THE EFFECT OF FACEBOOK'S CONTENT MODERATION ON HATE CRIME IN MYANMAR

By
Seng Moon Ja

Submitted to
Central European University
Department of Economic and Business

In partial fulfillment of the requirements for the degree of
Master Economic Policy in Global Markets

Supervisor: Mats Köster

Vienna, Austria
2023

Abstract

This paper investigates the relationship between content moderation and hate crime. The growth of hate content on social media platforms, especially targeting vulnerable communities, is being called under scrutiny. Despite efforts by social media companies to censor such content, evidence of the effectiveness of such content moderation measures is limited. In the context of Facebook's content moderation efforts in Myanmar in 2018, this research investigates the causal relationship between online content moderation and real-world crimes. With the difference in different methods, the paper finds that the townships with a higher percentage of Facebook users increase hate incidents after content moderation. Townships with a higher percentage of Facebook users witnessed an 11.59% rise in hate incidents. The results suggest that content moderation alone may not be a comprehensive solution to combat hate crimes, however social media still act as a space for the sharing of hate.

Acknowledgements

I would like to express my sincere gratitude to my thesis supervisor, Professor Mats Koster, for his guidance and constant support throughout this research endeavor. Also, I am grateful to the Civil Society Leadership Award for giving me the scholarship that enabled me to attend this program and author the thesis. Thanks everyone from my cohort.

Finally,
Thank you, Atwi.
Thank you, Anu, and Awa.
Let justice flow like a river.

Table of contents

1 List of Figures, Tables, or Illustrations

I. Introduction

The synonym for Facebook is “The Internet” in Myanmar. As half of the population of the country is an active user of Facebook, it is the best place to promote products and services. With 85% of social media penetration, most of the businesses in Myanmar are present on Facebook to engage or advertise their products to their customers and potential customers. Besides, majority of the Myanmar online shoppers are buying products from Facebook than other E-commerce. To create more effective marketing content, Facebook allows businesses to use algorithms to customize their advertising campaigns and target specific audiences to optimize customer engagement. Facebook bring a lot of blessings until, these algorithms have been weaponized by the military to spread misinformation, fuel extremism, hate speech and false information that resulted from genocide, violence, and displacement of nearly a million of the Rohingya people since 2012. Myanmar people realized it is a double-edged sword, the power of it lies in the hands of its users.

In 2018, following the UN Independent International Fact-Finding Mission on Myanmar's accusation of Facebook spreading hatred against the Rohingya (Miles, 2018), Facebook acted by removing hate figures and organizations from Myanmar. This included individuals and groups such as Wirathu, Thuseitta, Parmaukkha, Ma Ba Tha, the Buddha Dhamma Prahita Foundation, as well as 20 military-linked individuals, organizations, and Commander-in-Chief Min Aung Hlaing (Facebook Report on Myanmar). Although warnings from civil society organizations, media, and academia, Facebook failed to act appropriately to address the problem until 2018. Despite the introduction of additional content moderation measures and the increased hiring of Burmese language specialists to review content in 2018, it is reported that hateful content is still spreading on the platform. According to Global Witness's experiment result [1], Facebook ads service accept to promote eight ads in Burmese that contain real-life content of hate speech promoting violence and genocide against the Rohingya sourced from a UN fact-finding mission (Global Witness 2022). There is no empirical evidence on the effectiveness of content moderation on hate crimes in Myanmar.

In terms of content moderation in social media platforms, YouTube started removing user-generated content in around 2006 (Gillespie 2019), Facebook released content moderation rules (Klonick, K.,2018b) and Twitter introduce its suspension policy in 2015 (Twitter, 2015). Content moderation aims to stop the dissemination of offensive or unlawful content, such as hate speech, violence, harassment, misinformation, and other types of inappropriate or unappealing content. However, there is debate about introducing content moderation. Others claim platforms do not moderate content enough to create a safe and inclusive environment, while others worry about online potential censorship and infringement on free speech. The effectiveness of these policies in lowering online hate and attenuating its violent effects must be determined before judging their social acceptability.

Therefore, this paper examines the impact of Facebook's content moderation on hate crimes in Myanmar. Firstly, it investigates whether the implementation of Facebook's content moderation measures leads to a decrease in real-world hate incidents targeting the Rohingya and minority groups. The study uses a difference-in-differences methodology exploiting differential exposure to Facebook to analyze the effects of content moderation. Secondly, the paper explores the

[1] Global Witness article on digital thread (2022). Retrieved from https://www.globalwitness.org/en/blog/exposing-social-media-platforms-failures-to-protect-their-users/

relationship between Facebook usage and offline hate crimes. Furthermore, it examines the association between the followers of the Young Men Buddhist Association Facebook group and hate incidents.

The study examines the impact of Facebook's content moderation on hate crimes targeting Muslims and minorities, exploiting variations in township-level exposure to the Facebook platform. If the implementation of Facebook's content moderation leads to a decrease in hate crimes, I would expect to observe a reduction in the likelihood of hate incidents occurring in townships. By incorporating township fixed effects and controlling for a vector of township-level characteristics in the model, the findings indicate that the introduction of content moderation resulted in an increase in hate incidents in townships with a higher percentage of Facebook users. The estimates suggest that townships with a higher percentage of Facebook users, on average, have a 47% higher likelihood of occurrence of protests against Rohingya and Muslims after content moderation, which is contrary to expectations. The estimates also suggest that townships with a higher percentage of Facebook users and YMBA Facebook followers increase 20% and 22 % higher likelihood of the occurrence of the protest led by Ma Ba Tha (Organization for the Protection of Race and Religion), the far-right nationalist Buddhist.

Additionally, I use “hate incidents” as one variable which refers to various forms of real-world incidents, including hate protests, riots, and explosions specifically targeted towards Rohingya and Muslims. I study the link between Facebook and hate incidents by controlling with a vector of township-level characteristics. Such as general media exposure, socio-economic factors, and demographic controls. Without these controls, the relationship between Facebook and hate incidents is negative. However, with these controls, the model estimates suggest that townships with a higher percentage of Facebook users, on average, have a 4% to 11% higher likelihood of occurrence of hate incidents against Rohingya and Muslims.

The paper's findings are further validated by conducting robust checks using similar variables such as mobile phone and internet penetration. After the implementation of content moderation measures, townships with a higher percentage of mobile phone or internet users are more likely to experience hate protests, on average, by 52% and 47%, respectively. As part of the placebo check, the study investigates whether this hate crime trend is similar to other types of crime (e.g. theft, kidnapping, etc.) or whether it is influenced by other factors. I examine the relationship between Facebook content moderation and different types of crimes unlikely to be influenced by hate content or Facebook usage. Internet users are used as a proxy for Facebook users at the regional level because crime rate data is only available at a regional and annual level and Facebook user data is only available at the township level for Myanmar. The findings reveal that regions with a higher percentage of internet users observe an increase in riots, hate protests, and protests led by Ma Ba Tha. However, the interaction term between internet users and content moderation does not yield a statistically significant coefficient across the various models. To study heterogeneous effects, I studied the content moderation effect on townships with more than 50% of the population using Facebook. The study found that there is a significant increase of 47% in the likelihood of hate protests occurring after content moderation measures are implemented in this township. To examine the varying effects in separate locations, particularly those subjected to two types of interventions—curfews and internet cutouts—I selected two townships with a high percentage of hate incidents. The paper's findings reveal that even after the internet was cut off in the selected township, there was an increase in hate incidents. The township with an imposed curfew does not yield a statistically significant coefficient.

The paper is organized as follows: In section II, the literature relevant to the topic is introduced. Sections III and IV present the background and data used in the study. Section V explains the empirical strategy employed, while Section VI presents the main findings. Section VII discusses the mechanism underlying the paper's analysis, and finally, Section VIII provides the conclusion and section IX provides policy recommendation.

II. Literature Review

Much of the current literature on the effect of social media on real-world outcomes is increasing. This paper contributes to the growing literature on the impact of social media on xenophobia and hate crime. Firstly, this paper adds up contribution to the evidence of establishing the relationship between social media and real-world crimes. Previous studies have explored the relationships between social media and hate crime (Müller and Schwarz 2021) and found out social media platforms not only spread hateful ideas but motivate real-life actions. To better understand the mechanisms of higher penetration of social media and its effects on ethnic hate crime,(Bursztyn, Rao, and B. Y. n.d.) analyzed the causal effect of social media penetration on hate crimes in Russia from 2007-2015. A recent study by Williams et al. (2019)suggests that online hate victimization is part of a wider process that starts on social media and spreads to physical spaces.

Secondly, this paper’s findings are relating to the literature on the role of social media in social movements and protests worldwide. Social media platforms have been used to spread messages, organize protests, and bring attention to their causes. During the Hong Kong protests in 2019, Facebook was used to disseminate information about the protests and to encourage people to take part. A study by Brym, R., Godbout, M., Hoffbauer, A., Menard, G., & Zhang, T. H. (2014) found that individuals who used Facebook were more likely to participate in protests compare to the people who primarily received the protest information from face-to-face and phone contact. Similarly, a study by Larson, J. M., Nagler, J., Ronen, J., & Tucker, J. A. (2019) found that those who used social media platforms (Twitter) to search for protest-related information were more likely to participate in protests than those who did not use social media. A study by (Xie et al. 2017)found that users of social media developed a sense of solidarity and collective action against the government and other social forces. A study by Lee, S. (2018) found that individuals used social media as a platform to coordinate and organize protests, as well as to spread news and information regarding protests. Overall, there seems to be much evidence to indicate that social media can create collective action such as protests and riots. Relative to this literature, this paper not only explores the relationship between social media and hate crime, but it will also explore the relationship between collective action and hate crimes.

Thirdly, the findings of this paper contribute to our understanding of the potential consequences of social media use and content moderation on hate crimes. Jiménez Durán et al. (2022) studied the effect of content moderation on online and offline hate and found that content moderation significantly reduced hate crime in towns with more far-right Facebook users in Germany. Collectively, these studies outline the critical role of content moderation in real-world problems. To the best of my knowledge, this paper is one of the few papers which study the effect of content moderation on real-world problems such as hate crime in Myanmar.

Finally, the paper's result also relates to a number of papers on the relationship between mobile phones and collective action. Diamond (2010) studied the impact of mobile phones on collective action, and Lee, S. (2018) the role of social media in protest participation. In terms of social media in Myanmar, few have studied the role of social media in Myanmar, Aricat and Ling (2016) studied the promise and threat of mobile communication and the internet. Bergren and Bailard (2017) examined ICT and ethnic conflict in Myanmar.

III. Background

Myanmar is a hugely diverse country with 55.02 million in 2022 and it is home to over 100 ethnicities. The big ethnic Bamar make up 68% of the population and the remaining 32% are various ethnic minority groups. Buddhists make up 88% of the total population, 6% Christian, 4% Muslim and less than 1% Hindu and Animist.

Since 1962, the country has been under a military dictatorship run by the State Peace and Development Council (SPDC). The SPDC has suppressed pro-democracy activists and restricted freedom of expression and assembly. The military rulers have also been accused of human rights abuses and have imposed restrictions on the media. The military junta has maintained its grip on power through a variety of tactics, including a 2008 Constitution that gives it control over all branches of government. From 2015 to 2019, the country had been undergoing a transition to a civilian-led government, although the military held considerable influence. In 2021, the country experienced a military coup, as the military declared a state of emergency and arrested several civilian leaders, including State Counsellor Aung San Suu Kyi. The conflict between ethnic groups has been going on for 7 decades and many armed conflicts fought were between the Burmese military and different ethnic armed groups. These conflicts have resulted in hundreds of thousands of people being displaced and the deaths of thousands more. The military has been accused of committing human rights abuses against ethnic minorities, including the Rohingya, Kachin, Karen, and Shan. The Country was isolated from the rest of the world for decades and connected with the international community in 2011.

The Internet penetration rate in Myanmar is 45.9%, representing 25.28 million people with access to the Internet in 2022. Additionally, the number of mobile phone subscribers and sim cards in the country is estimated to be around 69.43 million, which amounts to a mobile phone penetration rate of 127.2% in 2021. In 2011, only 1% of the Myanmar population had access to the internet. This indicates that most of Myanmar's population is now connected to the rest of the world through the Internet after being isolated by the military dictatorship for over nearly 7 decades. The growth of internet connectivity in Myanmar has also contributed to an increase in the number of people using Facebook in the country. In early 2022, there are currently around 21 million (37.5% of the total population) active users on the platform in Myanmar. Facebook become the digital tea shop of Myanmar, the place where every piece of information is shared. It has become the main source of information platform for economic, political, educational, and social awareness topics. As well as it is a platform for citizens to share their views, gossip, and experiences besides the news. The average hour Myanmar people use the internet is 2.4 hours per day and 81.25 % of social media users as shown in figure (1) are using Facebook and other platforms is less than 9%.

Facebook is the source of social capital building in developing countries Raza, S. A., Qazi, W., & Umer, A. (2017). It is widely said that Facebook is “the internet” in Myanmar Asher, S. (2021). Whitten-Woodring, J., Kleinberg, M. S., Thawnghmung, A., & Thitsar, M. T. (2020) examined that 40% of people who access the internet said that they use Facebook as the primary source of news. Wai, H. L. (2019) examined that young people choose Facebook to start an online business in

Myanmar with zero capital and business owners investing more in Facebook Ads to promote their products. Wittekind, C. T., & Faxon, H. O. (2022) studied that information from Facebook plays a major role in increasing and decreasing the price of land in Myanmar. These papers tell the magnitude of the role of Facebook in Myanmar. These studies pointed out that Facebook play a vital role in socio-economy of Myanmar. The information is searched on Facebook and not on Google. The development of Myanmar's telecommunication infrastructure enabled affordable SIM cards, and affordable mobile data prices in 2015. The combination of cheap SIM cards, mobile data prices and Facebook's free basis to create an account, Facebook became a powerful tool for connecting people with shared interests, promoting social movements, and the place to access the news and information. Facebook users in Myanmar grew from 8.4 million in 2014 to 23.4 million in 2020. This represents an increase of 179% in the last 6 years.

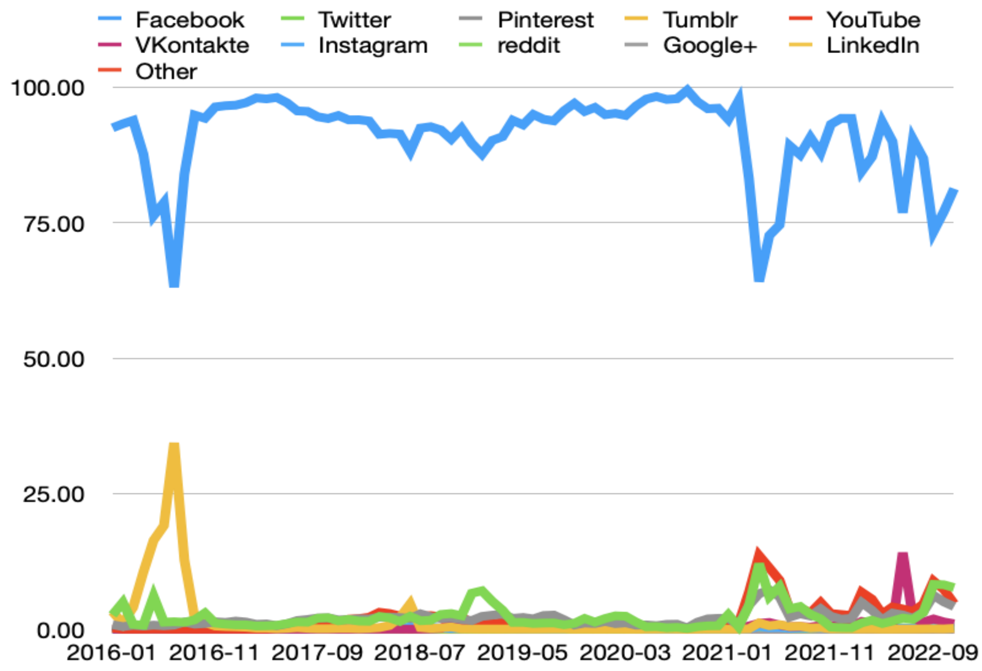


Figure 1. Different Social Media Platform Usage in Myanmar

Source: Statcounter.com

A. Facebook and Rohingya Genocide

The problems of Facebook in Myanmar are related to the spread of fake news from the Buddhist nationalist movement led by Buddhist monks. The Buddhist nationalist organization (Ma Ba Tha) which can be translated as the Organization for the Protection of Race and Religion was a network with more than 100,000 members present from different township levels. Facebook is the key weapon to raising the campaigns built on a stigmatizing and often dehumanizing narrative against Muslims. The lack of access to reliable information and the prevalence of conspiracy theories, hates speech and inflammatory rhetoric have fueled the fire of anti-Rohingya sentiment throughout the country. This has resulted in genocide, violence and displacement of nearly a million of the Rohingya people since 2017. One of the fake news that was spread by the extremist and anti-Muslim monk, Ashin Wirathu (Ma Ba Tha) in 2014 and a post which claimed a Buddhist girl had been raped by Muslim men went viral on Facebook. As a result of the violence that followed, two people were killed by a crowd targeting those accused of involvement. It turned out that the monk's accusation was completely false after a police investigation.

Reuters reported that when analyzing 1000 posts from Facebook, most of the content is Anti Rohingya comments, images, and videos in the Burmese language. The hate speech includes a call for Rohingya to be shot, set on fire, and fed to pigs; suggesting genocide – "kill Rohingya as the way Hitler killed Jews"; pornographic anti-Muslim images; description of the group as being dogs, maggots, and rapists.

B. The weaponization of Facebook in propaganda

Facebook has been weaponized as a propaganda tool by the Burmese military (a.k.a Tatmadaw), Ma Ba Tha. Ma Ba Tha was founded in 2013 as a rebranding of the 969 movement, a nationalist campaign that urged the boycott of Muslim-owned businesses in Myanmar. The 969 Movement, a Buddhist nationalist movement, emerged in response to perceived Islamic influence encroaching upon predominantly Buddhist Myanmar (Burma). Ma Ba Tha has expressed support for the military's political and economic interests, and some of its members have been appointed to positions of power within the government. There have also been allegations that the military has used Ma Ba Tha to promote its own nationalist agenda and to suppress dissent. Ma Ba Tha has regularly trained its members on Facebook features and content creation and used secret Facebook groups to coordinate their efforts.

Initially, they create news pages and pages on Facebook featuring Myanmar pop stars, models, and other celebrities, like a beauty queen who recited military propaganda. Through these pages, they attract more people, then the pages were taken over and dedicated to military-related news. As it become distribution channels, they started to spread false news, hoax, trolls, rumors, and fire-up posts, often aimed at Myanmar Muslims, opposite political opponent Daw Aung San Su Kyi and other ethnic armed groups. Another well-known technique is the use of fake accounts (usually with few followers), they post venomous comments beneath mainstream news posts or pages that have many followers and spread false information.

To operate this specific job such as creating, promoting content, and engaging in a social media platform, a large group of military officers were sent to Russia to get the training. The training is one of the curriculums of Defense Service Academy Information Warfare training. As of October 2018, it is estimated that 700 officers are working in those units.

To attract users' attention to stay longer on the platform, Facebook algorithms are designed to prioritize content from friends and family, as well as posts that are most engaging or trending and show in users' newsfeeds. By making users engage with the most engaging posts, targeted Ads are advertised or personalized. At that time, the most engaging posts turned out to be the posts from Ma Ba Tha and the military, which they were working on as their professions. While they are spreading fake news and hate speech constantly, Facebook fails to detect and remove hateful or offensive content instead its algorithm automatically promoted those contents assuming that these contents are useful content to increase engagement of the users. This dangerous algorithm greatly contributed to brutalities committed by the Myanmar military against the Rohingya people.

C. Facebook's Response

In 2018, 20 military-linked individuals and organizations were banned, including Commander-in Chief Min Aung Hlaing, for his role in committing human rights violations(Frankel, R.,2022, Facebook Report on Myanmar). In 2020, six Tatmadaw Coordinated Inauthentic Behavior networks were removed. In 2018, Myanmar hate figures and organizations on Facebook, including Wirathu, Thuseitta, Parmaukkha, Ma Ba Tha and the Buddha Dhamma Prahita Foundation were banned from Facebook (Facebook Report on Myanmar). As of December 2018, Facebook has removed 425 Facebook Pages, 17 Facebook Groups, 135 Facebook accounts and 15 Instagram accounts in Myanmar for engaging in coordinated inauthentic behavior on Facebook.

On the other hand, Facebook autodetected reporting tools had a low capacity to detect misinformation and hate speech. Besides, one of the biggest hindrances to detecting content is also because 90% of the phones in Myanmar use Zawgyi which is the nonstandard encoding for characters in the Burmese language. After Facebook find out the root cause of failure to detect words that are against Facebook Community Standard, they start hiring more Myanmar language experts to review content. As of June 2018, they hired over 60 from 2 to four moderators in 2015(Facebook Report on Myanmar).

In addition to Facebook response, there are main intervention on hate crime in Myanmar. As of Figure 2, Panzagar Campaign led by one of the civil society organizations start the movement that fight against hate speech on Facebook started on April 2014. After mob attack in the second largest of the country, local government impose curfew in July 2016. After rising number of riots and protests between the ethnic arm group, the military and Rohingya, the government cut off the internet in 2019 in Rakhine state.

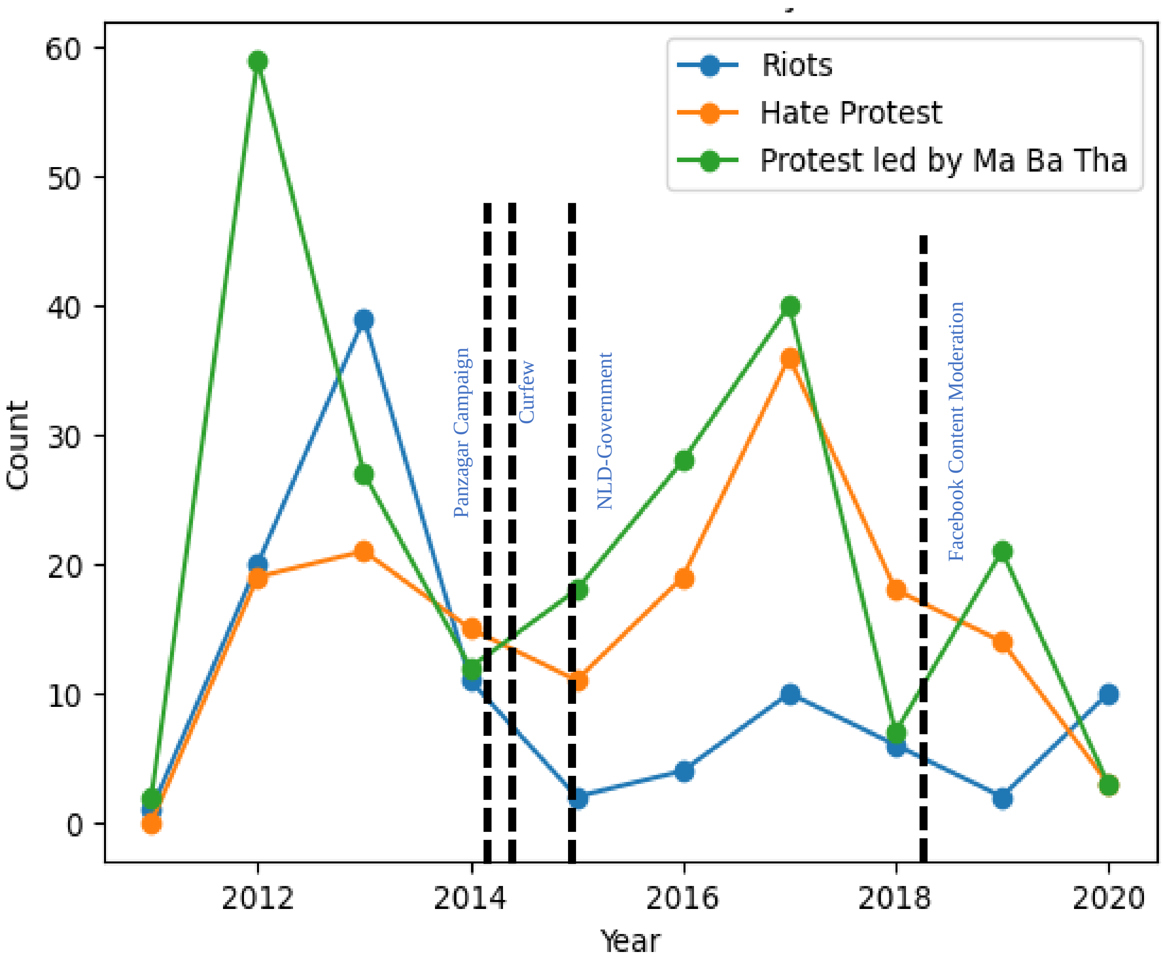


Figure 2 The Occurrence of Riots, Protest led by Ma Ba Tha and Hate protest

IV. Data

The data is constructed on the hate crime and Facebook activity in Myanmar with baseline control data. I combine data from different data sources. I organize (1) country-level conflict data (2) the number of Facebook users by township level using Facebook Marketing API, (3) hand collection of 600 YMBA Facebook group followers (3) township-level electricity coverage data, (4) economy data by township level and (5) internet access and mobile phone access data by township level, (6) township level demographic and education level data, (7) regional crime rate. Due to the military coup in February 2022, there was a significant increase in the number of riots, protests, and explosions therefore, I choose the period between 2010 and 2020 which is before Military Coup.

A. Country level conflict Data

To study the impact of Facebook on actual violence I use country-level data on conflict from Armed Conflict Location and Event Data Project (ACLED). The data is available from 2010 to present and includes information of political violence such as Battles, Protests, Violence against civilians, Strategic developments, Explosions/Remote violence, and Riots. Among these event types, I retrieve Riots, Protests and Explosions/Remote violence data by requesting ACLED's API. According to ACLED's code book 2021, 'Riots' are violent events where demonstrators or mobs engage in disruptive acts, including but not limited to rock throwing, property destruction, etc. 'Riots' are violent events where demonstrators or mobs engage in disruptive acts, including but not limited to rock throwing, property destruction, etc. But in the definition of 'Riots', contraries to armed groups, rioter do not use weapons such as guns, knives, or swords however 'crude bombs' used in the actions. Riots are classified into two sub-event types such as 'violent demonstration' and 'mob violence.' Violent demonstration involves violent actions such as vandalism, road-blocking, burning tires of destructive behavior. Mob violence is a mob is defined as "a large crowd of people, especially one that is disorderly and intent on causing trouble or violence" (ACLED code book 2021). The data also includes the information of two associated actors regarding the violent incident. Both associated actor1 and associated actor 2 can be the victims of an attack or the socio-political affiliation of demonstrators or ethno-religious identity of a civilian victim. In the data, out of many actors, frequent associated actors are Buddhist Group (Myanmar), Muslim Group (Myanmar) and Rakhine Ethnic Group (Myanmar).

I specifically analyze the data pertaining to hate crime-related conflicts, focusing on the main actors involved. Considering the nature of hate crimes, which often involve conflicts between Buddhist and Muslim, Buddhist and Christian, Rakhine and Rohingya, as well as Buddhist and Rohingya communities, I filter the dataset to include only those instances associated with these key actors. Then, I examine the primary causes of riots for each of the selected data points. If a particular incident is found to be unrelated to hate or religion-related conflicts, it is subsequently dropped from the analysis. This approach ensures that the dataset is refined to include only relevant cases that align with the research focus on hate crime dynamics. Table A.2: Hate Incidents from appendices section present details.

Furthermore, I examine the prevalence of protests directed against Rohingya, Muslims, and Christians. Even, in ACLED data, it is classified as "peaceful protest: when demonstrators are engaged in a protest while not engaging in violence or other forms of rioting behavior and are not faced with any sort of force or engagement" (ACLED code book 2021), I code as hate conflict related event that is creating tension as cause of violence. Table A.2: Hate Incidents from appendices section present details.

Moreover, in order to gain deeper insights into potential underlying mechanisms, I delve into the data regarding protests instigated by monks since the initiation of the 969 movement in 2012. A significant portion of these organizing events are facilitated through the platform of Facebook, which has served as a prominent medium for mobilization and coordination.

B. Township Level Facebook User and

In order to examine the potential relationship between Facebook users and real instances of hate crimes, I investigated using the Facebook Marketing API which allows to gather data on the number of Facebook users at the city or zip code level. Through this API, I obtained three key metrics for estimating reach: estimate reach, lower bound reach, and upper bound reach. The estimate reach metric provides an approximation of the potential audience size that a Facebook ad campaign could potentially reach. The lower and upper bound reach metrics indicate the minimum and maximum number of Facebook users that the ads might potentially reach, respectively Table A.2. I focused specifically on examining the number of Facebook users at the township level; therefore, I use the upper bound reach user information as a reference. This allowed me to gain insights into the potential maximum Facebook user size within each township.

C. YMBA Facebook Group Follower Data

In order to investigate the potential correlation between the number of Facebook users associated with extreme nationalist Facebook pages or groups and hate crimes, the Young Men Buddhist Association (YMBA) Facebook Group was selected as a subject of study due to its resemblance to Ma Ba Tha which played prominent role in inciting hate crimes and protests in Myanmar. I could not use Ma Ba Tha's pages and groups as all the pages and groups associated with Ma Ba Tha were deleted on January 2018 due to content moderation measures, and YMBA Facebook's group is the closet alternative for the analysis. Even with YMBA Facebook group, the privacy settings of the YMBA Facebook group prevented me from scraping all group members data. Consequently, I resorted to examining every post within the group and visiting the profiles of individuals who interacted with group posts (e.g., through likes, shares, and comments). Particularly, I manually gathered location data of each of group members. Nonetheless, after Facebook's introduction of the Profile Lock option on March 31st, 2021, many profiles began utilizing this feature, restricting access to their data. Consequently, I could only collect 600 users' data who make it publicly available.

D. Socioeconomic Data

I acquired socioeconomic data at the township level from the Myanmar Information Management Unit, which was accessible through the website (https://themimu.info/vulnerability-in-myanmar). Among the various indicators available, I specifically focused on the following: Percentage of Highest Education: At least Middle school (age 25 and above from population) and Literacy rate, Population, Population density and Wealth rank by township level. These variables were selected as they provide valuable insights into the socioeconomic characteristics of each township. The data obtained from these indicators will be used as control variable in analyzing.

To assess how many people are using mobile phones and computers or the internet at home, I use the percentage of using mobile phone by household and the percentage of using internet or having computer at home data. As I observed that townships utilizing electricity for lighting tend to exhibit a higher preference for using internet with mobile phone or computer compared to those that do not have electricity. Therefore, I obtained data on the percentage of households with access to electricity for lighting as an additional factor for analysis.

E. Crime Rate Data

To test placebo check, regional level yearly crime rate data of Myanmar from Myanmar Statistical Information Service is used. The crime data include the total crime of different crime such as murder, dacoity, robbery, burglary, kidnap, rape, theft, animal theft, forgery and coinage, narcotic offences, gambling, and trafficking in person.

F. Summary Statistics for Main Variable

The summary statistics table is presented in Table (1). The table compares the summary statistics for different variables based on whether the percentage of Facebook User is below or above the median in the township level. Main outcome variables include Protests Led by Ma Ba Tha, Riots, and Hate Incidents. Since 2011, the 969 movement has been on the rise, and in 2013, protests were led by Ma Ba Tha. From the available data, I specifically focus on protests led by Ma Ba Tha and Buddhist monks. Even if the data does not explicitly mention Ma Ba Tha, I include protests that are thematically similar and categorize them as "Protests led by Ma Ba Tha”. For "Riots", incidents related to religiously motivated mob violence and violent demonstrations are selected. Hate Incidents include a combination of riots, hate Protests which include both by Ma Ba Tha and ordinary civilians, and targeted explosions against Rohingya, Buddhist, Rakhine, and Muslim individuals.

The main variables include the number of mobile subscriptions per household, the number of internet subscriptions per household at the township level, and the number of YMBA group followers in the township. When comparing means of these variables, townships falling below the median of Facebook User in township level display lower than its townships above median value.

Regarding the supplementary control variables, Table (1) presents the differences between townships below and above the median percentage of Facebook users. Townships falling below the median exhibit lower levels of socioeconomic status across multiple dimensions. They have limited electricity availability, lower Wealth Rank, lower ownership of TVs and radios, and lower literacy rates. However, it is noteworthy that the mean occurrence of Hate Incidents is higher among these townships.

Table 2 Summary statistics

	Facebook User > Median			Facebook User < Median		
Variable	**Obs**	**Mean**	**SD**	**Obs**	**Mean**	**SD**
Hate Incidents						
Riots	202	0.32	0.47	201	0.19	0.41
Explosion	202	0.06	0.25	201	0.00	0.00
Hate Protest	202	0.30	0.46	201	0.48	0.52
Protest Led by Ma Ba Tha	202	0.31	0.46	201	0.72	0.47
Hate Incidents (Riots + Explosion + Hate Protest)	202	0.94	0.37	201	0.70	0.52
Main Variables						
Facebook User	202	16846.04	13636.49	201	2390977.11	2467501.56
Facebook User (%)	202	0.07	0.05	201	0.71	0.27
Household Internet User (%)	202	0.03	0.02	201	0.24	0.17
Household Mobile Phone (%)	202	0.21	0.09	201	0.65	0.19
YMBA Facebook Follower	15	2.07	1.58	187	143.80	139.50

V. Empirical Strategy and Main Results

To investigate whether Facebook content moderation reduced hate crime or not, I estimate difference-in-difference method as follows.

$$Hate\ Incident_{it} = \beta_0 + \beta_1 * fb_i + \beta_2 * content_m + \beta_3 * (fb_i * content_m) + \beta_4 * Control_i + Township\ FE_i + \varepsilon_{it}$$

In this equation, the symbols represent the following:

$\text{Hate Incident}_{\text{it}}$, the dependent variable represents the number of Hate Incidents in township i in the time of event *t*. The variable fb_{i} is a dummy variable that takes the value of 0 if the percentage of Facebook Users in a specific township *i* is below the median of Facebook user population. Otherwise, if it is equal to or above the median, fb_{i} takes 1. Therefore, the coefficient β_1 measures the difference in Hate Incidents, on average, between townships above the median in terms of Facebook users (fb_{i} = 1) and those below the median (fb_{i} = 0). It captures the average effect of being above the median in Facebook users on the occurrence of hate incidents, while controlling for other variables in the equation.

β^2 coefficient measures the average effect of content moderation that took place in 2018 when Facebook removed pages and groups associated with extreme nationalist monks and the military on the hate incidents $\text{Hate Incident}_{\text{it}}$ while holding other variables constant. Content Moderation is also dummy variable; it takes 1 if event date is after 2018-02-28, otherwise 0. It captures the average effect of after content moderation on the occurrence of hate incidents, while controlling for other variables in the equation. The coefficient β^3 represents the effect on $\text{Hate Incident}_{\text{it}}$ from the main independent variables, which is the interaction between the dummy variable of Facebook user ($\mathrm{fb_i}$) and the content moderation (content_m). This interaction can be interpreted as the effect of Facebook User ($\mathrm{fb_i}$) on $\mathrm{Hate\ Incident_{it}}$ differing depending on whether content moderation occurred before or after 2018.

The empirical strategy involves comparing the changes in the number of hate incident and hate crimes before and after the introduction of content moderation by Facebook in Myanmar, which involved the removal of extreme nationalist and military-related Facebook pages and groups. Intuitively, a decrease in the average occurrence of hate crimes is expected after the implementation of Facebook content moderation. The regression model includes a full set of township fixed effects, which controls for any baseline differences in the number of hate crime incidents across townships. The control variables used in the regression model include the general media exposure factors, socioeconomic factors, and demographic factors.

To measure the effect on hate crimes, I control variations in several key factors associated with higher exposure to hate speech on Facebook. Specifically, I examine the prevalence of Facebook users by controlling internet accessibility, mobile phone usage for internet access, and reliance on electricity for lighting purposes. These factors are presumed to be indicators of populations that are more susceptible to encountering hate speech on Facebook. Intuitively, I expect that regions characterized by a greater concentration of Facebook users may have a higher prevalence of hate speech on the platform would experience a disproportionate impact from hateful content originating on Facebook, compared to regions with a lower prevalence. By employing a difference-in-differences with ordinary least squares approach, I aim to explore whether, the content moderation has impact on hate crimes while all factors that may influence the occurrence of hate incidents before and after the implementation. Township fixed effect is applied to control for township-specific characteristics that may influence the occurrence of hate incidents. These measures ensure that the analysis considers unobserved variability between townships. Robust standard errors are clustered by township in all specifications to get more reliable inference.

VI. Findings

A. Content Moderation and Hate Crimes

Table (2) presents the regression results obtained from estimating Equation (1), which aims to analyze the impact of content moderation, Facebook user, and YMBA follower presence in the township on hate incidents such as protests led by Ma Ba Tha, riots, and hate protests. The Facebook User variable is a binary variable that takes the value 1 if townships are above the median of the Facebook user population which is 24% of the township population and otherwise it takes 0. Likewise, the Content Moderation variable is a binary variable that takes the value 1 if the incident occurs after 28th February 2018 and otherwise it is 0. The analysis is conducted at the township level, and the study includes different sets of supplementary control variables. 403 observations are considered in the analysis. To examine the relationship between the presence of Ma Ba Tha's followers on Facebook, YMBA Facebook group followers is considered as the closest alternatives. This choice allows for a comparative exploration of the impact of these follower groups on hate incidents. To ensure more accurate estimations and account for potential confounding factors, all models include controls for general media exposure, demographic factors, and socioeconomic indicators. Additionally, township fixed effects are included in all models to control township-specific characteristics that may influence the occurrence of hate incidents. These measures ensure that the analysis considers unobserved variability between townships.

Protests Led by Ma Ba Tha: Columns (1) to (2) present the regression results for variables related to protests led by Ma Ba Tha. These protests include hate protests as well as protesting the newly appointed government or leader. The inclusion of these variables aims to examine the magnitude of Facebook's role in organizing such protests through its platform. The analysis reveals that townships with a higher median level of Facebook followers show a negative relationship between Facebook User and the occurrence of protests led by Ma Ba Tha. However, this relationship is not statistically significant. Likewise, townships with more YMBA followers experience a negative impact on the likelihood of protests. Specifically, in townships where both the median level of Facebook users and YMBA followers is higher, the likelihood of Ma Ba Tha-led protests is significantly reduced by 27% and 21% respectively, after implementing content moderation measures. When considering the combined effect of Facebook User and Content Moderation in townships with a higher number of Facebook users and YMBA followers, it is found that after content moderation, the occurrence of Ma Ba Tha protests increases by 20% and 22%, respectively. These results suggest that while the presence of Facebook users alone does not have a statistically significant effect on protests led by Ma Ba Tha, the presence of YMBA followers does significantly decrease the likelihood of such protests. Additionally, the implementation of content moderation measures leads to an increase in protests.

Riots: Columns (3) and (4) investigate the connection between Riots and Facebook users and YMBA followers before and after the content moderation. Riots are violent events where demonstrators or mobs engage in disruptive acts. A higher number of Facebook users negatively impacts the estimated chance of riots. It is consistent with my baseline summary statistic, which reveals that riots occurred in areas with fewer Facebook users, as well as in areas with lower general media exposure as shown in Table A.1 from appendix section. The interaction effect between a Facebook user and content moderation is insignificant. But existence of higher YMBA followers in township could increase the occurrence of riots in the townships by 35% and statistically significant at 1% level.

Table (2). Protests led by Ma Ba Tha, Riots, Hate Protest and Facebook User and YMBA follower in Township level

	Protest led by MaBaTha		Riots		Hate Protests	
	Facebook User	YMBA Follower	Facebook User	YMBA Follower	Facebook User	YMBA Follower
	(1)	(2)	(3)	(4)	(5)	(6)
Facebook_User	-0.022		-0.026*		0.121***	
	(0.019)		(0.014)		(0.020)	
YMBA_Follower		-0.220***		0.359***		-0.308***
		(0.013)		(0.012)		(0.020)
Content_Moderation	-0.275***	-0.217***	0.021	-0.030	-0.310***	-0.155
	(0.076)	(0.066)	(0.090)	(0.065)	(0.116)	(0.095)
Facebook_User:Content_Moderation	0.208*		-0.113		0.477***	
	(0.115)		(0.107)		(0.144)	
YMBA_Follower: Content_Moderation		0.221**		-0.006		0.371***
		(0.089)		(0.077)		(0.134)
Township Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
General Media Exposure [4]	Yes	Yes	Yes	Yes	Yes	Yes
Demographic controls [7]	Yes	Yes	Yes	Yes	Yes	Yes
Socioeconomic factors [5]	Yes	Yes	Yes	Yes	Yes	Yes
Observations	403	403	403	403	403	403
R^{2}	0.546	0.543	0.527	0.523	0.370	0.339
Note:	*p<0.1; **p<0.05; ***p<0.01					

Note: This table presents regression coefficient estimates for the interaction of Facebook users and YMBA Facebook page followers on Protest led by Ma Ba Tha, Riots, and Hate Protest, as specified in equation (1). The dependent variables, Protest led by Ma Ba Tha, Riots, and Hate Protest, are transformed using the natural logarithm (log (1+number of events)). All independent variables are dummy variables, taking the value 1 if the percentage of the variable is above the median, and 0 otherwise. The models include all control variables, township fixed effects. Robust standard errors are clustered by township in all specifications.

Hate protests: In columns (5) and (6) of Table (2), the models present the effect of Facebook content moderation on hate protests. Hate protests are both organized by Ma Ba Tha and ordinary citizens against Muslims, Rohingya and some minorities. Column (5) illustrates that there are more hate protests when there are more Facebook users and the interaction term between Facebook users and content moderation has a coefficient of 0.477 (p<0.01), showing a significantly positive interaction effect. This shows that hate protests are increasing despite Facebook's content moderation. Even though the frequency of hate protests has decreased since content moderation was enacted, the hate protests have increased in higher number of Facebook user in townships. This can be interpreted as townships with a higher number of Facebook users have, on average, a 47% higher likelihood of experiencing hate incidents. Townships with higher number of YMBA f decrease the occurrence of hate protests in the townships however, the combined effect of YMBA follower and content moderation increase hate protests. In 2018, Facebook implemented content moderation efforts to remove Facebook pages and groups associated with hate icons such as Ma Ba Tha and the military. Despite these measures, hate protests persist.

The R-squared values range from 0.34 to 0.54, indicating that the models explain 34% to 54% of the variance in incidents. Robust standard errors are clustered by township in all specifications.

B. Facebook User and Hate Incidents

Table (3) presents the regression results from estimating Equation (1) which examines the relationship between hate incidents and Facebook User before and after the content moderation in 2018, at the township level with varying sets of different sets control variables and analysis 403 observations.

Hate Incidents. In this paper, the term "hate incidents" refers to various forms of real-world problems, including hate protests, riots, and explosions specifically targeted towards Rohingya and Muslims. By combining these incidents as one variable, I hope to assess how Facebook has influenced real-world problems. Regardless of the severity of each incident, they all represent tangible issues that can occur through Facebook. In Column (1), the model examines the relationship between Facebook User in townships and the occurrence of Hate Incidents, considering only township fixed effects as control.

The coefficient on Facebook User is negative (coefficient = -0.0349) and statistically significant at the 1% level, indicating a significant association between Facebook users and hate incidents at the township level. This suggests that an increase in the number of Facebook users is associated with a decrease of 3% in hate incidents. In Column (2), the coefficient for the interaction term between Facebook User and Content Moderation is presented. The coefficient of Content Moderation is positive (0.0107) across models (2) to (6), indicating a potential explanatory effect on hate incidents. However, it is not statistically significant with various types of additional controls, including general media exposure, socioeconomic factors, literacy rate, and demographic factors. This suggests that there is no clear relationship between content moderation and hate incidents. Similarly, the coefficient for the interaction term between Facebook User and Content Moderation in columns (2) to (6) is 0.0418, and it is not statistically significant, indicating that the presence of content moderation on Facebook does not have a significant direct effect on hate incidents in townships.

In Column (3), the model investigates the impact of Facebook usage on hate incidents while incorporating the control variable of general media exposure. The coefficient of Facebook User in this model is statistically significant at the 10% level, with a coefficient of 0.0416. This suggests that the presence of the general media exposure variable, which includes the share proportion of households with TV, the share proportion of households with radio, the percentage of household internet users, and the percentage of household mobile phone users, influences the relationship between Facebook users and hate incidents. Without this controls the coefficient is negative but with this general media exposure control it changes to positive, there is possibility that higher exposure to general media is positively correlated with both Facebook user and occurrence of hate incidents.

Table (3) Facebook User and Hate Incidents in Township

	Dependent variable: Hate Incidents					
	(1)	(2)	(3)	(4)	(5)	(6)
Facebook_User	-0.0349***	-0.0392***	0.0416*	0.0926***	0.0087	0.1159***
	(0.0000)	(0.0118)	(0.0226)	(0.0247)	(0.0259)	(0.0198)
Content_Moderation		0.0107	0.0107	0.0107	0.0107	0.0107
		(0.0175)	(0.0176)	(0.0176)	(0.0177)	(0.0179)
Facebook_User:Content_Moderation		0.0418	0.0418	0.0418	0.0418	0.0418
		(0.0932)	(0.0938)	(0.0941)	(0.0943)	(0.0956)
Township Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
General Media Exposure control [4]			Yes			Yes
Socioeconomic factors control[6]				Yes		Yes
Demographic controls [7]					Yes	Yes
Observations	403	403	403	403	403	403
R^{2}	0.4461	0.4476	0.4476	0.4476	0.4476	0.4476
Note:	*p<0.1; **p<0.05; ***p<0.01					

Note: Table reports the estimated coefficients from a regression of Hate_Incidents on percentage of Facebook User in township. The independent variable 'Facebook User' is a dummy variable. If the percentage of Facebook_User is above the median value, it takes the value 1, otherwise 0. Content_Moderation is also dummy variable; it takes 1 if event date is after 2018-02-28, otherwise 0. The dependent variables which is Hate_Incidents is transformed using the natural logarithm (log (1+number of events)).

By using media exposure as a control variable, the model explains that a higher number of Facebook users is associated with an increase in hate incidents within the township. These incidents can be protests, riots, or explosions. Controlling general media exposure allows us to isolate the specific impact of Facebook usage by considering the potential influence of other media types on people's beliefs and attitudes towards hate incidents. For instance, exposure to TV and radio broadcasts can increase exposure to news, discussions, and other content that can shape attitudes towards Rohingya and Muslims. Moreover, higher levels of internet and phone usage may facilitate access to social media platforms like Facebook, where hate content can be disseminated. By keeping these factors constant, the model can examine the distinctive effect of Facebook users on hate crimes. However, content moderation effects have no significant with these controls.

Columns (4) estimates the same model with only controlling for socioeconomic factors such as female and male workforce participations, availability of electricity in the townships, wealth rank, urban population proportion and literacy rate. The models' main coefficient, which is the interaction term between Facebook user and content moderation is not significant, the coefficient (0.0926) of Facebook user is positive and statistically significant at 1% level. In column (1), the models provide the relationship between Facebook User and hate incidents without any controls, but when I include socioeconomic factors in column (4) it gives positive coefficient, this means that the occurrence of hate incidents is not solely due to direct influence of Facebook User but rather it reflects the combined influences of control variables and Facebook users in the township. Similarly, in column (5), I include only demographic factors such as the age group structure, population density, and sex ratio of the township as controls. The coefficient of Facebook users shows no statistically significant relationship with hate incidents when considering the effects of demographic variables that may influence the relationship. This lack of significance could be attributed to the presence of multicollinearity between population variables and Facebook users.

Since my dataset consists of only 403 observations, obtaining more data would be necessary to address this issue. Even after removing both population density and population variables and retesting, the results still indicate a statistically insignificant coefficient. When including all the control variables related to various aspects in columns (6), the model estimates that having a Facebook user population higher than the median is associated with an approximate 11.59% increase in the expected occurrence of hate incidents. However, content moderation and the interaction between content moderation and Facebook users do not have a significant impact on hate incidents.

All models include township fixed effects to avoid unobserved characteristics specific to each township that may influence hate incidents. These measures ensure that the analysis considers unobserved variability between townships. The R-squared values range from 0.4461 to 0.4467, indicating that the models explain 44% of the variance in hate incidents.

VII. Mechanism

The difference-in-difference (DID) model identifies the causal impact of Facebook's content moderation by comparing the mean occurrence of hate incidents before and after 28th February 2018, both for townships with fewer and higher Facebook users. With this model, I try to look at two types of hate incidents such as hate protests which protest without violence and hate crime which involve violence.

The first mechanism is to investigate whether increased social media penetration can encourage the occurrence of hate crime. According to Müller and Schwarz (2021) and Williams et al. (2019), social media penetration increases hate crimes through coordinating hate crimes or learning about other people's desire to engage in such activities through the platform. To investigate this association, the sample is separated into two groups based on the number of Facebook users: those who use Facebook less than the median Facebook user and those above median. The models' analysis includes different sets of control variables. Table (3) demonstrates that the occurrence of hate incidents can be better explained when controlling for general media exposure and socioeconomic factors. However, the impact of demographic factors on hate crimes within the townships is found to be weak. This paper 'evidence contributes to the findings of social media penetration increased hate incidents. In Tables (A.6), similar models are explored using the number of mobile phone users and internet users as alternative measures. The findings indicate that higher penetration of mobile phones decreases hate incidents however it increases the protest organized by Ma Ba Tha. This study is aligned to previous evidence from Manacorda & Tesei (2018) which suggests that mobile phones contribute to the increase of social issues.

Secondly, this paper's mechanism investigates if the people who use social media are more exposed to current social issues and more likely to participate in hate protest. This mechanism is based on the existing evidence by Larson, J. M., Nagler, J., Ronen, & Tucker, J. A. (2019) who found that those who used Twitter to look for information on protests were more likely to take part in demonstrations than people who did not use the social media platform. To investigates this, I applied equation (1) with dependent variable such as hate protest and protests led by Ma Ba Tha and found out that there is a strong positive relationship between number of Facebook user and the occurrence of protest in that township. The results are also in line with evidence by (Xie et al. 2017) who argued that Users of social media have acquired a sense of solidarity and taken collective action. The number of protests is higher in township with higher numbers of Facebook users despite I add control for general media exposure, socioeconomic factors, and demographic. However, in general, collective actions such as protests strikes and demonstrations usually take place in metropolitan areas than villages. At time same time, usually number of Facebook users is usually higher in this area.

Third mechanism investigates if higher number of YMBA Facebook page followers in township increase real world crimes. According to the existing evidence DellaVigna et al.(2014) suggest that nationalist propaganda on radio can increase the prevalence of violence against minorities. Müller & Schwarz (2021) suggests that social media can act as a propagation mechanism for violent crimes by enabling the spread of extreme viewpoints. As well as Rucht and Neidhardt (2018) says that individuals used social media as a platform to coordinate and organize protests, as well as to spread news and information regarding protest. This mechanism suggests that having more YMBA Facebook page followers increases protests in that township. With this mechanism, this paper was expected to see higher YMBA Facebook page follower will lead to higher number of hate crimes. However, instead of higher riots and explosions against Rohingya, higher number of YMBA Facebook page followers increase hate protests and protests led by Ma Ba Tha but decrease riots.

Final mechanism is the effect of Facebook’s content moderation decrease hate crime as Jiménez Durán et al.(2022) explore the effect of content moderation on online and offline hate after Netwrok Enforcement Act in Germany. The paper found that NetzDG reduced hate crime against refugees in townships with more far-right Facebook users. In terms of hate protests in this paper, even after the content moderation in 2018, the township with higher number of Facebook user increase hate protests but there is no evidence on the relationship between content moderation and riots. After the Facebook's content moderation, hate protests still increase in Myanmar. According to Global Witness's experiment result [2], Facebook ads service accept to promote eight ads in Burmese that contain real-life content of hate speech promoting violence and genocide against the Rohingya sourced from a UN fact finding mission. This implies that despite Facebook's efforts to strengthen its content moderation measures and remove hate icons such as Ma Ba Tha and the military's propaganda pages, the effectiveness of these measures is still inadequate.

Limitation of the study: while the initial analysis reveals a negative link between the number of Facebook users and hate incidents, the coefficient becomes positive once I control with general media exposure and socioeconomic factors. More research is required to identify variables that influence and factors that contribute to hate crimes. For example, it would be great if I could get the data such as ethnic and diversity rate in township level.

A. Heterogeneous effects

If the effect of an increase measure of content moderation depends on the Facebook User, I would expect to see heterogeneity of my estimates by the higher number of Facebook user in townships. Therefore, instead of creating dummy of 1 if it is above median Facebook user which is 24% of township population, I adjust the dummy equal to 1 if Facebook user is above 50% of township

[2] Global Witness article on digital thread (2022). Retrieved from https://www.globalwitness.org/en/blog/exposing-social-media-platforms-failures-to-protect-their-users/

population. The impact of content moderation on protests led by Ma Ba Tha and riots is found to be statistically insignificant. However, there is a significant increase of 47% in the likelihood of hate protests occurring after content moderation measures are implemented.

If the effect of increased content moderation measures depends on the regions and their characteristics, for instance, the Rakhine region in the western part of Myanmar, which is the epicenter of conflicts, there may be different outcomes. In 2019, due to a surge in conflicts and protests against Rohingya and Muslims, the Myanmar government shut down the internet. I expected to observe a decrease in the occurrence of protests organized through Facebook after 2019 by including an internet shutdown dummy variable set to 1 for 2019 onwards. However, as shown in Table (A.10), the occurrence of hate protests and hate protests led by Ma Ba Tha did not decrease; in fact, they increased even after 2019. However, riots decreased after 2019. While the internet shutdown impacted the occurrence of violence, hate protests continued to increase.

Furthermore, in April 2014, activists in Myanmar initiated a campaign called "panzagar" or "flower speech" to counter hate speech. Following a deadly conflict in Mandalay in July 2014, the police implemented a curfew in the second largest city of Myanmar. To examine the combined effects of civil society and police interventions, I employed April 2014 as a dummy variable set to 1 for this period and onwards in Table (A.11). Additionally, I focused on the Mandalay region to analyze its specific effect. yielded results similar to the content moderation period and the internet shutdown in 2019. Despite these interventions, hate protests continued to occur without a statistically significant impact. No significant results were observed for riots.

B. Robustness

Table (4) presents a comprehensive analysis of the impact of Facebook content moderation on different types of crimes that are unlikely to be influenced by hate content or Facebook usage. The objective is to validate the findings by focusing on offenses that are not specifically affected by Facebook's content moderation controls. Considering the crime rates outlined in Section IV (E), the total crimes rate include kidnapping, burglary, robbery, murder, dacoity, and others. Due to the availability of crime rate data only at the regional and annual levels, the analysis is conducted at the regional level rather than the township level. Therefore, content moderation is adjusted with 2018 instead of event date “2018-02-28”.

Table (4) Comparing Hate Crime and other Crimes.

	Riots	Hate Protests	Protest led by Ma Ba Tha	Explosion	Crime Rate
	(1)	(2)	(3)	(4)	(5)
Internet_User	0.148***	0.471**	0.205***	0.029	-0.000
	(0.025)	(0.236)	(0.050)	(0.024)	(0.105)
Content_Moderation	-0.004	-0.994	-0.244	0.136	0.416***
	(0.112)	(1.269)	(0.228)	(0.141)	(0.085)
Internet_User:Content_Moderation	-0.154	1.327	0.059	-0.136	0.728
	(0.146)	(1.368)	(0.293)	(0.141)	(0.617)
Buddhist_proportion	0.046***	-0.084***	-0.083***	0.014***	-0.301***
	(0.002)	(0.020)	(0.004)	(0.002)	(0.009)
Islam_proportion	-0.020***	-0.086***	-0.016***	-0.004***	-0.033***
	(0.001)	(0.005)	(0.001)	(0.000)	(0.002)
Christian_proportion	-0.052***	0.089***	0.082***	-0.011***	0.204***
	(0.003)	(0.030)	(0.006)	(0.003)	(0.013)
Unemployment_Rate	0.073***	0.565***	0.088***	0.040***	-0.080***
	(0.000)	(0.003)	(0.001)	(0.000)	(0.000)
Electricity	0.007***	0.044***	0.013***	0.002***	-0.044***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)
Radio_proportion	-0.001***	-0.028***	-0.010***	-0.000***	-0.000
	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)
TV_proportion	-0.012***	-0.066***	-0.017***	-0.003***	-0.008***

	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)
Sex_Ratio	-0.020***	-0.070***	-0.032***	-0.000	-0.033***
	(0.000)	(0.002)	(0.001)	(0.000)	(0.001)
Literate	0.002***	-0.037***	-0.013***	-0.001***	-0.085***
	(0.000)	(0.002)	(0.000)	(0.000)	(0.001)
Region Fixed Effect	Yes	Yes	Yes	Yes	Yes
Observations	150	150	150	150	150
R^2	0.354	0.511	0.506	0.245	0.782

*p<0.1; **p<0.05; ***p<0.01

Note: Table reports the estimated coefficients from a regression of Hate_Incidents on percentage of Internet User in region. The independent variable 'Internet User' is a dummy variable. If the percentage of Internet_User is above the median value, it takes the value 1, otherwise 0. Content_Moderation is also dummy variable; it takes 1 if event date is after 2018, otherwise 0. All dependent variables except explosion are transformed using the natural logarithm (log (1+number of events)).

As regional-level data on Facebook users is not accessible for Myanmar, attempts to obtain it through the Facebook Marketing API were unsuccessful. As an alternative, the number of internet users in each region is used. The findings indicate that regions with a higher percentage of internet users experience an increase in riots, hate protests, and protests led by Ma Ba Tha. However, the interaction term between internet users and content moderation does not yield a statistically significant coefficient across the various models.

In Table (A.7), I explore the correlation between internet users and hate incidents at the township level, and the results align with the findings for Facebook users. As of 2020, 42 % and 36.8% of Myanmar population are using internet and Facebook respectively (World Bank data). However, for my analysis, I utilize data from the 2014 census when only 9% of the country's population had access to the internet. The relationship between internet users and hate incidents is examined and presented in Table (A.7). The result suggests that higher internet users increase protest led by Ma Ba Tha and hate protests.

Furthermore, I explore the impact of content moderation on mobile phone users and hate incidents. Table (A.8) demonstrates a strong positive association between mobile phone users and hate incidents. A higher percentage of mobile phone users leads to an increase in protests led by Ma Ba Tha and hate protests, while it reduces the occurrence of riots. Following the implementation of content moderation measures, both protests led by Ma Ba Tha and hate protests experienced an increase.

VIII. Conclusion

Facebook helps to drive Myanmar economic activities by connecting businesses and people and lowering marketing barriers. It brings greater transparency and accountability to the country's political system and aids the spread of democracy. Simultaneously, it has been used as a weapon to spread hate speech and promote ethnic violence, and genocide, and ultimately, the displaced millions of people in Myanmar.

This paper investigates whether Facebook's content moderation reduce hate crimes in Myanmar. My findings suggest that hate incidents such as protest against Rohingya and Muslims increase after the content moderation. Despite Facebook’s effort to remove hateful contents and increase hiring Burmese content moderator, the prevalence of hate crimes has not decreased but on average, have a 47% higher likelihood of occurrence of protests against Rohingya and Muslims after content moderation. Additionally, the paper explores the relationship between Facebook usage and offline hate crimes and estimates suggest that townships with higher percentage of Facebook user, on average, have 4% to 11% higher likelihood of prevalence of hate incident. Out of hate incident, the occurrence of riots decreased in Rakhine state however protests increase even after the internet cut out in 2019.

This evidence suggests that while the existence of Facebook may not be the primary factor leading to tragic events in Myanmar, it serves as a catalyst where the flames of hatred ignite and spread rapidly. Like a wildfire, it is not easy to extinguish once it starts to spread.

IX. Policy Recommendation

Digital Policing: Policymakers often underestimate or ignore the significance of online crimes. The evidence highlights that the origin of crime on social media should not be underestimated. The findings indicate that the implementation of content moderation in 2018 did not reduce hate protests.

Therefore, it is recommended to introduce stricter policing mechanisms in countries like Myanmar. Myanmar instantly become the victims of weaponization of Facebook just after it emerged from the strict dictatorship of 7 decades. The country was too young to understand the dark side of the digital world. To prevent the same type of tragic incident in the future, government regulation must be tightened. On November 1, 2022, the EU Digital Services Act (DSA) was entered into force. EU said, "It aims to protect the digital space against the spread of illegal content, and to ensure the protection of users' fundamental rights". Similarly, on September 1, 2017, in response to hate speech online against refugees in Germany, the country introduced the "Netzwerkdurchsetzungsgesetz," which translates as the Network Enforcement Act. This enforcement allows for imposing significant penalties of up to 50 million euros on social media companies if they fail to remove hateful content within 24 hours. Such type of policy should be introduced in Myanmar. However, the current military government is the main actor who is spreading hateful content, therefore, a more feasible alternative to that recommendation is building resilience among citizens.

Build Resilience: Firstly, to build resilience amongst the population and prevent ethnic conflicts, digital media literacy training should be more often provided by NGOs and civil society groups. For, in Myanmar, the current military government is the actor that weaponizes Facebook, and the responsibility of building resilience among conflicts is now on the shoulders of NGOs and civil society groups. These actors should raise awareness campaigns on the weaponization of social media both online and offline. As well as more seminars, networks and events for social cohesion-building activities should be organized. Especially to the areas with low socioeconomic areas. According to the findings, riots can occur in places with lower socioeconomic development, but protests are more often in urban areas.

Myanmar should have a mechanism that encounter weaponized information, hate speech, and fake news if this has already spread. Most often, hate speech and weaponized content can be spread during elections. Frequently, toxic comments have been spreading in the comment sections under the news of civil wars and conflict news. Civil society and NGOs should work together across countries to identify shared patterns of weaponized information and detect their tactic and encounter this hate contents through the same platform.

Appendices

Table A.1: Summary Statistics

	Facebook User > Median					Facebook User < Median				
Variable	**Obs**	**Max**	**Mean**	**Min**	**SD**	**Obs**	**Max**	**Mean**	**Min**	**SD**
Hate Incidents										
Riots	202	1.00	0.32	0.00	0.47	201	2.00	0.19	0.00	0.41
Explosion	202	1.00	0.06	0.00	0.25	201	0.00	0.00	0.00	0.00
Hate Protest	202	1.00	0.30	0.00	0.46	201	2.00	0.48	0.00	0.52
Protest Led by Ma Ba Tha	202	1.00	0.31	0.00	0.46	201	2.00	0.72	0.00	0.47
Hate Incidents (Riots + Explosion + Hate Protest)	202	3.00	0.94	0.00	0.37	201	2.00	0.70	0.00	0.52
Main Variables										
Facebook User	202	99300.00	16846.04	1000.00	13636.49	201	5500000.00	2390977.11	5000.00	2467501.56
Facebook User (%)	202	0.24	0.07	0.01	0.05	201	1.14	0.71	0.25	0.27
Household Internet User (%)	202	0.06	0.03	0.00	0.02	201	0.60	0.24	0.01	0.17
Household Mobile Phone (%)	202	0.40	0.21	0.01	0.09	201	0.84	0.65	0.23	0.19
YMBA Facebook Follower	15	5	2.07	1	1.58	187	308	143.80	1	139.50
Additional Control Variables										

Own TV Proportion (%)	202	0.67	0.28	0.08	0.12	201	0.96	0.73	0.43	0.19
Own Radio Proportion (%)	202	0.57	0.31	0.20	0.08	201	0.66	0.33	0.12	0.12
Share Urban Population (%)	202	0.37	0.18	0.03	0.09	201	1.00	0.78	0.05	0.27
Electricity (%)	202	0.57	0.22	0.04	0.13	201	0.92	0.72	0.19	0.21
Literacy Rate (%)	202	0.97	0.80	0.65	0.11	201	0.99	0.94	0.61	0.05
Male Labor Force (%)	202	0.87	0.71	0.58	0.05	201	0.78	0.69	0.64	0.03
Female Labor Force (%)	202	0.62	0.35	0.21	0.06	201	0.55	0.40	0.24	0.06
Wealth Rank	202	245.00	80.39	5.00	70.32	201	312.41	284.28	124.00	32.76
Population	202	531799	278570.10	67523	145399	201	7360703	3134040	20039	3319614
Population Density	202	358.59	184.40	11.91	101.99	201	19995.48	9619.72	5.89	8660.29
Share Population 0 to 14 (%)	202	40.60	31.52	21.24	3.55	201	41.36	22.31	10.92	5.41
Share Population 15 to 64 (%)	202	77.09	62.88	55.36	2.94	201	80.32	71.41	55.39	4.34
Share Population 64 + (%)	202	10.16	5.60	1.67	1.78	201	9.45	6.28	3.24	1.57

Note: This table shows the maximum, mean, minimum and standard deviation and number of observations of main outcome, variables, and additional controls.

Table A.2: Hate Incidents

event_date	sub_event_type	notes	assoc_actor_1	assoc_actor_2
01 July 2013	Mob violence	On 1 July 2013, Buddhist rioters continued rioting, burning down a Muslim owned home in Thandwe, Rakhine state. No fatalities.	Rakhine Ethnic Group (Myanmar); Buddhist Group (Myanmar)	Muslim Group (Myanmar)
01 July 2014	Mob violence	On 1 July 2014, in Chanayethazan Township, Mandalay, a group of monks and Buddhist lay people rioted, targeting Muslim owned shops and homes. Police fired rubber bullets into the air to disperse the crowd. Several people were injured.	Buddhist Group (Myanmar)	Civilians (Myanmar); Muslim Group (Myanmar)
01 May 2013	Mob violence	On 1 May 2013, in Win Kite village, a group of 100 Buddhists armed with sticks threatened Muslim villagers, saying they would burn the village and kill them. Police intervened and pushed the group back. No fatalities.	Buddhist Group (Myanmar)	Civilians (Myanmar); Muslim Group (Myanmar)

Table A.3: Occurrence of Hate Incidents

YEAR	Riots	Hate Protest	Protests led by monks
2011	1.0	0.0	2.0
2012	20.0	19.0	59.0
2013	39.0	21.0	27.0
2014	11.0	15.0	12.0
2015	2.0	11.0	18.0
2016	4.0	19.0	28.0
2017	10.0	36.0	40.0
2018	6.0	18.0	7.0
2019	2.0	14.0	21.0
2020	10.0	3.0	3.0

Figure A.1 Number of YMBA Facebook Follower in each region

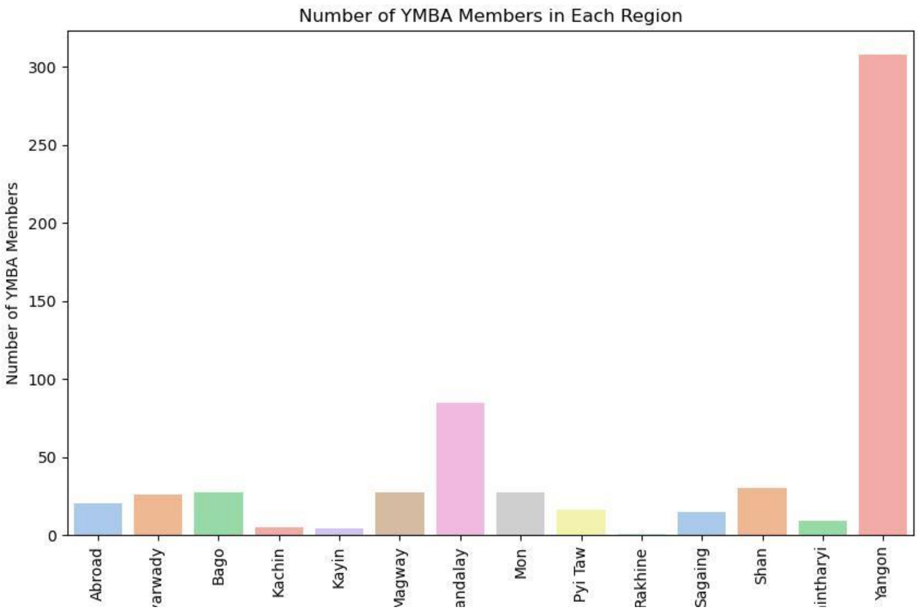
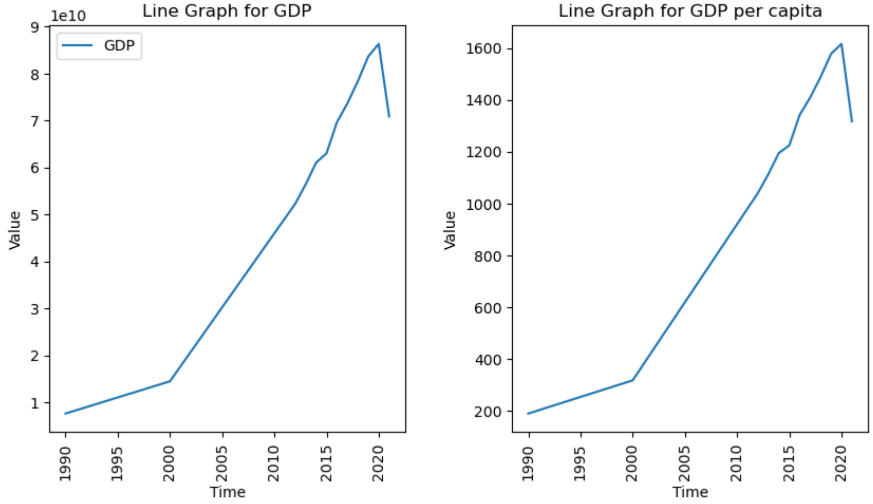


Table A.4: Facebook Marketing API and Township level Facebook User

City-Key	Region	Township	Estimate Reach	Estimate Reach Lower bound	Estimate Reach Upper bound	Type
1453681	Yangon	Ahlone	699	1800	2100	city
1453691	Ayeyarwady	Ahtaung	4668	6900	8100	city
1454121	Ayeyarwady	Athok	15184	22400	26400	city
1454256	Magway	Aunglan	15361	19500	22900	city
1454347	Sagaing	Ayadaw	166	1000	1000	city

Figure A.2: GDP and GDP per capita of Myanmar overtime



In 2012, SIM card prices in major cities such as Yangon and Mandalay dropped from over a thousand dollars to a few hundred dollars, making them more affordable, but still unattainable for many. With the growth of internet connectivity, social media usage also increased. In 2014, when Telenor and Ooredoo began operating in Myanmar, individuals waited in line for hours to purchase SIM cards that were priced at approximately one dollar.

Figure A.3: Internet user and mobile subscription

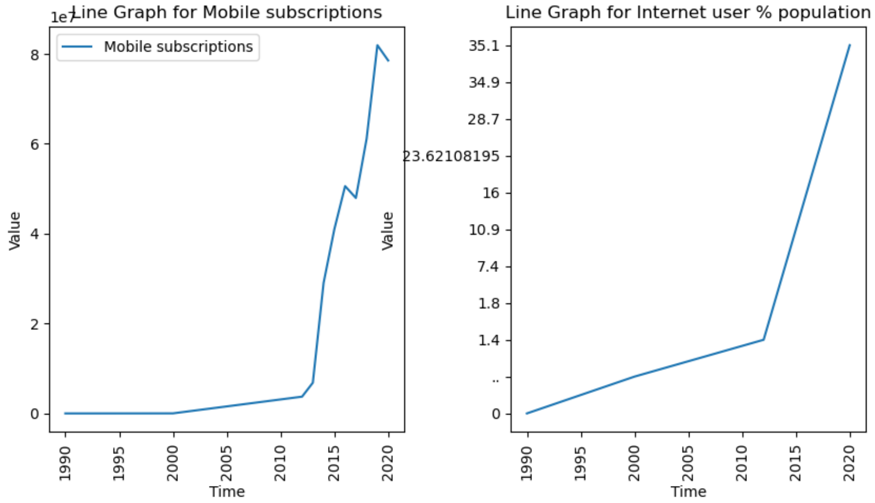


Table A.5: Facebook User and Riots in Township

	Dependent variable: Riots_					
	(1)	(2)	(3)	(4)	(5)	(6)
Facebook_User	-0.1946***	-0.1791***	-0.0520***	-0.0080	-0.0415**	-0.0008
	(0.0000)	(0.0191)	(0.0194)	(0.0187)	(0.0197)	(0.0169)
Content_Moderation		0.0211	0.0211	0.0211	0.0211	0.0211
		(0.0878)	(0.0884)	(0.0887)	(0.0888)	(0.0900)
Facebook_User:Content_Moderation		-0.1134	-0.1134	-0.1134	-0.1134	-0.1134
		(0.1042)	(0.1048)	(0.1052)	(0.1053)	(0.1068)
Township Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
General Media Exposure control [4]			Yes			Yes
Socioeconomic factors control [6]				Yes		Yes
Demographic controls [7]					Yes	Yes
Observations	403	403	403	403	403	403
R^{2}	0.5224	0.5269	0.5269	0.5269	0.5269	0.5269
Note:	*p<0.1; **p<0.05; ***p<0.01					

Table A.6: Hate Incidents and Facebook User, Internet User and Mobile Phone

	Dependent variable: Hate_Incidents		
	(1)	(2)	(3)
Facebook_User	0.1159***		
	(0.0198)		
Internet_User		0.0301	
		(0.0206)	
Mobile_Phone			-0.0566**
			(0.0262)
Content_Moderation	0.0107	-0.0106	0.0107
	(0.0179)	(0.0326)	(0.0179)
Facebook_User:Content_Moderation	0.0418		
	(0.0956)		
Internet_User:Content_Moderation		0.0909	
		(0.0948)	
Mobile_Phone:Content_Moderation			0.0418
			(0.0954)
Township Fixed Effect	Yes	Yes	Yes
All controls [17]	Yes	Yes	Yes
Observations	403	403	403
R^{2}	0.4476	0.4494	0.4476
Note:	*p<0.1; **p<0.05; ***p<0.01		

Table A.7: Protest led by Ma Ba Tha, Riots, Hate Protest and Internet User in Township level.

	Protests led by Ma Ba Tha	Riots	Hate_Protests
	(1)	(2)	(3)
Internet_User	-0.068***	-0.014	-0.053*
	(0.024)	(0.022)	(0.029)
Content_Moderation	-0.266***	0.020	-0.322***
	(0.072)	(0.087)	(0.115)
Internet_User:Content_Moderation	0.198*	-0.116	0.520***
	(0.116)	(0.106)	(0.140)
Township Fixed Effect	Yes	Yes	Yes
General Media Exposure [4]	Yes	Yes	Yes
Demographic controls [7]	Yes	Yes	Yes
Socioeconomic factors [5]	Yes	Yes	Yes
Observations	403	403	403
R^2	0.545	0.527	0.379
Adjusted R^2	0.414	0.391	0.200
Residual Std. Error	0.268 (df=312)	0.237 (df=312)	0.305 (df=312)
F Statistic	5.204*** (df=90; 312)	2.030*** (df=90; 312)	5.627*** (df=90; 312)
Note:	*p<0.1; **p<0.05; ***p<0.01		

Table A.8: Protest led by Ma Ba Tha, Riots, Hate Protest and Mobile Phone User in Township level.

	Protests led by Ma Ba Tha	Riots	Hate_Protests
	(1)	(2)	(3)
Mobile_Phone	0.046*	-0.106***	0.006
	(0.024)	(0.017)	(0.025)
Content_Moderation	-0.275***	0.021	-0.310***
	(0.076)	(0.090)	(0.116)
Mobile_Phone:Content_Moderation	0.208*	-0.113	0.477***
	(0.115)	(0.107)	(0.144)
Township Fixed Effect	Yes	Yes	Yes
General Media Exposure [4]	Yes	Yes	Yes
Demographic controls [7]	Yes	Yes	Yes
Socioeconomic factors [5]	Yes	Yes	Yes
Observations	403	403	403
R^2	0.546	0.527	0.370
Note:	*p<0.1; **p<0.05; ***p<0.01		

Table A.9: Protest led by Ma Ba Tha, Riots, Hate Protest and Facebook User in Township level.

	Protests led by Ma Ba Tha	Riots	Hate_Protests
	(1)	(2)	(3)
Facebook_User	-0.065*	0.016	0.005
	(0.035)	(0.028)	(0.031)
Content_Moderation	-0.236***	0.010	-0.252**
	(0.071)	(0.074)	(0.102)
Facebook_User : Content_Moderation	0.169	-0.121	0.477***
	(0.121)	(0.107)	(0.133)
Township Fixed Effect	Yes	Yes	Yes
General Media Exposure [4]	Yes	Yes	Yes
Demographic controls [7]	Yes	Yes	Yes
Socioeconomic factors [5]	Yes	Yes	Yes
Observations	403	403	403
R^2	0.542	0.527	0.365
Note:	*p<0.1; **p<0.05; ***p<0.01		

Note: This table presents regression coefficient estimates for the interaction of Facebook user and YMBA Facebook page follower on Protest led by Ma Ba Tha, Riots, and Hate Protest, as specified in equation (1). The dependent variables, Protest led by Ma Ba Tha, Riots, and Hate Protest, are transformed using the natural logarithm (log (1+number of events)). All independent variables are dummy variables, taking the value 1 if the percentage of the variable is above the median, and 0 otherwise. The models include all control variables, township fixed effects. Robust standard errors are clustered by township in all specifications. Facebook User is dummy variable, it takes 1 if it is equal or more than 50% of township population are using Facebook otherwise it is 0.

Table A.10: Protest led by Ma Ba Tha, Riots, Hate Protest, and Internet User in Rakhine

	Protests led by Ma Ba Tha	Riots	Hate_Protests
	(1)	(2)	(3)
Internet_User	-0.035***	0.061***	-0.006
	(0.002)	(0.003)	(0.004)
Content_Moderation	-0.277***	0.023	-0.327***
	(0.078)	(0.096)	(0.125)
Internet_User:Content_Moderation	0.503***	-0.162*	0.365***
	(0.078)	(0.096)	(0.125)
Township Fixed Effect	Yes	Yes	Yes
General Media Exposure [4]	Yes	Yes	Yes
Demographic controls [7]	Yes	Yes	Yes
Socioeconomic factors [5]	Yes	Yes	Yes
Observations	190	190	190
R^2	0.277	0.230	0.271
Adjusted R^2	0.210	0.158	0.203
Residual Std. Error	0.300 (df=173)	0.277 (df=173)	0.303 (df=173)
F Statistic	20.032*** (df=16; 173)	1.506 (df=16; 173)	4.194*** (df=16; 173)
Note:	*p<0.1; **p<0.05; ***p<0.01		

Table A.11: Protest led by Ma Ba Tha, Riots, Hate Protest and Facebook User after curfew in Mandalay.

	Protests led by Ma Ba Tha	Riots	Hate_Protests
	(1)	(2)	(3)
Facebook_User	0.189***	-0.165***	0.085***
	(0.003)	(0.004)	(0.004)
Content_Moderation	-0.087	-0.054	0.051
	(0.059)	(0.088)	(0.093)
Facebook_User:Content_Moderation	-0.087	-0.054	0.051
	(0.059)	(0.088)	(0.093)
Township Fixed Effect	Yes	Yes	Yes
General Media Exposure [4]	Yes	Yes	Yes
Demographic controls [7]	Yes	Yes	Yes
Socioeconomic factors [5]	Yes	Yes	Yes
Observations	52	52	52
R^2	0.700	0.672	0.235
Adjusted R^2	0.636	0.602	0.071
Residual Std. Error	0.201 (df=42)	0.203 (df=42)	0.325 (df=42)
F Statistic	4.916*** (df=9; 42)	0.249 (df=9; 42)	0.095 (df=9; 42)

Note:	*p<0.1; **p<0.05; ***p<0.01

Table A.12: Protest led by Ma Ba Tha, Riots, Hate Protest, and townships that is above 50% of population using Facebook.

	Protests led by Ma Ba Tha	Riots	Hate_Protests
	(1)	(2)	(3)
Facebook_User_75	-0.065*	0.016	0.005
	(0.035)	(0.028)	(0.031)
Content_Moderation	-0.236***	0.010	-0.252**
	(0.071)	(0.074)	(0.102)
Facebook_User_75: Content_Moderation	0.169	-0.121	0.477***
	(0.121)	(0.107)	(0.133)
Township Fixed Effect	Yes	Yes	Yes
General Media Exposure [4]	Yes	Yes	Yes
Demographic controls [7]	Yes	Yes	Yes
Socioeconomic factors [5]	Yes	Yes	Yes
Observations	403	403	403
R^2	0.542	0.527	0.365
Adjusted R^2	0.410	0.391	0.182
Residual Std. Error	0.269 (df=312)	0.237 (df=312)	0.308 (df=312)
F Statistic	4.820*** (df=90; 312)	1.694*** (df=90; 312)	4.259*** (df=90; 312)
Note:	*p<0.1; **p<0.05; ***p<0.01		

Bibliography or Reference List

Acemoglu, D., A. Ozdaglar, and J. Siderius (2021). Misinformation: Strategic Sharing, Homophily, and Endogenous Echo Chambers. Technical report, National Bureau of Economic Research.

ACLED code book (https://acleddata.com/acleddatanew/wpcontent/uploads/2021/11/ACLED_Codebook_v1_January-2021.pdf)

Asher, S. (2021) Myanmar coup: How Facebook became the 'Digital Tea Shop', BBC News. BBC. Available at: https://www.bbc.com/news/world-asia-55929654 (Accessed: December 18, 2022).

Bursztyn, Leonardo, Aakaash Rao, and George B. Y. n.d. “Social Media and Xenophobia: Evidence from Russia.”

Brym, R., Godbout, M., Hoffbauer, A., Menard, G., & Zhang, T. H. (2014). Social media in the 2011 E Egyptian uprising. The British Journal of Sociology, 65(2), 266-292.

DellaVigna, Stefano, Ruben Enikolopov, Vera Mironova, Maria Petrova, and Ekaterina Zhuravskaya. 2014. “Cross-Border Media and Nationalism: Evidence from Serbian Radio in Croatia.” *American Economic Journal: Applied Economics* 6(3):103–32. doi: 10.1257/app.6.3.103.

Frankel, R. (2022) An update on the situation in Myanmar, Meta. Available at: https://about.fb.com/news/2021/02/an-update-on-myanmar/ (Accessed: December 17, 2022).

Gillespie, Tarleton. 2019. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.

Jiménez Durán, Rafael, Karsten Müller, and Carlo Schwarz. 2022. “The Effect of Content Moderation on Online and Offline Hate: Evidence from Germany’s NetzDG.” *SSRN Electronic Journal*. doi: 10.2139/ssrn.4230296.

Müller, Karsten, and Carlo Schwarz. 2021. “Fanning the Flames of Hate: Social Media and Hate Crime.” *Journal of the European Economic Association* 19(4):2131–67. doi: 10.1093/jeea/jvaa045.

Williams, Matthew L., Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. 2019. “Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline

Racially and Religiously Aggravated Crime." *The British Journal of Criminology* azz049. doi: 10.1093/bjc/azz049.

Xie, Yungeng, Rui Qiao, Guosong Shao, and Hong Chen. 2017. "Research on Chinese Social Media Users' Communication Behaviors during Public Emergency Events." *Telematics and Informatics* 34(3):740–54. doi: 10.1016/j.tele.2016.05.023.

Twitter (2015). Fighting Abuse to Protect Freedom of Expression.

https://blog.twitter.com/en_us/a/2015/ fighting-abuse-to-protect-freedom-of-expression.

Mozur, P. (2018). A Genocide Incited on Facebook, With Posts from Myanmar's Military. New York Time

Wai, H. L. (2019). Effect of facebook application on purchase intention towards online shopping (Doctoral dissertation, MERAL Portal).

Wittekind, C. T., & Faxon, H. O. (2022). Networks of Speculation: Making Land Markets on Myanmar Facebook. Antipode.

Dias Oliva, T. (2020). Content moderation technologies: Applying human rights standards to protect freedom of expression. Human Rights Law Review, 20(4), 607-640. doi:10.1093/hrlr/ngaa032

Global Witness. (2022). Retrieved from https://www.globalwitness.org/en/blog/exposing-social media-platforms-failures-to-protect-their-users/

Jiḿenez Duŕan, R. (2022). The Economics of Content Moderation: Theory and Experimental Evidence from Hate Speech on Twitter. Available at SSRN.

Klonick, K. (2018). Retrieved from https://www.nytimes.com/2018/04/26/opinion/facebook-content-moderation-rules.html

Larson, J. M., Nagler, J., Ronen, J., & Tucker, J. A. (2019). Social networks and protest participation: Evidence from 130 million Twitter users. American Journal of Political Science, 63(3), 690-705.

Lee, S. (2018). The role of social media in protest participation: The case of candlelight vigils in South Korea. International Journal of Communication, 12, 18.

Liu, Y., P. Yildirim, and Z. J. Zhang (2021). Social Media, Content Moderation, and Technology. arXiv preprint arXiv:2101.04618.

Lwin, M., & Panchapakesan, C. (2015). Retrieved from https://www.researchgate.net/publication/286446503_The_Use_of_Mobile_Phones_Among_Trishaw_Operators_in_Myanmar

Madio, L. and M. Quinn (2021). Content Moderation and Advertising in Social Media Platforms. Available at SSRN 3551103.

Miles, T. (2018). Retrieved from https://www.reuters.com/article/us-myanmar-rohingya-facebook-idUSKCN1GO2PN

Myanmar Crime Rate (https://www.mmsis.gov.mm/)

Su, S. (2019) Update on Myanmar, Meta. Available at: https://about.fb.com/news/2018/08/update-on-myanmar/ (Accessed: December 17, 2022).

Whitten-Woodring, J., Kleinberg, M. S., Thawnghmung, A., & Thitsar, M. T. (2020). Poison if you don’t know how to use it: Facebook, democracy, and human rights in Myanmar. The International Journal of Press/Politics, 25(3), 407-425.

Wittekind, C. T., & Faxon, H. O. (2022). Networks of Speculation: Making Land Markets on Myanmar Facebook. Antipode.