Natural Language Engineering (2023), **29**, pp. 1495–1515
doi:10.1017/S1351324923000104

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

A transformer-based multi-task framework for joint detection of aggression and hate on social media data

Soumitra Ghosh[1], Amit Priyankar[1], Asif Ekbal[1,*] and Pushpak Bhattacharyya[2]

[1]Department of Computer Science and Engineering, IIT Patna, India and [2]Department of Computer Science and Engineering, IIT Bombay, India
*Corresponding author. E-mail: asif@iitp.ac.in

(Received 31 August 2021; revised 28 May 2022; accepted 13 June 2022; first published online 11 April 2023)

Abstract

Moderators often face a double challenge regarding reducing offensive and harmful content in social media. Despite the need to prevent the free circulation of such content, strict censorship on social media cannot be implemented due to a tricky dilemma – preserving free speech on the Internet while limiting them and how not to overreact. Existing systems do not essentially exploit the correlatedness of hate-offensive content and aggressive posts; instead, they attend to the tasks individually. As a result, the need for cost-effective, sophisticated multi-task systems to effectively detect aggressive and offensive content on social media is highly critical in recent times. This work presents a novel multifaceted transformer-based framework to identify aggressive and hate posts on social media. Through an end-to-end transformer-based multi-task network, our proposed approach addresses the following array of tasks: (a) aggression identification, (b) misogynistic aggression identification, (c) identifying hate-offensive and non-hate-offensive content, (d) identifying hate, profane, and offensive posts, (e) type of offense. We further investigate the role of emotion in improving the system's overall performance by learning the task of emotion detection jointly with the other tasks. We evaluate our approach on two popular benchmark datasets of aggression and hate speech, covering four languages, and compare the system performance with various state-of-the-art methods. Results indicate that our multi-task system performs significantly well for all the tasks across multiple languages, outperforming several benchmark methods. Moreover, the secondary task of emotion detection substantially improves the system performance for all the tasks, indicating strong correlatedness among the tasks of aggression, hate, and emotion, thus opening avenues for future research.

Keywords: Hate speech; Offensive language; Aggressive posts; Transformer

1. Introduction

Social media interactions are frequently a mirror of offline interactions. Online, there are no geographical or temporal restrictions since anybody may join a conversation at any time, no matter where they are in the world. As a result, individuals no longer have to be afraid of societal reactions while expressing their opinions online. Recently, social media has propagated hate speech, mainly based on religion, cyberbullying, trolling, offensive posts, etc. They also utilize it to spread misinformation and hate messages the hard way. The Internet is home to a wide range of extreme views. A social problem exists here, as well as a technical one. As a result, misleading information undermines the information-sharing ecology in society. To create fear, uncertainty, and discord during a huge epidemic such as COVID-19, social media may be utilized as an instrument of

mistrust.[a] Misuse of these platforms may lead to prejudice and even violence, as well as economic, psychological, and political repercussions[b] (Weinstein 2018).

Nowadays, online hate speech and other unpleasant and undesirable information are significant issues. While democratizing the Internet, social media platforms can nonetheless generate conflict by allowing erroneous information to spread at an unparalleled rate.[c] Harvard University researchers discovered in a 2017 study that fake news travels "further, quicker, and deeper" on social media networks (Vosoughi *et al.* 2018). Social media monitoring agencies do not have the resources available to detect and remove such information swiftly. Persons who engage in objective debates are undermined by offensive language such as disrespectful, harmful, disparaging, or filthy material. There is an increasing demand for study into the automatic categorization of hate speech into several categories of objectionable content on social media platforms. Specific groups may incubate and disseminate their hate towards any individual or group on social media. However, when their speech reaches particular people, it can escalate into real-world violence.

Several recent occurrences have shown that when online anger crosses into the real world, it can be deadly. Facebook, Twitter, and other social media platforms quickly become the new battlegrounds of hatred. However, it appears that the tendency is worldwide. As a result of the El Paso shooting, the Trump administration has finally woken up to the realities of internet extremism. Within a week after the massacre, the White House convened a conference of tech firms to examine if the world's Google, Facebook, and Twitter might create magical algorithms that could detect the next gunman and anticipate the subsequent mass killing. The trust of Governments in technology may be as naive as their concerns.

Hate speech[d] is defined by the United Nations as any type of communication (spoken or written) in which a person or group is attacked or disparaged because of who they are, such as their race, ethnicity, gender, or other identifying factors. On the other hand, *abusive language* is a phrase that covers a wide range of language patterns, including offensive language, aggressive language, and hate speech. For example, cyberbullying, racism, sexism, and trolling may be detected by noting the use of abusive language.

The considerable overlap among hate speech, offensive language, aggressive posts, and other correlated tasks motivates us to investigate the interrelation among these tasks through a single end-to-end deep neural multi-task framework. Moreover, existing systems address these tasks individually, specifically the primary tasks of hate speech and aggression identification, which leaves scope for learning the interplay among the tasks in a collaborative learning environment. Our proposed system presents an automated multi-task network where we leverage the effectiveness of a pre-trained language model to extract the shared features and an independent multi-head self-attention network to extract the private features for the various tasks. The system additionally performs emotion identification that aids the overall system performance on all the tasks.

More specifically, we propose a transformer-based multi-task framework that addresses the following tasks simultaneously:

1. Task A: Aggression identification;
2. Task B: Misogynistic Aggression Identification;
3. Task C: Identifying Hate-Offensive and Non-Hate-Offensive content;
4. Task D: Identifying Hate, Profane, and Offensive posts;
5. Task E: Type of Offense;

The system features a shared XLM-RoBERTa (XLMR) (Conneau *et al.* 2020) model to represent the common features among the tasks and separate multi-head self-attention networks to

[a]Reports of hate crimes in India – A Washington Post Report
[b]https://news.un.org/en/story/2019/09/1047102
[c]https://www.axios.com/hate-speech-social-media-youth-2020-b13a1744-844a-4602-9fd2-8f055315d92a.html
[d]Hyperlink: United Nations Strategy and Plan of Action on Hate Speech

https://doi.org/10.1017/S1351324923000104 Published online by Cambridge University Press

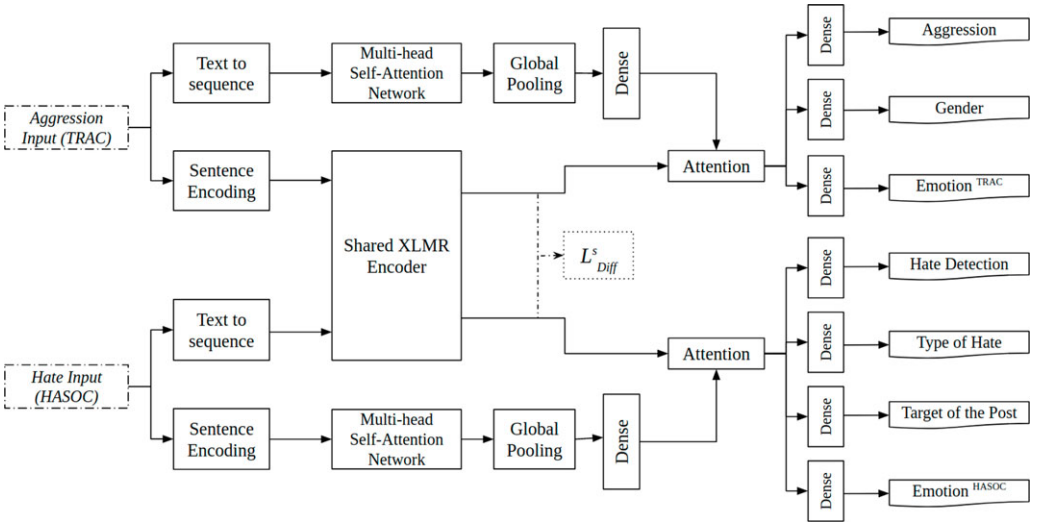

Figure 1. Overall architecture of the proposed transformer-based multi-task framework Detection *(MTFHAD)*.

describe the task-specific features. We consider the datasets introduced in HASOC-2019 (Mandl *et al.* 2019) and TRAC-2 2020 (Kumar *et al.* 2020) shared tasks to conduct our experiments. In addition to the above tasks, we also train our model to detect dataset-specific emotion (secondary task) for the input sentences, thus, learning two more tasks jointly, Emotion-TRAC (E-T) and Emotion-HASOC (E-H). We look at how the secondary task affects the overall performance of the primary tasks (Task A - E). The considered TRAC and HASOC datasets do not share any common task between them, particularly the emotion task, which we have included in this work by generating emotion labels through weak supervision (as discussed in Section 4.1.3). To include the emotion task in the overall training process, it is essential to associate the emotion task with each of the datasets; hence, two more tasks (secondary) are shown in the architecture (Figure 1). The evaluation findings reveal that our proposed multi-task system outperforms the current single-task benchmark setups on the majority of the tasks, demonstrating a high connection between the tasks evaluated.

The major contributions of this study are summarized as follows:

- We propose a single end-to-end Multi-task Transformer-based Framework for Hate speech and Aggressive Post Detection (*MTFHAD*) along with various correlated tasks.
- We investigate the role of the emotion identification task (secondary task) in increasing overall system performance for the primary tasks of recognizing hate-offensive material and aggressive posts when learned concurrently.
- We evaluate our proposed approach on two prominent multi-lingual datasets with four languages and find that it performs well on all tasks.

The rest of the paper is organized as follows. We cover prior research on the themes of hate speech, abusive language, and aggressive posts on social media in Section 2. In Section 3, we formulate our problem and explore our suggested framework. Section 4 discusses the datasets utilized in this study as well as the various experimental settings, as well as the results and discussion. In Section 5, we conclude our work and discuss future directions.

2. Related work

Previous studies on the topic have been conducted to automatically recognize certain related behaviors, such as trolling (Cambria *et al.* 2010), cyberbullying (Dinakar *et al.* 2012),

https://doi.org/10.1017/S1351324923000104 Published online by Cambridge University Press

abusive/offensive language (de la VegaandNg 2018), hate speech (Waseem and Hovy 2016; Malmasi and Zampieri 2018), racism (Greevy and Smeaton 2004), and others. These behaviors are deemed unpleasant, aggressive, and harmful for people on the receiving end. In addition, certain pragmatic studies on behavior, such as trolling, have been conducted (Hardaker 2010, 2013). Hardaker (2010) explains that trolling is designed to "create disturbance and/or instigate or aggravate conflict for their pleasure." A cyberbully is someone who engages in "humiliating and slandering actions towards other individuals" (Nitta *et al.* 2013). In a recent work by Jacobs *et al.* (2020), the authors propose to distinguish diverse participant roles involved in textual cyberbullying trials automatically. The work details the creation of two cyberbullying corpora (one in Dutch and one in English) that were manually annotated with bullying classes. A series of multiclass classification experiments are performed on the developed corpora to determine text-based cyberbullying participant role detection feasibility.

The SemEval-2019 Task 5 (Basile *et al.* 2019) introduced the task of detecting hate speech against immigrants and women. In another study, Tulkens *et al.* (2016) conducted a couple of experiments to identify racist discourse on Dutch social media. Each experiment used the same training data to train various classifiers. This training set used two public Belgian social media accounts containing Dutch postings that were likely to elicit racist reactions. At ELAVITA, the Hate Speech Detection task (HaSpeeDe) (Bosco *et al.* 2018) presented the shared challenge on Italian social media (Facebook and Twitter). Identifying and Categorizing Offensive Language on Social Media (OffensEval), Task 6 of SemEval-2019 (Zampieri *et al.* 2019b), introduced various sub to be undertaken on the Offensive Language Identification Dataset (OLID). The first sub-task involved distinguishing between offensive and non-offensive posts, whereas the second sub-task focussed on categorizing the type of offense. The purpose of Sub-task C was to identify the target of the offense. The GermEval Shared Job (Wiegand *et al.* 2018) on the Identification of Offensive Language established the task of classifying German tweets as offensive or non-offensive. Supervised classification techniques rely heavily on the annotated corpora. Several previous studies produced corpora that have been used for research purposes in the realm of hate speech. Many languages, including English, have shown substantial progress. HASOC, on the other hand, was the first shared task to introduce a labeled dataset for languages with minimal resources, such as Hindi and German. Both GermEval and OffensEval, two prior assessment forums, were the primary inspiration for creating HASOC.

Multi-tasking approaches have garnered the interest of researchers in recent times due to their capability to exploit the correlatedness among several tasks by effective knowledge sharing and provide superior performance on all the tasks compared to the single-task equivalent systems. Barnes *et al.* (2021) suggested a multi-task technique that outperforms learning negation in an end-to-end manner to directly add negation information into sentiment analysis. They described cascading and hierarchical neural networks with selective Long Short-Term Memory layers. It is demonstrated how explicit negation training improves sentiment analysis. Ghosh *et al.* (2022) proposed a multi-task framework for depression, sentiment, and multilabel emotion identification in suicide notes. The authors leveraged the cascading model mechanism with external knowledge infusion to improve the proposed system's performance on the primary task of multilabel emotion detection.

Anger or hostility against women is characterized as a misogynistic attitude.[e] Women-biased employment advertising is one form of online sexism that may be seen online. Shushkevich and Cardiff (2019) examined past research on automatic misogyny detection and discovered that classical machine learning methods, particularly ensembles, can outperform neural network-based approaches in several circumstances. However, because these studies were done on relatively small datasets, it is not guaranteed that the outcomes will be the same with a bigger dataset. Within

[e]https://www.dictionary.com/browse/misogyny?s=t

https://doi.org/10.1017/S1351324923000104 Published online by Cambridge University Press

this area, there have been several activities that have been shared, such as identifying misogynistic behavior and identifying specific types of sexism such as stereotyping, discredit, domination, sexual harassment, and threats of violence (Fersini *et al.* 2018).

Caselli *et al.* (2020b) worked on a recent English offensive language dataset, OLID/OffensEval (Zampieri *et al.* 2019a, 2019b) where the distinction between explicit and implicit signals was specifically highlighted, enhancing the data with a supplementary annotation layer. Also, new annotation guidelines were introduced and tested using OLID/OffensEval, resulting in AbuseEval v1.0. Some of the remaining difficulties in the annotation of offensive/abusive words were resolved by this newly developed English resource (e.g., message explicitness, the existence of a target, necessity for context, and interaction across multiple phenomena). The authors Poletto *et al.* (2017) detailed the creation of a social media corpus to represent and analyze hate speech directed towards certain minority groups in Italy. The study stresses the difficulties in creating a complex collection of labels that adequately reflect the fundamental elements of vocal hate utterances. A preliminary examination of the dataset and methods was also offered, along with an analysis of the annotators' disagreement. Caselli *et al.* (2020a) recently presented HateBERT, a pre-trained language model for abusive language phenomena in English. HateBERT consistently outperformed a generic BERT across a wide range of abusive language phenomena, including offensive language, abusive language, and hate speech. According to cross-dataset investigations, HateBERT was able to build robust representations of each abusive language phenomenon that it was fine-tuned against.

More recent systems, such as ToxicBERT (Hanu and Unitary team, 2020), fBERT (Sarkar *et al.* 2021), etc., are known to improve systems like BERT, HateBERT, etc. ToxicBERT is a BERT-based model that uses a transfer learning strategy to classify toxicity. It performed extremely well in the Toxic Comment Classification Challenge on Kaggle[f] with 93.64% F1 score. fBERT, which is also built using the BERT encoder, is trained on the SOLID dataset, containing over 1.4 million offensive instances. This model effectively infuses domain-specific offensive language and social media features, thus producing superior results than BERT and HateBERT on both OffensEval and HatEval tasks.

Numerous definitions and terminologies exist for the concepts of hate speech, offensive language, aggressive posts, etc. However, there appears to be a great deal of overlap in how each of these occurrences is perceived in different research. As a result of this overlap, insights from other fields may be useful in comprehending these seemingly different challenges. This work addresses hate-offensive content identification and aggression detection and their various nuances through a single end-to-end deep neural network model.

3. Methodology

This section explicitly defines our problem and presents the *MTFHAD* that we propose.

3.1. Task definition

Given a post (textual) of a social media user, identify the post as Hate and Offensive (*HOF*) or Non-Hate-Offensive (*NOT*) post and also classify the type of aggression from the following categories: Overtly Aggressive (*OAG*), Covertly Aggressive (*CAG*), and Non-Aggressive (*NAG*). In addition, any gender-directed aggression is classified as Gendered (*GEN*) and Non-Gendered (*NGEN*); the type of hate is categorized among the classes Hate speech (*HATE*), Offensive (*OFFN*), and Profanity (*PRFN*); type of offense is categorized as Targeted Insult (*TIN*) or Untargeted (*UNT*).

[f]https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

https://doi.org/10.1017/S1351324923000104 Published online by Cambridge University Press

Table 1. Examples of HOF and NOT instances from HASOC 2019 dataset.

Sentence	Emotion
By shitting yourself and taking the backdoor out, instead of fronting up to the public.	HOF
Bye Bye you foolish app.. Now sell your shit to these halala chaaps only ! #boycottzomato	HOF
John is one of our top Councillors and We thank you John and appreciate your hard Work..	NOT
You came, you saw . . .. we will look after the fort! Good luck!	NOT

Let $D = (s_1, s_2, \ldots, s_N)$ denote the entire dataset and s_1, s_2, $\ldots$, s_N be the instances in the dataset where The total number of instances is N. The task's goal is to maximize the value of the following function:

$$\underset{\theta}{argmax}\left(\Pi_{j=0}^{N} P\left(o_j^A, o_j^B, o_j^{E-T}, o_j^C, o_j^D, o_j^E, o_j^{E-H} | s_j; \theta\right)\right) \tag{1}$$

where s_j is the current user post whose output labels for Task A to E are $\left(o_j^A, o_j^B, o_j^C, o_j^D, o_j^E\right)$ to be predicted. o_j^t indicates the output label for task t. θ represents the model parameters that we want to optimize.

The overall architecture of the proposed methodology system is shown in Figure 1.

As described in Mandl *et al.* (2019), the Hate and Offensive (HOF) class constitutes all such posts that contain hate, offensive, and profane content and the Non-Hate-Offensive (NOT) class constitutes all such posts that do not contain any hate speech and/or offensive content. We show some examples for each HOF and NOT class in Table 1.

3.2. MTFHAD

We build an end-to-end deep multi-task framework that takes inputs from two channels: Aggression input from the TRAC-2 dataset and Hate-Offensive input from the HASOC dataset. The inputs are processed through an effective transformer-based shared-private network that generates rich contextualized feature representation and produces quality task-specific outputs.

Text to sequence: The text-to-sequence block prepares the raw inputs to be fed to the self-attention networks. Every input sequence (both for TRAC and HASOC data) $D^m = (w^1, w^2, \ldots, w^c)$ is a sequence consisting of c words. A word embedding layer and position encoding convert each token x^c in D^m into its vector representation. Instead of using pre-trained embeddings to initialize the embeddings weights for the words in the input sequence, each word is mapped to an "emb" dimensional vector that the model will learn during training. The constituents of such vectors are handled as model parameters, and back-propagation is used to optimize them just like any other weights.

$$w^c = Word\_Embedding(w^c) + Position\_Encoding(c) \tag{2}$$

$$E^c = [w^1, w^2, \ldots, w^c] \tag{3}$$

Private multi-head self-attention blocks: To extract private features from the aggression and hate inputs, we use two private multi-head self-attention networks that are multi-layer hierarchical transformer encoders. Each self-attention network block consists of three sequential transformers (Vaswani *et al.* 2017) encoders, each of which performs a multi-head self-attention operation on the embedding representation E^c, and the resulting output R^l is passed to point-wise fully linked feedforward network (FFN) layers to generate a knowledge representation (q^c).

https://doi.org/10.1017/S1351324923000104 Published online by Cambridge University Press

$$R^{(l)} = \text{MultiHeadAttn}(E^c) \tag{4}$$

$$S^{(l)} = \text{FFN}(R^{(l)}) \tag{5}$$

$$q^c = S^{(l)} \tag{6}$$

where l is the network's number of transformer encoders. The self-attention network output is routed through a global average pooling (GAP) layer, which is followed by a fully connected dense layer.

$$Q = \text{GAP}(q^c) \tag{7}$$

Shared XLMR Encoder: We consider XLMR (Conneau et al., 2020) as the shared encoder for the two datasets because of its ability to perform better on low-resource languages and model multi-lingual datasets. It is a massive multi-lingual model that was trained on 2.5 TB of CommonCrawl data in 100 distinct languages. It outperforms other transformer models, such as Bidirectional Encoder Representations from Transformers (BERT) and Multi-lingual BERT (mBERT), on cross-lingual benchmarks. The model performs well on multi-lingual datasets without sacrificing the competitive edge on monolingual benchmarks. We use the base version of the XLM-RoBERTa model that has 12 hidden layers, 250k parameters, 12 attention heads, and hidden dimension = 768.

As the input sentence, we take D^m (described previously), which is a token sequence of c words, and append [*CLS*] and [*SEP*] tokens at the beginning and end of the sequence, as shown below.

$$[CLS], w^1, w^2, \ldots, w^c, [SEP]$$

The [*CLS*] token denotes the start of the input sentence, while the [*SEP*] token denotes the end of the sentence. Each row in an input batch must have the same length. Thus, we add padding (or truncate sentences). Each token in the input sequence is replaced with a 768-dimensional word embedding vector during training. We consider the output from the special [*CLS*] token as the final hidden vector that gives the contextualized sentence representation.

Attention block: We apply additive-attention () between the private representation of each input (output from the self-attention network) with its shared equivalent (the output from the XLMR encoder) to get a weighted private representation that follows the dynamics of the shared space. Since the weight updates in the shared space are driven by the inputs from two distinct datasets, we wanted to make the private spaces aware of the shared correlation among the datasets. The intuition is to allow sufficient scope for the two broad correlated tasks (Hate and Aggression) to share knowledge and learn the inter-task relatedness from latent features while the model trains.

$$\alpha_i = \frac{exp\left(\gamma\left(S_* q_t^c\right)\right)}{\sum_{j=1}^{c} exp\left(\gamma\left(S_* q_j^c\right)\right)} \tag{8}$$

$$qw_t = \sum_{t=1}^{c} \alpha_i q_t^c \tag{9}$$

$$\gamma = W_3^T \tanh\left(W_1 S_* + W_2 q_t^c\right) \tag{10}$$

where W_1, W_2, W_3 are learnable weight matrices and *tanh* is a non-linear function. Here, "t" denotes the number of instances in the dataset and S_* denotes the dataset-specific shared output representation from the XLMR encoder. α_i, qw_t, and γ are the attention weights, context vector and attention vector, respectively.

https://doi.org/10.1017/S1351324923000104 Published online by Cambridge University Press

The output of the attention blocks is routed through task-specific dense layers, which are then routed to the relevant output layers. The attention output corresponding to the TRAC input captures the features of the aggression task and is passed to the task-specific layers corresponding to the sub for Aggression (Task A and Task B). Similarly, the attention block corresponding to the HASOC input-outputs the features for the hate input and passes them to the task-specific dense layers for the sub for Hate (Task C, Task D, and Task E). There are two task-specific dense and output layers, one each for the respective input-specific emotion detection tasks.

Loss function: We train our overall multi-task network to minimize the cross-entropy loss function shown below:

$$L_{task} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{n=1}^{k} y_{i,n}^{(t)} log\left(p_{i,n}^{(t)}\right) \tag{11}$$

where N is the number of samples, k is the number of task t classes, log is the natural logarithm, $y_{i,n}$ is 1 if sample i belongs to class n and 0 otherwise, and $p_{i,n}$ is the predicted probability that sample i belongs to class n. We assign equal weightage to the individual loss of each task and sum the losses to result in the overall system loss.

In addition to the task-specific cross-entropy losses, we compute the mean squared difference loss $\left(L_{Diff}^{s}\right)$ between the shared representations of the TRAC and HASOC datasets (H_i and S_i, outputs from the shared XLMR encoder), as shown in equation (12). Specifically, we compute the mean of the element-wise squared difference of ϕ_{TRAC} and ϕ_{HASOC} tensors, where ϕ depicts the output representation of a particular instance of the TRAC/HASOC dataset from the XLMR encoder. We employ the mean squared difference loss in particular to calculate the loss between the representations H_i and S_i due to one of its inherent drawbacks. Mean squared difference loss is known to heavily weigh the outliers as squaring of each term effectively weighs large errors more heavily than small ones Bermejo and Cabestany (2001). In our case, as the feature representations from the shared XLMR encoder are from two distinct inputs, hence L_{Diff}^{s} for any particular input pairs will supposedly be large; thus, incorporation of this loss to the overall training loss will help our model to train in a better way. Hence, putting it all together, the final loss function is represented as:

$$\mathrm{L}_{\mathrm{Diff}}^{\mathrm{s}} = \frac{1}{N}\sum_{i=1}^{N}(H_i - S_i)^2 \tag{12}$$

$$L = L_{task} + L_{Diff}^{s} \tag{13}$$

4. Datasets and experimental setting

The datasets utilized and the experimental setup are described in detail in this section.

4.1. Datasets

We evaluate our proposed method on the multi-lingual HASOC-2019 and TRAC-2 2020 datasets. We also prepare a consolidated emotion corpus from existing emotion corpora to train a weak emotion classifier for the generation of emotion labels on the HASOC (Mandl et al., 2019) and TRAC (Kumar et al., 2020) datasets.

https://doi.org/10.1017/S1351324923000104 Published online by Cambridge University Press

Table 2. Data distribution over train and test sets of HASOC 2019 for the 3 subtasks. '-' indicates no data as Subtask-C was not present for German language.

HASOC-2019	Sub-task A		Sub-task B			Sub-task C	
Tasks	NOT	HOF	HATE	OFFN	PRFN	TIN	UNT
Hindi Train	2196	2469	556	676	1237	1545	924
Hindi Test	713	605	190	197	218	542	63
English Train	3591	2261	1143	451	667	2041	220
English Test	865	288	124	71	93	245	43
German Train	3412	407	111	210	86	-	-
German Test	714	136	41	77	18	-	-

Table 3. Data distribution over train and test sets of TRAC-2 2020 for the 2 subtasks.

TRAC-2 2020	Sub-task A			Sub-task B	
Tasks	NAG	CAG	OAG	NGEN	GEN
Hindi Train	2823	1040	1118	4168	813
Hindi Test	316	215	669	700	500
English Train	4211	570	548	4947	382
English Test	690	224	286	1023	177
Bengali Train	2600	1116	1067	3880	903
Bengali Test	789	169	242	1005	195

4.1.1. HASOC-2019 shared task dataset (Mandl et al. 2019)

We utilize the multi-lingual datasets introduced in the HASOC[g] shared task. For each of the three languages (English, code-mixed Hindi, and German) presented in HASOC, there are three sub-tasks (Sub-task1, Sub-task2, and Sub-task3), and the data instances are garnered from Twitter and Facebook. Each English, Hindi, and German training set has 5852, 4665, and 3819 posts, respectively. There are 1153, 1318, and 850 posts in English, Hindi, and German test sets. Table 2 shows the data distribution of instances over the train and test sets for the HASOC shared task.

4.1.2. TRAC-2 2020 shared task dataset (Kumar et al. 2020)

This shared task competition has 5000 randomly chosen YouTube comments for training and 1000 comments for development. There are three categories for A (Aggression Identification): Overtly Aggressive (OAG), Covertly Aggressive (CAG), and Non-Aggressive (NAG). Misogynistic Aggression Identification is the emphasis of Sub-task B, which is a binary categorization between the two categories of GEN (gendered) and non-gendered misogynistic aggression (NGEN). Over 1000 comments are included in the test set. The statistics of the whole dataset in each language are displayed in Table 3.

[g]https://hasoc2019.github.io/

https://doi.org/10.1017/S1351324923000104 Published online by Cambridge University Press

Table 4. Data distribution over various emotion classes for the considered emotion datasets. '-' indicates the absence of a particular class in a dataset.

Datasets	Anger	Disgust	Fear	Joy	Sadness	Surprise	Others	Total
Emotion English	3124	2593	1067	4074	1765	1623	4500	18,746
Emotion Hindi	2286	665	171	2698	295	96	2877	9088
Disaster Hindi	658	-	500	782	1401	69	437	3847
Code-Mixed 1	471	-	304	490	324	-	-	1578
Code-Mixed 2	667	291	85	595	878	182	-	2698

4.1.3. Emotion recognition datasets for weak classifier

We generate emotion labels for each instance of the HASOC and TRAC-2 datasets using weak supervision. We train an XLMR-based emotion classifier on existing emotion datasets and generate predictions on the HASOC and TRAC-2 datasets. The following emotion datasets were used to prepare a consolidated emotion dataset with seven emotions and train the emotion classifier:

- Emotion Dataset in English (Ghosh *et al.* 2020): 18,746 instances
- Emotion Dataset in Hindi: 9088 instances
- Disaster Dataset in Hindi (Ahmad *et al.* 2020): 3847 instances
- Hindi-English Code-mixed data 1 (Singh 2021): 1578 instances
- Hindi-English Code-mixed data 2 (Vijay *et al.* 2018): 2698 instances

The emotion datasets in English and Hindi were created in-house as part of a larger study. Part of the emotion English dataset has been introduced in work by Ghosh *et al.* (2020). However, the distribution of cases across the various emotion classes in the provided dataset is substantially skewed. We considered an extended version of the dataset in this work, where we added additional instances in the under-represented emotion classes. We undersampled the *Others* class to attain a better distribution of instances. The data distribution over various emotion classes for the different emotion datasets is shown in Table 4. We split the overall dataset into the train, validation, and test sets in the ratios 80, 10, and 20, respectively. The classifier attains a test accuracy of 65.25% and a weighted F1 score of 65.07%. We show in Table 5 some sample emotion predictions on the instances from both the TRAC and HASOC datasets. Despite being trained on English and Hindi emotion data, manual evaluation of the predictions indicate that the XLMR cross-lingual classifier generates reliable emotion predictions for the HASOC-German (ge_H) and TRAC-Bengali (be_T) datasets as well.

4.2. Experimental setup

We use the Huggingface[h] Transformers package, a popular python-based library, to import the pre-trained XLMR model and also used Keras[i] and Scikit-learn[j] libraries at different stages of our implementation. Our experiments were carried out on an NVIDIA GeForce GTX 1080 Ti GPU. We set the input sentence length to 60 for both the XLMR and self-attention network inputs. We used the Adam (Kingma and Ba 2015) optimizer with a batch size of 16 to fully leverage the GPU.

[h] https://huggingface.co/transformers/
[i] https://keras.io/api/
[j] https://scikit-learn.org/

https://doi.org/10.1017/S1351324923000104 Published online by Cambridge University Press

Table 5. Sample emotion predictions on instances from different languages of the TRAC and HASOC datasets. The annotated labels for a particular instance from each dataset is shown in the square brackets. The text inside parentheses is the gloss for a particular non-English instance.

Dataset	Sentence	Emotion
$HASOC^{Ge}$	Hätte er mal nur #AllahuAkbar gerufen, dann wär's vielleicht weniger geworden? [HOF, HATE] *(Had he just only #AllahuAkbar called, then perhaps fewer become?)*	Surprise
$HASOC^{Ge}$	Schock für Angela Merkel: Mutter der Kanzlerin ist tot. Jeder reagiert anders im Schockzustand der großen Trauer. [NOT] *(Shock for Angela Merkel: Mother the chancellor is dead. Everyone reacts differently in Shock the great sadness.)*	Fear
$TRAC^{Be}$	Vikhari jmon Hoya uchit Proman hoye gache [CAG, NGEN] *(Beggar like be should, proof done was)*	Anger
$TRAC^{Be}$	story ta sotty korun,manusher mon j emon hai Keno [NAG, NGEN] *(story the truly sad, human mind the like that why)*	Sadness
$HASOC^{Hi}$	Aap to Pura family ko le dube Bhaiya ji. . . but koe Nhi hm sab [Aap ke] sath h. . . [HOF, HATE, UNT] *(You are whole family to take drowned Brother [salutation]. . . but any no we all your with is)*	Sadness
$HASOC^{Hi}$	Ahmed's dad:– beta aaj teri mammy kyu nahi baat [kr rhi] h. Ahmed. . .. [NOT] *(Ahmed's dad:– son today your mother why not talk doing is. Ahmed. . ..)*	Surprise
$HASOC^{En}$	These lists of banned substances have been around forever. Stupid boy. [NOT]	Disgust
$HASOC^{En}$	Shazia is the person I'm thinking of right now. She already lives under massive threat because of her stance on Islam, and now this from the Tommy team. An absolute disgrace. What is that chant they like to bellow out at marches? "shame on you". . . if the cap fits. [HOF, HATE, UNT]	Fear
$TRAC^{Hi}$	Yeh pagal aurat hai. . . Dnt listen to her. . . iske jaise aur bhi ghum rahe hai, jaha dikhe juta khol kar maro. [OAG, GEN] *(this mad lady is. . . Dnt listen to her. . . this like and also roam have been)*	Anger
$TRAC^{Hi}$	pagal tha jo mene kabhi virtual reality news par bahrosa kiya but now i am aware [NAG, NGEN] *(crazy was what i ever virtual reality news on believe did but now i am aware)*	Sadness
$TRAC^{En}$	He should have also killed that bitch [OAG, GEN]	Anger
$TRAC^{En}$	he is only 44? i was shocked to know that. looks 50+ atleast. [NAG, NGEN]	Surprise

We were able to set the number of epochs to 20 and the learning rate to 1e–5 by experimenting with [10,20,30] and [1e–3, 1e–4, 1e–5] and [1e–3, 1e–4, 1e–5]. The validation set's best model was preserved for testing. Each transformer in the knowledge encoding network featured eight self-attention heads, each having 512 embedding dimensions and 2048 feedforward dimensions. The hyper-parameters utilized in the experiments are shown in Table 6. We consider the weighted F1 and macro-F1 as the evaluation metrics for the TRAC-2 and HASOC-2019 datasets, respectively, as these are the official evaluation metrics as released by the task organizers.

4.3. Baselines

To assess the efficacy of our proposed technique, we compare our system's performance on the TRAC and HASOC datasets with several single-task and multi-task baselines, as well as specific state-of-the-art methodologies.

- TRAC-2 baselines
  - *Ms8qQxMbnjJMgYcw (Gordeev and Lykova 2020)*: The authors utilized a single BERT-based (Devlin *et al.* 2019) system with two outputs to perform all the tasks

https://doi.org/10.1017/S1351324923000104 Published online by Cambridge University Press

Table 6. Details of various hyper-parameters related to our experiments.

Parameters	Details
Self-Attention Network	Attention Heads: 8, Embedding dimension: 512,
	dimension inside transformer: 2048
	All dense layers (except output dense) have 100 neurons.
Number of Neurons	Output layers: Task A and Task E (3 neurons), Task B and
	Task C (2 neurons), Task D (4 neurons), Emotion tasks (7 neurons)
Hidden Activations	*ReLU* (Glorot *et al.* 2011) for dense layers
Output Activations	*Softmax* for all task-specific output layers.
Batch size	16
Epochs	20
Dropout (Srivastava *et al.* 2014)	25%
Loss	*Categorical Cross-Entropy* for all tasks;
Optimizer	*Adam* (Kingma and Ba 2015)

simultaneously. Results indicated that multi-task BERT fine-tuning for non-Indo-European languages might be seen as a promising method in this regard.

- *na14 (Safi Samghabadi et al. 2020)*: The authors demonstrated a BERT-based (Devlin et al., 2019) architecture with a multi-task approach. The proposed model leverages an attention mechanism over BERT to extract the relative relevance of words after fully connected layers and a final classification layer for each sub-task that predicted the class.
- *FlorUniTo (Koufakou et al. 2020)*: Using word embeddings that have been retrofitted to an abusive language vocabulary, an LSTM network model predicts the labels in this approach. The word embeddings have been changed such that terms from the same lexical categories are closer together in the vector space. When it comes to hate lexicons, the retrofitting technique has never been applied.
- *AI_ML_NIT_Patna (Kumari and Singh 2020)*: The authors suggested two deep learning systems based on Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) using FastText and one-hot embeddings (LSTM). The LSTM model with FastText embedding outperforms other models for the Hindi and Bangla datasets, whereas the CNN model with FastText embedding beats other models for the English dataset. It was also discovered that one-hot embedding and pre-trained FastText embedding perform similarly.

- HASOC-2019 baselines
  - *A3–108 (Mujadia et al. 2019)*: As part of the challenge, Team A3–108 submitted a range of machine learning and neural network-based models for all of the languages. Majority voting was utilized to create an ensemble model utilizing Support Vector Machine, Random Forest, and Adaboost classifiers.
  - *LGI2P (Mensonides et al. 2019)*: For each language and sub-task, the authors developed a different fastText-based model.
  - *RALIGRAPH (LuandNie 2019)*: The authors developed a vocabulary graph and employed graph convolutional networks as an embedding layer to add global

https://doi.org/10.1017/S1351324923000104 Published online by Cambridge University Press

information to the entire phrase, drawing inspiration from Text GCN (Yao *et al.* 2019). The BERT's self-attention encoder used vocabulary graph embedding and word embedding combined to encode the phrase with self-attention.

- *VITO (Nina-Alcocer 2019)*: Two techniques were explored to tackle this shared challenge. The results demonstrated that the initial way of employing CNNs and recurrent neural networks for n-gram processing and long-term dependencies did not produce satisfactory results. Improvement was noted when attention layers and part-of-speech vector representations were incorporated into the design process. The second approach, an ensemble of various machine learning, neural networks, and transformer-based models, provided the best overall performance.
- *HateMonitors (Saha et al. 2019)*: To detect abusive content using pre-trained BERT and LASER sentence embeddings, the authors used zero-shot transfer learning and pre-trained sentence embeddings. For the system to be language-independent, they employed the Gradient Boosting model coupled with BERT and LASER embeddings.
- *3Idiots (Mishra and Mishra 2019)*: BERT-based neural network models were used to refine the pre-trained monolingual and multi-lingual transformer models. Besides, the authors also looked at a method that relies on labels from all the sub together.

4.4. Results and analysis

The findings for the Hindi, English, Bengali, and German datasets of TRAC and HASOC common tasks are shown in Tables 7, 8, and 9, respectively. On Task B, C, and E, the overall findings show that our proposed *MTFHAD* system outperforms the baselines in all languages. We observe that our multi-task system provides strong performance for the misogynistic aggression identification task, irrespective of the language involved, and comfortably outperforms the baselines by greater than 2 points. The cross-lingual XLMR model in *MTFHAD* enables effective joint learning of the task features in two different languages (Bengali and German) and provides commendable results as depicted in Table 9.

While English and German are both members of the Indo-European language family's Germanic branch, Hindi and Bengali share Sanskrit roots. To understand the interplay among these pairs of languages, we performed an additional set of experiments considering the following dataset pairs: TRAC-English (en_T) and HASOC-German (ge_H), TRAC-Bengali (be_T) and HASOC-Hindi (hi_H). Table 10 displays the results. For the en_T-ge_H pair, we observe that learning the aggression tasks in English jointly with the hate task on German dataset proved to be beneficial for all the English tasks (tasks A and B) whereas not so for the Hate tasks on German.[k] We observe from the results that our model outperforms the previously attained scores on the Hindi tasks C and D,[l] when we consider the Bengali dataset of HASOC for joint training of the aggression and hate tasks. However, the performance on the Bengali tasks (task A and B) was better when be_T-ge_H dataset pair was considered. The languages presented in the TRAC (Hindi, English, Bengali) and HASOC (Hindi, English, German) datasets all come from the Indo-European language family. This may be a significant reason behind the performance improvement obtained by our proposed approach when the various tasks from different language types are learned jointly.

To account for the non-determinism of different TensorFlow GPU operations, we have reported F1 scores averaged across the 10 runs of the experiments. We conducted a Student's *t*-test with a 5% (0.05) significance level to illustrate that the scores obtained by the proposed *MTFHAD*

[k] comparing results between Tables 8, 9 and 10
[l] comparing results between Tables 7, 9 and 10

https://doi.org/10.1017/S1351324923000104 Published online by Cambridge University Press

Table 7. Results on TRAC-Hindi and HASOC-Hindi datasets. Values in bold indicate the maximum score for a perticular task. * indicates a model variant with no emotion task. '-' indicates no output for a particular task as it was not considered by the baseline system. Standard deviation values are shown inside parentheses.

		TRAC-2		HASOC-2019		
Models	Language	Task A	Task B	Task C	Task D	Task E
Baselines						
Ms8qQxMbnjJMgYcw (Gordeev and Lykova 2020)	hi_T	77.6	83.8	-	-	-
na14 (Safi Samghabadi et al., 2020)	hi_T	71.8	80.0	-	-	-
A3–108 (Mujadia *et al.* 2019)	hi_H	-	-	80.32	52.53	57.54
LGI2P (Mensonides et al., 2019)	hi_H	-	-	80.76	**56.17**	-
Proposed						
MTFHAD	hi_T-hi_H	**77.76** (± 0.4)	**86.57** (± 0.22)	81.09 (± 0.15)	53.26 (± 0.27)	**58.07** (± 0.16)
Ablation Experiments						
MTFHAD^{BERT}	hi_T-hi_H	77.73	83.39	80.22	55.97	57.04
MTFHAD*	hi_T-hi_H	75.70	81.72	**81.85**	51.32	53.41
MTFHAD†	hi_T-hi_H	71.60	78.61	81.27	39.58	54.16
MTFHAD§	hi_T-hi_H	75.33	82.59	80.55	55.89	56.24

system have not happened by chance. Specifically, we perform the test for significance on the *MTFHAD*-Hindi system on Tasks A, C, and E with the best-performing baselines (Mensonides et al., 2019; Gordeev and Lykova 2020; Mujadia *et al.* 2019) as the difference in scores is less than 1. The *p*-values attained are 0.036, 0.024, and 0.041, indicating that the obtained scores are statistically significant. We also perform the test for significance[m] on the results of the *MTFHAD*-English system on Task D against the VITO baseline. We observe a p-value of 0.038, indicating that the obtained result is statistically significant.

4.4.1. Comparison with the state-of-the-art

We observe from the reported results in Tables 7, 8, 9 that our proposed *MTFHAD* system significantly outperforms various existing methods on most of the tasks and produces a comparable performance on the rest. On the Hindi datasets, the *MTFHAD* model outperforms the baseline systems considerably on Task B, C, and E and gets equivalent results on Task A. However, it could not beat the system by Mensonides *et al.* (2019) on Task D. On all tasks except Task A, the *MTFHAD* system outperforms state-of-the-art approaches on the English datasets. FlorUniTo (Koufakou *et al.* 2020), which leveraged external task-specific lexicons in building their model, outperformed our system by 3 F1 points (approx.). On both the Bengali and German datasets, our suggested *MTFHAD* technique outperforms the baseline systems in all tasks. It is to be noted that many existing systems employ ensemble approaches (Mujadia *et al.* 2019; Nina-Alcocer 2019),

[m] Student's *t*-test

https://doi.org/10.1017/S1351324923000104 Published online by Cambridge University Press

Table 8. Results on TRAC-English and HASOC-English datasets. Values in bold indicate the maximum score for a particular task. * indicates a model variant with no emotion task. '-' indicates no output for a particular task as it was not considered by the baseline system. Standard deviation values are shown inside parentheses.

		TRAC-2		HASOC-2019		
Models	Language	Task A	Task B	Task C	Task D	Task E
Baselines						
FlorUniTo (Koufakou *et al.* 2020)	en_T	67.7	83.7	-	-	-
AI_ML_NIT_Patna (Kumari and Singh 2020)	en_T	66.0	82.2	-	-	-
RALIGRAPH (Lu and Nie 2019)	en_H	-	-	74.09	47.89	49.07
VITO (Nina-Alcocer 2019)	en_H	-	-	75.68	50.54	49.4
Proposed						
MTFHAD	en_T-en_H	64.52 ($\pm$ 0.45)	**86.12** ($\pm$ 0.29)	**76.69** ($\pm$ 0.23)	**50.69** ($\pm$ 0.23)	**50.72** ($\pm$ 0.33)
Ablation Experiments						
MTFHAD^{BERT}	en_T-en_H	62.44	**86.12**	75.11	48.70	47.97
MTFHAD*	en_T-en_H	61.80	85.40	72.30	47.76	48.14
MTFHAD†	en_T-en_H	**68.62**	85.69	73.97	50.48	47.80
MTFHAD§	en_T-en_H	66.72	85.66	74.96	47.08	47.99

which are resource and cost-intensive, often dependent on external datasets (Nina-Alcocer 2019) and lexicons (Koufakou *et al.* 2020) to boost their system performance. On the other hand, the proposed *MTFHAD* system is less resource-hungry and highly cost-effective. It delivers state-of-the-art performances on multiple tasks involving hate and aggression through a single end-to-end network.

4.4.2. Ablation study

To investigate the performance improvement of *MTFHAD* over the system proposed by Safi Samghabadi *et al.* (2020), where a simple transformer (mBERT) model with multiple heads was employed for each task (sharing the transformer model between each task), we developed *MTFHAD*^{BERT} by replacing the XLMR encoder in *MTFHAD* by mBERT. We present the results of *MTFHAD*^{BERT} in Table 7, 8, and 9. Results indicate that irrespective of the pre-trained transformer encoder (mBERT or XLMR) used in *MTFHAD*, both *MTFHAD* and *MTFHAD*^{BERT} outperform the baseline systems for the majority of the tasks. However, we observe that there is a significant performance drop over most of the tasks when we replaced the XLMR with mBERt in *MTFHAD* for the experiment with TRAC-Bengali and HASOC-German dataset pairs. The cross-lingual understanding ability of the XLMR encoder enables it to comprehend information from two different language pairs in a better way than mBERT in a single training setup. This ensures that the improvement of scores by the proposed system, when compared to the baselines, is mainly due to the underlying information-sharing architecture and not solely due to the shared document encoder employed.

To study the impact of emotion detection tasks in the overall learning process, as an ablation study, we develop *MTFHAD** for each language that does not consider the secondary task

https://doi.org/10.1017/S1351324923000104 Published online by Cambridge University Press

Table 9. Results on TRAC-Bengali and HASOC-German datasets. Values in bold indicate the maximum score for a particular task. Here, * indicates a model variant with no emotion task. '-' in the Baselines indicates no output for a particular task as it was not considered by the baseline system. '-' in the MTFHAD rows indicates no result as Subtask-C was not present for German language. Standard deviation values are shown inside parentheses.

		TRAC-2		HASOC-2019		
Models	Language	Task A	Task B	Task C	Task D	Task E
Baselines						
FlorUniTo (Koufakou *et al.* 2020)	be_T	74.5	86.8	-	-	-
AI_ML_NIT_Patna (Kumari and Singh 2020)	be_T	71.7	87.9	-	-	-
HateMonitors (Saha et al., 2019)	ge_H	-	-	61.62	27.69	-
3Idiots (Mishra and Mishra 2019)	ge_H	-	-	57.74	27.58	-
Proposed						
MTFHAD	be_T-ge_H	**78.35** ($\pm$ 0.47)	90.53 ($\pm$ 0.18)	**63.23** ($\pm$ 0.39)	**29.20** ($\pm$ 0.22)	-
Ablation Experiments						
MTFHAD^{BERT}	be_T-ge_H	74.96	90.99	57.03	27.73	-
MTFHAD*	be_T-ge_H	77.69	90.95	49.67	22.86	-
MTFHAD†	be_T-ge_H	71.59	90.53	59.22	25.76	-
MTFHAD§	be_T-ge_H	77.65	**91.18**	60.55	28.50	-

Table 10. Results considering the dataset pairs of the following language combination: TRAC-English and HASOC-German, TRAC-Bengali and HASOC-Hindi. '-' indicates no result as Subtask-C was not present for German language.

		TRAC-2		HASOC-2019		
Models	Language	Task A	Task B	Task C	Task D	Task E
MTFHAD	en_T-ge_H	68.13	86.54	62.70	29.44	-
MTFHAD	be_T-hi_H	74.53	89.23	81.47	56.50	57.08

of emotion detection. Results indicate that consideration of the emotion task significantly boosts the system performance on all the tasks, hinting at a strong correlation between aggression, hate, and emotion tasks. We conduct another set of ablation experiments to investigate the impact of the dataset-specific self-attention networks in obtaining the overall performance improvement by our proposed method when compared to the state-of-the-art systems. We develop *MTFHAD*† by removing the self-attention blocks and their following pooling and attention layers from *MTFHAD*, which leaves only the shared XLMR encoder to generate the input features before passing them to the task-specific dense layers. Tables 7, 8, and 9 depict the overall results. We observe notable performance deterioration for most of the tasks over the various datasets and language pairs, which indicates that private self-attention networks play a critical role in boosting the system's overall performance.

https://doi.org/10.1017/S1351324923000104 Published online by Cambridge University Press

We also examine the importance of the mean squared difference loss in the overall performance of our approach, *MTFHAD*, by removing L^{s}_{Diff} loss and developing *MTFHAD*[§]. For all three language pair setups, as shown in Tables 7, 8, and 9, we observe a notable fall in scores over most of the tasks on both the datasets. This depicts that the mean squared difference loss plays a crucial role in improving the system's overall performance. We observed average (over all the tasks) performance improvement of 1.23, 1.27, and 0.86 F1 score points for the Hindi-Hindi, English-English, and Bengali-German language pairs of the TRAC and HASOC datasets. The possible reasons for the lowest improvement score from the Bengali-German setup may be the languages belonging to the low-resource languages and being dissimilar language pairs.

4.4.3. Error analysis

To understand the limitations of our methodology, we conducted a rigorous qualitative analysis of the misclassifications made by our *MTFHAD* system. We categorize the challenges under the following points:

- *Errors in annotations:* We observed several wrongly annotated instances in both the TRAC and HASOC datasets which limited our system to train properly on the aforesaid datasets. Our model failed to make correct predictions on certain instances with conflicting annotations in the training data despite being similar contextually. For example, consider the first two sentences below from the TRAC-2 Hindi train set, which are similar in length and also carry the exact contextual meaning, yet the annotations are different. In such instances, with minimum context information and mention of a slang word, our proposed *MTFHAD* system identifies them as belonging to the *CAG* class. The third and fourth sentences are from the HASOC-2019 English train set, where we observe that both the sentences carry negative sentiment and are offensive. Still, they have conflicting annotations.

    1. *"Chutiya movie. . ." ([slang] movie. . .) – NAG*
    2. *"Chutiya bhakt." ([slang] devotee) – CAG*
    3. *"Let's be clear there is a deference between oppo-reasearch and foreign influence and there is a reason why it is discourage! Fuck trump and his disciples! That's right disciples!! #Fucktrump #impeachtrumpnow!" – NOT*
    4. *"Fuck Trump and anybody who voted for that Lyin POS! #FuckTrump https://t.co/sudpYAU1Eu" – HOF, PRFN, and TIN*

- *Noise in datasets:* Closer analysis of the instances in the datasets reveals that the language-specific datasets contain noise, such as, for a particular dataset in one language, instances of a different language are present in that same dataset. For example, the first example is a romanized Hindi post present in the TRAC-2 Bengali test set, and the second post is a Bengali post in the TRAC-2 Hindi test set.

    1. *kabhi time nikal ke mar ja na. . . kuttia. . . (sometimes time out of die go no. . . [slang]. . .)*
    2. *last duto line. . . just mon chue gelo boss. (last two line. . . just mind touch was boss.)*

1. Linguistic problems and a lack of clean code-mixed data pose severe challenges in building an efficient classifier to perform any downstream classification task on such data. Class-specific cleaner data would be required to eliminate the impact of spelling errors, stemmed phrases, and the usage of various contexts.

    - *Datasets with limited diversity in topics:* Almost all instances in the TRAC-2 Bengali dataset (Train, Validation, and Test sets) surprisingly involve posts directed towards

https://doi.org/10.1017/S1351324923000104 Published online by Cambridge University Press

a single person (Ranu Mondal). The TRAC-2 Hindi datasets, on the other hand, are limited to a handful of topics related to an individual (Akshay Khanna, Rape, Feminism, Movies).

- *Model biases towards over-represented classes:* Empirical evaluation showed that the classifiers performed well when classes were balanced and contained sufficient number of instances in the training set. On the other hand, the lack of under-represented classes such as low frequency of profane tweets over all the datasets for all the languages made it difficult for our model to predict them correctly. For the German dataset, our system performed poorly for all the s.

5. Conclusion

In this paper, we proposed *MTFHAD*, a novel, multi-task transformer-based architecture for identifying aggressive and hateful posts on social media. We employ a shared-private multi-task network to handle a variety of tasks, including the following: aggression identification, misogynistic aggression identification, identifying hate-offensive and non-hate-Offensive content, identifying hate, profane, and offensive posts, and Type of Offense. We assess our system on two popular benchmark datasets of four languages, TRAC and HASOC. Comprehensive evaluation indicates that our multi-tasking system outperforms several existing benchmark techniques for most tasks, regardless of the language used. Aside from that, the secondary job of emotion detection greatly enhances the system's performance for all tasks, suggesting that aggressiveness, hatred, and emotion are firmly connected, thus opening up new study paths. In terms of cost-effectiveness and resource requirements, our suggested *MTFHAD* system outperforms existing techniques. It can handle several tasks involving aggressive posts and hate speech over multiple languages through a single framework.

Future studies should leverage task-specific lexicons to elevate system performance and consider external knowledge sources to build knowledge graphs that may infuse valuable context/information in the learning process to make the proposed approach more generic and robust across different datasets. It would be interesting to see how sexual and gender identities affect the system's overall efficacy if considered during training. The presented results may also be improved if unequal weightage for the individual task losses is considered (instead of equal weightage) that would enable to find the right balance among the various participating tasks towards reaching an optimum system state.

Acknowledgement. The authors gratefully acknowledge partial support from the sponsored project *HELIOS* – Hate, Hyperpartisan, and Hyperpluralism Elicitation and Observer System, sponsored by Wipro. Asif Ekbal acknowledges the Young Faculty Research Fellowship, supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

Competing interests. The authors declare none.

References

Ahmad Z., Jindal R., Ekbal A. and Bhattachharyya P. (2020). Borrow from rich cousin: transfer learning for emotion detection using cross lingual embedding. *Expert Systems with Applications* **139**, 112851.

Bahdanau D., Cho K. and Bengio Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*.

Barnes J., Velldal E. and Øvrelid L. (2021). Improving sentiment analysis with multi-task learning of negation. *Natural Language Engineering* **27**(2), 249–269.

Basile V., Bosco C., Fersini E., Debora N., Patti V., Pardo F. M. R., Rosso P. and Sanguinetti M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, pp. 54–63.

https://doi.org/10.1017/S1351324923000104 Published online by Cambridge University Press

Bermejo S. and Cabestany J. (2001). Oriented principal component analysis for large margin classifiers. *Neural Networks* **14**(10), 1447–1461.

Bosco C., Dell'Orletta F., Poletto F., Sanguinetti M. and Tesconi M. (2018). Overview of the EVALITA 2018 hate speech detection task. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) Co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12–13, 2018, CEUR Workshop Proceedings*, Vol. **2263**. Available at https://ceur-ws.org/

Cambria E., Chandra P., Sharma A. and Hussain A. (2010). *Do Not Feel the Trolls*. Shanghai: ISWC.

Caselli T., Basile V., Mitrovic J. and Granitzer M. (2020a). Hatebert: Retraining BERT for abusive language detection in english. *CoRR*, abs/2010.12472.

Caselli T., Basile V., Mitrovic J., Kartoziya I. and Granitzer M. (2020b). I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11–16, 2020*. European Language Resources Association, pp. 6193–6202.

Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L. and Stoyanov V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020*. Association for Computational Linguistics, pp. 8440–8451.

de la Vega L. G. M. and Ng V. (2018). Modeling trolling in social media conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7–12, 2018*. European Language Resources Association (ELRA).

Devlin J., Chang M., Lee K. and Toutanova K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, pp. 4171–4186.

Dinakar K., Jones B., Havasi C., Lieberman H. and Picard R. W. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems* **2**(3), 30–30.

Fersini E., Rosso P. and Anzovino M. (2018). Overview of the task on automatic misogyny identification at ibereval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) Co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, CEUR Workshop Proceedings, Vol. **2150**, pp. 214–228. Available at https://ceur-ws.org/

Ghosh S., Ekbal A. and Bhattacharyya P. (2022). A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes. *Cognitive Computation* **14**, 110–129.

Ghosh S., Ekbal A., Bhattacharyya P., Saha S., Tyagi V., Kumar A., Srivastava S. and Kumar N. (2020). Annotated corpus of tweets in English from various domains for emotion detection. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*. Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI), pp. 460–469.

Glorot X., Bordes A. and Bengio Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11–13, 2011*, JMLR Proceedings, Vol. **15**, pp. 315–323. Available at https://www.jmlr.org/

Gordeev D. and Lykova O. (2020). BERT of all trades, master of some. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, Marseille, France*. European Language Resources Association (ELRA), pp. 93–98.

Greevy E. and Smeaton A. F. (2004). Classifying racist texts using a support vector machine. In *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25–29, 2004*. New York: ACM, pp. 468–469.

Hanu L. and Unitary Team (2020). *Detoxify*. Github. Available at https://github.com/unitaryai/detoxify

Hardaker C. (2010). Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research* **6**(2), 215J242.

Hardaker C. (2013). "uh.... not to be nitpicky, but. . . the past tense of drag is dragged, not drug.": An overview of trolling strategies. *Journal of Language Aggression and Conflict* **1**(1), 58–86.

Jacobs G., Van Hee C. and Hoste V. (2020). Automatic classification of participant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text? *Natural Language Engineering*, **28**, 141–166.

Kingma D. P. and Ba J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*.

Koufakou A., Basile V. and Patti V. (2020). FlorUniTo@TRAC-2: Retrofitting word embeddings on an abusive lexicon for aggressive language detection. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, Marseille, France*. European Language Resources Association (ELRA), pp. 106–112.

Kumar R., Ojha A. K., Malmasi S. and Zampieri M. (2020). Evaluating aggression identification in social media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, TRAC@LREC 2020, Marseille, France, May 2020*. European Language Resources Association (ELRA), pp. 1–5.

https://doi.org/10.1017/S1351324923000104 Published online by Cambridge University Press

Kumari K. and Singh J. P. (2020). AI_ML_NIT_Patna @ TRAC - 2: Deep learning approach for multi-lingual aggression identification. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pp. 113–119, Marseille, France*. European Language Resources Association (ELRA), pp. 113–119.

Lu Z. and Nie J. (2019). RALIGRAPH at HASOC 2019: VGCN-BERT: Augmenting BERT with graph embedding for offensive language detection. In *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12–15, 2019*, CEUR Workshop Proceedings, Vol. **2517**, pp. 221–228. Available at https://ceur-ws.org/

Malmasi S. and Zampieri M. (2018). Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence* **30**(2), 187–202.

Mandl T., Modha S., Majumder P., Patel D., Dave M., Mandlia C. and Patel A. (2019). 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, New York, NY, USA*. Association for Computing Machinery, pp. 14–17.

Mensonides J., Jean P., Tchechmedjiev A. and Harispe S. (2019). IMT mines ales at HASOC 2019: Automatic hate speech detection. In *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12–15, 2019*, CEUR Workshop Proceedings, Vol. **2517**, pp. 279–284. Available at https://ceur-ws.org/

Mishra S. and Mishra S. (2019). 3idiots at HASOC 2019: Fine-tuning transformer neural networks for hate speech identification in indo-european languages. In *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12–15, 2019*, CEUR Workshop Proceedings, Vol. **2517**, pp. 208–213. Available at https://ceur-ws.org/

Mujadia V., Mishra P. and Sharma D. M. (2019). Iiit-hyderabad at HASOC 2019: Hate speech detection. In *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12–15, 2019*, CEUR Workshop Proceedings, Vol. **2517**, pp. 271–278. Available at https://ceur-ws.org/

Nina-Alcocer V. (2019). Vito at HASOC 2019: Detecting hate speech and offensive content through ensembles. In *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12–15, 2019*, CEUR Workshop Proceedings, Vol. **2517**, pp. 214–220. Available at https://ceur-ws.org/

Nitta T., Masui F., Ptaszynski M., Kimura Y., Rzepka R. and Araki K. (2013). Detecting cyberbullying entries on informal school websites based on category relevance maximization. In *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14–18, 2013*. Asian Federation of Natural Language Processing/ ACL, pp. 579–586.

Poletto F., Stranisci M., Sanguinetti M., Patti V. and Bosco C. (2017). Hate speech annotation: Analysis of an italian twitter corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, December 11–13, 2017*, CEUR Workshop Proceedings, Vol. **2006**. Available at https://ceur-ws.org/

Safi Samghabadi N., Patwa P., S. P. Y. K. L., Mukherjee P., Das A. and Solorio T. (2020). Aggression and misogyny detection using BERT: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, Marseille, France*. European Language Resources Association (ELRA), pp. 126–131.

Saha P., Mathew B., Goyal P. and Mukherjee A. (2019). Hatemonitors: Language agnostic abuse detection in social media. In *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12–15, 2019*, CEUR Workshop Proceedings, Vol. **2517**, pp. 246–253. Available at https://ceur-ws.org/

Sarkar D., Zampieri M., Ranasinghe T. and Ororbia A. (2021). fBERT: A neural transformer for identifying offensive content. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 1792–1798.

Shushkevich E. and Cardiff J. (2019). Automatic misogyny detection in social media: A survey. *Computación y Sistemas* **23**(4). https://doi.org/10.13053/cys-23-4-3299

Singh D. (2021). Detection of emotions in hindi-english code mixed text data. *CoRR* abs/2105.09226.

Srivastava N., Hinton G. E., Krizhevsky A., Sutskever I. and Salakhutdinov R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* **15**(1), 1929–1958.

Tulkens S., Hilte L., Lodewyckx E., Verhoeven B. and Daelemans W. (2016). The automated detection of racist discourse in dutch social media. *Computational Linguistics in the Netherlands Journal* **6**, 3–20.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L. and Polosukhin I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA*, pp. 5998–6008,

Vijay D., Bohra A., Singh V., Akhtar S. S. and Shrivastava M. (2018). Corpus creation and emotion prediction for Hindi-English code-mixed social media text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. New Orleans, LA: Association for Computational Linguistics, pp. 128–135.

Vosoughi S., Roy D. and Aral S. (2018). The spread of true and false news online. *Science* **359**(6380), 1146–1151.

Waseem Z. and Hovy D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In *Proceedings of the Student Research Workshop, SRW@HLT-NAACL 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California, USA, June 12–17, 2016*. The Association for Computational Linguistics, pp. 88–93.

Weinstein J. (2018). *Hate Speech, Pornography, and the Radical Attack on Free Speech Doctrine*. New York: Routledge.

Wiegand M., **Siegel M. and Ruppenhofer J.** (2018). Overview of the germeval 2018 shared task on the identification of offensive language. In *14th Conference on Natural Language Processing KONVENS 2018*.

Yao L., **Mao C. and Luo Y.** (2019). Graph convolutional networks for text classification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27–February 1, 2019*. AAAI Press, pp. 7370–7377.

Zampieri M., **Malmasi S.**, **Nakov P.**, **Rosenthal S.**, **Farra N. and Kumar R.** (2019a). Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, pp. 1415–1420.

Zampieri M., **Malmasi S.**, **Nakov P.**, **Rosenthal S.**, **Farra N. and Kumar R.** (2019b). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 75–86.

Cite this article: Ghosh S, Priyankar A, Ekbal A and Bhattacharyya P (2023). A transformer-based multi-task framework for joint detection of aggression and hate on social media data. *Natural Language Engineering* **29**, 1495–1515. https://doi.org/10.1017/S1351324923000104