

# FEDERATED LEARNING BASED MULTILINGUAL EMOJI PREDICTION IN CLEAN AND ATTACK SCENARIOS

Karim Gamal, Ahmed Gaber, Hossam Amer

Queen's University, Microsoft

{21kgmm, 21amga}@queensu.ca, hossamamer@microsoft.com

## Abstract

Federated learning is a growing field in the machine learning community due to its decentralized and private design. Model training in federated learning is distributed over multiple clients giving access to lots of client data while maintaining privacy. Then, a server aggregates the training done on these multiple clients without access to their data, which could be emojis widely used in any social media service and instant messaging platforms to express users' sentiments. This paper proposes federated learning-based multilingual emoji prediction in both clean and attack scenarios. Emoji prediction data have been crawled from both Twitter and SemEval emoji datasets. This data is used to train and evaluate different transformer model sizes including a sparsely activated transformer with either the assumption of clean data in all clients or poisoned data via label flipping attack in some clients. Experimental results on these models show that federated learning in either clean or attacked scenarios performs similarly to centralized training in multilingual emoji prediction on seen and unseen languages under different data sources and distributions. Our trained transformers perform better than other techniques on the SemEval emoji dataset in addition to the privacy as well as distributed benefits of federated learning.

## 1 Introduction

Federated learning FL is an emerging field in machine learning first introduced by Google in 2017 (McMahan et al., 2017). FL is a distributed learning framework, where multiple distributed clients collaborate in training machine learning models under the orchestration of a central server. This server aggregates the trained models into a final global model. FL is a major shift from centralized machine learning to a distributed manner that uses many distributed computing resources such as devices and data. Policies such as General Data Protection Regulation (GDPR) (Regulation, 2018)

imposes data privacy rules among different organizations. Thus, FL is essential to enhance data privacy by keeping the raw data on the local device while taking into account that some clients may have a Non-IID distribution of data or poisoned data limiting federated learning accuracy.

Federated learning is used in many natural language processing applications (Li et al., 2020; Singh et al., 2022) such as adapting to pedestrian behavior in data generated by distributed sensors and devices, enabling real-time analysis and decision-making. In this paper, we focus on the task of emoji prediction because emojis enhance communication quality among users and associated text data is private. For example, the use of emojis in tweets has been steadily increasing over the years, with 35% of tweets in 2019 (McShane et al., 2021) containing at least one emoji, compared to 9.9% in 2012. This trend has led to a 25% increase in engagement for tweets that feature at least one emoji (McShane et al., 2021), as compared to those without any emojis. Moreover, the use of emojis in combination with brand names has increased by 49% since 2015, (Agnew, 2017).

While traditional next-word prediction models are limited by the relatively narrow range of options available, emoji prediction presents unique challenges due to the diverse and context-dependent nature of emoji usage. For example, when predicting the next word after 'you' in the sentence 'Thank you so much', the options are relatively limited and predictable, such as 'guys' or 'so'. However, for the next emoji prediction after 'you', the options are more diverse and context-dependent, ranging from different colored hearts to facial expressions.

Reviewing the literature on emoji prediction, we can find that several papers indeed proposed methods in either centralized or federated settings (Barbieri et al., 2018b; Weller et al., 2022; Ramaswamy et al., 2019; Gandhi et al., 2022; Lee et al., 2022; Tomihira et al., 2020; Peng and Zhao,

2021; Barbieri et al., 2018c; Yang et al., 2018; Barbieri et al., 2020; Venkit et al., 2021; Edwards et al., 2020; Camacho-Collados et al., 2022; Barbieri et al., 2022; Loureiro et al., 2022; Caldarola et al., 2022). For example, SemEval shared task in (Barbieri et al., 2018b) is dedicated to emoji prediction where multiple methods were developed in a centralized setting with a maximum F1 score of 35.99%. In this shared task, methods did not consider the widely used transformer architecture, multilinguality, and federation. In addition, Google successfully implemented a federated learning solution for Gboard’s emoji prediction using the LSTM architecture in (Ramaswamy et al., 2019). However, the results in this paper were only shown in English with 100 emoji classes while assuming that all clients have clean data, i.e. clean scenarios. Additionally, the findings presented in (Weller et al., 2022) indicate that multilingual federated learning can be achieved without significant performance degradation compared to centralized learning. However, the study did not evaluate the performance of their federated learning models in the case of unseen languages or poisoned data from some clients (i.e., attack scenarios). Notably, the study did not include sparsely activated MoE transformers (Kim et al., 2021), which showed success in NLP tasks. Although previous work by (Gandhi et al., 2022) achieved good results in predicting emojis in Hindi tweets using both centralized and federated learning approaches, our research aims to expand upon this approach by applying it to a multilingual context. It is worth noting that their study did not explore the effects of sparsely activated MoE transformers or evaluate the performance of their federated models on unseen languages or under poisoned data attack scenarios.

In a parallel research line, there were studies that investigated FL in an attack scenario, and proposed techniques to defend against label-flipping attacks and backdoor attacks (Lyu et al., 2020; Rodríguez-Barroso et al., 2022; Blanchard et al., 2017; Wang et al., 2020a; Fung et al., 2020; Jebreel et al., 2022; Manoel et al., 2022; Ma et al., 2020). However, their experiments were not carried out on the task of multilingual emoji prediction with multiple classes, and the results were not compared to the centralized setting.

This paper proposes federated learning-based multilingual emoji prediction in both clean and

attack scenarios<sup>1</sup>. Our two million training and testing examples are acquired from both Twitter and the standard SemEval emoji dataset (SemEval Data). For multilingual emoji prediction, we train publicly available pre-trained models of different sizes of dense and sparsely activated transformers, namely, Multilingual-MiniLM (M-MiniLM) (Wang et al., 2020b), Twitter-twihi-Bert-base (Bert-Base) (Zhang et al., 2022), Twitter-XLM-Roberta (XLM-R) (Barbieri et al., 2022), and switch-MoE with 8 experts (Fedus et al., 2021). To simulate attack scenarios in FL, we apply the label-flipping data poisoning attack to some clients and utilize different FL aggregation schemes to reverse this attack. Our experimental results led to the following findings:

- In either centralized or FL experiments, we achieved emoji prediction accuracy better than the teams reported in the SemEval emoji prediction shared task (Barbieri et al., 2018b).
- FL training on emoji data achieves similar accuracy performance to traditional centralized setup. This FL accuracy performance is confirmed in seen or unseen languages with IID and Non-IID data distributions in both unilingual and multilingual settings.
- When some clients’ emoji data is attacked via label flipping, FL’s K-representative unweighted median (Krum) aggregation scheme can restore the accuracy of the clean setting.

The rest of the paper is organized as follows: Section 2 describes the data, models, training algorithm, and label-flipping attack for FL. This is followed by explaining experiments carried out in this paper. Last but not least, Section 4 concludes the paper.

## 2 Methodology

### 2.1 Data Acquisition

We used the **Twitter API** to crawl over 2 million tweets that **contain only one emoji as well as 500k training and test data from SemEval**. The Twitter set encompasses 3 languages namely, Spanish, Italian, and French, while SemEval includes English. This data is filtered to remove stop words, hyperlinks, and duplicate special characters resulting in 1.3 million examples. Given an input sentence, we focus on predicting the 20 most popular emojis,

<sup>1</sup>Demo and Source code of this paper on GitHub ( [FEDERATED-LEARNING-BASED-MULTILINGUAL](#) )

which are the same emojis as the SemEval paper (Barbieri et al., 2018b) shown in Figure 1.



Figure 1: The 20 most frequent emojis.

## 2.2 Client Partitioning

Three different training setups were carried out in this paper: traditional centralized training with no FL<sup>2</sup>, FL with IID data where each client has a random subset of all data, and FL with Non-IID data where each client includes data from one language. The number of clients for FL experiments is four. We have selected four clients for our FL experiments based on previous experiments and to maintain consistency with the Non-IID setup of our problem. This decision is also influenced by the fact that we initially started our work with four languages. In either FL IID or Non-IID, we carry out an experiment while assuming that all clients have clean data (clean scenario) and another experiment while assuming that 25% and 50% of the clients have label-flipped data (attack scenario). For centralized training with no FL, we merge the clients' datasets into one pool and use label flipping to attack the same set of samples as FL to compare the results.

## 2.3 Attack Scenario: Label flipping Procedure

We flipped the first 10 emoji classes in Figure 1 into the last 10 emoji. Based on this flipping, we created two attack scenarios. The first scenario is to make 1 out of 4 FL clients toxic (i.e., 25%), while the second scenario is to make 2 out of 4 of FL clients toxic (i.e. 50%). We then apply these datasets to federated training (IID setup, and Non-IID setup) and centralized training as described in Section 2.2.

To elaborate on the IID and Non-IID setup, we divided the data into four parts for the IID setup, with each client taking one part. For the 25% FL clients toxic attack scenario, we flipped the first 10 emoji classes to create a toxic dataset for one of the four clients.

For the Non-IID setup, since we had four languages for training, we divided the data per language and assigned each client one language to

<sup>2</sup>In practical scenarios, publicly available datasets such as Twitter data or datasets similar to the private chat data can be used to train the server model for FL. This is particularly useful in chat problems where users may be hesitant to share their private chat data with the application.

work on. In this scenario, we made the English data toxic for the 25% FL clients attack scenario, while for the 50% FL clients attack scenario, we made both the English and French languages toxic. This approach allowed us to simulate a more realistic attack scenario, where the toxic data distribution is not uniform across all clients.

## 2.4 Federated Learning Methods

Federated Averaging (FedAVG) (McMahan et al., 2017) is an aggregation scheme used in federated learning, where each client trains a model on its local data using stochastic gradient descent (SGD), and the server aggregates the client weights by taking their average. This process is repeated for a fixed number of rounds, with each round consisting of clients training their models and the server aggregating their weights. On the other hand, Krum (Blanchard et al., 2017) is a robust federated learning algorithm that selects a subset of model updates, excluding those from malicious or incorrect clients, and computes the K-representative unweighted median of the selected updates. The selected update is the one that is closest to the center of the other updates while ignoring the updates that are far away from the center. The parameter K is specified by the user.

Krum (Krumble) is more computationally demanding than FedAvg (Federated Averaging) because it involves additional steps of distance calculation to select the best model updates.

In Krum, the server selects the best model updates from a subset of the participating devices (clients) based on the distances between the updates. The distances are calculated based on the number of disagreements between the updates and the other updates in the subset. This process requires more computation than FedAvg, where the server simply averages the model updates received from the clients.

Thus, the additional distance calculation step in Krum makes it more computationally demanding than FedAvg. However, Krum may provide better performance in certain scenarios where the participating clients may be potentially malicious or have poor quality updates.

Although, Krum can produce a more precise final model in attack scenarios where FedAVG may not be effective.

## 2.5 Models

Publicly available models with different numbers of model parameters are trained for the task of emoji prediction. In particular, we utilized 4 models from hugging face: (1) M-MiniLM (Wang et al., 2020b) is a 21M parameter transformer model pre-trained on 16 languages and distilled from Bert Base; (2) Bert-Base is a multi-lingual tweet language model (Zhang et al., 2022) that is trained on 7 billion Tweets from over 100 distinct languages and has 280M parameters; (3) XLM-R (Babieri et al., 2022) is trained on 198M multilingual tweets, has 278M parameters, and pre-trained in more than 30 Languages; (4) switch-MoE is an 8-expert MoE model trained on Masked Language Modeling (MLM) task. The model architecture is similar to the classic T5, but with the Feed Forward layers replaced by the Sparse MLP layers containing "experts" MLP. It has 619M parameters and it's pre-trained in the English-only language. In addition to the publicly available models, we also built an LSTM model from scratch for the emoji prediction task. The model has 18M parameters and consists of a 1D convolutional layer and 3 LSTM layers.

## 2.6 Training Description

We utilized the Flower framework for both federated training and evaluation, given its user-friendly interface and active community (Beutel et al., 2020). Additionally, we employed Hugging Face's transformers library (Wolf et al., 2019) to load pre-trained models, and PyTorch as the underlying differentiation framework (Paszke et al., 2019). We conducted training for each sparsely and densely activated transformer model for 30 epochs with the AdamW optimizer and (1e-3, 1e-4) learning rates. Our experiment runs in about 15-20 hours for dense transformers, while it takes a day for MoE training. For FL, we assigned four clients and conducted five rounds of training, with each client training for one epoch per round for both the Non-IID and IID setups. All these experiments are carried out on 256 GB of RAM and two NVIDIA A40 GPUs, an Intel(R) Xeon(R) Gold 6338 machine.

## 3 Experimental Results

Our experimental design consists of three stages:

1. Train the models under test in a centralized setting on the task of emoji prediction. This setting is **Baseline**, which is done to ensure

that the models under test are well-equipped for the task of emoji prediction before federated and centralized comparisons.

2. Take the trained models and train them in FL with a new dataset distributed to clients. In this setup, we mainly carry out an experiment while assuming that all clients have clean data (clean scenario) and another experiment while assuming that some clients have data that has been label-flipped (attack scenario). In both scenarios, we carry out experiments for both **IID** and **Non-IID** FL to simulate real scenarios. For further clarification, see Figure 2.

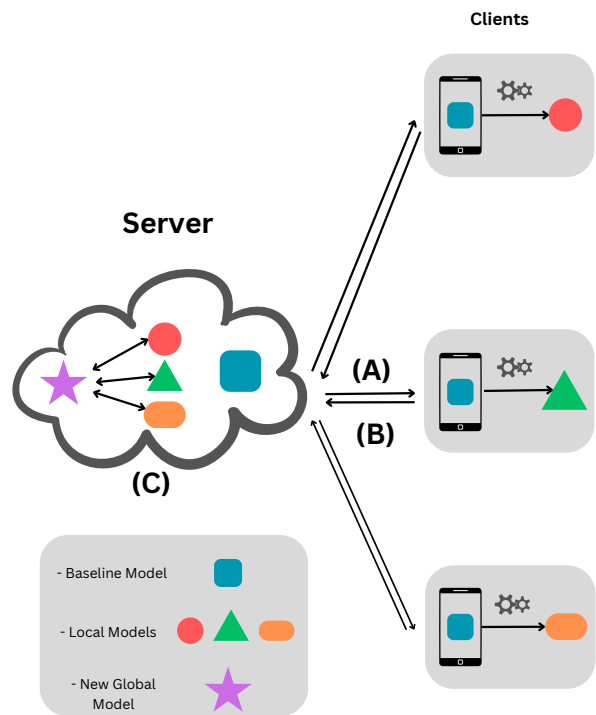


Figure 2: Second stage of our experimental design. (A) shows how the server pushes the baseline model to the clients for training on their local dataset. (B) demonstrates how the clients push back the tuned model to the server. Finally, (C) represents the process by which the server combines all models from the clients to build a new global model.

3. Take the trained models from the first stage and fine-tune these models in a centralized setting on the combination of all the distributed data used to train the FL. The third stage is the **Finetuned** setting, which is done to compare the performance of the federated learning distributed learning approach versus the traditional centralized learning approach in the task of emoji prediction. Our training set is



divided into two halves. The first half of the data is used for the Baseline stage, while the second is used in either the FL stage (both IID and non-IID) or the Finetuned stage. We carry out all these experiments in unilingual and multilingual setups while measuring the corresponding Macro-F1 score for the 20-class emojis.

Macro-F1 and Micro-F1 are two evaluation metrics used in multi-class classification problems. The Micro-F1 score gives equal weight to each individual instance in the dataset, while the Macro-F1 score gives each class an equal weight. In cases where we have unbalanced classes, the Macro-F1 score is often used as it gives equal importance to each class, regardless of its frequency in the dataset.

### 3.1 Experimenting with the Number of Clients in Federated Learning

We present the results of our initial experiments to determine the optimal number of clients in federated learning for our problem. We used an LSTM IID multilingual model and evaluated its performance on: Micro-F1 and Macro-F1 scores.

Metrics	Baseline	IID	Finetuned
Accuracy	43.4%	45.1%	44.8%
Macro-F1	27.6%	29.2%	29.1%

Table 1: Multilingual Centralized and Federated with 4 clients Accuracy and Macro-F1 Scores for SemEval test dataset.

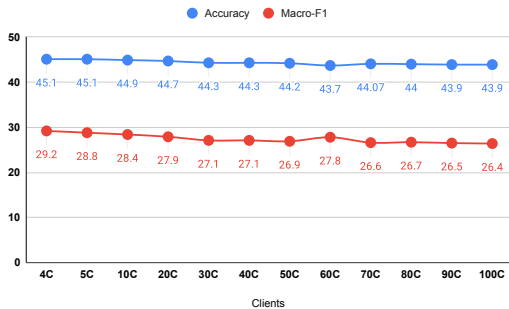


Figure 3: SemEval Test Results of LSTM IID Multilingual experiments with varying clients.

As shown in Figure 3, we observe that FL shows consistent performance across different number of clients in terms of Accuracy and Macro-F1 scores. There are slight drops in Macro-F1 because we use a fixed dataset distributed into multiple clients lead-

ing to a smaller data share per client. Smaller data may yield lower scores, which was also observed in the literature (Caldarola et al., 2022).

So our initial experiments suggest that federated learning can improve the performance of our model, and a smaller number of clients might be more appropriate for our problem.

However, we can obtain higher scores by using transformer models, which is what we will do in the upcoming experiments.

### 3.2 Unilingual and Multilingual Macro-F1 results in clean scenario

Table 2 presents the results of the unilingual models for the SemEval English test dataset. As can be seen, The Finetuned models generally perform better than their baseline and IID counterparts. For example, the Bert-Base model has a baseline score of 36.9%, while the IID score is 37.4%, and the Finetuned score is 38.1%. This shows that while the Finetuned process improves the model’s performance, the IID training approach is also effective and can achieve a score close to the Finetuned model.

Model	Baseline	IID	Finetuned
Switch-Base-8	33.2%	37.3%	36.6%
Bert-Base	36.9%	37.4%	38.1%
XLM-R	35.9%	36.7%	37.6%
M-MiniLM	33.3%	33.9%	35.9%

Table 2: Unilingual Centralized and Federated Macro-F1 Accuracy Scores for SemEval test dataset.

Turning to *multilingual*, Table 3 presents the Macro-F1 accuracy scores for centralized and FL multilingual models on the SemEval English dataset. The baseline accuracy for all models is improved by applying federated learning with the IID, which achieves slightly higher accuracy than the Non-IID setting. This suggests that data distribution has an impact on model performance. Additionally, finetuning the models further improves the Baseline performance with only small differences observed between federated and Finetuned results.

Model	Baseline	IID	Non-IID	Finetuned
Bert-Base	35.3%	36.8%	36.3%	35.7%
XLM-R	33.4%	34.9%	34.9%	34.1%
M-MiniLM	31.4%	32.7%	32.03%	32.9%

Table 3: Multilingual Centralized and Federated Macro-F1 Accuracy Scores for SemEval English Dataset.

When comparing the unilingual and multilingual models via SemEval, we can see a drop in the

Macro-F1 scores due to the mixed languages in the multilingual dataset. However, the drop was not significant, indicating that the models can handle multiple languages to a certain extent. Overall, the federated approach has shown to be a viable alternative to the centralized approach in terms of performance.

Using the Twitter multilingual dataset, Table 4 outlines the trained *multilingual* Macro-F1 scores in centralized and FL setups. Similar observations to the case of SemEval are seen.

Model	Average Test Results for Twitter Dataset			
	Baseline	IID	Non-IID	Finetuned
Bert-Base	29.2%	30.6%	30.7%	30.2%
XLM-R	26.5%	28.2%	28.02%	28.3%
M-MiniLM	24.2%	25.7%	26.1%	25.9%

Table 4: Centralized and Federated Learning average results for the Twitter Multilingual (e.g Spanish, French, Italian) Dataset.

In Table 5, we show the Macro-F1 score per language of the best-performing multilingual model Bert-Base in the centralized and federated setups. As shown, the accuracy does not significantly vary per language, which shows the effectiveness of our trained models.

Data	Baseline	IID	Non-IID	Finetuned
Spanish	27.7%	28.5%	28.8%	27.3%
French	29.6%	31.3%	31.1%	30.9%
Italian	30.2%	32.2%	32.3%	32.4%

Table 5: Centralized and Federated Learning Results for the Twitter multilingual dataset for Bert-Base model.

### 3.3 Comparison between our models and the literature

Looking at the results in Table 6, we can observe that most of our multilingual models perform better than the majority of models that were also trained on the SemEval training dataset from the literature in terms of both Micro-F1 and Macro-F1. Specifically, our multilingual models in most cases achieved more than 36% Macro-F1 or more than 49% Micro-F1, whereas the best-performing model from the literature, BERT(Twitter) (Edwards et al., 2020), achieved 40% Micro-F1. Furthermore, our unilingual model with Bert-Base achieved 38.1% Macro-F1, which is comparable to the performance of BERT(Twitter), which achieved 38% Macro-F1. Moreover, our federated models achieved more than 50% Micro-F1, which is better than the performance of BERT(Twitter). Following the Large

Language Models interest, we carried out an experiment using the Davinci-003 model (Brown et al., 2020) on the SemEval set in zero-shot. Davinci-003 achieved a Macro-F1 score of 16%, which also shows the promise of our trained FL models.

Model	Micro-F1	Macro-F1
BiLSTM (Venkit et al., 2021)	29.6%	21.3%
Proposed LSTM IID Multilingual	45.1%	29.2%
XLM-Tw (Barbieri et al., 2022)	-	30.9%
TweetNLP (Camacho-Collados et al., 2022)	-	34.0%
SemEval first team (Barbieri et al., 2018a)	47.1%	35.9%
BERT (Twitter) (Edwards et al., 2020)	40.0%	38.0%
Proposed Bert-Base Finetuned Multilingual	49.4%	35.7%
Proposed Bert-Base FL Non-IID Multilingual	49.5%	36.3%
Proposed Bert-Base FL IID Multilingual	50.3%	36.8%
Proposed Bert-Base FL IID Unilingual	50.9%	37.4%
Proposed Bert-Base Finetuned Unilingual	50.1%	38.1%

Table 6: Comparison between our models’ performance and models from the literature on emoji prediction task using the Micro-F1 and Macro-F1 metrics.

### 3.4 Multilingual Macro-F1 Results on unseen language

To investigate the trained models’ performance on unseen languages (i.e, zero-shot), we run inference on an unseen German dataset in Baseline, FL, and Finetune settings. Table 7 shows that there is some drop in performance due to the zero-shot setting. However, this experiment still shows that FL performs at least similarly to centralized settings even in unseen languages.

Model	Baseline	IID	Non-IID	Finetuned
Bert-Base	21.9%	23.1%	23.1%	21.5%
XLM-R	20.04%	20.9%	21.07%	19.2%
M-MiniLM	15.4%	16.5%	16.4%	17.1%

Table 7: The zero-shot inference results for Centralized and Federated Learning

### 3.5 Multilingual Macro-F1 Results in Label-flipping Attack Scenario

Table 8 and 9 depict the results of label flipping experiments when 50% clients are attacked (i.e 50% of the data is attacked) using the SemEval

and Twitter datasets, respectively. The tables compare the centralized and FL results for FedAVG and Krum. The results demonstrate that Krum performed better than FedAVG in both datasets, with higher accuracy rates. Krum was able to handle the label-flipping attack scenario and produced scores that were very close to the results obtained with the Finetune setting.

Table 8 presents the experiment results for three different models, namely Bert-Base, XLM-R, and M-MiniLM. However, we will focus on the results of the Bert-Base model in Figure 4, where Bert achieved 36.8% Macro-F1 in the clean IID scenario but dropped to 26.2% under FedAVG with Fed-IID due to label-flipping attacks. Traditional training and FL with FedAVG had low Macro-F1 of 24.4%, while Krum aggregation function achieved 36.5% Macro-F1, showing superior handling of label-flipping attacks and improving FL model performance. Appendix A shows similar results for 25% Toxic clients experiments.

Model	Setting	FedAVG	Krum
Bert-Base	Fed-IID	26.2%	36.5%
	Fed-Non-IID	28.1%	35.2%
	Finetuned	24.4%	
XLM-R	Fed-IID	FedAVG	Krum
	Fed-Non-IID	27.6%	34.8%
	Finetuned	27.5%	32.9%
		23.3%	
M-MiniLM	Fed-IID	FedAVG	Krum
	Fed-Non-IID	25.3%	32.6%
	Finetuned	26.7%	30.5%
		23.9%	

Table 8: Centralized and Federated Learning Results in Label-Flipping Attack Scenario for the SemEval English test dataset.

Model	Setting	FedAVG	Krum
Bert-Base	Fed-IID	21.8%	30.1%
	Fed-Non-IID	23.6%	29.4%
	Finetuned	21.8%	
XLM-R	Fed-IID	FedAVG	Krum
	Fed-Non-IID	20.4%	27.3%
	Finetuned	21.5%	26.6%
		20.8%	
M-MiniLM	Fed-IID	FedAVG	Krum
	Fed-Non-IID	19.02%	24.9%
	Finetuned	20.9%	23.9%
		20.3%	

Table 9: Centralized and Federated Learning Results in Label-Flipping Attack Scenario for the average results for the Twitter multilingual dataset.

Similar to the experiment in Table 8. Table 9 shows that Krum outperforms FedAVG in both the Non-IID and IID settings as well. Also in Table 10

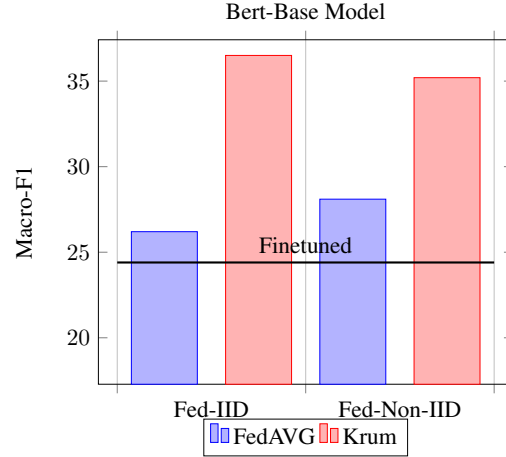


Figure 4: Centralized and Federated Learning Results in Label-Flipping Attack Scenario for the SemEval English test dataset for Bert-Base Model.

reports the results of applying label flipping to the German Zero-shot scenario,

Model	Setting	FedAVG	Krum
Bert-Base	Fed-IID	15.3%	23.5%
	Fed-Non-IID	15.7%	21.3%
	Finetuned	16.3%	
XLM-R	Fed-IID	FedAVG	Krum
	Fed-Non-IID	14.1%	20.9%
	Finetuned	14.04%	19.8%
		14.3%	
M-MiniLM	Fed-IID	FedAVG	Krum
	Fed-Non-IID	11.3%	16.2%
	Finetuned	11.9%	14.3%
		11.7%	

Table 10: Centralized and Federated Learning Results in Label-Flipping Attack Scenario for the German Zero-shot.

### 3.6 Comparing Results of Clean and Attack Scenarios using Various Aggregation Functions.

Table 11 provides valuable insights into the performance of a Bert-Base model in a federated multilingual setting under both clean and attack scenarios. The results highlight the impact of toxic clients on the federated setting, showing a decrease in performance for FedAVG, Fed-IID, and Fed-Non-IID. This decrease in performance is particularly evident when the toxic data constitutes 50% of the overall data. However, the use of the Krum aggregation function can mitigate this drop in performance, as previously observed in literature.

Overall, this table serves to provide a clear comparison of the performance under different scenarios, which can be useful for identifying the most

suitable aggregation function for a given federated learning scenario.

Setting		Results	
Cleaned FedAVG	Fed-IID	36.8	
	Fed-Non-IID	36.3	
FedAVG	Fed-IID	25% Toxic	50% Toxic
	Fed-Non-IID	36.3	26.2
Krum	Fed-IID	33.5	28.1
	Fed-Non-IID	36.5	36.5
Krum	Fed-IID	35.3	35.2
	Fed-Non-IID	35.3	35.2

Table 11: Macro-F1 Results of Bert-Base Model in Federated Multilingual Setting under Clean and Attack Scenarios for the SemEval English test dataset.

### 3.7 Computational Overhead Comparison between Krum and FedAVG

The exact computational overhead of the Krum algorithm compared to FedAvg or other aggregation algorithms depends on several factors, such as the number of participating clients, the size of the models, and the specific implementation of the algorithms. In some cases, the computational overhead of the Krum algorithm may be similar to that of FedAvg or even lower, depending on the specific scenario. Table 12 shows that the Fed-Non-IID algorithm takes longer than the Fed-IID algorithm due to the varying sizes of language datasets among clients. The Krum algorithm takes  $\sim 10\%$  more time than FedAVG in our experiments.

Setting		Time (Hours)
FedAVG	Fed-IID	$\sim 18$
	Fed-Non-IID	$\sim 22$
Krum	Fed-IID	$\sim 20$
	Fed-Non-IID	$\sim 24$

Table 12: Comparison of computational overhead between FedAVG and Krum algorithms in terms of training time for each federated learning experiment.

To estimate the communicated payload for each client in our federated learning model, the total size of model parameters needs to be calculated. Assuming that the local model is the M-MiniLM model of size 0.47 GB, and there are 4 clients participating in each round, then each client needs to transmit 0.47 GB of data to the server during each round of training. Since there are 5 rounds in total, the total amount of data transmitted per client for one epoch would be 2.35 GB (0.47 GB  $\times$  5). This is an approximate estimate and does not take into account factors such as compression techniques or network latency. The actual resource constraints

may vary depending on these factors.

Table 13 shows the estimated amount of data transmitted during one epoch of federated learning using different models. For the M-MiniLM model, the estimated amount of data transmitted per client for one epoch would be 2.35 GB, which is within the resource constraints for modern devices.

Model Name	Model Size	Data Transmitted
Switch-Base-8	1.24 GB	$1.24 \times 5 = 6.2$ GB
Bert-Base	1.12 GB	$1.12 \times 5 = 5.6$ GB
XLNet	1.11 GB	$1.11 \times 5 = 5.55$ GB
M-MiniLM	0.47 GB	$0.47 \times 5 = 2.35$ GB

Table 13: Estimated amount of data transmitted per client during one round of FL.

## 4 Conclusion

This paper proposes federated learning-based multilingual emoji prediction in clean and attack scenarios. Different transformer models with varying sizes are trained in centralized and federated for which we compare their corresponding emoji prediction accuracy. Our experiments were carried out in seen and unseen languages using different data sources and distributions. Due to federated performance, federated learning can act as a substitute for centralized settings to gain privacy and access to multiple data sources benefits. In addition, we showed that our federated learning performance is competitive with the SemEval shared task on multilingual emoji prediction. In the future, we wish to explore how to achieve similar accuracy performance while considering the communication efficiency of federated learning (Passban et al., 2022). We also believe that more active research in federated learning user personalization (Arivazhagan et al., 2019) given the subjectivity of emojis can be investigated.

## Acknowledgements

We sincerely thank our invaluable contributors for their unwavering support during the entire project. First and foremost, we thank Dr. Mohamed Afify, Principal Applied Scientist at Microsoft Advanced Technology Lab, for his continuous guidance and encouragement. Dr. Mona Farouk for her diligent supervision and contributions to the DEBI program. We extend our thanks to Dr. Yuanzhu Chen, Professor at Queen’s University, for generously providing us with his workstation. In addition, we would like to thank Dr. Muhammad Jabreel for his valuable



suggestions on attack scenarios. Moreover, we express our appreciation to Abdelrahman ElHamoly for his invaluable assistance in the initial phase. Finally, we would like to extend our thanks to Orion Weller for his assistance and prompt responses to our inquiries about federated learning.

## References

- P Agnew. 2017. Emoji report.
- Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. 2019. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*.
- Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266.
- Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.
- Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018a. Semeval 2018 task 2: Multilingual emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 24–33.
- Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa-Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018b. SemEval-2018 Task 2: Multilingual Emoji Prediction. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, United States. Association for Computational Linguistics.
- Francesco Barbieri, Luís Marujo, Pradeep Karuturi, and William Brendel. 2018c. Multi-task emoji learning. In *Wijeratne S, Kiciman E, Saggion H, Sheth A, editors. Proceedings of the 1st International Workshop on Emoji Understanding and Applications in Social Media (Emoji 2018) co-located with the 12th International AAAI Conference on Web and Social Media (ICWSM 2018); 2018 Jun 25; Stanford, CA.[Aachen]: CEUR; 2018*. CEUR Workshop Proceedings.
- Daniel J. Beutel, Taner Topal, Akhil Mathur, Xinchu Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, and Nicholas D. Lane. 2020. [Flower: A friendly federated learning research framework](#).
- Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. 2017. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems*, 30.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Debora Caldarola, Barbara Caputo, and Marco Ciccone. 2022. Improving generalization in federated learning by seeking flat minima. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII*, pages 654–672. Springer.
- Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa-Anke, Fangyu Liu, Eugenio Martínez-Cámara, et al. 2022. Tweetnlp: Cutting-edge natural language processing for social media. *arXiv preprint arXiv:2206.14774*.
- Aleksandra Edwards, Jose Camacho-Collados, Hélène De Ribaupierre, and Alun Preece. 2020. Go simple and pre-train on domain-specific corpora: On the role of training data for text classification. In *Proceedings of the 28th international conference on computational linguistics*, pages 5522–5529.
- William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23:1–40.
- Clement Fung, Chris JM Yoon, and Ivan Beschastnikh. 2020. The limitations of federated learning in sybil settings. In *RAID*, pages 301–316.
- Deep Gandhi, Jash Mehta, Niral Parekh, Karan Waghele, Lynette D’Mello, and Zeerak Talat. 2022. A federated approach to predicting emojis in hindi tweets. *arXiv preprint arXiv:2211.06401*.
- Najeeb Moharram Jebreel, Josep Domingo-Ferrer, David Sánchez, and Alberto Blanco-Justicia. 2022. Defending against the label-flipping attack in federated learning. *arXiv preprint arXiv:2207.01982*.
- Young Jin Kim, Ammar Ahmad Awan, Alexandre Muzio, Andres Felipe Cruz Salinas, Liyang Lu, Amr Hendy, Samyam Rajbhandari, Yuxiong He, and Hany Hassan Awadalla. 2021. Scalable and efficient moe training for multitask multilingual models. *arXiv preprint arXiv:2109.10465*.
- SangEun Lee, Dahye Jeong, and Eunil Park. 2022. Multiemo: Multi-task framework for emoji prediction. *Knowledge-Based Systems*, 242:108437.
- Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60.

- Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Timelms: Diachronic language models from twitter. *arXiv preprint arXiv:2202.03829*.
- Lingjuan Lyu, Han Yu, and Qiang Yang. 2020. Threats to federated learning: A survey. *arXiv preprint arXiv:2003.02133*.
- Chuan Ma, Jun Li, Ming Ding, Howard H Yang, Feng Shu, Tony QS Quek, and H Vincent Poor. 2020. On safeguarding privacy and security in the framework of federated learning. *IEEE network*, 34(4):242–248.
- Andre Manoel, Mirian Hipolito Garcia, Tal Baumel, Shize Su, Jiale Chen, Dan Miller, Danny Karmon, Robert Sim, and Dimitrios Dimitriadis. 2022. Federated multilingual models for medical transcript analysis. *arXiv preprint arXiv:2211.09722*.
- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.
- Lindsay McShane, Ethan Pancer, Maxwell Poole, and Qi Deng. 2021. Emoji, playfulness, and brand engagement on twitter. *Journal of Interactive Marketing*, 53(1):96–110.
- Peyman Passban, Tanya Roosta, Rahul Gupta, Ankit Chadha, and Clement Chung. 2022. Training mixed-domain translation models via federated learning. *arXiv preprint arXiv:2205.01557*.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.
- Dunlu Peng and Huimin Zhao. 2021. Seq2emoji: A hybrid sequence generation model for short text emoji prediction. *Knowledge-Based Systems*, 214:106727.
- Swaroop Ramaswamy, Rajiv Mathews, Kanishka Rao, and Franoise Beaufays. 2019. Federated learning for emoji prediction in a mobile keyboard. *arXiv preprint arXiv:1906.04329*.
- General Data Protection Regulation. 2018. General data protection regulation (gdpr). *Intersoft Consulting*, Accessed in October, 24(1).
- Nuria Rodr guez-Barroso, Eugenio Mart nez-C mara, M Victoria Luz n, and Francisco Herrera. 2022. Dynamic defense against byzantine poisoning attacks in federated learning. *Future Generation Computer Systems*, 133:1–9.
- Pushpa Singh, Murari Kumar Singh, Rajnesh Singh, and Narendra Singh. 2022. Federated learning: Challenges, methods, and future directions. In *Federated Learning for IoT Applications*, pages 199–214. Springer.
- Toshiki Tomihira, Atsushi Otsuka, Akihiro Yamashita, and Tetsuji Satoh. 2020. Multilingual emoji prediction using bert for sentiment analysis. *International Journal of Web Information Systems*, 16(3):265–280.
- Pranav Venkit, Zeba Karishma, Chi-Yang Hsu, Rahul Katiki, Kenneth Huang, Shomir Wilson, and Patrick Dudas. 2021. Asourceful’twist: Emoji prediction based on sentiment, hashtags and application source. *arXiv preprint arXiv:2103.07833*.
- Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. 2020a. Attack of the tails: Yes, you really can backdoor federated learning. *Advances in Neural Information Processing Systems*, 33:16070–16084.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020b. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.
- Orion Weller, Marc Marone, Vladimir Braverman, Dawn Lawrie, and Benjamin Van Durme. 2022. Pre-trained models for multilingual federated learning. *arXiv preprint arXiv:2206.02291*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pi ric Cistac, Tim Rault, R mi Louf, Morgan Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
- Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Franoise Beaufays. 2018. Applied federated learning: Improving google keyboard query suggestions. *arXiv preprint arXiv:1812.02903*.
- Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2022. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations. *arXiv preprint arXiv:2209.07562*.

## A Multilingual Macro-F1 Results in Label-flipping Attack Scenario Results

Tables 14 and 15 present a comparison of centralized and federated learning for different multilingual models in a label-flipping attack scenarios with an Attack Ratio of 25%. The models’ performance was evaluated using two different settings, Fed-IID and Fed-Non-IID, and two federated learning algorithms, FedAVG and Krum. Both tables show that the models’ performance decreases in the label-flipping attack scenario, and is worse for the Fed-Non-IID setting than for the Fed-IID setting. Furthermore, Krum outperforms the FedAVG

aggregator in all cases, and the Bert-Base model generally performs better than the other models.

Model	Setting	FedAVG	Krum
Bert-Base	Fed-IID	36.3%	36.5%
	Fed-Non-IID	33.5%	35.3%
	Finetuned	35.6%	
XLM-R	Fed-IID	FedAVG	Krum
	Fed-Non-IID	34.7%	34.9%
	Finetuned	32.7%	33.4%
M-MiniLM	Fed-IID	33.9%	
	Fed-Non-IID	FedAVG	Krum
	Finetuned	32.3%	32.2%
		30.4%	30.7%
		32.7%	

Table 14: Centralized and Federated Learning Results in Label-Flipping Attack Scenario for the SemEval English test dataset.

Model	Setting	FedAVG	Krum
Bert-Base	Fed-IID	28.2%	29.1%
	Fed-Non-IID	30.5%	29.7%
	Finetuned	28.9%	
XLM-R	Fed-IID	FedAVG	Krum
	Fed-Non-IID	26.1%	27.7%
	Finetuned	27.1%	27.3%
M-MiniLM	Fed-IID	26.7%	
	Fed-Non-IID	FedAVG	Krum
	Finetuned	22.9%	25.3%
		24.8%	24.6%
		24.6%	

Table 15: Centralized and Federated Learning Results in Label-Flipping Attack Scenario for the average results for the Twitter multilingual dataset.

Table 16 presents a comparison between centralized and federated learning approaches using FedAVG and Krum aggregation functions in a label-flipping attack scenario for three different models. We observed that Krum outperforms FedAVG in all cases, indicating that the choice of aggregation function has a significant impact on the model’s performance. However, we also found that the model’s performance is heavily dependent on the architecture and the type of aggregation function used.

The results of our experiments highlight the vulnerability of federated learning to label-flipping attacks. This vulnerability emphasizes the importance of carefully selecting the federated learning algorithm and aggregation function to mitigate such attacks. Additionally, our experiments revealed that the Fed-IID setting is less vulnerable to label-flipping attacks, which suggests that data distribution plays a critical role in the performance of federated learning models.

To further investigate the impact of label-flipping attacks on federated learning models, future experiments can explore the impact of different attack ratios and other attack scenarios. Moreover, it would be interesting to study the impact of other factors, such as the heterogeneity of data sources and the distribution of data samples, on the vulnerability of federated learning models to attacks. By gaining a better understanding of the vulnerabilities of federated learning, we can develop more robust and secure models that are better suited for real-world applications.

Model	Setting	FedAVG	Krum
Bert-Base	Fed-IID	20.67%	23.48%
	Fed-Non-IID	20.60%	21.62%
	Finetuned	21.6%	
XLM-R	Fed-IID	FedAVG	Krum
	Fed-Non-IID	18.60%	21.25%
	Finetuned	18.81%	19.93%
M-MiniLM	Fed-IID	18.2%	
	Fed-Non-IID	FedAVG	Krum
	Finetuned	13.78%	16.47%
		14.62%	15.38%
		14.7%	

Table 16: Centralized and Federated Learning Results in Label-Flipping Attack Scenario for the German Zero-shot.