Data sheet for
Misinformation Commands Attention: An English-Tagalog dataset of political discussions on Philippine Facebook Pages

Marc Riven Reyes Herrera, Kokil Jaidka, & Brendan Luyt

Motivation

For what purpose was the dataset created?

The dataset was created to investigate the dynamics of misinformation and engagement within political discourse on Facebook, specifically focusing on the Philippine political landscape. It aims to explore how misinformation influences political discussions and engagement, particularly during the post-election period.

Who created the dataset?

The dataset was curated and annotated by Marc Riven Reyes Herrera (Yale-NUS College), Kokil Jaidka (National University of Singapore), and Brendan Luyt (Nanyang Technological University).

Who funded the creation of the dataset?

This work was supported by the Singapore Ministry of Education AcRF TIER 3 Grant (MOE-MOET32022-0001), the MOE Tier 1 programme (WBS A-8000231-01-00), and an MoU of the Department of Communications and New Media, National University of Singapore with iSentia Brandtology Pte. Ltd.

Composition

What do the instances that comprise the dataset represent?

The instances in the dataset represent posts and comments from Filipino news outlets' Facebook pages, annotated for topics, misinformation, discussion quality, and incivility.

How many instances are there in total?

The dataset consists of 14,495 posts and 47,889 comments collected from Facebook.

Does the dataset contain all possible instances?

No, the dataset is a sample focused on posts and comments related to the 2022 Philippine Presidential Elections and its aftermath, spanning six months of political discourse.

What data does each instance consist of?

Each instance includes the text of the post or comment, engagement metrics (likes, shares, etc.), and annotations for misinformation, incivility, and discussion quality.

Is there a label or target associated with each instance?

Yes, labels include misinformation categories (e.g., factual error, unverified claims), civility levels, and discussion quality metrics.

Is any information missing from individual instances?

Instances are complete as per the dataset's defined scope. However, annotations may not capture all nuances of the discussions.

Are relationships between individual instances made explicit?

Yes, relationships are explicit, such as replies to specific posts or comments.

Are there recommended data splits?

No recommended splits are provided.

Are there any errors, sources of noise, or redundancies in the dataset?

The dataset may contain noise due to the inherent nature of social media data, including sarcasm, slang, and mixed languages (English and Tagalog).

Is the dataset self-contained, or does it link to external resources?

The dataset is self-contained. However, it relies on metadata and annotations derived from external analyses.

Does the dataset contain data that might be considered confidential?

No, the dataset only includes publicly available posts and comments.

Does the dataset contain data that might be offensive, insulting, or threatening?

Some comments may include uncivil or offensive language, reflecting the nature of political discussions online.

Collection Process

How was the data associated with each instance acquired?

Data was collected using the CrowdTangle API and sourced from Isentia, a commercial vendor. The collection focused on Facebook pages of Filipino news outlets.

What mechanisms were used to collect the data?

Posts and comments were identified using political keywords and annotated by trained Filipino annotators.

If the dataset is a sample, what was the sampling strategy?

A stratified sampling approach was used to balance posts about incumbents, opposition, and general political discourse.

Who was involved in the data collection process?

The authors and trained Filipino annotators with native-level proficiency in Tagalog.

Over what timeframe was the data collected?

The data spans six months following the 2022 Philippine Presidential Elections.

Were any ethical review processes conducted?

Yes, ethical considerations were followed, including anonymization and responsible data handling practices.

Preprocessing/Cleaning/Labeling

Was any preprocessing/cleaning/labeling of the data done?

Yes, preprocessing included filtering out empty posts and using tools like KeyBERT for keyword extraction. Labeling focused on misinformation types and civility.

Was the raw data saved in addition to the preprocessed data?

Yes, raw and preprocessed data were saved.

Is the software that was used to preprocess the data available?

Some tools, such as KeyBERT, were used for preprocessing. The code for annotation and analysis is documented but not explicitly shared.

Uses

Has the dataset been used for any tasks already?

Yes, it has been analyzed to study misinformation's impact on engagement and civility in political discourse.

Is there a repository linking to papers or systems that use the dataset?

Yes, the dataset is available at https://doi.org/10.7910/DVN/5IHQM5.

What (other) tasks could the dataset be used for?

The dataset can be used for:

- Studying misinformation dynamics.
- Analyzing political engagement patterns.
- Exploring the relationship between civility and misinformation.

Is there anything about the composition of the dataset that might impact future uses?

The dataset reflects the political and cultural context of the Philippines and may contain biases related to this specific setting.

Are there tasks for which the dataset should not be used?

The dataset is unsuitable for studying non-political or private data contexts.

Distribution

How will the dataset be distributed?

The dataset is publicly accessible via Harvard Dataverse at https://doi.org/10.7910/DVN/5IHQM5.

What license is the dataset distributed under?

The dataset is distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Maintenance

Who will be supporting/maintaining the dataset?

The authors will maintain the dataset.

Will the dataset be updated?

Future updates are not planned, but any revisions will be communicated through the distribution platform.