Journal of the Faculty of Engineering and Architecture of Gazi University 37:3 (2022) 1277-1292

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

A new content-free approach to identification of document language: Angle patterns

Tuba Noyan[1], Fatma Kuncan[1]*, Ramazan Tekin[2], Yılmaz Kaya[1]

[1]Department of Computer Engineering, Siirt University, Siirt, 56100, Turkey
[2]Department of Computer Engineering, Batman University, Batman, 72100, Turkey

Highlights:

- Language Identification in text mining is the process of detecting the natural language in which a document or part of it is written
- The proposed angle pattern method is used for feature extraction from texts
- The performance results of the proposed language recognition method in four different data sets are between 93,31% and 99,39%

Keywords:

- Text-based language identification
- Natural language processing
- Angle patterns
- Feature extraction

Article Info:

Research Article
Received: 21.12.2020
Accepted: 25.09.2021

DOI:

10.17341/gazimmfd.844700

Acknowledgement:

This study was performed in Siirt University Faculty of Engineering Machine Vision (MaVi) Laboratory. The authors of this article would like to thank the staff of MaVi Laboratory for their support.

Correspondence:

Author: Fatma Kuncan
e-mail: fatmakuncan@siirt.edu.tr
phone: +90 484 212 1111 / 3033

Graphical/Tabular Abstract

Language identification (LI) in text mining is the process of detecting the natural language in which a document or part of it is written. LI aims to mimic a human's ability to recognize certain languages from text by computer algorithms. LI can be defined as a classification problem subject based on the information used in word or character size for any document. When the literature is examined for LI application, it is seen that various linguistic or statistical-based approaches are used. Linguistic methods are methods that perform LI according to a special word or character of a language. These methods are applied based on the special rules of the languages. When we look at the statistical methods, it shows that the words or characters that make up the language depend on their frequency and distribution. The statistical approaches used are content-independent methods. The semantic context of the text is not concerned with its content. According to linguistic methods, it does not provide sufficient information about the content of the text. The proposed model in this study is a statistical approach.

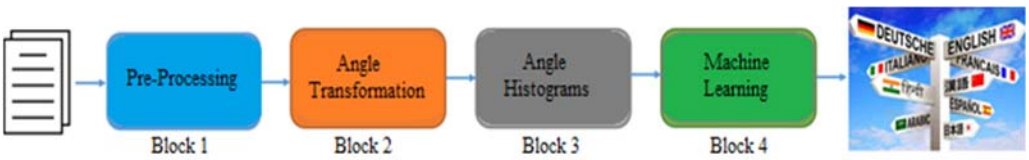


Figure A. Proposed block diagram for LI

Purpose: In this study, a new LI approach using the angle information between the UTF-8 values of the characters in the text is proposed. The proposed angle pattern method is used for feature extraction from texts. Angle patterns method is a statistical approach. In the angle method, there are two distance parameters, R and L, which express which neighborhood to look at from the reference point to the left and right.

Theory and Methods:
To test the proposed approach, four datasets, two created by the authors and two publicly available on the Internet, were used. By using the features obtained by the angle pattern method, classification process was carried out with different machine learning methods such as Random Forest, Support Vector Machine, Linear Discriminant Analysis, Naive Bayes and K-nearest neighbor. Language identification performance results determined from four different data sets were observed as 96.81%, 99.39%, 93.31% and 98.60%, respectively.

Results:
According to the performance results achieved as a result of the study, it has been determined that the proposed angle pattern method provides important distinguishing information in language identification application. It is thought that the proposed approach in this study can be used in many different text mining applications such as spam recognition, text categorization, as well as LI application.

Journal of the Faculty of Engineering and Architecture of Gazi University 37:3 (2022) 1277-1292

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Doküman dili tanıma için içerik bağımsız yeni bir yaklaşım: Açı örüntüler

Tuba Noyan[1], Fatma Kuncan[1]*, Ramazan Tekin[2], Yılmaz Kaya[1]
[1]Siirt Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 56100, Siirt, Türkiye
[2]Batman Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 72100, Batman, Türkiye

ÖNEÇIKANLAR

- Metin madenciliğinde dil tanıma, bir belgenin veya bir kısmının yazıldığı doğal dili algılama işlemidir
- Önerilen açı örüntüler yöntemi metinden öznitelik çıkarımı için kullanılmıştır
- Dört farklı veri seti kümesinde önerilen dil tanıma yöntemin başarım sonuçları %93,31 ile %99,39 arasındadır

Makale Bilgileri
Araştırma Makalesi
Geliş: 21.12.2020
Kabul: 29.09.202

DOI:
10.17341/gazimmfd.844700

Anahtar Kelimeler:
Metin tabanlı dil tanıma,
doğal dil işleme,
açı örüntüler,
öznitelik çıkarma

ÖZ

Metin madenciliğinde dil tanıma, bir belgenin veya bir kısmının yazıldığı doğal dili algılama işlemidir. Dil tanıma bilgisayar algoritmaları tarafından bir insanın metinden belirli dilleri tanıma yeteneğini taklit etmeyi amaçlamaktadır. Bu çalışmada, metin içindeki karakterlerin UTF-8 değerleri arasında kalan açı bilgisini kullanan yeni bir dil tanıma yaklaşımı önerilmiştir. Önerilen açı örüntüler yöntemi metinlerden öznitelik çıkarımı için kullanılmıştır. Açı örüntüler yöntemi istatistiksel bir yaklaşımdır. Açı yönteminde, referans noktadan sola ve sağa doğru kaçıncı komşuluğa bakılacağını ifade eden R ve L şeklinde iki uzaklık parametresi bulunmaktadır. Önerilen yaklaşımı test etmek amacıyla ikisi yazarlar tarafından oluşturulmuş ve ikisi internette herkesin kullanımına açık dört veri seti kullanılmıştır. Açı örüntüler yöntemi ile elde edilen öznitelikler kullanılarak Rastsal Orman, Destek Vektör Makinesi, Lineer Diskriminant Analiz, Naive Bayes ve K-en yakın komşu gibi farklı makine öğrenmesi yöntemler ile sınıflandırma işlemi gerçekleştirilmiştir. Dört farklı veri seti kümesinden belirlenen dil tanıma başarım sonuçları sırası ile %96,81, %99,39, %93,31 ve %98,60 olarak gözlenmiştir. Yapılan çalışma sonucunda ulaşılan başarım sonuçlarına göre önerilen açı örüntüler yönteminin dil tanıma uygulamasında önemli ayırt edici bilgiler sağladığı belirlenmiştir.

A new content-free approach to identification of document language: Angle patterns

HIGHLIGHTS

- Language Identification in text mining is the process of detecting the natural language in which a document or part of it is written
- The proposed angle pattern method is used for feature extraction from texts
- The performance results of the proposed language recognition method in four different data sets are between 93,31% and 99,39%

Article Info
Research Article
Received: 21.12.2020
Accepted: 25.09.2021

DOI:
10.17341/gazimmfd.844700

Keywords:
Text-based language identification,
natural language processing,
angle patterns,
feature extraction

ABSTRACT

Language Identification in text mining is the process of detecting the natural language in which a document or part of it is written. Language identification aims to mimic a human's ability to recognize certain languages by computer algorithms. In this study, a new language identification approach using the angle information between the UTF-8 values of the characters in the text is proposed. The proposed angle pattern method is used for feature extraction from texts. Angle patterns method is a statistical approach. In the angle method, there are two distance parameters, R and L, which express which neighborhood to look at from the reference point to the left and right. To test the proposed approach, four datasets, two created by the authors and two publicly available on the Internet, were used. By using the features obtained by the angle pattern method, classification process was carried out with different machine learning methods such as Random Forest, Support Vector Machine, Linear Discriminant Analysis, Naive Bayes and K-nearest neighbor. Language identification performance results determined from four different data sets were observed as 96,81%, 99,39%, 93,31% and 98,60%, respectively. According to the performance results achieved as a result of the study, it has been determined that the proposed angle pattern method provides important distinguishing information in language identification application.

*Sorumlu Yazar/Yazarlar / Corresponding Author/Authors : tubanoyan@yandex.com, *fatmakuncan@siirt.edu.tr, ramazan.tekin@batman.edu.tr, yilmazkaya@siirt.edu.tr / Tel: +90 484 212 1111 / 3033

1. GİRİŞ (INTRODUCTION)

Metin veya sesten dil tanıma (DT) problemi uzun yıllardır araştırma konusu olmuştur. Bir belgenin yazıldığı dili önceden bilmek, belgeden bilgi çıkarımı, özetleme, çeviri vs. gibi işlemlerin yapılması açısından önemlidir [1]. DT metin madenciliği [2] ve hesaplamalı dilbilim konularının önemli bir parçasıdır [3]. Otomatik DT bilgisayar algoritmaları [4] tarafından bir insanın belirli dilleri tanıma yeteneğini taklit etmeyi amaçlamaktadır [5]. DT herhangi bir doküman için kelime ya da karakter boyutunda kullanılan bilgileri temel alarak yapılan bir sınıflandırma problemi konusu olarak tanımlanabilir [6]. Yazılan dilin belirlenmesi problemleriyle web uygulamalarında sıklıkla karşılaşmaktayız [7]. Bunun yanında DT, duygu analizi [8], OCR (Optical Character Recognition, Optik karakter tanıma) sistemleri [9], otomatik çeviri sistemlerinde [10] ve metin özetleme algoritmaları için çok önemlidir. Çeviri işleminden önce metin dilinin belirlenmesi gerekmektedir [11].

Yıllar içinde, özel olarak tasarlanmış algoritmalar ve indeksleme yapıları kullanılarak, insan müdahalesine gerek kalmadan kullanılan dili tanıyabilen yaklaşımlar geliştirilmiştir. DT metin kategorize etme işlemi olarak görülebilir [12]. Metin kategorize etme işlemi, bir belgeyi önceden belirlenmiş bir sınıfa eşleme görevi olarak ifade edilir [13]. Son yıllarda internet teknolojilerinin gelişmesi nedeni ile web sayfalarından bilgi edinmek amaçlı yeni yaklaşımlara daha fazla yönelmenin gerektiği ortaya çıkmıştır. Bilgi elde edilmesi sürecinden önce belgenin dilinin belirlenmesi gerekmektedir.

DT uygulaması için literatür incelendiğinde dilbilimsel ya da istatistiksel tabanlı çeşitli yaklaşımlar kullanıldığı görülmektedir [13]. Dilbilimsel yöntemler bir dile ait özel bir kelime ya da karaktere göre DT işlemini gerçekleştiren metotlardır. Bu yöntemler dillere ait özel kuralları baz alarak uygulanmaktadır. İstatistiksel metotlara bakıldığında dili meydana getiren kelime ya da karakterlerin frekans ve dağılımlarına bağlı olduğunu göstermektedir [14]. İstatistiksel olarak kullanılan yaklaşımlar ise içerik-bağımsız yöntemlerdir [15]. Metnin anlamsal bağlamda içeriği ile ilgilenilmez. Metnin içeriği ile ilgili dilbilimsel yöntemlere göre yeterli bilgi vermemektedir [16].

DT çalışmalarında bakıldığında tekil karakter kombinasyonları [17], kısa kelime [18], n-gram [12] ve ASCII/Unicode karakter frekans vektörleri olmak üzere birçok farklı öznitelik çıkarım uygulamalarının kullanıldığı görülmüştür. Çoğunlukla, karakter seviyesinde çıkarılan özniteliklerin kullanılmasıyla ulaşılan sonuçlar kelime düzeyindeki özniteliklerin kullanılmasıyla ulaşılan sonuçlara göre daha başarılıdır [19]. Ayrıca karakter seviyesinde öznitelik çıkaran n-gram uygulamasının farklı öznitelik çıkarım yöntemlerine göre daha başarılı sonuçlar elde ettiği belirtilmiştir [20, 21]. Fakat n-gram uygulamasında aşırı öznitelik çıkarılmasından dolayı öznitelik uzayı belirgin olarak büyür ve bu belirgin olarak büyüme ise çok fazla işlem yükü ve bellek ihtiyacı gibi istenmeyen durumları meydana getirmektedir [22]. Bu nedenle genel olarak bu yöntemin beraberinde öznitelik seçme yöntemlerinin kullanılması gerekmektedir. DT'nın başarısı metinlerden çıkarılan özniteliklere, metin boyutuna, sınıflandırma metoduna, veri setindeki dil sayısına bağlıdır. Bu amaçla çalışmamızda metin içindeki karakterlerin UTF8 değerlerini bulduktan sonra bu değeler arasındaki açı bilgilerini kullanan yeni bir öznitelik çıkarım yöntemi olarak açı örüntüler yaklaşımı önerilmiştir. Bu çalışmanın motivasyonu şudur, her dil için karakterlerin ardışık diziliş biçimlerinin farklı olduğu düşünülmektedir. Dolayısıyla karakterler arasındaki açı değerleri her dil için farklı örüntüler oluşturacaktır. Bu uygulamada öncellikli olarak bütün karakterler Unicode değerlerine çevrilmeli ve sonrasında karakterler arasındaki açı değerleri bulunmaktadır. Bu dönüşümden sonra değerler aralığı 0-359 arasında değişen bir açı değer vektörü elde edilecektir. Yeni oluşan bu açılar bir sinyal olarak düşünüldüğünde bu sinyale ait histogram sınıflandırıcı metotlar için öznitelik vektörü olarak kullanılabilir. Açı örüntüler yönteminin metinlerden farklı örüntüleri yakalayabilmesi için merkez noktanın sağ ve solundan ne kadar uzağa bakılacağını ifade eden (R ve L) parametreleri bulunmaktadır. Önerilen yaklaşımı test etmek için çeşitli şekillerde oluşturulan dört veri seti kullanılmıştır. Elde edilen öznitelikler kullanılarak Rastsal Orman (RF, Random Forest), Destek Vektör Makinesi (SVM, Support Vector Machine), Lineer Diskriminant Analiz (LDA, Linear Discriminant Analysis), Naive Bayes (NB) ve K-en yakın komşu (Knn, k-nearest neighbors) olmak üzere farklı makine öğrenmesi yöntemler ile sınıflandırma işlemi gerçekleştirilmiştir. Gözlemlenen başarı performansları dikkate alındığında önerilen yaklaşımın dil tanımada başarılı bir şekilde kullanılabileceği görülmüştür.

2. LİTERATÜR ÇALIŞMALARI (LITERATURE STUDIES)

DT'nın amacı, hangi dilde yazıldığı bilinmeyen doküman ya da belgeyi işlemek ve o dokümanı tanımlamaktır. Bazı tekniklerin dilin ayırt edici özelliklerine uygulanmasıyla DT gerçekleştirilir. Bilgisayar tabanlı DT, insanın belirli dilleri tanıma yeteneğini taklit etmeyi amaçlar. Yıllar boyunca, özel olarak tasarlanmış algoritmalar ve indeksleme yapılarının kullanımı yoluyla, insan müdahalesine ihtiyaç duymadan kullanılan dili çıkarabilen DT algoritmaları geliştirilmiştir. Bir insan metinlerden birkaç dili tanıma kabiliyetine sahipken bilgisayar tabanlı DT sistemleri yüzlerce dili tanıma yeteneğine sahip olabilir [23]. Geniş anlamda, DT, konuşma, işaret dili ve el yazısı metin dahil olmak üzere herhangi bir dil yöntemi için geçerlidir ve dili, dijital veya başka türlü içeren tüm bilgi depolama araçlarıyla ilgilidir.

Bir belgenin yazıldığı dili doğru bir şekilde tespit etme yeteneği, verilere erişilebilirliği artırabilir. Örneğin, bir kullanıcının ana dilinde bilgi sunmanın, web sitesi ziyaretçilerini ilgili dilde içerik sunma önemli DT uygulama alanları olabilir [24]. Diğer önemli bir uygulama alanı otomatik çeviri işleminden önce dokümanın dilinin

belirlenmesi gerektirir. DT, metin sınıflandırmasının özel bir durumudur. Bir belgeyi önceden belirlenmiş bir sınıf kümesine eşleme görevi olarak görülebilir. Yapılan çalışmalara bakıldığında; Öztürk vd. yaptıkları çalışmada Türk müziği eserleri üzerinde dil tanıma uygulamasına çok benzer ve değişik bir uygulama olan makam tanıma çalışması gerçekleştirmişlerdir. Farklı Türk müziği eserlerinin farklı makama ait özellik göstermesinde yola çıkılarak Türk makam tanıma konusunda dil tanıma çalışmasına yakın bir faklı çalışma gerçekleştirmişlerdir [25]. Basile vd. yaptıkları çalışmada büyük harf frekanslarını kullanarak DT işlemi gerçekleştirmişlerdir [26]. Tohma ve Kutlu yaptıkları çalışmada doğal dil işleme üzerinde durarak, Türkçenin doğal dil işleme açısından dikkat çeken bir dil olduğuna değinmişlerdir. Özellikle sondan eklemeli bir dil yapısına sahip olması, ünlü-ünsüz uyumu v.b. birçok farklı özelliğe sahip olması sebebiyle dil tanıma konusunda araştırmacılar için ilginç bir yönü olduğuna değinmişlerdir [27]. Tian ve Suontausta'ın yaptıkları çalışmada sinir ağı tabanlı ölçeklenebilir yeni bir dil tanımlaması yöntemi önermişlerdir. Geliştirdikleri algoritmayı seyrek bellek kaynaklarına sahip gömülü uygulamalar için önermektedirler. Önerdikleri yaklaşımla birkaç dilde hem yüksek dil tanımlama hem de tanıma oranı olduğunu ifade etmişlerdir. Bu çalışmanın temel sistemin dil tanımlama doğruluğunu korurken hedef tarafından belirlenen bellek gereksinimlerini karşılayabilmesi gibi avantajı olduğunu belirtmişlerdir [28]. Özcan ve Baştürk yaptıkları çalışmalarında dil tanıma konusunda farklı bir çalışma yaparak konuşma ve duyma yetisi olmayan insanların kullanabilmesi için işaret dilinin tanınması konusunda akademik bir çalışma yaptıklarını belirtmişlerdir. Yazarlar yaptıkları çalışma sonucunda CNN modeli kullanarak %93 civarında bir doğruluk yakaladıklarını belirtmişlerdir [29].

Genellikle DT süreçleri, öznitelik çıkarımı [30], bilgi çıkarımı [31] ve sınıflandırma aşamalarından [32] oluşur [33]. Xiao arkadaşlarının çalışmalarında N-Gram maskeleme yöntemi olan Ernie-Gram yöntemini kullanmışlardır [34]. Bu yöntemle n-gramlar maskelenir ve bitişik diziler yerine açık n-gramlar kullanılarak doğrudan tahmin edilmesinin sağlandığını söylemişlerdir. Çalışmaları sonucunda Ernie-Gram yönteminin diğer yöntemlere göre çok daha iyi performans gösterdiğini ve güncel yöntemlerle karşılaştırılabilir sonuçlar elde ettiklerini vurgulamışlardır [35]. Castro vd. tarafından ele alınan çalışmada metnin dilinin belirlenmesi üzerine bir çalışma yapmışlardır. Yazarlar yaptıkları çalışmada n-gram tabanlı bir yöntem önermişlerdir ve yapılan çalışmalar sonucunda %92 civarında bir doğruluk elde ettiklerini ifade etmişlerdir [36]. Kumar vd. mevcut kelimenin konum bilgisini DT için kullanmışlar. Cümlelerdeki kelimelerin konumu dil hakkında bilgi verdiğini ortaya koymaya çalışmışlar [37]. Öznitelik çıkarım uygulamalarında genellikle n-gram [38], kısa kelime sıklıkları, kelime ve harf frekansları, dile ait özel karakter frekansları veya karakteristik yapılar kullanılmıştır [39]. Öznitelikleri kullanan sınıflandırma metodu [40] da başarıyı etkilemektedir [41, 42]. Markov modelleri [43], entropi tabanlı metotlar [44], Gaussian karışımlı modeller [45], karar ağaçları [46], yapay sinir ağları [47], destek vektör makineler [48], melez modeller [49], Knn [50], NB ve regresyon modeller [51] gibi makine öğrenmesi yöntemler kullanılmıştır [52]. DT çalışması için yapılan uygulamalarından belirli bir bölümünde kullanılan yöntemler ve bu yöntemler sonucunda ortaya çıkan başarım oranları Tablo 1'de kısaca verilmiştir. Yapılan çalışmalar incelendiğinde [53], DT uygulamasında çalışmaların çok önemli bir kısmında [54] öznitelik çıkarım tabanlı yaklaşımların önemli bir etkiye sahip olduğu belirlenmiştir [55].

3. VERİ SETLERİ (DATA SETS)

Bu çalışmada önerilen açı örüntülerin DT'da başarısını test için dört farklı veri seti kullanılmıştır. Çalışmada kullanılan iki veri seti (VS1 ve VS2) yazarlar tarafından oluşturulmuştur. Diğer iki veri seti (VS ve VS4) ise internette herkesin kullanımına açık veri setleridir.

Veri seti 1 (VS-1): Birinci veri seti kümesi aşk, iktidar, barış, bilgisayar, bilişim, teknoloji, insanlık, aile, sevgi, kanser, spor, uzay, para vb. çeşitli kelimeler ile ilgili metinlerden oluşmaktadır. Metinlerde özel kelimeler ve sayısal değerler bulunmamaktadır. Veri seti öncelikle Türkçe olarak toplatılmıştır. Türkçe Wikipedia kullanılarak 200 metin toplandıktan sonra bu metinler hazırlanan bir yazılım ile Google çeviri sistemi kullanılarak 60 farklı dile çevrilmiştir. Geliştirilen yazılıma ait ekran görüntüsü Şekil 1'de verilmiştir. Bu veri setinde toplamda 12000 metin bulunmaktadır.

Veri seti 2 (VS-2): Bu set VS-1'e benzerlik göstermektedir. VS-1'deki anahtar kelimelerin İngilizce karşılıklarına göre İngilizce olarak toplatılmıştır. İngilizce toplanan 200 metin daha sonra geliştirilen yazılım ile 60 farklı dili çevrilmiştir. Bu veri setinde de toplam 12000 metin bulunmaktadır.

Veri seti 3 (VS-3): Wili-2018, veri seti farklı dillerdeki Wikipedia sitelerinden oluşturulmuştur [45]. Farklı dillerin birbirinden ayırt edilmesi için oluşturulmuş herkese açık bir veri setidir. Bu veri setinden 50 dil seçilerek önerilen yöntem test edilmiştir. Her dil için 1000 metin kullanılmıştır. Bu veri seti için toplamda 50000 metin kullanılmıştır.

Veri seti 4 (VS-4): Dördüncü veri seti kümesi Baldwin ve Lui'nin Wikipedia ismiyle oluşturulan geniş çaplı olarak çok fazla çalışmada kıyaslama uygulamalarında tercih edilen çok dilli olan veri seti kümesi olarak bilinmektedir [4]. Önerilen çalışmamızda dördüncü veri seti kümesinde olan 250 karakterin altındaki metinlerin atılması sonucunda elde kalan 25 dile ait metinlerin kullanılmasıyla gerçekleştirilmiştir. Bu veri setinde 25 dile ait toplamda 1849 metin bulunmaktadır. Tüm veri setlerinde bulunan metinlerden noktalama işaretleri, boşluklar ve özel işaretler çıkarılmış ve tüm metinler tekil kodlara (Unikod) dönüştürüldükten sonra kullanılmıştır. Tekil kod, metni kodlamak için bir standarttır. Bir tekil kod tek bir karakteri benzersiz şekilde tanımlayan bir tamsayıdır.

Tablo 1. DT için yapılan çalışmalar ve elde edilen başarı oranları (Studies and success rates for LI)

Öznitelik Çıkarım Yöntemi	Sınıflandırma Yöntemi	Performans (%)	Referans
Karakter Frekansı	Ağırlık merkezi ve ters sınıf frekansı	%98	[34]
1-gram 2-gram	Bağıl Entropy	%78,2-99,4 %90,2-100	[36]
3-gram 4-gram 5-gram Kısa kelimeler Kısa kelimeler + 3-gram Kısa kelimeler + 4-gram	Linguini (vektorel mesafe tabanlı bir yöntem)	%68,8-100 %79,5-100 %83,6-100 %81,4-99,9 %61,3-100 %83,8-100 %84,9-100	[56]
Harf frekansı	Yapay sinir ağları ve bulanık mantık modeli	%99,75	[57]
n-gram	Ağırlık Merkezi Tabanlı DVM YSA k-ortalamalar Bulanık C ortalamaları	%93-93,5 %61-85 %77-91 %64-76,4	[58]
Birleşim tespit yöntemi	YSA	%77-86,6 %90-98	[59]
4 farklı veri seti için 1B-YİÖ	YSA, DVM ve ÇTNB	%86,20 %92,75 %100	[13]
N-Gram	mutual cross entropy uzaklığı	%89,77 %89-99	[60]
N-Gram	Naive Bayes	%80-100	[61]
Karakter, Kelime, Özel karakter, Tabanlı Yaklaşımlar	Benzerlik Ölçütleri (Kosinüs, Manhattan, Chi2, Canberra, Bhattacharyya, Korelasyon)	%96,09-97,92	[9]
N-Gram	NB, SVM	%74,0-98,4	[62]
Graf tabanlı N-gram	İstatistiksel Testler (T-testi)	%92,4-97,3	[63]
N-gram	decision tree	%90,9	[64]
Trigram istatistikler	NB	%90,2	[65]
n-gram	SVM, NB	%92-98	[66]
Yasaklı kelimeler	Markov Modeller	%99,18	[67]

4. YÖNTEM (METHOD)

4.1. Açı Örüntüler (Angle patterns)

Bu çalışmada DT için yeni bir yaklaşım önerilmiştir. Açı Örüntüler, metinlerin tek boyutlu sinyal olarak düşünülmesi ve bu sinyal örneklerinin karakterin tekil kod değerleriyle ifade edilmesi esasına dayanır. Buna göre, metinlerden etkin özelliklerin elde edilmesi için işaretler üzerindeki tekil kod değerlerin birbirleri ile oluşturdukları açı bilgilerini kullanan yeni bir yaklaşım önerilmektedir. Şekil 2'de metindeki karakterlerin tekil kod değerleri kullanılarak elde edilen sinyal üzerinde örnek noktalar gösterilmiştir.

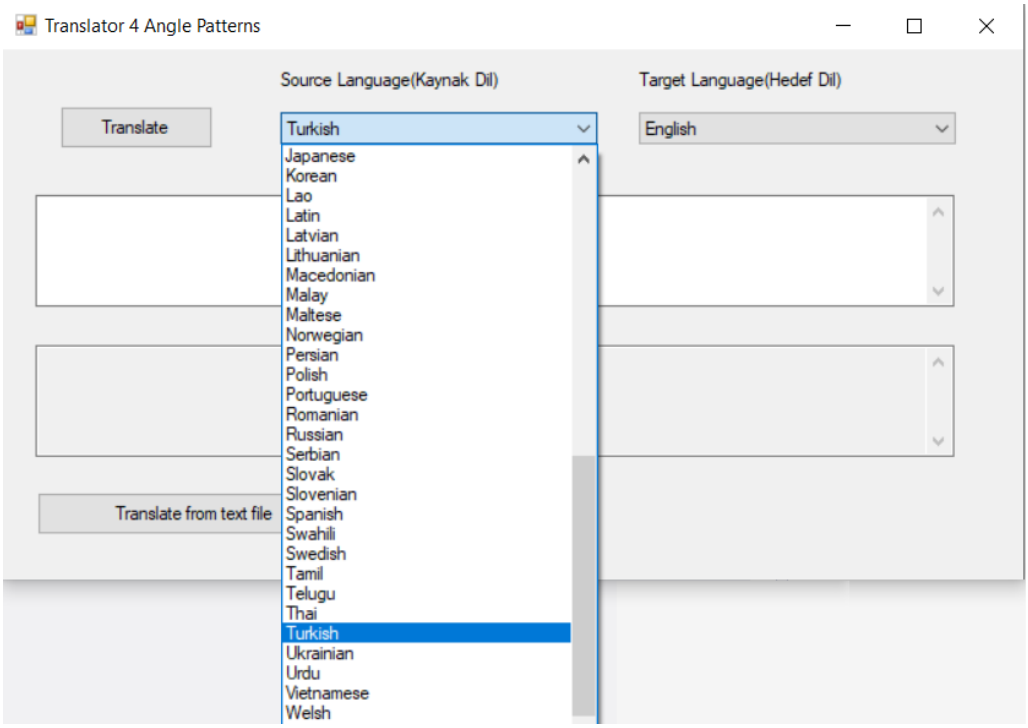
Bu yöntemde öncelikle metinlerin oluşturdukları işaretler üzerindeki sırası ile Şekil 2'de gösterildiği gibi her 3 nokta arasındaki açı değerleri hesaplamaktadır. Daha sonra oluşan açıların frekansları hesaplanmaktadır. Diğer bir deyişle yeni oluşan açı işaretlerine ait histogram elde edilmektedir. Komşu 3 nokta arasındaki aşağı bakan açı hesaplanmaktadır. Örnek açılar Şekil 3'te gösterilmiştir.

Zaman serisinde herhangi bir $P_i^{x,y}$ noktası için açı değeri Eş. 1 ile hesaplanmaktadır. ($P_i^x > 0$ olmak üzere):

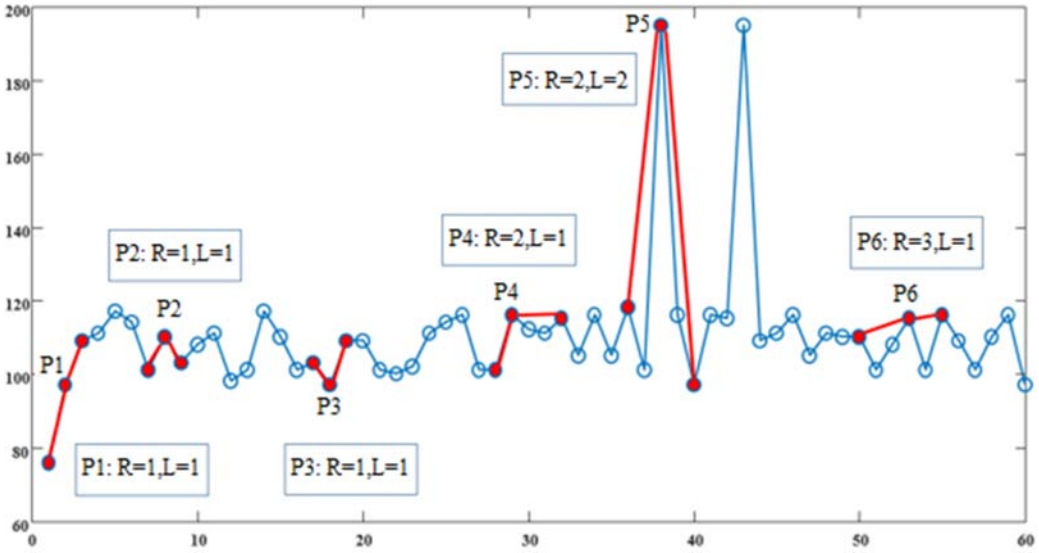
$$\theta_i = \arctan\left(\frac{\left|det\begin{vmatrix} P_{i-1}^x - P_i^x & P_{i-1}^y - P_i^y \\ P_i^x - P_{i+1}^x & P_i^y - P_{i+1}^y \end{vmatrix}\right|}{\left(P_{i-1}^x - P_i^x\right)\left(P_i^x - P_{i+1}^x\right) + \left(P_{i-1}^y - P_i^y\right)\left(P_i^y - P_{i+1}^y\right)}\right) * \frac{180}{pi} + 180 \quad (1)$$

Burada θ_i radyal olarak elde edilmektedir.

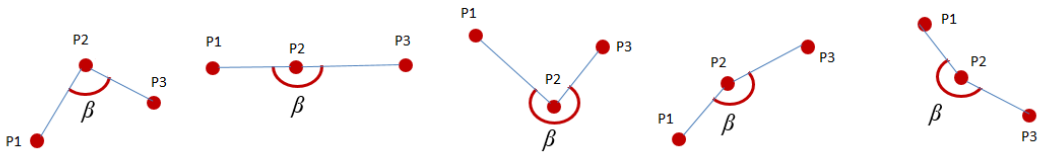
Açı yönteminde, referans noktadan sola ve sağa doğru kaçıncı komşuluğa bakılacağını ifade eden sırasıyla R ve L şeklinde iki uzaklık parametresi bulunmaktadır. L parametresi P1 noktasının P2 noktasına olan uzaklığını



Şekil 1. Çeviri yazılıma ait ekran (Screen of translation software)



Şekil 2. Bir metne ait tekil kod değerler ve açı örnekleri (Unicode values and angle examples of a text)



Şekil 3. Örnek açı örüntüler (Example angle patterns)

belirtir. Diğer bir deyişle, P1 noktasının P2 noktasına göre soldan kaçıncı nokta alınacağını belirtmektedir. R parametresi ise P3 noktasının P2 noktasına sağ taraftan olan uzaklığını belirtir. L ve R değerlerine göre P1, P2 ve P3 arasında farklı açılar oluşmaktadır. Bunlar da farklı örüntülerin oluşmasını sağlamaktadır. (Şekil 4) İşaretlere açı yöntemi uygulandıktan sonra işaretler 0 ile 359 arasındaki değerlere dönüşmektedir. Her açı değerinin frekansı bir açı

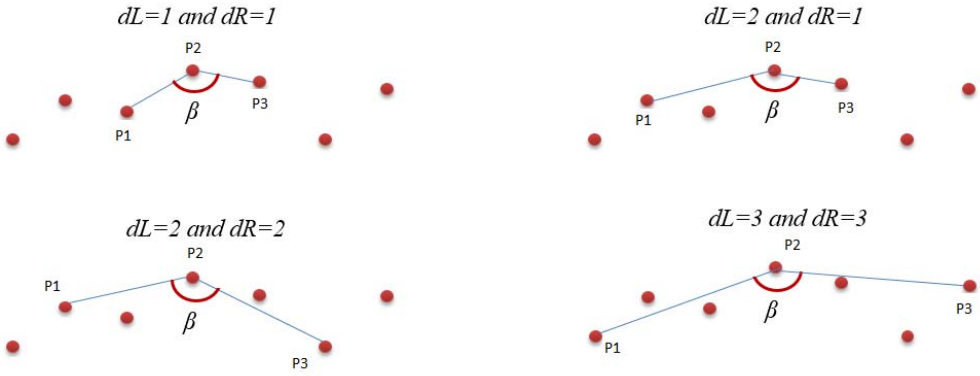
örüntüsü olarak ele alınmaktadır. Diğer bir deyişle yeni oluşan açı işaretlerine ait histogram öznitelik vektörü olarak ele alınmaktadır.

4.2. Dil Tanıma Diyagramı (Language Recognition Diagram)

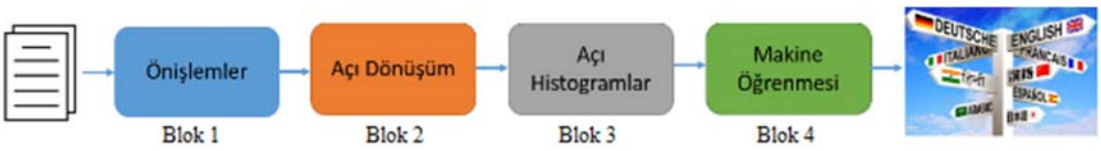
Dil tanıma için önerilen sistem şeması Şekil 5'te verilmiştir. Önerilen sistem 4 aşamadan oluşmaktadır. Her aşamada gerçekleştirilen işlemler Şekil 5'te özetlenmiştir.

Blok 1: Bu aşamada metin içindeki noktalama işaretleri, rakamlar, özel karakterler, boşluk karakterleri silinmektedir. Dil tanıma için gerekli karakterler dışındaki tüm karakterler metin içinden atılmaktadır. Yeni oluşan string ifade içindeki her karakterin tekil kodu bulunur. Bu aşamanın sonunda karakterlere karşılık gelen UTF-8 değerler elde edilmiş olur. Örnek bir metin Tablo 2'de verilmiştir. İlk satırda normal metin ikinci satırda ise önişlem kuralların uygulanmasından sonra elde edilen metin verilmiştir.

Blok 2: Bu aşamada UTF-8 tekil kod değerlerden oluşan vektöre açı yöntemi uygulanmaktadır. Yöntemin uygulanmasından sonra değerleri 0-359 arasında oluşan açı değerler elde edilmiş olur. Açı yöntemine ait uzaklık parametrelerinin farklı değerleri için farklı tekil kod histogramlar elde edilir. Farklı örüntülerin yakalanması için bu parametre değerleri önemli olmaktadır. Tablo 1'deki metne ait tekil kod değerler ve bu değerlere açı yöntemi uygulandıktan sonra elde edilen açı bilgilerinin bir kısmı Tablo 3'te verilmiştir. Tablo 3'teki ilk 3 değerlerin (90,97,109) arasında kalan açı değeri aşağıdaki gibi hesaplanmaktadır. Noktalar P1= (1,90), P2= (2,97), P3= (3,109) değerleri arasındaki açı değeri Şekil 6'da gösterilmiştir. P1, P2 ve P3 noktaları arasındaki açı değeri Eş. 2-Eş. 5'deki gibi hesaplanmaktadır.



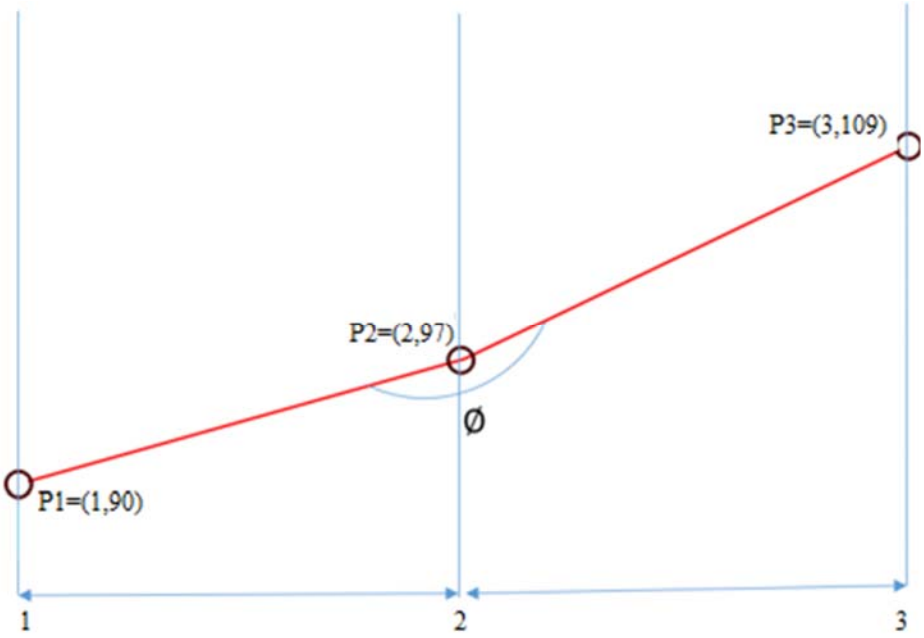
Şekil 4. L ve R parametrelerine göre oluşan örüntüler (Patterns based on L and R parameters)



Şekil 5. DT için önerilen blok diyagram (Suggested block diagram for LI)

Tablo 2. Örnek bir metin (An example text)

Zaman, göreceli bir kavramdır. Zaman içinde olduğumuz üç mekân ve bir zaman boyutlu uzay zamanın soyut olan boyutu olarak da kabul edilir. Zaman hareket eseri ortaya çıkmıştır o halde zaman hareketin ürünüdür. Bu görelilikte de böyle denilebilir. Zaman yolculuğun mümkün olup olmadığı birçok bilim adamı tarafından düşünülmektedir. Zaman, ışık hızı ile de dolaysız ilişki içinde olup maddenin ışık hızına yaklaşması durumunda zamanın yavaş akması, ışık hızında durması ve ışık hızı ötesinde de tersine akması; atom altı parçacıkların ışıktan hızlı hareket ettiği ve zamanlarının gelecekten geçmişe doğru aktığı veya içinde bulunduğumuz uzay zamandan başka sonsuz sayıda da ihtimalin olabileceği hipotezleri de modern fiziğin ve Görelilik Kuramının temelini oluşturan konulardandır.
ZamangörecelibirkavramdırZamaniçindeolduğumuzüçmekanvebirzamanboyutluuzayzamanınsoyutolan boyutuolarakdakabuledilirZamanhareketeseriortayaçıkmıştırohaldezamanhareketinürünüdürBu göreleliktedeböyledenilebilirzamanyolculuğunmümkünolupolmadığıbirçokbilimadamıtarafından düşünülmektedirZamanışıkhızıiledolaysızilişkiiçindeolupmaddeninışıkhızınayaklaşmasıdurumunda zamanınyavaşakmasıışıkhızındadurmasıveışıkhızıötesindedetersineakmasıatomaltıparçacıklarınışıktanhızlı hareketettiğivezamanlarınıngelecektengeçmişedoğruaktığıveyaiçindebulunduğumuzuzayzamandanbaşka sonsuzsayıdadaihtimalinolabileceğihipotezleridemodernfiziğinveGörelilikKuramınıntemelinioluşturan konulardandır



Şekil 6. Örnek noktalar arasındaki açı hesabı (Angle calculation between sample points)

Tablo 3. Tekil kod değerler ve bu değerler arasındaki açı bilgileri (Unicode values and angle information between these values)

Tekil Kodlar	90	97	109	97	110	103	246	114	101	99	101	
		108	105	98	105	114	107	97	118	114	97	
		109	100	253	114	90	97	109	97	110	105	
		231	105	110	100	101	111	108	100	117	240	
		117	109	117	122	252	231	109	101	107	97	
		110	118	101	98	105	114	122	97	109	97	
		110	98	111	121	117	116	108	117	117	122	97
		121	122	97	109	97	110	253	110	115	111	
		121	117	116	111	108	97	110	98	111	121	
		117	116	117	111	108	97	114	97			
Açı Değerler	183	10	351	13	352	1	184	203	307	199	27	
		170	344	182	15	178	352	17	170	352	12	
		354	1	182	350	184	10	351	16	349	1	
		349	18	310	220	25	169	350	183	1	187	
		346	176	191	4	178	187	344	16	350	178	11
		196	334	182	180	10	353	10	351	10	351	
		179	20	211	143	347	97	259	14	356	138	48
		353	10	351	184	1	349	26	341	20	211	
		147	188	167	351	10						
		351	179	20	211	270	55	189	167	352	7	

$$\theta_i = \arctan\left(\frac{\left|det\begin{vmatrix}1-2 & 90-97\\ 2-3 & 97-109\end{vmatrix}\right|}{(1-2)(2-3)+(90-97)(97-109)}\right) * \frac{180}{pi} + 180 \quad (2)$$

$$\theta_i = \arctan\left(\frac{|5|}{85}\right) * 180/pi + 180 \quad (3)$$

$$\theta_i = \arctan(0.0588) * \frac{180}{pi} + 180 \quad (4)$$

$$\theta_i = 183 \quad (5)$$

Blok 3: Bu aşamada önceki aşamada elde edilen UTF-8 tekil kod değerlerin histogramı oluşturulur. Histogramdaki her değer bir örüntü olarak kabul edilir. Diğer bir deyişle her örüntüyü bir öznitelik olarak kabul edebiliriz. Tablo 1'deki metne açı örüntüler uygulandıktan sonra elde edilen açı değerlerine ait histogram Şekil 7'de verilmiştir.

Blok 4: Bu aşama elde edilen öznitelik vektörlerine uygulanan makine öğrenmesi işlemlerini belirtir. RF, SVM, LDA, NB ve Knn sınıflandırma metotları kullanılmıştır. Sınıflandırma işlemleri 10 katlı çapraz geçerlilik testine göre gerçekleştirilmiştir. Sınıflandırma işlemleri için açık kaynak kodlu Weka yazılımı kullanılmıştır.

4.3. Performans Ölçütleri (Performance criteria)

Önerilen açı yönteminin performansını belirlemek amacıyla doğruluk, kesinlik, duyarlılık ve f-ölçütleri kullanılmıştır. Bu başarı ölçütleri Eş. 6-Eş. 9'daki gibi hesaplanmaktadır.

$$Doğruluk(Accuracy) = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$Kesinlik(Precision) = TP/(TP + FP) \quad (7)$$

$$Duyarlılık(Recall) = TP/(TP + FN) \quad (8)$$

$$F - Ölçütü(F - Measure) = 2(Duyarlılık * Kesinlik)/(Duyarlılık + Kesinlik) \quad (9)$$

Bu denklemlerde T, F, P ve N sırasıyla doğruyu, yanlışı, pozitifi ve negatifi ifade etmektedir. Örneğin, TP doğru sınıflandırılan pozitif örnek sayısını; FN ise yanlış sınıflandırılan negatif örnek sayısını göstermektedir.

Doğruluk: Başarının belirlenmesi amacıyla kullanılan en yaygın ve basit yöntemdir ve bu oran doğru sınıflandırılmış (TP+TN) örnek sayısının, toplam örnek sayısına (TP+TN+FP+FN) oranı olarak tanımlanmaktadır.

Kesinlik: Sınıflandırıcı sonucunun kesinlik derecesini vermektedir. Pozitif olarak etiketlenen örneklerin sayısının (TP) pozitif olarak sınıflandırılmış toplam örneklere (TP+FP) oranı olarak ifade edilmektedir.

Duyarlılık: Pozitif olarak etiketlenmiş örneklerin (TP) gerçekten pozitif olan örneklerin (TP+FN) toplam sayısına oranı olarak ifade edilmektedir.

F-Ölçütü: Kesinlik ve duyarlılık metriklerinin kullanılmasıyla hesaplanmaktadır. Sistemin, kesinlik veya duyarlılık yönüne doğru optimize edilmesi için kullanılmaktadır.

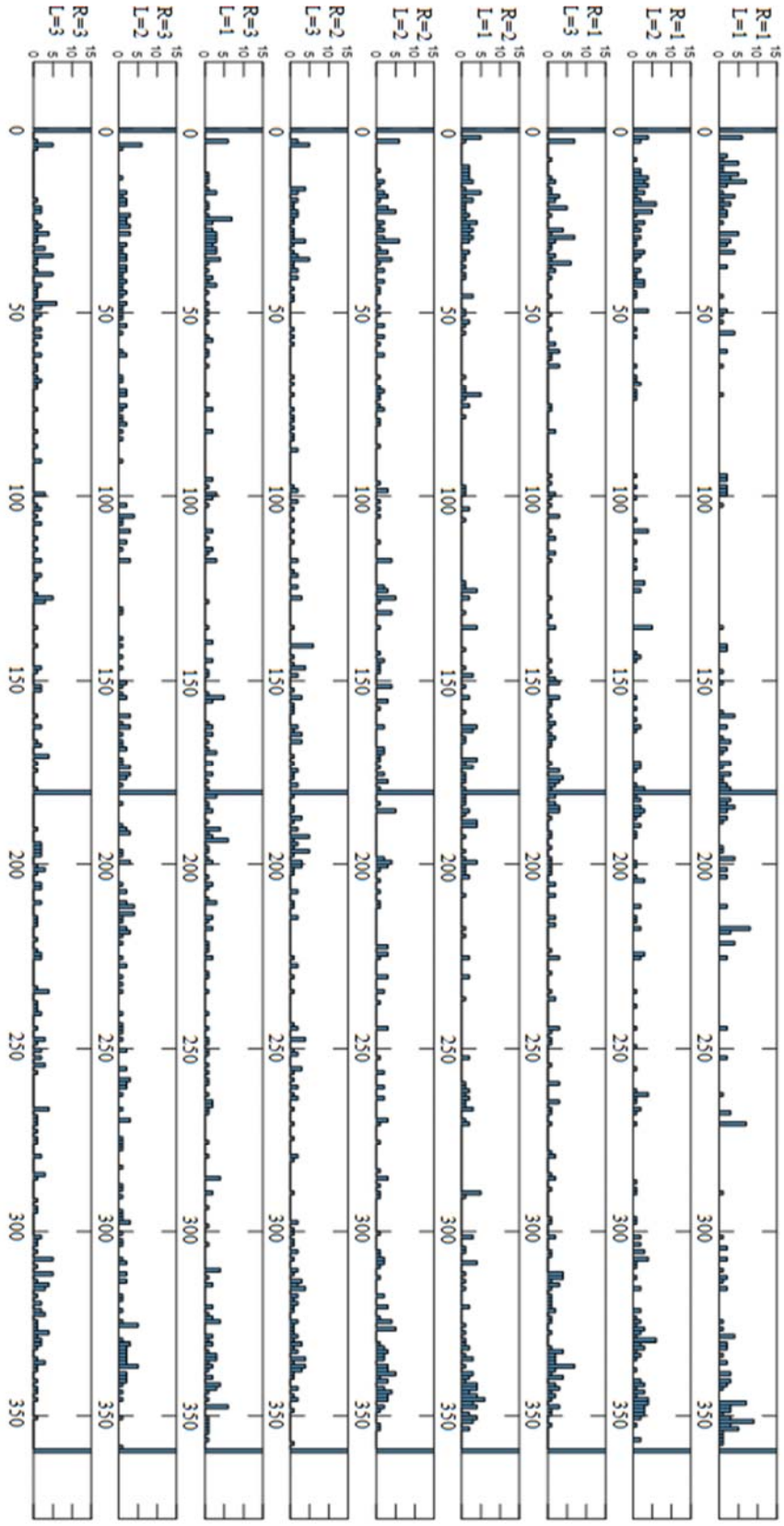
5. DENEYSEL SONUÇLAR (EXPERİMENTAL RESULTS)

Bu çalışmada DT amacıyla karakterlerin UTF-8 tekil kod sıralı değerleri arasındaki açı bilgilerini kullanan yeni bir yaklaşım önerilmiştir. Önerilen açı yönteminin uzaklık parametreleri (R ve L) ile farklı örüntülerin yakalanması amaçlanmaktadır. Örnek bir Fransızca metinde bu parametrelerin farklı değerlerine göre çıkarılan örüntülerin dağılımı Şekil 8'de verilmiştir. Şekilden anlaşıldığı üzere R ve L parametrelerin farklı değerlerine göre histogramların değiştiği gözlemlenmektedir. L ve R parametrelerine göre farklı örüntüler elde edilmiştir.

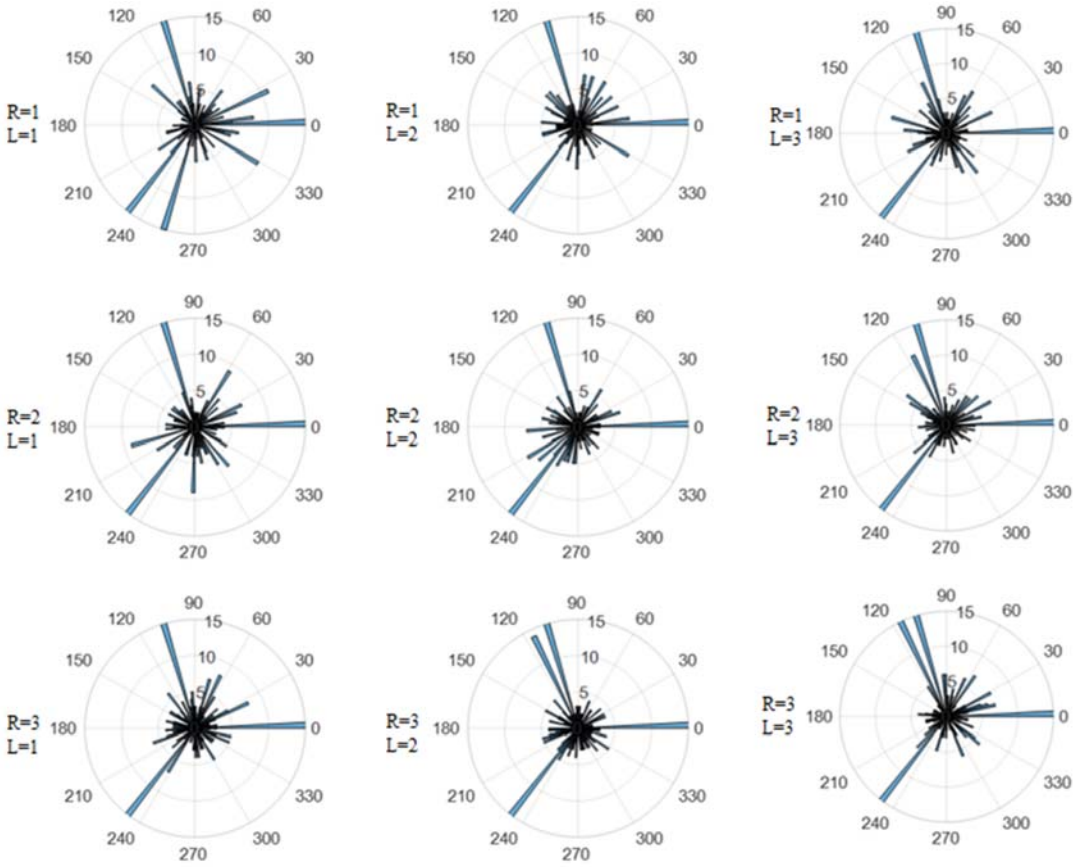
Karakterlerin yan yana gelme olasılıkları her dil için farklıdır. Aynı metnin farklı dillerdeki çevirilerine ait metinlerden çıkarılan açı örüntüler için histogramlar Şekil 9 ve gül histogramlar ise Şekil 9'da verilmiştir. Şekil 9'dan görüldüğü gibi her dil için dağılımın farklı olduğu, örüntülerin farklılık gösterdiği görülmektedir. Şekil 9'daki diller VS-1'den rastgele seçilerek verilmiştir. Dört farklı veri seti için R=1 ve L=1 için elde edilen örüntüler RF, SVM, LDA, NB ve Knn ile sınıflandırılmıştır. Sınıflandırma sonucundaki ulaşılan başarı oranları Tablo 4'te gösterilmiştir. Başarı oranlarına bakıldığında VS-1 için en iyi başarı %96,81 olarak NB sınıflandırıcı yöntemi ile edilmiştir. VS-2 için ise RF yöntemi ile %98,43 başarı gözlenmiştir. VS-3 için tabloya bakıldığında LDA ile %93,31 başarı oranı görülmektedir. Son olarak VS-4 için NB yöntemi ile %98,60 oranı gözlenmiştir. Genel olarak her veri seti için sınıflandırıcı yöntemler birbirine yakın performanslar sergilemişlerdir.

Farklı örüntülerin yakalanması için bu açı yöntemine ait R ve L parametrelerinin farklı değerleri için denemeler gerçekleştirilmiştir. Denemeler her veri seti için Tablo 1'deki en başarılı sınıflandırıcı yöntem ile gerçekleştirilmiştir. Veri setleri için farklı R ve L değerleri için elde edilen başarı oranları Tablo 5'te verilmiştir. En iyi R ve L değerlerine göre elde edilen performans değerleri ise Tablo 6'da verilmiştir. Başarı oranları ayrıca Şekil 10'da gösterilmiştir.

Tablo 5'e bakıldığında VS-1, VS-3 ve VS-4 için (R=1, L=1) iken en yüksek başarı oranları elde edilmiştir. VS-2 için en iyi başarı oranı (R=1, L=2) için gözlenmiştir. VS1, VS2 ve VS4 için yüksek başarı oranları gözlenmiştir. En düşük başarı oranları VS3 için elde edilmiştir. Bu veri seti için diğerlerine göre düşük çıkmasının nedeni veri setindeki her dil için metin uzunluklarının düzensiz olmasından kaynaklanmıştır. VS3(Wili-2018), veri seti farklı dillerdeki Wikipedia sitelerinden oluşturulmuştur. En uygun R ve L değerlerine denemeler sonucunda karar verilmelidir. Veri setine göre farklılık göstereceği düşünülmektedir. Tablo 5 ve/veya Şekil 10'da ayrıntılı incelendiğinde R değerleri artış gösterdiğinde başarı oranları çok az düşüş göstermektedir. Ancak L parametresi artış gösterdiğinde başarı oranlarının daha büyük oranda düşüş gösterdiği gözlenmiştir. Açı yönteminde teorik olarak 360 farklı örüntü oluşmaktadır. Ancak yan yana sıralı dizilmiş tekil kod karakterler arasındaki açı değerleri hiçbir zaman 0 veya 360 olmayacaktır. Bunu yanında DT için veri setlerinden çıkarılan tüm açı örüntülerin etkinliği aynı olmayacaktır. Başarı oranı üzerinde etkisiz özniteliklerin öznitelik matrisinden atılması gerekmektedir. (CfsSubsetEval+Greedy Stepvise) yöntemine göre en iyi öznitelikler elde edildikten sonra ulaşılan başarı oranları Tablo 7'de gösterilmiştir. Tablo 7'e bakıldığında VS-1 için öznitelik indirgeme işleminden sonra 87 öznitelik ile %93,61 başarı oranı elde edilmiştir. VS-2 için 100 adet öznitelik ile %98,73 gibi yüksek bir başarı oranı gözlenmiştir. VS-3 için ise 47 gibi az sayıda öznitelik ile %79,86 başarı oranı elde edilmiştir. Son olarak VS-4 için ise 64 öznitelik ile %96,57 başarı oranı gözlenmiştir. Veri setlerindeki metinler, metin uzunlukları birbirinden farklıdır. Dolayısıyla her veri seti için seçilen öznitelikler ve öznitelik sayıları birbirinden farklılık göstermiştir. Bu durum tüm özniteliklerin kullanılması ile elde edilen başarı oranlarına göre daha düşük başarı oranları gözlenmiştir. En çok düşüş VS3 için görülmüştür. Metin uzunluklarının açı yönteminin başarısına etkisini test etmek için tüm veri setleri için 100, 200, 300, 400 ve 500 karakter uzunluğundaki metinler için denemeler gerçekleştirilmiştir. Sınıflandırma sonucundaki ulaşılan başarı oranları Tablo 8 ve Şekil 11'de gösterilmiştir.



Şekil 8. R, L parametrelerinin farklı değerlerine göre elde edilen örüntüler
(Patterns obtained according to different values of R, L parameters)



Şekil 9. Farklı dillere ait bir metin için çıkarılan örüntülere ait gül histogramlar (1: Danimarka dili, 2: Flemenkçe, 3:Hindi, 4: İrlanda dili, 5:Japonca, 6:Korece, 7:Malayca, 8:Romence, 9:İsveççe, 10:Tükçe)
(Rose histograms of patterns extracted for text in different languages (1: Danish, 2:Dutch, 3:Hindi, 4:Irısh, 5:Japanese, 6:Korean, 7:Malay, 8:Romanian, 9:Swedish, 10:Turkish))

Tablo 4. Farklı sınıflandırıcılar için başarı oranları (Success rates for different classifiers)

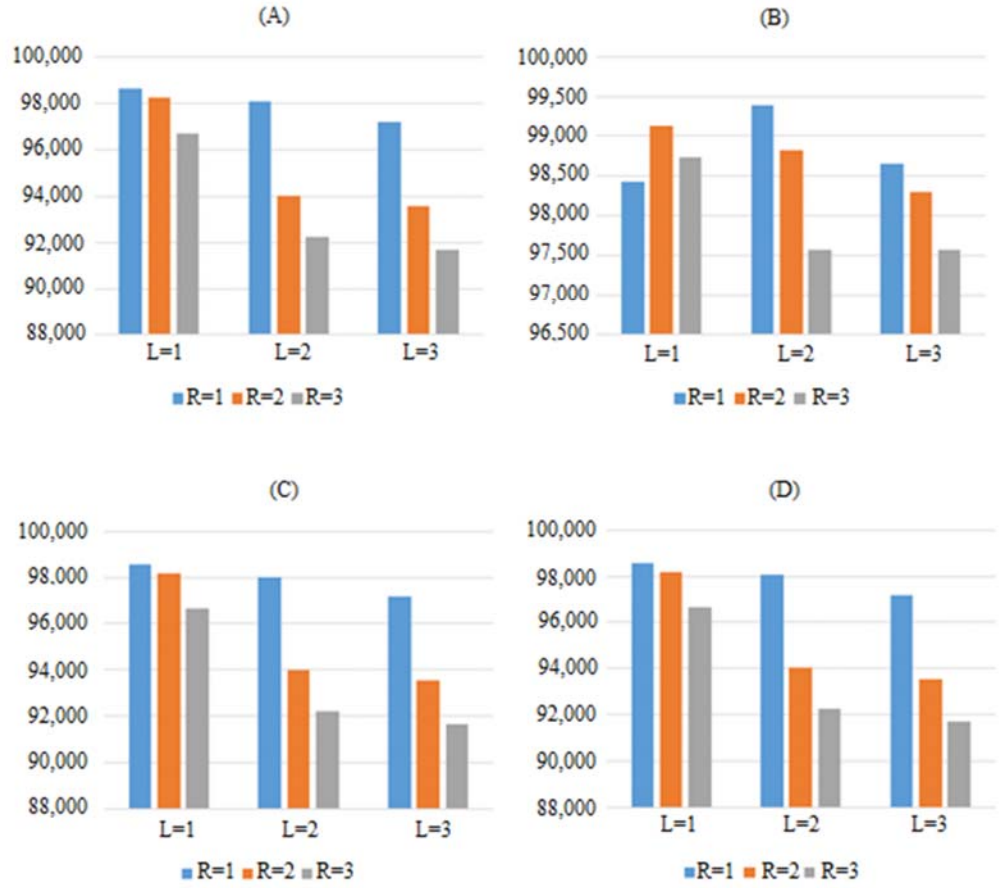
Veri Seti	RF	SVM	LDA	NB	Knn
VS-1	91,0364	96,768	96,7405	*96,8123*	94,05
VS-2	*98,4346*	96,265	96,4039	96,5449	95,7975
VS-3	91,1591	92,2955	*93,3182*	89,4545	89,5909
VS-4	95,6454	96,9673	96,112	*98,6003*	83,1726

Tablo 5. R ve L parametrelerinin farklı değerlerine göre başarı oranları (Success rates according to different values of R and L parameters)

R, L	VS-1	VS-2	VS-3	VS-4
R=1,L=1	*96,8123*	98,4346	*93,3182*	*98,6003*
R=1,L=2	96,3194	*99,3936*	90,7955	98,0560
R=1,L=3	94,9368	98,6603	89,4091	97,2006
R=2,L=1	96,2644	99,1257	91,1136	98,2115
R=2,L=2	92,0527	98,8295	86,4773	94,0124
R=2,L=3	89,0954	98,3077	84,3864	93,5459
R=3,L=1	89,0954	98,7449	89,7045	96,6563
R=3,L=2	88,9031	97,5744	84,8864	92,2240
R=3,L=3	85,9824	97,5744	82,9545	91,6796

Tablo 6. Performans Ölçütleri (Performance Metrics)

Veri Seti	Kesinlik (Precision)	Duyarlılık (Recall)	F-Ölçütü (F-Measure)	ROC Alan (ROC Area)
VS-1	0,987	0,986	0,986	0,986
VS-2	0,994	0,994	0,994	0,994
VS-3	0,937	0,933	0,934	0,931
VS-4	0,986	0,986	0,986	0,985



Şekil 10. Farklı R ve L değerlerine göre başarı oranları (Success rates according to different R and L values)

Tablo 7. Öznitelik seçiminden sonra elde edilen başarı oranları (Success rates after attribute selection)

Veri Seti	#Öznitelik	Accuracy	Kesinlik (Precision)	Duyarlılık (Recall)	F-Ölçütü (F-Measure)	ROC Alan (ROC Area)
VS-1	87	93,6184	0,937	0,936	0,936	0,935
VS-2	100	98,7308	0,987	0,987	0,987	0,987
VS-3	47	79,8636	0,815	0,799	0,803	0,795
VS-4	64	96,5785	0,966	0,966	0,966	0,963

Tablo 8. Farklı uzunlukta metinler için başarı oranları (Success rates for texts of different lengths)

Veri Seti	#karakter=100	#karakter=200	#karakter=300	#karakter=400	#karakter=500
VS-1	89,5898	95,8433	96,768	96,8504	96,8687
VS-2	92,5962	98,2936	99,0692	99,168	99,1539
VS-3	84,9773	92,7273	93,4318	93,3864	93,4318
VS-4	92,6128	97,5117	98,7558	98,7558	98,7558

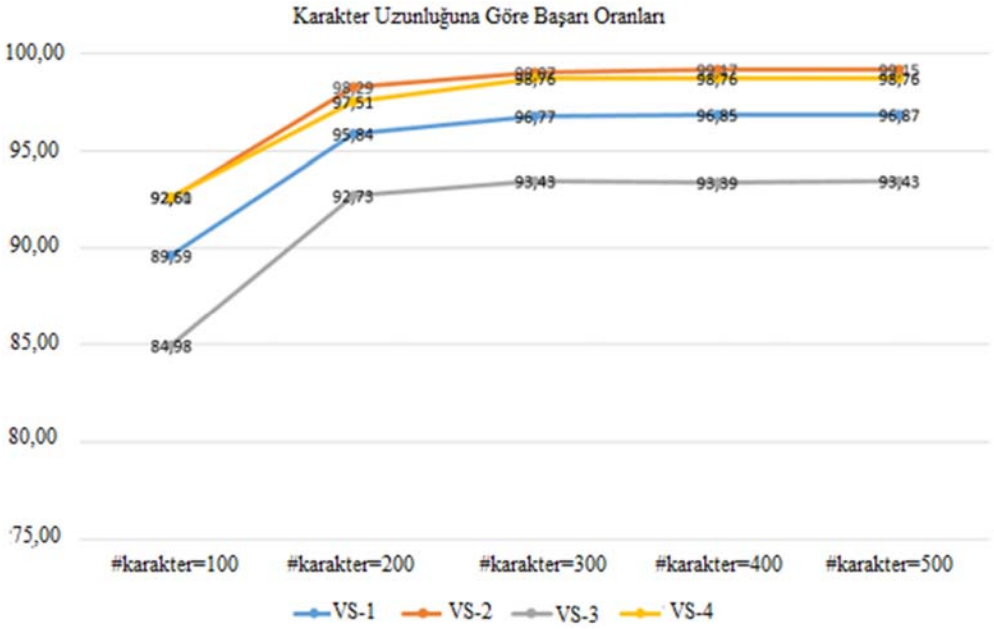
Tablo 8'den anlaşılacağı üzere DT uygulamasında başarı metin uzunluklarıyla doğrudan doğruya ilişkili olduğu görülmektedir. Başka bir ifadeyle, dil sınıflandırma başarısının metin karakter uzunluklarına bağlı olduğu söylenebilir. Karakter uzunluğu arttıkça başarı oranının da arttığı görülmektedir. Karakter uzunluğu 100, 200, 300, 400 ve 500 için kabul edilebilir başarı oranları gözlenmiştir.

Açı yönteminde 360 öznitelik ile DT işlemi gerçekleştirilmiştir. Açı yöntemi büyük boyutta öznitelik vektörleri oluşturan n-gram yöntemi ile karşılaştırılmıştır. VS-2 ve VS-4 için veri setlerinden 2-gram, 3-gram ve 4-gram ile gözlenen başarı oranları Tablo 9'da verilmiştir.

VS-2 için 2-gram ile 500265 kadar öznitelik çıkartılmıştır. VS-4 için ise 2-gram ile 17865 öznitelik çıkartılmıştır. VS-2 hem dil sayısı hem de metin sayısı daha fazla olduğundan yüksek sayıda öznitelik çıkartılmıştır. Tüm özniteliklerin sınıflandırma yöntemleri ile kullanılması hesaplama maliyeti açısından mümkün olmamıştır. Bu yüzden tüm n-gram değerleri için çıkartılan özniteliklerin frekansı yüksek ilk 1000, 2000 ve 3000 öznitelikler kullanılmıştır. Tablo 9'dan görüldüğü gibi önerilen yöntemin aksine n-gram ile yüksek başarı oranları gözlenmemiştir. Test ettiğimiz veri setleri üzerinde N-gram yönteminin hem performans hem de hesaplama maliyeti açısından önerilen yönteme göre başarısız kalmıştır.

Açı yöntemi DT için daha önce önerdiğimiz Bir (1) boyutlu yerel ikili örüntüler (1B-YİÖ) [13] yöntemi ile de karşılaştırılmıştır. Dört veri seti için 1B-YİÖ yöntemi ile çıkarılan öznitelikler NB ile sınıflandırılması sonucunda ulaşılan başarı oranları Tablo 10'da gösterilmiştir.

Sonuçlardan anlaşıldığı gibi genel olarak açı yöntemi 1B-YİÖ yönteminden daha başarılı görülmektedir. VS-2, VS-3



Şekil 11. Karakter uzunluklarına göre başarı oranları (Success rates based on character length)

Tablo 9. n-gram başarı oranları (n-gram success rates)

Veri Seti	Çıkarılan #Öznitelik	Kullanılan #Öznitelik	n=2 gram	n=3 gram	n=4 gram
		1000	21,5639	16,7521	18,7854
VS-2	500265	2000	46,3339	37,2562	42,8542
		3000	77,8305	68,4582	73,2561
		1000	52,9915	24,2424	32,8671
VS-4	17865	2000	78,0109	45,6099	41,3364
		3000	82,5952	72,9604	67,7540

Tablo 10. 1B-YİÖ metodu başarı oranları (1D-LBP method success rates)

Veri Seti	Başarı (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F-Ölçütü (F-Measure)	ROC Alan (ROC Area)
VS-1	97,419	0,975	0,974	0,974	1,000
VS-2	82,100	0,823	0,821	0,821	0,994
VS-3	85,340	0,851	0,853	0,848	0,993
VS-4	86,977	0,869	0,860	0,860	0,998

ve VS-4 için açı yöntemi daha başarılı sonuçlar vermiştir. Ancak VS-1 için 1B-YİÖ çok az farkla daha başarılı görülmektedir. Genel olarak açı yöntemi daha başarılı bulunmuştur.

6. SONUÇLAR VE TARTIŞMALAR (RESULTS AND DISCUSSIONS)

DT, bir metnin yazıldığı dili belirleme işlemidir. DT birçok doğal dil işleme uygulamaları (çeviri sistemleri, metin madenciliği, duygu analizi veya bilgi çıkarma) için önemli bir adımdır. Bu çalışmada, bir metnin yazıldığı dili tespit eden yeni bir yaklaşım önerilmiştir. Çalışmada, metinlerdeki karakterler UTF-8 tekil kod değerlerine göre sıralandıktan sonra bu değerler arasındaki açı bilgisine göre DT işlemi gerçekleştirilmektedir. DT için öncelikle açı yöntem ile metinlerden açı örüntüler elde edilmiş daha sonra bu örüntüler DT için RF, SVM NB, LDA ve Knn gibi farklı makine öğrenmesi yöntemler ile sınıflandırma işlemi gerçekleştirilmiştir. Önerilen yaklaşımı test etmek için farklı dil sayıları içeren 4 farklı veri kümesi kullanılmıştır. Veri setlerinden ikisi (VS-1 ve VS-2) tarafımızca oluşturulmuş ve diğer ikisi (VS-3 ve VS-4) herkese açık veri setleridir. Bu veri setleri için sırası ile %96,81, %99,39, %93,31 ve %98,60 başarı oranları gözlenmiştir. Bunun yanında önerilen yaklaşım karakter uzunluğuna göre test edilmiştir. #100, #200, #300, #400 ve #500 karakter sayısı içeren metinler için DT işlemi gerçekleştirilmiştir. Sonuçlar incelendiğinde metin uzunluğu arttıkça başarının daha iyi olduğu belirlenmiştir. Kabul edilebilecek seviyede bir başarı oranı elde etmek amacıyla metin uzunluklarının #200 karakter ve daha fazla uzunlukta olmasının gerektiği belirlenmiştir. Açı örüntüler yöntemi ile gözlenen sonuçlar literatürde sıklıkla kullanıldığı görülen n-gram yöntemine ait sonuçlar ile karşılaştırılmış ve önerilen yaklaşıma ait sonuçların daha iyi olduğu görülmüştür. Son olarak önerilen yöntem 1B-YİÖ yöntemi ile de karşılaştırılmıştır. Açı yöntemi 1B-YİÖ yönteminden de daha başarılı sonuçlar vermiştir. Bu çalışmada önerilen yaklaşımın DT uygulamasının yanında spam tanıma, metin kategorize etme gibi birçok farklı metin madenciliği uygulamalarında kullanılabileceğine düşünülmektedir. İleriki çalışmalarda, VS-1 ve VS-2 için metin sayıları ve dil sayıları artırılarak çeşitli denemeler gerçekleştirilecektir. Açı yöntemi farklı doğal dil işleme alanlarına uygulanacaktır.

KAYNAKLAR (REFERENCES)

1. Başkaya, F. & Aydin, İ., Classification of news texts by different text mining methods, In 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), 1-5, 2017.
2. Kul, S., Natural language processıng on the way to Turkısh lecturer artıfıcıal ıntellıgence, Journal of Management Information Systems, 6 (2), 43-56, 2020.
3. Ong, E.J., Cooper, H., Pugeault, N., Bowden, R., Sign language recognition using sequential pattern trees, Conference on Computer Vision and Pattern Recognition, Washington-USA, 2200–2207, 16-21 Haziran, 2012.
4. Aksu, M. Ç., Karaman, E., Comparison of fastText and Bag of Words Word Representation Methods by Using Turkish Reviews Conducted for Touristic Places, European Journal of Science and Technology, 20, 311-320, 2020.
5. Ali, C.B., Haddad, H., Slimani, Y., Empirical evaluation of compounds indexing for turkish texts, Computer Speech & Language, 56, 95-106, 2019.
6. Amasyali, M. F., Yıldırım, T., Automatic text categorization of news articles, Proceedings of the IEEE 12th Signal Processing and Communications Applications Conference, Kusadasi- Turkey, 224–226, 28-30 April, 2004.
7. Tang, B., He, H., Baggenstoss, P.M., Kay, S., A Bayesian Classification Approach Using Class-Specific Features for Text Categorization, IEEE Transactions on Knowledge and Data Engineering, 28 (6), 1602–1606, 2016.
8. Fragkou, P., Text segmentation for language identification in Greek forums, Procedia-Social and Behavioral Sciences, 147, 160-166, 2014.
9. Abainia, K., Ouamour, S., Sayoud, H., Effective language identification of forum texts based on statistical approaches, Information Processing & Management, 52 (4), 491-512, 2016.
10. Johnson, R., Zhang, T., Effective Use of Word Order for Text Categorization with Convolutional Neural Networks, arXiv:1412.1058v2, 2014.
11. Lui, M., Lau, J.H., Baldwin, T., Automatic detection and language identification of multilingual documents. Transactions of the Association for Computational Linguistics, 2, 27–40, 2014.
12. Cavnar, W.B., Trenkle, J.M., N-gram-based text categorization, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las VegasNevada-USA, 161–175, 11-13 April, 1994.
13. Kaya, Y., Ertuğrul, Ö.F., A novel feature extraction approach for text-based language identification: Binary patterns, Journal of the Faculty of Engineering and Architecture of Gazi University, 31 (4), 1085-1094, 2016.
14. Sarma, N., Singh, S.R., Goswami, D., Influence of social conversational features on language identification in highly multilingual online conversations, Information Processing & Management, 56 (1), 151-166, 2019.
15. Takçı, H., Ekinci, E., Minimal feature set in language identification and finding suitable classification method with it, Procedia Technology, 1, 444–448, 2012.
16. Gamallo, P., Pichel, J.R., Alegria, I., From language identification to language distance, Physica A: Statistical Mechanics and its Applications, 484, 152-162, 2017.
17. Takcı, H., Soğukpınar, İ., Letter based text scoring method for language identification, International Conference on Advances in Information Systems, İzmir-Türkiye, 283-290, 20-22 October, 2004.

18. Evans, D.A., Grefenstette, G.T., Tong X., Method of identifying the language of a textual passage using short word and/or n-gram comparisons, U.S. Patent No: US7359851, Washington, DC: U.S. Patent and Trademark Office, 15 April, 2008.

19. Popescu, M., Dinu, L.P., Kernel methods and string kernels for authorship identification: The federalist papers case, International Conference on Recent Advances in Natural Language Processing (RANLP-07), Borovets-Bulgaria, 27-29 September, 2007.

20. Popescu, M., Grozea, C., Kernel methods and string kernels for authorship analysis Notebook for PAN at CLEF, Conference and Labs of the Evaluation Forum, Rome-Italy, 17-20 September, 2012.

21. Popescu, M., Ionescu, R.T., The Story of the Characters, the DNA and the Native Language, Eighth Workshop on Innovative Use of NLP for Building Educational Applications, Atlanta-GA-USA, 270–278, 13 June, 2013.

22. Ahmed, B., Cha, S.H., Tappert, C., Language identification from text using n-gram based cumulative frequency addition, Proceedings of Student/Faculty Research Day, CSIS, Pace University, 12.1-12.8, 7 May, 2004.

23. Gary, F. Simons and Charles, D. Fennig, editors. Ethnologue: Languages of the World, Twentieth Edition. SIL International, Dallas, USA, 2017.

24. Acı, Ç., Çırak, A., Turkish News Articles Categorization Using Convolutional Neural Networks and Word2Vec, Journal of Information Technologies, 12 (3), 219-228, 2019.

25. Öztürk, Ö., Abidin, D., Özacar, T., Using classification algorithms for Turkish music makam recognition, Selcuk University Journal of Engineering, Science and Technology, 6 (3), 377-393, 2018.

26. Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., & Nissim, M., N-gram: New groningen author-profiling model, arXiv preprint arXiv:1707.03764, 2017.

27. Tohma, K., Kutlu, Y., Challenges Encountered in Turkish Natural Language Processing Studies, Natural and Engineering Sciences, 5 (3), 204-211 , 2020.

28. Tian, J., Suontausta, J., Scalable neural network based language identification from written text, In 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings.(ICASSP'03), 1, I-48, 2003.

29. Özcan T., Baştürk A., ERUSLR: A new Turkish sign language dataset and its recognition using hyperparameter optimization aided convolutional neural network, Journal of Gazi University Faculty of Engineering and Architecture, 36 (1), 527-542, 2020.

30. Kuncan F., Kaya Y., Kuncan, M., New approaches based on local binary patterns for gender identification from sensor signals, Journal of the Faculty of Engineering and Architecture of Gazi University, 34 (4), 2173-2185, 2019.

31. Li, G., Li, J., Ju, Z., Sun, Y., & Kong, J., A novel feature extraction method for machine learning based on surface electromyography from healthy brain, Neural Computing and Applications, 31 (12), 9013-9022, 2019.

32. Kuncan, M., Kaplan, K., Minaz, M. R., Kaya, Y., & Ertunc, H. M., A novel feature extraction method for bearing fault classification with one dimensional ternary patterns, ISA transactions, 100, 346-357, 2020.

33. Gumaei, A., Hassan, M. M., Hassan, M. R., Alelaiwi, A., & Fortino, G., A hybrid feature extraction method with regularized extreme learning machine for brain tumor classification, IEEE Access, 7, 36266-36273, 2019.

34. Takçı, H., Güngör, T., A high performance centroidbased classification approach for language identification, Pattern Recognition Letters, 33 (16), 2077-2084, 2012.

35. Xiao, D., Li, Y. K., Zhang, H., Sun, Y., Tian, H., Wu, H., & Wang, H., ERNIE-Gram: Pre-Training with Explicitly N-Gram Masked Language Modeling for Natural Language Understanding, arXiv preprint arXiv:2010.12148, 2020.

36. Suzuki, I., Mikami, Y., Ohsato A., Chubachi, Y., A language and character set determination method based on N-gram statistics, ACM Transactions on Asian Language Information Processing, 1 (3), 269-278, 2002.

37. Castro, D.W., Souza, E., Vitório, D., Santos, D., Oliveira, A. L., Smoothed n-gram based models for tweet language identification: A case study of the brazilian and european portuguese national varieties, Applied Soft Computing, 61, 1160-1172, 2017.

38. Zheng, L., Liang, B., Sign language recognition using depth images, 14th International Conference on Control, Automation, Robotics and Vision (ICARCV), Phuket-Thailand, 1-6, 13-15 Kasım, 2016.

39. Zhang, X., Zhao, J., LeCun, Y., Character-level Convolutional Networks for Text Classification, Advances in Neural Information Processing Systems, Curran Associates Inc., 649-657, 2015.

40. Güven Z., Diri B., Çakaloğlu T., Comparison of n-stage Latent Dirichlet Allocation versus other topic modeling methods for emotion analysis, Journal of the Faculty of Engineering and Architecture of Gazi University, 35 (4), 2135-2145, 2020.

41. Durmuş G., Soğukpınar İ., A novel approach for analyzing buffer overflow vulnerabilities in binary executables by using machine learning techniques, Journal of the Faculty of Engineering and Architecture of Gazi University, 34 (4), 1695-1704, 2019.

42. Yücesoy E., Nabiyev V.V., Determination of a speaker's age and gender with an SVM classifier based on GMM supervectors, Journal of the Faculty of Engineering and Architecture of Gazi University, 31 (3), 501-509, 2016.

43. Poutsma, A., Applying Monte Carlo techniques to language identification, In: Proceedings of Computational Linguistics in the Netherlands, 2001.

44. Binas, A., Markovian Time Series Models for Language Identification, Project Report, Available: http://www.cs.toronto.edu/ abinas/csc2515report.pdf (online), 2005.

45. Xafopoulos, A., Kotropoulos, C., Almpanidis, G., Pitas, I., Language identification in web documents using discrete HMMs, Pattern Recognition, 37 (3), 583-594, 2004.

46. Li, Q., Chen, Y.P., Personalized text snippet extraction using statistical language models, Pattern Recognition, 43 (1), 378-386, 2010.

47. Sibun, P., Reynar, J.C., Language identification: examining the issues, In: Proc.5th Symposium on Document Analysis and Information Retrieval, Las Vegas-Nevada-USA, 125–135, 15-17 April, 1996.

48. Song, Y., Dai, L., Wang, R.., An automatic language identification method based on subspace analysis, IEEE International Conference on Multimedia and Expo, New York-NY-USA, 598-601, 28 Jun - 03 Jul, 2009.

49. Takci H., Diagnosis of breast cancer by the help of centroid based classifiers, Journal of the Faculty of Engineering and Architecture of Gazi University, 31 (2), 323-330, 2016.

50. Sagiroglu, S., Yavanoglu, U., & Guven, E.N., Web based machine learning for language identification and translation. In Sixth International Conference on Machine Learning and Applications (ICMLA 2007), 280-285, 2007.

51. Selamat, A., Ng, C.C., Arabic script web page language identifications using decision tree neural networks, Pattern Recognition, 44 (1), 133-144, 2011.

52. Köklü M., Kahramanlı H., Allahverdi N., A new accurate and efficient approach to extract classification rules, Journal of the Faculty of Engineering and Architecture of Gazi University, 29 (3), 477-486, 2014.

53. Jo, T., Normalized table-matching algorithm as approach to text categorization, Soft Computing, 19 (4), 839–849, 2015.

54. Tan S., An effective refinement strategy for KNN text classifier, Expert Systems with Applications, 30 (2), 290-298, 2006.

55. Murthy, K.N., Kumar, G.B., Language identification from small text samples, Journal of Quantitative Linguistics, 13 (01), 57-80, 2006.

56. Jiang, C., Coenen, F., Sanderson, R., Zito, M., Text classification using graph mining-based feature extraction, Knowledge-Based Systems, 23 (4), 302-308, 2010.

57. Botha, G.R., Barnard, E., Factors that affect the accuracy of text-based language identification, Computer Speech & Language, 26 (5), 307-320, 2012.

58. Hayta, Ş.B., Takçı, H., Eminli M., Language Identification Based on n-Gram Feature Extraction Method by Using Classifiers, IU-Journal of Electrical & Electronics Engineering, 13 (2), 1629-1639, 2013.

59. Yavanoğlu U., Sağıroğlu Ş., Automatic web based language identification and translation system, Journal of the Faculty of Engineering and Architecture of Gazi University, 25 (3), 483-494, 2010.

60. Singh, A.K., Study of some distance measures for language and encoding identification, In Proceedings of the Workshop on Linguistic Distances, 63-72, 2006.

61. Gottron, T., Lipka, N., A comparison of language identification approaches on short, query-style texts, In European Conference on Information Retrieval, Springer, Berlin, Heidelberg, 611-614, March, 2010.

62. Baldwin, T., Lui, M., Language identification: The long and the short of the matter, In Human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics, 229-237, June, 2010.

63. Tromp, E., Pechenizkiy, M., Graph-based n-gram language identification on short texts, In Proc. 20th Machine Learning conference of Belgium and The Netherlands, 27-34, May, 2011.

64. Hakkinen, J., & Tian, J., N-gram and decision tree based language identification for written words, In IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU'01, 335-338, 2001.

65. Carreras, X., Chao, I., Padró, L., Padró, M., FreeLing: An Open-Source Suite of Language Analyzers, In LREC, 239-242, May, 2004.

66. Zhai, L.F., Siu, M., Yang, X., Gish, H., Discriminatively trained language models using support vector machines for language identification, In: IEEE Odyssey 2006: The Speaker and Language Recognition Workshop, 1–6, 2006.

67. Ljubesic, N., Mikelic, N., Boras, D., Language indentification: How to distinguish similar languages?, In 2007 29th International Conference on Information Technology Interfaces, 541-546, June, 2007.

68. Martin, T., The WiLI benchmark dataset for written language identification, https:// arxiv. org/ pdf / 1801 . 07779 . pdf, 2020.