HCMUT EE MACHINE LEARNING & IOT LAB

Twitter Sentiment Analysis

Presentation By: 5 anh em siêu nhân

Ho Chi Minh City – 2025

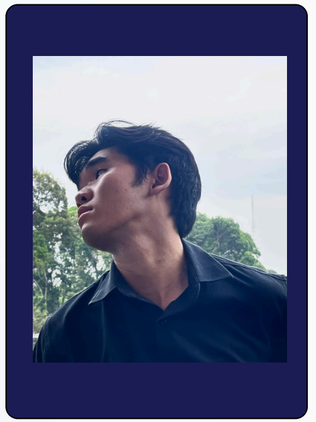
Team Members

Vũ Trí Khải

Võ Minh Khôi

Nguyễn Tiến Nam

Hoàng Minh Hải



Đoàn Trần Quốc Việt

Advisor/Mentor:

Trần Minh Huy

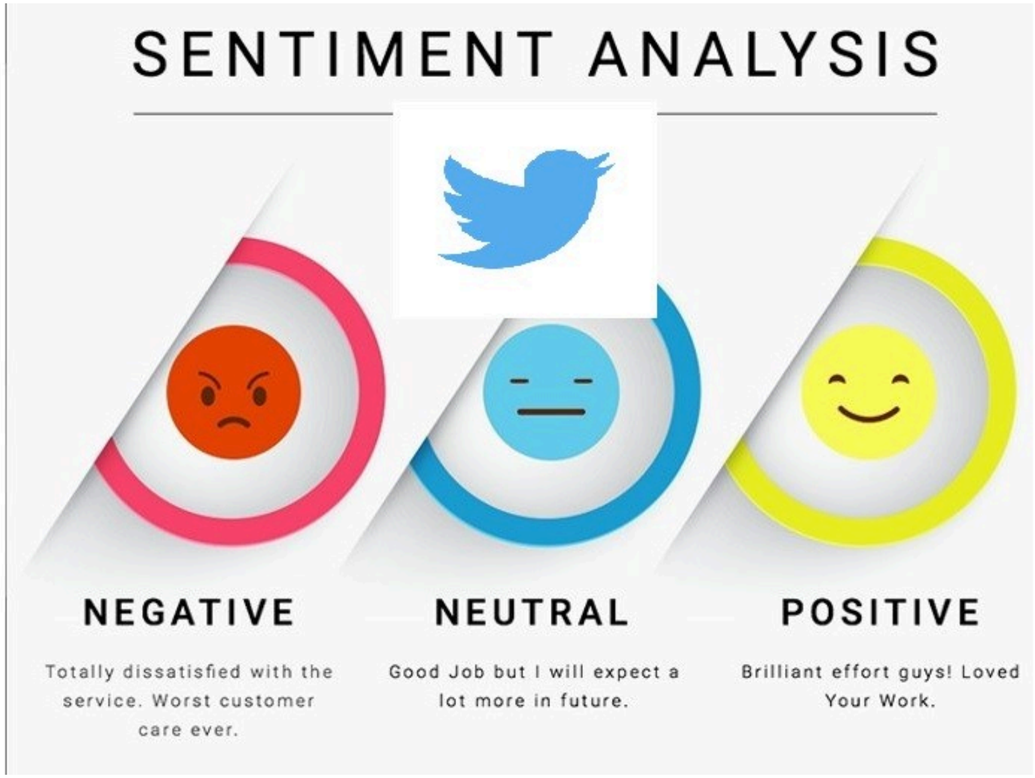
Table of Content

I. Giới thiệu

I. GIỚI THIỆU

Đặt vấn đề

Ngôn ngữ trên Twitter gặp tình trạng nhiễu, phi cấu trúc, nhiều slang, lỗi chính tả, emoji và hashtag, khiến mô hình truyền thống khó nắm bắt ngữ nghĩa và cảm xúc chính xác.

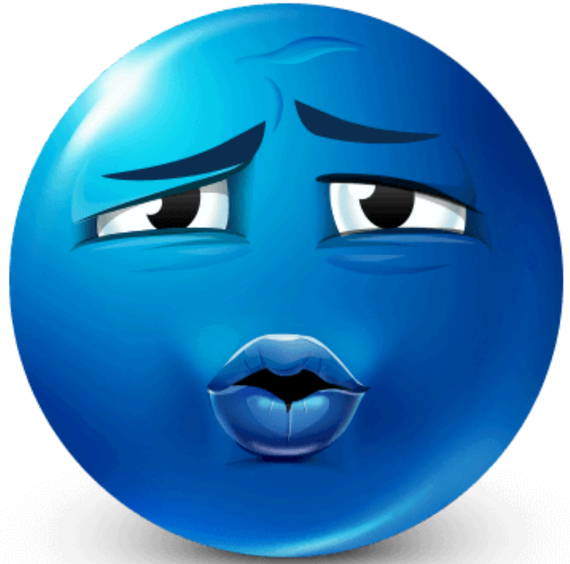


I. GIỚI THIỆU

Mục tiêu

Cần có một mô hình để xử lý các vấn đề phi cấu trúc, tiếng lóng, viết sai chính tả, các hashtag và phân tích những thành phần phi ngôn ngữ như Emoji cũng như nắm bắt các yếu tố khác như ngữ cảnh.

shutterstock.com · 1134592043



II. Tổng quan dữ liệu

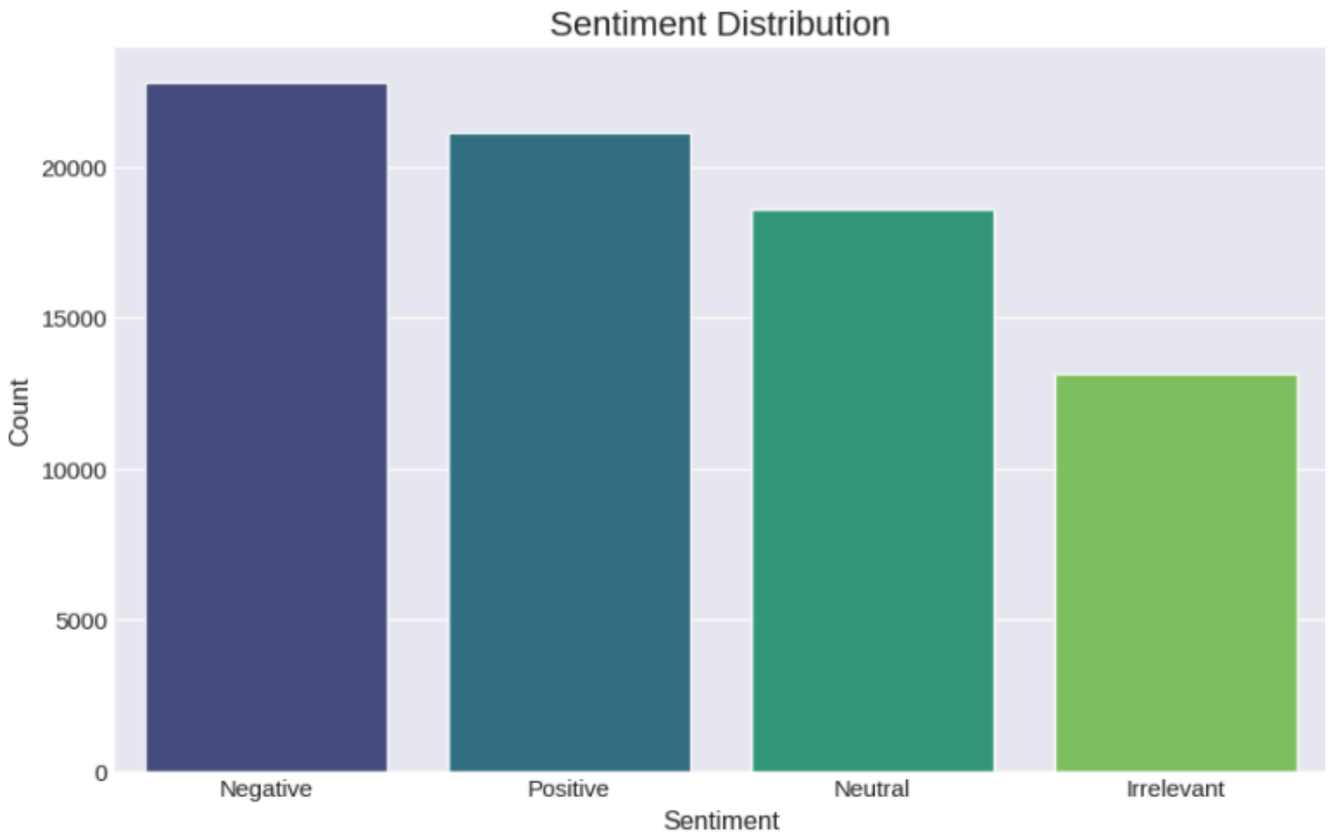
II. TỔNG QUAN DỮ LIỆU

Sample data

2401	Borderland	Positive	im getting on borderlands and i will murder you all ,
2401	Borderland	Positive	I am coming to the borders and I will kill you all,
2401	Borderland	Positive	im getting on borderlands and i will kill you all,
2401	Borderland	Positive	im coming on borderlands and i will murder you all,
2401	Borderland	Positive	im getting on borderlands 2 and i will murder you me all,
2401	Borderland	Positive	im getting into borderlands and i can murder you all,
2402	Borderland	Positive	So I spent a few hours making something for fun. . . If you don't know I am a HUGE @Borderlands fan and Maya is one of my favorite characters. So I decided to make mysel
2402	Borderland	Positive	So I spent a couple of hours doing something for fun... If you don't know that I'm a huge @ Borderlands fan and Maya is one of my favorite characters, I decided to make a w
2402	Borderland	Positive	So I spent a few hours doing something for fun... If you don't know I'm a HUGE @ Borderlands fan and Maya is one of my favorite characters.
2402	Borderland	Positive	So I spent a few hours making something for fun. . . If you don't know I am a HUGE RhandlerR fan and Maya is one of my favorite characters. So I decided to make myself a v
2402	Borderland	Positive	2010 So I spent a few hours making something for fun. . . If you don't know I am a HUGE RhandlerR fan and Maya is one of my favorite characters. So I decided to make mys

II. TỔNG QUAN DỮ LIỆU

A. SENTIMENTS DISTRIBUTION

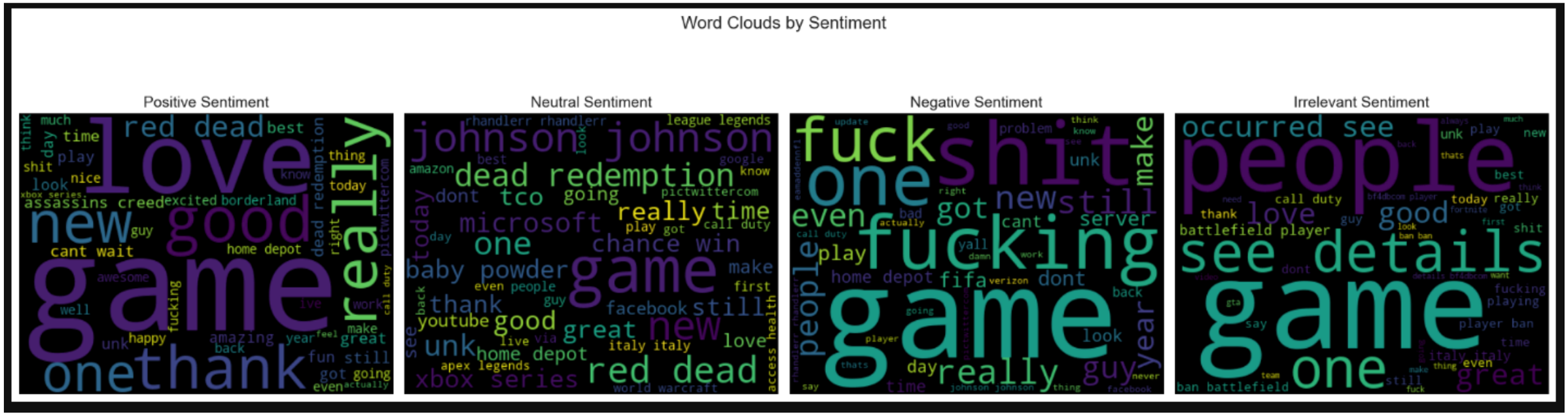


Vấn đề:

- Các sentiments phân bố không đều gây ra mất cân bằng cho bộ dữ liệu
- Khi train model dễ gây ra hiện tượng bias

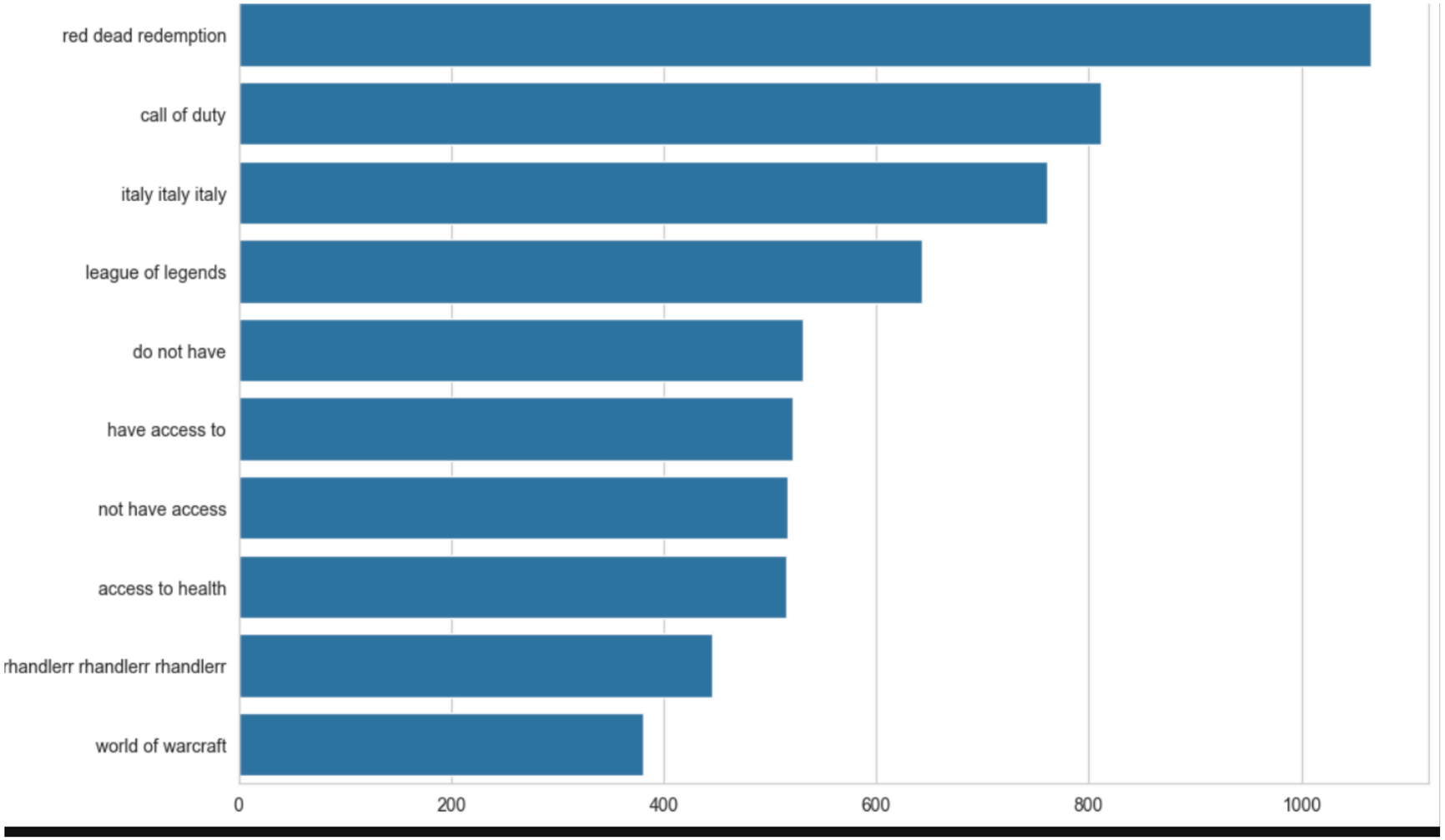
II. TỔNG QUAN DỮ LIỆU

B. WORDS DISTRIBUTION



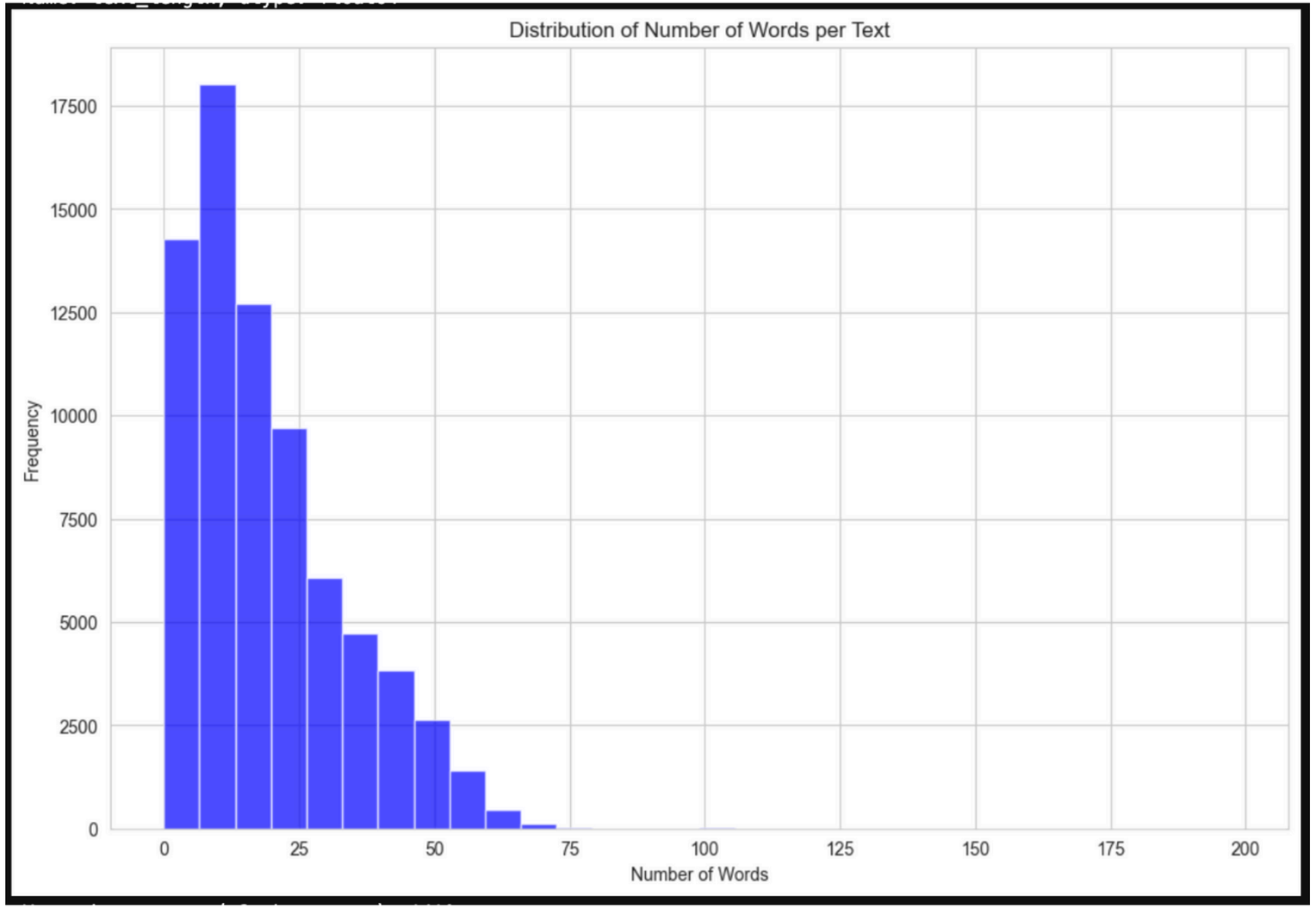
II. TỔNG QUAN DỮ LIỆU

B. WORDS DISTRIBUTION



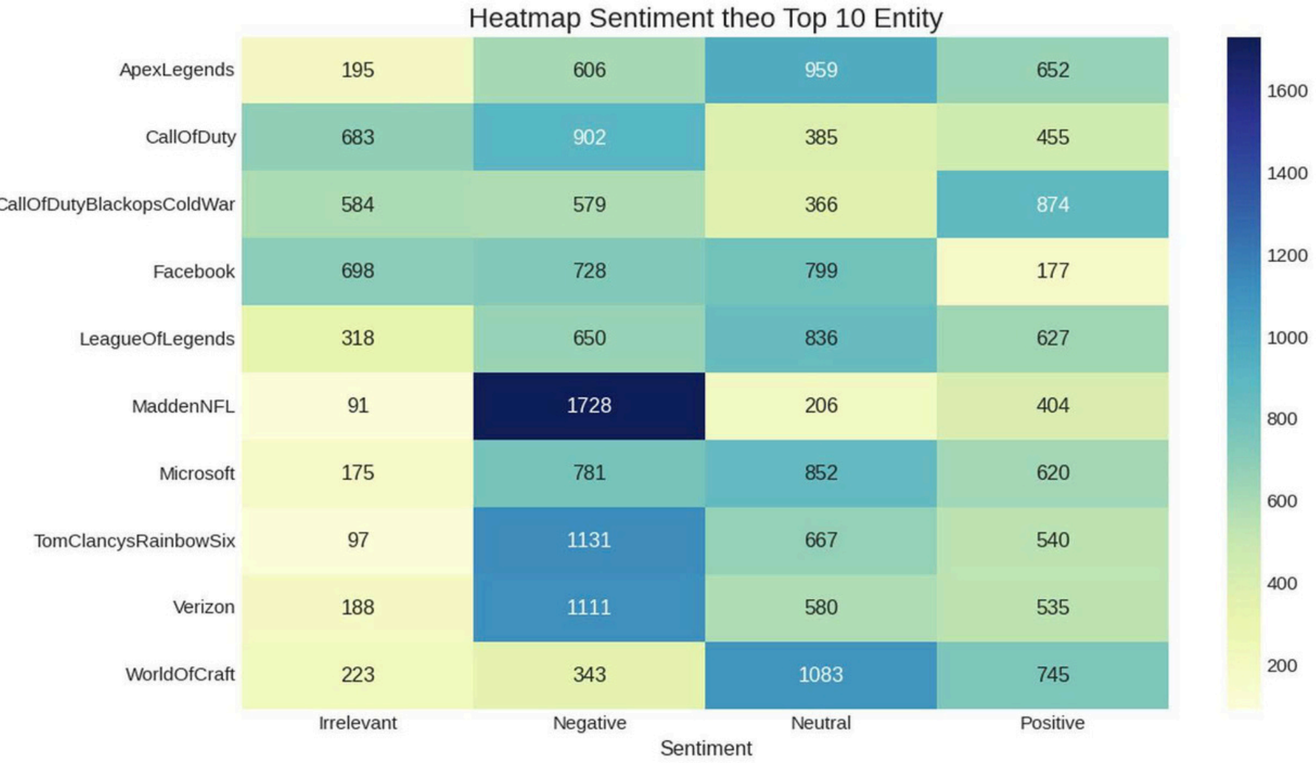
II. TỔNG QUAN DỮ LIỆU

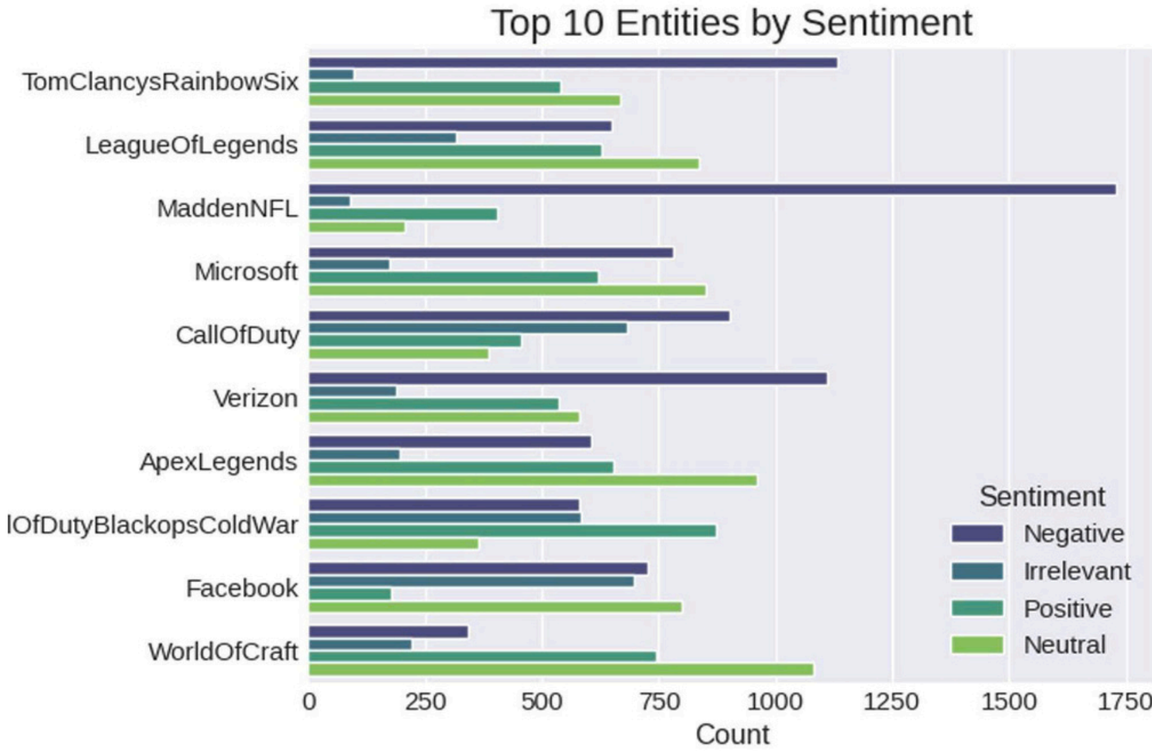
B. WORDS DISTRIBUTION



II. TỔNG QUAN DỮ LIỆU

C. SENTIMENT/ENTITY DISTRIBUTION





Sentiment-Aware Tokenization

!!!

Tiền xử lí dữ liệu

Spell Correctiom

Hashtag Processing

Sentiment Aware Tokenization

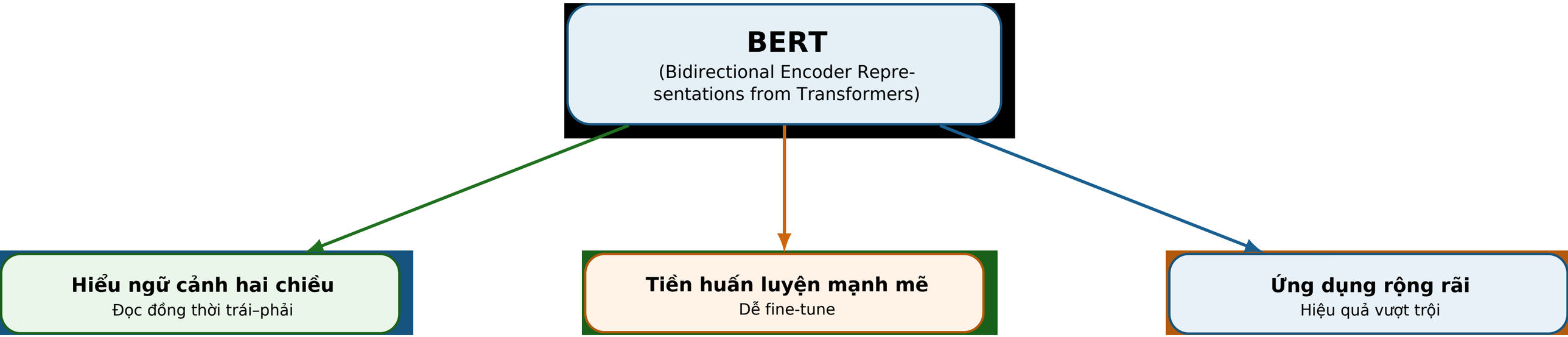
URL Handling

Mention Handling

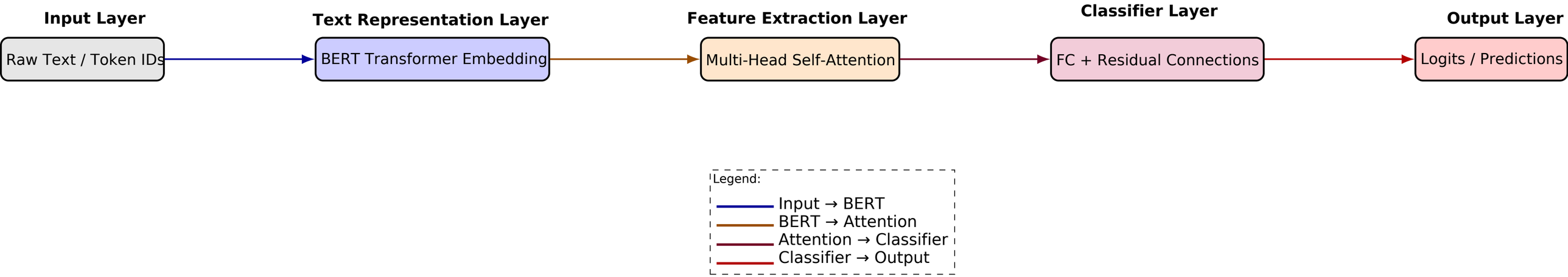
Special Character Deletion

III. Phương pháp thực hiện

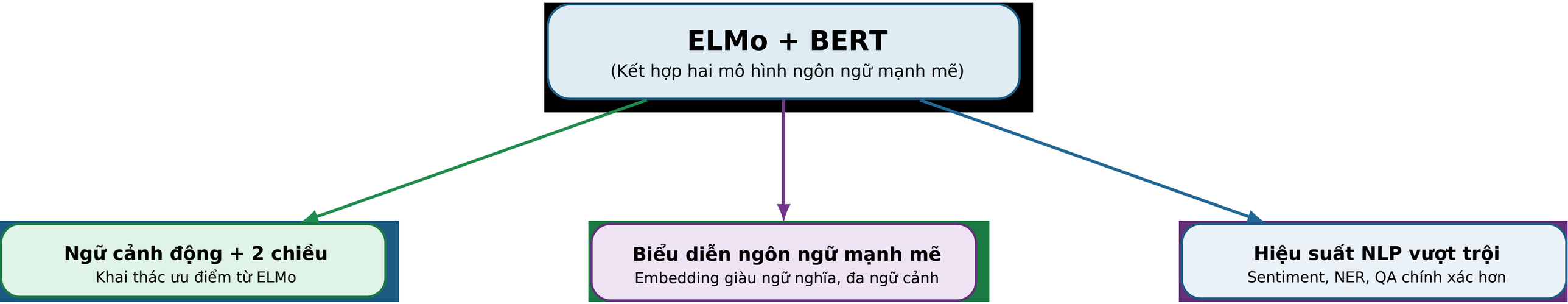
BERT


BERT
(Bidirectional Encoder Repre-
sentations from Transformers)
Hiểu ngữ cảnh hai chiều
Đọc đồng thời trái–phải
Tiền huấn luyện mạnh mẽ
Dễ fine-tune
Ứng dụng rộng rãi
Hiệu quả vượt trội

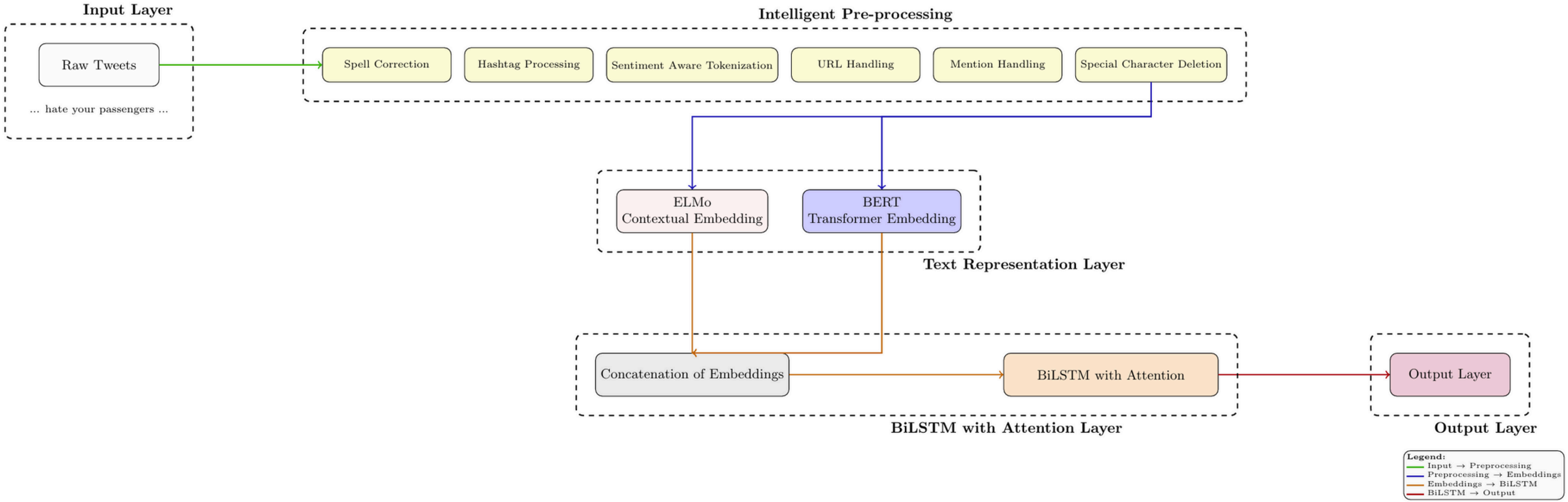
BERT


Input Layer
Raw Text / Token IDs
Text Representation Layer
BERT Transformer Embedding
Feature Extraction Layer
Multi-Head Self-Attention
Classifier Layer
FC + Residual Connections
Output Layer
Logits / Predictions
Legend:
Input → BERT
BERT → Attention
Attention → Classifier
Classifier → Output

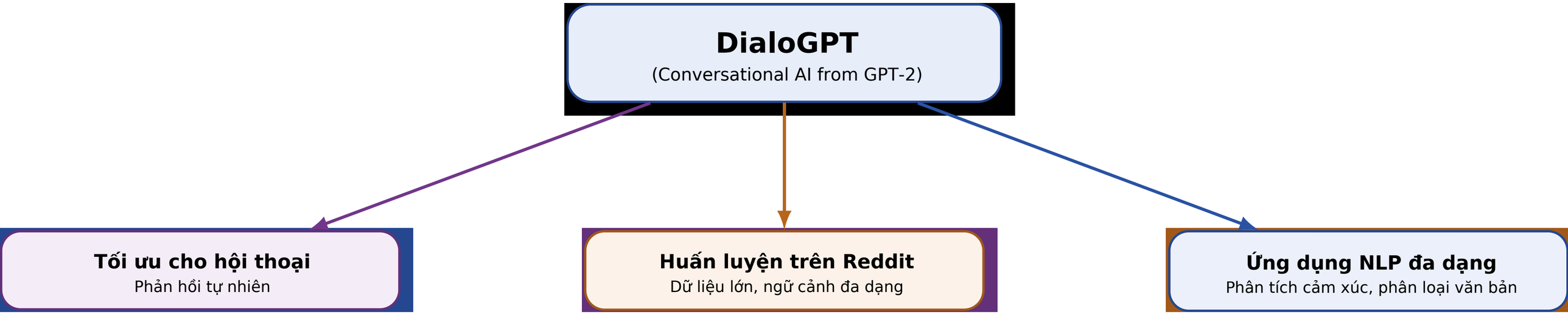
ELMo + BERT


ELMo + BERT
(Kết hợp hai mô hình ngôn ngữ mạnh mẽ)
Ngữ cảnh động + 2 chiều
Khai thác ưu điểm từ ELMo
Biểu diễn ngôn ngữ mạnh mẽ
Embedding giàu ngữ nghĩa, đa ngữ cảnh
Hiệu suất NLP vượt trội
Sentiment, NER, QA chính xác hơn

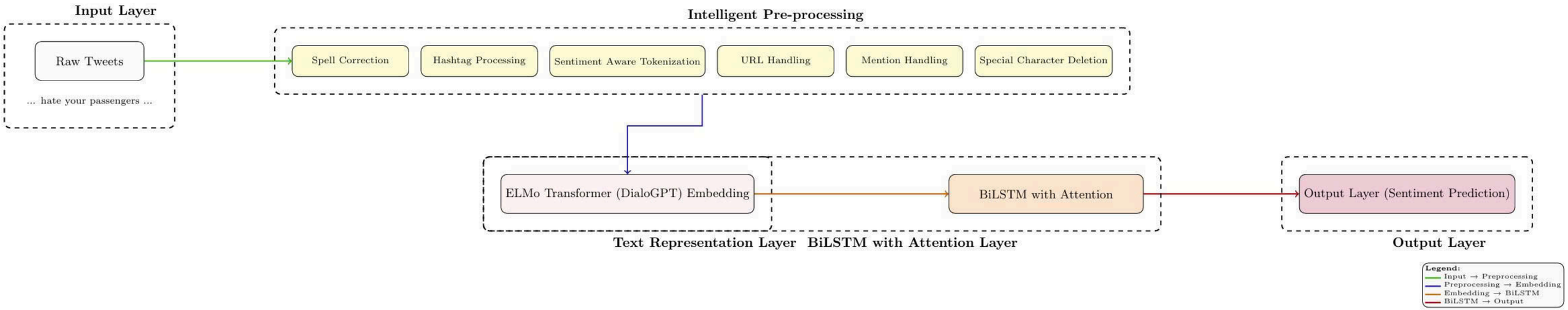
ELMo + BERT


Input Layer
Raw Tweets
... hate your passengers ...
Intelligent Pre-processing
Spell Correction
Hashtag Processing
Sentiment Aware Tokenization
URL Handling
Mention Handling
Special Character Deletion
ELMo
Contextual Embedding
BERT
Transformer Embedding
Text Representation Layer
Concatenation of Embeddings
BiLSTM with Attention
BiLSTM with Attention Layer
Output Layer
Output Layer
Legend:
Input → Preprocessing
Preprocessing → Embeddings
Embeddings → BiLSTM
BiLSTM → Output

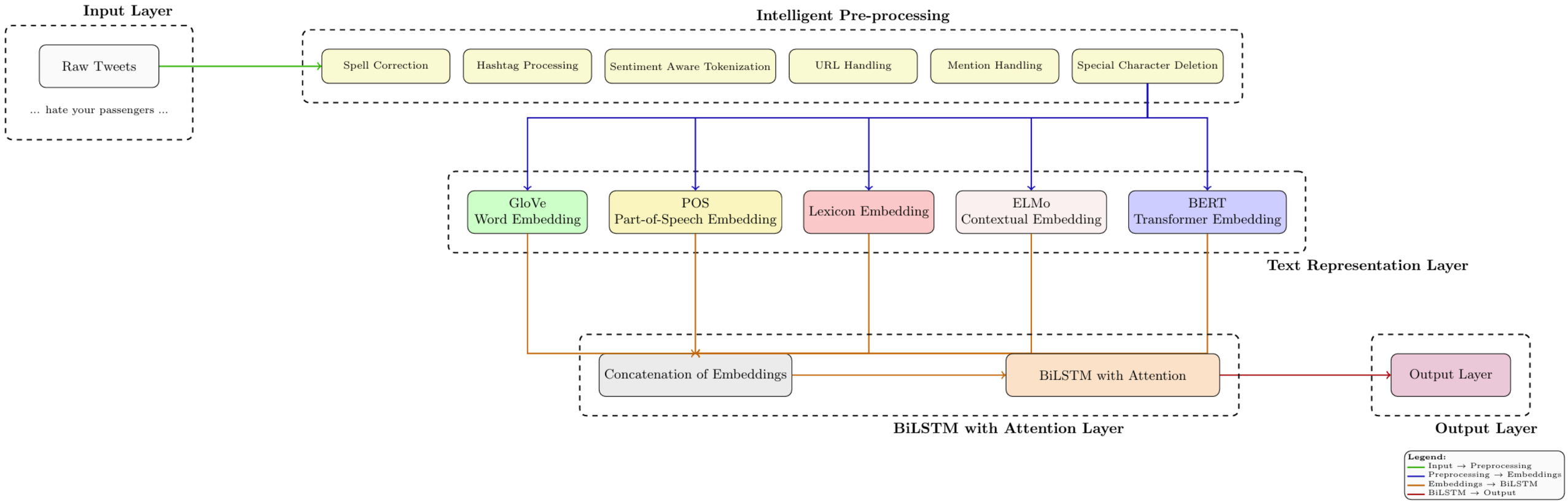
DialoGPT


DialoGPT
(Conversational AI from GPT-2)
Tối ưu cho hội thoại
Phản hồi tự nhiên
Huấn luyện trên Reddit
Dữ liệu lớn, ngữ cảnh đa dạng
Ứng dụng NLP đa dạng
Phân tích cảm xúc, phân loại văn bản

DialoGPT


Input Layer
Raw Tweets
... hate your passengers ...
Intelligent Pre-processing
Spell Correction
Hashtag Processing
Sentiment Aware Tokenization
URL Handling
Mention Handling
Special Character Deletion
ELMo Transformer (DialoGPT) Embedding
BiLSTM with Attention
Output Layer (Sentiment Prediction)
Text Representation Layer
BiLSTM with Attention Layer
Output Layer
Legend:
Input → Preprocessing
Preprocessing → Embedding
Embedding → BiLSTM
BiLSTM → Output

Five-Embedding Layers



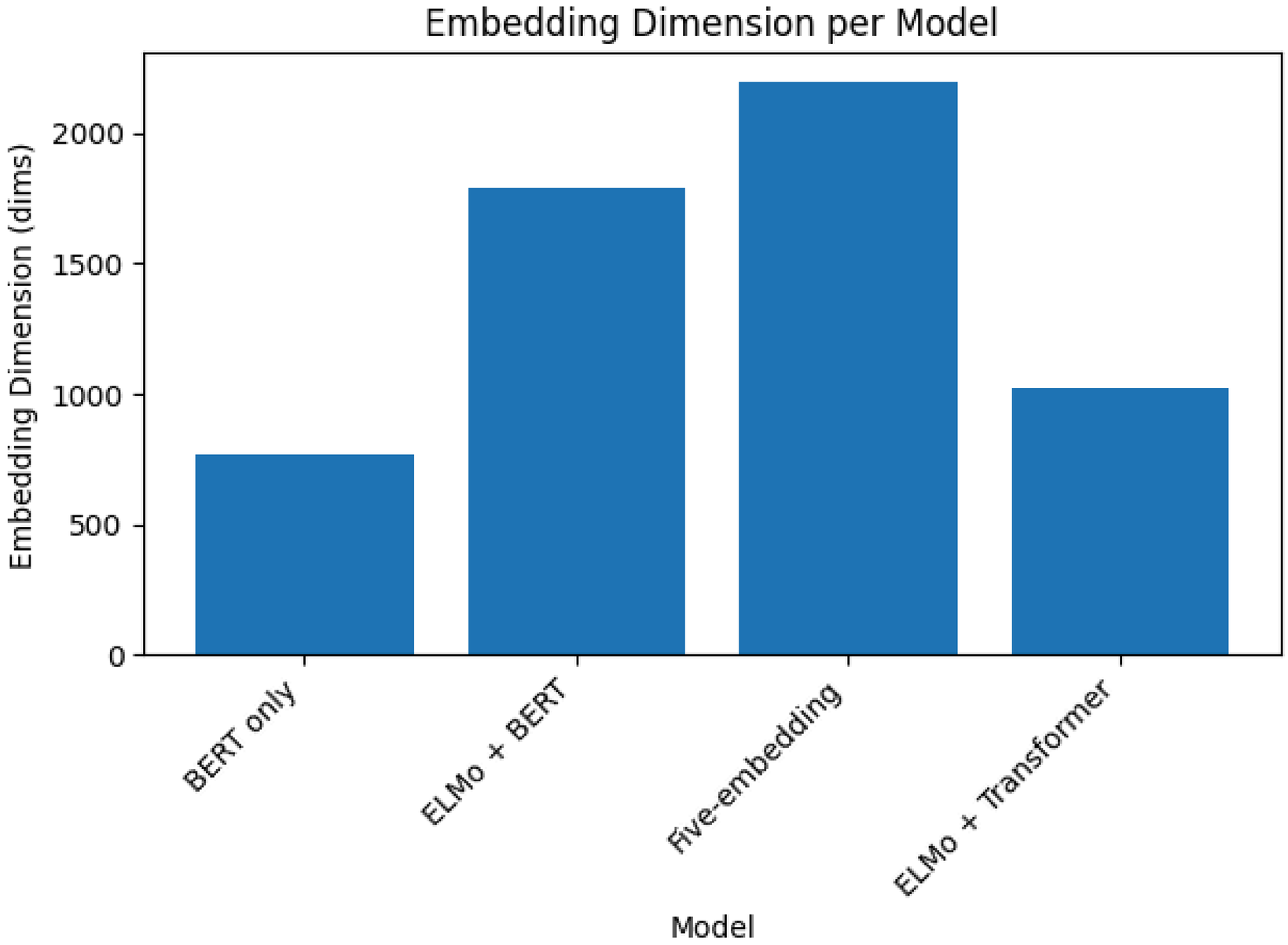
IV. Kết quả thực nghiệm

Configurations – Model Setups

Model	Input size	Batch size
BERT only	max sequence length = 128 tokens; hidden size 768 (BERT-base)	16
ELMo + BERT	max_bert_len = 64 tokens; concatenated embedding = BERT 768 + ELMo 1024 (total 1792 dims)	32
Five-embedding	max_bert_len = 64 tokens; combined embedding = 768 (BERT) + 1024 (ELMo) + 300 (GloVe) + 50 (POS) + 6 (lexicon) + 50 (char) = 2198 dims	32

ELMo + Transformer	input length = 64 tokens for DialoGPT-based ELMo alternative;
	transformer embedding dim = 1024 (passed as elmo_dim);
	Bi-LSTM hidden size = 150
		128

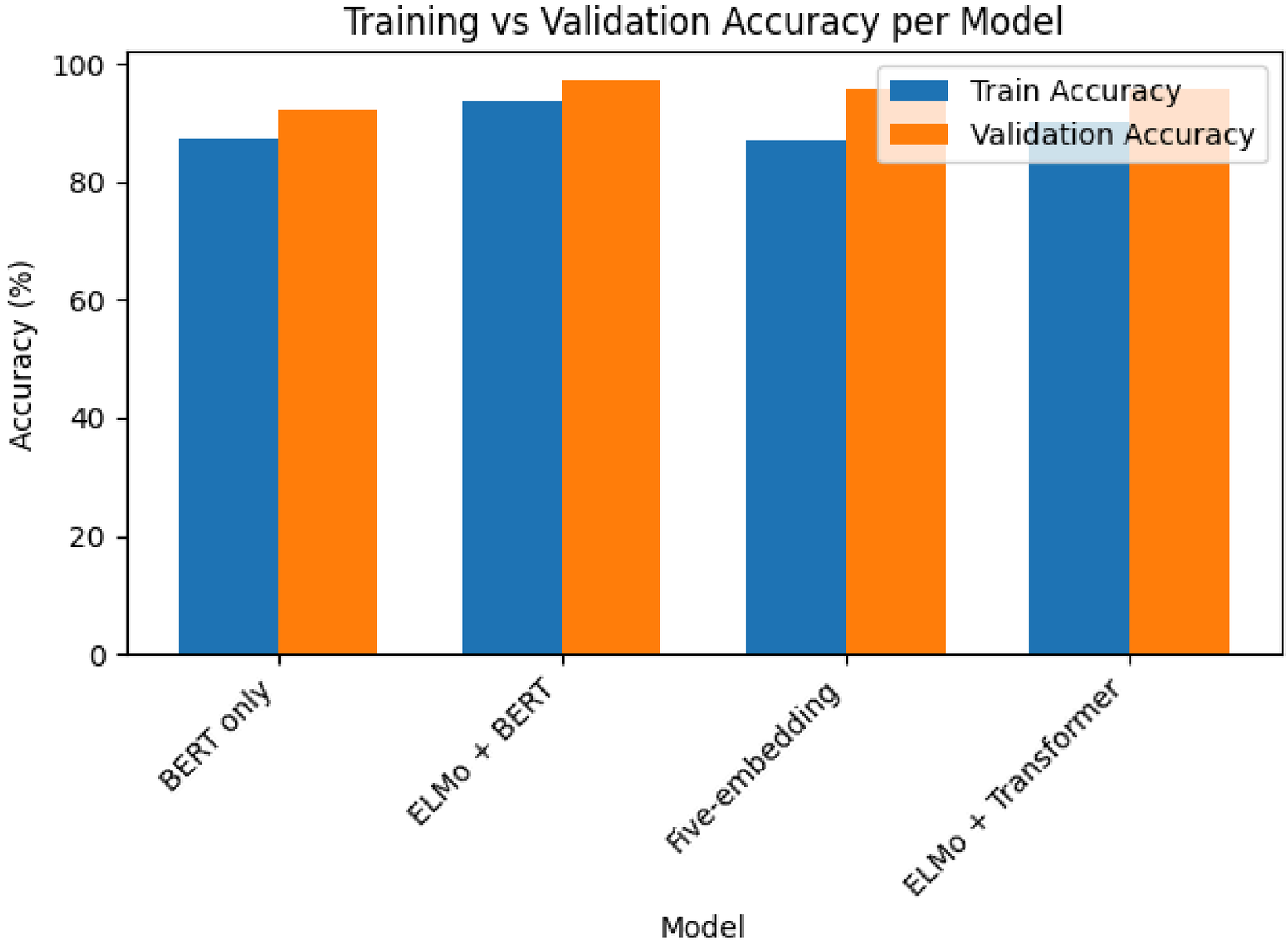
Some Visualizations...



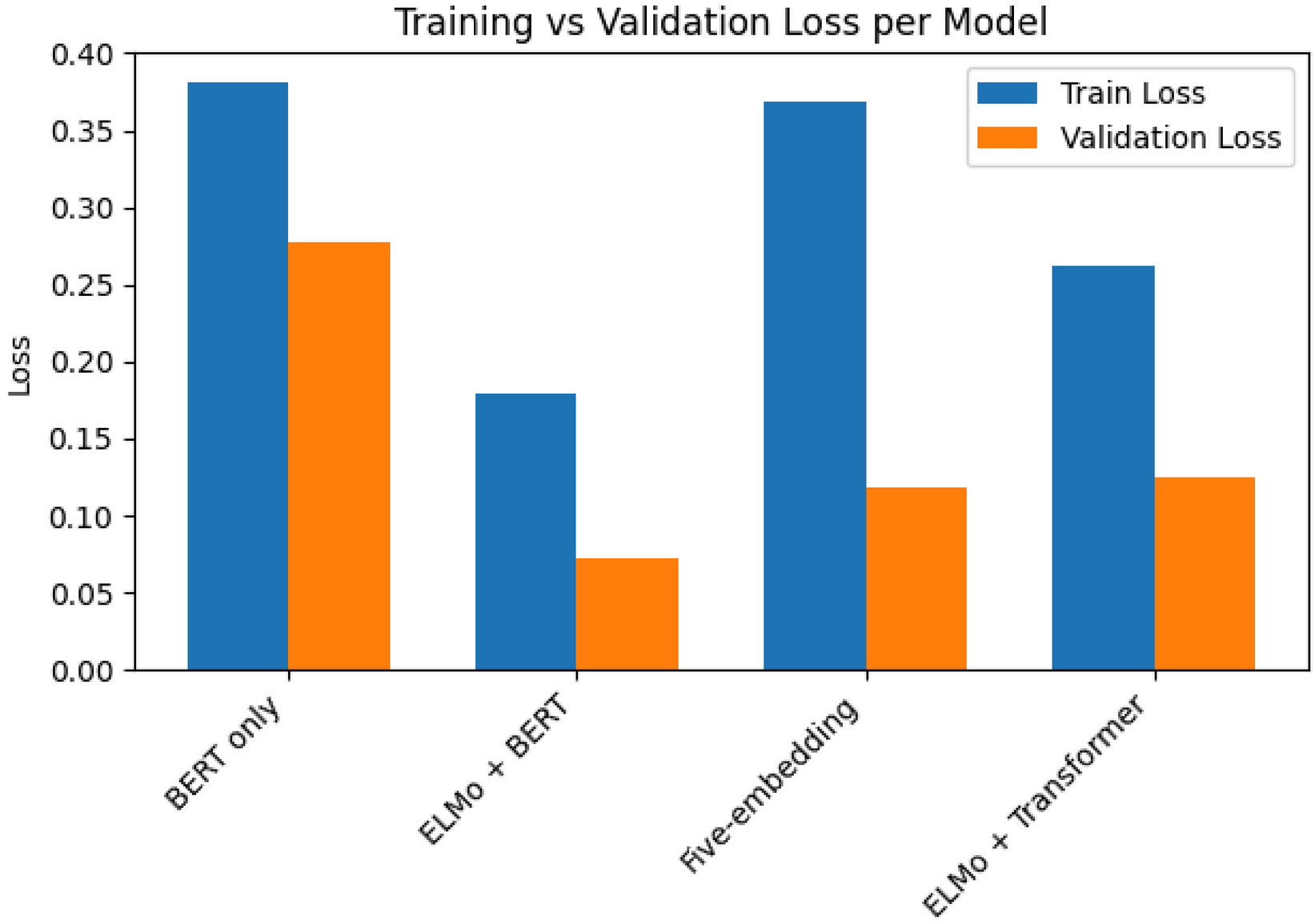
Configurations – Model Setups

Model	Optimizer & Loss	Regularisation techniques
BERT only	Optimizer: AdamW (lr = 2e-5, weight_decay = 0.01); Loss: Weighted cross-entropy	Linear warm-up LR scheduler; gradient clipping (clip_norm = 1.0); early stopping with patience = 3
ELMo + BERT	Optimizer: Adam (lr = 2e-5); Loss: Cross-entropy	Gaussian noise regularisation (σ = 0.3); dropout 0.25 in classifier
Five-embedding	Optimizer: Adam (lr = 2e-5); Loss: Cross-entropy	Gaussian noise (σ = 0.3) and dropout 0.25
ELMo + Transformer	Optimizer: Adam (lr = 0.001); Loss: Cross-entropy	Dropout 0.25 and Gaussian-noise regularisation σ = 0.3

Some Visualizations...



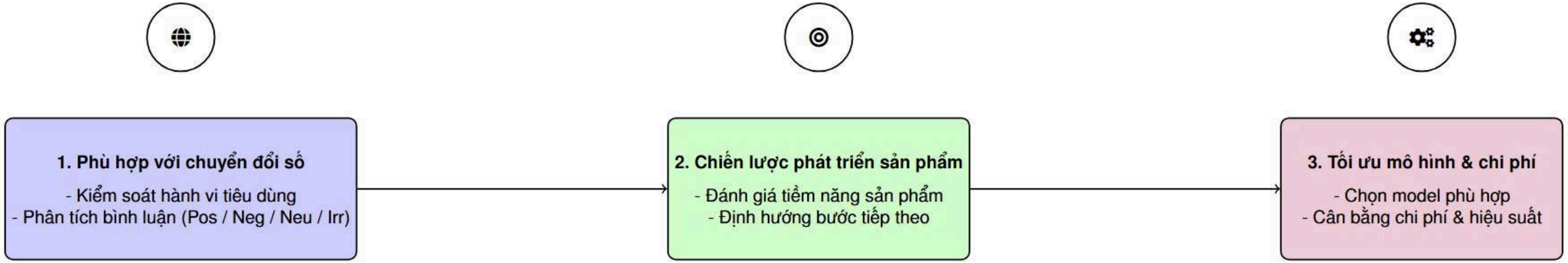
Some Visualizations...



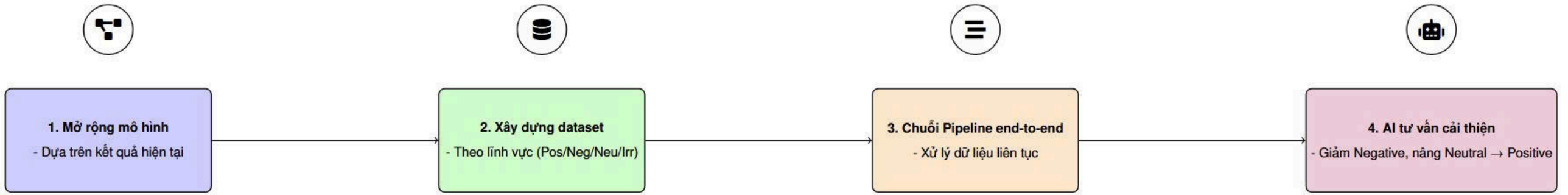
VI. Demo

V. Kết luận

Đánh giá tổng quan



Hướng phát triển



ML IoT
HCMUT EE MACHINE LEARNING & IOT LAB

THANK YOU