dkit@LT-EDI-2024: Detecting Homophobia and Transphobia in English Social Media Comments

Sargam Yadav, Abhishek Kaushik and **Kevin McDaid**
d00263026@student.dkit.ie, abhishek.kaushik@dkit.ie and kevin.mcdaid@dkit.ie
Dundalk Institute of Technology,
Dublin Rd, Marshes Upper
Dundalk, Co. Louth, A91 K584

Abstract

Machine learning and deep learning models have shown great potential in detecting hate speech from social media posts. This study focuses on the homophobia and transphobia detection task of LT-EDI-2024 in English. Several machine learning models, a Deep Neural Network (DNN), and the Bidirectional Encoder Representations from Transformers (BERT) model have been trained on the provided dataset using different feature vectorization techniques. We secured top rank with the best macro-F1 score of 0.4963, which was achieved by fine-tuning the BERT model on the English test set.

1 Introduction

The increase in popularity of social media has fostered hate speech in online discourse Paz et al. (2020) Fortuna and Nunes (2018). Social media posts produce a great volume of data which can be hard to moderate manually. Artificial Intelligence tools have proven to be useful in combating trolling Cheng et al. (2017), misinformation, cyberbullying Moreno et al. (2019), etc. Specifically, Large Language Models (LLMs) such as BERT, Cross-Lingual RoBERTa (XLM-RoBERTa) Conneau et al. (2019), and Multilingual Representations for Indian Languages (MuRIL) Khanuja et al. (2021) have been used in recent studies to counter different types of hate speech Mozafari et al. (2020a,b); Kumaresan et al. (2023). The Lesbian, Gay, Bisexual, and Transgender (LGBT+) community has been a prominent target for online hate speech in the past Hinduja and Patchin (2020). Homophobia is the expression of hate and negative attitudes towards people who identify as homosexuals. Transphobia is the expression of negative beliefs towards people who identify as transgenders. It is imperative to filter such toxic and abusive language towards the LGBT+ community, as it can be the cause of severe psychological distress, and can silence their online voices. Very few datasets are available online for homophobia and transphobia detection in code-mixed languages such as Malayalam and Hindi Kumaresan et al. (2023) Chakravarthi et al. (2023). In recent years, shared tasks have been conducted to promote research for different types of hate speech such as misogyny (Automatic Misogyny Identification) Fersini et al. (2020), hate speech in low-resource languages (Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages) Mandl et al. (2021), and code-mixed languages Satapara et al. (2021).

In this study, we focus on our participation in the LT-EDI-2024 shared task, which was was the detection of homophobia and transphobia from social media comments[1]. We have selected English data set for the task Kumaresan et al. (2024). The provided datasets were converted into feature vectors using techniques such as Term Frequency - Inverse Document Frequency (TF-IDF), count vectorizer, Word2Vec. Machine learning and deep learning models were then trained and evaluated on the datasets using empirical metrics such as accuracy, macro-F1 score, etc. The rest of the article is structured as follows: Section 2 discusses the relevant literature in sentiment analysis and hate speech detection. Section 3 provides details of the dataset, and the steps involved in the experiment such as feature vectorization, model training, fine-tuning, and evaluation. Section 4 discusses the results and findings of the study, and Section 5 concludes the study.

2 Related Works

In this section, we will discuss the relevant literature and previous work conducted in sentiment analysis and hate speech detection.

[1] https://codalab.lisn.upsaclay.fr/competitions/16056

Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 271–276
March 21, 2024

2.1 Sentiment Analysis

Natural Language Processing (NLP) tools have been extensively utilised to perform sentiment analysis on datasets in English and other languages Shah and Kaushik (2019); Shah et al. (2020); Kazhuparambil and Kaushik (2020a,b). Code-mixed languages present several challenges due to factors such as inconsistent spelling, lack of grammatical rules, and more Mathur et al. (2018). A novel dataset in code-mixed Hinglish was introduced by Kaur et al. (2019), who performed sentiment analysis on comments about cookery channels using machine learning models. Additionally, deep learning approaches such as multi-layer perceptron Donthula and Kaushik (2019) and Transformer-based models were also explored Yadav et al. (2021); Yadav and Kaushik (2022).

2.2 Hate Speech Detection

NLP models have seen significant success in hate speech detection Yadav et al. (2023a); Kumar et al. (2018); Yadav et al. (2023b); Chinnaudayar Navaneethakrishnan et al. (2022). Forum for Information Retrieval Evaluation (FIRE) 2022 organized task A for detecting sentiment analysis and task B for detecting homophobia Chinnaudayar Navaneethakrishnan et al. (2022). The highest accuracy of 93% and 91% was achieved by using XLM-RoBERTa and BERT respectively Manikandan et al. (2022). Authors Kumaresan et al. (2023) presented a novel dataset of YouTube comments for homophobia and transphobia in the following languages: Malayalam, Hindi, Tamil, English, and code-mixed Tamil and English. Chakravarthi (2023) introduce a dataset for homophobia and transphobia detection in English, Tamil and code-mixed Tamil and English. Another study Chakravarthi et al. (2022) expands on the baseline in Chakravarthi (2023) by evaluating the performance of multilingual language models. The second shared task on Homophobia and Transphobia Detection in Social Media Comments (LT-EDI@RANLP-2023) Chakravarthi et al. (2023) was conducted in the following 5 languages: English, Spanish, Tamil, Hindi, and Malayalam. For task A in Malayalam, Spanish, and Tamil, the best weighted F1 score achieved was 0.9976, 0.8883, and 0.9496 respectively, using a weight-space ensembling technique Ninalga (2023). A multilingual model was trained on the complete dataset consisting of all languages, and individual models were fine-tuned for each language. Linear interpolation was then performed between the weights of the fine-tuned and multilingual models. For task B in Malayalam, the best score of 0.8842 was achieved using a custom XLM-RoBERTa model, which was pre-trained with a random sample of 50,000 tweets. For Hindi, Malayalam, and Tamil, one-fourth of the tweets were Romanized to accommodate code-mixing Wong et al. (2023). The literature review suggests that machine learning and deep learning models should be further studied to develop efficient systems for detecting different aspects of hate speech, such as homophobia and transphobia.

Category	Train	Test	Dev
Non-anti-LGBT+	2,978	748	931
Homophobia	179	42	55
Transphobia	7	2	4

Table 1: Class distribution of the English datasets

3 Methodology

In the section, the methodology used in the task is discussed.

3.1 Task and Dataset Description

The Homophobia/Transphobia Detection in social media comments shared task at LT-EDI@EACL-2024 was available in several languages such as Hindi, Tamil, Telugu, Kannada, Gujarathi, Malayalam, Marathi and Tulu. The training dataset for English consisted of a total of 3,164 samples which was divided into the following three classes: 'Non-anti-LGBT+ content', 'Homophobia', and 'Transphobia'. The development set consists of a total of 792 samples, and the test set of 990 samples. Table 1 displays the class distribution of all the sets. Figure 1 displays an example comment from each class in the dataset. In Phase-1 of the study, the training and development sets were released. In Phase-2 of the study, the test comments were released and predictions on these comments were submitted to the shared task organisers for evaluation. Later on, the test set with labels was released so that the performance of all the models could be evaluated.

3.2 Experiment

In this subsection, we will discuss the steps involved in preprocessing the data, feature extraction, and model training. The free version of Google Co-

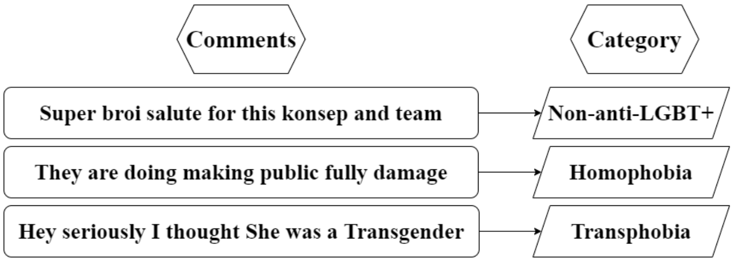

Figure 1: Example comments from each class for English

lab with GPU was used for experimentation. For all models, the training (TS) and development sets (DS) were combined into the merged training set (MTS), and finally split as a stratified sample into training and validation sets consisting of 70% and 30% of the data respectively. Stratified 10-fold cross validation (CV) and parameter tuning was performed using GridSearch CV on the 70% training set for machine learning models to find best parameters. The best models with optimal parameters are selected based on the macro-F1 score obtained by evaluating on the 30% validation set. Finally, the best model with optimal parameters trained on the 70% training set is used to evaluate model performance on the unseen test set (UTS). The results of top two models for each vectorization technique have been recorded. Figure 2 depicts the various steps of the experiment proposed in this study.

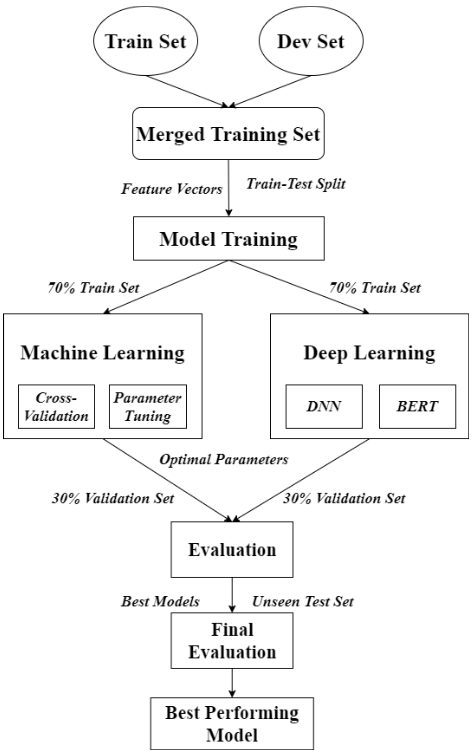

Figure 2: Flowchart of Experimental Methodology

3.2.1 Feature Engineering

For training the machine learning models, data cleaning and pre-processing was performed by removing all non-ASCII characters, user handles, hyperlinks, punctuation, extra whitespaces, stopwords, and newlines. The pre-processing steps were handled by the Natural Language Toolkit (NLTK) library[2]. The 'category' columns for all the sets were converted into numeric labels using Label Encoder. The following feature vectorization techniques were tested for machine learning models: TF-IDF, count vectorizer, and Word2Vec. For TF-IDF and count vectorizer, the maximum number of features has been limited to 2000. A custom Word2Vec model was trained on the merged training set with a vector size of 300, window of 10, and the skip-gram architecture McCormick (2016). The Word2Vec model was trained using the Gensim library[3]. Min-Max scaling to a feature range of 0 to 1 was performed on the Word2Vec embeddings to remove any negative values in the training data.

3.2.2 Machine Learning

The following machine learning models were trained on the resulting vectors: Logistic Regression (LR), Naive Bayes Bernoulli (NB-B), Gaussian (NB-G), and Multinomial (NB-M), Support Vector Machine Linear (SVM-L) and Radial Basis Function (SVM-R), Decision Trees (DT), and Random Forests (RF). For LR, SVM-L, and SVM-R, the value of the parameter C ranges from 10^{-3} to 10^{+3}. For LR, the lbfgs and liblinear solvers are considered. For NB-B and NB-M, the value of α considered is in the range of 10^{-3} to 10^{+3}. For SVM-R, a range of 10^{-3} to 10^{+3} is considered for the value of γ. For DT and RF, gini and entropy are considered as criterion. For DT, the maximum depth of the nodes is considered in the range of 40

[2]https://www.nltk.org/
[3]https://radimrehurek.com/gensim/models/word2vec.html

Model	Acc	Macro-F1	Prec	Rec
BERT	**0.9556**	**0.4963**	**0.5794**	**0.4585**
TF-IDF + DNN	0.9202	0.4295	0.4302	0.4288
TF-IDF + RF	0.9282	0.3482	0.3696	0.3461
TF-IDF + DT	0.8939	0.3551	0.3535	0.3567
Count Vec + DT	0.8797	0.3731	0.3667	0.3859
Count Vec + RF	0.8888	0.3707	0.3661	0.3778
Word2Vec + NB-G	0.6975	0.3428	0.3679	0.4797
Word2Vec + NB-M	0.9418	0.3233	0.3139	0.3333

Table 2: Top Model Results on the Unseen Test Set

to 60. For RF, the no. of estimators are considered in a range of 10 to 100 in steps of 10. For NB-G, var smoothing has been applied.

3.2.3 Deep Neural Network

The DNN model has been trained and evaluated using Tensorflow[4]. TF-IDF vectors have been used to train a DNN consisting of seven layers. The dense input layer has 128 neurons, 'relu' activation, and is followed by a dropout layer (dropout = 0.2). Next is another dense layer with 64 neurons and 'relu' activation, followed by a dropout layer (dropout = 0.2). This is followed by another dense layer with 32 neurons and 'relu' activation, followed by a dropout layer (dropout = 0.2). The final layer is a dense layer with 'softmax' activation to predict the classes. The Adam optimiser with a learning rate of 0.001 is used for optimization. The sparse categorical cross-entropy loss is used while training. The model is then trained for 15 epochs. A class weight dictionary has been calculated and used while training to account for the class imbalance.

3.2.4 Transformer-based models

The BERT (bert-base-uncased)[5] Devlin et al. (2018) English model consists of 12 layers and 110M parameters. It was fine-tuned using HuggingFace[6] and Pytorch[7]. All the comments have been encoded using a BERT tokenizer with the maximum sequence length of 128. The encodings have been converted into TensorDataset and batched using data loader. The hyperparamters used are as followings: number of epochs = 3, learning rate = 3e-5, and training and evaluation batch size = 2. The fine-tuned model was then evaluated on the 30% validation set and finally used to make predictions on the unseen test set.

[4]https://www.tensorflow.org/

[5]https://huggingface.co/bert-base-uncased

[6]https://huggingface.co/

[7]https://pytorch.org/

4 Results and Analysis

In this section, we will discuss the results of the experiment and analyse the findings. Table 2 displays the performance of the best two models for each vectorization technique based on the following evaluation criteria: Accuracy (Acc), Macro-F1, Precision (Prec), and Recall (Rec). The highest macro-F1 score of 0.4963 is achieved by the BERT model, followed by 0.4295 achieved by the DNN + TF-IDF model. Out of the machine learning models, DT performs the best with count vectorizer, achieving a macro-F1 score of 0.3731. Thus, the BERT model can be considered as the best model for homophobia and transphobia detection on this English dataset. The model has been made available[8]

5 Conclusion

In this study, homophobia and transphobia detection in English is conducted using different machine learning models, a DNN, and BERT. The highest macro-F1 score achieved is 0.4963 using the BERT model through simple fine-tuning. Transformer-based models have outperformed traditional machine learning models in this task of homophobia and transphobia detection. Further exploration can be carried out for online inclusivity through experimentation on different datasets and more complex model architectures.

Limitations

The dataset consists of YouTube comments in informal English. Informal English on social media platforms does not follow the linguistic rules of proper English. NLP models and tools have been pre-trained on internet sources written in formal English. Additionally, there is a scarcity of datasets that focus on homophobia and transphobia detection.

[8]https://huggingface.co/sam34738/BERT_homo

Ethics Statement

The annotated dataset used in this study has been taken from the LT-EDI@EACL 2024 shared task. The authors did not re-annotate the data, and only performed feature vectorization and model training. We respect all communities mentioned in the study.

References

Bharathi Raja Chakravarthi. 2023. Detection of homophobia and transphobia in YouTube comments. *International Journal of Data Science and Analytics*, pages 1–20.

Bharathi Raja Chakravarthi, Adeep Hande, Rahul Ponnusamy, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022. How can we detect Homophobia and Transphobia? Experiments in a multilingual code-mixed setting for social media governance. *International Journal of Information Management Data Insights*, 2(2):100119.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Malliga Subramanian, Paul Buitelaar, Miguel Ángel García-Cumbreras, Salud María Jiménez-Zafra, José Antonio García-Díaz, Rafael Valencia-García, and Nitesh Jindal. 2023. Overview of Second Shared Task on Homophobia and Transphobia Detection in English, Spanish, Hindi, Tamil, and Malayalam. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1217–1230.

Subalalitha Chinnaudayar Navaneethakrishnan, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Malliga Subramanian, Prasanna Kumar Kumaresan, Bharathi, Lavanya Sambath Kumar, and Rahul Ponnusamy. 2022. Findings of shared task on Sentiment Analysis and Homophobia Detection of YouTube Comments in Code-Mixed Dravidian Languages. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 18–21.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Suraj Kumar Donthula and Abhishek Kaushik. 2019. Man is what he eats: A research on Hinglish sentiments of YouTube cookery channels using deep learning. *International Journal of Recent Technology and Engineering (IJRTE)*, 8(2S11):930–937.

Elisabetta Fersini, Debora Nozza, Paolo Rosso, et al. 2020. AMI@ EVALITA2020: Automatic misogyny identification. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*. (seleziona...).

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Sameer Hinduja and Justin W Patchin. 2020. Bullying, cyberbullying, and LGBTQ students. *Cyberbullying Research Center. http://hdl. handle. net/20.500*, 11990:2073.

Gagandeep Kaur, Abhishek Kaushik, and Shubham Sharma. 2019. Cooking is creating emotion: A study on Hinglish sentiments of YouTube cookery channels using semi-supervised approach. *Big Data and Cognitive Computing*, 3(3):37.

Subramaniam Kazhuparambil and Abhishek Kaushik. 2020a. Classification of Malayalam-English mix-code comments using current state of art. In *2020 IEEE International Conference for Innovation in Technology (INOCON)*, pages 1–6. IEEE.

Subramaniam Kazhuparambil and Abhishek Kaushik. 2020b. Cooking is all about people: Comment classification on cookery channels using BERT and classification models (Malayalam-English mix-code). *arXiv preprint arXiv:2007.04249*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. MuRIL: Multilingual representations for Indian languages. *arXiv preprint arXiv:2103.10730*.

Ritesh Kumar, Guggilla Bhanodai, Rajendra Pamula, and Maheshwar Reddy Chennuru. 2018. TRAC-1 shared task on aggression identification: IIT (ISM)@ COLING'18. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pages 58–65.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Ruba Priyadharshini, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2023. Homophobia and transphobia detection for low-resourced languages in social media comments. *Natural Language Processing Journal*, page 100041.

Prasanna Kumar Kumaresan, Ruba Priyadharshini, Bharathi Raja Chakravarthi, Paul Buitelaar, Asha Hegde, Hosahalli Lakshmaiah Shashirekha, Saranya Rajiakodi, Miguel Ángel García-Cumbreras, Salud María Jiménez-Zafra, José Antonio García-Díaz, Rafael Valencia-García, Kishore Kumar Ponnusamy, Poorvi Shetty, and Daniel García-Baena. 2024. Overview of Third Shared Task on Homophobia and Transphobia Detection in Social Media Comments. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity and Inclusion*, Malta. European Chapter of the Association for Computational Linguistics.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schäfer, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini, et al. 2021. Overview of the HASOC subtrack at FIRE 2021: Hate speech and offensive content identification in English and Indo-aryan languages. *arXiv preprint arXiv:2112.09301*.

Deepalakshmi Manikandan, Malliga Subramanian, and Kogilavani Shanmugavadivel. 2022. A System For Detecting Abusive Contents Against LGBT Community Using Deep Learning Based Transformer Models. In *Working Notes of FIRE 2022-Forum for Information Retrieval Evaluation (Hybrid). CEUR*.

Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. 2018. Did you offend me? classification of offensive tweets in Hinglish language. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 138–148.

Chris McCormick. 2016. Word2vec tutorial-the skip-gram model. *Apr-2016.[Online]. Available: http://mccormickml. com/2016/04/19/word2vec-tutorial-the-skip-gram-model*.

Megan A Moreno, Aubrey D Gower, Heather Brittain, and Tracy Vaillancourt. 2019. Applying natural language processing to evaluate news media coverage of bullying and cyberbullying. *Prevention science*, 20(8):1274–1283.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2020a. A BERT-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*, pages 928–940. Springer.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020b. Hate speech detection and racial bias mitigation in social media based on BERT model. *PloS one*, 15(8):e0237861.

Dean Ninalga. 2023. Cordyceps@ LT-EDI: Patching Language-Specific Homophobia/Transphobia Classifiers with a Multilingual Understanding. *arXiv preprint arXiv:2309.13561*.

María Antonia Paz, Julio Montero-Díaz, and Alicia Moreno-Delgado. 2020. Hate speech: A systematized review. *Sage Open*, 10(4):2158244020973022.

Shrey Satapara, Sandip Modha, Thomas Mandl, Hiren Madhu, and Prasenjit Majumder. 2021. Overview of the HASOC subtrack at FIRE 2021: Conversational hate speech detection in code-mixed language. *Working Notes of FIRE*, pages 13–17.

Sonali Rajesh Shah and Abhishek Kaushik. 2019. Sentiment analysis on Indian indigenous languages: a review on multilingual opinion mining. *arXiv preprint arXiv:1911.12848*.

Sonali Rajesh Shah, Abhishek Kaushik, Shubham Sharma, and Janice Shah. 2020. Opinion-mining on Marglish and Devanagari comments of YouTube cookery channels using parametric and non-parametric learning models. *Big Data and Cognitive Computing*, 4(1):3.

Sidney Wong, Matthew Durward, Benjamin Adams, and Jonathan Dunn. 2023. cantnlp@ LT-EDI-2023: Homophobia/Transphobia Detection in Social Media Comments using Spatio-Temporally Retrained Language Models. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 103–108.

Sargam Yadav and Abhishek Kaushik. 2022. Contextualized Embeddings from Transformers for Sentiment Analysis on Code-Mixed Hinglish Data: An Expanded Approach with Explainable Artificial Intelligence. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 99–119. Springer.

Sargam Yadav, Abhishek Kaushik, and Kevin McDaid. 2023a. Hate Speech is not Free Speech: Explainable Machine Learning for Hate Speech Detection in Code-Mixed Languages. In *2023 IEEE International Symposium on Technology and Society (ISTAS)*, pages 1–8. IEEE.

Sargam Yadav, Abhishek Kaushik, and Shubham Sharma. 2021. Cooking well, with love, is an art: Transformers on YouTube Hinglish data. In *2021 International Conference on Computational Performance Evaluation (ComPE)*, pages 836–841. IEEE.

Sargam Yadav, Abhishek Kaushik, and Surbhi Sharma. 2023b. Comprehensive Analysis of the Artificial Intelligence Approaches for Detecting Misogynistic Mixed-Code Online Content in South Asian Countries: A Review. *Cyberfeminism and Gender Violence in Social Media*, pages 350–368.