Contents

1 Introduction 9

1.1 Problem . . . 9

1.2 Motivation . . . 9

1.3 Challenges . . . 10

1.4 Related Work . . . 11

1.5 Overview of the Project - DEADLINE: 9TH APRIL . . . 12

2 Preparation 13

2.1 Requirements Analysis . . . 13

2.2 Changes from the Initial Proposal . . . 13

2.3 Starting Point - DEADLINE: 11TH APRIL . . . 13

2.4 Introduction to Supervised Learning . . . 13

2.5 Introduction to the Naïve Bayes Classifier . . . 14

2.6 Introduction to Support Vector Machines . . . 15

2.7 Introduction to Evaluating Supervised Learning Systems . . . 20

2.7.1 Train/Test Split . . . 20

2.7.2 Cross Validation . . . 20

2.7.3 Evaluation Metrics . . . 20

2.8 Introduction to the Bag-of-Words Model . . . 21

2.9 Introduction to the House of Commons . . . 22

2.10 Software Engineering Techniques . . . 22

2.11 Implementation Approach . . . 22

2.12 Choice of Tools . . . 22

2.13 Summary - DEADLINE: 11TH APRIL . . . 22

3 Implementation 23

3.1 Overview . . . 23

3.2 Data Acquisition . . . 23

3.2.1 Scraping Data . . . 23

3.2.2 Cleaning Data . . . 23

3.2.3 Database - DEADLINE: 13TH APRIL . . . 23

3.3 Classifier . . . 23

3.3.1 Constructing Features . . . 23

3.3.2 Optimisations . . . 23

3.4 Summary - DEADLINE: 16TH APRIL . . . 23

4 Evaluation Code - DEADLINE: 20TH APRIL 25

5 Evaluation 27

5.1 Unit Testing 27

5.2 Internal Evaluation 27

5.2.1 Manual Checks 27

5.2.2 Comparison of Optimisations 27

5.2.3 Baseline Comparison - DEADLINE: 23RD APRIL 27

5.3 External Evaluation 27

5.3.1 Spam Email Dataset 27

5.3.2 Comparisons With Related Work 27

5.4 Evaluation of Project Goals 27

5.5 Summary - DEADLINE: 24TH APRIL 27

6 Conclusions 29

6.1 Achievements 29

6.2 Lessons Learned 29

6.3 Future Work - DEADLINE: 25TH APRIL 29

7 Diagrams - DEADLINE: 27TH APRIL 31

Bibliography 31

A Project Proposal 35