Indonesian Sentiment Analysis Pipeline Process Flow

This document outlines the comprehensive process flow of the Indonesian Sentiment Analysis Pipeline, designed to analyze sentiment in Indonesian e-commerce product reviews. The pipeline consists of eight main phases, each contributing to the overall sentiment analysis process.

PHASE 1: Data Loading

- **Dataset Source**: Loads Indonesian e-commerce product reviews from Hugging Face (`dipawidia/ecommerce-product-reviews-sentiment`).
- **Data Type**: Reviews in Indonesian with sentiment labels (0=Negative, 1=Positive).
- **Validation**: Ensures dataset integrity, displays statistics, and handles missing values.

PHASE 2: Text Preprocessing

- **Text Cleaning**: Converts text to lowercase, removes URLs, emails, and special characters, and handles extra whitespace.
- **Indonesian Language Processing**: Normalizes slang (e.g., "gak" → "tidak"), removes stopwords, and handles colloquial expressions.

PHASE 3: Model & Embeddings

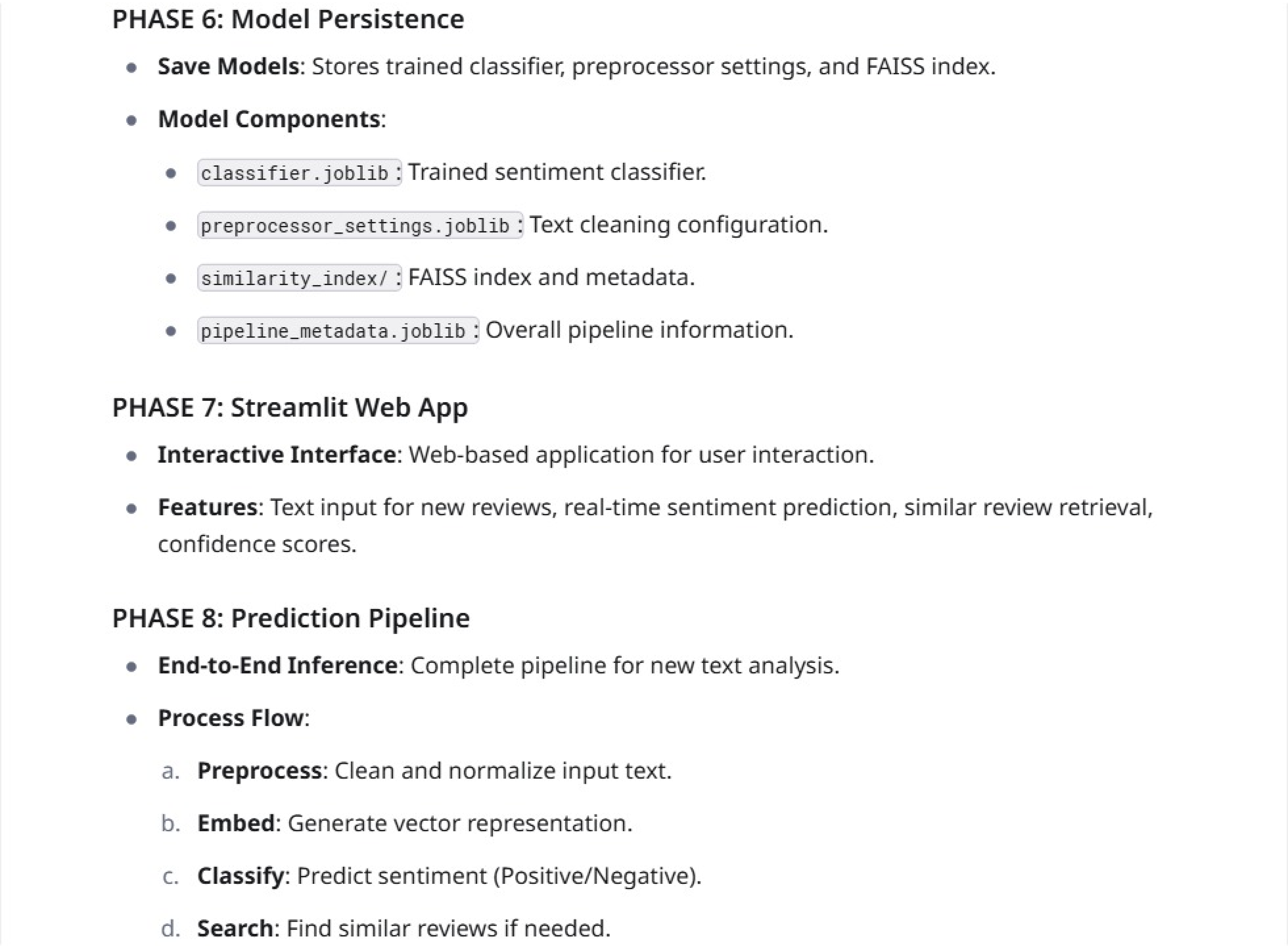
- **Sentence Transformer**: Utilizes `paraphrase-multilingual-mpnet-base-v2` model for multilingual support.
- **Vector Generation**: Converts cleaned text into 768-dimensional embeddings.
- **Batch Processing**: Efficiently processes large text volumes.

PHASE 4: Machine Learning

- **Sentiment Classifier**: Implements Logistic Regression with feature selection.
- **Training Process**: Splits data (80/20), applies cross-validation, and identifies important dimensions.
- **Model Evaluation**: Calculates accuracy, F1-score, precision, recall, and generates a confusion matrix.

PHASE 5: Similarity Search

- **FAISS Index**: Builds a fast similarity search index using FAISS.
- **Vector Similarity**: Uses cosine similarity for finding similar reviews.
- **Index Type**: `IndexFlatIP` for inner product similarity.



PHASE 6: Model Persistence

- **Save Models**: Stores trained classifier, preprocessor settings, and FAISS index.
- **Model Components**:
  - `classifier.joblib`: Trained sentiment classifier.
  - `preprocessor_settings.joblib`: Text cleaning configuration.
  - `similarity_index/`: FAISS index and metadata.
  - `pipeline_metadata.joblib`: Overall pipeline information.

PHASE 7: Streamlit Web App

- **Interactive Interface**: Web-based application for user interaction.
- **Features**: Text input for new reviews, real-time sentiment prediction, similar review retrieval, confidence scores.

PHASE 8: Prediction Pipeline

- **End-to-End Inference**: Complete pipeline for new text analysis.
- **Process Flow**:
  a. **Preprocess**: Clean and normalize input text.
  b. **Embed**: Generate vector representation.
  c. **Classify**: Predict sentiment (Positive/Negative).
  d. **Search**: Find similar reviews if needed.