Monitoring Hate Speech in Indonesia: An NLP-based Classification of Social Media Texts

Musa Izzanardi Wijanarko[*,1], Lucky Susanto[*,1], Prasetia Anugrah Pratama[2]
Ika Idris[1], Traci Hong[3], Derry Wijaya[3,1]

[*]Equal Contribution
[1]Monash University, [2]Independent Researcher, [3]Boston University

Abstract

Online hate speech propagation is a complex issue, deeply influenced by both the perpetrator and the target's cultural, historical, and societal contexts. Consequently, developing a universally robust hate speech classifier for diverse social media texts remains a challenging and unsolved task. The lack of mechanisms to track the spread and severity of hate speech further complicates the formulation of effective solutions. In response to this, to monitor hate speech in Indonesia during the recent 2024 presidential election, we have employed advanced Natural Language Processing (NLP) technologies to create an improved hate speech classifier tailored for a narrower subset of texts; specifically, texts that target vulnerable groups that have historically been the targets of hate speech in Indonesia. Our focus is on texts that mention these six vulnerable minority groups in Indonesia: Shia, Ahmadiyyah, Christians, LGBTQ+, Indonesian Chinese, and people with disabilities, as well as one additional group of interest: Jews. The insights gained from our dashboard have assisted stakeholders in devising more effective strategies to counteract hate speech. Notably, our dashboard has persuaded the General Election Supervisory Body in Indonesia (BAWASLU) to collaborate with our institution and the Alliance of Independent Journalists (AJI) to monitor social media hate speech in vulnerable areas in the country known for hate speech dissemination or hate-related violence in the upcoming Indonesian regional elections. This dashboard is available online at `https://aji.or.id/hate-speech-monitoring`.

1 Introduction

Indonesia's history is marked by the use of hate speech to incite discrimination and violence (George, 2016). This speech, often amplified during times of political tension such as during an election, targets people or groups based on their race, gender, ethnicity, religion, sexual orientation, and disability. The advent of social media has exacerbated this issue, as evidenced by a ten-fold increase in hate speech ratio during the 2024 Indonesian presidential election compared to 2021-2022 (CSIS, 2022).

Jews	LGBTQ+	Indo-Chinese
is ra hell	lesbong	cokin
setanyahu	eljibiti	cindo
joo	lghdtv+	chindo

Table 1: Words and phrases commonly appearing in Indonesian hate speech texts targeting each group.

Countering and mitigating hate speech is challenging due to its volume and the variation in content based on the cultural, historical, and societal contexts of both the perpetrator and the target (e.g., different words may be used to target different groups in different countries at different times (Table 1)). Hence, creating effective strategies to counter hate speech is hard. Detection may be the logical first step in combating hate speech. A hate speech monitoring tool for effective intervention and mitigation is therefore needed.

Neural networks (Devlin et al., 2019; Liu et al., 2019) and large language models (Touvron et al., 2023; OpenAI et al., 2024; Nguyen et al., 2024) are potential solutions for detecting hate speech. Indeed, they have been used in works such as Mathew et al. (2022) and Guo et al. (2024); but their performance is not yet satisfactory, with the highest performance benchmarked on English hate speech being a macro-F1 score of 0.73 by ChatGPT (Brown et al., 2020). Correspondingly, on the Indonesian hate speech we build, ChatGPT reaches a macro-F1 score of 0.63 (section 3.2).

In this work, we demonstrate that leveraging keywords for data collection and insights from minority groups can enhance hate speech detection, even with a smaller model. Specifically, we use keywords (Appendix A) obtained through focus group discussions (FGDs) involving Indonesian

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing
System Demonstrations, pages 142–152
November 12-16, 2024

minority groups to collect posts mentioning these groups. Then, representatives from the groups annotate samples of these posts for the presence of hate speech. The resulting annotated data is used to build our hate speech dataset, named IndoToxic2024[1] (Susanto et al., 2024). The IndoBERTweet (Koto et al., 2021) fine-tuned on this dataset achieves a 0.78 macro-F1 cross-validation score.

We introduce our hate speech dashboard[2], which is the result of the collaboration between Monash University Indonesia and the civil society organization the Indonesian Alliance of Independent Journalists (AJI). This dashboard is licensed under CC BY-SA 4.0[3]. We also publicly release the model used to construct the dashboard on Huggingface[4].

Using the fine-tuned IndoBERTweet model, our dashboard automatically detects hate speech in sources like X, Facebook, Instagram, and online articles, providing insights to stakeholders. Media stakeholders can use it to track hate speech trends against vulnerable groups, aiding in public reporting and impact mitigation. Social media platforms can gain insights into how their moderation policies impact hate speech toward vulnerable groups. Election organizers can use this tool to alert them on the severity of hate speech during elections, which can serve as a foundation for future strategies to mitigate hate speech, balance freedom of expression, guide staff, and establish ethical guidelines for election participants.

2 Related Work

2.1 Hate Speech Detection

Evolution in hate speech detection systems is attributed to the changes in what society perceives as hate speech (Delgado, 1982; Greenawalt, 1989; Nations, 2023; Paramadina and Mafindo, 2023). Initially, these systems were trained on data with unanimous agreement among annotators (Alfina et al., 2017a; Ibrohim and Budi, 2018). Recent research, however, has shifted focus to the role of subjectivity in hate speech classification (Fleisig et al., 2024; Susanto et al., 2024). Unfortunately, incorporating subjectivity into hate speech detection systems is still nascent, leading us to utilize a traditional hate speech detection system, taking only the text as its sole input.

[1] IndoToxic2024 Dataset
[2] AJI Website, containing our hate speech dashboard
[3] Attribution-ShareAlike 4.0 International
[4] Our Indonesian Hate Speech text classifier

Online hate speech, a growing problem linked to an increase in offline hate crime, has been the focus of numerous monitoring efforts (Williams et al., 2019). For instance, CSIS (2022) developed a dashboard to track hate speech on Twitter (now X) targeting Indonesian minority groups consisting of Ahmadiyyah, Shi'a, Tionghoa (Chinese Indonesians), Christians, and Ethnic Papuans; which was developed due to the groups receiving some of the worst campaigns of hate speech that cause significant harm to the groups and the violation of their rights (CSIS, 2022). Similarly, CIJ (2023) created a dashboard for monitoring hate speech during Malaysia's 15th general election, working with a broader definition of target groups consisting of "Gender and LGBTIQ", "Race", "Refugees and Migrants", "Religion", and "Royalty". CIJ (2023)'s dashboard emphasizes the severity of hate speech, where it circulates, and who created it. However, neither the models nor the datasets used to construct these dashboards were publicly released, limiting evaluations and future works for these monitoring efforts.

2.2 NNs as Hate Speech Classifier

Neural Networks (NNs) have gained much traction since the introduction of the transformer architecture (Vaswani et al., 2017), which was further popularized by the BERT model (Devlin et al., 2019) and other subsequent language models. These language models have been employed early on for text classification including sentiment analysis and hate speech detection in various languages, not only on English texts (Saleh et al., 2021), but also on other language texts such as Bengali (Keya et al., 2023), Vietnamese (Hoang et al., 2023), and Indonesian (Susanto et al., 2024).

2.3 LLMs as Hate Speech Classifier

Recent years have seen large language models (LLMs) excel in various tasks (Touvron et al., 2023; OpenAI et al., 2024) including hate speech classification (Guo et al., 2024). However, their performance tends to drop for non-English languages as they are predominantly trained on English language texts (Li et al., 2024). Most of the state-of-the-art LLMs perform poorly on Indonesian language tasks, with gpt-3.5 being an exception as of 2023 (Koto et al., 2023). Many recent works have therefore focused on the creation of language-specific LLMs for non-English languages, like SeaLLM for Southeast Asian languages (Nguyen et al., 2024).

3 Methodology

In this work, we adopt the definition of hate speech set by Indonesia's National Human Rights Commission, which includes any communication motivated by hatred against people based on their identities, intending to incite violence, death, and social unrest (Paramadina and Mafindo, 2023). Based on this definition and the domestic context of online hate speech and toxicity in Indonesia, we define five types of hate speech and toxic text in our work:

- **Profanity or obscenity**: Texts that utilize harsh and inappropriate language that offend the majority of the reader.
- **Insult**: Texts that utilize harsh and inappropriate language that intend to humiliate the target.
- **Incitement to violence**: Texts that intend to cause loss, danger, or difficulties to a person or a group, including physical violence, intimidation, or any other actions that cause fear and distress to the target.
- **Identity attack**: Texts that attack and demean others' identities which include ethnicity, religion, race, sexual orientation, and gender.
- **Sexual explicit**: Texts with the mention of sexual activities or sex organs that intend to harass the target.

Unlike prior hate speech detection efforts that focus primarily on detection models, we integrate insights from Indonesian vulnerable group about common online attacks targeted towards them. This was achieved through focus group discussions (FGDs), where we identified seven targeted vulnerable groups, comprising six minority groups: Shia, Ahmadiyyah, Christians, LGBTQ+ individuals, Tionghoa, and people with disabilities, along with one additional group of interest: Jews, due to the rising Israeli-Palestinian conflict.

Through the FGDs, we obtain keywords that are often used online to refer to each minority group as well as keywords used to target each vulnerable group (listed in Appendix A). Using these keywords, we use Brandwatch (www.brandwatch.com) to collect data mentioning the targeted vulnerable groups from X (formerly Twitter), and the now-deprecated Crowdtangle (https://crowdtangle.com/) to retrieve data from Facebook and Instagram. Due to X's download limit, we use a sampling rate of 23%, implying that for each post we gathered from the platform, approximately three posts were not collected. In collaboration with an Indonesian fact-checking organization Mafindo, we collect news articles containing misinformation that mention these groups from Cekfakta's article database (https://cekfakta.com/). The data totals 1.45 million texts (from 1 Sep 23 to 27 Mar 24).

3.1 IndoToxic2024 Hate Speech Dataset

Our IndoToxic2024 dataset was created by randomly sampling previously collected data, which was then annotated by 19 annotators from various backgrounds and ethnicities, including members of the six targeted minority groups. The dataset is multi-label, including a toxicity type label for each entry in the data. This dataset was then used to train and evaluate our hate speech detection model.

To train the model, we down-sample the imbalanced IndoToxic2024 dataset, which contains more non-hate speech texts than hate speech texts, to the ratio of one positive to three negative examples. We use the 6,807 positive and 20,421 negative samples; totaling 27,228 samples. Since the IndoToxic2024 dataset contains text multiple annotators annotate, there are samples with conflicting annotations for a singular text. This dataset therefore imitates the real-life complexity of hate speech messages in social media.

3.2 Model Comparison

We evaluate IndoBERTweet (Koto et al., 2021), SeaLLM (Nguyen et al., 2024), and gpt-3.5-turbo (Brown et al., 2020). IndoBERTweet, fine-tuned on the IndoToxic2024 dataset (Susanto et al., 2024), is assessed using **stratified 10-fold cross-validation**, ensuring no leakage during evaluation. Due to resource constraint, SeaLLM and gpt-3.5-turbo are evaluated in a zero-shot setup. gpt-3.5-turbo is also evaluated in a few-shot setup. IndoBERTweet is pre-trained on Indonesian texts, SeaLLM is primarily pre-trained on Southeast Asian languages, and gpt-3.5-turbo is mainly trained on English texts.

Model	Macro-F1
IndoBERTweet	**0.718**
gpt-3.5-turbo (zero-shot)	0.627
SeaLLM-7B-v2.5	0.517
gpt-3.5-turbo (few-shot)	0.429

Table 2: Performance of multiple models on the IndoToxic2024 Dataset.

The gpt-3.5-turbo's few-shot prompting setup involves providing the model with 15 static exam-

ples (provided in Appendix B), comprising eight positive and seven negative instances, maintaining a balanced ratio. The eight positive instances represent hate speech toward each of our seven targeted vulnerable groups, with the addition of Rohingya refugees in the IndoToxic2024 dataset. However, the performance significantly declined **from a macro-F1 score of 0.627 in the zero-shot setup to 0.429 in the few-shot setup** (Table 2). This drop may be attributed to the increased complexity of the prompt and its application to a non-English task (Li et al., 2024).

3.3 Model Selection

Classification Task	Accuracy	Macro-F1
Related to Election	0.96	0.93
Hate Speech	0.89	0.78
Identity Attack	0.75	0.80
Incitement to Violence	0.77	0.53
Insult	0.79	0.85
Profanity or Obscenity	0.81	0.70
Sexual Explicit	0.91	0.80

Table 3: Performance of the fine-tuned IndoBERTweet models for each text classification task in our dashboard.

We utilize IndoBERTweet models fine-tuned on the IndoToxic2024 dataset Susanto et al. (2024) in this work as our final classifier for the dashboard. The performance of the fine-tuned IndoBERTweet models for different classification tasks visualized in our dashboard is shown in Table 3.

IndoBERTweet itself is pre-trained by extending a monolingually-trained Indonesian BERT model, named IndoBERT (Koto et al., 2020), with additive domain-specific vocabulary specific to Indonesian Twitter texts. The model efficiently handles vocabulary mismatch, an important quality when handling social media texts as the vocabulary may drastically change with time. IndoBERTweet has been trained for various tasks in previous works, including hate speech detection, using data from Alfina et al. (2017b) and Ibrohim and Budi (2019).

3.4 Our Dashboard Pipeline

After scraping posts and articles containing mentions of the vulnerable groups using the keywords, we utilize the fine-tuned IndoBERTweet model for the various classification tasks. We then visualize the results on a dashboard created using Power BI.

4 System Description: Content of the Dashboard

At the time of this paper's submission, our dashboard has processed over 1.45 million online texts mentioning the identified vulnerable groups, dating from 1 September 2023 to 27 March 2024, from Facebook, X, Instagram, and online articles. The dashboard, created using Power BI, consists of the following 6 pages.

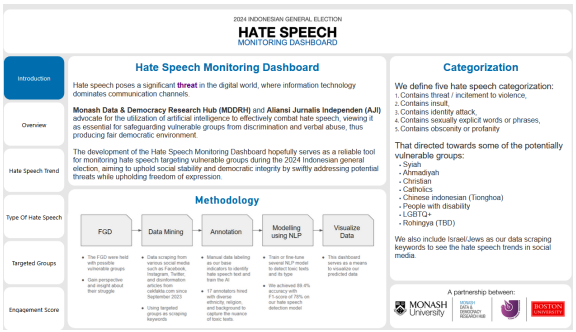

Figure 1: The Introduction Page

The Introduction Page outlines the motivation behind this dashboard, what we define as hate speech, the time frame of interest, where the data originate from, the target groups we focus on, and how we create this dashboard.

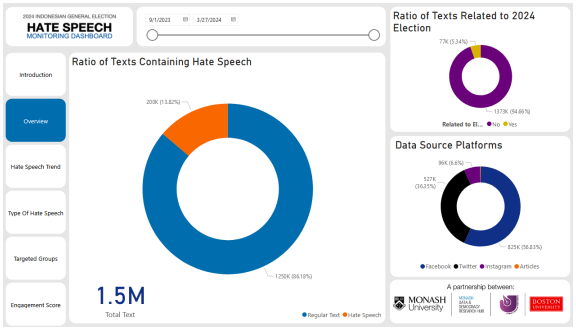

Figure 2: The Overview Page

The Overview Page serves as the main summary of information. At the top of the page exists a slider to filter the data date range. Additionally, there are three pie charts, each displaying the hate speech distribution, the distribution of texts related to the election (i.e., "Related to Pemilu 2024"), and the data source distribution.

The Hate Speech Trend Page shows the quantity of hate speech over time on multiple social media platforms. We also add filter options to enhance analysis capability: the date filter, platform filter, and related-to-election filter. These filters are also available in the following two pages.

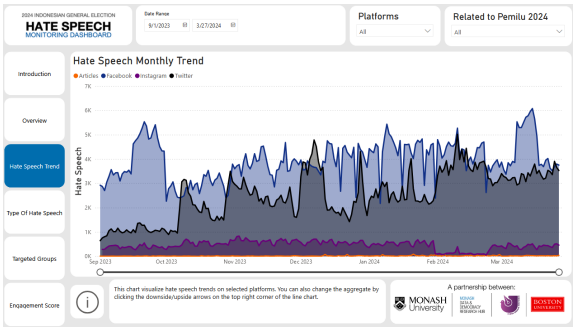

Figure 3: The Hate Speech Trend Page

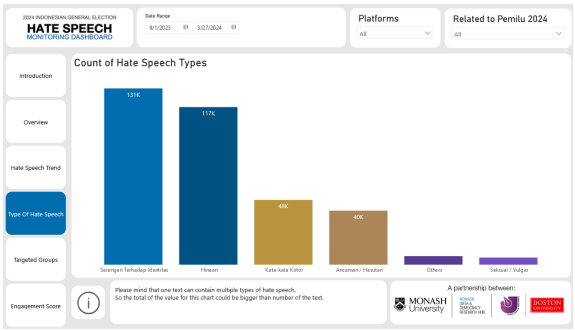

Figure 4: The Type of Hate Speech Page

The Type of Hate Speech Page functions to map the type of hate speech–identity attack, insult, profanity, threat/incitement to violence, or vulgarity–that our model predicts in the dataset. Since a text can potentially contain more than one type of hate speech, the total sum of data on this page will be above the hate speech count presented on the overview page.

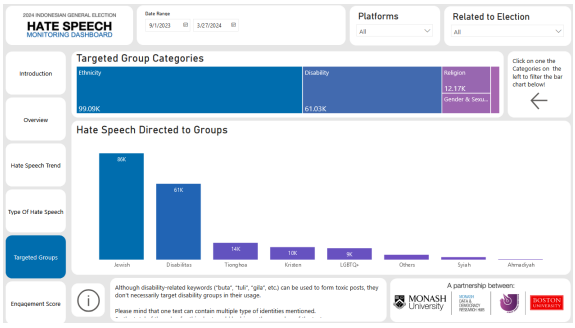

Figure 5: The Targeted Groups Page

The Targeted Groups Page shows the distribution of the targeted vulnerable groups in the detected hate speech. We also group these target groups into coarser categories such as ethnicity, religion, disability, and gender & sexuality.

The Engagement Score Page shows how much engagement hate speech texts collectively obtain from each platform. This page contains filters from previous pages, namely the (hate speech) target

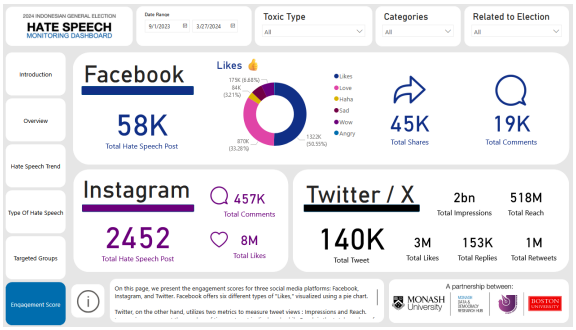

Figure 6: The Engagement Score Page

group category filter, the related-to-election filter, and the hate speech type filter.

5 Observation Results

From this monitoring tool, a non-exhaustive list of interesting observations can be made:

The 2023 Israel–Hamas war has affected the circulation of hate speech targeting Jews in Indonesia, shown in Figure 7. Before the war, which started on 7th October 2023, only 15K out of 189.9K (7.78%) texts were found to be hate speech. During this period, only 1.5K hate speech texts targeted Jews, while Chinese descendants in Indonesia (the Tionghoa ethnicity) had 4.1K hate speech texts targeting them. However, in November 2023, 42K out of 206.9K (20.21%) texts were found to be hate speech. During this period, hate speech texts against Tionghoa ethnicity dropped to only 1.25K texts, while hate speech texts targeting Jews sharply rose to 28K. This number means that two-thirds of hate speech texts in November 2023 targeted Jews.

Though the ratio of hate speech circulating in March 2024 on social media has returned to its previous level in September 2023, the number of overall hate speech has increased. Despite our constant sampling rate during data collection, the number of posts mentioning targeted vulnerable groups in Indonesia has increased in recent months, as shown in Figure 8. So, even though technically the ratio of hate speech to non-hate speech text mentioning vulnerable groups in Indonesia has fallen from 7.53% in September 2023 to 7.39% in March 2024, the total number of hate speech has increased from 12,465 to 16,395. Note that we did update our keywords to collect texts mentioning the Rohingya refugees in December 2023.

Some vulnerable groups are attacked for political reasons. Filtering our dashboard to texts related to the 2024 Indonesian presidential election,

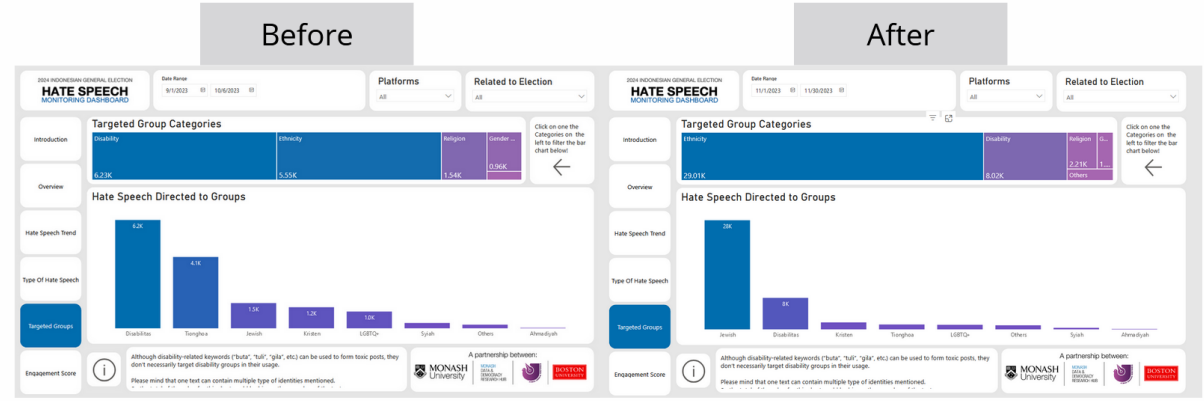

Figure 7: Hate Speech trend before and after the Israel-Hamas war on 7th October 2023, where a drastic increase of hate speech against Jews in Indonesia can be seen.

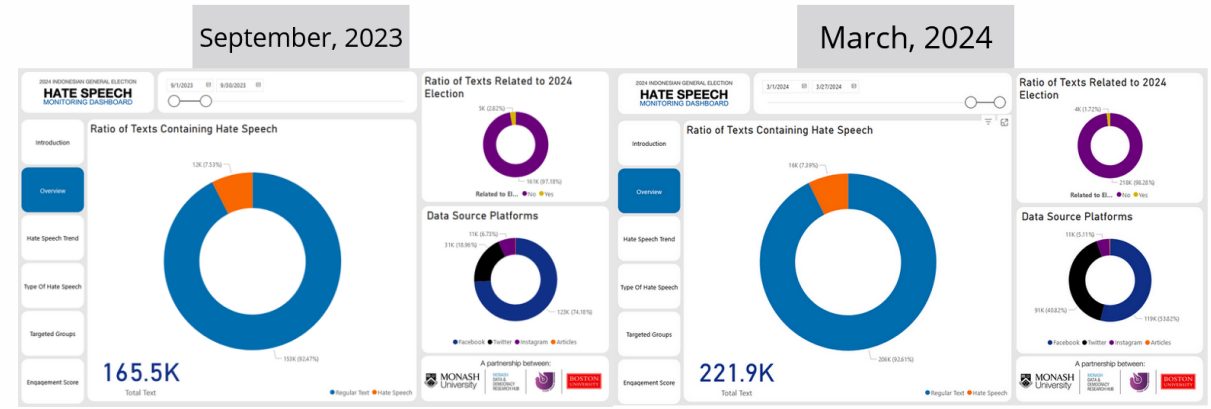

Figure 8: Hate Speech ratio on September 2023 and March 2024. The count of hate speech texts increases, though the percentage remains similar. In September 2023, Tionghoa ethnicity was the main target, but in March 2024, Jewish ethnicity became the main target of hate speech.

we see that the Tionghoa ethnicity is often the target of political (i.e., related-to-election) hate speech, as shown in Figure 9. After the Israel-Hamas war, the prominent target of political hate speech shifted to Jews. However, we noticed that during both the 4th and 5th presidential debates, aired on 21st January and 4th February 2024 respectively, the target of political hate speech returned to the Tionghoa ethnicity for a short while.

Meanwhile, other vulnerable groups are attacked for non-political reasons. The top three vulnerable groups often being targeted by political hate speech are Jewish, Tionghoa, and LGBTQ+ while the top three vulnerable groups often being targeted by hate speech in general are Jewish, Tionghoa, and Christians. Throughout the dashboard's monitoring, we only find 301 texts where Christians are the target of political hate speech; meanwhile, they are targeted by over 9765 non-political hate speech texts.

6 Conclusion and Recommendation

Correctly fighting hate speech is hard. Effective measures like stringent content filtering or social media bans should be reserved for extreme cases. But, knowing when we have reached those extreme cases is not trivial. This is why we reiterate the importance of a hate speech monitoring tool.

The General Election Supervisory Body in Indonesia (BAWASLU) has also monitored hate speech during Indonesia's 2024 presidential election. However, theirs was done manually with human annotators monitoring and collecting posts on multiple social media platforms. As expected, this approach to monitor hate speech lacks scalability. Comparatively, our dashboard allows for scalable monitoring, only requiring someone to download scraped social media posts and prepare them for the model to infer, which can be done by a single person. This was the basis of Monash University Indonesia's collaboration with BAWASLU, under-

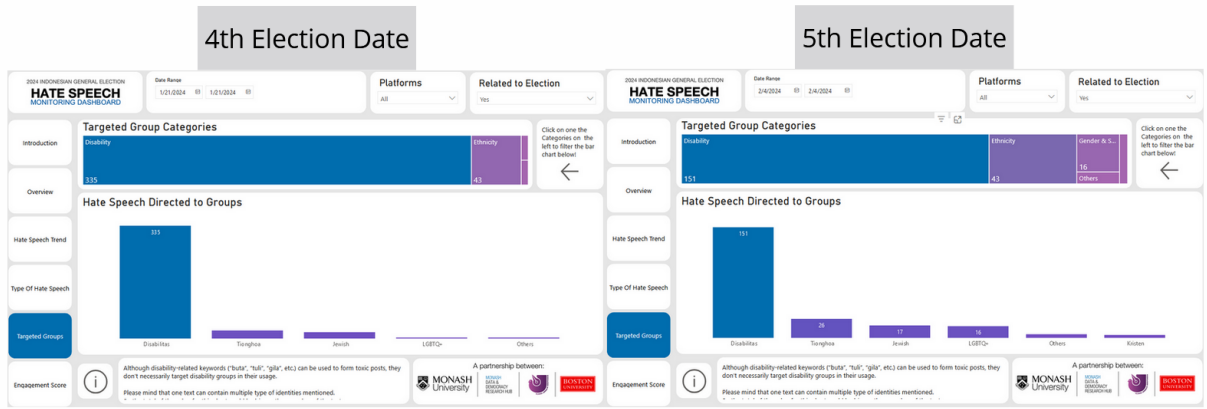

Figure 9: Targets of political hate speech on the 4th and 5th presidential debate, where Tionghoa ethnicity was the main target, overtaking Jewish ethnicity hate speech count slightly.

lining the importance of scalability and the application of NLP technologies for monitoring hate speech, which we explain further in the **Impact** section of our work below.

Based on our dashboard's findings from the 2024 election, we urge stakeholders - social media platforms, election organizers, media, and journalists - to intensify their efforts to prevent and mitigate online hate speech, particularly during political events like general elections.

Our recommendations for social media platforms are as follows:

1. **Map and identify** potential targets for online hate speech as a first step, since targets of hate speech may change over time, exemplified by the surge in anti-Semitic hate speech in the ongoing Israeli-Palestinian conflict.
2. **The inclusion of experts and vulnerable communities** in the development and throughout the hate speech monitoring can assist in the early detection of unpredictable events like the Rohingya refugee hate speech.
3. **Examine the social media algorithm's** impact on hate speech content promotion, particularly its inadvertent promotion of hate speech, to avoid echo chambers and filter bubbles.
4. **Utilize fact-checked databases** such as Cekfakta, annotated by neutral parties, to combat hate speech and discrimination.
5. **Collaborate with other platforms** to manage the cross-platform spread of hate speech.
6. **Promote credible news sources** like independent media and fact-checking organizations to inform the public accurately.
7. **Update community standards** to counter cyber-troops infiltrating the platform with fake accounts and troll content.
8. **Provide API access** to experts, researchers, and journalists for monitoring and analyzing hate speech trends on the platform.

Election organizers must remember that hate speech is context-dependent; influenced by historical, societal, and cultural contexts. Any action to prevent and mitigate hate speech must consider its impact on citizens' freedom of expression. Controversial regulations like Article 28 paragraph (2) of the 2016 Indonesian ITE Law (Law on Electronic Information and Transactions), often misused to silence marginalized minority groups, necessitate the exploration of non-regulatory solutions. To this end, we recommend the following:

1. **Strategic partnerships** with civil society, experts, and organizations are essential to address hate speech during political events.
2. **Monitoring and reporting** hate speech against each minority group is crucial, especially during political times, to prevent civil unrest and targeted violence.
3. **Training sessions** are necessary to equip local election organizers with the skills to monitor hate speech effectively.

Lastly, for the media and journalists, we recommend the following:

1. **Promote awareness**, maintain a vigilant watch, and report on the trends of hate speech on social media platforms, especially during periods of political unrest.
2. **Reinforce fact-checking culture** by verifying statements containing hate speech made by politicians, candidates, and their party.

Limitations of Our Work

Limited to Indonesian texts Our dashboard can only accurately infer Indonesian texts. It is well known that social media posts can sometimes contain code-switch texts such as a regional dialect. However, we did not conduct an extensive review of this phenomenon. We mitigate this by using IndoBERTweet, a model trained on informal Indonesian social media texts.

Not evaluated on general texts Though the model we used for hate speech detection boasts a 89% accuracy with a 78% macro-F1 score, this is only tested on texts already filtered by the keywords we use i.e., on texts mentioning targeted vulnerable groups. We did not evaluate its performance for general social media texts.

Not up-to-date with LLMs evaluation Our dashboard, launched online on 12th February 2024, may not reflect the rapid advancements in large language models, such as the cheaper and more efficient GPT-4o mini released on 18th July 2024. The performance gap between our model and the latest large language models may be smaller than reported.

The Impact of Our Dashboard

Acts as a catalyst in starting the collaboration between the General Election Supervisory Body in Indonesia (BAWASLU) and Monash University Indonesia After advocating our results to BAWASLU, Monash University Indonesia is now collaborating with the government agency, starting with a memorandum of understanding. This collaboration is proof that BAWASLU now wants to take a more proactive stance, collaborating to monitor social media hate speech in vulnerable locations known for abundant hate incidents, both online and offline.

Raising the issue of hate speech to Meta We have also advocated our results to Meta, which resulted in talks between Monash University Indonesia and the team at Meta. Particularly, they are interested on how we can collaborate to mitigate hate speech in the upcoming regional elections in Indonesia, where hate speech is predicted to spike again.

Increasing awareness and educating the masses on hate speech Our hate speech dashboard has garnered significant attention, with coverage from 32 national media outlets, including high-traffic media outlets like Kompas.com. This widespread media coverage has played a role in enhancing public awareness about the prevalence of hate speech in Indonesia. For quantifiable proof, we also checked the visit count and page view count where our dashboard went live. On 11th February 2024, a day before the dashboard's official release on AJI's homepage, we recorded 332 visits and 2,226 page views. The subsequent day, these numbers surged, with visits doubling to 667 and page views escalating to 5,045. The interest peaked on February 13, 2024 (the day before the presidential election), with 701 visits and a remarkable 15,545 page views. The high page view count also indicates a significant interest from visitors who are keen to understand more about the situation of hate speech in Indonesia.

Ethical Consideration

Weighing the Pros and Cons of monitoring hate speech Hate speech has continued to thrive in online social media platforms. However, tools to combat them effectively are still capable of improvements. Hate speech is a complex issue because it involves human emotions and biases, thus it cannot be solved correctly by relying only upon a machine solution. Of course, one extreme solution always exists, to remove any text that mentions any vulnerable groups; but this type of action can only end up hurting everyone and further marginalizing the already vulnerable groups. The phenomenon of hate speech not only poses a threat but also an opportunity to learn why it exists and how it can be mitigated or treated. The benefits of having a monitoring tool for this issue far outweigh the drawbacks of not having one, as it can be used to inform citizens, track the trend of hate speech, quantify the severity, and provide insights on how to mitigate it correctly.

Protection of the authors of the used data Our dashboard only reports on the statistics of the data, without any leak on who the original author of the data is. This act ensures that no authors can be traced and is protected.

Consideration of misuse Potential misuse of our dashboard, such as by malicious groups gauging their success, is deemed non-concerning as such groups do not require a monitoring tool for this purpose.

References

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017a. Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238.

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017b. Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *Neural Information Processing Systems*, 33:1877–1901.

CIJ. 2023. Election Monitoring.

CSIS. 2022. Hate speech dashboard.

R. Delgado. 1982. Words that wound: A tort action for racial insults, epithets, and name-calling. *Harvard Civil Rights-Civil Liberties Law Review*, 17:133–181.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2024. When the majority is wrong: Modeling annotator disagreement for subjective tasks.

Cherian George. 2016. *Hate Spin: The Manufacture of Religious Offense and Its Threat to Democracy*. The MIT Press, Cambridge.

K. Greenawalt. 1989. *Conflicts of Law and Morality*. Oxford University Press, New York.

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2024. An investigation of large language models for real-world hate speech detection.

Phu Gia Hoang, Canh Duc Luu, Khanh Quoc Tran, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2023. ViHOS: Hate speech spans detection for Vietnamese. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 652–669, Dubrovnik, Croatia. Association for Computational Linguistics.

Muhammad Okky Ibrohim and Indra Budi. 2018. A dataset and preliminaries study for abusive language detection in indonesian social media. *Procedia Computer Science*, 135:222–229. The 3rd International Conference on Computer Science and Computational Intelligence (ICCSCI 2018) : Empowering Smart Technology in Digital Era for a Better Life.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57, Florence, Italy. Association for Computational Linguistics.

Ashfia Jannat Keya, Md. Mohsin Kabir, Nusrat Jahan Shammey, M. F. Mridha, Md. Rashedul Islam, and Yutaka Watanobe. 2023. G-bert: An efficient method for identifying hate speech in bengali texts on social media. *IEEE Access*, 11:79697–79709.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10660–10668, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 757–770, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2024. Quantifying multilingual performance of large language models across languages.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2022. Hatexplain: A benchmark dataset for explainable hate speech detection.

United Nations. 2023. Hate speech and real harm | United Nations.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. Seallms – large language models for southeast asia.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and Irwan Bello et al. 2024. Gpt-4 technical report.

T. P. Paramadina and Mafindo. 2023. *Buku Panduan Melawan Hasutan Kebencian dan Hoax Edisi Perluasan*. PUSAD Paramadina, Jakarta.

Hind Saleh, Areej Alhothali, and Kawthar Moria. 2021. Detection of hate speech using bert and hate speech word embedding with deep model.

Lucky Susanto, Musa Izzanardi Wijanarko, Prasetia Anugrah Pratama, Traci Hong, Ika Idris, Alham Fikri Aji, and Derry Wijaya. 2024. Indotoxic2024: A demographically-enriched dataset of hate speech and toxicity types for indonesian language.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Matthew L Williams, 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*, 60(1):93–117.

A Keywords Used for Scraping

cina, china, tionghoa, chinese, cokin, cindo, chindo, shia, syiah, syia, ahmadiyya, ahmadiyah, ahmadiya, ahmadiyyah, transgender, queer, bisexual, bisex, gay, lesbian, lesbong, gangguan jiwa, gangguan mental, lgbt, eljibiti, lgbtq+, lghdtv+, katolik, khatolik, kristen, kris10, kr1st3n, buta, tuli, bisu, budek, conge, idiot, autis, orang gila, orgil, gila, gendut, cacat, odgj, zionis, israel, jewish, jew, yahudi, joo, anti-christ, anti kristus, anti christ, netanyahu, setanyahu, bangsa pengecut, is ra hell, rohingya, pengungsi, imigran, sakit jiwa, tuna netra, tuna rungu, sinting.

B Static 15 Few-shot Prompts

```
Input: Perbincangan mengenai Kristen Muhammadiyah sempat menjadi trending topic di media sosial. Hal ini terjadi selepas B
edah Buku berjudul Kristen Muhammadiyah yang menghadirkan Mendikbudristek Nadiem Makarim. #monitorday #kristenmuhammadiyah
Output: 0
Input: PEMILU / PILPRES YANG CACAT HUKUM❗ MERUBAH BATAS USIA "MELALUI MK" ADALAH KECURANGAN DAN AROGANSI TERANG BENDERAN
G DI DEPAN MATA YANG DI LAKUKAN PENGUASA
Output: 0
Input: Islam juga TPI knpa masih ada lagi nabi sesudah nabi Muhammad yaitu Ahmad /Ahmadiyah Trus gmna bacaan syahadatnya?
Output: 1
Input: YNHOR IFD WADUH !! Imigran Rohingya Dipindahkan Jauh dari Pemukiman Warga Karena Sering BAB Sembarangan INFOVIRAL K
omentar kalian guys ??
Output: 1
Input: Ada yang malemnya semangat banget mau summit pagi buta ngejar sunrise di puncak, tapi ternyata tidurnya sampe siang
. Akhirnya pas bangun lebih pilih ngopi dan bikin mie di tenda.. ada. MountNesia Selamat malam. #pendakingantukan
Output: 0
Input: Jalur kereta buatan China berdayakan warga lokal jadi lebih mandiri
Output: 0
Input: Knp siy banyak yg salty ke akun syiah yg terang2an belain yahudi? Padahal kan wajar sesama murid dajal harus saling
support. Kalo syiah bela muslim baru aneh.
Output: 1
Input: Ajari anak-anakmu bahwa Palestina sedang diduduki dan Masjid Aqsha sedang di nistakan. Entitas Zionis adalah musuh,
perlawanan adalah suatu kehormatan dan tidak ada negara yang disebut Isrel
Output: 1
Input: Hasil Akhir Qualification AFC Cup U23 , Indonesia 9:0 Chinese Taipei, skor akhir , TimNas diatas angin, tapi lawan
Timnas sesungguhnya adalah lawan Turkmenistan nanti #TimnasDay
Output: 0
Input: Ga abis pikir gw sama pemerintah knp ga bs lindungin warga negaranya? Udah tau mandarin tuh bukan bhs wajib dipljri
n di Indo tp mentang2 ambil investasi dr China skrg lowker bnyk yg hrsin itu?? Ga fair dong. Udah gila lu ye..
Output: 1
Input: KRISTEN sedunia ... penipu semuaaaa akal liciknya terbaca sama saya
Output: 1
Input: NEWS UPDATE kumparan Serangan Israel hantam Gereja Santo Porfiri yang lagi lindungi warga Islam di Gaza, 8 orang te
was. Serangan Israel di Gaza pada Kamis (19/10) menghantam sebuah gereja ortodoks. Tempat ibadah itu dipakai tempat berlin
dung warga Muslim dan Kristen di Gaza.

Keterangan Kementerian Dalam Negeri Hamas serangan Israel menyebabkan beberapa orang tewas dan terluka. AFP

Baca info selengkapnya di link bio. Jangan lupa follow Instagram @kumparancom untuk berita menarik lainnya!

#newsupdate #update #news #oneliner #gaza #israel #infoterkini #berita #beritaterkini #kumparan
Output: 0
Input: sorry.. paling ga respect sama gay or banci or apalah itu.. jiji njirr.. melawan kodrat.. aga redflag sih kalo ada
cewe yang temenan deket sama yang begitu..
Output: 1
Input: PARA PENDUKUNG ANIS 100% ORANG2 IDIOT/ MABUK AGAMA, COBA AJA SIMAK DARI MULAI POSTINGAN SAMPE OMONGAN NYA MIRIF ORA
NG DI HIPNOTIS
Output: 1
Input: MasyaAllah Alhamdulillah Bismillah... 👊 Zindabad bagi Muballigh Jemaat Ahmadiyah Indonesia🇮🇩 di Kabupaten Bone unt
uk UPAYA PROAKTIFnya memberi kontribusi dalam menjaga hubungan kemasyarakatan yang harnonis, sambil menjalin silaturahmi d
engan berbagai elemen masyarakat di daerahnya.
Output: 0
```

Figure 10: The Targeted Groups Page

The fifteen texts and annotations were chosen by the author manually. The order of prompt appearance is randomized using an integer seed of 42. The prompts contain 8 positive examples and 7 negative examples.