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 12

2 Preparation **13**
2.1 Requirements Analysis 13
2.2 Changes from the Initial Proposal 13
2.3 Starting Point 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 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 23
3.3 Classifier 23
3.3.1 Constructing Features 23
3.3.2 Optimisations 23
3.4 Summary 23

4 Evaluation **25**
4.1 Unit Testing . . . 25
4.2 Internal Evaluation . . . 25
4.2.1 Manual Checks . . . 25
4.2.2 Comparison of Optimisations . . . 25
4.2.3 Baseline Comparison . . . 25
4.3 External Evaluation . . . 25
4.3.1 Spam Email Dataset . . . 25
4.3.2 Comparisons With Related Work . . . 25
4.4 Evaluation of Project Goals . . . 25
4.5 Summary . . . 25

5 Conclusions **27**
5.1 Achievements . . . 27
5.2 Lessons Learned . . . 27
5.3 Future Work . . . 27

Bibliography **27**

A Project Proposal **31**