Phân tích quan điểm phản hồi sinh viên tiếng Việt trên tập dữ liệu UIT-VSFC

Vũ Nguyên Đan – 23020351
Hà Xuân Huy – 23020375
Hoàng Mạnh Hùng – 23020371

Viện Trí tuệ Nhân tạo – Trường Đại học Công nghệ (UET-VNU)

Môn học: Xử lý ngôn ngữ tự nhiên
Giảng viên hướng dẫn: PGS.TS. Nguyễn Phương Thái

Nội dung trình bày

- Tổng quan dự án & Bài toán
- Phân tích dữ liệu (UIT-VSFC)
- Phương pháp đề xuất (PhoBERT Large)
- Thực nghiệm & Kết quả
- Phân tích & Đánh giá
- Kết luận & Hướng phát triển

Tổng quan & Đặt vấn đề

- Bối cảnh: Việc phân tích phản hồi sinh viên thủ công tốn kém thời gian và dễ sai sót.
- Mục tiêu: Xây dựng hệ thống tự động phân loại cảm xúc phản hồi (Sentiment Analysis) sử dụng Deep Learning.
- Input: Câu phản hồi tiếng Việt (Ví dụ: "Giảng viên dạy nhiệt tình").
- Output: 3 nhãn cảm xúc:
  - Tiêu cực (Negative)
  - Trung tính (Neutral)
  - Tích cực (Positive)
- Thách thức chính:
  - Ngôn ngữ "teencode", viết tắt, icon.
  - Dữ liệu mất cân bằng (Imbalanced Data).
  - Đặc trưng ngôn ngữ tiếng Việt (từ ghép, phủ định).

Bộ dữ liệu UIT-VSFC

- Nguồn: Khảo sát sinh viên đại học (Nguyen et al., 2018).
- Kích thước: $>$ 16,000 câu.
- Phân chia dữ liệu:
  - Train (70%) – Dev (10%) – Test (20%).
- Chiến lược nhóm: Gộp Train + Dev để huấn luyện tối ưu.
- Đặc điểm phân bố:
  - Tích cực & Tiêu cực: Chiếm đa số ($\sim$95%).
  - Trung tính: Rất ít ($\sim$5%) $\rightarrow$ Thách thức lớn nhất.

Phương pháp đề xuất – Mô hình

- Mô hình: PhoBERT Large (Pre-trained Language Model).
- Lý do lựa chọn:
  - State-of-the-art cho tiếng Việt (huấn luyện trên 20GB văn bản).
  - Kiến trúc Transformer với cơ chế Attention hiểu ngữ cảnh tốt hơn n-gram/LSTM.
  - Phiên bản Large (370M tham số) trích xuất đặc trưng sâu hơn bản Base.
- Cơ chế: Input Token $\rightarrow$ BERT Encoder $\rightarrow$ [CLS] Token $\rightarrow$ Classifier (Linear).

Điểm mới – Tiền xử lý ngữ nghĩa (Semantic Preprocessing)

- Vấn đề: Dữ liệu chứa nhiều icon dạng text (colonlove, colonsmile) mang cảm xúc mạnh. Nếu xóa bỏ $\rightarrow$ mất thông tin.
- Giải pháp: Xây dựng từ điển ánh xạ (Mapping Dictionary) sang ngôn ngữ tự nhiên.
- Ví dụ:
  - colonlove $\rightarrow$ "yêu thích" (Tăng trọng số Positive).
  - colonsad $\rightarrow$ "buồn" (Tăng trọng số Negative).
  - wzjwz... $\rightarrow$ "giảng viên".
- Hiệu quả: Giúp PhoBERT "hiểu" được cảm xúc của các ký tự đặc biệt.

```python
def preprocess_line(line):
    # Regex replacements
    line = re.sub(r"wzjwz\d+", "giảng viên", line)
    line = line.replace("doubledot", ":")
    line = line.replace("colonlove", "yêu thích")
    line = line.replace("colonsmile", "vui vẻ")
    line = line.replace("colonsad", "buồn")
    line = line.replace("colonp", ":p")
    line = line.replace("colonb", ":b")
    line = line.replace("colond", ":d")
    line = line.replace("colonright", ")")
    line = line.replace("colonleft", "(")
    return line
```

Chiến lược huấn luyện (Training Strategy)

- Tối ưu hóa phần cứng:
  - Mixed Precision (FP16): Giảm VRAM, tăng tốc độ train.
  - Gradient Accumulation: Batch size ảo $= 16$ (giải quyết giới hạn bộ nhớ GPU).
- Hyperparameters:
  - Learning Rate: 1.5×10^{-5}
  - Epochs: 5
  - Scheduler: Linear Warmup (500 steps).

Kết quả thực nghiệm (Trên tập Test)

Bảng 1: Kết quả chi tiết theo từng lớp (Class-wise Metrics)

Lớp (Class)	**Precision**	**Recall**	**F1-Score**	**Support**
Negative	0.942	0.958	**0.950**	1,409
Neutral	0.681	0.551	0.609	167
Positive	0.951	0.956	**0.954**	1,590
Accuracy			**0.936**	3,166
Weighted Avg	0.933	0.936	**0.934**	3,166
Macro Avg	0.858	0.822	0.838	3,166

- Độ chính xác (Accuracy): 93.6%.
- Weighted F1-Score: 0.934.
- Kết quả từng lớp:
  - Negative: F1 0.950 (Rất cao).
  - Positive: F1 0.954 (Rất cao).
  - Neutral: F1 0.609 (Thấp hơn do thiếu dữ liệu).

So sánh với Baseline (Điểm nhấn)

Bảng 2: So sánh kết quả với baseline từ bài báo gốc

Phương pháp	**Accuracy (%)**	**Average F1 (%)**
Naive Bayes (baseline)	86.1	-
MaxEnt (baseline)	87.9	87.94
PhoBERT Large (đề xuất)	**93.6**	**93.4**

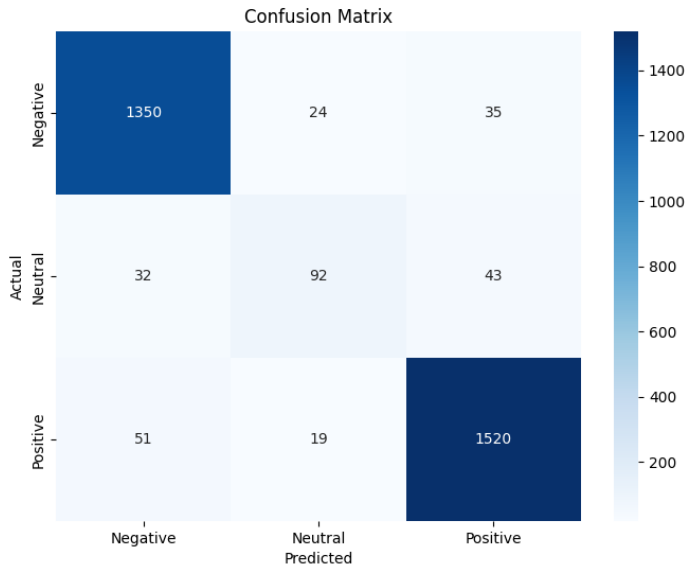
Bảng 3: So sánh chi tiết F1-Score theo từng lớp

Lớp	**MaxEnt (baseline)**	**PhoBERT Large**	**Cải thiện**
Positive	91.32%	95.4%	+4.08%
Negative	90.52%	95.0%	+4.48%
Neutral	33.99%	60.9%	+26.91%
Average	**87.94%**	**93.4%**	**+5.46%**

$\rightarrow$ Kết luận: Deep Learning vượt trội hoàn toàn, đặc biệt khả năng nhận diện lớp khó (Neutral) tăng gấp đôi hiệu năng.

Phân tích lỗi (Confusion Matrix)

- Ưu điểm: Phân biệt cực tốt giữa Tích cực và Tiêu cực (nhầm lẫn rất ít).
- Hạn chế: Lớp Neutral thường bị nhầm sang Positive/Negative.
- Nguyên nhân:
  - Mất cân bằng dữ liệu: Neutral chỉ chiếm 5.2% tập Test.
  - Tính nhập nhằng: "Thầy nhiệt tình nhưng..." $\rightarrow$ bắt từ "nhiệt tình" $\rightarrow$ Positive.



Tổng kết các điểm sáng tạo

- Tiền xử lý thông minh: Chuyển đổi emoji/teencode thay vì loại bỏ.
- Chiến lược dữ liệu: Gộp Train + Dev để tối đa hóa lượng mẫu huấn luyện.
- Phát hiện sâu: Sai lớp Neutral chủ yếu do bản chất dữ liệu, không chỉ do mô hình.

Kết luận & Hướng phát triển

Kết luận:

- Đã xây dựng thành công mô hình đạt độ chính xác 93.6%.
- Hoàn toàn đáp ứng tốt nhu cầu thực tế (lọc phản hồi Khen/Chê).

Hướng phát triển:

- Data Augmentation cho lớp Neutral.
- Weighted Loss / Focal Loss.
- Ensemble: Kết hợp nhiều mô hình để tăng độ ổn định.