National Hate Speech Dashboard Codebook

v.1.0.

CSIS Indonesia

Table of Content

Introduction to the Dashboard

The phenomenon of hate speech is not new to Indonesia. Since its democratic consolidation, various actors have periodically instigated hate speech campaigns against vulnerable groups, particularly religious and ethnic minorities, to invigorate recruits, build solidarity, and mobilize members in supporting their socio-political agenda (George, 2016). Notably, these hate speech and incitements have led to significant physical and systematic harm to minorities such as Ahmadiyyas, Shi'as, Christians, and Chinese Indonesians. (Ahnaf et. al., 2015; Burhani, 2013; Setijadi, 2017).

While offline forms of hate speech are still present, there has also been an alarming increase of online hate speech content that has spread through Indonesia internet spaces in the past decade. Between 2015 and 2020, Indonesia's Cyber Patrol unit recorded over 7,460 reported cases of people spreading provocative content (in which hate speech and incitements are included). The significance of online hate speech is most evident in campaign against Jakarta governor, Basuki Tjahaja Poernama, where online hateful posts ultimately led to mass protests and his imprisonment.

Amidst its increasing prevalence, however, no tools in Indonesia are currently sufficient to quantitatively aggregate and accurately identify online hate speech trends at the national level. Existing reports and studies on Indonesian hate speech largely focus on individual hate speech campaigns (As'ad, 2009; KontraS Surabaya, 2012; Winarni, Agussalim, and Bagir, 2019). Even if they try to establish trends across cases, they are mostly using qualitative rather than quantitative methodology (Panggabean and Fauzi, 2015; Ahnaf et. al., 2015).

The National Hate Speech Dashboard is developed to fill this gap and provide quantitative visualization over online hate speech trends in Indonesia. Specifically, this dashboard will provide visualizations on two important trends: the volume of hate speech by topic by time, and the network mapping of actors that actively post or share hate speech content. By providing these trends, this dashboard aims to provide a better sense of the magnitude, trend, and pattern of hate speech and incitements issues in Indonesia, which can help embolden the issue's urgency of to the policymakers and public.

Focus of Observation: Ahmadiyyah, Shi'a, and Chinese Indonesians

Notably, Indonesia has experienced multiple hate speech campaigns that targets numerous vulnerable communities. Among others are the minority religious sects within Islam (e.g., Shi'a, Ahmadiyya), ethnic Chinese groups (Tionghoa), ethnic Papuan groups, and Christian communities. This list is not exhaustive as hate speech has also targeted the LGBTQ community and even political groups, such as the alleged relatives/acquaintances of the banned Indonesian Communist Party.

However, this dashboard will limit the focus of its observations towards hate speech that are targeted to three actors: Indonesian Ahmadiyyah, Shi'a Indonesians, and Chinese Indonesians. These actors are selected because in the past these communities have been targeted by some of the worst campaigns of hate speech (i.e., the 2005-2011 campaign against Ahmadiyyah, the 2006-2012 campaign against Shi'a Sampang, and the 2016-2017 campaign against Jakarta governor Basuki Tjahaja Purnama) – each campaign resulting into significant, and normalized, violation of their rights.

Source of Observation: Twitter

While online hate speech can come from various platforms and website, the National Hate Speech Dashboard has limited its data collection to tweets from Twitter. Per January 2020 Twitter is the fifth most used social media platform in Indonesia, with 56% of the country's population recorded to have used the application and/or website (Kemp, 2020). While its monthly traffic is not as large as other social media applications (i.e., Facebook, Youtube, Instagram) Twitter holds the second highest engagement rate compared to its competitors. Notably, Twitter has the second longest average time per visit, only below Youtube, and has the highest average pages per visit compared to the other platforms (Kemp, 2020).

Data Collection and Analysis Methodology

The data used to visualize the graphs in the National Hate Speech Dashboard is collected and analyzed using an independently created machine learning model. This machine learning model was specifically developed by the CSIS team in collaboration with to recognize patterns of online hate speech in Indonesia. This machine learning model was developed through a three-step process.

Data Scrapping

The first step in developing the machine learning model is to conduct data scrapping of relevant public tweets. The team collected the tweets based on three criteria.

1. *First*, the team collected tweets that were tweeted between January 1st 2020 to April 30th 2020.
2. *Second*, tweets are only collected if they contain more than four words.
3. *Third*, tweets are collected if they contain phrases related to the focus of the model's observation (i.e., Ahmadiyyah, Shi'a, Chinese Indonesians). Various spellings of the phrases are used to accommodate the many ways that tweets can refer to them.
4. *Fourth*, tweets are only collected if they use the Indonesian language and is geotagged in Indonesia. This language and geotag filter are used to ensure that all the tweets that were used to train the model are tweets which originate and can be understood by Indonesian twitter users.

Label Annotation

The second step in developing the machine learning model is to annotate the collected tweets – annotations that would become the main reference used to train, fit, and test the model. In doing so, the CSIS team manually annotated over 4000 tweets using a two-step validation process.

1. *First*, a team of trained annotators manually annotated individual tweets to determine which tweets "contain hate speech" and which tweets do not, based on a standardized definition as written in the coding manual.
2. *Second*, these annotations would then be checked by a smaller team of validators that would thoroughly record, crosscheck, and standardize discrepancies between annotations to ensure a common parameter and definition of hate speech was upheld.

Model Training, Validation, and Testing

The third step in developing the machine learning model is to train, validate, and test the model using the annotated tweets. This is done through a four-step process.

1. *First*, the annotated dataset is separated into a training dataset, a validation dataset, and a test dataset. To ensure the representativeness of different tweet characteristics in all the dataset reflect tweets in a real scenario, the sampling for this separation is randomized.
2. *Second*, the team used the training dataset as the basis to fit the initial model. It is this dataset that became the primary basis for the model to learn and recognize patterns of tweets containing hate speech.
3. *Third*, the team used the trained model to predict tweets that contained hate speech within the validation dataset. This allowed the team to evaluate the model's error rate and then tweak the existing model's hyperparameters to minimized it.
4. *Lastly*, the team used the testing dataset to evaluate the final model fit. Through this process, the team was able to develop a model that predicted tweets containing hate speech with 85% precision and 91% recall.

Definition of Hate Speech

The National Hate Speech Dashboard defines hate speech as **"any tweet that uses phrases which legitimize hostile actions or ascribe negative qualities towards the identity of a vulnerable community."** Operationally, for a tweet to be considered as hate speech, it must

1. *First*, target a person based on their identity. Meaning an insult towards the character of an individual's behavior without referencing to their identity is not considered as hate speech.
2. *Second*, the tweet uses phrases that legitimize hostility. These include phrases that actively incite people to conduct hostile actions (e.g., "burn," "kill") and phrases that diminishes the repugnance of hostile actions (e.g., "this is self-defense").
3. Or, *third*, the tweet uses phrases that ascribe negative qualities to the target's identity. This includes overt slurs and contextually pejorative dog-whistling phrases (e.g., "he is a communist").

This definition is derived from synthesizing two main academic definitions of hate speech. *First*, it refers to the United Nations legal definition of hate speech as recorded in the United Nations Strategy and Plan of Action on Hate Speech (2019). The document defines hate speech as "as any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factors."

Second, this National Hate Speech Dashboard also uses a similar definition provided by Parekh (2012), which defines hate speech as any expression that fulfills three criteria. First, the expression "is directed against a specified or easily identifiable individual or, more commonly, a group of individuals based on an arbitrary or normatively irrelevant feature," second, the expression "stigmatizes the target group by implicitly or explicitly ascribing to it qualities widely regarded as undesirable," and third, the expression portrays the "target group as an undesirable presence and a legitimate object of hostility."

Table 1. List of Definitions

United Nations (2019)	Parekh (2012)	Dashboard (2021)
"As any kind of communication in speech, writing or behavior …"	"Any expression."	"Any tweet…"
"… that attacks…"	"Portrays the target group as an undesirable presence and a *legitimate object of hostility*."	"…that uses phrases which legitimize hostility…"
"… or uses pejorative or discriminatory language …"	"Stigmatizes the target group by implicitly or explicitly ascribing qualities widely regarded as highly undesirable."	"… or ascribe negative qualities…"
"… with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factors."	"Is directed against a specified or easily identifiable individual or, more commonly, a group of individuals based on an arbitrary and normatively irrelevant feature."	"… towards a vulnerable community's identity."

Attribution Policy

Data and statistics from The National Hate Speech Dashboard can be freely used and downloaded provided that the following attribution policy is followed.

1. *First*, when data and/or the statistics from The National Hate Speech Dashboard is used in any way, the data and/or statistics must be acknowledged. This acknowledgement should include 1) a footnote with the full citation that includes a link to the website, 2) in text acknowledgement of the National Hate Speech Dashboard and that the data are available for public use, and 3) clear citation on any visuals using data from the National Hate Speech Dashboard.
2. *Second*, to reference the National Hate Speech Dashboard in a footnote, please cite: Lina Alexandra, Alif Satria, Edbert Gani Suryahudaya, and Beltsazar Krisetya, "The National Hate Speech Dashboard," *CSIS Indonesia*, (2021). https://hatespeech.csis.or.id
3. *Third*, to reference the National Hate Speech Dashboard Codebook in a footnote, please cite: Lina Alexandra, Alif Satria, Edbert Gani Suryahudaya, and Beltsazar Krisetya, "The National Hate Speech Dashboard Codebook," *CSIS Indonesia*, (2021).

Bibliography

Ahnaf, Mohamad Iqbal, Samsul Maarif, Budi Asyhari-Afwan, Muhammad Afdillah. *Politik Lokal dan Konflik Keagamaan: Pilkada dan Struktur Kesempatan Politik dalam Konflik Keagamaan di Sampang, Bekasi, dan Kupang*. Yogyakarta: CRCS, 2015.

As'ad, Muhammad. "Ahmadiyah and the Freedom of Religion in Indonesia." *Journal of Indonesian Islam* 3, No. 2 (2009): 390-413.

Burhani, Ahmad Najib. "When Muslims are Not Muslims: The Ahmadiyya Community and the Discourse on Heresy in Indonesia." PhD diss., University of California Santa Barbara, 2013.

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

Kemp, Simon. "Digital 2020: Indonesia." *DATAREPORTAL*, February 18, 2020. https://datareportal.com/reports/digital-2020-indonesia.

KontraS Surabaya. "Laporan Investigasi dan Pemantauan Kasus Syi'ah Sampang." (2012): 1-17.

Panggabean, Rizal and Ihsan Ali-Fauzi. *Policing Religious Conflicts in Indonesia*. Jakarta: Center for the Study of Religion and Democracy (PUSAD), 2015.

Parekh, Bhiku. "Is There a Case for Banning Hate Speech." In *The Content and Context of Hate Speech: Rethinking Regulation and Responses*, edited by Michael Herz and Peter Molnar, 37-56. New York, NY: Cambridge University Press, 2012.

Setijadi, Charlotte. "Ahok's Downfall and the Rise of Islamist Populsim in Indonesia." *ISEAS Perspective*, No. 38 (2017): 1-39.

United Nations Office on Genocide Prevention and the Responsibility to Protect. *United Nations Strategy and Plan of Action on Hate Speech*. New York, NY: UN Headquarters, 2019.

Winarni, Leni, Dafri Agussalim, and Zainal Abidin Bagir. "Memoir of Hate Spin in 2017 Jakarta's Gubernatorial Election: A Political Challenge of Identity against Democracy in Indonesia." *Religio: Jurnal Studi Agama-agama* 9, No. 2 (2019): 134-156.