Task duration: 1 day

You are given a ham (0) - spam (1) dataset. Follow the instructions given below and answer the questions:

1. Build a tf-idf model and get the confusion matrix. Keep it as your baseline model. Do you notice anything worth looking into?
2. Use a transformer-based language model to extract keywords from your training dataset that are not already there in the baseline model's tf-idf vocabulary and may help in classification. Update the vocabulary of the tf-idf model.
3. Train another model with the updated vocabulary and see if it makes any difference to your confusion matrix. Please provide reasons for the changes observed. Perform any necessary analytics to support your conclusion.
4. If you have time, try different transformer-based keyword extraction methods, update the vocabulary, train the tf-idf model and compare your results. What do you see? Perform any necessary analytics to support your conclusion.

Note: Steps 1-3 are mandatory, step 4 is optional.

Instructions for solution:

1. You are allowed to use google search and AI assistants such as ChatGPT, Gemini, Copilot, etc. But if you use those, you have to show how your solution differs from the AI assistant's. Anything you copy-paste, you need to be able to understand and explain clearly. Extra points for unconventional thinking and solutions.
2. Please write clean and structured code in python. Use functional programming and OOPs concepts whenever necessary.
3. Comment your code generously.
4. Explain your thought process at the beginning of your notebook. You are supposed to write down clearly all the assumptions you are making to perform the above tasks and possible reasons for taking each step.
5. You can use Google Colab or Jupyter notebook for writing the code.
6. Save your notebook with name "coding_solution_<your_name>.ipynb" before sharing.