Journal of the Faculty of Engineering and Architecture of Gazi University 36:4 (2021) 2183-2196

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Topic modeling with latent Dirichlet allocation for cancer disease posts

Volkan Altıntaş[1]*, Mehmet Albayrak[2], Kamil Topal[3]

[1]Graduate School of Natural and Applied Sciences Computer Engineering Department, Süleyman Demirel University, Isparta, 32040, Turkey
[2]Department of Computer Programming, Isparta University of Applied Science, Isparta, 32040, Turkey
[3]Department of Computer Engineering, Balıkesir University, Balıkesir, 10100, Turkey

Highlights:

- Examination of social media platform shares using Natural Language Processing Methods
- Examination of the data set with the help of Text Mining techniques
- Determining the sub-headings in the data set with the Topic Modelling methods

Keywords:

- Natural Language Processing
- Topic Modelling
- Text Mining
- Latent Dirichlet Allocation
- Social Media

Article Info:

Research Article
Received: 09.05.2020
Accepted: 14.04.2021

DOI:

10.17341/gazimmfd.734730

Correspondence:

Author:Volkan Altıntaş
e-mail: volkanaltintas@gmail.com
phone: +90 505 817 6087

Graphical/Tabular Abstract

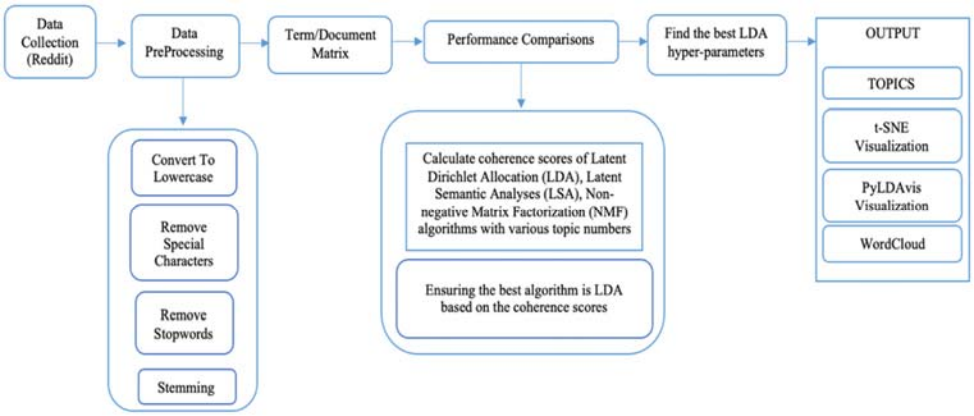


Figure A. System Architecture

Purpose: The aim of this paper is to reveal the main topics discussed by examining reddit user comments about cancer disease.

Theory and Methods:
After the preproccesing, user comments are divided into topics with the help of the latent dirichlet allocation method.

Results:
The proposed approach using LDA has created consistent and semantically meaningful topics and clusters from user shares. The obtained topics can not only help people to interpret the texts in a large sharing collection in a way that can be interpreted by human beings but can also help patients and doctors discover new content that may be neglected.

Conclusion:
The results obtained with the LDA algorithm consist of the diagnosis of cancer disease, treatment process, moral-motivation during the disease period, chemotherapy period and medical support.

Journal of the Faculty of Engineering and Architecture of Gazi University 36:4 (2021) 2183-2196

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Kanser hastalığı paylaşımları için Dirichlet ayrımı ile gizli konu modelleme

Volkan Altıntaş[1]*, Mehmet Albayrak[2], Kamil Topal[3]

[1]Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı, Isparta, 32040, Türkiye
[2]Isparta Uygulamalı Bilimler Üniversitesi, Uzaktan Eğitim Meslek Yüksekokulu, Bilgisayar Tekonolojileri Böl., Isparta, 32040, Türkiye
[3]Balıkesir Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Balıkesir, 10100, Türkiye

ÖNEÇIKANLAR

- Sosyal medya platform paylaşımlarının doğal dil işleme yöntemleri kullanılarak incelenmesi
- Metin madenciliği teknikleri yardımıyla elde edilen veri setinin incelenmesi
- Konu modelleme yöntemi ile veri seti içerisindeki alt başlıkların tespit edilmesi

Makale Bilgileri

Araştırma Makalesi
Geliş: 09.05.2020
Kabul: 14.04.2021

DOI:
10.17341/gazimmfd.734730

Anahtar Kelimeler:
Doğal dil işleme,
konu modelleme,
metin madenciliği,
gizli dirichlet ayrımı,
sosyal medya

ÖZ

Sosyal medya ortamlarında, kullanıcılar yaşadıkları olaylar ile ilgili edindikleri tecrübeleri paylaşmaktadır. Kişiler başlarından geçen bir olayı, yeni gördükleri bir şehri, okudukları kitabı vb. paylaşarak bu konular hakkında diğer kişilere deneyimlerini aktarmaktadır. Kullanıcıların konuştuğu konulardan biri de sağlık problemleri ve bu konudaki deneyimlerin paylaşılmasıdır. Sağlık problemi yaşayan bazı bireyler, geçirdikleri hastalıkları, gördüğü tedavileri ve sonuçlarını, her bir evresinde kazandıkları tecrübeleri sosyal ortamlarda yazarak paylaşabilmektedir. Bu paylaşımlar gerek bilgilendirici gerekse hastalıkla mücadelede moral/motivasyon için diğer hastalar açısından önem arz etmektedir. Paylaşım sayısının fazla olması, hastalıkların çeşitliği ve veri miktarının büyüklüğü nedeniyle insan tarafından manuel olarak yorumlanması imkânsız hale gelmektedir. Bu çalışmada, Reddit sosyal platformu üzerinden, kanser hastalığı ile ilgili paylaşımlar toplanarak bu veriler üzerinde çalışılmıştır. Bu paylaşımlar üzerinden yapay zekâ tabanlı konu modelleme algoritmalarından "Gizli Dirichlet Ayrımı (GDA)" algoritması ile konuşulan başlıca konu başlıkları bulunmuştur. Konu başlıklarının konuşulan konu ile ilişkisi incelenmiş ve içerik analizi yapılmıştır. Kanser hastalığı ile ilgili paylaşımlar içerisinde en fazla konuşulan içeriklerin belirlenmesi hedeflenmiştir. Ayrıca t-SNE tekniği kullanılarak konuların birbiri arasındaki ilişkisi incelenmiştir. GDA algoritması ile modelleme sonucunda elde edilen konu başlıklarında bulunan kelimelerin yapılan tutarlılık testinde uyumlu olduğu görülmüştür.

Topic modeling with latent Dirichlet allocation for cancer disease posts

HIGHLIGHTS

- Examination of social media platform shares using natural language processing methods
- Examination of the data set with the help of text mining techniques
- Determining the sub-headings in the data set with the topic modeling methods

Article Info

Research Article
Received: 09.05.2020
Accepted: 14.04.2021

DOI:
10.17341/gazimmfd.734730

Keywords:
Natural language processing,
topic modelling,
text mining,
latent Dirichlet allocation,
social media

ABSTRACT

In social media platforms, users share their experiences about the events they have experienced. People talk about a recent event, a city they have just seen, a book they read, etc. They post their experiences with other people about the same specific issues. One of the topics that users often talk about is health problems and sharing their experiences on this subject. Individuals with health problems can share their illnesses, treatments and results, and the experiences they have gained at each stage in social media platforms. These shares are important for other patients, both for informative and for morale / motivation in combating the disease. Manual analysis of the posts by human beings becomes impossible due to reasons such as the high number of posts, the variety of diseases and the amount of data. In this study, posts about cancer disease were collected on the Reddit social platform and these data were studied. The main topics discussed with the "Latent Dirichlet Allocation (LDA)" algorithm, one of the artificial intelligence-based topic modeling algorithms, were found through these posts. The relationship of the subject headings with the spoken subject was examined and content analysis was made. It is aimed to determine the most talked about contents among the posts about cancer disease. In addition, the relationship between the subjects was examined using the t-SNE technique. It was observed that the words in the topics obtained as a result of modeling with the LDA algorithm were compatible in the coherence test.

*Sorumlu Yazar/Yazarlar / Corresponding Author/Authors : *volkanaltintas@gmail.com, mehmetalbayrak@isparta.edu.tr, kamiltopal@balikesir.edu.tr / Tel: +90 505 817 6087

1. GİRİŞ (INTRODUCTION)

Dizüstü bilgisayar, cep telefonu, tablet vb. mobil cihazlar günlük hayatımızda çok sık kullanılmaktadır. Mobil cihazların kullanım oranındaki artış beraberinde internet kullanım oranlarında da artışı getirmektedir. Kullanıcıların interneti daha yoğun ve etkin şekilde kullanması ile ortaya çıkan veri miktarı üstel olarak katlanarak büyümektedir. Üretilen büyük veri; içerisinden anlamlı ve kişiye özel olarak kullanışlı bilgilerin ortaya çıkarılması amacıyla doğal dil işleme yöntemlerinden olan metin özetleme, konu modelleme, sınıflandırma, varlık ismi tanımlama, metin etiketleme, duygu analizi, bilgi geri getirimi gibi konulardaki çalışmalar ön plana çıkmaktadır. Son yıllarda yapısal olmayan veriler üzerinde doğal dil işleme çalışmalarının sayısının arttığı gözlemlenmektedir.

Yaşamımızın her alanına giren sosyal medya platformlarının kullanımı ve yeni teknolojiler ile kullanıcıların alışkanlıkları da değişmektedir. Kullanılan birçok elektronik cihaz barındırdığı sensörler ve yazılımlar yardımıyla kullanıcılar ve cihazları hakkında çok sayıda veri üretmektedir. Dünyadaki en büyük ağ olan, internet, aracılığı ile kullanıcıların yaptıkları paylaşımlar, yorumlar, alıp gönderilen e-postalar vb. büyük veri seti oluşturmaktadır. Sosyal medya platformlarında kullanıcılar tarafından paylaşılan bilgiler, metin, resim, konum, yapılan bir paylaşımın yeniden paylaşımı, haber paylaşımı ve paylaşımlara yapılan yorumlardan oluşur. Meta veriler çıkarıldığında, düz yazı içeren metinler büyük bir hacimde olup bu metinsel veriler yapısal olmayan doğal dil formundadır. Yapısal olmayan metin verilerinden anlamlı sonuçların otomatik olarak elde edilmesi metin analiz yöntemleri / algoritmaları ile yapılabilmektedir. Metin analiz yöntemlerinden birisi metin madenciliğidir. Metin madenciliği, doğal dil metninden anlamlı bilgiler elde etme işlemidir.

Metin madenciliğinin bir parçası olan doğal dil işleme (DDİ), bilgisayar ve insan (doğal) dilleri arasındaki etkileşimlerle ilgili bilgisayar bilimi ve yapay zekânın çalışma alanıdır. Metin madenciliği; metinlere ve konuşmaya makine öğrenmesi algoritmalarını uygulamak için kullanılmaktadır. DDİ dilbilimsel analiz için bilgisayarın insan dilini anlamasına yardımcı olmakta ve insan doğal dillerinin niteliğinden dolayı çözümü zor bir problemdir. Doğal dil ile aktarılan bilgilerin bilgisayarlar tarafından anlamasını beklemek kolay bir problem değildir. İnsan dilini kapsamlı bir şekilde anlamak için hem kelimelerin hem de kavramlarda iletilen mesajı anlamayı gerektirmektedir. DDİ ile, yapılandırılmamış doğal dil verilerinin bilgisayarların anlayabileceği bir formata dönüştürülmesi için doğal dil kurallarını tanımlama ve çıkarsamak için çeşitli algoritmalar kullanmaktadır. Bilgisayarlar, cümlenin anlamını tam ve doğru anlamlı halde beklenen doğruluk oranında yerine getiremediğinde istenilen sonuçlara ulaşmakta sorun olmaktadır. DDİ, Google Translate gibi çeviri uygulamalarında, kelime işlemci programlarında metin hatalarının düzeltilmesinde, çağrı merkezlerinde müşteri taleplerinin alınmasında, OK Google, Siri, Cortana ve Alexa gibi kişisel asistan uygulamalarında da yaygın olarak kullanılmaktadır. Günümüzde DDİ yardımıyla belge özetleme, varlık tanıma, makine çevirisi, konuşma tanıma, önemsiz e-posta algılama, konu modelleme gibi birçok uygulama alanlarında kullanılmaktadır.

Konu modelleme, veri topluluğu içerisinde var olan gizli başlıkları/konuları bulmak için kullanılan metin madenciliğinin istatistiksel yolu olarak tanımlanmaktadır. Metin madenciliği yöntemleri uygulanan konu modellemesi yaklaşımlarının avantaj ve dezavantajları Tablo 1'de gösterilmektedir. Ayrıca büyük metin koleksiyonlarındaki sözcük kümelerini bulmanın bilimsel bir yolu olarak tanımlanmaktadır [1]. Konu modelleme, duygu analizi, soru cevaplama, özetleme vb. doğal dil uygulamalarında sıklıkla kullanılmaktadır.

Tablo 1. Konu modellemesi tekniklerinin karşılaştırılması (Comparison of topic modeling techniques)

	Boyut Azaltma	Semantik Çıkarım	Karışık Modelleme	Genelleme Yeteneği
Gizli Anlam Analizi	✓	✓	✗	✗
Olasılıksal Gizli Anlam Analizi	✓	✓	✓	✗
Gizli Dirichlet Ayrımı	✓	✓	✓	✓

Kullanıcılar internet ortamında farklı platformlarda çeşitli olaylarla ilgili kişisel görüşlerini ve deneyimlerini paylaşmaktadır. Paylaşılan bu görüş ve deneyimler, aynı konu hakkında araştırma yapan kişiler için yol göstermektedir. Deneyimlerin paylaşıldığı alanların başında kişilerin yaşadığı sağlık sorunları ve sorunların üstesinden gelmek için uyguladıkları yöntem ve tedavi süreçleri gelmektedir. Kişiler yaşadıkları hastalıklar ile ilgili geçirdikleri evreleri detaylı olarak paylaşmaktadır. Aynı hastalık ile ilgili bilgi almak isteyen kişiler bu bilgilere ulaşarak daha detaylı bilgi alabilmekte ve farklı kişilerin tecrübelerinden faydalanmaktadırlar. Kendi durumlarını anlatıp gerektiğinde soru sorabilmektedirler. Bu platformlardan biri Reddit sosyal medya platformudur.

Reddit sosyal medya platformu dünyada yaygın olarak kullanılmaktadır. Reddit dünyadaki en çok ziyaret edilen 22., Amerika'da ise 6. web sitesi olarak öne çıkmaktadır [2]. Reddit metin ve resim paylaşımı için kullanılan forum tabanlı sosyal medya platformudur. Reddit konu başlıklarından ve her konuya ait alt başlıklardan oluşmaktadır. Platforma üye olan kişiler ilgili konu başlıklarında bulunan alt konu başlıklarına yorum yapabilir, yorumları okuyabilir veya farklı bir alt başlık açabilir.

Kullanıcılar resim, video, link ve makale bağlantıları gibi paylaşımlarda bulunabilmektedir. Platformun yaygın kullanılmasından ve belirtilen özelliklerinden dolayı farklı bölgelerde yaşayan kullanıcılar platform sayesinde bir araya gelmektedir. Reddit'te kullanıcı yorumlarının sayısı arttıkça faydalı yorumları bulmak zor olmaktadır. Kullanıcıların istenilen bilgiye ulaşması için uzun süreler ayırmaları gerekmektedir. Bu nedenle Reddit'te kullanıcı paylaşımları ile oluşan büyük verinin algoritmalar tarafından işlenip ayrıştırılmasına ihtiyaç duyulmaktadır.

Bu çalışmada; GDA [3] algoritmasını kullanarak Reddit platformu kullanıcıları tarafından paylaşılan kanser ile ilgili gönderilerin konu modelleme analizi yapılmıştır. "Kelime Torbası" mantığına dayanan GDA, gizli konuların karışımı olarak bir konunun kelimeler üzerinde çok terimli bir dağılıma sahip olduğu dokumanı temsil etmektedir. Her belgenin kendine özgü bir konu oranı vardır ve her konunun kendi kelime dağılımı vardır. Denetimsiz Bayesian [4] öğrenme algoritmasına dayanan GDA, yapılandırılmamış ve kullanıcı paylaşımları ile oluşan büyük veriyi temsil eden gizli konuları ortaya çıkarabilir.

Bu çalışma ile kanser hastalığı üzerine yapılan paylaşımlar için GDA ile analiz çalışması yapan literatürdeki ilk çalışma olması nedeni ile önemlidir. Çalışma tamamen çevrimiçi platform üzerinden yapılan kullanıcı paylaşımlarına dayalıdır. Kanser hastalığı ile ilgili yapılan paylaşımların konu modellemesi ve özellik çıkarımı yapılmıştır. Reddit'teki büyük veri hacmi ve konu çeşitliliği göz önüne alındığında, doktorlar/kullanıcılar için platformdan istenilen biçimde uygulanabilir bilgiler elde etmek zor bir iştir. Doktorların hastalık tespiti, tedavi süreci, hastaların hastalığa karşı bakışı, hastaların yaşadığı deneyimler vb. konular hakkında bilgi edinerek bir öngörü sağlaması basit hale gelmektedir. Ayrıca kanser hastalarına ışık tutacak, yol gösterecek hastalık ile ilgili bilgilerin ön plana çıkartılarak yardımcı olmakta amaçlanmaktadır. Çalışma, kanser hastalığı ile ilgili yorumların toplanması sonucu ortaya çıkarılan konuların açıklayıcı bir analizini de sağlamaktadır. Bununla birlikte, sonraki süreçlerde kanser hastalığı ile ilgili tedavi, teşhis vb. konuların gösterimini ortaya çıkarmak için öngörücü bir model üretme yeteneği ortaya koyarak metin madenciliği analizini daha güçlü kılmaktadır.

Literatürde konu modelleme üzerine birçok çalışma bulunmaktadır. Reddit platformundan Okon vd., 2005 ve 2017 yılları arasında paylaşılan dermatoloji ile ilgili 176K adet yorumu toplamıştır. Elde edilen yorumlara GDA algoritmasını uygulanmış ve dermatoloji başlığının alt başlıklarında hangi konular üzerinde paylaşımlar yapıldığını kullanıcıların hangi konular üzerinde konuştuğu bulunmuştur. Bu konular içinde tutarlı temalar ve terimlerin sıklığını oluşturmak için spektral kümelemeyi kullanmıştır. Reddit platformunun dermatoloji ile ilgili çeşitli konularda kullanıcılar tarafından oluşturulan veriler ile etkili bir veri kaynağı olduğu, doktorlar tarafından düzenli çalışmalar ile belirlenemeyecek eğilimlerin belirlendiği örneğin hastaların sıklıkla tartıştığı Homeopatik yöntemlerin yönetimsiz GDA modeli ile belirlendiği sonuçlarına ulaşılmıştır [5]. Ding vd., yaptıkları çalışmada diyabet hastalığı üzerine yeni bir GDA modeli önermektedir. Önerilen "seLDA" modeli, ile diyabetik komplikasyon tahmini yapılmıştır. Verilerin ön işlenmesinden sonra ilk olarak tıbbi kayıtlar arasındaki benzerliği tahmin edilmiştir. Daha sonra benzerlik kısıtlamalarına dayanarak seLDA modeli ile diyabetik komplikasyon konu madenciliği gerçekleştirilmiştir. Son olarak, destek vektör makineleri ile çoklu etiket sınıflandırma problemini çözerek bir tahmin modeli oluşturulmuştur. Deneysel sonuçlar, geleneksel GDA temelli yaklaşımlara göre %22,49 daha iyi bir performans sergilediğini göstermektedir [6]. Chen ve Ren yaptıkları çalışmada farklı forumlarda yapılan paylaşımları incelemişlerdir. Yapılan paylaşımların başlıkla ilgili olup olmadığı konusunda "Forum-LDA" modelini geliştirmişlerdir. Geliştirilen Forum-LDA modeli ana konu bilgilerini ve bunlara verilen yanıtların ilişkilerini inceleyerek yalnızca alakasız yanıt gönderilerini değil aynı zamanda ciddi ve bilinmeyen kullanıcıları da ortaya çıkarmaktadır [7]. Bastani vd., Amerika'da tüketici hakları bürosuna yapılan şikayetler ile ilgili konu modelleme çalışması için GDA algoritması kullanmıştır. Şikayetlerin otomatik olarak analiz edilmesi ve uzmanlara doğru bilgi aktarımının sağlanması amaçlanmıştır. Önerilen yaklaşım, tüketici bürolarına yapılan şikayetlerde gizli konuların çıkarılmasını ve zamanla ilişkili değişimleri araştırmaktadır. GDA algoritması kullanılarak önerilen yaklaşım, tüketici şikayetlerinden tutarlı ve semantik olarak anlamlı konular/kümeler oluşturduğu gözlemlenmiştir. Bu konular, geniş bir şikayet korpusunda bulunan metinlerin insan tarafından yorumlanabilir bir şekilde ayrıştırmasına değil, ayrıca uygulayıcıların tüketici hakları bürosunda çalışmakta olan kişiler tarafından gözden kaçırılmış olabilecek yeni içerikleri bulmalarına olanak sağlamaktadır [8]. Wang vd., iki rakip firmanın benzer ürünleri için müşterilerin yaptığı yorumlar üzerine GDA algoritması kullanmıştır. GDA algoritması yardımıyla iki ürün içinde konuşulan başlıca konular ortaya çıkarılmıştır. Sonuçlar iki ürünün benzersiz konularını göstermekte ve her iki ürünün de rekabetçi üstünlüklerini ve zayıflıklarını belirlemektedir. Bu çalışma, iki farklı markaya ait bilgisayar faresi ve yağ markası olmak üzere rakip ürünlerin analizinde uygulanmıştır. İki farenin ve yağ markasının benzersiz konuları, rekabet gücü ve zayıf yönleri belirlenmiştir [9]. Hagen, e-Dilekçe içeriklerinin analizi için GDA algoritmasını kullanmıştır. GDA algoritmasından elde edilen sonuçların %87 oranında insanların yorumladığı sonuçlar ile benzer olduğu sonucunu elde etmiştir [10]. Akademik yayınlar [11], mikroblog paylaşımları [12], Twitter paylaşımları [13] ve çevrimiçi forum siteleri [14] gibi farklı doküman koleksiyonları üzerinde GDA algoritması kullanılmıştır. Güven vd. konu modelleme algoritmalarından GDA, GAA ve Olasılıksal-Gizli Anlamsal Analiz (O-GAA) algoritmalarını Türkçe "tweet" lerden kişilerin duygularını ortaya çıkarmak için kullanmıştır. Ayrıca, mevcut yöntemler ile GDA algoritmasının geliştirilen farklı çeşitlerinin duygu analizindeki başarısı kıyaslanmıştır. Geliştirilen n-aşamalı GDA modelinin, GDA ve O-GAA'ya göre performans ve

süre olarak üstün olduğu tespit edilmiştir. GAA'nın ise en hızlı ve başarılı yöntem olduğu tespit edilmiştir[15].

2. TEORİK METOD (THEORETICAL METHOD)

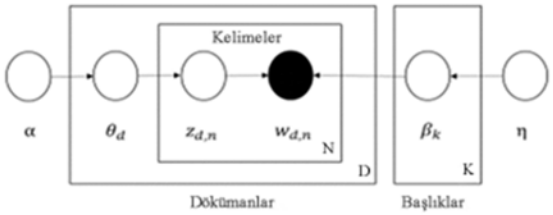
Bu bölümde kullanılan metot, veri toplama, veri ön işleme ve konu modelleme uygulaması hakkında bilgi verilmektedir.

2.1. Gizli Dirichlet Ayrımı (Latent Dirichlet Allocation)

Makine öğrenmesi ve metin madenciliği uygulamalarında tercih edilen önemli çalışma alanlarından biri doküman koleksiyonlarındaki gizli tematik bilgiyi daha küçük boyutlu yapılara dönüştürerek ortaya çıkaran algoritmalar olarak bilinen konu modelleme yöntemleridir [16]. Konu modelleme, etiketsiz halde bulunan dokümanların işlenerek gizli konuları ortaya çıkarılma çalışmasıdır. Konu oluşumunun dokümanda bulunan kelimelerin kullanımı ile ilişkili olduğu varsayımına göre hareket eder. Literatürde konu modelleme yöntemleri üzerine çalışmalar olmasına rağmen, teorik olarak anlaşılması güç bir konu olarak karşılaşılmaktadır.

Konu modelleme çalışma mantığı Şekil 1'de verilmiştir. Doküman okunurken karşılaşılan farklı temalara işaret eden anahtar kelimeler, vurgulamak için farklı renkler ile işaretlenmektedir. Makalenin her bir anahtar kelimesini okuduktan ve her bir anahtar kelimeyi vurgulayıcıyla renklendirdikten sonra, ortak bir renkle renklendirilmiş anahtar kelimeler toplanırsa, ortak bir renkle renklendirilmiş her bir anahtar kelime grubunun bir temayı temsil ettiği kabul edilebilir. Farklı renk sayılarının toplamı dokümanda yer alan konu sayısını vermektedir [1].

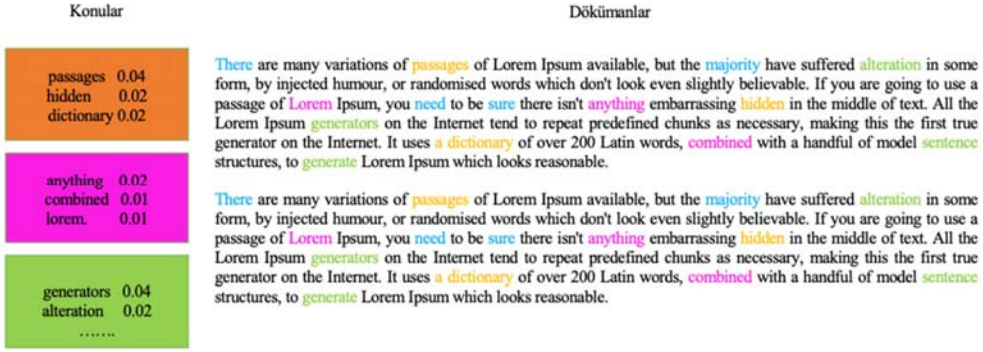
Konu modelleme için GDA, "Gizli Anlam Analizi (GAA)" [18] ve "Negatif Olmayan Matris Ayrıştırma (NMA)" [19] yöntemleri ön plana çıkmaktadır [20]. Konu modelleme yöntemleri arasında GDA yaygın bir şekilde kullanılmaktadır. GDA, metinsel belge koleksiyonunda bulunan gizli anlamsal yapıları ortaya çıkarmak için olasılıksal bir modelleme yaklaşımı olarak Blei vd. [3] tarafından geliştirilmiştir. Her bir doküman, her bir konunun doküman koleksiyonundaki benzersiz kelimeler üzerinde dağıtılmasıyla karakterize edilen gizli konuların bir karışımını gösterir. GDA'nın düzlemsel gösterimi (plate notasyonu) Şekil 2'de gösterilmektedir.



Şekil 2. Gizli Dirichlet ayrımı düzlemsel gösterimi (Latent Dirichlet allocation plate notation)

Dokümanda bulunan kelimeler $W_{d,n}$ ile gösterilmektedir. Gösterimdeki n kelimeyi ($\forall n$=1, …, N), d($\forall$d=1, …, N) ise dokümanı temsil etmektedir. β_k konu başlıklarının içeriğini, θ_d ise her dokümanın bulunan konu başlıklarına dağılımını göstermektedir. $Z_{d,n}$ ise her kelimenin bağlı olduğu konu başlığını temsil etmektedir. β_k, θ_d, $Z_{d,n}$ önceden bilinmeyen, $W_{d,n}$ bilinen değişkeninin işlenmesi sonucu elde edilen değerleri temsil etmektedir. η ve α parametreleri ise β_k ve θ_d parametrelerinin önceki değerlerini temsil etmektedir. Doğrusal gösterimde K kutusu konu başlıklarının sayısını, D kutusu doküman sayısını, N kutusu ise dokümanlarda yer alan benzersiz toplam kelime sayısını belirtmektedir.

Doğrusal gösterim tekrarlayan yapıları yani aynı tipte birden fazla nesnenin olduğu durumları ifade etmek için kullanılmaktadır. GDA için doğrusal gösterim ise gözlemlenen verinin rastgele değişkenler ve bu değişkenlerin yönlü kenarlar boyunca yayılımı üzerinden nasıl üretildiğini açıklamaktadır. Konu modellemedeki asıl amaç, doküman koleksiyonundan konuların çıkartılmasıdır. Bu işlem adımında sadece dokümanlar gözlenebilir durumda olup; kelimelerin konulara atanması, konuların dokümandaki ve kelimelerin konulardaki dağılımları gizlidir. Bu nedenle Şekil 2'de gözlemlenen değişkenler siyah renkle, gözlenemeyenler beyaz renk ile temsil edilmiştir. Tamamen denetimsiz bir yöntem olan GDA herhangi bir önbilgiye gerek kalmadan, kelime torbası yaklaşımına dayalı olarak çalışmaktadır. Kelimelerin doküman içerisindeki yerleşimi dikkate alınmazken,



Şekil 1. Olasılıksal konu modelleme gösterimi [17] (Probabilistic topic modeling representation)

kelimelerin birlikte bulunması bu yöntemde kullanılmaktadır.

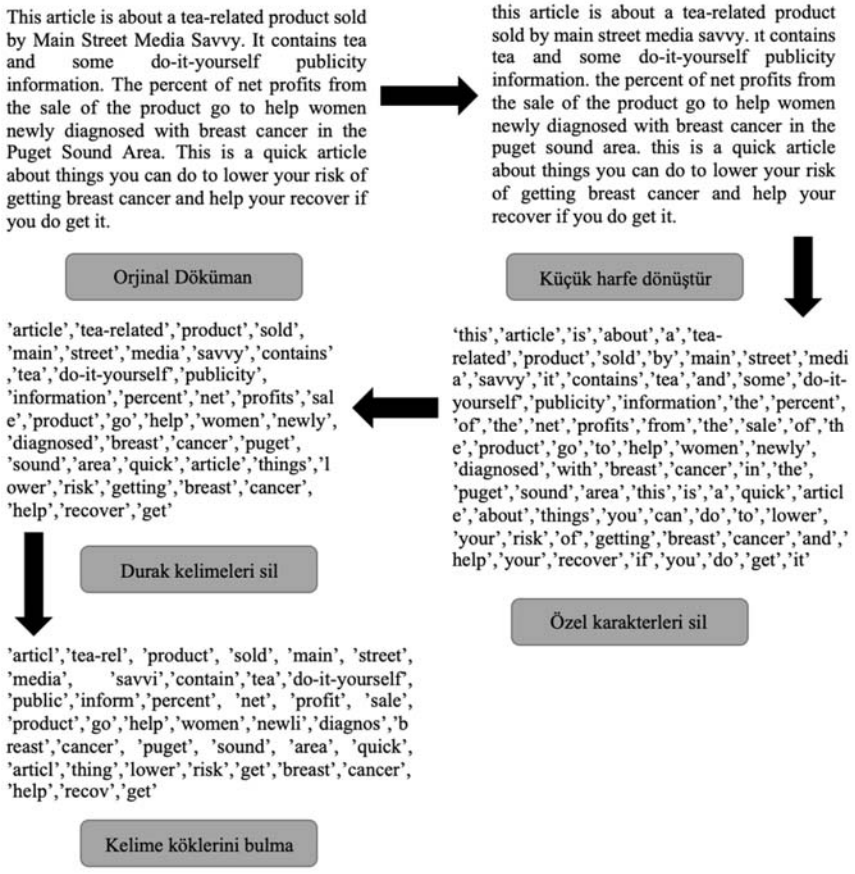
Gizli konuları bulma uygulamalarında hiper-parametre olarak α ve β değerleri önceden verilir. Bu parametrelerin alacağı değerler her bir konunun dağılımını ve yoğunluklarını belirler. α doküman-alt başlık yoğunluğunu belirlerken, β ise kelime-alt başlık yoğunluğunu belirler. Başka bir deyişle, yüksek α değeri fazla sayıda başlık bulabilirken, düşük α değeri daha az alt başlık bulur. Benzer şekilde, yüksek β değeri, bir başlığın fazla sayıda kelimelerden oluştuğunu gösterirken, düşük β değeri bir başlığın daha az kelimelerden oluştuğunu gösterir. Bu parametreler simetrik veya asimetrik dağılımla giriş değerleri verilebilir. Yapılan çalışmalara göre asimetrik α ve simetrik β parametreler daha iyi sonuç vermektedir [21].

2.2. Veri Toplama (Data Collection)

Yapılan çalışmada, kanser hastalığı ile ilgili paylaşımlar dünyanın en çok kullanılan 22., Amerika'nın ise 6. [2] sosyal paylaşım platformu olan Reddit [22] internet sitesinden elde edilmiştir. Reddit çevrimiçi bir platform olup yapılan yorumlar siteye giriş yapan her kullanıcı tarafından okunabilmektedir. Veriler Reddit platformu tarafından sunulan "PRAW" [23] kütüphanesi kullanılarak alınmıştır. Reddit sosyal platformundan 28,8K üyesi bulunan kanser "subReddit"'inde bulunan 109.243 yorum çekilerek *.txt uzantılı dosyada kayıt altına alınmıştır.

2.3. Veri Önişleme (Data Preprocessing)

Reddit platformundan elde edilen yorumlar DDİ teknikleri ile analiz edilmiştir. Öncelikle *.txt uzantılı dosyada saklanan veriler ön işlemeye tabii tutulmuştur. Bu adımlar sonucu elde edilen veriler Şekil 3'te gösterilmektedir. Birinci adım olarak bütün yorumların tutarlı bir şekilde olması için bütün harfler küçük harfe dönüştürülmüştür. Sonraki adımda metin madenciliği işleminde herhangi bir önemi bulunmayan noktalama işaretleri (ör.!% $ # & ∗?, / .; "\) ve sayıları içeren özel karakterler kaldırılmış, belge sadece terimler haline getirilmiştir. 3. adımda ise her dilde yaygın olarak kullanılan durak kelimeleri olarak adlandırılan kelimeler (a, an, the, this vb..), önemli kelimelere daha çok odaklanabilmek için kaldırılmıştır. Durak kelimelerin kaldırılması için "NLTK" [24] kütüphanesinde İngilizce dili için bulunan kelimeler listesi kullanılmıştır. Metin çeşitliliğini azaltmak ve kelimelerin köklerini bulmak için kök bulma (stemming) işlemi yapılmıştır. Kök bulma işlemi için NLTK kütüphanesi içinde bulunan "Snowball" kütüphanesi kullanılmıştır. Kök bulma adımı, metin madenciliği algoritmalarında karışıklığa yol açabilecek kelimelerin varyasyonları arasında ayrım



Şekil 3. Veri önişleme gösterimi (Data preprocessing representation)

yapmak yerine, analizin kelimelerin kök halinde düzenlemesine yardımcı olduğundan, metin madenciliği analizinde çok yaygındır. Veri ön işlemede son yapılan işlem doküman matrisi terimini oluşturmaktır. Bu matris dokümanlarda bulunan terimlerin sıklığını belirtmektedir. Doküman matrisi GDA algoritmasında girdi bilgisi olarak kullanılmaktadır.

2.4. Özellik Çıkarımı (Feature Extraction)

Ön işlemden geçen Reddit yorumlarının kelimeleri, sayısal dönüşüme hazırdır. Her bir kelime özniteliğine göre uygun ağırlıklar hesaplanarak, öğrenme algoritmalarına göndermeye hazır hale gelmelidirler. Bu işlem için en sık kullanılan yöntemlerden bir tanesi çanta modeli yaklaşımıdır. Her bir kelimenin kullanım sıklığı ve karşıt dokümanlarda geçme sıklığına bakılarak hesaplanan TF-IDF (Terim Sıklığı - Ters Doküman Sıklığı) vektör dönüşümü, her bir kelimenin içinde bulunduğu dokümanda ne kadar etkili olduğunu hesaplatır [25]. Öncelikle her bir ifadenin bir dokümandaki yer alma sıklığı Eş. 1 hesaplanır.

$$TF(w,d) = w \tag{1}$$

Daha sonra da her bir kelimenin bütün derlemdeki dokümanlar içindeki yer alma sıklığı Eş. 2 ile hesaplanır.

$$IDF(w) = \log\left(\frac{|D|}{DF(w)}\right) \tag{2}$$

bu denklemde w her bir kelimeyi, $|D|$ toplam doküman sayısını ve $DF(w)$ ise w kelimesinin tüm dokümanlarda en az bir kere yer alma sayısını verir. Yukarıdaki denklemde payda kısmında yer alan $DF(w)$ ifadesi ne kadar fazla olursa, o kelimenin IDF değeri düşer. Eğer, bir kelime ne kadar çok dokümanda yer alıyorsa, o kelimenin IDF değeri 0'a yakın olur. Son olarak her bir dokümanın TF-IDF skoru Eş. 3'te bulunan denklemle hesaplanır.

$$d^{(i)} = TF(w_i, d).IDF(w_i) \tag{3}$$

$d^{(i)}$ d dokümanında yer alan w_i kelimesinin (i. kelime) TF-IDF skorudur. Her bir d vektörü bir dokümanın vektöre çevrilmiş halini temsil eder. Böylece tüm dokümanların (Reddit yorumlarının) vektörlere çevrilmiş hali olan $m \times n$ boyutlarındaki (m doküman sayısı, n farklı kelime sayısı) D matrisi elde edilmiş olur.

2.5. GDA ve Diğer Konu Modelleme Algoritmaları ile Karşılaştırma
(Comparison with other Topic Modeling Algorithms.)

Bu bölümde Gizli Dirichlet Analizi algoritmasını karşılaştıracağı iki farklı denetimsiz algoritmadan bahsedilmiştir.

2.5.1. Negatif olmayan matris ayrıştırma
(Nonnegative matrix factorization)

Negatif olmayan Matris Ayrıştırma, negatif olmayan sayılardan oluşan bir matrisi iki negatif olmayan matrise ayrıştırır [26]. Bu özellik matrislerin yorumlanmasını kolaylaştırır. Eş. 4'te bu ayrışma gösterilmektedir.

$$D \approx WH \tag{4}$$

D matrisi $m \times n$ boyutunda, satırlarda her bir yorum, sütunlarda ise her bir kelime değerleri yer alıyordu. Ayrıştırma için verilen k parametresi, konu sayısını temsil eder. Ayrıştırma sonucu elde edilen W matrisi $m \times k$ ve H matrisi $k \times n$ boyutlarındadır. W matrisi konuların dokümanlarla olan ilişkisini, H matrisi de kelimelerin k adet konuyla olan ilişkilerini verir. Sayısal değerinin yüksek veya düşük olması, o kelimenin ilgili konuyla alakalı ilişkisini ortaya koyar.

2.5.2. Gizli anlam analizi (Latent semantic indexing)

Gizli Anlam Analizi için Tekil Değer Ayrıştırma işlemini Eş. 5'te gösterildiği gibi yapılması gerekmektedir [27].

$$D = U\Sigma V^T \tag{5}$$

D matrisi $m \times n$ boyutundadır. U matrisi $m \times m$; Σ matrisi $m \times n$ ve V^T matrisi $n \times n$ boyutları olacak şekilde bulunur. Σ matrisinin köşegeninde D'nin özdeğerleri yer alır. U matrisi dokümanlarla ilgili bilgileri verirken; V^T ise kelimeler ve dokümanlarla ilgili ilişkileri gösterir. Burada belirlenecek k konu sayısı ile U matrisinin ilk k sütunu alınır ve $m \times k$ boyutuna; V^T matrisinin ilk k satırı alınır ve $k \times n$ boyutlarına indirgenir. Böylece tüm k konu için kelime skorları V^T matrisinden elde edilmiş olunur.

2.6. Ölçme Teknikleri - Tutarlılık Puanı (Measurement Techniques-Coherence Measure)

Denetimsiz öğrenme yöntemlerinde, önceden verilmiş bir eğitim seti olmadığı için, sonuçlar genellikle birbirine benzer verileri kümeleyerek verilir. Bu kümeye ait özellikleri yorumlayacak uzman kişiler konunun ne olduğunu anlamaları gerekmektedir. Dahası, konu modelleme algoritmalarında veri setinde kaç farklı konu olabileceği bilgisi verilmesi gerekmektedir. Bu parametrenin ne kadar doğru olduğunu ölçmek için kullanılacak yöntemlerden en önemlilerinden bir tanesi tutarlılık puanıdır [28]. Bu çalışmada her bir konunun tutarlılık puanlarını hesaplamak için önce, belirli bir pencere aralığında, bir konunun en önemli n kelimesinin birlikteliklerinin olasılıkları hesaplanmıştır. Daha sonra bu olasılıkların Normalize Noktasal Karşılıklı Bilgi Katsayısı (Normalized Pointwise Mutual Information NPMI) skorları bulunmuştur. En önemli n kelimelerin birbirlerine olan uzaklıkları kosinüs benzerliği ile hesaplanıp, bu benzerliklerin aritmetik ortalaması her bir modelin tutarlılık katsayısını vermiştir. 0 - 1 aralığında olacak tutarlılık puanı; yüksek puanlarda daha tutarlı, düşük puanlarda ise tutarsız konuları belirlemiş bir modelin genel performansını ortaya koyar.

3. ARAŞTIRMA BULGULARI
(RESULTS RESEARCH FINDINGS)

3.1. Metotların Performans Değerlerinin Karşılaştırılması
(Comparison of Performance Values of Methods)

GDA, 2. Bölümde yer alan ve açıklanan metotlarla kıyaslanmıştır. Sonuçlar Şekil 4'te gösterilmektedir. Sonuçlara göre GDA algoritmasının, GAA ve NMA algoritmalarına göre 5 başlık için en iyi tutarlılık puanına sahip olduğu görülmektedir. Bu yüzden GDA bu çalışmanın geri kalanında incelenmiştir.

Şekil 4'te görüldüğü gibi k sayısı 2 ile 7 arasında değiştirilerek 3 farklı algoritmada denenmiştir. Çıkan sonuçlara göre GAA algoritması 2 ve 4 başlık için iyi çıkmıştır. NMA algoritması ise konu sayısı arttıkça iyi olmasına rağmen tutarlılık skoru olarak GDA algoritması için 5 başlık değeri en iyi sonuç olarak elde edilmiştir.

3.2. GDA İçin Farklı Parametrelerin Belirlenmesi
(Determination of Different Parameters of GDA)

Tablo 2'de konu sayısı 5 seçildikten sonra ($k = 5$) farklı α ve β değerleri için tutarlılık skorları hesaplanmıştır. Elde edilen tutarlılık değerlerine göre en iyi olan 5 tutarlılık skoruna ait parametreler gösterilmiştir. α ve β değerleri için 0 ile 1 arasında farklı değerlere göre elde edilen en iyi sonuç α için asimetrik, β için 0,61 değeri olmuştur. Bu sonuçlarda "yapılan çalışmalara göre asimetrik α ve simetrik β parametreler daha iyi sonuç vermektedir [28]" yorumunu destekler bir sonuç ortaya koymuştur.

3.3. GDA Sonuçlarının İncelenmesi
(Examination of GDA Results)

GDA algoritmasının çıktı olarak verdiği bilgiler dokümanın içerdiği konu başlıkları, her dokümanın konu başlıklarına katkısı ve başlıkların dağılım oranlarıdır. Konu başlıkları Tablo 3'de gösterilmektedir. Konuların içerdiği kelimelerden anlamlı etiketler çıkarılmaktadır. GDA algoritmasının denetimsiz olarak çalışmasından dolayı konuların otomatik olarak bulunması mümkün değildir. GDA algoritmasından elde edilen başlıkların içerdiği kelimeleri yorumlamak için dışarıdan bir etki gerekmektedir. GDA algoritmasından çıkarılan konulardaki kelimeler yorumlanarak Tablo 3'de etiket sütununda gösterilmektedir. Kelimeler ilgili konu başlığındaki ağırlıklarına göre sıralanmaktadır. Kelimelerin ağırlıklarına göre yorumlanarak etiket bilgisi oluşturulmuştur

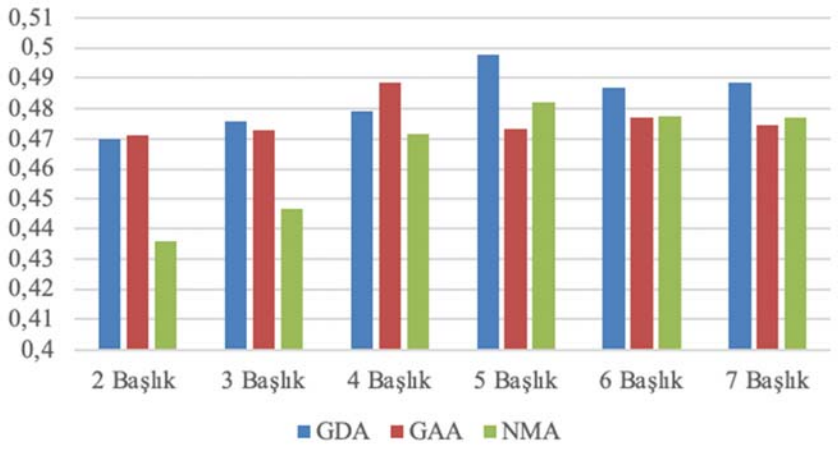
Tablo 2. $k = 5$ değeri için GDA algoritmasının farklı parametreler ile tutarlılık skorları
(Coherence scores of the GDA algorithm with different parameters)

Alfa	Beta	Tutarlılık Skoru
asym	0,61	0,538
asym	0,91	0,520
0,61	0,91	0,504
0,01	0,01	0,501
asym	0,01	0,498

Tablo 3. GDA algoritmasından elde edilen kelimeler ve çıkarılan başlıklar
(Extracted words and extracted titles from LDA algorithm)

No	Başlık	Etiket
1	feel, know, go, like, thing, help, peopl, get, cancer, want,	Diagnose of Disease
2	time, sorri, year, go, cancer, day, love, dad, mom, life	Togetherness
3	cancer, chemo, year, surgeri, stage, treatment, month, tumor, get, radiat	Chemotherapy
4	chemo,get,day,like,help,eat,week, take, pain, time	Nutrition
5	cancer, treatment, doctor, get, patient, help, work, medic, may, oncologist	Medical Aid

Bu çalışmada 109.243 kullanıcı yorumu analiz edilmiştir. Bundan dolayı bütün yorumların konu dağılımlarını göstermek imkansızdır. Her bir konu alt başlığın oluşması esnasında birçok yorum etkin rol oynamaktadır. Bazı yorumlar birden çok başlığa uymaktadır. Her bir başlığın



Şekil 4. Konu modelleme metotlarının karşılaştırılması (Comparison of topic modeling methods)

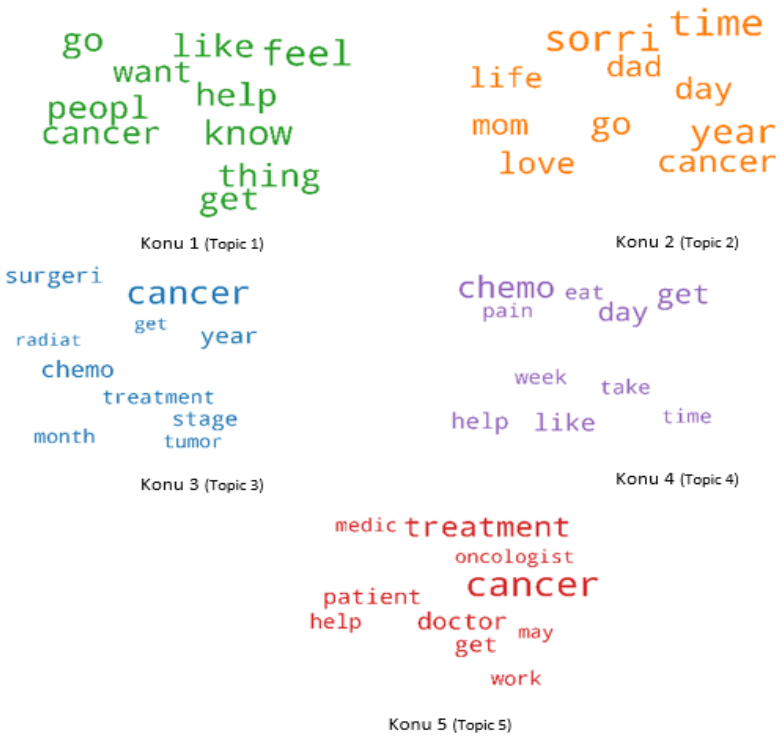
belirlenmesinde en fazla orana sahip olan yorumlar ve başlıklar Tablo 4'te gösterilmektedir. Tablo 4'te görüldüğü gibi her yorum %99 ve üstü oranlarda ilgili başlıkların oluşmasında etkin rol oynamıştır. Tablo kısıtlamalarından dolayı yorumların belli bir kısmı gösterilmektedir.

Kelime bulutları, birçok sosyal ağ analitik aracında güçlü bir görselleştirme aracı haline gelmiştir. Genellikle blog yazılarında, belgelerde, sosyal medya konuşmalarında tartışmanın "başlıkları" olarak adlandırılan kelimeleri göstermek için kullanılır. Kelime bulutları, altta yatan tartışma konularını gizleme eğilimindedir ve yalnızca en sık kullanılan sözcüklerin yüzeye çıkmasına izin verir. Şekil 5'de her bir konu başlığını oluşturan kelimeler için hazırlanmış olan kelime bulutu gösterilmektedir. Hangi başlıkta hangi kelimenin daha ağırlıklı olduğunu kelimenin bulut içerisindeki boyutu belirlemektedir. Her konu başlığı farklı bir renkte gösterilmektedir. Konu başlıklarında baskın bulunan kelimeler daha büyük boyutta gösterilmektedir. Kelimelerin ilgili konu başlığı üzerinde ağırlıkları azaldıkça kelime bulutu üzerinde gösterilen boyutları azalmaktadır. "t-SNE" yöntemi boyutun azaltılması için doğrusal olmayan bir

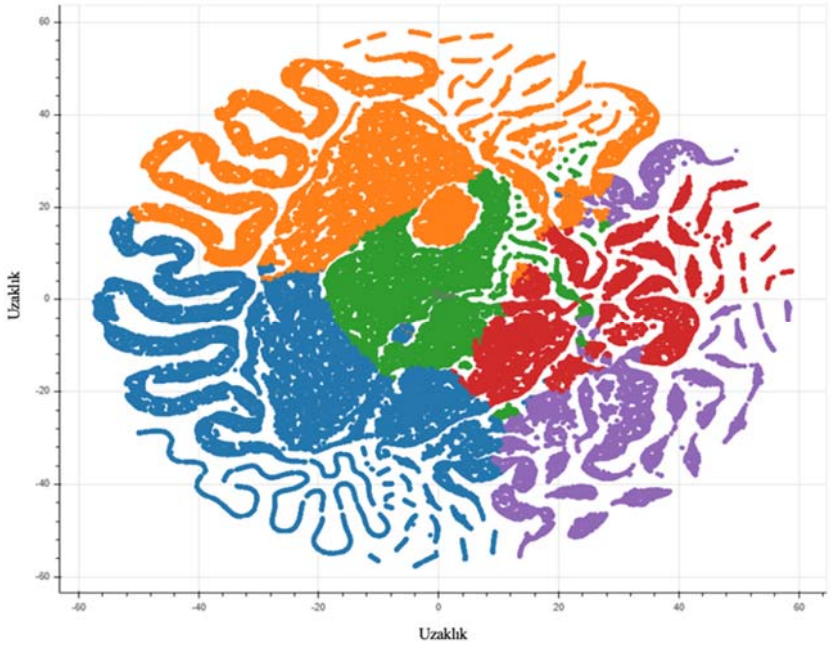
Tablo 4. Başlıklarda en fazla ağırlığa sahip yorumlar (Top weight reviews in titles)

No	Konu Başlığına Etkisi	Etiket	Yorum
1	0,9902	Diagnose of Disease	For me it started a year ago when I found some nodes on the side of my neck. After 3 months I went to my doctor, but he told me not to worry and didn't do any test. I went back around 2 months later because the nodes had grown but he told me again not to worry and only did a blood test. Later in April, I went to my doctor for the third time because some nodes appeared somewhere else and the we're growing and hurting. …
2	0,9937	Togetherness	It's so hard. I don't know exactly what you're feeling because it was my mom and not a partner but I'm sending you internet hugs and you'll be in my thoughts. Take care of yourself too right now. I could only imagine the pain your going through be strong for her. Wishing both you and her peace in this time. Fuck cancer.This is very hard and I am so sorry. What a wonderful thing you've done being with her through it all. I am sending you all of my love, you always have someone here to be with you.
3	0,9933	Chemotherapy	I can relate to this. I've told two friends that my cancer is terminal, and they've been sort of able to handle the news but I dread the thought of my larger friend group finding out because all our bar meet-ups and parties would just stop. Because that shit is awkward. And I enjoy those meet-ups and I don't want to lose them just yet. It really sucks seeing people who know the full extent of how bad it is (or just seeing people who know about the cancer) because you're immediately aware of their awkwardness and/or their concern and you suddenly feel like you have to manage their goddamn feelings on top of everything else you're going through.
4	0,9904	Nutrition	As for alternative treatments...facts are facts. And the facts show that no known alternative medicine can cure cancer. You will find anecdotal "evidence" of people being "cured" by alternative medicine, but if you do the research behind these claims, you will not find scientific evidence or any long term studies done that have found any alternative medicine effective enough to successfully treat cancer. For example, my aunt was pushing a "proven" cancer treatment on my mother.
5	0,9956	Medical Aid	-foot and lower leg massages for fluid retentionComfy PJs or other comfort items like a soft blanket, soft socks, or nice pillow for lounging on the sofa. My mouth got really dry so I kept a small humidifier by my bed at night. I also got some biotene gel for dry mouth. I stocked up on unscented lotion for dry skin, unscented/baby body wash, and hand sanitizer for when I was out and about and needed to wash my hands without easy sink access. Face masks if his immune system will be compromised. Aquaphor was a godsend during radiation.Sleep is of utmost importance. A quiet private place would be appreciated and help him recover from chemo. He may get diarrhea so the softest, gentlest TP is essential.

teknik olup özellikle yüksek boyutlu veri kümelerinin görselleştirilmesi için çok uygundur [29]. Görüntü işleme, DDİ, genomik veriler ve konuşma işlemede yaygın olarak uygulanır. t-SNE algoritması kelimelerin yakınlıklarını belirlemek için olasılık dağılımları üzerinden hareket etmektedir. 109243 adet yorumu 5 konu başlığına t-SNE algoritması yardımıyla ayırdığımız zaman elde edilen grafik Şekil 6'te gösterilmektedir. Her bir renk farklı bir alt konu başlığını temsil etmektedir. Grafikte görüldüğü gibi 5 farklı konu birbirinden ayrılmaktadır. Turuncu, Mavi ve mor renk



Şekil 5. Konu başlıkları için kelime bulutu (Word cloud for topics)



Şekil 6. t-SNE algoritması ile konu başlıklarının yakınlık gösterimi (Proximity display of topics with t-SNE algorithm)

ile temsil edilen konularla bazı bölgelerde yakınlaşmış hatta diğer renklerin arasına karışmış durumdadır. Bu yorumların diğer konuya çok yakın olduğunu göstermektedir. t-SNE algoritmasının yorumları başarı ile 5 farklı konu başlığına sınıflandırdığı görülmektedir.

"LDAvis" gösterimi [30], kullanıcıların bir metin modeline uygun bir konu modelindeki konuları yorumlamasına yardımcı olmak için tasarlanmıştır. "PyLDAvis", GDA algoritması kullanılarak tahmin edilen konuların web tabanlı etkileşimli bir görselleştirmesi olan LDAvis'i temel alarak oluşturulmuştur. PyLDAvis, kullanıcıların bir konu modelindeki konuları yorumlamasına yardımcı olan etkileşimli konu modeli görselleştirme için bir Python kütüphanesidir [31]. PyLDAvis iki farklı sütün içermektedir. İlk sütunda konuların birbirine olan uzaklıkları yer almaktadır. Sağ tarafta bulunan ikinci sütunda ise ilgili konu içerisinde yer alan kelimelerle ilgili veriler bulunmaktadır. PyLDAvis grafiklerinden elde edilen terimlerin konu başlıklarına dağılımı Tablo 5'te gösterilmektedir. GDA algoritmasından elde edilen konulara, veri önişleme sonucu kalan terimlerin en yüksek %28 oranda 1 nolu başlık ile eşleştiği, en düşük ise %14,5 oranında 5 nolu başlık ile eşleştiği görülmektedir.

Tablo 6-Tablo 10'da Reddit sosyal platformu "cancer subReddit" inden toplanan yorumlardan oluşturulan 5 konu başlığı için ayrı ayrı PyLDAvis grafiklerinden elde edilen sayısal değerler gösterilmektedir. Tabloda her başlık için en çok ilgili olan 10 terim gösterilmektedir. 1 nolu konu başlığına uyan terimler, tüm terimlerin %28'ini oluşturmaktadır. 5 Nolu konu başlığında gösterilen terimler, tüm terimlerin %14,5'ine karşılık gelmektedir. Tablo 6'da 1 nolu konu başlığı için 10 adet kelime listelenmektedir. Bu başlık için en etkili kelimeler "feel", "go" ve "like" kelimeleri olarak görülmektedir. Bu konu başlığında bulunan terimler bütün terimlerin %28'ini oluşturmaktadır. Çalışmada kullanılan veri seti için konu başlıkları içerisinde öne çıkan başlık olarak görülmektedir. 1 nolu başlığa en yakın başlık, 2 nolu başlıktır. Tablo 7'de 2 nolu konu başlığı kanser hastalarının hastalık zamanında moral-motivasyon ve çevrelerinden bekledikleri ilgi üzerine 10 adet kelime listelenmektedir. En etkili kelimeler "time", "sorri", "year" kelimeleridir. Bütün terimlerin %21,1'i bu konu başlığını oluşturmaktadır. Tablo 8'de kemoterapi başlığı üzerine 10 adet kelime gösterilmektedir. En etkili kelimeler "cancer", "chemo" ve "year" olarak listelenmiştir. 3 nolu konu başlığına en yakın konu başlığı olarak 4 nolu konu başlığı belirlenmiştir. Bütün terimlerin %18,3'ü bu konu başlığını oluşturmaktadır. Tablo 9'da ise hastalık döneminde beslenme ile ilgili konu başlığında etkili 10 terim listelenmektedir. Bütün terimlerin %18,1'i bu konu başlığını oluşturmaktadır. En yakın başlığı olarak 3 nolu başlık görülmektedir. Tablo 10'da tıbbi destek konusu ile ilgili 10 adet terim listelenmiştir. 5 nolu konu başlığı bütün konu başlıklarına uzak olarak görünmektedir. Bütün terimlerin %14,5'i bu konu başlığını oluşturmaktadır.

Tablo 5. Terimlerin konulara dağılımı
(Distribution of terms to topics)

No	Konu Başlığı	Yüzde
1	Diagnose of Disease	%28
2	Togetherness	%21,1
3	Chemotherapy	%18,3
4	Nutrition	%18,1
5	Medical Aid	%14,5

Tablo 6. GDA konu modelleme konu 1 pyldavis dağılımı
(LDA topic modelling topic 1 pyldavis distribution)

No	Terim	Genel Terim Sıklığı (adet)	Terim Sıklığı (adet)
1	feel	25.325	18.172
2	go	28.673	17.304
3	like	33.324	15.463
4	thing	21.765	15.281
5	help	26.983	14.783
6	peopl	19.173	14.498
7	get	18.452	14.395
8	cancer	62.786	13.901
9	want	20.754	13.590
10	need	18.476	12.874

Tablo 7. GDA konu modelleme konu 2 pyldavis dağılımı
(LDA topic modelling topic 2 pyldavis distribution)

No	Terim	Genel Terim Sıklığı (adet)	Terim Sıklığı (adet)
1	time	30.743	11.479
2	sorri	12.378	10.637
3	year	20.906	10.320
4	go	34.673	9.893
5	cancer	62.786	9.106
6	day	19.537	8.878
7	love	11.362	8.509
8	dad	9.763	8.253
9	mom	11.560	7.957
10	life	16.954	7.474

Tablo 8. GDA konu modelleme konu 3 pyldavis dağılımı
(LDA topic modelling topic 3 pyldavis distribution)

No	Terim	Genel Terim Sıklığı (adet)	Terim Sıklığı (adet)
1	cancer	62.786	22.893
2	chemo	24.708	10.432
3	year	20.906	10.404
4	surgeri	9.542	9.510
5	stage	11.754	8.703
6	treatment	22.768	7.598
7	month	12.453	7.376
8	tumor	6.905	6.905
9	get	18.452	6.817
10	radiat	9.768	6.616

Tablo 9. GDA konu modelleme konu 4 pyldavis dağılımı
(LDA topic modelling topic 4 pyldavis distribution)

No	Terim	Genel Terim Sıklığı (adet)	Terim Sıklığı (adet)
1	chemo	24.708	11.315
2	get	18.452	9.874
3	day	19.537	9.254
4	like	33.324	8.478
5	help	26.983	7.790
6	eat	7.435	7.435
7	week	7.250	7.250
8	take	15.698	6.943
9	pain	11.543	6.749
10	time	30.743	6.549

Tablo 10. GDA konu modelleme konu 5 pyldavis dağılımı
(LDA topic modelling topic 5 pyldavis distribution)

No	Terim	Genel Terim Sıklığı (adet)	Terim Sıklığı (adet)
1	cancer	62.786	14.786
2	treatment	22.768	9.129
3	doctor	11.906	8.109
4	get	18.452	7.245
5	patient	9.342	7.102
6	help	26.983	6.907
7	work	12.783	6.509
8	medic	8.078	6.106
9	may	10.984	5.905
10	oncologist	8.698	5.875

4. SONUÇLAR VE TARTIŞMALAR
(RESULTS AND DISCUSSIONS)

Araştırma bulgularından da görüleceği üzere, Reddit platformunda kullanıcı paylaşımları için konu modellemesinin faydasını ortaya koymaktadır. GDA kullanarak önerilen yaklaşım, kullanıcı paylaşımlarından tutarlı ve anlamsal olarak anlamlı konular başlıkları ve kümeleri oluşturmuştur. Bu konular, sadece büyük bir paylaşım koleksiyonunda bulunan metinlerin insanlar tarafından yorumlanabilir bir şekilde ayrıştırmasına değil, aynı zamanda hastaların ve doktorların ihmal etmiş olabileceği yeni içerikleri keşfetmelerine yardımcı olmaktadır. Aşağıda, çıkarılan konuların anlamsal olarak tutarlı olduğu, önerilen yaklaşımın paylaşımların etiketlenmesindeki, kanser hastalarının neler konuştuğunu ve son olarak önerilen yaklaşım kullanılarak konuların birbirine yakınlığını ve benzer konuların kümelenmesi ele alınmıştır.

Platform üzerindeki paylaşımlar, kanser hastalığı ile herhangi bir nedenden dolayı bağlantılı kişilerin paylaşımlarından oluşmaktadır. Bu konu genel başlık olarak çalışmanın ana konusu olarak seçilmiştir. Kullanıcılar kanser hastalığı ile ilgili görüşlerini, tecrübelerini çevrimiçi platform üzerinden kanser alt başlığında paylaşmaktadırlar. Alt başlık manuel olarak seçilmiştir. Bu sayfanın 29K üyesi bulunmaktadır ve 2008 yılından beri paylaşımlar yapılmaktadır. 109247 kullanıcı paylaşımı platformdan çekilerek analiz edilmiştir.

Veri seti üzerinde GDA, GAA ve NMA algoritmalarının hangi başlık sayısında tutarlı olduğunu belirlemek için başlık sayısı 2'den itibaren 7 başlığa kadar 3 algoritma için tutarlılık skorları hesaplanmıştır. Şekil 4'te görüldüğü gibi GDA algoritması 5 başlık için en iyi sonucu vermektedir. GDA algoritmasının en iyi tutarlılık skorlarını vermesinden dolayı, α ve β parametreleri için farklı değerler denenerek en iyi tutarlılık oranı olarak bulunan değerler ile elde edilen sonuçlar Tablo-3'de paylaşılmıştır. Tablo 3'de gösterilen kullanıcı paylaşımları sonucu elde edilen konu başlıklarının seçilen ana başlığa uygun olması kanser hastalığı ile ilgili paylaşımların anlamsal olarak da tutarlı olduğunu göstermektedir. GDA modeli tarafından çıkarılan 5 farklı konu genel olarak kanser hastalığı ile ilgili olduğu görülmektedir. Başlıklar kanser hastalığının teşhisi, tedavi süreci, hastalık dönemindeki moral-motivasyon, kemoterapi dönemi ve medikal olarak destek konularından oluşmaktadır. Bu sonuçlar doğrultusunda GDA algoritmasının doğru konuları tespit ettiğinin göstergesidir.

Tutarlılık testi öncelikle algoritmanın belirlenmesi amacıyla 3 farklı algoritmada uygulanmıştır. 3 algoritma arasından en iyi çıkan GDA algoritması için farklı parametrelerde denenen tutarlılık skorlarına göre 0,537 ile konu içerisindeki kelimelerin tutarlı oldukları gözlemlenmiştir. Bulunan tutarlılık değeri ortalamanın üzerinde ve yüksek doğruluk oranlı sonuç olarak literatürde yer almaktadır.

Sosyal platformdan elde edilen veriler ve GDA algoritması ile elde edilen sonuçlar Reddit platformunun kanser subReddit'inin, kanser hastaları ve hastalıkla ilgili olan kişilerin görüşlerini paylaştığı bir veri kaynağı olduğunu doğrulamaktadır. Bu alandaki doktorlar, hastaların hangi konular hakkında neler hissettiğini öğrenebilir. Bu platformdaki verilerin kullanılması, doktorların toplulukta tartışılan tedaviler hakkında bilgi edinmesine olanak sağlayabilir. Bu veriler doktorlar için resmi bir çalışmayı takip edebilecek herhangi bir hastalığı vurgulamamasına rağmen, yapılan paylaşımlar üzerinden hastalık ile ilgili trendler izlenebilir. Örneğin; 1 nolu konu başlığı elde edilen terimlerin %28'inden oluşmaktadır. Kanser hastalığı ile ilgili bu subReddit'te konuşanların çoğunluğunun kemoterapi üzerine paylaşımlar yaptığı görülmektedir. GDA algoritması ile hastalık ve tedavisinden farklı bir şekilde kanser hastalarının bu dönemde beraberlik ve desteğe olan ihtiyacı tespit edilmiştir.

Bu çalışma, çevrimiçi platformlarda paylaşılan verilerin tam olarak doğrulanamaması nedeniyle sınırlıdır. GDA denetlenemeyen bir modeldir, model çıktısının karşılaştırılabileceği temel bir gerçek yoktur. Bu yüzden modelden elde edilen sonuçlar nitel olarak analiz edilmelidir. Hastalık ile ilgili yapılan yorumların, denenip elde edilen sonuçların doğru olup olmadığını belirlemek zordur. Paylaşımları yapanlar genelde isimsiz olmalarından dolayı doğru bilgiyi paylaştığı bilinmemektedir.

Kanser hastalığı için çevrimiçi platformlar özellikle sosyal medya siteleri paylaşımları baz alınarak oluşturulan sistemler faydalı olabilir. Bu şekilde eğer herhangi bir yanlış bilgi varsa doktorlar tarafından bu bilgiler düzeltilebilir. Güncel bir GDA tabanlı web servisi ile paylaşımlar güncel olarak takip edilerek doktorların kullanabileceği bir uygulama yapılabilir. Uygulama hem doktorların hasta davranışı hakkındaki bilgilerini artırabilir hem de bunun üzerine hareket etme fırsatı sunabilir. Önerilen GDA modeli sadece sağlık alanı ile kısıtlı değildir. Eğitim, pazarlama, turizm sektörü gibi farklı alanlarda da uygulanabilir.

GDA algoritması avantajlarına rağmen büyük veri setleri için daha verimli ve etkinli yöntemler geliştirilebilir. Büyük veri setlerinde öncelikle özetleme işlemi yapılıp özet veriler üzerinden konu modelleme algoritması uygulanabilir. Bu yaklaşım ile verilerin işlenmesi ve sonuca ulaşma süresi kısalabilir. Ayrıca birçok alanda uygulanmakta olan derin öğrenme algoritmaları ile GDA algoritması birleştirilerek daha etkin bir hale getirilebilir.

KAYNAKLAR (REFERENCES)

1. Bhat, M. R., Kundroo, M. A., Tarray, T. A., ve Agarwal, B., Deep LDA: A new way to topic model, Journal of Information and Optimization Sciences, 41 (3), 823-834, 2019.
2. Alexa. Alexa Top 500 Global Sites. http://alexa.com/topsites/. Erişim Tarihi Şubat 10 2020.
3. Blei, D. M., Ng, A. Y., ve Jordan, M. I., Latent Dirichlet Allocation, Journal of Machine Learning Research, 3, 993-1022, 2003.
4. Dan Foresee, F. ve Hagan, M. T., Gauss-Newton approximation to Bayesian learning, Proceedings of International Conference on Neural Networks (ICNN'97), Houston-USA, 1930-1935, 12 Haziran, 1997.
5. Okon, E., Rachakonda, V., Hong, H. J., Callison-Burch, C., ve Lipoff, J., Natural Language Processing of Reddit Data to Evaluate Dermatology Patient Experiences and Therapeutics, Journal of the American Academy of Dermatology, 83 (3), 803-808, 2020.
6. Ding, S., Li, Z., Liu, X., Huang, H., ve Yang, S., Diabetic complication prediction using a similarity-enhanced latent Dirichlet allocation model, Information Sciences, 499, 12–24, 2019.
7. Chen, C. ve Ren, J., Forum latent Dirichlet allocation for user interest discovery, Knowledge-Based Systems, 126, 1–7, 2017.
8. Bastani, K., Namavari, H., ve Shaffer, J., Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints, Expert Systems with Applications, 127, 256–271, 2019.
9. Wang, W., Feng, Y., ve Dai, W., Topic analysis of online reviews for two competitive products using latent Dirichlet allocation, Electronic Commerce Research and Applications, 29, 142–156, May, 2018.
10. Hagen, L., Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models?, Information Processing & Management, 54 (6), 1292–1307, 2018.
11. Griffiths, T. L. ve Steyvers, M., Finding scientific topics, Proceedings of the National Academy of Sciences, 101 (1), 5228–5235, 2004.
12. Wang, Y., Agichtein, E., ve Benzi, M., TM-LDA: efficient online modeling of latent topic transitions in social media, KDD '12: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, Beijing- China, 123-131, 12-16 Ağustos, 2012.
13. Xu, Z., Ru, L., Xiang, L., ve Yang, Q., Discovering User Interest on Twitter with a Modified Author-Topic Model, 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Lyon-Fransa, 422-429, 22-27 Ağustos 2011.
14. Xu, H., Zhang, F., ve Wang, W., Implicit feature identification in Chinese reviews using explicit topic mining model, Knowledge-Based Systems, 76, 166–175, 2015.
15. Güven Z. A., Diri B., Çakaloğlu T., Comparison of n-stage Latent Dirichlet Allocation versus other topic modeling methods for emotion analysis, Journal of the Faculty of Engineering and Architecture of Gazi University, 35 (4), 2135-2145, 2020.
16. Lu, Y., Mei, Q., ve Zhai, C., Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA, Information Retrieval, 14 (2), 178–203, 2011.
17. Blei, D. M., Probabilistic topic models, Communications of the ACM, 55 (4), 77-86, 2012.
18. Landauer, T. K. ve Dutnais, S. T., a Solution to platos problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge, Psychological review,104 (2), 211–240, 1997.
19. Lee, D. D. ve Seung, H. S., Algorithms for non-negative matrix Factorization, 14th Annual Neural Information Processing Systems Conference, NIPS 2000 , Denver-USA, 556–562, 27 Kasım-2 Aralık, 2000.
20. Stevens, K., Kegelmeyer, P., Andrzejewski, D., ve Buttler, D., Exploring Topic Coherence over Many Models and Many Topics, Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea, 952–961, 5 Temmuz, 2012.
21. Wallach, H. M., Mimno, D., ve McCallum, A., Rethinking LDA: Why priors matter, Advances in Neural Information Processing Systems 22-Proceedings of the 2009 Conference, Vancouver- Kanada 1973–1981, 7-10 Aralık, 2009.
22. Reddit, http://www.reddit.com, Erişim Tarihi 10 Ocak 2020
23. PRAW, http://praw.readthedocs.io, Erişim Tarihi 17 Eylül 2019.
24. Nltk, http://www.nltk.org, Erişim Tarihi 17 Eylül 2019.
25. Joachims, T., A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorixation,

Proceedings of ICML97, San Francisco-USA,143-151, 8-12 Temmuz, 1997.

26. Lee, D. D. ve Seung, H. S., Learning the parts of objects by non-negative matrix factorization, Nature, 401, (6755), 788–791, 1999.

27. Altıntaş, V., Topal, K., ve Albayrak, M., Sosyal Medya Platformu Üzerinde Gizli Anlam Analizi, European Journal of Science and Technology, 16, 863–869, 2019.

28. Röder, M., Both, A., ve Hinneburg, A., Exploring the Space of Topic Coherence Measures, Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, New York-USA 399–408, 2-6 Şubat, 2015.

29. Maaten, L. van der ve Hinton, G., Visualizing Data using t-SNE, Journal of Machine Learning Research, 9, 2579–2605, 2008.

30. Sievert, C. ve Shirley, K., LDAvis: A method for visualizing and interpreting topics, Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Maryland-USA 63–70, 2014.

31. Hidayatullah, A. F., Ma'arif, M. R., Road traffic topic modeling on Twitter using latent Dirichlet allocation, International Conference on Sustainable Information Engineering and Technology (SIET17), Batu, Indonesia 47-52, 24-25 Kasım, 2017.