Marmara Fen Bilimleri Dergisi 2016, 3: 89-94
DOI: 10.7240/mufbed.69674

KNN Algoritması ve R Dili ile Metin Madenciliği Kullanılarak Bilimsel Makale Tasnifi

Deniz KILINÇ[1], Emin BORANDAĞ[1], Fatih YÜCALAR[1], Volkan TUNALI[1], Macit ŞİMŞEK[1], Akın ÖZÇİFT[1]

[1]Celal Bayar Üniversitesi Hasan Ferdi Turgutlu Teknoloji Fakültesi Yazılım Mühendisliği Bölümü, Manisa, Türkiye

ÖZET

Metin tabanlı veri setleri üzerinde analiz işlemi gerçekleştirebilmek için Veri Madenciliğinin alt alanı olan Metin Madenciliği (MM) alanındaki teknik ve yöntemler kullanılmaktadır. Bu çalışmada, akademik yayınlar üzerinde metin madenciliği yöntemleri kullanılarak akademik makalelerin sınıflara ayrılarak tasnif edilme başarısı ölçülmüştür. Bu amaçla bir akademik bilgi paylaşım ağı olan Research Gate üzerindeki belirli akademik yayınların özetleri, geliştirilen yazılım araçları kullanılarak elde edilmiş ve bu özetlerden bir veri seti oluşturulmuştur. Veri seti içerisindeki yayınlar "Materials Science & Engineering" ve "Social Sciences & Humanities" olmak üzere iki ayrı kategoride yer almaktadırlar. Veri seti üzerinde R dili ve R Studio araçlarından yararlanılarak sınıflandırma amacıyla K-En Yakın Komşu (KNN) algoritması kullanılmıştır. Çalışma sonucunda %96,67 oranında doğruluk (ACC) değeri bulunarak yayınların hangi sınıfa ait olduğu tespit edilmiştir.

Anahtar Kelimeler: Metin Madenciliği, R, R Studio, KNN, Metin Sınıflama.

Classification of Scientific Articles Using Text Mining with KNN Algorithm and R Language

ABSTRACT

In order to perform analysis on text-based datasets, the techniques and methods in Text Mining (TM) which is a subdomain of Data Mining are used. In this study, it is aimed to evaluate the classification accuracy of academic articles which are produced in academic domain. In accordance with this purpose, the abstracts of the academic articles are obtained and a dataset is created from an academic knowledge sharing network named Research Gate by using self-developed software tools. The academic articles in the dataset fall into two categories as "Materials Science & Engineering" and "Social Sciences & Humanities". KNN (k-nearest neighbors) classification algorithm is performed by utilizing R language and R Studio tools on the dataset. The experimental results show that the classification accuracy (ACC) of KNN is obtained as 96.67%.

Keywords: Text Mining, R, R Studio, KNN, Text Classification

I. GİRİŞ

Internet teknolojilerinin hızlı gelişimi, internet üzerinden yapılan paylaşımların hızlı şekilde artmasına ve büyük veri setlerinin oluşmasına yol açmıştır [1]. Bu veri setlerinin önemli bir kısmı yapısal olmayan formda, işlenmemiş ve analiz edilmemiş verileri içermektedir. Metinler, fotoğraflar, videolar, ses dosyaları bu verilerden bazılarıdır. Yapısal olmayan verilerin işlenmesi için makine öğrenim yöntemleri geliştirilmiştir. Bu yöntemler biyoenformatik, sistem tanıma, yüksek enerji fiziği, market analizi, görüntü işleme gibi çeşitli alanlarda kullanılmaktadır.

Metin formunda olan veri setleri üzerinde analiz işlemini gerçekleştirebilmek için Veri Madenciliğinin alt alanı olan Metin Madenciliği (MM) alanındaki teknik ve yöntemler kullanılmaktadır. MM genellikle yapısal halde olmayan metin verilerinden ilgi çekici bilgi ve anlam çıkarma işlemi olarak tanımlanır [2]. Günümüzde farklı veri kaynakları üzerinde (haberler, sosyal medya, çevrimiçi kütüphaneler vb.), farklı MM yöntemleri kullanılarak birçok bilimsel çalışma yapılmıştır [2-6].

2010 yılında, Yessenalina ve arkadaşları [8] yaptıkları çalışmada, seçtikleri Movie Review isimli veri setinde

Sorumlu yazar/Corresponding Author: Deniz KILINÇ, Tel: +90 (236) 314 10 10, e-posta: deniz.kilinc@cbu.edu.tr
Gönderilme/Submitted: 02/03/2016, **Düzenleme/Revised:**12/07/2016, **Kabul/Accepted:** 27/07/2016.

bulunan kelimelerin görüş sınıflandırmaya ilişkin ayırt etme özelliği kullanarak tümce etki değerlerini belirlemişlerdir. Veri seti üzerinde destek vektör makineleri yöntemi kullanarak %93.22'lik doğruluk başarımı elde etmişlerdir.

2011 yılında, Bai [9] yaptığı çalışmasında, Markov model sınıflandırıcı kullanan bir algoritma tasarlamıştır. Bu sayede sözcükler arası ilişkiler tanımlanmış ve sınıflandırma yönteminin performansında artış sağlamıştır. Yaptığı labaratuvar çalışmalarında %92.7 lik sınıflandırma başarımı elde etmiştir.

2013 yılında, Li YM, ve arkadaşı [10] Dilin sahip olduğu özelliklere dayalı öznitelik çıkarımı, TF-IDF terim puanlama, destek vektör makineleri yöntemlerini bir arada kullanarak sosyal medya veri seti üzerinde sınıflandırma gerçekleştirmişlerdir. Yaptıkları çalışma sonucunda %90.40'lık sınıflandırma başarımı elde etmişlerdir.

Bu çalışmada ise, akademik alanda üretilen yayınlar üzerinde metin madenciliği yöntemleri kullanılarak akademik makalelerin tasnif edilme başarısı ölçülmeye çalışılmıştır. Bu amaçla bir akademik bilgi paylaşım ağı olan Research Gate [11] üzerindeki belirli akademik yayınların özetleri, geliştirilen yazılım araçları kullanılarak elde edilmiş ve bu özetlerden bir veri seti oluşturulmuştur. Oluşturulan veri setindeki bilimsel yayınlar "Materials Science & Engineering" ve "Social Sciences & Humanities" olmak üzere iki ayrı sınıfta yer almaktadırlar. Veri seti üzerinde R dili ve R Studio [12] araçlarından yararlanılarak sınıf-landırma amacıyla K-En Yakın Komşu (KNN) algoritması kullanılmıştır [13]. Çalışma sonucunda %96.67 oranında doğruluk (ACC) [14,15] değeri bulunarak yayınların hangi sınıfa ait olduğu tespit edilmiştir.

Çalışmanın 2. Bölümünde sınıflandırma yöntemlerinden ve KNN algoritmasından bahsedilmiştir. 3. Bölümde metin madenciliği, veri ön işleme, doküman terim matrisi (DTM), değerlendirme kriterleri ile R dili ve R Studio konuları ele alınmaktadır. 4. Bölümde ise yapılan çalışma, geliştirilen yazılım, veri seti, veri ön işleme, kullanılan algoritma, yapılan çalışma ile deney sonuçlarından elde edilen bilgilere yer verilmiştir. Son bölümde ise sonuç ve öneriler yer almaktadır.

II. K-En Yakın Komşu (KNN) Algoritması

Sınıflandırmada temel amaç, nesnelerin sahip olduğu özelliklere bakılarak nesnelerin hangi sınıfa ait olduğunun belirlenme işlemidir. Çok farklı sınıflandırma türleri ve algoritmaları bulunmaktadır. Karar ağaçları, en yakın komşu, bayes, yapay sinir ağları bunlardan bazılarıdır [16, 17].

KNN algoritması ya da diğer adıyla K-En yakın komşu algoritması makine öğrenim algoritmaları içerisinde en çok bilinen ve kullanılan algoritmalardan biridir. Seçilen bir özelliğin kendine en yakın olan özellikle arasındaki yakınlığı kullanarak sınıflandırma yapılır. Burada bulunan K değeri örnek olarak 3 veya 5 gibi bir sayı ile ifade edilir. Nesneler arasındaki mesafelerin belirlenmesinde Denklem-1'deki formül kullanılmaktadır.

$$d_{(i,j)} = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2} \quad (1)$$

Çalışma şekline baktığımızda, tanımlanan verilere göre yeni bir tanımlanması gereken nesne geldiğinde öncelikle K değerine bakılır. Burada eşitlik olmaması için genellikle K sayısı tek sayı olarak seçilir. Yeni gelen veri ile diğer veriler arasındaki mesafeler hesaplanırken Kosinüs, Öklid ya da Manhattan uzaklığı gibi yöntemler kullanılır [1].

III. Metin Madenciliği

Metin madenciliği, temel olarak yapısal olmayan metinlerden bilgi içeren yapısal metinleri üretme işlemi olarak tanımlanabilir. Metinlerin işlenmesi ile anlamlı bilgilerin elde edilmesi için, veri ön işleme ve özellik çıkartımı gibi isimlerle adlandırılan bazı adımların gerçekleştirilmesi gerekmektedir. Bu aşamalardan sonra yapısal olmayan veriler, metin madenciliğinin kullanılacağı ve bilgisayarlar tarafından işlenen yapısal bir biçime dönüştürülebilir [1, 18]. Bu sayede büyük miktardaki veriler içerisinde bulunan değerli bilgiler keşfedilmiş olur [19]. Üretilen anlamlı bilgiler kullanılarak, kurum ya da kuruluşların faydalanacağı çeşitli sonuçlara ulaşılabilir. Metin madenciliği yöntemlerinin temelinde matematiksel ve istatiksel yöntemler bulunmaktadır. Metin madenciliği, yazar tanıma, metin sınıflama, fikir madenciliği, duygu analizi, anahtar kelime çıkartımı, başlık çıkartımı gibi farklı alanlarda da kullanılmaktadır.

3.1. Veri Ön İşleme ve Doküman Terim Matrisi

Sınıflandırma algoritmalarının çalıştırıla-bilmesi için veri ile ilgili bazı veri temizleme ve düzenleme işlemlerinin gerçekleştirilmesi gerekmektedir. Yapılan işlemler genellikle metin-tabanlı olmayan verileri metin içerisinden çıkartma (noktalama işaretleri, boşluk, özel karakterler vb.), küçük harfe dönüştürme, etkisiz kelimeleri (stop words) ayıklama gibi kısımlardan oluşmaktadır [20]. Etkisiz kelimeler bilgi çıkarımı ve sınıflandırma için bir anlam ifade etmeyen, örneğin; Türkçe dilindeki "ve", "veya", "bazı", "hepsi" gibi

kelimelerden oluşmaktadır. Diğer diller içinde ayrıca etkisiz kelime listeleri oluşturulmaktadır.

Veri ön işleme adımından sonra, doküman matrisi oluşturulması amacıyla bir Bilgi Getirimi (BG) yöntemi olan, her bir dokümanın bir vektörü temsil ettiği ve her bir elemanın ayrı bir terime karşılık geldiği klasik Vektör Uzay Modeli (VUM) kullanılmıştır [23]. Eğer bir terim dokümanın içinde bulunuyorsa, onun vektördeki değeri sıfırdan farklı olmaktadır. VUM'da, terim ağırlıklandırma, modern metin BG sistemlerinin önemli bir özelliğidir. Doküman içerisindeki bir terimin önemini etkileyen başlıca iki kısım bulunur: Terim Frekansı (TF) ve Ters Doküman Frekansı (TDF).

TF temel olarak bir kelimenin metin içerisinde geçme sıklığıdır. TDF değeri ise bu kelimenin bütün metinler içerisindeki geçme sıklığıdır. TF ve TDF değerlerini hesaplamak için kullanılan formüller Denklem-2 ve Denklem-3'te verilmiştir [21-23].

Her bir kelime için TF ve TDF değerleri bulunduktan sonra, Denklem-4'te verilen formül yardımıyla her bir kelimenin ağırlığı hesaplanarak doküman terim matrisi oluşturulur [22, 23].

Her bir kelime için yukarıdaki işlemlerin yapılmasından sonra, Doküman Terim Matrisi (Document Term Matrix – DTM) Şekil 1'de gösterildiği gibi oluşturulmuş olur. T değerleri bir terimi, D değeri ise bir dokümanı simgelemektedir.

$$TF_{(d,m)} = \frac{\text{m kelimesinin d dokümanında geçme sayısı}}{\text{Dokümandaki toplam kelime sayısı}} \quad (2)$$

$$TDF_{(m)} = \ln \frac{\text{Vektor modelindeki toplam dokuman sayısı}}{\text{İçerisinde m kelimesi bulunduran toplam dokuman sayısı}} \quad (3)$$

$$w_{(d,m)} = TF_{(d,m)} * TDF_{(m)} \quad (4)$$

$$\begin{array}{c} \\ D_1 \\ D_2 \\ D_3 \\ \vdots \\ D_m \end{array} \begin{array}{c} \begin{array}{cccccc} t_1 & t_2 & t_3 & \dots & t_m \end{array} \\ \begin{pmatrix} w_{11} & w_{12} & w_{13} & \dots & w_{1m} \\ w_{21} & w_{22} & w_{23} & \dots & w_{2m} \\ & & & & \\ & & & & \\ w_{m1} & w_{m2} & w_{m3} & \dots & w_{mm} \end{pmatrix} \end{array}$$

Şekil 1. Doküman Terim Matrisi.

3.3. Değerlendirme Kriterleri

Sınıflandırma algoritmalarının başarımının değerlendirilmesi için Hata Matrisi (Confusion Matrix) kullanılmaktadır. Tablo-1'de Hata Matrisi görülmektedir. Bu matriste DP "Doğru Pozitif", YP "Yanlış Pozitif", YN "Yanlış Negatif", DN ise "Doğru Negatif" değerini ifade etmektedir [24].

Doğruluk oranı (ACC), sınıflandırıcının sınıf ayırım başarısını ölçmek için geniş çapta kullanılan bir metriktir. Algoritma tarafından doğru olarak sınıflandırılan test örneklerinin yüzdesi olarak tanımlanır. Denklem 5'te verilen formül ile ACC oranı hesaplanır [14, 15].

$$ACC = \frac{(DP + DN)}{(DP + DN + YP + YN)} \quad (5)$$

Tablo 1. Hata Matrisi.

		Tahmin Edilen Sınıf	
		C1 (+)	C2 (-)
Gerçek Sınıf	C1(+) ∑ Pozitif	DP	YN
	C2(-) ∑ Negatif	YP	DN

3.4. R Dili ve R Stüdyo

R dili sayesinde karmaşık matematiksel problemler yazılım kodları ile kolayca kodlanabilir. Karmaşık bir kodlama yapısı olmadığından kolayca öğrenilebilir. Ayrıca matematikçiler ve istatistikçiler oluşturmuş oldukları algoritmaları R dilini kullanarak kolayca yazabilirler. R dili SPSS gibi bir paket program olmayıp bir yazılım geliştirme platformudur [12].

R dilini kullanarak programlar yazmak için R dili ve R Studio aracı ücretsiz olarak kullanılabilir [25- 27]. R dili ve R Studio aracı kendisine özgü bir yazılım ve bir yorumlayıcı ortamı sağlamaktadır.

IV. Yapılan Çalışma

Bu bölümde, gerçekleştirilen yazılımdan, elde edilen veri setinden, kullanılan yöntemden, değerlendirme kriterlerinden ve yapılan çalışma sonuçlarından bahsedilmiştir.

4.1. Veri Seti

Çalışmanın temelini oluşturan veri seti iki ayrı sınıftan oluşmaktadır. Bu sınıflar, "Materials Science & Engineering" ile "Social Sciences & Humanities" sınıflarıdır. Veri seti her bir sınıf için 25 adet farklı derginin her birinden ayrı ayrı seçilen kırkar adet İngilizce dilinde yazılmış makalelerin özetlerini içermektedir. Bu kapsamda veri seti 50 farklı dergiden yukarıda adı geçen sınıflara ait toplam 2000 adet makalenin özetlerinden oluşmaktadır. Veri setinin oluşturulması için standart yazılım geliştirme sürecindeki gerçekleştirim yöntemleri kullanılarak bir araç geliştirilmiştir. Tablo-2'de örnek olarak seçilen dergilerden bazılarının isimleri ve hangi kategoriye ait olduklarına dair bilgiler bulunmaktadır.

Tablo 2. Sınıflandırılmış Dergi İsimlerine Örnekler.

Materials Science & Engineering	Social Sciences & Humanities
Journal of Composite Materials	Critique of Anthropology
Journal of Thermoplastic Composite Materials	Cross-Cultural Research: The Journal of Comparative Social Science
Journal of Wide Bandgap Materials	Cultural Dynamics
Mathematics and Mechanics of Solids	Culture & Psychology
Measurement and Control	Ethnography

4.2. Deneysel Çalışma

Deneysel çalışmaya ait işlem adımları aşağıda sırası ile belirtilmiştir.

- Akademik makalelerden elde edilen veri seti üzerinde çalışmak için, yayın editörleri ve akademisyenler tarafından sınıflandırılması tamamlanmış, SAGE journals tarafından yayımlanan dergilerin isimleri ve kategorilerin bilgisi http://online.sagepub.com/browse/by/discipline isimli siteden 05.10.2015 tarihinde elde edilmiştir.
- İkinci adımda çevrimiçi dergilerden makale özetleri toplamaya yarayan bir araç geliştirilmiştir.
- Üçüncü adımda ise https://www.researchgate.net isimli, akademisyenlerin kendi yapmış oldukları yayınları paylaştıkları web sitesi üzerinden, geliştirilen yazılım aracı kullanılarak belirlenen 50 farklı dergiden 2000 adet makaleye ait özet bilgileri 10.10.2015 tarihinde toplanmıştır. Bu yapılan çalışma ile veri seti oluşturulma işlemi tamamlanmıştır.

Yukarıda adı geçen işlemler tamamlandıktan sonra, R Studio ve R dili kullanılarak veri seti KNN algoritması ile sınıflandırılmıştır.

- R dilinde TM kütüphane paketi içerisinde bulunan removePunctuation kullanılarak veri seti içerisinde bulunan noktalama işaretleri ve boşluk karakterleri kaldırılmıştır. Bütün metin bilgileri tm_map (corpus.tmp, tolower) fonksiyonu kullanılarak küçük harfe dönüştürülmüştür.
- Doküman içerisinde hazır olarak bulunan tm_map (corpus.tmp, removeWords, stopwords ("english")) fonksiyonu kullanılarak etkisiz kelimeler (stopwords) çıkartılmıştır.
- DTM oluşturulmuştur, DTM'de 2000 satır ve 38147 sütun bulunmaktadır.
- Veri setinin %70'lik kısmı ise eğitim, %30'luk kısmı ise test için kullanılmıştır.
- R ve R Studio aracı ile oluşturulan KNN algoritması kaynakçada verilen web adresine ayrıca yüklenmiştir [28].

4.3. Deneysel Sonuçlar

Deneysel çalışma sonucunda KNN algoritmasının doğruluk oranı (ACC) %96,67 olarak elde edilmiştir. Yapılan çalışmanın tekrarlanabilmesi için oluşturulan veri seti referansta verilen web adresine ayrıca yüklenmiştir [28].

KNN algoritmasının doğruluk oranlarını karşılaştırmak amacıyla, metin madenciliği ve metin sınıflandırılması alanında iyi bilinen Naive Bayes (NB) ve J48 algoritmaları kullanılmıştır [29]. NB, bu alanda iyi bilinen istatistik-tabanlı bir gözetimli öğrenme tekniği olup, Bayes teorisini esas almaktadır. Sınıflandırılacak olan dokümanın hangi kategoriye ait olacağı, eğitim setinin tamamı kullanılarak koşullu olasılıkların hesaplanması ile belirlenmektedir. J48, mevcut eğitim seti üzerinden bir karar ağacı oluşturarak girdi dokümanın ait olduğu kategoriyi belirmek üzere sınıflandırma işlemini gerçekleştiren denetimli öğrenme algoritmasıdır. Algoritmaların karşılaştırması Tablo 3'de gösterilmiştir.

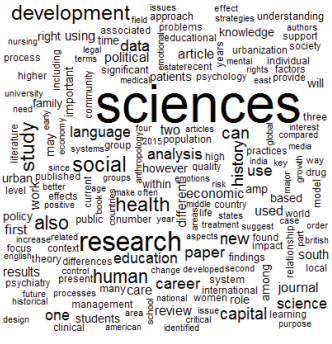
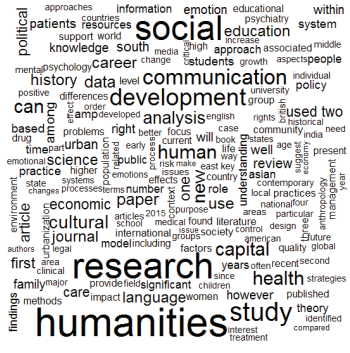
Tablo 3. Algoritma karşılaştırmaları.

Algoritma	Doğruluk Oranı
KNN	%96.67
NB	%91.32
J48	%86.53

Algoritmalar arasında en yüksek doğruluk oranına KNN'nin sahip olduğu görülmektedir. Çalışma sonucunda elde edilen doğruluk oranlarının tüm algoritmalarda genel olarak yüksek olması nedeniyle, sınıflar arasında ayrımı bariz bir şekilde ortaya koyan bazı kelimelerin olduğu düşünülerek, iki farklı sınıfa ait kelime bulutu R dili kodları ile oluşturulmuş ve kaynakçada verilen web adresine ayrıca yüklenmiştir [28]. Oluşturulan kelime bulutları Tablo 4'de gösterilmektedir.

Sınıflara ait kelime bulutlarına baktığımızda "Materials Science & Engineering" sınıfına ait kelime bulutunda en çok çıkan kelimelerin, sınıfın özelliğini de tanımlayan "sciences", "development" ve "research" gibi kelimeler olduğu görülmektedir. Benzer şekilde, "Social Sciences & Humanities" sınıfında ait kelime bulutuna baktığımızda sınıfın özelliklerini tanımlayan "social", "humanities" ve "research" kelimelerinin en fazla sayıda sınıf içerisinde tekrarlandığı görülmektedir.

Tablo 4. Materials Science & Engineering - Social Sciences & Humanities Sınıflarına Ait Kelime Bulutu.

Materials Science & Engineering	Social Sciences & Humanities
	

V. SONUÇLAR

Bu çalışmada akademik makaleleri toplamak için geliştirilen bir araç kullanılarak "Materials Science & Engineering" ve "Social Sciences & Humanities" kategorilerine ait toplam 50 farklı dergiden 2000 adet makalenin özet bilgileri toplanmıştır. Toplanan makale özetlerinden oluşturulan veri seti üzerinde MM çalışması gerçekleştirilmiştir. Elde edilen veri setinin değerlendirilmesi için KNN makine öğrenimi algoritması kullanılmıştır. Oluşturulan veri setinin sınıflandırma bilgisi gizlenerek, KNN sınıflandırması yapılarak doğruluk performans oranları elde edilmiştir. KNN algoritmasının veri seti üzerinde kullanımı için R dili ve R Studio aracı kullanılmıştır. Gerçekleştirilen deneysel çalışma sonucunda %96.67 oranında doğruluk (ACC) değeri bulunarak

yayınların hangi sınıfa ait olduğu tespit edilmiştir. KNN algoritmasının doğruluk oranlarını karşılaştırmak amacıyla, bu alanda iyi bilinen Naive Bayes (NB) ve J48 algoritmaları da kullanılmıştır. Sonuç olarak, MM alanında KNN algoritması oluşturulan veri seti üzerinde kullanıldığında, karşılaştırılan algoritmalara göre daha iyi sonuç alındığı görülmüştür.

Sonraki çalışmalarda geliştirilen yazılım aracı ile daha fazla kategoride makale toplanarak deneyin tekrarlanması düşünülmektedir. Ayrıca kelime frekansı yöntemine alternatif olarak "bilgi kazanımı" gibi özellik seçim yöntemleri kullanılabilir.

KAYNAKÇA

[1] Dolgun, M. Ö., Özdemir, T., Oğuz, D. (2009). "Veri madenciliği'nde yapısal olmayan verinin analizi: Metin ve web madenciliği". İstatistikçiler Dergisi. (2), s.48-58.

[2] Korhonen, A., Séaghdha, D. Ó., Silins, I., Sun, L., Högberg, J., Stenius, U. (2012). "Text mining for literature review and knowledge discovery in cancer risk assessment and research". PLoS One. 7(4) DOI: 10.1371/journal.pone.0033427.

[3] Acun, G., Bilgin, T. T. (2015). "Yazılım hata logları kullanılarak veri madenciliği uygulaması gerçekleştirilmesi". Marmara Fen Bilimleri Dergisi, 27(1).

[4] Ananiadou, S., McNaught, J. (2006). "Text mining for biology and biomedicine". Boston and London: Artech House. 33(1). 135-140.

[5] Zweigenbaum, P., Demner-Fushman, D., Yu, H., Cohen, K.B. (2007). "Frontiers of biomedical text mining: current progress". Briefings in Bioinformatics. 8(5). 358-375.

[6] Cohen, K. B., Yu, H., Bourne, P. E., Hirschman, L. (2008). "Translating biology: Text mining tools that work". Proceedings of the Pacific Symposium on Biocomputing (PSB-08). (13). 551-555.

[7] Onan A., Korukoğlu S. (2016) "Makine öğrenmesi yöntemlerinin görüş madenciliğinde kullanılması üzerine bir literatür araştırması" Pamukkale Univ Muhendislik Bilimleri Dergisi, 22 (2), 111-122

[8] Yassenalina A, Yue Y, Cardie C.(2010) "Multi-Level structured models for document-level entiment classification". Conference on Empirical Methods in Natural Language Processing (EMNLP), Boston, MA, USA, 9-11

[9] Bai X. (2011) "Predicting consumer sentiments from online text". Decision Support Systems, 50(4), 732-742

[10] Li YM, Li TY. (2013) "Deriving Market intelligence from microblogs". Decision Support Systems, 55(1), 206-217

[11] Çevrimiçi: https://www.researchgate.net

[12] Özdemir, A. F., Yıldıztepe, E., Binar, M. (2010). "İstatistiksel Yazılım Geliştirme Ortamı: R", Akademik Bilişim'10, Muğla Üniversitesi.

[13] Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonpara-metric regression". The American Statistician. 46 (3): 175–185.

[14] Ozcift, A., Gulten, A. (2011). "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms". Computer Methods and Programs in Biomedicine. Vol. 104. Issue 3. pp. 443–451.

[15] Faraggi, D., Reiser, B. (2002). "Estimation of the area under the ROC curve". Stat Med. John Wiley & Sons, Ltd. 21(20). pp.3093-3106 (2002).

[16] Karasulu, B. (2015). "Esnek Hesaplama - Melez Zeki Sistemler İçin Bir Rehber". Nobel Akademik Yayıncılık.

[17] Alpaydin, E. (2010). "Introduction to Machine Learning". MIT Press. 2nd Ed. ISBN 978-0-262-01243-0.

[18] Hotho, A., Nurnberger, A., Paaß, G. (2005). "A brief survey of text mining". LDV Forum - GLDV Journal for Computational Linguistics and Language Technology. 20(1). 19-62.

[19] Azzalini, A., Scarpa, B., Walton, G. (2012). "Data Analysis and Data Mining: An Introduction". Oxford University Press. New York.

[20] Fieldman, R., Sanger J. (2006). "The text mining handbook advanced approaches in advanced analyzing unstructured data". Cambridge University Press.

[21] Kılınç, D., Bozyiğit, F., Kut, A., Kaya, M. (2015). "Overview of Source Code Plagiarism in Programming Courses". International Journal of Soft Computing and Engineering (IJSCE). 5(2), 79-85.

[22] Can, F., Kocberber, S., Balcik, E., Kaynak, C., Ocalan H. C., and Vursavas O.M. (2008). "Information retrieval on Turkish texts". Journal of the American Society for Information Science and Technology. 59(3). 407-421.

[23] Salton, G., Wong, A., Yang, C.S. (1975). "A vector space model for information retrieval". Journal of the American Society for Information Science, 18(11). 613-620.

[24] Kılınç, D., Bozyiğit, F., Özçift, A., Yucalar, F., Borandağ, E. (2015). "Metin madenciliği kullanılarak yazılım kullanımına dair bulguların elde edilmesi". Ulusal Yazılım Mühendisliği Sempozyumu. Yaşar Üniversitesi. İzmir.

[25] R Language Download Available: https://www.r-project.org/

[26] R Studio Available: http://www.rstudio.com

[27] Mustafa Gökçe Baydoğan, Berk Orbay, Uzay Çetin, "R ile Programlamaya Giriş ve Uygulamalar" XIX. Türkiye'de İnternet Konferansı 2014 Yaşar Üniversitesi, İzmir.

[28] Çevrimiçi: https://sourceforge.net/projects/acedemic-data-set/files/latest/download

[29] Kılınç, D., Özçift, A., Bozyigit, F., Yıldırım, P., Yücalar, F., & Borandag, E. (2015). TTC-3600: A new benchmark dataset for Turkish text categorization. Journal of Information Science, 0165551515620551.