# Improving the Detection of Multilingual Online Attacks with Rich Social Media Data from Singapore

Janosch Haber[1*], Bertie Vidgen[2*], Matt Chapman[*], Vibhor Agarwal[3*],
Roy Ka-Wei Lee[4], Yong Keong Yap[5], and Paul Röttger[6*]

[1]Queen Mary University London [2]Alan Turing Institute [3]University of Surrey
[4]Singapore Univ. of Technology & Design [5]DSO National Laboratories [6]Univ. of Oxford

## Abstract

Toxic content is a global problem, but most resources for detecting toxic content are in English. When datasets are created in other languages, they often focus exclusively on one language or dialect. In many cultural and geographical settings, however, it is common to code-mix languages, combining and interchanging them throughout conversations. To shine a light on this practice, and enable more research into code-mixed toxic content, we introduce SOA, a new multilingual dataset of online attacks. Using the multilingual city-state of Singapore as a starting point, we collect a large corpus of Reddit comments in Indonesian, Malay, Singlish, and other languages, and provide fine-grained hierarchical labels for online attacks. We publish the corpus with rich metadata, as well as additional unlabelled data for domain adaptation. We share comprehensive baseline results, show how the metadata can be used for granular error analysis, and demonstrate the benefits of domain adaptation for detecting multilingual online attacks.

**Content warning**: This article contains illustrative examples of toxic content.

## 1 Introduction

Toxic content, such as hate speech and abuse, is a global problem, but most resources for detecting toxic content are in English (Vidgen and Derczynski, 2020; Poletto et al., 2021; Röttger et al., 2022a). This makes it difficult to develop effective models for detecting toxic content in other languages, and as a consequence, non-English speakers across the world are less protected against toxic content.

New datasets and models for non-English languages often focus exclusively on one language or dialect. In many cultural and geographical settings, however, languages and dialects are often *code-mixed*, i.e., combined or used interchangeably within a conversation or even a single utterance (Gibbons, 1987; Rijhwani et al., 2017). So far, this practice has received very limited attention in toxic content research, with most work on code-mixed content focusing on Hinglish, which is a mix of Hindi and English (e.g. Mathur et al., 2018a; Bohra et al., 2018; Sengupta et al., 2022).

In this article, we take a step towards addressing this issue by introducing **SOA**, a new multilingual dataset of **S**ingapore-centered **o**nline **a**ttacks. Singapore is a multilingual city-state in Southeast Asia, with five million inhabitants from a wide range of ethnic, religious, and cultural backgrounds. Singapore's official languages are English, Malay, Singaporean Mandarin and Tamil, but many other languages are widely spoken, including the code-mixed Singlish language and Indonesian, which is closely related to Malay. Using the r/Singapore subreddit as a starting point, we collect a large corpus of Reddit comments in Indonesian, Malay, Singlish and other languages. We select 15,000 comments for annotation with a diverse set of sampling methods. We provide fine-grained hierarchical labels for online attacks, as well as language identification, from trained, native-speaking annotators. We also publish rich metadata, such as timestamps, anonymised user IDs and source subreddit for all comments, and make available the complete unlabelled pool of 3,196,400 comments that the labelled data was sampled from.

For the new dataset, we share comprehensive baseline results for a suite of mono- and multilingual models, finding that Indonesian models adapted to Twitter data perform best out-of-the-box. We show how the rich metadata we provide can be leveraged for more granular error analysis, finding that the advantages of the Indonesian models over multilingual models stem from the language distribution in our data. Finally, we demonstrate how the unlabelled pool of comments we provide can be used for adapting models to the domain of our data, finding that this domain adaptation creates

[*]Work completed at Rewire.



*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*
*Volume 1: Long Papers*, pages 12705–12721
July 9-14, 2023 

clear performance benefits, especially for models not pre-trained on any social media data.

Overall, we make two main research contributions. 1) We publish a new dataset for multilingual online attacks in under-resourced languages with fine-grained hierarchical labels, rich metadata and additional resources. 2) We provide comprehensive baseline results, and demonstrate how metadata and additional resources can be used to evaluate and improve classification models. Together, we hope that these contributions will enable more research into code-mixed toxic content for under-resourced languages, and thus serve to improve how non-English speakers across the world are protected online.[1]

## 2 Taxonomy of Online Attacks

Research into toxic online content and its automated detection is marred by definitional challenges, with much disagreement about the exact characteristics of core concepts (Vidgen et al., 2019; Banko et al., 2020; Röttger et al., 2022). Following Poletto et al. (2021), we use toxicity as an umbrella term for various kinds of disruptive online content, with online attacks being a particular type of toxic content. Other types of toxic content include spam, sexually explicit language or the use of profanity. We define online attacks as *content that directs anger, aggression or maliciousness at an identifiable target. This includes insults, threats, and inciting harm and violence, and being overtly abusive. A range of entities, individuals and groups can be targeted by an attack.* Related concepts include abuse, which is a subset of online attacks directed at just individuals or groups, and hate speech, which is commonly defined as a subset of online attacks directed at groups with protected characteristics, such as race, gender or sexual orientation (Röttger et al., 2021).

The taxonomy of online attacks, which we introduce in this article and use for data annotation, is hierarchical and comprises two levels. The first level is binary, indicating whether content is an online attack based on our definition. If an online attack is present, the second level lists the potential targets of the attack, split into 1) individuals, 2) social groups, 3) the media, 4) institutions and government, and 5) other. Table 1 shows more details on each target, as well as example content. An online attack can have one or multiple targets.

[1]All data, annotation guidelines and code are available at github.com/rewire-online/singapore-online-attacks.

As it relates to other taxonomies of toxic content, our hierarchical setup takes inspiration from how Zampieri et al. (2019) classify offensive language. Talat et al. (2017) and Vidgen et al. (2021), like us, also differentiate between attacks targeting persons and attacks targeting groups. Vidgen et al. (2019) also separate out attacks against institutions. Our taxonomy is also more general, compared to work that focuses on specific targets of attacks, such as women (Guest et al., 2021; Zeinert et al., 2021), Muslims (Vidgen and Yasseri, 2020), or trans people (Lu and Jurgens, 2022).

## 3 Dataset

### 3.1 Data Collection

We collected all data from Reddit, a large online forum where discussions are organised into *subreddits* dedicated to particular topics or communities, via a public API (Baumgartner et al., 2020). To identify subreddits that are most relevant to the languages spoken in Singapore, we used a snowball sampling approach. We first identified the 1,000 users who made the most comments on the r/Singapore subreddit between August 2021 and August 2022. For each of these users, we collected their 1,000 most recent comments and extracted the name of the subreddit each comment was posted to, resulting in a list of circa 11,000 unique subreddits. This list was then filtered to subreddits which contained keywords related to Singapore as well as Singaporean languages in their names (e.g., "Sing", "SG", "Malay", "Indo", etc.). From this filtered list we manually selected the most relevant subreddits. This yielded a final list of 104 subreddits.

For each of the 104 subreddits, we collected all comments written before September 1st 2022, resulting in a total of 16,966,812 comments. Most of these comments were in English, reflecting Reddit's overall language bias. Since our project focuses on content written in Singaporean languages, such as Malay, Indonesian and Singlish, we used the Python language detection tool `lingua` to identify content which contained these languages. For each comment, `lingua` assigns a match probability to each of a set of specified languages. Identifying code-mixed content proved difficult, because it would often be predicted with high confidence as just one language, particularly English. We found that those comments that were predicted as Indonesian first and Malay second, or vice versa, were most likely to fit our language scope. Selecting only

| Target | Definition | Examples |
|---|---|---|
| Individuals | An identifiable individual that is either directly addressed or referred to. | "F*ck you dude", "Closed minded idiots", "He is a fool and a wh*re" |
| Social Groups | A group defined by protected characteristics such as race, gender or sexual orientation. | "The rapist is a black, typical behaviour", "Just like a pervert, he bats for both sides" |
| Media | Journalists, media organisations, and the media as a concept. | "Journos can s*ck my c*ck", "the media are infidels and satan worshippers" |
| Institutions & Govt. | Governments, official bodies, regulators, political bodies and political parties. | "Loong is a dangerous man, our PM is fricking n*nce", "Stupid goverment" |

Table 1: The four specified targets of online attacks in our taxonomy. Examples are in English, to be illustrative. This does not reflect the language distribution in our dataset, where English is excluded through filtering.

those comments resulted in a pool of 3,196,400 comments. From this pool, we selected 15,000 comments for model training and evaluation, using three sampling strategies.

**1) Keyword sampling** We sampled the first 9,000 comments using a keyword-search approach, to increase the proportion of online attacks in our dataset. For this purpose, we created a list of 229 attack-related key terms in Indonesian, Malay and Singlish with support from our native-speaking annotators. For instance, "bondol" is a Malay word that means "loony" in English, "chao" is a Singlish word which means "smelly", and "makan tai" is an Indonesian phrase which means "eat sh*t". We then filtered the unlabelled pool of 3,196,400 comments to include only those comments containing at least one keyword. This resulted in 138,361 comments, from which we sampled at equal rates using each keyword to obtain 9,000 comments.

**2) Active learning** We sampled the next 5,000 comments using two rounds of active learning. Active learning is a sampling method, whereby an initial model selects further entries for annotation that are expected to be particularly informative to it. This method has been shown to be effective for toxic content detection (Markov et al., 2022; Kirk et al., 2022). In the first round of active learning, we fine-tuned an initial XLM-R model (Conneau et al., 2020) on the 9,000 comments collected with keyword sampling, and ran inference with it over 100,000 random comments from the unlabelled pool. Then, we selected 3,000 comments – 1,500 comments about which the model was most uncertain, to address gaps in model coverage, and 1,500 comments for which the model was maximally certain they were online attacks, to address potential false positive issues. We trained another XLM-R model on the now-total 12,000 comments, and repeated the process, sampling another 1,000 comments of each type, so that in total we collected 5,000 comments with active learning. While the outcome of the active learning process is somewhat contingent on the model used for it, we found that complementing keyword-based sampling with this method resulted in more diverse data, which we expect to be useful for any model.

**3) Random sampling** We sampled a final 1,000 comments randomly from the unlabelled pool, so that we could have a portion of the test set that reflects a more realistic distribution of online attacks (see experimental setup in §4.1). In the dataset, we specify for each comment which sampling method was used to select it.

### 3.2 Data Annotation

We recruited a team of 14 annotators through Upwork, a crowdworking platform. All annotators were screened using a set of example annotation questions, and then onboarded and trained for our annotation task. We also asked all annotators to complete a short survey about themselves. 11 of the annotators are Indonesian, two are Malaysian and one declined to give this information. All an-

notators primarily resided in the country of their nationality. Annotators could enter their ethnicity as free-text. Two identified as Chinese, two as Malay, two as Gorontalo, and two as Asian. One identified as Minang, one as Javanese, one as being from Flores, and one from Sumatra. One annotator identified as mixed and another declined to give this information. All annotators were intermediate or fluent in English, and 12 were native, or near-native, in Indonesian. Six were intermediate or better in Malay, and many could speak other languages such as Japanese, Javanese, Korean and Tagalog. Ten of the annotators identify as women and four as men. Eight of the annotators are aged 18-29 years old and six are 30-39 years old.

Each annotator worked independently, labelling comments assigned to them according to extensive annotation guidelines based on our taxonomy of online attacks (§2). The annotators labelled whether comments contained an online attack or not, and if they did, they selected the target(s) of the attack. They also labelled the language(s) in which each comment was written. The 15,000 comments were annotated in six batches, with 10-14 annotators working on each batch. Annotation was *prescriptive* (Röttger et al., 2022), in the sense that we tasked annotators with applying the guidelines rather than their subjective beliefs. Each comment was annotated by three annotators. We make all annotations with anonymised annotator IDs available for each comment in order to enable further analysis of human label variation (Plank, 2022).

For the primary binary attack label, there was 3/3 agreement on 49.4% of comments, and 2/3 agreement on the rest. Fleiss' Kappa is 0.314.

For the language label, there was 3/3 agreement on 70.2% of comments. 3.9% of comments had three-way disagreement on the language label, which we resolved through expert annotation.

Throughout the annotation process, we followed guidelines by Davani et al. (2022) to protect the wellbeing of our annotators. Annotators were compensated at a rate of $16 per hour, well above the living wage in their countries of residence.

### 3.3 Descriptive Statistics

**Attack** 6,173 comments (41.2%) were majority-labelled as containing an attack, while 8,827 comments (58.8%) were majority-labelled as not containing an attack. Of the 6,173 attacks, based on majority labels, 4,356 attacks (70.6%) target an individual, 534 (8.7%) target a social group, 428 (6.9%) target an institution, 78 (1.3%) target the media, and 14 (0.2%) were labelled with another target in a free text field (e.g. "Animal", "Convenience Store", and "Place"). For 1,199 attacks (19.4%), there is no majority agreement on a target.

**Language** 12,212 comments (81.4%) were majority-labelled as Indonesian, followed by 1,635 comments (10.9%) labelled as Malay and 218 comments (1.5%) were majority-labelled as Singlish. The remaining 688 (4.6%) comments were marked as containing one of dozens of other languages spoken in or around Singapore, such as Javenese and Hokkien Chinese, and code-mixed combinations thereof. This imbalance is created by our language filtering, which favours Indonesian and Malay – the two languages being very similar (see Section 3.1). Both languages often code-mix with English (e.g. *"Straight outta horror movie, jangan2 kerasukan makhluk halus"*). For details on the language distribution, see Appendix A.1

**Subreddit** The 15,000 comments in our labelled datasets are from 26 different subreddits, out of 104 subreddits initially selected for data collection. A large majority of 12,561 comments (83.7%) is from the r/indonesia subreddit, followed by 1,389 comments (9.3%) from r/malaysia, 272 comments (1.8%) from r/malaygonewild and 239 comments (1.6%) from r/singapore. This skewed distribution is a consequence of our sampling methods and our language filtering, which did not explicitly account for subreddit sources. As a result, the largest subreddits with the most activity in in-scope languages, like r/indonesia, are most represented in our data. For details, see Appendix A.2

**Time** The oldest comment in our labelled dataset is from May 2011, and the most recent from August 31st 2022, which is the end of our sampling period. Most comments were written in more recent years, with 3,672 comments (24.5%) from 2022, 4,142 comments (27.6%) from 2021, and 3,028 comments (20.2%) from 2020. By contrast, only 293 comments (2.0%) were written before 2017. This reflects general growth trends in Reddit activity.[2] For details, see Appendix A.3

**Authorship** We replace comment author names with alphanumeric IDs. The 15,000 comments in our labelled dataset were written by 5,307 different

[2] See, for example, https://subredditstats.com/r/indonesia.

authors. 3,303 authors (62.2%) wrote just a single comment in the dataset. 763 authors (14.4%) wrote two comments, and 376 authors (7.1%) wrote three. 70 authors (1.3%) wrote ten or more comments, with 179 being the largest number of comments from a single author. For details, see Appendix A.4

## 4 Experiments

### 4.1 Experimental Setup

We show results for three sets of experiments. The task is always the binary distinction between content containing or not containing an online attack. Our primary goal is not to develop a best-performing classifier for our task, but rather to provide baseline results and demonstrate the usefulness of the additional resources and metadata we share along with the labelled dataset.

**Model Parameters** We use the same standard parameters across all models we evaluate. In training, the learning rate is 1e-05, and the batch size 16. The maximum input length is 256 tokens, which affects less than 1% of our data. We train for a maximum of 10 epochs, with early stopping based on development set cross-entropy loss, and a patience of three epochs. None of the models trained for more than six epochs. We do not perform any further hyperparameter optimisation.

**Data Splits** We split the 15,000 labelled comments into 10,000 comments for model training, 2,000 for validation and 3,000 for testing. The 3,000 comments for testing include all 1,000 comments selected with random sampling, to reflect a more realistic distribution of online attacks (§3.1). The test set therefore contains 945 comments (31.8%) labelled as online attacks.

**Preprocessing** For all comments, we collapse whitespaces, and remove linebreaks and HTML artefacts. We replace user mentions in the format of 'u/username' with a [USR] token, and URLs with a [URL] token.

**Evaluation Metrics** We use macro F1 as an overall measure of performance, and evaluate performance on attacks, i.e. the positive class, based on precision and recall, given as percentages.

### 4.2 Baseline Models

For our baseline experiments, we evaluate six models. Three models are multilingual models, chosen for their widespread use and/or competitive performance on toxic content detection tasks (see e.g. Röttger et al., 2022a): mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and XLM-T (Barbieri et al., 2021), which is XLM-R adapted to the Twitter domain through continued pre-training. Two models are monolingual Indonesian models, chosen because of the large amount of Indonesian content in our labelled dataset, and the high similarity between Indonesian and Malay: IndoBERT (Koto et al., 2020), and IndoBERTweet (Koto et al., 2021), which is IndoBERT adapted to Twitter, analogous to XLM-T and XLM-R.[3] Finally, we translate the train, validation and test set to English using the Google Translate API, and evaluate a monolingual English DeBERTA-v3 model (He et al., 2021). None of the models are case sensitive.

In addition to the six models, we show results for three naive baselines: one model that always predicts attack, one that never predicts attack, and one that predicts each label with equal probability. All results are shown in Table 2.

| Model | Prec. | Rec. | Macro F1 |
|---|---|---|---|
| mBERT | 61.9 | 58.9 | 71.3 |
| XLM-R | 65.8 | 68.2 | 75.6 |
| XLM-T | 71.0 | 68.0 | 77.9 |
| IndoBERT | 65.4 | 64.0 | 74.3 |
| IndoBERTweet | **73.8** | **68.2** | **79.1** |
| DeBERTa | 73.1 | 49.6 | 72.1 |
| Always attack | 31.8 | 100 | 24.1 |
| Never attack | 0.0 | 0.0 | 40.5 |
| Equal prob. | 31.8 | 50.0 | 48.3 |

Table 2: Baseline results. Precision and recall are for attacks, i.e. the positive class. Best model performance is highlighted in **bold** (excl. naive baselines).

We find that the Twitter-adapted models perform best overall, with 79.0 macro F1 for IndoBERTweet and 77.8 for XLM-T. Adaptation has a larger effect on the IndoBERT models (4.8 points difference) than on the XLM models (2.2 points difference). Precision on attacks is generally higher than recall across models, except for XLM-R, where precision is 65.8 and recall is 68.2. The worst-performing model is mBERT, with 71.3 macro F1. The DeBERTa model trained and evaluated on auto-

[3] Wilie et al. (2020) introduce another model also called IndoBERT, which we do not test in this article but would expect to give similar results.

translated data performs second-worst, with recall below the 50% from random guessing. The naive baselines perform strictly worse than all other models in terms of macro F1 and precision on attacks.

### 4.3 Error Analysis

Each comment in our dataset comes with rich metadata, which includes the comment language, timestamp, anonymised user ID, and the subreddit that the comment was posted to. This metadata can be used to perform fine-grained error analysis and diagnose specific model strengths and weaknesses. To demonstrate this, we analyse the predictions of XLM-T and IndoBERTweet, the two strongest baseline models, across different languages. Table 3 shows macro F1 scores for the different languages in our 3,000-comment test set.

| Language | n | XLM-T | IndoT |
|---|---|---|---|
| Indonesian | 2,476 | 77.1 | **78.9** |
| Malay | 276 | **80.1** | 78.6 |
| English + Indo. | 94 | 74.1 | **74.4** |
| Singlish | 38 | **71.9** | **71.9** |
| English | 37 | **100** | 82.6 |
| English + Malay | 19 | **84.2** | 72.5 |
| Other | 50 | **76.2** | 74.0 |

Table 3: Macro F1 for XLM-T and IndoBERTweet on the 3,000-comment test set, split by comment language. Best model performance is highlighted in **bold**.

We find that the IndoBERTweet model, which performs best overall (Table 2), outperforms XLM-T on Indonesian and code-mixed Indonesian content, but falls behind on Malay, English and Other languages. *"Jadi gini mbak, rasanya k*ntol saya pengen saya cekek deh liat mbak soalnya mbak ngomongnya dah kek k*ntol"*, for example, is correctly identified as an online attack by IndoBERTweet, but not by XLM-T. *"You sure, I used to be quite effeminate in sekolah rendah and got called p*ndan too."*, on the other hand, is classified correctly an non-attack by XLM-T and misclassified by IndoBERTweet.

Further, we can leverage the secondary labels, which specify for all online attacks which target is attacked, for error analysis. Table 4 shows accuracy on the 945 online attacks in our 3,000-comment test set, split by the five different target categories.[4]

| Target | n | XLM-T | IndoT |
|---|---|---|---|
| Individual | 679 | 70.3 | **72.5** |
| Social Group | 81 | **70.4** | 65.4 |
| Institution | 69 | **78.3** | 73.9 |
| Media | 11 | **72.7** | 63.6 |
| Other | 4 | 75.0 | **100** |

Table 4: Accuracy for XLM-T and IndoBERTweet on the 945 attacks in the test set, split by attack target. Best model performance is highlighted in **bold**.

We find that the IndoBERTweet model, which has the best precision and recall on attacks overall (Table 2), actually performs worse than XLM-T on all attack targets except for attacks on individuals, which are by far the most common type of attack in our dataset, and "other" targets, which are extremely rare. *"Damkar gak mau menanggapi panggilan darurat dan malah ngehalu ujungnya bakal gantian mereka yang dibakar massa"*, for example, which attacks an institution (the fire department), is misclassified by IndoBERTweet but not XLM-T. *"b*debah ini yg komen, sy bikin meme OG sendiri"*, on the other hand, attacks a person, and is classified correctly by IndoBERTweet while being misclassified by XLM-T.

### 4.4 Domain Adaptation

We also provide a large unlabelled pool of 3,196,400 comments along with the 15,000 labelled comments (§3.1). These unlabelled comments can be used to adapt pre-trained language models to the specific domain of our data through continued pre-training. This approach to domain adaptation has been found to improve model performance on a wide variety of downstream tasks (e.g. Alsentzer et al., 2019; Lee et al., 2020; Gururangan et al., 2020; Röttger and Pierrehumbert, 2021).

We randomly sample 100,000 comments for domain adaptation from the unlabelled pool, and then continue pre-training each of our baseline models on these comments for one epoch with default hyperparameters on a masked language modelling objective.[5] Then, we fine-tune these newly-adapted models in the same way as our baseline models. We show macro F1 comparisons in Table 5, and more detailed results in Appendix B.

We find that all models benefit from domain

[4]We show accuracy because all comments we test on here are online attacks, i.e. belong to the same class.

[5]We exclude the DeBERTa model because it would require translation of the larger unlabelled pool of comments.

| Model | Baseline | Adapted | Change |
|---|---|---|---|
| mBERT | 71.3 | 74.0 | +2.7 |
| XLM-R | 75.6 | 76.9 | +1.3 |
| XLM-T | 77.9 | 77.6 | -0.2 |
| IndoBERT | 74.3 | 77.5 | +3.3 |
| IndoBERTweet | 79.1 | **79.9** | +0.9 |

Table 5: Macro F1 for domain-adapted models compared to baselines. Best model performance is highlighted in **bold**, positive changes in blue, and negative changes in red. Change is in percentage points.

adaptation, except for XLM-T. Weaker models, like mBERT (+2.7 macro F1), tend to benefit more than stronger models, like XLM-R (+1.3 macro F1). Further, models already adapted to social media data from Twitter have very little benefit from domain adaptation with Reddit data (IndoBERTweet, +0.9 macro F1), or even suffer a slight performance decrease (XLM-T, -0.2 macro F1).

## 5 Discussion

The results for our baseline models suggest clear benefits from model scale for detecting online attacks in our dataset. XLM-R is much like mBERT, but it has more model parameters and was pre-trained on a larger corpus. Accordingly, it performs much better than mBERT.

Our results also show the benefits of adapting models to social media data, even if adaptation data and task data come from different social media platforms. XLM-R and XLM-T, like IndoBERT and IndoBERTweet, are the same, except for additional pre-training on Twitter data. In our baseline results (Table 2), this adaptation has a clear positive effect, with the adapted models outperforming all others. In our own domain adaptation experiments (Table 5), however, models that were already adapted to Twitter data did not substantially benefit from further adaptation with Reddit data. This suggests that most of the benefit of adaptation comes from capturing language use that is shared between Twitter and Reddit. On the other hand, we perform our own domain adaptation experiments with just 100,000 Reddit comments, whereas XLM-T and IndoBERTweet, respectively, are adapted with 198 million and 26 million tweets. On our dataset, XLM-R adapted with our Reddit comments performs roughly on par with XLM-T (Table B). This suggests that, even if large amounts of Twitter data are as useful for adaptation, it may be more efficient to learn from Reddit, the target platform.

Multilingual models do not appear to have an advantage over Indonesian monolingual models for our dataset. This can likely be explained by Indonesian content making up most of the dataset (§3.3), and other languages in the dataset, like Malay and Singlish, sharing a lot of similarity with Indonesian. As we found in our error analysis, the monolingual Indonesian IndoBERTweet model outperforms XLM-T, the strongest multilingual model, on Indonesian content (Table 3), but performs worse on most other languages. This aligns with evidence on the *curse of multilinguality* (Conneau et al., 2020; Pfeiffer et al., 2022), which describes the trade-off between language coverage and monolingual performance for fixed model sizes.

Overall, the dataset appears to be moderately challenging for models, with performance differences between baselines that align with general intuition and other research. However, there are also some limitations to our dataset and experiments, which we discuss in a separate Section following the Conclusion below.

## 6 Related Work

### 6.1 Multilingual Toxic Content Detection

Most resources for detecting toxic content focus on English only (Vidgen and Derczynski, 2020; Poletto et al., 2021; Röttger et al., 2022a), which mirrors an overall imbalance in natural language processing (Joshi et al., 2020). More recently, researchers have started to create more multilingual toxic content datasets, which usually consist of an English portion and separate portions in other languages. Basile et al. (2019), for example, collect Spanish and English hate speech against women and immigrants from Twitter. Modha et al. (2021) provide datasets for offensive language in English, Hindi and Marathi (see also Mandl et al., 2019, 2020). Ousidhoum et al. (2019) collect hate speech in English, French and Arabic, using separate sets of keywords. Röttger et al. (2022b) create functional test suites for hate speech detection models in ten different languages. By contrast, we create a single dataset, which includes a variety of languages (§3.3), and we explicitly filtered out English-only content, which is already well-represented in the research. We use a single sampling method to collect multilingual data from multilingual communities, rather than collecting data in different languages from different communities.

### 6.2 Cross-Lingual Toxic Content Detection

Another closely-related stream of research focuses on cross-lingual toxic content detection, where large multilingual language models are first fine-tuned on a resource-rich source language – often English – and then applied to another target language. This is relevant to our work, as our dataset contains large amounts of content in some languages, like Indonesian, and relatively little content in many other languages, like Singlish (see Appendix A.1). For detecting toxic content, like online attacks, research has generally found that some target language content is necessary for good performance, but very little data goes a long way (Leite et al., 2020; Stappen et al., 2020; Nozza, 2021; Bigoulaeva et al., 2021; Pelicon et al., 2021; Röttger et al., 2022a). Therefore, we would expect our dataset to be a useful resource for the wide range of languages and dialects that it covers, even if it only contains a few entries in some languages.

### 6.3 Code-Mixed Toxic Content

Code-mixed toxic content, where languages are combined and used interchangeably within conversations or single utterances, has received little research attention. Most work so far focuses on Hinglish, which is a mix of English and Hindi. Mathur et al. (2018b) and Mathur et al. (2018a) each create a dataset of offensive tweets in Hinglish, and train baseline models by first translating content to English, which resembles our translation baseline (§4.2). Kapoor et al. (2019) use the dataset released by Mathur et al. (2018b) to train stronger LSTM models. Bohra et al. (2018) create a dataset of Hinglish tweets labelled for hate speech. Kumar et al. (2018) annotate Hinglish content from Twitter and Facebook for aggression. Sengupta et al. (2022) train and evaluate simple transformer models across several of these datasets. By contrast, to our knowledge, we introduce the first dataset for code-mixed Singaporean languages, including Singlish as well as Indonesian and Malay content that borrows English words.

### 6.4 Toxic Content in Singaporean languages

Among the languages we focus on in this article, only Indonesian has received some dedicated attention in toxic content research. Alfina et al. (2017) share a small dataset of 520 Indonesian Twitter posts labelled for hate speech, along with baseline models. Pratiwi et al. (2018) create a dataset of 1,200 Indonesian Instagram comments, also labelled for hate speech. Ibrohim and Budi (2018) label 2,500 Indonesian tweets for abuse. Ibrohim and Budi (2019) then combine and expand the previous three datasets, and provide results for simple baseline models such as a random forest classifier. Similarly, Elisabeth et al. (2020) use the Ibrohim and Budi (2018) dataset, and provide additional annotations for implicit hate. Our dataset contains a large amount of Indonesian comments – more than any of the existing Indonesian datasets – but it also contains content in Malay, Singlish and other regional dialects, like Javanese. To our knowledge, our dataset is the first in toxic content research to cover these language.[6]

## 7 Conclusion

Online attacks and other forms of toxic content are a global problem. This is not reflected in the available resources for detecting toxic content, which are mostly in English. As a consequence, non-English models for toxic content detection are less effective, and non-English speakers across the world are less protected from toxic content. When non-English resources are created, they often focus on single languages. By contrast, in this article, we focused on multilingual code-mixed content.

We introduced a dataset of multilingual online attacks, using Reddit community of the multilingual city-state of Singapore as our starting point for data collection. From the unlabelled data we collected, which covers Indonesian, Malay, Singlish and other languages, we sampled 15,000 comments for annotation using diverse sampling methods. We provided fine-grained hierarchical labels for online attacks, and also shared rich metadata as well as the unlabelled pool of 3,196,400 comments along with the labelled data.

We shared comprehensive baseline results for the new dataset, finding strong out-of-the-box performance for multilingual and monolingual Indonesian models adapted to Twitter data. We conducted an error analysis, using language metadata and secondary attack labels to gain granular insights into model performance. Finally, we showed how the unlabelled data we provide can be used for domain adaptation, showing that this particularly benefits models not already adapted to social media data.

To our knowledge, our toxic content dataset and

[6]For a similar non-toxic resource relevant to Indonesian dialects, see the NusaX corpus (Winata et al., 2023).

experiments are the first for code-mixed Singaporean languages. With our contributions, we hope to enable more research into code-mixed toxic content, especially for such under-resourced language settings. This research is needed to develop more effective models for multilingual toxic content detection, and therefore to improve how billions of non-English are protected online.

## Acknowledgments

We thank all annotators for their work, and all reviewers for their constructive feedback.

## Limitations

**Dataset** All our data was sampled from a single social media platform, over a long but static time span. This limits the generalisability of models trained on our dataset, and the conclusions that can be drawn from model performance on our dataset.

Our sampling methods did not account for language and subreddit information. Therefore, the language and subreddit distributions in our labelled dataset are extremely skewed, broadly matching the distributions in our unlabelled pool. Most comments are in Indonesian, and from the r/Indonesia subreddit. While other languages and subreddits are represented, this still increases the specificity of our dataset, and limits the scope of our insights.

Despite the prescriptive annotation process and the training of native-speaking annotators, disagreement on the attack labels in our dataset is high. This suggests that there are many challenging cases in our dataset, as annotators tend to agree on more extreme cases (Salminen et al., 2019). The disagreement also likely creates some inconsistencies in the majority labels, which limits optimal model performance on the dataset.

**Experiments** The primary goal of our experiments was to 1) provide useful baseline results, and 2) demonstrate how the additional resources and metadata, which we share along with the labelled dataset, can be used to further improve the detection of multilingual online attacks. Therefore, we did not focus on optimising the performance of the models we trained and evaluated. It is very possible, that the same models we used could be more effective with different hyperparameters.

We also did not re-run our experiments for many different random seeds, which limits our ability to test for statistically significant differences in performance. Initial experiments did not reveal much randomness in performance, which is expected given the relatively large size of our labelled training set. Further, we see relatively large differences in performance across models, and the differences match clear intuitions.

## Ethical Considerations

**Annotator Wellbeing** As outlined in §3.2, we followed guidelines by Davani et al. (2022) to protect the wellbeing of our annotators. Annotators were clearly informed about the nature of the annotation task before commencing their work. They completed their work in batches, on their own schedules, and could decide to withdraw from the work at any point. Compensation for annotators was well above the living wage in their countries of residence, at $16 per hour. We do not release identifiable information about our annotators.

**Data Privacy** We used Reddit data made publicly available via the Pushshift API (Baumgartner et al., 2020) rather than scraping any new data ourselves. Comment author usernames are anonymised by replacing them with alphanumeric IDs.

**Environmental Impact** We only trained a handful of models in our experiments, and did not perform any hyperparameter tuning. Relative to the concerns raised around the environmental costs of pre-training large language models (Strubell et al., 2019; Henderson et al., 2020; Bender et al., 2021), or even larger-scale fine-tuning with hyperparameter tuning, we therefore consider the environmental costs of our work to be relatively minor.

## References

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. 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. IEEE.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A Unified Taxonomy of Harmful Content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 125–137. Association for Computational Linguistics.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2021. XLM-T: A multilingual language model toolkit for twitter. *arXiv preprint arXiv:2104.12250*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Irina Bigoulaeva, Viktor Hangya, and Alexander Fraser. 2021. Cross-lingual transfer learning for hate speech detection. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 15–25, Kyiv. Association for Computational Linguistics.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A Dataset of Hindi-English Code-Mixed Social Media Text for Hate Speech Detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 36–41. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Marco Del Tredici and Raquel Fernández. 2017. Semantic variation in online communities of practice. In *IWCS 2017 - 12th International Conference on Computational Semantics - Long papers*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Damayanti Elisabeth, Indra Budi, and Muhammad Okky Ibrohim. 2020. Hate code detection in indonesian tweets using machine learning approach: A dataset and preliminary study. In *2020 8th International Conference on Information and Communication Technology (ICoICT)*, pages 1–6. IEEE.

John Gibbons. 1987. *Code-mixing and code choice: A Hong Kong case study*, volume 27. Cambridge University Press.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43.

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.

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.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Raghav Kapoor, Yaman Kumar, Kshitij Rajput, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Roger Zimmermann. 2019. Mind your language: Abuse and offense detection for code-switched languages. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9951–9952.

Hannah Kirk, Bertie Vidgen, and Scott Hale. 2022. Is more data better? re-thinking the importance of efficiency in abusive language detection with transformers-based active learning. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 52–61, Gyeongju, Republic of Korea. 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.

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated corpus of Hindi-English code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Christina Lu and David Jurgens. 2022. The subtle language of exclusion: Identifying the toxic speech of trans-exclusionary radical feminists. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 79–91, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Forum for information retrieval evaluation*, pages 29–32.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th forum for information retrieval evaluation*, pages 14–17.

Todor Markov, Chong Zhang, Sandhini Agarwal, Tyna Eloundou, Teddy Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2022. A holistic approach to undesired content detection in the real world. *arXiv preprint arXiv:2208.03274*.

Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. 2018a. Did you offend me? classification of offensive tweets in hinglish language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 138–148.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018b. Detecting offensive tweets in hindi-english code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26.

Sandip Modha, Thomas Mandl, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Tharindu Ranasinghe, and Marcos Zampieri. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech. In *Forum for Information Retrieval Evaluation*, FIRE 2021, page 1–3, New York, NY, USA. Association for Computing Machinery.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

*9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Andraž Pelicon, Ravi Shekhar, Blaž Škrlj, Matthew Purver, and Senja Pollak. 2021. Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science*, 7:e559. Publisher: PeerJ Inc.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Nur Indah Pratiwi, Indra Budi, and Ika Alfina. 2018. Hate speech detection on indonesian instagram comments using fasttext approach. In *2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 447–450. IEEE.

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Shekhar Maddila. 2017. Estimating code-switching on Twitter with a novel generalized word-level language detection technique. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1971–1982, Vancouver, Canada. Association for Computational Linguistics.

Paul Röttger, Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022a. Data-efficient strategies for expanding hate speech detection into under-resourced languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5674–5691, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Paul Röttger and Janet Pierrehumbert. 2021. Temporal adaptation of BERT and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2400–2412, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022b. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Talat, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Joni Salminen, Hind Almerekhi, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J. Jansen. 2019. Online hate ratings vary by extremes: A statistical analysis. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, CHIIR '19, page 213–217, New York, NY, USA. Association for Computing Machinery.

Ayan Sengupta, Sourabh Kumar Bhattacharjee, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Does aggression lead to hate? detecting and reasoning offensive traits in hinglish code-mixed texts. *Neurocomputing*, 488:598–617.

Lukas Stappen, Fabian Brunn, and Björn W. Schuller. 2020. Cross-lingual zero- and few-shot hate speech detection utilising frozen transformer language models and AXEL. *CoRR*, abs/2004.13850.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Zeerak Talat, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Bertie Vidgen and Taha Yasseri. 2020. Detecting weak and strong islamophobic hate speech on social media. *Journal of Information Technology & Politics*, 17(1):66–78.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating online misogyny. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3181–3197, Online. Association for Computational Linguistics.

# A Additional Descriptive Statistics

## A.1 Language Distribution

The 15,000 comments in our labelled dataset comprise 69 unique combinations of languages.

| Language | n | % of Data |
|---|---|---|
| Indonesian | 12,212 | 81.4 |
| Malay | 1,635 | 10.9 |
| Indonesian and English | 396 | 2.6 |
| Singlish | 218 | 1.5 |
| Malay and English | 131 | 0.9 |
| Javanese | 92 | 0.6 |
| English | 85 | 0.6 |
| Sundanese | 46 | 0.3 |
| Javanese and Indonesian | 23 | 0.2 |
| Sundanese and Indonesian | 20 | 0.1 |
| Chinese | 11 | 0.1 |
| Other | 121 | 4.0 |

Table 6: Distribution of languages and language combinations for the 15,000 comments in our labelled dataset. Languages or language combinations present in fewer than 10 comments, such as Hokkien Chinese, Arabic and Russian, are combined as 'Other'.

## A.2 Subreddit Distribution

The 15,000 comments in our labelled dataset come from 26 different subreddits.

| Subreddit | n | % of Data | % Attacks |
|---|---|---|---|
| indonesia | 12,561 | 83.7 | 39.7 |
| malaysia | 1,389 | 9.3 | 51.1 |
| malaygonewild | 272 | 1.8 | 61.4 |
| singapore | 239 | 1.6 | 26.8 |
| MalaysGoneWild | 201 | 1.3 | 54.2 |
| Ajar_Malaysia | 89 | 0.6 | 23.6 |
| MalaysianFappers | 49 | 0.3 | 57.1 |
| malaysians | 35 | 0.2 | 37.1 |
| NegarakuMalaysia | 35 | 0.2 | 37.1 |
| SeksiArtisMalaysia | 24 | 0.2 | 79.2 |
| SingaporeRaw | 20 | 0.1 | 35.0 |
| malaysiasecretlab | 17 | 0.1 | 64.7 |
| MalaysNSFW | 15 | 0.1 | 60.0 |
| IndoR4R | 13 | 0.1 | 7.7 |
| NSFW_Malaysia | 11 | 0.1 | 63.6 |
| askSingapore | 8 | 0.1 | 12.5 |
| SGExams | 6 | 0.0 | 16.7 |
| Other | 11 | 0.4 | 25.0 |

Table 7: Subreddit distribution for the 15,000 comments in our labelled dataset. Subreddits from which we sampled fewer than 5 comments are combined as 'Other'

## A.3 Temporal Distribution

The earliest comment in the labelled dataset was published on May 19th 2011, and the most recent comment on August 31st 2022.

| Year | n | % of Data | % Attacks |
|---|---|---|---|
| 2022 | 3,672 | 24.5 | 38.7 |
| 2021 | 4,142 | 27.6 | 39.6 |
| 2020 | 3,028 | 20.2 | 42.1 |
| 2019 | 2,084 | 13.9 | 40.9 |
| 2018 | 1,076 | 7.2 | 46.7 |
| 2017 | 705 | 4.7 | 51.9 |
| 2016 | 101 | 0.7 | 44.6 |
| 2015 | 113 | 0.8 | 39.8 |
| 2014 | 61 | 0.4 | 32.8 |
| 2013 | 10 | 0.1 | 30.0 |
| 2012 | 5 | 0.0 | 60.0 |
| 2011 | 3 | 0.0 | 66.7 |

Table 8: Distribution of the 15,000 labelled comments across years covered by the dataset.

## A.4 Author Distribution

The most active author in our labelled dataset of 15,000 comments made 179 comments. This analysis is based on anonymised author IDs.

| Comments | Users | % of Users |
|---|---|---|
| 1 | 3,303 | 62.2 |
| 2 | 763 | 14.4 |
| 3 | 376 | 7.1 |
| 4 | 194 | 3.7 |
| 5 | 150 | 2.8 |
| 6 | 105 | 2.0 |
| 7 | 63 | 1.2 |
| 8 | 46 | 0.9 |
| 9 | 36 | 0.7 |
| 10+ | 70 | 1.3 |

Table 9: Distribution of comment counts for the 5,307 users contributing to the labelled dataset.

## A.5 Attack Types

6,173 (41.15%) out of 15,000 comments were majority-labelled as containing an online attack.

| Attack Target | n | % of Attacks |
|---|---|---|
| Person | 4,356 | 70.6 |
| Media | 78 | 1.3 |
| Social Group | 534 | 8.7 |
| Institution | 428 | 6.9 |
| Other | 14 | 0.2 |

Table 10: Distribution of attack types for the 6,173 comments labelled attacks. An attack type is assigned if a majority of annotators selected it for a given comment. Comments can be assigned multiple attack types.

## B Domain Adaptation Results

| Model | Prec. | Rec. | Macro F1 |
|---|---|---|---|
| mBERT | 61.7 (+4.5) | 61.7 (+2.8) | 74.0 (+2.7) |
| IndoBERT | 68.1 (+4.6) | 68.1 (+4.1) | 77.5 (+3.2) |
| IndoBERTweet | 67.8 (+2.5) | 67.8 (-0.4) | 79.9 (+0.9) |
| XLM-R | 67.0 (+3.7) | 67.0 (-1.2) | 76.9 (+1.3) |
| XLM-T | 66.4 (+0.8) | 66.4 (-1.6) | 77.6 (-0.2) |

Table 11: Domain adaption results for extended pretraining on 100,000 Reddit comments. Change, compared to baselines (Table 2), is in percentage points.

## C Community Context Results

Each comment in our dataset comes with rich metadata, which includes the comment timestamp, anonymised user ID and the source subreddit that the comment was posted to. Different subreddits will have different community guidelines and moderation practices, which can result in different propensities to share online attacks (see Figure 7). We also expect topical and semantic variation across online communities more generally (Del Tredici and Fernández, 2017). Therefore, we hypothesised that this kind of *community context*, as captured by information about the source subreddit of each comment, could be leveraged to improve classification.

To test this hypothesis for each of our baseline models, we take a simple approach using a support vector machine (SVM). For a given comment and a given baseline model, the input features for the SVM are 1) the prediction of the baseline model, and 2) the identity of the subreddit that the comment was posted to, encoded in a one-hot vector. Since the distribution of comments across subreddits in our dataset is heavily skewed (see Appendix A.2), we collapse all subreddits from which there are ten or fewer comments in our dataset into a single category. The SVM is then trained on the same training set and evaluated on the same test set as our baseline models. We use default parameters for the SVM, as given by the `scikit-learn` Python package, and training time is negligible. Results are shown in Table 12.

We find that adding community context as an additional feature using our SVM method does not improve model performance compared to the performance baselines. Performance differences are small, and mostly negative. Our hypothesis is that this negative result can mainly be attributed to the uneven distribution of subreddits in our dataset. Over 90% of labelled comments come the largest two subreddits (see Table 7). These two subreddits also have a similar rate of attacks (39.7% for r/indonesia and 51.1% for r/malaysia), which resembles the average proportion of attacks in the overall dataset (41.2%). As a consequence, the additional community context information will have minimal impact on the classifier's decision boundary. For the less-represented subreddits, on the other hand, the SVM will struggle to establish a better decision boundary than that based on text alone because of data scarcity. And even if the context-aware model did make better predictions on comments from less-represented subreddits, the impact on overall performance would be minimal.

| Model | Baseline | Context | Change |
|---|---|---|---|
| mBERT Base | 71.3 | 71.4 | +0.1 |
| XLM RoBERTa Base | 75.6 | 75.2 | -0.4 |
| Twitter XLM RoBERTa Base | 77.9 | 77.3 | -0.6 |
| IndoBERT Base | 74.3 | 73.3 | -1.0 |
| IndoBERTweet Base | **79.1** | 78.7 | -0.3 |

Table 12: Macro F1 for community context models compared to baselines. Best model performance is highlighted in **bold**, positive changes in blue, and negative changes in red. Change is in percentage points.

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*Limitations section after conclusion*

☑ A2. Did you discuss any potential risks of your work?
*Ethical considerations section after conclusion*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
*1*

☒ A4. Have you used AI writing assistants when working on this paper?
*Left blank.*

### B ☑ Did you use or create scientific artifacts?

*3*

☑ B1. Did you cite the creators of artifacts you used?
*3*

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*3*

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*3*

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*3*

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*3*

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*3*

### C ☑ Did you run computational experiments?

*4*

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*4*

---

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*4*

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*4*

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*4*

**D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?**
*3*

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*In supplementary materials*

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*3*

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*3*

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*internal ethics review*

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*3*