Contents

1 Introduction **9**
- 1.1 Motivation . . . 9
- 1.2 Problem Formulation . . . 9
- 1.3 Related Work . . . 10
- 1.4 Overview of the Project . . . 10

2 Preparation **13**
- 2.1 Requirements Analysis . . . 13
- 2.2 Changes from the Initial Proposal . . . 13
- 2.3 Starting Point - DEADLINE: 9TH APRIL . . . 15
- 2.4 Introduction to Supervised Learning . . . 15
- 2.5 Introduction to the Naïve Bayes Classifier . . . 16
- 2.6 Introduction to Support Vector Machines . . . 17
- 2.7 Introduction to Evaluating Supervised Learning Systems . . . 22
  - 2.7.1 Train/Test Split . . . 22
  - 2.7.2 Cross Validation . . . 22
  - 2.7.3 Evaluation Metrics . . . 22
- 2.8 Introduction to the Bag-of-Words Model . . . 23
- 2.9 Software Engineering Techniques . . . 24
- 2.10 Implementation Approach . . . 24
- 2.11 Choice of Tools . . . 24
- 2.12 Summary - DEADLINE: 11TH APRIL . . . 24

3 Implementation **25**
- 3.1 Overview . . . 25
- 3.2 Data Acquisition . . . 25
  - 3.2.1 Scraping Data . . . 25
  - 3.2.2 Cleaning Data . . . 25
  - 3.2.3 Database - DEADLINE: 13TH APRIL . . . 25
- 3.3 Classifier . . . 25
  - 3.3.1 Constructing Features . . . 25
  - 3.3.2 Optimisations . . . 25
- 3.4 Summary - DEADLINE: 16TH APRIL . . . 25

4 Evaluation Code - DEADLINE: 20TH APRIL **27**

5 Evaluation **29**
5.1 Unit Testing . . . 29
5.2 Internal Evaluation . . . 29
5.2.1 Manual Checks . . . 29
5.2.2 Comparison of Optimisations . . . 29
5.2.3 Baseline Comparison - DEADLINE: 23RD APRIL . . . 29
5.3 External Evaluation . . . 29
5.3.1 Spam Email Dataset . . . 29
5.3.2 Comparisons With Related Work . . . 29
5.4 Evaluation of Project Goals . . . 29
5.5 Summary - DEADLINE: 24TH APRIL . . . 29

6 Conclusions **31**
6.1 Achievements . . . 31
6.2 Lessons Learned . . . 31
6.3 Future Work - DEADLINE: 25TH APRIL . . . 31

7 Diagrams - DEADLINE: 27TH APRIL **33**

Bibliography **33**

A Project Proposal **37**