ĐỀ XUẤT THUẬT TOÁN DỰA TRÊN CHATBOT ĐỂ PHÁT HIỆN CÁC BÌNH LUẬN NHẠY CẢM

Nguyễn Hữu Phát*, Đỗ Mạnh Cầm* , Hoàng Văn Quang†
*Bộ môn Mạch và Xử lý tín hiệu, Viện Điện tử viễn thông, Trường Đại học Bách Khoa Hà Nội
† Phòng Tổ chức cán bộ, Trường Đại học Bách Khoa Hà Nội

Tóm tắt: Hiện nay văn hóa ứng xử trên mạng xã hội đang là một vấn đề gây nhức nhối dư luận trong khoảng thời gian qua. Các cơ quan nhà nước cũng đã và đang bắt tay vào để làm sạch môi trường mạng của đất nước. Hàng loạt bộ luật chương trình lên án và xử lý những video và clip có nội dung phản cảm. Nhưng đó chỉ là một phần rất nhỏ trong quá trình làm sạch môi trường mạng. Thứ chúng ta thấy hàng ngày đó chính là những bình luận mang nội dung phản cảm trên các trang mạng xã hội. Nó tồn tại ở bất cứ đâu từ mạng xã hội đến các cộng đồng, trong các game online. Hiện tại trên thị trường cũng không có quá nhiều trang mạng xã hội, game online. Do đó không quá khó cho các cộng đồng có các biện pháp hạn chế những từ ngữ không phù hợp xuất hiện. Xuất phát từ thực tế đó trong bài báo này, chúng tôi đề xuất thuật toán dựa trên chatbot để phát hiện các bình luận nhạy cảm với hi vọng có thể góp phần nào vào việc làm sạch môi trường mạng, mang đến sự thoải mái mỗi khi tham gia các trang mạng xã hội hiện thời. Kết quả chỉ ra rằng hệ thống đạt được độ chính xác lên đến 75% với 100.000 bình luận được thử nghiệm.

Từ khóa: Chatbot, bình luận phản cảm, văn hóa ứng xử, online, xử lý dữ liệu.

I. ĐẶT VẤN ĐỀ

Với sự bùng nổ của internet như hiện nay, số lượng người sử dụng ngày càng nhiều. Ví dụ như trang mạng xã hội lới nhất hiện nay Facebook, tính đến 31/3/2020 có đến 2,6 tỷ người sử dụng và 1,7 tỷ người sử dụng hàng ngày [1], [2]. Nguyên trên Việt Nam, với dân số trên 90 triệu dân thì có đến 64 triệu tài khoản FaceBook đủ để thấy số lượng người đang dùng các trạng mạng xã hội lớn như thế nào. Trong số đó không thiếu các thành phần luôn luôn để lại những lời bình luận đầy phản cảm, đi ngược lại dư luận khiến người đọc khó chịu. Để tránh những tác hại xấu đến tương lai, chúng ta phải thực hiện loại bỏ ngay. Vì thế chúng tôi đưa ra đề xuất chatbot quản lý bình luận để góp phần giải quyết vấn đề này.

Hệ thống đối thoại người máy hay còn gọi với thuật ngữ là chatbot [3]. ChatBot là một chương trình máy tính tiến hành cuộc trò chuyện thông qua nhắn tin nhanh, nó có thể tự động trả lời những câu hỏi hoặc xử lý tình huống. Phạm vi và sự phức tạp của ChatBot được xác định bởi thuật toán của người tạo nên chúng. ChatBot thường được ứng dụng trong nhiều lĩnh vực như thương mại điện tử, dịch vụ khách hàng, y tế, tài chính ngân hàng, các dịch vụ giải trí.

Chatbot có thể được chia thành 2 loại:

- Hệ thống hướng mục tiêu trên một miền ứng dụng (Task-Oriented)(hay còn gọi là Miền đóng (Close Domain))

Miền đóng (Close Domain): Mô hình trả lời tự động thuộc miền đóng thường tập trung vào trả lời các câu hỏi đối thoại liên quan đến một miền cụ thể, ví dụ như: Y tế, Giáo dục, Du lịch, Mua sắm, ..

Trong một miền đóng cụ thể, không gian các mẫu hỏi input và output là có giới hạn, bởi vì các hệ thống này đang cố gắng để đạt được một mục tiêu rất cụ thể. Hệ thống hỗ trợ kỹ thuật (Technical Customer Support) hay tư vấn và hỗ trợ mua hàng (Shopping Assistants) là các ứng dụng thuộc miền đóng. Các hệ thống này không thể đối thoại về "Chính trị" hay "Pháp luật", chúng chỉ cần thực hiện các nhiệm vụ cụ thể một cách hiệu quả nhất có thể. Chắc chắn, người dùng vẫn có thể hỏi đáp bất cứ gì, nhưng hệ thống không yêu cầu phải xử lý những trường hợp ngoại lệ này.

- Hệ thống không có định hướng mục tiêu (chit-chat)(hay còn gọi là Miền mở (Open Domain))

Miền mở (Open Domain): Mô hình trả lời tự động trên miền mở cho phép người dùng có thể tham gia trò chuyện với một chủ đề bất kỳ, không nhất thiết phải có một mục tiêu rõ ràng hay một ý định cụ thể nào. Các cuộc trò chuyện trên mạng xã hội như Facebook, Twitter thường là miền mở, chúng có thể đi vào tất cả các chủ đề. Số lượng các chủ đề thảo luận được đề cập đến là không giới hạn, do đó, tri thức yêu cầu được tạo ra để trả lời các câu đối thoại thuộc miền mở trở nên khó hơn. Tuy nhiên, việc thu thập trích rút dữ liệu từ miền này khá phong phú và đơn giản.

Mỗi cách tiếp cận bài toán đều có hướng giải quyết khác nhau dẫn tới các kỹ thuật sử dụng khác nhau.

Hiện nay, với việc các trang mạng xã hội ngày càng phổ biến, với việc các bình luận không được kiểm soát một cách triệt để thì những câu phản cảm, những câu nói không phù hợp xuất hiện ngày càng nhiều gây nhức mắt những người cùng tham gia cộng đồng.

Tác giả liên hệ: Nguyễn Hữu Phát
Email: phat.nguyenhuu@hust.edu.vn
Đến tòa soạn: 12/2020, chỉnh sửa: 3/2021, chấp nhận đăng: 4/2021

Tuy nhiên vấn đề khó khăn ở đây là các từ trong tiếng việt có khả năng kết hợp với nhau một cách kỳ diệu để tạo ra những câu nói vô cùng đa dạng.

- Chương trình có thể tích hợp vào nhiều loại ngôn ngữ lập trình khác nhau. Cần biến chat bot thành dạng như một lib có thể sử dụng rộng rãi.

Hình 1. Sơ đồ cấu trúc hệ thống

Ví dụ: Hổ mang bò lên núi. Có thể hiểu là con hổ mang con bò lên núi mà cũng có thể hiểu được là rắn Hổ Mang đang bò lên núi.

Điều này cũng áp dụng với những câu phản cảm. Tùy vào ngữ cảnh mà một câu có thể hiểu là câu phản cảm. Nếu như liệt kê tất cả các từ đó thành các từ cấm và kiểm soát là hoàn toàn có thể. Nhưng điều này cần một cơ sở dữ liệu rất là lớn. Mà còn chưa kể đến việc con người là những người rất biết lách luật. Cùng một cách diễn đạt thì bọn họ có thể diễn đạt kiểu khác như:

- Dùng từ trái nghĩa. Ví dụ: Ngu như bò với không thông minh bằng con bò.
- Dùng những từ đồng nghĩa. Ví dụ: Ngu như bò với dốt như heo.
- Dùng cách viết đánh vần.
- Dùng cách viết tắt.
- Dùng cách nói lái.
- Dùng cách thêm, bớt từ.

Với vô vàn cách để biểu diễn như vậy thì việc thống kê hết tất cả các trường hợp phản cảm nói tục là rất khó khăn. Chúng ta cần rất nhiều thời gian để thu thập và tổng hợp. Không chỉ thế còn cần liên tục bổ sung không ngừng để có thể bắt kịp thời đại. Công sức chúng ta bỏ ra chưa chắc đã thu về được hiệu quả. Do đó, chúng ta cần tìm một phương pháp khác để giải quyết vấn đề này.

Trong bài viết này, chúng tôi sẽ đề xuất phương pháp để giải quyết sự đa dạng trong việc phản cảm trên mạng, để từ đó tạo ra một mô hình chatbot có thể nhận diện và phân loại các câu nói không phù hợp trên mạng.

Nội dung bài báo được tổ chức như sau. Sau phần I giới thiệu, chúng tôi sẽ trình bày các phương hướng giải quyết vấn đề trong phần II. Phần III cho thấy kết quả thử nghiệm của thuật toán. Phần IV là kết luận và hướng phát triển mới của đề tài.

II. NỘI DUNG CẦN GIẢI QUYẾT

A. Xây dựng lý thuyết

Để thuận lợi cho việc thiết kế, ta cần phải đề ra những yêu cầu cho thuật toán (chatbot) cũng như kết quả cuối cùng chúng ta cần đạt được.

Ở đây, yêu cầu chúng tôi đặt ra với chatbot sẽ là:

- Tự động phát hiện các câu nói không phù hợp với độ chính xác cao từ 70\% trở lên.

Từ những yêu cầu, mục đích trên kết hợp thêm với hiểu biết về chatbot chúng tôi đưa ra sơ đồ cấu trúc của hệ thống như hình 1.

Trong đó:

- Khối tiền xử lý: Tiến hành chuyển đổi câu đầu vào thành một mảng chứa các từ có ý nghĩa. Nó gồm các bước: tách từ tiếng việt; làm sạch dữ liệu; xử lý các từ không có nghĩa; và cuối cùng là xác định ý nghĩa của từng từ.
- Khối xác định mức độ phản cảm: Dựa vào một mảng đã xác định ở trên, cộng thêm một quy chuẩn đề ra từ đó xác định mức độ phản cảm của cả câu.
- Trả lời: Từ mức độ phản cảm của câu và những thành phần cấu tạo nên điều này. Chatbot sẽ tiến hành đưa ra câu trả lời thích đáng nhất.

Trong phạm vi nghiên cứu, chúng tôi chưa tìm thấy tài liệu nào nghiên cứu về các thuật toán để xử lý từ nhạy cảm trong tiếng Việt. Từ sơ đồ cấu trúc hệ thống, ta tiến hành đi phân tích chi tiết từng vấn đề cần xử lý.

A.A.1 Thu thập dữ liệu

Khó khăn trong việc kiểm tra hiệu quả của chatbot chính là bộ dữ liệu những câu bình luận trên các trang mạng xã hội. Hiện tại chúng tôi không tìm thấy data những câu bình luận do đó đã tiến hành tự tạo dựa vào lấy bình luận trên facebook. Hiện tại, bộ dữ liệu của chúng tôi có khoảng 100.000 câu bình luận.

A.A.2 Tiền xử lý dữ liệu

Chúng tôi chia thành bốn bước:

Tách từ tiếng Việt:

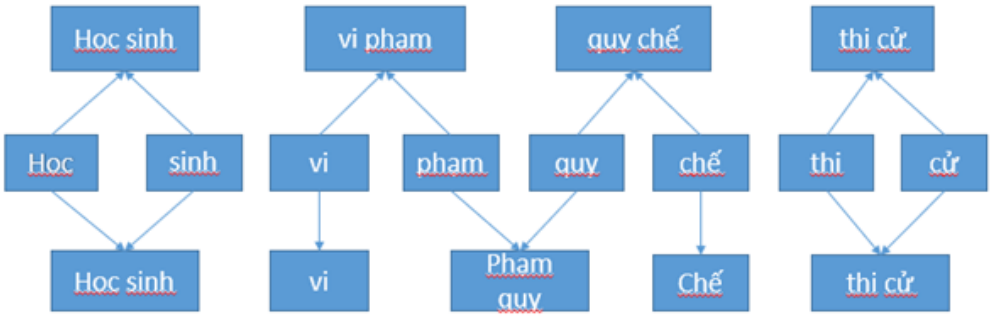
Xử lý ngôn ngữ tự nhiên bao gồm rất nhiều các bài toán như dịch tự động (machine translation), tóm tắt văn bản (text summarization), tìm kiếm thông tin (information retrieval), trích chọn thông tin (information extraction). Muốn giải quyết được các bài toán trên thì bài toán phân tách từ (word segmentation) là bài toán quan trọng nhất, nó quyết định thành công của các bài toán khác.

Để hiểu được vì sao cần một bài toán tách từ thì chúng ta cần biết một số đặc tính chính của từ trong tiếng Việt là:

- Từ ở dạng nguyên thể, hình thức và ý nghĩa của từ độc lập với cú pháp.
- Từ bao gồm từ đơn và từ phức, bao gồm từ láy, từ ghép.
- Từ được cấu trúc từ tiếng. Việc nhận biết từ trong tiếng Việt được gọi là phân cụm từ:

Trong hình 2, có nhiều hơn một cách để hiểu câu văn này:

1. (Học sinh) (vi phạm) (quy chế) (thi cử).



Hình 2. Vấn đề phân cụm từ trong tiếng Việt [6].

2. (Học sinh) (vi) (phạm quy) (chế) (thi cử).

Câu văn này không mang ý nghĩa.

Như chúng ta đã biết, văn bản tiếng Việt đặt dấu cách giữa các âm tiết chứ không phải giữa các từ. Một từ có thể có một, hai hoặc nhiều âm tiết nên có nhiều cách phân chia các âm tiết thành các từ, gây ra nhập nhằng. Việc phân giải nhập nhằng này gọi là bài toán tách từ.

Tiêu chí quan trọng nhất trong bài toán tách từ đương nhiên là độ chính xác. Hiện tại người ta đã đạt được độ chính xác lên đến 97% tính theo từ. Tuy nhiên nếu tính theo câu (số câu được tách hoàn toàn đúng/tổng số câu) thì độ chính xác chỉ khoảng 50%. Sự chênh lệch này nguyên nhân là do sự phức tạo của tiếng Việt.

Chúng ta lấy ví dụ một câu khá nổi tiếng về sự phức tạp của tiếng việt: Hổ mang bò lên núi.

Câu này tùy theo cách chia câu có thể hiểu theo hai cách.

- Hổ mang/ bò/ lên núi. Câu này có nghĩa là con rắn hổ mang đang bò lên núi.
- Hổ/ mang/ bò/ lên núi. Câu này có nghĩa là Con hổ đang mang con bò lên núi.

Cả 2 cách tách này đều đúng, đều có thể nhưng lại tạo ra những câu có ý nghĩa khác nhau. Do đó độ chính xác khi tính theo câu mới nhỏ như vậy.

Đây có thể nói là vấn đề khá là nghiêm trọng trong quá trình xác định ý nghĩa của câu bởi vì chỉ cần thay đổi một chút thì hoàn toàn có thể khiến câu có nghĩa khác hoàn toàn.

Hiện tại có một số cách tiếp cận bài toán tách từ như sau [4]:

- Ghép cực đại: Đặt các từ vào câu sao cho phủ hết được câu đó, thoả mãn một số heuristic nhất định. Phương pháp này các ưu điểm là rất nhanh, nhưng có rất nhiều hạn chế, ví dụ như độ chính xác thấp, không xử lý được những từ không có trong từ điển.
- Xây dựng tập luật bằng tay hoặc tự động để phân biệt các cách kết hợp được phép và không được phép.
- Đồ thị hoá: Xây dựng một đồ thị biểu diễn câu và giải bài toán tìm đường đi ngắn nhất trên đồ thị.
- Machne Learning: Coi như bài toán gán nhãn chuỗi. Cách này được sử dụng trong JVNSegmenter, Đông du.
- Dùng mô hình ngôn ngữ: Cho trước một số cách tách từ của toàn bộ câu, một mô hình ngôn ngữ có thể đánh giá được cách nào có khả năng cao hơn. Đây là cách tiếp cận của vnTokenizer.

Trong bài viết lần này, chúng tôi sử dụng phương pháp **Ghép cực đại**.

Làm sạch dữ liệu:

Sau khi tách từ, văn bản còn xuất hiện nhiều ký tự đặc biệt, dấu câu, … Những thành phần này làm giảm hiệu quả trong quá trình xử lí. Trong phần này, chúng tôi chuyển tất cả những từ in hoa về chữ in thường, xóa bỏ các dấu câu.

Xử lý những từ không có nghĩa:

Đây chính là điểm mấu chốt của bài báo lần này. Không như những bài viết đã có những quy chuẩn, các từ ngữ sử dụng phải chính xác.Trong các câu comment trên mạng thường xuyên sử dụng các từ viết tắt và cách nói lái nói tắt. Chỉ có một phần nhỏ là sử dụng thẳng thừng những câu phản cảm thông dụng.Nếu chỉ có tách từ và phân loại sẽ bỏ sót rất nhiều câu phản cảm vẫn đang tồn tại trên mạng xã hội.Mà một vài cách nói lái câu phản cảm thường dùng là

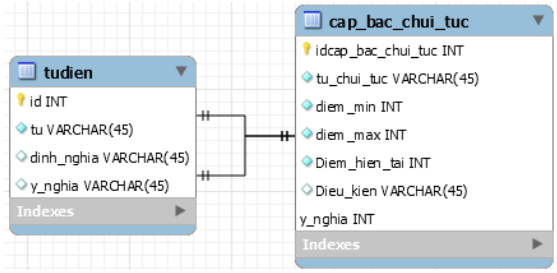
- Sử dụng các từ thay thế để nói lái câu đi.
- Sử dụng các dấu để vào giữa câu.

Những cách viết này đều có thể biểu đạt trọn vẹn nghĩa của từ phản cảm nhưng khi tách từ, nó sẽ không tạo ra được từ mang ý nghĩa phản cảm. Đây chính là những cách thường thấy để phản cảm. Điểm chung của hai cách này là những từ tách ra là những từ không có nghĩa hoặc là từ đơn. Từ điểm chung đó, ta tiến hành xử lý các từ đã được tách ra.

Như trong hình 2, chúng ta sẽ có 2 bước để xử lý vấn đề này.

- Ghép từ: Áp dụng với những từ có một đến hai chữ cái đứng cạnh nhau hoặc lớn hơn hai từ không có nghĩa

đứng cạnh nhau. Tiến hành ghép chúng lại thành một từ mới. Nếu từ đó có nghĩa thì tiến hành ghép lại. Ngược lại thì thực hiện bước xử lý số 2.



Hình 3. Thiết kế database xác định mức độ phản cảm.

- Đổi chỗ: Các chữ cái trong tiếng việt sẽ được chia thành nguyên âm và phụ âm. Các từ trong tiếng việt sẽ được cấu tạo từ những nguyên âm và phụ âm này.Trong đó có rất nhiều những từ không có nghĩa nhưng có nguyên âm và phụ âm giống với những tiếng phản cảm. Vì vậy nó sẽ được sử dụng như từ thay thế cho các từ phản cảm và người nghe vẫn có thể hiểu được ý nghĩa của từ đó. Nắm được điểm này, với những từ không có nghĩa, chúng ta tiến hành phân tách nguyên âm và phụ âm. Nếu ghép được các từ mang ý nghĩa phản cảm thì tiến hành cập nhật vào từ điển của bản thân.

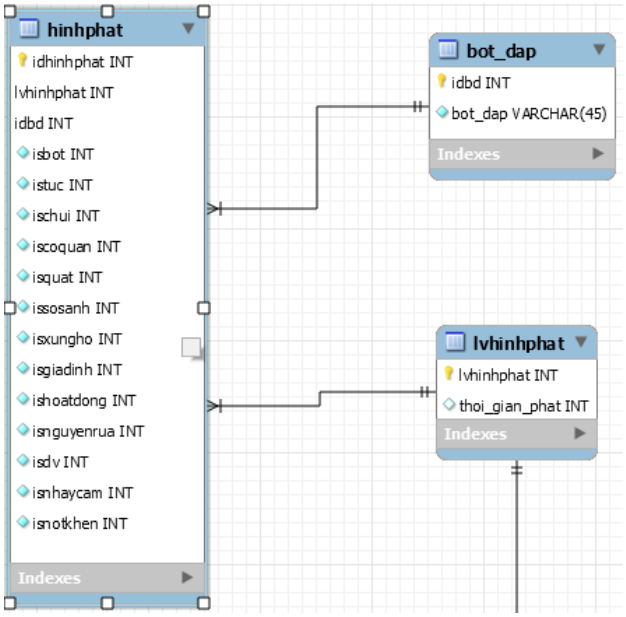
Qua bước này, chúng ta đã giải quyết được hai trong số vô số cách viết các từ phản cảm. Nâng cao khả năng chính xác trong quá trình tách từ những câu bình luận, câu nói hàng ngày.

Phân loại ý nghĩa của từ:

Như đã đề cập ở trên, một trong số những cách hay dùng nhất để phản cảm là sử dụng những từ đồng nghĩa và trái nghĩa.Để giải quyết vấn đề này, chúng tôi đề xuất cần phải nhóm các từ đồng nghĩa, trái nghĩa hay được sử dụng vào thành từng nhóm ý nghĩa.

Ở đây, chúng tôi đang phân chia các thành phần thường xuyên được sử dụng trong các từ phản cảm thành các nhóm sau.

- Từ phản cảm: Những từ mang ý nghĩa phản cảm.
- Từ chê bai: Những từ thường dùng khi mà xúc phạm người khác. Ví dụ: Ngu, dốt, ngốc
- Đại từ nhân xưng: Các từ được sử dụng để xưng hô những người thân trong nhà. Ví dụ: cha, mẹ
- Động vật: Các từ chỉ động vật. Ví dụ: chó, mèo
- Những từ xưng hô: Những từ xưng hô mang tính xuồng xã. Ví dụ: mày, tao, thằng
- Quát: Các từ ra lệnh, quát nạt người khác. Ví dụ:im, câm, cút, nín
- Cơ quan: Các từ chỉ cơ quan con người. Ví dụ: mắt, mũi, mồm.
- Từ nhạy cảm: Các từ liên quan đến vấn đề sinh lý của con người.



Hình 4. Thiết kế database câu trả lời chatbot

- Từ so sánh: Các từ dùng để so sánh. Ví dụ: giống, khác, hơn, kém
- Từ phủ định: Các từ mang nghĩa phủ định. Ví dụ: không, not
- Nguyền rủa: các từ liên quan đến bệnh tật, chết chóc. Ví dụ: chết, nguyền
- Hoạt động: Các từ chỉ hoạt động. Ví dụ: Đi, đứng ăn.
- Các từ chỉ hoạt động xuồng xã.

Xác định mức độ phản cảm:

Nhìn phân chia các từ đồng nghĩa trái nghĩa có lẽ rất nhiều người sẽ thắc mắc vì sao rất nhiều nhóm từ chẳng có tý phản cảm này như là hoạt động, so sánh hay là cơ quan. Nhưng chỉ vài từ đó cũng có thể tạo ra những câu phản cảm. Do một từ chỉ hoạt động và một đại từ quan hệ tạo ra. Nó vẫn có phần nào đó phản cảm. Không chỉ thế, có những câu phản cảm theo người này là vô cùng tục nhưng với người khác lại cảm thấy bình thường.Để giải quyết vấn đề này, chúng tôi tiến hành tạo một quy chuẩn để xác định mức độ phản cảm của câu. Quy chuẩn được thể hiện như trên bảng 1.

Điểm phản cảm sẽ là tổng của tất cả các điểm tục từ các từ có có ý nghĩa như trên. Từ đó chúng tôi đề xuất chia thành 6 cấp độ như sau:

- Mức 0(0-3 điểm): Câu không phản cảm, không tiến hành xử phạt.
- Mức 1(4-7 điểm): Những câu nói xuồng xã, những câu chửi không sử dụng các từ mang tính chê bai. Nhưng nếu lặp lại nhiều có thể lên đến mức hai.
- Mức 2(8-11 điểm): Những câu đã có mục đích xúc phạm người khác. Cần tiến hành cảnh cáo, xử phạt.

- Mức 3(12-15 điểm): những câu chửi mang tính phản cảm, xúc phạm, chứa đựng những từ ngữ mang tính tục. Cần xử phạt.

Bảng 1. Quy chuẩn xác định các câu không phù hợp

Ý nghĩa của từ	Ví dụ	Điểm tục	Điểm tục lớn nhất	Điều kiện
Từ phản cảm	***	14	Không giới hạn	Không
Từ dùng chửi, chê	Ngu, dốt, chảnh,…	7	14	Không
Đại từ nhân xưng	Cha, mẹ, anh chị...	2	2	không
Động vật	Chó, mèo...	2	2	không
Xưng hô không phù hợp	Mày, tao, thằng...	1	2	Đi liền với từ chỉ động vật thì điểm tục tăng 1
Quát	Im, nín, cút…	3	9	Không
Từ chỉ cơ quan người	Mắt, mũi miệng...	1	2	Từ chửi, động vật,
Từ nhạy cảm	***	9	9	không
Từ so sánh	Giống, như,…	2	2	Với những từ có điểm tục>=3
Phủ định	Mang nghĩa phủ định	2	2	Có tác dụng khi đi với những từ có ý nghĩa khen
Nguyền rủa	Chết, biến,…	7	7	Đi liền với đại từ nhân xưng
Hoạt động	Đi, đứng, ăn,…	1	1	Đi liền với từ nguyền rủa.
Những từ xuồng xã không phù hợp	Vãi, đù,…	4	8	Không

- Mức 4(16-19 điểm): Những câu chứa các từ ngữ phản cảm mang tính xúc phạm cao. Cần xử phạt mạnh tay
- Mức 5(>=20 điểm): Nhưng câu mang đầy những từ phản cảm không chấp nhận được. Cần răn đe để làm gương.

Qua bước này, chúng ta đã xác định được tính phản cảm cũng như mức độ phản cảm của các câu bình luận riêng biệt để từ đó có thể đưa ra những biện pháp xử lý cũng như cảnh cáo phù hợp.

Trả lời:

Đây là phần chatbot sẽ tương tác với người dùng. Chatbot sẽ tác động đến người dùng qua hai yếu tố sau:

- Hình phạt: Như đã đề cập ở phần trên, ứng với từng mức độ không phù hợp của các câu bình luận, tiến hành đưa ra những hình phạt thích đáng như: cấm chat 15 phút hay 1 tiếng. Cấm tài khoản 15 phút, 1 tiếng. Các hình phạt này tùy vào như cầu của admin có thể thay đổi.
- Câu cảnh báo: Bên cạnh hình phạt thì chatbot sẽ gửi trên kênh chat những câu cảnh báo, thông báo về hình phạt mà người dùng mắc phải. Các câu thông báo sẽ dựa vào 2 yếu tố.
- Mức độ của lời nói: Đây là yếu tố quyết định hình phạt cũng như mức độ nặng nhẹ của câu cảnh báo.
- Ý nghĩa của các từ: Mỗi câu nói không phù hợp có thể do nhiều từ có các yếu tố khác nhau cấu tạo thành. Dựa vào những yếu tố cấu thành đó, chúng tôi sẽ đưa ra những lời cảnh báo, thông báo khác nhau.\\

Tất cả thông tin này sẽ được lưu trữ dưới dạng database (Hình 3), thuận lợi cho người quản lý có thể thay đổi tùy theo ý muốn.

B. Thiết kế CSDL

Các mạng xã hội hiện giờ rất đa dạng, được viết bởi rất nhiều ngôn ngữ. Để chatbot này có thể phổ biến thì cần phải có thể sử dụng ở tất cả các loại ngôn ngữ.

Để làm được điều đó thì chương trình này không được sử dụng các thư viện riêng biệt của từng loại ngôn ngữ khác nhau mà chỉ sử dụng những thư viện phổ biến. Tuy nhiên điều đó là chưa đủ. Người Việt được đánh giá là khá thông minh và biết cách để lách luật, phát minh ra những cách nói khác nhau để thể hiện sự phản cảm mà lại không vi phạm. Do đó thì để có thể đảm bảo hiệu quả của chatbot, cần không ngừng cập nhật, mở rộng.

Ở đây sẽ có hai yếu tố cần thiết để đảm bảo việc này.

- Số lượng từ ngữ Việt Nam đã có phải đủ phong phú để không dẫn đến việc nhầm lẫn
- Cần một chương trình có thể thuận lợi cho việc cập nhật thường xuyên mà không cần những thao tác rườm rà.

Để thỏa mãn yếu tố đầu tiên là một nhiệm vụ rất khó. Với mười hai nguyên âm và mười bảy phụ âm thì số lượng từ có thể ghép từ chúng là một con số vô cùng lớn khó ai có thể thống kê hết. Chỉ có thể không ngừng cập nhật, không ngừng hoàn thiện theo thời gian, theo số lượng câu bình luận ngày càng nhiều và phong phú. Do đó chúng ta chú trọng vào điều thứ hai. Đó là làm sao để dễ dàng có thể cập nhật thường xuyên một cách dễ dàng không cần phải rườm rà, không cần mỗi lần chỉnh sửa lại phải bảo trì để cập nhật phiên bản.

Chúng tôi không có biện pháp nào để có thể tránh hoàn toàn điều này nhưng muốn hạn chế tối đa việc này. Chúng tôi cố gắng dựa nhiều vào cơ sở dữ liệu hiện có để xử lý như trên hình 3, trong đó:

- Diem_min, diem_max là điểm nhỏ nhất và lớn nhất ứng với từng nhóm đã tạo ở trên,
- Diem_hien_tai là điểm số của từng nhóm ý nghĩa,
- Dieu_kien là điều kiện của ý nghĩa đó(nếu có),
- Tu_chui_tuc là từ phản cảm ứng với nhóm ý nghĩa đó.

Dựa vào bảng từ điển, ta có thể chia các câu thành các từ và cụm từ mang ý nhĩa.

Bảng 2. Thống kê đánh giá kết quả training

Số lượng comment	Thời gian xử lý dự kiến(Tối đa)	Số lượng comment mang ý phản cảm được tìm ra	Số lượng chính xác về cấp bậc trong số comment phản cảm đã tìm ra	Tỷ lệ chính xác	Thời gian xử lý thực tế	Đánh giá
100.000	80h	10.211	7696	75.36%	115h	Chưa đạt

Sau đó dựa vào ý nghĩa đó. Ta tiến hành cập nhật điểm hiện tại vào bảng 1 ứng với các nhóm ý nghĩa nằm trong bảng đó. Từ đó xác định ra được cấp độ phản cảm của câu nói để tiến hành xử lý.

Một câu trả lời có thể dùng cho nhiều hình phạt. 1 lv trừng phạt có thể sử dụng cho nhiều hình phạt. Do đó cơ sở dữ liệu thể hiện cho chức năng hình phạt sẽ gồm ba bảng với hai quan hệ 1-n với nhau như hình 4, trong đó:

- thoi_gian_phat: Thời gian trừng phạt ứng với mỗi mức trừng phạt tính theo phút,
- bot_dap: câu trả lời của bot,
- isbot, istuc, ischui, iscoquan, isquat, issosanh, isxungho, isgiadinh, ishoatdong, isnguyenrua, isdv, isnhaycam, isnotkhen: sự tồn tại của các yếu tố kiểm tra câu có phải câu phản cảm hay không.

Như vậy, chúng tôi đã tiến hành dữ liệu hóa việc xác định cấp bậc cũng như cách chatbot đáp lại. Tùy vào ý muốn của người sử dụng có thể điều chỉnh theo ý muốn của bản thân một cách dễ dàng.

III. KẾT QUẢ VÀ THẢO LUẬN

Chúng tôi tiến hành thực hiện chạy thử 100.000 câu bình luận đã chuẩn bị dựa trên [5].

Cụ thể kế hoạch training như sau:

- Dữ liệu đầu vào: 100.000 câu bình luận lấy từ các bài viết trên facebook. Lưu dưới dạng các file .xlsx.
- Ngôn ngữ lập trình: php.
- Hệ thống cơ sở dữ liệu sử dụng: my Sql
- Công cụ training: 1 chương trình sử dụng ngôn ngữ lập trình php sẽ đọc các câu comment từ file đầu vào. Sau đó chạy qua chương trình chatbot đã chuẩn bị sẵn để tiến hành lấy dữ liệu output và lưu lại.
- Thông số máy tính để chạy chương trình training: CPU: Intel Core i7. Memory: 8192 Ram. System Model: Inspiron 3543.
- Dữ liệu đầu ra: 5 file ứng với các cấp bậc phản cảm. Chương trình sẽ phân chia câu bình luận thành cấp bậc phản cảm và lưu vào từng file tương ứng

Do không thể sàng lọc hết 100.000 câu này nên chúng tôi sẽ tiến hành đánh giá dựa theo kết quả thu được của từng mức độ. Kết quả như trên bảng 2.

Cụ thể kết quả chi tiết với từng cấp độ phản cảm như bảng 3:

Bảng 3. Thống kê kết quả thu được với từng cấp bậc nhạy cảm

Cấp bậc phản cảm	Số lượng câu comment	Tỷ lệ chính xác
1	4254	63.3%
2	2627	77.54%
3	1574	80.74%
4	1021	94%
5	735	100%

Câu này có từ bà là đại từ nhân xưng và từ chó là động vật=> Cấp độ 1. Các trường hợp kết quả sai thường gắn với 1 số loại từ như sau:

- Các câu có nhiều đại từ nhân xưng như bà, bố, mẹ..vv..
- Các câu liên quan đến động vật.
- Các câu chỉ các bộ phận trên cơ thể.

Còn những câu có mức độ phản cảm cao thì tỷ lệ chính xác rất cao vì nó đều có những từ mang ý nghĩa phản cảm hay nói tục. Nhưng số lượng phát hiện quá ít. Rất nhiều từ để dưới dạng viết tắt bị bỏ qua và không thể phát hiện.

Các trường hợp sai đối với những câu có mức phản cảm cao thường liên quan đến:

- Các từ xuồng xã như: vãi, dù...
- Các câu liên quan đến bệnh tật, nguyền rủa
- Những câu phủ định

Ngoài ra do chương trình cần lặp đi lặp lại xử lý câu nên với những câu dài, tốc độ xử lý là quá chậm. Với 100.000 câu bình luận. Thời gian tối đa để xử lý là khoảng 83 tiếng. Tương ứng với khoảng gần 3.5 ngày. Tuy nhiên, thời gian training thực tế lên đến con số gần 5 ngày.

Không chỉ thế, thuật toán chatbot này mới chỉ hoàn thành vấn đề phát hiện bình luận và đưa ra những hình phạt cũng như nhắc nhở. Đây mới chỉ là tương tác một chiều giữa chatbot đến người dùng. Cần phát triển thêm các tương tác mà người dùng có thể sử dụng chatbot như: khen, hỏi, chào, yêu cầu 1 số tác vụ như: hỏi thời gian, thống kê các câu bình luận không phù hợp trong tháng...

Do đó thuật toán này vẫn cần cập nhật để có thể áp dụng vào trong thực tế. Để có thể hướng tới điều này, cần phải làm được những công việc sau:

- Thiết lập lại bảng phân chia mức độ bình luận sao cho chặt chẽ hơn nữa. Có thể bao quát càng nhiều trường hợp cũng như có thể loại bỏ những trường hợp không chính xác như đã đề cập ở trên.

- Tối ưu hóa code, giảm thời gian xử lý xuống mức phù hợp. Tối đa xử lý cho 1 câu comment là 30 giây.
- Không ngừng cập nhật từ điển để bước xử lý các từ không có nghĩa có thể có độ chính xác cao nhất.
- Tìm cách ứng dụng machine learning và AI vào chatbot [7]÷[9] để chương trình trở nên thông minh hơn. Có thể dựa vào hoàn cảnh để đưa ra những nhận định chính xác nhất.
- Phát triển tương tác hai chiều giữa người dùng và chatbot.

Dựa vào những kinh nghiệm đã nhận được, chúng tôi có thể phát triển nó thành một phần mềm có tính thực dụng cao và có thể áp dụng vào trong thực tế, góp phần nho nhỏ vào công cuộc làm sạch thế giới mạng đang tràn đầy những lời ác ý như hiện nay.

IV. KẾT LUẬN

Những câu bình luận mang ý nghĩa phản cảm, không phù hợp tràn ngập trên mạng. Ở bất cứ trang mạng xã hội nào dù lớn dù nhỏ ta có thể dễ dàng thấy những câu đó ở một nơi nào đó. Càng là những vấn đề nóng hổi thì những câu bình luận phản cảm càng nhiều và càng nặng. Thậm chí các câu bình luận còn vượt qua biên giới mà xuất hiện ở các cộng đồng nước ngoài, làm xấu hình ảnh chúng ta trong mắt bạn bè quốc tế.

Vì hình ảnh của đất nước, chúng ta cần cấm hoặc chí ít là hạn chế những bình luận không phù hợp thuần phong mĩ tục tràn lan trên mạng. Hiện nay, một số game nổi tiếng đều có một số cách để che đi những câu bình luận không phù hợp. Nhưng vấn đề này dường như vẫn chưa xử lý triệt để. Các chương trình đó đại đa số chỉ là dựa vào các từ ngữ cụ thể để xác định. Do vậy tác dụng nó đem lại quả thật không lớn.

Đó chính là lý do chúng tôi đề xuất chương trình chatbot quản lý bình luận này. Trong đây, chúng tôi đã xử lý được một số cách viết lái các câu bình luận không phù hợp bao gồm:

- Dùng từ đồng nghĩa trái nghĩa,
- Tách từ bằng dấu cách hay các dấu câu,
- Dùng các từ thay thế.

Ngoài ra, chúng tôi định ra môt quy chuẩn để có thể xác định mức độ phản cảm của một câu bình luận để từ đó có thể đưa ra các biện pháp xử lý tối ưu.Trong tương lai chúng tôi sẽ tích hợp thêm các thuật toán mới dựa trên nền tảng trí tuệ nhân tạo để xử lý triệt để hơn [10] ÷[13].

TÀI LIỆU THAM KHẢO

[1] S. Phillips, “A brief history of facebook,” The Guardian, 01 2007.

[2] M. Zuckerberg, Facebook, 2020 (accessed Dec. 11, 2020.). [Online]. Available: https://www.facebook.com/

[3] M. Mauldin, Chatbot, 2020 (accessed Dec. 11, 2020.). [Online]. Available: https://en.wikipedia.org/wiki/Chatbot

[4] C.-T. Nguyen, T.-K. Nguyen, X.-H. Phan, L.-M. Nguyen, and Q.-T. Ha, “Vietnamese word segmentation with CRFs and SVMs: An investigation,” in Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation. Huazhong Normal University, Wuhan, China: Tsinghua University Press, Nov. 2006, pp. 215–222. [Online]. Available: https://www.aclweb.org/anthology/Y06-1028

[5] M. Burtsev, A. Seliverstov, R. Airapetyan, M. Arkhipov, D. Baymurzina, N. Bushkov, O. Gureenkova, T. Khakhulin, Y. Kuratov, D. Kuznetsov, A. Litinsky, V. Logacheva, A. Lymar, V. Malykh, M. Petrov, V. Polulyakh, L. Pugachev, A. Sorokin, M. Vikhreva, and M. Zaynutdinov, “Deeppavlov: Open-source library for dialogue systems,” 07 2018.

[6] L.-H. Phuong, H. Nguyen, A. Roussanaly, and T. Ho, “A hybrid approach to word segmentation of vietnamese texts,” in Language and Automata Theory and Applications. LATA 2008. Lecture Notes in Computer Science, vol. 5196, 12 2013, pp. 240–249.

[7] T. Kl¨uwer, From Chatbots to Dialogue Systems, 07 2011, pp. 1–22.

[8] Y.-N. Chen, C. Asli, and D. Hakkani-Tur, “Deep learning for dialogue systems,” 01 2017, pp. 8–14.

[9] K. van Deemter, E. Krahmer, and M. Theune, “Plan- based vs. template-based nlg: a false opposition?” 08, 1999.

[10] N. N. Khin and K. M. Soe, “University chatbot using artificial intelligence markup language,” in 2020 IEEE Conference on Computer Applications(ICCA), 2020, pp. 1–5.

[11] J. Bozic, O. A. Tazl, and F. Wotawa, “Chatbot testing using ai planning,” in 2019 IEEE International Conference On Artificial Intelligence Testing (AITest), 2019, pp. 37–44.

[12] N. Albayrak, A. ¨Ozdemir, and E. Zeydan, “An overview of artificial intelligence based chatbots and an example chatbot application,” in 2018 26th Signal Processing and Communications Applications Conference (SIU), 2018, pp. 1–4.

[13] S. J. du Preez, M. Lall, and S. Sinha, “An intelligent web-based voice chat bot,” in IEEE EUROCON 2009, 2009, pp. 386–391.

PROPOSING ALGORITHM BASED ON CHATBOT TO DETECT SENSITIVE COMMENTS

Abstract: Today, the cultural behavior is an issue of social concern. State and government have many policies to solve the problem in order to clean up the network environment. However, there are still many comments with offensive content on social networking sites and online games. Therefore, we propose an algorithm to detect sensitive comments in the paper. The chatbot-based algorithm automatically detects and warns unhealthy content as well as inappropriate comments. The results show that the algorithm achieves 75% accuracy with 100,000 comments that is applicable in practice.

Keywords: Chatbot, offensive comments, behavioral culture, online, data processing.

Nguyễn Hữu Phát, nhận bằng kỹ sư2003), thạc sỹ (2005) ngành Điện tử và Viễn thông tại Đại học Bách Khoa Hà Nội (HUST), Việt Nam và bằng tiến sĩ (2012) về Khoa học Máy tính tại Viện Công nghệ Shibaura, Nhật Bản. Hiện tại, đang là giảng viên tại Viện Điện tử Viễn thông, HUST, Việt Nam. Các nghiên cứu gồm xử lý hình ảnh và video, mạng không dây, big data, hệ thống giao thông thông minh (ITS), và internet của vạn vật (IoT). Ông đã nhận được giải thưởng bài báo hội nghị tốt nhất trong

SoftCOM (2011), giải thưởng tài trợ sinh viên tốt nhất trong APNOMS (2011), giải thưởng danh dự của Viện Công nghệ Shibaura (SIT).

Đỗ Mạnh Cầm, hiện tại là sinh viên Viện Điện tử Viễn thông, Trường Đại Học Bách Khoa Hà Nội. Hướng nghiên cứu gồm xử lý ngôn ngữ và các ứng dụng thông minh.

Hoàng Văn Quang, hiện tại là cán bộ phòng Tổ chức, Trường Đại Học Bách Khoa Hà Nội. Hướng nghiên cứu quan tâm gồm quan trắc môi trường, xử lý ngôn ngữ, và các ứng dụng thông minh.