International Journal of Data Science and Analytics (2024) 18:49–68
https://doi.org/10.1007/s41060-023-00400-0

REGULAR PAPER

Detection of homophobia and transphobia in YouTube comments

Bharathi Raja Chakravarthi[1]

Received: 22 June 2022 / Accepted: 6 June 2023 / Published online: 19 June 2023
© The Author(s) 2023

Abstract
Users of online platforms have negative effects on their mental health as a direct result of the spread of abusive content across social media networks. Homophobia are terms that refer to the fear, hatred, discomfort, or suspicion of or toward those who identify as homosexual or bisexual. Transphobia is fear, hatred, discomfort toward those who are transgenders. Homophobia/transphobia speechs are a sort of offensive language that can be summed up as hate speech directed toward LGBTQ+ persons, and it has become an increasing concern in recent years. The homophobia and transphobia found online are a serious societal issue that can make online platforms toxic and unwelcoming to LGBTQ+ individuals and hinder the eradication of equality, diversity, and inclusion. We present a new dataset for online homophobia and transphobia detection that has been annotated by experts, which will enable homophobic and transphobic content to be automatically recognized. The dataset includes 15,141 annotated comments written in English, Tamil, and both Tamil and English. Additionally, we provide the outcomes of our benchmark system in a variety of machine learning models. For the purpose of developing benchmark systems, we conducted a number of experiments utilizing a variety of cutting-edge machine and deep learning models. Furthermore, we discuss our shared task conducted at LTEDI-ACL 2022 workshop to improve the research in homophobia and transphobia detection. It garnered 10 systems for the Tamil language, 13 systems for the English language, and 11 systems for the combination of Tamil and English languages. The best systems for Tamil, English, and Tamil–English each received an average macro F1 score of 0.570, 0.870, and 0.610, respectively.

Keywords Homophobia · Transphobia · Inclusion · Multilingual · Hate Speech

1 Introduction

The social media platforms of the twenty-first century have evolved into the nerve center of divisive viewpoints, claims, and conflicts [1, 2]. The convenience of accessing information from social media not only contributes to the proliferation of good conversations, but it also makes phenomena such as cyberbullying and hate speech possible [3]. Despite the progress that has been made around LGBTQ+ rights, the internet continues to be an unwelcoming place for LGBTQ+ people. The increasing frequency, severity, and complications of hate crimes committed online are mirrored in the offline world in the following ways: Hate crimes directed at LGBTQ+ people and their allies have shown a sharp rise in the past three years.[1] A report on hate crimes committed online due to homophobia, biphobia, and transphobia was presented in 2020 in the UK by the LGBTQ+ anti-violence organization called Gallop.[2] The organization polled 700 persons who identified as LGBTQ+ and circulated the survey through online community networks of LGBTQ+ activists and individuals [4]. The findings offer cause for concern: In the past five years, eight out of ten people have been exposed to hate speech online, and one out of five people have reported being the target of online abuse at least 100 times. The percentage of transgender people who encounter online harassment is significantly higher (93%) than in the case of cisgender people (70%). It is particularly concerning that 18 percent of people indicated that offline occurrences were associated with online abuse that occurred on the internet [5]. These numbers paint a troubling picture of the reality that LGBTQ+ persons face on a daily basis.

✉ Bharathi Raja Chakravarthi
bharathi.raja@insight-centre.org

[1] Insight SFI Research Centre for Data Analytics, School of Computer Science, University of Galway, Galway, Ireland

[1] https://www.theguardian.com/world/2021/dec/03/recorded-homophobic-hate-crimes-soared-in-pandemic-figures-show.

[2] https://www.report-it.org.uk/files/online-crime-2020_0.pdf.

Homophobia/transphobia is a type of abuse that can take the shape of physical violence such as murder, mutilation, or beating; explicit sexual violence such as rape, molestation, or penetration; or a breach of privacy in the form of the disclosure of personal information [6–8]. The comment "Gays ought to be shot dead" is one of the examples. Other examples of homophobia/transphobia comments include "Gays should be stoned," "Someone should rape that lesbo to make her into straight," "You should kill yourself," "You lesbos, I know where you live, I will visit you tonight," and "Knock the gay out of him"; these are all comments that have been directed at socially vulnerable LGBTQ+ individuals.

Automatic recognition of homophobic and transphobic terminology on the internet could make it simpler to block damaging anti-LGBTQ+ content and advance the internet toward the achievement of equality, diversity, and inclusion. While considerable effort has been devoted to identifying aggression [9], misogyny [10, 11], and racism [12], homophobic or transphobic verbal abuse has received significantly less attention than racist or other hate speeches. The lack of annotated homophobic and transphobic data has hindered the creation of homophobic and transphobic speech detection systems. As in the rest of the world, socially vulnerable LGBTQ+ individuals in India are subjected to various kinds of online abuse that perpetuates and legitimizes homophobic attacks, the inferior social standing of LGBTQ+ individuals, sexual assault, assault, and mistreatment of them, and contempt toward them [13–18]. Moreover, the online harassment of these vulnerable persons may evolve into systematic bullying campaigns launched on social media to target and, in some cases, brutally threaten LGBTQ+ individuals who are famous on social media [19–21].

In this study, we introduce a dataset for the identification of homophobia/transphobia not only in English but also in under-resourced Tamil (ISO 639-3: tam) and code-switched Tamil–English languages.

- We propose the identification of homophobia/transphobia in online social media comments to remove hate speech toward socially vulnerable LGBTQ+ individuals.
- We apply the schema to create a multilingual homophobia/transphobia dataset for the identification of hate speech toward socially vulnerable LGBTQ+ individuals. This is a new large-scale dataset of English, Tamil, and Tamil–English (code-mixed) YouTube comments with high-quality annotation.
- We perform an experiment on our homophobia/transphobia dataset using different state-of-the-art machine and deep learning models to create benchmark systems.
- We also perform a shared task at LTEDI-ACL 2022 workshop to improve the research on the homophobia/transphobia detection in online comments. We present the results obtained by and the methodology undertaken by the international researchers who used our data.

2 Related works

Online social media platforms are becoming increasingly infested with hate speech, especially homophobic and transphobic speech. Hate speech is distinguished from other types of speech on social media by the fact that it is directed toward a particular group of people; from this point of view, it is also distinct from offensive speech, which consists solely of the use of language that is considered to be vulgar or otherwise inappropriate [22–24]. The detrimental impacts of online homophobic and transphobic speech on individual psychological well-being, as well as wider intergroup relations, have been the subject of empirical studies conducted by social scientists [25, 26]. The emotional toll that exposure to excluding and homophobic/transphobic speech takes on socially vulnerable LGBTQ+ people is significant [27–29]. Higher levels of bias may be observed in conjunction with expanding experiences of desensitization when populations consume online content with larger degrees of toxicity. In the context of online conversations, we define toxicity as comments that provoke immediate toxic responses. While these toxicity triggers may vary by group and issue due to varying linguistic norms and usages [30]. This is because the toxicity of the content that is consumed online is increasing. Furthermore, widespread hate speech, in the long run, adds to an increased possibility of radicalization directed toward socially vulnerable LGBTQ+ groups.

In the body of academic research, the definition of online hatred, homophobia, and transphobia has frequently been studied through a variety of theoretical perspectives and conceptual frameworks, including social psychology, human–computer interaction, politics, and aspects of legislation and regulation [31–34]. The identification of online hate speech toward LGBTQ+ people in large-scale interactions using methods that may be scaled up accordingly constitutes a significant challenge from a computational point of view. Recent developments in machine learning and natural language processing (NLP) have led to significant advancements in the area of automated hate speech identification [35]. These advancements have resulted in major progress. For instance, deep learning strategies have been effectively employed in cutting-edge methods for identifying hate speech, and these strategies have been successful in adequately accounting for the complex linguistic traits that characterize hate speech online [35, 36]. Based on a particular case study and/or the application of baseline datasets, the various machine learning approaches each employ their unique terminology to describe what constitutes hatred. Approaches that are lexicon- or embedding-based have also been created thanks to work that

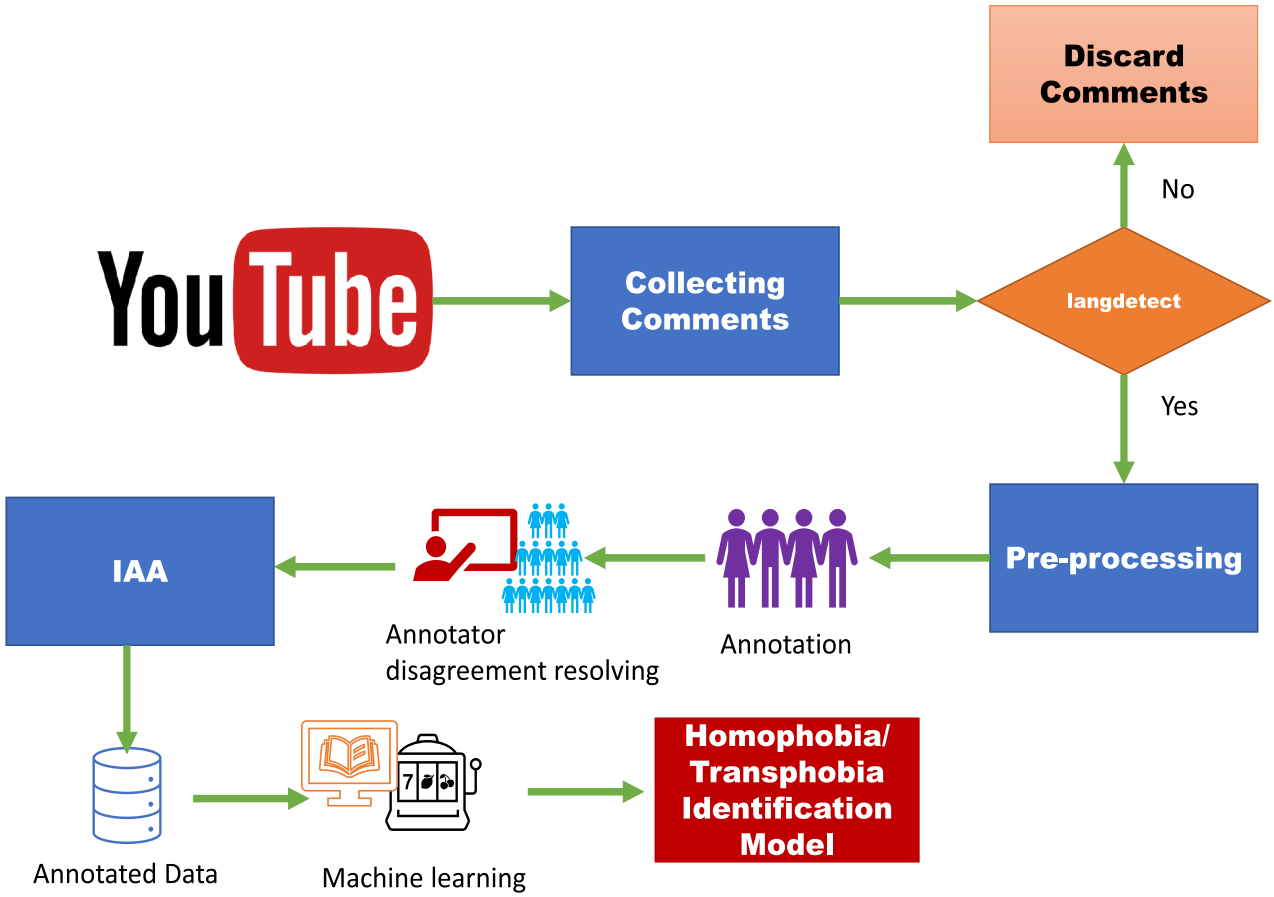


Fig. 1 Process flow diagram. IAA (Inter-annotator Agreement) calculation

is more theoretically motivated. These approaches have special applicability for investigating the targets of hate speech [37].

In general, solutions to these challenges are found through the application of machine learning-based text categorization algorithms. These methods can range from supervised learning to transfer learning and from traditional shallow machine learning to deep learning [38]. Simple approaches can determine whether an instance of communication constitutes hate speech based on whether it contains a potentially hateful keyword. However, these algorithms are unable to identify hateful content that is only implicitly hateful[3] and does not employ certain keywords [39–41]. In addition, some of these keywords may be used mockingly and are not always considered offensive (e.g., swine, trash, etc.). These procedures lead to the identification of a large number of false positives [42, 43]. Apart from a few exploratory studies, there is a lack of development and testing of models using data from multiple social media platforms. This is the case despite the fact that hate has been observed as a problem on multiple online social media platforms, such as Reddit, YouTube, Wikipedia, Twitter, and so on. Several different datasets have been developed to locate instances of hate speech [44], racial bias in hate speech [45], countering hate speech [46], hierarchically labeled hate speech [47], and abusive language [48]; detect bullying [49], and identify offensive language [50, 51].

While there is a substantial body of work in the field of NLP that deals with binary gender prejudice, hate speech, and abusive and offensive languages in general, the research landscape assessing the damages suffered by members of the LGBTQ+ community online is relatively limited. Wu et al. [52] investigated the linguistic behavior that may be discovered in LGBTQ+-generated Chinese literature and demonstrated that standard methods trained to discern gender from text fail in more complicated dimensions. Ljubešić et al. [53] developed emotion lexicons for Croatian, Dutch, and Slovene and then utilized these lexicons to search for texts that include or exclude socially undesirable speech on the subjects of migration and LGBTQ+. The fact that research on this subject is still in its infancy means that it is plagued by a number of shortcomings that are linked to both the particu-

[3] Implicit hate speech involves circumlocution, metaphor, or stereotypes to convey hatred of a specific group, and its hatred can be caught by analyzing its overall compositional meanings. Example from [39]: (3) Hillary's welfare army doesn't really want jobs. They want more freebies.

lar goals and nuances of offensive language directed toward homophobia and transphobia, as well as the nature of the classification task in general, which prevents systems from achieving ideal results. One of the most significant difficulties is the inherent difficulty in defining offensive language, as well as the pervasive ambiguity in the usage of similar phrases (such as abusive, poisonous, harmful, hateful, or violent language), which vary from culture to culture and are susceptible to highly subjective interpretations depending on the individual. If we utilize studies on "sissy boys" to explain transphobia, the former of which is one of the cruelest labels ever coined and gives rise to the perception that a transgender person suffers from a psychiatric disorder, then we may comprehend the relevance of this vital but basic contrast—between sexism and androcentrism. How researchers and members of society see homosexuality and homophobia during a time that the test of the study is being produced also has an effect on the measures of homophobia that are used.

3 Homophobia and transphobia

Homophobia refers to negative attitudes and reactions toward homosexuals. Homophobia has been variously described by authors as a cultural phenomena, a set of attitudes, and a psychological characteristic [8]. It is believed that cultural "homophobia" serves to preserve traditional sex role disparities. Homophobia is characterized in terms of attitudes as a collection of established, unfavorable attitudes toward homosexual people [54]. As a personality dimension, "homophobia" is associated with rigidity, authoritarianism, conservatism, and intolerance for ambiguity and deviation [55]. Fear, ignorance, and a lack of knowledge on and tolerance for sexual choice have given rise to a second group of ideas toward homosexuality. One is that gays molest children. The misconception that only gays would engage in such relationships further restricts such conduct [56]. Another example is the notion that homosexuals are promiscuous (many sexual encounters with multiple partners). As many people bear moral objections to promiscuity, associating homosexuality with promiscuous behavior reinforces traditional beliefs regarding ethical sexual activity. In this instance, homosexuality falls under the broader religious system of promiscuity. Such beliefs obfuscate the fact that many homosexuals engage in long-term partnerships.

Homophobia (prejudice toward lesbian and gay individuals) is distinguished from transphobia (prejudice against transgender individuals) based on the perceived social status challenges posed by lesbian and gay individuals vs transgender individuals [57]. The gender identification of one's sexual partners may influence one's own sexual orientation, which refers to the person(s) to whom one feels a strong sexual attraction to [58]. Transgender individuals are people who live with a gender identity that is different from traditional heteronormative definitions and may or may not also seek gender affirmation surgery. While gay and lesbian individuals are defined by their sexual orientation, transgender individuals have a gender identity that is different from traditional heteronormative definitions [59]. These people do not adhere to the accepted conventions of gender identities and gender roles or cross over from one gender to another. Transphobia, thus, focuses on non-heteronormative gender identity and possibly non-gender heteronormative gender roles, whereas homophobia focuses on non-heteronormative gender identity and sexual orientation [60]. Transphobia differs from homophobia in that it encompasses not only revulsion and irrational fear of transgender and transsexual individuals but also cross-dressers, feminine men, and masculine women. That is, it is concerned with gender roles and gender identity and not necessarily sexual orientation [61].

Both homophobia and transphobia are terms that refer to the negative attitudes toward people who identify as homosexual or transgender, respectively [7]. Transphobia is characterized as a serious issue that impacts the lives of a great number of people. It is the fear of and/or hatred toward transgender people. People who identify as transgender are typically excluded from and ignored in homosexual communities as well as heterosexual societies. Due to ignorance and animosity, many transgender people are prohibited from coming out or identifying themselves as trans, which further obscures the community of transgender people. We came up with a hierarchical taxonomy that has two levels of classification. In the first place, we will differentiate between content that is homophobic, content that is transphobic, and content that is not anti-LGBTQ+.

3.1 Homophobic content

Homophobic content can be described as "an attitude of animosity against male or female homosexuals." Lesbophobia, gayphobia, and biphobia are all families of phobias that target different subgroups of the LGBQI+ community. However, there is a difference between general homophobia and more specific forms of the condition. Under the umbrella term of homophobia, this article addresses lesbophobia, gayphobia, and biphobia. Homophobic content is a type of harassment that involves the use of pejorative labels (such as "fag" or "homo") or denigrative phrases (such as "don't be a homo" or "that's so gay" or "that's so lesbo") directed against people who are gay, lesbian, bisexual, queer, or gender non-conforming. Content that supports, promotes, urges, or incites violence against LGBQ+ individuals or groups suggests a purpose or desire to damage or cause harm to LGBQ+ individuals and is considered as homophobic content in our paper.

3.2 Transphobic content

"Transphobia" refers to hostile responses to people who are perceived to be "trans." The term "trans" is typically used to describe people whose designations of their gender are independent from either their assigned gender or from the administrative sex category listed on their original birth certificate [62]. Transphobia refers to hostile responses to people who are perceived to be "trans." It may be defined as a feeling of repulsion against those who do not comply with the gender standards of society. It manifests in the form of prejudice, discrimination, harassment, and, sometimes, acts of violence directed toward transgender people [63]. Although it is impossible to determine the whole scope of the problem, many people have been on the receiving end of acts of discrimination, aggression, victimization, and sexual assault that were motivated by the victim's gender identification. The brutal killing of hundreds of transgender people all around the world is perhaps the most horrifying manifestation of transphobia [64]. Pejorative terms that are used to degrade transgender individuals in a vulnerable state are known as transphobic pejoratives. It includes idioms that indicate implicit animosity or fury against transgender people, such as "she-male," "it," and "9," as well as phrases that are openly insulting and derogatory, such as "tranny," "trannie," "cross-dresser," or "drag." In a similar vein, the phrases "not man enough," "not women enough," "will never be a complete man," and "will never be full women" are all synonymous with one another. It includes not only declaring the desire to take action against transgender persons but also expressing preferences for how they should be treated, which may include using threatening language, engaging in physical violence, engaging in sexual assault, or invading someone's privacy.

3.3 Non-anti-LGBTQ+ content

This refers to content that does not contain any homophobic or transphobic slurs, pejoratives, or threats in the manner that was defined in previous sections. Most of the time, this issue has nothing to do with exploitation or socially vulnerable LGBTQ+ persons in general. For example, one may encourage people to like the video, subscribe to the channel, or like the comment one left on the video. On the other hand, it may include using different types of abusive words that are not anti-LGBTQ+.

4 Dataset construction

YouTube is popular across the Indian subcontinent as a result of the vast amount of content on the platform that can be accessed on the internet. Some of the content that can be found on YouTube includes music, courses, product evaluations, trailers, and other similar videos. YouTube enables people to upload their own content, which may then be discussed by other users. As a result, it allows for more content to be developed by users in languages that have few resources. This also applies to vulnerable members of the LGBTQ+ community who view videos and leave comments on the videos to which they relate. We made the decision to compile our data from the social media comments posted on YouTube,[4] which is the most widely used platform throughout the world for voicing one's opinion on a specific video.

In India, a country where the LGBTQ+ do not have equal marital rights, vulnerable young people in the LGBTQ+ community are defined as an "invisible" minority and one of the most significant "at-risk" groups of adolescents. This is an expected description. These people have no other way to locate persons with comparable experiences other than to search for them on social media. We did not utilize any comments from personal coming out stories by LGBTQ+ persons because they contained private information, and we did not want to disclose them. Instead, we compiled a collection of videos from well-known users on YouTube that explain LGBTQ+ issues in the hope that more people will have a positive outlook. To guarantee that our dataset has an adequate amount of homophobic and transphobic abuse, we started by selecting certain films of pranks uploaded to YouTube by users with usernames such as "Gay Prank," "Transgender Prank," and "Legalizing Homosexuality." There were some videos that discussed the advantages of transgenderism; nevertheless, the majority of the videos from both popular channels and news channels portrayed transgender individuals as persons who take advantage of others and start disputes. It was challenging to locate a video on YouTube that discussed LGBTQ+ concerns in Tamil, as the topic is still taboo, marital equality is not legal, and until recently, homosexuality was criminalized in India [65].

For the purpose of collecting the comments, *YouTube Comment Scraper tool*[5] was utilized. These comments were used leveraged by us in the process of creating our datasets with manual annotations. The gathering of Tamil comments was one of our primary objectives. However, we found that the text contained a significant amount of English as well as a mixture of other languages.

Code mixing is a common and natural phenomenon among Tamil speakers as a result of their bilingual and multilingual language use. Users of social media write in the Roman alphabet for convenience (phonetic typing), which increases the likelihood of code-mixing with a language that uses the Roman alphabet. The Tamil language has a native

[4] https://www.youtube.com/.

[5] https://github.com/philbot9/youtube-comment-scraper.

Table 1 Raw dataset statistics by language based on Langdect

Language	Number of comments	Number of tokens	Number of characters
English	7265	116,015	632,221
Tamil	5240	255,578	787,177
Tamil–English	10,319	88,303	628,077
Total	22,824	459,896	2,047,475

Fig. 2 World cloud for English training data

script that has its own unicode, but users still choose to write in the Roman alphabet instead [66].

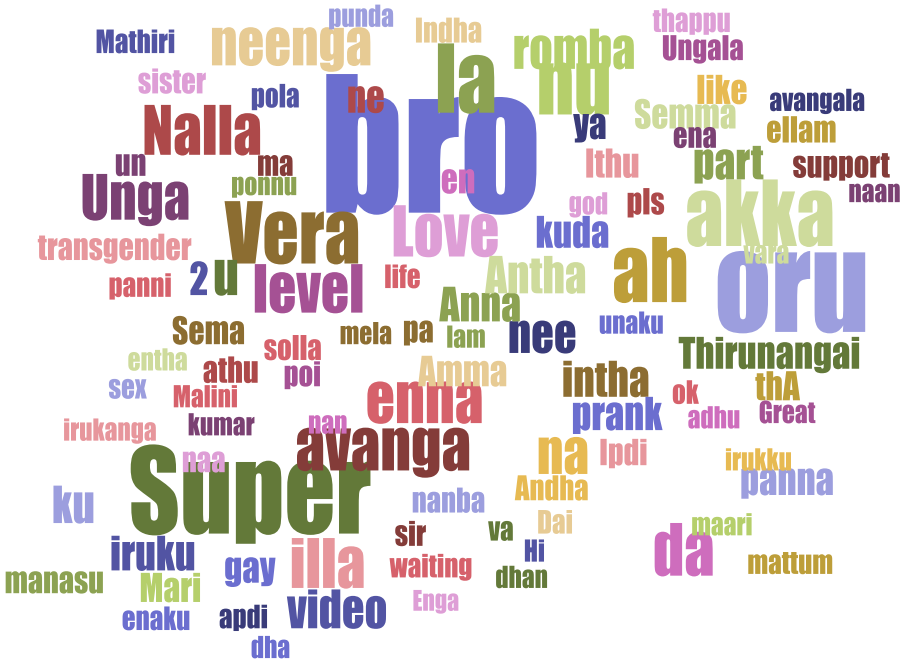
We had a hard time getting the most important information from the comment section that suited our goal. This was complicated by the fact that there were responses in languages that were not our target. We used the *Langdetect library*[6] to separate languages and tell them apart as part of the process of getting the data ready for cleaning. We split the data into three groups, such as Tamil and English-based *langdetect library*. We kept the code-mixed Tamil–English that was left. As per data privacy laws, we removed all information about users from the corpus. For preparing for the test, we got rid of extra information, such as URLs. Most of the comments were written in Roman letters and either used Tamil or English grammar mixed with Tamil words. Between August 2020 and February 2021, all the comments were gathered. After cleaning, our data reduced. The final dataset ready for annotation is shown in Table 1. The wordcloud in Figs. 2 and 3 presents the distribution of words in training data for the shared task.

4.1 Annotation

After pre-processing and cleaning the text, we reviewed it thoroughly to verify that it did not include any personally identifiable information. When developing training datasets for sensitive issues such as ours, producing high-quality annotations is a huge challenge. Several variables influence the production of such annotations, including the cultural influences and personal biases of the annotators. As LGBTQ+ affairs are a controversial subject, it was difficult to locate annotators. For English, we utilized members of our in-house LGBTQ+ groups. It was difficult to locate willing Tamil annotators; even Tamil speakers in culturally progressive regions such as Europe, the UK, and the USA did not want to be linked with this project out of fear that they would be considered part of the LGBTQ+ community. After a lengthy search, we discovered LGBTQ+ or

[6] https://pypi.org/project/langdetect/.

(a) World Cloud for Tamil Training Data



(b) World Cloud for Tamil-English Training Data

LGBTQ+ ally volunteer annotators. We trained them by showing them YouTube movies explaining what LGBTQ+ meant and whether it is normal.

We collected the annotator's email address using Google Forms so that they may only annotate once. Each Google Form allows for a maximum of 100 comments. For comments in the English language, we circulated the Google Form among members of the LGBTQ+ organization of the National University of Ireland, Galway. For Tamil and code-mixed language, we emailed the Google Form to several LGBTQ+ societies in Tamil Nadu, but we received very few answers. In the end, seven annotators offered to anno-

Table 2 Dataset statistics after annotation

Language	Number of comments	Number of tokens	Number of Characters
English	4946	82,111	438,980
Tamil	4161	197,237	539,559
Tamil–English	6034	66,731	435,890
Total	15,141	346,079	1,414,429

Table 3 Dataset distribution at class level

	English	Tamil	Tamil–English
Homophobic	276	723	465
Transphobic	13	233	184
Non-anti-LGBTQ+ content	4657	3205	5385
Total	4946	4161	6034

tate the combined Tamil and Tamil–English code. All of them are multilingual in Tamil and English and were ready to take the work seriously. Further, three people responded for English. Because we contacted people through college societies, every annotator was either a graduate or a postgraduate. Each annotator identifies as LGBTQ+ or as an ally of the LGBTQ+ community. At least three annotators annotated every remark. If more than three annotators agreed, the labels were accepted; otherwise, they were marked as disputes. Annotators and the writers of this work examined the dispute under the direction of the author, who established the annotation taxonomy, was familiar with the literature on homophobia and transphobia, and is fluent in Tamil and English. The facilitator was responsible for promoting debate among the annotators and ensuring that the final labels adhered to the taxonomy. The debate took place online using Google Meet. Each argument was resolved until the annotators reached consensus on the final label. If no consensus could be reached, those comments were eliminated from the study's dataset. We paid 300 rupees per hour to Indian annotators; the typical wage in India is around Rs. 16,000 per month (Rs. 533 per day), according to moneymint.com.[7] We removed the comments on which a consensus couldn't be reached ultimately. The final data statistics after annotation is shown in Table 2. Table 3 shows the classwise distribution of dataset. From Fig. 4, we can see that there is less transphobic and homophobic content in English compared to Tamil and Tamil–English.

4.2 Ethical concerns

Data from social media is very sensitive, especially when it has to do with the LGBTQ+ community. We took great care to reduce the risk of people being identified in the data by removing personal information such as names, but we did not remove celebrity names. However, to look into equality, diversity, and inclusion (EDI), we had to keep track of information on race, gender, sexual orientation, ethnicity, and philosophical views. Annotators could only see anonymous posts and promised not to get in touch with the person who made them. Researchers who want to use the dataset for research will only be able to do so if they agree to follow ethical rules.

4.3 Inter-annotator agreement (IAA)

We sought agreement from the majority of the annotators for aggregating the annotations on homophobic/transphobic comments; the comments that did not receive a majority agreement in the first round were collected and placed in a second Google Form so that more annotators may contribute them. Following the last round of annotation, we computed the inter-annotator agreement. Using Krippendorff's alpha (α), we measure the clarity of the annotation and report on inter-annotator agreement. Krippendorff's alpha is a statistical measure of annotator agreement that reveals how well the resultant data conforms to the underlying data [67]. Although Krippendorff's alpha is computationally expensive, it is more relevant in our case, as more than two annotators annotated the comments and not all phrases were annotated by the same annotator. Further, it is unaffected by missing data; permits flexibility in sample sizes, categories, and the number of raters; and may be used to any measurement level, including nominal, ordinal, interval, and ratio. Krippendorff's alpha is obtained by the following:

$$\alpha = 1 - \frac{D_o}{D_e} \tag{1}$$

where D_o is the observed disagreement between the homophobic/transphobic labels given by the annotators, and D_e is the disagreement that is predicted when the coding of homophobic/transphobic may be ascribed to chance rather than to an intrinsic quality of the homophobic/transphobic label itself.

$$D_o = \frac{1}{n} \sum_{c} \sum_{k} o_{ck\ metric}\ \delta^2_{ck} \tag{2}$$

[7] https://moneymint.com/what-is-average-salary-in-india/.

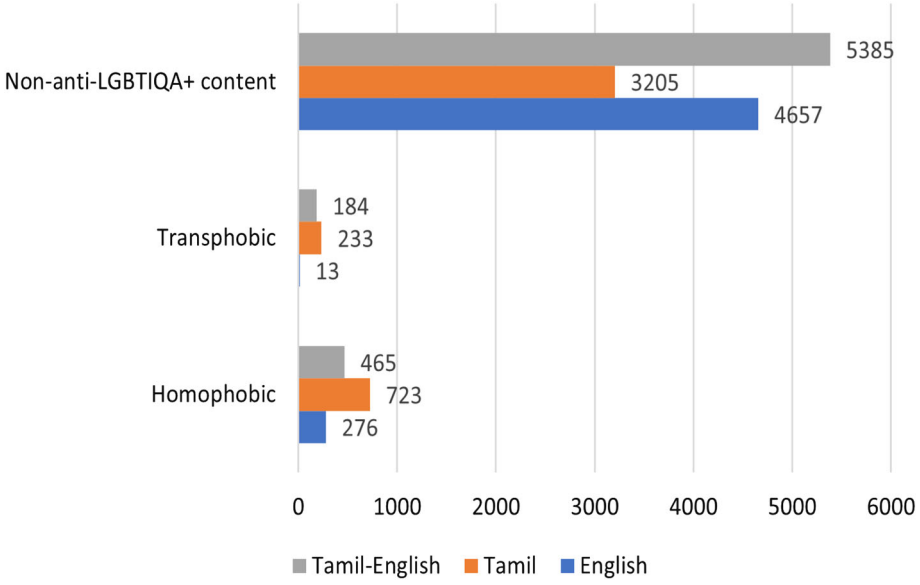


Fig. 4 Dataset distribution at class level

$$D_e = \frac{1}{n(n-1)} \sum_c \sum_k n_c \, . \, n_{k\,metric} \, \delta_{ck}^2 \tag{3}$$

Here, o_{ck} n_c n_k and n refer to the frequencies of values in the coincidence matrices, and $metric$ refers to any metric or level of measurement such as nominal, ordinal, interval, ratio, and others.

The values "0" and "1" are included in the range of α, which may be written as $1 \geq \alpha \geq 0$. When α is "1," the annotators are in complete agreement with one another, but when it is "0," the annotators' agreement is the result of solely random chance. $\alpha \geq .80$, as this is the standard requirement. While $0.67 \leq \alpha \leq 0.8$ is required as an acceptable rule of thumb that allows for preliminary inferences to be formed, $\alpha \geq .653$ is the lowest feasible limit. We used *nltk*[8] to compute (α). When we used the nominal measure to determine the level of agreement between our annotations, we obtained Krippendorf's alpha values of 0.67, 0.76, and 0.54 for English, Tamil, and Tamil–English, respectively. We have shown the details of the dataset in Table 2 and classwise distribution in Table 3.

5 Benchmark experiments

To examine our dataset, baseline models were created. We constructed three corpora, each of which includes monolingual texts in either Tamil or English, as well as multilingual texts in a code-mixed version of Tamil and English. As data was taken from social media, the text in the corpus contains a lot of noise. Therefore, various punctuation marks, tags, and symbols such as emojis and @ signs were included in the YouTube comments. To gather the data in a clean state, preprocessing procedures, namely the removal of punctuation, stop words, and tags, were utilized. We employed a stratified sampling approach using K-folds to divide the dataset into groups; every group had precisely the same percentage of labels. This allowed us to compare the results of our analysis more accurately. We decided to employ stratified sampling due to the imbalance in our dataset. For the purpose of cross-validation, we divided the data into five folds. Several different baseline models are constructed by employing various distinct feature collections and learning techniques. Machine learning models with different embeddings, such as TF-IDF, count vectorizer, BERT [68] embeddings, and fastText embeddings [69], are utilized for both monolingual and code-mixed datasets. Further, classifiers such as logistic regression, naive Bayes, random forest, support vector machines, and decision trees are utilized in the construction of baseline models with the aforementioned embeddings [70].

The deep learning model was constructed utilizing a bidirectional LSTM (BiLSTM) layer. The model consisted of an embedding layer where the input vectors are vectorized using BERT embeddings, followed by a BiLSTM layer, a flatten layer, and two dense layers. The model was developed with Keras layers [71]. Using a linear, fully connected layer with the Softmax activation function, the probability distribution across the classification classes was generated, and the class with the highest probability was selected as the final label. All of our machine learning and deep learning models were trained using Google Colab Pro.[9]

The performance of the classification model had to be evaluated once the technique and architecture of the classifier was

[8] https://www.nltk.org/.

[9] https://colab.research.google.com/.

Table 4 Results for English dataset

Classifier	Feature	Acc	Pmac	Rmac	F1mac	Pw	Rw	F1w
LR	TF-IDF (tri-gram)	0.908	0.392	0.388	0.388	0.910	0.908	0.908
LR	countvec (tri-gram)	0.916	0.388	0.378	0.382	0.908	0.916	0.912
LR	fastText	0.638	0.366	0.480	0.328	0.924	0.638	0.734
LR	BERT	0.904	0.442	0.504	0.466	0.926	0.904	0.914
NB	TF-IDF (tri-gram)	0.940	0.496	0.336	0.332	0.920	0.940	0.920
NB	countvec (tri-gram)	0.940	0.548	0.350	0.352	0.924	0.940	0.918
NB	fastText	0.738	0.360	0.424	0.344	0.910	0.738	0.804
NB	BERT	0.756	0.394	0.580	0.392	0.938	0.756	0.820
RF	TF-IDF (tri-gram)	0.812	0.426	0.362	0.354	0.908	0.812	0.832
RF	countvec (tri-gram)	0.684	0.386	0.330	0.306	0.900	0.684	0.736
RF	fastText	0.940	0.444	0.342	0.336	0.908	0.940	0.914
RF	BERT	0.944	0.534	0.424	0.442	0.928	0.940	0.926
SVM	TF-IDF (tri-gram)	0.934	0.422	0.370	0.380	0.910	0.934	0.920
SVM	countvec (tri-gram)	0.932	0.410	0.358	0.366	0.906	0.932	0.916
SVM	fastText	0.940	0.310	0.330	0.320	0.890	0.940	0.910
SVM	BERT	0.910	0.450	0.480	0.460	0.922	0.910	0.916
DT	TF-IDF (tri-gram)	0.940	0.344	0.332	0.322	0.894	0.940	0.910
DT	countvec (tri-gram)	0.940	0.412	0.334	0.326	0.904	0.940	0.912
DT	fastText	0.940	0.310	0.330	0.320	0.890	0.940	0.910
DT	BERT	0.934	0.400	0.374	0.380	0.910	0.934	0.918
BiLSTM	–	0.940	0.310	0.330	0.320	0.890	0.940	0.910
MBERT	–	0.060	0.020	0.333	0.040	0.00	0.060	0.010

chosen and constructed, respectively, to identify whether the classification model can correctly place unknown data into appropriate classes. To evaluate the efficacy of the classification algorithm, we made use of several different metrics, such as accuracy, precision, and recall, in addition to the F1 score, which are described as follows:

$$Recall(R) = \frac{TP}{(TP + FN)} \tag{4}$$

$$Precision(P) = \frac{TP}{(TP + FP)} \tag{5}$$

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{6}$$

$$F1 = \frac{(2 \times Precision \times Recall)}{(Precision + Recall)} \tag{7}$$

$$P_{mac} = \frac{1}{L}\sum_{i=1}^{L} P_i \tag{8}$$

$$R_{mac} = \frac{1}{L}\sum_{i=1}^{L} R_i \tag{9}$$

$$F1_{\text{mac}} = \frac{1}{L}\sum_{i=1}^{L} 2 \times \frac{P_{\text{mac}} \times R_{\text{mac}}}{P_{\text{mac}} + R_{\text{mac}}} \tag{10}$$

$$P_{\text{weighted}} = \sum_{i=1}^{L}(P_i \times Weight_i) \tag{11}$$

$$R_{\text{weighted}} = \sum_{i=1}^{L}(R_i \times Weight_i) \tag{12}$$

$$F1_{\text{weighted}} = \sum_{i=1}^{L}(F1_i \times Weight_i) \tag{13}$$

where TP, TN, FP, and FN refer to True Positive, True Negative, False Positive, and False Negative, respectively.

Tables 4, 5, and 6 illustrate the classification performance of several machine and deep learning models paired with a variety of features for a dataset with three classes. For English, the accuracy ranges anywhere from 0.63 to 0.94. The overall macro average score is lower than 0.4 for both accuracy and recall questions, as well as the F1 score for all three classes. We feel that there is a greater open field for future study on the identification of an ideal model for homophobia/transphobia detection because the macro average penalizes models that do not perform well with minority classes and because our dataset is severely unbalanced.

When it comes to Tamil, the accuracy ranges anywhere between 0.61 and 0.92. According to the data, it is clear that fastText with RF offers the most advantageous set of characteristics for the Tamil language. It has come to our

Table 5 Results for Tamil dataset

Classifier	Feature	Acc	Pmac	Rmac	F1mac	Pw	Rw	F1w
LR	TF-IDF (tri-gram)	0.836	0.690	0.598	0.632	0.824	0.836	0.824
LR	countvec (tri-gram)	0.846	0.722	0.596	0.642	0.834	0.846	0.832
LR	fastText	0.610	0.504	0.648	0.502	0.804	0.610	0.654
LR	BERT	0.706	0.560	0.722	0.590	0.812	0.706	0.736
NB	TF-IDF (tri-gram)	0.798	0.618	0.392	0.396	0.782	0.798	0.734
NB	countvec (tri-gram)	0.824	0.790	0.476	0.522	0.820	0.824	0.786
NB	fastText	0.720	0.544	0.648	0.568	0.798	0.720	0.748
NB	BERT	0.532	0.466	0.466	0.408	0.718	0.532	0.574
RF	TF-IDF (tri-gram)	0.468	0.628	0.610	0.466	0.880	0.468	0.576
RF	countvec (tri-gram)	0.462	0.626	0.598	0.460	0.876	0.462	0.568
RF	fastText	0.920	0.930	0.746	0.808	0.920	0.920	0.912
RF	BERT	0.882	0.796	0.728	0.752	0.882	0.882	0.880
SVM	TF-IDF (tri-gram)	0.864	0.818	0.538	0.588	0.848	0.848	0.814
SVM	countvec (tri-gram)	0.855	0.815	0.575	0.638	0.853	0.855	0.835
SVM	fastText	0.808	0.752	0.456	0.498	0.796	0.808	0.764
SVM	BERT	0.890	0.816	0.788	0.800	0.888	0.890	0.888
DT	TF-IDF (tri-gram)	0.810	0.652	0.418	0.420	0.772	0.792	0.732
DT	countvec (tri-gram)	0.780	0.554	0.368	0.358	0.734	0.780	0.708
DT	fastText	0.772	0.588	0.452	0.469	0.740	0.772	0.743
DT	BERT	0.786	0.662	0.474	0.470	0.770	0.752	0.724
BiLSTM	–	0.890	0.300	0.330	0.310	0.800	0.890	0.840
MBERT	–	0.168	0.328	0.434	0.282	0.252	0.168	0.142

Table 6 Results for Tamil–English Dataset

Classifier	Feature	Acc	Pmac	Rmac	F1mac	Pw	Rw	F1w
LR	TF-IDF (tri-gram)	0.890	0.522	0.340	0.328	0.826	0.890	0.840
LR	countvec (tri-gram)	0.890	0.508	0.340	0.326	0.822	0.890	0.840
LR	fastText	0.584	0.396	0.530	0.370	0.860	0.584	0.672
LR	BERT	0.812	0.490	0.604	0.524	0.876	0.812	0.838
NB	TF-IDF (tri-gram)	0.890	0.300	0.330	0.310	0.800	0.890	0.840
NB	countvec (tri-gram)	0.890	0.366	0.332	0.312	0.814	0.890	0.840
NB	fastText	0.676	0.408	0.460	0.382	0.848	0.676	0.738
NB	BERT	0.562	0.410	0.594	0.374	0.880	0.562	0.662
RF	TF-IDF (tri-gram)	0.428	0.478	0.356	0.184	0.864	0.428	0.434
RF	countvec (tri-gram)	0.272	0.420	0.360	0.132	0.878	0.272	0.296
RF	fastText	0.890	0.396	0.344	0.332	0.820	0.890	0.846
RF	BERT	0.890	0.610	0.366	0.380	0.846	0.890	0.852
SVM	TF-IDF (tri-gram)	0.890	0.530	0.338	0.324	0.826	0.890	0.840
SVM	countvec (tri-gram)	0.890	0.530	0.338	0.324	0.826	0.890	0.840
SVM	fastText	0.890	0.398	0.338	0.326	0.808	0.890	0.842
SVM	BERT	0.836	0.496	0.520	0.504	0.854	0.836	0.844
DT	TF-IDF (tri-gram)	0.890	0.564	0.338	0.324	0.832	0.890	0.840
DT	countvec (tri-gram)	0.890	0.498	0.336	0.322	0.818	0.890	0.840
DT	fastText	0.890	0.332	0.332	0.314	0.802	0.890	0.840
DT	BERT	0.888	0.620	0.380	0.396	0.850	0.888	0.850
BiLSTM	–	0.770	0.260	0.330	0.290	0.590	0.770	0.670
MBERT	–	0.030	0.010	0.330	0.020	0.000	0.030	0.000

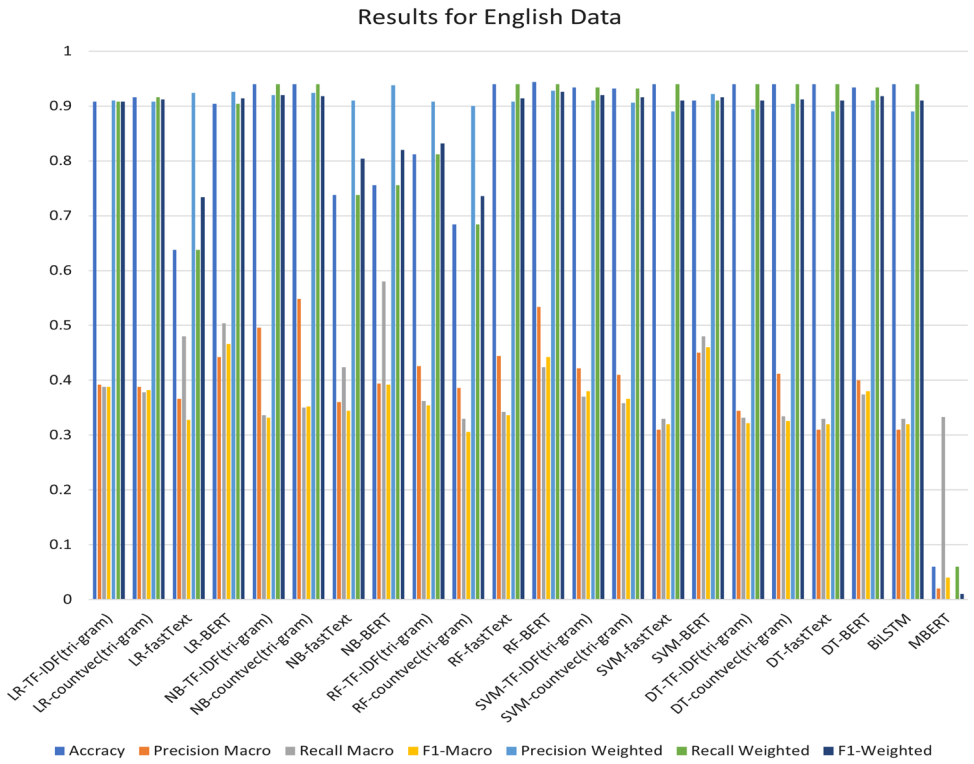


Fig. 5 Results for English data

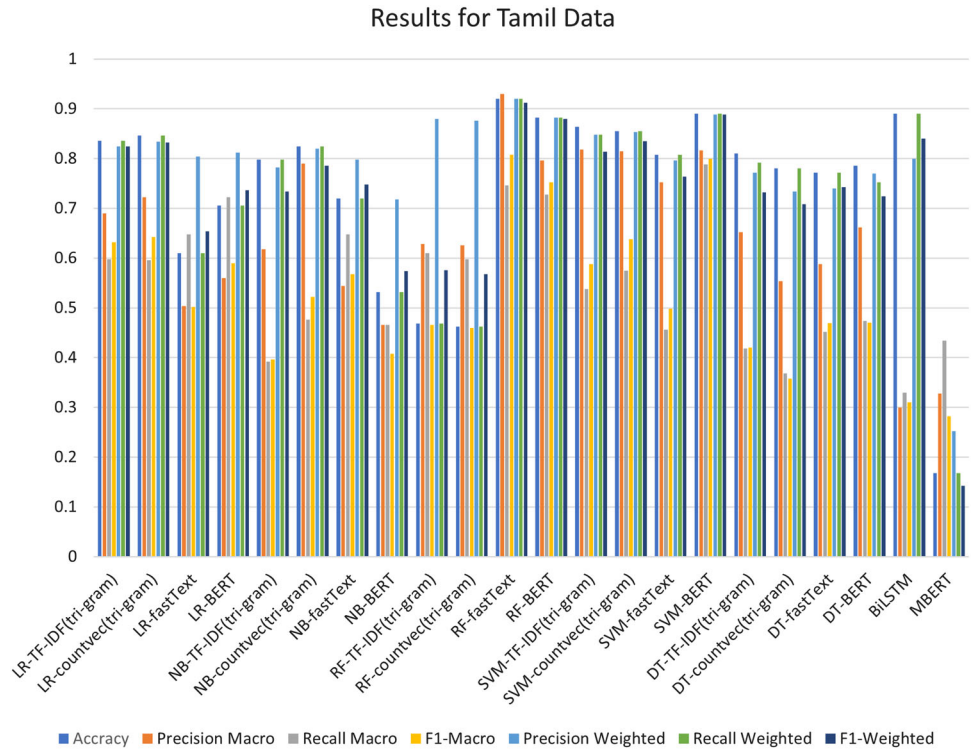


Fig. 6 Results for Tamil data

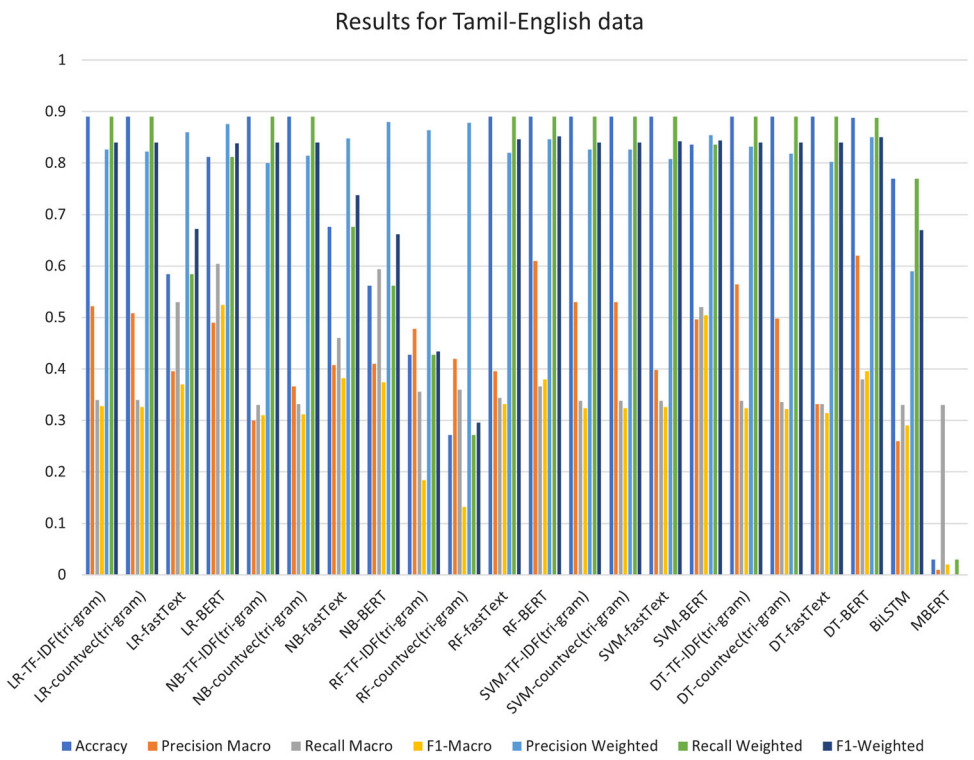


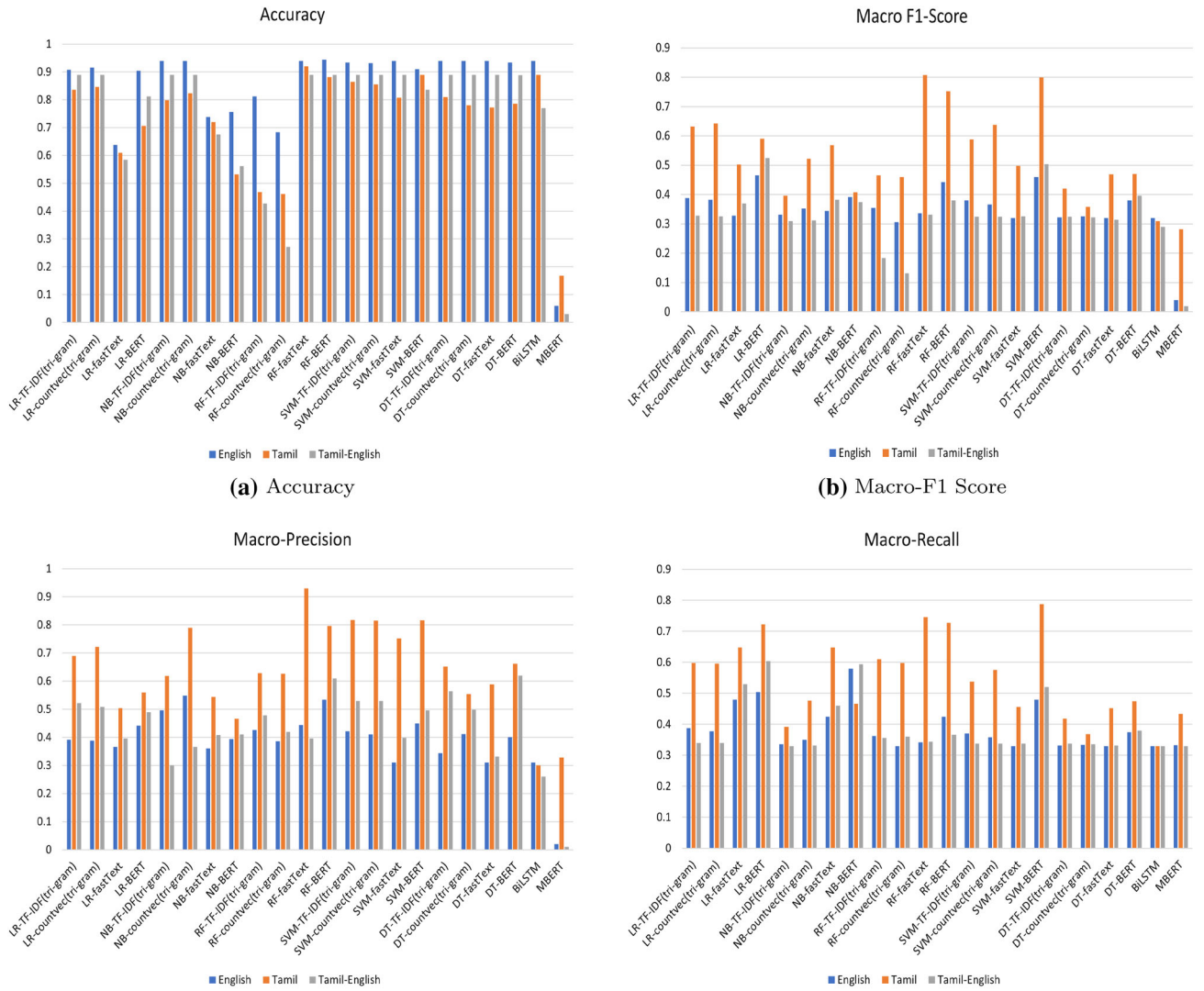
Fig. 7 Results for Tamil–English data

attention that, of all the model and feature combinations tested, random forest with BERT embedding achieves the greatest weighted F1 score in both the English and Tamil–English code-mixed scenarios. The combination of random forest with fastText embedding yields the greatest weighted F1 score for the Tamil language. As Tamil–English code-mixed settings make use of romanized writing and English words, they are practically indistinguishable from standard English. Based on the outcomes of our experiments with all three languages and the three different class label configurations, we find that a combination of deep learning and machine learning worked significantly better than either deep learning or machine learning alone. We have conducted experiments with only BiLSTM and multilingual BERT for the deep learning settings. In some of the configurations, the multilingual BERT's performance was inferior to that of every other classifier, so for the benchmarking performance, we left out the MBERT from the accuracy comparison. We have shown the experiment results of our benchmark on our new dataset, and this study could be used as a base for creating new resources on detecting homophobia and transphobia in other under-resources languages such as Kannada, Hindi, and Malay.

6 Task setting and evaluation setting

A dataset compiled from social media comments in Tamil, English, and Tamil–English will be analyzed to search for homophobic and transphobic utterances. This will be the core objective of this effort. This work involves classifying comments and posts at the comment/post level. A system must decide if a comment is homophobic, transphobic, or non-anti-LGBTQ+ content. Even if a single remark or post in the dataset is composed of many sentences, the average sentence length throughout the corpus is just one. Annotations at the level of comments and posts are included in the corpus. The participants were provided with datasets in Tamil, English, and Tamil–English for the purposes of creating, training, and testing homophobia/transphobia detection model.

The participants were provided with English, Tamil, and Tamil–English development, training, and test datasets. In the first phase, development, training, and validation, data were made available to the participants so that they may train and develop homophobia/transphobia detection systems for any of the three languages. Participants had the option of performing cross-validation on the training data or using the validation dataset for early assessments and the develop-



(a) Accuracy

(b) Macro-F1 Score

(c) Macro-Precision

(d) Macro-Recall

Fig. 8 Results for benchmarking systems

ment set for sharing hyperparameters. The objective of this phase was to verify that all participant-developed systems were ready for review prior to the release of test results. The application was selected for examination and for creating the ranking list. The accuracy of the predictions was measured against gold standard labels.

All the datasets have an unbalanced distribution of homophobia and transphobia classes. Most comments in the Tamil–English code-mixed dataset belong to the non-anti-LGBTQ+ content (5385) class, indicating a class imbalance as seen in the Table 3. In the monolingual dataset, the non-anti-LGBTQ+ content (Tamil: 3205 and English 4657) class emerged as the majority class, compared to the other two categories. This disparity was rectified by selecting the macro-averaged F1 score (F) official evaluation metric task significant variance number of instances in different classes. Macro-averaging gives the same weight to all classes, irrespective of their size. We utilized a Scikit learn classification report tool.[10] Participants submitted up to five test runs, with one of them serving as official runs that would be scored and shown on the leader board. If no official runs were specified, the most recent contributions from each team were assumed to be official. In their papers, we allowed groups to explore the distinctions between their systems. The goal was to compare the effectiveness of various setups on the test set.

[10] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html.

Table 7 Rank list for Tamil language

Team	Acc	Pmac	Rmac	F1mac	Pw	Rw	F1w	Rank
ARGUABLY	0.940	0.880	0.850	0.870	0.940	0.940	0.940	1
NAYEL [72]	0.920	0.860	0.810	0.840	0.920	0.920	0.920	2
UMUTeam [73]	0.920	0.850	0.800	0.820	0.920	0.920	0.920	3
hate-alert	0.900	0.830	0.750	0.780	0.900	0.900	0.900	4
Ablimet [74]	0.890	0.810	0.710	0.750	0.880	0.890	0.880	5
bitsa_nlp [75]	0.850	0.690	0.610	0.640	0.840	0.850	0.840	6
niksss	0.810	0.720	0.590	0.620	0.820	0.810	0.810	7
Sammaan [76]	0.880	0.520	0.580	0.550	0.850	0.880	0.860	8
SSNCSE_NLP [77]	0.770	0.550	0.470	0.500	0.740	0.770	0.750	9
SOA_NLP	0.690	0.360	0.360	0.360	0.670	0.690	0.680	10

6.1 Participants methodology

A total of 98 people signed up to take part in this shared task. Finally, we received a total of 10 submissions in the Tamil language, 13 submissions in the English language, and 11 submissions in the Tamil–English language. The processes to be followed and the results obtained from carrying out these activities have been outlined. The articles listed below should be referred to for additional in-depth information on the following topics:

ABLIMET [74] utilized a method that focuses on the fine-tuning of the pre-trained language model. This model performs processing on the target data and then normalizes the output of that processing using a layer normalization module. This is followed by two fully connected layers. They utilized the Roberta-base model for the English subtask and the Tamil-Roberta model for the Tamil and Tamil–English subtasks. All of these are pre-trained language models.

bitsa_nlp [75] used famous distinctive models primarily based on transformer architecture and a data augmentation approach for oversampling the English, Tamil, and Tamil–English datasets. They implemented various pre-trained language models based on transformer architectures, namely BERT, multilingual BERT (mBERT), XLM-RoBERTa, IndicBERT, and HateBERT, to classify the detection of homophobic and transphobic content.

For experiments with the code-mixed datasets, **SSNCSE_NLP** [77] used a mix of word embeddings, classifiers, and transformers. They used TF-IDF and a count vectorizers with some models, such as SVM, MLP, random forest, and K-nearest neighbors, and simple transformers, such as LaBSE, tamillion, and IndicBERT, to pull out the features.

To vectorize comments, **NAYEL** [72] tried out TF-IDF with bigram models. Then, they used a set of classification algorithms such as support vector machine, random forest, passive aggressive classifier, Gaussian naive Bayes, and multilayer perceptron. From these models, they chose support vector machine as the best because, among all the models, it was the most accurate.

Nozza [78] used finely tuned models to classify the task. They chose two large language models, BERT and RoBERTa. They chose HateBERT because it was more accurate than other models and provided better results than BERT. The team tried out ensemble modeling, made with a meta-classifier that uses each machine learning classifier's predicted label as a vote for the final label they present as a prediction. Moreover, they offered two ways to decide how an ensemble should work: majority voting and weighted voting.

Sammaan [76] constructed the classifier with a collection of transformer-based models. They placed second for English, eighth for Tamil, and 10th for Tamil–English. Experimentation was conducted using BERT, RoBERTa, HateBERT, IndicBERT, XGBoost, random forest classifier, and Bayesian optimization models.

UMUTeam [73] combined contextual and non-contextual sentence embeddings with linguistic components collected from a self-developed tool using neural networks. This team placed seventh in English, third in Tamil, and second in Tamil–English.

6.2 Results and discussion of shared task

A total of 98 individuals registered for this shared task. Fourteen teams presented conclusive results for the Tamil, English, and Tamil–English datasets. Tables 7, 8, and 9 provide the rank lists for Tamil, English, and Tamil–English, respectively. We ranked the teams using the average macro F1 score, which recognizes the F1 score in each label and determines their unweighted average. The runs were placed in decreasing order by the macro F1 scores. By fine-tuning a pre-trained language model, the ABLIMET team achieved the best results solely with the English dataset. For this English subtask, their pre-trained language model used the Roberta-base model. They offer RoBERTa as the best model

Table 8 Rank list for English language

Team	Acc	Pmac	Rmac	F1mac	Pw	Rw	F1w	Rank
Ablimet [74]	0.910	0.570	0.610	0.570	0.940	0.910	0.920	1
Sammaan [76]	0.940	0.520	0.470	0.490	0.930	0.940	0.940	2
Nozza [78]	0.950	0.580	0.450	0.480	0.940	0.950	0.940	3
hate-alert	0.940	0.510	0.450	0.470	0.920	0.940	0.930	4
LeaningTower	0.940	0.530	0.430	0.460	0.930	0.940	0.930	4
leaningtower	0.940	0.530	0.430	0.460	0.930	0.940	0.930	5
niksss	0.930	0.460	0.440	0.450	0.920	0.930	0.920	6
UMUTeam [73]	0.930	0.480	0.430	0.450	0.920	0.930	0.920	7
ARGUABLY	0.940	0.540	0.400	0.430	0.920	0.940	0.920	8
SOA_NLP	0.940	0.500	0.400	0.430	0.920	0.940	0.920	9
bitsa_nlp [75]	0.920	0.430	0.420	0.420	0.910	0.920	0.910	10
NAYEL [72]	0.940	0.510	0.370	0.390	0.910	0.940	0.910	11
SSNCSE_NLP [77]	0.930	0.480	0.370	0.390	0.910	0.930	0.910	12

Table 9 Rank list for Tamil–English dataset

Team	Acc	Pmac	Rmac	F1mac	Pw	Rw	F1w	Rank
ARGUABLY	0.890	0.630	0.600	0.610	0.890	0.890	0.890	1
UMUTeam [73]	0.850	0.540	0.670	0.580	0.900	0.850	0.870	2
bitsa_nlp [75]	0.880	0.610	0.560	0.580	0.890	0.880	0.880	3
hate-alert	0.830	0.540	0.630	0.560	0.890	0.830	0.850	4
SOA_NLP	0.900	0.650	0.500	0.540	0.890	0.900	0.890	5
Ablimet [74]	0.800	0.490	0.640	0.530	0.880	0.800	0.830	6
niksss	0.880	0.560	0.500	0.520	0.870	0.880	0.880	7
NAYEL [72]	0.900	0.620	0.470	0.510	0.870	0.900	0.880	8
SSNCSE_NLP [77]	0.890	0.660	0.430	0.470	0.870	0.890	0.870	9
Sammaan [76]	0.830	0.340	0.350	0.350	0.820	0.830	0.830	10
Ajetavya_Tamil–English	0.870	0.340	0.340	0.340	0.820	0.870	0.840	11

for this English dataset based on these models. Using the macro F1 score, we can determine that this transformer model performed well in comparison to other models. However, the team's performance in the Tamil and Tamil–English subtasks was quite poor. They ranked fifth in Tamil and sixth in Tamil–English because the accuracy of their models was inferior. As a consequence of performing data balancing in these activities for running the model, they achieved better outcomes than other teams. The ARGUABLY team did well in Tamil and Tamil–English classification challenges, utilizing machine and deep learning architectures. Other groups also fared better on this job, particularly those organized with a fine-tuning strategy, pre-trained models, and transformer models such as BERT [68], mBERT, XLM-RoBERTa [79], IndicBERT [80], HateBERT [81], etc. They include TF-IDF and count vectorizer, among others, for extracting features from datasets.

7 Conclusion

We propose a dataset that contains the high-quality, expert categorization of homophobic and transphobic content taken from comments posted in many languages on YouTube. In comparison to the numerous other annotated datasets utilized for various classifications, the one that was produced in this work offers very less information. Nevertheless, to the best of our knowledge, this is the first dataset that has been developed for the purpose of analyzing homophobia and transphobia in multilingual comments written in Tamil, English, and Tamil–English. Within the confines of a supervised classification framework, we carried out an exhaustive empirical investigation in which we evaluated a wide variety of feature selection approaches. These approaches included machine and deep learning methods. We also conducted a shared task to encourage research on homophobia and trans-

phobia detection systems. The results of our research showed that detecting homophobic and transphobic language in multilingual and multicultural contexts is a difficult challenge that needs to be tackled.

Acknowledgements The author was supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2 (Insight_2).

Funding Open Access funding provided by the IReL Consortium

Funding This research has not been funded by any company or organization.

Availability of data and material The datasets used in this paper were obtained from https://competitions.codalab.org/competitions/36394.

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest. The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals. The authors complied with ethical standards.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Weber, D., Nasim, M., Mitchell, L., Falzon, L.: Exploring the effect of streamed social media data variations on social network analysis. Soc. Netw. Anal. Min. **11**(1), 62 (2021)
2. Islam, M.M., Islam, M.M., Ahmed, F., Rumana, A.S.: Creative social media use for COVID-19 prevention in Bangladesh: a structural equation modeling approach. Soc. Netw. Anal. Min. **11**(1), 38 (2021)
3. Yenala, H., Jhanwar, A., Chinnakotla, M.K., Goyal, J.: Deep learning for detecting inappropriate content in text. Int. J. Data Sci. Anal. **6**(4), 273–286 (2018)
4. Nozza, D., Bianchi, F., Lauscher, A., Hovy, D.: Measuring harmful sentence completion in language models for LGBTQIA+ individuals. In: Proceedings of the 2nd Workshop on Language Technology for Equality, Diversity and Inclusion, pp. 26–34. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.ltedi-1.4. https://aclanthology.org/2022.ltedi-1.4
5. Tuna, T., Akbas, E., Aksoy, A., Canbaz, M.A., Karabiyik, U., Gonen, B., Aygun, R.: User characterization for online social networks. Soc. Netw. Anal. Min. **6**(1), 104 (2016)
6. O'Donohue, W., Caselles, C.E.: Homophobia: conceptual, definitional, and value issues. J. Psychopathol. Behav. Assess. **15**(3), 177–195 (1993)
7. Haaga, D.A.: Homophobia? J. Soc. Behav. Personal. **6**(1), 171 (1991)
8. Fyfe, B.: homophobia or homosexual bias reconsidered. Arch. Sex. Behav. **12**(6), 549–554 (1983)
9. Risch, J., Krestel, R.: Aggression identification using deep learning and data augmentation. In: Proceedings of the 1st Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), pp. 150–158. Association for Computational Linguistics, Santa Fe, New Mexico, USA (2018). https://www.aclweb.org/anthology/W18-4418
10. Fersini, E., Nozza, D., Boifava, G.: Profiling Italian misogynist: An empirical study. In: Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language, pp. 9–13. European Language Resources Association (ELRA), Marseille, France (2020). https://www.aclweb.org/anthology/2020.restup-1.3
11. Bhattacharya, S., Singh, S., Kumar, R., Bansal, A., Bhagat, A., Dawer, Y., Lahiri, B., Ojha, A.K.: Developing a multilingual annotated corpus of misogyny and aggression. In: Proceedings of the 2nd Workshop on Trolling, Aggression and Cyberbullying, pp. 158–168. European Language Resources Association (ELRA), Marseille, France (2020). https://www.aclweb.org/anthology/2020.trac-1.25
12. Waseem, Z.: Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In: Proceedings of the First Workshop on NLP and Computational Social Science, pp. 138–142. Association for Computational Linguistics, Austin, Texas (2016). https://doi.org/10.18653/v1/W16-5618. https://www.aclweb.org/anthology/W16-5618
13. Rao, T.S., Jacob, K.: The reversal on gay rights in India. Indian J. Psychiatry **56**(1), 1 (2014)
14. Chakrapani, V., Vijin, P.P., Logie, C.H., Newman, P.A., Shunmugam, M., Sivasubramanian, M., Samuel, M.: Understanding how sexual and gender minority stigmas influence depression among trans women and men who have sex with men in India. LGBT Health **4**(3), 217–226 (2017)
15. Kealy-Bateman, W.: The possible role of the psychiatrist: the lesbian, gay, bisexual, and transgender population in India. Indian J. psychiatry **60**(4), 489 (2018)
16. Kar, A.: Legal recognition and societal reaction on sexual minorities: reflections on moral policing and mental health of LGBT community in India. RESEARCH IN SOCIAL CHANGE p. 4 (2018)
17. Billies, M., Johnson, J., Murungi, K., Pugh, R.: Naming our reality: low-income LGBT people documenting violence, discrimination and assertions of justice. Fem. Psychol. **19**(3), 375–380 (2009)
18. Chauhan, V., Reddy-Best, K.L., Sagar, M., Sharma, A., Lamba, K.: Apparel consumption and embodied experiences of gay men and transgender women in India: variety and ambivalence, fit issues, LGBT-fashion brands, and affordability. J. Homosex. **68**(9), 1444–1470 (2021)
19. Garaigordobil, M.G., Larrain, E.L., Garaigordobil, M., Larrain, E.: Bullying and cyberbullying in LGBT adolescents: prevalence and effects on mental health. Comunicar. Media Edu. Res. J. **28**(1) (2020)
20. Mkhize, S., Nunlall, R., Gopal, N.: An examination of social media as a platform for cyber-violence against the LGBT+ population. Agenda **34**(1), 23–33 (2020)
21. Ybarra, M.L., Mitchell, K.J., Palmer, N.A., Reisner, S.L.: Online social support as a buffer against online and offline peer and sexual victimization among US LGBT and non-LGBT youth. Child Abuse Negl. **39**, 123–136 (2015). https://doi.org/10.1016/j.chiabu.2014.08.006
22. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Comput. Surv. (CSUR) **51**(4), 1–30 (2018)

23. Bashar, M.A., Nayak, R., Luong, K., Balasubramaniam, T.: Progressive domain adaptation for detecting hate speech on social media with small training set and its application to covid-19 concerned posts. Soc. Netw. Anal. Min. **11**(1), 1–18 (2021)
24. Miok, K., Škrlj, B., Zaharie, D., Robnik-Šikonja, M.: To ban or not to ban: Bayesian attention networks for reliable hate speech detection. Cognit. Comput. **14**(1), 353–371 (2022)
25. Gámez-Guadix, M., Incera, D.: Homophobia is online: sexual victimization and risks on the internet and mental health among bisexual, homosexual, pansexual, asexual, and queer adolescents. Comput. Hum. Behav. **119**, 106728 (2021)
26. Uyheng, J., Carley, K.M.: Characterizing network dynamics of online hate communities around the COVID-19 pandemic. Appl. Netw. Sci. **6**(1), 1–21 (2021)
27. Chard, A.N., Finneran, C., Sullivan, P.S., Stephenson, R.: Experiences of homophobia among gay and bisexual men: results from a cross-sectional study in seven countries. Cult. Health Sex. **17**(10), 1174–1189 (2015)
28. Awan, I., Zempi, I.: The affinity between online and offline anti-Muslim hate crime: dynamics and impacts. Aggress. Violent Behav. **27**, 1–8 (2016)
29. Marret, M.J., Choo, W.Y.: Factors associated with online victimisation among Malaysian adolescents who use social networking sites: a cross-sectional study. BMJ Open **7**(6), e014959 (2017)
30. Almerekhi, H., Kwak, H., Jansen, B.J., Salminen, J.: Detecting toxicity triggers in online discussions. In: Proceedings of the 30th ACM Conference on Hypertext and Social Media, HT '19, p. 291-292. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3342220.3344933
31. DePalma, R., Jennett, M.: Homophobia, transphobia and culture: Deconstructing heteronormativity in English primary schools. Intercult. Edu. **21**(1), 15–26 (2010)
32. Warriner, K., Nagoshi, C.T., Nagoshi, J.L.: Correlates of homophobia, transphobia, and internalized homophobia in gay or lesbian and heterosexual samples. J. homosex. **60**(9), 1297–1314 (2013)
33. Rasmussen, M.L., Sanjakdar, F., Allen, L., Quinlivan, K., Bromdal, A.: Homophobia, transphobia, young people and the question of responsibility. Discourse Stud. Cult. Polit. Edu. **38**(1), 30–42 (2017)
34. Tontodimamma, A., Nissi, E., Sarra, A., Fontanella, L.: Thirty years of research into hate speech: topics of interest and their evolution. Scientometrics **126**(1), 157–179 (2021)
35. Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., Patti, V.: Resources and benchmark corpora for hate speech detection: a systematic review. Lang. Resour. Eval. **55**(2), 477–523 (2021)
36. Pamungkas, E.W., Basile, V., Patti, V.: Towards multidomain and multilingual abusive language detection: a survey. Pers. Ubiquitous Comput. **27**(1), 17–43 (2023)
37. MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N., Frieder, O.: Hate speech detection: challenges and solutions. PLoS ONE **14**(8), e0221152 (2019)
38. Naseem, U., Razzak, I., Eklund, P.W.: A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter. Multimed. Tools Appl. **80**(28), 35239–35266 (2021). https://doi.org/10.1007/s11042-020-10082-6
39. Gao, L., Kuppersmith, A., Huang, R.: Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 774–782. Asian Federation of Natural Language Processing, Taipei, Taiwan (2017). https://www.aclweb.org/anthology/I17-1078
40. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Comput. Surv. (2018). https://doi.org/10.1145/3232676
41. Kim, Y., Park, S., Han, Y.S.: Generalizable implicit hate speech detection using contrastive learning. In: Proceedings of the 29th International Conference on Computational Linguistics, pp. 6667–6679. International Committee on Computational Linguistics, Gyeongju, Republic of Korea (2022). https://aclanthology.org/2022.coling-1.579
42. Arango, A., Pérez, J., Poblete, B.: Hate speech detection is not as easy as you may think: a closer look at model validation (extended version). Inf. Syst. **105**, 101584 (2022)
43. Ayo, F.E., Folorunso, O., Ibharalu, F.T., Osinuga, I.A.: Machine learning techniques for hate speech classification of twitter data: state-of-the-art, future challenges and research directions. Comput. Sci. Rev. **38**, 100311 (2020)
44. Burnap, P., Williams, M.L.: Us and them: identifying cyber hate on twitter across multiple protected characteristics. EPJ Data Sci. **5**(1), 11 (2016). https://doi.org/10.1140/epjds/s13688-016-0072-6
45. Davidson, T., Bhattacharya, D., Weber, I.: Racial bias in hate speech and abusive language detection datasets. In: Proceedings of the 3rd Workshop on Abusive Language Online, pp. 25–35. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/W19-3504. https://www.aclweb.org/anthology/W19-3504
46. Qian, J., Bethke, A., Liu, Y., Belding, E., Wang, W.Y.: A benchmark dataset for learning to intervene in online hate speech. 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), pp. 4755–4764. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1482. https://www.aclweb.org/anthology/D19-1482
47. Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., Nunes, S.: A hierarchically-labeled Portuguese hate speech dataset. In: Proceedings of the Third Workshop on Abusive Language Online, pp. 94–104. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/W19-3510. https://www.aclweb.org/anthology/W19-3510
48. Mulki, H., Haddad, H., Bechikh Ali, C., Alshabani, H.: L-HSAB: A Levantine Twitter dataset for hate speech and abusive language. In: Proceedings of the Third Workshop on Abusive Language Online, pp. 111–118. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/W19-3512. https://www.aclweb.org/anthology/W19-3512
49. Xu, J.M., Jun, K.S., Zhu, X., Bellmore, A.: Learning from bullying traces in social media. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 656–666. Association for Computational Linguistics, Montréal, Canada (2012). https://www.aclweb.org/anthology/N12-1084
50. Sigurbergsson, G.I., Derczynski, L.: Offensive language and hate speech detection for Danish. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 3498–3508. European Language Resources Association, Marseille, France (2020). https://www.aclweb.org/anthology/2020.lrec-1.430
51. Çöltekin, Ç.: A corpus of Turkish offensive language on social media. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 6174–6184. European Language Resources Association, Marseille, France (2020). https://www.aclweb.org/anthology/2020.lrec-1.758
52. Wu, H.H., Hsieh, S.K.: Exploring lavender tongue from social media texts[in Chinese]. In: Proceedings of the 29th Conference on Computational Linguistics and Speech Processing (ROCLING 2017), pp. 68–80. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Taipei, Taiwan (2017). https://www.aclweb.org/anthology/O17-1007

53. Ljubešić, N., Markov, I., Fišer, D., Daelemans, W.: 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, pp. 153–157. Association for Computational Linguistics, Barcelona, Spain (Online) (2020). https://www.aclweb.org/anthology/2020.peoples-1.15
54. Weinberger, L.E., Millham, J.: Attitudinal homophobia and support of traditional sex roles. J. Homosex. **4**(3), 237–246 (1979)
55. Smith, K.T.: Homophobia: a tentative personality profile. Psychol. Rep. **29**(3), 1091–1094 (1971)
56. MacDonald, A., Huggins, J., Young, S., Swanson, R.A.: Attitudes toward homosexuality: preservation of sex morality or the double standard? J. Consult. Clin. Psychol. **40**(1), 161 (1973)
57. Hill, D.B., Willoughby, B.L.: The development and validation of the genderism and transphobia scale. Sex Roles **53**(7), 531–544 (2005)
58. Bornstein, K., Bornstein, K.: Gender outlaw. Vintage Books New York (1994)
59. Nagoshi, C.T., Raven Cloud, J., Lindley, L.M., Nagoshi, J.L., Lothamer, L.J.: A test of the three-component model of gender-based prejudices: Homophobia and transphobia are affected by raters' and targets' assigned sex at birth. Sex Roles **80**(3), 137–146 (2019)
60. Worthen, M.G.: An argument for separate analyses of attitudes toward lesbian, gay, bisexual men, bisexual women, MtF and FtM transgender individuals. Sex Roles **68**(11), 703–723 (2013)
61. Worthen, M.G.: Hetero-cis-normativity and the gendering of transphobia. Int. J. Transgenderism **17**(1), 31–57 (2016)
62. Bandini, E., Maggi, M.: Transphobia. In: Emotional, Physical and Sexual Abuse, pp. 49–59. Springer (2014)
63. Ansara, Y.G., Friedman, E.J.: Transphobia. The Wiley Blackwell Encyclopedia of Gender and Sexuality Studies pp. 1–3 (2016)
64. Bettcher, T.M.: Transphobia. Transgender Stud. Q. **1**(1–2), 249–251 (2014)
65. Dasgupta, R.K.: Digital queer cultures in India: Politics, intimacies and belonging. Taylor & Francis, Milton Park (2017)
66. Chakravarthi, B.R., Muralidaran, V., Priyadharshini, R., McCrae, J.P.: Corpus creation for sentiment analysis in code-mixed Tamil-English text. In: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), pp. 202–210. European Language Resources association, Marseille, France (2020). https://www.aclweb.org/anthology/2020.sltu-1.28
67. Krippendorff, K.: Estimating the reliability, systematic error and random error of interval data. Edu. Psychol. Meas. **30**(1), 61–70 (1970). https://doi.org/10.1177/001316447003000105
68. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423. https://www.aclweb.org/anthology/N19-1423
69. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguis. **5**, 135–146 (2017)
70. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res **12**, 2825–2830 (2011)
71. Chollet, F., et al.: Keras. https://github.com/fchollet/keras (2015)
72. Ashraf, N., Taha, M., Abd Elfattah, A., Nayel, H.: NAYEL @LT-EDI-ACL2022: Homophobia/transphobia detection for equality, diversity, and inclusion using SVM. In: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pp. 287–290. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.ltedi-1.42. https://aclanthology.org/2022.ltedi-1.42
73. García-Díaz, J., Caparros-Laiz, C., Valencia-García, R.: UMUTeam@LT-EDI-ACL2022: Detecting homophobic and transphobic comments in Tamil. In: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pp. 140–144. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.ltedi-1.16. https://aclanthology.org/2022.ltedi-1.16
74. Maimaitituoheti, A.: ABLIMET @LT-EDI-ACL2022: A roberta based approach for homophobia/transphobia detection in social media. In: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pp. 155–160. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.ltedi-1.19. https://aclanthology.org/2022.ltedi-1.19
75. Bhandari, V., Goyal, P.: bitsa_nlp@LT-EDI-ACL2022: Leveraging pretrained language models for detecting homophobia and transphobia in social media comments. In: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pp. 149–154. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.ltedi-1.18. https://aclanthology.org/2022.ltedi-1.18
76. Upadhyay, I.S., Srivatsa, K.A., Mamidi, R.: Sammaan@LT-EDI-ACL2022: Ensembled transformers against homophobia and transphobia. In: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pp. 270–275. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.ltedi-1.39. https://aclanthology.org/2022.ltedi-1.39
77. Swaminathan, K., B, B., G L, G., Sampath, H.: SSNCSE_NLP@LT-EDI-ACL2022: Homophobia/transphobia detection in multiple languages using SVM classifiers and BERT-based transformers. In: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pp. 239–244. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.ltedi-1.34. https://aclanthology.org/2022.ltedi-1.34
78. Nozza, D.: Nozza@LT-EDI-ACL2022: Ensemble modeling for homophobia and transphobia detection. In: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pp. 258–264. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.ltedi-1.37. https://aclanthology.org/2022.ltedi-1.37
79. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8440–8451. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.747. https://www.aclweb.org/anthology/2020.acl-main.747
80. Kakwani, D., Kunchukuttan, A., Golla, S., N.C., G., Bhattacharyya, A., Khapra, M.M., Kumar, P.: IndicNLPSuite: monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 4948–4961. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.445. https://www.aclweb.org/anthology/2020.findings-emnlp.445

81. Caselli, T., Basile, V., Mitrović, J., Granitzer, M.: HateBERT: Retraining BERT for abusive language detection in English. In: Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), pp. 17–25. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.woah-1.3. https://aclanthology.org/2021.woah-1.3

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.