TỰ ĐỘNG PHÁT HIỆN TÀI KHOẢN SOCKPUPPET TRONG WIKIPEDIA

qua

Mostofa Najmus Sakib

Một luận án

nộp một phần để hoàn thành

của các yêu cầu về mức độ

Thạc sĩ Khoa học Máy tính

Đại học Boise State

Tháng 8 năm 2022

TRƯỜNG CAO ĐẲNG ĐẠI HỌC BANG BOISE

ỦY BAN QUỐC PHÒNG VÀ PHÊ DUYỆT ĐỌC CUỐI CÙNG

của luận án được nộp bởi

Mostofa Najmus Sakib

Tiêu đề luận án: Tự động phát hiện tài khoản Sockpuppet trong Wikipedia

Ngày thi vấn đáp cuối kỳ: 23 tháng 6 năm 2022

Những cá nhân sau đây đã đọc và thảo luận về luận án do sinh viên Mostofa Najmus Sakib nộp, và họ đã đánh giá bài thuyết trình và phản hồi của anh ấy đối với các câu hỏi trong kỳ thi vấn đáp cuối cùng. Họ thấy rằng sinh viên đã vượt qua kỳ thi vấn đáp cuối cùng.

Francesca Spezzano, Tiến sĩ	Chủ tịch, Ủy ban giám sát
Tiến sĩ Edoardo Serra	Thành viên, Ủy ban giám sát
Tiến sĩ Nasir Eisty	Thành viên, Ủy ban giám sát

Sự chấp thuận đọc cuối cùng của luận án đã được Francesca Spezzano, Tiến sĩ, Chủ tịch Ủy ban giám sát chấp thuận. Luận án đã được chấp thuận bởi Cao đẳng sau đại học.

CỐNG HIẾN

Dành tặng cho cha mẹ yêu quý của tôi và tất cả các nhà nghiên cứu đã cống hiến cả cuộc đời mình cho con đường khoa học.

LỜI CẢM Ơ N

Trư ớc hết, tôi muốn gửi lời cảm ơ n chân thành nhất tới cố vấn của tôi, Tiến sĩ. Francesca Spezzano, vì sự ủng hộ liên tục của cô ấy đối với thử thách như ng đáng nhớ này hành trình học tập. Cô ấy đã không ngừng hỗ trợ và động viên tôi đạt đư ợc mục tiêu hoàn thành luận án này. Sự hư ớng dẫn của cô ấy đã giúp tôi phát triển thành một nhà nghiên cứu độc lập. Tôi sẽ luôn trân trọng ký ức về chuyến đi này và sự tự do nghiên cứu. Số lư ợng việc học sẽ giúp ích cho tôi hư ớng tới một tư ơ ng lai tốt đẹp hơ n.

Tôi cũng muốn bày tỏ lòng biết ơ n của tôi đến Tiến sĩ Edoardo Serra và Tiến sĩ Nasir Eisty đã hào phóng phục vụ trong ủy ban luận án của tôi. Cùng với cố vấn của tôi, ủy ban các thành viên luôn hư ớng dẫn tôi bằng những lời khuyên và chỉ dẫn có giá trị. Tất cả các giảng viên khác các thành viên của Khoa Khoa học Máy tính tại Đại học Boise State luôn luôn cũng rất hữu ích và chư a bao giờ khiến tôi từ bỏ sự hư ớng dẫn và hỗ trợ của họ.

Cuối cùng, lòng biết ơ n và lời cảm ơ n của tôi đến gia đình và tất cả các bạn Khoa học máy tính Sinh viên sau đại học của khoa tại Đại học Boise State.

TÓM TẮT

Wikipedia là một bách khoa toàn thư miễn phí trên Internet được xây dựng và duy trì thông qua sự hợp tác nguồn mở của một cộng đồng tình nguyện viên. Mục đích của Wikipedia là mang lại lợi ích độc giả bằng cách hoạt động như một bách khoa toàn thư miễn phí và dễ tiếp cận, một văn bản toàn diện tóm tắt có chứa thông tin về tất cả các nhánh kiến thức đã được khám phá. Trang web có hàng triệu trang được duy trì bởi hàng ngàn biên tập viên tình nguyện. Thật không may, với định dạng biên tập mở, Wikipedia rất dễ bị tấn công bởi các hoạt động độc hại, bao gồm phá hoại, thư rác, biên tập trả phí không được tiết lộ, v.v.

Người dùng độc hại thường sử dụng tài khoản sockpuppet để tránh bị chặn hoặc bị cấm được áp đặt bởi người quản lý Wikipedia trên tài khoản gốc của người đó. Một con rối là một "danh tính trực tuyến được sử dụng cho mục đích lừa dối." Thông thường, một số tài khoản bù nhìn được điều khiển bởi một cá nhân (hoặc thực thể) duy nhất được gọi là người điều khiển rối. Hiện tại, các tài khoản sockpuppet bị nghi ngờ đang được Wikipedia xác minh thủ công người quản lý, khiến quá trình này chậm và kém hiệu quả.

Mục tiêu chính của nghiên cứu này là phát triển một ML tự động và mạng nơ-ron hệ thống dựa trên mạng để nhận dạng các mẫu tài khoản sockpuppet càng sớm càng tốt và đề xuất đình chỉ. Chúng tôi giải quyết vấn đề như một nhiệm vụ phân loại nhị phân và đề xuất một bộ tính năng mới để nắm bắt hành vi đáng ngờ có tính đến hoạt động của người dùng và phân tích nội dung đóng góp. Để tuân thủ công việc này, chúng tôi đã tập trung vào tính năng dựa trên tài khoản và dựa trên nội dung. Giải pháp của chúng tôi được chia thành phát triển một chiến lược tự động phát hiện và phân loại các chỉnh sửa đáng ngờ được thực hiện bởi cùng một tác giả

từ nhiều tài khoản. Chúng tôi đư a ra giả thuyết rằng “bạn có thể ẩn sau màn hình, như ng
tính cách không thể che giấu.” Ngoài phư ơ ng pháp nêu trên, chúng tôi cũng có
gặp phải bản chất tuần tự của công việc. Do đó, chúng tôi đã mở rộng phân tích của mình
với mô hình Bộ nhớ dài hạn ngắn (LSTM) để theo dõi mô hình tuần tự của
phong cách viết của ngư ời dùng.

Trong suốt quá trình nghiên cứu, chúng tôi cố gắng tự động hóa việc phát hiện tài khoản sockpuppet
hệ thống và phát triển các công cụ để giúp quản trị Wikipedia duy trì chất lư ợng
bài viết. Chúng tôi đã thử nghiệm hệ thống của mình trên một tập dữ liệu mà chúng tôi đã xây dựng có chứa 17K tài khoản đư ợc xác thực là
sockpuppets. Kết quả thử nghiệm cho thấy cách tiếp cận của chúng tôi đạt đư ợc điểm F1 là 0,82
và vư ợt trội hơ n các hệ thống khác đư ợc đề xuất trong tài liệu. Chúng tôi có kế hoạch cung cấp nghiên cứu của mình
cho các nhà quản lý Wikipedia để tích hợp nó vào hệ thống hiện có của họ.

MỤC LỤC

DANH SÁCH CÁC BẢNG

DANH SÁCH CÁC HÌNH ẢNH

DANH SÁCH CÁC TỪ VIẾT TẮT

AA	Quyền tác giả
Trí tuệ nhân tạo	Trí tuệ nhân tạo
Giao diện lập trình ứng dụng API	Giao diện lập trình ứng dụng
LSTM	Bộ nhớ dài hạn ngắn hạn
Quặng	Dịch vụ đánh giá sửa đổi khách quan
Tăng cường tuyến	Rừng ngẫu nhiên
RNN	Mạng nơ -ron hồi quy
Giao diện người dùng	Giao diện ngư ời dùng

CHƯƠNG MỘT: GIỚI THIỆU

Wikipedia là một bách khoa toàn thư miễn phí trên Internet được bắt đầu vào năm 2001 [9]. Nó hoạt động theo phong cách quản lý nguồn mở và được duy trì bởi tổ chức phi lợi nhuận Wikimedia Nền tảng [9]. Họ sử dụng phần mềm cộng tác được gọi là "wiki" giúp việc tạo ra dễ dàng hơn, phát triển và phân phối các bài viết. Mục tiêu của Wikipedia là mang lại lợi ích cho người đọc bằng cách hoạt động như một bộ bách khoa toàn thư có thể truy cập rộng rãi, miễn phí và là một tài liệu toàn diện bản tóm tắt có chứa thông tin về tất cả các nhánh kiến thức đã khám phá [20]. Hơn nữa, đối tượng chung có ít quyền truy cập vào thiết bị điện tử được hưởng lợi từ Các bài viết trên Wikipedia vì nó trình bày một bản tóm tắt được viết một cách trung lập về các thông tin có sẵn kiến thức chính thống duy trì tính chính xác và công bằng với một cách đơn giản, "chỉ- phong cách sự thật" [20].

Các dự án hợp tác như Wikipedia đã trở nên phổ biến trong thời gian gần đây. bách khoa toàn thư cộng đồng lớn nhất thế giới đã xuất hiện do tính chất phi tập trung của nó [29]. Với định dạng biên tập mở, Wikipedia rất dễ bị tấn công bởi phần mềm độc hại hoạt động, bao gồm phá hoại, thư rác, biên tập trả phí không được tiết lộ, v.v. [22, 23, 24]. Miễn phí diễn đàn trực tuyến như Wikipedia cung cấp một nền tảng tuyệt vời cho người dùng để giao tiếp và chia sẻ kiến thức. Mặt khác, nó cũng tạo điều kiện cho những kẻ phạm tội trực tuyến lừa đảo, gian lận và tăng nguy cơ của người dùng phổ thông. Theo chính sách của Wikipedia, mỗi người dùng được cho là chỉ tạo một tài khoản người dùng để duy trì sự rõ ràng và tăng cường lòng tin của cộng đồng. Tuy nhiên, Wikipedia không có quy định nghiêm ngặt về hệ thống một người dùng một tài khoản [30]. Do đó, người dùng có thể tự do tạo nhiều tài khoản theo lựa chọn của mình. Điều này

sự tự do tạo tài khoản ngư ời dùng với thông tin tối thiểu đã khiến những ngư ời dùng có ý đồ xấu
tạo ra nhiều danh tính và sử dụng chúng cho nhiều mục đích khác nhau, từ việc quảng bá
của sản phẩm, thúc đẩy quan điểm của một ngư ời, đư ợc trả tiền cho các bài viết, trốn tránh lệnh trừng phạt,
những tuyên bố sai lệch về ý kiến đa số, tránh sự giám sát, v.v. [21]. Nếu bất kỳ ngư ời dùng nào tạo ra một
giải thích cho các mục đích xấu nêu trên, nó đư ợc gọi là con rối. Trong
theo thuật ngữ kỹ thuật, một con rối là "một danh tính trực tuyến đư ợc sử dụng để tạo ra sự lừa dối" [21].
Thông thư ờng, một số tài khoản bù nhìn đư ợc kiểm soát bởi một cá nhân (hoặc thực thể) duy nhất
đư ợc gọi là ngư ời điều khiển rối.

Trong Wikipedia, bất kỳ ngư ời dùng nào đư ợc chứng minh là đóng góp thông tin sai lệch để tạo ra một
thanh toán, phá hoại các bài viết hiện có hoặc thao túng quan điểm chung thông qua việc làm sai lệch
thông tin đư ợc xác định là có tội. Bằng chứng như vậy có thể dẫn đến lệnh cấm ngay lập tức
bị áp dụng trong vài giờ đến một ngày, tùy thuộc vào mức độ nghiêm trọng của tội phạm.
Những ngư ời dùng có ác ý thư ờng sử dụng tài khoản sockpuppet để lách lệnh chặn hoặc lệnh cấm do
Ngư ời quản lý Wikipedia sử dụng tài khoản gốc của ngư ời đó cho mục đích không trung thực [29].

Thông thư ờng, các tài khoản sockpuppet hoặc địa chỉ IP khác nhau đư ợc vận hành để tiếp tục
những tác phẩm đư ợc trình bày rõ ràng như vậy bằng cách tận dụng chính sách tạo tài khoản thoải mái của Wikipedia.
Nếu có bất kỳ khiếu nại nào đư ợc chỉ ra đối với ngư ời dùng liên quan đến sockpuppetry, một cuộc điều tra sockpuppetry
vụ kiện đư ợc đệ trình. Không giống như các bư ớc tạo tài khoản đơn giản, khiếu nại yêu cầu bằng chứng đầy đủ
để dẫn đến lệnh cấm vĩnh viễn. Ngoài ra, những tuyên bố như vậy cần phải đư ợc hỗ trợ bằng các bằng chứng cụ thể
bằng chứng liên quan đến thao túng, phá hoại, thông tin quảng cáo, văn bản tư ơng tự
các mẫu, v.v. [29].

Mặc dù trong hầu hết các trư ờng hợp, nhiều tài khoản đư ợc tạo ra vì mục đích cá nhân, như ng vẫn có
một số tình huống cần phải duy trì nhiều hơn một tài khoản. Ví dụ, có

có thể là một kịch bản mà ngư ời tạo nội dung viết một bài viết liên quan đến các chủ đề nhạy cảm như như chính trị hoặc tôn giáo. Biên tập viên có thể cần sử dụng bút danh để đóng góp cho những trư ờng hợp như vậy vì việc tiết lộ danh tính thực sự của mình có thể gây ra sự thù hận và dẫn đến hậu quả đe dọa đến tính mạng. Ngoài ra, ngư ời dùng cũng đư ợc phép tạo thêm tài khoản cho các vấn đề riêng tư . Ví dụ, nếu tài khoản chính bị xâm phạm, duy trì bảo mật trong khi kết nối thông qua mạng không an toàn, giữ quyền riêng tư trong khi chỉnh sửa các chủ đề gây tranh cãi cao, một khởi đầu mới dư ới tên ngư ời dùng mới, tham gia vào mục đích giáo dục, kiểm tra sự xuất hiện của một tài khoản khác trong khi tạo nội dung, v.v. [21].

Hiện tại, các tài khoản sockpuppet bị nghi ngờ đang đư ợc Wikipedia xác minh thủ công quản trị viên, làm cho quá trình chậm và kém hiệu quả [29]. Các tác phẩm hiện có của phát hiện sockpuppetry từ các tài khoản đơ n lẻ hoặc nhiều trung thành đã tập trung vào các tính năng tập trung vào phong cách, cú pháp và mạng xã hội chủ yếu thông qua kiểm tra chéo sự giống nhau của những ngư ời nắm giữ tài khoản khác nhau. Ý nghĩa ngữ nghĩa đư ợc thừa hư ởng của các chỉnh sửa hiếm khi đư ợc các nhà nghiên cứu trư ớc đây xem xét. Cùng với tài khoản dựa trên đặc điểm về phong cách và cú pháp, chúng tôi sẽ nhấn mạnh trong nghiên cứu này về nội dung hoặc, nói cách khác từ ngữ, ý nghĩa ngữ nghĩa của các chỉnh sửa để điều tra các mô hình liên quan đến tài khoản sockpuppet do cùng một ngư ời dùng nắm giữ. Nghiên cứu của chúng tôi mở rộng các công trình trư ớc đó bằng cách mang lại ngữ nghĩa, tức là, mẫu viết của ngư ời dùng, giọng điệu và các yếu tố bổ sung của một chỉnh sửa để kết nối với nhiều chủ tài khoản.

1.1 Luận đề

Luận án này nhằm mục đích phát hiện sự hiện diện của các tài khoản bù nhìn trên Wikipedia.
tác phẩm áp dụng thuật toán học máy và học sâu để loại bỏ những tài khoản như vậy.
Trong suốt công trình này, chúng tôi tập trung vào việc tìm kiếm câu trả lời cho các nghiên cứu sau đây
câu hỏi.

RQ1: Những mô hình tài khoản bù nhìn do người điều khiển bù nhìn tạo ra là gì?

RQ2: Phân tích ngữ nghĩa từ các bản chỉnh sửa có nắm bắt được mô hình viết và đóng góp không?
các trang phức tạp hơn và xác định các tài khoản sockpuppet tốt hơn cú pháp,
phong cách và các tác phẩm dựa trên mạng lưới đồ thị và đưa ra mức độ ý nghĩa ngữ cảnh sâu sắc?

RQ3: Có thể phát hiện sớm các tài khoản sockpuppet và đề xuất đình chỉ không?

CHƯƠNG HAI: CÔNG TRÌNH LIÊN QUAN

Tài khoản Sockpuppet thường được sử dụng để tăng cường lưu lượng truy cập internet của nội dung ngách không mong muốn, bài đăng trả phí, chủ đề gây tranh cãi và các tài liệu không liên quan thao túng phiếu bầu và lượt xem nội dung [32]. Ngoài ra, những tài khoản bổ sung đó cũng được sử dụng cho hành vi độc hại cụ thể như các nỗ lực gian lận, gửi thư rác, xác định gian lận, và phân phối phần mềm độc hại. Nói chung, nhiều danh tính giả được tạo bởi người dùng để thao túng quan điểm của người dùng, trong khi các hình thức làm việc khác bao gồm một con rối nhóm. Một nhóm sockpuppet có thể là một nhóm tài khoản được tạo bởi một hoặc nhiều người dùng để chuyển hướng sự chú ý của khán giả đến các bài đăng mục tiêu và tạo ra ảo giác về sự ủng hộ [32].

Lịch sử nghiên cứu về những nỗ lực làm rối trên Wikipedia không phải là cũ. Cho đến khi thời gian gần đây khái niệm như vậy không được thiết lập. Với sự xuất hiện của phương tiện truyền thông xã hội và nền tảng trực tuyến, tạo nhiều danh tính và các nỗ lực gian lận trên nền tảng trực tuyến đã trở nên nổi bật hơn. Wikipedia đã thực hiện đánh giá dựa trên quản trị viên của mình tuyên bố về múa rối công khai. Theo truyền thống, các nhà nghiên cứu đã tận dụng những dữ liệu công khai đó để tiến hành cuộc điều tra về hoạt động điều khiển.

Trong tài liệu, một số tác phẩm đã phân tích và phát hiện các tài khoản bù nhìn trong mạng xã hội trực tuyến và diễn đàn thảo luận [25, 26, 27,28]. Cách tiếp cận ban đầu đối với Vấn đề phát hiện sockpuppet xoay quanh việc phát hiện xác định tác giả (AA). Tất cả các loại phát hiện AA đó thường tuân theo một khuôn khổ phân loại văn bản trong đó các tác giả là số lượng các lớp. Theo lịch sử, các tác phẩm như vậy bao gồm đơn giản và dễ dàng-

để triển khai các thuật toán học máy cho phân loại [1,2,3,4,5]. Cụ thể là
Wikipedia, Solorio et al. [29,30] đã giải quyết vấn đề phát hiện xem có hay không
cùng một ngư ời dùng duy trì hai tài khoản bằng cách sử dụng các tính năng xác định tác giả văn bản. Họ
đã tập trung nhiều vào các bình luận và chỉnh sửa trên các trang thảo luận và xem xét các tính năng
chẳng hạn như dấu câu, sử dụng biểu tư ợng cảm xúc, viết hoa và loại từ để
đặc trư ng cho phong cách viết của ngư ời dùng. Nhiều nghiên cứu trư ớc đó [29,30] đã thu hút
chú ý đến thực tế là các tính năng cấp thấp như n-gram ký tự có thể xác định thành công
phong cách viết độc đáo. Phân tích của họ nhấn mạnh lại rằng các đặc điểm ngữ nghĩa như túi-của-
từ ngữ, đặc điểm phong cách như dấu câu, sử dụng biểu tư ợng cảm xúc, viết hoa
thông tin và thông tin cú pháp như cấp độ từ loại, tất cả các loại này đều là
đặc biệt hữu ích cho việc phát hiện rối [29]. Một loại công việc khác đã theo sau
ý thức hệ của các phư ơng pháp tiếp cận dựa trên sự tư ơng đồng. Các tính năng cụ thể của tác giả đã hỗ trợ quá trình này theo cách như vậy
các trư ờng hợp như điểm số dựa trên sự tư ơng đồng thư ờng đư ợc tính toán từ chúng [6,7,8].

Yamak et al. [31] đã tập trung vào việc phân loại các con rối so với các tài khoản chính hãng theo
sử dụng hành vi phi ngôn ngữ và xem xét các mẫu chỉnh sửa. Họ coi Wikipedia-
các tính năng cụ thể, tức là số lần chỉnh sửa, tần suất hoàn nguyên sau mỗi lần đóng góp trong
cùng một bài viết, thời gian giữa đăng ký và chỉnh sửa, v.v. Tiếp tục công việc,
cùng tác giả cũng đề cập đến việc nhóm các tài khoản sockpuppet đư ợc phát hiện đư ợc tạo ra bởi
cùng một cá nhân [32]. Các tác giả đã phát triển các đồ thị quan hệ và kết hợp chúng với
thuật toán phát hiện cộng đồng và các thuộc tính tập trung vào tài khoản để bắt sockpuppet
nhóm. Tsikerdekis và Zeadally [33] đã thực hiện một phân tích tập trung vào Wikipedia để phát hiện
lừa dối danh tính thông qua việc sở hữu hoạt động ngư ời dùng phi ngôn ngữ. Thí nghiệm của họ phản ánh
trên 7.500 tài khoản bù nhìn với ít nhất một lần sửa đổi và tính toán phi ngôn từ

hành vi, bao gồm số lượng tổng số lần sửa đổi trên các trang Wikipedia khác nhau (bài viết, thảo luận bài viết, trang người dùng, trang thảo luận người dùng) và số byte trung bình được thêm vào hoặc bị xóa.

Zheng [34] đã thực hiện một phân tích sockpuppet bằng cách xem xét sockpuppet trong cùng một diễn đàn và đa nền tảng. Họ đã so sánh các hồ sơ tương đồng dựa trên từ khóa cho các bài đăng A1 và A2 trong hai diễn đàn khác nhau và đánh giá khả năng trở thành một cặp rối. Họ cho rằng những người điều khiển rối có xu hướng tuân theo các mẫu viết tương tự ngay cả khi họ sử dụng nhiều tài khoản.

Giống như Wikipedia, việc tạo nhiều tài khoản rất phổ biến trong các phương tiện truyền thông xã hội. Ví dụ, Maitry et al. [35] đã phân tích các tài khoản sockpuppet trên Twitter và Swati Adhikari [36] đã thực hiện một phát hiện sockpuppet tương tự trên dữ liệu Reddit. Ngoài ra, Maitry et al. [35] nhấn mạnh các tweet thời gian thực và các tính năng tập trung vào hồ sơ để xác định tài khoản dưới cùng một người dùng trong thời gian nhanh chóng, trong khi Swati Adhikari [36] bao gồm Reddit người dùng, bài đăng, subreddit và điểm karma của họ. Tuy nhiên, cả hai tác phẩm đều là nền tảng phụ thuộc và không thể được khái quát hóa trên các nền tảng chéo khác.

Một phân tích dựa trên cộng đồng trực tuyến đa dạng đã được Kumar và cộng sự thực hiện [28]. Các tác giả đã phân tích hành vi của những người làm rối tất trên chín cộng đồng khác nhau. phân tích sâu sắc cho thấy rằng những con rối khác với người dùng thông thường về mặt mô hình hoạt động truyền thông xã hội và cấu trúc mạng xã hội tương ứng. Ví dụ, họ chỉ ra rằng rối tất tuân theo những đặc điểm ngôn ngữ độc đáo (ngôi thứ nhất đơn lẻ hơn) và có nhiều cơ hội đăng bài trên cùng một cuộc thảo luận trong một khoảng thời gian ngắn. Ngoài ra, họ khẳng định cặp rối tất có phong cách viết và mẫu tương tự so với những người đóng góp thường xuyên.

Joshi et al. [24] đã điều tra việc sử dụng tài khoản bù nhìn để thực hiện các hành vi không đư ợc tiết lộ chỉnh sửa trả phí trên Wikipedia. Họ phát hiện ra rằng các tài khoản bù nhìn liên quan đến biên tập viên đư ợc trả lư ơ ng chỉ làm việc trên một số lư ợng hạn chế các tiêu đề Wikipedia mà họ quan tâm quảng bá, trong khi ngư ời dùng thực sự chỉnh sửa nhiều trang liên quan đến lĩnh vực chuyên môn của họ hơ n. Điều này cho thấy hành vi của các tài khoản bù nhìn trong Wikipedia khác với hành vi bù nhìn trong cộng đồng thảo luận trực tuyến, nơ i mục tiêu chính của những con rối là tư ơ ng tác với nhau để lừa dối ngư ời dùng khác [28].

CHƯƠNG BA: PHƯƠNG PHÁP NGHIÊN CỨU

Phần này mô tả cách chúng tôi xây dựng một tập dữ liệu chứa sockpuppet và lành tính tài khoản người dùng. Chúng tôi đã thu thập và phân tích dữ liệu điều tra sockpuppet thông qua API (Giao diện lập trình ứng dụng) thu thập thông tin có liên quan từ Wikipedia. Chương sau đây mô tả phương pháp luận của chúng tôi và hướng dẫn người đọc áp dụng các phương pháp tương tự cho các vấn đề khác. Chúng tôi bắt đầu bằng việc xác định quy trình quản lý tập dữ liệu và sau đó bao gồm mô tả tính năng và quá trình trích xuất.

3.1 Bộ dữ liệu

Để thu thập dữ liệu Wikipedia, chúng tôi đã sử dụng MediaWiki Action API [10]. MediaWiki Action API là một dịch vụ web cho phép truy cập vào một số tính năng wiki như xác thực, hoạt động trang và tìm kiếm. Ngoài ra, nó có thể cung cấp siêu thông tin về wiki và người dùng đã đăng nhập.

Để bắt đầu thu thập dữ liệu Wikipedia thông qua API, chúng tôi đã tìm kiếm tất cả các tiểu thể loại nằm trong thể loại chính “Những con rối Wikipedia bị nghi ngờ”. Tất cả các tiểu thể loại trong thể loại chính đã được thu thập cho đến ngày 28 tháng 5 năm 2022. Những các tiểu thể loại là các tài khoản bù nhìn được Wikipedia xác định. Tất cả các tiểu thể loại đó thường tuân theo quy ước đặt tên chuẩn của Wikipedia và bắt đầu bằng “Wikipedia sockpuppets of” theo sau là tên tài khoản. Ví dụ, “Wikipedia sockpuppets of -dantbh” là một tiểu thể loại của các trường hợp sockpuppetry. Khi tất cả các tiểu thể loại Wikipedia đã được trích xuất, chúng tôi tập trung vào các tài khoản người dùng trong từng tiểu thể loại sockpuppet. Thông thường, mỗi tiểu thể loại bù nhìn (ví dụ: bù nhìn Wikipedia của -dantbh)

có nhiều tài khoản ngư ời dùng dư ới cùng một tên tài khoản. Ví dụ đã chọn của chúng tôi
tiểu thể loại (Wikipedia sockpuppets của -dantbh) có 20 tài khoản ngư ời dùng khác nhau cho
cùng một tài khoản. Sau khi các tài khoản ngư ời dùng đư ợc lấy lại cho mỗi ngư ời dùng trong mỗi tiểu thể loại,
chúng tôi đã tìm kiếm những đóng góp hoặc chỉnh sửa của từng ngư ời dùng. Trọng tâm phân tích của chúng tôi là
đóng góp của mỗi ngư ời dùng. Đóng góp này bao gồm nhiều loại thông tin khác nhau cho mỗi
chỉnh sửa của ngư ời dùng. Dựa trên các thiết lập tham số mặc định cho ngư ời dùng, định dạng chung
và dữ liệu thu đư ợc trông giống như hình 3.1 đối với ngư ời dùng.

```
{
    "batchcomplete": "",
    "continue": {
        "uccontinue": "20190130180447|880978627",
        "continue": "-||"
    },
    "query": {
        "usercontribs": [
            {
                "userid": 24,
                "user": "Jimbo Wales",
                "pageid": 9870625,
                "revid": 881893498,
                "parentid": 881892978,
                "ns": 3,
                "title": "User talk:Jimbo Wales",
                "timestamp": "2019-02-05T14:05:11Z",
                "comment": "/* Fancy I edit Wikipedia T-Shirt */",
                "size": 29753
            },
            {
                "userid": 24,
                "user": "Jimbo Wales",
                "pageid": 9870625,
                "revid": 881282261,
                "parentid": 881270759,
                "ns": 3,
                "title": "User talk:Jimbo Wales",
                "timestamp": "2019-02-01T15:29:31Z",
                "comment": "/* Macedonian President Gorge Ivanov is
now in the House arrest */",
                "size": 60166
            },
```


Hình 3.1 Cấu trúc tập dữ liệu cơ bản từ Wikipedia API

Đối với mỗi lần chỉnh sửa, chúng tôi đã lấy thông tin sau: tên ngư ời dùng (ngư ời dùng), userid, id trang, id cha, id sửa đổi, không gian tên trang (Wikipedia nhóm bài viết vào nhiều danh mục hoặc không gian tên, cụ thể là bài viết, thảo luận bài viết, ngư ời dùng trang, trang thảo luận của ngư ời dùng, dự án, v.v.), tiêu đề trang, dấu thời gian chỉnh sửa, văn bản của

đóng góp của ngư ời dùng và quy mô đóng góp của ngư ời dùng. Danh sách các không gian tên Wikipedia là

đư ợc thể hiện trong hình 3.2.

Namespaces			
Subject namespaces		Talk namespaces	
0	(Main/Article)	Talk	1
2	User	User talk	3
4	Wikipedia	Wikipedia talk	5
6	File	File talk	7
8	MediaWiki	MediaWiki talk	9
10	Template	Template talk	11
12	Help	Help talk	13
14	Category	Category talk	15
100	Portal	Portal talk	101
118	Draft	Draft talk	119
710	TimedText	TimedText talk	711
828	Module	Module talk	829

Hình 3.2 Các danh mục Wikipedia theo không gian tên

Trong tập dữ liệu của chúng tôi, userid và user là id và tên duy nhất cho một tài khoản ngư ời dùng. Chúng tôi

đã xác định tất cả các tài khoản liên quan đến sockpuppetry là các tập dữ liệu tích cực cho sockpuppet

mục đích phát hiện. Ban đầu chúng tôi đã thu thập đư ợc tổng cộng 20.978 danh mục con rối

đư ợc đề cập trong danh mục con rối tất của Wikipedia. Tuy nhiên, sau khi vệ sinh kỹ lư ỡng

và loại bỏ các bình luận trống và nan, chúng tôi còn lại 17.180 sockpuppet hợp lệ

tài khoản.

3.2 Dữ liệu tiêu cực

Để đối chiếu với tài khoản bù nhìn tích cực hoặc đư ợc xác định, chúng tôi cũng cần một số

thông tin tài khoản đư ợc xác định là ngư ời dùng thực sự hoặc chư a bao giờ có bất kỳ khiếu nại nào

chống lại tài khoản của họ. Chúng tôi sẽ gọi những ví dụ như vậy là mẫu tiêu cực. Để có đư ợc

tập dữ liệu tiêu cực, chúng tôi dựa vào các tác phẩm của Kumar et al. [22]. Họ đã ghi lại 16.496

tài khoản tích cực và chúng tôi đã sử dụng các tài khoản được ghi lại của họ làm ví dụ về người dùng tiêu cực. Bộ dữ liệu được báo cáo của họ chứa tên người dùng được xác định là người dùng lành tính. Với cách tiếp cận tương tự với quá trình truy xuất dữ liệu từ Wikimedia API, sự đóng góp của người dùng lành tính được tải xuống như là tập hợp những người dùng tiêu cực. Để phù hợp với sockpuppet hoặc tập dữ liệu tích cực, chúng tôi đã trải qua cùng một quá trình làm sạch đối với tập dữ liệu lành tính người dùng và vẫn còn 16.043 trường hợp cuối cùng. Vì vậy, tập dữ liệu kết hợp của chúng tôi gần như cân bằng.

Đối với mỗi tài khoản được xem xét (cả tài khoản con rối và người dùng lành tính), chúng tôi đã lấy lại 20 lần chỉnh sửa đầu tiên của họ. Chúng tôi đã xem xét 20 lần chỉnh sửa cho mỗi người dùng vì mục tiêu của chúng tôi là xây dựng một hệ thống phát hiện tự động có thể xác định các tài khoản giả mạo càng sớm càng tốt.

3.3 Các tính năng dựa trên tài khoản để xác định người dùng Sockpuppet

Trong phần này, chúng tôi sẽ mô tả và liệt kê tất cả các tính năng chúng tôi đã sử dụng cho phát hiện sockpuppet.

Như đã đề cập trong quá trình trích xuất dữ liệu, chúng tôi có một loạt các thuộc tính tài khoản có sẵn từ phần đóng góp của tài khoản người dùng. Dựa trên thông tin đó, chúng tôi đã sửa một số tính năng có nguồn gốc từ tên tài khoản của người dùng. Từ trước văn học, rõ ràng là tên người dùng là một tính năng quan trọng để phát hiện những kẻ gửi thư rác, biên tập trả tiền không được tiết lộ, bù nhìn và các hành vi độc hại khác [23, 24, 37]. Do đó, chúng tôi đã xem xét các đặc điểm sau được trích xuất từ tên người dùng:

Số lượng chữ số trong tên người dùng: Để tạo nhiều tài khoản, người dùng sockpuppet đôi khi tập trung vào việc tạo tên tài khoản tương tự với các chữ số bổ sung như là sự khác biệt. Đó là lý do tại sao chúng tôi đã xem xét số chữ số trong tên người dùng như một chỉ báo có tác động mạnh mẽ của trò hề rối.

Tỷ lệ giữa các chữ số và tổng số ký tự chữ cái trong tên ngư ời dùng: Giống như các chữ số, ký tự cũng là một thành phần quan trọng của bất kỳ tên ngư ời dùng nào. Nhiều ngư ời dùng tài khoản thư ờng tạo một tài khoản bổ sung chỉ bằng cách điều chỉnh một số ký tự. Trong tính năng này, chúng tôi có tập trung vào tỷ lệ chữ số trên tổng số ký tự chữ cái trong tên ngư ời dùng để nắm bắt những điều tư ơ ng tự tên ngư ời dùng có những thay đổi nhỏ.

Số lư ợng chữ số đầu tiên trong tên ngư ời dùng: Để phân biệt giữa tên ngư ời dùng, ngư ời điều khiển rối đôi khi tạo tài khoản với các chữ số đứng đầu có thể phân biệt giữa các tên tài khoản. Để nắm bắt loại hành vi đó, chúng tôi cũng đã tập trung về số chữ số hàng đầu đã đư ợc sử dụng làm tên ngư ời dùng. Tuy nhiên, sử dụng chữ số hàng đầu chữ số là hành vi đặc biệt so với việc sử dụng số ở bất kỳ nơ i nào khác trong tên ngư ời dùng. Vì vậy, tổng số chữ số và tên ngư ời dùng có chữ số đứng đầu sẽ có khả năng nắm bắt hai mẫu quy ư ớc đặt tên khác nhau.

Tỷ lệ ký tự duy nhất trong tên ngư ời dùng: Tính năng này tập trung vào tính duy nhất tỷ lệ ký tự trong tên ngư ời dùng. Để có đư ợc tính năng này, chúng tôi đã tính toán các ký tự duy nhất của tên ngư ời dùng và chia cho tổng độ dài của tên ngư ời dùng.

Ngoài các tính năng tập trung vào tên ngư ời dùng, chúng tôi đã bao gồm các đặc điểm của ngư ời dùng để khám phá mô hình ẩn của ngư ời dùng sockpuppet. Các tính năng sau đây đư ợc trích xuất để xác định phong cách và chuẩn mực viết chung của ngư ời dùng.

Độ dài đóng góp trung bình: Một phần thông tin thiết yếu đư ợc lấy từ đóng góp của mỗi ngư ời dùng là những bình luận của họ về mỗi lần chỉnh sửa tiếp theo. Vì ngư ời dùng lành tính cố gắng cộng tác và đóng góp nhiều hơ n, độ dài của bình luận phải cao hơ n đối tác của họ. Đó là lý do tại sao chúng tôi coi độ dài bình luận là một tính năng quan trọng.

Độ dài tiêu đề trung bình: Chúng tôi xem xét độ dài trung bình của tiêu đề các trang a
do ngư ời dùng đóng góp.

Chênh lệch thời gian trung bình giữa hai lần chỉnh sửa liên tiếp: Hành vi theo thời gian
là một tính năng thiết yếu để phát hiện bất kỳ hoạt động gian lận nào [38]. Do đó, chúng tôi đã xem xét
sự chênh lệch thời gian trung bình giữa hai lần đóng góp liên tiếp là một tính năng khác.

Tất cả các tính năng đư ợc đề cập trư ớc đó đã đư ợc tính toán cho mỗi đóng góp của ngư ời dùng
tài khoản. Tuy nhiên, trọng tâm của chúng tôi là phát hiện ngư ời dùng sockpuppet, không phải đóng góp của họ. Để
phục vụ cho mục đích đó, chúng tôi đã tính trung bình các giá trị của tất cả các tính năng đư ợc mô tả trư ớc đó cho
mỗi ngư ời dùng. Vì vậy, các tính năng dựa trên tên ngư ời dùng sẽ hoàn toàn giống nhau đối với mỗi ngư ời dùng.
Tuy nhiên, độ dài bình luận hoặc tiêu đề cũng như thời gian chênh lệch của mỗi bài viết là khác nhau.
Vì vậy, đối với ba tính năng này, chúng tôi đã tính toán giá trị trung bình của chúng.

3.4 Các tính năng dựa trên nội dung để phát hiện Sockpuppet trong Wikipedia

Thể loại tính năng thứ hai mà chúng tôi kiểm tra là dựa trên nội dung, mà chúng tôi có
đánh giá nội dung chỉnh sửa. Mỗi lần chỉnh sửa đư ợc coi là một tài liệu duy nhất trong trư ờng hợp này và đư ợc thực hiện
thông qua quá trình đư ợc mô tả sau để đư a ra các đặc điểm dựa trên nội dung cho phân tích của chúng tôi.

Chúng tôi đã theo hai cách tiếp cận cơ bản để phân tích nội dung của ngư ời dùng
đóng góp cho việc phát hiện sockpuppetry. Một trong số đó bao gồm việc sử dụng mô hình biến đổi BERT
[39], và một cái khác là tích hợp mô hình chủ đề để thêm các chủ đề của bản chỉnh sửa làm các tính năng cho chúng tôi
phân tích. Động lực chính đằng sau việc áp dụng mô hình máy biến áp và chủ đề
mô hình hóa là để nắm bắt ngữ nghĩa và ý nghĩa của nội dung. Theo truyền thống
phát hiện rối và các nhiệm vụ NLP tư ơng tự chủ yếu tập trung vào việc nắm bắt
sự kế thừa cú pháp và phong cách của nội dung [29, 30]. Ít nhấn mạnh đã đư ợc đư a ra
về các tính năng tập trung vào ngữ nghĩa. Đóng góp chính của chúng tôi thông qua nghiên cứu này là đư a vào

nghĩa ngữ nghĩa có nghĩa là hiểu đư ợc sự kế thừa sâu sắc của nội dung hoặc chỉnh sửa. Cú pháp
là tập hợp các quy tắc cần thiết để đảm bảo một câu đúng ngữ pháp. Ngữ nghĩa, về
Mặt khác, là cách viết, cấu trúc ngữ pháp, giọng điệu và các yếu tố khác của một ngư ời
của một câu hợp nhất lại để truyền đạt ý nghĩa của nó.

Chúng tôi đư a ra giả thuyết rằng việc xem xét ngữ nghĩa của các chỉnh sửa của ngư ời dùng sẽ nắm bắt đư ợc
mẫu cấp độ sâu của nội dung từ các chỉnh sửa đư ợc thực hiện bởi cùng một ngư ời điều khiển rối. Đối với
Ví dụ, nếu ngư ời điều khiển rối tập trung vào một loại nội dung hoặc ngư ời cụ thể, thì tài khoản đó
ngư ời giữ sẽ chỉnh sửa hoặc xuất bản nội dung tư ơ ng tự từ nhiều tài khoản. Vì hành vi
mô hình của ngư ời điều khiển rối có thể đư ợc nắm bắt hiệu quả hơ n thông qua ngữ nghĩa, chúng ta
quyết định đư a nhúng BERT và mô hình hóa chủ đề vào nghiên cứu của chúng tôi. Ví dụ,
giả sử một ngư ời điều khiển rối hoặc ngư ời giữ tài khoản nhóm cố gắng chỉnh sửa các trang liên quan đến Barack
Obama. Trong trư ờng hợp đó, có khả năng cao là họ sẽ thực hiện chỉnh sửa tư ơ ng tự từ nhiều
tài khoản. Việc nắm bắt ý nghĩa ngữ nghĩa sẽ là bư ớc lý tư ởng để làm sáng tỏ một
vấn đề. Đó là lý do tại sao tiếp tục phân tích tập trung vào phong cách hoặc cú pháp của
các nhà nghiên cứu trư ớc đây, chúng tôi sẽ tiến hành nghiên cứu dựa trên ngữ nghĩa để cải thiện hơ n nữa.

3.4.1 Nhúng BERT

Trong cách tiếp cận này, chúng tôi đã tập trung vào máy biến áp hiện đại
Mô hình biến đổi hiện đư ợc sử dụng rộng rãi cho một số xử lý ngôn ngữ tự nhiên
nhiệm vụ, tức là, dịch máy [15], nhận dạng thực thể đư ợc đặt tên [16], trình tự sinh học
phân tích [17,18,19], v.v. Chúng tôi cũng muốn sử dụng một công nghệ tư ơ ng tự để xem máy biến áp
các mô hình có thể thực hiện tốt hơ n để hiểu các mẫu chỉnh sửa tuần tự so với
các phư ơ ng pháp tiếp cận hiện có đư ợc mô tả trong phần công việc liên quan. Mô hình BERT là lựa chọn của chúng tôi
cho nhiệm vụ này.

BERT là viết tắt của Bidirectional Encoder Representations from Transformers. Đây là một mô hình học sâu độc đáo hoạt động dựa trên quá trình chú ý. Mỗi phần tử đầu ra trong mô hình đư ợc kết nối với tất cả các yếu tố đầu vào và duy trì luồng thông tin bằng điều chỉnh trọng lư ợng. Quá trình kết nối độc đáo này đư ợc gọi là sự chú ý cơ chế và làm cho toàn bộ hệ thống trở nên mạnh mẽ và bền bỉ.

BERT thư ờng sử dụng cơ chế chú ý để hiểu ngữ cảnh mối quan hệ giữa các từ. Hai bư ớc riêng biệt (mã hóa văn bản và giải mã cho nhiệm vụ dự đoán) diễn ra một cách hài hòa và trích xuất mối quan hệ thừa kế sâu sắc giữa từ trong văn bản. Nó đặc biệt giúp giải quyết sự mơ hồ trong văn bản bằng cách tiết lộ ngữ cảnh. Không giống như mô hình định hư ớng, đọc các từ theo trình tự (từ trái sang phải hoặc từ phải sang trái) bên trái), bộ mã hóa BERT lấy toàn bộ câu làm một đầu vào. Chiến lư ợc đơ n giản này giúp để hiểu toàn bộ bối cảnh của một văn bản thay vì tập trung từng từ một. Cụ thể này công suất đư ợc bao gồm bằng cách đư a máy biến áp vào và đư ợc gọi là bi-tính định hư ớng.

Chúng tôi đã sử dụng mô hình BERT để tính toán mức độ nhúng của từng đóng góp của ngư ời dùng. Cụ thể, chúng tôi đã sử dụng BertTokenizer để phân mã hóa và chuyển đổi thành tenxơ và mô hình “cơ sở” BERT đư ợc đào tạo trên tiếng Anh viết thư ờng (12 lớp Transformer, 12 đầu tự chú ý, kích thư ớc ẩn 768) từ thư viện Huggingface [39]. Lựa chọn của chúng tôi của phư ơ ng pháp tiếp cận dựa trên tính năng ở đây bao gồm việc trích xuất các kích hoạt (hoặc ngữ cảnh nhúng hoặc biểu diễn mã thông báo hoặc tính năng) từ một hoặc nhiều lớp trong số 12 lớp mà không có tinh chỉnh bất kỳ tham số nào của BERT. Mô hình đóng góp 768 nhúng ngữ cảnh từ mỗi lớp, và đầu ra từ lớp cuối cùng đư ợc sử dụng làm đầu vào cho máy thông thư ờng

học tập và LSTM, tiếp theo là phân loại ngư ời dùng lành tính từ sockpuppet
tài khoản.

3.4.2 Mô hình chủ đề

Mô hình hóa chủ đề là một cách khám phá các chủ đề cấp cao thông qua thống kê
mô hình hóa liên quan đến việc thu thập tài liệu. Giả thuyết của chúng tôi là việc xác định
chủ đề nội dung có thể đóng góp đáng kể vào việc phát hiện nhiều danh tính. Ngư ời dùng có
đức tin thư ờng đóng góp vào nhiều loại nội dung khác nhau. Tuy nhiên, ngư ời dùng sockpuppet có xu hư ớng đăng
nội dung tư ơ ng tự ngay cả khi chúng đã bị xóa trư ớc đó. Để tuân thủ tiền đề này, chúng tôi đã
tận dụng lợi thế của kỹ thuật mô hình hóa chủ đề Phân bổ Dirichlet tiềm ẩn (LDA)
đư ợc cung cấp bởi thư viện Gensim (chúng tôi đã sử dụng WordNetLemmatizer và mô hình bigram)
[12]. LDA là một quá trình tạo chủ đề đơ n giản như ng mạnh mẽ từ một ngữ liệu nhất định.

Để sử dụng các kỹ thuật đư ợc đề cập ở trên, chúng tôi đã lấy lại bản tóm tắt của
nội dung hoặc đóng góp của ngư ời dùng một lần nữa thông qua MediaWiki Action API. Trư ớc đây
công việc, chúng tôi đã phân tích một bình luận hoặc chỉnh sửa duy nhất đư ợc thực hiện bởi mỗi ngư ời dùng. Tuy nhiên, chúng tôi yêu cầu
thêm thông tin để hiểu và tính toán các chủ đề thông qua LDA. MediaWiki API có
một tham số khác có tên là “extracts” trả về bất kỳ văn bản thuần túy hoặc HTML giới hạn nào của trang.
Thông qua quá trình thu thập dữ liệu tư ơ ng tự đư ợc mô tả trong phần 3.1, chúng tôi đã thu thập đư ợc
nội dung cho từng ngư ời dùng. Cuối cùng, chúng tôi sử dụng những nội dung đó để trích xuất chủ đề bằng LDA.

Nội dung mà chúng tôi nhận đư ợc thông qua API bao gồm các thẻ HTML, thêm
dấu câu và khoảng cách. Trư ớc khi đư a vào mô hình LDA, dữ liệu này cần nhiều
dọn dẹp. Đầu tiên, chúng tôi đã thực hiện theo quy trình dọn dẹp văn bản cơ bản, loại bỏ dấu câu,
khoảng trắng thừa và các ký tự đặc biệt bổ sung. Sau đó thông qua mã thông báo và
lemmatization, chúng tôi đã chuẩn bị các văn bản thô cho các bư ớc tiếp theo. Khi chúng tôi đã có các mã thông báo cho

mỗi điểm quan sát, chúng tôi đã phát triển một mô hình bigram theo sau là một ngữ liệu trên toàn bộ
bộ dữ liệu, kết hợp dữ liệu sockpuppet và dữ liệu lành tính. Cụ thể, chúng tôi đã đào tạo một LDA
mô hình với 20 chủ đề về tất cả các bình luận của ngư ời dùng và sau đó đư ợc gán cho mỗi bình luận
vectơ có phân phối chủ đề tư ơ ng ứng.

3.5 Mô hình phân loại

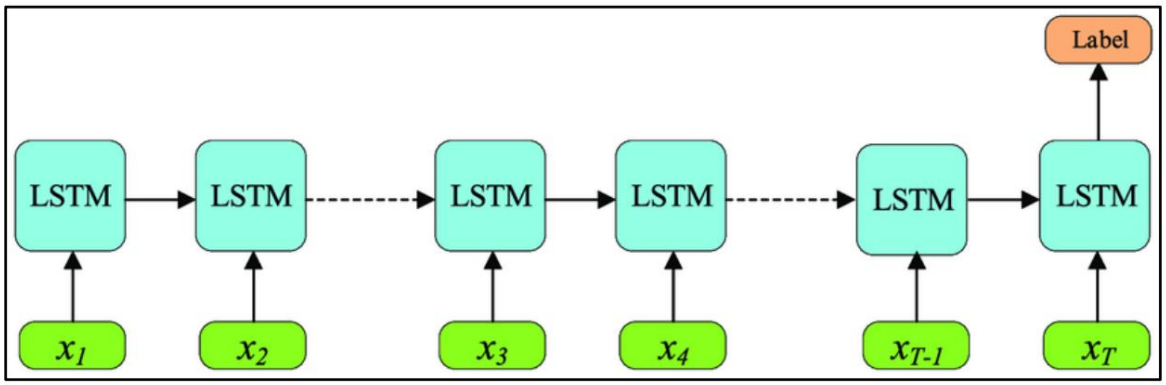
Để kiểm tra các tính năng chúng tôi đề xuất cho nhiệm vụ phát hiện tự động, chúng tôi
xem xét các bộ phân loại khác nhau, cụ thể là Hồi quy Logistic, Gaussian Naive Bayes,
Cây quyết định, Phân loại Multilayer Perceptron (MLP), Rừng ngẫu nhiên, ExtraTree
Bộ phân loại và Bộ nhớ dài hạn ngắn (LSTM). LSTM là một khu vực phức tạp của sâu
học tập mà mạng lư ới của nó là một loại mạng nơ -ron hồi quy có khả năng học thứ tự
phụ thuộc trong các vấn đề dự đoán trình tự. Điều này đạt đư ợc vì sự lặp lại
mô-đun của mô hình có sự kết hợp của các lớp tư ơ ng tác với nhau. Trên-
phư ơ ng pháp nói trên sẽ giúp chúng ta hiểu đư ợc độ tin cậy của
trình tự thời gian của các mẫu chỉnh sửa của ngư ời dùng. Chúng tôi đã đư a ra những cân nhắc sau đây trong
quyết định lựa chọn kiến trúc:

i. Vấn đề phát hiện bình luận của biên tập viên là một nhiệm vụ phân loại dựa trên
lịch sử chỉnh sửa dư ới dạng dữ liệu như vậy đư ợc tạo ra trong khi biên tập viên chỉnh sửa trong một khoảng thời gian.

ii. Để dự đoán trình tự hành vi chỉnh sửa của ngư ời dùng tại bất kỳ bư ớc thời gian nào, thì
cần thiết để học hỏi từ hành vi hoặc hành động của nó từ các bư ớc thời gian trư ớc đó. Điều này mang lại giải pháp cho chúng ta
phản hồi toàn diện từ các bư ớc thời gian trư ớc đến bư ớc hiện tại.

iii. Ngoài ra, LSTM có thể truyền tải luồng phản hồi liên tục mà không bị mất hoặc
bùng nổ trong một chuỗi dài.

Mặc dù LSTM là một dạng chính xác của Mạng nơ-ron hồi quy (RNN), không giống như RNN, LSTM kết hợp các cổng đầu vào, quên và đầu ra [13] có hiệu quả giải quyết vấn đề về độ dốc biến mất hoặc bùng nổ. Trong cách tiếp cận của chúng tôi, chúng tôi đã sử dụng mô hình LSTM kiến trúc với thiết lập nhiều-đến-một hoặc đầu ra lớp ẩn chỉ từ lớp cuối cùng, như được thể hiện trong hình 3.3.



Hình 3.3 Kiến trúc LSTM nhiều-một

Chúng tôi đã sử dụng trọng số cụ thể cho từng lớp để giải quyết tình trạng mất cân bằng lớp. Ngoài ra, quá trình này cho phép mô hình xem xét toàn bộ chuỗi của người đóng góp trước khi phân loại một bản chỉnh sửa. Với kiến trúc mô hình như vậy, một hàm mất mát entropy chéo tiêu chuẩn có dạng như thể hiện trong phương trình 3.1.

$$= \quad (\quad . h \ ) \qquad (3.1)$$

Đây,

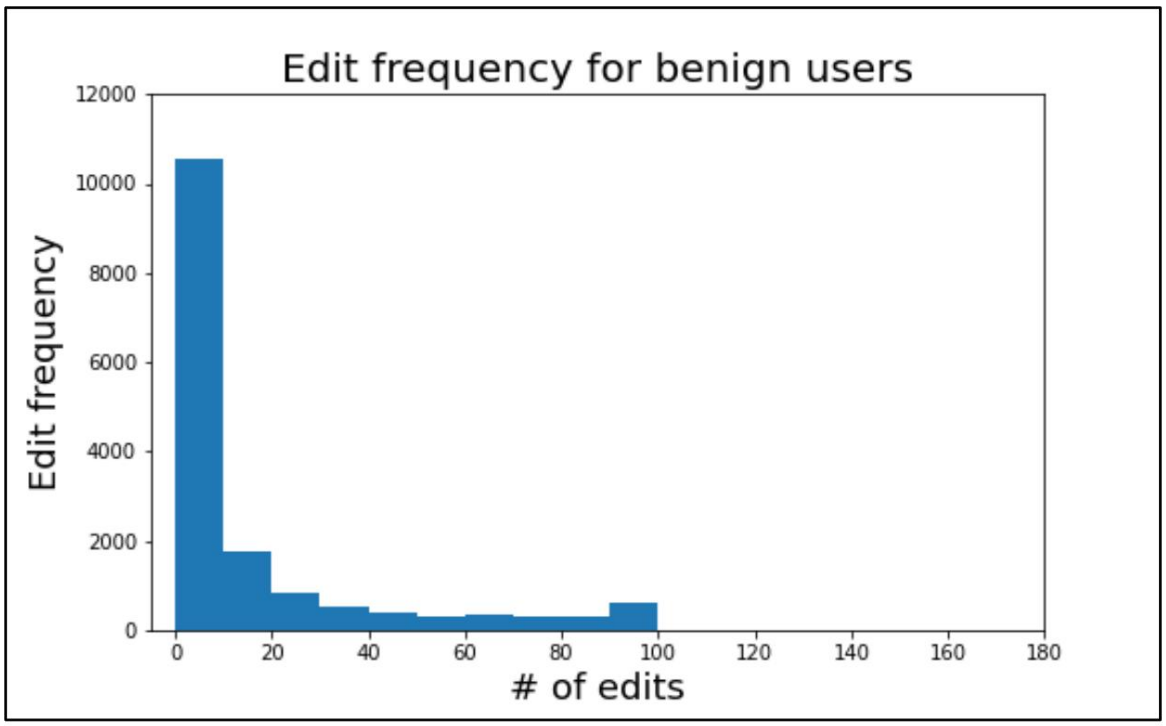
u € người dùng trong tập hợp người dùng U

L = độ dài của chuỗi chỉnh sửa của người dùng u

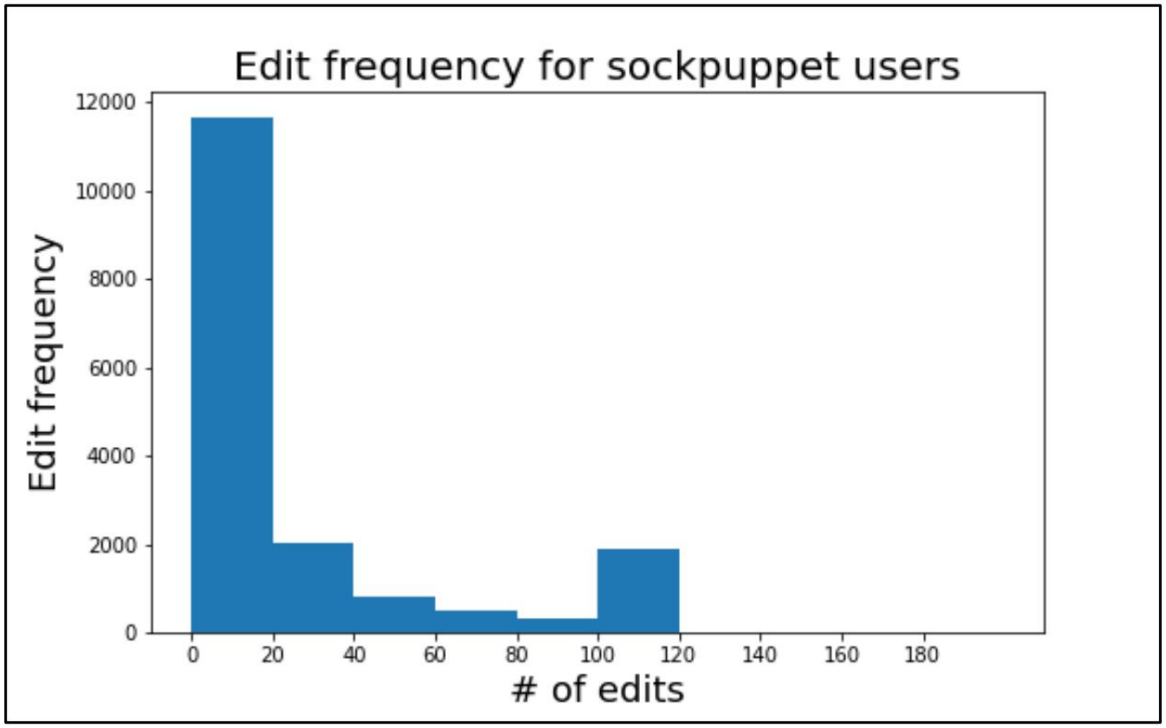
Đối với các mô hình học máy cổ điển, chúng tôi đã xem xét tất cả các tính năng được mô tả trong phần phương pháp luận cộng với vectơ trung bình của các nhúng BERT của người dùng đóng góp và vectơ trung bình của các chủ đề đóng góp của người dùng để nắm bắt ngữ nghĩa của người dùng. Một trong những đóng góp cơ bản của chúng tôi thông qua nghiên cứu này là phát hiện các tài khoản bù nhìn

nhanh chóng. Chúng tôi đã thử nghiệm chỉnh sửa từ một đến hai mươi lần lượt, kết quả là hai mươi
các kịch bản khác nhau. Ví dụ, trong kịch bản đầu tiên, chúng tôi chỉ thực hiện chỉnh sửa đầu tiên của mỗi
đóng góp của người dùng và đánh giá tất cả các tính năng cần thiết cho các mô hình cổ điển. Đối với
tình huống thứ hai, chúng tôi đã thực hiện hai lần chỉnh sửa liên tiếp và tính toán tương tự tất cả các tính năng
một lần nữa, và tính trung bình cho mỗi người dùng. Chúng tôi tiếp tục mô hình tương tự cho phần còn lại của các chỉnh sửa,
tăng số lần chỉnh sửa lên một lần mỗi lần. Cuối cùng, chúng tôi đã có kết quả cho k (1 đến 20)
chỉnh sửa ở cấp độ người dùng khi chúng tôi tính trung bình các tính năng ở cấp độ người dùng.

Đối với mô hình LSTM, chúng tôi xem xét đầu vào là trình tự các tính năng cho mỗi lần chỉnh sửa.
Đối với mỗi lần chỉnh sửa, chúng tôi xem xét độ dài đóng góp, độ dài tiêu đề, chênh lệch thời gian
giữa các bản chỉnh sửa hiện tại và trước đó, việc nhúng BERT của đóng góp và
vectơ các chủ đề của đóng góp. Cuối cùng, chúng tôi đã nối các tính năng dựa trên tên người dùng
để biểu diễn ô cuối cùng của LSTM và chuyển chúng đến phân loại
lớp. Sự đóng góp của các bài viết không đồng đều cho mỗi người dùng vì cả hai lý do lành tính
hoặc người dùng sockpuppet. Chúng tôi đã sử dụng phần đệm trong trường hợp có ít hơn 20 đóng góp của
các biên tập viên để làm cho mỗi đầu vào của người dùng vào LSTM có kích thước cố định của số lượng các tính năng X
20.



Hình 3.4 Tần suất chỉnh sửa cho ngư ời dùng lành tính



Hình 3.5 Tần suất chỉnh sửa cho ngư ời dùng sockpuppet

Hình 3.4 và 3.5 cho thấy sự phân bố các bình luận của từng ngư ời dùng đối với các
và các loại rối tất. Một mô hình thú vị có thể đư ợc nhìn thấy từ số lư ợng
đóng góp của biên tập viên trong cả hai trư ờng hợp. Sau 100 thư ờng là ngư ời dùng lành tính không có bất kỳ
đóng góp, như ng ngư ời dùng sockpuppet vẫn tiếp tục đóng góp. Vì có ít bình luận
sau 20 lần chỉnh sửa, chúng tôi đã xem xét 20 lần chỉnh sửa cho mỗi danh mục để tránh việc thêm nhiều giá trị bằng không
trong LSTM và phát hiện nhanh chóng hiện tư ợng rối ren.

3.6 Đánh giá các phư ơ ng pháp đề xuất

Phần này báo cáo về giao thức đánh giá của chúng tôi.

3.6.1 Số liệu

Để đánh giá hiệu suất của mô hình, chúng tôi đã sử dụng điểm F1. Điểm F1 là
trung bình có trọng số của Độ chính xác và Thu hồi. Độ chính xác đề cập đến tổng số chính xác
dữ liệu tích cực đư ợc phân loại so với tổng số dữ liệu tích cực. Thu hồi là tỷ lệ
của các quan sát tích cực đư ợc dự đoán chính xác so với tất cả các quan sát trong lớp thực tế - có.
Do đó, điểm F1 tính đến cả kết quả dư ơ ng tính giả và kết quả âm tính giả. Theo trực giác, nó
không dễ hiểu như độ chính xác, như ng F1 thư ờng hữu ích hơ n độ chính xác,
đặc biệt là nếu chúng ta đang giải quyết vấn đề phân bố giai cấp không đồng đều.

3.6.2 So sánh với công trình liên quan

Để so sánh công việc của chúng tôi với các kết quả trư ớc đó, chúng tôi cũng đã đư a vào các công việc đã thực hiện
bởi Solorio et al. [29] và Yamak et al. [31] trong nghiên cứu của chúng tôi. Cả hai đều cố gắng phát hiện sockpuppet
tài khoản sử dụng các bộ tính năng và phư ơ ng pháp tiếp cận khác nhau như ng có mục tiêu tư ơ ng tự như của chúng tôi.
công trình của Solorio et al. [29] là một trong những công trình sơ bộ đư ợc thực hiện về phát hiện rối,
trong khi cái cuối cùng là gần đây hơ n. Chúng tôi đã so sánh các số liệu đã đề cập trư ớc đó

thông qua phư ơ ng pháp luận của chúng tôi và cách tiếp cận của họ và cố gắng đư a ra một cách chính xác hơ n
phát hiện hành vi gian lận càng sớm càng tốt.

Solorio et al. [29] đã tiếp cận vấn đề từ việc xác định quyền tác giả
góc nhìn. Mỗi bình luận do ngư ời dùng đư a ra đư ợc coi là một tài liệu và chúng
đư ợc phân loại để kiểm tra các khiếu nại về rối. Họ làm việc theo hai bư ớc. Trong
Bư ớc đầu tiên, họ thu thập dự đoán mức độ bình luận cho mỗi tài khoản. Sau đó thông qua một
sơ đồ bỏ phiếu đa số, họ đư a tài khoản vào danh mục nghi ngờ hoặc lành tính.
các tính năng đư ợc đề cập đư ợc chỉ định dư ới đây.

Tổng số ký tự: Các tác giả đã tính toán tính năng này để mô hình hóa
hành vi viết của ngư ời đóng góp, cụ thể là các văn bản dài hoặc bình luận ngắn.

Tổng số câu: Tính năng này tính tổng số câu
trong các bình luận. Các tác giả cho rằng đây sẽ là một tính năng có giá trị để xác định
sự lựa chọn của ngư ời đóng góp về việc sắp xếp văn bản thành câu. Để đếm số câu, chúng ta có
tận dụng gói sent_tokenizer từ NLTK.

Tổng số mã thông báo: Tổng số mã thông báo không bao gồm khoảng trắng
đư ợc tính ở đây. Chúng tôi đã sử dụng gói word_tokenizer từ NLTK để tính toán điều này
tính năng.

Từ không có nguyên âm: Tỷ lệ từ không có nguyên âm có thể chỉ ra một tín hiệu
đối với một số ngư ời đóng góp. Ví dụ về các từ không có nguyên âm là: try, cry, fly, v.v.

Tổng số chữ cái: Tính năng này là tổng của tất cả các chữ cái
ký tự trong văn bản.

Tổng số dấu câu: Lựa chọn dấu câu của ngư ời dùng thư ờng thay đổi theo
cách độc đáo. Ví dụ, dấu chấm phẩy và dấu gạch nối thư ờng đư ợc một số ngư ời sử dụng

những ngư ời đóng góp, và phần còn lại bỏ qua chúng. Một số dấu câu cũng thay đổi theo cách nó đã đư ợc đư ợc sử dụng trên toàn thế giới. Ví dụ, việc sử dụng dấu phẩy là đặc biệt, một đặc điểm quan trọng trong phát hiện các mẫu chữ viết.

Hai hoặc ba dấu câu đư ợc tính: Ngày nay, nhiều văn bản trang trọng và không trang trọng viết bao gồm việc sử dụng nhiều dấu câu để nhấn mạnh tầm quan trọng hoặc đơ n giản là thể hiện cảm xúc. Những trư ờng hợp như vậy có thể đư ợc xác định bằng cách kiểm tra việc sử dụng nhiều dấu câu đư ợc sử dụng bởi những ngư ời đóng góp. Do đó, các tác giả tin rằng có nhiều cách khác nhau để thể hiện cảm xúc sẽ là một chỉ báo lý tư ởng về ngư ời dùng rối.

Tổng số lần co thắt: Các cơ n co thắt thư ờng đư ợc sử dụng để rút ngắn và kết hợp từ, tức là, đừng, nó là, và tôi là. Đư ợc sử dụng riêng biệt hoặc dạng rút gọn, cả hai trư ờng hợp đều đúng trong ngữ pháp tiếng Anh. Tuy nhiên, cách một ngư ời đóng góp viết hoặc đóng góp là một sự lựa chọn sở thích cá nhân và tính toán sự co lại là một cách lý tư ởng để trích xuất văn bản mẫu hình hoặc hành vi.

Đếm dấu ngoặc đơ n: Tính năng này là một cách chung để xác định tác giả sự ghi nhận và sẽ đóng vai trò quan trọng trong việc phân biệt những ngư ời đóng góp.

Số lư ợng từ viết hoa toàn bộ: Các tác giả đã đếm số lư ợng mã thông báo trong đó tất cả các từ là chữ in hoa. Theo truyền thống, những ngư ời đóng góp sử dụng tất cả các chữ cái in hoa như viết tắt hoặc để nhấn mạnh một số từ. Một số ví dụ là “USA” hoặc “the word was phát âm KHÔNG CHÍNH XÁC.”

Biểu tư ợng cảm xúc đư ợc tính: Trong đấu trư ờng ngày nay, biểu cảm và phong cách viết đư ợc sử dụng rộng rãi đư ợc chi phối bởi các biểu tư ợng cảm xúc, đặc biệt là trong các bài viết trên các trang web. Biểu tư ợng cảm xúc là một hình ảnh sự thể hiện cảm xúc, đặc biệt là biểu cảm khuôn mặt và cảm xúc bên trong. Các tác giả

đánh giá mô hình sử dụng lựa chọn biểu tượng cảm xúc bằng cách đếm tổng số
biểu tượng cảm xúc trong nội dung.

Biểu tượng cảm xúc vui vẻ có giá trị: Mọi người thường thiên vị khi chọn biểu tượng cảm xúc hoặc thể hiện cảm xúc. Nhiều người dùng chỉ thể hiện cảm xúc tích cực hoặc vui vẻ. Biểu tượng cảm xúc vui vẻ thống trị các bài viết như vậy. Các tác giả đã tính riêng các biểu tượng cảm xúc vui vẻ như :) và :-) để đánh giá những người đóng góp.

Số câu không có chữ in hoa ở đầu: Một số người đóng góp thích để bắt đầu viết bằng một chữ cái hoặc số nhỏ. Ví dụ về những trường hợp như vậy có thể là "1862 là năm" hoặc "to và đậm đều áp dụng cho nghi phạm của chúng tôi." Các tác giả tin rằng tính năng này cũng sẽ nắm bắt được kiểu chữ viết độc đáo.

Số lượng trích dẫn: Tương tự như số lượng trong ngoặc đơn, đóng góp của tác giả cũng là thiết yếu để phát hiện sự đóng góp của tác giả. Trong một kịch bản thực tế, người dùng được phân biệt với sự lựa chọn trích dẫn của họ. Vì vậy, số lượng trích dẫn sẽ giúp phân biệt các nhà văn với người khác.

Tần suất thẻ các phần của bài phát biểu (POS): Các tác giả đã xem xét 36 phần của bài phát biểu các thẻ từ bộ thẻ POS của Penn TREE-bank và xóa các dấu chấm câu vì chúng đã được xem xét thông qua các tính năng khác.

Tần suất của các chữ cái: Bảng chữ cái tiếng Anh có 26 chữ cái và tần suất của những chữ cái trong mỗi bình luận được tính toán như các tính năng riêng biệt. Số lượng là được chuẩn hóa theo tổng số ký tự không phải màu trắng trong mỗi bình luận.

Tần suất của từ chức năng: Lựa chọn từ chức năng là một cách tuyệt vời để gắn thẻ các nhà văn với các tác phẩm tương ứng của họ. Ví dụ, các tác giả đã xem xét một danh sách

các từ chức năng từ [11]. Lựa chọn này đã tạo ra 150 tính năng từ danh sách 150 chức năng từ.

Tất cả các tính năng được đề cập ở trên thường được sử dụng trong việc ghi nhận tác giả và các tác giả đã tích hợp thêm một số tính năng thông qua việc kiểm tra thủ công Wikipedia của họ tập dữ liệu.

Tần suất “i” nhỏ: “i” nhỏ thay cho “I” thường được một số người sử dụng Những người đóng góp cho Wikipedia. Thật thú vị khi những người đóng góp dễ mắc phải lỗi này.

Dấu chấm không có tần suất màu trắng: Nhiều người viết quên thêm khoảng trắng sau dấu chấm, và đây được coi là một đặc điểm để phân biệt các tài khoản bù nhìn.

Tần suất câu hỏi: Một số tác giả sử dụng dấu chấm hỏi thường xuyên hơn những tác giả khác. Vì vậy, đây là một tính năng đặc biệt vì các tác giả cho rằng một số nhà văn lạm dụng việc sử dụng dấu chấm hỏi cho các câu không cần dấu chấm hỏi hoặc sử dụng nhiều dấu hỏi dấu hiệu mà chỉ cần một dấu chấm hỏi là đủ.

Câu có tần suất chữ cái thường: Các tác giả quan sát thấy một sự đồng nhất mẫu viết không bắt đầu câu bằng chữ in hoa và họ coi đây là một tính năng để kiểm tra thói quen viết độc đáo.

Chữ cái, chữ số, chữ hoa, khoảng trắng và tần suất tab: Các tác giả đã đề cập rằng nhóm ký tự này thường thay đổi giữa những người đóng góp cho Wikipedia. Vì vậy, điều này sẽ nắm bắt các tùy chọn định dạng của văn bản như "số không" và "số một" thay vì "số 0" và "số 1" và chữ in hoa cho mỗi từ.

Tần suất lỗi “A” và “an”: Người dùng Wikipedia thường mắc lỗi trong khi gõ “a” và “an'. Nhiều người sáng tạo nội dung đã quen với những lỗi như vậy và việc xem xét những điều đó có thể giúp chúng ta phát hiện ra các trường hợp giả mạo.

Tần suất “anh ấy” và “cô ấy”: Việc lựa chọn “anh ấy” và “cô ấy” được ưu tiên cho mỗi người
người đóng góp. Các tác giả đã đề cập rằng bất kỳ người đóng góp nào sử dụng “anh ấy” hoặc “cô ấy” cho một
chủ đề không xác định được áp dụng thống nhất trong các lần chỉnh sửa hoặc bình luận ở nhiều bài viết hoặc trang thảo luận khác nhau.

Chúng tôi đã tính trung bình tất cả các tính năng được liệt kê ở trên trong số những đóng góp của người dùng giống nhau khi
đưa chúng vào đầu vào cho các bộ phân loại học máy cổ điển. Đồng thời, chúng tôi
xem xét chuỗi đặc điểm đầu vào của LSTM.

Yamak et al. [31] đã thử nghiệm một số loại tính năng trong công việc của họ. Những tính năng đó
được liệt kê dưới đây.

Số lượng đóng góp của người dùng theo không gian tên: Đóng góp của người dùng là
về cơ bản được phân loại thành sáu loại. Đây là bài viết, thảo luận bài viết, trang người dùng, người dùng
trang thảo luận, không gian tên dự án và các trang khác (tất cả các không gian tên khác đều nằm trong trang này)
Các tác giả cho rằng các danh mục được đề cập ở trên là quan trọng nhất
về việc phát hiện hành vi viết và sở thích của người dùng Wikipedia.

Trung bình số byte được thêm vào và xóa khỏi mỗi lần sửa đổi: Với mong muốn
xác định các mẫu văn bản về hành vi của người dùng, các tác giả đã tính toán mức trung bình của
số lượng byte thông tin được thêm vào bài viết cho tất cả các đóng góp
(sửa đổi) của mỗi tài khoản. Họ cũng tính toán số byte trung bình của
thông tin đã xóa trong các bài viết cho mỗi đóng góp của tài khoản. Giả thuyết của họ là
sự thao túng của những người đóng góp cho Wikipedia có thể được kiểm tra thông qua việc thêm/xóa
hành vi.

Đóng góp trung bình trong cùng một bài viết: Ý tưởng đằng sau việc đưa vào
Tính năng này dùng để tính toán số lần trung bình một tác giả đóng góp cho một bài viết.

Các tác giả cho rằng những kẻ thao túng thư ờng cố gắng thao túng cùng một bài viết nhiều lần
lần.

Khoảng thời gian giữa lần đăng ký của ngư ời dùng và lần đóng góp đầu tiên của ngư ời đó: Đối với điều này
tính năng, các tác giả đã tính toán sự khác biệt giữa thời điểm đăng ký và thời điểm
đóng góp đầu tiên trong EnWiki của mỗi tài khoản. Họ cho rằng ngư ời dùng sockpuppet tạo ra
nhiều tài khoản lúc đầu và sau đó để chúng không sử dụng. Tuy nhiên, những bản sao lư u này
tài khoản sẽ đư ợc cấp lại khi một tài khoản đang hoạt động bị chặn.

Tần suất quay lại sau mỗi đóng góp trong cùng một bài viết :
Giả thuyết cơ bản cho tính năng này là hầu hết sự thao túng của ngư ời dùng rồi
sẽ đư ợc hoàn nguyên bởi ngư ời dùng khác vì thông thư ờng có nhiều ngư ời đóng góp quản lý từng trang.
Bất cứ khi nào tìm thấy một đóng góp độc hại, họ thư ờng trực tiếp hoàn nguyên chúng.

Tính năng cuối cùng xem xét liệu một chỉnh sửa đã đư ợc ngư ời dùng khác hoàn nguyên hay chư a,
khiến việc phát hiện không hoàn toàn tự động vì cần có sự tham gia của con ngư ời. Như chúng tôi đề xuất
một phư ơ ng pháp phát hiện tự động không dựa vào đầu vào của con ngư ời, chúng tôi đã không bao gồm
tính năng dựa trên việc hoàn nguyên trong việc triển khai phư ơ ng pháp Yamak et al. [31] của chúng tôi để công bằng hơ n
so sánh. Chúng tôi cũng loại trừ khoảng thời gian giữa lần đăng ký của ngư ời dùng và lần đầu tiên
đóng góp vì chúng tôi không có thông tin này trong tập dữ liệu của mình.

3.6.3 So sánh với ORES

Dịch vụ đánh giá sửa đổi khách quan (ORES) là một dịch vụ dựa trên máy học
hệ thống dự đoán như một dịch vụ web cung cấp các dịch vụ cho các dự án Wikimedia như
Wikipedia và Wikidata. Một hệ thống như vậy đư ợc thiết kế để giúp các biên tập viên thực hiện
nhiệm vụ phức tạp trong khi xem xét Wikipedia như một nguồn thông tin [14]. Ngoài ra,
ORES có thể phát hiện hành vi phá hoại và xóa các chỉnh sửa không đư ợc thực hiện một cách thiện chí. ORES là

đư ợc phát triển bởi Nền tảng chấm điểm Wikimedia [14]. Họ là chuyên gia trong việc phát triển dễ dàng-
để truy cập các mô hình dựa trên AI (Trí tuệ nhân tạo) minh bạch và có đạo đức. Điều này
công cụ truy cập mở hỗ trợ con ngư ời ra quyết định.

ORES đư ợc thiết kế như một dịch vụ phụ trợ và đư ợc dự định để tạo ra các cấu trúc
thông tin của các nhà phát triển. Để lấy điểm ORES, một API điểm đơ n giản (Ứng dụng
Giao diện lập trình) và giao diện ngư ời dùng tham chiếu (Giao diện ngư ời dùng) có sẵn [14]. Nhiều
các nhà nghiên cứu cũng truy cập ORES thông qua các công cụ của bên thứ ba do các tình nguyện viên phát triển.

Chúng tôi cũng đã sử dụng API có sẵn để thu thập điểm ORES cho mỗi lần chỉnh sửa cho cả hai
những đóng góp của ngư ời dùng lành tính và ngư ời dùng bù nhìn Cụ thể hơ n, với một bản chỉnh sửa, ORES
cung cấp phân phối xác suất (điểm chất lư ợng bản thảo) của việc nằm trong một trong những điều sau đây
bốn lớp: thư rác, phá hoại, tấn công hoặc OK. Các loại có vấn đề nghiêm trọng nhanh hơ n
các bài viết dự thảo bị xóa, càng tốt. Chúng tôi đã tính trung bình điểm chất lư ợng dự thảo của tất cả các bản chỉnh sửa
của cùng một ngư ời dùng khi sử dụng các thuật toán học máy cổ điển trong khi chúng tôi xem xét
trình tự điểm chất lư ợng bản nháp cho các lần chỉnh sửa của cùng một ngư ời dùng khi nhập vào LSTM.

CHƯƠNG BỐN: KẾT QUẢ THÍ NGHIỆM

Phần này sẽ xem xét các kết quả thử nghiệm từ máy học của chúng tôi và phương pháp tiếp cận dựa trên mạng nơ-ron. Đối với điều này, chúng tôi đã sử dụng các tính năng được mô tả trong phần phương pháp luận. Chúng tôi đã thực hiện phân tích của mình bằng cách xem xét tất cả các tính năng được mô tả trong Phần 3 để xác định các tài khoản bù nhìn trên Wikipedia. Chi tiết của Các bước phân tích và công việc được trình bày trong chương này.

4.1 Kích thước tập dữ liệu cuối cùng

Chúng tôi đã sử dụng toàn bộ tập dữ liệu của các mẫu dương tính và âm tính được đề cập trong phần phương pháp luận. Sau khi thu thập và làm sạch, chúng tôi đã có một tập dữ liệu gần như cân bằng. Tuy nhiên, số lượng đóng góp của tổng số lần chỉnh sửa của các tài khoản đó lại khác nhau. Số lượng mẫu tập dữ liệu cuối cùng được liệt kê trong Bảng 4.1.

Bảng 4.1 Số lượng mẫu cuối cùng

	Dữ liệu tích cực	Dữ liệu tiêu cực
Số lượng người dùng	17.180	16.043
Tổng số lần chỉnh sửa	420,111	393,950

4.2 Quy trình và thiết lập thí nghiệm

Như đã mô tả trong phần phương pháp luận, chúng tôi đã sử dụng một số phân loại các thuật toán cho các tính năng của chúng tôi để xây dựng một mô hình lý tưởng cho việc tách các tài khoản thật khỏi nhiều chủ tài khoản. Bộ dữ liệu mà chúng tôi sử dụng khá cân bằng. Vì vậy, chúng tôi đã làm không cần sử dụng bất kỳ kỹ thuật mất cân bằng lớp nào. Tuy nhiên, để an toàn hơn, chúng ta có

tập trung vào xác thực chéo phân tầng. Chúng tôi đã thực hiện xác thực chéo 5 lần. Để đo lường
Về hiệu suất, chúng tôi đã xem xét điểm F1.

4.3 Kết quả của các tính năng được đề xuất của chúng tôi

Kết quả của các mô hình học máy khác nhau với các tính năng được đề xuất của chúng tôi được hiển thị
trong Bảng 4.2. Như chúng ta có thể thấy, trong số tất cả các mô hình học máy được xem xét, Random
Forest đạt được điểm F1 tốt nhất là 0,82. Hơn nữa, các mô hình này hoạt động tốt hơn
LSTM đạt điểm F1 thấp hơn là 0,75.

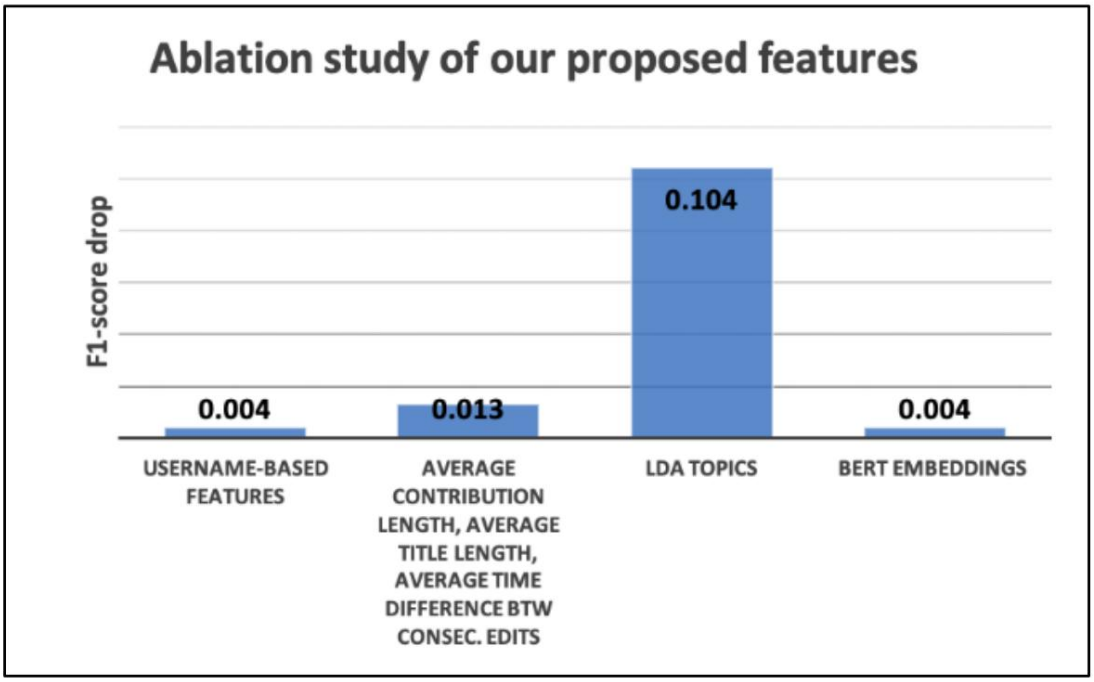
Bảng 4.2 So sánh điểm F1 của các mô hình học máy khác nhau với các tính năng được đề xuất của chúng tôi trong đầu vào để dự đoán các tài khoản sockpuppet. Điểm tốt nhất được in đậm.

Phân loại	Điểm F1
Rừng ngẫu nhiên	0,82
Hồi quy logistic	0,75
Phân loại cây bổ sung	0,75
Gaussian Naive Bayes	0,60
Cây quyết định	0,75
Bộ phân loại MLP	0,77
LSTM	0,75

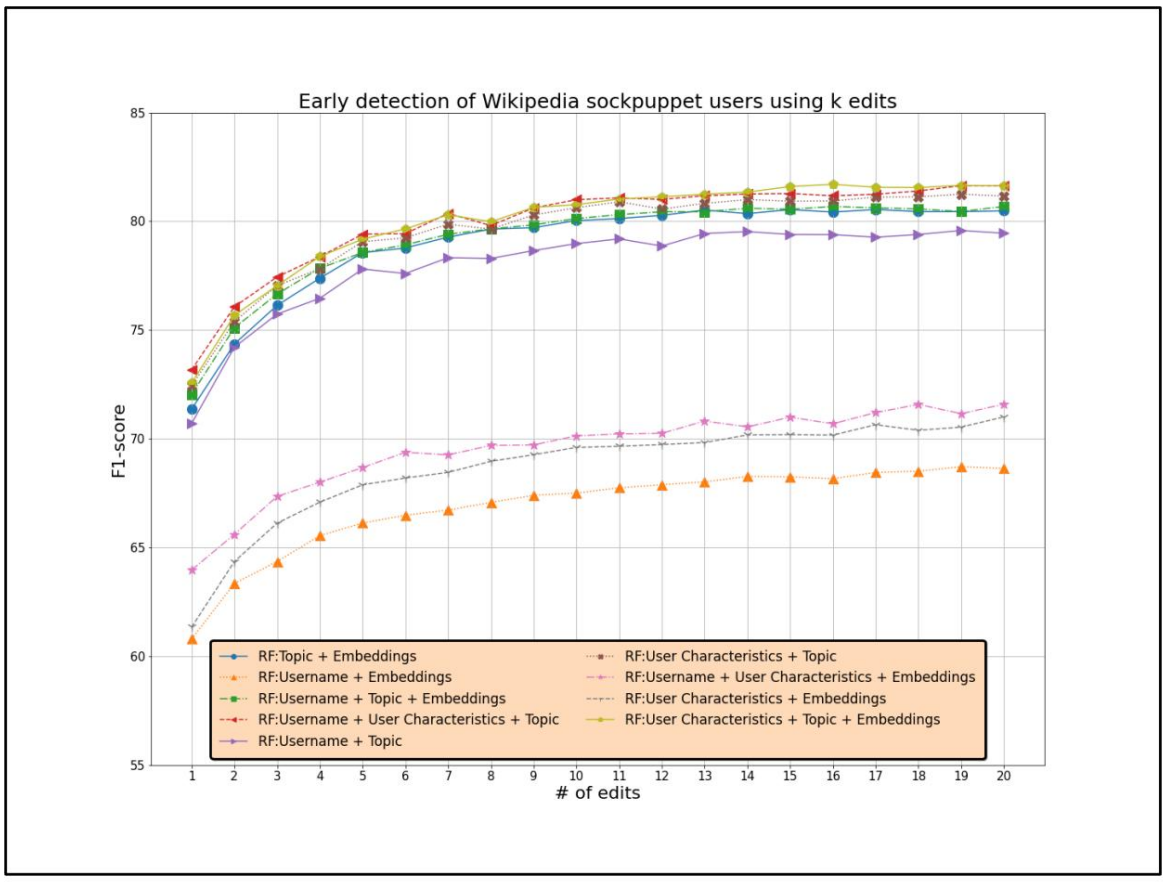
4.4 Phân tích tính năng

Để đo lường tầm quan trọng của tính năng, chúng tôi đã thực hiện loại bỏ tính năng, tức là đối với mỗi
nhóm, g các tính năng được xem xét đã được di chuyển và thực hiện phân loại với
các tính năng còn lại. Điểm F1 giảm càng cao thì nhóm càng quan trọng
các tính năng cho nhiệm vụ phân loại. Kết quả được hiển thị trong Hình 4.1 và 4.2. Như chúng ta có thể
xem, nhóm tính năng quan trọng nhất là nhóm chủ đề LDA, vì việc loại bỏ nó sẽ làm giảm

điểm F1 là 0,71 cho 20 lần chỉnh sửa. Nhóm tính năng quan trọng thứ hai chứa
độ dài đóng góp trung bình, độ dài tiêu đề trung bình và chênh lệch thời gian trung bình
giữa hai lần chỉnh sửa liên tiếp. Việc xóa nhóm tính năng này sẽ làm giảm điểm F1 xuống
0.81. Các tính năng dựa trên tên ngư ời dùng và nhúng BERT của bình luận ngư ời dùng đều như nhau
quan trọng, và việc loại bỏ một trong số chúng sẽ làm giảm nhẹ điểm F1. Việc loại bỏ cả hai
chúng làm giảm điểm F1 xuống 0,81. Hình 4.2 đảm bảo mô hình này là nhất quán ngay cả khi ít hơ n
các chỉnh sửa đang đư ợc xem xét.



Hình 4.1 Nghiên cứu cắt bỏ các tính năng đề xuất của chúng tôi: điểm F1 giảm cho mỗi
nhóm tính năng đư ợc xem xét



Hình 4.2 Nghiên cứu cắt bỏ các tính năng đư ợc đề xuất của chúng tôi cho k lần chỉnh sửa

4.5 So sánh phư ơ ng pháp đề xuất của chúng tôi với công trình liên quan

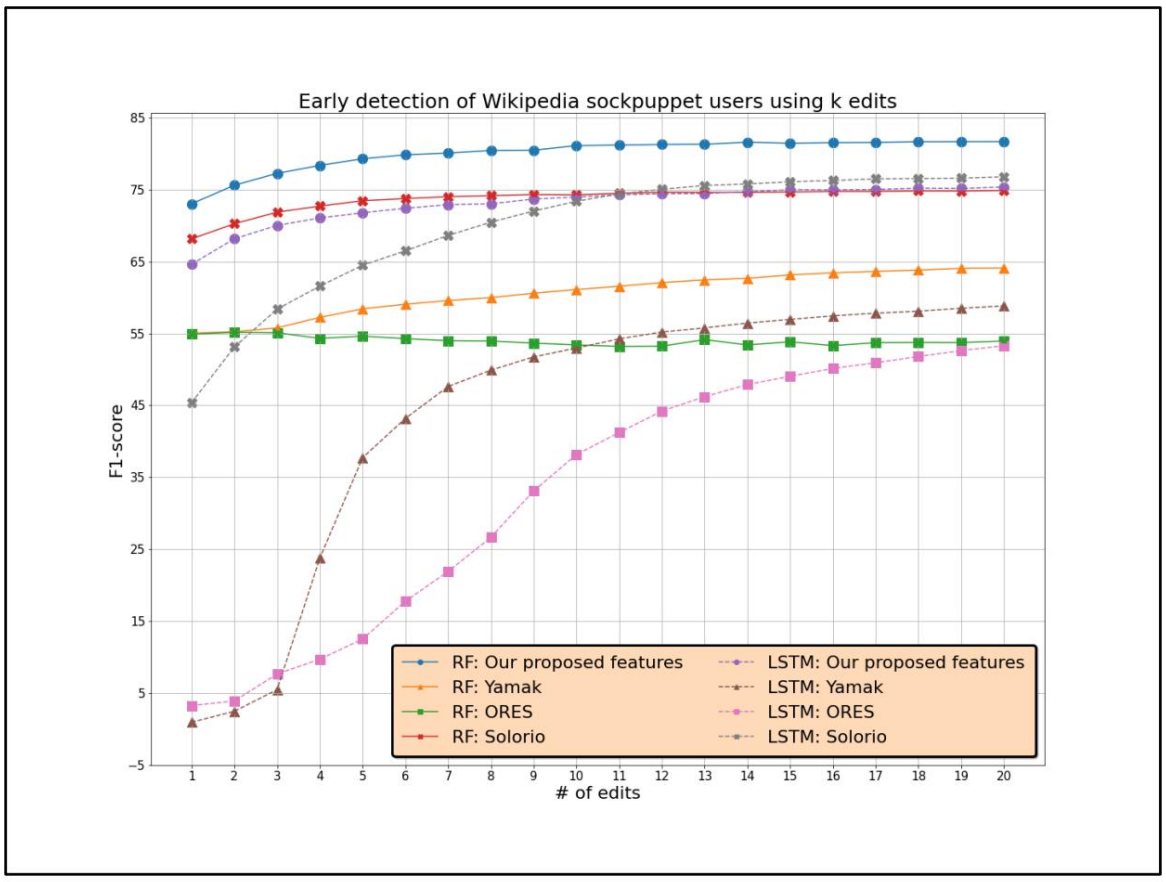
Điểm F1 của phư ơ ng pháp tiếp cận đư ợc đề xuất của chúng tôi và các đối thủ cạnh tranh đư ợc xem xét đư ợc hiển thị trong Bảng 4.3, trong đó chúng tôi cũng so sánh các tính năng trong đầu vào với máy cổ điển tốt nhất mô hình học tập (Random Forest trong trư ờng hợp của tất cả các đối thủ cạnh tranh) và LSTM. Như chúng ta có thể thấy, phư ơ ng pháp tiếp cận đư ợc đề xuất của chúng tôi đạt đư ợc điểm F1 cao hơ n là 0,82 so với ORES với Rừng ngẫu nhiên (RF), Yamak et al. [31] với RF, và Solorio et al. [29] với LSTM, trong đó đạt đư ợc điểm F1 lần lư ợt là 0,54, 0,64 và 0,77.

Bảng 4.3 so So sánh điểm F1 của các tính năng đề xuất của chúng tôi với công việc liên quan. Chúng tôi sánh các tính năng trong đầu vào của Random forest (mang lại thuật toán học máy cổ điển tốt nhất) và LSTM. Điểm số tốt nhất đư ợc in đậm.

	Rừng ngẫu nhiên	LSTM
Các tính năng chúng tôi đề xuất	0,82	0,75
Quặng	0,54	0,53
Yamak	0,64	0,59
Solorio	0,75	0,77

4.6 Phát hiện sớm các tài khoản Wikipedia Sockpuppet

Chúng tôi nghiên cứu tác động của những chỉnh sửa đầu tiên do ngư ời dùng thực hiện lên điểm số dự đoán F1. Hình 4.3 cho thấy sự thay đổi trong điểm số F1 khi k thay đổi từ 1 đến 20. Chúng tôi cho thấy các tính năng so với các tính năng công việc liên quan trong đầu vào của Random Forest và LSTM. bộ tính năng đư ợc đề xuất có thể phát hiện tài khoản sockpuppet có điểm F1 là 0,73 bằng chỉ cần xem xét lần chỉnh sửa đầu tiên của ngư ời dùng (so với 0,68 đạt đư ợc bởi Solorio et al. [29]) và F1 điểm 0,80 khi xem xét sáu lần chỉnh sửa đầu tiên. Hơ n nữa, Random Forest luôn tốt hơ n hơ n LSTM, đặc biệt là đối với dự đoán sớm. Ngoại lệ duy nhất đư ợc đư a ra bởi Solorio et al. [29], trong đó LSTM tốt hơ n một chút khi bắt đầu từ 12 lần chỉnh sửa.



Hình 4.3 Phát hiện sớm các tài khoản bù nhìn Wikipedia

4.7 Trả lời các câu hỏi nghiên cứu

Trong phần này, chúng tôi sẽ cố gắng tóm tắt những phát hiện của mình để trả lời cho nghiên cứu những câu hỏi đư ợc đặt ra ở phần đầu của nghiên cứu.

RQ1: Những mô hình tài khoản bù nhìn do ngư ời điều khiển bù nhìn tạo ra là gì?

Trả lời: Bằng cách phân tích các tính năng có trong nghiên cứu của chúng tôi, chúng tôi thấy rằng sockpuppet tài khoản đóng góp ngắn hơ n so với ngư ời dùng lành tính (trung bình trung bình độ dài đóng góp là 27 so với 31 ký tự) và chỉnh sửa các trang có tiêu đề dài hơ n (trung bình độ dài tiêu đề trung bình là 18 đối với sockpuppets so với 17 ký tự đối với ngư ời dùng lành tính), và chỉnh sửa thêm thư ờng xuyên (khoảng thời gian chênh lệch trung bình giữa hai lần chỉnh sửa liên tiếp là 3,5 ngày so với

17 ngày đối với ngư ời dùng lành tính). Vì vậy, nhìn chung, tài khoản bù nhìn của ngư ời điều khiển rối có mẫu đóng góp đặc biệt so với ngư ời dùng vô tội.

RQ2: Phân tích ngữ nghĩa từ các bản chỉnh sửa có nắm bắt đư ợc mô hình viết và các trang đóng góp tinh vi hơ n và xác định các tài khoản bù nhìn tốt hơ n hơ n là các tác phẩm dựa trên mạng cú pháp, phong cách và đồ thị và đư a ra một cấp độ sâu sắc có ý nghĩa theo ngữ cảnh?

Trả lời: Phân tích ngữ nghĩa đư ợc chọn lọc của chúng tôi từ các bản chỉnh sửa đã nắm bắt đư ợc các mẫu viết tốt hơ n hơ n các tác phẩm dựa trên mạng cú pháp, phong cách và đồ thị. Mô hình dựa trên RF của chúng tôi thực hiện tốt hơ n phư ơ ng pháp đã thiết lập và đư a ra mức độ ngữ cảnh sâu sắc nghĩa.

RQ3: Có thể phát hiện sớm các tài khoản sockpuppet và đề xuất không? đình chỉ?

Trả lời: Cách tiếp cận đư ợc mô tả của chúng tôi có thể phát hiện sớm các tài khoản bù nhìn bằng cách xem xét 20 lần chỉnh sửa đầu tiên của ngư ời dùng và đạt đư ợc điểm F1 là 0,73 chỉ bằng cách xem xét lần chỉnh sửa đầu tiên (so với số điểm 0,68 đạt đư ợc bởi đối thủ cạnh tranh tốt nhất). Vì vậy, có thể phát hiện ra sockpuppet tài khoản ngay sau khi họ bắt đầu đóng góp.

CHƯƠNG NĂM: KẾT LUẬN

5.1 Chúng tôi đã làm được gì cho đến nay?

Trong nghiên cứu này, chúng tôi trình bày phương pháp tiếp cận được đề xuất của mình để giải quyết vấn đề tự động xác định tài khoản con rối trên Wikipedia. Chúng tôi xử lý vấn đề như một nhiệm vụ phân loại nhị phân và đề xuất một bộ tính năng mới để nắm bắt hành vi đáng ngờ xem xét hoạt động của người dùng và phân tích nội dung đóng góp. Cụ thể, nội dung-các tính năng dựa trên chưa từng được xem xét trước đây và tạo nên sự mới lạ trong công việc của chúng tôi.

Chúng tôi đã thử nghiệm cách tiếp cận của mình trên một tập dữ liệu chúng tôi thu thập được có chứa 17.000 tài khoản được Wikipedia xác nhận là con rối. Kết quả thử nghiệm cho thấy đề xuất của chúng tôi phương pháp có thể phát hiện tài khoản sockpuppet với điểm F1 là 0,82 (so với điểm 0,77 đạt được bởi đối thủ cạnh tranh tốt nhất) bằng cách xem xét 20 lần chỉnh sửa đầu tiên của người dùng và 0,73 chỉ bằng xem xét lần chỉnh sửa đầu tiên (so với điểm 0,68 đạt được bởi đối thủ cạnh tranh tốt nhất). Chúng tôi cũng cho thấy rằng việc tính toán các chủ đề đóng góp của người dùng đặc biệt quan trọng đối với phát hiện những loại tài khoản độc hại này. Chúng tôi cũng có thể phân biệt mẫu chung của người dùng sockpuppet như độ dài đóng góp trung bình và thời gian trung bình sự khác biệt giữa hai lần chỉnh sửa liên tiếp khác biệt đáng kể so với người dùng xác thực tài khoản. Nhìn chung, chúng ta đã thấy tầm quan trọng của các tính năng cấp độ ngữ nghĩa đối với phát hiện rối sockpuppetry so với các phương pháp tiếp cận riêng biệt khác đã được thiết lập trước đó. Của chúng tôi phân tích cũng bao gồm việc phát hiện sớm các tài khoản không trung thực để loại bỏ chúng đóng góp trong thời gian nhanh chóng.

5.2 Hướng đi trong tương lai

Là một phần của công việc trong tương lai, chúng tôi dự định thử nghiệm các tính năng của mình để dự đoán xem hai tài khoản thuộc về cùng một cuộc điều tra con rối. Trong suốt công việc hiện tại, chúng tôi đã tập trung vào việc phát hiện xem một tài khoản có phải là con rối hay không. Để mở rộng hiện tượng như vậy, chúng tôi muốn làm việc trong tương lai để đánh giá xem hai tài khoản có được liên kết với nhau theo cùng một cuộc điều tra.

Chúng tôi cũng quan tâm đến các nền tảng đa phương tiện. Ví dụ, chúng tôi sẽ kiểm tra xem cùng một nhóm sockpuppets tồn tại trên cả Facebook và Twitter. Phân tích như vậy sẽ là điều cơ bản là phải nhận ra liệu người dùng lạm dụng chỉ tập trung vào một nền tảng hay thực hiện hành vi tương tự hành vi trên bất kỳ nền tảng nào khác. Động lực của nghiên cứu là kiểm tra xem các nhà quảng cáo, những kẻ gửi thư rác và người quảng bá, bất kể nền tảng xã hội nào, đều hoạt động theo một mô hình tương tự hoặc thành lập một nhóm để thực hiện hoạt động tàn ác như vậy. Một mô hình tập hợp có khả năng kết hợp dữ liệu từ nhiều nền tảng và phân tích sockpuppetry sẽ đảm bảo tính toàn diện cải thiện chức năng theo dõi nhiều chủ tài khoản.

TÀI LIỆU THAM KHẢO

[1] K. Luyckx và W. Daelemans, “Ghi nhận tác giả và xác minh với nhiều
tác giả và dữ liệu hạn chế,” trong Biên bản Hội nghị quốc tế lần thứ 22 về
Ngôn ngữ học tính toán (Coling 2008). Manchester, Vương quốc Anh: Coling 2008
Ban Tổ chức, tháng 8 năm 2008, trang 513–520. [Trực tuyến]. Có sẵn:
https://aclanthology.org/C08-1065

[2] K. Luyckx và W. Daelemans, “Personae: một tập hợp các tác giả và nhân cách
dự đoán từ văn bản,” trong Biên bản Hội nghị quốc tế lần thứ sáu về
Tài nguyên ngôn ngữ và đánh giá (LREC'08). Marrakech, Morocco: Châu Âu
Hiệp hội Tài nguyên Ngôn ngữ (ELRA), tháng 5 năm 2008. [Trực tuyến]. Có sẵn:
http://www.lrec-conf.org/proceedings/lrec2008/pdf/759paper.pdf

[3] K. Luyckx và W. Daelemans, “Ảnh hưởng của kích thước tập hợp tác giả và kích thước dữ liệu trong
ghi nhận quyền tác giả,” LLC, tập 26, trang 35–55, 04 2011.

[4] HJ Escalante, T. Solorio, và M. Montes-y Gómez, “Biểu đồ tần suất cục bộ của đặc điểm
n-gram để xác định tác giả,” trong Biên bản cuộc họp thường niên lần thứ 49 của
Hiệp hội Ngôn ngữ học tính toán: Công nghệ ngôn ngữ của con người.
Portland, Oregon, Hoa Kỳ: Hiệp hội Ngôn ngữ học tính toán, tháng 6 năm 2011,
trang 288–298. [Trực tuyến]. Có sẵn: https://aclanthology.org/P11-1030

[5] S. Raghavan, A. Kovashka và R. Mooney, “Ghi nhận tác giả bằng cách sử dụng
ngữ pháp không ngữ cảnh xác suất,” trong Biên bản báo cáo của ACL 2010
Bài báo ngắn về Hội nghị. Uppsala, Thụy Điển: Hiệp hội tính toán
Ngôn ngữ học, tháng 7 năm 2010, trang 38–42. [Trực tuyến]. Có sẵn:
https://aclanthology.org/P10-2008

[6] V. Keselj, F. Peng, N. Cercone và C. Thomas, “Hồ sơ tác giả dựa trên N-gram cho
ghi nhận quyền tác giả,” 2003.

[7] E. Stamatatos, “Xác định tác giả bằng cách sử dụng các văn bản đào tạo mất cân bằng và hạn chế,” trong
Hội thảo quốc tế lần thứ 18 về ứng dụng cơ sở dữ liệu và hệ thống chuyên gia
(DEXA 2007), 2007, trang 237–241.

[8] M. Koppel, J. Schler và S. Argamon, “Ghi nhận tác giả trong tự nhiên,” Lang.
Tài nguyên. Đánh giá, tập 45, số 1, trang 83–94, tháng 3 năm 2011. [Trực tuyến]. Có sẵn:
https://doi.org/10.1007/s10579-009-9111-2

[9] “Wikipedia,” https://www.britannica.com/topic/Wikipedia, truy cập: 2022-07-06.

[10] “Api:main page,” https://www.mediawiki.org/wiki/API:Mainpage, truy cập: 2022-
07-06.

[11] R. Zheng, J. Li, H. Chen và Z. Huang, “Một khuôn khổ để xác định tác giả
của các tin nhắn trực tuyến: Các đặc điểm về phong cách viết và các kỹ thuật phân loại,” Tạp chí
của Hiệp hội Khoa học và Công nghệ Thông tin Hoa Kỳ, tập 57, số 3,
trang 378–393, 2006. [Trực tuyến]. Có sẵn:
https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20316

[12] “Phân bổ Dirichlet tiềm ẩn,”
https://radimrehurek.com/gensim/models/ldamodel.html, truy cập: 2022-07-06.

[13] S. Hochreiter và J. Schmidhuber, “Bộ nhớ dài hạn ngắn”, Neural Comput., tập.
9, số 8, trang 1735–1780, tháng 11 năm 1997. Có sẵn:
https://doi.org/10.1162/neco.1997.9.8.1735

[14] “Ores,” https://www.mediawiki.org/wiki/ORES, truy cập: 2022-07-06.

[15] J. Zhu, Y. Xia, L. Wu, D. He, T. Qin, W. Zhou, H. Li, và T.-Y. Liu, “Kết hợp
bert vào dịch máy thần kinh”, bản in trư ớc arXiv arXiv:2002.06823, 2020.

[16] U. Zaratiana, P. Holat, N. Tomeh, và T. Charnois, “Máy biến áp phân cấp
Mô hình nhận dạng thực thể có tên khoa học,” bản in điện tử arXiv
arXiv:2203.14710, tháng 3 năm 2022.

[17] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, CL Zitnick,
J. Ma và R. Fergus, “Cấu trúc và chức năng sinh học xuất hiện từ việc mở rộng quy mô
học không giám sát tới 250 triệu chuỗi protein,” Biên bản báo cáo
Viện Hàn lâm Khoa học Quốc gia, tập 118, số 15, trang e2016239118, 2021. [Trực tuyến].
Có sẵn:
https://www.pnas.org/doi/abs/10.1073/pnas.2016239118

[18] A. Nambiar, S. Liu, M. Hopkins, M. Heflin, S. Maslov và A. Ritz, “Chuyển đổi
ngôn ngữ của sự sống: Mạng nơ -ron biến đổi cho nhiệm vụ dự đoán protein,”
bioRxiv, 2020. Có sẵn:
https://www.biorxiv.org/content/early/2020/06/16/2020.06.15.153643

[19] R. Rao, N. Bhattacharya, N. Thomas, Y. Duan, X. Chen, J. Canny, P. Abbeel, và Y.
S. Song, “Đánh giá quá trình học chuyển giao protein bằng băng”, bioRxiv, 2019.
[Trực tuyến]. Có sẵn: https://www.biorxiv.org/content/early/2019/06/20/676825

[20] “Wikipedia:mục đích,” https://en.wikipedia.org/wiki/Wikipedia:Mục đích, truy cập:
2022-07-06.

[21]“Wikipedia:sockpuppetry,”https://en.wikipedia.org/wiki/Wikipedia:Sockpuppetry,
truy cập: 2022-07-06.

[22] S. Kumar, F. Spezzano, và VS Subrahmanian, “VEWS: Một kẻ phá hoại wikipedia sớm
hệ thống cảnh báo,” trong Biên bản của Hội nghị quốc tế ACM SIGKDD lần thứ 21
Hội nghị về Khám phá tri thức và Khai thác dữ liệu, 2015. ACM, 2015, tr.
607–616.

[23] T. Green và F. Spezzano, “Xác định ngư ời dùng thư rác trong wikipedia thông qua chỉnh sửa
hành vi,” trong Biên bản Hội nghị quốc tế lần thứ mư ời một về Web và
Social Media, 2017. AAAI Press, 2017, trang 532–535.

[24] N. Joshi, F. Spezzano, M. Green và E. Hill, “Phát hiện biên tập trả tiền không đư ợc tiết lộ trong
wikipedia,” trong Biên bản Hội nghị Web 2020, 2020, trang 2899–2905.

[25] B. Viswanath, A. Post, KP Gummadi, và A. Mislove, “Một phân tích về xã hội
phòng thủ sybil dựa trên mạng,” ACM SIGCOMM Computer Communication
Tạp chí, tập 41, số 4, trang 363–374, 2011.

[26] Z. Bu, Z. Xia và J. Wang, “Một thuật toán phát hiện con rối tất trên không gian ảo,”
Hệ thống dựa trên tri thức, tập 37, trang 366–377, 2013.

[27] D. Liu, Q. Wu, W. Han, và B. Zhou, “Phát hiện băng đảng bù nhìn trên phư ơ ng tiện truyền thông xã hội
các trang web,” Biên giới của Khoa học máy tính, tập 10, số 1, trang 124–135, 2016.

[28] S. Kumar, J. Cheng, J. Leskovec, và VS Subrahmanian, “Một đội quân của tôi:
“Những con rối trong cộng đồng thảo luận trực tuyến,” trong Biên bản báo cáo của Hội nghị lần thứ 26
Hội nghị quốc tế về World Wide Web, WWW 2017, 2017, trang 857–866.

[29] T. Solorio, R. Hasan và M. Mizan, “Một nghiên cứu điển hình về phát hiện rối trong
wikipedia,” trong Biên bản Hội thảo về Phân tích Ngôn ngữ trong Xã hội
Phư ơ ng tiện truyền thông tại NAACL HTL, 2013, trang 59–68.

[30] T. Solorio, R. Hasan và M. Mizan, “Phát hiện con rối trong wikipedia: Một kho văn bản
của văn bản lừa dối trong thế giới thực để liên kết danh tính,” bản in trư ớc của arXiv
arXiv:1310.6772, 2013.

[31] Z. Yamak, J. Saunier và L. Vercouter, “Phát hiện sự thao túng nhiều danh tính
trong các dự án hợp tác,” Biên bản Hội nghị quốc tế lần thứ 25
Bạn đồng hành trên World Wide Web, 2016.

[32] Z. Yamak, J. Saunier và L. Vercouter, “Sockscatch: Phát hiện tự động và
nhóm rối trong phư ơ ng tiện truyền thông xã hội,” Hệ thống dựa trên kiến thức, tập 149,
trang 124–142, 2018.

[33] M. Tsikerdekis và S. Zeadally, “Phát hiện lừa đảo danh tính nhiều tài khoản trong
phư ơ ng tiện truyền thông xã hội sử dụng hành vi phi ngôn ngữ,” Giao dịch IEEE về thông tin
Khoa học pháp y và an ninh, tập 9, số 8, trang 1311–1321, 2014.

[34] X. Zheng, YM Lai, K. Chow, LC Hui và S. Yiu, “Phát hiện rối tất trong
diễn đàn thảo luận trực tuyến”, luận án tiến sĩ, Đại học Hồng Kông, 2011.

[35] SK Maity, A. Chakraborty, P. Goyal, và A. Mukherjee, “Phát hiện
“con rối trong phư ơ ng tiện truyền thông xã hội,” trong Companion của Hội nghị ACM năm 2017 về
Công việc hợp tác đư ợc hỗ trợ bằng máy tính và máy tính xã hội, 2017, trang 243–
246.

[36] S. Adhikari, “Phát hiện tài khoản sockpuppet trên reddit,” 2020.

[37] R. Zafarani và H. Liu, “10 điều bất ngờ: Phát hiện ngư ời dùng độc hại với
thông tin tối thiểu,” trong Biên bản báo cáo của Hội nghị quốc tế ACM lần thứ 24
Hội nghị về Quản lý thông tin và tri thức, CIKM 2015, 2015, tr.
423–431.

[38] K. Lee, BD Eoff, và J. Caverlee, “Bảy tháng với quỷ dữ: Một
nghiên cứu về những kẻ gây ô nhiễm nội dung trên Twitter,” trong Biên bản của Hội nghị quốc tế lần thứ năm
Hội nghị về Weblog và phư ơ ng tiện truyền thông xã hội, Barcelona, Catalonia, Tây Ban Nha, ngày 17 tháng 7
21, 2011, LA Adamic, R. Baeza-Yates, và S. Counts, Biên tập viên. AAAI Press,
2011.

[39] J. Devlin, M.-W. Chang, K. Lee, và K. Toutanova, “Bert: Đào tạo trư ớc về sâu
máy biến áp hai chiều để hiểu ngôn ngữ,” bản in trư ớc của arXiv
arXiv:1810.04805, 2018.