Cyprus University of Technology
Department of Electrical Engineering and Computer Engineering and Informatics

Bachelor Thesis

An Evaluation of a Federated Learning Framework for Detecting Hate Speech and Disinformation on Social Media Platforms

Michael Angelos Demou

Limassol, May 2024

CYPRUS UNIVERSITY OF TECHNOLOGY

Faculty of Engineering and Technology

Department of Electrical Engineering, Computer Engineering, and Informatics

Bachelor Thesis

An Evaluation of a Federated Learning Framework for Detecting Hate Speech and Disinformation on Social Media Platforms

Michael Angelos Demou

Advisor:

Dr Michael Sirivianos

Limassol, May 2024

Copyrights

The approval of the dissertation by the Department of Electrical Engineering, Computer Engineering, and Informaticsdoes not necessarily imply the approval by the Department of the views of the writer.

Acknowledgements

I would like to especially thank Dr. Michael Sirivianos, Dr. Nikos Salamanos and PhD student Pantelitsa Leonidou of the Department of Electrical Engineering, Computer Engineering and Informatics, for their help and support throughout this thesis. Thank you for giving me the opportunity to undertake a topic of such interest and importance.

ABSTRACT

The spread of harmful content and fake news on social media platforms has become a significant issue, highlighting the need for automatic detection and filtering tools. Machine learning, which enables computers to identify patterns, is frequently employed to develop these tools. However, it involves handling sensitive user data, which must be managed carefully to respect privacy. Federated Learning (FL) addresses this by allowing data to remain on the user's device, rather than being transferred to a central server. In FL, learning occurs locally on individual devices and only model updates are shared with the server, thus minimizing privacy risks. Despite its advantages, FL faces challenges such as statistical heterogeneity, where data is not uniformly distributed across devices. This can result in biased predictions when some classes are over-represented in a client's dataset or when data amounts vary, affecting model influence. This thesis develops a text classifier within a distributed Federated Learning (FL) setup to identify tweets containing harmful content or fake news, focusing on simulating two custom non-IID methods from another study [12] to create real a life scenario for detecting harmful content, as well as a real-life non-IID scenario for disinformation detection. Testing this model on five Twitter datasets representing various types of user misbehavior and misinformation yielded promising results, with F1 scores ranging from 70% to 82%. We further analyzed the model's performance with different numbers of clients and compared it to a centralized approach. Additionally, we experimented with various alpha values affecting the custom non-IID methods to assess the model's performance compared to IID conditions. This comprehensive evaluation confirms the robustness of our FL framework, making it a viable solution for detecting harmful content and disinformation in real-world, non-IID scenarios.

Keywords: Federated Learning, Non-IID, Disinformation, Harmful Content, Text Classification