

Technical documentation

By team 1 2022

**Spam Classifier**

Introduction: sur une selection de 5574 messages étiquetés nous avons créé un modèle de ML qui permet de classifier si le SMS reçu est un HAM ou un SPAM. Pour cela nous avons recherché les descripteurs (features) :

- nombre et proportion des caractères "!, ?, €, #" etc

- longueur du message

- nombre et proportion de majuscules

- présence d'URL

- et d'autres descripteurs

Après cela nous avons créé un Dataframe avec les features qui permet de différencier les messages SPAM et HAM.

Par la suite, nous avons choisi deux modèles de machine learning : KNN et naive bayes.

Nous avons choisi le naive bayes car ces performance sont plus élevé.

Et enfin nous avons créé une application sur steamlit.