TRESATA HACKATHON

Team 8

System Architecture

Approach:

The solution implements a **two-stage pipeline**:

1. **Semantic Classification**: Automatically identify column types (Phone, Company, Country, Date, Other)
2. **Parsing & Normalization**: Extract structured information from classified Phone Numbers and Company Names

The approach combines **multiple classification strategies**:

- **Feature-based ML** using Random Forest with hand-crafted features
- **Semantic embedding-based** classification using SBERT (Sentence-BERT)
- **Rule-based parsing** with specialized libraries for phone numbers and companies

Workflow

1. **Classification Pipeline** :
   Raw CSV → Feature Extraction → ML Classification → Semantic Validation
2. **Parsing Pipeline** :
   Classified Data → Specialized Parsers → Structured Output

System Architecture and Team Roles

Model Designs:

- **Random Forest**: 11 hand-crafted features (length, character types, for

Two-Stage Architecture

1. **Semantic Classification**

- **Dual Classifier Approa**mat indicators)
  - **SBERT + Logistic Regression**: 384D semantic embeddings from all-MiniLM-L6-v2
- **Output**: 5 classes (Phone, Company, Country, Date, Other)

2. **Specialized Parsing**

- **Phone Parser**: library → Country + National Number
- **Company Parser**: SBERT embeddings + cosine similarity → Base Name + Legal Suffix

Team Roles

Semantic Classification - Utkarsh Mathur
Parsing & Normalization - Pratyush Dubey
Training Data - Anushika Verma