TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
Tập 20, Số 6 (2023): 980-991
Website: https://journal.hcmue.edu.vn
ISSN: 2734-9918

HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
Vol. 20, No. 6 (2023): 980-991
https://doi.org/10.54607/hcmue.js.20.6.3629(2023)

Bài báo nghiên cứu

MỘT SỐ PHƯƠNG PHÁP PHÁT HIỆN TIN TỨC GIẢ MẠO TRONG NGÔN NGỮ TIẾNG VIỆT

Bùi Công Danh, Nguyễn Thị Diệu Hiền*

Khoa Công nghệ Thông tin, Đại học Công nghiệp Thực phẩm Thành phố Hồ Chí Minh, Việt Nam

**Tác giả liên hệ: Nguyễn Thị Diệu Hiền – Email:hienntd@hufi.edu.vn*

Ngày nhận bài: 11-10-2022; ngày nhận bài sửa: 21-02-2023; ngày duyệt đăng: 22-6-2023

TÓM TẮT

Với thực trạng tin giả chiếm tỉ lệ ngày càng cao trong số lượng tin tức được xuất ra hằng ngày trên Internet của nước ta hiện nay, việc đòi hỏi người đọc khi bắt gặp một tin tức cần phải nhận biết đó là tin tức có đáng tin cậy? Có nên chia sẻ và phát tán hay không? Là một việc rất khó khăn. Vì vậy, trong bài báo này chúng tôi nghiên cứu, xây dựng và đánh giá các mô hình học máy cũng như học sâu bao gồm: Naive Bayes (NB), Support Vector Machine (SVM), mạng hồi quy Long Short Term Memory (LSTM) để giải quyết bài toán phát hiện tin giả mạo trên bộ dữ liệu tiếng Việt. Kết quả của nghiên cứu rất khả quan với tỉ lệ chính xác tập kiểm thử trên 88% cho bộ dữ liệu tiếng việt VFND mở ra các hướng nghiên cứu ứng dụng nhận dạng tin tức giả mạo trong tiếng Việt.

Từ khóa: học sâu; giả mạo; tin tức giả

1. Giới thiệu

Hiện nay rất nhiều trang thông tin điện tử đưa các tin tức thời sự, hoạt động như một tờ báo điện tử, trong khi giấy phép hoạt động không phải là báo điện tử (Baomoi.com; kenh14.vn; Soha.vn; Tinnhanh24h…). Nhiều trang thông tin điện tử nội bộ, trang thông tin điện tử cá nhân, trang thông tin điện tử ứng dụng chuyên ngành (không phải xin cấp phép) lại chen chân vào tổng hợp và chuyển tải thông tin, thậm chí là thông tin thời sự như báo chí, càng làm khó người đọc (Tran & Pham, 2020).

Với sự phát triển của công nghệ thông tin, mạng Internet đã lan rộng và phủ sóng toàn cầu. Bên cạnh những lợi ích to lớn mà mạng xã hội mang lại, chúng ta đang đối mặt với nhiều nguy cơ, thách thức không nhỏ, thậm chí đe dọa đến an ninh quốc gia và trật tự an toàn xã hội. Trong đó, phải kể đến những ảnh hưởng tiêu cực từ các thông tin xấu, độc được lan truyền trên mạng xã hội cũng như vấn nạn tin giả – Fake News. Hệ lụy của việc lan truyền "tin giả" không chỉ dừng lại ở những cá nhân đơn lẻ, những nhóm người ở từng địa phương nhất định mà còn có tác động rộng lớn hơn, đe dọa trực tiếp tới an ninh quốc gia.

Cite this article as: Bui Cong Danh, & Nguyen Thi Dieu Hien (2023). Some methods to detect fake news in Vietnamese language. *Ho Chi Minh City University of Education Journal of Science, 20*(6), 980-991.

Tin giả lan tràn như con virus, dịch bệnh gây ra rất nhiều tổn thất không những với cá nhân mà với cả các tổ chức kinh tế. Như việc tung tin thất thiệt về dịch tả lợn châu Phi của một tài khoản Facebook khiến dư luận hoang mang, tẩy chay thịt lợn, ảnh hưởng tới chăn nuôi của người nông dân, khiến nhiều người hoảng loạn, mất phương hướng trong cuộc sống, hay mất lòng tin vào những chỉ đạo của các cơ quan quản lí nhà nước. Ngoài những thiệt hại về kinh tế, một trong những hậu quả nghiêm trọng nhất mà tin giả gây ra là làm suy giảm niềm tin của công chúng vào truyền thông nói chung và báo chí chủ lưu nói riêng. Tin giả khiến công chúng không xác định được đâu là những nguồn tin đáng tin cậy để tiếp nhận, luôn ở trạng thái ngờ vực, tham khảo cả những nguồn tin không chính thống dẫn đến bị nhiễu loạn.

Việc đưa Internet tiếp cận người dân trong thập niên 90 là một hành động nhằm cho phép họ truy cập thông tin. Theo thời gian, Internet đã phát triển đến một tầm cao không thể tưởng tượng được với hàng tấn thông tin xuất hiện mọi lúc cho phép Internet trở thành máy chủ lưu trữ nhiều thông tin đến mức những thông tin không được mong đợi, không trung thực và sai lệch có thể được tạo ra bởi bất kì ai.

Tin giả và các loại thông tin sai lệch khác có thể xuất hiện dưới nhiều dạng. Chúng cũng có thể có tác động lớn, bởi vì thông tin định hình thế giới quan và cách suy nghĩ của cá nhân, hơn nữa, việc ra quyết định cũng được dựa vào thông tin. Vì vậy, nếu thông tin được thấy trên Internet được tạo ra, phóng đại hoặc bị bóp méo, điều này có thể ảnh hưởng đến việc ra một quyết định đúng đắn. Tin giả có thể ảnh hưởng xấu đến các yếu tố tài chính, sức khỏe, nỗi sợ. Thậm chí tin giả cũng có thể tạo ra các định kiến về chủng tộc, hoặc dẫn đến hành vi bạo lực trên các kênh trực tuyến. Cuối cùng, điều tồi tệ là các quan điểm của độc giả có thể bị ảnh hưởng bởi tin giả, họ sẽ trở nên đa nghi hơn, và điều này làm mất niềm tin của họ vào truyền thông.

Phát hiện tin tức giả là một nhiệm vụ quan trọng và cũng là một thách thức lớn. Các nhà nghiên cứu đang cố gắng tìm ra phương pháp học máy tốt nhất để phát hiện tin tức giả (Kurasinski, 2020). Độ chính xác của phương pháp phụ thuộc vào công tác đào tạo của phương pháp này, một mô hình được đào tạo một cách tốt có thể cho độ chính xác cao hơn.

Rubin và các cộng sự (Rubin et al., 2016) đã đề xuất một mô hình để xác định các bài báo tin tức châm biến và hài hước. Họ kiểm tra và xem xét 360 bài báo châm biếm chủ yếu trong bốn lĩnh vực, bao gồm công dân, khoa học, kinh doanh và những gì họ gọi là "soft news" ("giải trí/tin đồn bài viết"). Họ đề xuất một mô hình phân loại SVM sử dụng chủ yếu năm đặc trưng được phát triển dựa trên phân tích của họ về tin tức châm biếm. Năm đặc trưng là vô lí, hài hước, ngữ pháp, ảnh hưởng tiêu cực và dấu chấm câu. Độ chính xác cao nhất của họ là 90% đã đạt được bằng cách sử dụng chỉ có ba kết hợp các tính năng là vô lí, ngữ pháp, và dấu chấm câu.

Aphiwongsophon và Chongstitvatana (2018), cho biết rằng mạng xã hội tạo ra một số lượng lớn các bài đăng. Bất kì ai có thể đăng kí trên các nền tảng này đều có thể đăng bài họ

muốn. Bài đăng này có thể chứa thông tin sai lệch chống lại một cá nhân hoặc tổ chức kinh doanh. Tác giả đã sử dụng ba phương pháp học máy để phân loại tin tức giả, Naive Bayes, mạng nơ-ron và SVM. Độ chính xác do Naive Bayes cung cấp là 96,08%. Mặt khác, hai phương pháp khác là mạng nơ-ron và SVM cho độ chính xác là 99,90%.

Dựa vào các công trình nghiên cứu trên, bài báo tập trung phát triển nghiên cứu các thuật toán học máy giải quyết bài toán phát hiện tin giả trên bộ dữ liệu tiếng Việt. Từ đó đưa ra các đánh giá, nhận xét một cách tổng quan nhất.

2. Nội dung

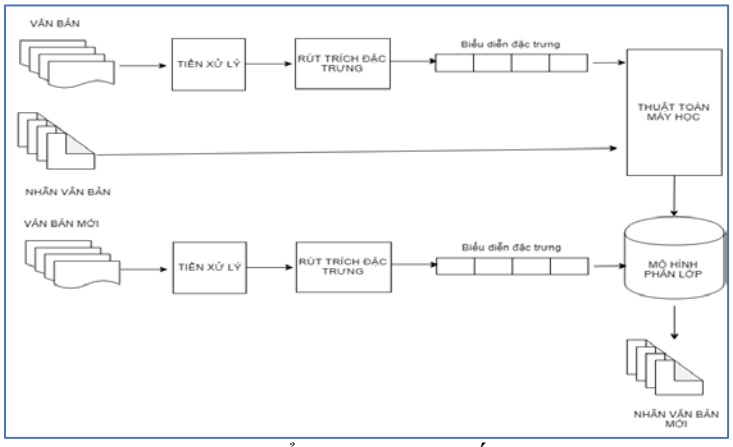
2.1. Khó khăn trong ngăn chặn tin tức giả mạo

Tin tức giả thường được phát tán rất nhanh, nhanh hơn gấp nhiều lần so với khả năng ngăn chặn và xử lí chúng. Với sự phát triển của công nghệ, người ta có thể dễ dàng tạo lập một website, một trang blog hay tài khoản hoặc fanpage trên các mạng xã hội với chi phí gần như bằng không. Đây chính là những công cụ hỗ trợ đắc lực cho việc phát tán tin tức giả. Do vậy, dù là vô tình hay cố ý, lực lượng tạo ra và phát tán tin tức giả có thể là bất cứ thành phần nào trong xã hội: từ cá nhân, tổ chức và thậm chí là có cả một ngành công nghiệp sản xuất tin tức giả ở một nơi như thị trấn Veles, thuộc Macedonia, nơi được xem như là cái nôi của ngành công nghiệp tin tức giả ăn theo chiến dịch tranh cử Mĩ. Cũng nhờ công nghệ tiên tiến, các đối tượng sản xuất tin giả có thể tìm ra những cách phát tán tin tức giả một cách nhanh chóng đến mức khó kiểm soát. Cách thức tạo ra và phát tán tin tức giả từ nghiệp dư đến chuyên nghiệp đều góp phần làm số lượng tin tức giả được phát tán trên trực tuyến là vô cùng lớn so với khả năng phát hiện và ngăn chặn chúng của các lực lượng chức năng liên quan. (Della Vedova et al., 2018)

Theo thống kê của Smartinsights.com, mỗi phút có khoảng 360 nghìn người dùng đăng kí mới trên Facebook, 150 nghìn tin nhắn được trao đổi, 300 nghìn status được cập nhật, 50 nghìn link được chia sẻ, 133.300 ảnh được đăng tải và 100 nghìn đề nghị kết bạn mới. Trong khi đó, trên Youtube, mỗi phút có hơn 400 giờ nội dung được đăng tải. Còn theo Google, công cụ tìm kiếm này nhận thấy số lượng tìm kiếm đã đạt đến hàng nghìn tỉ mỗi năm, trong đó 15% lượng tìm kiếm mỗi ngày có nội dung hoàn toàn mới. Ngày 01/11/2017, Facebook thừa nhận có tới 270 triệu tài khoản trên mạng xã hội này là không hợp pháp. Với số lượng tài khoản không hợp pháp và nội dung đăng tải lớn như trên, việc phát hiện và ngăn chặn những nội dung không đúng sự thật trên các nền tảng mạng xã hội là vô cùng khó. Tin tức giả chỗ này chặn chưa xong thì tin tức giả ở nhiều chỗ khác đã mọc lên như nấm sau mưa. (Shu et al., 2017)

2.2. Mô hình phân loại tin tức giả mạo

2.2.1. Đề xuất mô hình phân loại



Hình 1. *Mô hình tổng quan hệ thống phân loại tin*

Văn bản đầu vào: là các bài báo, bài post về thông tin thật hoặc giả.

Nhãn văn bản: Nhãn 0 hoặc 1 tương ứng với giá trị tin thật hoặc tin giả.

Tiền xử lí: Bao gồm các bước tiền xử lí trong văn bản như tách từ, loại bỏ các kí tự thừa, xoá khoảng trắng, xoá các từ stopword.

Rút trích đặc trưng: Là bước lựa chọn các đặc trưng mang nhiều thông tin của mỗi loại nhãn khác nhau để phân biệt giữa các nhãn trong dữ liệu huấn luyện. Việc rút trích đặc trưng có vai trò quan trọng ảnh hưởng đến chất lượng của bộ phân lớp.

Biểu diễn đặc trưng: Sau khi rút trích các đặc trưng ở văn bản đầu vào thì các đặc trưng này cần phải được biểu diễn ở dưới dạng văn bản số trước khi đưa vào các thuật toán học máy. Ở đây chúng ta có thể dùng các kĩ thuật như TFIDF, Countvector, Doc2vec.

Mô hình huấn luyện: Các thuật toán trong mô hình này của bài báo chúng tôi áp dụng mô hình các thuật toán như: SVM, Naive Bayes, LSTM.

Mô hình phân lớp: Sau khi các thuật toán như SVM, Naive Bayes, LSTM huấn luyện xong trên tập huấn luyện thì sẽ tạo thành mô hình phân lớp. Mô hình này sẽ nhận đầu vào là đặc trưng biểu diễn văn bản và dự đoán kết quả.

Nhãn văn bản mới: Kết quả trả về của mô hình phân lớp khi đưa dữ liệu huấn luyện vào.

2.2.2. BAYES

Định lí Bayes cho phép tính xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra. Xác suất này được kí hiệu là P(A|B), và đọc là "xác suất của A nếu có B". Đại lượng này được gọi xác suất có điều kiện hay xác suất hậu nghiệm vì nó được rút ra từ giá trị được cho của B hoặc phụ thuộc vào giá trị đó. (Julio et al., 2019)

Theo định lí Bayes, xác suất xảy ra A khi biết B sẽ phụ thuộc vào 3 yếu tố:

✓ Xác suất xảy ra A của riêng nó, không quan tâm đến B. Kí hiệu là P(A) và đọc là xác suất của A. Đây được gọi là xác suất biên duyên hay xác suất tiên nghiệm, nó là "tiên nghiệm"

theo nghĩa rằng nó không quan tâm đến bất kì thông tin nào về B.

✓ Xác suất xảy ra B của riêng nó, không quan tâm đến A. Kí hiệu là P(B) và đọc là "xác suất của B". Đại lượng này còn gọi là hằng số chuẩn hóa (normalising constant), vì nó luôn giống nhau, không phụ thuộc vào sự kiện A đang muốn biết.

✓ Xác suất xảy ra B khi biết A xảy ra. Kí hiệu là P(B|A) và đọc là "xác suất của B nếu có A". Đại lượng này gọi là khả năng (likelihood) xảy ra B khi biết A đã xảy ra. Chú ý không nhầm lẫn giữa khả năng xảy ra B khi biết A và xác suất xảy ra A khi biết B.

Các mô hình xác suất Bayes

✓ Một mô hình phân lớp là một mô hình Machine Learning dùng để phân loại các vật mẫu dựa trên các đặc tính đã xác định.

Naive Bayes là một thuật toán phân lớp được mô hình hoá dựa trên định lí Bayes trong xác suất thống kê: $P(y|X) = \frac{\mathrm{P(X|y)P(y)}}{P(X)}$ (1)

Trong đó:

+ P(y|X) gọi là posterior probability: xác suất của mục tiêu y với điều kiện có đặc trưng X.

+ P(X|y) gọi là likelihood: xác suất của đặc trưng X khi đã biết mục tiêu y.

+ P(y) gọi là prior probability của mục tiêu y

+ P(X) gọi là prior probability của đặc trưng X

Ở đây, X là vector các đặc trưng, có thể viết dưới dạng:

$\mathrm{X} = (x_1, x_2, x_3, \dots, x_n)$ (2)

Khi đó, đẳng thức Bayes trở thành:

$\mathrm{P(y}|x_1, \dots, x_n) = \frac{\mathrm{P}(x_1|\mathrm{y})\mathrm{P}(x_2|\mathrm{y}) \dots \mathrm{P}(x_n|\mathrm{y})P(y)}{P(x_1)P(x_2)\dots P(x_n)}$ (3)

✓ Trong mô hình Naive Bayes, có hai giả thiết được đặt ra:

+ Các đặc trưng đưa vào mô hình là độc lập với nhau. Tức là sự thay đổi giá trị của một đặc trưng không ảnh hưởng đến các đặc trưng còn lại.

+ Các đặc trưng đưa vào mô hình có ảnh hưởng ngang nhau đối với đầu ra mục tiêu.

Khi đó, kết quả mục tiêu y để P(y|X) đạt cực đại trở thành:

$\mathrm{y} = \mathrm{argmax}_y\ P(y)\pi_{i=1}^{n} P(x_i|y)$ (4)

Chính vì hai giả thiết gần như không tồn tại trong thực tế trên, mô hình này mới được gọi là naive (ngây thơ). Tuy nhiên, chính sự đơn giản của nó với việc dự đoán rất nhanh kết quả đầu ra khiến nó được sử dụng rất nhiều trong thực tế trên những bộ dữ liệu lớn, đem lại kết quả khả quan. Một vài ứng dụng của Naive Bayes có thể kể đến như: lọc thư rác, phân loại văn bản, dự đoán sắc thái văn bản,...

Thuật toán naive bayes

Dữ kiện cần có:

D: tập dữ liệu huấn luyện, được vector hoá dưới dạng $\vec{x}$ = ($x1$, $x2$, … , xn).

Ci : tập các tài liệu của D thuộc lớp Ci với i={1,2,3,…}.

Các $x1$, $x2$, … , xn độc lập xác suất đôi một với nhau.

Thuật toán Naïve Bayes cơ bản:

✓ Bước 1: Huấn luyện Naïve Bayes (dựa vào tập dữ liệu)

+ Tính xác suất P(C_i).

+ Tính xác suất P(x_k)|C_i).

✓ Bước 2: Phân lớp Xnew

+ Tính (Xne , C_i) = $P(C_i)$Π P($x_k|C_i$)

+ Xnew được gán vào lớp Cq sao cho (Xne , C_q) = max($F(Xnew$, C_i))

(x_i |) được tính như sau:

$$P(x_k|C_i) = \frac{C_{i,D}\{x_k\}}{|C_{i,D}|} \quad (5)$$

Trong đó:

C_i , số mẫu của tập dữ liệu huấn luyện D thuộc về lớp C_i.

C_i, { } số mẫu trong tập C_i, mà có nhân giá trị là x_k.

2.3.1. SVM

✓ Giới thiệu

Bài toán phân lớp (Classification) và dự đoán (Prediction) là hai bài toán cơ bản và có rất nhiều ứng dụng trong tất cả các lĩnh vực như: học máy, nhận dạng, trí tuệ nhân tạo… Trong bài báo này, chúng tôi sẽ đi sâu nghiên cứu phương pháp Support Vector Machines (SVM), một phương pháp rất hiệu quả hiện nay.

Phương pháp SVM được coi là công cụ mạnh cho những bài toán phân lớp phi tuyến tính được các tác giả Vapnik và Chervonenkis phát triển mạnh mẽ năm 1995.

Phương pháp này thực hiện phân lớp dựa trên nguyên lí cực tiểu hóa rủi ro có cấu trúc SRM (Structural Risk Minimization), được xem là một trong các phương pháp phân lớp giám sát không tham số tinh vi nhất cho đến nay. Các hàm công cụ đa dạng của SVM cho phép tạo không gian chuyên đổi để xây dựng mặt phẳng phân lớp.

✓ Định nghĩa

Là phương pháp dựa trên nền tảng của lí thuyết thống kê nên có một nền tảng toán học chặt chẽ để đảm bảo rằng kết quả tìm được là chính xác.

Là thuật toán học giám sát (supervied learning) được sử dụng cho phân lớp dữ liệu.

Là 1 phương pháp thử nghiệm, đưa ra một trong những phương pháp mạnh và chính xác nhất trong số các thuật toán nổi tiếng về phân lớp dữ liệu.

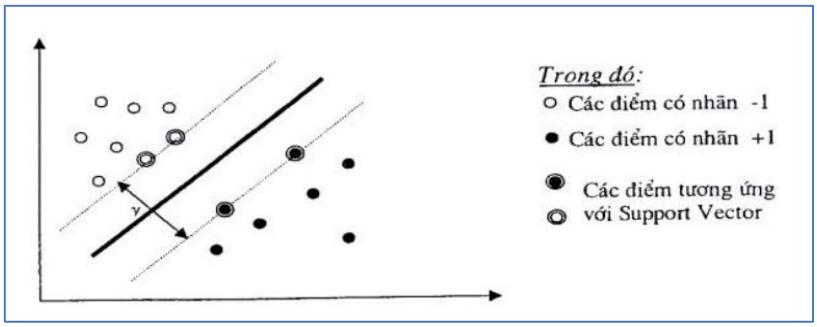
SVM là một phương pháp có tính tổng quát cao nên có thể được áp dụng cho nhiều loại bài toán nhận dạng và phân loại.

✓ Bài toán phân 2 lớp với SVM

Bài toán đặt ra là: Xác định hàm phân lớp để phân lớp các mẫu trong tương lai, nghĩa là với một mẫu dữ liệu mới xi thì cần phải xác định xi được phân vào lớp +1 hay lớp -1.

Để xác định hàm phân lớp dựa trên phương pháp SVM, ta sẽ tiến hành tìm hai siêu phẳng song song sao cho khoảng cách y giữa chúng là lớn nhất có thể để phân tách hai lớp

này ra làm hai phía. Hàm phân tách tương ứng với phương trình siêu phẳng nằm giữa hai siêu phẳng tìm được.



Hình 2. *Minh họa bài toán 2 phân lớp bằng phương pháp SVM*

Các điểm mà nằm trên hai siêu phẳng phân tách được gọi là các Support Vector. Các điểm này sẽ quyết định đến hàm phân tách dữ liệu.

✓ Các bước chính của phương pháp SVM

Phương pháp SVM yêu cầu dữ liệu được diễn tả như các vector của các số thực. Như vậy nếu đầu vào chưa phải là số thì ta cần phải tìm cách chuyển chúng về dạng số của SVM Tiền xử lí dữ liệu: Thực hiện biến đổi dữ liệu phù hợp cho quá trình tính toán, tránh các số quá lớn mô tả các thuộc tính. Thường nên co giãn (scaling) dữ liệu để chuyển về đoạn [-1, 1] hoặc [0, 1].

Chọn hàm hạt nhân: Lựa chọn hàm hạt nhân phù hợp tương ứng cho từng bài toán cụ thể để đạt được độ chính xác cao trong quá trình phân lớp. Thực hiện việc kiểm tra chéo để xác định các tham số cho ứng dụng.

Điều này cũng quyết định đến tính chính xác của quá trình phân lớp. Sử dụng các tham số cho việc huấn luyện với tập mẫu.

Trong quá trình huấn luyện sẽ sử dụng thuật toán tối ưu hóa khoảng cách giữa các siêu phẳng trong quá trình phân lớp, xác định hàm phân lớp trong không gian đặc trưng nhờ việc ánh xạ dữ liệu vào không gian đặc trưng bằng cách mô tả hạt nhân, giải quyết cho cả hai trường hợp dữ liệu là phân tách và không phân tách tuyến tính trong không gian đặc trưng.

Kiểm thử tập dữ liệu Test.

2.2.4. Mô hình học sâu

Deep learning là một nhánh của lĩnh vực học máy, dựa trên một tập hợp các thuật toán nhằm cố gắng mô hình dữ liệu trừu tượng hóa ở mức cao bằng cách sử dụng nhiều lớp xử lí với cấu trúc phức tạp, hoặc bao gồm nhiều biến đổi phi tuyến.

Deep learning là một lớp của các thuật toán máy học:

✓ Sử dụng một tầng (cascade) nhiều lớp các đơn vị xử lí phi tuyến để trích trọn đặc trưng và chuyển đổi. Mỗi lớp kế tiếp dùng đầu ra từ lớp trước làm đầu vào. Thuật toán này có thể được giám sát hoặc không cần giám sát và được ứng dụng cho các mô hình phân tích (không có giám sát) và phân loại (giám sát).

✓ Dựa trên học (không có giám sát) của nhiều cấp các đặc trưng hoặc đại diện của dữ

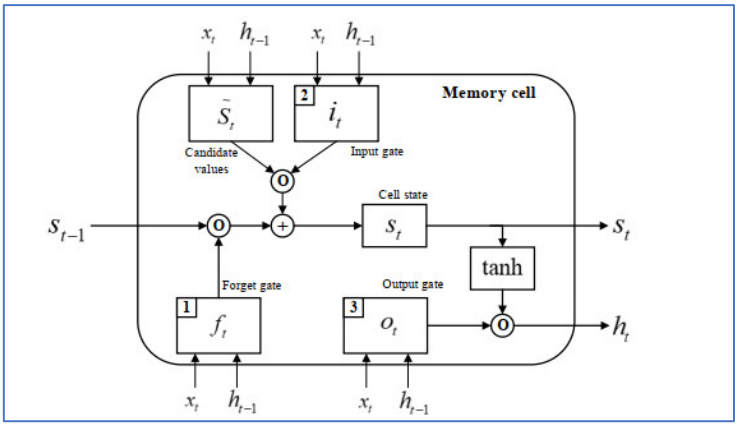
liệu. Các tính năng cao cấp bắt nguồn từ các tính năng thấp cấp hơn để tạo thành một đại diện thứ bậc.

- ✓ Là một phần của lĩnh vực máy học và rộng lớn hơn về việc học đại diện dữ liệu.
- ✓ Học nhiều cấp độ đại diện tương ứng với các mức độ trừu tượng khác nhau; các mức độ hình thành một hệ thống phân cấp của các khái niệm.

Deep learning còn là phương pháp nâng cao của mạng nơ-ron nhân tạo (Artificial Neural Networks) khai thác khả năng tính toán ngày càng rẻ từ các chip xử lí hiện đại. Phương pháp này nhắm tới việc xây dựng nhiều hơn các mạng nơ- ron phức tạp cũng như giải quyết bài toán semi-supervised do tập dữ liệu khổng lồ thường được gán nhãn không đầy đủ.

2.2.5. LSTM

LSTM là một phiên bản mở rộng của mạng Recurrent Neural Network (RNN), nó được thiết kế để giải quyết các bài toán về phụ thuộc xa (long-term dependencies). RNN là mạng nơ-ron có chứa vòng lặp. Mạng này có khả năng lưu trữ thông tin, thông tin được truyền từ lớp này sang lớp khác. Đầu ra của lớp ẩn phụ thuộc vào thông tin của các lớp tại mọi thời điểm. RNN đã được sử dụng phổ biến trong xử lí ngôn ngữ tự nhiên hay các bài toán có dữ liệu tuần tự. Tuy nhiên, do kiến trúc của RNN khá đơn giản nên khả năng liên kết các lớp có khoảng cách xa là không tốt. Nó cơ bản không có khả năng ghi nhớ thông tin từ các dữ liệu có khoảng cách xa, và do đó, những phần tử đầu tiên trong chuỗi đầu vào thường không có nhiều ảnh hưởng đến kết quả dự đoán phần tử cho chuỗi đầu ra các bước sau. Nguyên nhân của việc này là do RNN chịu ảnh hưởng bởi việc đạo hàm bị thấp dần trong quá trình học – biến mất đạo hàm (vanishing gradient). Mạng LSTM được thiết kế để khắc phục vấn đề này. Cơ chế hoạt động của LSTM là chỉ ghi nhớ những thông tin liên quan, quan trọng cho việc dự đoán, còn các thông tin khác sẽ được bỏ đi.



Hình 3. *Mô hình LSTM*

Mạng LSTM có thể bao gồm nhiều tế bào LSTM liên kết với nhau. Ý tưởng của LSTM là bổ sung thêm trạng thái bên trong tế bào (cell internal state) st và ba cổng sàng lọc thông tin đầu vào và đầu ra cho tế bào bao gồm cổng quên ft, cổng đầu vào it và cổng đầu ra ot.

Tại mỗi bước thời gian t, các cổng lần lượt nhận giá trị đầu vào xt (đại diện cho một phần tử trong chuỗi đầu vào) và giá trị ht-1 có được từ đầu ra của các ô nhớ từ bước thời gian trước đó *t – 1*. Các cổng đều có chức năng sàng lọc thông tin với mỗi mục đích khác nhau. Các cổng được định nghĩa như sau:

Cổng quên: Có chức năng loại bỏ những thông tin không cần thiết nhận được khỏi trạng thái tế bào bên trong.

Cổng đầu vào: Giúp sàng lọc những thông tin cần thiết để được thêm vào trạng thái tế bào bên trong.

Cổng đầu ra: Có chức năng xác định những thông tin nào từ các trạng thái tế bào bên trong được sử dụng như đầu ra.

Trong quá trình thực hiện, st và các giá trị đầu ra ht được tính toán như sau:

Ở bước đầu tiên, tế bào LSTM quyết định những thông tin cần được loại bỏ từ các trạng thái tế bào bên trong ở bước thời gian trước đó st-1. Giá trị ft của cổng quên tại bước thời gian t được tính toán dựa trên giá trị đầu vào hiện tại xt, giá trị đầu ra ht-1 từ tế bào LSTM ở bước trước đó và độ lệch (bias) bf của cổng quên. Hàm sigmoid biến đổi tất cả các giá trị kích hoạt (activation value) về miền giá trị trong khoảng từ 0 và 1 theo công thức:

ft = (Wf, + Wf,hht-1 + bf) (6)

Ở bước thứ 2, tế bào LSTM xác định những thông tin nào cần được thêm vào các trạng thái tế bào bên trong st. Bước này bao gồm hai quá trình tính toán đối với st và ft. st biểu diễn những thông tin có thể được thêm vào các trạng thái tế bào bên trong:

st = $tanh$(Ws,xxt + Ws,hht-1 + bs) (7)

Giá trị của cổng đầu vào tại bước thời gian t được tính:

it = (Wi, + Wi,hht-1 + bi) (8)

Ở bước tiếp theo, giá trị mới của trạng thái tế bào bên trong st được tính toán dựa trên kết quả thu được từ các bước trên:

st = ft ∗ st-1 + it ∗ st (9)

Cuối cùng, giá trị đầu ra ht:

ot = (Wo, + Wo,hht-1 + bo) (10)

ht = ot ∗ $tanh$(st) (11)

Trong đó: Ws,h, Wf,x, Wf,h, Wi,x, Wi,h là các ma trận trọng số trong mỗi tế bào LSTM. bs, bi, bo là các vector bias.

2.3. Kết quả thực nghiệm

2.3.1. Môi trường và cơ sở dữ liệu thực nghiệm

- ✓ Môi trường cài đặt

  + Ứng dụng minh họa cho thuật toán được thực hiện trên phần mềm Jupyter notebook.

  + Ứng dụng được viết bằng ngôn ngữ lập trình Python.

- ✓ Dữ liệu thực nghiệm

Trong bài báo này chúng tôi sử dụng 2 bộ dữ liệu thực nghiệm bao gồm:

+ Vietnamese fake news dataset – VFND

VFND là bộ dataset về các tin tức giả bằng ngôn ngữ tiếng Việt được tập hợp trong khoảng thời gian từ 2017 đến 2019 của Hồ Quang Thanh (Ho Quang Thanh and ninh-pm-se, 2019).

VFND có 226 tin tức.

Cấu trúc tên của 1 file bao gồm: VFND_{Source}_{Label}_{Number}.json.

Trong đó

+ {Source}:

Ac - nguồn bài báo từ các trang tin tức;

So - Nguồn từ các bài viết của người dùng trên mạng xã hội Facebook, Twitter, YouTube... có tính chất như nguồn tin tức.

+ {Label} thuộc tập {"Fake", "Real"}

Giới hạn các chủ đề tin tức trong tập dữ liệu:

Các tin tức sử dụng trong bộ dataset đều là tin tức tường thuật về 1 sự kiện. Lí do: Để có thể kiểm tra chéo giữa các nguồn tin để xác định được tin tức thật hoặc giả trong trường hợp mà cộng đồng chưa hỗ trợ phân loại tin tức.

Các chủ đề mà bộ dataset tập trung là: Thể thao, Văn hóa, Xã hội, Kinh tế, Pháp luật, Y tế... Các tin tức sẽ được kiểm tra chéo về nguồn gốc, nội dung, sự kiện để xác định thật và giả.

Một số chú ý: Một số tin tức sẽ được nhóm mặc định là tin giả: tin tức có tính chất mê tín, dị đoan; tin tức không xác nhận được nguồn tin; tin tức dựa trên những nguồn tin, kiến thức, học thuyết, luận thuyết được khoa học công nhận là sai lầm.

+ Fake or real news (FRN):

Đây là tập dữ liệu phân loại tin thật/giả bằng tiếng anh được lấy từ trang https://www.kaggle.com/.

Bộ dữ liệu có 6335 tin tỉ lệ thật/giả là 3171/3164.

Cột text lưu trữ tin tức, cột label chứa nhãn tin gồm 2 loại: 0: tin giả, 1: Tin thật.

✓ Cách đánh giá

Cách đơn giản và hay được sử dụng nhất là accuracy (độ chính xác). Cách đánh giá này đơn giản tính tỉ lệ giữa số mẫu dự đoán đúng và tổng số mẫu trong tập dữ liệu. Công thức:

$$\text{Accuracy} = \frac{TP+TN}{\text{Số lượng mẫu}} = \frac{TP+TN}{TP+TN+FP+FN} \quad (12)$$

Giả sử độ accuracy = 90% có nghĩa là trong số 100 mẫu thì có 90 mẫu được phân loại chính xác. Tuy nhiên đối với tập dữ liệu kiểm thử không cân bằng (nghĩa là số positive lớn hơn rất nhiêu so với negative) thì đánh giá có thể gây hiểm nhầm.

2.3.2. Kết quả thực nghiệm

Trong quá trình thực nghiệm chúng tôi chia tỉ lệ là 80% cho tập train và 20% cho tập test kết quả được ghi nhận như sau:

Tập dữ liệu VFND

Tổng số tin: 226, Số tin train: 181, Số tin test: 45.

Bảng 1. *Kết quả tập dữ liệu VFND*

	Accuracy Train	Accuracy Test
Bayes	0.977	0.88
SVM	1.0	0.86
LSTM	0.5419	0.565

Fake or real news (**FRN**)

Sau tiền xử lí tổng số tin: 3625, Số tin train: 5060, Số tin test: 1265.

Bảng 2. *Kết quả tập dữ liệu Fake or real news*

	Accuracy Train	Accuracy Test
Bayes	0.92	0.84
SVM	0.92	0.89
LSTM	0.5434	0.5

2.3.3. Đánh giá các mô hình

Từ Bảng 1 và 2 cho ta thấy tỉ lệ huấn luyện và kiểm thử của 2 thuật toán Bayes và SVM đạt tỉ lệ tốt hơn LSTM trên cả bộ dữ liệu nhị phân VFND và FRN. Tỉ lệ chênh lệch giữa 2 thuật toán Bayes và SVM là không đáng kể. Như vậy, trong tình huống phân loại tin tức dựa trên phân lớp nhị phân thì thuật toán SVM cho kết quả tốt nhất trên cả 02 bộ dữ liệu.

3. Kết luận, kiến nghị

Trong bài báo này, chúng tôi đã trình bày phương pháp sử dụng kĩ thuật học SVM, mô hình Bayes và LSTM để tiến hành xác định tin thật giả. Trong đó cả 2 mô hình SVM và Bayes đã đạt được những kết quả nhất định và có độ chính xác cao hơn LSTM trên cả 02 bộ dữ liệu thực nghiệm VFND và FRN.

Kết quả thực nghiệm cho chúng ta thấy sự hiệu quả của thuật tóa SVM trên bộ dữ liệu FRN với độ chính xác tập kiểm thử là trên 89%. Trong khi đó Bayes hiệu quả kiểm thử trên 88% trên bộ dữ liệu VFND.

Chúng tôi đã xây dựng được mô hình thuật toán nhằm phát hiện tin giả mạo bằng phương pháp học.

Kiến nghị trong các hướng nghiên cứu tiếp theo là xây dựng các bộ dữ liệu chuẩn tiếng Việt cho bài toán phát hiện tin thật giả và nghiên cứu kết hợp giữa hình ảnh video với nội dung văn bản.

❖ ***Tuyên bố về quyền lợi:*** *Các tác giả xác nhận hoàn toàn không có xung đột về quyền lợi.*

❖ ***Lời cảm ơn:*** *Cảm ơn Trường Đại học Công nghiệp Thực phẩm Thành phố Hồ Chí Minh đã hỗ trợ chúng tôi trong quá trình thực nghiên cứu này.*

TÀI LIỆU THAM KHẢO

Aphiwongsophon, S., & Chongstitvatana, P. (2018). Detecting Fake News with Machine Learning Method. *International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, (pp. 528-531).

Della Vedova, M. L., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M., & de Alfaro, L. (2018). Automatic online fake news detection combining content and social signals. *Proceedings of the 22st Conference of Open Innovations Association FRUCT*, (pp. 272-279).

Ho Quang Thanh and ninh-pm-se. (2019). A collection of Vietnamese articles and Facebook posts categorizing 2 labels Real & Fake [*Tap hop cac bai bao tieng Viet và cac bai post Facebook phan loai 2 nhan That & Gia (228 bai)*]. Zenodo.

Julio C. S. Reis, André Correia, Fabrício Murai, Adriano Veloso, & Fabrício Benevenuto (2019). Supervised Learning for Fake News Detection. *IEEE Intelligent Systems*, 76-81.

Kurasinski, L. (2020). *Machine Learning explainability in text classification for Fake News detection.* Malmö University Publications.

Rubin, V., Conroy, N., Chen, Y., & Cornwell, S. (2016). Fake news or truth? Using satirical cues to detect potentially misleading news. *Proceedings of the second workshop on computational approaches to deception detection.*

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 22-36.

Tran, T. Y. M., & Pham T. H. (2020). Some skills to verify information for internet and social network users [Mot so ki nang tham dinh thong tin cho nguoi dung internet va mang xa hoi]. *UED Journal of Social Sciences, Humanities & Education, 10*(2020), 112-119.

SOME METHODS TO DETECT FAKE NEWS IN VIETNAMESE LANGUAGE

Bui Cong Danh, Nguyen Thi Dieu Hien*

Falcuty of Information Technology, Ho Chi Minh City University of Food Industry, Vietnam

**Corresponding author: Nguyen Thi Dieu Hien – Email: hienntd@hufi.edu.vn*

Received: October 11, 2022; Revised: February 21, 2023; Accepted: June 22, 2023

ABSTRACT

With an increase of fake news published daily on the Internet in Vietnam, Readers, when readinga news story, need to ask themselves whether the news is reliable,Should I share and spread it? It is difficult to decide. Therefore, in this paper, we study, build, and evaluate machine learning and deep learning models including Naïve Bayes (NB), Support Vector Machine (SVM), and Long Short Term Memory (LSTM) regression networks to detect fake news on Vietnamese datasets. The results show that these models are able to detect fake news with an accuracy rate of more than 88% for the Vietnamese VFND dataset. This study could open up future research directions to detect fake news in Vietnamese.

Keywords: deep learning; fake; fake news