Detecção de Spam em SMS com Aprendizado Federado e Privacidade Diferencial

Relatório Final

Ana Flávia de Matos Souza - 2020006353
Carlos Magalhães Silva - 2021421885
Henrique da Fonseca Diniz Freitas - 2021031688
Renato Silva Santos - 2020006981

Junho de 2025

Contents

1 Introdução **3**

2 Metodologia **3**

2.1 Descrição do Problema 3

2.2 Dataset 3

2.3 Pré-processamento de Dados 3

2.4 Modelos de Machine Learning e Extração de Características 4

2.5 Configuração do Ambiente de Aprendizado Federado 4

2.6 Implementação Manual de Privacidade Diferencial 4

3 Desenho Experimental **5**

3.1 Experimento 1: Linha de Base Centralizada 5

3.2 Experimento 2: AF com Dados IID 5

3.3 Experimento 3: AF com Dados Não-IID 5

3.4 Experimento 4: AF Realista com Hashing Vectorizer 5

3.5 Experimento 5: AF com Privacidade Diferencial 5

4 Resultados e Análise **6**

4.1 Discussão dos Resultados 6

5 Conclusão **7**

1 Introdução

A proliferação de spam e ataques de *phishing* por SMS (conhecidos como *smishing*) representa uma ameaça crescente à segurança e privacidade dos usuários. Soluções tradicionais de detecção, geralmente baseadas em servidores centralizados, exigem acesso direto às mensagens dos usuários, o que levanta sérias preocupações éticas e legais.

O Aprendizado Federado (AF) surge como uma alternativa promissora, possibilitando o treinamento de modelos de *machine learning* sem que os dados saiam dos dispositivos dos usuários. Para fortalecer ainda mais as garantias de privacidade, este projeto incorporou também técnicas de Privacidade Diferencial (DP), buscando um equilíbrio entre desempenho e proteção de dados.

2 Metodologia

2.1 Descrição do Problema

O desafio central foi desenvolver um sistema de detecção de mensagens SMS, classificando-as como spam ou legítimas (*ham*), com o uso de Aprendizado Federado aliado a mecanismos de proteção de privacidade. O sistema precisava garantir boa acurácia, respeitar a heterogeneidade dos dados entre usuários e impedir a exposição de informações sensíveis ao servidor central.

2.2 Dataset

Utilizou-se o **SMS Spam Collection Dataset**, amplamente citado na literatura, contendo 5.574 mensagens em inglês, rotuladas como *ham* ou *spam*. O forte desbalanceamento (87% *ham* e 13% *spam*) exigiu estratégias específicas durante o treinamento, como o ajuste de pesos entre classes.

2.3 Pré-processamento de Dados

As mensagens passaram por um rigoroso processo de limpeza textual:

- Remoção de pontuação e caracteres especiais.
- Normalização para letras minúsculas.
- Eliminação de *stopwords*, com o objetivo de reduzir ruídos nas representações textuais.

Após a limpeza, os textos foram vetorizados usando duas abordagens distintas, detalhadas a seguir.

2.4 Modelos de Machine Learning e Extração de Características

O modelo escolhido foi a **Regressão Logística**, com o hiperparâmetro *class_weight* configurado como *balanced*, visando mitigar o impacto do desbalanceamento de classes.

As representações textuais utilizaram as seguintes técnicas:

1. **TF-IDF (Term Frequency-Inverse Document Frequency)**: Produz uma representação ponderada da frequência dos termos em relação ao corpus total. Embora eficaz, essa abordagem requer um vocabulário centralizado, o que pode comprometer a privacidade em contextos federados.
2. **Hashing Vectorizer**: Transforma textos em vetores de tamanho fixo de maneira não reversível, sem a necessidade de um vocabulário central. Essa técnica foi fundamental nos experimentos que priorizaram a privacidade.

2.5 Configuração do Ambiente de Aprendizado Federado

A simulação do ambiente federado foi realizada com o framework **Flower**, permitindo a criação de múltiplos clientes virtuais.

Principais parâmetros de configuração:

- **Número de Clientes:** 10.
- **Estratégia de Agregação:** Federated Averaging (FedAvg).
- **Número de Rodadas:** Até 50, conforme o experimento.
- **Distribuição de Dados:** Simulação de cenários IID e Não-IID, com controle explícito sobre a distribuição entre os clientes.

2.6 Implementação Manual de Privacidade Diferencial

Foi implementado um mecanismo de **Privacidade Diferencial** diretamente no processo de atualização de modelos locais, com as seguintes etapas:

1. **Clipping L2:** Limitação das atualizações de gradiente a uma norma L2 máxima de 5.0, controlando outliers.
2. **Adição de Ruído Gaussiano:** Aplicação de ruído com desvio padrão de 0.5, proporcional ao clipping, com um *noise multiplier* de 0.1.
3. **Atualização do Modelo Global:** As atualizações privatizadas foram agregadas ao modelo global de maneira consistente com a técnica FedAvg.

O objetivo foi combinar garantias formais de privacidade com viabilidade prática de treinamento.

3 Desenho Experimental

Foram definidos cinco experimentos principais, cada um refletindo diferentes níveis de dificuldade e restrição de privacidade.

3.1 Experimento 1: Linha de Base Centralizada

Modelo de Regressão Logística treinado centralizadamente com TF-IDF e divisão 80/20 entre treino e teste. Este cenário serve como limite teórico superior.

3.2 Experimento 2: AF com Dados IID

Distribuição aleatória e homogênea dos dados entre os clientes, criando um cenário IID. O objetivo foi avaliar a performance básica do AF sob condições ideais.

3.3 Experimento 3: AF com Dados Não-IID

Simulação de heterogeneidade realista, com distribuição desigual das classes:

- 8 clientes com 90% *ham*.
- 2 clientes com predominância de *spam*.

O objetivo foi testar a robustez do AF frente a distribuições Não-IID.

3.4 Experimento 4: AF Realista com Hashing Vectorizer

Neste cenário, adotou-se o Hashing Vectorizer para garantir privacidade extra, eliminando o compartilhamento de vocabulário entre os clientes. O impacto esperado foi a redução da performance como consequência do ganho em privacidade.

3.5 Experimento 5: AF com Privacidade Diferencial

Configuração mais restritiva do estudo, combinando a distribuição Não-IID, o uso de Hashing Vectorizer e a aplicação de DP em todas as rodadas.

Parâmetros de DP:

- **Clipping L2:** 5.0.
- **Noise Multiplier:** 0.1.
- **Rodadas:** 50.

4 Resultados e Análise

Table 1: Comparação de Métricas dos Experimentos

Métrica	**Exp. 1 Centralizado**	**Exp. 2 AF (IID)**	**Exp. 3 AF (Não-IID)**	**Exp. 4 AF (Hashing)**	**Exp. 5 AF + DP**
Acurácia	0.9812	0.9704	0.9713	0.9534	0.9600
Precisão	0.9384	0.9677	0.8774	0.7836	0.9400
Recall	0.9195	0.8053	0.9127	0.8993	0.7700
F1-Score	0.9288	0.8791	0.8947	0.8375	0.8400

4.1 Discussão dos Resultados

A análise dos experimentos destacou a influência de diferentes fatores sobre o desempenho.

O Experimento 1 confirmou o esperado: o treinamento centralizado, sem restrições, gerou os melhores resultados. A combinação de alta precisão e recall resultou em um F1-Score superior a 0.92.

Os Experimentos 2 e 3 evidenciaram o impacto da distribuição de dados. O cenário IID favoreceu a precisão, enquanto o Não-IID trouxe um ganho expressivo no recall, refletindo a especialização de alguns clientes.

No Experimento 4, a mudança para o Hashing Vectorizer resultou em queda geral de desempenho, confirmando a penalidade de privacidade nas representações de dados.

O Experimento 5 apresentou o efeito mais notável da Privacidade Diferencial. Apesar da manutenção de uma precisão elevada (0.94), o recall sofreu uma queda significativa (0.77), evidenciando o clássico trade-off entre utilidade e privacidade.

Importante destacar que, mesmo sob ruído, o sistema manteve a capacidade de convergência, demonstrando a eficácia do balanceamento entre clipping e adição de ruído.

5 Conclusão

Este trabalho demonstrou a viabilidade da detecção de spam em SMS por meio de Aprendizado Federado aliado a Privacidade Diferencial, com impactos controlados sobre a performance do modelo.

Apesar da penalidade observada nas métricas de recall, o desempenho geral permaneceu dentro de níveis aceitáveis para aplicações práticas. O F1-Score de aproximadamente 0.84 no cenário mais restritivo reforça essa conclusão.

Os resultados também evidenciaram a relevância de fatores como a distribuição dos dados e a escolha de técnicas de extração de características. Esses aspectos devem ser considerados cuidadosamente em futuras implementações.