A Corpus of Hindi-English Code-Mixed Posts to Hate Speech Detection

Prashant Kapil *
SCSET
Bennett University
prashant.kapil@bennett.edu.in

Asif Ekbal
Department of CSE
IIT Patna
asif@iitp.ac.in

Abstract

Social media content, such as blog posts, comments, and tweets, often contains offensive language, including racial hate speech, personal attacks, and sexual harassment. Detecting inappropriate language is crucial for user safety and prevention of hateful behavior and aggression. This study introduces the HECM (Hindi-English code-mixed tweets) to fill the gap in Hindi language resources. The corpus comprises approximately 9.4K tweets labeled as hateful and nonhateful. It includes detailed information on the data, such as the annotation schema, the label definitions, and an interannotator agreement score of 85%. The study evaluates the effectiveness of traditional machine learning, deep neural networks, and transformer encoder-based approaches. The results show a significant improvement in terms of macro-F1 and weighted F1 scores. Additionally, a lexicon containing 2000 lexicons tagged in 21 categories is created based on the multilingual HURTLEX lexicon. This lexicon is merged with the transformer encoder, resulting in a marginal improvement in macro-F1 and weighted-F1. The study also experiments with a Hindi-Devanagari dataset, HHSD, to assess the impact of the lexicon on performance metrics. The code and lexicon are available at https://github.com/imprasshant/ICON

1 Introduction

Social networks have become an intrinsic part of the lives of many people. A phenomenon that accompanies social media platforms with serious impacts on society is the presence of socially unacceptable language. Socially unacceptable language comprehends many different user-generated language phenomena, such as toxic language, offensive language, abusive language, and hate speech, among others. An approach to filter offensive content is to use human experts (e.g., moderators) and manually review the posts or comments as soon as they get posted. There has been a growing interest in computational linguistics (CL) and natural language processing (NLP), as manually monitoring and flagging these phenomena is impossible. The automatic identification of abusive language phenomena has followed a common trend in NLP: feature-based linear classifiers ((Bohra et al., 2018), (Maitra and Sarkhel, 2018),(Risch and Krestel, 2018), (Samghabadi et al., 2018)) neural network architectures ((Modha et al., 2018),(Raiyani et al., 2018), (Jha et al., 2020), (Baruah et al., 2020),(Bashar and Nayak, 2020),) (e.g., CNN or Bi-LSTM), and fine-tuning pre-trained language models, e.g., BERT, RoBERTa ((Mishra and Mishra, 2019), (Velankar et al., 2021), (Gupta et al., 2022), (Kapil et al., 2023)). This paper focuses on detecting offensive language in Hindi and Hindi-English code-mixed posts. Although there are numerous studies on automatic detection of offensive content in resource-rich languages such as English ((Davidson et al., 2017), (Waseem and Hovy, 2016), (Zampieri et al., 2019), (de Gibert et al., 2018), (Founta et al., 2018)), there are limited data and work available for a resource-poor Hindi language. The key contributions of this paper are listed below.

(i) Dataset: A new dataset called Hindi-English Code-Mixed (HECM) has been created by labeling 9.4K posts as hateful or non-hateful. This dataset will be shared with the research community.

*This work was conducted at IIT Patna.

Additionally, a lexicon has been developed based on HURTLEX, consisting of 2000 offensive words across 21 categories. The experiment also utilizes the existing dataset HHSD (Kapil et al., 2023).

(ii) Model: The experiments are conducted using numerous cutting-edge models, such as support vector machine (SVM), convolution neural network (CNN), multilingual-bert (M-BERT), multilingual representations for Indian languages (MuRIL), and XLM-RoBERTa. The experiment is done on HECM, HHSD, and HECM + HHSD. The lexicon features are infused with an encoder to enhance the performance.

(iii) Analysis: The effectiveness of the models is evaluated using a 5-fold cross-validation approach.

2 Related Work

The corpus for the Hindi covers both single-layer (Bohra et al., 2018) (Kumar et al., 2018) (Mathur et al., 2018) (Jha et al., 2020) (Bhardwaj et al., 2020) and multi-layer (Mandl et al., 2019) (Mandl et al., 2020) (Mandl et al., 2021)(Kapil et al., 2023) (Bhattacharya et al., 2020) textual-tagged data. (Kapil et al., 2023) crawled Hindi data in Devanagari script and tagged it hierarchically, covering multiple layers of hate. The experiment involved using M-BERT and MuRIL for multitask learning with related datasets from the English, Hindi, and Urdu domains, resulting in significant accuracy and f-score on the single-task learning framework.

The multi-channel multichannel transfer learning-based model (MIMCT) described in (Mathur et al., 2018) uses several feature inputs in conjunction with transfer learning to identify offensive (hate speech or abusive) Hinglish tweets from the proposed Hinglish Offensive Tweet (HOT) dataset. In order to address hate and offensive detection on the data by (Modha et al., 2021), (Velankar et al., 2021) investigated deep learning architectures such as CNN, LSTM, and variants of BERT like M-BERT, IndicBERT, and monolingual RoBERTa. A complementary approach to supervised learning towards the detection of abusive and offensive language is the use of language resources such as lexicons and dictionaries. (Bassignana et al., 2018) describe the creation of HurtLex, a multilingual lexicon of hate words. (Koufakou et al., 2020) proposes to utilize lexical features derived from a hate lexicon towards improving the performance of BERT in such tasks.

3 Corpus Creation

3.1 Data Crawling and Processing

The proposed data set is constructed from Hindi-English tweets crawled using the Twitter search API[1]. The findings from earlier studies by (Wiegand et al., 2018) and (Davidson et al., 2017) motivated us to create a hate speech dataset using a sampling method that requires less input. The data collection includes approximately 100,000 tweets from May 2021 to September 2021, covering keywords and topics in Roman Hindi script related to politics, religion, racism, and sexism. These topics were identified based on recent news and their potential to incite hate speech. Additionally, we gathered abusive lexicons in Roman Hindi script to identify explicit hate posts. Table 1 provides an overview of the important keywords and topics used for crawling the posts.

Selecting relevant tweets for annotation: A set of tweets is sampled out in order to choose the pertinent ones for the final annotation from the vast collection of 100K unlabeled data. Eight publicly accessible Hindi datasets (Bohra et al., 2018),(Kumar et al., 2018),(Jha et al., 2020), (Mathur et al., 2018), (Mandl et al., 2019),(Mandl et al., 2020), (Mandl et al., 2021),(Modha et al., 2021) are used to train a convolutional neural network C_i classifier. The trained models C_i process the unlabeled tweet i acquired during the crawling to produce a weak label based on the probability value p. The two sets of tweets (S_h) with p (hate)$\geq$ 0.65 and (S_{nh}) with p (non-hate)$\geq$ 0.85 were given to the annotators.

[1] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets

3.2 Preprocessing

Prior to training the models, we perform a few preprocessing steps as follows:

- All the characters like |:,? were removed along with the numbers and URLs.
- All the @ (ex. @abc) mentions were replaced with the common token, i.e user.
- Emojis: The emojis, emoticons, symbols, pictographs, transport, maps, dingbats, flags, etc. were removed.

3.3 Data Annotation

The annotation is initiated by hiring three annotators with strong linguistic and Hindi knowledge. The annotators had a higher level of education (master, PhD). Before starting the annotations, the annotators were informed of the content's offensiveness and hostility. The posts were categorized into two classes: hateful and Non-hateful.

Hateful: The Language that is intended to be disparaging, humiliating, or insulting to the members of the group or an individual based on race, gender, ethnic origin, sexual orientation, disability, religion, or colour (Davidson et al., 2017), (Founta et al., 2018))

Non-Hateful: Posts that do not contain any hateful content.

Inter Annotator Agreement: The class definitions, along with numerous examples, were provided to the annotators. Each of the three annotators received the same 200 tweets to annotate. The quality of the annotation is evaluated using the Fleiss Kappa score (Fleiss, 1971), which is a measure of inter-rater agreement used to determine the agreement among two or more raters. The Inter-annotator agreement (IAA) score obtained after annotation is 84%. Table 2 consists of the 21 categories for word-level tags, and Table 3 shows the data statistics used in the experiment.

4 Methodology

(i) Support vector machine (Cortes and Vapnik, 1995): We use scikit-learn's 4 linear SGD classifier with default hyperparameters and tf-idf weighting.

(ii) CNN: It is proposed by (Kim, 2014) and consists of five main layers: input layer, embedding layer, convolution, pooling, and fully connected layer.

(iii) M-BERT (Devlin et al., 2018): This model extracts features by employing bidirectional training of the transformer to understand the context of a word based on its surroundings, utilizing masked language modeling (MLM) and next sentence prediction (NSP).

(iv) MuRIL(Multilingual representations for Indian languages) (Khanuja et al., 2021): This language model was specifically created for Indian languages and trained using text corpora from 16 Indian languages known as "IN." The training objectives include MLM and TLM, among others.

(v) XLM-RoBERTa: (Conneau et al., 2019): This transformer model was trained by sampling streams of text in 100 languages and predicting the masked tokens in the input using the MLM objective.

The transformer encoder is enhanced with lexicon features. In the first architecture, Figure 1, we identify their categories in HurtLex and then generate a vector of HurtLex categories. This process is referred to as HurtLex encoding. The lexicon contains a total of 21 categories, so the dimensionality of the HurtLex encoding is 21. Each element in this vector represents a frequency count for the respective category in HurtLex. The second model explores the use of HurtLex embeddings with a Bi-LSTM, as depicted in Figure 2. The HurtLex embedding is a 21-dimensional one-hot encoding of the word presence in each of the lexicon categories. This model is named HurtLex Embedding.

5 Experiment Setup

The experiments were performed using a 5-fold cross-validation approach. The 4-fold training set is split into 15% validation and 85% training, while the last fold is treated as the test set to evaluate the model. All the deep learning models were implemented using Keras (Chollet et al., 2015) with Tensorflow (Abadi et al., 2016) as the backend. The number of filters used in CNN is 100, and the kernel width ranges from 1 to 4. For the BiLSTM, the number of hidden nodes is set to 100. Categorical cross-entropy is used as a loss function,

Topics
(CAA), (NRC), (article 370), (ram mandir), (beef ban), (triple talaaq), (award wapsi), (demonetization), (GST), (liquor ban), (mannkibaat), (pulwama attack), (saheenbagh), (swachh bharat), (sabrimala mandir), (fatwa), (love jihad), (AzadiMarch)

Table 1: Topics crawled to collect the HECM corpus

Categories
(Ethnic stereotype slur), (professions and occupations), (Physical disabilities and diversity), (Cognitive disabilities and diversity), (Moral and behavioral defects), (Social and economic disadvantage), (Words related to prostitution), (Obfuscation of slangs), (Animal Picturization), (Explicit slang), (Casteist), (Negative), (Threat), (Racial/Ethnic), (Sarcasm), (Negative activity), (Religion), (Emotion), (Mass protest), (places), (pronoun)

Table 2: Variants of hate attacks used to create the Lexicon

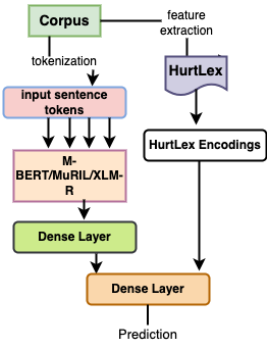

Figure 1: Encoder features +HurtLex-Encodings

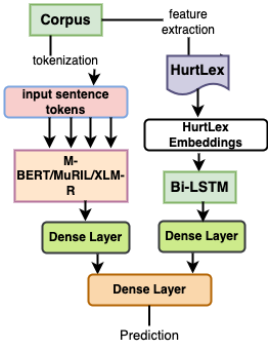

Figure 2: Encoder features +HurtLex-Embeddings

Dataset	labels and count
HHSD (Kapil et al., 2023)	Hateful: 7311 Non-Hateful: 7472
Hindi-English code mixed (HECM)	Hateful: 2823 Non-hateful: 6534
(HHSD + HECM)	Hateful: 10134 Non-Hateful: 14006

Table 3: Data statistics used in the experiment

and Adam (Kingma and Ba, 2015) optimizer is used for optimizing the network. We use a learning rate of 2e-5 for the transformer models. The batch size of 30 and an epoch of 2 are found to be optimal. The value of bias is randomly initialized to all zeros, the relu activation function is utilized at the intermediate layer, and Softmax is utilized in the last dense layer. The evaluation criteria used are macro-F1 and weighted-F1. The NVIDIA GPU is used for the evaluation.

6 Results and Analysis

Table 4 shows the results obtained on the proposed data in terms of macro-F1 and weighted-F1. For HECM, the M-BERT outperforms CNN, MuRIL, and XLM-R to obtain 86.82% macro-F1 and 86.95% weighted-F1 score. For HHSD as well, the performance of M-BERT surpasses the results obtained by CNN, MuRIL, and XLM-R to obtain an 86.16% macro-F1 and an 86.25% weighted-F1 score. The augmented data (HECM + HHSD) saw a great improvement in all the model performances. The inclusion of lexicon encodings and lexicon-embeddings to the M-BERT, MURIL, and XLM-R saw an enhancement in the macro-F1 and weighted-F1. The inclusion of lexicon embedding features outperforms all the models.

7 Conclusion and Future Works

This paper released approximately 9.4K posts tagged into hateful and non-hateful, named as HECM. Extensive experiments are conducted on HECM and the existing dataset HHSD to train SVM, CNN, M-BERT, MuRIL, and XLM-R. The experiment is also conducted over the merged data set (HECM + HHSD). The transformer encoder features are fused with lexicon features to obtain significant macro-F1 and weighted-F1. The future work intends to augment the dataset with additional boosted data. Since a lot of tweets require contextual information, localized knowledge graphs can be created for this by collecting intra-user and inter-user tweets to obtain valuable features. The contextual knowledge can easily be verified against this knowledge base.

Model	HECM		HHSD		HECM + HHSD	
	Macro(%)	Weighted(%)	Macro(%)	Weighted(%)	Macro(%)	Weighted(%)
SVM	76.12	76.88	75.52	77.92	77.61	72.23
CNN	82.23	82.77	81.28	80.99	82.61	82.83
M-BERT	86.12	86.84	85.82	85.67	87.30	87.59
MuRIL	84.42	84.62	84.50	84.46	84.89	85.35
XLM-R	79.23	79.12	75.07	75.21	79.91	79.71
M-BERT+lex-encodings	86.72	86.52	86.04	86.12	86.40	86.23
M-BERT + Lex-embeddings	86.82	86.95	86.16	86.25	87.84	87.96
MuRIL + lex-encodings 84.56	84.72	84.92	85.04	84.31	84.66	
MuRIL + lex-embeddings	84.62	84.85	85.06	85.22	84.23	84.72
XLM-R + lex-encodings	79.73	79.04	75.76	76.03	79.67	79.82
XLM-R + lex-embeddings	79.12	79.22	75.92	76.32	79.76	79.98

Table 4: Evaluation results on HECM, HHSD, and HECM + HHSD

8 Limitations

The model struggles to accurately classify hateful instances that require contextual understanding, as it lacks the ability to capture nuanced context. Additionally, the scarcity of resources for Hindi presents significant challenges in effective hate speech detection.

Acknowledgement

The Authors gratefully acknowledge the project "HELIOS - Hate, Hyperpartisan, and Hyperpluralism Elicitation and Observer System", sponsored by Wipro Ltd. Prashant Kapil acknowledges the University Grant Commission (UGC) of the Government of India for UGC NET-JRF/SRF fellowship.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.

Arup Baruah, Kaushik Das, Ferdous Barbhuiya, and Kuntal Dey. 2020. Aggression identification in english, hindi and bangla text using bert, roberta and svm. In Proceedings of the second workshop on trolling, aggression and cyberbullying, pages 76–82.

Md Abul Bashar and Richi Nayak. 2020. Qutnocturnal@ hasoc'19: Cnn for hate speech and offensive content identification in hindi language. arXiv preprint arXiv:2008.12448.

Elisa Bassignana, Valerio Basile, Viviana Patti, et al. 2018. Hurtlex: A multilingual lexicon of words to hurt. In CEUR Workshop proceedings, volume 2253, pages 1–6. CEUR-WS.

Mohit Bhardwaj, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Hostility detection dataset in hindi. arXiv preprint arXiv:2011.03588.

Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr Ojha. 2020. Developing a multilingual annotated corpus of misogyny and aggression. arXiv preprint arXiv:2003.07428.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of hindi-english code-mixed social media text for hate speech detection. In Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media, pages 36–41.

François Chollet et al. 2015. Keras. https://keras.io.

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

Corinna Cortes and Vladimir Vapnik. 1995. Support vector machine. Machine learning, 20(3):273–297.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. arXiv preprint arXiv:1703.04009.

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

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5):378.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization

of twitter abusive behavior. In Twelfth International AAAI Conference on Web and Social Media.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. arXiv preprint arXiv:1809.04444.

Vikram Gupta, Sumegh Roychowdhury, Mithun Das, Somnath Banerjee, Punyajoy Saha, Binny Mathew, Animesh Mukherjee, et al. 2022. Multilingual abusive comment detection at scale for indic languages. Advances in Neural Information Processing Systems, 35:26176–26191.

Vikas Kumar Jha, Pa Hrudya, PN Vinu, Vishnu Vijayan, and Pa Prabaharan. 2020. Dhot-repository and classification of offensive tweets in the hindi language. Procedia Computer Science, 171:2324–2333.

Prashant Kapil, Gitanjali Kumari, Asif Ekbal, Santanu Pal, Arindam Chatterjee, and BN Vinutha. 2023. Hhld: Hateful posts identification in hindi language leveraging multi task learning. IEEE Access.

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

Yoon Kim. 2014. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. CoRR, abs/1412.6980.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, Viviana Patti, et al. 2020. Hurtbert: incorporating lexical features with bert for the detection of abusive language. In Fourth Workshop on Online Abuse and Harms, pages 34–43. Association for Computational Linguistics.

Ritesh Kumar, Aishwarya N Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated corpus of hindi-english code-mixed data. arXiv preprint arXiv:1803.09402.

Promita Maitra and Ritesh Sarkhel. 2018. A k-competitive autoencoder for aggression detection in social media text. In Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018), pages 80–89.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In Forum for Information Retrieval Evaluation, pages 29–32.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. 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, pages 14–17.

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

Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. 2018. Did you offend me? classification of offensive tweets in hinglish language. In Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), pages 138–148.

Shubhanshu Mishra and Sudhanshu Mishra. 2019. 3idiots at hasoc 2019: Fine-tuning transformer neural networks for hate speech identification in indo-european languages. In FIRE (Working Notes), pages 208–213.

Sandip Modha, Prasenjit Majumder, and Thomas Mandl. 2018. Filtering aggression from the multilingual social media feed. In Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018), pages 199–207.

Sandip Modha, Thomas Mandl, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Tharindu Ranasinghe, and Marcos Zampieri. 2021. 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, pages 1–3.

Kashyap Raiyani, Teresa Gonçalves, Paulo Quaresma, and Vitor Beires Nogueira. 2018. Fully connected neural network with advance preprocessor to identify aggression over facebook and twitter. In Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018), pages 28–41.

Julian Risch and Ralf Krestel. 2018. Aggression identification using deep learning and data augmentation. In Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018), pages 150–158.

Niloofar Safi Samghabadi, Deepthi Mave, Sudipta Kar, and Thamar Solorio. 2018. Ritual-uh at trac 2018 shared task: aggression identification. arXiv preprint arXiv:1807.11712.

Abhishek Velankar, Hrushikesh Patil, Amol Gore, Shubham Salunke, and Raviraj Joshi. 2021. Hate and offensive speech detection in hindi and marathi. arXiv preprint arXiv:2110.12200.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL student research workshop, pages 88–93.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words–a feature-based approach.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. arXiv preprint arXiv:1902.09666.