PPFN: A dataset of misinformation and partisan rhetoric on Philippine Facebook Pages

Anonymous ACL submission

Abstract

We curated and annotated the Philippine Facebook News corpus, a new dataset to investigate the dynamics of misinformation and engagement within the political discourse on Facebook, focusing on the Philippine political landscape. Our dataset comprises 14,495 posts and 47,889 comments from Filipino news outlets' Facebook pages, which have been annotated for topical focus, partisanship, misinformation, discussion quality, and incivility. Our analyses show that much of the misinformation identified involves unverified claims or lack of context, contributing to a hostile online environment. Political posts, especially those containing misinformation, attract more engagement and often provoke uncivil comments; yet their impacts vary depending on their partisanship. Our dataset is a novel addition to the misinformation literature in the understudied Global South, and can be utilized to further examine the role of misinformation in political discourse and engagement and the prevalence of uncivil comments in political discussions. This Corpus also offers a unique perspective on the political landscape ahead of the 2028 Philippine general elections. The dataset is available at AnonymousURL.

1 Introduction

Social media has become a primary arena for political expression, offering inclusivity, accessibility, and vast reach. Platforms such as Facebook and Twitter enable citizens to engage with political content through everyday language, allowing deliberation, contestation, and identity signaling to unfold in public view (Dahlgren, 2005). Yet, research on these spaces remains uneven. Computational studies of political communication have largely focused on detecting misinformation or modeling engagement, while linguistic work has emphasized style and affect without systematically linking them to patterns of participation. As a result, we know little about how linguistic and rhetorical strategies in user–generated content shape civic engagement, especially in multilingual, non-Western contexts where misinformation and political polarization coexist with high levels of participation.

The Philippines provides a critical context for addressing this gap. It is one of the world's most active social media environments, where political actors increasingly rely on coordinated disinformation tactics, from troll networks to manipulative framing, to influence public discourse (Ong and Cabañes, 2018). These dynamics expose a broader epistemic challenge: citizens encounter abundant political information but limited cues for assessing its reliability. In such environments, linguistic fluency, emotional tone, and stance often serve as proxies for credibility, influencing how users evaluate and share political content. Understanding these linguistic correlates of engagement is essential to explain why misinformation thrives despite widespread awareness of its presence.

We present **PPFN – Philippine Facebook News**, a bilingual (English–Tagalog) dataset comprising 14,495 posts and 47,889 comments from major Filipino news outlets' Facebook pages, collected over six months following the 2022 Presidential Election. The dataset captures the discursive aftermath of a highly polarized campaign, where narratives of legitimacy, accountability, and governance were actively negotiated. Beyond misinformation detection, the corpus enables linguistic analyses of framing, code-mixing, and stance, allowing us to examine how language patterns relate to engagement and deliberative quality.

Our approach moves beyond fact-verification to investigate how misinformation and civic reasoning interact through discourse. Citizens in digital publics respond to linguistic signals of certainty, alignment, and affect when evaluating political information. Drawing on deliberative theory (Barber, 2003; Steenbergen et al., 2003; Jaidka, 2022), we conceptualize misinformation as both a cogni-

tive and communicative phenomenon - one that exploits linguistic heuristics to guide credibility judgments, mobilize engagement, and sustain partisan discourse. We have three principal contributions:

- A bilingual dataset annotated with discourse- and misinformation-related features, combining LLM-based and human validation. The A dataset preserves content from the now-defunct CNN Philippines Facebook page, providing a rare record of disappearing media discourse;
- A political lexicon for the Philippines, capturing references to incumbents, opposition figures, and neutral actors, and
- Empirical analyses linking misinformation exposure and discussion quality to engagement behavior.

Together, these contributions offer a foundation for studying how citizens deliberate and make sense of political information under conditions of uncertainty, advancing a socio-computational perspective on misinformation and civic reasoning in the Global South.

2 Research Questions

The 2016 and 2022 Philippine presidential elections illustrate how information manipulation can shape electoral outcomes (Ong, 2022). Scholars argue that the proliferation of political misinformation has reconfigured the country's democratic processes, transforming electoral discourse and weakening institutional accountability. This transformation coincides with a profound shift in the Philippine media landscape: Facebook has overtaken traditional broadcast media as the dominant channel for political news consumption, increasing information accessibility while eroding gatekeeping functions and informational integrity. Although prior research has examined political communication on social media, there remains limited understanding of how institutional power asymmetries—particularly incumbency advantage—affect the quality, credibility, and engagement patterns of political discussions in high–social-media democracies. Appendix C provides additional context on these dynamics. This section motivates the three research questions that guide our empirical analysis.

2.1 Partisan gaps in discussion quality

Research on Philippine campaigns reveals systematic disparities in online discourse across political camps. Content about incumbents typically receives more positive engagement, reflecting their symbolic authority and media visibility (Montiel and Uyheng, 2020). In contrast, opposition figures face heightened scrutiny and are disproportionately subjected to hostile or uncivil commentary (Cabbuag and Abidin, 2024). Such asymmetries undermine deliberative norms: adversarial rhetoric and personal attacks amplify polarization and reinforce "digitally mediated echo chambers" that insulate incumbents from criticism while amplifying hostility toward challengers (Ong and Cabañes, 2018; Miranda, 2021). These patterns motivate our first research question: ***How do discussions about political opposition candidates compare with those about political incumbents in terms of discourse quality and civility (RQ1)?***

2.2 Partisan gaps in the prevalence of misinformation

While misinformation dynamics have been well documented in Western contexts, they remain underexplored in Southeast Asia. Prior work shows that politically charged misinformation spreads more effectively than neutral content, driven by affective polarization and confirmation bias (Choudrie et al., 2021; Wilson et al., 2023; Yin et al., 2024). In the Philippines, misinformation operates as a strategic communication tool weaponized by both incumbents and opposition figures (Madraki et al., 2021; Lotto et al., 2023). Yet, the directionality of these effects remains unclear: does misinformation more frequently defend incumbents or attack challengers? To address this, we ask: ***How do posts about political incumbents compare against those about opposition candidates in terms of misinformation content (RQ2)?***

2.3 Partisan gaps in user engagement

Misinformation and incivility often heighten engagement by invoking moral outrage and emotional arousal (Tandoc Jr, 2020; Yin et al., 2024). These dynamics generate "negative engagement spirals," where divisive content is algorithmically amplified through likes, shares, and comments (Miranda, 2021). However, the distribution of these engagement benefits remains poorly understood. Do such dynamics reinforce incumbents' visibility and legit-

Political Affiliation	Keywords
Incumbent	marcos, malacañang, duterte, sara, romualdez, zubiri, davao, estrada, bato, jr, pacquiao, macapagal
Opposition	robredo, diokno, leni, poe, leila, kris, manuel
Neither	gma, pangulong, quezon, china, senate, pnp, afp, kapulisan, pna, pamahalaang, congress, gov, biden, dole, edca, donald, pagcor, doh, fda, maguindanao, putin, dti, protesta, pulisya, pdrrmo, qcbd, manila, barangay, garcia, negros

Table 1: Political classification keywords.

imacy, or can opposition figures also exploit them for reach and resonance? Our final research question thus asks: ***How does the engagement in misinformation posts by the political incumbents compare with the engagement in misinformation posts by the political challengers (RQ3)?***

3 Methodology

To address our three research questions, we constructed two complementary datasets capturing both institutional and participatory dimensions of Philippine political discourse on Facebook. To examine partisan differences in discourse quality and civility (**RQ1**), we analyze language patterns within news posts and public comments. To assess the prevalence and directional targeting of misinformation (**RQ2**), we identify and annotate false or misleading content appearing in these discussions. Finally, to evaluate whether misinformation amplifies engagement and visibility for incumbents relative to opposition candidates (**RQ3**), we model user interaction metrics—reactions, comments, and shares—associated with these posts. Together, these datasets provide a comprehensive foundation for testing how linguistic, informational, and behavioral dynamics intersect in shaping online political discussions.

We curated two datasets: the Philippines Media Posts dataset, comprising posts on Facebook pages collected using the CrowdTangle API, and the Philippines Political Comments dataset, comprising posts and comments on Facebook pages available from a commercial third-party vendor, <Anonymized>. The following paragraphs describe the process followed to create both datasets. The data statement for this project is reported in Appendix A. Key definitions relevant to the dataset curation and analysis are reported in Appendix B.

3.1 Keyword Extraction

We created an English-Tagalog political keyword lexicon through a two-step process. First, we extracted the top 1000 keywords from 36,693 Facebook posts from CrowdTangle using KeyBERT. We then filtered for proper nouns and excluding English stopwords. Five trained annotators labeled these as *incumbent*, *opposition*, or *neutral*, following the instructions in Appendix F.

To capture lower-frequency but relevant terms, we applied TF-IDF on the filtered posts to identify contextually significant keywords. Only TF-IDF terms that overlapped with validated political keywords or entities were retained. These were added to the final lexicon for analysis, treating all retained terms as political regardless of stance. Finally, we used GPT-4 to classify all political keywords into *incumbent*, *opposition*, or *neither*, followed by manual verification. Ultimately, we identified 13,255 keywords, of which the ones used to distinguish partisan affiliations are reported in Table 1. Full details of the keyword extraction procedure are reported in Appendix D.

3.2 Philippines Media Posts and Comments datasets

We used the keywords reported in Table 1 to obtain Philippine Facebook news pages posts and comments from a commercial third-party vendor, <Anonymized>. The data contained 14,495 posts and 47,889 comments from March 28, 2023 up to September 28, 2023 inclusive. We split these into posts and comments and annotated each for presence of misinformation and incivility, which resulted in the Philippines Media Posts dataset and Philippines Political comments dataset.

4 Content Analysis

We conducted a three-level analysis of political discourse on Philippine Facebook Pages that went beyond fact-checking to examine the rhetorical and discursive dynamics of online political engagement. The detailed analytical framework is reported in Appendix E.

At **Level 1**, we identified the types of misinformation present (i.e. presenting outdated information, unverified claims, factual errors, inaccurate context, and jokes/satires prone to misrepresentation) in posts and comments using the misinformation codebook from Twitter's (now X's) Community Notes project (X Corp., 2025). This process involved human annotation of 1,621 posts and 2,981 comments, supported by TF-IDF-based keyword analysis and majority-voted labels with moderate inter-annotator agreement (Fleiss' Kappa

Type	Example (Original)	Translated	Justification
Contains factual error and presents outdated information	For the ruling on Gigi Reyes to apply to Leila de Lima, De Lima has to be detained at a normal cell for women in Taguig where Gigi was detained, replete with a toilet that overflows...	For the Gigi Reyes ruling to apply to Leila de Lima, she must be in the same kind of women's cell in Taguig, with an overflowing toilet...	This mixes up the law and the facts. The ruling does not depend on the cell's location or toilet. It also uses old or wrong details about where and how she is detained.
Misrepresenting or missing important context	Pinaghahandaan ng Philippine National Police ang lahat ng senaryo kabilang na ang posibleng panggugulo ng mga komunista at pananamantala ng mga kriminal sa ikalawang State of the Nation Address ni Pangulong Ferdinand Marcos Jr...	The Philippine National Police is preparing for all scenarios, including possible disturbances by communists and exploitation by criminals during President Ferdinand Marcos Jr.'s second State of the Nation Address...	This cherry-picks threats and leaves out key context. It names groups without clear evidence and makes it sound like violence is expected. It also does not show data or an official threat report.
Presents unverified claims as fact	May tsansa pa rin daw ang 20-pesos na kada kilo ng bigas, ayon kay Pangulong Bongbong Marcos.	There is still a chance for rice to cost 20 pesos per kilo, according to President Bongbong Marcos.	This is a hope or target, not a confirmed fact. There is no timeline, plan, or proof.

Table 2: Examples of misinformation in posts, with simple justifications.

Type	Example (Original)	Translated	Justification
Contains factual error	Sa senado nga po lahat daw ingles at di pilipinong wika at wala namang mga banyaga na kasali kundi lahat naman pilipino	They say that in the Senate, everything is in English and not in Filipino, and there are no foreigners involved, only Filipinos	This is wrong. The Senate can use both Filipino and English. The part about foreigners is not related and only adds confusion.
Presents outdated information	Mahal n Pangulo Bongbong Marcos maraming salamat po na binuhay muli ang kadíwa na aking naaalala noong akoy bata pa sa kadiwa kami namimili ng nanay ko...	Dear President Bongbong Marcos, thank you very much for reviving the Kadiwa which I remember from my childhood where my mother and I used to shop...	This relies on old memories. Today's program, prices, and rules are not the same as before.
Misrepresenting or missing important context	si ate glue pandakikay ang nag mastermind sa mga false accusations kay former senator leila de lima bec she is insanely motivated by vendetta	Sister Glue Pandakikay was the mastermind behind the false accusations against former Senator Leila de Lima because she is insanely motivated by vendetta	It blames one person without proof and uses insults. It ignores legal steps, evidence, and other actors, so key context is missing.
Presents unverified claims as fact	benjamin diokno kagagawan ninyo yan during Digong Administration ngayon sasabihin mo mababankrupt ang gobyerno!? isa ka na rin bang kasama ni Digong na naging bilyonaryo?	Benjamin Diokno, that was your doing during the Digong Administration, now you say the government will go bankrupt!? Are you also one of Digong's associates who became a billionaire?	It makes serious claims without evidence. It also uses a leading question as if it were a fact. No data is shown.
Contains joke or satire that may be misrepresented	Sino kaya ang mananalo kapag nagsuntukan si Dave Batista at Manny Pacquiao?	I wonder who would win if Dave Batista and Manny Pacquiao were to fight?	This is a joke that could easily be misunderstood in fast-moving feeds. Lack of tone also makes it likely for others to not see that this is satire.
Other	Defeated Dra. Atty. PHD Leni Robredo. The Clown Queen.	Defeated Dra. Atty. PHD Leni Robredo. The Clown Queen.	This is name-calling and mockery and has no checkable claim or evidence.

Table 3: Examples of misinformation in comments, with simple justifications.

= 0.54). The annotation instructions are reported in Appendix G.

At **Level 2**, we evaluated the quality of 47,789 comments in terms of their deliberative characteristics—specifically, the presence of *positivity* (constructive tone), *reciprocity* (engagement with opposing views), and *justification* (providing reasons or evidence to support claims)—following the Discourse Quality Index framework (Steenbergen et al., 2003). This large-scale annotation was conducted using large language model (LLM)-assisted labeling following the codebook provided in Jaidka (2022). The LLM annotation achieved an average accuracy of 96.07% based on manual validation by the lead author. Leveraging LLMs enabled us to efficiently scale the annotation process to cover a large volume of data without compromising reliability. The LLM annotation instructions and examples are provided in Appendix E.

Building on the identification of misinformation types (Level 1), **Level 3** focused on the rhetorical and linguistic strategies that shaped how misinformation was presented, reinforced, or challenged. We subsampled the comments dataset and analyzed their linguistic and rhetorical functions using the Systemic Functional Linguistics (SFL) framework (Halliday and Matthiessen, 2014; Eggins, 2004), which situates language in terms of field (the issues being discussed), tenor (the social roles and relationships between speakers and audiences), and mode (the medium and organization of the discourse).

Table 2 presents examples of misinformation in posts, including their original and translated text. One post, containing both a factual error and outdated information, was noteworthy for its perceived credibility, originating from the page of a prominent political figure, written in English, and carefully edited. Another example highlights the absence of context, with vague terms such as "communists" and "criminals" that could incite unnecessary alarm and therefore may constitute fear speech (Saha et al., 2023). Lastly, a post presents an unverified claim as fact, lacking supporting evidence such as economic analyses, policy references, or statements from relevant authorities.

Table 3 analyzes examples of misinformation in comments, with original and translated text. One comment, flagged for factual error, inaccurately claimed that English is the sole language used in the Philippine Senate, disregarding the official use of Filipino. Another, tagged for outdated informa-

Table 4: Rhetorical strategies in misinformation comments (SFL-aligned).

SFL Dimension	Realization in Political Comments	Linguistic Indicators	Illustrative Examples
Field (Ideational: What's happening?)	Constructing political actions, outcomes, or issues through material and relational processes. Framing events with time and social context.	Verbs like "brought", "destroyed", "is", "was", "celebrate", "free"; time or place references like "noon", "ngayon", "in the time of"	"Marcos brought the golden age", "Free Leila de Lima", "Happy birthday Atty. Leni Robredo"
Tenor (Interpersonal: Who is speaking, and to whom?)	Enacting solidarity, moral evaluation, or antagonism. Using modality to express obligation or judgment. Using mood structures (commands, questions) to provoke alignment.	Modal verbs ("must", "should", "need"), evaluative language ("corrupt", "husay", "galang"), imperatives ("Wake up!", "Free Leila")	"They must be stopped", "Husay at Galang Leni Robredo", "Free Leila de Lima", "Sino ang binoto mo?"
Mode (Textual: How is it organized?)	Organizing discourse through thematic framing, multimodal cues (hashtags, emojis, CAPS), and interactional markers (fillers, repetition, rhetorical structures).	Phrases like "What they don't want you to know is…", repetition ("not then, not now"), hashtags, ALL CAPS	"What they don't want you to know is…", "#NeverAgain", "not then, not now", "TANGINA MO!"

tion, misrepresented the recent status of the Kadiwa system, a state-run retail initiative, ignoring its intermittent operation over the years. A comment marked for misrepresentation accused legal actions against a former senator of being purely personal, without evidence. Similarly, an unverified claim alleged corruption and economic mismanagement by a public official, lacking supporting proof. A satirical comment imagined an unrealistic conflict, risking misinterpretation as a serious claim. Lastly, a comment employing derogatory language, such as "The Clown Queen," was categorized as problematic for its divisive and harmful tone, though it fell outside conventional misinformation categories.

We see that a closer reading of the data reveals that tenor—how speakers position themselves in relation to others—was particularly consequential in shaping how misinformation spread. Across comments, we observed that rhetorical moves rarely functioned in isolation; rather, they were designed to pull audiences into shared narratives, making them feel implicated, morally aligned, or socially obligated to engage. In addition to these core strategies, we observed several recurring rhetorical and linguistic patterns that further shaped audience engagement and meaning-making. The detailed analysis is reported in a Appendix E, and the salient features are:

- **Outsourced attribution and projection:** Commenters often introduced claims using vague or collective attribution (e.g., *"Sabi nila…"*, *"I heard that…"*), which distanced the speaker from direct responsibility while lending claims an air of social legitimacy (Potter, 1992; Potter and Edwards, 2001).
- **Direct audience engagement:** Comments frequently positioned readers as active participants in the discourse, using second-person address or rhetorical questions (e.g., *"Kayo na ang maghusga"*, *"Tama ba ako?"*) to invite alignment or provoke reflection (Myers, 2020; Kavada, 2020).
- **Lexical bundling and repetition:** Repetitive use of slogans, hashtags, or copypasta-style comments (e.g., *"#LabanLeni #LabanLeni #LabanLeni"*) functioned as an amplification strategy to increase algorithmic visibility and signal mass support (Clark-Parsons, 2021; Wang et al., 2016).
- **Code-mixing as identity work:** Strategic switching between English and Filipino not only marked informality but also performed stance-taking and identity signaling (Nafa, 2015; Wawire, 2020; Azahari and Mohamad, 2023). For instance, commenters began in English to project formality or rationality (e.g., *"Let's look at the facts"*), then switched to Filipino for emotional impact (e.g., *"Walang hiya kayo!"*). Others mixed both to create a conversational, insider tone (e.g., *"Think before you vote, mga kababayan"*, *"So proud to be Pinoy, pero bakit ganito?"*).

5 Dataset Characteristics

Of a total of 36,693 posts, after filtering out non-empty posts and distinguishing the political posts, we obtained 5,850 posts, or 15.94% of posts that were relevant to politics. For these posts, Table 5 presents the total count of comments and their distribution across different categories based on political affiliation. It includes the average number of comments per post and the average number of comments per post identified as containing misinformation, rounded to the nearest whole number. We see that about 62% of the comments were rele-

A: Across Facebook pages

Facebook Page	Total Posts	Avg. Likes Pol.	Avg. Likes Non-Pol.	p value	Sig.
Bongbong Marcos	749	5176	4053	2.92e-2	*
GMA News	48278	999	939	1.71e-1	No
Harry Roque	726	789	807	8.89e-1	No
Inday Sara Duterte	320	3833	3150	1.39e-2	*
Philippine Daily Sun	7	N/A	N/A	N/A	N/A
Philstar.com	25580	119	264	1.24e-10	***
Philippine News Agency	13150	60	147	1.90e-3	**
Rappler	49243	207	265	3.28e-7	***
Rody Duterte	31	N/A	N/A	N/A	N/A
UNTV News and Rescue	13763	104	126	3.63e-8	***
VP Leni Robredo	225	3516	4812	2.55e-2	*
CNN Philippines	855	58	48	N/A	N/A

B: Across partisan affiliations

Category	Total Comments	Mean engagement All content	Mean engagement Misinfo Posts	Chi-Square Test
Incumbents	19613	53	63	$\chi^2(1, N=30252) = 6.15$,
Opposition	10639	56	12	$p = 0.013$

Table 5: User engagement across Facebook pages and political affiliations. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Category	Posts Not sure	Posts Yes	Comments Yes
Misrepresentation	9	43	314
Factual error	4	1	665
Unverified claim	4	11	1375
Digitally altered content	2	-	-
Outdated information	47	10	13
Joke or satire	-	-	574
Other	-	40	-

Table 6: Misinformation label frequency in posts and comments.

vant to politics. For incumbents, there were 19,613 comments, with an average of 53 comments per post and 63 comments per misinformation post. In contrast, the opposition had 10,639 comments, averaging 56 comments per post but only 12 comments per misinformation post. 211 comments discussed incumbents and opposition, while 17416 comments comprised unrelated posts. The data highlights a significant engagement gap in misinformation posts influenced by the political group being discussed.

Table 5a and b shows the total count and breakdown of each type of misinformation present in the annotated posts and comments, providing the foundation for Level 1 of our analysis. Out of 1,621 annotated posts, 44 posts (2.71%) were identified as containing factual inaccuracies or misleading framings, such as omitting key context. 147 posts (9.06%) were marked as containing potentially harmful content, either targeting individuals (n = 29), social groups (n = 88), or society at large (n = 30). We compared the engagement metrics (total number of likes) for political versus non-political posts from the same Facebook pages for the same period, using unpaired t-tests. This is shown in Table 5b: Facebook pages of incumbents Bongbong Marcos and Inday Sara Duterte have much higher average likes for political posts compared to non-political posts. However, the Facebook page of opposition Leni Robredo has much less average likes for political posts compared to non-political posts.

6 Results

To answer **RQ1** (how do posts about political opposition candidates compare with those about political incumbents in terms of discourse quality and civility), we annotated 47,789 comments in the Philippines Media Posts dataset using GPT-4 and analyzed the quality of discussions. We assessed statistical significance using chi-square tests. Our findings are presented in Table 7, Section (A). The first two rows show that comments about opposition candidates exhibit significantly higher positivity (30.9%) compared to comments about incumbents (16.7%), suggesting either public dissatisfaction with incumbents or a general preference for change. In contrast, incumbents have slightly higher rates of reciprocity (3.4%) compared to the opposition (3.0%), indicating that discussions about incumbents may be marginally more interactive. Both groups, however, have low rates of justification, and the majority of comments are not constructive overall.

Section (B) of Table 7 compares incivility across comments. The nature of uncivil comments differs: comments on posts about incumbents show higher rates of exaggeration (6.3%), while comments on posts about the opposition display higher rates of abuses (31.8%) and threats (1.6%).

To answer **RQ2** (How do posts about political incumbents compare against those about opposition candidates in terms of misinformation content?)), we analyzed 1,621 posts for the presence of misinformation. As shown in Section (C) of Table 7, posts about the opposition are more likely to contain misinformation (7.04%) compared to posts about incumbents (3.82%). This supports prior reports suggesting that misinformation is more frequently directed against opposition candidates.

Finally, to answer **RQ3** (How does the engage-

Classification	Total Count	Percentage	Chi-Square Test
(A) Discussion quality			
Positivity			
Incumbent	19,613	16.7%	χ^2(3, N=47799) = 875.71,
Opposition	10,639	30.9%	$p < 0.0001$
Reciprocity			
Incumbent	19,613	3.4%	χ^2(3, N=47799) = 875.71,
Opposition	10,639	30.9%	$p < 0.0001$
Justification			
Incumbent	19,613	3.3%	χ^2(3, N=47799) = 875.71,
Opposition	10,639	1.9%	$p < 0.0001$
(B) Incivility (Abuses / Threats / Exaggerations)			
Classification	**Civil Discussions**		
Incumbent	61.3%	31.2% / 1.2% / 6.3%	χ^2(6, N=47789) = 44.32,
Opposition	62.5%	31.8% / 1.6% / 4.1%	$p < 0.0001$
(C) Misinformation			
Incumbent	810	3.82%	χ^2(1, N=1621) = 7.5426,
Opposition	811	7.04%	$p < 0.01$

Table 7: Comparison of discussion quality, incivility, and misinformation in politically affiliated comments

Predictor	Estimate	Std. Error	t-value	p-value
Misinformation	-0.326	0.253	-1.287	0.198
Political affiliation	0.002	0.072	0.034	0.973
Sentiment (Ternary)	0.048	0.050	0.953	0.341
Misinfo. × Pol. affil.	0.765	0.315	2.427	0.015*
Additional Model Information				
Residual Std. Error: 1.372 on 1607 degrees of freedom				
Multiple R-Squared (full)	0.3296			
Adjusted R-Squared (full)	0.3242			
F-statistic (full): 60.79 on 13 and 1607 DF, $p = 0.001$				

Note: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

Table 8: Regression results for comment engagement

ment in misinformation posts by political incumbents compare with the engagement in misinformation posts by political challengers?), we conducted a regression analysis predicting the number of comments a post received based on its misinformation status and political alignment. The results, presented in Table 8, reveal a statistically significant interaction: misleading posts about incumbents receive significantly more engagement than misleading posts about opposition candidates. Specifically, the interaction term has an estimated coefficient of 0.765. When exponentiated ($e^{0.765} \approx 2.15$), this suggests that misleading posts about incumbents receive approximately **2.15 times more comments** than baseline posts. Applied to the dataset's median of 12 comments, this corresponds to **about 25.8 comments on average** ($12 \times 2.15 \approx 25.8$). This highlights a clear engagement advantage for incumbents when misinformation is present.

7 Discussion

Our findings reveal that political talk on Philippine Facebook pages functions not merely as the transmission of information but as an ongoing negotiation of meaning and belonging. Language was used to construct shared social realities, manage relational dynamics, and signal affective alignment within the public. Through rhetorical strategies such as code-mixing, humor, and direct address, commenters transformed political discourse into an interactive performance that blurred the line between news consumption and identity expression. These linguistic practices made political messaging—and misinformation—feel authentic, emotionally resonant, and personally relevant. In multilingual contexts such as the Philippines, where credibility often depends on linguistic fluency and tone rather than verified evidence, these stylistic cues serve as heuristic signals guiding how users assess truth and credibility (Dahlgren, 2005; Nafa, 2015; Wawire, 2020; Azahari and Mohamad, 2023).

Empirically, our results across the three research questions demonstrate how these linguistic and discursive dynamics interact with structural power. Discussions about incumbents consistently attracted higher engagement, particularly when misinformation was present. This pattern supports the notion of an "incumbency advantage" in networked political communication, where those

already in power benefit from greater visibility and algorithmic amplification. Engagement may be driven less by belief in the misinformation itself than by its rhetorical salience: emotionally charged, morally framed, or linguistically playful content stimulates participation regardless of factual accuracy (Tandoc Jr, 2020; Miranda, 2021). Meanwhile, opposition figures faced higher levels of incivility, echoing prior observations that challengers are more frequent targets of hostility and derision (Cabbuag and Abidin, 2024; Ong and Cabañes, 2018). Together, these findings illustrate that linguistic form—how something is said—can shape not only engagement but also the perceived legitimacy of the message. Finally, by preserving content from the now-defunct CNN Philippines Facebook page, the PPFN corpus serves as a historical archive of public discourse in a constrained media environment.

From a linguistic perspective, misinformation in this dataset operates as both discourse and performance. Building on the Systemic Functional Linguistics framework (Halliday and Matthiessen, 2014; Eggins, 2004), our qualitative analysis revealed how rhetorical functions—particularly tenor and mode—shape credibility and emotional uptake. Tenor-related moves such as the use of collective pronouns, vague attribution (*"Sabi nila"*, "they said"), and second-person address (*"Kayo na ang maghusga"*) established social proximity and distributed accountability (Potter, 1992; Potter and Edwards, 2001; Myers, 2020). Similarly, code-mixing between English and Filipino performed stance-taking and identity signaling, projecting both rationality and emotional authenticity. These forms of linguistic hybridity position speakers as both informed citizens and affective participants, reinforcing solidarity while inviting alignment. Even repetitive lexical bundling (e.g., slogans or hashtags) functioned less as redundancy and more as a rhythmic performance of belonging (Wang et al., 2016; Clark-Parsons, 2021). The combination of these features transformed misinformation from a static falsehood into an interactive practice of meaning-making.

Viewed through this lens, misinformation's success cannot be attributed solely to cognitive deficits or lack of media literacy. As Marres (2018) argues, false claims often endure because they express unarticulated social concerns.The Marcos-era myth of the "anti-typhoon rocket" (Talamayan, 2021) exemplifies this process: although scientifically implausible, the narrative draws symbolic power from real conditions of vulnerability to natural disasters and frustration with state inefficacy (Library of Congress, 1974; Government of the Philippines, 1977). Such stories transform misinformation into a vernacular form of political communication that blends hope, grievance, and critique.

The PPFN dataset offers scholars and policymakers a resource for understanding how digital platforms and government censorship can mediate power and control information in emerging democracies. Our findings highlight how misinformation is a complicated sociolinguistic phenomenon, which may explain why attempts to counter it through fact-checking alone might fail: interventions may address propositional content while neglecting the relational and performative dimensions of online talk.

8 Conclusion

The PPFN dataset provides a new empirical foundation for studying the intersection of misinformation, rhetoric, and engagement in Southeast Asian social media. By combining bilingual text, misinformation annotation, and discourse-quality measures, it enables analysis of how citizens and political actors co-construct meaning through language and interaction. Our findings reveal that misinformation targeting incumbents attracts disproportionate engagement, that opposition figures experience more incivility, and that these asymmetries are reinforced through linguistic and rhetorical strategies such as code-mixing, repetition, and audience address. Together, these results demonstrate how linguistic heuristics mediate credibility and participation in multilingual, high-participation environments.

Beyond its empirical contribution, this work suggests that in multilingual, high-participation environments, effective interventions must therefore engage with the linguistic fabric of civic discourse. Rather than treating misinformation as a pathology of irrational publics, scholars and practitioners should recognize it as an expressive register through which citizens negotiate belonging, critique authority, and imagine alternative futures. Addressing misinformation thus requires not only technological solutions but also attention to the linguistic, cultural, and affective structures that sustain interpretive trust in digital publics.

Limitations

While our study explores misinformation patterns and engagement in the Philippine context during the post-election period, we caution against interpreting these correlations as evidence of causation. The observed trends may be influenced by various external factors, such as heightened political discourse, shifts in public sentiment, or platform-level algorithmic changes during this period. Additionally, the post-election environment often amplifies polarized discussions and could shape how misinformation spreads, but these dynamics are not unique to elections and may not fully explain broader societal trends. Further research is necessary to disentangle the specific effects of the election period from underlying patterns in misinformation and engagement.

We would also like to acknowledge several important ethical and methodological considerations. We conducted thorough data documentation and implemented robust anonymization protocols to ensure responsible research practices. The reproducibility of our findings is supported by detailed methodology descriptions and carefully preserved experimental configurations. While our analysis demonstrates promising results, we recognize certain limitations and potential concerns. Specifically, we examined the number of misclassifications through an inter-coder reliability analysis and found that the high inter-annotator percentage agreements achieved can ameliorate any concerns around the same. To mitigate these risks, we implemented a two-week training period and established clear guidelines for responsible annotator protocol. Our data analysis accounts for population-specific distributions and potential artifacts, which we address through the stratified sampling of posts from all news pages and across political affiliations. All the findings constitute, and the data published comprises, deidentified data reported as cumulative, post-level effects.

Nonetheless, these findings carry important implications for practice. News organizations and media platforms can use these insights to inform their content moderation and verification strategies, particularly focusing on the subtle, but impactful issue of partial truths. The observed incumbency advantage in our study highlights the need for more equitable representation across political groups. Additionally, educators, curriculums, and organizations could integrate these findings into media literacy campaigns to help the public better recognize and critically assess misinformation. Finally, policymakers may consider these results in developing tailored, context-sensitive approaches to regulating discourse in social media, especially during key political periods, such as during election and regime changes.

In addition, our dataset offers a significant contribution to both researchers and practitioners. As one of the first multilingual, LLM- and human-annotated datasets focused on political discourse in the Philippines, it provides a valuable resource for studying misinformation, engagement, and linguistic dynamics across English, Filipino, and code-switched content. Notably, the dataset includes preserved content from the now-inaccessible CNN Philippines Facebook page, which would otherwise be lost to researchers. We hope that this dataset can support further inquiry and collaborative efforts to understand and address the complex challenges of digital political communication in Southeast Asia.

9 Ethics Statement

The data collection and annotation processes posed minimal risk to participants, as all data analyzed in this study consisted of publicly available, anonymous social media posts and comments. Personally identifying data about the authors of the posts and comments was never provided to the researchers, neither from the CrowdTangle API nor the commercial vendor.

While some of the content involved politically sensitive or offensive language, annotators were fully briefed on the nature of the task and were free to withdraw at any time without penalty. Annotators were compensated at an hourly rate of USD $15, which is consistent with fair compensation guidelines for crowd work. The annotation guidelines, examples, and data statistics are documented in the appendices of this paper.

Acknowledgment: AI was used to proofread a draft of this paper and format the tables.

References

Sacha Altay and colleagues. 2023. Misinformation and its correction: Cognitive mechanisms and recommendations for future research. *Current Opinion in Psychology*, 47:101407.

Benedict Anderson. 1988. Cacique democracy and the philippines: origins and dreams. *New Left Review*, 169:3–31.

Nur Khairunnisha Azahari and Nur Husna Serip Mohamad. 2023. Bridging identities: Analysing code-mixing in yuna's conversation with a local malaysian activist. *Journal for the Study of English Linguistics*, 11(1).

Benjamin R Barber. 2003. *Strong democracy: Participatory politics for a new age*. Univ of California Press.

Victor Felipe Bautista. 2018. The pervert's guide to historical revisionism: Traversing the marcos fantasy. *Philippine Studies: Historical & Ethnographic Viewpoints*, pages 273–300.

Walden Bello, Herbert Docena, Marissa de Guzman, and Mary Lou Malig. 2006. *The Anti-Development State: The Political Economy of Permanent Crisis in the Philippines*. Zed Books, London.

Samuel I Cabbuag and Crystal Abidin. 2024. Tiktok 'dogshows' and the amplification of online incivility among gen z influencers in the philippines. *International Journal of Cultural Studies*, page 13678779241302826.

Jyoti Choudrie, Snehasish Banerjee, Ketan Kotecha, Rahee Walambe, Hema Karende, and Juhi Ameta. 2021. Machine learning techniques and older adults processing of online information and misinformation: A covid 19 study. *Computers in human behavior*, 119:106716.

Rosemary Clark-Parsons. 2021. "i see you, i believe you, i stand with you":# metoo and the performance of networked feminist visibility. *Feminist Media Studies*, 21(3):362–380.

Cesi Cruz, Julien Labonne, and Pablo Querubin. 2017. Politician family networks and electoral outcomes: Evidence from the philippines. *American Economic Review*, 107(10):3006–3037.

Peter Dahlgren. 2005. The internet, public spheres, and political communication: Dispersion and deliberation. *Political communication*, 22(2):147–162.

Suzanne Eggins. 2004. *An Introduction to Systemic Functional Linguistics*, 2nd edition. Continuum.

Tatjana Radanović Felberg. 2008. Evaluation in media discourse: analysis of a newspaper corpus.

Government of the Philippines. 1977. Presidential decree no. 1149: Amending certain sections of presidential decree no. 78, otherwise known as 'the atmospheric, geophysical and astronomical science act of 1972'. Accessed from The LawPhil Project.

Michael Alexander Kirkwood Halliday and Christian MIM Matthiessen. 2014. *An Introduction to Functional Grammar*, 4th edition. Routledge.

Tim Highfield. 2017. *Social media and everyday politics*. John Wiley & Sons.

Paul David Hutchcroft. 1998. *Booty capitalism: The politics of banking in the Philippines*. Cornell University Press.

Kokil Jaidka. 2022. Talking politics: Building and validating data-driven lexica to measure political discussion quality. *Computational Communication Research*, 4(2):486–527.

Anastasia Kavada. 2020. Creating the collective: social media, the occupy movement and its constitution as a collective actor. In *Protest technologies and media revolutions*, pages 107–125. Emerald Publishing Limited.

Jay L Lemke. 2005. *Textual politics: Discourse and social dynamics*. Taylor & Francis.

Stephan Lewandowsky, Ullrich K H Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2013. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131.

Library of Congress. 1974. *Astronautics and Aeronautics, 1972: Chronology of Science, Technology, and Policy*. NASA, Washington, D.C.

Matheus Lotto, Olivia Santana Jorge, Maria Aparecida de Andrade Moreira Machado, and Thiago Cruvinel. 2023. Exploring online oral health misinformation: a content analysis. *Brazilian Oral Research*, 37:e049.

Brendan Luyt and Karryl Sagun-Trajano. 2023. Grappling with polarization on wikipedia: the case of the biography of ferdinand e. marcos. *Journal of Documentation*, 80(2):320–336.

Golshan Madraki, Isabella Grasso, Jacqueline M. Otala, Yu Liu, and Jeanna Matthews. 2021. Characterizing and comparing covid-19 misinformation across languages, countries and platforms. In *Companion proceedings of the web conference 2021*, pages 213–223.

Pedro C Magalhães. 2015. The impact of the economy on the vote in established and new democracies. *Electoral Studies*, 38:94–103.

Noortje Marres. 2018. Why we can't have our facts back. *Engaging Science, Technology, and Society*, 4:423–443.

James R Martin et al. 2001. Language, register and genre. *Analysing English in a global context: A reader*, pages 149–166.

Patricia Blardony Miranda. 2021. A disproportionately unequal playing field: Challenges to and prospect for campaign finance law and policy in the philippines. *Asia-Pacific Journal of Elections and Democracy*, 1(01):50–62.

Cristina Jayme Montiel and Joshua Uyheng. 2020. Mapping contentious collective emotions in a populist democracy: Duterte's push for philippine federalism. *Political Psychology*, 41(4):737–754.

Greg Myers. 2020. Conversation analysis. In *Researching Discourse*, pages 19–35. Routledge.

Hanan Omar A Ben Nafa. 2015. Code-switching and social identity construction among arabic-english bilinguals: A stance perspective. In *Papers from the 10th Lancaster University Postgraduate Conference in Linguistics & Language Teaching 2015*, page 1.

Jonathan Corpus Ong. 2022. Philippine elections 2022. *Contemporary Southeast Asia*, 44(3):396–403.

Jonathan Corpus Ong and Jason Vincent Cabañes. 2018. Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the philippines.

William H Overholt. 1986. The rise and fall of ferdinand marcos. *Asian Survey*, 26(11):1137–1163.

Jonathan Potter. 1992. *Representing Reality: Discourse, Rhetoric and Social Construction*. SAGE.

Jonathan Potter and Derek Edwards. 2001. Discursive social psychology. *The New Hanbook of Language and Social Psychology/Edited by WP Robinson and H. Giles.–Hoboken: Willey*, 688:2.

Diego Saez-Trumper, Mounia Lalmas, and Nicolas Castell. 2019. Fighting misinformation in social media: Characterizing the spread of rumors in whatsapp groups. In *Proceedings of the 10th ACM Conference on Web Science*, pages 145–154. ACM.

Punyajoy Saha, Kiran Garimella, Narla Komal Kalyan, Saurabh Kumar Pandey, Pauras Mangesh Meher, Binny Mathew, and Animesh Mukherjee. 2023. On the rise of fear speech in online social media. *Proceedings of the National Academy of Sciences*, 120(11):e2212270120.

Marco R Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1(1):21–48.

Fernan Talamayan. 2021. The politics of nostalgia and the marcos golden age in the philippines. *Asia Review*.

Edson C Tandoc. 2020. Differentiating misinformation, disinformation, and malinformation: A typology of problematic information. *Journalism & Mass Communication Quarterly*, 97(2):532–550.

Edson C Tandoc Jr. 2020. Tools of disinformation: How fake news gets to deceive. In *Disinformation and fake news*, pages 35–46. Springer.

Rong Wang, Wenlin Liu, and Shuyang Gao. 2016. Hashtags and information virality in networked social movement: Examining hashtag co-occurrence patterns. *Online information review*, 40(7):850–866.

Gorrety Nafula Wawire. 2020. English-swahili code switching: An intersection between stance and identity. Master's thesis, University of Mississippi, Oxford, MS.

Anna Wilson, Seb Wilkes, Yayoi Teramoto, and Scott Hale. 2023. Multimodal analysis of disinformation and misinformation. *Royal Society Open Science*, 10(12):230964.

X Corp. 2025. Community Notes Guide: Introduction. https://communitynotes.x.com/guide/en/about/introduction. Accessed: 2025-05-18.

Jason Yin, Tzu-Chin Wu, Chia-Yun Chen, Fen Lin, and Xiaohui Wang. 2024. Diffusion of vaccine misinformation in a taiwanese online community: The role of influencers and echo chambers (preprint). *JMIR Infodemiology*.

sectionappendix subsectionappendix subsubsectionappendix

A Datasheet for the PPFN Dataset

A.1 Motivation

For what purpose was the dataset created?
The PPFN dataset was created to investigate the dynamics of misinformation and engagement within political discourse on Facebook, specifically focusing on the Philippine political landscape. It aims to explore how misinformation influences political discussions and engagement, particularly during the post-election period.

Who created the dataset?
The dataset was curated and annotated by Anonymous.

Who funded the creation of the dataset?
This work was supported by Anonymous.

A.2 Composition

What do the instances that comprise the dataset represent?
Posts and comments from Filipino news outlets' Facebook pages, annotated for topics, misinformation, discussion quality, and incivility.

How many instances are there in total?
The dataset consists of 14,495 posts and 47,889 comments collected from Facebook.

Does the dataset contain all possible instances?
No, it is a sample focused on political content during the six months following the 2022 Philippine Presidential Elections.

What data does each instance consist of?
Each instance includes the post or comment text, engagement metrics (likes, shares, etc.), and annotations for misinformation, incivility, and discussion quality.

Is there a label or target associated with each instance?

Yes, including misinformation categories, civility levels, and discussion quality ratings.

Is any information missing from individual instances?
Instances are complete as per the defined scope, though annotation limitations apply.

Are relationships between individual instances made explicit?
Yes, relationships such as replies and comment chains are represented.

Are there recommended data splits?
No.

Are there any errors, sources of noise, or redundancies?
As with any social media data, instances may include sarcasm, slang, or code-switching between English and Tagalog.

Is the dataset self-contained, or does it link to external resources?
The dataset is self-contained but includes metadata derived from external analysis.

Does the dataset contain data that might be considered confidential?
No, all data is from publicly available Facebook pages.

Does the dataset contain data that might be offensive, insulting, or threatening?
Some comments may include uncivil or offensive language, reflecting real-world political discourse.

A.3 Collection Process

How was the data associated with each instance acquired?
Collected using the CrowdTangle API and sourced from Anonymous, covering Facebook pages of Filipino news outlets.

What mechanisms were used to collect the data?
Posts and comments were identified using political keywords and annotated by trained Filipino annotators.

If the dataset is a sample, what was the sampling strategy?
Stratified sampling to balance posts about incumbents, opposition, and general political discourse.

Who was involved in the data collection process?
The authors and trained Filipino annotators with native-level Tagalog proficiency.

Over what timeframe was the data collected?
Six months following the 2022 Philippine Presidential Elections.

Were any ethical review processes conducted?
Yes, with anonymization and responsible data handling protocols.

A.4 Preprocessing/Cleaning/Labeling

Was any preprocessing/cleaning/labeling of the data done?
Yes, including removal of empty posts and keyword extraction using KeyBERT.

Was the raw data saved in addition to the preprocessed data?
Yes.

Is the software that was used to preprocess the data available?
The code for annotation and analysis is documented but not publicly shared.

A.5 Uses

Has the dataset been used for any tasks already?
Yes, it has been analyzed for studying misinformation's impact on engagement and civility.

Is there a repository linking to papers or systems that use the dataset?
Yes, available at Anonymous.

What (other) tasks could the dataset be used for?

- Studying misinformation dynamics
- Analyzing political engagement patterns
- Exploring the relationship between civility and misinformation

Is there anything about the composition of the dataset that might impact future uses?
The dataset reflects the Philippine political and cultural context, which may introduce context-specific biases.

Are there tasks for which the dataset should not be used?
It is unsuitable for non-political or private data analysis contexts.

A.6 Distribution

How will the dataset be distributed?
Publicly accessible at Anonymous.

What license is the dataset distributed under?
Creative Commons Attribution 4.0 International (CC BY 4.0).

A.7 Maintenance

Who will be supporting/maintaining the dataset?
The authors.

Will the dataset be updated?
No updates are planned, but revisions will be announced via the distribution platform.

B Dataset definitions

In this paper, we adopt the following operational definitions for key concepts used in our annotation process:

1. **Misinformation** — Information that is false or misleading, regardless of the intent to deceive (Altay and colleagues, 2023). Misinformation may arise from incorrect facts, misinterpretations, or incomplete context and can spread unintentionally. For our analyses, we have considered misinformation as the systematic distortion of truth rather than merely the propagation of falsehoods.

2. **Disinformation** — Deliberately misleading information intended to create false beliefs or manipulate public perception (Saez-Trumper et al., 2019). Disinformation differs from misinformation in that it involves a conscious effort to deceive audiences.

3. **Post Quality** — The degree to which a post presents accurate, clear, and constructive content. High-quality posts are fact-checked, well-researched, and foster meaningful engagement, while low-quality posts may lack clarity, contain misinformation, or provoke incivility (Tandoc, 2020).

4. **Civility** — The degree of respect and politeness exhibited in communication. Civil discourse avoids inflammatory language, derogatory remarks, and personal attacks, promoting constructive engagement (Lewandowsky et al., 2013).

5. **Incumbents** — Individuals currently holding political office or positions of power, including members of political dynasties. Incumbents typically benefit from name recognition, established networks, and access to resources that increase their re-election advantages.

6. **Opposition** — Political candidates or parties that challenge incumbents or the ruling government. Opposition actors often face heightened scrutiny and may be frequent targets of misinformation and disinformation campaigns in polarized political environments (Magalhães, 2015).

C Background and Related Work

The Philippine political system The Philippines has had long experience of electoral democracy. While a limited franchise existed at the civic level of government under the Spanish, as a direct colony of the United States (1898-1935), electoral democracy was greatly expanded so that by the end of the period Filipino participation in the running of government was at a level unknown in the possessions of other colonial powers. And from the inauguration of the Philippine Commonwealth in 1935 to the Japanese occupation of 1941, the country was self-governing except in the fields of foreign affairs and defense. Electoral democracy continued after the war and national independence in 1946, only interrupted by a long period of martial law from 1972 to 1981 and shorter periods in Mindanao (the most recent being declared in 2017 and revoked in 2019).

Despite this history of electoral democracy, however, it is also apparent that the Philippine political system has significant weaknesses. Political parties tend to exist not for the implementation of any particular ideological or policy agenda, but rather for the advancement of the interests of broader family dynasties. In other words, they serve as mechanisms to channel the economic largesse of the state into private hands. On the other hand, the existence of a relatively competitive electoral democracy also serves to regulate the flow of state resources among the elite families that jockey for political power, curbing the potential for one group of families to monopolize the state in perpetuity (Anderson, 1988; Bello et al., 2006; Hutchcroft, 1998). Finally, it needs to be noted that to at least a limited extent electoral democracy is also a means for those in the middle and at the bottom of the economic pile to also register their demands and dissatisfaction, gaining something from the state in the process.

Electoral competition and liberal democracy are therefore a key element in stabilizing Philippine society so that anything that has the potential to upset this delicate system needs scrutiny. Social media in the Philippines, with its ability to disseminate half-truths and other information distortions, seems to be ripe for study as a result.

Misinformation during the Philippine General Elections 2022. In the Philippines, the recent electoral cycles, notably the presidential con-

tests of 2016 and 2022, demonstrate the strong impact of information manipulation on electoral outcomes (Ong, 2022). Scholars argue that the proliferation of misinformation in Philippine political discourse has fundamentally altered the nation's democratic processes, particularly within the context of electoral campaigns and policy deliberations. This was in part spurred by the transformation of the Philippine media landscape, characterized by the ascendancy of Facebook as the predominant source of news consumption, marking a decisive shift from traditional broadcast media. While this digital transformation has enhanced accessibility to information, it simultaneously presents substantial challenges to information integrity.

The 2022 presidential election, contested between Ferdinand Marcos Jr. and Leni Robredo, presents a compelling case study of this phenomenon. The historical context is significant: Ferdinand Marcos Jr.'s father, Ferdinand Marcos Sr., governed the Philippines from 1965 until his removal from power in 1986. While initially celebrated as a visionary leader who attempted to propel the Philippines down the same growth trajectory as the rest of the southeast Asian "miracle" economies, towards the end of his time in power, his regime was characterized by authoritarian governance, systematic human rights violations, widespread corruption, and the implementation of martial law (Overholt, 1986). Despite this contentious legacy, the Marcos family implemented a sophisticated social media strategy to reconstruct their historical narrative (Bautista, 2018). This digital campaign systematically promoted narratives emphasizing the purported economic achievements and infrastructural development during the Marcos Sr. administration, characterizing it as a period of unprecedented prosperity or 'golden age' (Luyt and Sagun-Trajano, 2023). The efficacy of this historical revisionism was evidenced by Marcos Jr.'s decisive electoral victory over Robredo, underscoring the potency of strategic information manipulation in contemporary Philippine politics.

The Philippine political landscape is distinguished by its deep integration into social and cultural structures, where political affiliations often intersect with familial and community relationships, generating strong emotional and loyalty-based connections. Research indicates that provincial affiliation significantly influences voting behavior, with voters demonstrating a marked preference for candidates from their home regions (Ong and Cabañes, 2018; Cruz et al., 2017). This sociocultural framework facilitates heightened engagement with political content, particularly within social media ecosystems. While extensive research has examined political discourse on social media platforms, there remains a critical gap in understanding how institutional power dynamics, particularly incumbency advantage, shape the quality and nature of online political discussions in democracies with high social media penetration rates.

D Keyword extraction

Manual keyword labeling:
The raw data resulted in 36,693 posts, which included columns describing the type of the post, the message, source, etc. The raw data was then filtered to remove empty posts (posts where the 'message' column is NaN). Afterwards, keywords were extracted from the posts using KeyBERT from the keybert library. Keywords were limited to proper nouns and stop words in English were excluded from the list. Finally, the keywords were ranked by their raw frequency in the dataset. The top 1000 entries were then reserved for five annotators for labeling political posts as 'incumbent,' 'opposition,' or 'neutral,' following the instructions reported in Appendix F.

Semisupervised keyword labeling: Not all relevant keywords might have been captured because the top 1000 entries were only graded by absolute frequency. To address this, we applied TF-IDF (Term Frequency–Inverse Document Frequency) on the filtered set of Facebook post messages (after removing empty posts) to extract additional keywords. TF-IDF scores were computed at the document level, where each document corresponds to a single Facebook post. This allowed us to identify terms that were relatively unique to individual posts but not widespread across the corpus—helping surface politically relevant but low-frequency terms that frequency-based ranking may have missed. While raw frequency favors commonly used terms, TF-IDF downweighs those and emphasizes terms with higher contextual significance. This was particularly useful in identifying named entities or political expressions tied to specific events or figures. To ensure consistency, only TF-IDF terms that overlapped with the broader political discourse (e.g., appeared alongside already validated political keywords or named political entities) were retained. These TF-IDF-derived keywords were then added

to the set of political keywords for downstream classification. Regardless of classification into incumbents, opposition, or neutral, all retained terms were treated as political for the purpose of analysis.

Classification into political incumbents and opposition:
Afterwards, the final list of political keywords was classified into incumbents, the opposition, and neither using GPT-4. We then verified the processed list to ensure accuracy of the classification. Table 1 shows the breakdown of the political keywords into their respective classifications.

E Analytical framework

E.1 Level 1: Annotating misinformation types

We began by preparing a random sample of 100 posts and 100 comments for exploratory annotation. Of these, 9% of posts and 3% of comments were written entirely in English, while 91% of posts and 88% of comments featured a mix of English and Filipino. A small percentage of comments (4%) were in other languages. To preserve the linguistic authenticity and rhetorical nuance of the data, we did not translate the content. Instead, we engaged Filipino annotators with native-level proficiency in Tagalog and English to ensure accurate interpretation.

Posts were first classified by political alignment as either supporting the incumbent, the opposition, or neither, based on the presence of specific keywords (reported in Table 1). Posts mentioning both were assigned to the category with the higher keyword count. We filtered the dataset to retain only posts about the incumbents or the opposition, resulting in a sample of 1,621 posts:

- 725 posts about incumbents
- 725 posts about the opposition, and
- 171 additional posts scraped from the CNN Philippines Facebook page (details in Appendix C)

We also processed 2,981 comments on these posts for annotation.

Filipino college students, trained by the authors, annotated whether posts and comments contained misleading information, using a codebook adapted from Twitter's Community Notes project (see grey boxes in Appendix G).

In the training stage, a random sample of 300 comments was annotated by all three annotators. After the annotators achieved at least 80% agreement on whether the comment was politically relevant and whether it was uncivil, the remaining 2,681 comments were split evenly among the three annotators (893, 894, and 894 comments each). Annotations on the 300 jointly labeled comments were finalized through majority voting and merged with the rest of the data. The Fleiss' Kappa for incivility annotations was 0.54, reflecting moderate agreement, likely due to the subjectivity of interpreting incivility and misinformation. To improve robustness, we supplemented annotations with TF-IDF-based keyword analysis, manual review, and spot checks to ensure sensible labeling across the dataset. The detailed inter-annotator reliability statistics are reported in Appendix I.1.

E.2 Level 2: Annotating Discussion Quality

To extend our annotation at scale, we used GPT-4 (via the OpenAI API) in a few-shot setting to label 47,789 comments for discussion quality. Our annotation prompt was derived from a codebook by Jaidka (2022) (see Appendix H). To minimize hallucination, we ran separate notebooks for each annotation category, and applied the annotations directly to the original, untranslated comments in Tagalog, English, or mixed code, preserving contextual meaning. The LLM-based annotations were validated by the lead author using random samples of 100 comments per category. The label distribution is reported in Table 7. The average annotation accuracy was 96.07%, supporting the reliability of the automated annotations. Full validation metrics and examples are reported in Appendix I.2.

E.3 Level 3: Analyzing Rhetorical and Linguistic Functions using SFL

Building on the insights from misinformation and discussion quality annotations, we turn to the linguistic and rhetorical strategies that characterize political discourse on Philippine Facebook Pages. While Levels 1 and 2 identify what is being said and how audiences respond, Level 3 asks how language itself functions to frame political realities, position social actors, and manage public engagement.

To guide this analysis, we apply the Systemic Functional Linguistics (SFL) framework (Halliday and Matthiessen, 2014; Eggins, 2004), which distinguishes three metafunctional dimensions of language:

- Field: What the discourse is about—its activities, issues, or narratives.
- Tenor: Who is involved—the roles, relationships, and alignments between speakers and audiences.
- Mode: How the discourse unfolds—its medium, style, and organization.

We adapted this framework to map the discursive functions observed in our dataset, focusing on how political actors and commenters frame issues (Field), position themselves and others (Tenor), and manage engagement (Mode) through linguistic and stylistic choices:

  - **Field:** Constructing political realities through verbs and participants that assign blame or praise (e.g., "Marcos brought the golden age", "Free Leila de Lima"), and using temporal framing to situate events in shared historical narratives (e.g., "in the time of Aquino") (Halliday and Matthiessen, 2014; Lemke, 2005).
  - **Tenor:** Enacting solidarity or antagonism through collective identity appeals (e.g., "tayo mga Pilipino", "taumbayan"), negative labeling (e.g., "dilawan", "DDS trolls"), and evaluative language (e.g., "husay at galang", "corrupt as hell") (Martin et al., 2001; Felberg, 2008). These moves often involve modality (e.g., "must", "should") and rhetorical questions or commands (e.g., "Sino ang binoto mo?", "Wake up!") to position the audience on one side of a moral divide.
  - **Mode:** Organizing engagement through thematic framing (e.g., "What they don't want you to know is. . . "), repetition (e.g., "not then, not now"), and multimodal cues such as hashtags, ALL CAPS, and emojis (e.g., "#NeverAgain", "TANGINA MO!") to amplify emotional impact and visibility (Clark-Parsons, 2021; Wang et al., 2016; Highfield, 2017).

Furthermore, our analysis shows that tenor—how language constructs social relationships and alignment—is a key mechanism through which misinformation becomes persuasive. Across posts and comments, we find that misinformation is rarely delivered in a neutral or objective tone. Instead, it is often embedded in alignment moves that seek to mobilize supporters, polarize audiences, or discredit opponents. For example, comments that amplify misleading claims frequently:

- Appeal to collective identity (e.g., "tayo mga Pilipino", "taumbayan"),
- Attack out-groups with derogatory labels (e.g., "dilawan", "DDS trolls"),
- Project certainty or moral obligation through modal verbs (e.g., "kailangan natin", "dapat gising na tayo"), and
- Use rhetorical questions or imperatives to provoke emotional engagement (e.g., "Seryoso ka ba dyan?", "Wake up!").

These strategies position the speaker and audience as part of a unified moral or political in-group, framing misinformation as shared knowledge or moral truth, rather than as contestable information. This may help explain if certain low-quality or misleading posts still generate high engagement—not because of their informational value, but because they offer identity-affirming narratives that emotionally and socially resonate with their intended audiences.

Table 4 summarizes these strategies with concrete linguistic indicators and examples from our dataset.

F Instructions for annotating political keywords

The following procedure was followed for annotator training and the final annotation of political keywords:

- **Training task:** Potential annotators were given a subset of the top 1000 keywords to label. They had to determine if a keyword was political and, if so, whether it related to Filipino politics, the ruling party, or referred to a person. Annotators who correctly labeled more than 80% of the training task keywords were approved for the main annotation task.
- **Main annotation task:** the training task helped shortlist 10 annotators down to 5. Approved annotators labeled all 1000 entries from the top keywords list. A majority vote approach coded in Python was used to resolve the annotator labels and identify relevant keywords. If more than 3 out of 5 annotators labeled the keyword as political, then the word was added to the final list of keywords.

G Instructions for annotating misinformation

1. Is this post misleading?
Answer: *Misleading* or *Not misleading*
2. Why is it misleading? (if applicable)

- Contains a factual error
- Contains a digitally altered photo or video
- Contains outdated information that may be misleading
- Misrepresents or misses important context
- Presents an unverified claim as fact
- Is a joke or satire that might be misinterpreted as fact

3. Potential Consequences if Believed

- No perceivable consequences
- Consequences to individuals
- Consequences to groups of people
- Consequences to society

4. Target of the Misinformation (if applicable)

- About the opposition
- About the incumbents
- About both
- Not applicable

1. Is this comment political?
Answer: *Yes (confident)*, *No*, *Not sure*
2. Is this comment uncivil?
Answer: *Yes (confident)*, *No*, *Not sure*
3. Why is it uncivil? (if applicable)

- Contains obscene language or vulgarity (e.g., "Tangina mo!")
- Contains name calling or insults (e.g., "Mga bobo")
- Contains extreme ideological language (e.g., "Mga libtard ka Leni Lugaw, Bobo kasi DDS")
- Contains stereotyping (e.g., "Tapon na yang mga bum-bay na yan")
- Contains exaggerated arguments (e.g., "Ang dali naman hindi magaral wag ka lang magdamit ng malaswa")

4. Is this comment misleading? (if applicable)

- Contains a factual error
- Contains outdated information that may be misleading
- Misrepresents or misses important context
- Presents an unverified claim as fact
- Is a joke or satire that might be misinterpreted as fact

H LLM Instructions for annotating discussion quality

Instruction:
You are a helpful annotator who is fluent in Filipino. This text is a comment as a reply to a post on a Facebook news page. Please classify this comment according to whether it is uncivil or not.
Guidelines for Uncivil Comments:

- **YES: Abuses and Sledging** — Uses ideological extremes like "yellowtard," "DDS," abusive terms like "bobo" or "tanga," or harmful stereotypes like "bakla," "palamunin," or "terorista."
- **YES: Threatening** — Threatens individual freedoms (e.g., "Mga tao tulad mo dapat patahimikin"), targets someone, or threatens democracy (e.g., "Bayan natin ang magpapatalsik sa mga buwayang ito").
- **YES: Exaggeration** — Uses exaggerated arguments (e.g., "Napakasimple lang, wag kang mag-anak kung hindi mo kayang buhayin!").
- **NO** — The comment is not particularly uncivil.

Notes:

- A comment can fall under multiple categories (e.g., abusive, threatening, and exaggerated).
- Respond with all applicable categories in this format: `"YES: Abuses and Sledging, YES: Threatening"`.
- If the comment is not uncivil, respond with `"NO"`.
- Only respond with the specified categories. Do not respond with anything else. Do not hallucinate.

Examples:
Example 1:
Comment: "Ang bobo mo naman, puro ka reklamo wala ka namang ginagawa para makatulong!"
Response: YES: Abuses and Sledging
Example 2:
Comment: "Mga katulad mo dapat ikulong para matutong sumunod sa gobyerno!"
Response: YES: Threatening
Example 3:
Comment: "Sa panahon ngayon, lahat ng mahirap ay tamad, kaya wag mag-anak kung di kaya buhayin!"
Response: YES: Exaggeration
Example 4:
Comment: "DDS ka pala, kaya hindi ka marunong mag-isip ng tama! Mga taong katulad mo dapat patahimikin."
Response: YES: Abuses and Sledging, YES: Threatening
Example 5:
Comment: "Mga buwaya sa gobyerno, dapat tanggalin at ipakulong lahat!"
Response: YES: Threatening
Example 6:
Comment: "Ang simpleng problema hindi nyo kayang solusyunan, kaya napapahamak ang bansa!"
Response: YES: Exaggeration

I Annotation validity

I.1 Human Annotation validity

Table 9: Inter-annotator reliability for post-level annotations

Category	Percentage Agreement
This post is misleading	0.92
This post contains a factual error	0.96
Contains digitally altered photo or video	0.99
Contains outdated information	0.99
Contains unverified claim presented as fact	0.94
Contains misrepresentation	0.93
Contains joke or satire that could be misrepresented	0.99
This post consequences	0.76
This post has misinformation	0.80

Table 10: Inter-annotator reliability for comment-level annotations

Category	Percentage Agreement
This comment is relevant to politics	0.84
This comment is uncivil	0.88
Contains obscene language/vulgarity	0.99
Contains insulting language/name calling	0.91
Contains ideologically extreme language	0.92
Contains stereotyping	0.97
Exaggerated argument	0.95
Comment is misleading	0.65

I.2 LLM Annotation validity

Table 11: LLM annotation accuracy across categories

Category	Accuracy
Positive	86%
Civil	93%
Uncivil: contains abuses and sledging	93%
Uncivil: threatening	98%
Uncivil: exaggeration	95%
Reciprocity	93%
No justification	95%
Contains fact-based justification	99%
Contains personal justification	96%
Not constructive	99%
Constructive: includes facts	100%
Constructive: presents a common ground	99%
Constructive: proposes solutions	99%