TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

ĐẠI HỌC QUỐC GIA TP.HCM

TÊN ĐỒ ÁN

Môn học: Python cho Khoa học Dữ liệu

Giảng viên hướng dẫn: ThS. Hà Minh Tuấn

Nhóm sinh viên thực hiện:

Nguyễn Thị Ngọc Anh	23280037
Trương Thị Quỳnh Giang	23280052
Trần Trung Kiên	23280066

TP.HCM, 2025

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

1. Mục tiêu và nội dung của đồ án

2. Kết quả đạt được

3. Ý thức làm việc của sinh viên:

TP. Hồ Chí Minh, ngày ... tháng ... năm 2025

Giảng viên hướng dẫn

ThS. Hà Minh Tuấn

LỜI CẢM ƠN

Nội dung

Mục lục

Bảng ký hiệu và chữ viết tắt

Danh sách bảng

Danh sách hình

Tóm tắt

Nhận diện và phân loại phát ngôn là một trong những bài toán quan trọng của xử lý ngôn ngữ tự nhiên, nơi hệ thống cần xác định đặc trưng và sắc thái biểu đạt của các phát ngôn được tạo ra trong môi trường trực tuyến. Các phát ngôn có thể mang nhiều dạng nội dung khác nhau như tích cực, trung tính hoặc tiêu cực, và thường đa dạng và phức tạp. Điều này khiến bài toán phân loại trở nên khó xử lý, đòi hỏi mô hình phải nắm bắt được ý nghĩa, mục đích và đặc điểm ngôn ngữ ẩn sau từng câu.

Trong lĩnh vực xử lý ngôn ngữ tự nhiên, các phương pháp học máy truyền thống giữ vai trò quan trọng nhờ tính ổn định và khả năng khái quát hóa tốt trong nhiều bài toán phân loại văn bản. Các mô hình này dựa trên việc học các đặc trưng ngôn ngữ được trích xuất từ dữ liệu, sau đó sử dụng những đặc trưng đó để suy luận và phân tách các nhóm phát ngôn khác nhau. Cách tiếp cận này cho phép mô hình nắm bắt được những tín hiệu ngữ nghĩa cốt lõi trong văn bản, hỗ trợ việc nhận diện nội dung và sắc thái biểu đạt một cách nhất quán.

Trong đồ án này, ba thuật toán phân loại phổ biến gồm Support Vector Machine (SVM), Naive Bayes và Logistic Regression được triển khai nhằm khảo sát khả năng nhận biết các kiểu tính chất thường gặp trong văn bản tiếng Việt trên không gian số. Các mô hình này đại diện cho những hướng tiếp cận phổ biến trong bài toán phân loại văn bản, nhờ vào khả năng học các đặc trưng ngôn ngữ quan trọng và đưa ra dự đoán ổn định trong quá trình phân tích.

Giới thiệu

Phân loại phát ngôn đóng vai trò quan trọng trong việc phân tích nội dung trao đổi của người dùng trên mạng xã hội, nơi ngôn ngữ thường được thể hiện qua nhiều sắc thái và biến thể khác nhau. Một nhiệm vụ trọng yếu trong phân loại phát ngôn là nhận diện loại nội dung mà một phát ngôn thể hiện ở các mức độ khác nhau, từ toàn văn bản đến từng câu hay từng kiểu biểu đạt - nhằm xác định liệu thông điệp đó mang sắc thái tích cực, tiêu cực, trung tính hay thuộc các hình thức diễn đạt đặc thù như mỉa mai hay công kích. Nhờ đó, hệ thống có thể hỗ trợ doanh nghiệp nắm bắt phản hồi của khách hàng, theo dõi xu hướng đánh giá về sản phẩm và nhận diện sớm các chủ đề tiêu cực liên quan đến thương hiệu. Đối với các nhà nghiên cứu, việc phân loại phát ngôn còn giúp phân tích thái độ xã hội và hành vi giao tiếp trong các bối cảnh trực tuyến.

Kết quả thu được từ đồ án tạo nền tảng cho việc phát triển các hệ thống phân tích chuyên sâu hơn trong tương lai. Trên cơ sở đó, các ứng dụng như phát hiện và sàng lọc nội dung độc hại, nhận diện khuynh hướng cảm xúc hay phân tích xu hướng thảo luận trên mạng xã hội có thể được mở rộng và hoàn thiện.

Đặt vấn đề

Sự phát triển mạnh mẽ của Internet và các nền tảng mạng xã hội đã làm gia tăng đáng kể khối lượng phát ngôn được tạo ra mỗi ngày, tạo nên nhu cầu lớn đối với việc thu thập và phân tích nội dung người dùng. Các mạng xã hội, diễn đàn thảo luận hay hệ thống bình luận trực tuyến liên tục xuất hiện những ý kiến thể hiện quan điểm, thái độ và cảm xúc của người dùng về nhiều chủ đề khác nhau.

Vai trò then chốt của việc phân loại phát ngôn:

- Đối với người dùng: Phân loại phát ngôn giúp họ nhanh chóng nhận diện xu hướng thảo luận, tiếp cận các ý kiến trước đó và đưa ra đánh giá chính xác hơn về một vấn đề, sự kiện hay sản phẩm. Người dùng có thể xem tổng quan ý kiến tích cực – tiêu cực của cộng đồng để hỗ trợ việc ra quyết định.

- Đối với tổ chức, doanh nghiệp: Các phát ngôn trên mạng xã hội là nguồn dữ liệu quan trọng giúp doanh nghiệp lắng nghe phản hồi, phát hiện các chủ đề nhạy cảm, theo dõi danh tiếng thương hiệu và điều chỉnh chiến lược truyền thông. Nhờ nắm bắt kịp thời thái độ người dùng, các tổ chức có thể nâng cao chất lượng dịch vụ, tối ưu hoạt động quản lý cộng đồng và hạn chế các xu hướng tiêu cực lan rộng.

Tuy nhiên, việc xử lý và phân loại phát ngôn trực tuyến gặp phải nhiều thách thức:

- Khối lượng dữ liệu lớn: Số lượng phát ngôn phát sinh liên tục từ nhiều nền tảng khiến việc phân tích thủ công trở nên tốn kém và thiếu hiệu quả.

- Ngôn ngữ phức tạp và đa dạng: Phát ngôn trên mạng xã hội thường mang tính chủ quan, đa dạng về cách diễn đạt và sử dụng ngôn ngữ, gây khó khăn cho việc phân tích tự động

Do đó, việc xây dựng các hệ thống phân loại phát ngôn tự động trở nên cần thiết nhằm hỗ trợ nhận diện nhanh chóng nội dung và xu hướng trao đổi trên môi trường trực tuyến. Trước khi người dùng tiếp cận một chủ đề hay doanh nghiệp tiến hành phân tích phản hồi, những hệ thống này có thể cung cấp góc nhìn tổng quát về thái độ và mức độ quan tâm của cộng đồng. Bằng cách xác định đặc trưng của từng nhóm phát ngôn, các mô hình phân loại giúp khai thác hiệu quả thông tin ngôn ngữ do người dùng tạo ra, từ đó phục vụ tốt hơn cho các hoạt động nghiên cứu và ứng dụng trong lĩnh vực phân tích dữ liệu xã hội.

Lý do chọn đề tài

Ngày nay, cùng với sự phát triển mạnh mẽ của Internet và các nền tảng mạng xã hội, lượng thông tin và phát ngôn do người dùng tạo ra ngày càng trở nên phong phú và đa dạng. Trên các trang mạng xã hội, diễn đàn trực tuyến hay các nền tảng thảo luận công cộng, người dùng liên tục bày tỏ quan điểm, cảm xúc và thái độ của mình về nhiều vấn đề trong đời sống. Những phát ngôn này không chỉ phản ánh nhận thức và trải nghiệm cá nhân mà còn góp phần hình thành các xu hướng, ảnh hưởng đến hành vi của cộng đồng cũng như tác động trực tiếp đến doanh nghiệp và tổ chức.

Tuy nhiên, sự gia tăng nhanh chóng của khối lượng dữ liệu cùng đặc điểm ngôn ngữ thiếu chuẩn hóa trên mạng xã hội - như teencode, emoji, ký tự viết tắt hoặc các lối diễn đạt mang tính cảm xúc - khiến việc phân tích thủ công trở nên không khả thi. Trong nhiều trường hợp, các phát ngôn tiêu cực hoặc nội dung gây tranh cãi có thể lan truyền nhanh chóng, gây ảnh hưởng đến danh tiếng thương hiệu, tạo áp lực cho công tác quản lý cộng đồng và đặt ra yêu cầu cấp thiết cho các hệ thống xử lý tự động. Do đó, nhu cầu xây dựng các mô hình phân loại phát ngôn nhằm hỗ trợ nhận diện nhanh chóng nội dung, theo dõi xu hướng thảo luận và phát hiện những biểu hiện bất thường trở nên vô cùng quan trọng.

Xuất phát từ thực tiễn đó, việc nghiên cứu và xây dựng một hệ thống phân loại phát ngôn trên mạng xã hội tiếng Việt không chỉ có ý nghĩa về mặt khoa học mà còn mang tính ứng dụng cao. Hệ thống này giúp doanh nghiệp nắm bắt phản hồi khách hàng, hỗ trợ nhà quản lý giám sát tương tác trực tuyến, đồng thời là công cụ hữu ích cho các nhà nghiên cứu trong việc phân tích hành vi ngôn ngữ của cộng đồng. Với mong muốn góp phần giải quyết những thách thức nêu trên, nhóm chúng em quyết định lựa chọn đề tài "Nhận diện và phân loại phát ngôn trên mạng xã hội" làm hướng nghiên cứu và triển khai trong đồ án này.

Phát biểu bài toán

Trong nghiên cứu này, mục tiêu chính là xây dựng và đánh giá mô hình có khả năng nhận diện và phân loại các phát ngôn được tạo ra trên mạng xã hội. Bài toán hướng đến việc xác định nhóm nội dung hoặc sắc thái biểu đạt của từng phát ngôn, với các lớp thường được quan tâm như tích cực, tiêu cực và trung tính.

Dữ liệu đầu vào của bài toán gồm gần 7.000 phát ngôn tiếng Việt được thu thập từ mạng xã hội, đã được gán nhãn cảm xúc và lưu trữ dưới dạng văn bản thô sau khi loại bỏ các yếu tố nhiễu cơ bản.

Trong đồ án này, chúng em sử dụng ba mô hình học máy truyền thống gồm Support Vector Machine (SVM), Naive Bayes và Logistic Regression để giải quyết bài toán.

Cấu trúc đồ án

PHẦN 1: CƠ SỞ LÝ THUYẾT

Trình bày các khái niệm liên quan đến bài toán xử lý ngôn ngữ tự nhiên, những đặc điểm của ngôn ngữ trên mạng xã hội, cùng các phương pháp phân loại văn bản cơ bản thường được sử dụng trong NLP.

PHẦN 2: THU THẬP, XỬ LÝ VÀ KHÁM PHÁ DỮ LIỆU

Mô tả quy trình thu thập dữ liệu phát ngôn từ mạng xã hội, các bước tiền xử lý văn bản như làm sạch, chuẩn hóa, xử lý emoji, teencode và stopwords, đồng thời thực hiện phân tích khám phá dữ liệu (EDA) nhằm hiểu rõ các đặc trưng của tập dữ liệu.

PHẦN 3: XÂY DỰNG MÔ HÌNH PHÂN TÍCH

Trình bày các mô hình học máy được sử dụng trong nghiên cứu gồm Support Vector Machine (SVM), Naive Bayes và Logistic Regression; mô tả cách trích xuất đặc trưng, thiết lập mô hình và quy trình huấn luyện.

PHẦN 4: KẾT QUẢ THỰC NGHIỆM

Tiến hành thử nghiệm, trình bày và so sánh kết quả mô hình dựa trên các thước đo đánh giá, trực quan hóa kết quả và đưa ra nhận xét về hiệu quả phân loại phát ngôn.

Chương 1 CƠ SỞ LÝ THUYẾT

1.1 Các phương pháp vectorize dữ liệu

1.1.1 One Hot

1.1.2 TF IDF

1.1.3 PhoBert

1.1.4 Một số phương pháp khác

1.2 Các mô hình học máy

1.2.1 Logistic Regression

1.2.2 Support Vector Machine

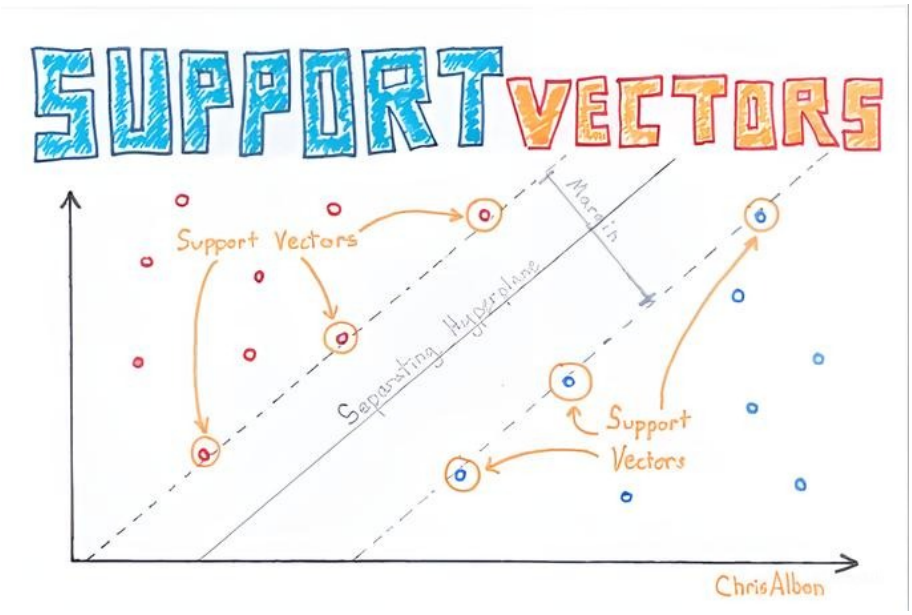
Support Vector Machine (SVM) là mô hình học máy giám sát, mục tiêu của thuật toán là tìm ra một siêu phẳng (hyperplane) sao cho có thể phân tách tối ưu các điểm dữ liệu thuộc các lớp khác nhau. "Tối ưu" ở đây nghĩa là tìm ra siêu phẳng tạo ra khoảng cách lớn nhất (margin) giữa các lớp dữ liệu. Các điểm dữ liệu có khoảng cách nhỏ nhất đến siêu mặt phẳng (các điểm gần nhất) được gọi là các vector hỗ trợ (support vectors) Margin là độ rộng tối đa của dải không chứa bất kỳ điểm dữ liệu nào và song song với siêu phẳng.

Việc tính toán phân tách dữ liệu phụ thuộc vào hàm kernel. Có nhiều hàm kernel khác nhau: Tuyến tính (linear kernel), đa thức (polynomial kernel), gaussian, RBF (radial basis function), sigmoid kernel. Các hàm này xác định độ mượt và hiệu quả trong việc phân tách lớp, và việc tùy biến các siêu tham số của chúng có thể dẫn đến hiện tượng quá khớp (overfitting) hoặc thiếu khớp (underfitting).

Trong khuôn khổ báo cáo này, nhóm không trình bày chi tiết các biểu thức toán học của từng loại kernel, bạn có thể tham khảo công thức đầy đủ ở tài liệu sau: [1] V. Cortes, C. & Vapnik, "Support-vector networks," Machine Learning, 1995.

SVM không chỉ hỗ trợ phân loại nhị phân và tách các điểm dữ liệu thành hai lớp, SVM còn mở rộng để hỗ trợ các bài toán phân loại đa lớp thông qua cơ chế chia nhỏ bài toán đa lớp thành nhiều mô hình nhị phân.

a) Phương pháp OvO (One-to-One)



Hình 1.1: Support Vector Machine

Phương pháp OvO xây dựng bộ phân loại nhị phân cho mỗi cặp lớp trong tập dữ liệu. Với tập dữ liệu gồm K lớp, số lượng mô hình được tạo ra là *công thức toán học* [K(K-1)]/2. Mỗi mô hình được huấn luyện để phân biệt hai lớp cụ thể. Khi dự đoán, mỗi mô hình đưa ra một phiếu bầu cho một trong hai lớp, và lớp nhận được nhiều phiếu nhất sẽ được chọn làm kết quả cuối cùng. Phương pháp OvO thường hiệu quả khi số lượng lớp không quá lớn và mỗi mô hình nhị phân tương đối nhẹ.

b) Phương pháp OvR (One-to-Rest)

Phương pháp OvR (OvA) chia dữ liệu đa lớp thành nhiều bài toán phân loại nhị phân, sau đó mỗi bộ phân loại nhị phân được huấn luyện trên mỗi bài toán phân loại nhị phân và đưa ra dự đoán bằng cách sử dụng mô hình có độ tin cậy cao nhất. Phương pháp này yêu cầu tạo ra một mô hình cho mỗi lớp.

1.2.3 Naive Bayes

Chương 2 THU THẬP VÀ KHÁM PHÁ DỮ LIỆU

2.1 Thu thập dữ liệu

2.2 Khám phá bộ dữ liệu

Chương 3 XÂY DỰNG MÔ HÌNH PHÂN TÍCH

3.1 Module Loader

3.2 Module Preprocessor

3.3 Module feature

3.4 Tối ưu tham số

3.4.1 Logistic Regression

3.4.2 Support Vector Machine

3.4.3 Naive Bayes

3.5 Huấn luyện mô hình

3.5.1 Logistic Regression

3.5.2 Support Vector Machine

3.5.3 Naive Bayes

Chương 4 KẾT QUẢ THỰC NGHIỆM

4.1 Đánh giá mô hình

4.2 So sánh mô hình

4.3 Ứng dụng thực tế

Tài liệu tham khảo

[1] V. Cortes, C. & Vapnik, "Support-vector networks," *Machine Learning*, 1995.