A Federated Learning Approach to Privacy Preserving Offensive Language Identification

Marcos Zampieri[1]**, Damith Premasiri**[2]**, Tharindu Ranasinghe**[3]**,**
[1]George Mason University, USA, [2]Lancaster University, UK, [3]Aston University, UK
mzampier@gmu.edu

Abstract

The spread of various forms of offensive speech online is an important concern in social media. While platforms have been investing heavily in ways of coping with this problem, the question of privacy remains largely unaddressed. Models trained to detect offensive language on social media are trained and/or fine-tuned using large amounts of data often stored in centralized servers. Since most social media data originates from end users, we propose a privacy preserving decentralized architecture for identifying offensive language online by introducing Federated Learning (FL) in the context of offensive language identification. FL is a decentralized architecture that allows multiple models to be trained locally without the need for data sharing hence preserving users' privacy. We propose a model fusion approach to perform FL. We trained multiple deep learning models on four publicly available English benchmark datasets (AHSD, HASOC, HateXplain, OLID) and evaluated their performance in detail. We also present initial cross-lingual experiments in English and Spanish. We show that the proposed model fusion approach outperforms baselines in all the datasets while preserving privacy.

Keywords: federated learning, offensive language identification, privacy

1. Introduction

NLP systems relying on modern deep learning paradigms are trained on very large amounts of data. In several applications and domains (e.g., social media), most data used to train machine learning models comes from end users. Such confidential data often cannot be shared without compromising users' privacy. This is an important concern for organizations that handle large amounts of confidential data, such as financial institutions, healthcare facilities, law firms, and many others. With the widespread use of personal computing devices (e.g., PCs, smartphones, and virtual assistants), data privacy also became a great concern to individuals, which motivated several countries to pass legislation aiming to protect users' privacy such as the European Union General Data Protection Regulation (GDPR)[1] and the Swiss Datenschutzgesetz (DSG).[2]

The need for privacy-preserving machine learning models that can handle confidential data while protecting organizations' and users' privacy emerges from this situation. To address this important challenge, Federated Learning (FL) has become an increasingly popular machine learning paradigm (McMahan et al., 2017) as it allows us to train robust machine learning models across multiple devices or servers without exchanging data. In FL, multiple clients work together under the coordination of a central server. Each client's data is stored locally and not exchanged among clients or with the central server. FL, therefore, offers the possibility of training robust machine learning models on large numbers of decentralized local data repositories without compromising privacy. FL models have been successfully applied in a wide range of applications in computer networks (Lim et al., 2020), computer vision (Yan et al., 2021), information retrieval (Wang et al., 2021), NLP (Chen et al., 2019), and many others.

In this paper, we explore the use of FL in offensive language identification online through a model fusion technique (Choshen et al., 2022). Datasets containing the various forms of offensive speech (e.g., hate speech, cyberbullying, etc.) are sensitive in nature, which creates an interesting use case for FL. The use of FL and other privacy-preserving paradigms allows social media platforms to work together to solve this important issue without the need to exchange confidential information, thus preserving users' privacy. While FL has recently started to be explored in NLP (Chen et al., 2019; Lin et al., 2022b), including the workshop on Federated Learning for NLP (FL4NLP) at ACL-2022 (Lin et al., 2022a), to the best of our knowledge, no studies have yet explored the use of FL in the context of offensive language identification. Our work fills this gap by introducing FL in the context of offensive language identification online and by providing the community with an evaluation of FL methods using four publicly available English offensive language benchmark datasets presented in Section 3.

One recent study (Gala et al., 2023) proposed

[1] https://gdpr.eu/

[2] https://www.edoeb.admin.ch/edoeb/de/home/datenschutz/ueberblick/datenschutz.html

TRAC-2024 Workshop, pages 12–20

	Training		Testing		
Dataset	Inst.	OFF %	Inst.	OFF %	Data Sources
AHSD (Davidson et al., 2017)	19,822	0.83	4,956	0.82	Twitter
HASOC (Mandl et al., 2020)	5,604	0.36	1,401	0.35	Twitter, Facebook
HateXplain (Mathew et al., 2021)	11,535	0.59	3,844	0.58	Twitter, Gab
OLID (Zampieri et al., 2019a)	13,240	0.33	860	0.27	Twitter

Table 1: The four datasets, including the number of instances (Inst.) in the training and testing sets, the OFF % in each set and the data source.

FL in offensive language identification, but it lacks the consideration of combining different data. Their architecture solely focuses on distributed training on the same dataset with multiple clients and evaluating *fedopt (Reddi et al., 2021), fedprox (Sahu et al., 2019)* algorithms to optimise the global model. Our main focus in this study is on combining multiple models using FL, which could identify offensive content in different data.

2. Related Work

Offensive Language Identification The task of automatically identifying offensive language online has been substantially explored in the literature (MacAvaney et al., 2019; Melton et al., 2020; Zia et al., 2022; Weerasooriya et al., 2023). Multiple types of offensive content have been addressed, such as *aggression*, *cyberbulling*, and *hate speech* using classical machine learning classifiers (e.g., Support Vector Machines) (Malmasi and Zampieri, 2017, 2018), neural networks (Gambäck and Sikdar, 2017; Djuric et al., 2015; Hettiarachchi and Ranasinghe, 2019), pre-trained general-purpose transformer-based language models (Ranasinghe and Zampieri, 2020, 2021), and fine-tuned language models on offensive language datasets (Caselli et al., 2020; Sarkar et al., 2021). The vast majority of studies addressed offensive content in English and other widely-spoken resource-rich languages such as Arabic (Mubarak et al., 2021), Portuguese (Fortuna et al., 2019) and Turkish (Çöltekin, 2020) while a few studies dealt with low-resource languages (Fišer et al., 2017; Gaikwad et al., 2021; Raihan et al., 2023). Multiple competitions on this topic have been organized creating important benchmark datasets such as OffensEval (Zampieri et al., 2019b, 2020), HASOC (Mandl et al., 2020; Modha et al., 2021; Satapara et al., 2022), TRAC (Kumar et al., 2018, 2020), and HatEval (Basile et al., 2019). While substantial progress has been made in the past few years, to the best of our knowledge, none of the aforementioned studies or competitions has addressed the question of data privacy.

Federated Learning in NLP With the goal of preserving users' data privacy, FL architectures have been extensively studied in a variety of domains (Wang et al., 2021) in the past several years. Only more recently, however, FL has been explored for text and speech processing (Lin et al., 2022b; Silva et al., 2023; Zhang et al., 2023; Che et al., 2023). Recent workshops co-located with top-tier conferences confirm this growing interest in FL and privacy in general. The workshop on Privacy in Natural Language Processing (PrivateNLP) (Feyisetan et al., 2022), which is currently in its fourth edition, addressed the interplay between NLP and data privacy while the aforementioned FL4NLP workshop (Lin et al., 2022a) co-located with ACL-2022 was the first workshop organized focusing exclusively on FL for NLP. Most papers presented in the workshop, however, dealt with language modelling and learning representation rather than with downstream tasks and applications such as offensive language identification. As we mentioned before, a recent study applied different FL strategies in offensive language identification (Gala et al., 2023). However, their study focuses on distributed training on the same dataset (Sahu et al., 2019).

3. Data

We use four popular publicly available datasets containing English data summarized in Table 1. As the datasets were annotated using different guidelines and labels, following the methodology described in previous work (Ranasinghe and Zampieri, 2020), we map all labels to OLID level A (Zampieri et al., 2019a), which contains the labels offensive (OFF) vs. not offensive (NOT). We choose OLID due to the flexibility provided by its general three-level hierarchical taxonomy below, where the OFF class contains all types of offensive content, from general profanity to hate speech, while the NOT class contains non-offensive examples. The OLID taxonomy is presented next:

- **Level A:** Offensive (OFF) vs. Non-offensive (NOT).
- **Level B:** Classification of the type of offensive (OFF) tweet - Targeted (TIN) vs. Untargeted (UNT).
- **Level C:** Classification of the target of a targeted (TIN) tweet - Individual (IND) vs. Group (GRP) vs. Other (OTH).

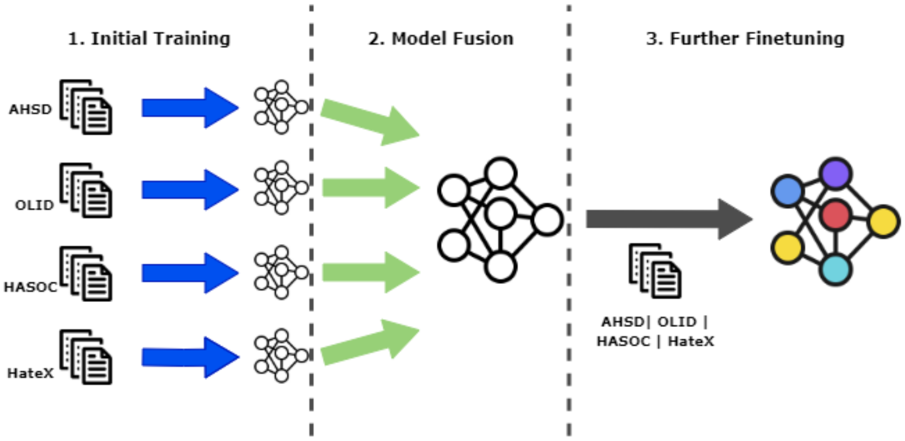

Figure 1: The three stages of the FL pipeline in the proposed fused model.

In the OLID taxonomy, offensive (OFF) posts targeted (TIN) at an individual are often cyberbulling whereas offensive (OFF) posts targeted (TIN) at a group is often hate speech.

AHSD (Davidson et al., 2017) is one of the most popular hate speech datasets available. The dataset contains data retrieved from Twitter and it was annotated using crowdsourcing. The annotation taxonomy contains three classes; Offensive, Hate, and Neither. We conflate Offensive and Hate under a class OFF while neither class corresponds to OLID's NOT class.

OLID (Zampieri et al., 2019a) is the official dataset of the SemEval-2019 Task 6 (OffensEval) (Zampieri et al., 2019b). It contains data from Twitter annotated with a three-level hierarchical annotation in which level A classifies posts into offensive and not offensive; level B differentiates between targeted pots (insults and threats) and untargeted posts (general profanity); and level C classifies them into three targets: individual, group, or other. We adopt the labels in OLID level A as our classification labels.

HASOC (Mandl et al., 2020) is the dataset used in the HASOC shared task 2020. It contains posts retrieved from Twitter and Facebook. The upper level of the annotation taxonomy used in HASOC is the same as OLID's level A, which allows us to directly use the same labels in our models.

HateXplain (Mathew et al., 2021) is a recent dataset collected for the explainability of hate speech. It contains both token- and post-level annotation of Twitter and Gab posts. The annotation taxonomy contains three classes; hate speech, offensive speech, and normal. Following the annotation guidelines of OLID (Zampieri et al., 2019a), we mapped the hate speech and offensive speech classes to offensive (OFF) and normal class to not offensive (NOT).

4. Methodology

The proposed FL pipeline contains three steps depicted in Figure 1. We describe these steps below.
Initial Model Training Transformer models have achieved state-of-the-art performance in many NLP tasks (Devlin et al., 2019), including offensive language identification (Ranasinghe et al., 2019; Sarkar et al., 2021). Therefore, our methodology in this paper builds around pre-trained transformers. For the text classification tasks such as offensive language identification, we use the pre-trained transformer models by utilizing the hidden representation of the classification token (CLS) as shown in Figure 2. For this task, we implemented a softmax layer on top of the CLS token, i.e., the predicted probabilities are $\boldsymbol{y}^{(B)} = \mathrm{softmax}(W\boldsymbol{h} + b)$, where $W \in \mathbb{R}^{k \times d}$ is the softmax weight matrix, and k is the number of labels. which in our case is always equal to two.

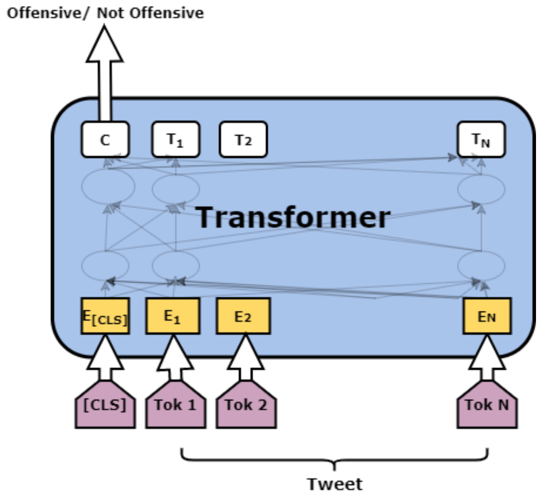

Figure 2: A sample transformer model for offensive language identification (Ranasinghe and Zampieri, 2020) predicting offensive and not offensive labels.

We used this text classification architecture to build separate models for each dataset that we introduced in the previous section. We trained the model using the training sets of each dataset. We employed a batch-size of 16, Adam optimiser with learning rate $4\mathrm{e}{-5}$, and a linear learning rate warm-up over 10% of the training data. During the training process, the parameters of the transformer model and the parameters of the subsequent layers were updated. The models 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 three evaluation steps. All the models were trained for three epochs.

We repeated this process with two popular pre-trained transformer models; *bert-large-cased* (Devlin et al., 2019) and *fBERT* (Sarkar et al., 2021). The *bert-large-cased* is a general purpose pre-trained transformer model while *fBERT* is a domain-specific pre-trained transformer model for offensive language identification that has been trained on over 1.4 million offensive tweets in SOLID dataset (Rosenthal et al., 2021) and has shown state-of-the-art results in several offensive language identification benchmarks (Sarkar et al., 2021).

Model Fusion In order to combine the different models created using different datasets, we followed a recent approach named model fusion (Choshen et al., 2022). Model Fusion is the process of taking several fine-tuned models and creating a new base model. Formally, given an initialization base model P and n models fine-tuned on it, let $W_1, W_2 \ldots W_n \in \mathbb{R}^d$ be the weights fine-tuned by the models over P. Fusing is a function

$$W_{fuse} = f(W_1, W_2, \ldots, W_n) \quad \mathbb{R}^d \times \mathbb{R}^d \times \ldots \times \mathbb{R}^d \rightarrow \mathbb{R}^d \tag{1}$$

In this work, we propose the simplest form of fusion. For each weight shared by all models, assign the average weight to the model.

$$W_{fuse} = f\left(W_1, W_2, \ldots, W_n\right) \quad = \frac{W_1 + W_2 + \ldots + W_n}{n} \tag{2}$$

In order to empirically evaluate model fusion in offensive language identification, we consider all possible seven combinations. These include different combinations of two models, such as $AHSD + OLID$ and $HASOC + HateX$, different combinations of three models, such as $AHSD + OLID + HASOC$ and $AHSD + OLID + HASOC$ and finally, the combination of all four models.

Further Finetunning The weights of the fused model resulting from step 2 can be anomalous as we followed a naive averaging method. Therefore, we performed a further finetuning step on the fused model. In this step, we fine-tuned the fused model using only one available dataset in a particular environment. We followed the same classification objective described in step 1 and used the same configurations. However, to avoid the model being biased toward the finetuning dataset, we only used 20% of the available training data in the finetuning step.

The whole pipeline described above simulates a real-life scenario where the data can not be shared. The machine learning models are trained in separate environments using their own data, as in the first step. In the second step, with model fusion, we combined the models. In the final step. We further fine-tuned the fused model on a particular dataset where we repeated the process for all four datasets. Therefore with this pipeline, the datasets are not shared, and privacy is preserved among the different environments.

4.1. Baseline Models

We compared our fusion-based approach to two baseline models.

Non-fused Baseline We train a transformer-based baseline using the training set of one of the datasets and evaluate it on the test set of that particular dataset as well as on the test sets of other datasets. We repeated the process for all four datasets with two transformer models; *bert-large-cased* (Devlin et al., 2019) and *fBERT* (Sarkar et al., 2021). This baseline reflects the most common approach in offensive language detection, where a model is trained on a dataset available for a particular environment, but evaluated on other datasets in different environments as well.

Ensemble Baseline We also used an ensemble baseline; where we trained four separate transformer models on each dataset. For each test instance, we predicted values from all four models, and the final label is the label predicted with the highest probability from all four models. Similar to our previous experiments we repeated the process for *bert-large-cased* (Devlin et al., 2019) and *fBERT* (Sarkar et al., 2021).

5. Results and Discussion

In Table 2, we present the best results from each approach for each dataset. We show the results for fBERT as it provided better overall results. For the AHSD test set, the best result, 0.921 Macro F1 score, is obtained when fBERT models are trained on AHSD and OLID and fused, then further fine-tuned on AHSD. For OLID the best result, 0.839 Macro F1 score was provided when BERT-large-cased models trained on AHSD and OLID were fused and further fine-tuned on AHSD. Similarly, for HateX the best result, 0.777 was provided when the

Dataset	Approach	Models				Macro F1
AHSD	non-fused	AHSD		-	-	0.931 ±0.01
	fusion with FT	AHSD	OLID	-	-	0.921 ±0.00
	fusion without FT	AHSD	OLID	-	-	0.866 ±0.00
	ensemble	AHSD	OLID	-	-	0.845 ±0.01
OLID	non-fused	-	OLID	-	-	0.854 ±0.00
	fusion with FT	AHSD	OLID	-	-	0.837 ±0.03
	fusion without FT	AHSD	OLID	-	-	0.836 ±0.00
	ensemble		OLID	-	HateX	0.785 ±0.04
HASOC	non-fused	-	-	HASOC	-	0.798 ±0.01
	fusion without FT	AHSD	OLID	HASOC	-	0.770 ±0.01
	fusion with FT	AHSD	OLID	HASOC	-	0.754 ±0.07
	ensemble	AHSD		HASOC	-	0.647±0.02
HateX	non-fused	-		-	HateX	0.795 ±0.01
	fusion with FT	AHSD	-	-	HateX	0.777 ±0.00
	fusion without FT	AHSD	OLID	-	HateX	0.772 ±0.01
	ensemble	-	-	HASOC	HateX	0.654 ±0.01

Table 2: The best result for each dataset for each approach; non-fused models, fused models with fine-tuning (FT), fused models without finetuning and ensemble. We only report the results with fBERT. The results are ordered from Macro F1.

fBERT models trained on AHSD and HateX were fused and further fine-tuned on HateX. However, HASOC follows a different pattern, and the best result was produced when fBERT models trained on AHSD, OLID and HASOC were fused, and further fine-tuned on AHSD. Overall, fBERT models provided slightly better results than BERT-large-cased models in most experiments. This is mainly because the fBERT model was trained on domain-specific data on offensive language identification. Finally, we present all results of the fused models and the non-fused model baseline in Table 3 in terms of Macro F1 score.

5.1. Discussion

We discuss the following four main findings from our results;

(1) The fused model performs better when evaluated on the same dataset used in further finetunning. All the datasets except for HASOC, the best result was produced when the fused model was further fine-tuned on that particular dataset. For HASOC too, when the fBERT model trained on AHSD, OLID and HASOC were fused and further fine-tuned on HASOC provided 0.754 Macro F1 score, which is very close to the best result (0.770). With the results, we can conclude that the fused model performs better when evaluated on the same dataset used in further finetunning. This observation reflects an ideal scenario in real-world applications where we want an ML model to perform excellently in data specific to our environment/ platform. This objective can be achieved successfully with model fusion and finetunning as we see in the results.

(2) The fused model generalizes well across datasets even when it is not used in finetunning. One drawback of fused models is that the result slightly decreases compared to the non-fused models trained only using a particular dataset. In the results, this is clear as there is a decrease in the Macro F1 score between underlined values and bolded values. Furthermore as you can see in Table 2, the best result in all the datasets were produced with the non-fused baseline. However, after further investigating this, it is clear that non-fused models do not often generalise well across other datasets. For example in Table 3, the non-fused model trained on AHSD only provides 0.699 Macro F1 score for OLID. However, AHSD and OLID fused model further fine-tuned on AHSD provides 0.830 Macro F1 score. This is similar to the majority of the experiments, and fused models provide better results than non-fused models in other datasets. This observation again reflects an ideal scenario in real-world applications where we want an ML model to perform well across data not specific to our environment/ platform. As we see in the results, this objective can be achieved successfully with model fusion.

(3) The Fused model outperforms the ensemble baseline in all the datasets. As shown in Table 2, model fusion approaches with and without finetunning on a particular dataset outperform the best ensemble model. For HASOC, there is a large gap between the ensemble model and fused models as the ensemble model produces only 0.670 Macro F1 score while the fused model provides 0.770 Macro F1 score. The other datasets also follow a similar pattern. This is a key observation because we have presented a fusion based approach for FL that

Fine-tuned Dataset	Fused Models				BERT-large-cased				fBERT			
					AHSD	OLID	HASOC	HATEX	AHSD	OLID	HASOC	HATEX
AHSD	AHSD	OLID	-	-	0.900±0.00	0.830±0.07	0.610±0.00	0.554±0.06	**0.921±0.00**	0.836±0.09	0.627±0.00	0.628±0.00
	AHSD	-	HASOC	-	0.778±0.14	0.627±0.00	0.637±0.00	0.607±0.02	0.776±0.04	0.722±0.00	0.632±0.00	0.677±0.05
	AHSD	-	-	HATEX	0.727±0.03	0.697±0.00	0.660±0.04	0.594±0.00	0.781±0.03	0.707±0.00	0.673±0.03	0.648±0.00
	AHSD	OLID	HASOC	-	**0.919±0.00**	0.837±0.08	**0.766±0.02**	0.636±0.00	0.915±0.00	0.835±0.08	**0.770±0.01**	0.623±0.00
	AHSD	-	HASOC	HATEX	0.705±0.06	0.674±0.00	0.595±0.03	0.565±0.00	0.734±0.03	0.704±0.00	0.643±0.00	0.643±0.00
	AHSD	OLID	-	HATEX	0.905±0.00	0.813±0.09	0.628±0.00	0.719±0.05	0.914±0.00	0.834±0.08	0.627±0.00	0.772±0.01
	AHSD	OLID	HASOC	HATEX	0.716±0.03	0.708±0.00	0.646±0.05	0.652±0.06	0.730±0.01	0.724±0.00	0.668±0.04	0.684±0.04
	Non-fused Baseline				0.926±0.01	0.699±0.03	0.630±0.05	0.586±0.06	0.931±0.01	0.743±0.03	0.682±0.04	0.606±0.06
OLID	AHSD	OLID	-	-	0.893±0.00	**0.839±0.05**	0.647±0.00	0.621±0.03	0.866±0.00	**0.837±0.03**	0.601±0.00	0.598±0.00
	-	OLID	HASOC	-	0.715±0.00	0.405±0.01	0.392±0.00	0.651±0.06	0.718±0.00	0.725±0.07	0.655±0.00	0.667±0.05
	-	OLID	-	HATEX	0.696±0.00	0.692±0.08	0.656±0.04	0.616±0.00	0.679±0.07	0.723±0.07	0.611±0.00	0.650±0.00
	AHSD	OLID	HASOC	-	0.868±0.00	0.826±0.04	0.756±0.00	0.608±0.00	0.840±0.00	0.819±0.02	0.759±0.09	0.606±0.00
	-	OLID	HASOC	HATEX	0.687±0.00	0.649±0.09	0.586±0.01	0.596±0.00	0.729±0.00	0.694±0.08	0.637±0.01	0.630±0.00
	AHSD	OLID	-	HATEX	0.847±0.00	0.812±0.04	0.642±0.00	0.751±0.09	0.861±0.00	0.831±0.03	0.615±0.00	0.752±0.01
	AHSD	OLID	HASOC	HATEX	0.713±0.00	0.777±0.00	0.672±0.07	0.699±0.08	0.708±0.08	0.793±0.00	0.682±0.08	0.707±0.09
	Non-fused Baseline				0.685±0.02	0.845±0.00	0.636±0.05	0.620±0.06	0.702±0.01	0.851±0.00	0.653±0.05	0.645±0.08
HASOC	AHSD	-	HASOC	-	0.777±0.13	0.419±0.00	0.652±0.00	0.356±0.06	0.792±0.11	0.785±0.05	0.680±0.00	0.708±0.08
	-	OLID	HASOC	-	0.147±0.00	0.707±0.05	0.656±0.00	0.220±0.07	0.717±0.00	0.734±0.05	0.683±0.00	0.673±0.04
	-	-	HASOC	HATEX	0.530±0.05	0.480±0.00	0.695±0.04	0.738±0.00	0.761±0.03	0.791±0.00	0.689±0.00	0.690±0.00
	AHSD	OLID	HASOC	-	0.864±0.00	0.812±0.05	0.763±0.08	0.624±0.00	0.805±0.00	0.801±0.00	0.754±0.07	0.635±0.00
	AHSD	-	HASOC	HATEX	0.754±0.01	0.419±0.00	0.686±0.01	0.698±0.00	0.734±0.09	0.780±0.00	0.668±0.01	0.661±0.00
	-	OLID	HASOC	HATEX	0.732±0.00	0.700±0.04	0.675±0.01	0.686±0.00	0.736±0.00	0.712±0.06	0.671±0.00	0.676±0.00
	AHSD	OLID	HASOC	HATEX	0.703±0.09	0.647±0.00	0.651±0.00	0.651±0.00	0.719±0.06	0.781±0.00	0.702±0.06	0.718±0.06
	Non-fused Baseline				0.620±0.03	0.492±0.01	0.788±0.01	0.555±0.06	0.645±0.02	0.532±0.01	0.798±0.01	0.575±0.05
HATEX	AHSD	-	-	HATEX	0.758±0.01	0.449±0.00	0.531±0.08	0.744±0.00	0.671±0.01	0.591±0.00	0.587±0.00	**0.777±0.00**
	-	OLID	-	HATEX	0.650±0.00	0.689±0.06	0.557±0.09	0.749±0.00	0.584±0.02	0.668±0.01	0.599±0.00	0.775±0.00
	-	-	HASOC	HATEX	0.538±0.01	0.545±0.0	0.710±0.05	**0.756±0.00**	0.527±0.05	0.573±0.00	0.707±0.07	0.772±0.00
	AHSD	-	HASOC	HATEX	0.692±0.04	0.529±0.00	0.693±0.05	0.741±0.00	0.636±0.10	0.588±0.00	0.688±0.08	0.767±0.00
	-	OLID	HASOC	HATEX	0.561±0.00	0.640±0.09	0.690±0.06	0.755±0.00	0.526±0.00	0.664±0.08	0.689±0.08	0.772±0.00
	AHSD	OLID	-	HATEX	0.522±0.00	0.597±0.08	0.607±0.00	0.645±0.09	0.532±0.00	0.563±0.03	0.613±0.00	0.633±0.10
	AHSD	OLID	HASOC	HATEX	0.627±0.08	0.532±0.00	0.635±0.09	0.642±0.11	0.631±0.09	0.565±0.00	0.652±0.09	0.671±0.11
	Non-fused Baseline				0.569±0.03	0.504±0.01	0.604±0.02	0.782±0.02	0.581±0.01	0.523±0.01	0.612±0.01	0.795±0.01

Table 3: Macro F1 score results for the fuse models (BERT-large-cased and fBERT) compared to the baseline systems fine-tuned on the four datasets. Results are reported on 10 runs along with standard deviation. The best results from the fused approach for each model are in bold. Results for the non-fused baseline model evaluated on the same dataset are underlined.

can surpass an ensemble based model preserving privacy across different datasets. The platforms/ environments that are interested in developing a FL approach should focus on model fusion based strategies that outperform ensemble based models as we showed in the results.

(4) The Fused model performance heavily depends on the datasets it was trained on. Our final observation is that the fused model performance depends on the datasets that it was trained on. For example, when the model fusion was performed between AHSD and OLID, the final model provided excellent results on both datasets. This is due to the general nature of these two datasets covering multiple types of offensive content rather than focusing on a particular type of offensive content. On the other hand, results are not the same when the model fusion was performed between AHSD and HASOC where the final model did not provide good results for both datasets. This can be explained by the demography of the dataset as HASOC data is collected on Twitter users based in India. It is clear that model fusion would thrive in similar kinds of datasets, but would not perform well with different kinds of data.

Overall, model fusion produces excellent results on the dataset that it was fine-tuned on, and it generalizes well across other datasets. Fused models outperform both of our baselines in all the datasets. Therefore, model fusion provides a successful approach to FL.

5.2. Multilingual Experiments

We conducted initial multilingual experiments with the same FL setting. We used OffendES (Plaza-del Arco et al., 2021), a Spanish offensive language identification dataset. For English we used the OLID dataset described before. Each instance in OffendES is labelled as belonging to one of the five classes; Offensive and targeted to a person (OFP), Offensive and targeted to a group (OFG), Offensive and not targeted to a person or a group (OFO), Non-offensive, but with expletive language (NOE), and Non-offensive (NO). We map the instances belonging to the OFP, OFG, OFO, and NOE to OLID OFF, and the NO class as NOT. Even though, the label NOE is considered non-offensive in OffendES, it contains profanity so we map it to OLID label OFF to conform with the OLID guidelines.

Instead of the monolingual BERT models we used in the previous experiments, we use cross-lingual models, specifically XLM-R (Conneau et al., 2019). We used the same FL settings and compared it with ensemble baseline. The results are shown in Table 4.

Dataset	Approach	Macro F1
English	non-fused	0.845 ±0.01
	fusion with FT	0.829 ±0.03
	fusion without FT	0.831 ±0.00
	ensemble	0.776 ±0.02
Spanish	non-fused	0.812 ±0.04
	fusion with FT	0.809 ±0.02
	fusion without FT	0.792 ±0.01
	ensemble	0.761 ±0.02

Table 4: The results for multilingual experiments on English and Spanish; non-fused models, fused models with fine-tuning (FT), fused models without finetuning and ensemble. We report the results with xlm-roberta. The results are ordered from Macro F1.

The results show that fusion based FL outperforms ensemble baseline in multilingual settings too. This opens new avenues for privacy preserving models for languages other than English and more specifically, low-resource languages.

6. Conclusion and Future Work

This paper introduced FL in the context of combining different offensive language identification models. While a recent study (Gala et al., 2023) uses FL learning in offensive language identification, their work is limited to distributed training on the same dataset with multiple clients. As far as we know, our research is the first study to use FL in combining multiple offensive language identification models. We evaluated a fusion-based FL architecture using a general BERT model and a fine-tuned fBERT model on four publicly available English benchmark datasets. We also presented initial cross-lingual experiments in English and Spanish. Our results show that the fusion model performances outperform the performance of an ensemble baseline model. We also show that the fused model generalizes well across all datasets tested. As the FL architecture does not require data sharing, we believe that FL is a promising research direction in offensive language identification due to its privacy preserving nature.

In future work, we would like to explore other FL architectures and compare their performance to the fused model proposed in this paper. Finally. we would like to evaluate the performance of recently proposed large language models (LLMs) (e.g., GPT-4, LLama 2) for this task in FL settings.

Bibliographical References

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*.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. In *Proceedings of WOAH*.

Çağrı Çöltekin. 2020. A Corpus of Turkish Offensive Language on Social Media. In *Proceedings of LREC*.

Tianshi Che, Ji Liu, Yang Zhou, Jiaxiang Ren, Jiwen Zhou, Victor Sheng, Huaiyu Dai, and Dejing Dou. 2023. Federated learning of large language models with parameter-efficient prompt tuning and adaptive optimization. In *Proceedings of EMNLP*.

Mingqing Chen, Ananda Theertha Suresh, Rajiv Mathews, Adeline Wong, Cyril Allauzen, Françoise Beaufays, and Michael Riley. 2019. Federated learning of n-gram language models. In *Proceedings of CoNLL*.

Leshem Choshen, Elad Venezian, Noam Slonim, and Yoav Katz. 2022. Fusing finetuned models for better pretraining. *arXiv preprint arXiv:2204.03044*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

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*.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of WWW*.

Oluwaseyi Feyisetan, Sepideh Ghanavati, Patricia Thaine, Ivan Habernal, and Fatemehsadat Mireshghallah, editors. 2022. *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*. ACL.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2017. Legal Framework, Dataset and Annotation Schema for Socially Unacceptable On-line Discourse Practices in Slovene. In *Proceedings ALW*.

Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. 2019. A Hierarchically-labeled Portuguese Hate Speech Dataset. In *Proceedings of ALW*.

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

Jay Gala, Deep Gandhi, Jash Mehta, and Zeerak Talat. 2023. A federated approach for hate speech detection. In *Proceedings of EACL*.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using Convolutional Neural Networks to Classify Hate-speech. In *Proceedings of ALW*.

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*.

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*.

Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. 2020. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 22(3):2031–2063.

Bill Yuchen Lin, Chaoyang He, Chulin Xie, Fatemehsadat Mireshghallah, Ninareh Mehrabi, Tian Li, Mahdi Soltanolkotabi, and Xiang Ren, editors. 2022a. *Proceedings FL4NLP*. ACL.

Bill Yuchen Lin, Chaoyang He, Zihang Ze, Hulin Wang, Yufen Hua, Christophe Dupuy, Rahul Gupta, Mahdi Soltanolkotabi, Xiang Ren, and Salman Avestimehr. 2022b. Fednlp: Benchmarking federated learning methods for natural language processing tasks. In *Findings of NAACL*.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

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*, 30:1–16.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. 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*.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of AISTATS*.

Joshua Melton, Arunkumar Bagavathi, and Siddharth Krishnan. 2020. Del-hate: a deep learning tunable ensemble for hate speech detection. In *Proceedings of ICMLA*.

Sandip Modha, Thomas Mandl, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Tharindu Ranasinghe, and Marcos Zampieri. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech. In *Proceedings of FIRE*.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2021. Arabic Offensive Language on Twitter: Analysis and Experiments. In *Proceedings of WANLP*.

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

Md Nishat Raihan, Umma Tanmoy, Anika Binte Islam, Kai North, Tharindu Ranasinghe, Antonios Anastasopoulos, and Marcos Zampieri. 2023. Offensive language identification in transliterated and code-mixed bangla. In *Proceedings of BLP*.

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

Tharindu Ranasinghe and Marcos Zampieri. 2021. MUDES: Multilingual Detection of Offensive Spans. In *Proceedings of NAACL*.

Tharindu Ranasinghe, Marcos Zampieri, and Hansi Hettiarachchi. 2019. BRUMS at HASOC 2019: Deep Learning Models for Multilingual Hate Speech and Offensive Language Identification. In *Proceedings of FIRE*.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. 2021. Adaptive federated optimization. In *Proceedings of ICLR*.

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 ACL*.

Anit Kumar Sahu, Tian Li, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith. 2019. Federated optimization for heterogeneous networks. In *Proceedings of AMTL*.

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

Shrey Satapara, Prasenjit Majumder, Thomas Mandl, Sandip Modha, Hiren Madhu, Tharindu Ranasinghe, Marcos Zampieri, Kai North, and Damith Premasiri. 2022. Overview of the hasoc subtrack at fire 2022: Hate speech and offensive content identification in english and indo-aryan languages. In *Proceedings of FIRE*.

Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. 2023. Fedperc: Federated learning for language generation with personal and context preference embeddings. In *Findings of EACL*.

Yansheng Wang, Yongxin Tong, Dingyuan Shi, and Ke Xu. 2021. An efficient approach for crosssilo federated learning to rank. In *Proceedings of ICDE*.

Tharindu Weerasooriya, Sujan Dutta, Tharindu Ranasinghe, Marcos Zampieri, Christopher Homan, and Ashiqur Khudabukhsh. 2023. Vicarious offense and noise audit of offensive speech classifiers: Unifying human and machine disagreement on what is offensive. In *Proceedings of EMNLP*.

Bingjie Yan, Jun Wang, Jieren Cheng, Yize Zhou, Yixian Zhang, Yifan Yang, Li Liu, Haojiang Zhao, Chunjuan Wang, and Boyi Liu. 2021. Experiments of federated learning for covid-19 chest x-ray images. In *Proceedings of ICAIS*.

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, 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*.

Zhuo Zhang, Xiangjing Hu, Jingyuan Zhang, Yating Zhang, Hui Wang, Lizhen Qu, and Zenglin Xu. 2023. Fedlegal: The first real-world federated learning benchmark for legal nlp. In *Proceedings of ACL*.

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*.