# Exploring Homophobic Discourse in Tamil Transcripts: A Comparative Analysis of Sentiment Classification Models using TF-IDF Vectorization

Soubraylu Sivakumar
*Computing Technologies,*
*SRM Institute of Science and Technology,*
SRM Nagar, Kattankulathur,
Tamilnadu, India.
sivas.postbox@gmail.com

Jyothi Prakash
*Computing Technologies*
*SRM Institute of Science and Technology,*
SRM Nagar, Kattankulathur,
Tamilnadu, India.
jb4412@srmist.edu.in

P. Saravanan
*Computing Technologies,*
*SRM Institute of Science and Technology,*
SRM Nagar, Kattankulathur,
Tamilnadu, India.
saravanp7@srmist.edu.in

T. Rajesh Kumar
*Computer Science and Engineering,*
*Saveetha School of Engineering,*
*Saveetha Institute of Medical and Technical Sciences,*
Chennai, Tamil Nadu, India.
t.rajesh61074@gmail.com

Selvanayaki Kolandapalayam Shanmugam
*Mathematics and Computer Science,*
*Ashland University, OH,*
*United States of America.*
skolanda@ashland.edu

Ratnavel Rajalakshmi
*School of Computer Science and Engineering,*
*Vellore Institute of Technology*,
Chennai, India.
rajalakshmi.r@vit.ac.in

2024 Second International Conference on Advances in Information Technology (ICAIT) | 979-8-3503-8386-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICAIT61638.2024.10690521

*Abstract*— **This research delves into the linguistic patterns of homophobic discourse within Tamil transcripts, employing sentiment analysis methods and TF-IDF vectorization for feature representation. Homophobia, characterized by prejudice, discrimination, or antagonism directed against individuals based on their sexual orientation, is a pervasive issue addressed in this study. Various ML algorithms, such as Support Vector Machine Classifier (SVM), Naive Bayes Classifier (NBC), Logistic Regression, and Gradient Boosting (GB), are utilized to discern and assess their effectiveness in distinguishing between homophobic and non-homophobic language. The dataset has been derived from various online platforms, prominently including YouTube, alongside other sources. Performance metrics including accuracy, F1 score, precision, and support are employed to gauge the models' performance. The outcomes shed light on the prevalent discriminatory language within the Tamil-speaking community and offer insights into combating homophobia through computational linguistic techniques. Notably, SVM demonstrated the highest accuracy of 88.18% in the classification task.**

***Keywords— Homophobic discourse, Sentiment analysis, Support Vector Machine (SVM), Gradient Boosting (GB), Naive Bayes Classifier (NBC), Logistic Regression (LR), Tamil-speaking community***

## I. Introduction

In recent years, an increasing fascination has emerged regarding the utilization of ML methods to examine discriminatory language within diverse cultural and linguistic settings. A crucial area of concentration pertains to identifying and scrutinizing homophobic rhetoric, a prevalent issue on a global scale. Among the Tamil-speaking population, manifestations of homophobia are frequently articulated through language, perpetuating bias and discrimination. Comprehending the linguistic structures of homophobic discourse is imperative for formulating efficacious approaches to combat these detrimental attitudes and advocate for inclusivity.

This study aims to identify the most effective machine learning models for detecting homophobic discourse within Tamil transcripts. It endeavours to leverage sentiment analysis techniques and TF-IDF vectorization to identify and analyse instances of homophobic expressions accurately. Through the utilization of diverse machine learning algorithms, the research aims to assess their efficacy in discriminating between homophobic and non-homophobic speech specifically in the context of Tamil text. The primary objective of the study is to uncover prevalent discriminatory language in the Tamil-speaking community and explore the potential of computational linguistic strategies in addressing homophobia. Additionally, the study seeks to provide a comprehensive analysis of the performance of various machine learning models, including Support Vector Machine (SVM), Gradient Boosting Machine (GBM), Logistic Regression (LR), and Naive Bayes (NB), in detecting homophobic language within Tamil transcripts. By evaluating the effectiveness of these models through metrics viz. precision, sensitivity, F1 score and accuracy, the research aims to identify the optimal approach for combating homophobia through computational linguistic techniques tailored for Tamil text analysis.

This study's main contribution is to:

- Apply sentiment analysis techniques and TF-IDF vectorization for feature representation in Tamil transcripts to detect homophobic language.
- Examine the effectiveness of different machine learning algorithms, such as SVM, NBC, Logistic Regression, and GB, in accurately distinguishing between homophobic and non-homophobic language.
- Evaluate the performance measures of the utilized models, including accuracy, F1 score, precision, and support, to ascertain their efficacy in recognizing homophobic discourse in Tamil written materials.

979-8-3503-0692-7/23/$31.00 ©2023 IEEE

This investigation holds considerable consequences for academia as well as society. Through the revelation of widespread discriminatory discourse in the Tamil-speaking population, the research aids in enhancing understanding regarding the enduring existence of homophobia. Furthermore, the utilization of machine learning methodologies provides tangible perspectives on addressing discriminatory mindsets via automated scrutiny of text-based information. Ultimately, the research endeavours to nurture inclusiveness and advocate for tolerance within the Tamil-speaking populace and other communities by utilizing computational linguistic strategies.

Section 2's literature review delves into the analysis of various classification tasks performed on Tamil text. This section comprehensively explores the various methodologies, their utilization, and practical applications in detail. Moving on to Section 3, a detailed examination of the proposed flowchart designed for the classification process is provided. Section 4 is dedicated to the discussion of the experimental setup implemented and the corresponding findings. Furthermore, Sections 5 and 6 are allocated for the thorough exploration of the conclusions drawn and the future enhancements that can be made in this study.

## I. Related Work

As part of the LT-EDI-ACL2022 shared task, teams demonstrated a variety of techniques for identifying homophobia and transphobia in social media comments in a variety of languages, including Tamil, English, and datasets in code-mixed Tamil and English. Every team devised distinct strategies, utilising cutting-edge methods and models to improve the identification of objectionable material in comments.

Performance metrics were a vital reference point for assessing how well the various strategies worked. Achieving an impressive 85.06% of macro F1-score for the Tamil dataset, the UMUTeam led by José A et al. [1] stood out by demonstrating promising results in identifying homophobic and transphobic words within this linguistic context. Bitsa_nlp by Vitthal B et al. [2] also showed competitive performance, demonstrating their ability to navigate the challenges of recognizing offending content across languages with 0.64 and 0.42 macro-averaged F1-scores for Tamil and English respectively.

The teams' approaches were as creative and varied as they were varied, displaying a wide range of strategies applied to the problem at hand. Many systems relied on extensive preparation of textual data from social media platforms, including YouTube comments and tweets. To extract the subtleties of homophobic and transphobic language found in the data, feature extraction approaches such linguistic features, pretrained word embedding's, FastText sentence embedding's, and BERT sentence embedding's were widely used as Muskaan S et al. [3], Bharathi R et al. [4], Nsrin A et al. [5], Ishan S et al. [6] respectively.

Furthermore, several teams adopted ensemble learning as a common strategy to enhance model performance. By combining multiple models or classifiers, teams leveraged the strengths of each individual model while mitigating their respective weaknesses. In order to address class imbalance in the datasets and maintain the models' effectiveness and robustness in a variety of linguistic contexts, data augmentation techniques were also utilized.

Among the standout approaches was ABLIMET's innovative use of a RoBERTa-based strategy, focusing on training models with balanced datasets through techniques like Random Over Sampler, significantly enhanced model performance, emphasizing the crucial role of dataset balancing in improving the robustness of models for detecting offensive language was proposed by Abulimiti M et al. [7].Similarly, SSNCSE_NLP's hybrid approach, which combined word embedding's, BERT-based transformers and SVM classifiers, exhibited high weighted F1-scores across various datasets. This indicates the potential of integrating diverse techniques for effective detection across languages, as demonstrated by Krithika S et al. [8].

However, despite notable achievements, challenges such as dataset class imbalance and biases introduced during preprocessing persisted, necessitating further research and refinement of methodologies to address these hurdles. Additionally, the nuanced nature of homophobic and transphobic discourse presented unique challenges, demanding ongoing investigation and improvement of existing approaches.

Several teams in the LT-EDI-ACL2022 shared task demonstrated significant success in identifying homophobic and transphobic language in social media comments. According to bitsa_nlp's remarkable macro-averaged F1-scores of 0.42 for English and 0.64 for Tamil, as well as UMUTeam's exceptional performance in the Tamil dataset with a remarkable macro F1-score of 85.06 by José A et al. [1], Vitthal B et al. [2] have demonstrated the effectiveness of advanced techniques in combating hate speech online.

Furthermore, ABLIMET's strategic use of a RoBERTa-based approach, emphasizing balanced dataset training, further exemplified the potential of sophisticated methodologies in addressing challenges related to detecting offensive language in social media comments Abulimiti M et al. [7]

In summary, the LT-EDI-ACL2022 shared task showcased researchers' collective efforts in advancing hate speech detection in social media comments [10-11]. Each approach contributed unique insights, pushing the boundaries of natural language processing. Continued collaboration and innovation are crucial for addressing existing challenges and enhancing the efficacy of models in detecting offensive language across diverse linguistic contexts.

Building upon the survey, the study aims to explore the effectiveness of TF-IDF in detecting hate speech in social media. By combining TF-IDF with classification algorithms, the goal is to improve hate speech detection accuracy. A comparative analysis [12] will be conducted to examine the advantages and limitations of using TF-IDF for hate speech detection. The research aims to enhance computational linguistic techniques for identifying offensive content in different contexts, contributing to natural language processing and social media moderation efforts.

## I. Materials And Method

The system functions in accordance with the procedure shown in Fig. 1.

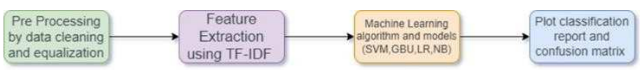



Fig 1. Proposed Homophobia disclosure block diagram.

979-8-3503-0692-7/23/$31.00 ©2023 IEEE

### A. Hardware Environment

TABLE I. HARDWARE ENVIRONMENT

| | |
|---|---|
| **GPU** | NVIDIA GeForce GTX 1050 ti |
| **CPU** | Intel(R) Core(TM) i5-9300H |
| **Installed Memory(RAM)** | 8GB |
| **Operating System** | Windows 11 (64-bit OS) |

### B. Software Environment

The programming language of choice for creating ML and DL algorithms is Python. Model training takes place in an ML environment called Jupyter Notebook.

### C. Vectorization Explanation

This section presents a visualisation of the suggested process for converting textual data into numerical vectors for machine learning algorithms. TF-IDF vectorization assigns weights based on term frequency in documents. This approach highlights important terms for distinguishing homophobic and non-homophobic language. Text representation in numerical format enables analysis [13] by machine learning algorithms for classifying linguistic patterns of homophobia in Tamil transcripts.

```
"இந்த மாதிரி பண்ணாதிங்க இந்த பாவத்த நீங்க அனுபவ",
"ரொம்ப கஷ்டமா இருக்க இந்த வீடியோ",
"சிரிப்பு காட்டாத பா",
"இவள்ங்களை துடைப்பங்கட்டையா அடிக்கனும்",
```

| | அட | அன | இந | இர | இவள | கட | கன |
|---|---|---|---|---|---|---|---|
| 1 | 0.000000 | 0.392644 | 0.619130 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 2 | 0.000000 | 0.000000 | 0.414289 | 0.525473 | 0.000000 | 0.000000 | 0.000000 |
| 3 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 4 | 0.408248 | 0.000000 | 0.000000 | 0.000000 | 0.408248 | 0.408248 | 0.408248 |

| | கள | கஷ | டம | பங | பண | பவ | வத |
|---|---|---|---|---|---|---|---|
| 1 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.392644 | 0.392644 | 0.392644 |
| 2 | 0.000000 | 0.525473 | 0.525473 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 3 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 4 | 0.408248 | 0.000000 | 0.000000 | 0.408248 | 0.000000 | 0.000000 | 0.000000 |

Fig 2. Text before and after TF-IDF vectorization

### D. Machine Learning Algorithm

Considering a training dataset $D = \{(x_i, y_i)\}$, where $x_i \in R^d$ is the feature vector and $y \in \{-1, +1\}$ is the corresponding class label. The steps involved in SVM algorithm is given below:

Step 1: Kernel Selection

The RBF kernel, also known as the Gaussian kernel, is a widely used kernel function in SVMs.

$$K(x_i, x_j) = exp(-\gamma \|x_i - x_j\|^2) \quad (1)$$

where $K(x_i, x_j)$ is the kernel function; $x_i, x_j$ are feature vectors; $\|x_i - x_j\|^2$ is the squared Euclidean distance between $x_i$ and $x_j$; and $\gamma$ is a parameter that defines the influence of a single training example.

Step 2: Formulate Optimization Problem

The optimization problem can be expressed [14] as:

$$\max imize \quad W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (2)$$

$$subject \text{ to } \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$0 \le \alpha_i \le C \quad \forall i = 1, \cdots, n$$

Here C is the tuning parameter, and $\alpha_i$ are the Lagrange multipliers.

Step 3: Solve the Optimization Problem
Use quadratic programming (QP) solvers or Sequential Minimal Optimization (SMO) to find the optimal values of $\alpha$.

Step 4: Identify Support Vectors (SV)
SV are the data points for which $\alpha_i > 0$. Compute the weight vector m and bias c (for linear kernel) or proceed to construct the decision function directly (for non-linear kernels).

Step 5: Construct Decision Function

The decision function for a linear SVM is given by:

$$fun(x) = m \cdot x + c \quad (3)$$

where c is the constant term, m · x is the dot product between the Coefficient vector m and the input feature vector $x$, and fun(x) is the Scoring function output.

Step 6: Output

The Scoring function $(x)$ for predicting the class label of a new instance $x$:

$$y = sign(fun(x)) \quad (4)$$

### E. Dataset Description

The dataset comprises text data gathered from various online platforms, prominently including YouTube, alongside other sources. This corpus is carefully categorized into three distinct classes: (i) Non-anti-LGBT+ content, (ii) Transphobia, and (iii) Homophobia with corresponding instance counts of 2064, 145, and 453 respectively. To facilitate model development and evaluation, the dataset undergoes meticulous partitioning into a training subset, encompassing approximately 1651 instances, and a testing subset, comprising roughly 413 instances. Fig 3. Gives some example texts available

| | text | category |
|---|---|---|
| 0 | இந்த மாதிரி பண்ணாதிங்க இந்த பாவத்த நீங்க அனுபவ... | Non-anti-LGBT+ content |
| 1 | ரொம்ப கஷ்டமா இருக்க இந்த வீடியோ | Non-anti-LGBT+ content |
| 2 | சிரிப்பு காட்டாத பா | Non-anti-LGBT+ content |
| 3 | மற்றவர்களின் உணர்ச்சிகளையும் மதிப்போம் | Non-anti-LGBT+ content |
| 4 | இவள்ங்களை துடைப்பங்கட்டையா அடிக்கனும் | Transphobia |

Fig 3. Example text of the dataset

### F. Classification Metrics

Accuracy:

The accuracy of the model in classifying both homophobic

979-8-3503-0692-7/23/$31.00 ©2023 IEEE

and non-homophobic content is exemplified by its homophobia detection accuracy, which is calculated as the ratio of correctly identified homophobic language instances to the total number of instances assessed.

$$Accuracy = \frac{Correct \text{ Prediction of Homophobic Disclosure}}{Total \text{ Prediction of Homophobic Disclosure}} \quad (5)$$

Precision:

The percentage of successfully recognised homophobic language instances among all instances projected to be homophobic is known as precision in homophobia detection, and it highlights the model's ability to prevent false positives.

$$Precision = \frac{Hit \text{ of Homophobic Disclosure}}{Predicted \text{ Positives of Homophobic Disclosure}} \quad (6)$$

Recall (Sensitivity):

Recall in homophobia detection evaluates the percentage of accurately identified instances of homophobic language among all instances actually containing homophobic language, emphasizing the model's capability to minimize false negatives.

$$Recall = \frac{Correct \text{ Prediction of Homophobic Disclosure}}{Actual \text{ Positives of Homophobic Disclosure}} \quad (7)$$

F1 Score:

The F1 score, which calculates the harmonic mean of precision and sensitivity, accurately evaluates a model's ability to identify homophobic language by considering both type I and II errors.

$$F1 = \frac{2 \times Sensitivity \times Precision}{Sensitivity + Precision} \quad (8)$$

## II. RESULT ANALYSIS

TABLE II. PERFORMANCE COMAPISION OF VARIOUS METHODS

| METHODS | ACC | PREC | REC | F1 |
|---|---|---|---|---|
| Naïve Bayes (NB) | 78.99 | 78.53 | 78.98 | 69.89 |
| Logistic Regression (LR) | 83.68 | 81.11 | 83.67 | 79.05 |
| Gradient Boosting Machine (GBM) | 85.37 | 85.51 | 85.36 | 83.06 |
| Support Vector Machine (SVM) | 88.18 | 89.72 | 88.18 | 86.09 |

The data shown in Table II offers insightful information on how well different ML algorithms detect homophobic language [15] in Tamil transcripts. Each model was assessed using important metrics like precision, recall, accuracy, and F1 score, which provided useful information on how well the models identified homophobic discourse situations.

Naive Bayes (NB) illustrated the least effectiveness among the models, achieving an accuracy of 78.99%, precision of 78.53%, and F1 score of 69.89%. Though NB contributed insights into recognizing homophobic language, its inferior performance compared to other models, particularly in precision and F1 score, implies constraints in precisely classifying instances of homophobic discourse in the dataset.

When contrasting Logistic Regression (LR) with Naive Bayes (NB), LR demonstrated superior performance, with an accuracy of 83.68%, surpassing NB's accuracy of 78.99%. LR also attained higher precision (81.11% versus NB's 78.53%) and a slightly enhanced F1 score (79.05% versus NB's 69.89%). While both models exhibited similar recall rates, LR's overall performance indicates its edge over NB in identifying instances of homophobic language within the dataset.

On the other hand, GBM [16] performed better than LR when compared to Logistic Regression (LR), attaining an accuracy of 85.37% as opposed to LR's 83.68%. Additionally, GBM showed much higher F1 score (83.06% versus LR's 79.05%) and higher precision (85.51% compared 81.11%). Despite similar recall rates in both models, GBM's overall performance implies its superiority over LR in accurately identifying instances of homophobic discourse.

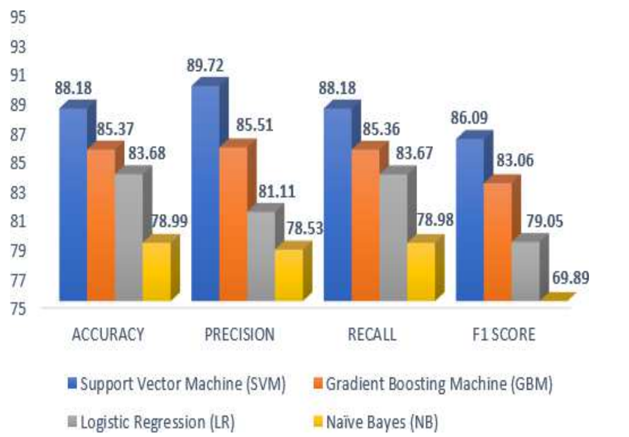
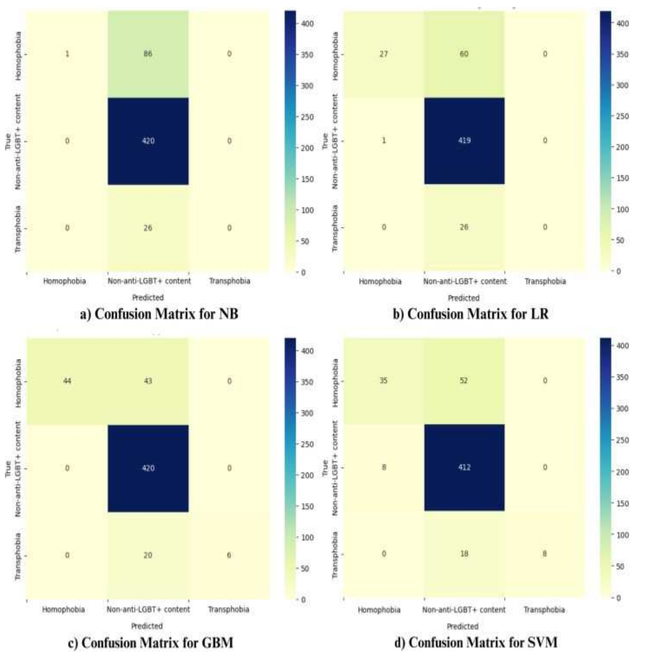



Fig 4. Performance metric comparison for homophobia detection in Tamil scripts

Conversely, in the comparison between SVM and GBM, SVM displayed higher accuracy (88.18% versus GBM's 85.37%) and precision [17] (89.72% versus GBM's 85.51%). Additionally, SVM exhibited a slightly higher F1 score (86.09% versus GBM's 83.06%). While both models showed comparable recall rates, SVM's overall performance suggests its supremacy over GBM in effectively identifying instances of homophobic language within the dataset.

In Fig 4. and Fig 5., the findings underline SVM's effectiveness as the favored model for detecting homophobic language in Tamil transcripts, closely trailed by GBM. These results underscore [18] the potential of ML algorithms in combating homophobia through computational linguistic methods, providing avenues for further investigation and practical application to foster inclusivity and acceptance within the Tamil-speaking community and beyond.

979-8-3503-0692-7/23/$31.00 ©2023 IEEE




a) Confusion Matrix for NB

b) Confusion Matrix for LR

c) Confusion Matrix for GBM

d) Confusion Matrix for SVM

Fig 5. Confusion Matrix for all the model

## III. Discussion

The results of this study offer important new information about how well different machine learning models identify homophobic speech in Tamil transcripts. Logistic Regression outperformed Naive Bayes; Gradient Boosting Machine surpassed LR. SVM has achieved an accuracy of 88.18% and precision of 89.72%. These results suggest that SVM, with its ability to discern subtle linguistic patterns, is well-suited for identifying instances of homophobic language in Tamil text.

The implications of these findings are significant, particularly in the context of addressing homophobia within the Tamil-speaking community. By employing computational linguistic techniques, it is possible to develop automated systems capable of flagging and addressing instances of discriminatory language in real-time. This has the potential to not only raise awareness about the prevalence of homophobia but also facilitate interventions aimed at fostering inclusivity and acceptance.

## IV. Limitations

Tamil language processing constrains the diversity and size of the training data and may affect model performance. Additionally, the linguistic characteristics and cultural nuances of Tamil discourse may influence the applicability of the methodologies. Notwithstanding these drawbacks, teamwork is required to overcome these obstacles. In order to enhance the precision and applicability of machine learning models for identifying discriminating language in Tamil, forthcoming studies ought to concentrate on constructing extensive datasets, establishing tools for language processing, and investigating different approaches.

## V. Conclusion

This study underscores the promise of ML methodologies in combating homophobia through linguistic analysis of Tamil transcripts. By leveraging computational techniques, deeper insights into the linguistic patterns of discriminatory discourse can be gained, and effective strategies can be developed for promoting inclusivity and acceptance within the Tamil-speaking community and beyond.

## References

[1]. José Antonio García-Díaz, Camilo Caparrós-Laiz Rafael, Valencia-García, "Detecting homophobic and transphobic comments in Tamil", Association for Computational Linguistics 2022 Second Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI-ACL2022), pp. 140- 144, May 27 2022, doi: 10.18653/v1/2022.ltedi-1.16.

[2]. Vitthal Bhandari, Poonam Goyal, "Leveraging Pretrained Language Models for Detecting Homophobia and Transphobia in Social Media Comments", Association for Computational Linguistics 2022 Second Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI-ACL2022), pp. 149- 154, May 27 2022, doi: 10.18653/v1/2022.ltedi-1.18.

[3]. Muskaan Singh, Petr Motlicek, "Homophobia/Transphobia Detection in social media comments", Association for Computational Linguistics 2022 Second Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI-ACL2022), pp. 356- 361, May 27 2022, doi: 10.18653/v1/2022.ltedi-1.55.

[4]. Bharathi Raja Chakravarthi, Ruba Priyadharshini, Durairaj Thenmozhi, John Phillip McCrae, Paul Buitelaar, Rahul Ponnusamy, Prasanna Kumar Kumaresan, "Overview of The Shared Task on Homophobia and Transphobia Detection in Social Media Comments", Association for Computational Linguistics 2022 Second Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI-ACL2022), pp. 356- 361, May 27 2022, doi: 10.18653/v1/2022.ltedi-1.57.

[5]. Nsrin Ashraf, Mohamed Taha, Ahmed Taha, Hamada Nayel, "Homophobia/Transphobia Detection for Equality, Diversity, and Inclusion using SVM", Association for Computational Linguistics 2022 Second Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI-ACL2022), pp. 287- 290, May 27 2022, doi: 10.18653/v1/2022.ltedi-1.42.

[6]. Ishan Sanjeev Upadhyay, KV Aditya Srivatsa, Radhika Mamid, "Ensembled Transformers Against Homophobia and Transphobia", Association for Computational Linguistics 2022 Second Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI-ACL2022), pp. 270- 275, May 27 2022, doi: 10.18653/v1/2022.ltedi-1.39.

[7]. Abulimiti Maimaitituoheti, Yang Yong, Fan Xiaochao, "A RoBERTa basedApproach for Homophobia/Transphobia Detection in Social Media", Association for Computational Linguistics 2022 Second Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI-ACL2022), pp. 155- 160, May 27 2022, doi: 10.18653/v1/2022.ltedi-1.19.

[8]. Krithika Swaminathan, Hrishik Sampath, Gayathri G L, B. Bharathi, "Homophobia/Transphobia Detection in Multiple Languages using SVM Classifiers and BERT-based Transformers", Association for Computational Linguistics 2022 Second Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI-ACL2022), pp. 239- 244, May 27 2022, doi: 10.18653/v1/2022.ltedi-1.34.

[9]. Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kayalvizhi Sampath, Durairaj Thenmozhi , Sathiyaraj Thangasamy, Rajendran Nallathambi, John Phillip McCrae, "Dataset for Identification of Homophobia and Transophobia in Multilingual YouTube Comments", ArXiv 2021, 1 September 2021, doi: https://doi.org/10.48550/arXiv.2109.00227.

[10]. Chakravarthi B., Priyadharshini R., Thavareesan S., Chinnappa D., Thenmozhi D., Sherly E., Mccrae J., Hande A., Ponnusamy R., Banerjee S., Vasantharajan , "indings of the Sentiment Analysis of Dravidian Languages in Code-Mixed Text", ArXiv 2021, 18 September 2021, doi: https://doi.org/10.48550/arXiv.2111.09811.

[11]. Bharathi Raja Chakravarthi , "Detection of homophobia and transphobia in YouTube comments", International Journal of Data Science and Analytics, 19 June 2023, doi: https://doi.org/10.1007/s41060-023-00400-0.

[12]. Soubraylu, Sivakumar & Kumar, Naveen, "Attention-Based Convolution Bidirectional Recurrent Neural Network for Sentiment Analysis", International Journal of Decision Support System Technology (IJDSST-Scopus, ESCI), Volume 14, no.1, pp. 1–21, 2022. DOI : 10.4018/IJDSST.300368.

979-8-3503-0692-7/23/$31.00 ©2023 IEEE

[13]. S. Sivakumar and R. Rajalakshmi, Comparative Evaluation of various feature weighting methods on movie reviews, Springer 2017 International Conference on Computational Intelligence in Data Mining (ICCIDM), pp. 721-730, Nov 2017. DOI: 10.1007/978-981-10-8055-5_64.

[14]. S. Sivakumar, L. S. Videla, T. Rajesh Kumar, J. Nagaraj, S. Itnal and D. Haritha, Review on Word2Vec Word Embedding Neural Net, 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2020, pp. 282-290, doi: 10.1109/ICOSEC49089.2020.9215319.

[15]. Soubraylu S, Rajalakshmi R. Hybrid convolutional bidirectional recurrent neural network based sentiment analysis on movie reviews. Computational Intelligence. 2020;1–23. https://doi.org/10.1111/coin.12400.

[16]. Soubraylu Sivakumar and Ratnavel Rajalakshmi, Self-attention based sentiment analysis with effective embedding techniques, Int. J. Computer Applications in Technology (Scopus, ESCI), Vol. 65, No. 1, pp.65–77, 2021. DOI: 10.1504/IJCAT.2021.113651.

[17]. Soubraylu Sivakumar and Ratnavel Rajalakshmi, 2021. Analysis of Sentiment on Movie Reviews Using Word Embedding Self-Attentive LSTM, International Journal of Ambient Computing and Intelligence (IJACI), IGI Global, vol. 12(2), pages 33-52, April. DOI : 10.4018/IJACI.2021040103.

[18]. Sivakumar, S., Rajalakshmi, R. Context-aware sentiment analysis with attention-enhanced features from bidirectional transformers. Soc. Netw. Anal. Min. 12, 104 (2022). Doi : 10.1007/s13278-022-00910-y.

979-8-3503-0692-7/23/$31.00 ©2023 IEEE