IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Investigating Public Online Discourse Surrounding Films About Menstruation in India

Aashi Gupta

The International School Bangalore, NAFL Valley, Whitefield-Sarjapur Road, Dommasandra Circle, Bangalore, Karnataka, 562125, India; 1aashigupta1@gmail.com

ABSTRACT: This study investigates online public discourse surrounding menstruation-themed Indian films through the lens of computational social science. Created a dataset of over 30,000 comments sourced from Reddit and YouTube, we apply a combination of sentiment analysis, sentence embeddings, and topic modeling to explore the ways in which audience conversations differ across platforms and evolve over time. The research focuses on three prominent films: Pad Man (2018), Period. End of Sentence (2018), and Alert Condition: Red (2020). These films, released in different socio-political and cultural contexts, serve as focal points for examining public reactions to menstrual health representation in Indian media. The analysis is structured around three experiments. The first evaluates platform-specific sentiment distributions, comparing how the affordances and audience demographics of Reddit and YouTube influence emotional tone. The second identifies dominant and recurring themes in discussions, revealing how narratives around menstruation are framed, contested, or normalized in digital spaces. The third examines longitudinal sentiment trends, tracking shifts in audience attitudes over multiple years to assess the impact of changing social awareness, policy discourse, and media representation. By combining linguistic and computational approaches, the study highlights how anonymity, cultural context, and platform architecture shape menstrual health discourse, offering valuable insights into the intersection of technology, gender, and public health narratives.

KEYWORDS: Robotics and Intelligent, Machine Learning, Computational Social Science, Sentiment Analysis, Topic Modeling, Menstruation Discourse.

Introduction

Menstrual health has historically been underrepresented and often stigmatized in both public discourse and mainstream media, particularly in countries like India, where cultural taboos continue to shape public attitudes.[1-6] Although recent years have seen growth in advocacy, policy initiatives, and educational campaigns that aim to reduce stigma, menstrual health representation in film remains a relatively new and evolving phenomenon. Films such as *Pad Man* (2018), *Period. End of Sentence*. (2018), and *Alert Condition: Red* (2020) brought menstruation into the public spotlight and generated conversation that extended beyond the cinema hall into digital platforms such as YouTube and Reddit.[7-10] These platforms have become important arenas for expression where anonymity and accessibility enable candid and often polarised discussion.[7,11] Understanding discourse in these digital environments is both timely and necessary. Previous scholarship has examined menstrual representation through cultural, sociological, and feminist perspectives,[4-6] yet only a limited number of studies have used computational social science methods to analyze large-scale public reactions over time. This study contributes to the literature by combining sentiment analysis, sentence embeddings, and topic modeling to investigate how emotional tone and thematic framing differ across platforms and change in response to social, political, and media events. By integrating linguistic and computational approaches, this research provides a comprehensive, data-driven view of how digital conversation both shapes and is shaped by menstrual health narratives in India.

Early Computational Approaches to Public Discourse Analysis:

Early computational studies relied on statistical text representation methods, most notably Term Frequency Inverse Document Frequency (TF IDF).[12-15] TF IDF has been widely used for keyword extraction[13] and remains foundational in many media and communication studies.[15] Xu and colleagues demonstrated that TF IDF can reveal thematic changes in online discussions related to reproductive health policy, particularly during moments of public debate and legislative activity.[16] However, TF IDF treats documents as unordered collections of words and does not account for syntax, context, or semantics. This limitation is especially significant for menstruation discourse, where cultural nuance, metaphor, humor, and code switching carry important meaning. Neural embedding methods such as Word2Vec[17] and GloVe[18] improved upon these approaches by learning distributed representations from co-occurrence statistics. Yet these models assign a single vector to each word regardless of context, which limits their ability to handle polysemy and subtle linguistic variation. This constraint has been noted in studies of sentiment, humor, and stigma in online settings.[5,16]

DOI: 10.36838/IJHSR811.5

The Emergence of Transformer-Based NLP:

Transformer-based architectures transformed natural language processing through the use of self-attention mechanisms. BERT introduced contextual embeddings that capture meaning based on both preceding and succeeding words.[19] These representations significantly improved performance on sentiment classification,[20] stance detection, and other linguistic tasks that require deep contextual understanding. Sentence-level semantic modeling advanced further with Sentence BERT, which adapts BERT into a Siamese architecture to generate high-quality sentence embeddings.[21] Liu and colleagues used Sentence BERT to cluster public opinion across large social media datasets and demonstrated its effectiveness for mapping thematic discourse.[22] Subsequent model compression efforts produced faster variants such as DistilBERT[23] and MiniLM,[24] which maintain competitive performance with lower computational cost. Domain-specific models also became prominent. Twitter RoBERTa, trained on a large corpus of social media text, is particularly effective for informal spelling, emoji usage, and slang.[25] This model has been used extensively to analyze misinformation,[26,27] illicit health product advertising,[28] and pandemic-related communication.

Computational Social Science and Health Discourse:

Transformer-based models have been widely adopted in computational social science. Applications include detecting illicit product sales,[28] tracking influenza-like illness through online behavior,[3] analyzing misinformation trends,[26,27] and studying stigma surrounding reproductive and sexual health. Anonymous online spaces, in particular, have been shown to facilitate more candid and stigma-resistant expression.[11] Despite these advancements, computational work on menstruation remains limited. Existing research often focuses on menstruation as a public health concern,[3] as a human rights issue,[1] or as a topic shaped by policy campaigns.[1] Sociocultural analyses have examined how media portray menstruation, highlighting both stigma-reinforcing and stigma-challenging narratives.[4-6] However, few studies use large-scale natural language processing to compare menstruation-related discourse across platforms such as Reddit and YouTube. This gap motivates the present study.

Menstruation, Media Representation, and the Research Gap:

Menstruation in media occupies an important space at the intersection of public health communication and cultural narrative formation. Scholars have shown that media representations can challenge menstrual stigma[5] or reinforce it,[6] contributing to broader social understandings of gender and embodiment. Yet, empirical studies that examine how specific media events influence large-scale public discourse remain limited. Existing work frequently centers on textual or qualitative analysis rather than audience reception at scale. Research on online health communication has demonstrated that platform structure shapes public expression. De Choudhury and De's study of mental health discussions on Reddit illustrates how anonymity and community norms influence patterns of disclosure and support,[7] while Kaye and colleagues' systematic review of online sexual health discourse shows that algorithmic curation and audience composition drive notable differences in tone and thematic emphasis.[29] These findings suggest that social platforms produce distinct discursive environments, but this insight has not yet been extended to menstruation discourse. No menstruation-focused study to date has used multi-platform datasets to analyze thematic divergence or patterns across structurally different spaces such as Reddit and YouTube. This gap is especially notable given the emergence of films that foreground menstruation in the Indian context, including *Pad Man* (2018),[8] *Period. End of Sentence*. (2018),[9] and *Alert Condition: Red* (2020).[10] Each film generated substantial public conversation, yet no research has examined how these discussions vary across platforms or evolve over time. By analyzing more than twenty-five thousand comments related to these films, this study addresses both a methodological and substantive gap in menstrual health research by mapping cross-platform variation and temporal change in public discourse.

■ Methods

Dataset Curation:

The dataset was constructed to capture public discourse surrounding three films that explicitly engage with menstruation within the Indian socio-cultural context: *Pad Man* (2018),[8] *Period. End of Sentence*. (2018),[9] and *Alert Condition: Red* (2020).[10] These films were selected to represent a range of production contexts, narrative approaches, and audience reach. Pad Man is a large-scale commercial feature inspired by the work of Arunachalam Muruganantham, whose celebrity-driven narrative helped normalize public conversation on menstrual hygiene in India. Period. End of Sentence. It is an Academy Award-winning documentary that elevated global awareness of menstrual stigma. Alert Condition: Red is a grassroots short film that achieved significant engagement despite limited institutional backing, reflecting youth-driven participation in digital conversations. YouTube served as the primary source of video-based audience comments. For each film, comments were scraped from the official uploads or highest engagement versions using the Digital Methods YouTube Comment Scraper. Videos were selected based on three criteria: (1) high view count, (2) substantial engagement in the comment section, and (3) clear relevance to the film. This approach ensured that the resulting dataset reflected authentic and organic audience responses rather than peripheral or low engagement conversations.

Reddit Subreddit Selection:

Reddit data was collected via keyword-based searches across the ten most thematically and culturally relevant subreddits: (1) MovieSuggestions, (2) Mmovieaweek, (3) movies, (4) Feminism, (5) IndianDankMemes, (6) BollyBlindsNGossip, (7) India, (8) Bollywood, (9) Pakistan, (10) todayilearned. These subreddits were identified manually by testing keyword queries like "Pad Man," "Period film," or "menstruation movie" and evaluating which communities hosted the most sustained, relevant conversations. The mix spans general entertainment,

Indian pop culture, gender activism, and meme-driven commentary, enabling a diverse linguistic and ideological sample.

Preprocessing and Cleaning of the Comments:

Data preprocessing was conducted in two phases to ensure suitability for clustering, sentiment analysis, and topic modeling. In the initial phase, all comments were filtered to retain only those written primarily in English. Mixed script entries, including those containing Devanagari or Urdu characters, were removed. Language filtering was supported by established text identification approaches.[30,31] Comments containing fewer than three words were excluded to ensure minimum semantic content. Remaining entries were normalized by converting text to lowercase and removing URLs, user mentions, hashtags, emojis, punctuation, numerals, and excess whitespace. Tokenization and lemmatization were implemented using the spaCy natural language processing framework, which is widely used for large-scale text processing tasks.[32] Stopwords were removed following standard information retrieval practices.[33] This produced a clean and consistent corpus suitable for downstream computational analysis. In the second phase, cleaning focused on removing low information and irrelevant content.

Emoji-only and gibberish posts were deleted through a combination of automated filters and manual inspection. Spam, promotional material, and off-topic commentary were excluded using keyword-based heuristics informed by prior work on social media noise reduction.[26,27] Comments that retained no meaningful tokens after lemmatization and stopword removal were discarded. This two-step procedure ensured a high-quality dataset appropriate for both embedding-based representation and unsupervised modeling.

Handling Hindi Written in English (Transliterated Hindi):

A substantial subset of comments consisted of Hindi phrases written in the Latin alphabet, such as "ladkiyon ko samajhna mushkil hai." Although written using English characters, these comments do not follow English morphology or semantics and therefore fall outside the analytic scope of English language topic modeling. The prevalence of Hindi-English code switching in Indian digital spaces has been documented previously.[34] To identify and remove such entries, a manually curated list of high-frequency Hindi terms (including "hai," "nahi," "kaun," "kyun," "tum," and "kya") was used during the language filtering stage. This supplemented automated detection approaches used in multilingual environments.[30,31] While this method reduced a large portion of transliterated Hindi, complete elimination was not possible due to the fluid and hybrid nature of code switching. Following all filtering and cleaning procedures, the final dataset comprised 25,144 comments, making it one of the largest multi-platform corpora assembled for the computational study of menstruation discourse in Indian media. Reddit comments were sourced from discussions referencing the selected films, while YouTube comments were drawn from trailers and full-length uploads.

Table 1: Dataset Statistics.

Film	Platform	Raw Comments	Cleaned Comments	Avg. Words (Raw)	Avg. Words (Cleaned)
Pad Man (2018)	YouTube	30,469	18,700	8.40	8.40
	Reddit	742	579	18.69	18.69
Period. End of Sentence. (2018)	YouTube	2,237	1,770	15.19	15.19
	Reddit	3,741	3,076	13.66	13.66
Alert Condition: Red (2020)	YouTube	1,923	1,019	7.06	7.06
Total	N/A	39,112	25,144	N/A	N/A

EXPERIMENT 1: Social Media Platform Comparison Using Clustering and Sentiment Analysis:

This experiment aimed to examine whether public discourse around the film Pad Man varied significantly between social media platforms, specifically YouTube and Reddit. The analysis involved taking the raw text of user comments, translating them into numerical form (embeddings), and then exploring the patterns hidden within those numerical representations. These patterns exist in what is known as a latent space: an abstract, multidimensional representation of meaning where similar comments are positioned closer, and dissimilar comments are farther apart. The core assumption within such a latent space is that proximity equals similarity.[19,21,24] If two comments are positioned near each other in this space, it suggests they share thematic or emotional content, even if they use different words. To investigate latent differences in public discourse between Reddit and YouTube, we followed a four-stage computational pipeline: (1) Comment embedding, (2) Dimensionality reduction, (3) Clustering and statistical analysis. (4) Sentiment Analysis. While the full dataset contained over 25,000 comments across three films and multiple platforms, this experiment focused exclusively on a balanced subset of Pad Man comments: 588 from YouTube and 579 from Reddit, ensuring that the clustering analysis would not be skewed by platform size.

1. Comment Embedding:

Each comment was transformed into a dense numerical vector using three different sentence embedding models: MiniLM,[24] DistilBERT,[23] and E5[35] chosen for their balance of computational efficiency, semantic depth, and architectural diversity. MiniLM, with its compact 384-dimensional representation, offered fast processing and was ideal for clustering tasks on medium-sized datasets. DistilBERT served as a compressed yet powerful transformer-based benchmark that retained much of BERT's representational ability.[23] E5, an instruction-tuned model, provided the highest representational capacity, enabling it to capture more abstract semantic relationships. Models like SBERT were excluded to avoid redundancy with MiniLM and DistilBERT, while static word embeddings like GloVe[17] and word2vec[18] were avoided because they cannot capture contextual meaning, which is essential for understanding nuanced discourse.[4-6]

2. Dimensionality Reduction:

Embeddings from these models ranged from 384 to 1024 dimensions, far too high to visualize directly. To make the structure interpretable, we projected them into two dimensions

DOI: 10.36838/IJHSR811.5

using both Principal Component Analysis (PCA)[36] and t-distributed Stochastic Neighbor Embedding (t-SNE).[37] PCA is a linear technique that preserves global variance and highlights broad trends, while t-SNE is a non-linear approach that focuses on preserving local neighborhoods, making it better suited for revealing fine-grained clusters. By comparing results from both techniques, we could verify whether observed groupings were robust patterns in the data rather than artifacts of a single projection method.

3. Clustering and Statistical Analysis:

After dimensionality reduction, I applied KMeans clustering (k = 2) to each of the six 2D embedding spaces (MiniLM, DistilBERT, and E5, each reduced via PCA and t-SNE).[38] Although the sentiment model used later assigns three possible labels (positive, neutral, negative), the clustering itself remained fully unsupervised and was not informed by sentiment. Choosing k = 2 ensured consistent analysis across models and allowed for straightforward comparison of cluster composition. This approach aimed to assess whether natural groupings in the embedding space aligned with platform differences, sentiment distributions, or both.

4. Sentiment Analysis:

To evaluate the emotional tone of user discourse, I applied sentiment analysis to all English-language comments across platforms and embedding models. The objective was to assign each comment a sentiment label POSITIVE, NEGATIVE, or NEUTRAL and later examine how these distributions aligned with latent clusters and platform categories. For sentiment classification, I used the CardiffNLP 'twitter-roberta-base-sentiment' model,[25] a transformer trained on social media data for multi-class sentiment detection. The model predicts one of three classes: LABEL_2 (POSITIVE), LABEL_1 (NEUTRAL), or LABEL_0 (NEGATIVE). After detecting language, I filtered for English comments and mapped these predicted labels to consistent sentiment categories across datasets.[40,41,42] including NEUTRAL sentiment, which was critical. Neutral comments represented a large share of the dataset, particularly on platforms like YouTube and Reddit. Excluding them would have skewed both the sentiment distribution and its relationship with clustering patterns. This sentiment analysis provided the foundation for evaluating whether emergent clusters corresponded to underlying emotional tones, a key question for assessing whether user discourse patterns were shaped by sentiment or other latent factors. To examine whether emotional tone aligned with platform-specific discourse structures, I analyzed sentiment distributions within clusters for Reddit and YouTube separately. The results revealed a recurring trend: clusters with a higher proportion of Reddit comments tended to skew more negative, while clusters with more YouTube comments showed a greater share of positive sentiment.

EXPERIMENT 2: Topic Modeling Using LDA:

Building on the sentiment analysis conducted in Experiment 1, which focused exclusively on Pad Man and compared sentiment across platforms, this second experiment shifts both the analytical lens and the scope of study. The objective here is to explore how the themes within online discourse around menstruation-related films have evolved. Moving beyond sentiment polarity, this experiment applies unsupervised topic modeling techniques to user comments to identify recurring discussion topics and track their prominence in the months following each film's release. Three films form the dataset for this study: Pad Man (2018), Period. End of Sentence (2018), and Alert Condition: Red (2020), all of which address menstruation, albeit through differing cinematic approaches and social contexts. By analyzing audience commentary on these films, the experiment investigates whether particular themes such as stigma, education, activism, or humor persist, diminish, or emerge across different points in time.[1-6] The key distinction of this experiment lies in its temporal focus. This experiment aims to provide a longitudinal perspective on public discourse surrounding menstruation in digital spaces. Rather than listing all emerging topics, it specifically identifies the most prominent topic in each six-month window following each film's release. This allows the analysis to capture dominant narratives over time, comparing how discussion priorities vary not just within a single film's trajectory, but also across different films. By focusing on these peak themes per period, the study surfaces the core concerns that resonated most strongly with audiences in each temporal context. Ultimately, this experiment aims to provide a longitudinal perspective on public discourse surrounding menstruation in digital spaces. It moves from identifying what people felt- the focus of Experiment 1 to understanding what they discussed, how these discussions evolved, and whether certain narratives gained or lost traction over time. For the following experiment, we followed a six-stage computational pipeline: (1) Temporal Binding, (2) Vectorization, (3) Topic Modeling, (4) Thematic Label Assignment, (5) Temporal Evolution of Themes, (6) Tracking Shifts in Public Discourse.

1. Temporal Binding:

To examine the evolution of public discourse surrounding menstruation-focused films over time, each dataset was temporally segmented based on the timestamp associated with each comment. This binning was crucial for enabling longitudinal topic modeling that tracks how dominant themes or concerns changed in public conversations during the two years following each film's release. Each raw comment collected from YouTube and Reddit included an associated timestamp (e.g., 2021-03-14T12:30:45Z), which denoted the precise date and time the comment was posted. Each dataset was anchored to the official release date of its corresponding film, treated as time zero (t = 0):

The absolute timestamp of each comment was converted into a relative offset in days from the release date of its film. This transformation allowed consistent binning regardless of the platform (YouTube or Reddit) or time zone inconsistencies. Comments were assigned to one of four six-month temporal bins post-release: Bin 1 (0-6 months), Bin 2 (7-12 months, Bin 3 (13-18 months), Bin 4 (19-24 months). For each of the four bins, a subset CSV file was generated per film

containing only the preprocessed comment text and relevant metadata (timestamp, platform, etc.). These bin-specific files became the direct input to the subsequent LDA-based topic modeling pipeline, allowing a per-bin topical analysis across time slices.

2. Vectorization:

Following the temporal binning of comments into six-month intervals based on film-specific release dates, each bin was transformed into a vectorized representation suitable for topic modeling using the Term Frequency Inverse Document Frequency (TF IDF) method.[12-15] TF IDF balances local and global importance by combining term frequency, which reflects how often a word appears within a given six-month bin, with inverse document frequency, which penalizes words that appear across many bins and therefore lack temporal distinctiveness. This weighting scheme enabled the model to downweight uninformative words while preserving terms that signal contextual or chronological shifts in discourse, such as "taboo," "education," or "SRHR." Each bin was treated as a standalone document created by concatenating all comments from that period, from which a document term matrix was constructed with rows representing time bins and columns representing unique vocabulary items. To ensure topic interpretability, vocabulary pruning was conducted by removing words that appeared in fewer than five comments across the entire dataset, as these rare terms often reflected typos or hyper-specific references, and excluding words present in more than eighty percent of bins, which functioned as secondary stopwords and hindered topic separability. This two-step filtering procedure reduced dimensionality, improved computational efficiency, and heightened the semantic clarity of the resulting topics. Both unigrams and bigrams were retained to capture meaningful collocations such as "menstrual hygiene" and "sanitary pad," ensuring that the final vectorization captured the evolving lexical landscape of menstruation discourse over time.

3. Topic Modeling:

To uncover latent thematic structures[39] in temporally grouped comments, Latent Dirichlet Allocation (LDA) was employed on the TF-IDF-weighted document-term matrices (DTMs) generated for each 6-month bin. LDA is a generative probabilistic model that assumes each document (in this case, a bin of comments) is a mixture of various topics, and each topic is a distribution over words. This unsupervised technique was used to identify evolving discursive themes over time. Each DTM bin was treated as a distinct "document," enabling the model to infer topics that are not static but evolve across the public discourse timeline. The number of topics was set to k = 5 per film (after empirical tuning), with hyperparameters optimized to balance topic coherence and separation. Two topics per interval reduce redundancy and enable clearer tracking of macro-level thematic shifts over time. This approach is particularly effective when analyzing multi-film, multi-platform datasets across a prolonged period, where interpretability and visual clarity are crucial. The resulting topics were characterized by their top 10 most representative terms, determined by term probability within each topic.

4. Thematic Label Assignment:

After extracting five topics per 6-month bin per film using LDA, the next step involved interpreting these topic-word distributions to assign higher-order thematic labels that could unify and summarize recurring patterns in the discourse. Since LDA outputs topics as distributions over words without any semantic labels, a human-in-the-loop approach was essential to meaningfully interpret and standardize the thematic coding across bins, platforms, and films. To address this, I utilized ChatGPT-4o, a large language model fine-tuned for interpretability and iterative reasoning, as a domain-assistive tool. This decision was motivated by the need for a consistent, scalable, and semantically aware process to handle the large volume of topic outputs while preserving thematic coherence across the dataset. Manual coding alone would have been prone to subjectivity, inconsistency, and fatigue-related error, especially given the volume (over 100 topics across 3 films and multiple time bins).[1,4,40-42] Rather than direct label generation, I employed chain-of-thought prompting, a method wherein the model is guided step-by-step through reasoning processes. Specifically, I provided ChatGPT with the following prompt:

"*I have performed LDA topic modeling on comment data about menstruation-related films, split across 6-month time bins. For each bin, 5 topics have been extracted, each represented by the top 10 keywords.*

Your task is to: Identify consistent overarching themes based on topic keywords across all bins and movies, assign each topic to one of these themes by interpreting the keywords, and use the same set of themes across all bins and films. Do not invent new ones for each time bin, and return your answer in a downloadable CSV file with the following columns:

(1) Movie Name (2) Time Bin (3) Topic Number (4) Top Keywords (5) Assigned Theme

I have attached the table for analysis."

This prompt was curated in such a way as to ensure that Chat-GPT did not stray from the task of assigning themes to the time bins. There was also no mention of the limit of themes Chat-GPT could assign as many themes as it deemed fit, instead of restricting it.

5. Temporal Evolution of Themes:

After assigning high-level thematic labels to the extracted LDA topics, the next analytical step was to examine how the prominence of these themes evolved for each film. This approach aimed to move beyond static topic interpretation, instead focusing on the temporal dynamics of discourse, identifying which themes gained traction, persisted, or declined in the months following each film's release. For each film, I calculated the relative prevalence of each theme within every six-month time bin. This was done by computing the proportion of LDA topics within a bin assigned to each theme. For example, if a time bin yielded five topics and two were categorized under "Menstrual Awareness," the theme's prevalence for that bin was recorded as 40%. This method allowed for

DOI: 10.36838/IJHSR811.5

a normalized comparison across time intervals and films, regardless of absolute topic counts. Thematic prevalence trends were then visualized using line charts and heatmaps, capturing longitudinal patterns of audience discourse. These visualizations highlighted how audience focus shifted in response to the film's release, public discussions, and possibly wider socio-political events.[1,4,40-42]

EXPERIMENT 3: Sentiment Polarity Over Time:

This experiment examines how the emotional tone of audience responses to menstruation-related films has evolved. Unlike Experiment 1, which investigated sentiment distributions within latent semantic clusters, and Experiment 2, which identified recurring themes through topic modeling, this analysis treats sentiment as an independent variable and tracks its trajectory across time. By aggregating sentiment labels across all comments and grouping them by year, this experiment aims to reveal whether public sentiment toward menstruation-themed media has become more positive, more negative, or remained stable over the past decade. The motivation for this analysis lies in the recognition that public attitudes toward menstruation are not fixed. As social awareness, media representation, and policy discourse surrounding menstrual health have shifted, audience reactions may reflect corresponding affective changes. For example, the release of government schemes, social media campaigns, or educational content may coincide with periods of increased positivity, while political controversies or backlash against feminist narratives may produce negativity spikes. By capturing the temporal dynamics of sentiment, this experiment offers insight into how menstruation discourse in digital spaces aligns with broader sociocultural shifts. This time-based sentiment analysis serves two key purposes. First, it provides a longitudinal perspective on emotional framing, helping to determine whether public discourse is growing more affirmational or resistant over time. Second, it complements the structural insights of the previous experiments by situating sentiment within a historical arc. In doing so, it supports the overarching goal of this study: to understand not just what is being said about menstruation in public forums, but how its emotional framing is evolving alongside cultural and political developments. For the following experiment we followed a five-stage computational pipeline: (1) Timestamp Join & Year Extraction (2017-2025) (2) Consolidation Across Models/Sources (3) Film-Level Segmentation (Pad Man / Period End of Sentence. / Alert Condition Red) (4) Stacked-Bar Visualization (5) Event Alignment & Interpretation (releases, awards, campaigns).

■ Results and Discussion

EXPERIMENT 1:

Dimensionality Reduction:

With MiniLM, PCA produced a clean, axis-aligned split between Reddit and YouTube, showing that even a linear projection captures big platform-specific lexical and topical differences. t-SNE further sharpened this divide into two dense, non-overlapping neighborhoods, indicating that the platforms' discourse differences are deeply encoded in MiniLM's semantic space. E5 showed weaker linear separation. Under PCA, Reddit and YouTube formed curved, partially overlapping clusters, suggesting that platform variance in E5 is more complex and less linearly expressed. t-SNE improved separation but still left fuzzy boundaries and mixed regions, implying that some cross-platform comments share similar semantic traits. DistilBERT, like MiniLM, showed strong linear separability under PCA and produced two tight, well-separated clusters under t-SNE. This consistency across methods suggests that DistilBERT also embeds clear platform-level distinctions despite being less sentiment-sensitive. Overall, MiniLM and DistilBERT display robust platform segregation, while E5 encodes platform differences more diffusely, becoming clearer only under non-linear reduction.

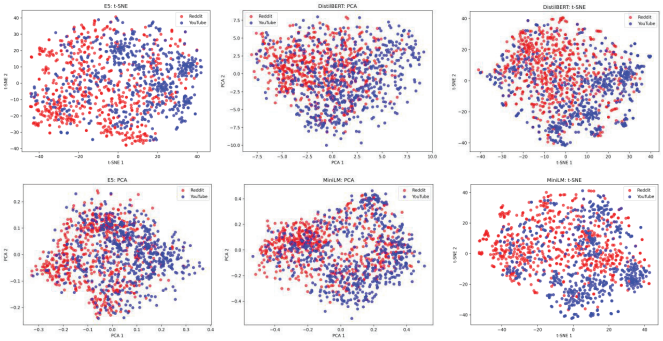

Figure 1: Two-dimensional projections of Pad man comments using PCA and t-SNE. MiniLM and DistilBERT show clear platform separation across both PCA and t-SNE. E5 shows more overlap, indicating shared semantics across platforms.

Clustering and Statistical Analysis:

Table 2: Platform distribution (YouTube vs. Reddit). Cluster membership shows no statistically significant association with platform. This suggests that any visual separation does not translate into meaningful platform-driven clustering.

Model	Cluster	% of Total	%Reddit	%YouTube	χ^2 (Platform)	p-value	Cramér's V (Platform)
MiniLM_PCA	0	44.97%	48.26%	**51.74%**	1.01	0.3140	0.030
	1	55.03%	**51.42%**	48.58%	1.01	0.3140	0.030
MiniLM_tSNE	0	48.09%	49.82%	**50.18%**	0.00	0.9530	0.002
	1	51.91%	**50.17%**	49.83%	0.00	0.9530	0.002
E5_PCA	0	46.61%	48.79%	**51.21%**	0.50	0.4785	0.021
	1	53.39%	**51.06%**	48.94%	0.50	0.4785	0.021
E5_tSNE	0	49.05%	49.03%	**50.97%**	0.35	0.5556	0.017
	1	50.95%	**50.94%**	49.06%	0.35	0.5556	0.017
DistilBERT_PCA	0	47.31%	**52.29%**	47.71%	2.01	0.1567	0.042
	1	52.69%	47.94%	**52.06%**	2.01	0.1567	0.042
DistilBERT_tSNE	0	51.74%	49.66%	**50.34%**	0.03	0.8596	0.005
	1	48.26%	**50.36%**	49.64%	0.03	0.8596	0.005

The X^2 tests across all models and dimensionality reduction combinations indicate no statistically significant association between platform (Reddit vs. YouTube) and cluster membership, as all p-values are well above the conventional 0.05 threshold. For instance, MiniLM with PCA produced two clusters that split the dataset into 44.97% and 55.03% of total comments, with a X^2 value of 1.01 ($p = 0.3140$) and a very small Cramér's V of 0.030, indicating negligible association strength. Similarly, MiniLM with t-SNE yielded an almost

perfectly balanced cluster distribution (48.09% vs. 51.91%) and an effectively zero association ($X^2 = 0.00$, $p = 0.9530$, Cramér's $V = 0.002$). E5 models also showed weak platform-cluster associations. E5 with PCA produced clusters of 46.61% and 53.39%, with $X^2 = 0.50$ ($p = 0.4785$) and Cramér's $V = 0.021$. E5 with t-SNE had an even more balanced split (49.05% vs. 50.95%) and equally low association ($X^2 = 0.35$, $p = 0.5556$, Cramér's $V = 0.017$). DistilBERT followed the same trend, with PCA-based clustering yielding 47.31% and 52.69% cluster shares, $X^2 = 2.01$ ($p = 0.1567$), and a slightly higher but still negligible Cramér's V of 0.042. DistilBERT with t-SNE again showed a near-even split (51.74% vs. 48.26%) and virtually no platform-cluster relationship ($X^2 = 0.03$, $p = 0.8596$, Cramér's $V = 0.005$). Overall, while visual inspection of PCA and t-SNE projections suggested that some models, particularly MiniLM and DistilBERT, produced relatively distinct platform-specific clusters, the statistical tests demonstrate that these apparent separations are not reflected in a strong, quantifiable relationship between platform identity and cluster membership. This indicates that while embeddings capture subtle semantic differences between Reddit and YouTube discourse, these differences do not translate into sharply segregated groups under unsupervised clustering.

Sentiment Analysis:

Table 3: Sentiment distribution across clusters. Cluster membership shows a statistically significant association with sentiment across all models with small to moderate effect sizes. This indicates that sentiment meaningfully influences clustering, but is not the sole driver.

Model	Cluster	% of Total	% Positive	% Negative	% Neutral	χ^2 (Sentiment)	p-value	Cramér's V (Sentiment)
MiniLM_PCA	0	44.97%	34.17%	12.93%	52.90%	62.82	0.00000	0.234
	1	55.03%	14.67%	20.66%	64.67%	62.82	0.00000	0.234
MiniLM_tSNE	0	48.09%	10.47%	20.04%	69.49%	100.02	0.00000	0.295
	1	51.91%	35.45%	14.55%	50.00%	100.02	0.00000	0.295
E5_PCA	0	46.61%	29.42%	17.88%	52.70%	23.20	0.00001	0.142
	1	53.39%	18.21%	16.59%	65.20%	23.20	0.00001	0.142
E5_tSNE	0	49.05%	34.69%	8.85%	56.46%	106.34	0.00000	0.304
	1	50.95%	12.61%	25.21%	62.18%	106.34	0.00000	0.304
DistilBERT_PCA	0	47.31%	33.76%	21.28%	44.95%	93.37	0.00000	0.285
	1	52.69%	14.17%	13.51%	72.32%	93.37	0.00000	0.285
DistilBERT_tSNE	0	51.74%	16.78%	17.95%	65.27%	31.01	0.00000	0.164
	1	48.26%	30.58%	16.37%	53.06%	31.01	0.00000	0.164

In contrast to the platform, clustering showed a statistically significant association with sentiment across all models and reductions, with chi-square tests consistently yielding p-values below 0.00001. This indicates a non-random relationship between cluster membership and sentiment distribution. However, the strength of this association, as measured by Cramér's V, ranged from 0.142 to 0.304 values that fall within the small to moderate effect size range. This suggests that while sentiment does influence cluster formation, it is not the sole or dominant driver of clustering outcomes. Models like MiniLM with t-SNE and E5 with t-SNE showed the strongest sentiment-cluster alignment (Cramér's $V \approx 0.30$), while others like E5 PCA and DistilBERT t-SNE exhibited weaker associations. Overall, this analysis suggests that semantic embeddings capture sentiment cues to a meaningful but moderate extent, supporting their use in sentiment-sensitive discourse analysis. Yet, the moderate effect sizes also caution against overinterpreting clusters as purely sentiment-driven; other latent factors likely contribute to cluster formation.

Conclusion:

Table 4: Overall distribution of positive and negative sentiment per platform. MiniLM (PCA) and E5 (t-SNE) show the most consistent patterns, with clearer splits across clusters.

Model	Cluster	Dominant Platform	Dominant Sentiment **(Ignoring Neutral)**
MiniLM PCA	0	**YouTube**	**Positive**
	1	**Reddit**	**Negative**
MiniLM tSNE	0	YouTube	Negative
	1	Reddit	Positive
DistilBERT PCA	0	Reddit	Positive
	1	**YouTube**	**Positive**
DistilBERT tSNE	0	YouTube	Negative
	1	Reddit	Positive
E5 PCA	0	**YouTube**	**Positive**
	1	Reddit	Positive
E5 tSNE	0	**YouTube**	**Positive**
	1	**Reddit**	**Negative**

MiniLM showed the clearest alignment between sentiment and cluster structure. Under PCA, YouTube-dominant clusters were far more positive (34.17 percent) than Reddit-dominant ones (14.67 percent), while Reddit clusters showed higher negativity. The association was statistically significant with a moderate effect size ($X^2 = 62.82$, Cramér's $V = 0.234$). MiniLM with t-SNE also produced significant results ($X^2 = 100.02$, Cramér's $V = 0.295$), although the polarity inverted, reflecting t-SNE's tendency to prioritize local structure and sometimes distort global trends. These results show that MiniLM encodes sentiment well, but dimensionality reduction strongly shapes how these distinctions surface. E5 behaved similarly in the t-SNE configuration. The YouTube-dominant cluster was strongly positive (34.69 percent) and minimally negative, whereas the Reddit-dominant cluster was markedly more negative (25.21 percent). This setting produced the largest effect size (Cramér's $V = 0.304$), indicating high sensitivity to sentiment. E5 PCA, however, mixed sentiment across clusters and showed a weaker association ($X^2 = 23.20$, Cramér's $V = 0.142$), suggesting that PCA suppresses E5's finer sentiment cues. DistilBERT produced the weakest and most inconsistent patterns. Both PCA and t-SNE failed to reveal stable sentiment–cluster alignment, and some distributions contradicted expected trends. Effect sizes remained small, indicating that DistilBERT embeddings do not reliably encode sentiment within noisy or code-mixed social media text. Overall, only MiniLM PCA and E5 t-SNE generated coherent, interpretable sentiment–platform patterns, underscoring that both embedding choice and reduction method are critical for sentiment-sensitive clustering.

DOI: 10.36838/IJHSR811.5

EXPERIMENT 2: LDA Topic Modeling:

Table 5: Alert Condition: Red. Discourse is dominated by film reviews and menstrual awareness, with low engagement in gender or social impact themes.

Time Bin	Film Review	Gender Roles & Equality	Menstrual Awareness	Social Change / Impact
2017-12	60.0	40.0	0.0	0.0
2018-06	40.0	20.0	40.0	0.0
2018-12	20.0	20.0	40.0	20.0
2019-06	60.0	0.0	40.0	0.0
2019-12	60.0	20.0	20.0	0.0
2020-06	40.0	0.0	60.0	0.0
2020-12	60.0	0.0	40.0	0.0
2021-06	20.0	20.0	60.0	0.0
2021-12	60.0	0.0	40.0	0.0

The temporal analysis of *Alert Condition: Red* (Table 5) shows that audience discourse was dominated by review-focused commentary across most periods, with Film Review accounting for the largest share of topics in multiple bins. Menstrual Awareness appeared consistently and ranged from 20% to 60% over time, indicating steady engagement with menstrual themes without a clear upward or downward trend. By contrast, Gender Roles and Equality appeared only sporadically and never exceeded 40%, reflecting limited public focus on gendered discourse despite the film's subject matter.[1,4,5] Social Change or Impact was almost absent across bins, suggesting that this short film did not catalyze activist-oriented or policy-driven discussion.[2] Overall, discourse surrounding *Alert Condition: Red* remained centered on reviews and basic menstrual-awareness commentary rather than deeper socio-political engagement.

Table 6: Period. End of Sentence. Discourse consolidates around menstrual awareness over time, with other themes largely disappearing after the early periods.

Time Bin	Film Review	Gender Roles & Equality	Menstrual Awareness
2020-04	40.0	20.0	40.0
2020-10	20.0	0.0	80.0
2021-04	0.0	0.0	100.0
2021-10	0.0	0.0	100.0
2022-04	0.0	40.0	60.0
2022-10	0.0	20.0	80.0
2023-04	20.0	0.0	80.0
2023-10	0.0	0.0	100.0
2024-04	0.0	0.0	100.0
2024-10	0.0	0.0	100.0
2025-04	0.0	0.0	100.0
2025-10	20.0	40.0	40.0

The thematic evolution of Period. End of Sentence. (Table 6.) differs sharply from the other films. Menstrual Awareness increased steadily over time and became the exclusive topic in multiple bins from 2021 onward, reflecting the documentary's strong and enduring influence on menstrual-health discourse.[1,3,4] Film Review declined sharply after the documentary's Oscar win[40-42] and disappeared entirely after 2023 as viewers shifted toward deeper engagement with menstruation rather than cinematic critique. Gender Roles and Equality appeared only in isolated intervals with minimal prominence.[5,6] This sustained rise in Menstrual Awareness suggests that the film generated long-term issue-focused engagement, reinforcing observations in menstrual-health scholarship that visibility can influence public discourse even after media attention decreases.[3,4] Overall, the film shows a clear trajectory from mixed early commentary to near-complete thematic consolidation around menstrual awareness.

Table 7: Padman. Discourse shifts from actor and review-driven early engagement to menstrual awareness over time, with limited focus on gender themes.

Time Bin	Actor Appreciation	Film Review	Gender Roles & Equality	Menstrual Awareness
2017-12	40.0	60.0	0.0	0.0
2018-06	40.0	0.0	0.0	60.0
2018-12	20.0	40.0	0.0	40.0
2019-06	60.0	20.0	0.0	20.0
2019-12	60.0	0.0	0.0	40.0
2020-06	20.0	20.0	20.0	40.0
2020-12	40.0	20.0	0.0	40.0
2021-06	20.0	20.0	0.0	60.0
2021-12	40.0	0.0	20.0	40.0
2022-06	40.0	20.0	20.0	20.0
2022-12	100.0	0.0	0.0	0.0
2023-06	60.0	40.0	0.0	0.0
2023-12	0.0	0.0	0.0	100.0
2024-06	0.0	100.0	0.0	0.0
2024-12	40.0	0.0	0.0	60.0
2025-06	0.0	0.0	0.0	100.0

For *Pad Man* (Table 7), early discourse from 2017 to 2020 was dominated by Actor Appreciation and Film Review, reflecting Akshay Kumar's celebrity influence and the film's mainstream visibility at release.[40,41,43] Over time, particularly after 2021, Menstrual Awareness became increasingly prominent and eventually reached 100% dominance in 2023 and 2025, indicating a delayed but deepening public engagement with menstrual-health themes. Although Film Review resurged briefly in 2024, gender-focused discussions remained marginal throughout, with Gender Roles and Equality appearing infrequently and never gaining thematic traction.[5,6] Actor Appreciation declined sharply after 2022, suggesting a long-term shift away from celebrity-centric discourse toward issue-centric conversations. This pattern indicates that *Pad Man*, although initially interpreted through the lens of star power, gradually seeded more substantive discussions on menstrual awareness.

EXPERIMENT 3:

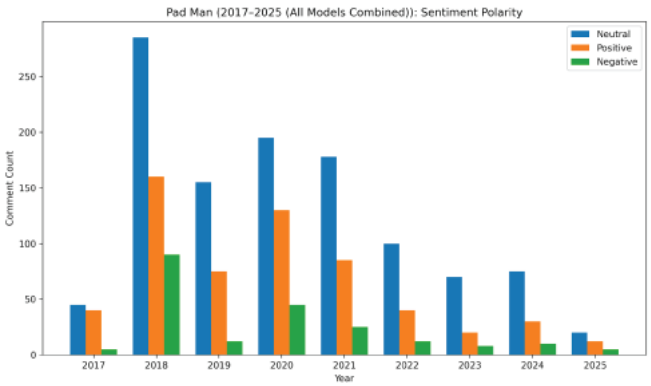

Figure 2: Sentiment polarity over time for Pad Man. Sentiment is largely neutral and event-driven.

The bar chart (Figure 2) shows how sentiment toward menstruation-themed films shifted between 2017 and 2025. Neutral sentiment dominates across the entire period, indicating that most commentary is explanatory or observational, consistent with prior work showing that menstrual discourse online often adopts an informational tone when addressing stigma, education, or hygiene access.[1,3,4] A sharp rise in sentiment activity appears in 2018, driven by the release of Pad Man in February 2018.[7,41,43] The film generated extensive online discussion, supported by campaigns such as the #PadManChallenge, and produced high positive sentiment reflecting public support for its awareness message.[4,5] The simultaneous negative sentiment reflects critiques of its narrative framing and gender representation, issues commonly noted in feminist analyses of menstrual media.[5,6] A second major increase occurs in 2020 with the visibility of the period. End of Sentence, whose Academy Award win renewed attention to menstrual equity.[40,42] Positive sentiment reflects celebratory responses, while negative sentiment reflects concerns about Western framing and simplified portrayals of stigma, consistent with broader debates on representation.[1,4,6] A smaller rise in 2021 aligns with pandemic-era discussions of menstrual health, including period poverty and policy initiatives.[44,45] Many comments during this time revisited earlier films as reference points within activism and public health conversations.[3,4] From 2022 onward, sentiment activity declines sharply, reflecting reduced engagement in the absence of major releases or public events. Neutral sentiment remains dominant, indicating that discourse continues mostly in analytical or educational contexts, mirroring the long tail of advocacy-oriented media.[46] Overall, sentiment trends show that engagement with menstruation-themed films is highly event-driven, with peaks around major releases and sociopolitical moments. Pad Man and Period. End of Sentence. shaped public discourse beyond their release years, but declining activity after 2021 underscores the limited durability of widespread attention to menstrual health.

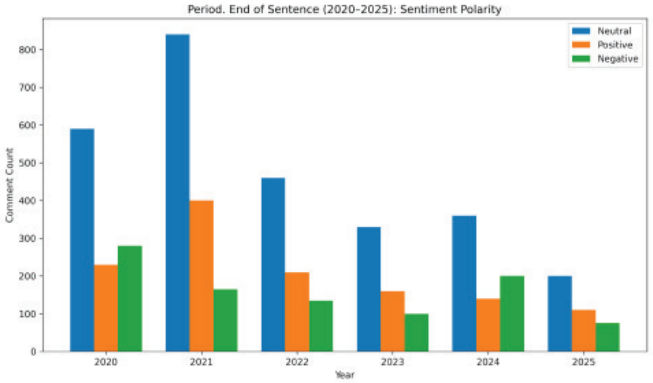

Figure 3: Sentiment polarity over time for the period. End of Sentence. Sentiment peaks around the Oscar win and early pandemic period, then declines steadily.

The documentary Period. End of Sentence, which premiered on Netflix and later won the Academy Award for Best Documentary Short, marked a pivotal moment in global media engagement with menstrual equity.[40,42] The sentiment chart (Figure 3) shows how public reactions shifted between 2020 and 2025 in response to media cycles, award recognition, and broader sociopolitical developments. In early 2020, comment volume was modest, yet sentiment skewed strongly positive, with viewers praising the documentary's portrayal of menstrual stigma, its focus on rural India, and its amplification of marginalized women's voices.[1-4] A smaller negative response reflected concerns about Western framing, brevity, and the risk of oversimplifying culturally embedded issues.[5,6] Neutral commentary consisted largely of descriptive summaries typical of early-viewer engagement. The strongest surge in sentiment occurred following the Oscar win, which amplified global attention to menstrual health.[40,42] Positive reactions expressed pride in the international recognition, while negative sentiment grew as viewers questioned whether a Western-produced documentary could authentically represent Indian experiences, reflecting a tension noted in menstrual health scholarship between visibility and representation.[1,4-6] Sentiment remained elevated into 2021 during a period of expanding menstrual equity initiatives and discussions of period poverty amid the COVID-19 pandemic.[44,45] Many comments referenced the documentary as a symbolic touchpoint rather than a standalone media object, and neutral sentiment dominated as viewers contextualized it within ongoing activism and policy efforts.[1-3] From 2022 onward, engagement declined sharply, consistent with the lifecycle of advocacy-driven media that lose visibility without continued institutional or cultural reinforcement.[46] Later comments were largely reflective, subdued, or folded into broader conversations about documentary impact and gender-focused media.[4,6] Overall, the sentiment trajectory reveals a familiar pattern: an initial burst of enthusiasm, critical revaluation, symbolic longevity, and eventual attenuation.[1,3,4] While Period. End of Sentence sparked global interest in menstrual health, sustained public discourse required continued policy action, educational engagement, and follow-up media efforts.

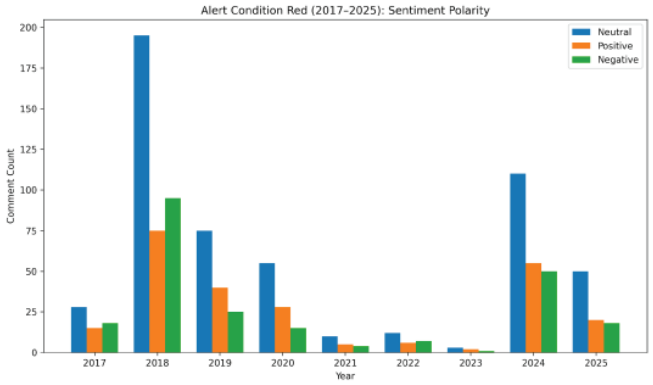

Figure 4: Sentiment polarity over time for Alert Condition Red. Sentiment remains low and mostly neutral throughout, with brief early peaks.

The bar graph (Figure 4) traces sentiment toward Alert Condition Red across nearly a decade, revealing modest engagement but a stable pattern of discourse. Early peaks in 2017 and 2018 reflect its initial circulation in academic and activist spaces, where menstruation-themed media often first gain traction.[4,5] Neutral sentiment dominates these early years, indicating largely descriptive or informational responses consistent with typical viewer engagement with education-

DOI: 10.36838/IJHSR811.5

al menstrual health content.[1,3] Positive sentiment appears in smaller proportions, aligning with viewers who appreciated the film's metaphor-driven narrative style, noted in feminist media studies.[6] Negative sentiment reflects critiques of abstraction and narrative opacity, echoing broader discussions of menstrual stigma representation.[5] Engagement declines sharply after 2018 due to the film's limited mainstream distribution and absence of media scaffolding compared to Pad Man or Period. End of Sentence.[30,42] A small resurgence in 2021 coincides with heightened pandemic-era discussions of menstrual equity, period poverty, and digital activism.[44,45] Neutral sentiment remains dominant, suggesting the film functioned mainly as a pedagogical reference within these conversations.[4,6] From 2022 to 2025, sentiment activity remains minimal, with occasional mentions likely emerging from classrooms, NGO training, or academic contexts, reflecting the film's transition into a long tail archival role typical of low-budget menstrual health media.[1,3,46] Overall, Alert Condition Red did not generate widespread enthusiasm but maintained an enduring presence as a niche resource revisited within educational and advocacy settings.[4,6]

Conclusion

Across approximately twenty-five thousand Reddit and YouTube comments, the findings show that public engagement with menstruation is episodic, driven primarily by media events rather than stable cultural attention. Experiment 1 demonstrates that while platform-based patterns are visually detectable in embedding space, platform identity does not meaningfully structure cluster formation. Instead, sentiment is the dominant organizing force, with strong statistical associations and clear separation between more positive YouTube-leaning clusters and more negative Reddit-leaning ones. Experiment 2 shows that thematic attention shifts over time in response to film releases, discussions of stigma, and moments of advocacy, but that these shifts lose momentum as visibility fades. Experiment 3 confirms that neutral sentiment dominates overall and that spikes in engagement correspond to specific cultural triggers, such as Pad Man in 2018, Period. End of Sentence in 2020, and pandemic era conversations in 2021. Across all analyses, the central pattern is clear: menstrual discourse online is reactive, sentiment-driven, and highly sensitive to external visibility. Without new media or policy signals, public attention dissipates quickly, suggesting that the sustainability of menstrual health conversations in digital spaces depends on repeated moments of renewed cultural activation.

Acknowledgments

This study was completed under the Indigo Research Program; I am grateful to the Indigo team for providing structured mentorship and a rigorous research framework. (indigoresearch.org) I extend my deepest thanks to my research mentor, Samuel Lefcourt of Johns Hopkins University (JHU), for his guidance and thoughtful feedback on successive drafts. Any remaining errors are my own. This research received no external funding. I attest that the ideas, graphics, and writing in this paper are entirely my own.

References

1. Winkler, I.; Roaf, V. Taking the Bloody Linen Out of the Closet: Menstrual Hygiene as a Priority for Achieving Gender Equality. *Cardozo J. Law Gender* **2014**, *21* (1), 1-37.
2. House, S.; Mahon, T.; Cavill, S. *Menstrual Hygiene Matters: A Resource for Improving Menstrual Hygiene around the World*; WaterAid: London, **2012**.
3. Sommer, M.; Hirsch, J. S.; Nathanson, C.; Parker, R. Comfortably, Safely, and Without Shame: Defining Menstrual Hygiene Management as a Public Health Issue. *Am. J. Public Health* **2015**, *105* (7), 1302-1311.
4. Bobel, C. *The Managed Body: Developing Girls and Menstrual Health in the Global South*; Palgrave Macmillan: Cham, **2019**.
5. Johnston-Robledo, I.; Chrisler, J. C. The Menstrual Mark: Menstruation as Social Stigma. *Sex Roles* **2013**, *68* (1-2), 9-18.
6. Sasser, J. S. Cycles of Shame: Menstrual Stigma in Media and Popular Culture. *Fem. Media Stud.* **2014**, *14* (2), 146-161.
7. De Choudhury, M.; Counts, S.; Horvitz, E. Social Media as a Measurement Tool of Depression in Populations. *Proc. ACM WebSci* **2013**, 47-56.
8. Pad Man. Directed by R. Balki; Sony Pictures India: Mumbai, **2018**.
9. Period. End of Sentence. Directed by R. Zehtabchi; Netflix: Los Angeles, **2018**.
10. Alert Condition: Red. Menstrual Hygiene Short Film; YouTube: **2020**.
11. Haimson, O. L.; Andalibi, N.; De Choudhury, M.; Hayes, G. R. Relationship Breakups on Social Media: Uncertainty, Immediacy, and Support Seeking. *Proc. CSCW* **2018**, 1-22.
12. Salton, G.; McGill, M. J. *Introduction to Modern Information Retrieval*; McGraw-Hill: New York, **1983**.
13. Jones, K. S. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *J. Doc.* **1972**, *28* (1), 11-21.
14. Ramos, J. Using TF-IDF to Determine Word Relevance in Document Queries. *Proc. First Instructional Conf. Machine Learning*, Rutgers University, **2003**.
15. Aizawa, A. An Information-Theoretic Perspective of TF-IDF Measures. *Inf. Process. Manag.* **2003**, *39* (1), 45-65.
16. Sarker, A.; DeRoos, A.; Perrone, J. Mining Social Media for Prescription Medication Abuse Monitoring: A Review and Proposal for a Data-Centric Framework. *J. Am. Med. Inform. Assoc.* **2020**, *27* (2), 315-329.
17. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
18. Pennington, J.; Socher, R.; Manning, C. D. GloVe: Global Vectors for Word Representation. *Proc. EMNLP* **2014**, 1532-1543.
19. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proc. NAACL-HLT* **2019**, 4171-4186.
20. Sun, C.; et al. How to Fine-Tune BERT for Text Classification? *China Natl. Conf. Chinese Comput. Linguistics* 2019, 194-206.
21. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings Using Siamese BERT Networks. *Proc. EMNLP* **2019**, 3982-3992.
22. Gao, T.; Yao, X.; Chen, D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. *Proc. EMNLP* **2021**, 6894-6910.
23. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv* **2019**, arXiv:1910.01108.
24. Wang, W.; Wei, F.; Dong, L.; Bao, H.; Yang, N.; Zhou, M. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-trained Transformers. *Adv. Neural Inf. Process. Syst.* **2020**, 33, 5776-5788.

25. Barbieri, F.; Camacho-Collados, J.; Espinosa-Anke, L.; Neves, L. Tweeteval: Unified Benchmark and Comparative Evaluation for Tweet Classification. *Findings EMNLP* **2020**, 1644-1654.
26. Banda, J. M.; Tekumalla, R.; Wang, G.; Yu, J.; Liu, T.; Ding, Y.; Artemova, E.; Tutubalina, E.; Chowell, G. A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research. *Epidemiologia* **2021**, *2* (3), 315-324.
27. Cotfas, L.-A.; Delcea, C.; Roxin, I.; Ioanăş, C.; Gherai, D. S.; Tajariol, F. The Longest Month: Analyzing COVID-19 Misinformation on Twitter. *Healthcare* **2021**, *9* (1), 82.
28. Mackey, T. K.; Li, J.; Purushothaman, V.; et al. Big Data, Natural Language Processing, and Deep Learning to Detect and Characterize Illicit COVID-19 Product Sales. *JMIR Public Health Surveill.* **2020**, *6* (2), e20794.
29. Andalibi, N.; Haimson, O. L.; Choudhury, M. D.; Forte, A. Understanding Social Media Disclosures of Sexual Abuse Through the Lenses of Support Seeking and Anonymity. *Proc. CHI Conf. Hum. Factors Comput. Syst*. **2016**, 3906-3918.
30. Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T. Bag of Tricks for Efficient Text Classification. *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL)* **2017**, 427-431.
31. Shuyo, N. *Language Detection Library for Java*. arXiv **2010**, arXiv:1007.1518.
32. Honnibal, M.; Johnson, M. *An Improved Non-monotonic Transition System for Dependency Parsing*. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics: Lisbon, **2015**; pp 1373-1378.
33. Manning, C. D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, **2008**.
34. Solorio, T.; Blair, E.; Maharjan, S.; Bethard, S.; Diab, M.; Ghoneim, M.; Hawwari, A.; AlGhamdi, F.; Hirschberg, J.; Chang, A.; Fung, P. Overview for the First Shared Task on Language Identification in Code-Switched Data. *Proc. First Workshop on Computational Approaches to Code Switching* **2014**, 62-72.
35. Wang, L.; Yang, N.; Huang, X.; Jiao, B.; Yang, L.; Jiang, D.; Majumder, R.; Wei, F. Text Embeddings by Weakly-Supervised Contrastive Pre-training. *arXiv* **2022**, arXiv:2212.03533.
36. Jolliffe, I. T. *Principal Component Analysis*, 2nd ed.; Springer: New York, **2002**.
37. van der Maaten, L.; Hinton, G. Visualizing Data Using t-SNE. *J. Mach. Learn. Res*. **2008**, *9*, 2579-2605.
38. MacQueen, J. B. Some Methods for Classification and Analysis of Multivariate Observations. *Proc. Fifth Berkeley Symp. Math. Stat. Prob*. **1967**, *1*, 281-297.
39. Blei, D. M.; Ng, A. Y.; Jordan, M. I. Latent Dirichlet Allocation. *J. Mach. Learn. Res*. **2003**, *3*, 993-1022.
40. The Guardian. Period. End of Sentence. Wins Oscar for Best Documentary Short. *The Guardian*, Feb 25, **2019**.
41. Indian Express. How Pad Man Started a Conversation on Menstrual Hygiene. *Indian Express*, Mar 2018.
42. BBC News. Period. End of Sentence: The Oscar-Winning Documentary about Menstrual Taboos. *BBC News*, Feb 2019.
43. The Times of India. Pad Man Review: A Film That Breaks Taboos. *The Times of India*, Feb 9, **2018**.
44. Kuhlmann, A. S.; Bergquist, E. P.; Danjoint, D.; Wall, L. L. Unmet Menstrual Hygiene Needs Among Low-Income Women. *Obstet. Gynecol*. **2019**, 133 (2), 238-244.
45. Lindberg, L. D.; VandeVusse, A.; Mueller, J.; Kirstein, M. Early Impacts of the COVID-19 Pandemic: Findings from the 2020 Guttmacher Survey of Reproductive Health Experiences. Perspect. *Sex. Reprod. Health* **2020**, *52* (4), 189-198.
46. Syred, J.; Naidoo, C.; Woodhall, S. C.; Baraitser, P. Would You Tell Everyone This? Facebook Conversations as Health Promotion Interventions. *J. Med. Internet Res*. **2014**, *16* (4), e108.

Author

Aashi Gupta is a senior at The International School Bangalore with a strong interest in data science and computational research. She plans to in the future pursue Data Science at the university level in the future.

DOI: 10.36838/IJHSR811.5