

# JU at HASOC 2020: Deep Learning with RoBERTa and Random Forest for Hate Speech and Offensive Content Identification in Indo-European Languages

Biswarup Ray, Avishek Garain

*Department of Computer Science and Engineering,  
Jadavpur University,  
Kolkata-700032, West Bengal, India*

## Abstract

The identification of Hate Speech in Social Media has received much attention in research recently. There has been an ever-growing increase in demand particularly for research in languages other than English. The Hate Speech and Offensive Content (HASOC) track has created resources for Hate Speech Identification in three different languages namely Hindi, German, and English. We have participated in both Sub-tasks A and B of the 2020 shared task on hate speech and offensive content identification in Indo-European languages. Our approach relies on a combined model of multilingual RoBERTa (a Robustly Optimized BERT Pretraining Approach) model with pre-trained vectors and a Random Forest model using Word2Vec, TF-IDF, and other textual features as input. Our system has achieved a maximum Macro F1-score of 50.28% for English Sub-task A which is quite satisfactory relative to the performance of other systems and secured 8th position among participating teams.

## Keywords

Hate Speech, RoBERTa, Random Forest, TF-IDF, Word2Vec

## 1. Introduction

Accuracy and efficiency of any supervised classification method are heavily dependent on the corpora on which it is trained to make the dataset a very important entity for such classification tasks. Hate Speech as such is a topic that has attracted the attention of researchers time and again resulting in several previous initiatives of corpora creation [1]. There has been significant work in several languages, in particular for English. However, for languages other than English, such as Hindi standard datasets are not available as such. There is a huge demand for resources for many languages other than English. HASOC is such a shared task that developed a resource for three languages altogether and which encourages attaining results in terms of multilingual research. In this paper, we have proposed a model for the new HASOC task for hateful and offensive speech classification on texts from three different languages (English, Hindi, and German). A combined classifier model has been proposed, which uses the pre-trained multilingual model RoBERTa as

---


*FIRE '20, Forum for Information Retrieval Evaluation, December 16–20, 2020, Hyderabad, India*

✉ raybiswarup9@gmail.com (B. Ray); avishekgarain@gmail.com (A. Garain)

ORCID 0000-0001-6225-3343 (A. Garain)



© 2020 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

the base for the contextual word representations and a 2 class sentiment analysis task. Among the 2 classes, the Hate and Offensive class are further classified into 3 classes by a Random Forest classifier using textual features as input features.

The rest of the paper has been organized as follows. Section 2 describes the data on which the task was performed. The methodology followed is described in Section 3. This is followed by the results and concluding remarks in Section 4 and 5 respectively.

## 2. Data

The corpus that has been used for this task contains texts in 3 different languages namely English, German, and Hindi, and consists of 3794, 2452, and 2963 tweets respectively. We have divided the dataset in the ratio 80:20 for training and validation purposes respectively. The distribution of data instances is given in Tables 1 and 2.

### 2.1. Sub-task A

The three possible categories established in the dataset under this Sub-task are:

- (NOT) Non-Hate-Offensive - This post does not contain any Hate speech, profane, offensive content.
- (HOF) Hate and Offensive - This post contains Hate, offensive, and profane content.

Language	Label	Train	Validation
English	NOT	1477	385
	HOF	1588	374
	All	3035	759
German	NOT	1474	377
	HOF	487	114
	All	1961	491
Hindi	NOT	1698	418
	HOF	672	175
	All	2370	593

Table 1: Distribution of the labels in the dataset in Sub-task-A

### 2.2. Sub-task B

The three possible categories established in the dataset under this Sub-task are:

- (HATE) Hate speech:- Posts under this class contain Hate speech content.
- (OFFN) Offensive:- Posts under this class contain offensive content.

- (PRFN) Profane:- These posts contain profane words.

Language	Label	Train	Validation
English	HATE	124	30
	OFFN	251	60
	PRFN	1106	269
	All	1481	359
German	HATE	85	17
	OFFN	101	25
	PRFN	292	72
	All	478	114
Hindi	HATE	192	42
	OFFN	369	96
	PRFN	111	37
	All	672	175

Table 2: Distribution of the labels in the dataset in Sub-task-B

### 3. Methodology

For HASOC Task, our method uses a RoBERTa model for classification of reviews in each language into Hate and Offensive (HOF), Non- Hate, and offensive (NOT) sentiment labels. Then a method based on a Random Forest classifier with Word2Vec embeddings and TF-IDF (Term Frequency-Inverse Document Frequency) of commonly recurring words as input features were used to further classify the HOF sentiment into (HATE) Hate speech, (OFFN) Offensive and (PRFN) Profane. The workflow of our methodology is shown in Figure 1.

#### 3.1. Preprocessing

It consisted of the following steps:

1. Replacing emojis and emoticons by their corresponding meanings [2]
2. Removing mentions
3. Removing URLs
4. Contracting whitespace
5. Extracting words from hashtags [3]
6. Normalizing numeronyms [4]

#### 3.2. RoBERTa Model

RoBERTa is equipped with the BERT’s language masking method, i.e. the system intentionally learns to predict sections of text which are hidden. Implementation of RoBERTa was done in PyTorch. This allows modification of the key hyperparameters in BERT, which include

training with much larger mini-batches and learning rates and eliminating BERT's objective of next-sentence pre-training. This leads the path in the improvement of the masked language modeling objective for RoBERTa compared with BERT and also leads to better downstream task performance.

For the sentiment analysis task from the Huggingface team transformers library, the pre-trained RoBERTa model has been accessed. XLM RoBERTa base model has been used for the task. Since it has been pre-trained on 100 different languages, the same model could be used for all three languages (Hindi, English, German) datasets. The RoBERTa base uses the BERT-base architecture hence it has 12-layer, 768-hidden, 12-heads, 125M parameters. The pre-trained RoBERTaTokenizer for the RoBERTa large model has been used to get the token representations. The learning rate of  $1e - 5$  has been selected and the model is trained for 10 epochs. The batch size has been set to be 32. For the training procedure, a Dropout Layer for some regularization and a fully-connected layer for the output is used in the model. The Dropout Layer reduces overfitting in the model by preventing complex co-adaptations on training data.

### 3.3. Random Forest Model

Different textual features are extracted from each of the text presents in the HOF and those features are fed into a Random Forest classifier. The textual extracted features added to the model for classification are:

1. The vector representations were obtained using Word2Vec.
2. The TF-IDF (Term Frequency-Inverse Document Frequency) [5] for the words that frequently occur in the text are also added to the feature list. The TF computes the number of times a word recurs in the dataset, and IDF computes the relative importance of the word which depends on how many times the word can be found, and is added as features to filter and reduce the size of the final output [6].
3. Normalized counts of words with positive sentiment, negative sentiment, and neutral sentiment in the corresponding language by dividing with a word count of the corresponding sentence [7].
4. Normalized Frequency of auxiliary verbs by dividing by word count of the sentence.
5. Subjectivity score of the text (calculated using predefined libraries) [8].

The features extracted through the various above mentioned processes are selected by using the feature importance rankings for each feature. The features having a higher feature importance ranking were added to train the Random Forest model. The predictions given by the model for the test dataset were checked if they match with the HOF predictions given by the RoBERTa model. All the predictions which align with the HOF predictions are given by the RoBERTa model were kept as final outputs.

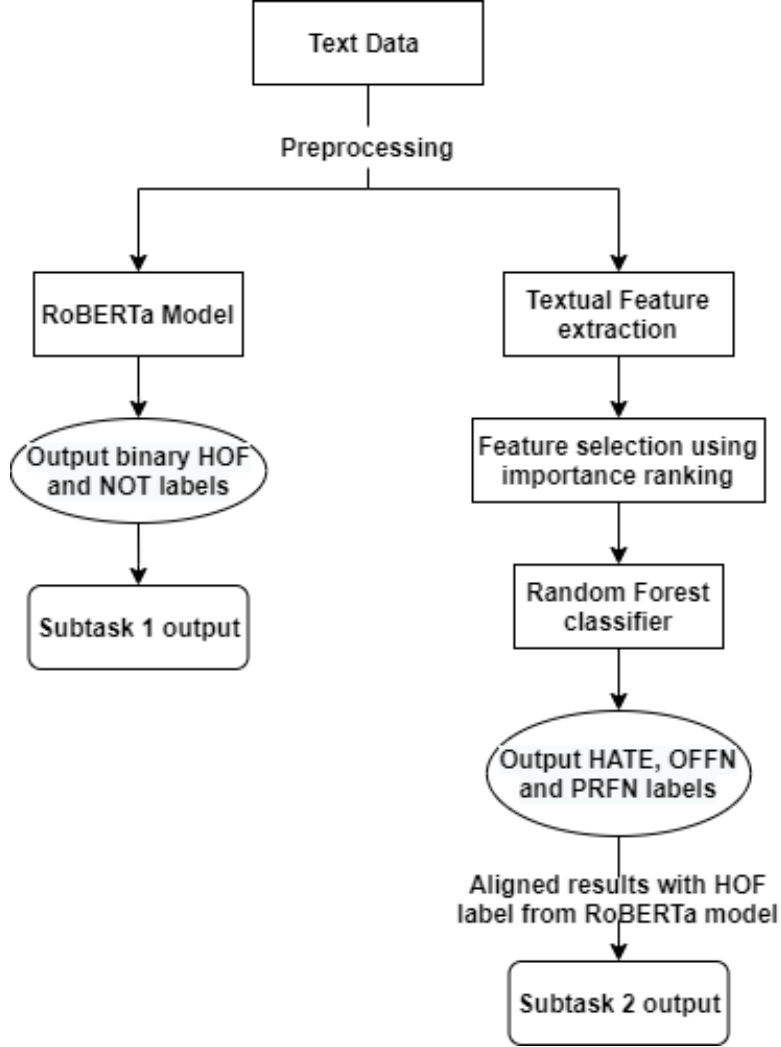


Figure 1: Overview of the methodology

## 4. Results

Our model secured 8<sup>th</sup> and 23<sup>rd</sup> positions in English Sub-task A and B respectively. For German Sub-tasks A and B, our model secured 20<sup>th</sup> and 14<sup>th</sup> positions respectively and for Hindi Sub-task A and B, our model secured 20<sup>th</sup> and 16<sup>th</sup> positions respectively. The performance of our model in terms of Macro F1-score is shown in Table 3.

Sub-task	English	German	Hindi
A	0.5028	0.3231	0.4599
B	0.1623	0.0984	0.1600

Table 3: Performance in terms of Macro F1-score for various tasks

## 5. Conclusion

We have presented the system that we have used for participating in the 2020 shared task on hate speech and offensive content (HASOC) identification in Indo-European languages. Considering previous approaches, our approach is a comparatively different approach in terms of architecture as well as the methodology of feature extraction. It is a generalized and versatile framework and has shown satisfactory performance among all participating systems during the HASOC evaluations. In future works, we will further fine-tune the classification models to increase the performance and we will further experiment with the model in other languages.

## References

- [1] A. Garain, A. Basu, The titans at semeval-2019 task 5: Detection of hate speech against immigrants and women in twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 494–497.
- [2] A. Garain, Humor analysis based on human annotation (haha)-2019: Humor analysis at tweet level using deep learning (2019).
- [3] A. Garain, A. Basu, The titans at semeval-2019 task 6: Offensive language identification, categorization and target identification, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 759–762.
- [4] A. Garain, S. K. Mahata, S. Dutta, Normalization of numeronyms using nlp techniques, in: 2020 IEEE Calcutta Conference (CALCON), IEEE, 2020, pp. 7–9.
- [5] J. Ramos, et al., Using tf-idf to determine word relevance in document queries, in: Proceedings of the first instructional conference on machine learning, volume 242, New Jersey, USA, 2003, pp. 133–142.
- [6] B. Ray, A. Garain, Factuality classification using bert embeddings and support vector machines (2020).
- [7] A. Garain, S. K. Mahata, Sentiment analysis at sepln (tass)-2019: Sentiment analysis at tweet level using deep learning (2019).
- [8] A. Garain, S. K. Mahata, S. Dutta, Normalization of numeronyms using nlp techniques, in: 2020 IEEE Calcutta Conference (CALCON), 2020, pp. 7–9.
- [9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv (2019) arXiv-1907.
- [10] A. Liaw, M. Wiener, et al., Classification and regression by randomforest, R news 2 (2002) 18–22.
- [11] 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, FIRE '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 14–17. URL: <https://doi.org/10.1145/3368567.3368584>. doi:10.1145/3368567.3368584.
- [12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

- [13] T. Mandl, S. Modha, G. K. Shahi, A. K. Jaiswal, D. Nandini, D. Patel, P. Majumder, J. Schäfer, Overview of the HASOC track at FIRE 2020: Hate Speech and Offensive Content Identification in Indo-European Languages), in: Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, CEUR, 2020.
- [14] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew, Huggingface’s transformers: State-of-the-art natural language processing, CoRR abs/1910.03771 (2019). URL: <http://arxiv.org/abs/1910.03771>.  
`arXiv:1910.03771`.
- [15] J. Howard, et al., fastai, <https://github.com/fastai/fastai>, 2018.
- [16] A. Garain, A. Basu, R. Dawn, S. K. Naskar, Sentence simplification using syntactic parse trees, in: 2019 4th International Conference on Information Systems and Computer Networks (ISCON), 2019, pp. 672–676.