ESTUDAM Bilişim Dergisi
Cilt 3, Sayı 2, 46-53, 2022

Journal of ESTUDAM Information
Volume 3, Issue 2, 46-53, 2022

DOI:10.53608/estudambilisim.1097978

(Geliş Tarihi / Received Date: 03.04.2022, Kabul Tarihi/ Accepted Date: 27.04.2022)

(**Araştırma Makalesi**)

Konu Modelleme Yöntemlerinin Karşılaştırılması

Ahmet KAYA[*1], Eyyüp GÜLBANDILAR[2]

[1]Eskişehir Osmangazi Üniversitesi, Mühendislik-Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Eskişehir,
ORCID No : http://orcid.org/0000-0002-5222-0887
[2] Eskişehir Osmangazi Üniversitesi, Mühendislik-Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Eskişehir,
ORCID No : http://orcid.org/0000-0001-5559-5281

Anahtar Kelimeler:
Konu Modelleme, İlişkisel Konu Modeli, Yapısal Konu Modeli, Konu Tutarlılığı, Çapraşıklık

Özet: Son zamanlarda internet üzerinde üretilen veriler her geçen gün artmaktadır. Bu verilerin önemli bir çoğunluğunu da metinler oluşturmaktadır. Metinlerin çoğunlukta olması, bilim insanlarını bu alanda daha fazla çalışma yapmaya yönlendirmiştir. Metinler üzerinde yapılan çalışmaların en popüler olanı Konu Modelleme (KM) yöntemleridir. Konu modelleme yöntemleri metinlerin içerisinde gizli veya açık geçen konuları tespit etmektir. Bu çalışma kapsamında elde edilen metin veri kümeleri üzerinde Gizli Dirichlet Ayrımı (GDA), ilişkisel konu modeli (İKM) ve yapısal konu modeli (YKM) yöntemleri uygulanmıştır. Ayrıca çalışma da konu modelleme yöntemlerinin sonuçlarını karşılaştırabilmek için konu tutarlılığı ve çapraşıklık değerleri kullanılmıştır. Kaynak olarak kullanılan yayındaki yöntemlerin sonuçları ile çalışmada elde edilen sonuçların aynı olduğu görülmüştür. Çapraşıklık değerine ek olarak kullandığımız tutarlılık değeri de aynı şekilde YKM yönteminde daha başarılı sonuçlar elde edildiği gösterilmiştir. Tutarlılık değeri 0.509 olarak YKM tip 3 yöntemi en iyi sonucu vermiştir. Ayrıca bundan sonra yapılacak çalışmalar içinde karşılaştırma yöntemi gösterilmiştir.

(**Research Article**)

Comparison of Topic Modeling Methods

Keywords:
Topic Modelling , Correlated Topic Model, Structural Topic Model, Topic Coherence, Perplexity

Abstract: The amount of data created on the internet has been steadily expanding in recent years. Texts make up a large portion of the data. The fact that the texts are in the majority has led scientists to do more studies in this field. The Topic Modeling (TM) approach is the most frequently used method for studying literature. Methods for identifying hidden or open subjects in texts are known as "topic modeling." The text datasets collected for this investigation were subjected to the Latent Dirichlet Allocation (LDA), Correlated Topic Model (CTM), and Structural Topic Model (STM) approaches. In addition, the findings of the subject modeling approaches in the study were compared using subject consistency and complexity values. The results of the published techniques employed as a source and the findings of the study were found to be identical. Moreover to the crowding value, the consistency value we employed revealed that the STM technique produces more successful outcomes. The STM type 3 approach produced the best results, with a consistency score of 0.509. In addition, future investigations will reveal the comparative approach.

1. GİRİŞ

İnsanlar tarih boyunca nesilden nesile bildiklerini gelecek nesillere aktarmak istemişlerdir. Tarih öncesi dönemlerde resim veya semboller kullanılmıştır. Bunlar tarih öncesi dönemleri anlamak için günümüzde bile hala başvurduğumuz kaynaklardır. Özellikle yazının bulunması ile sonraki nesillere bırakılan mirasın epey arttığı söylenebilir. Artık, günümüzde özellikle internet çağı sonrası bu mirasın her geçen gün daha hızlı bir şekilde arttığı görülmektedir. Bırakılan miras her yıl katlanarak artmaya devam etmektedir. Tarih boyunca

*İlgili yazar/Corresponding Author: ahmetkaya.3408@gmail.com

üretilen kaynakların artık günümüzde kısa süre içerisinde üretildiğini görülmektedir.

Son zamanlarda üretim kaynaklarını takip etmekte zorlanırken, geçmiş olaylardan gelen birikimlerini incelemek, takip etmek veya onlardan bir değerli bilgiyi üretmek oldukça zorlaşmıştır.

Metin Madenciliği-Text Mining (MM), yapısal olmayan ve düzensiz yapıdaki elektronik metin yığınlarından; önceden bilinmeyen, potansiyel olarak kullanışlı, yapısal ve düzenli veri elde etme sürecidir. Elde edilen bilgiyle, analiz edilen metin kaynaklarında açık olarak görülmeyen ilişkiler, hipotezler ve eğilimler tespit edilir. Metin madenciliği, veri madenciliğinin bir parçası olarak düşünülmesine rağmen, alışılagelen veri madenciliğinden farklıdır. Ana farklılık, MM örüntülerin olay tabanlı veri tabanlarından daha çok, doğal dil metinlerinden çıkarılmasıdır. En basit anlamda MM çalışmaları, veri kaynağı olarak metinleri kabul eden veri madenciliği çalışmasıdır ve metinler üzerinden yapısallaştırılmış veri elde etmeyi amaçlar. Örneğin; metinlerin sınıflandırılması (classfication), bölütlenmesi (clustering), metinlerden konu çıkarılması (entity extraction), sınıf taneciklerinin üretilmesi (production of granular taxonomy), duygusal analiz (sentimental analysis), metin özetleme (document summarization), ve varlık ilişki modellemesi (entity relationship modelling) ve konu modelleme (topic modelling) gibi çalışma alanları bulunmaktadır [1].

Konu modelleme (KM), MM alt alanlarından biridir. Asıl hedefi belgeler içerisindeki gizli veya açık konuları ortaya çıkarmaktır. Özellikle son yıllarda artan önemi ve yapılan çalışmalarla birlikte metin madenciliği alt alanları arasında önemli bir yer tutmaktadır. KM belgelerde denetimsiz olarak konu sınıflandırması yapıldığı yöntemdir. Bu çalışma kapsamında veri kümeleri üzerinde 3 farklı konu modeller uygulanmıştır. Denetimsiz olması sonuçların değerlendirilmesi konusunda zorluklar çıkarmaktadır. Sonuçların karşılaştırılması içinde tutarlılık değeri ile bir karşılaştırma yöntemi gösterilmiştir. Çalışmanın amacı tutarlılık değerinin konu modelleme yöntemlerinin karşılaştırılması için kullanışlı bir değerlendirme kriteri olduğunu göstermek ve hangi konu modelleme yönteminin bu kritere göre daha başarılı olduğunu göstermektir. Konu modelleri ile illgili yapılan çalışmalar sırasıyla İlişkili Konu Modeli (İKM) ve Yapısal Konu Modeli (YKM) şeklinde aşağıdaki şekilde incelenmiştir.

Alghamdi [2] çalışmasında konu modelleme alanında kullanılan yöntemleri ve konuların tespitinin nasıl yapıldığı incelenmektedir. Xiao [3] çalışmasında konu modelleme yöntemlerinde konu dağılımlarının hangi yöntemde daha iyi sonuç verdiğini ölçebilmek için bir sıralama algoritması önerilmiştir.

Hilmi [4] çalışmasında 1973 yılından 2019 yılına kadar zaman aralığındaki 1824 adet makale üzerinde GDA yöntemi ile konu modelleme çalışması yapılmıştır. Bu çalışmanın temel amacı, makalelerde geçen gelişmekte olan sektörlerdeki konuların tespitini yapmaktır. Z. A. Güven [5] çalışmasında Türkçe veri seti üzerinde konu modelleme yöntemlerini uygulamıştır. Tutarlılık değerleri kullanılarak konu modelleme yöntemleri sonuçları birbirleri ile karşılaştırılmıştır. Bu çalışmada GDA yönteminin diğer yöntemlere göre daha kötü sonuçlar verdiği gösterilmiştir. Ekinci vd. [6] çalışmasında; kullanıcı yorumlarından ürün özelliklerini çıkarmada GDA yöntemi kullanılmıştır. Türkçe otel yorumları üzerinden elde edilen sonuçlar, GDA'nın özellik çıkarmada başarılı olduğunu göstermiştir. E. S. Negara [7] çalışmasında Twitterda yapılan yorumların hangi konular ile ilgili olduğunu belirlemek için yapılmıştır. Bu çalışmada atılan twitlerden konu tahmini yapılarak; 4 farklı konu seçilmiştir. Yapılan çalışmada GDA konu modelinin çalışmadaki diğer konu modellerine göre daha başarılı olduğu gösterilmiştir. F. Zhang [8] tarafından GDA konu modeline dayanan, kısa metinler üzerinde kelimelerin birlikte bulunma bilgisi eksikliği sorununu hafifletecek bir yöntem önerilmiştir. Yöntemin sonuçları değerlendirildiğinde konu tutarlılığı sonuçlarına göre önemli gelişmeler sağlandığı tespit edilmiştir. GDA'nın bir sınırlaması, örneğin sporla ilgili bir belgenin uluslararası finanstan ziyade sağlıkla ilgili olması daha olası olmasına rağmen, konu korelasyonunu modelleyememesidir. Bu sınırlama, konu oranları arasındaki değişkenliği modellemek için Dirichlet dağılımının kullanılmasından kaynaklanmaktadır.

Blei D. [9] çalışmasında konu oranlarının lojistik normal dağılım yoluyla korelasyon sergilediği İKM yöntemi önerilmiştir. İKM yöntemi sonucuna göre, GDA yönteminden daha iyi sonuçlar verdiği gözlenmiştir. Blei D. [10] çalışmasında İKM konu modelini 57 milyon kelimeden oluşan ve 1990-1999 yılları arasında yayınlanan Science makaleleri üzerinde uygulamıştır. İKM yönteminin, GDA yönteminden daha iyi sonuçlar verdiği gösterilmiştir. Liu L. [11] çalışmasında konu dağılımlarını yeniden şekillendirerek konular arasında korelasyon yakalamak için yeni bir yöntem önermiştir. Yapılan çalışmada önerilen yöntemin etkinliği doğrulanmıştır. Bu çalışmada konu modellerini sonuçlarını karşılaştırmak için konu tutarlılığı ve çapraşıklık değerleri kullanılmıştır. Fu [12], çalışmasında ikinci derece istatistiklerini kullanarak yeni bir konu belirleme kriteri önermiştir. Önerilen yöntemin çeşitli değerlendirme ölçütleri altında kelime tabanlı yaklaşıma kıyasla gelişmeler gösterdiği görülmüştür. He [13] çalışmasında kompakt konu yerleştirmelerini öğrenen ve konu vektörleri arasında yakınlık yoluyla konu korelasyonlarını yakalayan yeni bir model önerilmiştir. Çalışmada İKM yöntemi, yüksek hesaplama maliyeti ve zayıf ölçekleme nedeniyle küçük model ve probem boyutlarıyla sınırlandırılmıştır. Çalışma sonucunda belge sınıflandırma ve almada rekabetçi veya üstün performans sağlayarak modelleme kalitesinden ödün vermeden, mevcut korelasyon sonuçlarından birkaç kat daha büyük olan model ve veri ölçeklerinin işlenebildiği gösterilmiştir. Xu [14] , çalışmasında görüntü veri seti üzerinde konu modelleme algoritmalarını uygulayıp performans değerlendirmesi yapmıştır. Çalışmada GDA yönteminde karşılaşılan konu korelasyonu eksikliği problemine çözüm olarak daha esnek model dağılımının kovaryans yapısı ile konu korelasyonunu yakalamak için doğal dil işlemeden İKM yöntemini genişleterek

kullanılmıştır. Varyasyon yöntemine dayalı yaklaşık bir çıkarım ve tahmin algoritması türetilmiştir. Çalışmada uygulanan modelin performansı iki karşılaştırılmalı görüntü veri setinde incelenmiş ve İKM yöteminin GDA yöntemine göre kelime korelasyonu için daha gelişmiş performans gösterdiği gösterilmiştir. Funnell [15] , çalışmasında İKM yöntemini kanser veri seti üzerinde uygulamış ve daha iyi sonuçlar elde etmiştir. Çalışmada tıp alanında hasta sınıflandırması için gelecekteki çalışmalar için faydalı olabileceği gösterilmiştir.

Esmizadeh [16], çalışmasında ekonomi verileri üzerinde YKM konu modelini uygulamıştır. Çalışmada daha önceki çalışmalara yeni bileşenler ekleyerek paylaşım ekonomisi bilgi tabanına katkıda bulunulmuştur. Roberts [17], makalesinde YKM yöntemi için R yazılım dilinde stm paketinin nasıl kullanıldığı gösterilmiştir. Stm paketi konuları keşfetmek, belirsizliği tahmin etmek ve ilgi miktarlarını görselleştirmek için zengin yollar dahil olmak üzere birçok faydalı özellik sağlamaktadır. Liu S. [18] , çalışmasında farklı veri kümeleri üzerinde YKM yöntemini uygulayıp sonuçları üzerinde değerlendirmeler yapılmıştır. Çalışma sonuçları ile geçmiş bilgilerden yeni bilgilerin keşfedilmesindeki potansiyel ortaya konulmuştur. Bai [19], çalışmasında deniz taşımacılığı alanındaki literatür üzerinde trend temaları tespit etmek için YKM yöntemi kullanılmıştır. Çalışmada 30 yıl süre ile yayınlanan 3199 makale analiz edilmiştir. Sonuçlar, deniz taşımacılığının başlıca akademik kaygılarının liman yönetimi, konteyner operasyonları ve gemi taşımacılığı yönetimi ile ilgili olduğu göstermiştir. Elde edilen sonuçlar, YKM yönteminden keşfedilen ana araştırma konuları ve ortaya çıkan trendlerin, bu alan ile ilgili olan kişilerin veya kurumların daha bilinçli karar vermelerine yardımcı olabileceğini gösterilmiştir. Sim [20] çalışmasında uzaktan öğrenme faktörlerini belirlemek için YKM yöntemi kullanmıştır. Araştırmacılar öğrenme faktörlerini belirlemişlerdir ve öğrenme faktörleri arasındaki ilişkisel yapıyı ortaya koymuşlardır. Ma [21] , çalışmasında çince haberler üzerinde YKM yöntemi uygulanmıştır. Bu çalışma ile mevcut Çince dilinde mahremmiyet kavramının farklı konularda nasıl kullanıldığını ve etki alanlarını ortaya koyan deneysel kapsamlı ilk keşiflerinden biri olmuştur. Çalışmada elde edilen bulgular bu alandaki araştırmaları iletletmek için kullanılabilir. Hu [22] çalışmasında otel yorumları üzerinde YKM konu modeli uygulanmıştır. Çalışmada New York şehrindeki 27.864 otel için müşteri yorumları kullanılarak müşteri şikayet ve memnuniyet konuları tespit edilmiştir. Çalışma sonuçlarında olumlu ve olumsuz yorumlar ayrı ayrı incelenip sonuçlar değerlendirilmiştir. Müşteri şikayetlerinin farklı otel sınıflarında nasıl değiştiği de incelenmiştir. Sonuçlar, yüksek kaliteli oteller için müşteri şikayetlerinin esas olarak hizmet sorunlarıyla ilgili olduğunu, düşük kaliteli otellerin müşterilerinin ise tesisle ilgili sorunlardan sık sık rahatsız olduğunu göstermiştir. Bu çalışma otellerin müşteri memnuniyetsizliğinin hangi alanlarda olduğunu tespit etmeye çalışan literatür çalışmalarına da katkıda bulunmaktadır.

2. MATERYAL VE METOT

2.1. Konu Modelleme

Konu modelleme, verilerdeki gizli veya açık temaları tanımlayan denetimsiz bir yöntemdir. Denetimsiz olmasında dolayı etiketli verilere ihtiyaç duyulmaz. Bilgi çıkarmak için doğrudan metin belgelerine uygulanır. Konu modelleme, bir metin veri kümesi içinde yer alan temaları bulurken keşfedici ve üretken bir şekilde çalışır. Belgelerde gözlenen kelimelere dayanarak metinlerdeki temaları çıkarmak için olasılık çerçevesini kullanıp konular tespit edilir. Büyük hacimli metin verilerini eleme sürecini otomatikleştirmek veya anlamaya yardımcı olmak için kullanılabilir. Anahtar konular keşfedildikten sonra, eğilimleri belirlemek için veya sınıflandırma yapabilmek için kullanılabilir.

2.1.1. Gizli Dirichlet Ayrımı

Belgeler içerisinden ilişkili konuları çıkarmaya yarayan üretken ve olasılıksal bir konu modelleme yöntemidir. Dayandığı temel fikir, belgelerin konuların karışımı olduğu, aynı zamanda konuların da kelimelerin bir karışımı olduğu üzerinedir. Konu modelleme yöntemlerinin elde ettiği popülariteye en fazla katkı veren yöntemdir. Sonraki çalışmalarda yapılan çalışmalara öncülük etmiş ve büyük katkılar sağlamıştır. Günümüzde bile yöntem ile ilgili bir çok yeni çalışma yapılmaktadır. Gizli Dirichlet Ayrımı (GDA), belgelerdeki gizli temaları keşfeder. Bunu yapabilmek için üretken bir olasılık modeli ve dirichlet dağılımlarını kullanır. İki farklı dirichlet parametresi kullanılır. Parametrelerden biri konuların belgelerdeki 7 dağılımını etkilerken, diğer parametre ise kelimelerin konulardaki dağılımı üzerinde etkilidir. Dirichlet değerlerinin yüksekliği ve düşüklüğü yapılan dağılımların etkisini etkilemektedir. GDA yöntemi, Bayes çerçevesine dayanmaktadır. Blei, D. vd. [23] İşlemler öncesi herhangi bir bilgiye ihtiyaç duymaz. GDA kelimeleri bir araya toplayarak işleme alır. Bu işlemde kelimelerin cümle içi yerleşimleri önemli değildir. Kelimelerin birlikte bulunma durumları dikkate alınmamaktadır. Ekinci vd. [24] Şekil 1.'de yöntemin grafiksel gösterimi verilmiştir.

2.1.2. İlişkili Konu Modeli

Konu modelleme yöntemleri belge koleksiyonlarını istatiksel analiz edebilmek için faydalı yöntemlerdir. İstatistiksel yöntemlerden biri olan İlişkili Konu Modeli (İKM), belgelerdeki konuları bulmaya çalışır. İKM, konuların tek tek belgeler üzerinde nasıl dağıldığını açıklamaya çalışan üretken bir modeldir. Blei D. [9] İlişkili konu modeli (İKM), GDA yönteminin gelişmiş bir versiyonudur. GDA yöntemi tespit edilen konuların birbiri ile ilişkisini dikkate almaz. Bu sınırlama konu oranları arasındaki değişkenliği hesaplamak için dirichlet dağılımının kullanmasından kaynaklıdır. Ancak konular birbiri ile ilişkili olabilirler. Örneğin çıkarım yapılan konulardan biri yemek ile ilgili iken; diğer konular da sağlık veya spor ile ilgili olabilir. Konu seçimi yaparken rastgele seçilen konular yerine ilişkili olduğu tespit edilen konulara atamak daha iyi sonuçlar elde etmemizi

sağlayacaktır. [25] İKM konu ilişkilerini hesaplamak için lojistik normal dağılımı kullanır. Lojistik normal, çok değişkenli normal bir rastgele değişkeni dönüştürerek bileşenler arasında genel bir değişkenlik modeline izin veren tek yönlü bir dağılımdır. Blei D. [9] bu dağıtımda hedef, her konu için bir değer üretmektir. Lojistik normal dağılım, kovaryans matrisi ile gizli konular arasındaki korelasyonları gösterir. Lojistik normal, İKM'nin çıkarım sürecine karmaşıklık katar. Şekil 1.'de yöntemin grafiksel gösterimi verilmiştir. Çıkarım sürecinde konu oranlarının bir dirichlet yerine lojistik normalden çekilmesi dışında, GDA'nın üretici süreciyle aynıdır. D bir belge koleksiyonu , K konuları ve K boyutlu normal ortalama dağılımı ve kovaryans matrisi $N(\mu, \Sigma)$ verildiğinde, İKM belgeleri aşağıdaki üretken sürece göre oluşturur. Konu oranı(θ) her belge için lojistik normal 8 dönüşümden elde edilir. Blei D. [9], Oo, M. K.vd. [26] Lojistik normal dağılım için Denklem 1 kullanılır.

$$\theta = f(\eta) = \exp\{\eta\} \sum \exp\{\eta i\} \quad (1)$$

Dirichlet dağılımı yerine lojistik normal dağılım kullanmanın bir sonucu olarak aynı özelliklere sahip sonuçlar elde edilmemektedir. Bu nedenle, GDA'da kullanılan standart varyasyon çıkarım tekniği yerine İKM yönteminde ortalama varyasyon çıkarımı kullanılır. Bu yöntemi optimize edebilmek için yinelemeli yöntemler tercih edilir. Bu nedenle İKM yöntemi GDA yöntemine göre çalışma süreleri açısından daha uzun sürmektedir. İKM yönteminde uzun eğitim sonucu daha iyi konulara sahip olmakla kalınmaz; aynı zamanda konular arasında ilişki de takip edilmektedir. Konuların birbiri ile illişkili olduğu durumlarda kullanıldığında fayda sağladığı görülmüştür. [25]

2.1.3. Yapısal Konu Modeli

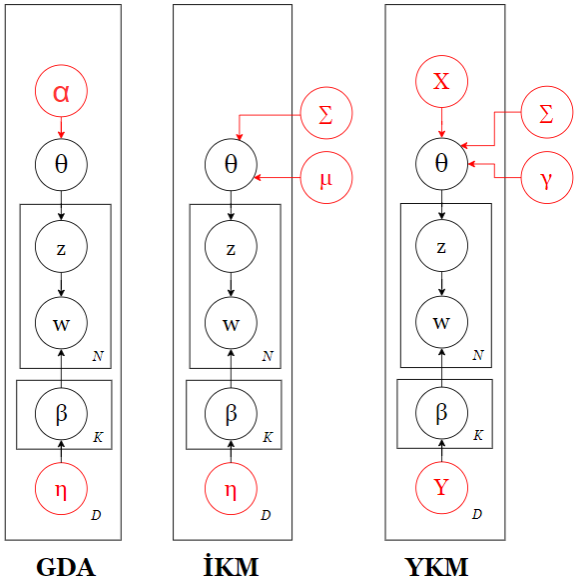
Konu modelleme yötemleri belgeler içindeki konuları ortaya çıkarmayı hedefler. Çıkarım yapılırken sadece metne bakılır, diğer hiçbir bilgi dikkate alınmaz. Metin ile birlikte diğer bilgiler (yazar, tarih, kaynak, puanı vb.) de konu çıkarımı için etkili olabilir. Metinle birlikte diğer bilgileri de dikkate alarak konu modelleme işlemlerini yapmak için Yapısal Konu Modeli (YKM) ortaya çıkmıştır. YKM işlemlerinde sadece metin içeriğini kullanmak yerine metin ile birlikte bulunan meta verileri de yönteme dahil edilir. YKM üzerine yeni özellikler ekleyerek İKM yöntemini kullanır. [25] Şekil 3.1.'e bakıldığında YKM yönteminin İKM yöntemi ile farkları görülebilir. GDA konu dağılımlarını hesaplarken dirichlet dağılımını kullanır. YKM'de ise belge ve belge meta verileri konu dağılımlarında kullanılır.[34] YKM sadece yüksek kaliteli bir model değildir. Aynı zamanda meta verilerin bir belgenin bir konu içinde kullandığı kelimeleri nasıl etkilediğini tespit etmeye yardımcı olmaktadır. Roberts, M. E. vd. [27], "Yapısal Konu Modeli, belge düzeyinde ortak değişken bilgileriyle konu modelleme için genel bir çerçevedir" yargısını ortaya koymuştur. Ortak değişkenler, çıkarımı ve nitel yorumlanabilirliği iyileştirebilir ve tüm adımları etkilemesine izin verilir. [28] Şekil 1.'de yöntemin grafiksel gösterimi verilmiştir.

2.2. Değerlendirme Metrikleri

2.2.1. Çapraşıklık (Perplexity)

Konu modellerini değerlendirmek için geleneksel bir ölçüm olan çapraşıklık[29] (perplexity) kullanılır. Aynı zamanda beklenen olasılık (held out likelihood) olarak adlandırılır. Çapraşıklık metriği bir eğitim kümesinin eğitildikten sonra elde edilen sonuç kümesini tahmin etme yeteneğidir.

Çapraşıklık, eğitilmiş bir konu modelinin yeni verileri ne kadar başarılı tahmin ettiğinin bir ölçüsüdür. Yüksek çapraşıklık değeri iyi bir konu modeli anlamına gelmektedir. Bu sezgisel bir anlam ifade etse de konu modelleri tarafından oluşturulur. Bu nedenle, kafa karışıklığı konu modellerini değerlendirmek için matematiksel olarak sağlam bir yaklaşım olsa da , insan tarafından yorumlanabilir konuların iyi bir göstergesi değildir.



Şekil 1. Konu Modelleme Yöntemlerinin Grafiksel Gösterimi

Tablo 1. Şekil 1.'deki Parametrelerin Açıklaması

Parametre	Açıklaması
D	Toplam doküman sayısı
K	Konuların sayısı
N	Sözlükte bulunan toplam kelime sayısı
α	Dirichlet parametresi
η	Dirichlet parametresi
θ	Konuların dokümanlardaki dağılımı
β	Kelimelerin konulardaki dağılımı
Dm	m. dokümanın boyutu
Zm,n	m. dokümandaki n. kelimenin konusu
Wm,n	m. dokümandaki n. kelime
∑	Konular arasındaki korelasyonlar
μ	K boyutlu ortalama matris
X	Belge meta verilerin matrisi
γ	Konu dağılımını etkileyen meta veriler
Y	Belge meta verilerin matrisi

Çapraşıklık metriği, insan yorumu için her zaman iyi sonuçlar sağlamaz. Araştırmalar, çapraşıklık metriğinin insan yorumları ile karşılaştırıldığında ilişkili olmadığını göstermiştir. Jonathan Chang vd. [30] tarafından yapılan ve çapraşıklığın konuların tutarlı olup olmadığını anlamak için kullanılmayacağını göstermiştir. Çapraşıklığı, kelime ihlali ve konuya müdahale gibi insan yargısı yaklaşımlarıyla karşılaştırırken, aralarında negatif bir ilişki olduğunu gösterilmiştir. Konu modelleme yöntemlerinde çapraşıklık metriği yanıltıcı olarak görülmektedir. [30]

2.2.2. Konu Tutarlılığı (Topic Coherence)

Konu tutarlılığı, kelime kümelerinin bir kelime veya alt kelime grupları arasındaki uyumu ölçmektedir. Kelime kümesi alt gruplara ayrılır. Alt gruplar için olasılık hesabı yapılır. Doğrulama ölçütleri belirlenir. Değerler toplanır. Tutarlılık değeri elde edilmiş olur. [31]

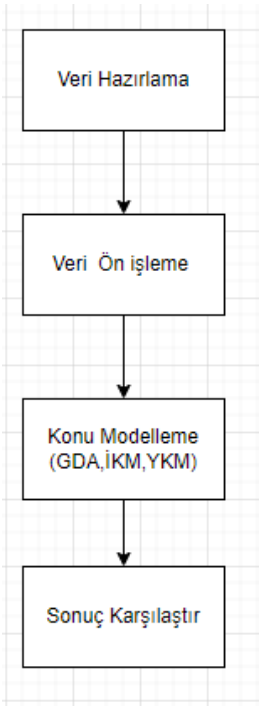
Tutarlılık, konu modeli tarafından oluşturulan konulardaki kelimeler arasındaki anlamsal benzerlik derecesini ölçer. Konu içindeki kelimeler ne kadar benzerse, tutarlılık puanı da o kadar yüksek olur. Karşılaştırma için kelimeleri gruplamak, kelime birlikte bulunma olasılıklarını hesaplamak ve nihai bir tutarlılık ölçüsünde toplamak için farklı yöntemlere dayalı olarak tutarlılığı hesaplamanın birkaç yolu bulunmaktadır. Tutarlılık, konu modellerini nicel olarak değerlendirmek için popüler bir yaklaşımdır ve Python ve Java gibi kodlama dillerinde yaygın uygulamaları bulunmaktadır.

Şaşkınlığın eksikliklerinden biri, bağlamı yakalamamasıdır. Çapraşıklık, bir konudaki kelimeler veya bir belgedeki konular arasındaki ilişkiyi yakalamaz. Anlamsal bağlam fikri, insan anlayışı için önemlidir. Bunun üstesinden gelmek için, bir konudaki kelimeler arasındaki bağlamı yakalamaya çalışan yaklaşımlar geliştirilmiştir. Bir konudaki kelimelerin bir arada bulunmasının olasılığı (log-likelihood) yerine koşullu olasılığı (likelihood) gibi ölçümler kullanırlar. Bu yaklaşımlar topluca 'tutarlılık' olarak adlandırılır. Tutarlılık, konu modellerini nicel olarak değerlendirmenin popüler bir yoludur.

Tutarlılık değeri makalesinde farklı yaklaşımlar bulunmaktadır. Bu yaklaşımlar arasında c_v, c_p, c_umass c_one-any, c_uci, c_npmi, c_a yöntemleri gösterilebilir. Bu yöntemler arasında tutarlılık değerini en iyi hesaplayan yöntem olarak c_v yöntemi gösterilmektedir. En kötü yöntem olarak da c_umass yöntemi gösterilmiştir. [31] Bu çalışma kapsamında tutarlılık yöntem olarak en iyi ve en kötü değerlendirme yöntemleri seçilip sonuçları değerlendirilecektir.

2.2. Yapılan Çalışma

Çalışmamızda izlenen adımlar dört başlıkta toplanabilir. Veri hazırlama, veri ön işleme, konu modelleme uygulanması ve sonuçların analizi şeklindedir. Çalışmanın amacı, Roberts, M. E. vd. [27] çalışmasını tekrarlayıp elde edilen sonuçlarla karşılaştırma yapmaktır. Tablo 2'de Roberts, M. E. vd. [27] makalesindeki çalışmanın sonuçları verilmiştir. Doğrudan bir karşılaştırma sağlamak için makaledeki veri kümesi ve kodlar aynı şekilde kullanılmıştır. Farklı yöntem eklenerek hem makaledeki sonuçlar hem de eklenilen yöntemlerin sonuçları değerlendirilecektir. Konu modelleme uygulanması adımında GDA, İKM ve YKM yöntemleri farklı method ve parametreler kullanılarak uygulanmıştır. Sonuç karşılaştırma işlemlerinde ise çalışma süresi, çapraşıklık ve tutarlılık değerleri hesaplanarak çalışmadaki yöntemlerin ne kadar başarılı olduğu oratay konulmuştur. Çalışmadaki işlem süreçleri Şekil 2. de gösterilmiştir



Şekil 2. Çalışma İşlem Adımları

Çalışmada GDA, İKM ve YKM konu yöntemleri kullanılmıştır. Yöntemleri çalıştırmak için R yazılım dili tercih edilmiştir. GDA ve İKM yöntemleri için topicmodels* kütüphanesi tercih edilmiştir. YKM yöntemi için stm† kütüphanesi tercih edilmiştir. Çalışmada kullanılan methodları ve değişkenleri seçerken Roberts vd. [27] çalışmasındaki methodlar ve değişkenler tercih edilmiştir. GDA için 4, İKM için 2 ve YKM için ise 3 adet farklı değişken değerleri kullanılmıştır. GDA yönteminde Varyasyonlu Beklenti Maksimizasyonu (Variational Expectation Maximization, VEM) ve Gibbs Örnekleme (Gibbs Sampling) methodu kullanılmıştır. İKM yöntemi için kütüphanede sadece Varyasyonlu Beklenti Maksimizasyonu (VBM) desteklenmektedir. Bu nedenle İKM yönteminde sadece VBM yöntemi kullanılmıştır. YKM yönteminde "init.type" parametresi için spektral (spectral) ve rastgele (random) metotları tercih edilmiştir.

Sonuçların değerlendirilmesi kısmında çalışma süresi değerlendirmesi için sadece yöntemin başladığı ve bittiği

*https://cran.rproject.org/web/packages/topicmodels/index.html

† https://www.structuraltopicmodel.com/

kod satırları arasındaki süre hesaplanmıştır. Çapraşıklık değerinin hesaplanması için stm kütüphanesi içerisindeki *eval.heldout* fonksiyonu kullanılmıştır. Çalışmada Roberts, M. E. vd. [27] makalesindeki kıyaslama metodunda yer alan değerlere tutarlılık değerleri de eklenerek değerlendirmelere katkı sağlanması amaçlanmıştır. Tutarlılık değerinin hesaplanması için python dilinde yazılmış gensim kütüphanesi kullanılmıştır. Tutarlılık değeri için *u_mass, c_v* tipleri kullanılmıştır.

YKM yönteminde tutarlılık değerinin hesaplanması adımında konulardaki kelimeleri tespit ederken dört farklı sonuç üzerinden tutarlık değerleri elde edilmiştir. Bunlar prob, frex, lift ve score olarak dörde ayrılmıştır. Prob, en yüksek olasılıklı kelimelerin matrisini verir. Frex, en yüksek dereceli kelimelerinin matrisini verir. Lift, en yüksek puanlı kelimelerin matrisini verir. Score, puana göre en iyi kelimelerin matrisini verir. Dört farklı sonucunda tutarlılık değerlerine nasıl etki ettiği de ayrıca incelenmiştir.

Çalışmada sonuçları düzgün değerlendirebilmek için sonuçlara etki edebilecek ortak değerler aynı tutulmuştur. Konu sayısı (K) tüm yöntemlerde 20 olarak belirlenmiştir. Tutarlılık değerlerini hesaplarken her konu modeli yöntemi için 10 adet kelime alınmıştır.

Tablo 2. Roberts, M. E. vd. [27] makalesindeki Çalışma Sonuçları

Yöntem	İterasyon Sayısı	Toplam Süre (dakika)	İterasyon süresi (saniye)	Çapraşıklık
İKM 1	34	96.7	170.7	-6.935
İKM 2	15	6.1	24.3	-7.040
YKM 1	21	0.7	1.9	-6.900
YKM 2	16	0.3	2.7	-6.905
YKM 3	280	7.2	1.5	-6.92

Yöntemde stm kütüphanesinin içinde hazır bulunan veri (poliblog5k) üzerinde modeller denenmiştir. Veri setinde 5000 satır ve 4 kolon (rating, day, blog, text) bulunmaktadır. Veri setinde örnek olması için Şekil 3.'de verilmiştir.

rating	day	blog	text
Conservative	182	ha	How happy do you think Team Barry is that, thanks
Conservative	299	ha	Tough stuff, but conservative passions this year
Liberal	345	db	Epic Ideological Failby digbyBefore you listen to
Conservative	90	ha	Excellent work as usual from a guy who s never
Conservative	321	at	The headline at the Los Angeles Times blog says
Conservative	271	at	To those of us of a certain age, Paul Newman will
Conservative	76	at	Barack Obama seems to have the power to make
Conservative	112	ha	If you thought that Barack Obama s
Liberal	71	tpm	The battle over Ferraro is rapidly heating up
Conservative	116	at	The New York Post is reporting that things are

Şekil 3. Poliblogs5k Veri Seti

3. BULGULAR

Çalışmada elde edilen tüm sonuçlar Tablo 3'de verilmiştir. Sonuçlar üzerinden hangi konu modellerinin daha başarılı olduğu tespit edilmeye çalışılmaktadır.

Roberts, M. E. vd. [27] makalesindeki calismalar birebir tekrarlanarak kendi yontemlerimizle birebir performans karsılastırması sağlanmıstır. Tablo 2'deki sonuçlar incelendiğinde çalışmadaki benzer yöntemlerin makale [27] ile aynı çapraşıklık değerine sahip olduğu tespit edilmiştir. Çalışma süreleri karşılaştırıldığında makalede [27] belirtilen sürelerden daha yavaş sonuçlar alınmıştır. Bu yavaşlığın çalışma ortamındaki donanım kaynaklı olabileceği düşünülmektedir. İKM yönteminde iki katına kadar süreler uzarken YKM yönteminde daha yakın sürelerde işlemler tamamlanmıştır. Çalışma sürelerinin ortama bağlı değişkenlik göstermesi YKM yönteminin İKM yöntemine göre tercih edilme sebeplerinden biri olarak gösterilebilir.

Makaleye[27] ek olarak GDA için 4 yöntem daha eklenip birlikte sonuçları tekrar karşılaştırdığımızda çapraşıklık değeri üzerinden değerlendirildiğinde YKM, İKM yöntemine göre; İKM ise GDA yöntemine göre daha iyi sonuçlar vermiştir.

Tutarlılık (coherence) değerine bakıldığında u_mass değeri için GDA tip 1 yöntemi daha iyi olduğu tespit edilmiştir. C_v tutarlılık değeri için YKM tip 1 ve tip 2 için frex olarak seçilen konu matrislerinde daha iyi sonuçlar verdiği tespit edilmiştir.

Sonuçlar, İKM yönteminin stm uygulamasının üstün performansını açıkça göstermektedir. Daha az iterasyon ve iterasyon başına sürelere bakıldığında daha iyi çözümler sunduğu görülmektedir.

Çapraşıklık değerine göre değerlendirildiğinde YKM yöntemi en iyi sonuçları vermiştir. U-mass tutarlılık değerine göre en iyi sonucu GDA tip 2 yöntemi vermiştir. C-v tutarlılık değerine göre en iyi değerleri YKM yöntemleri vermiştir.

4. TARTIŞMA VE SONUÇ

Konu modelleme yöntemleri son yıllarda çok fazla gelişme göstermiştir. Bu gelişmeler sonucunda modellerin etkinliğini tespit etmek oldukça önem kazanmıştır. Bu çalışma kapsamında konu modelleme yöntemlerinin karşılaştırılması için tutarlılık değeri kullanılmıştır. Aynı zamanda modele bağımlı kalmadan giriş ve çıkış değerleri ile konu modellerini değerlendirmenin bir yöntemi sunulmuştur. Uygulanan bu yöntem performans ve etkinlik açısından ileri de yapılacak çalışmalarda da kullanılabilir.

Tablo 3. Uygulanan Yöntemlerin Sonuçları

Yöntem		İterasyon Sayısı	Toplam Süre	İterasyon süresi	Çapraşıklık	u_mass	c_v
GDA tip 1		11	1.7 dk	9.4 sn	-7.195	-10.908	0.363
GDA tip 2		5	1.1 dk	12.9 sn	-7.195	**-10.608**	0.365
GDA tip 3		2000	2.7 dk	0.1 sn	-7	-16.135	0.460
GDA tip 4		2000	2.7 dk	0.1 sn	-7	-16.135	0.460
İKM tip 1		34	192.2 dk	339.2 sn	-6.935	-15.781	0.456
İKM tip 2		15	12.5 dk	49.8 sn	-7.04	**-14.958**	0.455
YKM tip 1	Frex	21	0.8 dk	2.4 sn	**-6.9**	-18.527	**0.508**
	Lift					-19.028	0.499
	Prob					-15.451	0.432
	score					-17.608	0.463
YKM tip 2	Frex	6	0.3 dk	3.3 sn	-6.905	-18.227	0.492
	Lift					-18.717	0.498
	Prob					**-14.871**	0.426
	score					-16.761	0.435
YKM tip 3	Frex	280	9.8 dk	2.1 sn	-6.918	-18.462	**0.509**
	Lift					-18.937	0.484
	Prob					-16.027	0.458
	score					-18.148	0.492

YKM modelinin çalışma süresi ve değerlendirme kriterleri açısından bakıldığında başarılı bir konu modelleme yöntemi olduğu görülmektedir. Sonraki çalışmalarda bu yöntem ile ilgili farklı veri kümeleri üzerinde çalışmalar yapılmalıdır. Aynı zamanda bu yöntemin değişkenlerinin sonuçlara olan etkisi üzerinde çalışmalar yapılabilir. YKM modelinde öne çıkan meta verilerin sonuçlara olan etkisi de incelenebilir.

Konu modelleme yöntemlerinde tutarlılık değerleri kıyaslama için çok önemlidir. Yöntemlerde sonuçların değerlendirilmesinde çapraşıklık değeri haricinde tutarlılık değeri de tercih edilebilir. Tutarlılık değerleri arasında c_v değeri u_mass değerine göre daha iyi sonuçlar verdiği gözlenmiştir. İleride yapılacak çalışmalarda c_v tutarlılık değeri tercih edilmelidir. Ayrıca diğer tutarlılık değerlerininde etkisi incelenebilir. Tutarlılık değerlerinin hesaplamalarındaki bölümlerin etkileri de incelenmelidir.

Çalışma kapsamında konu modelleme alanında farklı yöntemlerin birbirleri ile karşılaştırılması için yöntem ortaya konulmuştur. Konu modelleme alanında yapılan çalışmaların farklı yöntemlerde performans analizi yapılabilir. İlerleyen çalışmalarda farklı veriler ile benzer çalışmalar yapılarak elde edilen sonuçlar desteklenmelidir.

Finansman
Kar amacı gütmeyen herhangi bir kuruluştan çalışma ile ilgili fon alınmamıştır.

Çıkar çatışması
Çalışma ile ilgili herhangi bir kişi veya kurumla çıkar çatışmasının bulunmadığını yazarlar olarak onaylıyoruz.

KAYNAKÇA

[1] Metin madenciligi. http://www.metinmadenciligi.com (Erişim Tarihi: 19.03.2022).

[2] Alghamdi, R., Alfalqi, K. 2015. A Survey of Topic Modeling in Text Mining. International Journal of Advanced Computer Science and Applications, 6.

[3] Xiao, Z. 2014. CorrRank: Correlation based ranking topic model. Journal of Computational Information Systems, 10.

[4] Hilmi, M. F., Mustapha, Y.,Omar, M. T. C. 2020. Innovation in an Emerging Market: A Bibliometric and Latent Dirichlet Allocation Based Topic Modeling Study.

[5] Güven,Z., Diri, B., Çakaloğlu,T. 2019. Comparison of Topic Modeling Methods for Type Detection of Turkish News. 2019 4th International Conference on Computer Science and Engineering (UBMK), 150-154.

[6] Ekinci, E., Omurca, S. 2017. Ürün Özelliklerinin Konu Modelleme Yöntemi ile Çıkartılması. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 51-58.

[7] Negara, Edi Surya., Triadi, D. 2019. Topic Modelling Twitter Data with Latent Dirichlet Allocation Method. International Conference on Electrical Engineering and Computer Science (ICECOS), 386-390.

[8] Zhang, F., Gao, W., Fang, Y. , Zhang, B. 2020. Enhancing Short Text Topic Modeling with FastText Embeddings. Artificial Intelligence and Internet of Things Engineering (ICBAIE), 255-259.

[9] Blei, D., Lafferty, J. 2005. Correlated topic models. Advances in neural information processing systems, 18, 147.

[10] Blei, D., Lafferty, J. . 2007. A correlated topic model of Science. The Annals of Applied Statistics.

[11] Liu, L. &. 2019. Neural Variational Correlated Topic Modeling. WWW '19: The World Wide Web Conference, 1142-1152. 28

[12] Fu, X., Huang, K., Sidiropoulos, N. D., Shi, Q., Hong, M. 2018. Anchor-Free Correlated Topic Modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1-1.

[13] He, J., Hu, Z., Berg-Kirkpatrick, T., Huang, Y., Xing, E. P. . 2017. Efficient Correlated Topic Modeling with Topic Embedding. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 225-233.

[14] Xu, X., Shimada, A., Taniguchi, R. 2013. Correlated topic model for image annotation. In The 19th KoreaJapan Joint Workshop on Frontiers of Computer Vision , 201-208.

[15] Funnell, T., Zhang, A. W., Grewal, D., McKinney, S., Bashashati, A., Wang, Y. K., Shah, S. P. 2019. Integrated structural variation and point mutation signatures in cancer genomes using correlated topic models. PLOS Computational Biology, 15.

[16] Esmizadeh, Y., Canziani, B., Nemati, H.R., Mondaresnezhad, M. 2020. Sharing Economy: Application of Structural Topic Models.

[17] Roberts, M. E., Stewart, B. M., Tingley, D. 2019. stm : An R Package for Structural Topic Models. Journal of Statistical Software, 91.

[18] Liu, S., Yao, Y., Hu, Q. 2021. Characterization of Idea Relations in Text: Investigation with Topic Modelling and Structural Topic Modelling. 10.

[19] Bai, X., Zhang, X., Li, K. X., Zhou, Y., Yuen, K. F. 2021. Research Topics and Trends in the Maritime Transport: a Structural Topic Model. Transport Policy.

[20] Sim, S. H., Choi, H. G. 2016. The Structured Topic Model for E-Learning System. Advanced Science Letters.

[21] Ma, Y. 2021. A Structural Topic Model Analysis of Privacy in Mandarin Chinese News: 2010–2019. Proceedings of the Association for Information Science and Technology, 58(1):792-794.

[22] Hu, N., Zhang, T., Gao, B., Bose, I. 2019. What do hotel customers complain about? Text analysis using structural topic model. Tourism Management, 72, 417-426.

[23] Blei, D. M., Ng, A. Y. 2003. Latent dirichlet allocation. the Journal of machine Learning research , 3, 993-1022.

[24] Ekinci, E., Omurca, S. İ., KIRIK, E., TAŞÇI, Ş. 2020. Tıp Veri Kümesi için Gizli Dirichlet Ayrımı. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, 22(64), 67-80.

[25] towards data science https://towardsdatascience.com/intuitive-guide-tocorrelated-topic-models-76d5baef03d3 (Erişim Tarihi: 19.03.2022).

[26] Oo, M. K., Khine, M. A. 2020. Topic extraction of crawled documents collection using correlated topic model in mapreduce framework. arXiv preprint arXiv:2001.01669.

[27] Roberts, M. E., Stewart, B. M., Tingley, D. (2019). stm: An R Package for Structural Topic Models. Journal of Statistical Software, 91(2), 1–40.

[28] structural topic model https://www.structuraltopicmodel.com (Erişim Tarihi: 19.03.2022).

[29] Yıldıztepe, E. & Uzun, V. (2018). Olasılıksal Yöntemler ile Türkçe Metinlerin Anlamsal Benzerliğinin Belirlenmesi. Sinop Üniversitesi Fen Bilimleri Dergisi , 3 (2) , 66-78 . DOI: 10.33484/sinopfbd.350445

[30] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., Blei, D. 2009. Reading tea leaves: How humans interpret topic models. Advances in neural information processing systems, 22.

[31] Michael R., Andreas, B., Alexander H. 2015. Exploring the Space of Topic Coherence Measures. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15), 399–408.