



CLASSIFYING AND PREDICTING THE RATING SENTIMENT OF WOMEN'S E-COMMERCE CLOTHING REVIEWS

ICORIS 2023 - PAPER 86





OUR TEAM

Immanuel
Yabes

Computer Science Departement
Bina Nusantara University

Novita
Aryanti

Computer Science Departement
Bina Nusantara University

Rayes Jordan
Pradana

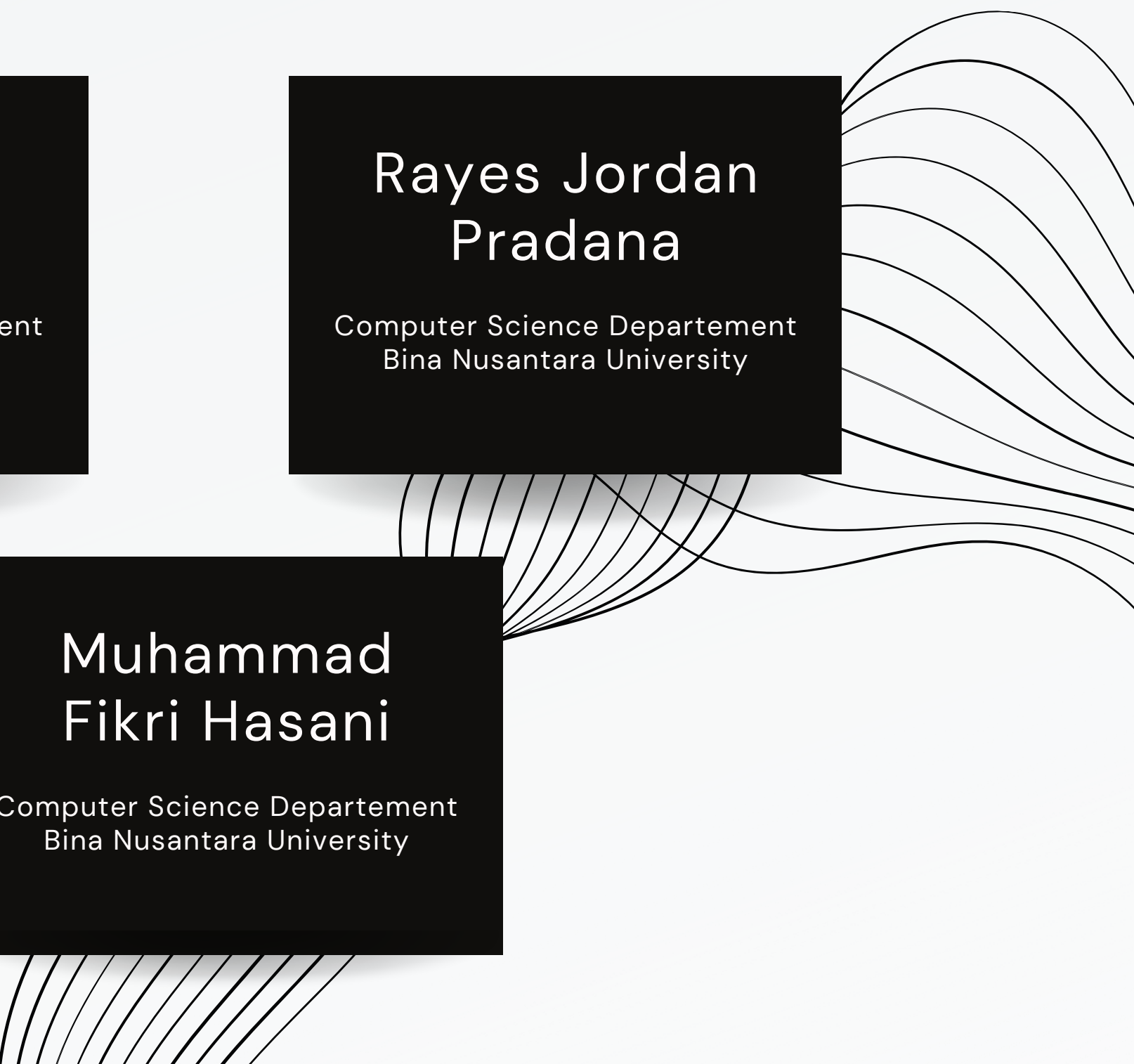
Computer Science Departement
Bina Nusantara University

Karli Eka
Setiawan

Computer Science Departement
Bina Nusantara University

Muhammad
Fikri Hasani

Computer Science Departement
Bina Nusantara University



INTRODUCTION

Rating Prediction



- Ratings given by consumers may not always align with their comments, requiring the need for review rating prediction.
- The performance of models, including Support Vector Machine (SVM), Artificial Neural Network (ANN), and Bidirectional Encoder Representations from Transformers (BERT), is compared.

- The experiments utilize a clothing e-commerce review dataset, focusing on accuracy and F1 Score results.
- The dataset undergoes pre-processing and feature selection, selecting four key features: Review Text, Rating, Recommended IND, and Positive Feedback Count

Sentiment Analysis



METHOD: DATASET

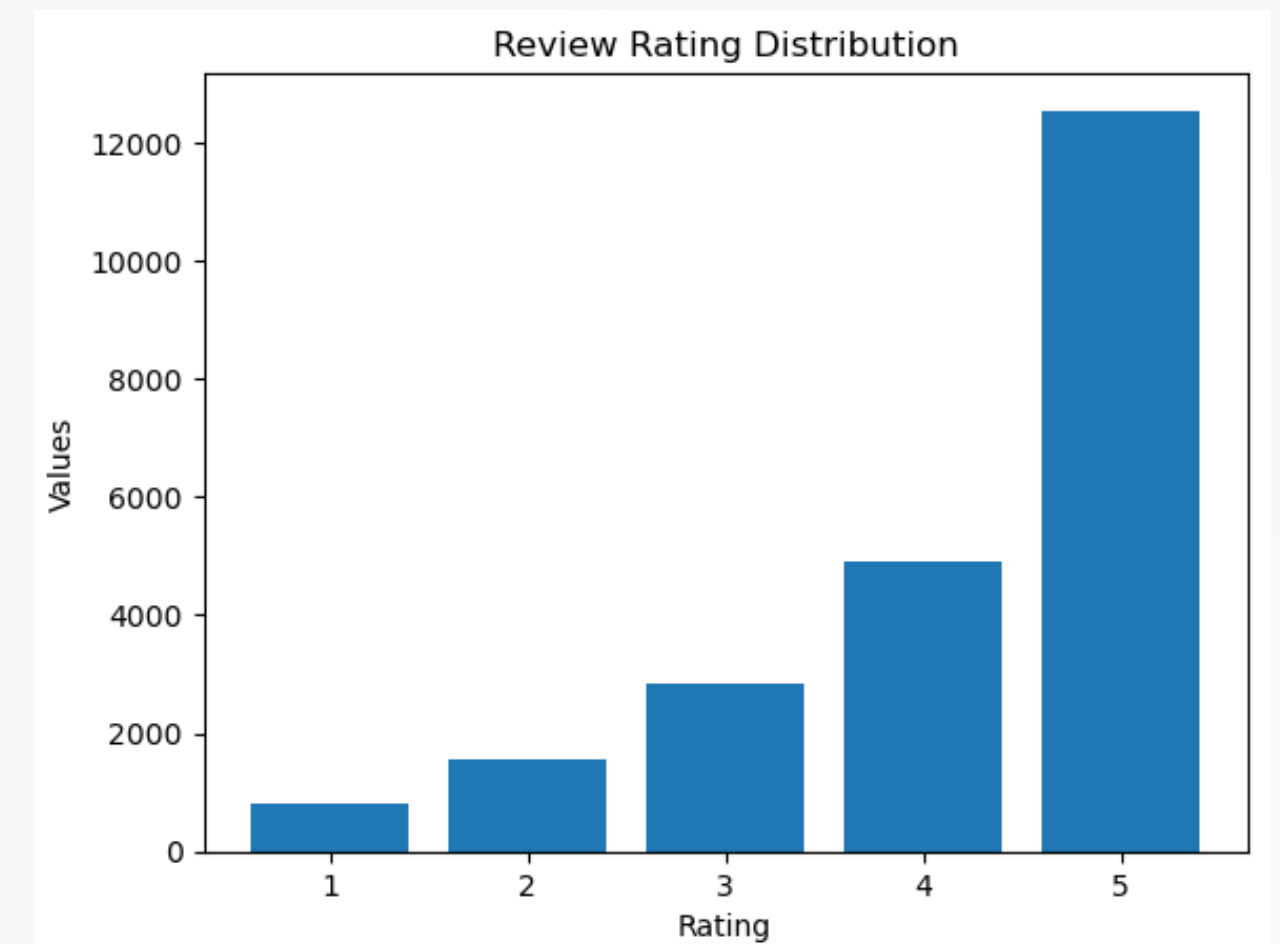


Women's E-Commerce Clothing Reviews

23,000 Customer Reviews and Ratings

[kaggle.com](https://www.kaggle.com)

- The dataset has 23486 rows and 10 feature, contains approximately 23,000 Customer Reviews and Ratings.
- The features Clothing ID, Age, Title, Review Text, Rating, Recommended IND, Positive Feedback Count, Division Name, Department Name, and Class Name represent a customer review.
- The research focused on 4 features, "Review Text, Rating, Recommended IND, and Positive Feedback Count"





METHOD: TEXT VECTORIZATION

TF-IDF

TF-IDF is a numerical statistic used in natural language processing and information retrieval to measure the importance of a term in a document. It is obtained by multiplying the TF and IDF values for a term, and the higher the TF-IDF value, the more effective the term is to the document. It can be used in various applications, such as text mining, information retrieval, document classification, and search engine ranking.

TF-IDF is a numerical statistic used in natural language processing and information retrieval to measure the importance of a term in a document. It is obtained by multiplying the TF and IDF values for a term, and the higher the TF-IDF value, the more effective the term is to the document. It can be used in various applications, such as text mining, information retrieval, document classification, and search engine ranking.

SentenceBERT

METHOD: MODEL

SVM

A supervised machine learning algorithm used for classification and regression tasks, finding an optimal hyperplane that separates data points of different classes. It adjusts parameters during training to find the optimal hyperplane, and can classify new data points based on their position relative to the decision boundary.

ANN

ANNs are machine learning algorithms inspired by biological neural networks. They consist of interconnected nodes organized in layers, and learn through backpropagation. They are powerful for tasks such as classification, regression, and pattern recognition. They can handle large amounts of data and have been successful in various domains.

BERT

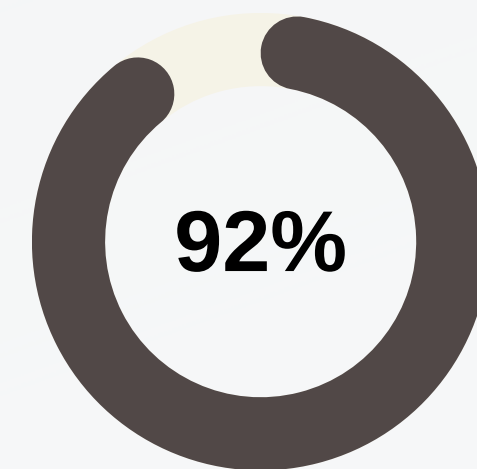
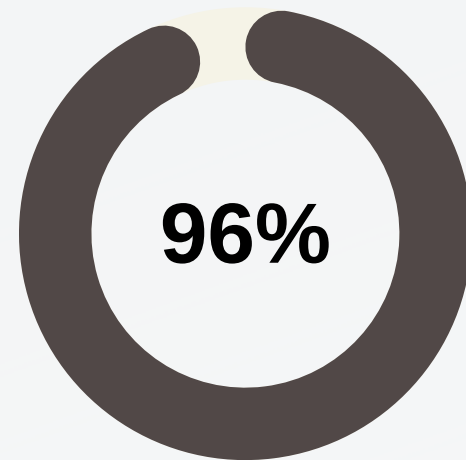
BERT is a transformer-based language model designed to capture contextual information and provide deep understanding of language by pre-training on large amounts of unlabeled text. It is pre-trained using two tasks: masked language modeling (MLM) and next sentence prediction (NSP). After pre-training, BERT can be fine-tuned on downstream tasks, such as text classification, named entity recognition, and question answering.

RESULT

TABLE I. Accuracy and F1 Score Comparison of Different Model

Model	Accuracy	F1 Score
TF-IDF - SVM	0.923	0.961
TF-IDF - ANN	0.933	0.963
sentenceBERT - BERT	0.844	0.842

CONCLUSION



- The TF-IDF text vectorization with ANN model achieved the highest F1 score of 0.963, followed by TF-IDF with SVM (F1 score: 0.961) and BERT (F1 score: 0.842).
- In terms of accuracy, the TF-IDF text vectorization with ANN model also performed the best with a score of 0.933, followed by TF-IDF with SVM (accuracy: 0.923) and BERT (accuracy: 0.844).
- Therefore, the TF-IDF text vectorization with ANN model demonstrated the best overall performance in terms of both F1 score and accuracy.

**THANK YOU
FOR WATCHING**

