DEU FMD 22(64), 67-80, 2020

Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi
Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering

Basılı/Printed ISSN: 1302-9304. Elektronik/Online ISSN: 2547-958X

Tıp Veri Kümesi için Gizli Dirichlet Ayrımı

Latent Dirichlet Allocation for Medical Dataset

Ekin Ekinci [1*], Sevinç İlhan Omurca [2], Elif Kırık [1.2], Şeymanur Taşçı [1]

[1] Kocaeli Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Kocaeli, TÜRKİYE
[2] Enuygun, İstanbul, TÜRKİYE
*Sorumlu Yazar / Corresponding Author **: silhan@kocaeli.edu.tr

Geliş Tarihi / Received: 12.03.2019 *Araştırma Makalesi/Research Article*
Kabul Tarihi / Accepted: 20.09.2019 *DOI: 10.21205/deufmd.2020226408*

Atıf şekli/ How to cite: EKİNCİ, E., OMURCA, S.O., KIRIK, E., TAŞÇI, Ş. (2020). Tıp Veri Kümesi için Gizli Dirichlet Ayrımı. DEUFMD 22(64),67-80.

Öz

Bilimsel çalışmalarda ilgili alandaki literatürün incelenmesi oldukça önemli bir aşamadır. Literatür insan tarafından tarandığında, geniş kapsamlı bir inceleme yapılması mümkün olamamakta, ya da böyle bir arama çok uzun zaman almaktadır. Öte yandan literatürün otomatik olarak taranması derinlemesine bir anlamsal analizi mümkün kılmamaktadır. Bu çalışma kapsamında Türkiye'deki araştırmacılar tarafından yayınlanmış tıp makalelerinin otomatik ve anlamsal analizini gerçekleştiren bir konu modelleme yöntemi olan Gizli Dirichlet Ayrımı (GDA) uygulanmıştır. Deneysel çalışma, yıllara göre bir tıp veritabanı olan PubMed'den elde edilen son 11 (on bir) yıldaki yayınla tıp literatüründeki makaleler üzerinde gerçeklenmiştir. Deneysel sonuçlar incelendiğinde, son 11 (on bir) yılda trend olan çalışma başlıklarının başarılı bir şekilde keşfedildiği gözlenmiştir.

Anahtar Kelimeler: *Konu modelleme, Gizli Dirichlet Ayrımı (GDA), Tıp literatürü, PubMed*

Abstract

Examination of the literature in the relevant field is a very important stage in scientific studies. When the literature is reviewed manually, it is not possible to perform a comprehensive review or such a search takes a very long time. On the other hand, the automatic search of the literature does not enable in-depth semantic search. In this study, a topic modelling method Latent Dirichlet Allocation (LDA), that performs the automatic and semantic analysis of medical articles published by researchers in Turkey, is applied. The experimental study was carried out on articles in the medical literature in the last 11 (eleven) years from PubMed, which is a medical database based on years. When the experimental results are analyzed, it has been observed that the titles, which have trend in the last 11 (eleven) years, have been discovered successfully.

Keywords: *Topic modelling, Latent Dirichlet Allocation (LDA), Medical literature, PubMed*

1. Giriş

Günümüzde insanların bilgiye erişiminde internetin önemi çok büyüktür. Web sayfaları, haberler, bloglar, sosyal medya platformları, e-ticaret siteleri, bilimsel makaleler gibi çok farklı kaynaklar her alanda bilgiye erişimi oldukça kolaylaştırmaktadır. Bunun sonucu olarak, internet kaynaklarında depolanan verinin boyutu da her geçen gün artmaktadır. Depolanan verinin çok büyük oranda metin verisi olduğu göz önüne alındığında, metin verilerinin otomatik analizi oldukça önemli bir araştırma problemi haline dönüşmektedir. Bu doğrultuda,

geniş ölçekli metin verilerinde arama, anlama ve işleme görevlerini yerine getirecek otomatik araçlara ihtiyaç ortaya çıkmıştır. Bu boşluğu doldurmak için konu modelleme yöntemleri; makine öğrenmesi, doğal dil işleme ve bilgi çıkarımı süreçlerinde yaygın şekilde uygulanmaya başlanmıştır. Konu modelleme, geniş ölçekli doküman koleksiyonlarından anlamsal bilgiye erişimde uygulanan denetimsiz bir makine öğrenmesi yöntemidir. Literatürde, araştırmacılar tarafından geliştirilen mevcut pek çok konu modeli bulunmaktadır. Bu konu modelleri arasında en yaygın ve tam olanı Gizli Dirichlet Ayrımı (GDA)'dır [1-4]. Dokümanlar gibi ayrık verileri modellemek için geliştirilen üretici grafiksel bir model olan GDA dokümanı oluşturan gizli konuları ortaya çıkarmaktadır [5]. GDA'nın dayandığı temel fikir, konuların sabit bir sözlük üzerinden olasılık dağılımına sahip olması ve dokümanların gizli konuların rastgele bileşiminden oluşmasıdır. Bu temel fikre göre GDA, doküman koleksiyonundaki konuları, konuları oluşturan kelimelerin konular altındaki olasılıklarını, dokümanlar için o dokümanı oluşturan kelimelerin hangi konulara atandığını ve her doküman için bu dokümandaki konuların dağılımını öğrenmektedir [6].

Tıp alanındaki çevrimsel araştırmalar, temel laboratuvar bilimini etkili hasta terapilerine mümkün olduğunca çabuk dönüştürmekle ilgilidir. Etkili tedavilerin geliştirilmesi, fizyolojik, hücresel ve moleküler düzeyde tıp, farmakoloji, biyoloji ve kimya disiplinler arası bir anlayış gerektirir. Örneğin, biyomedikal alanda kimyasallar, genler ve hastalıklar arasındaki ilişkilerin hızlı bir şekilde keşfi, ilaç keşifleri gibi araştırmalarda çok değerlidir. Bugünün tıp doktorları ve biyomedikal araştırmacıları, çeşitli kaynaklardan giderek artan miktarda yüksek boyutlu, heterojen ve karmaşık verilerle karşı karşıya kalmaktadır [7]. Özgün çalışmaların gerçekleştirilebilmesi için bu çeşitli alanlarda yayınlanmış literatürün derinine incelenmesi önem arz etmektedir. Bunun yanı sıra karmaşık disiplinler arası araştırmalar, yayınlanmış literatürde yer alan en son bilgileri, eğilimleri ve bulguları makul bir süre içinde verimli bir şekilde keşfetmeyi, değerlendirmeyi ve sentezlemeyi zorlaştırmaktadır. Bu nedenle, özetlerin ve tam metinli dergi makalelerinin sistematik analizi yoluyla bilgi keşfini kolaylaştırmak için yararlı yaklaşımlar üretmek önemli ve devam eden bir sorun haline gelmiştir. Bu çalışmada, Türkiye'de tıp alanındaki araştırmacılar için hangi konularda daha aktif çalışıldığını saptamanın yaralı olacağı düşünülerek; tıp alanında son 11 (on bir) yılda trend olan konu başlıklarının otomatik olarak keşfedilmesi sağlanmıştır. Bunun yanı sıra yıl bazında trend olan çalışma konuları kolayca keşfedilebilmekte ve bu konularda yayınlanan belirgin makalelerin otomatik ve kolay erişimi de sağlanabilmektedir.

Makalenin geri kalan kısmı; ikinci bölümde mevcut çalışmalar, üçüncü bölümde deneysel yöntemin, GDA ve hata analizi başlıkları altında ayrıntılı şekilde incelenmesi; dördüncü bölümde veri kaynaklarının elde edilmesi ve bu veri kaynaklarına yapılan önişleme adımlarının, yapılan parametre ayarlarının ve deneysel çalışmanın anlatılması ve son olarak da tartışma ve sonuçlar şeklinde verilmiştir.

2. Mevcut Çalışmalar

PubMed, dergilerde ve kitaplarda yayınlanan makaleler için 28 (yirmi sekiz) milyondan fazla alıntı sağlayan çevrimiçi bir kaynaktır ve çoğu kısa özetlerle ilişkilendirilirken, giderek artan sayıda ücretsiz tam metin makalelerine de erişilebilmektedir. Bu çalışma kapsamında literatüre erişim PubMed veri tabanı üzerinden gerçekleştirilmiştir. Mevcut literatür incelendiğinde benzer çalışmaların yer aldığı gözlenmiştir.

[8] nörobilimin genel bir analizini elde etmek için 2001-2006 yılları arasında Society for Neuroscience (SFN) yıllık toplantı özetlerinden özet metinleri elde edip bu metinleri işlemişlerdir. Doğal dil işleme, metin madenciliği ve diğer veri analizi tekniklerini kullanarak, bilimsel işbirliği ağının demografik ve yapısını, zaman içindeki alanın dinamiklerini, araştırma eğilimlerini incelemişlerdir. Elde ettikleri SFN özet kümesinin içerdiği konu başlıklarını tanımlamak için Gizli Anlamsal Analiz (GAA) yöntemini kullanmışlardır.

[9] PubMed veri tabanındaki özetleri kullanarak kelimeler, MeSH terimleri, dokümanlar ve konu başlıkları arasındaki ilişkileri keşfetmek için Topic-concept model adını verdikleri bir konu modelleme yaklaşımı önermişlerdir. [10] 2647 diyabet hastasına ait vakaları incelemek üzere LinkLDA modeli önermişlerdir. Hastalığa ait bazı semptomlara ve bu semptomları tedavi edebilecek bazı bitkilere ait konu başlıkları keşfetmeyi hedeflemişlerdir.

[11] elektronik hasta kayıtlarını analiz etmek için konu modelleme, temel bileşenler analizi ve ikili kümeleme tekniklerini uygulamışlardır. Hastalara ait kanser kayıtlarına ve kanserle ilişkili olabilecek biyolojik etkenleri keşfetme konusuna odaklanmışlardır. [12] son 5 (beş) yıl içinde artış gösteren 5 (beş) farklı kanser türü hakkındaki anahtar konu başlıklarının çıkarılması için biyomedikal literatürüne GDA algoritmasını uygulamışlardır. GDA parametrelerinin güncellenmesini Gibbs örnekleme yöntemi ile gerçekleştirmişlerdir.

[13] hiyerarşik Dirichlet süreci kullanarak konuları otomatik olarak çıkartan bir yapı önermişlerdir. Sonrasında çıkan konular arasında bir ilişki ağı oluşturmuşlardır. Modelin etkinliğini kanıtlamak için 18.000 dokümandan oluşan autism spectrum disorder (ASD) literatürünü kullanmışlardır. [14] Alzheimer hastalığına ait literatürü incelemek için PubMed veri tabanında de yayınlanan 96.081 makale üzerinde "concept graph-based network analysis" yaklaşımı ile bir inceleme gerçekleştirmişlerdir. Aynı zamanda zamana bağlı trend konu çıkarımını "Dirichlet multinomial regression topic modeling" yöntemi ile gerçekleştirmişlerdir. 2013 yılının bu alanda en çok yayın yapılan yıl olduğu sonucuna varmışlardır. [15] PubMed ve PubMed Central veritabanlarından elde ettikleri geniş ölçekli biyomedikal yayınlar üzerinde GDA yöntemini büyük veri kavramı ile ilgili başlıkları keşfetmek üzere uygulamışlardır. Sonrasında 7 (yedi) uzman kişi, model tarafından çıkarılan başlıkları büyük veri tanımları ile eşleştirmiştir.

[16] Konu modelleme yöntemlerini kullanarak, serbest metinler olarak yazılmış klinik dokümanlardan klinik kavramları temsil eden kelime öbeklerini keşfetmeyi amaçlamışlardır. [17] 1998-2016 yılları arasında yayınlanmış biyoinformatik literatürünü analiz etmek için konu modelleme yaklaşımını kullanmışlardır. Her yıla ait en yaygın kullanılmış olan anahtar kelimeler analiz edilmiş ve sonrasında daha detaylı analiz için konular otomatik olarak keşfedilmiştir. Çalışmaları sonucunda kanser alanında büyük veri tekniklerini kullanarak çözüm arayan araştırmaların son yıllarda artış gösterdiği sonucuna varmışlardır.

[18] PubMed veritabanında 2016 ve 2017 yılları arasında yayınlanmış olan sadece elektronik sağlık literatürüne ait trend konuları keşfetmek üzere klasik GDA yöntemini kullanmışlardır. Bu Yaptıkları çalışma ile, elektronik sağlık alanda hangi konuların bir yıl içerisinde pozitif trend hangilerinin ise negatif trend gösterdiği tespit edilmiştir.

Biyomedikal literatürün analizine ilişkin çalışmalar incelendiğinde, araştırma makalelerinde, disiplinler arası işbirliklerinden dolayı giderek artan bir konu karmaşıklığı olduğu açıktır. Bu durum da, biyomedikal metinlerde birlikte ya da sık çalışılan konu başlıklarının belirlenmesini zorlaştırmaktadır. Bildiğimiz kadarı ile Türkiye'de tıp alanında yayınlanmış bilimsel makalelerden otomatik konu keşfeden bir çalışma yer almamaktadır. Çalışmamız bu anlamda gerçekleştirilen ilk çalışmadır. Ayrıca 2007-2017 yılları arasında yayınlanmış olan makalelerin incelendiği göz önüne alındığında oldukça geniş kapsamlı bir analiz gerçekleştirilmiştir.

3. Materyal ve Metot

3.1. Gizli dirichlet ayrımı

Doküman koleksiyonlarındaki gizli tematik bilgiyi küçük boyutlu uzaya çevirerek keşfeden bir grup algoritma olarak tanımlanan olasılıksal konu modelleri son yıllarda makine öğrenmesi ve metin madenciliği alanlarında büyük önem kazanan araştırma konularından birisi haline gelmiştir [5,19-21]. Bu yöntemler altında yatan temel fikir kelimeler üzerinde olasılık dağılımı gösteren konuların rastgele bir araya gelerek dokümanları oluşturması şeklinde açıklanmaktadır [21]. Konu ise dokümanda tartışılan temel fikirdir yani dokümanın temasıdır.

Konu modelleme yöntemleri üzerine literatürde pek çok başarılı çalışma olmakla birlikte, hala daha teorikte anlaşılması güç bir konu olarak karşımıza çıkmaktadır. Konu modelleri ilk olarak [22] tarafından geliştirilen GAA yöntemi ile ortaya çıkmıştır. GAA'da doküman koleksiyonundaki gizli anlamsal ilişkiler keşfedilip düşük boyutlu anlamsal bir uzay elde etmek amacıyla doküman terim matrisi üzerinden tekil değer ayrışımı uygulanmaktadır. GAA'nın olasılık tabanlı formu olarak 1999 yılında geliştirilen Olasılıksal Gizli Anlamsal Analiz (Olasılıksal GAA) üretici grafiksel bir konu modelidir [23]. Olasılıksal GAA dokümanların gizli konuların rastgele bir araya gelerek dokümanı oluşturduğu fikrine dayalı ilk yöntemdir. Model sadece kelime seviyesinde bir olasılık modeli sunmaktadır bu da modelin tam

bir üretici model olmasını engellemektedir. Ayrıca aşırı öğrenmeye eğilimi olması ve yeni gördüğü dokümanlar üzerinde genelleme yapamaması modelin dezavantajlarıdır [24]. [5] tarafından geliştirilen Gizli Dirichlet Ayırımı (GDA) kelimelerin konulardaki, konuların da dokümanlardaki dağılımını Dirichlet dağılımından elde ederek Olasılıksal GAA yöntemini geliştirip tam bir üretici modeli sunmaktadır.

GDA, doküman gibi ayrık verileri modellemek ve dokümanı oluşturan gizli konuları ortaya çıkarmak için geliştirilmiş üretici grafiksel bir modeldir [5]. GDA'daki gizli dokümanı oluşturan gizli konuları tanımlamaktadır. Dirichlet dağılımı Beta dağılımının çok değişkenli versiyonu olup, bir simpleks ile sınırlandırılmış rastgele vektörler için çok değişkenli bir dağılım olarak tanımlanmaktadır [25]. Tamamen denetimsiz bir yöntem olan GDA herhangi bir önbilgiye ihtiyaç duymaz. Kelime torbası yaklaşımına dayalı yöntemde kelimelerin doküman içerisindeki yerleşimi göz ardı edilmekte, kelimelerin birlikte bulunması ise göz önünde bulundurulmaktadır. GDA'nın üretici bir model olması demek basit bir olasılıksal süreç ile dokümandaki kelimelerin rastgele değişkenler ile oluşturulmasıdır, yani dokümanın oluşturulmasıdır [19]. GDA için üretici model Şekil 1'de verilmiştir.

1. Her konu k için $k \in [1, K]$

   a. Kelimelerin konular içerisindeki dağılımını belirle: $\varphi_k \sim Dirichlet(\beta)$

2. Her doküman m için $m \in [1, M]$

   a. Konuların doküman içerisindeki dağılımını belirle: $\theta_m \sim Dirichlet(\alpha)$

   b. Doküman m'deki her kelime w_n için

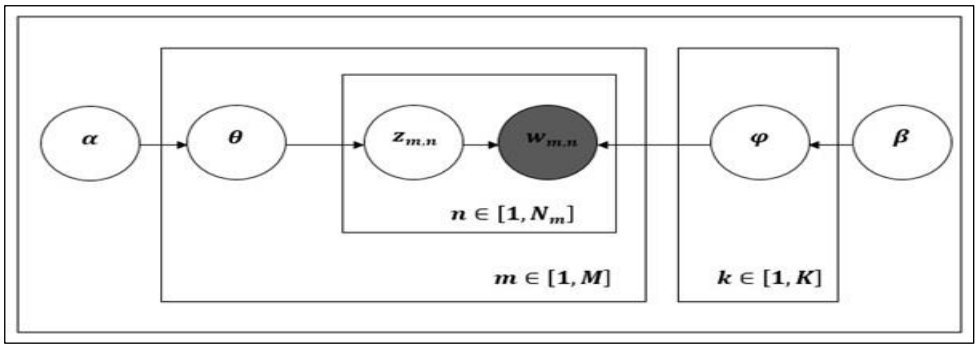
      i. Rastgele bir konu belirle $z_{m,n} \sim Mult(\theta_m)$

      ii. Rastgele bir kelime seç $w_{m,n} \sim Mult(\varphi_k, z_{m,n})$

Şekil 1. GDA için üretici model [26]

Üretici model ilk olarak doküman koleksiyonu içerisindeki tüm kelimeleri kapsayan sabit bir sözlük olan V'nin bütün konular içerisindeki dağılımını Dirichlet dağılımına göre belirler. Toplam K konu içerisinden k. konudaki kelimelerin dağılımı φ_k ile temsil edilir ve φ_k Dirichlet parametresi olan β'ya göre hesaplanır. İkinci adımda toplam boyutu M olan koleksiyondaki her doküman m için ilk olarak Dirichlet parametresi olan α'ya göre konuların doküman içerisindeki dağılımı θ_m belirlenir. Toplamda N_m tane kelime içeren m dokümanındaki her kelime wm,n için bu kelimenin konusu zm,n çok terimli dağılıma göre örneklenir. Son olarak da yine çok terimli dağılıma göre ilgili konu için rastgele bir kelime örneklenir.

GDA aynı zamanda grafiksel bir modeldir ve grafiksel temsilinde plate notasyonu kullanılmaktadır. Plate notasyonu aynı tipte birden fazla nesnenin olduğu durumlar için kullanılmaktadır. Plate notasyonu GDA'da ise gözlemlenen verinin gizli değişkenler ve bu gizli değişkenlerin yönlü kenarlar üzerinden yayılımı ile nasıl üretildiğini göstermek amacıyla kullanılmaktadır [27]. Burada gözlemlenen veriler ile kastedilen dokümanı oluşturan kelimeler iken, gizli değişkenler ile kastedilen ise kelimelerin konulardaki dağılımı, konular ve konuların dokümandaki dağılımıdır. GDA'ya ait plate notasyonu Şekil 2'de verilmiştir. Grafiksel modelde gözlemlenen değişkenler gri renkle temsil edilirken gizli değişkenler beyaz renk ile temsil edilmiştir.



Şekil 2. GDA'ya ait plate notasyonu [28]

Şekil 2'deki grafiksel modele göre gözlemlenen ve gizli değişkenlere için bileşik dağılımı $p(\varphi_{1:K}, \theta_{1:M}, z_{1:M}, w_{1:M})$ Denklem 1'de verilmiştir.

$$\left(\prod_{k=1}^{K} p(\varphi_k|\beta)\right)\left(\prod_{m=1}^{M} p(\theta_m|\alpha)\right) \left(\prod_{n=1}^{N_m} p(z_{m,n}|\theta_m)(w_{m,n}|z_{m,n}, \varphi_k)\right) \quad (1)$$

GDA'nın hedefi gizli değişkenlerin elde edilmesidir. Gizli değişkenlerin $p(\varphi_{1:K}, \theta_{1:M}, z_{1:M}|w_{1:M})$ elde edilmesinde Denklem 2 kullanılmaktadır.

$$\frac{p(\varphi_{1:K}, \theta_{1:M}, z_{1:M}, w_{1:M})}{p(w_{1:M})} \quad (2)$$

Denklem 2'deki payın tüm rastgele değişkenlerin birleşik dağılımı olduğu ve kolayca hesaplanabileceği görülmektedir. Ancak payda gözlemlerin marjinal olasılığıdır dolayısıyla bu olasılık doküman kümesinin herhangi bir konu modeli altındaki olasılığını ifade etmektedir. Hesaplanabilmesi için de gizli konu yapısının tüm örnekleri üzerinden birleşik dağılımı toplamak gerekmektedir. Yalnız bu toplamın hesaplanması mümkün konu yapısının oldukça fazla olmasından ötürü mümkün değildir. Bu nedenle sonsal dağılıma yakınsamak için örnekleme algoritmalarından yararlanılmaktadır. GDA için kullanılan örnekleme algoritması ise Gibbs örneklemenin standart bir gerçekleştirimi olan Collapsed Gibbs Örnekleme (CGÖ)'dir.

Griffiths ve Steyvers tarafından önerilen CGÖ model parametreleri olan θ ve φ'yi dışarlamakta, kelimelere konu atamada kullanılan parametre z önce dokümandaki her kelime için daha sonra koleksiyonundaki diğer tüm dokümanlarda yer alan kelimeler için iteratif olarak yeniden örneklenmektedir [29]. CGÖ'de model parametreleri dışarlandığı için koleksiyonundaki kelimelerin konulara atanmasında diğer kelimeler model parametrelerinin vekili olarak kullanılmaktadır. Örnekleme ile belli bir kelimenin hangi konuya atanacağı Denklem 3 ile hesaplanmaktadır.

$$p(z_i = k|w_i = v, m, \alpha, \beta, .) = \frac{n_{v,k} + \beta}{\sum_{w \in V} n_{w,k} + V\beta} \frac{n_{m,k} + \alpha}{N_m - 1 + \alpha K} \quad (3)$$

Denklem 3'ün sol tarafında w_i, m, α ve β diğer tüm kelimelerin hangi konuya atanmış oldukları ('.' ile temsil ediliyor) biliniyor iken z_i=k olma olasılığı bulunmaktadır. Eşitliğin sağ tarafında ise $n_{v,k}$ ile v kelimesinin k. konuya tüm koleksiyonda kaç kere atandığı hesaplanmakta, $n_{w,k}$ k. konunun tüm koleksiyonda kaç kere kullanıldığını göstermekte, $n_{m,k}$ m. yorumda k. konuya atanan kelime sayısını temsil etmektedir. Başlangıç değerleri kullanılarak Denklem 3 doküman koleksiyonundaki tüm dokümanlardaki tüm kelimeler için yapıldıktan sonra θ ve φ değerleri Denklem 4 ve Denklem 5'e göre güncellenmektedir.

$$\varphi_{v,k} = \frac{n_{v,k} + \beta}{\sum_{w \in V} n_{w,k} + V\beta} \quad (4)$$

$$\theta_{m,k} = \frac{n_{m,k} + \alpha}{N_m - 1 + \alpha K} \quad (5)$$

3.2. Hata analizi

GDA ile elde edilen konuların kendi içlerindeki anlamsal uyumluluğunu ölçmek amacıyla konu uyumluluğu ölçütünden yararlanılmaktadır [30]. Konu uyumluluğu bir konuyu oluşturan sözcükler arasındaki ortalama anlamsal ilişki olarak tanımlar [31]. Konu uyumluluğu Denklem 6'ya göre hesaplanmaktadır.

$$C(k; V^k) = \sum_{n=2}^{N} \sum_{l=1}^{n-1} log \frac{D\left(v_n^{(k)}, v_l^{(k)}\right) + 1}{D\left(v_l^{(k)}\right)} \qquad (6)$$

$V^{(k)}=(v_1^{(k)}, \ldots, v_S^{(k)})$ k. konudaki en mümkün S kelimenin listesidir. $D(v_n^{(k)}, v_l^{(k)})$; v_n ve v_l kelimelerinin birlikte geçtiği doküman sayısıdır. 1 ise yumuşatma için kullanılmıştır. $D(v_l^{(k)})$, v_l'nin tüm koleksiyondaki frekansıdır. Yüksek konu uyumluluğu modelin başarılı olduğunu göstermektedir.

4. Bulgular

4.1. Veri kaynakları ve önişleme

Bu çalışmada kullanılan veri kümesi PubMed Biyomedikal Veritabanı kaynağından elde edilen İngilizce makalelerden oluşturulmuştur. Ülkemizde çalışılan alanların tespit edilmesini amaçladığımız için de makalelerin tünü Türkiye'de bulunan doktorlara aittir. PubMed, ücretsiz bir biyomedikal veritabanıdır. Sitede yer alan başvuru kitapları moleküler biyoloji, genetik ve tıp bilimleriyle ilgili konulara ışık tutmaktadır. PubMed, biyomedikal literatür için MEDLINE'dan 20 (yirmi) milyondan fazla atıf bilgisi ile yaşam bilimleri dergileri ve çevrimiçi kitapları içermektedir. PubMed ayrıca ilişkili web sayfaları ile NCBI (National Center for Biotechnology Information)'un diğer moleküler biyoloji kaynaklarına erişim sağlayıcı bağlantılar sunmaktadır.

Veri kaynaklarının eXtensible Markup Language (XML) formatında elde edilmesinde açık kaynak bir veri analizi, raporlama ve uyum platformu olan Konstanz Information Miner (KNIME) kullanılmıştır [32]. Verinin XML formatında elde edilmesinin amacı ise verileri kolayca erişilebilecek ve paylaşılabilecek bir biçimde saklamasıdır. KNIME ile çekilen verilerin formatına bir örnek Şekil 3'te verilmiştir.

Veri kümesi 2007-2017 yılları arasında yayınlanmış makalelerin özetlerini içermektedir ve yıllara göre makale sayıları Tablo 1.'de özetlenmektedir.

Tablo 1. Yıllara göre elde edilen makale sayıları.

Yıl	Elde Edilen Makale Sayısı
2007	10502
2008	10806
2009	10807
2010	10900
2011	11864
2012	13362
2013	15044
2014	16814
2015	18348
2016	18080
2017	15698

İnternet üzerindeki veriler kolay elde edilebilir olmalarına karşın kolayca kullanılabilecek bir formata sahip olmayabilirler. Bu durumda verilerin uygun formlara dönüştürülmeleri için ön işleme adımlarının uygulanması gerekmektedir. Çalışmamızda gerçekleştirilen ön işleme adımları sırası ile şu şekilde gerçekleştirilmiştir: (I) Model büyük küçük harf duyarlı olmadığı için ilk olarak metinlerdeki verilerin hepsi küçük harfe çevrilmiştir. (II) Sonrasında küçük harfe çevrilen metin verileri birimlendirici ile birimlerine ayrılmıştır. (III) Metinlerdeki tüm noktalama işaretleri konuların çıkartılmasında bir anlam ifade edilmediği için kaldırılmıştır. (IV) İngilizce doğal dil işlemede durak kelimeleri olarak listelenen kelimeler (the, am, who, …) çıkartılmış ve doğal dil işleme aracı olan Natural Language Toolkit (NLTK)'nın durak kelime temizleme modülü olan stopword modülü ile metinler bu kelimelerden temizlenmişlerdir. (V) Son olarak modelin doğru çalışabilmesi için kelimeler gövdelemeye tabii tutulmuşlar ve yine NLTK'nın sunmuş olduğu WordNetLemmatizer modülü bu aşamada kullanılmıştır. NLTK doğal dildeki metinleri işleyen Python için geliştirilmiş bir kütüphanedir.

```
        <Pagination>
            <MedlinePgn>465-9</MedlinePgn>
        </Pagination>
        <Abstract>
            <AbstractText>The present investigation was undertaken to assess the effects of aflatoxin (AF) containing
tes were observed in the periphery of capillaries and around endocrine islets in the experimental groups. Furthermore, cap
        </Abstract>
```

Şekil 3. Veri formatı örneği

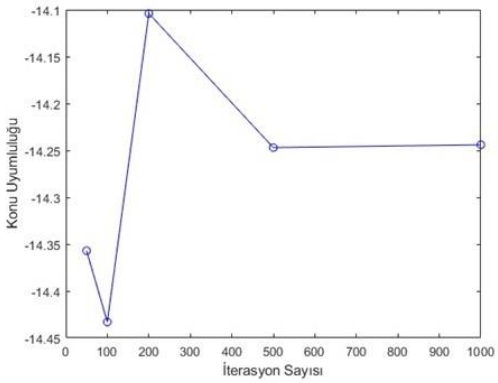
4.2. Parametre ayarları

GDA için Gibbs örnekleme 50, 100, 200, 500 ve 1000 iterasyonda çalıştırılmıştır. K ile temsil edilen konu sayısı 20 (yirmi) olarak belirlenmiştir. Dirichlet hiperparametreleri α ve β sırasıyla 50/K=2.5 ve 0.01 olarak belirlenmiştir. Model parametreleri metin sayısı ile toplam kelime sayısının çarpımı kadar güncellenmektedir. Çıkartılan her konu ise ilk 10 (on) kelimesiyle temsil edilmiştir. Her yıl için 20 (yirmi) konudan anlamlı olma durumlarına göre 4 (dört) ya da 5 (beş) tanesi o yılı temsil etmek için seçilmiştir.

4.3. Deneysel sonuçlar

Bu çalışma kapsamında GDA farklı sayılarda metinlerden oluşan 11 farklı veri kümesine uygulanmış ve sonuçlar hem niceliksel hem de niteliksel olarak değerlendirilmiştir. Niceliksel değerlendirmede konu uyumluluğu ölçütünden yararlanılmışken, niteliksel değerlendirme için uzman tıp doktoru konuların geçerliliğini kontrol etmiş ve konu etiketlerini belirlemiştir. Böylece yıllara göre hangi konular üzerine araştırmalar yapıldığı ve hangi konuların popüler olduğu belirlenmiştir.

Modelin yıllara göre elde edilen ortalama konu uyumluluğu iterasyon sayısına bağlı olarak Şekil 4'te verilmiştir.



Şekil 4. İterasyon sayısına göre ortalama konu uyumluluğu

Şekil 4 incelendiğinde modelin konu uyumluluğu açısından en başarılı olduğu iterasyon sayısının 200 olduğu görülmektedir. -14.244 konu uyumluluğu ile 1000 iterasyon ikinci sırada yer almaktadır. Ancak GDA değerlendirilirken genelde 1000 iterasyon sonucunda elde edilen konu uyumluluğu ve konular dikkate alınmaktadır. Bu nedenle niteliksel değerlendirmeler 1000 iterasyona göre yapılmıştır. Modelin yıllara göre niteliksel olarak değerlendirilmesi Tablo 2, Tablo 3, Tablo 4, Tablo 5, Tablo 6, Tablo 7, Tablo 8, Tablo 9, Tablo 10, Tablo 11 ve Tablo 12'de verilmiştir.

Tablo 2. 2007 yılına ait konular.

Nöroloji	Plazma Sonuçları	Burun Hasarı	Türkiye'deki Virüsler	Akciğer veya Böbrek Komplikasyonları
nerve	p	technique	Turkey	complication
muscle	control	surgical	population	blood
type	risk	procedure	data	injury
h	serum	surgery	region	undergo
spinal	high	graft	sequence	mortality
area	factor	perform	virus	pulmonary
positiom	v	flap	gene	hospital
pylorus	0001	defect	identify	rate
side	subject	nasal	Turkish	time
hip	plasma	fracture	country	renal

Tablo 3. 2008 yılına ait konular.

Kalp Damar Cerrahisi	Göğüs Kanseri	Gebelik	Çocuklarda Kemik Yaşı	Genetik
leave	tumor	value	child	gene
artery	cancer	rate	test	isolate
nerve	stage	number	bone	infection
right	survival	first	year	mutation
coronary	garde	pregnancy	find	polymorphism
graft	carcinoma	two	clinical	frequency
verticular	breast	phase	age	strain
flap	median	observe	disease	find
defect	thyroid	find	include	genotype
cardiac	metastasis	increase	evaluate	activity

Tablo 4. 2009 yılına ait konular.

Genetik	Göğüs Kanseri	Ortopedik Operasyon	Farelerle İlgili Çalışma
gene	tumor	bone	rat
frequency	cancer	mm	effect
polymorphism	case	technique	activity
mutation	primary	measurement	control
protein	woman	fracture	tissue
population	breast	degree	increase
genotype	carcinoma	lateral	injury
risk	survival	measure	decrease
factor	stage	implant	stress
expression	thyroid	mean	liver

Tablo 5. 2010 yılına ait konular.

Farelerle İlgili Çalışma	Göğüs Kanseri	Genetik	Gebelik, Çocuklar	Cerrahi Çalışma
rat	cell	gene	woman	case
effect	cancer	frequency	pregnancy	diagnosis
increase	expression	protein	weight	lesion
control	tumor	polymorphism	first	examination
activity	tissue	method	age	present
decrease	breast	genotype	infant	rare
day	carcinoma	investigate	period	surgery
tissue	stage	dna	early	imaging
liver	show	complex	high	mass
stress	lung	find	week	undergo

Tablo 6. 2011 yılına ait konular.

Kalp damar cerrahi	Ortopedik operasyon	Türkiye'deki gen çalışması	Göğüs kanserinin görüntülenmesi	Kronik böbrek hastalığı
case	mm	gene	tumor	disease
complication	bone	isolate	diagnosis	renal
report	fracture	sample	cancer	chronic
present	measurement	turkey	lesion	acute
surgical	implant	infection	case	b
surgery	image	strain	mass	cell
artery	technique	mutation	diagnose	infection
procedure	type	test	breast	antibody
leave	root	detect	imaging	positive
right	mean	identify	examination	liver

Tablo 7. 2012 yılına ait konular.

Genetik ve gebelik	Enfeksiyon hastalıkları	Yeni doğanlarda solunum yetmezliği	Göğüs kanseri ve tiroit	Mikrobiyolojik ajanların boyanması
gene	Turkey	mortality	cancer	isolate
frequency	region	acute	tumor	test
woman	data	pulmonary	breast	infection
mutation	specie	infant	stage	sample
polymorphism	virus	early	survival	detect
risk	country	period	diagnosis	method
control	area	age	metastasis	strain
genotype	population	rate	carcinoma	positive
association	sequence	unit	thyroid	find
pregnancy	identify	failure	case	resistance

Tablo 8. 2013 yılına ait konular.

Plastik Cerrahi	Türkiye'deki enfeksiyon hastalıkları	Görüntüleme yöntemleri	Cerrahi çalışma	Farelerle ilgili çalışma
procedure	infection	find	fracture	rat
technique	isolate	lesion	bone	activity
perform	Turkey	imaging	nasal	effect
graft	sample	examination	Side	control
complication	strain	reveal	range	increase
skin	test	loss	nerve	stress
day	positive	present	type	decrease
defect	detect	normal	right	tissue
ii	b	image	surgery	oxidative
flap	specie	magnetic	anterior	antioxidant

Tablo 9. 2014 yılına ait konular.

Cerrahi teknik sonrası ağrı	Öğrencilerde depresyon	Farelerle ilgili çalışma	Kardiyoloji	Kanser
surgery	health	rat	leave	tumor
postoperative	disorder	effect	artery	cancer
pain	data	activity	blood	stage
time	scale	control	pressure	survival
procedure	student	increase	function	treatment
undergo	depression	decrease	heart	grade
complication	care	day	cardiac	month
surgical	research	c	measure	primary
technique	Turkish	tissue	increase	carcinoma
perform	participant	liver	ventricular	median

Tablo 10. 2015 yılına ait konular.

Gen mutasyonları	Türkiye'de yapılan çalışmalarla ilgili bir özet	Göğüs kanseri	Vaka (nadir görülen) çalışması	Cerrahi teknik sonrası ağrı
gene	review	cancer	case	pain
expression	report	tumor	report	surgery
mutation	country	lesion	present	procedure
isolate	research	breast	rare	postoperative
infection	turkey	diagnosis	artery	time
protein	health	find	leave	complication
identify	medical	positive	right	undergo
strain	data	biopsy	diagnosis	technique
polymorphism	management	carcinoma	cause	2
sequence	literature	vitamin	due	operation

Tablo 11. 2016 yılına ait konular.

Oksidatif stres	Cerrahi tekniğin sinirlere etkisi	Ameliyat sonrası kalp fonksiyonları	Hastane enfeksiyonları	Vaka, tümör
rat	surgery	time	treatment	case
effect	surgical	blood	infection	report
tissue	technique	leave	hospital	present
increase	mean	pressure	mortality	diagnosis
stress	complication	postoperative	therapy	tumor
decrease	procedure	heart	rate	lesion
oxidative	nerve	cardiac	clinical	mass
control	follow-up	function	day	disease
activity	fracture	v	failure	cause
damage	perform	undergo	include	rare

Tablo 12. 2017 yılına ait konular.

İlaç Çalışması	Göğüs kanseri	Gebelik	Çocuk hastalıkları	Ameliyat sonrası komplikasyonlar
cell	cancer	number	case	surgery
activity	tumor	see	disease	postoperative
acid	lesion	observe	report	surgical
protein	case	pregnancy	clinical	technique
compound	breast	data	present	undergo
drug	stage	two	diagnosis	complication
show	survival	energy	child	perform
property	diagnosis	compare	syndrome	mean
effect	imaging	cycle	treatment	procedure
extract	carcinoma	high	symptom	fracture

Çıkartılan konular tıp doktoru tarafından niteliksel olarak değerlendirilmiş ve etiketlenmiştir. Değerlendirmeler sonucunda modelin tıp makaleleri üzerinde oldukça başarılı bir şekilde çalıştığı, anlamlı konuların elde edildiği ve konuların kendi içerisinde tutarlı olduğu gözlemlenmiştir. Çıkartılan konuların yıllara göre özeti ise Tablo 13'te verilmiştir.

Tablo 13. Konuların yıllara göre özeti.

Yıllar	Çalışma Konuları				
2007	Nöroloji	Plazma Sonuçları	Burun Hasarı	Türkiye'deki Virüsler	Akciğer veya Böbrek Komplikasyonu
2008	Kalp Damar Cerrahisi	Göğüs Kanseri	Gebelik	Çocuklarda kemik yaşı	Genetik
2009	Genetik	Göğüs Kanseri	Ortopedik operasyon	Farelerle ilgili çalışma	
2010	Farelerle İlgili Çalışma	Göğüs kanseri	Genetik	Gebelik, Çocuklar	Cerrahi çalışma
2011	Kalp Damar Cerrahisi	Ortopedik operasyon	Türkiye'deki gen çalışması	Göğüs kanserinin görüntülenmesi	Kronik böbrek hastalığı
2012	Genetik ve Gebelik	Enfeksiyon hastalıkları	Yeni doğanlarda solunum yetmezliği	Göğüs kanseri ve tiroit	Mikrobiyolojik ajanların boyanması
2013	Plastik Cerrahi	Türkiye'deki enfeksiyon hastalıkları	Görüntüleme yöntemleri	Cerrahi çalışma	Farelerle ilgili çalışma
2014	Cerrahi Teknik Sonrası Ağrı	Öğrencilerde depresyon	Farelerle ilgili çalışma	Kardiyoloji	Kanser
2015	Gen Mutasyonları	Türkiye'de yapılan çalışmalarla ilgili bir özet	Göğüs kanseri	Vaka (nadir görülen) çalışması	Cerrahi teknik sonrası ağrı
2016	Oksidatif Stres	Cerrahi tekniğin sinirlere etkisi	Ameliyat sonrası kalp fonksiyonları	Hastane enfeksiyonları	Vaka, tümör
2017	İlaç Çalışması	Göğüs kanseri	Gebelik	Çocuk hastalıkları	Ameliyat sonrası komplikasyonlar

Tablo 13 genel olarak incelendiğinde 11 yıllık süreçte göğüs kanseri ve göğüs kanseri ile ilişkili konuların en sık araştırma yapılan başlıklar olduğu görülmüştür. 2009-2014 yılları arasında farelerle ilgili çalışmaların revaçta olduğu görülmektedir. 2013 yılı itibariyle cerrahi teknik ve ameliyat sonrası oluşan durumlar literatürde yer bulmaya başlamıştır. Genetik ve genetikle ilgili konular da göğüs kanseri gibi 11 yıldır neredeyse her yıl çalışılmıştır. Gebelik ve gebelikle ilişkili konuların da araştırmacıların ilgi alanına girdiği yine Tablo 13 üzerinden varılan sonuçlardan birisi olmuştur. Sadece Türkiye'nin örnek vaka seçildiği çalışmalar da 11 yıllık süreçte çalışılan önemli başlıklardan biri olarak literatürde yerini almıştır.

5. Tartışma ve Sonuç

Bilimsel literatür her geçen gün hızlı bir ivme ile genişlemektedir. Bu büyük veri kümesi üzerinden bir yargıya varmak, hangi yıllarda hangi konuların ağırlıklı çalışıldığının tespit edilmesi ise insan gücü ile gerçekleştirilmesi zor bir görevdir. Tüm bu durumlar göz önünde

bulundurulduğunda araştırmacılara katkı sağlamak ve yol gösterici olmak amacıyla 2007-2017 yılları arasında tıp alanındaki Türkiye'deki araştırmacılar tarafından oluşturulan literatür elde edilmiştir. Elde edilen literatür GDA'ya girdi olarak verilmiş ve sonucunda hangi yıllarda hangi konuların ağırlıklı çalışıldığı tespit edilmiştir. Model hem niceliksel hem de niteliksel olarak değerlendirilmiş ve GDA'nın hem otomatik hem de anlamsal analiz açısından başarılı olduğu gözlemlenmiştir.

Çalışmamız bu anlamda Türk tıp literatürü için gerçekleştirilen ilk çalışma olup, tıp alanındaki araştırmacılara fikir oluşturması açısından, oldukça önemlidir. Ayrıca, elde edilen konular üzerinden hangi hastalıkların ülkemizde 11 (on bir) yıl içerisinde sık görüldüğü ve önlem alma açısından hangi konuların daha çok göz önünde bulundurulması gerektiğinin tespiti açısından da çalışma oldukça değerlidir.

Teşekkür

Modelin niteliksel değerlendirilmesi ve konuların etiketlenmesi aşamasında bize her türlü yardımda bulunan Kocaeli Üniversitesi Tıp Fakültesi Göğüs Hastalıkları Bölümü'nden Doç. Dr. Haşim Boyacı Hocamıza teşekkürlerimizi sunarız.

Kaynakça

[1] Schwarz, C. 2018. ldagibbs: A command for topic modeling in Stata using latent Dirichlet allocation, The Stata Journal, Cilt. 18, s. 101-117. DOI: 10.1177/1536867X1801800107

[2] Sun, M., Zheng, H. 2018. Topic Detection for Post Bar Based on LDA Model. ss 136-149. Zhou, Q., Miao, Q., Wang, H., Xie, W., Wang, Y., Lu, Z., ed. 2018. Communications in Computer and Information Science, Springer Singapore, Singapore, 1396s.

[3] Shah, A.H. 2019. How episodic frames gave way to thematic frames over time: A topic modeling study of the Indian media's reporting of rape post the 2012 Delhi gang-rape, Poetics, Cilt. 72, s. 54-69. DOI: 10.1016/j.poetic.2018.12.001

[4] Karami, A., Ghasemi, M., Sen, S., Moraes, M.F., Shah, V. 2019. Exploring diseases and syndromes in neurology case reports from 1955 to 2017 with text mining, Computers in Biology and Medicine, Cilt. 109, s. 322-332. DOI: 10.1016/j.compbiomed.2019.04.008

[5] Blei, D.M., Ng, A.Y. 2003. Latent dirichlet allocation, The Journal of Machine Learning Research, Cilt. 3, s. 993-1022.

[6] Agrawal, A., Fu, W., Menzies, T. 2018. What is wrong with topic modeling? And how to fix it using search-based software engineering, Information and Software Technology, Cilt. 98, s. 74-88. DOI: 10.1016/j.infsof.2018.02.005

[7] Holzinger, A., Dehmer, M., Jurisica, I. 2014. Knowledge discovery and interactive data mining in bioinformatics: State-of-the-art, future challenges and research directions, BMC Bioinformatics, Cilt. 15, s. 1-9. DOI: 10.1186/1471-2105-15-S6-I1

[8] Lin, J.M., Bohland, J.V., Andrews, P., Burns, G.A.P.C., Allen, C.B., Mitra, P.P. 2008. An Analysis of the Abstracts Presented at the Annual Meetings of the Society for Neuroscience from 2001 to 2006, PLoS One, Cilt. 3, s. 1-9. DOI: 10.1371/journal.pone.0002052

[9] Bundschus, M., Tresp, V., Kriegel, H.P. 2009. Topic Models for Semantically Annotated Document Collections. NIPS Workshop: Applications for Topic Models: Text and Beyond, 7-10 Aralık, Whistler, 1-4.

[10] Jiang, Z., Zhou, X., Zhang, X., Chen, S. 2012. Using Link Topic Model to Analyze Traditional Chinese Medicine Clinical Symptom-Herb Regularities. 2012 IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom), 10-13 Ekim, Pekin, 15-18.

[11] Redfield, C.K., Lou, X., Karaletsos, T., Crosbie, C., Gardos, S., Artz, D., Ratsch, G. 2013. An empirical analysis of topic modeling for mining cancer clinical notes. 2013 IEEE 13th International Conference on Data Mining Workshops, 7-10 Aralık, Dallas, 56-63.

[12] Cui, M., Liang, Y., Li, Y., Guan, R. 2015. Exploring Trends of Cancer Research Based on Topic Model. 1st International Workshop on Semantic Technologies, 9-12 Mart, Changchun, 7-18.

[13] Beykikhoshk, A., Arandjelović, O., Venkatesh, S., Phung, D. 2015. Hierarchical Dirichlet Process for Tracking Complex Topical Structure Evolution and Its Application to Autism Research Literature. ss 550-562. Cao, T., Lim, E.P., Zhou, Z.H., Ho, T.B., Cheung, D., Motoda, H., ed. 2015. Advances in Knowledge Discovery and Data Mining, Springer Cham, Switzerland, 763s.

[14] Song, M., Heo, G.E., Lee, D. 2015. Identifying the landscape of Alzheimer's disease research with network and content analysis, Scientometrics, Cilt. 102, s. 905-927. DOI: 10.1007/s11192-014-1372-x

[15] van Altena, A.J., Moerland, P.D., Zwinderman, A.H., Olabarriaga, S.D. 2016. Understanding big data themes from scientific biomedical literature through topic modeling, Journal of Big Data, Cilt. 3, s. 1-21. DOI: 10.1186/s40537-016-0057-0

[16] Speier, W., Ong, M.K., Arnold, C.W. 2016. Using phrases and document metadata to improve topic modeling of clinical reports, Journal of Biomedical Informatics, Cilt. 61C, s. 260-266. DOI: 10.1016/j.jbi.2016.04.005

[17] Hahn, A., Mohanty, S.D., Manda, P. 2017. What's Hot and What's Not? - Exploring Trends in Bioinformatics Literature Using Topic Modeling and Keyword Analysis. ss 279-290. Cai, Z., Daescu, O., Li, M., ed. 2017. Bioinformatics Research and Applications, Springer Cham, Switzerland, 499s.

[18] Drosatos, G., Kavvadias, S.E., Kaldoudi, E. 2018. Topics and Trends Analysis in eHealth Literature. ss 563-566. Eskola, H., Väisänen, O., Viik, J., Hyttinen, J., ed. 2018. IFMBE Proceedings, Springer Singapore, Singapore, 1139s.

[19] Steyvers, M., Griffiths, T. 2007. Probabilistic Topic Models. ss 1-15. Landauer, T., McNamara, D.S., Dennis, S., Kintsch, W., ed. 2007. Handbook of Latent Semantic Analysis: A Road to Meaning, Psychology Press, USA, 544s.

[20] Lu, Y., Mei, Q., Zhai, C.X. 2011. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA, Information Retrieval, Cilt. 14, s. 178-203. DOI: 10.1007/s10791-010-9141-9

[21] Blei, D.M. 2012. Probabilistic Topic Models, Communucations of the ACM, Cilt. 55, s. 77-84. DOI: 10.1145/2133806.2133826

[22] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harsman, R. 1990. Indexing by Latent Semantic Analysis, Journal of the American Society for Information Science, Cilt. 41, s. 391-407. DOI: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

[23] Hoffman, T. 1999. Probabilistic Latent Semantic Analysis. Fifteenth Conference on Uncertainty in Artificial Intelligence, 20 Temmuz-1 Ağustos, Stockholm, 289-296.

[24] Popescul, A., Ungar, L., Pennock, D., Lawrence, S. 2001. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. 17th Conference in Uncertainty in Artificial Intelligence, 2-5 Ağustos, Washington, 437-444.

[25] Ng, K.W., Tian, G.L., Tang, M.L. 2011. Dirichlet and Related Distributions: Theory, Methods and Applications. Wiley, New York, 337s.

[26] Ekinci, E., İlhan Omurca, S. 2018. An Aspect-Sentiment Pair Extraction Approach Based on Latent Dirichlet Allocation for Turkish, International Journal of Intelligent Systems and Applications in Engineering, Cilt. 6, s. 209-213. DOI: 10.18201/ijisae.2018644779

[27] Ekinci, E., İlhan Omurca, S. 2017. Ürün Özelliklerinin Konu Modelleme Yöntemi ile Çıkarılması, Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, Cilt. 9, s. 51-58.

[28] Ekinci, E., İlhan Omurca, S. 2019. Concept-LDA: Incorporating Babelfy into LDA for aspect extraction, Journal of Information Science. DOI: 10.1177/0165551519845854

[29] Griffiths, T.L., Steyvers, M. 2004. Finding Scientific Topics, Proceedings of the National Academy of Sciences of the United States of America, Cilt. 101, s. 5228-5235. DOI: 10.1073/pnas.0307752101

[30] Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A. 2011. Optimizing semantic coherence in topic models. Conference on empirical methods in natural language processing, 27-31 Temmuz, Edinburgh, 262-272.

[31] Newman, D., Lau, J.H., Grieser, K., Baldwin, T., McCallum, A. 2010. Automatic evaluation of topic coherence. The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2-4 Haziran, California, 100-108.

[32] Atıcı, B., İlhan Omurca, S., Ekinci, E. 2017. Kullanıcı Şikayetlerindeki Ürün Özelliklerinin Gizli Dirichlet Ayırımı ile Saptanması. 2017 Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı, 5-8 Ekim, Antalya, 250-254.