Phân tích cảm xúc các bình luận xúc phạm trên mạng xã hội

Nguyễn Xuân Vinh
18521655

Hồ Anh Dũng
18520630

Trần Minh Quân
18521288

Abstract

Phân tích cảm xúc các bình luận trên mạng xã hội hay phát hiện xúc phạm trong ngôn ngữ là một bài toán đang được nghiên cứu khá nhiều trong thời gian hiện nay. Đồ án của chúng tôi thực hiện việc phát triển mô hình cho bộ dữ liệu ViHSD với các phương pháp học máy và học sâu. Cùng với đó là việc tiền xử lý để tăng cao kết quả cho mô hình. Với việc kết hợp giữa tiền xử lý và học sâu chúng tôi đã có mô hình PhoBert và Bert4News có kết quả được cải thiện hơn so với mô hình cơ sở khoảng 1% đến 2 %.

1 Giới thiệu

Kể từ khi đại dịch Covid-19 diễn ra, xu hướng làm việc tại nhà trở nên phổ biến, mọi người dành nhiều thời gian hơn vào internet, liên lạc, kết nối, tương tác với nhau qua các nền tảng mạng xã hội. Cùng với rất nhiều lợi ích của mạng xã hội đem lại cho chúng ta, tồn tại một vấn nạn gây ra không ít ảnh hưởng xấu đến những người dùng mạng xã hội, đó là sử dụng ngôn ngữ kích động thù địch, xúc phạm, độc hại một cách bừa bãi hoặc có tổ chức trên các nền tảng mạng xã hội. Vấn nạn này đã tồn tại trước đó từ lâu, nhưng kể từ lúc đại dịch Covid-19 diễn ra, thì nó xảy ra ngày càng nghiệm trọng hơn, gây ra nhiều hiện lụy xấu đến xã hội.

Vì vậy trong các năm gần đây bài toán phân tích bình luận xúc phạm trên mạng xã hội đang được nghiên cứu trên nhiều nước, nhiều ngôn ngữ. Ở Việt Nam chúng ta cũng có các bộ dữ liệu về phân tích bình luận xúc phạm trên mạng xã hội, tiêu biểu là bộ dữ liệu ViSHD của tác giả Lưu Thanh Sơn và cộng sự được công bố năm 2021. Chúng tôi sẽ phát triển mô hình, cải thiện độ chính xác cho bộ dữ liệu này.

Bài toán của chúng tôi sẽ được mô tả như sau:

- Input: Tập các bình luận Tiếng Việt trên mạng xã hội
- Output: Nhãn tương ứng cho từng câu bình luận. Một trong 3 nhãn: CLEAN, OFFENSIVE, HATE.

2 Công trình nghiên cứu liên quan

Các năm gần đây đã có nhiều nghiên cứu trong việc phân tích lời nói, văn bản xúc phạm của người sử dụng mạng xã hội trên nhiều ngôn ngữ khác nhau.

Năm 2017, (Badjatiya et al., 2017) đã công bố bài báo "Deep Learning for Hate Speech Detection in Tweets". Nhóm tác giả thực hiện các thử nghiệm mở rộng với nhiều kiến trúc học sâu để tìm hiểu cách nhúng từ ngữ nghĩa để xử lý sự phức tạp trong việc xác định những tweet gây thù hận. Các thử nghiệm của nhóm tác giả trên tập dữ liệu điểm chuẩn gồm 16K tweet (Waseem and Hovy, 2016) có nhãn, cho thấy rằng các phương pháp học sâu như vậy vượt trội hơn so với các phương pháp char/word n-gram hiện đại.

Năm 2021, trong xử lý ngôn ngữ tự nhiên tiếng Việt, nghiên cứu của (Luu et al., 2021) đã công bố bộ dữ liệu ViHSD hơn 33.000 câu với 3 nhãn (CLEAN, OFENSIVE, HATE) về các bình luận xúc phạm trên mạng xã hội. Trong nghiên cứu, nhóm tác giả sử dụng mô hình học sâu Text-CNN, các mô hình transformer đạt kết quả cao nhất với F1-macro là 62.69% với mô hình BERT(bert-base-multilingual-cased). Cùng trong năm này một nghiên cứu khác của (Nguyen et al., 2021) cũng đã công bố bộ dữ liệu UIT-ViCTSD, kích thước 10.000 câu với 2 nhãn (Non-toxic, Toxic) về các bình luận mang tính xúc phạm trên mạng xã hội, mô hình cơ sở trong bài báo có kết quả cao nhất với F1-macro là 59.40%.

Table 1: Một số ví dụ từ tập dữ liệu ViHSD

STT	Câu bình luận	Nhãn
1	Nhanh thực sự	CLEAN
2	phải thầy k ta	CLEAN
3	Cận thận nhà ae	CLEAN
4	Minh mẫn vcl	OFFENSIVE
5	Đấu khẩu - Chim lợn	OFFENSIVE
6	Bóng dơ	HATE
7	Im mẹ đi thằng mặt lon	HATE

Từ các nghiên cứu phía trên, chúng tôi sẽ kết thừa và phát triển với mục đích cải thiện độ chính xác trong việc phát hiện các bình luận xúc phạm, độc hại trên mạng xã hội.

3 Bộ dữ liệu

3.1 Chuẩn bị bộ dữ liệu

Tập dữ liệu mà chúng tôi sử dụng trong bài này là tập dữ liệu phát hiện lời nói căm thù của ViHSD-Tiếng Việt. Tập dữ liệu này được thu thập từ các nhận xét của người dùng về giải trí, người nổi tiếng, các vấn đề xã hội và chính trị từ các trang Facebook và video Youtube khác nhau có tỉ lệ tương tác cao và không hạn chế bình luận của Việt Nam, sau đó xóa tên các người bình luận để duy trì tính ẩn danh.

3.2 Chú thích

Tập dữ liệu ViHSD chứa ba nhãn: HATE, OFFENSIVE và CLEAN (Non HATE). Mỗi nhãn sẽ được gán cho một dòng bình luận trong tập dữ liệu.Trong ViHSDdataset, có hai nhãn biểu thị cho nhận xét ngôn từ kích động thù địch và một nhãn biểu thị cho nhận xét bình thường. Ý nghĩa chi tiết về các nhãn và ví dụ cho mỗi nhãn được mô tả trong bảng 1 dưới đây.

Trong bộ dữ liệu có nhiều nhận xét mà ở trong đó có nhiều từ bị viết dưới dạng không chính thức, viết tắt hoặc sai chính tả. Ví dụ như từ "tao" thường bị viết tắt hoặc viết sai thành là "t" và "taooo", cụm từ "vãi cả l*n" thường bị viết thành "vcl", "vcd" và "vãi cả đ*i", từ "đ*t" thường bị viết thành "đuyt", "disss" và "disme". Những từ trên thường xuyền được một nhóm người dùng Facebook và Youtube Việt Nam sử dụng trong các bài đăng hoặc các bình luận tiêu cực trên mạng xã hội nhằm gây mâu thuẫn, chia rẽ và xúc phạm một số cá nhân hay tổ chức nào đó.

3.3 Tổng quan về tập dữ liệu

Tập dữ liệu ViHSD chứa 33.400 bình luận. Mỗi câu được gắn nhãn là CLEAN (0), OFFENSIVE (1) và HATE (2). Bảng 2 hiển thị một số ví dụ từ tập dữ liệu ViHSD. Sau đó, chúng tôi chia tập dữ liệu của mình thành các tập huấn luyện (train), phát triển (dev) và test, tương ứng với tỷ lệ: 7-1-2 như hình 1 dưới đây.

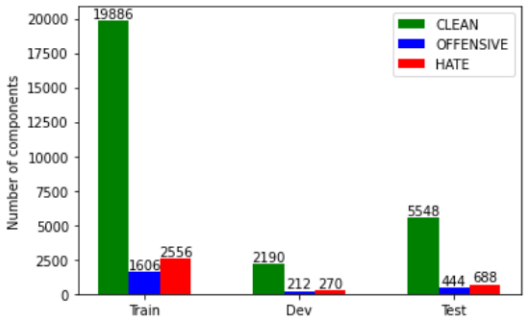

Figure 1: Biểu đồ thể hiện sự phân bô các nhãn trên các tập train, dev, test

4 Phương pháp

4.1 Tiền xử lý

Chúng tôi thực hiện các bước tiền xử lý với bộ dữ liệu: chuẩn hóa unicode tiếng Việt, lọc các link spam, chuyển đổi các danh xưng về người thành NER, lọc các icon và ký tự đặc biệt, xóa URL, đưa về dạng chữ thường, loại bỏ stop word và đặc biệt là chuẩn hóa các từ sai chính tả, teencode về chuẩn như Bảng 2 vì phần lớn các câu bình luận trên mạng đều chứa các từ viết tắt, sai chính tả, không như các văn bản thuần túy.

4.2 Bi-LSTM

Long Short Term Memory là một dạng đặc biệt của mạng nơ-ron hồi qui – Recurrent Neural Network (RNN). Nó có khả năng học được những phụ thuộc xa. LSTM được giới thiệu bởi Hochreiter và Schmidhuber (1997), và sau đó đã được cải tiến và phổ biến bởi rất nhiều người trong ngành. Chúng hoạt động cực kì hiệu quả trên nhiều bài toán khác

Table 2: Chuẩn hóa dữ liệu

Từ gốc(sai chính tả)	Từ chuyển đổi
cf, coffee, cafe, caphe	cà phê
uk, um, uhm, ukm, uh	ừ
del, méo, đíu, đel	đéo
ko, hông, k, hôngg, khôg	không
.....	.....
kbg, kbh	không bao giờ

nhau nên dần đã trở nên phổ biến như hiện nay. LSTM được hoạt động dựa trên các cell state và các hàm sigmoid, tanh và các cổng: forget, input, output.

Bi-LSTM là một biến thể của LSTM. Bi-LSTM thêm vào một layer đặc biệt gọi là backward recurrent layer kết nối layer đó cùng với forward layer để cho ra output. Bằng cách này, một unit của Bi-LSTM có thể học được các thông tin ở cả quá khứ (trước nó) và tương lai (unit kế tiếp). Kiến trúc của Bi-LSTM được thể hiện như hình 2.

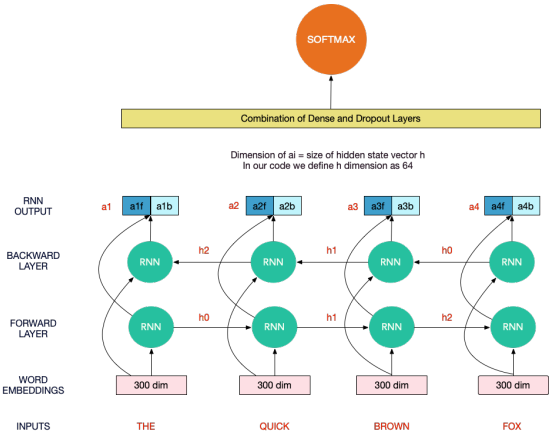

Figure 2: Kiến trúc mô hình Bi-LSTM

4.3 Text-CNN

CNN là một lớp mạng nơ-ron nhân tạo sâu, chuyển tiếp (nơi kết nối giữa các nút không tạo thành chu kỳ) sử dụng một biến thể của các perceptron nhiều lớp được thiết kế để yêu cầu xử lý trước tối thiểu. Chúng được lấy cảm hứng từ vỏ não thị giác của động vật.

CNNs được thường được sử dụng trong các tác vụ xử lý ảnh tuy nhiên đối với NLP (Zhang and Wallace, 2015) CNN cũng mang đến đầy hứa hẹn.Hình 3 là kiến trúc CNN đối với NLP.

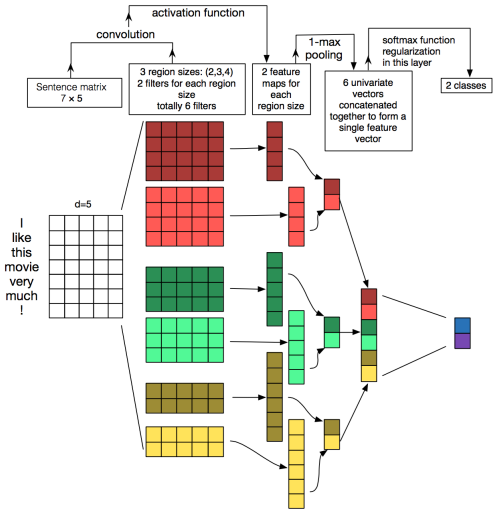

Figure 3: Kiến trúc mô hình Text-CNN

4.4 PhoBERT

PhoBERT là một pre-trained được huấn luyện monolingual language dành riêng cho tiếng Việt. Việc huấn luyện dựa trên kiến trúc và cách tiếp cận giống RoBERTa (hình 4) của Facebook được Facebook giới thiệu vào năm 2019. Tương tự như BERT, PhoBERT cũng có 2 phiên bản là PhoBERTbase với 12 transformers block và PhoBERTlarge với 24 transformers block. PhoBERT được train trên khoảng 20GB dữ liệu bao gồm khoảng 1GB Vietnamese Wikipeia corpus và 19GB còn lại lấy từ Vietnamese news corpus. PhoBERT sử dụng RDRSegmenter của VnCoreNLP để tách từ cho dữ liệu đầu vào trước khi qua BPE encoder. Do tiếp cận theo kiến trúc RoBERTa, PhoBERT chỉ ứng dụng task Masked Language Model để train, bỏ đi task Next Sentence Prediction.

Nhóm truncate decoder của BERT, giữ nguyên kiến trúc encoder của transformer và sau đó trích suất ra biểu diễn véc tơ của token CLS đánh dấu vị trí đầu tiên. Véc tơ này sẽ được sử dụng làm đầu vào cho thuật toán classifier bằng cách thêm một linear projection layer (cũng chính là fully connected layer) ở cuối có kích thước bằng với số classes cần phân loại.

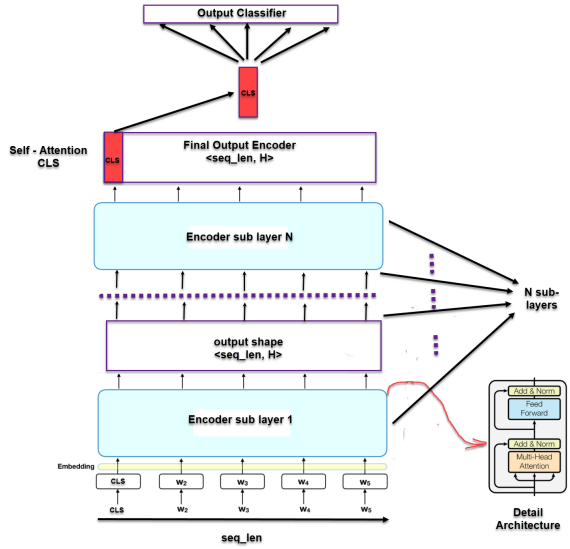

Figure 4: Kiến trúc mô hình RoBERTa

5 Thực nghiệm

5.1 Huấn luyện mô hình

Tất cả các mô hình với dữ liệu đầu vào sẽ đều được tiền xử lý như chúng tôi đã đề cập ở 4.1. Trong nghiên cứu này chúng tôi sử dụng các bộ nhúng từ fastText, W2V và PhoW2V để so sánh, các bộ nhúng từ này được huấn luyện cấp độ word và các phương pháp chúng tôi thực hiện đều ở cấp độ word.

Với cài đặt mô hình Bi-LSTM, Chúng tối sử dụng bộ nhúng từ W2V và fastText, xây dựng mô hình Bi-LSTM với các thông số như sau: Thông số ban đầu: maxlen=100, embed_size=300. Input: maxlen. Embedding: input_dim=max_features, output_dim=embed_size, weights=[embedding_matrix], trainable=True. Mô hình Bi-LSTM với 64 units, return_sequences=True, dropout=recurrent_dropout=0.2. Dense: 3(tương ứng với 3 nhãn đầu ra), activation=relu.

Với cài đặt mô hình Text-CNN, chúng tôi sử dụng fastText, W2V và PhoW2V tạo ma trận biểu diễn từ cho dữ liệu, tạo ma trận biểu diễn từ cho các từ xúc phạm xuất hiện trong dữ liệu. Mô hình chúng tôi sử dụng lớp tích chập 2D(2D Convolution Layer) với số filter là 32 và kích cỡ 3, 4, 5, hàm kích hoạt là softmax.

Cuối cùng chúng tôi sử dụng mô hình PhoBert Pre-trained với các thông số được tinh tỉnh như sau: batch_size=32, epochs=5, weight_decay: 0.001, metrics là accuracy và f1_score ('macro').

5.2 Kết quả

Các mô hình thực nghiệm các phương pháp với bộ dữ liệu ViHSD được sử dụng các độ đo Accuracy và F1-score macro để đánh giá hiệu suất mô hình. Kết quả khi kết hợp các phương pháp đề xuất trên, mô hình của chúng tôi có cả thiện so với mô hình cơ sở từ 1% đến 2%. Kết quả mỗi mô hình được trình bày tại Bảng 3 dưới đây:

Table 3: Kết quả đánh giá trên tập test của bộ dữ liệu ViHSD

Model	**Accuracy**	**F1-score macro**
Navie Bayes + TF-IDF	0.83	0.31
Logictic Regression + TF-IDF	0.86	0.51
SVM + TF-IDF	0.83	0.30
Bi-LSTM + Embedding tự tạo	0.86	0.58
Bi-LSTM + W2V	0.83	0.30
Bi-LSTM + fastText	0.83	0.30
Text-CNN + Embedding tự tạo	0.86	0.59
Text-CNN + W2V	0.86	0.60
Text-CNN + fastText	0.86	0.60
Text-CNN + PhoW2V	0.86	0.60
PhoBERT	0,86	**0.64**
BERT4News	0,86	**0.63**

Từ kết quả trên, mô hình PhoBERT mô hình có kết quả cao nhất với f1-score macro bằng 0.64.

6 Phân tích lỗi sai

Qua quá trình thực nghiệm chúng tôi thấy rằng các mô hình phân loại nhãn CLEAN tốt hơn so với 2 nhãn OFFENSIVE và HATE. Hình 5 thể hiện Confusion Matrix của model PhoBERT , ta có thể thấy rằng hầu hết các câu có nhãn OFFENSIVE được dự đoán nhãn là CLEAN so với câu nhãn HATE được dự đoán đúng nhiều hơn. Việc này được giải thích là do sự mất cân bằng lớn trong

dữ liêu, số lượng câu nhãn CLEAN là lớn hơn rất nhiều so với 2 nhãn còn lại. Phương pháp xử lý dữ liệu chưa được tốt, có nhiều từ viết tắt, teencode, sai chính tả trong câu mà tiền xử lý không thể bao quát hết được. Tóm gọn lại các mô hình chúng tôi sử dụng chưa phân loại tốt trên các nhãn OFFENSIVE và HATE.

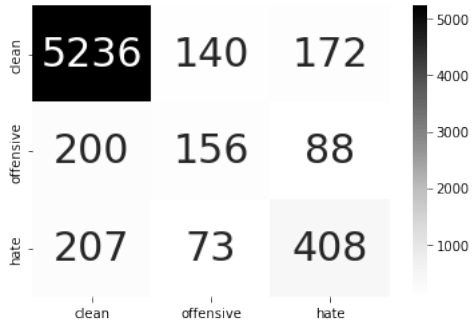

Figure 5: Confusion Matrix mô hình PhoBERT

7 Kết luận và hướng phát triển

7.1 Kết luận

Trong đồ án này, chúng tôi đã tiến hành phân tích bình luận xúc phạm trên bộ dữ liệu Tiếng Việt ViHSD với việc tiền xử lý bộ dữ liệu kết hợp với các mô hình máy học và mô hình học sâu để cải thiện độ chính xác của mô hình. Sau khi chúng tôi thực hiện các bước tiền xử lý thông thường và chuẩn hóa các từ sai chính tả, teencode kết hợp với các mô hình học sau thì chúng tôi thu được kết quả F1-score macro cải thiện khoảng 1% đến 2% ở mô hình PhoBERT và BERT4News.

Cuối cùng, nhóm đã thu được mô hình tốt nhất là mô hình PhoBERT với F1-score macro là 64% , cao hơn mô hình cơ sở với F1-score macro là 62,69%.

7.2 Hướng phát triển

Đã có nhiều bài nghiên cứu về bài toán phân loại bình luận xúc phạm trên mạng xã hội trên nhiều ngôn ngữ khác nhau những hầu hết đều cho kết quả không được tốt. Bởi vì ngôn ngữ trên mạng xã hội của chúng ta đa dạng, phong phú, ngoài ra còn biến đổi liên tục bằng nhiều cách như là viết tắt, icon, teencode ,các ngôn ngữ mới,.... Những thay đổi đó làm cho độ chính xác của mô hình bị giảm. Mặc dù mô hình mà chúng em đề ra trong đồ án này đã có sự cải thiện so với các mô hình cơ sở, nhưng hiệu vẫn còn chưa cao và đây cũng là một thách thức đối với các nghiên cứu trong tương lai để cải thiện hiệu suất của mô hình cho nhiệm vụ phát hiện lời nói căm thù ở Việt Nam. Trong tương lại, để cải thiện độ chính xác của mô hình, chúng em đề xuất các phương pháp như là tăng số lương, kích thước và độ đa dạng của tập từ vựng nhằm cập nhật mới nhất các từ mới xuất hiện trên mạng xã hội, cải thiện phương pháp tiền xử lý dữ liệu nhằm tăng độ chính xác, áp dụng những mô hình máy học và học sâu thích hợp hơn để cải thiện độ chính xác của mô hình.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Son T Luu, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. A large-scale dataset for hate speech detection on vietnamese social media texts. *arXiv preprint arXiv:2103.11528*.

Luan Thanh Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Constructive and toxic speech detection for open-domain social media comments in vietnamese. *arXiv preprint arXiv:2103.10069*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Ye Zhang and Byron Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.