Table 10: Comparison of MACD with existing abusive speech datasets. We compare with **Indic** and **Non-Indic** datasets.

| Dataset | Language | Samples |
|---|---|---|
| | **Non-Indic** | |
| [7] | Spanish | 7K |
| [11] | Italian | 8K |
| [22] | English | 25K |
| [29] | English | 80K |
| [34] | English | 35K |
| [66] | English | 1M |
| [24] | English | 11K |
| [40] | English | 13K |
| [7] | English | 13K |
| [28] | English | 58K |
| [66] | Turkish | 1M |
| [41] | Indonesian | 13K |
| [54] | Polish | 10K |
| [53] | Greek | 1.6M |
| [2] | Arabic | 20K |
| | **Indic** | |
| [48] | Hindi | 6K |
| [47] | Hindi | 5K |
| [42] | Hindi | 4.5K |
| [62] | Hindi | 2K |
| [70] | Hindi | 4.5K |
| [70] | Marathi | 2K |
| [42] | Marathi | 2K |
| [9] | Bengali | 4K |
| [1] | Urdu | 2K |
| [59] | Bengali | 30K |
| | MACD | |
| MACD | Hindi | 33K |
| MACD | Tamil | 30K |
| MACD | Telugu | 30K |
| MACD | Kannada | 33K |
| MACD | Malayalam | 25K |
| MACD | ALL | 150K |

# A Appendix

## A.1 Comparison with existing datasets

In this section, we compare our proposed MACD with existing datasets in detail in Table 10. We note that large scale datasets containing more than 50K samples exist for some **non-Indic** languages like English, Greek and Turkish language. These datasets enable large-scale study of abuse detection for these languages. However, for other languages, presence of large-scale datasets is still lacking. Next, we compare with **Indic** datasets and note that **Indic** datasets are small-scale as compared to **non-Indic** datasets. This shows that there is an immediate requirement for a dataset like MACD to fill this gap and foster advancements in abuse detection in **Indic** languages. Overall and at language level, MACD is one of the largest dataset for studying **Indic** languages.

## A.2 MACD dataset

**Explicit warning**: We want to urge the community to be mindful of the fact that our dataset MACD contains comments which express abusive behaviour towards religion, region, gender etc. that might be abusive and depressing to the researchers. We did not censor such harmful words/phrases because

that would defeat the purpose of the study. Kindly use your discretion while following up on our work.

**Respecting user privacy**: Protecting the privacy of users is a core value for *ShareChat* and we took measures for ensuring that no Personally Identifiable Information (PII) is made public.

**Informed consent**: First, the comments in our dataset MACD are publicly available on the *ShareChat* application. These comments are published by users of the platform for public consumption and informed consent is requested by the platform for broadcasting them.

**Anonymized comments**: To further address privacy concerns, we anonymize our dataset. User mentions and email addresses contain "@" and we use regular expressions to search for matching patterns and replace them with the tag <*email*>. Similarly, phone numbers are also removed using regular expressions and replacement with <*number*>. We repeat the same process for urls in the dataset.

**Anonymized metadata**: We also replace the user and post indentifiers present in MACD with a random string to preventing of back-tracking of content to the users.

**Raw data**: We do not store the raw data used for this study. Only the anonymized data will be made available for future research.

### A.3 Models

In this section, we discuss details of the token-level baselines and contextual models that we used for the experiments in this work.

TF-IDF: As token-level baseline, we extract the TF-IDF features from our dataset. The features are then classified using Linear SVM & logistic regression.

mBERT: mBERT [26] has been trained over more than 104 languages using Masked Language Modelling (MLM) task and follows BERT base architecture [26]. mBERT use word pieces which are shared across the languages for creating the vocabulary.

XLM-R: XLM-R [20] is a multilingual version trained with similar settings as XLM [21] but does not use the Translation Language Modelling (TLM) loss and has been trained over 100 languages using Masked Language Modelling (MLM) loss. XLM-R use byte pair encoding (BPE) which encourages improved sharing between languages.

MuRIL: MuRIL [44] follows similar architecture as BERT base [26] and has been pretrained from scratch over 17 languages (16 Indic and English). Monolingual corpora is used with Masked Language Modelling (MLM) loss while translated and transliterated segment pairs are additionally used using Translation Language Modelling (TLM) loss during pretraining. Translated segment pairs are created using the monolingual sentences and their English translation. Transliterated pairs are constructed using the native script and it's Latin transliteration.

### A.4 Translation experiments

In this section, we explore the performance of models pretrained on Indic languages and finetuned over English abuse detection datasets to understand if abuse detection in Indic languages can be addressed using advancements in English abuse detection and translation algorithms. We finetune the MuRIL model over abuse detection datasets from English language by combining datasets from [22, 29, 49]. We combine the *hate* and *offensive* categories in these datasets for training a binary classification model. We choose MuRIL as it has been trained exclusively over 17 Indic languages and shows competitive performance in our experiments. We then translate the test set of MACD into English using Google translator[10] and evaluate the performance in Table11. These results show that it is challenging to achieve competitive performance by using transfer learning (non-Indic to Indic languages) and translation based approaches. A possible reason is that abusive behaviour is driven by cultural, political and religious beliefs due to which models trained on non-Indic languages and context do not transfer so well for Indic context.

[10] https://translate.google.co.in/

Table 11: **Translation:** Accuracy (Acc) and F1-macro score (F1) using MuRIL fine-tuned over English abuse detection datasets and evaluated on English translated test splits of MACD.

| Hindi | | Tamil | | Telugu | | Kannada | | Malayalam | |
|---|---|---|---|---|---|---|---|---|---|
| Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 |
| 45.31 | 32.57 | 43.7 | 32.19 | 50.88 | 34.92 | 45.13 | 33.04 | 43.40 | 32.49 |

### A.5 Training details

We finetune our models for 15 epochs with a batch size of 8 and learning rate of 2e-5 using Adam optimizer. We truncate the longer comments to 512 tokens and pad the smaller comments with zeros. We use the training set for learning the model parameters and report the results on the test set using the best epoch selected by utilizing the validation set. We use A100 GPUs with 40GB of GPU memory hosted by Google Cloud Platform[11] and funded by ShareChat.

### A.6 Error analysis

In this section, we discuss errors across different languages. As noted previously, we find that XLM-R model fails in cases where specific words (more frequent in abusive comments) are present. For example, इसकी सबको निंदा करनी चाहिए ।हिन्दू मुस्लिम ना करे। *(Condemn this action! Do not make it a matter of Hindu and Islam)* is spoken in good faith to neutralize the conversation but similar tokens are majorly used in abusive comments. Similarly, पुष्पा पुष्पराज झुके गा नहीं साला (bloody pushpa pushraj will not bow down) is a famous movie dialogue which is non-abusive but since the word *bloody* is used in abusive context, the model fails. We also note that certain numbers like 9 have abusive references in different cultures. For example, in Tamil, நீ ஒரு 9 புடவை கட்டின ஆம்பளை 9 பயலே *(You're a man wearing a saree[12], eunuch !)* is calling a man eunuch signified by the use of numerals which the model is not able to detect. Other cases like செம தொடை அழகி *(You have beautiful thighs)* and ಕುಂಡಿ ಚನ್ನಾಗಿ ಬೆಳೆಸಿದ್ದೀರಾ 👌 *(Your buttocks have grown well)* are indirect/implicit sexual references where our model fails to identify abusive behaviour. Also there are some ambiguous cases like അതെന്താ ബനിയന്റെ ഉള്ളിൽ Balloon വല്ലതും ആണോ 😂😂 *(Is there a balloon inside t-shirt ?)* and ആൻറി വേറെ മൂഡിലാ *(Aunty is in different mood)* which can be interpreted in both abusive and non-abusive manner depending on the context. These error cases highlight the complexity of MACD dataset and generate insights for further research.

### A.7 Emoticon and code-mixing statistics

Our dataset contains emoticons as we sourced the comments from a social media platform. We show the percentage (%) of emoticons present in our dataset MACD in Table 12. In future work, we will investigate in detail about the impact of emoticons on abuse detection. Most of the existing large-scale Indic datasets (Table 10) are code-mixed in nature and thus reduce the coverage of Indic language data. We use explicit language-specific character sets to filter out comments thus creating a high-quality Indic dataset without much code-mixing as evident from Table 12.

| **Language** | **%Code-mixing** | **%Emoticons** |
|---|---|---|
| Hindi | 2.23 | 4.32 |
| Kannada | 3.23 | 2.89 |
| Malayalam | 3.86 | 5.22 |
| Tamil | 1.53 | 3.22 |
| Telugu | 2.68 | 3.02 |

Table 12: Code-mixing and emoticon statistics.

[11] https://cloud.google.com/compute/

[12] Traditional Indian dress primarily worn by females.

Table 13: Comparison of MACD with existing code-mixed abusive speech datasets.

| Dataset | Language | Samples | Abuse% |
|---|---|---|---|
| **Various code-mixed Indic datasets** | | | |
| [10] | Hindi-English | 4.5K | 36% |
| [39] | Kannada | 7K | 43% |
| [16] | Tamil | 44K | 28% |
| [16] | Malayalam | 12K | 12% |
| [16] | Kannada | 44K | 44% |
| [15] | Tamil | 15K | 33% |
| [1] | Urdu | 10K | 50% |
| [17] | Tamil-English | 44K | 25% |
| [17] | Kannada-English | 7K | 38% |
| [17] | Malayalam-English | 20K | 15% |
| Kaggle Dataset | 10+ Indic languages | 740K | 42% |
| MACD (only Indic) | | | |
| MACD | Hindi | 33K | 52% |
| MACD | Tamil | 30K | 46% |
| MACD | Telugu | 30K | 52% |
| MACD | Kannada | 33K | 49% |
| MACD | Malayalam | 25K | 45% |

### A.8 Limitations

MACD comprises only five Indic languages which falls short of the 1369 rationalized languages and dialects used by Indian population. Second, our abuse detection model, AbuseXLMR is pretrained over a small subset of one year data which were reported by users in the platform to be potentially abusive in nature. It could be used to generate abusive words/sentences given an initial prompt since it has been pretrained using potentially abusive data (refer Section 4). Thus these models should not be used in-the-wild without presence of any explicit control mechanism (possibly human-in-the-loop).

## B Dataset documentation

We follow the framework proposed by [32].

### B.1 Dataset documentation and intended uses

**For what purpose was the dataset created? Was there a specific task in mind? Was there a 4 specific gap that needed to be filled? Please provide a description.** Abusive content detection is an important problem because abusive behaviour is increasing exponentially on social media platforms. Due to this scale of this problem, manually moderating every interaction is not feasible and thus development of automated abuse detection algorithms is extremely important. However, due to the limited *scale* and *diversity* of abuse detection datasets in Indic languages, development of these models for Indic languages has been severely impeded. This motivates us to release a large scale (150k), linguistically diverse (5 Indic languages) and expertly-annotated dataset for Indic languages. We hope that this dataset would pave the way for large scale abusive content detection for Indic languages.

**Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., 15 company, institution, organization)?** This dataset is created by researchers at ShareChat, India and Indian Institute of Technology, Kharagpur, India.

**Who funded the creation of the dataset?** The creation of dataset was funded by ShareChat, India.

**Any other comments?** N/A

### B.2 Composition

**What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.** The dataset contains the comments and their annotations (abusive/not-abusive) shared by social medial users in response to the multimedia content uploaded on the platform. Along with comments, the masked user identifiers and masked identifier for the multimedia post are also shared.

**How many instances are there in total (of each type, if appropriate)?** `MACD` contains 150K comments from five Indic languages). We further describe the dataset statistics in detail in Section 3.3. **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).** The dataset is sampled from the comments posted on ShareChat, which forms the larger set and contains comments posted in more than 10 Indic languages. The dataset is not a true representation of the larger set because it captures only 5 Indic languages, 150K samples and 70K users which are small in comparison to the whole dataset. For this study, we also sampled a well-balanced dataset, which is not a true representation of larger set because in larger set, the abusive comments are much lesser than normal comments.

**What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.** Each datapoint contains the textual comment posted by social media user and *abusive/not-abusive* label. The masked user identifier and masked identifier to the original multimedia post is also provided.

**Is there a label or target associated with each instance?If so, please provide a description.** Yes each comment has a label - *abusive/not-abusive*

**Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.** Metadata like age/gender of user, identity of user, date of comment etc. are intentionally removed. Moreover, any PII (personally identifiable information) has been removed from the dataset.

**Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.** Comments are also accompanied with masked user and post identifier which form an explicit relationship between them.

**Are there recommended data splits (e.g., training, development/validation,testing)? If so, please provide a description of these splits, explaining the rationale behind them.** The training, test and validation split have been standardized and will be shared with the community along with the dataset for reproducing our results. We randomly split the data into these splits.

**Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.** Since the dataset has been sourced from social media, we observe spelling mistakes, grammatical mistakes etc. We removed the comments which were duplicates of each other, so there is no redundancy.

**Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** MACD is self-contained.

**Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?If so, please provide a description.** No. We completely anonymized the dataset for protecting the privacy of the users.

**Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.** Yes. Owing to the nature of the problem, the dataset contains *abusive* comments which might be offensive in nature.

**Does the dataset relate to people? If not, you may skip the remaining questions in this section.** No. The dataset contains comments posted by people on social media platform. In particular, the dataset does contain any Personally Identifiable Information that could be mapped back to an individual.

### B.3 Collection process

**How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.** The annotation labels associate with comments were acquired via manual annotation using a multilingual annotation team. Each sample was annotated by two annotators and disagreements were resolved by an expert-annotator. We report the annotator agreement in the Section 3.2. The data was directly observable by the annotators.

**What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?** Social media users share their comments on the multimedia posts uploaded by other users on *ShareChat* application. We collect these comments from the platform. Along with the comments, we also procure the metadata (user and post identifier) for MACD.

**If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?** For each language, the dataset was randomly sampled from the larger set and annotated. To create a well balanced dataset, we further sampled abusive and non-abusive comments in balanced manner. For preventing the user-bias, comments from individual users were restricted to 500.

**Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the time-frame in which the data associated with the instances was created.** The comments posted between February, 2021 and September, 2022 were collected for MACD and this timeframe aligns with the data creation.

**Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.** Not applicable.

**Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.** No.

### B.4 Preprocessing/cleaning/labeling

**Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.** PII (personally identifiable information) has been anonymized from the dataset and each comment has been annotated with *abusive/non-abusive* label. Duplicate comments have also been removed.

**Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.** Raw data was not saved to prevent misuse.

**Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.** We have used Python language to implement the models. All the packages can be installed via pip/conda. We will share the package details in our codebase.

### B.5 Uses

**Has the dataset been used for any tasks already? If so, please provide a description** No.

**Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.** No.

**What (other) tasks could the dataset be used for?** The dataset can be used to identify abusive speech in online social media platforms to keep the platforms safe and healthy. There is also a potential for using the dataset to synthetically generate more abusive content; however, such practices should be strictly avoided unless needed for research purposes.

**Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?** Dataset captures only 5 Indic languages. Future user could increase the coverage of `MACD`.

**Are there tasks for which the dataset should not be used? If so, please provide a description.** NA

#### B.5.1 Distribution

**Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?If so, please provide a description.** `MACD` dataset and codebases for reproducing the experiments are available at: `https://github.com/ShareChatAI/MACD`

**How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?** The dataset is available on GitHub repository.

**When will the dataset be distributed?** The dataset will be available for public download after acceptance.

**Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.** The code and dataset is available for only research purposes and any commercial usage is strictly prohibited. The dataset `MACD` and model `AbuseXLMR` is distributed under CC BY-NC-SA license. CC BY-NC-SA allows reusers to distribute, remix, adapt, and build upon the material in any medium or format for noncommercial purposes only,

and only so long as attribution is given to the creator. If you remix, adapt, or build upon the material, you must license the modified material under identical terms.

**Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.** No.

**Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.** No.

### B.5.2 Maintenance

**Who is supporting/hosting/maintaining the dataset?** Authors of this work bear all responsibility in case of violation of rights. Vikram Gupta (vikram.nov.14@gmail.com) and Hastagiri Vanchinathan (hasta@sharechat.co) will be responsible for maintaining this dataset.

**How can the owner/curator/manager of the dataset be contacted (e.g., email address)?** If you wish to extend or contribute to our dataset `MACD`, please contact us via email - Vikram Gupta (vikram.nov.14@gmail.com) and Hastagiri Vanchinathan (hasta@sharechat.co)

**Is there an erratum? If so, please provide a link or other access point.** Any updates to the dataset wiil be shared via GitHub.

**Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list,GitHub)?** If we find inconsistencies in the dataset or extend the dataset, we will release the new version via GitHub.

**If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** N/A

**Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users** All versions of `MACD` will be continue to be supported and maintained on GitHub. We will post the updates on the GitHub repository.

**If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.** Yes. Please contact the authors of this paper for building upon this dataset.

### B.5.3 Responsibility

The authors bear all responsibility in case of violation of rights, etc. We confirm that the dataset is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.