Classification of Hate, Offensive and Profane content from Tweets using an Ensemble of Deep Contextualized and Domain Specific Representations

Basavraj Chinagundi[1], Muskaan Singh[2], Tirthankar Ghosal[2], Prashant Singh Rana[1] and Guneet Singh Kohli[1]

[1]*Thapar Institute of Engineering and Technology, India*

[2]*Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Czech Republic*

Abstract

The explosive growth of social media has also resulted in unfortunate emergence of hate, offensive, and profane content on the web. A certain conversational thread can contain hate, offensive, and profane content, which is not apparent from a standalone or single tweet or replies but can be identified if given the context of the parent content. Such social media content is spread in many different languages, including code-mixed languages like hinglish (English code-mixed with Hindi). So it becomes a huge responsibility for the social media sites to identify such hate content before it gets disseminated to the general population, which may trigger havoc. The hate speech and offensive content identification track (HASOC)[1] in FIRE 2021 English Subtask A track provides a forum and a data challenge for multilingual research on the identification of such problematic content. In this paper, we describe our submission for the above track. Our proposed approach uses a transformer-based embedding with HateBERT and achieves the Macro F1 score of 79% on the test data, which is 3.96% behind the best-performing system. We make our system run available at https://github.com/basavraj-chinagundi/HASOC_2021

Keywords

hate Speech, Text Classification, Profane Content, HateBERT

1. Introduction

Social media sites like Twitter and Facebook, being user-friendly and a free source , provide opportunities for people to air their voices. People, irrespective of age group, use these sites to share every moment of their lives, making these sites flooded with data. Apart from these commendable features of social media, they also have downsides as well. Due to the lack of restrictions set by these sites for their users to express their views as they like, anyone can make adverse and unrealistic comments in abusive language against anybody with an ulterior motive to tarnish one's image and status in society. A conversational thread can also contain hate content, which is not apparent just from a single comment or the reply to a comment, but can be identified if given the context of the parent content. Furthermore, the contents on such social media are spread in so many different languages, including code-mixed languages such as hinglish. So it becomes a huge responsibility for these sites to identify such hate content before it disseminates to the masses. The best performing model in our study is based on

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

Transformer contextual embedding and HateBERT architecture. When compared to traditional and ensembled machine learning models, the presented solution enhances accuracy by 6-9% on average. The HateBERT based model achieves a competitive f1 score of 0.7909, demonstrating potential for further improvement in performance

2. Related Work

Most of the previous works in the field use surface features, word embedding features, lexical resources, meta-information, linguistic study, cross-domain information, dealing with biases and multi-task learning. The problem of classifying any sentence as hate is challenging. There might be cases when a sentence containing slang might be classified as hate, because of the words having different meanings in a different context. It affects the right to freedom of speech with the kind of words being used on social media. [2] used SVM with syntactic and semantic information of word-n-grams. [3] presents logistic regression model performance with feeding TF-IDF values and unigram, bigrams, trigrams featured weights achieve 90% precision with 61% correctly predicting hate class. [4] classified ontological classes of harmful speech based on different parameters such as degree of content, intent and its effect on social media. [5] annotated a dataset of 16K tweets using race theory releases publicly. They also did performance analysis on geographic, word length distribution does not have a significant impact while gender information combined with char-n-grams shows some significant improvement. [6] uses four features (such as linguistic, syntactic, distributional, n-gram) to differentiate between abusive and clean features in news and financial data. [7] identified racist and radicalized intent on Tumblr microblogging website using semantic, sentiment and linguistic features with cascaded ensemble learning classifier. [8] provided an annotated corpora of 80K tweets categorized into 8 labels, for studying different types of abusive behaviour. [9] uses features such as sentiment, semantic, unigrams and pattern-based to classify 2010 sentences. [10] released a data set of 2435 tweets on refugees and Muslims and a new novel approach using CNN-GRU architecture. Their approach shows promising results for 6 out of 7 data sets. outperforming other state-of-art by 13 F1 scores. [11] applied bag-of-words to learn a classifier for the labels racist and non-racist with 76% accuracy. [12] combines LSTM model and neural-based GBDT word embedding on dataset [5]. [13] combined char-CNN and word-CNN by formulating a hybrid CNN which performed well than classic methods like logistic regression and SVM on a data set of 16k tweets by [5]. It firstly detected the abusive language and then classified it into specific types of abuse. Further [14] also use CNN with random vectors, word vectors based on semantic information, word vectors combined with character 4-grams. It also presents a comparative performance analysis.

3. Task Description

A conversational thread contains hate, offensive, and profane content, which is not apparent from a standalone or single tweet or comment or the reply to a comment, but can be identified if given the context of the parent content is known. In reference to fig:example the screenshot from Twitter describes the problem at hand effectively. The parent/source tweet, which was posted at 2:30 am on May 11th, expresses Hate and profanity towards Muslim countries regarding the

Figure 1: Example of Contextual misinterpretation[15]

controversy happening during the recent Israel-Palestine conflict. The 2 comments on the tweet have written "Amine", which means trustworthy or honest in Arabic. If the 2 comments were to be analyzed for hate or offensive speech without the context of the parent tweet, they would not be classified as hate or offensive content. But if we take the context of the conversation, then we can say that the comments support the hate expressed in the parent tweet. So those comments are labelled as hate/offensive/profane. The English sub-task A [16] focused on the binary classification of such conversational tweets with tree-structured data into:

- **(NOT) Non hate-Offensive** This tweet, comment, or reply does not contain any hate speech, profane, offensive content.
- **(HOF) hate and Offensive** This tweet, comment, or reply contains hate, offensive, and profane content in itself or supports hate expressed in the parent tweet.

Another such example with code mixed text. The Source Tweet: "Modi Ji COVID situation ko solve karne ke liye ideas maang rahe the. Mera idea hai resignation dedo please. "

- Translation : Modi ji (PM of India) was asking for ideas to solve the covid situation of India. My idea to him is to resign.
- The Comment: Doctors aur Scientists se manga hai. Chutiyo se nahi. Baith niche. [HOF]
- Translation: They have asked Doctors and Scientists. Not fuckers. Sit down. [HOF]

- The reply: You totally nailed it, can't stop laughing. [HOF]

The reply has a positive sentiment. But it is positive in favour of the hate expressed towards the author of the source tweet in the comment. Hence, it is supporting the hate expressed in the comment. Hence, it is also hate speech. This is the type of problem we're aiming to solve via this shared task.

4. Dataset Description

We experiment with a collection of diverse datasets comprising of foul and offensive tweets, comments acquired from various sources. The dataset[17] used for training consists a total of 76601 texts which are either hate speech and offensive (40823) or normal (35778). We have collected these samples from namely 6 sources like:

1. **HSOL[18]**: HSOL is a dataset for hate speech identification that includes a hate speech lexicon including words and phrases recognised as hate speech by internet users and collected by hatebase.org. Using the Twitter API, they searched for tweets containing phrases from the lexicon, yielding a sample of tweets from 33,458 Twitter users. They retrieved the timeline for each user, yielding a collection of 85.4 million tweets. From this dataset, they selected a random sample of 25k tweets containing lexical words and had them manually coded by CrowdFlower (CF) employees. Workers were asked to categorise each tweet into one of three categories: hate speech, offensive but not hate speech, or neither offensive nor hate speech.
2. **OLID[19]**:OLID is a hierarchical dataset to identify the type and the target of offensive texts in social media. The dataset was compiled via Twitter and is freely accessible to the public. There are 14,100 tweets in all, with 13,240 in the training set and 860 in the test set. There are three degrees of labelling for each tweet: (A) Offensive/Not-Offensive, (B) Targeted-Insult/Untargeted, and (C) Individual/Group/Other. If a tweet is offensive, it might have a target or no target. If it is offensive to a specified target, the target might be an individual, a group, or any other thing. This dataset is utilised in the OffensEval-2019 competition at SemEval-2019.
3. **hatespeech [20]**: Dataset of hate speech annotated at the sentence level from Internet forum postings in English. Stormfront, a prominent online community of white nationalists, is where the source forum can be found. A total of 10,568 sentences were taken from Storm front and labelled as hate speech or not.
4. **TRAC[21]**: The data set consists of 15,000 aggression-annotated Facebook Posts and Comments that include labels for three-way categorization of text data into 'Overtly Aggressive, ' 'Covertly Aggressive, ' and 'Non-aggressive. '
5. **ETHOS[22]**: ETHOS is a data set for detecting hate specks. It is made up of YouTube and Reddit comments that have been verified using a crowd sourcing tool. It is divided into two subsets: one for binary classification and one for multi-label classification. The one used for our experiment is the binary subset. The former has 998 comments, whereas the latter has 433 comments with fine-grained hate-speech annotations.

6. **HASOC[23]**: The dataset focuses on hate speech and offensive language detection in English. The data set is classified into two classes, namely: hate and Offensive (HOF) consisting 5051 tweets and Non- hate and offensive (NOT) consisting 5798 tweets respectively.

5. Methodology

In the methodology we describe the pre-processing and explain different baselines along with submitted experiments for classifying hate content as shown in Figure 2.

5.1. Pre-processing

We first went with lower casing each tweet/comment in the data set. Secondly, hashtags are very critical while retrieving sentiment of a text, therefore we preprocess the hashtags using a tailored technique. We start by creating a data frame of all hashtags in a column and their counts. After that we remove numbers and segment multiple words using hash fix function which basically splits the word into segments using the word segment library. Finally we create a dictionary of the hashtags and their clean strings. For example if we have hashtags consisting of multiple words such as #fuckdick it will split the above token into fuck and dick respectively, enhancing the ability to retrieve significant words which are critical for classification of the text as a negative sentiment. We further remove other irrelevant parts of the texts such as usernames, some special characters, retweet tags etc.

5.2. Model Description

The next step is extracting features from the text for which we use TF-IDF vectorizer to transform text into a meaningful representation of numbers which is then used to fit machine algorithm such as NB, LR, KNN, SVM, DT, RF, Bagging, AdaBoost and voting in Table 1 for classifying our text as hate speech or not. We also experiment with another word embedding technique GloVe (840B tokens, 2.2M vocab, cased, 300 dimension vectors) which is an unsupervised learning algorithm for obtaining vector representations for words.

We test out multiple machine learning algorithms and also use ensemble learning in order to produce one optimal predictive model. Now to produce even better results, we try out transformer based pre-trained models:

1. Ernie 2.0 [24]
2. Twitter Roberta Base Offensive [25]
3. HateBERT [20]

The deep learning based models have their own embeddings which were used to extract features from the text. In the final step we fine tune these three models on our combined dataset and boost the results for the classification of text as hateful and offensive.

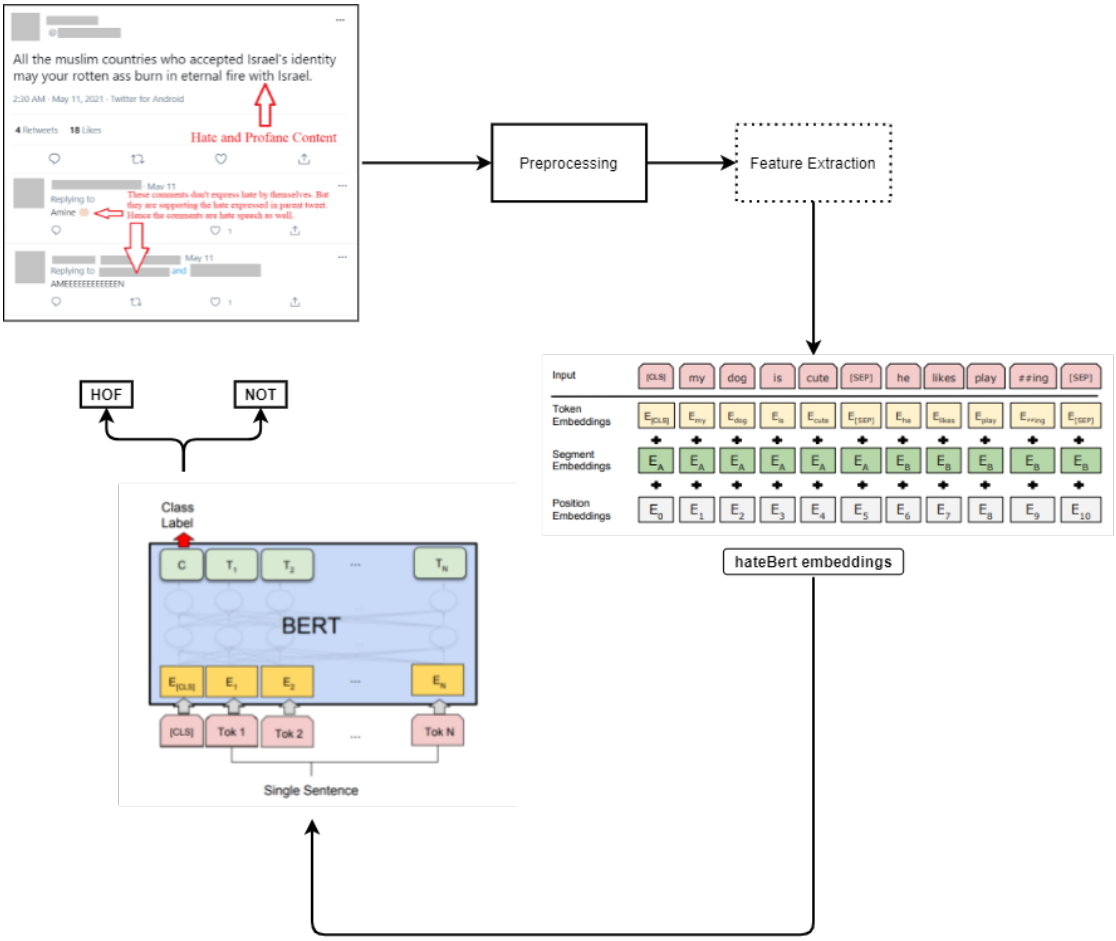

Figure 2: Our Proposed Architecture

6. Experiment and Results

After extracting features with TF-IDF We first use a logistic regression with L2 regularization as it disperses the error terms in all the weights and leads to more accurate customized final models. We then test a variety of models that have been used in prior work: logistic regression, naive Bayes, decision trees, random forests, k-nearest neighbors (KNN), and linear SVMs. We then try bagging method using decision tree classifier with parameters set as *max_samples=0.5, max_features=1.0, n_estimators=10.* We also check AdaBoostClassifier using decision trees with parameters *min_samples_split=10, max_depth=4, n_estimators=10, learning_rate=0.6.* We tested each model using 5-fold cross validation, holding out 10 percent of the sample for evaluation to help prevent over-fitting. After using a grid-search to iterate over the models and parameters we find that the Logistic regression, naive bayes, random forest and Linear SVM tend to perform significantly better than other models. So by ensembling these we make another model using voting classifier. When comparing all these models we see that logistic regression

with tfidf vector representation performs the best having 0.77 accuracy, macro avg f1 score of 0.75 and weighted avg f1 score of 0.76 respectively. We then experiment using GloVe represtation and ensembling ML algorithms namely naive bayes, logistic regression and multilayer perceptron and find out that it doesn't necessarily boost the performance of our model achieving accuracy of 0.72, macro avg f1 score of 0.71 and weighted avg f1 score of 0.72. Thus we move onto our final set of experimentations. We use transformer based pretrained models as a transformer is to able to parallely process the words in the sentences and get contexualized embeddings. This parallel processing is not possible in LSTMs or RNNs or GRUs as they take words of the input sentence as input one by one. We ran all three pretrained models for 5 epochs by fine tuning it with hyperparameters having batch size 16 and Adam optimizer with learning rate 1e-5 , eps=1e-8. Ernie 2.0 improved the performance in comparison to the previous experiments we ran using word embeddings and machine learning algorithms, achieving 0.80 accuracy, macro avg f1 score of 0.78 and weighted avg f1 score of 0.80 respectively. TwitterRobertaBaseOffensive which is a pretrained model trained on 58M tweets and finetuned for offensive language identification with the TweetEval benchmark further increased the performance attaining 0.81 accuracy, macro avg f1 score of 0.79 and weighted avg f1 score of 0.81. Finally we test another pretrained model HateBERT which was was trained on RAL-E, a large-scale dataset of Reddit comments in English from communities banned for being offensive, abusive, or hateful that we have collected and made available to the public. It performed the best optimally amongst all the pretrained models achieving same accuracy and f1 scores as TwitterRobertaBaseOffensive but having slightly better performance in recall metrics for classifying text as HOF and NOT hate respectively, overall leading to a much more balanced model. The Google colab's Tesla P100-pcie-16GB with 8 core CPU and 32GB RAM was used for the experimental setup.

7. Conclusion

In this paper, we present our submission to classify hate content from the tweets and comments. The recent trend of hateful speech has increased and has posed a lot of challenge in discriminating hate speech against freedom of speech. One post can mean differently in different context as there is no universally accepted definition of hate speech.

There are different benchmark depending upon demography, social influence and cultural factors. We propose, a deep learning model based on Transformer contextual embedding and HateBERT architecture. We pre-processed the tweet from HASOC 2021 data set, extracted features embedding and trained our system to classify into hate speech or not with 79% macro F1 score. The work compiled showcases the scope of HateBERT in being employed for further experimentation and being optimised for better performances by focusing on newer embedding combinations and ensemble approaches.

Table 1
TF-IDF embeddings based model performance comparison

MODELS	precision	recall	macro f1	weighted f1	accuracy
NB	0.72	0.72	0.72	0.74	0.74
LR	0.75	0.74	0.75	0.76	0.77
KNN	0.59	0.56	0.46	0.43	0.48
SVM	0.73	0.72	0.72	0.74	0.74
DT	0.67	0.66	0.66	0.68	0.69
RF	0.76	0.60	0.58	0.63	0.69
Bagging	0.71	0.70	0.70	0.72	0.72
AdaBoost	0.64	0.64	0.63	0.64	0.63
Voting	**0.74**	**0.73**	**0.74**	**0.75**	**0.75**

Table 2
GloVe embeddings based ensemble model performance

MODEL	precision	recall	macro f1	weighted f1	accuracy
glove+Voting	0.71	0.72	0.71	0.72	0.72

Table 3
Transformer based pretrained models comparison

Submission	MODELS	precision	recall	macro f1	weighted f1	accuracy
Baseline	*ERNIE 2.0*	0.80	0.77	0.78	0.80	0.80
Baseline	*TwitterRobertaOff*	0.81	0.78	0.79	0.81	0.81
Submitted	**hateBERT**	**0.81**	**0.79**	**0.79**	**0.81**	**0.81**

Table 4
Comparison of our submission with the other submission in the HASOC@FIRE2021[26] English Subtask-A Shared task[1] with team giniUS

Maximum F1 across all submission	0.83
Minimum F1 across all submission	0.50
Average F1 across all submission	0.75
Our submission	**0.79**

References

[1] 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, 2021.

[2] H. Chen, S. McKeever, S. J. Delany, Abusive text detection using neural networks., in: Artificial Intelligence and Cognitive Science (AICS), 2017, pp. 258–260.

[3] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 11, 2017.

[4] S. Sharma, S. Agrawal, M. Shrivastava, Degree based classification of harmful speech using

twitter data, arXiv preprint arXiv:1806.04197 (2018).
[5] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL student research workshop, 2016, pp. 88–93.
[6] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive language detection in online user content, in: Proceedings of the 25th international conference on world wide web, 2016, pp. 145–153.
[7] S. Agarwal, A. Sureka, Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website, arXiv preprint arXiv:1701.04931 (2017).
[8] A. M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, N. Kourtellis, Large scale crowdsourcing and characterization of twitter abusive behavior, in: Twelfth International AAAI Conference on Web and Social Media, 2018.
[9] 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.
[10] Z. Zhang, D. Robinson, J. Tepper, Detecting hate speech on twitter using a convolution-gru based deep neural network, in: European semantic web conference, Springer, 2018, pp. 745–760.
[11] I. Kwok, Y. Wang, Locate the hate: Detecting tweets against blacks, in: Twenty-seventh AAAI conference on artificial intelligence, 2013.
[12] 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.
[13] J. H. Park, P. Fung, One-step and two-step classification for abusive language detection on twitter, arXiv preprint arXiv:1706.01206 (2017).
[14] B. Gambäck, U. K. Sikdar, Using convolutional neural networks to classify hate-speech, in: Proceedings of the first workshop on abusive language online, 2017, pp. 85–90.
[15] Hasoc, https://hasocfire.github.io/hasoc/2021/index.html, 2021. Accessed: 2021-11-27.
[16] 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).
[17] B. Dave, S. Bhat, P. Majumder, Irnlp_daiict@ lt-edi-eacl2021: Hope speech detection in code mixed text using tf-idf char n-grams and muril, in: Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, 2021, pp. 114–117.
[18] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: Proceedings of the 11th International AAAI Conference on Web and Social Media, ICWSM '17, 2017, pp. 512–515.
[19] S. Rosenthal, P. Atanasova, G. Karadzhov, M. Zampieri, P. Nakov, A large-scale semi-supervised dataset for offensive language identification, arXiv preprint arXiv:2004.14454 (2020).
[20] T. Caselli, V. Basile, J. Mitrović, M. Granitzer, Hatebert: Retraining bert for abusive

language detection in english, arXiv preprint arXiv:2010.12472 (2020).
[21] Trac, https://sites.google.com/view/trac1/home, 2021. Accessed: 2021-11-27.
[22] I. Mollas, Z. Chrysopoulou, S. Karlos, G. Tsoumakas, Ethos: an online hate speech detection dataset, arXiv preprint arXiv:2006.08328 (2020).
[23] P. Alonso, R. Saini, G. Kovács, Hate speech detection using transformer ensembles on the hasoc dataset, in: International Conference on Speech and Computer, Springer, 2020, pp. 13–21.
[24] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, H. Wang, Ernie 2.0: A continual pre-training framework for language understanding, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 8968–8975.
[25] F. Barbieri, J. Camacho-Collados, L. Neves, L. Espinosa-Anke, Tweeteval: Unified benchmark and comparative evaluation for tweet classification, arXiv preprint arXiv:2010.12421 (2020).
[26] S. Satapara, S. Modha, T. Mandl, H. Madhu, P. Majumder, Overview of the hasoc subtrack at fire 2021: Conversational hate speech detection in code-mixed language, Working Notes of FIRE (2021).