Email Spam Detection using Hidden Markov Model, Naïve Bayes, and Support Vector Machine

Lucas Franson Barros

Dann Anthony Astillero

# Abstract

This project presents a hybrid machine learning approach to email spam detection, combining the strengths of Hidden Markov Models (HMMs), Naïve Bayes, and Support Vector Machines (SVMs). Using the SpamAssassin Public Corpus, we aim to preprocess and analyze email content to classify messages as spam or ham effectively. HMMs are utilized for sequential pattern recognition, Naïve Bayes serves as a baseline classifier using probabilistic methods, and SVMs refine the results through high-dimensional feature separation. The goal is to enhance classification performance and accuracy using a multi-model strategy, with experiments evaluated through metrics such as accuracy, precision, recall, and F1-score.

# I. Introduction

Email spam detection is a crucial component in cybersecurity and information filtering systems. In this project, we propose the development of an efficient spam detection model using a hybrid machine learning approach. Our method combines three powerful algorithms—Hidden Markov Models (HMMs), Naïve Bayes, and Support Vector Machines (SVMs)—to leverage their individual strengths and achieve improved classification performance.  
  
We will use the SpamAssassin Public Corpus, a well-established dataset that contains a balanced mix of spam and non-spam (ham) emails. This corpus provides a solid foundation for training and evaluating the effectiveness of our models.

# II. Related Work

Prior research in spam detection has explored various machine learning algorithms. Naïve Bayes classifiers are widely used due to their simplicity and robust performance with text classification problems. SVMs have also proven effective, especially with high-dimensional feature spaces like those encountered in email filtering. Hidden Markov Models, commonly applied in natural language processing, offer a unique advantage in modeling word sequences and contextual information in emails.

We plan to base our work on insights from the following papers:  
1. A Comparative Study on Email Spam Filtering Techniques  
2. Machine Learning Approaches for Spam Detection in Email Communication  
3. Hidden Markov Models and their Applications in NLP

# III. Dataset and Features

Our primary dataset is the SpamAssassin Public Corpus. This dataset includes labeled examples of spam and ham emails, which are essential for supervised learning.

Feature extraction will include:  
- Tokenization and preprocessing of raw email content  
- Bag-of-words and TF-IDF features for Naïve Bayes and SVM  
- Sequential features for HMMs that capture linguistic structure and word dependencies  
- Additional engineered features based on exploratory data analysis.

# IV. Methodology

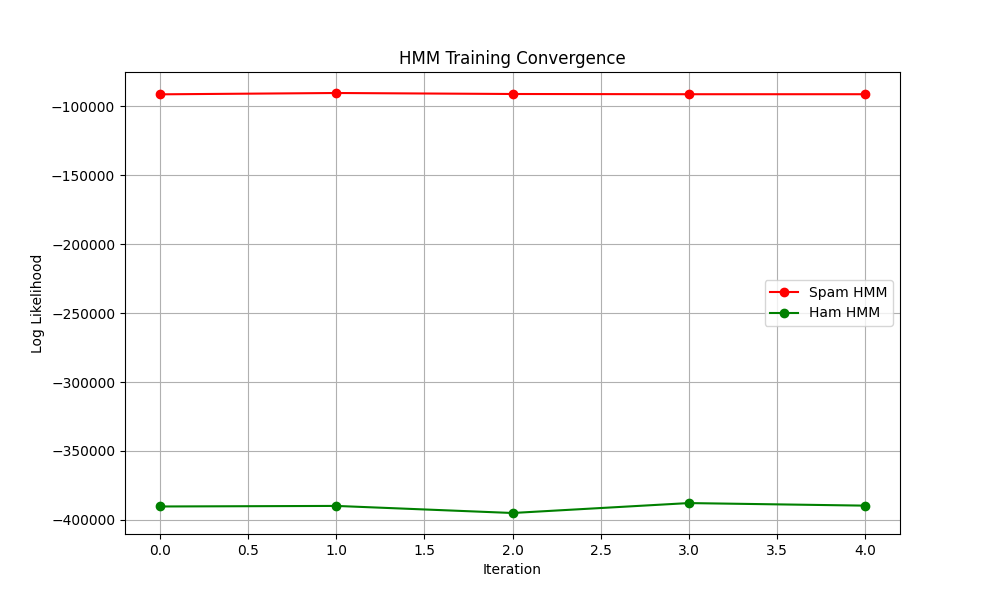
Our hybrid approach consists of the following components:  
  
1. Preprocessing: Scripts will clean, normalize, and tokenize emails.  
2. HMM: Used to model sequences and capture contextual word patterns. Implemented using hmm learn or a custom implementation.  
3. Naïve Bayes: Acts as a baseline probabilistic model using bag-of-words features.  
4. SVM: Refines classification accuracy using features generated from earlier stages.  
  
Each model will be trained and evaluated independently, followed by experimentation with combining outputs or using ensemble methods.

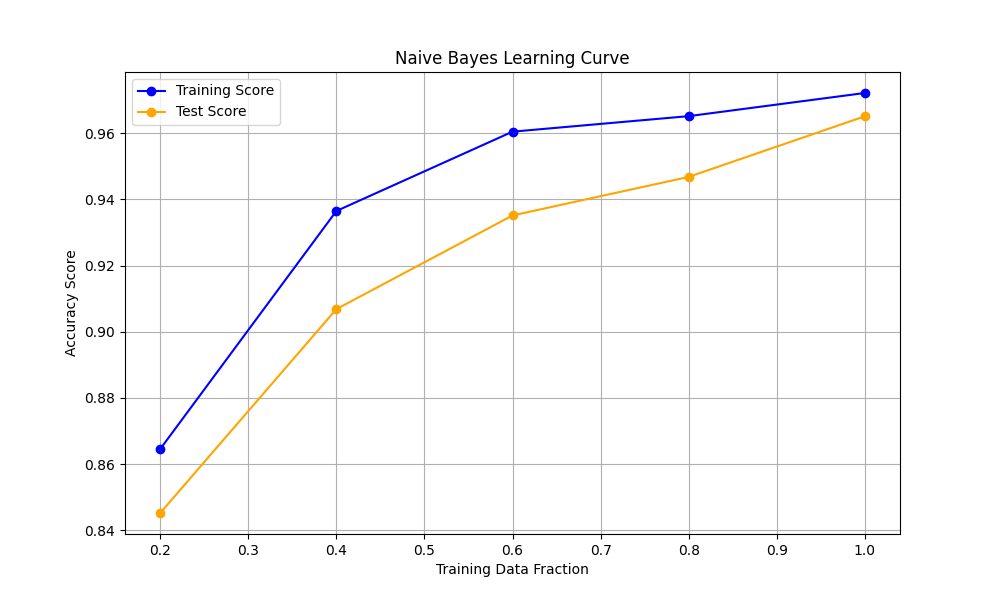
# V. Experiments

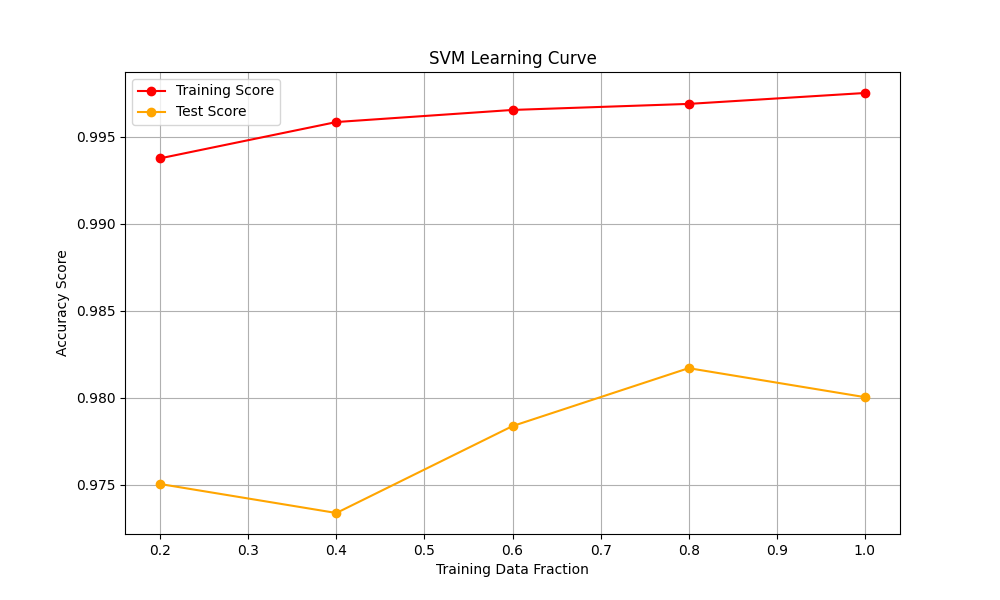
We will:  
- Conduct exploratory data analysis (EDA) on the dataset.  
- Train and evaluate each classifier (HMM, Naïve Bayes, SVM).  
- Measure performance using metrics such as accuracy, precision, recall, and F1 score.  
- Visualize classifier comparisons through graphs and confusion matrices.

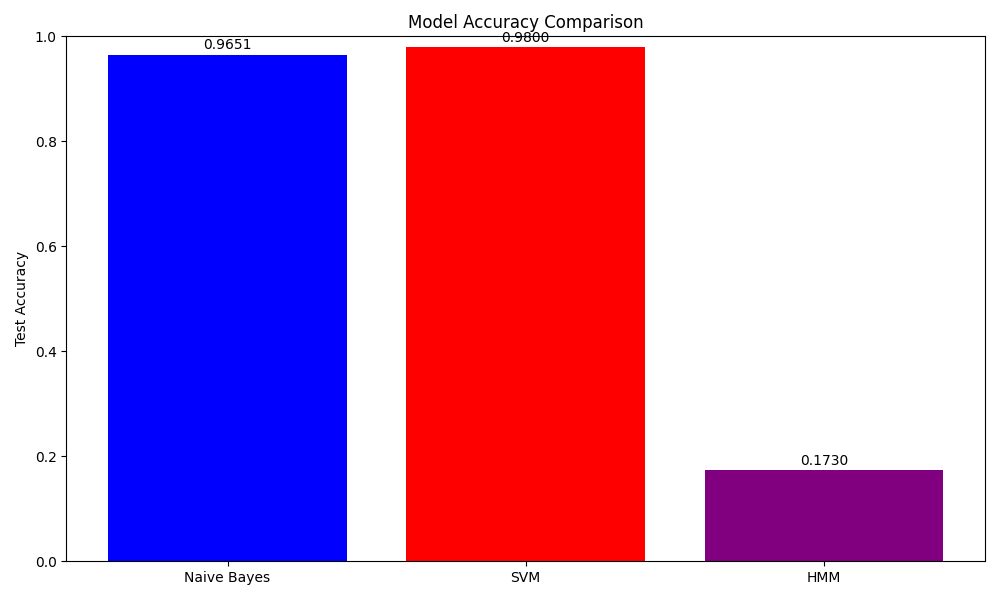
# VI. Results

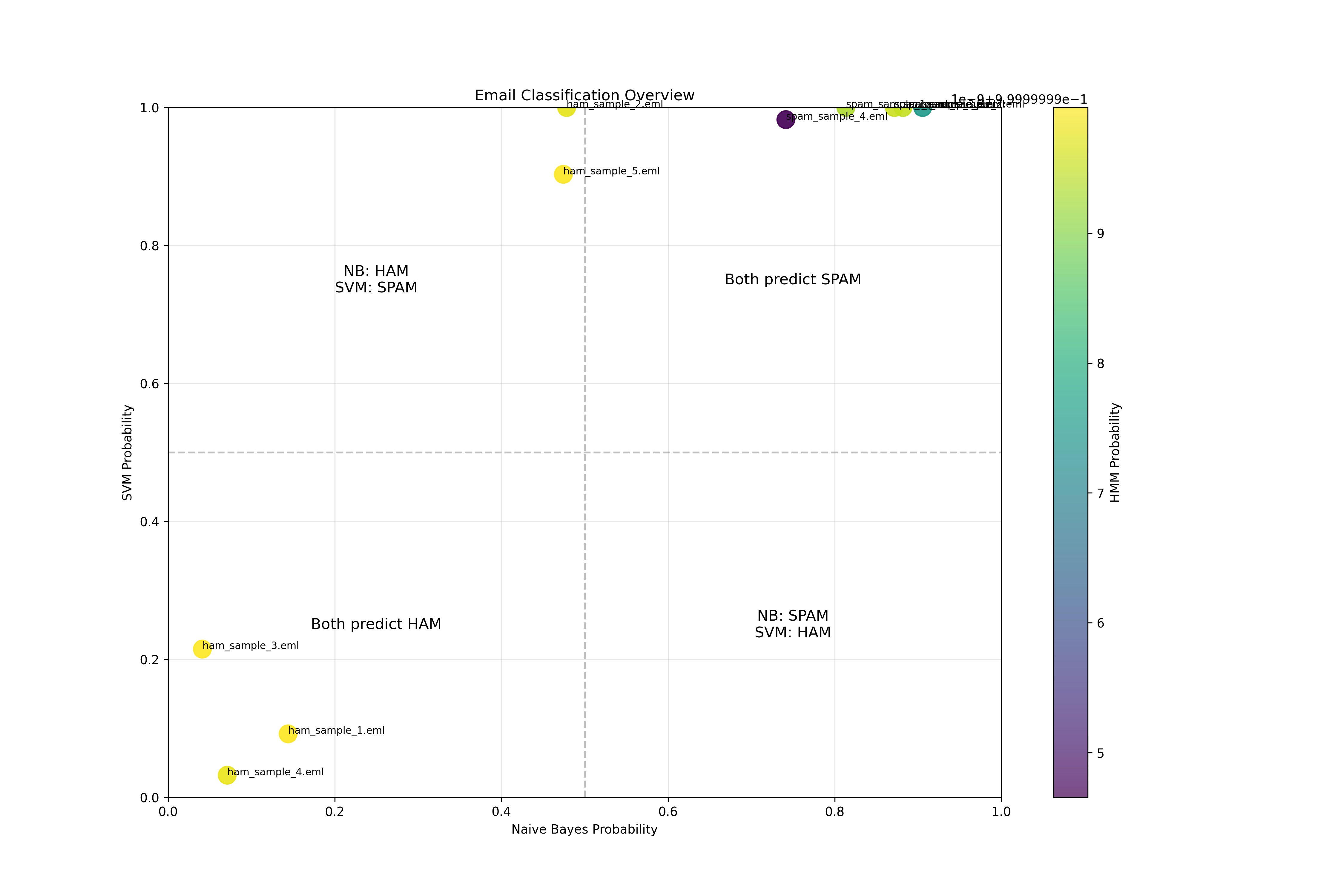
By the fourth week of the project, we aim to complete:  
- Preprocessing and feature extraction  
- Implementation and evaluation of Naïve Bayes as the baseline  
- Initial SVM classifier training and evaluation  
- HMM integration strategy  
- Preliminary results for Naïve Bayes and SVM (including confusion matrices and accuracy metrics)











# VII. Conclusion

This project aims to create a robust spam detection system by leveraging the complementary strengths of HMMs, Naïve Bayes, and SVMs. The initial phase will establish baseline performance and explore how contextual modeling with HMMs can enhance spam detection.

# VIII. References

- SpamAssassin Dataset: https://spamassassin.apache.org/old/publiccorpus/  
- Scikit-learn: https://scikit-learn.org/  
- hmmlearn Documentation: https://hmmlearn.readthedocs.io/  
- Research papers listed in Section II