

TABLE OF CONTENT

| | |
|---|-------------|
| HALAMAN PENGESAHAN..... | iii |
| PLAGIARISM STATEMENT..... | iv |
| PREFACE..... | v |
| TABLE OF CONTENT..... | vi |
| LIST OF TABLES..... | viii |
| LIST OF FIGURES..... | x |
| INTISARI..... | xii |
| ABSTRACT..... | xiii |
| CHAPTER 1 | |
| INTRODUCTION..... | 1 |
| 1.1. Research Background..... | 1 |
| 1.2. Research Problem..... | 2 |
| 1.3. Research Objective..... | 2 |
| 1.4. Research Scope..... | 3 |
| 1.5. Benefits of the Research..... | 3 |
| 1.6. Research Writing System..... | 3 |
| CHAPTER 2 | |
| LITERATURE REVIEW..... | 5 |
| CHAPTER 3 | |
| THEORETICAL BASIS..... | 14 |
| 3.1. Sentiment Analysis..... | 14 |
| 3.2. Lexicon-based Approach..... | 15 |
| 3.3. Negation Handling..... | 17 |
| 3.4. Code-Mixed Sentiment Analysis..... | 18 |
| 3.5. Sirekap Application..... | 18 |
| 3.6. DANA Application..... | 21 |
| 3.7. Data Preprocessing..... | 22 |
| 3.8. Feature Extraction..... | 23 |
| 3.9. TF-IDF..... | 24 |
| 3.10. Machine Learning Classifier..... | 26 |
| 3.11. Support Vector Machine (SVM)..... | 27 |
| 3.12. Naive Bayes..... | 30 |
| 3.13. Performance Evaluation..... | 31 |
| 3.14. Multi-Class Performance Evaluation..... | 34 |
| CHAPTER 4 | |
| METHODOLOGY..... | 38 |
| 4.1. Research Description..... | 38 |
| 4.2. Research Step..... | 38 |

| | |
|---------------------------------------|-----------|
| 4.2.1. Data Collection..... | 40 |
| 4.2.2. Handling Code-Mixed..... | 42 |
| 4.2.3. Data Preprocessing..... | 42 |
| 4.2.4. Negation Handling..... | 45 |
| 4.2.5. Data Labeling..... | 46 |
| 4.2.6. Feature Extraction..... | 47 |
| 4.2.6. Split and Train Dataset..... | 49 |
| 4.2.7. SVM Model..... | 49 |
| 4.2.8. Naive Bayes Model..... | 50 |
| 4.2.9. Performance Evaluation..... | 50 |
| CHAPTER 5 | |
| IMPLEMENTATION..... | 54 |
| 5.1. Data Collection..... | 54 |
| 5.1.1. Sirekap Dataset..... | 54 |
| 5.1.2. DANA Dataset..... | 56 |
| 5.2. Data Preprocessing..... | 57 |
| 5.2.1. Handling Code-Mixed..... | 57 |
| 5.2.2. Case Folding..... | 59 |
| 5.2.3. Non-ASCII Removal..... | 59 |
| 5.2.4. Emoji Removal..... | 59 |
| 5.2.5. Non-Alpha Numeric Removal..... | 60 |
| 5.2.6. Tokenization..... | 61 |
| 5.2.7. Slang Words Conversion..... | 61 |
| 5.2.8. Negation Handling..... | 62 |
| 5.2.9. Stopword Removal..... | 64 |
| 5.2.10. Stemming..... | 66 |
| 5.3. Data Labeling..... | 67 |
| 5.4. Split and Train Dataset..... | 70 |
| 5.5. Feature Extraction..... | 71 |
| 5.6. SVM Model..... | 71 |
| 5.7. Naive Bayes Model..... | 73 |
| 5.7. Performance Evaluation..... | 74 |
| CHAPTER 6 | |
| RESULT AND DISCUSSION..... | 76 |
| 6.1. Data Collection..... | 76 |
| 6.2. Data Preprocessing..... | 77 |
| 6.3. Negation Handling..... | 78 |
| 6.4. Data Labeling..... | 78 |
| 6.5. Split Dataset Result..... | 80 |
| 6.6. SVM Model..... | 81 |

| | |
|--|-----------|
| 6.7. Naive Bayes Model..... | 82 |
| 6.8. Performance Evaluation Results..... | 82 |
| CHAPTER 7 | |
| CONCLUSION..... | 85 |
| 7.1. Conclusion..... | 85 |
| 7.2. Future Works..... | 85 |
| REFERENCES..... | 86 |

LIST OF TABLES

| | |
|---|----|
| Table 2.1 Comparison of Naïve Bayes and SVM (Patel, 2017)..... | 8 |
| Table 2.2 Literature Review..... | 12 |
| Table 3.1 Extracted Data from Sirekap App in Google Play Store..... | 20 |
| Table 3.2 Extracted Data from DANA App in Google Play Store..... | 21 |
| Table 3.3 NLP Feature Extraction Techniques Comparison (Eskandar, 2023)..... | 23 |
| Table 3.4 Comparison of Supervised Learning Classifiers with SVM (IBM, 2023)..... | 29 |
| Table 3.5 Confusion Matrix..... | 32 |
| Table 4.1 Sample Dataset of Sirekap Google Play Store Reviews..... | 41 |
| Table 4.2 Sample Dataset of DANA Google Play Store Reviews..... | 42 |
| Table 4.3 Illustration of Preprocessing Stages..... | 44 |
| Table 4.4 Illustration of Negation Handling Process..... | 45 |
| Table 4.5 Star Rating to Sentiment Label Conversion..... | 46 |
| Table 4.6 Example of Manually Labeling..... | 46 |
| Table 4.7 Example of TF and IDF Calculation..... | 48 |
| Table 4.8 Example of TF-IDF Calculation..... | 48 |
| Table 4.9 Confusion Matrix Sentiment Class using SVM (Patel, 2017)..... | 51 |
| Table 5.1 Sample of Labeled Sirekap Dataset..... | 68 |
| Table 5.2 Sample of Labeled DANA Dataset..... | 69 |
| Table 6.1 Column Descriptions for Sirekap Dataset..... | 76 |
| Table 6.2 Column Descriptions for DANA Dataset..... | 77 |
| Table 6.3 Data Size Comparison..... | 77 |
| Table 6.4 Negation Handling Counter Results..... | 78 |
| Table 6.5 Sentiment Label Distribution..... | 80 |
| Table 6.6 Sentiment Label Distribution in Sirekap Train and Test Sets..... | 80 |
| Table 6.7 Sentiment Label Distribution in DANA Train and Test Sets..... | 81 |
| Table 6.8 Hyperparameter Tuning Results..... | 82 |
| Table 6.9 Hyperparameter Tuning Results..... | 82 |
| Table 6.10 Performance Evaluation Results..... | 83 |

LIST OF FIGURES

| | |
|---|----|
| Figure 2.1 R2 Result Comparison (Pröllochs et al., 2020)..... | 7 |
| Figure 2.2 The F1 Scores of Various Experiments (Duong & Thi, 2021)..... | 9 |
| Figure 2.3 Performance Comparison of Classifiers (Mukherjee et al., 2021)..... | 10 |
| Figure 2.4 Performance of Classifiers on New Dataset (Mukherjee et al., 2021)..... | 11 |
| Figure 3.1 Sentiment Analysis Process Steps (Aqlan et al., 2019)..... | 15 |
| Figure 3.2 Lexicon-Based Approach Flow (Thangavel & Lourdusamy, 2023)..... | 15 |
| Figure 3.3 Sentiment Classification (Aqlan et al., 2019)..... | 17 |
| Figure 3.4 User Interface of Sirekap Application (KPU, 2024)..... | 19 |
| Figure 3.5 Web Scraping Process for Sirekap App (Google Play Store)..... | 20 |
| Figure 3.6 Support Vector Machine (IBM, 2023)..... | 27 |
| Figure 3.7 Bayes Theorem..... | 30 |
| Figure 3.8 Confusion Matrix 3x3..... | 33 |
| Figure 3.9 Multi-class Visual Example (Evidently AI, n.d.)..... | 34 |
| Figure 3.10 Calculation of All Micro-Average Metrics (Leung, 2022)..... | 36 |
| Figure 3.11 Calculation of Weighted F1 Score (Leung, 2022)..... | 37 |
| Figure 4.1 Research Steps..... | 39 |
| Figure 4.2 Sirekap Application in Google Play Store..... | 40 |
| Figure 4.3 Preprocessing Steps..... | 43 |
| Figure 4.4 TF-IDF Flowchart..... | 47 |
| Figure 4.5 Hyperparameter Tuning SVM..... | 50 |
| Figure 4.6 Hyperparameter Tuning Naive Bayes..... | 50 |
| Figure 5.1 Installing Google-play-scraper Library..... | 54 |
| Figure 5.2 Scraping Application Reviews..... | 55 |
| Figure 5.3 Sample Application Reviews..... | 55 |
| Figure 5.4 Sample DANA Reviews..... | 56 |
| Figure 5.5 Data Preprocessing Scenarios..... | 57 |
| Figure 5.6 Handling Code-Mixed Using Googletrans..... | 58 |
| Figure 5.7 Applying Translate Function..... | 58 |
| Figure 5.8 Case Folding Function..... | 59 |
| Figure 5.9 Non-ASCII Removal Function..... | 59 |
| Figure 5.10 Emoji Removal Function..... | 60 |
| Figure 5.11 Non-Alpha Numeric Removal Function..... | 60 |
| Figure 5.12 Tokenization Function..... | 61 |
| Figure 5.13 Indonesian Slang Words Dictionary..... | 61 |
| Figure 5.14 Indonesian Slang Words Dictionary..... | 62 |
| Figure 5.15 Negation Handling Function..... | 63 |

| | |
|--|----|
| Figure 5.16 Apply Negation Handling..... | 63 |
| Figure 5.17 Negation Handling Counter..... | 64 |
| Figure 5.18 Stopword List..... | 65 |
| Figure 5.19 Stopword Removal Function..... | 65 |
| Figure 5.20 Stemming Function..... | 66 |
| Figure 5.20 Splitting the Dataset..... | 70 |
| Figure 5.21 Applying TF-IDF..... | 71 |
| Figure 5.22 Hyperparameter Tuning the SVM Model..... | 72 |
| Figure 5.23 Running the SVM Model..... | 73 |
| Figure 5.24 Bernoulli Naive Bayes..... | 74 |
| Figure 5.25 Multinomial Naive Bayes..... | 74 |
| Figure 5.26 Macro-Average Performance Evaluation..... | 75 |
| Figure 6.1 Sirekap Sentiment Label Distribution Graph..... | 79 |
| Figure 6.2 DANA Sentiment Label Distribution Graph..... | 79 |