Hate Speech and Offensive Content Identification in English Tweets

Ritesh Kumar[1], Vishesh Gupta[2] and Rajendra Pamula[2]

[1]*Department of Computer Science and Engineering, National Institute of Technology Jamshedpur, India*
[2]*Department of Computer Science and Engineering, Indian Institute of Technology (Indian School of Mines), Dhanbad, India*

Abstract

Hate speech is a prevalent practice that society has to struggle with everyday. The freedom of speech and ease of anonymity granted by social media has also resulted in incitement to hatred. This presents the need for automatic detection of hate speeches or tweets on social media. In this paper, we have presented the machine learning models that can detect hate Speech and offensive content. Specifically, we described the model submitted for the shared task on hate Speech and offensive content identification in English Tweets at HASOC 2021 and our team name is Vishesh Gupta.The problem concentrates on hate speech detection in English language. The challenge is divided into two tasks of different granularity: (1) coarse-grained binary classification in which participating system are required to classify tweets into two class, namely: Hate and Offensive (HOF) and Non-Hate and offensive (NOT). (2) to predict one of the three types of hate speeches present. Overall, our performance is good but it needs some improvement, our scores are encouraging enough to work for better results in future.

Keywords

LSTM, GRU, Random Forest, TFIDF, XGBoost, Logistic regression, Ktrain

1. Introduction

Nowadays, social media has become a significant part of our lives and just like everything it has its pros and cons. Various benefits of social media come with several challenges including hate speech, offensive and profane content getting published targeting an individual, a group or a society. Hate speech and other offensive content in online socialization have seriously affected daily life of people. Social media companies such as, YouTube, Facebook, and Twitter have their own approaches to eliminate the hate speech content or anything which negatively affects the society. However, detecting such objectionable content at the earliest to curb the menace of spreading such news online is still a major challenge faced by social media companies and researchers. It is very essential to detect such behaviour. The amount of data generated on social media sites can be estimated from the fact that, every second,on average, around 6,000 tweets are generated. Content moderation of such a huge data is difficult to achieve exclusively through man power. Social networking sites are struggling with content moderation.

Forum for Information Retrieval Evaluation, December 13-17, 2021, India
ritesh.cse@nitjsr.ac.in (R. Kumar); me.guptavishesh@gmail.com (V. Gupta); rajendrapamula@gmail.com (R. Pamula)

Artficial Intelligence and different Machine Learning techniques can be exploited for hate speech Detection.

In this paper, we have explored various Machine Learning (ML) algorithms for hate speech and offensive content identification in English language in a shared task called HASOC 2021 [1] of Forum for Information Retrieval Evaluation (FIRE) 2021 and our team name is Vishesh Gupta. As per requirement of HASOC 2021 Subtask-1 [2], we have submitted five runs for Subtask-1A and four runs for Subtask-1B. We have extracted dfferent lexical and non lexical features from the text for the classification. Our best run in subtask-1A has achieved Macro-F1 score of 0.7680. For subtask-1B, our best run was with an F1-score of 0.5871.

2. Related Work

Several works have been proposed to detect hate speech and offensive content across social platforms. Hajim et. al. [3] proposed a approach to collect hateful and offensive expressions and perform Hate Speech Detection. Muhammad Okky Ibrohim et. al. [4] proposed a Multi-label Hate Speech and Abusive Language Detection in Twitter. M. Ali Fauzi [5] used Ensemble Method for Indonesian Twitter Hate Speech Detection.Anusha M D et. al. [6] proposed an Ensemble Model for Hate Speech and Offensive Content Identification in Indo-European Languages in HASOC 2020 in which they combine CountVectorizer and TF-IDF transformer with additional text-based features to build an ensemble of Gradient Boosting, Random Forest and XGBoost classifiers, with soft voting.

Tharindu Ranasinghe et. al. [7] submitted thier model in HASOC-2019 in which they evaluated six different neural network architectures for the classification tasks: pooled Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM) and GRU with Attention , 2D Convolution with Pooling, GRU with Capsule and LSTM with Capsule and Attention.

Urmi Saha . [8] submitted her model in HASOC 2019 in which she used a list of hate words for feature engineering to build ML approaches for English and their approach on the test set provided by HASOC 2019 achieved accuracies of 0.68, 0.65 and 0.66 for English language subtasks 1, 2 and 3 respectively.

Sarthak Gupte et. al. [9] participated in HASOC-2020 task 1 i.e. Offensive comment identification in Code-mixed Malayalam Youtube comments in which they used cross-lingual contextual word embeddings and transfer learning to make predictions to Malayalam data. P.Karthikeyan et. al. [10] proposed a research paper on Hate Speech Detection with Hateful and Offensive Expressions on Twitter using various Machine Learning Techniques where they show various concepts of sentiment analysis.

3. Task and Dataset Description

In this section, we have described the hate speech and offensive content identification shared task and the dataset provided to the participants.

HASOC 2021 shared task is divided into two subtasks(Subtask-1 and Subtask-2). Subtask-1 is further divided into two subtasks (Subtask-1A and Subtask-1B).

SubTask-1A basically aims to classify tweets into two class, namely: Hate and Offensive (HOF) and Non-Hate and offensive (NOT). (shown in Table 1):

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

Table 1
Categorization of Tweets in Subtask-1A with Example

Category	Example
NOT	It's heartbreaking to see Osama struggling alone for the judicial killing of his father.
NOT	Oh wow, I had my first vaccine yesterday, today I feel awful
HOF	@ndtv Shameless PM. What else can we say?
HOF	Violence = IslaM Love = "behayayi" It's totally unethical ,unislamic.

In Subtask-1B, hate-speech and offensive posts from the Subtask-1A are further classified into four categories. (shown in Table 2):

1. (HATE) Hate speech :- Posts under this class contain Hate speech content.
2. (OFFN) Offensive :- Posts under this class contain offensive content.
3. (PRFN) Profane :- These posts contain profane words.
4. (NONE) Not Hate-offensive:- This post does not contain any Hate speech, profane, offensive content.

Table 2
Categorization of Tweets in Subtask-1B with Example

Category	Example
NONE	It's heartbreaking to see Osama struggling alone for the judicial killing of his father.
OFFN	@ndtv Shameless PM. What else can we say?
HATE	Violence = IslaM Love = "behayayi" It's totally unethical ,unislamic.
PRFN	God, did you hear about that stupid guy who works at Burger King who is a total asshole?

We have used the dataset available at HASOC 2021. The dataset consists of 3,843 tweets for training and 1,281 tweets for testing with balanced distribution of each classes. Data was interpreted at two different levels of granularity. First, each text was labelled as 'HOF' or 'NOT'. Secondly, levels are further divided as 'NONE', 'HATE', 'OFFN' and 'PRFN'.

4. System Description

4.1. Text Preprocessing

We have removed all the links, punctuations, numbers and stop words. We have used lemmatization for grouping together the different forms of a word into a single word. NLTK wordnet [11] is used for lemmatization. Both Subkask-1A and Subtask-1B uses same preprocessing.

4.2. Feature Extraction

TfidfVectorizer [12] is used for converting the text into numerical features.Pipeline[1] is used for doing TfidfVectorizer and classification in pipelined manner.Tokenizer by keras library is used for LSTM. For Logistic regression , Random Forest and XGBoost, we have used TfidfVectorizer from scikit-learn library. Glove [13] is used to create word embeddings and GRU model is used with this glove for classification.

4.3. Machine Learning Models

For Subtask-1A, we have submitted five runs based on five different algorithms, namely- Logistic Regression [14], LSTM [15], ktrain [16], XGBoost [17] and Glove+GRU [18]. We have used the scikit-learn library for logistic regression based models and Keras for LSTM. GloVe is an unsupervised learning algorithm for getting vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus. We have used pre-trained word vectors of twitter data for training in Glove+GRU model. We scored maximum Macro F1 score of 0.7680 using GRU for subtask-1A. We have used the following values of the parameter :

1. For TfidfVectorizer, we have used mindf=20, maxfeatures=2000 and maxdf=0.6.
2. In XGBoost, we have used learning rate=0.1, max depth=7 and n estimators=150.
3. For LSTM and GRU, we have used batch size of 256 and 10 epochs for training data .

For Subtask-1B, we have submitted four runs based on four different algorithms, namely Logistic regression,Random Forest, Ktrain and XGBoost.We have treated subtask 1B as a multi-classification problem with 4 categories.The parameter values were the same as mentioned above. And Random Forest was implemented using n estimators=1000. We have scored maximum Macro F1 score of 0.5871 using XGBoost for subtask-1B.

5. Results and Discussion

The results of Subtask-1A are represented in terms of Macro-F1,Macro Precision, Macro Recall and Accuracy (shown in Table 3), and the results of Subtask-1B are also in terms of Macro-F1,Macro Precision, Macro Recall and Accuracy (shown in Table 4). The best score as Macro-F1, we get from Subtask-1A is 0.7680. For Subtask-1B we get best score as Macro-F1 is 0.5871. Table 3 and 4 shows the score of our submissions based on HASOC official ranking. Our best system was ranked 27 in Subtask-1A and 22 for Subtask-1B.

For Subtask-1A, the Glove + GRU system have performed better than all other models.For Subtask-1B, the XGBoost system have performed better than all other models.The accuracy and Macro-F1 score obtained in subtask-1B was lower than that of subtask-1A due to more number of categories of classifiction in subtask-1B. The same situation could be found in results of all the teams. This can be due to the fact that classification of the hate speeches text into finer granularity is a much more difficult task than detecting instances of hate speech. Also, as

[1]https://chrisfotache.medium.com/text-classification-in-python-pipelines-nlp-nltk-tf-idf-xgboost-and-more-b83451a327e0

Table 3
Results for Subtask-1A: The official Evaluation measure is Macro F1. The best score obtained by us is mentioned in bold

Run	Macro F1	Macro Precision	Macro Recall	Accuracy
Glove + GRU	**0.7680**	**0.7733**	**0.7741**	**78.37%**
XGBoost	0.7514	0.7881	0.7404	77.98%
Logistic	0.7331	0.7696	0.7215	76.42%
Ktrain	0.7345	0.7543	0.7324	75.95%
LSTM	0.7144	0.7329	0.7181	74.23%

Table 4
Results for subtask-1B: The official Evaluation measure is Macro F1. The best score obtained by us is mentioned in bold

Run	Macro F1	Macro Precision	Macro Recall	Accuracy
XGBoost	**0.5871**	**0.5883**	**0.6137**	**65.496%**
Random Forest	0.5827	0.5810	0.6108	65.417%
Logistic	0.5648	0.5674	0.5982	64.871%
Ktrain	0.5330	0.5399	0.5390	58.782%

subtask-1B was multi-class classification problem , it's macro F1 score was lower as compared to the binary classification problem in subtask-1A.

6. Conclusion and Future Work

We have completed the task using various classification algorithms and evaluated the performance of different classification algorithms for Hate Speech and Offensive Content in English Tweets this year's shared task.Our overall rank is **27** for subtask-1A and **22** for subtask-1B which were average as compared to other submissions obtained in the HASOC 2021 shared task. We look forward to experimenting with different advance algorithm or neural network models. Also, fine tuning the parameters of the algorithm can help in improvement of the overall performance. And the results of more than one classification algorithm can be combined to generate an overall better score. We shall be exploring these tasks in the coming days.

References

[1] 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.

[2] 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/.

[3] H. Watanabe, M. Bouazizi, T. Ohtsuki, Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection, IEEE Access 6 (2018) 13825–13835. doi:10.1109/ACCESS.2018.2806394.

[4] M. O. Ibrohim, I. Budi, Multi-label hate speech and abusive language detection in indonesian twitter (2019) 46–57.

[5] M. A. Fauzi, A. Yuniarti, Ensemble method for indonesian twitter hate speech detection, Indonesian Journal of Electrical Engineering and Computer Science 11 (2018) 294–299.

[6] M. Anusha, H. Shashirekha, An ensemble model for hate speech and offensive content identification in indo-european languages. (2020) 253–259.

[7] T. Ranasinghe, M. Zampieri, H. Hettiarachchi, Brums at hasoc 2019: Deep learning models for multilingual hate speech and offensive language identification. (2019) 199–207.

[8] U. Saha, A. Dubey, P. Bhattacharyya, Iit bombay at hasoc 2019: Supervised hate speech and offensive content detection in indo-european languages. (2019).

[9] T. Ranasinghe, S. Gupte, M. Zampieri, I. Nwogu, Wlv-rit at hasoc-dravidian-codemix-fire2020: Offensive language identification in code-switched youtube comments, arXiv preprint arXiv:2011.00559 (2020).

[10] P. KARTHIKEYAN, B. JYOTHI, Hate speech detection with hateful and offensive expressions on twitter using various machine learning techniques (????).

[11] E. Loper, S. Bird, Nltk: The natural language toolkit, arXiv preprint cs/0205028 (2002).

[12] V. Kumar, B. Subba, A tfidfvectorizer and svm based sentiment analysis framework for text data corpus, in: 2020 National Conference on Communications (NCC), 2020, pp. 1–6. doi:10.1109/NCC48643.2020.9056085.

[13] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[14] Logistic regression (2010) 631–631. URL: https://doi.org/10.1007/978-0-387-30164-8_493. doi:10.1007/978-0-387-30164-8_493.

[15] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.

[16] A. S. Maiya, ktrain: A low-code library for augmented machine learning, arXiv preprint arXiv:2004.10703 (2020).

[17] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[18] R. Ni, H. Cao, Sentiment analysis based on glove and lstm-gru, in: 2020 39th Chinese Control Conference (CCC), 2020, pp. 7492–7497. doi:10.23919/CCC50068.2020.9188578.