TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

Đồ án khoa học dữ liệu và ứng dụng 2

Phân tích cảm xúc của các bình luận tiếng Việt theo khía cạnh

Giáo viên hướng dẫn:
ThS. Nguyễn Văn Kiệt
CN. Lưu Thanh Sơn

Sinh viên thực hiện:
Trần Nhật Tiến - 18521493
Kiên Tiến Đạt – 18520568
Nguyễn Tấn Phong-18521239
Hồ Đình Long – 18521022
Đỗ Hùng Dũng - 18520629

NỘI DUNG

1. Giới thiệu bài toán
2. Quy trình tạo bộ dữ liệu
3. Bộ dữ liệu
4. Tiền xử lý dữ liệu
5. Các thuật toán áp dụng và kết quả
6. Kết luận và hướng phát triển

Giới thiệu bài toán

Mục đích bài toán: xác định khía cạnh mà khách hàng đang nói đến và cảm xúc của khách hàng

Input: Giao diện rối mắt, khó chịu.

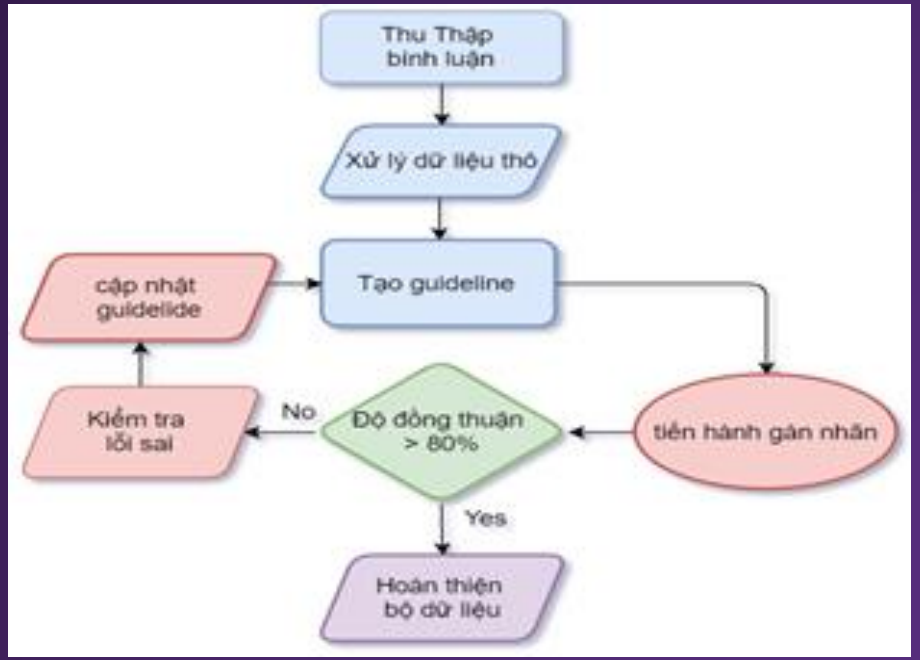
Output: {giao diện, tiêu cực}

Quá trình tạo dữ liệu

- Nguồn thu thập : Google Play
- Công cụ thu thập : google-play-scraper
- Số dòng dữ liệu : 4969
- Số người gán nhãn: 5

Quá trình tạo dữ liệu

Quy trình gán nhãn:



Phân tích bộ dữ liệu

Số nhãn khía cạnh: 8

Giao diện, chỉnh sửa, tính năng, sao lưu, cập nhật, độ ổn định, chung, khác.

Số nhãn cảm xúc: 3

Tích cực, tiêu cực, trung tính.

Phân tích bộ dữ liệu

	Chung	Giao diện	Chỉnh sửa	Tính năng	Sao lưu	Cập nhật	Độ ổn định	Khác
Tích cực	1918	88	91	362	1478	27	49	26
Tiêu cực	155	131	74	495	662	404	872	75
Trung tính	84	69	71	377	144	25	19	15
Tổng	2157	288	236	1234	2284	456	940	116

Bảng 1. Thống kê bộ dữ liệu

Bộ dữ liệu

	comment	labels
0	rất tuyệt vời khi bạn mất điện thoại vẫn không...	[backup_positive]
1	rất thích khả năng lưu trữ nhiều ảnh và chức n...	[feature_positive, backup_positive]
2	tôi không hiểu tại sao bao nhiêu năm tôi dùng ...	[stability_negative]
3	sao lưu nhanh miễn phí tạo ảnh hoạt cảnh pa...	[edit_positive, backup_positive]
4	gogle đệ tam tuyệt đẹp hoàn hảo rất có í...	[general_positive]
5	tải về rồi mở không lên lưu hình giờ không xe...	[feature_negative, stability_negative]
6	quyết định chuyển từ flickr sang thấy cái này...	[general_positive]
7	tôi yêu phần mềm này nếu mất gmail tôi sẽ khó...	[general_positive]
8	rất thích và trên cả tuyệt vời cám ơn gogle	[general_positive]
9	cặp nhan xg không tu dong sao lu ma no tu dong...	[feature_negative, backup_negative, update_neg...

Tiền xử lý dữ liệu

- Đưa các chữ viết hoa về dạng chữ thường: A → a
- Loại bỏ các dấu câu và các ký tự đặc biệt
- Xoá các kí từ kéo dài: ghêêêê → ghê
- Sửa lỗi chính tả: ko → không

  ứg dụg → ứng dụng

  chằm Zn → trầm cảm

Tiền xử lý dữ liệu

- Chuẩn hoá Unicode
- Xoá các ký tự thừa
- Chuyển các biểu tượng cảm xúc thành “tích cực” hoặc “tiêu cực”

 → tích cực

 → tiêu cực

- Tách từ bằng thư viện VnCoreNLP

Các thuật toán áp dụng

Trích xuất đặc trưng:

- TF-IDF

Các mô hình:

- SVM
- PhoBERT
- Bert4news
- Multilingual BERT (uncased)

Đánh giá mô hình dựa trên Precision, Recall, F1-score.

Kết quả thực nghiệm

	Precision	Recall	F1-score
TF-IDF + SVM	0.41	0.33	0.36
PhoBERT	0.77	0.43	0.55
Bert4news	0.76	0.50	0.60
Multilingual BERT	0.80	0.51	0.62

Bảng 2. Kết quả thực nghiệm

Kết quả theo từng nhãn

Mô hình PhoBERT

	precision	recall	f1-score	support
backup_negative	0.82	0.12	0.22	72
backup_neutral	0.00	0.00	0.00	11
backup_positive	0.86	0.85	0.85	152
display_negative	0.00	0.00	0.00	7
display_neutral	0.00	0.00	0.00	4
display_positive	0.00	0.00	0.00	7
edit_negative	0.00	0.00	0.00	6
edit_neutral	0.00	0.00	0.00	5
edit_positive	0.00	0.00	0.00	13
feature_negative	0.00	0.00	0.00	43
feature_neutral	0.00	0.00	0.00	46
feature_positive	0.00	0.00	0.00	35
general_negative	0.00	0.00	0.00	17
general_neutral	0.00	0.00	0.00	6
general_positive	0.76	0.79	0.77	187
other_negative	0.00	0.00	0.00	6
other_neutral	0.00	0.00	0.00	2
other_positive	0.00	0.00	0.00	4
stability_negative	0.64	0.47	0.54	90
stability_neutral	0.00	0.00	0.00	1
stability_positive	0.00	0.00	0.00	3
update_negative	0.00	0.00	0.00	40
update_neutral	0.00	0.00	0.00	2
update_positive	0.00	0.00	0.00	3

Kết quả theo từng nhãn

Mô hình Bert4news

	precision	recall	f1-score	support
backup_negative	0.74	0.49	0.59	72
backup_neutral	0.00	0.00	0.00	11
backup_positive	0.85	0.86	0.85	152
display_negative	0.00	0.00	0.00	7
display_neutral	0.00	0.00	0.00	4
display_positive	0.00	0.00	0.00	7
edit_negative	0.00	0.00	0.00	6
edit_neutral	0.00	0.00	0.00	5
edit_positive	0.00	0.00	0.00	13
feature_negative	0.00	0.00	0.00	43
feature_neutral	0.50	0.09	0.15	46
feature_positive	0.00	0.00	0.00	35
general_negative	0.00	0.00	0.00	17
general_neutral	0.00	0.00	0.00	6
general_positive	0.76	0.78	0.77	187
other_negative	0.00	0.00	0.00	6
other_neutral	0.00	0.00	0.00	2
other_positive	0.00	0.00	0.00	4
stability_negative	0.57	0.46	0.51	90
stability_neutral	0.00	0.00	0.00	1
stability_positive	0.00	0.00	0.00	3
update_negative	1.00	0.60	0.75	40
update_neutral	0.00	0.00	0.00	2
update_positive	0.00	0.00	0.00	3

Kết quả theo từng nhãn

Mô hình
Multilingual BERT

backup_negative	0.71	0.42	0.53	72
backup_neutral	0.00	0.00	0.00	11
backup_positive	0.88	0.89	0.89	152
display_negative	0.00	0.00	0.00	7
display_neutral	0.00	0.00	0.00	4
display_positive	0.00	0.00	0.00	7
edit_negative	0.00	0.00	0.00	6
edit_neutral	0.00	0.00	0.00	5
edit_positive	0.00	0.00	0.00	13
feature_negative	0.00	0.00	0.00	43
feature_neutral	0.69	0.20	0.31	46
feature_positive	0.00	0.00	0.00	35
general_negative	0.00	0.00	0.00	17
general_neutral	0.00	0.00	0.00	6
general_positive	0.81	0.75	0.78	187
other_negative	0.00	0.00	0.00	6
other_neutral	0.00	0.00	0.00	2
other_positive	0.00	0.00	0.00	4
stability_negative	0.62	0.50	0.56	90
stability_neutral	0.00	0.00	0.00	1
stability_positive	0.00	0.00	0.00	3
update_negative	0.95	1.00	0.98	40
update_neutral	0.00	0.00	0.00	2
update_positive	0.00	0.00	0.00	3

Kết luận

- Độ chính xác của các mô hình còn hạn chế, trong đó mô hình Bert-base-multilingual có kết quả cao nhất (0.62 f1-score).

- Dữ liệu chênh lệch lớn giữa các nhãn khiến cho các mô hình hoạt động không tốt. Dự đoán trên các nhãn chiếm tỉ lệ thấp Không chính xác.

Hướng phát triển

- Nghiên cứu và cải tiến các phương pháp tiền xử lý.
- Tăng kích thước và cải thiện chất lượng của bộ dữ liệu.
- Thử nghiêm thêm các mô hình SOTA.

CẢM ƠN THẦY VÀ CÁC BẠN ĐÃ LẮNG NGHE !