Topic Modeling in Feminist Debates on Instagram: A Generative AI Approach

Thalia Almeida[1,2], **Keila Barbosa** [1,3], **Sheyla Fernandes**[4], **and André Aquino**[1]

[1]Orion Lab., Computer Institute – Federal University of Alagoas
Maceió, AL, Brazil

[2]Department of Computer Science – Federal University of Minas Gerais
Belo Horizonte, MG, Brazil

[3]Department of Computer Science – Federal University of Pernambuco
Recife, PE, Brazil

[4]Psychology Institute – Federal University of Alagoas
Maceió, AL, Brazil

{thalia.almeida,keilabarbosa,alla}@orion.ufal.br, sheyla.fernandes@ip.ufal.br

Abstract. *This study investigates the identification of feminist themes addressed by Brazilian profiles on Instagram using Natural Language Processing (NLP) techniques and topic modeling. To analyze recurring discussions and their variations, we apply the BERTopic technique for topic modeling, while we use the large language model (LLM) LLaMA for labeling the identified themes. The modeling process resulted in 90 topics, highlighting issues such as domestic violence, reproductive rights, and mental health, reflecting current debates in the Brazilian context. Additionally, an experiment compared topic labeling performed by human participants and LLM, analyzing the similarity between both responses. The BERTScore metric, which assesses semantic similarity, yielded the highest results, with values between* 0.68 *and* 0.79*. This result indicates that the LLM produced semantically similar responses to human ones. The results emphasize the role of NLP techniques and language models in identifying complex social themes, providing a solid foundation for future studies on the impact of social media on awareness and social change promotion.*

1. Introduction

Topic modeling is a technique used to identify latent themes in large text collections by grouping frequently co-occurring words, revealing the main subjects under discussion [Blei et al. 2003]. This approach is widely applied in social media analysis, enabling researchers to uncover patterns in posts, comments, and interactions across platforms. It is particularly relevant in fields such as sociology, communication, and data science [Wahid et al. 2025]. Common applications include monitoring public opinion, analyzing online discourse, and detecting hate speech. Topic modeling can also help examine ideological movements like digital feminism. Within this realm, Cyberfeminism emerges as a key concept—defined as a contemporary branch of feminism that leverages digital technologies for empowerment and mobilization [Paasonen 2011]. It explores the intersection of gender and technology, advocating for female presence in digital spaces and promoting activism and access to knowledge through online networks. Instagram, with over 1

billion active users worldwide [Statista 2024], serves as a powerful channel for feminist communication. It provides a platform for movements to grow, reach broader audiences, and strengthen engagement.

Understanding this digital landscape—both in terms of topics and underlying platform policies—is essential for evaluating the impact of digital feminism. While feminist activism has historically confronted systems of oppression, its online presence introduces new tensions that warrant critical reflection [Vachhani 2024].

Recent events in Brazil underscore the urgency of this analysis. In September 2023, Minister Rosa Weber of the Supreme Federal Court (STF)[1] voted to decriminalize abortion up to 12 weeks of pregnancy. Additionally, 2024 data from the Laboratory for Feminicide Studies at the State University of Londrina (UEL)[2] showed a sharp rise in femicide cases. By June 2024, Brazil recorded 750 confirmed femicides and 1,693 combined confirmed and attempted cases—an average of 6.05 per day, up from 2.7 in 2023. These numbers highlight the importance of analyzing how feminist voices respond to and disseminate such critical issues online.

Based on these considerations, this study applies topic modeling to short-text data from feminist Instagram profiles and explores the potential of Large Language Models (LLMs) for topic labeling. It addresses two research questions:

> **RQ1 [Topic Modeling on Feminist Discourse]**: *Can we use topic modeling techniques to identify the main themes discussed by feminist profiles on Brazilian Instagram?*
> **RQ2 [Generative AI for Topic Labeling]**: *Can generative AI replace human effort in labeling these topics?*

The first question aims to investigate whether topic modeling algorithms are capable of extracting coherent and relevant themes from short-text social media content. By focusing on feminist discourse on Brazilian Instagram, the study seeks to explore the feasibility and effectiveness of these techniques in a context marked by brevity, informal language, and dynamic interactions. The goal is to understand whether automated methods can support the thematic analysis of digital feminist activism. The second one explores the potential of Large Language Models (LLMs) to perform the task of topic labeling in place of human annotators. The objective is to assess the quality, consistency, and interpretability of labels generated by LLMs when applied to topics extracted from feminist content on Instagram. By comparing machine-generated labels with those produced by human evaluators, the study aims to understand the viability of incorporating generative AI into the workflow of social media analysis, particularly in feminist narratives.

For this study, data was collected directly from feminist Instagram profiles, focusing on the textual content of their posts. After undergoing preprocessing, we apply the BERTopic model to extract key topics. For the labeling stage, we compared annotations produced by human participants and an LLM-based annotator. Then, for the evaluation phase of the experiment, we use different metrics to assess textual similarity between responses. Notably, the BERTScore metric, which measures semantic similarity, achieved the highest values, with an average score of 0.73. The findings highlight the effectiveness

[1] `https://shorturl.at/6rNav`, in Portuguese, Accessed March 2025
[2] `https://shorturl.at/ZWXDB`, in Portuguese, Accessed March 2024

of NLP techniques and LLMs in analyzing complex social issues, offering a valuable basis for future research on the influence of social media in raising awareness and driving social change.

2. Related Work

The literature presents several applications of topic extraction. These studies aim to infer real-world environments by exploring the relationship between topics and other contextual factors [Joo et al. 2020]. [Kurten and Beullens 2021] sought to determine whether the number of *Tweets* varied according to the pandemic timeline and its phases, as well as how the content of these tweets evolved. In another study, [Nobles et al. 2020] applied topic modeling to understand how individuals who self-identified as living with Human Immunodeficiency Virus (HIV) expressed their experiences with the disease. Additionally, [Peres et al. 2023] analyzed homophobic hate speech on Twitter during the COVID-19 pandemic using NLP techniques and topic modeling, contributing to discussions on stress and public discourse in Brazil.

The fragmented nature of texts on platforms such as Instagram and Twitter, where post captions are often short and accompanied by images and *hashtags*, presents significant challenges for traditional topic modeling methods such as *Latent Dirichlet Allocation* (LDA) [Blei et al. 2003], designed to analyze longer, more coherent text corpora. The study by [Mazarura and de Waal 2016] compares two topic modeling approaches applied to short texts: LDA and the *Dirichlet Multinomial Mixture Model* (GSDMM). While LDA performs well for long texts, such as books and academic articles, its effectiveness is considerably lower when applied to short texts, such as social media posts.

With advancements in topic modeling techniques, a study published in 2022 introduced BERTopic [Grootendorst 2022], a model that integrates language *embeddings*, specifically *Bidirectional Encoder Representations from Transformers* (BERT), with hierarchical clustering and dimensionality reduction techniques to generate more coherent topics [Devlin et al. 2019]. The research highlights that BERTopic excels in capturing word context through *embedding* generation, enabling a deeper understanding of topics. In some studies, topic labeling is based on the authors' knowledge of the dataset, subject matter, or other contextual insights, as reported by [Ibrahim and Wang 2019]. Approximately 63.64% of the analyzed studies employed this approach [Laureate et al. 2023]. However, studies such as [Brown 2019] and [Bérubé et al. 2020] have raised concerns about potential biases and the depth of insights derived from this method. For this reason, [Ibrahim and Wang 2019] suggests that future research should consider these critical aspects and explore additional or alternative methods for topic interpretation.

Recent research has increasingly explored the use of *Large Language Models* (LLMs), such as GPT-4 and LLaMA-3, in text analysis. For instance, the study by [Kwon et al. 2024] analyzed 1.26 million tweets about nuclear energy in the U.S. between 2008 and 2023, leveraging language models to classify public sentiment on political and energy-related topics. Few studies have explored the use of LLMs in topic modeling. For example, [Kirilenko and Stepchenkova 2024] focused on evaluating the ability of large language models (LLMs), such as ChatGPT, to extract key themes from viewer reactions to popular videos about a rural destination in China. When comparing the results with the LDA model, we observed that LLMs demonstrated superior accuracy,

specificity, and ability to distinguish discussed topics, especially in short text contexts. This finding reinforces the limitations of traditional methods, such as LDA, when dealing with fragmented documents, such as blog comments or social media posts. In this sense, our approach stands out by integrating state-of-the-art language models for topic extraction and interpretation, providing a more contextualized, coherent analysis aligned with the characteristics of contemporary data.

3. Methodology

Figure 1 illustrates our proposed flow. It considers four main phases: data collection, text preprocessing, topic modeling, and topic labeling. We divided the latter into two approaches: human labeling and generative AI-assisted labeling.

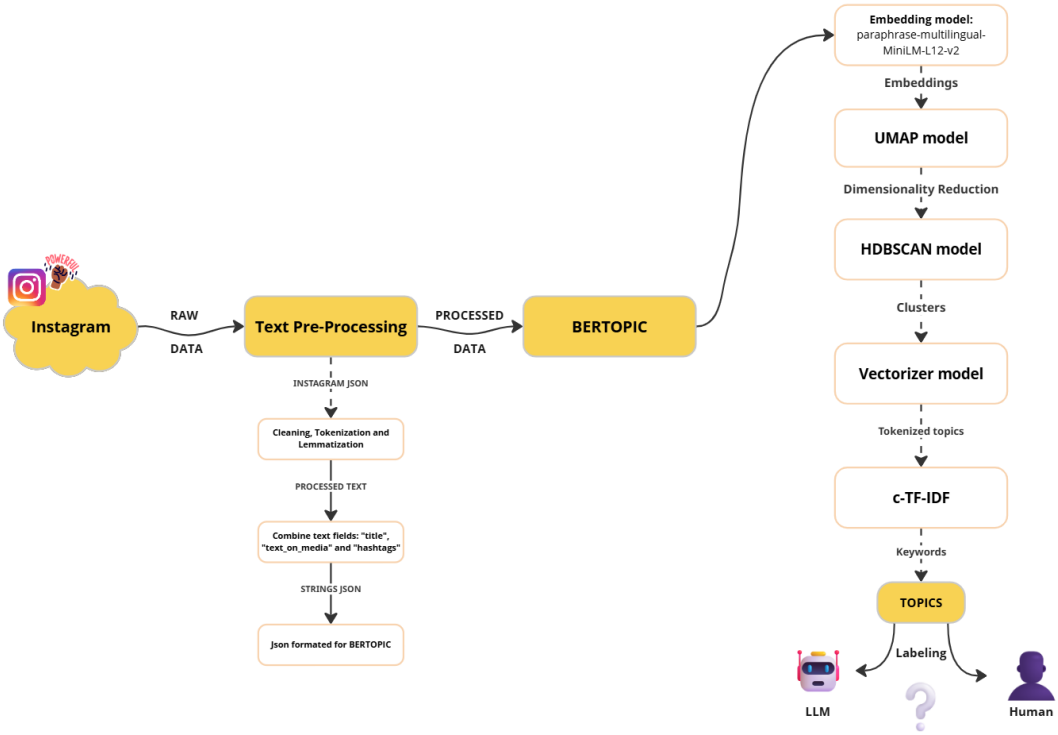


Figure 1. Flowchart of the methodology applied for topic modeling in feminist profiles on Instagram.

Data Collection allows the selection of feminist profiles using Instagram's search bar with the descriptors "feminism" and "feminist". We considered only public profiles with at least 10,000 followers, active within the past 12 months, and directly aligned with the study theme. This threshold was chosen as a proxy for influence, under the assumption that accounts with larger audiences are more likely to shape public discourse and reach broader segments of the online community. Data was collected using an Open Source Intelligence (OSINT) API [Lowenthal 2020], which queries Instagram's API to extract public post data. The final dataset includes posts from the following profiles: @planetaella, @feminiismo, @coletivo_feminista, @feminismo_semdemagogia, @arquivosfeministas, and @revistatpm. The data collection resulted in a total of 11,582 posts collected.

Text Preprocessing presents four steps: i. relevant content extraction from each post, including captions, hashtags, and text embedded in images. The latter was obtained directly through Instagram's API, via the *"text_on_media"* attribute, which provides pre-extracted textual content from images when available. Therefore, additional image processing was not required; ii. a text-cleaning process removed unwanted characters, line breaks, URLs, mentions, and punctuation; iii. Stopwords were eliminated, followed by lemmatization and stemming to reduce words to their root form [Manning et al. 2008]. These steps were performed using the NLTK and spaCy libraries, both configured for Portuguese, and iv. concatenation of all processed content into a single string per post. After preprocessing and cleaning, we excluded posts with no usable textual content, reducing the dataset from 11,582 to 11,070 posts prepared for analysis. Each string represented a post containing information from captions, text present in images, and associated hashtags, forming the corpus used in topic modeling techniques.

Topic Modeling employs the BERTopic model [Grootendorst 2022], which integrates transformer-based embeddings using the pre-trained *paraphrase-multilingual-MiniLM-L12-v2* model, UMAP for dimensionality reduction, and HDBSCAN clustering with BERTopic's default hyperparameters. After clustering, the texts are tokenized and a class-based TF-IDF (c-TF-IDF) is applied to identify representative keywords which highlights terms that distinguish one cluster from others [Manning et al. 2008]. This approach yields semantically coherent topic groups with relevant keywords.

Topic Labeling (Human-based) Interpreting topics through human labeling requires more than identifying keywords. We conducted a human-labeling experiment using a Google Forms survey. Ten topics were selected: five with the highest number of posts and five randomly chosen. The survey was open to the general public, with no restrictions on participant profile. Most respondents were university students or members of the academic community. No prior expertise in feminism or topic modeling was required. Each question included the top 10 keywords and a representative post excerpt, asking participants to suggest a label for the topic, as shown in Figure 2.

Example of the Question Asked to Respondents

1. Keywords Representing the Topic (in Portuguese):
estupro, aborto, feminismo, feminista, violência, afirmar, gênero, menina, legal, crime.

Part of a Post Caption (in Portuguese):
"NÃO É UM CASO ISOLADO — Entenda quais os direitos de crianças e adolescentes nessa situação no vídeo acima com anônimo, diretora do Instituto Liberta e doutora especialista em Direito Constitucional. Mais de 21 mil meninas entre 10 e 14 anos engravidam por ano no Brasil..."

Question: How would you name this topic? (It can be a single word or a phrase.)

Figure 2. Example of the labeling task presented to respondents.

Topic Labeling (LLM-based) applies Meta AI's large-scale language model, LLaMA-3, to automatically generate topic labels. This approach aimed to evaluate the consistency, specificity, and clarity of AI-generated labels compared to those proposed by human participants. This step is an essential contribution to our work.

We used prompt engineering techniques to formulate structured input prompts to

accomplish this. These prompts adopted a chat-style format with two components: *role*, defining the model's function (assistant or expert), and *content*, specifying instructions and the input data [White et al. 2023]. We used the model in a few-shot learning setup, including examples within the prompt to help guide the generation process, as shown in Figure 3.

Example of Prompt Sent to LLaMA

You are an assistant responsible for identifying topics based on a list of keywords and a representative tokenized text.
I will provide a list of keywords and a token string representing the topic's context. Your task is to analyze the keyword list and the representative text to determine the most relevant topic.

1. Analyze the list of keywords and the associated token string.
2. Use the keywords and representative text to identify the most relevant topic.
3. Return a JSON object that maps the index to the corresponding topic name.

Input Model:

```
"keywords": ["word1", "word2"],
"text": "token string"
```

Response Format:

```
"0": "Topic Name"
```

Now, analyze this input:

Figure 3. Example of prompt used to generate topic labels with the LLaMA language model.

We formatted the inputs sent to LLaMA in JSON and derived directly from the BERTopic output, enabling seamless integration between the models. Likewise, the responses returned by the LLM followed the same structure, facilitating future reuse in various applications and automated pipelines.

For a long time, researchers have considered subjective evaluations the gold standard in natural language generation tasks. However, automatic metrics have proven advantageous in terms of cost and time efficiency and consistency in large-scale assessments [Papineni et al. 2002, Lin 2004, Zhang et al. 2019]. We used the following four metrics to compare the human-labeled topics with those generated by the LLaMA model: i. **BLEU (Bilingual Evaluation Understudy)** measures the precision of overlapping n-grams between the reference (human) and candidate (LLM) outputs,

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right),$$

where BP is the brevity penalty, w_n are weights, and p_n is the modified precision for n-grams [Papineni et al. 2002]. ii. **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)** or ROUGE-L, a commonly used variant, consider the longest common subsequence (LCS) between the reference and candidate texts. It emphasizes re-

call [Lin 2004]

$$\text{ROUGE-L} = \frac{(1+\beta^2) \cdot \text{LCS\_Precision} \cdot \text{LCS\_Recall}}{\text{LCS\_Precision} + \beta^2 \cdot \text{LCS\_Recall}}$$

iii. **Cosine Similarity** used after converting both human and LLM-generated labels into TF-IDF vectors, the cosine similarity is calculated as

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|},$$

where A and B are the respective TF-IDF vectors of the two texts. iv. **BERTScore** measures semantic similarity by aligning tokens using cosine similarity in embedding space, leveraging contextual embeddings from BERT

$$\text{BERTScore}_{\text{F1}} = \frac{2 \cdot P \cdot R}{P + R},$$

where P and R represent precision and recall over embedding alignments [Zhang et al. 2019].

These metrics provide complementary perspectives on the comparison between human and LLM-generated topic labels. While BLEU and ROUGE focus on lexical overlap and structural similarity, Cosine Similarity and BERTScore allow for a deeper semantic analysis, capturing nuances beyond exact word matches. By employing this diverse set of assessment tools, we ensure a more comprehensive and robust assessment of topic labeling quality.

4. Results and Discussion

Applying the BERTopic technique led to the identification of 90 distinct topics, each represented by 10 keywords. These keywords are the most relevant terms within each group of posts, reflecting the key concepts discussed in each topic. Figure 4 presents the Topic Word Scores for the 12 most representative topics, i.e., those with the highest volume of associated posts.

The Topic Word Score is a metric that indicates the degree of importance of each word within a topic relative to the entire corpus, based on the TF-IDF calculation. The higher the score, the more relevant the word is for defining that particular topic. In the figure, the most significant words for each of the top 12 topics are listed and ordered according to their relative importance within each group. These topics contain the most significant posts, concentrating on the most widely discussed themes in the analyzed dataset. During the topic modeling process with BERTopic, a unique topic, referred to as Topic -1, was identified. This topic encompasses posts that did not fit into the others, making it a collection of outliers or atypical posts. Topic-1 contains 5,373 documents, making it the topic with the most extensive content. Although these posts were not grouped into more homogeneous topics, Topic-1 remains relevant and should not be disregarded.

Figure 5 displays a word cloud highlighting the key terms associated with Topic -1, where the most frequent and representative words appear more prominently. This result suggests that, despite its fragmented nature, these posts may address important themes that contribute to a broader understanding of feminist topics.

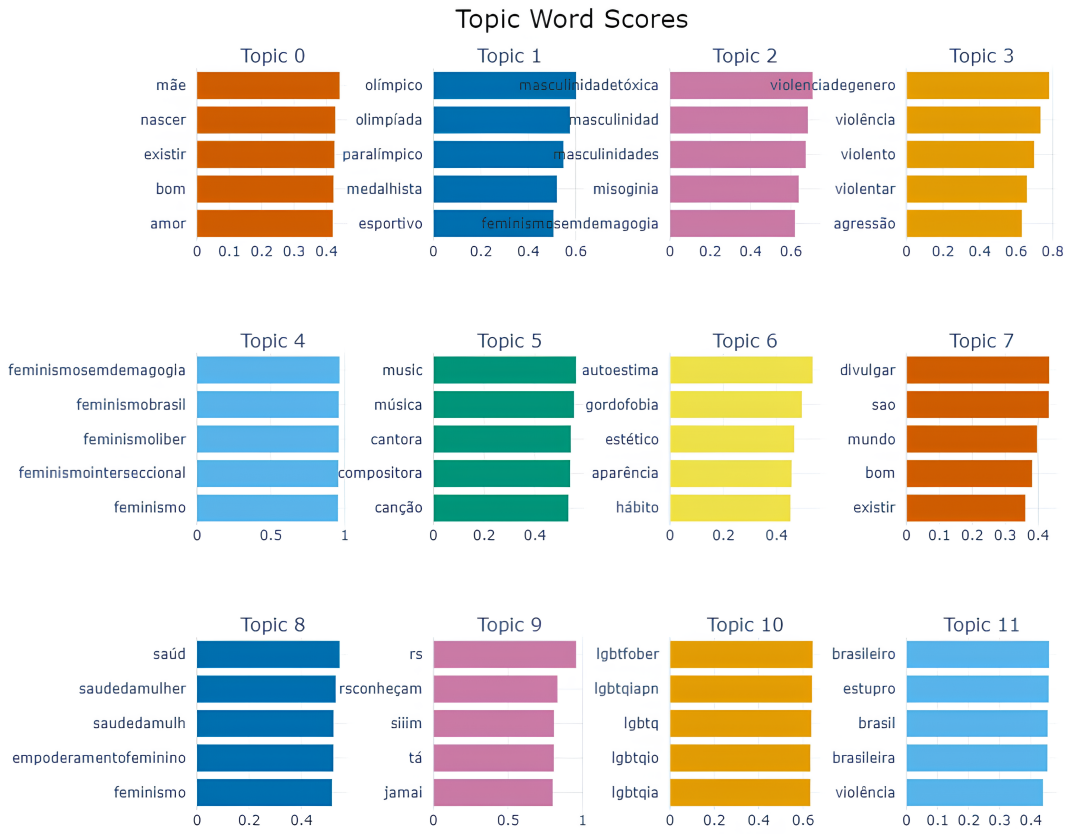


Figure 4. Topic Word Scores for the 12 Most Relevant Topics Identified by BERTopic (from a total of 90).

To validate the interpretability of the topics generated by the LLM, we conducted a human annotation experiment using Google Forms. The form remained open for three weeks and received 38 valid responses. Table 1 summarizes the comparison between the topic labels generated by the LLM and those provided by human participants. We used four evaluation metrics, categorized into two perspectives: **lexical similarity** — measured by BLEU and ROUGE — and **semantic similarity** — evaluated using Cosine Similarity and BERTScore.

From the lexical perspective, the **BLEU scores** are consistently close to zero across all topics, with most values below machine precision (e.g., 1.2750×10^{-231} in *Feminismo e Aborto* and *Maquiagem e Aparência*). These results indicate an extremely low rate of n-gram overlap between LLM-generated and human-provided labels, suggesting that the LLM does not replicate the exact vocabulary or phrasing used by human respondents, even if it captures the overall thematic structure.

In contrast, the **ROUGE scores** reveal greater variation. Topics such as *Violência Doméstica* (0.3858) and *Maternidade* (0.3138) achieved the highest ROUGE scores, indicating that the LLM reproduced key terms and recurring expressions found in human labels. On the other hand, topics like *Transtornos Mentais e Psicologia* (0.0525) and *Maquiagem e Aparência* (0.0531) showed minimal overlap, aligning with the low BLEU scores and reinforcing the conclusion that lexical similarity is topic-dependent.

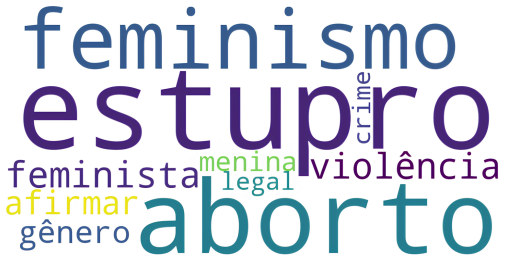


Figure 5. A word cloud displaying the key terms associated with Topic -1 (in Portuguese).

Table 1. Comparison between LLM model and human labeling.

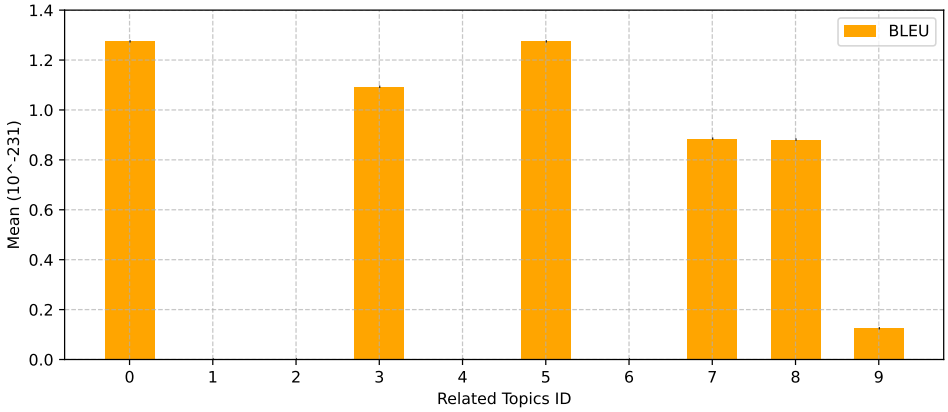
ID	**LLM's answer** (in Portuguese)	**Metrics** (average — standard deviation)			
		BLEU	**ROUGE**	**Cosine Similarity**	**BERTScore**
0	Feminismo e Aborto	$\simeq 0.0$ — 0.0	0.0853 — 0.1013	0.0364 — 0.0967	0.7256 — 0.0344
1	Maternidade	$\simeq 0.0$ — 0.0	0.3138 — 0.2745	**0.3769** — 0.2820	**0.7869** — 0.0855
2	Esporte Olímpico	$\simeq 0.0$ — 0.0	0.1352 — 0.1550	0.0861 — 0.1330	0.6803 — 0.0416
3	Masculinidades e feminismo	$\simeq 0.0$ — 0.0	0.0970 — 0.1496	0.0000 — 0.0000	0.7361 — 0.0496
4	Violência Doméstica	$\simeq 0.0$ — 0.0	**0.3858** — 0.3272	0.3076 — **0.3263**	0.7657 — 0.0923
5	Maquiagem e Aparência	$\simeq 0.0$ — 0.0	0.0531 — 0.1254	0.0215 — 0.0757	0.6900 — 0.0320
6	Feminismo	$\simeq 0.0$ — 0.0	0.0417 — 0.1730	0.0481 — 0.1844	0.7506 — 0.0625
7	Luta contra a homofobia e discriminação LGBT	$\simeq 0.0$ — 0.0	0.1257 — 0.1455	0.0745 — 0.1374	0.6835 — 0.0685
8	Discriminação racial e racismo	$\simeq 0.0$ — 0.0	0.2036 — 0.1476	0.1979 — 0.1535	0.7830 — 0.0567
9	Transtornos mentais e Psicologia	$\simeq 0.0$ — 0.0	0.0525 — 0.0961	0.0179 — 0.0534	0.7279 — 0.0403

The semantic perspective by **Cosine Similarity** values were relatively low overall, ranging from 0.0000 to 0.3769. The topic *Maternidade* again leads with the highest cosine value (0.3769), followed by *Violência Doméstica* (0.3076), which suggests that in these cases, the LLM's choice of terms showed some alignment with the TF-IDF-weighted vocabulary used by human annotators. However, the low scores for topics such as *Masculinidades e Feminismo* (0.0000) and *Transtornos Mentais e Psicologia* (0.0179) imply a significant divergence in term usage and semantic focus.

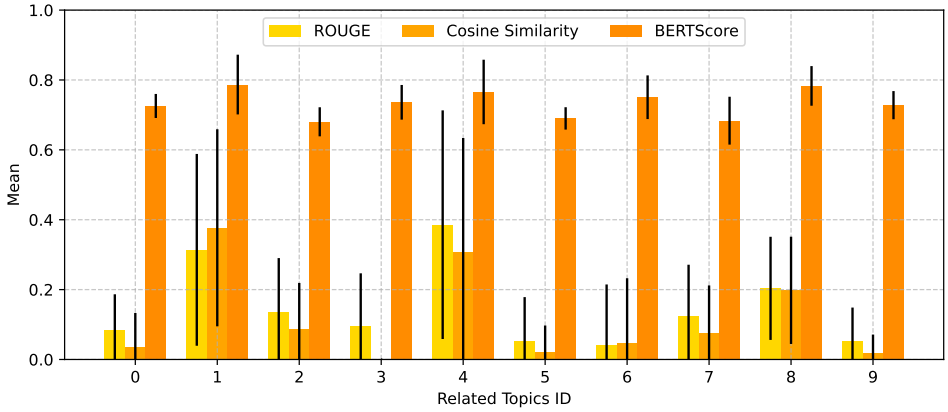
A different picture emerges with **BERTScore**, which consistently produced higher values across all topics — ranging from 0.6835 to 0.7869. Once again, *Maternidade* (0.7869), *Discriminação Racial e Racismo* (0.7830), and *Violência Doméstica* (0.7657) stood out as the most semantically aligned with human responses. This result reinforces the model's ability to convey thematic meaning even when the surface-level vocabulary differs. The lower end of the scale, represented by *Esporte Olímpico* (0.6803) and *Luta contra a homophobia e discriminação LGBT* (0.6835), still reflects moderate semantic alignment, albeit with less conceptual overlap.

Figure 6 reinforced these findings. It illustrates the distribution of metric scores across all topics (ID from Table 1). The figures display the average performance of each metric across ten thematic categories, providing insights into how different aspects of textual coherence and semantic similarity vary by topic. The top chart highlights the uniform near-zero BLEU values, while the bottom chart reveals more nuanced variations in ROUGE, Cosine Similarity, and BERTScore. Notably, the BERTScore bars are consistently higher, reflecting the robustness of embedding-based models in capturing contextual meaning.

The results emphasize an apparent discrepancy between lexical and semantic eval-



(a) BLEU metric



(b) ROUGE, Cosine Similarity, and BERTScore

Figure 6. Average scores and variability of coherence and semantic similarity metrics across feminist-related topics.

uations. While BLEU and Cosine Similarity are useful for surface-level comparison, they fail to recognize deeper thematic equivalences. BERTScore, on the other hand, demonstrates that LLM-generated labels can achieve high semantic congruence with human responses, even when expressed differently. These insights underscore the importance of using embedding-based metrics in evaluating generative models, particularly in complex social themes and short-text environments.

Figure 7 shows the topic-level comparison between Cosine Similarity and BERTScore metrics. It is evident that BERTScore consistently achieves high values across all topics, indicating that the semantic meaning conveyed by the LLM-generated labels is closely aligned with the human-labeled counterparts. This consistency reinforces the ability of transformer-based models to capture nuanced thematic information, even when there is limited lexical overlap. Cosine Similarity, on the other hand, shows considerably more variation across topics. Its lower values suggest that, despite conveying similar meanings, the LLM does not frequently replicate the same terms or surface-level vocabulary used by human respondents. Topics such as *Maternidade*, *Violência Doméstica*, and

Discriminação Racial e Racismo stood out with the highest semantic alignment across both metrics.

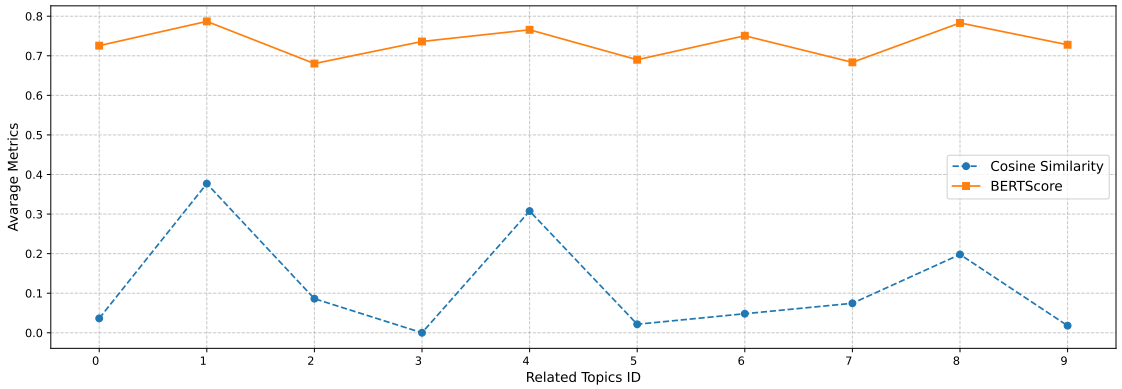


Figure 7. Comparison of Semantic Similarity by Topic (Cosine Similarity vs. BERTScore).

To complement the topic-level view, Figure 8 summarizes the average performance of all four metrics, with error bars representing the standard deviation across topics. The BLEU score confirms negligible n-gram overlap (average: 0.00), while ROUGE (average: 0.17) and Cosine Similarity (average: 0.12) indicate moderate alignment in terms of keyword overlap. However, BERTScore (average: 0.74) clearly outperforms all other metrics in capturing semantic similarity, reinforcing its appropriateness for evaluating topic labels in informal short-text contexts such as Instagram posts.

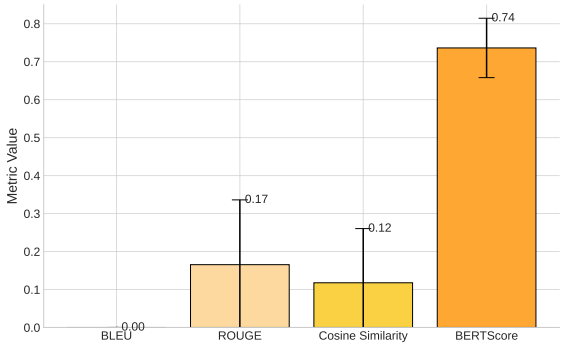


Figure 8. Overall metric averages with error bars representing the standard deviation.

The results affirm that topic modeling techniques, particularly BERTopic, are effective in identifying coherent and relevant themes from short-text feminist content on Brazilian Instagram. This conclusion answers satisfactory our research question **RQ1 — Topic Modeling on Feminist Discourse**. Despite the challenges posed by informal language and short textual length, the clustering and keyword extraction process grouped semantically similar posts into coherent topics. The clear thematic distinctions identified in topics such as *Maternidade*, *Violência Doméstica*, and *Discriminação Racial* demonstrate the model's ability to capture key aspects of feminist discourse. Therefore, automated topic modeling suggests that BERTopic is a viable tool for supporting the thematic analysis of digital feminist activism.

The comparative evaluation between human-labeled and LLM-generated topic names shows that generative AI, huge language models like LLaMA, can produce high-quality and semantically accurate labels. This conclusion answers satisfactory our second research question **RQ2 — Generative AI for Topic Labeling**. While lexical overlap is low (as reflected by BLEU and ROUGE), the consistently high BERTScore values suggest that the LLM captured underlying meanings similar to those intended by human annotators. This result suggests that LLMs can complement or support human in topic-labeling tasks, offering scalability and efficiency for large-scale analyses in computational social science.

5. Conclusion and Future Work

This study aimed to identify and interpret the themes discussed by feminist profiles on Brazilian Instagram by applying Natural Language Processing (NLP) techniques and topic modeling. Using the BERTopic algorithm, we extracted 90 distinct themes from short-text social media posts, with twelve key topics emerging prominently. These themes reflect core concerns of contemporary feminist discourse in Brazil, a country marked by persistent gender-based challenges and vibrant digital activism.

To assess the interpretability of the generated topics, we conducted a comparative analysis between human-labeled topics and those produced by the LLaMA language model. While lexical similarity measured through BLEU and ROUGE was generally low, the BERTScore results demonstrated a high semantic alignment between human and machine-generated labels. This outcome suggests that although the model does not replicate exact terms, it is capable of capturing the intended meaning behind human annotations. In this context, the divergence in vocabulary is not a limitation but a reflection of the model's ability to rephrase concepts while preserving their core semantics.

To our knowledge, our hybrid approach combining BERTopic with a generative Large Language Model for topic interpretation is among the first to systematically evaluate the effectiveness of LLMs in labeling social media topics related to feminist discourse. It demonstrates the feasibility of integrating unsupervised topic modeling with generative AI to reduce human effort in labeling without sacrificing interpretability or thematic coherence. Beyond methodological contributions, this research has practical implications. The ability to automatically identify and interpret key themes in feminist narratives can support institutions, researchers, and civil society organizations in monitoring public discourse, informing communication strategies, and guiding data-driven public policy development focused on gender equity and social justice.

For future work, we suggest exploring the use of LLMs as standalone tools for topic extraction and interpretation, potentially eliminating the need for separate clustering steps. Additionally, multilingual and cross-platform studies may further validate the scalability of this approach and contribute to broader analyses of feminist movements across digital ecosystems. Access our GitHub repository: https://github.com/thalialmeida/tcc/

Acknowledgments

This study was partly financed by the Research Foundation of the State of Alagoas (FAPEAL) under grants E:60030.0000000352/2021 and the National Council for Scientific and Technological Development (CNPq) under grant 407515/2022-4.

References

Bérubé, M., Tang, T.-U., Fortin, F., Ozalp, S., Williams, M. L., and Burnap, P. (2020). Social media forensics applied to assessment of post–critical incident social reaction: The case of the 2017 manchester arena terrorist attack. *Forensic science international*, 313:110364.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Brown, N. M. (2019). Methodological cyborg as black feminist technology: constructing the social self using computational digital autoethnography and social media. *Cultural Studies: Critical Methodologies*, 19(1):55–67.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Ibrahim, N. F. and Wang, X. (2019). A text analytics approach for online retailing service improvement: Evidence from twitter. *Decision Support Systems*, 121:37–50.

Joo, S., Lu, K., and Lee, T. (2020). Analysis of content topics, user engagement and library factors in public library social media based on text mining. *Online information review*, 44(1):258–277.

Kirilenko, A. and Stepchenkova, S. (2024). Automated topic analysis with large language models. In *ENTER e-Tourism Conference*, pages 29–34. Springer.

Kurten, S. and Beullens, K. (2021). # coronavirus: Monitoring the belgian twitter discourse on the severe acute respiratory syndrome coronavirus 2 pandemic. *Cyberpsychology, Behavior, and Social Networking*, 24(2):117–122.

Kwon, O. H., Vu, K., Bhargava, N., Radaideh, M. I., Cooper, J., Joynt, V., and Radaideh, M. I. (2024). Sentiment analysis of the united states public support of nuclear power on social media using large language models. *Renewable and Sustainable Energy Reviews*, 200:114570.

Laureate, C. D. P., Buntine, W., and Linger, H. (2023). A systematic review of the use of topic models for short text social media analysis. *Artificial Intelligence Review*, 56(12):14223–14255.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. ACL. Workshop of the ACL 2004.

Lowenthal, M. M. (2020). *Intelligence: From Secrets to Policy*. CQ Press.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Mazarura, J. and de Waal, A. (2016). A comparison of the performance of latent dirichlet allocation and the dirichlet multinomial mixture model on short text. In *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, pages 1–6.

Nobles, A. L., Leas, E. C., Latkin, C. A., Dredze, M., Strathdee, S. A., and Ayers, J. W. (2020). # hiv: alignment of hiv-related visual content on instagram with public health priorities in the us. *AIDS and Behavior*, 24:2045–2053.

Paasonen, S. (2011). Revisiting cyberfeminism. *Communications*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. ACL.

Peres, D., Silva, G., Faria, E., and Barioni, M. (2023). Análise do estresse e tópicos discutidos no twitter durante a pandemia da covid-19 no brasil. In *Anais do XII Brazilian Workshop on Social Network Analysis and Mining*, pages 43–54, Porto Alegre, RS, Brasil. SBC.

Statista (2024). Instagram: number of global users 2020-2025. Available at: https://www.statista.com/statistics/183585/instagram-number-of-global-users/. Accessed on: March 29, 2025.

Vachhani, S. J. (2024). Networked feminism in a digital age—mobilizing vulnerability and reconfiguring feminist politics in digital activism. *Gender, Work & Organization*, 31(3):1031–1048.

Wahid, J. A., Xu, M., Ayoub, M., Jiang, X., Lei, S., Gao, Y., Hussain, S., and Yang, Y. (2025). Ai-driven social media text analysis during crisis: A review for natural disasters and pandemics. *Applied Soft Computing*, page 112774.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.