

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/394784288>

Büyük Dil Modelleriyle Diplomatik Metin Analizi: Analysis of Diplomatic Texts with LLMs: A Hybrid Approach in Computational Social Sciences

Conference Paper · May 2025

DOI: 10.1109/SIU66497.2025.11112347

CITATIONS

0

READS

34

2 authors:



Ahmet Kurnaz

Çanakkale Onsekiz Mart University

10 PUBLICATIONS 16 CITATIONS

SEE PROFILE



Hamid Akın Ünver

Özyeğin University

70 PUBLICATIONS 462 CITATIONS

SEE PROFILE

Büyük Dil Modelleriyle Diplomatik Metin Analizi: Hesaplamalı Sosyal Bilimlerde Hibrit Bir Yaklaşım

Analysis of Diplomatic Texts with LLMs: A Hybrid Approach in Computational Social Sciences

Ahmet Kurnaz
Siyaset Bilimi ve Kamu Yönetimi Bölümü
Çanakkale Onsekiz Mart Üniversitesi
Çanakkale, Türkiye
ahmetkurnaz@comu.edu.tr

Akın Ünver
Uluslararası İlişkiler Bölümü
Özyeğin Üniversitesi
İstanbul, Türkiye
akin.unver@ozyegin.edu.tr

Özetçe—Bu çalışma, Türkiye Cumhuriyeti Dışişleri Bakanlığı tarafından çevrimiçi paylaşılan metinleri kullanarak büyük dil modellerinin metin madenciliği ve konu modelleme süreçlerindeki pratik uygulamasını sunmaktadır. Araştırmada, alan uzmanı ile büyük dil modellerinin birlikte çalıştığı döngüsel ve hibrit bir algoritma tanımlanmaktadır. Bulgular, iyi tanımlanmış temaların model başarısında kritik bir rol oynadığına işaret etmektedir. Sosyal bilimlerde, konu modelleme süreçlerinin yalnızca sayısal çıktılar üzerinden değerlendirilmesi yerine, araştırmacının veri kıymetlendirme süreçlerinde bağlamsal yorumlarını da içeren “terzi-dikim” bir yaklaşımla ele alınması gerektiği vurgulanmaktadır. Sonuç olarak, önerilen algoritma, metinlerin nitel veya nicel analizine uygun konu etiketleri üreterek, sosyal bilimsel araştırmalarda kapsamlı ve esnek bir analiz imkânı sunmaktadır.

Anahtar Kelimeler — konu modelleme; büyük dil modelleri; doğal dil işleme.

Abstract— This paper presents the practical application of big language models in text mining and topic modeling processes using texts shared online by the Ministry of Foreign Affairs of the Republic of Türkiye. The research describes an iterative and hybrid algorithm in which domain experts and large language models work together. The findings indicate that well-defined themes play a critical role in model success. It is emphasized that in social sciences, topic modeling processes should be handled with a “tailor-made” approach that includes the contextual interpretations of the researcher in the data evaluation processes, rather than being evaluated only through numerical outputs. In conclusion, the proposed algorithm offers a comprehensive and flexible analysis opportunity in social scientific research by generating topic labels suitable for qualitative or quantitative analysis of texts.

Keywords — topic modelling; LLMs; natural language processing.

I. Giriş

Konu modelleme algoritmaları yaklaşık yirmi yıldır içerik analizi için doğal dil işleme (DDİ) yaklaşımlarının merkezi

araştırma konularından bir tanesi olmuştur. Araştırmacı konu modelleme yöntemiyle, keşifsel ya da nedensel çalışmalarda araştırma sorusunu yanıtlamak için elindeki büyük ve etiketlenmemiş veri setlerini faydalı sınıflandırmalara indirger. Daha sonra yapacağı analizlerle (post-hoc) bu hesaplamalı sosyal bilimler yaklaşımını geleneksel nitel veya nicel yöntemlere bağlar.

Bu bağlamda LDA ile birlikte popülerlik kazanan algoritmalar zaman içerisinde STM gibi meta-verileri de kullanabilecek modellere evrilmiştir. Ancak transformers mimarisiyle birlikte [1] büyük dil modellerinin (BDM veya LLM) ortaya çıkışı DDİ'nin diğer alanlarında olduğu gibi içerik analizi için konu modellemede de bir sıçramaya yol açmıştır. Son yıllarda BDM tabanlı konu modelleme algoritmaları kısa ve uzun metinlerin çözümlemesi için daha sık kullanılmaktadır [2], [3], [4], [5], [6].

En yaygın kullanılan geleneksel modeller LDA [7] ve STM [8], LLM konu modellerine ise modelleri arasında BERTopic [9] ve TopicGPT [10] verilebilir. Geleneksel ve BDM konu modelleme algoritmalarının performansları son yıllarda oldukça sık karşılaştırılmaktadır [11], [12], [13], [14], [15]. Bu karşılaştırmalar metin ön-işleme ve hiper-parametre hassasiyeti, metin uzunluğunun ve farklı dillerin çıktılara etkisi, çıktılarının anlamsal bütünlük, ilgililik ve yorumlanabilirliği gibi boyutlarda yapılmaktadır. Algoritmaların işlemci ve RAM tüketimi gibi hesaplama performansları da kıyaslanırsa da sosyal bilimlerin merkezi problemlerinden bir tanesi değildir.

Her ne kadar BERTopic gibi BDM tabanlı algoritmalar kural temelli geleneksel modellere üstünlük kurmuş olsa da sosyal araştırmaların doğası gereği girdilerin ve çıktılarının sadece kantitatif değerlendirmeye tabi tutulması şu anda mümkün görünmemektedir. Buna ek olarak sosyal ilişkilerin metin ile aktarılamayan yönleri de bulunmaktadır. Sonuç olarak analizlerde girdi olarak kullanılan metinlerin kalite ve içeriğinin tespit edilmesi ve çıktılarının anlamlılık düzeylerinin değerlendirilmesi için insan aklının çıkarım yapabilen bilişsel

979-8-3315-6655-5/25/\$31.00 ©2025 IEEE

kapasitesine, alan uzmanının derinlemesine ve bağlamlar arası ilişkisel bilgisine günümüzde hala ihtiyaç duyulmaktadır [16].

Bu nedenlerden ötürü sosyal metin analizinde araştırma sorusunu yanıtlamada agnostik kabullerin daha doğru olduğuna yönelik yaklaşımı benimsemektediriz [17]. Araştırma sorusu, araştırmacının ulaşabildiği ve kıymetlendireceği (data-repurposing) verinin boyutu, kalitesi ve dokusu, araştırma yapılan disiplinin doğası gibi dış etmenler seçilecek modelin performansını doğrudan etkileyeceği için modelleri doğrudan kıyaslamak yerine tüm modellere şüpheli yaklaşmak ve onları evrensel doğrular yerine faydalı birer alet olarak görmenin gerekli olduğuna inanıyoruz.

Yukarıdaki açıklamalar ışığında araştırmacıların kendi çalışmalarına uygun modeli geliştirmelerinin onları terzi-dikim sonuçlara ulaştırabileceğini değerlendirmekteyiz. Günümüzde yapay zeka teknolojilerinde yaşanan gelişmeler ve HuggingFace gibi platformlar üzerinden tüm DDİ sorunlarının çözümüne yönelik ve açık kaynak olarak paylaşılan büyük dil modelleri sayesinde model geliştirmek çok daha kolay hale gelmiştir.

Bu çalışmada siyaset bilimi ve uluslararası ilişkiler alanında HuggingFace üzerinden paylaşılan ön-egitimli (pre-trained) modellerle terzi-dikim konu modelleme algoritmasının nasıl oluşturulabileceğine dair bir örnek sunmaktayız.

II. YÖNTEM

A. Veri

Bu çalışmada Dışişleri Bakanlığı web sitesi üzerinden (mfa.gov.tr) kamuoyuna açık biçimde paylaşılan veriler kullanılmıştır. Paylaşılan verilerin kategorileri *güncel açıklamalar, güncel gelişmeler, basın toplantıları, açıklamalar, bakanlık duyuruları ve basın bilgi notlarıdır*. 2002 ve 2024 yıllarını kapsayan toplam 8,304 internet sayfasından dinamik web kazıma yöntemiyle elde edilmiştir. HuggingFace üzerinden model seçiminde daha fazla seçeneğe sahip olmak için tüm metinler İngilizce olarak toplanmıştır.

İçerik analizini yapabilmek için veriyi çeşitli ön-işleme tabii tutmamız gerektiğine karar verdik. Bunun öncelikli sebebi metinlerin çok farklı uzunluklara sahip olmasıydı. Bir belgede geçen konu ve duygu yoğunluğunu tespit ederken uygulayacağımız normalizasyon süreçlerinin analizini açıklanmasını zorlaştıracağını değerlendirdik. Buna karşın cümle düzeyinde yapılacak bir analiz bağlamdan kopması sebebiyle tercih etmedik. Sonuç olarak internet sayfasındaki tüm metin yerine paragrafları analiz etmeye karar verdik.

Bu karar, bir paragraftan diğerine geçişin genellikle bağlamda bir değişime işaret ettiği gözlemine dayanıyordu. Ayrıca, analiz birimi olarak cümlelerin kullanılması, bağlamın doğru bir şekilde çıkarılmasını engelleyebileceği için tercih edilmemiştir. Bunun en önemli sebebi izole edilmiş cümlelerin diplomatik dilin doğasından kaynaklanan zincirleme anlatıyı inşa sürecini içermemesidir. Örneğin, diplomatik ilişkileri tartışan olumlu bir cümle, belirli bir olaya atıfta bulunan sonraki olumsuz bir cümle ile yan yana gelebilmekte veya olumlu olarak başlayan ifadeler daha sonra olumsuz bir olaya işaret edebilmektedir. Bu yüzden paragraf bütünlüğünün orijinal bağlamı koruduğu gözlemine dayanarak ve metin

uzunluğunda standartlaşmayı da sağladığı için cümle veya tüm metinleri kullanmak yerine paragrafları tercih ettik.

Analiz birimlerine karar verdikten sonra metinleri basit bir ön işleme tabii tutarak paragraflar şeklindeki analiz birimlerine indirgeedik. Bu adım sonrası “belgeler” veya “veri” olarak anılacak olan toplam 43.129 analiz birimi ortaya çıkarmıştır.

B. Yaklaşım

Bu çalışmada farklı büyük dil modellerinin ve alan uzmanının birlikte çalıştığı bir algoritma geliştirdik. Araştırmacı konuların çerçevesinin çizilmesi ve konu sayısının belirlenmesi noktasında girdilerde bulunurken büyük dil modelleri konu etiketleme için vektörizasyon ve sınıflandırma işlemlerini yerine getirmektedir. Araştırmacının bu yarı-yönlendirilmiş hibrit sistemdeki rolünü belirleyen King vd.’nin bilgisayar destekli anahtar kelime belirleme algoritmasından esinlendik [18].

İçerik analizi algoritmamızda büyük dil modelleri kullandığımız için kapsamlı bir metin ön işleme yapmadık. Genel olarak metin içine karışmış HTML etiketlerinin silinmesi, kısaltmalardaki noktalıların kaldırılması, liste numaralandırma ve harflendirmelerinin kaldırılması, Roma rakamlarının ve fazla boşlukların silinmesini içermektedir. Lemmatizasyon veya budama (stemming) gibi işlemler uygulanmamıştır.

Konuları sınıflandırırken farklı algoritmaların zincirleme biçimde kullanılmasının daha işlevsel olduğunu tecrübe ettiğimiz için bir soru-cevap (SC) modeli ve bir genel transformer modelini birlikte kullandık.

SC modelini araştırmacının farklı konuları tespit etmesi için kullandık. Buna göre her bir belgenin ana fikrini türetmek için belgeye soru sorduğumuz bir yaklaşım benimsedik. Modele “bu belgede bahsedilen ana konu nedir” sorusunu sorduk. Bu soruyu yanıtlamak için *deepset/roberta-base-squad2* modeli kullandık. Elde edilen cevaplar araştırmacı tarafından manuel olarak incelenmesi ve konuların çerçevesinin çizilmesi için temel oluşturmuştur.

SC modelinden bağlamsal verilerin çıkarılmasının ardından, bu bağlamları farklı temalar halinde sınıflandırmak ve standartlaştırmak için yinelemeli ve döngüsel bir prosedür uyguladık. Bunun için yeni tematik kategorilerin ortaya çıkmadığından emin olmak için 43.129 benzersiz bağlamın yaklaşık %5’ini manuel olarak inceledik. Bu süreç, 23 tema altında gruplanan 101 kategorik anahtar kelimenin belirlenmesiyle sonuçlanmıştır.

Son olarak, belgelere konu etiketleri atamak için kelime vektörlerini (embeddings) ve tam eşleşme puanlamasını içeren özel bir konu sınıflandırma algoritması geliştirdik. Algoritmamızda *sentence-transformers/all-mpnet-base-v2* modelini kullandık. Konuları sınıflandırırken kelime benzerlikleri ve tam kelime eşleştirmesinden gelen ağırlıklı puanları birleştirerek her belge için tüm konu puanlarını hesapladık. Bu ham puanlar 0 ile 1 arasında olacak şekilde normalleştirilerek ve konu dağılımı çerçevesi üretilmiştir.

Daha sonra normalleştirilmiş puanlara 0,5’lik bir eşik uygulanmıştır - yalnızca 0,5’e eşit veya daha yüksek puanlara

sahip konular ilgili belgeye atanmıştır. Böylece bir belgede birden fazla konu olabileceği gibi hiçbir konuyla ilgili olmayan belgelerin de bulunduğu karma-üyelik (mixed-membership) çıktılarına ulaştık.

C. Algoritma

Geliştirdiğimiz konu sınıflandırma algoritmasının adımları aşağıdaki şekilde ifade edilebilir:

(1) N adet belge olduğunu varsayalım ve dokümanlar $i=1, \dots, N$ şeklinde sıralanır. Tematik kategoriler kümesi C ve her $c \in C$ için, K_c o kategoriye ait anahtar kelimeler kümesini ifade etmektedir. Kelime benzerlikleri için verdiğimiz ağırlık α ve tam eşleşme ağırlığı da β ile gösterilmektedir.

(2) Her $c \in C$ kategorisi için, temayı temsil eden tek bir vektör, anahtar kelimelerin birleştirilip gömülü modele kodlanmasıyla elde edilir. Burada, $\text{concat}(K_c)$ ifadesi K_c içindeki tüm anahtar kelimelerin birleştirilmiş halini belirtir.

$$e_c = \text{encode}(\text{concat}(K_c))$$

(3) Her metin parçası belge i için, vektörel değer (embeddings) hesaplanır:

$$e_i = \text{encode}(\text{belge}_i)$$

(4) Her metin parçası i ve kategori c için, belge vektörü ile tema vektörü arasındaki kosinüs benzerliği hesaplanır. Bu $S_{i,c}$ değeri, i numaralı parça için C kategorisinin gömülü benzerlik skorunu oluşturur.

$$S_{i,c} = \cos(e_i, e_c) = \frac{e_i \cdot e_c}{\|e_i\| \|e_c\|}$$

(5) Her parça i ve kategori C için, belge i içerisinde K_c kümesindeki tüm anahtar kelimelerin frekansı sayılır. Bir anahtar kelime k 'nin belge i içerisindeki sayısını $f_{i,k}$ olarak ifade edelim. Böylece, i numaralı parça için C

$$r_{i,c} = \sum_{k \in K_c} f_{i,k}$$

kategorisinin ham anahtar kelime skoru:

(6) Her parça i için, tüm kategoriler arasındaki maksimum ham anahtar kelime skoru:

(7) Daha sonra, i için C kategorisinin normalleştirilmiş anahtar kelime skoru. Buradaki $\max(M_i, 1)$ ifadesi, sifra

$$M_i = \max_{c \in C} r_{i,c}$$

$$k_{i,c} = \frac{r_{i,c}}{\max(M_i, 1)}$$

bölünmeyi önlemek amacıyla kullanılmıştır.

(8) Her parça i ve kategori C için, gömülü benzerlik skoru ile normalleştirilmiş anahtar kelime skoru, α ve β ağırlıkları kullanılarak birleştirilir. Bu $S_{i,c}$ değeri, i numaralı parça için C kategorisinin genel skorunu oluşturur.

$$S_{i,c} = \alpha S_{i,c} + \beta k_{i,c}$$

(9) Her parça i için, en yüksek birleşik skora sahip kategori seçilir:

$$C_i^* = \operatorname{argmax}_{c \in C} S_{i,c}$$

(10) Böylece, belge i 'ye atanan kategori C_i^* olur ve i numaralı parça için çıktı şu şekilde tanımlanır:

$$(\text{belge}_i, C_i^*, \{S_{i,c}\}_{c \in C})$$

(11) Algoritma, her belge i ve kategori C için iki bileşene dayalı olarak birleştirilmiş bir skor $S_{i,c}$ hesaplar. Gömülü benzerlik skoru: $s_{i,c}$, parça gömüsü ile kategori gömüsü arasındaki kosinüs benzerliği olarak hesaplanır. Normalleştirilmiş anahtar kelime skoru: $k_{i,c}$, anahtar kelime sayımlarının normalleştirilmesiyle elde edilir. Bu skorlar, aşağıdaki şekilde ağırlıklandırılarak toplanır:

$$S_{i,c} = 0.7 \cdot s_{i,c} + 0.3 \cdot k_{i,c}$$

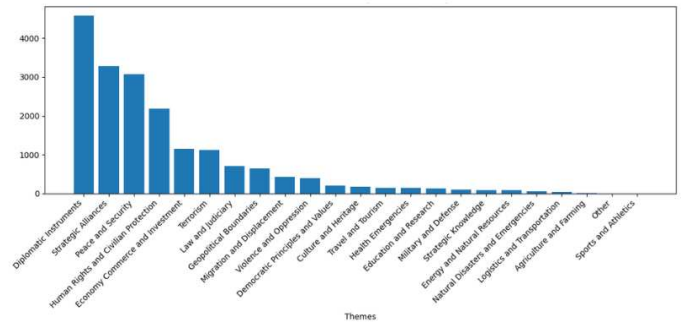
III. BULGULAR

Her ne kadar kullandığımız veriyle çok başarılı sonuçlar üretse ve kullanışlı olsa da bu algoritmanın başarısının sırrı 23 temanın anlamlı ve kavramsal olarak bütünlük sağlayan kelime setleriyle tanımlanmış olmasında yatmaktadır. Farklı durumlarda temaların iyi tanımlanamaması durumunda daha küçük ve anlamsal olarak bütünlük sağlayan tema setleri tanımlanabilir.

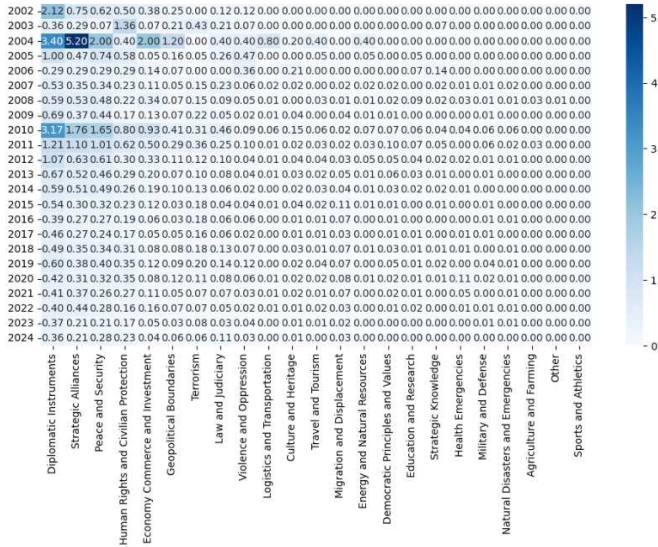
Bir başka durum ele alınan metinlerin Dışişleri Bakanlığı verisi olması sebebiyle bütün belgelerde (paragraflarda) çok benzer ifadelerin kullanılması algoritmanın başarısına katkıda bulunmuştur. Farklı bağlamlarda kullanıldığında araştırmacının farklı kaynaklardan topladığı verilerle test etmesi gereklidir.

Son olarak farklı uzunluklardaki metinlerle çalışırken normalizasyon mekanizmasına dikkat edilmesi gereklidir. Bizim elimizdeki veriyi aşağı-yukarı benzer uzunluklara indirgeme şansımız olduğu için maksimum kelime sayısı ile normalize ettiğimizde başarılı sonuçlar elde ettik. Fakat çok uzun ve çok kısa metinlerin bir arada kullanılması sonuçları etkileyecektir.

Elde ettiğimiz konuları aşağıdaki örneklerde olduğu gibi birikimli veya zaman serisi olarak görselleştirilebilir. Şekil 1'de tüm algoritmanın ürettiği puan değeri 0.5'i aşan temaların frekansı verilmiştir. Şekil 2'de ise zaman içinde bir temanın görünümündeki değişikliklerle gösterilmektedir.



Şekil 1 – Temaların Frekans Dağılımı



Şekil 2 – Yıllara Göre Tema Yoğunluğunun Isı Haritası

IV. SONUÇ

Geleneksel ve/ya kural tabanlı yaklaşımların yerini büyük dil modellerine dayanan konu modelleme algoritmaları almaktadır. Yapılan tüm karşılaştırmalarda transformers mimarisine geliştirilmiş modellerin üstünlükleri ortaya koyulmuştur.

Ancak sosyal metin analizine agnostik yaklaşım model karşılaştırmalarının her durum için yeterli sonuç üreteceğine olan şüpheyi de beraberinde getirmektedir. Araştırmacının model seçimindeki rolü sadece sayısal değerler üzerinden kıyaslama yapmanın ötesinde anlamlara dayanan bilişsel işlevleri de içermektedir. Bu yüzden araştırmaya uygun terzidikim algoritmaların geliştirilmesi çok daha iyi sonuçlar verme potansiyelini taşımaktadır.

HuggingFace üzerinden açık erişimli olarak paylaşılan ve DDİ'nin farklı sorularını yanıtlamak için ön eğitim süreçlerinden geçirilmiş modeller sosyal bilimlerde araştırmalarında araştırmacıya daha fazla esneklik sağlamaktadır. Böylece sosyal araştırmacı bu modelleri bir yap-bozun parçaları gibi kullanabilir.

Bu çalışmada Türkiye Cumhuriyeti Dışişleri Bakanlığı web sitesinde yayınlanan metinleri kullanarak iki farklı modelin içerik analizinde konu sınıflandırması için nasıl kullanılabileceğine bir örnek verdik. Sonuçta anlamsal bütünlüğe sahip ve verinin içinde geçen ve bir eldiven gibi tam olarak kapsayan sayıda konu sınıflandırılarak bir sonraki analiz aşamasına hazır hale getirilmiştir. Bundan sonra bu konu modelleri nitel veya nicel olarak analiz edilebilir, açıklayıcı ya da nedensel çıkarımlar için kullanılabilir.

KAYNAKLAR

[1] A. Vaswani vd., "Attention Is All You Need", 2017, *arXiv*: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.
 [2] A. Najafi ve O. Varol, "TurkishBERTweet: Fast and Reliable Large Language Model for Social Media Analysis", 29 Kasım 2023, *arXiv*: arXiv:2311.18063. doi: 10.48550/arXiv.2311.18063.

[3] T. Schmiedel, O. Müller, ve J. vom Brocke, "Topic Modeling as a Strategy of Inquiry in Organizational Research: A Tutorial With an Application Example on Organizational Culture", *Organ. Res. Methods*, c. 22, sy 4, ss. 941-968, Eki. 2019, doi: 10.1177/1094428118773858.
 [4] L. Valentine, S. D'Alfonso, ve R. Lederman, "Recommender systems for mental health apps: advantages and ethical challenges", *Ai Soc.*, ss. 1-12, Oca. 2022, doi: 10.1007/s00146-021-01322-w.
 [5] R. P. da Silva, J. T. Polletini, ve A. Pazin Filho, "Unsupervised natural language processing in the identification of patients with suspected COVID-19 infection", *Cad. Saude Publica*, c. 39, sy 11, s. e00243722, 2023, doi: 10.1590/0102-311XPT243722.
 [6] "User Voice Mining and Prediction: Topic Evolution and Prediction on E-commerce Using Hybrid BERTopic-LDA-RF Model". Erişim: 07 Mart 2025. [Çevrimiçi]. Erişim adresi: <http://ouci.dntb.gov.ua/en/works/IRMJKbP4/>
 [7] D. M. Blei, A. Y. Ng, ve M. I. Jordan, "Latent Dirichlet Allocation", *J. Mach. Learn. Res.*, c. 3, sy Jan, ss. 993-1022, 2003.
 [8] M. E. Roberts, B. M. Stewart, ve D. Tingley, "stm: An R Package for Structural Topic Models", *J. Stat. Softw.*, c. 91, sy 1, Art. sy 1, Eki. 2019, doi: 10.18637/jss.v091.i02.
 [9] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure", 11 Mart 2022, *arXiv*: arXiv:2203.05794. Erişim: 03 Ocak 2024. [Çevrimiçi]. Erişim adresi: <http://arxiv.org/abs/2203.05794>
 [10] C. M. Pham, A. Hoyle, S. Sun, P. Resnik, ve M. Iyey, "TopicGPT: A Prompt-based Topic Modeling Framework", 01 Nisan 2024, *arXiv*: arXiv:2311.01449. doi: 10.48550/arXiv.2311.01449.
 [11] O. El-Gayar, M. Al-Ramahi, A. Wahbeh, T. Nasrallah, ve A. Elnoshokaty, "A Comparative Analysis of the Interpretability of LDA and LLM for Topic Modeling: The Case of Healthcare Apps", 2024.
 [12] R. Egger ve J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts", *Front. Sociol.*, c. 7, s. 886498, 2022, doi: 10.3389/fsoc.2022.886498.
 [13] O. Babalola, B. Ojokoh, ve O. Boyinbode, "Comprehensive Evaluation of LDA, NMF, and BERTopic's Performance on News Headline Topic Modeling", *J. Comput. Theor. Appl.*, c. 2, sy 2, Art. sy 2, Kas. 2024, doi: 10.62411/jcta.11635.
 [14] A. Amaro ve F. Bacao, "Topic Modeling: A Consistent Framework for Comparative Studies", *Emerg. Sci. J.*, c. 8, sy 1, Art. sy 1, Şub. 2024, doi: 10.28991/ESJ-2024-08-01-09.
 [15] A. Kaur ve J. R. Wallace, "Moving Beyond LDA: A Comparison of Unsupervised Topic Modelling Techniques for Qualitative Data Analysis of Online Communities", 19 Aralık 2024, *arXiv*: arXiv:2412.14486. doi: 10.48550/arXiv.2412.14486.
 [16] Stanford Online, Andrew Ng and Chris Manning *Discuss Natural Language Processing*, (07 Haziran 2023). Erişim: 17 Temmuz 2024. [Çevrimiçi Video]. Erişim adresi: <https://www.youtube.com/watch?v=6w0Po83ZmjA>
 [17] J. Grimmer, B. M. Stewart, ve M. E. Roberts, *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton, 2021.
 [18] G. King, P. Lam, ve M. E. Roberts, "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text: KEYWORD AND DOCUMENT SET DISCOVERY", *Am. J. Polit. Sci.*, c. 61, sy 4, ss. 971-988, Eki. 2017, doi: 10.1111/ajps.12291.