Trích xuất các thực thể y tế từ phư ơ ng tiện truyền thông xã hội

Sanja Šćepanović
Phòng thí nghiệm Nokia
Bell Cambridge,
Vư ơ ng quốc Anh sanja.scepanovic@nokia-bell-labs.com

Enrique Martín-López Nokia
Bell Labs
Cambridge,
Vư ơ ng quốc Anh enrique.martin-lopez@nokia-bell-labs.com

Daniele Quercia
Nokia Bell Labs
Cambridge, Vư ơ ng
quốc Anh quercia@cantab.net

Khan Baykaner
Phòng thí nghiệm Nokia
Bell Cambridge,
Vư ơ ng quốc Anh khan.baykaner@nokia-bell-labs.com

TRỪU TƯỢNG

Trích xuất chính xác các thực thể y tế từ phư ơ ng tiện truyền thông xã hội là một thách thức vì mọi ngư ời sử dụng ngôn ngữ thông thư ờng với các cách diễn đạt khác nhau cho cùng một khái niệm và họ cũng mắc lỗi chính tả. Công việc trư ớc đây hoặc tập trung vào các bệnh cụ thể (ví dụ: trầm cảm) hoặc thuốc (ví dụ: opioid) hoặc, nếu làm việc với nhiều tổ chức y tế, thì chỉ xử lý các bộ dữ liệu chuẩn cá nhân và quy mô nhỏ (ví dụ: AskaPatient). Trong công việc này, trư ớc tiên chúng tôi đã trình bày cách trích xuất chính xác nhiều loại thực thể y tế như triệu chứng, bệnh và tên thuốc trên ba bộ dữ liệu chuẩn từ các nguồn truyền thông xã hội khác nhau, sau đó cũng xác thực phư ơ ng pháp này trên bộ dữ liệu Reddit quy mô lớn .

Trư ớc tiên, chúng tôi đã triển khai một phư ơ ng pháp học sâu bằng cách sử dụng các nhúng theo ngữ cảnh dựa trên hai bộ dữ liệu điểm chuẩn hiện có, một bộ chứa các bài đăng của AskaPatient đư ợc chú thích (CADEC) và bộ còn lại chứa các tweet có chú thích (Micromed), vư ợt trội so với các phư ơ ng pháp tiên tiến nhất hiện có. Thứ hai, chúng tôi đã tạo tập dữ liệu điểm chuẩn bổ sung bằng cách chú thích các thực thể y tế trong các bài đăng 2K Reddit (đư ợc cung cấp công khai dư ới tên MedRed) và cho thấy rằng phư ơ ng pháp của chúng tôi cũng hoạt động tốt trên tập dữ liệu mới này.

Cuối cùng, để xác thực tính chính xác của phư ơ ng pháp của chúng tôi trên quy mô lớn, chúng tôi đã áp dụng mô hình đư ợc đào tạo trư ớc trên MedRed cho nửa triệu bài đăng trên Reddit. Các bài đăng đến từ các subreddit dành riêng cho bệnh nên chúng tôi có thể phân loại chúng thành 18 bệnh dựa trên subreddit đó. Sau đó, chúng tôi đã đào tạo một bộ phân loại học máy để dự đoán danh mục của bài đăng chỉ từ các thực thể y tế đư ợc trích xuất. Điểm F1 trung bình giữa các danh mục là 0,87. Những kết quả này mở ra những cơ hội hiệu quả về chi phí mới để lập mô hình, theo dõi và thậm chí dự đoán hành vi sức khỏe trên quy mô lớn.

KHÁI NIỆM CCS

• Tin học ứng dụng Tin học y tế; • Các phư ơ ng pháp điện toán Xử lý ngôn ngữ tự nhiên.

ACM CHIL '20, ngày 2–4 tháng 4 năm 2020, Toronto, ON,
Canada © 2020 Bản quyền thuộc về chủ sở hữu/(các) tác giả. Quyền xuất bản đư ợc cấp phép cho ACM.
ACM ISBN 978-1-4503-7046-2/20/04. . . $15,00
https://doi.org/10.1145/3368555.3384467

TỪ KHÓA

thảo luận về sức khỏe khai thác, học sâu, Reddit

Định dạng tham chiếu ACM:
Sanja Šćepanović, Enrique Martín-López, Daniele Quercia và Khan Baykaner. 2020. Trích xuất các thực thể y tế từ mạng xã hội. Trong Hội nghị ACM về Sức khỏe, Suy luận và Học tập (ACM CHIL '20), ngày 2–4 tháng 4 năm 2020, Toronto, ON, Canada. ACM, New York, NY, USA, 12 trang. https://doi.org/10.1145/3368555.3384467

1. GIỚI THIỆU

Ngày càng có nhiều ngư ời sử dụng các diễn đàn trực tuyến để thảo luận về sức khỏe của họ. Những diễn đàn như vậy bao gồm từ những nơ i mà bệnh nhân xin lời khuyên từ các chuyên gia y tế (The Body, Health24), đến những nơ i họ nói chuyện với nhau (AskaPatient, MedHelp) hoặc những nơ i họ nói chuyện với công chúng (Reddit). Trong đó, mọi ngư ời có xu hư ớng thảo luận về các bệnh và triệu chứng mà họ gặp phải cũng như các loại thuốc và biện pháp khắc phục khác nhau mà họ thấy hữu ích. Không chỉ bệnh nhân mà cả các chuyên gia y tế [21, 24, 39, 56, 60] và các tổ chức y tế [14, 21, 22, 31, 68] đang sử dụng mạng xã hội ngày càng nhiều.

Moorhead et al. [44] đã thực sự chỉ ra rằng phư ơ ng tiện truyền thông xã hội đã cung cấp một nguồn thông tin thay thế cho các cuộc phỏng vấn bệnh nhân truyền thống, báo cáo lâm sàng và hồ sơ sức khỏe điện tử.

Một trong những công cụ đư ợc áp dụng rộng rãi nhất để trích xuất các thực thể y tế từ các văn bản y tế chính thức như hồ sơ sức khỏe điện tử là MetaMap.1 Điều này sử dụng phư ơ ng pháp tiếp cận kiến thức chuyên sâu (tư ợng trư ng) dựa trên Hệ thống ngôn ngữ y tế thống nhất (UMLS). Các nghiên cứu gần đây đã chỉ ra rằng MetaMap không hoạt động tốt trên dữ liệu truyền thông xã hội như trên văn bản y tế chính thức [11, 20, 69].

Kết quả là, như chúng ta sẽ thấy trong Phần 2, trọng tâm nghiên cứu đã chuyển từ các công cụ truyền thống như MetaMap sang các kỹ thuật NLP thống kê, bao gồm cả học sâu. Những kỹ thuật này hứa hẹn sẽ khắc phục một số điểm yếu của cách tiếp cận tư ợng trư ng bằng cách tính toán tốt hơ n hai khía cạnh then chốt để phân tích chính xác văn bản truyền thông xã hội: ngữ cảnh và các kiểu diễn đạt không chính thức tinh tế.

Tuy nhiên, những kỹ thuật này yêu cầu các bộ dữ liệu đư ợc chú thích chuyên nghiệp để đào tạo, vốn khan hiếm, hạn chế về kích thư ớc và không phải lúc nào cũng có sẵn công khai. Do đó, không có nghiên cứu nào điều tra khả năng ứng dụng của các mô hình trích xuất thực thể y tế ngoài các bộ dữ liệu nhỏ, đơ n lẻ như CADEC [28] (1250 bài đăng đư ợc chú thích từ AskaPatient) hoặc Micromed [27] (1300 tweet đư ợc chú thích).

Mục tiêu của công việc này là xây dựng một khuôn khổ trích xuất chính xác nhiều thực thể y tế từ các trang truyền thông xã hội khác nhau,

1https://metamap.nlm.nih.gov/

và để chứng minh khả năng ứng dụng của nó trên quy mô lớn. Để đạt đư ợc mục tiêu này, chúng tôi đã thực hiện ba đóng góp

chính: (1) Chúng tôi đã thiết kế một khung dựa trên học sâu bằng cách sử dụng các nhúng theo ngữ cảnh để trích xuất chính xác nhiều loại thực thể y tế như bệnh, triệu chứng và tên thuốc (Phần 3). Phư ơ ng pháp này vư ợt trội so với các hiệu suất đư ợc công bố tốt nhất [67, 73] trên hai bộ dữ liệu điểm chuẩn (CADEC và Micromed) đạt đư ợc điểm F1 là 0,82 trên các bài đăng của AskaPatient và 0,72 trên các tweet (Phần 5).

(2) Để bổ sung cho các bộ dữ liệu điểm chuẩn hiện có, chúng tôi chú thích từ đám đông các bài đăng trên Reddit năm 1977 về các bệnh, triệu chứng và tên thuốc (Phần 4.4). Bộ dữ liệu này đư ợc gọi là MedRed và hiện đã có sẵn công khai2 . Chúng tôi đã đánh giá phư ơ ng pháp của mình trên bộ dữ liệu MedRed (Phần 5) và nó đã đạt đư ợc điểm F1 là 0,73.

(3) Cuối cùng, chúng tôi đã xác thực phư ơ ng pháp của mình trên nửa triệu bài đăng Reddit đư ợc phân loại dựa trên các subreddits dành riêng cho bệnh mà chúng đư ợc đăng. Chúng tôi đã sử dụng phư ơ ng pháp của mình để trích xuất các thực thể y tế và đối với mỗi bài đăng, chúng tôi dự đoán danh mục của bài đăng (tức là bệnh) chỉ dựa trên các thực thể đư ợc trích xuất (Phần 6). Với việc sử dụng từ vựng đư ợc chấp nhận rộng rãi trong tài liệu truyền thông xã hội , để làm cơ sở, chúng tôi đã tạo ra một từ vựng (Dis-LIWC) chứa các triệu chứng và tên thuốc liên quan đến 18 bệnh có trong các bài đăng trên Reddit. Sau đó, chúng tôi đã trích xuất các thực thể y tế từ các bài đăng Reddit bằng Dis-LIWC, cũng như sử dụng MetaMap.
Trên 18 bệnh, điểm F1 trung bình phân loại bằng cách sử dụng các thực thể đư ợc trích xuất theo phư ơ ng pháp của chúng tôi là 0,87 so với 0,61 theo các thực thể của MetaMap và 0,45 theo các thực thể của Dis-LIWC.

2 CÔNG VIỆC LIÊN QUAN

Chúng tôi xem xét cả hai phư ơ ng pháp hiện có để trích xuất các thực thể y tế từ phư ơ ng tiện truyền thông xã hội (Phần 2.1) và các ứng dụng của chúng (Phần 2.2).

2.1 Trích xuất thực thể y tế Các phư ơ ng pháp tiếp

cận ban đầu để trích xuất thực thể y tế từ phư ơ ng tiện truyền thông xã hội là dựa trên từ khóa [20, 34, 39, 58], sau đó là các phư ơ ng pháp dựa trên từ vựng dành riêng cho miền [20, 47, 50, 69, 71] . Mặc dù các phư ơ ng pháp này có thể hoạt động khá tốt trên các văn bản y tế chính thức, như ng chúng có những hạn chế nổi tiếng khi áp dụng cho dữ liệu truyền thông xã hội: chúng không nắm bắt đư ợc tính không đồng nhất về ngữ nghĩa của các biểu thức của ngư ời dùng và thích ứng với sự thay đổi của ngôn ngữ thông thư ờng và lỗi chính tả [10, 51].

Do đó, các phư ơ ng pháp học máy như Trư ờng ngẫu nhiên có điều kiện (CRF) ngày càng đư ợc áp dụng để khai thác văn bản truyền thông xã hội [20, 35, 46, 72]. Tuy nhiên, gần đây hơ n, các phư ơ ng pháp học sâu như Mạng thần kinh tái phát (RNN) đã trở nên phổ biến hơ n CRF [61, 63] và đã trở thành kỹ thuật tiếp theo để trích xuất các yêu cầu y tế từ phư ơ ng tiện truyền thông xã hội [67]. Hạ và cộng sự. [70] đã đề xuất một mô hình RNN đư ợc tăng cư ờng bằng các nhúng đư ợc đào tạo trên kho dữ liệu y tế và đánh giá nó trên tập dữ liệu gồm khoảng 6 nghìn bài đăng, không có sẵn công khai.

Chúng tôi đã so sánh công việc của mình với hai cách tiếp cận đã cho thấy kết quả đư ợc công bố tốt nhất trên Bộ dữ liệu AskaPatient (CADEC) [67] và trên Bộ dữ liệu Twitter (Micromed) [73], đây là hai bộ dữ liệu có sẵn công khai:

(1) Các tweet từ Micromed đư ợc khai thác bởi Yepes và MacKinlay [73]'s Long Short-Term Memory (LSTM) RNN.

(2) Các bài đăng của AskaPatient từ CADEC đã đư ợc khai thác bởi bộ kỹ thuật của Tutubalina và Nikolenko [67] bao gồm LSTM, Đơ n vị tái phát có cổng (GRU) và đơ n vị Mạng thần kinh chuyển đổi (CNN).

2.2 Ứng dụng truyền thông xã hội Khai thác sức khỏe

không chỉ đư ợc áp dụng cho các diễn đàn sức khỏe mà còn cho các trang truyền thông xã hội của Twitter và Reddit [23, 45, 51].
Twitter. Sarker et al. [57] đã phát triển một bộ phân loại đư ợc giám sát để phát hiện các tweet đề cập đến lạm dụng ma túy và Karisani và Agichtein [29] để phát hiện các tweet có bất kỳ đề cập nào về sức khỏe cá nhân. Tổng quát hơ n, MacKinlay et al. [40] đã áp dụng một phư ơ ng pháp để trích xuất các thực thể y tế từ Twitter và sau đó nghiên cứu sự xuất hiện đồng thời của các triệu chứng khác nhau đư ợc đề cập với một loại thuốc ibuprofen có tên là Advil. Theo cách tư ơ ng tự, Yepes et al. [74] áp dụng mô hình chủ đề để theo dõi các triệu chứng liên quan đến sự mệt mỏi từ các tweet đư ợc định vị địa lý của một thành phố trong hơ n một năm. Tất cả các nghiên cứu này đều mang tính chất mô tả và do đó, chúng không tập trung vào các thách thức kỹ thuật liên quan đến việc khai thác xuyên suốt một tập hợp các thực thể y tế đa dạng.

Reddit. Công viên và cộng sự. [48] so sánh ba cộng đồng Reddit liên quan đến sức khỏe tâm thần (r/Trầm cảm, r/Lo lắng và r/PTSD) và nhận thấy rằng họ có bốn chủ đề chung: chia sẻ những cảm xúc tích cực, lòng biết ơ n vì đã nhận đư ợc sự hỗ trợ về mặt tinh thần, các vấn đề về giấc ngủ và công việc -vấn đề liên quan. Choudhury và De [9] mô tả diễn ngôn về sức khỏe tâm thần trên Reddit bằng cách sử dụng kết hợp tìm kiếm dựa trên từ khóa và từ vựng LIWC. Park và Conway [47] đã theo dõi các cuộc trò chuyện xung quanh bốn chủ đề đư ợc công chúng quan tâm trong hơ n 8 năm: Ebola, thuốc lá điện tử, cúm và cần sa. Đúng như dự đoán , các cuộc thảo luận đã tăng lên đáng kể để đáp ứng với các sự kiện ngoại sinh như trư ờng hợp đầu tiên mắc bệnh Ebola đư ợc chẩn đoán và chủng vi rút cúm H1N1 lần đầu tiên đư ợc xác định.

Gkotsis et al. [18] dự đoán căn bệnh liên quan đến một bài đăng Reddit nhất định từ toàn bộ nội dung của nó, không chỉ từ các thực thể y tế có trong đó – như chúng tôi hạn chế làm. Bò tót et al. [17] dự đoán nguy cơ tự tử của ngư ời dùng Reddit dựa trên các bài đăng của anh ấy/cô ấy. Cuối cùng, bằng cách tham khảo địa lý ngư ời dùng Reddit, Balsamo et al. [7] ư ớc tính tỷ lệ tiêu thụ thuốc phiện trên khắp các bang của Hoa Kỳ.

Để tổng kết bài đánh giá tài liệu này, chúng tôi thấy rằng có hai cách tiếp cận nghiên cứu: một cách tiếp cận tập trung vào việc thiết kế các giải pháp học máy hiện đại, cuối cùng đư ợc áp dụng cho các bộ dữ liệu hạn chế đư ợc dán nhãn cẩn thận; và cái còn lại tập trung vào dữ liệu truyền thông xã hội quy mô lớn như ng trong các trư ờng hợp sử dụng hạn chế. Có rất ít công việc trong việc kết hợp hai cách tiếp cận.

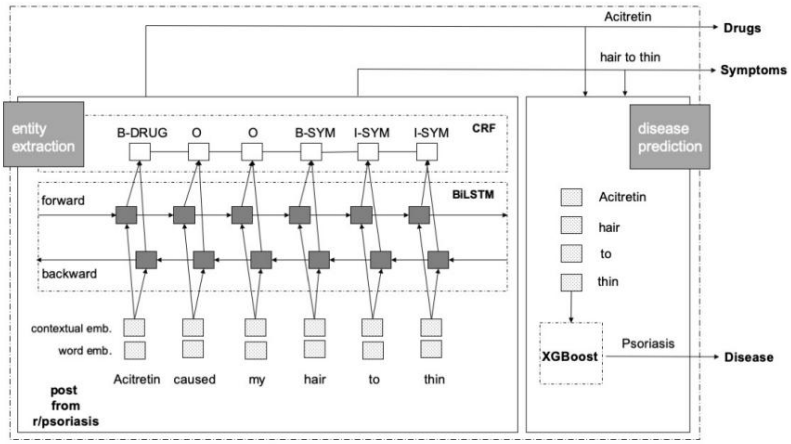
3 PHƯ Ơ NG PHÁP

Để giải quyết lỗ hổng nghiên cứu đó, chúng tôi đã thiết kế và cung cấp công khai3 một khung (Hình 1) để khai thác các cuộc thảo luận về sức khỏe dựa trên học sâu và nhúng theo ngữ cảnh phục vụ hai chức năng: trích xuất các thực thể y tế từ văn bản truyền thông xã hội và dự đoán bệnh đư ợc thảo luận .

Đối với mô-đun trích xuất thực thể (hình chữ nhật bên trái trong Hình 1), chúng tôi đã sử dụng kiến trúc ghi nhãn trình tự BiLSTM-CRF kết hợp với nhúng theo ngữ cảnh.

2http://goodcitylife.org/Humane-AI

3http://goodcitylife.org/Humane-AI



Hình 1: Khung của chúng tôi xử lý câu “Acitretin khiến tóc tôi mỏng đi”, đư ợc lấy từ subreddit r/psoriasis. Khung có hai mô-đun. Lần lư ợt, mô-đun trích xuất thực thể có hai lớp: lớp BiLSTM với các đơ n vị LSTM của nó đư ợc biểu thị dư ới dạng hình vuông tối và lớp CRF với các đơ n vị của nó đư ợc biểu thị dư ới dạng hình vuông màu trắng . Các phần nhúng theo ngữ cảnh và từ đư ợc sử dụng cho đầu vào đư ợc biểu diễn dư ới dạng các ô vuông chấm. Trong lớp CRF, các triệu chứng, bệnh và tên thuốc đư ợc trích xuất đư ợc hiển thị: mỗi từ đư ợc đánh dấu là DRUG/SYM, nếu nó là một phần của thực thể thuốc/triệu chứng hoặc là O nếu nó không phải là một trong hai. Mô-đun thứ hai – mô-đun dự đoán bệnh – sử dụng bộ phân loại XGBoost lấy các triệu chứng đư ợc trích xuất và tên thuốc làm đầu vào và dự đoán bệnh tư ơ ng ứng.

BiLSTM-CRF. Kiến trúc này đư ợc giới thiệu bởi Huang et al. [26], đã nhiều lần đư ợc hiển thị để trích xuất chính xác các thực thể [2, 54, 62] và bao gồm hai lớp. Lớp đầu tiên là mạng BiLSMT (hình chữ nhật nét đứt trong Hình 1), viết tắt của LSTM hai hư ớng. Nó đư ợc gọi là “hai chiều” vì có hai bộ đơ n vị LSTM (các ô vuông nhỏ tối màu trong Hình 1) qua đó “kích hoạt mạng” chảy theo hai hư ớng, tiến và lùi. Các đầu ra của BiLSTM sau đó đư ợc chuyển đến lớp thứ hai: lớp CRF (đư ợc đặt trong hình chữ nhật nét đứt khác). Các dự đoán của lớp này (các ô vuông màu trắng) đại diện cho đầu ra của mô-đun trích xuất thực thể. Để trích xuất các thực thể y tế của các triệu chứng và tên thuốc, BiLSTM-CRF cần đư ợc đào tạo với dữ liệu đư ợc dán nhãn. Để làm đư ợc điều này, chúng tôi đã sử dụng sơ đồ gắn thẻ IoB [55, 65].

Mỗi từ trong sơ đồ này đư ợc dán nhãn là một triệu chứng (SYM), một loại thuốc (DRUG) hoặc không có từ nào trong hai (O); và vị trí của nó đư ợc đánh dấu là ở đầu thực thể (B-) hay không (I-). Ví dụ, trong câu ở Hình 1, từ “Acitretin” đư ợc dán nhãn là B-DRUG, ba từ “tóc to mỏng” là B-SYM, I-SYM, I-SYM, trong khi các từ còn lại đư ợc dán nhãn là Ô.

Có dữ liệu đư ợc gắn nhãn, mô-đun trích xuất thực thể trải qua hai giai đoạn đào tạo và thử nghiệm điển hình. Trong giai đoạn đào tạo, chúng tôi đã liên kết mỗi từ đư ợc gắn nhãn với một biểu diễn nhúng (hình vuông chấm), là một vectơ phản ánh vị trí của từ trong một không gian ngữ nghĩa. Các cặp kết quả (nhúng ding, nhãn) đư ợc đư a tuần tự vào BiLSTM, cập nhật trọng số của mạng. Cuối cùng, lớp CRF đã cải thiện hơ n nữa các liên kết liên quan đến tất cả các cặp (nhúng, nhãn).

Trong giai đoạn thử nghiệm (tức là giai đoạn mà mô-đun phải trích xuất các thực thể từ mỗi câu của bộ thử nghiệm), chúng tôi đã liên kết từng từ trong một câu trong bộ thử nghiệm với phần nhúng của nó và cung cấp phần nhúng thông qua các trọng số đã học trư ớc đó. Điều này dẫn đến một chuỗi các nhãn (đư ợc xuất ra bởi các ô vuông màu trắng trong hình, nghĩa là bởi các nút trong biểu đồ của CRF): các nhãn đư ợc đánh dấu bằng SYM hoặc DRUG đại diện cho cả hai triệu chứng (bao gồm cả bệnh) và tên thuốc đư ợc trích xuất. Nếu chúng ta tự giới hạn mình trong LSTM một chiều, thì mạng sẽ chỉ dự đoán nhãn của từ dựa trên các từ trư ớc đó. Thay vào đó, bằng cách sử dụng LSTM hai chiều, mạng đã dự đoán nhãn của từ dựa trên toàn bộ câu. Trong ví dụ đang chạy của chúng tôi, từ tóc đư ợc tách riêng có thể biểu thị những thứ khác nhau, bao gồm một bộ phận cơ thể hoặc một triệu chứng. Ngư ợc lại, trong LSTM hai chiều, từ tóc đư ợc sử dụng trong ngữ cảnh (trong ngữ cảnh của từ gầy) và do đó , đư ợc hiểu là một phần của triệu chứng.

Nhúng theo ngữ cảnh. Các liên kết trư ớc đây giữa các từ và nhúng có thể đư ợc thực hiện với các loại nhúng khác nhau.

Các nhúng đư ợc sử dụng phổ biến nhất là Vectơ toàn cầu cho biểu diễn từ (GloVe) [53] và Biểu diễn phân tán của từ (word2vec) [43]. Tuy nhiên, những điều này không tính đến ngữ cảnh của một từ. Ví dụ, từ 'đau' có thể là một triệu chứng (ví dụ: 'Tôi cảm thấy đau khắp ngư ời') hoặc có thể đư ợc sử dụng theo nghĩa bóng (ví dụ: 'Anh ấy thật là một nỗi đau phải đối mặt'). Để giải thích cho ngữ cảnh, nghiên cứu gần đây đã chuyển từ nhúng dựa trên từ sang nhúng theo ngữ cảnh. Chúng tôi đã thử nghiệm bốn loại nhúng như vậy (Phần 5.6):

Nhúng từ Mô hình Ngôn ngữ (ELMo). Các nhúng ELMo đư ợc xây dựng bằng cách gán cho mỗi mã thông báo một biểu diễn thu đư ợc bằng cách huấn luyện BiLSTM trên một kho văn bản lớn. Peters và cộng sự. [54] đã chỉ ra rằng các trạng thái LSTM cấp thấp hơ n nắm bắt các khía cạnh của cú pháp từ, trong khi các trạng thái cấp cao hơ n nắm bắt ngữ nghĩa.

Nhúng tinh tế. Các phần nhúng này đư ợc xây dựng bằng cách học cách dự đoán ký tự tiếp theo trong một chuỗi ký tự. Akbik và cộng sự. [3] đã chỉ ra rằng theo cách đó, các khái niệm ngôn ngữ như từ, câu và thậm chí cả tình cảm đư ợc tự động tiếp thu. Ở đó

cũng là phiên bản cải tiến của nhúng này có tên là Pooled Flair Embeddings.

Biểu diễn bộ mã hóa hai chiều từ bộ đệm của Transformers (BERT) . BERT đư ợc đào tạo trư ớc về hai nhiệm vụ không giám sát. Trong tác vụ đầu tiên, một số phần trăm mã thông báo đầu vào đư ợc che một cách ngẫu nhiên, sau đó mô hình đư ợc huấn luyện để dự đoán các mã thông báo bị che. Trong nhiệm vụ thứ hai, đư a ra hai câu, mô hình đư ợc đào tạo để dự đoán xem câu này có nối tiếp câu kia trong một đoạn văn bản hay không. Do đó, BERT là một mô hình NLP chung [12]. Lư u và cộng sự. [38] đã sao chép mô hình BERT với các tham số và lựa chọn thiết kế khác nhau để tạo ra một phiên bản nâng cao có tên là phư ơ ng pháp nhúng Phư ơ ng pháp tiếp cận trư ớc đào tạo BERT đư ợc tối ư u hóa mạnh mẽ (RoBERTa).

nhúng BERT y tế chuyên dụng. Bằng cách đào tạo trư ớc một mô hình BERT trên khối dữ liệu y sinh lớn thay vì khối dữ liệu chung, Lee et al. [36] đã tạo BioBERT. Việc sử dụng BioBERT vư ợt trội đáng kể so với công nghệ tiên tiến nhất trong ba nhiệm vụ khai thác văn bản y sinh tiêu biểu, bao gồm cả khai thác thực thể y sinh. Alsentzer et al. [4] đã đào tạo trư ớc một mô hình BERT trên một loại tập đoàn khác – ghi chú lâm sàng. Họ đã chỉ ra rằng ClinicalBERT vư ợt trội hơ n cả BERT và BioBERT trong các nhiệm vụ cụ thể, chẳng hạn như dự đoán tái nhập viện dựa trên các ghi chú lâm sàng hoặc tóm tắt xuất viện.

Với kích thư ớc nhỏ của bộ dữ liệu đào tạo của chúng tôi (Phần 4), chúng tôi đã sử dụng các phần nhúng theo ngữ cảnh làm đầu vào cho kiến trúc Bi-LSTM-CRF mà không cần đào tạo chúng thêm nữa.

Mô-đun Dự báo Bệnh tật. Đối với mỗi bài đăng trên mạng xã hội, mô-đun thứ hai (hình chữ nhật bên phải trong Hình 1) sau đó nhận đư ợc các thực thể đư ợc trích xuất từ bài đăng bởi mô-đun đầu tiên (cụ thể hơ n là các nhúng từ xếp chồng lên nhau của chúng) trong đầu vào và dự đoán khả năng xảy ra cao nhất . bệnh thảo luận trong bài viết. Thành phần cốt lõi của nó là một bộ phân loại dựa trên một tập hợp các cây quyết định

tăng cư ờng độ dốc (XGBoost) [16]. XGBoost đã đư ợc phát hiện là cung cấp hiệu suất tốt một cách nhất quán bằng cách kết hợp lặp đi lặp lại các kết quả của một tập hợp các cây phân loại và hồi quy vào một bộ học mạnh duy nhất [8].

4 BỘ DỮ LIỆU

Để nghiên cứu khả năng ứng dụng chung của phư ơ ng pháp của chúng tôi, chúng tôi cần đánh giá nó dựa trên nhiều bộ dữ liệu khác nhau. Ngoài hai bộ dữ liệu điểm chuẩn từ AskaPatient và Twitter (Phần 4.1 và 4.2), chúng tôi đã phân loại nửa triệu bài đăng trên Reddit về các loại bệnh (Phần 4.3) và các thực thể thuốc và triệu chứng đư ợc chú thích trong một tập hợp con của những bài đăng đó (Phần 4.4) .

4.1 Bộ dữ liệu AskaPatient (CADEC)

CADEC đã trở thành kho dữ liệu tiêu chuẩn để trích xuất thực thể y tế từ phư ơ ng tiện truyền thông xã hội và việc sử dụng nó giúp so sánh kết quả của chúng tôi với kết quả trong các nghiên cứu trư ớc đây. Nó bao gồm hơ n 1 nghìn bài đăng có chú thích từ AskaPatient4 (Bảng 1). Trong diễn đàn này, bệnh nhân thảo luận về kinh nghiệm của chính họ liên quan đến nhiều khía cạnh liên quan đến sức khỏe. Trong kho văn bản, đề cập đến phản ứng có hại của thuốc (6318 thực thể), triệu chứng (275), kết quả lâm sàng (435), bệnh (283) và tên thuốc (1800) đư ợc chú thích [28].

4https://www.askaopathy.com

Bảng 1: Thống kê cho bộ dữ liệu Reddit Medical Entities (MedRed) so với bộ dữ liệu AskaPatient (CADEC) [28] và Micromed [27] đã đư ợc sử dụng làm tài liệu tham khảo để tạo MedRed.

	MedRed	CADEC	Micromed
#bài #gửi.	1977 8794	1321 7632	734 1027
# từ	147,915	101,486	15.690
khoảng thời gian	Jan-Jun 2017	Jan 2001-Sep 2013	May 2014
liên lạc.	18 subreddits	lipitor, diclofenac	Twitter
thực thể	4485	9111	757

4.2 Bộ dữ liệu Twitter (Micromed)

Một bộ dữ liệu khác hiện có như ng ít đư ợc chấp nhận hơ n là Micromed. Jimeno Yepes và cộng sự. [27] đã chú thích 1300 tweet về triệu chứng (764 thực thể), bệnh (253) và dư ợc chất (233). Vì các từ (bao gồm cả những từ liên quan đến sức khỏe) có thể đư ợc sử dụng theo nghĩa bóng trên phư ơ ng tiện truyền thông xã hội [40], nên Micromed cũng đi kèm với một cờ cho biết liệu mỗi từ đư ợc trích xuất (ví dụ: đau) có thực sự là một thực thể y tế hay không (ví dụ: 'Tôi cảm thấy đau khắp ngư ời cơ thể của tôi') hay không (ví dụ: 'Anh ấy thật khó chịu'). Vì bộ dữ liệu chỉ cung cấp các số nhận dạng tweet (chứ không phải các tweet thực tế), nên chúng tôi đã thu thập dữ liệu các số nhận dạng đó và tại thời điểm viết bài, 734 trong số 1300 tweet ban đầu vẫn có sẵn.

4.3 Bộ dữ liệu Reddit đầy đủ (Subreddits về bệnh)

Do tính đặc hiệu của bộ dữ liệu AskaPatient (chỉ bao gồm hai loại thuốc) và kích thư ớc hạn chế của bộ dữ liệu Twitter, chúng tôi đã chuyển sang Reddit để thu thập thêm dữ liệu. Reddit là một nền tảng xã hội phổ biến về tin tức và giải trí, nơ i ngư ời dùng thảo luận về nhiều chủ đề khác nhau.
Theo thống kê chính thức từ trang web,5 Reddit có hơ n 330 triệu ngư ời dùng hoạt động trung bình hàng tháng, hơ n 138 nghìn cộng đồng đang hoạt động (subreddits) và phần lớn ngư ời dùng đến từ Hoa Kỳ, Canada và Vư ơ ng quốc Anh. Các cuộc thảo luận đư ợc nhóm thành các subreddits (nhóm con) theo chủ đề (ví dụ: trầm cảm). Subreddits bao gồm các chủ đề thảo luận đư ợc bắt đầu bởi bài đăng của ngư ời dùng, thư ờng đư ợc theo sau bởi các nhận xét/câu trả lời từ những ngư ời dùng khác.

Ngư ời dùng thảo luận về nhiều chủ đề khác nhau, bao gồm các vấn đề sức khỏe và chia sẻ kinh nghiệm của bản thân, xin lời khuyên và học hỏi từ những ngư ời khác. Có toàn bộ danh mục 'sức khỏe' trên Reddit và nó chứa tổng cộng khoảng 40 subreddits. Trong số các subreddits này, chúng tôi đã chọn những subreddits đáp ứng hai tiêu chí chính:

(1) Đã hoạt động, tức là những ngư ời có trung bình ít nhất hàng trăm bài đăng mỗi tháng. Tiêu chí đầu tiên này cho phép lựa chọn các bệnh đư ợc quan tâm chung.
(2) Tập trung vào một bệnh cụ thể, chẳng hạn như r/trầm cảm và r/sỏi thận. Tiêu chí thứ hai này tạo ra mối quan hệ rõ ràng giữa subreddit và bệnh tật, do đó, cho phép chúng tôi chú thích ngầm các bài đăng trên Reddit tùy thuộc vào subreddit nào chúng xuất hiện.

Bằng cách áp dụng cả hai tiêu chí, chúng tôi còn lại 18 subreddits, từ đó chúng tôi tải xuống các bài đăng trong sáu tháng đầu năm 2017 (Bảng 2).

5https://www.redditinc.com/press

Bảng 2: Bộ dữ liệu Reddit (đư ợc gắn nhãn cho Bệnh tật): tổng số bài đăng trong mỗi subreddit từ tháng 1 đến tháng 6 năm 2017.

phụ bản	tên bệnh	bài viết
r/bpd	Rối loạn nhân cách ranh giới r/cfs	48000
	Hội chứng mệt mỏi mãn tính r/	10711
crohnsdisease	Bệnh Crohn r/chứng	30774
mất trí nhớ	Sa sút trí	1979
tuệ r/trầm cảm	trầm cảm r /	286968
bệnh tiểu đư ờng	Đái tháo đư ờng r /	54285
chứng khó đọc	Rối loạn hệ thống thần kinh tự trị r / liệt	1655
dạ dày	Hội chứng liệt dạ dày r/suy	679
giáp Suy giáp r/ibs		7990
	Hội chứng ruột kích thích r/	19497
viêm bàng quang kẽ	Viêm bàng quang kẽ mãn tính r/sỏi	1851
thận	Bệnh thận r/	1301
menieres	Bệnh menieres r/đa	613
xơ cứng Đa xơ cứng r/parkinsons r/vẩy nến r/		12996
thấp khớp r/	bệnh Parkinson	703
ngư ng thở	Bệnh vẩy nến	5734
khi ngủ	Viêm khớp dạng thấp	3736
	Chứng ngư ng thở lúc ngủ	7486
tổng cộng		496958

4.4 Các thực thể y tế trong Reddit (MedRed) tập dữ liệu

Đối với 1980 bài đăng từ Reddit (110 bài đăng đư ợc lấy mẫu ngẫu nhiên từ mỗi trong số 18 subreddits), chúng tôi đã thiết lập một thử nghiệm Mechanical Turk (MT) để chú thích các thực thể y tế trong đó. Vì CADEC là lớn nhất

bộ dữ liệu điểm chuẩn thư ờng đư ợc sử dụng, đư ợc chú thích bởi các chuyên gia, chúng tôi đã sử dụng một số bài đăng của nó để đảm bảo chất lư ợng của các chú thích của chúng tôi.

Mỗi nhiệm vụ MT bao gồm sáu bài đăng đư ợc gắn nhãn: bốn bài đăng đư ợc chọn ngẫu nhiên từ các bài đăng Reddit năm 1980; một là ' bài đăng kiểm soát' đư ợc lấy mẫu cẩn thận từ CADEC để giống với bài đăng Reddit điển hình; và cái cuối cùng là một 'bài bẫy' đư ợc tạo thủ công chứa chính xác một triệu chứng và một tên thuốc. Vị trí của các loại bài đăng khác nhau đư ợc sắp xếp ngẫu nhiên trong mỗi nhiệm vụ.

Chúng tôi đã hư ớng dẫn nhân viên trích xuất các thực thể thuộc hai loại: i) triệu chứng/bệnh và ii) tên thuốc. Hư ớng dẫn về những gì cấu thành một thực thể có liên quan tư ơ ng tự như hư ớng dẫn của Karimi et al. khi tạo tập dữ liệu CADEC.

Để đảm bảo chú thích chất lư ợng cao:

(1) Một nhiệm vụ chỉ có thể đư ợc thực hiện bởi những ngư ời lao động có sự chấp thuận tỷ lệ trên 95%.

(2) Một kết quả nhiệm vụ chỉ đư ợc chấp nhận nếu cả triệu chứng và tên thuốc của bài bẫy đều đư ợc xác định chính xác (loại bỏ 21% phản hồi).

(3) Mỗi bài đăng đư ợc dán nhãn bởi ít nhất 10 công nhân khác nhau.

(4) Mỗi bài đăng chỉ đư ợc chú thích với các thực thể đư ợc trích xuất bởi ít nhất hai công nhân một cách độc lập, đây đư ợc coi là một con số thỏa thuận tốt bởi công việc trư ớc đó [30].

Cuối cùng, chúng tôi đã sử dụng bài đăng kiểm soát từ CADEC trong số 6 bài đăng trong mỗi tác vụ để tính toán thỏa thuận theo cặp của các chú thích, như sau:

$$\text{Adr}(i, j) = \frac{\text{trận đấu}(Ai, Aj)}{\max(nAi, nAj)},$$

Bảng 3: Số lư ợng thực thể đư ợc gắn nhãn trong bộ dữ liệu MedRed.

phụ bản	thuốc	triệu chứng	tất cả
r/bpd	6	152	158
r/	33	226	259
cfs r/bệnh crohns	51	134	185
r/chứng mất	10	184	194
trí nhớ r/trầm	11	65	76
cảm r/tiểu	46	93	139
đư ờng r/rối loạn	54	333	387
tự chủ r/liệt dạ	69	251	320
dày r/suy giáp r/ibs	76	200	276
39 r/viêm bàng quang kẽ	84	135	174
r/sỏi thận r/menieres r/đa		252	336
xơ cứng 44 r/	55	223	278
parkinsons	43	306	349
76 r /vẩy nến r/thấp khớp r/		161	205
ngư ng thở khi ngủ		259	335
	93	148	241
	143	236	379
	41	158	199
kho văn bản	974	3511	4485

trong đó Ai là danh sách các thực thể y tế đư ợc nhân viên MT trích xuất, Aj là danh sách các thực thể y tế đư ợc các chuyên gia CADEC trích xuất (các vị trí kiểm soát), nAi là số lư ợng thực thể y tế trong Ai , nAj là số lư ợng thực thể y tế trong Aj , và match(Ai ,Aj) là số lư ợng thực thể y tế đư ợc trích xuất bởi cả nhân viên MT và chuyên gia. Thỏa thuận theo cặp trung bình ở dạng chặt chẽ, nghĩa là khi chỉ cho phép đối sánh chính xác, là 0,62 đối với các triệu chứng và 0,75 đối với thuốc, trong khi ở dạng thoải mái, tức là khi cho phép các thực thể chồng lên nhau (ví dụ: 'đau ' trùng với 'cơ n đau dữ dội'), là 0,77 đối với các triệu chứng và 0,83 đối với tên thuốc. Những điểm số này có thể so sánh với những điểm số đạt đư ợc trư ớc đây đối với các chú thích của chuyên gia CADEC [28], do đó xác nhận chất lư ợng của các chú thích MedRed.

MedRed là bộ dữ liệu điểm chuẩn mới để trích xuất thực thể y tế công khai cho các nhà nghiên cứu khác [59]6 .

5 ĐÁNH GIÁ

Mục tiêu chính của đánh giá của chúng tôi là đánh giá xem phư ơ ng pháp của chúng tôi có hoạt động cạnh tranh trên các bộ dữ liệu từ nhiều nguồn khác nhau hay không.

5.1 Số liệu đánh giá

Chỉ số đánh giá của chúng tôi là điểm F 1, F 1 = 2P · R/(P + R), tức là giá trị trung bình hài hòa của độ chính xác P và R thu hồi, trong đó:

$$\text{xác} \quad P = \frac{\#\text{ thực thể y tế đư ợc phân loại chính}}{\#\text{tổng số thực thể đư ợc phân loại là y tế}},$$

Và

$$R = \frac{\#\text{thực thể y tế đư ợc phân loại chính xác}}{\#\text{tổng số thực thể y tế}}.$$

Để thận trọng, chúng tôi chỉ tính là "đư ợc phân loại chính xác" đối với các thực thể khớp chính xác với nhãn sự thật cơ bản. Điều này có nghĩa là chúng tôi đã sử dụng điểm số F1 nghiêm ngặt, trái ngư ợc với phiên bản thoải mái đôi khi đư ợc sử dụng khi báo cáo kết quả trích xuất thực thể. Ngoài ra, do dữ liệu của chúng tôi đi kèm với sự mất cân bằng lớp (tức là,

6http://goodcitylife.org/Humane-AI

mã thông báo văn bản không tư ơ ng ứng như nhau với các triệu chứng, thuốc hoặc thực thể phi y tế), chúng tôi đã khắc phục điều đó bằng cách tính toán P và R sử dụng trung bình vi mô [6].

5.2 MetaMap MetaMap là

một công cụ đư ợc thiết lập tốt để trích xuất các khái niệm y tế từ văn bản bằng cách sử dụng NLP tư ợng trư ng và các kỹ thuật ngôn ngữ tính toán [5], và đã trở thành một phư ơ ng pháp cơ sở thực tế cho các nghiên cứu NLP liên quan đến sức khỏe [66]. MetaMap thực hiện trích xuất thực thể bằng cách tuân theo cách tiếp cận dựa trên quy tắc chuyên sâu về kiến thức với UMLS Metathesaurus làm nguồn kiến thức. Do đó, khi xử lý một câu, nó sẽ trả về một danh sách các mã thông báo tư ơ ng ứng với các thực thể y tế mà nó tìm thấy trong câu. Các thực thể này là

hoặc là tên thuốc – mà chúng tôi đã xác định là danh mục của MetaMap7 về Kháng sinh, Thuốc lâm sàng và Dư ợc chất – hoặc triệu chứng – nằm trong Bệnh hoặc Hội chứng, Phát hiện và Dấu hiệu hoặc Triệu chứng. Ngoài loại xử lý hậu kỳ này, chúng tôi cũng giới hạn kết quả của mình ở hai nguồn từ vựng: SNOMEDCT_US cho các triệu chứng và RxNorm cho thuốc.

5.3 TaggerOne TaggerOne là

một công cụ máy học sử dụng các mô hình bán Markov để cùng thực hiện hai tác vụ: trích xuất thực thể và chuẩn hóa thực thể. Công cụ này làm như vậy bằng cách sử dụng từ điển y khoa [33]. Tuy nhiên, vì chúng tôi có dữ liệu đào tạo để trích xuất thực thể chứ không phải dữ liệu dành cho quá trình hóa bình thư ờng nên chúng tôi không thể đào tạo TaggerOne trên dữ liệu của mình. Do đó, chúng tôi đã sử dụng phiên bản của nó đã đư ợc đào tạo trư ớc đây về kho dữ liệu y sinh "BioCreative V CDR corpus" [37] làm một trong những cơ sở của chúng tôi và chúng tôi có thể làm như vậy vì các thực thể y tế đư ợc trích xuất của nó tư ơ ng tự như của chúng tôi.

5.4 Các phư ơ ng pháp học sâu trư ớc đây Các mô hình học sâu

(DL) ngày càng trở thành giải pháp tiên tiến nhất để trích xuất thực thể y tế. Cách tiếp cận của chúng tôi đư ợc đánh giá dựa trên hai cách tiếp cận hiện có với kết quả tốt nhất trên các bộ dữ liệu tư ơ ng ứng: CADEC DL[67]. Tutubalina và Nikolenko [67] đã áp

dụng BiLSTM CRF bằng cách sử dụng từ nhúng chuyên dụng HealthVec [41] trên Bộ dữ liệu AskaPatient (CADEC). Do đó, chúng tôi gọi phư ơ ng pháp của họ là CADEC DL.

Micromed DL [73]. Yepes và MacKinlay [73] đã đề xuất một LSTM RNN có đầu ra đư ợc chuyển đến một bộ phân loại tuyến tính đư ợc đào tạo bằng cách sử dụng mất bản lề đa lớp. Chúng tôi gọi phư ơ ng pháp của họ là Micromed DL.

Do không có sẵn mã nguồn, chúng tôi đã lấy kết quả về hình thức tốt nhất cho hai cách tiếp cận này từ các ấn phẩm tư ơ ng ứng [67, 73], đảm bảo phân tích so sánh công bằng.

5.5 Cài đặt triển khai và đào tạo Để triển khai mô hình BiLSTM-CRF,

chúng tôi đã sử dụng Python với thư viện Flair [1] và chư ơ ng trình phụ trợ Pytorch [49]. Đối với mỗi phần nhúng đang đư ợc xem xét, chúng tôi đã sử dụng các mô hình ngôn ngữ của chúng từ các kho lư u trữ nguồn mở tư ơ ng ứng. Mạng đư ợc thiết lập với các thông số sau: 256 ẩn đơ n vị, tốc độ học

7https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml

Bảng 4: Điểm F1 (P/R) cho phư ơ ng pháp của chúng tôi khi sử dụng các cách nhúng khác nhau để trích xuất các thực thể từ ba bộ dữ liệu.

nhúng	HỏiBệnh nhân (CADEC)	Twitter (Micromed)	reddit (MedRed)
	Emb theo ngữ cảnh cá nhân.		
ELMo	.80 (.79/.80)	.69 (.69/.69)	.79
Tư cách tổng hợp sự tinh tế	(.79/.78)	.62 (.66/.58)	.80
	(.81/.79)	.63	
BERT			.70 (.69/.71)
RoBERTa			.69 (.68/.69)
BioBERT			.70 (.77/.64)
BERT lâm sàng			.70 (.74/.66)
			.73 (.77/.69)
			.66
	(.66 /.61)	.80(.79/.82) .70(.66/.75)	.81(.81/.81) .74 (.65/.84) .80(.79/.81) . 59
	Kết hợp ngữ cảnh và từ emb.		
RoBERTa + GloVe	.82 (.81/.82)	.72 (.69/.74)	.73 (.75/.70)

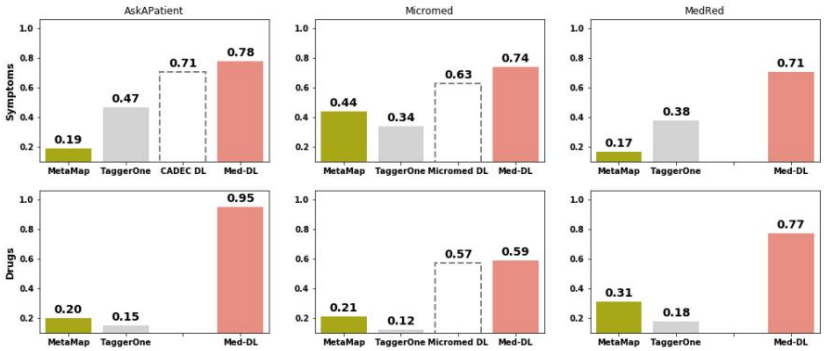
bắt đầu từ 0,1 và giảm dần một nửa mỗi khi không có sự cải thiện nào sau 3 kỷ nguyên, kích thư ớc lô là 4 và chúng tôi đã đào tạo bằng cách sử dụng cả bộ đào tạo và bộ phát triển. Quá trình đào tạo đư ợc thực hiện trên một GPU GeForce GTX 1080 duy nhất trong tối đa 200 kỷ nguyên hoặc trư ớc khi tốc độ học trở nên quá nhỏ (≤ 0,0001).

5.6 Kết quả

Trư ớc tiên, chúng tôi đã thử nghiệm các phiên bản khác nhau của khung sử dụng dif nhúng theo ngữ cảnh (không có bất kỳ từ nhúng nào) và đã làm như vậy trên từng bộ dữ liệu trong số ba bộ dữ liệu (ba cột trong Bảng 4). Sự khác biệt về điểm số là rất nhỏ trong trư ờng hợp của AskaPatient, như ng đáng chú ý trong trư ờng hợp của Twitter và Reddit. Trên khắp ba bộ dữ liệu, các phần nhúng chuyên dụng cho lĩnh vực y tế (tức là BioBERT và Clinical BERT) đã bị vư ợt trội so với hai trong số các phần nhúng chung (RoBERTA mang lại kết quả tốt nhất trên AskaP atient và Reddit, trong khi BERT đã làm như vậy trên Twitter). Điều đó chủ yếu là do các nhúng chuyên biệt nắm bắt ngôn ngữ y tế chính thức, trong khi các biểu hiện về sức khỏe trên phư ơ ng tiện truyền thông xã hội có tính chất không chính thức hơ n. Bằng cách xếp chồng RoBERTa với các nhúng từ GloVe [53], khung của chúng tôi mang lại hiệu suất tốt nhất trên cả ba bộ dữ liệu. Do đó, trong phần tiếp theo, chúng tôi báo cáo kết quả cho phiên bản sử dụng kết hợp nhúng RoBERTA và GloVE (Hình 2).

Bệnh nhân Aska (CADEC). Bộ dữ liệu AskaPatient (CADEC) đư ợc chia thành các bộ đào tạo (60%), nhà phát triển (20%) và thử nghiệm (20%), một sự phân chia đư ợc sử dụng bởi công việc trư ớc đó [42]. Bằng cách xem xét điểm F1 cho bộ dữ liệu AskaP atient (CADEC) (Hình 2), chúng tôi thấy rằng trên các triệu chứng, phư ơ ng pháp của chúng tôi có F1 là 0,78 và do đó, vư ợt trội so với MetaMap (0,19) và TaggerOne (0,47) và hoạt động tốt hơ n CADEC DL (.71), mặc dù CADEC DL bị hạn chế trong việc trích xuất các ADR hơ n là các triệu chứng hoặc bệnh tật. Kết quả cuối cùng này cho thấy rằng việc sử dụng hai loại nhúng - GloVe cho từ và RoBERTa cho ngữ cảnh - đã tạo ra sự khác biệt đáng kể. Cuối cùng, về mặt khai thác thuốc, chúng tôi chỉ có thể so sánh phư ơ ng pháp của mình với MetaMap (vư ợt trội hơ n rất nhiều), vì CADEC DL không báo cáo bất kỳ kết quả nào về nó (do đó có các vị trí trống trong Hình 2).

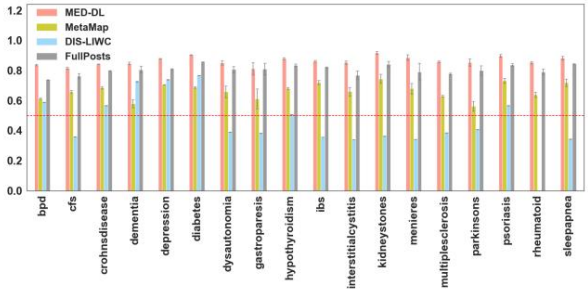
Twitter (Micromed). Tập dữ liệu Twitter Micromed đư ợc chia thành các tập đào tạo (50%), nhà phát triển (25%) và thử nghiệm (25%) (chúng tôi giữ ít nhất 25% tập dữ liệu để xác thực và thử nghiệm vì nó nhỏ).



Hình 2: Đánh giá phư ơ ng pháp của chúng tôi so với đư ờng cơ sở trong việc trích xuất các thực thể y tế trên ba bộ dữ liệu. Điểm F1 đư ợc hiển thị riêng cho các thực thể triệu chứng và thuốc. Các đư ờng đứt nét cho Micromed DL và CADEC DL biểu thị rằng những kết quả này không đư ợc chúng tôi tính toán mà đến từ các ấn phẩm tư ơ ng ứng ([67],[73]) và các thanh trống cho biết rằng các kết quả này không có sẵn từ các ấn phẩm này. Đối với Micromed DL trong trư ờng hợp có triệu chứng, chúng tôi lấy điểm trung bình có trọng số đư ợc báo cáo riêng cho các bệnh và triệu chứng. CADEC DL chỉ trích xuất ADR.

Bằng cách xem xét điểm số F1 cho tập dữ liệu Twitter (Micromed) (Hình 2), chúng tôi thấy rằng đối với các triệu chứng, phư ơ ng pháp của chúng tôi có điểm số F1 là 0,74 và do đó, vư ợt trội so với MetaMap (0,44) và TaggerOne (0,34) và đư ợc thực hiện tốt hơ n MicroMed DL (.63). Đối với thuốc, hiệu quả của phư ơ ng pháp của chúng tôi cao hơ n một chút so với Micromed DL (0,59 so với 0,57). Thật thú vị, MetaMap hoạt động tốt hơ n đáng kể trên Twitter so với trên nền tảng chuyên biệt hơ n của AskaPatient. Điều đó một phần là do bộ dữ liệu Micromed ban đầu đư ợc xây dựng bằng cách tìm kiếm các tweet cho các thuật ngữ UMLS[27] và MetaMap dựa trên UMLS.

Reddit (MedRed). Tập dữ liệu MedRed đư ợc chia thành các tập đào tạo (50%), nhà phát triển (25%) và kiểm tra (25%) (một lần nữa, chúng tôi giữ ít nhất 25% tập dữ liệu để xác thực và thử nghiệm vì tập dữ liệu này cũng nhỏ). Bằng cách xem xét điểm số F1 trên bộ dữ liệu MedRed (Hình 2), chúng tôi thấy rằng phư ơ ng pháp của chúng tôi có điểm số là 0,71 đối với các triệu chứng và 0,77 đối với thuốc, vư ợt trội so với MetaMap/TaggerOne (với điểm số F1 là 0,17/, 38 và .31/.18, tư ơ ng ứng).



Hình 3: Điểm số F1 cho 18 bộ phân loại nhị phân dành riêng cho bệnh, dự đoán các bệnh liên quan đến các bài đăng trên Reddit hoàn toàn dựa trên các thực thể y tế của bài đăng và cũng dựa trên toàn bộ nội dung của bài đăng. Các thanh lỗi hiển thị độ lệch chuẩn cho xác thực chéo 5 lần. Đư ờng màu đỏ biểu thị đư ờng cơ sở phân loại ngẫu nhiên.

6 XÁC NHẬN: DỰ ĐOÁN BỆNH

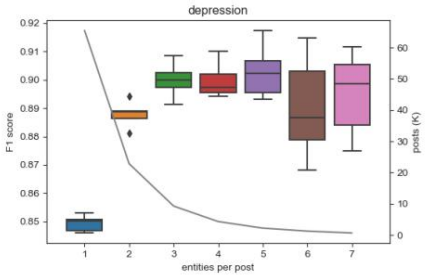
Các kết quả trư ớc đó cho thấy phư ơ ng pháp của chúng tôi hoạt động tốt trên ba bộ dữ liệu, mỗi bộ từ một nền tảng khác nhau. Tuy nhiên, chúng tôi phải thừa nhận rằng những kết quả này không mang tính kết luận vì chúng đư ợc tạo ra trên các tập hợp đư ợc dán nhãn phong phú như ng có kích thư ớc hạn chế. Với khả năng có sẵn của tập hợp lớn hơ n các bài đăng Reddit đư ợc phân loại thành 18 bệnh (Phần 4.3), chúng tôi chuyển sang nhiệm vụ dự đoán cuối cùng: đó là dự đoán bệnh của bài đăng Reddit từ tập hợp các thực thể y tế chứa trong đó. Chúng tôi kỳ vọng rằng nếu các thực thể đư ợc trích xuất theo phư ơ ng pháp của chúng tôi là chính xác, thì bộ phân loại sẽ có thể phân biệt các bài đăng thuộc các danh mục khác nhau, vì các triệu chứng và thuốc liên quan đến 18 bệnh là khác nhau. Trong thực tế, việc dự đoán bệnh chỉ từ một vài thực thể y tế mà một bài đăng nhất định phải chứa thay vì sử dụng toàn bộ nội dung văn bản của nó (như phần lớn công việc trư ớc đây đã thực hiện) có lợi ích là kết quả dự đoán không có xu hư ớng bị ảnh hư ởng bởi các mối tư ơ ng quan giả, khiến chúng mạnh mẽ hơ n đối với các sự kiện ngoại sinh và có khả năng khái quát hóa hơ n.

6.1 Thiết lập phân loại Chúng tôi đã

chọn mô hình hoạt động tốt nhất trong số các mô hình học sâu (nghĩa là sử dụng RoBERTa và GloVe).

Chúng tôi đã sử dụng MetaMap và phư ơ ng pháp đã chọn của chúng tôi để trích xuất các thực thể y tế từ các bài đăng trên Reddit (tạo hai nhóm thực thể y tế riêng biệt) và chỉ giữ lại các bài đăng trong đó tìm thấy ít nhất một thực thể y tế. Những bài đăng này sau đó đư ợc sắp xếp thành 18 bộ dữ liệu cân bằng, mỗi bộ cho một trong số 18 bệnh. Mỗi tập hợp bệnh chứa một tập hợp con các ví dụ tích cực (tất cả các bài đăng liên quan đến bệnh) và một tập hợp con các ví dụ tiêu cực (các bài đăng đư ợc lấy mẫu ngẫu nhiên từ 17 tập hợp bệnh còn lại).

Sau đó, chúng tôi đã đào tạo 36 bộ phân loại nhị phân XGBoost (với n = 1000 bộ ư ớc tính và độ sâu cây tối đa là 4): đối với mỗi bệnh trong số 18 bệnh, chúng tôi đã đào tạo ba bộ phân loại, một bộ dựa trên các thực thể y tế đư ợc trích xuất bởi MetaMap và bộ kia dựa trên các thực thể y tế đư ợc trích xuất bằng phư ơ ng pháp của chúng tôi.



Hình 4: Khả năng dự đoán liệu một bài đăng Reddit có liên quan đến trầm cảm hay không (điểm F1, trục y ở phía bên trái) tăng lên theo số lượng tổ chức y tế mà bài đăng đề cập. Đường màu xám biểu thị số lượng bài đăng (số lượng K bài đăng, trục y ở phía bên tay phải) chứa một số thực thể nhất định.

6.2 Số liệu phân loại

Đào tạo và kiểm tra được thực hiện bằng xác thực chéo 5 lần và độ chính xác được đo bằng điểm F 1. Đây là giá trị trung bình điều hòa của độ chính xác P của bộ phân loại và thu hồi R:

$$P = \frac{\#\text{bài đăng về bệnh được phân loại chính xác}}{\#\text{tổng số bài đăng được phân loại liên quan đến bệnh}},$$

Và

$$R = \frac{\#\text{bài viết phân loại đúng bệnh}}{\#\text{tổng hợp bài viết liên quan đến bệnh}}.$$

6.3 Bộ phân loại dựa trên từ điển (DIS-LIWC)

Do các phương pháp tiếp cận dựa trên từ điển (ví dụ: so khớp các từ trong LIWC - Truy vấn ngôn ngữ và Đếm từ - từ điển [52]) đã được sử dụng rộng rãi trong các nghiên cứu truyền thông xã hội và đã cho thấy hiệu quả tốt, nên chúng tôi đã thêm một phương pháp như vậy làm cơ sở. Chúng tôi đã tạo ra Dis-LIWC (Điều tra ngôn ngữ bệnh và đếm từ):8 một bộ gồm 1493 từ phản ánh các triệu chứng và tên thuốc, đồng thời được sắp xếp theo 18 loại bệnh. Chúng tôi đã làm như vậy bằng cách thu thập cho từng bệnh:

biểu thức y tế chính thức cho các triệu chứng. Chúng tôi đã thu thập các biểu hiện này từ hơn 100 nghìn cặp triệu chứng-bệnh từ bộ dữ liệu Mạng lưới bệnh tật ở người (HDN) [19], bao gồm các cặp thường được đề cập cùng nhau trong các ấn phẩm được lập chỉ mục bởi PubMed.

Biểu thức thông tục cho các triệu chứng. Chúng tôi đã viết thủ công tất cả các triệu chứng xuất hiện trong các trang mô tả chính của bệnh trên MedScape9 , WebMed10 và Wikipedia.

Tên thuốc. Chúng tôi đã thu thập thông tin tên thuốc và các bệnh tương ứng từ toàn bộ cơ sở dữ liệu của DrugBank11, dẫn đến tổng cộng hơn 100 tên.

6.4 Bộ phân loại dựa trên bài viết đầy đủ (FullPosts)

Người ta có thể tự hỏi mức độ mà một bộ phân loại lấy toàn bộ nội dung văn bản của bài đăng (không chỉ các thực thể y tế của bài đăng) là đầu vào sẽ có thể dự đoán bệnh của bài viết. Để xác định điều đó, chúng tôi đã tạo đường cơ sở FullPosts. Quá trình này đã mã hóa tất cả các bài đăng có ít hơn 512 mã thông báo chung (không nhất thiết là mã thông báo y tế) thành tài liệu nhúng Roberta+Glove và dẫn đến 18 bộ dữ liệu cân bằng và bộ phân loại nhị phân theo cách tương tự như cách thiết lập trong Phần 6.1. Chúng tôi mong đợi FullPosts trả về các dự đoán chính xác (vì nó hoạt động trên nhiều thực thể, không chỉ y tế), nhưng chúng tôi cũng mong đợi nó ít nguyên tắc hơn và do đó, ít khái quát hơn (ví dụ: nó có thể được đào tạo để liên kết từ "súp" với bệnh cúm).

6.5 Kết quả

Từ kết quả (Hình 3), chúng tôi thấy rằng, dựa trên đầu vào của các thực thể y tế được trích xuất bằng phương pháp của chúng tôi, người ta có thể dự đoán một cách đáng tin cậy tất cả 18 bệnh: đối với tất cả chúng, điểm F1 đều cao hơn 0,80. Những điểm số này thậm chí còn cao hơn một chút so với FullPosts' [38], sử dụng toàn văn có trong các bài đăng trên Reddit. Mặt khác, trên đầu vào của các thực thể y tế được trích xuất bởi MetaMap, người ta vẫn có thể dự đoán phần lớn các bệnh, tuy nhiên, các thực thể theo phương pháp của chúng tôi luôn liên quan đến độ chính xác dự đoán tăng từ 15% đến 20%. Ngược lại, trên đầu vào của các thực thể được Dis-LIWC trích xuất, hầu hết các bệnh không thể được xác định. Tất cả những kết quả này gợi ý rằng: (1) Hành vi dự kiến của Dis-LIWC là trích xuất tất cả các thực thể phù hợp với nội dung từ điển của nó. Tuy nhiên, những trích xuất như vậy hóa ra lại có ít khả năng dự đoán hơn so với những trích xuất từ phương pháp của chúng tôi, vốn không dựa trên từ vựng chuyên ngành. (2) Yêu cầu mức độ chính xác cao để dự đoán bệnh, vì dự đoán được thực hiện ở cấp độ hậu kỳ, có thể chỉ từ một vài tổ chức y tế. Đó là bởi vì, như người ta mong đợi, khả năng dự đoán bệnh của một bài đăng tăng theo số lượng thực thể được tìm thấy trong bài đăng (Hình 4).

6.6 Các yếu tố ảnh hưởng đến dự đoán bệnh Điểm F1 khác nhau

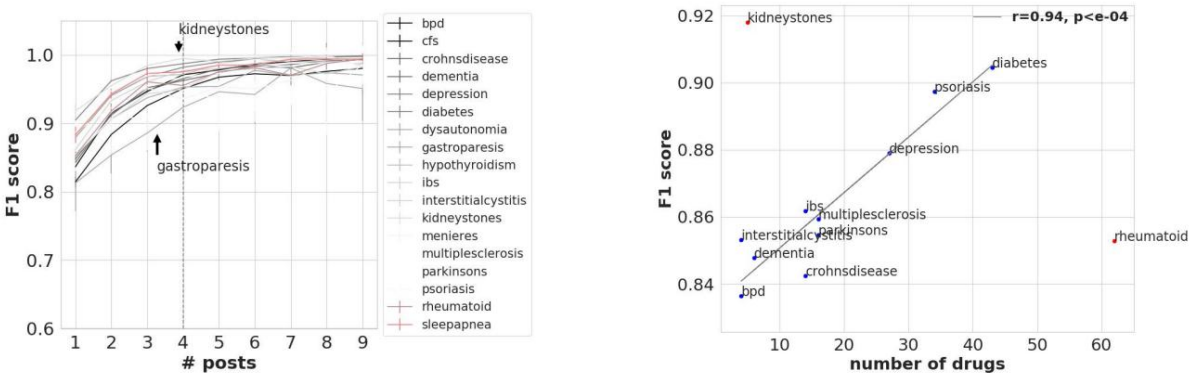
giữa các bệnh (Hình 3): có sự khác biệt gần 10% giữa bệnh có điểm cao nhất (sỏi thận) và bệnh có điểm thấp nhất (liệt dạ dày). Bây giờ người ta có thể tự hỏi những yếu tố nào giải thích sự thay đổi đó.

Kích thước của dữ liệu đào tạo. Người ta có thể kỳ vọng rằng càng nhiều dữ liệu huấn luyện cho một bệnh thì càng dễ dự đoán. Chúng tôi đã so sánh F1 của một bệnh với logarit của số lượng bài đăng liên quan đến nó. Không có mối tương quan ($r = 0,006$, $p > 0,98$). Thật vậy, ngay cả đối với căn bệnh có số lượng bài đăng thấp nhất (r/menieres với 613 bài đăng), chúng tôi có điểm F1 là 0,89.

Kích thước đầu vào từ 1 đến một bộ n bài đăng. Để đánh giá tác động của kích thước đầu vào đối với dự đoán bệnh, chúng tôi cũng đã đào tạo 7 phiên bản khác nhau của 18 bộ phân loại bệnh nhị phân với kích thước của đơn vị đầu vào tăng dần, từ n = 1 đến n = 7 bài đăng. N bài đăng trong mỗi đơn vị đào tạo / kiểm tra được lấy mẫu ngẫu nhiên mà không cần thay thế. Thực sự dễ dự đoán một bệnh trên đầu vào của 2 bài đăng hơn là 1 (bảng điều khiển bên trái trong Hình 5) và chỉ với n = 4 bài đăng, tất cả các bộ phân loại đều vượt quá điểm F 1 là 0,90.

Số triệu chứng/thuốc cho từng bệnh. Việc phân loại của chúng tôi được thực hiện dựa trên thông tin đầu vào của các tổ chức y tế, bao gồm các triệu chứng và tên thuốc. Người ta có thể mong đợi rằng số lượng tên thuốc hoặc triệu chứng liên quan đến bệnh càng nhiều thì bệnh càng được phân loại độc đáo. Để kiểm tra điều đó, chúng tôi lấy số triệu chứng và số tên thuốc liên quan đến từng triệu chứng.

8Dis-LIWC được cung cấp công khai tại http://goodcitylife.org/Humane-AI 9https://www.medscape.com 10https://www.webmd.com 11https://www.drugbank.ca



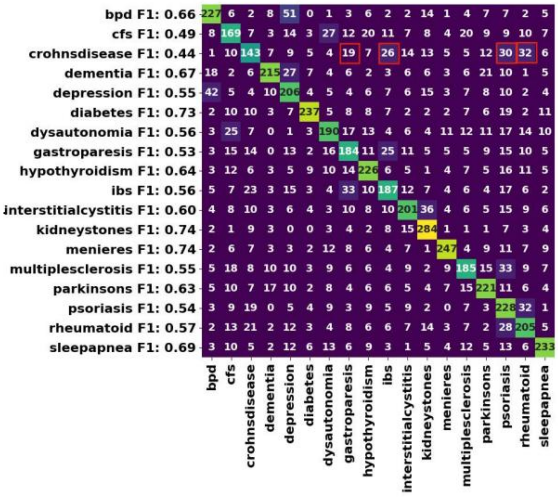
Hình 5: Khả năng dự đoán bệnh (điểm F1) tùy thuộc vào: (trái) số lư ợng bài viết làm đơ n vị đầu vào; và (phải) mức độ mà một căn bệnh có liên quan đến một số lư ợng lớn các loại thuốc.

bệnh (từ DIS-LIWC của chúng tôi trong Phần 6.3), và tính toán mối tư ơ ng quan của chúng với điểm số F 1 của từng bệnh. Như chúng tôi đã đư a ra giả thuyết, một bệnh có liên quan đến số lư ợng triệu chứng càng cao thì càng dễ dự đoán (r = 0,59, p < 0,01); theo cách tư ơ ng tự, số lư ợng tên thuốc liên quan đến một bệnh12 càng cao thì càng dễ dự đoán (bảng bên phải trong Hình 5), với mối tư ơ ng quan cao tới r = 0,94 (p < 0,0001). Hệ quả tất yếu của kết quả này là các bệnh thông thư ờng (có xu hư ớng đư ợc điều trị bằng nhiều loại thuốc)13 dễ bị phát hiện trên mạng xã hội hơ n những bệnh ít phổ biến hơ n.

Thú vị hơ n, có hai trư ờng hợp ngoại lệ đối với giả thuyết của chúng tôi (hai chấm đỏ ở bảng bên phải của Hình 5):

(1) Viêm khớp dạng thấp đư ợc điều trị bằng nhiều loại thuốc như ng tư ơ ng đối khó phân loại. Tình trạng này đư ợc biết là khó chẩn đoán vì các triệu chứng của nó có liên quan đến nhiều bệnh khác.14
(2) Sỏi thận chỉ

đư ợc điều trị bằng một số loại thuốc như ng tư ơ ng đối dễ phân loại. Đó là bởi vì nó liên quan đến một cơ quan trong cơ thể và đi kèm với các triệu chứng đư ợc xác định rõ ràng và các loại thuốc cụ thể. Điều này phân biệt sỏi thận với bất kỳ tình trạng nào khác.



Hình 6: Ma trận nhầm lẫn cho bộ phân loại nhiều lớp dự đoán 18 bệnh liên quan đến các bài đăng trên Reddit hoàn toàn dựa trên các thực thể y tế đư ợc trích xuất theo phư ơ ng pháp của chúng tôi.

Triệu chứng chung giữa các bệnh. Để kiểm tra những bệnh nào khó phân biệt với nhau hơ n, ngoài việc có một bộ phân loại nhị phân cho từng bệnh trong số 18 bệnh, chúng tôi còn có một bộ phân loại đa lớp duy nhất cho 18 bệnh. Để tạo tập dữ liệu cân bằng, đối với mỗi subreddit, chúng tôi lấy ngẫu nhiên số lư ợng bài đăng bằng với số lư ợng bài đăng của subreddit nhỏ nhất. Ma trận nhầm lẫn trong Hình 6 báo cáo điểm F1 cho một bệnh trên mỗi hàng (đư ờng cơ sở ngẫu nhiên trong trư ờng hợp nhiều lớp có điểm F1 là 0,06), đư ờng chéo của nó tư ơ ng ứng với thời điểm bệnh i đư ợc dự đoán chính xác và (i, j) phần tử báo số lư ợng bệnh j bài đăng sai nhãn asi. Hãy xem xét bệnh Crohn (viêm ruột), có điểm F1 thấp nhất là 0,44, vẫn cao hơ n nhiều so với điểm cơ bản là 0,06. Thay vào đó , hầu hết các bài đăng bị dán nhãn sai là của Crohn đề cập đến hai bệnh đư ờng ruột khác - "hội chứng viêm ruột" (26 bài) hoặc "liệt dạ dày" (19 bài); một căn bệnh mà nó chia sẻ các cơ chế sinh học cơ bản đư ợc gọi là "bệnh vẩy nến" (30 bài đăng) [15]; hoặc "viêm khớp dạng thấp" (32 bài viết), là một biến chứng của bệnh Crohn ngoài đư ờng tiêu hóa. Với kết quả này, chúng tôi đư a ra giả thuyết rằng hai bệnh khó phân biệt với nhau, nếu chúng có xu hư ớng chia sẻ các triệu chứng. Để kiểm tra điều đó, chúng tôi đã tính toán số lư ợng các triệu chứng đư ợc chia sẻ cho mỗi cặp bệnh (D1,D2): JD1,D2 = trong đó $\frac{SD1 \cap SD2}{SD1 \cup SD2}$, SD1 và SD2) là tập hợp các triệu chứng của hai bệnh (đư ợc tính từ danh sách triệu chứng Dis-LIWC trong Phần 6.3), và J là chỉ số Jaccard (chỉ số tư ơ ng tự) của hai bộ. Như chúng tôi đã đư a ra giả thuyết, có một mối tư ơ ng quan thuận và có ý nghĩa thống kê (r = 0,31) giữa

12Khi liên hệ tên thuốc với các bệnh trong Mục 6.3, chúng ta có thể tìm thấy tên thuốc của 12 bệnh trong tổng số 18 bệnh và như vậy, Hình 5 (bên phải) sẽ hiển thị kết quả cho 12 bệnh này .

13Các công ty dư ợc nhắm vào các bệnh có lợi nhuận https://www.focusforhealth. org/big-pharma-creates-diseases-medications-big-business 14https://www.nhs.uk/conditions/rheumatoid-arthritis/diagnosis

J (sự giống nhau giữa hai bệnh về số lượng triệu chứng chung) và phân loại sai.

7 THẢO LUẬN

Việc áp dụng trích xuất thực thể y tế cho phương tiện truyền thông xã hội khó khăn hơn so với việc áp dụng nó cho hồ sơ sức khỏe điện tử. Điều này là do cách người dùng thể hiện bản thân khác thường, có thể mắc lỗi chính tả và sử dụng tiếng lóng trên internet [63]. Khả năng học sâu đã được chứng minh để trích xuất các triệu chứng trong bối cảnh truyền thông xã hội này, được cho là có nhiều dữ liệu hơn và cũng dễ tiếp cận hơn, có nhiều ý nghĩa lý thuyết và thực tiễn. Tuy nhiên, trước khi có thể nhận ra những hàm ý đó, các nhà nghiên cứu phải giải quyết hai hạn chế chính.

Chú thích y tế là tốn kém. Các thuật toán học máy cần dữ liệu và trong trường hợp cụ thể là các ứng dụng y tế, việc tạo chú thích rất tốn kém và khó khăn: sử dụng chuyên gia chú thích rất tốn kém và nhân viên đám đông có thể không phải là ứng cử viên tốt nhất cho các chú thích y tế chuyên môn cao. Cần có các phương pháp tiếp cận cộng đồng mới giúp giảm thiểu chi phí chú thích mà không ảnh hưởng đến chất lượng. Cuối cùng, mức độ mà các kỹ thuật trò chơi hóa có thể được giới thiệu trong lĩnh vực chăm sóc sức khỏe nên được khám phá trong tương lai [13].

Sự khó hiểu của phương tiện truyền thông xã hội. Vẫn còn nhiều chỗ để cải thiện độ chính xác. Để bắt đầu, kết quả của chúng tôi cho thấy rằng việc sử dụng ngôn ngữ theo nghĩa bóng hiện diện trên các nền tảng truyền thông xã hội và phương pháp của chúng tôi cần được cải thiện hơn nữa để giải quyết vấn đề đó. Ngoài ra, nghiên cứu này chưa xử lý thư rác hoặc nội dung độc hại và việc tích hợp các kỹ thuật lọc nội dung, ví dụ, dựa trên tính nhất quán theo chủ đề hoặc độ tin cậy của người tạo nội dung có thể nâng cao hiệu suất [64].

Ý nghĩa lý thuyết. Chúng tôi đã xác định được những hàm ý lý thuyết tức thì trong hai lĩnh vực y học:

(1) Mạng lưới bệnh tật ở người kiểu hình. Trong loại mạng này, các nút là bệnh và trọng số liên kết phản ánh mức độ mà các cặp bệnh tương ứng có cùng triệu chứng. Những cách mới để trích xuất các triệu chứng từ phương tiện truyền thông xã hội, chẳng hạn như phương pháp được trình bày ở đây sẽ dẫn đến việc tạo ra các mạng lưới bệnh tật kiểu hình mới ở người, có lẽ mở ra một lĩnh vực mới có thể được gọi là 'mạng lưới bệnh tật kiểu hình xã hội'. Với các mạng kiểu hình đa dạng và các phương pháp do cộng đồng nghiên cứu 'mạng phức hợp' phát triển, những người nghiên cứu lại có thể có 'cái nhìn mới mẻ' về sự tương tác phức tạp giữa các triệu chứng và bệnh tật .
(2) Nghiên cứu di truyền. Những nghiên cứu như vậy ngày càng có khả năng mô tả các bệnh có mối liên hệ di truyền phổ biến. Đặc điểm hơn nữa có thể dựa trên các triệu chứng được chia sẻ bởi bệnh . Hai bài báo gần đây đã đạt được điều này bằng cách khai thác tóm tắt các ấn phẩm nghiên cứu y khoa [25, 75]. Khi làm như vậy, họ phát hiện ra rằng thực sự “sự giống nhau dựa trên triệu chứng của các bệnh tương quan chặt chẽ với số lượng các mối liên hệ di truyền được chia sẻ và mức độ mà các protein liên quan của chúng tương tác”. Khả năng trích xuất các triệu chứng từ phương tiện truyền thông xã hội của học sâu có thể đóng góp thêm cho các nghiên cứu di truyền.

Ý nghĩa thực tiễn. Phương pháp của chúng tôi được đánh giá cho 18 bệnh nhưng bệnh nào cũng áp dụng được. Đó là bởi vì loại nhúng cụ thể cho phép nhận ra các đề cập không có trong dữ liệu huấn luyện. Để xem làm thế nào, chúng ta hãy kể tên một vài ví dụ trong số rất nhiều ví dụ mà chúng tôi gặp phải: ' mất thính giác thần kinh cảm giác tần số trung bình', 'tiếng tim đập thình thịch trong tai', 'đường không giảm', 'cơ thể thay đổi đột ngột'. lượng đường trong máu', 'vai tôi đặt quá cao trên bàn cho đến khi cánh tay tôi bắt đầu tê liệt', mà còn cả những từ viết tắt, chẳng hạn như 'DKA' (đối với nhiễm toan ceton do tiểu đường) và 'OCD' đối với (rối loạn ám ảnh cưỡng chế). Với khả năng khái quát hóa của nó, phương pháp của chúng tôi có thể được sử dụng cho:

(1) Cảnh giác dược. Cảnh giác dược đòi hỏi khả năng xác định một loạt các triệu chứng cụ thể được gọi là 'Phản ứng có hại của thuốc ' (ADRs). Vì phương pháp của chúng tôi có thể trích xuất các triệu chứng ở cấp độ chi tiết của bài đăng, nên nó có thể được sử dụng cho cảnh giác phar và ứng dụng của nó trên phương tiện truyền thông xã hội cũng có thể dẫn đến việc phát hiện ra các ADR chưa biết.
(2) Theo dõi bệnh theo thời gian và không gian. Nghiên cứu đáng chú ý đã chỉ ra rằng phương tiện truyền thông xã hội có thể được sử dụng để theo dõi bệnh theo thời gian và trên các khu vực địa lý bằng cách khai thác các bài đăng đi kèm với vị trí và dấu thời gian [47, 57]. Ở đây chúng tôi đã đề xuất một phương pháp mới để xác định bệnh – một phương pháp dựa trên việc trích xuất các triệu chứng và dựa vào đó để dự đoán căn bệnh đang được thảo luận. Cách xác định bệnh này là nguyên tắc: trái ngược với các phương pháp hiện có, bằng cách hạn chế các từ được trích xuất đối với các thực thể y tế, nó không gây ra sự nhầm lẫn giữa các từ phi y tế và các bệnh. Đây là lý do chính khiến Google chấm dứt dự án 'xu hướng cúm', trong đó các đợt bùng phát dịch cúm được theo dõi từ các tìm kiếm của mọi người – bằng cách xem xét bất kỳ từ nào hữu ích để dự đoán, hệ thống cuối cùng đã không thể mạnh mẽ đối với các sự kiện ngoại sinh [32].
(3) Tái sử dụng thuốc. Bằng cách khai thác các triệu chứng từ các bài đăng trên mạng xã hội đề cập đến tên thuốc cụ thể (ví dụ: từ các diễn đàn đánh giá thuốc), các công ty dược phẩm có thể khám phá ra các ứng cử viên tiềm năng cho cái mà trong ngành gọi là 'tái sử dụng thuốc', nghĩa là xác định thêm bệnh/triệu chứng nào có thể được điều trị bằng các loại thuốc hiện đang được kê đơn cho các tình trạng khác .

8 KẾT LUẬN

Chúng tôi đã trình bày một khung học sâu có thể áp dụng rộng rãi để trích xuất các thực thể y tế một cách đáng tin cậy như các triệu chứng và tên thuốc cũng như để dự đoán chính xác các bệnh hoàn toàn từ các thực thể y tế được trích xuất. Bằng cách đánh giá nó trên ba bộ dữ liệu có nguồn gốc từ AskaPatient, Twitter và Reddit, chúng tôi đã chỉ ra rằng nó luôn vượt trội so với các phương pháp cơ bản và hiện đại, cho thấy kết quả có thể khái quát hóa. Trong tương lai, cần nghiên cứu thêm về: các nỗ lực thu thập dữ liệu mới, bao gồm thiết kế các giải pháp tìm nguồn cung ứng đám đông mới phù hợp với lĩnh vực y tế chuyên môn cao; các kỹ thuật khai thác văn bản có khả năng xử lý việc sử dụng ngôn ngữ theo nghĩa bóng; và các kỹ thuật khai thác phương tiện truyền thông xã hội có thể lọc nội dung độc hại và không chính xác theo những cách mạnh mẽ. Tất cả công việc này có thể cho phép các ứng dụng theo dõi sức khỏe quy mô lớn, dẫn đến việc xây dựng các mạng lưới bệnh tật kiểu hình mới ở người và thậm chí tác động đến các nghiên cứu di truyền.

NGƯ ỜI GIỚI THIỆU

[1] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter và Roland Vollgraf. 2019. FLAIR: Khung dễ sử dụng cho NLP hiện đại. Trong Kỷ yếu Hội nghị của Chư ơ ng Bắc Mỹ của Hiệp hội Ngôn ngữ học Tính toán. 54–59.

[2] Alan Akbik, Tanja Bergmann, và Roland Vollgraf. 2019. Các nội dung nhúng theo ngữ cảnh đư ợc tổng hợp cho Nhận dạng thực thể đư ợc đặt tên. Trong Kỷ yếu Hội nghị của Chư ơ ng Bắc Mỹ của Hiệp hội Ngôn ngữ học Máy tính: Công nghệ Ngôn ngữ Con ngư ời, Tập 1. 724–728.

[3] Alan Akbik, Duncan Blythe, và Roland Vollgraf. 2018. Bộ đệm chuỗi theo ngữ cảnh để ghi nhãn trình tự. Trong Kỷ yếu Hội nghị Quốc tế về Ngôn ngữ học Tính toán của Hiệp hội Ngôn ngữ học Tính toán. 1638–1649.

[4] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann và Matthew McDermott. 2019. Các nhúng BERT lâm sàng có sẵn công khai. Trong Kỷ yếu của Hội thảo Xử lý Ngôn ngữ Tự nhiên Lâm sàng. 72–78.

[5] Alan R Aronson và François-Michel Lang. 2010. Tổng quan về MetaMap: quan điểm lịch sử và những tiến bộ gần đây. Tạp chí của Hiệp hội Tin học Y tế Hoa Kỳ 17, 3 (2010), 229–236.

[6] Vincent Van Asch. 2013. Các biện pháp đánh giá trung bình vĩ mô và vi mô. công nghệ Báo cáo chính thức (2013).

[7] Duilio Balsamo, Paolo Bajardi và André Panisson. 2019. Lạm dụng thuốc phiện trực tiếp trên mạng xã hội: Theo dõi các mẫu sở thích về không gian địa lý thông qua một nhóm kỹ thuật số. Trong Kỷ yếu của Hội nghị ACM World Wide Web. 2572–2579.

[8] Tianqi Chen và Carlos Guestrin. 2016. XGBoost: Hệ thống tăng cư ờng cây có thể mở rộng. Trong Kỷ yếu của Hội nghị ACM SIGKDD về Khám phá Tri thức và Khai thác Dữ liệu. 785–794.

[9] Munmun De Choudhury và Sushovan De. 2014. Bài diễn văn về Sức khỏe Tâm thần trên reddit: Tiết lộ Bản thân, Hỗ trợ Xã hội và Ẩn danh. Trong Kỷ yếu của Hội nghị AAAI Quốc tế về Weblog và Truyền thông Xã hội.

[10] Aaron M Cohen và William R Hersh. 2005. Một cuộc khảo sát về công việc hiện tại trong khai thác văn bản y sinh. Briefings in bioinformatics 6, 1 (2005), 57–71.

[11] Kerstin Denecke. 2014. Trích xuất các khái niệm y tế từ phư ơ ng tiện truyền thông xã hội y tế bằng các công cụ NLP lâm sàng: một nghiên cứu định tính. Trong Kỷ yếu Hội thảo về Xây dựng và Đánh giá Tài nguyên cho Xử lý Văn bản Y tế và Y sinh.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, và Kristina Toutanova. 2019. BERT: Đào tạo trư ớc về Máy biến áp hai chiều sâu để hiểu ngôn ngữ. Trong Kỷ yếu Hội nghị của Chư ơ ng Bắc Mỹ của Hiệp hội Ngôn ngữ học Máy tính: Công nghệ Ngôn ngữ Con ngư ời, Tập 1. 4171–4186.

[13] Emilia Duarte, Pedro Pereira, Francisco Rebelo và Paulo Noriega. 2014. Đánh giá về Gamification cho các bối cảnh liên quan đến sức khỏe. Trong Kỷ yếu của Hội nghị Quốc tế lần thứ 3 về Thiết kế, Trải nghiệm Ngư ời dùng và Khả năng Sử dụng. 742–753.

[14] Tara Fenwick. 2014. Truyền thông xã hội và chuyên môn y tế: suy nghĩ lại về cuộc tranh luận và con đư ờng phía trư ớc. Học viện Y học 89, 10 (2014), 1331–1334.

[15] Gionata Fiorino và Paolo D Omodei. 2015. Bệnh vẩy nến và bệnh viêm ruột : Hai mặt của cùng một đồng tiền? Tạp chí Bệnh Crohn & Viêm đại tràng 9, 9 (2015), 697–698.

[16] Jerome H Friedman. 2001. Xấp xỉ hàm tham lam: máy tăng độ dốc. Biên niên sử thống kê (2001), 1189–1232.

[17] Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, và Jyotishman Pathak. 2019. Đánh giá nhận thức về mức độ nghiêm trọng của nguy cơ tự tử để can thiệp sớm. Trong Kỷ yếu của Hội nghị ACM World Wide Web. 514–525.

[18] George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim JP Hub bard, Richard JB Dobson, và Rina Dutta. 2017. Đặc điểm của tình trạng sức khỏe tâm thần trên mạng xã hội bằng cách sử dụng Học sâu có hiểu biết. Báo cáo khoa học 7 (2017), 45141.

[19] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, và Albert-László Barabási. 2007. Mạng lư ới bệnh tật ở ngư ời. Kỷ yếu của Viện Hàn lâm Khoa học Quốc gia 104, 21 (2007), 8685–8690.

[20] G Gonzalez-Hernandez, A Sarker, K O'Connor, và G Savova. 2017. Nắm bắt quan điểm của bệnh nhân: Đánh giá về những tiến bộ trong xử lý ngôn ngữ tự nhiên của văn bản liên quan đến sức khỏe. Niên giám Tin học Y tế 26, 01 (2017), 214–227.

[21] Frances Griffiths, Jonathan Cave, Felicity Boardman, Justin Ren, Teresa Paw likowska, Robin Ball, Aileen Clarke và Alan Cohen. 2012. Mạng xã hội – tư ơ ng lai của việc cung cấp dịch vụ chăm sóc sức khỏe. Khoa học Xã hội & Y học 75, 12 (2012), 2233–2241.

[22] Carleen Hawn. 2009. Uống hai viên aspirin và tweet cho tôi vào buổi sáng: Twitter, Facebook và các phư ơ ng tiện truyền thông xã hội khác đang định hình lại dịch vụ chăm sóc sức khỏe như thế nào. Vấn đề sức khỏe 28, 2 (2009), 361–368.

[23] Matthew Herland, Taghi M Khoshgoftaar, và Randall Wald. 2014. Đánh giá về khai thác dữ liệu sử dụng dữ liệu lớn trong tin học y tế. Tạp chí Dữ liệu lớn 1, 1 (2014), 2.

[24] William R. Hersh. 2002. Tin học y tế: cải thiện chăm sóc sức khỏe thông qua thông tin. Jama 288, 16 (2002), 1955–1958.

[25] Robert Hoehndorf, Paul N Schofield, và Georgios V Gkoutos. 2015. Phân tích bệnh ở ngư ời sử dụng sự giống nhau về kiểu hình giữa gen chung, gen, và các bệnh truyền nhiễm. Báo cáo khoa học 5 (2015), 10888.

[26] Zhiheng Huang, Wei Xu, và Kai Yu. 2015. Các mô hình LSTM-CRF hai chiều để gắn thẻ trình tự. bản in trư ớc arXiv arXiv:1508.01991 (2015).

[27] Antonio Jimeno-Yepes, Andrew MacKinlay, Bo Han, và Qiang Chen. 2015. Xác định Bệnh tật, Thuốc và Triệu chứng trên Twitter. Nghiên cứu Công nghệ Y tế và Tin học 216 (2015), 643.

[28] Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, và Chen Wang. 2015. Cadec: Tập hợp các chú thích về tác dụng phụ của thuốc. Tạp chí Tin học Y sinh 55 (2015), 73–81.

[29] Payam Karisani và Eugene Agichtein. 2018. Bạn có thực sự vừa lên cơ n đau tim không?: hư ớng tới việc phát hiện mạnh mẽ các đề cập về sức khỏe cá nhân trên mạng xã hội. Trong Kỷ yếu của Hội nghị ACM World Wide Web. Ban Chỉ đạo Hội nghị World Wide Web Quốc tế , 137–146.

[30] Nolan Lawson, Kevin Eustice, Mike Perkowitz, và Meliha Yetisgen-Yildiz. 2010. Chú thích bộ dữ liệu email lớn để nhận dạng thực thể đư ợc đặt tên bằng Mechanical Turk. Trong Kỷ yếu Hội thảo ACM NAACL HLT về Tạo dữ liệu giọng nói và ngôn ngữ với Mechanical Turk của Amazon. 71–79.

[31] Allison J Lazard, Emily Scheinfeld, Jay M Bernhardt, Gary B Wilcox, và Melissa Suran. 2015. Phát hiện các chủ đề đư ợc công chúng quan tâm: một phân tích khai thác văn bản của cuộc trò chuyện Twitter trực tiếp về Ebola của Trung tâm Kiểm soát và Phòng ngừa Dịch bệnh. Tạp chí Kiểm soát Nhiễm trùng Hoa Kỳ 43, 10 (2015), 1109–1111.

[32] David Lazer, Ryan Kennedy, Gary King, và Alessandro Vespignani. 2014. Câu chuyện ngụ ngôn về Google Flu: bẫy trong phân tích dữ liệu lớn. Khoa học 343, 6176 (2014), 1203–1205 .

[33] Robert Leaman và Zhiyong Lu. 2016. TaggerOne: nhận dạng và chuẩn hóa thực thể đư ợc đặt tên chung với Mô hình bán Markov. Tin sinh học 32, 18 (2016), 2839–2846.

[34] Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, và Graciela Gonzalez. 2010. Hư ớng tới cảnh giác dư ợc thời đại internet: trích xuất các phản ứng có hại của thuốc từ các bài đăng của ngư ời dùng lên các mạng xã hội liên quan đến sức khỏe. Trong Pro cceedings of the Worshop về Xử lý ngôn ngữ tự nhiên y sinh. Hiệp hội Ngôn ngữ học Tính toán, 117–125.

[35] Hsin-Chun Lee, Yi-Yu Hsu, và Hung-Yu Kao. 2015. Một hệ thống dựa trên CRF nâng cao để nhận dạng và chuẩn hóa thực thể tên bệnh trong Nhiệm vụ BioCreative V DNER. Trong Kỷ yếu Hội thảo Đánh giá Thách thức Sáng tạo Sinh học. 226–233.

[36] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, và Jaewoo Kang. 2019. BioBERT: mô hình biểu diễn ngôn ngữ y sinh đư ợc đào tạo trư ớc để khai thác văn bản y sinh. Tin sinh học (2019).

[37] Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, và Zhiyong Lu. 2016. Kho tác vụ BioCreative V CDR: nguồn tài nguyên để khai thác mối quan hệ bệnh hóa học. Cơ sở dữ liệu 2016 (2016).

[38] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, và Veselin Stoyanov. 2019. RoBERTa: Phư ơ ng pháp tiếp cận trư ớc đào tạo BERT đư ợc tối ư u hóa mạnh mẽ. bản in trư ớc arXiv arXiv:1907.11692 (2019).

[39] Yingjie Lu, Yang Wu, Jingfang Liu, Jia Li, và Pengzhu Zhang. 2017. Hiểu rõ việc sử dụng phư ơ ng tiện truyền thông xã hội chăm sóc sức khỏe từ các quan điểm của các bên liên quan khác nhau: phân tích nội dung của một cộng đồng y tế trực tuyến. Tạp chí nghiên cứu Internet y tế 19, 4 (2017).

[40] Andrew MacKinlay, Antonio Jimeno Yepes, và Bo Han. 2015. Xác định và phân tích các trư ờng hợp xảy ra cùng lúc với thực thể y tế trên Twitter. Trong Kỷ yếu Hội thảo Quốc tế ACM về Khai thác Dữ liệu và Văn bản trong Tin học Y sinh. 22–22.

[41] Zulfat Miftahutdinov, Elena Tutubalina, và Alexander Tropsha. 2017. Xác định các Biểu hiện liên quan đến Dịch bệnh trong Đánh giá bằng Trư ờng ngẫu nhiên có điều kiện. Trong Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog, Vol. 1. 155–167.

[42] Zulfat Miftahutdinov, Elena Tutubalina, và Alexander Tropsha. 2017. Xác định các Biểu hiện liên quan đến Dịch bệnh trong Đánh giá bằng Trư ờng ngẫu nhiên có điều kiện. Trong Kỷ yếu Đối thoại Hội nghị Quốc tế, Vol. 1. 155–167.

[43] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, và Jeff Dean. 2013. Biểu diễn phân tán của các từ và cụm từ và thành phần của chúng. Trong Kỷ yếu hội nghị về những tiến bộ trong hệ thống xử lý thông tin thần kinh. 3111–3119.

[44] S Anne Moorhead, Diane E Hazlett, Laura Harrison, Jennifer K Carroll, Anthea Irwin, và Ciska Hoving. 2013. Một khía cạnh mới của chăm sóc sức khỏe: đánh giá có hệ thống về việc sử dụng, lợi ích và hạn chế của phư ơ ng tiện truyền thông xã hội đối với truyền thông sức khỏe. Tạp chí nghiên cứu Internet y tế 15, 4 (2013).

[45] Ramona Nelson và Nancy Staggers. 2016. Tin học y tế: Cách tiếp cận liên ngành. Khoa học sức khỏe Elsevier.

[46] Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor, Rachel Ginn, và Graciela Gonzalez. 2015. Cảnh giác dư ợc từ phư ơ ng tiện truyền thông xã hội: khai thác để cập đến phản ứng có hại của thuốc bằng cách sử dụng ghi nhãn trình tự với các tính năng cụm nhúng từ. Tạp chí của Hiệp hội Tin học Y tế Hoa Kỳ 22, 3 (2015), 671–681.

[47] Công viên Albert và Mike Conway. 2017. Theo dõi các cuộc thảo luận liên quan đến sức khỏe trên Reddit cho các ứng dụng sức khỏe cộng đồng. Trong Kỷ yếu hội thảo thư ờng niên của AMIA,

tập 2017. Hiệp hội Tin học Y tế Hoa Kỳ, 1362.

[48] Albert Park, Mike Conway, và Annie T Chen. 2018. Kiểm tra sự giống nhau , khác biệt và tư cách thành viên theo chủ đề trong ba cộng đồng sức khỏe tâm thần trực tuyến từ Reddit: một cách tiếp cận trực quan và khai thác văn bản. Máy tính trong hành vi của con ngư ời 78 (2018), 98–112.

[49] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, và Adam Lerer. 2017. Tự động phân biệt trong PyTorch. Trong Kỷ yếu về những tiến bộ trong Hội thảo Autodiff của Hệ thống xử lý thông tin thần kinh.

[50] Michael J. Paul và Mark Dredze. 2011. You are what you Tweet: Phân tích Twitter vì sức khỏe cộng đồng. Kỷ yếu Hội nghị AAAI Quốc tế về Web và Truyền thông Xã hội 20 (2011), 265–272.

[51] Michael J Paul, Abeed Sarker, John S Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen L Smith, và Graciela Gonzalez. 2016. Khai thác phư ơng tiện truyền thông xã hội để theo dõi và giám sát sức khỏe cộng đồng. Trong Điện toán sinh học: Kỷ yếu của hội nghị chuyên đề Thái Bình Dư ơng. 468–479.

[52] James W Pennebaker, Martha E Francis, và Roger J Booth. 2001. Truy vấn ngôn ngữ và đếm từ: LIWC 2001. Mahway: Lawrence Erlbaum Associates 71 (2001).

[53] Jeffrey Pennington, Richard Socher, và Christopher Manning. 2014. GloVe: Các vectơ toàn cục để biểu diễn từ. Trong Kỷ yếu hội thảo về các phư ơng pháp thực nghiệm trong xử lý ngôn ngữ tự nhiên. Hiệp hội Ngôn ngữ học Tính toán, 1532–1543.

[54] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, và Luke Zettlemoyer. 2018. Biểu thị từ đư ợc ngữ cảnh hóa sâu sắc . Trong Kỷ yếu Hội nghị Thư ờng niên của Chư ơng Bắc Mỹ của Hiệp hội Ngôn ngữ học Tính toán. 2227–2237.

[55] Lance A Ramshaw và Mitchell P Marcus. 1995. Phân đoạn văn bản sử dụng học tập dựa trên chuyển đổi. CoRR. arXiv in sẵn cmp-lg/9505040 50 (1995).

[56] Müi khoan Frederic G. 2015. Công tác xã hội lâm sàng trong môi trư ờng kỹ thuật số: Những thách thức về đạo đức và quản lý rủi ro. Tạp chí Công tác xã hội lâm sàng 43, 2 (2015), 120–132.

[57] Abeed Sarker, Karen O'Connor, Rachel Ginn, Matthew Scotch, Karen Smith, Dan Malone, và Graciela Gonzalez. 2016. Khai thác phư ơng tiện truyền thông xã hội để cảnh giác với chất độc: giám sát tự động việc lạm dụng thuốc theo toa từ Twitter. An toàn thuốc 39, 3 (2016), 231–240.

[58] Daniel Scanfeld, Vanessa Scanfeld, và Elaine L Larson. 2010. Phổ biến thông tin y tế qua mạng xã hội: Twitter và thuốc kháng sinh. Tạp chí kiểm soát nhiễm trùng Hoa Kỳ 38, 3 (2010), 182–188.

[59] Sanja Scepanovic, Enrique Martin-Lopez, và Daniele Quercia. 2020. MedRed. https://doi.org/10.7910/DVN/8YVINU [60] Wendy Sinclair,

Moira McLoughlin và Tony Warne. 2015. Đến Twitter để thu hút: Khai thác sức mạnh của (một số) phư ơng tiện truyền thông xã hội trong giáo dục y tá để nâng cao trải nghiệm của học sinh. Thực hành giáo dục y tá 15, 6 (2015), 507–511.

[61] Gabriel Stanovsky, Daniel Gruhl, và Pablo Mendes. 2017. Ghi nhận các đề cập về phản ứng có hại của thuốc trên mạng xã hội bằng cách sử dụng các mô hình lặp lại đư ợc truyền tải kiến thức. Trong Kỷ yếu Hội nghị của Chư ơng Châu Âu của Hiệp hội Ngôn ngữ học Tính toán, Vol. 1. 142–151.

[62] Jana Straková, Milan Straka, và Jan Hajic. 2019. Kiến trúc thần kinh cho NER lồng nhau thông qua tuyến tính hóa. Trong Kỷ yếu Hội nghị của Hiệp hội Ngôn ngữ học Tính toán. 5326–5331.

[63] Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, và Wei Xu. 2016. Kết quả của nhiệm vụ chia sẻ WNUT16 Named Entity Recognition. Trong Kỷ yếu Hội thảo ACM về Văn bản ồn ào do ngư ời dùng tạo. 138–144.

[64] Gianluca Stringhini, Christopher Kruegel, và Giovanni Vigna. 2010. Phát hiện những kẻ gửi thư rác trên mạng xã hội. Trong Kỷ yếu Hội nghị Ứng dụng Bảo mật Máy tính Thư ờng niên của ACM. 1–9.

[65] Erik F Tjong Kim Sang và Fien De Meulder. 2003. Giới thiệu về tác vụ chia sẻ CoNLL 2003: nhận dạng thực thể đư ợc đặt tên độc lập với ngôn ngữ. Trong Kỷ yếu Hội nghị ACM về Học ngôn ngữ tự nhiên tại HLT-NAACL. 142–147.

[66] Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, và Valentin Malykh. 2018. Bình thư ờng hóa khái niệm y tế trong các bài đăng trên mạng xã hội với mạng nơ -ron hồi quy. Tạp chí Tin học Y sinh (2018).

[67] Elena Tutubalina và Sergey Nikolenko. 2017. Sự kết hợp của mạng lư ới thần kinh tái phát sâu và các trư ờng ngẫu nhiên có điều kiện để trích xuất các phản ứng có hại của thuốc từ đánh giá của ngư ời dùng. Tạp chí Kỹ thuật Y tế 2017 (2017).

[68] Lee Ventola. 2014. Truyền thông xã hội và các chuyên gia chăm sóc sức khỏe: lợi ích, rủi ro và thực tiễn tốt nhất. Dư ợc và Trị liệu 39, 7 (2014), 491.

[69] Matthew T Wiley, Canghong Jin, Vagelis Hristidis, và Kevin M Esterling. 2014. Thuốc tân dư ợc bàn tán xôn xao trên mạng xã hội. Tạp chí tin học y sinh 49 (2014), 245–254.

[70] Long Xia, G Alan Wang, và Weiguo Fan. 2017. Phư ơng pháp tiếp cận nhận dạng thực thể đư ợc đặt tên dựa trên Deep Learning để xác định và trích xuất các tác dụng phụ của thuốc trong phư ơng tiện truyền thông xã hội về sức khỏe. Trong Kỷ yếu của Hội nghị Quốc tế Springer về Sức khỏe Thông minh. 237–248.

[71] Christopher C Yang, Haodong Yang, Ling Jiang, và Mi Zhang. 2012. Khai thác phư ơng tiện truyền thông xã hội để phát hiện tín hiệu an toàn thuốc. Trong Kỷ yếu hội thảo Quốc tế ACM về Sức khỏe và phúc lợi thông minh. 33–40.

[72] Andrew Yates, Nazli Goharian, và Ophir Frieder. 2015. Trích xuất các phản ứng có hại của thuốc từ mạng xã hội. Trong Kỷ yếu Hội nghị AAAI về Trí tuệ nhân tạo, Tập. 15. 2460–2467.

[73] Antonio Jimeno Yepes và Andrew MacKinlay. 2016. NER cho các thực thể y tế trong Twitter bằng cách sử dụng trình tự để sắp xếp các mạng lư ới thần kinh. Trong Kỷ yếu của Hội thảo Hiệp hội Công nghệ Ngôn ngữ Úc. 138–142.

[74] Antonio Jimeno Yepes, Andrew MacKinlay, và Bo Han. 2015. Điều tra giám sát sức khỏe cộng đồng bằng twitter. Trong Kỷ yếu Hội thảo Xử lý Ngôn ngữ Tự nhiên Y sinh. 164–170.

[75] XueZhong Zhou, Jörg Menche, Albert-László Barabási, và Amitabh Sharma. 2014. Mạng lư ới triệu chứng-bệnh tật ở ngư ời. Truyền thông tự nhiên 5 (2014), 4212.