|  |
| --- |
| **TRƯỜNG ĐẠI HỌC THỦY LỢI**  **ĐỀ TÀI:**  **Sinh viên thực hiện:**  **Lớp:**  **Khoa:**  ***Hà Nội,*** *……* |

**NGHIÊN CỨU MÔ HÌNH ……**

*(Tên đề tài font chữ thường in hoa, đậm, căn giữa)*

|  |  |
| --- | --- |
| *SVTH:* | *Đàm Khôi Nguyên – 60PM1* |
| *GVHD:* | *Nguyễn Mạnh Hiển* |

*(Phần này nên kẻ bảng như mẫu hướng dẫn để thuận lợi cho việc format chung)*

1. Mở đầu:

Tổng quan tình hình nghiên cứu thuộc lĩnh vực đề tài:

đa

lý do chọn đề tài:

mục tiêu đề tài:

phương pháp nghiên cứu:

đối tượng và phạm vi nghiên cứu:

1. Nội dung và kết quả nghiên cứu
   1. Nội dung:

Chương 1: Giới thiệu

Bộ phân loại văn bản bán giám sát đã đạt được thành công đáng kể trong vài năm qua do khả năng tổng quát hóa của mạng nơ-ron cao. Mặc dù các bộ phân loại hiện đại thường dựa vào các tập huấn luyện lớn, việc giới thiệu các phép nhúng từ ngữ theo ngữ cảnh và đào tạo trước mô hình ngôn ngữ [18, 41, 43] đã làm giảm đáng kể nhu cầu chú thích dữ liệu thủ công. Tuy nhiên, các mô hình thần kinh hiện đại vẫn có xu hướng bị quá tải, đặc biệt là ở những khu vực có mô hình ngôn ngữ chuyên biệt và thưa thớt. Các lĩnh vực này bao gồm, nhưng không giới hạn ở: lĩnh vực pháp lý [26], lĩnh vực y tế [34] và lĩnh vực truyền thông xã hội [3].

Tùy thuộc vào nhiệm vụ hiện tại, một giải pháp để giải quyết vấn đề này là tự động xây dựng một tập dữ liệu lớn – và có lẽ ồn ào [20], tuy nhiên, điều này không phải lúc nào cũng khả thi [59]. Một cách tiếp cận phương pháp hơn là sử dụng các kỹ thuật cải thiện khả năng tổng quát hóa. Các kỹ thuật này bao gồm khai thác các phép nhúng từ thần kinh [29], tăng dữ liệu [6], và thích ứng miền [2]. Khai thác dữ liệu không được gắn nhãn [9, 58] cũng là một cách tiếp cận bổ sung. Trong nghiên cứu này, chúng tôi bổ sung vào nội dung tài liệu về học bán giám sát bằng cách sử dụng các thuộc tính của mạng nơ-ron và đề xuất một cách mới để sử dụng dữ liệu không được gắn nhãn. Chúng tôi tập trung vào một trong những lĩnh vực được báo cáo là thiếu đủ dữ liệu đào tạo, tức là khai thác mạng xã hội. Ngoài việc thiếu dữ liệu đào tạo, tiến trình trong lĩnh vực này còn bị cản trở bởi độ dài tài liệu ngắn, mô hình ngôn ngữ không chính thức và sự lựa chọn từ vựng điển hình là mơ hồ. Những phẩm chất này làm cho các nhiệm vụ trên mạng xã hội trở thành một giường thử nghiệm phù hợp để đánh giá các thuật toán học bán giám sát.

Thuật toán của chúng tôi, được gọi là Tự tiền thử nghiệm, được lấy cảm hứng từ mô hình tự đào tạo [58]. Tương tự như tự đào tạo, thuật toán của chúng tôi là lặp đi lặp lại và trong mỗi lần lặp lại chọn một tập hợp các tài liệu chưa được gắn nhãn để gắn nhãn. Tuy nhiên, trái ngược với tự đào tạo, thuật toán của chúng tôi là không có ngưỡng. Do đó, nó không xếp hạng các tài liệu không được gắn nhãn dựa trên những dự đoán của họ. Điều này làm cho thuật toán của chúng tôi đặc biệt phù hợp với các mô hình mạng nơ-ron do các đầu ra được hiệu chỉnh kém của chúng [35]. Ngoài ra, thuật toán của chúng tôi có thể đối phó với vấn đề trôi dạt ngữ nghĩa [12]. Đó là, nó có khả năng chống lại tiếng ồn trongnhãn giảkhi số lần lặp lại tăng lên và tỷ lệ lỗi của trình phân loại bên dưới tăng lên. Hơn nữa, Tự tiền đào tạo có khả năng sửa đổi nhãn của các tài liệu đã được dán nhãn trước đó. Để đạt được những điều này, mô hình của chúng tôi sử dụng quy trình chưng cất lặp đi lặp lại, tức là, trong mỗi lần lặp, thông tin thu được trong các lần lặp trước đó được chắt lọc vào bộ phân loại. Nó chuyển một giả thuyết qua các lần lặp lại và sử dụng mô hình học hai giai đoạn, trong đó tập hợp các nhãn giả được sử dụng để khởi tạo trình phân loại và tập hợp các tài liệu có nhãn được sử dụng để tinh chỉnh trình phân loại. Ngoài ra, Self-Pretraining điều chỉnh một lịch trình tốc độ học tập mới để tích hợp hiệu quả hai bộ ví dụ đào tạo ồn ào và không gây tiếng ồn. Cuối cùng, để giảm thiểu tác động của các nhãn giả ồn ào trong mỗi lần lặp lại,

Các thử nghiệm của chúng tôi trong ba tập dữ liệu Twitter công khai cho thấy rằng Tự thử nghiệm vượt trội so với hiện đại trong nhiều cài đặt nơi chỉ có vài trăm tài liệu được gắn nhãn sẵn có. Điều này rất quan trọng, xét rằng trình phân loại cơ bản của thuật toán của chúng tôi và tất cả các mô hình cơ sở là BERT [18] đã sử dụng đào tạo trước mô hình ngôn ngữ, và do đó, thực hiện bất kỳ cải tiến nào so với đường cơ sở là rất khó khăn. Chúng tôi cũng thực hiện một bộ thí nghiệm toàn diện để hiểu rõ hơn về các phẩm chất của Tự thử nghiệm. Đặc biệt, chúng tôi chứng minh sự mạnh mẽ của mô hình của chúng tôi đối với tiếng ồn trong các nhãn giả. Những đóng góp trong nghiên cứu của chúng tôi như sau:1)Chúng tôi đề xuất một khung học tập bán giám sát mới có tên là Tự tiền đào tạo. Mô hình của chúng tôi dựa trên mô hình tự đào tạo, tuy nhiên, nó không có ngưỡng, nó có thể đối phó với vấn đề trôi dạt ngữ nghĩa và cũng có thể sửa đổi các tài liệu đã được gắn nhãn trước đó. Theo hiểu biết của chúng tôi, Tự tiền đào tạo là mô hình đầu tiên giải quyết những nhược điểm này trong một khuôn khổ thống nhất.2)Chúng tôi đề xuất một lịch trình tốc độ học tập mới để tích hợp hiệu quả quy trình tối ưu hóa với quy trình học tập bán giám sát hai giai đoạn của chúng tôi.3)Để giảm thiểu hơn nữa vấn đề trôi dạt ngữ nghĩa, chúng tôi lập mô hình phân phối lớp của các nhãn giả như một quá trình ngẫu nhiên qua các lần lặp khởi động và đề xuất một cách tiếp cận mới để chuyển đổi các phân phối lớp.4)Chúng tôi thực hiện một bộ thử nghiệm toàn diện trên ba tập dữ liệu Twitter công khai và chứng minh rằng mô hình của chúng tôi hoạt động tốt hơn một số đường cơ sở hiện đại nhất trong nhiều cài đặt.

Nghiên cứu của chúng tôi rõ ràng đã thúc đẩy sự tiên tiến trong phân loại văn bản bán giám sát. Chúng tôi tin rằng các ý tưởng được trình bày trong bài báo của chúng tôi có thể được áp dụng cho các lĩnh vực khác, ví dụ: phân loại hình ảnh. Công việc trong tương lai có thể khám phá hướng này. Trong phần tiếp theo, chúng tôi cung cấp một cái nhìn tổng quan về các nghiên cứu liên quan và làm nổi bật các phẩm chất của Tự tiền đào tạo.

Chương 2: Thông tin liên quan

Dữ liệu chưa được gắn nhãn trong học tập bán giám sát.Dữ liệu không được gắn nhãn có thể được khai thác theo nhiều cách. Nó có thể được sử dụng như một siêu nguồn thông tin [21], nó có thể được sử dụng như một bộ điều chỉnh [56], hoặc nó có thể được sử dụng trong cài đặt điều chỉnh miền để tạo mối tương quan giữa nguồn và dữ liệu đích [51]. Mối quan tâm gần đây hơn đối với văn học làtự giám sát, trong đó một tác vụ khép kín được xác định sao cho không cần chú thích thủ công. Ví dụ của các nhiệm vụ như vậy là đào tạo trước mô hình ngôn ngữ [18, 41] trong NLP, và học tương phản trong xử lý hình ảnh [15, 46]. Từ một góc độ khác, các nghiên cứu tự giám sát có thể được phân loại thành các phương pháp tiếp cận theo nhiệm vụ [10] và theo nhiệm vụ cụ thể [22]. Điều này đã làm nảy sinh khái niệm về mô hình “prerain, sau đó là finetune”. Chúng tôi tích hợp mô hình này vào thuật toán tự đào tạo.

Khởi động trong học tập bán giám sát.Tự đào tạo là phương pháp học tập bán giám sát lâu đời nhất [14] ra đời từ năm 1965 [48]. Ý tưởng này tái xuất hiện trong công việc chính của Yarowsky [58] cho các nhiệm vụ NLP vào năm 1995, và một lần nữa trong cộng đồng thị giác máy tính vào năm 2013 nhưdán nhãn giả[33]. Thuật toán này là một trình bao bọc liên tục sử dụng một thuật toán được giám sát làm mô hình cơ bản. Có nhiều giả định trong đó việc tự đào tạo – và nói chung là học tập bán giám sát – được kỳ vọng sẽ hoạt động tốt. Ví dụ,giả định êm áiđiều đó cho biết nếu hai điểm dữ liệu 1và 2 gần, sau đó là dự đoán của họ 1

và 2cũng nên gần - giả định này là cơ sở của các thuật toán như MixUp [60] và MixMatch [8]. Như chúng ta thảo luận trong phần tiếp theo, một khía cạnh không thỏa mãn của việc tự đào tạo là nó dựa vào các thuộc tính của mô hình dự đoán cơ bản, ví dụ, các phân phối đầu ra của mô hình. Đã có những nỗ lực để giải quyết nhược điểm này. Ví dụ, điều chỉnh [1] có thể được sử dụng để giảm bớt ảnh hưởng của các ứng viên ồn ào, hoặc trong bối cảnh học tập chuyển đổi, mật độ của các điểm dữ liệu không được gắn nhãn có thể được kết hợp để giảm thiểu vấn đề này [49].

Trong vài năm qua, các nghiên cứu đã khám phá hiệu quả của mạng nơ-ron như một mô hình dự đoán cơ bản trong quá trình tự đào tạo. Một biến thể mạng nơron của đồng đào tạo [9] được đề xuất trong [42]. Trong [30], các tác giả đề xuất một khuôn khổ để tích hợp tri thức nhân loại với đồng đào tạo. Trong [55], một biến thể học tăng cường của đồng đào tạo được đề xuất. Trong [44], một biến thể mạng nơ-ron của

đào tạo tri với bất đồng [50] được trình bày, và cho thấy rằng sự kết hợp là một cơ sở mạnh mẽ đáng ngạc nhiên trong bối cảnh thích ứng miền. Các tác giả trong [13] đề xuất sử dụng điểm phân vị thay vì điểm tin cậy để chọn các nhãn giả tốt nhất; và các tác giả trong [39] sử dụng mạng nơ-ron Bayes để chọn các nhãn giả ít tin cậy nhất trong mỗi lần lặp lại. Trong [4], một chiến lược lấy mẫu tài liệu mới để tự đào tạo được đề xuất. Mô hình, ngoài độ tin cậy của trình phân loại, sử dụng các kỷ nguyên đào tạo trong đó các tài liệu không được gắn nhãn được gắn nhãn gần đúng. Trong [5], các tác giả đề xuất tích hợp MixUp [60] với việc lấy mẫu quá mức các ví dụ đào tạo được gắn nhãn. Họ cho thấy rằng tự đào tạo thực sự là một cơ sở rất tốt so với các kỹ thuật chính quy hóa và tăng cường dữ liệu thông thường. So với các nghiên cứu này, Tự thử nghiệm là mô hình đầu tiên sử dụng quá trình chưng cất mô hình [25] cùng với một giả thuyết để chuyển thông tin qua các lần lặp lại, cho phép nó có khả năng sửa đổi các nhãn giả. Nó tích hợp mô hình đào tạo trước / tinh chỉnh với tự đào tạo, sử dụng quy trình tối ưu hóa hiệu quả cùng với kỹ thuật nhiễu loạn để giảm thiểu tác động tiêu cực của các nhãn giả ồn ào.

Các nghiên cứu liên quan chặt chẽ khác.Ngoài các nghiên cứu trên, Self- Pretraining cũng liên quan đến các nghiên cứu về chưng cất mô hình [25] và tổ hợp thời gian [32]. Chưng cất mô hình đã được đề xuất trong [11, 25] để chuyển kiến thức từ mô hình này sang mô hình khác. Trong [16], các tác giả chỉ ra rằng việc chuyển giao kiến thức của một mạng lớn, được đào tạo bởi một nhiệm vụ tự giám sát, sang một mạng nhỏ cải thiện khả năng tổng quát hóa. Đóng góp chính của họ là cho thấy rằng các mô hình lớn được đào tạo dễ dàng hơn, và do đó, có thể được sử dụng như một proxy để đào tạo các mạng nhỏ. Mô hình của họ không lặp lại và không khám phá dữ liệu chưa được gắn nhãn để trích xuất thông tin mới. Các mạng tái sinh đã được đề xuất trong [19], các tác giả cho thấy rằng chỉ cần chưng cất một mạng nơ-ron vào chính nó sẽ cải thiện hiệu suất. Mô hình của họ không phải là một thuật toán bán giám sát và không được đề xuất để khai thác dữ liệu không được gắn nhãn. Các tác giả trong [57] chỉ ra rằng thuật toán tự huấn luyện thần kinh thường xuyên có thể được cải thiện bằng cách thêm nhiễu vào mô hình. Tương tự như công việc của chúng tôi, chúng cho phép các nhãn giả phát triển qua các lần lặp lại. Ngoài bước này, họ không đề xuất bất kỳ sửa đổi nào đối với thuật toán tự đào tạo. Ngoài ra, hiệu quả của mô hình của họ không được khám phá trong bối cảnh bán giám sát. Một cách tiếp cận rất gần với nghiên cứu này được trình bày trong [23], nơi các tác giả một lần nữa chỉ ra rằng việc thêm nhiễu vào biểu diễn bên trong của mô hình sẽ nâng cao hiệu suất tự đào tạo. Dàn nhạc tạm thời đã được đề xuất trong [32]. Các tác giả đề xuất duy trì mức trung bình dự đoán trên mỗi mẫu của dữ liệu không được gắn nhãn trong các kỷ nguyên và hạn chế phương sai dự đoán. Mô hình của họ không dựa trên việc tự đào tạo, không có chiến lược để tách biệt dữ liệu được gắn nhãn khỏi dữ liệu không được gắn nhãn, và trở nên khó sử dụng khi sử dụng các tập dữ liệu lớn. Các tác giả trong [52] giải quyết độ phức tạp cao của việc tổng hợp theo thời gian bằng cách cập nhật trọng số của mô hình qua các kỷ nguyên, thay vì lưu trữ các dự đoán.

Chương 3: Các thuật toán được sử dụng để

* 1. Kết quả nghiên cứu

*(Chỉ nêu những nội dung chính, không liệt kê chương, mục)*

**3. Kết luận và kiến nghị**:

*(Nêu rõ kết quả nghiên cứu và kiến nghị nếu có)*

Chú ý: Không viết lời cám ơn.

Ý kiến xác nhận của GVHD