Mapping Global Protest Tendencies: Geolocating Trends and Topics Through Wikipedia Analysis

Jiyun Beak, Ludovic Moncla

Mapping Global Protest Tendencies: Geolocating Trends and Topics Through Wikipedia Analysis

Jiyun Beak[1,2,*], Ludovic Moncla[2]

[1]Korea Advanced Institute of Science and Technology
[2]INSA Lyon, CNRS, UCBL, LIRIS, UMR 5205, F-69621

Abstract

This study investigates the diverse manifestations of 'protest' across cultures and regions, aiming to provide a nuanced understanding of global dynamics and their impact on human rights. Utilizing topic modeling methods, we extract a substantial corpus of documents from the English Wikipedia, employing precise clustering techniques to categorize various types of protests based on semantic elements such as race, gender, and language. Through cartographic visualization, we illustrate the frequency and distribution of different protest topics. The primary goal is to identify geographic hotspots of human rights conflict, offering a detailed analysis of regional differences in protest propensity. This research serves as an initial step towards a comprehensive global understanding of protest dynamics and their implications for human rights worldwide.

Keywords

Geographical Topic Modelling, Zero-shot topic modeling, Semi-supervised topic modeling

1. Introduction

Throughout history, protests have sparked numerous revolutions around the world or at least marked a historic moment for countries. Recent examples include the Black Lives Matter movement in the United States [1], the Candlelight protests in South Korea [2] or the yellow vest protest in France [3]. However, the nature of protests can vary greatly depending on the cultural context and some of the most lasting impacts of social movements are not only in the political realm, but also in everyday life [4].

This article describes the use of topic modeling and geographic mapping from Wikipedia article content focusing on social movements to analyze protest activity in different countries[1]. Topic modeling is commonly used in open data sources such as social media and Wikipedia. It helps to discover themes that recur in texts and to understand the evolution of topics in social media data. Social movements and protests are actively discussed online and thus offer a large amount of interesting data for topic modeling methods. For example, Latent Dirichlet Allocation (LDA) topic modeling has been used to conduct a comparative study of the #BlackLivesMatter and #StopAsianHate movements [5], which were actively pursued on social media. Topic

GeoExT 2024: Second International Workshop on Geographic Information Extraction from Texts at ECIR 2024, March 24, 2024, Glasgow, Scotland
*Corresponding author.
bamjy99@naver.com (J. Beak); ludovic.moncla@insa-lyon.fr (L. Moncla)
0000-0002-1590-9546 (L. Moncla)

[1]Dataset and code are available on GitHub: https://github.com/ludovicmoncla/mapping-global-protest-tendencies

modeling combined with geographic information science have been used in several studies such as for mapping tweets [6] or track online discussions geographically over time [7].

Our methodology aims to highlight how often and where different protest topics occur based on Wikipedia articles. Our main objective is to pinpoint geographic areas where human rights conflicts are most intense. This study seeks to provide a first step towards the understanding of the dynamics of protest movements and their influence on human rights globally.

2. Methodology

2.1. Topic Modeling

This study explores the application of topic modeling to identify themes related to protests and human rights in Wikipedia articles. Simultaneously, it clusters articles within these identified topics. For this purpose, we experimented the BERTopic[2] framework [8], a deep learning-based topic modeling approach. BERTopic aims to overcome the limitations of traditional topic modeling methods like LDA and NMF. Unlike Bag-of-Words models that ignore semantic relationships, BERTopic uses embeddings (default: BERT Sentence Transformers), Dimensionality Reduction (UMAP), clustering (HDBSCAN), Tokenizer, and Weighting Scheme (c-TF-IDF) in a sequence to create coherent topics. This approach allows for more meaningful topic representations. BERTopic variations include semi-supervised, multimodal, hierarchical, and dynamic, with additional models like KeyBERT or GPT for enhanced topic fine-tuning. OpenAI's Topic Modeling utilizes GPT-3.5 to improve clustering by generating synthetic text samples as topic labels.

This study explores 'Zero Shot Topic Modeling', a semi-supervised method that identifies predefined topics in texts with selected labels and uncovers new topics when documents diverge from these labels. It flexibly handles different outcomes, including the identification of both zero-shot and clustered topics, solely zero-shot topics, or none at all. Zero-shot topics are pinpointed through cosine similarity with predefined labels, and a combined BERTopic model integrates both zero-shot and traditional topics.

2.2. Geographical Mapping

The second major step in our research pipeline involves the process of geographically mapping the identified topics within Wikipedia articles related to different forms of protest. Our first experiment is conducted at the country level, with the goal of gaining a comprehensive understanding of the distribution of protest-related topics across different regions of the world. We systematically extract all instances of country names and their corresponding adjectival forms from the articles. This comprehensive compilation serves as the basis for our georeferencing efforts. We then delve into the analysis of the frequency with which each country name is associated with specific topics. This granular examination allows us to discern the prevalence and distribution patterns of protest-related discourse within the context of individual countries, facilitating a nuanced exploration of how these themes manifest and resonate across different geopolitical landscapes.

[2] https://maartengr.github.io/BERTopic/

3. Experiments

Our experiment started with the gathering of nearly 10,000 Wikipedia entries featuring the term "protest" in their title or content. Subsequently, we employed the Wikipedia API Python wrapper[3] to retrieve the textual content of these entries. Then, a language dependent preprocessing phase involving traditional natural language processing techniques (i.e., tokenization, lemmatization, and the removal of stopwords) was performed in order to prepare the dataset for topic modeling.

In an initial experiment, we used the zero-shot BERTopic method to compare predefined topic embeddings with document embeddings via cosine similarity. Based on a threshold, documents were either assigned to these zero-shot topics or clustered by the standard BERTopic model. This approach helps identify both expected and unexpected topic clusters. For the analysis of predefined protest tendencies we define seven labels: gender, nationality, ethnicity, race, language, religion and disability. The minimum topic size has been set to 50 documents in order to limit the number of new topics and the General Text Embeddings[9] has been used. Several values for the minimum cosine similarity have been tested. Results show that, for 0.7, no new cluster is found, while 30 and 40 new clusters are identified for 0.8 and 0.85, respectively. Having 40 new clusters reduces the size of predefined topics, in this case the nationality topic don't even exist. Then, among the 30 new topics, 20 seem to refer directly to locations (mainly countries) rather than themes, such as [knesset palestine gaza israel], [syria syrian damascus assad], [ukraine russia protest crimea], ...Other topics more related to forms of protests, activism or specific political themes rather than human rights are also interesting such as [song performed music album], [statue monument erected plaza], or [nuclear protest antinuclear opposition]. However, the zero-shot learning showed his limits as some of the new clusters should have been grouped with predefined one such as [apartheid opposition protest africa]. Additionally, because of the minimum topic size limit, some articles are considered as outliers and not clustered such as "School Strike for Climate"[4].

For the geocoding part of our process, we counted the number of occurrences of each country name (and their corresponding adjectival forms) within each document and topic. Country level is a first step and allows us to reduce ambiguity of extracting place names with global coverage. We used the GeoPandas[5] and Matplotlib[6] Python libraries to visualize the frequency of each country name per topic (see Figure 1). Frequency mapping shows that topics have a different worldwide distribution. The results show that USA is in the top five of the most frequent country names for all the seven predefined topics while, Israel, India, Iran, United Kingdom and Russia are in the top 5 for 3 topics. Also, France is in the top 5 for *nationality* and *language*, Ireland for *nationality* and *disability*, Canada for *disability* and *language*, China for *language*, and South Africa and Australia for *race*. The countries with the highest number of protests are known for their diversity based on historical background. For example, Israel is known for being a multi-ethic, multi-language country [10, 11]. Moreover, ethnic and national identities such as "American" often conflict with the diverse realities within countries, complicating state-building

[3] https://github.com/martin-majlis/Wikipedia-API
[4] https://en.wikipedia.org/wiki/School_Strike_for_Climate
[5] https://geopandas.org/en/stable/
[6] https://matplotlib.org

and reviving ethnic movements with the historical forces of nationalism [12]. Furthermore, in the contemporary world, inequalities at the intersections of global capitalism, race, and class are intensifying [13], suggesting that one protest can escalate others in similar contexts.

Enhancements are needed for accurately recognizing multi-word country names and various spellings in texts. Additionally, a detailed analysis is required to determine the context and relevance of the country names to the topics discussed.

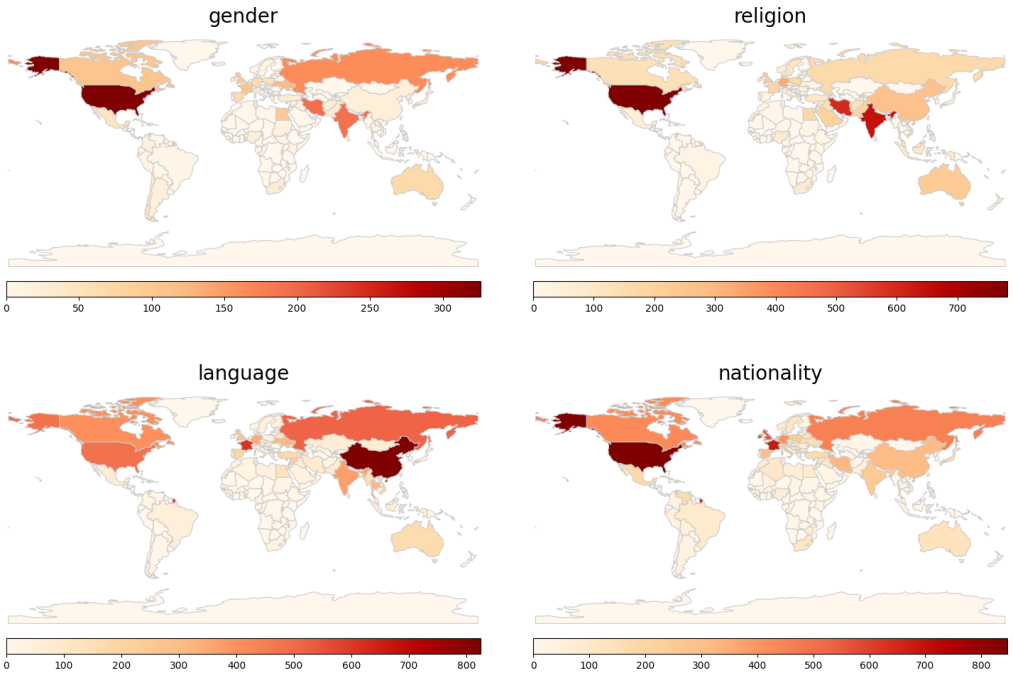

Figure 1: Country name frequency per topic

4. Conclusion

In conclusion, this preliminary study shows promising results in the use of topic modeling and geographical mapping to analyze protest activity across different countries in order to provide a nuanced understanding of the dynamics of social movements and their impact on human rights globally. By employing innovative methodologies such as BERTopic and zero-shot topic modeling, this research provides insights into the prevalence and distribution patterns of protest-related discourse, shedding light on the varying cultural contexts and historical influences shaping these movements. Further investigation into the role of specific countries within each protest topic is essential for a deeper understanding of their societal and historical implications. Through continued research and analysis, we can strive towards fostering greater awareness and advocacy for human rights issues worldwide.

References

[1] F. Megan Ming, Can black lives matter within us democracy?, The ANNALS of the American Academy of Political and Social Science 699 (2022) 186–199. doi:10.1177/00027162221078340.

[2] A. L. Park, A Recycling of the Past or the Pathway to the New? Framing the South Korean Candlelight Protest Movement, Journal of Asian Studies 81 (2022) 101–105. doi:10.1017/S0021911821001480.

[3] S. Kipfer, What colour is your vest? reflections on the yellow vest movement in france, Studies in Political Economy 100 (2019) 209–231. doi:10.1080/07078552.2019.1682780.

[4] A. Morris, C. M. Mueller, Frontiers in social movement theory, Yale University Press, 1992.

[5] X. Tong, Y. Li, J. Li, R. Bei, L. Zhang, What are people talking about in #blacklivesmatter and #stopasianhate? exploring and categorizing twitter topics emerged in online social movements through the latent dirichlet allocation model, in: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22), Oxford, United Kingdom, 2022, p. 723–738. doi:10.1145/3514094.3534202.

[6] D. Ghosh, R. Guha, What are we 'tweeting'about obesity? mapping tweets with topic modeling and geographic information system, Cartography and geographic information science 40 (2013) 90–102. doi:10.1080/15230406.2013.776210.

[7] M. G. Lozano, J. Schreiber, J. Brynielsson, Tracking geographical locations using a geo-aware topic model for analyzing social media data, Decision Support Systems 99 (2017) 18–29. doi:10.1016/j.dss.2017.05.006.

[8] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, arXiv preprint arXiv:2203.05794 (2022). doi:10.48550/arXiv.2203.05794.

[9] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, M. Zhang, Towards general text embeddings with multi-stage contrastive learning, 2023. arXiv:2308.03281.

[10] E. Ben-Rafael, E. Shohamy, M. Hasan Amara, N. Trumper-Hecht, Linguistic landscape as symbolic construction of the public space: The case of israel, International journal of multilingualism 3 (2006) 7–30.

[11] B. Spolsky, R. L. Cooper, The languages of Jerusalem, Oxford University Press, 1991.

[12] S. Olzak, Ethnic protest in core and periphery states, Ethnic and racial studies 21 (1998) 187–217.

[13] S. Nazneen, A. Okech, Introduction: Feminist protests and politics in a world in crisis, 2021.