SN Computer Science (2021) 2:26
https://doi.org/10.1007/s42979-020-00414-6

ORIGINAL RESEARCH

Aggressive and Offensive Language Identification in Hindi, Bangla, and English: A Comparative Study

Ritesh Kumar[1] · Bornini Lahiri[2] · Atul Kr. Ojha[3,4]

Received: 15 May 2020 / Accepted: 25 November 2020

Abstract

In the present paper, we carry out a comparative study between offensive and aggressive language and attempt to understand their inter-relationship. To carry out this study, we develop classifiers for offensive and aggressive language identification in Hindi, Bangla, and English using the datasets released for the languages as part of the two shared tasks: hate speech and offensive content identification in Indo-European languages (HASOC) and aggression and misogyny identification task at TRAC-2. The HASOC dataset is annotated with the information about offensive language and TRAC-2 dataset is annotated with the information about aggressive language. We experiment with SVM as well as BERT and its different derivatives such as ALBERT and DistilBERT for developing the classifiers. The best classifiers achieve an impressive *F*-score in between 0.70 and 0.80 for different tasks. We use these classifiers to cross-annotate the two datasets, and look at the co-occurrence of different sub-categories of aggression and offense. The study shows that even though aggression and offense significantly overlaps, but still one does not entail the other.

Keywords Aggression · Offensive language · Hindi · Bangla · English · Comparison · TRAC · HASOC · BERT

Introduction

The surge in the users as well as activity on social media has co-occurred with an increase in the use of aggressive and offensive language in the last decade. The velocity and volume of content generation on social media has rendered the traditional, manual methods of content governance, and content moderation largely ineffective and insufficient. However, the problem of offensive and aggressive content on social media has been widely recognised by NLP researchers across the globe. A large number of researchers have been conducting research into automatic identification of related behaviours such as such as trolling [8, 29, 34, 35], cyberbullying [11, 12, 16, 23, 38], flaming/insults [37, 43], abusive/offensive language [9, 39, 49, 51, 52], hate speech [3, 7, 13, 17–19, 31, 45, 48, 50], radicalisation [1, 2], racism [20, 21], and others. Quite a few of the researchers have also organised shared tasks focussed on automatic detection of offensive language (GermEval [46]; OffensEval [53, 54]; HASOC [32]), hate (HatEval [4]), and aggression (TRAC 1 [28] and TRAC 2). These have resulted in the availability of large datasets for developing automatic recognition systems for some of the world's major languages such as German, English, Spanish, Hindi, Bangla, and others.

The availability of datasets from different perspectives and annotated with information of different kinds presents a unique opportunity to explore and understand the same/similar phenomenon from various angles and develop a multi-pronged, nuanced understanding of what we are dealing with. This, in turn, will contribute towards developing better systems for the automatic identification of aggressive/offensive/abusive/hate(ful) speech on social media. One of the major challenges in the field, however, is a lack of understanding of the interactions among these different

This article is part of the topical collection "Social Media Analytics and its Evaluation" guest edited by Thomas Mandl, Sandip Modha, and Prasenjit Majumder.

✉ Ritesh Kumar
ritesh78_llh@jnu.ac.in

[1] Department of Linguistics, K. M. Institute of Hindi and Linguistics, Dr. Bhimrao Ambedkar University, Agra, India

[2] Department of Humanities and Social Sciences, Indian Institute of Technology, Kharagpur, India

[3] Panlingua Language Processing LLP, New Delhi, India

[4] DSI, NUIG, Galway, Ireland

Published online: 09 January 2021

phenomena. Moreover, it has also resulted in duplication of research, to certain extent, and a certain kind of lack of focus and reusability of datasets across different strands of research. To make improvements towards solving a complex phenomenon like this, it is of utmost importance that some kind of uniform understanding of problem be achieved, so that, at least, standardised datasets and an understanding of different approaches to solving the problem may be developed.

Díaz-Torres et al. [15] have proposed a three-way distinction between offensive, aggressive, and vulgar language. They propose that offensive language aims at insulting or humiliating a group or individual, usually using derogatory terms; on the other hand, aggressive language seeks to hurt a group of individual by inciting violence. Significantly, authors also recognise that both may overlap in the same text; however, they do not make a very clear distinction in between the two phenomena, at least, as far as their annotation process is concerned. In their annotation flowchart, the question for deciding aggression reads like the following—'does it refer to or incite violence against the referent, or does it somehow force their will?'—and one of the questions for deciding offense reads like this—'is its intention to insult, humiliate, hurt or harm the referent in any way?' Both these questions are quite similar in the sense that they both refer to the ideas of hurt, violence, and harm. However, the authors also mention the use of 'pejorative, derogatory or negative intensifiers of a term' as offensive (but not aggressive). Thus, even though authors try to make a distinction between the two phenomena, their annotation process ultimately ends up positing aggressive language as a sub-category of offensive language. In this paper, we try to tease apart this elusive distinction between offensive and aggressive language.

Waseem et al. [49] makes an attempt to unify the different trends of research in what may be considered a significantly overlapping field and proposes a two-way typology for understanding what they call 'abusive language' over the web. They propose two scales on which abusive language could be categorised—the target of the abuse (an individual or a group) and the nature of the language (explicit or implicit). While this proposal looks good, it tries to arrive at a completely new and generic typology. It does not explain the inter-relationship across different phenomena being studied within the field nor does it take into consideration their complexity.

There are two broad goals of the present study—first is to develop an offensive and aggressive language identification system for Hindi, Bangla, and English; second is to address the issues related to approaching the problem from multiple perspectives and look at the inter-relationship between two phenomena of offensive language and aggressive language. We use the datasets released in HASOC shared task and TRAC-2 shared task for this study. We develop different classifiers for the different sub-levels in both datasets. We experiment with different kinds of classifiers including the state-of-the-art BERT models as well as its derivatives (such as ALBERT and DistilBERT). The classifiers are then used to cross-annotate both the train and test of the other task, i.e., the HASOC dataset is annotated with the information about aggression and TRAC-2 dataset is annotated with the information about offensive and hateful language. This yielded a corpus that was annotated with information about both aggressive as well as offensive language. We use this annotated dataset for comparison of co-occurrence of the two phenomena. In the following sections, we discuss the development of the two classifiers and the results of the comparative study between the two phenomena.

Datasets

The HASOC dataset consists of approximately 8000 posts from Twitter and Facebook in each of the three languages—German, English, and Hindi. It is annotated at three levels. At the first level, the posts are annotated as 'hate and offensive' (HOF) vs 'non hate-offensive' (NOT). The 'hate and offensive' posts are further classified as 'hate speech' (HATE), 'offensive' (OFFN), and 'profane' (PRFN) at the second level. At the third level, the HOF posts are further classified as targeted insult (TIN) vs untargeted insult (UNT).

The TRAC-2 dataset consists of approximately 5000 comments from YouTube comments in the three languages—Hindi, Bangla, and English. The dataset is annotated at two levels—at the first level, the comments are annotated as overtly aggressive, covertly aggressive, and non-aggressive. At the second level, it is annotated for being gendered or not. Our study is based on the first two levels of annotations for English and Hindi posts from HASOC dataset and the first level of annotation for English, Hindi, and Bangla comments from the TRAC-2 dataset. The statistics about the two datasets, as used in our experiments, are given in Table 1[1].

Developing the Classifiers

We experimented with broadly two kinds of systems—an SVM classifier and different BERT-based classifiers including DistilBERT [42] and ALBERT [30] for both the tasks in

[1] Please note that in both these tasks train, dev/validation and test sets were provided separately. In this table, the figures for train refer to the total of train and dev datasets—in our experiments also, we use the two datasets together for training the system and use the test set for reporting the system performance.

Table 1 HASOC and TRAC-2 datasets

Language		HASOC					TRAC-2			
		TOTAL	NOT	HATE	OFFN	PRFN	TOTAL	NAG	CAG	OAG
Bangla	Train	NA	NA	NA	NA	NA	4783	2600	1116	1067
	Test	NA	NA	NA	NA	NA	1200	789	169	242
Hindi	Train	4665	2196	556	676	1237	4981	2823	1040	1118
	Test	1318	713	190	197	218	1200	316	215	669
English	Train	5852	3591	1143	451	667	5329	4211	570	548
	Test	1153	865	124	71	93	1200	690	224	286

all the languages.[2] We used the scikit-learn implementation of SVM ([40, 6]) and Fast-Bert library[3] (which itself is based on Hugging Face pytorch-transformers library[4]) for the experiments. The implementation details of the Fast-Bert are given in the two blogs by the author[5],[6].

Support vector machines [22] are one of the most successful classic machine learning models used for various kinds of text classification tasks. On the other hand, bidirectional encoder representations from transformers (BERT) [14] make use of a masked language model (MLM), which enables it to fuse the left and right context, thereby allowing to pre-train a deep bidirectional transformer. These pretrained models could be fine-tuned for specific tasks. BERT models are demonstrated to have given significant improvements in several NLP tasks including general language understanding, question answering, next sentence prediction/text generation, as well as some text classification tasks. DistilBERT makes use of the technique of knowledge distillation whereby a smaller model is trained to reproduce the behaviour of the larger model. Sanh et al. [42] trains a 40% smaller (in the sense that it has 40% less parameters in comparison to BERT) and 60% quicker transformer under the supervision of BERT. This new model retains 97% language understanding capabilities of the original BERT. ALBERT employs two parameter reduction techniques to reduce the parameters of the model. The first technique, called factorized embedding parameterisation, involves decomposing the large vocabulary embedding matrix into two smaller matrices, thereby separating the size of the hidden layers from that of vocabulary embedding. This allowed for training with a larger dataset (and higher vocabulary size) without significantly increasing the parameter size of the vocabulary embeddings. The second technique called cross-layer parameter sharing prevents the parameter from growing with the depth of the network. With the application of these techniques, Lan et al. [30] managed to train a BERT-large like configuration with 18× less parameters than BERT in 1.7× less time. The ALBERT models are shown to perform better than BERT in several downstream tasks.

Given the prior claims, we would expect ALBERT to give the best performance while some loss of performance with DistilBERT. We conducted experiments with these different models to explore the usefulness and efficacy of transformer models vis-a-vis SVMs and see if transformers could be helpful in the specific tasks of offensive and aggressive speech detection. We also explored if BERT, DistilBERT, and ALBERT have any significant (dis)advantages over each other for these tasks.

For both the classifiers, we used the train set for training and test set for testing the classifiers.

We also carried out a basic preprocessing—removing the links and anonymising the mentions—for all the experiments. We did not carry out any kind of normalisation, stemming, or lowercasing, since all of these might result in loss of information necessary for identification of offensive and aggressive language.

Experiments with SVM

For SVM, we used fivefold cross-validation for figuring out the optimum model. We experimented with the following sets of features:

1. Word *n*-grams (unigrams, bigrams, and trigrams).
2. Character *n*-grams (bigrams to 5 g).
3. A combination of different word *n*-grams and character *n*-grams features.

A grid search was performed for *C* values from 0.0001 up to 10 (with a 10× interval in between two *C* values) for each of the feature combination and each of the sub-tasks.

[2] In this section, we discuss the experiments related to the development of classifiers for sub-task A and sub-task B of HASOC shared task and sub-task A of the TRAC-2 shared task. For experiments related to sub-task C of HASOC and sub-task B of TRAC-2, you may refer to [33] and [5, 27].

[3] https://github.com/kaushaltrivedi/fast-bert.

[4] https://github.com/huggingface/pytorch-transformers.

[5] https://medium.com/huggingface/introducing-fastbert-a-simple-deep-learning-library-for-bert-models-89ff763ad384.

[6] https://medium.com/huggingface/multi-label-text-classification-using-bert-the-mighty-transformer-69714fa3fb3d.

Experiments with BERT and Other Transformers

We experimented with BERT, ALBERT, and DistilBERT in both sub-task A and sub-task B of HASOC dataset—offensive language identification and fine-grained classification of offensive language—and sub-task A of TRAC-2 dataset—aggression identification. We fine-tuned the pre-trained BERT-base-uncased, ALBERT-base-uncased, and DistilBERT-base-uncased models released by the respective research teams for the English dataset. We fine-tuned the BERT-multilingual-based-uncased and DistilBERT-base-uncased models for Hindi and Bangla datasets. Since ALBERT still does not provide a multilingual pre-trained model that could be used for languages like Hindi and Bangla, we did not experiment with ALBERT for these languages. The fine-tuning for all the models was carried out on a standard Google Colab GPU system. For all the sub-tasks, the models were trained for 10 epochs and used the LAMB optimizer. All the other hyperparameters were kept as recommended by the respective research teams for fine-tuning and were default settings in the fast-bert library that we used for fine-tuning.

We also experimented with the multilingual joint fine-tuning of the BERT and DistilBERT models. However, since the two languages—Hindi and English—are quite far apart from each other, all the models overfitted for the majority class in Hindi, resulting in rather trivial performance where all instances were labelled OAG/HOF. For English, there was no significant gain over the monolingual fine-tuning. As such, we are not reporting these results and we will not be discussed those any further in this paper.

Results and Discussion

The results of the feature ablation study for each sub-task in each language are given in Table 2. The details and discussion of these experiments with the HASOC dataset are discussed in [26]. We discuss these results and give a comparative overview in the following subsections.

HASOC Task A Results

For HASOC task A, BERT gives the best performance for English. For Hindi, and the SVM model with character 5-g and word unigram gives the best performance. However, a McNemar's test to acertain whether these features result in a significantly better performance vis-a-vis reveal that the overall performance is similar to those attained by a character trigram model. The test reveals that the character trigram model outperforms character bigram model significantly. However, beyond this neither word nor additional character *n*-grams prove to be helpful in the task. Moreover, word *n*-grams perform worse that all the character *n*-grams across the board and they do not provide any additional value as far as the system performance is concerned. In case of English, among the SVM models, however, it is the other way round, such that word *n*-grams prove to be significantly better than the character *n*-grams and character *n*-grams do not seem to give any additional information to the classifier. These results are on the expected lines considering the fact that morphological properties do prove to be distinguishing features in Hindi (captured by the character *n*-grams), while in English, lexical items differ in offensive and not offensive texts.

HASOC Task B Results

In HASOC task B, since the dataset was pretty small in size, the BERT-based models were completely outperformed by the SVM classifiers. These results are consistent with the results reported by other scholars in low-data conditions for similar tasks [36, 44, 46]. From among the SVM models, those trained with character 5-g for English and character 4 g for Hindi prove to be the best models. Adding word unigram to the English model and word bigram to the Hindi model slightly improves precision and recall; however, the McNemar's test reveal that these gains are not statistically significant. Unlike in Task A, higher order character n-grams prove to be useful for English as well in this case.

TRAC Results

As with HASOC task B dataset, SVMs outperform all the BERT-based systems with the TRAC dataset. For English, character bigram combined with word unigram gives a perfect score for the task. For Hindi and Bangla, considering the fact that it is a three-class classification task, the performance of the SVMs is surprisingly good. For both the language, character trigram combined with the word unigram gives the best performance. For both the languages, character trigrams and word unigrams individually perform significantly worse than when the features are combined together. This trend is observed in other combinations of character and word *n*-gram features where combining the two kinds of features leads to a significantly better performance than the individual features.

Comparison Across Tasks and Techniques

In general, SVM classifiers have performed better than the neural-network-based classifiers like the fine-tuned BERT, DistilBERT, and ALBERT for both the Indian languages—Hindi and Bangla—in all the sub-tasks in both HASOC and TRAC datasets. In case of English, BERT-based classifier outperforms the SVM classifiers in only sub-task A in HASOC. However, in the TRAC-2 dataset, SVM classifiers

Table 2 Overall experiment results

Algo		Lang	HASOC						TRAC-2		
			Sub-task A			Sub-task B			Sub-task A		
			Precision	Recall	*F*-score	Precision	Recall	*F*-score	Precision	Recall	*F*-score
SVM	C2	BEN	–	–	–	–	–	–	0.79	0.79	0.76
		HIN	0.76	0.76	0.76	0.56	0.58	0.55	0.79	0.77	0.77
		ENG	0.74	0.71	0.72	0.60	0.58	0.52	0.86	0.56	0.86
	C3	BEN	–	–	–	–	–	–	0.84	0.85	0.83
		HIN	0.79	0.79	0.79	0.61	0.62	0.60	0.84	0.82	0.82
		ENG	0.74	0.69	0.71	0.68	0.62	0.56	0.91	0.91	0.91
	C4	BEN	–	–	–	–	–	–	0.87	0.87	0.86
		HIN	0.80	0.79	0.79	0.62	0.63	0.62	0.86	0.85	0.85
		ENG	0.75	0.69	0.71	0.65	0.65	0.59	0.89	0.89	0.89
	C5	BEN	–	–	–	–	–	–	0.87	0.87	0.86
		HIN	0.79	079	0.79	0.63	0.63	0.62	0.87	0.86	0.86
		ENG	0.75	0.68	0.70	0.65	0.65	0.61	0.88	0.88	0.88
	W1	BEN	–	–	–	–	–	–	0.79	0.80	0.78
		HIN	0.76	0.74	0.74	0.48	0.44	0.37	0.80	0.77	0.77
		ENG	0.76	0.73	0.74	0.39	0.51	0.41	0.79	0.77	0.75
	W2	BEN	–	–	–	–	–	–	0.80	0.80	0.78
		HIN	0.74	0.73	0.73	0.49	0.45	0.38	0.79	0.76	0.76
		ENG	0.75	0.74	0.74	0.38	0.51	0.41	0.80	0.78	0.76
	W3	BEN	–	–	–	–	–	–	0.80	0.80	0.78
		HIN	0.73	0.72	0.72	0.49	0.45	0.38	0.78	0.74	0.75
		ENG	**0.75**	**0.74**	**0.75**	0.38	0.50	0.41	0.79	0.77	0.75
	C2+W1	BEN	–	–	–	–	–	–	0.88	0.89	0.89
		HIN	0.74	0.74	0.74	0.58	0.59	0.57	0.84	0.83	0.83
		ENG	0.76	0.71	0.72	0.66	0.60	0.54	**1.0**	**1.0**	**1.0**
	C2+W2	BEN	–	–	–	–	–	–	0.89	0.89	0.89
		HIN	0.76	0.75	0.75	0.58	0.59	0.57	0.84	0.83	0.83
		ENG	0.75	0.71	0.72	0.66	0.60	0.54	0.95	0.95	0.95
	C2+W3	BEN	–	–	–	–	–	–	0.88	0.89	0.88
		HIN	0.76	0.76	0.76	0.58	0.59	0.57	0.87	0.87	0.87
		ENG	0.76	0.72	0.73	0.66	0.60	0.54	0.93	0.93	0.93
	C3+W1	BEN	–	–	–	–	–	–	**0.93**	**0.93**	**0.93**
		HIN	0.79	0.79	0.79	0.61	0.62	0.60	**0.88**	**0.87**	**0.87**
		ENG	0.74	0.68	0.69	0.66	0.63	0.57	0.91	0.91	0.91
	C3+W2	BEN	–	–	–	–	–	–	0.92	0.92	0.92
		HIN	0.79	0.79	0.79	0.61	0.62	0.60	0.85	0.83	0.83
		ENG	0.74	0.68	0.69	0.63	0.62	0.56	0.91	0.91	0.91
	C3+W3	BEN	–	–	–	–	–	–	0.93	0.92	0.92
		HIN	0.79	0.79	0.79	0.61	0.62	0.60	0.85	0.84	0.84
		ENG	0.74	0.68	0.70	0.63	0.62	0.56	0.90	0.91	0.90
	C4+W1	BEN	–	–	–	–	–	–	0.93	0.93	0.93
		HIN	0.80	0.79	0.79	0.63	0.64	0.62	0.85	0.84	0.84
		ENG	0.74	0.68	0.70	0.64	0.64	0.59	0.89	0.89	0.89
	C4+W2	BEN	–	–	–	–	–	–	0.93	0.93	0.93
		HIN	0.79	0.79	0.79	**0.63**	**0.64**	**0.63**	0.87	0.86	0.86
		ENG	0.75	0.68	0.70	0.64	0.64	0.59	0.89	0.89	0.89

Table 2 (continued)

Algo		Lang	HASOC						TRAC-2		
			Sub-task A			Sub-task B			Sub-task A		
			Precision	Recall	*F*-score	Precision	Recall	*F*-score	Precision	Recall	*F*-score
	C4+W3	BEN	–	–	–	–	–	–	0.93	0.93	0.93
		HIN	0.80	0.79	0.79	0.63	0.64	0.62	0.87	0.86	0.86
		ENG	0.75	0.68	0.70	0.63	0.64	0.59	0.90	0.90	0.90
	C5+W1	BEN	–	–	–	–	–	–	0.92	0.92	0.92
		HIN	0.80	0.79	0.79	0.63	0.63	0.62	0.88	0.87	0.87
		ENG	0.74	0.68	0.70	**0.66**	**0.65**	**0.61**	0.88	0.88	0.88
	C5+W2	BEN	–	–	–	–	–	–	0.93	0.93	0.93
		HIN	**0.80**	**0.80**	**0.80**	0.62	0.63	0.62	0.88	0.87	0.87
		ENG	0.75	0.68	0.70	0.65	0.65	0.60	0.88	0.88	0.88
	C5+W3	BEN	–	–	–	–	–	–	0.93	0.93	0.93
		HIN	0.80	0.80	0.80	0.63	0.63	0.62	0.88	0.87	0.87
		ENG	0.75	0.68	0.70	0.65	0.65	0.50	0.88	0.88	0.88
BERT	ENGM	ENG	**0.81**	**0.79**	**0.80**	0.44	0.58	0.49	0.69	0.70	0.67
	MULT	BEN	–	–	–	–	–	–	**0.73**	**0.75**	**0.73**
		HIN	0.42	0.42	0.38	0.21	0.34	0.21	**0.74**	**0.69**	**0.71**
		ENG	0.80	0.69	0.71	0.34	0.35	0.30	**0.73**	**0.75**	**0.73**
DistilBERT	ENGM	ENG	0.65	0.34	0.30	0.25	0.32	0.17	0.65	0.68	0.63
	MULT	BEN	–	–	–	–	–	–	0.73	0.75	0.73
		HIN	**0.73**	**0.73**	**0.73**	**0.46**	**0.48**	**0.40**	0.67	0.65	0.65
		ENG	0.81	0.72	0.74	**0.40**	**0.52**	**0.45**	0.67	0.69	0.64
ALBERT	ENGM	ENG	0.75	0.71	0.72	0.42	0.47	0.44	0.59	0.62	0.59

The values in bold show the best performance of the system with that specific classification technique

C2 character bigram, *C3* character bi and trigrams, *C4* character bi, tri and 4 g, *C5* character bi to 5 g, *W1* word unigram, *W2* word uni and bigrams, *W3* word uni, bi and trigrams, *ENGM* pre-trained English model, *MULT* pre-trained multilingual model, *BEN* Bangla dataset, *HIN* Hindi dataset, *ENG* English dataset

perform significantly better than all of the neural-network classifiers. In sub-task B of HASOC, we see that SVM outperforms all the neural-network-based classifiers by a huge margin. The best SVM model for English—using character bi to 5 g and word unigram—gets an overall *F*-score of 0.61[7], while the best neural-network classifier—multilingual model of DistilBERT—manages an *F*-score of just around 0.45, with recall of 0.52. The difference for Hindi is even larger, with SVM and multilingual DistilBERT getting best scores of 0.63 and 0.40, respectively.

Among the neural-network classifiers, even though it is expected that English-only models would work better than multilingual models (for English dataset), we get mixed results. For HASOC dataset, the overall performance (and more specifically 'recall' scores) of English-only models of BERT is marginally better than multilingual ones; however, in case of DistilBERT, multilingual models outperform the English-only models by a huge margin. In fact, multilingual DistilBERT models, as against our expectations, in most of the cases, either stand at par with the multilingual BERT models or significantly outperform those, especially in the two sub-tasks of HASOC.

Let us now take a closer look at HASOC sub-task B, in which the performance for all the classifiers remain low, in comparison to both sub-task A as well as TRAC-2 sub-task A. Given the fact that it was a three-class classification problem, the scores do not look very promising for any of the classifiers in the HASOC Task B (compare this with TRAC-2 performances, which was also a three-class classification task). This could be attributed to the tiny amount of training data. If we look at the class-wise precision and recall (Table 3), then we see that the neural-network-based models completely fail to learn the OFFN class. For HATE, especially for English, we see a reasonably good recall but

[7] The precision, recall and *F*-score reported in this paper are calculated using the scitkit-learn classification_report function, which gives an averaged *F*-score weighted by the total 'support' for each class—this, as expected, has resulted in some *F*-scores that do not lie in between the precision and recall values.

Table 3 Class-wise scores of the best models

Model	Lang		HASOC							TRAC-2			
			Sub-task A			Sub-task B				Sub-task A			
			Overall	HOF	NOT	Overall	HATE	OFFN	PRFN	Overall	OAG	CAG	NAG
SVM	BEN	Precision	–	–	–	–	–	–	–	0.88	0.84	0.79	0.92
		Recall	–	–	–	–	–	–	–	0.89	0.80	0.74	0.95
		F-score	–	–	–	–	–	–	–	0.89	0.82	0.76	0.93
	HIN	Precision	0.80	0.76	0.83	0.63	0.62	0.61	0.67	0.88	0.93	0.89	0.77
		Recall	0.80	0.81	0.78	0.64	0.49	0.48	0.91	0.87	0.90	0.64	0.95
		F-score	0.80	0.78	0.80	0.63	0.55	0.54	0.77	0.87	0.92	0.75	0.85
	ENG	Precision	0.75	0.48	0.84	0.66	0.66	0.67	0.65	1.0	1.0	1.0	1.0
		Recall	0.74	0.55	0.80	0.65	0.85	0.17	0.76	1.0	0.99	1.0	1.0
		F-score	0.52	0.82	0.75	0.61	0.74	0.27	0.70	1.0	0.99	1.0	1.0
BERT	BEN	Precision	–	–	–	–	–	–	–	0.73	0.76	0.50	0.79
		Recall	–	–	–	–	–	–	–	0.75	0.63	0.37	0.90
		F-score	–	–	–	–	–	–	–	0.73	0.69	0.43	0.84
	HIN	Precision	0.42	0.42	0.42	0.21	0.26	0.00	0.35	0.74	0.89	0.41	0.61
		Recall	0.42	0.67	0.20	0.34	0.06	0.00	0.90	0.69	0.71	0.55	0.76
		F-score	0.38	0.51	0.27	0.21	0.10	0.00	0.50	0.71	0.79	0.47	0.68
	ENG	Precision	0.80	0.44	0.92	0.44	0.57	0.00	0.60	0.73	0.73	0.49	0.81
		Recall	0.69	0.84	0.65	0.58	0.90	0.00	0.60	0.75	0.64	0.37	0.92
		F-score	0.71	0.58	0.76	0.49	0.70	0.00	0.60	0.73	0.68	0.42	0.86
DistilBERT	BEN	Precision	–	–	–	–	–	–	–	0.73	0.72	0.57	0.79
		Recall	–	–	–	–	–	–	–	0.75	0.62	0.36	0.92
		F-score	–	–	–	–	–	–	–	0.73	0.67	0.44	0.85
	HIN	Precision	0.73	0.72	0.74	0.46	0.40	0.41	0.56	0.67	0.81	0.40	0.55
		Recall	0.73	0.68	0.78	0.48	0.01	0.66	0.72	0.65	0.67	0.34	0.80
		F-score	0.73	0.70	0.76	0.40	0.02	0.51	0.63	0.65	0.73	0.65	0.36
	ENG	Precision	0.81	0.46	0.92	0.40	0.60	0.00	0.44	0.67	0.86	0.36	0.69
		Recall	0.72	0.83	0.68	0.52	0.70	0.00	0.68	0.69	0.50	0.10	0.97
		F-score	0.74	0.59	0.78	0.45	0.65	0.00	0.53	0.64	0.63	0.15	0.81
ALBERT	ENG	Precision	0.75	0.44	0.86	0.42	0.53	0.17	0.46	0.59	0.54	0.35	0.70
		Recall	0.71	0.64	0.73	0.47	0.65	0.07	0.54	0.62	0.24	0.32	0.88
		F-score	0.72	0.52	0.79	0.44	0.58	0.10	0.50	0.59	0.33	0.34	0.78

low precision in case of BERT and both low recall and low precision in case of other two models. For Hindi, we see similar performance for PRFN. In case of SVMs, the classifier for English gets a really bad score for OFFN, but still manages to perform better than all the neural-network-based classifiers. The overall results for Hindi are similar, but we see a much improved performance for OFFN, but the recall for HATE and OFFN are still less than half of that for PRFN. This shows that there were sufficient features for learning PRFN (as well as, to certain extent, HATE) in this small dataset, but for OFFN, none of the classifiers could generalise well.

In case of TRAC-2 dataset, we see an astounding performance for English (a perfect score) and pretty high scores for Bangla and Hindi. However, if we look closer at class-wise scores (Table 3), we see that all the classifiers perform worst in recognising instances of CAG and score comparatively low recall. This is on expected lines, since, by definition, CAGs will not provide many surface-level features and so difficult to predict using such features. However, we had hoped for a better performance with neural-network-based classifiers, which are known to be better at discovering the ’covert’ features.

Given these, the following remarks could be made about SVM classifiers, transformer-based classifiers, as well as the nature of these tasks:

– Even though the transformer-based models like BERT and ALBERT have created a lot of hype within the academic circles (as well as outside it), given the computational resources required in training these huge models and lack of dataset for most of the world's approximately 7000 languages, they have some serious practical limitations. The fact that multilingual models performed better than the English models (in most of the cases) shows the sensitivity of these models to variation in language. The TRAC-2 dataset and a significant proportion of the HASOC dataset consisted of Indian English (and not the usual British or American English), collected from social media (while these models are trained on a much cleaner Wikipedia dataset). Therefore, despite 'English' model being the 'recommended' one for training English data, the multilingual model may actually work better with non-Wikipedia-like English, as is found in most of the outer and expanding circle countries with English-speaking population ([24, 25]). Similarly, remarks could also be made for the social media data. Therefore, while these pre-trained models may perform well for a large number of tasks (because of the huge amount of data being used for training), they may still be quite fragile in 'out-of-domain' situations.

- Even though it has been claimed that the pre-trained models could be fine-tuned efficiently with relatively smaller datasets, it may still require more data than classifiers like SVM (known to perform especially well in low-data situations) to perform at par with those. We see that it was only in HASOC sub-task A in English (which is a combination of relatively bigger dataset as well as data in a relatively 'standard' form) that BERT performs better than SVM.
- In HASOC sub-task B, we see a general inability of the classifier to predict OFFN class; rather, it is generally assigned PRFN. Similarly, a lot of instances of HATE has been marked as PRFN. We discuss more about this in "Error Analysis". These results are also somewhat related to the amount of data available for training—in Hindi dataset, we have comparatively larger samples for PRFN, while in English, it is for HATE. OFFN is underrepresented in both the languages. However, in general, these performances may not be completely attributable to the amount of dataset available, since PRFN and HATE have relatively similar performance in both the languages, which is significantly higher than that of OFFN and may show a possible issue with the tagset where PRFN may be a more overarching category and not directly distinguishable from OFFN and HATE; PRFN could and often does co-occur with OFFN and HATE and may also occur independent of each other. As such it might be a better idea to identify PRFN separately from HATE and OFFN and that may lead to an improvement in the overall performance.
- For TRAC-2, it may be the case that the test set is pretty similar to the train and dev set (since all the datasets were drawn from YouTube comments) and so the performance was much better for certain classifiers. However, the consistently low scores for neural-network-based classifiers, in comparison to the SVM classifiers may be indicative of one of the two things—either SVMs overfitted (while NNs tried to learn and generalise) or the NN models just could not generalise well enough with a relatively small, 'non-canonical' dataset. Since the dataset was not huge, it is also not very likely that SVMs would overfit, especially in a three-class classification task. However, this could only be confirmed by looking at the loss of these classifiers in case of an out-of-domain test set and is subject to future investigation.

Error Analysis

In the results of the offensive and aggressive language identification systems, there were various errors. The majority of the errors correspond to the following:

1. Errors in marking HATE, OFFN, and PRFN.
2. Errors in identifying NAG, CAG, and OAG.

In this section, we present the underlying causes of the errors in detail.

1. Ungrammatical/nonsensical sentence: When the sentences or the phrases are ungrammatical and hence made no sense then the errors are created, in all the three languages. For example,
   i. bhikA.Tima
   ii. khetapur mere dam.
   iii. Arlene is not logical, nor is the DUP, if they think you are diluting the union, it is curtains.

   There is no meaning of (i), (ii), and (iii). As (i) and (ii) are nonsensical sentences, they cannot be aggressive, but the system recognised those as OAG which is wrong. Similarly, in (iii), the system marked it as HATE which is not correct.
2. Insufficient information: When there are small phrases which do not carry sufficient information, then the tagging was wrongly done by the system. It is again common for all the three languages:
   i. #Repost free.wicked • • • • • #freewicked #freethekids #terrorist #Americanterrorist #fuck-trump #donaldtrump #notothewall #nowall #ak47 #shooting #justiceforthekids #nomoregunviolence #gunviolence ......
   ii. dongi...faltu mohila...... (Bangla)
   (characterless woman)
3. Use of unconventional spelling: Generally, a variety of spellings can be found to be used in the comments. It

is more so when Bangla or Hindi is written in Roman script instead of their canonical scripts, Bangla and Devanagari, respectively. However, writing Bangla and Hindi in Roman script also follows a convention. When this convention is not maintained and the spelling looks non-canonical, then the system failed to correctly tag the data. The use of unconventional spellings created errors. For example:

"oree juta diye marr ree vaii...." (beat her with a shoe).

In this example, the word "vaii" (brother) is a non-canonical spelling. The canonical spelling is "bhai". This is marked as NAG by the system. However, it should be marked as CAG.

In the following sentence again, "vhii" (bhai) is used for brother "dada vhii mood off 6ilo ....tmi puro chill kore dile" (brother, I was sad, but you changed my mood). This sentence was marked CAG; however, the correct tag is NAG. When the canonical script was used but with misspelled words, then also the system failed to tag the sentence correctly. For example:

"jo shabda tUma Aja kisI aura aurata ke lie yUja kara rahe vo bacAkara rakhanA........ ye salAya nahIM saccAI hai" (the words you are using for other women today, save those...this is not a suggestion but a truth).

In the following sentence of Hindi written in Devanagari script, one word "salAya" (suggestion) is misspelled, so the system had an error in tagging it.

4. Figurative use of language: The comments used figurative use of language with metaphors, satires, or proverbs. In such cases, the system failed to identify the correct tag. For example:

   "pAẏera juto pAẏe rAkhA bhAlo." (shoes should be kept in one's feet)

   The pragmatics of the sentence is very different from its literal meaning. The writer of the sentence is referring to someone as a shoe. Here, shoe means a person of very low class. The sentence means that a lower class person should be kept at the lower level of society. This sentence should have been marked as CAG. However it was marked as NAG. In most of such sentences where metaphorical or proverbial language was used the system tagged it as NAG perhaps, because the literal meaning was never aggressive. Another similar example can be seen in the following sentence. "r aei mohila neighbour agune ghee chetate eseche", (and this lady neighbour is there to put oil/ghee in the fire); the sentence means that the lady neighbour is there to intentionally increase the problem as one increases the fire by adding oil to it. This sentence too should have been marked as CAG, but it was marked as NAG by the system. Similarly in the following example, the sentence refers to the popular personalities who know Hindi well but satirically points at them as people without the knowledge of Hindi, e.g.,

   "saima pitrodA kI hindI KarAba hai maNi shaMkara kI hindI KarAba hai shashi tharUra kI hindI KarAba hai adhIra raMjana caudharI ki hindI KarAba hai yahA.n taka kI rAhula gAMdhI bhI hindI ke kaThina shabdoM ko mobAila meM khojateM haiM AnaMda sharmA ke mutAbiqa . para pIema modI gAlI deneM vAle tevara KarAba nahIM hue ." (Som Pitroda's Hindi is bad, Mani Shankar's Hindi is bad, Adhir Ranjan's Hindi is bad, even Rahul Gandhi looks for the meanings of the difficult Hindi words in his mobile, according to Anand Sharma. But PM Modi's abusive speech is not bad).
5. Lack of world knowledge: Every discourse is context bound. These contexts at times refer to world knowledge which may include some real-life events. If someone is ignorant of the world knowledge, then he/she may not be able to fully understand the discourse. In this case too, similar thing can be witnessed. The discourse which referred to world knowledge was wrongly tagged by the system. For example:

   "ei rakama mahilAra eka samaẏ pAna pAtAra paẏsA juTabonA", (this lady will not be able to afford betel leaf).

   Betel leaf is a very cheap commodity in this part of the world and which is commonly used by the Bengali ladies. The sentence basically means that the lady will become so poor that she will not be able to afford even a betel leaf. It should have been marked as CAG according to the native speakers of Bangla; however, it was marked as OAG by the system. Similarly, in the following sentence, though no one has been mentioned but by the context it is known that the sentence refers to a lady beggar who became a singer. There was aggression in the public against this lady as from a poor beggar she became a rich singer. Therefore, someone writes, "thik sobsmy otit k mathy rekhe cholte hoi" (one should always remember one's past). The sentence looks non-aggressive without the context; however, in the given context, it looks covertly aggressive. The system tagged it as NAG.
6. Selected words: It was noticed that in some of the sentences in English and Hindi though not in Bangla whenever some set of words occurred in the sentence, the system marked it as HATE or HOF. These set of words are Muslim, Mulla, Khuda, and other words related to Muslim community and religion. For example:

   "bahuta pyAra karate haiM tumase sanama kasama cAhe le lo khudA kI kasama" (I love you a lot, God knows).

   This is a line from a popular Hindi song which should have been marked as NONE, but it was marked as HOF by the system as the sentence has a word "Khuda" (Islamic God). Similarly, in the following sentences, "Kolkata Muslims write to Mamata Banerjee: Do not let culprits who attacked doctors and Ushoshie get

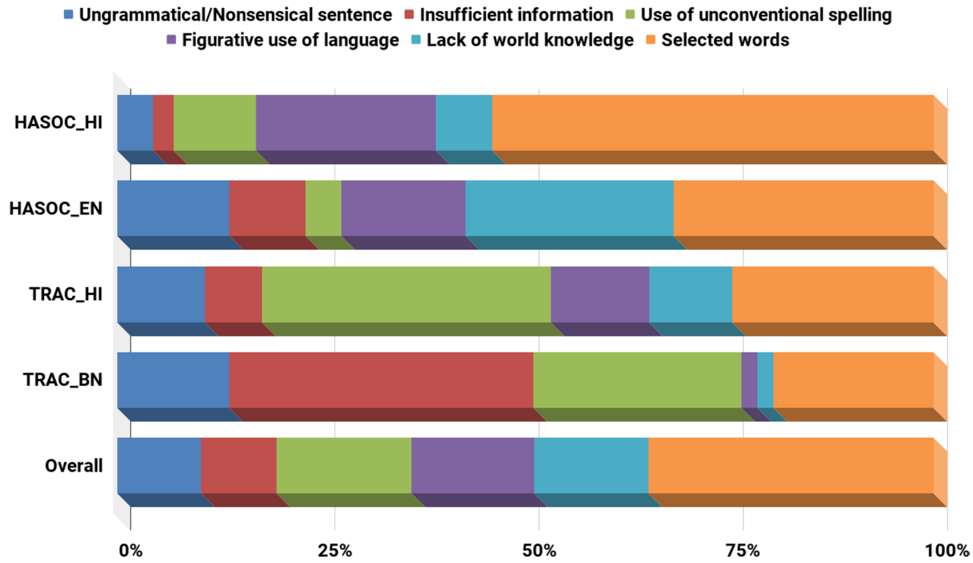

Fig. 1 Proportion of errors across tasks and languages

away just, because they are Muslims", word "Muslims" appears, and the sentence is marked as HATE rather than being marked as NONE.

Comparison of Errors

To understand the nature of errors made by different classifiers, we classified each of the errors in HASOC sub-task A and TRAC. In TRAC, the results of English are not included in this analysis, since there were just a couple of wrong predictions for this. We also do not include HASOC sub-task B because of the extremely low amount of data and a rather poor performance of the classifiers because of underfitting, which would render this kind of analysis rather biased and trivial.

Figure 1 gives a comparison of the proportion of each error kind in English and Hindi classifiers for HASOC sub-task A and Hindi and Bangla for TRAC. It shows that all the classifiers overgeneralises for certain lexical items across all languages. It is most prominent in case of HASOC Hindi dataset than in the other datasets. In addition to this, we see that more than 25% of errors in Bangla is because of lack of sufficient information—this is obvious given the fact that the average length of comments in Bangla is much smaller than those of Hindi and English (see Table 5) and as such the classifiers would not get sufficient discriminating features in a large number of cases. other than this, there are quite a large number of errors because of the use of unconventional spelling—this is also explainable by the fact that the TRAC dataset consists of YouTube comments, which probably make use of more unconventional spelling than Twitter or Facebook (source of HASOC dataset). Besides these, other sources of errors such as use of ungrammatical/nonsensical sentences are comparable across languages and datasets.

Merging and Preparing the Final Dataset

Using the classifiers discussed in the previous sections, we annotated the HASOC dataset with the TRAC-2 labels (OAG, CAG, and NAG) and the TRAC-2 dataset with HASOC labels (HOF, NOT, HATE, PRFN, and OFFN). This yielded a dataset of approximately 10,000 instances annotated with both aggression as well as different kinds of offensive language. Since HASOC did not have Bangla dataset, we could not automatically annotate the Bangla dataset with both the labels. However, for this study, we manually annotated the TRAC test dataset with the offensive labels and used for analysis. The manual annotation of a little over 1000 Bangla comments was carried out by two annotators and, in case of disagreement, a third annotation was taken and the majority tag was used for the study. Since it was a small sample that we were getting annotated by the same annotators who worked with the TRAC-2 dataset, we did not carry out IAA experiments separately for these annotations; however, we assume that it would be similar to that for the annotation of the TRAC-2 dataset. Table 4 shows the complete dataset in all the languages used for this study.

Aggressive and Offensive Language

Most of the studies on offensive and aggressive language [and other similar kinds of phenomena, viz., abusive language, toxicity, hate(ful) language, etc.] within the field of

Table 4 Final dataset

Language	Total	NOT	HATE	PRFN	OFFN	OAG	CAG	NAG
Bangla	1183	779	37	175	181	240	162	781
Hindi	20,023	12,799	1716	3525	1983	4896	4548	10,579
English	14,687	8851	2381	2851	604	1795	1086	11,806

NLP have focussed on one aspect of these related but differently-labelled phenomena and have either assumed them to be independent of each other or synonymous with each other[8]. The extreme interest in the field has resulted in a spurt of a large number of studies around these phenomena as well as development of large number of publicly available datasets. However, since these datasets are annotated with different labels and we have little understanding of how they are inter-related, it has made the use of dataset annotated with one kind of label almost impossible with another task. At this point, it need not be stressed too much that an understanding of inter-relationship among these phenomena might result in better utilisation of resources as well as it may prove to be helpful in arriving at better systems that could deal with these different hues of objectionable language.

On the other side of the things, these phenomena and their pragmatic structure have been greatly debated and theorised within the field of sociopragmatics and interactional sociolinguistics. Pragmaticians have tried to understand the micro as well as macro-level distinctions, similarities, and overlaps in between such related phenomena as impoliteness, aggression, rudeness, insulting language, etc. In the following subsections, we will discuss some of the insights from these studies, and also present the results and analysis of our comparative study between aggressive and offensive language, based on the dataset that we prepared in “Datasets”.

Statistical Overview of Aggressive and Offensive Language

Table 5 gives a statistical comparison of the different kinds of aggressive and offensive language across the three languages. While the figures about the total utterances[9] or total words may not be directly comparable in this because of cross-linguistic as well as cross-category variation in the absolute numbers, looking at the mean utterance per comment, mean token per comment, mean token per utterance, and mean character per comment yielded some useful insights and cross-linguistic generalisations into the aggressive and offensive language usage on social media.

If we look at HOF and NOT, then across all the three languages, NOT has more utterances per comment; however, HOF is longer in terms of mean number of tokens in each comment. In case of the three aggressive language categories, except Bangla, NAG comments/tweets are shortest in length, while CAG are the longest. In Bangla, CAG are still the longest comments, while NAG and OAG are of almost equal lengths. In terms of mean utterances per comment, CAG again has the maximum number of utterances, while NAG has the least. In this case as well, Bangla seems to be an outlier as OAG has less number of utterances than NAG. Among the three sub-categories of HOF, OFFN has least number of utterances per comment, while HATE has the highest across the three languages. In terms of number of tokens per comments, PRFN is shortest, while comments marked as HATE are the longest ones. In Bangla, this pattern is slightly different—however, given the extremely low number of data points in Bangla, these results are not reliable for Bangla. From amongst the two categories of aggressive and offensive language, there is no consistent pattern that could be observed across languages. This shows that the sub-categories across the two phenomena share at least some features and they overlap in significant number of cases. This aspect is explored further in the following subsections.

Pragmatics of Offense and Offensive Languages

In general, ‘offense’ has been considered as a result of the ‘impolite’ language [10]—“Impoliteness is a negative attitude towards specific behaviours occurring in specific contexts. It is sustained by expectations, desires, and/or beliefs about social organisation, including, in particular, how one person’s or a group’s identities are mediated by others in interaction. Situated behaviours are viewed negatively—considered ‘impolite’—when they conflict with how one expects them to be, how one wants them to be, and/or how one thinks they ought to be. Such behaviours always have or are presumed to have emotional consequences for at least one participant; that is, they cause or are presumed to cause

[8] I could not find any previous study that directly compares or studies the inter-relation between these phenomena and the assumption about their independence or being synonymous is largely implicit in the silence of the most of the researchers working in these areas. However, I have discussed some notable exceptions in “ Introduction”.

[9] Utterances are generally considered to be present only in spoken speech. However, considering social media comments to be a close approximation of speech, it would be more logical to divide these comments into utterances than sentences. In speech, utterances are characterised by short pauses while speaking. We considered the presence of single or multiple sentence-terminating punctuation, viz., full stop, exclamation mark, and question mark, as indicative of one utterance. Thus, these are used for calculating the numbers about punctuation.

Table 5 Aggressive and offensive language statistics

Metric	Language	NOT	HOF	HATE	PRFN	OFFN	OAG	CAG	NAG
Total utterances	Bangla	1406	696	58	251	368	349	399	1363
	Hindi	35,328	17,798	4956	7440	5402	12,953	14,490	25,683
	English	25,048	20,293	9281	8998	2014	7165	5232	32,944
Total tokens	Bangla	5681	3181	326	754	2034	1482	2522	4911
	Hindi	2,86,807	1,83,445	55,856	76,234	51,265	1,14,209	1,58,398	1,97,645
	English	1,52,810	1,62,901	82,690	66,063	14,148	51,451	52,204	2,12,056
Unique tokens	Bangla	2444	1828	248	514	1267	956	1491	2117
	Hindi	31,531	18,423	8635	9546	7799	15,771	18,589	23,937
	English	23,174	22,033	13,972	11,681	4128	10,201	8821	28,954
Mean utterances	Bangla	1.80	1.72	1.57	1.43	2.03	1.45	2.43	1.74
	Hindi	2.76	2.46	2.90	2.11	2.72	2.65	3.19	2.43
	English	2.83	3.48	3.90	3.16	3.33	3.99	4.81	2.79
Mean token per utterance	Bangla	4.04	4.57	5.62	3.00	5.27	4.25	6.32	3.60
	Hindi	8.12	10.31	11.27	10.26	9.49	8.82	10.93	7.70
	English	6.10	8.03	8.91	7.34	7.02	7.18	9.98	6.44
Mean token per comment	Bangla	7.20	7.77	8.78	4.25	11.08	6.12	15.15	6.18
	Hindi	22.27	25.45	32.60	21.72	25.88	23.26	34.76	18.61
	English	16.92	27.63	34.47	22.90	23.05	28.34	47.89	17.63
Mean character per comment	Bangla	40.91	42.55	48.97	22.93	60.88	33.37	83.77	35.29
	Hindi	116.32	123.99	161.33	104.75	125.86	117.39	171.81	97.20
	English	101.26	156.32	196.22	127.84	133.47	162.54	268.83	103.75

offence.” Culpeper [10] goes on to define a list of factors that define the degree of offense as well as the quality of offense. Some of these factors include:

- Attitudinal factors: These factors decide the extent to which the hearers’ expectations, desires, etc. are infringed.
- Linguistic–pragmatic factors: These include the degree of offense that is conventionally associated with the linguistic formula used in the impolite expression.
- Contextual and co-textual factors: These include the extent to which the behaviour is positively or negatively valued in a specific culture, whether the behaviour is in-group or out-group, the degree of intentionality ascribed to the text, the perspective of the person taking offense, and other factors.

In general, it has been argued that the symptom for offensive language lies both in the speakers’ utterances as well as the negative emotional reaction of the person taking the offense. In his formulation, aggressive is one of the impoliteness-related labels, along with impolite, rude, ill-mannered, etc., which may be applied to refer to and also partly shape the impoliteness attitude. Thus, aggression is one of the sub-types of impoliteness, while offense is the emotional reaction of the hearer towards impoliteness. Tedeschi and Felson [47] posit that the notion of ‘social harm’ is central to aggression, which is defined as below:

“Social harm involves damage to the social identity of target persons and a lowering of their power or status. Social harm may be imposed by insults, reproaches, sarcasm, and various types of impolite behaviour” [47].

Culpeper [10] also carries out an analysis of the terms ‘impolite’, ‘offensive’, ‘aggressive’, ‘abusive’, and ‘rude’ based on their synonyms in seven different thesauri of English. The analysis shows “that these items represent a relatively cohesive set, although the connection with (verbally) aggressive is somewhat weaker; that rude and impolite overlap a great deal; and that (verbally) abusive and offensive relate to effects, the former having more to do with the effects on one’s reputation and the latter being a general term for any extremely unpleasant effects.”. He also notes that all the synonyms of ‘aggressive’ “[...] have a sense of forcefulness and in some cases violence”. In case of ‘impolite’, there is a greater emphasis on negating with elements associated with ’polite’ (such as ‘dis-courteous’, ‘dis-respectful’, etc., besides sharing a great deal semantic trend with ‘rude’ (which has an additional sense of roughness, shown by synonyms like ‘crude’, ‘rough’, etc.). These analyses, thus, establish a clear distinction between ‘aggression’ (which is generally related to forcefulness and also violence) and offense (which is an ‘effect’ of rudeness/impoliteness/any general unpleasant effect).

Usage of Offensive, Aggressive, and More

Inspired by Culpeper [10], we conducted a similarity study of nine different related lexical, viz., ‘offensive’, ‘aggressive’, ‘violent’, ‘threatening’, ‘hateful’, ‘abusive’, ‘rude’, ‘impolite’, and ‘toxic’. The similarity study was done using the Common Crawl (840 B tokens, 2.2 M vocab, cased, 300 d vector) model of Global Vectors for Word Representation GLoVe [41]. This study was markedly different from that of [10] in the fact that Culpeper made use of Thesaurus entries for understanding the similarities between different lexical items, which did not represent the actual usage of these lexical items by speakers; rather, they represent a second-order, theoretical, meta-representation of the similarities by a specific group of speakers. However, since GLoVe vectors are based on a humongous corpora of actual language usage by the speakers, the analysis presented here could be taken as a good approximation to the understanding and usage of these lexical items by the speakers.

A visualisation of the similarity between these lexical items is shown in Fig. 2[10]. The analysis showed some results which were similar to that presented by Culpeper [10] such as the close relationship of ’aggression’ with ‘violence’ and ‘threat’ and the closeness of ‘impolite’ and ‘rude’. However, unlike the results presented by Culpeper [10], offensive and impolite seem to be quite distinct from each other. Similarly, ‘toxic’, ‘aggressive’, ‘offensive’, and ‘impolite’ lie at different extremes and are semantically quite distinct from each other. Since ‘offense’ is a result of ‘impoliteness’ (which is used as a meta-term for different phenomena like aggression, rudeness, etc.), it is expected that semantically they are distinct and ‘offense’ almost seems to be an outlier in the group and almost equidistant from words like ‘impolite’, ‘rude’, or ‘abusive’. The study shows that ‘aggressive’, ‘offensive’, and ‘toxic’ are semantically quite distinct from each other, and as such, it may not be the best idea to conflate them together as a single/uniform phenomena; it might be a worthwhile effort to study these separately. Moreover, ‘hateful’ and ‘abusive’ seem to cluster together, and are almost equally distant from ‘rude’/‘impolite’ and ‘aggressive’/‘violent’. This is also depicted in the fact that ’abusive’/‘hateful’ language regularly co-occurs with both ‘impolite’ as well as ‘aggressive’ language usage.

We would like to reiterate here that these different terminologies, with varied semantics, if used for representing the same phenomena, are very likely to introduce ’noise’ in the annotation process and the final annotated data. This will result in multiple ‘noisy’ datasets that may not used with each other in a meaningful way. We believe that in the annotation of pragmatic phenomena like aggression or offensive language, the choice of terminology/tags is an extremely important step. Since the annotators are not and should not be given ‘exact’ guidelines on how to annotate (as in case of grammatical annotations like POS and morphology); rather, a lot of annotation decisions are left to their own subjective judgements—as such, it is of utmost importance that tags are defined in a way that they are consistent with their usage among the speakers. Now given this, using non-standard terminology such as ’toxicity’ for language is likely to be perceived differently by different speakers (leading to low inter-annotator agreement, low validity of the final dataset, and its low compatibility with datasets annotated with other similar sounding but different terminologies), and as such, the annotations are likely to be internally inconsistent (leading to a poorer performance of the system) as well as externally incompatible with other kinds of datasets such as those annotated with more commonly used and understood terminology such as ‘aggression’ or ‘offense’ or even ‘hate speech’. By way of this analysis, we seek to showcase the terminologies which are used in the similar contexts (and so likely to be semantically similar) and, hence, the datasets annotated with those tags might be complementary to each other; while the datasets annotated with semantically distant terminologies may not very compatible with each other. The analysis also shows that similar terminologies are more likely to co-occur with each in most of the cases, while dissimilar ones are less likely to be so. Since aggressive and offensive are quite distinct from each other, it predicts that they may not co-occur in a large number of cases (even though it does not completely rule out that possibility).

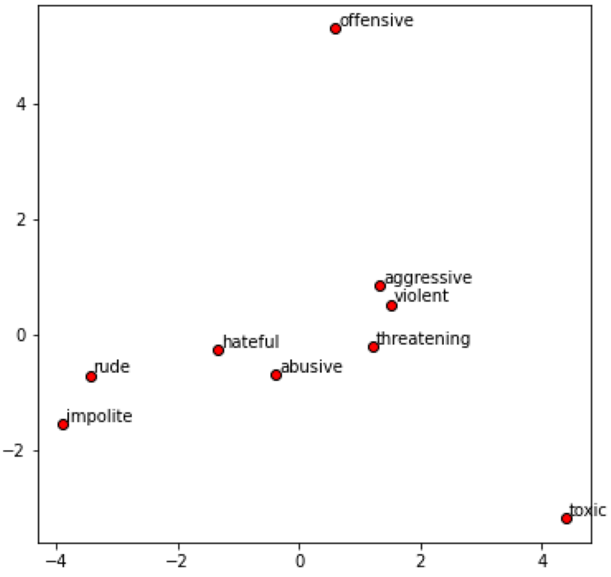

Fig. 2 Similarity among different lexical items related to offensive language

[10] The 300-dimensional vectors of each of the lexical items in the GLoVe model were reduced to 2-dimension using principal component analysis (PCA) and then plotted as a scatter diagram to generate this visualisation—it is a standard and well-accepted method for visualising high-dimensional vectors in lower dimensions while studying similarity.

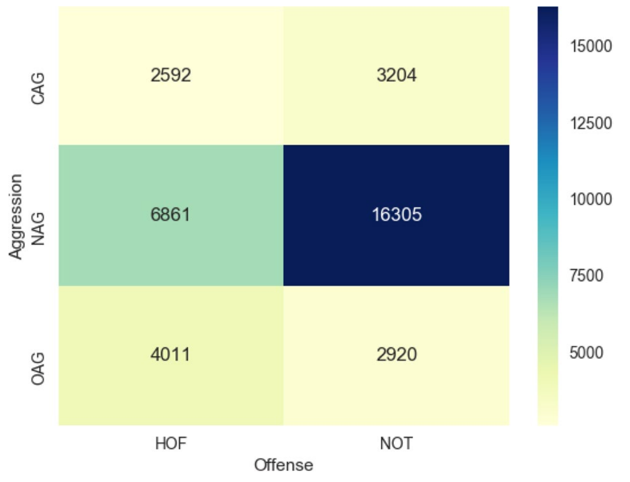

Fig. 3 Heatmap showing co-occurrence of aggressive and offensive language across languages

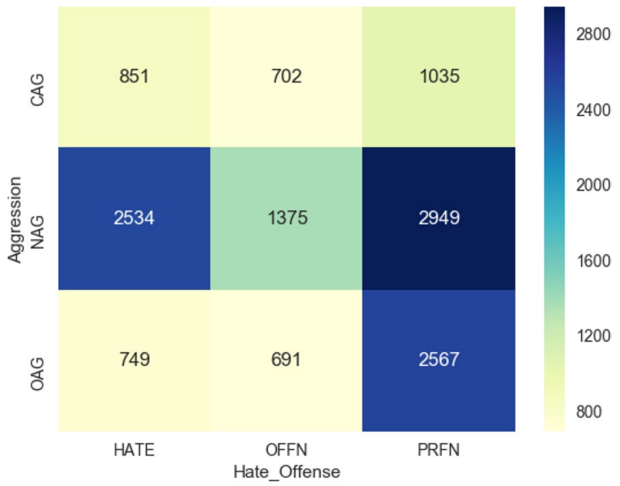

Fig. 4 Heatmap showing co-occurrence of aggressive and offensive language sub-classes across languages

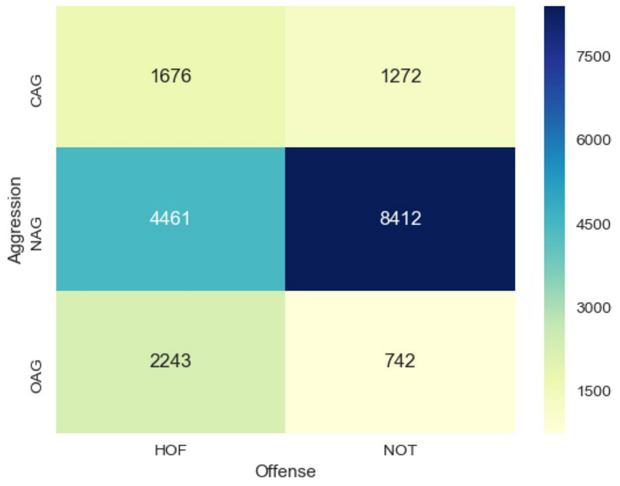

Fig. 5 Heatmap showing co-occurrence of aggressive and offensive language in Bangla

Co-occurrence of Aggressive and Offensive Language

In the previous section, we established that ‘aggressive’ and ‘offensive’ are semantically quite distinct from each other. In this section, using our dataset, annotated with both aggressive and offensive labels, we look at the co-occurrence of the different ‘aggressive labels’ (OAG, CAG, and NAG) with the offensive labels (HOF, NOT, HATE, PRFN,and OFFN). Figures 3 and 4 show the quantitative overlap in between these two sets of labels across all the three languages in our dataset, Figs. 5 and 6 show it for Bangla, Figs. 7 and 8 show it for Hindi, and Figs. 9 and 10 for English.

A closer look at these heatmaps points us towards some interesting patterns. It can be seen that there is a strong overlap between NAG and NOT, and there are relatively fewer examples of NAG and HOF or OAG/CAG and NOT. Most of the data which are NAG are also NOT across languages. Overall, approximately 70% of NAG comments are also NOT. However, at the same time, it is also worth noting that over 45% of PRFN comments are also NAG. This needs to be taken with a pinch of salt, since the results are based on a relatively smaller dataset in case of the fine-grained classifications and the labels are also assigned by classifiers which have not performed very well. However, in case of the major classes—CAG/OAG and HOF, both of these do not hold and we believe that the general trends shown by our dataset are likely to be representative of the linguistic facts, in general.

To understand how far the performance of the classifiers might have influenced the final analysis, we also looked at the HASOC and TRAC datasets separately. Since TRAC-2 classifiers have performed much better, it was expected that the HASOC dataset annotated with the TRAC-2 classifiers might give a better picture than the other way round. Figures 11 and 12 show the overall co-occurrence pattern in the HASOC dataset, and Figs. 13 and 14 show the co-occurrence pattern in the TRAC dataset. A comparison of the two datasets with each and with the overall distribution indeed shows some differences in the exact percentages of overlap; however, the general trends whereby NAG and NOT are very strongly correlated, while NOT is almost equally distributed in between OAG and CAG. In case of offensive language sub-classes, however, the results of the TRAC

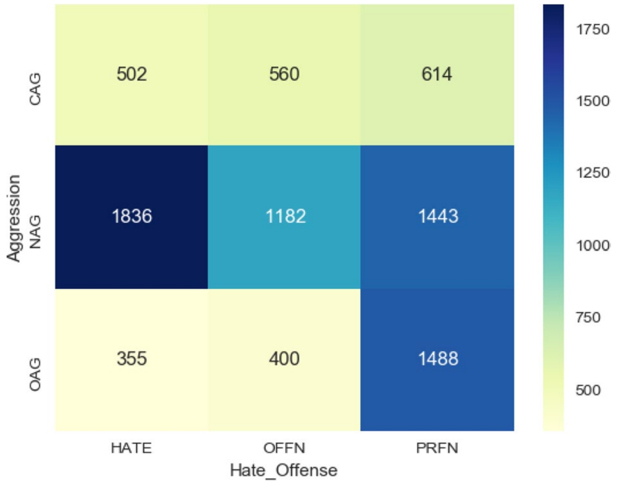

Fig. 6 Heatmap showing co-occurrence of aggressive and offensive language sub-classes in Bangla

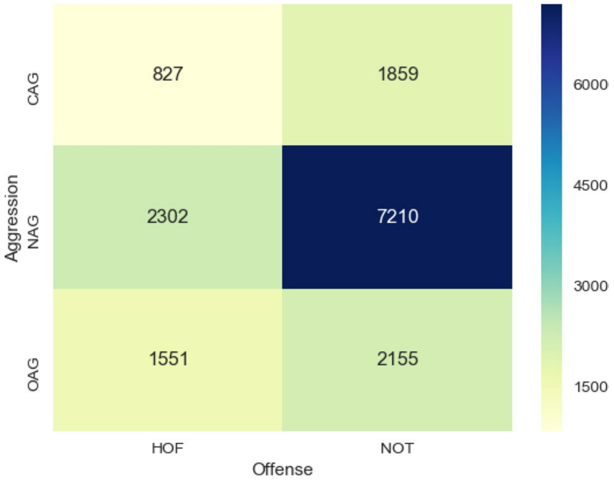

Fig. 8 Heatmap showing co-occurrence of aggressive and offensive language sub-classes in Hindi

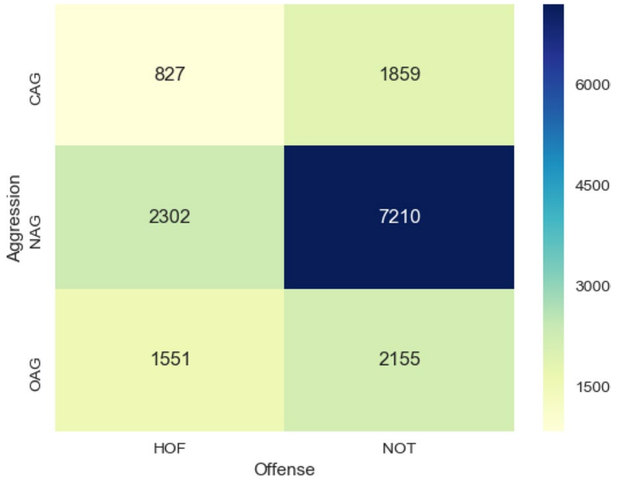

Fig. 7 Heatmap showing co-occurrence of aggressive and offensive language in Hindi

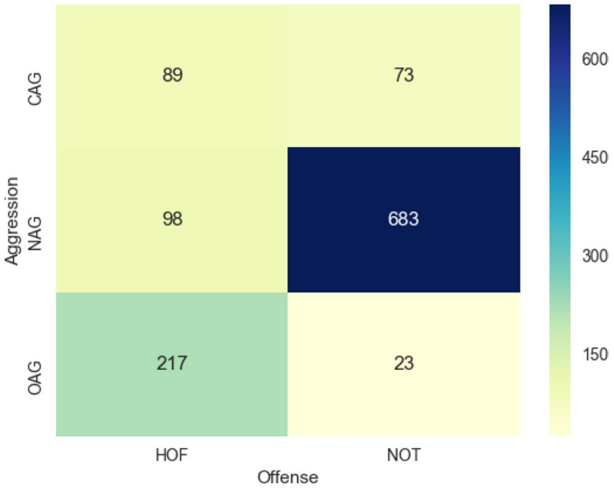

Fig. 9 Heatmap showing co-occurrence of aggressive and offensive language in English

dataset only marginally concur with the overall trends as well as the HASOC dataset. This clearly depicts the need for better annotated and larger dataset for more definitive conclusions.

If we look at the language—wise distribution—it can be seen that in Bangla the overlap between OAG and PRFN is the highest in number across the other sub-divisions. Over 50% OAG comments also contain profanity, while almost 40% are OFFN. We find a similar pattern in Hindi as well, though it is not same like Bangla. In Hindi too, the most number of OAG sentences overlap with PRFN, but HATE and OFFN are almost equally frequent in cases of OAG. It is also interesting to note that the OAG–PRFN overlap is followed by NAG–PRFN overlap in both Hindi and Bangla (though again in Bangla, NAG and CAG show similar distribution with respect to PRFN). In English, however, the data show a different pattern. The overlap between NAG and PRFN is highest (over 60% PRFN comments are considered NAG) unlike Bangla and Hindi. Moreover. out of all the NAG comments that are also HOF, almost 50% is also HATE. This points towards the fact a good amount of English sentences of the data is non-aggressive yet hateful, which indicates that aggression and hate do not always go together and one can be present without the other. It is also interesting to observe that, in English, a significantly larger proportion of NAG (over 1/3) is also HOF, while in Hindi, it is just over 1/4, and in Bangla, it is just over 1/8. It must be stressed at this point that some of these differences across languages may be attributed to the difference in our dataset owing to the different performance of the classifiers (used

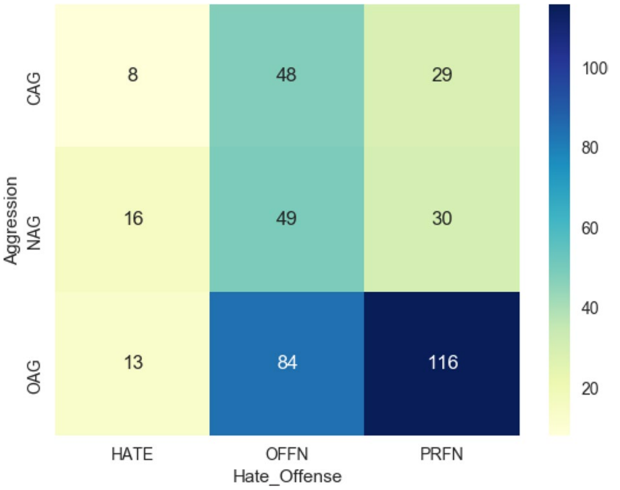

Fig. 10 Heatmap showing co-occurrence of aggressive and offensive language sub-classes in English

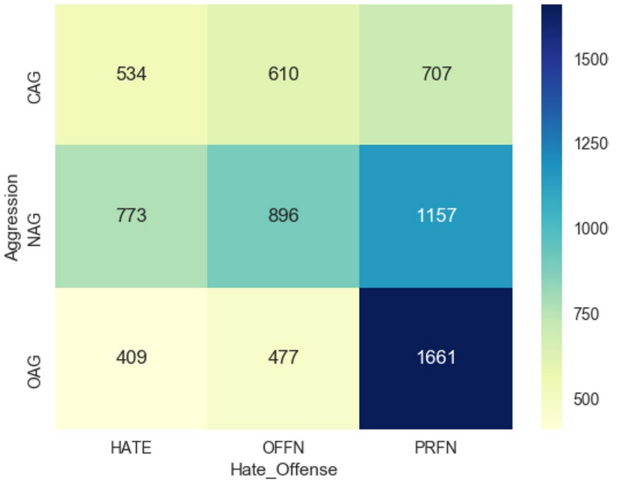

Fig. 12 Heatmap showing co-occurrence of aggressive and offensive language sub-classes in the HASOC dataset

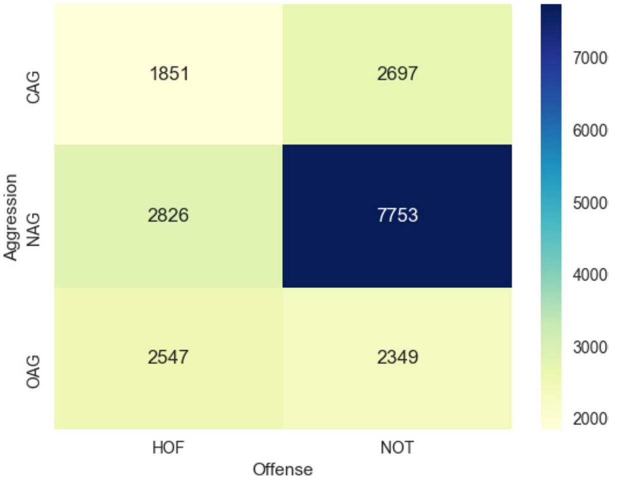

Fig. 11 Heatmap showing co-occurrence of aggressive and offensive language in the HASOC dataset

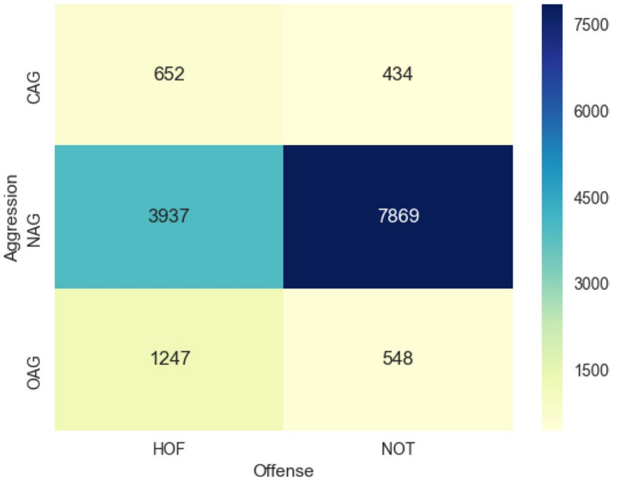

Fig. 13 Heatmap showing co-occurrence of aggressive and offensive language in the TRAC dataset

for generating the dataset used in the study), difference in the dataset sizes and also the fact that Bangla data, though significantly less in number, was completely manually annotated and, hence, may be considered the most robust representation.

The overlap pattern between different categories of aggression and offense can be summed up in the following points:

- NAG–NOT overlap is significantly high across languages, which is predictable. However, there is no such correlation between aggression and HOF as we see HOF being distributed quite evenly both across NAG and CAG/OAG.
- In Bangla and in Hindi, the overlap between OAG and PRFN is the highest in number, which is predictable, but there is also a significant overlap in between NAG and PRFN (more so in English), which shows that PRFN may not be very strongly correlated with aggression.

As this analysis shows, in a large number of the cases, aggression (both OAG and CAG) and HOF co-occur, yet there were significantly high number of instances where it can be seen that the sentences can be marked as CAG but not as HOF. Similarly, there were sentences which were HOF but not aggressive (NAG). Though it is difficult to mark the difference between HOF and aggression, as there is a thin line creating the division, but this analysis shows that the

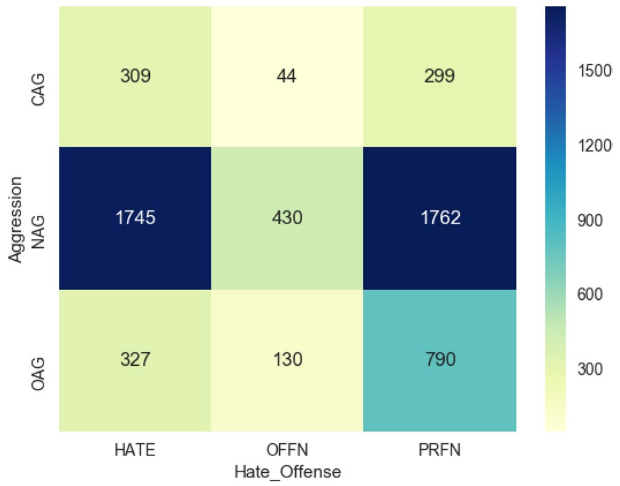

Fig. 14 Heatmap showing co-occurrence of aggressive and offensive language sub-classes in the TRAC dataset

line exists. We tried to explore this dividing line by looking more closely at the dataset.

It was observed that when the commenter ordered or suggested an individual or a group to do something, then the sentences were marked as CAG or OAG (but NOT), whereas when the sentences carried offensive adjectives or described the feature or character for addressing or describing a group, then the sentences were marked as HOF (but NAG). The following examples will explain the point:

1. "I think Arun dhoti should changed (change) her name kumfhu gadhi..." (I suggest Arundhoti should change her name to donkey).
2. "Should be cut the penies (penises) those criminals" (penises of those criminals should be cut).
3. "biswobhusan Sar isan ko jyada dhoka deta hai biwi aur girlfriend" (Biswobhusan sir, people are mostly cheated by their wives or girlfriends).
4. "chutiya ho tum nrc aanewala kagaj jama karo" (You are stupid, NRC is coming, submit your papers).

In example (1), this code-mixed data of English and Hindi, the writer suggests that Arundhuti should change her name to donkey which makes the sentence CAG. In example (2), written in Indian English, the writer suggests some action which should be performed on the criminals. These two examples (1 and 2) can be contrasted with the other two examples (3 and 4). It is apparent that while the first two examples mention some actions (changing of name and cutting of penises) to be performed, the other two examples describe a group. In example (3), female partners are targeted and described as cheaters, while in example (4), the addressee is mentioned as stupid.

The first two sentences are marked as CAG and NOT, i.e., they are covertly aggressive but still not offensive for the readers, since in both these cases, an individual or a group (who are generally considered to be 'bad' for the society) is being attacked and 'suggestions' are given to handle/punish them. Therefore, example (1) involves some kind of social harm for the target (hence, aggressive), while example (2) relates to a violent act (hence, aggressive). However, neither of these are taken insulting/impolite by the readers (because of the contextual factors which demand such treatment to them) and hence not offensive (or hateful). On the other hand, in the last two examples, neither social harm nor a violent act is being implied (hence not aggressive), but they are taken as an insult to the wives/girlfriends (and the contextual factors also reinforce this interpretation as 'insult') in example (3) and to the addressee in example (4), hence generating offense. Moreover, example (4) also makes use of profane word, which is interpreted as 'offensive'[11]. It was observed that in a large number of cases, OAG and HOF co-occur; however, there is no such relationship in between CAG and HOF (as was demonstrated in our quantitative study above as well). As such, we could posit CAG as one of the dividing lines between aggression and HOF.

Conclusion

In this paper, we have tried to understand and tease apart the distinction between aggression and offensiveness. To understand this distinction, we used the two datasets—released as part of the HASOC shared task and TRAC-2 shared task—marked for offensiveness and aggression. We conducted extensive experiments with these two datasets to develop automatic classifiers for aggression and different categories of offensiveness (as defined in [32]). The best classifiers in all instances achieved an $F1$-score in the approximate range of 0.70–0.80 [with some outliers on both sides of this range and the possible reason(s) for that is also discussed in the paper], which is reasonably good given the complexity of the task as well as the fact that most of the tasks were multi-class classification task. These experiments and their results gave valuable and interesting insights into how these derivatives of BERT perform in the present task as well as the effect of different kinds of pre-trained models on the final performance of the system. In general, it was shown that for non-standard varieties of English as well as the Indian languages, BERT is generally outperformed by classifiers like SVM because of the lack of sufficient data as well as the absence of good, relevant pre-trained models for

[11] We would like to reiterate here that simply use of profane word do not make the comment offensive—profanity is one of the factors that may lead to offense. However, as has been shown in this study neither of the two entail the other.

these languages. Hence, in such situations, it will always be a good idea to experiment with SVM-like classifiers (that are known to perform well in low-data scenarios) instead of relying on neural-network-based classifiers. We also carried out an extensive error analysis of these systems, which helped us in understanding the issues with these classifiers and also proved to be helpful in our analysis of the distinction between aggression and offense.

We used these classifiers to annotate a large dataset with both aggression and offense marked. This dataset was used to carry out a study of the distinction between aggression and offense. The word similarity-based study reinforced what has been proposed in the theoretical sociopragmatic literature—aggression is close to threatening and violence and relatively closer to impoliteness or rudeness; however, offense is an emotion which results from discursive practices like impoliteness or aggression. A more fine-grained quantitative and qualitative study of aggression and offense showed that two facts: first, the co-occurrence of aggression and offense may differ cross-linguistically; second, aggression and offense significantly overlaps, but still one does not entail the other. The distinction between aggression and offense becomes most clear in case of CAG and HOF, where a large number of CAG instances cannot be categorised as HOF.

While the analyses show some other distinctions as well, lack of sufficient data in cases of fine-grained distinctions (such as HATE, OFFN and PRFN) as well as the errors made by the classifiers in automatic annotation (which resulted in the dataset used for this study) make those distinctions unreliable. The errors in case of top-level categories are significantly less and the overall dataset is also large enough that the statistics would gloss over the classifiers' errors and it could be assumed that the general trends in the dataset is accurately presented (which is also cross-verified using our qualitative study of the dataset). We plan to include a larger dataset for the fine-grained labels in our future study, which will help both in reduction of the classifier errors as well as a better understanding of the differences between aggression and offense.

Acknowledgements We would like to thank the organisers of HASOC and TRAC-2 shared tasks for making the datasets publicly available, which has enabled us to carry out this research.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Agarwal S, Sureka A. Using knn and svm based one-class classifier for detecting online radicalization on twitter. In: International conference on distributed computing and internet technology. Springer; 2015. pp. 431–442.
2. Agarwal S, Sureka A. Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website. arXiv preprint. 2017;arXiv:1701.04931.
3. Badjatiya P, Gupta S, Gupta M, Varma V. Deep learning for hate speech detection in tweets. In: Proceedings of the 26th international conference on world wide web companion, International World Wide Web Conferences Steering Committee. 2017. pp. 759–760.
4. Basile V, Bosco C, Fersini E, Nozza D, Patti V, Rangel Pardo FM, Rosso P, Sanguinetti M. SemEval-2019 task 5: multilingual detection of hate speech against immigrants and women in twitter. In: Proceedings of the 13th international workshop on semantic evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA. 2019. pp. 54–63. https://doi.org/10.18653/v1/S19-2007. https://www.aclweb.org/anthology/S19-2007.
5. Bhattacharya S, Singh S, Kumar R, Bansal A, Bhagat A, Dawer Y, Lahiri B, Ojha AK. Developing a multilingual annotated corpus of misogyny and aggression. In: Proceedings of the second workshop on trolling, aggression and cyberbullying, European Language Resources Association (ELRA), Marseille, France. 2020. pp. 158–168.
6. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, Niculae V, Prettenhofer P, Gramfort A, Grobler J, Layton R, VanderPlas J, Joly A, Holt B, Varoquaux G. API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD workshop: languages for data mining and machine learning. 2013. pp. 108–122.
7. Burnap P, Williams ML. Hate speech, machine classification and statistical modelling of information flows on twitter: interpretation and communication for policy decision making. In: Proceedings of internet, policy and politics. 2014. pp. 1–18.
8. Cambria E, Chandra P, Sharma A, Hussain A. Do not feel the trolls. In: ISWC, Shanghai. 2010.
9. Chen Y, Zhou Y, Zhu S, Xu H. Detecting offensive language in social media to protect adolescent online safety. privacy, security, risk and trust (passat). In: International conference on social computing (SocialCom). 2012. pp. 71–80.
10. Culpeper J. Impoliteness: using language to cause offence. Cambridge: Cambridge University Press; 2011
11. Dadvar M, Trieschnigg D, de Jong F. Experts and machines against bullies: a hybrid approach to detect cyberbullies. In: Advances in artificial intelligence. Berlin: Springer; 2014. pp. 275–281.
12. Dadvar M, Trieschnigg D, Ordelman R, de Jong F. Improving cyberbullying detection with user context. In: Advances in information retrieval. Springer; 2013. pp. 693–696.
13. Davidson T, Warmsley D, Macy M, Weber I. Automated hate speech detection and the problem of offensive language. In: Proceedings of ICWSM. 2017.
14. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint. 2018; arXiv:1810.04805.
15. Díaz-Torres MJ, Morán-Méndez PA, Villasenor-Pineda L, Montes-y Gómez M, Aguilera J, Meneses-Lerín L. Automatic detection of offensive language in social media: defining linguistic criteria to build a Mexican Spanish dataset. In: Proceedings of the second workshop on trolling, aggression and cyberbullying, European Language Resources Association (ELRA), Marseille,

France. 2020. pp. 132–136. https://www.aclweb.org/anthology/2020.trac-1.21.

16. Dinakar K, Jones B, Lieberman CHH, Picard R. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Trans Interact Intell Syst (TiiS). 2012;2(3):18:1–30.
17. Djuric N, Zhou J, Morris R, Grbovic M, Radosavljevic V, Bhamidipati N. Hate speech detection with comment embeddings. In: Proceedings of the 24th international conference on world wide web. 2015. pp. 29–30.
18. Fortana P. Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes. Master's thesis, Faculdade de Engenharia da Universidade do Porto. 2017.
19. Gitari ND, Zuping Z, Damien H, Long J. A lexicon- based approach for hate speech detection. Int J Multimed Ubiquitous Eng. 2015;10(4):215–30.
20. Greevy E. Automatic text categorisation of racist webpages. Ph.D. thesis, Dublin City University. 2004.
21. Greevy E, Smeaton AF. Classifying racist texts using a support vector machine. In: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, ACM. 2004. pp. 468–469.
22. Hearst MA. Support vector machines. IEEE Intell Syst. 1998;13(4):18–28. https://doi.org/10.1109/5254.708428.
23. Hee CV, Lefever E, Verhoeven B, Mennes J, Desmet B, Pauw GD, Daelemans W, Hoste V. Detection and fine-grained classification of cyberbullying events. In: Proceedings of international conference recent advances in natural language processing (RANLP). 2015. pp. 672–680.
24. Kachru BB. The other tongue: English across Cultures. Urbana: University of Illinois Press; 1982.
25. Kachru BB. The Indianization of English: the English language in India. New Delhi: Oxford University Press; 1983.
26. Kumar R, Ojha AK. Kmi-panlingua at HASOC 2019: SVM vs BERT for hate speech and offensive content detection. In: Mehta P, Rosso P, Majumder P, Mitra M, editors. Working notes of FIRE 2019 - forum for information retrieval evaluation, Kolkata, India, December 12–15, 2019, CEUR workshop proceedings, vol. 2517, pp. 285–292. CEUR-WS.org. 2019. http://ceur-ws.org/Vol-2517/T3-14.pdf.
27. Kumar R, Ojha AK, Malmasi S, Zampieri M. 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.
28. Kumar R, Reganti AN, Bhatia A, Maheshwari T. Aggression-annotated corpus of hindi-english code-mixed data. In: Chair NCC, Choukri K, Cieri C, Declerck T, Goggi S, Hasida K, Isahara H, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S, Tokunaga T, editors. Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), European Language Resources Association (ELRA), Paris, France. 2018.
29. Kumar S, Spezzano F, Subrahmanian V. Accurately detecting trolls in slashdot zoo via decluttering. In: Proceedings of IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). 2014. pp. 188–195.
30. Lan Z, Chen M, Goodma S, Gimpel K, Sharma P, Soricut R. Albert: a lite bert for self-supervised learning of language representations. 2019.
31. Malmasi S, Zampieri M. Challenges in discriminating profanity from hate speech. J Exp Theor Artif Intell. 2018;30:1–16.
32. Mandl T, Modha S, Majumder P, Patel D, Dave M, Mandlia C, Patel A. 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. pp. 14–17. https://doi.org/10.1145/3368567.3368584.
33. Mandl T, Modha S, Majumder P, Patel D, Dave M, Mandlia C, Patel A. 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.
34. Mihaylov T, Georgiev GD, Ontotext A, Nakov P. Finding opinion manipulation trolls in news community forums. In: Proceedings of the nineteenth conference on computational natural language learning, CoNLL. 2015. pp. 310–314.
35. Mojica LG. Modeling trolling in social media conversations. 2016. arXiv:1612.05310 [cs.CL]. https://arxiv.org/pdf/1612.05310.pdf.
36. Montani JP, Schüller P. Tuwienkbs19 at germeval task 2, 2019: ensemble learning for german offensive language detection. In: Proceedings of the 15th conference on natural language processing (KONVENS 2019), German Society for Computational Linguistics and Language Technology, Erlangen, Germany. 2019. pp. 418–422.
37. Nitin Bansal A, Sharma SM, Kumar K, Aggarwal A, Goyal S, Choudhary K, Chawla K, Jain K, Bhasinar M. Classification of flames in computer mediated communications (2012). arXiv:1202.0617 [cs.SI]. https://arxiv.org/pdf/1202.0617.pdf.
38. Nitta T, Masui F, Ptaszynski M, Kimura Y, Rzepka R, Araki K. Detecting cyberbullying entries on informal school websites based on category relevance maximization. In: Proceedings of IJCNLP. 2013. pp. 579–586.
39. Nobata C, Tetreault J, Thomas A, Mehdad Y, Chang Y. Abusive language detection in online user content. In: Proceedings of the 25th international conference on world wide web, International World Wide Web Conferences Steering Committee. 2016. pp. 145–153.
40. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.
41. Pennington J, Socher R, Manning CD. Glove: global vectors for word representation. In: Empirical methods in natural language processing (EMNLP). 2014. pp. 1532–1543. http://www.aclweb.org/anthology/D14-1162
42. Sanh V, Debut L, Chaumond J, Wolf T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. 2019.
43. Sax S. Flame wars: automatic insult detection. Tech. rep., Stanford University. 2016.
44. Schmid F, Thielemann J, Mantwill A, Xi J, Labudde D, Spranger M. Fosil - offensive language classification of german tweets combining svms and deep learning techniques. In: Proceedings of the 15th conference on natural language processing (KONVENS 2019), German Society for Computational Linguistics and Language Technology, Erlangen, Germany. 2019. pp. 382–386.
45. Schmidt A, Wiegand M. 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.
46. Struß JM, Siegel M, Ruppenhofer J, Wiegand M, Klenner M. Overview of germeval task 2, 2019 shared task on the identification of offensive language. In: Proceedings of the 15th conference on natural language processing (KONVENS 2019), German Society for Computational Linguistics and Language Technology, Erlangen, Germany. 2019. pp. 354–365.
47. Tedeschi JT, Felson RB. Violence, aggression, and coercive actions. Washington: American Psychological Association; 1994.
48. Vigna FD, Cimino A, Dell'Orletta F, Petrocchi M, Tesconi M. Hate me, hate me not: hate speech detection on facebook. In:

Proceedings of the first Italian conference on cybersecurity, 2017. pp. 86–95.

49. Waseem Z, Davidson T, Warmsley D, Weber I. Understanding abuse: a typology of abusive language detection subtasks. In: Proceedings of the first workshop on abusive language online, Association for Computational Linguistics. 2017. pp. 78–84. http://aclweb.org/anthology/W17-3012.

50. Waseem Z, Hovy D. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: Proceedings of NAACL-HLT. 2016. pp. 88–93.

51. Wiegand M, Siegel M, Ruppenhofer J. Overview of the GermEval 2018 shared task on the identification of offensive language. In: Proceedings of GermEval. 2018.

52. Zampieri M, Malmasi S, Nakov P, Rosenthal S, Farra N, Kumar R. Predicting the type and target of offensive posts in social media. In: Proceedings of the annual conference of the North American chapter of the association for computational linguistics: human language technology (NAACL-HLT). 2019.

53. Zampieri M, Malmasi S, Nakov P, Rosenthal S, Farra N, Kumar R. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In: Proceedings of the 13th international workshop on semantic evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA. 2019. pp. 75–86. https://doi.org/10.18653/v1/S19-2010. https://www.aclweb.org/anthology/S19-2010.

54. Zampieri M, Nakov P, Rosenthal S, Atanasova P, Karadzhov G, Mubarak H, Derczynski L, Pitenis Z, Çöltekin c. SemEval-2020 task 12: multilingual offensive language identification in social media (OffensEval 2020). In: Proceedings of SemEval. 2020.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.