NHẬN DIỆN CẢM XÚC TỪ VĂN BẢN BẰNG MÔ HÌNH HỌC SÂU

Hồ Hưng Lộc, Nguyễn Trường Sinh, Hoàng Sỹ Khiêm

INTRODUCTION

Problems definition:
- Input: Text segments (reviews, feedback, etc.)
- Output: Probability distribution of sentiments (Positive, Neutral, Negative)
- Example:
- "No regrets on downloading this... Love it!" → Positive (99%)
- "I was disappointed with this product..." → Negative (Sadness, Anger)

Research Significance:
- Market analysis
- Customer psychology analysis
- Educational state analysis
- Feedback in healthcare and government policies

Challenges:
- Noisy training data (spelling errors, abbreviations, non-standard language)
- Context-dependent word meanings (e.g., same phrase, different sentiments)
- Complex languages like Vietnamese require advanced models and high-quality data

Thesis Goal:
Evaluate and compare the performance of deep learning models (CNN, LSTM, BERT) in text sentiment analysis, focusing on accuracy and robustness.
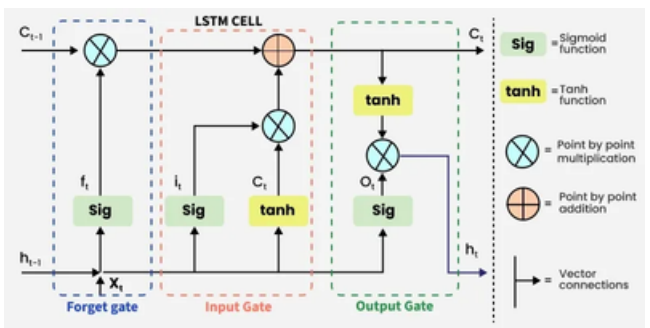
DATASET
- Source: Amazon Reviews, "All Beauty" category (https://amazon-reviews-2023.github.io/)
- Size: 701,500 reviews, 561,136 reviews (training set) and 70,143 samples (test set).
- Purpose: Training and testing deep learning models for sentiment classification (Love, Joy, Neutral, Sadness, Angry)
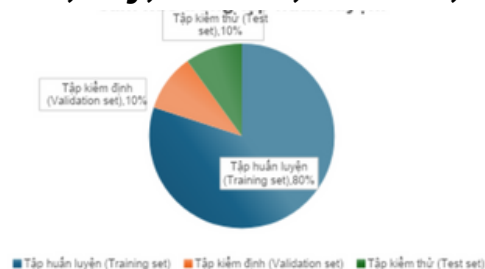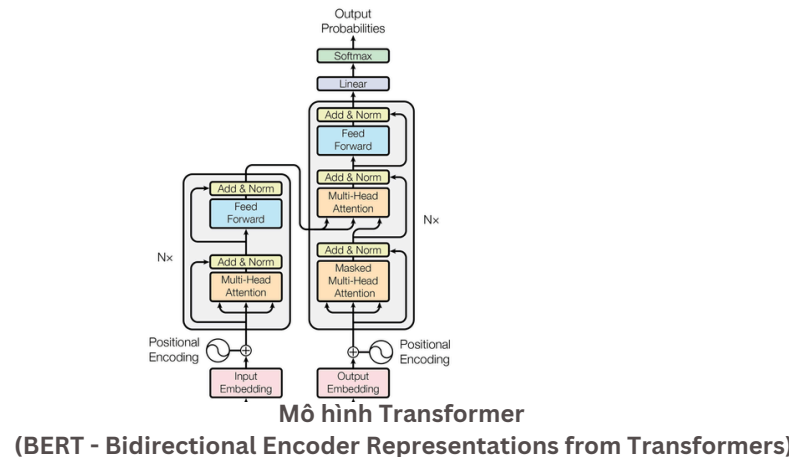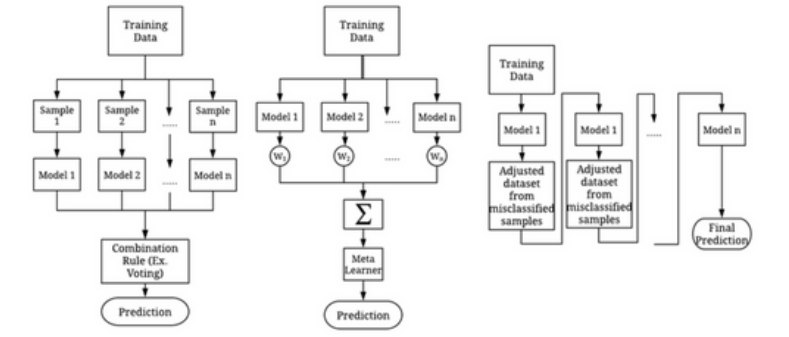

PROPOSED METHOD
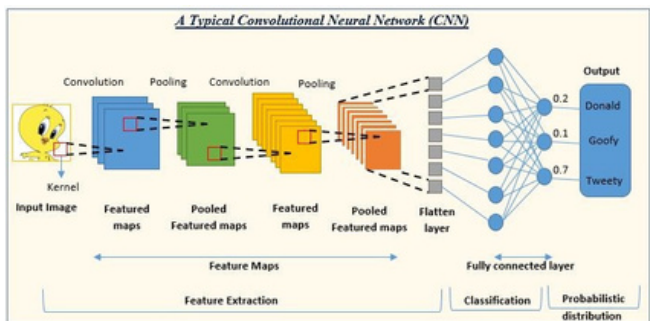
Mô hình học sâu CNN (Convolutional Neural Network)

Mô hình LSTM (Long Short Term Memory)

Mô hình Transformer
(BERT - Bidirectional Encoder Representations from Transformers)

Các phương pháp kết hợp các mô hình

1. Data Preprocessing: Clean text, handle noise (spelling, abbreviations)
2. Feature Extraction: Convert text to numerical representations (e.g., word embeddings)
   - Model Training:
   - Convolutional Neural Network (CNN)
   - Long Short-Term Memory (LSTM)
3. Bidirectional Encoder Representations from Transformers (BERT)
4. Model Evaluation: Compare performance on accuracy, robustness, and training time
5. Optimization: Apply data augmentation to improve model accuracy

EXPERIMENTS (LSTM)


Hình 4.1: Phân phối dữ liệu cảm xúc huấn luyện
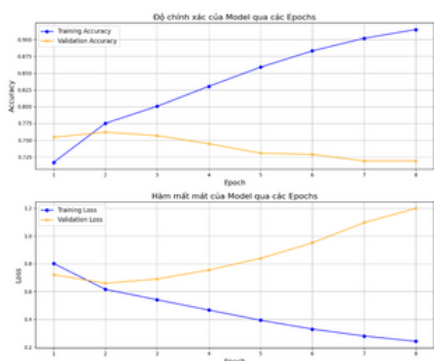
Hình 4.2: Độ chính xác (Accuracy) và Hàm mất mát (Loss) của mô hình trên tập huấn luyện và tập kiểm định qua các Epochs

Bảng 4.2: Báo cáo phân loại chi tiết cho 5 lớp cảm xúc trên tập Test

| Lớp Cảm Xúc	Precision	Recall	F1-score	Support
Anger	0.15	0.94	0.26	10,133
Joy	0.18	0.07	0.10	7,994
Love	0.58	0.01	0.02	42,004
Sadness	0.09	0.01	0.03	4,301
Surprise	0.06	0.00	0.01	5,711
Micro avg	0.16	0.15	0.15	70,143
Macro avg	0.21	0.21	0.08	70,143
Weighted avg	0.40	0.15	0.06	70,143
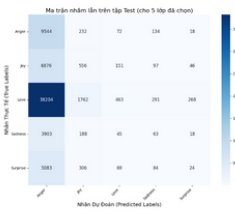
Hình 4.3: Ma trận nhầm lẫn của mô hình trên tập Test (cho 5 lớp cảm xúc)

LSTM model training results

EXPECTED RESULTS

1. **Model Performance:**
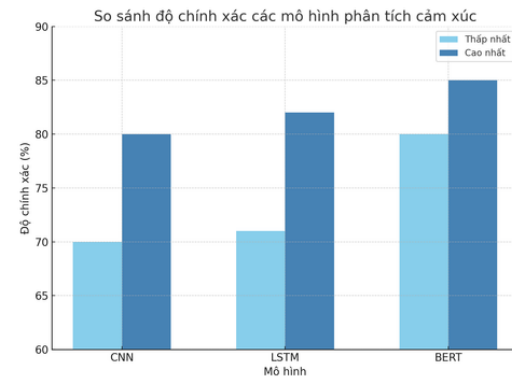
2. **Comparative Analysis:**
- Identify strengths and weaknesses of each model
- Determine the optimal model for sentiment analysis

3. **Tools and Optimization:**
- Use data augmentation for improved accuracy
- Provide open-source code and documentation

4. **Future Work:**
- Propose improvements for complex languages like Vietnamese
- Extend models to real-world applications