Machine Learning Approach for Hate Speech and Offensive Content Identification in English and Indo Aryan Code-Mixed Languages

Kirti Kumari[1,*], Jyoti Prakash Singh[2]

[1]Indian Institute of Information Technology Ranchi, Ranchi, Jharkhand, India.
[2]National Institute of Technology Patna, Patna, Bihar, India

Abstract

In current times, social media is the most widely used platform, and everyone has the right to express their speculations, ideas and thoughts. In such a case, it is often seen that hate speech and offensive contents are spreading like wildfire, making a detrimental impact on the world. It is important to identify and eradicate such offensive content from social media. This paper is a contribution to the Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC) 2022 shared task by the $AI\_ML\_IIITRanchi$ team. We experimented with machine learning models to detect hate speech and offensive content in all three code-mixed languages English, German and Marathi as provided. Our experimental results show that a Logistic Regression, Support Vector Machine and Random Forest classifier can achieve good results for multilingual hate speech and offensive content identification. Overall, our team participated on all the tasks and ranked 3^{rd},$5^{th} and 7^{th}$ on Marathi C, Marathi B and Marathi A tasks respectively. Our team ranked 8^{th} and 9^{th} on ICHCL-Multiclass and ICHCL-Binary class shared tasks, respectively.

Keywords

HASOC, Machine Learning, Logistic Regression, Support Vector Machine, Random Forest

1. Introduction

People are voicing themselves through social media sites such as Twitter and Facebook, which are user-friendly and easily available. People of various ages use all these sites to continue sharing every detail of their daily life, filling them with personal data and which gives us a huge pool of data. Every technology has advantages and disadvantages, and social media platforms are no exception. The prevalence of hate speech and other offensive and objectionable information on the web has posed an enormous threat to society. Derogatory, hurtful, insulting, or obscene language directed from one person to another person and also openly available to others impairs the objectivity of conversations. As this kind of communication becomes more prevalent online, disputes become more extreme. The democratic process may be threatened by objectionable content. Open societies must also come up with an appropriate remedy to such content that avoids enforcing strict censorship laws.

Forum for Information Retrieval Evaluation, December 9-13, 2022, India
*Corresponding author.
kirti@iiitranchi.ac.in (K. Kumari); jps@nitp.ac.in (J. P. Singh)
0000-0003-3714-7607 (K. Kumari)

Study of hate speech and abusive language Identification is gradually gaining momentum, mostly as a result of the aggregation of numerous shared tasks [1, 2, 3, 4]. The Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC) 2022 is also continuation and addition to previous shared task. This has prompted many social media networks to scrutinize what people are posting. As a result, techniques to detect questionable posts automatically become essential.

The current HASOC 2019, HASOC 2020, HASOC 2021 and HASOC 2022 shared tasks given the opportunity for the researchers to cope with code-mixing and script mixing different multilingual Indian languages.

In this work, we tried machine learning approach for the HASOC 2022 all the shared tasks. We tried for all the given tasks which are in code-mixed of English, Hindi and Marathi languages and achieved the good results by the team $AI\_ML\_IIITRanchi$.

The rest of the paper organized as follows: Section 2 discuss a brief about related works done in the area of Hate and Offensive language identification. Section 3 discussed the dataset as well as task description. Section 4 provides the detail about proposed approach. Section 5 presents the results and finding of our work. Finally, we concluded in Section 6.

2. Related Work

Automatic Hate Speech and Offensive language identification is an active area from the Natural Language Processing (NLP) research community [1, 2, 3]. The wide range of interrelated previous works have been done in this field [5, 6] but early works on these fields are mainly for mono-lingual English language. Recently some of the shared tasks are focused on code-mixed and multilingual regional languages [1, 2, 3, 7, 8, 9]. In the above mentioned shared tasks were tried to address the multilingual problems on automated identification of Hate Speech, Aggression and Offensive languages. The HASOC 2019 [1] and HASOC 2020 [2] shared tasks are focused on three languages: English, Hindi and German with similar tasks as current task. Next, HASOC 2021 [3, 10] added a one more Marathi language; which is similar to Hindi, spoken by millions of Indian people and also added the one more task: Conversational Hate Speech detection [11]. The TRAC 2018 [8], TRACK 2020 and TRAC 2022[8, 9] Some of the potential works in Hate Speech and Aggression identification areas are [12, 13, 14, 15]. In this work, we also tried to address same issue using machine learning approach, which discussed in subsequent sections. The wide range of interrelated previous works have been done in this field [5, 6] but early works on these fields are mainly for mono-lingual English language. Recently some of the shared tasks are focused on code-mixed and multilingual regional languages [1, 2, 3, 7, 8, 9]. In the above mentioned shared tasks were tried to address the multilingual problems on automated identification of Hate Speech, Aggression and Offensive languages. The HASOC 2019 [1] and HASOC 2020 [2] shared tasks are focused on three languages: English, Hindi and German with similar tasks as current task. Next, HASOC 2021 [3, 10] added a one more Marathi language; which is similar to Hindi, spoken by millions of Indian people and also added the one more task: Conversational Hate Speech detection [11]. The TRAC 2018 [8], TRACK 2020 [8], and TRAC 2022 [9] are focused on different types of aggression detection in multilingual scenarios.

Some of the potential works in Hate Speech and Aggression identification areas are [12, 13, 14]

Table 1
Distribution of Datasets for different task

Task	Class	#sample
ICHCL-Binary	HOF	2612
	NOT	2609
ICHCL-Multiclass	NONE	2390
	SHOF	1636
	CHOF	888
3A-Marathi	NOT	2034
	OFF	1069
3B-Marathi	NONE	2035
	UNT	327
	TIN	741
3C-Marathi	NONE	2363
	IND	502
	GPR	157
	OTH	80

are applied deep learning approaches with different embedding techniques and achieved good results. A recent work on aggression identification [15] utilized the machine leaning approach and ranked first position on TRAC 2022 shared task. So, motivating with the work [15], we tried machine learning approach to tackle HASOC 2022 shared tasks in this work, which discussed in subsequent sections.

3. Dataset

The HASOC 2022 shared task[1] has two main tasks which are:

- Identification of Conversational Hate-Speech in Code-Mixed Languages (ICHCL)
- Offensive Language Identification in Marathi

Identification of Conversational Hate-Speech in Code-Mixed Languages: this task has two subtasks: Subtask 1 and Subtask 2. Subtask 1 is about Binary classification and these include two categories Hate Speech and Offensive (HOF) and Not Hate Speech and Offensive (NOT). The comment includes the Hindi and English (Hinglish) as well as the German Languages words. Subtask 2 is about multiclass problems and contains Non-Hate (NONE), Contextual Hate (CHOF) and Standalone Hate(SHOF). Includes only the Hinglish Language in the Dataset.

Offensive Language Identification in Marathi: this has three subtasks Task 3A, Task 3B and Task 3C. Task 3A contains NOT and OFF named classes. Task 3B contains TIN (targeted insult) and UNT (untargeted insult) classes. Task 3C contains IND (individual), GRP (group) and OTH (others) classes.

The detail distribution of samples for each tasks can be seen in Table1.

More details about the all shared tasks and datasets used can be seen in [16, 4, 17, 18, 10, 11, 19, 20].

[1]https://hasocfire.github.io/hasoc/2022/dataset.html

4. Methodology

This section describes the methods used in this work on the given HASOC 2022 shared tasks [2]. The subsequent content describes the approach used for the further classification of hate speech into different categories as explained in the previous section. We begin by explaining the each steps of the dataset for each of the three languages followed by the machine learning models used.

4.1. Preprocessing

The preprocessing of text data for three languages has been done in the following ways. For the Hinglish language, we first converted the texts to lowercase, and texts such as URLs and punctuation symbols. Stemming was done on the dataset using 'SnowballStemmer'. Every tweet had comments and replies and every comment and tweet is to be predicted, so the comments and the replies were padded with the original tweet so the correct meaning of the tweets and comments is revealed. All the sentences in Marathi were lemmatized. Lemmatization is a part of stemming, stemming truncates the words harshly, but lemmatization keeps the word meaningful. All the emojis were removed from the sentences using regular expressions.

Some other preprocessing were done as: Stopwords removal, Stemming and Tweets processing are discussed in the following subsections.

4.1.1. Removal of Stopwords

The stopwords of the English language and Hindi languages are removed from the dataset. Our observation on this dataset that during offensive language detection, stopwords do not play any important role. So, we removed here.

4.1.2. Stemming

We used the stemmer to stem from the root word, which increased the efficiency of the model greatly.

4.1.3. Tweets Processing

Every tweet had comments and replies, so the comments and the replies were padded with the original tweet so the correct meaning of the sentences is revealed. We used TF-IDF-Vectorizer to minimize the running time of the code. We used a train test split to split the training data as 80:20 ratio for our validation phase.

4.2. Models Used

We tried different types of machine learning classifiers such as Support Vector Machine, Logistic Regression, Multinomial Naïve Bayes, Decision Tree and Random Forest. We found that Random

[2]https://hasocfire.github.io/hasoc/2022/ichcl.html

Table 2
Validation results for Task 1 (ICHCL-Binary Task) on Random Forest Classifier

Class	Precision	Recall	F1-Score
HOF	0.68	0.78	0.73
NOT	0.72	0.64	0.68

Table 3
Validation results for Task 2 (ICHCL-Multiclass Task) on Support Vector Machine

Class	Precision	Recall	F1-Score
NONE	0.51	0.47	0.49
SHOF	0.73	0.52	0.61
SHOF	0.65	0.76	0.70

Forest, Logistic Regression and Support Vector Machine are better classifiers in our case of experiments.

4.3. Model selection

The tasks which had binary classification problem, logistic regression gave the better result, as the Sigmoid function used in the logistic regression function which predicts zero or one. The dataset was also linearly separable into two classes, which was also a reason why Logistic Regression performed so well in our experimentations. For other two datasets, Random Forest worked better as data was high dimensional data and Random Forest works with subsets of data. It is faster to train than Decision Trees because we are working only on a subset of features in this model, so we can easily work with hundreds of features. Those tasks which had more than two classes Support Vector Machine performed very good in those cases as there was a clear separation between the classes, and the dataset was sufficient large to train the model.

5. Results and Analysis

In this section we presented our experimental results as well as analysis of our models.

Before the organisers made the test set accessible, we assessed the performance of our suggested models using validation data (20% of random data taken from training set of each shared task). We used the aforementioned validation data to develop the model when unlabeled test data was released, and final predictions are made utilising such models.

Our experimental results shown in the Table 2 and Table 3 for Task 1 (ICHCL-Binary Task) and Task 2 (ICHCL-Multiclass Task). We found that Random Forest is performing better for Task 1 and Support Vector Machine for Task 2 from the models that we experimented with.

Our further experimental results shown in the Table 4, Table 5 and Table 6 for Task 3A, Task 3B and Task 3C, respectively. We found that Logistic Regression is performing best for Task 3A and Task 3B and Support Vector Machine for Task 3C.

Table 4
Validation results for Task 3A Marathi on Logistic Regression

Class	Precision	Recall	F1-Score
HOF	0.79	0.82	0.84
NOT	0.76	0.80	0.78

Table 5
Validation results for Task 3B Marathi on Logistic Regression

Class	Precision	Recall	F1-Score
NaN	0.85	0.91	0.79
TIN	0.69	0.84	0.77
UNT	0.49	0.51	0.50

Table 6
Validation results for Task 3C Marathi on Support Vector Machine

Class	Precision	Recall	F1-Score
NaN	0.89	0.97	0.81
IND	0.71	0.93	0.81
GRP	0.82	0.23	0.36
OTH	0.42	0.38	0.40

Table 7
F1-Score of top three classifiers on validation data for all given tasks

Task	LR	SVM	RF
ICHCL-Binary Task	0.70	0.71	**0.74**
ICHCL-Multiclass Task	0.57	**0.60**	0.58
3A-Marathi	**0.82**	0.79	0.76
3B-Marathi	**0.60**	0.56	0.55
3C-Marathi	0.58	**0.60**	0.55

For each tasks, we present the results for the evolution of experimented models and the final model of each shared tasks. The observations are analyzed and compared in greater detail and after that the best model was submitted based on a comparison of our model's performance. A summary of the results for each of the tasks are evaluated using the average macro F1-Score. The best three models results can be seen in Table 7 and Table 8 on validation data and testing data, respectively. In Table 8, blank shows that we have missed the submission on test data for that specific classifier due to lack of time. We can observed from the Table 7 and Table 8, Random Forest classifier is performing better for ICHCL-Binary task, Support Vector Machine for ICHCL-Multiclass and 3C- Marathi tasks and Logistic Regression for 3A Marathi and 3B Marathi tasks that we experimented with.

Table 8
F1-Score of top three classifiers on test data for all given tasks with comparison of best reported team

Task	LR	SVM	RF	Result of Rank #1 Team [4]
ICHCL-Binary Task	0.55	-	**0.60**	**0.71**
ICHCL-Multiclass	0.410	**0.416**	-	**0.49**
3A-Marathi	**0.92**	0.82	-	**0.97**
3B-Marathi	-	**0.44**	0.33	**0.92**
3C-Marathi	0.49	**0.74**	-	**0.96**

The reason for Random Forest classifier out-performing the other algorithms on binary class problem is because it offers us relative feature importance which allows us to select the most contributing features. The difficulty faced during experimentation's that we were not able to properly pre-process the German data of ICHCL tasks due to lack of resources and lack of time. The difficulty faced for Marathi tasks that we have not able to pre-process some parts such as stopwords removal could be done for the Marathi language which led to low F1 Scores.

6. Conclusion

In this work we, $AI\_ML\_IIITRanchi$ team participated on all the shared tasks as very few teams participated on all the shared tasks. Here, we have presented our machine learning approach to address all the five different shared tasks of HASOC 2022. We found that Logistic Regression, Support Vector Machine and Random Forest classifiers are performing better in our case of experiments. Overall, our top models ranked 3^{rd},5^{th} and 7^{th} on Marathi C, Marathi B and Marathi A tasks, respectively. Our team ranked 8^{th} and 9^{th} on ICHCL-Multiclass and ICHCL-Binary class shared tasks, respectively.

Acknowledgments

We are thank full to our undergraduate students Ayush Kumar Singh and Mrinmoy Mahato for their help in prepossessing steps. A very special thanks to academics and Management of Indian Institute of Information Technology Ranchi for providing the necessary resources and encouragement.

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] S. Modha, P. Majumder, T. Mandl, C. Mandalia, Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance, Expert Systems with Applications 161 (2020) 113725.

[3] 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: Forum for Information Retrieval Evaluation, 2021, pp. 1–3.

[4] T. Ranasinghe, K. North, D. Premasiri, M. Zampieri, Overview of the HASOC subtrack at FIRE 2022: Offensive Language Identification in Marathi, in: Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, CEUR, 2022.

[5] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 1–10. URL: https://aclanthology.org/W17-1101. doi:10.18653/v1/W17-1101.

[6] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (2018) 1–30.

[7] R. Kumar, A. N. Reganti, A. Bhatia, T. Maheshwari, Aggression-annotated Corpus of Hindi-English Code-mixed Data, in: N. C. C. chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018.

[8] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri, Evaluating aggression identification in social media, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 1–5. URL: https://aclanthology.org/2020.trac-1.1.

[9] R. Kumar, S. Ratan, S. Singh, E. Nandi, L. N. Devi, A. Bhagat, Y. Dawer, b. lahiri, A. Bansal, A. K. Ojha, The comma dataset v0.2: Annotating aggression and bias in multilingual social media discourse, in: Proceedings of the Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 4149–4161. URL: https://aclanthology.org/2022.lrec-1.441.

[10] Modha, Sandip and Mandl, Thomas and Shahi, Gautam Kishore and Madhu, Hiren and Satapara, Shrey and Ranasinghe, Tharindu and Zampieri, Marcos, 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, CEUR, 2021, pp. 1–3.

[11] M. S. Satapara, Shrey, T. Mandl, H. Madhu, P. Majumder, Overview of the HASOC Subtrack at FIRE 2021: Conversational Hate Speech Detection in Code-mixed language , in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021, pp. 20–31.

[12] K. Kumari, J. P. Singh, Ai ml nit patna at hasoc 2019: Deep learning approach for identification of abusive content., FIRE (working notes) 2517 (2019) 328–335.

[13] K. Kumari, J. P. Singh, Ai_ml_nit_patna@ hasoc 2020: Bert models for hate speech identification in indo-european languages., in: FIRE (Working Notes), 2020, pp. 319–324.

[14] K. Kumari, J. P. Singh, AI_ML_NIT_Patna @ TRAC - 2: Deep learning approach for multilingual aggression identification, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Mar-

seille, France, 2020, pp. 113–119. URL: https://aclanthology.org/2020.trac-1.18.

[15] K. Kumari, S. Srivastav, R. R. Suman, Bias, threat and aggression identification using machine learning techniques on multilingual comments, in: Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022), Association for Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 30–36. URL: https://aclanthology.org/2022.trac-1.4.

[16] Satapara, Shrey and Majumder, Prasenjit and Mandl, Thomas and Modha, Sandip and Madhu, Hiren and Ranasinghe, Tharindu and Zampieri, Marcos and North, Kai and Premasiri, Damith, Overview of the HASOC Subtrack at FIRE 2022: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages, in: FIRE 2022: Forum for Information Retrieval Evaluation, Virtual Event, 9th-13th December 2022, ACM, 2022.

[17] S. Modha, T. Mandl, P. Majumder, S. Satapara, T. Patel, H. Madhu, Overview of the HASOC Subtrack at FIRE 2022: Identification of Conversational Hate-Speech in Hindi-English Code-Mixed and German Language, in: Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, CEUR, 2022.

[18] M. Zampieri, T. Ranasinghe, M. Chaudhari, S. Gaikwad, P. Krishna, M. Nene, S. Paygude, Predicting the type and target of offensive social media posts in marathi, Social Network Analysis and Mining 12 (2022) 77. URL: https://doi.org/10.1007/s13278-022-00906-8. doi:10.1007/s13278-022-00906-8.

[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, pp. 1–19.

[20] S. S. Gaikwad, T. Ranasinghe, M. Zampieri, C. Homan, Cross-lingual offensive language identification for low resource languages: The case of Marathi, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), INCOMA Ltd., Held Online, 2021, pp. 437–443. URL: https://aclanthology.org/2021.ranlp-1.50.