Hate speech classification for Sinhalese and Gujarati

Muhammad Deedahwar Mazhar Qureshi [*1,3], Madhuri Sawant [*1,3], M. Atif Qureshi[1,3], Wael Rashwan[1], Arjumand Younus[2,3] and Simon Caton[2,3]

[1]eXplainable Analytics Group, Faculty of Business, Technological University Dublin, Dublin, Ireland
[2]University College Dublin, Dublin, Ireland
[3]Science Foundation Ireland, Centre for Research Training in Machine Learning (ML-Labs)

Abstract

We, representing Team "XAG-TUD," participated in HASOC 2023, focusing on Task 1, which comprises subtasks 1A and 1B. Task 1A revolves around coarse-grained binary classification, specifically discriminating between content falling into the categories of HOF (Hateful or Offensive) and NOT for Sinhalese, a low-resource language. Similarly, Task 1B involves a similar classification for Gujarati, another low-resource language. In this paper, we provide detailed insights into our solutions for both sub-tasks within Task 1. Notably, our observations reveal that the LaBSE (Language-agnostic BERT Sentence Embedding) model consistently outperformed the XLM-R model for both sub-tasks, demonstrating its effectiveness in addressing hate speech classification challenges in these languages.

Code/Datasets for the paper are available on GitHub.[1]

1. Introduction

The ubiquitous use of social media platforms like Twitter and Facebook transcends age groups and diverse communities. While these platforms serve as a conduit for individuals to share moments from their lives, they also present an array of challenges. Within the vast expanse of content disseminated on these platforms, a significant portion is unsuitable for general audiences, often characterized by its offensive, hateful, insulting, or misleading nature, with specific targets within society. This proliferation of problematic content not only jeopardizes individual well-being but also disrupts the harmony of society as a whole. The challenge becomes especially pronounced in languages other than English, further exacerbated in low-resource language contexts, where identifying such problematic content proves to be an exceptionally formidable task.

The burgeoning presence of offensive content on the internet has propelled researchers to develop robust systems capable of automatically detecting and mitigating it. International competitions have been instrumental in advancing the field of offensive content identification.

[*]These authors contributed equally to this work.

Forum for Information Retrieval Evaluation, December 15-18, 2023, Goa, India

D22124696@mytudublin.ie (M. D. M. Q. *); D22130161@mytudublin.ie (M. S. *); atif.qureshi@tudublin.ie (M. A. Qureshi); wael.rashwan@tudublin.ie (W. Rashwan); arjumand.younus@ucd.ie (A. Younus); simon.caton@ucd.ie (S. Caton)

0009-0002-4878-9226 (M. D. M. Q. *); 0009-0004-4866-1710 (M. S. *); 0000–0003–4413–4476 (M. A. Qureshi); 0000–0002–2661–1892 (W. Rashwan); 0000-0001-7748-2050 (A. Younus); 0000-0001-9379-3879 (S. Caton)

[1]https://github.com/DeedahwarMazhar/HASOC-23-XAG-TUD

Among these competitions, HASOC (Hate Speech and Offensive Content Identification), initiated in 2019, stands as a significant milestone. In its fifth iteration, 2023, HASOC introduced four distinct tasks. This paper focuses on the findings of Task 1 of HASOC, which encompasses two sub-tasks.

Sub-Task 1A: Identifying Hate, Offensive, and Profane Content in Sinhala. This task primarily centers on identifying hate speech and offensive language in Sinhala. Sinhala, a low-resource Indo-Aryan language spoken by over 17 million people in Sri Lanka and one of the two official languages in the country, serves as a challenging linguistic context. This task involves coarse-grained binary classification to categorise tweets into two classes:

1. Hate and Offensive (HOF): Denoting posts containing hate, offensive, and profane content.
2. Non-Hate and Offensive (NOT): Indicating posts devoid of hate speech, profanity, and offensive content.

Sub-Task 1B: Identifying Hate, Offensive, and Profane Content in Gujarati. This task focuses on identifying hate speech and offensive language in Gujarati. Gujarati, another low-resource Indo-Aryan language, boasts approximately 50 million native speakers and holds the status of one of India's 22 official languages. Similar to sub-Task 1A, this task entails coarse-grained classification.

2. Related Work

This section discusses the challenges related to hate speech addressed in previous HASOC iterations. It is followed by an exploration of hate speech research within low-resource settings and then a general research background on hate speech.

2.1. HASOC challenge

2.1.1. HASOC for Hindi Language

Mandl et al. [1] provided a comprehensive overview of the HASOC 2019 iteration. They highlighted the popularity of Long-Short-Term Memory (LSTM) networks, effectively employing distributed word representations for text analysis. In the same competition, the QutNocturnal team [2] secured a noteworthy achievement with a Macro F1 score of 0.8149. Their success underscored the superiority of Convolutional Neural Networks (CNN) over LSTM when integrating transfer learning through word embeddings.

During the 2020 iteration, Raj et al. [3] embarked on an exploration of different approaches, incorporating both CNN and Bidirectional LSTM (BiLSTM). Among these, a single BiLSTM layer, coupled with fastText embeddings, emerged as a competitive solution, achieving a Macro-avg F1 score of 0.67 for hate speech classification.

In the subsequent 2021 iteration, Banerjee et al. [4] elevated the bar by fine-tuning a multilingual BERT model. Their approach included the addition of a classifier layer in the final phase, which was trained over 20 epochs. This rigorous methodology culminated in an outstanding achievement, boasting a Macro F1 score of 0.7797 and securing top honours in the competition.

2.1.2. HASOC for Marathi Language

In the HASOC 2021 iteration, Nene et al. [5] fine-tuned the XLM-R large model with a simple softmax layer and achieved a macro F1 score of 0.9144. In the same iteration, Glazkova et al. [6] proposed a system based on the Language-Agnostic BERT Sentence Embedding (LABSE), securing the second-best result with an F1 score of 0.8776. In the 2022 iteration, Chavan et al. [7] developed a BERT-based model pre-trained on a large monolingual dataset comprising tweets in the Marathi language called 'MahaTweetBert' and achieved a macro F1 score of 0.9156.

2.1.3. HASOC for German Language

In the 2019 HASOC iteration, Saha et al. [8] employed Multilingual BERT embeddings and LASER embeddings and attained a macro F1 score of 0.62. Subsequently, during the 2020 HASOC iteration, Mandl et al. [9] shed light on the notable achievements of the winning team, Comma@FIRE 2020 [10]. This team employed a joint fine-tuning approach involving mBERT, DistilBERT, RoBERTa, and XLM-R, ultimately achieving a macro F1 score of 0.5235.

2.2. Hate speech in a low-resource setting beyond HASOC challenge

In the context of the Sinhalese language, Ranasinghe et al. [11] directed their efforts toward classifying offensive content. For sentence-level offensive content identification, the XLM-R model emerged as the top performer, achieving an impressive 0.83 Macro F1 score. For token-level offensive language identification, XLM-R performed best with a 0.72 macro F1 score.

Kakwani et al. [12] introduced IndicNLPSuite, a collection of large-scale, general-domain, sentence-level corpora of 8.9 billion words across 11 Indian languages along with pre-trained models (IndicFT, IndicBERT) and NLU benchmarks (IndicGLUE)[2]. This has been used in hate speech detection tasks concerning languages of Indian origin.

Nkemelu et al. [13] undertook the task of developing machine learning models for the Burmese language, which is classified as a low-resource language. Their primary objective was to automatically detect hate speech posted on social media, focusing on the context of the Myanmar general election. Notably, they collected real-time data from Facebook. Similarly, Ishmam et al. [14] directed their efforts toward classifying Bengali comments found on Facebook pages. Their classification schema encompassed six distinct categories, including hate speech, communal attack, inciteful comments, religious hatred, political comments, and religious comments. They employed several machine learning algorithms, and they achieved noteworthy accuracy improvements, notably through the implementation of a Gated Recurrent Unit (GRU) based deep neural network.

Moy et al. [15] addressed hate speech detection in the Malay language, specifically targeting the Malaysian community. Their approach involved fine-tuning a pre-trained BERT model, which effectively adapted the model to the nuances of the Malay language and the local context. In a different linguistic context, Karunanayake et al. [16] employed a Convolutional Neural Network (CNN) in conjunction with Automatic Speech Recognition Systems (ASR) trained in the English language. Their goal was to classify the Sinhala and Tamil low-resource datasets,

[2] https://huggingface.co/ai4bharat/indic-bert

showcasing an innovative approach that leveraged existing technologies for language classification. Similarly, Mubarak et al. [17] focused on detecting vulgar and obscene speech within Arabic social media. Their research aimed to tackle offensive content, shedding light on the challenges of maintaining a respectful online environment in the Arabic language context.

2.3. Hate speech research beyond HASOC

Poletto et al. [18] provided a comprehensive overview of the datasets, lexicon, and evaluation campaigns focusing on hate speech. Their work provides detailed insights into hate speech corpora, shared tasks (such as open scientific competitions), hate speech lexicon, and various languages used in these contexts.

Naseem et al. [19] conducted a study on the impact of twelve different pre-processing techniques for tweet classification using three different labelled datasets focusing on Twitter hate speech and abusive language. Their research not only highlights the best-performing techniques but also identifies the least effective ones, offering valuable recommendations for optimising pre-processing techniques in individual use cases.

Burnap et al. [20] developed a classifier for hateful and antagonistic content on Twitter. This classifier served as a tool to assist policy and decision-makers in addressing the challenges of online hate speech. Furthermore, they applied an ensemble machine learning classifier to combat cyber hate, demonstrating the potential of machine learning in mitigating online hostility.

Matamoros et al. [21] conducted a systematic literature review and critique of academic articles published between 2014 and 2018, focusing on racism and hate speech on social media. Their work provides valuable insights into the scholarly discussions and trends regarding these issues during that period, shedding light on the evolving landscape of online hate speech.

3. Dataset and Data Pre-processing

The HASOC competition provided datasets to participating teams, comprising social posts sourced from Twitter[3]. The dataset encompasses two languages[4]: Sinhalese and Gujarati. [22], [23]

Sinhalese Dataset: The Sinhalese training dataset comprises 7500 records, while the test dataset comprises 2500 records. These datasets are based on the SOLD dataset by Ranasinghe et al. [11], which served as a foundational resource for the competition. The Sinhalese dataset has a subtle majority (57.6%) of tweets belonging to the *NOT* class, with the rest being *HOF* (Hate or Offensive).

Gujarati Dataset: In contrast, the Gujarati dataset consists of 200 records for training and approximately 1200 records for testing purposes. The Gujarati training dataset is completely balanced with 100 samples each for training and testing, as shown in Table 1.

Given the low-resource nature of these languages, a unique data pre-processing strategy was adopted. To enhance the dataset, each post was translated into English using the Google

[3]Now known as X

[4]Both language datasets are accessible on the HASOC website https://hasocfire.github.io/hasoc/2023/

Table 1
Class label distribution of the dataset

Language	HOF	NOT	Total
Sinhalese	3176	4324	7500
Gujarati	100	100	200

Translate API. This process resulted in a post-translation pairing for each entry. The suitability of Google Translate has been examined in various academic [24] and medical [25] contexts. In academic settings that do not demand intricate technical communication, it has demonstrated adequate semantic accuracy, albeit with occasional grammatical issues [24]. In domains like medicine, where precise grammar and syntax are crucial, it has shown to be less effective [25]. The unique nature of social media User-Generated Content (UGC) allows for the use of such translation tools without significantly compromising semantic understanding, a fact exemplified by its growing popularity in social networking environments [26]. Furthermore, it has proven to be a valuable tool for translations in machine translation settings, as suggested by de Vries in the context of the CBOW (Comparative Bag-of-Words) approach [27].

Subsequently, we employed the LaBSE Fast Tokenizer to tokenize these post-translation pairs, adding additional padding while maintaining a maximum token length of 512. From this tokenized input, we extracted input-IDs and attention masks, which served as inputs for our model. This data pre-processing approach facilitated a more comprehensive representation of the content within these low-resource language datasets, thereby intuitively assisting in the subsequent modelling and classification tasks.

4. Implementation

For the implementation of the model(s), we utilized the following hardware configuration:

1. To train our models, we employed an Nvidia T4 GPU equipped with 16GB of VRAM.
2. For testing the model's performance, we utilized an Intel Xeon CPU (2 vCPUs) with 13GB of RAM.

Our experimentation encompassed two distinct models: LaBSE and XLM-R. Both of these models have demonstrated their efficacy for low-resource languages, making them well-suited for our subtasks in Sinhalese and Gujarati.

LaBSE[5] language agnostic BERT sentence embedding model supporting 109 languages. Feng et al. [28] demonstrated that LaBSE excels even in languages where it lacks explicit training data, thanks to its language similarity and multilingual capabilities. LaBSE's dual-encoder approach, ideal for learning bilingual sentence embeddings, has been widely recognised. Additionally, LaBSE offers extensive language coverage and has been rigorously evaluated across various languages and their English translations.

In our binary classification approach, we adopted a strategy of translating the original language into English. Subsequently, we combined both the English translation and the original

[5]https://tfhub.dev/google/LaBSE/2

language before applying the LaBSE model. This approach has been successfully used previously for multilingual classifications using *Roman-script languages (English, Italian, French, German, and Spanish).* Balahur et al. [29] While there are not exactly low-resource languages in the modern age, they were still relevantly less extensively covered at the time of writing. The authors here have shown the potential of appending translations in multiple languages to improve classification performance for Twitter sentiment analysis. This approach effectively harnessed the power of cross-lingual embeddings, leveraging the synergy between tweets and their translations, as prevalent in our problem. Additionally, even with translations, the original text is still needed, which would otherwise degrade classification performance. This is consistent with the findings of Poncelas [30], who discovered that using translations alone runs the risk of degrading classification performance.

XLM-R[6] is a transformer-based multilingual masked language model, pre-trained on text from 100 languages. It has demonstrated its superiority over models like mBERT on cross-lingual classification tasks, particularly in low-resource language scenarios [31]. XLM-R boasts significant improvements in various NLP tasks, including classification, sequence labelling, and question answering, making it an excellent choice for our experiments, given its track record of success in low-resource language contexts.

Various prior studies have employed LaBSE and XLM-R, underscoring the efficacy of these models in tackling language-related tasks. Gamage et al.[32] highlighted XLM-R's superiority with an F1-score of 0.764 in the Sinhala language, and Pranith et al.[33] achieved promising results with LaBSE for English and IndicBERT models for Tamil and Malayalam. Dhananjaya et al.[34] further demonstrated XLM-R's strong performance for Sinhala text classification, and Heffernan et al.[35] reported positive outcomes using LaBSE and XLM-R for very low-resource African languages. Additionally, LaBSE's consistent outperformance of mBERT and XLM-R in language-English pairs, as observed in Feng et al. [28], further supports our choice of LaBSE and XLM-R for our Sinhalese and Gujarati subtasks. This wealth of evidence underscores the suitability of these models for our low-resource languages, leading us to utilize them in our study.

To maximise the potential of our models in the low-resource language context, we employed a technique based on Shi et al. [36], which automatically constructs text classifiers in a new language by leveraging labelled data from another language. This method transfers classification knowledge by translating model features. Our approach involved translating both of our low-resource languages into English and then appending the translated text to the original language data before model implementation.

While previous multilingual transformer models like mBERT and XLM have their merits, their limited scalability for low-resource languages led us to favour the LaBSE and XLM-R models for our experiments.

Table 2 provides a comparative overview of memory requirements and the prediction times for both transformer architectures in our test environment. An overview of the implementation is presented in the flow diagram in Figure 1.

[6]https://huggingface.co/xlm-roberta-base

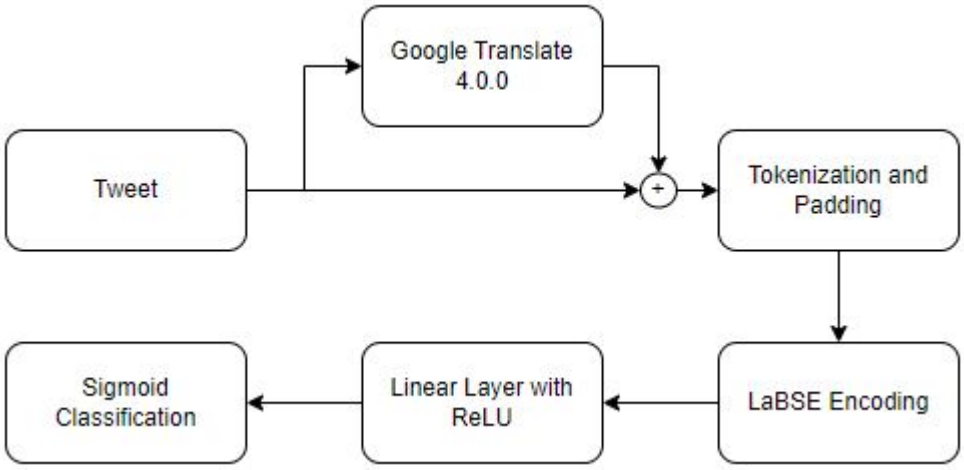

Figure 1: The process adopted for classification on low-resource language tweets.

4.1. Sinhalese

For Sinhalese, we flattened the *'pooler output'* of the transformer and appended it to a linear layer of size 256 with a LeakyReLU activation function ($\alpha = 0.01$), followed by a dropout layer with a rate of 0.3. The classification head consisted of a Sigmoid Layer. Our model was optimised using an AdamW Optimizer, with a learning rate set to $2e^{-5}$ and a Binary Cross-Entropy loss function.

The model is trained with a batch size of 4 for 15 epochs, with 15% of the dataset reserved for validation. To prevent overfitting, early stopping was implemented when the validation accuracy dropped for two consecutive epochs.

Our BERT model was sourced from the HuggingFace Transformers library.[7] We employed the BertTokenizerFast to construct a FAST BERT tokenizer, inheriting from PreTrainedTokenizerFast. Both the transformer and the tokenizer were retrieved from the uploaded version of *"setu4993/LaBSE"*.

4.2. Gujarati

In our Gujarati implementation, we harnessed the power of LaBSE in conjunction with SETFIT (Sentence Transformer Fine-tuning),[8] a highly efficient and prompt-free framework tailored for few-shot fine-tuning of sentence transformers (ST). This innovative framework, as showcased by Tunstall et al. [37], operates seamlessly without the need for prompts or verbalizers and achieves high accuracy with fewer parameters. Notably, it stands out for its faster training times compared to other few-shot techniques. SETFIT's versatility extends to multilingual settings, making it an ideal choice for our Gujarati coarse-grained binary classification task, where we grappled with limited data and the need for effective few-shot learning strategies.

[7] https://huggingface.co/transformers/v3.0.2/model_doc/bert.html
[8] https://huggingface.co/blog/setfit

Our architecture for Gujarati was structured with a linear layer of size 256, incorporating the LeakyReLU activation function ($\alpha = 0.01$). Following this, a sigmoid layer was employed at the classification head. For loss computation, we utilized the CosineSimilarityLoss function, with a batch size set at 32. The model underwent training for 7 epochs, after which it embarked on a few-shot learning phase consisting of 20 iterations, each involving 32 samples of a different split of the training data. To ensure model robustness and avoid overfitting, 15% of the training dataset was reserved for validation.

5. Experimental Results and Analysis

The test datasets for both Sinhalese and Gujarati are available on the HASOC website ([22], [23]).[9] In our experiments, we implemented two transformer architectures, LaBSE and XLM-R, for handling the Sinhalese dataset. For the Gujarati dataset, given its limited training data, we employed LaBSE with additional few-shot learning. LaBSE consistently delivered superior overall accuracy in both cases, as assessed on the validation dataset. The most promising architectures for each subtask were submitted for evaluation on the test dataset as part of the HASOC competition.

Tables 3 and 4 provide an overview of the validation accuracy and the test performance for both subtasks. Based on validation accuracy, we found that LaBSE model is more effective for both subtasks than the XLM-R models; therefore, we submitted the LaBSE model for the test run and reported the results as shown in table 4. Notably, for the test run, the higher F1-Score achieved for Sinhalese can be attributed to the availability of sufficient training data for this language. It's important to note that the test scores presented here are the actual run-submission results at HASOC. Given the lack of labels in the test set, we refrained from conducting multiple runs or cross-validation, leaving these as potential avenues for future exploration in our research. Furthermore, all the models discussed here use the same random state for the train-test split.

Upon a closer analysis of the results, it becomes evident that the model's predictions are significantly influenced by the presence or absence of disrespectful or indecent words within the tweet data. Two key scenarios emerged:

- False Positives: In some cases, benign posts containing terms from hate lexicons or words typically associated with hate speech were, perhaps, erroneously[10] classified as *HOF*. This misclassification often occurred due to the high representation of such terms in the training data.
- False Negatives: Conversely, hateful posts that did not contain the stereotypical hate terms were incorrectly flagged as *NOT*. This issue highlights the challenges of identifying subtle or less overt forms of hate speech.

Figure 3 and 4 represent the model's predictions on some examples of the test data set. Some tweets were classfied incorrectly by the model.

Furthermore, the presence or absence of specific indecent words had a notable impact on the model's predictions. Some tweets teetered on the borderline between *HOF* and *NOT* based on their content, making it challenging for the model to provide accurate classifications.

[9] https://hasocfire.github.io/hasoc/2023/

[10] or otherwise suffers from subjectivity

Table 2
Details of each transformer architecture in terms of Prediction Time and Model Size.

Model	Predication Time Per Sample	Model Size
LaBSE	0.53 seconds	500M Parameters
XLM-R	0.58 seconds	550M Parameters

Table 3
Validation accuracy for tasks

Task Name	Validation Accuracy
Sinhalese using LaBSE	0.8435
Gujarati using LaBSE with SETFIT	0.7667
Sinhalese using XLM-R	0.81
Gujarati using XLM-R	0.7333

Table 4
Comparative Analysis of Output on test dataset

Task Name	Macro F1	Precision	Recall	Run Name
Sinhalese using LaBSE	0.8127	0.8177	0.8092	Sinhalese LaBSE XAG-TUD
Gujarati using LaBSE with SETFIT	0.7799	0.7717	0.7958	Gujarati LaBSE XAG-TUD

The model also exhibited biases towards certain gender and religious domains, leading to misclassifications in cases involving stereotypical biases. This aspect highlights the need for further work in addressing model biases and ensuring fair and unbiased classifications.

Additionally, there were instances where tweets in the original low-resource language (Sinhalese/Gujarati) were incorrectly classified as *NOT*. This could potentially be attributed to nuances or contextual factors specific to the original language. To mitigate this, our approach of appending translations to the original tweets proved beneficial, as it allowed the tokenizer and the transformer to capture additional contextual cues from the translated text.

6. Conclusion and Future Directions

In this study, we undertook the challenging task of coarse-grained binary hate speech classification in low-resource languages, specifically Sinhalese and Gujarati, as part of two sub-tasks in Task 1 of HASOC 2023. Our findings revealed that the LaBSE BERT model consistently outperformed other transformer-based systems employed in our experiments. Our approach involved translating the target languages into English and then appending them to the original text before model implementation. For Gujarati, we leveraged the SETFIT model, well-suited for few-shot fine-tuning, enhancing our model's performance.

While this research marks a step in addressing hate speech classification in low-resource languages, it also opens up avenues for future exploration and improvement.

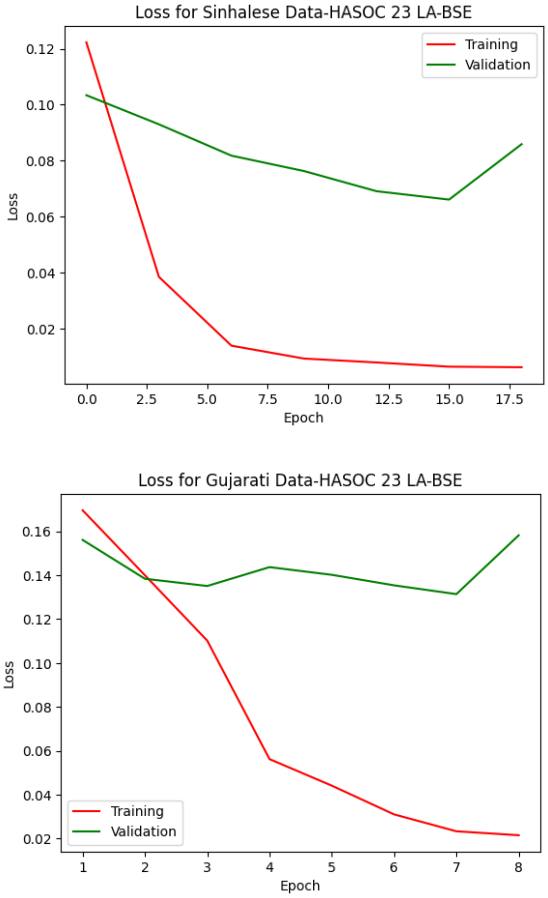

Figure 2: (a) Analysis of Sinhalese dataset (b) Analysis of Gujarati dataset

- Positional Impact of Translations: An intriguing future investigation would involve exploring how the positioning of translations within the text data might influence inference. Also, understanding and mitigating translation biases that may affect classification outcomes could be crucial for further refining the approach.
- Comparison of Translation APIs: While this study employed Google Translate for translation, future research could delve into the use of alternative translation APIs to provide a comparative analysis. Different translation services may exhibit variations in performance and biases, warranting a comprehensive assessment.
- Bias Mitigation: Given that most linguistic models, including ours, are influenced by dataset biases, addressing and mitigating biases in translated data is essential. Investigating methods to minimize gender and other biases introduced during translation can contribute to fairer and more reliable hate speech classification. The focus here could be to mitigate bias in datasets and predictions, especially gender bias, which has been found to be prevalent in multiple datasets. ([38], [39], [40], [41]).
- Cross-Validation and Comparative Analysis: With access to the full datasets beyond

Original Tweet	Google Translated Tweet	Result
තෝ නම් මහ සාපයක් බල්ලා පොඩිනිලමේ @USER · 9 Feb 2021 I did my second TikTok!	You are a cur*** d** @USER · 9 Feb 2021 I did my second TikTok!	False Negative
@USER කතන්දර අවශ්‍ය නැත හේලාදැකීම් අවශ්‍ය නැත පාකිස්තානය සමග සියලුම සම්බන්ධතා වහාම නවතා දැමිය යුතුය සියලුම දේශපාලන පක්ෂ පාට නොබලා එයට සහය දිය යුතුය ඒ ශ්‍රී ලාංකිකයා වෙනුවෙන් සියලුම ලාංකීය ජනතාව එකාවන්ව බේද අතහැර එක්විය යුතුය	@USER No need for stories No need for comments All relations with Pakistan should be stopped immediately All political parties should support it regardless of color All Sri Lankans should leave their differences and unite for the sake of that Sri Lankan	False Negative

Original Tweet	Google Translated Tweet	Result
@AUTHOR તમે આવા મૂર્ખ છો. મોદીજીને નફરત કરતી વખતે તમે લોકો તેમની સાથે જોડાયેલી દરેક વસ્તુનો વિરોધ કરવા માંડો છો.	@AUTHOR You are such an idiot. While hating Modi ji you people start opposing everything related to him.	False Negative
ગુજરાત માં ભરૂચ જિલ્લામાં આવેલો જંબુસર તાલુકામાં નગરપાલિકા ના કર્મચારીઓ કામ ચોર છે ને ગ્રાન્ડ પાસ થાય છે પણ કામકાજ નંથી થતાં તો રૂપિયા ક્યાં જાય છે ઈકવાઈરી બેસાળો નગરપાલિકા ના કર્મચારીઓ ઉપર	In Jambusar taluka of Bharuch district in Gujarat, the employees of the municipality are thieves and they get a grand pass, but if the work is not done, where does the money go?	False Negative
@AUTHOR કારણ કે ગુજરાત ના લોકો કોઈ ની પાછળ લાઇન લગાવી ને ઉભા રહી શકે છે ચોર ચકો લફંગો ગાંડો લેભાગુ કોઈ પણ નેતાબની જશે અને સત્યવાન અને પ્રજા હિત ધરાવતા લોકો ઘરે બેસી રહેશે એ હું નહિ દેવ શ્રી મામાઇ દેવ ની વાણી છે જે હજારો વર્ષ પહેલાં ઉંટબોધન આપ્યું છે જય ભારત	@AUTHOR Because the people of Gujarat can line up and stand behind anyone, thieves, thieves, fools, almost anyone will go to Netabni and the honest and public interest people will sit at home. It is the words of I Nahi Dev Sri Mamai Dev thousands of years ago. Jai Bharat has been declared	False Negative
કહેવાતા ઠગ ટોળકી પત્રકારો હજારો જાતના વિરોધ કરશે પણ જે મુખ્ય હકીકત છે, જે vision ને લઈને આ સભાઓ થાય છે એના માટે એક શબ્દ પણ નહીં ઉચ્ચારે. જ્યારે કોઈ હિન્દુ ધર્માંતર વિશે અવાજ ઉઠાવે ત્યારે બધા હરામી એક સાથે એનો વિરોધ કરે છે. અત્યારે ગુજરામાં પણ આવા દલાલ પત્રકારો તૂટી પડ્યા છે. @URL	The so-called thug gang of journalists will protest thousands of times but will not utter a single word for the main fact, the vision of which these meetings are held. When a Hindu raises his voice about conversion, all the b******* oppose it together. At present, even in Gujarat, such broker journalists have collapsed. @URL	False Negative

Figure 3: (a) False Negative examples of Sinhalese (b) False negative examples of Gujarati

the competition's constraints, future work can explore k-fold cross-validation to assess model robustness. Additionally, a deeper exploration of the literature for diverse classifiers and thorough comparative analyses can provide insights into refining hate speech classification models further.

- Code-Mixed Conversations and Non-Binary Classification: Extending the approach to address hate speech in code-mixed conversations, such as English-Hindi (Hinglish) or English-French, presents an intriguing challenge. Moreover, the research, initially focused on coarse-grained binary classification, can be expanded to tackle hate speech classification involving multiple non-binary classes.
- Conversational Hate Speech and Few-Shot Settings: The study's success in few-shot settings for low-resource languages opens doors for further experimentation. Future research can explore conversational hate speech detection in these languages and apply few-shot learning techniques to other low-resource languages, expanding the scope of

Original Tweet	Google Translated Tweet	Result
1927 ගාන්ධි තුමාගේ ශ්‍රීලංකා සංචාරයට පෙර "සැබවින්ම නිදහස් ඉන්දියාව තම අසල්වාසීන්ට උපකාර කිරීමට බැඳී සිටියි" යනුවෙන් පවසන ලදී. කලාශිල්පීන් වන @USER @USER විසින් නිදහස් චතුරශ්‍රයේදී ගායනා කරන ලදභජන් ගීතය “වෛෂ්ණවී ජාන තෝ” ජනාධිපතිතුමා @USER විසින් එළිදක්වන ලදී ,746	Before Gandhi's visit to Sri Lanka in 1927, he said, "Truly free India is bound to help her neighbours." Bhajan song “Vaishnav Jana Tho” sung at Freedom Square by artist @USER @USER was unveiled by President @USER ,746	False Positive
@USER @USER අනූස තමුසෙ දේශපාලනේට තමයි හොඳ...	@USER @USER Anusa you are good for politics...	False Positive

Original Tweet	Google Translated Tweet	Result
@AUTHOR દુનિયા નું સૌથી ધીરજ વાળું કોઈ જીવ હોઈ તો કાચબો છે. અને ગાળા મા નાગ છે મતલબ દુનિયા જેને ધિક્કારે છે એને શિવ ધરેણું બનાવે છે. બળદ મતલબ દુનિયા નું સુધી મહેનતુ પ્રાણી કેવાનું મતલબ એ કે બળદ જેવા મહેનતુ. કાચબા જેવી ધીરજ. અને જેને દુનિયા ધિક્કારે એનો સ્વીકાર કરો પછી જ શિવ ના દર્શન થાય🙏	@AUTHOR The most patient creature in the world is the turtle. And the period is Naga, which means what the world hates, Shiva makes it an ornament. Bull means hardworking animal till the end of the world. And only after you accept what the world hates, you will see Shiva	False positive
@AUTHOR @USER શાંત ગદાધારી ભીમ શાંત, એ નવરી બજાર એટલા ગુસ્સા ને પણ લાયક નથી. શેક્યો પાપડ તો એનાથી ભાંગે એમ છે નહી, તો ખાલી ફેન્ટાસિ માં રમ્યા કરે છે.	@AUTHOR @USER Shant Gadadhari Bhim Shant, Navari Bazar is not even worthy of such anger. Baked papad is not destroyed by it, otherwise it is just playing in fantasy.	False positive

Figure 4: (a) False positive examples of Sinhalese (b) False positive examples of Gujarati

this work.

Acknowledgments

This publication has emanated from research supported in part by a grant from Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission

References

[1] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, A. Patel, Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages, in: Proceedings of the 11th Forum for Information Retrieval Evaluation, 2019, pp. 14–17. URL: https://dl.acm.org/doi/abs/10.1145/3368567.3368584. arXiv:https://dl.acm.org/doi/pdf/10.1145/3368567.3368584.

[2] M. A. Bashar, R. Nayak, Qutnocturnal@ hasoc'19: Cnn for hate speech and offensive content identification in hindi language, arXiv preprint arXiv:2008.12448 (2020).

[3] R. Raj, S. Srivastava, S. Saumya, Nsit & iiitdwd@ hasoc 2020: Deep learning model for

hate-speech identification in indo-european languages., in: FIRE (Working Notes), 2020, pp. 161–167.
[4] S. Banerjee, M. Sarkar, N. Agrawal, P. Saha, M. Das, Exploring transformer based models to identify hate speech and offensive content in english and indo-aryan languages, arXiv preprint arXiv:2111.13974 (2021).
[5] M. Nene, K. North, T. Ranasinghe, M. Zampieri, Transformer models for offensive language identification in marathi, in: Forum for Information Retrieval Evaluation (Working Notes)(FIRE), CEUR-WS. org, 2021.
[6] A. Glazkova, M. Kadantsev, M. Glazkov, Fine-tuning of pre-trained transformers for hate, offensive, and profane content detection in english and marathi, arXiv preprint arXiv:2110.12687 (2021).
[7] T. Chavan, S. Patankar, A. Kane, O. Gokhale, R. Joshi, A twitter bert approach for offensive language detection in marathi, arXiv preprint arXiv:2212.10039 (2022).
[8] P. Saha, B. Mathew, P. Goyal, A. Mukherjee, Hatemonitors: Language agnostic abuse detection in social media, arXiv preprint arXiv:1909.12642 (2019).
[9] T. Mandl, S. Modha, A. Kumar M, B. R. Chakravarthi, Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german, FIRE '20, Association for Computing Machinery, New York, NY, USA, 2021, p. 29–32. URL: https://doi.org/10.1145/3441501.3441517. doi:10.1145/3441501.3441517.
[10] R. Kumar, B. Lahiri, A. K. Ojha, A. Bansal, Comma@ fire 2020: Exploring multilingual joint training across different classification tasks., in: FIRE (Working Notes), 2020, pp. 823–828.
[11] T. Ranasinghe, I. Anuradha, D. Premasiri, K. Silva, H. Hettiarachchi, L. Uyangodage, M. Zampieri, Sold: Sinhala offensive language dataset, arXiv preprint arXiv:2212.00851 (2022).
[12] D. Kakwani, A. Kunchukuttan, S. Golla, N. Gokul, A. Bhattacharyya, M. M. Khapra, P. Kumar, Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 4948–4961.
[13] D. Nkemelu, H. Shah, M. Best, I. Essa, Tackling hate speech in low-resource languages with context experts, in: Proceedings of the 2022 International Conference on Information and Communication Technologies and Development, 2022, pp. 1–11.
[14] A. M. Ishmam, S. Sharmin, Hateful speech detection in public facebook pages for the bengali language, in: 2019 18th IEEE international conference on machine learning and applications (ICMLA), IEEE, 2019, pp. 555–560.
[15] T. X. Moy, M. Rahem, R. Logeswaran, Multilingual hate speech detection, International Journal of Multidisciplinary Research and Publications 4 (2022).
[16] Y. Karunanayake, U. Thayasivam, S. Ranathunga, Transfer learning based free-form speech command classification for low-resource languages, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, Florence, Italy, 2019, pp. 288–294. URL: https://aclanthology.org/P19-2040. doi:10.18653/v1/P19-2040.
[17] H. Mubarak, K. Darwish, W. Magdy, Abusive language detection on arabic social media, in: Proceedings of the first workshop on abusive language online, 2017, pp. 52–56.
[18] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora

for hate speech detection: a systematic review, Language Resources and Evaluation 55 (2021) 477–523.
[19] U. Naseem, I. Razzak, P. W. Eklund, A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter, Multimedia Tools and Applications 80 (2021) 35239–35266.
[20] P. Burnap, M. L. Williams, Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making, Policy & internet 7 (2015) 223–242.
[21] A. Matamoros-Fernández, J. Farkas, Racism, hate speech, and social media: A systematic review and critique, Television & New Media 22 (2021) 205–224.
[22] S. Satapara, H. Madhu, T. Ranasinghe, A. E. Dmonte, M. Zampieri, P. Pandya, N. Shah, M. Sandip, P. Majumder, T. Mandl, Overview of the hasoc subtrack at fire 2023: Hate-speech identification in sinhala and gujarati, in: K. Ghosh, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, Goa, India. December 15-18, 2023, CEUR Workshop Proceedings, CEUR-WS.org, 2023.
[23] T. Ranasinghe, K. Ghosh, A. S. Pal, A. Senapati, A. E. Dmonte, M. Zampieri, S. Modha, S. Satapara, Overview of the HASOC subtracks at FIRE 2023: Hate speech and offensive content identification in assamese, bengali, bodo, gujarati and sinhala, in: Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE 2023, Goa, India. December 15-18, 2023, ACM, 2023.
[24] M. Groves, K. Mundt, Friend or foe? google translate in language for academic purposes, English for Specific Purposes 37 (2015) 112–121.
[25] S. Patil, P. Davies, Use of google translate in medical communication: evaluation of accuracy, Bmj 349 (2014).
[26] N. Bin Dahmash, 'i can't live without google translate': A close look at the use of google translate app by second language learners in saudi arabia, Arab World English Journal (AWEJ) Volume 11 (2020).
[27] E. De Vries, M. Schoonvelde, G. Schumacher, No longer lost in translation: Evidence that google translate works for comparative bag-of-words text applications, Political Analysis 26 (2018) 417–430.
[28] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic bert sentence embedding, arXiv preprint arXiv:2007.01852 (2020).
[29] A. Balahur, M. Turchi, Improving sentiment analysis in twitter using multilingual machine translated data, in: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, 2013, pp. 49–55.
[30] A. Poncelas, P. Lohar, A. Way, J. Hadley, The impact of indirect machine translation on sentiment classification, arXiv preprint arXiv:2008.11257 (2020).
[31] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).
[32] K. Gamage, V. Welgama, R. Weerasinghe, Improving sinhala hate speech detection using deep learning, in: 2022 22nd International Conference on Advances in ICT for Emerging Regions (ICTer), 2022, pp. 045–050. doi:10.1109/ICTer58063.2022.10024103.
[33] P. Pranith, V. Samhita, D. Sarath, D. Thenmozhi, Homophobia and transphobia detection

of youtube comments in code-mixed dravidian languages using deep learning (2022).
[34] V. Dhananjaya, P. Demotte, S. Ranathunga, S. Jayasena, Bertifying sinhala–a comprehensive analysis of pre-trained language models for sinhala text classification, arXiv preprint arXiv:2208.07864 (2022).
[35] K. Heffernan, O. Çelebi, H. Schwenk, Bitext mining using distilled sentence representations for low-resource languages, arXiv preprint arXiv:2205.12654 (2022).
[36] L. Shi, R. Mihalcea, M. Tian, Cross language text classification by model translation and semi-supervised learning, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010, pp. 1057–1067.
[37] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, O. Pereg, Efficient few-shot learning without prompts, arXiv preprint arXiv:2209.11055 (2022).
[38] E. M. Bender, B. Friedman, Data statements for natural language processing: Toward mitigating system bias and enabling better science, Transactions of the Association for Computational Linguistics 6 (2018) 587–604.
[39] D. Shah, H. A. Schwartz, D. Hovy, Predictive biases in natural language processing models: A conceptual framework and overview, arXiv preprint arXiv:1912.11078 (2019).
[40] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, W. Y. Wang, Mitigating gender bias in natural language processing: Literature review, arXiv preprint arXiv:1906.08976 (2019).
[41] S. L. Blodgett, S. Barocas, H. Daumé III, H. Wallach, Language (technology) is power: A critical survey of" bias" in nlp, arXiv preprint arXiv:2005.14050 (2020).