Sci

MDPI

Article

T5 for Hate Speech, Augmented Data, and Ensemble

Tosin Adewumi *, Sana Sabah Sabry, Nosheen Abid, Foteini Liwicki and Marcus Liwicki

ML Group, EISLAB, Luleå University of Technology, 97187 Luleå, Sweden; sana.al-azzawi@ltu.se (S.S.S.); foteini.liwicki@ltu.se (F.L.); marcus.liwicki@ltu.se (M.L.)
* Correspondence: tosin.adewumi@ltu.se

Abstract: We conduct relatively extensive investigations of automatic hate speech (HS) detection using different State-of-The-Art (SoTA) baselines across 11 subtasks spanning six different datasets. Our motivation is to determine which of the recent SoTA models is best for automatic hate speech detection and what advantage methods, such as data augmentation and ensemble, may have on the best model, if any. We carry out six cross-task investigations. We achieve new SoTA results on two subtasks—macro F1 scores of 91.73% and 53.21% for subtasks A and B of the HASOC 2020 dataset, surpassing previous SoTA scores of 51.52% and 26.52%, respectively. We achieve near-SoTA results on two others—macro F1 scores of 81.66% for subtask A of the OLID 2019 and 82.54% for subtask A of the HASOC 2021, in comparison to SoTA results of 82.9% and 83.05%, respectively. We perform error analysis and use two eXplainable Artificial Intelligence (XAI) algorithms (Integrated Gradient (IG) and SHapley Additive exPlanations (SHAP)) to reveal how two of the models (Bi-Directional Long Short-Term Memory Network (Bi-LSTM) and Text-to-Text-Transfer Transformer (T5)) make the predictions they do by using examples. Other contributions of this work are: (1) the introduction of a simple, novel mechanism for correcting Out-of-Class (OoC) predictions in T5, (2) a detailed description of the data augmentation methods, and (3) the revelation of the poor data annotations in the HASOC 2021 dataset by using several examples and XAI (buttressing the need for better quality control). We publicly release our model checkpoints and codes to foster transparency.

Keywords: hate speech; NLP; T5; LSTM; RoBERTa

Citation: Adewumi, T.; Sabry, S.S.; Abid, N.; Liwicki, F.; Liwicki, M. T5 for Hate Speech, Augmented Data, and Ensemble. *Sci* **2023**, *5*, 37. https://doi.org/10.3390/sci5040037

Academic Editors: Carson K. Leung, Haridimos Kondylakis

Received: 5 July 2023
Revised: 27 August 2023
Accepted: 20 September 2023
Published: 22 September 2023

1. Introduction

Any unethical disparaging remark or expression targeted at an individual or group, based on identities such as race, sexual orientation, religion, or similar factors, is usually considered Hate Speech (HS) [1–5]. It is motivated by hatred or bias and aims to offend the target. This work considerably extends the authors' work in [6]. Manual detection of HS content is a tedious task that can result in delays in stopping harmful behavior (bbc.com/news/world-europe-35105003, accessed on 5 February 2023). Automatic hate speech detection is, therefore, crucial and has been gaining increasing importance because of the rising influence of social media. It will facilitate the elimination/prevention of undesirable characteristics in data and, by extension, AI technologies, such as conversational systems [7,8]. HS examples that may incite others to violence in the Offensive Language Identification Dataset (OLID) [9] are given in Table 1.

The datasets in this work were selected based on the important subtasks covered with regard to HS or abusive language. The architectures employed include the Bi-Directional Long Short-Term Memory Network (Bi-LSTM), the Convolutional Neural Network (CNN), Robustly optimized BERT approach (RoBERTa)-Base, Text-to-Text-Transfer Transformer (T5)-Base, where the last two are pre-trained models from the HuggingFace hub. As the best-performing baseline model, T5-Base is then used on the augmented data for the HASOC 2021 and for an ensemble. In addition, we compare results from HateBERT, a re-trained BERT model for abusive language detection [10]. We release our codes publicly (github.com/LTU-Machine-Learning/hatespeech_t5, accessed on 5 July 2023).

Table 1. Inciteful examples from the OLID 2019 training set (parts of offensive words masked with "*").

id	Tweet
23352	@USER Antifa simply wants us to k*ll them. By the way. Most of us carry a back up. And a knife
61110	@USER @USER Her life is crappy because she is crappy. And she's threatening to k*ll everyone. Another nut job. . . Listen up FBI!
68130	@USER @USER @USER @USER @USER Yes usually in THOSE countries people k*ll gays cuz religion advise them to do it and try to point this out and antifa will beat you. No matter how u try in america to help gay in those countries it will have no effect cuz those ppl hate america.

The rest of this paper is structured as follows: Section 2 provides an overview of HS and prior work in the field. Section 3 explains the methods used in this study. The results, critical analysis with XAI, and discussion are in Section 4. Section 5 gives the conclusion and possible future work.

2. Related Work

Significant efforts have gone into addressing automatic HS detection [11,12]. Zampieri et al. [9] extended the OLID dataset to annotate the distinction between explicit and implicit messages. This effort was important given the many forms through which HS can be expressed [5]. Caselli et al. [13] performed cross-domain experiments on HatEval [14]. Mutanga et al. [15] experimented with different Transformer-based architectures using only the HSO dataset. However, their preprocessing approach, which involves removing low-frequency words, may result in newly introduced hate terms escaping detection.

The Transformer architecture by Vaswani et al. [16] has been very influential in recent progress with various NLP tasks. The attention mechanism on which it is based makes it possible for it to handle long-term dependencies [16,17]. Hence, Transformer-based models have gained increased attention in HS detection and classification [12,15,18,19]. Table 2 compares some of the methods that have been employed for automatic HS detection in the literature. Despite the introduction of these models, there seems to be a gap where recent SoTA models are not compared across many HS datasets. We address that in this work.

Table 2. Comparison of some methods for automatic HS detection. Some of these methods are also used in a hybrid fashion [20].

Method	Description with Pros/Cons
Bag of Words	This considers the multiplicity of words in a document and is usually applied with some classical ML methods [21]. It does not consider word order or context; hence, it is less effective than modern approaches.
Lexicon-based	This is based on a list of keywords that are identified as hateful or offensive [22]. The lack of context makes this approach more prone to false positives.
Term Frequency–Inverse Document Frequency (TF-IDF)	TF-IDF is based on the importance of a word or term to a document [20]. Therefore, it gives less frequently used words greater weight compared to common words, e.g., *the*. It is commonly used together with classical Machine Learning (ML) approaches, such as Random Forest, Decision Trees, Logistic Regression, etc. [21]. It is less effective than modern ML approaches because it does not consider the context of words in a document.
Recurrent Neural Network (RNN)	Neural networks based on sequential recurring time-steps are used in automatic HS detection [23,24]. An improved variant of the RNN is the Bi-LSTM. The vanishing gradient problem and the lack of parallelization in RNNs have placed limitations on their performance.
Convolutional Neural Network (CNN)	CNNs are typically used in computer vision and are a deep learning approach. They employ filters and pooling [24–26]. The architectural design of CNNs seems to make them less efficient for text than the Transformer [16].
Bidirectional Encoder Representations from Transformers (BERT) [27]	BERT has been employed in automatic HS detection in various forms, including hybrid methods [24,28]. Variants of BERT include RoBERTa [29], ALBERT [30], and DeBERTa [31]. These are deep learning models that generate deep contextual representations. They use only the encoder stacks of the Transformer architecture [16], unlike the T5 [32], which uses both encoder and decoder stacks.

3. Materials and Methods

All of the experiments were conducted on a shared DGX-1 machine with 8 × 32 GB Nvidia V100 GPUs. The Operating System (OS) of the server is Ubuntu 18, and it has 80 CPU cores. Each experiment is conducted 3 times and the average results are reported. Only limited hyperparameters are explored, through manual tuning through a limited grid search (without the use of automatic tools for exploration) due to resource constraints. Hence, six is the total number of epochs for each experiment, as this is a good tradeoff, especially for pre-trained models [33], and we noticed overfitting for higher epochs in the pilot study. Moreover, the model checkpoint with the lowest validation loss is saved and used for evaluation of the test set, where available. A linear schedule with a warm-up is used for the Learning Rate (LR) adjustment for T5 and RoBERTa. The average time per epoch for training and evaluation on the validation set is 83.52, 7.82, and 22.29 s for the OLID, HASOC 2020, and HASOC 2021 datasets, respectively (Restrictions (*cpulimit*) were implemented to avoid server overloading, in fairness to other users. Hence, the average time for the test sets ranges from 2 to over 24 h).

We report both weighted and macro F1 scores because of past studies. The F1 score is the harmonic mean of the precision and recall. The relative contribution to the F1 from precision and recall are equal. Macro-F1 does not take label imbalance into account, unlike weighted-F1, which accounts for label imbalance by finding the labels' mean weighted by support (each label's true instance) [34] (scikit-learn.org/../generated/sklearn.metrics.f1_score.html, accessed on 7 January 2023).

3.1. Preprocessing

We carried out automatic preprocessing on all the data to remove duplicates and unwanted strings or characters. In some of the datasets, such as OLID (task C), there are "nans" (empty entries) in some columns of the labels. These cause problems for the models by dropping model performance. We, therefore, dropped such rows during the preprocessing step. To prepare the text for the models, the following standard preprocessing steps are applied to all the datasets:

- URLs are removed.
- Emails are removed.
- IP addresses are removed.
- Numbers are removed.
- All characters are changed to lowercase.
- Excess spaces are removed.
- Special characters, such as hashtags(#) and mention symbols (@), are removed.

3.2. Data

The following are the datasets considered in this work:

1. HASOC 2020

The English dataset is composed of 3708 tweets for training and 1592 for testing. The dataset includes the following subtasks: (1) task_1 (A), which identifies hate and offensive text, and (2) task_2 (B), which is a further classification for the previous task to categorize the hateful and offensive content into either hate content (HATE), offensive (OFFN), or profane (PRFN). Mandl et al. [35] collected the dataset and used a trained SVM classifier and human judgment to label the data.

2. HASOC 2021

This third edition of HASOC Mandl et al. [36] provided another set of tweets dataset with the same subtasks as HASOC 2020. The English dataset consists of 3843 training samples and 1281 samples in the test set. The dataset has COVID-related topics since the data were gathered during the COVID-19 pandemic. A total of 10% of the training set is split as the dev set in this work for evaluation after each epoch.

3. HatEval 2019

Basile et al. [14] prepared this dataset of tweets to detect hateful content against women and immigrants. It contains 13,000 English tweets, distributed as 9000 for training, 1000 for development, and 3000 for testing. The dataset includes two subtasks: subtask A identifies the presence of hate speech, and subtask B is the average of three binary classification tasks. The 3 binary subtasks under subtask B include (1) HS, (2) whether the hate speech targets a group of people or an individual (TR), and whether the HS contains aggressive content or not (AG).

4. OLID 2019

The SemEval 2019 shared task 6 dataset is based on the OLID dataset. It has 14,200 annotated English tweets and encompasses the following three sub-tasks: (a) offensive language detection, (b) categorization of offensive language as to whether it is targeted at someone (or a group) or not, and (c) offensive language target identification, where distinctions are made between the individual, group, and other entities, such as an organization [9]. Crowd-workers performed its data annotation, and the original data-split was into training and test sets only. As we did with HASOC 2021, we split 10% of the training set as the dev set for evaluation after each epoch.

5. Hate Speech and Offensive HSO

Davidson et al. [11] gathered tweets based on a hate speech lexicon and employed crowd-sourcing efforts to annotate them. They make a distinction between hate speech and offensive language, choosing a narrower definition of hate speech, as opposed to some general views, such as that of Zampieri et al. [9]. Three categories are present in the labeled data: hate speech, only offensive language, and neither. Of the 24,802 labeled tweets, resulting in the HSO data, 5% were labeled as containing hate speech, while 1.3% were by unanimous decision.

6. Trolling, Aggression, and Cyberbullying (TRAC)

Kumar et al. [37] introduced TRAC. The second version, in 2020, contains two subtasks. Three categories are present in the first subtask: Overtly Aggressive, Covertly Aggressive, and Non-aggressive. The English version of this task contains 5000 samples for training and evaluation, which is the same as the Bangla and Hindi versions. The second subtask is a binary classification to identify gendered or non-gendered text. Our focus was on the first subtask only in this work. Elsafoury et al. [38] distinguished this dataset from other HS datasets. However, they also acknowledged that there are some similarities (such as abusive language) between aggression and HS. It is based on this that we selected the dataset.

3.3. Models

1. The Bi-LSTM is one form of Recurrent Neural Network (RNN) [39]. It is an improved variant of the vanilla RNN. Its input text flows forward and backward, providing more contextual information, and thereby improving the network performance [40]. We used 2 bi-directional layers and pre-trained Glove [41] word embeddings of 100 dimensions. We also applied a dropout layer to prevent overfitting. This model has 1,317,721 parameters. Word and subword embeddings have been shown to improve the performance of downstream tasks [41–43].
2. The CNN is common in computer vision or image processing. The author of [44] shows the effectiveness of CNNs in capturing text local patterns on different NLP tasks. Both the Bi-LSTM and CNN architectures are used as feature-based models, where for each tweet, we computed embeddings using the pre-trained Glove, before using the embeddings as an input to the baseline model. The CNN model is composed of 3 convolution layers with 100 filters each. The filter size for the first layer is 2×100, the filter size for the second layer is 3×100, and the filter size is 4×100 for the third layer. We use the ReLU activation function and max-pooling after each convolution

layer. We perform dropout for regularization. The total trainable parameters for the CNN are 1,386,201.

3. RoBERTa is based on the replication study of BERT. It differs from BERT in the following ways: (1) training for longer over more data, (2) removing the next sentence prediction objective, and (3) using longer sequences for training [29]. The base version of the model, which we use, has 12 layers and 110 M parameters. For our study, we use a batch size of 32, an initial learning rate of 1×10^{-5}, and a maximum sequence length of 256. We restricted the number of tasks to only binary tasks for this model.
4. The T5 [32] is based on the transformer architecture by Vaswani et al. [16]. However, a different layer normalization is applied, in which there is no additive bias applied, and the activations are only rescaled. Causal or autoregressive self-attention is used in the decoder for it to attend to past outputs. The T5-Base model has about twice the number of parameters as that of BERT-Base. It has 220 M parameters and 12 layers each in the encoder and decoder blocks, while the smaller version has 60 M parameters [32]. The T5 training method uses teacher forcing (i.e., standard maximum likelihood) and a cross-entropy loss. T5-Base required more memory and would not fit on a single V100 GPU for the batch size of 64; hence, we lowered the batch size to 16 but kept the batch size at 64 for T5-Small. The task prefix we use is 'classification' for all the tasks, as the model takes a hyperparameter called a task prefix.

3.4. Solving OoC Predictions in T5

The OoC intervention is a by-product of this work since it is not the main motivation. Raffel et al. [32] introduced T5 and noted the possibility of OoC predictions in the model. This is when the model predicts text (or empty string) seen during training but is not among the class labels. This issue appears to be more common in the initial epochs of training and may not occur at all sometimes. In order to solve this, first we introduced integers (explicitly type-cast as string) for class labels, which appeared to make the model predictions more stable. The issue is reduced by about 50% in pilot studies when it occurred. For example, for the HASOC datasets, we substituted "1" and "0" for the labels "NOT" and "HOF", respectively. As a second step, a simple correction we introduced is to replace the OoC prediction (if it occurs) with the label of the largest class in the training set.

3.5. Data Augmentation

The objective of data augmentation is to increase the number of training data samples in order to improve the performance of models on the evaluation set [45]. We experimented with 2 techniques: (1) word-level deletion of the start and end words per sample and (2) conversational AI text generation (Table 3). Our work may be the first to use conversational AI for data augmentation in automatic HS detection. It doubles the number of samples and provides diversity. The average number of new words generated per sample prompt is around 16 words.

The first technique involves the use of the list of offensive words available from an academic online resource (cs.cmu.edu/~biglou/resources/, accessed on 5 January 2022). This list is used to ensure offensive tokens are not deleted during the pass through the training set. From the original list of 1383 English words, we removed 160 words that we decided may not qualify as offensive words because they are nationalities/geographical locations or adjectives of emotions. A total of 1223 words are left in the document used for our experiment. Examples of words removed are: *European, African, Canadian, American, Arab, angry* , and many other unharmful words. Samples ending or starting with offensive words are kept as they are in the new augmented training data and, therefore, are dropped when merged with the original, to avoid duplicates.

The second technique involves the use of the dialogue (conversation) model checkpoint by Adewumi et al. [8], which was finetuned on the Multi-Domain Wizard-of-Oz (MultiWOZ) dataset by Eric et al. [46]. It is an autoregressive model based on the pretrained DialoGPT-medium model by Zhang et al. [7]. An autoregressive model conditions

each output word on previously generated outputs sequentially [47]. Each sample from the training data is used as a prompt for the model to generate a response, which is then concatenated to the prompt to form a new version of the prompt that was supplied. This ensures the original label of each sample is unchanged, as the offensive content, if any, is still retained in the new sample. As demonstrated by Adewumi [48], the possibility of generating an offensive token is small for this model because the MultiWOZ dataset it is trained on is reputed to be non-toxic.

This second technique literally doubled the training set size of the HASOC 2021 dataset. A random quality inspection was carried out on the augmented data. Examples of the original and augmented samples from the HASOC 2021 dataset, using this second technique, are given in Table 3. The offensive words (masked with ***) are retained in the new samples. The top p and k variables of the decoding algorithm for the model were set as $p = 0.7$ and k = 100, respectively. Additional hyperparameters include *maximum decoding length*, set to 200 tokens; *temperature*, set to 0.8; and *maximum ngram repeat limit*, set to 3. These hyperparameters are based on previous work, as they have been shown to perform well [8].

Table 3. Original and conversational AI-augmented examples from the HASOC 2021 dataset (offensive words masked with "***").

Type	Sample
original	Son of a *** wrong "you're"
augmented	son of a *** wrong youre No, that's Saint Johns Chop House. I need a taxi to take me from the hotel to the restaurant, leaving the first at 5:45.
original	SO EXCITED TO GET MY CovidVaccine I hate you covid!
augmented	so excited to get my covidvaccine i hate you covid You should probably get that checked out by a gastroenterology department.
original	ModiKaVaccineJumla Who is responsible for oxygen? ModiResign Do you agree with me? Don't you agree with me?
augmented	modikavaccinejumla who is responsible for oxygen modiresign do you agree with me dont you agree with me Yes, I definitely do not want to work with them again. I appreciate your help..

3.6. The Ensemble

The ensemble is a majority-voting system comprising theT5-Base, T5-Small, and RoBERTa-Base models. The saved model checkpoint from each trained model is used to make a prediction on each sample of the test set of the HASOC 2021 dataset for subtask A. The prediction ("HOF" or "NOT") with more than one vote (2 or 3) is recorded as the prediction for that sample. The weighted- and macro-F1 scores are then calculated with the *scikit-learn* library [34], as in all other cases. The computational effort required for the ensemble is quite a lot, requiring more time for evaluation. This is typical of ensembles.

4. Results

Tables 4 and 5 show baseline results and additional results using the best model (T5). More results using cross-tasks can be found in Table A1 (in Appendix B). Table A2 (in Appendix B), shows results for other datasets and the HateBERT model [10]. The HatEval task is the only comparable one in our work, as is that by Caselli et al. [10].

The Transformer-based models (T5 and RoBERTa) generally perform better than the other baselines (LSTM and CNN) [49], except for RoBERTa on the OLID subtask B and HASOC 2021 subtask A. T5 outperforms RoBERTa on all tasks. Based on the test set results, the LSTM obtains better results than the CNN in the OLID subtasks A, HASOC 2020 subtask A, and HASOC 2021 subtask A, while the CNN does better than it on the others. The T5-Base model achieves new best scores on the HASOC 2020 subtasks. The augmented data, using the conversational AI technique, improve the results on HASOC 2021 (The first technique is not reported because there was no improvement. This may be because

the number of total samples is smaller than that of the conversational AI technique). This challenge of the relatively high false negative (160) in Figure 1 is not unique to the T5 model alone and is even more pronounced in the other models. A possible way to reduce it is by oversampling to balance the dataset before training. In spam (or similar) detection, users are usually given the opportunity to mediate because of such imperfections.

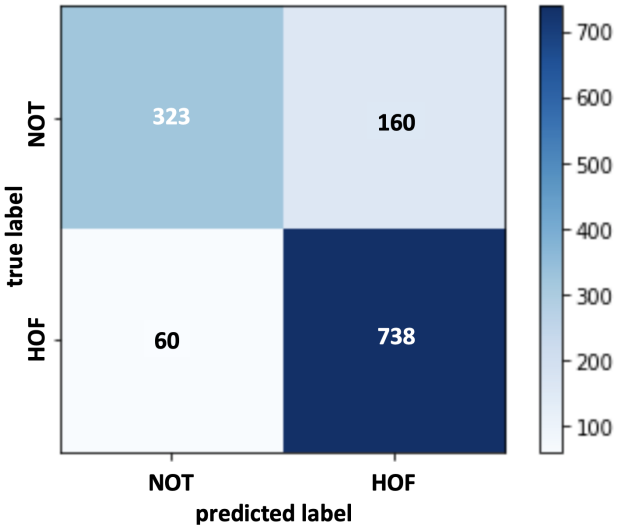

Figure 1. Confusion/error matrix of T5 on Hasoc 2021 test set [6].

Table 4. Mean scores of model baselines for different subtasks [6] (sd: standard deviation; bold values are best scores for a given task; '-' implies no information available).

Task	Weighted F1 (%)		Macro F1(%)	
	Dev (sd)	Test (sd)	Dev (sd)	Test (sd)
CNN				
OLID A	79.10 (0.26)	82.47 (0.56)	77.61 (0.39)	78.46 (0)
OLID B	82.43 (0.49)	83.46 (0)	46.76 (0)	47.88 (0)
OLID C	47.54 (1.36)	38.09 (3.91)	35.65 (0)	36.85 (0)
Hasoc 2021 A	77.22 (0.52)	77.63 (0.70)	74.28 (0.58)	75.67 (0)
Hasoc 2021 B	55.60 (0.61)	59.84 (0.41)	50.41 (0.41)	54.99 (0)
Bi-LSTM				
OLID A	79.59 (0.89)	83.89 (0.57)	78.48 (1.52)	79.49 (0)
OLID B	82.50 (1.70)	83.46 (0)	46.76 (0)	47.32 (0)
OLID C	49.75 (3.95)	43.82 (9.63)	35.65 (2.81)	36.82 (0)
Hasoc 2021 A	78.05 (0.85)	78.43 (0.84)	77.99 (1.79)	77.19 (0)
Hasoc 2021 B	50.65 (1.34)	52.19 (1.95)	43.19 (2.09)	42.25 (0)
RoBERTa				
OLID A	82.70 (0.55)	84.62 (0)	80.51 (0.76)	80.34 (0)
OLID B	82.70 (0.13)	83.46 (0)	46.76 (0.04)	47.02 (0)
Hasoc 2021 A	79.9 (0.57)	76.2 (0)	77.77 (0.75)	74 (0)
T5-Base				
OLID A	92.90 (1.37)	85.57 (0)	92.93 (1.42)	81.66 (0)
OLID B	99.75 (0.43)	86.81 (0)	99.77 (0.44)	53.78 (0)
OLID C	58.35 (1.22)	54.99 (0)	33.09 (0.76)	43.12 (0)
Hasoc 2021 A	94.60 (1.98)	82.3 (0)	94.73 (5.26)	80.81 (0)
Hasoc 2021 B	65.40 (0.82)	62.74 (0)	62.43 (6.32)	59.21 (0)
[49] best scores				
OLID A				**82.90** (-)
OLID B				**75.50** (-)
OLID C				**66** (-)

Table 5. T5 variants' mean scores over HASOC data. Augmented data improves accuracy (compare T5 baseline on HASOC 2021 in Table 4 (sd: standard deviation; bold values are best scores for a given task; '-' implies no information available; the ensemble result is restricted to task A due to computational constraints).

Task	Weighted F1 (%)		Macro F1(%)	
	Dev (sd)	Test (sd)	Dev (sd)	Test (sd)
T5-Base				
Hasoc 2020 A	96.77 (0.54)	91.12 (0.2)	96.76 (0.54)	**91.12** (0.2)
Hasoc 2020 B	83.36 (1.59)	79.08 (1.15)	56.38 (5.09)	**53.21** (2.87)
T5-Base+Augmented Data				
Hasoc 2021 A	95.5 (3.27)	83 (0)	92.97 (2.20)	82.54 (0)
Hasoc 2021 B	64.74 (3.84)	66.85 (0)	65.56 (1.48)	62.71 (0)
Ensemble				
Hasoc 2021 A		80.78 (0)		79.05 (0)
[35] best scores				
Hasoc 2020 A				51.52 (-)
Hasoc 2020 B				26.52 (-)
[36] best scores				
Hasoc 2021 A				**83.05** (-)
Hasoc 2021 B				**66.57** (-)

The ensemble macro-F1 result (79.05%) is closer to the T5-Base result (80.81%) and farther from the RoBERTa result (74%). The deciding factor is the T5-Small. Hence, a voting ensemble may not perform better than the strongest model in the collection if the other models are weaker at prediction.

4.1. Cross-Task Training

We obtained a new SoTA result (91.73%) for the HASOC 2020 subtask A after initial training on the OLID subtask A. The reason we outperform the previous SoTA result is that they used an LSTM with Glove embeddings [35], instead of a pre-trained deep model with the attention mechanism [17] that gives a transfer learning advantage. The p-value ($p < 0.0001$) obtained for the difference of the two means of the two-sample t-test is smaller than the alpha (0.05), showing that the results are statistically significant.

4.2. HASOC 2021 Annotation Issues

An inspection of some of the samples predicted by the T5 model reveals challenges with the quality of data annotation in the HASOC 2021 dataset. The model was intelligent enough to detect these. Table 6 gives several (10) examples of tweets incorrectly labeled as "NOT" ('1') by the annotators but which are clearly offensive (HOF ('0')), in our view, and are also correctly predicted as such by the model. More cases similar to these exist within the dataset than shown in the table. This issue makes a strong case for having better Quality Control (QC) with data annotation, given the possible implications, including the poor assessments that may result from the competitions organized using such a dataset. We provide SHAP explanations of the T5 model predictions for some of these suspicious examples (Appendix B.3).

Table 6. Hasoc 2021 test set samples with suspicious labels but correctly classified, in our view, as HOF by T5 (parts of offensive words masked with "*").

_id	Text	Task_1	Task_2
60c5d6bf5659ea5e55def8b3	When you're the main b*tch https://t.co//HWlNpEUiwS, accessed on 7 January 2023.	NOT	NONE
60c5d6bf5659ea5e55df0242	miya four creeps into every thought i have what the f*ck	NOT	NONE
60c5d6bf5659ea5e55defe58	At least we're being freed from the shambles of the evangelical, but d*mn y'all couldn't just go w the flow? Y'all just had to find jebus through crystals and astrology? Smdh	NOT	NONE
60c5d6bf5659ea5e55defe24	@kumarmbayar @Actor_Siddharth He is mentally Upset, Why he's not talking about #BengalBurning and why stupid talks about Tejasvi_Surya	NOT	NONE
60c5d6bf5659ea5e55df0210	@amemoott @Ripple @UCT_news @qz @coinbase @cz_binance låve been holding XRP for years. On Coinbase and binance . If you not going to help shut the f*ck up	NOT	NONE
60c5d6bf5659ea5e55def288	Like this tweet for a DM SC: jayjayson4225 #dmme #horny #hornydm #nsfwtwitter #nsfw #twitterafterdark #whitedick #whitecock #cockrate #nudes #naughtydm #dmsopen #bwc #cock #dick #nsfwtwt #porn #sex #dickpic #dickpics #cumslut #cum #slut #whore #hotwife	NOT	NONE
60c5d6bf5659ea5e55defdc1	happy p*ss off old people month!	NOT	NONE
60c5d6bf5659ea5e55defc79	#China is such a worst country first they unleashed #Chinesevirus on whole world and #ChinaLiedPeopleDied and now india is struggling with #Chinesevirus but they are mocking people who are dieing of #ChinesVirus in India https:/t.coV0AiuJV3lm	NOT	NONE
60c5d6bf5659ea5e55df01aa	@globaltimesnews Communist must feel shame God is watching you what you did with world. #CommunistVirus	NOT	NONE
60c5d6bf5659ea5e55def750	@RapidWolfReturn @Utd_76 @MenachoManuel1 @ITheKTrainI @UnitedStandMUFC Yeah. . . sh*t move, but as has been said; heat of the moment, stupid comment he probably doesn't really back.. should've just explained it, owned it, and moved on. He's a w*nker, regar	NOT	NONE

4.3. Error Analysis

The confusion matrix for the T5 on HASOC 2021 is given in Figure 1. It reveals that 33% (160) of the "NOT" class (not offensive) was misclassified as offensive while only 8% (60) of the "HOF" (hate or offensive) was misclassified as "NOT". The higher percentage of misclassification for the "NOT" class is very likely due to the fact that the training set is imbalanced, as there are more "HOF" samples (2251) compared to "NOT" (1207). Hence, the model is better at identifying samples of "HOF". Correction of the imbalance in the dataset through oversampling before training can help to improve performance.

4.4. Explainable Artificial Intelligence (XAI)

XAI helps us understand how a model arrives at a prediction and identify any incompleteness in the model [50]. This can add to the justification for using ML models and the trust in their predictions. In this study, rather than compare two XAI algorithms on one model, we focus on separate explanations from two XAI algorithms on two different models, using the same examples from the HASOC 2021 test set subtask A (Table A3). The XAI algorithms are Integrated Gradient (IG) and SHapley Additive exPlanations (SHAP).

We apply IG to the Bi-LSTM. It is an attribution method that is based on two fundamental axioms— Sensitivity and Implementation Invariance [51]. Generally, integrated gradients aggregate the gradients along the straight line between the baseline and the input. A good baseline (of a zero-input embedding vector, in this case) is very important. Models trained using gradient descent are differentiable, and IG can be applied to these. IG has the advantage of being relatively faster than SHAP computationally. Appendix B.2, in the appendix, shows IG explanations for examples of five correctly classified (Figure A11) and five incorrectly classified samples (Figure A12), based on the provided annotations. The attribution shows which input words affect the model prediction and how strongly. Important words are highlighted in shades of green or red, such that words in green contribute to non-hate speech, while those in red contribute to hate speech. In Figure A12, the second

tweet has what may be considered an offensive word, but it is incorrectly annotated as "NOT". The Bi-LSTM, however, predicts this correctly.

SHAP assigns each feature an importance value for a particular prediction [52]. The exact computation of SHAP values is challenging. However, by combining insights from current additive feature attribution methods, one can approximate them. Its novel components include: (1) the identification of a new class of additive feature importance measures and (2) theoretical results showing there is a unique solution in this class with a set of desirable properties [52]. It unifies six existing methods: LIME, DeepLIFT, Layer-Wise Relevance Propagation, Shapley regression values, Shapley sampling values, and Quantitative Input Influence. The last three are based on classic cooperative game theory [53]. This provides an improved performance and consistency with human intuition. SHAP has the advantage that it can be applied to models whose training algorithm is differentiable as well as those based on non-differentiable algorithms, such as trees.

SHAP functionality is employed in this work by passing the supported HuggingFace Transformers' T5 pipeline (*text2text-generation*) to SHAP. Important words or subwords are highlighted in shades of red or blue, such that words in red are those that contribute to a resulting prediction, while those in blue contribute to what would be an alternative prediction. The thicker the shade, the stronger the contribution, as also indicated by the real values above each word or subword. Figures A1–A5 present examples using the same samples from Table A3. Additional examples are provided in Appendix B.1, in the appendix. We observe that 7 out of the 10 are correctly predicted by the T5, as explained by SHAP, compared to the five correct predictions by the Bi-LSTM.

5. Conclusions

In this study, we address the gap in which recent SoTA models are not compared across many HS datasets and demonstrate the benefits of synthetic data augmentation through conversational AI and the possibilities with an ensemble for automatic HS detection. We achieved new SoTA results on the HASOC 2020 subtasks A and B. We also achieved near-SoTA results for both the subtask A of the OLID 2019 and HASOC 2021 datasets. As a by-product, we solve the OoC problem in T5 using a simple two-step approach. We reveal, with examples and XAI, the shortcomings of the HASOC 2021 dataset and make a case for better quality control with data annotation. IG and SHAP are also used to explain the predictions of some of the same examples from the HASOC 2021 dataset. Future work that compares performance with models, which are pre-trained on large volumes of tweets, such as BERTweet [54], may be worth investigating. Releasing our source codes and model checkpoints provides the opportunity for the community to reproduce our results and foster transparency.

Limitations

The datasets used in this study are all in the English language. The results are, therefore, limited to the English language. It is unclear how the models will perform with other languages. Many of the datasets are also based on tweets, which are usually short. Hence, there might be low scalability of the models to long text. Furthermore, none of the models has 100% performance on the short tweets. Moreover, all of the models were trained on GPU, and this requirement is necessary to train the models to speed up training time.

Author Contributions: Conceptualization, T.A. and S.S.S.; methodology, T.A.; writing—original draft preparation, T.A.; writing—review and editing, S.S.S., N.A., F.L. and M.L.; supervision, F.L. and M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1. SHAP Explanations

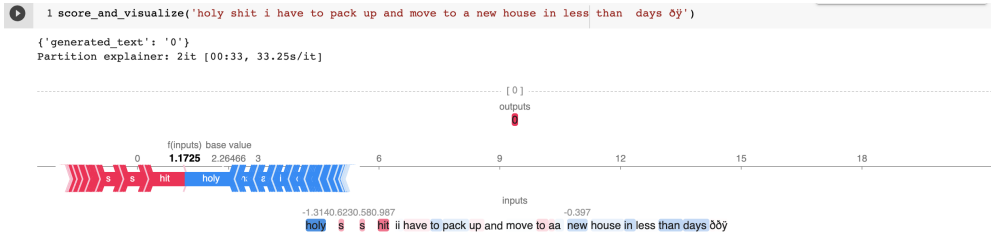

Figure A1. SHAP explanation of the T5 model prediction.

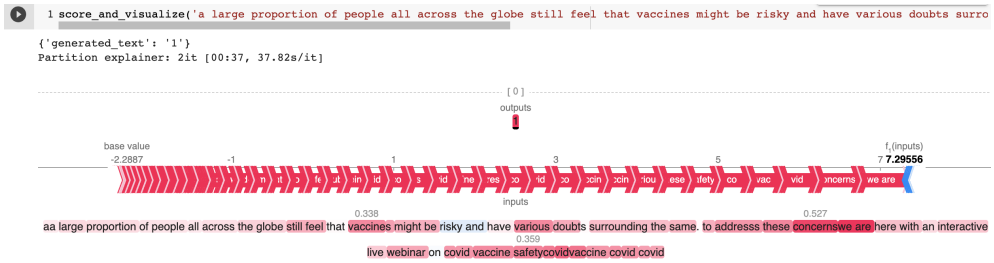

Figure A2. SHAP explanation of the T5 model prediction.

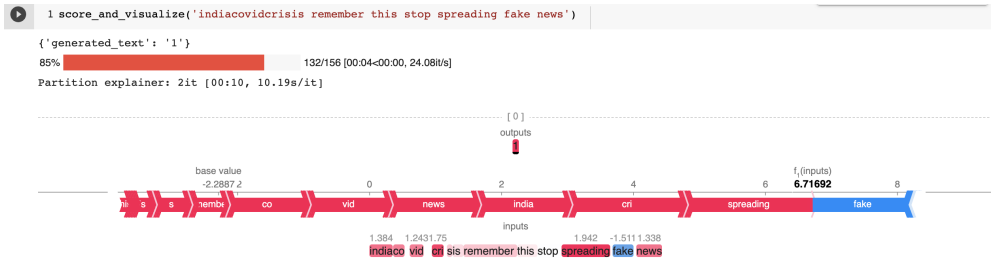

Figure A3. SHAP explanation of the T5 model prediction.

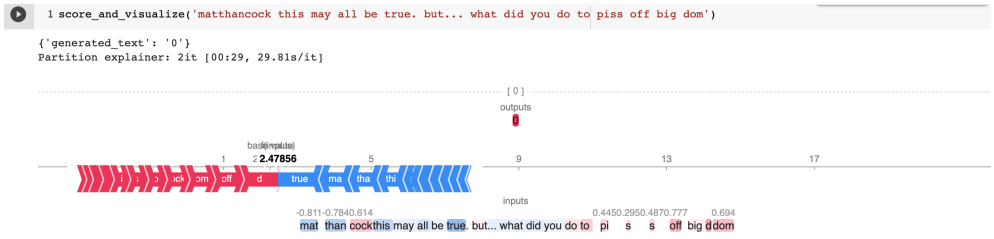

Figure A4. SHAP explanation of the T5 model prediction.

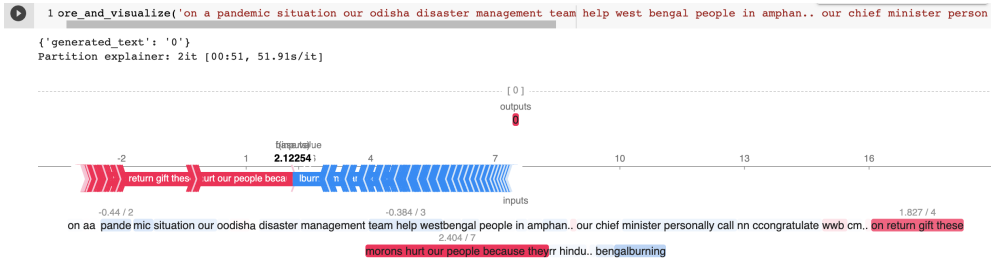

Figure A5. SHAP explanation of the T5 model prediction.

Appendix A.2. Cross-Task Training

We performed cross-task training to ascertain whether there will be performance gains on a target subtask. We discovered that cross-task training can improve performance, however, not always. Six subtasks from three datasets are selected for this purpose due

to resource constraints. The subtasks are all subtasks A (binary classification) from three datasets: OLID, HASOC 2020, and HASOC 2021. Only the T5 model is used for these experiments. We finetune an initial (source) subtask and then further finetune it on a final (target) subtask of a different dataset before evaluating the test set of the target subtask.

Appendix B

Table A1. Cross-task inference using T5.

Cross-Task	Weighted F1 (%)		Macro F1 (%)	
	Dev (sd)	Test (sd)	Dev (sd)	Test (sd)
Hasoc 2020 A -> OLID A	90.35 (0.01)	83.94 (0.72)	88.82 (0.01)	79.81 (0.85)
Hasoc 2021 A -> OLID A	91.82 (0.01)	83.52 (0.48)	90.57 (0.01)	79.22 (1.01)
Hasoc 2021 A -> Hasoc 2020 A	95.87 (0)	90.14 (0.85)	95.87 (0)	90.13 (0.85)
OLID A -> Hasoc 2020 A	96.59 (0.68)	91.73 (0.25)	96.58 (0.68)	91.73 (0.26)
OLID A -> Hasoc 2021 A	86.82 (0.01)	80.91 (0.53)	84.91 (0.02)	79.32 (0.55)
Hasoc 2020 A -> Hasoc 2021 A	87.2 (0.03)	81.75 (0.29)	87.37 (0.01)	80.4 (0.3)

Table A2. Model comparison of mean scores for other HS datasets. (sd: standard deviation; '-' implies no information available).

Task	Weighted F1 (%)		Macro F1(%)	
	Dev (sd)	Test (sd)	Dev (sd)	Test (sd)
Bi-LSTM				
HatEval SemEval 2019 A	-	72.38 (0.54)	-	72.12 (0.72)
HatEval SemEval 2019 B	-	77.74 (2.8)	-	73.11 (0.44)
Hasoc 2020 A	88.6 (0.15)	89.30 (0.15)	89.47 (1.47)	90.28 (0.20)
Hasoc 2020 B	75.80 (0.56)	74.39 (2.31)	42.99 (0.15)	42.97 (0.06)
HSO	90.19 (0.03)	-	68.77 (1.93)	-
Trolling, Aggression	68.69 (0.36)	-	36.00 (0.27)	-
CNN				
HatEval SemEval 2019 A	-	73.95 (0.64)	-	71.67 (0.43)
HatEval SemEval 2019 B	-	78.88 (0.55)	-	71.13 (0.43)
Hasoc 2020 A	88.06 (0.41)	89.76 (0.44)	88.21 (0.41)	90.08 (0.46)
Hasoc 2020 B	76.38 (0.63)	76.48 (0.61)	49.15 (1.25)	47.58 (0.85)
HSO	88.52 (0.62)	-	71.27 (0.74)	-
TRAC	71.01 (1.73)	-	40.24 (0.43)	-
T5-Base				
HatEval SemEval 2019 A	-	87.07 (4.81)	-	86.52 (5.11)
HatEval SemEval 2019 B	-	99.93 (0)	-	99.88 (0)
TRAC	80.84 (3.96)	-	56.97 (8.34)	-
HateBERT				
HatEval SemEval 2019 A	-	-	-	0.516 (0.007)

Table A3. Pre-processed examples from HASOC 2021 test set for XAI in Appendices B.1 and B.2.

_id	Text	Task_1	Task_2
60c5d6bf5659ea5e55def475	holy shit i have to pack up and move to a new house in less than days ðÿ	HOF	PRFN
60c5d6bf5659ea5e55df026b	a large proportion of people all across the globe still feel that vaccines might be risky and have various doubts surrounding the same. to addresss these concerns we are here with an interactive live webinar on covid vaccine safety covidvaccine covid covid	NOT	NONE
60c5d6bf5659ea5e55deff15	indiacovidcrisis remember this stop spreading fake news	NOT	NONE
60c5d6bf5659ea5e55defc3e	one thing which epitomise the level of incompetency of delhi amp centre is that both gvts failed to utilise existing healthcare infra built by earlier gvts to its capacity even failed to supply required oxygen this is sheer incompetency amp negligence. both should resign resignmodi	NOT	NONE

Table A3. *Cont.*

_id	Text	Task_1	Task_2
60c5d6bf5659ea5e55df028c	matthancock this may all be true. but…what did you do to piss off big dom	HOF	PRFN
60c5d6bf5659ea5e55defb7f	on a pandemic situation our odisha disaster management team help west bengal people in amphan.. our chief minister personally call n congratulate wb cm.. on return gift these morons hurt our people because they r hindu.. bengalburning	HOF	HATE
60c5d6bf5659ea5e55defca7	diovavl shit	NOT	NONE
60c5d6bf5659ea5e55def240	cancelthboardexams resign_pm_modi pmoindia because of your overconfidence and ignorance hundreds of indian citizens are dying everyday and now you are ignoring lakhs of students daily plea to cancel exam…cancelthboardexams	NOT	NONE
60c5d6bf5659ea5e55defa7d	china must be punished for unleashing the chinesevirus starting a biological war. ban and boycott everything sources from the animal country covidsecondwave	HOF	HATE
60c5d6bf5659ea5e55def5a2	globaltimesnews china is not at all a trustworthy nation. the epidemic caused by chinesevirus have wreaked havoc worldwide and not only in india. if china really wants to help it should accept its blunder of creating this chinesevirus and spreading it all over intentionally. boycottchina	HOF	HATE

Appendix B.1. Cherry-Picked Examples from the HASOC 2021 Test Set for T5 Explained by SHAP

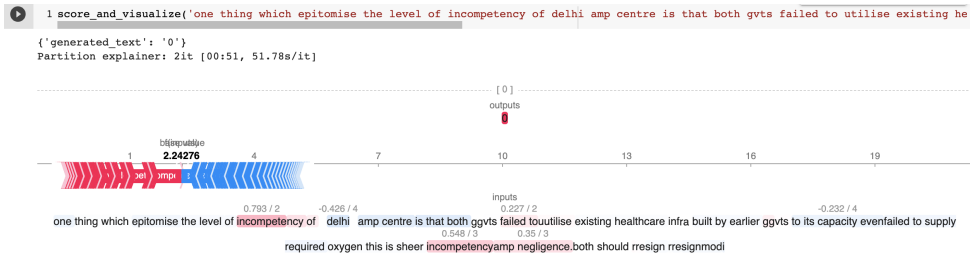

Figure A6. SHAP explanation of an incorrect T5 model prediction.

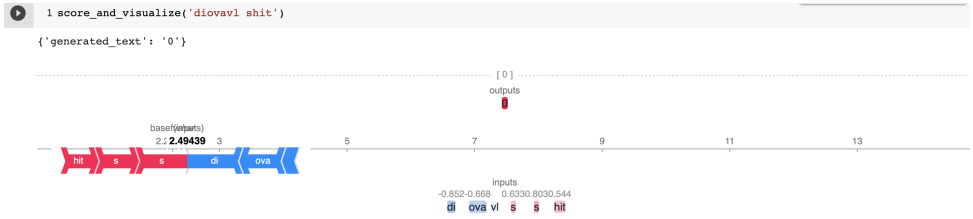

Figure A7. SHAP explanation of the T5 model prediction.

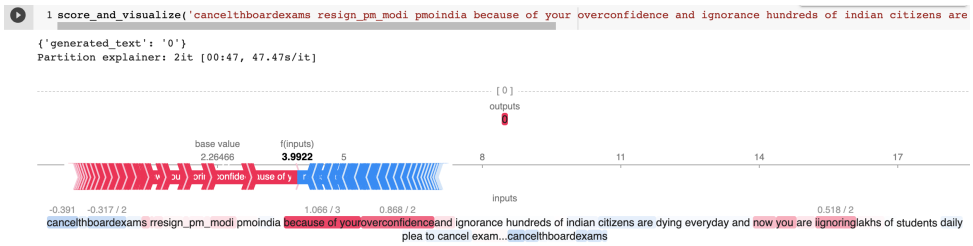

Figure A8. SHAP explanation of the T5 model prediction.

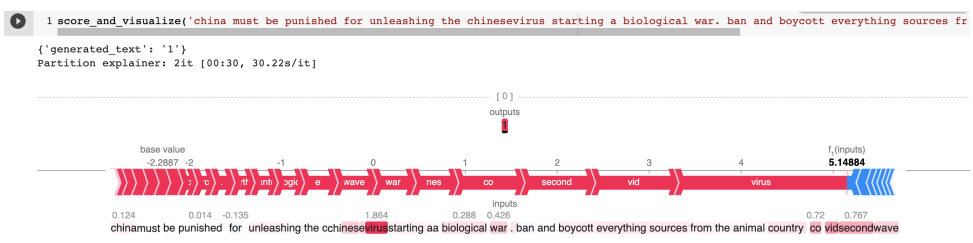

Figure A9. SHAP explanation of an incorrect T5 model prediction.

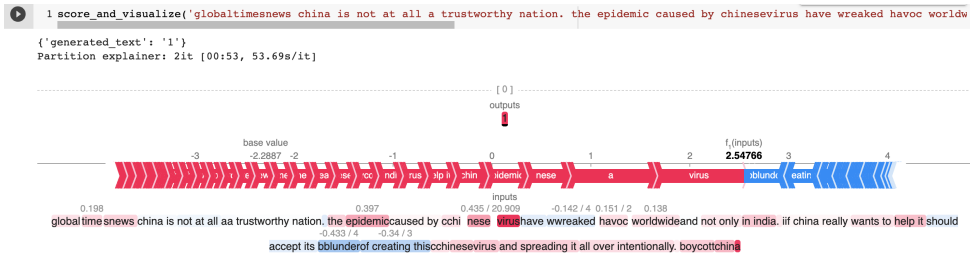

Figure A10. SHAP explanation of an incorrect T5 model prediction.

Appendix B.2. Cherry-Picked Examples from the HASOC 2021 Test Set for Bi-LSTM Explained by IG

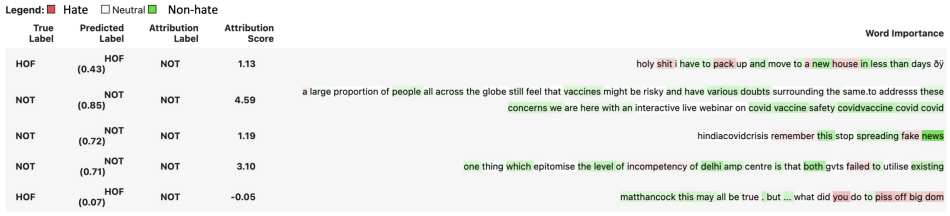

Figure A11. Visualize attributions for Bi-LSTM on HASOC 2021 test set (correct-classification).

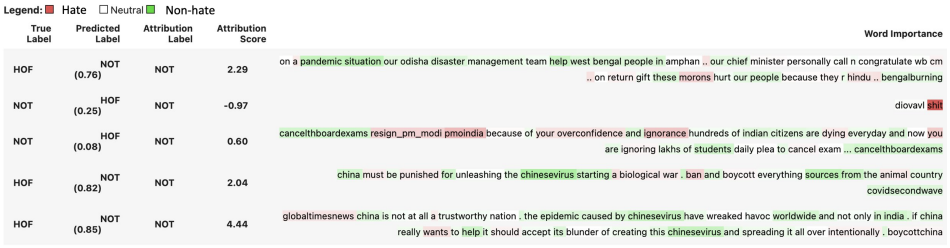

Figure A12. Visualize attributions for Bi-LSTM on Hasoc 2021 test set (misclassification).

Appendix B.3. Some Incorrect HASOC 2021 Annotations Correctly Classified by T5 and Explained by SHAP

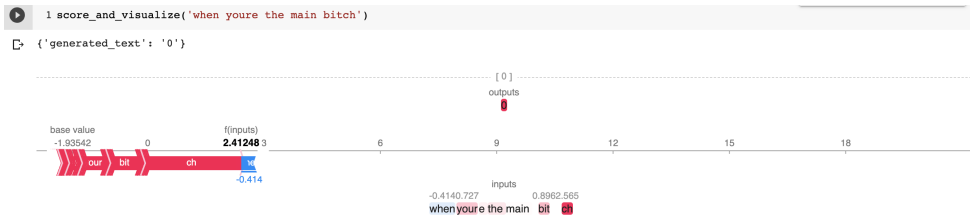

Figure A13. Incorrectly annotated but correctly classified by T5, as explained by SHAP.

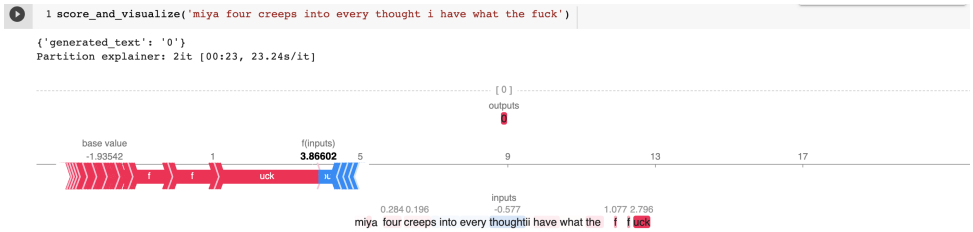

Figure A14. Incorrectly annotated but correctly classified by T5, as explained by SHAP.

References

1. Brison, S.J. The autonomy defense of free speech. *Ethics* **1998**, *108*, 312–339. [CrossRef]
2. Nockleby, J.T. Hate speech. *Encycl. Am. Const.* **2000**, *3*, 1277–1279.

3. Brown, A. What is hate speech? Part 1: The myth of hate. *Law Philos.* **2017**, *36*, 419–468. [CrossRef]
4. Quintel, T.; Ullrich, C. Self-regulation of fundamental rights? The EU Code of Conduct on Hate Speech, related initiatives and beyond. In *Fundamental Rights Protection Online*; Edward Elgar Publishing: Cheltenham, UK, 2020.
5. Anderson, L.; Barnes, M. Hate Speech. In *The Stanford Encyclopedia of Philosophy*, Spring 2022 ed.; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2022.
6. Sabry, S.S.; Adewumi, T.; Abid, N.; Kovács, G.; Liwicki, F.; Liwicki, M. HaT5: Hate Language Identification using Text-to-Text Transfer Transformer. In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 18–23 July 2022; pp. 1–7. [CrossRef]
7. Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; Dolan, B. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Online, 5–10 July 2020; pp. 270–278.
8. Adewumi, T.; Abid, N.; Pahlavan, M.; Brännvall, R.; Sabry, S.S.; Liwicki, F.; Liwicki, M. Smaprat: DialoGPT for Natural Language Generation of Swedish Dialogue by Transfer Learning. *arXiv* **2021**, arXiv:2110.06273.
9. 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 NAACL 2019, Minneapolis, MN, USA, 2–7 June 2019.
10. Caselli, T.; Basile, V.; Mitrović, J.; Granitzer, M. HateBERT: Retraining BERT for Abusive Language Detection in English. In Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), Online, 6 August 2021; pp. 17–25. [CrossRef]
11. Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. Automated hate speech detection and the problem of offensive language. In Proceedings of the International AAAI Conference on Web and Social Media, Montreal, QC, Canada, 15–18 May 2017; Volume 11.
12. Mathew, B.; Saha1, P.; Yimam, S.M.; Biemann, C.; Goyal1, P.; Mukherjee1, A. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In Proceedings of the 35th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence, Online, 2–9 February 2021.
13. Caselli, T.; Basile, V.; Mitrović, J.; Kartoziya, I.; Granitzer, M. 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, Marseille, France, 11–16 May 2020; pp. 6193–6202.
14. Basile, V.; Bosco, C.; Fersini, E.; Debora, N.; Patti, V.; Pardo, F.M.R.; 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, MN, USA, 6–7 June 2019; pp. 54–63.
15. Mutanga, R.T.; Naicker, N.; Olugbara, O.O. Hate speech detection in twitter using transformer methods. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 614–620. [CrossRef]
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
17. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proceedings of the International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
18. Kovács, G.; Alonso, P.; Saini, R. Challenges of Hate Speech Detection in Social Media. *SN Comput. Sci.* **2021**, *2*, 95. [CrossRef]
19. Elsafoury, F.; Katsigiannis, S.; Wilson, S.R.; Ramzan, N. Does BERT pay attention to cyberbullying? In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Online, 11–15 July 2021; pp. 1900–1904.
20. Alkomah, F.; Ma, X. A literature review of textual hate speech detection methods and datasets. *Information* **2022**, *13*, 273. [CrossRef]
21. Akuma, S.; Lubem, T.; Adom, I.T. Comparing Bag of Words and TF-IDF with different models for hate speech detection from live tweets. *Int. J. Inf. Technol.* **2022**, *14*, 3629–3635. [CrossRef]
22. Gitari, N.D.; Zuping, Z.; Damien, H.; Long, J. A lexicon-based approach for hate speech detection. *Int. J. Multimed. Ubiquitous Eng.* **2015**, *10*, 215–230. [CrossRef]
23. Pitsilis, G.K.; Ramampiaro, H.; Langseth, H. Effective hate-speech detection in Twitter data using recurrent neural networks. *Appl. Intell.* **2018**, *48*, 4730–4742. [CrossRef]
24. Khan, S.; Fazil, M.; Sejwal, V.K.; Alshara, M.A.; Alotaibi, R.M.; Kamal, A.; Baig, A.R. BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 4335–4344. [CrossRef]
25. Gambäck, B.; Sikdar, U.K. Using Convolutional Neural Networks to Classify Hate-Speech. In Proceedings of the First Workshop on Abusive Language Online, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 85–90. [CrossRef]
26. Roy, P.K.; Tripathy, A.K.; Das, T.K.; Gao, X.Z. A framework for hate speech detection using deep convolutional neural network. *IEEE Access* **2020**, *8*, 204951–204962. [CrossRef]
27. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]

28. Mozafari, M.; Farahbakhsh, R.; Crespi, N. A BERT-based transfer learning approach for hate speech detection in online social media. In Proceedings of the Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019, Lisbon, Portugal, 10–12 December 2019; Springer: Cham, Switzerland, 2020; pp. 928–940.
29. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692 .
30. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv* **2019**, arXiv:1909.11942.
31. He, P.; Liu, X.; Gao, J.; Chen, W. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv***2020**, arXiv:2006.03654.
32. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
33. Beltagy, I.; Lo, K.; Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3615–3620. [CrossRef]
34. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
35. Mandl, T.; Modha, S.; Kumar M, A.; Chakravarthi, B.R. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In Proceedings of the Forum for Information Retrieval Evaluation, Hyderabad, India, 16 –20 December 2020; pp. 29–32.
36. Mandl, T.; Modha, S.; Shahi, G.K.; Madhu, H.; Satapara, S.; Majumder, P.; Schaefer, J.; Ranasinghe, T.; Zampieri, M.; Nandini, D.; et al. Overview of the HASOC subtrack at FIRE 2021: Hate speech and offensive content identification in English and Indo-Aryan languages. *arXiv* **2021**, arXiv:2112.09301.
37. Kumar, R.; Ojha, A.K.; Malmasi, S.; Zampieri, M. Evaluating Aggression Identification in Social Media. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, Marseille, France, 11–16 May 2020; pp. 1–5.
38. Elsafoury, F.; Katsigiannis, S.; Pervez, Z.; Ramzan, N. When the timeline meets the pipeline: A survey on automated cyberbullying detection. *IEEE Access* **2021**, *9*, 103541–103563. [CrossRef]
39. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
40. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **2005**, *18*, 602–610. [CrossRef] [PubMed]
41. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
42. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–8 December 2013; pp. 3111–3119.
43. Adewumi, T.; Liwicki, F.; Liwicki, M. Word2Vec: Optimal hyperparameters and their impact on natural language processing downstream tasks. *Open Comput. Sci.* **2022**, *12*, 134–141. [CrossRef]
44. Kim, Y. Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1746–1751. [CrossRef]
45. Feng, S.Y.; Gangal, V.; Wei, J.; Chandar, S.; Vosoughi, S.; Mitamura, T.; Hovy, E. A Survey of Data Augmentation Approaches for NLP. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 1–6 August 2021; pp. 968–988. [CrossRef]
46. Eric, M.; Goel, R.; Paul, S.; Sethi, A.; Agarwal, S.; Gao, S.; Kumar, A.; Goyal, A.; Ku, P.; Hakkani-Tur, D. MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 422–428.
47. Zou, Y.; Liu, Z.; Hu, X.; Zhang, Q. Thinking Clearly, Talking Fast: Concept-Guided Non-Autoregressive Generation for Open-Domain Dialogue Systems. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online, Dominican Republic, 7–11 November 2021; pp. 2215–2226. . [CrossRef]
48. Adewumi, O. Vector Representations of Idioms in Data-Driven Chatbots for Robust Assistance. Ph.D. Thesis, Luleå University of Technology, Luleå, Sweden, 2022.
49. 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, Minneapolis, MN, USA, 6–7 June 2019; pp. 75–86. [CrossRef]
50. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv* **2017**, arXiv:1702.08608.
51. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. In Proceedings of the International Conference on Machine Learning. PMLR, Sydney, NSW, Australia, 6–11 August 2017; pp. 3319–3328.
52. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4765–4774.

53. Shapley, L.S. *Notes on the n-Person Game—II: The Value of an n-Person Game*; RAND Corporation: Santa Monica, CA, USA, 1951.
54. Nguyen, D.Q.; Vu, T.; Tuan Nguyen, A. BERTweet: A pre-trained language model for English Tweets. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 9–14. [CrossRef]