
V _{2G}	1789	1699	1739	1682	1887	1874	1800	1780	1731
V _{2G_CFS}	26	40	33	46	28	32	34	38	35
V _{2G_YTU}	2	5	10	8	20	18	16	14	12
V _{3G}	14842	12955	14747	13960	17138	16897	16015	15705	14998
V _{3G_CFS}	122	145	78	107	113	119	129	133	142
V _{3G_YTU}	2	5	10	8	20	18	16	14	12
V _{KÖK}	6323	4789	5664	5331	7103	7017	6533	6401	6087
V _{KÖK_CFS}	88	73	48	67	62	71	80	81	86
V _{KÖK_YTU}	2	5	10	8	20	18	16	14	12
V _{GÖVDE}	7676	5873	6844	6473	8597	8488	7954	7790	7404
V _{GÖVDE_CFS}	100	74	53	69	72	66	78	83	87
V _{GÖVDE_YTU}	2	5	10	8	20	18	16	14	12

6. DENEYSEL SONUÇLAR

Bu bölümde dokümanların türünü, yazarlarını, yazarlarının cinsiyetini belirlemek için yapılan deneysel çalışmaların sonuçlarına değinilmiştir. Deneysel sonuçlar elde edilirken 5. bölümde detaylı olarak açıklanan V_{IST} , V_{IST_CFS} , V_{IST_YTU} , V_{KZ} , V_{DB} , V_{IK} , V_{IK_CFS} , V_{IK_YTU} , V_{2G_CFS} , V_{2G_YTU} , V_{3G_CFS} , V_{3G_YTU} , $V_{KÖK_CFS}$, $V_{KÖK_YTU}$, $V_{GÖVDE_CFS}$, $V_{GÖVDE_YTU}$ özellik vektörleri kullanılmıştır. WEKA içerisinde bulunan, YBS algoritmaları *AIRS1*, *AIRS2* ve *AIRS2Parallel* ile eğiticili öğrenme yöntemleri olan *Naive Bayes*, *Destek Vektör Makinesi*, *K-En Yakın Komşuluk* ve *Rastgele Orman* ile sınıflandırma gerçekleştirilmiştir. K-En Yakın Komşuluk algoritmasında k değeri 1 ve 3 alınarak test edilmiştir. Elde edilen sonuçlar veri kümeleri ve sınıflandırma yöntemleri bazında gösterilmiş, sınıflandırma yöntemlerinin başarı performansları karşılaştırılmıştır.

Deneylerde 10 Kat Çapraz Geçerlilik (10 Fold Cross Validation) uygulanmıştır. K Kat Çapraz Geçerlilik olarak da bilinen bu test yönteminde elimizdeki veri k adet alt kümeye bölünür, bunun $k-1$ parçası eğitim, kalan bir parçası da test kümesi olarak kullanılır. Her seferinde test kümesi değiştirilip kalan $k-1$ parça ile eğitilerek, sistem k defa eğitilmiş olur.

6.1 Cinsiyet Belirleme İçin Deneysel Sonuçlar (VKc)

Çizelge 6.1’de VKc veri kümesinin, 16 özellik vektörü kullanılarak farklı sınıflandırıcılardan alınmış olan başarı sonuçları gösterilmektedir. Buna göre cinsiyet belirlemedeki en yüksek başarı %99.71 ile V_{3G_YTU} özellik vektörü kullanılarak AIRS1 ve AIRS2Parallel ile alınmıştır. Aynı özellik vektörüyle diğer YBS sınıflandırıcısı olan AIRS2 de %99.43 başarı oranı ile KNN ($k=3$) ve DVM ile aynı performansı göstermiştir.

Özellik vektörleri arasında ortalama en yüksek başarı %99.39 ile V_{3G_YTU} vektörüyle elde edilmiştir. Onu %96.36 ile V_{2G_YTU} özellik vektörü takip etmektedir. Türkçe dokümanlar üzerinde yazar cinsiyeti belirleme sistemlerinde, YTU boyut indirgeme yöntemi karakter n-gram’larına (Amasyalı ve Diri, 2006).uygulandığında çok iyi sonuçlar vermektedir. YBS sınıflandırıcıları bu vektörlerle en iyi performansı vermektedir.

Sınıflandırma yöntemlerinin tüm özellik vektörleri ile ortalama başarılarına bakıldığında ise DVM sınıflandırıcısı %84.19 ile en iyi ortalamaı vermektedir. YBS sınıflandırıcıları içerisinde en iyi ortalama %74.74 ile AIRS2Parallel vermiştir.

Sırasıyla Çizelge 6.2’de boyut indirgemesine gidilmeyen dört özellik vektörünün V_{IST} (26 özellik), V_{KZ} (3 özellik), V_{DB} (8 özellik) ve V_{IK} ’nın (154 özellik) başarı sonuçları gösterilmektedir. En yüksek başarı istatistiksel özellikler kullanılarak elde edilen özellik vektörünün, Rastgele Orman sınıflandırıcısından almış olduğu %93.70’lik başarıdır.

Çizelge 6.3’te, Weka içerisinde yer alan CfsSubsetEval boyut indirgeme fonksiyonu V_{IST} , V_{IK} , V_{2G} , V_{3G} , $V_{KÖK}$ ve $V_{GÖVDE}$ özellik vektörlerine uygulandığında elde edilmiş olan yeni altı farklı özellik vektörünün aynı sınıflandırıcılar üzerindeki başarıları gösterilmektedir. Burada en yüksek başarı V_{3G_CFS} özellik vektörü kullanılarak, %97.42 ile DVM sınıflandırıcısından alınmıştır. Çizelge 6.4 ise, aynı özellik vektörlerinin YTU yöntemi kullanılarak boyut azaltılmasına gidildiğinde elde edilen altı yeni özellik vektörünü göstermektedir. Yine en başarılı özellik vektörü V_{3G} ’nin boyutu düşürülmüş halidir. Ancak en iyi sınıflandırıcı AIRS’ler olmuştur. DVM’nin başarısının yükselmesine rağmen, AIRS1 ve AIRS2Parallel ile %99.71’lik bir başarı alınmıştır. CfsSubsetEval kullanıldığında AIRS sınıflandırıcılarının başarıları YTU’ya göre oldukça düşük kalmaktadır.

Çizelge 6.5’te istatistiksel özelliklerden oluşan üç özellik vektörü gruplandırılarak verilmiştir. En iyi performans boyut azaltılmadan V_{IST} ile Rastgele Orman sınıflandırıcısından %93.70 ile alınmıştır. Çizelge 6.6’da işlevsel kelimelerden elde edilmiş üç özellik vektörü verilmiş olup, en yüksek başarı CfsSubsetEval kullanılarak elde edilmiş özellik vektörü ile DVM sınıflandırıcısından %83.38 ile alınmıştır. Çizelge 6.7 ve Çizelge 6.8’de 2 ve 3-gram’lardan elde edilen özellik vektörleri gruplandırılarak verilmiştir. Her ikisinde de en iyi performans YTU kullanılarak alınmıştır. Ancak 2-gram’larda KNN’den ($k=3$) %97.71’lik başarı alınırken, 3-gram’larda AIRS1 ve paralel sınıflandırıcıları %99.71 ile en yüksek performansı vermişlerdir.

Çizelge 6.9’da kelime köklerinden, Çizelge 6.10’da da kelimenin gövdesinden oluşmuş özellik vektörleri verilmektedir. Bu vektörlerin özellik sayıları çok fazla olduğundan denemelerde sadece boyutları azaltılmış olan hali kullanılmaktadır. Her ikisinde de en iyi performans, CfsSubsetEval ile boyutun indirgenmiş halinden elde edilen özellik vektörleri ile DVM sınıflandırıcısından alınmış olup, sırasıyla %93.41 ve %94.56’dır.

Çizelge 6.1: VKc için Genel Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V _{IST}	71.06	78.51	78.51	89.40	91.12	64.18	93.70	80.23	80.84
V _{IST_CFS}	76.22	77.08	76.22	79.66	80.23	71.35	87.68	78.22	78.33
V _{IST_YTU}	69.71	68.29	70	72	75.43	77.14	71.14	80	72.96
V _{KZ}	63.04	60.17	58.45	58.45	60.46	60.46	59.03	56.73	59.60
V _{DB}	75.93	77.65	74.50	73.64	75.64	67.34	76.79	77.08	74.82
V _{IK}	55.59	54.73	51.29	63.04	65.90	77.36	80.52	80.23	66.08
V _{IK_CFS}	74.21	75.07	78.51	72.78	77.36	78.80	78.22	83.38	77.29
V _{IK_YTU}	68.86	67.71	72.29	71.43	76	77.71	71.43	79.71	73.14
V _{2G_CFS}	74.79	79.08	79.66	79.08	82.81	77.36	84.81	85.10	80.34
V _{2G_YTU}	95.71	97.14	94.29	96	97.71	96.86	96	97.14	96.36
V _{3G_CFS}	77.36	77.65	75.93	94.56	95.70	94.84	91.69	97.42	88.14
V _{3G_YTU}	99.71	99.43	99.71	99.14	99.43	99.14	99.14	99.43	99.39
V _{KÖK_CFS}	64.47	70.49	67.05	82.24	84.24	90.26	85.96	93.41	79.77
V _{KÖK_YTU}	78.80	69.63	73.64	75.64	77.36	82.24	76.79	82.52	77.08
V _{GÖVDE_CFS}	63.32	67.62	70.77	83.38	80.80	90.26	90.26	94.56	80.12
V _{GÖVDE_YTU}	73.93	73.64	75.07	77.94	81.95	82.52	77.94	81.95	78.12
Ortalama	73.92	74.62	74.74	79.27	81.38	80.49	82.57	84.19	

Çizelge 6.2: VKc için Boyut İndirgeme Yapılmayan Özellik Vektörleri İçin Başarı Oranları

	AIRS1	AIRS2	AIRS2Paralel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{IST}	71.06	78.51	78.51	89.40	91.12	64.18	93.70	80.23	80.84
V_{KZ}	63.04	60.17	58.45	58.45	60.46	60.46	59.03	56.73	59.60
V_{DB}	75.93	77.65	74.50	73.64	75.64	67.34	76.79	77.08	74.82
V_{İK}	55.59	54.73	51.29	63.04	65.90	77.36	80.52	80.23	66.08
Ortalama	66.41	67.77	65.69	71.13	73.28	67.34	77.51	73.57	

Çizelge 6.3: VKc için CFS ile Boyut İndirgenmiş Özellik Vektörleri İçin Başarı Oranları

	AIRS1	AIRS2	AIRS2Paralel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{IST_CFS}	76.22	77.08	76.22	79.66	80.23	71.35	87.68	78.22	78.33
V_{İK_CFS}	74.21	75.07	78.51	72.78	77.36	78.80	78.22	83.38	77.29
V_{2G_CFS}	74.79	79.08	79.66	79.08	82.81	77.36	84.81	85.10	80.34
V_{3G_CFS}	77.36	77.65	75.93	94.56	95.70	94.84	91.69	97.42	88.14
V_{KÖK_CFS}	64.47	70.49	67.05	82.24	84.24	90.26	85.96	93.41	79.77
V_{GÖVDE_CFS}	63.32	67.62	70.77	83.38	80.80	90.26	90.26	94.56	80.12
Ortalama	71.73	74.50	74.69	81.95	83.52	83.81	86.44	88.68	

Çizelge 6.4: VKc için YTu ile Boyut İndirgenmiş Özellik Vektörleri İçin Başarı Oranları

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{İST_YTU}	69.71	68.29	70	72	75.43	77.14	71.14	80	72.96
V_{İK_YTU}	68.86	67.71	72.29	71.43	76	77.71	71.43	79.71	73.14
V_{2G_YTU}	95.71	97.14	94.29	96	97.71	96.86	96	97.14	96.36
V_{3G_YTU}	99.71	99.43	99.71	99.14	99.43	99.14	99.14	99.43	99.39
V_{KÖK_YTU}	78.80	69.63	73.64	75.64	77.36	82.24	76.79	82.52	77.08
V_{GÖVDE_YTU}	73.93	73.64	75.07	77.94	81.95	82.52	77.94	81.95	78.12
Ortalama	73.92	74.62	74.74	79.27	81.38	80.49	82.57	84.19	

Çizelge 6.5: VKc için İstatistiksel Özellik Vektörleri İçin Başarı Oranları

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{İST}	71.06	78.51	78.51	89.40	91.12	64.18	93.70	80.23	80.84
V_{İST_CFS}	76.22	77.08	76.22	79.66	80.23	71.35	87.68	78.22	78.33
V_{İST_YTU}	69.71	68.29	70	72	75.43	77.14	71.14	80	72.96

Çizelge 6.6: VKc için İşlevsel Kelimeler Özellik Vektörleri İçin Başarı Oranları

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{IK}	55.59	54.73	51.29	63.04	65.90	77.36	80.52	80.23	66.08
V_{IK_CFS}	74.21	75.07	78.51	72.78	77.36	78.80	78.22	83.38	77.29
V_{IK_YTU}	68.86	67.71	72.29	71.43	76	77.71	71.43	79.71	73.14

Çizelge 6.7: VKc için 2-gram Özellik Vektörleri İçin Başarı Oranları

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{2G_CFS}	74.79	79.08	79.66	79.08	82.81	77.36	84.81	85.10	80.34
V_{2G_YTU}	95.71	97.14	94.29	96	97.71	96.86	96	97.14	96.36

Çizelge 6.8: VKc için 3-gram Özellik Vektörleri İçin Başarı Oranları

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{3G_CFS}	77.36	77.65	75.93	94.56	95.70	94.84	91.69	97.42	88.14
V_{3G_YTU}	99.71	99.43	99.71	99.14	99.43	99.14	99.14	99.43	99.39

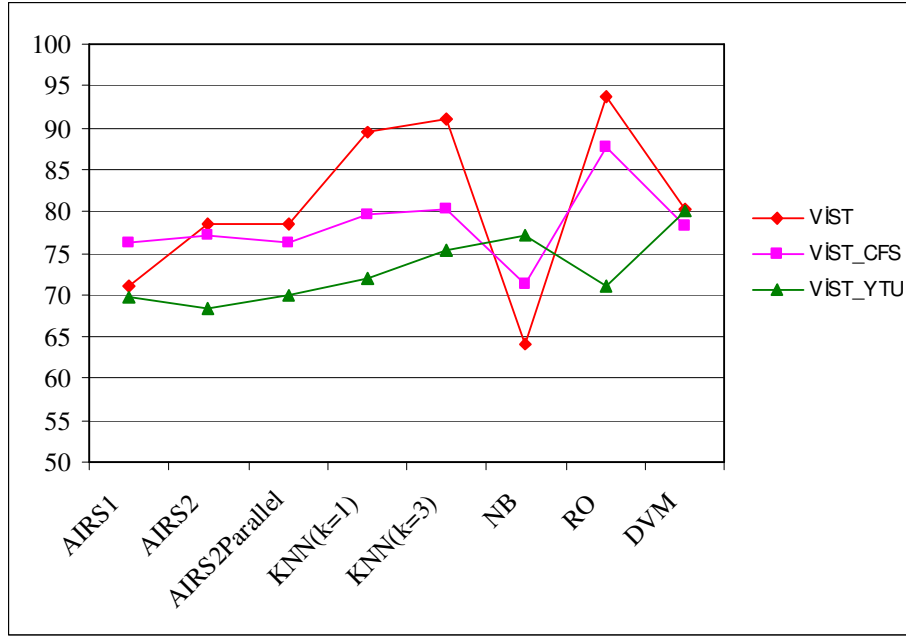
Çizelge 6.9: VKc için Kök Özellik Vektörleri İçin Başarı Oranları

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{KÖK_CFS}	64.47	70.49	67.05	82.24	84.24	90.26	85.96	93.41	79.77
V_{KÖK_YTU}	78.80	69.63	73.64	75.64	77.36	82.24	76.79	82.52	77.08

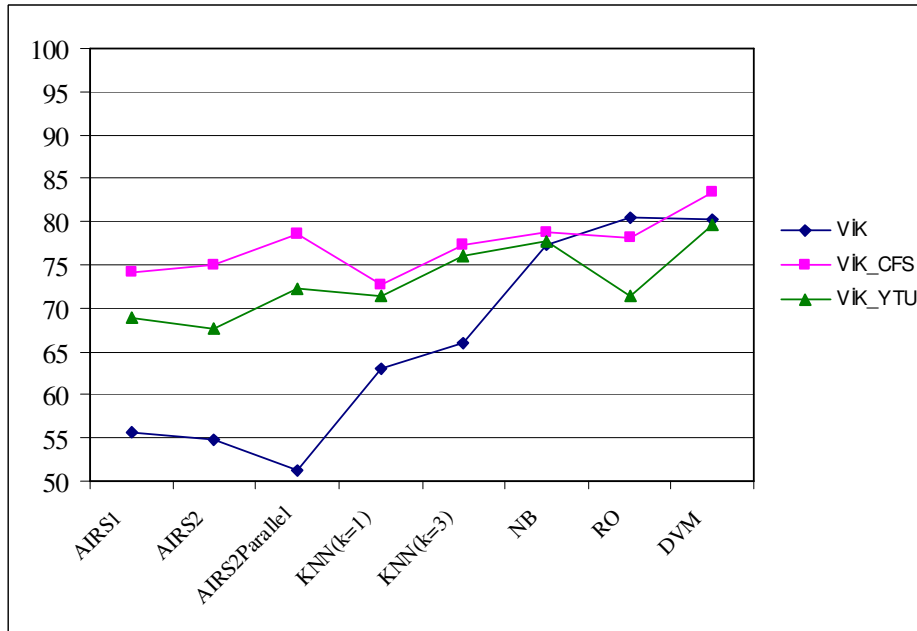
Çizelge 6.10: VKc için Gövde Özellik Vektörleri İçin Başarı Oranları

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{GÖVDE_CFS}	63.32	67.62	70.77	83.38	80.80	90.26	90.26	94.56	80.12
V_{GÖVDE_YTU}	73.93	73.64	75.07	77.94	81.95	82.52	77.94	81.95	78.12

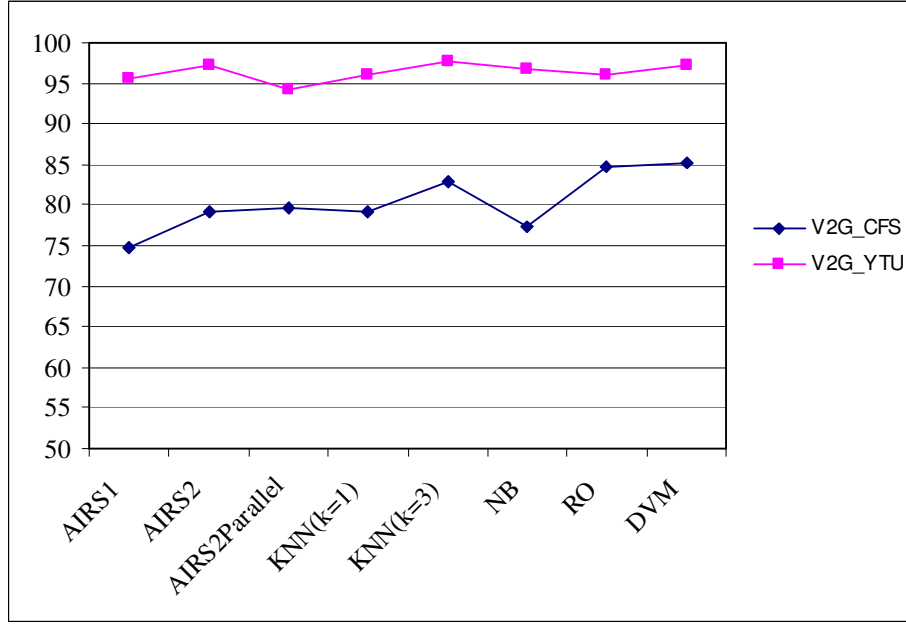
Şekil 6.1'e baktığımızda, V_{IST} özellik vektörünün NB sınıflandırıcısında oldukça başarısız olduğunu, CfsSubsetEval'in YTU'ya göre genelde daha başarılı olduğu gözlenmektedir. Şekil 6.2'de ise, CfsSubsetEval ile boyutu indirgenmiş olan işlevsel kelimelerden oluşan özellik vektörünün, YTU'ya göre genelde daha yüksek başarı verdiği görülmektedir.



Şekil 6.1 VKc için İstatistiksel Özellik Vektörleri Başarı Grafiği

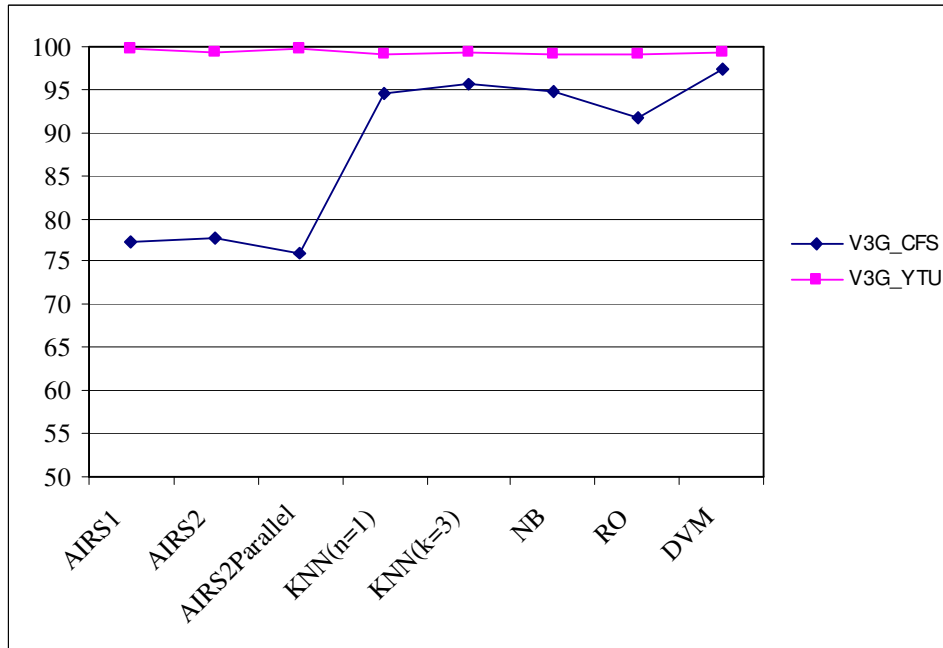


Şekil 6.2 VKc için İşlevsel Kelimeler Özellik Vektörleri Başarı Grafiği

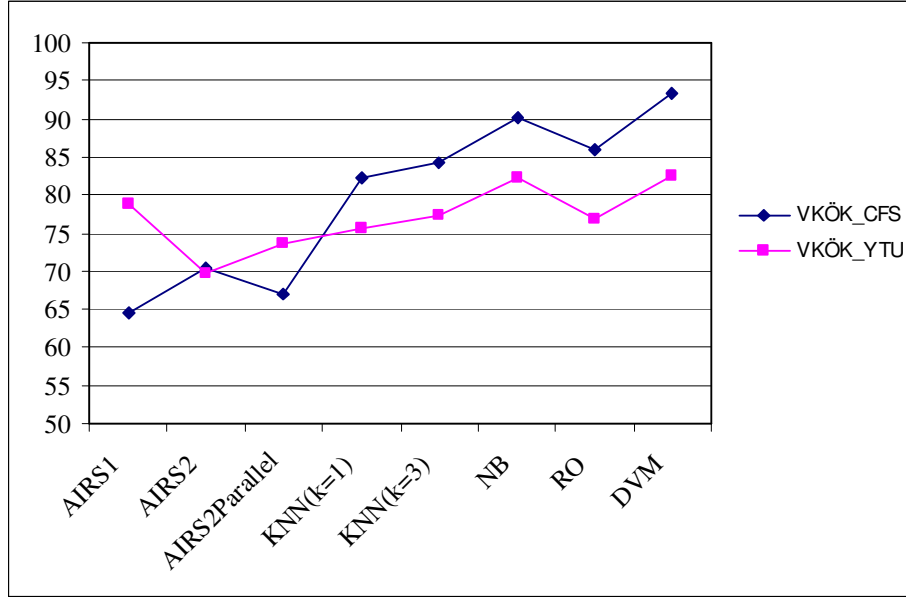


Şekil 6.3 VKc için 2-gram Özellik Vektörleri Başarı Grafiği

Şekil 6.3 ve Şekil 6.4'te ise YTU ile indirgenmiş olan özellik vektörlerinin tüm sınıflandırıcılarda CfsSubsetEval'e göre daha başarılı olduğu görülmektedir.

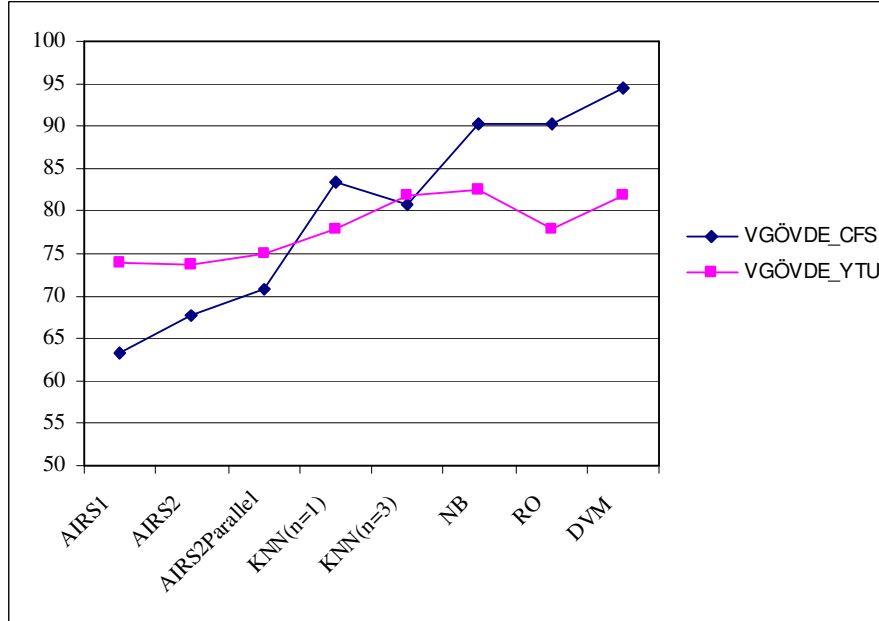


Şekil 6.4 VKc için 3-gram Özellik Vektörleri Başarı Grafiği

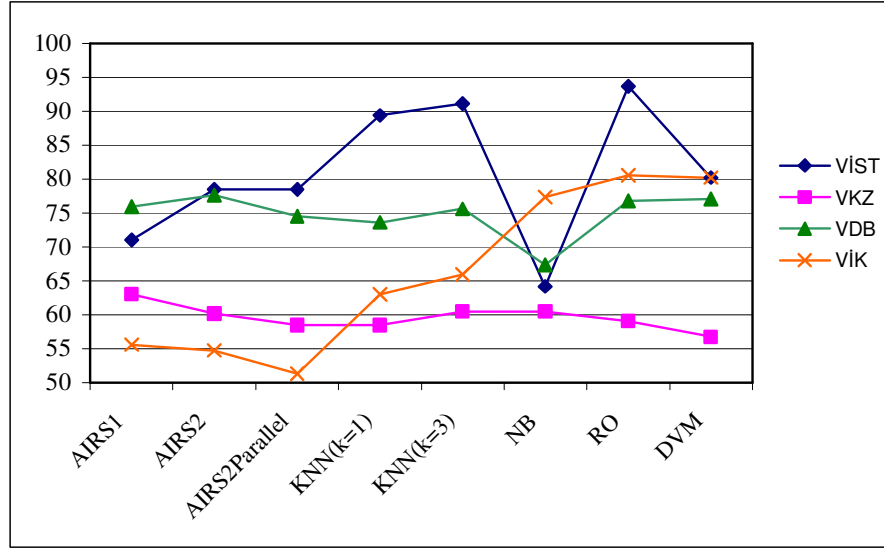


Şekil 6.5 VKc için Kök Özellik Vektörleri Başarı Grafiği

Şekil 6.5'te köklerin ve Şekil 6.6'da, gövdelerin kullanıldığı özellik vektörlerinde büyük bir çoğunlukla CfsSubsetEval daha başarılı olurken, gövdelerin özellik olarak kullanıldığı vektörlerde YTU özellikle AIRS sınıflandırıcılarında başarılı olmuştur.

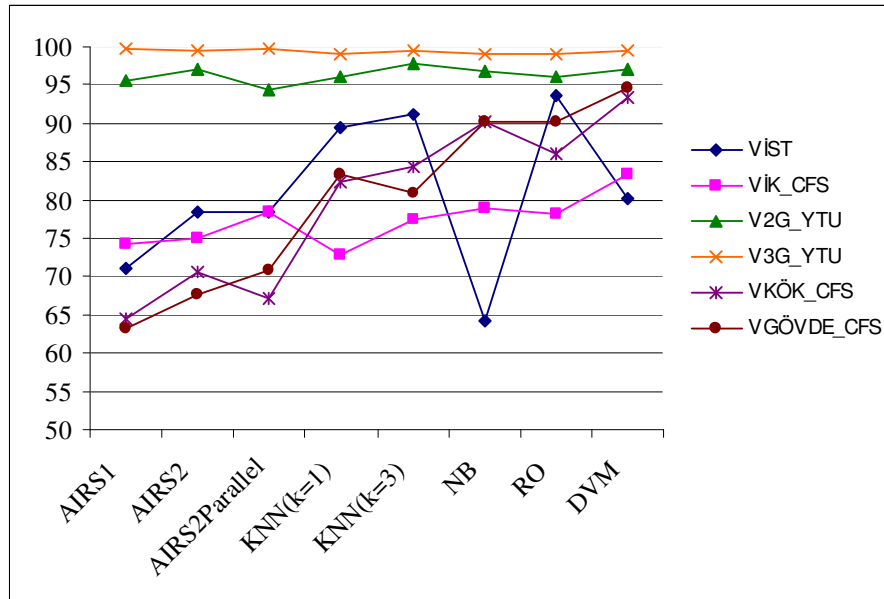


Şekil 6.6 VKc için Gövde Özellik Vektörleri Başarı Grafiği



Şekil 6.7 VKc için Boyut İndirgeme Yapılmayan Özellik Vektörleri Başarı Grafiği

Şekil 6.7’de ise, boyut indirgeme yapılmadan dört farklı özellik vektörünün başarı grafiği gösterilmektedir. Genelde V_{IST} özellik vektörünün başarılı olduğu gözlenmektedir. Şekil 6.8’de ise, kendi gruplarında en yüksek başarıyı veren özellik vektörleri gösterilmektedir. Kısaca, cinsiyet belirlemede en başarılı özellik vektörü V_{3G_YTU} olup, %99.71 ile AIRS1 sınıflandırıcısı vermiştir.



Şekil 6.8 VKc için En Başarılı Özellik Vektörleri Başarı Grafiği

6.2 Tür Belirleme İçin Deneysel Sonuçlar (VK₁)

Doküman türünü belirleme çalışmasında kullanılan sınıflandırıcıların, hangi özellik vektörleriyle yüzde kaçlık sınıflandırma başarıları gösterdikleri Çizelge 6.11’de gösterilmiştir. Buna göre tür belirlemedeki en yüksek başarı %100 ile V_{2G_YTU} özellik vektörü kullanılarak, NB sınıflandırıcısı ile alınmıştır. Aynı özellik vektörüyle YBS sınıflandırıcısı olan AIRS2 de %99.6 başarı oranı ile KNN (k=3) ile aynı performansı göstermiştir.

Özellik vektörleri arasında ortalama en yüksek başarı %99.39 ile V_{3G_YTU} vektörüyle elde edilmiştir. Onu %99.30 ile $V_{GÖVDE_YTU}$, %99.10 ile $V_{KÖK_YTU}$, %98.70 ile V_{2G_YTU} , özellik vektörü takip etmektedir. Türkçe dokümanlar üzerinde doküman türü belirleme sistemlerinde, YTU boyut indirgeme yöntemi karakter n-gram’larına, kelime gövdelerine ve kelime köklerine uygulandığında çok iyi sonuçlar vermektedir. YBS sınıflandırıcıları da bu vektörlerle çok iyi sınıflandırma performansları vermektedir. Örneğin V_{3G_YTU} vektörüyle AIRS1 ve AIRS2Paralel sınıflandırıcıları %99.71 başarı oranıyla en yüksek sınıflandırma performansını vermişlerdir. $V_{GÖVDE_YTU}$ özellik vektörüyle de AIRS2 sınıflandırıcısı %99.6’lık başarı oranıyla KNN ve RO ile birlikte en yüksek sınıflandırma başarılarını sağlamıştır. Yine AIRS2 sınıflandırıcısı V_{2G_YTU} vektörüyle %99.6 başarı oranıyla NB sınıflandırıcısından sonra ikinci en yüksek başarı performansını veren sınıflandırıcı olarak göze çarpmaktadır.

Sınıflandırma yöntemlerinin tüm özellik vektörleri ile ortalama başarılarına bakıldığında ise RO sınıflandırıcısı %82.27 ile en iyi ortalamaı vermektedir. YBS sınıflandırıcıları içerisinde en iyi ortalamaı ise %70.28 ile AIRS2Paralel vermiştir.

Cinsiyet belirleme çalışmasında olduğu gibi tür belirleme çalışmasında da özellik sayısı çok fazla olan özellik vektörlerine boyut indirgeme yöntemleri uygulandığında sınıflandırma performanslarında artış olduğu görülmüştür. Buna karşın özellik sayısı daha az olan özellik vektörlerine CFS ile boyut indirgendiğinde artış, YTU ile boyut değiştirildiğinde ise azalma olmuştur. Örneğin, tür belirleme çalışmasında AIRS1 sınıflandırıcısı için 26 özellik içeren V_{IST} özellik vektörü %58.4, CFS ile özellik indirgenmiş 9 özellik içeren V_{IST_CFS} özellik vektörü %70, YTU yöntemiyle boyut değiştirilmiş 5 özellikten oluşan V_{IST_YTU} özellik vektörüyle ise %46.8 başarı elde edilmiştir. Bunun yanında 152 özellik içeren V_{IK} özellik vektörü %23.6, CFS ile özellik indirgenmiş 19

özellik içeren V_{IK_CFS} özellik vektörü %54.8, YTU yöntemiyle boyut değiştirilmiş 5 özellikten oluşan V_{IK_YTU} özellik vektörüyle ise %76.8 başarı elde edilmiştir.

Sırasıyla Çizelge 6.12’de boyut indirgemesi yapılmayan dört özellik vektörünün V_{IST} (26 özellik), V_{KZ} (3 özellik), V_{DB} (8 özellik) ve V_{IK} ’nın (154 özellik) başarı sonuçları gösterilmektedir. En yüksek başarı yine istatistiksel özellikler kullanılarak elde edilen özellik vektörünün, Rastgele Orman sınıflandırıcısından almış olduğu %87.60’lık başarıdır. Bu başarının cinsiyet tanımadan daha düşük olmasının sebebi, iki sınıf yerine beş sınıf olmasıdır.

Çizelge 6.13’te, WEKA içerisinde yer alan CfsSubsetEval boyut indirgeme fonksiyonu V_{IST} , V_{IK} , V_{2G} , V_{3G} , $V_{KÖK}$ ve $V_{GÖVDE}$ özellik vektörlerine uygulandığında elde edilmiş olan yeni altı farklı özellik vektörünün aynı sınıflandırıcılar üzerindeki başarıları gösterilmektedir. Burada en yüksek başarı $V_{GÖVDE_CFS}$ özellik vektörü kullanılarak, %94 ile DVM sınıflandırıcısından alınmıştır. Cinsiyet belirlemede de en iyi sınıflandırıcı DVM olmasına karşılık başarı daha yüksekti. Çizelge 6.14 ise, aynı özellik vektörlerinin YTU yöntemi kullanılarak boyut azaltılmasına gidildiğinde elde edilen altı yeni özellik vektörünü göstermektedir. En başarılı özellik vektörü V_{2G_YTU} olup, NB sınıflandırıcısı ile %100’lük başarı elde edilmiştir. Cinsiyet belirlemede en iyi başarıyı veren V_{3G_YTU} özellik vektörü burada da NB’in arkasından Yapay Bağışılık sınıflandırıcıları ile alınmıştır. Hem AIRS1 hem de AIRS2Parallel ile %99.71’lik bir başarı vermiştir.

Çizelge 6.15’te istatistiksel özelliklerden oluşan üç özellik vektörü gruplandırılarak verilmiştir. En iyi performans boyut azaltılmadan V_{IST} ile Rastgele Orman ve V_{IST_CFS} ile KNN ($k=3$) sınıflandırıcısından %87.6 ile alınmıştır. Çizelge 6.16’da işlevsel kelimelerden elde edilmiş üç özellik vektörü verilmiş olup, en yüksek başarı cinsiyet tanımadakinin tersine YTU kullanılarak elde edilmiş özellik vektörü ile yine DVM sınıflandırıcısından %82.4 ile alınmıştır. Çizelge 6.17 ve Çizelge 6.18’de 2 ve 3-gram’lardan elde edilen özellik vektörleri gruplandırılarak verilmiştir. Her ikisinde de en iyi performans YTU kullanılarak alınmıştır. Ancak 2-gram’larda NB ile %100’lük başarı alınırken, AIRS2, %99.6 ile ikinci sırada yer almaktadır. 3-gram’larda ise AIRS1 ve paralel sınıflandırıcıları %99.71 ile en yüksek performansı vermişlerdir.

Çizelge 6.19’da kelime köklerinden, Çizelge 6.20’de de kelimenin gövdesinden oluşmuş özellik vektörleri verilmektedir. Bu vektörlerin özellik sayıları çok fazla olduğundan

denemelerde sadece boyutları azaltılmış olan hali kullanılmaktadır. Her ikisinde de en iyi performans cinsiyet tanımanın tersine YTU ile boyutun indirgenmiş halinden elde edilen özellik vektörlerinden alınmıştır. $V_{KÖK_YTU}$ özellik vektörü en yüksek başarıyı %99.6 ile AIRS2Parallel ve KNN (k=3) vermiş iken $V_{GÖVDE_YTU}$ vektörü ise KNN, RO ve AIRS2 ile aynı başarıyı vermiştir.

Çizelge 6.11: VK_i için Genel Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V _{IST}	58.4	64	63.6	80	76.8	83.6	87.6	78.4	74.05
V _{IST_CFS}	70	73.2	74.8	84	87.6	84.8	86	72.4	79.10
V _{IST_YTU}	46.8	42.4	54.4	60.8	58.8	51.6	63.6	35.2	51.70
V _{KZ}	44.4	47.6	44	48.8	50.8	52	55.2	44.4	48.40
V _{DB}	48.8	56.4	58	62.8	63.6	55.2	68.4	40.8	56.75
V _{IK}	23.6	30	30.8	46.8	48	66	64	67.2	47.05
V _{IK_CFS}	54.8	50.4	56.4	60.8	63.6	68	65.6	67.2	60.85
V _{IK_YTU}	76.8	78.8	78.4	77.6	78.8	81.6	79.2	82.4	79.20
V _{2G_CFS}	74.4	76.4	76.4	61.2	88.4	88.4	85.2	87.6	79.75
V _{2G_YTU}	98.8	99.6	99.2	96	99.6	100	98.4	98	98.70
V _{3G_CFS}	44.4	52	65.2	58.4	91.2	90	92.4	87.2	72.60
V _{3G_YTU}	99.71	99.43	99.71	99.14	99.43	99.14	99.14	99.43	99.39
V _{KÖK_CFS}	62	56.8	63.2	80.4	84	92.8	87.2	92.4	77.35
V _{KÖK_YTU}	99.2	99.2	99.6	99.2	99.6	98.4	98.4	99.2	99.10
V _{GÖVDE_CFS}	60.4	58.4	61.6	85.2	84	93.6	86.4	94	77.95
V _{GÖVDE_YTU}	99.2	99.6	99.2	99.6	99.6	98.8	99.6	98.8	99.30
Ortalama	66.36	67.76	70.28	75.05	79.61	81.50	82.27	77.79	

Çizelge 6.12: VK_i için Boyut İndirgeme Yapılmayan Özellik Vektörleri İçin Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V _{IST}	58.4	64	63.6	80	76.8	83.6	87.6	78.4	74.05
V _{KZ}	44.4	47.6	44	48.8	50.8	52	55.2	44.4	48.40
V _{DB}	48.8	56.4	58	62.8	63.6	55.2	68.4	40.8	56.75
V _{IK}	23.6	30	30.8	46.8	48	66	64	67.2	47.05
Ortalama	43.8	49.5	49.1	59.6	59.8	64.2	82.27	68.8	

Çizelge 6.13: VK_i için CFS ile Boyut İndirgenmiş Özellik Vektörleri İçin Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V _{IST_CFS}	70	73.2	74.8	84	87.6	84.8	86	72.4	79.10
V _{IK_CFS}	54.8	50.4	56.4	60.8	63.6	68	65.6	67.2	60.85
V _{2G_CFS}	74.4	76.4	76.4	61.2	88.4	88.4	85.2	87.6	79.75
V _{3G_CFS}	44.4	52	65.2	58.4	91.2	90	92.4	87.2	72.60
V _{KÖK_CFS}	62	56.8	63.2	80.4	84	92.8	87.2	92.4	77.35
V _{GÖVDE_CFS}	60.4	58.4	61.6	85.2	84	93.6	86.4	94	77.95
Ortalama	61	61.2	66.27	71.67	83.13	86.27	83.8	83.47	

Çizelge 6.14: VK_i için YTU ile Boyut İndirgenmiş Özellik Vektörleri İçin Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Paralel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V _{İST_YTU}	46.8	42.4	54.4	60.8	58.8	51.6	63.6	35.2	51.70
V _{İK_YTU}	76.8	78.8	78.4	77.6	78.8	81.6	79.2	82.4	79.20
V _{2G_YTU}	98.8	99.6	99.2	96	99.6	100	98.4	98	98.70
V _{3G_YTU}	99.71	99.43	99.71	99.14	99.43	99.14	99.14	99.43	99.39
V _{KÖK_YTU}	99.2	99.2	99.6	99.2	99.6	98.4	98.4	99.2	99.10
V _{GÖVDE_YTU}	99.2	99.6	99.2	99.6	99.6	98.8	99.6	98.8	99.30
Ortalama	86.75	86.51	88.42	88.72	89.31	88.26	89.72	85.51	

Çizelge 6.15: VK_i için İstatistiksel Özellik Vektörleri İçin Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Paralel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V _{İST}	58.4	64	63.6	80	76.8	83.6	87.6	78.4	74.05
V _{İST_CFS}	70	73.2	74.8	84	87.6	84.8	86	72.4	79.10
V _{İST_YTU}	46.8	42.4	54.4	60.8	58.8	51.6	63.6	35.2	51.70

Çizelge 6.16: VK_t için İşlevsel Kelimeler Özellik Vektörleri İçin Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{IK}	23.6	30	30.8	46.8	48	66	64	67.2	47.05
V_{IK_CFS}	54.8	50.4	56.4	60.8	63.6	68	65.6	67.2	60.85
V_{IK_YTU}	76.8	78.8	78.4	77.6	78.8	81.6	79.2	82.4	79.20

Çizelge 6.17: VK_t için 2-Gram Özellik Vektörleri İçin Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{2G_CFS}	74.4	76.4	76.4	61.2	88.4	88.4	85.2	87.6	79.75
V_{2G_YTU}	98.8	99.6	99.2	96	99.6	100	98.4	98	98.70

Çizelge 6.18: VK_t için 3-Gram Özellik Vektörleri İçin Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{3G_CFS}	44.4	52	65.2	58.4	91.2	90	92.4	87.2	72.60
V_{3G_YTU}	99.71	99.43	99.71	99.14	99.43	99.14	99.14	99.43	99.39

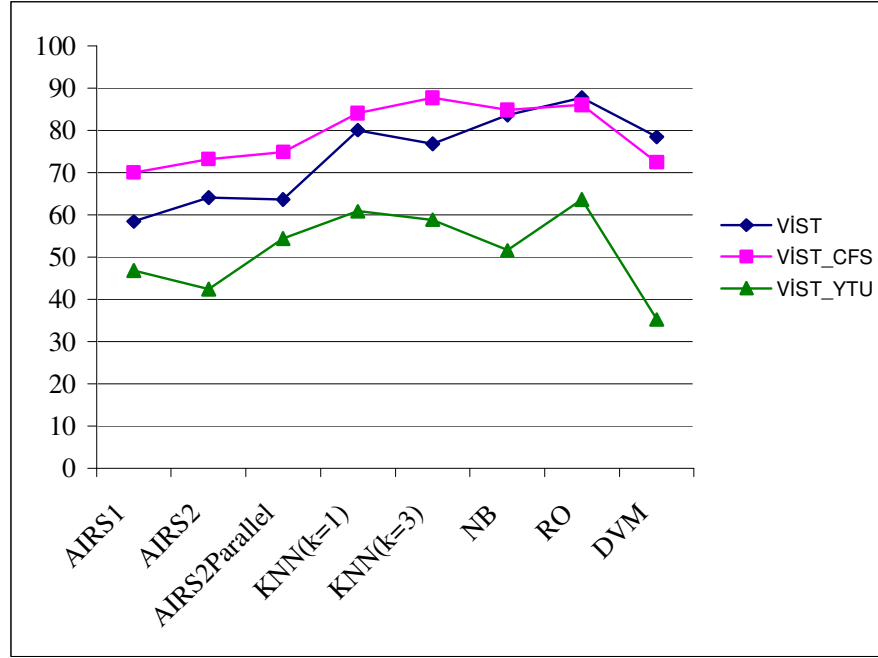
Çizelge 6.19: VK_t için Kök Özellik Vektörleri İçin Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{KÖK_CFS}	62	56.8	63.2	80.4	84	92.8	87.2	92.4	77.35
V_{KÖK_YTU}	99.2	99.2	99.6	99.2	99.6	98.4	98.4	99.2	99.10

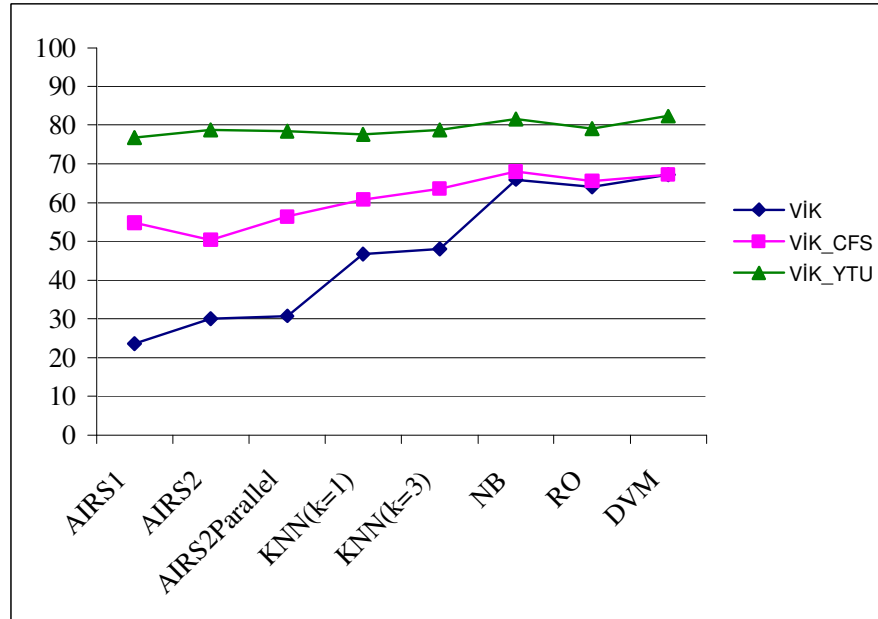
Çizelge 6.20: VK_t için Gövde Özellik Vektörleri İçin Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{GÖVDE_CFS}	60.4	58.4	61.6	85.2	84	93.6	86.4	94	77.95
V_{GÖVDE_YTU}	99.2	99.6	99.2	99.6	99.6	98.8	99.6	98.8	99.30

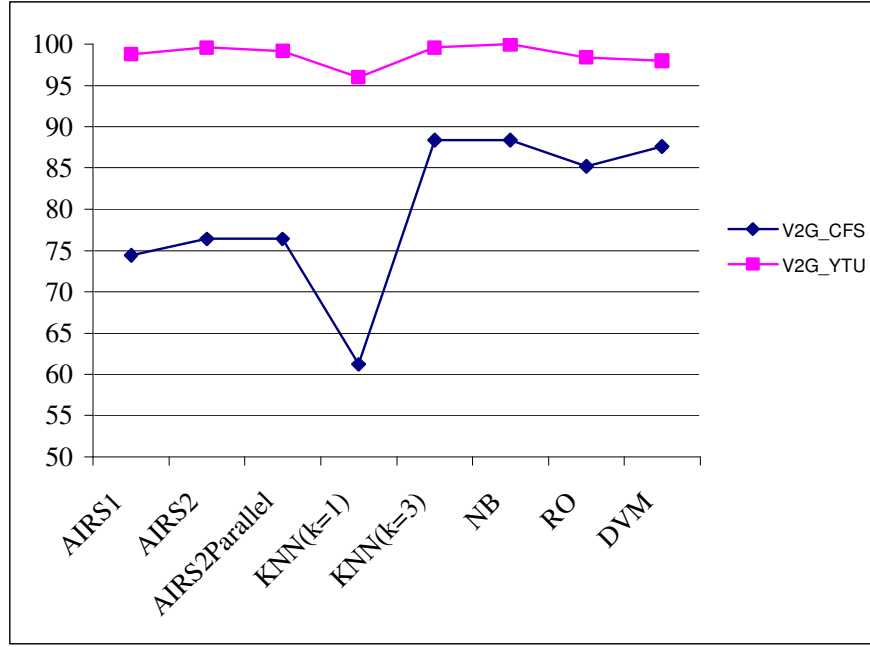
Şekil 6.9'a baktığımızda, CfsSubsetEval kullanılarak özellik sayısı azaltılan V_{IST} vektörünün daha başarılı olduğunu, Şekil 6.10'da da işlevsel kelimelerde boyut azaltmada YTU yönteminin kullanılması başarıyı gözle görülür bir şekilde yükseltmiştir.



Şekil 6.9 V_{K_i} için İstatistiksel Özellik Vektörleri Başarı Grafiği

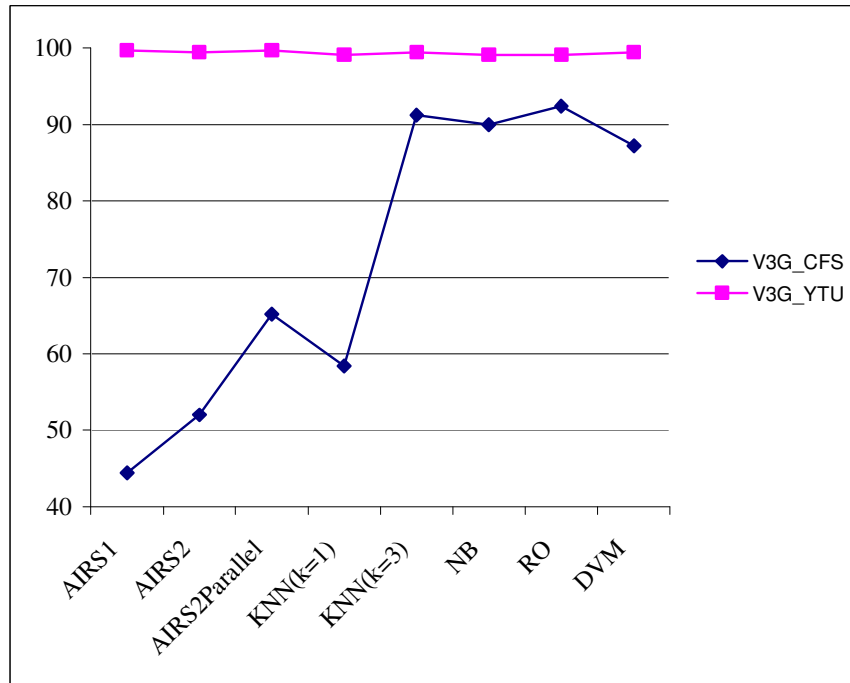


Şekil 6.10 V_{K_i} için İşlevsel Kelimeler Özellik Vektörleri Başarı Grafiği

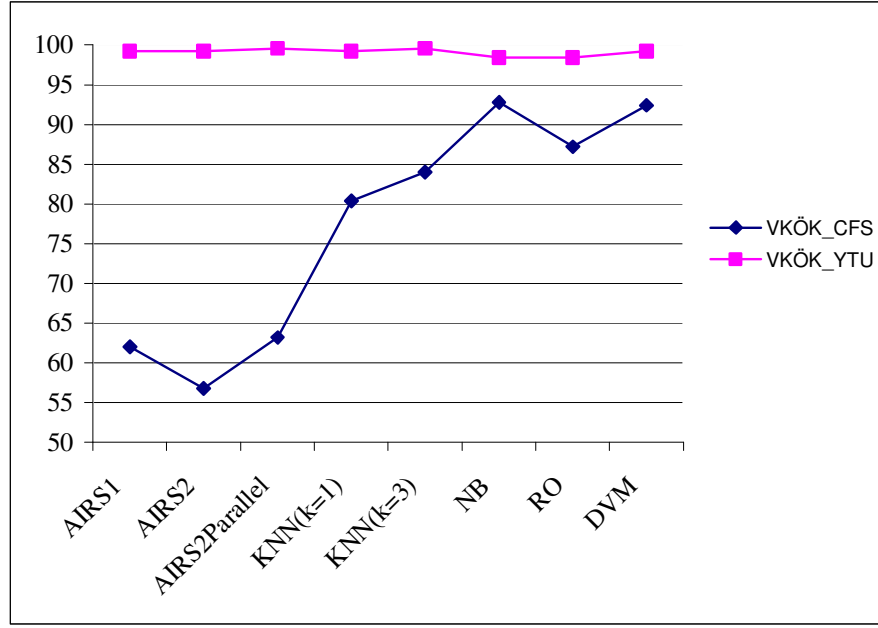


Şekil 6.11 VK_t için 2-gram Özellik Vektörleri Başarı Grafiği

Şekil 6.11 ve Şekil 6.12’de ise yine cinsiyet belirlemede olduğu gibi Y2U ile indirgenmiş olan özellik vektörlerinin tüm sınıflandırıcılarda CfsSubsetEval’e göre daha başarılı olduğu görülmektedir.

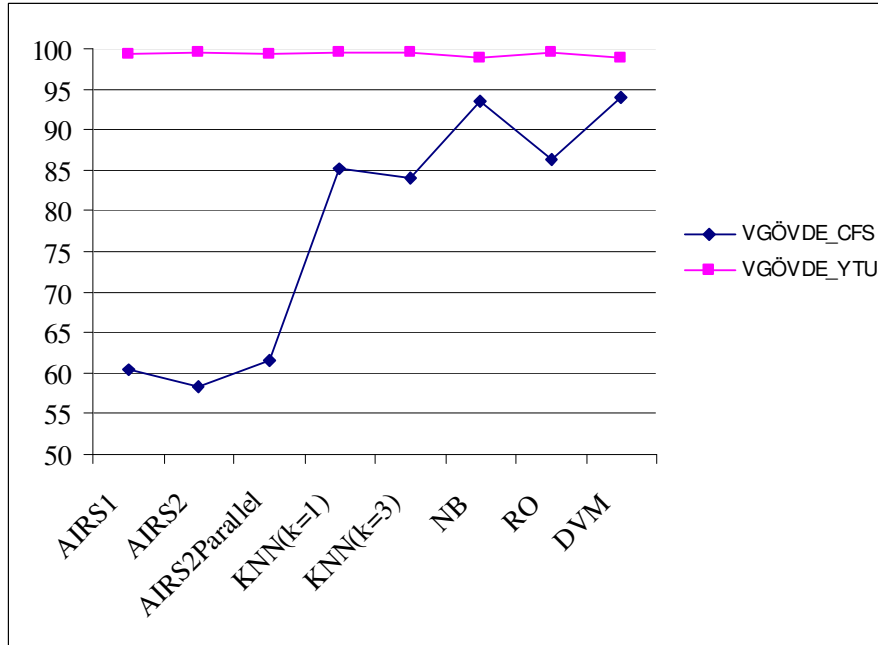


Şekil 6.12 VK_t için 3-gram Özellik Vektörleri Başarı Grafiği

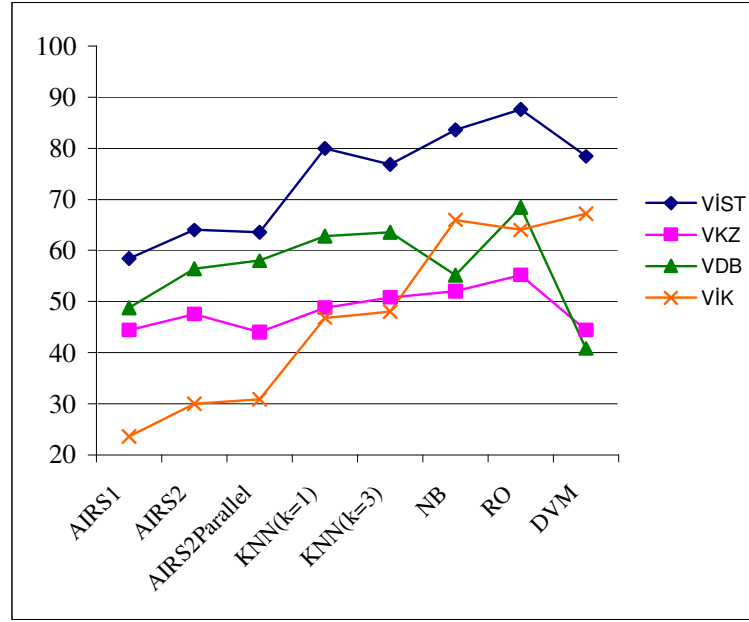


Şekil 6.13 VK_t için Kök Özellik Vektörleri Başarı Grafiği

Şekil 6.13'te köklerin ve Şekil 6.14'te gövdelerin kullanıldığı özellik vektörlerinde YTU'nun her sınıflandırıcıda yüksek bir performans sergilediği açıkça görülmektedir.

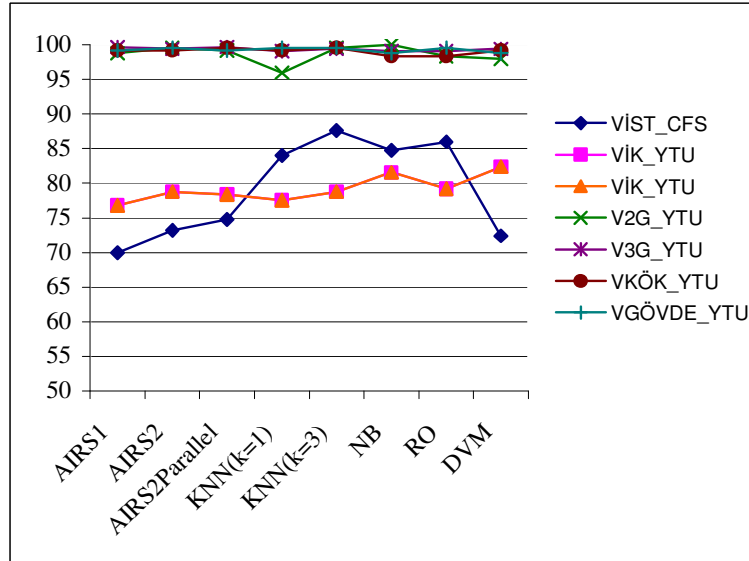


Şekil 6.14 VK_t için Gövde Özellik Vektörleri Başarı Grafiği



Şekil 6.15 VK_t için Boyut İndirgeme Yapılmayan Özellik Vektörleri Başarı Grafiği

Şekil 6.15’de, boyut indirgeme yapılmadan dört farklı özellik vektörünün başarı grafiği gösterilmektedir. V_{IST} özellik vektörü cinsiyet belirlemede olduğu gibi tür tanımda da başarılı olmuştur. Şekil 6.16’da ise, kendi gruplarında en yüksek başarıyı veren özellik vektörleri gösterilmektedir. Tür belirlemede V_{3G_YTU} , V_{2G_YTU} , $V_{KÖK_YTU}$ ve $V_{GÖVDE_YTU}$ özellik vektörleri bütün sınıflandırıcılarda birbirine yakın başarılar elde etmişlerdir.



Şekil 6.16 VK_t için En Başarılı Özellik Vektörleri Başarı Grafiği

6.3 Yazar Tanıma İçin Deneysel Sonuçlar

Yazar tanıma çalışması için deneysel sonuçlar aynı türde yazan yazarlar ve farklı türlerde yazan yazarlar olmak üzere iki kısımda incelenmiştir.

6.3.1 Aynı Türde Yazar Tanıma İçin Deneysel Sonuçlar

Aynı türde yazan yazar tanıma çalışması için deneysel sonuçlar 10 yazar ve 8 yazar içeren iki grup altında toplanmıştır.

6.3.1.1 Aynı Türde Yazar Tanıma İçin Deneysel Sonuçlar ($V_{K_{ya10}}$)

Aynı türde yazılar yazan 10 yazardan oluşan yazar tanıma çalışmasında kullanılan sınıflandırıcıların, hangi özellik vektörleriyle yüzde kaçlık sınıflandırma başarısı gösterdikleri Çizelge 6.21’de gösterilmiştir. Buna göre en yüksek başarı %99.71 ile V_{3G_YTU} özellik vektörü kullanılarak AIRS1, AIRS2, AIRS2Parallel ve KNN sınıflandırıcıları ile alınmıştır. YBS sınıflandırıcılarının tümünün V_{3G_YTU} özellik vektörüyle %99.71 başarı oranı vererek en iyi performansı göstermesi, yazar tanıma çalışmalarında YBS sınıflandırıcılarının uygun özellik vektörleri kullanılarak tercih edilebileceğini göstermektedir.

Tüm özellik vektörleri arasında ortalama en yüksek başarı %99.46 ile V_{3G_YTU} vektörüyle elde edilmiştir. Onu %93.64 ile V_{2G_YTU} özellik vektörü takip etmektedir. Cinsiyet ve tür belirleme çalışmalarında karakter n-gram’larıyla yakın başarı performansı sergileyen kök ve gövde özellik vektörleri ise yazar tanıma çalışmalarında sınıf sayısının artmasına paralel olarak n-gram özellik vektörlerinin gerisinde kalmıştır. $V_{KÖK_YTU}$ özellik vektörü %83.21, $V_{GÖVDE_YTU}$ özellik vektörü ise % 87.22 ortalama sınıflandırma başarısı göstermiştir.

Sınıflandırma yöntemlerinin tüm özellik vektörleri ile ortalama başarılarına bakıldığında ise DVM sınıflandırıcısı %70.95 ile en iyi ortalamayı vermektedir. YBS sınıflandırıcıları içerisinde en iyi ortalamayı ise %57.88 ile AIRS2Parallel vermiştir.

Çizelge 6.21: VK_{ya10} için Genel Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Paralel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V _{IST}	64.67	66.95	63.82	75.21	78.63	82.62	82.91	80.91	74.47
V _{IST_CFS}	70.94	74.07	72.93	76.64	79.49	82.05	83.19	79.20	77.31
V _{IST_YTU}	52.57	53.71	57.43	43.71	58.29	58.86	56.86	59.71	55.14
V _{KZ}	40.17	42.17	43.02	49.86	48.43	45.30	53.56	29.34	43.98
V _{DB}	40.29	35.43	38.29	40.86	38.29	50.29	55.14	42	42.57
V _{IK}	13.39	18.52	20.80	38.46	34.47	51.85	41.31	57.27	34.51
V _{IK_CFS}	38.46	35.90	34.76	37.61	34.47	52.42	39.32	45.58	39.82
V _{IK_YTU}	51.43	50.86	49.14	34.86	50.57	51.71	58.86	59.71	50.89
V _{2G_CFS}	49.14	52.29	51.14	55.71	59.71	73.14	70.57	72.29	60.50
V _{2G_YTU}	93.71	93.71	93.71	97.43	97.14	83.14	94	96.29	93.64
V _{3G_CFS}	45.71	45.43	54.57	68.29	69.14	80.57	76.57	85.43	65.71
V _{3G_YTU}	99.71	99.71	99.71	99.43	99.71	99.43	98.57	99.43	99.46
V _{KÖK_CFS}	33.33	36.18	39.60	52.71	46.44	69.80	55.27	71.23	50.57
V _{KÖK_YTU}	80	82	82.29	80.57	85.71	81.71	84	89.43	83.21
V _{GÖVDE_CFS}	34.29	35.71	38	52.57	48.29	72.86	53.14	75.71	53.75
V _{GÖVDE_YTU}	86.29	87.14	86.86	86.29	88	85.43	86	91.71	87.22
Ortalama	57.32	56.86	57.88	61.89	63.55	70.07	68.08	70.95	

Sırasıyla Çizelge 6.22’de boyut indirgemesi yapılmayan dört özellik vektörünün V_{IST} (26 özellik), V_{KZ} (3 özellik), V_{DB} (8 özellik) ve V_{IK} ’nın (154 özellik) başarı sonuçları gösterilmektedir. En yüksek başarı cinsiyet ve tür tanımda olduğu gibi yine istatistiksel özellikler kullanılarak elde edilen özellik vektörünün, Rastgele Orman sınıflandırıcısından almış olduğu %82.91’lik başarıdır. Yapay Bağışıklık sınıflandırıcılarının performansı oldukça düşüktür.

Çizelge 6.23’te, Weka içerisinde yer alan CfsSubsetEval boyut indirgeme fonksiyonu V_{IST} , V_{IK} , V_{2G} , V_{3G} , $V_{KÖK}$ ve $V_{GÖVDE}$ özellik vektörlerine uygulandığında elde edilmiş olan yeni altı farklı özellik vektörünün aynı sınıflandırıcılar üzerindeki başarıları gösterilmektedir. Burada en yüksek başarı V_{3G_CFS} özellik vektörü kullanılarak, %85.43 ile DVM sınıflandırıcısından alınmıştır. Cinsiyet ve tür belirlemede de en iyi sınıflandırıcı DVM olmuştur. Yine burada Yapay Bağışıklık sınıflandırıcıları başarısız kalmışlardır. Çizelge 6.24 ise, aynı özellik vektörlerinin YTU yöntemi kullanılarak boyut azaltılmasına gidildiğinde elde edilen altı yeni özellik vektörünü göstermektedir. En başarılı özellik vektörü V_{3G_YTU} olup, KNN (k=3) ve AIRS sınıflandırıcıları %99.71 ile aynı başarıyı göstermişlerdir. Tür ve cinsiyet belirlemede de en iyi başarıyı veren V_{3G_YTU} özellik vektörü olmuştur.

Çizelge 6.25’te istatistiksel özelliklerden oluşan üç özellik vektörü gruplandırılarak verilmiştir. En iyi performans, V_{IST_CFS} vektörü ile %83.19 ile RO sınıflandırıcısından alınmıştır. Çizelge 6.26’da işlevsel kelimelerden elde edilmiş üç özellik vektörü verilmiş olup, en yüksek başarı tür tanımda olduğu gibi YTU kullanılarak elde edilmiş özellik vektörü ile RO sınıflandırıcısından %58.86 ile alınmıştır. İşlevsel Kelimelerden oluşan özellik vektörleri aynı türde yazan yazarların tanınmasında oldukça başarısız kalmıştır. Çizelge 6.27 ve Çizelge 6.28’de 2 ve 3-gram’lardan elde edilen özellik vektörleri gruplandırılarak verilmiştir. Her ikisinde de en iyi performans YTU kullanılarak, AIRS sınıflandırıcıları ile alınmıştır. V_{2G_YTU} kullanıldığında AIRS sınıflandırıcıları %93.71’lik bir başarı verirken, V_{3G_YTU} kullanılarak aynı sınıflandırıcılardan %99.71’lik bir başarı elde edilmiştir.

Çizelge 6.29’da kelime köklerinden, Çizelge 6.30’da da kelimenin gövdesinden oluşmuş özellik vektörleri verilmektedir. Her ikisinde de en iyi performans, YTU ile boyutun indirgenmiş halinden elde edilen özellik vektörlerinden alınmıştır. $V_{KÖK_YTU}$ özellik vektörü en yüksek başarıyı DVM sınıflandırıcı ile %89.43 olarak elde ederken yine aynı sınıflandırıcı ile $V_{GÖVDE_YTU}$ vektörü %91.71’lik bir başarı vermiştir.

Çizelge 6.22: VK_{ya10} için Boyut İndirgeme Yapılmayan Özellik Vektörleri İçin Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{IST}	64.67	66.95	63.82	75.21	78.63	82.62	82.91	80.91	74.47
V_{KZ}	40.17	42.17	43.02	49.86	48.43	45.30	53.56	29.34	43.98
V_{DB}	40.29	35.43	38.29	40.86	38.29	50.29	55.14	42	42.57
V_{IK}	13.39	18.52	20.80	38.46	34.47	51.85	41.31	57.27	34.51
Ortalama	39.63	40.77	41.48	51.09	49.95	57.51	58.23	52.38	

Çizelge 6.23: VK_{ya10} için CFS ile Boyut İndirgenmiş Özellik Vektörleri İçin Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{IST_CFS}	70.94	74.07	72.93	76.64	79.49	82.05	83.19	79.20	77.31
V_{IK_CFS}	38.46	35.90	34.76	37.61	34.47	52.42	39.32	45.58	39.82
V_{2G_CFS}	49.14	52.29	51.14	55.71	59.71	73.14	70.57	72.29	60.50
V_{3G_CFS}	45.71	45.43	54.57	68.29	69.14	80.57	76.57	85.43	65.71
V_{KÖK_CFS}	33.33	36.18	39.60	52.71	46.44	69.80	55.27	71.23	50.57
V_{GÖVDE_CFS}	34.29	35.71	38	52.57	48.29	72.86	53.14	75.71	53.75
Ortalama	45.31	46.59	48.5	57.25	56.25	71.80	63.01	71.57	

Çizelge 6.24: VK_{ya10} için YTU ile Boyut İndirgenmiş Özellik Vektörleri İçin Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{IST_YTU}	52.57	53.71	57.43	43.71	58.29	58.86	56.86	59.71	55.14
V_{IK_YTU}	51.43	50.86	49.14	34.86	50.57	51.71	58.86	59.71	50.89
V_{2G_YTU}	93.71	93.71	93.71	97.43	97.14	83.14	94	96.29	93.64
V_{3G_YTU}	99.71	99.71	99.71	99.43	99.71	99.43	98.57	99.43	99.46
V_{KÖK_YTU}	80	82	82.29	80.57	85.71	81.71	84	89.43	83.21
V_{GÖVDE_YTU}	86.29	87.14	86.86	86.29	88	85.43	86	91.71	87.22
Ortalama	57.32	56.86	57.88	61.89	63.55	70.07	68.08	70.95	

Çizelge 6.25: VK_{ya10} için İstatistiksel Özellik Vektörleri İçin Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{IST}	64.67	66.95	63.82	75.21	78.63	82.62	82.91	80.91	74.47
V_{IST_CFS}	70.94	74.07	72.93	76.64	79.49	82.05	83.19	79.20	77.31
V_{IST_YTU}	52.57	53.71	57.43	43.71	58.29	58.86	56.86	59.71	55.14

Çizelge 6.26: VK_{ya10} için İşlevsel Kelimeler Özellik Vektörleri İçin Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Paralel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{iK}	13.39	18.52	20.80	38.46	34.47	51.85	41.31	57.27	34.51
V_{iK_CFS}	38.46	35.90	34.76	37.61	34.47	52.42	39.32	45.58	39.82
V_{iK_YTU}	51.43	50.86	49.14	34.86	50.57	51.71	58.86	59.71	50.89

Çizelge 6.27: VK_{ya10} için 2-gram Özellik Vektörleri İçin Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Paralel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{2G_CFS}	49.14	52.29	51.14	55.71	59.71	73.14	70.57	72.29	60.50
V_{2G_YTU}	93.71	93.71	93.71	97.43	97.14	83.14	94	96.29	93.64

Çizelge 6.28: VK_{ya10} için 3-gram Özellik Vektörleri İçin Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Paralel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{3G_CFS}	45.71	45.43	54.57	68.29	69.14	80.57	76.57	85.43	65.71
V_{3G_YTU}	99.71	99.71	99.71	99.43	99.71	99.43	98.57	99.43	99.46

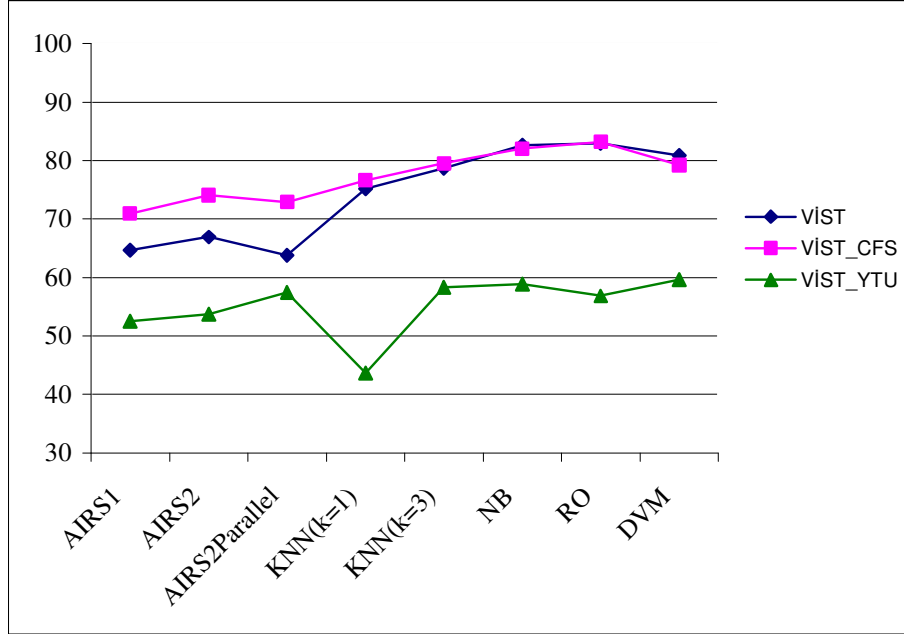
Çizelge 6.29: VK_{ya10} için Kök Özellik Vektörleri İçin Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{KÖK_CFS}	33.33	36.18	39.60	52.71	46.44	69.80	55.27	71.23	50.57
V_{KÖK_YTU}	80	82	82.29	80.57	85.71	81.71	84	89.43	83.21

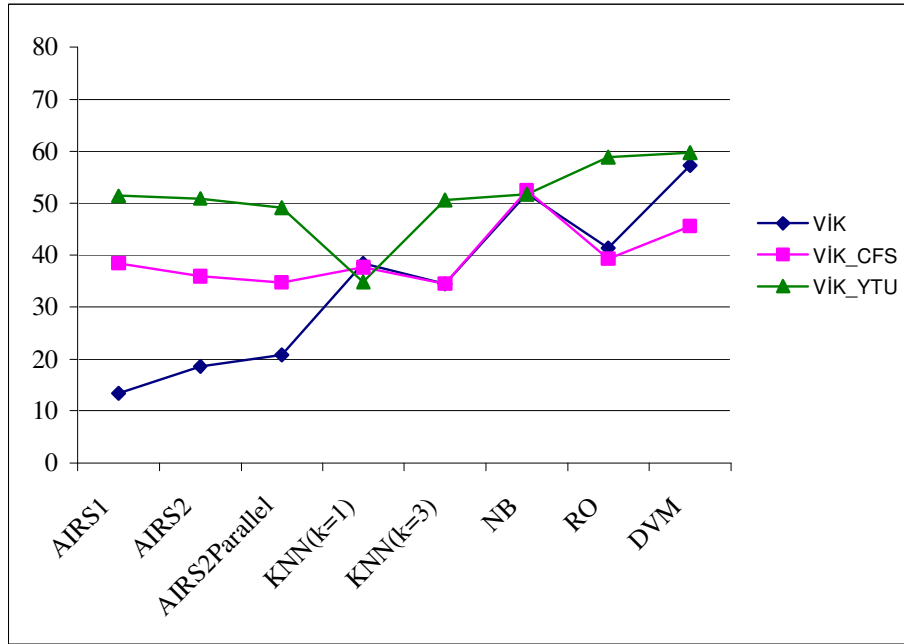
Çizelge 6.30: VK_{ya10} için Gövde Özellik Vektörleri İçin Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{GÖVDE_CFS}	34.29	35.71	38	52.57	48.29	72.86	53.14	75.71	53.75
V_{GÖVDE_YTU}	86.29	87.14	86.86	86.29	88	85.43	86	91.71	87.22

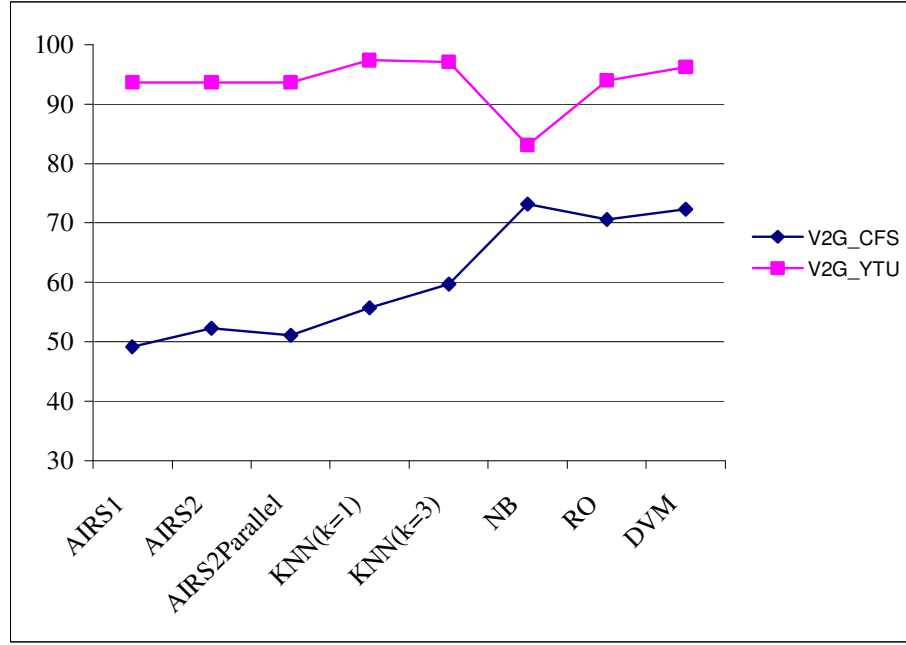
Şekil 6.17'ye baktığımızda, CfsSubsetEval kullanılarak özellik sayısı azaltılan V_{IST} vektörünün daha başarılı olduğunu, Şekil 6.18'de de işlevsel kelimelerde boyut azaltmada YTU yönteminin kullanılması KNN (k=1) haricinde başarıyı gözle görülür bir şekilde yükseltmiştir.



Şekil 6.17 VK_{ya10} için İstatistiksel Özellik Vektörleri Başarı Grafiği

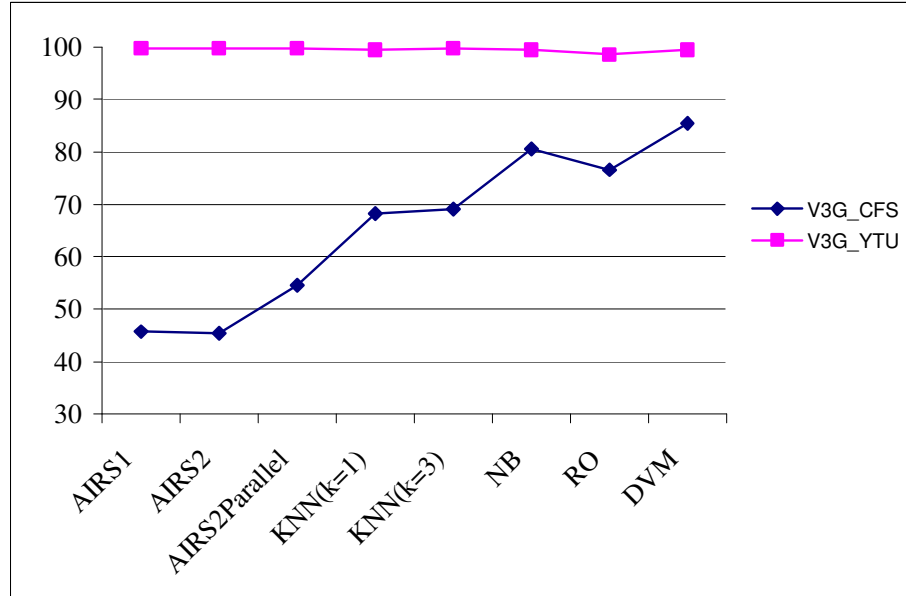


Şekil 6.18 VK_{ya10} için İşlevsel Kelimeler Özellik Vektörleri Başarı Grafiği

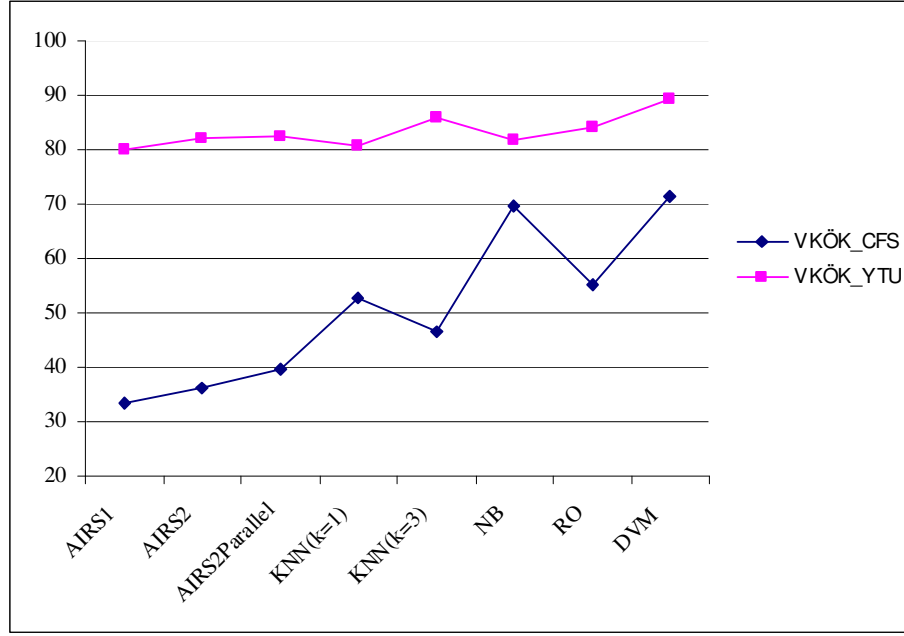


Şekil 6.19 VK_{ya10} için 2-gram Özellik Vektörleri Başarı Grafiği

Şekil 6.19 ve Şekil 6.20’de ise yine tür ve cinsiyet belirlemede olduğu gibi YTU ile indirgenmiş olan özellik vektörlerinin tüm sınıflandırıcılarda CfsSubsetEval’e göre daha başarılı olduğu görülmektedir.

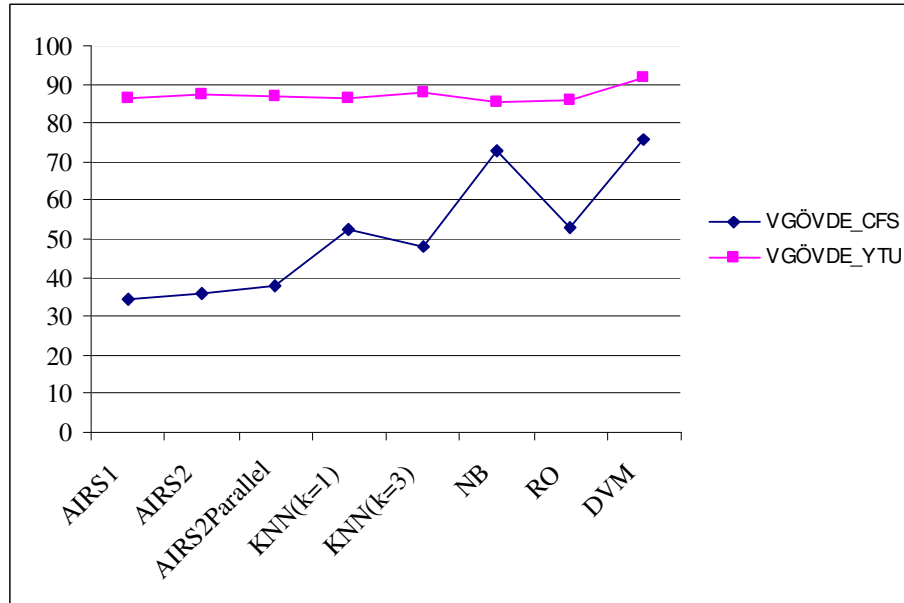


Şekil 6.20 VK_{ya10} için 3-gram Özellik Vektörleri Başarı Grafiği

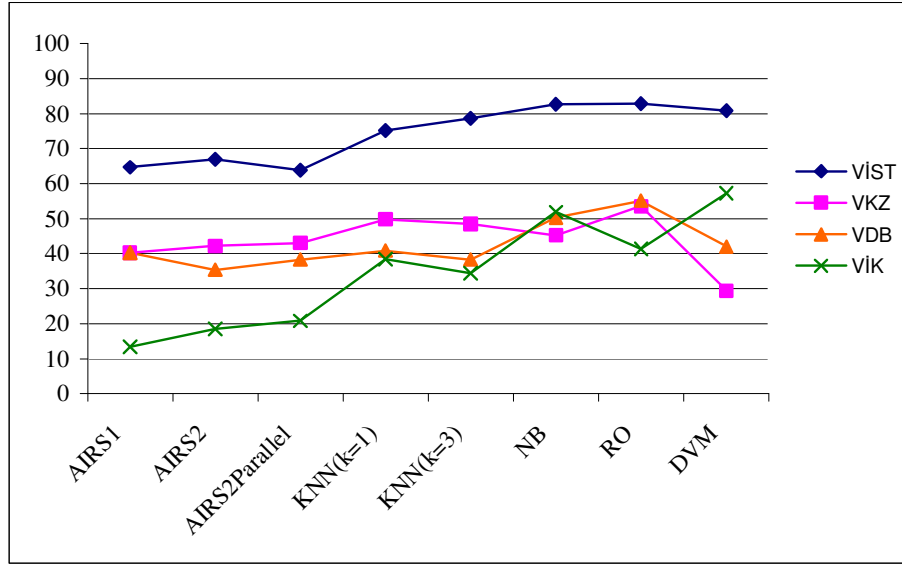


Şekil 6.21 VK_{ya10} için Kök Özellik Vektörleri Başarı Grafiği

Şekil 6.21’de köklerin ve Şekil 6.22’de gövdelerin kullanıldığı özellik vektörlerinde YTU’nun her sınıflandırıcıda yüksek bir performans sergilediği açıkça görülmektedir.

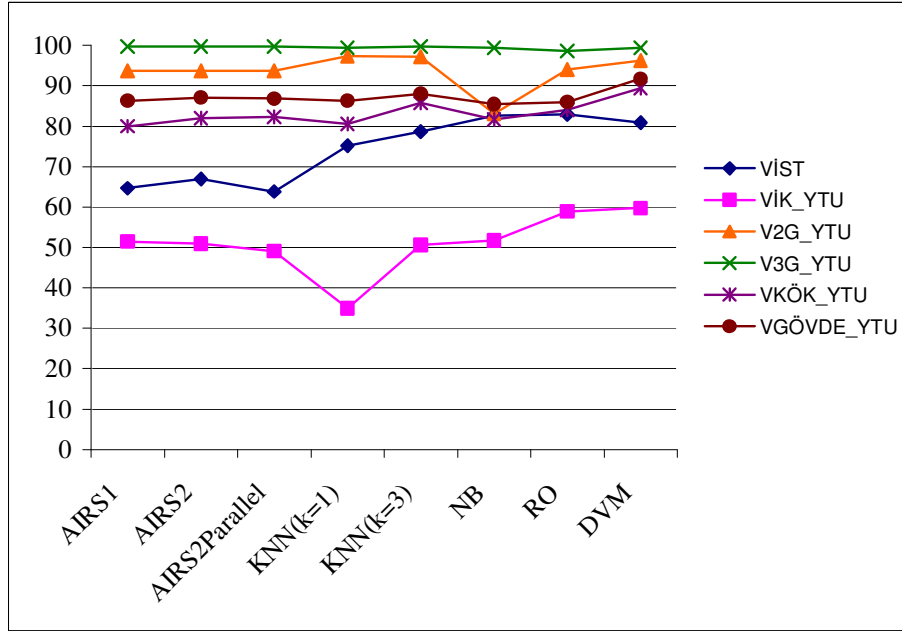


Şekil 6.22 VK_{ya10} için Gövde Özellik Vektörleri Başarı Grafiği



Şekil 6.23 VK_{ya10} için Boyut İndirgeme Yapılmayan Özellik Vektörleri Başarı Grafiği

Şekil 6.23'te, boyut indirgeme yapılmadan dört farklı özellik vektörünün başarı grafiği gösterilmektedir. V_{IST} özellik vektörü hem tür hem de cinsiyet belirlemede olduğu gibi aynı türde yazan yazarların tanınmasında da başarılı olmuştur. Şekil 6.24'te ise, kendi gruplarında en yüksek başarıyı veren özellik vektörleri gösterilmektedir. V_{3G_YTU} özellik vektörü bütün sınıflandırıcılarda yüksek başarılar elde etmiştir.



Şekil 6.24 VK_{ya10} için En Başarılı Özellik Vektörleri Başarı Grafiği

6.3.1.2 Aynı Türde Yazar Tanıma İçin Deneysel Sonuçlar ($V_{K_{ya8}}$)

Aynı türde yazılar yazan 8 yazardan oluşan yazar tanıma çalışmasında kullanılan sınıflandırıcıların, hangi özellik vektörleriyle yüzde kaçlık sınıflandırma başarısı gösterdikleri Çizelge 6.31’de gösterilmiştir. Buna göre en yüksek başarı %99.64 ile V_{3G_YTU} özellik vektörü kullanılarak AIRS2, AIRS2Parallel, NB ve RO sınıflandırıcıları ile alınmıştır. YBS sınıflandırıcılarından AIRS2 ve AIRS2Parallel bu veri kümesinde de en iyi performansı göstermesi, yazar tanıma çalışmalarında YBS sınıflandırıcılarının uygun özellik vektörleri kullanılarak tercih edilebileceğini göstermektedir.

Bu veri kümesi için tüm özellik vektörleri arasında ortalama en yüksek başarı %99.47 ile V_{3G_YTU} vektörüyle elde edilmiştir. Onu %95.09 ile V_{2G_YTU} özellik vektörü takip etmektedir. 10 yazardan oluşan veri kümesi ile karşılaştırıldığında 8 yazardan oluşan veri kümesinde daha yüksek başarı performansları sergilendiği görülmektedir. Örneğin, $V_{KÖK_YTU}$ özellik vektörünün ortalama başarısı %83.21’den %87.37’ye, $V_{GÖVDE_YTU}$ özellik vektörünün ortalama başarısı ise % 87.22’den %89.51’e yükselmiştir. Bu da sınıf sayısı azaldığında sınıflandırma performansının arttığını göstermektedir.

Sınıflandırma yöntemlerinin tüm özellik vektörleri ile ortalama başarılarına bakıldığında ise DVM sınıflandırıcısı %76.81 ile en iyi ortalamaı vermektedir. YBS sınıflandırıcıları içerisinde en iyi ortalamaı ise %63.28 ile AIRS2Parallel vermiştir.

Çizelge 6.31: VK_{ya8} için Genel Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V _{IST}	66.19	68.68	67.62	76.87	80.07	83.27	81.49	80.07	75.53
V _{IST_CFS}	73.67	73.67	77.58	78.29	81.14	83.99	80.43	77.94	78.34
V _{IST_YTU}	58.21	57.14	62.86	51.07	63.93	61.79	60	67.14	60.27
V _{KZ}	51.25	50.53	53.38	53.74	56.94	57.65	61.57	46.98	54.01
V _{DB}	48.21	45	50.36	47.86	45.36	58.21	54.29	51.43	50.09
V _{IK}	12.46	19.93	25.27	39.50	36.30	60.85	46.62	63.35	38.04
V _{IK_CFS}	43.77	48.40	40.93	45.55	45.20	60.85	50.53	59.43	49.33
V _{IK_YTU}	57.5	61.43	62.86	45.36	55.36	59.29	71.79	67.86	60.18
V _{2G_CFS}	58.21	61.79	63.21	74.29	74.64	81.07	71.79	82.86	70.98
V _{2G_YTU}	94.64	94.29	95.71	97.86	96.79	88.21	96.07	97.14	95.09
V _{3G_CFS}	50.36	49.29	54.29	76.43	74.64	84.64	73.21	86.79	68.71
V _{3G_YTU}	99.29	99.64	99.64	99.29	99.29	99.64	99.64	99.29	99.47
V _{KÖK_CFS}	41.99	39.86	40.57	55.87	54.45	80.07	59.79	82.92	56.94
V _{KÖK_YTU}	87.14	86.43	87.14	85.71	87.86	86.79	87.5	90.36	87.37
V _{GÖVDE_CFS}	38.93	43.21	40	57.86	53.57	75.36	66.43	82.86	57.28
V _{GÖVDE_YTU}	87.14	88.57	91.07	88.21	91.79	88.93	87.86	92.5	89.51
Ortalama	60.56	61.74	63.28	67.11	68.58	75.66	71.81	76.81	

Çizelge 6.32’de boyut indirgemesi yapılmayan dört özellik vektörünün V_{IST} (26 özellik), V_{KZ} (3 özellik), V_{DB} (8 özellik) ve V_{IK} ’nın (154 özellik) başarı sonuçları gösterilmektedir. En yüksek başarı tüm veri kümelerinde olduğu gibi yine istatistiksel özellikler kullanılarak elde edilen özellik vektörünün, Naive Bayes sınıflandırıcısından almış olduğu %83.27’lik başarıdır. Yapay Bağışıklık sınıflandırıcılarının performansı bu özellik vektöründe düşük kalmıştır.

Çizelge 6.33’te, Weka içerisinde yer alan CfsSubsetEval boyut indirgeme fonksiyonu V_{IST} , V_{IK} , V_{2G} , V_{3G} , $V_{KÖK}$ ve $V_{GÖVDE}$ özellik vektörlerine uygulandığında elde edilmiş olan yeni altı farklı özellik vektörünün aynı sınıflandırıcılar üzerindeki başarıları gösterilmektedir. Burada en yüksek başarı genelde diğer veri kümelerinde de olduğu üzere V_{3G_CFS} özellik vektörü kullanılarak, %86.79 ile DVM sınıflandırıcısından alınmıştır. Aynı şartlar altında şimdiye kadar analiz edilen tüm veri kümelerinde Destek Vektör Makinesi en yüksek başarıyı vermiştir. Yine burada da Yapay Bağışıklık sınıflandırıcıları diğer sınıflandırıcılara göre bu özellik vektörlerinde istenilen başarıyı elde edememiştir. Çizelge 6.34 ise, aynı özellik vektörlerinin YTU yöntemi kullanılarak boyut azaltılmasına gidildiğinde elde edilen altı yeni özellik vektörünü göstermektedir. En başarılı özellik vektörü V_{3G_YTU} olup, NB, RO ve AIRS sınıflandırıcıları %99.64 ile aynı başarıyı göstermişlerdir. Tür ve cinsiyet belirlemede de en iyi başarıyı veren V_{3G_YTU} özellik vektörü olmuştur.

Çizelge 6.35’te istatistiksel özelliklerden oluşan üç özellik vektörü gruplandırılarak verilmiştir. En iyi performans, V_{IST_CFS} vektörü ile %83.99 ile NB sınıflandırıcısından alınmıştır. Çizelge 6.36’da işlevsel kelimelerden elde edilmiş üç özellik vektörü verilmiş olup, en yüksek başarı tür ve cinsiyet tanımda olduğu gibi YTU kullanılarak elde edilmiş özellik vektörü ile yine RO sınıflandırıcısından %71.79 ile alınmıştır. Çizelge 6.37 ve Çizelge 6.38’de 2 ve 3-gram’lardan elde edilen özellik vektörleri gruplandırılarak verilmiştir. Her ikisinde de en iyi performans YTU kullanılarak alınmıştır. V_{2G_YTU} kullanıldığında KNN ($k=1$) sınıflandırıcısı %97.86’lık bir başarı verirken, AIRS sınıflandırıcılarının da başarıları yüksektir. V_{3G_YTU} kullanılarak NB, RO, AIRS2 ve AIRS2Paralel’den %99.64’lük bir başarı elde edilmiştir. Çizelge 6.39’da kelime köklerinden, Çizelge 6.40’da da kelimenin gövdesinden oluşmuş özellik vektörleri verilmektedir. Her ikisinde de en iyi performans, YTU ile boyutun indirgenmiş halinden elde edilen özellik vektörlerinden alınmıştır. $V_{KÖK_YTU}$ ve $V_{GÖVDE_YTU}$ özellik vektörlerinin her ikisinde en yüksek başarıyı DVM ile %90.36 ve %92.5 ile vermişken ikinci sırada sırasıyla %87.14 ve %91.07 ile AIRS2Paralel vermiştir.

Çizelge 6.32: VK_{ya8} için Boyut İndirgeme Yapılmayan Özellik Vektörleri Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V _{IST}	66.19	68.68	67.62	76.87	80.07	83.27	81.49	80.07	75.53
V _{KZ}	51.25	50.53	53.38	53.74	56.94	57.65	61.57	46.98	54.01
V _{DB}	48.21	45	50.36	47.86	45.36	58.21	54.29	51.43	50.09
V _{IK}	12.46	19.93	25.27	39.50	36.30	60.85	46.62	63.35	38.04
Ortalama	44.53	46.04	49.16	54.49	54.67	65.00	60.99	60.46	

Çizelge 6.33: VK_{ya8} için CFS ile Boyut İndirgenmiş Özellik Vektörleri Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V _{IST_CFS}	73.67	73.67	77.58	78.29	81.14	83.99	80.43	77.94	78.34
V _{IK_CFS}	43.77	48.40	40.93	45.55	45.20	60.85	50.53	59.43	49.33
V _{2G_CFS}	58.21	61.79	63.21	74.29	74.64	81.07	71.79	82.86	70.98
V _{3G_CFS}	50.36	49.29	54.29	76.43	74.64	84.64	73.21	86.79	68.71
V _{KÖK_CFS}	41.99	39.86	40.57	55.87	54.45	80.07	59.79	82.92	56.94
V _{GÖVDE_CFS}	38.93	43.21	40	57.86	53.57	75.36	66.43	82.86	57.28
Ortalama	51.16	52.70	52.76	64.72	63.94	77.66	67.03	78.80	

Çizelge 6.34: VK_{ya8} için YTU ile Boyut İndirgenmiş Özellik Vektörleri Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Paralel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{IST_YTU}	58.21	57.14	62.86	51.07	63.93	61.79	60	67.14	60.27
V_{IK_YTU}	57.5	61.43	62.86	45.36	55.36	59.29	71.79	67.86	60.18
V_{2G_YTU}	94.64	94.29	95.71	97.86	96.79	88.21	96.07	97.14	95.09
V_{3G_YTU}	99.29	99.64	99.64	99.29	99.29	99.64	99.64	99.29	99.47
V_{KÖK_YTU}	87.14	86.43	87.14	85.71	87.86	86.79	87.5	90.36	87.37
V_{GÖVDE_YTU}	87.14	88.57	91.07	88.21	91.79	88.93	87.86	92.5	89.51
Ortalama	80.65	81.25	83.21	77.92	82.50	80.78	83.81	85.72	

Çizelge 6.35: VK_{ya8} için İstatistiksel Özellik Vektörleri Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Paralel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{IST}	66.19	68.68	67.62	76.87	80.07	83.27	81.49	80.07	75.53
V_{IST_CFS}	73.67	73.67	77.58	78.29	81.14	83.99	80.43	77.94	78.34
V_{IST_YTU}	58.21	57.14	62.86	51.07	63.93	61.79	60	67.14	60.27

Çizelge 6.36: VK_{ya8} için İşlevsel Kelimeler Özellik Vektörleri Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Paralel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V _{IK}	12.46	19.93	25.27	39.50	36.30	60.85	46.62	63.35	38.04
V _{IK_CFS}	43.77	48.40	40.93	45.55	45.20	60.85	50.53	59.43	49.33
V _{IK_YTU}	57.5	61.43	62.86	45.36	55.36	59.29	71.79	67.86	60.18

Çizelge 6.37: VK_{ya8} için 2-Gram Özellik Vektörleri Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Paralel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V _{2G_CFS}	58.21	61.79	63.21	74.29	74.64	81.07	71.79	82.86	70.98
V _{2G_YTU}	94.64	94.29	95.71	97.86	96.79	88.21	96.07	97.14	95.09

Çizelge 6.38: VK_{ya8} için 3-Gram Özellik Vektörleri Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Paralel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V _{3G_CFS}	50.36	49.29	54.29	76.43	74.64	84.64	73.21	86.79	68.71
V _{3G_YTU}	99.29	99.64	99.64	99.29	99.29	99.64	99.64	99.29	99.47

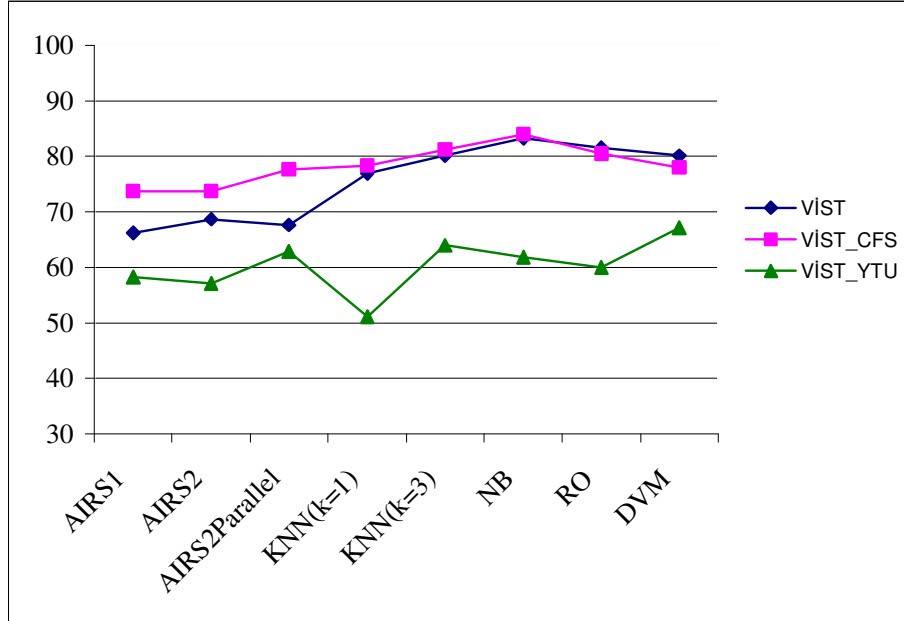
Çizelge 6.39: VK_{ya8} için Kök Özellik Vektörleri Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{KÖK_CFS}	41.99	39.86	40.57	55.87	54.45	80.07	59.79	82.92	56.94
V_{KÖK_YTU}	87.14	86.43	87.14	85.71	87.86	86.79	87.5	90.36	87.37

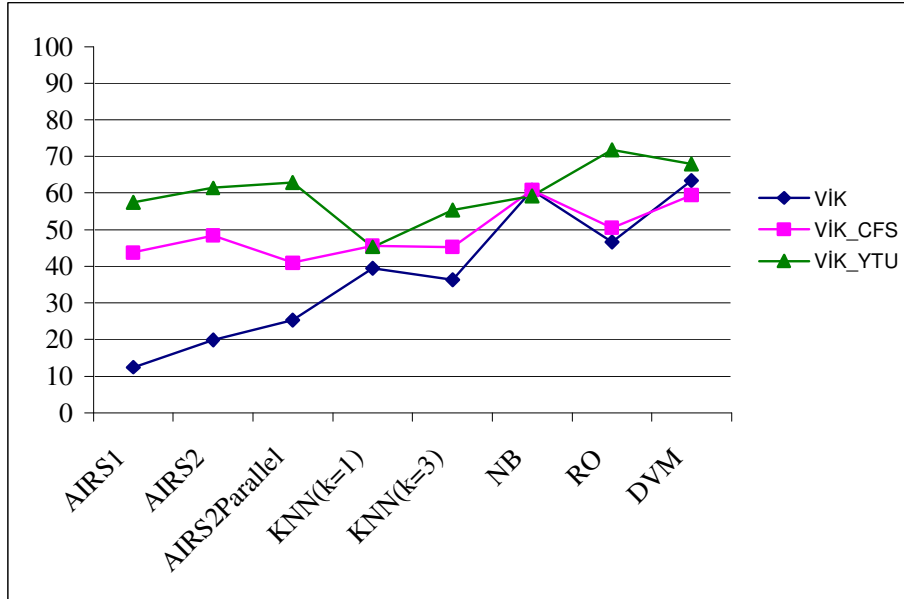
Çizelge 6.40: VK_{ya8} için Kök Özellik Vektörleri Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Parallel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{GÖVDE_CFS}	38.93	43.21	40	57.86	53.57	75.36	66.43	82.86	57.28
V_{GÖVDE_YTU}	87.14	88.57	91.07	88.21	91.79	88.93	87.86	92.5	89.51

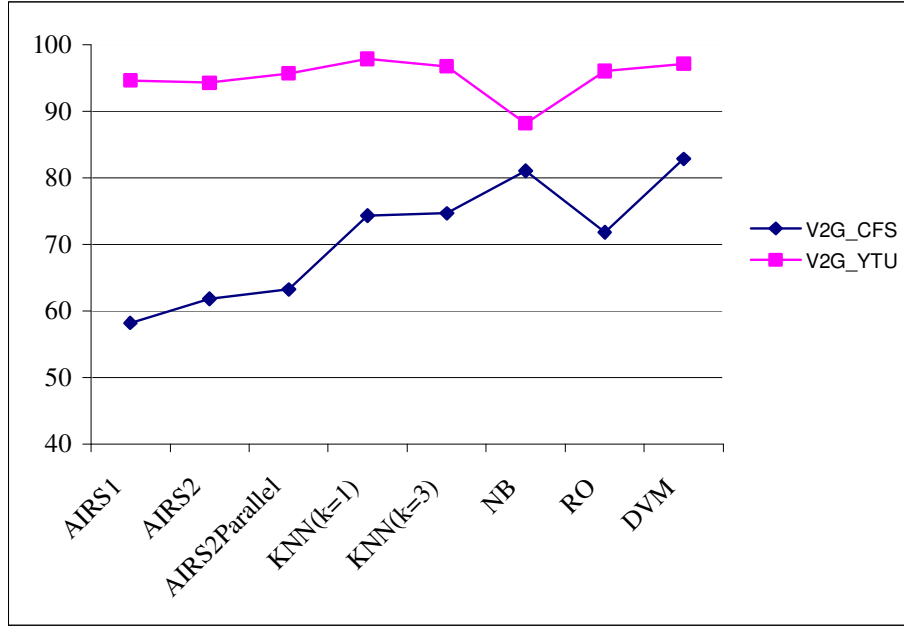
Şekil 6.25'e baktığımızda, CfsSubsetEval kullanılarak özellik sayısı azaltılan V_{IST} vektörünün daha başarılı olduğunu, Şekil 6.26'de de işlevsel kelimelerde boyut azaltmada YTU yönteminin kullanılması, NB haricindeki bütün sınıflandırıcılarda yüksek başarı vermiştir. NB sınıflandırıcısı bütün özellik vektörlerinde yaklaşık bir sonuç vermiştir.



Şekil 6.25 $V_{K_{ya8}}$ için İstatistiksel Özellik Vektörleri Başarı Grafiği

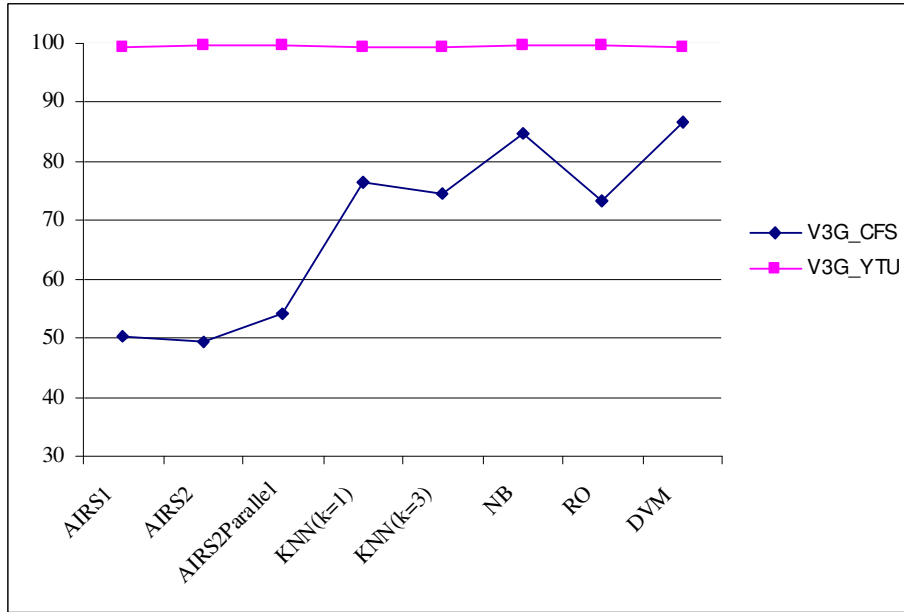


Şekil 6.26 $V_{K_{ya8}}$ için İşlevsel Kelimeler Özellik Vektörleri Başarı Grafiği

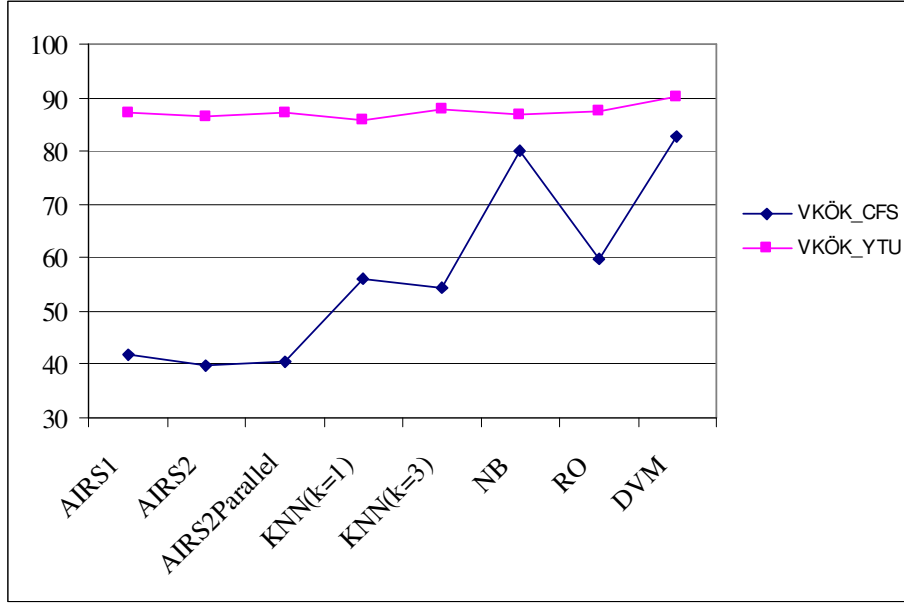


Şekil 6.27 VK_{ya8} için 2-gram Özellik Vektörleri Başarı Grafiği

Şekil 6.27 ve Şekil 6.28’de ise yine tür, cinsiyet ve aynı türde yazan 10 yazar veri kümesinde olduğu gibi YTU ile indirgenmiş olan özellik vektörlerinin tüm sınıflandırıcılarda CfsSubsetEval’e göre daha başarılı olduğu görülmektedir.

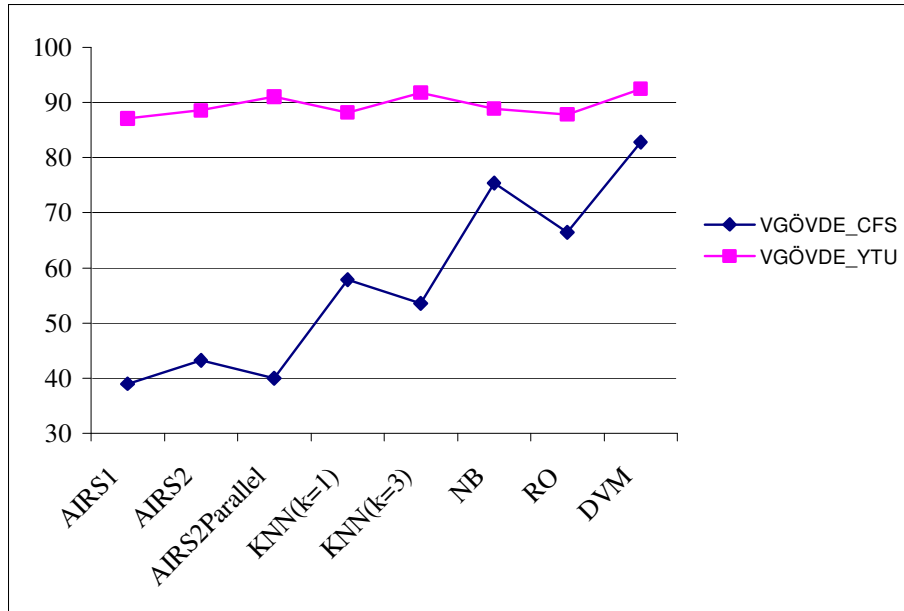


Şekil 6.28 VK_{ya8} için 3-gram Özellik Vektörleri Başarı Grafiği

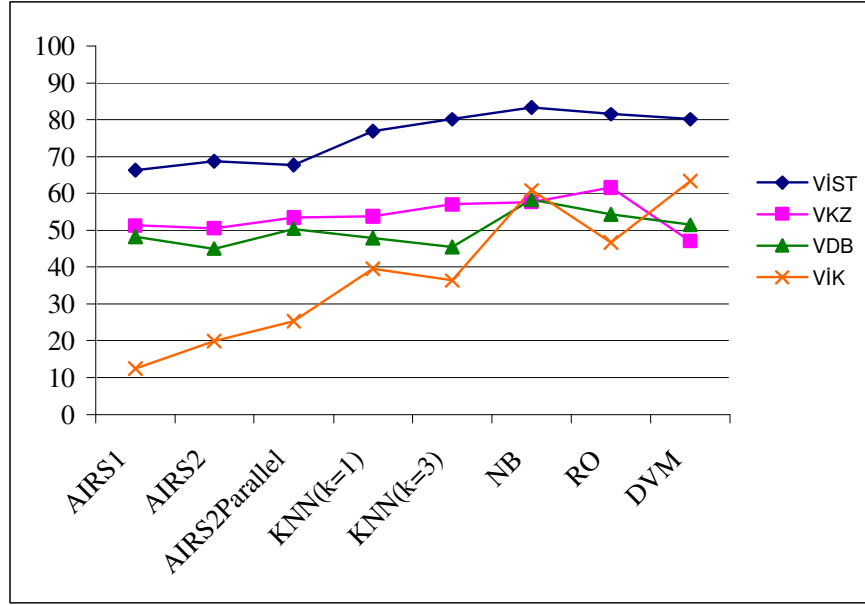


Şekil 6.29 VK_{ya8} için Kök Özellik Vektörleri Başarı Grafiği

Şekil 6.29’da köklerin ve Şekil 6.30’da gövdelerin kullanıldığı özellik vektörlerinde YTU’nun her sınıflandırıcıda yüksek bir performans sergilediği açıkça görülmektedir.

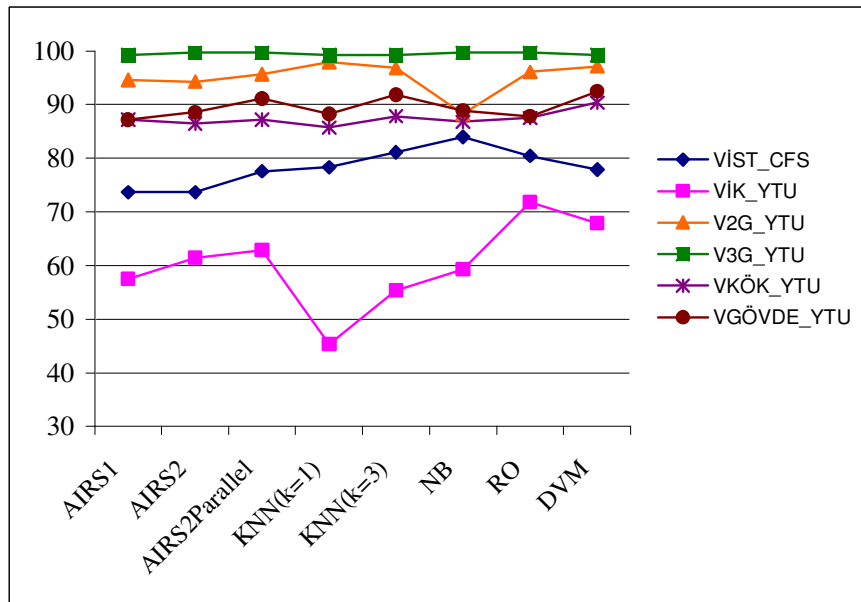


Şekil 6.30 VK_{ya8} için Gövde Özellik Vektörleri Başarı Grafiği



Şekil 6.31 VK_{ya8} için Boyut İndirgeme Uygulanmayan Özellik Vektörleri Başarı Grafiği

Şekil 6.31’de, boyut indirgeme yapılmadan dört farklı özellik vektörünün başarı grafiği gösterilmektedir. V_{IST} özellik vektörü şimdiye kadar ki bütün veri kümelerinde olduğu gibi aynı türde yazan yazarların tanınmasında da başarılı olmuştur. Şekil 6.32’de ise, kendi gruplarında en yüksek başarıyı veren özellik vektörleri gösterilmektedir. V_{3G_YTU} özellik vektörü bütün sınıflandırıcılarda yüksek başarılar elde etmiştir.



Şekil 6.32 VK_{ya8} için En Başarılı Özellik Vektörleri Başarı Grafiği

6.3.2 Farklı Türde Yazar Tanıma İçin Deneysel Sonuçlar

Farklı türlerde yazan yazar tanıma çalışması için deneysel sonuçlar 20, 18, 16, 14 ve 12 yazar içeren beş grup altında toplanmıştır.

6.3.2.1 Farklı Türde Yazar Tanıma İçin Deneysel Sonuçlar ($V_{K_{y20}}$)

Farklı türde yazılar yazan 20 yazardan oluşan yazar tanıma çalışmasında kullanılan sınıflandırıcıların, hangi özellik vektörleriyle yüzde kaçlık sınıflandırma başarısı gösterdikleri Çizelge 6.41’de gösterilmiştir. Buna göre en yüksek başarı %99.57 ile V_{3G_YTU} özellik vektörü kullanılarak, KNN ($k=1$) sınıflandırıcısı ile alınmıştır. YBS sınıflandırıcılarından AIRS2Parallel, V_{3G_YTU} özellik vektörüyle %98.43 başarı oranı vererek YBS sınıflandırıcıları içinde en iyi performansı vermiştir.

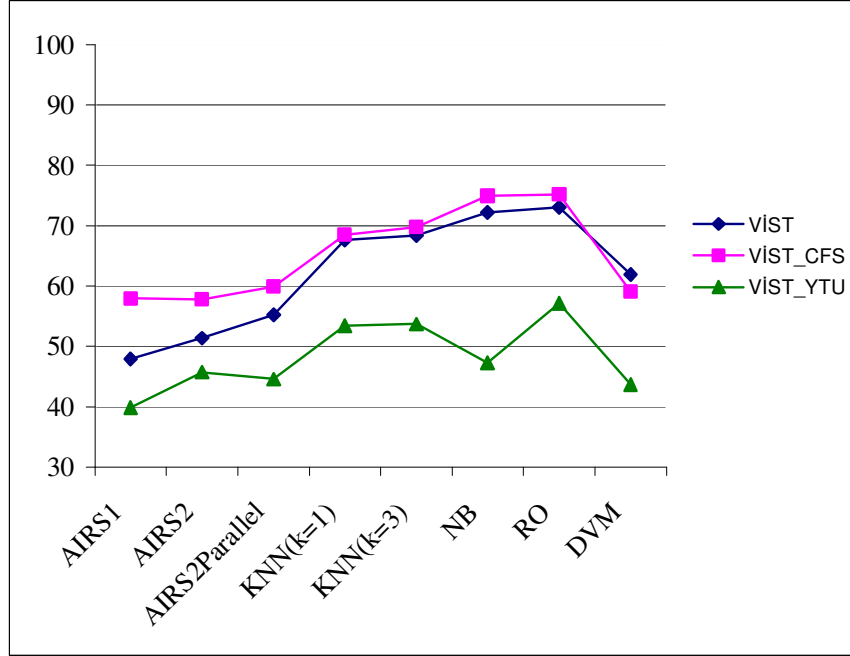
Tüm özellik vektörleri arasında ortalama en yüksek başarı %97.70 ile V_{3G_YTU} vektörüyle elde edilmiştir. Onu %94.45 ile $V_{KÖK_YTU}$ özellik vektörü takip etmektedir. $V_{GÖVDE_YTU}$ özellik vektörü ise % 73.57 ortalama sınıflandırma başarısı göstermiştir. En düşük başarı oranı ise %21.91 ile V_{IK} özellik vektörüyle alınmıştır. Bu da Türkçe için işlevsel kelimelerin tek başlarına kullanıldıklarında yazar tanıma çalışmaları için iyi bir ayırt edici olmadığını göstermektedir. Karakter n-gram’lar ise cinsiyet ve tür belirlemede olduğu gibi yazar tanıma çalışmasında da yüksek başarı oranları vermiştir. 2-gram özellik vektörü V_{2G_YTU} ile %86.07’lik başarı ortalaması alınmıştır. 3-gram’ların Türkçe metinlerde yazar tanıma çalışmalarında 2-gram’lara göre daha iyi performans verdikleri söylenebilir.

Sınıflandırma yöntemlerinin tüm özellik vektörleri ile ortalama başarılarına bakıldığında ise NB sınıflandırıcısı %63.61 ile en iyi ortalamayı vermektedir. YBS sınıflandırıcıları içerisinde en iyi ortalamayı ise %50.36 ile AIRS2Parallel vermiştir.

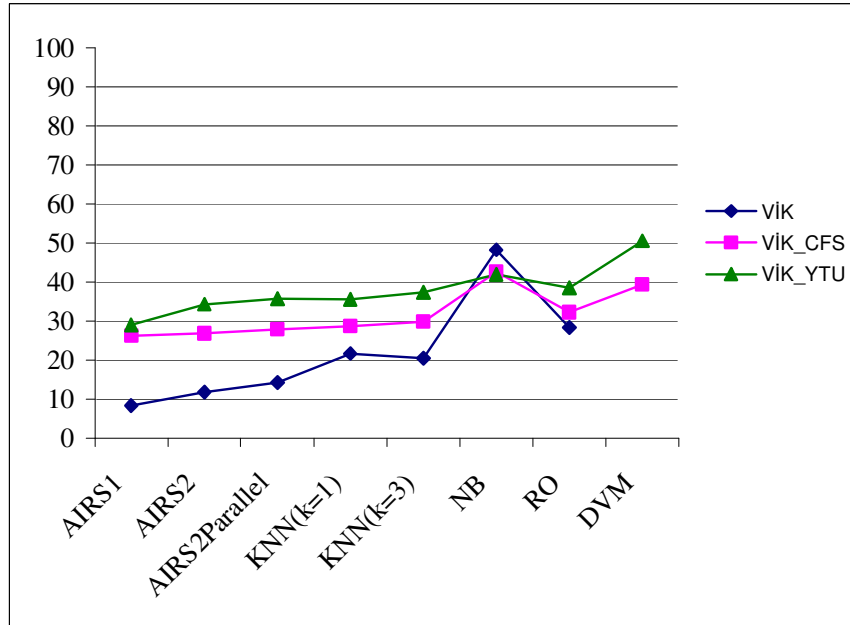
Çizelge 6.41: VK_{yk20} için Genel Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Paralel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V _{IST}	47.93	51.36	55.21	67.62	68.33	72.18	73.04	61.91	62.20
V _{IST_CFS}	57.92	57.77	59.91	68.47	69.76	74.89	75.18	59.06	65.37
V _{IST_YTU}	39.86	45.71	44.57	53.43	53.71	47.29	57.14	43.71	48.18
V _{KZ}	23.82	21.83	23.26	22.97	26.11	34.10	32.24	18.40	25.34
V _{DB}	28.57	31	34	35.29	38.86	45	38.57	21	34.04
V _{IK}	8.42	11.84	14.27	21.68	20.54	48.22	28.39	NA	21.91
V _{IK_CFS}	26.25	26.82	27.96	28.67	29.81	42.65	32.24	39.37	31.72
V _{IK_YTU}	29	34.29	35.71	35.57	37.29	42	38.43	50.57	37.86
V _{2G_CFS}	42	46.14	46	53.57	58	68.71	61	65.29	55.09
V _{2G_YTU}	80.71	84.14	85.43	93.29	92.29	78.71	84.43	89.57	86.07
V _{3G_CFS}	29.14	38.29	43.71	66.71	66.86	78.71	68.29	85.14	59.61
V _{3G_YTU}	97.86	97.14	98.43	99.57	99.43	93.71	96.14	99.29	97.70
V _{KÖK_CFS}	29.10	33.24	35.24	51.78	50.21	64.76	50.21	67.76	47.79
V _{KÖK_YTU}	92.29	91	93.29	98.43	98.43	90.43	92.57	99.14	94.45
V _{GÖVDE_CFS}	25.71	32.57	38.71	53.71	53.14	68.86	51.29	70.14	49.27
V _{GÖVDE_YTU}	67.14	67.57	70	80.86	80.14	67.57	72.14	83.14	73.57
Ortalama	45.36	48.17	50.36	58.23	58.93	63.61	59.46	63.57	

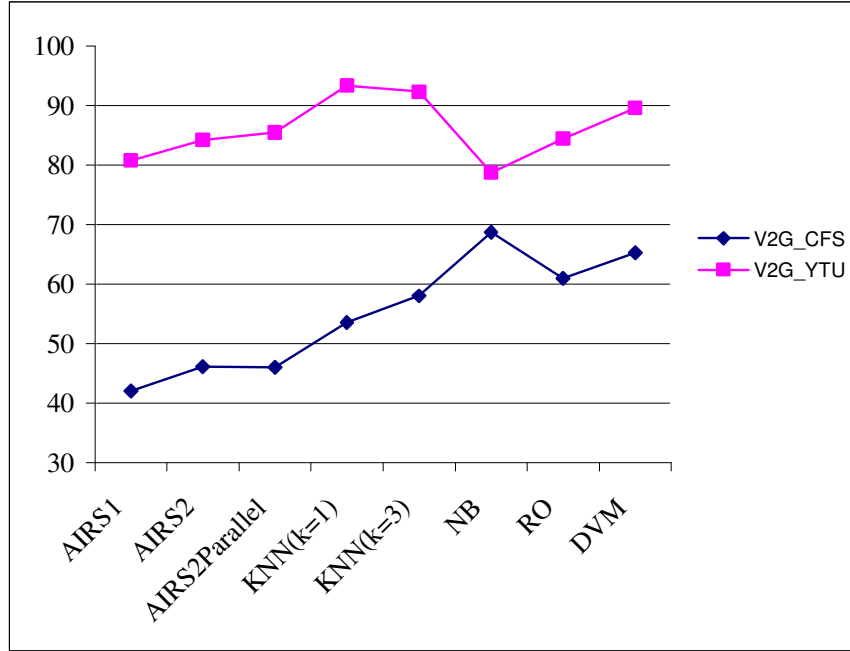
Şekil 6.33'e baktığımızda, CfsSubsetEval kullanılarak özellik sayısı azaltılan V_{IST} vektörünün daha başarılı olduğunu, Şekil 6.34'de de işlevsel kelimelerde boyut azaltmada YTU yönteminin kullanılması bütün sınıflandırıcılarda yüksek başarı vermiştir. NB sınıflandırıcısı, V_{IK} özellik vektöründe daha yüksek bir sonuç vermiştir.



Şekil 6.33 $V_{K_{yk20}}$ için İstatistiksel Özellik Vektörleri Başarı Grafiği

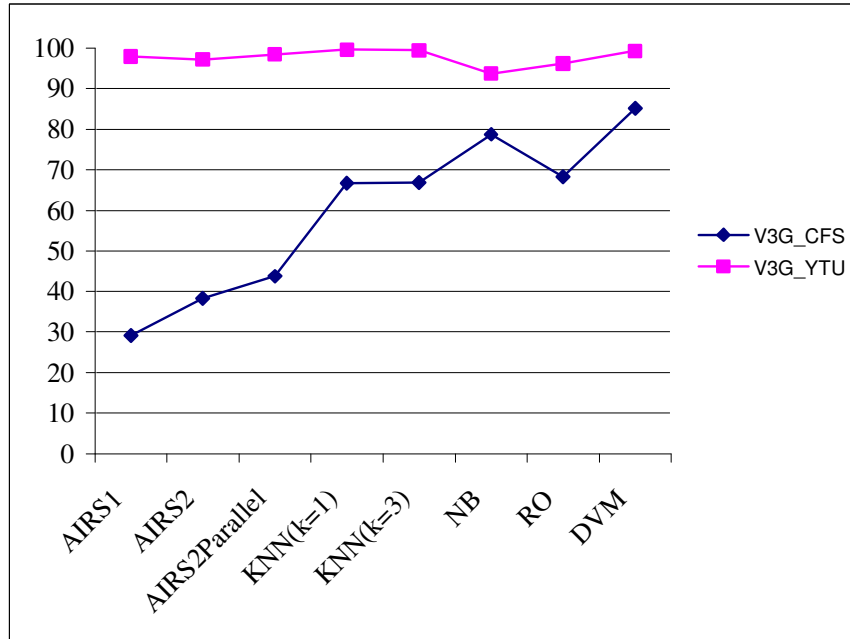


Şekil 6.34 $V_{K_{yk20}}$ için İşlevsel Kelimeler Özellik Vektörleri Başarı Grafiği

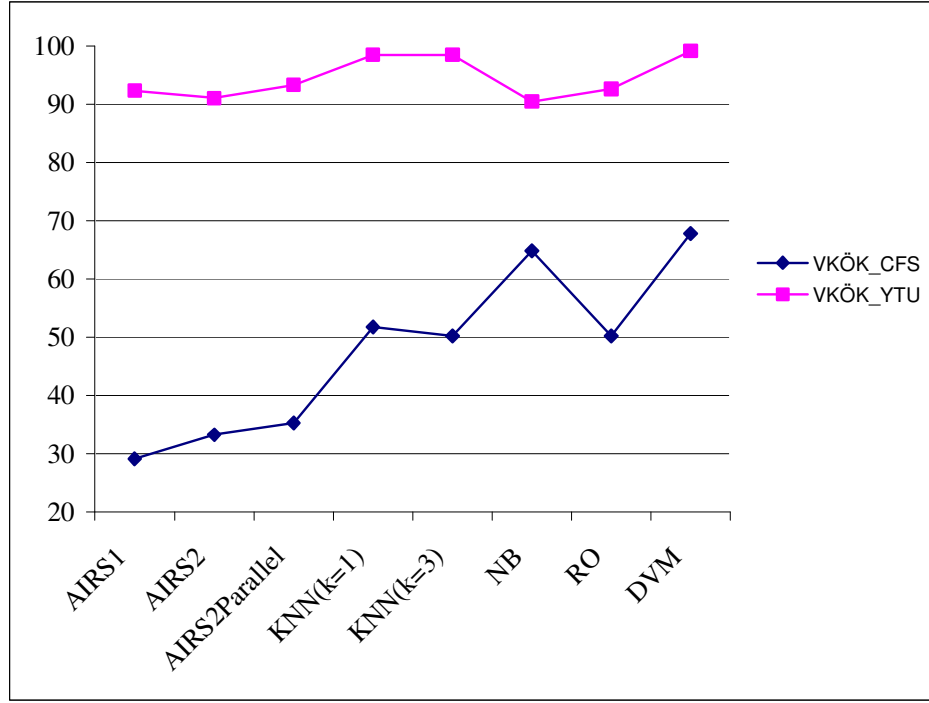


Şekil 6.35 VK_{yk20} için 2-gram Özellik Vektörleri Başarı Grafiği

Şekil 6.35 ve Şekil 6.36’da ise yine tür, cinsiyet ve aynı türde yazan 10 ve 8 yazar veri kümesinde olduğu gibi YTU ile indirgenmiş olan özellik vektörlerinin tüm sınıflandırıcılarda CfsSubsetEval’e göre daha başarılı olduğu görülmektedir.

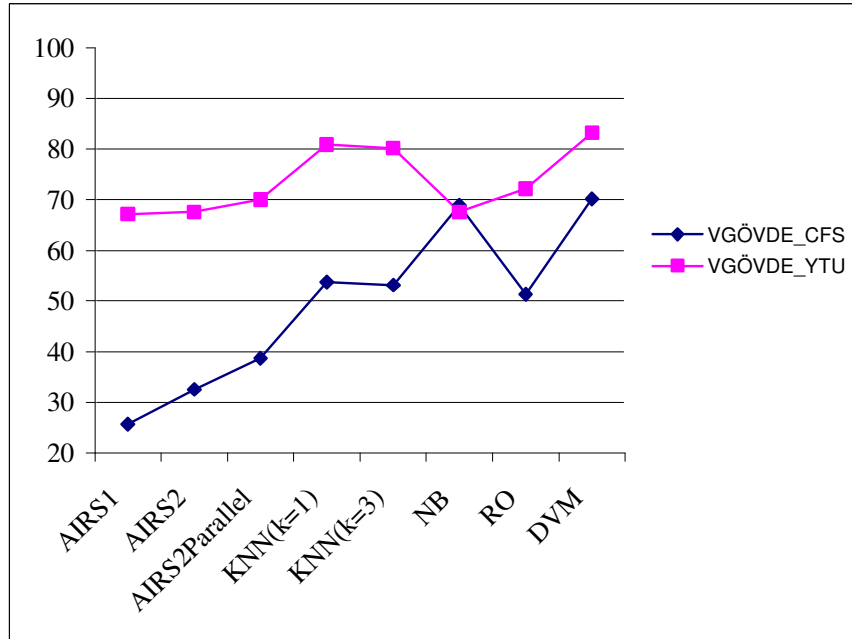


Şekil 6.36 VK_{yk20} için 3-gram Özellik Vektörleri Başarı Grafiği

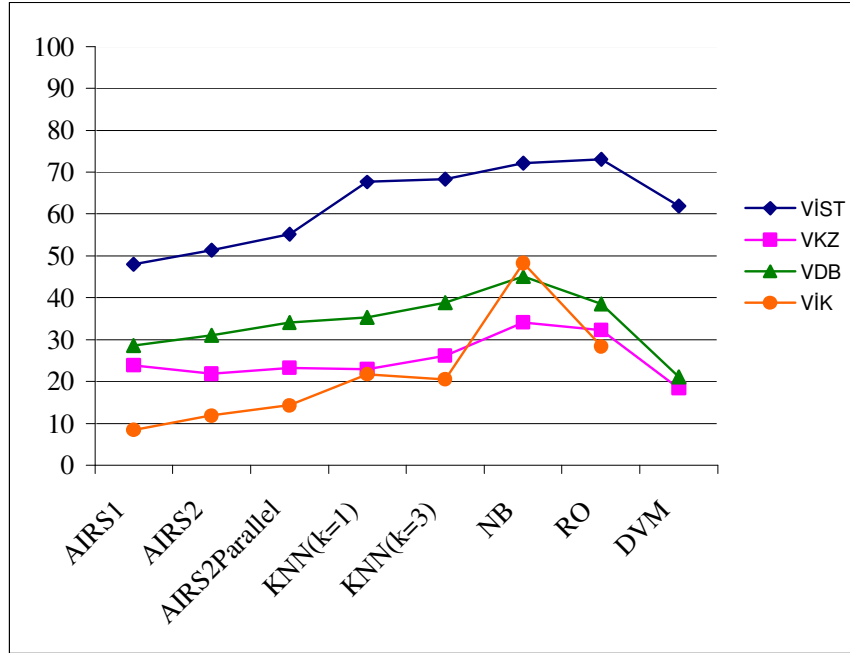


Şekil 6.37 VK_{yk20} için Kök Özellik Vektörleri Başarı Grafiği

Şekil 6.37’de köklerin ve Şekil 6.38’de gövdelerin kullanıldığı özellik vektörlerinde YTU’nun her sınıflandırıcıda yüksek bir performans sergilediği açıkça görülmektedir. NB yönteminde V_{GÖVDE_CFS} ve V_{GÖVDE_YTU} aynı başarıyı vermiştir.

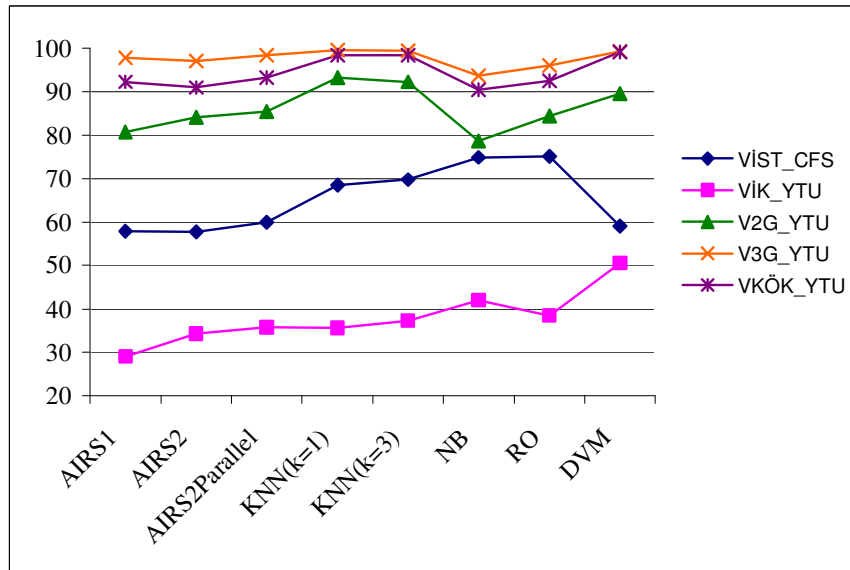


Şekil 6.38 VK_{yk20} için Gövde Özellik Vektörleri Başarı Grafiği



Şekil 6.39 VK_{yk20} için Boyut İndirgeme Yapılmayan Özellik Vektörleri Başarı Grafiği

Şekil 6.39'da, boyut indirgeme yapılmadan dört farklı özellik vektörünün başarı grafiği gösterilmektedir. V_{IST} özellik vektörü şimdiye kadar ki bütün veri kümelerinde olduğu gibi karma türde yazan yazarların tanınmasında da başarılı olmuştur. Şekil 6.40'da ise, kendi gruplarında en yüksek başarıyı veren özellik vektörleri gösterilmektedir. V_{3G_YTU} özellik vektörü bütün sınıflandırıcılarda yüksek başarılar elde vermiştir.



Şekil 6.40 VK_{yk20} için En Başarılı Özellik Vektörleri Başarı Grafiği

6.3.2.2 Farklı Türde Yazar Tanıma İçin Deneysel Sonuçlar ($V_{K_{yk18}}$)

Farklı türde yazılar yazan 18 yazardan oluşan yazar tanıma çalışmasında kullanılan sınıflandırıcıların, hangi özellik vektörleriyle yüzde kaçlık sınıflandırma başarısı gösterdikleri Çizelge 6.42’de gösterilmiştir. Buna göre en yüksek başarı %99.68 ile V_{3G_YTU} özellik vektörü kullanılarak KNN ($k=1$) sınıflandırıcısı ile alınmıştır. YBS sınıflandırıcılarından AIRS2Parallel, V_{3G_YTU} özellik vektörüyle %98.89 başarı oranı vererek, YBS sınıflandırıcıları içinde en iyi performansı vermiştir.

Tüm özellik vektörleri arasında ortalama en yüksek başarı %98.25 ile V_{3G_YTU} vektörüyle elde edilmiştir. Onu %96.57 ile $V_{GÖVDE_YTU}$ özellik vektörü takip etmektedir. $V_{KÖK_YTU}$ özellik vektörü ise % 96.11 ortalama sınıflandırma başarısı göstermiştir. En düşük başarı oranı ise %27.18 ile V_{KZ} özellik vektörüyle alınmıştır. Bu da Türkçe için kelime zenginliğine dayalı özelliklerin tek başlarına kullanıldıklarında yazar tanıma çalışmaları için iyi bir ayırt edici olmadığını göstermektedir. Karakter n-gram’lar ise cinsiyet ve tür belirlemede olduğu gibi yazar tanıma çalışmasında da yüksek başarı oranları vermiştir. 2-gram özellik vektörü V_{2G_YTU} %89.28’lik başarı ortalaması sağlamıştır. 3-gram’ların Türkçe metinlerde yazar tanıma çalışmalarında 2-gram’lara göre daha iyi performans verdikleri söylenebilir.

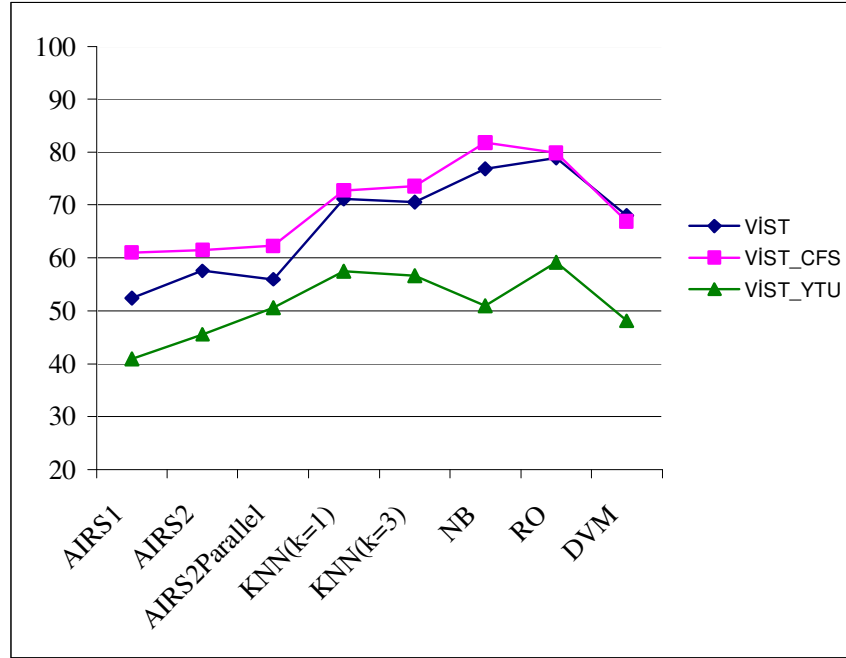
Deneylerde yazar sayısı azaldıkça sınıflandırma performansının tüm yöntemlerde arttığını söyleyebiliriz. Örneğin V_{IST} özellik vektörü 20 yazar için ortalama %66.2 başarı vermişken, 18 yazar için bu oran %66.42’ye çıkmıştır.

Sınıflandırma yöntemlerinin tüm özellik vektörleri ile ortalama başarılarına bakıldığında ise NB sınıflandırıcısı %67.86 ile en iyi ortalamaı vermektedir. YBS sınıflandırıcıları içerisinde en iyi ortalamaı ise %54.01 ile AIRS2Parallel vermiştir.

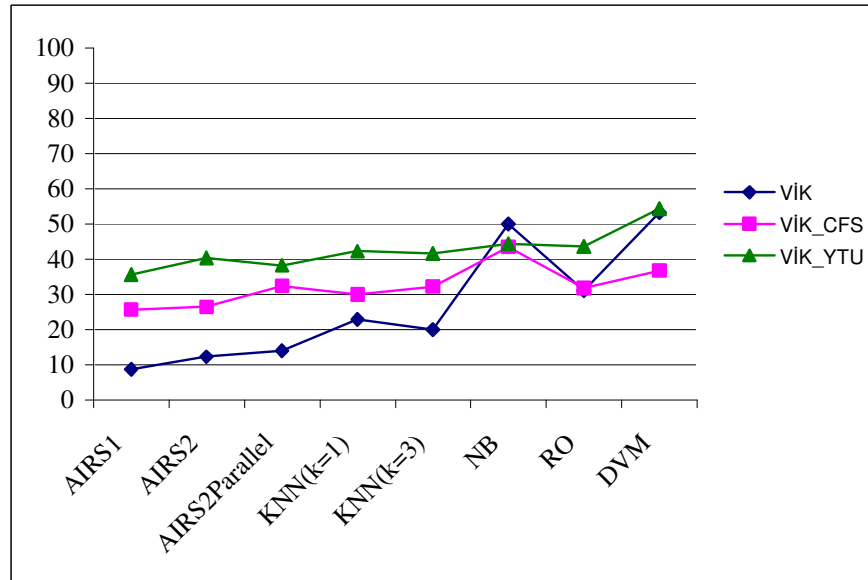
Çizelge 6.42: VK_{yk18} için Genel Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Paralel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V _{IST}	52.46	57.53	55.94	71.16	70.52	76.86	78.92	67.99	66.42
V _{IST_CFS}	61.01	61.49	62.28	72.74	73.53	81.78	79.87	66.88	69.95
V _{IST_YTU}	40.95	45.56	50.63	57.46	56.67	50.95	59.21	48.10	51.19
V _{KZ}	25.20	25.36	24.72	25.99	27.89	34.55	32.01	21.71	27.18
V _{DB}	32.70	32.38	35.71	37.94	39.52	45.56	40.48	25.08	36.17
V _{IK}	8.72	12.20	13.95	22.82	19.97	49.92	31.06	53.25	26.49
V _{IK_CFS}	25.67	26.47	32.33	29.95	32.17	43.42	31.70	36.77	32.31
V _{IK_YTU}	35.71	40.32	38.25	42.22	41.59	44.29	43.65	54.44	42.56
V _{2G_CFS}	46.51	50.16	48.41	63.02	64.13	73.17	66.19	71.75	60.42
V _{2G_YTU}	86.98	88.25	88.57	95.08	93.49	82.06	88.57	91.27	89.28
V _{3G_CFS}	31.27	38.25	48.57	73.65	73.65	82.06	68.89	88.41	63.09
V _{3G_YTU}	98.57	98.41	98.89	99.68	99.52	96.51	95.08	99.37	98.25
V _{KÖK_CFS}	28.21	33.12	37.56	55.47	54.99	69.57	51.19	74.17	50.54
V _{KÖK_YTU}	96.19	94.76	96.03	98.89	98.57	93.02	92.38	99.05	96.11
V _{GÖVDE_CFS}	26.83	34.13	34.13	53.02	53.65	69.37	53.65	72.70	49.69
V _{GÖVDE_YTU}	95.87	95.56	98.25	98.89	98.57	92.70	93.49	99.21	96.57
Ortalama	49.55	52.12	54.01	62.37	62.40	67.86	62.90	66.88	

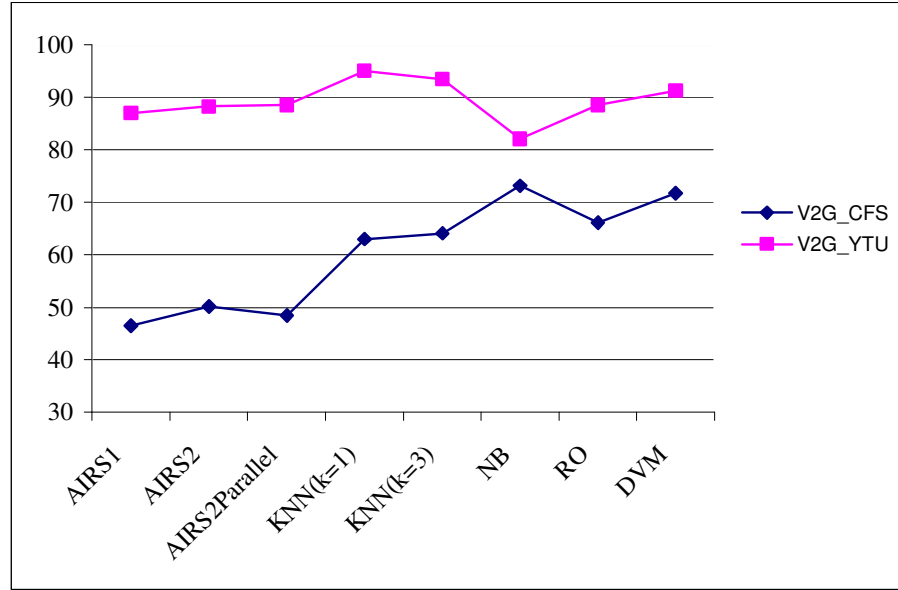
Şekil 6.41'e baktığımızda, CfsSubsetEval kullanılarak özellik sayısı azaltılan V_{IST} vektörünün daha başarılı olduğunu, Şekil 6.42'de de işlevsel kelimelerde boyut azaltmada YTU yönteminin kullanılması NB sınıflandırıcısı hariç diğer sınıflandırıcılarda yüksek başarı vermiştir. NB sınıflandırıcısı V_{IK} özellik vektöründe daha yüksek bir sonuç vermiştir.



Şekil 6.41 $V_{K_{yk18}}$ için İstatistiksel Özellik Vektörleri Başarı Grafiği

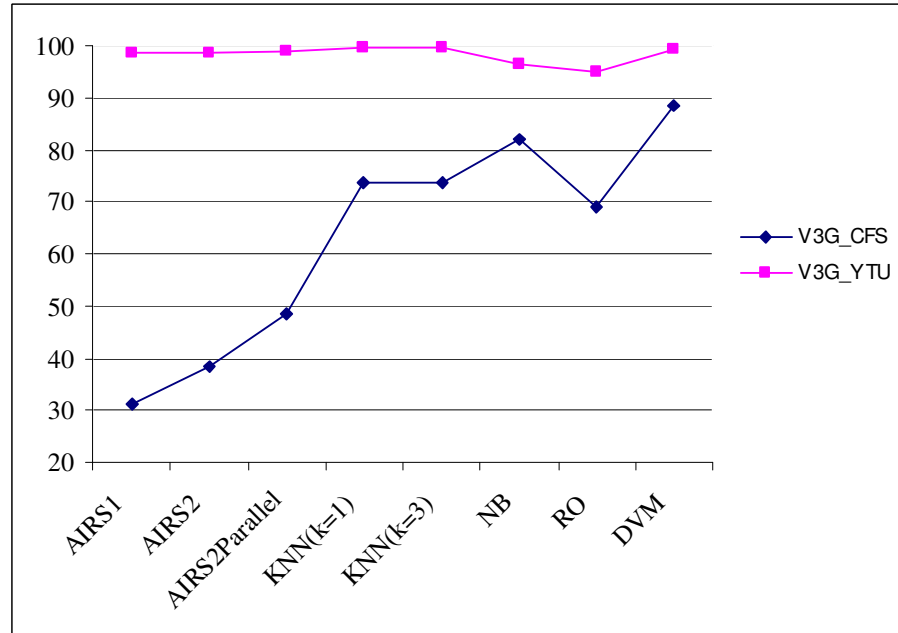


Şekil 6.42 $V_{K_{yk18}}$ için İşlevsel Kelimeler Özellik Vektörleri Başarı Grafiği

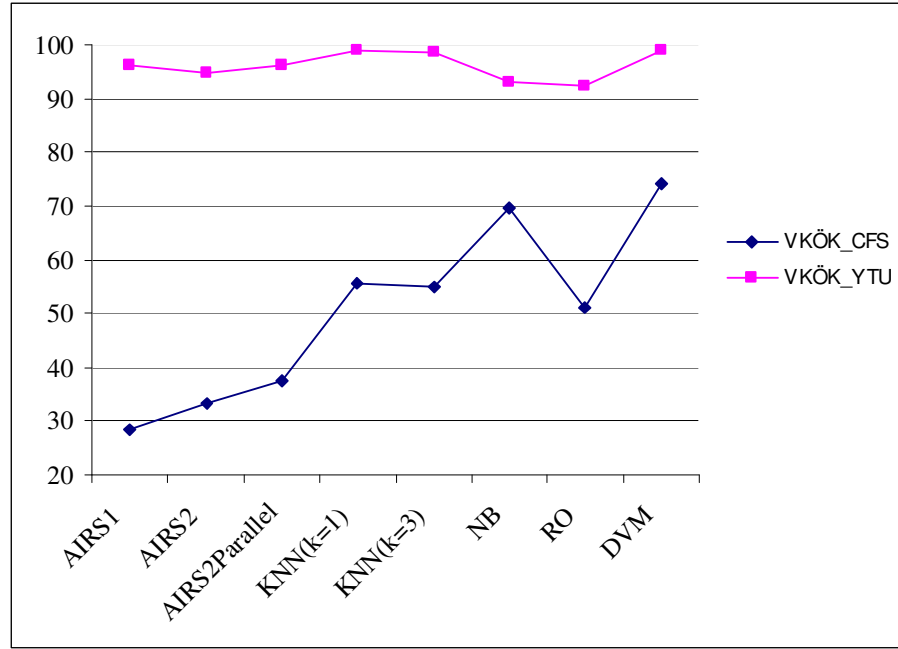


Şekil 6.43 VK_{yk18} için 2-gram Özellik Vektörleri Başarı Grafiği

Şekil 6.43 ve Şekil 6.44’de ise yine tür, cinsiyet, aynı türde yazan 10 ve 8 yazar, karışık türde yazan 20 yazar veri kümesinde olduğu gibi YTU ile indirgenmiş olan özellik vektörlerinin tüm sınıflandırıcılarda CfsSubsetEval’e göre daha başarılı olduğu görülmektedir.

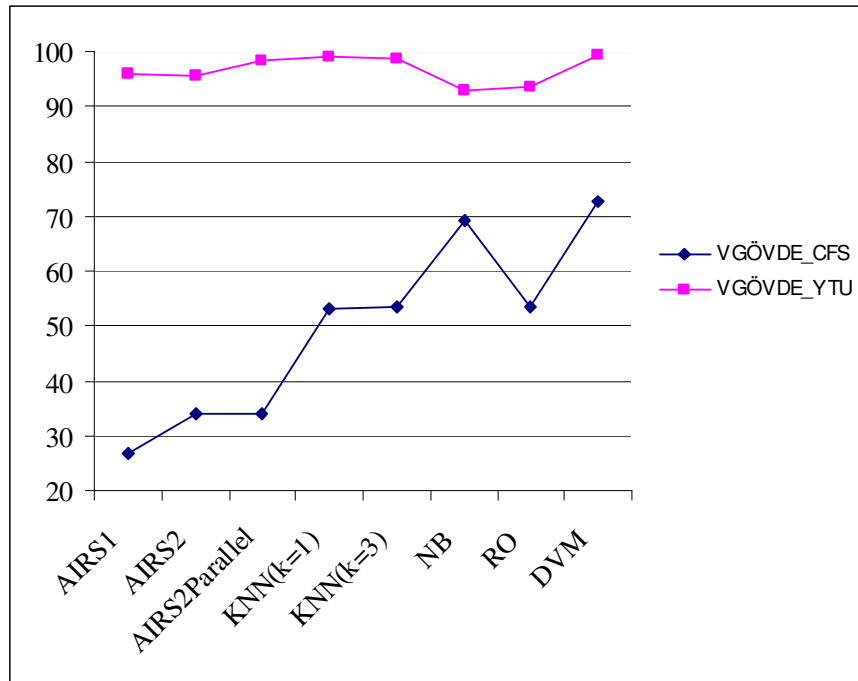


Şekil 6.44 VK_{yk18} için 3-gram Özellik Vektörleri Başarı Grafiği

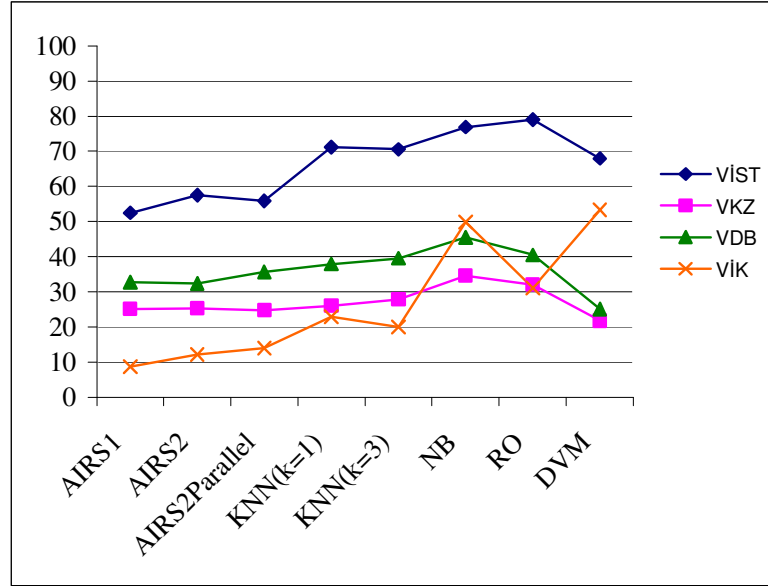


Şekil 6.45 VK_{yk18} için Kök Özellik Vektörleri Başarı Grafiği

Şekil 6.45’de köklerin ve Şekil 6.46’da gövdelerin kullanıldığı özellik vektörlerinde YTU’nun her sınıflandırıcıda yüksek bir performans sergilediği açıkça görülmektedir.

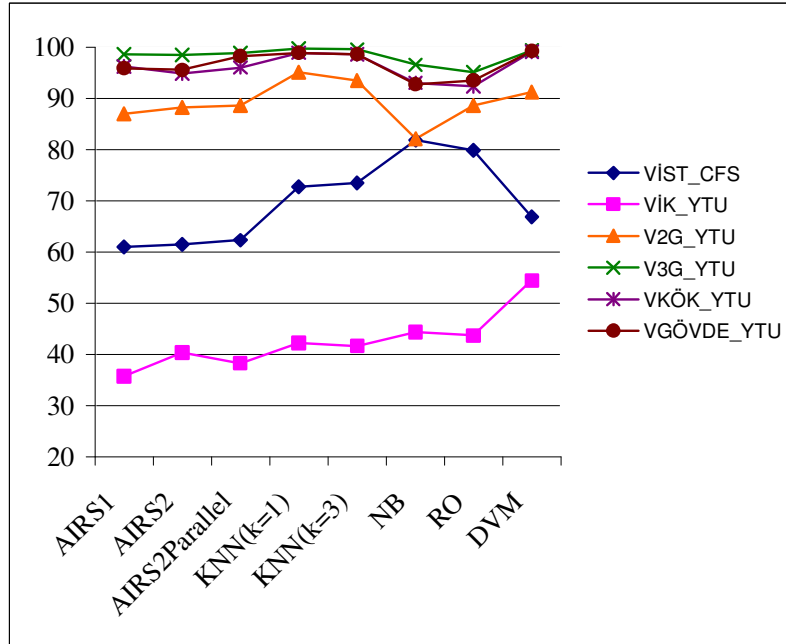


Şekil 6.46 VK_{yk18} için Gövde Özellik Vektörleri Başarı Grafiği



Şekil 6.47 VK_{yk18} için Boyut İndirgeme Yapılmayan Özellik Vektörleri Başarı Grafiği

Şekil 6.47’de, boyut indirgeme yapılmadan dört farklı özellik vektörünün başarı grafiği gösterilmektedir. V_{IST} özellik vektörü şimdiye kadar ki bütün veri kümelerinde olduğu gibi karma türde yazan yazarların tanınmasında da başarılı olmuştur. Şekil 6.48’de ise, kendi gruplarında en yüksek başarıyı veren özellik vektörleri gösterilmektedir. V_{3G_YTU} özellik vektörü bütün sınıflandırıcılarda yüksek başarılar elde vermiştir.



Şekil 6.48 VK_{yk18} için En Başarılı Özellik Vektörleri Başarı Grafiği

6.3.2.3 Farklı Türde Yazar Tanıma İçin Deneysel Sonuçlar ($V_{K_{yk16}}$)

Farklı türde yazılar yazan 16 yazardan oluşan yazar tanıma çalışmasında kullanılan sınıflandırıcıların, hangi özellik vektörleriyle yüzde kaçlık sınıflandırma başarısı gösterdikleri Çizelge 6.43’de gösterilmiştir. Buna göre en yüksek başarı %99.82 ile V_{3G_YTU} özellik vektörü kullanılarak KNN ($k=1$) sınıflandırıcısı ile alınmıştır. YBS sınıflandırıcılarından AIRS2Parallel, V_{3G_YTU} özellik vektörüyle %99.29 başarı oranı ile YBS sınıflandırıcıları içinde en iyi performansı vermiştir.

Tüm özellik vektörleri arasında ortalama en yüksek başarı %98.59 ile V_{3G_YTU} vektörüyle elde edilmiştir. Onu %97.28 ile $V_{KÖK_YTU}$ özellik vektörü takip etmektedir. $V_{GÖVDE_YTU}$ özellik vektörü ise % 81.01 ortalama sınıflandırma başarısı göstermiştir. En düşük başarı oranı ise %28.93 ile V_{IK} özellik vektörüyle alınmıştır. Bu da Türkçe için işlevsel kelimelerin tek başlarına kullanıldıklarında yazar tanıma çalışmaları için iyi bir ayırt edici olmadığını göstermektedir. Karakter n-gram’lar ise cinsiyet ve tür belirlemede olduğu gibi yazar tanıma çalışmasında da yüksek başarı oranları sağlamıştır. 2-gram özellik vektörü V_{2G_YTU} %89.15’lik başarı ortalaması sağlamıştır. 3-gram’ların Türkçe metinlerde yazar tanıma çalışmalarında 2-gram’lara göre daha iyi performans verdikleri söylenebilir.

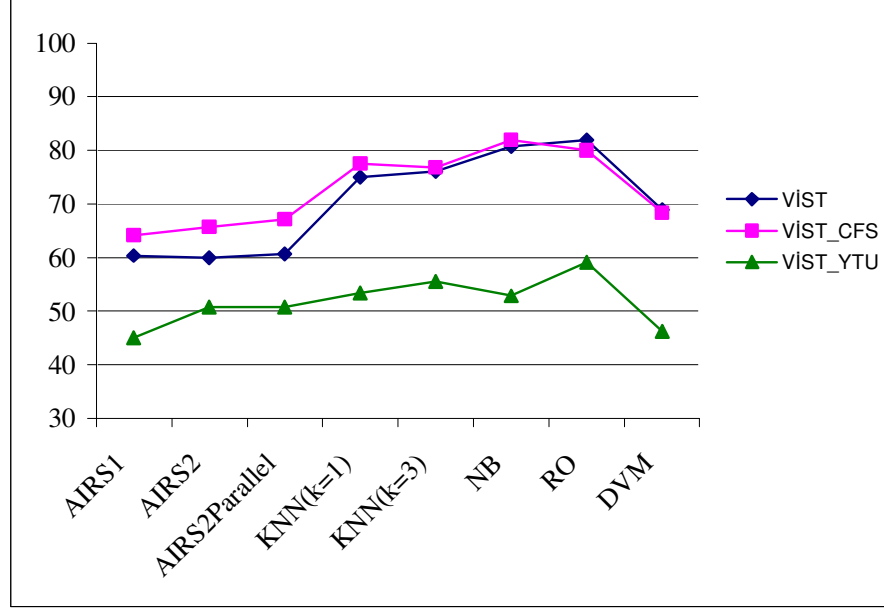
Deneysel yazar sayısı azaldıkça sınıflandırma performansının tüm yöntemlerde arttığını söyleyebiliriz. Örneğin, V_{IST} özellik vektörü 20 yazar için ortalama %66.2 başarı vermişken, 18 yazar için bu oran %66.42’ye çıkmıştır. 16 yazar içinse bu oran %70.47’ye yükselmiştir.

Sınıflandırma yöntemlerinin tüm özellik vektörleri ile ortalama başarılarına bakıldığında ise NB sınıflandırıcısı %69.50 ile en iyi ortalamaı vermektedir. YBS sınıflandırıcıları içerisinde en iyi ortalamaı ise %55.07 ile AIRS2Parallel vermiştir.

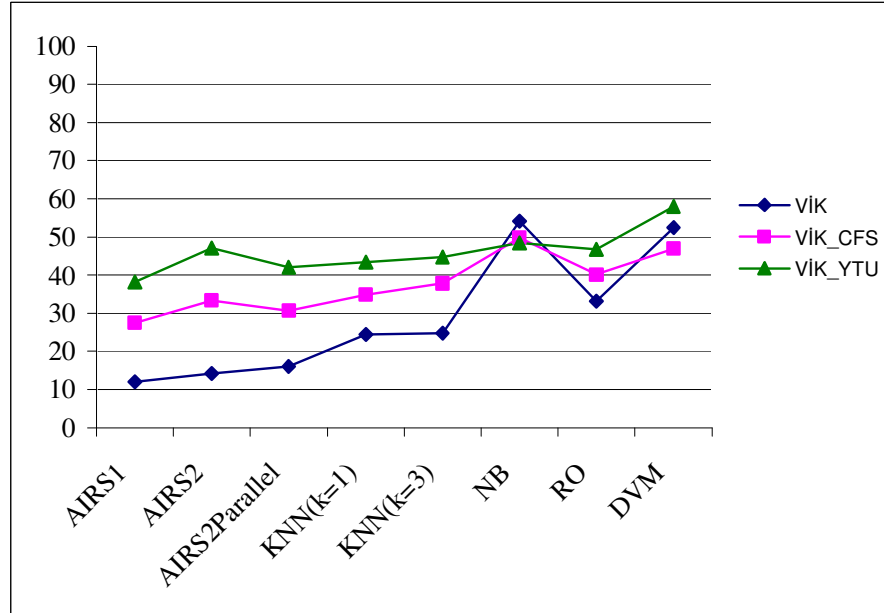
Çizelge 6.43: VK_{yk16} için Genel Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Paralel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V _{IST}	60.36	60	60.71	75	76.07	80.71	81.96	68.93	70.47
V _{IST_CFS}	64.11	65.71	67.14	77.5	76.79	81.96	80	68.39	72.70
V _{IST_YTU}	45	50.71	50.71	53.39	55.54	52.86	59.11	46.25	51.70
V _{KZ}	28.93	30.54	33.39	30.18	33.39	41.79	40.89	27.14	33.28
V _{DB}	35.38	35	39.82	40.71	45.18	50	45	25.54	39.58
V _{IK}	11.96	14.29	16.07	24.46	24.82	54.11	33.21	52.5	28.93
V _{IK_CFS}	27.5	33.39	30.71	34.82	37.86	49.82	40.18	46.96	37.66
V _{IK_YTU}	38.21	47.14	42.14	43.39	44.82	48.39	46.79	58.04	46.12
V _{2G_CFS}	45.71	47.5	45.71	56.07	57.5	74.29	67.32	70.18	58.04
V _{2G_YTU}	88.04	88.57	86.96	94.46	93.75	81.07	90	90.36	89.15
V _{3G_CFS}	35.71	44.46	46.43	74.64	73.93	83.39	71.79	89.29	64.96
V _{3G_YTU}	98.04	98.57	99.29	99.82	99.64	96.96	96.79	99.64	98.59
V _{KÖK_CFS}	31.96	36.61	42.32	60.71	61.96	76.61	57.86	79.82	55.98
V _{KÖK_YTU}	96.43	95.71	97.86	99.29	99.29	95.18	95	99.46	97.28
V _{GÖVDE_CFS}	33.04	38.04	43.39	58.21	61.25	70.89	55	78.93	54.84
V _{GÖVDE_YTU}	76.61	80	78.39	85	85.36	73.93	82.14	86.61	81.01
Ortalama	51.06	54.14	55.07	62.98	64.20	69.50	65.19	68.00	

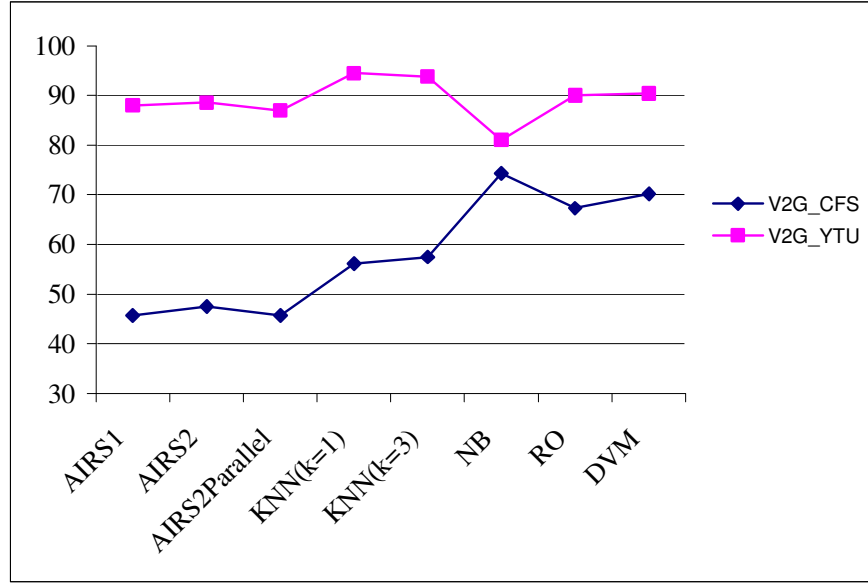
Şekil 6.49'a baktığımızda, CfsSubsetEval kullanılarak özellik sayısı azaltılan V_{IST} vektörünün RO haricinde başarılı olduğunu, Şekil 6.50'de de işlevsel kelimelerde boyut azaltmada YTU yönteminin kullanılması NB sınıflandırıcısı hariç hepsinde yüksek başarı vermiştir.



Şekil 6.49 VK_{yk16} için İstatistiksel Özellik Vektörleri Başarı Grafiği

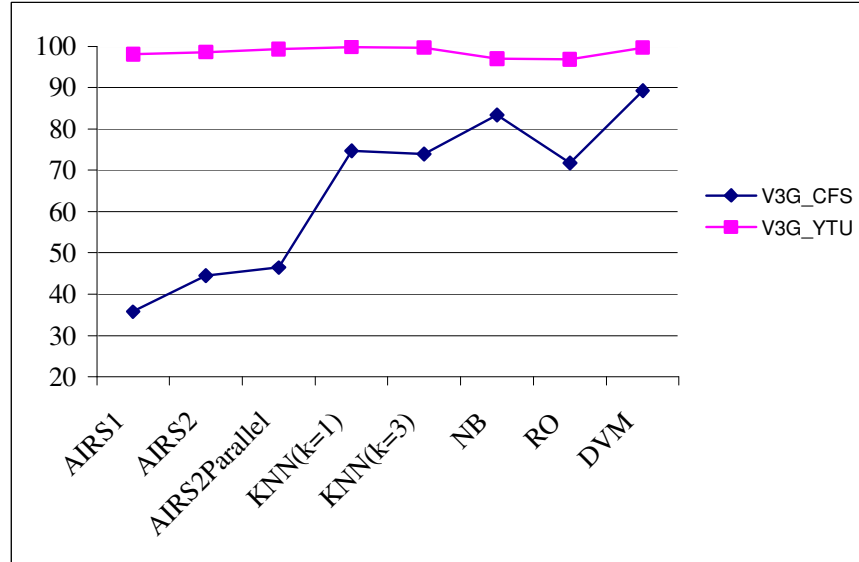


Şekil 6.50 VK_{yk16} için İşlevsel Kelimeler Özellik Vektörleri Başarı Grafiği

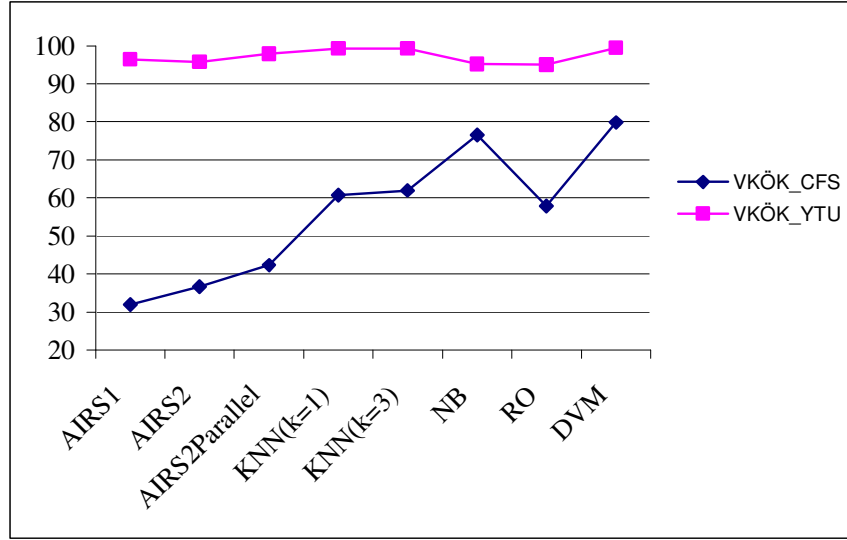


Şekil 6.51 VK_{yk16} için 2-gram Kelimeler Özellik Vektörleri Başarı Grafiği

Şekil 6.51 ve Şekil 6.52’de ise yine tür, cinsiyet, aynı türde yazan 10 ve 8 yazar, karışık türde yazan 20 ve 18 yazar veri kümesinde olduğu gibi Y2U ile indirgenmiş olan özellik vektörlerinin tüm sınıflandırıcılarda CfsSubsetEval’e göre daha başarılı olduğu görölmektedir.

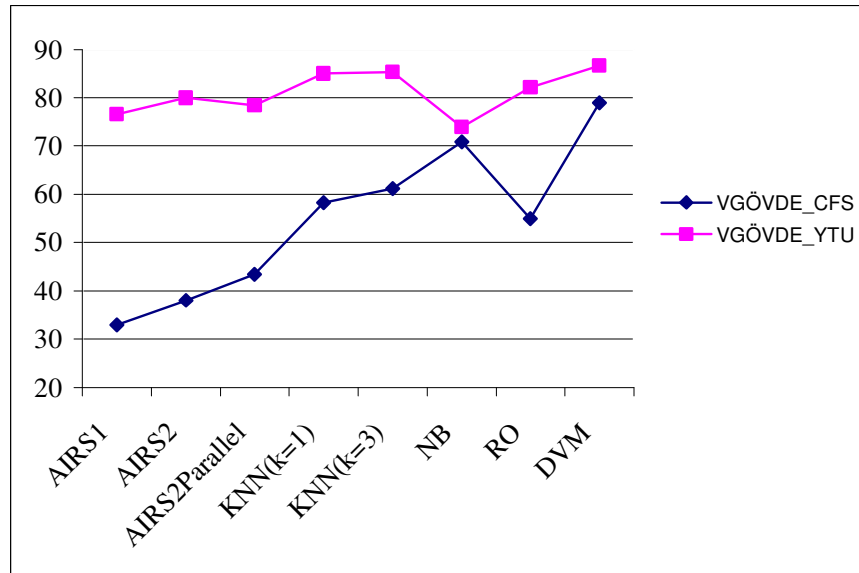


Şekil 6.52 VK_{yk16} için 3-gram Kelimeler Özellik Vektörleri Başarı Grafiği

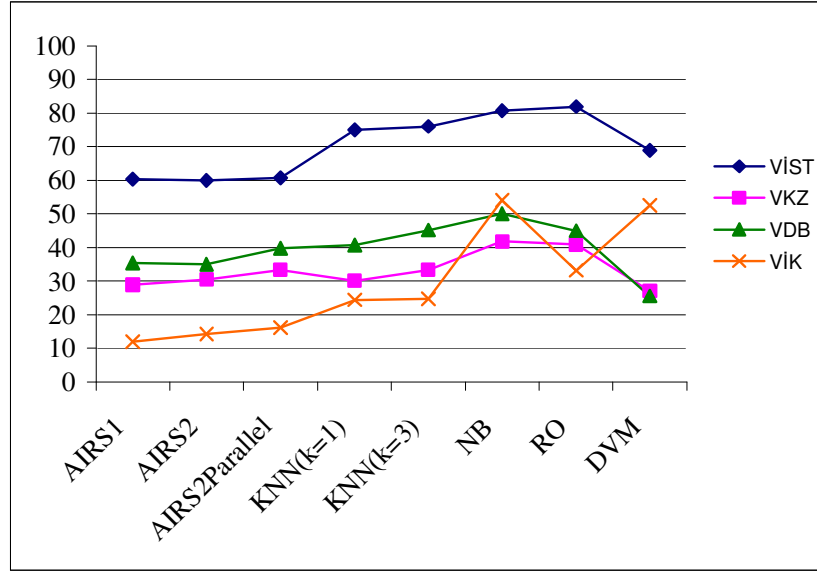


Şekil 6.53 VK_{yk16} için Kök Özellik Vektörleri Başarı Grafiği

Şekil 6.53’de köklerin ve Şekil 6.54’de gövdelerin kullanıldığı özellik vektörlerinde YTU’nun her sınıflandırıcıda yüksek bir performans sergilediği açıkça görülmektedir. NB yönteminde $V_{GÖVDE_CFS}$ ve $V_{GÖVDE_YTU}$ birbirine yakın bir başarı vermiştir.

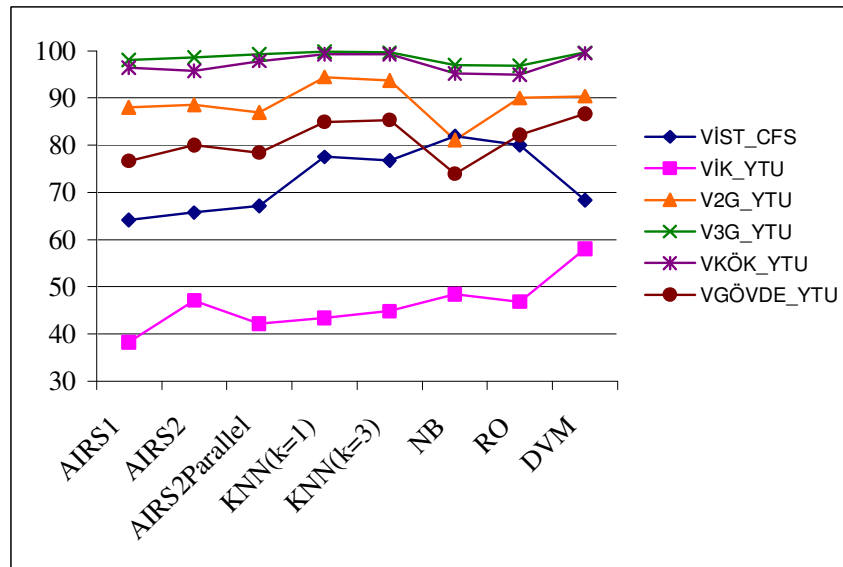


Şekil 6.54 VK_{yk16} için Gövde Özellik Vektörleri Başarı Grafiği



Şekil 6.55 $V_{K_{yk16}}$ için Boyut İndirgeme Yapılmayan Özellik Vektörleri Başarı Grafiği

Şekil 6.55’de, boyut indirgeme yapılmadan dört farklı özellik vektörünün başarı grafiği gösterilmektedir. V_{IST} özellik vektörü şimdiye kadar ki bütün veri kümelerinde olduğu gibi karma türde yazan yazarların tanınmasında da başarılı olmuştur. Şekil 6.56’da ise, kendi gruplarında en yüksek başarıyı veren özellik vektörleri gösterilmektedir. Yine V_{3G_YTU} özellik vektörü bütün sınıflandırıcılarda yüksek başarılar elde vermiştir. $V_{KÖK_YTU}$ da bazı sınıflandırıcılarda V_{3G_YTU} yakın başarı vermiştir.



Şekil 6.56 $V_{K_{yk16}}$ için En Başarılı Özellik Vektörleri Başarı Grafiği

6.3.2.4 Farklı Türde Yazar Tanıma İçin Deneysel Sonuçlar ($V_{K_{y_{k14}}}$)

Farklı türde yazılar yazan 14 yazardan oluşan yazar tanıma çalışmasında kullanılan sınıflandırıcıların, hangi özellik vektörleriyle yüzde kaçlık sınıflandırma başarısı gösterdikleri Çizelge 6.44'te gösterilmiştir. Buna göre en yüksek başarı %100 ile V_{3G_YTU} özellik vektörü kullanılarak AIRS1, KNN ve DVM sınıflandırıcıları ile alınmıştır. YBS sınıflandırıcılarından AIRS2Parallel, V_{3G_YTU} özellik vektörüyle %99.59, AIRS2 ise %99.80 başarı performansı vermiştir.

Tüm özellik vektörleri arasında ortalama en yüksek başarı %99.44 ile V_{3G_YTU} vektörüyle elde edilmiştir. Onu %98.62 ile $V_{KÖK_YTU}$ özellik vektörü takip etmektedir. $V_{GÖVDE_YTU}$ özellik vektörü ise % 84.92 ortalama sınıflandırma başarısı göstermiştir. En düşük başarı oranı ise %37.14 ile V_{KZ} özellik vektörüyle alınmıştır. Bu da Türkçe için kelime zenginliğine dayalı özelliklerin tek başlarına kullanıldıklarında yazar tanıma çalışmaları için iyi bir ayırt edici olmadığını göstermektedir. Karakter n-gram'lar ise cinsiyet ve tür belirlemede olduğu gibi yazar tanıma çalışmasında da yüksek başarı oranları sağlamıştır. 2-gram özellik vektörü V_{2G_YTU} ile %91.43'lik başarı ortalaması alınmıştır. 3-gram'ların Türkçe metinlerde yazar tanıma çalışmalarında 2-gram'lara göre daha iyi performans verdikleri söylenebilir.

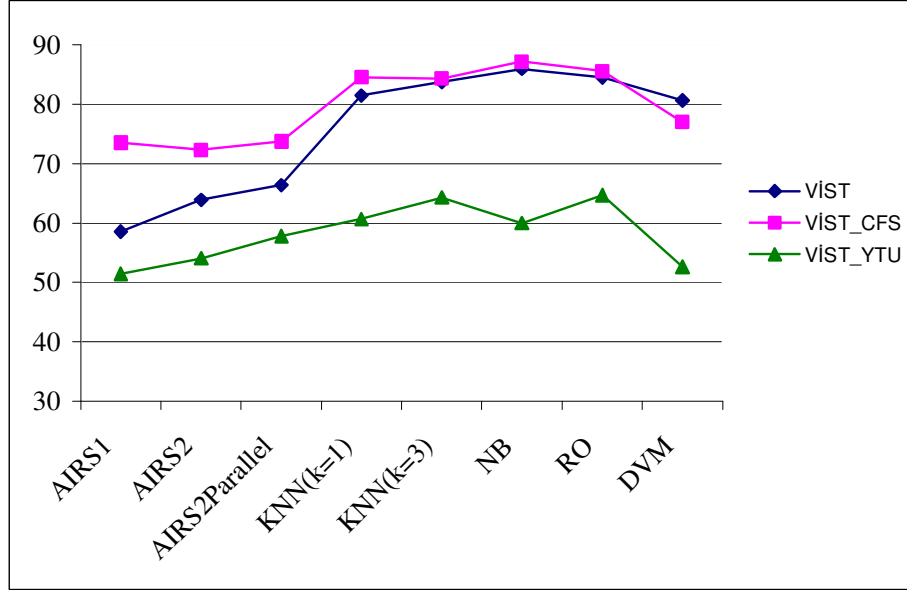
Deneylerde yazar sayısı azaldıkça sınıflandırma performansının tüm yöntemlerde arttığını söyleyebiliriz. Örneğin, V_{IST} özellik vektörü 20 yazar için ortalama %66.2, 18 yazar için %66.42, 16 yazar için %70.47, 14 yazar içinse %75.61 başarı performansı göstermiştir.

Sınıflandırma yöntemlerinin tüm özellik vektörleri ile ortalama başarılarına bakıldığında ise NB sınıflandırıcısı %73.71 ile en iyi ortalamayı vermektedir. YBS sınıflandırıcıları içerisinde en iyi ortalamayı ise %58.28 ile AIRS2Parallel vermiştir.

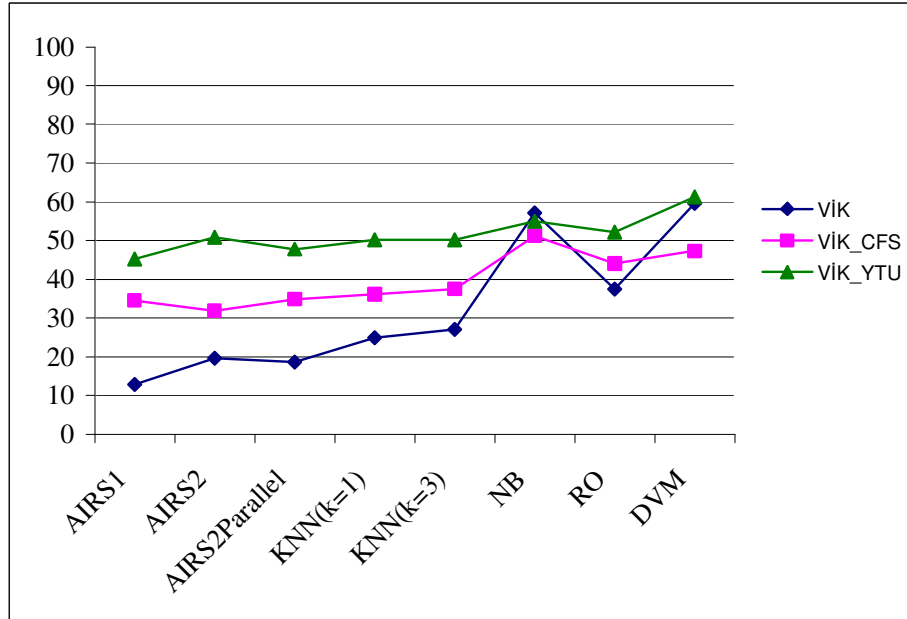
Çizelge 6.44: VK_{yk14} için Genel Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Paralel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V_{IST}	58.57	63.88	66.33	81.43	83.67	85.92	84.49	80.61	75.61
V_{IST_CFS}	73.47	72.24	73.67	84.49	84.29	87.14	85.51	76.94	79.72
V_{IST_YTU}	51.43	54.08	57.76	60.61	64.29	60	64.69	52.65	58.19
V_{KZ}	30.20	34.49	33.06	35.92	38.16	47.14	44.69	33.47	37.14
V_{DB}	36.12	45.71	44.69	42.65	48.78	55.31	49.39	28.98	43.95
V_{IK}	12.86	19.59	18.57	24.90	27.14	57.14	37.55	59.59	32.17
V_{IK_CFS}	34.49	31.84	34.90	36.12	37.55	51.22	44.08	47.35	39.69
V_{IK_YTU}	45.31	50.82	47.76	50.20	50.20	54.90	52.24	61.22	51.58
V_{2G_CFS}	48.57	54.69	50.82	62.45	63.27	76.53	71.43	77.76	63.19
V_{2G_YTU}	90.61	90.20	90.20	95.92	95.10	87.55	90.20	91.63	91.43
V_{3G_CFS}	31.84	45.71	50.61	78.37	78.78	85.31	74.49	91.84	67.12
V_{3G_YTU}	100	99.80	99.59	100	100	98.57	97.55	100	99.44
$V_{KÖK_CFS}$	35.10	42.86	42.65	55.92	62.04	76.73	64.49	84.49	58.04
$V_{KÖK_YTU}$	97.96	99.18	98.57	99.80	99.39	97.35	96.94	99.80	98.62
$V_{GÖVDE_CFS}$	34.08	41.02	38.57	61.84	64.08	78.98	64.69	83.06	58.29
$V_{GÖVDE_YTU}$	81.22	84.29	84.69	87.96	88.78	79.59	83.27	89.59	84.92
Ortalama	53.86	58.15	58.28	66.16	67.85	73.71	69.11	72.44	

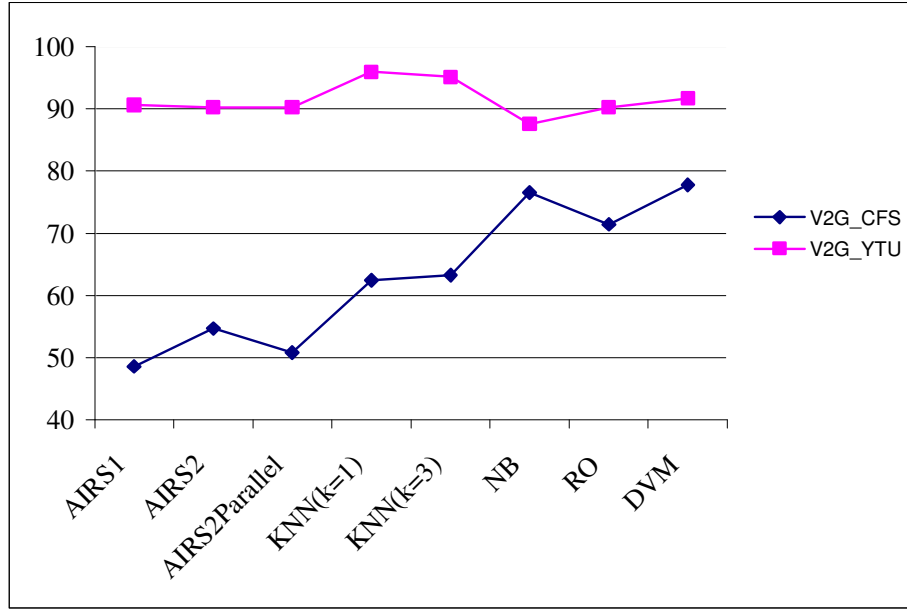
Şekil 6.57'ye baktığımızda, CfsSubsetEval kullanılarak özellik sayısı azaltılan V_{IST} vektörünün DVM haricinde başarılı olduğunu, Şekil 6.58'de de işlevsel kelimelerde boyut azaltmada YTU yönteminin kullanılması NB sınıflandırıcısı hariç hepsinde yüksek başarı vermiştir.



Şekil 6.57 VK_{yk14} için İstatistiksel Özellik Vektörleri Başarı Grafiği

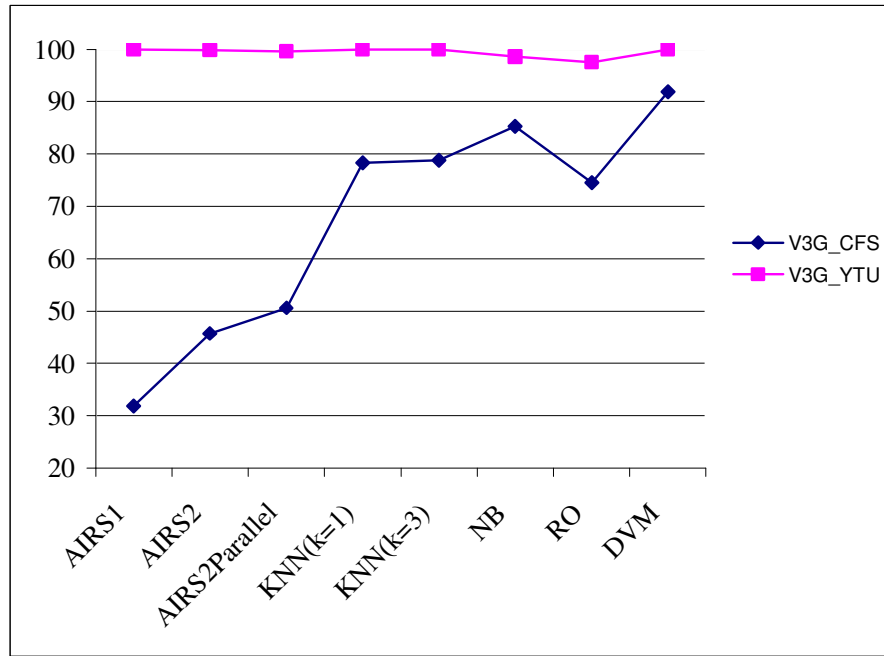


Şekil 6.58 VK_{yk14} için İşlevsel Kelimeler Özellik Vektörleri Başarı Grafiği

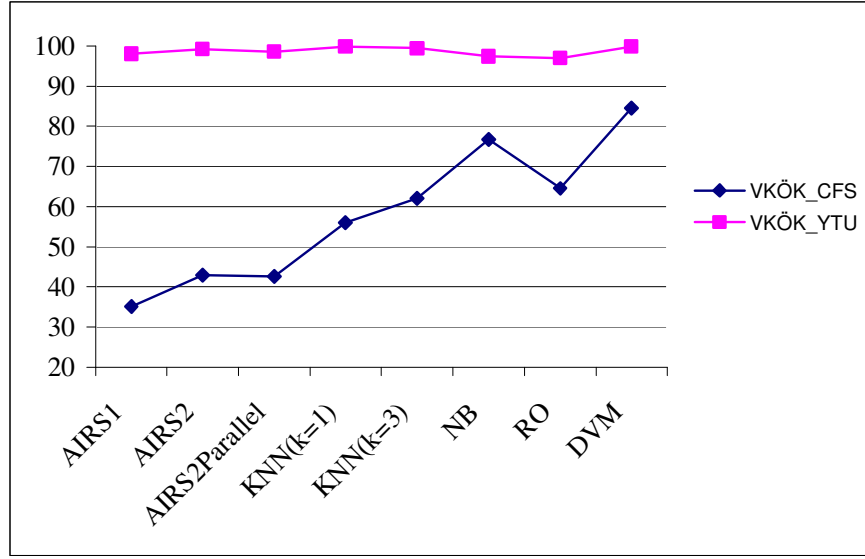


Şekil 6.59 VK_{yk14} için 2-gram Özellik Vektörleri Başarı Grafiği

Şekil 6.59 ve Şekil 6.60'da ise yine tür, cinsiyet, aynı türde yazan 10 ve 8 yazar, karışık türde yazan 20,18 ve 16 yazar veri kümesinde olduğu gibi YTU ile indirgenmiş olan özellik vektörlerinin tüm sınıflandırıcılarda CfsSubsetEval'e göre daha başarılı olduğu görülmektedir.

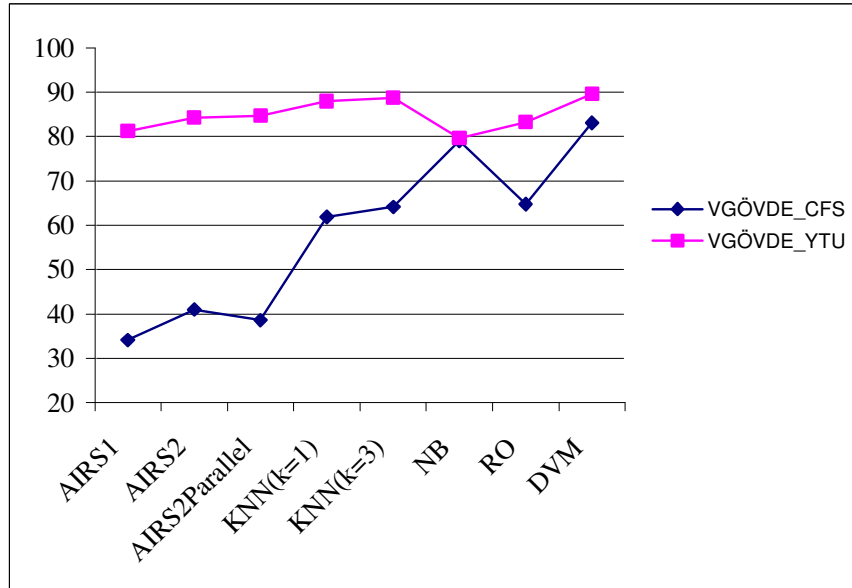


Şekil 6.60 VK_{yk14} için 3-gram Özellik Vektörleri Başarı Grafiği

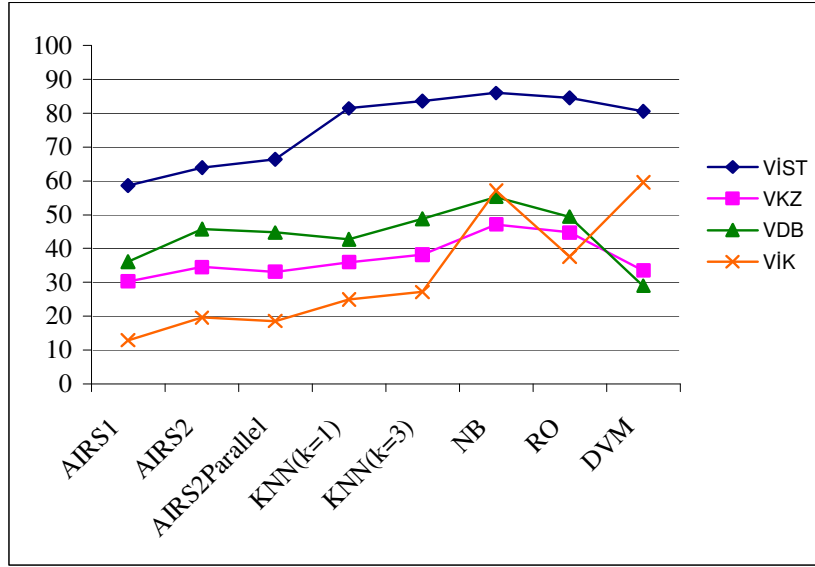


Şekil 6.61 VK_{yk14} için Kök Özellik Vektörleri Başarı Grafiği

Şekil 6.61’de köklerin ve Şekil 6.62’de gövdelerin kullanıldığı özellik vektörlerinde YTU’nun her sınıflandırıcıda yüksek bir performans sergilediği açıkça görülmektedir. NB yönteminde $V_{GÖVDE_CFS}$ ve $V_{GÖVDE_YTU}$ aynı başarıyı vermiştir.

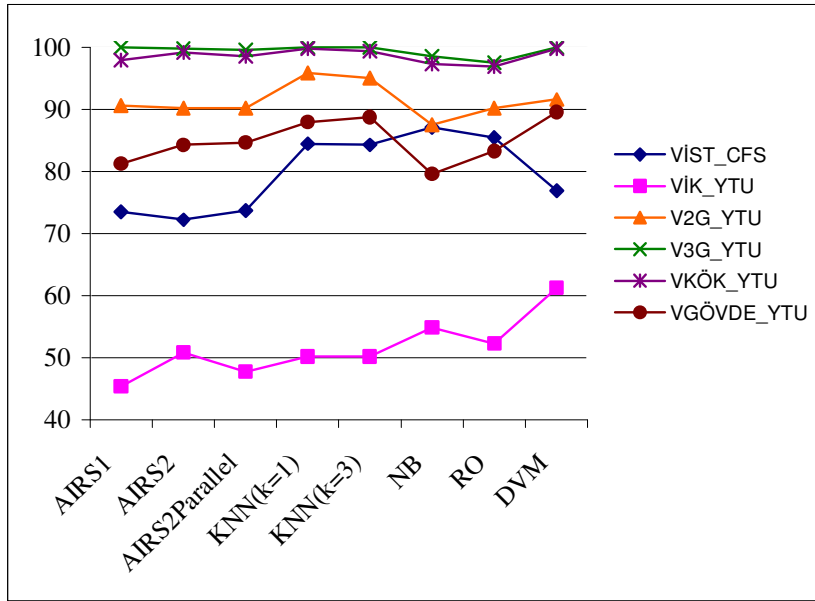


Şekil 6.62 VK_{yk14} için Gövde Özellik Vektörleri Başarı Grafiği



Şekil 6.63 VK_{yk14} için Boyut İndirgeme Yapılmayan Özellik Vektörleri Başarı Grafiği

Şekil 6.63'te, boyut indirgeme yapılmadan dört farklı özellik vektörünün başarı grafiği gösterilmektedir. V_{IST} özellik vektörü şimdiye kadar ki bütün veri kümelerinde olduğu gibi karma türde yazan yazarların tanınmasında da başarılı olmuştur. Şekil 6.64'de ise, kendi gruplarında en yüksek başarıyı veren özellik vektörleri gösterilmektedir. Yine V_{3G_YTU} özellik vektörü bütün sınıflandırıcılarda yüksek başarılar elde vermiştir. V_{KÖK_YTU} 'da bazı sınıflandırıcılarda V_{3G_YTU} yakın başarı vermiştir.



Şekil 6.64 VK_{yk14} için En Başarılı Özellik Vektörleri Başarı Grafiği

6.3.2.5 Farklı Türde Yazar Tanıma İçin Deneysel Sonuçlar ($V_{K_{yk12}}$)

Farklı türde yazılar yazan 12 yazardan oluşan yazar tanıma çalışmasında kullanılan sınıflandırıcıların, hangi özellik vektörleriyle yüzde kaçlık sınıflandırma başarısı gösterdikleri Çizelge 6.45’de gösterilmiştir. Buna göre en yüksek başarı %100 ile V_{3G_YTU} özellik vektörü kullanılarak AIRS2, KNN ($k=1$) ve DVM sınıflandırıcıları ile alınmıştır. YBS sınıflandırıcılarından AIRS1 ve AIRS2Parallel, V_{3G_YTU} özellik vektörüyle %99.76 başarı performansı vermiştir.

Tüm özellik vektörleri arasında ortalama en yüksek başarı %99.73 ile V_{3G_YTU} vektörüyle elde edilmiştir. Onu %98.75 ile $V_{KÖK_YTU}$ özellik vektörü takip etmektedir. $V_{GÖVDE_YTU}$ özellik vektörü ise % 88.93 ortalama sınıflandırma başarısı göstermiştir. En düşük başarı oranı ise %34.49 ile V_{IK} özellik vektörüyle alınmıştır. Bu da Türkçe için işlevsel kelimelerin tek başlarına kullanıldıklarında yazar tanıma çalışmaları için iyi bir ayırt edici olmadığını göstermektedir. Karakter n-gram’lar ise cinsiyet ve tür belirlemede olduğu gibi yazar tanıma çalışmasında da yüksek başarı oranları sağlamıştır. 2-gram özellik vektörü V_{2G_YTU} %94.38’lik başarı ortalaması sağlamıştır. 3-gram’ların Türkçe metinlerde yazar tanıma çalışmalarında 2-gram’lara göre daha iyi performans verdikleri söylenebilir.

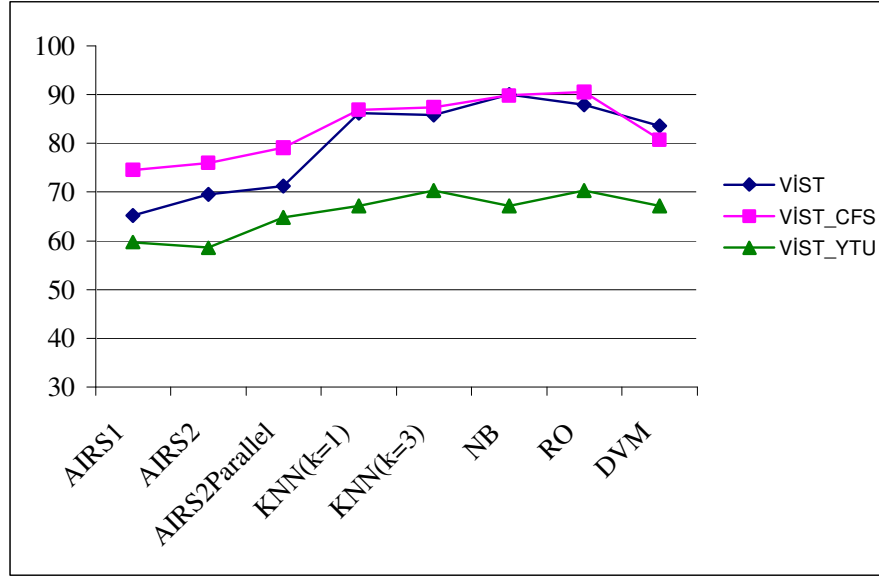
Deneylerde yazar sayısı azaldıkça sınıflandırma performansının tüm yöntemlerde arttığını bu denemelerde de görebiliriz. Örneğin, V_{IST} özellik vektörü 20 yazar için ortalama %66.2, 18 yazar için %66.42, 16 yazar için %70.47, 14 yazar için %75.61, 12 yazar içinse %79.91 başarı performansı göstermiştir.

Sınıflandırma yöntemlerinin tüm özellik vektörleri ile ortalama başarılarına bakıldığında ise NB sınıflandırıcısı %76.85 ile en iyi ortalamaı vermektedir. YBS sınıflandırıcıları içerisinde en iyi ortalamaı ise %63.15 ile AIRS2Parallel vermiştir.

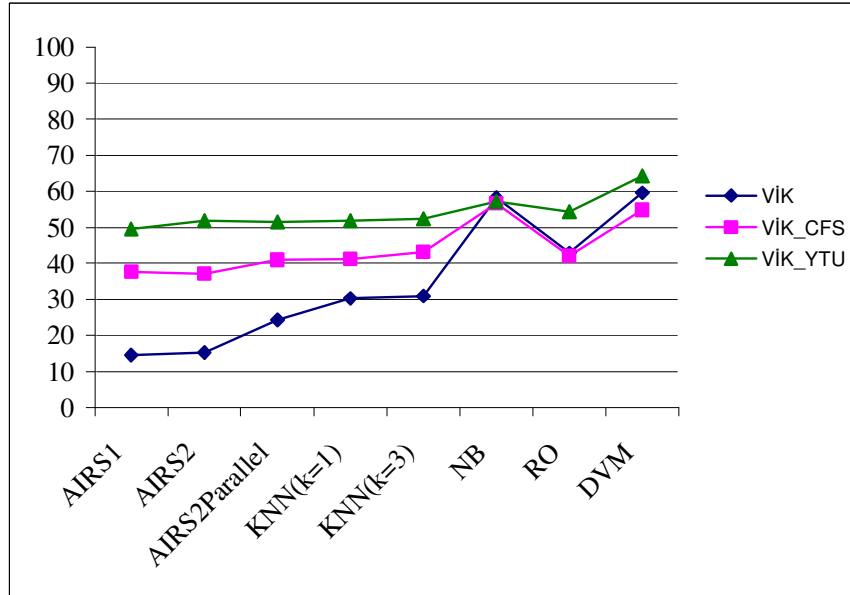
Çizelge 6.45: VK_{yk12} için Genel Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2Paralel	KNN(k=1)	KNN(k=3)	NB	RO	DVM	Ortalama
V _{IST}	65.24	69.52	71.19	86.19	85.71	90	87.86	83.57	79.91
V _{IST_CFS}	74.52	75.95	79.05	86.90	87.38	89.76	90.48	80.71	83.09
V _{IST_YTU}	59.76	58.57	64.76	67.14	70.24	67.14	70.24	67.14	65.62
V _{KZ}	37.14	36.43	39.05	39.52	41.90	51.19	47.62	36.43	41.16
V _{DB}	42.62	49.52	50.71	50.24	53.33	58.81	57.86	30.95	49.26
V _{IK}	14.52	15.24	24.29	30.24	30.95	58.33	42.86	59.52	34.49
V _{IK_CFS}	37.62	37.14	40.95	41.19	43.10	56.67	42.14	54.76	44.20
V _{IK_YTU}	49.52	51.90	51.43	51.90	52.38	57.14	54.29	64.29	54.11
V _{2G_CFS}	52.38	59.76	58.33	69.05	69.52	79.76	75.24	80.95	68.12
V _{2G_YTU}	94.05	93.33	93.10	97.38	96.67	91.19	95.24	94.05	94.38
V _{3G_CFS}	37.38	47.62	57.14	77.14	79.29	86.43	83.10	93.57	70.21
V _{3G_YTU}	99.76	100	99.76	100	100	99.52	98.81	100	99.73
V _{KÖK_CFS}	37.86	44.29	47.14	64.05	64.52	78.57	65	85.48	60.86
V _{KÖK_YTU}	98.81	97.14	99.52	99.52	99.76	98.10	97.62	99.52	98.75
V _{GÖVDE_CFS}	41.90	43.81	43.57	63.10	65.95	81.43	69.52	85.71	61.87
V _{GÖVDE_YTU}	87.38	88.10	90.48	90.24	91.67	85.48	85.48	92.62	88.93
Ortalama	58.15	60.52	63.15	69.61	70.77	76.85	72.71	75.58	

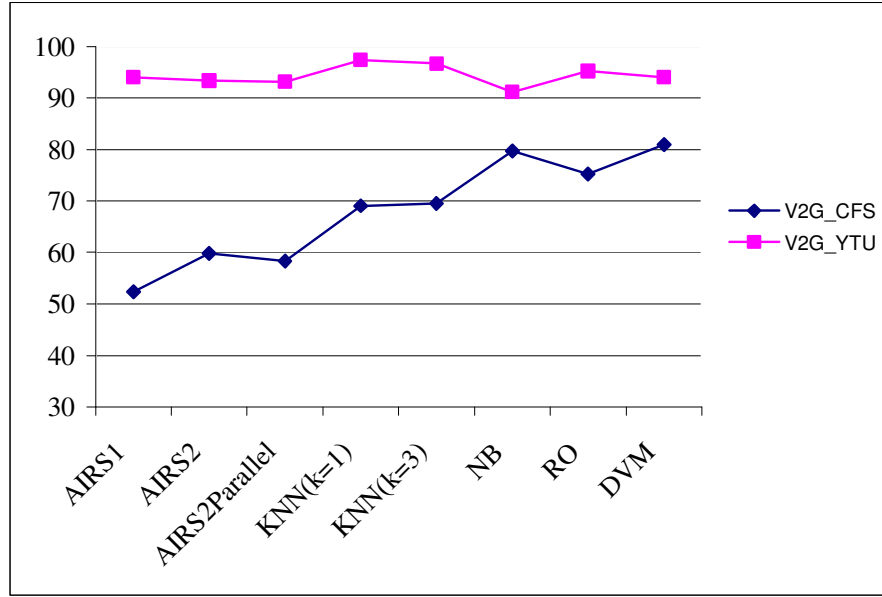
Şekil 6.65'e baktığımızda, CfsSubsetEval kullanılarak özellik sayısı azaltılan V_{IST} vektörünün DVM haricinde başarılı olduğunu, Şekil 6.66'da da işlevsel kelimelerde boyut azaltmada YTU yönteminin bütün sınıflandırıcılarda yüksek başarı verdiği görülmüştür.



Şekil 6.65 $V_{K_{12}}$ için İstatistiksel Özellik Vektörleri Başarı Grafiği

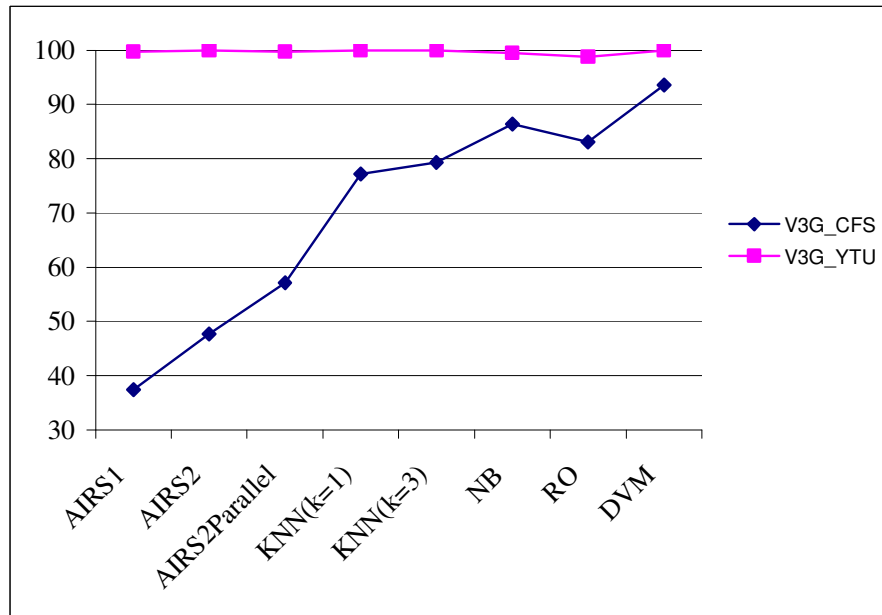


Şekil 6.66 $V_{K_{12}}$ için İşlevsel Kelimeler Özellik Vektörleri Başarı Grafiği

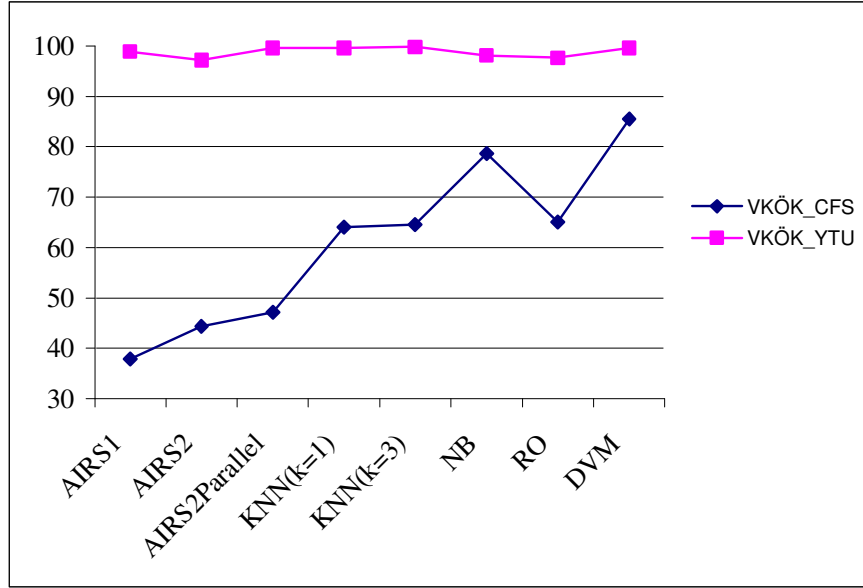


Şekil 6.67 VK_{yk12} için 2-gram Özellik Vektörleri Başarı Grafiği

Şekil 6.67 ve Şekil 6.68’da ise yine tür, cinsiyet, aynı türde yazan 10 ve 8 yazar, karışık türde yazan 20,18,16 ve 14 12 yazar veri kümesinde olduğu gibi Y2U ile indirgenmiş olan özellik vektörlerinin tüm sınıflandırıcılarda CfsSubsetEval’e göre daha başarılı olduğu görülmektedir.

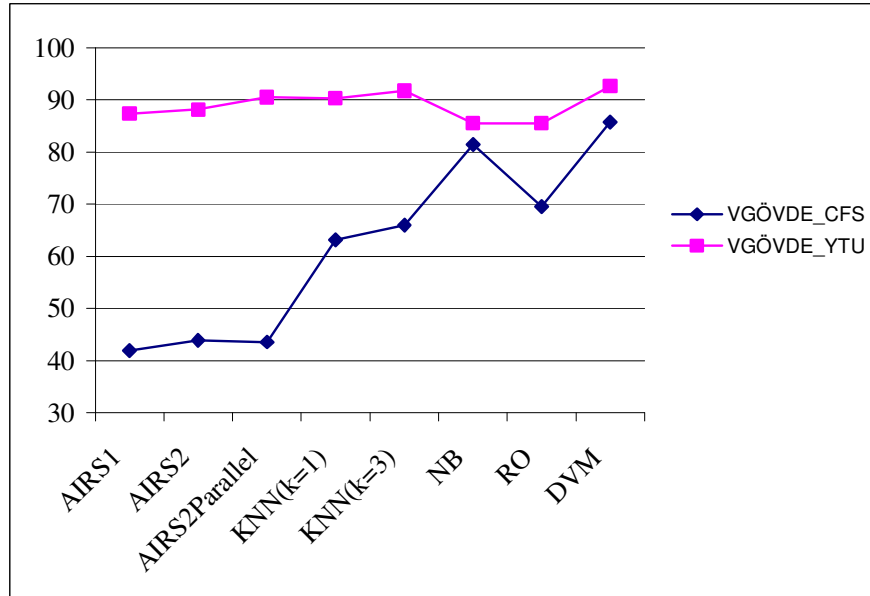


Şekil 6.68 VK_{yk12} için 3-gram Özellik Vektörleri Başarı Grafiği

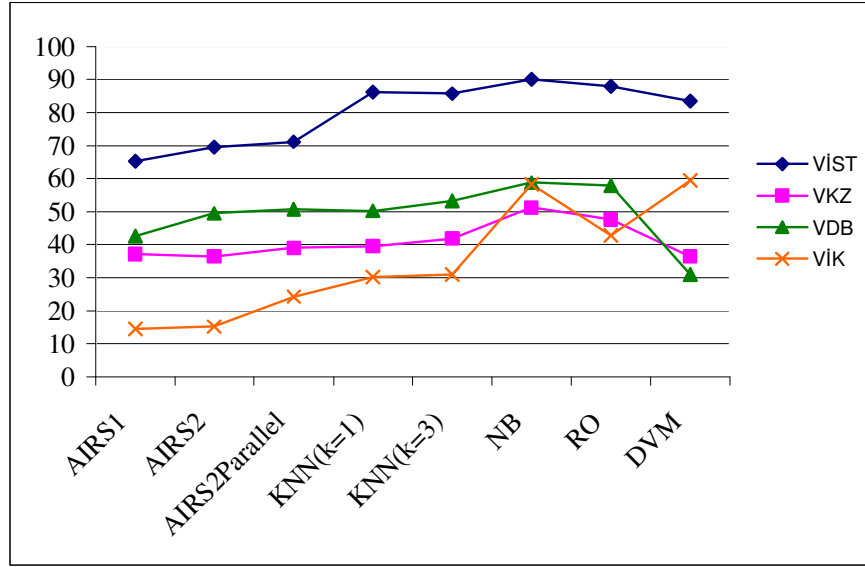


Şekil 6.69 VK_{yk12} için Kök Özellik Vektörleri Başarı Grafiği

Şekil 6.69’da köklerin ve Şekil 6.70’de gövdelerin kullanıldığı özellik vektörlerinde YTU’nun her sınıflandırıcıda yüksek bir performans sergilediği açıkça görülmektedir.

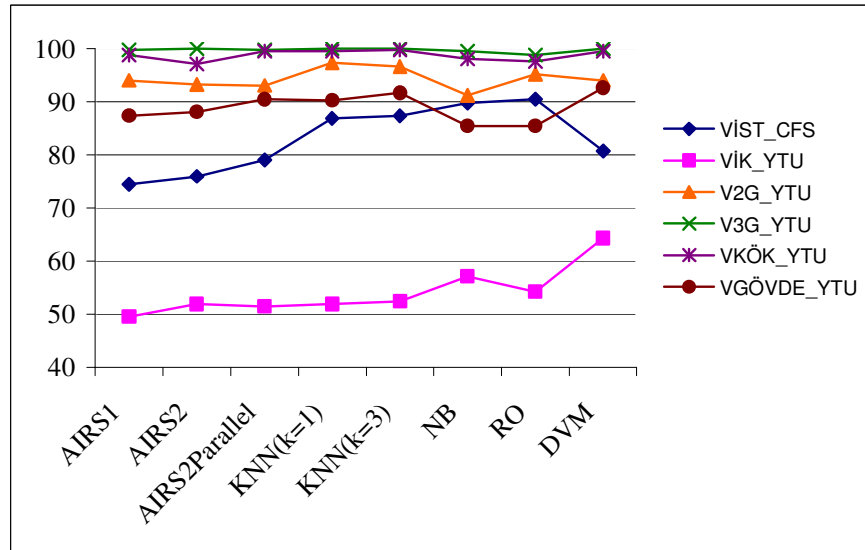


Şekil 6.70 VK_{yk12} için Gövde Özellik Vektörleri Başarı Grafiği

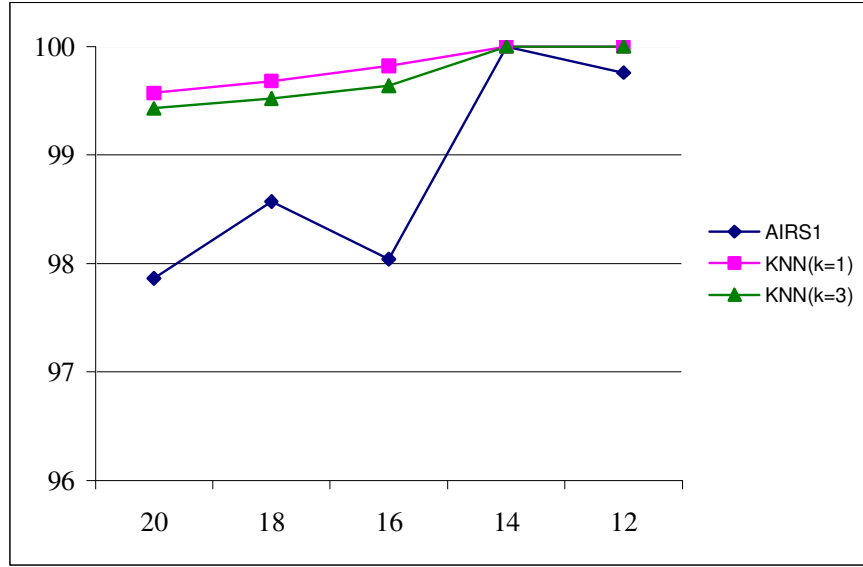


Şekil 6.71 VK_{yk12} için Boyut İndirgeme Yapılmayan Özellik Vektörleri Başarı Grafiği

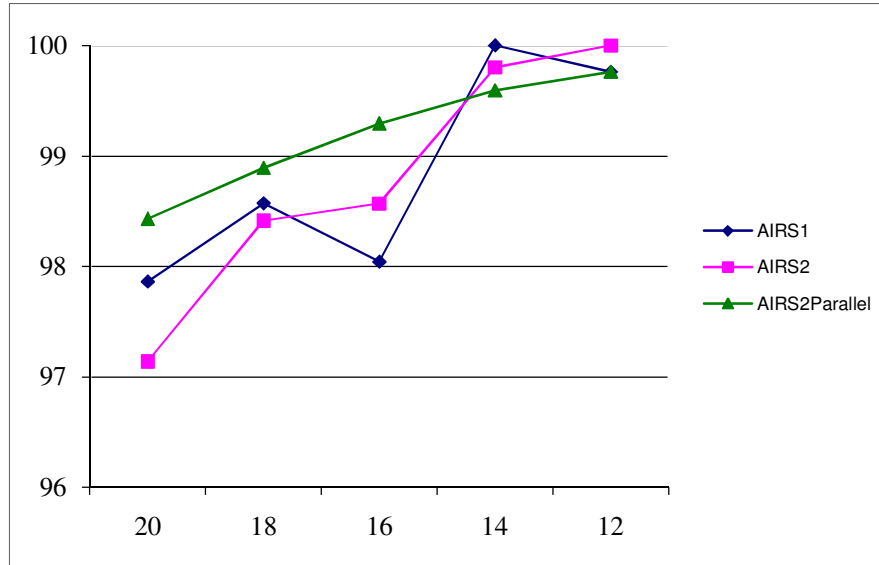
Şekil 6.71’de, boyut indirgeme yapılmadan dört farklı özellik vektörünün başarı grafiği gösterilmektedir. V_{IST} özellik vektörü şimdiye kadar ki bütün veri kümelerinde olduğu gibi karma türde yazan yazarların tanınmasında da başarılı olmuştur. Şekil 6.72’de ise, kendi gruplarında en yüksek başarıyı veren özellik vektörleri gösterilmektedir. Yine V_{3G_YTU} özellik vektörü bütün sınıflandırıcılarda yüksek başarılar elde vermiştir. $V_{KÖK_YTU}$ bazı sınıflandırıcılarda V_{3G_YTU} ‘ya yakın başarı vermiştir.



Şekil 6.72 VK_{yk12} için En Başarılı Özellik Vektörleri Başarı Grafiği



Şekil 6.73 V_{3G_YTU} Özellik Vektörü ile AIRS1 ve KNN Algoritmalarının Başarı Grafiği



Şekil 6.74 V_{3G_YTU} Özellik Vektörü ile YBS Algoritmalarının Başarı Grafiği

Şekil 6.73'te, V_{3G_YTU} Özellik Vektörü ile en yüksek başarı oranı veren algoritmalar olan AIRS1 ve KNN'in, Şekil 6.74'de ise, V_{3G_YTU} özellik vektörü ile YBS sınıflandırıcılarının başarı oranları verilmiştir. Yazar sayısı azaldıkça başarı oranlarının arttığı ve YBS algoritmalarının KNN'in göstermiş olduğu performansa yaklaştığı görülmektedir.

6.4 YTU Boyut Azaltma Yönteminin Farklı Bir Dil için Kullanımı

Türkçe dokümanlar kullanılarak gerçekleştirilen yazar tanıma, yazarın cinsiyetini belirleme ve tür tanıma çalışmalarında, seçilen özellik vektörlerinin YTU boyut indirgeme yöntemi ile özellik sayısı azaltılarak yapılmış olan sınıflandırma işlemlerinde başarılı sonuçlar elde edilmiştir.

Bu tez çalışmasında kullanılan YTU boyut indirgeme yönteminin Türkçe diline özgü bir başarı göstermediğini kanıtlamak amacıyla, İngilizce dokümanlar üzerinde tür belirleme çalışması yapılmıştır. Kullandığımız Reuters veri kümesi [8] içerisinde yazar tanıma ve cinsiyet belirleme için gerekli düzenlemeler olmadığından, sadece tür belirleme için kullanılması uygun görülmüştür.

6.4.1 Reuters Veri Kümesi

Reuters dünyaca ünlü bir haber ajansı olup, 23 dilde günde yaklaşık 11,000 haber metni üretmektedir. Reuters RCV1 veri kümesi 1996-1997 yılları arasında İngilizce haberlerden derlenmiş XML formatında bir veri kümesidir. Veri kümesindeki her dokümanın çeşitli konu kodları vardır. Konu kodları o dokümanın hangi türde olduğu hakkında bize fikir vermektedir. Reuters veri kümesi 4 ana konu başlığından oluşur. Bunlar CCAT (Endüstriyel), ECAT (Ekonomi), GCAT (Siyasi / Sosyal) ve MCAT (Pazarlama) konularıdır. Veri kümesi ayrıca bu ana konu başlıklarının detaylandırılmasından oluşan, toplam 126 konu kodu içermektedir. Bir doküman birden fazla konu kodu içerebilmektedir. Örneğin, CCAT-C15 konu kodlu bir doküman endüstriyel ve performans konuları hakkında bir içeriğe sahiptir. (Lewis vd., 2004).

Bu çalışmada bu konu kodlarını kullanarak 5 farklı tür belirlenmiştir. Bunlar:

- CCAT-C15 (Endüstriyel - Performans)
- ECAT-E21 (Ekonomi - Finans)
- GCAT-GSPO (Sosyal - Spor)
- MCAT-M11 (Pazarlama – Özsermaye Pazarları)
- GCAT-GPOL (Siyaset – İç Siyaset)

Her türden 500'er doküman seçilerek toplam 2500 dokümanın türü belirlenmeye çalışılmıştır.

6.4.2 Özellik Vektörü

Özellik olarak Reuters veri kümesinden elde edilen 5 farklı türdeki dokümanların kelime gövdeleri alınmıştır. İngilizce kelimelerin gövdelerini çıkarabilmek için Porter Gövde Algoritmasından [9] (The Porter Stemming Algorithm) faydalanılmıştır. Tüm dokümanlardan ilgili algoritma kullanılarak 21488 adet gövde elde edilmiştir.

Çalışmada 3 farklı özellik vektörü kullanılmıştır. Bunlardan VK_{TEM} 21488 adet gövdeden en çok kullanılan 2000 gövde seçilerek oluşturulmuştur. VK_{TEM_CFS} , Korelasyon Tabanlı Özellik Seçme kullanılarak, özellik sayısının 62'ye indirildiği bir özellik vektörüdür. VK_{TEM_YTU} ise YTU boyut azaltıcısı ile elde edilmiş, sınıf sayısına eşit 5 özellikten oluşan bir özellik vektörüdür.

6.4.3 Deneysel Sonuçlar

Yukarıdaki bölümlerde de bahsedildiği gibi Türkçe için tür tanımda 250 dokümandan 5878 tane farklı kelime gövdesi elde edilmiş ve bunların tamamı alınmıştır. CfsSubsetEval ile bu özellik sayısı 74 indirilmiştir. YTU yöntemi ile de sınıf sayısına düşürülmüştür. İngilizcede ise 2500 dokümandan 21488 tane farklı kelime gövdesi elde edilmiş olup, frekans dağılımına göre ilk 2000 tanesi alınmıştır. CfsSubsetEval ile bu özellik sayısı 62'ye indirilmiş ve yine YTU yöntemi ile sınıf sayısı olan 5'e düşürülmüştür.

Çizelge 6.46'da İngilizce ve Türkçe dokümanlar üzerinde gövde özellikleri ile dokümanın türünü belirleme çalışmasında elde edilen başarı sonuçları verilmektedir.

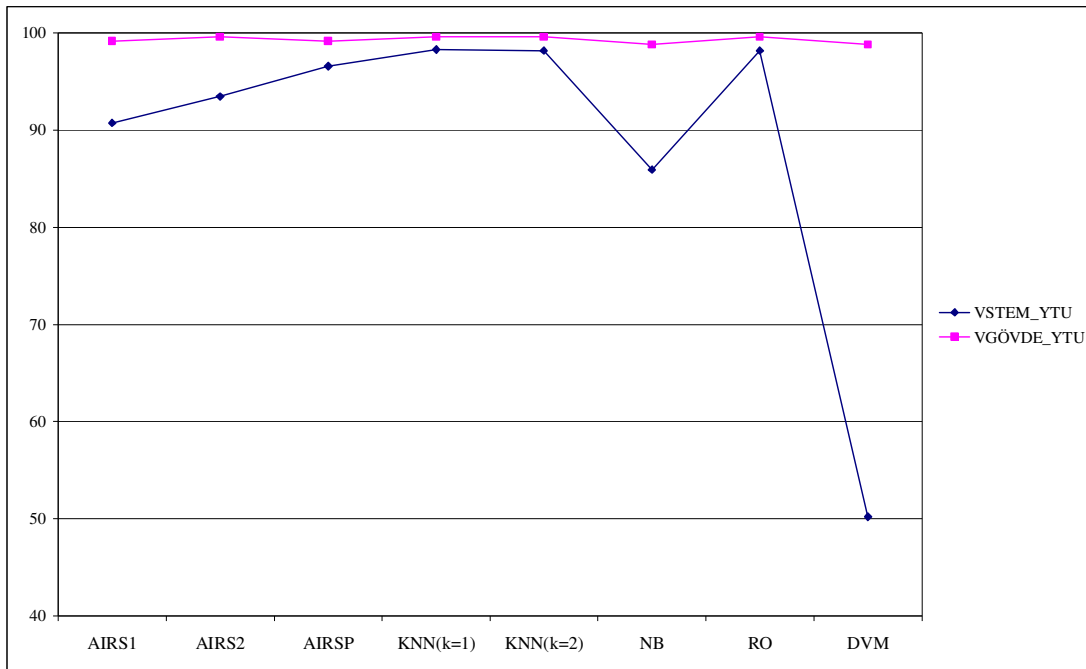
Çizelge 6.46: İngilizce ve Türkçe Gövde Özellikleri ile Tür Belirleme Başarı Oranları (%)

	AIRS1	AIRS2	AIRS2P	KNN (k=1)	KNN (k=3)	NB	RO	DVM
VK_{TEM}	NA	28.92	28.36	79.12	74.8	79.6	NA	93.7
VK_{TEM_CFS}	55.4	64.1	63.6	87.7	88.3	86.3	91.2	88.2
VK_{TEM_YTU}	90.72	93.48	96.6	98.3	98.2	85.9	98.2	50.2
$V_{GÖVDE_CFS}$	60.4	58.4	61.6	85.2	84	93.6	86.4	94
$V_{GÖVDE_YTU}$	99.2	99.6	99.2	99.6	99.6	98.8	99.6	98.8

İngilizce dilinde yapılan çalışmada NB ve DVM sınıflandırıcıları hariç, VK_{TEM_YTU} , VK_{TEM_CFS} 'den daha başarılı sonuçlar vermiştir. Türkçe dilinde yapılan çalışmada ise YTU yöntemi bütün sınıflandırıcılarda başarılı olmuştur.

Türkçe'de en başarılı sınıflandırıcı Yapay Bağışıklık Sistemlerinde %99.6 ile AIRS2, diğer sınıflandırıcılarda ise yine aynı başarı oranı ile KNN ve RO olmuştur. İngilizce'de Yapay Bağışıklık Sistemlerinde en başarılı sınıflandırıcı %96.6 ile AIRSP olmuş, diğer sınıflandırıcılarda ise %98.3 ile KNN (k=1) olmuştur.

Şekil 6.75'te her iki dilde yaklaşık en iyi performansı veren YTU'nun bütün sınıflandırıcılardan tür tanıma için almış olduğu başarı performansı gösterilmektedir.



Şekil 6.75 VK_{TEM_YTU} ve $VGÖVDE_YTU$ Özellik Vektörlerinin Tür Tanıma Başarı Grafiği (%)

İngilizce ile yapılmış olan tür tanıma çalışmasının sınıflandırma başarısı Türkçe'ye göre biraz daha düşük çıkmıştır. Bunun sebebi olarak 1) İngilizce için kullanılan veri kümesinin Türkçe için kullanılmış olan veri kümesinden daha büyük olduğunu gösterebiliriz. Türkçe'de toplamda 250 doküman ile çalışılırken İngilizce'de 2500 doküman ile çalışılmıştır. Buna karşılık Türkçe dokümanların içeriği daha zengindir. 2) Türkçe sonradan eklemeli bir dildir. Kelimeler köke yapılan ekler ile türetilir. Bizde bu çalışmada kelimenin gövdesini aldığımızdan İngilizce'ye göre daha az sayıda gövde elde edilir.

Buna karşılık $VK_{\text{TEM_YTU}}$ 'nın DVM'de neden bu kadar düşük bir başarı verdiğini açıklamada zorlanmaktayız. Kısacası önerilen bu yeni boyut azaltma yöntemi bütün doğal diller için kullanılabilir.

7. SONUÇ

Bu çalışmanın amacı Türkçe metinler üzerinde Yapay Bağışıklık Sistemlerinin dokümanların türünü, yazarını ve yazarlarının cinsiyetlerini belirlemede kullanılabileceğinin gösterilmesidir. Çalışma üç ana başlık altında toplanmıştır. Bunlar doküman yazarının cinsiyetini belirleme, dokümanın türünü belirleme ve dokümanın yazarını tanımadır. Her üç başlık için veri kümeleri oluşturulmuş ve çeşitli özellik vektörleri kullanılarak Yapay Bağışıklık Sistemlerinin bugüne kadar doküman sınıflandırmada sıkça kullanılan sınıflandırıcılar olan K-En Yakın Komşuluk, Naive Bayes, Destek Vektör Makinesi ve Rastgele Orman ile sınıflandırma performansları karşılaştırılmıştır.

Çalışmada kullanılan özellik vektörleri literatürde bugüne kadar kullanılmış ve başarısı kanıtlanmış özelliklerden oluşturulmuştur. Bunlar istatistiksel özellikler, kelime zenginliğine dayalı özellikler, dilbilgisi özellikleri, işlevsel kelimeler, karakter n-gram'ları ile kelime kök ve gövdeleridir. Bunlara ek olarak iki farklı boyut indirgeme yöntemiyle (CFS ve YTU) oluşturulan bu özellik vektörlerinden yeni özellik vektörleri türetilmiş ve bunların da sınıflandırma performansına etkileri gözlenmiştir.

Cinsiyet belirleme çalışmasında en yüksek başarı %99.71 ile V_{3G_YTU} özellik vektörü kullanılarak Yapay Bağışıklık Sistemi sınıflandırıcıları olan AIRS1 ve AIRS2Parallel ile alınmıştır. Karakter n-gram özelliklerinin, YTU yöntemiyle boyut indirgendikten sonra elde edilen özellik vektörleri V_{3G_YTU} ve V_{2G_YTU} , cinsiyet belirleme çalışmalarında tüm yöntemler için en başarılı sonuçları vermiştir. Bunların içinde en yüksek olanın Yapay Bağışıklık Sistemi algoritmaları olması bu özellik vektörleriyle Yapay Bağışıklık Sistemlerinin Türkçe dokümanların yazarlarının cinsiyetlerini belirlemede kullanılabileceğini göstermiştir.

Dokümanın türünü belirleme çalışmasında ise en yüksek başarı %100 ile V_{3G_YTU} özellik vektörü kullanılarak Naive Bayes sınıflandırıcısı ile alınsa da bu özellik vektörüyle %99.6'lık bir başarı performansı veren AIRS2 sınıflandırıcısı yine etkili bir sınıflandırma başarısı sağlamıştır. Doküman türü belirlemede kelime kök ve gövdelerinden oluşan özellik vektörlerinin, YTU ile boyut indirgendikten sonra elde edilen özellik vektörleri de tüm sınıflandırıcılar ile yüksek başarı oranları vermiş olup, Yapay Bağışıklık Sistemi sınıflandırıcılarından AIRS2Parallel $V_{KÖK_YTU}$ için %99.6, AIRS2 ise $V_{GÖVDE_YTU}$ için

%99.6'lık başarı performansı ile en yüksek başarıyı veren sınıflandırıcılardan olmuştur. Doküman türü belirleme çalışmalarında Yapay Bağışıklık Sistemi algoritmalarının, hem karakter n-gram'larının, hem de kelime kök ve gövdelerinden oluşmuş özellik vektörlerinin YTU ile boyut indirgenmesiyle oluşturulmuş özellik vektörleriyle birlikte kullanıldığında çok başarılı sonuçlar verdiği gözlenmiştir.

Doküman yazarını tanıma amaçlı yapılan çalışma ise iki ana bölüme ayrılmıştır. Birinci bölümde aynı türde yazan yazarları tanıma üzerinde durulmuştur. Bu bölüm de kendi içinde 10 ve 8 yazardan oluşan iki alt bölümde incelenmiştir. 10 yazardan oluşan veri kümesinde en yüksek başarı %99.71 ile V_{3G_YTU} özellik vektörü kullanılarak AIRS1, AIRS2, AIRS2Parallel ve KNN sınıflandırıcıları ile alınmıştır. YBS sınıflandırıcılarının tümü V_{3G_YTU} özellik vektörüyle %99.71 başarı oranı vererek çok iyi bir performans göstermiştir. 8 yazardan oluşan veri kümesinde ise en yüksek başarı %99.64 ile V_{3G_YTU} özellik vektörü kullanılarak AIRS2, AIRS2Parallel, NB ve RO sınıflandırıcıları ile alınmıştır. Bu oranlara da bakarak Yapay Bağışıklık Sistemi sınıflandırıcılarının, yazar tanıma için de karakter n-gram'larından oluşmuş özellik vektörlerinin YTU ile boyut indirgenmesiyle oluşturulmuş özellik vektörleriyle birlikte kullanıldığında çok başarılı sonuçlar verdiği gözlenmiştir. Yazar sayısı azaldıkça ortalama başarı performanslarının arttığı gözlenmiştir. Örneğin tüm özellik vektörleri dikkate alındığında 10 yazar için AIRS2Parallel sınıflandırıcısı ortalama %57.88 başarı göstermişken, aynı sınıflandırıcı 8 yazar için %63.28'lik başarı sağlamıştır.

Yazar tanıma çalışmasının ikinci bölümünde ise farklı türlerde yazan yazarları tanıma üzerinde durulmuştur. Bu bölüm de kendi içinde 20,18,16,14 ve 12 yazardan oluşan beş alt bölümde incelenmiştir. 20 yazar için en yüksek başarı %99.57 ile V_{3G_YTU} özellik vektörü kullanılarak KNN sınıflandırıcısı ile alınmıştır. YBS sınıflandırıcılarından AIRS2Parallel V_{3G_YTU} özellik vektörüyle %98.43 başarı oranı vermiştir. 18 yazar için en yüksek başarı %99.68 ile V_{3G_YTU} özellik vektörü kullanılarak KNN sınıflandırıcısı ile alınmıştır. YBS sınıflandırıcılarından AIRS2Parallel, V_{3G_YTU} özellik vektörüyle %98.89 başarı oranı vermiştir. 16 yazar için en yüksek başarı %99.82 ile V_{3G_YTU} özellik vektörü kullanılarak KNN sınıflandırıcısı ile alınmıştır. YBS sınıflandırıcılarından AIRS2Parallel, V_{3G_YTU} özellik vektörüyle %99.29 başarı vermiştir. 14 yazar için en yüksek başarı %100 ile V_{3G_YTU} özellik vektörü kullanılarak AIRS1, KNN ve DVM sınıflandırıcıları ile alınmıştır. 12 yazar için ise en yüksek başarı %100 ile V_{3G_YTU} özellik vektörü kullanılarak AIRS2,

KNN ve DVM sınıflandırıcıları ile alınmıştır. Aynı türde yazan yazarlarda olduğu gibi, farklı türde yazan yazarları tanıma çalışmasında da sınıflandırıcıların yazar sayısı azaldıkça tüm özellik vektörleri alınarak elde edilen ortalama başarı oranlarında artış olduğu gözlenmiştir. Diğer bölümlerde de olduğu gibi yazar tanıma çalışmalarında da Yapay Bağışıklık Sistemi sınıflandırıcılarının karakter n-gram'larından oluşmuş özellik vektörlerinin YTU ile boyut indirgenmesiyle oluşturulmuş özellik vektörleriyle birlikte kullanıldığında çok başarılı sonuçlar verdiği gözlenmiştir.

Doküman türü, doküman yazarı ve doküman yazarının cinsiyetini belirlemede CFS gibi bir özellik indirgeme yöntemi veya YTU gibi bir boyut azaltıcı kullanıldığında sınıflandırma yöntemlerinin başarılarının arttığı görülmüştür. Türkçe metinler üzerinde en etkili sınıflandırma performansları karakter n-gram'ları, kelime kökleri ve kelime gövdelerinden oluşturulmuş özellik vektörleriyle elde edilmiştir. Buna karşın kelime zenginliğine dayalı özellik vektörüyle dilbilgisel özelliklerden oluşan özellik vektörünün sınıflandırma performanslarının diğer vektörlerden daha düşük olduğu ortaya çıkmıştır. İşlevsel kelimelerden oluşan özellik vektörünün performansı düşük olmakla birlikte bu vektörden boyut indirgemeyle oluşturulan yeni vektörler daha iyi sonuç vermektedir. İstatistiksel özelliklerden oluşan özellik vektörünün de başarısı boyut indirgemeyle artmakla birlikte boyut indirgeme işlemi gerçekleştirilmeyen özellik vektörleri içerisinde en iyi performansı da istatistiksel özelliklerden oluşan özellik vektörü vermektedir.

Özetle bu çalışmada Yapay Bağışıklık Sistemi sınıflandırıcılarının, karakter n-gram'ları, kelime kök veya gövdeleri gibi uygun özelliklerin YTU gibi bir boyut indirgeme yöntemiyle oluşturulmuş özellik vektörleriyle birlikte kullanıldığında Türkçe dokümanlar üzerinde doküman türü, doküman yazarı ve doküman yazarının cinsiyetini belirlemede kullanılabileceği gösterilmiştir.

KAYNAKLAR

- Aha, D. ve Kibler, D., (1991), "Instance-based learning algorithms", *Machine Learning*, vol.6, pp. 37-66.
- Alataş, B., Aydın, İ., Akın, E., (2005), "Asenkron Motorların Hata Teşhisinde Yapay Bağışıklık Sistemi Yaklaşımı.", II. Mühendislik Bilimleri Genç Araştırmacılar Kongresi, MBGAK 2005.
- Amasyalı, M.F., Diri, B., (2006), "Automatic Turkish Text Categorization in Terms of Author, Genre and Gender.", 11th International Conference on Applications of Natural Language to Information Systems, NLDB 2006, Austria, LNCS Volume 3999 Springer.
- Amasyalı, M.F., Diri, B., (2003), "Automatic Author Detection for Turkish Texts.", pp. 138-141, ICANN/ICONIP 2003, İstanbul.
- Bıyıklioğlu, O., (2004), Use of Artificial Immune Systems for Network Intrusion Detection, M.Sc. Thesis.
- Blanz, V., Scholkopf B., Bulthoff H., Burges C., Vapnik V., Vetter T., (1996), "Comparison of View-based Object Recognition Algorithms Using Realistic 3d Models." In C. von der Malsburg, W. von Seelen, J. C. Vorbrüggen, and B. Sendhoff, editors, *Artificial Neural Networks — ICANN'96*, pages 251 – 256, Berlin, 1996. Springer Lecture Notes in Computer Science, Vol. 1112.
- Breiman, L., (2001), "Random Forests", *Machine Learning* 45 (1), 5-32.
- Burnet, (1978), "Clonal Selection and After."
- Clark, J., Hannon, C., (2007), "A Classifier System for Author Recognition Using Synonym-Based Features.", Sixth Mexican International Conference on Artificial Intelligence, November 2007. Aguascalientes, Mexico.
- Cortes, C., Vapnik, V. (1995) "Support Vector Networks." *Machine Learning*, 20:273–297, 1995.
- Cunningham, P., Delany, S.J., (2007), "k-Nearest Neighborhood Classifiers."
- Çatal, Ç., Diri, B., (2007), "Application and Benchmarking of Artificial Immune Systems to Classify Fault-Prone Moduls for Software Development Projects.", IADIS International Conference Applied Computing 2007.
- Dasgupta, D., (1997), "Artificial Neural Networks and Artificial Immune Systems: Similarities and Differences", *Proc. of the IEEE SMC*, 1, pp. 873-878.
- De Castro, L.N. and Timmis J., (2002), "A Novel Approach to Pattern Recognition". In L. Alonso, J. Corchado, and C. Fyfe, editors, *Artificial Neural Networks in pattern Recognition*, University of Paisley: 67-84.
- De Castro L. N. & Von Zuben F. J. (1999), "Artificial Immune Systems: Part I – Basic Theory And Applications", RT DCA 01/99.
- de Vel, O., Anderson, A., Corney, M., Mohay, G., (2002), "Language and Gender Author Cohort Analysis of E-mail for Computer Forensics." *Proceedings Digital Forensic Research Workshop*.
- Diri, B., Amasyalı, M.F., (2003), "Automatic Author Detection for Turkish Texts.", 13th International Conference on Neural Information Processing, Turkey.

- Döyen, A., (2004), Akış Tipi Çizelgeleme Problemlerinin Yapay Bağışıklık Sistemleri ile Çözümü ve Parametre Optimizasyonu, Yüksek Lisans Tezi, Konya.
- Engin, O., Döyen, A., (2004), "Artificial Immune Systems and Applications in Industrial Areas.", G.U. Journal of Science 17(1): 71-84, 2004.
- Farmer, J., Packard, N., Perelson, A., (1986), "The Immune System, Adaptation and Machine Learning", Physica D, 22, 187-204.
- Forrest, S., Perelson, A.S., Allen, L., Cherekuri, R., (1994), "Self-non-self Discrimination in a Computer." In Proceedings of the IEEE Computer Society symposium on security and privacy held in Oakland, California, May 16-18, 1994, 202-212.
- Hall, M.A., (1998), "Correlation-based Feature Selection for Machine Learning." PhD thesis, Department of Computer Science, University of Waikato, New Zealand.
- Hall, M., Smith, L. A., "Practical feature Subset Selection for Machine Learning", Proceedings of the Australian Computer Science Conference (University of Western Australia), February 1996.
- Ho, T.K., (1995). "Random Decision Forest". Proc. of the 3rd Int'l Conf. on Document Analysis and Recognition, Montreal, Canada, August 14-18, 1995, 278-282.
- Hofmeyr, S.A., Forrest, S., (1999), "Immunity by Design: An Artificial Immune System." Proceedings of the Genetic and Evolutionary Computation Conference (GECCO), San Francisco, CA, pp.1289-1296.
- Hunt, J. ve Cooke, D.E. (1996), "Learning Using an Artificial Immune System, Journal of Network and Computer Applications", 19, pp. 189-212.
- Ishida, Y., Okamoto, T. (1999), "Multiagent Approach Against Computer Virus: An Immunity-Based System", Proc. of the AROB'99, pp. 69-72.
- Joachims, T., (1998), "Text Categorization with Support Vector Machines: Learning with Many Relevant Features", 10th European Conference on Machine Learning, pages 137-142, Heidelberg, Germany.
- Kaban, Z., Diri, B., (2008) "Yapay Bağışıklık Sistemleri ile Türkçe Metinlerde Tür ve Yazar Tanıma", IEEE SIU 2008 16. Sinyal İşleme, İletişim ve Uygulamaları Kurultayı.
- Koppel, M., Argamon, S., Shimoni, A., (2002), "Automatically Categorizing Written Texts by Author Gender", Lit Linguist Computing, 17: 401-412.
- Koppel, M., Schler, J., (2003), "Exploiting Stylistic Idiosyncrasies for Authorship Attribution." In Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, Acapulco, Mexico.
- Lewis, D. D., Yang, Y., Rose, T., and Li, F. RCV1: "A New Benchmark Collection for Text Categorization Research", Journal of Machine Learning Research, 5:361-397, 2004.
- Lowd, D., Domingos, P., Naive Bayes Models for Probability Estimation.
- Mc Callum, A., Nigam, K., "A Comparison of Event Models for Naive Bayes Text Classification".
- Morales, R.M., Pineda, L., Gomez, M., Rosso, P., (2004), "Authorship Attribution Using Word Sequences", Progress in Pattern Recognition, Image Analysis and Applications, pp. 844-853.

- Osuna, E., Girosi, F., (1998), "Reducing the Run-time Complexity of Support Vector Machines." In International Conference on Pattern Recognition (submitted).
- Peng, F., Schuurmans, D., (2004), "Combining Naive Bayes and n-Gram Language Models for Text Classification", ECIR 335-350.
- Perelson, A.S., Oster, G.F., (1979), "Theoretical Studies of Clonal Selection: Minimal Antibody Repertoire Size and Reliability of Self-Nonself Discrimination", J. theor.Biol, 81, pp. 645-670.
- Polat, K., (2004), Özellik Seçme ile Yapay Bağışıklık Tanıma Sistemi Kullanılarak Medikal Teşhise Gidiş, Yüksek Lisans Tezi, Konya.
- Schmidt, M., (1996), "Identifying Speaker with Support Vector Networks." In Interface '96 Proceedings, Sydney, 1996.
- Stamatatos, E., Fakotakis, N., Kokkinakis, G., (1999), "Automatic Authorship Attribution". EACL, 1999.
- Yıldız, H. K., Gençtay, M., Usta N., Diri, B. ve Amasyalı ,M.F., (2007), "Metin Sınıflandırmada Yeni Özellik Çıkarımı", IEEE SIU 2007 15. Sinyal İşleme, İletişim ve Uygulamaları Kurultayı.
- Zhang, H., The Optimality of Naive Bayes.
- Zhao, Y. & Zobel, J. (2005), "Effective Authorship Attribution Using Function Word", in 'Proc. 2nd AIRS Asian Information Retrieval Symposium', Springer, pp. 174-190.

İNTERNET KAYNAKLARI

- [1] www.cs.waikato.ac.nz/ml/weka/
- [2] www.hurriyet.com.tr
- [3] www.milliyet.com.tr
- [4] www.sabah.com.tr
- [5] www.tdk.org.tr
- [6] <http://code.google.com/p/zemberek/>
- [7] <http://en.wikipedia.org/>
- [8] <http://trec.nist.gov/data/reuters/reuters.html>
- [9] <http://tartarus.org/~martin/PorterStemmer/>

ÖZGEÇMİŞ

Doğum tarihi 29.08.1982

Doğum yeri İzmir

Lise 1996–2000

İzmir Özel Türk Fen Lisesi

Lisans 2000–2004

İzmir Yüksek Teknoloji Enstitüsü Mühendislik
Fakültesi.Bilgisayar Mühendisliği Bölümü

Yüksek Lisans 2005–2008

Yıldız Teknik Üniversitesi Fen Bilimleri
Enstitüsü Bilgisayar Mühendisliği

Çalıştığı kurumlar

2005-Devam ediyor ARGELA Yazılım ve Bilişim Tek., Yazılım
Mühendisi