E-Mail Spam Classification

YZV 311E Term Project Proposal

Abdullah Bilici
Artificial Intelligence & Data Engineering
Istanbul Technical University
bilicia@itu.edu.tr
150200330

Bora Boyacıoğlu
Artificial Intelligence & Data Engineering
Istanbul Technical University
boyacioglu20@itu.edu.tr
150200310

***Abstract*—This documents is the intermediate report for the YZV 311E Data Mining course term project. It is about an E-Mail classification model using techniques like Natural Language Processing. It is being held by Abdullah Bilici and Bora Boyacıoğlu, Istanbul Technical University. Team number is 8.**

***Index Terms*—email, spam, classification, natural language processing**

I. Introduction

II. Related Work

III. Proposed Work

A. Model Selection

There are four models defined in our project proposal: **Naive Bayes**, **Support Vector Machines (SVM)**, **Random Forest** and **Logistic Regression**. Before using them, it is crucial to know their strengths and weaknesses. We will explain them in this section.

1) Naive Bayes: This model is based on Bayes' Theorem. It is a probabilistic model that uses the probability of each attribute belonging to each class to make a prediction. It is called naive because it assumes that all the attributes are independent of each other. This assumption is not true in real life but it is still a good model for classification problems. It is also a fast model to train and predict. For text classification problems, it is a preferable model.

2) Support Vector Machines (SVM): This model is based on the idea of finding a hyperplane that separates the data into classes. It is a supervised learning model that can be used for both classification and regression problems. For complex problems, SVM is a powerful model to use.

3) Random Forest: This model is based on the idea of creating multiple decision trees and combining them to get a better result. It is a supervised learning model that can be used for both classification and regression problems. It is a powerful model that can be used for complex problems.

4) Logistic Regression: This model is based on the idea of finding the best fitting S-shaped curve for the data.

IV. Experimental Results

We tried four different models with the goal of finding the one that pursues our goal the best. As stated in the project proposal, our models are **Naive Bayes**, **Support Vector Machines (SVM)**, **Random Forest** and **Logistic Regression**. Each one has their own advantages and disadvantages. We will try to find the best one for our problem.

A. Naive Bayes

At the first trial, we used the **Multinomial Naive Bayes** model for our problem. It is a variation of the Naive Bayes model that is used for text classification problems. It is based on the multinomial distribution of the data. The results were not that great, however. One thing about Naive Bayes is that it performs poorly when the data violates the independence assumption.

We saw low results, especially with the recall value. You can see the Validation Set results in Table I and the Test Set results in Table II.

TABLE I
NAIVE BAYES VALIDATION SET RESULTS

Accuracy	**Precision**	**Recall**	**F1 Score**
0.8392	1.0	0.3881	0.5592

True Positives: 111
False Positives: 0
False Negatives: 175
True Negatives: 802

TABLE II
NAIVE BAYES TEST SET RESULTS

Accuracy	**Precision**	**Recall**	**F1 Score**
0.8566	1.0	0.4046	0.5761

True Positives: 106
False Positives: 0
False Negatives: 156
True Negatives: 826

B. Support Vector Machines (SVM)

In our experiment, we used **SVM** after we tried Naive Bayes. There definitely is an improvement. You can see the Validation Set results in Table III and the Test Set results in Table IV. However, the results may be better. We will continue trying to find a better model.

TABLE III
SVM VALIDATION SET RESULTS

Accuracy	Precision	Recall	F1 Score
0.9586	0.9959	0.8462	0.9149

True Positives: 242

False Positives: 1

False Negatives: 44

True Negatives: 801

TABLE IV
SVM TEST SET RESULTS

Accuracy	Precision	Recall	F1 Score
0.9504	0.9906	0.8015	0.8861

True Positives: 210

False Positives: 2

False Negatives: 52

True Negatives: 824

C. Random Forest

This time, all scores are above 90%, which is a sign that this is the correct model. Especially with hyperparameter tuning, these scores may be improved. Of course we will try yet another model, but Random Forest seems to be the best one so far. You can see the Validation Set results in Table V and the Test Set results in Table VI.

We used `n_estimators=100` and `random_state=42` as our parameters for Random Forest.

TABLE V
RANDOM FOREST VALIDATION SET RESULTS

Accuracy	Precision	Recall	F1 Score
0.9779	0.9852	0.9301	0.9586

True Positives: 266

False Positives: 4

False Negatives: 20

True Negatives: 798

TABLE VI
RANDOM FOREST TEST SET RESULTS

Accuracy	Precision	Recall	F1 Score
0.9798	0.9839	0.9313	0.9569

True Positives: 244

False Positives: 4

False Negatives: 18

True Negatives: 822

D. Logistic Regression

After Random Forest, this model was a disappointment. It is a good model but it did not perform well for our problem. You can see the Validation Set results in Table VII and the Test Set results in Table VIII.

We used `max_iter=10000` as our parameters for Logistic Regression.

TABLE VII
LOGISTIC REGRESSION VALIDATION SET RESULTS

Accuracy	Precision	Recall	F1 Score
0.8860	1.0	0.5664	0.7232

True Positives: 162

False Positives: 0

False Negatives: 124

True Negatives: 802

TABLE VIII
LOGISTIC REGRESSION TEST SET RESULTS

Accuracy	Precision	Recall	F1 Score
0.8898	1.0	0.5840	0.7373

True Positives: 153

False Positives: 0

False Negatives: 109

True Negatives: 826

V. CONCLUSION

REFERENCES

[1] None