Orclever Proceedings of Research and Development
https://doi.org/10.56038/oprd.v4i1.470
OR CLEVER Science & Research Group

Konferans Bildirisi

Türkçe Şiir Verikümesi Üzerinde Denetimli Öğrenme ile Doğal Dil İşleme Tabanlı Şair Tanıma

Serkan Korkmaz [1]*, Fehim Köylü[2]

[1] Harran Üniversitesi, Orcid ID: https://orcid.org/0000-0003-2523-8819, E-mail: skorkmaz@harran.edu.tr,
[2] Erciyes Üniversitesi, Orcid ID: https://orcid.org/0000-0001-7991-5841, E-mail: fehimkoylu@erciyes.edu.tr
* Sorumlu Yazar: skorkmaz@harran@edu.tr;

4th International Conference on Access to Recent Advances in Engineering and Digitalization
May 27 - 28, 2024

Received 04 March 2024
In final form 21 May 2024

Reference: Korkmaz, S., Köylü, F. Türkçe Şiir Verikümesi Üzerinde Denetimli Öğrenme ile Doğal Dil İşleme Tabanlı Şair Tanıma. Orclever Proceedings of Research and Development, 4(1), 115-122.

Özet

Doğal dil işleme temelli çalışmalar günümüzde oldukça popüler durumda olmakla birlikte Türkçe temelli güncel çalışmalarda da artış görülmektedir. Yazar tanıma problemi, isimsiz bir metnin popüler yazarlardan birine ait olup olmadığının tespit edilmesine dayanmaktadır. Her bir yazarın eserinin yazara ait entellektüel bilgi dağarcığı hakkında temel bazı özellikleri yansıtacağı ve böylece yazarları birbirinden ayırabilmenin mümkün olabileceği fikri üzerine bu araştırma problemi ortaya çıkmıştır. Bu çalışmada Türk Edebiyatından 5 farklı şairin 50 şiiri alınmış ve bir verikümesi elde edilmiştir. Verikümesi üzerinde 9 farklı sınıflandırıcı yöntem kullanılarak deneyler gerçekleştirilmiştir. İlerleyen çalışmalara temel teşkil edecek bir ön çalışma ortaya konulmuştur.

Anahtar: Doğal dil işleme, metin madenciliği, denetimli öğrenme, destek vektör makinası, Bayes sınıflandırıcılar, karar ağaçları, rastgele orman

Orclever Proceedings of Research and Development

https://doi.org/10.56038/oprd.v4i1.470

OR CLEVER Science & Research Group

Conference Article

Natural Language Processing Based Poet Recognition with Supervised Learning on Turkish Poetry Dataset

Serkan Korkmaz [1]*, Fehim KÖYLÜ[2]

[1] Harran University, Orcid ID: https://orcid.org/0000-0003-2523-8819, E-mail: skorkmaz@harran.edu.tr,
[2] Erciyes University, Orcid ID: https://orcid.org/0000-0001-7991-5841, E-mail: fehimkoylu@erciyes.edu.tr
* Correspondence: : skorkmaz@harran@edu.tr;

Received 12 December 2023
Received in revised form 16 January 2024
Accepted 15 February 2024

4th International Conference on Access to Recent Advances in Engineering and Digitalization
May 27 - 28, 2024

Reference: will be edited by editorial

Abstract

Natural language processing-based studies become popular nowadays and Turkish based studies are increasing. The problem of author classification is based on determining whether an anonymous text belongs to one of the popular authors. This research problem is motivated by the idea that each author's work will reflect some basic features about the author's intellectual vocabulary and thus it should be possible to distinguish between authors. In this study, 50 poems of 5 different poets from Turkish Literature were taken and a dataset was obtained. Experiments were performed on the dataset using 9 different classifier methods. This is a preliminary study that will serve as a basis for future studies.

Keywords: Natural language processing, text mining, supervised learning, support vector machine, Bayesian classifiers, decision trees, random forest

https://doi.org/10.56038/oprd.v4i1.470

1. Giriş

Doğal Dil İşleme (DDİ), insanlar tarafından konuşulan dilin kurallarını çözümleyip bilgisayar-insan etkileşimini daha doğal bir hale getirebilecek en temel yapay zeka alanlarından biri olarak günümüzde karşımıza çıkmaktadır. İnsan dilinin karmaşık yapısını kavrayabilecek, insan dilini anlama ve doğal dilde karşılık üretme yeteneğini geliştirecek çalışmalar günümüzde yaygınlık kazanmıştır. IDC tarafından ortaya konulan rehberde sayısal olarak depolanan toplam veri miktarının 2018'de 33 zettabayttan (ZB) 2025 yılına kadar 175 ZB'a kadar artacağı ifade edilmiştir. Toplam verinin %20 gibi önemli bir kısmının web sayfaları, sosyal medya gönderileri vb. düz metin biçiminde olacağı tahmin edilmektedir. Dolayısıyla metin işlemeye dayalı DDİ yöntemleri oldukça önem kazanmıştır [1-3].

DDİ sosyal medya üzerinde yer alan mesajların analiz edilmesiyle devletler düzeyinde açık istihbarat çalışmalarına ve toplumların yönlendirilmesine kadar uygulama alanı bulabilmektedir. DDİ günlük hayatta birçok iş dalında insanların yerini alabilmesi için robotların gerekli olan insan etkileşimi için hem söyleneni anlamlandırması, duyguyu anlaması ona uygun veriyi bulup gerekli davranışı sergilemesi sağlanabilecektir. Farklı dillerdeki doğal dilin yapısının analiz edilmesi ve birbiri ile ilişkilendirilebilmesi ile günümüzde video ve ses uygulamalarının eş zamanlı çevirilerle farklı dillere tercüme edilmesi hatta sesli ve videolu hale dönüştürülmesi mümkün olabilmektedir. Yapılandırılmamış düz metin verisi önişlemlerden geçirilip DDİ yöntemleri ile amaca yönelik olarak analiz edilir. DDİ alanında bu dönüştürme işlemi için sesbilim, biçimbilim, sözdizim, anlambilim ve faydabilim olmak üzere farklı seviyelerde yöntemler geliştirilmiştir [4-10].

Yazar tanıma problemi, anonim eserlerin veya yazarının kim olduğu konusunda şüpheye düşülen edebi metinlerin yazarının bilgisayar destekli olarak belirlenmesi olarak ifade edilir [5]. Herhangi bir metin asıl yazarının adı geçmeden izinsiz biçimde farklı kişilerin adıyla paylaşılabilmektedir. Bundan dolayı bir metnin kim tarafından yazılmış olduğunun belirlenmesi önemlidir. Doküman içeriğine metin sınıflandırma uygulanmasına rağmen metin sınıflandırmadan farklı olarak uygulanan bir sınıflandırıcı tekniğidir [6,7]. Kemal Oflazer 1994 yılında yayınladığı makalesinde, Türkçe kelime yapılarının iki seviyeli morfolojik tanımını tam olarak açıklamaktadır. Yaklaşık 23.000 kök kelimeden oluşan bir kök kelime sözlüğü ortaya çıkarılmıştır. Sonraki çalışmalarda sentetik sözcük üretimi yöntemi, n-gram öznitelik çıkarımı, öznitelik vektör çıkartımı yöntemlerine dayalı çalışmalar yapılmıştır [8-14].

https://doi.org/10.56038/oprd.v4i1.470

Bu çalışmada istatistiksel olarak şiirlerin analiz edilmesi ve şairlerin belirlenmesi amacıyla deneysel bir çalışma yapılmıştır. Bu çalışmada Türkçe dilinde yazılmış ve Türk edebiyatçısı beş şaire ait toplam elli farklı şiir derlenmiştir. Makalenin geri kalan kısımları şu şekilde yapılandırılmıştır. İkinci bölümde toplanan veri kümesi ve kullanılan sınıflandırma yöntemleri açıklanmıştır. Üçüncü bölümde toplanan veri kümesi üzerindeki yapılan deneylerin sonuçları verilmiştir. Çalışmanın sonuçları ve gelecekte yapılacak çalışmalar dördüncü bölümde verilmiştir.

2. Materyal ve Yöntem

Öncelikle her bir yazar için 10 şiir olmak üzere 5 farklı şaire ait toplam 50 eser metni toplanmış ve uygun biçimde temizlenerek dönüştürülmüştür. Daha sonra çeşitli sınıflandırıcı yöntemlerle yazar tanıma deneyleri gerçekleştirilmiştir.

2.1. Veri kümesi

Türk edebiyatı şairlerinden Nahit Ulvi Akgün, Melih Cevdet Anday, Neşet Ertaş, Yılmaz Güney ve Neyzen Tevfik olmak üzere beşi seçilmiş ve her birine ait onar şiir metni toplanmıştır. Bu metinler içinde geçen kelimeler üzerinde önişlem yapılarak, metindeki ekler temizlenmiş ve kelime kökleri çıkartılmıştır. Daha sonra bütün köklerin listesi çıkartılarak en çok kullanılan 219 terim seçilmiştir. Her bir eser içinde geçen terimlerin frekansları öznitelik değeri olarak yer almıştır. Böylece her bir eser 219 giriş özniteliği ve eserin ait olduğu şair bilgisi de çıkış özniteliği olarak veri kümesinde yer almıştır. Böylece 5 sınıflı 50 örnekli ve 220 özniteliğe sahip bir seyrek matris veri kümesi elde edilmiştir. Seyrek matris biçimine getirilen veri kümesinin sınıfı eser hangi şaire ait ise ona ait bir kod numarası olacak biçimde 5 farklı değer içermektedir.

2.2. Yöntem

Çok terimli saf Bayes sınıflandırıcı (MNB), Bernoulli saf Bayes sınıflandırıcı (BNB), Gauss saf Bayes sınıflandırıcı (GNB), doğrusal destek vektör makinası (SVM), Nu-destek vektör makinası (NuSVM), C-destek vektör makinası (CSVM), karar ağacı (DT), rastgele orman (RF), Aşırı rastgele ağaç (extremely randomized trees, ERT) sınıflandırma algoritmaları uygulanmıştır [15-19]. Bayes sınıflandırıcılar Bayes teoremine dayalı olarak olasılık tabanlı sınıflandırma yapmaktadır. Saf Bayes sınıflandırıcılar özellikle metin sınıflandırma işlemlerinde yaygın bir kullanımı bulunmaktadır. Karar destek makinaları çok boyutlu giriş özniteliklerinin her birini sınıflara göre değerlerini ayıracak biçimde hiper düzlemler tespit ederek çalışmaktadır. Bir çekirdek fonksiyonuna bağlı olarak öznitelik değerlerinin sınırını belirlemektedir. Karar ağaçları sınıflandırma probleminin

https://doi.org/10.56038/oprd.v4i1.470

çözümü için öznitelikler ve aldıkları değerlerden meydana gelen bir karar ağacı keşfi algoritmasıdır. Veri kümesini homojen biçimde aynı sınıfa sahip örnekler alt kümesi haline getirecek biçimde farklı yapraklara bölmektedir. RF ve ERT algoritmaları topluluk öğrenme algoritması olarak karar ağaçlarına dayalı geliştirilmiş yöntemlerdir.

Her bir yöntemle elde edilen sınıflandırma sonuçları kesinlik (precision), duyarlılık (recall) ve F1 skoru (F1 score) ölçütleri ile değerlendirilmiştir. Kesinlik doğru tahmin edilen pozitif gözlemlerin, tahmin edilen toplam pozitif gözlemlere oranı ölçüsüdür. Pozitif tahmin değeri (positive predictive value) olarak da ifade edilmektedir. Kesinlik değerinin düşük olması, fazla miktarda yanlış pozitif (FP) değeri olduğunu ifade eder. Duyarlılık, doğru tahmin edilen pozitif örneklerin sayısının, toplam pozitif örnek sayısına oranıdır. Sınıflandırıcının bütünlüğünün bir ölçüsü olarak değerlendirilir. Yanlışlıkla tespit edilemeyen (FN) örnek oranı fazla olduğunda duyarlılık değeri düşük olacaktır. F1 skoru, kesinlik ve duyarlılık değerlerinin harmonik ortalamasıdır ve yanlış pozitif (FP) ve yanlış negatif (FN) hatalı sınıflandırma değerlerinin her ikisini de hesaba katmaktadır. Sınıf dağılımlarının düzensiz olduğu durumlarda F1 skoru değerine bakılması, doğruluk değerine bakılmasından daha anlamlı olmaktadır [20].

3. Sonuçlar

Veri kümesi bir önceki bölümde detayları verilen 9 farklı denetimli öğrenme algoritması ile eğitilmiştir. Her bir sınıflandırma algoritmasının başarıları incelenmiş ve elde edilen sonuçlar birbirleriyle karşılaştırılmıştır. Sınıflandırıcıların örnekler karşısında genel performanslarını tespit etmek amacıyla 10-kat çapraz doğrulama yöntemi kullanılmıştır. Tablo I.'de her bir sınıf için ayrı olacak biçimde ve genel ortalama olarak 10-kat doğrulama yöntemi ile elde edilen kesinlik değerleri verilmiştir. Tablodan da görüldüğü üzere veri kümesinin genel ortalaması %96 ile en yüksek Nu-destek vektör makinası, en düşük %60 Gauss saf Bayes sınıflandırıcı yöntemi ile elde edilmiştir. Tablo II'de her bir sınıf için ayrı olacak biçimde ve genel ortalama olarak 10-kat doğrulama yöntemi ile elde edilen duyarlılık değerleri verilmiştir. Tablodan da görüldüğü üzere veri kümesinin genel ortalaması %74 ile en yüksek Nu-destek vektör makinası, en düşük %52 ile Bernoulli saf Bayes sınıflandırıcı yöntemi ile elde edilmiştir. Tablo III'de her bir sınıf için ayrı olacak biçimde ve genel ortalama olarak 10-kat doğrulama yöntemi ile elde edilen F1 skoru değerleri verilmiştir. Tablodan da görüldüğü üzere veri kümesinin genel ortalaması %68 ile en yüksek Nu-destek vektör makinası, en düşük %51 ile Bernoulli saf Bayes sınıflandırıcı yöntemi ile elde edilmiştir.

https://doi.org/10.56038/oprd.v4i1.470

Tablo 1: Sınıflandırıcı Performansları Kesinlik (Precision) Değerleri.

Yazar Sınıfı	MNB	BNB	GNB	LSVM	NuSVM	CSVM	DT	RF	ERT
N.U. Akgün	%43	%50	%86	%42	100%	%71	%38	%57	56%
M. C. Anday	%60	100%	%91	%80	%80	%83	%57	100%	100%
N. Ertaş	%.83	100%	%62	100%	100%	%88	100%	100%	100%
Y. Güney	%80	100%	%78	%70	100%	100%	%83	100%	100%
N. Tevfik	100%	%39	%86	%44	100%	%47	%45	%37	45%
Ortalama	71%	75%	60%	64%	96%	77%	71%	78%	83%

Tablo 2: Sınıflandırıcı Performansları Duyarlılık (Recall) Değerleri.

Yazar Sınıfı	MNB	BNB	GNB	LSVM	NuSVM	CSVM	DT	RF	ERT
N.U. Akgün	%60	%50	%60	%50	%50	%90	%30	%50	%50
M.C. Anday	%60	%50	%100	%40	%100	%50	%40	%40	%40
N. Ertaş	%100	%80	%100	%50	%70	%70	%80	%80	%100
Y. Güney	%80	%40	%70	%70	%50	%60	%50	%70	%80
N. Tevfik	%40	%90	%60	%80	%90	%100	%90	%90	%90
Genel Ortalama	%66	%52	%6	%58	%74	%72	%60	%56	%66

Tablo 3: Sınıflandırıcı Performansları F1 Skor Değerleri.

Yazar Sınıfı	MNB	BNB	GNB	LSVM	NuSVM	CSVM	DT	RF	ERT
N. U. Akgün	%50	%50	%71	%45	%56	%59	%33	%50	%53
M. C. Anday	%60	%65	%95	%53	%59	%62	%47	%53	%53
N. Ertaş	%91	%89	%77	%67	%74	%78	%89	%89	%100
Y. Güney	%80	%57	%74	%70	%62	%75	%62	%70	%80
N. Tevfik	%57	%55	%71	%57	%62	%62	%60	%51	%60
Genel Ortalama	%65	%51	%58	%58	%68	%61	%61	%55	%66

4. Tartışma

Bu çalışmada yazar tanıma problemi için metin üzerinden çıkarılan öznitelikler ile şiir özelinde eser sahibinin tahmin edilmesine çalışılmıştır. Bunun için beş farklı şaire ait onar farklı şiir alınmış, ön işlemden geçirildikten sonra elde edilen veri kümesi üzerinde 9 farklı sınıflandırma yöntemi denenmiş ve sonuçlar karşılaştırılmıştır. Hem yanlışlıkla dahil edilen hem de dışlanan örnekleri bir arada değerlendiren F1 skoruna bakıldığında

elde edilen sonuçlar %51 ile %68 arasında değişmektedir. Dolayısıyla yeterli başarı elde etmek amacıyla çalışmanın daha yüksek performanslı sonuçlarını görmek amacıyla daha fazla sınıf ve örnek içerecek biçimde veri kümesinin geliştirilmesi ve derin öğrenme yaklaşımlarına dayalı güncel modellerin kullanılması sonraki çalışmalar olarak planlanmıştır.

5. Teşekkür

Bu çalışma Erciyes Üniversitesi Bilimsel Araştırma Projeleri Koordinasyon Birimi (BAP) tarafından FYL-2020-9769 nolu proje kapsamında desteklenmiştir. Yazarlar olarak Erciyes Üniversitesi BAP birimine desteklerinden dolayı teşekkür ederiz.

Referanslar

[1] D. Reinsel, J. Gantz, J. Rydning, "The digitization of the world from edge to core". International Data Corporation, 16, 1-28, 2018.
[2] A. Oğuzlar, "Temel Metin Madenciliği". Dora Yayınları, 2011.
[3] E. Adalı, "Türkçe Doğal Dil İşleme". Akçağ Yayınları, 2020.
[4] Z. Korkmaz, "Türkiye Türkçesi Grameri Şekil Bilgisi". Türk Dil Kurumu Yayınları, 2009.
[5] C. M. Stamatatos, "Automatic authorship attribution". Ninth Conference of the European Chapter of the Association for Computational Linguistics, 1999.
[6] D. Ünal, Ş. E. Şeker, "Metin Madenciliğinde Yazar Tanıma (Author Recognition in Text Mining)". BS Ansiklopedisi, 2018. Ninth Conference of the European Chapter of the Association for Computational Linguistics, 1999.
[7] F. Mosteller, D. L. Wallace, "Applied Bayesian and Classical Inference: The Case of the Federalist Papers". Addison-Wesley, 1984.
[8] K. Oflazer, Two-level description of Turkish morphology. In Literary and linguistic computing, volume 752, pages 137-148. Madison, WI, 1998.
[9] G. Cebiroğlu, "Sentetik Türkçe Sözcük Kökleri Üretimi". International XII. Turkish Symposium on Artificial Intelligence and Neural Networks–TAINN, 2003.
[10] İ. Büyukkuşcu, E. Adalı, "Heceleme Yöntemiyle Kök Sözcük Üretme". International XII. Turkish Symposium on Artificial Intelligence and Neural Networks–TAINN, 2003.
[11] C. M. Tan, Y. F. Wang, et al. "The use of bigrams to enhance text categorization". Information Processing & Management 38(4), 2002.
[12] B. Diri, F. Amasyalı, "Automatic author detection for Turkish texts". Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP), 2003.
[13] F. Amasyalı, B. Diri, et al. "Farklı özellik vektörleri ile Türkçe dokümanların yazarlarının belirlenmesi". 15th Turkish Symposium on Artificial Intelligence and Neural Network, Muğla, Türkiye, 2006.

https://doi.org/10.56038/oprd.v4i1.470

[14] İ. N. Bozkurt, O. Baghoglu, et al. "Authorship attribution performance of various features and classification methods". 22nd International Symposium on Computer and Information Sciences, 2007.

[15] A. McCallum, K. Nigam, et al. A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization, volume 752, pages 41–48. Madison, WI, 1998.

[16] B. Schölkopf, A. J. Smola, et al. "New support vector algorithms." Neural Computation, 12(5):1207–1245, 2000.

[17] L. Breiman, J. H. Friedman, et al. Classification And Regression Trees. Routledge, October 2017.

[18] L. Breiman. Random forests. Machine learning, 45:5–32, 2001.

[19] P. Geurts, D. Ernst, L. Wehenkel. Extremely randomized trees. Machine learning, 63:3–42, 2006.

[20] E. Şahin. Makine öğrenme yöntemleri ve kelime kümesi tekniği ile İstenmeyen e-posta/e-posta sınıflaması. Master's thesis, Hacettepe Üniversitesi, 2018