PHÁT HIỆN BÌNH LUẬN CHỐNG PHÁ, XÚC PHẠM NHÀ NƯỚC TRÊN MẠNG XÃ HỘI

Trần Trung Anh - 18520473 - CS114.K21.KHTN

Link Github:
https://github.com/AVL1/CS114.K21.KHTN/blob/master/ML_Capstone.ipynb

Tóm tắt

- Phát hiện bình luận chống phá, xúc phạm Nhà nước trên mạng xã hội
- Tự xây dựng và xử lý bộ dữ liệu, sử dụng các phương pháp vectơ hoá văn bản để trích xuất đặc trưng, sử dụng thuật toán máy học phân loại và Grid Search để tối ưu tham số.
- Mô hình đạt kết quả cao nhất: 73,74%
- Link demo: https://drive.google.com/file/d/1iQ-8RmikVEiEMstt_XgtTRIiZHSgauxY/view?usp=sharing
- Ảnh của thành viên:

Trần Trung Anh - 18520473

Bài toán

Input: một câu bình luận

Output: 1 (Positive), nếu bình luận trên có tính chất xúc phạm, chống đối.

Ví dụ: Cộng sản Việt không hèn nhát mới là chuyện lạ

0 (Neutral), nếu bình luận trung tính, không liên quan.

Ví dụ: Có bằng chứng về tài sản hối lộ của ông ấy dc ko?

1 (Negative), nếu bình luận không mang tính chất xúc phạm, chống đối.

Ví dụ: hoàn toàn tin tưởng vào sự lãnh đạo của đảng CS VN

Bộ dữ liệu

Công cụ thu thập dữ liệu:

Python + Selenium

Dữ liệu chủ yếu lấy từ 2 trang Facebook:

Việt Tân và *Thông tin chính phủ*

Đọc từng câu bình luận và gán nhãn (số người tham gia: 3).

Bộ dữ liệu

Số lượng từng nhãn:

0: 3086

1: 1039

- 1: 451

Tổng số: 4576

Phân chia train/test: 8/2

	Text	value
0	bậy nào, Cali giờ đóng cửa lại rồi, sao cạo mó...	0
1	nói thế dân Việt Nam mình ngu à.	1
2	Đương nhiên là cảnh sát chìm rồi. Dân nào lại ...	1
3	Cảnh sát chìm ngày xưa chỉ đi lùng tội phạm gi...	1
4	thái lan mà nghèo	0
5	ai giết Bắc Hà vậy ku ? Có hợp hiến pháp chxhc...	1
6	tao là người việt nam đây,,và tao sẽ đứng lên ...	1
7	chính xác rồi, dân ko bị ai giật dây hết nên đ...	1
8	nói thiệt ước gì lúc đó mình cũng đc giúp sức ...	1
9	thì ra phía bên kia chiến tuyến người ta ca ng...	1

Tiền xử lý dữ liệu

Gồm các thao tác:

Loại bỏ dấu kết thúc câu, kí tự đặc biệt.

Chuyển những kí tự viết hoa thành viết thường.

Chuyển các từ viết từ viết tắt thành dạng bình thường.

Loại bỏ stopword.

Tách từ.

Tiền xử lý dữ liệu

Dữ liệu sau quá trình tiền xử lý

	Text	value
0	bậy nào cali giờ đóng_cửa sao cạo móng mới mở_...	0
1	nói thế dân việt nam mình ngu à	1
2	đương_nhiên cảnh_sát chìm dân nào đi phản đồng...	1
3	cảnh_sát chìm ngày_xưa đi lùng tội_phạm giựt d...	1
4	thái lan nghèo	0
5	ai giết bắc hà ku hợp_hiến_pháp cộng_hoà xã_hộ...	1
6	tao người việt nam đây tao đứng đầu thằng_cha ...	1
7	chính_xác dân ai giật_dây hết đừng vu thế_lực ...	1
8	nói thiệt ước mình đc giúp_sức đập bọn chó ngu...	1
9	phía bên kia chiến_tuyến người_ta công_an ngợi...	1

Trích xuất đặc trưng

Count Vectorize

- Xây dựng một bộ từ điển gồm các từ khác nhau, tổng hợp từ bộ dữ liệu kèm với số lần từ đó xuất hiện trong toàn bộ dữ liệu.
- Trong mỗi câu, với mỗi từ trong từ điển, ta đếm số lần xuất hiện của từ đó trong câu.
- Vector đặc trưng gồm số lần xuất hiện của mỗi từ.

Trích xuất đặc trưng

TF-IDF Vectorize

Giá trị TF-IDF của một từ là một giá trị thu được bằng cách thống kê mức độ quan trọng của từ này trong một câu và trong toàn bộ dữ liệu.

TF-IDF value = TF * log(IDF), trong đó

- TF (Term frequency): tần số xuất hiện của một từ trong một câu.
- IDF (Inverse Document Frequency): nghịch đảo tần số của một từ trong toàn bộ các câu.

Thuật toán máy học

Multinomial Naive Bayes:

Được xây dựng dựa trên định lý Bayes với giả định các đặc trưng hoàn toàn độc lập với nhau. MNB được chủ yếu dùng cho bài toán phân loại nhiều nhãn.

Support Vector Machine:

Tìm các siêu mặt phẳng trong không gian N chiều sao cho chúng có thể phân biệt nhóm các điểm dữ liệu với nhau. Khoảng cách giữa các điểm dữ liệu của các nhãn khác nhau đến mặt phẳng càng cân bằng thì kết quả dự đoán càng tốt.

Kết quả thực nghiệm

Kết quả Accuracy của các mô hình đã xây dựng.

LSVC: Linear SVC

MNB: Multinomial Naive Bayes

CV: CountVectorize

TV: TF-IDFVectorize

	LSVC	MNB
CV	0.728166	0.713974
TV	0.734716	0.658297

Tối ưu tham số

Sử dụng thuật toán Grid Search để tối ưu tham số cho mô hình có accuracy cao nhất: LinearSVC + TF-IDFVectorize.

	precision	recall	f1-score	support
-1	0.59	0.40	0.48	96
0	0.76	0.85	0.80	595
1	0.58	0.48	0.52	225
accuracy			0.71	916
macro avg	0.64	0.57	0.60	916
weighted avg	0.70	0.71	0.70	916

Đánh giá, kết luận

Phân loại đúng một số câu bình luận phổ biến.

Phân loại dữ liệu nhập mới từ người dùng chưa thực sự chính xác.

Các mô hình cho kết quả khá thấp vì các nguyên nhân:

- Bộ dữ liệu mất cân bằng, tiền xử lý không hoàn toàn chính xác cho từng câu, quan trọng nhất là vấn đề sai chính tả, văn phạm.
- Công cụ tách từ chưa tốt cho dữ liệu Tiếng Việt.
- Số lượng dữ liệu còn hạn chế, các bình luận có thể không đủ ý.
- Có trường hợp một số câu gán nhãn chưa đúng.
- Chưa xử lý được trường hợp overfitting do một từ xuất hiện quá nhiều.

Hướng phát triển

Xây dựng, phát triển bộ dữ liệu:

- Tăng số lượng, cân bằng các nhãn.
- Tiền xử lý các câu sai chính tả, tên riêng.
- Thu thập từ nhiều nguồn khác nhau.
- Đảm bảo đồng bộ, chính xác trong quá trình gán nhãn.

Ứng dụng các phương pháp học sâu (Deep Learning) vào trích xuất đặc trưng.

Cải thiện việc tối ưu tham số.