A Text-to-Text Model for Multilingual Offensive Language Identification

Tharindu Ranasinghe[*]
Aston University
Birmingham, UK
t.ranasinghe@aston.ac.uk

Marcos Zampieri[*]
George Mason University
Fairfax, VA, USA
mzampier@gmu.edu

Abstract

The ubiquity of offensive content on social media is a growing cause for concern among companies and government organizations. Recently, transformer-based models such as BERT, XLNET, and XLM-R have achieved state-of-the-art performance in detecting various forms of offensive content (e.g. hate speech, cyberbullying, and cyberaggression). However, the majority of these models are limited in their capabilities due to their encoder-only architecture, which restricts the number and types of labels in downstream tasks. Addressing these limitations, this study presents the first pre-trained model with encoder-decoder architecture for offensive language identification with text-to-text transformers (T5) trained on two large offensive language identification datasets; SOLID and CCTK. We investigate the effectiveness of combining two datasets and selecting an optimal threshold in semi-supervised instances in SOLID in the T5 retraining step. Our pre-trained T5 model outperforms other transformer-based models fine-tuned for offensive language detection, such as fBERT and HateBERT, in multiple English benchmarks. Following a similar approach, we also train the first multilingual pre-trained model for offensive language identification using mT5 and evaluate its performance on a set of six different languages (German, Hindi, Korean, Marathi, Sinhala, and Spanish). The results demonstrate that this multilingual model achieves a new state-of-the-art on all the above datasets, showing its usefulness in multilingual scenarios. Our proposed T5-based models will be made freely available to the community.

1 Introduction

The widespread of offensive posts on social media platforms can have detrimental effects on users' mental health among other undesirable consequences. The relation between offensive language and mental health along with potential risks of self harm and depression has been widely addressed by previous studies (Bonanno and Hymel, 2013; Bannink et al., 2014; Bucur et al., 2021). To address this important issue, one of the most commonly employed strategies is to train systems to identify offensive content (Pavlopoulos et al., 2021a) mitigating its spread on social media platforms. By proactively identifying potentially harmful content, social media platforms aim to establish a safer and more inclusive environment for all users.

Early approaches to identifying offensive language ranged from classical machine learning models, such as support vector machines (Malmasi and Zampieri, 2017, 2018), to deep learning models based on word embeddings (Hettiarachchi and Ranasinghe, 2019). With the introduction of BERT (Devlin et al., 2019), transformer models have shown excellent results in offensive language identification (Zia et al., 2022). More recently, domain-specific language models for offensive language identification, such as fBERT (Sarkar et al., 2021), HateBERT (Caselli et al., 2021), and ToxicBERT[1]. have provided state-of-the-art in multiple offensive language identification benchmarks.

The aforementioned models can be grouped into two main categories following their training strategies. Models such as ToxicBERT have been trained using a classification objective by adding a classification layer on top of a BERT model and training on a large offensive language dataset. A clear limitation of this approach is that the trained model can only predict the classes that appear on the dataset. On the other hand, models such as fBERT (Sarkar et al., 2021) and HateBERT (Caselli et al., 2021) have been trained with a masked language modelling (MLM) objective. fBERT (Sarkar et al., 2021) has been trained on the offensive tweets in

WARNING: This paper contains examples that are offensive in nature.

[*] The two authors contributed equally to this work.

[1] ToxicBERT is available at https://huggingface.co/unitary/toxic-bert

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023, pages 375–384
November 1–4, 2023. ©2023 Asian Federation of Natural Language Processing

the SOLID (Rosenthal et al., 2021) dataset, while HateBERT (Caselli et al., 2021) has been trained on banned posts from Reddit Abusive Language dataset. The MLM strategy is not dependent on the number of classes present in the dataset. However, it is not possible to concatenate two datasets annotated with different annotation taxonomies under this strategy without mapping them into a common label (e.g. a general offensive class). This is a critical issue in offensive language identification as different datasets use different annotation schemes and problem formulations (e.g. hate speech, offensive, toxic, profanity). As a result, MLM based models are only trained on one dataset, which can limit their capabilities.

To address this important shortcoming, we introduce FT5, a pre-trained T5 model (Raffel et al., 2020) trained on two large-scale offensive language identification datasets. Since T5 follows a text-to-text approach, it does not rely on a classification layer. Therefore T5 (Raffel et al., 2020) can be used to train an offensive language identification model using different datasets without relying on the number of classes. We show that the proposed FT5 outperforms the plain T5 implementation as well as HateBERT (Caselli et al., 2021) and fBERT (Sarkar et al., 2021) on various offensive and hate speech detection tasks. To the best of our knowledge, this is the first pre-trained offensive language identification model based on T5.

All the previous models, such as ToxicBERT, HateBERT (Caselli et al., 2021) and fBERT (Sarkar et al., 2021) only supports English and training a large language model using similar approaches in low-resource languages can be difficult due to data scarcity. In this paper, we address this limitation by training a multilingual offensive language model, mFT5, which uses mT5 (Xue et al., 2021) as the base model. the results confirm that fine-tuned mFT5 produces state-of-the-art results in six languages, outperforming strong transformer-based models. To the best of our knowledge, mFT5 is the first multilingual model on offensive language opening exciting avenues for a multitude of languages.

The contributions of this paper are as follows:

1. An empirical evaluation of semi-supervised learning techniques that can be applied to train text-to-text models such as T5 (Raffel et al., 2020) and mT5 (Xue et al., 2021) in offensive language identification

2. A comprehensive evaluation of the effect of combining different datasets in pre-training text-to-text models.

3. The first-ever cross-lingual evaluation of mT5 (Xue et al., 2021) model in both high-resource and low-resource language settings.

4. The release of the FT5 and mFT5 made freely available to the research community, which are high-performing, state-of-the-art pre-trained models based on T5 for English and multilingual offensive language identification[2].

2 Related Work

2.1 Offensive Language Identification

The use of large pre-trained transformer models has become prevalent in NLP. This includes the development of various offensive language identification systems, which are based on transformer architectures, such as BERT (Devlin et al., 2019). These systems have demonstrated top performance in well-known competitions such as HASOC (Mandl et al., 2019), HatEval (Basile et al., 2019), OffensEval (Zampieri et al., 2019b, 2020), TRAC (Kumar et al., 2018), and TSD (Pavlopoulos et al., 2021b). Many of these competitions feature multilingual datasets opening opportunities for the use of cross-lingual models (Ranasinghe and Zampieri, 2020, 2021; Nozza, 2021). The outstanding results achieved by these systems provide concrete evidence that pre-trained transformer models are well-suited for detecting offensive content in both monolingual and multilingual settings.

User-generated content and offensive language online possess unique characteristics that are often not adequately captured by models trained on standard texts. Consequently, research has focused on the task of fine-tuning pre-trained models specifically for this challenging domain. There are several transformer models such as HateBERT (Caselli et al., 2021), fBERT (Sarkar et al., 2021) built for this purpose. In this study, we address the aforementioned limitations of fine-tuned transformer-based models. We propose the first multilingual domain-specific pre-trained offensive language identification model.

[2] https://github.com/TharinduDR/FT5

2.2 T5 Models

T5 models introduced by Raffel et al. (2020) have been widely used in many NLP tasks such as text classification (Bird et al., 2023), semantic similarity (Ni et al., 2022) and named entity recognition (Tavan and Najafi, 2022). As the T5 architecture follows a text-to-text approach, multi task learning can be used to improve the results with t5 (Raffel et al., 2020). Following the initial T5 model, multilingual T5 (mT5) models have also been proposed by Xue et al. (2021) which has provided excellent results in multilingual benchmarks. Several studies have used T5 in offensive language identification (Sabry et al., 2022; Adewumi et al., 2022). However, these studies only fine-tune the general T5 models for offensive language identification. To the best of our knowledge, this paper presents the first pre-trained domain specific T5 model for offensive language identification.

3 Methodology

Training Data In this study, we use two large offensive language identification datasets to retrain the T5 models; SOLID (Rosenthal et al., 2021) with over 9 million English tweets and CCTK with over 1.8 million posts from the civil comments platform. SOLID was the official dataset of SemEval-2020 Task 12 (OffensEval) (Zampieri et al., 2020). SOLID's annotation follows the OLID taxonomy (Zampieri et al., 2019a), which uses its level A (offensive vs non-offensive). Each instance in SOLID has been annotated semi-automatically with the mean and the standard deviation (STD) of the value for offensiveness predicted by four different machine learning models. CCTK was released for the Jigsaw Unintended Bias in Toxicity Classification Kaggle competition[3]. Each instance in CCTK has been annotated with one of the binary labels (toxic vs not toxic). Finally, we used multiple datasets for testing presented in Sections 4 and 5.

Retraining T5 We select the *t5-large*[4] (Raffel et al., 2020) and train it using the instances from SOLID (Rosenthal et al., 2021) and CCTK. For the instances in SOLID, the input texts to the model were tweets, and the output texts were "OFF" if the mean value in SOLID is above 0.5 and "NOT" otherwise. We used different thresholds (0.05, 0.1, 0.15 and 0.2) for the STD value to filter the most confident examples from SOLID. For each threshold, we consider appending/ not appending the CCTK dataset. For the instances in CCTK, the input texts to the model were the posts, and the output texts were "TOX" if the text is toxic and "NOT" if they are not toxic. As shown in Figure 1 we use "OLID_A" prefix for SOLID instances and "CCTK" prefix for CCTK instances. To create mFT5, we select *mt5-large* (Xue et al., 2021)[5] and repeat the same process. For both models, we use the same configurations; a batch-size of 16, Adam optimiser with learning rate 1e−4, and a linear learning rate warm-up over 10% of the training data and trained the models over ten epochs. We use a cluster of four GeForce RTX 3090 GPUs to train the models.

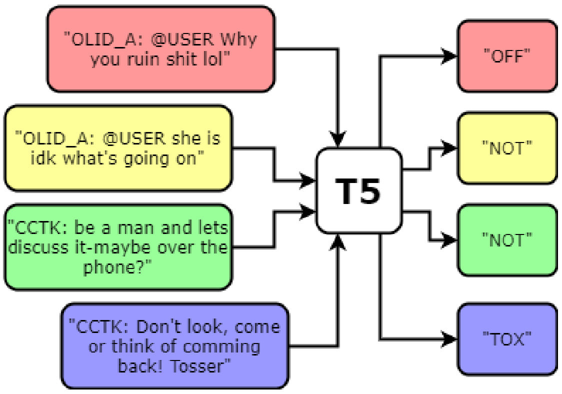
"OLID_A: @USER Why you ruin shit lol"
"OLID_A: @USER she is idk what's going on"
"CCTK: be a man and lets discuss it-maybe over the phone?"
"CCTK: Don't look, come or think of comming back! Tosser"
T5
"OFF"
"NOT"
"NOT"
"TOX"

Figure 1: T5/ MT5 pre-training process

4 English Experiments and Results

To determine the effectiveness and portability of the trained FT5, we conducted a series of experiments using benchmark datasets in English and compared our model with a general-purpose T5 model. We used the same set of configurations for all the datasets evaluated in order to ensure consistency between all the experiments. This also provides a good starting configuration for researchers who intend to use FT5 on a new dataset. For the sentence-level tasks; the input to the model is the text and the output is the related label. For the token level tasks the input to the model is the text and the output is the text with "[OFF]" placeholders infront of the offensive tokens as shown in Figure 2.

For each dataset, we used different task spe-

[3] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

[4] t5-large model is available in HuggingFace at https://huggingface.co/t5-large

[5] mt5-large model is available in HuggingFace at https://huggingface.co/google/mt5-large

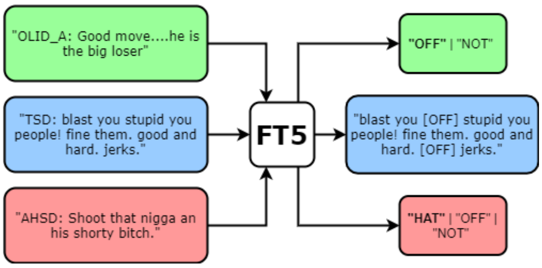
"OLID_A: Good move....he is the big loser"
"TSD: blast you stupid you people! fine them. good and hard. jerks."
"AHSD: Shoot that nigga an his shorty bitch."
FT5
"OFF" | "NOT"
"blast you [OFF] stupid you people! fine them. good and hard. [OFF] jerks."
"HAT" | "OFF" | "NOT"

Figure 2: FT5 fine-tuning process for different tasks.

cific prefixes (OLID_A for OLID, TSD for toxic spans detection etc.). We used a batch-size of eight, Adam optimizer with learning rate 1e−4, and a linear learning rate warm-up over 10% of the training data. During the training process, the parameters of the transformer model were updated. The models were trained using only training data. Furthermore, they were evaluated while training using an evaluation set that had one fifth of the rows in training data. We performed early stopping if the evaluation loss did not improve over ten evaluation steps. All the models were trained for three epochs. These experiments were also conducted in a GeForce RTX 3090 GPU. All the experiments were conducted for ten times and we report the mean and standard deviation for each experiment.

AHSD In fine-grain aggression detection, classifying offensive language and hate speech is challenging. Hate speech contains explicit instances targeted towards a specific group of people intended to degrade or insult. Davidson et al. (2017) compiled a 24, 783 English tweets dataset annotated with one of three labels – "hate speech", "only offensive", and "neither". The dataset contains 1, 430 hate speech, 19, 190 only offensive, and 4, 163 instances that are neither. We further split the dataset into training and test sets in a 4:1 ratio.

OLID We use OLID, the official dataset for OffensEval 2019 (Zampieri et al., 2019b), one of the the most popular offensive language identification shared tasks. The dataset has 13, 240 training and 860 test instances. There are 4, 400 and 240 offensive posts in the training and test dataset, respectively. For the experiment, we chose sub-task A, a binary classification task between offensive and non-offensive posts.

TRAC TRAC was released for TRAC shared task 2020 (Kumar et al., 2020). The dataset has 4200 training and 1200 test instances with three classes: overtly aggressive, covertly aggressive and non-aggressive. TRAC is the most heterogeneous dataset we used in terms of data sources containing posts from Facebook, Twitter, and YouTube.

TSD For token-level prediction we use TSD, released within the scope of SemEval-2021 Task 5: Toxic Spans Detection for English (Pavlopoulos et al., 2021a). The dataset contains 10,000 posts (comments) from the publicly available Civil Comments dataset (Borkan et al., 2019). If a post is toxic, it has been annotated for its toxic spans.

HateX HateXplain dataset (Mathew et al., 2021) was also for offensive language identification at the token level. The dataset contains 11535 training and 3844 testing instances from GAB and twitter. We only used the word level annotations.

4.1 Sentence-level Offensive Language Identification

To evaluate the sentence-level tasks, we used the macro F1 score computed on predicted sentence labels and gold sentence labels.

We present the results for the SOLID data selection thresholds and data augmentation with CCTK in Table 1 in terms of F_1 Macro. For most of the datasets tested, the 0.1 STD SOLID threshold combined with the CCTK provided the best results. Having a large number of training instances provides better results up to a certain STD threshold in SOLID, and results do not improve with adding further training instances. Furthermore, the results show that the combination of CCTK and SOLID provides better results than having one dataset in the training set. This confirms our previous assumption that T5 can take advantage of multiple datasets via text to text transfer learning.

We select the T5 model retrained on SOLID filtered with 0.1 STD combined with the CCTK dataset as the FT5 model, which provided the best result in most of the datasets. We then compare the performance of FT5 with fBERT and HateBERT. As can be seen in Table 2, FT5 outperforms fBERT and HateBERT in all of the datasets. Since the datasets contain offensive language identification, fine grained offensive language identification and fine-grained aggression identification, we can validate the effectiveness of the proposed FT5 model for offensive and aggressive language sentence-level classification tasks.

			Sentence-level			Token-level	
Train Dataset(s)	**STD**	**Inst.**	**AHSD**	**OLID**	**TRAC**	**TSD**	**HateX**
SOLID	0.05	18,169	0.832 ±0.009	0.807±0.006	0.849 ±0.005	0.542 ±0.004	0.801 ±0.006
	0.1	215,602	0.846 ±0.005	0.819±0.002	0.854 ±0.003	0.601 ±0.005	0.816 ±0.005
	0.15	1,282,474	0.870 ±0.005	0.813±0.003	**0.870 ±0.002**	0.591 ±0.008	0.801 ±0.006
	0.2	6,595,397	0.859 ±0.003	0.805±0.005	0.865 ±0.005	0.561 ±0.005	0.795 ±0.006
SOLID+CCTK	0.05	1,823,043	0.865 ±0.004	0.819±0.003	0.859 ±0.007	0.609 ±0.004	0.815 ±0.004
	0.1	2,020,476	**0.886 ±0.004**	**0.823 ±0.002**	0.869 ±0.005	**0.648 ±0.005**	**0.825 ±0.004**
	0.15	3,087,348	0.872 ±0.005	0.816±0.004	0.864 ±0.005	0.605 ±0.006	0.812 ±0.007
	0.2	8,400,271	0.868 ±0.005	0.809±0.006	0.858 ±0.006	0.589 ±0.012	0.808 ±0.009
CCTK	NA	1,804,874	0.832 ±0.009	0.813±0.006	0.842 ±0.005	0.595±0.004	0.809 ±0.006

Table 1: FT5 results for different sentence-level and token-level offensive language detection benchmarks.

Dataset	**Model**	**Macro F1**
AHSD	**FT5**	**0.886±0.004**
	fBERT	0.878±0.005
	HateBERT	0.846±0.009
	T5	0.821±0.012
OLID	**FT5**	**0.823±0.002**
	fBERT	0.810±0.005
	HateBERT	0.803±0.009
	T5	0.775±0.006
TRAC	**FT5**	**0.869±0.003**
	fBERT	0.859±0.005
	HateBERT	0.848±0.006
	T5	0.846±0.010

Table 2: The test set macro F_1 scores for sentence-level datasets and models. Results are ordered by performance. Best results are shown in bold font.

Dataset	**Model**	**Macro F1**
TSD	**FT5**	**0.648±0.012**
	fBERT	0.530±0.021
	T5	0.421±0.019
	HateBERT	0.410±0.027
HateX	**FT5**	**0.825±0.008**
	fBERT	0.812±0.009
	HateBERT	0.792±0.016
	T5	0.775±0.025

Table 3: The test set macro F_1 scores for sentence-level datasets and models. Results are ordered by performance. Best results are shown in bold font.

4.2 Token-level Offensive Language Identification

Multiple studies on token-level offensive language identification has discussed the need for accurate token-level predictions for improved model explainability (Mathew et al., 2021; Zampieri et al., 2023). Motivated by recent studies, we investigate our model performance on token-level offensive language identification. The token-level tasks were evaluated using the macro F1 score computed on predicted character offsets and gold character offsets (Da San Martino et al., 2019).

FT5 outperforms fBERT and HateBERT in all of the token-level offensive language identification datasets too as can be seen in Table 3. There is a clear improvement with the TSD dataset where the FT5 model outperforms the fBERT model by 0.11 macro F1 score which is over a 20% boost.

5 Multilingual Experiments and Results

To determine the effectiveness and portability of our multilingual model; mFT5, we conducted a series of experiments using benchmark datasets covering high-resource, mid-resource and low-resource languages. These datasets are summarised in Table 4. We used the same set of configurations we used for English experiments. The models were trained using the training set and evaluated on the test sets of each dataset.

5.1 Sentence-level Offensive Language Identification

For sentence-level offensive language identification, we mapped the labels of each dataset to its closet annotation scheme as we did for English

Language	Source(s)	Train, dev, test size	Labels	Sentence-level	Token-level
German (Risch et al., 2021)	Facebook	2076, 519, 649	Toxic Not-toxic	TOX, NOT	NA
Spanish (Plaza-del Arco et al., 2021)	Twitter Instagram Youtube	30163, 7540, 9425	Offensive individual target Offensive group target Offensive other target Expletive language Non-offensive	OFF, NOT	NA
Hindi (Mandl et al., 2019)	Twitter	5120, 1280, 1600	Offensive Not offensive	OFF, NOT	NA
Korean (Jeong et al., 2022)	Naver YouTube	25876, 6468, 8085	Offensive Not offensive	OFF, NOT	Available
Sinhala (Ranasinghe et al., 2022)	Twitter	6000, 1500, 2500	Offensive Not offensive	OFF, NOT	Available
Marathi (Gaikwad et al., 2021)	Twitter	2889, 722, 510	Offensive Not offensive	OFF, NOT	NA

Table 4: Datasets that were used to evaluate mFt5 model. **Source** column displays the platform data extracted,**Train, dev, test size** column shows the number of instances of the train, dev and test sets. **Label** column shows the original labels and **Sentence-level** column show the output label in sentence-level experimenst discussed in Section 5. **Token-level** column shows the availability of the token-level data.

Train Dataset(s)	STD	Inst.	High Resource		Mid Resource		Low Resource	
			German	Spanish	Hindi	Korean	Sinhala	Marathi
SOLID	0.05	18,169	0.567 ±0.010	0.832 ±0.009	0.799 ±0.005	0.735 ±0.004	0.748 ±0.008	0.789 ±0.016
	0.1	215,602	0.601 ±0.002	0.846 ±0.005	0.807 ±0.003	0.776 ±0.005	0.756 ±0.008	0.825 ±0.009
	0.15	1,282,474	0.598 ±0.007	0.870 ±0.005	0.839 ±0.002	0.781 ±0.008	0.784 ±0.007	**0.858 ±0.006**
	0.2	6,595,397	0.588 ±0.006	0.859 ±0.003	0.825 ±0.005	0.757 ±0.005	0.766 ±0.008	0.844 ±0.010
SOLID +CCTK	0.05	1,823,043	0.611 ±0.004	0.852±0.005	0.859 ±0.007	0.769 ±0.004	0.812±0.016	0.835 ±0.009
	0.1	2,020,476	**0.653±0.031**	**0.886 ±0.004**	**0.845 ±0.005**	**0.799±0.004**	**0.856±0.007**	0.854±0.006
	0.15	3,087,348	0.642±0.005	0.872 ±0.005	0.822 ±0.005	0.778 ±0.006	0.856 ±0.006	0.849±0.008
	0.2	8,400,271	0.611 ±0.005	0.843±0.012	0.818 ±0.006	0.765 ±0.012	0.836 ±0.006	0.811±0.005
CCTK	NA	1,804,874	0.628 ±0.009	0.829±0.006	0.825 ±0.005	0.775±0.004	0.796 ±0.005	0.801 ±0.003

Table 5: mFT5 results for different multilingual offensive language detection benchmarks.

benchmarks; OLID level A (offensive, not offensive) or CCTK (toxic, non toxic). Following this, Spanish, Hindi, Korean, Sinhala and Marathi labels were mapped to OLID level A, and German labels were mapped to CCTK as shown in *sentence-level* column in Table 4. We use "OLID_A" prefix for Spanish, Hindi, Korean, Sinhala and Marathi instances and "CCTK" prefix for German instances. In the training process, we started with different pre-trained models on the configurations described in Section 3. The input to the model is the text preceded by the relevant prefix, and the output is the related label. We performed individual experiments for each language separately.

We present the results for the SOLID data selection thresholds and data augmentation with CCTK in Table 5 in terms of F_1 Macro for each test set. For most of the datasets tested, the 0.1 STD SOLID threshold combined with the CCTK provided the best results. Having a large number of training instances provides better results up to a certain STD threshold in SOLID, and results do not improve with adding further training instances. Furthermore, the results show that the combination of CCTK and SOLID provides better results than having one dataset in the training set. This confirms our previous assumption that T5 can take advantage of multiple datasets via text to text transfer learning. We select the T5 model retrained on SOLID filtered with 0.1 STD combined with the CCTK dataset as the FT5 model, which provided the best result in five out of six datasets.

We then compare the performance of mFT5 with mBERT and XLM-R base models. These models are trained on the training set of each dataset by adding a classification layer on top of the transformer model. As shown in Table 6, mFT5 outperforms mBERT and XLM-R in all of the datasets. Since these datasets contain high-resource and low-resource languages as well as data from different social media platforms, we can validate the effectiveness of the proposed mFT5 model for offensive language identification in multiple languages and platforms.

Dataset	Model	Macro F1
German	**mFT5**	**0.653±0.031**
	XLM-R	0.621±0.023
	mBERT	0.572±0.013
	mFT5*	0.438±0.015
	mT5	0.398±0.016
Spanish	**mFT5**	**0.886±0.004**
	XLM-R	0.853±0.005
	mBERT	0.821±0.008
	mFT5*	0.785±0.005
	mT5	0.626±0.027
Hindi	**mFT5**	**0.845±0.003**
	XLM-R	0.811±0.007
	mBERT	0.798±0.006
	mFT5*	0.745±0.007
	mT5	0.612±0.012
Korean	**mFT5**	**0.799±0.004**
	XLM-R	0.765±0.006
	mBERT	0.755±0.008
	mFT5*	0.736±0.008
	mT5	0.715±0.005
Sinhala	**mFT5**	**0.856±0.007**
	XLM-R	0.834±0.005
	mFT5*	0.746±0.009
	mT5	0.538±0.029
	mBERT	0.531±0.013
Marathi	**mFT5**	**0.854±0.006**
	XLM-R	0.843±0.003
	mBERT	0.821±0.006
	mFT5*	0.708±0.012
	mT5	0.421±0.017

Table 6: The test set macro F_1 scores for coarse-grained offensive language detection. Results are ordered by performance. The best results are shown in bold font.

Zero-shot Offensive Language Identification - We also experimented with zero-shot cross-lingual offensive language identification with the mFT5 model. With this setting, we did not train the mFT5 on the language-specific training data. The results are shown in mFT5* rows in Table 6. While zero-shot cross-lingual experiments did not provide the best results, they provided very competitive results compared to the baselines. The results confirm the strong cross-lingual nature of the pre-trained mFT5 model in detecting offensive language. It should also be noted that an MLM approach similar to fBERT and HateBERT needs labelled data to fine-tune and will not be able to provide zero-shot offensive language identification. Therefore, mFT5 is useful for low-resource languages where the training data is scarce.

5.2 Token-level Offensive Language Identification

We also experimented with token-level offensive language identification in the datasets where the token-level labels are available; Korean and Sinhala. The input to the model is the text, and the output is the text with "[OFF]" placeholders in front of the offensive tokens. To evaluate the models, we used the macro F1 score computed on predicted offensive tokens and gold offensive tokens. We compared the results with token classification architecture in mBERT and XLM-R large models.

Dataset	Model	Macro F1
Korean	**mFT5**	**0.489±0.004**
	XLM-R	0.466±0.009
	mBERT	0.453±0.013
	mT5	0.311±0.016
Sinhala	**mFT5**	**0.743±0.009**
	XLM-R	0.723±0.013
	mT5	0.316±0.023
	mBERT	0.00

Table 7: The test set macro F_1 scores for token-level offensive language detection. Results are ordered by performance. The best results are shown in bold font.

As shown in Table 7, mFT5 outperforms mBERT and XLM-R in all of the token-level offensive language identification datasets too. It should be noted that pre-trained models with task-specific heads such as toxic-bert will not be able to perform token-level tasks. However, our text-to-text approach in mFT5 provided state-of-the-art results at token-level too.

6 Conclusion and Future Work

Neural transformer models have outperformed previous state-of-the-art deep learning models across different NLP tasks including offensive language identification. Following impressive results in international benchmark competitions, domain-specific pre-trained neural transformers such as ToxicBERT, fBERT and HateBERT have been proposed for offensive language identification. As discussed in this paper, these models have limitations which makes it difficult extend them in to different datasets. We address these limitations by proposing FT5, a *t5-large* model that has been trained using over 2 million instances from the SOLID and CCTK datasets. The FT5 model achieves better results in both sentence-level and token-level tasks across different offensive language identification benchmarks.

This paper also introduced mFT5. To the best of our knowledge, mFT5 is the first pre-trained multilingual offensive language detection model. The model uses the *mt5-large* and is trained using the same data used to train FT5. We show that the proposed FmT5 model achieves better results in both sentence-level and token-level tasks compared to the mBERT, XLM-R, and the vanilla mT5 model across different offensive language identification benchmark datasets. We show that the model performs consistently well across different languages and platforms. Furthermore, the model showed strong zero-shot cross-lingual results opening exciting new avenues for multilingual offensive language detection.

In future work, we would like to extend the proposed FT5 model to the identification of offensive spans along with their targets using the recently released TBO dataset (Zampieri et al., 2023). We believe that modelling targets and offensive expressions jointly is an important step towards improving explainability in offensive language identification systems. Another important direction we have been exploring is the computational efficiency. We have recently experimented with teacher-student architectures using knowledge distillation (KD) and we have shown that the use of KD results in lightweight models that are more computationally efficient and perform on par with larger models (Ranasinghe and Zampieri, 2023). We would like to investigate teacher-student architectures using the proposed FT5 model. Finally, we are interested in the application of multilingual models to low-resource scenarios. Thousands of languages and dialects are spoken in the world, but research on offensive language identification is still (mostly) restricted to English and a few other high-resource languages. We believe that the release of mT5 will encourage research on offensive language identification models for low-resource languages, dialects, and other challenging linguistic scenarios such as code-mixed texts.

Limitations

Training a T5 model requires a large amount of computing resources. We noted that training a T5 model can take more GPU resources than training a BERT model such as fBERT (Sarkar et al., 2021). Therefore, we did not experiment with large T5 models such as T5-XL and T5-XXL. While these models might perform better than T5-Large models we experimented with, they would consume more resources, limiting their potential use cases.

Ethics Statement

FT5 and mFT5 are essentially T5 models for offensive language identification, which is trained on multiple publicly available datasets. We used multiple datasets referenced in this paper which were previously collected and annotated to evaluate the models. No new data collection has been carried out as part of this work. We have not collected or processed writers'/users' information, nor have we carried out any form of user profiling to protect users' privacy and identity.

We believe that content moderation should be a trustworthy and transparent process applied to clearly harmful content so it does not hinder individual freedom of expression rights. We encourage research in automatically detecting offensive content on the web trough a trustworthy and transparent process. Using our proposed models for this purpose will alleviate the psychological burden for social media moderators who are exposed to large amounts offensive content while ensuring a more transparent moderation process.

Acknowledgments

We would like to thank the anonymous AACL-IJCNLP reviewers for their insightful feedback. We further thank the creators of the datasets used in this paper for making the datasets publicly available for our research.

The computational experiments in this paper were conducted on an Aston EPS Machine Learning Server, funded by the EPSRC Core Equipment Fund, Grant EP/V036106/1.

References

Tosin Adewumi, Sana Sabah Sabry, Nosheen Abid, Foteini Liwicki, and Marcus Liwicki. 2022. T5 for hate speech, augmented data and ensemble. *arXiv preprint arXiv:2210.05480*.

Rienke Bannink, Suzanne Broeren, Petra M van de Looij-Jansen, Frouwkje G de Waart, and Hein Raat. 2014. Cyber and Traditional Bullying Victimization as a Risk Factor for Mental Health Problems and Suicidal Ideation in Adolescents. *PloS one*, 9(4).

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of SemEval*.

Jordan J. Bird, Anikó Ekárt, and Diego R. Faria. 2023. Chatbot interaction with artificial intelligence: human data augmentation with t5 and language transformer ensemble for text classification. *Journal of Ambient Intelligence and Humanized Computing*, 14(4):3129–3144.

Rina A Bonanno and Shelley Hymel. 2013. Cyber bullying and internalizing difficulties: Above and beyond the impact of traditional forms of bullying. *Journal of youth and adolescence*, 42(5):685–697.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Proceedings of WWW*.

Ana-Maria Bucur, Marcos Zampieri, and Liviu P. Dinu. 2021. An exploratory analysis of the relation between offensive language and mental health. In *Findings of ACL*.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. Hatebert: Retraining bert for abusive language detection in english. In *Proceedings of WOAH*.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of EMNLP*.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of ICWSM*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Saurabh Sampatrao Gaikwad, Tharindu Ranasinghe, Marcos Zampieri, and Christopher Homan. 2021. Cross-lingual offensive language identification for low resource languages: The case of Marathi. In *Proceedings of RANLP*.

Hansi Hettiarachchi and Tharindu Ranasinghe. 2019. Emoji powered capsule network to detect type and target of offensive posts in social media. In *Proceedings of RANLP*.

Younghoon Jeong, Juhyun Oh, Jaimeen Ahn, Jongwon Lee, Jihyung Mon, Sungjoon Park, and Alice Oh. 2022. Kold: Korean offensive language dataset. In *Findings of the ACL*.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of TRAC*.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating aggression identification in social media. In *Proceedings of TRAC*.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting hate speech in social media. In *Proceedings of RANLP*.

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

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of FIRE*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In *Proceedings of AAAI*.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pretrained text-to-text models. In *Findings of the ACL*.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of ACL*.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021a. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of SemEval*.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021b. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of SemEval*.

Flor Miriam Plaza-del Arco, Arturo Montejo-Ráez, L. Alfonso Ureña-López, and María-Teresa Martín-Valdivia. 2021. OffendES: A new corpus in Spanish for offensive language research. In *Proceedings of RANLP*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Tharindu Ranasinghe, Isuri Anuradha, Damith Premasiri, Kanishka Silva, Hansi Hettiarachchi, Lasitha Uyangodage, and Marcos Zampieri. 2022. Sold: Sinhala offensive language dataset. *arXiv preprint arXiv:2212.00851*.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual Offensive Language Identification with Cross-lingual Embeddings. In *Proceedings of EMNLP*.

Tharindu Ranasinghe and Marcos Zampieri. 2021. Multilingual offensive language identification for low-resource languages. *Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–13.

Tharindu Ranasinghe and Marcos Zampieri. 2023. Teacher and student models of offensive language in social media. In *Findings of the ACL*.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval*.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. SOLID: A Large-Scale Weakly Supervised Dataset for Offensive Language Identification. In *Findings of the ACL*.

Sana Sabah Sabry, Tosin Adewumi, Nosheen Abid, György Kovács, Foteini Liwicki, and Marcus Liwicki. 2022. Hat5: Hate language identification using text-to-text transfer transformer. *arXiv preprint arXiv:2202.05690*.

Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia. 2021. fbert: A neural transformer for identifying offensive content. In *Findings of EMNLP*.

Ehsan Tavan and Maryam Najafi. 2022. MarSan at SemEval-2022 task 11: Multilingual complex named entity recognition using t5 and transformer encoder. In *Proceedings of SemEval*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of NAACL*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of NAACL*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of SemEval*.

Marcos Zampieri, Skye Morgan, Kai North, Tharindu Ranasinghe, Austin Simmmons, Paridhi Khandelwal, Sara Rosenthal, and Preslav Nakov. 2023. Target-based offensive language identification. In *Proceedings of ACL*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*.

Haris Bin Zia, Ignacio Castro, Arkaitz Zubiaga, and Gareth Tyson. 2022. Improving zero-shot cross-lingual hate speech detection with pseudo-label fine-tuning of transformer language models. In *Proceedings of ICWSM*.