Analyzing Social Media Responses Across Various Languages Based On RG Kar Incident

Sooraj Kakkanatt Manoj , Yogesh Kuchimanchi
Rochester Institute of Technology, Department of Data Science, Rochester, NY, USA
sk2716@g.rit.edu, yy6706@g.rit.edu

Abstract

This study explores the multilingual dynamics of public discourse surrounding the RG Kar Hospital incident, an event that brought governance, safety, and justice issues into the spotlight. Using data from YouTube comments in five languages—translated and preprocessed for uniform analysis—the research employs sentiment analysis with VADER and topic modeling using Latent Dirichlet Allocation (LDA) to uncover recurring themes. Contrary to initial expectations of a focus on misogyny, the analysis revealed significant public dissatisfaction and an unexpected dominance of political discourse, particularly around administrative failures. These findings emphasize the power of public sentiment analysis in highlighting societal concerns and underline the importance of addressing both governance issues and underlying sociopolitical dissatisfaction.

Keywords: Social Media Analysis, Multilingual Sentiment, Governance, Public Discourse, Topic Modeling

1 Introduction

Kolkata, India – The RG Kar Medical College and Hospital, one of the premier healthcare institutions in West Bengal, has become the epicenter of a harrowing controversy that has gripped the nation. Allegations of systemic inefficiencies, administrative negligence, and a deeply disturbing case of sexual violence have raised urgent questions about governance, accountability, and public safety.

The crisis erupted when reports surfaced of an alleged sexual assault on the hospital premises, sending shockwaves through the medical community and society at large. This incident came to light amid ongoing concerns about the state of public healthcare infrastructure and its capacity to ensure both patient safety and employee security. Public outrage intensified when initial reports suggested a lack of swift administrative action, further fueling skepticism about the commitment to justice.

In response, doctors, nurses, and medical students organized widespread protests, demanding accountability and the implementation of stricter safety protocols. Their demonstrations, marked by resilience and unity, echoed beyond hospital walls, sparking debates about the treatment of women, the responsibilities of authorities, and the role of protests in effecting systemic change.

The protests also became a political flashpoint, with opposing factions attempting to capitalize on the public discontent. Critics accused the ruling administration of failing to address long-standing issues within the healthcare system, while supporters pointed to efforts aimed at overhauling safety measures.

Social media platforms, particularly YouTube, became arenas for public discourse, with netizens expressing their anguish, anger, and demands for reform. Multilingual discussions, spanning Bengali, Hindi, Tamil, Telugu, and English, reflected the diversity of India's voices and their shared discontent over the incident. Common themes included a demand for justice, critiques of governance, and calls for institutional accountability.

As the RG Kar Hospital incident continues to unfold, it underscores the urgent need for reform in public institutions. The collective outcry demonstrates that the public is no longer willing to accept systemic failures without questioning their roots.

This incident serves as a stark reminder of the challenges faced by India's public healthcare system and the pressing need for reforms that prioritize safety, transparency, and accountability. It also highlights the power of public sentiment and collective action in holding authorities accountable. As investigations proceed and policy discussions unfold, the RG Kar Hospital case may well mark a turning point in how such incidents are addressed in the future.

2 Methodology

To analyze public sentiment and uncover key themes related to the RG Kar Hospital incident, this study adopted a systematic approach that involved data collection, preprocessing, sentiment analysis, and topic modeling. The methods used ensure the multilingual and diverse nature of the comments

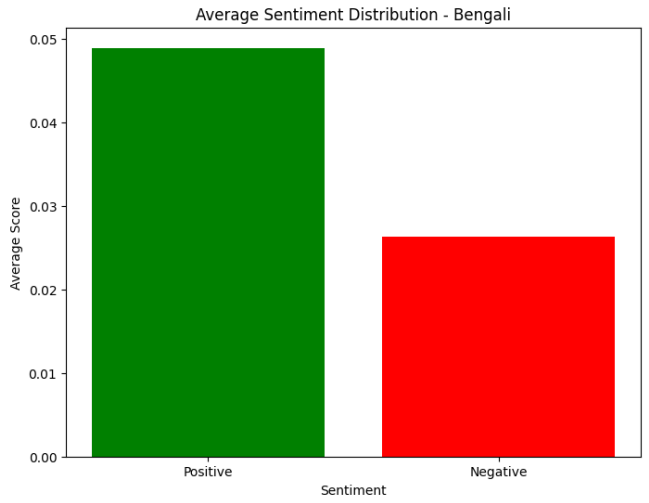

Figure 1: Sentiment distribution for Bengali comments using VADER.

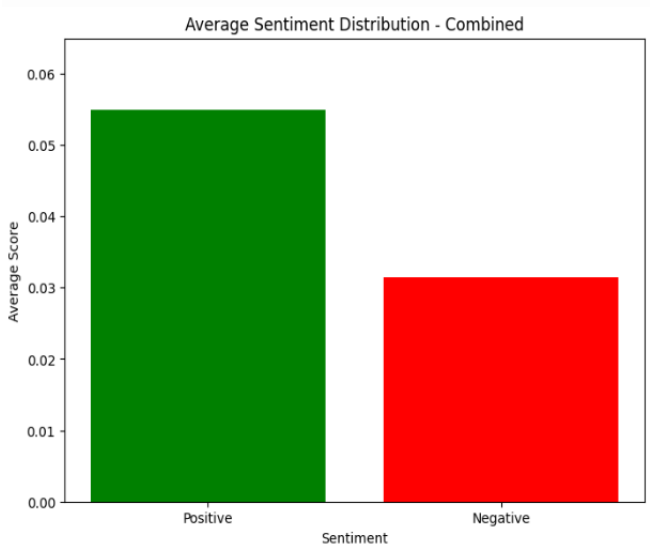

Figure 2: Sentiment distribution for all comments using VADER.

was appropriately processed and analyzed. The following subsections detail the steps taken to conduct this research.

2.1 Data Collection

The data collection process for this study was conducted in three sequential steps to ensure relevance and accuracy in capturing public sentiment regarding the RG Kar Hospital incident.

First, video IDs were extracted using the *ScrapeTube* library. This involved targeting YouTube channel IDs of various national and regional news channels. By programmatically scraping these channels, a comprehensive list of video IDs related to recent news events was generated.

Next, the *YouTube Data API v3* was employed to gather metadata for the extracted video IDs. This included details such as video titles, descriptions, and publication dates. The dataset was then filtered to retain only videos uploaded on or after *August 5, 2024*, the date of the RG Kar Hospital incident. This step ensured temporal relevance, focusing the analysis on public discourse surrounding the event and its aftermath.

In the final step, the *text-embedding-ada-002 model* from OpenAI was used to refine the dataset further. Video titles and descriptions were embedded and compared against a set of predefined contextual keywords, including *"RG Kar," "hospital rape case," "Kolkata doctors,"* and *"medical protests."* Only videos with high contextual relevance to these keywords were retained for analysis.

This multi-step approach ensured the collection of a targeted and high-quality dataset representing diverse public reactions across languages, setting the stage for subsequent sentiment analysis and topic modeling.

2.2 Data Preprocessing

To ensure the data was clean and relevant for analysis, the following preprocessing steps were performed:

- Translation of non-English comments using Google Translator API.
- Removal of non-alphanumeric characters, redundant whitespace, and irrelevant text.
- Filtering comments to retain those discussing the RG Kar Hospital incident, using keywords such as *"RG Kar," "hospital rape case," "Kolkata doctors," and "medical protests."*

2.3 Sentiment Analysis

Sentiment analysis was conducted using the **VADER** (Valence Aware Dictionary and Sentiment Reasoner) tool. VADER was chosen due to its ability to analyze short, social-media style texts effectively, making it particularly suitable for comments on platforms like YouTube. Comments were classified into three categories:

- Positive
- Neutral
- Negative

The sentiment scores were aggregated to analyze public sentiment trends across languages, providing insights into public reactions to the incident and associated protests.

2.4 Topic Modeling

To identify key themes within the dataset, the **Latent Dirichlet Allocation (LDA)** algorithm was applied. LDA is a widely used unsupervised machine learning technique that effectively uncovers latent topics within unstructured text data. This approach was chosen due to its ability to handle multilingual comments and extract meaningful patterns, making it a robust method for analyzing public discourse.

The preprocessing steps for topic modeling included:

Topic	Top Keywords
Topic 1	governance, accountability, protest, action, reform
Topic 2	justice, safety, rights, victim, incident
Topic 3	political, dissatisfaction, leadership, administration, change
Topic 4	public, healthcare, support, response, reform
Topic 5	demand, action, anger, corruption, failure

Table 1: Topic modeling results with associated keywords.

- **Tokenization:** Splitting the text into individual tokens (words) for analysis.
- **Stopword Removal:** Removing common stopwords to focus on meaningful terms while retaining politically relevant terms. Custom stopwords used in this analysis included words such as *"govt," "mamta," "tmc," "bjp,"* and *"banerjee,"* which were retained to ensure the political context of the discussion was preserved.
- **Bag-of-Words Representation:** Converting the preprocessed text into a bag-of-words representation, where each comment was represented as a collection of word frequencies for analysis.

The LDA model was applied to this processed data to uncover recurring themes and topics. For each topic, the model identified the most relevant keywords, which were analyzed to interpret underlying public concerns. This approach allowed the study to reveal themes such as governance issues, calls for justice, and political dissatisfaction, highlighting the depth of societal reactions to the RG Kar Hospital incident.

3 Results and Discussion

The analysis of YouTube comments provided valuable insights into public sentiment and recurring themes surrounding the RG Kar Hospital incident. The results highlighted the multilingual and multifaceted nature of public discourse, emphasizing a strong focus on governance, justice, and political dissatisfaction. The following subsections detail the findings from sentiment analysis and topic modeling.

3.1 Sentiment Analysis Results

Sentiment analysis using the VADER tool provided insights into public sentiment across the five languages studied (Figure 2). The results indicated that negative sentiments dominated across all languages, highlighting widespread dissatisfaction with the RG Kar Hospital incident and subsequent administrative responses.

Malayalam and Bengali displayed particularly strong negative sentiments, reflecting distrust and anger toward governance and justice systems (Figure 1). Positive sentiments, though less prevalent, often acknowledged the resilience of protesters or supported calls for reform. Neutral sentiments captured informative or descriptive comments that provided context to the discussions.

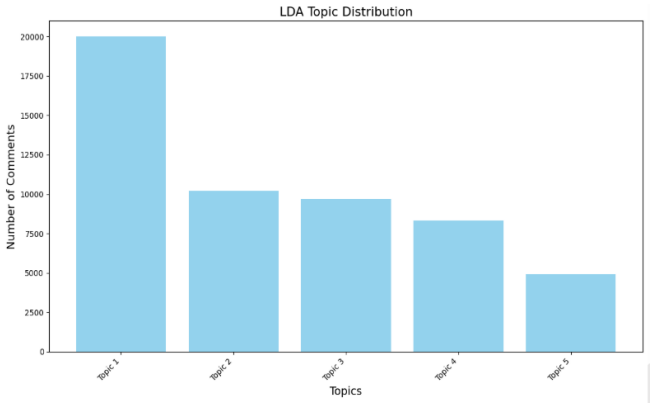

Figure 3: Topic distribution across multilingual comments.

3.2 Topic Modeling Results

The LDA model extracted five key themes from the multilingual comments, shedding light on public concerns (Figure 3):

- **Topic 1: Governance issues and systemic failures.** This topic included discussions about administrative inefficiencies, lack of safety protocols, and calls for accountability in public institutions.
- **Topic 2: Justice and accountability.** Comments in this category emphasized the need for justice for victims and strict punishment for perpetrators, highlighting public anger over perceived delays in action.
- **Topic 3: Political dissatisfaction and protests.** A significant portion of the comments criticized the ruling administration and political actors, reflecting frustration with governance and political opportunism during the crisis.
- **Topic 4: Social and cultural dynamics.** Comments in this topic addressed societal attitudes, cultural perceptions, and the role of public protests in driving change.
- **Topic 5: Institutional reforms.** This theme captured calls for systemic reforms, improved safety measures, and better governance, emphasizing the need for structural changes in public institutions.

The prevalence of political discourse, often tied to administrative inefficiencies, was unexpected, given the initial assumption of a focus on misogyny. Instead, the analysis revealed a broader dissatisfaction with governance and the political landscape.

3.3 Key Findings

The analysis revealed two primary outcomes:

1. Negative sentiment dominated, emphasizing public dissatisfaction.
2. Political discourse overshadowed the expected focus on misogyny, reflecting deep-seated societal frustrations.

These findings highlight the role of sentiment analysis in identifying governance and safety concerns, even in incidents tied to gender-based violence, and underline the interconnectedness of social, cultural, and political issues.

3.4 Limitations

This study has several limitations. First, relying solely on YouTube comments may not capture the full range of public opinion across other social media platforms, potentially introducing demographic biases.

Second, automated translations, while enabling multilingual analysis, may misinterpret cultural nuances or idiomatic expressions, affecting sentiment and topic modeling accuracy.

Third, VADER, though effective for short-form text, struggles with detecting sarcasm and complex emotional contexts. Similarly, LDA assumes distinct topics, potentially oversimplifying overlapping topics in public discourse.

Future work should expand data sources, use advanced multilingual tools, and explore alternative modeling techniques for a more comprehensive analysis.

4 Conclusion

This study investigated the public sentiment and key themes surrounding the RG Kar Hospital incident using multilingual data from YouTube comments. The analysis provided critical insights into societal reactions, uncovering widespread public dissatisfaction and an unexpected focus on political discourse.

Sentiment analysis revealed that negative sentiments overwhelmingly dominated, reflecting deep frustration with governance and administrative inefficiencies. While the initial expectation was a focus on misogyny and gender-related issues, the prevalence of political themes highlighted the complex interplay between social concerns and political dissatisfaction.

The use of multilingual sentiment analysis and topic modeling proved instrumental in capturing diverse public opinions. These findings underscore the importance of leveraging social media discourse to understand public sentiment on sensitive socio-political issues.

Future research could explore the evolution of public sentiment over time and analyze the impact of administrative actions on societal trust. Additionally, expanding the dataset to include other social media platforms could provide a more comprehensive understanding of public opinion.

Acknowledgments

We express our heartfelt gratitude to our professor, Ashique KhudaBukhsh, for his invaluable mentorship, insightful guidance, and constant encouragement throughout this research. His feedback significantly enriched the quality and depth of this study.

We also acknowledge the use of essential tools and resources, including ScrapeTube, the YouTube Data API, and OpenAI's text-embedding model, which played a crucial role in data collection and analysis.

References

[Blei *et al.*, 2003] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Gottlob *et al.*, 2002] G. Gottlob, N. Leone, and F. Scarcello. Hypertree decompositions and tractable queries. *Journal of Computer and System Sciences*, 64(3):579–627, 2002.

[Indian Medical Association, 2019] Indian Medical Association. Over 75% of doctors have faced violence at work, study finds. 2019.

[Levesque, 1984] H. J. Levesque. Foundations of a functional approach to knowledge representation. *Artificial Intelligence*, 23(2):155–212, 1984.

[OpenAI Embedding Models, 2024] OpenAI Embedding Models. Documentation for text embedding models, including text-embedding-ada-002, 2024.

[Ranjan, 2017] R. Ranjan. Epidemiology of violence against medical practitioners in a developing country (2006–2017). *Journal of Health Research and Reviews*, 4(3):123–130, 2017.

[ScrapeTube Documentation, 2024] ScrapeTube Documentation. A python library for extracting youtube video metadata, 2024.

[The Lancet, 2024] The Lancet. The rg kar tragedy: Violence against healthcare professionals in india. *Telegraph India*, 2024.

[VADER Sentiment Analysis Tool, 2024] VADER Sentiment Analysis Tool. An open-source tool for sentiment analysis, 2024.

[YouTube Data API Documentation, 2024] YouTube Data API Documentation. Official api for retrieving youtube data, 2024.