Türkçe Dokümanlar İçin Yazar Tanıma

Özcan KOLYİĞİT, Rıfat AŞLIYAN, Korhan GÜNEL

Adnan Menderes Üniversitesi, Matematik Bölümü Bölümü, Aydın

okolyigit@gmail.com, rasliyan@adu.edu.tr, kgunel@adu.edu.tr

Özet: Yazar Tanıma çalışmaları, teknolojinin gelişmesi ve bilginin yaygınlaşması ile ortaya çıkan bir takım sorunlara çözüm üretmek için yapılmaktadır. Bu sorunlardan bazıları yazarı belli olmayan dokümanların yazarlarının belirlenmesi ve yazarının kim olduğundan tam olarak emin olunamayan metinlerin yazarlarının belirlenmesidir. Bu çalışmada Türkçe dokümanlar için yazar tanıma sistemi geliştirilmeye çalışılmıştır. Günlük gazetelerden seçilen 5 yazara ait köşe yazıları kullanılmıştır. Yazarların 70'er yazısından oluşan 350 dokümandan oluşan bir derlem hazırlanmıştır. Bu dokümanlardan 20'şer tanesi eğitim için 50'şer tanesi test için kullanılmıştır. İlk olarak 5 yazara ait dokümanlar toplanmış, daha sonra her yazara ait 20 doküman birleştirilerek tek bir doküman haline getirilmiştir. Bu şekilde elde edilen 5 doküman için sözcük ve gövde öznitelik vektörleri belirlenmiştir. Öznitelik vektörleri belirlenirken her yazar için vektör uzunlukları 20, 30, ve 40 olarak seçilmiş, oluşan öznitelik vektörleri için *K*-En Yakın Komşu algoritmasıyla test edilmiştir. Sonuç olarak, sözcük ve gövde öznitelik vektörlerine göre ortalama %77 başarı elde edilmiştir.

Anahtar Sözcükler: Yazar Tanıma, Veri Madenciliği, *K*-Enyakın Komşu Metodu, Metin Sınıflandırma.

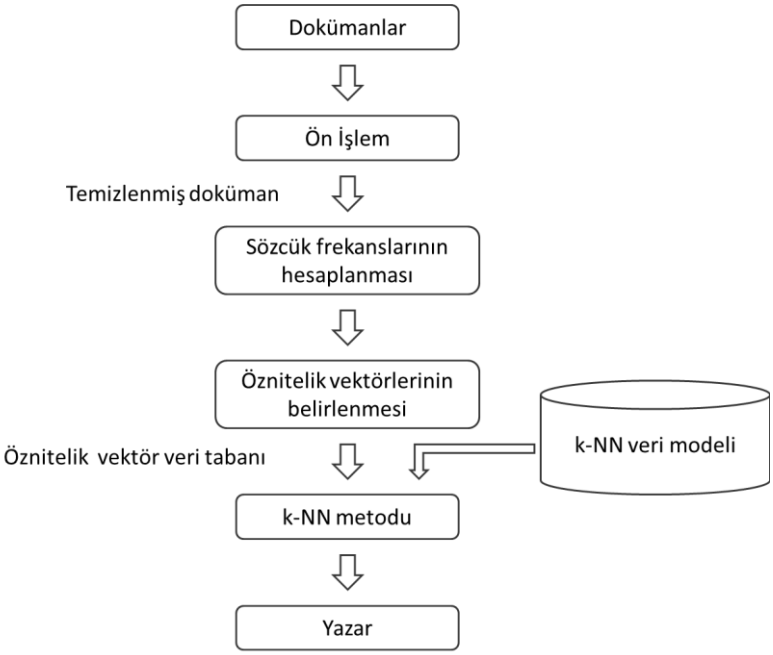
1. Giriş

Yazar tanıma çalışmalarında amaç yazarı bilinmeyen metinlerin yazarını tespit etmek veya yazarının kim olduğundan tam olarak emin olunamayan metinlerin yazarlarının belirlenmesidir [7]. Türkçe için yazar tanıma alanında ilk çalışmalar 1999 yılında yapılmaya başlanmış ve günümüzde, yapılan çalışmaların sayısı hızla artmaktadır.

Aynı doküman üzerinde, yazarlık iddia eden iki kişiden hangisinin dokümanın gerçek yazarı olduğunun tespiti için yazar tanıma uygulamalarından faydalanılır [4].

Bu çalışmada 5 yazara ait 70'er doküman, toplamda 350 doküman kullanılmıştır. Bu dokümanlardan 20'şer tanesi sistemin eğitilmesi için 50'şer tanesi sistemin test edilmesinde kullanılmıştır. Dokümanlar ilk olarak ön işlemden geçirilmiştir. Dokümanlardaki tüm harfler küçük harfe dönüştürülmüş, noktalama işaretleri ve rakamlar silinmiştir.

Daha sonra her yazarın herhangi 20 dokümanı birleştirilmiş ve dokümanlarda geçen sözcüklerin frekansları çıkarılmış ve değerler 0 ile 1 aralığında normalize edilmiştir. Her yazar için, o yazarın dokümanlarında yüksek frekansa sahip ve diğer yazarların dokümanlarında düşük frekansa sahip sözcükler ve sözcüğün en uzun gövdeleri [9] belirlenmiştir. Örneğin "kitaplıklarımızdan" sözcüğünün en uzun gövdesi "kitaplık" olarak tespit ediliyor. Bu sözcükler ve en uzun gövdeler öznitelik vektörleri olarak ele alınmıştır. Daha sonra test aşaması için ayrılan 50'şer doküman K-En Yakın Komşu metodu (*K*-NN) kullanılarak öznitelik vektör veri tabanındaki değerlerle karşılaştırılarak dokümana ait yazar belirlenmiştir. Sistemin yapısı genel olarak Şekil 1.1'de gösterilmiştir.



Şekil 1.1 Yazar tanıma sisteminin genel yapısı

3. Yazar Tanıma Çalışmaları

İlk yazar tanıma çalışmaları Stamatatos ve arkadaşları tarafından yapılmıştır [7]. Sözdizimsel stil özelliklerinin çeşitli kombinasyonlarını kullanarak dokümanların yazarlarını belirlemeye yönelik bir çalışma yapmışlardır. Yunanca dokümanlar üzerinde çalışmışlardır.

Peng ve arkadaşları, her yazarın en çok kullandığı belli sayıdaki *n*-gramlardan oluşan bir vektör oluşturmuş, daha sonra en yakın komşuluk algoritmasını kullanarak dokümanların yazarlarını belirleyen bir çalışma yapmışlardır [6].

Fung, Federalist yayınlarının yazarlık özelliklendirilmesi için Destek Vektör Makinesi sınıflandırıcısını kullanılmıştır. Çalışmada Federalist yayınlar "as", "of" ve "on" kelimelerinin üç boyutlu uzayında bir düzlemle ayrılmışlardır. Bir takım fonksiyonel kelimeler kullanılarak Destek Vektör Makinesi uygulanmış ve yayınlar birbirinden ayrılmıştır [4].

Diri ve Amasyalı, Türkçe metinler üzerinde ilk çalışmaları yapmışlardır. Dokümanın içeriğini ve belirlenen 22 farklı stil özelliğini kullanarak dokümanların yazarlarını belirleyen sınıflandırma yöntemleri ile çalışmışlardır. 18 yazara ait 20'şer dokümandan oluşan bir derlem oluşturmuşlardır.

Doküman içeriğine bağlı sınıflandırmada Naive Bayes metodunu kullanmışlardır. Stil özelliklerine göre sınıflandırmada kendi geliştirdikleri Automatic Author Detection for Turkish Text (AADTT) metodunu kullanmışlardır [3].

Diri ve Amasyalı, Türkçe metinlerde yazar, tür ve cinsiyete bağlı sınıflandırma yapan bir sistem geliştirmişlerdir. Bu çalışmalarında da Naive Bayes, Destek Vektör Makineleri, C 4.5 ve Rastgele Orman yöntemlerini kullanmışlardır [1].

Diri, Amasyalı ve Türkoğlu, farklı öznitelik vektörleri kullanarak Türkçe dokümanların yazarlarının belirlenmesini amaçlayan bir çalışma yapmışlardır. Türkçenin 2 ve 3-gram'larını, Türkçede sık geçen sözcükleri, dilbilgisel ve istatistiksel özellikleri kullanarak 10 farklı öznitelik vektörü çıkarmışlardır. Daha sonra yine Naive Bayes, Destek Vektör Makineleri, C 4.5 ve Rastgele Orman yöntemlerini kullanmışlardır [2].

3. K- En Yakın Komşu Algoritması

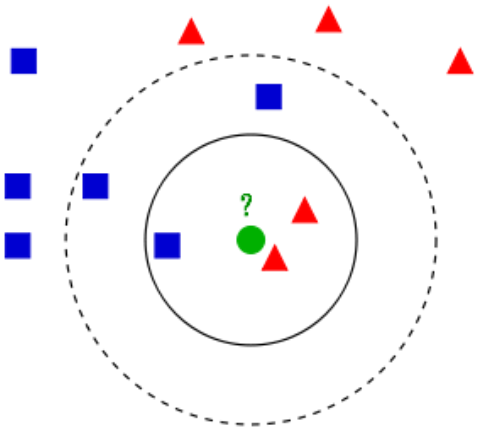
Sınıfları belli olan bir örnek kümesindeki gözlem değerlerinden yararlanarak, örneğe katılacak yeni bir gözlemin hangi sınıfa ait olduğunu belirlemek amacı ile *K*-En Yakın Komşu algoritması (*K*-Nearest Neighbors Algorithm) kullanılmaktadır.

Bu yöntem, örnek kümedeki gözlemlerin her birinin, sonradan belirlenen bir gözlem değerine olan uzaklıklarının hesaplanması ve en küçük uzaklığa sahip *K* sayıda gözlemin bulunduğu sınıfın seçilmesi esasına dayanmaktadır.

Örneğin, *K*=3 için yeni bir eleman sınıflandırılmak istensin. bu durumda eski sınıflandırılmış elemanlardan en yakın 3 tanesi alınır. Bu elamanlar hangi sınıfa dahilse, yeni eleman da o sınıfa dahil edilir.
Uzaklıkların hesaplanmasında Öklid uzaklık formülü kullanılabilir. Aralarındaki uzaklık hesaplanacak *i* ve *j* noktaları için aşağıdaki Öklid uzaklık formülü kullanılabilir:

$$d(i,j) = \sqrt{\sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2} \qquad (3.1)$$

Şekil 3.1' de *K*-NN algoritması ile ilgili basit bir örnek verilmiştir. Mavi karelerden ve kırmızı üçgenlerden oluşan iki sınıfımız olsun. Yeşil daire ise sınıfını belirlemek istediğimiz test verimiz olsun. Eğer *K*=3 seçilirse dairemize yakın iki üçgen bir kare olduğundan üçgen sınıfını seçmeliyiz. Fakat *K*=5 seçilirse dairemize yakın 3 kare 2 üçgen olduğundan kare sınıfını seçmeliyiz. Bu nedenle *K*'nın seçimi kritiktir.



Şekil 3.1 *K*-en yakın komşu algoritması

4. Sistemin Yapısı ve Uygulanması

Bu çalışmada ilk olarak 5 yazara ait dokümanlar toplanmış, daha sonra her yazara ait 20 doküman birleştirilerek tek bir doküman haline getirilmiştir. Bu şekilde elde edilen 5 dokümanlara göre sözcük ve en uzun gövde öznitelik vektörleri belirlenmiştir.
Öznitelik vektörü oluşturmak için öncelikle dokümanlar üzerinde, dokümanlarda geçen noktalama işaretleri ve sayılar temizlenmesi, tüm harflerin küçük harfe dönüştürülmesi işlemlerinin uygulandığı bir ön işlem yapılmıştır.

Ön işlem uygulanmış dokümanlardaki sözcüklerin ve en uzun gövdelerin frekansları normalize edilerek hesaplanmıştır. 5 yazara ait frekanslar oluşturulduktan sonra her yazar için diğer yazarlar tarafından daha az tercih

Kelime	Frekans
'takım'	45
'fenerbahçe'	23
'maç'	37
'galatasaray'	19
'golü'	15
'teknik'	15
'beşiktaş'	16
'ikinci'	25
'rağmen'	19
'pozisyon'	18
.	.
.	.
.	.
.	.
'karşısında'	10

Kelime	Frekans
'yüzde'	59
'ise'	42
'ile'	40
'gol'	33
'euro'	28
'ilk'	24
'avrupa'	21
'maçın'	20
'ancak'	20
'orta'	19
.	.
.	.
.	.
.	.
'sargent'	10

Kelime	Frekans
'kadın'	41
'hem'	22
'iş'	27
'var'	29
'dünya'	23
'milyar'	23
'başkanı'	21
'ilgili'	19
'yönetim'	16
'birlikte'	16
.	.
.	.
.	.
.	.
uluslararası'	9

Tablo 4.1 Yazarlara ait öznitelik vektörleri

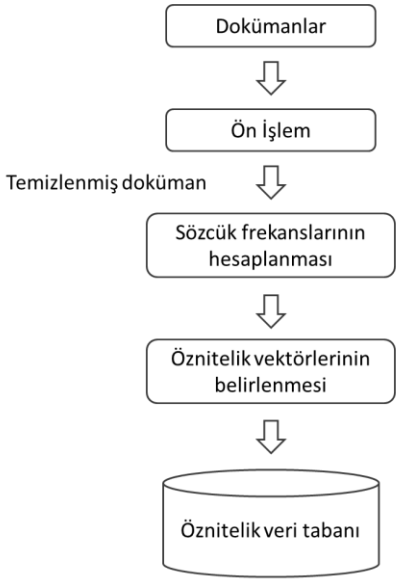
edilen (%50 daha az) sözcükler ve gövdeler seçilerek öznitelik vektörleri oluşturulmuştur.

Her yazar için oluşturulan öznitelik vektörleri birleştirilerek öznitelik veri tabanı oluşturulmuştur. Öznitelik veri tabanı oluşturulma aşamaları Şekil 4.1'de gösterilmiştir.

Her yazar için öznitelik sözcük ve en uzun gövde sayısı 20, 30 ve 40 seçilerek 3 farklı öznitelik vektörü oluşturulmuş, ayrı ayrı test edilmiştir.

Eğitim seti ve test seti oluşturulurken her dokümanın sözcük öznitelik vektöründeki 100, 150 ve 200 sözcük ve en uzun gövde için frekansları hesaplanmıştır. Oluşan 350 vektörün (her yazar için 70 adet) 20'şer tanesi eğitim için ayrılmış geri kalanları test için kullanılmıştır.

Test için *K*-NN metodu kullanılmıştır. Test setindeki her bir dokümanın eğitim setindeki her bir doküman ile arasındaki uzaklık Öklid uzaklık formülü kullanılarak hesaplanmıştır. *K*=1, *K*=3 ve *K*=5 için *K*-NN metodu uygulanmıştır.



Şekil 4.1 Öznitelik veri tabanının oluşturulması

5. Tartışma ve Sonuçlar

Bu çalışmada, veri madenciliği yöntemlerinden *K*-NN metodu kullanılarak Türkçe dokümanlar için yazar tanıma sistemi geliştirilmiştir. Günlük gazetelerden seçilen 5 yazarın yazılarından bir derlem oluşturulmuştur. Sözcük frekansları hesaplanarak sözcük öznitelik vektörleri oluşturulmuş, eğitim ve test için *K*-NN metodu kullanılarak dokümanların yazarları belirlenmeye çalışılmıştır. Öznitelik vektörü için seçilen sözcük ve en uzun gövde sayısı ve *K*'nın seçimine göre farklı başarı oranları elde edilmiştir. Tablo 5.1, 5.2, 5.3, 5.4, 5.5 ve 5.6'da başarı oranları ve ortalamaları verilmiştir.

Şekil 5.1'deki başarı oranlarına bakıldığında, *K*-NN metodunun *K*=1 alındığında, *K*=3 ve *K*=5 değerlerine göre daha başarılı olduğu görülmektedir. Öznitelik vektör boyu 20 olduğunda 30 ve 40'a göre daha başarılı sonuç vermiştir. En yüksek doğru tanıma oranı %77,2 olmuştur. Sözcük tabanlı ve en uzun gövde tabanlı sistemler arasında çok belirgin bir fark görülmemiştir. Her iki yaklaşıma göre en başarılı oran %77,2'dir.

Yazarlar	**Doğru Tanıma Başarı Oranları**		
	***K*=1**	***K*=3**	**K=5**
Yazar 1	% 70	% 64	% 70
Yazar 2	% 62	% 60	% 60
Yazar 3	% 98	% 98	% 96
Yazar 4	% 96	% 98	% 98
Yazar 5	% 60	% 58	% 56
Bütün Yazarlar	% 77,2	% 75,6	% 74

Tablo 5.1 Sözcük tabanlı *K*-NN metoduyla yazar tanıma başarı oranları (Öznitelik vektör boyu: 20)

Daha sonraki çalışmalarda, çok katmanlı algılayıcı ve destek vektör makinesi metotlarıyla uygulamalar geliştireceğiz ve başarı oranlarını karşılaştıracağız. Aynı zamanda, karakter, kök ve hece tabanlı yazar tanıma çalışmalarını yapacağız ve hangisinin en başarılı olduğunu tespit edeceğiz.

Yazarlar	**Doğru Tanıma Başarı Oranları**		
	***K*=1**	***K*=3**	***K*=5**
Yazar 1	% 70	% 66	% 54
Yazar 2	% 60	% 64	% 64
Yazar 3	% 100	% 100	% 96
Yazar 4	% 96	% 100	% 100
Yazar 5	% 54	% 52	% 48
Bütün Yazarlar	% 76	% 76,4	% 72,4

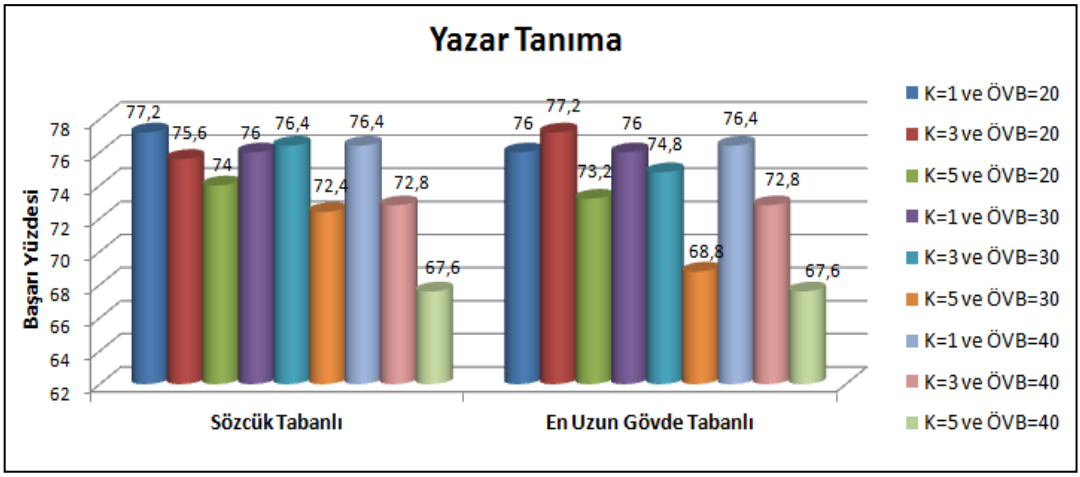
Tablo 5.2 Sözcük tabanlı *K*-NN metoduyla yazar tanıma başarı oranları (Öznitelik vektör boyu: 30)

Yazarlar	**Doğru Tanıma Başarı Oranları**		
	***K*=1**	***K*=3**	***K*=5**
Yazar 1	% 68	% 36	% 22
Yazar 2	% 52	% 54	% 44
Yazar 3	% 100	% 100	% 100
Yazar 4	% 88	% 98	% 98
Yazar 5	% 74	% 76	% 74
Bütün Yazarlar	% 76,4	% 72,8	% 67,6

Tablo 5.3 Sözcük tabanlı *K*-NN metoduyla yazar tanıma başarı oranları (Öznitelik vektör boyu: 40)

Yazarlar	**Doğru Tanıma Başarı Oranları**		
	***K*=1**	***K*=3**	***K*=5**
Yazar 1	% 58	% 54	% 40
Yazar 2	% 56	% 58	% 52
Yazar 3	% 100	% 100	% 100
Yazar 4	% 98	% 98	% 98
Yazar 5	% 68	% 76	% 76
Bütün Yazarlar	% 76	% 77,2	% 73,2

Tablo 5.4 En uzun gövde tabanlı *K*-NN metoduyla yazar tanıma başarı oranları (Öznitelik vektör boyu: 20)



Şekil 5.1 *K* ve ÖVB (Öznitelik Vektör Boyutu) değerlerine göre sözcük ve en uzun gövde tabanlı yazar tanıma başarı oranları

Yazarlar	**Doğru Tanıma Başarı Oranları**		
	***K*=1**	***K*=3**	***K*=5**
Yazar 1	% 58	% 46	% 28
Yazar 2	% 50	% 56	% 44
Yazar 3	% 100	% 100	% 100
Yazar 4	% 98	% 98	% 98
Yazar 5	% 74	% 74	% 74
Bütün Yazarlar	% 76	% 74,8	% 68,8

Tablo 5.5 En uzun gövde tabanlı *K*-NN metoduyla yazar tanıma başarı oranları (Öznitelik vektör boyu: 30)

Yazarlar	**Doğru Tanıma Başarı Oranları**		
	***K*=1**	***K*=3**	***K*=5**
Yazar 1	% 68	% 36	% 22
Yazar 2	% 52	% 54	% 44
Yazar 3	% 100	% 100	% 100
Yazar 4	% 88	% 98	% 98
Yazar 5	% 74	% 76	% 74
Bütün Yazarlar	% 76,4	% 72,8	% 67,6

Tablo 5.6 En uzun gövde tabanlı *K*-NN metoduyla yazar tanıma başarı oranları (Öznitelik vektör boyu: 40)

6. Kaynaklar

[1] Amasyalı M.F., Diri B., "Automatic Written Turkish Text Categorization in Terms of Author, Genre and Gender", **11th International Conference on Applications of Natural Language to Information Systems**, Austria (2006).

[2] Amasyalı M.F.,Diri B., Türkoğlu F., "Farklı Özellik Vektörleri ile Türkçe Dokümanların Yazarlarının Belirlenmesi", (2006).

[3] Diri B., Amasyalı M.F., "Automatic Author Detection for Turkish Texts", **Artificial Neural Networks and Neural Information Processing**, 138-141 (2003).

[4] Fung, G., Mangasarian, O., "The Disputed Federalist Papers: SVM Feature Selection via Concave Minimization", **Proceedings of the 2003 Conference of Diversity in Computing,** Atlanta, Georgia, USA, 42-46 (2003).

[5] Gerritsen, C.M., "Authorship Attribution Using Lexical Attraction", **Master Thesis Department of Electrical**

Engineering and Computer Science, MIT, (2003).

[6] Peng, F., Schuurmans, D., Keselj, V., Wang, S., "Language Independent Authorship Attribution using Character Level Language Models", **EACL**, 267-274 (2003).

[7] Stamatatos, E., Fakotakis, N., Kokkinakis, G., "Automatic Authorship Attribution", **EACL**, (1999).

[8] Stamatatos, E., Fakotakis, N., Kokkinakis, G., "Automatic Text Categorization in Terms of Genre and Author", **Computational Linguistics,** 471-495 (2000).

[9] Kut, A., Alpkoçak, A., Özkarahan, E., "Bilgi bulma sistemleri için otomatik Türkçe dizinleme yöntemi", **Bilişim Bildirileri**, Dokuz Eylül Üniversitesi, İzmir, (1995).