Evolutionary Computation Project

Intent Classification Using Evolutionary Algorithm

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

IIT (ISM) DHANBAD

Final Evaluation Report

Course Code: CSD405

Professor: Associate Professor Annavarapu Chandra Sekhara Rao

Session: 2023-24

Group: 9

INTRODUCTION

Social media platforms like Facebook, Twitter, and Amazon gather vast amounts of data every day. But much of this data is in a mix of English and Hindi, called Hinglish. Sorting through this mix to understand what users want is tough. That's where intent classification comes in. It helps figure out what users are asking for.

In this report, we focus on classifying intents in Hinglish data. Our goal is to use evolutionary computation methods to make this process smarter. By doing so, we aim to improve how we understand and respond to users on social media.

Problem Statement

Taking data from any popular social media apps like FB/Twitter/Amazon, where data is in Hinglish format (English and Hindi combined), prepare a dataset creating two columns, where users' queries are in one column and labelled intent (purpose) in another and then perform intent classification using any of the EA.

Goals and Objectives

Goal

Demonstrate the effectiveness of Evolutionary Algorithms (EAs) in enhancing the accuracy and robustness of intent classification models.

Objective 1

Provide an overview of intent classification and its significance in natural language processing tasks, especially in the context of multilingual social media data.

Objective 2

Present the methodology for creating a dataset consisting of user queries and their labelled intents extracted from social media platforms.

Objective 3

Discuss the implementation of Evolutionary Algorithms (EAs) for feature

selection, hyperparameter optimization, and model selection in intent classification.

What has been done:

1. **Divide Intents**:
    a. For our project, we have used four Labels
        i. Customer Support
        ii. Technical Assistance
        iii. Feedback
        iv. Shopping Inquiry

2. **Data Collection** (The Idea is to collect data from various open source datasets and relabel them as per our needs, along with adding some of our own data points) Following are the sources:
    a. Manually

213	Mujhe yeh software kis version ka hai?	Technical Assistance
214	Mujhe yeh product kis price range mein hai?	Feedback
215	Mujhe yeh app kis language mein available hai?	Feedback
216	Mujhe yeh game kis processor requirement ko meet karta hai	Technical Assistance
217	Kya yeh product aaj deliver ho sakta hai?	Shopping Inquiry
218	Is phone ka warranty kab tak hai?	Shopping Inquiry
219	Yeh issue kyun aa raha hai?	Technical Assistance
220	Kya yeh device international shipping ke liye available hai?	Shopping Inquiry
221	Feedback dena chahta hoon, yeh product bahut accha hai.	Feedback
222	Kya yeh book out of stock hai?	Shopping Inquiry
223	Mere account ka password reset karna hai, kaise hoga?	Technical Assistance
224	Yeh movie ticket online book karne ka process kya hai?	Shopping Inquiry
225	Yeh product kis brand ka hai?	Feedback
226	Is feature ko kaise use karte hain?	Technical Assistance
227	Kya yeh app Android ke liye bhi available hai?	Technical Assistance
228	Kya yeh game multiplayer mode mein khela ja sakta hai?	Technical Assistance
229	Is device ke saath accessories bhi milte hain?	Feedback
230	Yeh software update kyun nahi ho raha hai?	Technical Assistance
231	Kya yeh product kisi specific location tak ship kiya ja sakta ha	Shopping Inquiry

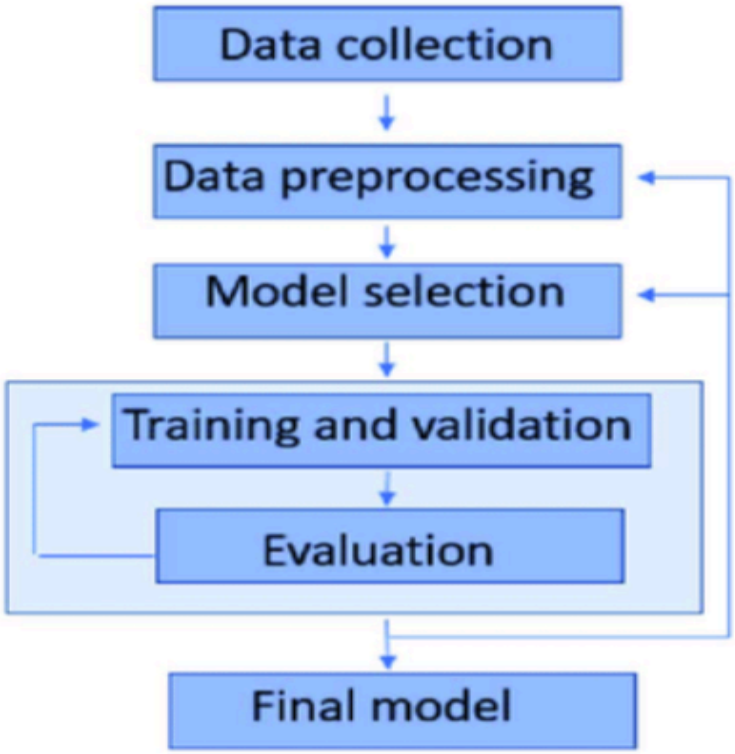
3. **Data Preprocessing**: Most of the available data comes with their own labels, which may or may not match the labels for our model. In this stage, we alter the labels to match our requirements.

- **Data Cleaning**
  - Remove any irrelevant metadata or formatting tags from the collected data.
  - Handle any encoding issues to ensure uniformity in text representation.
  - Remove duplicate entries to ensure data quality.
  - Remove common stopwords from both English and Hindi segments to reduce noise in the data.
- **Tokenization**
  - Tokenize the Hinglish text into words or subwords. Since Hinglish combines both English and Hindi, you may need a specialized tokenizer that can handle multilingual text.
- **Normalization**
  - Normalize the text by converting any Hinglish-specific variations, slang, or informal spellings to their standard forms. For example, "kya" and "क्या" can be standardized to "what".
  - Convert all text to a consistent case, usually lowercase, to avoid redundancy in vocabulary.
- **Stemming and Lemmatization**
  - Apply stemming or lemmatization to normalize words to their root forms, especially for Hindi text. This helps in reducing the vocabulary size and capturing the essence of the text.
- **Handling Code-Mixed Text**
  - Develop strategies to handle code-mixed text where English and Hindi are used interchangeably within a single sentence. This might involve segmenting the text into English and Hindi parts and applying language-specific preprocessing steps.

Work Done After Midsem

1. Model Development

Model development in machine learning involves several key steps, from data preprocessing to model evaluation.

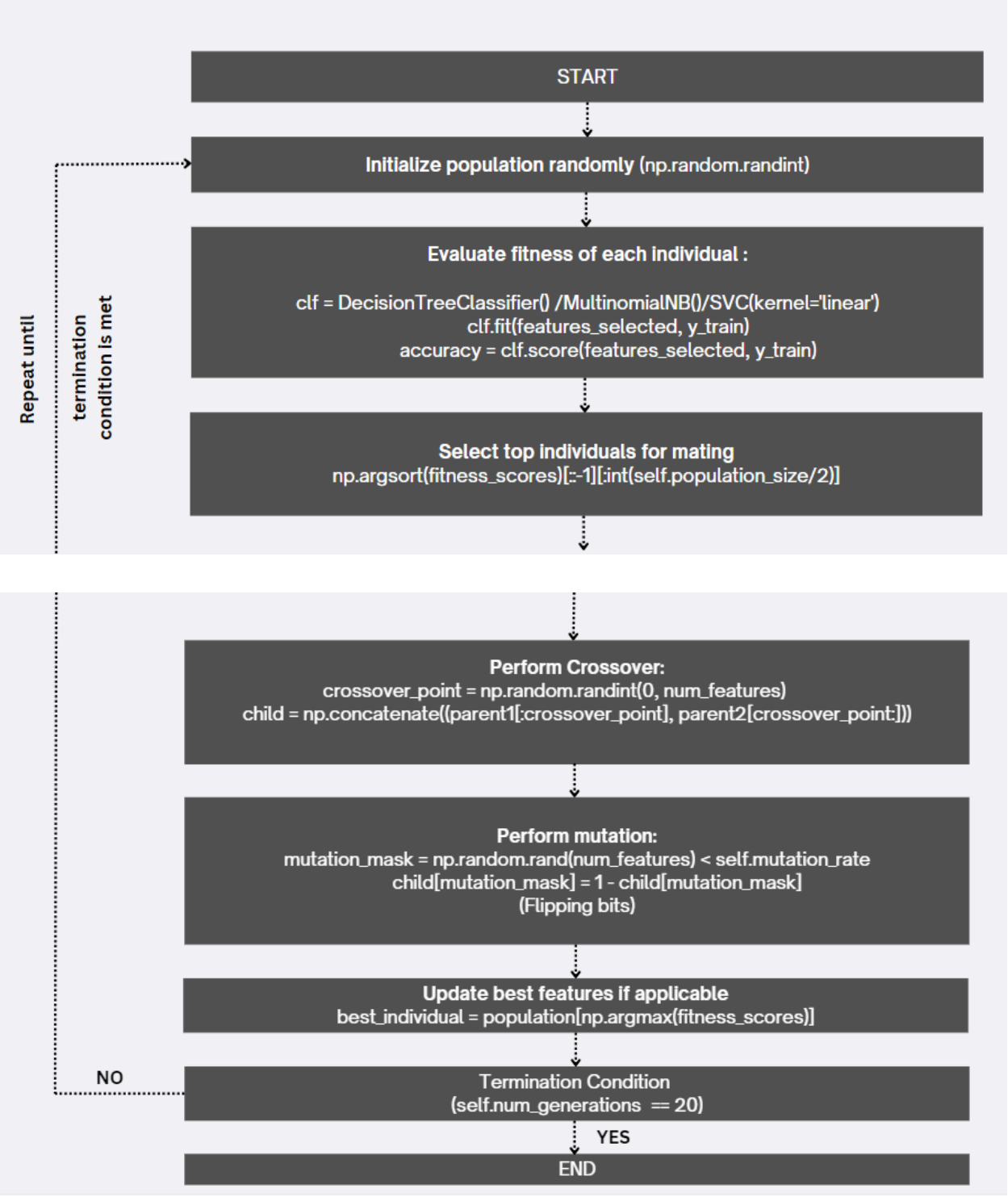


2. Feature Selection using EA

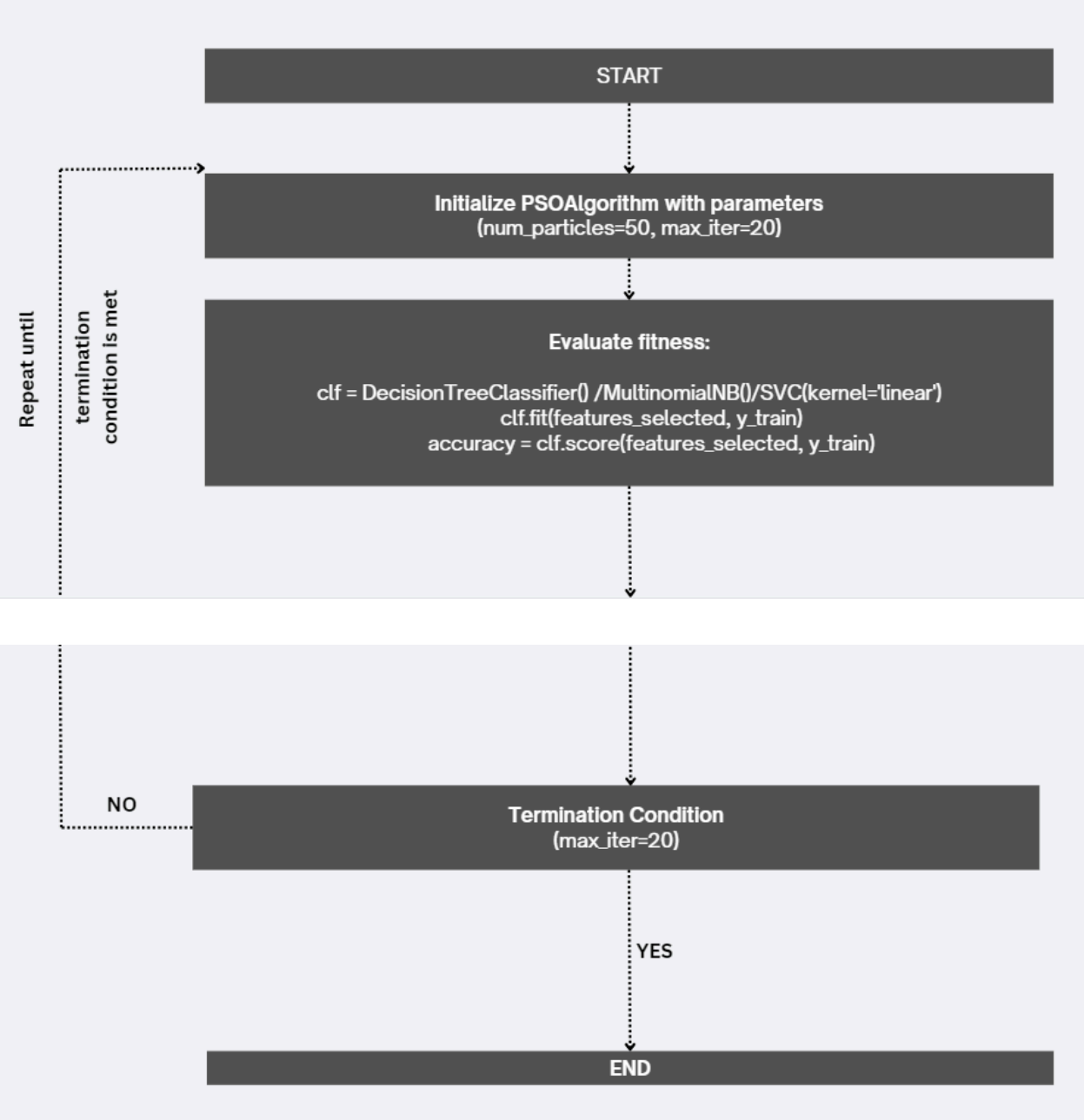
Feature selection using Evolutionary Algorithms (EA) is a powerful technique employed in machine learning to enhance model performance and efficiency by identifying the most relevant features from a large pool of potential predictors. By mimicking biological evolution, EA iteratively generates candidate subsets of features, evaluates their performance using a predefined fitness function, and

selects the fittest subsets for further refinement. This iterative process continues until an optimal subset of features, maximizing predictive accuracy while minimizing complexity, is identified. EA-based feature selection offers a flexible and automated approach, particularly beneficial for high-dimensional datasets where manual feature selection methods may be impractical or suboptimal. With its ability to handle complex search spaces and nonlinear relationships between features, EA empowers machine learning models to extract meaningful patterns and insights from data, contributing to more robust and interpretable predictive models.

Flow Chart of Feature Selection using Genetic Algorithm



Flow Chart of Feature Selection using PSO



Models Made:

1. **MNB with GA**

```
Enter your text: Product ka return process kya hai?
Predicted intent: Customer Support
```

```
Classification Report:
                    precision    recall  f1-score   support

    Customer Support       0.89      0.89      0.89        72
            Feedback       0.82      0.73      0.77        70
    Shopping Inquiry       0.82      0.83      0.82        82
Technical Assistance       0.78      0.84      0.81        82

            accuracy                           0.82       306
           macro avg       0.83      0.82      0.82       306
        weighted avg       0.82      0.82      0.82       306
```

2. **MNB with PSO**

```
Classification Report:
                    precision    recall  f1-score   support

    Customer Support       0.86      0.94      0.90        72
            Feedback       0.91      0.71      0.80        70
    Shopping Inquiry       0.80      0.87      0.83        82
Technical Assistance       0.82      0.83      0.82        82

            accuracy                           0.84       306
           macro avg       0.85      0.84      0.84       306
        weighted avg       0.84      0.84      0.84       306
```

3. **SVM with GA**

```
Classification Report:
                      precision    recall  f1-score   support

    Customer Support       0.98      0.89      0.93        72
            Feedback       0.85      0.90      0.88        70
    Shopping Inquiry       0.87      0.95      0.91        82
Technical Assistance       0.88      0.83      0.86        82

            accuracy                           0.89       306
           macro avg       0.90      0.89      0.89       306
        weighted avg       0.90      0.89      0.89       306
```

4. **SVM with PSO**

```
Classification Report:
                      precision    recall  f1-score   support

    Customer Support       0.84      0.86      0.85        72
            Feedback       0.90      0.77      0.83        70
    Shopping Inquiry       0.81      0.89      0.85        82
Technical Assistance       0.85      0.85      0.85        82

            accuracy                           0.85       306
           macro avg       0.85      0.84      0.85       306
        weighted avg       0.85      0.85      0.85       306
```

5. **Decision Tree with GA**

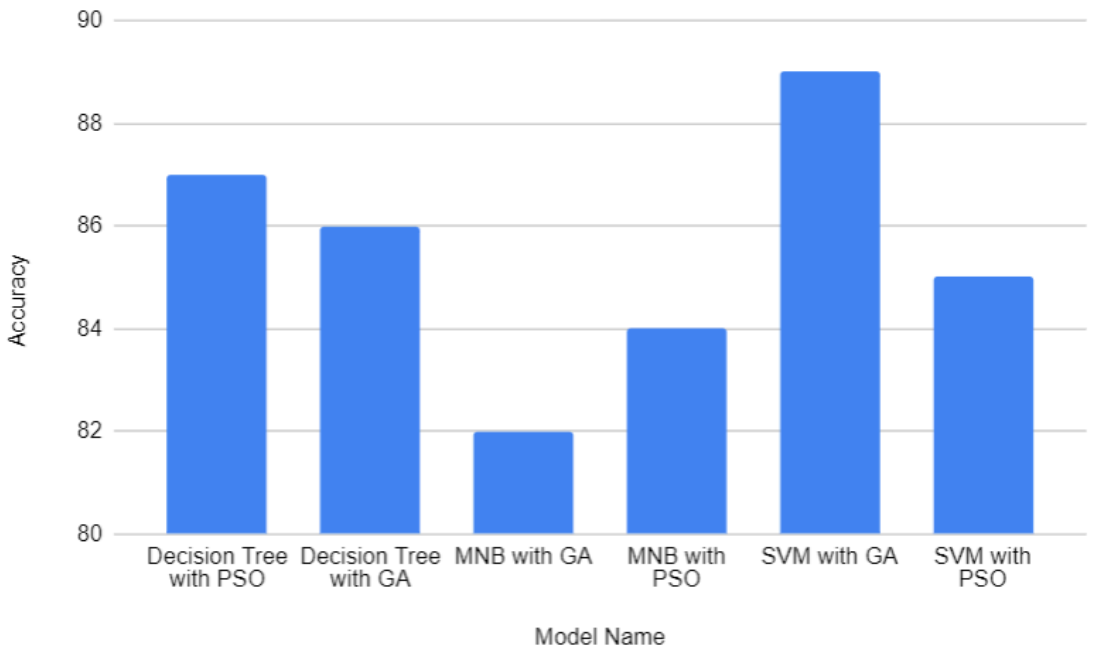
```
Classification Report:
                      precision    recall  f1-score   support

    Customer Support       0.88      0.90      0.89        72
            Feedback       0.82      0.86      0.84        70
    Shopping Inquiry       0.87      0.89      0.88        82
Technical Assistance       0.87      0.79      0.83        82

            accuracy                           0.86       306
           macro avg       0.86      0.86      0.86       306
        weighted avg       0.86      0.86      0.86       306
```

6 . Decision Tree with PSO

Classification Report:

	precision	recall	f1-score	support
Customer Support	0.88	0.88	0.88	72
Feedback	0.86	0.86	0.86	70
Shopping Inquiry	0.88	0.89	0.88	82
Technical Assistance	0.86	0.85	0.86	82
accuracy			0.87	306
macro avg	0.87	0.87	0.87	306
weighted avg	0.87	0.87	0.87	306

Overall Accuracy between all Models :



Contribution :

- **Dataset Collection: Siddharth Dhiman (20JE0948), Soutrik Das (20JE0971) , Varun Parihar (20JE1062), Ayush Ranjan (21JE0215)**
- **DataSet Preprocessing: Chirag Gajana (20JE0296), Siddharth Dhiman (20JE0948)**
- **Feature Selection: Avinash Kumar (20JE0217), Aviral Kumar Singh (20JE0220), Varun Parihar (20JE1062), Ayush Ranjan (21JE0215)**
- **Model Development: Avinash Kumar (20JE0217), Anurag Kumar(20JE0169), Soutrik Das (20JE0971), Piyush Mishra (21JE0654)**
- **Documentation: Soutrik Das (20JE0971) , Piyush Mishra (21JE0654)**

Source Code :

Project Link