Hate Speech and Offensive Content Detection in Indo-Aryan Languages: A Battle of LSTM and Transformers

Nikhil Narayan[1], Mrutyunjay Biswal[1], Pramod Goyal[1] and Abhranta Panigrahi[1]

[1]Z-AGI Labs, India

Abstract

Social media platforms serve as accessible outlets for individuals to express their thoughts and experiences, resulting in an influx of user-generated data spanning all age groups. While these platforms enable free expression, they also present significant challenges, including the proliferation of hate speech and offensive content. Such objectionable language disrupts objective discourse and can lead to radicalization of debates, ultimately threatening democratic values. Consequently, social media platforms have taken steps to monitor and curb abusive behavior, necessitating automated methods for identifying suspicious posts. This paper contributes to Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC) 2023 shared tasks track for Hate Speech Detection in Low-Resource Languages. We, team Z-AGI Labs, conduct a comprehensive comparative analysis of hate speech classification across five distinct languages—Bengali, Assamese, Bodo, Sinhala, and Gujarati—within the context of the HASOC competition. Our study encompasses a wide range of pre-trained models, including Bert variants, XLM-R, and LSTM models, to assess their performance in identifying hate speech across these languages. Results reveal intriguing variations in model performance. Notably, Bert Base Multilingual Cased emerges as a strong performer across languages, achieving an F1 score of 0.67027 for Bengali and 0.70525 for Assamese. At the same time, it significantly outperforms other models with an impressive F1 score of 0.83009 for Bodo. In Sinhala, XLM-R stands out with an F1 score of 0.83493, whereas for Gujarati, a custom LSTM-based model outshined with an F1 score of 0.76601. This study offers valuable insights into the suitability of various pre-trained models for hate speech detection in multilingual settings. By considering the nuances of each, our research contributes to an informed model selection for building robust hate speech detection systems.

Keywords

Multilingual Models, Low-Resource Languages, Hate Speech, Indic Languages, HASOC-FIRE, CEUR-WS

1. Introduction

In the era of expanding global connectivity through social media, platforms such as Facebook, X (Formerly Twitter), YouTube, and Instagram have grappled with a disturbing surge in hate speech perpetuated by individuals and organized groups. With the surge of Influencers and Content Creators, also commonly known as the Creator Economy[1], there has been an alarming rise in incidents concerning targeted attacks on individuals based on their opinions, appearance, and ethnicity[2]. The consequences of such pervasive abusive language are far-reaching, often

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

nikhilnarayan73@gmail.com (N. Narayan); mrutyunjay.biswal.hmu@gmail.com (M. Biswal); goyalpramod1729@gmail.com (P. Goyal); abhranta.panigrahi@gmail.com (A. Panigrahi)

resulting in public humiliation and significant personal and professional consequences[3, 4, 5] for victims. The escalating prevalence of online hate speech, often characterized by anonymity, scale, and overwhelming volumes that challenge human moderators, underscores the pressing need for social media platforms to strike a balance between safeguarding freedom of expression and fostering an environment of inclusiveness and respect.

The need for content moderation[6] by detecting Hate and Offensive engagement has pushed organizations and research groups to develop systems and solutions at scale. Significant work has been done to identify toxic, profane, and offensive comments. However, a majority of contributions focus predominantly on Resource-heavy languages such as English[7, 8, 9, 10, 11, 12, 13]. This brings constraints to hate and offensive speech content detection and moderation in low-resource languages. The lack of large-scale corpus and pre-trained models makes it extremely difficult to tackle Natural Language Understanding (NLU) downstream tasks.

In response to these challenges, Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC) presented four shared tasks as a part of its 5th edition, 2023. Out of these, Task 1 and Task 4[14] focus on detecting hate and offensive content in 5 low-resource languages as follows:- Bodo[15], Bengali[16], and Assamese[17] in Task 4[18], Gujarati and Sinhala in Task 1[19]. Each language has its corresponding dataset, evaluation metric, competition page, and leaderboard. The challenges present one problem statement, that is to classify a given content into one of the following classes:

- **HOF** (Hate and/or Offensive): This content is hate speech, offensive, and/or profane.
- **NOT** (Non-Hate and/or Offensive): This content is not hate speech, offensive, and/or profane.

In this paper, we describe our approach to tackle the challenges. Here's the summary of our contribution:

- Adequate preprocessing techniques for datasets.
- Provide a strong LSTM with an attention head baseline model.
- Comparative analysis of pre-trained models in zero-shot and few-shot settings.
- Fine-tuning large multi-lingual models on the given datasets.

From here, the report continues in the following manner: In section 2, we highlight previous approaches as related work. In section 3, we give an overview of the dataset for each language and describe the challenge at hand. In section 4, we present our approach in detail, covering the nitty gritty of our experimental set-up, cross-validation strategy, models used, and intuition behind them. In section 5, we brief the results from the experiments section. Then, we conclude in section 6 with the final takeaways, our standings, and the scope of future work. The implementation details can be found in the following GitHub repository[1].

2. Related Work

Detecting hate speech poses a formidable challenge in the realm of research, with existing literature encompassing diverse methodologies, such as dictionary-based approaches[20], the

[1]https://github.com/The-Originalz/fire-hasoc-2023

utilization of distributional semantics[21], and the recent exploration of the efficacy of neural network architectures[22]. However, it is notable that a substantial portion of this research predominantly focuses on hate speech detection within the English language. Conversely, there is limited scholarly attention directed towards other foreign languages[23, 24, 25, 26] and the intricacies of code-switched text[27, 28]. Despite the significant impact of regional low-resource languages on online hate speech, this domain remains relatively uncharted, with recent investigations exploring the utility of transformers[29] and author profiling through the application of graph neural networks[30].

Historically, numerous strategies have been employed to address the challenge of identifying hate speech. Kwok and Wang[31] experimented with a straightforward bag of words (BOW) methodology to detect hate speech, but these lightweight models yielded subpar results characterized by elevated false positive rates. Enhancing these models with various fundamental natural language processing (NLP) components, such as part-of-speech tags[32] and N-gram graphs, contributed to improved performance. Lexical techniques employing TF-IDF in conjunction with Support Vector Machines (SVM) as a classification model surprisingly achieved commendable results[33].

The advent of embedding words into distributed representations marked a pivotal shift, as researchers harnessed word embeddings like Glove[34] and FastText[35] to project discrete text into a latent space, surpassing the performance of conventional BOW and lexical approaches.

Recurrent Neural Networks (RNNs) remained the go-to method for tackling various natural language challenges over an extended period. For instance, the winning approach in the 2020 HASOC competition for Hindi[36] employed a one-layer Bidirectional Long Short-Term Memory (BiLSTM) model with FastText embeddings to discern hate speech. Likewise, the most accurate model for English[37] adopted an LSTM architecture with Glove embeddings to represent textual inputs. Mohtaj et al.[38] also embraced a character-based LSTM, aligning with this prevailing trend.

In recent times, self-attention-based transformer models[29], and their derivatives such as BERT[39], derived from extensive corpus-trained encoders, have exhibited greater potential than traditional RNNs across a multitude of NLP tasks. BERT-like models have garnered substantial attention due to their remarkable transfer learning capabilities, outperforming alternative approaches consistently[40].

Despite the substantial body of research on hate speech detection, experiments dedicated to low-resource languages remain relatively scarce. Notably, simple logistic regression using LASER embeddings demonstrated superior performance to BERT-based models[41], underscoring the necessity for more precise multilingual base language models. Consequently, we have witnessed the ascendancy of multilingual language models like XLM-Roberta[42]. Following the trend, region-specific low-resource language models are developed. Some of the notable contributions are MuRIL[43], SinBERT[44] for Sinhala, BanglaBERT[45], Indic-BERT[46], and XLM-Indic[47] variants. Authors in [48] provide a detailed study on performance of monolingual and multi-lingual in the context of cross-lingual evaluation for hate speech identification. Previous editions of HASOC[49, 50] have witnessed a significant effort towards improving performance in low-resource languages such as Hindi[51], Marathi[52], etc. In the ensuing sections, we will elucidate our approach, which leverages several multilingual models for hate speech identification, accompanied by an exhaustive comparative analysis against alternative

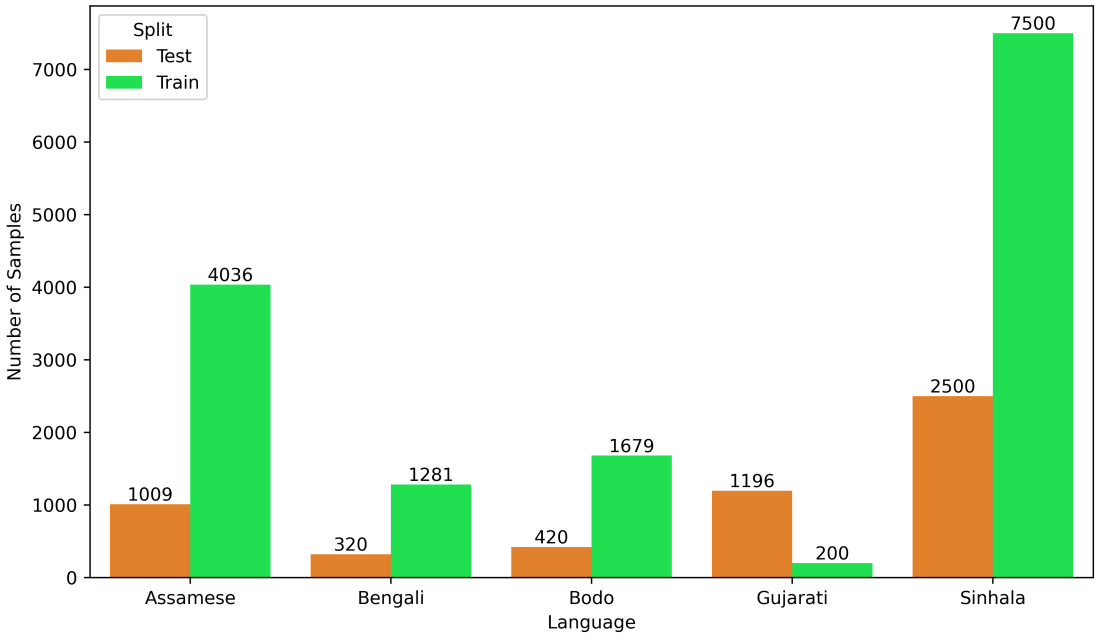

Figure 1: Train-Test Split for the respective Tasks

methodologies.

3. Dataset Description

3.1. Sinhala

Task 1 consists of 2 sub-tasks, one for Sinhala, and the other for Gujarati. Sinhala, one of the two official languages in Sri Lanka, is spoken by over 17 million people. This edition of HASOC brings the first-ever shared task for the aforementioned Indo-Aryan low-resource language. The train and test sets for Sinhala are based on the SOLD: Sinhala Offensive Language Detection dataset[53]. The SOLD consists of 10,000 manually annotated tweets divided into two classes: Offensive and Not offensive, both at the token level and the sentence level. However, the dataset provided for the task contains 7500 samples in the train set, each labeled as HOF or NOT. The test set contains 2500 samples.

3.2. Gujarati

Gujarati is one of the 22 official languages of India with over 50M native speakers. The train set for this task contains 200 tweets whereas the test set contains 1196 tweets. This is also a coarse-grained binary classification problem but in a few-shot setting. The train data frame is made up of 5 columns, named as follows: tweet_id, created_at, text, user_screen_time, and label. The test set contains only tweet_id and the text column.

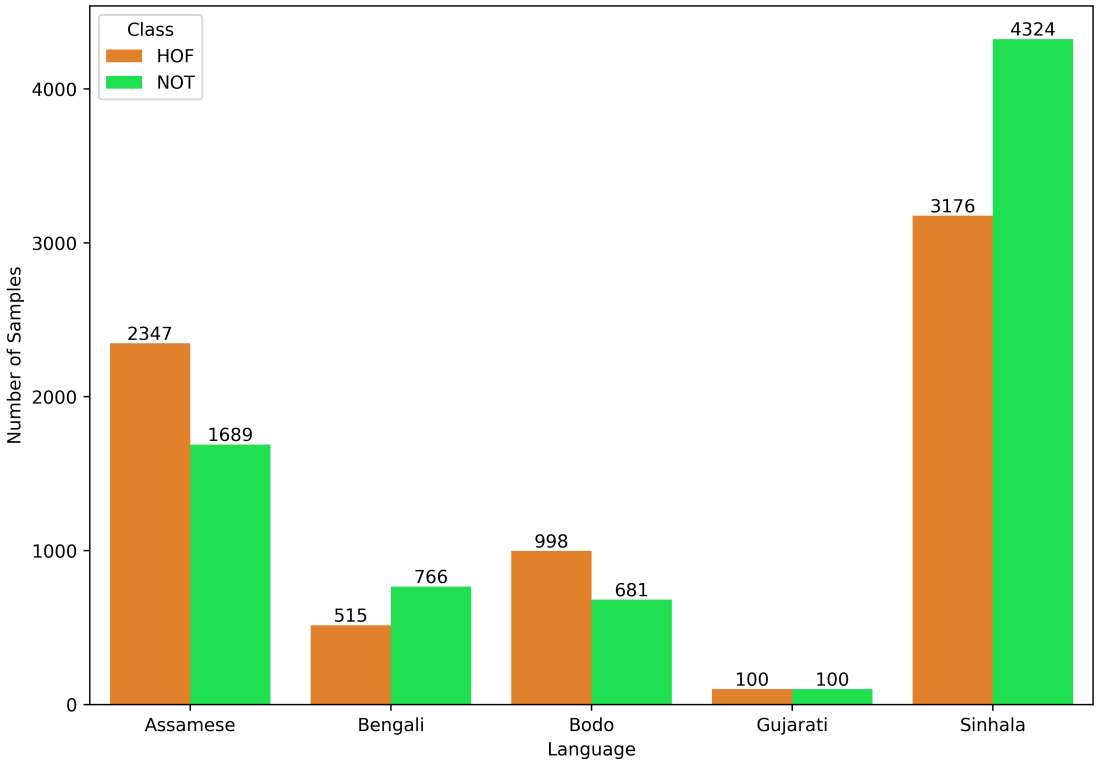

Figure 2: Target Distribution for the respective tasks.

3.3. Assamese, Bodo, and Bengali

Task 4 consists of 3 Kaggle competitions, each corresponding to one of the following languages: Assamese, Bodo, and Bengali. The primary sources for data collection are X (formerly Twitter), Facebook, and YouTube Comments. Each train set contains tweet-label pairs, with only tweets in the test set to predict the targets.

4. Experimental Set-up

In this section, we discuss our approach and explain the experimental set-up details. We start with creating a validation strategy for each dataset. As all the datasets are fairly balanced (refer to figure 3.3), we opt for K-Fold cross-validation with 5 folds. And while creating the splits, we set the random seed to 2023.

4.1. Preprocessing

The preprocessing step involves cleaning the contents for feature extraction. We notice that the Bodo tweets do not contain any emoji or repetitive punctuation marks that need attention.

However, there are instances where some of the content contains English words in a code-mixed manner. Note that, such instances are seen across all the datasets. For the other two datasets in Task 4, though the majority of the content is cleaned, some of them contain emojis and a few repetitive punctuation characters. The repetition is removed, and the emojis are converted into their respective textual description using the emot[2] library. The Task 1 datasets for Sinhala and Gujarati contain usernames as @USER and the usernames are made available in a separate column. Along with that, the datasets also contain code-mixed English words, repetitive punctuation characters, emojis, and hashtags. Note that, there are no URLs or hyperlinks associated with any of the content. The usernames are removed along with repetitive punctuation characters. The hashtags are further processed with Ekphrasis[3] tokenizer to segment them into meaningful tokens. The emoji2vec[54] embeddings are used on top of emojis, as it has shown competitive results in previous works.

4.2. Modeling

To start with, we create an LSTM-attention-based baseline with 2 bi-direction LSTM layers followed by an attention block. The attention head is further connected through 2 dense layers with sigmoid activation in the last layer. The model is trained with Adam optimizer and Binary Cross Entropy as the loss function. The hyperparameters involved such as the batch size, number of epochs, learning rate, vocab size, embedding dimension, and maximum length of the input sequence are varied and tuned on a case-to-case basis.

As the available datasets have less number of samples per language (refer to figure3.1), we also leverage Transformer-based language models for the downstream task at hand. The available training data are used to fine-tune the encoder layers of the transformer-based models, leaving the embedding layers frozen. Note that to incorporate emojis semantics, we concatenate the embedding layers with the emoji2vec-generated embeddings. We experiment with various multi-lingual transformer-based models for fine-tuning such as Bert-Base-Multilingual (Cased and Uncased), DistilBert-Base-Multilingual-Cased, XLM-Roberta-Base, Muril-Base, and XLM-Indic-Base (UniScript[4] and Multi-Script[5]). Other than that, we also present the results from a couple of language-specific models such as Bangla-BERT and SinhalaBERTo[6].

Note that, the Huggingface implementation for the models[7] is used via TFAutoModelForSequenceClassification with corresponding hyperparameters for each. All the training and inference are done using Kaggle runtime and MacBook Air M2 with 24GB unified memory.

For inference, we ensemble the models from each fold with equal weightage on the logits, then take a threshold of 0.5 to classify into HOF or NOT. The labels are mapped to numbers as follows: HOF is mapped to 0, and NOT is mapped to 1.

[2]https://github.com/NeelShah18/emot
[3]https://github.com/cbaziotis/ekphrasis
[4]https://huggingface.co/ibraheemmoosa/xlmindic-base-uniscript
[5]https://huggingface.co/ibraheemmoosa/xlmindic-base-multiscript
[6]https://huggingface.co/keshan/SinhalaBERTo
[7]https://huggingface.co/models

Table 1
Model aliases used a reference in the result graphs.

Model Name	Alias Name
LSTM Baseline	LSTM
Bert Base Multilingual Cased	mBert C
Bert Base Multilingual Uncased	mBert U
DistilBert Base Multilingual Cased	mDistil C
XLM Roberta Base	XLM-R
Muril Base Cased	MuRIL
XLM Indic Base Multiscript	XLM-I M
XLM Indic Base Uniscript	XLM-I U

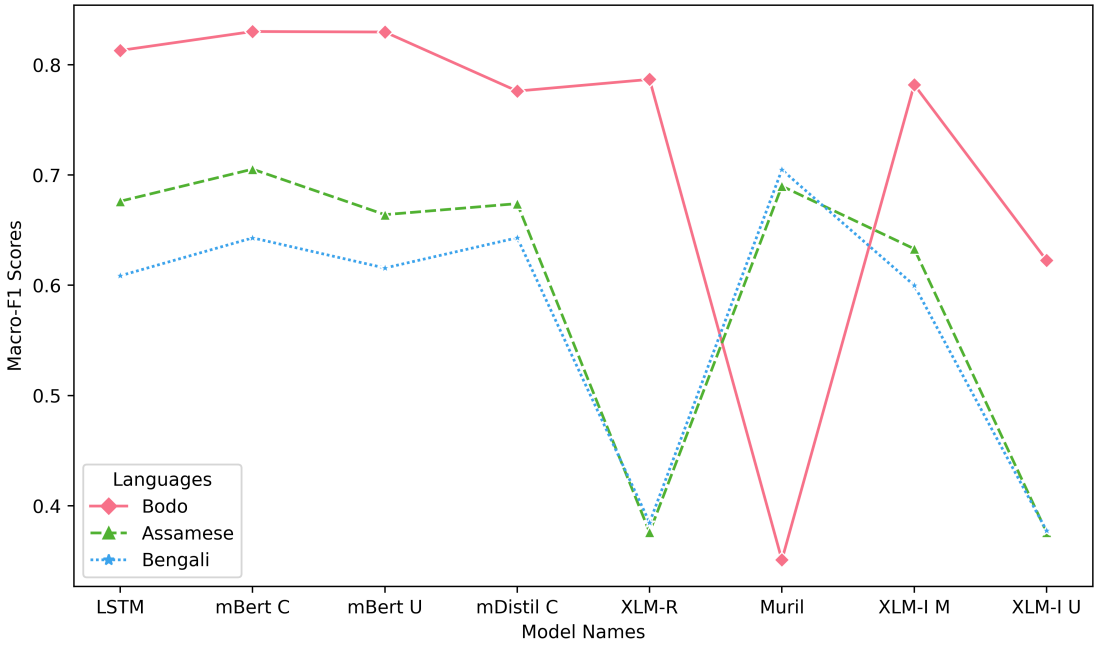

Figure 3: Leaderboard macro-F1 scores of different models for respective languages in Task 4.

5. Results

The competitions for Task 4 are evaluated on macro-f1 metrics, whereas the Task 1 challenges are evaluated on macro-f1, Precision, and Recall.

It is evident from the results matrix (refer to table 3) that, the LSTM baseline poses a strong competition in performance for all the languages. Even going a step further for Gujarati, the LSTM-based Model scores highest amongst XLM-Indic-Base-MultiScript and Bert-Base-Multilingual. This results in the highest F1-Score of 0.76601 and a Recall of 0.79704, with a marginal gain of 0.001 for F1 and 0.03 for Recall against the second-best performing model for

Table 2
Leaderboard macro-f1 scores in Task 4.

Models	Bodo	Assamese	Bengali
LSTM Baseline	0.81291	0.67616	0.60856
Bert Base Multilingual Cased	**0.83009**	**0.70525**	0.64294
Bert Base Multilingual Uncased	0.82962	0.66398	0.61561
DistilBert Base Multilingual Cased	0.77606	0.67399	0.643
XLM Roberta Base	0.78668	0.376	0.38476
Muril Base Cased	0.35085	0.68985	0.70467
XLM Indic Base Multiscript	0.78186	0.633	0.59989
XLM Indic Base Uniscript	0.62257	0.376	0.37743
csebuetnlp/banglabert	-	-	**0.75625**

Table 3
Leaderboard macro-f1 scores in Task 1.

Models	Gujarati	Sinhala
LSTM Baseline	**0.7660**	0.7530
Bert Multilingual base Cased	0.7656	0.8095
XLM Roberta Base	-	**0.8349**
XLM Indic Base Multiscript	0.7235	-

the task. For Sinhala, XLM-Roberta seems to be the winner beating our LSTM Baseline and Sinhala Bert with a considerable margin. Due to time constraints and run submission limits, we experimented with a handful of BERT-based and Roberta-based models for fine-tuning along with the LSTM-with-Attention baseline Model.

For Task 4, we have varied candidate models for experimentation as mentioned in section 4 (refer to table 1). Our LSTM baseline poses as one of the top performers for Bodo and Assamese by yielding the third-highest F1-Score, beating XLM-Roberta-based models. For Bengali, however, BanglaBert beat the rest of the models with a staggering 0.75625 f1-score. The best-performing model for Bodo and Assamese is Bert-Base-Multilingual-Cased with an f1-score of 0.83009, and 0.70525 respectively. Note that, all the scores mentioned above (refer to table2) are the performance on the hidden test set, and directly taken from the system-run report provided by the Organizers after a finalized leaderboard. A visual representation of comparative performance for Task 4 is shown in figure 5.

The obtained results help us climb to 3rd/20 for Bengali, 5th/16 for Sinhala, 7th/17 for Gujarati, 8th/20 for Assamese, and 12th/19 for Bodo.

6. Conclusion

This work has been submitted to the CEUR-2023 Workshop Proceedings for Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC) track. In this report, we entail our approach to solving two tasks for the Track on classifying a given content

into Hate and Offensive (HOF) or NOT. We experiment with various models starting from simple LSTM-based architecture to pretrained transformer-based multilingual models. Though the transformer shows sheer dominance in the majority of the tasks, our LSTM baseline emerges as a strong competitor with promising results, even resulting in the highest f1-score amongst our candidate models for Gujarati. We plan to further extend our work to other low-resource indic languages, and explore the possibilities of zero-shot, few-shot, and cross-lingual transfer learning scenarios. We also aim to develop a unified language model to incorporate all the languages such as NLLB[55] for similar downstream tasks to strengthen content moderation.

References

[1] Wikipedia contributors, Creator economy — Wikipedia, the free encyclopedia, 2023. URL: https://en.wikipedia.org/w/index.php?title=Creator_economy&oldid=1159137601.

[2] Taniya Roy, Hindutva's circulation of anti-muslim hate aided by digital platforms, finds report, 2022. URL: https://thewire.in/communalism/india-anti-muslim-hate-twitter-facebook-whatsapp-hindutva-modi-bjp.

[3] K. E. Riehm, K. A. Feder, K. N. Tormohlen, R. M. Crum, A. S. Young, K. M. Green, L. R. Pacek, L. N. La Flair, R. Mojtabai, Associations Between Time Spent Using Social Media and Internalizing and Externalizing Problems Among US Youth, JAMA Psychiatry 76 (2019) 1266–1273. URL: https://doi.org/10.1001/jamapsychiatry.2019.2325. doi:10.1001/jamapsychiatry.2019.2325.

[4] J. Naslund, A. Bondre, J. Torous, K. Aschbrenner, Social media and mental health: Benefits, risks, and opportunities for research and practice, Journal of Technology in Behavioral Science 5 (2020). doi:10.1007/s41347-020-00134-x.

[5] R. Bannink, S. Broeren, P. Looij-Jansen, F. Waart, H. Raat, Cyber and traditional bullying victimization as a risk factor for mental health problems and suicidal ideation in adolescents, PloS one 9 (2014) e94026. doi:10.1371/journal.pone.0094026.

[6] Dylan Walsh, As content booms, how can platforms protect kids from hateful speech?, 2022. URL: https://mitsloan.mit.edu/ideas-made-to-matter/content-booms-how-can-platforms-protect-kids-hate-speech.

[7] H. H. Saeed, K. Shahzad, F. Kamiran, Overlapping toxic sentiment classification using deep neural architectures, in: 2018 IEEE International Conference on Data Mining Workshops (ICDMW), 2018, pp. 1361–1366. doi:10.1109/ICDMW.2018.00193.

[8] A. Vaidya, F. Mai, Y. Ning, Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection, Proceedings of the International AAAI Conference on Web and Social Media 14 (2020) 683–693. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/7334. doi:10.1609/icwsm.v14i1.7334.

[9] S. M. Carta, A. Corriga, R. Mulas, D. R. Recupero, R. Saia, A supervised multi-class multi-label word embeddings approach for toxic comment classification, in: International Conference on Knowledge Discovery and Information Retrieval, 2019. URL: https://api.semanticscholar.org/CorpusID:204754719.

[10] T. Tran, Y. Hu, C. Hu, K. Yen, F. Tan, K. Lee, S. Park, Habertor: An efficient and effective deep hatespeech detector, 2020. arXiv:2010.08865.

[11] O. Kamal, A. Kumar, T. Vaidhya, Hostility detection in hindi leveraging pre-trained language models, 2021. arXiv:2101.05494.
[12] Hitkul, K. Aggarwal, P. Bamdev, D. Mahata, R. R. Shah, P. Kumaraguru, Trawling for trolling: A dataset, 2020. arXiv:2008.00525.
[13] A.-M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, N. Kourtellis, Large scale crowdsourcing and characterization of twitter abusive behavior, 2018. arXiv:1802.00393.
[14] 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.
[15] K. G. Aditya Shankar Pal, Annihilate hates (bodo), 2023. URL: https://kaggle.com/competitions/annihilate-hates-bodo.
[16] K. Ghosh, Annihilate hates (bengali), 2023. URL: https://kaggle.com/competitions/annihilate-hates-bengali.
[17] K. G. Aditya Shankar Pal, Annihilate hates (assamese), 2023. URL: https://kaggle.com/competitions/annihilate-hates-assamese.
[18] 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.
[19] 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.
[20] R. Guermazi, M. Hammami, A. B. Hamadou, Using a semi-automatic keyword dictionary for improving violent web site filtering, in: 2007 Third International IEEE Conference on Signal-Image Technologies and Internet-Based System, 2007, pp. 337–344. doi:10.1109/SITIS.2007.137.
[21] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, N. Bhamidipati, Hate speech detection with comment embeddings, in: Proceedings of the 24th International Conference on World Wide Web, WWW '15 Companion, Association for Computing Machinery, New York, NY, USA, 2015, p. 29–30. URL: https://doi.org/10.1145/2740908.2742760. doi:10.1145/2740908.2742760.
[22] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in: Proceedings of the 26th International Conference on World Wide Web Companion, WWW '17 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2017, p. 759–760. URL: https://doi.org/10.1145/3041021.3054223. doi:10.1145/3041021.3054223.
[23] J. A. Leite, D. Silva, K. Bontcheva, C. Scarton, Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing,

Association for Computational Linguistics, Suzhou, China, 2020, pp. 914–924. URL: https://aclanthology.org/2020.aacl-main.91.

[24] A. Saroj, S. Pal, An Indian language social media collection for hate and offensive speech, in: Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 2–8. URL: https://aclanthology.org/2020.restup-1.2.

[25] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. URL: https://aclanthology.org/S19-2007. doi:10.18653/v1/S19-2007.

[26] A. Ghosh Chowdhury, A. Didolkar, R. Sawhney, R. R. Shah, ARHNet - leveraging community interaction for detection of religious hate speech in Arabic, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, Florence, Italy, 2019, pp. 273–280. URL: https://aclanthology.org/P19-2038. doi:10.18653/v1/P19-2038.

[27] S. Chopra, R. Sawhney, P. Mathur, R. Ratn Shah, Hindi-english hate speech detection: Author profiling, debiasing, and practical perspectives, Proceedings of the AAAI Conference on Artificial Intelligence 34 (2020) 386–393. URL: https://ojs.aaai.org/index.php/AAAI/article/view/5374. doi:10.1609/aaai.v34i01.5374.

[28] R. Kapoor, Y. Kumar, K. Rajput, R. R. Shah, P. Kumaraguru, R. Zimmermann, Mind your language: Abuse and offense detection for code-switched languages, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) 9951–9952. URL: https://ojs.aaai.org/index.php/AAAI/article/view/5112. doi:10.1609/aaai.v33i01.33019951.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[30] T. Ranasinghe, S. Gupte, M. Zampieri, I. Nwogu, Wlv-rit at hasoc-dravidian-codemix-fire2020: Offensive language identification in code-switched youtube comments, 2020. arXiv:2011.00559.

[31] I. Kwok, Y. Wang, Locate the hate: Detecting tweets against blacks, in: Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, AAAI'13, AAAI Press, 2013, p. 1621–1622.

[32] Y. Chen, Y. Zhou, S. Zhu, H. Xu, Detecting offensive language in social media to protect adolescent online safety, in: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, 2012, pp. 71–80. doi:10.1109/SocialCom-PASSAT.2012.55.

[33] R. Rajalakshmi, Y. Reddy, Dlrg@hasoc 2020: A hybrid approach for hate and offensive content identification in multilingual tweets, in: Fire, 2020. URL: https://api.semanticscholar.org/CorpusID:232314467.

[34] J. Pennington, R. Socher, C. Manning, GloVe: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing

(EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. URL: https://aclanthology.org/D14-1162. doi:10.3115/v1/D14-1162.

[35] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146. URL: https://aclanthology.org/Q17-1010. doi:10.1162/tacl_a_00051.

[36] R. Raj, S. Srivastava, S. Saumya, Nsit & iiitdwd @ hasoc 2020: Deep learning model for hate-speech identification in indo-european languages, in: Fire, 2020. URL: https://api.semanticscholar.org/CorpusID:232313876.

[37] A. K. Mishra, S. Saumya, A. Kumar, Iiit_dwd@hasoc 2020: Identifying offensive content in indo-european languages, in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 139–144. URL: https://ceur-ws.org/Vol-2826/T2-5.pdf.

[38] S. Mohtaj, V. Woloszyn, S. Möller, Tub at hasoc 2020: Character based lstm for hate speech detection in indo-european languages, in: Fire, 2020. URL: https://api.semanticscholar.org/CorpusID:232314731.

[39] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.

[40] M. Mozafari, R. Farahbakhsh, N. Crespi, A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media, 2019, pp. 928–940. doi:10.1007/978-3-030-36687-2_77.

[41] S. S. Aluru, B. Mathew, P. Saha, A. Mukherjee, Deep learning models for multilingual hate speech detection, 2020. arXiv:2004.06465.

[42] 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, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

[43] S. Khanuja, D. Bansal, S. Mehtani, S. Khosla, A. Dey, B. Gopalan, D. K. Margam, P. Aggarwal, R. T. Nagipogu, S. Dave, et al., Muril: Multilingual representations for indian languages, arXiv preprint arXiv:2103.10730 (2021).

[44] V. Dhananjaya, P. Demotte, S. Ranathunga, S. Jayasena, BERTifying Sinhala - a comprehensive analysis of pre-trained language models for Sinhala text classification, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7377–7385. URL: https://aclanthology.org/2022.lrec-1.803.

[45] T. Hasan, A. Bhattacharjee, K. Samin, M. Hasan, M. Basak, M. S. Rahman, R. Shahriyar, Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation, in: Proceedings of the 2020 Conference on Empirical Methods

in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 2612–2623. URL: https://aclanthology.org/2020.emnlp-main.207. doi:10.18653/v1/2020.emnlp-main.207.

[46] 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.

[47] I. M. Moosa, M. E. Akhter, A. B. Habib, Does transliteration help multilingual language modeling?, in: Findings of the Association for Computational Linguistics: EACL 2023, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 670–685. URL: https://aclanthology.org/2023.findings-eacl.50. doi:10.18653/v1/2023.findings-eacl.50.

[48] 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.

[49] S. Satapara, P. Majumder, T. Mandl, S. Modha, H. Madhu, T. Ranasinghe, M. Zampieri, K. North, D. Premasiri, Overview of the hasoc subtrack at fire 2022: Hate speech and offensive content identification in english and indo-aryan languages, in: Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '22, Association for Computing Machinery, New York, NY, USA, 2023, p. 4–7. URL: https://doi.org/10.1145/3574318.3574326. doi:10.1145/3574318.3574326.

[50] S. Modha, T. Mandl, G. K. Shahi, H. Madhu, S. Satapara, T. Ranasinghe, M. Zampieri, Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech, in: Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '21, Association for Computing Machinery, New York, NY, USA, 2022, p. 1–3. URL: https://doi.org/10.1145/3503162.3503176. doi:10.1145/3503162.3503176.

[51] M. Bhatia, T. S. Bhotia, A. Agarwal, P. Ramesh, S. Gupta, K. Shridhar, F. Laumann, A. Dash, One to rule them all: Towards joint indic language hate speech detection, 2021. arXiv:2109.13711.

[52] T. Ranasinghe, K. North, D. Premasiri, M. Zampieri, Overview of the hasoc subtrack at fire 2022: Offensive language identification in marathi, 2022. arXiv:2211.10163.

[53] T. Ranasinghe, I. Anuradha, D. Premasiri, K. Silva, H. Hettiarachchi, L. Uyangodage, M. Zampieri, Sold: Sinhala offensive language dataset, 2022. arXiv:2212.00851.

[54] B. Eisner, T. Rocktäschel, I. Augenstein, M. Bošnjak, S. Riedel, emoji2vec: Learning emoji representations from their description, in: Proceedings of the Fourth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Austin, TX, USA, 2016, pp. 48–54. URL: https://aclanthology.org/W16-6208. doi:10.18653/v1/W16-6208.

[55] N. Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk,

J. Wang, No language left behind: Scaling human-centered machine translation, 2022. arXiv:2207.04672.