

## TABLE OF CONTENTS

UNDERGRADUATE THESIS .....	1
TABLE OF CONTENTS .....	3
LIST OF TABLES.....	6
LIST OF FIGURES.....	7
ABSTRACT.....	9
CHAPTER 1 INTRODUCTION .....	10
1.1 Research Background.....	10
1.2 Research Problem.....	12
1.3 Research Scope .....	12
1.4 Research Objective.....	12
1.5 Research Benefits .....	12
CHAPTER II LITERATURE REVIEW .....	14
CHAPTER III THEORETICAL BASIS.....	27
3.1 Natural Language Processing.....	27
3.2 Sentiment Analysis.....	27
3.3 Preprocessing .....	29
3.3.1 Text Cleaning .....	29
3.3.2 Tokenization.....	29
3.3.3 Stopword Removal.....	30
3.3.4 Stemming/Lemmatization .....	30
3.4 Term Frequency-Inverse Document Frequency (TF-IDF).....	30
3.4 Bag-of-Words.....	32
3.5 Naïve Bayes Classification .....	34
3.5.1 Multinomial Naïve Bayes .....	35
3.5.2 Gaussian Naïve Bayes.....	36

3.5.3 Bernoulli Naïve Bayes .....	37
3.6 KNN Algorithm.....	38
3.7 Support Vector Machines .....	41
3.7.1 Linear Kernel .....	42
3.7.2 Polynomial Kernel .....	43
3.8 K-Fold Cross Validation.....	45
3.9 Confusion Matrix .....	47
CHAPTER IV RESEARCH AND ANALYSIS .....	50
4.1 Research Description .....	50
4.2 Research Dataset .....	51
4.3. Preprocessing .....	53
4.4. Feature Extraction .....	55
4.5 Data Splitting .....	56
4.6 K-Fold Cross Validation.....	56
4.7 Model Training and Construction .....	56
4.8 Model Evaluation .....	58
4.9 Results Analysis .....	58
CHAPTER V IMPLEMENTATION .....	59
5.1 Hardware and Software Specifications .....	59
5.2 Text Analysis & Data Visualization .....	59
5.3 Text Preprocessing .....	62
5.4. Feature Extraction .....	64
5.5 Data Splitting and K-Fold Cross Validation.....	65
5.6 Model Building and Evaluation .....	66
5.6.1 Naïve Bayes Algorithm .....	66
5.6.2 KNN Algorithm.....	71
5.6.3 Support Vector Machines .....	73



CHAPTER VI RESULTS AND DISCUSSION .....	77
6.1 Dataset Splitting Results .....	77
6.2 Naïve Bayes Model Results .....	77
6.3 K Nearest Neighbor Model Results .....	82
6.4 Support Vector Machines Model Results .....	85
6.5 Results Comparison Between All Models.....	89
CHAPTER VII CONCLUSION AND SUGGESTION.....	92
7.1 Conclusion .....	92
7.2 Suggestions .....	92
REFERENCES.....	93

## LIST OF TABLES

Table 2.1 Comparison between research studies.....	21
Table 4.1 Example Process of Text Cleaning.....	53
Table 4.2 Example Process of Tokenization .....	54
Table 4.3 Example Process of Stopword Removal .....	54
Table 4.4 Example Process of Stemming.....	54
Table 4.5 Illustration of the preprocessing phase using an example data. ....	55
Table 4.6 Hyperparameter Values .....	57
Table 6.2 The Results of The Implementation of Naïve Bayes.....	80
Table 6.3 K-Fold Cross Validation Results for Naïve Bayes Classification .....	81
Table 6.4 Confusion Matrix Results for the KKN Algorithm.....	84
Table 6.5 K-Fold Cross Validation Results for the KNN Algorithm .....	85
Table 6.6 Comparison of Results of Both Linear and Polynomial SVM.....	87
Table 6.7 K-Fold Cross Validation for The Support Vector Machines Models .....	88
Table 6.8 Confusion Matrix Results for All Models.....	89
Table 6.9 K-Fold Cross Validation Results for All Models.....	90

## LIST OF FIGURES

Figure 3.1 Similar data points existing close to each other (Harrison, 2018).....	38
Figure 3.2 Overview of SVM (Gandhi, 2018). .....	41
Figure 3.3 The Cycle of K-Fold Cross Validation (Pandian, 2023).....	46
Figure 3.4 The Flow of the K-fold-defined size. ....	46
Figure 3.5 K-Fold Cross Validation flow (Pandian, 2023). ....	46
Figure 4.1 Research Flow .....	51
Figure 4.2 A Sample of The Dataset .....	51
Figure 4.3 The Amount of Data for Each Sentiment. ....	52
Figure 4.4 The Pie Chart for Total Percentage of Each Sentiment. ....	52
Figure 4.5 The Pie Chart of Two Sentiments. ....	53
Figure 5.1 Dropping Rows in the Sentiment Column with Missing Values. ....	59
Figure 5.2 Replacing the Irrelevant Sentiment with Neutral Sentiment. ....	59
Figure 5.3 Removing the Neutral Sentiments .....	60
Figure 5.4 Dataset Description.....	60
Figure 5.5 Number of Characters and Words Visualization.....	61
Figure 5.6 WordCloud for the Negative Sentiment.....	61
Figure 5.7 WordCloud for the Positive Sentiment .....	62
Figure 5.8 Training Dataset defined as DF .....	62
Figure 5.9 Preprocessing Phase.....	63
Figure 5.10 Applying the Preprocessing Methods into the Dataset. ....	64
Figure 5.11 The results of the Preprocessing Phase. ....	64
Figure 5.12 TF-IDF for Feature Extraction.....	64
Figure 5.13 TF-IDF Usage. ....	65
Figure 5.14 Data Splitting .....	65
Figure 5.15 K-Fold Cross Validation .....	65

Figure 5.16 Naïve Bayes Parameter Distribution. ....	67
Figure 5.17. RandomizedSearchCV for Hyperparameter Tuning.....	68
Figure 5.18. The Initiation of the Naïve Bayes Models .....	69
Figure 5.19. Confusion Matrix Evaluation .....	70
Figure 5.20. The process of determining the Cross Validation Score .....	70
Figure 5.21 Hyperparameter Distribution for KNN.....	71
Figure 5.22 KNN Hyperparameter Tuning. ....	72
Figure 5.23 KKN Model Evaluation .....	73
Figure 5.24. Linear Support Vector Machines Parameter Distribution.....	74
Figure 5.25. Polynomial Support Vector Machines Distribution.....	75
Figure 5.26. Support Vector Machines Model Evaluation. ....	76
Figure 5.27. The Function to Obtain the K-Fold Cross Validation Score. ....	76
Figure 6.1 Multinomial Naïve Bayes Confusion Matrix .....	78
Figure 6.2 Gaussian Naïve Bayes Confusion Matrix.....	78
Figure 6.3 Bernoulli Naïve Bayes Confusion Matrix. ....	79
Figure 6.4 KNN Confusion Matrix .....	82
Figure 6.5 KNN Confusion Matrix with The Manhattan Distance.....	83
Figure 6.6 Linear Support Vector Machines Confusion Matrix.....	86
Figure 6.7 Polynomial Support Vector Machines Confusion Matrix.....	87