Machine Translated by Google

scientific reports

Check for updates

MỞ

Phương pháp phát hiện hiệu quả các trang web lừa đảo sử dụng tính năng URL và HTML

AliAljofey1,2, Qingshan Jiang1*, Abdur Rasool1,2, Hui Chen1,2, Wenyin Liu3 , Khương Qu1 & YangWang4

Các trang web lừa đảo đang phát triển ngày nay đặt ra những mối đe dọa đáng kể do rủi ro cực kỳ khó phát hiện. Họ dự đoán người dùng Internet sẽ nhầm chúng là hàng thật nhằm tiết lộ thông tin và quyền riêng tư của người dùng, chẳng hạn như id đăng nhập, mật khẩu, số thẻ tín dụng, v.v. mà không cần thông báo trước. Bài viết này đề xuất một cách tiếp cận mới để giải quyết vấn đề chống lừa đảo. Các tính năng mới của phương pháp này có thể được biểu thị bằng chuỗi ký tự URL mà không cần biết trước về lừa đảo, các thông tin siêu liên kết khác nhau và nội dung văn bản của trang web, được kết hợp và cung cấp để huấn luyện bộ phân loại XGBoost. Một trong những đóng góp chính của bài viết này là việc lựa chọn các tính năng mới khác nhau, có đủ khả năng phát hiện các cuộc tấn công 0-h và các tính năng này không phụ thuộc vào bất kỳ dịch vụ nào của bên thứ ba. Cụ thể, chúng tôi trích xuất các tính năng Tần số nghịch đảo tần số tài liệu (TF-IDF) ở cấp độ ký tự từ các phần nhiễu của HTML và văn bản gốc của trang web nhất định. Hơn nữa, các tính năng siêu liên kết được đề xuất của chúng tôi sẽ xác định mối quan hệ giữa nội dung và URL của trang web. Do không có tập dữ liệu lừa đảo lớn có sẵn công khai nên chúng tôi cần tạo tập dữ liệu của riêng mình với 60.252 trang web để xác thực giải pháp được đề xuất. Dữ liệu này chứa 32.972 trang web lành tính và 27.280 trang web lừa đảo. Để đánh giá, hiệu suất của từng danh mục của bộ tính năng đề xuất sẽ được đánh giá và các thuật toán phân loại khác nhau được sử dụng. Từ kết quả thực nghiệm, người ta nhận thấy rằng các tính năng riêng lẻ được đề xuất có giá trị để phát hiện lừa đảo. Tuy nhiên, việc tích hợp tất cả các tính năng sẽ cải thiện khả năng phát hiện các trang web lừa đảo với độ chính xác đáng kể. Phương pháp đề xuất đạt được độ chính xác 96,76% với tỷ lệ dương tính giả chỉ 1,39% trên tập dữ liệu của chúng tôi và độ chính xác 98,48% với tỷ lệ dương tính giả 2,09% trên tập dữ liệu chuẩn, vượt trội so với các phương pháp cơ bản hiện có.

Tội lừa đảo ngày càng gia tăng, gây thiệt hại hàng tỷ USD1 . Trong các cuộc tấn công này, người dùng nhập thông tin quan trọng của họ (tức là chi tiết thẻ tín dụng, mật khẩu, v.v.) vào trang web giả mạo có vẻ hợp pháp. Phần mềm dưới dạng dịch vụ (SaaS) và các trang webmail là mục tiêu phổ biến nhất của lừa đảo2 . Kẻ lừa đảo tạo ra các trang web trông rất giống với các trang web lành tính. Sau đó, liên kết trang web lừa đảo được gửi đến hàng triệu người dùng internet qua email và các phương tiện truyền thông khác. Những kiểu tấn công mạng này thường được kích hoạt bằng email, tin nhắn tức thời hoặc cuộc gọi điện thoại3 . Mục đích của cuộc tấn công lừa đảo không chỉ là đánh cắp nhân cách của nạn nhân mà còn có thể được thực hiện để phát tán các loại phần mềm độc hại khác như ransomware, khai thác điểm yếu của cách tiếp cận hoặc để nhận lợi nhuận bằng tiền4 . Theo báo cáo của Nhóm công tác chống lừa đảo (APWG) trong Quý 3 năm 2020, số vụ tấn công lừa đảo đã tăng lên kể từ tháng 3 và 28.093 trang web lừa đảo duy nhất đã được phát hiện từ tháng 7 đến ngày 2 tháng 9 . Số tiền trung bình được yêu cầu trong các cuộc tấn công Thỏa thuận E-mail Kinh doanh (BEC) chuyển khoản ngân hàng là 48.000 USD trong quý 3, giảm từ 80.000 USD trong quý 2 và 54.000 USD trong quý đầu tiên.

Việc phát hiện và ngăn chặn hành vi lừa đảo là một thách thức lớn đối với các nhà nghiên cứu do cách những kẻ lừa đảo thực hiện cuộc tấn công để vượt qua các kỹ thuật chống lừa đảo hiện có. Hơn nữa, kẻ lừa đảo thậm chí có thể nhắm mục tiêu vào một số người dùng có trình độ học vấn và kinh nghiệm bằng cách sử dụng các chiêu trò lừa đảo mới. Vì vậy, các kỹ thuật phát hiện lừa đảo dựa trên phần mềm được ưa thích hơn để chống lại cuộc tấn công lừa đảo. Hầu hết các phương pháp sẵn có để phát hiện các cuộc tấn công lừa đảo là danh sách đen/danh sách trắng5 , xử lý ngôn ngữ tự nhiên6 , tương tự hình ảnh7 , quy tắc8 , kỹ thuật học máy 9,10, v.v. Các kỹ thuật dựa trên danh sách đen/danh sách trắng không phát hiện được các trang web lừa đảo không được liệt kê (tức là các cuộc tấn công 0-h)

1Phòng thí nghiệm trọng điểm Thâm Quyến về khai thác dữ liệu hiệu suất cao, Viện Công nghệ tiên tiến Thâm Quyến, Viện Hàn lâm Khoa học Trung Quốc, Thâm Quyến 518055, Trung Quốc. 2Trường Cao đẳng Công nghệ Tiên tiến Thâm Quyến, Đại học Viện Hàn lâm Khoa học Trung Quốc, Bắc Kinh 100049, Trung Quốc. 3Khoa Khoa học Máy tính, Đại học Công nghệ Quảng Đông, Quảng Châu, Trung Quốc. 4Trung tâm Điện toán Đám mây, Viện Công nghệ Tiên tiến Thâm Quyến, Trung Quốc Viện Khoa học, Thâm Quyến 518055, Trung Quốc. *email: qs.jiang@siat.ac.cn

Machine Translated by Google
www.nature.com/scientificreports/

cũng như các phương pháp này không thành công khi URL nằm trong danh sách đen gặp phải những thay đổi nhỏ. Trong các kỹ thuật dựa trên máy học, mô hình phân loại được đào tạo bằng cách sử dụng nhiều tính năng phỏng đoán khác nhau (ví dụ: URL, nội dung trang web, lưu lượng truy cập trang web, công cụ tìm kiếm, bản ghi WHOIS và Xếp hạng trang) để cải thiện hiệu quả phát hiện. Tuy nhiên, những tính năng phỏng đoán này không được đảm bảo sẽ xuất hiện trên tất cả các trang web lừa đảo và cũng có thể xuất hiện trên các trang web lành tính, điều này có thể gây ra lỗi phân loại. Hơn nữa, một số tính năng heuristic khó truy cập và phụ thuộc vào bên thứ ba. Một số dịch vụ của bên thứ ba (ví dụ: xếp hạng trang, lập chỉ mục công cụ tìm kiếm, WHOIS, v.v.) có thể không đủ khả năng để xác định các trang web lừa đảo được lưu trữ trên máy chủ bị tấn công và các trang web này bị xác định không chính xác là trang web lành tính vì chúng có trong kết quả tìm kiếm. Các trang web được lưu trữ trên các máy chủ bị xâm nhập thường tồn tại hơn một ngày, không giống như các trang web lừa đảo khác chỉ mất vài giờ. Ngoài ra, các dịch vụ này xác định không chính xác trang web lành tính mới là trang web lừa đảo do thiếu tuổi miền. Các kỹ thuật chẩn đoán dựa trên sự tương đồng về hình ảnh sẽ so sánh trang web mới với chữ ký được lưu trữ trước của trang web. Chữ ký trực quan của trang web bao gồm ảnh chụp màn hình, kiểu phông chữ, hình ảnh, bố cục trang, biểu tượng, v.v. Do đó, những kỹ thuật này không thể xác định các trang web lừa đảo mới và tạo ra tỷ lệ âm tính giả cao (lừa đảo lành tính). Kỹ thuật dựa trên URL không xem xét HTML của trang web và có thể đánh giá sai một số trang web độc hại được lưu trữ trên các máy chủ miễn phí hoặc bị xâm nhập. Nhiều cách tiếp cận hiện có11-13 trích xuất các đặc điểm dựa trên URL được tạo thủ công, ví dụ: số dấu chấm, sự hiện diện của ký hiệu “@”, “#”, “-” đặc biệt, độ dài URL, tên thương hiệu trong URL, vị trí của tên miền Cấp cao nhất, kiểm tra tên máy chủ cho địa chỉ IP, sự hiện diện của nhiều TLD, v.v. Tuy nhiên, vẫn còn những trở ngại trong việc trích xuất các tính năng URL thủ công do nỗ lực của con người đòi hỏi thời gian và chi phí nhân công bảo trì bổ sung. Việc phát hiện và ngăn chặn hành vi phạm tội lừa đảo là một trở ngại lớn đối với các nhà nghiên cứu vì kẻ lừa đảo thực hiện những hành vi phạm tội này theo cách có thể tránh được các phương pháp chống lừa đảo hiện tại. Do đó, người quản lý an ninh mạng rất khuyến khích việc sử dụng các phương pháp kết hợp thay vì một phương pháp duy nhất.

Bài viết này cung cấp một giải pháp hiệu quả để phát hiện lừa đảo bằng cách trích xuất các tính năng từ mã nguồn URL và HTML của trang web. Cụ thể, chúng tôi đã đề xuất một bộ tính năng kết hợp bao gồm các tính năng chuỗi ký tự URL mà không có kiến thức chuyên môn, các thông tin siêu liên kết khác nhau, văn bản gốc và các tính năng dựa trên dữ liệu HTML ồn ào trong mã nguồn HTML. Sau đó, các tính năng này được sử dụng để tạo vectơ đặc trưng cần thiết cho việc huấn luyện phương pháp được đề xuất bởi phân loại XGBoost. Các thử nghiệm mở rộng cho thấy phương pháp chống lừa đảo được đề xuất đã đạt được hiệu suất cạnh tranh trên tập dữ liệu thực về mặt thống kê đánh giá khác nhau.

Phương pháp chống lừa đảo của chúng tôi được thiết kế để đáp ứng các yêu cầu sau.

- Hiệu quả phát hiện cao: Để mang lại hiệu quả phát hiện cao, việc phân loại không chính xác các trang web lành tính là lừa đảo (dương tính giả) phải ở mức tối thiểu và phân loại chính xác các trang web lừa đảo (dương tính thật) phải cao.

- Phát hiện theo thời gian thực: Phải cung cấp dự đoán về phương pháp phát hiện lừa đảo trước khi phát hiện thông tin cá nhân của người dùng trên trang web lừa đảo.
- Độc lập với mục tiêu: Do các tính năng được trích xuất từ cả URL và HTML nên cách tiếp cận được đề xuất có thể phát hiện các trang web lừa đảo mới nhắm mục tiêu vào bất kỳ trang web lành tính nào (tấn công zero-day).
- Độc lập với bên thứ ba: Bộ tính năng được xác định trong công việc của chúng tôi rất nhẹ và có thể thích ứng phía khách hàng, không phụ thuộc vào các dịch vụ của bên thứ ba như danh sách đen/danh sách trắng, bản ghi Hệ thống tên miền (DNS), bản ghi WHOIS (tuổi tên miền) , lập chỉ mục công cụ tìm kiếm, các biện pháp lưu lượng truy cập mạng, v.v. Các dịch vụ khó khăn của bên thứ ba có thể nâng cao hiệu quả của phương pháp phát hiện, chúng có thể phân loại sai các trang web lành tính nếu một trang web lành tính mới được đăng ký. Hơn nữa, cơ sở dữ liệu DNS và bản ghi tuổi tên miền có thể bị nhiễm độc và dẫn đến kết quả âm tính giả (lừa đảo vô hại).

  Do đó, cần có một kỹ thuật gọn nhẹ để phát hiện các trang web lừa đảo có thể thích ứng ở phía khách hàng. Những đóng góp chính trong bài viết này được chia thành từng khoản như sau.
- Chúng tôi đề xuất phương pháp phát hiện lừa đảo, phương pháp này trích xuất các tính năng hiệu quả từ URL và HTML của trang web nhất định mà không cần dựa vào dịch vụ của bên thứ ba. Vì vậy, nó có thể thích ứng ở phía máy khách và chỉ định quyền riêng tư tốt hơn.
- Chúng tôi đề xuất tám tính năng mới bao gồm các tính năng chuỗi ký tự URL (F1), cấp độ ký tự nội dung văn bản (F2), các tính năng siêu liên kết khác nhau (F3, F4, F5, F6, F7 và F14) cùng với bảy tính năng hiện có được áp dụng từ tài liệu.

- Chúng tôi đã tiến hành nhiều thử nghiệm rộng rãi bằng cách sử dụng nhiều thuật toán học máy khác nhau để đo lường hiệu quả của các tính năng được đề xuất. Kết quả đánh giá cho thấy phương pháp đề xuất xác định chính xác các trang web hợp pháp vì nó có tỷ lệ âm tính thực cao và tỷ lệ dương tính giả rất ít.
- Chúng tôi phát hành bộ dữ liệu phát hiện trang web lừa đảo thực sự để các nhà nghiên cứu khác sử dụng về chủ đề này.

Phần còn lại của bài viết này được cấu trúc như sau: Phần "Công việc liên quan" trước tiên sẽ xem xét các công việc liên quan về phát hiện lừa đảo. Mười phần "Phương pháp tiếp cận được đề xuất" trình bày tổng quan về giải pháp được đề xuất của chúng tôi và mô tả các tính năng được đề xuất được thiết lập để huấn luyện các thuật toán học máy. Phần "Thí nghiệm và phân tích kết quả" giới thiệu các thử nghiệm mở rộng bao gồm tập dữ liệu thử nghiệm và đánh giá kết quả.
Hơn nữa, phần "Thảo luận và hạn chế" chứa nội dung thảo luận và những hạn chế của phương pháp được đề xuất. Cuối cùng, phần “Kết luận” kết luận bài viết và thảo luận về công việc trong tương lai.

Công việc có liên quan

Phần này cung cấp cái nhìn tổng quan về các kỹ thuật phát hiện lừa đảo được đề xuất trong tài liệu. Các phương thức lừa đảo được chia thành hai loại; mở rộng nhận thức của người dùng để phân biệt các đặc điểm của trang web lừa đảo và trang web lành tính14, đồng thời sử dụng một số phần mềm bổ sung. Các kỹ thuật dựa trên phần mềm còn được phân loại thành phát hiện dựa trên danh sách và phát hiện dựa trên máy học. Tuy nhiên, vấn đề lừa đảo rất phức tạp

Machine Translated by Google
www.nature.com/scientificreports/

rằng không có giải pháp dứt khoát nào có thể vượt qua mọi mối đe dọa một cách hiệu quả; do đó, nhiều kỹ thuật thường được dành riêng để hạn chế các hành vi lừa đảo cụ thể.

Phát hiện dựa trên danh sách. Các phương pháp phát hiện lừa đảo dựa trên danh sách sử dụng kỹ thuật dựa trên danh sách trắng hoặc danh sách đen. Danh sách đen chứa danh sách các miền, URL và địa chỉ IP đáng ngờ, được sử dụng để xác thực xem URL có gian lận hay không. Đồng thời, danh sách trắng là danh sách các miền, URL và địa chỉ IP hợp pháp được sử dụng để xác thực một URL bị nghi ngờ. Wang và cộng sự15, Jain và Gupta5 và Han và cộng sự16 sử dụng phương pháp dựa trên danh sách trắng để phát hiện URL đáng ngờ. Các phương pháp dựa trên danh sách đen được sử dụng rộng rãi trong các thanh công cụ chống lừa đảo có sẵn công khai, chẳng hạn như duyệt web an toàn của Google, thanh công cụ này duy trì danh sách đen các URL và đưa ra cảnh báo cho người dùng khi một URL bị coi là lừa đảo. Prakash và cộng sự17 đã đề xuất một kỹ thuật dự đoán các URL lừa đảo có tên là Phishnet. Trong kỹ thuật này, các URL lừa đảo được xác định từ các URL nằm trong danh sách cấm hiện có bằng cách sử dụng cấu trúc thư mục, địa chỉ IP tương đương và tên thương hiệu. Felegyhazi và cộng sự18 đã phát triển một phương pháp so sánh thông tin tên miền và máy chủ định danh của các URL đáng ngờ mới với thông tin của các URL nằm trong danh sách đen cho quá trình phân loại. Sheng và cộng sự19 đã chứng minh rằng một miền giả mạo đã được thêm vào danh sách đen sau một khoảng thời gian đáng kể và khoảng 50–80% miền giả mạo đã được thêm vào sau khi cuộc tấn công được thực hiện. Vì có hàng nghìn trang web lừa đảo được tung ra mỗi ngày nên danh sách đen yêu cầu phải được cập nhật định kỳ từ nguồn của nó. Vì vậy, các kỹ thuật phát hiện dựa trên máy học sẽ hiệu quả hơn trong việc xử lý các hành vi lừa đảo.

Phát hiện dựa trên học máy. Kỹ thuật khai thác dữ liệu đã mang lại hiệu suất vượt trội trong nhiều ứng dụng, ví dụ: bảo mật dữ liệu và quyền riêng tư20, lý thuyết trò chơi21, hệ thống chuỗi khối22, chăm sóc sức khỏe23, v.v. Do sự phát triển gần đây của các phương pháp phát hiện lừa đảo, nhiều kỹ thuật dựa trên máy học khác nhau cũng đã được sử dụng6,9 ,10,13 để điều tra tính hợp pháp của các trang web. Hiệu quả của các phương pháp này phụ thuộc vào việc thu thập tính năng, dữ liệu huấn luyện và thuật toán phân loại. Bộ sưu tập tính năng được trích xuất từ các nguồn khác nhau, ví dụ: URL, nội dung trang web, dịch vụ của bên thứ ba, v.v. Tuy nhiên, một số tính năng heuristic khó truy cập và tốn thời gian, khiến một số phương pháp học máy yêu cầu tính toán cao để trích xuất những tính năng này. đặc trưng.

Jain và Gupta24 đã đề xuất một phương pháp chống lừa đảo trích xuất các tính năng từ URL và mã nguồn của trang web và không dựa vào bất kỳ dịch vụ nào của bên thứ ba. Mặc dù phương pháp được đề xuất đạt được độ chính xác cao trong việc phát hiện các trang web lừa đảo nhưng nó sử dụng bộ dữ liệu hạn chế (2141 trang web lừa đảo và 1918 trang web hợp pháp).
Các tác giả này9 cũng trình bày một phương pháp phát hiện lừa đảo có thể xác định các cuộc tấn công lừa đảo bằng cách phân tích các siêu liên kết được trích xuất từ HTML của trang web. Phương pháp được đề xuất là giải pháp phía máy khách và không phụ thuộc vào ngôn ngữ. Tuy nhiên, nó hoàn toàn phụ thuộc vào HTML của trang web và có thể phân loại không chính xác các trang web lừa đảo nếu kẻ tấn công thay đổi tất cả các tham chiếu tài nguyên trang web (ví dụ: Javascript, CSS, hình ảnh, v.v.). Rao và Pais25
đã đề xuất một kỹ thuật chống lừa đảo hai cấp độ gọi là BlackPhish. Ở cấp độ đầu tiên, danh sách đen chữ ký được tạo bằng cách sử dụng các tính năng dựa trên sự tương đồng về mặt hình ảnh (ví dụ: tên tập tin, đường dẫn và ảnh chụp màn hình) thay vì sử dụng danh sách URL đen. Ở cấp độ thứ hai, các tính năng phỏng đoán được trích xuất từ URL và HTML để xác định các trang web lừa đảo ghi đè bộ lọc cấp độ đầu tiên. Mặc dù vậy, các trang web hợp pháp luôn trải qua quá trình rung chuyển hai cấp độ. Trong một số nghiên cứu26 tác giả đã sử dụng cơ chế dựa trên công cụ tìm kiếm để xác thực trang web dưới dạng xác thực cấp độ đầu tiên. Trong xác thực cấp độ thứ hai, nhiều siêu liên kết khác nhau trong HTML của trang web được xử lý để phát hiện trang web lừa đảo. Mặc dù việc sử dụng các kỹ thuật dựa trên công cụ tìm kiếm làm tăng số lượng trang web hợp pháp được xác định chính xác là hợp pháp nhưng nó cũng làm tăng số lượng trang web hợp pháp được xác định không chính xác là lừa đảo khi các trang web xác thực mới tạo không được tìm thấy trong kết quả hàng đầu của công cụ tìm kiếm.

Các phương pháp tiếp cận dựa trên tìm kiếm giả định rằng trang web chính hãng xuất hiện trong kết quả tìm kiếm hàng đầu.

Trong một nghiên cứu gần đây, Rao và cộng sự27 đã đề xuất một phương pháp phát hiện trang web lừa đảo mới bằng cách nhúng từ được trích xuất từ văn bản thuần túy và văn bản cụ thể tên miền của mã nguồn html. Họ đã triển khai cách nhúng từ khác nhau để đánh giá mô hình của họ bằng cách sử dụng các kỹ thuật tổng hợp và đa phương thức. Tuy nhiên, phương pháp được đề xuất hoàn toàn phụ thuộc vào văn bản thuần túy và văn bản miền cụ thể và có thể thất bại khi văn bản được thay thế bằng hình ảnh.
Một số nhà nghiên cứu đã cố gắng xác định các cuộc tấn công lừa đảo bằng cách trích xuất các mối quan hệ siêu liên kết khác nhau từ các trang web. Guo và cộng sự28 đã đề xuất một phương pháp phát hiện trang web lừa đảo mà họ gọi là HinPhish. Cách tiếp cận này thiết lập một mạng thông tin không đồng nhất (HIN) dựa trên các nút miền và các nút tài tài nguyên và thiết lập ba mối quan hệ giữa bốn siêu liên kết: liên kết ngoài, liên kết trống, liên kết nội bộ và liên kết tương đối. Mười, họ đã áp dụng thuật toán xếp hạng thẩm quyền để tính toán tác động của các mối quan hệ khác nhau và đạt được điểm định lượng cho mỗi nút.

Trong công việc của Sahingoz và cộng sự6 , cách biểu diễn phân tán của các từ được sử dụng trong một URL cụ thể và sau đó bảy phân loại học máy khác nhau được sử dụng để xác định xem một URL đáng ngờ có phải là trang web lừa đảo hay không. Rao và cộng sự13 đã đề xuất một kỹ thuật chống lừa đảo có tên là CatchPhish. Họ đã trích xuất các tính năng được tạo thủ công và Tần suất tài liệu nghịch đảo tần số thuật ngữ (TF-IDF) từ các URL, sau đó huấn luyện một trình phân loại về các tính năng này bằng thuật toán rừng ngẫu nhiên. Mặc dù các phương pháp trên đã cho thấy hiệu suất khả quan nhưng chúng gặp phải những hạn chế sau: (1) không có khả năng xử lý các ký tự không được quan sát vì các URL thường chứa các từ vô nghĩa và không xác định không có trong tập huấn luyện; (2) họ không xem xét nội dung của trang web. Theo đó, một số URL có tính phân biệt với các URL khác nhưng bắt chước các trang web hợp pháp có thể không được xác định dựa trên chuỗi URL. Vì công việc của họ chỉ dựa trên các tính năng URL nên không đủ để phát hiện các trang web lừa đảo. Tuy nhiên, chúng tôi đã cung cấp một giải pháp hiệu quả bằng cách đề xuất cách tiếp cận của chúng tôi đối với miền này bằng cách sử dụng ba loại tính năng khác nhau để phát hiện trang web lừa đảo hiệu quả hơn. Cụ thể, chúng tôi đã đề xuất một bộ tính năng kết hợp bao gồm chuỗi ký tự URL, các thông tin siêu liên kết khác nhau và các tính năng dựa trên nội dung văn bản.

Các phương pháp học sâu đã được sử dụng để phát hiện lừa đảo, ví dụ: Mạng thần kinh chuyển đổi (CNN), Mạng thần kinh sâu (DNN), Mạng thần kinh tái phát (RNN) và Mạng thần kinh chuyển đổi tái phát

Machine Translated by Google
www.nature.com/scientificreports/

(RCNN) do sự thành công của Xử lý ngôn ngữ tự nhiên (NLP) đạt được bằng các kỹ thuật này. Tuy nhiên, phương pháp học sâu không được sử dụng nhiều trong việc phát hiện lừa đảo do thời gian đào tạo toàn diện. Aljofey và cộng sự 3 đã đề xuất một phương pháp phát hiện lừa đảo bằng mạng nơ ron tích chập ở cấp độ ký tự dựa trên URL.
Phương pháp đề xuất được so sánh bằng cách sử dụng nhiều thuật toán máy và học sâu khác nhau cũng như các loại tính năng khác nhau như ký tự TF-IDF, vectơ đếm và các tính năng được tạo thủ công. Le và cộng sự29 đã cung cấp phương pháp URLNet để phát hiện trang web lừa đảo từ URL. Họ trích xuất các đặc điểm cấp độ ký tự và cấp độ từ từ chuỗi URL và sử dụng mạng CNN để đào tạo và thử nghiệm. Chatterjee và Namin30 đã giới thiệu kỹ thuật phát hiện lừa đảo dựa trên học tăng cường sâu để xác định các URL lừa đảo. Tey đã sử dụng mô hình của họ trên một tập dữ liệu cân bằng, được gắn nhãn gồm các URL lành tính và lừa đảo, trích xuất 14 tính năng được tạo thủ công từ các URL nhất định để đào tạo mô hình được đề xuất. Trong các nghiên cứu gần đây, Xiao và cộng sự31 đã đề xuất phương pháp phát hiện trang web lừa đảo có tên CNN-MHSA. Mạng CNN được ứng dụng để trích xuất các đặc điểm ký tự từ URL. Trong khi đó, cơ chế tự chú ý nhiều đầu (MHSA) được sử dụng để tính toán các trọng số tương ứng cho các tính năng đã học của CNN. Zheng và cộng sự32 đã đề xuất Mạng lưới thần kinh Kim tự tháp sâu xa lộ (HDP-CNN) mới, là một mạng tích chập sâu tích hợp cả biểu diễn nhúng ở cấp độ ký tự và cấp độ từ để xác định xem một URL nhất định là lừa đảo hay hợp pháp. Mặc dù các phương pháp trên đã cho thấy hiệu quả có giá trị nhưng chúng có thể phân loại sai các trang web lừa đảo được lưu trữ trên máy chủ bị xâm nhập do các tính năng chỉ được trích xuất từ URL của trang web.

Các tính năng được trích xuất trong một số nghiên cứu trước đây dựa trên công việc thủ công và cần nỗ lực nhiều hơn vì các tính năng này cần được đặt lại theo tập dữ liệu, điều này có thể dẫn đến việc sử dụng quá mức các giải pháp chống lừa đảo. Chúng tôi lấy động lực từ các nghiên cứu nêu trên và đề xuất cách tiếp cận của mình. Trong đó, tác phẩm hiện tại có tính năng trích xuất chuỗi ký tự từ URL mà không cần can thiệp thủ công. Hơn nữa, cách tiếp cận của chúng tôi sử dụng dữ liệu ồn ào về thông tin HTML, văn bản gốc và siêu liên kết của trang web nhằm mục đích xác định các trang web lừa đảo mới. Bảng 1 trình bày so sánh chi tiết các phương pháp phát hiện lừa đảo dựa trên học máy hiện có.

Phương pháp đề xuất. Phương pháp tiếp cận của chúng tôi trích xuất và phân tích các đặc điểm khác nhau của các trang web bị nghi ngờ để xác định hiệu quả các hành vi lừa đảo quy mô lớn. Đóng góp chính của bài viết này là việc sử dụng kết hợp các bộ tính năng này. Để cải thiện độ chính xác khi phát hiện các trang web lừa đảo, chúng tôi đã đề xuất tám tính năng mới. Các tính năng được đề xuất của chúng tôi xác định mối quan hệ giữa URL của trang web và trang web nội dung.

Kiến Trúc Hệ Thống. Kiến trúc tổng thể của phương pháp đề xuất được chia thành ba giai đoạn. Trong giai đoạn đầu tiên, tất cả các tính năng cần thiết sẽ được trích xuất và mã nguồn HTML sẽ được thu thập thông tin. Giai đoạn thứ hai áp dụng vectơ đặc trưng để tạo vectơ đặc trưng cụ thể cho từng trang web. Giai đoạn thứ ba xác định xem trang web nhất định có phải là lừa đảo hay không. Hình 1 thể hiện cấu trúc hệ thống của phương pháp đề xuất. Chi tiết của từng giai đoạn được mô tả như sau.

Tạo tính năng. Các tính năng Te được tạo ra trong thành phần này. Các tính năng của chúng tôi dựa trên URL và mã nguồn HTML của trang web. Cây Mô hình đối tượng tài liệu (DOM) của trang web được sử dụng để tự động trích xuất các tính năng siêu liên kết và nội dung văn bản bằng trình thu thập dữ liệu web. Các đặc điểm trong cách tiếp cận của chúng tôi được phân loại thành bốn nhóm như được mô tả trong Bảng 2. Cụ thể, các đặc điểm F1–F7 và F14 là mới và do chúng tôi đề xuất; Các tính năng F8–F13 và F15 được lấy từ các phương pháp khác9,11,12,24,33 nhưng chúng tôi đã điều chỉnh chúng để có kết quả tốt hơn. Hơn nữa, phương pháp và chiến lược quan sát liên quan đến việc giải thích các đặc điểm này được áp dụng khác nhau trong cách tiếp cận của chúng tôi. Giải thích chi tiết về các tính năng được đề xuất được cung cấp trong phần trích xuất tính năng của bài viết này.

Vector hóa đặc tính. Sau khi các đặc điểm được trích xuất, chúng tôi áp dụng vectơ hóa đặc điểm để tạo ra một vectơ đặc trưng cụ thể cho mỗi trang web nhằm tạo tập dữ liệu được gắn nhãn. Chúng tôi tích hợp các tính năng chuỗi ký tự URL với các tính năng TF-IDF nội dung văn bản và các tính năng thông tin siêu liên kết để tạo ra vectơ đặc trưng cần thiết cho việc huấn luyện phương pháp đề xuất. Siêu liên kết có tính năng kết hợp đầu ra vectơ đặc tính 13 chiều dưới dạng FH = f3, f4, f5, ... , f15 và chuỗi ký tự URL có tính năng kết hợp đầu ra vectơ đặc tính 200 chiều dưới dạng FU = c1,c2,c3, .. . ,c200, chúng tôi đặt độ dài URL cố định thành 200. Nếu độ dài URL lớn hơn 200, phần bổ sung sẽ bị bỏ qua. Nếu không, chúng tôi đặt số 0 vào phần còn lại của chuỗi URL. Cài đặt của giá trị này phụ thuộc vào việc phân phối độ dài URL trong tập dữ liệu của chúng tôi. Chúng tôi nhận thấy rằng hầu hết độ dài URL đều nhỏ hơn 200, điều đó có nghĩa là khi vectơ dài, nó có thể chứa thông tin vô ích, ngược lại khi vectơ đặc trưng quá ngắn, nó có thể chứa không đủ tính năng. Tổ hợp cấp độ ký tự TF-IDF tạo ra vectơ đặc trưng D chiều dưới dạng FT = t1, t2, t3, ... , tD trong đó D là kích thước của từ điển được tính toán từ kho nội dung văn bản. Từ phân tích thực nghiệm cho thấy kích thước của từ điển D=20.332 và kích thước tăng lên khi số lượng kho văn bản tăng lên. Te ở trên ba vectơ đặc trưng được kết hợp để tạo ra vectơ đặc trưng fnal FV = FT FU FH = t1, t2, ... , tD,c1,c2 ... ,c200, f3, f4, f5, ... , f15 được cung cấp làm đầu vào cho các thuật toán học máy để phân loại trang web.

Mô-đun phát hiện. Giai đoạn Phát hiện Te bao gồm việc xây dựng một bộ phân loại mạnh bằng cách sử dụng phương pháp tăng cường, bộ phân loại XGBoost. Boosting tích hợp nhiều bộ phân loại yếu và tương đối chính xác để xây dựng một bộ phân loại mạnh và trước đó mạnh mẽ để phát hiện các hành vi lừa đảo. Việc tăng cường cũng giúp kết hợp các tính năng đa dạng giúp cải thiện hiệu suất phân loại34. Ở đây, trình phân loại XGBoost được sử dụng trên các bộ tính năng tích hợp của chuỗi ký tự URL FU, thông tin siêu liên kết khác nhau FH, tính năng biểu mẫu đăng nhập FL và các tính năng dựa trên nội dung văn bản FT để xây dựng trình phân loại mạnh mẽ nhằm phát hiện lừa đảo. Trong giai đoạn huấn luyện, phân loại XGBoost được huấn luyện

Machine Translated by Google
www.nature.com/scientificreports/

Tiếp cận	Sự miêu tả	tập dữ liệu	Hạn chế
Jain và Gupta24	Cách tiếp cận này lọc các trang web lừa đảo ở phía khách hàng dựa trên các tính năng URL thủ công, tính năng siêu liên kết và tính năng từ khóa nhận dạng bằng cách sử dụng Rừng ngẫu nhiên	Tập dữ liệu riêng gồm 2141 trang web lừa đảo và 1918 trang web lành tính	Nó trích xuất các tính năng URL được thiết kế thủ công, cần nỗ lực của con người Các đặc điểm nhận dạng phụ thuộc vào ngôn ngữ trong đó các từ khóa hàng đầu được trích xuất từ trang web
Jain và Gupta9	Đề xuất phươ ng pháp chống lừa đảo bằng logistic dựa trên các trang web siêu liên kết khác nhau các kết quả được trích xuất từ nội dung HTML của trang web	Một tập dữ liệu riêng về 1428 lừa đảo và 1116 hồi quy lành tính,	Tập dữ liệu hạn chế Bộ tính năng hoàn toàn phụ thuộc vào nội dung trang web bị lỗi khi nội dung được thay thế bằng Hình ảnh
Rao và Pais25	Các tác giả đã phát triển một kỹ thuật tán tính hai cấp độ để phát hiện các trang web lừa đảo bằng cách sử dụng danh sách đen nâng cao và phươ ng pháp phỏng đoán đặc trưng	Một tập dữ liệu công khai gồm 5438 trang web lành tính và 4097 trang web lừa đảo	Các site lành tính luôn trải qua hai cấp độ rung chuyển
Jain và Gupta26	Một cách tiếp cận để phân loại các trang web dựa trên xác thực hai cấp độ: thông tin công cụ tìm kiếm và siêu liên kết	Một tập dữ liệu riêng gồm 2000 trang web lành tính và 2000 trang web lừa đảo	Thất bại ở cấp độ đầu tiên khi các trang web lành tính mới được xây dựng không xuất hiện trong kết quả tìm kiếm hàng đầu Không thành công khi nội dung của trang web được thay thế bằng hình ảnh
Sahingoz và cộng sự6	Sử dụng các tính năng dựa trên NLP, vectơ từ và tính năng kết hợp, sau đó bảy thuật toán học máy khác nhau được sử dụng để phân loại URL	Tập dữ liệu công khai gồm 36.400 URL lành tính và 37.175 URL lừa đảo	Không thể xử lý các ký tự không nhìn thấy được trong URL Phươ ng pháp Te có thể không phát hiện được các URL ngắn hơ n
Rao và cộng sự13	Kỹ thuật này đề xuất các tính năng URL được tạo thủ công và các tính năng dựa trên TF-IDF và việc sử dụng các tính năng này sẽ phân loại các URL bằng cách sử dụng bộ phân loại nhóm ngẫu nhiên	Tập dữ liệu công khai gồm 85.409 URL lành tính và 40.668 URL lừa đảo	Trích xuất các tính năng URL được tạo thủ công, cần nỗ lực của con người và chi phí lao động bảo trì bổ sung Mô hình có thể bị lỗi khi các trang web lừa đảo được lưu trữ trên các máy chủ lưu trữ miễn phí hoặc bị xâm nhập
Aljofey và cộng sự.3	Một mô hình học sâu nhanh dựa trên URL sử dụng CNN cấp ký tự được đề xuất để phát hiện lừa đảo	Tập dữ liệu riêng gồm 157.626 URL lành tính và 161.016 URL lừa đảo	Nó hoàn toàn phụ thuộc vào URL của trang web Nó không quan tâm liệu URL của trang web có còn tồn tại hay không hoặc có lỗi hay không
Lê và cộng sự29	Kỹ thuật này áp dụng mạng CNN cho cả ký tự và từ của chuỗi URL để phát hiện URL độc hại	Tập dữ liệu riêng gồm 4.683.425 URL lành tính và 9.366.850 URL độc hại	Do mô hình deep learning được triển khai bằng cả nhúng cấp độ từ và cấp độ ký tự nên cần có đủ bộ nhớ
Xiao và cộng sự31	Đề xuất một kỹ thuật có tên CNN-MHSA, kết hợp mạng nơ ron tích chập (CNN) và cơ chế tự chú ý nhiều đầu (MHSA) với nhau để tìm hiểu các tính năng trong URL và phát hiện lừa đảo	Một tập dữ liệu riêng tư trong đó 45.000 là lành tính và 43.984 là lừa đảo	Tham số độ dài URL có thể ảnh hưởng đến tính mạnh mẽ của mô hình
Zheng và cộng sự32	Đề xuất Mạng thần kinh phân cấp đường cao tốc (HDP-CNN) mới để phát hiện các URL lừa đảo. Phươ ng pháp này sử dụng nhúng cấp độ từ cùng với nhúng cấp độ ký tự để thể hiện hiệu suất tốt hơ n	Một tập dữ liệu riêng tư chứa 344.794 URL lành tính và 71.556 URL lừa đảo	Vấn đề mất cân bằng dữ liệu nghiêm trọng có thể khiến mô hình bị tràn trên các tập dữ liệu lớn
Rao và cộng sự27	Kỹ thuật nghiêng máy sử dụng thuật toán nhúng từ để tạo vectơ đặc trưng bằng văn bản thuần túy và văn bản miền được trích xuất từ nội dung trang web	Một tập dữ liệu công khai bao gồm 5438 trang web lừa đảo và 5076 trang web lành tính có URL của chúng	Kỹ thuật Te phụ thuộc vào ngôn ngữ Nó không thành công khi nội dung của trang web được thay thế bằng một hình ảnh
Guo và cộng sự28	Phươ ng pháp phát hiện lừa đảo tạo ra các mạng thông tin không đồng nhất dựa trên các nút miền, nút tài nguyên trang và mối quan hệ giữa các siêu liên kết	Một tập dữ liệu công khai chứa 29.496 mẫu lừa đảo và 30.649 mẫu lành tính	Cách tiếp cận này có thể cho thấy hiệu suất kém khi trang web chứa một số siêu liên kết
Phươ ng pháp đề xuất	Phươ ng pháp học máy bao gồm một bộ tính năng kết hợp bao gồm chuỗi ký tự URL, các tính năng siêu liên kết khác nhau và các tính năng cấp độ ký tự TF-IDF từ phần văn bản gốc và phần nhiều của HTML của trang web nhất định	Một tập dữ liệu công khai bao gồm 27.280 lừa đảo URL có mã HTML và 32.972 trang lành tính	Tính năng dựa trên văn bản đơ n giản của trang web là dựa trên ngôn ngữ Cần truy cập mã nguồn HTML của trang web

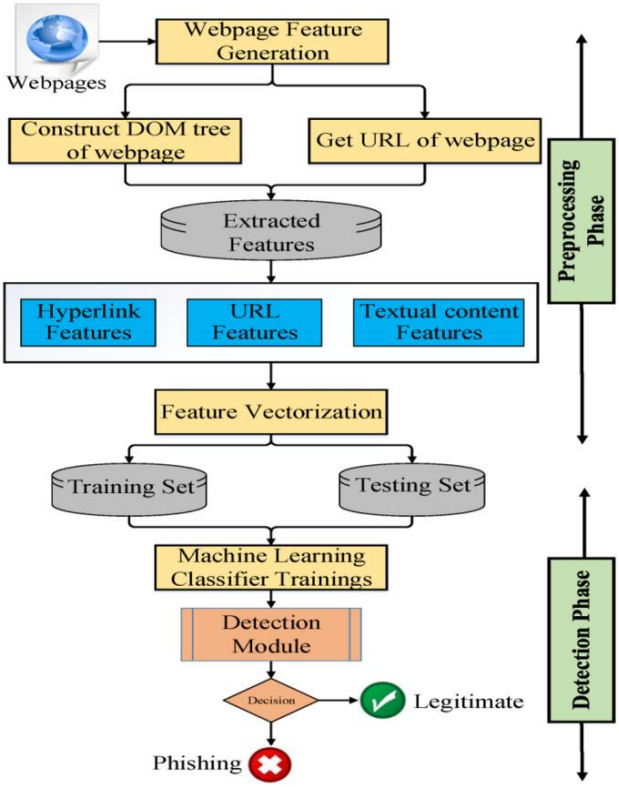
Bảng 1. So sánh các phươ ng pháp phát hiện lừa đảo dựa trên học máy.

sử dụng vectơ đặc trưng (FU FH FL FT) được thu thập từ mỗi bản ghi trong tập dữ liệu huấn luyện. Ở giai đoạn thử nghiệm, trình phân loại sẽ phát hiện xem một trang web cụ thể có phải là trang web độc hại hay không. Mô tả chi tiết được hiển thị trong Hình 2.

Trích xuất tính năng. Do công cụ tìm kiếm hạn chế và các phươ ng pháp của bên thứ ba được thảo luận trong tài liệu, chúng tôi trích xuất các tính năng cụ thể từ phía khách hàng trong cách tiếp cận của mình. Chúng tôi đã giới thiệu 11 tính năng siêu liên kết (F3–F13), hai tính năng biểu mẫu đăng nhập (F14 và F15), tính năng TF-IDF cấp ký tự (F2) và tính năng chuỗi ký tự URL (F1). Tất cả các tính năng này sẽ được thảo luận trong các phần phụ sau.

Tính năng chuỗi ký tự URL (F1). URL Te là viết tắt của Bộ định vị tài nguyên thống nhất. Nó được sử dụng để cung cấp vị trí của các tài nguyên trên web như hình ảnh, tập tin, siêu văn bản, video, v.v. URL. Mỗi URL bắt đầu bằng một giao thức (http, https và fp) được sử dụng để truy cập tài nguyên được yêu cầu. Trong phần này, chúng tôi trích xuất các đặc điểm chuỗi ký tự từ URL. Chúng tôi sử dụng phươ ng pháp được sử dụng trong35 để xử lý URL ở cấp độ ký tự. Thông tin thêm được chứa ở cấp độ ký tự. Những kẻ lừa đảo cũng bắt chước URL của các trang web hợp pháp bằng cách thay đổi nhiều ký tự không thể xác định được, ví dụ: “www.icbc.com” là “www.1cbc.com”. Xử lý URL cấp ký tự là một giải pháp cho vấn đề hết từ vựng. Chuỗi cấp độ ký tự xác định thông tin quan trọng từ các nhóm ký tự cụ thể xuất hiện cùng nhau. Đây có thể là dấu hiệu của lừa đảo. Nói chung, URL là một chuỗ

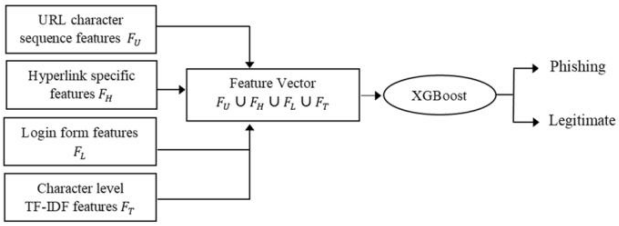
Machine Translated by Google
www.nature.com/scientificreports/



Hình 1. Cấu trúc chung của phương pháp đề xuất.

Loại	KHÔNG	Tên
Các tính năng dựa trên URL	F1	Vectơ chuỗi ký tự
Tính năng nội dung văn bản	F2	Ký tự N-gram vectơ TF-IDF
Thông tin siêu liên kết	F3, F4, F5, F6, F7, F8, F9, F10, F11, F12 và F13	Script_fles, CSS_fles, img_fles, a_fles, a_Null_hyperlinks, Null_hyperlinks, Total_hyperlinks, Internal_hyperlinks, external_hyperlinks, external_hyperlinks/Internal_hyperlinks và Error_hyperlinks
Thông tin form đăng nhập	F14 và F15	Total_forms và Suspicious_form

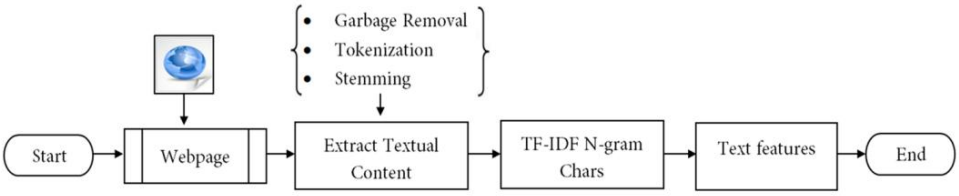
Bảng 2. Các đặc điểm được sử dụng trong phương pháp đề xuất.



Hình 2. Thuật toán phát hiện lừa đảo.

hành động hoặc từ mà một số từ có ít ý nghĩa ngữ nghĩa. Chuỗi ký tự giúp tìm ra thông tin nhạy cảm này và cải thiện hiệu quả phát hiện URL lừa đảo. Trong quá trình thực hiện nhiệm vụ học, các kỹ thuật học máy có thể được áp dụng trực tiếp bằng cách sử dụng các đặc điểm chuỗi ký tự được trích xuất mà không cần sự can thiệp của chuyên gia. Các quy trình chính của việc tạo chuỗi ký tự bao gồm: chuẩn bị từ vựng ký tự, tạo đối tượng mã thông báo bằng gói tiền xử lý Keras (https://Keras.io) để xử lý URL ở cấp char và thêm mã thông báo "UNK" vào từ vựng ở mức tối đa giá trị của từ điển ký tự, chuyển đổi văn bản của URL thành chuỗi

Machine Translated by Google
www.nature.com/scientificreports/



Hình 3. Quá trình tạo đặc điểm văn bản.

mã thông báo và đệm chuỗi URL để đảm bảo các vectơ có độ dài bằng nhau. Mô tả về trích xuất các tính năng URL được hiển thị trong Thuật toán 1.

Algorithm 1: Extract the URL based features

Input: URL of suspicious website U

Output: Character sequences vector $F_U = < c_1, c_2, c_3, \dots, c_{200} > \in F1$

Start

1. Initialize Tokenizer (char_level=True, oov_token='UNK'), Alphabet="abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUV WXYZ0123456789,;.!?:'\"/\\|_@#$%^&*~`+-=<>()[]{}" char_dict = { }
2. **for** *i, char* in enumerate(*Alphabet*) **do**
3. char_dict [char] = i + 1
4. Tokenizer.word_index = char_dict
5. **end for**
   Add 'UNK' to the vocabulary
6. Tokenizer.word_index [oov_token]= max_value_of_char_dict + 1
7. U_character_sequences = Tokenizer.texts_to_sequences (U)
8. **If** len (U_character_sequences) < 200 **then**
   the remaining part will be filled as 0
9. $F_U = < c_1, c_2, c_3, \dots, 0, 0, 0, c_{200} >$
10. **else**
    the part longer will be truncated
11. $F_U = < c_1, c_2, c_3, \dots, c_{200} >$
12. **end if**
13. Return F_U

End

Các tính năng của HTML. Mã nguồn trang web là chương trình đằng sau bất kỳ trang web hoặc phần mềm nào. Trong trường hợp trang web, bất kỳ ai cũng có thể xem mã này bằng nhiều công cụ khác nhau, ngay cả trong chính trình duyệt web. Trong phần này, chúng tôi trích xuất các tính năng văn bản và siêu liên kết hiện có trong mã nguồn HTML của trang web.

Các tính năng dựa trên nội dung văn bản (F2). TF-IDF là viết tắt của Tần số tài liệu nghịch đảo tần số. Trọng số TF IDF là thước đo thống kê cho chúng ta biết tầm quan trọng của một thuật ngữ trong kho tài liệu36. Các vectơ TF-IDF có thể được tạo ở nhiều cấp độ khác nhau của mã thông báo đầu vào (từ, ký tự, n-gram) 37. Người ta quan sát thấy rằng kỹ thuật TF IDF đã được triển khai theo nhiều cách tiếp cận để bắt lừa đảo các trang web bằng cách kiểm tra URL 13, lấy thông tin gián tiếp các liên kết liên quan38, trang web mục tiêu11 và tính hợp lệ của trang web bị nghi ngờ 39. Mặc dù kỹ thuật TF-IDF trích xuất các từ khóa nổi bật từ nội dung văn bản của trang web nhưng nó vẫn có một số hạn chế. Một trong những hạn chế là kỹ thuật TF-IDF không thành công khi từ khóa được trích xuất là vô nghĩa, sai chính tả, bỏ qua hoặc thay thế bằng hình ảnh. Do dữ liệu văn bản gốc và dữ liệu nhiễu (tức là các giá trị thuộc tính cho thẻ div, h1, h2, body và form) được trích xuất theo cách tiếp cận của chúng tôi từ trang web nhất định bằng cách sử dụng trình phân tích cú pháp BeautifulSoup, nên kỹ thuật cấp ký tự TF-IDF được áp dụng với số lượng tính năng tối đa là 25.000. Để có được thông tin văn bản hợp lệ, các phần bổ sung (ví dụ: mã JavaScript, mã CSS, ký hiệu dấu chấm câu và số) của trang web sẽ bị xóa thông qua các biểu thức thông thường, bao gồm các gói Xử lý ngôn ngữ tự nhiên (http://www.nltk.org/nltk_data/) chẳng hạn như phân đoạn câu, mã thông báo từ, từ vựng văn bản và bắt nguồn từ như trong Hình 3.

Những kẻ lừa đảo thường bắt chước nội dung văn bản của trang web mục tiêu để lừa người dùng. Hơn nữa, những kẻ lừa đảo có thể nhầm lẫn hoặc ghi đè một số văn bản (ví dụ: tiêu đề, bản quyền, siêu dữ liệu, v.v.) và thẻ trong các trang web lừa đảo để bỏ qua việc tiết lộ thông tin nhận dạng thực tế của trang web. Tuy nhiên, các thuộc tính thẻ vẫn giữ nguyên để duy trì sự tương đồng về mặt hình ảnh giữa trang web lừa đảo và trang web được nhắm mục tiêu bằng cách sử dụng cùng một phong cách và chủ đề như trang web lành tính. Vì vậy, cần phải trích xuất các đặc điểm văn bản (văn bản thuần và phần nhiễu của HTML) của trang web. Cơ bản của

Machine Translated by Google
www.nature.com/scientificreports/

Bước này là trích xuất biểu diễn dạng vectơ của văn bản và nội dung trang web hiệu quả. Đối tượng TF-IDF được sử dụng để vector hóa văn bản của trang web. Quy trình chi tiết của thuật toán tạo vectơ văn bản như sau.

Algorithm 2: Extract the textual content features of a webpage

Input: A HTML document *doc*

Output: TF-IDF vector N-gram chars $F_T = < t_1, t_2, t_3, \dots, t_D > \in F2$

Start

1. P_T = getPlaintext (doc)
2. N_T =getTagAttributesValues (doc)
   (DIV, IMG, Body, Footer, a, Link, Article, Label, H1…H5, Template...etc.)
3. $T_1 = P_T \cup N_T$
4. Text cleaning and preprocessing
   - T_2=∅ ,T_3=∅
   - T_2 = Text_cleaner (T_1)
     (Remove punctuations symbols, numbers, spaces, newline, character that are not English)
   - **for** *token* in T_2 **do**
   - **if** *token* not in STOPWORDS and len (*token*) > 3 **then**
   - $T_3 = T_3 \cup$ lemmatize_stemming (*token*)
   - **end if**
   - **end for**
5. F_T =TF-IDF_Ngram_chars_Transform (T_3)
6. Return F_T

End

Tập lệnh, CSS, img và các tập tin neo (F3, F4, F5 và F6). Các tập tin JavaScript bên ngoài hoặc các tập tin Cascading Style Sheets (CSS) bên ngoài là các tập tin riêng biệt có thể được truy cập bằng cách tạo một liên kết trong phần đầu của trang web. Tệp JavaS cript, CSS, hình ảnh, v.v. có thể chứa mã độc khi tải trang web hoặc nhấp vào một liên kết cụ thể.
Hơ n nữa, các trang web lừa đảo có nội dung dễ vỡ và thiếu chuyên nghiệp khi số lượng siêu liên kết trỏ đến một tên miền khác tăng lên. Chúng ta có thể sử dụng thẻ <img> và <script> có thuộc tính "src" để trích xuất hình ảnh và các tập tin JavaScript bên ngoài trong trang web. Tươ ng tự, CSS và các tập tin neo nằm trong thuộc tính "href " trong thẻ <link> và <a>. Trong các phươ ng trình. (1-4), về cơ bản, chúng tôi đã tính toán tỷ lệ thẻ img và script có thuộc tính "src", thẻ liên kết và thẻ neo có thuộc tính "href" trên tổng số siêu liên kết có sẵn trong một trang web, các thẻ này thường liên kết đến hình ảnh, Cần có các tập tin Javascript, neo và CSS cho một trang web

$$F3 = \begin{cases} \frac{FScript\_files}{FTotal} & \text{nếu FTotal} > 0 \\ 0 & \text{nếu FTotal} = 0 \end{cases} \quad (1)$$

$$F4 = \begin{cases} \frac{FCSS\_files}{\text{Tổng cộng } 0} & \text{nếu FTotal} > 0 \\ & \text{nếu FTotal} = 0 \end{cases} \quad (2)$$

$$F5 = \begin{cases} \frac{FImg\_files}{\text{Tổng cộng } 0} & \text{nếu FTotal} > 0 \\ & \text{nếu FTotal} = 0 \end{cases} \quad (3)$$

$$F6 = \begin{cases} \frac{Fa\_files}{\text{Tổng cộng } 0} & \text{nếu FTotal} > 0 \\ & \text{nếu FTotal} = 0 \end{cases} \quad (4)$$

trong đó FScript_files, FCSS_files, FImg_files, Fa_files là số lượng Javascript, CSS, hình ảnh, các tập tin neo hiện có trong một trang web và FTotal là tổng số siêu liên kết có sẵn trong một trang web.

Siêu liên kết trống (F7 và F8). Trong siêu liên kết trống, thuộc tính "href" hoặc "src" của thẻ neo, liên kết, tập lệnh hoặc img không chứa bất kỳ URL nào. Liên kết trống sẽ quay trở lại cùng một trang web khi người dùng nhấp vào nó.
Một trang web lành tính chứa nhiều trang web; do đó, kẻ lừa đảo không đặt bất kỳ giá trị nào trong các siêu liên kết để làm cho trang web lừa đảo hoạt động giống như trang web lành tính và các siêu liên kết trông có vẻ đang hoạt động trên trang web lừa đảo. Ví dụ: <a href="#">, <a href="#content"> và <a href="javascript:void(0);"> Mã hóa HTML được sử dụng để thiết kế các siêu liên kết rỗng 24. Để thiết lập các siêu liên kết trống các tính năng siêu liên kết, chúng tôi xác định tỷ lệ siêu liên kết trống tới

Machine Translated by Google
www.nature.com/scientificreports/

tổng số siêu liên kết có sẵn trong một trang web và tỷ lệ thẻ liên kết không có thuộc tính "href" trên tổng số siêu liên kết trong một trang web. Các công thức sau đây được sử dụng để tính toán các tính năng siêu liên kết trống

$$F7 = \begin{cases} \frac{Fa\_null}{Tổng\ cộng\ 0} & \text{nếu FTotal} > 0 \\ \text{nếu FTotal} = 0 \end{cases} \quad (5)$$

$$F8 = \begin{cases} \frac{Null}{Tổng\ số\ F} & \text{nếu FTotal} > 0 \\ 0 \text{ nếu FTotal} = 0 \end{cases} \quad (6)$$

trong đó Fa_null và Fnull là số lượng thẻ liên kết không có thuộc tính href và siêu liên kết rỗng trong một trang web.

Tính năng tổng siêu liên kết (F9). Các trang web lừa đảo thường chứa các trang tối thiểu so với các trang web lành tính.
Hơ n nữa, đôi khi trang web lừa đảo không chứa bất kỳ siêu liên kết nào vì những kẻ lừa đảo thường chỉ tạo một trang đăng nhập. Phươ ng trình (7) tính toán số lượng siêu liên kết trong một trang web bằng cách trích xuất các siêu liên kết từ thẻ neo, liên kết, tập lệnh và img trong mã nguồn HTML.

$$F9 = \text{Tổng số siêu liên kết có trong một trang web} \quad (7)$$

Các siêu liên kết bên trong và bên ngoài (F10, F11 và F12). Tên miền cơ sở trong siêu liên kết bên ngoài khác với tên miền trang web, không giống như siêu liên kết bên trong; tên miền cơ sở giống với tên miền của trang web. Các trang web lừa đảo có thể chứa nhiều siêu liên kết bên ngoài dẫn đến các trang web mục tiêu do tội phạm mạng thường sao chép mã HTML từ các trang web được ủy quyền được nhắm mục tiêu để tạo các trang web lừa đảo của chúng. Hầu hết các siêu liên kết trong một trang web lành tính đều chứa tên miền cơ sở tươ ng tự, trong khi nhiều siêu liên kết trong một trang web lừa đảo có thể bao gồm tên miền trang web lành tính tươ ng ứng. Theo cách tiếp cận của chúng tôi, các siêu liên kết bên trong và bên ngoài được trích xuất từ thuộc tính "src" của img, script, thẻ khung, thuộc tính "hành động" của thẻ biểu mẫu và thuộc tính "href" của thẻ neo và thẻ liên kết. Chúng tôi tính toán tỷ lệ siêu liên kết bên trong trên tổng số liên kết có sẵn trong một trang web (Phươ ng trình 8) để thiết lập tính năng siêu liên kết bên trong và tỷ lệ siêu liên kết bên ngoài trên tổng số liên kết (Phươ ng trình 9) để đặt tính năng siêu liên kết bên ngoài. Ngoài ra, để thiết lập tính năng siêu liên kết bên ngoài/nội bộ, chúng tôi tính toán tỷ lệ siêu liên kết bên ngoài so với siêu liên kết bên trong (Phươ ng trình 10). Một số lượng cụ thể đã được sử dụng như một cách để phát hiện các trang web bị nghi ngờ trong một số nghiên cứu trước đây5,9,24 mà các tính năng này được sử dụng để phân loại. Ví dụ: nếu tỷ lệ siêu liên kết bên ngoài so với tổng số liên kết lớn hơ n 0,5 thì điều đó sẽ cho thấy trang web đó đang lừa đảo. Tuy nhiên, việc xác định một số cụ thể làm phát hiện tham số có thể gây ra lỗi trong phân loại.

$$F10 = \begin{cases} \frac{FNội\ bộ}{Tổng\ cộng\ 0} & \text{nếu FTotal} > 0 \\ \text{nếu FTotal} = 0 \end{cases} \quad (số\ 8)$$

$$F11 = \begin{cases} \frac{F\ bên\ ngoài}{Tổng\ cộng\ 0} & \text{nếu FTotal} > 0 \\ \text{nếu FTotal} = 0 \end{cases} \quad (9)$$

$$F12 = \begin{cases} \frac{F\ bên\ ngoài}{FINội\ bộ\ 0} & \text{nếu FInternal} > 0 \\ \text{nếu FInternal} = 0 \end{cases} \quad (10)$$

trong đó FInternal, FExternal và FTotal là số lượng siêu liên kết bên ngoài, nội bộ và tổng số trong một trang web.

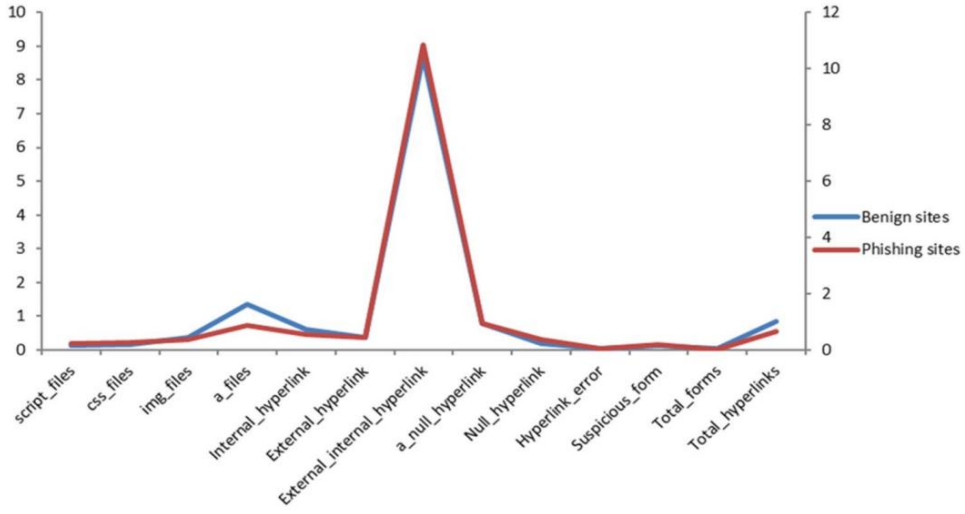
Lỗi siêu liên kết (F13). Những kẻ lừa đảo đôi khi thêm một số siêu liên kết vào trang web giả mạo là các liên kết đã chết hoặc bị hỏng.
Trong tính năng lỗi siêu liên kết, chúng tôi kiểm tra xem siêu liên kết có phải là URL hợp lệ trên trang web hay không. Chúng tôi không xem xét mã phản hồi lỗi 403 và 404 của siêu liên kết do tốn thời gian truy cập internet để lấy mã phản hồi của từng liên kết. Lỗi siêu liên kết được xác định bằng cách chia tổng số liên kết không hợp lệ cho tổng số liên kết như được biểu thị trong biểu thức. (11)

$$F13 = \begin{cases} \frac{FLỗi}{Tổng\ cộng\ 0} & \text{nếu FTotal} > 0 \\ \text{nếu FTotal} = 0 \end{cases} \quad (11)$$

trong đó FError là tổng số siêu liên kết không hợp lệ.

Tính năng biểu mẫu đăng nhập (F14 và F15). Trong trang web lừa đảo, thủ thuật phổ biến để lấy thông tin cá nhân của người dùng là đưa vào biểu mẫu đăng nhập. Trong trang web lành tính, thuộc tính hành động của biểu mẫu đăng nhập thường bao gồm một siêu liên kết có miền cơ sở tươ ng tự như xuất hiện trong địa chỉ trình duyệt bar24. Tuy nhiên, trong các trang web lừa đảo, thuộc tính hành động biểu mẫu bao gồm một URL có miền cơ sở khác (liên kết ngoài), liên kết trống hoặc URL không hợp lệ (Phươ ng trình 13). Đặc điểm biểu mẫu đáng ngờ (Phươ ng trình 14) được xác định bằng cách chia tổng số biểu mẫu đáng ngờ S cho tổng số biểu mẫu có sẵn trong một trang web (Phươ ng trình 12)

$$F14 = \text{Tổng số biểu mẫu có trong một trang web} \quad (12)$$



Hình 4. Phân phối các tính năng dựa trên siêu liên kết trong dữ liệu của chúng tôi.

$$S = \begin{cases} 1 \text{ nếu URL của trường hành động là Null} \\ 1 \text{ nếu URL của trường hành động không hợp lệ} \\ 1 \text{ nếu URL của hành động được gửi là liên kết bên ngoài} \\ 0 \text{ Ngược lại} \end{cases} \tag{13}$$

$$F15 = \begin{cases} \frac{FS}{\text{Tổng cộng}} & \text{nếu LTotal} > 0 \\ 0 \text{ nếu LTotal} = 0 \end{cases} \tag{14}$$

trong đó FS và LTotal là số lượng biểu mẫu đáng ngờ và tổng số biểu mẫu có trong một trang web.

Hình 4 cho thấy sự so sánh giữa các tính năng siêu liên kết lành tính và fshing dựa trên tỷ lệ xuất hiện trung bình trên mỗi tính năng trong mỗi trang web trong tập dữ liệu của chúng tôi. Từ hình ảnh, chúng tôi nhận thấy rằng tỷ lệ siêu liên kết bên ngoài so với siêu liên kết bên trong và siêu liên kết rỗng trong các trang web lừa đảo cao hơ n so với các trang web lành tính. Trong khi đó, các trang web lành tính chứa nhiều liên kết neo, siêu liên kết nội bộ và tổng số siêu liên kết.

Các thuật toán phân loại. Để đo lường hiệu quả của các tính năng được đề xuất, chúng tôi đã sử dụng nhiều bộ phân loại học máy khác nhau như eXtreme gradient Boosting (XGBoost), Random Forest, Logistic Reres sion, Naïve Bayes và Ensemble of Random Forest và Adaboost để huấn luyện phươ ng pháp đề xuất của chúng tôi. Mục đích chính của việc so sánh các trình phân loại khác nhau là để tìm ra trình phân loại tốt nhất cho tập tính năng của chúng tôi. Để áp dụng các trình phân loại học máy khác nhau, gói Scikit-learn.org được sử dụng và Python được sử dụng để trích xuất tính năng. Từ kết quả thực nghiệm, chúng tôi nhận thấy XGBoost vượt trội hơ n các công cụ phân loại khác. Thuật toán XGBoost là một loại trình phân loại tổng hợp, giúp chuyển đổi những người học yếu thành những người mạnh mẽ và thuận tiện cho bộ tính năng được đề xuất của chúng tôi, do đó nó có hiệu suất cao.

XGBoost (tăng cường độ dốc cực cao) là một hệ thống máy học có thể mở rộng để tăng cường cây được đề xuất bởi xi Rd là Chen có N trang web trong tập dữ liệu xi, yi |i = 1, 2, ..., N ,và Guestrin40. Giả sử được trích xuất các đặc điểm liên quan đến trang web thứ i thứ , yi {0, 1} là nhãn lớp, sao cho yi = 1 khi và chỉ nếu trang web đó là trang web lừa đảo được gắn nhãn. Đầu ra cuối cùng fK (x) của mô hình như sau41,46:

$$fk(x) = l\ yi, fk(x) = \sum_{toi=1}^{N} l\ yi, fk\ \ 1(xi) + Gk(xi) + \Omega(Gk) \tag{15}$$

trong đó l là hàm mất huấn luyện và $\Omega(Gk) = \gamma T + \frac{1}{2}\sum_{t=1}^{T} \omega 2_t$ là thuật ngữ quy định, vì XGBoost giới thiệu chươ ng trình đào tạo bổ sung và tất cả những người học cơ sở k-1 trước đó đều được sửa đổi, nên ở đây chúng tôi giả định rằng chúng tôi đang ở bước k tối ưu hóa hàm fk(x) của chúng ta, T là số nút lá trong trình học cơ sở Gk, γ là độ phức tạp của mỗi lá, λ là tham số để chia tỷ lệ hình phạt và ωt là giá trị đầu ra tại mỗi nút lá fnal . Nếu áp dụng khai triển Taylor để khai triển hàm Loss tại fk-1(x), chúng ta sẽ có41:

$$\begin{aligned} ly, fk\ \ 1(x) + Gk(x) &\approx \sum_{toi=1}^{N} l\ yi, fk\ \ 1(xi) + Gk(xi) \\ &= \sum_{toi=1}^{N} l\ yi, fk\ \ 1(xi) + giGk(xi) + \frac{1}{2}hiG2\ k(xi) + \gamma T + \frac{1}{2}\sum_{t=1}^{T} \omega 2_t \end{aligned} \tag{16}$$

Machine Translated by Google
www.nature.com/scientificreports/

Tập dữ liệu	Các trang web lành tính	Các trang web lừa đảo	Tổng cộng
D1	32.972	27.280	60.252
D2	36.400	37.175	73.575

Bảng 3. Phân phối dữ liệu.

Công thức	Sự miêu tả
TPR = $\frac{Np\ \ p}{Np} \times 100$	Tỷ lệ các trang web lừa đảo được phân loại là lừa đảo trên tổng số trang web lừa đảo
TNR = $\frac{NB\ \ B}{NB} \times 100$	Tỷ lệ các website lành tính được phân loại là lành tính trên tổng số website lành tính
FPR = $\frac{NB\ \ p}{NB} \times 100$	Tỷ lệ các trang web lành tính được phân loại là lừa đảo trên tổng số trang web lành tính
FNR = $\frac{Np\ \ B}{Np} \times 100$	Tỷ lệ các trang web lừa đảo được phân loại là lành tính trên tổng số trang web lừa đảo
độ chính xác = $\frac{(TP+TN)}{(TP+TN+FP+FN)} \times 100$	Tỷ lệ các trang web lừa đảo và lành tính được phân loại chính xác trên tổng số trang web
Độ chính xác = $\frac{(TP)}{(TP+FP)} \times 100$	Tỷ lệ các trang web lừa đảo được phân loại chính xác là lừa đảo trên tổng số trang web lừa đảo được phân loại là lừa đảo và các trang web lành tính được phân loại là lừa đảo
Nhớ lại = $\frac{(TP)}{(TP+FN)} \times 100$	Tỷ lệ các trang web lừa đảo được phân loại chính xác trên tổng số trang web lừa đảo
F Điểm = (độ $\frac{(2\times độ\ chính\ xác\times thu\ hồi)}{chính\ xác\ +\ thu\ hồi)} \times 100$	Trung bình điều hòa có trọng số của độ chính xác và tỷ lệ thu hồi

Bảng 4. Các số liệu thống kê khác nhau được sử dụng để đo lường hiệu suất của phương pháp tiếp cận của chúng tôi.

trong đó gi = $\frac{l(yi,fk\ \ 1(xi))}{fk\ \ 1(x)}$, hi $\frac{l(yi,fk\ \ 1(xi))}{fk\ \ 1(x)}$ lần lượt là đạo hàm thứ nhất và thứ hai của hàm Mất mát.

= XGBoost classifer là một loại trình phân loại tổng hợp, giúp chuyển đổi những người học yếu thành những người mạnh mẽ và thuận tiện cho bộ tính năng được đề xuất của chúng tôi để dự đoán các trang web lừa đảo, do đó nó có hiệu suất cao. Hơn nữa, XGBoost cung cấp một số lợi thế, một số lợi thế bao gồm: (i) Sức mạnh để xử lý các giá trị còn thiếu trong tập huấn luyện, (ii) xử lý các tập dữ liệu khổng lồ không đưa vào bộ nhớ và (iii) Để tính toán nhanh hơn, XGBoost có thể sử dụng nhiều lõi trên CPU. Các trang web được phân loại thành hai loại: lừa đảo và vô hại bằng cách sử dụng trình phân loại nhị phân. Khi người dùng yêu cầu một trang web mới, bộ phân loại XGBoost đã được đào tạo sẽ xác định tính hợp lệ của một trang web cụ thể từ vectơ đặc trưng đã tạo.

Thí nghiệm và phân tích kết quả

Trong phần này, chúng tôi mô tả tập dữ liệu đào tạo và thử nghiệm, số liệu hiệu suất, chi tiết triển khai và kết quả của phương pháp của chúng tôi. Các tính năng được đề xuất được mô tả trong phần "Trích xuất tính năng" được sử dụng để xây dựng bộ phân loại nhị phân, giúp phân loại chính xác các trang web lừa đảo và lành tính.

Bộ dữ liệu. Chúng tôi đã thu thập dữ liệu từ hai nguồn để triển khai thử nghiệm. Các trang web lành tính được thu thập vào tháng 2 năm 2020 từ Stuf Gate42, trong khi các trang web lừa đảo được thu thập từ PhishTank43, đã được xác thực từ tháng 8 năm 2016 đến tháng 4 năm 2020. Tập dữ liệu của chúng tôi bao gồm 60.252 trang web và mã nguồn HTML của chúng, trong đó có 27.280 trang web lừa đảo. và 32.972 trường hợp là lành tính. Bảng 3 cung cấp sự phân bố của các trường hợp lành tính và lừa đảo. Chúng tôi đã chia tập dữ liệu thành hai nhóm trong đó D1 là tập dữ liệu của chúng tôi và D2 là tập dữ liệu được sử dụng trong tài liệu hiện có6 . Hệ thống quản lý cơ sở dữ liệu (tức là pgAdmin) đã được sử dụng với python để nhập và xử lý trước dữ liệu. Các tập dữ liệu được chia ngẫu nhiên theo tỷ lệ 80:20 tương ứng cho việc huấn luyện và kiểm tra.

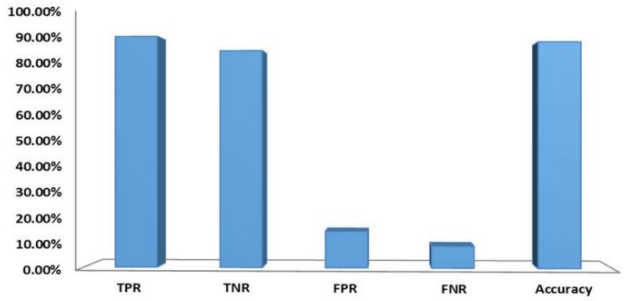
Chỉ số hiệu suất. Để đo lường hiệu suất của phương pháp chống lừa đảo được đề xuất, chúng tôi đã sử dụng các số liệu thống kê khác nhau như tỷ lệ dương tính thật (TPR), tỷ lệ âm tính thật (TNR), tỷ lệ dương tính giả (FPR), tỷ lệ âm tính giả (FNR), độ nhạy hoặc thu hồi, độ chính xác (Acc), độ chính xác (Pre), F-Score, AUC và chúng được trình bày trong Bảng 4. NB và NP lần lượt biểu thị tổng số trang web lành tính và lừa đảo. NB B là các trang web lành tính được đánh dấu chính xác là lành tính, NB P là các trang web lành tính được đánh dấu không chính xác là lừa đảo, NP P là các trang web lừa đảo được đánh dấu chính xác là lừa đảo và NP B là các trang web lừa đảo không chính xác được đánh dấu là lành tính. Vòm đặc tính vận hành máy thu (ROC) và AUC thường được sử dụng để đánh giá các thước đo của bộ phân loại nhị phân. Tọa độ ngang của vòm ROC là FPR, cho biết khả năng trang web lành tính bị phân loại sai là lừa đảo; thứ tự là TPR, cho biết khả năng trang web lừa đảo được xác định là lừa đảo.

Đánh giá các tính năng. Trong phần này, chúng tôi đã đánh giá hiệu suất của các tính năng được đề xuất (URL và HTML). Chúng tôi đã triển khai các trình phân loại Học máy (ML) khác nhau để đánh giá tính năng được sử dụng trong phương pháp của chúng tôi. Trong Bảng 5, chúng tôi đã trích xuất các đặc điểm văn bản khác nhau như cấp độ từ TF-IDF, cấp độ N-gram TF-IDF (độ dài của n-gram trong khoảng từ 2 đến 3), cấp độ ký tự TF-IDF, vectơ đếm (bag-of- từ), vectơ chuỗi từ, nhúng từ được đào tạo trước từ toàn cầu sang vectơ (GloVe), nhúng từ được đào tạo, chuỗi ký tự vec

Machine Translated by Google
www.nature.com/scientificreports/

phân loại	Đặc điểm nội dung văn bản	Trước (%)	Thu hồi (%)	Điểm F (%)	AUC (%)	Acc (%)
LR	Cấp độ từ TF-IDF	85,68	88,25	86,95	85,38	85,62
	Mức N-gram TF-IDF	85,23	85,42	85,33	83,93	84,05
	Cấp độ ký tự TF-IDF	84,55	87,15	85,83	84,13	84,39
	Đếm vectơ	86,84	79,12	82,80	82,45	82,16
	Vectơ chuỗi từ	55,87	83,27	66,87	52,61	55,23
XGBoost	Cấp độ từ TF-IDF	88,44	88,56	88,50	87,41	87,52
	Mức N-gram TF-IDF	87,77	86,51	87,13	86.10	86,14
	Cấp độ ký tự TF-IDF	89,01	90,58	89,79	88,65	88,82
	Vectơ chuỗi từ	82,66	85,87	84,23	82,24	82,55
	Đếm vectơ	88,26	87,75	88.00	86,95	87,02
	Vectơ chuỗi ký tự	81,47	87,81	84,52	82.05	82,54
RF	Cấp độ từ TF-IDF	85,94	92,67	89,18	87,34	87,80
	Mức N-gram TF-IDF	86,77	89,57	88,14	86,68	86,93
	Cấp độ ký tự TF-IDF	85,44	92,81	88,97	87,02	87,51
	Đếm vectơ	85,81	93,08	89,30	87,41	87,90
	Vectơ chuỗi từ	81,56	90,71	85,89	83,19	83,83
	Vectơ chuỗi ký tự	79,51	93,91	86,11	82,60	83,56
NB	Cấp độ từ TF-IDF	84,50	79,12	81,72	80,95	80,79
	Mức N-gram TF-IDF	82,45	71,16	76,39	76,59	76,13
	Cấp độ ký tự TF-IDF	76,45	81,89	79,08	75,98	76,49
	Đếm vectơ	82,62	71,63	76,74	76,88	76,43
	Vectơ chuỗi từ	62,89	42,66	50,83	56,39	55.22
DNN	Cấp độ từ TF-IDF	87,08	91.20	89,09	87,57	87,88
	Mức N-gram TF-IDF	88,12	84,29	86,17	85,40	85,31
	Cấp độ ký tự TF-IDF	88,32	91,62	89,94	88,62	88,40
	Đếm vectơ	87,49	89,46	88,47	87,14	87,34
	Vectơ chuỗi từ	54,26	100,0	70,35	50,0	54,26
	Vectơ chuỗi ký tự	76,41	91,43	83,25	78,97	80.03
LSTM	Nhúng từ được đào tạo trước của GloVe	87.05	90,79	88,88	87,38	87,67
	Nhúng từ được đào tạo	88,14	89,20	88,66	87,48	87,62
CNN	Nhúng ký tự Nhúng từ	82.06	89,34	85,54	83,08	83,61
	được đào tạo	89,57	85,00	87,22	86,62	86,49

Bảng 5. Hiệu suất của các đặc điểm dựa trên văn bản khác nhau trên tập dữ liệu D1 với các bộ phân loại khác nhau. Các giá trị quan trọng nằm ở [đậm].



Hình 5. Hiệu suất của các tính năng nội dung văn bản.

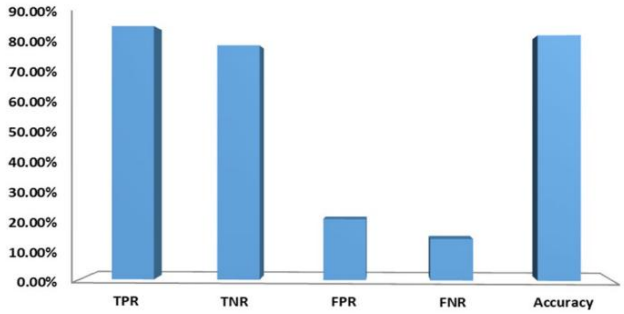
tors và triển khai nhiều trình phân loại khác nhau như XGBoost, Rừng ngẫu nhiên, hồi quy logistic, Naïve Bayes, Mạng thần kinh sâu (DNN), Mạng thần kinh chuyển đổi (CNN) và mạng Bộ nhớ ngắn hạn dài (LSTM). Mục đích chính của thử nghiệm này là tiết lộ các tính năng nội dung văn bản tốt nhất thuận tiện cho dữ liệu của chúng tôi. Từ kết quả thử nghiệm, người ta nhận thấy rằng các tính năng ở cấp độ ký tự TF-IDF vượt trội hơn các tính năng khác về độ chính xác, độ chính xác, F-Score, Recall và AUC đáng kể khi sử dụng bộ phân loại XGBoost và DNN.

Do đó, chúng tôi đã triển khai kỹ thuật cấp ký tự TF-IDF để tạo các đặc điểm văn bản (F2) của trang web. Hình 5

Machine Translated by Google
www.nature.com/scientificreports/

phân loại	Độ chính xác (%)	Thu hồi (%)	F_Measure (%)	AUC (%)	Độ chính xác (%)
RF	77,59	86.10	81,63	82,57	82,27
Hòa tấu 77,39		86,23	81,57	82,50	82,18
LR	69.05	55,67	61,65	67,32	68,31
NB	68,31	31:60	43.21	59,62	62.01
XGBoost	75,55	84,77	79,90	80,82	80,49

Bảng 6. Hiệu suất của các tính năng siêu liên kết được đề xuất trên D1 với các phân loại khác nhau. Các giá trị quan trọng nằm ở [đậm].



Hình 6. Hiệu suất của các tính năng dựa trên siêu liên kết.

phân loại	Đặc trưng	Trước (%)	Thu hồi (%)	Điểm F (%)	AUC (%)	ACC (%)	
LR	LỖI	74,67	67,92	71.13	74,25	74,79	
	FHTML	83,50	81,98	82,74	84,16	84,35	
	FURL+ HTML	77,71	68,74	72,95	76,06	76,68	
NB	LỖI	81,41	22.09	34,76	58,92	62.06	
	FHTML	65,67	87,57	75,06	74,49	73,38	
	FHTM + URL	86,99	62,15	72,51	77,16	78,44	
hòa tấu	LỖI	98,42	92,05	95,13	95,40	95,69	
	FHTML	90,22	82.01	85,92	87,25	87,70	
	FURL+ HTML	93,89	87,85	90,77	91,52	91,83	
RF	LỖI	98,54	92,14	95,23	95,49	95,78	
	FHTML	90,77	81,98	86,16	87,48	87,95	
	LÔNG + HTML	93,81	86,79	90.16	90,98	91,34	
XGBoost	LỖI	99,58	92,27	95,79	95,97	96,29	
	FHTML	88,21	87,68	87,94	88,90	89,01	
	LÔNG + HTML	98,28	94,56	96,38	96,58	96,76	

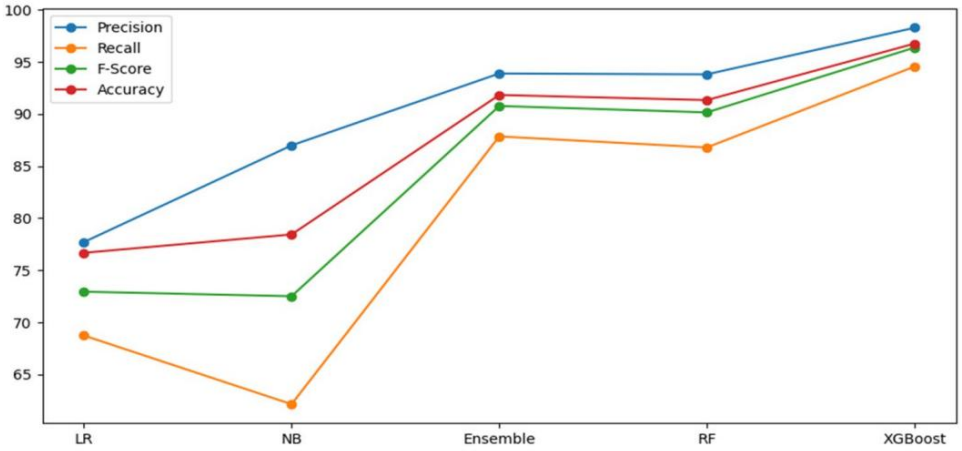
Bảng 7. Hiệu suất của các tổ hợp đặc trưng khác nhau trên tập dữ liệu D1 với các bộ phân loại khác nhau. Các giá trị quan trọng nằm ở [đậm].

trình bày hiệu suất của các tính năng dựa trên nội dung văn bản. Như được hiển thị trong hình, các tính năng văn bản có thể chặn một lượng lớn trang web lừa đảo một cách chính xác và đạt độ chính xác 88,82%.

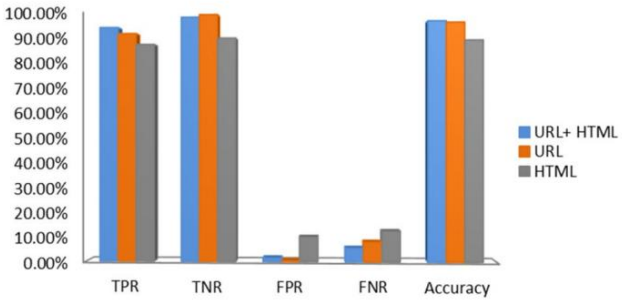
Bảng 6 thể hiện kết quả thử nghiệm với tính năng siêu liên kết. Từ kết quả thực nghiệm, người ta nhận thấy rằng bộ phân loại Rừng ngẫu nhiên vượt trội hơ n các bộ phân loại khác với độ chính xác 82,27%, độ chính xác 77,59%, F_Measure 81,63%, thu hồi 86,10% và AUC là 82,57%. Người ta cũng nhận thấy rằng các lớp đồng bộ và XGBoost đạt độ chính xác tốt lần lượt là 82,18% và 80,49%. Hình 6 trình bày kết quả phân loại của các tính năng dựa trên siêu liên kết (F3–F15). Như được hiển thị trong hình, các tính năng dựa trên siêu liên kết có thể làm rõ chính xác 79,04% trang web lành tính và 86,10% trang web lừa đảo.

Trong Bảng 7, chúng tôi đã tích hợp các tính năng của URL và HTML (siêu liên kết và văn bản) bằng cách sử dụng nhiều bộ phân loại khác nhau để xác minh hành vi bổ sung trong việc phát hiện các trang web lừa đảo. Từ kết quả thực nghiệm, người ta nhận thấy rằng phân loại LR có đủ độ chính xác, độ chính xác, F-Score, AUC và khả năng thu hồi về các tính năng HTML. Ngược lại, NB classifer có độ chính xác, độ chính xác, F-Score, AUC và khả năng thu hồi tốt khi kết hợp tất cả các tính năng. Các bộ phân loại RF và tập hợp đã đạt được độ chính xác, khả năng thu hồi, Điểm F và AUC cao đối với các tính năng dựa trên URL.

Machine Translated by Google
www.nature.com/scientificreports/



Hình 7. Kết quả thử nghiệm của các bộ phân loại khác nhau đối với các đặc điểm kết hợp.



Hình 8. Hiệu suất của các kết hợp tính năng khác nhau bằng XGBoost trên tập dữ liệu D1.

Ma trận hỗn loạn	Dự đoán	
	PN	
Các tính năng dựa trên URL		
Thật sự		
P	5086	426
N	21	6518
Các tính năng dựa trên HTML		
Thật sự		
P	4833	679
N	646	5893
Các tính năng dựa trên URL+HTML		
Thật sự		
P	5212	300
N	91	6448

Bảng 8. Ma trận nhầm lẫn của phươ ng pháp đề xuất trên tập dữ liệu D1.

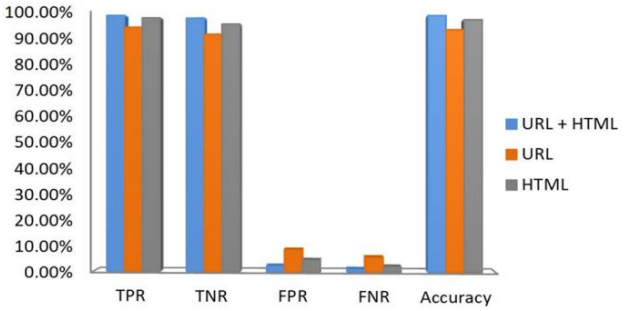
Phân loại XGBoost vượt trội hơ n các phân loại khác với độ chính xác 96,76%, Điểm F là 96,38%, AUC là 96,58% và thu hồi 94,56% khi kết hợp tất cả các tính năng. Người ta nhận thấy rằng các tính năng URL và HTML có giá trị trong việc phát hiện lừa đảo. Tuy nhiên, một loại tính năng không phù hợp để xác định tất cả các loại trang web lừa đảo và không mang lại độ chính xác cao. Vì vậy, chúng tôi đã kết hợp tất cả các tính năng để có được các tính năng toàn diện hơ n.
Kết quả về các bộ phân loại khác nhau của bộ tính năng kết hợp cũng được hiển thị trong Hình 7. Trong Hình 8 , chúng tôi so sánh ba bộ tính năng về độ chính xác, TNR, FPR, FNR và TPR.

Ma trận nhầm lẫn được sử dụng để đo lường kết quả trong đó mỗi hàng của ma trận biểu thị các cá thể trong một lớp được dự đoán, trong khi mỗi cột biểu thị các cá thể trong một lớp thực tế (hoặc ngược lại). Ma trận nhầm lẫn của phươ ng pháp đề xuất được tạo như được trình bày trong Bảng 8. Từ kết quả, kết hợp tất cả các loại tính năng lại với nhau như một thực thể được xác định chính xác 5212 trên 5512 trang web lừa đảo và 6448 trên 6539

Machine Translated by Google
www.nature.com/scientificreports/

Đặc trưng	Trước (%) Thu hồi (%) Điểm F (%) AUC (%) Acc (%)				
LỖI	92.16	94,58	93,36	93.11	93,14
FHTML	95,84	98,05	96,94	96,82	96,84
LÔNG + HTML	98.01	99,04	98,52	98,47	98,48

Bảng 9. Kết quả của phương pháp đề xuất trên tập dữ liệu D2. Các giá trị quan trọng nằm ở [đậm].



Hình 9. Hiệu suất của phương pháp đề xuất trên tập dữ liệu D2.

Tác giả	Trước (%) Thu hồi (%) Điểm F (%) Acc (%)			
Sahingoz và cộng sự.6	97,00	99,00	98,00	97,98
Rao và cộng sự.13	98.04	98,42	98,23	98,25
Chatterjee và Namin30 86.71 Phương		88.00	87.30	90.10
pháp đề xuất	98.01	99,04	98,52	98,48

Bảng 10. So sánh phương pháp đề xuất với các phương pháp tiêu chuẩn khác trên tập dữ liệu D2. Các giá trị quan trọng nằm ở [đậm].

Tác giả	Trước (%)	Nhớ lại (%)	Điểm F (%)	Tích lũy (%)	TNR (%)	FPR (%)
Lê và cộng sự29	96,38	90.06	93,12	93,91	97,15	2,84
Aljofey và cộng sự.3	94,84	97,86	96,33	95,94	93,64	6,35
Phương pháp đề xuất	98,28	94,56	96,38	96,76	98,61	1,39

Bảng 11. So sánh phương pháp đề xuất với các phương pháp tiêu chuẩn khác trên tập dữ liệu D1. Các giá trị quan trọng nằm ở [đậm].

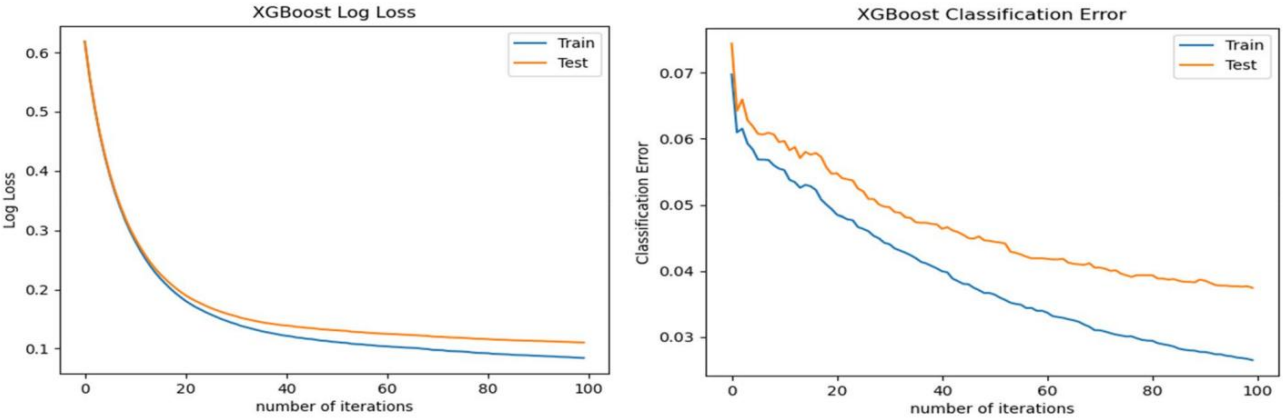
các trang web lành tính và đạt độ chính xác 96,76%. Cách tiếp cận của chúng tôi mang lại tỷ lệ dương tính giả thấp (tức là ít hơn 1,39% các trang web lành tính được phân loại không chính xác là lừa đảo) và tỷ lệ dương tính thật cao (tức là hơn 94,56% trang web lừa đảo được phân loại chính xác). Chúng tôi cũng đã thử nghiệm các bộ tính năng (URL và HTML) trên tập dữ liệu D2 hiện có. Vì tập dữ liệu D2 chỉ chứa các URL hợp pháp và độc hại nên chúng tôi cần trích xuất các tính năng mã nguồn HTML cho các URL này. Kết quả được trình bày trong Bảng 9 và Hình 9. Từ kết quả, người ta nhận thấy rằng việc kết hợp tất cả các loại tính năng đã vượt trội hơn các bộ tính năng khác với độ chính xác đáng kể là 98,48%, TPR là 99,04% và FPR là 2,09%.

So sánh với các phương pháp hiện có. Trong thử nghiệm này, chúng tôi so sánh phương pháp của chúng tôi với các phương pháp chống lừa đảo hiện có. Lưu ý rằng chúng tôi đã áp dụng Le et al.29 và Aljofey et al.3 làm việc trên tập dữ liệu D1 để đánh giá hiệu quả của phương pháp được đề xuất. Trong khi so sánh phương pháp đề xuất với Sahingoz et al.6
, Rao et al.13, Chatterjee và Namin30 hoạt động, chúng tôi đã đánh giá cách tiếp cận của chúng tôi trên tập dữ liệu chuẩn D26,13,30 dựa trên bốn số liệu thống kê được sử dụng trong bài báo. Kết quả so sánh được trình bày trong Bảng 10. Từ kết quả, có thể thấy rằng phương pháp của chúng tôi mang lại hiệu suất tốt hơn các phương pháp khác được thảo luận trong tài liệu, điều này cho thấy hiệu quả của việc phát hiện các trang web lừa đảo so với các phương pháp hiện có.

Trong Bảng 11, chúng tôi đã triển khai các phương pháp của Le và cộng sự.29 và Aljofey và cộng sự.3 cho tập dữ liệu D1 của chúng tôi và phương pháp tiếp cận của chúng tôi đã thực hiện các phương pháp khác với độ chính xác là 96,76%, độ chính xác là 98,28% và Điểm F là 96,38%. Cũng cần đề cập rằng Aljofey et al. phương pháp đạt được tỷ lệ thu hồi 97,86%, cao hơn 3,3% so với phương pháp của chúng tôi, trong khi

Đặc trưng	Thời gian đào tạo (s)	Thời gian kiểm tra (s)	Thời gian phát hiện (s)
LỖI	12.7969	0,0548	0,356995
FHTML	1467,12	23.328	26.380,12
LÔNG + HTML	9352,64	33.338	26.371,05

Bảng 12. Thời gian huấn luyện, thử nghiệm và phát hiện phươ ng pháp đề xuất trên D1.



Hình 10. Đường cong học XGBoost về mất logarit và lỗi phân loại trên tập dữ liệu D1.

Phươ ng pháp này mang lại TNR cao hơ n 4,97% và FPR thấp hơ n 4,96%. Cách tiếp cận của chúng tôi xác định chính xác các trang web hợp pháp có TNR cao và FPR thấp. Một số phươ ng pháp phát hiện lừa đảo có khả năng thu hồi cao, tuy nhiên việc phân loại không chính xác các trang web hợp pháp còn nghiêm trọng hơ n so với việc phân loại không chính xác các trang web lừa đảo.

Thảo luận và hạn chế

Trang web lừa đảo có vẻ giống với trang web chính thức lành tính của nó và vấn đề là làm thế nào để phân biệt giữa chúng. Bài báo này đã đề xuất một phươ ng pháp chống lừa đảo mới, bao gồm các tính năng khác nhau (URL, siêu liên kết và văn bản) chưa bao giờ được xem xét. Cách tiếp cận được đề xuất là một giải pháp hoàn toàn phía khách hàng. Chúng tôi đã áp dụng các tính năng này trên nhiều thuật toán học máy khác nhau và nhận thấy rằng XGBoost đạt được hiệu suất tốt nhất. Mục đích chính của chúng tôi là thiết kế một phươ ng pháp tiếp cận thời gian thực, có tỷ lệ âm tính thật cao và tỷ lệ dươ ng tính giả thấp. Kết quả cho thấy phươ ng pháp tiếp cận của chúng tôi đã lọc chính xác các trang web lành tính với số lượng nhỏ các trang web lành tính được phân loại không chính xác là lừa đảo. Trong quá trình phân loại trang web lừa đảo, chúng tôi xây dựng tập dữ liệu bằng cách trích xuất các tính năng hữu ích và có liên quan từ các trang web lừa đảo và lành tính.

Một máy tính để bàn có bộ xử lý core™ i7 với tốc độ xung nhịp 3,4 GHz và RAM 16 GB được sử dụng để thực thi phươ ng pháp chống lừa đảo được đề xuất. Vì Python cung cấp sự hỗ trợ tuyệt vời cho các thư viện của nó và có thời gian biên dịch hợp lý nên phươ ng pháp đề xuất được triển khai bằng ngôn ngữ lập trình Python. Thư viện BeautifulSoup được sử dụng để phân tích HTML của URL được chỉ định. Thời gian phát hiện là thời gian từ khi nhập URL đến khi tạo kết quả đầu ra. Khi URL được nhập dưới dạng tham số, phươ ng pháp này sẽ cố gắng tìm nạp tất cả các tính năng cụ thể từ URL và mã HTML của trang web như được tranh luận trong phần trích xuất tính năng. Tiếp theo là phân loại URL hiện tại dưới dạng lành tính hoặc lừa đảo dựa trên giá trị của tính năng được trích xuất. Tổng thời gian thực hiện phươ ng pháp phát hiện trang web lừa đảo của chúng tôi là khoảng 2-3 giây, khá thấp và có thể chấp nhận được trong môi trường thời gian thực. Thời gian phản hồi phụ thuộc vào các yếu tố khác nhau, chẳng hạn như kích thước đầu vào, tốc độ Internet và cấu hình máy chủ. Sử dụng dữ liệu D1 của chúng tôi, chúng tôi cũng đã cố gắng tính toán thời gian dành cho việc đào tạo, thử nghiệm và phát hiện phươ ng pháp được đề xuất (tất cả các kết hợp tính năng) để phân loại trang web. Kết quả được đưa ra trong Bảng 12.

Để hiểu rõ hơ n về khả năng học tập, chúng tôi cũng trình bày lỗi phân loại cũng như mất nhật ký liên quan đến số lần lặp do XGBoost triển khai. Mất log, viết tắt của mất logarit là một hàm mất để phân loại cho biết cái giá phải trả cho sự thiếu chính xác của các dự đoán trong các vấn đề phân loại. Hình 10 cho thấy sự mất logarit và lỗi phân loại của phươ ng pháp XGBoost cho từng kỷ nguyên trên tập dữ liệu huấn luyện và kiểm tra D1. Từ việc xem xét fgure, chúng ta có thể lưu ý rằng thuật toán học đang hội tụ sau khoảng 100 lần lặp.

Hạn chế. Mặc dù phươ ng pháp đề xuất của chúng tôi đã đạt được độ chính xác vượt trội nhưng nó vẫn có một số hạn chế. Hạn chế đầu tiên là đặc điểm văn bản của phươ ng pháp phát hiện lừa đảo của chúng tôi phụ thuộc vào ngôn ngữ tiếng Anh. Điều này có thể gây ra lỗi trong việc tạo ra kết quả phân loại hiệu quả khi trang web đáng ngờ sử dụng ngôn ngữ khác ngoài tiếng Anh. Khoảng một nửa (60,5%) trang web sử dụng tiếng Anh làm ngôn ngữ văn bản44. Tuy nhiên, cách tiếp cận của chúng tôi sử dụng URL, phần ồn ào của HTML và các tính năng dựa trên siêu liên kết, độc lập với ngôn ngữ

Machine Translated by Google
www.nature.com/scientificreports/

đặc trưng. Hạn chế thứ hai là mặc dù phương pháp được đề xuất sử dụng các tính năng dựa trên URL, nhưng phương pháp của chúng tôi có thể không xác định được các trang web lừa đảo trong trường hợp kẻ lừa đảo sử dụng các đối tượng được nhúng (ví dụ: Javascript, hình ảnh, Flash, v.v.) để che khuất nội dung văn bản và Mã hóa HTML từ các giải pháp chống lừa đảo. Nhiều kẻ tấn công sử dụng tập lệnh phía máy chủ để ẩn mã nguồn HTML. Dựa trên thử nghiệm của chúng tôi, chúng tôi nhận thấy rằng các trang hợp pháp thường chứa các tính năng nội dung văn bản phong phú và số lượng siêu liên kết cao (Ít nhất một siêu liên kết trong mã nguồn HTML). Hiện nay, một số trang web lừa đảo có chứa phần mềm độc hại, chẳng hạn như ngựa Trojan cài đặt trên hệ thống của người dùng khi người dùng mở trang web. Do đó, hạn chế tiếp theo của phương pháp này là nó không đủ khả năng phát hiện phần mềm độc hại đính kèm vì phương pháp của chúng tôi không đọc và xử lý nội dung từ các tập tin bên ngoài của trang web, cho dù chúng có phải là tên miền chéo hay không. Cuối cùng, thời gian huấn luyện phương pháp của chúng tôi tương đối dài do vectơ chiều cao được tạo ra bởi các đặc điểm nội dung văn bản. Tuy nhiên, phương pháp được đào tạo tốt hơn nhiều so với các phương pháp cơ bản hiện có về độ chính xác.

Kết luận và công việc trong tương lai

Các cuộc tấn công vào trang web lừa đảo là một thách thức lớn đối với các nhà nghiên cứu và chúng tiếp tục cho thấy xu hướng gia tăng trong những năm gần đây. Kỹ thuật danh sách đen/danh sách trắng là cách truyền thống để giảm bớt các mối đe dọa như vậy. Tuy nhiên, các phương pháp này không phát hiện được các trang web lừa đảo không nằm trong danh sách đen (tức là các cuộc tấn công 0 ngày). Để cải tiến, các kỹ thuật học máy đang được sử dụng để tăng hiệu quả phát hiện và giảm tỷ lệ phân loại sai. Tuy nhiên, một số trong số chúng trích xuất các tính năng từ dịch vụ của bên thứ ba, công cụ tìm kiếm, lưu lượng truy cập trang web, v.v., rất phức tạp và khó truy cập. Trong bài viết này, chúng tôi đề xuất một phương pháp tiếp cận dựa trên máy học có thể phát hiện nhanh chóng và chính xác các trang web lừa đảo bằng cách sử dụng các tính năng URL và HTML của trang web nhất định. Cách tiếp cận được đề xuất là giải pháp hoàn toàn phía khách hàng và không dựa vào bất kỳ dịch vụ nào của bên thứ ba. Nó sử dụng các tính năng chuỗi ký tự URL mà không cần sự can thiệp của chuyên gia và các tính năng cụ thể của siêu liên kết xác định mối quan hệ giữa nội dung và URL của trang web. Hơn nữa, cách tiếp cận của chúng tôi trích xuất các đặc điểm cấp độ ký tự TF-IDF từ phần văn bản gốc và phần nhiễu của HTML của trang web nhất định.

Một tập dữ liệu mới được xây dựng để đo lường hiệu suất của phương pháp phát hiện lừa đảo và nhiều thuật toán phân loại khác nhau được sử dụng. Hơn nữa, hiệu suất của từng loại của bộ tính năng được đề xuất cũng được đánh giá. Theo kết quả thực nghiệm và so sánh từ các thuật toán phân loại được triển khai, bộ phân loại XGBoost với việc tích hợp tất cả các loại tính năng sẽ mang lại hiệu suất tốt nhất. Nó đạt được tỷ lệ dương tính giả là 1,39% và độ chính xác phát hiện tổng thể là 96,76% trên tập dữ liệu của chúng tôi. Độ chính xác 98,48% với tỷ lệ dương tính giả 2,09% trên tập dữ liệu chuẩn.

Trong công việc trong tương lai, chúng tôi dự định sẽ đưa vào một số tính năng mới để phát hiện các trang web lừa đảo có chứa phần mềm độc hại. Như chúng tôi đã nói trong phần "Hạn chế", phương pháp tiếp cận của chúng tôi không thể phát hiện phần mềm độc hại đính kèm với trang web lừa đảo. Ngày nay, công nghệ blockchain ngày càng phổ biến và dường như là mục tiêu hoàn hảo cho các cuộc tấn công lừa đảo như lừa đảo trên blockchain. Blockchain là một sổ cái mở và phân tán, có thể đăng ký các giao dịch giữa bên nhận và bên gửi một cách hiệu quả, một cách rõ ràng và liên tục, khiến nó trở nên phổ biến đối với các nhà đầu tư45. Vì vậy, việc phát hiện các trò gian lận lừa đảo trong môi trường blockchain là một biện pháp bảo vệ cho nhiều nghiên cứu và phát triển hơn. Hơn nữa, việc phát hiện các cuộc tấn công lừa đảo trên thiết bị di động là một chủ đề quan trọng khác trong lĩnh vực này do sự phổ biến của điện thoại thông minh47, khiến chúng trở thành mục tiêu chung của các hành vi lừa đảo.

Tính khả dụng của dữ liệu

Tập dữ liệu được tạo trong nghiên cứu hiện tại có sẵn trong kho lưu trữ Google Drive: https://drive.google.com/file/d/18ZZHsCeMmF9HKTaL_yd41oJ_3Fgk0gWE/view?usp=sharing.

Đã nhận: ngày 17 tháng 12 năm 2021; Được chấp nhận: ngày 6 tháng 4 năm 2022
Published online: 25 May 2022

Người giới thiệu

1. RSA. Báo cáo gian lận Rsa. https://go.rsa.com/l/797543/2020-07-08/3njln/797543/48525/RSA_Fraud_Report_Q1_2020.pdf (2020) (Truy cập ngày 14 tháng 1 năm 2021).
2. APWG. Báo cáo xu hướng tấn công lừa đảo, ngày 24 tháng 11 năm 2020. https://docs.apwg.org/reports/apwg_trends_report_q3_2020.pdf (2020) (Truy cập ngày 14 tháng 1 năm 2021).
3. Aljofey, A., Jiang, Q., Qu, Q., Huang, M. & Niyigena, J.-P. Một mô hình phát hiện lừa đảo hiệu quả dựa trên mạng nơ-ron tích chập ở cấp độ ký tự từ URL. Điện tử 9, 1514 (2020).
4. Dhamija, R., Tygar, JD, & Hearst, M. Tại sao lừa đảo lại hoạt động. trong Kỷ yếu của Hội nghị SIGCHI về các yếu tố con người trong máy tính-ing Systems, Montreal, QC, Canada, 22–27 tháng 4 năm 2006, 581–590 (2006).
5. Jain, AK & Gupta, BB Một cách tiếp cận mới để bảo vệ chống lại các cuộc tấn công lừa đảo từ phía khách hàng bằng cách sử dụng danh sách trắng được cập nhật tự động. EURASIP J. trên Thông tin. Bảo vệ. 9, 1–11. https://doi.org/10.1186/s13635-016-0034-3 (2016).
6. Sahingoz, OK, Buber, E., Demir, O. & Diri, B. Phát hiện lừa đảo dựa trên máy học từ URL. Hệ thống chuyên gia ứng dụng 2019(117), 345–357 (2019).
, Asahina, H., & Sasase, I. Lược đồ phát hiện lừa đảo dựa trên sự tương đồng về hình ảnh bằng hình ảnh và CSS với trang web mục tiêu 7. Haruta, S. Người tìm kiếm. 978-1-5090-5019-2/17/$31,00 ©2017 IEEE (2017).
8. Cook, DL, Gurbani, VK, & Daniluk, M. Phishwish: Một kẻ lừa đảo phi quốc tịch sử dụng các quy tắc tối thiểu. trong Mật mã tài chính và bảo mật dữ liệu, (ed. Gene Tsudik) 324, (Berlin, Heidelberg, Springer-Verlag, 2008).
9. Jain, AK & Gupta, BB Một phương pháp tiếp cận dựa trên máy học để phát hiện lừa đảo bằng cách sử dụng thông tin siêu liên kết. J. Môi trường xung quanh. Trí tuệ. Nhân hóa. Điện toán. https://doi.org/10.1007/s12652-018-0798-z (2018).
10. Li, Y., Yang, Z., Chen, X., Yuan, H. & Liu, W. Một mô hình xếp chồng sử dụng các tính năng URL và HTML để phát hiện trang web lừa đảo. Tương lai. Tương. Điện toán. Hệ thống. 94, 27–39 (2019).
11. Xiang, G., Hong, J., Rose, CP & Cranor, L. CANTINA+: một khung máy học giàu tính năng để phát hiện các trang web lừa đảo. ACM Trans. thông tin liên lạc Hệ thống. An toàn. 14(2), 1–28. https://doi.org/10.1145/2019599.2019606 (2011).
12. Zhang, W., Jiang, Q., Chen, L. & Li, C. ELM hai giai đoạn để phát hiện các trang Web lừa đảo bằng cách sử dụng các tính năng kết hợp. Mạng toàn cầu 20(4), 797–813 (2017).

Machine Translated by Google
www.nature.com/scientificreports/

13. Rao, RS, Vaishnavi, T. & Pais, AR CatchPhish: Phát hiện các trang web lừa đảo bằng cách kiểm tra URL. J. Môi trường xung quanh. Trí tuệ. Máy tính nhân bản. 11, 813–825 (2019).
14. Arachchilage, NAG, Love, S. & Beznosov, K. Hành vi tránh mối đe dọa lừa đảo: Một cuộc điều tra thực nghiệm. Điện toán. Hùm. cư xử. 60, 185–197 (2016).
15. Wang, Y., Agrawal, R., & Choi, BY Chống lừa đảo nhẹ với việc đưa người dùng vào danh sách trắng trong trình duyệt web. tại hội nghị Khu vực 5, 2008 IEEE, IEEE, 1–4 (2008).
16. Han, W., Cao, Y., Bertino, E. & Yong, J. Sử dụng danh sách trắng cá nhân tự động để bảo vệ danh tính kỹ thuật số trên web. Hệ thống chuyên gia ứng dụng 39(15), 11861–11869 (2012).
17. Prakash, P., Kumar, M., Kompella, RR, Gupta, M. Phishnet: Đưa vào danh sách đen dự đoán để phát hiện các cuộc tấn công lừa đảo. trong INFOCOM, Kỷ yếu năm 2010 IEEE, IEEE, 1–5. https://doi.org/10.1109/INFCOM.2010.5462216 (2010)
18. Felegyhazi, M., Kreibich, C. & Paxson, V. Về khả năng chủ động đưa tên miền vào danh sách đen. LEET 10, 6–6 (2010).
19. Sheng, S., Wardman, B., Warner, G., Cranor, LF, Hong, J., & Zhang, C. Một phân tích thực nghiệm về danh sách đen lừa đảo. trong Kỷ yếu của Hội nghị lần thứ 6 về Email và Chống thư rác (CEAS'09) (2010).
20. Qi, L. và cộng sự. Dự đoán và kết hợp dữ liệu nhận biết quyền riêng tư với bối cảnh không gian-thời gian cho môi trường công nghiệp của thành phố thông minh. IEEE Dịch. Ấn Độ Thông báo. 17(6), 4159–4167. https://doi.org/10.1109/TII.2020.3012157 (2021).
21. Liu, Y. và cộng sự. Một nhãn nhiễu loạn và thiếu khung bổ sung dựa trên lý thuyết trò chơi. Cộng đồng kỹ thuật số. Mạng. https://doi.org/10.1016/j.dcan.2021.12.008 (2022).
22. Muzammal, M., Qu, Q. & Nasrulin B. Đổi mới blockchain với cơ sở dữ liệu phân tán: Một hệ thống nguồn mở. Thế hệ tương lai. Điện toán. Hệ thống. 90, 105–117. https://doi.org/10.1016/j.future.2018.07.042 (2019).
23. Liu, Y. và cộng sự. Dự đoán danh mục POI tiếp theo dựa trên mạng GRU hai chiều dành cho chăm sóc sức khỏe. quốc tế J. Trí tuệ. Hệ thống. https://doi.org/10.1002/int.22710 (2021).
24. Jain, AK & Gupta, BB Hướng tới việc phát hiện các trang web lừa đảo ở phía khách hàng bằng cách sử dụng phương pháp tiếp cận dựa trên máy học. Viễn thông-mun. Hệ thống. https://doi.org/10.1007/s11235-017-0414-0 (2017).
25. Rao, RS & Pais, AR Cơ chế lọc hai cấp độ để phát hiện các trang web lừa đảo bằng cách sử dụng phương pháp tiếp cận tương tự về mặt hình ảnh. J. Ambi-ent. Trí tuệ. Nhân hóa. Điện toán. https://doi.org/10.1007/s12652-019-01637-z (2019).
26. Jain, AK & Gupta, BB Phương pháp xác thực hai cấp để bảo vệ khỏi các cuộc tấn công lừa đảo trong thời gian thực. J. Môi trường xung quanh. Trí tuệ. Máy tính của con người. https://doi.org/10.1007/s12652-017-0616-z (2017).
27. Rao, RS, Umarekar, A. & Pais, AR Ứng dụng nhúng từ và học máy trong việc phát hiện các trang web lừa đảo. Viễn thông. Hệ thống. 79, 33–45. https://doi.org/10.1007/s11235-021-00850-6 (2022).
28. Guo, B. và cộng sự. HinPhish: Một phương pháp phát hiện lừa đảo hiệu quả dựa trên các mạng thông tin không đồng nhất. ứng dụng Khoa học. 11(20), 9733. https://doi.org/10.3390/app11209733 (2021).
29. Le, H., Pham, Q., Sahoo, D., & Hoi, SCH Urlnet: Học cách biểu diễn URL bằng học sâu để phát hiện URL độc hại. arXiv 2018, arXiv: 1802.03162 (2018).
30. Chatterjee, M., & Namin, AS Phát hiện các trang web lừa đảo thông qua học tăng cường sâu. tại Hội nghị ứng dụng và phần mềm máy tính thường niên lần thứ 43 của IEEE (COMPSAC) năm 2019. 978-1-7281-2607-4/19/$31,00 ©2019 IEEE. (Hiệp hội máy tính IEE, 2019). https://doi.org/10.1109/COMPSAC.2019.10211.
31. Xiao, X., Zhang, D., Hu, G., Jiang, Y. & Xia, S. CNN-MHSA: Mạng nơ-ron tích chập và phương pháp tiếp cận kết hợp tự chú ý nhiều đầu để phát hiện các trang web lừa đảo. Mạng lưới thần kinh. 125, 303–312. https://doi.org/10.1016/j.neunet.2020.02.013 (2020).
32. Zheng, F., Yan Q., Victor CM Leung, F. Richard Yu, Ming Z. HDP-CNN: Mạng thần kinh tích chập kim tự tháp sâu trên đường cao tốc kết hợp các biểu diễn cấp độ từ và cấp độ ký tự để phát hiện trang web lừa đảo, máy tính và bảo mật . https://doi.org/10.1016/j.cose.2021.102584 (2021)
33. Mohammad, RM, Tabtah, F. & McCluskey, L. Dự đoán các trang web lừa đảo dựa trên mạng thần kinh tự cấu trúc. Máy tính thần kinh. ứng dụng 25(2), 443–458 (2014).
34. Ramanathan, V. & Wechsler, H. Phát hiện lừa đảo và phát hiện thực thể mạo danh bằng cách sử dụng Trường ngẫu nhiên có điều kiện và Phân bổ Dirichlet tiềm ẩn. Điện toán. Bảo vệ. 34, 123–139 (2013).
35. Zhang, X., Zhao, J., & LeCun, Y. Mạng tích chập cấp độ ký tự để phân loại văn bản. trong Kỷ yếu của những tiến bộ trong Hệ thống xử lý thông tin thần kinh 28 (NIPS 2015), Montreal, QC, Canada, ngày 7–12 tháng 12 năm 2015 (2015).
36. Stecanella, B. TF-IDF là gì? https://monkeylearn.com/blog/what-is-tf-idf/. (2019) (Truy cập ngày 20 tháng 12 năm 2020).
37. Bansal, SA Hướng dẫn toàn diện để hiểu và triển khai phân loại văn bản trong python. https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-hiểu-andimplement-text-classifcation-in-python/ (2018) (Truy cập ngày 1 tháng 7 năm 2020).
38. Ramesh, G., Krishnamurthi, I. & Kumar, KSS Một phương pháp hiệu quả để phát hiện các trang web lừa đảo thông qua nhận dạng tên miền mục tiêu. Quyết định. Hệ thống hỗ trợ 2014(61), 12–22 (2014).
39. Zhang, Y., Hong, JI, & Cranor, LF Cantina: Cách tiếp cận dựa trên nội dung để phát hiện các trang web lừa đảo. trong Kỷ yếu ngày 16 Hội nghị quốc tế về World Wide Web, Banf, AB, Canada, 8–12 tháng 5 năm 2007, 639–648 (2007).
40. Chen, T., & Guestrin, C.: Xgboost: Hệ thống thúc đẩy cây có thể mở rộng. trong Kỷ yếu của Hội nghị quốc tế ACM Sigkdd lần thứ 22 về Khám phá tri thức và Khai thác dữ liệu. ACM, 785–794 (2016)
41. Aljofey, A., Jiang, Q. & Qu, Q. Một mô hình học tập có giám sát để phát hiện các hợp đồng Ponzi trong Ethereum Blockchain. Trong Dữ liệu lớn và Bảo mật. ICBDS 2021. Truyền thông trong Khoa học Thông tin và Máy tính Tập. 1563 (eds Tian, Y. và cộng sự) (Springer, 2022). https://doi.org/10.1007/978-981-19-0852-1_52.
42. http://stufgate.com/stuf/website/. (Truy cập tháng 2 năm 2020).
43. http://www.phishtank.com. (Truy cập tháng 4 năm 2020).
44. Sử dụng ngôn ngữ nội dung cho trang web. https://w3techs.com/technologists/overview/content_lingu/all. (2021) (Truy cập 19 tháng 1 năm 2021).
45. Iansiti, M. & Lakhani, KR Te sự thật về blockchain. Xe buýt Harvard. Rev. 95(1), 118–127 (2017).
46. https://github.com/YC-Coder-Chen/Tree-Math/blob/master/XGboost.md. (Truy cập tháng 9 năm 2021).
47. Qu, Q., Liu, S., Yang, B. & Jensen, CS Tìm kiếm địa phương không gian top-k hiệu quả cho các đối tượng web không gian cùng vị trí. IEEE lần thứ 15 năm 2014 Hội nghị quốc tế về quản lý dữ liệu di động. 1, 269–278 (2014).

Sự nhìn nhận

Công việc nghiên cứu này được hỗ trợ bởi Chương trình Nghiên cứu và Phát triển Trọng điểm Quốc gia của Trung Quốc. 2021YFF1200104 và 2021YFF1200100.

Sự đóng góp của tác giả

Quản lý dữ liệu, AA và QJ; Mua lại tài trợ, QJ và QQ; Điều tra, QJ và QQ; Phương pháp luận, AA và QJ; Quản lý dự án, QJ; Phần mềm, AA; Giám sát, QJ; Xác nhận, AR và HC; Viết-
bản thảo gốc, AA; Viết–đánh giá và chỉnh sửa, QJ, WL, YW và QQ; Tất cả các tác giả đã xem xét bản thảo.

Machine Translated by Google
www.nature.com/scientificreports/

Lợi ích cạnh tranh

Các tác giả tuyên bố không có lợi ích cạnh tranh.

Thông tin thêm

Thư từ và yêu cầu về tài liệu phải được gửi tới QJ

In lại và thông tin về quyền có sẵn tại www.nature.com/reprints.

Ghi chú của nhà xuất bản Springer Nature vẫn giữ thái độ trung lập đối với các tuyên bố về quyền tài phán trong các bản đồ đã xuất bản và các liên kết thể chế.

Bài viết Truy cập Mở Tis được cấp phép theo Giấy phép Quốc tế Creative Commons Ghi công 4.0, cho phép sử dụng, chia sẻ, điều chỉnh, phân phối và sao chép dưới bất kỳ phương tiện hoặc định dạng nào, miễn là bạn ghi công phù hợp cho (các) tác giả gốc và nguồn, cung cấp liên kết tới giấy phép Creative Commons và cho biết liệu các thay đổi có được thực hiện hay không. Hình ảnh hoặc tài liệu của bên thứ ba khác trong bài viết này được bao gồm trong giấy phép Creative Commons của bài viết, trừ khi có quy định khác trong hạn mức tín dụng cho tài liệu. Nếu tài liệu không có trong giấy phép Creative Commons của bài viết và mục đích sử dụng dự định của bạn không được quy định pháp luật cho phép hoặc vượt quá mức sử dụng được phép, bạn sẽ cần phải xin phép trực tiếp từ người giữ bản quyền. Để xem bản sao của giấy phép này, hãy truy cập http://creativecommons.org/licenses/by/4.0/.

© Te Tác giả 2022