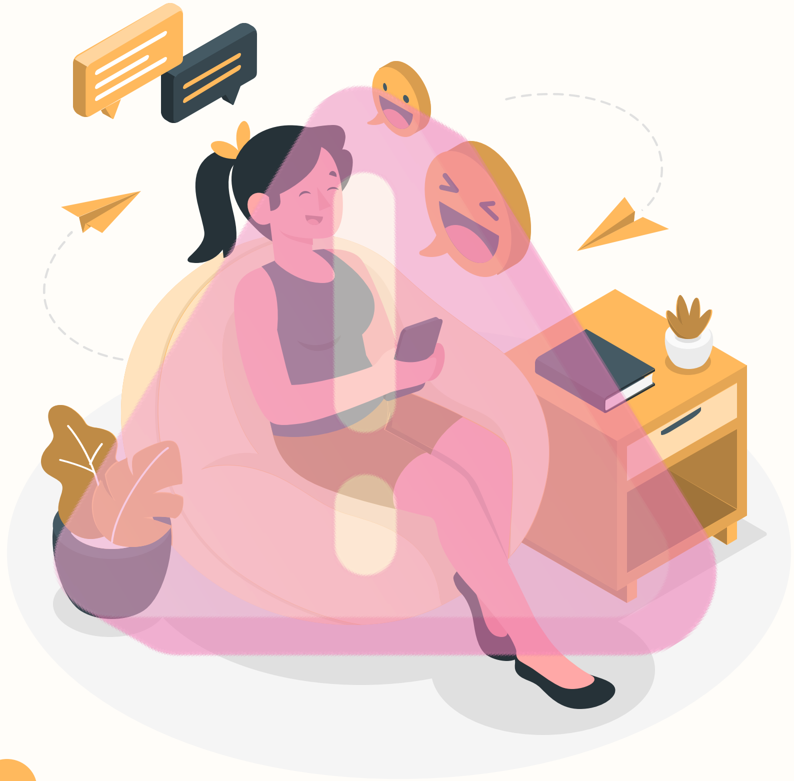


Hate Speech Recognition

Centralized Vs Federated Learning

+ Ashkan Vedadi Gargary
862352576
+ Aditya Mohan Gupta
862467699

An Italian is forced to watch the first time pineapple is added to a pizza, Brooklyn, New York 1924.

Table of Contents

01

Intro + BG

Related Works

Introduction

- The rise of social media platforms has significantly increased the prevalence of user-generated content, bringing the issue of hate speech to the forefront.
- Hate speech detection is crucial for maintaining safe and inclusive online environments, as it helps prevent the spread of harmful and offensive content.
- This project explores the use of advanced machine learning models, specifically centralized and federated learning approaches, to effectively detect hate speech while preserving user privacy.

Background

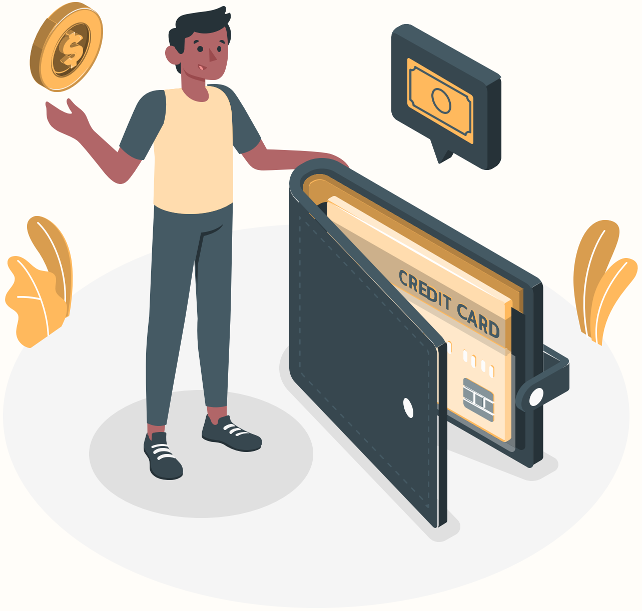
- The Traditional hate speech detection methods rely on centralized machine learning models, such as BERT, which require large, aggregated datasets, raising concerns about data privacy and security.
- Federated learning offers a novel approach by enabling models to be trained across multiple decentralized devices, ensuring that sensitive data remains local and private.
- This project showcases the performance of centralized models like TinyBERT with federated models such as FedProx and FedAVG, highlighting the trade-offs between accuracy and data privacy.


EVERYONE IS ENTITLED TO THEIR OPINION,
UNLESS IT MAKES ME TRIGGERED, THEN IT'S HATE SPEECH
imgflip.com

02

Motivation

Why do we need it?



Motivation

- **Enhancing Online Safety:** The proliferation of hate speech on social media platforms poses significant threats to the safety and well-being of users. Developing effective detection models is crucial to mitigate the spread of harmful content and foster healthier online communities.

- **Balancing Privacy and Performance:** Traditional centralized machine learning models, while effective, often compromise user privacy. Federated learning offers a promising alternative by allowing data to remain local, thus addressing privacy concerns while still achieving high model performance.

03

Problem

Problem Definition and Objectives

Problem Definition

- We propose an **FL-based model for Hate Speech recognition** and show that it can outperform centralized models by achieving over **6% higher accuracy**.
- We also try to compare different methods of Federated-based recognition with TinyBert Classifier.
- We mainly focus on the datasets that classify a prompt or message into two groups:
  (i) Hate-Full
  (ii) Non-Hate-Full

Objectives

1

Finding Related Datasets

2

Data Pre-Processing

3

Re-implementing SOTA Centralized Methods

4

Federating Best Centralized Models

5

Reproducing Current Federated Models

6

Comparing All Models

04

Methods

Centralized
Federated

Methodology

Centralized

Logistic Regression
Decision Trees
Random Forest
K-Nearest Neighbors
TinyBert Classifier

Federated

Neural Network + FedProx
TinyBert + FedAVG

Centralized Methods – Classic ML

Decision Tree

Logistic Regression

Random Forest

K-NN

Centralized Methods – TinyBert

TinyBert Classifier

For our BERT-based model, we employed **Tiny-BERT**, a compact version of the BERT model.

- A total of **4 million parameters**
- **2** hidden layers with **128** hidden dimension.
- Tiny-Bert is **7x smaller** and **9x faster** than BERT while achieving **96%** of its performance.

Centralized Methods

Datasets	Model	Validation (20%)		Test (30%)	
		Accuracy	F1	Accuracy	F1
Kaggle (1)	DT	97.49%	97.40%	67.24%	79.89%
	LR	94.03%	93.90%	51.86%	66.62%
	KNN	96.57%	96.43%	67.24%	79.89%
	RF	98.67%	98.64%	71.03%	82.75%
	TB	**98.36%**	**98.36%**	**95.33%**	**95.47%**
Kaggle (2)	DT	46.70%	52.61%	51.82%	66.60%
	LR	90.83%	91.01%	78.67%	87.91%
	KNN	90.09%	89.18%	87.33%	93.22%
	RF	98.44%	98.42%	58.85%	73.06%
	TB	**97.73%**	**97.72%**	**94.00%**	**94.43%**
Davidson	DT	95.95%	95.80%	54.35%	67.90%
	LR	96.12%	96.00%	55.42%	68.52%
	KNN	85.29%	86.05%	54.35%	67.90%
	RF	96.01%	95.86%	54.89%	68.34%
	TB	**92.13%**	**92.13%**	**91.90%**	**91.96%**
Merge	DT	96.36%	96.24%	60.78%	74.59%
	LR	94.96%	97.26%	85.35%	88.04%
	KNN	91.46%	95.37%	60.78%	74.59%
	RF	98.12%	98.09%	73.87%	84.61%
	TB	**97.20%**	**97.20%**	**93.31%**	**93.96%**

Overfitting on Classical ML:

- Strong Validation Accuracy, Weak Test Accuracy

Random Forest (RF) achieves the best performance among classical machine learning models.

TinyBERT achieves over 91% accuracy across all three datasets. We achieved an accuracy and F1 score of 93% for the merged dataset.

Federated Methods

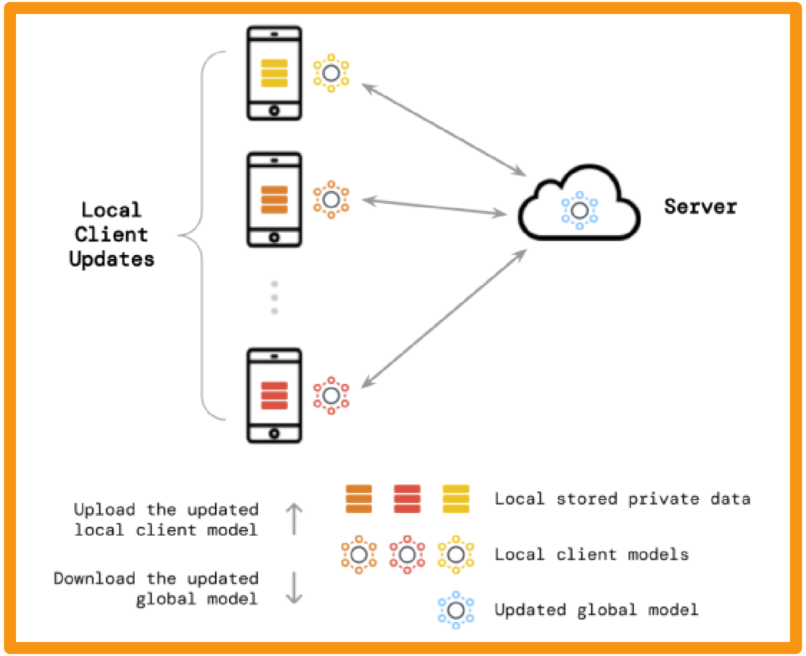
Neural Networks

FedPROX

TinyBert

FedAVG

Federated Methods



Local Training: Each client trains on its local dataset and sends the model updates to a centralized server locally.

Aggregation: The server receives model updates from all participating clients and performs secure aggregation over the uploaded parameters without learning local information.

Aggregated Parameters Broadcasting: The server broadcasts the aggregated parameters of model updates to all clients.

Updating Local Models: Each client updates its local model with the aggregate parameter received from the server, thereby improving its performance

Federated Methods
Neural Network + FedProx

Neural Networks

FedPROX

Customization and Flexibility: FedProx introduces a proximal term to the standard federated learning objective, allowing for customization of local updates. This term helps in handling heterogeneous data distributions across different clients, making the model more robust and adaptable to varied local datasets.

Improved Convergence: By incorporating the proximal term, FedProx stabilizes the training process across multiple clients with diverse data, leading to improved convergence rates. This ensures that even with non-IID (non-Independent and Identically Distributed) data, the model converges effectively, enhancing overall performance.

Federated Methods
Neural Network + FedProx

Neural Networks

FedPROX

Balanced Performance and Privacy:

FedProx maintains high accuracy in hate speech detection while ensuring data privacy. By performing computations locally on clients' devices and only sharing model updates (not raw data), FedProx strikes a balance between achieving robust model performance and preserving user privacy.

```
Epoch 1, Loss: 0.3179404742801294
Accuracy: 0.9035631536101196
Epoch 2, Loss: 0.31761063959482166
Accuracy: 0.9035631536101196
Epoch 3, Loss: 0.31744151197849435
Accuracy: 0.9035631536101196
Epoch 4, Loss: 0.31727424332398796
Accuracy: 0.9035631536101196
Epoch 5, Loss: 0.31700114741902774
Accuracy: 0.9035631536101196
Epoch 6, Loss: 0.31693555000651913
Accuracy: 0.9035631536101196
Epoch 7, Loss: 0.31688254936802224
Accuracy: 0.9035631536101196
Epoch 8, Loss: 0.3167625699290197
Accuracy: 0.9035631536101196
Epoch 9, Loss: 0.31640649866848924
Accuracy: 0.9035631536101196
Epoch 10, Loss: 0.3164629475382601
Accuracy: 0.9035631536101196
```

Federated Methods
TinyBert + FedAVG

TinyBert

FedAVG

Utilize Best Centralized Method:

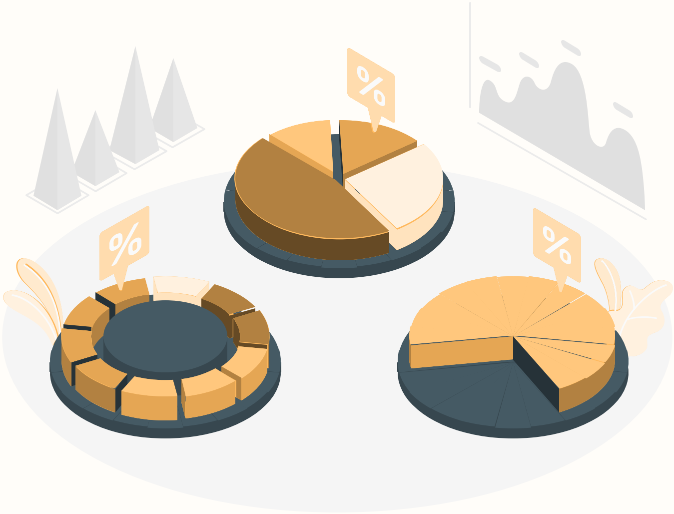
We utilize our best-centralized model, TinyBERT, for our federated setup. We simulate the federated environment by using virtual machines in the PySyft library.

```
Epoch 1, Loss: 0.6426772759563621
Accuracy: 0.7336170212765958
Epoch 2, Loss: 0.5326418862945732
Accuracy: 0.7948936170212766
Epoch 3, Loss: 0.4502756828549265
Accuracy: 0.8561702127659574
Epoch 4, Loss: 0.3864084206435872
Accuracy: 0.8680851063829788
Epoch 5, Loss: 0.3358551733110143
Accuracy: 0.8825531914893617
Epoch 6, Loss: 0.2989364254406129
Accuracy: 0.8910638297872341
Epoch 7, Loss: 0.27172867988032856
Accuracy: 0.8978723404255319
Epoch 8, Loss: 0.24658665780363412
Accuracy: 0.9012765957446809
Epoch 9, Loss: 0.2316553740837108
Accuracy: 0.9038297872340425
Epoch 10, Loss: 0.21480954070200867
Accuracy: 0.9080851063829787
```

05

Evaluation

Does it worth it?



Experiment Setups

HW& SW

Google Colab Free
Python
PySyft/Scikit-Learn
Hugginface

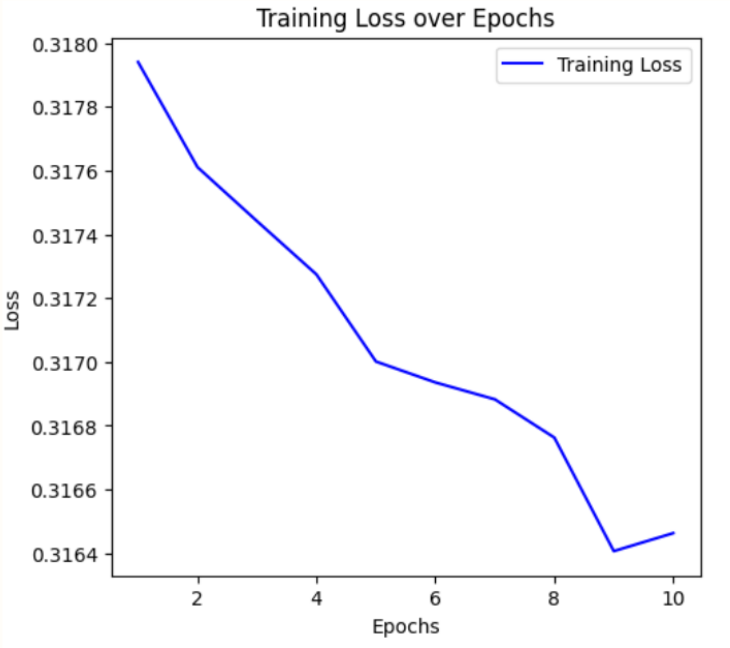
Datasets

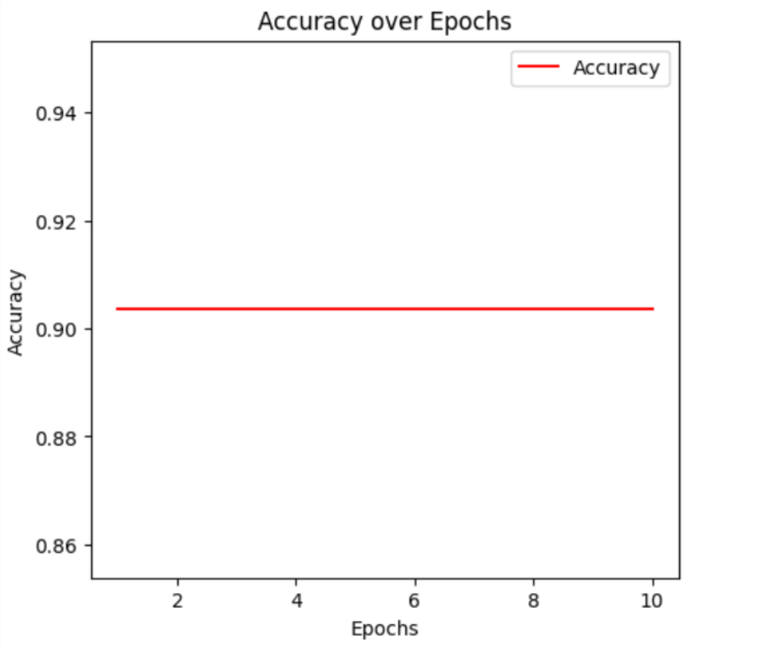
Kaggle (1)
Kaggle (2)
Davidson
Merged

Measurement

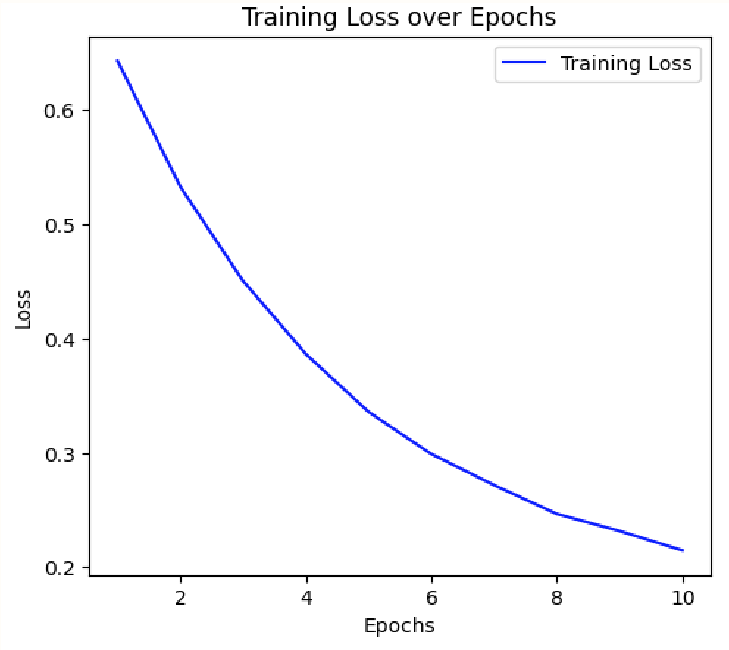
Accuracy
F1-Score

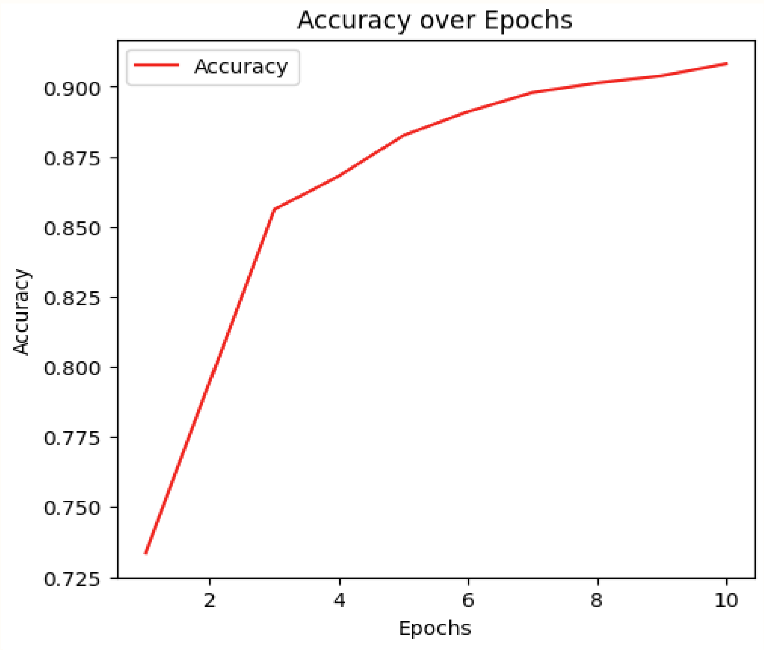
Federated Neural Network + FedProx





Federated TinyBert + FedAVG





Centralized Vs. Federated

Training Model	Aggregation Method	Test Accuracy
Centralized DT	N/A	60.78%
Centralized RF	N/A	73.87%
Centralized LR	N/A	85.35%
Centralized KNN	N/A	60.78%
Centralized TinyBert	N/A	93.31%
Federated Neural Network	FedProx	90.35%
Federated TinyBert	FedAVG	90.80%

Federated NN Vs. Federated TB

- Both Federated Method after 10 iterations achieve 91% accuracy.

Federated Vs. Centralized Classical ML

- Our implementation of both federated methods after 10 iterations outperforms Classical Centralized Methods.

Federated VS. Centralized TinyBert

- Same accuracy after 10 iterations while providing enhanced privacy.

06

Conclusion

Discussion, Revisiting Objectives, Implications, Feature Works, Limitations

Discussion

Centralized vs. Federated Learning: The study demonstrates that while centralized models like TinyBERT achieve high accuracy in hate speech detection, federated models such as FedProx offer comparable performance with the added benefit of enhanced data privacy.

Privacy without Compromise: Federated learning models manage to maintain user data privacy without significantly compromising on detection accuracy, proving to be a viable solution for privacy-sensitive applications.

Conclusion

Effective Hate Speech Detection: The integration of federated learning into hate speech detection systems provides a robust and privacy-preserving alternative to traditional methods, making online platforms safer.

Future Potential: The promising results of both Federated Methods shows its potential for broader applications, suggesting future research could explore its use in other domains where data privacy is crucial.

Limitations

Resources:

- There are many other encoders for classifying each tweet, which we could not implement due to a lack of resources and time limitations.
- Epoch, Iterations, Different Params

Time:

- We could not reproduce more SOTA papers.
- We only considered a virtual machine federated setup without implementing the real server-client model transfer

Future Works

Multi-Class Classification

Instead of Binary Classification

Effect of Client Size

The effect of the number of clients.

Privacy and Security

Comparing existing methods regarding privacy and security is crucial.

Different Aggregation

Many other well-performing aggregation methods exist and can be helpful.

Materials

GitHub:

- https://github.com/ashkanvg/Hate-Speech-Recognizer

Documentation:

- More information is available in the documentation.

Dataset:

- https://drive.google.com/drive/folders/1yrHJnPINYEEe674dYridJc23odxYG5h6


EVERY GROUP PROJECT
DOES 99% OF THE WORK
HAS NO IDEA WHAT'S GOING ON THE WHOLE TIME
SAYS HE'S GOING TO HELP BUT HE'S NOT
DISAPPEAR AT THE VERY BEGINNING AND DOESN'T SHOW UP AGAIN TIL THE VERY END
VIA 9GAG.COM
IN SCHOOL YOU HAVE EVER DONE

Our Team

Ashkan Vedadi Gargary

- Paper Reviews
- Objectives
- Centralized Methods
- Federated TinyBert
- Centralized Evaluation
- Federated Evaluation
- Limitation
- Feature Works

Aditya Mohan Gupta

- Paper Reviews
- Background
- Introduction
- Objectives
- Federated NN
- Federated Evaluation
- Discussion
- Implications

Thanks!

Aditya Mohan Gupta
Ashkan Vedadi Gargary

Resources

1. Slide Template is from SLIDESGO: https://slidesgo.com/
2. Illustration by OlFi from Ouch! Icons8.com