Highlights

Modeling Homophobia and Transphobia detection using Data-Augmentation in a Multilingual Code-mixed Setting

- A overview of Homophobia and Transphobia.
- A novel data augmentation to detect Homophobia and Transphobia in multilingual code mixed setting
- An extensive evaluation on benchmark dataset with very positive results.

Modeling Homophobia and Transphobia detection using Data-Augmentation in a Multilingual Code-mixed Setting

ARTICLE INFO

Keywords:
Homophobia detection
Transphobia detection
Hate Speech Detection

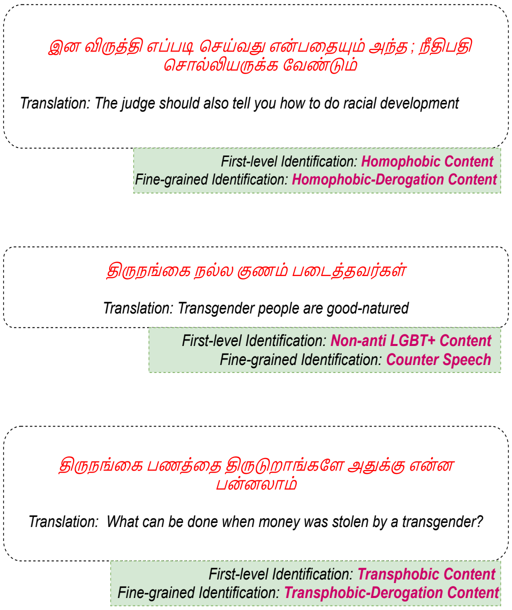
ABSTRACT

There has been a surge of abusive content on social media, which has harmed online users. Homophobia or Transphobia can be defined as the hatred, discomfort, or dislike of lesbian, gay, transgender or bisexual people. Studies have shown that these individuals were more likely to develop mental health issues, likely due to being subjected to more forms of abuse on social media. Hence there is an ardent need to develop automated abusive speech detection systems to tackle the ever-proliferating abusive content on social media. There has been an elevation in hate speech or abuse towards members of the LGBTQIA+ communities on online forums, and this paper focuses on the LGBTQIA+ community. Our work formulates the task of detecting homophobic/transphobic speech in a code-mixed multilingual setting. Due to the shortage of resources in the said study area, we hypothesize that data augmentation via Pseudolabeling by transliterating the code-mixed text to the parent language will improve the models' performances on the newly constructed dataset. We use data from (Chakravarthi et al., 2021) to test out our approach We put our hypothesis into testing, and studied the performances of several multilingual language models for our cause. Additionally, we carried out statistical tests to infer the significance of the performance of the models on both datasets. This is one of the first studies that attempt to model Homophobia and Transphobia across multiple levels of study. We show how data augmentation has increased our models' performances and provide statistical proofs to support our hypothesis.

1. Introduction

The emergence of social media over the past ten years has brought all of the world's regions together in one centre for the facilitation of communication. People can communicate their ideas and opinions on any subject on social media, (**?**) forecasting. This phenomena has made it possible for academics to conduct focused studies on social media users in order to obtain understanding of enduring difficulties including interpreting, monitoring, perceiving, and measuring user behaviour toward topics or occurrences (Hürriyetoğlu et al., 2021). The enormous amount of user-generated data that is made available online every day has also benefited recent advances in big language models and prepared the way for a thorough behavioural analysis of social media users citepwang-etal-2019-topic-aware.

People have shared a wide variety of information and ideas to express their views as a result of the lack of moderation on social media applications. However, the language that was used to communicate the facts can be offensive. It may contain unpleasant and rude language like racism, sexism, homophobia, and transphobia, and even disparage a specific group of individuals or organisations(Zampieri et al., 2020; Sufi, 2022). The growth of social media has created a huge issue that researchers must address since automatic moderation solutions are urgently needed. The terminology used on social media can be extremely hurtful and upsetting to lesbian, gay, bisexual, transgender, queer, and other related communities (LGBTQIA+) due to its subjective character. It may have more significant ramifications for these communities (Uppunda et al., 2021). Because of who they love, or how they appear, or who they identify themselves as, the LGBTQIA+ community is constantly under scrutiny, and is often discriminated against or subjected to abuse (Thurlow, 2001). These are vital aspects of our identities, and should not be used as a means of segregation, as it is against ones' fundamental rights. However, the same does not hold true for several other countries. Being identified as LGBTQIA+ in many countries are in very stressful situations and conditions once they identify their true self.



இன விருத்தி எப்படி செய்வது என்பதையும் அந்த ; நீதிபதி சொல்லியருக்க வேண்டும்

Translation: The judge should also tell you how to do racial development

First-level Identification: **Homophobic Content**
Fine-grained Identification: **Homophobic-Derogation Content**

திருநங்கை நல்ல குணம் படைத்தவர்கள்

Translation: Transgender people are good-natured

First-level Identification: **Non-anti LGBT+ Content**
Fine-grained Identification: **Counter Speech**

திருநங்கை பணத்தை திருடுறாங்களே அதுக்கு என்ன பன்னலாம்

Translation: What can be done when money was stolen by a transgender?

First-level Identification: **Transphobic Content**
Fine-grained Identification: **Transphobic-Derogation Content**

Figure 1: Examples showing the multilingual comments and its English translation for the identification of homophobia and Transphobia at different levels of study,

ORCID(s):

Recent studies on the mental health of people belonging to the LGBTQIA+ communities revealed that these individuals were found to show high rates of mental health concerns, presumably due to being subjected to many forms of online abuse, homophobic attacks, sexual assaults, and disrespect (Wandrekar and Nigudkar, 2020a). Furthermore, the online abuses hurled towards these individuals may lead to systematic bullying campaigns that are organized to spread hatred towards the community (Chintalapudi et al., 2021). In some instances, it results in threats towards people who are vocal towards the better treatment of LGBTQIA+ communities on social media (Mkhize et al., 2020). They also target the individuals of these communities who find solace on the Internet, disrupting their lives. Therefore, there is a need to filter out the homophobic and transphobic comments online.

Homophobia and Transphobia are two concepts that tend to provide unfavourable perceptions towards homosexual and transsexual people. Automation of transphobic and homophobic speech identification on social media can reduce the amount of the hate speech towards these communities and move towards equality, diversity, and inclusion (Chakravarthi et al., 2021; Luo and Mu, 2022). At the same time, much work has been explored on collecting and analysis of data among the domains of hate speech (Tontodimma et al., 2020; Yang et al., 2020), aggression identification (Kumar et al., 2020), misogyny (Guest et al., 2021; Coria et al., 2020), sexism (Singh et al., 2021), and racism (Field et al., 2021), there has been little amounts of work on homophobic and transphobic abuse. The complex nature of the speech on these stigmatized issues provide researchers with an opportunity to analyse how people express their opinions on a sensitive topic in an informal setting on social media. The study provides the social media moderators with a chance to stimulate the amalgamation of several communities to improve an individual's perception about a stigmatized issue (Sawhney et al., 2021).

Despite India recently recognizing all types of queer sexualities, nevertheless, the stigma and taboo surrounding the Indian LGBTQIA+ community is persistent. This can be considered to mirror the scarcity of data and scientific literature focusing on the community, inclusive of both social and medical studies (Wandrekar and Nigudkar, 2020b). Henceforth, there is a clear need to improve the research on these stigmatized issues and develop automated systems that make social media a safe space for the people belonging to these communities. In multilingual countries like India, people's engagement in social media involves a lot of informal texts that do not follow any grammatical rules, and at times tend to be written in code-mixed or non-native scripts. Code-mixing refers to pairing linguistic units from two or more languages into a single conversation (Bali et al., 2014). Our work aims to classify the homophobic and transphobic comments on social media for English, Tamil, and code-mixed Tamil by treating it as a sequence classification task. We analyze and report our findings on a dataset constructed for identifying homophobic and transphobic comments from YouTube, which was constructed to instigate research on mass instances of Homophobia and Transphobia, focusing more on the Indian communities (Chakravarthi et al., 2021). The Tamil (ISO 639-3: tam) language belongs to the family of Dravidian languages predominantly spoken in Tamil Nadu, Sri Lanka, Singapore, Malaysia.

To classify the homophobic and transphobic speech identification, the paper proposes a semi-supervised approach for three languages, Tamil, English, and code-mixed Tamil. The main idea behind the approach is to transliterate the code-mixed Tamil set into Tamil and English using transliteration APIs[1], by training the model with the training set and subsequently generating the pseudo-labels treating the transliterated dataset as the test set. The transliterated dataset is now combined with the former training set to establish a more extensive dataset to overcome these issues' lack of annotated datasets for these issues. After generating the pseudo-labels, we perform the fine-grained analysis of the newly constructed dataset at several levels, as shown in Table 1, Table 2, and Table 3. To the best of our knowledge, this work is the first attempt at modeling Homophobia and Transphobia speech identification at different levels, and among aims to instigate more research into the behavioural analysis of social media on such stigmatized issues.

2. Related Work

Cyberspace enables communication and expression of one's views (Kumar et al., 2021b; Hossain et al., 2022). However, contemporary social media platforms are frequently abused to promote violent messages, remarks, and hate speech. This is referred to as online hate speech, which is defined as any communication that disparages an individual or a group based on their race, color, ethnic origin, gender, sexual orientation, nationality, religion, or political affiliation. Several works have focused on identifying aggression (Kumar et al., 2018), sexism (Parikh et al., 2019), racism (Guest et al., 2021), harassment and violence (Ghosh Chowdhury et al., 2019; Calderwood et al., 2017) in social media by analysing the people's engagement on these issues through posts, videos, and comments. However, there have been relatively fewer amounts of research on identifying homophobic and transphobic speech online, and researchers tend to generalize these works under hate speech or offensive language (Kumar et al., 2021a).

Researchers investigated the linguistic behaviours among homosexual individuals in China by constructing a corpus comprising of the texts from these individuals (Wu and Hsieh, 2017). Emotion lexicons were constructed to differentiate between the socially acceptable and unacceptable discourse on sensitive issues of migrants and LGBTQ+ in the languages of English, Croatian, Dutch, and Slovene (Ljubešić et al., 2020). Researchers collected 1,784 comments from posts on Facebook that involved anything about LGBTQ+ in the state of Mato Grosso do Sul in Brazil

[1] https://pypi.org/project/ai4bharat-transliteration/

and classified 477 comments as hate speech with prevalence towards 'repulsion for the existence or repudiation of LGBT people's attitudes' and 'discrediting journalistic information' (Silva and Silva, 2021). A manually annotated corpus comprising Homophobia and Transphobia speech in a multilingual setting was scraped from YouTube. The dataset consists of 22,824 comments in three languages, with 7,265 in English, 5,240 in Tamil, and 10,319 in code-mixed Tamil-English and aims to classify the homophobic and transphobic speech several different levels (Chakravarthi et al., 2021). While there are several other studies on the said issue, most of them are psychological studies focusing on homophobic bullying (Bacchini et al., 2021), verbal victimization (Elipe et al., 2021), and impacts on mental health (Ventriglio et al., 2021) among other studies.

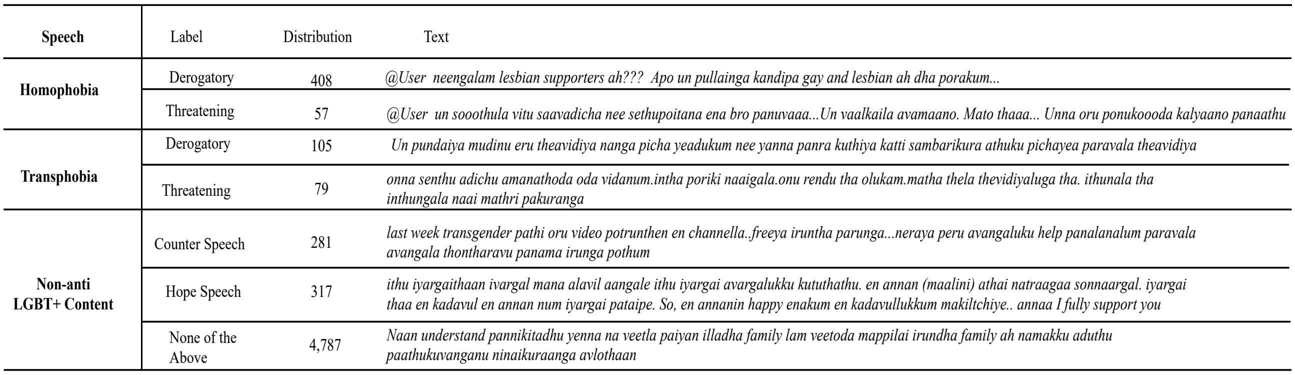
According to European Union Commission directives, hate speech should be criminalized (Fortuna and Nunes, 2018). The European Union has directed social media platforms to enhance their automated hate speech detection methods to ensure that no unacceptable information remains online for more than 24 hours. However, homophobic or transphobic comments in social media is not detected properly and India like developing countries need to start a plan to protect their vulnerable individuals. These factors combine to make the study of online homophobia/transphobia detection a critical field of research. Indeed, there are several theoretical gaps regarding the explanation of this behavior, and there is an insufficient amount of empirical evidence to comprehend this phenomena and its relationship to Internet use. On the basis of this requirement, the primary contribution of this article is to definition of Homophobia and Transphobia based on various publications.

When it comes to detecting hate speech or offensive language in a code-mixed setting, researchers have begun scraping code-mixed data from social media applications owing to the surge in social media users in recent years (Lashkarashvili and Tsintsadze, 2022; Hande et al., 2020; Mahdikhani, 2022). However, considering polyglottic countries such as India, where there are several languages spoken around, there is a lack of relevant data in low-resourced languages Roy et al. (2022, 2020). To overcome the lack of data, Hande et al. (2022) explored the significance of Multi-task Learning for closely related downstream NLP tasks of Offensive Language Identification and Sentiment Analysis in code-mixed languages of Kannada, Malayalam, and Tamil (Neogi et al., 2021; Mohamed Ridhwan and Hargreaves, 2021). However, the study is limited to closely related auxiliary tasks. Biradar et al. (2022) used translational systems to convert all texts to a single language (English) and then fine-tuned the language model for hate speech classification in Hinglish (Hindi - English). However, the translation fails to capture context and sarcasm into account, and this would result in incorrect predictions. One of the main challenging aspects of dealing with code-mixed data is with pretrained language models failing to capture the contextual between the two languages (Yasaswini et al., 2021).

3. Dataset

The dataset is classified at three levels, namely, first level, second-level, and third-level identification. Fig 2 and Fig 1 show some examples and the distribution of data across several levels respectively. The class-wise distribution of the dataset at three different levels are tabulated in Table 1, Table 2, and Table 3. The dataset used Kripendorff's inter annotator agreement and had a score of 0.67, 0.76, and 0.54 for English, Tamil and Tamil-English respectively. The three levels of classification include:

- **Homophobic Content:** Any comment that is deemed gender-based and involves the use of pejorative labels (e.g, “fag", "homo") or defamatory phrases. All comments of lesbophobia, gayphobia and biphobia belong to this class. They can be further classified into two types:
  - **Homophobic Derogation:** They are the phrases used to belittle, convey a low opinion or a lacking of respect towards the LGBIQIA+ community.
  - **Homophobic Threatening language:** Any phrase with an intent/desire to harm/hurt LGBTQIA+, or encourage, stimulate, or advocate such violence.
- **Transphobic Content:** Any comment that involves the use of derogatory labels (e.g, "cross-dresser", "not-manly enough") or any other phrases that involve offensive language about the people belonging to the transgender community. They can be further classified into:
  - **Transphobic derogation:** Any phrase that is used to disparage vulnerable transgender people and are usually very offensive.
  - **Transphobic Threatening Langauge:** Any phrase with an intent/desire to harm/hurt or encourage, stimulate, or advocate such violence towards the transgender people.
- **Non-anti-LGBTQ+ Content:** They can be classified into three groups as follows:
  - **Counter Speech:** It can be defined as a non-aggressive reaction that provides feedback through fact-based arguments (e.g., "That's terribly homophobic" or "what you said in unacceptable").
  - **Hope Speech:**Hope speech incites optimism and resilience that positively influences the readers.
  - **None of the Categories:** The comment does not include any homophobic or transphobic insults, derogatory speech towards LGBTQ+ communities, counter-speech or hope speech. Nevertheless, any form of offensive language that does not belong to the other categories can be included.



Speech	Label	Distribution	Text
Homophobia	Derogatory	408	*@User neengalam lesbian supporters ah??? Apo un pullainga kandipa gay and lesbian ah dha porakum...*
	Threatening	57	*@User un sooothula vitu saavadicha nee sethupoitana ena bro panuvaaa...Un vaalkaila avamaano. Mato thaaa... Unna oru ponukoooda kalyaano panaathu*
Transphobia	Derogatory	105	*Un pundaiya mudinu eru theavidiya nanga picha yeadukum nee yanna panra kuthiya katti sambarikura athuku pichayea paravala theavidiya*
	Threatening	79	*onna senthu adichu amanathoda oda vidanum.intha poriki naaigala.onu rendu tha olukam.matha thela thevidiyaluga tha. ithunala tha inthungala naai mathri pakuranga*
Non-anti LGBT+ Content	Counter Speech	281	*last week transgender pathi oru video potrunthen en channella..freeya iruntha parunga...neraya peru avangaluku help panalanalum paravala avangala thontharavu panama irunga pothum*
	Hope Speech	317	*ithu iyargaithaan ivargal mana alavil aangale ithu iyargai avargalukku kututhathu. en annan (maalini) athai natraagaa sonnaargal. iyargai thaa en kadavul en annan num iyargai pataipe. So, en annanin happy enakum en kadavullukkum makiltchiye.. annaa I fully support you*
	None of the Above	4,787	*Naan understand pannikitadhu yenna na veetla paiyan illadha family lam veetoda mappilai irundha family ah namakku aduthu paathukuvanganu ninaikuraanga avlothaan*

Figure 2: Distribution of labels and examples for all levels in the Homophobia and Transphobia dataset. The names of the users have been censored, and texts about it have been rephrased. We want to caution the readers that examples in this paper, though censored for profanity, might contain offensive and hate-driven language.

- **First Level(Content):**
  - Homophobic Content:
  - Transphobic Content:
  - Non-anti-LGBT+ Content:
- **Second Level:**
  - Homophobic Content
  - Transphobic Content
  - Counter Speech
  - Hope Speech
  - None of the Above
- **Third Level:**
  - Homophobic-Derogatory
  - Homophobic-Threatening
  - Transphobic-Derogatory
  - Transphobic-Threatening
  - Counter Speech
  - Hope Speech
  - None of the Above

Class	HT Dataset	DA-HT Dataset
Homophobic	465	619
Transphobic	184	246
Non-anti-LGBT+	5,385	11,293
Total	6,034	12, 068

Table 1
Dataset distribution at the first level. Eng: English, Tam: Tamil, Tam-Eng: Code-mixed Tamil-English

Class	HT Dataset	DA-HT Dataset
Homophobic	465	1251
Transphobic	184	185
Counter speech	281	317
Hope speech	317	365
NOTA	4,787	9,950
Total	6,034	12,068

Table 2
Dataset distribution at the second level. NOTA: None of the above

4. Methodology

4.1. Data Augmentation via Pseudo-labeling

As observed from Table 1, it is evident that there is a substantial lack of resources for Homophobia and Transphobia detection. To overcome the absence of the resources, we propose a strategy of data augmentation using the available code-mixed texts but with its transliterated texts to the native language using an automatic transliteration tool[2] from (Kakwani et al., 2021), inspired by its effectiveness in low-resource ASR (Khare et al., 2021). Transliteration can be defined as the process of writing and describing words or letters by using letters of a different alphabet or language (Regmi et al., 2010), and the plausibility of sentences having a secondary meaning. We use three language models to generate the pseudo-labels for the transliterated dataset, which is established on the effectiveness of transformer-based models in NLP. We train the language models on the code-mixed dataset and generate pseudo-labels on the best performing model. For the sake of relevance in the upcoming sections, we now refer the Data Augmented via Pseudolabeling Homophobia & Transophobic dataset as **DA-HT** and the former as **HT** dataset. We choose MuRIL (Khanuja et al., 2021), mBERT (Pires et al., 2019), and XLM-R (Conneau et al., 2020). MuRIL is a multilingual language model specifically built for Indian languages, supporting 16 Indian languages and English. Unlike mBERT, which is trained on the Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) objectives, MuRIL is pretrained on MLM and Translational Language Modeling, with the TLM objective utilizing both translated and transliterated pairs. mBERT is the multilingual version of BERT (Devlin et al., 2019), pretrained on 104 languages over texts

[2] https://pypi.org/project/ai4bharat-transliteration/

Class	HT Dataset	DA-HT Dataset
H-Derogation	408	410
H-Threatening	57	59
T-Derogation	105	107
T-Threatening	79	81
Counter speech	281	317
Hope speech	317	320
NOTA	4,787	10,774

Table 3
Dataset distribution at the third level. H-Derogation: Homophobic-Derogation, H-Threatening: Homophobic Threatening, T-Derogation: Transphobic-Derogation, T-Threatening: Transphobic Threatening

crawled from Wikipedia dumps. XLM-R is pretrained over two terabytes of CommonCrawl data spanning over 100 languages to boost the performances in cross-lingual tasks. All these models use transformers as the building blocks of their architecture. The vital attribute is self-attention at token-level (K, Q, and V) (Vaswani et al., 2017), enabling them to generate contextualized word embeddings. For any comment c_i from the dataset, let $\{m_1, m_2, ..., m_n\}$ represent the tokens after being tokenized by the tokenizer. We extract the word embeddings from the final layer of the models and obtain the embeddings E_i, E_j, and E_k for MuRIL, mBERT, and XLM-R, respectively.

$$E_i = MuRIL(c_i)$$
$$E_j = mBERT(c_i)$$
$$E_k = XLMR(c_i)$$

Let $C = c_1, c_2, .., c_n$ *where* $C_i \in (Ta - En)$ *be the sentences in the dataset* These word embeddings are passed through a stacked BiLSTM decoder. We use dropout as means of regularization and then a linear output layer to get the output O. We train the dataset for the three models.

$$h_t = BiLSTM(c_t, h_{t-1}) \tag{1}$$

$$h_t = BiLSTM(c_t, h_{t+1}) \tag{2}$$

To generate the pseudo-labels, we treat the task of identifying at multiple levels as an MTL task, by treating the second level categorization as an auxiliary task to the primary level identification. We employ a complex parameter sharing model to generate the pseudo-labels on the dataset. We employ a familiar encoder that is shared and updated by the three tasks, primary level, and the secondary classification. We treat the tasks equal and, subsequently, the losses too.

$$L = \frac{1}{j} \sum_{i=1}^{j} \mathcal{L}_{\geqq}(P, y) \tag{3}$$

Where, $\mathcal{L}(P, y)$ refers to the loss class balanced focal loss. Based on the embedding E_is' performance on the validation set, we generate the pseudo-labels on the transliterated set by treating it as the test set. Our experiments concluded that $E_k = XLMR(c_i)$ performed the best of all.

Let $C = \{c_1, c_2, .., c_n\}$ where $C_i \in (Ta - En)$ be the sentences in the Homophobia and Transphobia dataset and $T = \{t_1, t_2, ..., t_n\}$ where $T_i \in (Ta)$, we define *Data Augmented Homophobia and Transphobia dataset*, $DA - HT = \{c_1, c_2, .., c_n, t_1, t_2, .., t_n\}$, which is the amalgamation of the transilterated and the code-mixed dataset. We remove any plausible common sentences from the dataset. Our main intention to perform pseudo-labeling is to overcome the low resources on Homophobia and Transphobia speech available.

4.2. Homophobia and Transphobia Detection

We treat the task of identifying types of speech at different levels and categorize them related to speech as - *Homophobia*, *Transphobia*, and *Non-anti LGBT+ Content* independently. We present our findings on the two datasets, namely, the code-mixed dataset and the combined dataset. We extract the contextualized word embeddings from the dataset and feed them as input to a BiLSTM decoder (Wadud et al., 2022). Unlike the approach undertaken for pseudo-labeling, we treat the two levels of study independently. We use a softmax activation function for both the tasks for the final output layer.

Optimization : To address the severe case of class imbalance present among the labels, we employ *Class-balanced focal loss* to optimize the model (Cui et al., 2019), as shown in Equation 4. We consider the weights of each class and pass the weights as a tensor to the parameter while computing the loss. We define a weighted Class-balanced focal loss as:

$$CB(P, y) = \frac{1 - \beta}{1 - \beta^{n_y}} \mathcal{L}(P, y) \tag{4}$$

where $\mathcal{L}(P, y)$ is the α balanced variation of focal loss (Lin et al., 2017), formulated as

$$\mathcal{L}(P, y) = -\alpha_t (1 - P_t)^{\gamma} \log P_t. \tag{5}$$

Where N is the number of training samples, y and $P(y)$ denote target and predicted labels respectively.

4.3. Experiments Setup

Preprocessing: Due to the nature of the dataset, we make the usernames and any other personally identifiable information anonymous by (a) replacing that information with *User* and (b) normalize the *URLs*, and finally (c) translating the emoticon into the text to capture the intention of the texts. We use the Huggingface's *AutoNLP* (Wolf et al., 2020) library to import the tokenizers and tokenize the texts.

Hyperparameters: We train our models using the huggingface Transformers library using the Pytorch backend for implementation. We fine-tune four pretrained language models, namely, multilingualBERT (Pires et al., 2019), XLM-RoBERTa (Conneau et al., 2020), MuRIL (Khanuja et al., 2021), Indic-BERT (Kakwani et al., 2021). All models follow the architecture of BERT. We extract a 768-dimensional token-level embedding from these pretrained language models. We employ *randomized search* to find the

First-level														
	HT Dataset							DA-HT Dataset						
Embeddings	P_w	R_w	$F1_w$	P_m	R_m	$F1_m$	Acc	P_w	R_w	$F1_w$	P_m	R_m	$F1_m$	Acc
mBERT	0.922	0.932	0.928	0.680	0.670	0.664	0.932	**0.936**	**0.946**	**0.940**	**0.724**	**0.640**	**0.666**	**0.946**
MuRIL	0.800	0.890	0.840	0.300	0.330	0.331	0.890	0.896	0.936	0.912	0.440	0.420	0.416	0.934
IndicBERT	0.828	0.896	0.854	0.388	0.360	0.358	0.896	0.908	0.938	0.916	0.650	0.492	0.528	0.938
XLM-R	0.810	0.898	0.850	0.310	0.342	0.322	0.888	0.858	0.928	0.890	0.316	0.334	0.320	0.926
Second-level														
mBERT	0.764	0.834	0.796	0.316	0.270	0.258	0.808	**0.930**	**0.932**	**0.902**	**0.726**	**0.666**	**0.724**	**0.932**
MuRIL	0.630	0.790	0.700	0.160	0.200	0.180	0.790	0.780	0.848	0.796	0.256	0.282	0.264	0.858
IndicBERT	0.258	0.712	0.808	0.316	0.270	0.258	0.808	0.852	0.884	0.858	0.556	0.362	0.386	0.884
XLM-R	0.660	0.798	0.716	0.228	0.230	0.218	0.798	0.814	0.866	0.782	0.356	0.314	0.324	0.866
Third-level														
mBERT	0.748	0.822	0.784	0.262	0.284	0.250	0.822	**0.888**	**0.912**	**0.898**	**0.344**	**0.366**	**0.318**	**0.912**
MuRIL	0.642	0.794	0.722	0.136	0.158	0.146	0.794	0.802	0.888	0.838	0.128	0.138	0.128	0.886
IndicBERT	0.716	0.814	0.760	0.222	0.236	0.220	0.814	0.806	0.890	0.842	0.148	0.144	0.138	0.890
XLM-R	0.704	0.800	0.732	0.192	0.194	0.178	0.800	0.804	0.890	0.842	0.158	0.152	0.148	0.892

Table 4
Classification reports of the language models on both datasets. P_w: Weighted-average of Precision, P_m: Macro-average of Precision. R_w, R_m, $F1_w$, $F1_m$ follow similarly for Recall and F1-score respectively. Acc: Accuracy.

First-level								
	HT Dataset				DA-HT Dataset			
Embeddings	P	R	$F1$	Acc	P	R	$F1$	Acc
IndicBERT	0.022	0.051	0.002	0.003	0.029	0.049	0.021	0.048
MuRIL	0.001	0.038	0.001	0.001	0.028	0.045	0.031	0.055
XLM-R	0.001	0.039	0.001	0.001	0.001	0.019	0.001	0.011
Second-level								
IndicBERT	0.424	0.023	0.049	0.028	0.001	0.004	0.115	0.001
MuRIL	0.001	0.007	0.001	0.001	0.003	0.002	0.054	0.002
XLM-R	0.143	0.007	0.025	0.001	0.003	0.004	0.034	0.004
Third-level								
MuRIL	0.012	0.013	0.024	0.012	0.037	0.014	0.091	0.022
IndicBERT	0.141	0.168	0.136	0.169	0.018	0.019	0.011	0.0194
XLM-R	0.165	0.021	0.014	0.021	0.001	0.013	0.069	0.055

Table 5
Values of T-Test when the models are compared with mBERT (Best performing model).

optimal setting of the hyperparameters. They are summarized as follows: BiLSTM layers size $\in \{128, 256, 512\}$, α-based class-balanced focal loss: $\{\alpha = 0.25\}$. dropout, $\delta \in \{0.1, 0.2, 0.3, 0.4\}$, weight decay $\omega \in \{10^{-5}, 10^{-6}\}$, optimizer $\{AdamW\}$, batch size $b \in \{16, 32, 64, 128\}$. We set the learning rate to $2e-5$. We set the maximum length of input text to be tokenized at a time, $Max\_len \in \{64, 128\}$.

For Data augmentation via pseudo-labeling as discussed in Section 4.1, we employ multitask learning under a similar experimental setup as above. However, we treat the first-level task as the *primary task*, and the other detection levels as the *additional tasks*. The main idea behind this is that the *auxiliary tasks* assist in improving the overall performance of the *primary task*. We use *Adam* optimizer with *weight decay* as the optimizer, and we give equal priority to speech detection at all the three levels.

Training: All experiments were carried out using stratified 5-fold cross-validation. Additionally, we conduct a student's T-Test on the outputs obtained on the five folds during training, to examine the statistical significance of the improvement. The main intention behind using statistical tests is to examine if there is any significant increase in the performance primarily due to the algorithm at hand, or whether it was based on randomization done while setting the folds during training.

T-test (One tail, Type I): For the population means of two sets of measurements, this test examines whether the population means differ from each other, while assuming that both samples are a derivative of normal distribution (Dror et al., 2018). Our motive to add a T-Test is to address the extreme class imbalance in the dataset, and assess whether the precedence of a language model can be backed by statistical significance testing, and the results are not merely coincidental. Practically, we apply weighted F1-score as the evaluation metric, to compare the correct predictions over an input sample. The T-test is based on the Central Limit Theorem (CLT), which establishes that the normalized sum when the independent variables are added tend towards a

Language Model	Increase/Decrease
First Level	
mBERT	0.946(+**0.014**)
MuRIL	0.934(+**0.044**)
IndicBERT	0.938(+**0.042**)
XLM-R	0.926(+**0.038**)
Second Level	
mBERT	0.932(+**0.124**)
MuRIL	0.858(+**0.068**)
IndicBERT	0.884(+**0.076**)
XLM-R	0.866(+**0.068**)
Third Level	
mBERT	0.912(+**0.090**)
MuRIL	0.886(+**0.092**)
IndicBERT	0.890(+**0.076**)
XLM-R	0.892(+**0.092**)

Table 6
Comparisons of the models' accuracies on the two datasets. Δ(Acc) = Acc(DA-HT) - Acc (HT) w.r.t to Acc(HT)

normal distribution despite the original variables not being normally distributed. We perform One tail, Type 1 T-Test with $\alpha = 0.5$. It is formulated as:

$$T = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{S^2(\frac{1}{n_1} + \frac{1}{n_2})}} \tag{6}$$

Where $\bar{x_1} \& \bar{x_2}$ are the population means of two different sets, namely, the best model and the model that is being compared, while n_1 and n_2 are the numbers of observations in the group, and S is the standard error between the groups. We initially confirmed if the two population means are normally distributed.

5. Results and Discussion

The primary intention behind this paper is to evaluate the performance of the models by studying and comparing the effectiveness of the different methodologies towards the research community. As per our earlier statements, we observe that mBERT is one of the best performing models, and hence we conduct $T - Test$ of all other models relative to it.we conduct T-Test to check the difference in the predictions between the data augemented dataset and the HT dataset. We examine if the language models, when fine-tuned on the code-mixed dataset and the augmented dataset, have a difference in the performance, if in a good way. The tarefcr comprises of the classification reports of different levels of Homophobia and Transphobia detection, and observe the changes, if any, in the performances of the models on the newly constructed dataset. From Table 4, we observe that mBERT achieves the best performance on the three levels of studies on Homophobia and Transphobia. As stated earlier, we perform statistical analysis on all folds, to make sure that the models did not achieve better performance based on randomization during 5-fold cross validation. The results of the *T-Test* are tabulated in Table 5. The lower the value of when computing the *T-test* between two sets of populations (performances on the evaluation metrics of other language models when compared to the best performing model, i.e, mBERT), the closer the two populations are, in terms of the performance. XLM-RoBERTa uses the *RoBERTaTokenizer*, which uses a byte-level BPE tokenizer for tokenizing and encoding the tokenized sentences. As Tamil is a morphologically rich language. BPE tends to have a poor morphological alignment with the original text (Jain et al., 2020). *DA-HT* dataset is also heavily reliant on the quality of transliterations produced by the APIs.

We observe that there is a substantial increase in the performance of the same language models when fine-tuned on the *DA-HT Dataset* (Data-Augmented Homophobia and Transliterated Dataset), on a set of evaluation metrics tabulated in Table 4. We can observe from Table 6 that the increase in performance commensurates with the increased level of study. The statistical tests carried out on the models' performances are tabulated in Table 5, confirming our hypothesis. However, due to the extreme class imbalance, we observe a low macro average scores among all models, especially, XLM-RoBERTa, and MuRIL. Due to the additional Tamil data, XLM-RoBERTa tends to predict only two classes in the third-level of study. Similarly, $MuRIL$ can not generalize well on code-mixed sentences, and correctly predicts on the transliterated texts, but not on the code-mixed sentences. *IndicBERT* performs better than these two language models, owing to the data it has been pretrained on. Unlike mBERT, and XLM-R, IndicBERT, an ALBERT-based model was pretrained on 16 Indic languages, and English. *MuRIL* too, is pretrained around Indic languages. However, the model did not perform well on the dataset as it is expected to do so. Therefore, a limitation of carrying out Data augmentation via Pseudolabeling is the inability to address the mixture of monolingual and code-mixed text to influence effective cross-lingual transfer between the two languages and create negative shared representations, which eventually, hinders the performance.

6. Conclusion

The increasing amount of abusive speech towards the LGBTQIA+ community and relatively fewer approaches dedicated to detecting the content motivated us to develop approaches aimed towards improved social life of the people belonging to these communities. Due to the dearth of studies and data available at the time, we propose a Data Augmentation via pseudolabeling on the transliterated texts. Our work is specialized towards detecting the types of abusive speech in a multilingual code-mixed setting. We observe a substantial increase in the models' performance on all evaluation metrics when fine-tuned on the augmented dataset. We carry out statistical significance tests such as *T-Test* to validate the increase in the models' performances. We intend to optimize the loss functions for future work and develop a multilingual Homophobia and Transphobia detection system for several languages.

References

Bacchini, D., Esposito, C., Affuso, G., Amodeo, A., 2021. The impact of personal values, gender stereotypes, and school climate on homophobic bullying: a multilevel analysis. Sexuality Research and Social Policy 18. doi:10.1007/s13178-020-00484-4.

Bali, K., Sharma, J., Choudhury, M., Vyas, Y., 2014. "I am borrowing ya mixing ?" an analysis of English-Hindi code mixing in Facebook, in: Proceedings of the First Workshop on Computational Approaches to Code Switching, Association for Computational Linguistics, Doha, Qatar. pp. 116–126. URL: https://aclanthology.org/W14-3914, doi:10.3115/v1/W14-3914.

Biradar, S., Saumya, S., et al., 2022. Fighting hate speech from bilingual hinglish speaker's perspective, a transformer-and translation-based approach. Social Network Analysis and Mining 12, 1–10.

Calderwood, A., Pruett, E.A., Ptucha, R., Homan, C., Ovesdotter Alm, C., 2017. Understanding the semantics of narratives of interpersonal violence through reader annotations and physiological reactions, in: Proceedings of the Workshop Computational Semantics Beyond Events and Roles, Association for Computational Linguistics, Valencia, Spain. pp. 1–9. URL: https://aclanthology.org/W17-1801, doi:10.18653/v1/W17-1801.

Chakravarthi, B.R., Priyadharshini, R., Ponnusamy, R., Kumaresan, P.K., Sampath, K., Thenmozhi, D., Thangasamy, S., Nallathambi, R., McCrae, J.P., 2021. Dataset for identification of homophobia and transophobia in multilingual youtube comments. arXiv:2109.00227.

Chintalapudi, N., Battineni, G., Canio, M.D., Sagaro, G.G., Amenta, F., 2021. Text mining with sentiment analysis on seafarers' medical documents. International Journal of Information Management Data Insights 1, 100005. URL: https://www.sciencedirect.com/science/article/pii/S2667096820300057, doi:https://doi.org/10.1016/j.jjimei.2020.100005.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V., 2020. Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online. pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747, doi:10.18653/v1/2020.acl-main.747.

Coria, J.M., Ghannay, S., Rosset, S., Bredin, H., 2020. A metric learning approach to misogyny categorization, in: Proceedings of the 5th Workshop on Representation Learning for NLP, Association for Computational Linguistics, Online. pp. 89–94. URL: https://aclanthology.org/2020.repl4nlp-1.12, doi:10.18653/v1/2020.repl4nlp-1.12.

Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S., 2019. Class-balanced loss based on effective number of samples, pp. 9260–9269. doi:10.1109/CVPR.2019.00949.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 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), Association for Computational Linguistics, Minneapolis, Minnesota. pp. 4171–4186. URL: https://aclanthology.org/N19-1423, doi:10.18653/v1/N19-1423.

Dror, R., Baumer, G., Shlomov, S., Reichart, R., 2018. The hitchhiker's guide to testing statistical significance in natural language processing, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia. pp. 1383–1392. URL: https://aclanthology.org/P18-1128, doi:10.18653/v1/P18-1128.

Elipe, P., Espelage, D., Del Rey, R., 2021. Homophobic verbal and bullying victimization: Overlap and emotional impact. Sexuality Research and Social Policy , 1–12doi:10.1007/s13178-021-00613-7.

Field, A., Blodgett, S.L., Waseem, Z., Tsvetkov, Y., 2021. A survey of race, racism, and anti-racism in NLP, 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), Association for Computational Linguistics, Online. pp. 1905–1925. URL: https://aclanthology.org/2021.acl-long.149, doi:10.18653/v1/2021.acl-long.149.

Fortuna, P., Nunes, S., 2018. A survey on automatic detection of hate speech in text. ACM Comput. Surv. 51. URL: https://doi.org/10.1145/3232676, doi:10.1145/3232676.

Ghosh Chowdhury, A., Sawhney, R., Shah, R.R., Mahata, D., 2019. #YouToo? detection of personal recollections of sexual harassment on social media, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy. pp. 2527–2537. URL: https://aclanthology.org/P19-1241, doi:10.18653/v1/P19-1241.

Guest, E., Vidgen, B., Mittos, A., Sastry, N., Tyson, G., Margetts, H., 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, Association for Computational Linguistics, Online. pp. 1336–1350. URL: https://aclanthology.org/2021.eacl-main.114.

Hande, A., Hegde, S.U., Chakravarthi, B.R., 2022. Multi-task learning in under-resourced dravidian languages. Journal of Data, Information and Management , 1–29.

Hande, A., Priyadharshini, R., Chakravarthi, B.R., 2020. Kancmd: Kannada codemixed dataset for sentiment analysis and offensive language detection, in: Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media, pp. 54–63.

Hossain, M.A., Quaddus, M., Warren, M., Akter, S., Pappas, I., 2022. Are you a cyberbully on social media? exploring the personality traits using a fuzzy-set configurational approach. International Journal of Information Management 66, 102537. URL: https://www.sciencedirect.com/science/article/pii/S0268401222000718, doi:https://doi.org/10.1016/j.ijinfomgt.2022.102537.

Hürriyetoğlu, A., Tanev, H., Zavarella, V., Piskorski, J., Yeniterzi, R., Yuret, D., Villavicencio, A., 2021. Challenges and applications of automated extraction of socio-political events from text (CASE 2021): Workshop and shared task report, in: Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021), Association for Computational Linguistics, Online. pp. 1–9. URL: https://aclanthology.org/2021.case-1.1, doi:10.18653/v1/2021.case-1.1.

Jain, K., Deshpande, A., Shridhar, K., Laumann, F., Dash, A., 2020. Indic-transformers: An analysis of transformer language models for indian languages. CoRR abs/2011.02323. URL: https://arxiv.org/abs/2011.02323, arXiv:2011.02323.

Kakwani, D., Kunchukuttan, A., Golla, S., N.C., G., Bhattacharyya, A., Khapra, M.M., Kumar, P., 2021. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online. pp. 4948–4961. URL: https://aclanthology.org/2020.findings-emnlp.445, doi:10.18653/v1/2020.findings-emnlp.445.

Khanuja, S., Bansal, D., Mehtani, S., Khosla, S., Dey, A., Gopalan, B., Margam, D.K., Aggarwal, P., Nagipogu, R.T., Dave, S., Gupta, S., Gali, S.C.B., Subramanian, V., Talukdar, P., 2021. Muril: Multilingual representations for indian languages. arXiv:2103.10730.

Khare, S., Mittal, A., Diwan, A., Sarawagi, S., Jyothi, P., Bharadwaj, S., 2021. Low resource asr: The surprising effectiveness of high resource transliteration. Proc. Interspeech 2021 , 1529–1533.

Kumar, G., Singh, J.P., Kumar, A., 2021a. A deep multi-modal neural network for the identification of hate speech from social media, in: Dennehy, D., Griva, A., Pouloudi, N., Dwivedi, Y.K., Pappas, I., Mäntymäki, M. (Eds.), Responsible AI and Analytics for an Ethical and Inclusive Digitized Society, Springer International Publishing, Cham. pp. 670–680.

Kumar, R., Ojha, A.K., Malmasi, S., Zampieri, M., 2020. Evaluating aggression identification in social media, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France. pp. 1–5. URL: https://aclanthology.org/2020.trac-1.1.

Kumar, R., Reganti, A.N., Bhatia, A., Maheshwari, T., 2018. Aggression-annotated corpus of Hindi-English code-mixed data, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan. URL: https://aclanthology.org/L18-1226.

Kumar, S., Kar, A.K., Ilavarasan, P.V., 2021b. Applications of text mining in services management: A systematic literature review. International Journal of Information Management Data Insights 1, 100008. URL: https://www.sciencedirect.com/science/article/pii/S266709682100001X, doi:https://doi.org/10.1016/j.jjimei.2021.100008.

Lashkarashvili, N., Tsintsadze, M., 2022. Toxicity detection in online georgian discussions. International Journal of Information Management Data Insights 2, 100062. URL: https://www.sciencedirect.com/science/article/pii/S2667096822000064, doi:https://doi.org/10.1016/j.jjimei.2022.100062.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P., 2017. Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV).

Ljubešić, N., Markov, I., Fišer, D., Daelemans, W., 2020. The LiLaH emotion lexicon of Croatian, Dutch and Slovene, in: Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media, Association for Computational Linguistics, Barcelona, Spain (Online). pp. 153–157. URL: https://aclanthology.org/2020.peoples-1.15.

Luo, M., Mu, X., 2022. Entity sentiment analysis in the news: A case study based on negative sentiment smoothing model (nssm). International Journal of Information Management Data Insights 2, 100060. URL: https://www.sciencedirect.com/science/article/pii/S2667096822000040, doi:https://doi.org/10.1016/j.jjimei.2022.100060.

Mahdikhani, M., 2022. Predicting the popularity of tweets by analyzing public opinion and emotions in different stages of covid-19 pandemic. International Journal of Information Management Data Insights 2, 100053. URL: https://www.sciencedirect.com/science/article/pii/S266709682100046X, doi:https://doi.org/10.1016/j.jjimei.2021.100053.

Mkhize, S., Nunlall, R., Gopal, N., 2020. An examination of social media as a platform for cyber-violence against the lgbt+ population. Agenda 34, 1–11. doi:10.1080/10130950.2019.1704485.

Mohamed Ridhwan, K., Hargreaves, C.A., 2021. Leveraging twitter data to understand public sentiment for the covid-19 outbreak in singapore. International Journal of Information Management Data Insights 1, 100021. URL: https://www.sciencedirect.com/science/article/pii/S2667096821000148, doi:https://doi.org/10.1016/j.jjimei.2021.100021.

Neogi, A.S., Garg, K.A., Mishra, R.K., Dwivedi, Y.K., 2021. Sentiment analysis and classification of indian farmers' protest using twitter data. International Journal of Information Management Data Insights 1, 100019. URL: https://www.sciencedirect.com/science/article/pii/S2667096821000124, doi:https://doi.org/10.1016/j.jjimei.2021.100019.

Parikh, P., Abburi, H., Badjatiya, P., Krishnan, R., Chhaya, N., Gupta, M., Varma, V., 2019. Multi-label categorization of accounts of sexism using a neural framework, 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), Association for Computational Linguistics, Hong Kong, China. pp. 1642–1652. URL: https://aclanthology.org/D19-1174, doi:10.18653/v1/D19-1174.

Pires, T., Schlinger, E., Garrette, D., 2019. How multilingual is multilingual BERT?, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy. pp. 4996–5001. URL: https://aclanthology.org/P19-1493, doi:10.18653/v1/P19-1493.

Regmi, K., Naidoo, J., Pilkington, P., 2010. Understanding the processes of translation and transliteration in qualitative research. International Journal of Qualitative Methods 9, 16 – 26.

Roy, P.K., Bhawal, S., Subalalitha, C.N., 2022. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. Computer Speech & Language 75, 101386. URL: https://www.sciencedirect.com/science/article/pii/S0885230822000250, doi:https://doi.org/10.1016/j.csl.2022.101386.

Roy, P.K., Tripathy, A.K., Das, T.K., Gao, X.Z., 2020. A framework for hate speech detection using deep convolutional neural network. IEEE Access 8, 204951–204962. doi:10.1109/ACCESS.2020.3037073.

Sawhney, R., Mathur, P., Jain, T., Gautam, A.K., Shah, R.R., 2021. Multitask learning for emotionally analyzing sexual abuse disclosures, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online. pp. 4881–4892. URL: https://aclanthology.org/2021.naacl-main.387, doi:10.18653/v1/2021.naacl-main.387.

Silva, M., Silva, L., 2021. Hate speech dissemination in news comments: analysis of news about lgbt universe on facebook cybermedia from mato grosso do sul. Intercom Revista Brasileira de Ciências da Comunicação 44, 137.

Singh, S., Anand, T., Ghosh Chowdhury, A., Waseem, Z., 2021. "hold on honey, men at work": A semi-supervised approach to detecting sexism in sitcoms, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, Association for Computational Linguistics, Online. pp. 180–185. URL: https://aclanthology.org/2021.acl-srw.19, doi:10.18653/v1/2021.acl-srw.19.

Sufi, F.K., 2022. Identifying the drivers of negative news with sentiment, entity and regression analysis. International Journal of Information Management Data Insights 2, 100074. URL: https://www.sciencedirect.com/science/article/pii/S2667096822000179, doi:https://doi.org/10.1016/j.jjimei.2022.100074.

Thurlow, C., 2001. Naming the "outsider within": homophobic pejoratives and the verbal abuse of lesbian, gay and bisexual high-school pupils. Journal of Adolescence 24, 25–38. URL: https://www.sciencedirect.com/science/article/pii/S0140197100903713, doi:https://doi.org/10.1006/jado.2000.0371.

Tontodimma, A., Nissi, E., Sarra, A., Fontanella, L., 2020. Thirty years of research into hate speech: topics of interest and their evolution. Scientometrics doi:10.1007/s11192-020-03737-6.

Uppunda, A., Cochran, S., Foster, J., Arseniev-Koehler, A., Mays, V., Chang, K.W., 2021. Adapting coreference resolution for processing violent death narratives, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online. pp. 4553–4559. URL: https://aclanthology.org/2021.naacl-main.361, doi:10.18653/v1/2021.naacl-main.361.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I., 2017. Attention is all you need, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Ventriglio, A., Castaldelli-Maia, J.M., Torales, J., De Berardis, D., Bhugra, D., 2021. Homophobia and mental health: a scourge of modern era. Epidemiology and Psychiatric Sciences 30, e52. doi:10.1017/S2045796021000391.

Wadud, M.A.H., Kabir, M.M., Mridha, M., Ali, M.A., Hamid, M.A., Monowar, M.M., 2022. How can we manage offensive text in social media - a text classification approach using lstm-boost. International Journal of Information Management Data Insights 2, 100095. URL: https://www.sciencedirect.com/science/article/pii/S2667096822000386, doi:https://doi.org/10.1016/j.jjimei.2022.100095.

Wandrekar, J., Nigudkar, A., 2020a. What do we know about lgbtqia+ mental health in india? a review of research from 2009 to 2019. Journal of Psychosexual Health 2, 26–36. doi:10.1177/2631831820918129.

Wandrekar, J., Nigudkar, A., 2020b. What do we know about lgbtqia+ mental health in india? a review of research from 2009 to 2019. Journal of Psychosexual Health 2, 26–36. doi:10.1177/2631831820918129.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A., 2020. Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on

Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online. pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6, doi:10.18653/v1/2020.emnlp-demos.6.

Wu, H.H., Hsieh, S.K., 2017. Exploring lavender tongue from social media texts[in Chinese], in: Proceedings of the 29th Conference on Computational Linguistics and Speech Processing (ROCLING 2017), The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Taipei, Taiwan. pp. 68–80. URL: https://aclanthology.org/O17-1007.

Yang, X., McEwen, R., Ong, L.R., Zihayat, M., 2020. A big data analytics framework for detecting user-level depression from social networks. International Journal of Information Management 54, 102141. URL: https://www.sciencedirect.com/science/article/pii/S0268401219313325, doi:https://doi.org/10.1016/j.ijinfomgt.2020.102141.

Yasaswini, K., Puranik, K., Hande, A., Priyadharshini, R., Thavareesan, S., Chakravarthi, B.R., 2021. Iiitt@ dravidianlangtech-eacl2021: Transfer learning for offensive language detection in dravidian languages, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pp. 187–194.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., Çöltekin, Ç., 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020), in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online). pp. 1425–1447. URL: https://aclanthology.org/2020.semeval-1.188.