Hate Speech Campaigns in the 2016 Philippine Elections on Facebook

Sudhamshu Hosamane, Kiran Garimella

Rutgers University
sudhamshu.hosamane@rutgers.edu, kiran.garimella@rutgers.edu

Abstract

The paper presents a comprehensive analysis of hate speech and trolling campaigns on Facebook during the 2016 national elections in the Philippines. Employing a vast dataset of hundreds of millions of Facebook comments, we uncover the first empirical evidence of coordinated online hate speech campaigns in this political context. Our findings reveal that over 12% of comments on political pages contained hate speech, predominantly originating from supporters of then-candidate Rodrigo Duterte and his affiliates. We further examine the relationship between offline political events and online hate speech, finding a surge in hateful commenting following the launch of Duterte's campaign, though similar spikes were not observed after some of his later controversial remarks. Alarmingly, we observe a "spillover effect": regular social media users, after exposure to orchestrated hate speech by highly active "troll" accounts, began emulating these behaviors. This contagion effect highlights a worrying trend in which hate speech normalizes and spreads within online communities. Overall, our results shed light on the dynamics of digital political campaigns and their implications for democracy and public discourse. Given Facebook's ubiquity in the Philippines, these findings raise significant concerns about social media's influence on electoral politics and the health of online civic dialogue.

1 Introduction

The advent of social media has transformed the global political landscape, introducing new dynamics in how information is disseminated and public opinion is shaped. The Philippines, with its exceptionally high social media usage, serves as a critical case study in understanding these changes. Nearly the entire population is active on platforms like Facebook, making it an influential arena for political discourse. This paper examines the problem of orchestrated hate speech campaigns on social media, particularly during the 2016 election of Rodrigo Duterte.

President Rodrigo Duterte's campaign provides a stark illustration of how national leaders can harness social media to achieve political objectives. Maria Ressa, a prominent journalist and Nobel Peace Prize laureate, has detailed how both real and fake Facebook accounts were used in the Philippines to spread disinformation and manipulate public opinion under Duterte's regime (Borschmann 2017). This strategy effectively flooded the information space with falsehoods, distorting public understanding of facts. Ressa's observations underscore the need for interventions by tech companies to preserve factual integrity, especially during elections (Pazzanese 2021). This issue is particularly compelling due to the unprecedented scale and sophistication of these social media strategies. The Philippines' experience offers valuable insights into the broader implications of social media in politics, especially considering similar tactics were later observed in major political events globally, like Brexit and the US elections (Gorrell et al. 2019; Badawy, Ferrara, and Lerman 2018).

Identifying and analyzing hate speech in online discourse is inherently complex due to the nuanced and context-dependent nature of language, particularly more so across diverse cultural and linguistic settings (Paz, Montero-Díaz, and Moreno-Delgado 2020). This complexity is exacerbated by limited access to comprehensive data: social media platforms often restrict data sharing, making it difficult to gather sufficiently large and representative datasets for rigorous analysis (Lukito 2024). Additionally, there is a geographic imbalance in existing research, with limited studies focusing on regions like the Global South, including the Philippines (Badrinathan and Chauchard 2023; Septiandri, Constantinides, and Quercia 2024).

Previous research in this area has predominantly employed qualitative methods such as literature reviews (Cabañes and Cornelio 2017), interviews, and online participant observations (Ong and Cabañes 2018; Ragragio 2022). These studies have provided crucial insights, identifying paid troll farms engaged in strategic political manipulation (Ong and Cabañes 2018), documenting the optimization of Duterte's Facebook ecosystem for campaign messaging (Karunungan 2023), and highlighting anecdotal evidence of increased hate speech during elections (Montiel, Uyheng, and de Leon 2022). Yet, their qualitative focus and limited datasets do not quantify the broader scale, dynamics, and implications of these phenomena. Our study bridges this gap by presenting a comprehensive, large-scale quantitative analysis, assessing the prevalence, propagation, and spillover effects of hate speech and coordinated trolling during the 2016 Philippine elections. By doing so, we offer empirical evidence that complements existing qualitative in-

sights, deepening the understanding of digital political manipulation in this underrepresented context.

Our approach leverages millions of Facebook comments to quantitatively capture the scale and dynamics of hate speech and coordinated trolling. Methodologically, we developed a high-precision hate speech detection model specifically tailored for code-mixed Filipino-English content. Using this model, we assessed hate speech prevalence over time and identified coordinated harassment campaigns, particularly those orchestrated by pro-Duterte supporters. Our key findings include:

- On average, about 12% of comments on political posts were classified as hate speech. In absolute terms, this amounts to tens of millions of comments, indicating a substantial volume of hate-fueled discourse on Facebook.
- A significant portion of this hate speech originated from Duterte's supporters. Our analysis further reveals that these supporters engaged in coordinated campaigns, often involving the repetitive posting of identical messages containing hate speech.
- External events such as the beginning of Duetere's presidential campaign play a significant role in causing an increase in hate speech. However, the effects do not generalize to other offline events, such as Duterte's attacks on journalists or other female politicians. The data does not conclusively demonstrate a direct causal relationship between offline activity and online hate speech.
- We observed an interesting 'spillover effect,' where highly active trolls' comments tended to attract more hate speech. This indicates a contagion effect, suggesting that exposure to orchestrated hate speech can influence regular users' behavior on the platform.

Overall, our work contributes new insights into the intersection of digital politicking and democratic discourse, offering a robust, data-driven perspective on social media's role in political campaigns. For researchers, our results underscore the potential and effectiveness of computational models tailored to analyzing extensive code-mixed datasets, particularly in under-resourced linguistic contexts. This methodological advancement helps bridge the gap between technical capabilities and practical application in regions that remain underrepresented in existing research.

2 Relevant Literature

2.1 Background

Rodrigo Duterte became the 16th president of the Philippines after a highly divisive election in 2016. His critics included members of the liberal party such as Presidential candidate Mar A. Roxas, Vice President Leni Robredo,[1] and Leila De Lima, a senator. The election was heavily influenced by social media, which Duterte's campaign leveraged to bypass traditional media and speak directly to voters. Duterte's tough-talking, anti-establishment persona resonated with many citizens frustrated by crime and corruption. He cultivated a loyal online following of supporters who aggressively attacked dissenting voices and amplified his populist, often incendiary rhetoric. Critics have argued that Duterte's camp manipulated online networks to smear opponents and bolster his image (Juego 2017), exemplifying a modern authoritarian style that blends traditional politics with the expansive reach of social media.

[1]The Vice President can be from a different political party in the Philippines.

2.2 Hate Speech and Trolling in Elections

Hate speech and coordinated trolling have unfortunately become staples of recent elections worldwide. Researchers have developed methods to detect online coordinated inauthentic behavior by analyzing account metadata, linguistic patterns, and network connections (Stella, Ferrara, and Domenico 2018; Ratkiewicz et al. 2021; Keller et al. 2019). These studies show that many supposed grassroots movements are actually astroturfing operations–orchestrated campaigns using fake accounts and sometimes troll-like bots. The goals of such trolling campaigns include suppressing opposition voices, spreading disinformation, and amplifying extremist narratives (Cabañes and Cornelio 2017; Bradshaw and Howard 2018). There is evidence that political actors in various countries hire paid human trolls to operate fake profiles and coordinate attacks (Zannettou et al. 2019). Such tactics pose a serious threat to democratic discourse by creating a hostile information environment (Akhtar and Morrison 2019). As these tactics evolve, detecting and curbing them demands coordinated efforts from researchers, platforms, governments, and civil society.

Consistent with previous literature defining trolls as intentional disruptors of online communities (Donath 1998; Hardaker 2010; Binns 2012), we define *trolls* as active users who consistently post a high proportion of hateful speech, often within coordinated campaigns. We distinguish these from general *hate speech posters*, who may engage in hateful comments occasionally and without coordination. While many trolls employ hate speech, not all hate speech arises from organized trolling efforts.

2.3 Elite Cueing and Political Rhetoric

Elite cueing refers to the process by which public figures (elites) send signals that influence the attitudes and behaviors of their followers (McGraw 2003). Decades of political science research show that when a charismatic leader normalizes extreme views, supporters often follow suit, sometimes even radicalizing further in their expressions (Brass 2011; Tilly 2003). Duterte's presidency offers a case of this: he routinely used incendiary rhetoric–for example, urging violence against drug suspects–which appeared to embolden certain groups. Qualitative accounts suggest that Duterte's violent words empowered vigilantes and led police to take extreme actions in the drug war (Reyes 2016). This implies that when the head of state encourages hate or violence, it can legitimize such behavior among supporters

At the same time, elite cues can provoke a backlash from opponents. Studies on polarization show that when a polarizing figure makes extreme statements, those who oppose him may react with equal ferocity (Lupia 1995; Mondak 1994).

In the online context, Duterte's harsh rhetoric not only rallied his base but also galvanized critics who responded with their own hateful remarks directed at Duterte and his supporters. This dynamic can create a snowballing effect where both sides escalate in hostility, further deepening divisions.

A growing body of work suggests that online hate speech can spill over into real-world violence. For example, a recent study found that President Trump's anti-Muslim tweets correlated with subsequent spikes in anti-Muslim hate crimes in the U.S. (Müller and Schwarz 2023). Specifically, counties exposed to Trump's inflammatory tweets saw measurable increases in hate crimes in the days following those online outbursts. In Germany, a similar pattern has been documented: localities with surges of anti-refugee posts on Facebook experienced corresponding increases in violent attacks against refugees (Müller and Schwarz 2021). Conversely, online events can also be reactions to offline triggers. For instance, the onset of the COVID-19 pandemic in early 2020 sparked a wave of anti-Asian hate speech online, as extremist narratives sought to blame certain communities for the virus (Tahmasbi et al. 2021).

It is important to note that elite cueing does not operate in isolation. Whether such cues result in real-world harm depends on many factors: institutional checks, societal norms, and the media ecosystem. For example, one study found only mixed evidence that Trump's polarizing campaign rhetoric caused enduring spikes in hate speech on Twitter, as any surges tended to be temporary (Siegel et al. 2021). Similarly, in the Philippines, while Duterte's aggressive comments dominated the discourse, it has been unclear how much they directly increased online hate speech. Our study is the first to empirically test this in the Philippine context, by examining data for changes in hate speech volume around key moments like Duterte's campaign launch and his verbal attacks on specific targets.

2.4 Impact of Hate Speech and Trolling

Understanding the broader implications of hate speech and trolling is critical, particularly in politically charged online environments. Research indicates that hate speech is often self-reinforcing; initial hateful comments tend to attract more similar replies, perpetuating a cycle of negativity within discussions (Munger 2017; Mathew et al. 2020). Such behavior can normalize hostility, making it easier for others to participate in hateful interactions.

Furthermore, studies have highlighted a clustering effect, where hate speech tends to concentrate in specific threads or online communities. For instance, Mathew et al. (2019) demonstrated how hateful content frequently aggregates within particular online spaces, creating echo chambers that amplify negativity and hostility. This clustering not only exacerbates online divisions but also intensifies the overall prevalence of hostile discourse. Trolling and targeted hate campaigns also impose significant psychological and behavioral impacts. These campaigns often induce self-censorship among individuals, especially those from marginalized communities, who fear becoming targets of harassment themselves. Saha, Chandrasekharan, and De Choudhury (2019) and Phillips (2015) documented how organized trolling campaigns systematically silence dissenting or minority voices, significantly constraining open and healthy public discourse. Additionally, trolls can significantly influence the behavior of regular users not directly involved in hateful activities. Troll-driven interactions often alter online discourse's nature and tone, potentially normalizing aggressive or hostile communication styles among broader user groups. For example, Buckels, Trapnell, and Paulhus (2014) found that trolls, characterized by personality traits such as psychopathy, can indirectly affect the wider community by reshaping the informational landscape and influencing interactions among non-troll users. Given these complexities, it becomes crucial to explore how these theoretical insights translate empirically within specific political contexts, such as the 2016 Philippine elections.

3 Dataset

Our dataset consists of posts and comments from public Facebook pages and groups that were active around the time of the 2016 Philippine elections. Facebook pages serve as public broadcast channels, while groups are designed for group discussions.

The careful selection of the groups and pages and the data collection from those was done manually by journalists and editors at a top online Filipino news portal, Rappler[2], founded by Nobel peace prize winner Maria Ressa. Through their field work, journalists at Rappler identified an initial seed of 26 Facebook groups suspected to be operated by paid political operatives (troll networks) spanning different sides of the political spectrum. In the process of checking those paid troll groups, they identified 51 related groups with similar names. They then added more groups through monitoring the popular group links shared in these 77 groups. This iterative process eventually identified over 300 relevant groups, ensuring the representation of diverse political viewpoints. Pages were selected through a different but equally thorough approach. Rappler editors began with a curated list of the country's top news sites and expanded it to include pages frequently shared within the selected groups. Additional pages, including those identified as propagandist, were added based on their prominence in these groups. This resulted in a list of approximately 1,000 relevant pages.

This comprehensive, manually curated selection of groups and pages by editors and our research team ensured a balanced representation. Overall, the dataset contains 1,284 groups and pages,[3] which included around 5 million posts and 400 million comments on these posts using the Facebook Graph API. The data was actively collected from June 2015 to June 2018.[4] The dataset spans from August 2008

[2]Though Rappler has faced conflict with Duterte's government, including a 2018 shutdown order, MediaBiasFactCheck.com rates it as center-left and highly factual, with no false reports from 2017 to 2022

[3]For simplicity, for the rest of the paper, we refer to groups and pages as only 'pages'.

[4]The Facebook Graph API was functional until June 2018 and was shut down after the Cambridge Analytica scandal. See details here: https://techcrunch.com/2018/07/02/facebook-rolls-out-

to June 2018. For contextually relevant data, we focused on comments made between January 1, 2014, and March 1, 2018. This period was chosen because the study centers on the 2016 Filipino elections, held in May 2016. To capture a meaningful timeframe around this event, we included two years before and after the election. Comments from 2008 to 2013 totaled 0.8% of all comments, with an average hate speech proportion of just under 5.5%, while 90% of the total comments were posted between 2014 and 2018. The process unearthed how Duterte's campaign was exceptionally meticulous in ensuring that grassroots support was cultivated. They created hyperlocal 'chapters' of Facebook groups and pages and had a sophisticated top down operation (Ong and Cabañes 2018). For instance, our dataset included over 100 groups which just had a title of the format 'DUTERTE DIEHARD SUPPORTERS - [LOCATION]', where location could be various cities/towns in the Philippines). Some of these groups were seeded by content from paid trolls who would post content (Ong and Cabañes 2018) and some of them were organically formed by Duterte supporters. A complete list of the groups and pages analyzed in our study is available at https://bit.ly/philippines-pages-groups.

In assessing our dataset, it is important to transparently acknowledge potential limitations and sampling biases. Our methodology, while designed to achieve high precision, inherently does not guarantee representative coverage. Such limitations are common in social media studies, as true recall rates can only be accurately assessed by platform owners themselves (Olteanu et al. 2019). The selection of pages and groups was conducted through an expert-driven iterative process beginning with an initial focus on pro-Duterte pages, which were subsequently expanded via snowball sampling. Although this approach effectively captures significant activity related to Duterte's campaign, it likely results in an inherent bias toward more organized or vocal factions, particularly within the pro-Duterte sphere. Consequently, the dataset may underrepresent certain political viewpoints, and we acknowledge that it does not fully encompass the entire spectrum of political discourse active on Facebook during the 2016 Philippine elections. Despite these limitations, our dataset provides critical insights into the dynamics of online political discourse, especially regarding hate speech and coordinated trolling. Given Facebook's substantial role in shaping public discourse, the insights drawn from this dataset remain valuable. To our knowledge, no other publicly available resource provides comparable granularity and scale for analyzing social media's impact on political engagement during this pivotal electoral period.

3.1 Annotating Pages

We annotated a total of 1,284 Facebook pages and groups to determine their political affiliation. The URLs for these pages/groups were provided in an Excel sheet. Annotators visited each page/group using the provided URL and verified the political leaning based on specific guidelines described in the appendix (A.5). To determine the affiliation of each group/page, annotators were asked to first use the information from the title and description, and if that was insufficient, they were to rely on the post content and comments, both from 2016-2017.

Each page/group was coded using one of four values: **pro-duterte** (for pages/groups supporting ex-President Duterte), **anti-duterte** (for those opposing him), **neutral** (for pages/-groups not politically oriented or without clear support for any particular party), and **unknown** (if the affiliation was unclear or unknown). Pages with broken or unavailable links were coded as unknown as well.

We recruited four Filipino natives familiar with Filipino politics through the freelancing platform Upwork.com. In 2021, one annotator was employed for this task, and in 2024, three additional annotators completed the process independently. The inter-annotator agreement, measured by Fleiss' kappa, was 0.6865, indicating substantial agreement. Final labels were assigned based on majority vote. Annotators were asked to provide notes on each page/group and generate their own page/group category labels, as no predefined labels were given. For the 66 pages/groups without majority agreement, the authors manually reviewed the labels and assigned a final label based on additional information based on conversation with the annotators.

Our final affiliation labels showed that 62.5% (802 out of 1,284) of the pages were pro-Duterte, reflecting his substantial influence and support on social media. In contrast, 10.3% (132) were anti-Duterte, and 11.6% (149) were neutral. The affiliation of the remaining 15.6% of pages/groups could not be definitively determined.

In addition, the page category labels from the most informative annotator revealed that, 17.9% of all pages/groups were those that posted original content, 16.1% were fan pages/groups of popular politicians and their families and 13.1% were community based groups and pages. Details of the number of pages in each category as noted by this annotator is provided in Table 13.

3.2 Identification of User Support

We assigned political support to users based on the hashtags they used. We started by visualising word clouds of hashtags and recorded the most noticeable hashtags supporting or opposing one of the following national politicians or associations of interest (e.g. #isupportduterte, #notoduterte, #ihatedelima, etc). Using these manually curated hashtags as reference, we used the measure developed in (Garimella et al. 2018) to compute the similarity between two hashtags, which relies on co-occurring words and hashtags, to find hashtags that were commonly used along with the hashtags in the reference set. Hashtags that didn't explicitly support or oppose a person or a group were excluded from this analysis. We categorized a person as supporting or opposing one of the above-mentioned groups or people only if they used more than one hashtag from each group that implied support or opposition. Using this approach, we were able to identify 66,750 users (and their support for different political entities) of the 14,373,527 unique users in the dataset. Although this only covers 0.46% of all the users, they account for 10.3% of the total comments. For users with multiple affiliations, users were assigned to a single cate-

more-api-restrictions-and-shutdowns

gory based on the affiliation of the majority of the hashtags they used (see Table 3 for the full list). The order of the prevalence of hashtags with at least 100 users (Table 4) was used to resolve tie-breakers. Appendix tables 5 and 6 show the full list of hashtags we used. In table 4, we show that members affiliated with either Pro or Anti Duterte groups post substantial hashtags that belong to groups that align with their political leaning. To make our analysis simpler we grouped all the users with affiliations 'pro-duterte', 'anti-leni', 'pro-marcos', 'anti-delima', 'anti-roxas', and 'anti-rappler' as *pro-Duterte*. This gave us high confidence that the pro-Duterte group consisted only of Duterte's supporters and opponents of Rappler and the Liberal Party. We considered the other users as *anti-Duterte*. Pro-Duterte supporters accounted for over 85.3% of all of our final user labels.

While this hashtag-based approach prioritizes precision over recall, we are confident that it allowed us to identify highly likely supporters and opponents of key political actors. The manual curation of hashtags, the requirement of multiple signals per user, and a principled tie-breaking logic all contributed to the reliability of our classifications.

That said, we acknowledge that this one-dimensional classification into pro- and anti-Duterte categories cannot fully capture the complexity and nuance of political identities. Hashtags are often used for tactical, ironic, or oppositional purposes, and future work could build on interaction networks or discourse cues to refine and extend these labels.

4 Hate Speech Detection

Hate speech is a concept that is defined and interpreted differently across various communities, platforms, and regions. Different legal systems, online platforms, and research bodies offer distinct definitions based on cultural, social, and political contexts (Siegel 2020; Cohen-Almagor 2011). Online platforms have also developed their own definitions, which evolve over time to reflect their unique global user bases. In this study, we follow Facebook's June 2021 definition of hate speech, which identifies it as content that directly attacks individuals based on their "protected characteristics," such as race, ethnicity, national origin, religious affiliation, sexual orientation, gender identity, and other key traits (Meta 2021; Allan 2017). Based on this definition, our approach includes harassment, attack and toxic speech directed against any of the predefined protected characteristics as hate speech and users who post such content as hate-speech posters.

Hate speech detection has been a prominent area of research for several years (Davidson et al. 2017), with significant advancements achieved, especially in the context of the English language. The advent of large language models (LLMs) has markedly improved detection capabilities in English, as demonstrated in studies such as (Yin and Zubiaga 2021), which provided a review of existing approaches to automated hate speech detection. However, the scenario is notably different for non-English languages. While progress is being made, as evidenced by Aluru et al. (2020), who explored deep learning techniques for hate speech detection in non-English contexts, the availability of resources and research is still limited. The challenge becomes even more pronounced when dealing with code-mixed text, particularly in low-resource languages like Filipino. Code-mixing, the phenomenon where two or more languages are intermingled in communication, is common in multilingual societies but poses unique difficulties for hate speech detection.

Datasets. We started with a dataset from Cruz and Cheng (2020), features over 110,000 annotations for hate speech for approximately 11,000 tweets. However, an initial examination revealed a significant limitation: more than half of the dataset comprised tweets annotated by only a single user, and the quality of these annotations was not great, raising concerns about the reliability of these annotations. To enhance the robustness of our dataset, we implemented a rigorous filtering process. We eliminated all tweets with solitary annotations, opting to include only those with majority agreement among annotators. This approach, while enhancing data quality, reduced our dataset size substantially, from the original 11,000 to about 2,000 samples. Recognizing the potential for model over fitting due to this reduced dataset size, particularly when applying contemporary transformer models, we introduced an additional layer of annotation to expand the dataset.

We curated a more diverse dataset comprising 4,000 comments, stratified across four categories to ensure a broad representation of potential hate speech contexts. These categories included: a random sample of 1,000 comments; 1,000 comments with the highest like count; 1,000 comments sampled from users involved in coordinated posting activities (identified as detailed in Section 5.2); and 1,000 comments containing explicitly threatening keywords (e.g., 'rape', 'kill'). This stratification approach was designed to capture a wide spectrum of hate speech occurrences, thereby enhancing the representativeness of our training dataset. The annotation dataset included both hateful and non-hateful comments. 24.9% of our annotated dataset was hate speech. We recruited four native Filipino speakers as annotators for this task through the freelancer platform Upwork.com in 2021. These annotators were different from the ones recruited for page annotation. The annotators received initial examples and underwent extensive training provided by one of the authors through individual Skype sessions. This training involved iterative feedback and discussions based on small batches of annotations to ensure consistency and accuracy in the annotation process. The inter-annotator agreement (Fleiss Kappa) was 0.63, which is high for a task like hate speech detection (Del Vigna et al. 2017; Ousidhoum et al. 2019). The annotation process, both for original and pseudo-labels, is detailed in the Appendix (A.6). Each data point was labeled as either hate speech or not, based on the criteria outlined therein.

Models. We tested a variety of models for our hate speech classification. We began by establishing baseline performance using traditional machine learning techniques. Ensemble tree-based models, specifically XGBoost and Random Forest, in conjunction with TF-IDF vectorization, served as our starting point. These models yielded accuracy rates ranging from 64% to 71%.

Subsequently, we shifted our focus to more advanced methods, particularly the fine-tuning of transformer models, a standard approach for sequence-tagging tasks. Typically,

this involves training a transformer encoder with a classification head – a linear layer with dropout. Our initial attempt utilized a pre-trained BERT model fine-tuned for the Filipino language, as provided by Cruz and Cheng (2019). However, this model, trained on Wikipedia datasets, demonstrated poor zero-shot performance on our hate speech detection task, achieving only 67% accuracy. We attributed this to a mismatch between the nature of our code-mixed dataset and the predominantly Filipino text of the Wikipedia dataset.

We fine-tuned this Filipino BERT model on our combined dataset (2,000 tweets and 4,000 comments). The performance improved but remained suboptimal, with accuracy peaking at 78%. Analysis revealed a significant discrepancy in subword token distribution between our code-mixed corpus and the largely Filipino Wikipedia corpus. This highlighted the limitation of merely re-training the model without addressing the pre-trained tokenizer's inability to effectively segment our code-mixed text.

Our final refined pipeline included a RoBERTa model (Cruz and Cheng 2019) trained from scratch with a linear tuning head, pre-trained on a 30M random sample set of comments from our dataset and fine-tuned on the combined dataset of annotated data. To enhance the model's accuracy, particularly in reducing false positives, we reincorporated the TF-IDF driven Random Forest model. The lexical nature of this model, despite its marginally lower classification performance, proved adept at identifying key hateful tokens. This strategy retained most hateful comments while effectively filtering out false positives. Since our goal was to apply this classifier on the rest of our dataset, we aimed for high precision even while sacrificing on recall. This means that our estimates for hate speech prevalence are a lower bound of the amount of hate speech. Our best model obtains a 0.92 F1-score on a hold out set. Detailed evaluation metrics of our model are shown in the Appendix in Section A.1.

5 Analysis

In light of the growing concern over the misuse of comment sections for disseminating political propaganda and hate speech, as highlighted by Jeong, Kang, and Moon (2020), this study focuses on analyzing comment data.

5.1 Hate Speech Volume

The results of our analysis, as depicted in Figure 1, paint a striking picture of hate speech prevalence in the comments. The figure shows both the total count and the proportion of comments classified as hateful. Notably, the red line representing the actual count of hate speech comments reveals a staggering number, exceeding 100,000 daily during the election period, with an average of around 32,000 hateful comments per day. However, the raw count alone does not fully capture changes in prevalence. To account for any overall growth in commenting, we also examined the proportion of comments containing hate speech over time.

Our findings indicate that, on average, 11.8% of comments were hateful, with a marked increase following the commencement of the campaign and continuing into Duterte's presidency. This rate is alarmingly high, especially when compared to other platforms known for minimal content moderation. For instance, Mathew et al. (2019) found that less than 1% of the content on Gab, a platform with low moderation and a far-right user base, constituted hate speech. The prevalence of hate speech in our dataset is exceptionally high and unprecedented. The scale of our dataset indicates tens of millions of hateful comments, suggesting Facebook comments section had become a cesspool of hate.[5]

Interestingly, the proportion of hate speech, which hovered around 10% before the elections, surged significantly during the election period beginning in January 2016 and remained elevated thereafter. This sustained trend into Duterte's presidency, which commenced in June 2016, highlights a continuous and aggressive use of hate speech on social media. The persistent high levels of hate speech, emerging during the election period and continuing throughout Duterte's presidency, reveal a significant and concerning dynamic in online political discourse. This phenomenon suggests a state of perpetual conflict on social media, where the hate speech tactics employed during the electoral campaign were not only sustained but possibly intensified during Duterte's tenure as president.

This continuation suggests a strategic and deliberate use of hate speech as a tool for political influence and control, extending beyond the confines of electioneering into the day-to-day governance and political discourse (Ragragio 2022). The use of online platforms for spreading hate speech and propaganda has been a tactic observed in various political contexts globally. In the case of Duterte's presidency, it seems these digital strategies were not just confined to garnering support during elections but became a characteristic feature of the political landscape under his administration.

This sustained use of hate speech in the digital public sphere raises critical concerns about the long-term impacts on democratic discourse, social harmony, and the normalization of aggressive political rhetoric. It underscores the need for more robust mechanisms to counteract the spread of hate speech and highlights the vital role of digital literacy and critical media consumption in modern democracies.

5.2 Who Is Posting the Hate Speech?

For this analysis we examined the commenting behaviour of pro- and anti-Duterte supporters (identified in Section 3.2).

At first glance the two camps look symmetrical: the share of hateful comments *per identified user* is similar across groups, and although pro- and anti-Duterte supporters make up only 0.40 % and 0.056 % of the overall user base, they account for 8 % and 0.9 % of all hateful comments, respectively.

[5] Given these exceptionally high numbers, we wanted to be sure that our classifier is doing a good job on detecting hate speech. To validate the accuracy of our hate speech detection model, we manually reviewed a sample of 1,000 comments that the model had classified as hateful. In this hand-coding process, we found that the model correctly identified hate speech in these comments with an accuracy of approximately 93%. This high accuracy on a hand-coded sample provides reassurance that our model is successfully identifying hateful content within our dataset.

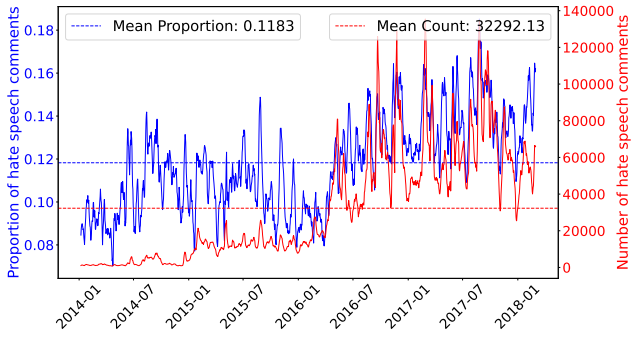

Figure 1: Trends in the total volume (red) and proportion (blue) of hateful comments in our dataset. The lines show a 7-day binned average.

A closer look at Figure 2 compares the *volume* of comments per supporter. For non-hate content the two CDFs almost overlap; the large sample nonetheless produces a statistically significant KS statistic ($D = 0.038,\ p = 2.7\times10^{-9}$). In practical terms the shift is modest: the medians differ by 21 posts and the 95th-percentile counts are nearly identical. The picture changes for hateful content. The median pro-Duterte supporter has authored *23* hateful comments versus *16* for an anti-Duterte supporter, and 20.5% of pro-Duterte supporters exceed 100 hateful comments compared with 16.9% of their counterparts. The hateful-comments distributions differ more strongly ($D = 0.072,\ p = 1.3 \times 10^{-27}$), indicating a heavier and more concentrated reservoir of hate among pro-Duterte supporters. Thus, while both camps produce hateful speech, the pro-Duterte side does so *more frequently and at higher volumes*.

Our third key finding concerns the *fraction* of each supporter's comments that are hateful. A Welch's *t*-test (Welch 1947) shows a significant difference in the mean per-user hate fraction (Figure 3): on average, an individual pro-Duterte supporter posts almost twice as much hateful speech as an anti-Duterte supporter.

Coordinated Posting We next examine the role of coordinated posting on social media—a phenomenon well documented in prior qualitative work highlighting the presence of troll farms and paid online operatives (Ong and Cabañes 2018). To detect coordinated behavior at scale, we applied Locality Sensitive Hashing (LSH) (Gionis, Indyk, and Motwani 1999), a technique designed to efficiently identify near-duplicate messages.

Our analysis surfaced extensive coordination: we identified 5,673 clusters where a near-same message was reposted more than 50 times. Some campaigns were massive in scope, with the largest exceeding 7,000 posts. A detailed breakdown of these campaigns appears in Figure 9 (Appendix).

Political affiliation analysis reveals an asymmetry. Pro-Duterte supporters were involved in 46.7% of all the identified coordinated campaigns, compared to 10.1% for anti-Duterte supporters. Moreover, 20.5% of all coordinated campaigns involved hate speech. Of these, 42% were tied to pro-Duterte users, while only 9.9% were associated with anti-Duterte groups. These figures suggest a disproportionate use of coordinated activity by pro-Duterte networks to amplify hateful content.

Campaign size also varied by group: as shown in Figure 8 (Appendix, Section A.2), coordinated efforts involving pro-Duterte supporters were, on average, substantially larger than those from the opposing side.

These patterns align with previous reports of centralized, top-down digital operations (Ong and Cabañes 2018), potentially involving professional trolls or hyper-partisan volunteers. While coordinated messaging is not unique to the Philippines, its scale during the 2015–2016 period was unprecedented—and likely instrumental in shaping public opinion in an environment where social media served as a key battleground for political influence.

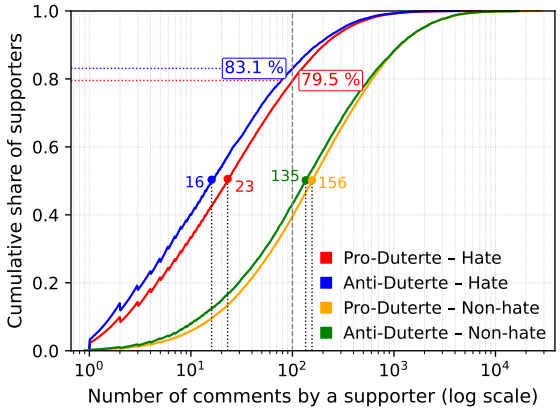

Figure 2: Cumulative distribution of per-supporter commenting activity, comparing pro- and anti-Duterte supporters. Each curve shows the share of supporters who authored up to x comments. The dotted lines highlight key statistics discussed in the text

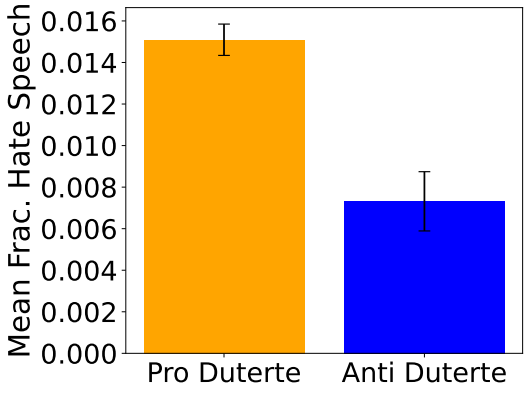

Figure 3: Proportion of hateful comments posted by pro- and anti-Duterte supporters. Roughly 1.5% of the comments by pro-Duterte supporters were hateful, whereas, for anti-Duterte, it was significantly less. Error bars show 95% confidence intervals. The difference is statistically significant ($p < 0.001$).

5.3 Where Is the Hate Speech Being Posted?

Next, we focused on *where* the hate speech campaigns were being posted, specifically looking at the leaning of the pages (from Section 3.1). Figures 4 and 5, show the results. Firstly, perhaps surprisingly, we observed that the majority of hate speech by each group was concentrated on pages aligned with their respective political affiliations. This phenomenon points to a pronounced echo chamber effect, where individuals primarily interact with content and communities that reinforce their existing beliefs. This insularity results in minimal cross-party information exchange and contributes to the intensification of partisan views. Secondly, as observed previously, the figures show the volume of hate speech posted from anti-Duterte supporters was significantly (almost an order of magnitude) lower compared to that of pro-Duterte supporters. Thirdly, a considerable portion of hate speech, nearly 20%, was directed at neutral pages, such as news portals. This consistent targeting of neutral platforms by both pro and anti-Duterte supporters indicates a strategic use of hate speech to influence or disrupt broader public discourse. Finally, an interesting temporal pattern emerged as well: the Duterte campaign's hate speech significantly increased following the commencement of their campaign. While the volume of hate speech on anti-Duterte pages remained relatively stable, the proportion of hate speech on pro-Duterte pages showed a steady increase. This contrast in the trajectory of hate speech output between the two groups provides insights into how political campaigns can influence online behavior. Our qualitative examination of the creation dates of several of these pages revealed that most had been established well before Duterte's presidency, with origins tracing back to at least 2014, when Duterte was still a city mayor. This indicates that the platforms for these online activities were in place long before the height of the political campaigns. It is important to note a limitation of this analysis: we did not examine the stance of the hateful comments or whether they were responses to opposing views. This presents an opportunity for future research to explore the dynamics of hate speech interactions in more detail.

5.4 Evidence of Elite Cueing

Anecdotal evidence in the press about the increase in hate crimes and hate speech coinciding with Duterte's political ascent suggests that his influence may have contributed to normalizing extremist dialogue (Montiel, Uyheng, and de Leon 2022).

In this section, we aim to determine whether Duterte's election campaign and subsequent verbal rhetoric have influenced the behavior of politically engaged Filipino Facebook users. Specifically, we are interested in whether interventions such as the kickoff of Duterte's campaign or his attacks on journalists and female politicians have had a significant *causal* impact on the increase in online hate speech.

Given the challenge of isolating the individual effects of Duterte's online and offline campaigns, as well as the influence of his followers and opponents on the rise in hate speech, we focus on measuring the macroscopic changes in the daily proportion of hateful Facebook comments. By conditioning on specific interventions, we analyze these changes across all political groups and pages from which we have collected comments.

To achieve this goal, we model proportion of daily hateful comments using a segmented regression model—Interrupted Time Series Analysis (ITSA)—as described in Bernal, Cummins, and Gasparrini (2016). Segmented regression refers to a model with different intercept and slope coefficients for the pre- and post-intervention time periods. We follow a similar analysis to that of Siegel et al. (2021) to model the effects of elite cueing on the proportion of hateful comments.

At the outset of Duterte's official campaign in February 2016, we conducted an ITSA spanning from January 2014 to March 2018. The analysis revealed a significant immediate increase in the proportion of hate speech, as illustrated in Figure 6. To ensure robustness, we also verified that both a quadratic ITSA model (Figure 10) and a first-order autoregressive ITSA model (Figure 11) showed significant results for the immediate increase in hate speech. The full ITSA coefficients are tabulated in Table 7 in the Appendix. Apart from showing an abrupt increase in hate speech, the findings also suggest that this upward trend continued after the campaign and persisted into Duterte's term.

We also looked at whether this overall effect applies to something specific, like Duterte's personal attacks on female politicians or journalists but do not find any significant effects. Refer to Section A.4 in the Appendix for more examples of ITSA analysis and effect sizes, and see Figures 12 and 13. As mentioned in section 2.3, evidence of elite cueing is mixed. A lack of a significant result is likely because of the constant nature of the attacks leading the pre and post treatment periods to be too narrow to show an effect.

5.5 Analyzing Potential Spillovers of Hate Speech

In this section, we explore the relationship between commenting activity and the volume of hateful (sub)comments it attracts, along with identifying the sources of these comments. Our goal is to understand the potential "spillover" effect of trolling activities, examining whether hateful trolls incite further hateful responses from both trolls and non-trolls alike.

The core of our inquiry revolves around two hypotheses. First, we question whether hate speech posted by popular trolls leads to a higher volume of hate speech in responses, particularly by non-troll users. This would indicate a spillover effect where the aggressive or hateful tone set by trolls catalyzes similar behavior in other users' replies. Second, we explore the possibility that hate speech from these popular trolls attracts more responses from other popular trolls, thereby creating a concentrated network of hate speech propagation.

Our hypothesis posits that hate speech posted by popular trolls may lead to an increase in hate speech responses and potentially attract other popular trolls, thereby affecting the overall distribution of replies within the network. To examine these dynamics, we identified popular comment threads in our dataset and distinguished users who exhibited troll-like and non-troll-like behavior.

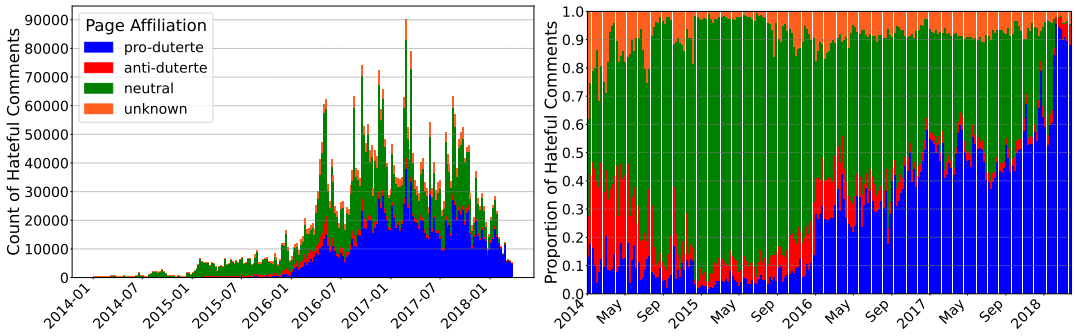

Figure 4: (a). Counts of hateful comments made by pro-Duterte supporters. (b). shows the proportion. Both plots show 7-day binned averages.

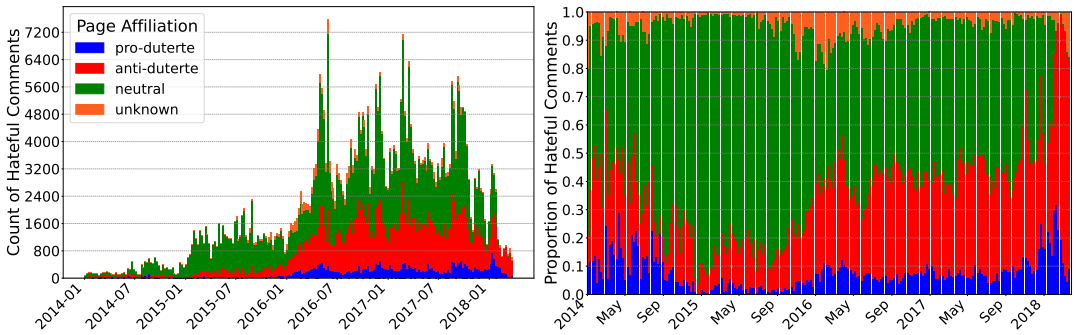

Figure 5: (a). Counts of hateful comments made by anti-Duterte supporters. (b). shows the proportion.

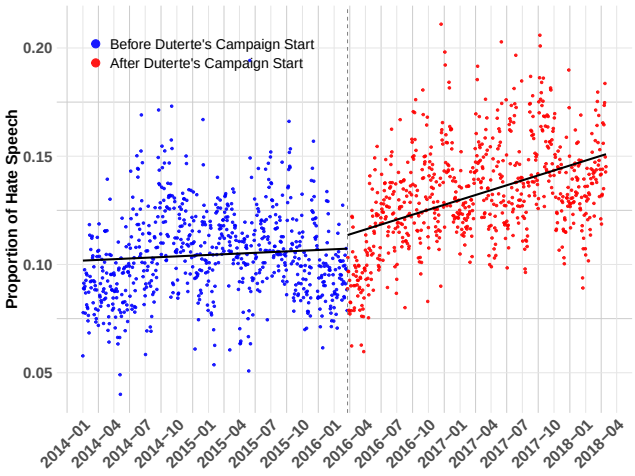

Figure 6: Interrupted time series analysis of proportion of hateful speech from the announcement of Duterte's election campaign in February 2016.

To identify the thread-like structure among comments, we applied the depth-first search algorithm to all comments that were direct replies to a post (i.e., those without a parent comment). Using these comments as root nodes, we traced all possible paths to the leaf nodes, organizing them as independent threads. We confirmed that all identified threads maintained a two-level hierarchy, consistent with the visual structure of comments on Facebook. To test our hypotheses, we compared the means of the proportions of hateful comments in threads initiated by trolls versus those started by non-trolls. To avoid biasing the proportion of hateful comments, we selected a subset of threads with more than 4 comments, focusing on the top 5% of threads with the most subcomments. The largest thread in our dataset contained 2,511 subcomments.

We define troll-like users (trolls) as those who are highly active (comment frequently) and have a higher proportion of hateful comments. Specifically, we identified users who made more than 10 comments between 2014 and 2018 (top 20% of all commenters) and who have a hate comment proportion exceeding 25% (top 10% of hateful commenters by proportion among users who have commented more than 10 times). Similarly, we define non-trolls as users whose hateful comments constitute less than 10% of their total comments, regardless of the total number of comments. We excluded threads started by users who did not fit into either group. Using these definitions, we identified 352K users exhibiting troll-like behavior and 11.9M non-troll users. We also confirmed that there was no overlap between the two groups.

We recognized that dependencies within comments of a single thread could affect the calculation of proportions within that thread (clustering behavior). Additionally, the posts and pages on which the comments are made might impact the independence assumption of the threads, crucial for conducting significance tests. To mitigate these dependencies, we performed a permutation test (using 10000

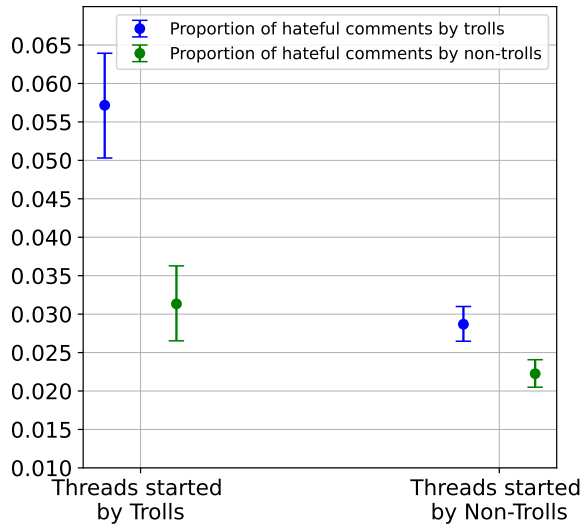

Figure 7: Mean fraction of hate speech in threads started by Trolls and Non-Trolls. Error bars show 95% bootstrap confidence intervals.

permutations) to assess the significance of the difference in the means of the proportion of hateful comments between threads started by trolls and those started by non-trolls.

The results of our study, as depicted in Figure 7, show a statistically significant difference in the average proportion of hate speech in replies between threads started by trolls and non-trolls. Threads started by trolls attract hateful comments from other trolls 2.85% more frequently than threads started by non-trolls ($p < 10^{-10}$). Additionally, threads started by trolls attract 0.91% more hateful comments from non-trolls compared to threads started by non-trolls ($p = 0.0002$).

Our findings reveal an intriguing pattern: threads started by trolls tend to have a higher proportion of hateful comments from both trolls and non-trolls. This occurs even though there is no significant difference in the proportion of hateful subcomments made by trolls (permutation-test $p = 0.87$) or non-trolls (permutation-test $p = 0.048$) when responding to a hateful comment by a troll compared to a non-troll. Moreover, threads with any initial hateful comment (regardless of the commenter) already attract significantly (permutation-test $p < 10^{-10}$) more hateful subcomments, compared to threads started by a non-hateful comment. These findings indicate that threads initiated by trolls inherently foster a greater volume of hate speech, suggesting the presence of nuanced linguistic or contextual triggers that our classifier does not fully capture. Importantly, the results are not sensitive to the specific threshold values used to define trolls and non-trolls: we tested a range of alternative thresholds, and the observed spillover effects remained consistent. This underscores the unique and potent role trolls play in shaping the tone of online discourse and amplifying hate speech on social media platforms.

6 Conclusion

This study provides a large-scale analysis of political discourse on social media in the Philippines, addressing the growing role of these platforms in shaping political landscapes

We acknowledge limitations in our study. Our analysis focuses exclusively on Facebook comments, potentially missing relevant dynamics occurring on other platforms. Additionally, the detection of coordinated campaigns relies on algorithmic patterns, which may not fully capture subtle nuances in human communication.

Our research contribution goes beyond developing a high-precision hate speech detection model for code-mixed Filipino text. This study addresses a notable gap in the field—the limited availability of descriptive analytics in politically and culturally complex regions like the Philippines. It also offers a framework that can be used for similar studies in other parts of the global south, where such detailed analyses are less common. Our findings provide a valuable dataset and a methodology that can be built upon and expanded in future research. This research opens avenues for designing targeted and scalable interventions to mitigate hate speech and online manipulation and methods to understand causal impacts of issues like hate speech on election outcomes.

The analysis of Facebook comments, which are now difficult to access, underscores the challenges researchers face in studying the dynamics of online platforms that can foster hate speech, particularly in political contexts. The fact that such data is no longer easily accessible does not imply that the issue of hate speech has lessened; rather, it highlights the persistent and pervasive nature of the problem, even if it is now less visible. In response, our team made significant efforts to compile a dataset that captures the scale and intensity of hate speech during the 2016 elections. While our dataset is comprehensive, we recognize that it may not fully represent all political viewpoints in Filipino politics. Nonetheless, it offers valuable insights into the dynamics and prevalence of hate speech on Facebook and emphasizes the ongoing need for sustained access to comprehensive data for academic research. Additionally, we introduce a novel approach to accurately identify user political leanings.

The ethical implications of collecting and analyzing a dataset that cannot be reproduced due to changes in Facebook's terms of service pose a significant dilemma. Is it ethical to work with data that is obtained under such restrictive conditions? We argue that it is not only ethical but necessary. Such data are crucial for informing regulatory actions, such as those proposed under Europe's Digital Services Act (DSA), which aim to enhance transparency and accountability of major tech platforms. By making such datasets available to researchers, regulators can better understand the prevalence and impact of hate speech, thereby applying pressure on platforms like Facebook to take more robust actions. Furthermore, this approach challenges the platforms' possible use of privacy as a veneer to restrict data access, urging them to balance user privacy with the necessity of transparency to combat hate speech effectively. This expanded access could lead to more informed policymaking and improved platform governance, ultimately contributing

to a healthier online discourse environment.

The study of the 2016 Philippine elections, although nearly a decade old, remains profoundly relevant today. It offers a historical perspective that is crucial for understanding the evolution of digital political campaigns and predicting the adoption of similar tactics in various global contexts. This is especially important in places like the Philippines, where the omnipresence of social media significantly impacts political dynamics. Despite the passage of time, descriptive analyses of such events are scarce, particularly in the Global South, making this study a valuable resource. Moreover, the insights gained from examining these elections are invaluable to academic communities such ICWSM, where research focusing on the Global South is often underrepresented. By providing detailed accounts of the strategies used, this work not only enhances our understanding of digital campaigning but also serves as a historical record that documents these evolving tactics.

Additionally, this study addresses the effects of hate speech and online trolling, which are poorly understood but critically important. Issues like the persistence of hate speech in political campaigns and the spillover effects of online behaviors remain prevalent. By examining these phenomena, the research contributes to a broader understanding of digital political strategies, equipping stakeholders to manage and counteract negative campaigning more effectively.

References

Akhtar, S.; and Morrison, C. M. 2019. The prevalence and impact of online trolling of UK members of parliament. *Computers in Human Behavior*, 99: 322–327.

Allan, R. 2017. Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community? https://about.fb.com/news/2017/06/hard-questions-hate-speech/. Accessed: 2021-07-15.

Aluru, S. S.; Mathew, B.; Saha, P.; and Mukherjee, A. 2020. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*.

Badawy, A.; Ferrara, E.; and Lerman, K. 2018. Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 258–265.

Badrinathan, S.; and Chauchard, S. 2023. Researching and countering misinformation in the Global South. *Current Opinion in Psychology*, 101733.

Bernal, J. L.; Cummins, S.; and Gasparrini, A. 2016. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International Journal of Epidemiology*, 46(1): 348–355.

Binns, A. 2012. DON'T FEED THE TROLLS! Managing Troublemakers in Magazines' Online Communities. *Journalism Practice*, 6(4): 547–562.

Borschmann, G. 2017. Philippines government propaganda war on the internet: journalist. https://www.abc.net.au/listen/programs/radionational-breakfast/philippines-government-propaganda-war-on-the-internet/8855088. Radio National Breakfast, ABC Radio National.

Bradshaw, S.; and Howard, P. N. 2018. Challenging truth and trust: A global inventory of organized social media manipulation. *The computational propaganda project*, 1: 1–26.

Brass, P. R. 2011. *The production of Hindu-Muslim violence in contemporary India*. University of Washington Press.

Buckels, E. E.; Trapnell, P. D.; and Paulhus, D. L. 2014. Trolls just want to have fun. *Personality and individual Differences*, 67: 97–102.

Cabañes, J.; and Cornelio, J. 2017. The rise of trolls in the Philippines (and what we can do about it). *A Duterte reader: Critical essays on the early presidency of Rodrigo Duterte*.

Cohen-Almagor, R. 2011. Fighting Hate and Bigotry on the Internet. *Policy and Internet*, 3(3): 1–26.

Cruz, J. C. B.; and Cheng, C. 2019. Evaluating Language Model Finetuning Techniques for Low-resource Languages. *arXiv preprint arXiv:1907.00409*.

Cruz, J. C. B.; and Cheng, C. 2020. Establishing Baselines for Text Classification in Low-Resource Languages. *arXiv preprint arXiv:2005.02068*.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, 512–515.

Del Vigna, F.; Cimino, A.; Dell'Orletta, F.; Petrocchi, M.; and Tesconi, M. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first Italian conference on cybersecurity (ITASEC17)*, 86–95.

Donath, J. S. 1998. Identity and Deception in the Virtual Community. In *Communities in Cyberspace*, 27–57. Routledge, 1st edition edition. ISBN 9780203194959.

Garimella, K.; Morales, G. D. F.; Gionis, A.; and Mathioudakis, M. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1): 1–27.

Gionis, A.; Indyk, P.; and Motwani, R. 1999. Similarity Search in High Dimensions via Hashing. In *VLDB*.

Gorrell, G.; Bakir, M. E.; Roberts, I.; Greenwood, M. A.; Iavarone, B.; and Bontcheva, K. 2019. Partisanship, Propaganda and Post-Truth Politics: Quantifying Impact in Online Debate. *The Journal of Web Science*, 7.

Hardaker, C. 2010. Trolling in Asynchronous Computer-Mediated Communication: From User Discussions to Academic Definitions. *Journal of Politeness Research*, 6(2): 215–242.

Jeong, J.; Kang, J.-h.; and Moon, S. 2020. Identifying and quantifying coordinated manipulation of upvotes and downvotes in Naver News comments. In *ICWSM*, volume 14.

Juego, B. 2017. The Philippines 2017: Duterte-led authoritarian populism and its liberal-democratic roots. *Asia Maior*, XXVIII: 129–164.

Karunungan, R. 2023. *The role of Facebook influencers in shaping the narrative of the Duterte era*. Ph.D. thesis, Loughborough University.

Keller, F. B.; Schoch, D.; Stier, S.; and Yang, J. 2019. Political Astroturfing on Twitter: How to Coordinate a Disinformation Campaign. *Political Communication*, 37: 256 – 280.

Lukito, J. 2024. Platform research ethics for academic research. *Center for media engagement. https://mediaengagement. org/research/platform-research-ethics/. Accessed*, 30.

Lupia, A. 1995. Who Can Persuade?: A Formal Theory, A Survey and Implications for Democracy. Prepared for the Annual Meetings of the Midwest Political Science Association, Chicago, IL, April 6–8.

Mathew, B.; Dutt, R.; Goyal, P.; and Mukherjee, A. 2019. Spread of hate speech in online social media. In *WebScience*.

Mathew, B.; Illendula, A.; Saha, P.; Sarkar, S.; Goyal, P.; and Mukherjee, A. 2020. Hate begets Hate: A Temporal Study of Hate Speech. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2).

McGraw, K. M. 2003. Political impressions: Formation and management.

Meta. 2021. Community Standards: Hate Speech. https://transparency.meta.com/policies/community-standards/hate-speech/. Revision notes, June 23, 2021.

Mondak, J. J. 1994. Question wording and mass policy preferences: The comparative impact of substantive information and peripheral cues. *Political Communication*, 11(2).

Montiel, C. J.; Uyheng, J.; and de Leon, N. 2022. Presidential Profanity in Duterte's Philippines: How Swearing Discursively Constructs a Populist Regime. *Journal of Language and Social Psychology*, 41(4): 428–449.

Müller, K.; and Schwarz, C. 2023. From Hashtag to Hate Crime: Twitter and Antiminority Sentiment. *American Economic Journal: Applied Economics*, 15(3): 270–312.

Munger, K. 2017. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*.

Müller, K.; and Schwarz, C. 2021. Fanning the Flames of Hate: Social Media and Hate Crime. *Journal of the European Economic Association*, 19(4): 2131–2167.

Olteanu, A.; Castillo, C.; Diaz, F.; and Kıcıman, E. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, 2: 13.

Ong, J. C.; and Cabañes, J. V. A. 2018. Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the Philippines. *Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the Philippines*.

Ousidhoum, N.; Lin, Z.; Zhang, H.; Song, Y.; and Yeung, D.-Y. 2019. Multilingual and Multi-Aspect Hate Speech Analysis. In *EMNLP*, 4675–4684.

Paz, M. A.; Montero-Díaz, J.; and Moreno-Delgado, A. 2020. Hate speech: A systematized review. *Sage Open*, 10(4): 2158244020973022.

Pazzanese, C. 2021. Maria Ressa warns of authoritarians, social media, disinformation. https://news.harvard.edu/gazette/story/2021/11/maria-ressa-warns-of-authoritarians-social-media-disinformation/. Harvard Gazette.

Phillips, W. 2015. *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. Mit Press.

Ragragio, J. L. D. 2022. Facebook populism: mediatized narratives of exclusionary nationalism in the Philippines. *Asian Journal of Communication*, 32(3): 234–250.

Ratkiewicz, J.; Conover, M.; Meiss, M.; Goncalves, B.; Flammini, A.; and Menczer, F. 2021. Detecting and Tracking Political Abuse in Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1): 297–304.

Reyes, D. A. 2016. The Spectacle of Violence in Duterte's "War on Drugs". *Journal of Current Southeast Asian Affairs*, 35: 111 – 137.

Saha, K.; Chandrasekharan, E.; and De Choudhury, M. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM conference on web science*, 255–264.

Septiandri, A. A.; Constantinides, M.; and Quercia, D. 2024. How Western, Educated, Industrialized, Rich, and Democratic is Social Computing Research? *arXiv preprint arXiv:2406.02090*.

Siegel, A. A. 2020. Online Hate Speech. In Persily, N.; and Tucker, J. A., eds., *Social Media and Democracy*, 56–88. Cambridge: Cambridge University Press.

Siegel, A. A.; Nikitin, E.; Barberá, P.; Sterling, J.; Pullen, B.; Bonneau, R.; Nagler, J.; Tucker, J. A.; et al. 2021. Trumping hate on Twitter? Online hate speech in the 2016 US election campaign and its aftermath. *Quarterly Journal of Political Science*, 16(1): 71–104.

Stella, M.; Ferrara, E.; and Domenico, M. D. 2018. Bots sustain and inflate striking opposition in online social systems. *CoRR*, abs/1802.07292.

Tahmasbi, F.; Schild, L.; Ling, C.; Blackburn, J.; Stringhini, G.; Zhang, Y.; and Zannettou, S. 2021. "Go eat a bat, Chang!": On the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19. In *Proceedings of the web conference 2021*, 1122–1133.

Tilly, C. 2003. *The politics of collective violence*. Cambridge University Press.

Welch, B. L. 1947. The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, 34(1/2): 28–35.

Yin, W.; and Zubiaga, A. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7: e598.

Zannettou, S.; Caulfield, T.; Setzer, W.; Sirivianos, M.; Stringhini, G.; and Blackburn, J. 2019. Who Let The Trolls Out? Towards Understanding State-Sponsored Trolls. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, 353–362. New York, NY, USA: Association for Computing Machinery. ISBN 9781450362023.

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes.
   (e) Did you describe the limitations of your work? Yes, in the relevant sections.
   (f) Did you discuss any potential negative societal impacts of your work? Yes.
   (g) Did you discuss any potential misuse of your work? No.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes
   (b) Have you provided justifications for all theoretical results? Yes
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? No
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes
   (e) Did you address potential biases or limitations in your theoretical framework? Yes
   (f) Have you related your theoretical results to the existing literature in social science? Yes
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? No
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? Yes
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? Yes
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? No
   (d) Did you discuss how data is stored, shared, and deidentified? Yes

A Appendix

A.1 Hate Speech Detection Performance

The performance of our final ensemble model is shown in Table 1.

Table 1: Accuracy of the best hate speech classifier

Label	Precision	Recall	F1-Score
Hate	0.92	0.93	0.93
Not hate	0.93	0.92	0.92
Overall Metrics			
Accuracy		0.93	
Macro Avg	0.93	0.92	0.92
Weighted Avg	0.93	0.93	0.92

A.2 Coordinated Posting

More information on coordinated posting can be found in Figures 8, and 9.

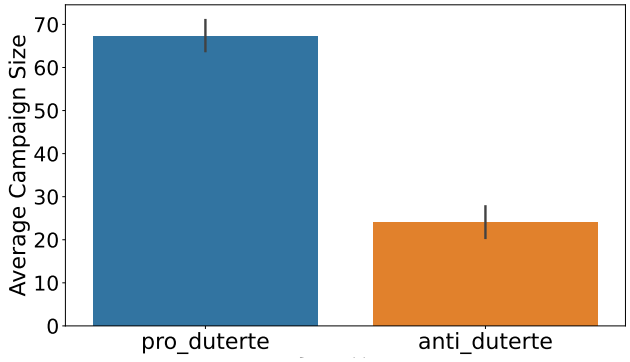

Figure 8: Average coordinated campaign size for pro and anti duterte supporters. Error bars indicate 95% confidence intervals.

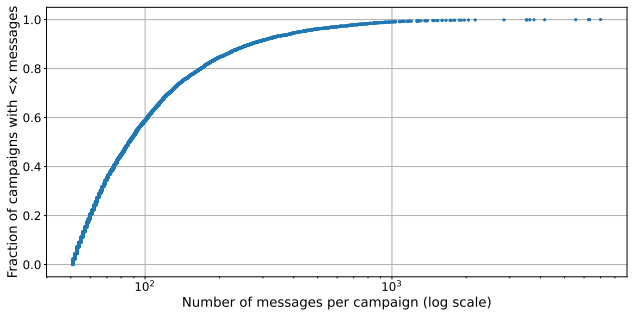

Figure 9: Coordinated campaigns size. Over 60% of the 5700 campaigns are less than a hundred messages but there are some massive campaigns with over 7000 messages.

A.3 Pro and Anti Duterte Supporters

As detailed in Section 3.2, we curated hand made list of hashtags to identify who users support. The details of the user leaning assignment are shown in Table 3. The exact hashtags used are shown in Tables 5 and 6. We note that we did not integrate native Filipino speakers' knowledge in identifying affiliations for commonly occurring Tagalog hashtags. Instead, we relied on Google Translate to understand their meanings and verified the context of extremely popular hashtags through news coverage. For other hashtags, we manually checked the context on social media and used Google Translate for the main posts or comments in Tagalog to ensure they weren't used in support of the opposing group. This approach may have limitations in accurately capturing the nuances of the local language and context.

Table 2: Details of coordinated posting. We can see that the largest campaign involved 3300 users posting on 113 pages.

	#users	#pages	#posts
count	5673.000000	5673.000000	5673.000000
mean	22.250308	8.902168	108.116164
std	86.628386	10.054372	135.500591
min	0.000000	1.000000	0.000000
25%	1.000000	3.000000	52.000000
50%	2.000000	6.000000	69.000000
75%	15.000000	11.000000	115.000000
max	3353.000000	113.000000	2005.000000

The heuristics used to identify trolls and non-trolls might not be perfectly accurate. Our definitions and thresholds for classifying users may overlook nuances in user behavior, suggesting the need for more sophisticated methods to distinguish between different types of online actors.

We acknowledge that many comments could be attributed to bots. However, identifying whether a comment was made by a human or a bot was not the focus of this paper. If a bot is identified with a strong political leaning (as discussed in section 3.2), it could be used to attack opponents, promote the party's agenda, and gain support. Our main concern is the impact on the general public—whether comments from bots or humans influence regular users. Including all commenting agents in our study is valid, as all content is visible to the public.

Table 3: Affiliation and Number of Users

Affiliation	Number of Users
Pro-Duterte	44279
Anti-Leni	11506
Pro-Santiago	2939
Pro-Leni	2367
Pro-Marcos	1953
Default	1078
Anti-Duterte	820
Anti-Delima	764
Pro-Delima	311
Pro-Rappler	305
Anti-Marcos	230
Anti-Roxas	198

Table 4: Affiliation and Number of Users

Affiliation	Number of Users
Pro-Duterte	37181
Anti-Leni	10548
Anti-Leni, Pro-Duterte	3307
Pro-Santiago	2738
Pro-Leni	1915
Pro-Marcos	1908
Pro-Cayetano, Pro-Duterte	969
Anti-Delima	761
Anti-Duterte	646
Anti-Leni, Pro-Marcos	533
Pro-Roxas	448
Pro-Duterte, Pro-Marcos	442
Pro-Duterte, Pro-Santiago	340
Anti-Binay	336
Pro-Rappler	295
Anti-Leni, Pro-Duterte, Pro-Marcos	284
Pro-Delima	283
Anti-Marcos	223
Anti-Leni, Pro-Cayetano, Pro-Duterte	216
Anti-Roxas	177
Anti-Roxas, Pro-Duterte	144
Anti-Delima, Pro-Duterte	126
Pro-Marcos, Pro-Santiago	118
Pro-Duterte, Pro-Roxas	106
Anti-Leni, Anti-Roxas, Pro-Duterte	104

A.4 Elite Cueing Results

Model The conducted OLS Interrupted Time Series Analysis is given by the below model -

$$h = \beta_0 + \beta_1 \times \text{intervention} + \beta_2 \times \text{time} + \beta_3 \times \text{intervention} \times \text{time} \quad (1)$$

where,

- h is the proportion of hateful Facebook comments.
- *intervention* is an indicator variable for Duterte's public interventions (start of campaign, attacks, apologies, etc.).
- *time* is time in days since the intervention (relative).
- β_0 is the baseline level of the outcome variable when the treatment (represented by the variable *intervention*) hasn't been applied and *time* is zero.
- β_1 is the effect of intervention – shows how much h changes with the treatment, holding other factors constant.
- β_2 is the time trend – shows how the outcome variable h changes over time, independent of the treatment.
- β_3 is the effect of intervention on time trend – measures how the effect of the treatment (*intervention*) on the outcome variable h changes over time.

We should note that our ITSA model might have potential violations that are not fully accounted for, such as seasonality, time-varying confounders, and higher-order auto-correlation. These factors could influence the observed trends in the proportion of hateful comments, and addressing them in future analyses would strengthen the robustness of the findings.

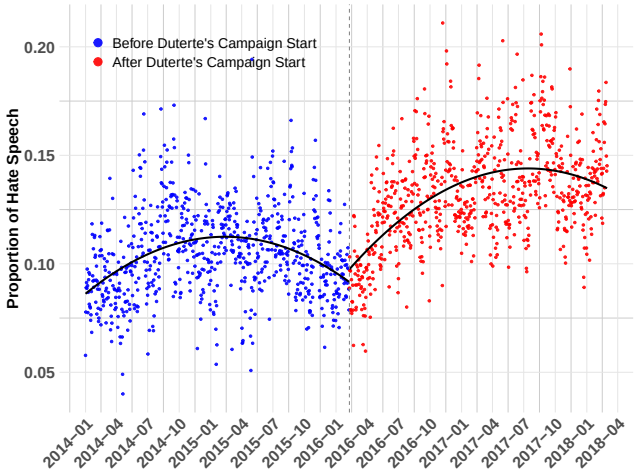

Figure 10: Interrupted Time Series Analysis of proportion of hateful speech from the announcement of Duterte's election campaign (quadratic fit).

Intervention 2: Justifies killing of journalists An ITSA was conducted for 31 May 2016, with a two week window before and after the intervention, to analyze the effect of Duterte's public justification of killing journalists he deemed as corrupt. We expected a rise in hate speech targeted towards journalists, but on the contrary found a negative effect on the level of hate speech immediately after the intervention. The results are shown in Table 10.

Intervention 3: Personal attacks at Senator Leila De Lima On August 17 2016 President Rodrigo Duterte hurled personal abuses at Senator Leila De Lima (reference in footnote) that was largely covered by popular Philippine media. We found that there wasn't any conclusive evidence of an immediate increase in hate speech by Duterte's supporters following his offline attacks. 11.

Model 2: Post-Pledge Reduction in Profanity The regression discontinuity analysis conducted on October 28, 2016—subsequent to Duterte's public commitment to refrain from swearing—exhibits a statistically significant diminution in the proportion of hate speech. This aligns with the anticipated outcomes premised on elite cueing theory. However, the temporal proximity post-intervention is notably truncated, casting doubt on the long-term efficacy of the intervention. This curtailed bandwidth is attributable to subsequent overlapping interventions.

Model 3: Reversion to Profanity The third model evaluates the regression discontinuity associated with November 3, 2016, when Duterte reneged on his vow to avoid public use of profanity. Contrary to the hypothesized immediate amplification in hate speech among Duterte's adherents,

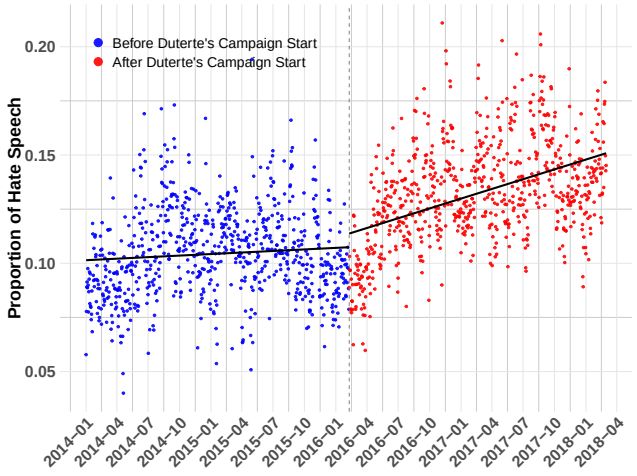

Figure 11: Interrupted Time Series Analysis of proportion of hateful speech from the announcement of Duterte's election campaign (AR(1) fit).

the results, while indicating an uptick following the intervention, did not reach statistical significance. Nonetheless, the post-intervention trend does suggest a statistically significant increase in the frequency of hate speech.

Model 4: Unpublicized Pledge Against Profanity On February 21, 2018, Duterte once more avowed to abstain from profanity, an event that failed to capture widespread media attention. Analogous to the observations in Model 2, a decline in hate speech was anticipated. Contrariwise, the analysis registers a statistically significant surge in hate speech, as evidenced in Table 7.

Model 5: Davos Night Market Explosion The regression discontinuity analysis for September 2, 2016—the date of the Davos Night Market explosion—was anticipated to exhibit an elevation in hate speech, premised on the theory of elite cueing. Surprisingly, the empirical evidence suggested a significant decline in the proportion of hate speech post-intervention. This unanticipated outcome contradicts the expected increase and indicates a complexity in the relationship between elite rhetoric and hate speech propagation that warrants further investigation.

A.5 Page Annotation

Job Description The same job description and annotator criteria were used for both 2021 (when one annotator was hired) and 2024 (when four were hired). The following job description was posted on Upwork.com to recruit annotators:

"We are seeking individuals knowledgeable about Filipino politics and fluent in Tagalog to assist in annotating the political leanings of various Facebook pages. This project involves reviewing and analyzing the content of these pages to determine their political affiliations. Ideal candidates should have a deep understanding of Filipino politics,

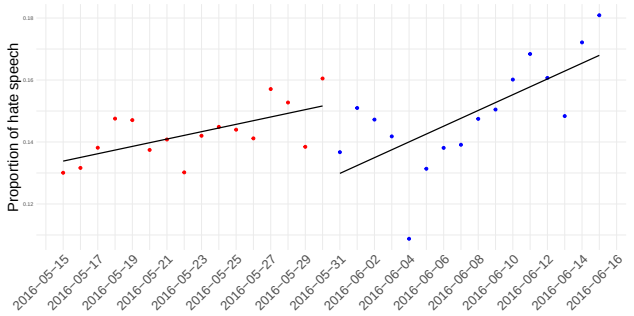

Figure 12: Interrupted Time Series Analysis of proportion of hateful speech from when Duterte justified killing of journalists (May-31-2016). Blue dots indicate the proportion of hateful speech in a two weeks time-span after Duterte's justification. Red dots represent the proportion of hateful speech in comments two weeks prior to Duterte's justification.

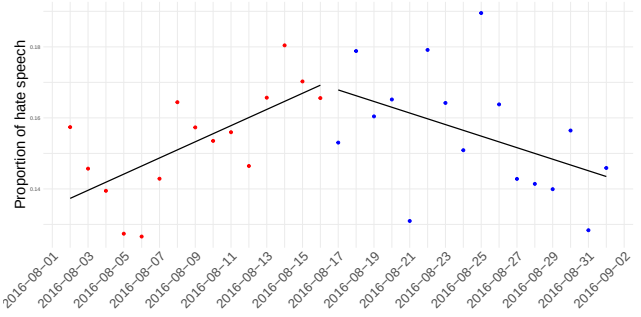

Figure 13: Interrupted Time Series Analysis of proportion of hateful speech for when Duterte attacked Leila De Lima at a press conference (Aug-17-2016). Blue dots indicate the proportion of hateful speech in a two weeks time-span after Duterte's attack at De Lima. Red dots represent the proportion of hateful speech in comments two weeks prior to Duterte's attack.

proficiency in reading and writing Tagalog, and strong research and analytical skills. This is a remote position that can be performed from anywhere, and detailed instructions will be provided. If you are passionate about politics and possess the necessary expertise, we would love to hear from you."

Skills Required:

- In-depth knowledge of Filipino politics
- Fluency in Tagalog
- Strong research and analytical skills

Application Requirements:

When submitting a proposal, applicants were asked to specify:

1. Are you well aware of the political parties and politicians in the Philippines?
2. Are you familiar with the political events that took place during the 2016 and 2022 elections?

3. Do you use Facebook?
4. Can you read and understand Filipino/Tagalog text well?

Applicants were also requested to describe their recent experience with similar projects. The submitted curricula vitae and proposals were vetted by our team to select suitable candidates.

Annotation Instructions All the four annotators (albeit hired at different times) were provided with the following instructions:

"We have 1,284 Facebook pages and groups to be annotated. The URLs for all these pages/groups are provided in an Excel sheet. You need to visit each page/group using the URL and verify its political affiliation. Each page/group should be coded using one of the following four values (all lowercase):

- **pro-duterte**: for pages and groups that support ex-President Duterte
- **anti-duterte**: for pages and groups that are against ex-President Duterte
- **neutral**: for pages that are not politically oriented or do not have clear support for any particular party
- **unknown**: if the affiliation is not clear or unknown

*For pages that do not currently exist, please mark them as **unknown** as well. Please use the exact annotations as provided (all lowercase)."*

Determining Affiliation The following was mentioned:

"Importantly, you should only use information from the name/title of the page/group, the description, the content of posts, and the content of comments on posts. Only comments and posts from 2016–2017 should be used for assigning political affiliation. It is crucial not to use any prior knowledge or biases; your decisions should be based solely on the evidence provided in the descriptions and content of the pages and groups.

For pages that are not currently available or whose links are broken, if you feel the leaning is obvious or quite evident from the title/name of the page, please assign the respective political leaning code. Remember, this applies only to pages with broken or unavailable links and is intended to maximize the information we can obtain. For all other pages, please use the strategy described above.

If you agree to work on this task, we will provide you with a sample of 20 URLs in an Excel sheet to assess your ability to perform the task correctly. Please let us know if you are interested in the task and if you have any questions."

The sample of 20 URLs included an equal representation from the four affiliation categories and was drawn from a subset of affiliations provided by journalists. Annotators were asked to provide additional notes on the affiliation of each page/group and categorize them using labels they deemed appropriate. The ad stated that a bonus would be offered for highly descriptive category labels. The first four applicants who achieved significant inter-coder agreement with the actual test set labels (Cohen's $\kappa > 0.75$) were recruited for the task.

Page Categories A detailed description of page categories and the number of pages affiliated to each is provided in Table 13. We used the labels of the annotator that provided the most descriptive information. This annotator explained that when a page or group fit multiple categories or the category was unclear, they assigned the most relevant category based on recent posts. The annotator informed us that labels were developed through an iterative thematic analysis, with the categories divided at the level of granularity the they deemed most appropriate.

A.6 Hate Speech Annotation

As outlined in Facebook's policy (Meta 2021), "We define hate speech as a direct attack against people—rather than concepts or institutions—on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease." Additionally, "We define a hate speech attack as dehumanizing speech; statements of inferiority, expressions of contempt or disgust; cursing; and calls for exclusion or segregation." Based on this definition, our approach includes harassment, attack and toxic speech directed against any of the predefined protected characteristics as hate speech.

Details provided to annotators on Upwork:

Help us identify derogatory content on social media!

In this task, we show you a comment posted by a user along with the original post on which the comment was made. Our goal is to identify Facebook comments that insult, harass, or attack others based on their protected characteristics, such as race, ethnicity, gender, disability, or religion. According to Facebook's policy,[6] derogatory content includes dehumanizing speech, harmful stereotypes, and expressions of contempt or disgust aimed at individuals or groups, which can create an environment of exclusion and intimidation.

- 1. Content intended to cause disruption, trigger conflict or insult for amusement. Users who participate or conduct trolling are called trolls. e.g. "You look like the generic gay hipster that has too high of an ego. Du30 will lock you all up".
- 2. Derogatory content: Insults and messages that are offensive and directed to any group or individual. e.g. "O FUCK YOU U MATHRFUKER BITCH PRESSTITUTE. ALL JOURNALISTS are idiots".
- 3. Profanity: Comments that contain profane words. e.g. "you are a fucking moron", or "I will rape you, bitch".
- 4. Hate speech: An expression of hatred towards individuals or groups on the grounds of their identity. e.g. "I'm going to start killing these assholes. Chin chin."
- 5. Explicit threats: e.g. "Protect the president PRDU3O and kill all destabilizers".
- 6. Attacks on specific groups like journalists or politicians, including leaders currently or formerly in power

[6] https://transparency.meta.com/policies/community-standards/hate-speech/

or in the opposition. e.g. “#LeniResign #LeniResign #LeniResign #LeniResign #LeniResign #LeniResign #LeniResign #LeniResign #LeniResign #LeniResign”.

Even if part of the comment contains such speech, please mark the comment as derogatory.

The data annotation has been much harder than we had anticipated. Finding annotators who speak the Filipino language has been an issue. We tried Amazon Mechanical Turk, Prolific and Appen which is usually used for crowdsourced annotations. However, since our content is in the Filipino language, we did not succeed. To overcome this, we recruited volunteers through the gig working platform Upwork.com. We recruited four local language professionals who were native Filipino speakers and were well versed with the Filipino politics.

Table 5: Hashtags used in support of politician (part 1)

Subsection Title	Hashtags
Pro-Duterte	duteteforpresident, duterte2016, dutertecayetano, dc2016, dutertecayetano2016, ducay, ducay2016, soliddurterte2016, presduterte2016, wesupportduterteadministration, phvoteduterte, du302016, du30cayetano, dc, godu30, solidduterte, duterteparin, welovedigong, teamduterte, uniteddds, du30forpresident, phvotesduterte, dds, ducos, solidducayaqsapagbabagongbansa, supportduterte, prouddds, soliddu30, goduterte, saludoduterte, du30forpresedent2016, teamdavao, voteduterte2016, duterteyouth, du30parasapagbabago, phduterte, duteronlyhope, gotataydigong, dutertepamore, duterteismypresident, presidentdu30, du304life, isupportduterte, duterteforpresident, du30ftw, dubong2016, allpinoy4duterte, isuportdu30, pdu30, pduterte, duteete, votedutertecayetano, presidentrodrigoduterte, digong, uniteforduterte, soliduterte, solidutertehere, wesalutedu30, changeishere, mypresidentdigong, produterte, team_du30, dutertemarcosthebesttandem, du30bbm, fightfordu30, dutertebestpresident, d30, presidentduterte, changeiscoming, duterteako, duriam, labandu30, yestoduterte, du30, partnerforchange, iloveduterte, dutertemarcos2016, dutertemypresident, duterteornothing, radicalchangeiscoming, dutertenatayo, dutertenakami, dutertenaako, duterte-cayetano, foreverduterte, phvoteducay, prayforduterte, pray4duterte, peoplescallforduterte, tataydigong, duriampamore, isupportduterteadministration, ilovepresidentduterte, isupportpresidentduterte, dutertesolid, changehascome, duterteadministration, fight4duterte, duterteuntilmylastbreath, forthewinduterte, ilovemypresidentdu30
Anti-Duterte	angtagalmaimpeachduterte, impeachduterte, impeachdigong, notoduterte, notodutertes, nomoredutertesever, digongresign, dutirty, changescamming, impeachd30, oustduterte, resignduterte, duterteresign, no4duterte, dutertard, dutertetard, duterteistheworstpresidentever, unfitpresident, regretiscoming, diedutertards, no2du30dq, impeachditerte, trollking, dutertetroll, dilawan_trolls, duterteisanaddict, insecureduterte, duterteatraitor, duterteisacriminal, kupalsiduterte, notoduterte2016, dictator, dutertemassmurderer
Pro-Cayetano	cayetanoforvp, cayetano, phvotecayetano, dutertecayetano, alanpetercayetanovp, phvoteducay, sencayetano, cayetanoangvpko
Pro-Delima	angtagalmaimpeachduterte, impeachduterte, impeachdigong, notoduterte, notodutertes, nomoredutertesever, digongresign, dutirty, changescamming, impeachd30, oustduterte, resignduterte, duterteresign, no4duterte, dutertard, dutertetard, duterteistheworstpresidentever, unfitpresident, regretiscoming, diedutertards, no2du30dq, impeachditerte, trollking, dutertetroll, dilawan_trolls, duterteisanaddict, insecureduterte, duterteatraitor, duterteisacriminal, kupalsiduterte, notoduterte2016, dictator, dutertemassmurderer
Anti-Delima	ihatedelima, delimaresign, delimabringthetruth, noneforleila, ripleila, sabaforleila, saba4leila, resigndelima, impeachdelima, drugprotectordelima, thiefdelima, adultererdelima, sexmaniacdelima, liardelima, pcosmachinedelima, guiltydelima, lairdelima, whoredelima, drugtraderprotectordelima, drugtraderprotector, corruptdelimacohorts
Pro-Binay	binay2016, binayparin2016, onlybinayknows, onlybinay, binaythealienmovement, binayforpresident2016, binayforthepoor, binaynihan, binayforpresident
Anti-Binay	notobinay, binayresign, notobinay2016, stopbinay, ripbinay, binaybigfatliar, impeachbinaynow, anyonebutbinay, binaysucks, stoppoliticaldynasty, binaygotohell, deflectingyourfamilyscorruption
Pro-Santiago	mds, phvotesantiago, miriam2016, switch2miriam, miriamforever, angatkaymiriam, santiago2016, mdsforlife, switchtomiriam, miriamforpresident, miriamparin, mds2016, miriamdefensorsantiago, miriam, duriam, duriampamore, voteformiriam, youthformiriam, miriamparin, mds2016, miriamforpresident, mdsforpresident2016, youthformiriam2016movements, mdsforpresident, iamformiriam, miriammagic, miriamfight, miriamtuloyanglaban, miriamsantiago
Pro-Marcos	bbm4thewin, solidmarcos, ducos, bbmvp, vpbbm, bbmtruevp, bbmtherealvp, bbm4vp, bbmrealvp, dubong2016, marcosparin, bbmforvp, dutertemarcosthebesttandem, bongbongmarcos, yesbbm, bbm2016, dutertemarcos, dutertemarcos2016, bbmrealvp, bbmrealvicepresident, bbmmyrealvicepresident, fight4bbm, bbmforever, phvotebbm, wevotedbbm, votebbm, ilovebongbong, victoryformarcoses, marcosishero, marcosinnocent

Table 6: Hashtags used in support of politician (part 2)

Subsection Title	Hashtags
Anti-Marcos	marcosmagnanakaw, marcossohungryforpower, bbmoutofthepicture, byebyemarcos, marcosisnotahero, marcosnotahero, notomarcos, nomoremarcoseinmalacanang, marcosthebiggestthief, notomarcosjr, notobbm, marcosfakehero, notomarcoses, crynabbm, delusionalbbm, marcosisacriminal, gotojailmarcos, marcosburial
Pro-Leni	leni4vp, lenizoned, protectvpleni, vpleni, congratsvpleni, lenimyvp, leniforthewin, leniismyvp, labanleni, lenirobredotherealvicepresident, leniforvp, lenitherealvp, women4leni, oneforleni, liberalforever, lenibeatsnotcheats, kapitleni, leaveLenialone, installrobredo, marleni2016, protectleni, ivoteleni, lenirobredovp, womanwithintegrity, myvpleni, ipaglabansileni, labaleni, palagleni, yestoleni, wewillprotectleni, weloveyouvpleni, defendvpleni, oneforvpleni, roxasrobredoforthewin, ivotedforleni, lenirobredo2016
Anti-Leni	resignedleni, impeachleni, resignfakevp, resignleni, oustleni, impeachlenirobredo, fakevp, lenipowergraber, leninomore, lenipabigatsabayan, lenipowergrabber, impeanchlenilugaw, boboleni, leniresign, impeachlenilugaw, impeachlenilugawnow, impeachlenirobredonow, vpvoterrecount, notoleni, leniresign, notolp, impeachleninow, notolenirobredo, lenilangsot, lenilastog, leniletche, leniloko, lenileaks, recountvp, leniimpeach, lenipambansangtraydor, lenirobreopowergrabber, vprecount, fakevplenirobredo, lenirobredoresign, whorefakevplenilugawfraudredo, nomoreyellowtards, nomoreyellowtae, notoliberalparty, impeachlleni, leniresignfakevp, lenistopdemonizingourgovt, impeachtheyellowturd, impeachfakevp, lenipowergrabber, powegrabber, oustrobredo, fakevp, disbarleni, powergrabberlenilugaw, oustlenirobredo, recount, yellowtard, yellowtards, yellowshit
Pro-Roxas	roxas, roro, teamroro, teamroxas, solidroxas, youaretheonemrpalengke, nsdmar4president2016, mrpalengke, marroxa, marroxas, yestomarroxas, marleni2016, marthebest, welleducatedwellmanneredwellraised, yestolp, marroxas2016, roxasforpresident, goroxas, roxasrobredoforthewin, orasnaroxasna, phvoteroxas, orasnaroxas, phvotemarroxas2016, roxasalltheway, onlyroro
Anti-Roxas	notomar, notomarroxas, notoroxas, roxasmandaraya, asapamoreroxas, notolp, nomoreyellowtards, nomoreyellowtae, roxasrapist
Pro-Rappler	supportrappler, istandwithrappler, supportpressfreedom, defendpressfreedom, fightforpressfreedom, standwithrappler, supportreesa, istandforrappler, isupportrappler, supportrealjournalism, supportfairhonestjournalism, supportfreedomofthepress, standwitrappler, isupportthetruth, supportfreedomofexpression, blessyourappler, istandforpressfreedom, upholdrealjournalism, labanrappler, pressfreedomisaright, supportpressfreedom, standwithrappler, isupportrapper
Anti-Rappler	supporttostoprappler, nevertrustrappler, notofakenews, standnotforrappler, fakerappler, riprappler, fakenewsisrappler, shutdownrappler, stopfakenews, neveragainrappler, abolishrappler, oustrappler, nomorefakenews, rappler_is_a_law_breaker, notorappler, goodbyerappler, karmarappler, onenightstandwithrappler, istandwiththeconstitution, stoppressmanipulation, unsubscribedrappler, isupporttheconstitution, boykootrappler, upholdtheconstitution, arrestmariaressa, thenurve, terriblecult, unfollowrappler, unfollowingrappler

Table 7: OLS ITS model coefficients for Figure 6 in Section 5.4

	Estimates
β_0	0.1073^{***}
	(0.0014)
β_1	0.0115^{***}
	(0.0021)
β_2	0.0000^{*}
	(0.0000)
β_3	0.0000^{***}
	(0.0000)
R^2	0.3202
Adj. R^2	0.3189
Num. obs.	1637
RMSE	0.0215

$^{***}p < 0.001; ^{**}p < 0.01; ^{*}p < 0.05$

Table 8: ITS model coefficients (quadratic), for Figure 10) in Section A.4

	Estimates
β_0	0.0904^{***}
	(0.0019)
β_1	0.0155^{***}
	(0.0020)
β_2	-0.0002^{***}
	(0.0000)
β_{22}	-0.0000^{***}
	(0.0000)
β_3	0.0004^{***}
	(0.0000)
R^2	0.6031
Adj. R^2	0.6021
Num. obs.	1550
RMSE	0.0206

$^{***}p < 0.001; ^{**}p < 0.01; ^{*}p < 0.05$

Table 9: ITS model coefficients (AR(1), for Figure 11) in Section A.4

	Estimates
β_0	0.0848^{***}
	(0.0029)
β_1	0.0000
	(0.0000)
β_2	0.0045^{*}
	(0.0021)
β_{T-1}	0.2068^{***}
	(0.0244)
β_3	0.0000^{***}
	(0.0000)
R^2	0.3938
Adj. R^2	0.3922
Num. obs.	1520
RMSE	0.0146

$^{***}p < 0.001; ^{**}p < 0.01; ^{*}p < 0.05$

Table 10: Coefficients for Figure 12 (quadratic model)

	Estimates
β_0	0.1621^{***}
	(0.0068)
β_1	-0.0259^{**}
	(0.0083)
β_2	0.0043^{**}
	(0.0014)
β_{22}	0.0002^{*}
	(0.0001)
β_3	-0.0045
	(0.0026)
R^2	0.5560
Adj. R^2	0.4902
Num. obs.	32
RMSE	0.0100

$^{***}p < 0.001; ^{**}p < 0.01; ^{*}p < 0.05$

Table 11: Table for Figure 13 (quadratic model)

	Estimates
β_0	0.1703^{***}
	(0.0077)
β_1	-0.0033
	(0.0096)
β_2	0.0019
	(0.0023)
β_{22}	-0.0000
	(0.0002)
β_3	-0.0031
	(0.0047)
R^2	0.3036
Adj. R^2	0.1965
Num. obs.	31
RMSE	0.0146

$^{***}p < 0.001; ^{**}p < 0.01; ^{*}p < 0.05$

Table 12: Table of coefficients for the ITS models from 2 to 5

	Model 2	Model 3	Model 4	Model 5
β_0	0.1436^{***}	0.1074^{***}	0.1346^{***}	0.1471^{***}
	(0.0024)	(0.0048)	(0.0086)	(0.0058)
β_1	-0.0220^{**}	0.0168	0.0315^{*}	-0.0263^{**}
	(0.0072)	(0.0157)	(0.0117)	(0.0080)
β_2	0.0002^{***}	-0.0033^{*}	-0.0012	0.0006^{**}
	(0.0000)	(0.0011)	(0.0007)	(0.0002)
β_3	-0.0002	0.0102^{**}	-0.0006	-0.0000
	(0.0028)	(0.0028)	(0.0010)	(0.0003)
R^2	0.3569	0.4038	0.2609	0.1648
Adj. R^2	0.3496	0.2249	0.1976	0.1366
Num. obs.	269	14	39	93
RMSE	0.0184	0.0275	0.0155	0.0188

$^{***}p < 0.001; ^{**}p < 0.01; ^{*}p < 0.05$

Table 13: Description of page categories and number of pages affiliated to each

Page Type	Description	Count
Unavailable Pages	Pages that are no longer accessible or have been deleted.	365
Content based	Pages producing original content including general information, memes or satirical posts, acting as their own entities rather than representing individuals.	230
Supporter Pages	Pages promoting a person, family, or political clan in a positive light.	224
Community based	Groups where members contribute content centered around specific interests or topics.	168
News and Media Pages	Pages primarily sharing news content; quality and biases may vary.	86
Personal Pages	Pages owned by individuals using their real names to interact with fans or share personal content.	57
Government Pages	Pages owned by government institutions or officials; may or may not display political leanings.	32
Advocacy Pages	Pages promoting specific causes or advocating for societal or political change.	27
Political Campaign Pages	Pages sharing political content, usually supporting a candidate or political message.	17
Local Area Pages	Pages focused on content related to a specific geographic area, such as a town or region.	17
Content Aggregator Pages	Pages primarily sharing content from other sources with little original material.	16
Influencer Pages	Pages owned by individuals using pseudonyms or stage names, sharing opinions and branded content.	14
Organization Pages	Pages representing organizations, posting content relevant to their members or audience.	11
Marketplace Pages	Pages serving as online marketplaces, promoting products or services for sale.	10
Engagement Bait Pages	Pages seeking reactions and likes without focused or consistent content.	5
Political Party Pages	Pages representing political parties, promoting their values and candidates.	2
Religious Pages	Pages focused on religious content or affiliated with religious institutions.	1
Product Promotion Pages	Pages dedicated to specific products or offerings.	1
Event Promotion Pages	Pages promoting events such as concerts, festivals, or community gatherings.	1