KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

WEB TABANLI METİNLERDE YAZARIN ANADİLİNİ TANIMLAMA

YÜKSEK LİSANS TEZİ

Bilgisayar Müh. Parham MOHAMMADALIPOUR TOFIGHI

AĞUSTOS 2012
TRABZON

KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

WEB TABANLI METİNLERDE YAZARIN ANADİLİNİ TANIMLAMA

Parham MOHAMMADALIPOUR TOFIGHI

Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsünce
"BİLGİSAYAR YÜKSEK MÜHENDİSİ"
Unvanı Verilmesi İçin Kabul Edilen Tezdir.

Tezin Enstitüye Verildiği Tarih : 17.07.2012
Tezin Savunma Tarihi : 01.08.2012

Tez Danışmanı : Doç. Dr. Cemal KÖSE

Trabzon 2012

Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsü
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
Parham MOHAMMAD ALI POUR TOFIGHI

WEB TABANLI METİNLERDE YAZARIN ANADİLİNİ TANIMLAMA

başlıklı bu çalışma, Enstitü Yönetim Kurulunun 24 / 07 / 2012 gün ve 1467 sayılı kararıyla oluşturulan jüri tarafından yapılan sınavda
YÜKSEK LİSANS TEZİ
olarak kabul edilmiştir.

Jüri Üyeleri

Başkan : Prof. Dr. Sefa AKPINAR …...……

Üye : Doç. Dr. Cemal KÖSE …...……

Üye : Yrd. Doç. Dr. Bekir DİZDAROĞLU ……...……

Prof. Dr. Sadettin KORKMAZ

Enstitü Müdürü

ÖNSÖZ

Bir yabancı dilde yazı yazma her zaman zor bir uğraş olmuştur. Hatta yıllarca mekaniği ve gramer kuralları çalıştıktan sonra bile anadil yazarlarının yazılarında bulunan aynı doğal akış ile yazmak, yabancı yazarlar için çok zor bir iştir. Ayrıca, yazarlar yabancı bir dilde yazarken, genellikle yanlış ya da ilk dillerinden etkilendikleri bir kalıp veya doku olup olmadığı konusunda hata yaparlar.

Yazarlık Analizi; yazarlık hakkındaki kararlarına dikkat çekmek için metnin özelliklerini inceleme sürecidir. Yazarlık Analizinin kökleri stylometry'e dayanmaktadır. Stylometry'nin belirtici niteliği, Yazarlık Analizinin en etkin ayrımcısı olan yazım tarzı göstergeleridir. Bu çalışma, bir yazarın Anadil Tanımlaması için Stylometry özellikleri kullanmaktadır. Bu alanda ufak bazı çalışmalar yapılmış olmasına rağmen yazarların anadillerini bulma nispeten yeni bir konudur.

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

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

Parham MOHAMMADALIPOUR TOFIGHI

Trabzon 2012

TEZ BEYANNAMESİ

Yüksek Lisans Tezi olarak sunduğum “Web Tabanli Metinlerde Yazarın Anadilini Tanımlama” başlıklı bu çalışmayı baştan sona kadar danışmanım Doç. Dr. Cemal KÖSE ‘nin sorumluluğunda tamamladığımı, verileri/örnekleri kendim topladığımı, deneyleri/analizleri ilgili laboratuarlarda yaptığımı/yaptırdığımı, başka kaynaklardan aldığım bilgileri metinde ve kaynakçada eksiksiz olarak gösterdiğimi, çalışma sürecinde bilimsel araştırma ve etik kurallara uygun olarak davrandığımı ve aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul ettiğimi beyan ederim. 17/07/2012

Parham MOHAMMADALIPOUR TOFIGHI

İÇİNDEKİLER

Sayfa No

Yüksek Lisans

ÖZET

WEB TABANLI METİNLERDE YAZARIN ANADİLİNİ TANIMLAMA

Parham MOHAMMADALIPOUR TOFIGHI

Karadeniz Teknik Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim dalı
Danışman: Doç. Dr. Cemal KÖSE
2012, 56 Sayfa, 4 Ek Sayfa

İnternet teknolojileri ve uygulamalarının hızlı gelişimine rağmen metinsel sayfalar hala en yaygın internet ortamlarıdır. Bunun en önemli örnekleri olarak, çoğunlukla metin tabanlı olan Twitter, Facebook, vb. sosyal ağ uygulamaları ile haber grupları, e-posta, blog, vb. gibi web uygulamaları verilebilir. Dolayısıyla, bu çalışmada, Metinsel veri Madenciliği ve Belge Sınıflandırma çerçevesinde, yazarların ana dillerini belirlemeye bir giriş çalışması yapılmıştır. Özellikle, birçok internet uygulamasında olduğu gibi İngilizcede yazılmış bir metinin yazarının ana dilini belirlemek için bir sistem geliştirilmiştir.

Bu çalışmada, stylometry ve geleneksel makine öğrenmesi gibi alanlardan teknikler kullanarak bir yazarın ana dilinin belirlenmesi için bir araç geliştirilmiştir. Burada, bir yazarın tarzı, metinden çeşitli stylometric özelliklerin ölçümleri yapılarak bir örüntü (pattern) tanıma işlemine dönüştürülmektedir.

Bir stilistik metnin özelliklerini dört türde (Sözcüksel, Sözdizimsel, Yapısal ve İçeriğe özgü özellikleri) ve makine öğrenme algoritmasını da üç türde (destek vektör makinesi, karar ağacı ve Naïve Bayes) ele alınmış ve daha sonra amaçlanan özelliklere dayanarak yazarın Anadilinin Tanımlaması işlemi gerçeklenmiştir. Yapılan çalışmada, dört farklı anadilden yazarlar (Türkçe, Almanca, Farsçanın ve İngilizce) tarafından yazılan çevrimiçi haber sayfalarından oluşan bir veritabanı kullanılmıştır.

Anahtar Kelimeler: Metin Madenciliği, Sınıflandırma Teknikleri, Anadili Tanımlama, Web Tabanlı Metinler.

Master Thesis

SUMMARY

AUTHOR'S NATIVE LANGUAGE IDENTIFICATION IN WEB MEDIUMS

Parham MOHAMMADALIPOUR TOFIGHI

Karadeniz Technical University
The Graduate School of Science
Computer Engineering Graduate Program
Supervisor: Assoc. Prof. Dr. Cemal KÖSE
2012, 56 Pages, 4 Pages Appendix

In the domain of Text Mining and Document Classification, an introduction into the field of Authorship Attribution is presented. On the other hand, with the rapid growth of Internet technologies and applications, text is still the most common Internet medium. Examples of this include social networking applications such as Twitter, Facebook, etc. and web applications such as newsgroups, email, blog, etc. are also mostly text based. We developed a framework to determine an anonymous author's native language for short length and multi-genre writing in English such as the ones found in many Internet applications.

This thesis describes the development of such a tool using techniques from the fields of stylometry and traditional machine learning techniques. An author's style can be reduced to a pattern by making measurements of various stylometric features from the text. In this framework, four types of stylistic text features (Lexical, Syntactic, Structural, and Content-Specific Features) are extracted and two machine learning algorithms (Decision Tree, Support Vector Machine and Naïve Bayesian) are designed for author's native language identification based on the proposed features. For this research, we used four different collections of writings online news messages by speakers of four different nationalities: native English as well as speakers of Turkish, German, and Persian.

Key Words: Text Mining, Classification Techniques, Native Language Identification, Stylometry, Web-based Texts.

ŞEKİLLER DİZİNİ

Sayfa No

TABLOLAR DİZİNİ

Sayfa No

SEMBOLLER DİZİNİ

DVM : Destek Vektör Makinaları

SVM : Support Vector Machines

KA : Karar Ağaçları

PD : Pozitif Doğru

NY : Negatif Yanlış

PY : Pozitif Yanlış

ND : Negatif Doğru

CSV : Comma Separated Values

SMO : Sequential Minimal Optimization

KNN : K Nearest Neighbor

1. GİRİŞ

İnternet teknolojileri ve uygulamalarının hızlı gelişimi ve çoğalması zaman ve mekan içinde bilgileri paylaşmak için yeni bir yol oluşturmuştur. Online sosyal ağlar (Twitter, Facebook gibi), e-ticaret kullanımı (e-Bay, Amazon gibi), haber grupları vb. daha fazla önem kazanmaktadır.

İnternet üzerinde kaynak paylaşımı için fiili iletişim, basit bilgi alışverişi ve e-ticaret faaliyetleri ile çeşitlenen geniş çaplı aktiviteler geliştirilmiştir. Özellikle, internet sohbet odaları, e-posta, internet haber grupları ve Web siteleri gibi Web tabanlı kanallar üzerinde bilgi dağıtmak için büyük oranda çevrimiçi mesajlar kullanılmaktadır. İnternet genellikle el yazımı mektuplar ve denemeler gibi geleneksel olarak uzun yazı formların daha çok iletişimin kısa formlarını kullanmıştır. Maalesef bu gelişme aynı zamanda gereksiz ileti, saldırgan mesajlar ve yanlış bilgi alma gibi uygunsuz bilgi dağıtımı için farklı yanlış kullanımlara neden olur. Ayrıca suçlular, hakkı gerektiren malzemeler, çocuk pornografileri, çalıntı eşyalar vb. yasa dışı materyalleri dağıtmak için Web tabanlı uygulamaları kullanmışlardır.

Terör örgütleri de kendi büyük iletişim istasyonlarından biri olarak online sistemleri kullanıyorlar. Bu eğilimler, küresel elektronik ağlar üzerinden yasadışı bilgisayar merkezli faaliyetler olarak "siber suç" gibi yeni bir kavram oluşturdu [1]. Siber suç, büyüyen bir suç alanıdır ve birçok hükümet tarafından bir tercih olarak kabul edilmiştir. Online metinlerin önemli bir özelliği anonim olmalarıdır [2], [3], [28]. İnsanlar genellikle ad, yaş, cinsiyet ve adres gibi gerçek kimlik bilgilerini vermeye ihtiyaç duymayabilir. Siber alanda birçok yanlış kullanım veya suç göndericiler kendi gerçek kimliklerinin tespit edilmesinden sakındıkları için gizlemeye çalışacaklardır. Bu nedenle, çevrimiçi metinlerin anonimliği, siber uzayda takip için eşsiz bir mücadele gerektirecektir. Böyle kısa metinlerin doğru yazarlık profilini oluşturma bu suçluları yargılamada yardımcı olur.

Yazarlık profili oluşturması; bir yazarın cinsiyetini, anadilini, yaşını veya diğer niteliklerini belirlemeyle ilgilidir. Bu araştırmada özellikle, yazarın anadiliyle yazılan metindeki dilin onun anadili olmadığı düşünülerek ilgilenilir. Bir yazarın anadilini belirleme Yazarlık Özelliği (Authorship Attribution) sorununun bir türüdür. Listelenen yazarlar arasından bir yazarı tespit etmek yerine, belirli bir anadili paylaşan bu yazarlardan

bir grubu belirlemeyi umuyoruz. Yerli olmayan yazarlar için İngilizce metin yazma her zaman zor bir uğraş olmuştur. Hatta yıllarca gramer kuralları çalıştıktan sonra bile bu yerli yazarların yazılarında bulunan aynı doğal akış ile yazmak yerli olmayan yazarlar için çok zor bir iştir. Ayrıca, yerli olmayan konuşucular yazarken yanlış olduğu ya da ilk dillerinden etkilendikleri bir kalıp (pattern) olup olmadığı konusunda hata yaparlar. Örneğin, bazı dillerde, "a" veya "the" gibi harfi tarifleri kullanma anlayışı yoktur, bu yüzden yerli olmayan yazarlar için bunları yanlış veya yanlış yerde kullanması muhtemeldir. Yerli olmayan yazarların yazarken tek bir kalıpa odaklanarak yazmaları, onları farklı dilleri konuşan diğer insanlardan ayırdığına dair bir model oluşturabiliriz.

Eğer teorik olarak tüm diller için bu modeller oluşturabiliriz, biz onların İngilizcedeki yazım tarzlarına dayanarak yazarların anadillerini belirleyebileceğiz. Bildiğimiz gibi herhangi bir alanda yazarın eserine dayanarak onun anadilini bulan bir sistem türü, önemli bir uygulama olmamasına rağmen, fakat az önce bahsettiğimiz gibi dünya her zamankinden daha fazla birbirine bağlıdır ve bilgi paylaşımı daha da kolaylaşmıştır. Tehdit edici bir mesajın yazarının anadilinin keşfedilebilmesi, böyle tehditten sorumlu olan kişilerin yakalaması için önemli bir araçtır. Örneğin, bir mesajın tanımlanması muhtemel bir terörist faaliyetini durdurursa, tehdit hakkında muhtemel tehdidin arkasında kimin olabileceği gibi daha çok bilgiyi öğrenmek için otomatik dil bulucusunu kullanabilirler.

Halbuki Yazarların Anadilinin otomatik olarak bulunabilmesi için tam bir çözüm geliştirmek çok zordur. Böylece, böyle bilgi varlıkları ile hiçbir önemli deneme olmadığı için bu tezin amacı web tabanlı çevrimiçi metinlerin stylometrik yaklaşımını tanıtmaktır.

Yazarların anadillerini bulma nispeten yeni bir konudur. Bu alanda ufak bir araştırma yapılmıştır, ancak yukarıda açıklanan görevleri yerine getirmekten uzaktır. Bu araştırmada, üç soruyu cevaplamak istiyoruz:

1. Yazarın Anadilini tanımlama Web tabanlı metinlere uygulanabilir mi?
2. Yazarın Anadilinin belirlenmesi için kuvvetli yazım tarzı özellik seti nedir?
3. Hangi sınıflandırma teknikleri çevrimiçi metinlerde yazarın anadilinin tespiti için etkilidir?

2. GENEL BİLGİLLER

2.1. Metin Madenciliği Alanındaki Yazarlık Analizi

Bir metnin yazarının anadilinin otomatik olarak nasıl belirleneceği hakkında herhangi bir teknik ayrıntıya girmeden önce Metin Madenciliği alanında Yazarlık Özelliği (Atıf) konusunu, açık bir şekilde ele almalıyız.

Yazarın Anadil Tanımlama uygulamasının asıl amacı, otomatik olarak metinden özellikleri ayıklamaktır, bu nedenle Metin Madenciliği bu işlemin temel bir dalıdır. Bu, büyük miktardaki metinleri analiz ederek yararlı bilgiler çıkarmaya çalışma ve bir sözcüğün alışılagelmiş yazılış ve söyleniş kalıplarını ortaya çıkarma anlamına gelir [30]. Metin Madenciliği alanında önemli bir aşama olan Metin Sınıflandırma aynı zamanda Doküman Sınıflandırma olarak adlandırılır.

Metin sınıflandırma, internetin ve Worldwide Web'den bilgi erişimi için arama motorlarının ortaya çıkmasıyla yaygın bir araştırma sorunu haline gelmiştir. Yakın zamanda, web sayfaları, yürütülen bir araştırmada ilgili dokümanların elde edilebilmesi için bir konuyu tespit etmek zorundadır. Metin sınıflandırma, bir dizi metin belgesini içeriklerine ya da başlığına göre kategorize etmeye çalışır. Metin sınıflandırma için birçok metot önerilmiştir ve bunlardan en çok kullanılan bir belgedeki kelimelerin sırasının o belgenin içeriğinin ya da konusunun belirlenmesi için daha az öneme sahip olduğunu tespit eden "Bag of Words" yaklaşımıdır. Bu yaklaşımda, bir korpusun her dokümanındaki her bir kelime ayırt edici bir özellik olarak belirlenir. Her bir dokümanın sınıflandırılması için bir kelime vektör özelliği temsili, kelimelerin frekanslarını dayanarak oluşturulur.

Metin Sınıflandırmada, dokümanın önceden tanımlanmış kategori kümesine göre etiketlenir. Bir kategorinin etiket değeri, her bir çift için D dokümanların kümesi ve C önceden tanımlanmış bir kategorilerin (sınıflar) kümesi yani $<d_j, c_i> \in D \times C$ olarak belirlenir. Metin Sınıflandırmanın çok sayıda uygulaması vardır, Bunlardan en önemlileri: doküman indeksleme ve filtreleme, web sayfaları ve web arama motorlarının hiyerarşik kategorizasyonudur [30] .

Yazarın Anadil Tanımlaması, Yazarlık Analizi gibi bir sınıflandırma sorunudur. Bir metin dosyasından başlayarak, bu iş otomatik olarak metnin yazarı hakkındaki bilgileri ortaya çıkarmaktır. Araştırma veya tartışmaya göre önceden tanımlanmış kategoriler, yazarların yerli veya ilk dilleridir. Yazarlık Analizi için çok sık kullanılan bir teknik, metin tarzı veya stylometry incelenmesidir. Araştırmacılar, tüm yazarların bilinçli kontrolü dışında olan belirli üslup özelliklere sahip olduğunu farz ederler.

2.2. Stylometry ve Yazarlık Analizi

Yazarlık Analizinin kökleri stylometrydedir; Stylometric'in belirtici niteliği, Yazarlık Analizinin en etkin ayrımcısı olan yazım tarzı göstergelerdir. Stylometry alanı, yazınsal tarzın bir gelişimidir ve istatistiksel analizi olarak tanımlanabilir [33].

Bu, bir yazarın metin içi kelime kullanımı, cümle karmaşıklığı ve ifade tarzı gibi ayırt edici yazma alışkanlıkları sergilediğine dair temel varsayımdır. Başka bir varsayım, bu alışkanlıkların bilinçsiz ve kökleşmiş olduğudur, yani birisi kendi tarzını gizlemek için bilinçli bir çaba harcasa da bunu başarması zor olacaktır.

Stylometry bir yazarın üslup özelliklerini belirlemeye çalışır ve bu özellikleri ölçmek için istatistiksel metotlar belirler böylece iki veya daha fazla metin parçaları arasındaki benzerlikler analiz edilebilir. Bu varsayımlar, bu tezde yürütülen araştırma için temel çekirdek olarak kabul edilir. Biz, yazarın metinindeki özelliklerinin profilini belirlenmeye (belirlemeye) çalışacağız. Bu profil sayesinde yazarın cinsiyetini, anadilini, eğitim durumunu, kültürel özgeçmişini ve vb. ortaya çıkarabiliriz.

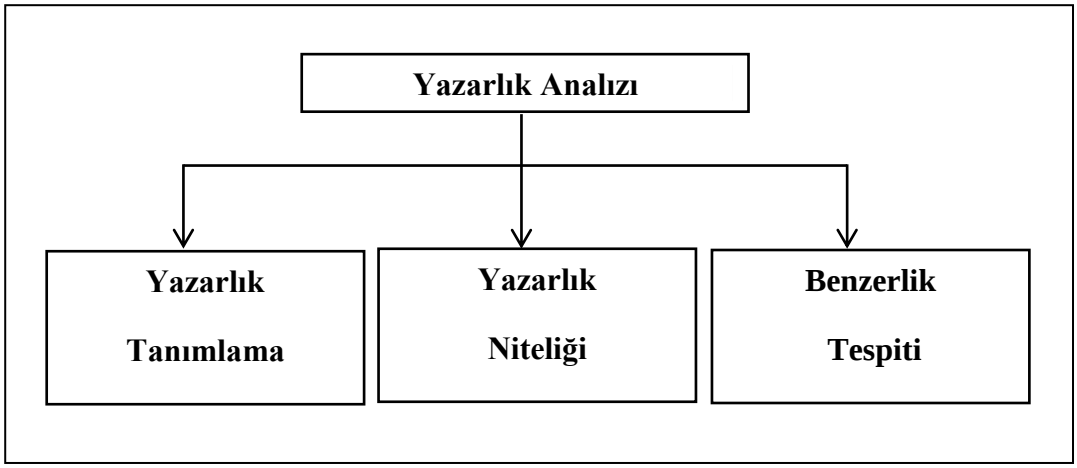
2.3. Yazarlık Analizi

Yazarlık Analizi, yazarlık hakkındaki kararlarına dikkat çekmek için metnin özelliklerini inceleme sürecidir. Kökleri stylometry olarak adlandırılan, yazınsal tarzın istatistiksel analizine değinen dilsel bir araştırma alanına dayanmaktadır.

Makine öğrenme teknikleri gibi daha gelişmiş teknikler olarak bu alanda uygulanmıştır, genellikle bu araştırma alanı yazarlık analizi olarak tanınmıştır. Yazarlık

analizi, edebiyatta yazarların tanımlanması, program kodu ve suç davalarında adli analiz için birkaç uygulama alanında kullanılmıştır. Yazarlık analizi uygulaması, yayınlanan makalelerin ve edebi eserlerin yazarlık atıflandırılmasında geniş bir şekilde incelenmiştir. Tanınmış çalışmalar, Shakespeare'in tartışmalı eserlerinin (Merriam [22]) ve Federalist makalelerin (Mosteller and Wallace [24], Holmes and Forsyth [14], Tweedie et al. [33]) atfedilmesini kapsar.

Yazarlık Analizine, Yazarlık Tanımlama, Yazarlık Niteliklerini (Karekterizasyonu) ve Benzerlik Tespiti gibi birçok farklı sorundan birkaçıyla başlanabilir. Bu problemler arasındaki ilişki Şekil 1'de gösterilmiştir.



Şekil 1. Yazarlık Analizi için Taksonomi

2.3.1. Yazarlık Tanımlama

Yazarlın Tanımlama (Authorship Identification) veya yazarlık teşhisi bir metin yazarını belirleme işi olarak tanımlanabilir. Bir metnin o yazar tarafından yazılmış olduğunu tespit etmek için bazı kanıtlara dayanır. Bu kanıtlar, aynı yazar tarafından oluşturulan diğer metin örnekleri olacaktır. Ayrıca bazı edebiyatlarda özellikle dilbilim araştırmacıları tarafından "Yazarlık Atfetme" (Authorship Attribution) olarak adlandırılır.

Yazarlık teşhisi/kimlik saptaması şöyle ifade edilebilir. Birkaç yazardan verilen bir dizi metninden birinin yazısının bilinmeyen bir parçasının tespit edilmesidir. İstatistiksel hipotez sınaması (test) veya bir sınıflandırma sorunu göz önünde bulundurulabilir. Bu

sınıflandırmanın özü, aynı kişi tarafından oluşturulan çok sayıda eser için nispeten sabit kalan bir dizi özelliklerin belirlenmesidir.

2.3.2. Yazarlık Niteliği

Yazarlık Niteliği (Authorship Characterization), bir yazarı diğerlerinden ayıran özellikleri özetler ve onun yazılarına dayalı olarak bir yazar profili oluşturur. Bu özelliklerin bazıları; cinsiyet, eğitim bilgisi ve kültürel özgeçmiş, anadili belirlemedir. Bu yeni araştırma yönü nispeten yazarlık-teşhiş/kimlik saptaması çalışmalarından ileri gelmektedir.

2.3.3. Benzerlik Tespiti

Benzerlik tespiti (Similarity Detection), çok sayıda eser örneği karşılaştırır ve gerçekten yazarı tespit etmeden onların tek bir yazar tarafından yazılıp yazılmadığını belirler. Bu kategorideki birçok çalışma, intihal (eser hırsızlığı) tespiti ile ilişkilidir.

İntihal, eserin bir kısmının veya tamamının, asıl yazarının izni olmaksızın kullanılmasıdır. İntihal tespiti, eserlerin iki örneği arasındaki benzerlikler incelenerek intihal faaliyetini belirleme çalışmasıdır. Benzerlik tespiti çeşitli yönlerden yazar tanımlamadan çok farklı olduğu için bu araştırmanın kapsamı dışındadır.

2.3.4. Özellikler Kümesi

Yazarlık Analizde önemli kaygılardan biri, yazarlar birbirinden ayırabilen, ölçülebilir özellikleri aramaktadır. Genişletilmiş özellik gruplarının kullanımı, yazım tarzı analizinin ölçeklenebilirliğini, büyük ayırt etme yeteneği sağlayarak daha geniş sınıf grupları arasında geliştirebilir. Araştırmamızın önemli bir bileşeni, özellik kümesi (Feature sets) olduğu gibi bu, yazarın anadilini belirleme performansı üzerinde de çok önemli etkiye sahiptir.

Holmes [8], yaklaşık 1000 yazma stili özelliğinin Yazarlık Analiz uygulamalarındaki kullanımlarını özetlemiştir. En iyi özellik grubu hakkında geniş bir kullanım alanı için bir fikir birliği yoktur. Ayrıca, genel olarak yazar analizi performansının seçilen özellikler kombinasyonuna ve analitik tekniklere bağlı olduğu kabul edilmektedir.

Web Tabanlı metinlerin kendine özgü nitelikleri için özel özellikler seçimi gereklidir. Bizim özellik grubumuz, dört çeşit özelliği kapsamaktadır: sözlüksel, sözdizimsel, yapısal, ve içeriğe özgü özellikler. Web tabanlı uygulamaların çoğunda otomatik sözcük kontrolü uygulandığı için İdiyosinkratik Özelliklerin yanlış yazım içermesi ve diğer kullanım türü olan anomaliler, bu çalışmada dikkate alınmamıştır.

2.4.1. Sözlüksel Özellikler

Sözcüksel özellikler (Lexical Features), aşağıdaki gibi karakter tabanlı ve kelime tabanlı özellikler olarak ayrılabilir:

2.4.1.1. Karakter Tabanlı Özellikler

Bu özellikler, frekansları veya boşluk oranlarını, rakam, büyük harf karakterler ve noktalamaları kapsayacaktır [8], [11]. Tablo 1' de gösterilir.

Tablo 1. Sözlüksel Özelliklerin Karakter Tabanlı Özellikler Listesi

Özellik Numarası	Özellik Açıklaması
1	Karakter Toplam Sayısı (C)
2	Harflerin sayısı (a-z)/C
3	Büyük karakter toplam sayısı /C
4	Dijital karakterlerin toplam sayısı /C
5	Boşluk karakterlerin toplam sayısı /C
6	Sekme boşluk karakterleri toplam sayısı /C
7-30	Özel karakter toplam sayısı (%, &, etc.) /C (24 Özellik)

2.4.1.2. Kelime Tabanlı Özellikler

Farklı kelimelerin toplam sayısı ve kelimelerin toplam sayısı hesaplanabilir ve özellikler olarak kullanılabilir. Bu özellik kümesi, Tablo 2' de gösterilir.

Tablo 2. Sözlüksel Özellikler Kelime Tabanlı Özellikler Listesi

Özellik Numarası	Özellik Açıklaması
31	Kelimelerin toplam sayısı (N)
32	Kelime başına ortalama uzunlukları (karakterlerle)
33	Kelime zenginliği (Toplam farklı kelimeler /N)
34	6 karakterden daha uzun Kelimeler sayısı /N
35	Kısa kelime toplam sayısı (1-3 karakterlerle)/N
36	Hapax legomena/N
37	Hapax dislegomena/N
38	Yule'nin K ölçütü
39	Simpson'nun D ölçütü
40	Sichel'nin S ölçütü
41	Honore'nin R ölçütü
42	Entropy ölçüsü
43	Kısaltma kelime sayısı /N
44-63	Kelime uzunluğu frekans dağılımı /N (20 özellik)

Sözcük Zenginliği:

Yazarlar, kullandıkları Kelime varlığının çeşitliliğiyle birbirlerinden farklılaşırlar. Temel varsayım, yazarın çoğu diğerleri tarafından da kullanılan belirli kelime hazinesine sahip olduğudur. Farklı kelime sayısının toplam oranı ve kelime sayısının toplamı, bizim sözcük zenginliğimizi (Vocabulary Richness) belirler [11]. Böylece farklı yazarların metinlerindeki Kelime varlığı zenginliğini karşılaştırabilirz. Bu her yazarın kendi Kelime Sözlüğünün Zenginliğine sahip olduğunu açıklayabilir.

Bu araştırmadabiz, isatistiksel dilbilimi analizinde yaygın olan diğer kelime sözlüğünün zenginliğini kullandık. Bu ölçü, Tablo 2'de gösterilmiştir. Kelime varlığı ölçülerinden herbirinin tanımı Ek A'de yer alır. Ayrıca, Yüksek frekanslı kelimelerde olduğu kadar düşük frekanslı kelimele de yazarları ayırabilir. Hapax Legomena, bir metinde sadece bir kez ortaya çıkan kelimelerdir. Bir metinde sadece bir kez yer alan çok sayıdaki kelimeyle karşılaştırıldığında, sık sık tekrar eden sadece birkaç kelime vardır. Hapax Dislegomena, metin içinde iki kez görülen kelimelerdir [33].

Kelime Uzunluğu Frekans Dağılımı:

Kelime uzunluklarının frekans dağılımı (Word length frequency distribution), bir belge üzerinden belirlenebilir ve her kelime uzunluğunun frekansı, bireysel bir özellik olarak kullanılabilir. Bu araştırmada, 1 ile 20 arasında kelime uzunlukları dikkate alınmıştır [11], [14]. Önceki çalışmalarda, Dil Tanımlama için dilsel bir özellik olarak, verilen kelimelerin uzunluklarının dağılımının kullanıldığı, kelime uzunluklarının frekans dağılımı yönteminin oldukça yüksek bir doğruluğa sahip olduğu görülmüştür [12] .

Toplam Kısaltma Kelimelerin Sayısı:

Kısaltma Kelimelerin frekansları, hesaplanabilir ve bu frekansların metindeki toplam kelime sayısına oranı, özellik değeri olarak kullanılabilir. Kısaltma kelimelerin listesi ek C'de yer almaktadır.

2.4.2. Sözdizimsel Özellikler

Sözdizimsel özellikler (Syntactic Features), fonksiyon kelimeler ve noktalama işaretlerini kapsar ve bir yazarın cümle düzeyindeki yazma tarzını engelleyebilir. Sözdizimsel gücün ayırt ediciliği özellkleri, insanların farklı cümle kurma alışkanlıklarından kaynaklanır. Sözdizimsel özellikleri düzenli noktalama işaretlerini içerebilir (örneğin; virgül, nokta, vb.). Tablo 3'de bu çalışmada kullanılan Sözdizimsel Karakterleri ile ilgili özellikleri listelenmiştir.

Fonksiyon kelimeler, online metinler için önemli bir ayırıcı özellikti, Fonksiyon kelimeler (veya gramere ait kelimeler), daha az sözcüksel veya belirsiz anlamlara sahip kelimelerdir. Ancak bunun yerine, bir cümle içinde diğer kelimelerle gramer ilişkileri ortaya çıkarmak için kullanabilir ya da yazarın ruhsal durumunu açıkça belirtirler. Dolayısıyla Fonksiyon kelimeler yazar analizi için kullanışlıdır [20], [17].

Belirli Fonksiyon kelimelerinin, verilen bir dilde bu kelimeler varlığında ya da yokluğunda bağlı olarak yerli yazarlar ve yerli olmayan yazarlar tarafından az ya da çok sık kullanılabilme ihtimali olduğu için bu tür kelimeler, anadil tanımlamasında da yararlı olacaktır. Genellikle dilin yerli konuşucuları tarafından daha az sıklıkla kullanılan "The" kelimesi buna iyi bir örnektir. Yapılan çalışmada yazarın bir yazarın anadilini

tanımlama sürecinde 308 sözdizimsel özellik kullanılmıştrı. Bu özellikler, Tablo 4'de verilmiştir. Fonksiyon kelimelerinin listesi Ek B'de verilmiştir.

Tablo 3. Sözdizimsel Özellikler Karakterler Listesi

Özellik Numarası	Özellik Açıklaması
64	Tek Tırnak sayısı (') /C
65	Virgül sayısı (,) /C
66	Nokta sayısı (.) /C
67	Kolon sayısı (:) /C
68	Semi-Kolon sayısı (;) /C
69	Soru İşaretleri sayısı (?) /C
70	Ünlem İşareti sayısı (!)/C
71	Elips sayısı (. . .) /C

Tablo 4. Sözdizimsel Özellikler – Fonksiyon Kelimeler Listesi

Özellik Numarası	Özellik Açıklaması
72-75	Article words sayısı /N (3 özellik)
76-80	Pro-sentence words sayısı /N (4 özellik)
81- 154	Pronoun words sayısı /N (74 özellik)
155-201	Auxiliary verbs sayısı /N (47 özellik)
202-223	Conjunction words sayısı /N (22 özellik)
224-332	Interjection words sayısı /N (109 özellik)
333-456	Adposition words sayısı /N (124 özellik)

2.4.3. Yapısal Özellikler

Metin düzenlemede insanlar çeşitli alışkanlıklara sahipler. Yapısal Özellikler (Structural Features) bir yazarın çeşitli yapısal özellikleri içeren yazının bir kısmının düzenini hangi yolla yaptığını gösterir [17]. Bu özellikler, cümlelerin toplam sayısı (S)/ dizeler, paragraf uzunluğu, cümle uzunluğu, her bir paragrafa düşen kelime sayısı ve karşılaşılanların kullanımı gibi faklı yerli dillerin yazma tarzıyla yazarın yazarlığa ait güçlü delilleri olabilir. Bu, daha az bilgi içeren ancak daha esnek yapılar veya zengin stilistik bilgi içeren online dokümanlarda daha belirgindir. Bu çalışmada, Tablo 5'de listelen 13 yapısal özellik kullanılmıştır.

Tablo 5. Yapısal özellikler Listesi

Özellik Numarası	Özellik Açıklaması
457	Satır toplam sayısı
458	Cümle toplam sayısı (S)
459	Paragraf toplam sayısı
460	Paragraf başına Cümleler ortalama sayısı
461	Paragraf başına Kelime ortalama sayısı
462	Paragraf başına Karakter ortalama sayısı
463	Cümlenin başına kelime ortalama sayısı
464	Cümle sayısı büyük harf ile başlayan /S
465	Cümle sayısı küçük harf ile başlayan /S
466	Boş satır sayısı / Satır toplam sayısı
467	Boş olmayan satır uzunluğu ortalama
468	Selamlama kelime sayısı
469	Veda kelime sayısı

2.4.4. İçeriğe Özgü Özellikler

Karakterlerin veya kelimelerin bir N-Gramı, n tane karakter veya kelime dizisidir. N-gramlar, metin tokenizasyonundan kaçınabilir ve bu yazar analizi için bir avantajdır. Bazı Yazarlık Atfetme çalışmaları, harflerin n-gram oluşma olasılıklarının ve frekanslarının bir yazarın tarzını karakterize etmede yararlı olduğunu kanıtlamaya çalışmaktadırlar [10], [18].

İçeriğe Özgü Özellikler (Content-specific Features), önemli anahtar kelime ve terimlerden ve bir n-gram sözcük gibi belirli konulardaki cümle gruplarından oluşmaktadır [11], [12]. Yazarın profilini geliştirmek için n-gram kullanılması, yazarlar için bir dizi model içindeki metinlerin Derlemi (Corpus) çevirmenin başarılı bir yöntem olduğu kanıtlanmıştır. Anadil Tanımlama durumunda modeller, n-gram sözcük dağılımları gibi oluşturulmuştur. yazara en iyi eşleşme, bir dokümanda en sık görülen N-gramları bulmayla karar verilir ve onlar bu durum bir metinde sık sık oluşur. Özellikler vektör boyutu azaltılması amacıyla, bu çalışmada sadece tüm Derlem de en az on kez ortaya çıkan Bi-gramlarını seçilmiştir.

2.5. Önceki ve İlgili Çalışmalar

Yazar Tanıma tarihi, Mendenhall [19] şu an stylometrics'in ne olarak adlandırıldığına dair çalışmaları ile on dokuzuncu yüzyıla kadar uzanır. 40 yıldan daha fazla bir süreden sonra Yule [34] ve Zipf [35], sırasıyla Yule yönettiği K istatistiği ve Zipf'in dağıtımı, onların metinsel istatistikleri daha önceki çalışmaların niteliğini etkilemiştir. Bu alandaki çalışmalar başlangıçta edebiyat metinleri üzerinde odaklanırken modern yazarlık özelliği çalışmaları veya geleneksel olmayan Yazar Tanıma Mosteller ve Wallace [36] tarafından yazılan Federalist Papers adlı çalışma ile 1964 yılında başladı. Ayrıca bilgisayar programlarının Yazarlık Analizi, Gary Sallis ve Macdonell [13] ile Krsul ve Spafford [18] tarafından yirminci yüzyılın sonlarında başlatıldı. E-posta yazarlığı, Devel öğrencisi Corney ve Argamon [8]'un çalışmaları ile yirmi birinci yüzyılda başladı,Yazarlık Analizi, son yıllarda Online mesajlara uygulanmıştır.

Örneğin, De Vel [19] e-posta iletilerinin yazarlarını belirlemek için bir dizi deney gerçekleştirdi. Onlar e-posta üzerinde yazarlığı tanımlamanın bazı yeni özelliklerini araştırdılar ve onlar sonuçları online topluluğa geleneksel Yazar Tanımlama yöntemlerini uygulamak için bir umut ışığıdır.

De Vel'in [37] çalışmasından farklı olarak biz, online metinlerin yazarlık karakterizasyonu için yazım tarzı özelliklerinin farklı türlerini karşılaştırmayı ile ve sınıflandırma tekniklerini vorgulayan bir metod oluşturmayı amaçladık.

İlk veya anadili İngilizcede yerli olmayan yazarların yazım tarzını etkileyebildiğini biliyoruz ve ayrıca bir kişinin yazımını başkalarından ayırabilen özellik setlerinin çeşitli tipleri tartışılmıştır. Bu stylometric özellikler, çok başarı bir şekilde Yazarlık Analizi problemlerinde kullanılmaktadır ve Stamatatos [38], yaygın olarak kullanılan özellik setlerinden bazılarının neden daha iyi ayrıklaştırma yaptığını açıklar.

Bu bölümün geri kalanında daha önce bu çalışmayla ilgili yapılan araştırmalar tanıtılmaktadır, hangi özellik setlerinin kullanıldığı gözden geçirilmiş ve bazı özellik setlerinin ne kadar iyi çalıştığı gözlemlemiştir.

2.5.1. Koppel

Otomatik olarak bir Yazarın Ana Dilinin tespit üzerine yayınlanmış ilk çalışma 2005 yılında Moşe Koppel tarafından yapılmıştır. Koppel, yazarın eserinde stilistik idiosyncrasies'i inceleyerek bilinmeyen bir Yazarın Anadilini Belirlemek için çalışmıştır.

Koppel, Uluslararası İngilizce Öğrencilerin verilerini kullanmıştır, sınıf olarak Çekçe , Fransızca, Bulgarca, Rusça ve İspanyolca'dan katkıda bulunayı düşündü. Her sınıfta 258 deneme kullanıldı ve her yazının uzunluğu 579 ile 846 kelime arasındaydı. Koppel, Fonksiyon Kelimeler, mektup n-gram, hatalar ve idiosyncrasies gibi çeşitli stilistik özellik setlerini kullandı [21].

1. Fonksiyon kelimeler: 400 özel fonksiyonu kelime seçilmiştir, ama Koppel, kullanılan kelimeleri listelememiştir.
2. Harf n-gram: 200 belirli n-gram seçilmiştir.
3. Hatalar ve idiosyncrasies: Koppel bir dizi yazım hatasını, neolojizmi ve Part-Of-Speech (POS)/Konuşmanın Bir Kısmı bigramsı göz önünde bulundurdu ve onu baştan sona doğru 185 hata türü ve özellik setleri olarak 250 nadir POS bigramla sınırlandırmıştır

Koppel, bu araştırmada araştırma sınıflandırma aracı olarak çok sınıflı lineer destek vektör makinalarını (DVM) kullandı ve seçilen özellik setleriyle yazarların anadillerini doğru bir şekilde sınıflandırarak toplamda % 80.2 kesinlik elde etti. Koppel, bazı özelliklerin diğer sınıflardan ziyade bir sınıfta daha sık görüldüğünü fark etti.

Koppel Yazarların Anadili Belirlenmesinde toplam %80 kesinlik sağlanmıştır. O verilerin bütünüyle tutarlı yazarların yeterliliğiyle oluştuğunu varsaymıştır ve deneme uzunluklarına göre özellikleri normalleştirilmiştir. Ancak, O İspanyolca korpusun hatalara Bulgarcadan daha yatkın olduğunu keşfetti. Daha eksiksiz bir model oluşturmak amacıyla Koppel, özelliklerin tüm korpustan hata frekansıyla normarleştirilmesini önermektedir.

2.5.2. Rappoport

Ari Rappoport, yazarların ana dillerinin belirlenmesi sadece özellik kümesi olarak Bi-gram karakter kullanımı konusunda Koppel'in araştırmasını tekrar incelenmiştir. Öncelikle Rappoport, bütün korpuslardan en sık kullanılan 200 Bi-gramı seçti ve

bu özellik seti, 3.99 bir standart sapma ile % 65.6 doğruluk elde etmek için kullanıldı [39]. Daha sonra, sadece tüm korpusda en az 20 kez görünen Bi-gramı seçerek daha da ileri gitti veya 84 Bi-gram ve bu bigramlardan sadece % 61,38'i sınıflandırma doğruluğuna ulaşmak için kullanıldı.

Bi-gram frekansların içeriği, önyargı konusu olabilir, Rappoport, alt korpusdaki tüm baskın kelimeleri değerlendirmek ve kaldırmak için istatistiksel bir ölçü kullandı ve ardından sınıflandırma deneyler tekrarlamıştır. Sınıflandırma doğruluğu olan sonuç, (sadece % 2 düştü) aslında aynıdır. Rappoport, ayrıca bütün fonksiyon kelimelerini, onların etkilerini ortadan kaldırmak için deney yaptı ve 62, 92% doğruluk elde etti. Son olarak, alt korpusun ikisinden Fransızca ve İspanyolca ile Hollandaca ve İtalyancanın yerini değiştirdi. Yeni veri seti sayesinde Rappoport, aslında orijinal veri seti ile de aynı olan 64,66% doğruluk elde etti. Rappoport, karakter bigrams'ın temel ses ve kısa ses dizileri seviyesinde dil transfer etkilerini yakalayabileceği sonucuna vardı.

2.5.3. Wong

Sze-Meng Jojo Wong, temel olarak Koppel'in Anadil Tanımlama üzerine yaptığı araştırmayı kullandı ve bundan başka ek bir özellik olarak sözdizimsel (syntactic) hataları da dahil araştırmıştır [40]. Wong, Koppel'in veri setinin son sürümünü iki ek alt-korpus (sınıf) olan Japonca ve Çincede kullanmıştır. Wong, yerli olmayan konuşucuların yapmaya daha yatkın oldukları sözdizimsel hataların üç türünü özellik olarak seçmiştir. Seçilen sözdizimsel üç hata tipi: özne-fiil uyumsuzluğu, isim sayı çelişkisi ve belirteçlerin yanlış kullanımıdır.

Wong, bu sözdizimsel özellikleri kullanılarak iki farklı inceleme yapmıştır. İlk olarak, Wong, yedi yerli dile göre tür sözdizimsel hatanın üç türünün frekansını inceledi ve özellik seti olarak sadece üç sözdizimsel hatayı kullanılarak sınıflandırma tekniklerini uyguladı. Wong, öğrenme aracı olarak lib SVM'yi ve grammer ölçmek için Queequeg olarak adlandırılan bir aracı kullanmıştır. Wong'un araştırmasının ikinci bölümünde, Koppel'in araştırmasında % 80 doğruluk geliştirdiği, sözdizimsel üç özelliği birleştirebilseydi, Koppel'in çoğaltılan çalışmaları ve araştırma için üç sözdizimsel özellik ile onun çoğaltılmış çalışma versiyonunu birleştirecekti.

Fonksiyon kelimeler ve N-Gram özellik setleri olarak kullanıldığında en iyi sınıflandırma doğruluğu elde edildi. Ayrıca, bir özellik seti gibi n-gramlara karakter eklenmesi doğruluğu geliştirmedi. Wong, sözdizimsel üç hatanın tam doğruluğunu geliştirilememesini ya yeterli hata türünün kullanılmamasına ya da sözdizimsel hataların yazarın anadilini belirlemede iyi bir gösterge olmamasına bağlamıştır.

2.6. Web Tabanlı Metinlerin Nitelikleri

Anadil Tanımlanmasının önceki hedefleriyle karşılaştırıldığında, web tabanlı metinlerin Anadil Tanımlanmasının ve online metinlerin uzunluğu sınırlandırılmış mesajların bir mücadelesidir [12]. İddia edildiği gibi yazarlık nitelikleri, 500 kelimenin altında güçlü bir şekilde belirgin olmaz. İlgili çalışmalarda yazıların boyutunun incelenmesine dayanılarak 250 kelimeden az bir metin bir yazara tanıma çok zor olduğu bulundu [13].

Online metinlerin kısalığı, normal metinlerdeki bazı tanımlayıcı özelliklerinin etkisiz kalmasına neden olabilir. Diğer taraftan, yazarın anadilini doğru bir şekilde Tanımlamada, Web Tabanlı metinler üzerinde yazarın yazım tarzını ortaya çıkarmak için yardımcı olabilecek bazı özelliklere sahiptir. Haber-grubu, e-posta, bloglar ve sohbet odaları gibi web tabanlı ortamlara nispeten resmi yayınlarla karşılaştırıldığında yazıların kendilerine ait "yazma baskılar" (write-prints) bırakması olasılığı daha yüksektir. Örneğin, online metinlerde kullanılan kuruluş biçimi veya kompozisyon tarzı yaygın metin dokümanlardan genellikle farklıdır. Yapısal olarak plan çizme özellikleri, alışılmamış içerik belirleyicileri ve düzensiz dilde kullanım gibi bazı özel özellikler, uygun bir koleksion özelliğinin geliştirilmesinde yararlı olabilir.

İngilizce web sayfalarının binlercesinin bir korpusunun analizi, anadili İngilizce olan yazarlar ve yerli olmayan yazarlar arasındaki yazım tarzında ve içerikte önemli farklılıklar gösterir [15]. Bu tür farklılıklar, bir web sayfasının metininin temelinde bilinmeyen bir yazarın anadilini belirlemek için kullanılabilir. İnternet global bir ağdır, ve Cyberkullanıcıları siberuzay üzerinde çoğunlukla İngilizce mesajlar dağıtmaktadılar. Online metinler, çeşitli yerli dille birçok ülkeler arasında internet üzerinde dolaştırırlır. Yazım-tarzı (Writing-style) özellikleri, çoğunlukla dile bağlıdır. Bu nedenle, Yazarın Anadilini Tanımlamada tahmin gücü farklı ana diller için hemen bir endişe kaynağıdır.

2.7. Sınıflandırma Teknikleri

Stylometry, önerilen özelliklerle ayırt edici ölçümler elde eder, bu nedenle bir örüntü içinde belli bir yazarın profil tarzını azaltır [5]. Bir örüntü eşleştirme sorunu makine öğrenimi için özellikle uygundur. Makine öğrenimi, görünmeyen verilerin sınıflandırılmasını daha önceden görülen verilerden öğrenmiş bilgiye dayalı bir model oluşturarak sağlar. Makine öğrenim algoritmasının sonuçları, modelin doğruluğunu belirlemek ve değerlendirmesi için sırayla ölçme gerektirir.

Bu algoritmanın, bilgiyi temsil eden modelin bazı türlerini üretmesi gerekir ve biz onun ne kadar iyi bir model olduğu performansını ölçerek veya bilinmeyen örnekleri sınıflandırma onun kabiliyeti belirleyerek elde ediyoruz. Çeşitli makine öğrenme teknikleri, yazarlık atfında kullanılır. Bu çalışmada çok kullanışlı ve çok güçlü olan C4.5, Naïve Bayes ve Destek Vektör makinesi gibi üç sınıflandırma tekniğini kullanıldı.

2.7.1 Naïve Bayes Metodu

En etkili ve yaygın olan sınıflandırma algoritmalarından biri Naïve Bayes veya Navie Bayesian algoritmasıdır. Özellikler arasında serbestliği varsaydığından uygulanması çok kolaydır. Eğer özellikler birbirini etkiliyorsa olasılığı hesaplamak o zaman zorlaşmaktadır. Naïve Bayes modeli veya basitçe Bayes sınıflandırma (Simple Bayesian Classifier) algoritması aşağıdaki gibi gerçekleşir.

1. Eğer etiketlenmiş sınıfla ilişkili bir gurup küme içinde D'nin bir eğitim seti olduğunu farzedersek normal şekilde, her küme n-boyutlu özellik vektörü; $X=(x_1,x_2,\ldots,x_n)$, olarak temsil edilir, ve $A_1, A_2, \ldots, A_n$, sırayla, n özellikten oluşan kümelerin üzerinde yapılmış n ölçüyü gösterir.

2. $C_1, C_2, \ldots, C_m$, M sınıflarının olduğu varsayarak, C ve X den ulaşan bir tuple (İkili), olasılık fonksiyonunu bularak sınıflandırır. X koşularına göre bu X, en yüksek posteriyor olasılığına sahip olan sınıfa aittir. Yani, Naïve Bayesian bir olasılığı sınıflandırıcı yöntemdir ki X, C_i sınıfına aittir eğer ve sadece eğer;

$$P(C_i|X) > P(C_j|X) \quad \text{eğer } 1 \leq j \leq m\,, \qquad j \neq i. \tag{1}$$

Böylece, $P(C_i \mid X)$ en yüksek olasılığa sahip olan sonrakidir. C_i sınıfı herhangi $P(C_i \mid X)$ için maksimum sonraki hipotezdir. Bayes teoremi (Denklem (2));

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{2}$$

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X} \tag{3}$$

3. $P(X)$ tam sınıflar için sabittir, sadece $P(X|C_i)P(C_i)$'in maksimize edilmesine gerek vardır. Eğer sınıfın önceki olasılıklar bilinmiyorsa, o zaman yaygın olarak o sınıfların eşit olasılıklara sahip oldukları kabul edilir, yani $P(C_1) = P(C_2) = \ldots = P(C_m)$, bu nedenle $P(X \mid C_i)$, en üst düzeye sahip olandır. Aksi halde, maksimum olan $P(X|Ci)P(Ci)$'dir.

Unutmayalım ki, sınıfın ilk ihtimali $P(C_i)=|C_i,D|/|D|$'e göre tahmin edilebilir, burada $|C_i,D|$, Ci sınıfı ve D eğitim sayısı tuple'larıdır.

4. Birçok özelliğe sahip olan veri setleri için $P(X|C_i)$'nin hesaplanması son derecede pahalı olacaktır. $P(X|C_i)$'i değerlendirmede hesaplamayı azaltmak için sınıf koşulu serbestliğine bağlı olan Naïve varsayımı yapılır. Bu, tuple'ın verilen sınıf etiketi içinde (yani, nitelikler arasında herhangi bir bağımlılık ilişkisi yoktur) niteliklerin değerlerinin bir diğerinin bağımsızlık şartına bağlı olduğunu varsayar. Böylece, birbirlerinin eşdeğeri olan, sırasıtla denklem (4) ve denklem (5) eşitliklerini dönüştürülür.

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \tag{4}$$

$$P(X|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \cdots \times P(x_n|C_i) \tag{5}$$

$P(x_1|C_i)$, $P(x_2|C_i)$, … , $P(x_n|C_i)$ olasılıklarını, eğitim tuplerinin üzerinden kolayca tahmin edebiliriz. Burada tuple X için X_k değeri A_k özniteliğini ifade ettiğini hatırlayalım. Her öznitelik için, özniteliğin bir kategorik veya sürekli değerli olup olmadığını incelemekteyiz. Örneğin, $P(X|C_i)$'i hesaplamak için, aşağıdaki adımları dikkate alırız:

(a) A_k kategorisel ise, o zaman, $P(x_k|C_i)$, Ci sınıfı ile D tupleler dizisinde A_k için x_k değerine sahip olan sayıdır, $|C_i, D|$, C_i sınıf ile D tupleların sayısına göre bölünür.

(b) A_k sürekli-değerli ise, o zaman biraz daha fazla çalışmamız gerekir, ama hesaplama oldukça basittir. Sürekli-değerli bir özniteliği genelde μ ortalaması ve ∂ standart sapması ile bir Gauss dağılımının olduğunu varsayılır. Böyle tanımlanmış: denlem (6) ile ifede edeilir.

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6}$$

Öyle ki

$$P(x_k|C_i) = g(x_k, \mu_{Ci}, \sigma_{Ci}) \tag{7}$$

Bu denklemler, zor görülebilir, bizim μ_{Ci} ve ∂_{Ci}'y$_i$ hesaplamamız gerekir, ki bunlar, C_i sınıfı için eğitimi tuplelar içinde öznitelik A_k değerleri olarak, sırasıyla ortalama ve standart sapma anlamına gelir. Daha sonra, bu iki miktarları denkleme (4) içinde birbiriyle ve x_k için $P(x_k | C_i)$ tahmin edilir.

5. X'in etiketli sınıfını tahmin etmek amacıyla, her C_i sınıfı için $P(X|C_i)P(C_i)$ değerlendirilir. Tahmin edici sınıflandırma tuple X'deki etiketlenmiş sınıfın C_i olan sınıf olduğunu tahmin eder eğer ve sadece eğer denkleme (8) :

$$P(X|Ci)P(Ci) > P(X|Cj)P(Cj) \text{ eğer } 1 \leq j \leq m, j \neq i \tag{8}$$

Diğer bir ifadeyle, tahmin edilen sınıf etiketi, maksimum olan herhangi $P(X|C_i)P(C_i)$ için C_i sınıfıdır.

"Bayes sınıflandırıcılar ne kadar etkilidir ?" Bu sınıflandırıcıların çeşitli deneysel çalışmaları, karar ağaçları ve yapay sinir ağı sınıflandırıcıları ile bazı uygulama alanlarında

karşılaştırabilir. Teorik olarak, Bayes sınıflandırıcı araçları diğer tüm sınıflandırıcı araçlara göre en az hata oranına sahiptir. Fakat, bunun kullanımı için yapılan varsayımlardaki, yani sınıf bağımsızlık koşullu ve verilerin uygun olasılığının eksikliği gibi yanlışlıklar sebebiyle bu uygulamalarda her zaman geçerli değildir. Ayrıca Bayesian sınıflandırma araçları, açıkça Bayes teoremi kullanmayan diğer sınıflandırıcılar için teorik bir gerekçe sağlamada yararlıdırlar. Örneğin, bu belirli varsayımlar altında, birçok sinir ağı ve eğriye uyan algoritmaların sonucu, Naïve Bayes'de olduğu gibi maksimum sonraki hipotez (posteriori hypothesis) olduğunu gösterir.

2.7.2. Destek Vektör Makinesi

Destek Vektör Makinaları (DVM), yazarlık atfı için daha sık kullanılan bir teknik haline gelmektedir. 1995 yılında Vladimir Vapnik tarafından geliştirilmiştir, SVM özelliklerinin bir grubunu alır ve yeni bir uzaya ulaşılması için burada, yeni uzayın içinde öznitelik vektörleri bölmek için bir hiper düzlemi belirleyebilen bazı hesaplamalar yapar. İdeal hiper düzlemi öznitelik vektörleri, maksimum mesafe (tolerans) ile ayırır. Çünkü hiper düzlemin yalnızca iki yüzü vardır ve bu teknik, ikili sorunları için çok uygundur.

Destek Vektör Makineleri (DVMs) doğrusal modeller ile sınıflandırma kavramını genişletir. Giriş verilerin vektör değerleri, bu iki sınıftan olmak üzere sınırlıdır ve doğrusal olmayanlar da Öklid uzaydan yeni yüksek boyutlu bir Hilbert uzaya dönüştürülmüş olabilir. Bu model, daha sonra bu yeni alanda inşa edilen en çok mesafesi olan hiper düzlemidir. Maksimum toleransli hiper düzlem, iki sınıf arasındaki en büyük ayırıma sahip olan modeldir. O hiper düzlemde yerleştirilmiş bu veri vektörleri tarafından tanımlanır. Bu veri vektörlerine de "destek vektörleri" denir.

Test veri vektörleri, aynı şekilde yeni bir uzay haline dönüştürülür ve maksimum toleranslı hiper düzlem tarafından belirlenip sınıflandırılarak bunlar arasında yer alır. Maksimum marjin hiper düzlem tarafından tanımlanan model bir çekirdek (kernel) fonksiyonu üzerine kuruludur. Bu, bir radyal tabanlı fonksiyon çekirdeği veya bir sigmoid çekirdeği gibi doğrusal bir kernel fonksiyonunu, bir polinom kernel fonksiyonunu veya diğer doğrusal olmayan fonksiyon olabilir. En iyi kernel fonksiyonunu genellikle deneysel olarak belirlenir.

Tartışılan metin sınıflandırma araştırma sonuçları, SVM metin sınıflandırma problem alanında yukarıda tartışılan oranla: Naive Bayes sınıflandırıcı ve C4.5 sınıflandırıcı gibi yöntemlerden daha iyi bir makine öğrenme sınıflandırıcısı olarak ortaya çıktığını göstermektedir. Metin sınıflandırma genelde yaklaşım özellik kümesini sık sık görülen kelime sayısını kullanacak biçimde ağırlıklandırır. Bu araştırmada kullanılan özelliklerin birçoğu bazı tiplerin frekanslarına dayalıdır ve bunun SVM Web tabanlı metinlerde yazarın anadilinin keşfi için uygun bir sınıflandırıcı olduğuna inanılmaktadır.

2.7.3. Karar Ağaçları

Karar Ağaçları (KA), esası yukarıdan aşağıya bir yaklaşım kullanarak bilgi edinimine (Information Gain) [22] bağlı sonuç çıkarabilir. Ağacın her seviyesinde kök düğümünden başlayarak, kendi sınıflarından maksimum bilgi kazanç oranını sağlayan özelliği seçilir. Bu, minimum yapıya sahip bir karar ağacı üretir. Quinlan, C4.5 sınıflandırıcısını bu yaklaşımı kullanarak üretir ve özellik değerleri tamamen sayısal olduğu zaman ve eksik özellik değerleri ile başa çıkabilmek, o verilerin ayrıştırılmasını gerçekleştirebilmek için onu optimize eder.

Karar Ağacı Sınıflayıcılar, kolaylıkla yorumlanabilen öğrenme algoritmalarıdır. Çünkü onlar iç içe geçmişse o kurallarından oluşur. İç düğümler, terimler tarafından etiketlenir, bunlardan uzaklaşan dallar, deney belgesinde terimi olan ağırlık testleri üzerinde etiketlenir ve yapraklar kategorilere göre etiketlenir. KA öğrenme algoritmasının amacı, farklı kategoriler içinde en iyi sınıflandırma problemlerın ayırımını yapan testleri bulmaktır. Örnekler arasında en iyi ayrım gücüne sahip olan kök düğümü veya test olmalıdır. Yoksa tüm eğitim örnekleri aynı etikete sahipse (örneğin reddetmek), bir takım kategoriler içinde bütün eğitim örnekleri birbirinden ayrılana kadar yeni testlerin tasarlanması gerekir [30].

Başvurulan bir yazarlık atıf işlemi için, KA düğümleri veya testler tarafından temsil edilen ayırıcı özellik uygulanır. Dalları üzerindeki değerler, eşikler olabilir. Kural tabanlı öğrenim ve karar ağaçları, her özellik için bir değerler grubuyla sınırlandırılan veya her sayısal özellik için çeşitlilik gösteren bir grup içinde ayrıklaştırılan veri değerleridir. Verilerin ayrıştırılması bu tekniklerin başarısıyla eleştirilebilir.

K yakın komşu (K-NN) tekniği gibi örnek tabanlı öğrenme, verim sınıfını en iyi şekilde eşlemeyi bulmak için her eğitim örneğinin her test örneğiyle karşılaştırılabilmesi için yoğun hesaplamalar olabilir. Bu nedenlerle, bu makina öğrenme tekniklerini, bu araştırmada kullanmayı düşünmüyoruz.

2.8. Sınıflandırma Performansının Ölçütleri

Sonuçları ölçmek için birçok farklı yol vardır ve farklı yöntemler verilen bir problemin farklı yönlerindeki başarıyı gösterir. Bu bölümde, araştırmada kullanılan değerlendirme ölçütleri anlatılmıştır. Genel olarak model başarısının değerlendirmesinde kullanılan temel kavramlar hata oranı, kesinlik, hatırlama ve F-ölçütüdür. Modelin başarısı, doğru sınıfa atanan örnek sayısı ve yanlış sınıfa atılan örnek sayısı nicelikleriyle ilgilidir.

Sınıflandırma performansını hesaplamadan önce, sınıflandırma sonuçlarının aşağıdaki 4 sonuç türü kullanılarak bilinmesi gereklidir.

- Pozitif Doğru (PD): Pozitif bir sınıf olarak sınıflandırıcı tarafından doğru bir şekilde işaretlenmiş pozitif sınıftır.
- Negatif Yanlış (NY): Olumsuz bir sınıf olarak sınıflandırıcı tarafından doğru bir şekilde işaretlenmiş doğru sınıftır.
- Pozitif Yanlış (PY): Olumlu bir sınıf olarak sınıflandırıcı tarafından yanlış bir şekilde işaretlenmiş yanlış sınıftır.
- Negatif Doğru (ND): Olumsuz bir sınıf olarak sınıflandırıcı tarafından doğru bir şekilde işaretlenmiş yanlış sınıftır.

2.8.1. Kesinlik ve Duyarlılık

Kesinlik, sınıflandırılmış sınıf için doğru sınıflandırılmış öğeler ve öğelerin toplam sayısına oranı ölçülerek hesaplanır, bu oranı kullanarak doğruluğu ölçer. Diğer bir ifadeyle, kesinlik, sınıfı 1 olarak tahmin edilen True Pozitif örnek sayısının, sınıfı 1 olarak tahmin edilen tüm örnek sayısına oranıdır:

Yani kesinlik ölçümü, bütün durumların dışında bir belgenin doğru bir şekilde kaç defa sınıflandırılmış oluğunu gösterir.

Kesinlik, hedeflenen sınıf için sınıflandırılan tüm öğelerin dışında sadece kaç öğenin doğru şekilde sınıflandırılmış olduğuna dikkat eder, böylece doğru bir şekilde sınıflanmayan gerçek öğeler hakkında hiçbir şey söylemez. Bu yüzden kesinlik ölçümünün sadece doğruluğu ölçtüğünü söyleyebiliriz. Denklem 9 Kesinlik hesaplamayı gösterir:

$$\text{Kesinlik} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

Gerçek gibi yanlış olarak sınıflandırılan öğelerin göz ardı edilmesinin aksine Duyarlılık, sınıf olarak kategorize edilmiş öğelerin toplam sayısıyla ilişkilendirildiğinde doğru bir şekilde sınıflandırılan öğelerin sayısını ölçer. Diğer bir deyişle, Duyarlılık sadece bütünlüğü ölçer. Denklem 10 Duyarlılık hesaplamayı gösterir:

$$\text{Duyarlılık} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

2.8.2. Doğruluk

Doğruluk, her durumda fakat doğru sınıflandırılmış maddelerin sayısını ölçer. Doğruluğu tanımlamak için bir başka yol, doğru değerlerinin yakınlık derecesinin ölçülmesidir. Doğruluk, verilerin toplamını, olayların toplam sayısına bölmeyle hesaplanır. Denklem 11'de gösterilmiştir. Doğru sınıflandırılmış pozitif örnek sayısının, toplam pozitif örnek sayısına oranıdır.

$$\text{Doğruluk} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{11}$$

2.8.3. F-ölçütü

F-ölçütü, hem Hatırlatma hem de Kesinlik tartmayla doğruluğu ölçmenin diğer bir yoludur. Hatırlama ve Kesinliğin harmonik anlamındadır ki hem kesinlik hem de hatırlatma değerleri yüksekse o da yüksek bir değere erişir. Eğer hatırlatma ya da kesinlik

düşükse, F-ölçütü da cezalandırılacaktır. F-ölçütü, doğruluk değerlerini etkileyen gerçek yanlışları göz ardı etme ve eşit olarak Kesinlik ve Hatırlatmayı tartmayla zarar gören diğer ölçülere çözüm konusu olan bir ölçüdür.

Kesinlik ve Duyarlılık ölçütleri tek başlarına anlamlı bir karşılaştırma sonucu çıkarılması için yeterli değildir. Her iki ölçütü de beraber değerlendirmek daha doğru sonuçlar verir. Bu sebeple f-ölçütü tanımlanmıştır. F-ölçütü, kesinlik ve duyarlılığın harmonik ortalamasıdır. Denklem 12 F-ölçütü hesaplamayı gösterir.

$$F - \text{Ölçütü} = \frac{2 \times Duyarlilik \times Kesinlik}{Duyarlılık + Kesinlik} \tag{12}$$

3. YAPILAN ÇALIŞMALAR

3.1. Giriş

Bu bölümde deneme kurulumunun detayları açıklanacaktır. Öncelikle, Web tabanlı metinlerde ve bizim önerdiğimiz çerçevede Yazarın Anadili Tanımlamasını tartışacağız. Bu aşamada, bizim çerçevemizin dört adımını ve veri kümesi toplanmasını açıklayacağız. Son olarak da bizim korpusumuzun istatistiksel özellikleri analiz edeceğiz.

Ayrıca, özellik çıkarma aşamasında makine öğrenme araçlarına göre kullanılabilir format için veri kümesindeki bilginin nasıl dönüştürüldüğünü açıklayacağız. Bir sonraki adımda, geliştirilen Yazarın Anadilini Tanımlama uygulaması hakkında yorum yapacağız. Son olarak, Weka makine öğrenme aracı üzerinde ve çıkarılan özellik vektörü denemelerini sınıflandırma algoritmaların kabiliyeti için bir bildirim yapacağız.

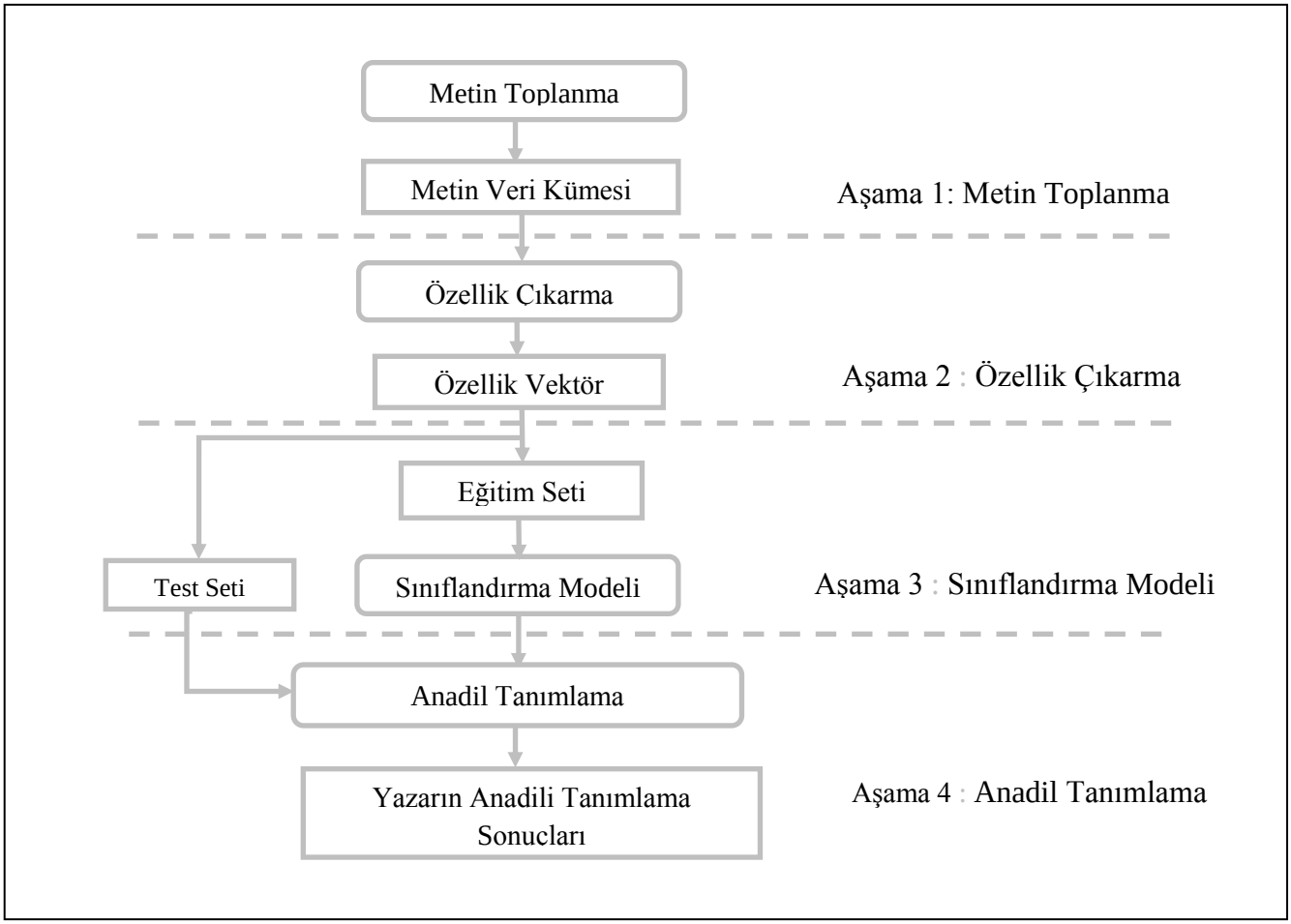
3.2. Web Tabanlı Metinlerde Yazarın Anadilini Tanımlama

Bir Yazarın Anadilini Tanımlama, yazarlık analizi sorununun bir türüdür. Listelenen yazarlar arasından belirli bir yazarın belirlemesi yerine, özel bir yerli dili paylaşan bir grup yazarı tanımlamayı ümit ediyoruz. Yazarlık analizi son yılarda çevrimiçi metinlere uygulanmıştır [11], bu çalışmada Anadil Tanımlamada geniş kapsamda stylometric özelliklerinin kullanımı hakkında bir inceleme yapmayı amaçladık. Ayıca, bir sınıflandırma sorunu olan Yazarın Anadilini Tanımlama aşağıdaki gibi gelişebilir, İngilizcede farklı dillere sahip yazarlardan bir dizi metnin sağlanması ve bir yazarın anadili sınıfının belirlenmesi için bilinmeyen yeni bir metin belirleme. Bu hipotezi uygulamak için aynı anadilin yazarları tarafından yazılmış çok sayıda metin oldukça sabit kalan bu özellikler kümesi seçilir. Özellik grubu seçildiğinde, verilen bir metin, N özelliklerinin toplam sayısı olan N-boyutlu bir vektör ile temsil edilebilir. Önceden sınıflandırılmış metinlerin bir grubu verilmişse, yeni metne dayanarak oluşturulan yeni vektörün kategorisini belirlemek için birçok tekniğe başvurabiliriz.

3.4.Web Tabanlı Metinlerde Yazarın Anadilinin Tanımlaması İçin Bir Çerçeve

Yazarın Anadilini Tanımlama işlemi için önerdiğimiz çerçeve dört aşamadan oluşur: (bkz. Şekil 2)

1) Metin Toplanma: Yazı stillerini profillemek için web tabanlı metinlerin uygun bir korpusa toplaması.
2) Özellik Çıkarma: Otomatik metinlerden bir özellik seti çıkarma işlemidir.
3) Sınıflandırma Modeli: Test örneklerini sınıflandırma için kullanılan eğitim veri grubuyla sınıflandırıcı eğitim tarafından bir sınıflandırma modeli oluşturma.
4) Anadil Tanımlama: En son bu test dokümanlarının anadilini tahmin etmek için kullanılacak sınıflandırma modeli tasarlama.



Şekil 2.Yazarın Anadil Tanımlama işlemi için önerilen çerçeve

3.4.1. Korpusu Toplama

Önceki Anadili Tanımlama çalışmalarda en çok çeşitli ülkelerden yerli olmayanların İngilizce yazı çalışmalarının doğru amaçlar için toplanan İngilizce Öğrencilerinin Uluslararası Corpusu (Derlemi) [16] kullanılmıştır.

Bu çalışmanın amacı için Web tabanlı metinlerin farklı türleri arasından ulaşılabilen bütün Web tabanlı metinleri inceledik, kişisel e-posta ve chat mesajları genellikle kişisel gizlilik sorunları içerir ve toplanmaları zordur. Ayrıca, anadili bilinen yazarların metinlerinin toplanması sıkıntılıdır. Bu nedenle, bu çalışmada Derlem olarak seçilen ve toplanan İngilizcede herkese açık haber ajansları metinleri (haberler) Ana Dilleri İngilizce, Farsça, Türkçe ve Almanca yazarlar tarafından yazılmıştır. İndirmelerden her biri sağlanan anadilin göstergelerini içerir ve indirilen her bir metin haber şeklindedir.

Bu metinlerin bazılarında İngilizce dışında diğer dillerde ek metinler içermektedir; bu metinler hedefimiz için dikkate alınmamıştır. Her bir dil için (Türkçe, İngilizce, Farsça ve Almanca) 150'şer metin toplanmıştır.

3.4.2. Özellik Çıkarma

Özellik Çıkarma (Feature Extraction) aşaması, oluşturulan özellik kümlerinden tüm özelliklerin çıkarılmasını içermektedir. Yazarların yazım tarzı özellikleri yapısız metinlerden çıkarılma ihtiyacı duyar ve ardından özellik çıkarıcı, özellikleri temsil etmek için 968 boyutlu vöktör üretir. Önceki bölümde açıkladığımız gibi çevrimiçi metinler için 968 63 Sözlüksel, 393 Sözdizimsel, 13 Yapısal ve 500 İçeriğe Özgü özellikler seçilmiştir.

Çalışmada bir otomatik bir özellik çıkarma uygulaması geliştirilmiştir. Daha sonra rastgele seçilen her dil sınıfı için 10 metnin özellikleri çıkarılmış ve doğruluğunu onaylamak için elde edilen özellikler el ile orijinal belgede incelenmiştir. Bura bazı özelliklerin, bilhassa da yapısal özelliklerin doğru bir şekilde elde edilmesi zordur. Örneğin, bir yazar selamlama için nadir bir sözcük kullanıyorsa program bu özelliği doğru olarak çıkaramayabilir.

3.4.1.1. Veri Kümesinin İstatistiksel Analizi

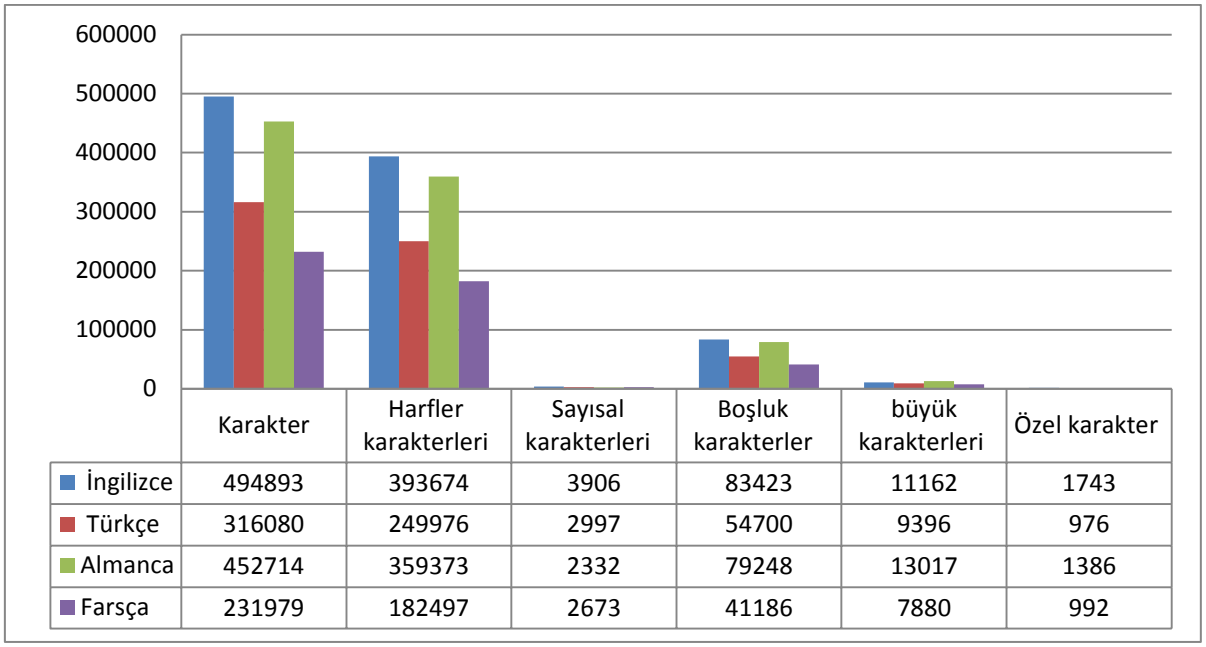
Bu bölümde, verilerin açıklaması yapılmıştır. Bu bölümün önemli bir kısmı istatistiksel analiz veri kümesini içeriyor. İstatistiksel sonuçlar, kurulan otomatik Yazarlık Analizi sistemleri için kullanılabilen özelliklerin bazıları Anadildeki farklılıkları ortaya çıkarırlar. Aşağıdaki gibi Tablo 5'te sonuçlar gösterilmiştir:

Tablo 5. Veri Kümesi Boyutu

Anadil Sınıfı	Doküman Sayısı
İngilizce	150
Türkçe	150
Almanca	150
Farsça	150
Toplam	600

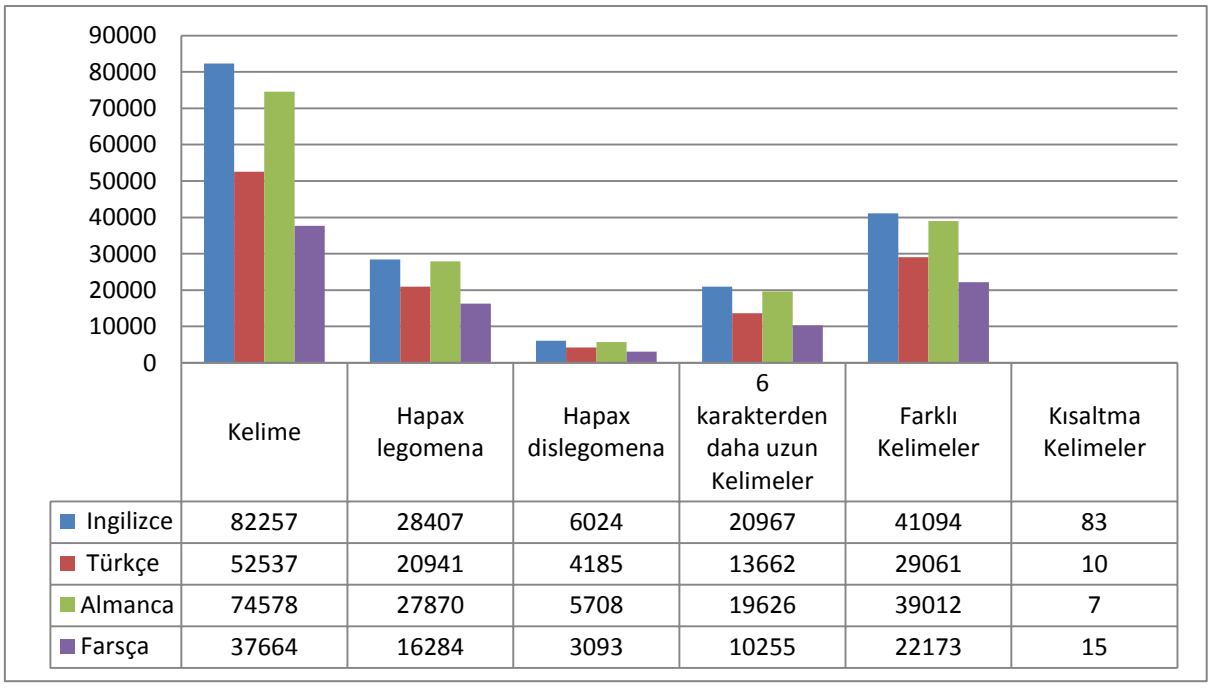
3.4.1.2. Sözlüksel Özellikler

Şekil 3 ve 4'de önceki bölümde (Bölüm 2.4) önerilen sözlüksel özelliklerin bazıları gösterilmiştir. Bu tablolar Karakter tabanlı ve Kelime tabanlı eğitim veri kümesinde (Derlemde) mevcut Sınıflardaki özelliklerin istatistiksel analizini göstermektedir. İngilizce sınıfında karakterin toplam sayısı diğer sınıflara göre daha uzundur ve diğer karakter tipleri Şekil 3 'de incelenmiştir.



	Karakter	Harfler karakterleri	Sayısal karakterleri	Boşluk karakterler	büyük karakterleri	Özel karakter
İngilizce	494893	393674	3906	83423	11162	1743
Türkçe	316080	249976	2997	54700	9396	976
Almanca	452714	359373	2332	79248	13017	1386
Farsça	231979	182497	2673	41186	7880	992

Şekil 3. Karakter Tabanlı Sözlüksel Özellikleri İstatistiksel Analizi



	Kelime	Hapax legomena	Hapax dislegomena	6 karakterden daha uzun Kelimeler	Farklı Kelimeler	Kısaltma Kelimeler
Ingilizce	82257	28407	6024	20967	41094	83
Türkçe	52537	20941	4185	13662	29061	10
Almanca	74578	27870	5708	19626	39012	7
Farsça	37664	16284	3093	10255	22173	15

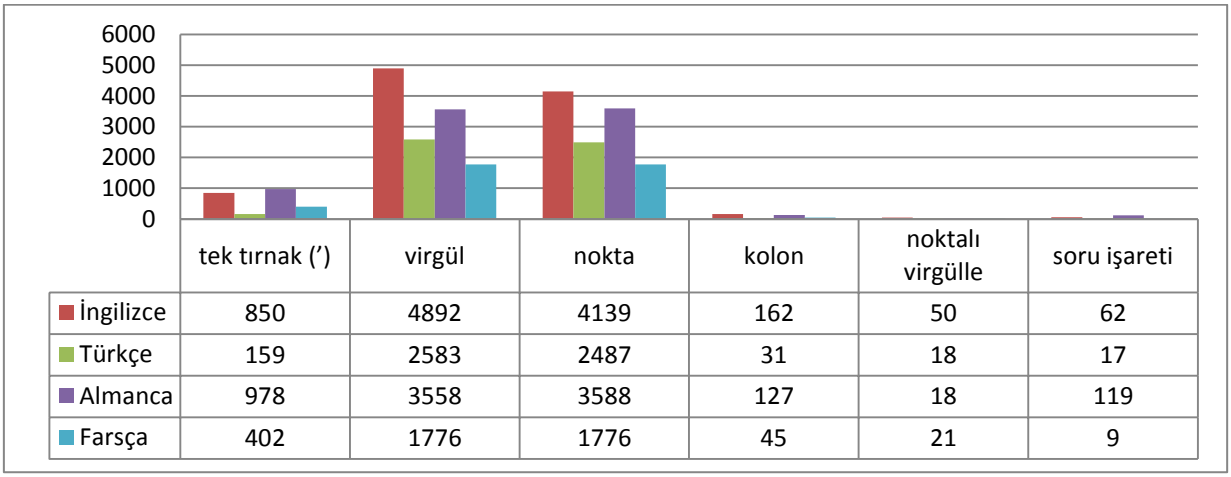
Şekil 4. Kelime Tabanlı Sözlüksel Özellikler İstatistiksel Analizi

Kelime Tabanlı Sözlüksel Özellikler yazar tespiti için önemli bir ölçüttür. İngiliz yazarlardan uzun metinler kullanmak farklı kelime sayısı oranı ve kelimelerin toplam sayısı İngilizce, Almanca, Türkçe ve Farsça için sırasıyla 0.49, 0.52, 0.55, 0.58'dir. Hapax legomena oranı ve kelimelerin toplam sayısı İngilizce, Almanca, Türkçe ve Farsça için

sırasıyla 0.34, 0.37, 0.39 ve 0.43'tür. İngiliz yazarlar tarafından kullanılan kısaltma kelimelerin toplam sayısı çoğunlukla farklıdır ve diğer yerli olmayan yazarlar arasında İngiliz yazar daha çok kısaltma kelime kullanmıştır.

3.4.1.3. Sözdizimsel Özellikler

Bahsedildiğimiz gibi Sözdizimsel Karakterlerin ve Fonksiyon Kelimelerin oluşturduğu Sözdizimsel listeyi sunduk. Şekil 5'de veri kümesi her bir sınıftaki sözdizimsel karakterlerin toplam sayısını göstermektedir. Şekil 5 ve Şekil 6 'de veri kümemiz, her sınıftaki Sözdizimsel karakterlerin ve Sözdizimsel karakterler toplam sayısı açıklanmaktadır.

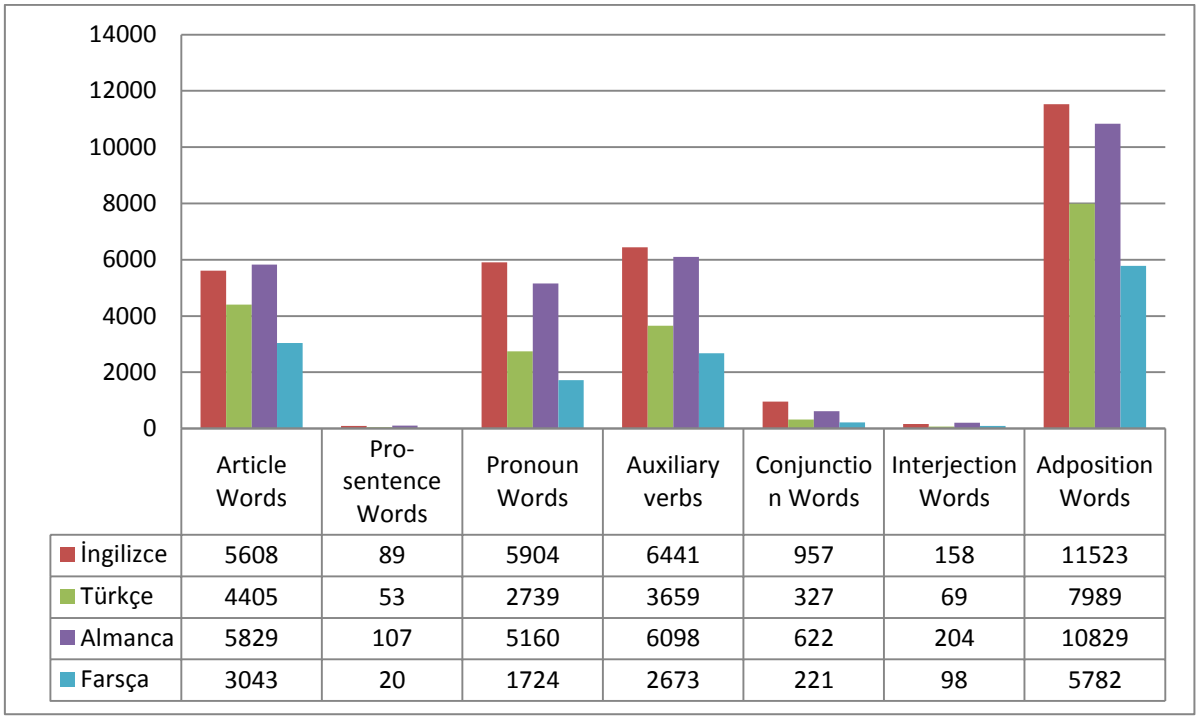


	tek tırnak (')	virgül	nokta	kolon	noktalı virgülle	soru işareti
İngilizce	850	4892	4139	162	50	62
Türkçe	159	2583	2487	31	18	17
Almanca	978	3558	3588	127	18	119
Farsça	402	1776	1776	45	21	9

Şekil 5. Sözdizimsel Karakterlerin İstatistiksel Analizi

Şekil 5'in sonuçlarına bakarak, onların yazılarında Alman yazarların daha çok (') ve (?) karakterini kullandığını, diğer taraftan İngiliz yazar, diğer yazarlardan daha çok noktalı virgülü kullandığını görebiliriz. Sınıf başına nokta işaretinin toplam sayısının oranı ve sınıf başına karakter sayısı İngilizce, Almanca, Türkçe ve Farsça için sırasıyla 0,0083, 0,0079, 0,0078, 0,0076'dır.

Aynı zamanda her dil sınıfında önceden belirlenmiş yedi kategorideki Fonksiyon kelimelerin toplam sayısı hesaplanmıştır. Şekil 6 göz önüne alınarak, çeşitli dillerde konuşan yazarlardaki farklı fonksiyon kelimelerinin dağılımını analiz edebiliriz.



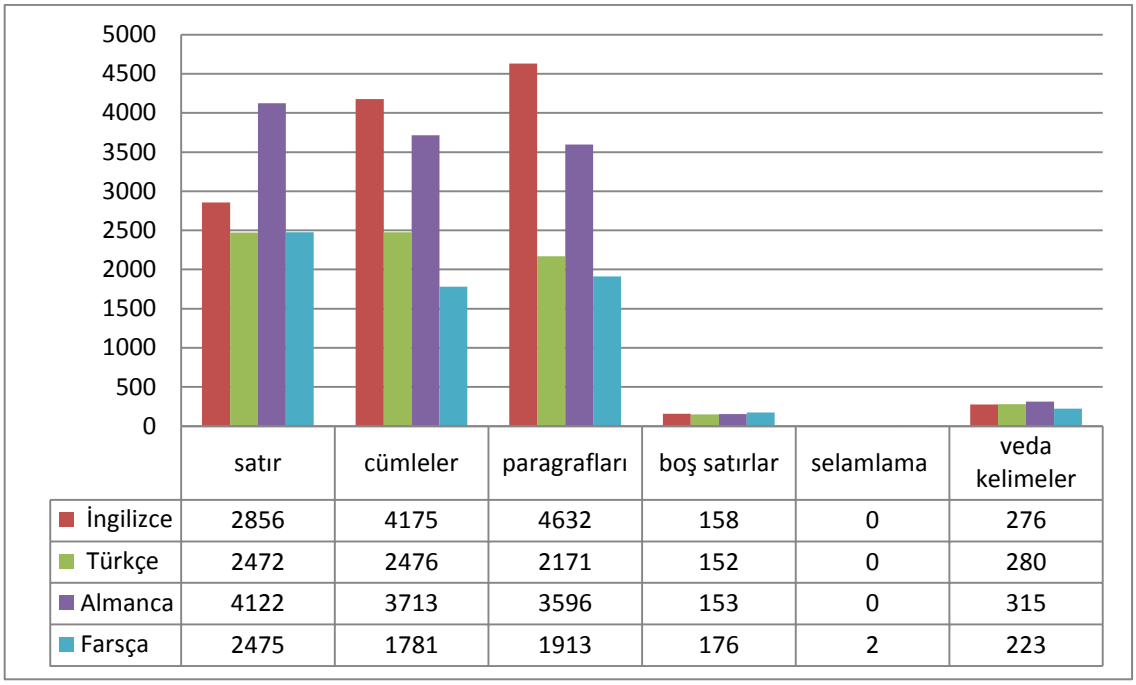
	Article Words	Pro-sentence Words	Pronoun Words	Auxiliary verbs	Conjunction Words	Interjection Words	Adposition Words
İngilizce	5608	89	5904	6441	957	158	11523
Türkçe	4405	53	2739	3659	327	69	7989
Almanca	5829	107	5160	6098	622	204	10829
Farsça	3043	20	1724	2673	221	98	5782

Şekil 6. Fonksiyon kelimelerin İstatistiksel Analizi

Daha önce de belirttiğimiz gibi Fonksiyon Kelimeler, bir yazarı diğerlerinden ayırt etmek için önemli bir ölçüttür. Şekil 6'da farklı Anadildeki Yazarların kendi anadilinin yapısından etkilenerek İngilizce gramer ve kelimeleri farklı bir şekilde kullandığını görürüz. Örneğin; sınıf başına Article words kelimelerinin toplam sayısı ise kelimelerin toplam sayısının oranı İngilizce, Almanca, Türkçe ve Farsça için sırasıyla 0.068, 0.078, 0.083, 0.080'dir. Şekil 6, İngilizce ve Almanca Yazarlar ünlem (Interjection) kelimelerini ve Türkçe ve Farsça yazarlar da cümle öncesi (Pro-sentence) kelimeleri çok daha fazla kullanmışlardır.

3.4.1.4 Yapısal Özellikler

Yapısal özellikler (Structural Features), yazarın tasarladığı metin için yapısal özelliklere dayanan çeşitli türleri kullanmasını ele alır. Şekil 7'de yapısal özelliklerin çeşitlenen en önemli özellikleri gösterilmiştir.



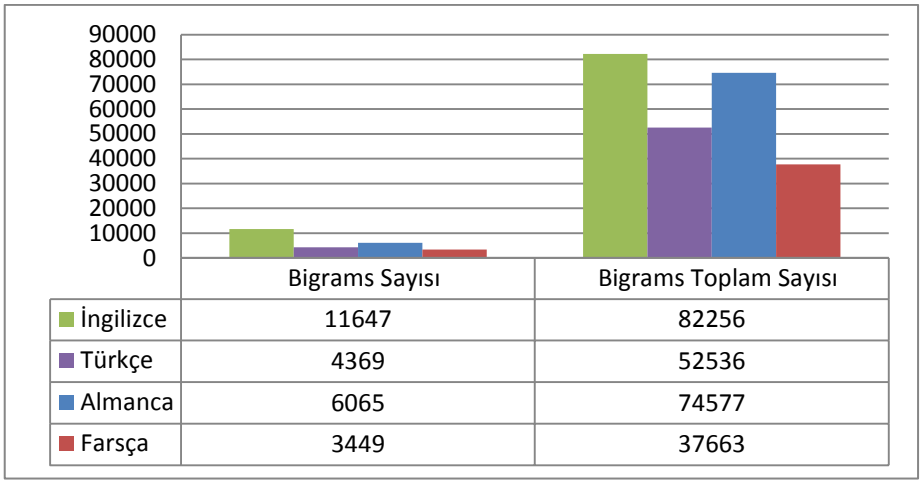
	satır	cümleler	paragrafları	boş satırlar	selamlama	veda kelimeler
İngilizce	2856	4175	4632	158	0	276
Türkçe	2472	2476	2171	152	0	280
Almanca	4122	3713	3596	153	0	315
Farsça	2475	1781	1913	176	2	223

Şekil 7. Yapısal Özelliklerin İstatistiksel Analizi

Şekil 7, Almanların yazarlar yazılarında diğer yazarlar ile karşılaştırılarak daha fazla satır kullanmak olduğunu gösterir, demek ki Alman yazarlar yazılarını düzenleme için satırdan yararlanıyorlar. Diğer taraftan, İngilizce anadili olan yazarların daha fazla paragraf kullanılmaktadırlar. Ayrıca Alman yazarların diğer yazarlara göre daha çok Veda kelimeler kullanmışlar.

3.4.1.5. İçeriğe Özgü Özellikler

İçeriğe Özgü özellikler olarak en az 10 kez görünen 500 Bigram kelime veri kümesi içinde kullanılmıştır. Şekil 8'de her bir sınıf etiketinde Bigram toplam sayısı gösterilmektedir. Sınıf başına kullanılan Bigram kelimelerinin toplam sayısı, toplam Bigram kelimelerinin sayısının oranı İngilizce, Almanca, Türkçe ve Farsça için sırasıyla 0.141, 0.083, 0.081, 0.091'dir.



	Bigrams Sayısı	Bigrams Toplam Sayısı
İngilizce	11647	82256
Türkçe	4369	52536
Almanca	6065	74577
Farsça	3449	37663

Şekil 8. İçeriğe Özgü Özellikler İstatistiksel Analizi

3.7. Sınıflama Modeli

Standart bir sınıflandırma öğrenme işleminde olduğu için çevrimiçi metin koleksiyonu, iki alt kümeye bölünmüştür. İlk alt kümesi, eğitim kümesi denilen ki sınıflandırma modelini (Classification model) eğitmek için kullanılmaktadır. Bu işlemde uygulanan sınıflandırma teknikleri, farklı tahmin gücündeki modellere sebep olabilir. Diğer alt kümesi ise sınıflama modeli tarafından oluşturulan yazarın anadilinin tahmin edilme gücünü doğrulamak için kullanılan deneme kümesidir. Tekrarlanan bir eğitim ve test süreci, iyi bir yazarın anadili tahmin modelini elde etmede gerekli olabilir.

3.8. Yazarın Ana Dilinin Tanımlama

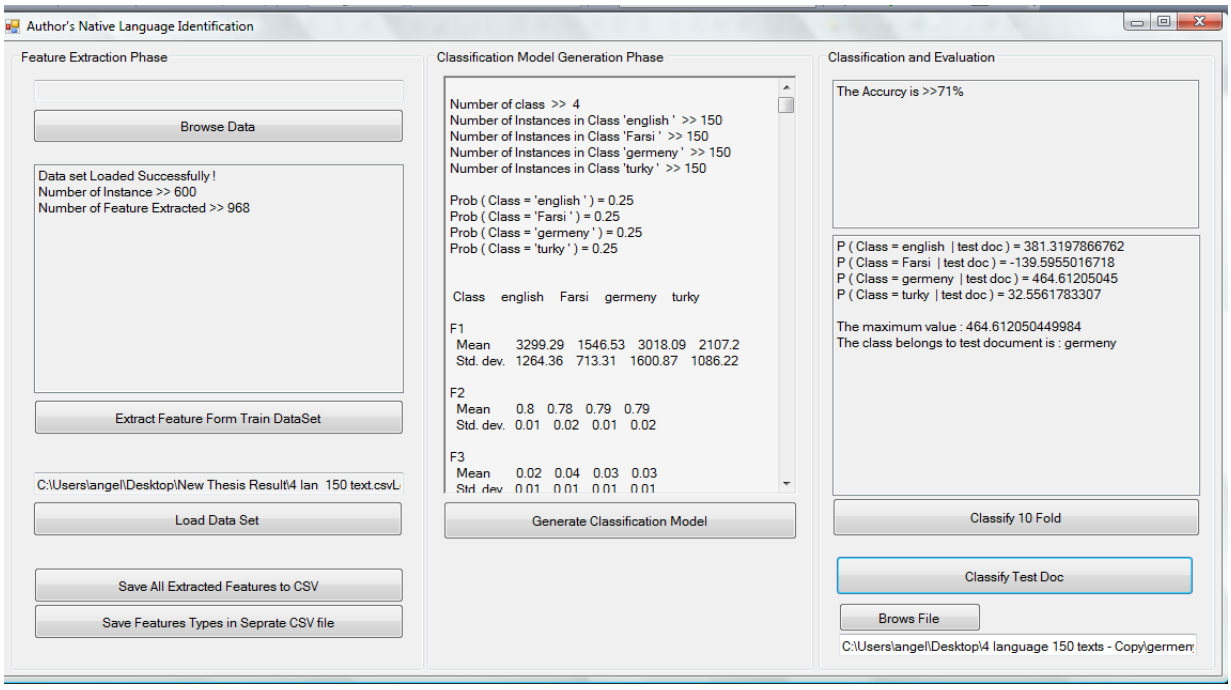
Yazarın Anadilinin Tanımlama modeli geliştirildikten sonar, Sınıflandırma için 10 kat çapraz doğrulama (10-fold cross validation) kullandık, bu verilerin eşit büyüklükteki 10 parçaya bölünmesi anlamına gelir. Sınıflandırma görevi, 10 kez yapılır, her defasında farklı bir parça test verisi olarak kullanılmaktadır. Geriye kalan 9 parça eğitim verisi olarak kullanılır. Bu işlemde her parça yalnızca bir kez test verisi olarak kullanılmaktadır. Bu 10 sınıflandırmanın sonuç ortalaması alınır. Yazarın Anadilini Tanımlama işlemini farklı kümelerden web tabanlı metinler ve yerli olmayan çeşitli İngiliz yazarların ilgili çalışmaları üzerinde geliştirebilir.

3.9. Yazarın Anadilinin Tanımlama Geliştirme

Bu bölümde, Yazarın Anadili Tanımlaması için önerilen çerçevenin geliştirilmesi analizi edilecektir. Bu uygulamada programlama dili olarak Visual C#. Net (Sürüm 2010) tasarlanmıştır.

3.9.1. Grafik Kullanıcı Arayüzü

Geliştirilen uygulamanın arayüzü Şekil 9'de verilmiştir. Uygulama aşağıdaki gibi üç birim veya aşmadan oluşur:



Şekil 9. Uygulamamızın Arayüzü

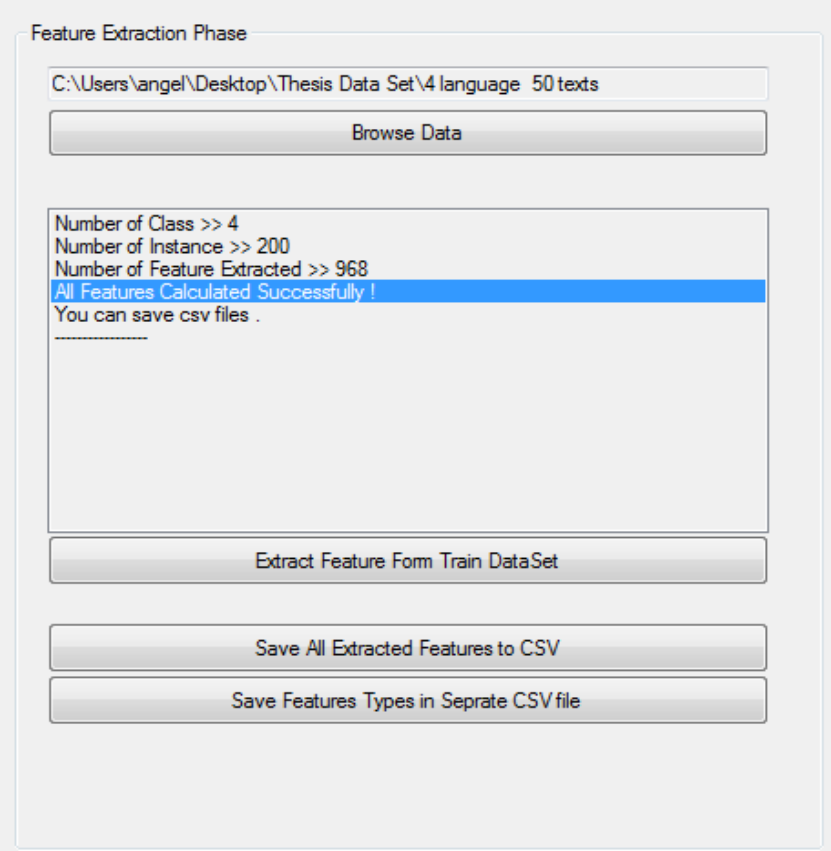
Aşama 1: Özellik Çıkarma

Aşama 2: Sınıflandırma Model Oluşturma

Aşama 3: Sınıflandırma ve Sonuçların değerlendirilmesi

3.9.1.1. Özellik Çıkarma Aşaması

Bu bölümde, Veri Kümemize yükleme yapabiliriz. Bu teklif için Veri Yükleme (Browse Data) düğmesi veri yükleme uygulamasında bize imkan sağlar. Veri Yükleme düğmesine tıkladığımızda, bir iletişim kutusu yerel dosya sisteminde veri dosyası için yüklemeye izin verir.



Şekil 10. Özellik Çıkarma Aşaması

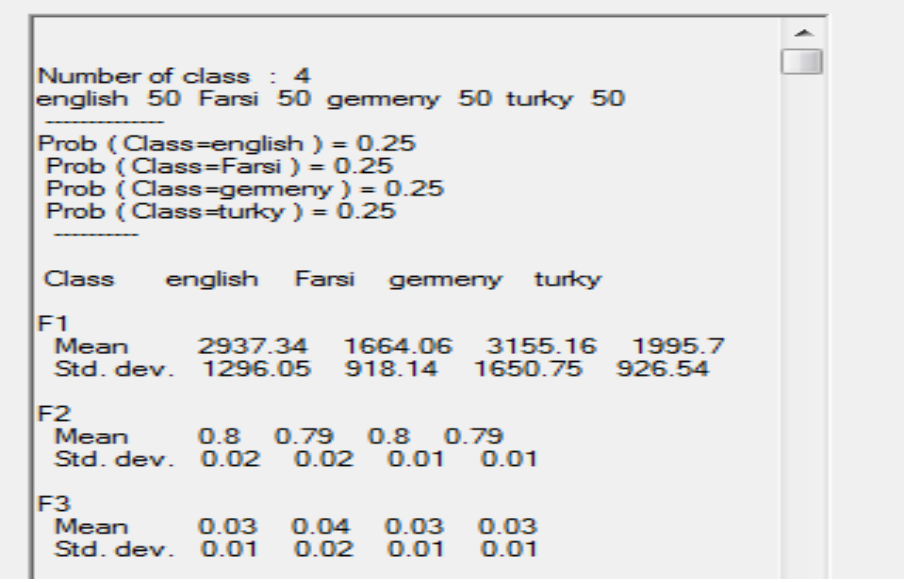
Başarılı bir veri yüklemesinden sonra, bölüm 2.4'te bizim açıkladığımız önerilen özellikleri seçebiliriz. Bu işlem uygulamak için kullanılan düğme (Features Extract) ve sonuçlar panelde incelenecektir. Özellik setlerinin her dört türü için hesaplama özelliğini bitirdikten sonra Panel, diller için veri seti içeren sınıfın sayısını ve adını, örneğin numarasını (metin dosyaları) ve bizim uygulama yüklediğimiz veri setlerinden seçilen özelliklerin sayısını gösterecektir. Şekil 10'de Özellik Çıkarımının Örnek Sonuçları bulunmaktadır.

Özellikleri değerlendirmek ve diğer makine öğrenme araçlarında kullanmak için Kaydetme düğmesini kullanarak bir CSV dosyasındaki özellik vektörünü kaydedebiliriz. CSV dosyalarında özellik setlerinin farklı türlerini kurtarmak için başka bir düğme bulunmaktadır.

3.9.1.2. Naïve Bayes Sınıflandırıcı İçin Sınıflandırma Modeli Oluşturma Aşaması

Önceki aşamada, özellik vektörü oluşturulur ve oluşturulan vektör sınıflandırma modeli oluşturmak için hazırlanaır. Naïve Bayes Sınıflandırıcı için sınıflandırma modeli üretmede her özellik ve sınıf için Matris ve Standart Sapma (standard deviation) Matrisin ortalaması bulunmalıdır.

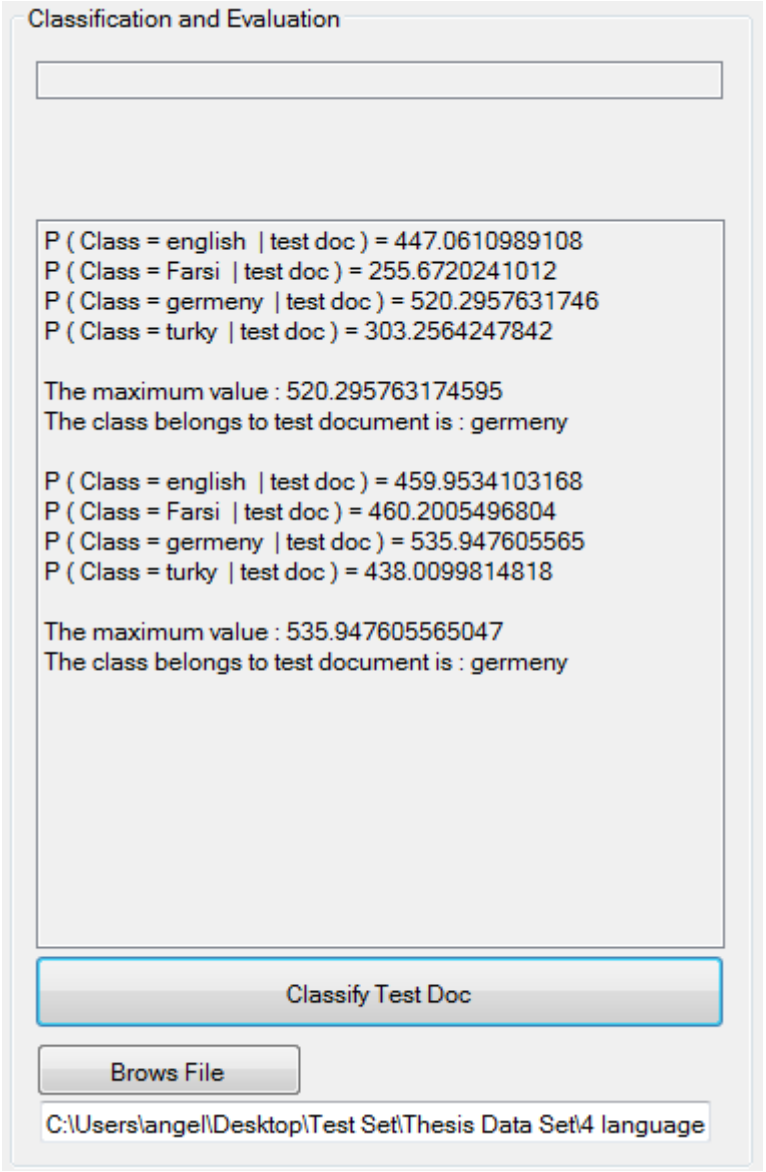
Şekil 11'de, her sınıfta özellikler için ortalama ve standart sapmanın örnek sonuçlarını göstermektedir. Bu örnekte İngilizce, Almanca, Farsça, Türkçe gibi dört dil ayıklanmış ve her sınıf 50 örnek (metin dosyası) içerir. Bunların hepsinin olasılığı aynı olacaktır ve bu değer 0.25'dir.



Şekil 11. Özellikler İçin Ortalama ve Standart Sapmanın Hesaplanması

3.9.1.3. Sınıflandırma ve Sonuçların Değerlendirmesi Aşaması

Son aşama sınıflandırma işlemi ve sonuçların değerlendirilmesidir. Buna ulaşmak için daha açıkladığımız 10 kat çapraz doğrulama kullanılacağız. Şekil 12'de hesaplanan sonuçlar ve değerlendirmelerin panelini göstermektedir. Özellik değerlerinin hesaplanmasında Gauss (normal) dağılımı kullanılır ve her belgenin olasılığını hesaplayarak bir sınıf etiketi için daha yüksek olasılığa sahip olan seçilir.



Şekil 12. Sınıflandırma İşleminin Sonuçları

3.10. Weka Aracı

WEKA aracı, araştırmacılar ve uygulayıcılar için veri ön işlemenin ve makine öğrenme algoritmalarının kapsamlı bir koleksiyonu sağlamayı amaçlamaktadır. Bu, kullanıcıların hızlı bir şekilde denemesini ve yeni veri setleri üzerinde farklı makine öğrenme yöntemlerini karşılaştırmasını sağlar. Wekanın modüler olması, genişleyebilen yapının sağlanan temel öğrenme algoritmaları ve araçlarının geniş bir koleksiyondan yapılması için gelişmiş veri madenciliğine imkan sağlar. WEKA grafik kullanıcı arayüzleri ile yeni öğrenme algoritmaları entegrasyonu otomatik hale getiren eklenti mekanizmaları ve kuruluşlar basit bir API sayesinde aracın genişletilmesi kolaylaştırmıştır.

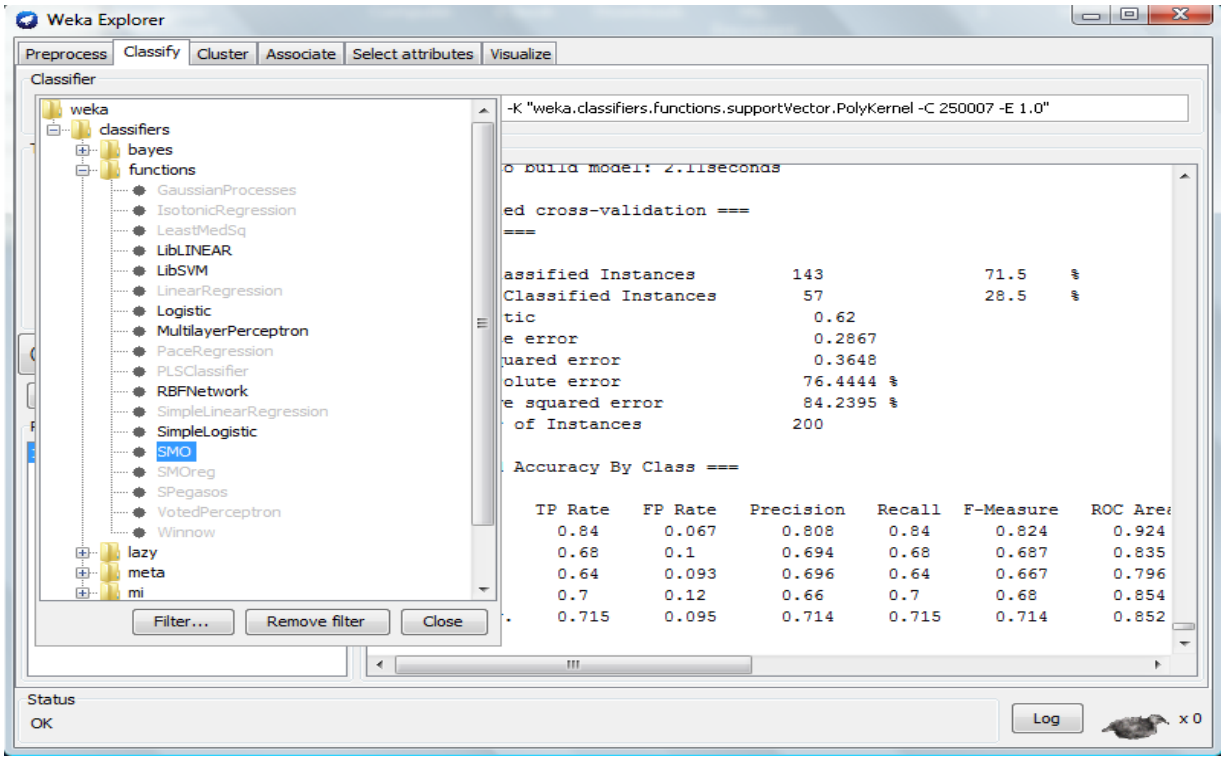
Weka aracı, sınıflandırma kümeleme, birliktelik kuralı madenciliği ve özellik seçme algoritmaları için regresyon içermektedir. Verilerin ön araştırması, veri görselleştirme imkanları ve birçok ön işleme araçlarıyla iyi bir şekilde desteklenebilir. Bilgi Analizi için Waikato Çevre (WEKA), birleşik bir aracın algılamaya ihtiyaç duyulmasıyla gerçekleşir. Araştırmacılar makine öğrenmede yapılan tekniklere kolayca erişim olanağı verecektir. 1992 yılında projenin kurulduğu zaman öğrenme algoritmaları, farklı platformlarda üzerinde kullanmak için ve verilerin çeşitli biçimleri işletilen farklı dillerde bulunmamaktaydı. Veri setleri koleksiyonu üzerine karşılaştırmalı bir çalışma yapmak için öğrenme planlarını toplamak temel bir işlemdir.

WEKA'nın sadece öğrenme algoritmaları için bir araç teşkil etmeyeceğini aynı zamanda araştırmacıların veri işleme ve değerlendirme planı için altyapı desteğine ihtiyaç duymaksızın yeni algoritmaların yaptığı bir çerçeve olduğu düşünülebilir. Şimdilerde WEKA veri madenciliği ve makine öğrenimi açısından bir dönüm noktası olarak tanınmaktadır. Bu, akademi ve iş çevrelerinde yaygın olarak kabul görmüş ve veri madenciliği araştırmalar için yaygın olarak kullanılan bir araç haline gelmiştir.

3.10.1. Weka ile Veri Sınıflandırma

Bu aşamaya gelmeden önceden özellik vektörleri haline getirilen metin dokümanları CSV formatında kaydedilerek kolayca sınıflandırma işlemi uygulanmıştır. Bir önceki aşamada bahsettiğimiz 600 dokümanı içeren veri dosyasını sınıflandırmak için veri dosyası yüklendikten sonra sınıflandırma işlemi için "Classify" sekmesi tıklanır.

Aşağıdaki ekran görüntüsü şeklinde de görüldüğü gibi, listeden bir sınıflandırma algoritması seçilir. Örnekte SMO sınıflandırma algoritması seçilmiştir ve Şekil 13'te göstermektedir.



Şekil 13. Weka Aracının Veri Sınıflandırma Arayüzü

Sınıflandırma algoritması "SMO" olarak seçildikten sonra, algoritma ile ilgili test seçenekleri belirlenir. Bu seçenekler, sınıflandırıcı algoritmayı seçtiğimiz kısmın hemen altında "Test options" kısmında verilmektedir. Yazarın Anadilini Tanımlamayı gerçekleştirmek için çeşitli sınıflandırma algoritmaları kullanabilir. Bu çalışmada, popüler ve güçlü üç sınıflandırıcı: C4.5 karar ağaçları, Naïve Bayes ve SMO kabul etmiştir.

WEKA'da veri madenciliği araçları, standart bir C4.5 algoritması uygulanmaktadır. C4.5, ID3 algoritmasının bir uzantısı, bölme ve fethetme stratejisini ve entropi ölçüsünü nesnenin sınıflandırılması için benimseyen Quinlan [27] tarafından geliştirilen bir karar ağacı algoritmasıdır. Onun amacı karışık nesneleri onların ilgili sınıflarına yani nesnelerin özellikler değerlerine dayanarak sınıflandırmaktır.

Wekada bir sonraki sınıflandı olarak SMO seçilmiştir. Destek vektörü sınıflandırıcı için Platt'inin (1999) Sequential Minimal Optimization algoritması uygulanmıştır [20]. Bu

uygulama genel olarak tüm kayıp değerleri değiştirir ve nominal özellikleri binary haline dönüştürür. Aynı zamanda tüm özellikleri varsayılan olarak normalleştirir. Çoklu sınıf problemleri, kullanılan çokterimli çekirdek sınıflandırma modeli kurulduğu zaman tek-karşısında-hepsi (one against all) yöntemi kullanılarak çözülmüştür.

4. SONUÇLAR

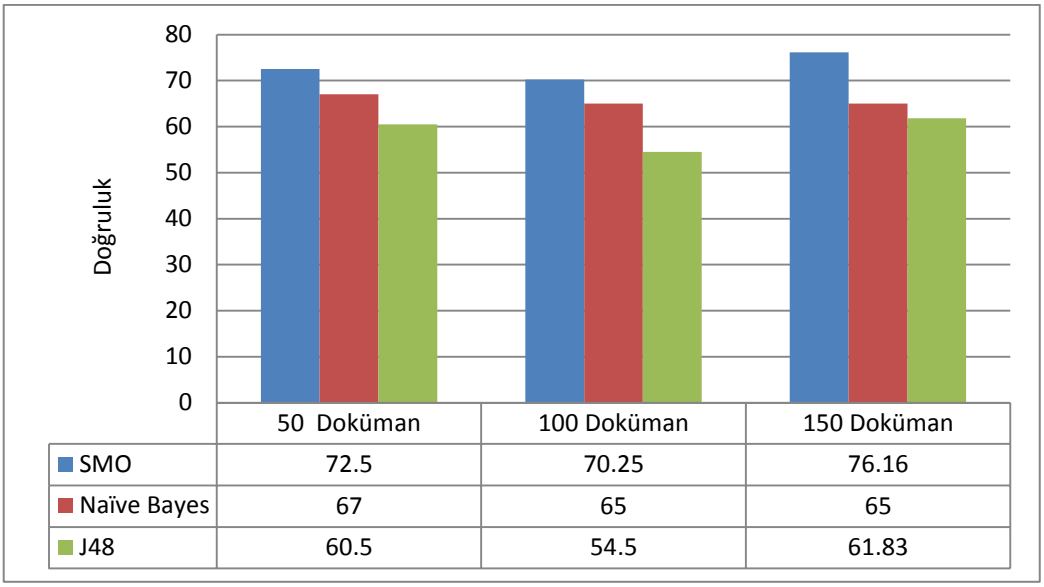
4.1. Giriş

Bu bölümde gerçeklenen uygulamaların sonuçlarını hem de önemlerine bağlı olarak tartışılmasını sunulacaktır. İlk olarak, geliştirilen uygulamanın sonuçları ve değerlendirmek farklı önlemler sınıflandırma sonuçları verilecektir. Bura farklı özellik kümeleri sonuçlarını tartışılacak ve Karar ağacı, Destek Vektör Makinesi ve Naive Bayes sınıflandırıcıları performansları karşılaştırılacaktır. Böylece özellik kümlerinin farklı etkisini ve veri kümesinin ve metin sayısı analiz edilecektir. Bu bölümdeki son uygulama olarak 3 yabancı dilden birini ve ikisini dışlayarak, üç ve iki dil için elde edilen sonuçları değerlendirilmiştir.

4.2. Sınıf Başına Doküman Sayısının Karşılaştırılması ve Sınıflandırma Teknikleri

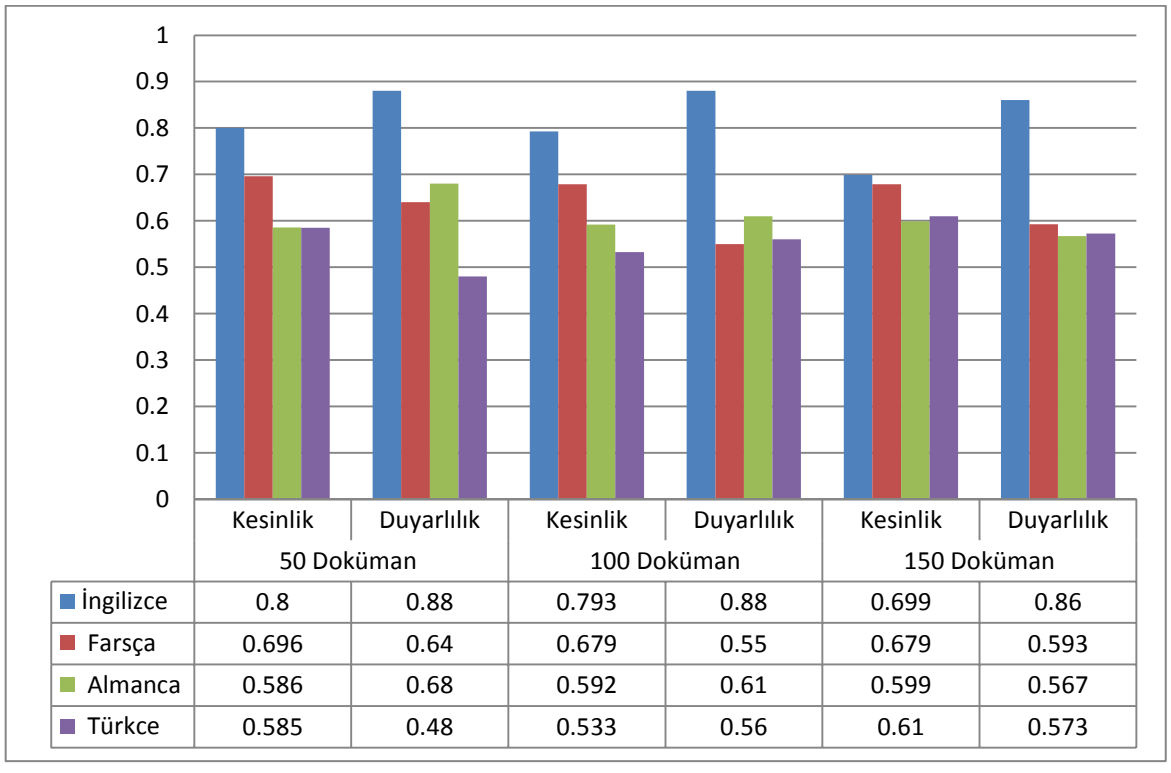
Sınıflandırma teknikleri performansları (Naïve Bayes, C4.5 ve SMO) karşılaştırmak için Yazarın Anadil Tanımlaması doğruluğunu ölçmek için sınıflandırıcıları ayrı ayrı uygulanmıştır. Farklı boyutlu veri kümesi için ve her dört tip özellikler için kullanıldı. Sonuçlar Şekil 14'de gösterildiği gibi SMO diğer iki yöntemden daha iyi performans göstermektedir.

Şekil 14'de veri kümesinin boyutuna göre sınıflandırma tekniklerinin performans değişimleri gösterilmektedir. SMO sınıflandırıcı, tarafından üretilen en iyi sınıflandırma sonucu 150 metinde oluşan veri kümesi için %76.16 doğruluk sağlamaktadır. J48 sınıflandırıcı performansı üzerinde önemli bir değişim var.



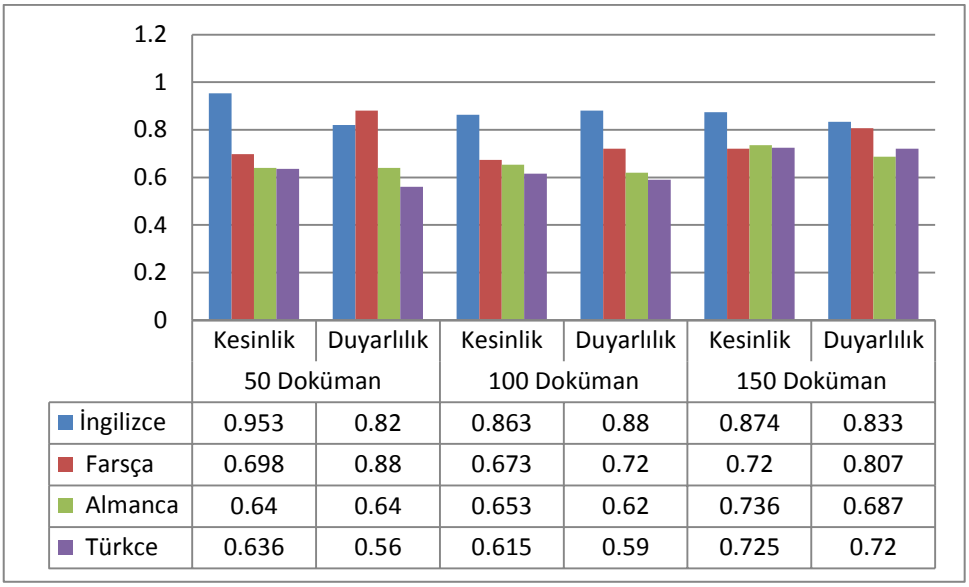
	50 Doküman	100 Doküman	150 Doküman
SMO	72.5	70.25	76.16
Naïve Bayes	67	65	65
J48	60.5	54.5	61.83

Şekil 14. Sınıflandırma Teknikleri Doğruluk Performanslarının karşılaştırılması



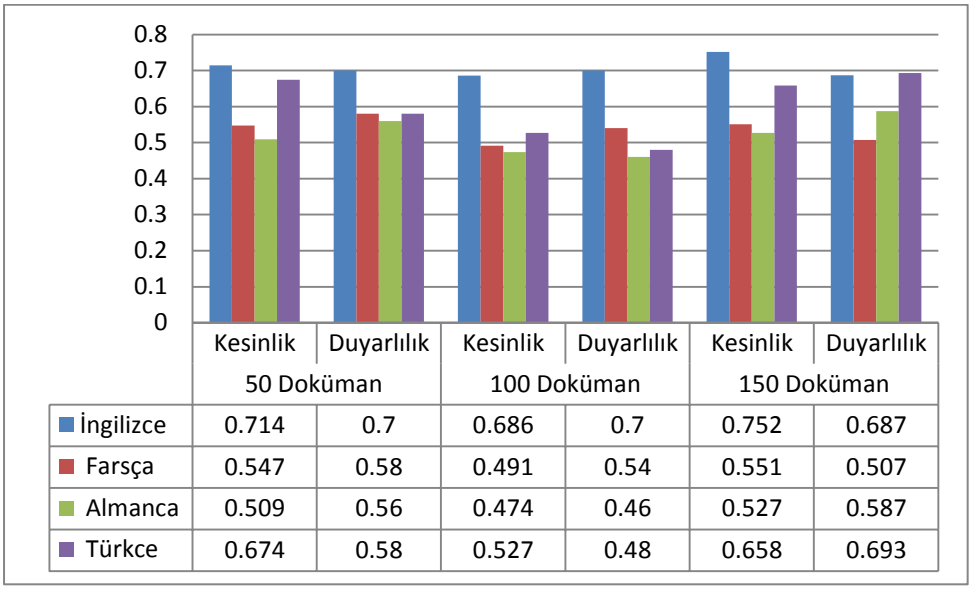
	Kesinlik	Duyarlılık	Kesinlik	Duyarlılık	Kesinlik	Duyarlılık
	50 Doküman		100 Doküman		150 Doküman	
İngilizce	0.8	0.88	0.793	0.88	0.699	0.86
Farsça	0.696	0.64	0.679	0.55	0.679	0.593
Almanca	0.586	0.68	0.592	0.61	0.599	0.567
Türkce	0.585	0.48	0.533	0.56	0.61	0.573

Şekil 15. Naïve Bayes Sınıflandırıcının Kesinlik ve Duyarlılık ölçümleri



	Kesinlik	Duyarlılık	Kesinlik	Duyarlılık	Kesinlik	Duyarlılık
	50 Doküman		100 Doküman		150 Doküman	
İngilizce	0.953	0.82	0.863	0.88	0.874	0.833
Farsça	0.698	0.88	0.673	0.72	0.72	0.807
Almanca	0.64	0.64	0.653	0.62	0.736	0.687
Türkce	0.636	0.56	0.615	0.59	0.725	0.72

Şekil 16. SMO Sınıflandırıcın Kesinlik ve Duyarlılık ölçümleri



	Kesinlik	Duyarlılık	Kesinlik	Duyarlılık	Kesinlik	Duyarlılık
	50 Doküman		100 Doküman		150 Doküman	
İngilizce	0.714	0.7	0.686	0.7	0.752	0.687
Farsça	0.547	0.58	0.491	0.54	0.551	0.507
Almanca	0.509	0.56	0.474	0.46	0.527	0.587
Türkce	0.674	0.58	0.527	0.48	0.658	0.693

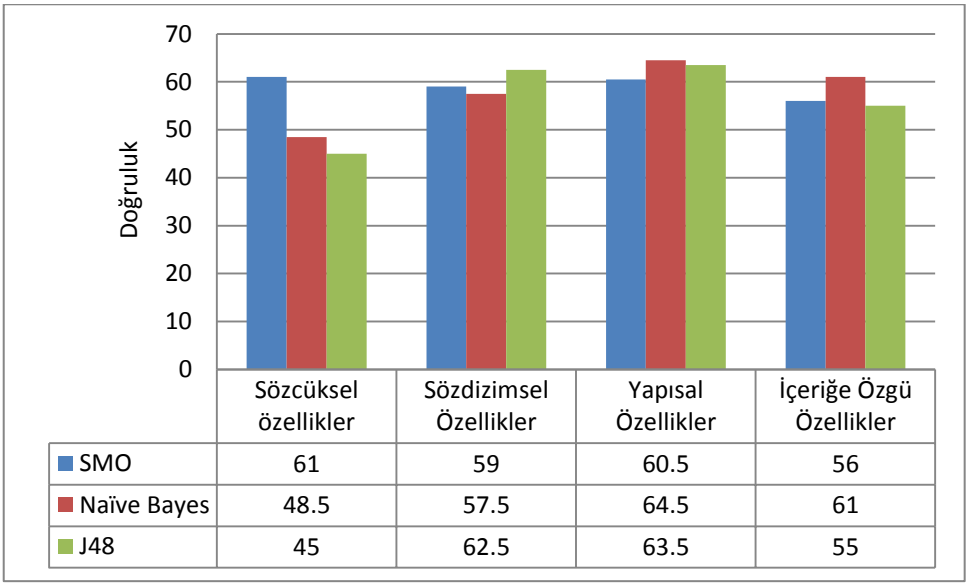
Şekil 17. J48 Sınıflandırıcıyla Kesinlik ve Duyarlılık ölçümleri

4.3. Özellik Tiplerinin Etkisi

Bu çalışmada, bir yazarın anadilinin tanımlanmasının doğruluğu, önerilen özellik türlerinin önemine ve kullanılan sınıflandırma tekniklerinin performansına bağlı olarak araştırılmıştır. Şekil 18'de bir yazarın anadiline belirlenmesinde kullanılan özellik türlerinin katkısı değerlendirilmiş ve gösterilmiştir.

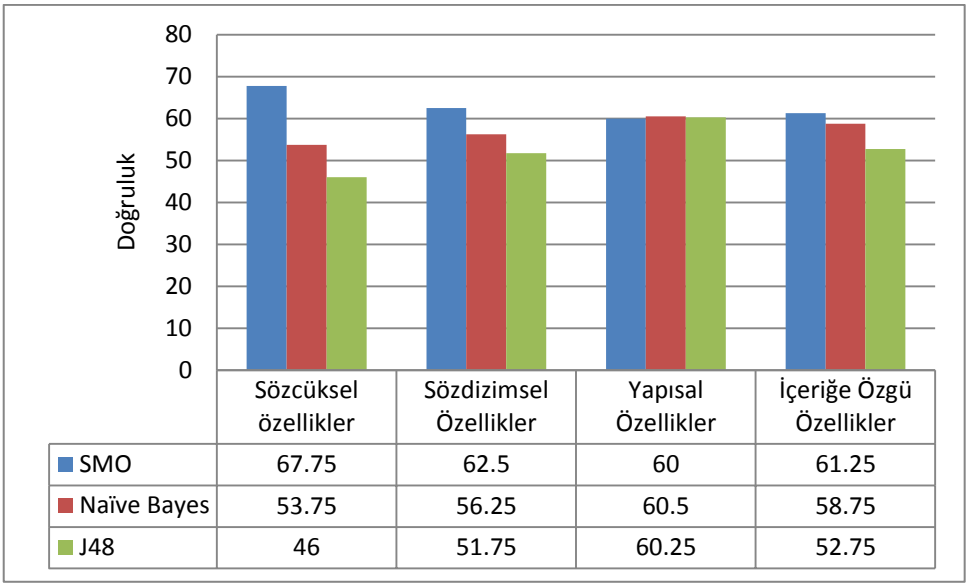
Sözlüksel ve sözdizimsel özellikleri anadil tanımlamada iyi bir ayrıt edicilik yeteneği göstermektedir [27]. Özellik kümesi olarak az sayıda yapısal özellikleri kullanılmasına rağmen, bu özelliklerin sınıflandırma performansına önemli katkıları vardır. Bu durum anadili aynı olan yazarların tutarlı yazım biçimlerinin yapısal özelliklere yansıtıldığını göstermektedir.

Her sınıftan 100 metinle yaptığımız testlerde, bütün sınıflandırıcılar için iyi bir doğruluk seviyesi elde edilmiştir. Farklı boyutlarda üç farklı veri kümesi için SMO sınıflandırıcı en yüksek doğruluğu vermiştir. Naïve Bayes yapısal özellikler için 64.5 doğruluk ile en iyi performansı sağlamaktadır. Sırasıyla 50, 100 ve 150 doküman üzerinde yapılan test sonuçları Şekil 18, 19, ve 20 de verilmiştir.



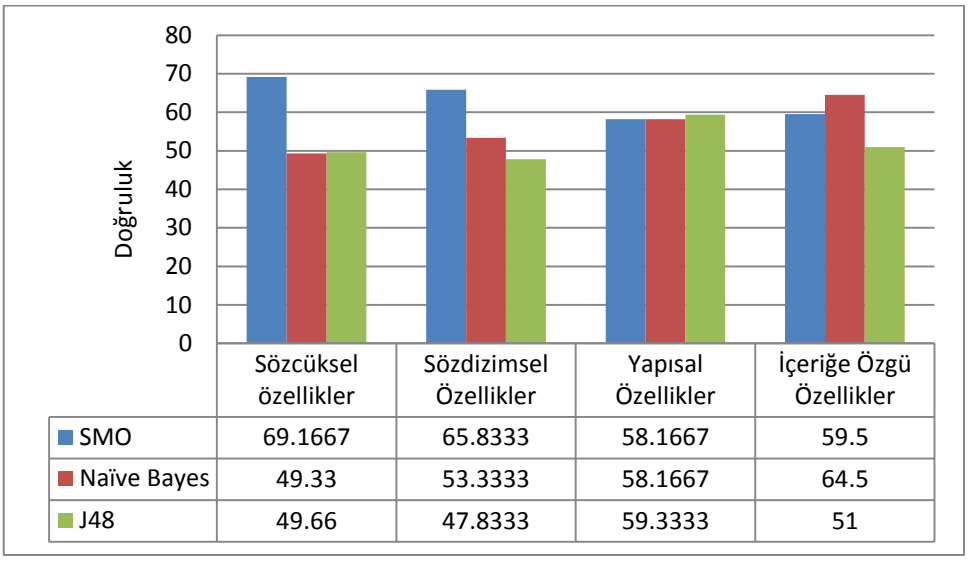
	Sözcüksel özellikler	Sözdizimsel Özellikler	Yapısal Özellikler	İçeriğe Özgü Özellikler
SMO	61	59	60.5	56
Naïve Bayes	48.5	57.5	64.5	61
J48	45	62.5	63.5	55

Şekil 18. 50 doküman için özellikler Tipleri Etkisi ve Sınıflandırma Teknikleri Performanslarını Doğruluk ölçümleri sonuçları



	Sözcüksel özellikler	Sözdizimsel Özellikler	Yapısal Özellikler	İçeriğe Özgü Özellikler
SMO	67.75	62.5	60	61.25
Naïve Bayes	53.75	56.25	60.5	58.75
J48	46	51.75	60.25	52.75

Şekil 19. 100 doküman için özellikler Tipleri Etkisi ve Sınıflandırma Teknikleri Performanslarını Doğruluk ölçümleri sonuçları



	Sözcüksel özellikler	Sözdizimsel Özellikler	Yapısal Özellikler	İçeriğe Özgü Özellikler
SMO	69.1667	65.8333	58.1667	59.5
Naïve Bayes	49.33	53.3333	58.1667	64.5
J48	49.66	47.8333	59.3333	51

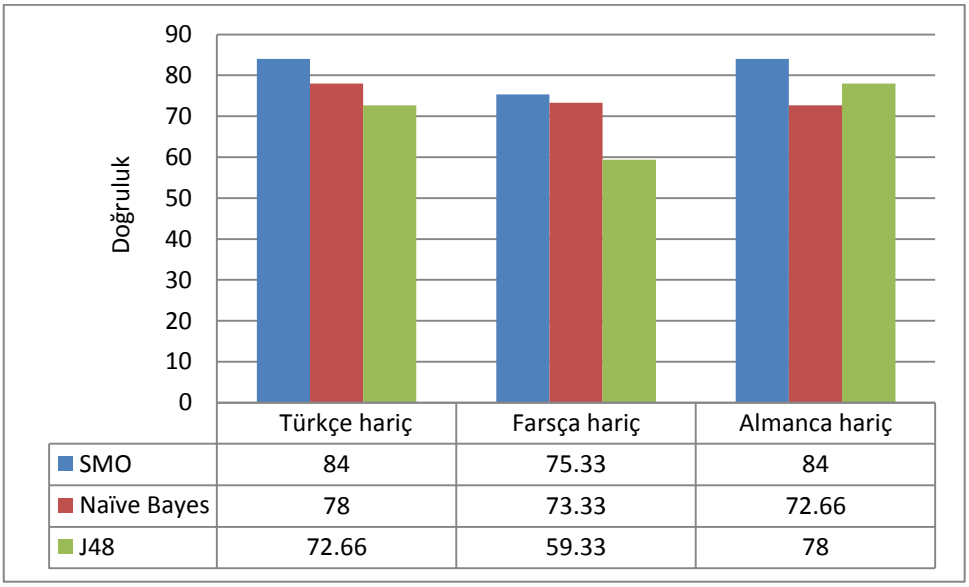
Şekil 20. 150 doküman için özellikler Tipleri Etkisi ve Sınıflandırma Teknikleri Performanslarını Doğruluk ölçümleri sonuçları

4.4. Kullanılan Dil Sayısının Etkisinin Değerlendirilmesi

Bu bölümde veri kümesindeki dil sınıfları sayısının etkisi değerlendirilmiştir. Uygulamada, adım adım ve her bir adımda test için kullanılan dillerin sayısı azaltılmıştır.

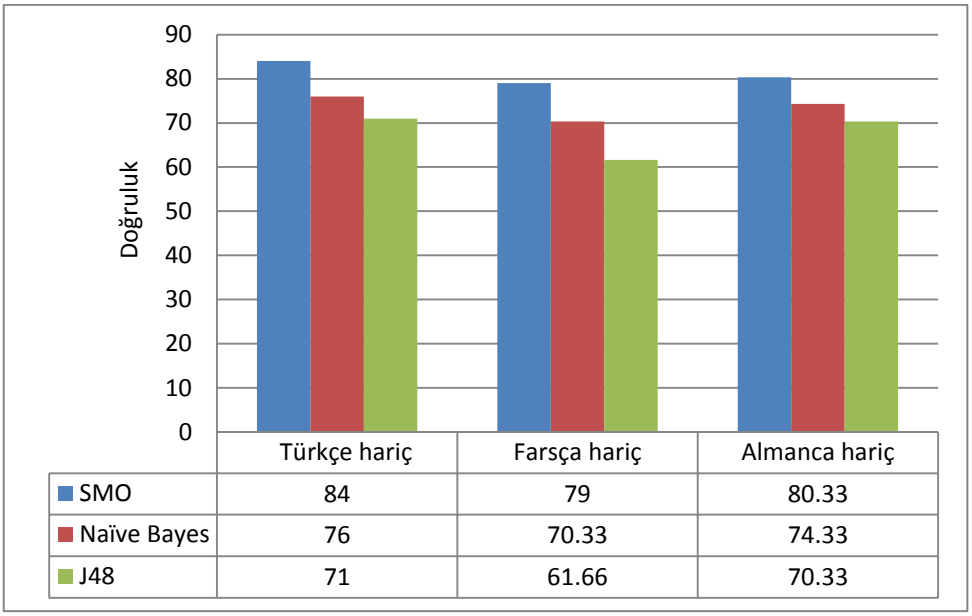
Şekil 21, 22 ve 23'te; test sonuçları verilmiştir. Şekillerde verildiği gibi, veri kümesinden Türkçe ve Farsça örneklerin çıkartılması sistem performansını önemli ölçüde artırmaktadır. Burada Alman kökenli yazarların İngiliz kökenli yazarların daha benzer yazdıklarını ima etmektedir. Dolayısıyla bir dilin çıkartılması ile kalan dillerden yazarların arasından daha iyi performansla anadil ayırt edilebilmektedir.

150 veriden oluşan bir veri kümesi kullanılması durumunda, en iyi doğruluk elde edilmiş ve üç sınıflandırma teknikleri performansın önemli ölçüde arttığı görülmüştür. Naive Bayes, J48, ve SMO sınıflandırıcıların performans artışları Şekil 23'de verilmiştir.



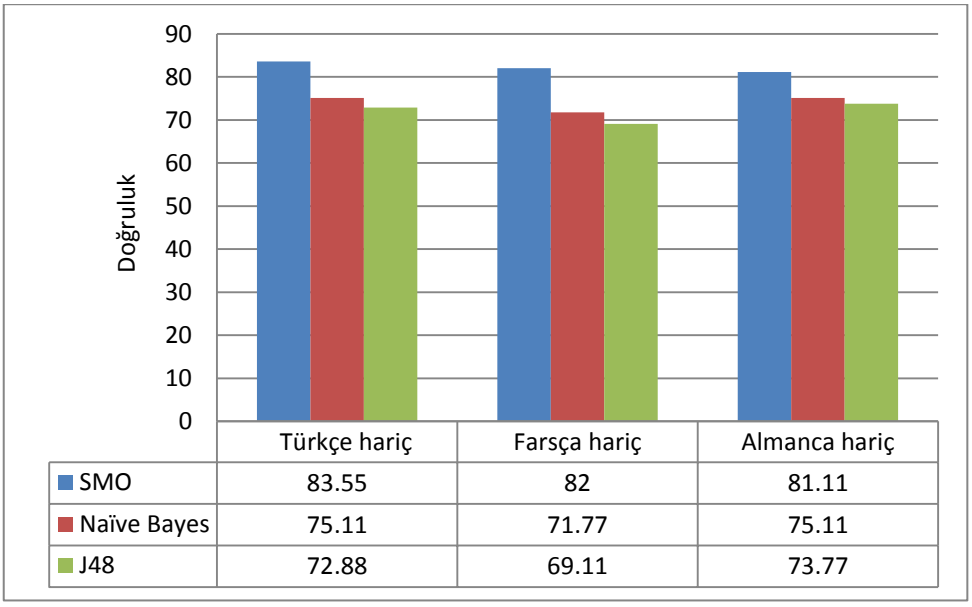
	Türkçe hariç	Farsça hariç	Almanca hariç
SMO	84	75.33	84
Naïve Bayes	78	73.33	72.66
J48	72.66	59.33	78

Şekil 21. 50 Dokümandan oluşan veri kümesi için bir dil hariç tutulması durumunda Sınıflandırma Doğruluklarının karşılaştırılması



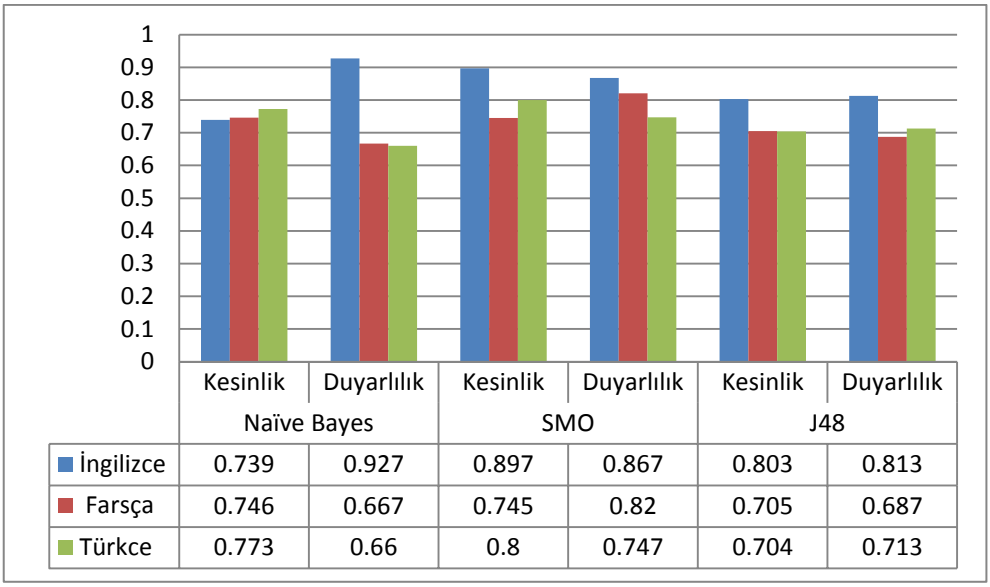
	Türkçe hariç	Farsça hariç	Almanca hariç
SMO	84	79	80.33
Naïve Bayes	76	70.33	74.33
J48	71	61.66	70.33

Şekil 22. 100 Dokümandan oluşan veri kümesi için bir dil hariç tutulması durumunda Sınıflandırma Doğruluklarının karşılaştırılması



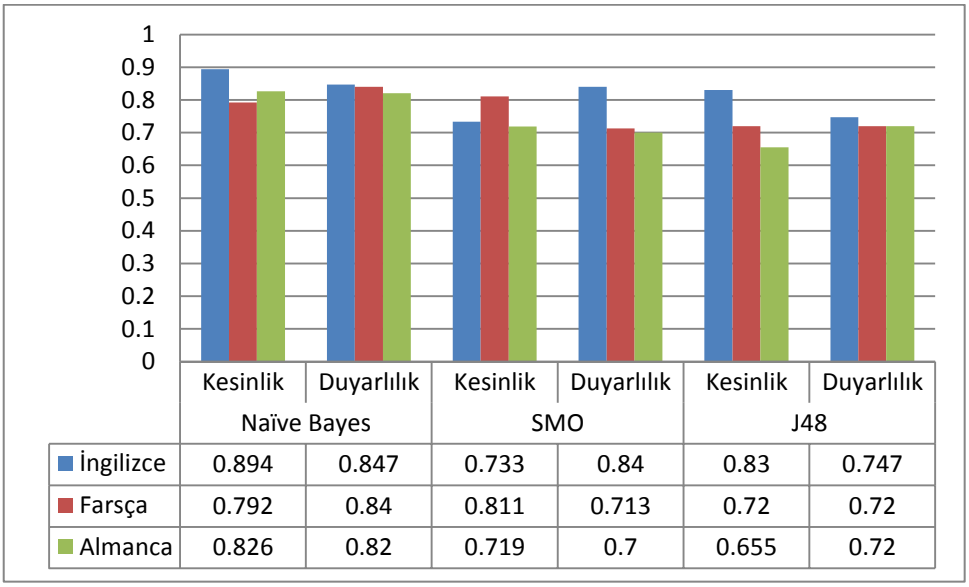
	Türkçe hariç	Farsça hariç	Almanca hariç
SMO	83.55	82	81.11
Naïve Bayes	75.11	71.77	75.11
J48	72.88	69.11	73.77

Şekil 23. 150 Dokümandan oluşan veri kümesi için bir dil hariç tutulması durumunda Sınıflandırma Doğruluklarının karşılaştırılması



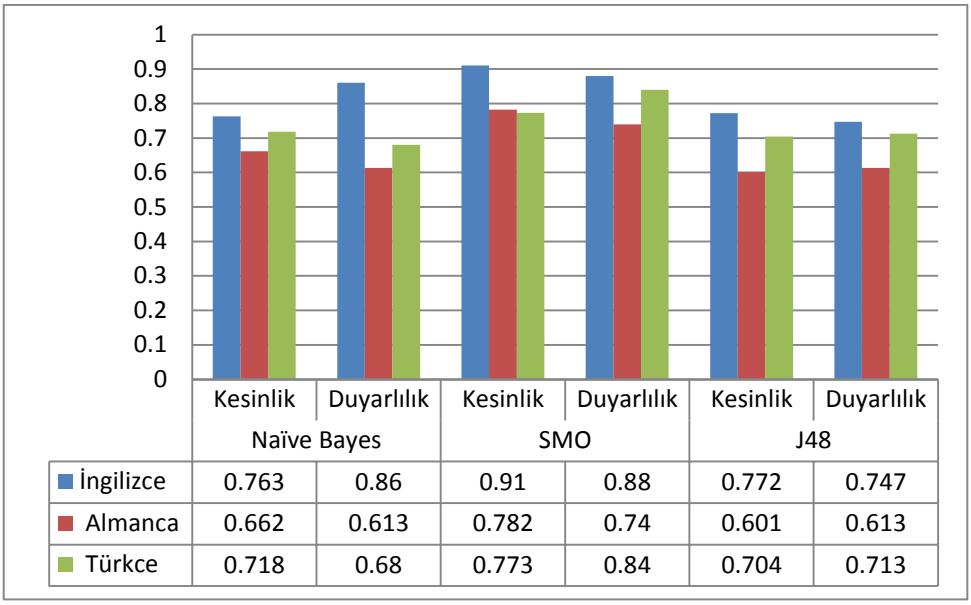
	Naïve Bayes		SMO		J48	
	Kesinlik	Duyarlılık	Kesinlik	Duyarlılık	Kesinlik	Duyarlılık
İngilizce	0.739	0.927	0.897	0.867	0.803	0.813
Farsça	0.746	0.667	0.745	0.82	0.705	0.687
Türkce	0.773	0.66	0.8	0.747	0.704	0.713

Şekil 24. Üç dil (Almanca hariç) için 150 Doküman üzerindeki Kesinlik ve Duyarlılık ölçüm sonuçları



	Naïve Bayes		SMO		J48	
	Kesinlik	Duyarlılık	Kesinlik	Duyarlılık	Kesinlik	Duyarlılık
İngilizce	0.894	0.847	0.733	0.84	0.83	0.747
Farsça	0.792	0.84	0.811	0.713	0.72	0.72
Almanca	0.826	0.82	0.719	0.7	0.655	0.72

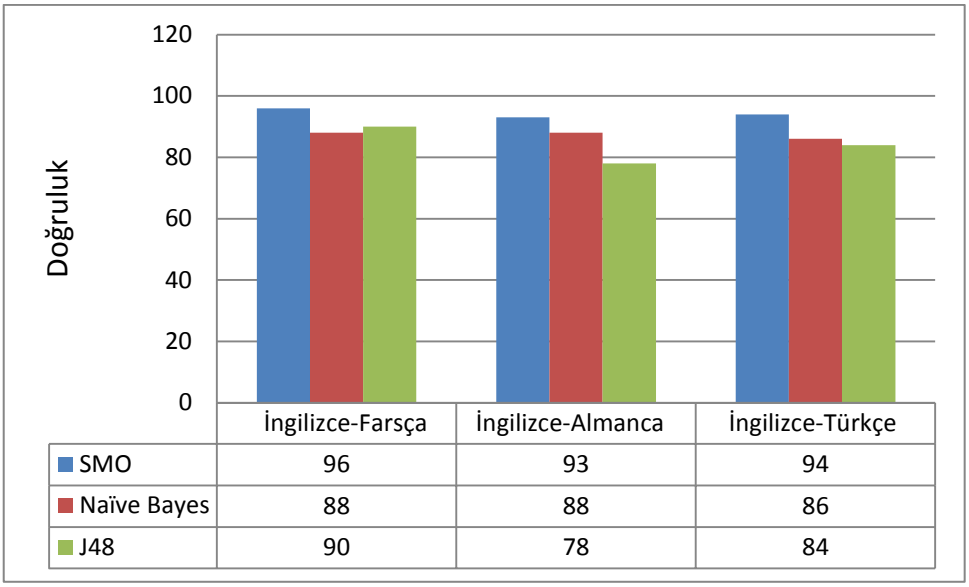
Şekil 25. Üç dil (Türkce hariç) için 150 Doküman üzerinde Kesinlik ve Duyarlılık ölçüm sonuçları



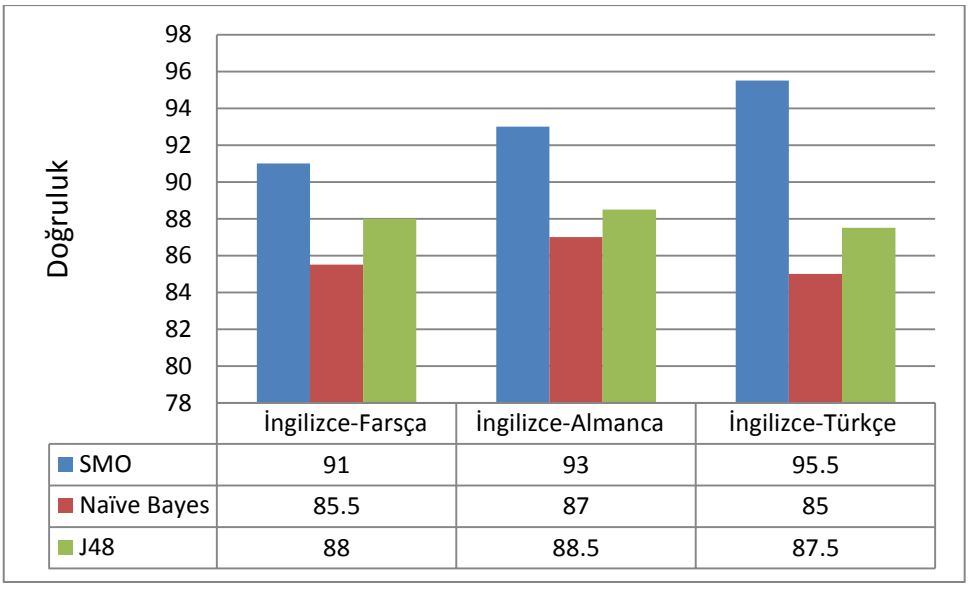
	Naïve Bayes		SMO		J48	
	Kesinlik	Duyarlılık	Kesinlik	Duyarlılık	Kesinlik	Duyarlılık
İngilizce	0.763	0.86	0.91	0.88	0.772	0.747
Almanca	0.662	0.613	0.782	0.74	0.601	0.613
Türkce	0.718	0.68	0.773	0.84	0.704	0.713

Şekil 26. Üç dil (Farsça hariç) için 150 Doküman üzerinde Kesinlik ve Duyarlılık ölçüm sonuçları

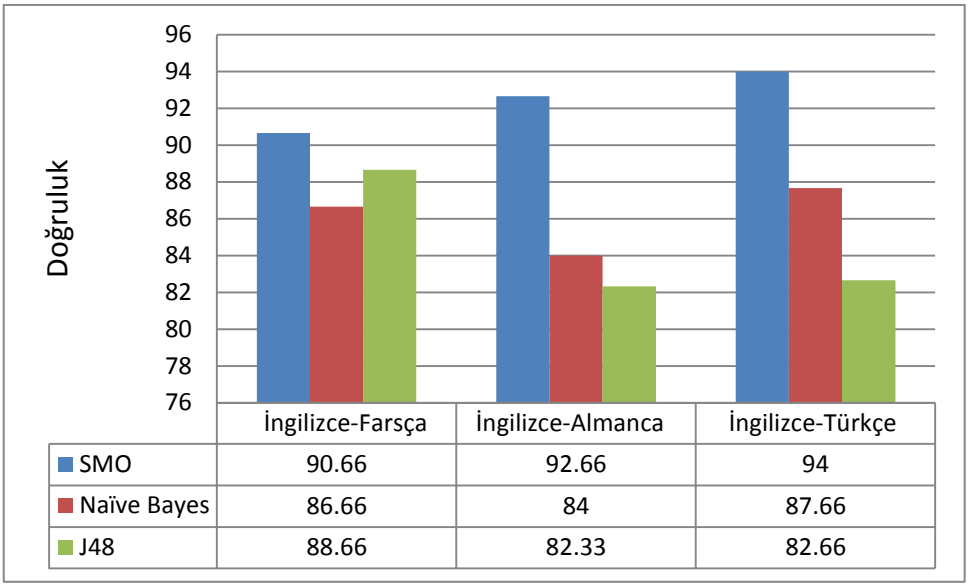
Şekil 27, 28 ve 29'da, dil sayısı ikiye düşürülmüş ve her aşamada bir dil İngiliz köken yazarlarla karşılaştırılmıştır. Sonuçlara göstermektedir ki veri kümesinden iki dil çıkartılması durumunda sistem performansını önemli ölçüde artırmaktadır. Elde edilen sonuçlara Alman kökenli yazarlarla İngiliz kökenli yazarların daha benzer yazdıklarını ima etmektedir. Ancak, Fars kökenli yazarlar ile diğer kökenli yazarlar arasında en düşük benzerlik olduğu görülmüştür. Dolayısıyla, iki dilin çıkartılması durumunda, kalan dillerden İngilizce yazarların ana dilleri daha iyi performansla ayırt edilebilmektedir.



Şekil 27. 50 Dokümandan oluşan veri kümesi için iki dil tutulması durumunda Sınıflandırma Doğruluklarının karşılaştırılması

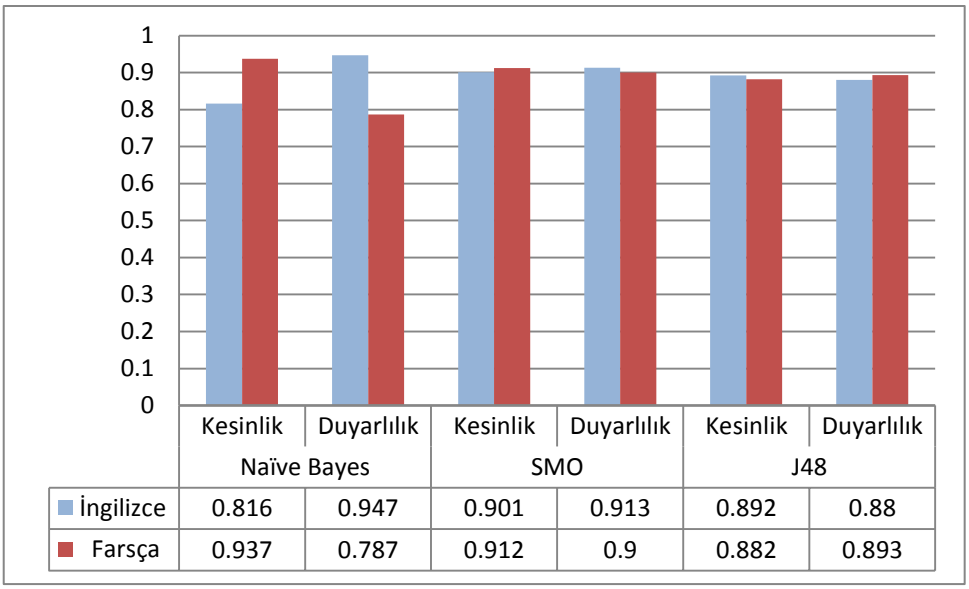


Şekil 28. 100 Dokümandan oluşan veri kümesi için iki dil tutulması durumunda Sınıflandırma Doğruluklarının karşılaştırılması



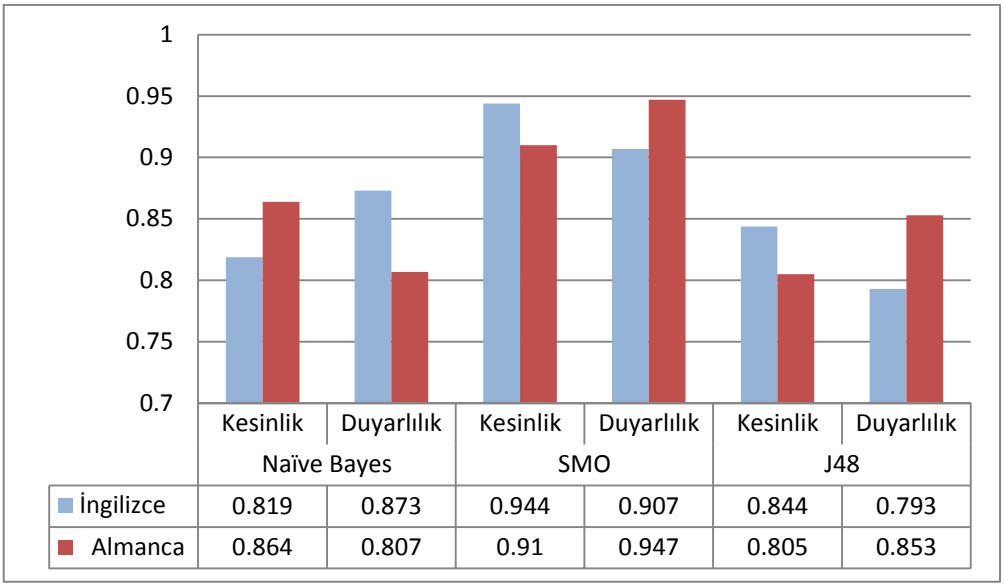
	İngilizce-Farsça	İngilizce-Almanca	İngilizce-Türkçe
SMO	90.66	92.66	94
Naïve Bayes	86.66	84	87.66
J48	88.66	82.33	82.66

Şekil 29. 150 Dokümandan oluşan veri kümesi için iki dil tutulması durumunda Sınıflandırma Doğruluklarının karşılaştırılması



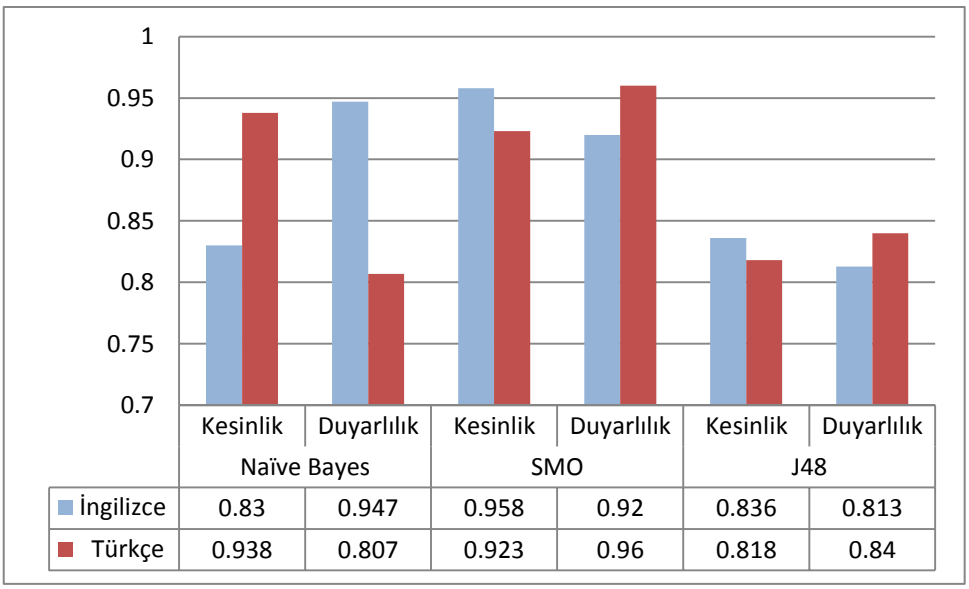
	Naïve Bayes		SMO		J48	
	Kesinlik	Duyarlılık	Kesinlik	Duyarlılık	Kesinlik	Duyarlılık
İngilizce	0.816	0.947	0.901	0.913	0.892	0.88
Farsça	0.937	0.787	0.912	0.9	0.882	0.893

Şekil 30. İki dil (İngilizce-Farsça) 150 Doküman üzerinde Kesinlik ve Duyarlılık ölçüsü ile karşılaştırmaktadır



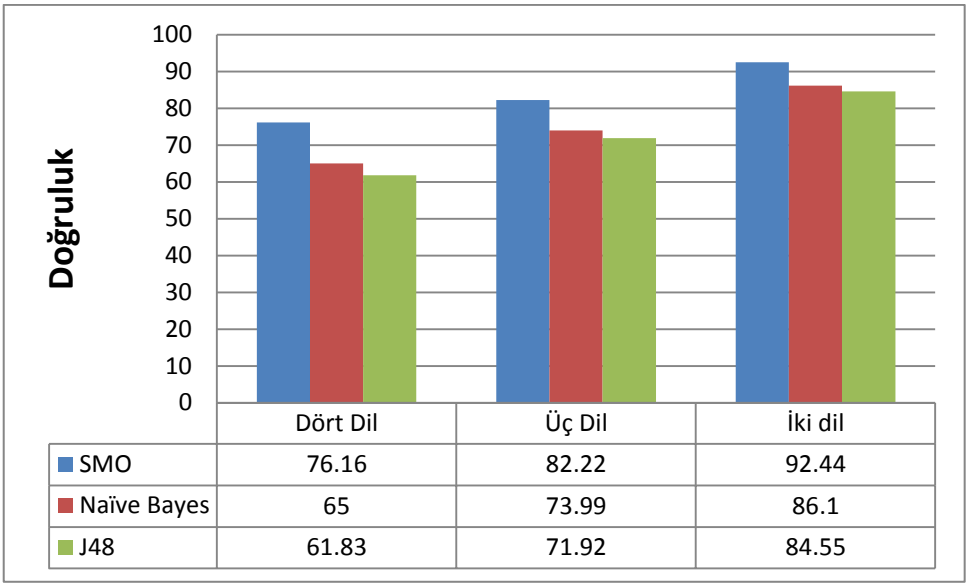
	Naïve Bayes		SMO		J48	
	Kesinlik	Duyarlılık	Kesinlik	Duyarlılık	Kesinlik	Duyarlılık
İngilizce	0.819	0.873	0.944	0.907	0.844	0.793
Almanca	0.864	0.807	0.91	0.947	0.805	0.853

Şekil 31. İki dil (İngilizce-Almanca) 150 Doküman üzerinde Kesinlik ve Duyarlılık ölçüsü ile karşılaştırmaktadır



	Naïve Bayes		SMO		J48	
	Kesinlik	Duyarlılık	Kesinlik	Duyarlılık	Kesinlik	Duyarlılık
İngilizce	0.83	0.947	0.958	0.92	0.836	0.813
Türkçe	0.938	0.807	0.923	0.96	0.818	0.84

Şekil 32. İki dil (İngilizce-Türkçe) 150 Doküman üzerinde Kesinlik ve Duyarlılık ölçüsü ile karşılaştırmaktadır



Şekil 33. 150 Dokümandan oluşan veri kümesindeki dil sayısının etkisi Doğruluklarının karşılaştırılması

Şekil 30, 31 ve 32 sadece iki dil için sınıflandırıcıların kesinlik ve duyarlılık ölçüm sonuçları verilmiştir. Şekil 33'de ise önerilen yöntemi daha kapsamlı biçimde değerlendirmek için değişen dil sayısının etkisinin ölçüm sonuçları gösterilmiştir. Burada, değerlendirmede çok sayıda veya az sayıda dil kullanımı durumları göz önüne alınarak, sistemin performansı ölçülmüştür. Dolayısıyla, dil sayısı her aşamada bir azaltılmış ve kalan diller için karşılaştırılmalar yapılmıştır. 150'şer dokümandan oluşan veri kümesi üzerinde elde edilen doğruluk sonuçları şekilde verilmiştir. Beklendiği gibi dil sayısı azaltıldığında, sınıflandırma performansı artmaktadır. Yine, SMO sınıflandırıcı en iyi performansı göstermektedir.

5. ÖNERİLER VE GELECEK ÇALIŞMALAR

Bu Çalışmada, Web Tabanlı metinlerin yazarının Anadilini Tanımlaması için bir yöntem önerilmiştir. Önerilen yöntem ve yaklaşımlar ile yapılan çalışmanın sonuçları, çevrimiçi metinlerin Yazarların Anadillerinin %80 üzerinde bir doğrulukla tespit edebileceğini göstermektedir. Aynı zamanda elde edilen sonuçlar, sözlüksel ve içeriğe özgü özelliklerin online metin yazarlarının belirlenmesinde, belli bir düzeyde ayırt edicilik gösterdikleri tespit edilmiştir.

SMO çevrimiçi metinlerin yazarlarının anadillerinin belirlenmesinde Naïve Bayes ve C4.5 metotlarına göre daha başarılı olmuştur. Sonuçlar göstermektedir ki, önerilen yöntem ve yaklaşımlar, siber uzaydaki yazarların anadil kimliğini otomatik olarak takip edilmesi ve belirlenmesi yeteneğine sahiptir.

Günümüzde, çevrimiçi ortamlarda yazarların analizi ve kimliklendirilmesi yeni bir çalışma alanını oluşturmaktadır. Bu özellikle yazarların kimliklendirilmesi konusu önümüzdeki yıllarda daha sonraki bir çalışma olarak daha ayrıntılı incelenebilir. Diğer yandan, yazar karakterizasyonu araştırmacılar için çok daha ilginç çalışma alanını oluşturmaktadır. Yazarın Anadilini Tanımlama, cinsiyet ve Yazarın Psikoloji durum gelecekte yapılacak diğer ilişkili araştırma konularıdır. Gelecekte, bu konularda özelliklerin en uygun kümesi tanımlamak Web tabanlı metinlerde Yazarın Anadilinin, cinsiyetinin ve diğer özelliklerinin Tanımlaması gelecek çalışmaların en önemli konularını oluşturacaktır. Yine daha fazla ve farklı dillerde çok daha kısa metinlerin (yani Facebook mesajları, Twitter tweets) yazarlarının analiz edilmesi ve tanımlanması diğer gelecek çalışmalarımızı oluşturacaktır.

6. KAYNAKLAR

1. Abbasi A., and Chen H., "Visualizing Authorship for Identification", IEEE International Conference on Intelligence and Security Informatics, 60–71, 2006.

2. Orebaugh A., and Allnutt J., "Classification of Instant Messaging Communications for Forensics Analysis", The International Journal of Forensic Computer Science, 22–28, 2009.

3. Boser B., Guyon I., and Vapnik V., "A Training Algorithm for Optimal Margin Classifiers", in Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, ACM Press, 144–152, 1992.

4. Baayen R.H., Halteren H.V., and Tweedie F.J., "Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution", Literary and Linguistic Computing, 110–120, 1996.

5. Cortes C., and Vapnik V., "Support Vector Networks in Machine Learning", 273–297, 1995.

6. Köse C., Özyurt Ö., and İkibaş C., "A Comparison of Textual Data Mining Methods for Sex Identification in Chat Conversations", Springer: Lecture Notes in Computer Science, LNCS, 4993, 638-643, 2008.

7. Chen H., "Applying Authorship Analysis to Extremist Group Web Forum Messages", IEEE Intelligent Systems: Special Issue on AI for Homeland Security 67-75, 2005.

8. Holmes D.I., "The Evolution of Stylometry in Humanities Scholarship", Literary and Linguistic Computing, 111–117, 1998.

9. De Vel O., "Mining E-mail Authorship", Paper presented at the Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining, 2000.

10. De Vel O., Corney M., Anderson A., and Mohay G., "Language and Gender Author Cohort Analysis of E-mail for Computer Forensics", In Proceedings of Digital Forensic Research Workshop, 2002.

11. Dietterich T.G., Hild H., and Bakiri G.A., "Comparative Study of ID3 and Backpropagation for English Text-to-Speech Mapping", In Proceedings of the Seventh International Conference on Machine Learning, 24–31, 1990.

12. Stamatatos E., "A Survey of Modern Authorship Attribution Methods", Journal of the American Society for Information Science and Technology, 538–556, 2009.

13. Iqbal F., Hadjidj R., Fung B.C.M., and Debbabi M., "A Novel Approach of Mining Write-Prints for Authorship Attribution in E-mail Forensics", digital investigation, 42–51, 2008.

14. Forsyth R.S., and Holmes D.I., "Feature Finding for Text Classification", Literary and Linguistic Computing, 163–174, 1996.

15. Granger S., Dagneaux E., and Meunier F., "The International Corpus of Learner English", Louvain-la-Neuve: Presses Universitaires de Louvain, 2002.

16. Mohtasseb H., Lincoln U., and Ahmed A., "Mining Online Diaries for Blogger Identification", Proceedings of the World Congress on Engineering, 2009.

17. Lee J., and Seneff S., "An Analysis of Grammatical Errors in Non-Native Speech in English", In Proceedings of the 2008 Spoken Language Technology Workshop, 2008.

18. Li J., Zheng R., and Chen H., "From Fingerprint to Writeprint", Commun. ACM, 76–82, 2006.

19. Joel R., "Tetreault and Martin Chodorow Examining the Use of Region Web Counts for ESL Error Detection", Web as Corpus Workshop (WAC-5), 2009.

20. Keerthi S.S., Shevade S.K., Bhattacharyya C., and Murthy K.R.K., "Improvements to Platt's SMO Algorithm for SVM Classifier Design", Neural Computation, 637–649, 2001.

21. Koppel M., Schler J., and Zigdon K., "Determining an Author's Native Language by Mining a Text for Errors", Proceedings of KDD, 2005.

22. Ledger G.R., and Merriam T.V.N., "Shakespeare, Fletcher, and the two Noble Kinsmen", Literary and Linguistic Computing, 235–248, 1994.

23. Mendenhall T.C., "The Characteristic Curves of Composition", Science, 237–249, 1887.

24. Mosteller F., and Wallace D.L., "Inference and Disputed Authorship: The Federalist", Series in behavioral science: Quantitative methods edition, 1964.

25. Cheng N., Chandramouli R., and Subbalakshmi K.P., "Author Gender Identification from Text", IEEE Symposium on Computational Intelligence and Data Mining Conference, 2009.

26. Tsur O., and Rappoport A., "Using Classifier Features for Studying the Effect of Native Language on the Choice of Written Second Language Words", Proceedings, ACL 2007 Workshop on Cognitive Aspects of Computational Language Acquisition, 2007.

27. Quinlan J.R., "Induction of decision trees", Machine Learning, 81–106, 1986.

28. Argamon S., Koppel M., Pennebaker J. W., and Schler J., "Automatically profiling the author of an anonymous text", Commun. ACM, 119–123, 2009.

29. Sebastiani F., "Machine Learning in Automated Text Categorization", ACM Computing Surveys, 1-47, 2002.

30. Wong S.J., and Dras M., "Contrastive Analysis and Native Language Identification", 2009.

31. Thomas D., and Loader B.D., "Cybercrime: Law Enforcement, Security and Surveillance in the Information Age", 2000.

32. Tweedie F.J., and Baayen R.H., "How Variable May a Constant be Measures of Lexical Richness in Perspective", Computers and the Humanities, 323–352, 1998.

33. Tweedie F.J., Singh S., and Holmes D.I., "Neural Network Applications in Stylometry", The Federalist Papers Computers and the Humanities, 1–10, 1996.

34. Vapnik V., "The Nature of Statistical Learning Theory", New York Springer Verlag, 1995.

35. De Vel O., Corney M., Anderson A., and Mohay G., "Language and Gender Author Cohort Analysis of E-mail for Computer Forensics", Digital Forensic Research Workshop, 2002.

36. Weka System, www.cs.waikato.ac.nz/ml/weka , 12/04/2012.

37. Yule G.U., "Statistical Study of Literary Vocabulary", Cambridge, U. Press, 1944.

38. Zhang H., and Sheng S., "Learning Weighted Naïve Bayes with Accurate Ranking", Data Mining (ICDM'04), 567-570, 2004.

39. Zipf G.K., "Selected Studies of the Principle of Relative Frequency in Language", Harvard University Press, Cambridge MA, 1932.

7. EKLER

Ek A. Kelime Zenginliği Ölçütleri

Yule'nin K ölçütü:

$$Yules\ K = 10^4 \left(-\frac{1}{N} + \sum_{i=1}^{V} V_i \left(\frac{i}{N} \right)^2 \right)$$

Simpson D ölçütü:

$$Simpsons\ D = \sum_{i=1}^{V} V_i \frac{i}{N} \frac{i-1}{N-1}$$

Sichel S ölçütü:

$$Sichels\ S = \frac{Hapax\ Dislegomena\ sayımı}{V}$$

Honore R ölçütü:

$$Honores\ R = \frac{100 \log_{10} N}{1 - \frac{Hapax\ legomena\ sayımı}{V}}$$

Entropy ölçütü:

$$Entropy = \sum_{i-1}^{N} V_i (-\log_{10} \frac{i}{N}) \frac{i}{N}$$

V: Farklı kelimelerin sayısı

V_i: i kez görülen farklı kelimelerin sayısı

N: Kelimelerin toplam sayısı

Hapax Dislegomena: sadece iki kez meydana gelen kelimeler

Hapax Legomena: sadece bir kez meydana gelen kelimeler

Ek B. Fonksiyon Kelimelerin Listesi

Article Words		
a	an	the

Pro-Sentence Words			
yes	no	okay	OK

Pronoun Words							
all	everybody	his	most	other	that	what	your
another	everyone	I	much	others	theirs	whatever	yours
any	everything	it	myself	ours	them	which	yourself
anybody	few	its	neither	ourselves	no	one	several
themselves	anyone	he	itself	who	anything	she	these
whichever	her	little	nobody	they	whoever	some	whom
yourselves	both	many	none	hers	this	each	whose
herself	me	nothing	other	somebody	whomever	those	something
each	him	mine	one	someone	us	either	himself
more	one	another	we	you			

Auxiliary Verbs							
are	shouldn’t	has	don’t	’re	were	’s	had
can	won’t	’ve	can’t	could	wouldn’t	shan’t	have
didn’t	aren’t	mightn’t	mustn’t	does	be	been	may
hadn’t	cannot	was	wasn’t	hasn’t	couldn’t	weren’t	should
haven’t	do	’ll	would	isn’t	doesn’t	’d	will
might	’d	ain’t	Is	shall	’s	did	

Conjunction Words						
and	or	though	now	that	if	while
because	yet	unless	even	though	now	that
nor	so	when	although	only	if	whether
in	order	whereas	even	or	that	in
if	not	case	until			

Interjection Words							
adios	bah	dear	Ha-ha	howdy	oops	tush	whoosh
ah	begorra	doh	hail	hoy	ouch	tut	wow
aha	behold	duh	hallelujah	huh	phew	Tutetut	yay
ahem	bejesus	eh	heigh-ho	humph	phooey	ugh	yikes
ahoy	bingo	encore	hello	hurray	pipepip	uh-huh	yippee
alack	bleep	eureka	hem	hush	pooh	uh-oh	yo
alas	boo	fie	hey	indeed	pshaw	uheuh	yoicks
all	hail	bravo	gee	hey	presto	jeepers	creepers
rats	viva	yoo-hoo	whiz	hi	jeez	righto	voila
alleluia	bye	gee	hip	lo	and	behold	scat
yuk	aloha	cheerio	gesundheit	dog	ooh	Touch	whoops
amen	cheers	goodness	hmm	man	shoo	well	zap
wahoo	yummy	gosh	ho	my	word	shoot	whoa
attaboy	ciao	aw	crikey	great	ho	hum	now
so	long	whoopee	ay	cripes	hah	hot	

Adposition Words						
aboard	astride	down	of	through	worth	on account of
about	at	during	off	throughout	according to	on behalf of
above	athwart	except	on	till	ahead to	Out from
absent	atop	failing	onto	considering	as to	out of
across	barring	following	opposite	toward	aside from	Outside of
after	before	for	out	towards	Because of	Owing to
against	behind	from	outside	under	close to	Prior to
along	below	in	over	underneath	due to	Pursuant to
alongside	beneath	inside	past	unlike	except for	Regardless of
amid	beside	into	per	until	far from	Subsequent to
amidst	besides	like	plus	up	in to	as far as
among	between	mid	upon	regarding	into	regarding
amongst	beyond	minus	round	via	considering	concerning
around	but	near	save	with	instead of	regarding
as	by	next	since	within	near to	in front of
aslant	despite	versus	than	without	next to	in lieu of
in place of	in spite of	on to	onto	on top of	versus	concerning
apart from	to	Not with standing				

Ek C. Kısaltma Kelimeler Listesi

Kısaltma Kelimeler Listesi							
Dr.	Pres.	Jr.	Mr.	Mrs.	Ms.	etc.	Inc.
Co.	dept.	hr.	Mon.	Tues.	Wed.	Thurs.	Fri.
Sat.	Sun.	in.	ft.	yd.	cm	mm	cc
mm	cm	kw	ft.	gr.	in.	lb.	oz.
pt.	qt.	tsp.	tbsp.	yd.	St.	Ave.	Ct.
Ln.	Blvd.	Cir.	mtn.	Long.	Rd.	ft.	Terr.
Hwy.	Pkwy.						

ÖZGEÇMİŞ

Parham MOHAMMADALIPOUR TOFIGHI, 1984 yılında Tahran'da doğdu. Sırası ile Şeykh Etar İlkokulu, Çemran Ortaokulu ve Ferdosi Lisesini bitirdi. 2006 yılında Shabester Üniversitesi Bilgisayar Mühendisliği Bölümünü bitirdikten sonra, 2010 yılında Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim dalında yüksek lisans yapmaya başladı. Çokiyi derecede İngilizce bilmektedir.