Whose Palestine Is It? A Topic Modelling Approach to National Framing in Academic Research

Maida Aizaz[1], **Taegyoon Kim**[2], **Lanu Kim**[2]
[1]Graduate School of Data Science, KAIST
[2]School of Digital Humanities and Computational Social Sciences, KAIST
{maidaa25, taegyoon, lanukim}@kaist.ac.kr

Abstract

In this study, we investigate how author affiliation shapes academic discourse, proposing it as an effective proxy for author perspective in understanding what topics are studied, how nations are framed, and whose realities are prioritised. Using Palestine as a case study, we apply BERTopic and Structural Topic Modelling (STM) to 29,536 English-language academic articles collected from the OpenAlex database. We find that domestic authors focus on practical, local issues like healthcare, education, and the environment, while foreign authors emphasise legal, historical, and geopolitical discussions. These differences, in our interpretation, reflect lived proximity to war and crisis. We also note that while BERTopic captures greater lexical nuance, STM enables covariate-aware comparisons, offering deeper insight into how affiliation correlates with thematic emphasis. We propose extending this framework to other underrepresented countries, including a future study focused on Gaza post-October 7.

1 Introduction

In academia, countries are studied not only by their own scholars but also by scholars from other countries. Yet, the institutional location of a researcher may shape how a nation is studied – what issues are highlighted, what is left unsaid. In this study, we ask: do researchers in different countries emphasise different topics when studying the same country? This question is crucial because academic research plays a great role in shaping global narratives, and overlooking how author perspectives shape national discourse – in addition to traditionally-studied aspects such as race, class and gender – may lead to incomplete or skewed understandings of the nation being studied. We explore this question using Palestine as a case study, owing to its history as one of the most politically-charged and contested nations (Irving, 2023). With a long history of occupation, resistance, and conflict, Palestine stands as not only a subject of study but a site of deep symbolism, particularly for scholars with direct ties to the nation.

We argue that author affiliation, domestic vs foreign, serves as an effective proxy for author perspective, shaping academic attention just as significantly as other social factors like race, gender, and class. Through understanding these influences, we can better assess knowledge construction, especially for marginalised and geopolitically-oppressed nations such as Palestine.

To this end, we compare two different legacy topic modelling frameworks – namely BERTopic (Grootendorst, 2022), and structured topic modelling (STM) (Roberts et al., 2019) – to a corpus of over 29,000 English-language academic articles on Palestine, and ask whether – and how – authors from Palestinian and non-Palestinian institutions differ in their topical focus when studying the country.

Our findings reveal that domestic scholars tend to focus on applied, survival-oriented themes like resistance, public health, education and infrastructure, whereas foreign scholars emphasise more abstract, geopolitical topics like conflict, war and law. Not only do these patterns reflect different proximities to the crises, but they also raise important questions about whose voices get to define which aspects of national narratives. In obtaining these findings, we note that BERTopic captures a higher detail of nuance in the text, while STM allows for covariate-level comparisons and thus provides a deeper look into how affiliation compares with topic.

2 Related Work

Previous literature in the fields of science of science and computational social science have examined how researcher identity influences topic

Proceedings of the 9th Widening NLP Workshop, pages 28–40
November 8, 2025 ©2025 Association for Computational Linguistics

selection. Gender, in particular, is a strong reason – an example is how fewer women study NSTEM (natural sciences, technology, engineering and mathematics) due to their systematic exclusion from the field (Kim et al., 2022). Other studies have shown that women are more likely to pursue gender-related fields such as families, gender-based violence and LGBTQIA+ studies than men are, due to their direct connection to the topics at hand (Thelwall et al., 2019; Kozlowski et al., 2022). Similarly, African American/Black scholars tend to study topics pertinent to their own communities, such as socioeconomic studies, health care and disparity, more than other topics (Hoppe et al., 2019). This tendency towards certain topics by certain groups, ostensibly self-serving, is an integral part of addressing issues in equity, as remarked by scholars like Gardner et al (2017). Diversity in scholarship is not merely ethical; it affects what questions are asked, how they are framed, and which narratives are centered – and by whom.

While previous works shed light on diversity manifested in forms such as gender and race, much less work looks into how geographical affiliation affects the academic representation of a country – a question particularly relevant for countries like Palestine, where scholars are simultaneously knowledge producers and subjects of crisis. In such contexts, studying how author affiliation influences topical emphasis reveals whose realities are being prioritised in academic discourse. Our work builds on this line of inquiry by empirically comparing the research topics of domestic and international scholars writing about Palestine, showing how geographical distance shapes academic narratives.

3 Methods

Data. In line with previous bibliometric studies, we make use of OpenAlex (Priem et al., 2022), the leading open-source catalogue of academic papers following the discontinuation of Microsoft Academic Graph. Using the API, we scrape the title, abstract inverted index, publication year and authorship data for all English-language journal articles on OpenAlex that explicitly mention Palestine or Gaza, their variations or demonyms (i.e., *Gazan, Gazans, Palestinian, Palestinians*). We include *Gaza* in addition to *Palestine* owing to its significance as both the centre of conflict between Israel and Palestinian, and the target of Israel's recent genocide (Umar and ur Rahman, 2025). Furthermore, manual checking reveals that when other major Palestinian cities such as West Bank, Hebron, and Ramallah are mentioned, Palestine as a country is often mentioned too – yet there are many studies, such as Aldabbour's (2025), that mention Gaza alone without including Palestine.

After dropping articles without valid title, abstract and/or authorship, we divide the ensuing data into two subsets – one where at least one author is affiliated with a Palestinian institution (thereafter domestic), and another where none are (thereafter foreign). As a result, we are left with a dataset of 29,536 papers – 6,748 domestic and 22,788 foreign – published between 1900 and 2025, with the majority of them published after the spike in 2000, the year that marked the beginning of the Second Intifada – a major Palestinian uprising against Israeli occupation (BBC, 2004) (see Appendix A for the distribution). Owing to the dataset being bibliometric data, we also create our own list of custom stopwords, slightly different for both BERTopic and STM due to their algorithmic variation (see Appendix B for more details).

Model Choice. Due to their frequency of usage and effective performance in computational social science, we use BERTopic (Grootendorst, 2022) in Python and STM (Roberts et al., 2019) in R, two complementary topic modelling frameworks: BERTopic yields lexically-nuanced, interpretable topics, whereas STM enables covariate-aware statistical analysis. We run separate BERTopic models for domestic and foreign authors to explore whether the underlying topics differ by affiliation. In contrast, STM's strength lies in analysing a shared topic space with group-level prevalence variation, which is why we use the same model for both author groups. Note that since we intend to infer the topics from the corpus itself, we do not have predefined topics and thus do not adopt BERTopic's semi-supervised topic modelling approach.

BERTopic. We separate the two subsets into different dataframes, preprocess them by lowercasing, tokenising and removing stopwords, and proceed to apply BERTopic to each of the two subsets separately. We thus generate 15 topics for each, which we inspect manually and do not label. To quantify the difference between the two sets of topics, we calculate the Jensen-Shannon divergence (Lin, 1991) of the foreign subset from the domestic one.

STM. We concatenate the two subsets into a single dataframe and then similarly preprocess, and fit the STM model with the affiliation (foreign vs domes-

tic) as the document-level covariate – allowing for statistical modelling of its effect on topic distribution, i.e., here, owing to STM explicitly incorporating metadata into topic estimation, we generate the topics and then observe how their proportion varies between foreign and domestically-written papers as opposed to our approach with BERTopic, where we model the topics for the two subsets separately. We generate 15 topics, which we characterise with the score-based keywords (from amongst probability, FREX, lift and score) based on our manual inspection. We label these topics using the manually-collected consensus of four large language models (LLMs): Anthropic's Claude-Sonnet-4 (Anthropic, 2025), Google's Gemini-2.5-Flash (Comanici et al., 2025), Meta's Llama-3.3-70B-Instruct (Grattafiori et al., 2024), and OpenAI's GPT-4.1 (OpenAI et al., 2024). Prompting details are available in Appendix D. We then estimate the effect of the affiliation group on topic prevalence in addition to manually inspecting the results and drawing inferences.

4 Results

BERTopic. The 15 topics generated by BERTopic are given in Figure 1 for articles by domestic authors, and in Figure 2 for foreign authors. The top keywords for each topic, by the class-based TF-IDF score (c-TF-IDF), can be found in Appendix C.

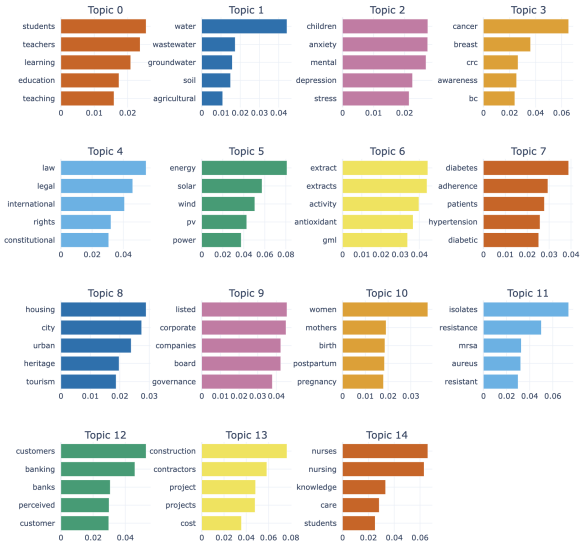

Figure 1: Top 15 topics with top 5 keywords per topic for articles on Palestine written by **domestic** authors.

Quantitatively comparing the two groups, we calculate the Jensen-Shannon divergence of the foreign affiliation articles from the domestic ones to

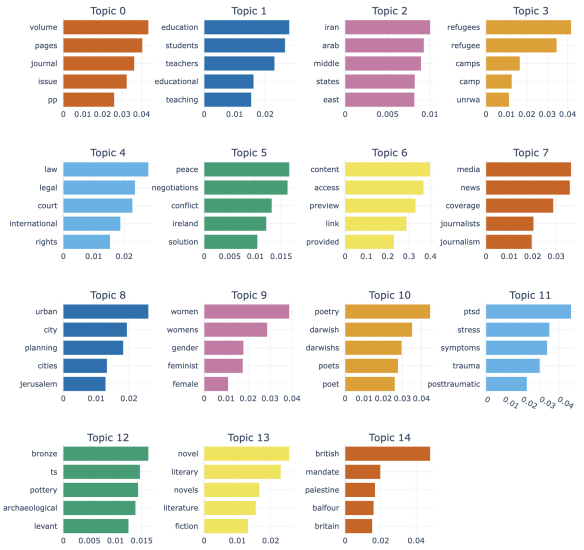

Figure 2: Top 15 topics with top 5 keywords per topic for articles on Palestine written by **foreign** authors.

be 0.319, indicating a moderate divergence (which may in part be attributed to the two low-information or "garbage" topics 0 and 6 in Figure 2); the topics of discussion are not identical but not totally disjoint.

In order to investigate this further, we look at the topics and their keywords in detail. For both domestic and foreign BERTopic outputs, we collect the top 10 keywords per topic, and create a list of 244 unique words. For each of these 244 unique words, we compute two scores: domestic score, and foreign score. These are each the sums of the c-TF-IDF scores as given by the BERTopic model for the foreign and domestic models respectively, with a score of zero if the word did not appear. For instance, if *woman* has a score of 0.1 in domestic topic 1, 0.2 in domestic topic 3, 0.5 in foreign topic 4 and 0.1 in foreign topic 9, then it has a domestic score of 0.3 and a foreign score of 0.6. However, if a word only appears in the domestic model, its foreign score would be zero.

Using these two scores, we calculate a simple bias metric – by subtracting the foreign score from the domestic score – to classify the words as domestic- or foreign-biased. As such, foreign-biased words have positive scores, whilst domestic-biased words have negative scores. Figure 3 details the results (excluding the garbage topics) of the top 45 keywords by cumulative (foreign plus domestic) c-TF-IDF score. The dotted grey line is where domestic bias equals foreign bias, i.e., the domestic and foreign word scores are the same.

The overlap between the two groups is small –

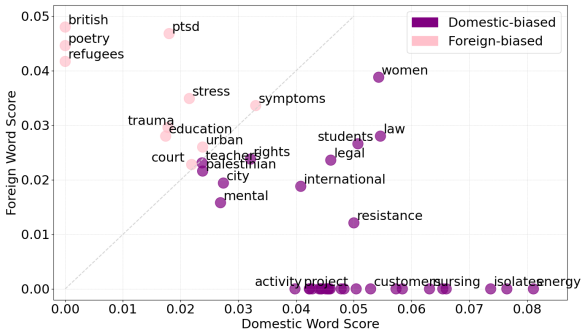

Figure 3: Domestic vs foreign word scores of the top 45 keywords, by cumulative c-TF-IDF score, across both foreign and domestic topics.

education (topic 0 in Figure 1 and topic 1 in Figure 2), law (topic 4 in both) and urban studies (topic 8 in both) – with these three topics appearing to be studied significantly by both foreign and domestic authors. According to Figure 3, a word-level analysis reveals that terms like *legal*, *rights*, and *resistance* are slightly more prominent in domestic works, whereas words like *education*, *court* and *urban* are used a little more frequently by foreign authors, despite being prominent in topics discussed by both author groups. Words like *teachers*, *court* and *symptoms*, however, lie very close to the line, signalling their equal importance to foreign and domestic scholars.

Figures 1 and 2 further reveal that, on both topic- and word-levels, domestic-biased studies pertain to a diverse collection of 'local' topics, such as energy and water, medicine and health, construction and finance, directly relevant to Palestinian society, largely bypassing politics and war. This suggests a more granular, applied focus on daily survival, resistance, and local infrastructure. In contrast, while foreign authors also study Palestine in multiple contexts – such as history, poetry, and politics – most foreign-authored topics tends to frame the country within geopolitical narratives, addressing topics such as diplomacy, occupation, refugees, trauma, and international law. This reinforces our idea of author positionality's impact on topical emphasis; for domestic scholars, the ongoing humanitarian crises may push their research toward practical, community-rooted needs. Meanwhile, foreign scholars – while perhaps motivated by advocacy – may be more inclined to frame Palestine as a site of conflict and resistance, engaging international audiences. In other words, for domestic scholars, the crises and war are not an abstract subject to be studied, but a daily reality to be endured.

However, despite capturing nuanced topics across the two groups, we find BERTopic to have several limitation. It often includes repetitive or redundant topic words (such as topic 10, which contains both *darwish* and *darwishs*), lacks details on topic prevalence and does not support covariate analysis. To address these, we turn to STM, allowing us to formally model the relationship between author affiliation and topic prevalence.

STM. The topic prevalence of the 15 topics generated by STM, with the top keywords per topic, along with the detailed topic words, are visualised in Appendix E. Based on the top 20 keywords, we used human evaluation on the results of four LLMs to label the topics, which are detailed in Table 4 in Appendix D.

In terms of general prevalence, *Israeli-Palestinian War* is, intuitively, most frequently discussed by both groups. Upon examination of the keywords, it appears that topics 4 and 7 – namely *Name Formatting Systems* and *Academic Publishing Locations* – may be garbage topics (as seen in BERTopic's results as well), containing boilerplate academic terms rather than thematic content. Removing the two, we proceed to estimate the effect of the affiliation group on topic prevalence – our main result for this paper – as shown in Figure 4, with positive values indicating stronger association with domestic authors, and negative values with international ones.

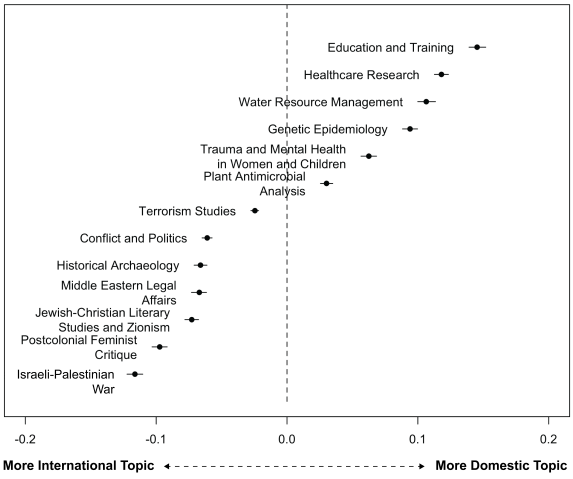

Figure 4: Relationship between author affiliation and topic prevalence, with topics ordered by coefficient size.

In line with our observations from BERTopic, we find that domestic authors are more likely to

study topics related to Palestine's internal contexts, such as environmental studies, healthcare, trauma response. It is worth noting here that while BERTopic's word-level analysis classified the keyword education as foreign-biased, STM reveals education to be a strongly domestic topic – highlighting the distinction between word- and topic-level analyses.

The domestically-prevalent topics appear to reflect concerns pertaining to public health, environment, and social welfare, rooted in local realities as a result of the ongoing humanitarian crises since as far back in history as the 1948 Nakba – the expulsion and forced displacement of over 700,000 Palestinians from their homes by Zionist militant groups that later formed today's Israel Defense Forces (IDF) (Natour, 2016) – marking the beginning of the resistance that goes on to this very day.

In contrast to their domestic counterparts, we see that foreign authors are more likely to engage with externally-oriented themes like feminist critique, legal affairs, archaeology, but most prominently, the ongoing war and ensuing politics. This divergence may be a product of the differential proximity to crisis; for domestic scholars, the war is an existential condition, so their response seems to be survival-oriented scholarship. Resource limitations and institutional pressures may be additional factors pushing them to prioritise healthcare, leaving discussions such as those of prominent poets like Mahmoud Darwish to foreign (or perhaps internationally-established Palestinian) authors. For these foreign authors, Palestine is a symbolic site – used as a lens for broader theoretical, legal, or comparative debates. Some may be motivated by solidarity – such as Umar and ur Rahman's work (2025) – using academic work to expose the injustice and inhumane activities carried out against the Gazans, or shed light on objects of protest like Port (2024) does.

Based on our comparison of the two topic modelling approaches for this task, we note that while STM offers statistical modelling the effect of author affiliation and identifies broader topic prevalence, it tends to yield coarser-grained themes. Compared to BERTopic, it captures fewer nuanced or culturally specific topic keywords – such as *negotiations* or *refugees* – that emerge clearly in BERTopic outputs, instead yielding more low-information topics. However, STM's strength lies in its ability to support covariate-informed analysis, offering a detailed look into the structural relationships across topics.

5 Conclusion

In this study, we asked whether the institutional affiliation of researchers affects how a country is represented in academic work, using Palestine as a case study. By applying two topic modelling frameworks – BERTopic and STM – to a corpus of over 29,000 articles, we found consistent evidence that domestic and international scholars frame Palestine differently, likely as a result of lived proximity to crises; domestic authors centre on internal realities, such as public health, education, and environmental issues, that reflect immediate societal needs as a result of the ongoing crises, wars and recent genocide. In contrast, foreign scholars adopt theoretical and external-facing framings with legal, historical and geopolitical conflict-related topics.

In our comparison of the two topic modelling frameworks, we note that BERTopic allows for richer lexical, word-level nuance, whereas STM supports structured comparisons and statistical inference. Together, our findings suggest that author affiliation is not merely a background detail; it is a factor that shapes the thematic landscape of national academic discourse. Our framework – combining bibliometric filtering, topic modelling and affiliation-based comparison – is easily adaptable to other countries. Applying it to other underrepresented or geopolitically-oppressed regions – including a further study on Gaza pre- and post-October 7, 2023 – could further highlight how knowledge production is shaped by researcher's positionality.

6 Limitations

While our findings reveal significant differences in topic prevalence between domestic and foreign authors, several limitations remain to be addressed. First, our classification of foreign authors is based solely on institutional affiliation, which may include Palestinian-origin researchers working abroad. This potentially mixes positionality with geographic affiliation, and future work could explore author ethnicity or language to disambiguate, perhaps with a scholar migration dataset such as Akbaritabar et al's (2024), to determine how the studies from Palestinian researchers abroad differ from those by non-Palestinian researchers. Additionally, our current findings do not take into account the field of study, even though topical emphases may vary between domains (e.g., medicine

vs humanities). We also do not account for time as a covariate; this limits the ability to track how the discourse has shifted following the escalation of the genocide in Gaza post-October 7, 2023. These limitations lay the grounds for future research.

7 Ethical Considerations

Our study uses publicly available bibliometric and abstract data from OpenAlex; no full-text content, private author metadata, or sensitive personal information are used. Institutional affiliation is treated as a proxy for author country, which may not always align with lived identity (e.g., Palestinian-origin researchers working abroad); this approximation is acknowledged as a limitation. In our analysis, we take care to avoid normative judgments about the "value" of foreign or domestic research, and to treat all topical patterns merely descriptively. We also actively resist abstracting the suffering of the Palestinian people, opting to instead frame domestic scholarship as rooted in lived crisis. As with all NLP studies involving demographic inference or group comparisons, we stress that observed differences are contextual and not causal. This framework is intended to spark discussion about scholarly narratives, and not to assign traits to authorship groups.

References

Aliakbar Akbaritabar, Tom Theile, and Emilio Zagheni. 2024. Bilateral flows and rates of international migration of scholars for 210 countries for the period 1998-2020. *Sci Data*, 11(1):816. Publisher: Nature Publishing Group.

Belal Aldabbour, Samah Elamassie, Saher Mahdi, Haytham Abuzaid, Tamer Abed, Yaser Tannira, Khaleel Skaik, Yousef Abu Zaydah, Abdelkareem Elkolak, Mohammed Alhabashi, Adham Abualqumboz, Abdelrahman Alwali, Heba Alagha, Mahmoud Eid, Shireen Abed, and Bettina Bottcher. 2025. Exploring maternal and neonatal health in a conflict-affected setting: cross-sectional findings from Gaza. *Conflict and Health*, 19(1):45.

Anthropic. 2025. Introducing Claude 4.

Robert Barron. 2019. Palestinian Politics Timeline: Since the 2006 Election.

BBC. 2004. Al-Aqsa Intifada timeline.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint*. ArXiv:2507.06261 [cs].

Jean-Pierre Filiu. 2014. *Gaza: A History*. Oxford University Press, New York, New York.

Susan K. Gardner, Jeni Hart, Jennifer Ng, Rebecca Ropers-Huilman, Kelly Ward, and Lisa Wolf-Wendel. 2017. "Me-search": Challenges and opportunities regarding subjectivity in knowledge construction. *Studies in Graduate and Postdoctoral Education*, 8(2):88–108.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783 [cs].

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint*. ArXiv:2203.05794 [cs].

Travis A. Hoppe, Aviva Litovitz, Kristine A. Willis, Rebecca A. Meseroll, Matthew J. Perkins, B. Ian Hutchins, Alison F. Davis, Michael S. Lauer, Hannah A. Valantine, James M. Anderson, and George M. Santangelo. 2019. Topic choice contributes to the lower rate of NIH awards to African-American/black scientists. *Science Advances*, 5(10):eaaw7238. Publisher: American Association for the Advancement of Science.

Sarah Irving, editor. 2023. *The Social and Cultural History of Palestine: Essays in Honour of Salim Tamari*. Edinburgh University Press.

Lanu Kim, Daniel Scott Smith, Bas Hofstra, and Daniel A. McFarland. 2022. Gendered knowledge in fields and academic careers. *Research Policy*, 51(1):104411.

Diego Kozlowski, Vincent Larivière, Cassidy R. Sugimoto, and Thema Monroe-White. 2022. Intersectional inequalities in science. *Proceedings of the National Academy of Sciences*, 119(2):e2113067119. Publisher: Proceedings of the National Academy of Sciences.

J. Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Ghaleb Natour. 2016. The Nakba—Flight and Expulsion of the Palestinians in 1948. In Andreas Hoppe, editor, *Catastrophes: Views from Natural and Human Sciences*, pages 81–104. Springer International Publishing, Cham.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774 [cs].

Matthew Porter. 2024. Black, white, & read all over: is wearing a keffiyeh enough for Palestinian justice? *Cultural Studies*, 0(0):1–21. Publisher: Routledge _eprint: https://doi.org/10.1080/09502386.2024.2445022.

Jason Priem, Heather Piwowar, and Richard Orr. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint*. ArXiv:2205.01833 [cs].

Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. 2019. stm: An R Package for Structural Topic Models. *Journal of Statistical Software*, 91:1–40.

Mike Thelwall, Carol Bailey, Catherine Tobin, and Noel-Ann Bradshaw. 2019. Gender differences in research areas, methods and topics: Can people and thing orientations explain the results? *Journal of Informetrics*, 13(1):149–169.

Khadija Umar and Zia ur Rahman. 2025. Genocide in Real Time: A Critical Analysis of The Political Logic of Civilian Destruction in Gaza. *Research Journal for Social Affairs*, 3(5):1–7. Number: 5.

A Yearly Publications

Figure 5 shows the yearly distribution of publications in our dataset; we mark some of the important years in the history of Palestine with dashed lines, and the start of the present-day Israeli occupation in 2023, which began on 7th October, with a dash-dot line.

1948 is the year of the 'Nakba' – over 700,000 Palestinians were expelled or forced to flee from their homes by Zionist militant groups that later formed today's Israel Defense Forces (IDF) (Natour, 2016). 1987 and 2000 are the years of the First and Second Intifada respectively; these were major uprisings of the Palestinian people against Israeli occupation (BBC, 2004). 2006 marks Hamas' legislative election win; the Palestinian national movement then fractured into two rival governments, with Hamas controlling Gaza, and Fatah leading the West Bank (Barron, 2019). In late 2008, the Gaza War began, resulting in the destruction of over 46,000 homes in Gaza, and making more than 100,000 Gazans homeless (Filiu, 2014).

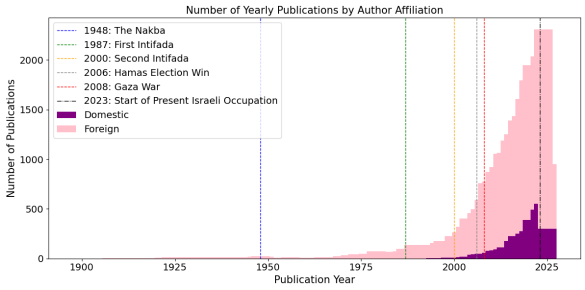

Figure 5: Number of yearly publications, with important years in the history of Palestine marked.

B Custom Stopwords

For BERTopic, in addition to Python package NLTK's English-language stopwords, we added the following custom stopwords: study, result, data, paper, method, analysis, country, et, al, altmetric, publication, researcher, data, objective, abstract, research, results, used, pdf, altmetrics, citation, author, academic, oxford, works, words, search, abstracts, crossref, doi, updated, score, metrics, article, describe, described, model.

For the STM model in R, in addition to its default stopwords, we added the following: study, result, data, paper, method, analysis, country, et, al, altmetric, publication, researcher, data, objective, abstract, research, results, used, pdf, altmetrics, citation, author, academic, oxford, works, words, search, abstracts, crossref, doi, updated, score, metrics, article, describe, described, model, south, chapter, book, report, volume, issue, number, journal, title, english, review, science, publish, google. Please note that we do not add Palestine-related words to either list, as we noticed in our experiments that doing so removed related terms, such as Israel or conflict, from the topic words altogether, resulting in a loss of information.

C BERTopic Topic Words

Table 1 describes the top 10 words in topics found by BERTopic for domestic authors, while Table 2 shows the same for foreign authors.

D STM Topic Labelling Prompts

Once we had the topic words as given above, we used four different LLMs to label the topics: Anthropic's Claude-Sonnet-4 (Anthropic, 2025), Google's Gemini-2.5-Flash (Comanici et al., 2025), Meta's Llama-3.3-70B-Instruct (Grattafiori et al., 2024), and OpenAI's GPT-4.1 (OpenAI et al., 2024). These models were chosen due to their

Topic	Top 10 Keywords
0	students, teachers, learning, education, teaching, university, universities, english, skills, educational
1	water, wastewater, groundwater, soil, agricultural, gaza, area, aquifer, samples, strip
2	children, anxiety, mental, depression, stress, psychological, ptsd, symptoms, trauma, traumatic
3	cancer, breast, crc, awareness, bc, patients, women, symptoms, risk, participants
4	law, legal, international, rights, constitutional, palestinian, state, court, judicial, states
5	energy, solar, wind, pv, power, electricity, renewable, photovoltaic, systems, speed
6	extract, extracts, activity, antioxidant, gml, plant, ic, plants, antibacterial, leaves
7	diabetes, adherence, patients, hypertension, diabetic, blood, tdm, medication, mets, glucose
8	housing, city, urban, heritage, tourism, architectural, spaces, historical, cultural, archaeological
9	listed, corporate, companies, board, governance, firms, accounting, financial, audit, exchange
10	women, mothers, birth, postpartum, pregnancy, pregnant, care, breastfeeding, maternal, childbirth
11	isolates, resistance, mrsa, aureus, resistant, genes, infections, antimicrobial, coli, antibiotic
12	customers, banking, banks, perceived, customer, adoption, intention, mobile, ecommerce, services
13	construction, contractors, project, projects, cost, productivity, factors, management, industry, materials
14	nurses, nursing, knowledge, care, students, competency, practice, bls, training, caring

Table 1: Top 10 keywords for each topic for articles written by **domestic** authors.

Topic	Top 10 Keywords
0	volume, pages, journal, issue, pp, published, university, palestine, google, scholar
1	education, students, teachers, educational, teaching, schools, learning, school, language, teacher
2	iran, arab, middle, states, east, syria, regional, relations, policy, nuclear
3	refugees, refugee, camps, camp, unrwa, lebanon, syrian, migration, protection, displaced
4	law, legal, court, international, rights, icc, jurisdiction, courts, occupation, criminal
5	peace, negotiations, conflict, ireland, solution, process, oslo, israelipalestinian, negotiation, parties
6	content, access, preview, link, provided, available, information, use, copy, permalink
7	media, news, coverage, journalists, journalism, reporting, framing, newspapers, conflict, journalistic
8	urban, city, planning, cities, jerusalem, architecture, marathon, architectural, landscape, housing
9	women, womens, gender, feminist, female, gendered, palestinian, rights, patriarchal, feminism
10	poetry, darwish, darwishs, poets, poet, poems, poem, poetic, mahmoud, resistance
11	ptsd, stress, symptoms, trauma, posttraumatic, exposure, traumatic, mental, depression, coping
12	bronze, ts, pottery, archaeological, levant, trade, age, amphora, bc, ceramic
13	novel, literary, novels, literature, fiction, postcolonial, palestinian, writer, writers, writing
14	british, mandate, palestine, balfour, britain, declaration, britains, tna, iwm, colonial

Table 2: Top 10 keywords for each topic for articles written by **foreign** authors.

cost-effectiveness as well as performance. With the temperature at 0.2 and the seed set to 8282, the system role was as follows: You are an expert in linguistics. Provide your answer in a single word or short phrase under four words.

The user role was set in the following manner: The following keywords are extracted from research articles. Based on these keywords, suggest a short, descriptive topic label: {*prompt*}. Here, *prompt* denotes the top 20 words by score, and this was repeated for each of the 15 topics.

The results of the topic labelling are given in Table 3. Based on these generated labels, we manually made the determination to create the final topic labels, detailed in Table 4.

E STM Topic Words

Figure 6 details the topic prevalence of the top 15 generated by STM, with the top 5 keywords per topic, whilst figures 7 to 21 show the wordclouds for topics 1 to 15 as extracted by STM.

Topic	Claude-Sonnet-4	Gemini-2.5-Flash	GPT-4.1	Llama-3.3-70B-Instruct
1	Byzantine-Ottoman Archaeology	Historical Archaeology	Historical Archaeology of the Holy Land	Historical Archaeology
2	Water Resource Management	Water Management	Water Resources Management	Water Resources Management
3	Military Conflict Politics	Political Conflict	Conflict & Politics	Conflict Politics
4	LaTeX Formatting Parameters	Naming Conventions	Name Formatting Systems	Name Formatting
5	Healthcare & Medical Practice	Clinical Health Research	Healthcare Research	Healthcare Research
6	International Refugee Law	International Law/Politics	Middle Eastern Legal Studies	Middle East Law
7	Academic Publishing	Academic Publishing Context	Academic Publishing Locations	Academic Publishing
8	Postcolonial Resistance Narratives	Critical Identity Narratives	Postcolonial Feminist Critique	Postcolonial Studies
9	Educational Training Management	Education & Training	Education & Training	E-learning Management
10	Genetic Disease Biomarkers	Disease Genetics	Genetic Epidemiology	Genetic Diabetes Research
11	Israeli-Palestinian Conflict	Israeli-Palestinian Conflict	Israeli-Palestinian Conflict	Israeli-Palestinian Conflict
12	Jewish Literary Theology	Religious Literary Studies	Jewish-Christian Literary Studies	Jewish Studies
13	Mental Health Trauma Research	Trauma & Gender	Women's & Children's Mental Health	Trauma in Women
14	Plant Antimicrobial Compounds	Bioactive Plant Compounds	Plant Antimicrobial Analysis	Plant Antimicrobials
15	Terrorism and Security Studies	Terrorism Studies	Terrorism Studies	Terrorism Studies

Table 3: Topic number and generated topic labels with each of our four chosen labels.

Topic	Label
1	Historical Archaeology
2	Water Resource Management
3	Conflict and Politics
4	Name Formatting Systems
5	Healthcare Research
6	Middle Eastern Legal Affairs
7	Academic Publishing Locations
8	Postcolonial Feminist Critique
9	Education and Training
10	Genetic Epidemiology
11	Israeli-Palestinian War
12	Jewish-Christian Literary Studies & Zionism
13	Trauma & Mental Health in Women & Children
14	Plant Antimicrobial Analysis
15	Terrorism Studies

Table 4: Topic number and final chosen topic label.

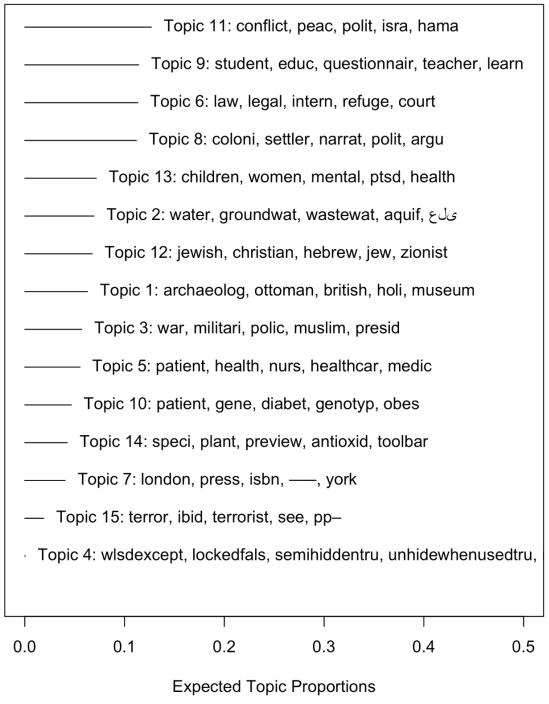

Figure 6: Prevalence of the 15 topics with top 5 keywords per topics.

Figure 7: Wordcloud for Topic 1 - Historical Archaeology

Figure 8: Wordcloud for Topic 2 - Water Resource Management

Figure 9: Wordcloud for Topic 3 - Conflict and Politics

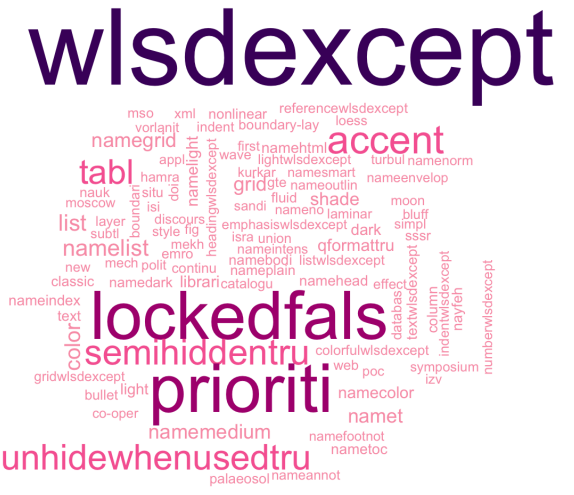

Figure 10: Wordcloud for Topic 4 - Name Formatting Systems

Figure 11: Wordcloud for Topic 5 - Healthcare Research

Figure 13: Wordcloud for Topic 7 - Academic Publishing Locations

Figure 12: Wordcloud for Topic 6 - Middle Eastern Legal Affairs

Figure 14: Wordcloud for Topic 8 - Postcolonial Feminist Critique

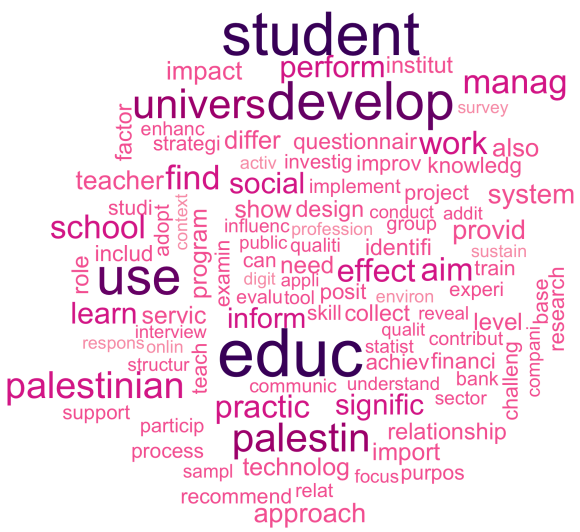

Figure 15: Wordcloud for Topic 9 - Education and Training

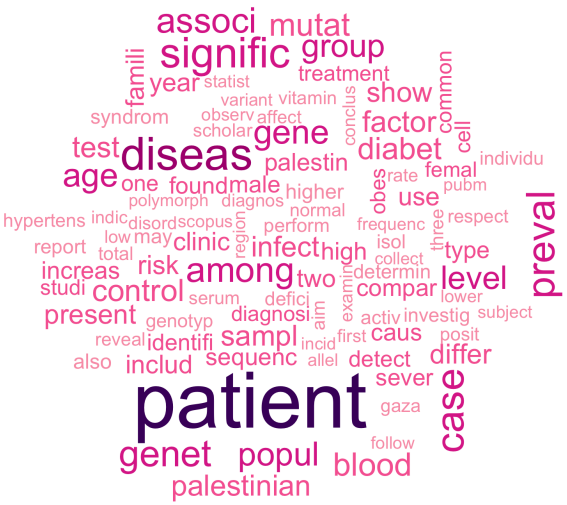

Figure 16: Wordcloud for Topic 10 - Genetic Epidemiology

Figure 17: Wordcloud for Topic 11 - Israeli-Palestinian War

Figure 18: Wordcloud for Topic 12 - Jewish-Christian Literary Studies & Zionism

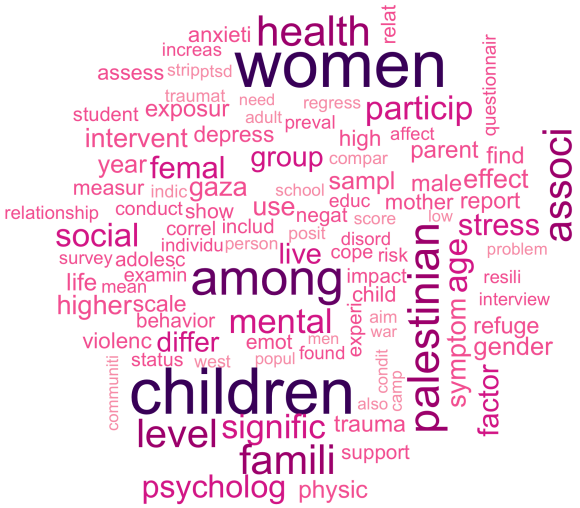

Figure 19: Wordcloud for Topic 13 - Trauma & Mental Health in Women & Children

Figure 20: Wordcloud for Topic 14 - Plant Antimicrobial Analysis

Figure 21: Wordcloud for Topic 15 - Terrorism Studies