Hate Speech Detection in Low Resource Indo-Aryan Languages

Sougata Saha[1], Michael Sullivan[1] and Rohini Srihari[1]

[1]*State University of New York at Buffalo, NY 14260, United States*

Abstract
This report outlines the problem formulation and methodology employed by team Chetona for identifying hate-speech in low resource languages from social media comments. We focus on HASSOC 2023 Task 4, which involves binary classification of Twitter, Facebook, and Youtube comments for hate speech in Bengali, Bodo, and Assamese languages. We propose ensembling IndicBERT and Naive Bayes, along with synthetic data upsampling techniques, and attain macro F1 scores of 0.73, 0.68, and 0.84 for Assamese, Bengali, and Bodo. The scores are significant improvements over existing baselines, placing us within the top 10 of the leaderboard for all languages. The code and method is available on Github.[1]

Keywords
Hate Speech, Low resource language, Assamese, Bengali, Bodo

1. Introduction

Hateful comments are prevalent on social media platforms. Although tools for automatically detecting, flagging, and blocking such false, offensive, and harmful content online have matured lately, research in low-resource languages such as Assamese, Bengali, and Bodo is still lacking [1, 2, 3]. Most prior research pertains to detecting offensive text in social media while neglecting the subtler and broader task of identifying hateful comments. This document delineates the problem definition and the approach incorporated to tackle the challenges presented by HASOC 2023 Task 4 [4, 5], which revolves around identifying hate speech in low-resource Indo-Aryan textual content sourced from social media platforms.

2. Annihilate Hates (Task 4)

2.1. Problem Statement

This task aims to identify hate speech from social media (Twitter, Facebook, and YouTube) text comments spanning the low resource languages of Bengali, Bodo, and Assamese from Eastern India. The datasets for these languages contain sentences labeled as hate/offensive (HOF) or not hate (NOT), and the goal is to develop robust machine learning models that can reliably

[1] https://github.com/sougata-ub/team_chetona_hasoc_2023

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

sougatas@buffalo.edu (S. Saha); mjs227@buffalo.edu (M. Sullivan); rohini@buffalo.edu (R. Srihari)
https://sougata-ub.github.io/ (S. Saha); https://www.acsu.buffalo.edu/~mjs227 (M. Sullivan);
https://www.acsu.buffalo.edu/~rohini/ (R. Srihari)

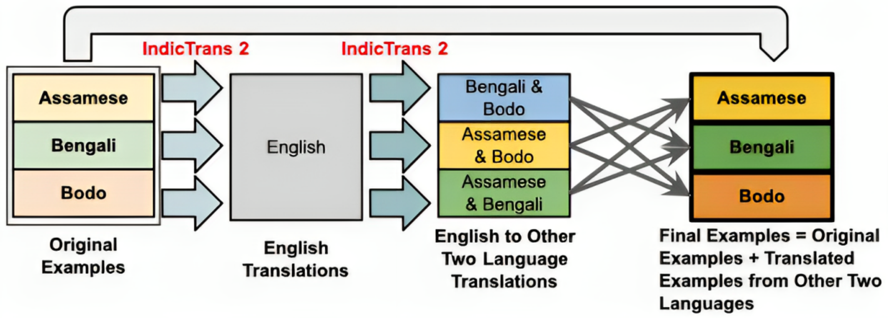

Figure 1: Task 4 Data Augmentation Pipeline.

predict the correct binary class for each comment. Models are evaluated and compared using the Macro F1 score.

2.2. Methodology

Our approach incorporates robust data augmentation followed by an ensemble approach for hate-speech detection.

2.2.1. Data Augmentation

The provided training data comprises 1281 Bengali, 4036 Assamese, and 1679 Bodo samples. We up-sample the training examples of each language by translating the examples from other two languages to the given language. For a language, we use the IndicTrans2 [6] Indic to English translation model to first translate examples from the other two languages to English. Next we use the English to Indic translation model to translate the English text to the desired source language. This method helps us generate additional 5715 Bengali, 2960 Assamese, and 5317 Bodo noisy training data, which we add to the original training data. We remove all emojis from the comments, and truncate them to 50 tokens. Figure 1 illustrates our data augmentation pipeline.

2.2.2. Model Architecture and Training

We implement an ensemble approach using IndicBERT [7] and Multinomal Naive Bayes. IndicBERT is a multilingual language model with 278 million parameters, and was trained on the IndicCorp v2 dataset [8] and evaluated on the IndicXTREME [9] benchmark. This model is versatile, supporting 23 Indic languages as well as English. We use the IndicBERT-MLM+Samanantar variant, a BERT-style model [10] trained on IndicCorp v2 and the Samanantar Parallel Corpus [11] focused on the MLM (Masked Language Model) objective. The final HOF/NOT predictions are made by passing the pooler output through a dropout layer with 0.1 probability, followed by a single linear layer. We also train a multinomal Naive Bayes (NB) model for each language using Sklearn. The NB model inputs tf-idf representations of comment tokens and predicts the binary HOF/NOT class. We weigh the predicted probabilities from the BERT-based model and

Naive Bayes model using a ratio of 4:1, and classify the text as HOF if the weighted sum is above a threshold of 0.5.

Since multi-task learning generally yields better results, we train a single model for all three languages. Furthermore, to distinguish between the original and translated examples, we prepend each sample text with *<language> + "original:" or "translated:"*, e.g. *"bengali original:<followed by the original text>", "bodo translated:<followed by the translated text>".* We fine-tuned IndicBert's 'ai4bharat/IndicBERTv2-MLM-Sam-TLM' version from the Hugging Face library using Pytorch. The model was optimized using AdamW [12] and a learning rate of 2e-5. We trained the model for 20 epochs on a single A5000 GPU with a batch size of 16. The training took approximately 1 hour with early stopping if the validation loss didn't decrease for 2 epochs. The Naive Bayes models were trained with the default Sklearn settings.

2.2.3. Results and Evaluation

We report our results in Table 1 and compare the test set Macro F1 score against teams that submitted predictions for all three languages. We highlight in bold the best F1 score for each language and underline our results. We attain the best results for Assamese and competitive results for Bengali and Bodo. Since most Indian languages originate from Devanagari, we experimented with using the IndicBERT-SS IndicBERT variant instead of IndicBERT-MLM+Samanantar. IndicBERT-SS was trained with the MLM objective on an Indic language to Devanagari corpus to encourage better lexical sharing among languages. We did not note any improvements in results on the validation and test sets (Note: We do not report the numbers here).

Team	Assamese	Bengali	Bodo
AI Alchemists	0.7075	0.7257	0.8437
Avigail Stekel	0.6988	0.6650	0.8507
Chen876	0.6812	0.6603	0.8427
Ours	**0.7346**	0.6786	0.8438
CIT TEAM	0.4684	0.3755	0.4153
CNLP-NITS-PP	0.5949	0.6011	0.6693
Code Fellas	0.6973	0.7196	0.8351
FiRC-NLP	0.7252	**0.7642**	0.8484
InclusiveTechies	0.3469	0.3583	0.3148
IRLab@IITBHU	0.6967	0.6527	0.7427
Komar99	0.6946	0.6467	0.8507
Michal Stekel	0.6862	0.6569	0.8379
MUCS	0.6884	0.6683	0.8368
Ravens	0.6621	0.6089	0.8434
SATLab	0.7151	0.6708	**0.8565**
Team +1	0.4832	0.4709	0.4952
TeamBD	0.7222	0.7349	0.7630

Table 1
Task 4 results Macro F1 comparison. Best F1 scores are highlighted in bold. Our results are underlined. (Note: The table is sorted alphabetically to accommodate the results from all three subtasks.)

3. Conclusion

This report outlines the problem formulation and the methodology employed to address the HASOC 2023 competition Task 4, focused on hate speech detection in low resource text from social media. Our proposed method implements a weighted ensemble of IndicBERT v2 and Multinomal Naive Bayes and incorporates a translation-based data augmentation approach. Our results indicate that our implementations can robustly detect social media hate speech in low resource Indo-Aryan languages, thus promoting a safer and more inclusive online environment.

References

[1] K. Ghosh, A. Senapati, U. Garain, Baseline bert models for conversational hate speech detection in code-mixed tweets utilizing data augmentation and offensive language identification in marathi, in: Fire, 2022. URL: https://api.semanticscholar.org/CorpusID:259123570.

[2] K. Ghosh, D. A. Senapati, Hate speech detection: a comparison of mono and multilingual transformer model with cross-language evaluation, in: Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation, De La Salle University, Manila, Philippines, 2022, pp. 853–865. URL: https://aclanthology.org/2022.paclic-1.94.

[3] K. Ghosh, D. Sonowal, A. Basumatary, B. Gogoi, A. Senapati, Transformer-based hate speech detection in assamese, in: 2023 IEEE Guwahati Subsection Conference (GCON), 2023, pp. 1–5. doi:10.1109/GCON58516.2023.10183497.

[4] K. Ghosh, A. Senapati, A. S. Pal, Annihilate Hates (Task 4, HASOC 2023): Hate Speech Detection in Assamese, Bengali, and Bodo languages, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.

[5] 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.

[6] AI4Bharat, J. Gala, P. A. Chitale, R. AK, S. Doddapaneni, V. Gumma, A. Kumar, J. Nawale, A. Sujatha, R. Puduppully, V. Raghavan, P. Kumar, M. M. Khapra, R. Dabre, A. Kunchukuttan, Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages, arXiv preprint arXiv: 2305.16307 (2023).

[7] S. Doddapaneni, R. Aralikatte, G. Ramesh, S. Goyal, M. M. Khapra, A. Kunchukuttan, P. Kumar, Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages, ArXiv abs/2212.05409 (2022).

[8] D. Kakwani, A. Kunchukuttan, S. Golla, G. N.C., 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, Association for Computational Linguistics, Online, 2020, pp. 4948–4961. URL: https://aclanthology.org/2020.findings-emnlp.445. doi:10.18653/v1/2020.findings-emnlp.445.

[9] S. Doddapaneni, R. Aralikatte, G. Ramesh, S. Goyal, M. M. Khapra, A. Kunchukuttan,

P. Kumar, Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages, 2023. arXiv:2212.05409.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[11] G. Ramesh, S. Doddapaneni, A. Bheemaraj, M. Jobanputra, R. AK, A. Sharma, S. Sahoo, H. Diddee, M. J, D. Kakwani, N. Kumar, A. Pradeep, S. Nagaraj, K. Deepak, V. Raghavan, A. Kunchukuttan, P. Kumar, M. S. Khapra, Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages, Transactions of the Association for Computational Linguistics 10 (2022) 145–162. URL: https://aclanthology.org/2022.tacl-1.9. doi:10.1162/tacl_a_00452.

[12] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).