Sosyal Bilimlerde Dönemsel Konu Modelleri: STM Üzerine Bir Çalışma
Temporary Topic Models in Social Sciences: A Study on STM

2022 30th Signal Processing and Communications Applications Conference (SIU) | 978-1-6654-5092-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/SIU55565.2022.9864923

Ahmet Kurnaz
Siyaset Bilimi ve Kamu Yönetimi Bölümü
Çanakkale Onsekiz Mart Üniversitesi
Çanakkale, Türkiye
ahmetkurnaz@hotmail.com

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

***Özetçe* —Konu modelleri son yıllarda sosyal bilimlerde hızla popülerleşmektedir. Ancak konu modellerini kullanırken araştırmacıların dikkat etmesi gereken çeşitli noktalar vardır. Bunlar arasında kullanılacak metnin biçim ve içerik özellikleri, dili, eşdeğişkenlerin varlığı, ön işleme adımları sayılabilir. Bu araştırmada farklı tipteki veri setleri ve ön işleme adımları kullanılarak Structural Topic Modelling (STM) algoritmasıyla elde edilen sonuçlar özetlenmiştir. Sonuç olarak STM kullanırken araştırma sorusuna bağlı olarak ön işleme adımları dramatik farklılıklar yaratabilmektedir. Ayrıca farklı dillerdeki metinlerin aynı veri setinde kullanılması konu modelinin performansı üzerinde olumsuz etkiye sahiptir. Kısa ve uzun metinler arasında performans farkı araştırmacılar tarafından tespit edilmiştir. Son olarak eşdeğişkenlerin model içinde kullanımıyla sosyal bilimlerde işlevselliği artmaktadır.**

Anahtar Kelimeler—konu modelleri, STM, içerik analizi, metin madenciliği, sosyal medya

***Abstract*—Topic models are rapidly becoming popular in social sciences. However, researchers should pay attention to some critical steps while using these models. The format and content of the textual data, language, existence of covariates, and preprocessing steps are the most crucial elements of a topic model analysis. This study inspects the effect of various datasets and preprocessing steps on Structural Topic Models (STM). Results shows that preprocessing, which depends on the research question, profoundly affects the model performance. Besides, the existence of multilingual data weakens the topic quality. Also, the algorithm performance is different among long and short texts. Last, the potential usage of covariates in the model enhances its functionality in social science.**

Keywords—topic models, STM, content analysis, text mining, social media

I. Giriş

Yazılı ve sözlü iletişimden devşirilen metinler sosyal bilimler araştırmaları için en merkezi ve önemli veri kaynaklarının başında gelmektedir. Ancak dijital veriye erişim imkanları arttıkça analiz edilmesi gereken veri devasa boyutlara ulaşmış ve bunlarin nitel tekniklerle incelenmesi imkansız hale gelmiştir. Metinleri sınıflandırma görevleri basit olsalar bile problem öznel olduğunda çoklu kodlayıcıları (*crowdsource*) kullanmak mümkün değildir [1]. Bu yüzden, nitel incelemelerin yerini alması amacıyla değil ancak onları desteklemek için otomatik içerik analizi yöntemleri geliştirilmekte ve her geçen gün daha da sofistike hale getirilmektedir [2]–[4].

En popüler otomatik içerik analiz yaklaşımlarının başında konu modelleme gelmektedir. Konu modelleme algoritmaları, belirli sayıda dokümanı girdi olarak alan ve çıktı olarak da bu doküman setinin hangi konulardan meydana geldiğini dönen yönlendirilmemiş makine öğrenmesi algoritmaları olarak tanımlanabilir. Geçmişi Latent Semantic Analysis'e kadar götürülse de bilinen anlamda ilk konu modelleme algoritması Latent Dirichlet Allocation (LDA) olarak kabul edilir. Takip eden yıllarda LDA'in varyasyonları başta olmak üzere farklı çalışma prensiplerine sahip algoritmalar geliştirilirken algoritmaların hesaplama performansları da yıllar içinde iyileştirilmiştir [5].

Konu modelleme algoritmalarını kullanarak yeniden üretilebilir ve genellenebilir çalışmalar yapmak, onları birbirleriyle kıyaslamak oldukça zordur. Bunun için bazı nicel metrikler geliştirilmiş olsa da son tahlilde hala algoritma ve konu sayısı seçiminde araştırmacının değerlemesine ihtiyaç vardır [3].

Sosyal bilimler çalışmalarında konu modelleme algoritmalarını seçerken ve kullanırken dikkat edilmesi gereken pek çok önemli husus vardır. Bunlardan birincisi incelenecek metnin uzunluğu ve yapısıdır. Kitaplar, konuşma metinleri, tutanaklar, gazeteler veya sosyal medya paylaşımları gibi farklı uzunluklara, amaçlara ve içeriklere sahip yazılı dokümanların sınıflandırılması ve anlaşılması için farklı algoritmalar seçmek gereklidir. Buna ek olarak eğer araştırmacı zamana bağlı bir çalışma yapıyorsa dönemsel (temporal) algoritmalar arasından birini tercih etmesi gerekir. Ayrıca analiz edilecek metinlerin önişleme adımları, konu sayısının seçimi gibi algoritma parametrelerinin doğru ayarlanması, analiz sonrası nicel ve nitel doğrulama adımları konu modellerinin kalitesini, yeniden üretilebilirliğini ve genellenebilirliğini doğrudan etkiler. Son olarak, konu modelleme ve diğer otomatik içerik analizinde incelenen dilin özellikleri de göz önünde bulundurulmalıdır. Örneğin, İngilizce ve Türkçe doğal dil işleme adımları birbirinden farklılık gösterir [6]. Farklı dillerdeki metinleri bir arada analiz etmek için, otomatik çeviri araçlarını kullanmak gibi, çeşitli stratejiler kullanılabilir [7]–[9].

978-1-6654-5092-8/22/$31.00 ©2022 IEEE

Bu araştırmada farklı uzunluklarda, farklı içerik özelliklerine sahip, ve dönemsel veriyle oluşturulabilecek konu modelleri ve ön işleme adımlarının modele etkisi incelenmektedir. Konu modelleme için metverinin algoritmaya dahil edilebilirliği sayesinde zamana bağlı çıkarımların yanı sıra kategorik değişkelerin de kullanılabildiği ve performans yönünden LDA gibi popüler algoritmaların önünde olan Structural Topic Models (STM) seçilmiştir [10]. STM algoritmalasının temel özellikleri açıklandıktan sonra veri toplama ve temizleme adımları açıklanmıştır. Daha sonra STM uygulamalarına ilişkin sonuçlar özetlenmiştir. Çalışma ilgili algoritmanın artıları ve eksilerinin tartışmasıyla son bulmaktadır.

STM kelime sayıları üzerinde işleyen üretken (*generative*) bir karma aidiyet (*mixed-membership*) modelidir. STM algoritmasını diğerlerinden ayıran en önemli özelliklerinin başında her döküman için eşdeğişkenlerin (*covariate*) de modele dahil edilebilmesi gelmektedir. Böylece model sonuçları hipotez testi için kullanılabilir. Model için geliştirilmiş bir R paketi bulunmaktadır [10].

STM 2013 yılında ortaya çıkışından günümüze farklı uzunluklardaki ve farklı yapıya sahip metinleri analiz etmek için kullanılmıştır. Karşılaştırmalı siyaset çözümlemelerinde çok dilli metinler [11], açık uçlu anket soruları [12], [13], yargı mensuplarının sosyal medya verileri [14], toplumsal kimlik analizleri bağlamında TED konuşmaları [15], terör saldırılarına verilen sosyal medya tepkileri [16], kentleşme çalışmalarında yatırımcılarla yapılan yarı yapılandırılmış mülakatlar [17], tarihi dönemlerin anlaşılması için dönem metinleri [18], göç krizinin bağlamında parlamento tutanakları [19], yabancı firmalara yönelik yanlılık bağlamında gazeteler [20] ve daha pek çok farklı türde metnin farklı bağlamlarda çözümlenmesi için kullanılmıştır.

II. VERI

Araştırma için NATO'nun resmi web sitesi ve Twitter hesapları üzerinden veri toplanmıştır. Daha önce benzer veri setleri belge özetleme [21], duygu analizi [22], doğal dil işleme ve belge sınıflandırma [23], ve konu modellemeyle içerik analizi [24] yapmak amacıyla kullanılmıştır.

Uzun metinler konuşma metinleri (speeches), basın açıklamaları (press releases), görüşler (reviews), resmi metinler (official texts), arşivler (archives) ve yayınlardan (publications) oluşan 3921 dokümandır. İlgili metinler web sitesi üzerinden web kazıma yöntemiyle toplandıktan içinde 300 karakterden daha az içerik olan metinler silinmiştir. Sonuçta 1941-2021 yılları arasındaki 76 periyodu (yıl) kapsayan 3844 dokümana ulaşılmıştır.

Kısa metinler NATO'nun 42 resmi Twitter hesabından 2014-2021 yılları arasında atılmış 264,071 İngilizce tweeti içermektedir. Metinler ay ve yıla göre birleştirilmiştir. Bu şekilde veri, her ay-yıl ikilisi için o ayda atılmış tweetlerin olduğu bir metin dokümanı olacak şekilde düzenlenmiştir.

Kısa ve uzun metin verilerinden rakam ve noktalama işaretleri silindikten sonra tamamı latinize edilmiştir. Araştırmada konu modelleme algoritmalarının ön işlem yapılmasına göre performansları da incelendiği için iki metin verisinin ikişer farklı versiyonu kullanılmıştır. Birinci tipte yukarıdakilere ek herhangi bir işlem yapılmamıştır. Bu veriler *ham* olarak gösterilmektedir. İkinci versiyonda ise kelimeler köklerine indirgenmiş (*stemming*) ve işlevsiz kelimeler (*stop words*) çıkarılmıştır. İşlevsiz kelimeler için SMART sözlükçesinin İngilizce versiyonu kullanılmıştır. Bu veri setleri önişlemden geçtiği için *ön* olarak gösterilmiştir. Sonuçta dört metin verisi Tablo-1'de gösterilmiştir.

	Kaynak	Yöntem	Ön İşleme	Belge	Terim
NATO_uzun_ham	web sitesi	web kazıma	Temel	3844	10270
NATO_uzun_ön	web sitesi	web kazıma	Sınırlı	3844	6174
NATO_kısa_ham	Twitter	dijital dinleme	Temel	147	10789
NATO_kısa_ön	Twitter	dijital dinleme	Sınırlı	147	7940

TABLO I: Veri setlerini gösterir tablo

Son olarak tüm veri setleri STM algoritmasına sokulmadan önce en az 10 dokümanda yer almayan veya belgelerin %70'inden daha fazlasında yer alan kelimeler veriseti sözcükçelerinden (*vocabulary*) çıkarılmıştır. Sonuç olarak herbir veriseti aynı parametre ayarları kullanılarak ayrı ayrı STM algoritmasıyla modellenmiştir.

III. STM UYGULAMASI

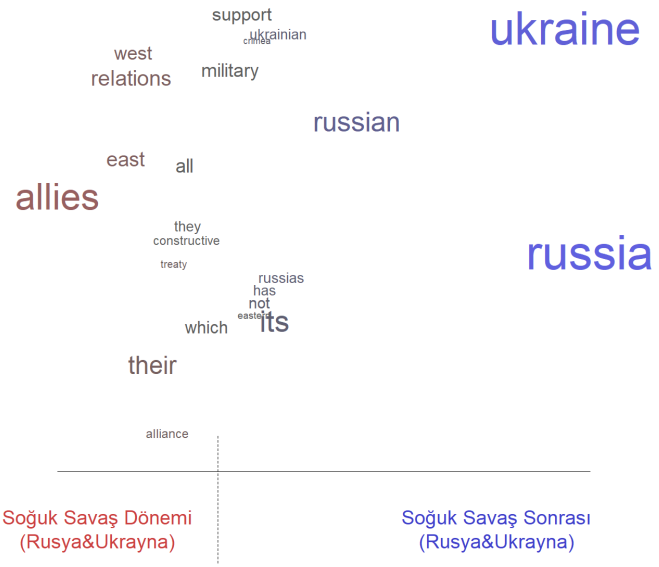
STM uygulamasında R için geliştirilmiş kütüphaneden yararlanılmıştır. Ayrıca doküman-terim matrisinin hazırlanması için de quanteda paketi kullanılmıştır. Konu sayısı seçiminde anlamsal bütünlük (*semantic coherence*), artanlar (*residuals*) ve dışarıda tutulanların kestirimi (*held-out likelihood*) parametreleri incelenmiştir. Tüm veri setleri için modellerin 50 konu çevresinde en iyi sonuçlarına yaklaştığı değerledirilmiştir. Veriler tarihlerine göre Soğuk Savaş ve Soğuk Savaş Sonrası şeklinde dönemsel ikili bir değişkenle etiketlenmiştir. Model eğitilirken bu ikili kategorik değişkenin yanı sıra yıl bazında tarihler de eşdeğişken olarak kullanılmıştır.

STM algoritmasıyla sosyal bilimlerde elde edilebilecek sonuçları örnekledirmek için NATO'nun resmi web sitesinden toplanan verilerin işlenmemiş halleri üzerinden oluşturulmuş konu modeline ilişkin grafikler Şekil-1, Şekil-2 ve Şekil-3'te gösterilmiştir.

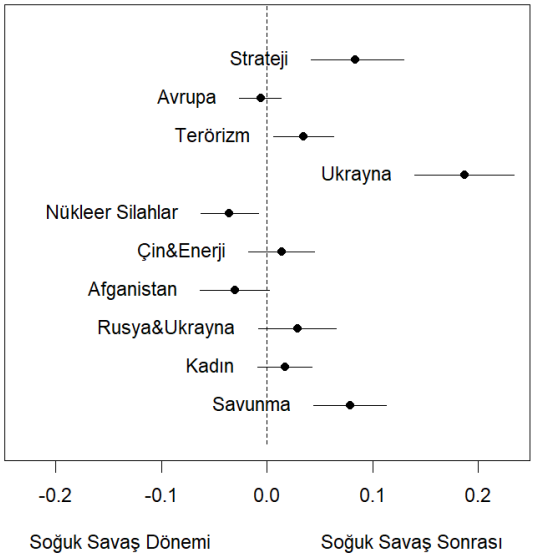
Şekil 1'de Doğu Avrupa ile ilgili konudaki kelimelerin dönemsel dağılımı gösterilmiştir. Dönemler algoritmaya kategorik değişken olarak verilmiştir. Kırmızı renkli terimler soğuk savaş dönemiyle ilişkiliyken mavi renkli kelimeler soğuk savaş sonrası dönemde öne çıkmıştır. Terimlerin büyüklükleri görece önemlerini ifade etmektedir. Buna göre soğuk savaş döneminde Batı-Doğu terimleri öne çıkarken 1991 sonrası dönemde Ukrayna-Rusya terimleri öne çıkmıştır.

Şekil-2 konu modelinin nitel olarak incelendikten sonra etiketlenen 10 konunun 1991 öncesi ve sonrası dönemdeki beklenen konumlarını %95 güven aralığında göstermektedir. Stratejik savunma, savaş alanlarında kadınların durumu, terörizm ve Ukrayna konularının Soğuk Savaş Sonrası daha fazla gözlendiği görülmektedir.

Şekil-3'te Terörizm ve UkraynaRusya konularının zamana bağlı dağılımları gösterilmektedir. 2000'li yıllardan sonra terörizm söylemi yükselirken, Rusyanın Kırım'ı ilhakından sonra NATO'nun söylemi Rusya çevresinde gelişmiştir.



Şekil 1: Doğu Avrupa konusu içindeki kelimelerin dönemsel ağılımı

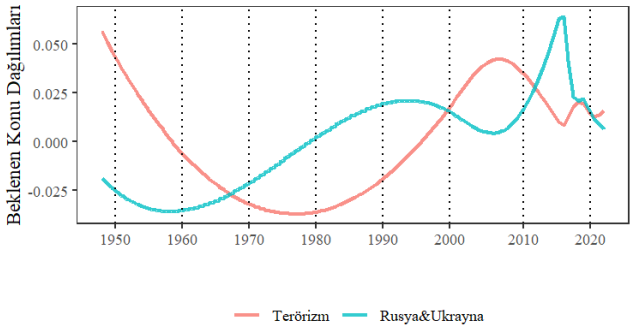


Şekil 2: Konuların dönemsel dağılımı

IV. Sonuç ve Tartışma

Sonuç olarak uzun ve kısa, ön işleme tabi tutulmuş ve tutulmamış metinlerin konu modellemesi yapılmış ve aralarında çeşitli farklılıkla tespit edilmiştir. Bunun yanı sıra STM algoritmasının sosyal bilimlerde farklı doğal dil işleme adımlarıyla birlikte nasıl kullanılabileceğine dair öneriler getirilmiştir.

Öncelikle, STM uygulamasını kullanmak alternatif modellere göre daha kolaydır. R paketine ilişkin detaylı bir rehber makalenin varlığı sosyal bilimler araştırmacısının işlerini kolaylaştırmaktadır. Kategori ve sürekli değişkenleri bir arada kullanarak bu değişkenlerle konular arasındaki ilişkiyi incelemek ve görselleştirmek mümkündür.



Şekil 3: Terörizm ve Ukrayna&Rusya konuların zamana göre bekelenen dağılımları

İkinci olarak, STM algoritması modeli oluştururken ilk 10 bin kelimeyi almaktadır. Bu yüzden ön işleme adımlarında yapılacak dikkatsizlikler modelde dramatik değişikliklere yol açabilir. Bu yüzden STM ile birlikte iyi tanımlanmış bir ön işleme reçetesine ihtiyaç vardır.

Üçüncü olarak, kısa metin performansı nitel olarak incelendiğinde konu kalitesinin uzun metinlerin oldukça gerisinde olduğu görülmüştür. Sadece NATO'nun resmi twitter hesaplarından veri toplanmasına rağmen konuların bağlam hakkında derinlemesine bilgi vermediği gözlenmiştir. Buna karşın konuların isimlendirilebilir olmaları STM algoritmasının kısa metinlerle yeterli performansı gösterdiğine işaret etmektedir.

Dördüncü olarak, bu araştırmada benimsenen ön işlemenin konu modellerine etkisinin çok fazla olmadığı görülmektedir. Bunda doğal dil işleme uygulamalarının sınırlı tutulması önemli rol oynamıştır. Ancak farklı dillerdeki belgelerden gelen terimlerin yarattığı gürültü konu kalitesini etkilemiştir. Bu yüzden araştırmacıların analiz öncesi dil çeşitliliğine yönelik geliştireceği yaklaşım önemlidir. Ek olarak sosyal bilimlerde STM ile birlikte farklı doğal dil işleme yaklaşımları kullanılarak çok daha sofistike sonuçlar elde edilebilir. Örneğin özel ve yer adları etiketlenmesi gibi sözlüksel yaklaşımların, sosyal medya paylaşımları üzerinden yapılacak alan adı, anahtar kavram, kullanıcı ağ analizleri, konu modellerinin vektörizasyonla birlikte kullanılarak anahtar kavram tespit edilmesi gibi farklı senaryolarda konu modelleme işlevselliştirilebilir.

Sonuç olarak bu araştırmada sosyal bilimlerde zaman eşdeğişkeni kullanarak yapılacak bir konu modelleme yaklaşımında dikkat edilmesi gereken adımlar ele alınmıştır. Araştırmanın eksik yönleri arasında tek bir agoritma seçilmesi ve sonuçların sadece nitel olarak değerlendirilmesi gelmektedir. Bu yüzden gelecekte aynı verilerle Dynamic Embedded Topic Models (D-ETM) ve Topics over Time (TOT) gibi farklı modeller kullanılarak modeller arasında karşılaştırma yapılabilir. Ayrıca, sonuçların yorumlanmasında nitel yaklaşımda word intrusion gibi yöntemler kullanılabilir. Nitel karşılaştırmaya ek olarak farklı modellerin aynı verilerle ürettiği sonuçların benzerlik ve farklılıkları literatürde ele alınan nicel yöntemlerle hesaplanabilir.

KAYNAKLAR

[1] M. J. Salganik, *Bit by Bit Social Research in the Digital Age*, 1st ed. New Jersey, USA: Princeton University Press, 2018.

[2] J. Grimmer and B. M. Stewart, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis*, vol. 21, no. 3, pp. 267–297, 2013, publisher: Cambridge University Press.

[3] J. Chuang, J. D. Wilkerson, R. Weiss, D. Tingley, B. M. Stewart, M. E. Roberts, F. Poursabzi-Sangdeh, J. Grimmer, L. Findlater, and J. Boyd-Graber, "Computer-Assisted Content Analysis: Topic Models for Exploring Multiple Subjective Interpretations," p. 9, 2014.

[4] J. Grimmer, M. E. Roberts, and B. M. Stewart, "Machine Learning for Social Science: An Agnostic Approach," *Annual Review of Political Science*, vol. 24, no. 1, pp. 395–419, 2021, _eprint: https://doi.org/10.1146/annurev-polisci-053119-015921.

[5] R. Churchill and L. Singh, "The Evolution of Topic Modeling," *ACM Computing Surveys*, Dec. 2021, just Accepted.

[6] K. Oflazer and M. Saraçlar, *Turkish Natural Language Processing (Theory and Applications of Natural Language Processing)*, 1st ed. Switzerland: Springer International Publishing, 2018.

[7] U. Reber, "Overcoming Language Barriers: Assessing the Potential of Machine Translation and Topic Modeling for the Comparative Analysis of Multilingual Text Corpora," *Communication Methods and Measures*, vol. 13, no. 2, pp. 102–125, Apr. 2019, publisher: Routledge _eprint: https://doi.org/10.1080/19312458.2018.1555798.

[8] D. Maier, C. Baden, D. Stoltenberg, M. De Vries-Kedem, and A. Waldherr, "Machine Translation Vs. Multilingual Dictionaries Assessing Two Strategies for the Topic Modeling of Multilingual Text Collections," *Communication Methods and Measures*, vol. 16, no. 1, pp. 19–38, Jan. 2022, publisher: Routledge _eprint: https://doi.org/10.1080/19312458.2021.1955845.

[9] F. Lind, J.-M. Eberl, O. Eisele, T. Heidenreich, S. Galyga, and H. G. Boomgaarden, "Building the Bridge: Topic Modeling for Comparative Research," *Communication Methods and Measures*, vol. 0, no. 0, pp. 1–19, Sep. 2021, publisher: Routledge _eprint: https://doi.org/10.1080/19312458.2021.1965973.

[10] M. E. Roberts, B. M. Stewart, and D. Tingley, "stm: R Package for Structural Topic Models," *Journal of Statistical Software*, vol. 91, no. 1, pp. 1–40, 2019.

[11] C. Lucas, R. A. Nielsen, M. E. Roberts, B. M. Stewart, A. Storer, and D. Tingley, "Computer-Assisted Text Analysis for Comparative Politics," *Political Analysis*, vol. 23, no. 2, pp. 254–277, 2015.

[12] M. E. Roberts, B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand, "Structural Topic Models for Open-Ended Survey Responses," *American Journal of Political Science*, vol. 58, no. 4, pp. 1064–1082, Oct. 2014.

[13] S. M. Mourtgos and I. T. Adams, "The rhetoric of de-policing: Evaluating open-ended survey responses from police officers with machine learning-based structural topic modeling," *Journal of Criminal Justice*, vol. 64, p. 101627, Sep. 2019.

[14] T. A. Curry and M. P. Fix, "May it please the twitterverse: The use of Twitter by state high court judges," *Journal of Information Technology & Politics*, vol. 16, no. 4, pp. 379–393, Oct. 2019, publisher: Routledge _eprint: https://doi.org/10.1080/19331681.2019.1657048.

[15] C. Schwemmer and S. Jungkunz, "Whose ideas are worth spreading? The representation of women and ethnic groups in TED talks," *Political Research Exchange*, vol. 1, no. 1, pp. 1–23, Jan. 2019, publisher: Routledge _eprint: https://doi.org/10.1080/2474736X.2019.1646102.

[16] D. Fischer-Preßler, C. Schwemmer, and K. Fischbach, "Collective sense-making in times of crisis: Connecting terror management theory with Twitter user reactions to the Berlin terrorist attack," *Computers in Human Behavior*, vol. 100, pp. 138–151, Nov. 2019.

[17] V. Anzoise, D. Slanzi, and I. Poli, "Local stakeholders' narratives about large-scale urban development: The Zhejiang Hangzhou Future Sci-Tech City," *Urban Studies*, vol. 57, no. 3, pp. 655–671, Feb. 2020, publisher: SAGE Publications Ltd.

[18] P. Grajzl and P. Murrell, "Toward understanding 17th century English culture: A structural topic model of Francis Bacon's ideas," *Journal of Comparative Economics*, vol. 47, no. 1, pp. 111–135, Mar. 2019.

[19] L. Geese, "Immigration-related Speechmaking in a Party-constrained Parliament: Evidence from the 'Refugee Crisis' of the 18th German Bundestag (2013–2017)," *German Politics*, Jan. 2019, publisher: Routledge.

[20] S. E. Kim, "Media Bias against Foreign Firms as a Veiled Trade Barrier: Evidence from Chinese Newspapers," *American Political Science Review*, vol. 112, no. 4, pp. 954–970, Nov. 2018, publisher: Cambridge University Press.

[21] P. T. Eles, B. Pennell, and M. Richter, "Assessing NATO policy alignment through text analysis: An initial study," in *2016 International Conference on Military Communications and Information Systems (ICMCIS)*, May 2016, pp. 1–7.

[22] E. M. Fay, "Measuring Indirect External Threat: A Time Series Sentiment Analysis of NATO Resolutions," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, Oct. 2020, pp. 767–768.

[23] I. I. Mestric, A. Kok, G. Valiyev, M. Street, P. Lenk, M. Racovita, and F. Vieira, "Extracting Value from NATO Data Sets through Machine Learning and Advanced Data Analytics," in *IST-178 Specialists meeting on Big data challenges: situational awareness and decision support.*, 2019.

[24] A. Unver and A. Kurnaz, "Securitization of Disinformation in NATO Lexicon: A Computational Text Analysis," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper 4040148, Feb. 2022.