Türkçe Metinlerde Yazar Doğrulama için Yazar Modelleme Yaklaşımı
Authorship Modelling Approach for Authorship Verification on the Turkish Texts

Pelin CANBAY
Bilgisayar Mühendisliği Bölümü
Hacettepe Üniversitesi
Ankara, Türkiye
pelin@cs.hacettepe.edu.tr

Ebru AKÇAPINAR SEZER
Bilgisayar Mühendisliği Bölümü
Hacettepe Üniversitesi
Ankara, Türkiye
ebru@hacettepe.edu.tr

Hayri SEVER
Bilgisayar Mühendisliği Bölümü
Çankaya Üniversitesi
Ankara, Türkiye
sever@cankaya.edu.tr

***Özetçe*— Yazar niteliklendirme, metin formundaki verileri analiz ederek metnin yazarı ile ilgili bilgi çıkarmayı amaçlayan ve yıllardan beri ilgi gören zorlu bir çalışma alanıdır. Bu alanda ele alınan verinin sınırlı olması bu çalışmaları daha da zorlaştırmaktadır. Yazar doğrulama adı altında yürütülen bu çalışmalara olan ihtiyaç, elektronik ortamlarda anonim yazarların çoğalmasıyla gün geçtikçe artmaktadır. Bu çalışmada Türkçe metinlerde yazar doğrulama problemine model tabanlı bir çözüm yaklaşımı sunulmuştur. Sunulan yaklaşım ile yazar doğrulama çalışmalarında dikkate alınması gereken başarı aralığının ne olması gerektiği tespit edilmiştir.**

Anahtar Kelimeler — yazar niteliklendirme, tanımlama, doğrulama, modelleme

***Abstract*— Authorship attribution which aims to extract information about an author by analyzing the text of the author is a challenging field that has been studied for years. This study becomes even more difficult when there is limited data on this field. The need for this study carried out under the name of Authorship Verification is increasing day by day with the increase of anonymous authors in the electronic environments. In this study, a model-based solution approach is presented for the authorship verification problem. With the presented approach, it was determined what should be the success interval to be considered in the authorship verification problem.**

Keywords — Authorship attribution, identification, verification, modeling

I. GİRİŞ

Bir dokümanın karakteristiğinin analiz edilmesi ile o dokümanın yazarı ile ilgili bilgi çıkarma işlemleri Yazar Niteliklendirme (Authorship Attribution) [1] veya Yazar Analizi (Authorship Analysis) [2] adı altında ele alınmaktadır. 19. yy'den beri istatistiksel ve hesaplamalı yöntemlerin kullanıldığı bu çalışmalar [3], farklı türlerde ve farklı kişiler tarafından sürekli olarak metinsel veriler üretilmesiyle günümüzde hala popülerliğini korumaktadır. Bu alanda 2007 yılından beri faaliyet gösteren PAN [4] organizasyonunun da her yıl alana özgü veri yayını yapması ve yapılan çalışmalar arası başarıyı sıralaması da alana ilgiyi arttırmıştır. Yazar niteliklendirme çalışmaları temelde 3 ana kola ayrılabilir. Bu alanda en yaygın ele alınan çalışmalar Yazar Tanımlama (Authorship Identification) çalışmalarıdır. Yazar tanımlamada genellikle bir grup yazara ait birçok doküman işlenerek o yazarlara ait sınıflar elde edilmeye çalışılır. Daha sonra yazarı bilinmeyen bir dokümanın bu aday yazarlardan hangisine ait olduğu sorgulanır. Sorgulanan doküman hangi yazarın sınıfına daha yakın olursa o yazara ait olarak etiketlenir. Yazar tanımlamada, başta İngilizce olmak üzere farklı diller üzerine başarılı çalışmalar yapılmaktadır [5]. Yazar Profil Çıkarımı (Author Profiling) [6], yazar niteliklendirme çalışmalarının ana kollarından biri olup, bir yazarın dokümanları üzerinden o yazara ait yaş, cinsiyet, psikolojik durum gibi bilgilerin çıkarıldığı çalışmalardır.

Yazar Doğrulama (Authorship Verification), yazar niteliklendirme çalışmalarının en önemli ve en zor koludur, çünkü tüm yazar niteliklendirme problemleri tek bir yazar doğrulama problemine indirgenebilir [7]. Adli bilişimin bir alt dalı olarak ele alınan bu çalışmalar, yapısı gereği gerçek dünya problemidir. Çünkü gerçek dünyada bir yazar ile ilgili çıkarım yapmak için arka planda her zaman o yazar tarafından yazılmış, bilgi çıkarılabilecek yüzlerce veya binlerce doküman yoktur. Yazar doğrulama çalışmaları temelde, yazarı bilinen bir doküman ile yazarı bilinmeyen bir dokümanı karşılaştırıp, ikisinin aynı yazar tarafından yazılıp yazılmadığı bilgisine ulaşmaya çalışır [7]. Çalışmaların zorluğu kullanılan veri setinin sınırlı olmasındandır.

Bu çalışmada, yazar doğrulama problemi Türkçe metinler üzerinden ele alınmıştır. Problemin çözümü için yazar modelleme yaklaşımı kullanılmıştır. 5 farklı öznitelik seti kullanılarak 12 yazar örüntüsü çıkarılmıştır. Elde edilen örüntüler yazarları temsil eden modeller olup söz konusu yazarlara ait dokümanlara olan uzaklıkları hesaplanmış ve yazar doğrulamada dikkate alınması gereken benzerlik aralığının ne olması gerektiği tespit edilmiştir. Bu çalışma ile hem yazar doğrulama çalışması olarak farklı bir yaklaşım ele alınmış, hem de Türkçe yazar niteliklendirme çalışmalarına katkı sağlanmıştır.

978-1-5386-1501-0/18/$31.00 © 2018 IEEE

II. İLGİLİ ÇALIŞMALAR

Yazar doğrulama probleminin çözümü pratik olarak diğer yazar niteliklendirme problemlerinin çözümünü de kapsayacağından, yazar niteliklendirme alanında yapılan tüm başarılı çalışmalar yazar doğrulama problemine katkı sağlamaktadır. Dolayısı ile Yazar doğrulama probleminin çözümü için yazar niteliklendirme alanındaki tüm çalışmalar kapsam dahilindedir.

Türkçe üzerine yazar niteliklendirmede yazar tanımlama alanında farklı öznitelik setleri, farklı veri setleri ve farklı algoritmalar kullanılarak başarılı çalışmalar yapılmıştır [8]. Türkçe köşe yazılarının kullanıldığı yazar tanımlama çalışmalarında, bu alanda en ayırt edici öznitelik setini ve en başarılı sınıflandırma metodunu belirlemek için çalışmalar yapılmıştır [9, 10]. Bir doktora tezi olarak ele alınan yazar tanımlama probleminde en ayırt edici öznitelik setinin belirlendiği çalışmanın çıktısı olan noktalamalar seti [11], bu çalışmada bir ön bilgi olarak kullanılacaktır (1.Set).

Uluslararası literatürde yazar niteliklendirme alanında geniş çaplı çalışmalar yapılmıştır [2, 3]. Farklı veri setleri [12], farklı dil ve konularda [13] yazar doğrulama çalışmaları da ele alınmıştır. Yazar doğrulama problemi sınırlı veri içerdiğinden bu problemin çözümü genellikle farklı problemlere benzetilerek çözülmeye çalışılmıştır. Yapı olarak doğrulama problemini tek sınıflı bir sınıflandırma problemi gibi ele alan çalışmalar [1, 14], sınıf içeriğini zenginleştirmek için metni parçalara ayırmak gibi [3] farklı yöntemler kullanmaktadır. Sahte dokümanlar ile farklı sınıflar elde edilip problemi çok sınıflı sınıflandırma problemine dönüştürerek çözen çalışmalarda da başarılı sonuçlara ulaşılmıştır [7, 15].

Bu çalışmada, yazar doğrulama problemlerinde çözüm aralığının ne olması gerektiği tespit edilmeye çalışılmıştır. Bu amaç doğrultusunda öncelikle bir grup yazara ait yazılar kullanılarak o yazarlara ait yazar modeli (yazı örüntüsü) elde edilmiştir. Elde edilen bu modeller ile bir dokümanın ait olduğu yazar modeli ile benzerliğine bakılarak yazar doğrulamada dikkate alınması gereken benzerlik aralığının ne olması gerektiğine karar verilmiştir. Böylece bir yazarın kendi yazıları kullanılarak elde edilen model ile söz konusu yazarın yazıları arasındaki benzerlikler karşılaştırılarak, yazar doğrulama çalışmalarında elde edilebilecek en düşük ve en yüksek benzerlik değerleri belirlenmiştir. Verilen 2 dokümanın hangi oranda benzerliği sonucunda aynı yazar tarafından yazılıp yazılmadığına karar verilebilir sorusu, bir grup deneysel çalışma sonunda cevaplanmıştır.

III. DENEYSEL ÇALIŞMA

Yazar doğrulama çalışmaları için başarı sınırlarının belirlenmesi amacıyla veri seti olarak, aynı alandan yazan (güncel-siyaset) 12 köşe yazarının 2012 – 2014 yılları arasında yazılmış ve rastgele seçilmiş 100'er yazısı kullanılmıştır. Toplanan bu veri setindeki her yazı 5 farklı öznitelik seti kullanılarak 5 farklı yapıda ele alınmıştır. Kullanılan öznitelik setlerinin hangi özniteliklerden oluştuğu Tablo 1'de gösterilmektedir.

Yazar modelleme aşamasında Tablo1'de verilen özniteliklerin frekansları alınıp, her öznitelik değeri Denklem (1) kullanılarak [0-1] aralığına ölçeklendirilmiştir. Denklem (1)'de X değeri öznitelikleri temsil etmektedir. Ölçeklenen değerler ile tüm dokümanların vektörel gösterimleri elde edilmiştir.

$$X' = (X - X_{min}) / (X_{max} - X_{min}) \quad (1)$$

Yazar modelleri, her yazar için o yazara ait dokümanların vektörel toplamları alınarak elde edilmiştir. 12 yazar için 5 farklı öznitelik seti ile 5 farklı model oluşturulmuştur. Yazar doğrulama çalışmalarında az sayıda doküman olduğu gerçeğinden yola çıkarak oluşturulan modeller ait olduğu yazarı temsil eden bir doküman gibi ele alınmıştır. Her bir yazar modeli ile söz konusu yazarın diğer dokümanları arasındaki uzaklık, Denklem (2) (kosinüs uzaklığı) kullanılarak ölçülmüştür.

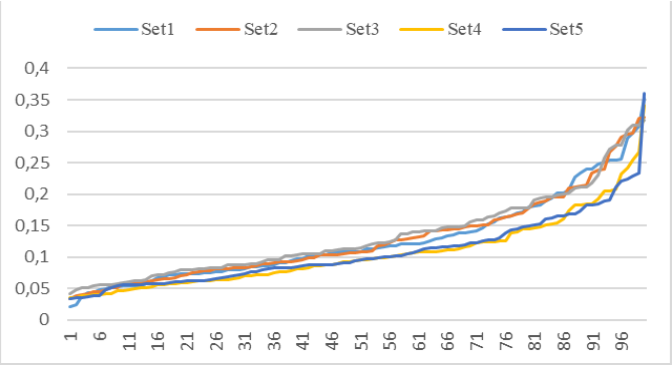
1.Set	**2.Set**	**3.Set**	**4.Set**	**5.Set**
Nokta Virgül SoruIsaret UcNokta TekTirnak CiftTirnak UnlemIsaret IkiNoktaUst NoktaliVirg Tire AltTire Slash TersSlash Parantez Ampersand	1.Set + Argo Osmanlı Kısaltma OzelIsim Yansima Zaman	2.Set + Devrik Edilgen Tek-kelimelik	3.Set + Cümle Ayrık-kelime Kelime Paragraf Sayı	4.Set + Bağlaç Edat Fiil Mastar İsim Sıfat

Tablo 1. Yazar Modellemede Kullanılan Öznitelik Setleri

Denklem (2)'deki A ve B değerleri dokuman veya model vektörlerini temsil etmektedir. A.B değeri iki vektörün noktasal çarpımını temsil ederken, $\|A\|.\|B\|$ değeri iki vektörün vektörel çarpımını temsil etmektedir. Şekil 1'de bir yazara ait 5 farklı öznitelik seti ile oluşturulmuş 5 farklı modelin o yazarın 100 dokümanına olan uzaklık değerleri gösterilmektedir.

$$\text{Kosinüs Uzaklığı} = 1 - (A.B)/(\|A\|.\|B\|) \quad (2)$$

Şekil 1'deki sonuçlar modellerin dokümanlara uzaklıklarını vermektedir. Hesaplanan kosinüs uzaklığının tersi model doküman arasındaki benzerliği verecektir. Kosinüs benzerliği açısından bakacak olursak, Şekil 1'de tüm modellerin yazarın dokümanları arasındaki benzerlik değeri %97 - %63 aralığındadır. Modeller arası uzaklık değerleri arasında büyük farklar olmamakla birlikte en iyi sonuç 4. öznitelik seti ile oluşturulan modelde elde edilmiştir. Diğer 11 yazar modellerinde de benzer sonuçlar elde edildiğinden grafiksel gösterimlerinin eklenmesine gerek duyulmamış ve 4 numaralı öznitelik seti yazar modeli oluşturmada en başarılı öznitelik seti olarak kabul edilmiştir.



Şekil 1. Bir yazara ait modellerin yazarın dokümanlarına olan uzaklıkları

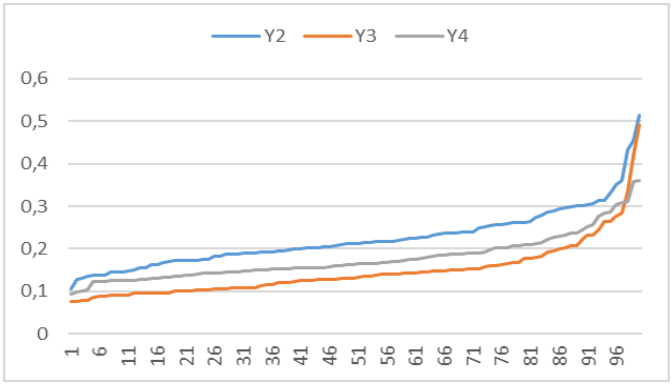
Bir yazarı temsil etmede kullanılan en başarılı modelin, yazar doğrulamada da yüksek başarı göstereceği varsayımından yola çıkarak yazar doğrulama çalışmalarında da 4 numaralı öznitelik setinin az farkla diğer setlerden daha başarılı sonuçlar vereceği çıkarımı yapılabilmektedir.

Şekil 1'deki grafik incelendiğinde yazarların bazı dokümanlarının standartların biraz uzağında olduğu görülmektedir. Örneğin 4.set için 100 dokümandan 97 tanesi %97 - %75 aralığında modelle benzerlik gösterirken 3 doküman %75 - %65 aralığında benzerlik göstermiştir. Şekil 1'deki yazar en iyi sonuçları veren yazardır. En kötü sonuçlarda ise yazar modeli 4.set için 100 dokümandan 92 tanesi %96 -%75 arası benzerlik gösterirken, 8 tanesi %75 - %50 arası benzerlik göstermiştir. Bu durum diğer 10 yazar için de benzer özelliklerdedir.

Bir yazarı temsil etmede kullanılan en başarılı modelin, yazar doğrulamada da yüksek başarı göstereceği varsayımından yola çıkarak yazar doğrulama çalışmalarında da 4 numaralı öznitelik setinin az farkla diğer setlerden daha başarılı sonuçlar vereceği çıkarımı yapılabilmektedir.

Yazar modeli ile modelin ait olduğu yazarın dokümanları arasındaki uzaklık düşük çıkmaktadır, bu beklenen ve istenen bir durumdur fakat modelin bir yazarı diğerinden ayırt etmede ne kadar başarılı sonuç verdiği de yazar doğrulama çalışmaları için önemli bir bilgidir. Bu bilgiye ulaşabilmek için Y1 yazarının 100 yazısının Y2, Y3 ve Y4 yazar modellerine olan uzaklık değerleri ölçülmüştür. Sonuçlar Şekil 2'deki grafikte gösterilmektedir.

Yazar modelleri ikili sınıflandırma söz konusu olduğunda ait oldukları yazarı temsil etmektedir çünkü elde edilen sonuçlar makuldür fakat yazar modelleri birbirine benzediğinden dolayı bir yazarın modeli başka yazarların dokümanları için de yakın benzerlik göstermektedir.



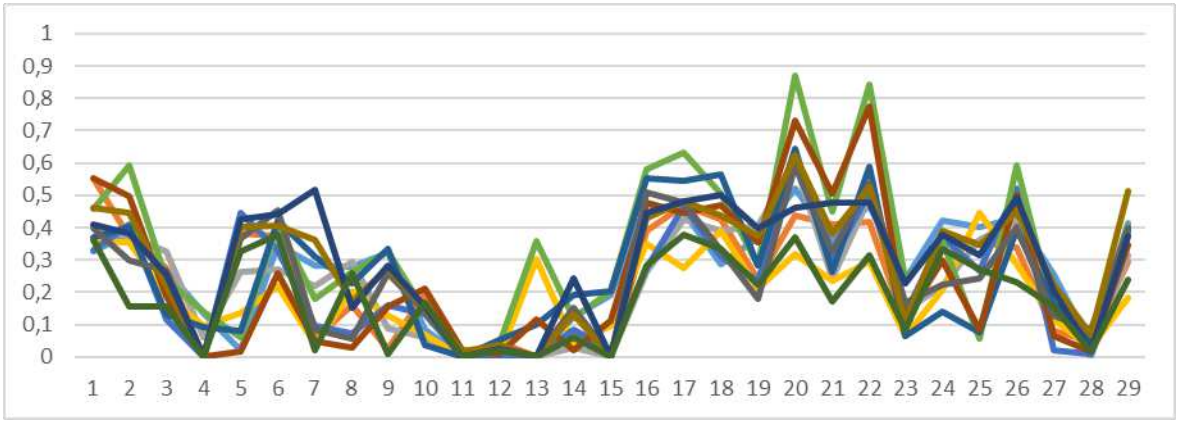
Şekil 2. Y1 yazarına ait dokümanların Y2, Y3 ve Y4 yazar modellerine olan uzaklıkları

Şekil 2'de Y1 ile Y2 arası benzerlik değerleri %88 - %48 arası, Y1 ile Y3 arası benzerlik değerleri %92 - %49 arası, Y1 ile Y4 arası benzerlik değerleri %90-62 arası çıkmıştır.

Ele alınan 29 öznitelik için yazar modellerinin aldığı değerler Şekil 3'te gösterilmektedir. Şekil 3'te görüldüğü üzere, belirlenen özniteliklerin elde edilen değerlerine göre veri setindeki yazarlar benzer örüntüler sergilemektedirler. Yani, veri setindeki yazarlar içerik, üslup ve yapı olarak benzer şekilde yazılarını oluşturmaktadırlar. Bu durum yazarlar arası ayırt ediciliğin başarısını düşürmektedir.

IV. TARTIŞMA VE SONUÇ

Yapılan deneysel çalışmalar sonucunda; yazar doğrulama çalışmalarında ele alınan iki doküman arası benzerlik değerinin, %100 ile %75 arası yüksek doğruluk, %75 ile %50 arası orta doğruluk, %50 ve aşağısı düşük doğruluk olarak değerlendirilebileceği sonucu kararlaştırılmıştır. Kullanılan öznitelik setleri ile oluşturulan yazar modelleri yazarları temsil etmede başarılı sonuçlar verirken, yazarlar arası ayırt edicilikte başarısız olmuştur. Bu durum hem ele alınan verilerin yapı ve işlev olarak benzer olmasından (güncel siyaset alanından köşe yazıları) hem de çıkarılan özniteliklerin bu veriler için yeteri kadar ayırt edicilik özelliğinde olmamasından kaynaklanmaktadır. İleriki çalışmalarda, Türkçe metinler için yazar doğrulama problemine farklı veri setleri ve farklı yaklaşımlar ile çözümler sunulacaktır.



Şekil 3. 12 yazar modelinin 29 öznitelik için aldığı değerler

KAYNAKLAR

[1] M. Koppel, J. Schler, and S. Argamon, "Computational methods in authorship attribution," Journal of the Association for Information Science and Technology, vol. 60, no. 1, pp. 9-26, 2009.

[2] S. E. M. El and I. Kassou, "Authorship analysis studies: A survey," International Journal of Computer Applications, vol. 86, no. 12, 2014.

[3] E. Stamatatos, "A survey of modern authorship attribution methods," Journal of the Association for Information Science and Technology, vol. 60, no. 3, pp. 538-556, 2009.

[4] P. S. Martin, Benno; Rosso, Paolo; Stamatatos, Efstathios (2018). PAN is a series of scientific events and shared tasks on digital text forensics. Available: http://pan.webis.de/index.html

[5] E. Stamatatos et al., "Overview of the author identification task at PAN 2015," in CLEF 2015 Evaluation Labs and Workshop Working Notes Papers, 2015, pp. 1-17.

[6] F. Rangel, P. Rosso, M. Potthast, B. Stein, and W. Daelemans, "Overview of the 3rd Author Profiling Task at PAN 2015," in CLEF, 2015, p. 2015: sn.

[7] M. Koppel and Y. Winter, "Determining if two documents are written by the same author," Journal of the Association for Information Science and Technology, vol. 65, no. 1, pp. 178-187, 2014.

[8] F. Türkoğlu, "Author Attribution of Turkish Documents with Hybrid Approaches," Msc, Computer Engineering, Yıldız Technical University, 2006.

[9] I. N. Bozkurt, O. Baglioglu, and E. Uyar, "Authorship attribution," in Computer and information sciences, 2007. ISCIS 2007. 22nd international symposium on, 2007, pp. 1-5: IEEE.

[10] F. Türkoğlu, B. Diri, and M. F. Amasyalı, "Author attribution of turkish texts by feature mining," in International Conference on Intelligent Computing, 2007, pp. 1086-1093: Springer.

[11] O. Aslantürk, "Turkish Authorship Analysis with an Incremental and Adaptive Model," Phd, Computer Engineering, Hacettepe University, 2014.

[12] M. L. Brocardo, I. Traore, and I. Woungang, "Authorship verification of e-mail and tweet messages applied for continuous authentication," Journal of Computer and System Sciences, vol. 81, no. 8, pp. 1429-1440, 2015.

[13] O. Halvani, C. Winter, and A. Pflug, "Authorship verification for different languages, genres and topics," Digital Investigation, vol. 16, pp. S33-S43, 2016.

[14] M. Koppel and J. Schler, "Authorship verification as a one-class classification problem," in Proceedings of the twenty-first international conference on Machine learning, 2004, p. 62: ACM.

[15] N. Potha and E. Stamatatos, "An Improved Impostors Method for Authorship Verification," in International Conference of the Cross-Language Evaluation Forum for European Languages, 2017, pp. 138-144: Springer.