SENTIMENT ANALYSIS OF VIETNAMESE STUDENT FEEDBACK: EXPERIMENTS ON UIT-VSFC

Nhóm 15
Nguyễn Đức Minh
Đồng Mạnh Hùng
Lường Minh Trí

MỤC LỤC

01

Giới thiệu

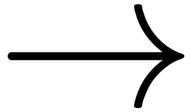
1 | Giới thiệu

Bối cảnh & Bài toán

Thực trạng: Các trường Đại học nhận hàng ngàn phản hồi từ sinh viên mỗi kỳ.
Thách thức: Việc đọc và phân loại thủ công tốn nhiều thời gian, nhân lực và mang tính chủ quan.
Nhu cầu: Cần một hệ thống tự động hóa để "lắng nghe" và hiểu thái độ sinh viên nhanh chóng.

Dữ liệu

Bộ dữ liệu: UIT-VSFC (Vietnamese Students' Feedback Corpus).
Quy mô: Hơn 16.000 câu phản hồi thực tế.
Đặc trưng:
Dữ liệu đời thực, phi cấu trúc.
Chứa nhiều nhiễu (Noise): Teencode ("ko", "dc"), từ viết tắt, icon cảm xúc, từ lóng tiếng Anh chuyên ngành.
Dữ liệu mất cân bằng (Imbalanced).



Mục tiêu đề tài

Phân loại cảm xúc (Sentiment Classification) thành 3 nhãn: Tích cực - Tiêu cực - Trung tính.

02

Tiền xử lý dữ liệu

2 | Tiền xử lý dữ liệu

Sự mất cân bằng dữ liệu nghiêm trọng.
Làm sạch thô (Raw Cleaning):

- Loại bỏ mã ẩn danh rác (wzjwz...), HTML tags.
- Chuẩn hóa khoảng trắng thừa.

Chuẩn hóa Teencode & Từ viết tắt (Normalization):

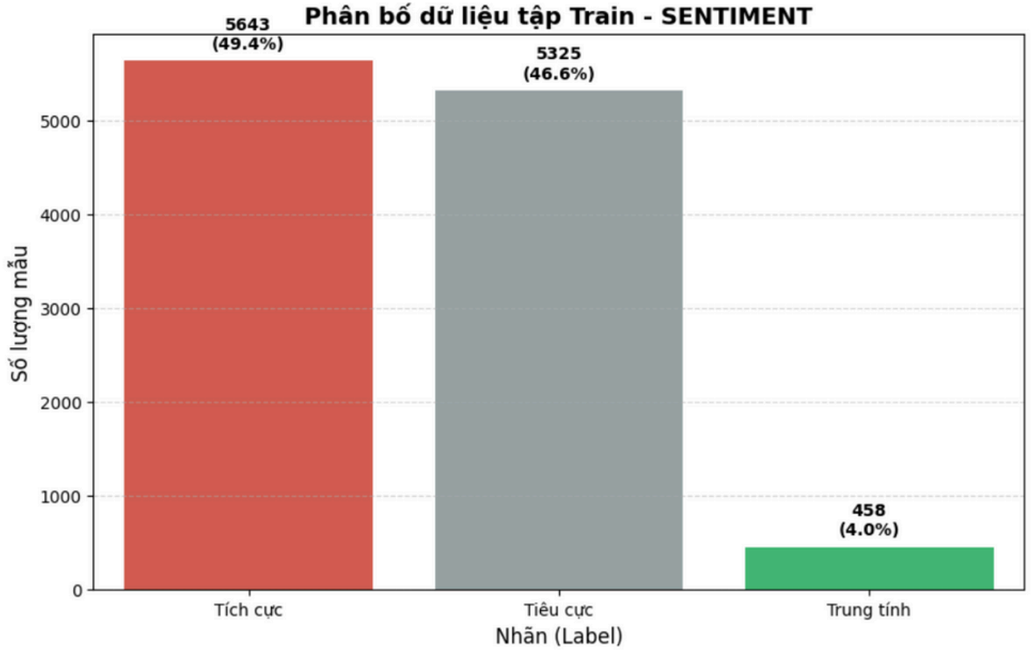
- Xây dựng từ điển ánh xạ (Mapping Dictionary) cho các từ phổ biến: ko → không, dc → được, wa → quá...

Xử lý Ký tự đặc biệt & Icon (Special Handling):

- Giữ nguyên các thuật ngữ chuyên ngành IT: C++, C#, .NET, deadline, bug.
- Giữ nguyên Icon/Artifacts thành từ ngữ mô tả cảm xúc để giữ ngữ cảnh: :)), :(, <3, ...

Tách từ (Word Segmentation):

- Sử dụng thư viện Underthesea để ghép các từ đơn thành từ ghép có nghĩa (sinh viên → sinh_viên).



03

Học máy (SVM + XGBoost)

3.1 Lựa chọn Model & Data Augmentation

Lựa chọn Models

SVM: Hiệu quả với dữ liệu chiều cao (TF-IDF)

XGBoost: Mạnh với dữ liệu không cân bằng

→ Ensemble: Kết hợp ưu điểm của cả hai

Vấn đề

Neutral chỉ 4% (458/11,426) → F1 thấp

↓

Giải pháp: Data Augmentation

Random swap & deletion → Tăng Neutral 4% → 8.5%

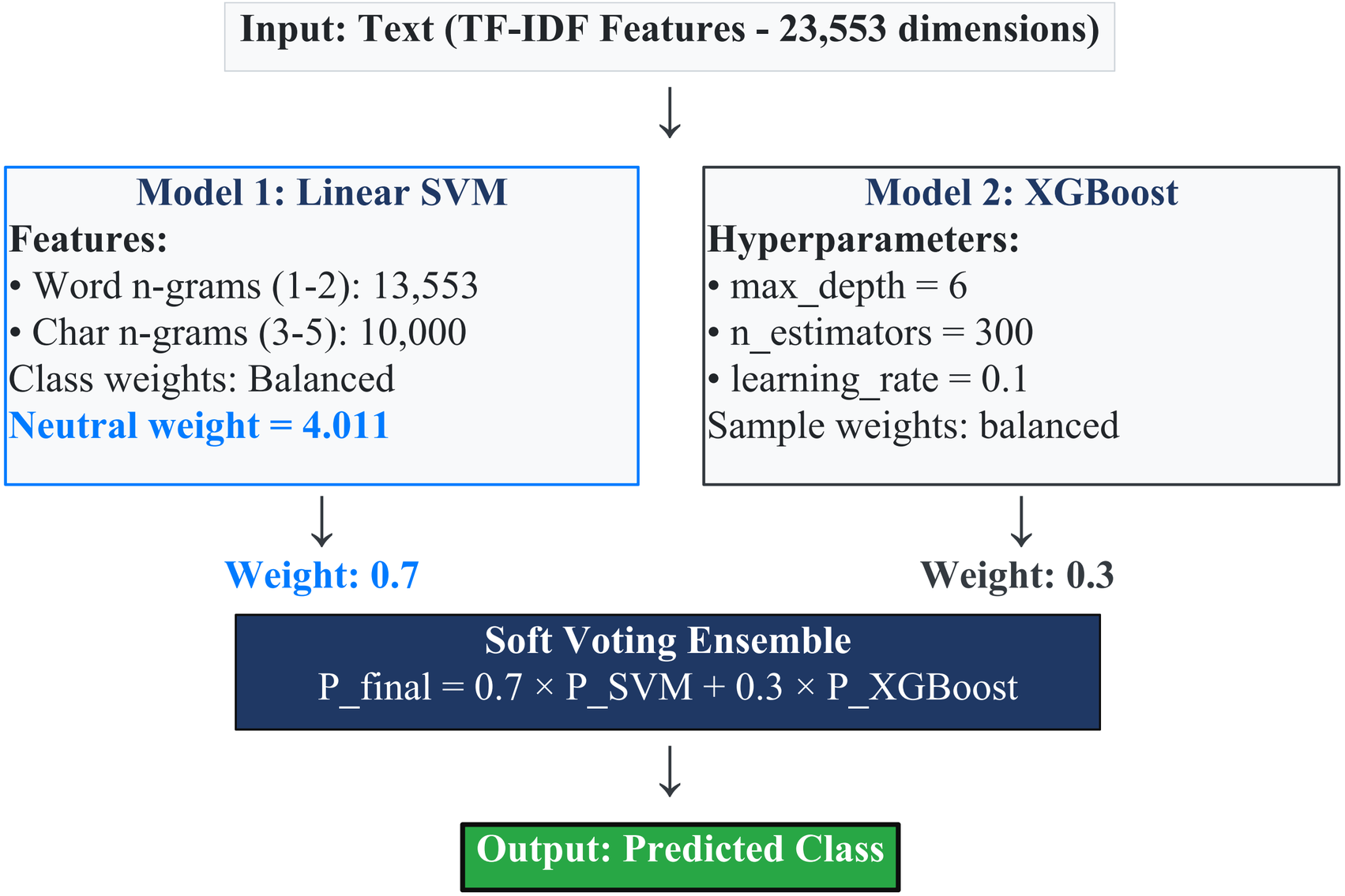
Augmentation Results

	Before	After	Change
Neutral	458 (4.0%)	994 (8.3%)	+116%
Total	11,426	11,962	+536

Distribution sau augmentation:

- Negative: 5,325 (44.5%)
- Neutral: 994 (8.3%)
- Positive: 5,643 (47.2%)

3.2 Kiến trúc Ensemble



3.3 Kết quả

ENSEMBLE RESULTS (SVM=0.7, XGB=0.3)

Class	Precision	Recall	F1-Score
Negative	88.65%	94.82%	91.63%
Neutral	51.30%	35.33%	41.84%
Positive	93.78%	91.07%	92.41%
Overall	89.26%	89.80%	89.39%
Baseline	87.71%	88.66%	87.94%
Improvement	+1.55%	+1.14%	+1.45%

04

Học sâu (PhoBERT)

4.1 Giới thiệu chung

- Nguồn gốc: Phát triển bởi VinAI Research.
- Kiến trúc: Dựa trên kiến trúc RoBERTa (một phiên bản tối ưu hóa mạnh mẽ hơn của BERT).
- Dữ liệu huấn luyện: Được học trên 20GB dữ liệu văn bản tiếng Việt (gồm báo chí, Wikipedia, văn bản pháp luật...).
- Cơ chế: Sử dụng cơ chế Self-Attention (Tự chú ý) của Transformer để hiểu ngữ cảnh hai chiều (Bidirectional).

4.2 Tại sao chọn PhoBERT thay vì BERT/mBERT?

Tối ưu cho Tiếng Việt (Vietnamese-Specific):

- mBERT (Google): Tách từ theo âm tiết đơn lẻ (Syllable-based). Ví dụ: "sinh viên" $\rightarrow$ ["sinh", "viên"].
  $\rightarrow$ Mất ngữ nghĩa từ ghép.
- PhoBERT: Tách từ theo cụm từ ghép (Word-level). Ví dụ: "sinh viên" $\rightarrow$ ["sinh_viên"].
  $\rightarrow$ Giữ trọn vẹn ngữ nghĩa.

Hiệu năng vượt trội:

- PhoBERT đạt kết quả SOTA trên các tác vụ NLP tiếng Việt (POS tagging, NER, Sentiment Analysis) tại thời điểm công bố.

05

Kết luận

5.1 Kết quả

Bảng 1: So sánh kết quả thực nghiệm giữa các mô hình trên tập dữ liệu UIT-VSFC

Mô hình (Model)	**Precision-Weighted**	**Recall-Weighted**	**F1-Weighted**	**F1-Neutral**
Baseline	87.71%	88.66%	87.94%	33.99%
Ensemble ML (SVM + XGB)	89.65%	90.05%	89.73%	46.53%
XLM-RoBERTa (Multilingual)	93.33%	93.68%	93.42%	58.82%
PhoBERT (Xử lý ít/Raw)	93.01%	92.99%	93.00%	59.52%
PhoBERT (Cleaned - Xử lý kỹ)	**93.77%**	**93.87%**	**93.81%**	**63.98%**

5.2 Thách thức về dữ liệu và phân tích lỗi

Sự đa nghĩa và Ẩn dụ

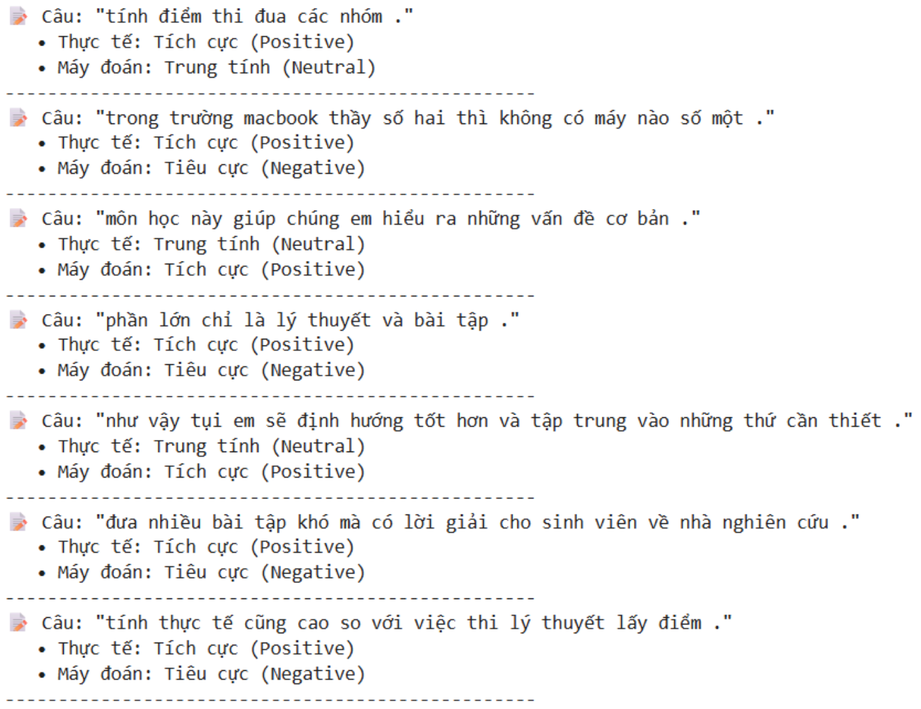
- Model bị bối rối bởi các cấu trúc so sánh kép hoặc nói giảm nói tránh.
- Ví dụ: Câu "Macbook thầy số hai thì không có máy nào số một".
  - Máy thấy: Từ "không", "số hai" . Đoán: Tiêu cực/Trung tính.
  - Thực tế: Khen máy thầy (số một). Nhãn: Tích cực.

2. Nhiễu nhãn

- Dữ liệu gốc (Ground Truth) được gán nhãn chủ quan bởi con người, đôi khi không nhất quán.
- Ví dụ: "Phần lớn chỉ là lý thuyết và bài tập".
  - Người gán nhãn cho là Tích cực.
  - Tuy nhiên, cụm từ "chỉ là" thường mang hàm ý chê (ít thực hành) .Model đoán Tiêu cực là có cơ sở.

3. Ngữ cảnh đặc thù (Context Dependency)

- Các từ vựng mang sắc thái tiêu cực ("khó", "nhiều bài tập") nhưng trong bối cảnh học thuật lại được sinh viên giỏi coi là tốt.
- Ví dụ: "Bài tập khó".
  - Model bắt key "khó": Tiêu cực.
  - Sinh viên thích thử thách :Tích cực.




Thank
You