

Contents

1	Summary	4
1.1	Acknowledgements	4
2	Introduction	5
2.1	Aims and requirements	5
2.1.1	Aim	5
2.1.2	Product	5
2.1.3	Evaluation of product	5
2.1.4	Objectives	5
2.1.5	Minimum requirements	6
2.1.6	Possible extensions	6
2.2	The task	6
3	Background research	7
3.1	Background reading	7
3.2	Data sets	7
3.3	Current research	8
3.4	Methodologies	10
3.4.1	Supervised learning	10
3.4.2	Unsupervised learning	11
3.5	Technologies	12
3.5.1	WEKA	12
3.5.2	GATE	13
3.5.3	NLTK	13
3.5.4	Stop lists	13
4	Design and Project management	15
4.1	Project plan	15
4.1.1	Project steps	15
4.2	Methodology	19
4.2.1	Agile Methodology	20
4.3	Choice of programming language	22
4.4	Design choices	23
4.4.1	Data set	23
4.4.2	Features	23
4.5	Classifiers	25

4.5.1	Naive Bayes	25
4.5.2	Decision trees	25
4.5.3	Planned classifiers	27
5	Implementation	28
5.1	Overview	28
5.2	Implementation processes	28
5.2.1	Data set preparation and preprocessing	29
5.2.2	Development of ARFF creation program	29
5.2.3	Experimental testing	30
5.2.4	Production of results	31
5.3	Overview of Implementation	31
6	Evaluation	32
6.1	Experimental evaluation	32
6.2	Evaluation by comparison	32
6.3	Accuracy	35
6.4	Classifier and feature set effectiveness	35
6.4.1	Features	40
6.5	Overview	40
6.5.1	Future work	42
	Bibliography	42
	Appendices	46
A	Reflection	46
B	WEKA experiment results example	47