

Machine Learning based hate speech identification for English and Indo-Aryan languages

Anirudh Anand¹, Jeet Golecha¹, B.Bharathi¹, Bhuvana Jayaraman¹ and Mirnalinee T.T¹

¹Department of CSE

Sri Sivasubramaniya Nadar College of Engineering,
Chennai, Tamil Nadu, India

Abstract

Social media platforms pave way for the public to express their opinions. These opinions are mostly on the events and happenings across the world. These comments are most often unbiased and cross the individual boundaries that cause hurt to the people involved. Some comments are intentionally delivered through these platforms with the purpose of offending the party concerned. An automatic technique is needed to identify offensive comments to prevent unwanted consequences. Our work is a part of Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages, where the offensive public opinions on social media platforms are to be identified. We have devised a system that uses both machine learning and deep learning techniques to detect the offensive comments. Random Forest has obtained 78% macro F1 score, in Hindi Recurrent Neural Network performed well with 73% and Support vector machine with 75% macro F1 score.

Keywords

Indo-Aryan Languages, Offensive comments, Text classification, Machine learning, Deep Learning

1. Introduction

With the development of technologies such as Artificial Intelligence, Machine learning, email, internet applications, came the social media and instant messaging applications. With the right to speech, people used these platforms to convey their thoughts and opinions. Most of the time these opinionated thoughts would cross their rightful boundaries and became offensive. Comments made on the social media platform get viral and publicised to get more attention from the public which are often hurtful, and derogatory. The hateful messages are mostly on the premise of religion, one's identity, gender, race, nationality, inclination towards one's ideology and so on. Such comments may incite violence and imbalance to the peacefulness of the society in its worst case. For an individual, the insulting comments may lead to anxiety, depression and mental instabilities.

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

✉ anirudh19015@cse.ssn.edu.in (A. Anand); jeetgolecha19043@cse.ssn.edu.in (J. Golecha); bharathib@ssn.edu.in (B. Bharathi); bhuvanaj@ssn.edu.in (B. Jayaraman); mirnalineett@ssn.edu.in (M. T.T)

🌐 <https://www.ssn.edu.in/staff-members/dr-b-bharathi/> (B. Bharathi);


<https://www.ssn.edu.in/staff-members/dr-j-bhuvana/> (B. Jayaraman);

<https://www.ssn.edu.in/staff-members/dr-t-t-mirnalinee/> (M. T.T)

🆔 0000-0001-7279-5357 (B. Bharathi); 0000-0002-9328-6989 (B. Jayaraman); 0000-0001-6403-3520 (M. T.T)



© 2021 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 hateful expressions can be handled in a few ways manually namely, calling out the inaccuracies in the comments or messages, tackling them politely, challenging them back and refuting them to show the truthful side or facts. Most of the time, handling social media harassment's went out of hand and demands the need for automatic handling of such messages. On the other hand the issue can be taken up on a legal way and could be handled to counter the effects caused by the hurtful messages.

Necessity towards building a computational model to handle the offensive social media comments has been increased in recent times. Social media organizations are working towards automatic handling of such offensive contents in order to preserve the sanity of their platforms. A single generalized model to identify the offensive comments irrespective of the languages they are based on, is the need of the day.

In this paper we investigated the performance of different machine learning and deep learning techniques in classifying the social media comments in English and Indo-Aryan languages into hate and offensive. This work is a part of Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC, 2021) under Forum for Information Retrieval Evaluation (FIRE, 2021). The section 2 provides the survey of existing work, section 3 discusses about the proposed system, experimental results and related discussion are given in section 4 and conclusion with direction for future enhancement in section 5.

2. Literature Survey

Machine learning and deep learning algorithms [1] have been widely used in identifying the offensive comments. This section reports similar such works that have been carried out recently in Indo-Aryan and English languages. For text classification machine learning approaches have performed well in literature so far [2, 3, 4]. Advancements in technologies brought the computationally intensive deep learning algorithms to reality. In text classification [5], [6] TF-IDF based vectorization, and transfer learning-based multilingual BERT technique have given a noticeable performance.

MOLD, the Marathi Offensive Language Dataset has been created to classify the offensive comments in Marathi language [7] having 2500 tweets that are annotated. Authors observed the predictions of closely related languages like Bengali and Hindi to classify the offensive tweets in Marathi. This corpus has three levels of information namely, whether the tweets are offensive or not, categorizes as threat, profanity, insult and to whom it is targeted as individual or group. Several machine learning and deep learning classifiers are applied to identify the tweets of this corpus namely SVM, Logistic Regression, Naive Bayes, CNN, RNN, etc.

A corpus is constructed with linguistic taboos and euphemisms in Nepali language [8]. These are based on racial discrimination, religion and disability with 1000 different taboos. This corpus can be used to identify any hurtful or derogatory comments.

An USAD (Urdu Slang and Abusive words Detection) automatic lexicon-based system was designed [9] to detect the hurtful and offensive slang words in Perso-Arabic-scripted Urdu Tweets. USAD constitute two phases with lexicon building and testing. The model attained a precision of 72.6% tweets were identified as abusive.

Indo-Aryan and Dravidian languages are investigated [10], using multilingual transformers to

identify offensive comments. The work has used cross-lingual word embeddings while designing the model and compared with single multilingual model. Authors have studied the language similarity and typology impacts along with zero shot and few-shot learning techniques. XLM-R with a softmax added as a final layer used for text classification.

3. Methodology Adopted

The proposed system of detecting offensive content from the HASOC2021 Subtask 1 data is described in the following sections. The steps involved in the proposed system are as follows:

1. Data preprocessing
2. Feature extraction
3. Model training
4. Testing

3.1. Pre-processing

Pre-processing is done to prepare the input data for further processing namely by removing the words that do not give meaning of the tweet, removing the special characters and getting the root word from the derived word, etc. The following preprocessing steps are carried out for the training data of all the three languages:

1. Stop word removal
2. Remove numbers
3. Remove special characters
4. Lemmatization through wordnet lemmatizer
5. Stemming through port stemmer

3.2. Feature extraction

Term Frequency Inverse Document Frequency vectors are extracted from the text. The char TF-IDF is calculated as explained in equation 1.

For each char a in a document t from the document set D , TF-IDF is calculated as:
 N is the total number of documents.

$$tfidf(a, t, D) = tf(a, t).idf(a, D) \quad (1)$$

where

$$tf(a, t) = \log(1 + freq(a, t)) \quad (2)$$

$$idf(a, D) = \log\left(\frac{N}{count(d \in D : a \in d)}\right) \quad (3)$$

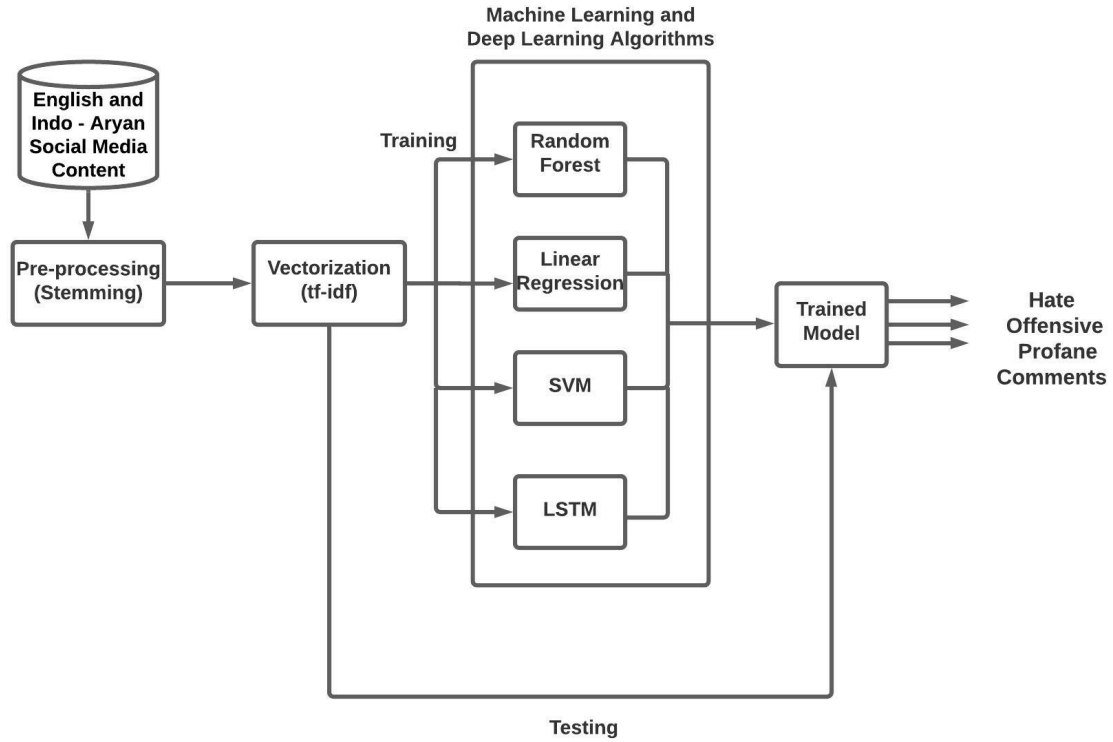


Figure 1: Overview of proposed Hate Speech and Offensive Content Identification model

3.3. Model training

The extracted features (TF-IDF) are used for training the machine learning algorithms such as Random forest, Linear regression, Support vector machine and Recurrent neural network. For Recurrent neural network, an embedding layer is included to get the embeddings for the text that uses default embedding in Keras. The model parameters used for these algorithms are given in Section 4. From the given dataset, 80% of the data used for training and remaining 20% of the data used for model tuning. The machine learning model implementations and the metrics of comparison is used from Scikit-learn ¹.

3.4. Model Testing

The four trained models using Random forest, Linear regression, SVM and LSTM are tested with the test data for evaluating their performance of classification. The metrics observed during the testing are discussed in the following section.

¹<https://scikit-learn.org/stable/>

4. Experimental results and Discussion

4.1. Dataset Description

In this proposed work, Subtask 1A data of HASOC 2021 is used. The data sets are given in three languages namely English, Hindi and Marathi. The datasets are sampled from Twitter. More details about the dataset are given in [11, 12]. Number of samples available for each class the dataset are described in Table 1.

Table 1

Data description

Language	Not Hate offensive (NOT)	Hate and offensive (HOF)
English	1342	2501
Hindi	3161	1433
Marathi	1205	669

Sub-task A focus on Hate speech and Offensive language identification offered for English, Hindi, and Marathi. The objective of this task is to classify the tweets into two classes, namely: Hate and Offensive (HOF) and Non- Hate, and offensive (NOT). In the proposed approach, the preprocessed text is trained using the machine learning algorithms such as Random forest, Linear regression, Support vector machine, and Recurrent neural network for all three languages. For a Random forest classifier, a number of jobs are assigned as 2 and the random state is initialized as 0. For the Support vector machine, the linear kernel is used. For logistic regression classifier, the random state is initialized as 0. For Recurrent neural network, the embedding layer is included to get the embeddings of preprocessed text. The number of LSTM units is 40, adam optimizer is used.

The cross-validation accuracy for English is given in Table 2. From Table 2, it has been noted

Table 2

Performance of proposed system for Subtask 1 - English language

Classifier	Cross validation accuracy (in %)	F1-score	Precision	Recall
Random Forest	80.2	0.68	0.76	0.63
Linear Regression	77.07	0.60	0.78	0.49
Support vector machine	78.02	0.61	0.75	0.67
Recurrent Neural Network	76	0.63	0.77	0.69

that Random forest classifier produces the F1-score of 0.68 and accuracy of 80.2%.

The cross-validation accuracy for Hindi is given in Table 3. From Table 3, it has been noted that Random forest classifier produces the F1-score of 0.86 and accuracy of 79.2%.

The cross-validation accuracy for Marathi is given in Table 4.

Table 3

Performance of proposed system for Subtask 1 - Hindi language

Classifier	Cross validation accuracy (in %)	F1-score	Precision	Recall
Random Forest	79.2	0.86	0.81	0.91
Linear Regression	78.45	0.85	0.79	0.94
Support vector machine	78.34	0.85	0.76	0.93
Recurrent Neural Network	78.69	0.54	0.75	0.90

Table 4

Performance of proposed system for Subtask 1 - Marathi language

Classifier	Cross validation accuracy (in %)	F1-score	Precision	Recall
Random Forest	77.06	0.83	0.80	0.85
Linear Regression	74.93	0.83	0.74	0.95
Support vector machine	79.72	0.84	0.76	0.93
Recurrent Neural Network	78.69	0.75	0.71	0.87

From Table 4, it has been inferred that the performance of support vector machine is better than other models.

The performance of the proposed system using test data is shown in Table 5.

Table 5

Performance of proposed system for sub task 1 using test data

Language	Model used	Macro F1
English	Support Vector Machine	0.75
Hindi	Recurrent Neural Network	0.73
Marathi	Random Forest	0.78

5. Conclusion

The need for automatic identification of offensive comments across social media platforms motivated us to devise a system that uses both machine learning and deep learning techniques. This work has been submitted as a part of Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC, 2021). One generalized approach cannot be designed to perform this classification in English and Indo-Aryan Languages. In the proposed automatic system for offensive comment identification, Random Forest, SVM and RNN have

outperformed with 78%, 75% and 73% of macro F1 score in Marathi, English and Hindi respectively. This can be further explored using multilingual word embeddings using transfer learning deep neural networks and by exploring the various parameters involved in those networks.

References

- [1] L. Deng, Y. Liu, Deep learning in natural language processing, Springer, 2018.
- [2] B. Bharathi, J. Bhuvana, N. N. A. Balaji, Ssnscse_nlp@ evalita2020: Textual and contextual stance detection from tweets using machine learning approach, EVALITA Evaluation of NLP and Speech Tools for Italian-December 17th, 2020 (2020) 224.
- [3] B. Bharathi, M. Anirudh, J. Bhuvana, Bharathi ssn@ inli-fire-2017: Svm based approach for indian native language identification., in: FIRE (Working Notes), 2017, pp. 110–112.
- [4] H. Sandaruwan, S. Lorensuhewa, M. Kalyani, Identification of abusive sinhala comments in social media using text mining and machine learning techniques, International Journal on Advances in ICT for Emerging Regions 13 (2020) 1.
- [5] N. N. A. Balaji, B. Bharathi, J. Bhuvana, Ssnscse_nlp@ dravidian-codemix-fire2020: Sentiment analysis for dravidian languages in code-mixed text., in: FIRE (Working Notes), 2020, pp. 554–559.
- [6] K. Yasaswini, K. Puranik, A. Hande, R. Priyadharshini, S. Thavareesan, B. R. Chakravarthi, Iiitt@ dravidianlangtech-eacl2021: Transfer learning for offensive language detection in dravidian languages, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, 2021, pp. 187–194.
- [7] S. Gaikwad, T. Ranasinghe, M. Zampieri, C. M. Homan, Cross-lingual offensive language identification for low resource languages: The case of marathi, Proceedings of RANLP (2021).
- [8] N. B. Niraula, S. Dulal, D. Koirala, Linguistic taboos and euphemisms in nepali, arXiv preprint arXiv:2007.13798 (2020).
- [9] N. U. Haq, M. Ullah, R. Khan, A. Ahmad, A. Almogren, B. Hayat, B. Shafi, Usad: an intelligent system for slang and abusive text detection in perso-arabic-scripted urdu, Complexity 2020 (2020).
- [10] T. Ranasinghe, M. Zampieri, An evaluation of multilingual offensive language identification methods for the languages of india, Information 12 (2021) 306.
- [11] S. Modha, T. Mandl, G. 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, 2021.
- [12] T. Mandl, S. Modha, G. Shahi, H. Madhu, S. Satapara, P. Majumder, J. Schäfer, T. Ranasinghe, M. Zampieri, D. Nandini, et al., Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages, Working Notes of FIRE (2021).