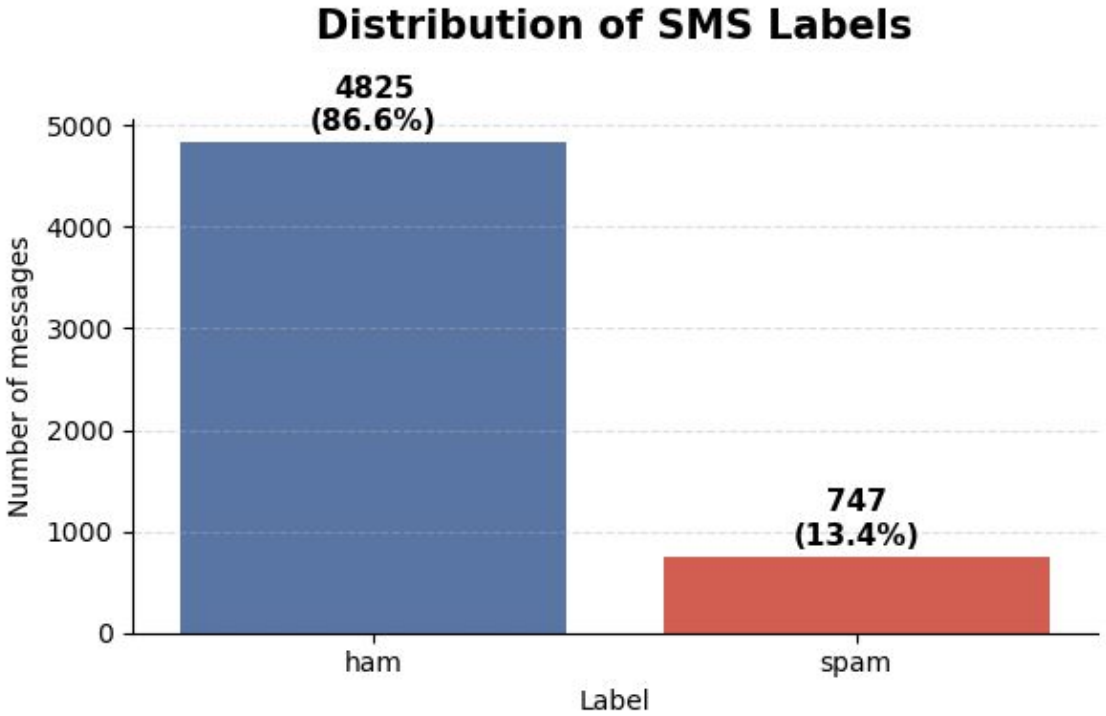
Spam Detector

Automatically classify SMS messages as spam or ham

Project Overview

5572 messages



Data Cleaning & Preprocessing

- **Standardize messages:**
  - lower casing
  - remove punctuation
  - keep digits
- **Convert messages into a comparable format:**
  - same length and structure for all messages
- **Learn common words from the data:**
  - ~7 600 unique words

Model Architecture

- ***Embedding***

  → understands word meaning
- ***LSTM***

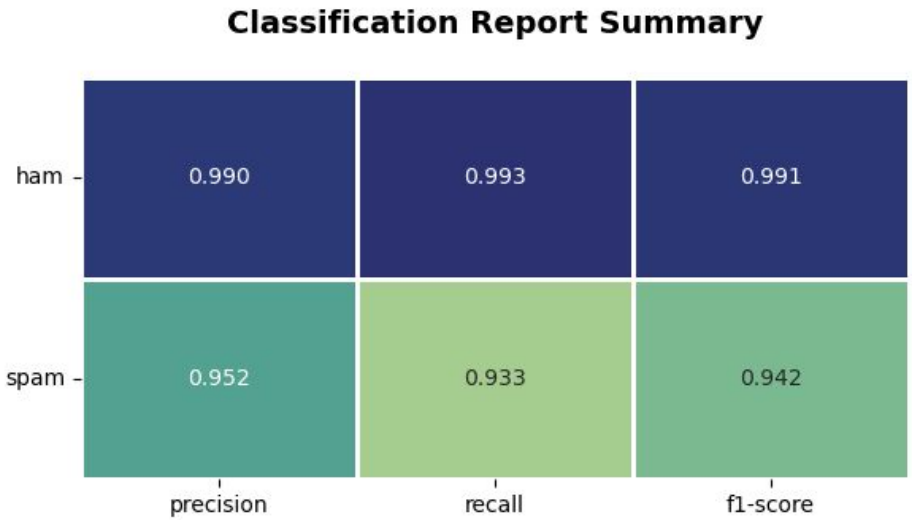
  → understands word order and context
- ***Dropout***

  → prevents overfitting
- ***Dense (sigmoïde)***

  → outputs spam probability

Training Results: heatmap

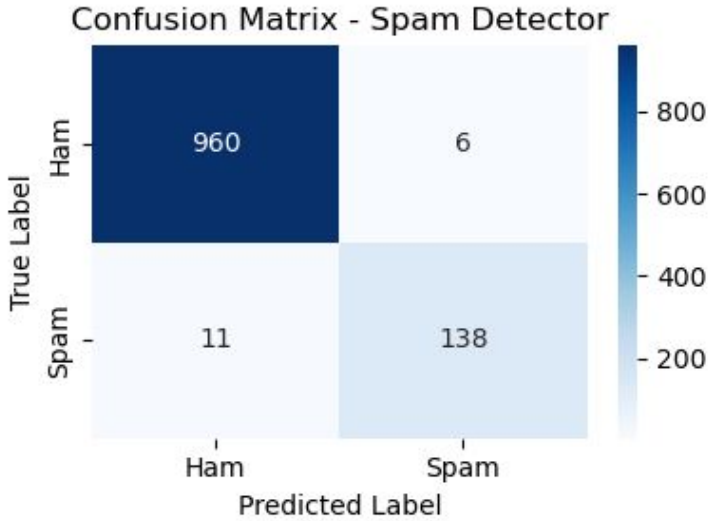
Each cell shows how well the model performs for **ham** and **spam** in terms of **precision**, **recall**, and **F1-score**.



F1-score combines **precision** and **recall** – near-perfect values indicate a robust model.

Training Results: Confusion matrix

- Each cell shows the number of messages falling into a specific category of prediction vs reality.



True Negative Real *ham* predicted as *ham*	**False Positive** Real *ham* predicted as *spam*
False Negative Real *spam* predicted as *ham*	**True Positive** Real *spam* predicted as *spam*

Error Analysis Examples

False Positives
(predicted spam but actually ham):

- [0.87] waiting for your **call**
- [0.81] **nokia phone** is lovly
- [0.81] height of confidence all the aeronautics professors wer **call**d amp they wer askd 2 sit in an aeroplane aftr they sat they...
- [0.52] **unlimited texts** limited minutes

False Negatives
(predicted ham but actually spam):

- [0.00] sorry i missed your **call** let's talk when you have the time i'm on **07090201529**
- [0.00] for **sale** arsenal dartboard good condition but no doubles or trebles
- [0.18] latest **news** police station toilet stolen cops have nothing to go on

True Negative Real *ham* predicted as *ham*	**False Positive** Real *ham* predicted as *spam*
False Negative Real *spam* predicted as *ham*	**True Positive** Real *spam* predicted as *spam*

Prediction Tool: Real-time Spam Detection

This simple prediction interface allows testing the model with new unseen messages.

- The tool takes raw text input (SMS) and predicts its **probability of being spam**.
- Messages are color-coded: **green for HAM**, **red for SPAM**.

	Message	Predicted Label	Spam Probability
0	Congratulations! You've won a new iPhone, click here to claim!	SPAM	94.97%
1	Hi John, can you send me the report by tomorrow?	HAM	0.06%
2	Urgent! Your bank account has been locked, verify immediately.	SPAM	90.82%
3	Ok cool, I'll bring the cake for Saturday.	HAM	0.21%

Thank you for your attention
– any questions?