An Ensemble Approach for Hate and Offensive Language Identification in English and Indo-Aryan Languages

Abhinav Kumar[1], Pradeep Kumar Roy[2] and Sunil Saumya[3]

[1]Department of Computer Science & Engineering, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, India
[2]Department of Computer Science & Engineering, Indian Institute of Information Technology Surat, Gujarat, India
[3]Department of Computer Science & Engineering, Indian Institute of Information Technology Dharwad, India

Abstract

The freedom to upload and the lack of effective social media monitoring have resulted in a slew of societal issues such as cyberbullying, offensive content, and hate speech. Due to this, identifying hate and abusive language on social media is one of the trendiest research topics these days. This work proposes an ensemble-based model for detecting hate and offensive language in English and Hindi social media postings, which combines a support vector machine, logistic regression, random forest, gradient boosting, and Adaboost classifiers. The use of word-level n-gram features performed significantly well in the English dataset, with macro F_1-scores of 0.79 and 0.59 for two different tasks, while character-level n-gram features performed significantly well in the Hindi dataset, with macro F_1-scores of 0.75 and 0.47 for two different tasks.

Keywords

Hate speech, Offensive content, Deep learning, Machine learning, Ensemble learning

1. Introduction

The rise of mobility and the accessibility of the Internet has enticed people all over the world to utilize social media platforms for communication [1, 2]. The majority of Internet users used at least one social media network today, such as Facebook, Twitter, Instagram, YouTube, or others. Because communication on these platforms is inexpensive, people are publishing an endless amount of content [3, 4]. In recent years, the freedom to upload and the lack of effective monitoring has led to a slew of societal issues, including cyberbullying, offensive content, and hate speech [5, 6, 7, 8, 9]. Because of anonymity and mobility provided by the social platforms, the cultivation and spread of hate speech eventually leading to hate crime has become easy in a virtual landscape beyond the reach of traditional law enforcement. Hate speech may be defined as "any communication that disparages a person or a group on the basis of their gender, sexual orientation, nationality, religion, or other characteristics" [10, 11, 12].

Forum for Information Retrieval Evaluation, December 13-17, 2021, India

abhinavanand05@gmail.com (A. Kumar); pradeep.roy@iiitsurat.ac.in (P. K. Roy); sunil.saumya@iiitdwd.ac.in (S. Saumya)

Hate speech is considered harmful by several online forums, including Facebook[1], YouTube[2], and Twitter[3], which have policies in place to delete hate speech content. There is significant motivation to explore automatic hate speech detection because of societal concern and how ubiquitous hate speech is becoming on the Internet [5, 13, 14]. The distribution of nasty content can be prevented by automating its identification. Automatic detection of hate speech over the social platform is needed in the current scenario; however, it has several challenges starting from the definition of hate speech itself. Recent social posts containing code-mixed languages, such as English-Hindi, English-Malayalam, or any other code-mixed languages. If a model was developed with a unimodal dataset like English, it might not detect the hate speech post having code-mixed languages effectively.

Several works [13, 14, 10, 8, 15, 16, 17] have been proposed by researchers to identify hate speech from social media. Kumari and Singh [15] presented a model based on convolutional neural networks for detecting hate, obscenity, and abusive language in English and Hindi tweets. To recognize hatred, offensive, and profanity in English, Hindi, and German tweets, Mishra and Pal [16] developed an attention-based bidirectional long-short-term memory network. Mujadia et al. [17] developed an ensemble-based model comprised of a support vector machine, random forest, and Adaboost classifiers to identify hate content in tweets written in English, Hindi, and German. Roy et al. [10] proposed a convolutional neural network-based model for the identification of hate content from social media. Kumar et al. [13] proposed a fine-tuned BERT model whereas [14] used conventional machine learning models for the hate speech identification. Saumya et al. [8] experimented with several conventional machine learning and deep learning models for the hate speech identification from Dravidian social media posts. They found character N-gram features with conventional machine learning classifiers performing better than the complex deep learning models.

In line with these works, the current paper proposes an ensemble-based machine learning model for the identification of hate and offensive content from English and Hindi social media posts. The dataset published for the FIRE-2021 workshop [18, 19] is used to validate the proposed ensemble-based model.

The rest of the sections are organized as follows: Section 2 discusses the proposed methodology in detail. Section 3 lists the findings and finally the paper is concluded in section 4.

2. Methodology

The detailed diagram of the proposed ensemble-based model can be seen in Figure 1. The proposed ensemble-based model consists of five different classifiers: (i) Support Vector Machine (SVM), (ii) Logistic Regression (LR), (iii) Random Forest (RF), (iv) Gradient Boosting (GB), and (v) AdaBoost. The proposed model is validated with the dataset published in FIRE-2021 [18]. Two different sub-tasks were given: (i) a coarse-grained binary classification of tweets in Hate and Offensive (HOF) and Non-Hate and offensive (NOT) classes, (ii) the further classification of Hate and Offensive (HOF) tweets into Hate (HATE), profane (PRFN) and offensive (OFFN) posts.

[1] https://www.facebook.com/communitystandards/objectionable$_{c}ontent$.

[2] https://support.google.com/youtube/answer/2801939.

[3] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

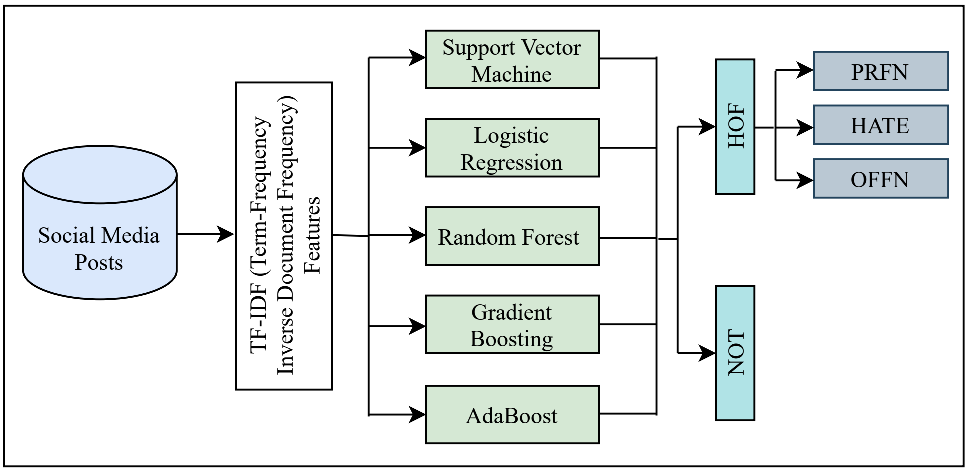

Figure 1: Proposed model for the hate and offensive language identification from social media

Table 1
Data statistic used to validate the proposed system

	Class	English		Hindi	
		Train	Test	Train	Test
Task-A	HOF	2,501	798	1,433	505
	NOT	1,342	483	3,161	1,027
	Total	3,843	1,281	4,594	1,532
Task-B	NONE	1,342	483	3,161	1,027
	PRFN	1,196	379	213	74
	HATE	683	224	566	215
	OFFN	622	195	654	216
	Total	3,843	1,281	4,594	1,532

The overall data statistic for both the task can be seen in Table 1.

In the experimentation, the aforementioned classifiers performed well individually in the identification of hate and offensive content, due to this, we utilized them to construct an ensemble-based model that can train efficiently to identify hate and offensive content on social media. To provide input to the proposed model, we experimented with different combinations of word and character n-gram TF-IDF (Term-Frequency Inverse Document Frequency) features for both English and Hindi datasets.

- **English Task-A and Task-B:** TF-IDF is retrieved from the textual contents of the social media post to provide input to the suggested ensemble-based model. In the case of English language posts, we found that the first 50,000 uni-gram, bi-gram, and tri-gram word-level TF-IDF features performed well with the model in classifying posts into the various hate classes, compared to other n-gram combinations of word-level and character-level TF-IDF features.

Table 2
Results for hate and offensive language identification from English and Hindi social media posts

Task	Class	Precision	Recall	F_1-score	Accuracy
English Task-A	HOF	0.79	0.89	0.84	78.22
	NOT	0.77	0.60	0.67	
	Macro Average	0.78	0.75	0.76	
English Task-B	HATE	0.56	0.42	0.48	65.96
	NONE	0.68	0.72	0.70	
	OFFN	0.59	0.32	0.42	
	PRFN	0.69	0.90	0.78	
	Macro Average	0.63	0.59	0.59	
Hindi Task-A	HOF	0.79	0.55	0.65	80.22
	NOT	0.81	0.93	0.86	
	Macro Average	0.80	0.74	0.75	
Hindi Task-B	HATE	0.44	0.08	0.14	73.56
	NONE	0.77	0.97	0.86	
	OFFN	0.57	0.44	0.49	
	PRFN	0.70	0.28	0.40	
	Macro Average	0.62	0.44	0.47	

- **Hindi Task-A and Task-B:** In the case of Hindi social media posts, we found that the first 70,000 one-to-six gram character-level TF-IDF features performed the best when compared to other word-level and char-level n-gram features.

3. Results

The performance of the proposed model is measured in terms of macro precision, macro recall, macro F_1-score, and accuracy. The results for both the sub-tasks for the English and Hindi dataset are listed in Table 2. In the case of English Task-A, the proposed ensemble-based model achieved a macro precision of 0.78, macro recall of 0.75, macro F_1-score of 0.76, and accuracy of 78.22%. The confusion matrix for English Task-A can be seen in Figure 2.

In the case of English Task-B, the proposed ensemble-based model achieved a macro precision of 0.63, macro recall of 0.59, macro F_1-score of 0.59, and accuracy of 65.96%. The confusion matrix for English Task-B can be seen Figure 3. For Hindi Task-A, the proposed model achieved a macro precision of 0.80, a macro recall of 0.74, macro F_1-score of 0.75, and accuracy of 80.22%. The confusion matrix for the Hindi Task-B can be seen in Figure 4. Similarly, for Hindi Task-B, the proposed model achieved a macro precision of 0.62, macro recall of 0.44, macro F_1-score of 0.47, and accuracy of 73.56%. The confusion matrix for the Hindi Task-B can be seen in Figure 5.

4. Conclusion

The detection of hate speech on social media poses significant problems. This paper investigates the usefulness of TF-IDF features at the word and character levels using an ensemble-based

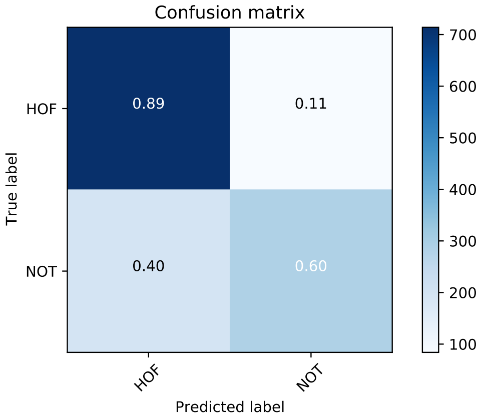

Figure 2: Confusion matrix for English Task-A

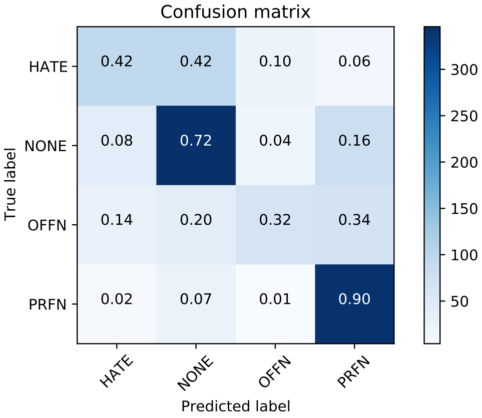

Figure 3: Confusion matrix for English Task-B

machine learning approach. The proposed ensemble-based model achieved macro $F1$-scores of 0.79 and 0.59 for English task-A and task-B, respectively, and 0.75 and 0.47 for Hindi task-A and task-B, respectively. In the future, some other deep learning-based ensemble models can be implemented for the identification of hate and offensive content from social media posts.

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.

[2] A. Kumar, J. P. Singh, Disaster severity prediction from twitter images, in: Intelligence Enabled Research, Springer, 2021, pp. 65–73.

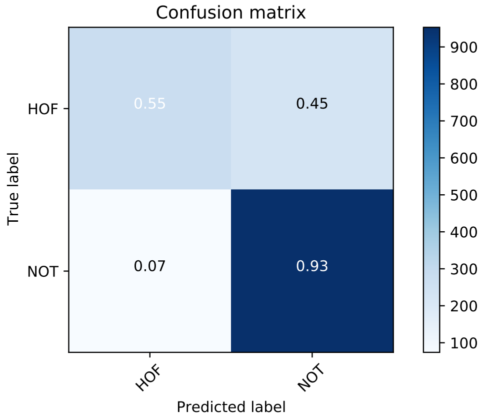

Figure 4: Confusion matrix for Hindi Task-A

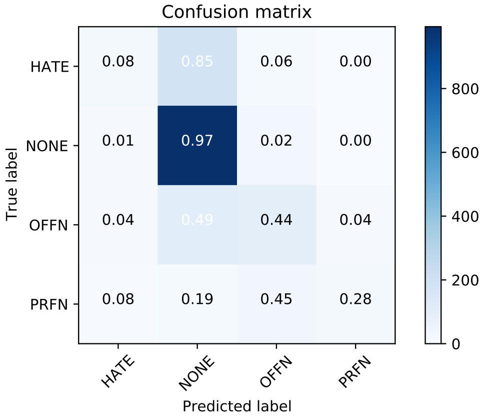

Figure 5: Confusion matrix for Hindi Task-B

[3] A. Priya, A. Kumar, Deep ensemble approach for COVID-19 fake news detection from social media, in: 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN), IEEE, 2021, pp. 396–401.

[4] A. Kumar, J. P. Singh, S. Saumya, A comparative analysis of machine learning techniques for disaster-related tweet classification, in: 2019 IEEE R10 Humanitarian Technology Conference (R10-HTC)(47129), IEEE, 2019, pp. 222–227.

[5] M. Mondal, L. A. Silva, F. Benevenuto, A measurement study of hate speech in social media, in: Proceedings of the 28th ACM conference on hypertext and social media, 2017, pp. 85–94.

[6] G. Kumar, J. P. Singh, A. Kumar, A deep multi-modal neural network for the identification of hate speech from social media, in: Conference on e-Business, e-Services and e-Society, Springer, 2021, pp. 670–680.

[7] A. K. Mishra, S. Saumya, A. Kumar, IIIT_DWD@ HASOC 2020: Identifying offensive

content in Indo-European languages, in: FIRE (Working Notes), 2020.

[8] S. Saumya, A. Kumar, J. P. Singh, Offensive language identification in Dravidian code mixed social media text, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, 2021, pp. 36–45.

[9] 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, 2017, pp. 759–760.

[10] P. K. Roy, A. K. Tripathy, T. K. Das, X.-Z. Gao, A framework for hate speech detection using deep convolutional neural network, IEEE Access 8 (2020) 204951–204962.

[11] 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.

[12] R. Jain, D. Goel, P. Sahu, A. Kumar, J. Singh, Profiling Hate Speech Spreaders on Twitter—Notebook for PAN at CLEF 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-2936/paper-175.pdf.

[13] A. Kumar, S. Saumya, J. P. Singh, NITP-AI-NLP@ HASOC-FIRE2020: Fine tuned BERT for the hate speech and offensive content identification from social media., in: FIRE (Working Notes), 2020, pp. 266–273.

[14] A. Kumar, S. Saumya, J. P. Singh, NITP-AI-NLP@ HASOC-Dravidian-CodeMix-FIRE2020: A machine learning approach to identify offensive languages from Dravidian code-mixed text., in: FIRE (Working Notes), 2020, pp. 384–390.

[15] K. Kumari, J. P. Singh, AI ML NIT Patna at HASOC 2019: Deep learning approach for identification of abusive content., in: FIRE (Working Notes), 2019, pp. 328–335.

[16] A. Mishra, S. Pal, IIT Varanasi at HASOC 2019: Hate speech and offensive content identification in Indo-European languages., in: FIRE (Working Notes), 2019, pp. 344–351.

[17] V. Mujadia, P. Mishra, D. M. Sharma, IIIT-Hyderabad at HASOC 2019: Hate speech detection., in: FIRE (Working Notes), 2019, pp. 271–278.

[18] 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: FIRE 2021: Forum for Information Retrieval Evaluation, Virtual Event, 13th-17th December 2021, ACM, 2021.

[19] T. Mandl, S. Modha, G. K. Shahi, H. Madhu, S. Satapara, P. Majumder, J. Schäfer, T. Ranasinghe, M. Zampieri, D. Nandini, A. K. Jaiswal, Overview of the HASOC subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021. URL: http://ceur-ws.org/.