JOURNAL OF COMMUNICATION, LANGUAGE AND CULTURE

AI and Human Interpretation of Multilingual Feminist Discourse: A Study of Aurat March Comments

Esha Shakoor[1*]

[1]Department of English, COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan

*Corresponding author: ech76507@gmail.com; ORCID iD: 0009-0005-7837-8454

ABSTRACT

Artificial intelligence is widely used to examine social media discourse; however, its ability to understand complex feminist discourse across various languages has been largely overlooked in the existing scholarship. Utilising AI tools, this study analyses social media discourse on Aurat March which is the most prominent feminist activist moment in Pakistan, to show how an AI-assisted discussion of different cultural frameworks, particularly in the Global South environment, can be useful. This study examines how artificial intelligence (AI) systems interpret multilingual feminist online content. It compares the AI results with human analysis. Human analysis used a framework to evaluate AI outputs against human-coded standards. This approach tracked the incorrect classification of user comments on the Aurat March in Pakistan. This research uses technofeminism and Relational Content Analysis (RCA) to study 450 comments on the Facebook, Instagram, and YouTube pages of the film. A multilingual approach was necessary because users responded to the Aurat March posts in English, Urdu, and Romanised Urdu. The research results clearly distinguish between human and AI interpretations. The AI tool (ChatGPT-5) successfully categorised comments into three different languages; however, it frequently oversimplified sentiment categories and failed to identify rhetorical devices such as sarcasm and humour, while its accuracy decreased when evaluating Roman Urdu and code-mixed content. AI tools can assess sentiment across multiple languages; however, their performance declines in multilingual contexts, especially when they need to analyse code-switching and specific cultural languages such as Roman Urdu. The research suggests that, despite AI's potential, it requires further refinement to effectively handle politically and culturally sensitive discourse in multilingual contexts.

Keywords: artificial intelligence, feminist discourse, multilingualism, digital spaces, social media

Received: 2 January 2026, **Accepted:** 19 May 2026, **Published:** 30 July 2026

doi: https://doi.org/10.33093/jclc.2026.6.2.27

Published by MMU Press. URL: https://journals.mmupress.com/jclc

1.0 Introduction

The massive growth of AI in digital communications research has changed the way research examines, interprets, and identifies patterns in large volumes of content generated by users on the Internet. For example, the use of AI-based tools, such as sentiment classifiers and content moderation systems, breaks down barriers to efficiently gathering and identifying data while maintaining quality control. However, a growing body of literature has identified limitations in these systems due to their lack of understanding of the cultural, ideological, and contextual factors that influence how users interpret meaning related to social interaction (Bender et al., 2021; Hovy & Spruit, 2016).

A distinct example of this disconnect is evident in how various user communities interact with one another on the various forms of social media, where there is frequent usage of hybrid linguistic expressions, emotional expression, satire, and culturally grounded expressions. The use of AI systems in digital spaces where gender and rights debates are taking place can be seen as a limitation of this study. This questions the reliability of automated interpretation in situations that require a good understanding of culture and political context.

Since 2018, Pakistani cities have hosted the Aurat March which occurs every year on International Women's Day to protest gender-based violence, economic inequality, and women's social rights. The march evolved from its original form as a protest into a complete social movement which works to create gender equality while challenging all patriarchal systems which exist in Pakistani society (Kelso, 2024). Public discussions about the march are not only held in physical demonstrations but also on digital platforms, where supporters and critics debate the movement's demands, slogans, and political implications (Hassan et al., 2023). The discussions create two main responses which show the social conflicts which exist between feminism, religion, morality, and Pakistani cultural values (Hassan et al., 2023). People use Facebook, Instagram, and YouTube as their main platforms to share their views about the movement while communicating in multiple languages, including Urdu, English, and Roman Urdu. The widespread use of Roman Urdu in online communication further reflects the multilingual nature of digital discourse in Pakistan and presents additional challenges for automated language analyses (Khan et al., 2020).

People in South Asia use multilingual communication methods, including code switching and code mixing, for their daily lives and Internet interactions. In Pakistani digital discourse, people use English and Urdu for switching in their social media posts because they want to achieve specific communicative outcomes, not because they speak English and Urdu as their native languages (Jamali et al., 2022; Shah et al., 2025). Users who use multiple languages to express their emotions, social ties, and political beliefs through their multilingual practices demonstrate their cultural and social identity development. The online environment of South Asia uses Urdish which combines Roman script Urdu with English to create hybrid language forms, demonstrating how people use multiple languages to build their online identity. According to Liaqat et al. (2022), Urdu and Roman Urdu sentiment analysis remains comparatively underdeveloped due to a scarcity of large corpora, language parsers, and pre-trained models relative to high-resource languages such as English. Consistent with this resource gap, the present study finds that AI systems frequently turn Roman Urdu expressions into literal or decontextualized sentiment categories, missing communicative nuance that human coders readily identify. This linguistic diversity creates a unique interpretative environment that requires awareness and respect for the various cultural norms of different communities.

Aurat March is Pakistan's largest and most visible feminist movement. This led to a national conversation on topics such as gender rights, human autonomy, labour, domestic roles, and political participation. Social media discourse related to the Aurat March is highly polarised, with emotionally charged reactions related to the movement. Supporters of the Aurat March often used humour, satire, and intertextuality referencing international feminist movements, while opponents of the Aurat March relied on religious reasoning, cultural reasoning, and moral critiques to challenge the movement's feminist slogans (Zia 2020). The complexity of the Aurat March debate stems from both the arguments and the language used. For instance, people often switch between English and Urdu languages. They use Roman Urdu to show informality or emphasise points. Rhetoric, such as sarcasm and exaggeration,

is also common. These characteristics make texts particularly challenging for AI systems that lack the cultural context, interpretive history, or pragmatic cues.

In contrast, human interpreters depend on cultural knowledge, personal experience, and contextual cues to appreciate how meaning is constructed in online spaces. According to Ahmad Ghani and Rahmat (2023), confirmatory biases, ideologically driven statement alignment, and culturally created expectations are examples of how humans interpret the meanings of an online message. Humans can derive the meanings of messages not only from what has been written (the literal words) but also through tone, symbolism, humour, and socio-political socialisation (socio-political). This abundance of interpretative meanings helps humans discover the embedded meanings of feminine discourse. For example, humans critique the support for violence, recognise the irony in gun control, and feel emotionally charged connections in protests against guns. All these aspects relate back to the cultural context and experience.

Social media platforms have become central spaces where users produce and exchange content that effects public perceptions and interactions. Previous researchers have highlighted that digital communication is not separate from a person's identity or culture. It is influenced by many factors. These include how someone classifies their culture and their ability to understand and respond to another person's identity and culture. Balakrishnan (2022) highlights that intercultural communication styles arise from how a culture views its own values and norms. In addition, (Tan Jia Xin & Md. Noor, 2025) also provide evidence that people interpret messages sent via a digital channel differently based on the way they perceive themselves in relation to their own culture (socio-cultural orientation) and their level of skepticism. Therefore, these studies show that the process whereby people create meaning in their own minds is influenced by social context and therefore using AI tools to analyse political discourse and/or discourse with cultural meaning is not practical. Without cultural context and understanding of the intended meaning of a particular message or comment, the classification of AI tools runs at the risk of oversimplification, misinterpretation, or ideological misrepresentation. These same challenges to AI systems exist when examining the use of language technology more broadly. In recent studies Aremu (2024) shows that AI systems have difficulty processing or interpreting low-resource languages due to a lack of adequate training data and a lack of an adequate contextual model of the language. This is especially true for the Roman Urdu language, which has no standardised corpus and, therefore, has a highly variable spelling system. AI systems are inconsistent when processing Roman Urdu and frequently misclassify emotional comments as neutral or misclassify humorous/sarcastic comments as negative in sentiment. These limitations affect activist discussions because when automated systems suppress, misinterpret, or misidentify feminist voices, they undermine the visibility and acceptance of these ideas.

Aurat March's study provides an insightful framework for analysing these aspects, as it is remarkably rich in cultural cues and multilingual expressions that AI fails to capture. How and why, AI misinterprets this type of content requires a combination of insights from computational linguistics, digital feminism, and Techno feminist theory. Techno feminism highlights that AI, as a technology, reflects the social structures and biases present in the data on which it is trained Wajcman (2004), D'Ignazio and Klein (2020). AI systems are not neutral classifiers but are socio technical systems trained on inequalities embedded in data, gendered traditions and dominant linguistics norms. So, AI's misinterpretations of feminist discourse are not a technical issue, but it is linked to the gap between feminist understanding and algorithmic interpretation. There is a significant gap in AI research when it comes to analysing feminist discourse in languages outside a specific cultural context. A comparative study of AI systems and human researchers is needed to understand how AI works across different linguistic and cultural contexts, particularly in feminist discourse. Most existing studies in NLP and sentiment analysis have focused on high-resource languages and monolingual datasets, while code-mixed and low-resource language processing remains a challenge which has not been solved according to Nazir et al. (2026) and Srivastava and Singh (2021). Current AI systems struggle to understand complex social and cultural discussions. They lack suitable tools to analyze low-resource and code-mixed languages. These languages often face challenges due to limited data and differences in language structure and script. Generative AI models primarily train on English and high-resource languages according to research which shows this training approach limits their capability to understand languages

that are underrepresented in developing countries according to Kshetri (2024). Researchers need to conduct systematic studies which investigate how multilingualism with its code-mixing and Roman Urdu script differences affects AI systems when they analyze feminist commentary.

This paper fills these gaps by comparing the AI and human interpretations of Instagram, Facebook, and YouTube comments in response to the Aurat March in English, Urdu, and Roman Urdu. Drawing on relational content analysis within a Techno feminist framing, this research investigates how meaning is constructed, flattened, or misrepresented by AI as compared to human coders. This method not only defines the boundaries of automated interpretation but also sheds light on the socio-cultural knowledge used to comprehend feminist discourse in the digital public sphere of Pakistan. These interpretive gaps are important to keep in mind for developing more culturally sensitive artificial intelligence systems, and to avoid misclassifying feminist voices from the Global South in algorithmic environments.

1.1 Research Questions

1. What differences exist between AI and human coders in interpreting multilingual social media discourse related to the Aurat March in Pakistan?
2. What linguistic, cultural, and rhetorical factors contribute to the interpretive gaps between AI and human analysis of Aurat March related feminist discourse?

2.0 Literature Review

2.1 AI Interpretation of Digital Communication

Studies of AI interpretation have found that standard computer models have difficulty with discourses that rely on socio-cultural understanding. Artificial intelligence systems are only aware of probabilistic links between linguistic tokens and do not work with the meanings of the situated contexts, culture and ideology (Hovy & Spruit, 2016). At the core of this approach, these systems are likely to reduce human communication to lexical patterns that constrain the detection of layered meanings, emotional subtexts, or ideological stances (Bender et al., 2021). This limitation is more evident in social media where communication routines shape textual, visual and cultural cues and users often resort to mixed linguistic expressions to express emotions, identities and criticism.

This issue is especially pronounced in online contexts, where discourse is influenced by bias, affect, and social identity. Ahmad Ghani and Rahmat (2023) found that personal opinions on social media are shaped by factors such as algorithmic exposure, conformity, and self-control, highlighting that how people engage with online content is filtered through subjective, pre-existing biases rather relying on their ideological frames, personal beliefs and cultural cues. Implicit interpretive processes are not accessible to artificial intelligence systems that are trained on generalised corpora. This helps to verify the argument that meaning on sites like Instagram, Facebook and YouTube cannot be derived by relying on a cursory analysis of texts. Therefore, the literal reading of AI does not allow him to see any depth created while sharing cultural knowledge, sarcasm, intertextual references, and rhetorical play.

Moreover, Tan Jia Xin and Md. Noor (2025) showed that the interpretation of participants was different based on their social context, cynicism, and experience with different types of digital content. This research shows that the interpretation of user-generated or influencer-generated communication is influenced by emotional and cultural factors. This further emphasises the larger issue of how AI lacks the ability to grasp the contextual intricacies essential for interpreting meanings in digital feminist activism. AI, when confronted with emotionally charged content such as the discourse of Aurat March, provides flattened or incomplete readings because it does not have access to the socio-cultural frameworks that inform human meaning-making. Milios and BehnamGhader (2022) found that methods for detecting bias in language models are highly language-dependent, showing that identifying meaning accurately requires cultural insight into how it is expressed differently across languages — often through coded language, indirect phrasing, and other culturally specific rhetorical devices. This supports the present study's findings, which show that AI's literal interpretation approach consistently failed to register meaning embedded in sarcasm, irony, and culturally coded moral or religious critique.

The findings of these studies show that AI is limited when dealing with discourses driven by ideology, culture, and identity, specifically the conditions inherent to feminist digital communication.

2.2 Algorithmic Bias and Misinterpretation of Identity-Related Discourse

AI systems inherit structural biases in training data, and the lack of critical interpretation and consideration leads to common distortions in discourse related to identity (Birhane, 2021). This bias reflects the unequal distribution of data and linguistic ideologies which leads to the absence of contextual markers that differentiate between criticism, humour, or resistance. Nadeem et al. (2021) showed that AI systems often rely heavily on the labels assigned to their training data to determine the attributes associated with a statement. Therefore, identical statements that used gendered language can be assigned different labels depending on the gendered language. AI systems trained on this data mislabel gendered comments negatively and may accurately assign positive labels to gendered neutral comments. Sap et al. (2019) also showed that classifiers for hate speech disproportionately mislabel toxicity for speakers of minority languages. This finding suggests that AI systems lack the cultural grounding necessary to classify hate speech accurately.

Yi and Al-Saggaf (2025) shows that analysis of immigration satire that satirical discourse greatly relies on shared cultural knowledge, inter textual cues, and recognition of the speaker's social positioning. Interpreters who are unfamiliar with such ideas will read satire literally and risk its actual intent. Therefore, this idea directly supports the current study's argument that AI's inability of AI to recognise satire and rhetorical resistance results from its lack of cultural understanding. AI does not simply misunderstand feminist discourse; it also misinterprets why these ideas are expressed in the way they are, which weakens the social and political intent behind feminist discourse.

2.3 Multilingual Digital Discourse and Code-Mixing

Multilingual communication is considered a great feature of digital communication in South Asia, enabling users to communicate in languages such as English, Urdu, and Roman Urdu. Every language has its own socio-cultural identifier; for instance, a globalised or cosmopolitan self, or identification is English, while Urdu and Roman Urdu provide culturally determined representations of emotion and humour, allowing individuals to communicate their culture. Roman Urdu presents interesting problems because of its unstandardised orthography and variable phonetic spellings, such as "acha", "achaa", and "accha", which have the same meaning but are visually different. Computational models that have been trained mostly on standard language corpora face the hardest time in identifying these distinctions because the meaning is based on the socio-linguistic context instead of formal grammar.

Researchers have provided helpful insights into this situation. Balakrishnan (2022) argues that cultural dimensions are the most powerful predictors of communication style, and therefore directly influence the manner in which people express acknowledgement, criticism or emotional positioning. This perspective supports the present research assumption that Roman Urdu and code-mixed expressions are cultural references that cannot be understood precisely without cultural knowledge. A recent study by Aremu (2024) shows that AI has difficulty with endangered languages that have limited resources. These findings strengthen the argument that conventional artificial intelligence systems face challenges in dealing with Roman Urdu due to language variation, cultural context, and language use.

Besides the cultural consequences of AI development for interaction with under-resourced languages, the authors also point out the intricacies of code-mixing. Mixed language gives rise to many additional layers of meaning that are beyond AI's capability to interpret. For example, the use of English in South Asian countries may imply power, whereas Urdu connotes strong feelings. AI perceives these texts as fragmented, whereas humans recognise code mixing as a rhetorical device used for a certain purpose. The gap between humans and AI demonstrates the need for advanced analysis of multilingual and culturally rich discourses.

2.4 Digital Feminist Activism and Aurat March Discourse

Digital feminism is defined by emotional expression, symbolic language and creative communication (Baer, 2016; Mendes et al., 2019). Online digital feminist activism often employs humour, irony, visual

metaphors, and provocative phrases to resist oppression. The Aurat March in Pakistan exemplifies this online feminist discourse. Reactions to the event were marked by sarcasm, criticism of religion, calls for moral regulation and debates over cultural values (Zia, 2020). The reactions to the Aurat March are a reflection of the power relations and the dominant narratives about women and society. To understand this discourse, one needs to be familiar with its cultural and linguistic setting. For instance, phrases such as 'Islam ne izzat di hai aurat ko' (Islam has given honour to women) and 'Mera jism meri marzi' (What I do with my body is my choice) contain narratives about morality, religious identity and ownership. Human interpreters know the context of these statements in history. AI might just label these statements as positive or negative.

Previous studies on the discourse of Aurat March reveal that the digital audiences interpreted the responses through a blend of cognitive and ideological lenses thus affecting the process of meaning-making and meaning-making itself. Ahmad Ghani and Rahmat (2023) show that this is how digital audiences interpret feminist discourse in its cultural and political dimensions. Tan Jia Xin and Md. Noor (2025) support this perspective by noting that emotions, scepticism, and social positioning shape how meanings are made in digital space; therefore, AI cannot understand feminist messaging in a culturally coherent manner. Rhetorical strategies such as exaggeration and intertextual references to global movements, as well as humour that critiques patriarchy, are present in the comment sections related to the Aurat March and feminist discourse. These strategies also present an additional challenge to AI systems, which tend to interpret such comments literally.

Although research on algorithmic bias and natural language processing has increased, most studies have focused on English-dominant settings or general algorithms for bias recognition in language models. Blodgett et al. (2020) critical survey of bias in NLP systems also emphasises that language technologies can encode the hierarchies of a society and reproduce them rather than represent linguistic patterns. Puttick et al. (2025) conducted a systematic review of bias detection strategies in non-English word embeddings and language models, with limited research and practical methods focusing on under-represented languages beyond English. The empirical data also suggest that such social biases exist in multilingual contextual models, as Miliots and Behnam Ghader (2022) observe that variations of BERT still display biased behaviour across multiple languages. Studies highlighting the gender effects of AI also highlight its negative effects. Srivastava and Singh (2021) show trends in gender bias in machine learning, while Wyer and Black (2025) found that GPT-3 models can generate sexualized and gendered content about women even without being prompted to, suggesting that the same model family may carry embedded gendered assumptions that could plausibly extend to how it interprets gender-related discourse, not only how it generates it.

Simultaneously, previous studies on digital feminism highlight the complexities of online activism. Baer (2016) examines feminist media practices by focusing on emotions and creativity as key strategies, whereas Mendes et al. (2018) focus on the communication principles and participatory methods of feminist digital activism. Recent studies have also highlighted the issue of rhetorical techniques and identified positions in social media. Bharti et al. (2020) highlighted that even advanced deep learning methods struggle to detect English and English-Hindi code-mixed posts. They highlighted that linguistic creativity, irony, and the use of hybrid languages can significantly reduce model accuracy, especially in multilingual settings. Similarly, Mohammad (2020) demonstrated that automated models exhibit a high rate of divergence from human coders in auto-tagging stance and sentiment in political discourse, requiring advanced contextual and ideological knowledge, which illustrates that AI can easily be outsmarted. Collectively, these studies make it clear that current AI solutions cannot identify subtle interpretations, particularly in the analysis of multilingual feminist discourse online. Nevertheless, the comparison of human and AI interpretations of multilingual feminist discourse has not been researched, especially in contexts such as Pakistan and the Global South. Although previous researchers have examined the concept of algorithmic misclassification in English or bias in multilingual NLP models, no one has historically examined the effects of AI on culturally based meanings, code-mixed text, and rhetorical strategies in feminist discourse across languages. Similarly, the available literature on online feminist activism has not integrated computational analysis with human coding to assess the interpretive accuracy. This wide gap indicates the need for comparative research on both AI and human

interpretation, where understanding of digital discourse is grounded in cultural norms and linguistic varieties.

2.5 Gaps in Existing Research

The study examined feminist discourse on social media, use of artificial intelligence in content analysis and the reasons for multilingual communication on social media. It also included general discussion of the Aurat March and feminist activist representation on the Internet. Also, there are many studies which have investigated the effectiveness of AI to sentiment analysis and discourse classification using large-scale social media datasets. But the gaps in the literature are still immense. There are limited studies that examine AI systems for interpreting multilingual feminist texts, especially in areas where Urdu and Roman Urdu are common. In addition, limited work has been done on the comparative analysis of human and AI interpretations of feminist discourse, which limits the knowledge of AI in identifying complex meanings. Furthermore, most of the current literature focuses on the West or monolinguals, leaving the Global South underrepresented in AI and digital discourse research. This study fills these gaps by using Relational Content Analysis and AI-based sentiment and rhetorical classification to contrast the interpretation of multilingual comments on the Aurat March by humans and AI on Instagram, Facebook, and YouTube. This study provides new insights into how AI struggles to understand feminist speech that is deeply culturally rooted. It examines remarks in English, Urdu, and Romanised Urdu. This challenges the demand for more inclusive AI that understands language and culture to support digital activism.

2.6 Technofeminism as the Guiding Framework

According to techno-feminism, the technology we create reflects the social, cultural, and political context in which it was created. In the view of techno-feminism, Wajcman (2004) and D'Ignazio and Klein (2020) argue that artificial intelligence (AI) cannot be viewed as a neutral tool, but rather as an investment of society and is therefore susceptible to reflecting social power structures and social value systems. This means that the errors made by AI when interpreting feminist discourse are not technical failures but the results of inconsistencies embedded in the underlying data, algorithms, and linguistic concepts. Balakrishnan (2022) and Ahmad Ghani and Rahmat (2023) support the technofeminism argument and highlight that communication and practice are shaped by ideologies and cultural norms. The inability of artificial intelligence systems to identify cultural information, feminist language, or public expressions reveals the limitations of using technology to classify all types of communication using only Western norms and paradigms through the lens of Western worldview. From a techno feminist perspective, these limitations help explain why conventional artificial intelligence systems tend to misunderstand or misclassify the discourse around the Aurat March. Such systems lack the cultural and historical knowledge of the feminist movement and the ideological context important for correctly interpreting feminist discourse in South Asia. There are large gaps in the current literature on certain key themes and issues. First, AI's understanding of multilingual feminist discourse remains under-researched, especially in countries within the Global South. Additionally, existing studies have not conducted a systematic comparative analysis between AI-based and human interpretations of Aurat March commentary, despite its cultural and political significance. Third, despite being essential features of digital discourse in Pakistan, research on Roman Urdu and code-mixing is limited. Finally, there has been little empirical application of technofeminism to examine AI-based misinterpretations of South Asian feminist activism. This study addresses these gaps by examining how both AI systems and human coders interpret multilingual Aurat March commentary and situating the differences in interpretation within a techno-feminist framework.

3.0 Methods

This study employs mixed methods to explore how AI and human coders understand multilingual feminist discourse about the Aurat March. The research framework uses relational content analysis (RCA) and insights from technofeminism. RCA examines how textual data create conceptual meanings and establish relationships between different concepts. This research method enables the researcher to study emotional expressions alongside ideological positions and cultural references in social media comments. The techno-feminism framework was formally proposed by Judy Wajcman in 2004,

enabling researchers to study how social and cultural factors, including gender power dynamics, shape the development of AI technological systems. The methodology conducts a systematic assessment of how humans and artificial intelligence understand information while studying how different cultural and linguistic elements shape meaning. The structured research method establishes how human coders and AI systems share a common understanding while also showing their different approaches to interpreting multilingual feminist discourses across different cultural contexts. The structured research method establishes how human coders and AI systems share a common understanding while also showing their different approaches to interpreting multilingual feminist discourses across different cultural contexts.

3.1 Data Collection

Across the three platforms (YouTube, Facebook, and Instagram), 450 publicly available comments in response to Aurat March related posts were collected between August and September 2025, drawn from authentic Aurat March source posts (150 per platform) spanning the 2023, 2024, and 2025 Aurat March events. Posts were identified through the hashtags #AuratMarch, #WomenRights, #AuratMarch2023, #AuratMarch2024, #AuratMarch2025, and #FeminismPakistan, and were retained only where they originated from public accounts with openly accessible comment. Comment selection was purposive rather than random: for each qualifying post, a single comment was retained if it met three criteria, it responded directly to Aurat March-related content or slogans; it was not composed solely of emojis, hyperlinks, or repeated spam text; and it was not a duplicate of an already-selected comment. Only top-level comments were included, but the thread context was referred to during coding as necessary to clarify meaning; no more than one comment per user was retained so as to avoid over-representation of individual voices. Language classification was executed by manual review by the researcher. Code-mixed comments were not excluded but assigned according to the dominant language. The final data set consisted of 450 comments including each in English, Urdu and Roman Urdu, the three most dominant languages of online communication in Pakistan with Roman Urdu being especially popular in informal digital interactions. The researchers stored all data in an anonymous format, which blocked users from being identified through their personal information or usernames. The method was created to maintain ethical digital research standards while researchers built a dataset to study AI and human coders' information processing variations.

3.2 Rationale for Multilingual Dataset

The multilingual nature of Pakistani social media influenced the choice of the dataset. English was selected as the language because it serves as a global feminist discourse medium and is commonly used by urban users. Urdu served as a medium for expressing national and religious identity. Roman Urdu, which lacks established writing rules, shows an emotional informal style that users employ for their personal communication. The spelling variations, hybrid forms, and phonetic representations of Roman Urdu create difficulties for computing systems to interpret the language, but they should be used to assess AI performance across different language situations.

3.3 Human Coding Procedures

The process of human coding used three stages which followed the rules of the Relational Content Analysis (RCA) method. The coder used open coding to study all dataset content while searching for common phrases which showed emotional, ideological, and rhetorical patterns about the Aurat March. The coder applied axial coding to arrange the discovered codes into primary thematic groups, as shown in Table 1. The categories included empowerment, backlash, humour, satire, religious justification, moral regulation, and gender critique. The study established clear category definitions which included specific quotes to ensure both consistency and transparency during the research process. The relational coding method examined how different themes were connected to each other through comments because people used multiple rhetorical methods to express their ideas through humourous and critical content, which included religious and ethical reasoning. The researchers evaluated the coding procedure through intercoder reliability testing which involved two coders who separately interpreted the same dataset. For a stratified subsample of 100 comments, the two coders coded independently, with no discussion beforehand; remaining differences across the full dataset were then resolved through

discussion between the coders to reach consensus. The study used Scott's Pi (Scott, 1955) to assess intercoder agreement on this independently coded subsample, yielding Po = 96.55% and $\pi = 0.65$, indicating substantial agreement. The coding process cannot function as a benchmarking tool for NLP models. Because human coding reflects an interpretive judgment rather than a definitive ground truth, the comparison between human and AI coding reported in Section 3.3.2 is presented as a measure of agreement and divergence between interpretive standpoints, not as an assessment of AI accuracy against a fixed benchmark. This study uses an interpretive sociotechnical framework to demonstrate how AI systems fail to comprehend complex aspects such as sarcasm and cultural references, together with moral framing.

3.3.1 Coders

The author, who is a graduate student in English Linguistics, handled most of the coding work while receiving feedback from the peers. The coder completed all initial coding stages; open, axial, and relational based on her understanding of feminist theory. She developed the discourse surrounding Aurat March, and the sociocultural context of Pakistan through her studies and independent research. Elements such as sarcasm, humour and rhetorical features were classified according to contextual interpretation including tone, contradictions, exaggerations and culturally specific cues present in the comments. Peer-level secondary coders reviewed the coding decisions and discussed any discrepancies and helped clarify any uncertainties. Due to lack of any formal training in RCA, the coder had to rely on her expertise and knowledge of the data set to accomplish this process. To assess reliability, the primary coder and one peer coder independently coded a stratified subsample of 100 comments selected to achieve proper representation across all three languages and platforms with no discussion between them prior to coding. Intercoder agreement on these independent codes was then formally calculated using Scott's Pi (Scott, 1955), a statistical measure suitable for nominal data and two coders; discrepancies were resolved through discussion afterward to inform consensus coding of the full dataset. The researchers calculated Percent Agreement (Po) and Expected Agreement (Pe) and used the formula $\pi = (Po - Pe) / (1 - Pe)$ to compute Scott's Pi. The results demonstrated a 96.55% agreement between coders, while Scott's Pi coefficient reached 0.65, indicating substantial coder agreement. The full calculation of Scott’s Pi is provided in Appendix B.

3.3.2 AI–Human Comparison Framework

Across the dataset (N = 450), human and AI sentiment coding diverged on 343 comments (76.2%) and agreed on 107 comments (23.8%). By platform, divergence occurred for 116 of 150 Facebook comments (77.3%), 108 of 150 Instagram comments (72.0%), and 119 of 150 YouTube comments (79.3%). These figures totalled 343 comments, consistent with the overall result. By language, divergence occurred for 61 of 84 English comments (72.6%), 209 of 280 code-mixed English/Roman Urdu comments (74.6%), and 73 of 86 Urdu-script comments (84.9%). These figures also totalled 343 comments, confirming that both groupings represented the same overall divergence pattern. For instance: the Facebook comment *"Aurat March ka asal maqsad samajhna chahiye, lekin har saal isay sirf controversy bana diya jata hai. Auraton ke huqooq ki baat karna ghalat nahi, magar apni values aur culture ko bhi nazar mein rakhna chahiye"* was human-coded Neutral, reflecting balanced criticism paired with support for women's rights, but classified Negative by AI likely caused by lexical cues such as "controversy" and the values/culture framing, without captured the comment's original tone.

Table 1

Thematic Categories, Definitions, and Example Comments

Theme / Category	Definition	Example Comment
Empowerment	Expressions highlighting women's rights, agency, or self-determination	"We need more women in leadership roles; their voices matter."
Backlash / Critique	Comments opposing the Aurat March or feminist messages	"یہ ٹھکرائی ہوئی گندگی ہے زمین کی" (This I rejected filth of the earth.)
Humor / Satire	Comments using jokes, irony, or sarcastic tones	"The irony is that Muslim feminists ask for Hijab rights while others want freedom from it."
Religious Justification	Comments referencing Islamic teachings or moral guidance	"Kis Islam main Likha h nagi ho kr ghomo" *(*In which Islam is it written to roam around naked?)
Moral Regulation	Comments focusing on societal norms, decency, or propriety	"Mera jism meri marzi ka matlab ye nhi ke aurton Allah ke hukm ke khilaaf bol rhi hain." (My body, my choice' does not mean that women are speaking against Allah's commands.)
Gendered Critique	Comments emphasizing gender roles or expectations	"all elites in march and masses are in home"

3.4 AI Classification

The AI system generated textual results through its use of ChatGPT-5, which functions as an extensive language model developed by OpenAI. The researchers used ChatGPT-5 (OpenAI) desktop application as an artificial intelligence language model to analyse the comments application during September 2025. The model was accessed via a standard interface with a uniform English prompt for the whole data set (see Appendix A). All comments received the same instructions to generate comparable and reproducible AI results. The application does not expose adjustable parameters (temperature, top-p); default settings were used throughout. Each comment was analysed once; multiple runs were not conducted, and this is disclosed as a limitation rather than a claim of guaranteed reproducibility. AI-generated responses were manually copied by the researcher from the ChatGPT interface into a structured Excel spreadsheet built for this purpose, which was then used for the AI–human comparison reported in Section 3.3.2.

The study was aimed at evaluating the AI's multilingual classification abilities by using comments in languages other than English, provided in their original form as Urdu and Roman Urdu without any translation. The team uses ChatGPT-5 for its natural language processing abilities, enabling it to work with multiple languages and understand contextual information. It would be used to directly compare the artificial intelligence generated interpretation to the encoded classification of humans from a single date set. To ensure comparability of AI results, a consistent prompt was used in English, Urdu and Roman Urdu. The prompts trained the model to identify sentiment, stance, emotion, and rhetorical features such as sarcasm and humour, which enabled the researchers to assess the performance of the AI. No post processing or manual corrections were applied to keep the consistency. AI results were also recorded as they were produced for a reliable comparison between computational and human

interpretations. This process revealed systematic misclassifications and interpretive gaps across different linguistic forms. This approach set AI classification consistency and allowed researchers to evaluate the accuracy of human coded interpretations which were studied in a controlled environment.

3.5 Comparative Analytical Procedure

The human coders primarily conducted a comparison to identify the similarities and differences between the AI-generated sentiment classifications and the human-coded sentiment classifications. The research is based on the places where the AI system failed to recognise gender-related discrimination, sarcastic comments, religious and moral references, and expressions specific to cultures. The researchers looked at English, Urdu and Roman Urdu to see if artificial intelligence systems consistently classified language in these languages. The researchers analysed content from Instagram, Facebook and YouTube platforms to see if different language patterns affected AI system performance. The research studied the misclassification patterns in various categories like sarcasm, Roman Urdu interpretation, religious context and culturally embedded meanings. The study looked at how artificial intelligence systems dealt with information and the results were different than humans using different languages and content assessment online platforms. The researchers defined AI misclassification as, where AI systems produced incorrect sentiment and rhetorical analysis results which differed from human evaluation results.

3.6 Ethical Considerations

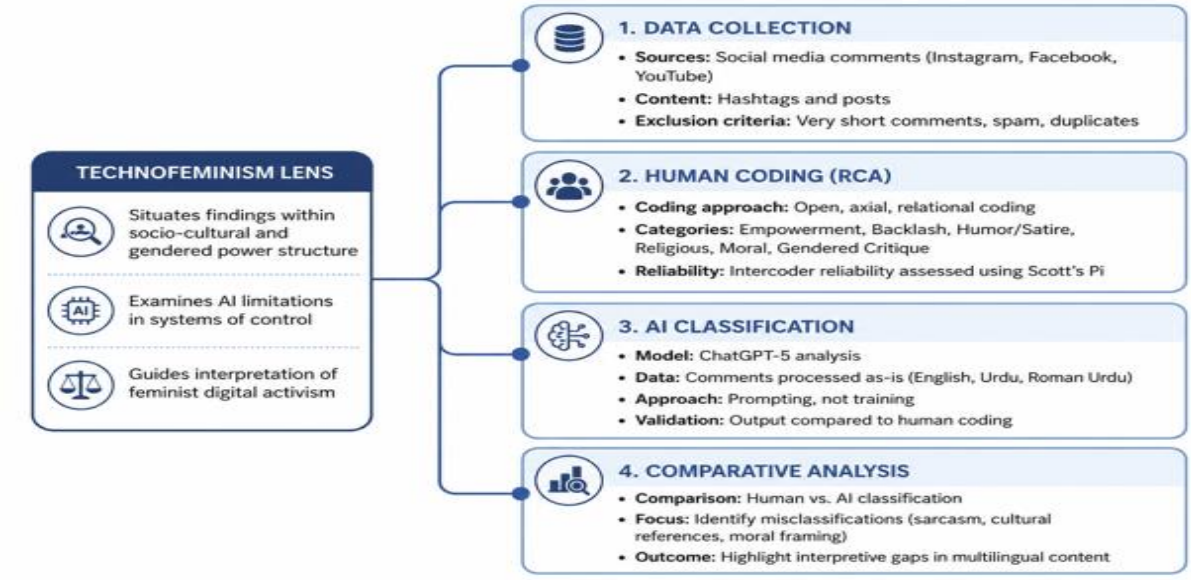
The comments included in the analysis are publicly available. The analysis is conducted on comments which do not contain personal and sensitive information. The analysis is done by following the ethical standards of digital discourse research. It is ensured that there was no interaction between commenters and the research. The researchers did not attempt to identify or follow individual commenters on the internet.

3.7 Methodological Alignment

The research method integrates data collection, human coding, AI classification, and comparative analysis into a structured analytical framework. Relational Content Analysis identifies thematic and relational patterns. The study results need to be understood through Techno-feminism which connects to existing cultural and gendered social power systems. The comparison between human and AI interpretation highlights interpretive gaps. The combined method shows how human intelligence and artificial intelligence systems understand multilingual feminist discourse in different ways. Figure 1 summarises the methodological alignment of the study.

Figure 1

Methodological Alignment



4.0 Results

Analysis showed a clear difference between the AI and human responses. The artificial intelligence system evaluated comments in three different languages, but the results did not correspond to human understanding. The comments on the Aurat March showed this pattern most strongly, since those who wrote them had either cultural or rhetorical knowledge of the event. The three main findings showed that human detect sentiment and position categories, while the model struggled to identify rhetorical elements, and its accuracy dropped when assessing Roman Urdu and code-mixed material. The dataset provides examples which demonstrate the gaps in interpretive understanding. The two coders achieved strong agreement through their inter-coder reliability assessment which resulted in a 96.55% Percent Agreement and 0.65 Scott's Pi (Scott 1955).

4.1 Divergence in Sentiment

Total of 450 comments analysed, AI sentiment labels matched human coding in only 107 instances (23.8%), diverging on the remaining 343 (76.2%). This pattern held broadly across all three platforms, though YouTube showed the highest rate of divergence at 119/150 (79.3%), followed by Facebook at 116/150 (77.3%) and Instagram somewhat lower at 108/150 (72.0%). For example: the Facebook comment *"Aurat March ka asal maqsad samajhna chahiye, lekin har saal isay sirf controversy bana diya jata hai. Auraton ke huqooq ki baat karna ghalat nahi, magar apni values aur culture ko bhi nazar mein rakhna chahiye"* was coded Neutral by the human, who saw it as a balanced critique of the way the march is framed annually while still supporting women's rights, but AI labelled the same comment Negative, apparently picking up on isolated words like "controversy" and "culture" rather than the overall measured tone of the sentence. Table 2 shows that divergence was lowest for English comments (72.6%), followed by code-mixed English/Roman Urdu comments (74.6%), and highest for Urdu-script comments (84.9%). Thus, agreement between AI and human coding was lower for less standardised and more culturally embedded content.

Table 2

AI vs Human Interpretation Divergence by Language (N=450)

Language	Total Comments	Misclassified by AI	AI Divergence (%)
English	84	61	72.6%
Code-mixed English/Roman Urdu	280	209	74.6%
Urdu-script	86	73	84.9%
Total	450	343	76.2%

Note. Divergence was calculated as the number of comments where AI sentiment labelling did not agree with human sentiment coding, over total comments in that language group.

4.2 Challenges in Detecting Rhetorical Nuance

AI face difficulty to recognize rhetorical devices, including sarcasm, irony, and moral or religious reasoning, often interpreting comments literally where human coders understood the underlying meaning. The YouTube comment *"The irony is that Muslim feminists around the world are asking for rights to wear Hijab while women over here are wanting the right to walk around without Hijab"* was labelled by human coders as ironic criticism but AI labelled it as Neutral, completely missing the irony. Similarly, the Instagram comment *"Mast joke mara rey"* a short, dismissive Roman Urdu was human coded as mocking ridicule but interpreted by AI as Neutral. Religious and moral reasoning followed a similar pattern – rather than interpreting comments referring to Islamic values or societal morality as negative, AI most often flattened them to Neutral, missing the underlying ideological or cultural argument human coders identified.

The results indicate that AI face difficulty interpreting rhetorical and socially constructed meaning in online feminist discourse, defaulting to a neutral interpretation when meaning is based on inference rather than explicit wording. The AI system encountered its highest difficulties when analysing Roman Urdu and code-mixed comments which contained non-standard English and Urdu words. The data in Table 2 show that the AI systems made the most incorrect identifications when processing Roman Urdu, followed by Urdu and English. Human coders identified the comment *"Aurat March bilkul sahi hai, warna humesha chup rehna parta hai"* (Aurat March is absolutely right; otherwise, women always have to remain silent) as a strong support for women's rights.

However, AI systems classified it as neutral because the Roman Urdu spelling blocked sentiment detection. Human coders assessed the comment *"Feminism is fine lekin extreme slogans se log offend ho jate hain"* (Feminism is fine, but extreme slogans offend people) demonstrating an ambivalent or mixed stance. The text received a negative classification from AI systems because the systems could not detect the linguistic elements which reduced the severity of criticism while showing partial affirmation of the statement. The research results show that AI systems experience difficulties with informal speech patterns, code-switching, and non-standard spelling because these elements appear throughout digital feminist communication used by Pakistanis. The Roman Urdu language system shows a high rate of incorrect classifications which need improvement through the enhanced representation of mixed language forms in AI technology.

4.3 Platform-Specific Patterns

Divergence patterns varied somewhat by platform, though all three showed AI diverging from human coding on the large majority of comments: Facebook (77.3%), YouTube (79.3%), and Instagram (72.0%), the lowest of the three. On Instagram, short, colloquial comments were frequently misread for example, *"This march should be banned.."* was human-coded Negative but classified Positive by AI,

showing the divergence was not limited to under-detection of negativity but ran in both directions. On YouTube, sarcasm and direct derogation were both consistently missed: beyond the irony example above, the comment *"Dil ki gehraiyon say lanat wasool kren"* ("receive curses from the depths of my heart" — a direct insult) was human-coded as Abuse/Derogation but classified Neutral by AI. Facebook comments, often longer and more discursive, showed a related pattern in Section 4.1 hedged, multi-part arguments being reduced to a single negative-triggering lexical cue. Taken together, these patterns suggest AI's difficulty is not tied to any single platform's format, but recurs wherever rhetorical indirection, code-mixing, or argumentative hedging is present, regardless of comment length or platform convention.

5.0 Discussion

The theoretical explanation of these results shows that misclassification errors are caused by both technical limitations and broader socio-technical problems. The Technofeminist framework shows how conventional artificial intelligence systems are unable to account for cultural backgrounds, linguistic differences and gendered communication styles which leads to the misrepresentation of feminist voices in digital spaces (Wajcman, 2004).

5.1 RQ1: Human vs AI Interpretation

The research revealed major differences in how AI and humans read multilingual feminist discourse. The comments were coded by humans to identify emotional expressions, persuasive techniques and ideological beliefs. This was demonstrated with humour, sarcasm, moral reasoning and contingent support. Despite the commendable performance of GAI (Generative Artificial Intelligence). in select developed languages like French, Spanish, and Japanese, it struggles to deliver comparable results in low-resource languages (LRLs) such as Bengali, Hindi, and Swahili" (Kshetri, 2024, p. 95). Hovy and Spruit (2016) have shown that social knowledge and cultural understanding are required to understand language, but today's artificial intelligence systems are not able to do that. Artificial intelligence systems might be able to parse complex expressions to determine basic emotional states positive, negative, neutral but they would not understand the deeper cultural meaning of those states. The worst misclassification rates were found in Urdu and Roman Urdu, due to the lack of informal spelling, hybrid language usage, and South Asian languages in the training datasets. AI struggled with code-mixed comments not just because of the language itself, but because code-mixing carries social meaning that a literal reading misses. Shahzadi (2025) found that Pakistani ESL students use code-switching to express identity, showing that mixing languages is a meaningful choice, not random. Digital platforms generate new ways of communicating that influence how users communicate humour, sarcasm and critical remarks. The human coders changed their views on the matter after getting some context, but the artificial intelligence systems used the same rules across both platforms. This led to more errors in both the assessment of sentiment and the rhetorical evaluation. Artificial intelligence systems rely on statistical data and therefore cannot grasp cultural nuances and are unable to comprehend intricate discourses that demand a nuanced cultural comprehension.

5.2 RQ2: Technofeminist Perspective and Sociotechnical Context

Technofeminists highlights that AI misinterpretations are not only technical issues but also show wider social and technical inequalities. Wajcman (2004) states that technology mirrors social relations and power dynamics in society, especially those influenced by gender. Such inequalities are visible in the inability of AI to properly understand feminist discourse. Automated systems can sometimes inadvertently amplify some voices and marginalise others D'Ignazio and Klein (2020) argue that data-driven systems reinforce existing gender hierarchies instead of challenging them. One of the major reasons for AI misinterpretation is the multilingual and code-mixed nature of Pakistani social media. Roman Urdu, which is often not standardized in its spelling, is less well represented in AI training data, hence there is higher rates of misclassification. Joshi et al. (2020) described this as a "resource scarcity problem", where non-English languages are not often included in the main AI systems. In such context, the mixing of languages is not simply a technical issue but also a cultural practice, as meaning in feminist discussions often emerges through the interplay of several languages within the text. AI limitations are exposed in platform-specific communication patterns. Each social media site has a

different way of communication which influences the way people express emotions, humour and criticism. Human coders in a study noted these differences across Instagram, Facebook and YouTube.

5.3 *Limitations*

The research study contains multiple restrictions because it focuses on one artificial intelligence system. The analysis used a sample of 450 comments from three social media platforms, all from one country (Pakistan), which may limit generalisability. The qualitative coding process was affected by its subjective nature because the study used peer review and high intercoder reliability as validation methods.

5.4 *Implications*

The results have important implications. Digital feminist activism employs code-mixing and creative language techniques (Baer, 2016; Mendes et al., 2019). The misclassification of these strategies by AI demonstrates the need for an improved understanding of the cultural meanings that exist in multilingual environments. The findings support three areas of application: content moderation, political communication analysis, and digital rights advocacy, because activist voices are distorted through misinterpretation. The combination of human judgment with AI systems will result in a better understanding of context which will improve multilingual discourse analysis accuracy**.**

6.0 Conclusion

This study examined how humans and AI interpret multilingual feminist discussions of the Aurat March in Pakistan. The results show that the AI system can process a large number of texts, but it failed to interpret rhetorical techniques, cultural, and contextual elements that are important to online feminist discussions. Human coders effectively identify humour, sarcasm, moral reasoning, and ideological views, whereas AI simplifies this expression into positive, negative, and neutral sentiments. This study also highlights the importance of code mixing and multilingualism, especially Roman Urdu, in shaping digital feminist conversations. The high rates of misinterpretations for these languages highlight the underrepresentation of Global South languages in AI systems and reveal broader structural and sociotechnical challenges (Wajcman, 2004). The study uses Relational Content Analysis and a techno-feminism lens to examine digital activism, algorithmic bias, and multilingual online communication. These results have significant implications. AI developers should design systems that understand cultural meanings in multilingual contexts and digital platforms should take these limitations into account when moderating content and analysing activist discussions. Interpreting accuracy can be enhanced by combining human knowledge and AI tools, particularly for code-mixed and non-standard languages. Future research should apply these methods to other social movements, languages and platforms. This should include direct collaboration with activists. These efforts will help make sure that AI-powered discourse analysis supports the meanings conveyed in feminist digital activism. Thus, technology can better reflect the complexities of multilingual and culturally rooted online communication.

Acknowledgement

The author would like to express sincere appreciation to Dr. Javaria Farooqui from Comstats University Islamabad, Lahore Campus for valuable guidance.

Conflict of Interest

The author has declared that no competing interests exist.

Author Contribution Statement

ES: Conceptualization, Data Collection, Data Curation, Methodology, Formal Analysis, Validation, Writing – Original Draft Preparation, Writing – Review & Editing.

Funding

This research received no external funding.

Ethics Statement

This research did not require IRB approval because it used publicly available data with no identifying personal information.

Data Access Statement:

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Author Biography

Esha Shakoor is a graduate of the Department of Humanities at COMSATS University Islamabad, Lahore Campus. Her research interests include sentiment analysis and feminist studies

References

Ahmad Ghani, A. N. H., & Rahmat, H. (2023). Confirmation bias in our opinions on social media: A qualitative approach. *Journal of Communication, Language and Culture*, *3*(1), 44–53. https://doi.org/10.33093/jclc.2023.3.1.4

Aremu, A. O. (2024). Utilising AI-powered chatbots for learning endangered Nigerian languages and considerations for their development. *Journal of Communication, Language and Culture*, *4*(2), 41–56. https://doi.org/10.33093/jclc.2024.4.2.3

Baer, H. (2016). Redoing feminism: Digital activism, body politics, and neoliberalism. *Feminist Media Studies*, *16*(1), 17–34. https://doi.org/10.1080/14680777.2015.1093070

Balakrishnan, K. (2022). Influence of cultural dimensions on intercultural communication styles: Ethnicity in a moderating role. *Journal of Communication, Language and Culture*, *2*(1), 44–58. https://doi.org/10.33093/jclc.2022.2.1.4

Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns*, *2*(2), 100205. https://doi.org/10.1016/j.patter.2021.100205

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association*

for Computational Linguistics (pp. 5454–5476). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT Press. https://doi.org/10.7551/mitpress/11805.001.0001

Hassan, H., Riaz, M., & Zia, M. H. (2023). Analytical review of the slogans of Aurat March expressing the right of freedom. *Pakistan Languages and Humanities Review*, *7*(2), 384–400. https://doi.org/10.47205/plhr.2023(7-II)33

Hovy, D., & Spruit, S. L. (2016). The social impact of natural language processing. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 591–598. https://doi.org/10.18653/v1/P16-2096

Jamali, J., Rasool, M., & Batool, H. (2022). Code-switching by multilingual Pakistanis on Twitter: A qualitative analysis. *Khazar Journal of Humanities and Social Sciences*, *25*(2), 22–44. https://doi.org/10.5782/2223-2621.2022.25.2.22

Jia Xin, J. T., & Md. Noor, S. (2025). The effects of user-generated and influencer-generated content on beauty product purchases: Navigating scepticism in Malaysia. *Journal of Communication, Language and Culture*, *5*(2), 254–275. https://doi.org/10.33093/jclc.2025.5.2.15

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

Kelso, E. (2024). Women marching for solidarity: 5 years of Aurat March in the Islamic Republic of Pakistan. *Dialectical Anthropology, 48*(1), 113–132. https://doi.org/10.1007/s10624-024-09720-4

Khan, A. R., Karim, A., Sajjad, H., Kamiran, F., & Xu, J. (2022). *A clustering framework for lexical normalization of Roman Urdu*. *Natural Language Engineering, 28*(1), 93–123. https://doi.org/10.1017/S1351324920000285

Kshetri, N. (2024). Linguistic challenges in generative artificial intelligence: Implications for low-resource languages in the developing world. *Journal of Global Information Technology Management, 27*(2), 95–99. https://doi.org/10.1080/1097198X.2024.2341496

Liaqat, M. I., Awais Hassan, M., Shoaib, M., Khurshid, S. K., & Shamseldin, M. A. (2022). Sentiment analysis techniques, challenges, and opportunities: Urdu language-based analytical study. *PeerJ Computer Science*, *8*, Article e1032. https://doi.org/10.7717/peerj-cs.1032

Mendes, K., Keller, J., & Ringrose, J. (2019). *Digital feminist activism: Girls and women fight back against rape culture*. Oxford University Press. https://doi.org/10.1093/oso/9780190697846.001.0001

Milios, A., & BehnamGhader, P. (2022). *An analysis of social biases present in BERT variants across multiple languages*. arXiv. https://doi.org/10.48550/arXiv.2211.14402

Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 5356–5371. https://doi.org/10.18653/v1/2021.acl-long.416

Nazir, M. K., Bilal, M., & Shongwe, S. C. (2026). Sentiment analysis for code-mixed low-resource languages: A systematic review of approaches, techniques, applications, challenges, and future directions. *Social Network Analysis and Mining*, *16*, Article 47. https://doi.org/10.1007/s13278-026-01588-2

Puttick, A., Ikae, C., Rigotti, C., Fosch-Villaronga, E., Kharas, M. W., Søraa, R. A., & Kurpicz-Briki, M. (2025). A systematic review of bias detection methods for non-English word embeddings and language models. *Artificial Intelligence Review, 58*, Article 370. https://doi.org/10.1007/s10462-025-11375-8

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1668–1678). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1163

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19*(3), 321–325. https://doi.org/10.1086/266577

Shah, S. R., Irfan, H., & Akram, A. (2025). Pakistani English and cultural identity: A critical discourse analysis of English-Urdu code-mixing in social media advertisements. *Social Sciences Spectrum*, *4*(4), 455–470. https://doi.org/10.71085/sss.04.04.427

Shahzadi, I. (2025). Negotiating social identities: The role of code-switching and code-mixing in ESL classrooms in Pakistan. *South Asian Journal of Arts, Humanities and Society, 1*(1), 21–29. https://journals.usa.edu.pk/sajahs/article/view/3

Srivastava, V., & Singh, M. (2021). *Challenges and considerations with code-mixed NLP for multilingual societies*. arXiv. https://doi.org/10.48550/arXiv.2106.07823

Wajcman, J. (2004). *Technofeminism*. Polity Press.

Wyer, S., & Black, S. (2025). Algorithmic bias: Sexualized violence against women in GPT-3 models. *AI and Ethics*, *5*, 3293–3310. https://doi.org/10.1007/s43681-024-00641-0

Yi, T., & Al-Saggaf, M. A. (2025). A critical discourse analysis of immigration satire in Joe Wong's stand-up comedy. *Journal of Communication, Language and Culture, 5*(2), 1–12. https://doi.org/10.33093/jclc.2025.5.2.1

Zia, A. S. (2022). Feminists as cultural 'assassinators' of Pakistan. *Journal of International Women's Studies*, *24*(2), Article 15. https://vc.bridgew.edu/jiws/vol24/iss2/15

APPENDIX A

ChatGPT Prompt Used for Content Analysis

You are an AI tool trained to analyse social media comments related to feminist discourse.

Please analyse the following comment and identify:

1. The overall sentiment (positive, negative, neutral, or mixed)
2. Key themes or topics discussed
3. Any presence of sarcasm, irony, or humour
4. Use of code-switching (mixing of languages, e.g., Urdu and English)
5. Emotional tone (e.g., anger, support, frustration, empowerment)

The prompt was kept consistent across all data to ensure reliability in AI-generated analysis.

APPENDIX B

Intercoder Reliability Calculation (Scott's Pi)

Intercoder reliability was calculated on a stratified subset of 100 comments. The comparison between the two coders was used to compute Percent Agreement (Po), Expected Agreement (Pe), and Scott's Pi (π).

Observed Agreement (Po):
Number of agreements ÷ total comments
= 96.55%

Expected Agreement (Pe):
Calculated based on category proportions across coders
= 0.9013

Scott's Pi (π):
$\pi = (Po - Pe) / (1 - Pe)$
$\pi = (0.9655 - 0.9013) / (1 - 0.9013)$
$\pi = 0.65$

This indicates strong agreement between coders.