Transfer Learning Based Youtube Toxic Comments Identification

S. Santhiya(✉), P. Jayadharshini, and S. V. Kogilavani

Department of Artificial Intelligence, Kongu Engineering College, Perundurai 638060, India
jayadharshini.ai@kongu.edu

Abstract. Online users are negatively affected by the spread of offensive content on social media sites. A fear, dislike, unease, or distrust of lesbian, gay, bisexual, or transgender persons is known as homophobia or transphobia. Homophobic/transphobic speech, which can be summed up as bigotry directed towards LGBT+ people, has grown to be a significant problem in recent years. The major social problem of online homopho- bia/transphobia threatens to eliminate equity, diversification, and acceptance while also making online places toxic and unwelcoming for LGBT+ people. It is found to be sensitive subject and untrained crowd sourced annotators have trouble in identifying homophobia due to cultural and other preconceptions. As a result, annotators had been educated and provided them with thorough annotation standards. 15,141 multilingual annotated comments make up the dataset. The proposed work identifies the best Machine Learning Classifier with BERT embedding model for the Code-Mixed Dravidian Languages in order to identify the toxic languages directed towards LGBTQ+ individuals. Adaboost classifier outperforms other three classifiers in terms of accuracy.

Keywords: Dravidian languages · Code-Mixed Language · BERT · Mixed-Feelings · Machine Learning Classifiers

1 Introduction

In the digital age, social media is crucial for online communication because it enables users to publish content, share it with others, and voice their opinions whenever they want. NLP academics have access to a large amount of data that allows them to tackle more difficult, enduring issues like comprehending, analyzing, and tracking user actions toward particular topics or events. Additionally, the rapid development of deep learning-based NLP and the enormous volume of user-generated content that is readily available online, particularly on social media, offer reliable and effective methods to analyses users' behaviors. Such tactics can be employed for purposes like acquiring data for affective behavior research or sexism detection. Online, there are many unpleasant statements, including those that are sexist, homophobic, racist, and racial slurs, as well as threats and insults that are directed at particular people or organizations. The proliferation of

Supported by organization x.

Anand Kumar M et al. (Eds.): SPELLL 2022, CCIS 1802, pp. 220–230, 2023.
https://doi.org/10.1007/978-3-031-33231-9_15

online content has made it a serious issue for online communities. Online profanity has been noted as a global phenomenon that has spread over social media sites like Facebook, YouTube, and Twitter during the past ten years [1]. It is even more disturbing for vulnerable Lesbian, Gay, Bisexual, Transgender, and other (LGBT+) individuals. LGBT+ people endure violence, injustice, suffering, and sometimes even assassination because of what they love, how they appear, or who they are. The Internet has, however, given everyone the ability to significantly influence the lives of other people by utilizing some of its distinctive features, such as anonymity. Homophobic and transphobic content attacks the LGBT+ individuals frequently. LGBT+ individuals who seek support online experience abuse or assault, which has a serious negative impact on their mental health [2, 3]. An original study on the automatic detection of homophobic and transphobic content on social media for LGBT+ groups, particularly among Tamil people. English, Tamil, and code-mixed Tamil English are all included in the datasets. The Codemixed dataset includes symbols, tags, punctuation and symbols. Stop words and tag were used for preprocessing to clean the data. First work, consist of five classes. They are Mixed feeling, Neutral, Positive, Negative and unknown state. Second work consist of three classes namely homophobic, transphobic, and non-anti-LGBT+ content labels. Embedding technique namely BERT embedding has been used for both the proposed work. Models are built with BERT embedding using following Classifiers. Ada Boost classifier, Logistic Regression Classifier, K-Nearest Neighbor Classifier, Random Forest Classifier. Word embedding features from the BERT vectorized the text for both the task A and task B. The classifiers are used after vectorizing the text to build the models.

2 Related Work

The use of ict infrastructure, particularly social media, has altered how individuals connect with one another and communicate on a global scale as a result of the widespread usage of social media apps. As an illustration, the popular networking sites site YouTube allows users to build their own profiles, upload videos, and leave comments. Numerous individuals may view each video or comment due to "liking" and "sharing" strategies, offering cyberbullies a simple opportunity to disseminate offensive or unwelcome information about their victims. A Platform [4, 5] has been provided for antisocial behaviors like racism, sexism, homophobia, and transphobia. Later, Code-mixed datasets [6, 7] were in scarce in terms of quantity, size, and accessibility. Gender bias in NLP [6] has been actively mitigated for the English language using several techniques. The studies examined [8, 9] gender discrimination not only for English language and also for other language including French and other languages. One of the first experiments on Tamil abusive language recognition was carried out in 2020 by HASOC Dravidian CodeMix [10, 11]. A Tamil dataset of disgusting comments was created and supplied to the shared task's participants afterwards, as reported by Dravidian LangTech [12]. Social media activity in local languages with mixed codes has dramatically expanded over the past several years as a result of cheaper internet and more people using smartphones. A significant amount of these exchanges are contributed by the 215 million speakers of Dravidian languages (4, many of whom are multilingual with English because it is India's national language). The examination of code-mixed text in Dravidian languages is hence becoming more and more necessary. The majority of current research on offensive language detection and sentiment analysis has been done on social media platforms using high resource languages. Empathy and offensiveness can be predicted by models that have been trained on such rich monolingual data. However, because bilingual people use social media more regularly, a system trained some under code mixed data is necessary.

3 Proposed System

Dataset Description. The dataset gathered by the organizers consists of 15,141 YouTube comments from different languages that have been categorized as homophobic, transphobic, or non-anti-LGBT+ content. In a multilingual culture, code-mixing is a common occurrence, and the writings that result from it are occasionally written in scripts that are not native to the speaker's language [23]. Systems trained on monolingual data struggle with code-mixed material because it is challenging to switch codes at lexical and syntactic levels in the text. The Common Task uses text that is code-mixed in Dravidian languages - A introduces a fresh corpus of unmatched quality for sentiment analysis (Tamil, English, Tamil- English). Task - B, which is shared, addresses homophobia and transphobia. The goal of detection is to isolate non-anti-LGBT+ content and homophobic, transphobic, and other offensive language from the corpus [24]. The destructive rhetoric used against LGBTQ+ people is known as "hate speech," and it includes both homophobia and transphobia.

Homophobic Language. It as a specific type of gender-based harassment statement that includes the practice of derogatory terms such as "fag," "homo," or "that's so gay" in reference to anyone who identify as gay, lesbian, bisexual, queer, or gender nonconforming [13, 14]. A posture of hatred against homosexuals, male or female is the most popular definition of homophobia [15]. Lesbophobia, gayphobia, and biphobia are three families of phobias that target various target groups. However, there is a distinction between general and specific homophobia.

Transphobic Language. Although there are minute differences, the reader may wonder why homophobia and transphobia shouldn't be included together. Contrary to popular belief, transphobia and homophobia [16] are not the same. A person who was given the gender of a woman at birth but now identifies as a man is an example of a heterosexual person. Nowadays, a lot of transgender persons refer to their gender identity in the present tense rather than their gender at birth by using the terminology of sexual orientation. Teenage transgender people face much greater marginalization and lack access to resources than their LGB counterparts in a number of different nations throughout the world. Given that numerous laws intended to protect LGB people do not include protections related to gender uniqueness or appearance [17, 18]. People who are transphobic may or may not also be homophobic. They may be homosexual or straight. In India, transgender persons have a constitutional basis because of mythology and their affiliations to Hindu gods, they must be accorded particular treatment. LGBQ individuals, however, are unable to be married in India. In Tamil Nadu, the word "homophobia" is forbidden than their LGB colleagues in some global regions. People who are transphobic may or may not also be homophobic. They may be homosexual or straight. LGBQ individuals, however, are unable to be married in India. In Tamil Nadu, the word “homophobia” is forbidden.

Non-LGBT+hating Material. Information that is not anti-LGBT+ can be divided into three categories, all of which are crucial for the study of homophobia and transphobia. The toxic online disinhibition and a lack of empathy [3] are linked to homophobic/transphobic cyberbullying. Second, by examining non-anti-LGBT+ remarks, Development of preventative and interventionary programs aid in changing the online behaviors and opinions of social media users.

3.1 Models

Word Embedding –BERT Model. A method for minimizing the amount of elements in the input is feature selection. Variety of feature extraction approaches are employed with word embedding such as BERT in the proposed work. BERT uses masked language models to enable pretrained deep bidirectional interpretations.

BERT embedding is word embedding that generates vectors depend on both the sentence’s context and the word’s meaning [19]. The universal language model BERT generates a summarised word vectors at the inter- and intra level. By using bidirectional self-attention transformers, BERT, as opposed to static, non-contextualized word embedding, captures both short- and long-span contextual dependency in the input text. The [CLS] token and [SEP] token are concatenated at the beginning and end of the sentence after it has been initially tokenized in the BERT embedding, respectively. Then, each token has a 768-byte embedding created for it (Figs. 1 and 2).

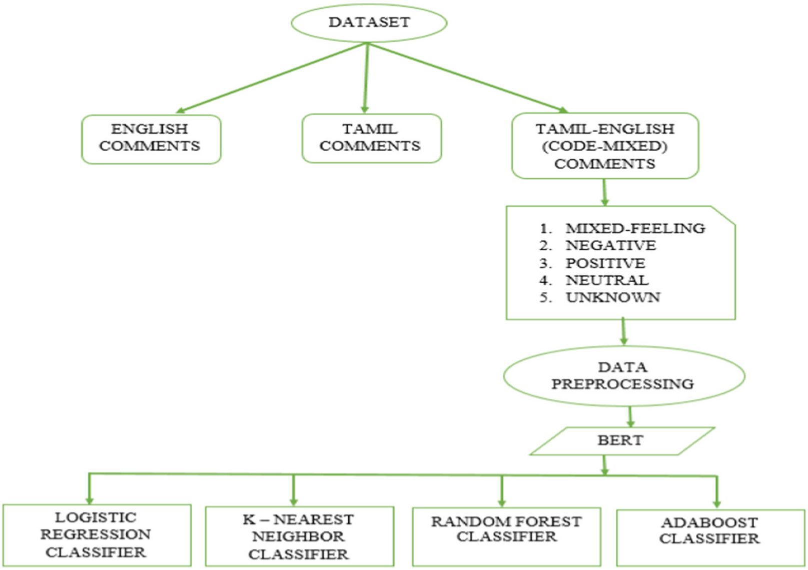

Fig. 1. Task A workflow

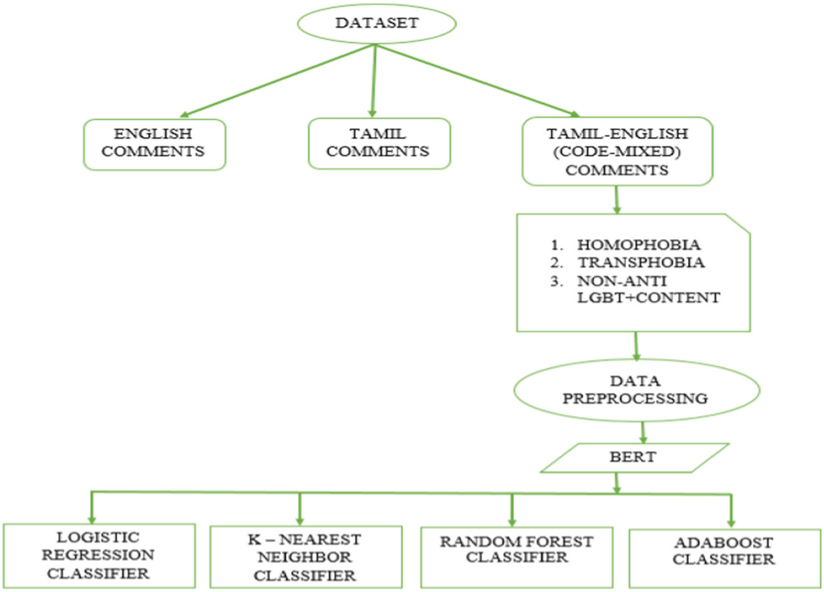

Fig. 2. Task B workflow

3.2 Classifiers

Logistic Regression One of the core machine learning algorithms is LR, a probabilistic classifier used for the purpose of classifying data. In essence, it is the logistic function-based transformed a linear regression style. In order to determine the class probability, it first accepts legitimate data as input, increases each by a load, and then delivers the generated summation to the nonlinear function, also known as the logistic function [20]. To predict the result, the classifier uses linear combinations of input. The logistic function is used to predict the likelihood of the specific class. The outcome of a logistic regression depends on the input and the associated system. Given that neural pathways can be seen as a collection of several LR classifications, logistic regression and neural networks have a tight link. As opposed to the operational classifier Nave Bayes, LR is a computational method. LR is clearly more resistant to correlated features [21] but Nave Bayes keeps strict restrictions on conditional independence. It indicates that the weight W will be distributed among the features as W1, W2, and W3, respectively, when there are numerous features, such as F1, F2, and F3, that are positively correlated.

Random Forest. As an ensemble classifier, random forest uses bootstraps, which are samples drawn at random from the training set, to construct its predictions. Bootstraps are a group of different decision trees that have all been trained using training datasets that are the same size as the training set. Following the construction of a tree, a set of bootstraps is used as the test set, which omits any particular record from the baseline dataset (out-of-bag (OOB) samples). The classification error rate across all test sets is the OOB estimate of the predictive performance. RF showed substantial advantages over other systems in managing extremely nonlinearly connected data, noise tolerance, adjustment easiness, and the capability for efficient massively parallel processing. Another crucial element that RF offers is an intrinsic feature selection step that is used before the classification task in order to condense the space of variables by assigning a value to each feature's relevance. In comparison to other machine learning algorithms, for tree structures, tree pairing, identity, and comment, RF follows specific guidelines. Additionally, it is resistant to overloading and is considered to be more stable when outliers are present and in very high-dimensional parameter spaces. With the same properties as DT, RF model has been assessed.

K-Nearest Neighbor. The classification task is where KNN is most frequently utilized, while it can also be used for regression problems. The KNN algorithm maintains all available data and categorizes new data points based on similarities. It suggests that as fresh data arrives; it can be easily categorized using the KNN algorithm into a suitable category. The KNN approach places the new case in the category that most closely resembles the categories that are currently accessible because it expects that the new incoming data will be linked to the existing examples. KNN is a quasi method [22]. Since it makes no assumptions about the underlying data, Because it saves the dataset and performs an operation on it when classifying data, this method is frequently referred to as a sloppy learner's algorithm rather than automatically recognizing from the training set. KNN method only retains the dataset during training; as new data is encountered, it sorts it into groups that are fairly similar to the present dataset. Applying consistent weights, KNN has been employed for classification with 3, 4, and 9 neighbors.

AdaBoost Classifier. AdaBoost derives its feature importance from the feature importance that its base classifier provides. The average feature importance provided by each Decision Tree, assuming you utilize one as your base classifier, determines how important a feature is to AdaBoost. It is very similar to the widespread method of assessing feature relevance by looking at a forest of trees. It takes advantage of the fact that a bigger proportion of input samples generate final predictions as a result of features identified at the top of the tree, and expected fraction may be used to calculate the relative relevance of a feature. Adaptive Boosting, often known as AdaBoost algorithm, is a Boosting method used as an Ensemble Method in Machine Learning. It is known as adaptive boosting because each instance receives a new set of weights, with higher weights given to examples that were mistakenly categorized. As the input parameters are not jointly optimized, Adaboost is less susceptible to overfitting. Adaboost can be used to increase the accuracy of weak classifiers. Adaboost is now used to categories text and graphics instead of binary classification issues.

4 Performance Evaluation

Precision, recall, F1-score, Support and Accuracy results from tests using the BERT Embedding model which displayed are machine learning classifiers for sentiment classification and identifying inappropriate language. The table describes the metrics (precision, recall, and F1-score) are annually calculated for each class, then combined using a macro-average. As a result, the statistic does not take into consideration the property of misclassification and treats all classes equally. A weighted sum employs metrics from each class, like a macro average does, but its contribution to the average is weighted based on how many examples are available for that class (Tables 1, 2, 3, 4, 5 and Figs. 3, 4).

Table 1. BERT Embedding using Logistic Regression Classifier

Class	Precision	Recall	F1 Score	Support
Mixed - Emotions	0.25	0.04	0.07	52
Negative	0.55	0.51	0.53	131
Positive	0.65	0.80	0.72	321
Neutral	0.61	0.67	0.64	110
Unknown	0.47	0.26	0.34	69
Mac	0.57	0.46	0.46	691
Weighted	0.58	0.61	0.58	

Table 2. BERT Embedding using K-Nearest Neighbor Classifier

Class	Precision	Recall	F1 Score	Support
Mixed - Emotions	0.32	0.13	0.19	52
Negative	0.58	0.52	0.55	139
Positive	0.66	0.86	0.72	321
Neutral	0.61	0.65	0.63	110
Unknown	0.56	0.32	0.41	69
Mac	0.55	0.48	0.50	691
Weighted	0.60	0.62	0.60	

Table 3. BERT Embedding using Random Forest Classifier

Class	Precision	Recall	F1 Score	Support
Mixed - Emotions	0.25	0.12	0.16	52
Negative	0.48	0.42	0.44	139
Positive	0.60	0.78	0.68	321
Neutral	0.58	0.51	0.54	110

(*continued*)

Table 3. (*continued*)

Class	Precision	Recall	F1 Score	Support
Unknown	0.43	0.22	0.29	69
Mac	0.47	0.41	0.42	691
Weighted	0.53	0.56	0.53	

Table 4. BERT Embedding using AdaBoost Classifier

Class	Precision	Recall	F1 Score	Support
Mixed - Emotions	0.22	0.10	0.13	52
Negative	0.58	0.47	0.52	139
Positive	0.65	0.78	0.71	321
Neutral	0.58	0.56	0.57	110
Unknown	0.38	0.33	0.35	69
Mac	0.48	0.45	0.46	691
Weighted	0.57	0.59	0.57	

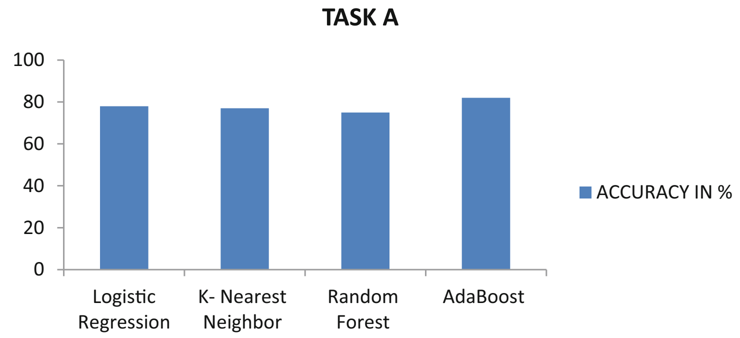

Fig. 3. Evaluation of Different Classifier with BERT Embedding for Task A

Table. 5. BERT Embedding using Different Classifiers for 3 Class Dataset

Classifier	Pmac	Rmac	F1mac	Pw	Rw	F1w
Logistic Regression	0.61	0.36	0.38	0.84	0.89	0.85
K – Nearest Neighbor	0.49	0.52	0.50	0.85	0.83	0.84

(*continued*)

Table. 5. (*continued*)

Classifier	Pmac	Rmac	F1mac	Pw	Rw	F1w
Random Forest	0.49	0.60	0.52	0.87	0.81	0.83
AdaBoost	0.62	0.38	0.39	0.85	0.88	0.85

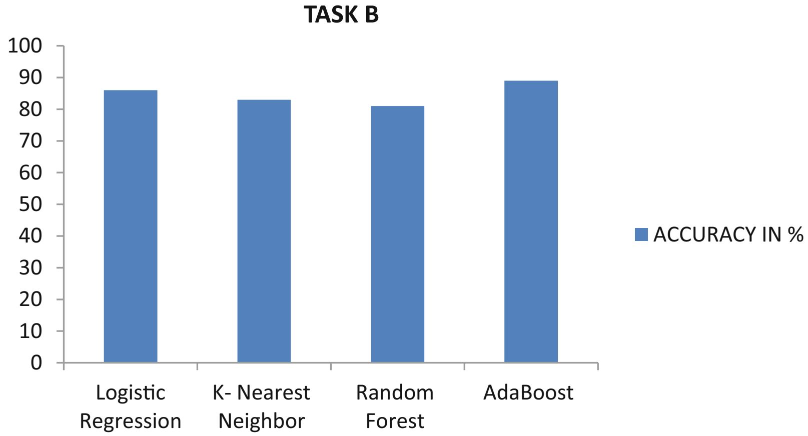

Fig. 4. Evaluation of Different Classifier with BERT Embedding for Task B

5 Conclusion

The work includes a dataset with elevated, trained evaluation of homophobic and transphobic content from linguistic YouTube comments. Many efforts to detect umbrella hate speech have not focused on detecting homophobia and transphobia. In a dataset of English and Tamil-language YouTube comments, Efficiency of big transformer-based pre-trained models has been tested to identify homophobia and transphobia. The final dataset is trivial in comparison to other labelled data used for other classification. The findings of the experiments showed that multilingual BERT, which had not previously been exposed to code mixing, excelled in both language challenges and the code-mixed test. From the above analysis, AdaBoost Classifier gives the better accuracy in place of detecting the toxic languages which is offensive towards Homophobia and Transphobia individuals using BERT Model. Future options for the work include building the dataset for more Dravidian languages. By crawling and annotating more social media data sets, Tamil dataset have to expanded significantly. To enhance the performance of the classifiers, it is intended to investigate semi-supervised and incremental methods. Additionally, it is decided to investigate the relationship between sarcasm and anti-LGBT+ comments as our manual analysis revealed that many of these comments are ironic.

References

1. Gao, Z., Yada, S., Wakamiya, S., Aramaki, E.: Offensive language detection on video live streaming chat. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 1936–1940, Barcelona, Spain (Online). International Committee on Computational Linguistics (2020)
2. McConnell, E.A., Clifford, A., Korpak, A.K., Phillips, G., Birkett, M.: Identity, victimization, and support: Facebook experiences and mental health among lgbtq youth. Comput. Hum. Behav. **76**, 237–244 (2017). https://doi.org/10.10007/1234567890
3. Wright, M.F., Wachs, S.: Does empathy and toxic online disinhibition moderate the longitudinal association between witnessing and perpetrating homophobic cyberbullying? Int. J. Bully. Prevent. **3**(1) (2021)
4. Diefendorf, S., Bridges, T.: On the enduring relationship between masculinity and homophobia. Sexualities **23**(7), 1264–1284 (2020)
5. Larimore, S., Kennedy, I., Haskett, B., Arseniev Koehler, A.: Reconsidering annotator disagreement about racist language: Noise or signal? In: Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, pp. 81–90, Online. Association for Computational Linguistics (2021)
6. Ranjan, P., Raja, B., Priyadharshini, R., Balabantaray, R.C.: A comparative study on code-mixed data of Indian social media vs formal text. In: 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), pp. 608–611 (2016). https://doi.org/10.1109/IC3I.2016.7918035
7. Jose, N., Chakravarthi, B.R., Suryawanshi, S., Sherly, E., McCrae, J.P.A.: survey of current datasets for code-switching research. In: 2020 6th International Conference on Advanced Computing Communication Systems (ICACCS) (2020)
8. Vanmassenhove, E., Hardmeier, C., Way, A.: Getting gender right in neural machine translation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3003–3008. Brussels, Belgium. Association for Computational Linguistics (2018)
9. Prates, M.O.R., Avelar, P.H., Lamb, L.C.: Assessing gender bias in machine translation: a case study with google translate. Neural Comput. Appl. **32**(10), 6363–6381 (2020)
10. Chakravarthi, B.R.: HopeEDI: a multilingual hope speech detection dataset for equality, diversity, and inclusion. In: Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media, pp.41–53. Barcelona, Spain (Online). Association for Computational Linguistics (2020)
11. Mandl, T., Modha, S., Kumar, M.A., Chakravarthi, B.R.: Overview of the HASOC Track at FIRE 2020: hate speech and offensive language identification in Tamil, Malayalam, Hindi, English and German. In: Forum for Information Retrieval Evaluation pp. 29–32. Association for Computing Machinery, New York, NY, USA, FIRE 2020 (2020). https://doi.org/10.1145/3441501.3441517
12. Chakravarthi, B.R., et al.: Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pp. 133–145. Kyiv. Association for Computational Linguistics (2021)
13. Meyer, E.J.: Gendered harassment in secondary schools: understanding teachers'(non) I nterventions. Gender Educ. **20**(6), 555–570
14. Poteat, V.P., Rivers, I.: The use of homophobic language across bullying roles during adolescence. J. Appl. Develop. Psychol. **31**(2), 166–172 (2008)
15. Fra¨ıss´e, C., Barrientos, J.: The concept of homophobia: a psychosocial perspective. Sexologies **25**(4), e65–e69 (2016)

16. Graham, R., Berkowitz, B., Blum, R., Bockting, W., Bradford, J., de Vries, B., Makadon, H.: The health of lesbian, gay, bisexual, and transgender people: building a foundation for better understanding, vol. 10, p. 13128. Institute of Medicine, Washington, DC (2011)
17. McGuire, J.K., Anderson, C.R., Toomey, R.B., Russell, S.T.: School climate for transgender youth: a mixed method investigation of student experiences and school responses. J. Youth Adolesc. **39**(10), 1175–1188 (2010)
18. Hatchel, T., Valido, A., De Pedro, K.T., Huang, Y., Espelage, D.L.: Minority stress among transgender adolescents: the role of peer victimization, school belonging, and ethnicity. J. Child Fam. Stud. **28**(9), 2467–2476 (2019)
19. 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. Minneapolis, Minnesota. Association for Computational Linguistics (2019)
20. Shah, K., Patel, H., Sanghvi, D., Shah, M.: A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augment. Hum. Res. **5**(1), 1–16 (2020)
21. Jin, S., Pedersen, T.: Duluth UROP at SemEval-2018 task 2: Multilingual emoji prediction with ensemble learning and oversampling. In: Proceedings of The 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, pp. 482–485 (2018). https://doi.org/10.18653/v1/S18-1077, https://www.aclweb.org/anthology/S18-1077
22. Nongmeikapam, K., Kumar, W., Singh, M.P.: Exploring an efficient handwritten Manipuri meetei-mayek character recognition using gradient feature extractor and cosine distance based multiclass k-nearest neighbor classifier. In: Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017), pp. 328–337. NLP Association of India, Kolkata, India (2017). https://www.aclweb.org/anthology/W17-7541
23. Chakravarthi, B.R., et al.: Dravidiancodemix: Sentiment analysis and offensive language identification dataset for Dravidian languages in code-mixed text. Lang. Resources Evalu. 1–42 (2022)
24. Stakic, I.: Homophobia and hate speech in Serbian public discourse: how nationalist myths and stereotypes influence prejudices against the LGBT minority. Master's thesis, Universitetet (2011)