Dissecting Hate Speech Campaigns in the 2016 Philippine Elections on Facebook

Sudhamshu Hosamane
Rutgers University
New Brunswick, USA
sudhamshu.hosamane@rutgers.edu

Kiran Garimella
Rutgers University
New Brunswick, USA
kiran.garimella@rutgers.edu

ABSTRACT

This 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 hate speech campaigns in this digital political arena. Our findings reveal that over 12% of comments on political pages were hate speech, predominantly linked to the Duterte campaign and its affiliates, with Duterte's supporters initiating over 90% of these hate speech comments. We further explore the relation between offline political events and online hate speech, identifying mixed evidence of causality following Duterte's public criticisms of journalists and politicians. Alarmingly, we observe a 'spillover effect' where regular social media users, after exposure to orchestrated hate speech, began emulating troll-like behaviors. This contagion effect highlights a worrying trend in social media's influence on public opinion and discourse. The results of our research are crucial for understanding the dynamics of digital political campaigns and their extensive implications for democracy and public discourse. Particularly, considering Facebook's extensive usage in the Philippines, we contend that the platform's widespread employment for these activities presents significant concerns.

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.

The social media strategy employed by President Rodrigo Duterte's campaign in the Philippines serves as a significant illustration of how national leaders can effectively harness these platforms to achieve political objectives. Maria Ressa, a prominent journalist and Nobel Peace Prize laureate, has highlighted the extensive use of both real and fake Facebook accounts in the Philippines to disseminate disinformation and manipulate public opinion under Duterte's regime. This strategy, which effectively floods the information space with lies and distorts the public's understanding of facts, serves as a cautionary tale for other countries. Ressa's observations point to the critical need for interventions by tech companies to address these challenges and preserve the integrity of facts, especially in the context of elections [13]. 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 [3, 34].

The challenge in addressing this problem lies in the sheer complexity and subtlety of online discourse making it difficult to categorically identify and analyze hate speech. This is compounded by the nuanced cultural and linguistic contexts within which such communications occur, often requiring deep local knowledge. Furthermore, access to comprehensive data poses a significant hurdle. Social media platforms like Facebook often restrict data availability, making it challenging to obtain a dataset extensive enough for meaningful analysis. Studies focusing on the global south, such as the Philippines, are particularly scarce, reflecting a geographic bias in existing research and a lack of resources devoted to these areas within the computational social science research community. This gap in research and data exacerbates the challenge, leaving critical aspects of global digital political campaigning largely unexplored and poorly understood. Our study seeks to bridge these gaps by leveraging a large-scale dataset, providing rare insights into the digital political landscape in a less-studied region, and highlighting the unique challenges faced in understanding and mitigating hate speech and online trolling in such contexts.

Previous research in this domain has predominantly centered on qualitative analyses or case studies, often constrained by limited datasets, which only illuminate a fraction of the broader phenomenon [7, 30, 32]. While these studies have made significant strides, they have limitations. Ong and Cabañes [30] reveals the presence of paid troll farms, highlighting a strategic use of social media for political manipulation. However, this revelation is largely qualitative. Karunungan [19] delves into the robustness of Duterte's Facebook ecosystem, illustrating how it was optimized for campaign messaging, but stops short of quantifying the impact. Montiel et al. [25] provides anecdotal instances indicating a rise in hate speech during the election period, but lacks a comprehensive, data-driven analysis. These studies, while pivotal, primarily offer surface-level insights without extensive empirical evidence or statistical analysis to measure the full extent and ramifications of these digital strategies. Our research aims to fill these gaps by providing a thorough, quantitative assessment of the scale and effects of hate speech and social media manipulation during the 2016 Philippine elections.

Our approach distinctively employs a large-scale, quantitative analysis, utilizing millions of Facebook comments to facilitate a comprehensive understanding of the prevalence, structure, and impact of hate speech and trolling campaigns. Methodologically, we developed a high-precision hate speech detection model tailored for code-mixed Filipino text. This model was applied to the extensive

comments dataset, enabling us to trace the prevalence and spread of hate speech, with a specific focus on pro-Duterte supporters' activities.

Our key findings include:

- We discovered that, on average, 12% of the comments on political posts constituted hate speech. In absolute terms, this represents tens of millions of comments, indicating a substantial volume of hate-fueled discourse on social media.
- 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, this paper contributes to the understanding of digital politicking's impact on democracy and public discourse, offering a novel, data-driven perspective on the challenges and implications of social media in political campaigns. For the research community, our results showcase the effectiveness of tailored computational models in processing and interpreting vast amounts of code-mixed language data, in under resourced contexts thus bridging the gap between technology and practice.

2 RELEVANT LITERATURE

2.1 Background

Rodrigo Duterte became the 16th president of the Philippines after a highly divisive and controversial election in 2016. His political opponents included members of the liberal party such as Presidential candidate Mar A. Roxas, Vice President Leni Robredo,[1] and Leila De Lima, a senator and vocal critic of Duterte. The election was dominated by social media, which served as a primary platform for candidates to reach voters, spread their messages, and discredit opponents. Duterte in particular galvanized support among citizens tired of corruption and high crime through inflammatory speeches and no-holds-barred Facebook live streams. His brash, tough-on-crime persona resonated with masses drawn to his unconventional style untouched by political correctness. Duterte commanded a loyal following of social media warriors who attacked dissenting voices and propagated his contentious views. Critics charged that he manipulated online disinformation networks to smear adversaries and buoy his populist platform [22]. The prevalence of social media introduced new dynamics into Philippine politics, allowing Duterte to circumnavigate traditional media to speak directly to the people. His campaign's mastery of this medium was instrumental to his eventual electoral triumph [13].

[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 campaigns have become an unfortunate staple of recent elections worldwide. Researchers have developed methods to identify coordinated inauthentic behavior on social media by analyzing account metadata, linguistic patterns, and network connections [36, 39]. This research reveals that many supposed grassroots movements are actually astroturfing operations using fake accounts and automation. The goals of such coordinated trolling are multifaceted, including suppressing opposition voices, spreading disinformation, and amplifying extreme narratives [4, 7]. There is evidence that such tactics are often orchestrated by political parties using paid human trolls and bots [33]. The campaigns target both domestic populations and international audiences, exploiting social divisions and digital openness. Emerging evidence shows coordinated trolling represents a serious threat to democratic discourse [1] and elections worldwide. While the tactics keep evolving, researchers and platforms are developing tools to identify and mitigate such inauthentic behavior. However, effectively combating online hate and disinformation will require cooperation across civil society, government, and the technology industry.

2.3 Elite Cueing

Elite cueing refers to the process by which political leaders use cues and signaling to influence followers' attitudes and behaviors [24]. There is extensive research showing how populist politicians can incite collective action in their supporters through indirect rhetorical cues, without directly calling for violence or illegal acts [5, 41]. This phenomenon operates similarly in digital spaces, as seen in how former US President Trump utilized Twitter to energize his base and legitimize far-right viewpoints [38].

There is growing recognition that online hate speech and extremist rhetoric from elites can spill over to enable real-world violence and unrest. For instance, analysis shows Trump's tweets about Muslims and immigrants preceded statistically significant increases in anti-Muslim and anti-immigrant hate crimes in the U.S. [16, 37]. Other work reveals spikes in anti-refugee attacks in Germany following anti-immigrant Facebook posts by far-right groups [26]. However, the relationships are complex, as offline events can also spark increases in online vitriol, seen in anti-Asian hate speech following the initial COVID-19 outbreak [40]. However, the impacts of such elite cueing are complex and context-dependent. While messages from influential political leaders can create an opening for extremist movements, this does not inevitably lead to violence or sustained increases in online hate speech. This aligns with theories on how institutional constraints can moderate the impacts of extremist cueing from elites [15, 17, 20, 21]. In the Philippines, President Duterte has similarly used incendiary rhetoric to legitimize violence against drug dealers and addicts. Qualitative studies have documented how this empowers vigilante groups and police to take extreme actions [10, 30].

2.4 Impact of hate speech and trolling

The impact of trolling, particularly in the context of online discussions, is a multifaceted issue that has garnered significant attention in academic research. One of the key questions explored is whether hate begets hate; that is, if a thread starts off with hateful comments, does it tend to attract more hateful replies? Studies have indicated that initial hate speech in a thread can indeed set a tone that encourages similar behavior. Munger [27] demonstrated that receiving negative social feedback online, such as hate speech, can significantly increase the probability of the individual exhibiting similar behavior. This phenomenon suggests a sort of normalization of hate speech within certain threads, where initial instances of hate speech lower the barrier for others to contribute similarly. Research has also shown that hate speech posts often cluster together. A study by Mathew et al. [23] found that hate speech begets more hate speech, leading to a clustering effect where such posts are concentrated in certain threads or communities. This clustering can create echo chambers of negativity and hostility, exacerbating the problem [28]. The chilling effect of trolls and hate speech campaigns on their targets is also a significant concern. Trolling and targeted hate campaigns can lead to self-censorship among users who fear becoming targets themselves. This effect was highlighted in a study by Phillips [31], which discussed how organized trolling campaigns can silence and intimidate individuals, particularly those from marginalized groups [35]. Finally, the impact of trolls on the behavior of normal users who are neither paid trolls nor invested actors is another critical area of importance. Trolls can significantly alter the tone and nature of online discourse, influencing how ordinary users interact. Buckels et al. [6] found that trolls have certain personality traits like psychopathy and can influence other users who they interact with, indicating that even users who are not directly engaged in trolling can be influenced by the altered informational landscape that trolls help create.

3 DATASET

In this study, we investigate the potential impact of Rodrigo Duterte's presidential campaign on the proliferation of hate speech on Facebook. Our analysis is grounded in an extensive dataset derived from Facebook, comprising text data from both individual posts and and large scale broadcasts by organizations. This dataset is a rich assembly of posts and comments from a variety of sources, meticulously gathered to ensure a comprehensive understanding of the digital landscape during this period. Our dataset includes data from both facebook groups and pages. Facebook pages represent public broadcast channels while groups represent channels for group discussions. Careful consideration went into selecting relevant pages and groups for inclusion.

The selection of the groups and pages was done manually by journalists and editors at a top online Philippino news portal, Rappler, founded by Nobel peace prize winner Maria Ressa. Through their field work, journalists at Rappler initially identified 26 groups being operated by paid political operatives on covering 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. The process was iterated a few more times, expanding the selection to over 300 relevant groups, ensuring a diverse representation of political viewpoints. The selection of pages followed a different yet equally rigorous approach. Starting with a list of hand curated pages by Rappler editors featuring the country's top news sites, we expanded our dataset to include pages shared frequently within our selected groups. Additional pages, including those identified as propagandist, were added based on their prevalence in the groups initially selected. This gave us a list of around 1,000 relevant pages.

This comprehensive, manually curated selection of groups and pages by editors and our research team ensured a balanced representation. Overall, we collected 1,285 groups and pages,[2] which included around 5 million posts and 400 million comments on these posts using the Facebook Graph API.[3] The data spanned over 10 years, starting from 2008. To consider data that is contextually relevant, we consider comments from Jan 01, 2014, to March 1, 2018 in our analysis. 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 [30]. 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 [30] 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 this link.

In assessing the integrity and implications of our dataset, it is crucial to acknowledge potential sampling biases. Our sampling methodology, while high in precision, does not guarantee full coverage. This is a common limitation in social media studies, where recall can only be accurately estimated by the platform itself [29]. The pages and groups were selected through an iterative process by experts seeking to document relevant online activity in good faith. However, there may be missing pages, particularly from certain political factions. Without full platform access, expert curation is the best available method for constructing a relevant dataset. Even given potential gaps, these results represent a lower bound on true activity. Given Facebook's profound impact on societal discourse and political dynamics, our study offers crucial insights, even within the constraints of our sampling method. Because the page and group selection methodology emphasizes relevance and balanced political representation, this dataset enables uniquely valuable analysis of social media's role in Duterte's election. No other publicly available data offers a comparable window into this sphere of online activity, making it an invaluable resource for understanding the complex interplay between social media and political engagement.

3.1 Annotating pages

We manually annotated the 1,285 Facebook pages, categorizing them into four distinct groups: Pro-Duterte, Anti-Duterte, Neutral (predominantly news websites), and Unknown. The Unknown category included pages with ambiguous or generic titles, such as "Philippines Politics" or "Death Penalty in the Philippines." Our

[2]For simplicity, for the rest of the paper, we refer to groups and pages as only 'pages'.
[3]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-more-api-restrictions-and-shutdowns

initial assessment relied on page titles; where clarity was lacking, we delved deeper, examining the page's bio and a selection of posts. Pages that had been deleted were also assigned to the Unknown category. For this task, we enlisted native Filipino speakers with a strong understanding of Filipino politics, recruited via the freelancing platform Upwork.com.

Our analysis revealed that 53.4% of the examined pages were Pro-Duterte, underscoring his significant presence and influence within the social media landscape. Conversely, only a smaller segment of 9.9% was identified as Anti-Duterte. Neutral pages, primarily online news sources, constituted 18.4% of our dataset. The remaining 13.5% fell into the Unknown category. While Pro-Duterte sentiment was dominant, the results demonstrated a diverse range of political views represented on these pages.

3.2 Identification of user support

We assigned political support of users using the hashtags they use. 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 [12] 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 0.4% of all the users, they account for 10.3% of the total comments. For users with multiple affiliations, users were assigned to a single category 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 (Section A.3) were used to resolve tie-breakers. Appendix tables 4 and 5 show the full list of hashtags we used. 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 consisted over 66.3% of all of our final user labels.

Though the approach of using hashtags may not provide high recall, we wanted to be sure that we identified supporters with high precision. Given the manual care taken at defining the hashtags, we are confident that our categorization helped us identify true supporters of various sides. Future work could look at interaction networks to extend labels of annotated users (e.g. replying to each other frequently).

4 HATE SPEECH DETECTION

Hate speech detection has been a prominent area of research for several years [11], 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 [42], 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. [2], 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 [9], 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. 24.9% of our annotated dataset was hate speech. The annotation process, both for original and pseudo-labels, is detailed comprehensively in the Appendix (Section A.5). Each data point was labeled as either hate speech or not, based on the criteria outlined therein. As we can see in Section A.5, our definition of hate speech is expansive and includes trolling, profanity, explict threats, etc. Using this broad definition, for simplicity, we use trolls and hate speech posters interchangeably in the rest of the paper.

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 [8]. 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 [8] 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 et al. [18], 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 hate speech 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. [23] 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.[4]

Interestingly, while the proportion of hate speech hovered around 10% before the elections, it surged significantly by during the election period (starting January 2016) and remained elevated thereafter. This sustained trend into Duterte's presidency, which began in June 2016, underscores a continuous and aggressive use of hate speech on social media. The persistent trend of high levels of hate speech, which not only emerged during the election period but also continued throughout Duterte's presidency starting in June 2016, reveals a significant and concerning dynamic in the realm of online political discourse. This phenomenon indicates a 'constantly being at war' situation on social media, where the tactics of hate speech campaigns, initially deployed during the electoral campaign, were sustained and possibly even 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 [32]. 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.

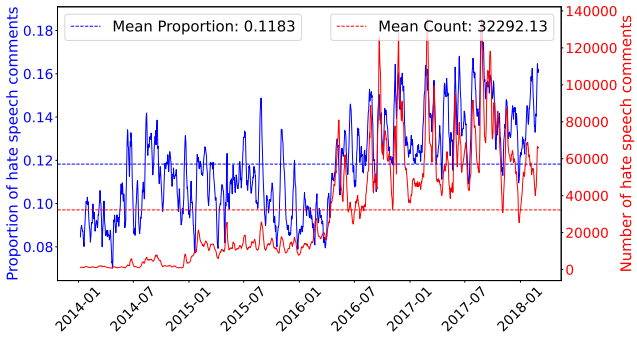

Figure 1: Trends in the total volume (red) and proportion (blue) of hateful comments in our dataset. The lines show a 7 day moving average.

[4]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.

5.2 Who is posting the hate speech?

Next, we looked at who is posting hate speech. For this, we split the users into pro and anti Duterte supporters (from Section 3.2). Our findings revealed a stark disparity in the distribution of hate speech among these two groups. An overwhelming 93.1% of hate speech comments were attributed to Pro-Duterte supporters, while only a marginal 6.9% originated from the Anti-Duterte faction. This distribution was unexpected, particularly considering the intensity of political discourse surrounding Duterte's presidency.

The CDF of posting behaviors in Figure 2 offers insightful distinctions in the hate speech posting behavior of pro and anti-Duterte supporters. While the patterns for posting non-hate content are quite similar between the two groups, the divergence becomes starkly evident in their hate speech posting behaviors. A significant finding is that over 40% of pro-Duterte supporters have shared more than 100 hateful posts, in stark contrast to approximately 15% of anti-Duterte supporters exhibiting the same level of hate speech activity. Furthermore, about 20% of Pro-Duterte supporters have shared upwards of 1000 hateful posts. This discrepancy in the volume of hate speech posts between the two groups is not only substantial but also indicative of deeper underlying factors.

Our third key finding pertains to the fraction of posts categorized as hate speech by each user group. This metric offers a revealing perspective at the individual user level, as shown in Figure 3. The data indicates that, on average, individual pro-Duterte supporters post almost double the amount of hate speech compared to their anti-Duterte counterparts.

Coordinated Posting. Next, we explore coordinated posting on social media, a phenomenon underscored by the presence of troll factories and private, for-hire personnel as noted by Ong and Cabañes [30]. To detect instances of coordinated posting, we employed Locality Sensitive Hashing (LSH) [14], a technique effective in identifying near-similar posts. Our application of LSH revealed a significant amount of coordinated activity: we identified 5,673 instances where the same message was reposted more than 50 times. Some campaigns were particularly extensive, with the largest comprising over 7,000 messages. Further details of these findings, including the distribution of these campaigns, are presented in Figure 9 in the appendix. In terms of political affiliation, pro-Duterte supporters were present in 46.7% of these coordinated campaigns, while 10.1% involved anti-Duterte messages. We also examined the overlap between coordinated campaigns and hate speech. Our analysis showed that 20.5% of the coordinated campaigns involved hate speech. Breaking this down further, 42% of hate speech campaigns were linked to pro-Duterte supporters, and 9.9% were associated with anti-Duterte groups. The average size of these coordinated campaigns, as depicted in Figure 8 (Appendix, Section A.2), was notably larger for those involving pro-Duterte supporters compared to anti-Duterte supporters.

An important aspect to consider in interpreting these results, particularly in context of previous qualitative work [30], is the role of top down, organized digital campaigns, possibly involving paid trolls or devoted supporters, in disseminating hate speech to reinforce Duterte's political narrative. This tactic is not uncommon in modern political campaigns, but in 2015, campaigns at this scale were unheard of and might have played an important role where social media is used as a battleground for shaping public opinion.

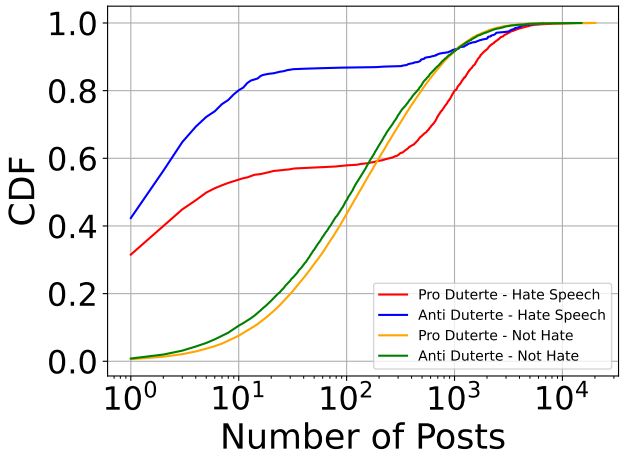

Figure 2: CDF of posts – hate speech and otherwise – for pro- and anti-Duterte supporters. We can see a clear difference in the posting behavior of hate speech posts between pro- and anti-Duterte supporters.

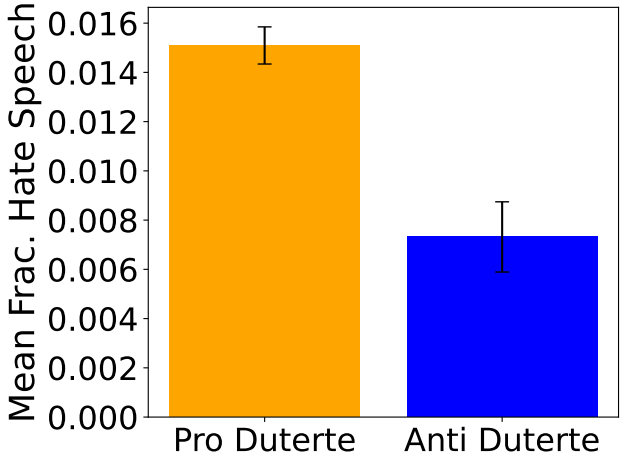

Figure 3: Fraction of hate speech by pro- and anti-Duterte supporters. Roughly 1.5% of the posts by pro-Duterte supporters were hateful, where as for anti Duterte it was significantly less. Error bars show 95% confidence intervals. The difference is statistically significant ($p < 0.001$).

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 posts 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.

5.4 Evidence of elite cueing

In this section, we want to understand whether Duterte's actions offline have an impact on his followers' behaviors online, particularly in the context of posting hate speech. 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 [25]. We are particularly interested in whether Duterte's significant actions, like the kick off of his campaign, or attacks on journalists/female politicians has had a significant *causal* impact on increase in online hate speech. We specifically focus on the commenting activity of 58,700 pro-Duterte supporters identified in Section 3.2. We study the proportion of hateful comments on Facebook across different pages by these users, over time, and see if there is a sudden increase in the proportion of hateful speech during dates that correspond to Duterte's actions in the real world. To achieve this goal, we use Interrupted Time Series Analysis (ITSA).

At the outset of Duterte's official campaign (February 2016), an ITSA was conducted, encompassing the full data spectrum from Jan 2014 to March 2018. The analysis revealed a conspicuous elevation in hate speech at a macroscopic level, as visible in Figure 6. The full ITSA coefficients are tabulated in Table 6 in the Appendix. The findings not only denote a significant level change, indicating an abrupt intensification of hate speech, but they also suggest a post-campaign commencement of a sustained upward trend in hate speech, persisting into Duterte's actual 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 11 and 12. 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 delve into the relationship between posting activity and the amount of hate speech a post attracts, seeking to understand the potential 'spillover' effect of trolling activities. 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 constructed a user network derived from a bipartite network of users and the pages on which they post. On this network, we calculated the PageRank of the users, to identify users with high levels of posting activity. High centrality scores indicate users who are not only active but also influential within the network. Analyzing our hypothesis for users belonging to varying centrality levels is of key interest for us. We considered the top 20% users with highest centrality as *High centrality* users (who are presumed to be more influential or popular) and the lowest 20% as *Low centrality* users. Note that the centrality is computed just based on activity, and not necessarily troll behavior, though they might correlate with each other.

We conducted a comparative analysis between threads initiated by high centrality users and those started by low centrality. The key variable of interest was the proportion of hate speech within the comments section of each thread. By analyzing the content of replies in threads initiated by users with varying levels of centrality, we aimed to discern patterns that could shed light on the spillover effect of trolling and the role of influential users in propagating hate speech. To control for external factors that might influence the visibility and engagement of a post, such as timing and popularity, we controlled for the number of shares per post and relative time in our regression model to find the average difference in hate speech proportion in posts across these two groups.

The results of our study are depicted in Figure 7, showing a statistically significant difference in the average proportion of hate speech in replies between threads started by high centrality users (average = 0.1392 and those by low centrality users (average = 0.09481). This indicates that threads initiated by more influential users tend to attract more hate speech. We tested if the difference was statistically significant using permutation tests, a non-parametric statistical method that doesn't assume data independence. The permutation tests supported our hypothesis, confirming that discussions started by high centrality users are significantly more likely to contain hate speech in their replies. This suggests a spillover effect where the behavior of prominent users influences the overall tone of the conversation, leading to an increase in hate speech ($p < 0.001$).

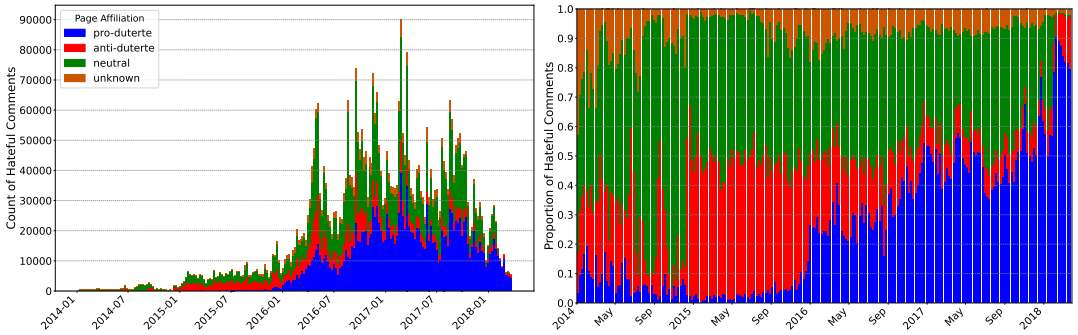

Figure 4: (a). Counts of hateful comments made by pro-Duterte supporters. (b). shows the proportion. Both plots show 7-day moving averages.

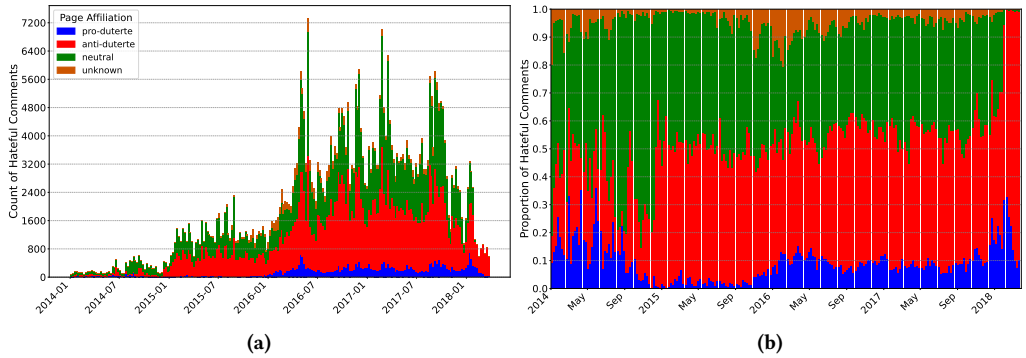

Figure 5: (a). Counts of hateful comments made by anti-Duterte supporters. (b). shows the proportion.

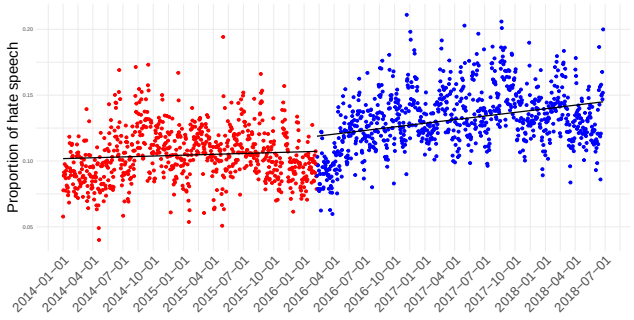

Figure 6: Interrupted time series analysis of proportion of hateful speech from the announcement of Duterte's election campaign in February 2016.

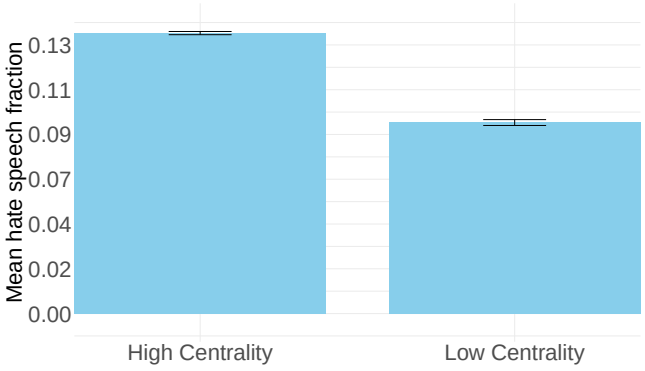

Figure 7: Mean fraction of hate speech in the top and bottom deciles. Error bars show 95% confidence intervals.

We also compared the distribution of centrality of users responding to hate speech threads by high and low centrality users. The results support the hypothesis that high centrality trolls are significantly more likely to respond to other high centrality trolls. Additional results are shown in Section A.6 in the Appendix.

6 CONCLUSION

In this study, we performed a large-scale analysis of political discourse on social media in the global south, with a focus on the Philippines. This choice of subject is not only pioneering but also timely, considering the ever-increasing role of social media in shaping political landscapes worldwide. Our research contribution extends beyond the technical achievement of developing a high-precision hate speech detection model for code-mixed Filipino text. This study addresses a fundamental gap in the field – the lack of descriptive analytics in politically and culturally complex regions like the Philippines. In doing so, it offers a template for similar studies in other parts of the global south, where such in-depth analyses are

scarce. Our findings provide a rich dataset and a methodological framework that can be replicated and expanded upon 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, although no longer readily accessible, reveals that such platforms can become breeding grounds for hate speech, particularly in the context of political discourse. Despite the 2016 Philippine elections being almost 10 years ago, the study's relevance remains high. It provides a historical lens through which we can understand current and future digital political campaigns. This longitudinal perspective is crucial for predicting and preparing for similar tactics in other contexts.

REFERENCES

[1] Shazia Akhtar and Catriona M Morrison. 2019. The prevalence and impact of online trolling of UK members of parliament. *Computers in Human Behavior* 99 (2019), 322–327.

[2] Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465* (2020).

[3] Aries A Arugay. 2022. *Foreign Policy & Disinformation Narratives in the 2022 Philippine Election Campaign.* ISEAS-Yusof Ishak Institute.

[4] Samantha Bradshaw and Philip N Howard. 2018. Challenging truth and trust: A global inventory of organized social media manipulation. *The computational propaganda project* 1 (2018), 1–26.

[5] Paul R Brass. 2011. *The production of Hindu-Muslim violence in contemporary India.* University of Washington Press.

[6] Erin E Buckels, Paul D Trapnell, and Delroy L Paulhus. 2014. Trolls just want to have fun. *Personality and individual Differences* 67 (2014), 97–102.

[7] Jason Cabañes and Jayeel Cornelio. 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* (2017), 233–252.

[8] Jan Christian Blaise Cruz and Charibeth Cheng. 2019. Evaluating Language Model Finetuning Techniques for Low-resource Languages. *arXiv preprint arXiv:1907.00409* (2019).

[9] Jan Christian Blaise Cruz and Charibeth Cheng. 2020. Establishing Baselines for Text Classification in Low-Resource Languages. *arXiv preprint arXiv:2005.02068* (2020).

[10] Nicole Curato. 2017. We need to talk about Rody. *A Duterte reader: Critical essays on Rodrigo Duterte's early presidency* (2017), 1–36.

[11] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, Vol. 11. 512–515.

[12] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing* 1, 1 (2018), 1–27.

[13] Harvard Gazette. 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/. Accessed: 2024-02-05.

[14] Aristides Gionis, Piotr Indyk, and Rajeev Motwani. 1999. Similarity Search in High Dimensions via Hashing. In *Proceedings of the 25th International Conference on Very Large Data Bases.* 518–529.

[15] Marco Giugni, Ruud Koopmans, Florence Passy, and Paul Statham. 2005. Institutional and discursive opportunities for extreme-right mobilization in five countries. *Mobilization: An International Quarterly* 10, 1 (2005), 145–162.

[16] Jessica Guynn. 2016. 'Massive Rise'in Hate Speech on Twitter during Presidential Election. *USA Today* 21 (2016).

[17] Lindsay Pérez Huber. 2016. Make America great again: Donald Trump, racist nativism and the virulent adherence to white supremecy amid US demographic change. *Charleston L. Rev.* 10 (2016), 215.

[18] Jiwan Jeong, Jeong-han Kang, and Sue Moon. 2020. Identifying and quantifying coordinated manipulation of upvotes and downvotes in Naver News comments. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 303–314.

[19] Renee Karunungan. 2023. *The role of Facebook influencers in shaping the narrative of the Duterte era.* Ph. D. Dissertation. Loughborough University.

[20] Ruud Koopmans and Jasper Muis. 2009. The rise of right-wing populist Pim Fortuyn in the Netherlands: A discursive opportunity approach. *European Journal of Political Research* 48, 5 (2009), 642–664.

[21] Christian Lamour and Renáta Varga. 2020. The border as a resource in right-wing populist discourse: Viktor Orbán and the diasporas in a multi-scalar Europe. *Journal of borderlands studies* 35, 3 (2020), 335–350.

[22] Asia Maior. 2017. The Philippines 2017: Duterte-led authoritarian populism and its liberal-democratic roots. https://www.asiamaior.org/the-journal/asia-maior-vol-xxviii-2017/the-philippines-2017.html. Accessed: 2024-02-05.

[23] Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science.* 173–182.

[24] Kathleen M McGraw. 2003. Political impressions: Formation and management. (2003).

[25] Cristina Jayme Montiel, Joshua Uyheng, and Nmanuel de Leon. 2022. Presidential Profanity in Duterte's Philippines: How Swearing Discursively Constructs a Populist Regime. *Journal of Language and Social Psychology* 41, 4 (2022), 428–449.

[26] Karsten Müller and Carlo Schwarz. 2023. From hashtag to hate crime: Twitter and antiminority sentiment. *American Economic Journal: Applied Economics* 15, 3 (2023), 270–312.

[27] Kevin Munger. 2017. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior* 39 (2017), 629–649.

[28] Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush Varshney. 2018. The effect of extremist violence on hateful speech online. In *Proceedings of the international AAAI conference on web and social media*, Vol. 12.

[29] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data* 2 (2019), 13.

[30] Jonathan Corpus Ong and Jason Vincent A Cabañes. 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* (2018).

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

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

[33] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. 2011. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web.* 249–252.

[34] Maria Ressa. 2022. *How to Stand Up to a Dictator: The Fight for Our Future.* HarperCollins.

[35] Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM conference on web science.* 255–264.

[36] Karishma Sharma, Yizhou Zhang, Emilio Ferrara, and Yan Liu. 2021. Identifying Coordinated Accounts on Social Media through Hidden Influence and Group Behaviours. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, Feida Zhu, Beng Chin Ooi, and Chunyan Miao (Eds.). ACM, 1441–1451. https://doi.org/10.1145/3447548.3467391

[37] Alexandra A Siegel and Vivienne Badaan. 2020. # No2Sectarianism: Experimental approaches to reducing sectarian hate speech online. *American Political Science Review* 114, 3 (2020), 837–855.

[38] Alexandra A Siegel, Evgenii Nikitin, Pablo Barberá, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler, Joshua A Tucker, 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 (2021), 71–104.

[39] Massimo Stella, Emilio Ferrara, and Manlio De Domenico. 2018. Bots sustain and inflate striking opposition in online social systems. *CoRR* abs/1802.07292 (2018). arXiv:1802.07292 http://arxiv.org/abs/1802.07292

[40] Fatemeh Tahmasbi, Leonard Schild, Chen Ling, Jeremy Blackburn, Gianluca Stringhini, Yang Zhang, and Savvas Zannettou. 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.

[41] Charles Tilly. 2003. *The politics of collective violence.* Cambridge University Press.

[42] Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science* 7 (2021), e598.

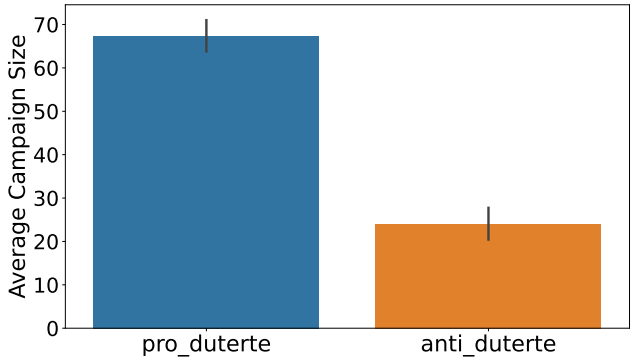

Figure 8: Average coordinated campaign size for pro and anti duterte supporters. Error bars indicate 95% confidence intervals.

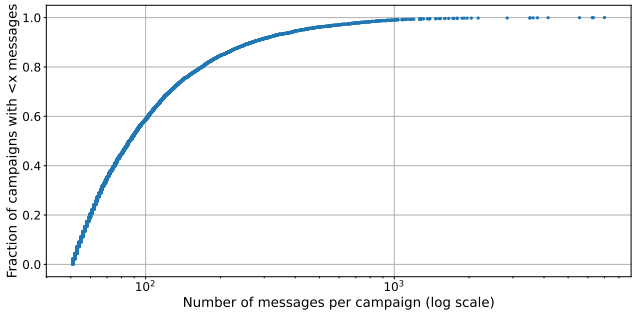

Figure 9: Coordinated campaigns size. Over 60% of the 5700 campaigns are less than a hundred messages but there are some massive campaigns with over 7000 messages.

A 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.

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

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 4 and 5.

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

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

A.4.1 Model. The conducted 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.

A.4.2 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 8.

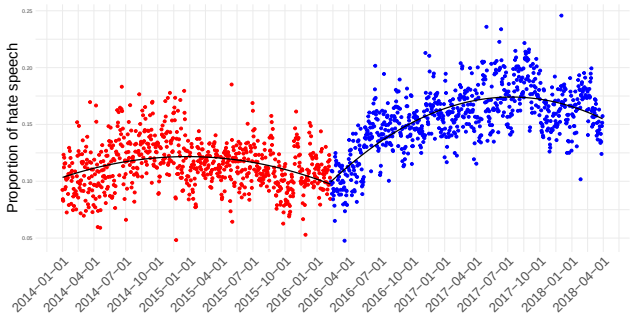

Figure 10: Interrupted Time Series Analysis of proportion of hateful speech from the announcement of Duterte's election campaign (quadratic fit).

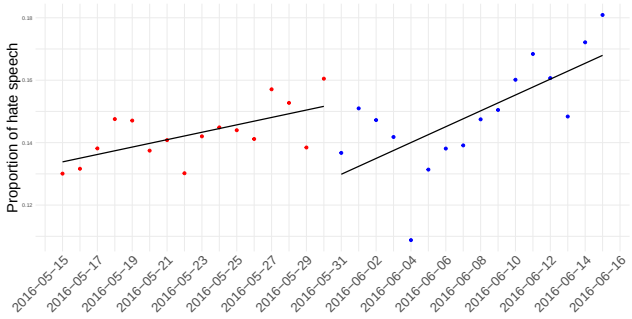

Figure 11: Interrupted Time Series Analysis of proportion of hateful speech from when Duterte justified killing of journalists (May-31-2016)

A.4.3 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. 9.

A.4.4 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.

A.4.5 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, the results, while indicating

Table 4: Hashtags by politician (part 1)

Subsection Title	Hashtags
Pro-Duterte	duteteforpresident, duterte2016, dutertecayetano, dc2016, dutertecayetano2016, ducay, ducay2016, solidurterte2016, 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

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.

A.4.6 Model 4: Unpublicized Pledge Against Profanity. On February 21, 2018, Duterte once more avowed to abstain from profanity, an

Table 5: Hashtags by 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, isupportttheconstitution, boykootrappler, upholdtheconstitution, arrestmariaressa, thenurve, terriblecult, unfollowrappler, unfollowingrappler

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 6.

A.4.7 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 Hate speech annotation

Details provided to annotators on Upwork:

- 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".

Table 6: ITS model (linear) coefficients for Figure 6 in Section 5.4

	Model 1
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 7: ITS model coefficients (quadratic), for Figure 10) in Section 5.4

	Model 1
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

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

- 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 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.

Table 8: Coeffieints for Figure 11 (quadratic model)

	Model 1
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 9: Table for Figure 12 (quadratic model)

	Model 1
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 10: 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

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

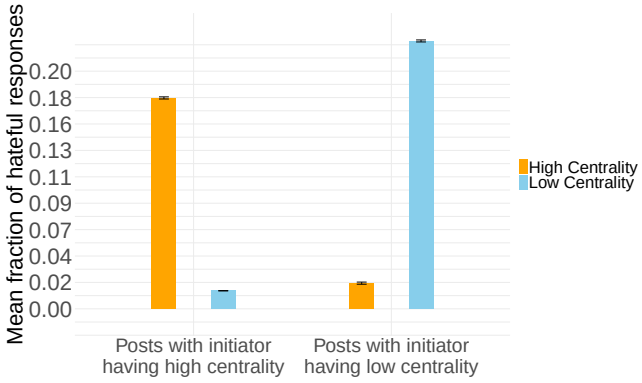

Figure 13: Mean fraction of high and low centrality responses for posts with a high and low centrality hate speech response.

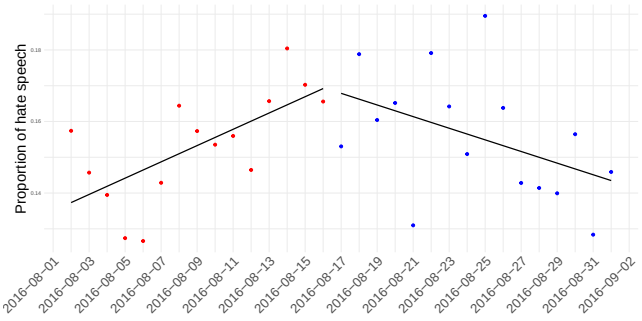

Figure 12: Interrupted Time Series Analysis of proportion of hateful speech for when Duterte attacked Leila De Lima at a press conference (Aug-17-2016)

To overcome this, we are now recruiting 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.

A.6 Spillover effects

Next, we compared threads started by high centrality users with those initiated by low centrality users, both involving hate speech, to assess the distribution of centrality among respondents in these threads. Our hypothesis was centered on determining whether hate speech posts by high centrality users tend to attract responses from other high centrality users (potentially other trolls or influential users) or from 'normal', less active users. The findings, illustrated in Figure 13, provide crucial insights: (i) In threads where a high centrality user started a hate speech post, the majority of responses came from other high centrality users. This pattern suggests a sort of clustering among active or influential users, where they are more likely to interact with each other. (ii) The tendency of high centrality users to engage predominantly with other high centrality users hints at the presence of an echo chamber effect, and (iii) Conversely, when low centrality users initiated hate speech threads, they predominantly attracted responses from other low centrality users. This indicates that less active or influential users are more likely to interact within their own circles, mirroring the pattern observed among high centrality users.