Eğitim İçerikleri için Sezgisel Metin Bölütlemeye Dayalı Çoklu Etiketleme Stratejisi: M.E.B. Sanat Tarihi Kitabı için Bir Durum Çalışması

Araştırma Makalesi/Research Article

Selcan KAYAHAN[1]*, Korhan GÜNEL[2], Urfat NURİYEV[3,4]

[1]Bilgisayar Bilimleri Anabilim Dalı, Fen Bilimleri Enstitüsü, Ege Üniversitesi, İzmir, Türkiye
[2]Matematik Bölümü, Aydın Adnan Menderes Üniversitesi, Aydın, Türkiye
[3]Matematik Bölümü, Ege Üniversitesi, İzmir, Türkiye
[4]Institute of Control Systems of ANAS, Baku, Azerbaijan
selcan.kayahan@gmail.com, kgunel@adu.edu.tr, urfatnuriyev@ege.edu.tr
(Geliş/Received:19.11.2021; Kabul/Accepted:14.02.2022)
DOI: 10.17671/gazibtd.1026142

Özet— Bu çalışmada, eğitim içeriklerinden otomatik öğretim kavramlarının tespit edilerek, metnin anlamsal bütünlük arz eden ve birbiriyle çakışan metin bloklarına bölütlenmesi ve metin blokları içindeki paragrafların öncelik derecesine bağlı olarak birden fazla öğretim kavramı ile etiketlendirilmesine amaçlanmıştır. Çalışmada T.C. Millî Eğitim Bakanlığı'na bağlı okullarda okutulan Sanat Tarihi kitabı kullanılmıştır. Kitap üzerine doğal dil işleme ve sezgisel kümeleme yaklaşımları uygulanmış ve dokümanın her bir paragrafının hangi öğretim kavramıyla ilişkili olduğunun belirlenmesi hedeflenmiştir. Hedef doğrultusunda, ayrıştırılan metin bloklarını temsil eden öznitelik vektörleri çıkartılmış ve bu öznitelik vektörleri üzerine Temel Bileşen Analizi uygulandıktan sonra Parçacık Sürü Optimizasyonu (Particle Swarm Optimization, PSO) yaklaşımı ile kümeleme işlemi gerçekleştirilmiştir. Bununla birlikte, önerilen sistemin başarım oranlarının belirlenmesi için bölütlendirilmiş metin blokları alan uzmanı tarafından kitap içinde sunulan öğretim kavramları ile eşleştirilmiştir. Ardından uzman görüşleri ve sistem çıktıları karşılaştırılarak ağırlıklandırılmış ortalama karesel hata değeri hesaplanmıştır. Elde edilen sonuç, eğitim içeriklerinin birden fazla öğretim kavramı ile etiketlenmiş metin bloklarına ayrıştırılabileceği konusunda umut vermektedir.

Anahtar Kelimeler— sezgisel metin bölütleme, çoklu etiketleme, parçacık sürü optimizasyonu, eğitim teknolojileri

Multi-Labeling Strategy based on a Heuristic Text Segmentation for Educational Contents: a Case Study for M.E.B. History of Art Book

Abstract— In this study, it is aimed to extract the learning concepts from the educational contents, to segment the context into some overlapped text blocks that have semantic integrity, and to label the paragraphs within the text blocks with multiple learning concepts. The study uses the Art History book taught in schools affiliated to the Republic of Turkey Ministry of National Education. Natural language processing and heuristic clustering techniques are applied on the book and it is aimed to determine which learning concepts are associated with each paragraph of the document. For this purpose, feature vectors representing the parsed text blocks are extracted and the Particle Swarm Optimization clustering technique is applied after applying Principal Component Analysis on the feature vectors. In addition, the segmented text blocks matched with the learning concepts presented by an expert in the book to make a performance analysis of the proposed system. Then, the weighted mean squared error is calculated by comparing expert opinions and system outputs. The obtained results give hope about educational content can be decomposed into text blocks labeled with more than one learning concept.

Keywords— heuristic text segmentation, multi-labeling, particle swarm optimization, educational technology

1. GİRİŞ (INTRODUCTION)

Kavram; nesnelerin veya olayların ortak özelliklerini kapsayan ve bir ortak ad altında toplayan genel tasarımdır [1]. Kavramlar, insanların yaşamlarında gereksinim duydukları temel zihinsel oluşumlar olarak da ifade edilebilir. Kavramlar yolu ile nesneler, olaylar, olgular insan zihninde ayırt edilir ve anlaşılır. Bireyin her bir deneyimi kavramların, dolayısıyla sahnelerin, şemaların ve anlam ağlarının yeniden yapılandırılması anlamına gelmektedir [2]. Okullarda okutulan eğitim dokümanları (ders kitapları) incelendiğinde, her bir ünite için veya kitabın tamamı için bir 'kavramlar listesi' verildiği görülmektedir. Bu listenin verilmesindeki amaç; üniteye veya bölüme çalışmaya başlamadan önce öğrenenin o ünitenin veya bölümün ilgili olduğu kavramları bilmesini sağlamaktır. Böylece öğrenen; öğrenmeye başlayacağı konunun genel olarak 'ne' veya 'neler' ile ilgili olduğunu anlamaktadır. Kısaca öğrenen neleri öğrenmesi gerektiğinin önceden farkında olur. Bu durum ise; öğrenenin zihnindeki sahneleri, şemaları ve anlam ağlarını yeniden yapılandırmasına yardımcı olmakta ve konuyu öğrenmesini kolaylaştırmaktadır. Dolayısıyla ders kitaplarında verilen kavramların tam, doğru ve tüm konuyu kapsayacak şekilde verilmesi önem arz etmektedir.

Bu çalışmada önerilen model ile metin bölütleme yaklaşımı kullanılarak tek eğitim dokümanı metin bloklarına ayrıştırılmış ve bu metin bloklarının birden çok öğrenim kavramı ile otomatik olarak etiketlendirilmesi sağlanmıştır. Bu doğrultuda sezgisel/sezgi üstü optimizasyon teknikleri, makine öğrenmesi ve istatistiksel dil modelleri kullanılarak öncelikle eğitim içerikleri içindeki öğretim kavramlarının minimal bir kümesi tespit edilmiştir. Ardından doküman, birbiri ile çakışan metin bloklarına ayrıştırılmış ve bu metin blokları üzerinden metnin her bir paragrafının öğretim kavramları ile etiketlendirilmesi sağlanmıştır.

Çalışmaya farklı bir bakış açısıyla bakıldığında, önerilen model ile dokümanların arama motorlarınca indekslenmesi mantığından yola çıkılarak, tek bir doküman içinde metin bloklarının indekslenmesine olanak sağlandığı yorumu yapılabilir. Bir üst çerçeveden bakıldığında, eğitim dokümanları için özelleştirilmiş bir arama motoru modeli önerilmiştir. Örneklemek gerekirse; eğer öğrenen matematikte "kapalı fonksiyonun türevi" hakkında bilgi edinmek istiyorsa geliştirilen model öğreneni "türev" konusunu içeren dokümana yönlendirmek yerine, belirtilen kavram ve ilintili olan alt kavramların (örneğin "zincir kuralı") anlatıldığı tüm metin bloklarını listeleyebilecektir.

Dijitalleşen dünyada genel olarak hızla artan veri miktarı eğitim alanında da etkisini göstermiştir. Eğitim bilimlerinde, öğretici ve öğrenenlerden toplanan verinin analiz edilmesi temel adımlardan biridir. Bu analiz sayesinde öğrenenlere rehberlik edilerek öğrenme süreçlerini efektif olarak devam ettirebilmeleri için yol haritaları sunulabilir. Luckin ve Çukurova [3], eğitim bilimleri alanında toplanan büyük veriyi hızla analiz edebilen yapay zeka algoritmaları tasarlamanın, bireyselleştirilmiş eğitim sistemleri için alt yapı ve standardize edilmiş çatılar oluşturmada bizlere yardımcı olacağını vurgulamaktadır. Renz ve Hilbig [4], eğitim teknolojilerinin yapay zeka yeteneklerine sahip olabilmeleri sağlamaları gereken ön koşulları tartışmışlardır.

Ouyang ve Jiao [5], yapay zeka tekniklerinin eğitim teknolojilerinde uygun olarak kullanılabilmesi için üç farklı paradigma önermişlerdir. Önerdikleri düşünce, yapay zekanın bilgi ve bilişsel öğrenme modellerini temsil etmede birer araç olarak kullanılması ve öğrenenlerin ise bu servislerin kullanıcı olması hipotezini savunur. İkinci yaklaşımlarında, öğrenenler yapay zeka ile işbirliği içerisinde sistemin eğitiminde rol alır ve böylece yapay zeka eğitim destek sistemi olarak çalışır. Son düşünce ise, yapay zeka öğrenenler tarafından yönetilen ve öğrenenlerin kendi seviyelerini geliştirmede kullandıkları bir araç olarak tanımlanmıştır. Zhai ve arkadaşları [6] ise eğitim alanında uygulanması gerçekleştirilecek yapay zeka yöntemleri için hiyerarşik bir model sundukları bir literatür incelemesi sunmuşlardır. Bununla birlikte Nabiyev ve Erümit [7], yapay zekanın eğitim ortamlarına teknik ve pedagojik etkilerinin neler olduğunu belirlemeye yönelik çalışmalar henüz yeterince olgunlaşmadığını belirtmektedirler. Bu bağlamda, Akdeniz ve Özdinç [8], makine öğrenmesi tekniklerinin eğitim teknolojilerinde uygulanmasının önemini vurgulayan ve Türkiye adresli çalışmaların incelendiği sistematik bir literatür çalışmasına imza atmışlardır.

Bu çalışmada önerilen makine öğrenmesi ile çoklu etiketlemeli metin bölütleme yaklaşımının eğitim teknolojilerinde nasıl kullanılabileceğinin vurgusu yapılmıştır. Makine öğrenmesinde, doküman sınıflandırma ve etiketlendirme yeni bir araştırma konusu sayılmaz ve alan yazında çok sayıda çalışmaya rastlamak mümkündür. Bununla birlikte, çoklu etiketleme pek çok yönden klasik sınıflandırma yaklaşımlarından ayrışır. Burkhardt ve Kramer [9], çoklu-etiketlemeli sınıflandırma problemlerinin, problem dönüşüm yaklaşımları ve uyarlanabilir algoritma yaklaşımları olmak üzere iki sınıfa ayrılan makine öğrenmesi metotlarıyla çözülebileceğini belirtmişlerdir. Problem dönüşüm yaklaşımları, verinin birden fazla sınıflandırıcı ile etiketlendirilmesi mantığıyla oldukça basit bir temele dayanır. Uyarlanabilir algoritma yaklaşımlarında ise tek bir sınıflandırıcı çevrimiçi öğrenme (online learning) yaklaşımı ile veriye göre algoritmanın çalışması için ihtiyaç duyulan keyfi parametre değerlerini sürekli güncelleyecek şekilde çalışır.

Moyano ve arkadaşları [10], çoklu etiketlemeli sınıflandırıcılar için bir literatür inceleme makalesi sunmuşlardır. Tarekegn ve arkadaşları [11] tarafından gerçekleştirilen benzer çalışmada dengeli olmayan veri setlerinde çoklu etiketleme yaklaşımları için bir literatür çalışması sunulmuştur. Her iki çalışmada da veriyi tek bir etiketle etiketleme ile çoklu etiketleme yaklaşımları arasındaki farklılıklar ve ortaya çıkan zorluklar vurgulanmıştır. Tarekegn ve arkadaşları [11] çoklu etiketlendirme yaklaşımlarını yeniden örnekleme

(resampling), uyarlanabilir sınıflandırma (adaptable classification), topluluk (ensemble) ve maliyete duyarlı (cost-sensitive) yöntemler olarak dört farklı gruba ayırmışlardır. Yöntemlerin performans analizleri için alan yazında sunulan ölçütleri ise örnek (example), etiket (label) ve sıralama (ranking) tabanlı ölçütler olarak üç ana başlıkta incelemişlerdir.

Liu ve arkadaşları [12], mikro blog yazıları üzerinde gerçekleştirdikleri duygusallık analizi çalışmasında çoklu etiketleme yaklaşımı önermişler, ancak 0.344 doğruluk (accuracy) oranı elde etmişlerdir. Bununla birlikte, Tarekegn ve arkadaşlarının [11] literatür inceleme çalışmalarında da bahsi geçen Hamming kaybı (Hamming loss) skoru, altküme doğruluk (subset accuracy) değeri, örnek tabanlı F1 (example based F1) skoru, micro ve makro F1 skorları, ortalama kesinlik (average precision) değeri, kapsama (coverage) ve bir hata (one error) değerlerini performans ölçütü olarak değerlendirmişlerdir.

Kumar ve arkadaşları [13], MLC-HMF adını verdikleri yaklaşımlarında hiyerarşik yerleştirme mantığını temel alarak çoklu etiketlemeli sınıflandırma yaklaşımını önermişler ve bu yaklaşımı farklı alanlarda oluşturulmuş hazır veri setleri ile deneyimlemişlerdir. Çalışmada vurgulanması gereken özellikle eğitim alanında oluşturulan veri seti için elde edilen doğruluk (accuracy) oranının 0.294 ± 0.021 ile sınırlı kalmasıdır.

Lee ve arkadaşları [14], doküman sınıflandırma için ayırt edici özniteliklerin belirlenmesi amacıyla memetik bir arama algoritması geliştirmişlerdir. Bahsi geçen çalışmada yazarlar etiket frekanslarındaki farklılıkları baz alarak dokümanları Sanat, Bilgisayar, Eğitim gibi oldukça genel sayılabilecek on bir farklı kategoriden bir veya birden fazla kategori içinde sınıflandırmışlardır. Önerdikleri yaklaşımı yedi farklı filtre ve sarmal tabanlı öznitelik seçim yaklaşımı (filter and wrapper based feature selection methods) ile karşılaştırmışlardır. Yöntemin performans ölçütü olarak doğruluk (accuracy) değerleri her bir kategori için ayrı ayrı hesaplanmıştır. Çalışmada farklı kategoriler için elde edilen doğruluk oranlarının 0.1322 ± 0.0082 ile 0.5345 ± 0.0082 değerleri arasında değiştiği görülmektedir.

Yang ve Liu [15], çoklu etiketlemeli metin sınıflandırma için paralel kodlama, seri kod çözme yaklaşımına dayanan bir yöntem geliştirmişlerdir. Geliştirdikleri model, aslında evrişimsel yapay sinir ağları ile kaynak metinden yerel komşuluk bilgisi ile global etkileşim bilgisini çıkartan kodlayıcının birleşiminden oluşur. Bu model üç farklı veri seti üzerinde test edilmiş ve her veri seti için sırasıyla 0.893, 0.725 ve 0.825 mikro-F1 skorları elde edilmiştir. Bu değerler çoklu etiketleme uygulayan doküman sınıflandırıcılar için oldukça makul değerler olarak görülmektedir.

Benzer bir çalışma gerçekleştiren Aljedani ve arkadaşları [16], hiyerarşik bir yaklaşımla çoklu etiketleme ile Arabi metinler üzerinde etiketleme çalışması gerçekleştirmişlerdir. Birden fazla sınıflandırıcının kullanıldığı HOMER (Hierarchy Of Multilabel ClassifiER) algoritmasının parametrelerini optimize eden önerileriyle çoklu etiketleme için 0.758 doğruluk oranı ve ortalama 0.853 mikro-F1 skoru elde etmeyi başarmışlardır.

Tüm bu çalışmalarda da kolaylıkla gözlemlenebileceği gibi, çoklu etiketleme yaklaşımlarının başarı oranları klasik tek etiketli sınıflandırma ve kümeleme yaklaşımlarına oranla oldukça düşüktür ve halen yeterli olgunluğa ulaşılamamıştır. Ek olarak, alan yazında çoklu etiketlemeli sınıflandırıcıların eğitim teknolojilerine uygulandığı bir çalışmaya rastlanılmamıştır.

Bununla birlikte doküman sınıflandırma çalışmalarının hemen hemen tümü bir derlem altında toplanmış dokümanlar koleksiyonunun üzerinde gerçekleştirilmiştir. Tek bir doküman üzerinden gerçekleştirilen doğal dil işleme çalışmaları yok denecek kadar azdır. Çalışmada klasik doküman sınıflandırmanın bir adım ötesine geçilerek, tek bir doküman içeriğinden otomatik olarak çıkarılan öğretim kavramları ile doküman içindeki metin bloklarının etiketlendirilmesi ve böylece dokümanın birden fazla kavramla etiketlendirilmesi fikri özgündür. Doküman içindeki metin blokları üzerinde içerik analizi yapmak, dokümanın tümünde içerik analizi yapmaktan görece daha zordur. Bunun temel nedeni metin bloklarının içerdiği sözcük sayılarının çok daha az olmasıdır. Bu sorunun üstesinden gelebilmek için, dokümanın birbiriyle ortak paragraflar içerecek şekilde metin bloklarına ayrıştırılması fikri ortaya çıkmıştır. Literatürde bu tarz yaklaşımla karşılaşılmamıştır. Bu bağlamda, çalışmanın bilimsel olarak doğal dil işleme ve makine öğrenmesi alanlarında literatüre özgün bir değer sunacağı düşünülmektedir.

Doğal Dil İşleme, yapay zekanın bir alt dalıdır ve doğal dilde yazılan metinlerin işlenmesine dayalı yöntem ve teknikleri içerir. Metin bölütleme ise, doğal dil işleme alanının çalışma konularından biridir. Metin bölütleme, doküman içinde birbiriyle ilişkili tutarlı metin bloklarını çıkarma işlemidir [17]. Daha açık ifade etmek gerekirse metin bölütleme, belirli konulara ilişkin bitişik bölümler arasındaki sınırların belirlenmesini, böylece büyük bir metnin içindeki anlamsal bütünlüğün ya da farklılığın ortaya çıkmasını sağlar [18]. Beeferman ve arkadaşları [19], metin bölütleme problemini kötü-tanımlı (ill-defined) problem olarak tanıtmışlar ve metni tutarlı alt segmentlere ayırabilmek için *n*-li sözcük dizilimlerinin (*n*-grams) sıklık değerlerini temel alan öznitelikleri kullanan istatistiksel modelleri incelemişlerdir.

Belgeyi bölütlemenin metin analizi için kullanışlı olmasının birçok nedeni vardır. Temel sebeplerden biri, belgelerin tamamından daha küçük ve daha tutarlı olmalarıdır. Diğer bir neden ise, her bölümün analiz ve erişim birimleri olarak kullanılmasıdır [20]. Anlamsal analizde metin bölütlemeyi uygulayan çalışmalar da vardır. Hoon, Keong ve Kong [21], arama sonuçlarını iyileştirmek için anlambilimin kullanıldığı bir yöntem önermişlerdir. Hoon ve Wei [22], arama algoritmasını geliştirmek için bilginin değerler olarak adlandırılan segmentler halinde düzenlendiği bir teknik kullanmışlardır. Duan ve

arkadaşları [23] ise, görüş madenciliği alanında kullanıcı görüşlerine hangi yönde eğilim gösterdiğinin belirlenmesi ve ilişkili özellikleri belirlemek için metin bölütlemesi uygulamışlardır. Metin bölütleme, belge içeriğine otomatik olarak açıklayıcı notlar ekleme (document annotation) ve metin özetleme (text summarization) uygulamalarının da temelini oluşturur. Doğal dil işlemede dokümana açıklayıcı notlar ekleme genellikle metin içindeki önemli görülen kısımların üst-veri (meta-data) ile etiketlendirilmesi veya renklendirilmesi ile gerçekleştirilir. Bu işlem ağırlıklı olarak, sözcük veya cümle düzeyinde gerçekleştirilir.

Belgeleri küçük alt bölümlere ayırarak bu alt bölümler arası ilişkilerin belirlenmesini sağlayan metin bölütleme yöntemleri ile eğitim materyalleri üzerindeki anlamsal benzerlik ve farklılıkların belirlenmesi sağlanabilir. Nitekim Wentao ve Scheepers [24], metin bölütlemenin bilişsel öğrenme sürecini kolaylaştırıp kolaylaştırmadığını ve ritmik okuma hızına etkisini belirlemek amacıyla, Çince şiirler üzerinde metin bölütleme çalışması yürütmüşlerdir.

Bu çalışma, basitçe eğitim dokümanları birbiriyle çakışan paragraflar halinde metin bloklarına bölütlenerek her bloğun ayrı ayrı etiketlendirilmesi mantığına dayanmaktadır. Böylelikle her bir metin bloğu içindeki paragrafların bir veya birden fazla öğretim kavramıyla etiketlendirilmesi sağlanmıştır.

Eğitim dokümanları, ders kitapları, kaynak kitaplar gibi özellikle k-12 düzeyinde kullanım alanı oldukça fazla olan yazılı materyallerde öğretim kavramları bulunmaktadır. Öğretim kavramlarını belirli bir öğrenme alanı içine çıkarmak çoğu zaman, bu alanda uzman bir kişi için bile zor, tartışmalı, zaman alıcı ve önemsiz görülen bir süreçtir [25]. Dokümanların veri işleme yöntemleri ile incelenmesi, eğitimde kullanılan kaynakların anlamsal analizinin yapılmasına, dokümanların 'ne' ifade ettiğinin tespit edilmesine ve öğretim kavramları ile kaynağın geneli arasındaki ilişkilerin belirlenmesine yardımcı olmaktadır. Metin bölütleme, belgeleri küçük alt bölümlere ayırarak bu alt bölümler arası ilişkilerin görülmesini sağlayan bir yöntemdir. Metin bölütleme yöntemleri ile eğitim materyalleri üzerindeki anlamsal benzerlik ve farklılıkların belirlenmesi sağlanabilir.

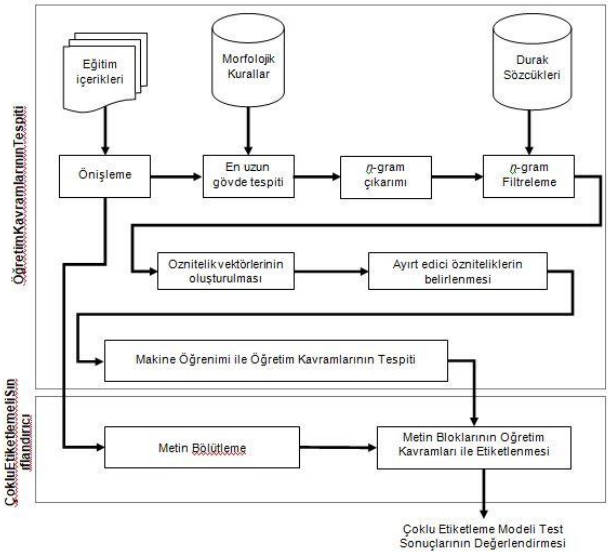
Bu çalışmada; Millî Eğitim Bakanlığı tarafından hazırlanan ve açık erişimli olarak paylaşılan kitaplar analiz edilerek, öğrenenler için öğrenme alanından bağımsız olarak eğitim dokümanındaki metin bloklarının öğretim kavramları ile etiketlendirilmesi amaçlanmıştır. Bu amaç doğrultusunda doğal dil işleme, makine öğrenmesi ve sezgisel veya sezgisel olmayan en iyileme yaklaşımları kullanılarak içeriklerin analizi yapılmıştır.

2. YÖNTEM (METHODOLOGY)

Bir belgedeki bilgiler temel olarak sözcüklerin anlambiliminden oluşur [26]. Bilgi geri getirim alanında sıklıkla kullanılan vektör tabanlı yöntemlerde, metinler içerdikleri sözcükler ve/veya metne ait farklı özelliklerden oluşan vektörler ile temsil edilirler [27, 28].

Çoğu birey için ilgi duyduğu bir konuda çalışmaya nereden başlayacağını belirlemek oldukça güçtür ve öğrenme hedeflerine ulaşabilmek için bir uzman gözetiminde genel olarak oluşturulmuş bir yol haritasına ihtiyaç duyulmaktadır. Ancak bireysel öğrenme sürecinde konu özelleştikçe bireyin kişisel bilgi düzeyine dayanarak bu yol haritasının oluşturulması ve bu haritayı oluşturacak uzman nitelikte kişinin bulunması da giderek zorlaşır. Üstelik her uzman kendi deneyimine göre farklı bir yol haritası oluşturabilir. Zubrinic ve arkadaşları [29], bilgi erişim ve yönetim sistemlerinin bir uzman gibi davranarak öğrenme planının yani yol haritalarının oluşturulmasında yardımcı olabileceğini öne sürmüşlerdir. Birey, bu yol haritasını kullanarak önermeleri gerçeklerse, yani kavramları ve aralarındaki ilişkileri öğrenirse, öğrenme hedeflerine de ulaşmış olur.

Çalışmada; yapılandırılmamış ham veri halindeki metin içerikli öğretim materyallerinden öğretim kavramlarını tespit eden, ardından metni bölütleyerek öğretim kavramları ile etiketlendirilmesi amaçlanmıştır. Eğitim dokümanlarının bölütlenerek öğretim kavramları ile etiketlendirilmesi için geliştirilmesi önerilen modelin adımları Şekil 1'de gösterilmiştir.

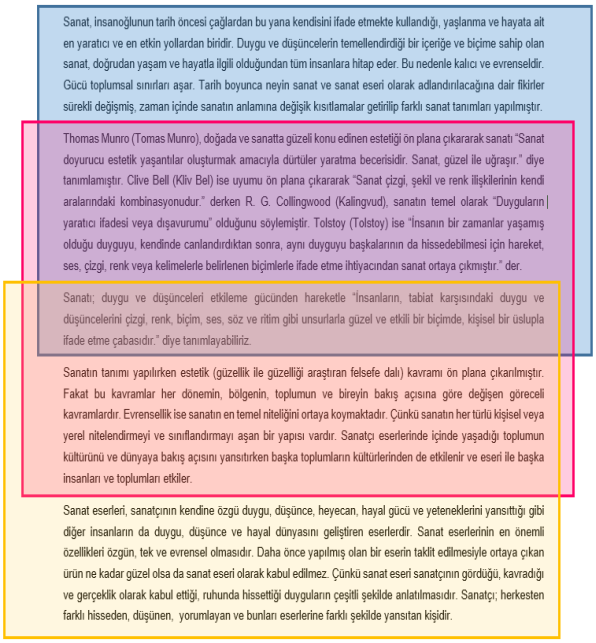


Şekil 1. Çoklu etiketlemeli metin bölütleme ile eğitim dokümanlarının öğretim kavramları ile etiketlenmesi
(Labeling of educational documents with instructional concepts with multi-labeled text segmentation)

Önişlem: Çalışmada, tüm eğitim içerikleri birer sözcük dizileri olarak ele alınmaktadır. Ön işlem aşamasında, dokümanlardaki tüm sözcükler öncelikle küçük harflere dönüştürülmüş, ardından, birden fazla boşluk ve satır atlama karakterleri dokümanlardan çıkarılmıştır. Dokümanlar, tüm matematiksel formüller, sayılar, değişkenler ve sembollerden de arındırılmıştır. Ardından #, %, & ve $ gibi özel karakterler ve noktalama işaretleri dokümanlardan temizlenmiştir. Ön işleme sonunda eğitim içerikleri ardışık sözcükler arasında sadece birer boşluk kalacak şekilde bir sözcük dizilimine dönüşmüş olur.

En uzun gövde tespiti: Bilgi geri getirim ve doğal dil işlemenin alanlarının temel problemlerinden olan gövdeleme, biçim-birimsel yapısı çok karışık olmayan İngilizce gibi dillerde nazaran daha kolay yapılabilmekte iken Türkçe gibi eklemeli dillerde ekin cümle içerisinde anlamsal ve yapısal görev değişikliklerinin görülmesi nedeniyle oldukça zordur. Çalışmada, en uzun gövde tespiti için hece tabanlı gövdeleme yaklaşımı kullanılmıştır. Bu yaklaşımda, Türkçe heceleme algoritması ilk olarak sesli harflerin konumunu belirlemekte, sesli harf yoksa ilk sesli harfin bulunduğu indisten itibaren sözcüğü hecelerine ayırmaktadır. Heceleme işleminin ardından ismin halleri (bulunma, ayrılma, belirtme halleri), ekleri ile sözcüğünün yazımı ve sert sessiz yumuşamasını göz önüne alarak ek almış bir sözcüğün en uzun gövdesi bulunmaya çalışılmıştır. Sözcükten atılan her ekten sonra ikili arama algoritması kullanılarak geri kalan kısmın sözlükte olup olmadığı kontrol edilmiştir. Eğer arta kalan kısım sözlükte yer alıyorsa kelimenin en uzun gövdesi olarak kabul edilmiştir.

Sonraki adımda, önişlemden geçirilen dokümanların bölütlenebilmesi için sezgisel bir yaklaşım önermekteyiz. Yaklaşımın ilk adımında Şekil 2'de görüldüğü üzere metin birbiriyle ortak paragraf içeren alt metin bloklarına ayrıştırılır. Şekil 2'de T.C. Millî Eğitim Bakanlığı'na bağlı okullarda okutulan Sanat Tarihi kitabından alıntılanma yapılarak dokümandan birbiriyle çakışan alt metin bloklarının çıkarıldığı görsel olarak gösterilmiştir. Birbiriyle çakışan metin blokları oluşturma fikri metin blokları arasındaki ilişkilerin ortaya konması ve çoklu etiketlemede kolaylık sağlaması açısından özgün bir fikirdir.


Sanat, insanoğlunun tarih öncesi çağlardan bu yana kendisini ifade etmekte kullandığı, yaşlanma ve hayata ait en yaratıcı ve en etkin yollardan biridir. Duygu ve düşüncelerin temellendirdiği bir içeriğe ve biçime sahip olan sanat, doğrudan yaşam ve hayatla ilgili olduğundan tüm insanlara hitap eder. Bu nedenle kalıcı ve evrenseldir. Gücü toplumsal sınırları aşar. Tarih boyunca neyin sanat ve sanat eseri olarak adlandırılacağına dair fikirler sürekli değişmiş, zaman içinde sanatın anlamına değişik kısıtlamalar getirilip farklı sanat tanımları yapılmıştır.

Thomas Munro (Tomas Munro), doğada ve sanatta güzeli konu edinen estetiği ön plana çıkararak sanatı "Sanat doyurucu estetik yaşantılar oluşturmak amacıyla dürtüler yaratma becerisidir. Sanat, güzel ile uğraşır." diye tanımlamıştır. Clive Bell (Kliv Bel) ise uyumu ön plana çıkararak "Sanat çizgi, şekil ve renk ilişkilerinin kendi aralarındaki kombinasyonudur." derken R. G. Collingwood (Kalingvud), sanatın temel olarak "Duyguların yaratıcı ifadesi veya dışavurumu" olduğunu söylemiştir. Tolstoy (Tolstoy) ise "İnsanın bir zamanlar yaşamış olduğu duyguyu, kendinde canlandırdıktan sonra, aynı duyguyu başkalarının da hissedebilmesi için hareket, ses, çizgi, renk veya kelimelerle belirlenen biçimlerle ifade etme ihtiyacından sanat ortaya çıkmıştır." der.

Sanatı; duygu ve düşünceleri etkileme gücünden hareketle "İnsanların, tabiat karşısındaki duygu ve düşüncelerini çizgi, renk, biçim, ses, söz ve ritim gibi unsurlarla güzel ve etkili bir biçimde, kişisel bir üslupla ifade etme çabasıdır." diye tanımlayabiliriz.

Sanatın tanımı yapılırken estetik (güzellik ile güzelliği araştıran felsefe dalı) kavramı ön plana çıkarılmıştır. Fakat bu kavramlar her dönemin, bölgenin, toplumun ve bireyin bakış açısına göre değişen göreceli kavramlardır. Evrensellik ise sanatın en temel niteliğini ortaya koymaktadır. Çünkü sanatın her türlü kişisel veya yerel nitelendirmeyi ve sınıflandırmayı aşan bir yapısı vardır. Sanatçı eserlerinde içinde yaşadığı toplumun kültürünü ve dünyaya bakış açısını yansıtırken başka toplumların kültürlerinden de etkilenir ve eseri ile başka insanları ve toplumları etkiler.

Sanat eserleri, sanatçının kendine özgü duygu, düşünce, heyecan, hayal gücü ve yeteneklerini yansıttığı gibi diğer insanların da duygu, düşünce ve hayal dünyasını geliştiren eserlerdir. Sanat eserlerinin en önemli özellikleri özgün, tek ve evrensel olmasıdır. Daha önce yapılmış olan bir eserin taklit edilmesiyle ortaya çıkan ürün ne kadar güzel olsa da sanat eseri olarak kabul edilmez. Çünkü sanat eseri sanatçının gördüğü, kavradığı ve gerçeklik olarak kabul ettiği, ruhunda hissettiği duyguların çeşitli şekilde anlatılmasıdır. Sanatçı; herkesten farklı hisseden, düşünen, yorumlayan ve bunları eserlerine farklı şekilde yansıtan kişidir.

Şekil 2.Ortak paragraf içeren metin bloklarının ayrıştırılması (Text blocks with common paragraphs)

Tek bir doküman üzerinde gerçekleştirilen bu işlemden sonra oluşan her bir metin bloğu ayrı bir metin dosyası gibi düşünülerek bu metin blokları için öznitelik çıkarma işlemi uygulanır. Metin bloklarının dokümanla karşılaştırıldığında çok daha az sayıda sözcük içerdiği aşikârdır. Bu nedenle, metin bloklarını karakterize eden öznitelik vektörleri çıkartılırken monogram ve bigramlar (1-li ve 2-li sözcük dizilimleri) kullanılarak sözcük-torbası modeli (bags-of-words) modeli oluşturulur. Bu noktada ortaya çıkan araştırma problemlerinden biri sözcük-torbası modelinin boyutunun belirlenmesidir. Her bir metin bloğunun birbirinden farklı sayıda monogram içerdiği düşünüldüğünde sözcük-torbası boyutunun en fazla, en az sayıda monogram içeren metin bloğunun sözcük sayısı kadar olabileceği açıktır. Bununla birlikte sözcük torbasının boyutu sınıflandırma veya kümeleme için belirleyici olabilir. Metin bloklarına ait öznitelik vektörleri çıkarıldıktan sonra, Temel Bileşen Analizi (Principal Component Analysis) kullanılarak boyut indirgeme işlemi uygulanmıştır.

Bu aşamanın adımları aşağıdaki şekilde devam etmektedir:

n-gram çıkarımı: Eğitim içerikleri ön işlemden geçirildikten sonra, bu içeriklerden $1 \leq n \leq 6$ için n-li terimlerin çıkartılması sağlanmıştır. Burada n-li terimden kasıt ardışık n adet sözcük dizilimidir. Her bir n-li terimin özniteliklerinin oluşturulabilmesi için öncelikle n-li terimlerin metin içindeki görüntülenme sıklıklarının çıkartılması gerekir. Bu işlemin hızlı biçimde yapılabilmesi için, n adet ardışık sözcüğün metin içindeki başlangıç ve bitiş konumları ayrı iki dizide saklanır. Ardından n-li terimler alfabetik olarak hızlı sıralama algoritması (quicksort) ile sıralanırken eş zamanlı olarak terimlerin başlangıç ve bitiş konumlarını gösteren dizilerde sıralanır. Hızlı sıralama algoritmasının zaman karmaşıklığı m sıralanacak terim sayısını belirtmek üzere en iyi durumda $O(m.\log(m))$ ve en kötü durumda ise $O(m^2)$ olur. Böylelikle aynı terimler arka arkaya sıralanır ve sadece başlangıç ve bitiş konumları arasındaki farkları kullanarak ardışık iki terimin aynı olup olmadığı kanaatine kolaylıkla varılır. Eğer terimlerin uzunlukları aynı ise ardışık iki terimin aynı olup olmadığına Boyer-Moore metin eşleme algoritması kullanılarak (sağdan sola doğru karakter karakter tarama yapılarak) karar verilir. Karşılaştırılacak metinler aynı uzunlukta olduğu için eğer m sözcüğün içerdiği karakter sayısı ise bu işlemin zaman karmaşıklığı en iyi durumda $O(1)$ ve en kötü durumda $O(m)$ olacaktır. Eğer ardışık iki sözcük aynı ise frekansları 1 arttırılır ve bir sonraki ardışık iki n-li terim karşılaştırılır. Eğer sözcükler aynı değilse ilk n-li terim sözcük havuzuna frekansı ile birlikte eklenir. n-gram modeline göre tüm dokümandaki monogram ve bigramlar çıkarılmıştır. N-gram modeli, olasılıksal bir dil modelidir [30] ve aşağıdaki eşitliğe göre çalışır [25]:

$$P\left(w_i \mid w_{i-(n-1)} \dots w_{i-1}\right) = \frac{C\left(w_{i-(n-1)} \dots w_{i-1} w_i\right)}{C\left(w_{i-(n-1)} \dots w_{i-1}\right)}$$

n-gram Filtreleme: n-li terimlerin birer aday öğretim kavramı olarak çıkarımı ile ilgili önemli sorunlardan biri, n-li sözcük diziliminin ilk ya da son sözcüğünün durak sözcüğü olup olmamasıdır. Eğer öyle ise o zaman bu n-

gram göz ardı edilmelidir. Çünkü eğitim içeriklerinde yer alan öğrenim kavramları bir durak sözcükle başlayamaz ya da bitemez. Örneğin, "sürü zekâsı" ve "parçacığın hızı" ifadeleri "Global Optimizasyon" alanında birer öğretim kavramı iken "ile sürü zekası" ve "parçacığın hızı ve" sözcük dizilimleri öğretim kavramı değildirler.

Durak sözcükler doküman içinde tek başlarına anlam ifade etmeyen ve metnin tümü dikkate alındığında düşük ayırt edici değere (low discrimination value) sahip olan sözcüklerdir. Bununla birlikte doküman içerisinde yüksek frekansla görüntülenmelerine rağmen, cümle içinde sadece söz-dizimsel olarak görevleri mevcuttur, anlam-bilimsel olarak bilgi çıkarım sistemlerine fazla katkıları bulunmaz. Bu nedenlerle çoğunlukla göz ardı edilirler. Durak sözcüklerin çıkarımı için literatürde farklı yaklaşımlar bulunmaktadır. Klasik yöntemlerde sözcükler durak sözcüğü havuzundaki sözcüklerle karşılaştırılırlar. Zipf yasasına dayalı yaklaşımlarda doküman içerisinde yüksek frekansa sahip tek bir sözcükten oluşan ifadeler eğer ters doküman frekansı da düşükse birer durak sözcük olarak görülüp silinirler. Karşılıklı Bilgi (Mutual Information) yaklaşımlarında ise problem bir sınıflandırma problemi olarak ele alınıp düşük karşılıklı bilgi değerine sahip tekli sözcükler durak sözcük olarak düşünülür. Terim tabanlı rassal örnekleme yaklaşımında ise rassal olarak seçilen metin öbekleri üzerinde Kullback-Leibler ayrışma ölçüsü kullanılarak durak sözcükleri tespit etmeye çalışılır. Literatürde temelde bu dört yaklaşımı baz alan farklı teknikler mevcuttur. Çalışmada bu yöntemlerden ikisinin hibritlenerek durak sözcüklerin belirlenmesi sağlanmıştır. Kumova ve Karaoğlan [31], ilk veya son sözcüğü durak sözcük olan n-li terimler elimine edildikten sonra yazım denetimi aşamasına geçilmiştir. Bu aşamada Aşlıyan, Günel ve Yakhno [32] tarafından önerilen istatistiksel yaklaşım kullanılmıştır. Belirtilen yaklaşım, Türkçe harflerin birbiri ardına gelme olasılıkları üzerinden hatalı yazılmış sözcüklerin tespitine dayanır. Aşlıyan, Günel ve Yakhno [32], 3-lü harf dizilimlerinin birbiri ardına gelme olasılıkları kullanılarak yaklaşık %97 doğruluk oranıyla hatalı yazılmış sözcüklerin tespiti gerçekleştirilmiştir. Bu çalışmada ise yazım denetimini geçemeyen *n*-li sözcük dizilimleri de aday öğretim kavramı olmaktan çıkarılmıştır.

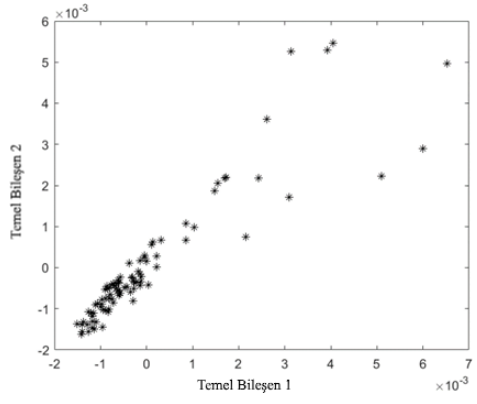
Öznitelik vektörlerinin oluşturulması: Çalışmada, öğretim kavramı çıkarımı problemi için doğal dil işleme uygulamalarında kullanılan özniteliklerden bazıları uygun bir şekilde değiştirilmiştir. Kullanılan öznitelikler kısaca aşağıdaki gibi özetlenmiştir:

- n: aday öğretim kavramındaki sözcük sayısı, $t_j = w_1 w_2 \dots w_n$
- Frekans (Frequency): aday öğretim kavramı olan t_j'nin eğitim içeriğinde geçme sayısı, $f(t_j)$
- Terim Frekansı (Term Frequency): t_j sözcük dizileminin n-gram frekansının eğitim içeriğine ait tüm n-gramlarının toplam frekansına oranı, $tf(t_j)$
- Metin Bloğu Frekansı (Text Block Frequency): t_j sözcük dizilimini içeren metin bloklarının sayısı, $df(t_j)$. Bu çalışmada, $df(t_j)$ hesaplanmıştır.
- *Ters Metin Bloğu Frekansı (Inverse Text Block Frequency):* Bu çalışmada doküman sınıflama ve kümelere yaklaşımlarında vazgeçilmez özniteliklerden biri olan ters doküman frekansı değeri, tek bir doküman üzerinde çalışıldığından, Ters Metin Bloğu Frekansı olarak yeniden yorumlanmıştır. Bu değer, N, d dokümanı içindeki tüm metin blokların sayısı olmak üzere $itbf(t_j) = \log\left(\frac{N}{df(t_j)}\right)$ eşitliği ile hesaplanır.
- *Metin Bloğu Frekansı – Ters Metin Bloğu Frekansı (Text Block Frequency – Inverse Text Block Frequency):* Bir önceki öznitelikte belirtilen benzer yaklaşım terim frekansı – ters doküman frekansı özniteliği için kullanılarak metin bloğu frekansı – ters metin metin bloğu frekansı değerleri elde edilmiştir. Bu değer her bir t_j sözcük dizilimi için $tbf - itbf(t_j) = tbf(t_j) \times itbf(t_j)$ olarak hesaplanır.

Normalizasyon: Çalışmada yukarıda belirtilen öznitelik değerleri ile tanımlanan her bir aday öğretim kavramı oluşan öznitelik vektörlerine min-max normalizasyonu kullanılmıştır. Böylelikle öznitelik değerlerinin [0,1] aralığında değer alması sağlanmıştır. Bununla birlikte her bir metin bloğunun içerdiği aday öğretim kavramı sayılarının minimumu hesaplanmış ve hesaplanan sayıda sözcük diziliminin öznitelik vektörleri kullanılarak kelime-torbası (bags-of-words) yaklaşımı ile metin blokların öznitelik vektörleri oluşturulmuştur.

Boyut İndirgeme: Paragrafların normalize edilen öznitelik vektörleri üzerinde temel bileşen analizi (Principal Component Analysis, PCA) ile boyut küçültme uygulanmıştır. Büyük çok değişkenli veri kümeleri analiz edildiğinde genellikle boyutlarının azaltılması istenir ve temel bileşen analizi, bunu yapmak için bir tekniktir [33].

Çalışmada, kullanılan dokümanda bölütlenen 92 metin bloğuna ait öznitelik vektörleri üzerine PCA boyut indirgenme yaklaşımı uygulandığında, temel bileşenlerin belirlenmesi ve metin bloklarının düzlemde konumlandırılması sonucu Şekil 3'de gösterilmiştir.



Şekil 3.Sanat Tarihi ders kitabı için PCA işleminden sonra metin bloklarının düzlemde konumlandırılması
(Positioning of text blocks after PCA)

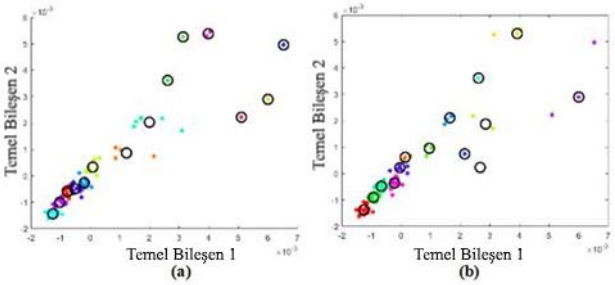
Şekil 3'de görüleceği üzere metin blokları hemen hemen doğrusal bir dağılımla arama uzayında konumlanmıştır. Bununla birlikte verinin büyük bir kısmının birbirine oldukça yakın olduğu düşünüldüğünde verinin kümelere ayrıştırılmasının kolay olmadığı gözlenmektedir. Çalışmanın bir sonraki aşamasında, alandan (doküman içeriğinden) bağımsız etiketlenme yapılması hedeflendiğinden denetimsiz kümeleme yaklaşımı yapılması uygun bulunmuştur.

Kümeleme: Bu çalışmada sınıflandırma yaklaşımı yerine, etiketleme sürecini gözetimsiz olarak otomatikleştirmek için kümeleme yaklaşımı kullanılması tercih edilmiştir. Kümeleme, bir denetimsiz öğrenme yöntemidir. Çalışmada kümeleme yaklaşımlarının tercih edilme sebebi Yann Le Cun'ın denetimli ve denetimsiz öğrenmeyi açıklayan ve çok bilinen bir cümlesiyle açıklanabilir: "Zekâ bir pasta olsaydı, denetimsiz öğrenme pastanın keki olurdu, denetimli öğrenme pastanın kreması olurdu ve pekiştirmeli öğrenme pastadaki kiraz olurdu [34].

Bu çalışmada, normalize edilen veriler kullanılarak, sürü zekâsı yaklaşımına dayanan tekniklerden biri olan Parçacık Sürü Optimizasyonu (Particle Swarm Optimization, PSO) yöntemi ile sezgisel olarak ayrıştırılan metin blokları üzerinde kümeleme yapılmıştır. PSO çok yönlü ve popülasyon tabanlı stokastik bir optimizasyon tekniğidir [35]. PSO yaklaşımında, popülasyonu oluşturan tüm parçacıklar başlangıçta rassal olarak arama uzayında konumlandırılır. Sürü içinde parçacıklar kendi bilişsel bilgileri ile sürüden elde ettikleri bilgiyi bir arada kullanarak hızlı bir şekilde en iyi konuma sahip parçacığın etrafında toplanır.

Bu çalışmada parçacıkların konum vektörleri, küme merkezlerinin koordinatlarını belirtmektedir. :alışmada kullanılan doküman 92 farklı metin bloğuna ayrıştırıldıktan sonra her metin bloğu bir öznitelik vektörü ile betimlenmiştir. Metnin 14 farklı öğretim kavramı ile etiketlendirileceği düşünülerek, 92 metin bloğunun 14 küme altında toplanması hedeflenmiştir. Bu amaçla PSO'un ilk adımında, öznitelik vektörleri bileşenlerinin minimum ve maksimum değerleri aralığı ile sınırlandırılan arama uzayında rassal olarak dağılım gösterecek şekilde bir parçacık sürüsü oluşturulmuştur. Sürüdeki her bir parçacığının konum vektörü, 14 aday küme merkezinin koordinatlarını içerecek şekilde tanımlanmıştır. Parçacık sürüsü boyutları 14×2 olan toplam 50 birey içermektedir. Algoritmanın her adımında metin bloklarına ait öznitelik vektörlerinin merkezlere olan uzaklıkları hesaplanır. Dolayısıyla 92 paragraf ve 14 küme için 92×14 tipinde bir uzaklık matrisi oluşturulur. Ardından her metin bloğunun en yakın olduğu küme merkezi tespit edilerek metin bloğu bu kümeye dâhil edilir. Her parçacık için bu işlem gerçekleştirildikten sonra tüm metin bloklarının öznitelik vektörlerinin dâhil oldukları küme merkezlerine olan toplam uzaklıkları ile parçacığın maliyeti belirlenir. Parçacıkların maliyetleri belirlendikten sonra sürü içinde en düşük maliyete sahip olan parçacık belirlenir. Bu aşamadan sonra klasik PSO yaklaşımı ile her iterasyonda parçacıkların konumları güncellenir. Maksimum iterasyon sayısına ulaşıldığında tüm parçacıkların sürü içinde en iyi konuma, yani küme merkezlerinin en uygun konumuna, sahip parçacık civarında toplandığı gözlenebilir. Sonuç olarak bu konum küme merkezlerini belirlemede kullanılır ve metin blokları bu parçacığın konum vektörü kullanılarak otomatik olarak etiketlendirilir.

Çalışmada *k*-means ve PSO olmak üzere iki farklı kümeleme yaklaşımı test edilmiştir. Çalışmada Şekil 3'de gösterilen verinin *k*-means ve PSO yaklaşımı ile kümelendirilmesiyle gerçekleştirilen etiketlendirme sırasıyla Şekil 4a ve Şekil 4b ile gösterilmiştir. Şekil 4'de her renk ayrı bir öğretim kavramını (kümeyi) belirtirken, oval olan şekiller ise küme merkezlerini ifade etmektedir.



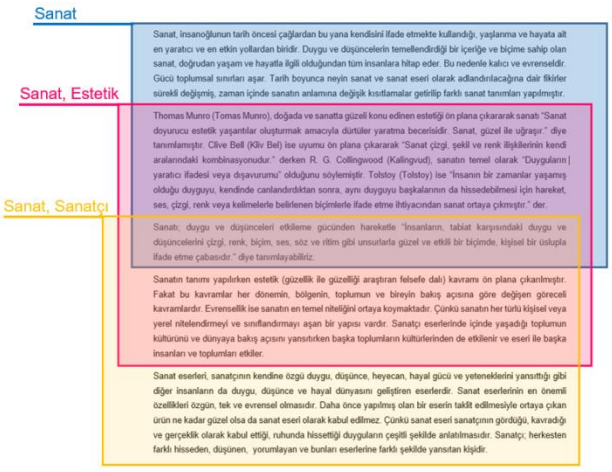
Şekil 4.Sanat Tarihi ders kitabı için metin bloklarının **(a)** *k*-means **(b)** PSO ile kümelenmesi (Clustering of text blocks (a) k-means (b) with PSO for Art History book)

Tablo 1. Doküman hakkında bazı istatistiği bilgiler
(Some statistical information about the document)

Nitelik	Değer
Toplam sözcük sayısı	6213
Durak sözcüklerin sayısı	1196
Paragraf sayısı	91
Monogram sayısı	5017
Tekil monogram sayısı	1546
Bigram sayısı	5351
Tekil bigram sayısı	4843
Metin bloğu başına ortalama monogram sayısı	369,856
Metin bloğu başına ortalama tekil monogram sayısı	138,747
Paragraf başına ortalama monogram sayısı	68,659
Paragraf başına ortalama tekil monogram sayısı	55,131
Öğretim kavramı sayısı	14
Öğretim kavramları sayısının tekil monogramlara oranı	0,009
Uzman tarafından etiketlenen ve paragraf başına ortalama öğretim kavramı sayısı	1,626
Uzman tarafından tek öğretim kavramıyla etiketlendirilen metin bloğu sayısı	49
Uzman tarafından çift öğretim kavramıyla etiketlendirilen metin bloğu sayısı	27
Uzman tarafından üç öğretim kavramıyla etiketlendirilen metin bloğu sayısı	15
Sistem tarafından etiketlenen ve paragraf başına ortalama öğretim kavramı sayısı	2,286
Uzman tarafından tek öğretim kavramıyla etiketlendirilen metin bloğu sayısı	12
Uzman tarafından çift öğretim kavramıyla etiketlendirilen metin bloğu sayısı	41
Uzman tarafından üç öğretim kavramıyla etiketlendirilen metin bloğu sayısı	38

3. DENEYSEL SONUÇLAR (EXPERIMENTAL RESULTS)

Çalışmada T.C. Millî Eğitim Bakanlığı'na bağlı okullarda okutulan Sanat Tarihi kitabı kullanılmıştır. "Estetik", "Heykel", "Mimari", "Resim", "Sanat", "Sanatçı", "Zanaatkâr", "Tarih", "Tunç", "Tablet", "Karum", "Tümülüs", "Megaron" ve "Sikke" kavramları Sanat Tarihi kitabında öğrenilmesi gereken on dört öğretim kavramı olarak tanıtılmıştır. Kitap içeriği ile ilgili bazı istatistiği veriler Tablo 1'de sunulmuştur. Kitap içeriği ön işlemeden geçirildikten sonra, çakışan 91 metin bloğuna ayrıştırılmış ve bu metin bloklarının alan uzmanı tarafından bahsi geçen kavramlarla etiketlendirilmesi istenmiştir. Tablo 1'de de görüleceği üzere, manüel etiketleme sonrası, belirtilen 14 kavramı kullanan uzmanın kitabın her bir metin bloğunu ortalama 1,626 öğretim kavramı ile etiketlendirdiği gözlenmiştir. Uzman 91 metin bloğundan sadece 15'ini üç farklı öğretim kavramıyla etiketlenmiştir. 27 metin bloğu ise 2 farklı öğretim kavramı ile etiketlenmiştir. Geri kalan 49 metin bloğunun ise sadece 1 öğretim kavramı ile etiketlendiği görülmüştür. Uzman görüşleri alındıktan sonra; metin blokları için sırasıyla *n*-gram çıkarımı, öznitelik vektörlerinin oluşturulması, Temel Bileşen Analizi ile boyut indirgeme işlemleri yapıldıktan sonra Parçacık Sürü Optimizasyonu (PSO) yaklaşımı ile kümeleme uygulanmıştır. Bu yaklaşımın tercih edilmesinin temel nedeni, alandan bağımsız olarak metin bloklarının otomatik olarak etiketlenmesinin hedeflenmesidir. Parçacık Sürü Optimizasyonu ile denetimsiz öğrenme gerçekleştirilerek kümeleme yapılır. Kümeleme sonucu metin blokları Şekil 5'de görülebileceği gibi etiketlenmiştir. Şekil 5'de görüleceği üzere çalışmada her bir paragraf en az bir en çok üç kavramla etiketlenmiştir. Uzman görüşlerini de kapsayacak şekilde dokümandan üretilen veriseti XML formatında erişime açılmıştır[1]. Sistem performans ölçümü için ise uzman görüşleri ile sistem çıktıları karşılaştırılmıştır. Performans ölçütü için izlenen yol kısaca aşağıda özetlenmiştir.



Şekil 5.Ortak paragraf içeren metin bloklarının etiketlendirilmesi (Labeling text blocks with common paragraphs)

[1] Dökümandan üretilen veriseti: https://github.com/kgunel/veriseti-Sanat-Tarihi

LC kümesi, N paragraf içeren d dokümanından elde edilen öğretim kavramlarının kümesi olsun. $k = 1, 2, \dots, N$ için,

$$T_k = \{t_{k,i} \in LC \mid i \in \{1,2,3\} \text{ için } t_{k,i}\ k.\text{paragraf için alan uzmanı tarafından belirlenen etiket}\}$$

$$O_k = \{o_{k,j} \in LC \mid j \in \{1,2,3\} \text{ için } o_{k,j}\ k.\text{paragraf için sistem tarafından belirlenen etiket}\}$$

kümeleri için $i, j \in \{1, 2, 3\}$ olacak şekilde

$$\chi_k(i,j) = \begin{cases} 0, & o_{k,j} = t_{k,i} \text{ ise} \\ 1, & \text{Aksi halde} \end{cases}$$

fonksiyonu tanımlansın. Bu durumda çoklu etiketleme sistemi için ortalama karesel hata (Mean Squared Error, MSE)

$$\text{MSE} = \frac{1}{N}\sum_{k=1}^{N}\left(\chi_k(i,j)\right)^2$$

eşitliği ile hesaplanır. Ek olarak, etiketler için önem sıralarına göre sezgisel bir ağırlıklandırma yaklaşımı gerçekleştirilmiştir. Buna göre her paragrafın en az bir en çok üç kavramla etiketlendirildiği düşünülerek, önem derecesine göre hedef etiketler sırasıyla $i = 1,2,3$ ile indislendirilmiş ve ağırlıklar sırayla 0,6, 0,3 ve 0.1 değerleriyle derecelendirilmiştir. Bu durumda ağırlıklandırılmış ortalama karesel hata miktarı (Weighted Mean Squared Error, wMSE)

$$\text{wMSE} = \frac{1}{N}\sum_{k=1}^{\text{N}}(0{,}6\chi_k(1,j)^2 + 0{,}3\chi_k(2,j)^2 + 0{,}1\chi_k(3,j)^2)$$

eşitliği ile hesaplanabilir.

Çalışmada numaralandırılan paragraflar ilk paragraftan itibaren başlayarak üçer paragraftan oluşan metin bloklarına ayrıştırılmıştır. Böylece sezgisel olarak oluşan metin blokları 1-3., 2-4., 3-5., … paragrafları kapsayacak şekilde oluşmuştur. Bununla birlikte dokümanın son paragrafına ulaşıldığında doküman tekrar başa dönüyormuş gibi yapılarak devirsel bir metin bloklama yaklaşımı kullanılmıştır. Bu yaklaşımın temel mantığı, dokümanın giriş ve sonuç bölümlerinde benzer kavramlardan bahsedilebileceği fikrine dayandırılmıştır. Dolayısıyla oluşturulan son metin blokları 90, 91 ve 1. paragraflar ile 91, 1 ve 2. paragrafları kapsar. Çalışma sonucu ortalama karesel hatalar k-means için $\text{wMSE} = 0{,}557$ ve PSO kümeleme yaklaşımı için ise $\text{wMSE} = 0{,}527$ olarak elde edilmiştir. Bu değer, her paragraf için birincil derecede ağırlıklandırılmış öğretim kavramının net olarak belirlendiğini, ikinci derece ağırlıklandırılmış

öğretim kavramlarının da paragrafların bir kısmında başarıyla tespit edildiği şeklinde yorumlanabilir. Ağırlıklandırılmış ortalama karesel hata formülünden anlaşılacağı üzere eğer bir paragrafa ait hiçbir öğretim kavramı tespit edilemezse, yani bu paragrafa dair yapılan etiketlemelerin hepsi hatalıysa 1 birimlik bir hata elde edilir. Dolayısıyla N paragraf için en fazla N birimlik bir hata elde edilebilir. Bu durum her paragrafın kesin olarak üç farklı öğretim kavramı ile etiketlenmesi ile elde edilebilecek maksimum hata miktarının oluşması ile görülebilir. Eğer paragrafların bazıları daha az sayıda öğretim kavramı ile etiketlendirilmişse etiket ağırlıkları güncellenerek elde edilecek maksimum hata miktarının yine N olması sağlanabilir. Bu şartın sağlanabilmesi için öğretim kavramlarının metin blokları ile ilişkilerini derecelendiren ağırlıklarının toplamının 1 olması yeterlidir.

Paragraf başına düşen ortalama hata miktarı ise yine 1 birimdir. Dolayısıyla hata wMSE değerinin [0,1] aralığında olması gerektedir. Haliyle 0,5 birimlik ağırlıklandırılmış ortalama karesel hata, her paragraf için üç etiketten birinin net olarak tespit edilebildiği, 2. etiketin ise bir kısmının uzman görüşüyle uyuştuğunu göstermektedir. Görüleceği üzere, önerilen modelde kullanılan her iki kümeleme yönteminde de wMSE değerleri birbirine oldukça yakın olarak hesaplanmıştır. Kümeleme yaklaşımından bağımsız olarak benzer etiketlemenin yapılması yaklaşımın tutarlılığı gösterebilmek adına önemlidir.

4. TARTIŞMA VE SONUÇ (DISCUSSION AND CONCLUSION)

Doğal dil işleme uygulamalarının hemen her alanda önem kazandığı düşünüldüğünde, eğitsel dokümanlar üzerinde de metin işleme yöntemlerine ihtiyaç duyulduğu görülecektir. Özellikle ders kitabı olarak okutulan materyallerin hangi konudan 'bahsettiğini' ortaya koymak, önemli bir problem alanı olarak görülmektedir. Kitaplardaki kavramların tam, doğru ve tüm konuyu kapsayacak şekilde belirlenmesi, gerçekleştirilen eğitim öğretimin nitelikli olmasını sağlayacaktır. Bu çalışmada, liselerde okutulan bir kitaptaki bölümlerin hangi kavramlarla ne derecede ilişkili olduğunun tespit edilebilmesi amaçlanmıştır. Çalışmada veri kaynağı olarak kullanılan Sanat Tarihi kitabında en çok geçen sözcüklerin ve kavramların 'sanat' ve 'tarih' olması, çalışmaya belli bir sınırlılık getirmiştir. Önerilen yöntemin farklı alanlar üzerinde ve doküman çeşitliliği artırılarak yapılması ve test edilmesi düşünülmektedir. Çalışma sonunda; bir metin bölütünün kaç paragraf içermesi gerektiği, paragraf sayısının her metin bloğu için farklı olacak şekilde dinamik olarak belirlenip belirlenemeyeceği, seçilecek kümeleme yaklaşımının performans üzerine etkisinin incelenmesi, metin bloklarının etiketlenmesinde kullanılacak öğretim kavramı sayısının belirlenmesi gibi pek çok araştırma problemi ortaya çıkmaktadır.

Bu çalışmanın kısıtlarından biri, önerilen yaklaşımın tek bir doküman üzerinde uygulanmasıdır. Her ne kadar çalışmanın kurgusu bu doğrultuda yapılmış olsa da, çalışmanın devamında alandan bağımsız olarak farklı konulardaki dokümanlar içinde verimli metin bölütleme ve etiketleme yapabilmesi adına yöntemin geliştirilmesi planlanmaktadır.

Ayrıca çalışmanın devamında, çoklu etiketlemeli metin bölütleme yaklaşımı ile uyarlanabilir öz-sınama (self-testing) sistem modeli geliştirilmesi ve modelin eğitim destek sistemlerine entegrasyonunun sağlanması için gerekli standartların belirlenmesi hedeflenmektedir. Öğrenenlerin öz-sınamaları sürecinde sisteme sağladıkları geri-dönütler doğrultusunda etiketleme sisteminin kendini güncellemesi hedeflenmektedir. Bu yolla, öğrenenlerin eğitim dokümanında sunulan konunun bütününden ziyade, konu içinde tanıtılan öğrenim kavramlarını ve ilintili alt öğrenme kavramlarının bulunduğu metin bölütlerine yönlendirilmesi hedeflenmektedir. Böylece öğrenene bireysel öğrenme sürecinde, konudan ziyade konu içinde yer alan öğretim kavramları düzeyine inilerek, görsel bir yol haritası sunulmuş olacaktır.

KAYNAKLAR (REFERENCES)

[1] İnternet: http://www.tdk.gov.tr, 27.04.2021

[2] B. Ü. Bozkurt, "Kavram, Kavramsallaştırma Yaklaşımları ve Kavram Öğretimi Modelleri: Kuramsal Bir Derleme ve Sözcük Öğretimi Açısından Bir Değerlendirme", *Ankara Üniversitesi Dil Dergisi*, 2018.

[3] R. Luckin and M. Cukurova, "Designing educational technologies in the age of AI: A learning sciences-driven approach", *British Journal of Educational Technology*, 50 (6), 2824-2838, 2019. https://doi.org/10.1111/bjet.12861

[4] A. Renz and R. Hilbig, "Prerequisites for artificial intelligence in further education: identification of drivers, barriers, and business models of educational technology companies", *International Journal of Educational Technology in Higher Education*, 17, 14 (2020). https://doi.org/10.1186/s41239-020-00193-3.

[5] F. Ouyang and P. Jiao, "Artificial intelligence in education: The three paradigms", *Computers and Education: Artificial Intelligence*, 2, 100020, 2021.

[6] X. Zhai et. al, "A Review of Artificial Intelligence (AI) in Education from 2010 to 2020", *Complexity*. https://doi.org/10.1155/2021/8812542

[7] V. Nabiyev and A.K. Erümit, **Eğitimde Yapay Zeka: Kuramdan Uygulamaya**, 2020.

[8] M. Akdeniz and F. Özdinç , "Eğitimde Yapay Zeka Konusunda Türkiye Adresli Çalışmaların İncelenmesi", *Van Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi*, 18(1), 912-932, 2021. https://doi.org/10.33711/yyuefd.938734

[9] S. Burkhardt and S. Kramer, "Online multi-label dependency topic models for text classification", *Mach Learn*, 107, 859–886, 2018.

[10] J. M. Moyano, E.L. Gibaja, K. J. Cios and S. Ventura, "Review of ensembles of multi-label classifiers: Models, experimental study and prospects", *Information Fusion*, 44, 33-45, 2018.

[11] A.N. Tarekegn, M. Giacobini, and K. Michalak, "A review of methods for imbalanced multi-label classification", *Pattern Recognition*, 118, 2021.

[12] L. Liu, J. Kang, J. Yu, and Z. Wang, "A comparative study on unsupervised feature selection methods for text clustering", **In Proc. 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering**, 597–601, 2005.

[13] V. Kumar, A. K. Pujari, V. Padmanaphan, S. K. Sahu, and V. R. Kagita, "Multi-label classification using hierarchical embedding', *Expert System with Application,* 91, 263-269, 2018.

[14] J. Lee, I. Yu, P. Park and D.W. Kim, "Memetic feature selection for multilabel text categorization using label frequency difference", *Information Sciences,* 485, 263-280, 2019.

[15] Z. Yang and G. Liu, "Hierarchical Sequence-to-Sequence Model for Multi-Label Text Classification", *IEEE Access,* 7, 153012-153020, 2019.

[16] N. Aljedani, R. Alotaibi and M. Taileb, "HMATC: Hierarchical multi-label Arabic text classification model using machine learning", *Egyptian Informatics Journal,* 2020.

[17] P. Deepak, K. Visweswariah, N. Wiratunga, and S. Sani, "Two-part segmentation of text documents", **Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM '12),** Association for Computing Machinery, New York, NY, USA, 793–802, 2012.

[18] I. Pak and L. P. Teh, "Text Segmentation Techniques: A Critical Review", *Springer International Publishing,* 2017.

[19] D. Beeferman, A. Berger and J. Lafferty, "Statistical Models for Text Segmentation", *Machine Learning* 34, 177–210, 1999.

[20] H. Oh, S. H. Myaeng, and M. G. Jang, "Semantic passage segmentation based on sentence topics for question answering', *Information Science (Ny),* 177, 3696–3717, 2007.

[21] G. K. Hoon, P. K. Keong and T. E. Kong, "A Semantic Learning Approach for Mapping Unstructured Query to Web Resources", **IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings) (WI'06),** 494-497, https://doi.org/10.1109/WI.2006.24, 2006.

[22] G. K. Hoon and T. C. Wei, "Flexible facets generation for faceted search", **First EAI International Conference on Computer Science Engineering EAI** 1–3 Penang: Malaysia, 2017.

[23] D. Duan, W. Qian and S. Pan, "VISA: A visual sentiment analysis system", **Proceedings 5th International Symposium Visa Information Communicate Interaction— VINCI'12**, 22–28, ACM: Hangzhou, 2012.

[24] C. Q. G. Wentao and C. Scheepers, "Effects of Text Segmentation On Silent Reading Of Chinese Regulated Poems: Evidence From Eye Movements", *The Journal of Chinese Linguistics,* 44(2) 265-286, 2017.

[25] K. Günel, R. Polat and M. Kurt, "Analyzing Learning Concepts in Intelligent Tutoring Systems", *The International Arab Journal of Information Technology*, 13(2), 2016.

[26] B. T. Dinçer and B. Karaoğlan, "Stemming in Agglutinative Languages: A Probabilistic Stemmer for Turkish", **Computer and Information Sciences - ISCIS 2003, 18th International Symposium,** Antalya, Turkey, 2003.

[27] C. Meadow, B. Boyce and D. Kraft, **Text Information Retrieval Systems**, second ed. Academic Press, 2000.

[28] M. Bilgin, "Kelime Vektörü Yöntemlerinin Model Oluşturma Sürelerinin Karşılaştırılması", *Bilişim Teknolojileri Dergisi*, Cilt 12(2), 141 – 146, 2019, doi:10.17671/gazibtd.472226.

[29] K. Zubrinik, D. Kalpic and M. Milicevic, "The automatic creation of concept maps from documents written using morphologically rich languages", *Expert System with Aplications,* 39(16), 12709-12718, 2012.

[30] C. Manning, P. Raghavan and H. Schütze, "An Introduction to Information Retrieval", *Cambridge University Press,* 2009.

[31] S. Kumova, S. and B. Karaoğlan, "Stop Word Detection as A Binary Classification Problem", *Anadolu University Journal of Science and Technology,* 18(2), 346 – 359, 2017.

[32] R. Aşliyan R., K. Günel and T. Yakhno, Detecting Misspelled Words in Turkish Text Using Syllable n-gram Frequencies, In: Ghosh A., De R.K., Pal S.K. (eds), Pattern Recognition and Machine Intelligence, PReMI 2007, Lecture Notes in Computer Science, 4815, Springer, Berlin, Heidelberg, 2007.

[33] I. Jolliffe, "Principal Component Analysis", *Encyclopedia of Statistics in Behavioral Science,* 648, 2005.

[34] A. Geron, **Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems,** Oreilly, 2nd Edition, 2019.

[35] S. X. Yang and C. H. Li, "A Clustering Particle Swarm Optimizer for Locating and Tracking Multiple Optima in Dynamic Environments", *Ieee Transactions On Evolutionary Computation,* 2010.