ỦY BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC SÀI GÒN
KHOA CÔNG NGHỆ THÔNG TIN

Tiểu luận học phần Đồ án chuyên ngành

NGHIÊN CỨU VỀ THUẬT TOÁN SUPPORT VECTOR MACHINE (SVM) VÀ ÁP DỤNG GIẢI BÀI TOÁN NHẬN DIỆN TIN GIẢ

Sinh viên: Nguyễn Đình Hoàng Khang
MSSV: 3120411074
GV Hướng dẫn: TS. Phan Tấn Quốc

THÀNH PHỐ HỒ CHÍ MINH, THÁNG 5 NĂM 2024

LỜI CAM ĐOAN

Tôi tên Nguyễn Đình Hoàng Khang, cam đoan rằng: Đề tài tiểu luận học phần Đồ án chuyên ngành "Nghiên cứu về thuật toán Support Vector Machines (SVM) và áp dụng giải bài toán nhận diện tin giả" là công trình nghiên cứu của bản thân dưới sự hướng dẫn của TS. Phan Tấn Quốc. Những kết quả nghiên cứu của các tác giả khác và số liệu được sử dụng trong tiểu luận đều được trích dẫn đầy đủ. Các kết quả và số liệu được trình bày trong tiểu luận là hoàn toàn trung thực và chưa được công bố trong bất cứ công trình nghiên cứu nào khác.

MỤC LỤC

DANH MỤC CÁC CHỮ VIẾT TẮT

Từ viết tắt	Từ tiếng anh	Từ tiếng việt
RNN	Recurrent Neural Network	Mạng thần kinh hồi quy
LSTM	Long Short Term Memory	Bộ nhớ dài-ngắn hạn
CNN	Convolutional Neural Networks	Mạng thần kinh tích chập
SVM	Support Vector Machine	Máy vectơ hỗ trợ
KNN	K-Nearest Neighbors	K hàng xóm gần nhất
RF	Random Forest	Rừng quyết định ngẫu nhiên
DT	Decision Tree	Cây quyết định
LR	Logistic Regression	Hồi quy logistic
RBF	Radial Basis Function	Mạng hàm cơ sở bán kính
NB	Naïve Bayes	
GBDT	Gradient Boosted Decision Trees	

DANH MỤC CÁC HÌNH ẢNH

LỜI MỞ ĐẦU

1. Lý do chọn đề tài

Vì sự phát triển ngày càng mạnh mẽ của công nghệ số, các trang mạng xã hội, truyền thông tin tức càng dễ tiếp cận tới nhiều người dùng hơn, thì việc này vừa có lợi cũng vừa có hại. Mặt lợi, việc tìm kiếm tin tức trên mạng xã hội dễ tiếp cận hơn, nhanh hơn và không tốn nhiều phí. Mặt hại, tạo điều kiện cho những tin tức không chính xác /tin giả, tức là tin có thông tin sai lệch được lan truyền nhanh chóng với chủ đích xấu. Những thông tin sai lệch này có thể gây ra những tác động tiêu cực tới người dùng. Vì thế, bài toán nhận diện tin giả trên mạng xã hội đang được quan tâm nghiên cứu nhiều hơn trong thời gian gần đây.

Bài toán nhận diện tin giả là quá trình xử lý ngôn ngữ tự nhiên, gồm việc xác định và phân loại những tin tức, bài viết hoặc các loại văn bản khác là thật hay giả. Mục đích chung của bài toán là phát triển các thuật toán để có thể tự xác định những thông tin sai lệch, nhằm hạn chế tin giả và phổ biến những thông tin chính xác hơn.

Dưới sự hướng dẫn của thầy Phan Tấn Quốc, học viên thực hiện đề tài "Nghiên cứu về thuật toán Support Vector Machines (SVM) giải bài toán nhận diện tin giả" để làm tiểu luận cho học phần đồ án chuyên ngành.

2. Lịch sử nghiên cứu vấn đề

Hiện nay có rất nhiều kỹ thuật có thể nhận diện tin giả, nhưng không có kỹ thuật nào có thể hoàn toàn phân biệt được vì những hạn chế như thiếu dữ liệu, dữ liệu quá lớn, dữ liệu quá đa dạng, … Có 2 hướng tiếp cận chính trong việc nhận diện tin giả là: Thủ công hoặc tự động.

Kỹ thuật nhận diện thủ công

Tin tức sẽ được kiểm tra bằng cách thủ công bằng việc con người sẽ tự xác minh thông tin qua việc đọc và xem xét những thông tin trong tin đó. Quá trình kiểm tra thông tin thủ công được phân loại thành hai loại:

- Kiểm tra thủ công dựa trên chuyên gia (EXPERT-BASED MANUAL FACT-CHECKING): Việc xác minh thông tin được thực hiện bởi những chuyên gia như phóng viên để kiểm tra xem thông tin đó có được xác thực hay không. Cách truyền

thống này sẽ tốn thời gian, tốn chi phí hơn. Có nhiều trang web sử dụng kỹ thuật này, ví dụ như PolitiFact, FactCheck, Snopes, TruthOrFiction, …

- Kiểm tra thủ công dựa trên cộng đồng (CROWD-SOURCED MANUAL FACT-CHECKING): Cộng đồng ở đây là nhiều người bình thường không phải chuyên gia thực hiện viện xác thực thông tin. Vì là một cộng đồng nhiều người như vậy, nên kỹ thuật này cần thực hiện việc thanh lọc những người không đáng tin cậy, và khắc phục những xung đột xảy ra do xung đột kết quả giữa người với người. Kỹ thuật này sẽ khó duy trì cũng như không hiệu quả bằng kỹ thuật đầu tiên. Ví dụ trang web sử dụng kỹ thuật này là Fiskkit

Kỹ thuật nhận diện tự động

Kỹ thuật này sẽ giúp tiết kiệm thời gian cũng như chi phí hơn so với cách trên. Có 2 hướng chính để giải quyết bài toán theo hướng tự động là học máy và học sâu.

- Học máy (Machine Learning): Là một lĩnh vực thuộc trí tuệ nhân tạo (AI), cho máy tính khả năng tự học hỏi dựa trên dữ liệu đầu vào, từ đó đưa ra dự đoán. Đa số các thuật toán giải bài toán nhận diện tin giả theo hướng này đều là thuật toán học máy giám sát (Supervised Learning), một số thuật toán trong học máy được sử dụng nhiều như: Naïve Bayes, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, …

- Học sâu (Deep Learning): Cũng là một lĩnh vực thuộc trí tuệ nhân tạo (AI), sử dụng nhiều lớp neural networks, gọi là deep neural network để mô phỏng lại bộ não của con người. Một số kỹ thuật để giải bài toán nhận diện tin giả như: Convolutional neural networks (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), …

Phương pháp được dùng nhiều hơn là học sâu, và nó cũng cho ra kết quả có độ chính xác cao hơn học máy. Tuy nhiên, việc phát triển và cài đặt một hệ thống dựa trên học sâu phức tạp và tốn kém hơn.

3. Mục đích và nhiệm vụ nghiên cứu

Tìm hiểu, nghiên cứu thuật toán Support Vector Machines (SVM), xây dựng và huấn luyện mô hình, sau đó áp dụng vào giải bài toán nhận diện tin giả.

Bài toán nhận diện tin giả sẽ nhận vào dữ liệu lẫn lộn giữa tin thật và tin giả. Mục tiêu là áp dụng thuật toán SVM tìm ra lời giải của bài toán, tìm ra đâu là tin thật và đâu là tin giả với độ chính xác cao nhất có thể.

Nghiên cứu về thuật toán Support Vector Machines (SVM) để giải bài toán nhận diện tin giả. Nghiên cứu các yếu tố có thể ảnh hưởng tới độ chính xác của thuật toán trong việc giải bài toán nhận diện tin giả như kích thước và chất lượng của dữ liệu đầu vào, tiền xử lý dữ liệu, chọn mô hình huấn luyện, các tham số tuỳ chọn của thuật toán SVM.

Các nội dung nghiên cứu cụ thể:

- Tổng quan về tin giả, bài toán nhận diện tin giả
- Khảo sát các nghiên cứu khác có liên quan tới bài toán nhận diện tin giả
- Nghiên cứu dữ liệu thực nghiệm cho bài toán nhận diện tin giả
- Nghiên cứu về học máy, cụ thể là về thuật toán SVM (Cơ sở lý thuyết, nguyên tắc hoạt động), và các biến thể khác của thuật toán SVM.
- Xây dựng thuật toán SVM áp dụng vào bài toán nhận diện tin giả
- Thực nghiệm và đánh giá thực nghiệm.

4. Đối tượng nghiên cứu và phạm vi nghiên cứu

Khách thể nghiên cứu

- Áp dụng thuật toán SVM giải bài toán nhận diện tin giả.

Đối tượng nghiên cứu

- Thuật toán SVM.
- Bài toán nhận diện tin giả.

Phạm vi nghiên cứu

- Một thuật toán trong machine learning cụ thể là thuật toán SVM để giải bài toán nhận diện tin giả.
- Bài toán nhận diện tin giả.

5. Phương pháp nghiên cứu

- Lý thuyết: Tìm đọc, nghiên cứu các tài liệu, tạp chí, các bài báo sau đó tổng hợp lại những lý thuyết có liên quan tới thuật toán SVM và bài toán nhận diện tin giả.

- Thực nghiệm: Cài đặt thử nghiệm bằng ngôn ngữ Python, dữ liệu thực nghiệm được lấy từ Kaggle, sau đó sẽ đánh giá hiệu quả thuật toán.

- Định lượng: Đánh giá kết quả lời giải dựa trên công thức tính toán độ chính xác. Sau đó so sánh kết quả bên trên với một số thuật toán khác.

6. Những đóng góp mới của đề tài

Ý nghĩa khoa học

- Đóng góp chung vào việc nghiên cứu về những thuật toán học máy, ở đây là thuật toán SVM.

- Đóng góp vào việc nghiên cứu những giải pháp để giải bài toán nhận diện tin giả.

Ý nghĩa thực tiễn

- Trong thời đại, mạng xã hội phát triển rộng rãi tới người dùng thì bài toán nhận diện tin giả sẽ giúp ích nhiều trong việc hạn chế tác hại của tin giả.

- Đề xuất một lời giải cho bài toán nhận diện tin giả, góp phần nâng cao hiệu quả của lĩnh vực này.

7. Cấu trúc của tiểu luận

Ngoài phần mở đầu, kết luận, danh mục tài liệu tham khảo, tiểu luận gồm có 3 chương:

Chương 1: Giới thiệu về tin giả và bài toán nhận diện tin giả. Chương này giới thiệu tổng quan về tin giả, phát biểu về bài toán nhận diện tin giả, ứng dụng của bài toán trong thực tiễn. Khảo sát dữ liệu thực nghiệm và công trình nghiên cứu liên quan tới bài toán.

Chương 2: Xây dựng thuật toán Support Vector Machine (SVM) để giải bài toán nhận diện tin giả. Chương này sẽ nói tổng quan về thuật toán SVM, sau đó nghiên cứu áp dụng SVM vào giải bài toán.

Chương 3: Cài đặt thuật toán SVM giải bài toán nhận diện tin giả. Đánh giá và so sánh kết quả thực nghiệm.

CHƯƠNG 1. GIỚI THIỆU VỀ BÀI TOÁN NHẬN DIỆN TIN GIẢ

1.1. Giới thiệu về tin giả

Các phương tiện truyền thông đang trở nên ngày càng đa dạng như tạp chí, sách, báo hay truyền hình và đặc biệt hơn hết là sự phát triển ngày càng mạnh mẽ của mạng xã hội, người dùng cũng dễ tiếp cận lượng thông tin khổng lồ trên những nền tảng đó. Bên cạnh lợi ích mang lại thì những điều trên cũng dẫn đến một vấn đề là "tin giả", "thông tin sai lệch" hay "thông tin sai sự thật" đang xuất hiện nhiều và cũng dễ dàng lan truyền hơn trên các phương tiện truyền thông ngày nay.

1.1.1. Tin giả

Tin giả (fake news), hay còn được gọi là tin tức giả mạo, tin rác, thuật ngữ này bắt nguồn từ thuật ngữ "fake news" của báo chí Âu, Mỹ và đã tồn tại rất lâu từ trước. Một ví dụ điển hình về những thông tin sai lệch phổ biến bắt nguồn vào nắm 1938, vụ việc chương trình radio phát vở kịch "The War of the Worlds" lấy cảm hứng từ cuốn tiểu thuyết của H.G. Wells, đã gây ra hoảng loạn cho khoảng một triệu cư dân. Tuy nhiên cho đến nay vẫn chưa có định nghĩa chung thống nhất cho thuật ngữ này.

Theo từ điển Cambridge "Tin giả là những tin tức có nội dung câu chuyện sai sự thật, được lan truyền trên internet hoặc các phương tiện truyền thông khác, thường được tạo ra để gây ảnh hưởng tới những quan điểm chính trị hoặc là một trò đùa".

Định nghĩa theo "Tiêu chuẩn cộng đồng" của Facebook:

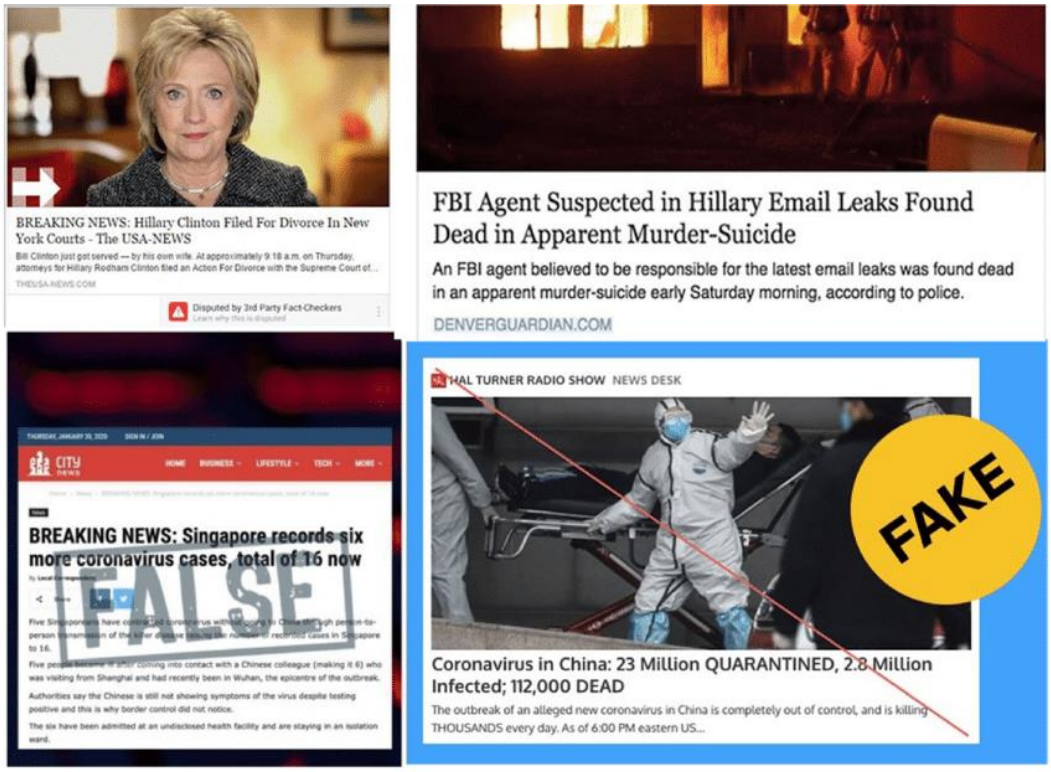
+ Thông tin sai lệch là nội dung chứa tuyên bố mà bên thứ ba đáng tin cậy xác định là sai sự thật.

+ Tin đồn không thể xác minh là tuyên bố mà đối tác chuyên môn tại nguồn xác nhận là rất khó hoặc không thể truy vết, trong trường hợp không có nguồn đáng tin cậy, nội dung tuyên bố không đủ cụ thể để vạch trần hoặc tuyên bố đó đáng ngờ/phi lý đến mức khó tin.

Bộ Thông tin và Truyền thông đưa ra định nghĩa về tin giả trên không gian mạng như sau: Tin giả trên không gian mạng là những thông tin sai sự thật được cố ý đăng tải, lan truyền nhằm mục đích không chính đáng, gây hiểu lầm cho người đọc, người xem hoặc những thông tin có một phần sự thật nhưng không hoàn toàn chính xác do không

được kiểm chứng, xác minh hoặc bị phóng đại, suy diễn, làm thay đổi bản chất của sự việc, thường xuất hiện dưới dạng tin tức và được lan truyền chủ yếu trên mạng xã hội.

PGS.TS Nguyễn Thị Trường Giang, Phó Giám đốc Học viện Báo chí và Tuyên truyền nhận định: "Tin giả giống như một loại virus. Tiếp xúc với tin giả nhiều lần hoặc nó đến từ một người nổi tiếng, có ảnh hưởng sẽ khiến công chúng bị thuyết phục, tin theo". Tin giả ngày nay thường gắn liền với những sự kiện hay từ khoá dễ đánh vào tâm lý người đọc, khiến cho người đọc dễ dàng bị thuyết phục đó là tin thật.



Hình 1.1. Một số tin tức giả trên mạng xã hội

(Ảnh: https://www.researchgate.net/figure/Examples-of-some-Fake-News-spread-over-social-media-Source-Facebook-and-Twitter_fig1_354203307)

Một ví dụ về tin giả ở Việt Nam như vào năm 2023, có một đoạn video khoảng 15 giây được lan truyền trên trang UEH Confession với nội dung "có nữ sinh HUFLIT bị xâm hại tập thể và nhảy lầu tự tử tại Trường Quân sự Quân khu 7". Vụ việc này được lan truyền với tốc độ chóng mặt, làm cho rất nhiều người tin là sự thật.

Với bối cảnh internet và các trang phương tiện truyền thông phát triển như hiện nay, thì nguồn thông tin lại càng khó kiểm soát cũng như kéo theo ảnh hưởng tới từ chính người đọc. Hơn nữa, các đối tượng càng có nhiều thủ đoạn hơn trong việc truyền tải

thông tin giả mạo, nguy hiểm nhất là việc sử dụng "khoảng trống thông tin" để tấn công vào sự hiếu kỳ của công chúng và làm mới thông tin cũ, bịa đặt thông tin mới. Nhiều thông tin bị xuyên tạc, bóp méo sự thật, nhất là vấn đề liên quan đến nội bộ Đảng, Nhà nước, tham nhũng, tiêu cực. Thông thường, tin giả được tạo ra có mục đích vụ lợi, thu hút lượt xem, lượt thích của cộng đồng mạng, tạo ra lợi nhuận. Tuy nhiên, nhiều tin giả được tạo ra với mục đích xâm phạm an ninh quốc gia, trật tự an toàn xã hội, quyền và lợi ích của tổ chức, cá nhân.

Nhìn chung, tin giả trong thời đại ngày nay đang là một vấn đề cần được lưu ý, vì nếu tin tức giả được tạo ra một cách có tổ chức, có mục đích rõ ràng, có nguồn lực hỗ trợ phía sau cũng như sử dụng những công nghệ hiện đại sẽ làm cho người dùng mạng xã hội rất khó phân biệt đâu là tin giả.

1.1.2. Phân loại tin giả

Hiện có rất nhiều ý kiến khác nhau về việc xác định cũng như phân loại tin tức giả mạo. Trên thực tế, tin giả có thể xuất hiện dưới nhiều hình thức, nhiều định dạng khác nhau như trên báo điện tử, trên mạng xã hội, các video trên Youtube, các hình ảnh trên các phương tiện truyền thông, ... Do sự đa dạng của tin giả nên việc nhận diện và phân loại tin giả là không hề đơn giản, nhưng cũng có nhiều nghiên cứu đã cố gắng để phân loại tin giả [1][2]. Chủ yếu là phân loại ra 3 kiểu tin giả sau:

- Thông tin giả mạo (Disinformation): Là một loại thông tin được tạo ra và chia sẻ có mục đích là gây hiểu lầm cho người đọc. Thông tin này sẽ gây nhiều ảnh hưởng vì người tạo ra và chia sẻ là có chủ đích và cố gắng đánh lừa người đọc.

- Thông tin không đúng sự thật (Misinformation): Thông tin này không chắc chắn, mơ hồ, không rõ ràng gây hiểu lầm cho người đọc. Đây vẫn có thể là thông tin chính xác, do có thể được tạo ra và chia sẻ một cách vô tình, không cố ý.

- Thông tin độc hại (Malinformation): Là thông tin thật nhưng được sử dụng với mục đích gây tổn hại cho một cá nhân, tổ chức hoặc là một quốc gia.

Theo bài viết của Claire Wardle trên First Draft News, có 7 loại của Misinformation và Disinformation là:

- Tin tức châm biếm/nhại lại (Satire/Parody): Không có chủ đích gây hại nhưng thông tin sẽ gây hiểu lầm cho người đọc

- Tin tức có các yếu tố không đúng (False Connection): Tin tức có tiêu đề, hình ảnh hay chú thích không phù hợp với nội dụng

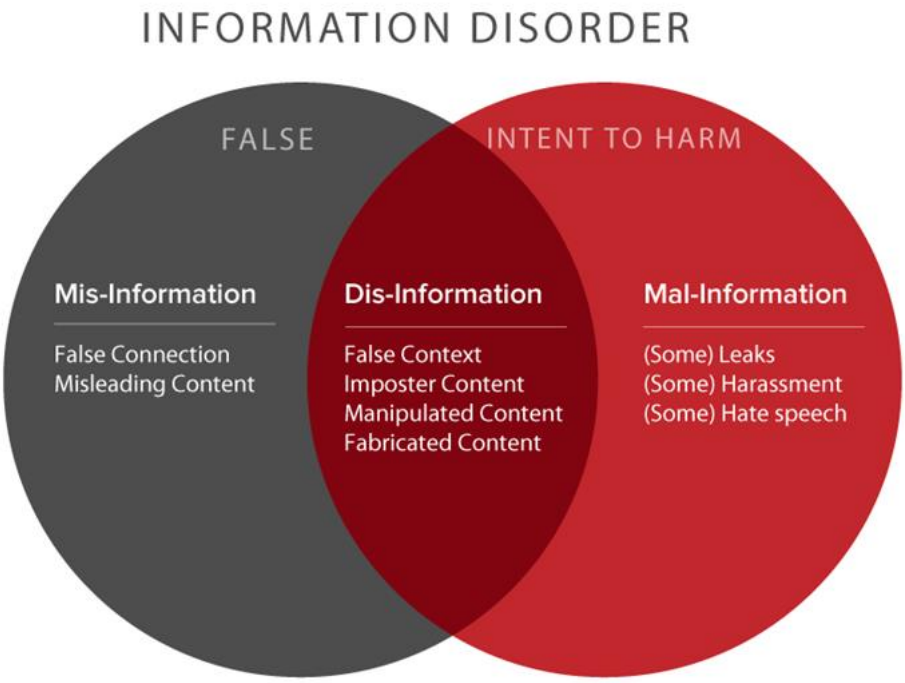
- Tin tức gây hiểu lầm (Misleading Content): Sử dụng thông tin sai lệch để đánh giá một vấn đề hoặc cá nhân nào đó.

- Tin tức có bối cảnh không đúng (False context): Nội dung tin tức được chia sẻ với bối cảnh không đúng.

- Tin tức mạo danh (Imposter content): Tin tức có nguồn không đáng tin, mạo danh.

- Tin tức thao túng (Manipulated content): Tin tức có thông tin hoặc hình ảnh sử dụng để thao túng, đánh lừa người đọc

- Tin tức bịa đặt (Fabricated content): Tin tức có nội dung hoàn toàn sai, nhằm mục đích lừa đảo và trục lợi.



Hình 1.2. Phân loại tin giả

(Ảnh: https://shorensteincenter.org/information-disorder-framework-for-research-and-policymaking/)

Ngoài ra còn có một cách phân loại khác từ hiệp hội EAVI (European Association for Viewers Interests), phân loại tin giả thành 10 loại: Tuyên truyền (Propaganda), clickbait, nội dung được tài trợ (sponsored content), châm biếm/lừa bịp (Satire/Hoax),

tin lỗi (error), tin thiên vị (Partisan news), thuyết âm mưu (Conspiracy theory), nguỵ khoa học (Pseudoscience), sai sự thật (Misinformation), không có thật (Bogus).

1.1.3. Đặc điểm của tin giả

Tin giả thường có 3 đặc điểm cơ bản như sau:

- Số lượng của tin giả: Vì không có giai đoạn xác thực tin tức, nên bất cứ ai cũng có thể viết ra những thông tin sai lệch trên internet, ngoài ra cũng có những website sinh ra để lan truyền những tin tức giả mạo. Do đó, có một lượng lớn tin giả sẽ được lan truyền qua internet.

- Sự đa dạng của tin giả: Có rất nhiều định nghĩa gần giống với tin giả như tin đồn, tin châm biếm, tin sai sự thật, thuyết âm mưu, ... Với sự phát triển của mạng xã hội, làm cho tin giả dễ dàng thao túng tâm lý nhiều người dùng hơn, việc này sẽ ảnh hưởng tới cách người dùng phản ứng với tin thật. Tin giả đang ngày càng ảnh hưởng sâu sắc tới nhiều khía cạnh trong cuộc sống.

- Tốc độ lan truyền của tin giả: Tin giả mạo có vòng đời ngắn, ví dụ có thể thấy nhiều trang tin tức giả mạo đã sập và còn có thể truy cập được nữa. Đa số những tin tức giả mạo trên mạng xã hội hiện tại tập trung vào sự kiện mới nhất, vấn đề nóng để thu hút người dùng, khiến cho tin giả ngày càng khó phát hiện hơn.

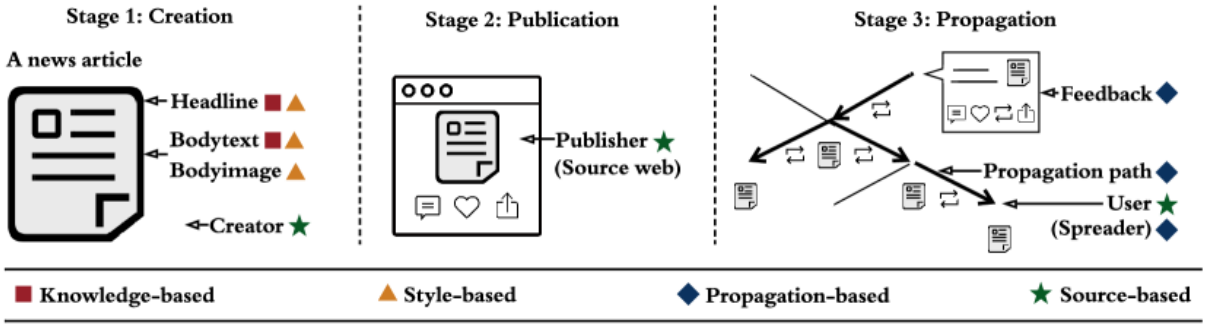
Ví dụ một tin tức giả sẽ bao gồm những thành phần sau:

- Người đăng thông tin/Người phát tán: Người đăng có thể là người thật hoặc không.

- Nạn nhân: Mục tiêu của những người đăng có thể là người dùng mạng xã hội hoặc các nền tảng trực tuyến khác. Tuỳ vào mục đích, nạn nhân có thể là học sinh, người lớn, người già, ...

- Nội dung tin tức: Bao gồm tiêu đề, nội dung, các phương tiện truyền thông.

- Ngữ cảnh xã hội: Cho biết nội dung được phân phối ra sao trên internet. Bao gồm phân tích mạng người dùng, phân tích mô hình phát tán thông tin.



Hình 1.3. Vòng đời của tin giả

(Ảnh: https://ar5iv.labs.arxiv.org/html/1812.00315)

1.1.4. Các tác động của tin giả

Sự phát triển nhanh chóng của các phương tiện truyền thông, đặc biệt là các nền tảng mạng xã hội tăng nguy cơ thường xuyên tiếp xúc với tin giả của người dụng. Việc tiếp cận với tin giả sẽ ảnh hưởng tới cộng đồng trên nhiều mặt khác nhau và dẫn tới hậu quả khó lường. Theo sách "Cẩm nang phòng chống tin giả, tin sai sự thật trên không gian mạng" của Bộ Thông tin và Truyền thông [3], tin giả có thể tác động tới những vấn đề như sau:

1.1.4.1. Tác động tới an ninh – kinh tế

Tin giả hay những tin đồn sai sự thật có thể gây ảnh hưởng về mặt kinh tế rất lớn, làm mất niềm tin vào các định chế lớn, làm tổn hại uy tín, mất hình ảnh thương hiệu của doanh nghiệp, thậm chí ảnh hưởng đến sự ổn định của nền kinh tế trong những giai đoạn nhạy cảm... Một tin xấu có thể khiến giá cổ phiếu lao dốc, hủy hoại danh tiếng của doanh nghiệp hoặc gây ra những kỳ vọng vô lý của khách hàng. Các doanh nghiệp phi đạo đức cũng có thể tạo ra tin tức hoặc đánh giá giả mạo để nâng cao vị thế hoặc lợi nhuận của chính họ.

Vào tháng 9 năm 2008, một bài báo cách đây 6 năm về vụ phá sản năm 2002 của công ty mẹ United Airlines lại xuất hiện trên Internet và bị nhầm lẫn là đang báo cáo một hồ sơ phá sản mới của công ty. Tình tiết này đã khiến giá cổ phiếu của công ty giảm tới 76% chỉ trong vài phút, trước khi NASDAQ (National Association of Securities Dealers Automated Quotation System) - một sàn giao dịch chứng khoán Hoa Kỳ tạm dừng giao dịch. Sau khi tin tức đó được xác định là sai, giá cổ phiếu đã tăng trở lại, nhưng vẫn kết thúc ngày ở mức thấp hơn 11,2% so với giá đóng cửa trước đó.

Thời điểm cuối tháng 3, đầu tháng 4/2022, loạt tin đồn thất thiệt trên mạng xã hội về việc một số doanh nghiệp ngoài nhà nước là các công ty đại chúng niêm yết trên sàn chứng khoán bị thanh tra chuyên đề hoạt động phát hành trái phiếu và việc thực hiện pháp luật về kế toán, thuế, chứng khoán đã gây hoang mang cho nhà đầu tư, khiến cổ phiếu các doanh nghiệp xuất hiện trong tin đồn bị ảnh hưởng nghiêm trọng. Một số công ty đã có thông cáo báo chí để đính chính, trấn an các cổ đông, nhà đầu tư về tin đồn này nhưng cổ phiếu của các công ty này vẫn theo đà tiếp tục giảm mạnh.

Hình 1.4. Tin tức không đúng về các doanh nghiệp
(Ảnh: https://vtv.vn/kinh-te/doanh-nghiep-dieu-dung-vi-tin-don-20221107235452187.htm)

1.1.4.2. Tác động tới an ninh quốc gia

Tin giả được hình thành trên cơ sở những vấn đề được người dân quan tâm, những chủ đề "nóng, nhạy cảm", gây hoang mang, mất niềm tin vào cán bộ, hệ thống lãnh đạo, Nhà nước, Đảng. Ví dụ như những vấn đề liên quan tới công tác nhân sự, tác động đến tâm lý cử tri, gây ảnh hưởng tới kết quả bầu cử, bỏ phiếu tín nhiệm cán bộ cấp cao từ đó ảnh hưởng tới an ninh quốc gia theo nhiều góc độ khác nhau. Trước mỗi kỳ bỏ phiếu, bầu cử, các thế lực phản động thù địch sẽ tung những bài viết xuyên tạc, chống phá nhằm mục đích nhiễu loạn thông tin, khiến người dân hoang mang, nghi ngờ.

Hình 1.5. Những thông tin xuyên tạc liên quan tới bầu cử

(Ảnh: http://svhttdl.phutho.gov.vn/tin/canh-giac-voi-nhung-luan-dieu-xuyen-tac-truoc-them-bau-cu_1931.html)

Ví dụ vào năm 2021 khi dịch COVID vẫn đang có những diễn biến căng thẳng tại Việt Nam, trên mạng xã hội có hàng loạt những bài viết xuyên tạc, mang tính chất phản động, phản đối Chính phủ nhằm mục đích kích động người dân, đã gây ra nhiều khó khăn trong công cuộc chống dịch của Nhà nước ta.

Hình 1.6. Thông tin sai sự thật về tình hình dịch COVID-19
(Ảnh: https://lacduong.lamdong.dcs.vn/thong-tin-can-biet/type/detail/id/28113/task/1473)

Một ví dụ khác, vào ngày 19/7/2021, trên Facebook đã lan truyền hình ảnh và video một người đàn ông tự thiêu trên địa bàn phường Trường Thọ, TP. Thủ Đức. Bài viết này do đối tượng Phan Hữu Điệp Anh đăng tải với dòng tiêu đề: "Bức xúc vì cách thức chống dịch COVID-19… người dân phẫn uất ngay giữa đường bức bách, tự thiêu". Nhưng thực chất người đàn ông trong video là một bệnh nhân tâm thần có giấy chứng nhận khuyết tật thần kinh – tâm thần 2, sau đó nạn nhân cũng đã được đưa đi điều trị tại Bệnh viện Chợ Rẫy. Còn đối tượng Anh cũng đã bị bắt ngay sau đó, cuối cùng bị tuyên phạt 1 năm 6 tháng tù về tội "lợi dụng các quyền tự do dân chủ xâm phạm lợi ích của Nhà nước, quyền, lợi ích hợp pháp của tổ chức, cá nhân".

Hình 1.7. Thông tin một người đàn ông tự thiêu vì bức xúc cách chống dịch (Ảnh: https://congan.com.vn/doi-song/tphcm-bac-tin-nguoi-dan-phan-uat-tu-thieu-vi-buc-xuc-cach-chong-dich-covid-19_116649.html)

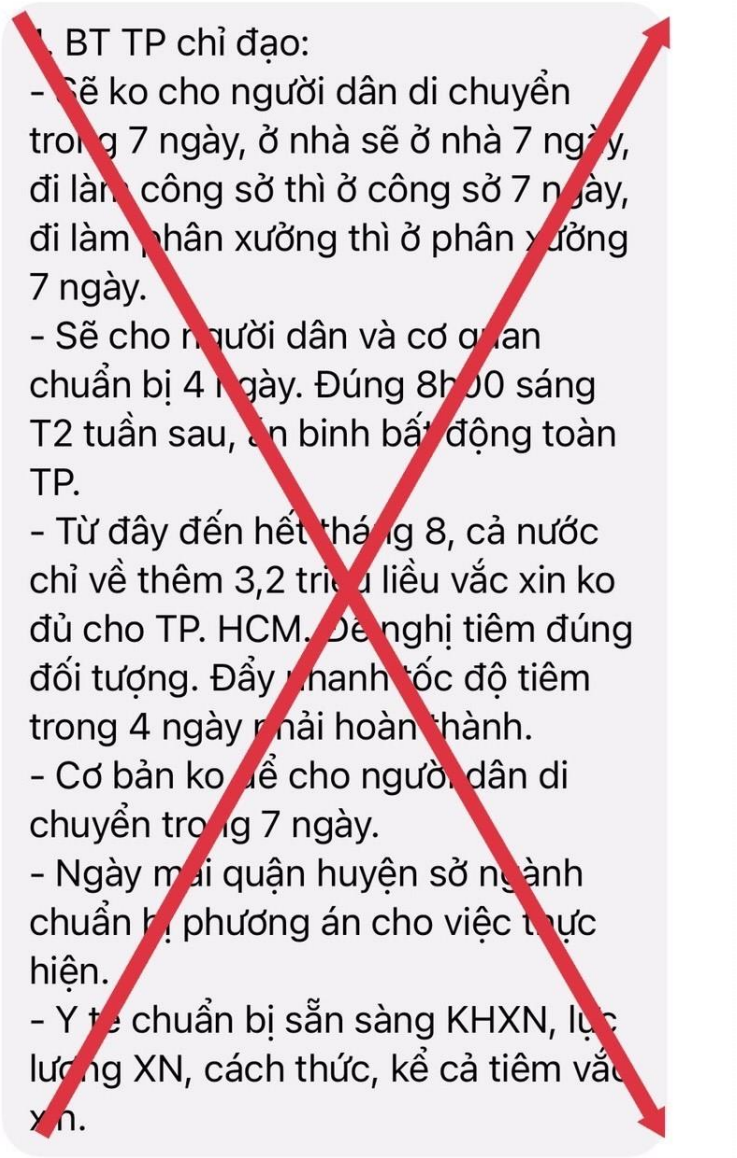
1.1.4.3. Tác động đến tâm lý, niềm tin của người dân

Tin giả có thể gây ra sự hoang mang, căng thẳng và lo lắng cho người đọc, và có thể dẫn đến những hậu quả tiêu cực về tâm lý và cảm xúc của họ. Tin giả có thể ảnh hưởng tới tinh thần, sức khoẻ tâm lý của người đọc như gây hoang mang, lo lắng. Tin giả cũng có thể gây ra mất mát, đặc biệt là khi chúng liên quan đến những sự kiện quan trọng.

Tin giả còn làm cho người dân suy giảm niềm tin khi đọc tin tức trên báo chí truyền thống hay các trang tin tức chính thống. Trong thời đại công nghệ phát triển, với ngày càng nhiều các nền tảng xã hội, nhắn tin ngày càng nhiều thì càng khó kiểm soát chất lượng tin tức hơn, làm cho việc các đối tượng giả mạo các trang tin tức chính thống dễ dàng hơn. Đối tượng có thể tung tin giả dưới dạng bài viết hoặc video clip bằng cách chỉnh sửa, lồng ghép hình ảnh một cá nhân vào bối cảnh có thực (ví dụ sử dụng hình ảnh logo của VTV lồng ghép để quảng cáo các thực phẩm chức năng không có nguồn gốc rõ ràng), làm cho người dân lầm tưởng là sản phẩm uy tín và lan truyền chúng trên các trang mạng xã hội.

Hình 1.8. Giả mạo kênh Quốc phòng Việt Nam để quảng cáo thực phẩm chức năng (Ảnh: https://tingia.gov.vn/kenh-gia-mao-truyen-hinh-quoc-phong-viet-nam-de-quang-cao-thuc-pham-chuc-nang.html)

Tin giả còn có thể tác động tới hành động của người dân, ví dụ vào sáng ngày 12/8/2021 trên mạng xã hội đang lan truyền thông tin có nội dung: “Sẽ ko cho người dân di chuyển trong 7 ngày, ở nhà sẽ ở nhà 7 ngày, đi làm công sở thì ở công sở 7 ngày, đi làm phân xưởng thì ở phân xưởng 7 ngày; Cơ bản ko để cho người dân di chuyển trong 7…”. Làm cho người dân hoang mang, đổ xô đi mua nhu yếu phẩm để dự trữ, khiến cho hàng hoá bỗng nhiên trở nên khan hiếm, ảnh hưởng tới hoạt động buôn bán.



Hình 1.9. Thông tin giả mạo cơ quan chức năng để thông báo

(Ảnh: https://moit.gov.vn/tin-tuc/bao-chi-voi-nguoi-dan/tp-hcm-khong-cho-nguoi-dan-di-chuyen-trong-7-ngay-la-tin-gia.html)

1.1.4.4. Tác động tới sức khoẻ cộng đồng

Theo nghiên cứu của Paul Hunter và Julii Brainard từ trường Đại học East Angila (UEA) [4], Tin giả được lan truyền có thể gây hại cho sức khỏe con người, đặc biệt là trong thời gian bùng phát dịch bệnh. Tin giả có thể làm các đợt bùng phát dịch bệnh thêm trầm trọng.

Chuyên gia COVID-19, giáo sư Paul Hunter và tiến sĩ Julii Brainard của thuộc Trường Y khoa Norwich của UEA, đã thử nghiệm tác động từ việc chia sẻ thông tin sai sự thật đối với sức khoẻ của người bệnh trong thời kỳ một dịch bệnh bùng phát. Và nghiên

cứu đã cho thấy 40% người dân ở Anh tin vào ít nhất một thuyết âm mưu, tỉ lệ này thậm chí còn cao hơn khi ở Hoa Kỳ và các nước khác. Giáo sư Hunter nhận xét "tin tức giả được tạo ra không tôn trọng tính chính xác và thường dựa trên các thuyết âm mưu". Giáo sư Hunter còn nói thêm một thực trạng đáng lo ngại, "mọi người thường chia sẻ lời khuyên sai trên mạng xã hội hơn là chia sẻ lời khuyên từ các nguồn đáng tin cậy như Cơ quan Y tế công cộng Anh hoặc Tổ chức Y tế thế giới". Công chúng có xu hướng tương tác với các "bong bóng thông tin" trên mạng do đó thông tin được chia sẻ có nhiều khả năng là những thông tin sai lệch hơn là thông tin chính xác.

Ngay sau khi dịch Covid-19 bùng nổ vào tháng 2/2020, tin giả về dịch bệnh Covid-19 tràn ngập trên các trang mạng xã hội như Facebook, YouTube, Twitter, … Tổng Giám đốc Tổ chức Y tế thế giới (WHO) - Tedros Adhanom Ghebreyesus gọi đây là "infodemic" (viết tắt của "information pandemic", tức "đại dịch thông tin") và tin giả là "căn bệnh thứ hai" tồn tại cùng COVID-19.

Một số hậu quả của việc cung cấp, chia sẻ tin giả, tin sai sự thật có thể kể đến như:

+ Rạn nứt các mối quan hệ.

+ Gây ra những phiền toái, phân biệt đối xử, cô lập, xa lánh... cho những người liên quan.

+ Ảnh hưởng đến uy tín, danh dự của bản thân và người khác

+ Tổn thương đến sức khỏe cả về thể chất và tinh thần.

+ Bị phạt tiền hoặc khởi tố hình sự tùy theo mức độ nghiêm trọng của sự việc theo quy định pháp luật.

+ Tổn thương đến sức khỏe cả về thể chất và tinh thần.

+ Bị phạt tiền hoặc khởi tố hình sự tùy theo mức độ nghiêm trọng của sự việc theo quy định pháp luật.

Hình 1.10. Một số tin giả trên mạng xã hội

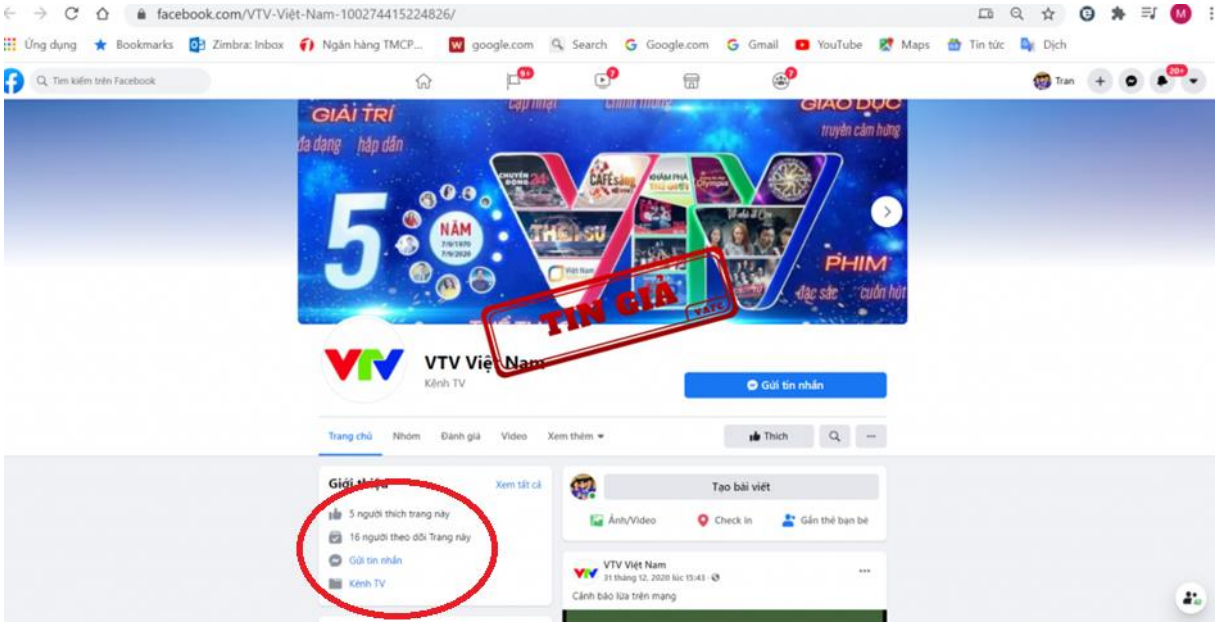
(Ảnh: https://infographics.vn/tin-gia-ve-covid-19-gay-hoang-mang-buc-xuc-du-luan/21152.vna)

1.1.5. Nhận biết, cách xử lý và ngăn chặn tin giả

1.1.5.1. Cách nhận biết tin giả

Với việc internet phát triển như ngày nay, thì càng có thêm vô số luồng thông tin để người dùng tiếp cận, thông tin sai thật, tin giả cũng được "làm giả" một cách tinh vi và lan truyền trên các nền tảng mạng xã hội sẽ khiến người dùng bị rối loạn trước quá nhiều luồng thông tin và không biết phải tin vào điều gì. Do đó, mỗi người dùng cũng phải tự trang bị cho mình kiến thức về tin giả, khả năng tự chọn lọc thông tin.

Người đọc cần phải kiểm tra nguồn tin, xem nguồn tin có đến từ website chính thống hay không, cần cảnh giác nếu như là những website không rõ nguồn gốc, không được xác thực. Kiểm tra tác giả/người đăng bài có uy tín hay không, nếu trên mạng xã hội thì có thể kiểm tra tài khoản của người đăng bài, kiểm tra lượt tương tác, số lượng bạn bè, … tránh trường hợp là tài khoản giả mạo (clone account). Ngoài ra, nên đối chiếu những thông tin trên mạng xã hội với tin tức trên truyền hình hoặc từ những trang báo uy tín, có nguồn gốc rõ ràng và có thương hiệu hoặc từ các cổng/trang thông tin điện tử chính thức của cơ quan nhà nước để kiểm tra tính xác thực của thông tin.



Hình 1.11. Một trang fanpage giả mạo VTV

(Ảnh: https://tingia.gov.vn/trang-fanpage-gia-mao-vtv.html)

Ngoài ra, tin giả còn có thể xuất hiện dưới nhiều hình thức khác như hình ảnh, hoặc sử dụng các đường liên kết sai, không liên quan tới nội dung của bài viết để lừa người dùng bấm vào (clickbait). Do đó, cần kiểm tra xem nội dung, bối cảnh của hình ảnh có đúng như nội dung của bài viết không, kiểm tra đường dẫn có dẫn đến trang có nguồn gốc rõ ràng hay không. Ví dụ, vào ngày 20/10/2020, tài khoản Facebook tên Luan Nguyen đã đăng tải một bài viết với nội dung như hình 1.12, dù bài đăng có nhiều lượt

tương tác nhưng đây là tin sai sự thật. Thực chất, hình ảnh bắt nguồn từ một câu chuyện ở Trung Quốc. Một trận động đất mạnh 6,1 độ richter đã xảy ra vào ngày 30/8/2008 tại Lương Sơn, Tứ Xuyên.

Hình 1.12. Hình ảnh bài viết mẹ ôm con tại Quảng Trị
(Ảnh: https://baohatinh.vn/hinh-anh-nguoi-me-om-con-sau-sat-lo-tai-quang-tri-la-tin-gia-post200488.html)

Những bài viết với nội dung sai sự thật thường được biên soạn và định dạng với mốc thời gian không đúng với thực tế. Do đó, cần kiểm tra kỹ các mốc thời gian diễn ra sự kiện có trong bài viết, người dùng cần đặc biệt kiểm tra tin cũ được đăng lại nhưng với mục đích xấu.

Đa số tin giả được xây dựng dựa trên những sự kiện đang nóng, nhạy cảm vào thời điểm đó, thường đi với những tiêu đề giật gân nhằm hấp dẫn, thu hút sự chú ý từ người đọc. Ngoài ra, cần phải đọc kỹ và phân biết được bài viết đó là tin tức thật hay chỉ là những câu chuyện phiếm, trò đùa từ người đăng bài. Giới hạn giữa câu chuyện phiếm, lời đùa, câu chuyện mang tính giải trí và thông tin giả, bịa đặt là rất mơ hồ, mong manh. Cũng như phải đọc kỹ xem những thông tin trong bài viết như sự kiện, tên nhân vật, địa phương, thời gian xảy ra sự việc, … Với những thông tin chung chung kể trên nếu có sai sót, người đọc cần phải cẩn thận vì có thể là thông tin giả.

Những chuyên gia cũng khuyên người dùng nên tiếp cận nguồn thông tin từ các trang tin tức chính thống hoặc các trang mạng xã hội của Nhà nước. Hoặc phải kiểm chứng thông tin trên mạng xã hội trước khi chia sẻ, lan truyền, tránh trở thành nạn nhân của thông tin giả đó.

Bên cạnh đó, người đọc khi tiếp cận những thông tin chưa biết thật hay giả, có thể tham khảo từ các tin tức tương tự từ các trang tin tức chính thống để đối chiếu, kiểm tra tính xác thực của thông tin. Trong trường hợp phát hiện thông tin có dấu hiệu sai sự thật, vi phạm pháp luật có thể báo cho cơ quan chức năng để tiến hành xử lý.

Đặc biệt, người dùng cần lưu ý khi chia sẻ, tương tác với những thông tin, tin tức chưa được xác thực trên mạng xã hội, tránh cho việc vô tình lan truyền tin giả đi rộng rãi hơn. Chỉ nên chia sẻ những thông tin đã được xác thực, được đăng tải từ những trang chính thống.

1.1.5.2. Cách xử lý và ngăn chặn tin giả

Trong những năm trở lại đây, đã có nhiều quốc gia ban hành luật để chống lại tin giả, hoặc là nghiên cứu về các cách xác minh thông tin trực tuyến. Việt Nam cũng có một số bộ luật như:

+ Luật An ninh mạng năm 2018 quy định về hoạt động bảo vệ an ninh quốc gia và bảo đảm trật tự, an toàn xã hội trên không gian mạng; trách nhiệm của cơ quan, tổ chức, cá nhân có liên quan;

+ Luật Công nghệ thông tin năm 2017 quy định về hoạt động ứng dụng và phát triển công nghệ thông tin, các biện pháp bảo đảm ứng dụng và phát triển công nghệ thông tin, quyền và nghĩa vụ của cơ quan, tổ chức, cá nhân (sau đây gọi chung là tổ chức, cá nhân) tham gia hoạt động ứng dụng và phát triển công nghệ thông tin;

+ Luật An toàn thông tin mạng năm 2015 quy định về hoạt động an toàn thông tin mạng, quyền, trách nhiệm của cơ quan, tổ chức, cá nhân trong việc bảo đảm an toàn thông tin mạng; mật mã dân sự; tiêu chuẩn, quy chuẩn kỹ thuật về an toàn thông tin mạng; kinh doanh trong lĩnh vực an toàn thông tin mạng; phát triển nguồn nhân lực an toàn thông tin mạng; quản lý nhà nước về an toàn thông tin mạng

QUỐC HỘI

Luật số: 24/2018/QH14

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

LUẬT
AN NINH MẠNG

Căn cứ Hiến pháp nước Cộng hòa xã hội chủ nghĩa Việt Nam;
Quốc hội ban hành Luật An ninh mạng.

Chương I
NHỮNG QUY ĐỊNH CHUNG

Điều 1. Phạm vi điều chỉnh

Luật này quy định về hoạt động bảo vệ an ninh quốc gia và bảo đảm trật tự, an toàn xã hội trên không gian mạng; trách nhiệm của cơ quan, tổ chức, cá nhân có liên quan.

Điều 2. Giải thích từ ngữ

Trong Luật này, các từ ngữ dưới đây được hiểu như sau:

1. *An ninh mạng* là sự bảo đảm hoạt động trên không gian mạng không gây phương hại đến an ninh quốc gia, trật tự, an toàn xã hội, quyền và lợi ích hợp pháp của cơ quan, tổ chức, cá nhân.

2. *Bảo vệ an ninh mạng* là phòng ngừa, phát hiện, ngăn chặn, xử lý hành vi xâm phạm an ninh mạng.

3. *Không gian mạng* là mạng lưới kết nối của cơ sở hạ tầng công nghệ thông tin, bao gồm mạng viễn thông, mạng Internet, mạng máy tính, hệ thống thông tin, hệ thống xử lý và điều khiển thông tin, cơ sở dữ liệu; là nơi con người thực hiện các hành vi xã hội không bị giới hạn bởi không gian và thời gian.

4. *Không gian mạng quốc gia* là không gian mạng do Chính phủ xác lập, quản lý và kiểm soát.

5. *Cơ sở hạ tầng không gian mạng quốc gia* là hệ thống cơ sở vật chất, kỹ thuật để tạo lập, truyền đưa, thu thập, xử lý, lưu trữ và trao đổi thông tin trên không gian mạng quốc gia, bao gồm:

Hình 1.13. Luật An ninh mạng được ban hành vào năm 2018

Ngoài ra, mỗi cá nhân cũng cần tự trang bị kiến thức về tin giả cho bản thân, phải tuân thủ quy định của luật pháp về An ninh mạng và phải tôn trọng và thực hiện các quy tắc

ứng xử trên không gian mạng do cơ quan có thẩm quyền ban hành với mục đích nhằm nâng cao chuẩn mực đạo đức về hành vi, ứng xử trên mạng xã hội cho người dân khi sử dụng mạng xã hội. Một số bộ quy tắc có thể kể đến như: Bộ Quy tắc ứng xử trên mạng xã hội do Bộ Thông tin và Truyền thông ban hành; quy tắc sử dụng Mạng xã hội của người làm báo Việt Nam; quy tắc ứng xử của người hoạt động trong lĩnh vực nghệ thuật do Bộ Văn hóa, Thể thao và Du lịch ban hành. Nếu người dùng sử dụng các mạng xã hội như Facebook, Tiktok, Twitter, Youtube ... người dùng còn cần phải tuân theo tiêu chuẩn cộng đồng riêng của mỗi mạng xã hội trên.

Ngoài ra, khi phát hiện thông tin có dấu hiệu sai sự thật, vi phạm pháp luật có thể báo cho cơ quan chức năng để tiến hành xử lý. Quy trình tiếp nhận xử lý tin giả trên không gian mạng có thể tóm tắt lại như hình sau.

Hình 1.14. Quy trình tiếp nhận và xử lý tin giả

(Ảnh: https://tingia.gov.vn/trang/huong-dan.html)

Một số cơ quan tiếp nhận phản ánh tin giả như: Trung tâm Xử lý tin giả Việt Nam (VAFC) thuộc Cục Phát thanh, truyền hình và thông tin điện tử, Sở Thông tin và Truyền thông hoặc Văn phòng UBND các tỉnh, thành đối với tin giả liên quan tới địa phương, - Các Bộ, cơ quan ngang Bộ, tổ chức đối với tin giả có liên quan đến ngành, lĩnh vực quản lý hoặc liên quan đến chức năng, nhiệm vụ của cơ quan, tổ chức.

Về cơ quan công bố tin giả là Trung tâm Xử lý tin giả Việt Nam (VAFC), Sở Thông tin và Truyền thông địa phương hoặc cơ quan, tổ chức có trách nhiệm tiếp nhận và xác minh tin giả

Quy trình cơ bản về tiếp nhận và xử lý tin giả của cơ quan nhà nước có thẩm quyền:

Bước 1: Tiếp nhận phản ánh tin giả qua cơ quan, tổ chức, cá nhân. Công bố cách thức liên hệ qua: Địa chỉ email, số điện thoại đường dây nóng, trang web hoặc fanpage trên các mạng xã hội.

Bước 2: Tiến hành phân loại tin giả

Bước 3: Cơ quan có thẩm quyền thẩm định, xác minh và công bố tin giả; chuyển cơ quan chức năng xử lý theo quy định.

Về phần xử phạt khi vi phạm quy định, một số hành vi theo khoản 1 Điều 101 Nghị định số 15/2020/ NĐ-CP của Chính phủ như hình sau:

1.	Cung cấp, chia sẻ **thông tin giả mạo, sai sự thật, xuyên tạc, vu khống**, xúc phạm uy tín của cơ quan, tổ chức, danh dự, nhân phẩm của cá nhân.	**10 - 20 triệu đồng**		Điểm a khoản 1 Điều 101 Nghị định số 15/2020/ NĐ-CP của Chính phủ.	Nếu gây thiệt hại nghiêm trọng có thể bị truy cứu: *(1) Tội làm nhục người khác (Điều 155 Bộ luật Hình sự):* - Phạt tù: 03 tháng - 02 năm. - Hình thức xử phạt bổ sung: 10 - 30 triệu đồng; cấm đảm nhiệm chức vụ, cấm hành nghề 01 - 05 năm. *(2) Tội vu khống (Điều 156 Bộ luật Hình sự):* - Phạt tù: 01 - 03 năm. - Hình thức xử phạt bổ sung: 10 - 50 triệu đồng; cấm đảm nhiệm chức vụ, cấm hành nghề 01 - 05 năm. *(3) Tội Tội lợi dụng các quyền tự do dân chủ xâm phạm lợi ích của Nhà nước, quyền, lợi ích hợp pháp của tổ chức, cá nhân (Điều 331 Bộ luật Hình sự):*
2.	Cung cấp, chia sẻ **thông tin bịa đặt**, gây hoang mang trong Nhân dân, kích động bạo lực, tội ác, tệ nạn xã hội, đánh bạc hoặc phục vụ đánh bạc.	**10 - 20 triệu đồng**	Buộc gỡ bỏ thông tin.	Điểm d khoản 1 Điều 101 Nghị định số 15/2020/ NĐ-CP ngày 03/02/2020 của Chính phủ.	1. Người nào lợi dụng các quyền tự do ngôn luận, tự do báo chí, tự do tín ngưỡng, tôn giáo, tự do hội họp, lập hội và các quyền tự do dân chủ khác xâm phạm lợi ích của Nhà nước, quyền, lợi ích hợp pháp của tổ chức, cá nhân, thì bị phạt cảnh cáo, phạt cải tạo không giam giữ đến 03 năm hoặc phạt tù từ 06 tháng - 03 năm. 2. Phạm tội gây ảnh hưởng xấu đến an ninh, trật tự, an toàn xã hội, thì bị phạt tù từ 02 năm đến 07 năm.

Hình 1.15. Một số quy định của pháp luật về hành vi đăng tải, lan truyền tin giả

1.2. Bài toán nhận diện tin giả

Nỗ lực trong việc cố gắng tìm cách nhận diện, phát hiện ra tin giả đã tồn tại từ khoảng đầu năm 2000, dựa trên đặc điểm của ngôn ngữ, phong cách viết và nguồn của tin tức đó để dự đoán. Nhưng trong những năm trở lại đây, bài toán nhận diện tin giả mới bắt

đầu được nhiều người nghiên cứu hơn vì sự phát triển mạnh của các phương tiện truyền thông.

Bài toán đặt ra ở đây là với một dữ liệu đầu vào bất kỳ, thông qua bài toán sẽ cho ra kết quả dự đoán dữ liệu đó là tin giả hay tin thật. Mục đích của bài toán là có thể tìm ra cách phân biệt giữa tin giả/tin thật để ngăn chặn sự lan truyền của những thông tin sai lệch, nâng cao chất lượng tin tức.

1.2.1. Lý do nhận diện tin giả

Nhận diện tin giả trong thời đại là một vấn đề quan trọng, vì những tin tức hiện nay sẽ tác động rất lớn đến hành vi của người đọc. Nếu để những tin tức giả mạo, sai sự thật được lan truyền rộng rãi, trở nên phổ biến hơn thì sẽ gây ra rất nhiều hậu quả không thể lường trước được đối với cộng đồng. Việc nhận diện tin giả lại càng trở nên quan trọng hơn hết trong thời đại mà các phương tiện truyền thông đang trở nên phổ biến trong cuộc sống hàng ngày như Facebook, Youtube, Twitter, Instagram hay các trang tin tức.

Có rất nhiều lý do để nhận diện tin giả, một trong nhiều lý do quan trọng nhất là việc tin giả mang hơi hướng chính trị sẽ ảnh hưởng tới những quan điểm của người đọc. Chẳng hạn việc tin giả có thể được sử dụng như một công cụ để thổi phồng những tư tưởng sai lệch, thay đổi quan điểm của người đọc, có thể gây ảnh hưởng xấu tới cá nhân, cộng đồng hay thậm chí là quốc gia. Những tin giả như trên có thể sẽ gây bất đồng quan điểm, gây tranh cãi tới mối quan hệ giữa nhiều quốc gia. Ví dụ, đầu năm 2016, trên mạng xã hội (MXH) xuất hiện tin giả về một bé gái người Nga 13 tuổi sống ở Đức, bị cưỡng hiếp tập thể ở Berlin. Câu chuyện được nhiều hãng tin chính thống Nga và Đức đăng lại, khiến quan hệ ngoại giao Nga - Đức rơi vào khủng hoảng trầm trọng. Nga cáo buộc Đức “tìm cách cho chìm xuồng vụ việc”. Đức đáp trả rằng đây là “một thủ đoạn kiểu KGB” của Nga.

Tương tự, tin giả xuất hiện trên nhiều lĩnh vực của đời sống xã hội, nhất là những vấn đề, lĩnh vực "nóng", "nhạy cảm" liên quan đến công tác nhân sự, bầu cử, bỏ phiếu tín nhiệm, xử lý kỷ luật cán bộ cấp cao... Tin giả cũng thường xuất hiện trước những sự kiện chính trị trọng đại của Đảng, đất nước, Quân đội với tần suất ngày càng nhiều và mức độ nguy hiểm, phức tạp khó lường.

Lý do khác để nhận diện tin giả là để bảo vệ quyền lợi của người đọc, tin giả có thể được sử dụng nhằm công kích tôn giáo, chủng tộc hay một cá nhân cụ thể. Điều này gây ra hậu quả là sự kỳ thị, phân biệt đối xử. Ví dụ, khi bệnh dịch COVID-19 xuất hiện, Maatje Benassi - một nữ quân nhân dự bị người Mỹ, bị đổ lỗi là nguồn gốc mang dịch bệnh từ Trung Quốc vào quốc gia này. Lời cáo buộc sai trái tràn lan mỗi ngày, thu hút hàng trăm ngàn lượt xem. Mặc dù chẳng hề có triệu chứng nào và cũng không bị dương tính với virus SARS-CoV-2, cuộc sống của hai vợ chồng Benassi vẫn bị ảnh hưởng rất nhiều.

1.2.2. Xây dựng vấn đề

Phần này, xây dựng vấn đề bài toán nhận diện tin giả theo hướng toán học, cụ thể xây dựng ra một hàm toán học cho bài toán có tham khảo qua bài nghiên cứu [5].

- Một bài tin tức trên mạng xã hội bất kỳ sẽ có 2 thành phần chính: Nhà xuất bản (Publisher) và nội dung (Content). Publisher $\vec{p}$ là một tập hợp có các thuộc tính về tác giả như tên, tuổi và các thuộc tính khác. Nội dung $\vec{c}$ là tập hợp gồm các thuộc tính như tiêu đề, văn bản, hình ảnh.

- Giả sử, ta có $N = \{n_1, n_2, n_3, ..., n_m\}$ biểu diễn một số lượng M tin tức. Mỗi tin tức n_j sẽ có số lượng K user $U = \{u_1, u_2, u_3, ..., u_k\}$ lan truyền tin tức với bài viết tương ứng $P = \{p_1, p_2, p_3, ..., p_k\}$ của user trên mạng xã hội về bài báo n_j. Trong vấn đề nhận diện tin giả, mục tiêu là đạt được hàm học $f(\text{label} \mid n_j, N, P, U, \theta)$ để nhận diện tin giả, trong đó θ biểu thị các siêu tham số. Hàm f có thể được biểu diễn như sau:

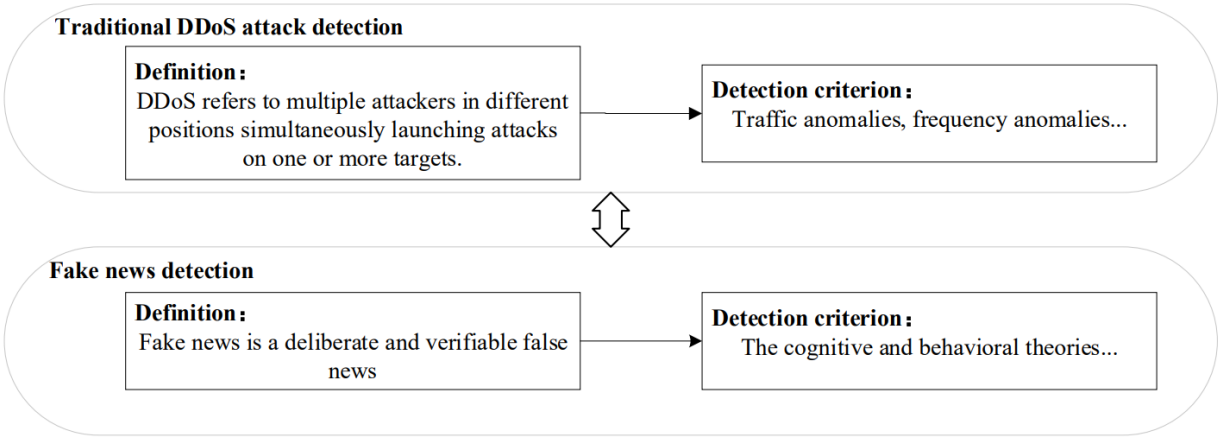
$$f(n_j) = \begin{cases} 1, \text{ nếu } n_j \text{ là tin giả} \\ 0, \text{ nếu ngược lại} \end{cases}$$

Một số phương pháp nhận diện tin giả chỉ dựa trên chính nội dung của tin tức đó, một số khác thì sẽ thu thập luôn cả thông tin của publisher để dự đoán.

1.2.3. Lý thuyết nhận diện tin giả

Các lý thuyết về nhận thức và hành vi trong khoa học xã hội và kinh tế, là những lý thuyết nền tảng của tin giả, từ đó có thể cung cấp những giá trị sâu sắc trong việc phân tích, nhận diện tin giả

Để hiểu rõ hơn về lý thuyết nhận thức và hành vi trong nhận diện tin giả, có thể xem qua sự so sánh giữa 2 vấn đề là nhận diện tin giả và phát hiện tấn công DDOS (Distributed denial of service).

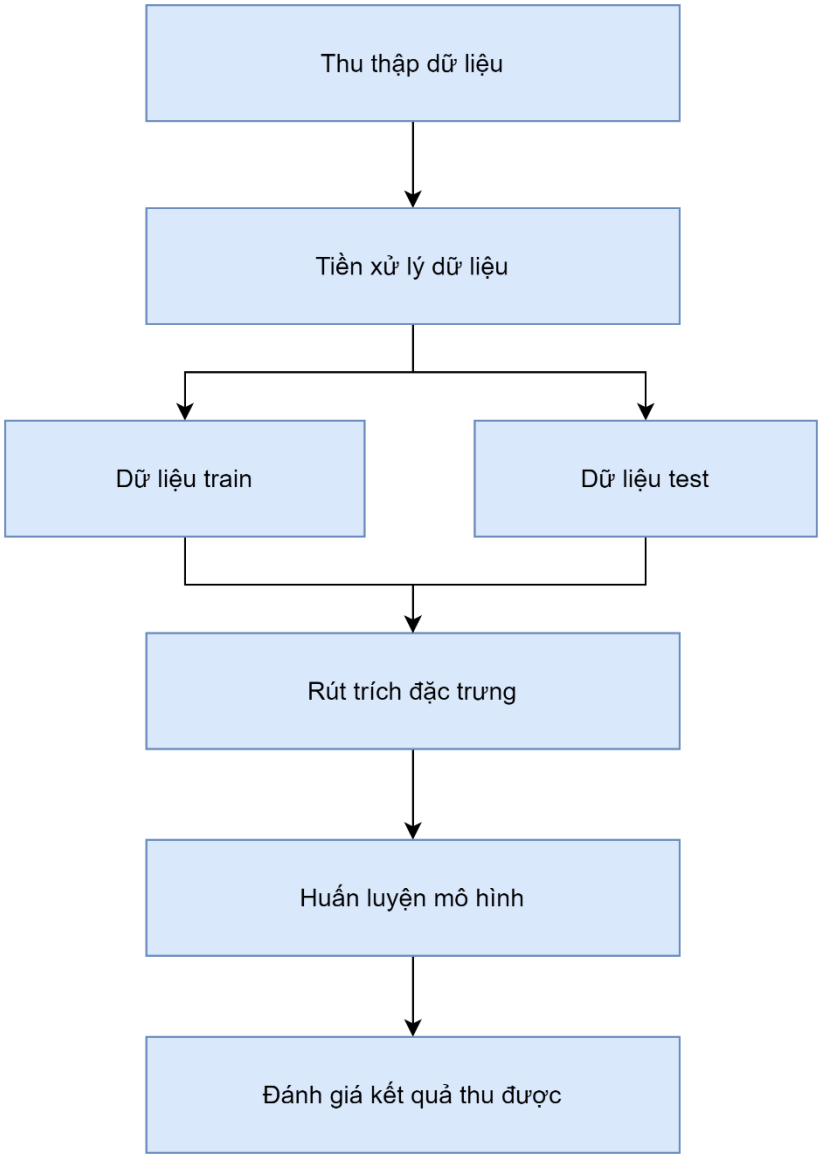


Hình 1.16. So sánh giữa DDOS và nhận diện tin giả

(Ảnh: Từ bài báo "Fake News Detection on Social Networks: A Survey" của tác giả Y. Shen và cộng sự)

Hai mô hình này được so sánh với nhau, vì cả hai đều là những vấn đề rất quan trọng trong lĩnh vực an toàn thông tin. Trước khi tiến hành nhận diện, cần phải xác định trước được những đặc điểm của tin giả và DDOS, do đó cần phải biết được định nghĩa của cả hai, sau đó tóm tắt lại những đặc trưng của chúng như là tiêu chí để phát hiện ra cả hai vấn đề. Ví dụ, lưu lượng truy cập bất thường và tần suất truy cập là tiêu chí để phát hiện các cuộc tấn công DDOS. Trong khi đó, định nghĩa về tin giả và tin thật cho ta thấy được sự khác nhau về hình thức trình bày, cách lan truyền. Tiêu chí phát hiện tin giả là dựa vào lý thuyết về nhận thức và hành vi. Ví dụ, lý thuyết the four-factory cho biết những lời nói dối sẽ thể hiện cảm xúc, hành vi khác. Có thể sử dụng các đặc điểm về nội dung hoặc lan truyền dựa trên các lý thuyết trên để phát hiện tin giả.

Mô hình xử lý của bài toán nhận diện tin giả như hình sau:



Hình 1.17. Mô hình cho bài toán nhận diện tin giả

1.2.4. Những thách thức của bài toán nhận diện tin giả

Việc nhận diện tin giả là một bài toán phức tạp với nhiều thách thức, do bản chất của tin giả luôn thay đổi và tinh vi hơn theo thời gian. Bài toán nhận diện tin giả có thể gặp một số thách thức:

- Định nghĩa của tin giả: Cho đến nay, vẫn chưa có một định nghĩa chung chính thức về thế nào là tin giả, điều này tạo ra khó khăn trong việc xác định và phân loại tin giả.

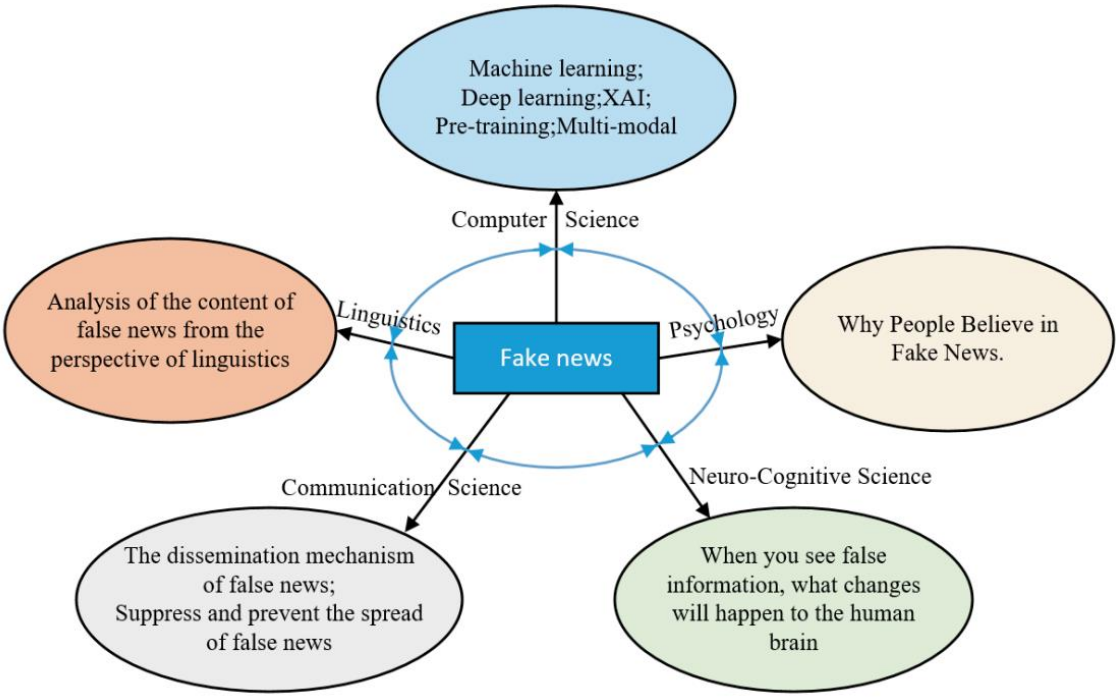
- Vấn đề về dữ liệu dùng cho bài toán: Việc thu thập dữ liệu để huấn luyện các mô hình nhận diện tin giả là một thách thức lớn. Do tin giả thường bị gỡ bỏ nhanh chóng sau khi phát hiện, dữ liệu phải luôn được cập nhật kịp thời và trong thực tế thì lượng tin tức giả là vô cùng lớn.

- Vấn đề về mô hình huấn luyện: Nếu dữ liệu được cập nhật thường xuyên thì sẽ kéo theo việc phải huấn luyện lại mô hình, phải tính toán và cập nhật lại mô hình sao cho vẫn giữ được hiệu quả tốt là một thách thức. Ngoài ra còn phải tránh trường hợp mô hình bị overfitting hoặc underfitting.

- Sự đa dạng, biến hoá của tin giả: Tin giả có thể xuất hiện dưới nhiều dạng khác nhau như bài viết, hình ảnh, video trên mạng xã hội và bằng nhiều ngôn ngữ khác nhau. Hơn nữa, các kỹ thuật tạo ra, phát tán tin giả ngày càng trở nên tinh vi, khiến cho việc phân biệt tin giả với tin thật trở nên khó khăn hơn.

1.3. Ứng dụng của bài toán

Trong nhiều năm qua, đã có nhiều nghiên cứu về tin giả và áp dụng bài toán nhận diện tin giả ở nhiều lĩnh vực khác nhau như khoa học máy tính, tâm lý học, ngôn ngữ học, truyền thông và khoa học nhận thức thần kinh [6][7]. Mỗi lĩnh vực sẽ có cách thức tiếp cận và phương pháp giải bài toán nhận diện tin giả khác nhau. Cũng có những nghiên cứu kết hợp nhiều lĩnh vực lại để nhận diện tin giả [7].



Hình 1.18. Tin giả trong nhiều lĩnh vực khác nhau

(Ảnh: Từ bài báo "Sustainable Development of Information Dissemination: A Review of Current Fake News Detection Research and Practice" của tác giả L. Yuan và cộng sự)

1.3.1. Ứng dụng trên mạng xã hội và truyền thông

Áp dụng bài toán nhận diện tin giả sẽ giúp hạn chế lại sự lan truyền tin giả trên mạng xã hội, nâng cao chất lượng thông tin, giúp cho người đọc không phải lo lắng hay nghi ngờ tính xác thực của thông tin. Tránh những ảnh hưởng tiêu cực tới người đọc

1.3.2. Ứng dụng vào giáo dục

Áp dụng bài toán nhận diện tin giả sẽ giúp học sinh, sinh viên hay giáo viên nâng cao chất lượng tìm kiếm tin tức, tránh đọc phải những thông tin sai sự thật. Giúp tìm kiếm tư liệu học tập nâng cao.

1.3.3. Ứng dụng vào pháp luật và an ninh quốc gia

Áp dụng vào để hạn chế phát tán tin giả, nâng cao an ninh của quốc gia, ngăn chặn việc lan truyền thông tin làm cho người dân bị xao động, gây ra hậu quả tiêu cực. Từ đó ngăn chặn được các vụ bạo động do tin tức sai lệch.

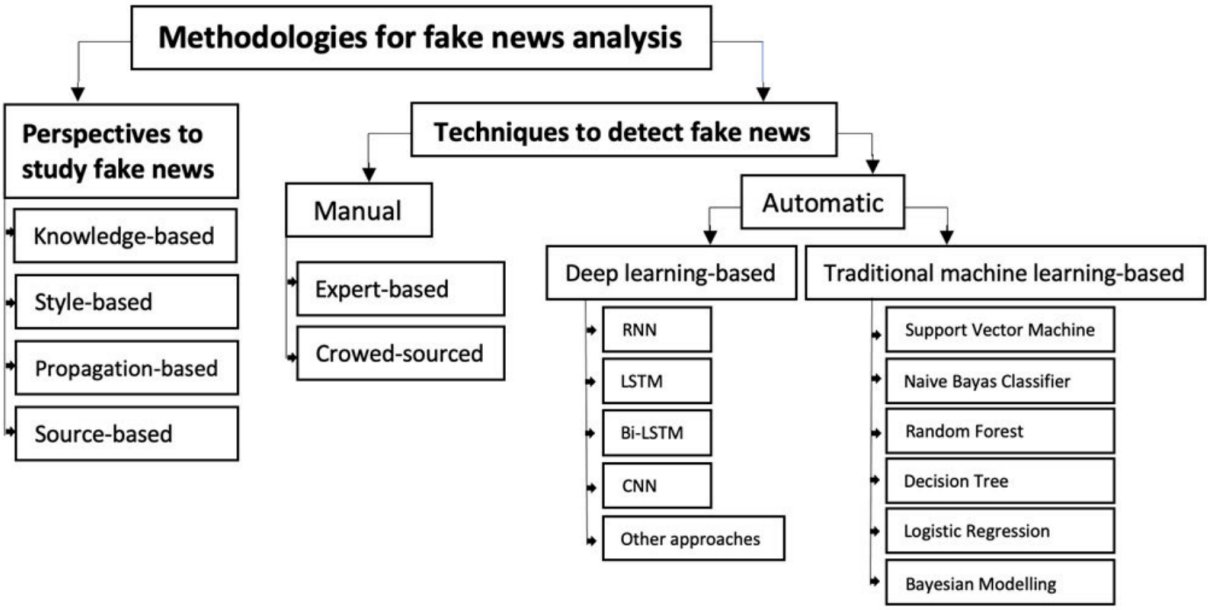
1.4. Khảo sát các hướng tiếp cận từ các công trình nghiên cứu liên quan

Có rất nhiều hướng tiếp cận để giải quyết bài toán nhận diện tin giả, hình 1.19 sẽ cho thấy được tổng quát về những hướng tiếp cận của bài toán này. Các phương pháp này được phân loại như sau.

Tin giả có thể được nghiên cứu theo bốn góc độ chính: Dựa trên kiến thức (Knowledge-based), dựa trên phong cách (Style-based), dựa trên sự lan truyền (Propagation-based), dựa trên nguồn tin (Source-based).

Ngoài ra, bài toán nhận diện tin giả có thể được thực hiện thủ công hoặc là tự động. Phương pháp thủ công bao gồm: Kỹ thuật dựa trên chuyên gia (Expert-based technique) hoặc kỹ thuật dựa trên cộng đồng (Crowded sourced based technique). Kỹ thuật thủ công phổ biến vào giai đoạn khi công nghệ chưa quá phát triển và khi kỹ thuật tự động chưa ra đời.

Khi công nghệ dần phát triển, các phương pháp hiện đại hơn cụ thể là kỹ thuật tự động ra đời. Kỹ thuật này được thực hiện bằng cách sử dụng phương pháp học sâu hoặc phương pháp học máy. Một số phương pháp học sâu có thể kể đến như RNN, LSTM, CNN, ... Còn học máy sử dụng một số giải thuật như SVM, KNN, RF, DT, LR, ...



Hình 1.19. Phân loại các phương pháp

(Ảnh: https://www.researchgate.net/figure/A-classification-of-the-overall-methodologies-in-fake-news-analysis_fig2_372363091)

1.4.1. Các góc nhìn về nghiên cứu tin giả

Trước khi bắt đầu vào nhận diện tin giả, người nghiên cứu phải hiểu rõ về nền tảng, về sự đa dạng của tin giả. Ngoài ra, việc hiểu về động cơ đằng sau việc lan truyền tin giả cũng quan trọng, hiểu về những tác động của tin giả. Sau khi đã nắm rõ về tin giả, các nhà nghiên cứu mới bắt đầu vào việc nghiên cứu cách thức phát hiện tin giả.Mỗi nhà nghiên cứu khác nhau sẽ có quan điểm, góc nhìn khác sau. Có bốn góc độ về nhận diện tin giả [6][8]. Có hai góc độ, tin giả được nghiên cứu trong quá trình tạo ra, là knowledge-based và style-based, còn hai góc độ còn lại, tin giả được nghiên cứu sau khi nó đã được lan truyền, propagation-based và source-based.

- **Knowledge-based**: Hay còn gọi là kiểm tra thực tế (fact-checking), gồm những công việc như thu thập dữ kiện thô từ những trang web. Tuy nhiên dữ liệu này cần phải được xử lý trước khi sử dụng để giải quyết các vấn đề như trùng lặp, không đáng tin cậy, không hợp lệ. Cần phải kiểm tra dữ kiện trước khi sử dụng, dữ kiện cần phải vượt qua 5 bài kiểm tra sau: giải quyết thực thể (entity resolution), ghi nhận thời gian (time recording), đánh giá tính nhất quán (consistency evaluation), đánh giá tính hoàn chỉnh (completeness evaluation), và đánh giá tính chính xác (accuracy evaluation) [6].

- **Style-based:** Mục đích là xem xét liệu tin tức có cố tình đánh lừa người đọc hay không dựa trên cách viết nội dung của tin tức đó (style). Kỹ thuật này là khả thi dựa trên một số đặc điểm nhất định, tin giả sẽ được xác định dựa trên lý thuyết, mẫu, chiến lược, nội dung và hình ảnh.

- **Propagation-based:** Nghiên cứu này cố gắng hiểu cách tin tức được lan truyền trên mạng nơi tin giả được tạo ra. Chủ yếu là nghiên cứu về cách một người dùng lan truyền tin giả. Đầu vào của nghiên cứu này có thể là một chuỗi tin tức (một biểu diễn trực tiếp của quá trình lan truyền tin tức), hay một đồ thị tự định nghĩa ra (biểu diễn gián tiếp của quá trình lan truyền tin tức).

- **Sourced-based:** Nhận diện tin giả bằng cách kiểm tra độ tin cậy từ những người tạo ra và lan truyền tin tức đó. Nghiên cứu này dựa trên việc người dùng tương tác với tin giả và vai trò của người dùng trong việc tạo ra tin giả và lan truyền nó trên mạng xã hội.

1.4.2. Các kỹ thuật để nhận diện tin giả

Hiện nay có rất nhiều kỹ thuật có thể nhận diện tin giả, nhưng không có kỹ thuật nào có thể hoàn toàn phân biệt được vì những hạn chế như thiếu dữ liệu, dữ liệu quá lớn, dữ liệu quá đa dạng, … Có 2 hướng tiếp cận chính trong việc nhận diện tin giả là: Thủ công hoặc tự động.

1.4.2.1. Kỹ thuật nhận diện thủ công

Tin tức sẽ được kiểm tra bằng cách thủ công bằng việc con người sẽ tự xác minh thông tin qua việc đọc và xem xét những thông tin trong tin đó. Hiện nay có rất nhiều trang web dùng để kiểm tra thông tin có đúng hay không, có thể dùng trang web Duke Reporter’s Lab để tìm kiếm các trang web trên. Quá trình kiểm tra thông tin thủ công được phân loại thành 2 loại:

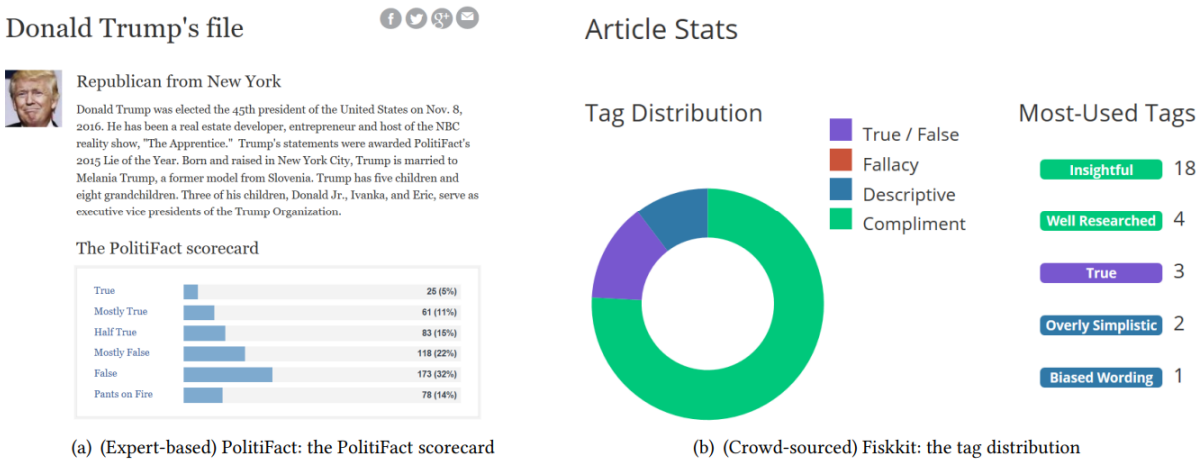
- **Kiểm tra thủ công dựa trên chuyên gia (Expert-based manual fact-checking)**

Ý tưởng chính của kỹ thuật này là dựa trên những chuyên gia trong lĩnh vực này để thực hiện bài toán nhận diện tin giả. Việc này được thực hiện bởi những nhóm chuyên gia nhỏ có độ tin cậy cao, vì sẽ dễ quản lý hơn. Đây là cách truyền thống để giải quyết bài toán, cách này sẽ khá tốn thời gian cũng như kinh phí [6]. Nhiều website có thể kể tới sử dụng kỹ thuật này như PolitiFact, FactCheck, Snopes, TruthOrFiction, …

- **Kiểm tra thủ công dựa trên cộng đồng (Crowd-sourced manual fact-checking)**

Cộng đồng ở đây là nhiều người bình thường không phải chuyên gia thực hiện viện xác thực thông tin. Vì là một cộng đồng nhiều người như vậy, nên kỹ thuật này cần thực hiện việc thanh lọc những người không đáng tin cậy, và khắc phục những xung đột xảy ra do mâu thuẫn lẫn nhau [6]. Kỹ thuật này sẽ khó duy trì cũng như không hiệu quả bằng kỹ thuật đầu tiên. Ví dụ trang web sử dụng kỹ thuật này là Fiskkit

Theo nhiều nghiên cứu, thì con người không giỏi trong việc đánh giá một thông tin là thật hay là giả nên kỹ thuật thủ công có độ chính xác không cao [9][10]. Độ chính xác rơi vào khoảng 46% tới 70%.



(a) (Expert-based) PolitiFact: the PolitiFact scorecard

(b) (Crowd-sourced) Fiskkit: the tag distribution

Hình 1.20. Ví dụ về kỹ thuật nhận diện thủ công

(Ảnh: https://ar5iv.labs.arxiv.org/html/1812.00315)

1.4.2.2. Kỹ thuật nhận diện tự động

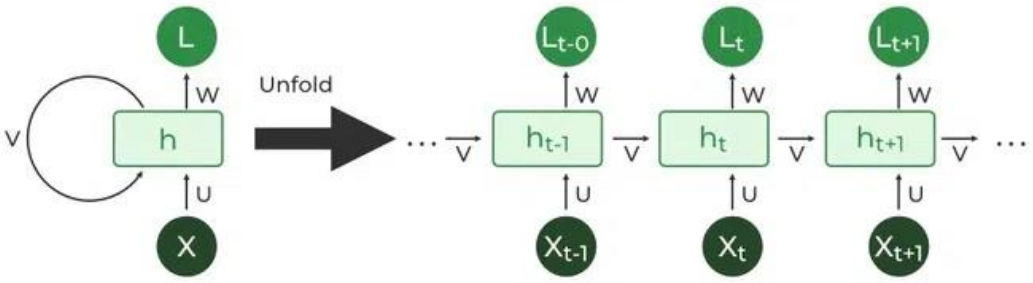
Kỹ thuật này sẽ giúp tiết kiệm thời gian cũng như chi phí hơn so với cách trên. Có 2 hướng chính để giải quyết bài toán theo hướng tự động là học sâu và học máy [2].

- **Học sâu (Deep Learning)**

Là một lĩnh vực thuộc trí tuệ nhân tạo (AI), sử dụng nhiều lớp neural networks, gọi là deep neural network để mô phỏng lại bộ não của con người. Phương pháp của kỹ thuật này là sử dụng các mạng nơ-ron chuyển tiếp sâu được gọi là mạng nơron truyền thẳng hoặc các perceptron nhiều lớp. Một số kỹ thuật để giải bài toán nhận diện tin giả như:

- **RNN**: Recurrent Neural Networks là phương pháp thường được dùng để giải quyết bài toán nhận diện tin giả đặc biệt là về nội dung của tin tức, lời nói hay

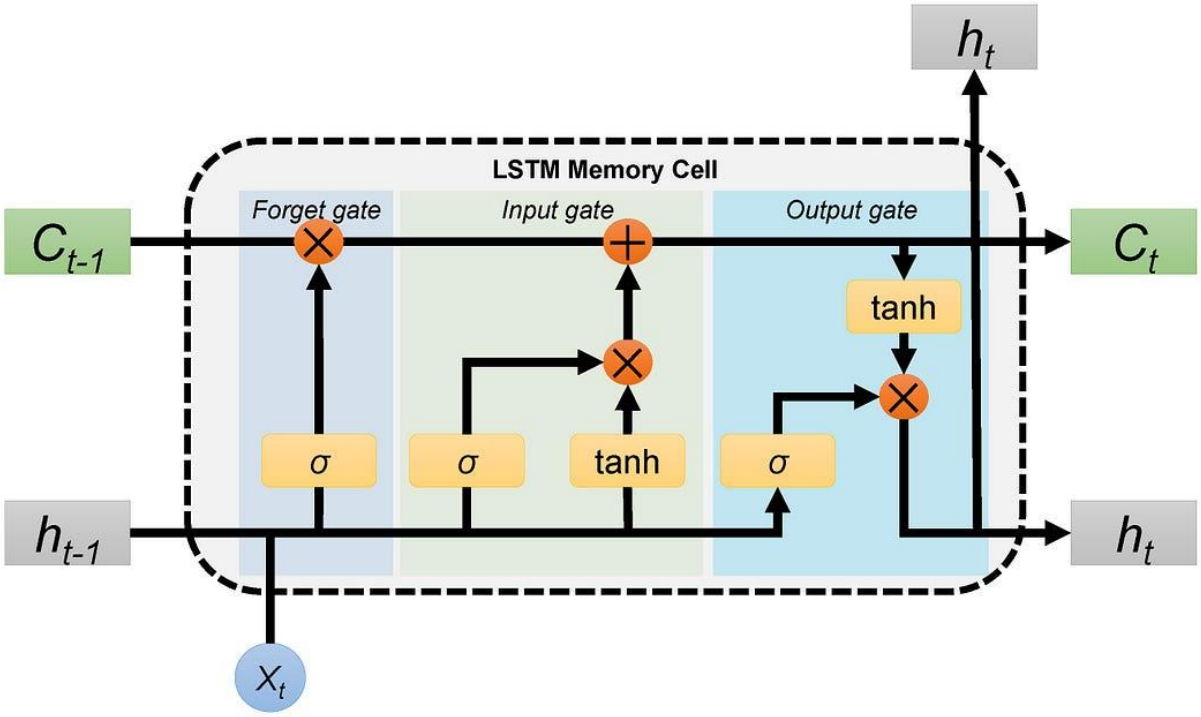
video. RNN coi dữ liệu đầu vào là một chuỗi liên tục và có thứ tự (Sequence), nối tiếp nhau theo thứ tự thời gian, cho phép nó học hỏi dữ liệu trước đó, nên sẽ rất hiệu quả khi sử dụng trong bài toán nhận diện tin giả [11]. RNN rất phù hợp để giải quyết vấn đề phát hiện tin tức giả, đặc biệt liên quan đến nội dung văn bản, lời nói và video. Ví dụ có một đoạn văn bản, có thể coi đoạn đó là một chuỗi các từ hoặc chuỗi các ký tự. Tại thời điểm t, với dữ liệu đầu vào X_t ta có kết quả output là L_t, L_t được sử dụng là input để tính kết quả output cho thời điểm (t+1). Mô hình hoạt động của RNN có thể được mô tả như hình sau



Hình 1.21. Cách xử lý của RNN

(Ảnh: https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/)

- **LSTM**: Long Short-Term Memory là một trong những công cụ mạnh nhất hiện nay dùng để giải quyết bài toán nhận diện tin giả. LSTM là một mạng cải tiến của RNN và có thể lưu trữ thông tin trong thời gian dài hơn và hợp hơn trong vấn đề ghi nhớ lại giá trị các lớp trước đó. Việc ghi nhớ thông tin trong thời gian dài là tính chất mặc định của LSTM, mỗi nút mạng có thể ghi nhớ được mà không cần bất kì can thiệp nào khác. Ngoài ra trong nhận diện tin giả, mạng LSTM còn được dùng để xác thực độ tin cậy của nguồn tin, bằng cách phân tích nội dung tìm các từ ngữ phổ biến. Mô hình xử lý của một nút mạng trong LSTM như sau:



Hình 1.22. Mô tả một nút mạng trong LSTM

(Ảnh: https://pub.towardsai.net/introduction-to-deep-learning-part-2-rnns-and-ltsm-fc65c230713d)

Trong đó, f_t, i_t, o_t tương ứng với forget gate, input gate và output gate

Forget gate: $f_t = \sigma(U_f * x_t + W_f + h_{t-1} + b_f)$

Input gate: $i_t = \sigma(U_i * x_t + W_i * h_{t-1} + b_i)$

Output gate: $o_t = \sigma(U_o * x_t + W_o + h_{t-1} + b_o)$

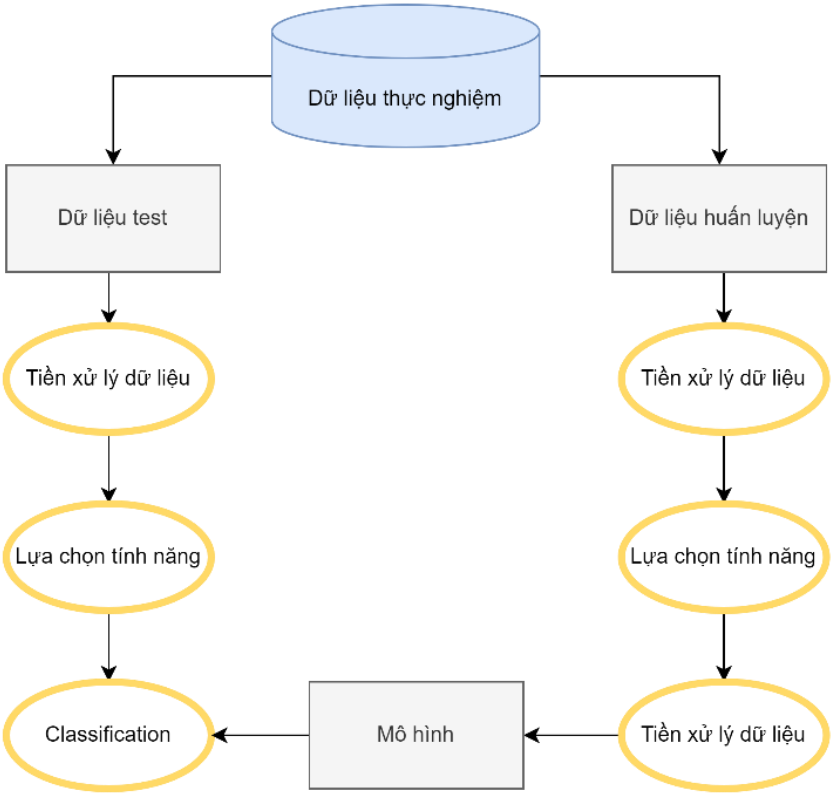
- **CNN**: Convolutional neural network là mạng nơ-ron đặc biệt được sử dụng để xử lý dữ liệu, những dữ liệu này được biểu diễn chính dưới dạng ma trận. Tận dụng khả năng trích xuất các thuộc tính của CNN, ta có thể áp dụng nó vào bài toán nhận diện tin giả. Mô hình CNN có thể xử lý dữ liệu lớn một cách nhanh chóng và chính xác. Giả sử dữ liệu trong bài toán là tập hơn m văn bản, mỗi văn bản được xử lý và biểu diễn dưới dạng vector n chiều. Như vậy dữ liệu đầu vào sẽ là m × n chiều

Có nhiều nghiên cứu đã kết hợp sử dụng CNN, RNN và LSTM lại để tận dụng những khả năng của mỗi kỹ thuật để có được kết quả tốt hơn. Ví dụ, nghiên cứu [11] đã đề xuất một hướng tiếp cận kết hợp giữa CNN và RNN cho bài toán nhận diện tin giả. Và

cho ra kết quả là 60% với dataset FA-KES, và 99% với dataset ISOT, cao hơn so với sử dụng riêng lẻ CNN hoặc RNN.

- **Học máy (Machine Learning)**

Cũng là một lĩnh vực thuộc trí tuệ nhân tạo (AI), cho máy tính khả năng tự học hỏi dựa trên dữ liệu đầu vào, từ đó đưa ra dự đoán. Đa số các thuật toán giải bài toán nhận diện tin giả theo hướng này đều là thuật toán học máy giám sát (Supervised Learning) để tạo và huấn luyện mô hình nhằm mục đích giải bài toán nhận diện tin giả. Mô hình tổng quát của hướng tiếp cận này như hình sau:



Hình 1.23. Mô hình học máy nhận diện tin giả

Có nhiều thuật toán được sử dụng để giải bài toán nhận diện tin giả như:

- **KNN**: Đây là một thuật toán học có giám sát đơn giản được sử dụng trong phân loại, dùng để phân loại các điểm mới bằng cách tìm điểm tương đồng giữa quan sát mới này với dữ liệu đã có. Ý tưởng cơ bản của thuật toán là dữ liệu tương tự nhau sẽ có khoảng cách gần nhau trong không gian, từ đó việc của chúng ta là tìm k điểm gần với dữ liệu cần test nhất. Công thức tìm khoảng cách giữa 2 điểm dữ liệu x, y có k thuộc tính:

Euclidean: $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan: $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski: $(\sum_{i=1}^{k}(|x_i - y_i|)^q)^{1/q}$

- **Decision Tree**: Đây là một thuật toán học máy hiệu quả trong vấn đề nhận diện tin giả, vì nó có khả năng phân tích và xử lý dữ liệu lớn nhanh chóng và chính xác. Một trong những đặc điểm chính của thuật toán này là nó có khả năng xác định mối quan hệ giữa các điểm dữ liệu. Ví dụ về bài toán nhận diện tin giả, nó có thể phân tích nguồn của tin tức, ngôn ngữ được sử dụng, nội dung và bối cảnh của tin tức đó, qua đó dự đoán được kết quả bài toán.

- **Naïve Bayes**: Thuật toán này dựa trên định lý Bayes, thuật toán có khả năng xử lý dữ liệu lớn và xác định các mẫu mà có thể có tin tức giả mạo. Ý tưởng của thuật toán đơn giản là giả sử rằng, một thuộc tính trong danh mục này không liên quan gì đến các thuộc tính khác. Ví dụ, trái cây sẽ được xác định là táo nếu có màu đỏ, hình dạng xoáy, đường kính khoảng 3 inches. Bất kể các thuộc tính này có phụ thuộc lẫn nhau hay phụ thuộc vào các thuộc tính khác hay không, Naïve Bayes giả định rằng những thuộc tính này đều riêng biệt.

Xác xuất Naïve Bayes:

$$P(c|x) = \frac{P(X|C)P(c)}{P(x)}$$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \ldots \times P(c)$$

Trong đó:

P(c|X): Là xác xuất hậu nghiệm.

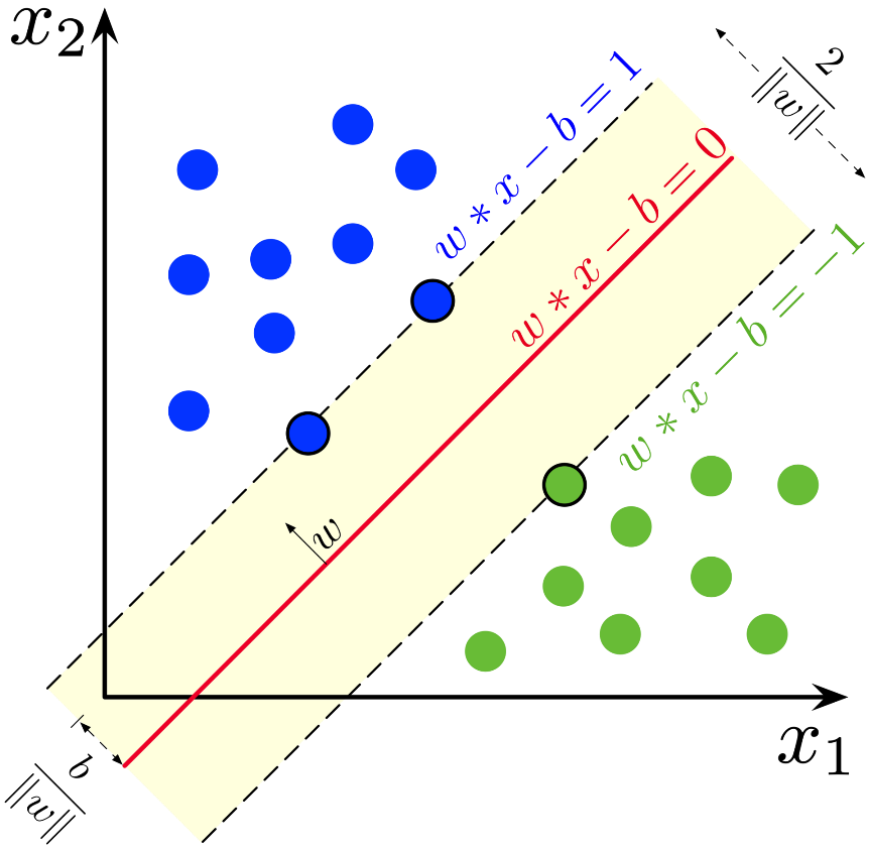
P(x|c): Độ tin cậy.

P(c): Xác xuất của lớp.

P(x): Xác xuất của x.

- **SVM**: Support Vector Machines được dùng trong bài toán nhận diện tin giả thông qua những đặc trưng như phong cách viết văn, nguồn tin tức và nội dung bài viết. Cách tiếp cận cơ bản để giải bài toán là cung cấp văn bản cần phân tích và đo lường các đặc điểm liên quan. Khi đã xác định được đặc điểm, SVM có thể giải bài toán nhận diện tin giả. Ý tưởng cơ bản là phân loại dữ liệu thành hai lớp

khác nhau, trong bài toán này là hai lớp tin giả và tin thật. Với một bộ dữ liệu mẫu thuộc hai thể loại cho trước, thuật toán huấn luyện SVM xây dựng một mô hình SVM để phân loại các mẫu test vào một trong hai mẫu trên. Thuật toán SVM chia hai lớp dữ liệu bằng một siêu mặt phẳng d–1 chiều khi số chiều của dữ liệu huấn luyện là d. Trong đó, wx–b=0 là siêu mặt phẳng thể hiện sự phân tách dữ liệu.



Hình 1.24. Mô hình phân lớp dữ liệu SVM

(Ảnh: https://en.wikipedia.org/wiki/Support_vector_machine)

Nhìn chung, Phương pháp được dùng nhiều hơn là học sâu, và nó cũng cho ra kết quả có độ chính xác cao hơn học máy. Tuy nhiên, việc phát triển và cài đặt một hệ thống dựa trên học sâu phức tạp và tốn kém hơn.

1.5. Dữ liệu thực nghiệm

Có hai loại dữ liệu thực nghiệm cho bài toán nhận diện tin giả được sử dụng phổ biến được chia ra như sau [12]:

- **Bài báo**

- **Polifact14** [13] là một trong những datasets sớm nhất được phát triển cho bài toán nhận diện tin giả với 221 mẫu, là tuyên bố (hay tiêu đề) có 5 nhãn: true, mostly true, half true, mostly false, false

- Horne và cộng sự [14] đã xây dựng nên 2 bộ dữ liệu: **Buzzfeed_political**, chủ đề chính là bầu cử tổng thống Mỹ năm 2016, được xây dựng từ dataset năm 2016 của Buzzfeed về tin tức bầu cử giả mạo. Bộ dữ liệu gồm 36 tin thật và 35 tin giả. Và **Random_political**, chủ đề về tin tức chính trị, gồm 75 tin thật, 75 tin giả và 75 tin tức châm biếm. Ngoài ra, còn nhiều bộ dữ liệu khác như **Celebrity** [9] là bộ dữ liệu được lấy từ tạp chí điện tử như Entertainment Weekly, People Magazine, RadarOnline, … gồm 250 tin thật và 250 tin giả, **LIAR** [15], ...

- **Bài đăng trên mạng xã hội**

- Những bộ dữ liệu tiêu chuẩn có thể kể tới như **Twitter15** và **Twitter16**. Ngoài ra còn có bộ dữ liệu như **FakeNewsNet** [16] là tập dữ liệu cho bài toán nhận diện tin giả được sử dụng nhiều trong nghiên cứu, tập dữ liệu gồm các bài viết và tweet liên quan đã được PolitiFact, GossipCop xác minh. Họ truy xuất 467 ngàn tweets của PolitiFact dataset và 1.25 triệu trong GossipCop, gắn nhãn real/fake. **BuzzFace** [17] là bộ dữ liệu lấy từ những bài viết trên Facebook, dựa trên bộ dữ liệu BuzzFeed, gồm 2,282 bài báo cùng với các thuộc tính (như lượt like youtube). Có 4 nhãn: no factual content, a mixture of true and false, mostly true, or mostly false. **COVID-19** [18] là bộ dữ liệu gồm những bài đăng trên mạng xã hội liên quan tới dịch COVID, bộ dữ liệu gồm khoảng hơn 10 ngàn dữ liệu với 2 nhãn là real/fake.

1.6. Tiêu chí đánh giá

Trong bài tiểu luận này, confusion matrix sẽ được sử dụng để đánh giá hiệu suất của bài toán nhận diện tin giả:

- True Positive (TP): Khi dự đoán chính xác là tin giả

- True Negative (TN): Khi dự đoán chính xác là tin thật

- False Positive (FP): Khi dự đoán là tin giả nhưng đó là tin thật

- False Negative (FN): Khi dự đoán là tin thật nhưng đó là tin giả

Các công thức để tính số liệu:

$$Precision = \frac{|TP|}{|TP| + |FP|}$$

$$Recall = \frac{|TP|}{|TP| + |FN|}$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$Accuracy = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|}$$

Những số liệu này thường được sử dụng trong học máy để đánh giá hiệu suất các thuật toán phân loại.

1.7. Tóm tắt chương 1

Trong chương 1, tôi đã giới thiệu về tin giả và bài toán nhận diện tin giả, xây dựng công thức cho bài toán, ứng dụng của bài toán vào nhiều lĩnh vực. Sau đó, tiến hành khảo sát một số hướng tiếp cận bài toán. Có 2 hướng chính để giải bài toán nhận diện tin giả là thủ công và tự động, kỹ thuật thủ công có độ chính xác không cao và chi phí cũng tốn nhiều hơn kỹ thuật tự động. Kỹ thuật tự động có hai hướng là học máy và học sâu, học sâu sẽ cho ra kết quả chính xác hơn nhưng lại khó khăn trong việc cài đặt và tốn chi phí hơn học máy. Đồng thời chương này cũng trình bày dữ liệu thực nghiệm của bài toán và những công thức để đánh giá kết quả thực nghiệm được sử dụng sau này.

CHƯƠNG 2. THUẬT TOÁN SVM GIẢI BÀI TOÁN NHẬN DIỆN TIN GIẢ

2.1. Tổng quan về thuật toán SVM

Support Vector Machine (SVM) là một thuật toán học máy, là một thuật toán có khả năng mạnh mẽ trong phân loại (classification) tuyến tính hoặc phi tuyến tính, hồi quy (regression) và còn các trong những việc nhận dạng. SVM có thể được dùng cho nhiều tác vụ khác nhau như phân loại văn bản, hình ảnh, phát hiện spam, phát hiện tin giả, nhận diện chữ viết tay, ...

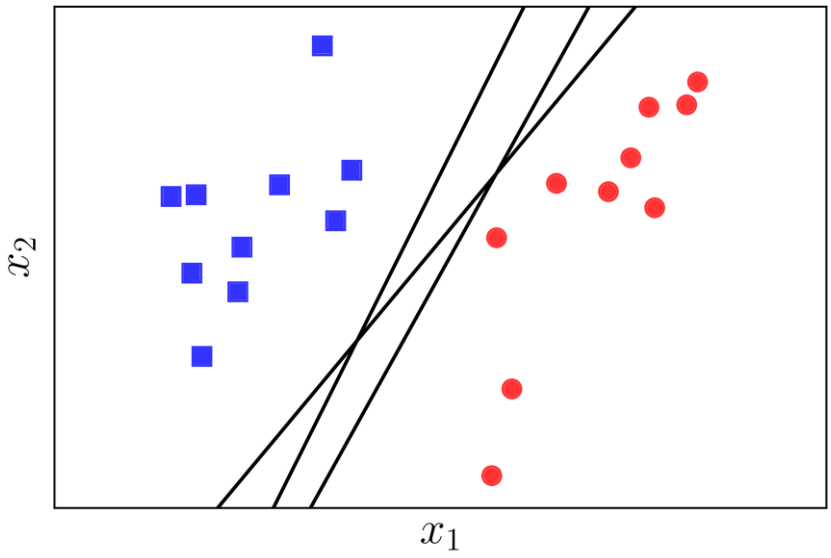
2.1.1. Giới thiệu về SVM

Support Vector Machine (SVM) lần đầu được nhắc tới vào năm 1992, được Boser, Guyon và Vapnik giới thiệu trong COLT-92 [19]. SVM là một thuật toán học máy, thuật toán học máy có 3 loại chính: Có giám sát (Supervised), không giám sát (Unsupervised) và học tăng cường (reinforcement learning) . SVM là thuật toán học máy có giám sát được sử dụng trong phân loại và hồi quy, nhưng được sử dụng chủ yếu trong việc phân loại nhằm muc tối đa hoá độ chính xác đồng thời tránh khỏi việc tình trạng over-fit.

Mục tiêu của thuật toán SVM là tìm được một hyperplane (siêu phẳng) trong không gian N chiều có khả năng phân tách các điểm dữ liệu thuộc các lớp khác nhau trong không gian đặc trưng. Hyperplane sẽ cố gắng để làm cho khoảng cách giữa các điểm thuộc các lớp khác nhau lớn nhất có thể. Nhờ vậy, SVM có khả năng phân chia các lớp dữ liệu một cách hiệu quả, đặc biệt là trong trường hợp dữ liệu có nhiều chiều, chiều của một siêu phẳng sẽ dựa trên số lượng đặc trưng đầu vào.

2.1.2. Cơ sở lý thuyết của SVM

Giả sử, ta có các điểm dữ liệu thuộc hai lớp khác nhau trong không gian N chiều, hai lớp này có thể phân biệt tuyến tính, nghĩa là sẽ tồn tại một hyperplane có thể phân tách chính xác hai lớp này như Hình 2.1 dưới đây.

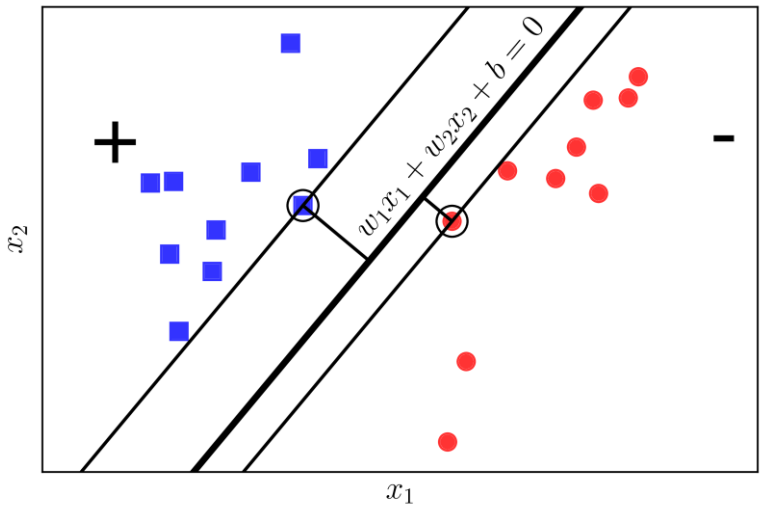


Hình 2.1. Các mặt phân cách hai lớp

(Ảnh: https://machinelearningcoban.com/2017/04/09/smv/)

Qua hình trên, có thể thấy rõ có nhiều mặt phân chia các điểm dữ liệu thành hai lớp màu xanh và màu đỏ. Vậy câu hỏi đặt ra ở đây là: trong vô số các mặt phân chia đó, đâu là mặt phân chia hay đâu là hyperplane phân tách các điểm dữ liệu tốt nhất?

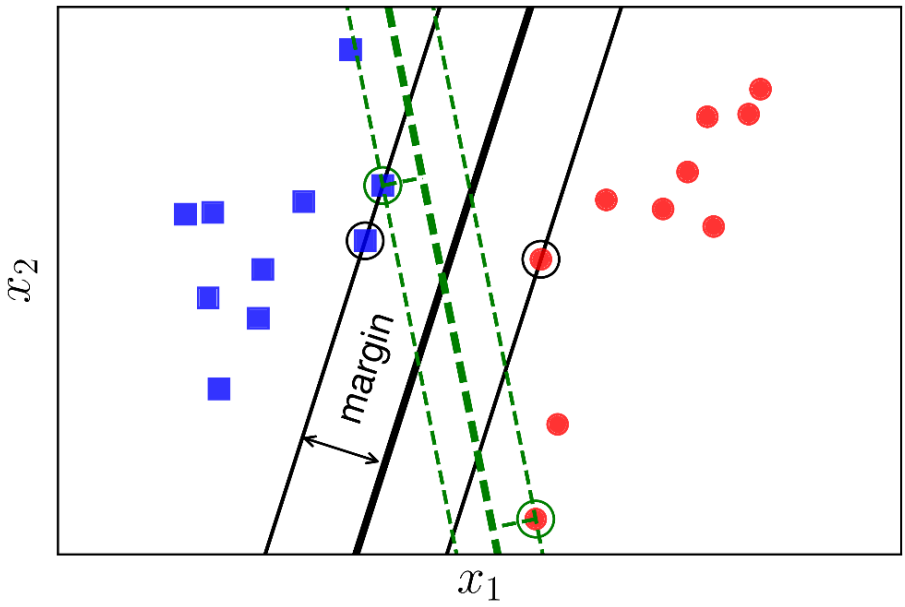
Để tìm ra được một hyperplane tốt nhất trong nhiều hyperplane, ta phải tìm theo một tiêu chuẩn nào đó. Hãy xem hình 2.2 dưới đây.



Hình 2.2. Điểm dữ liệu gần nhất của hai lớp

(Ảnh: https://machinelearningcoban.com/2017/04/09/smv/)

Thuật toán SVM sẽ tìm ra điểm dữ liệu gần hyperplane nhất từ cả hai lớp, những điểm dữ liệu này được gọi là vectơ hỗ trợ (những điểm khoanh tròn). Nhưng theo hình trên, hyperplane vẫn gần lớp màu đỏ hơn lớp màu xanh rất nhiều, chúng ta cần một hyperplane sao cho khoảng cách từ các vectơ hỗ trợ tới hyperplane là bằng nhau. Khoảng cách bằng nhau này được gọi là margin (lề), mục tiêu của SVM là tối đa hoá margin này. Hãy xem tiếp hình 2.3 dưới đây.



Hình 2.3. Margin của hai lớp bằng nhau và lớn nhất

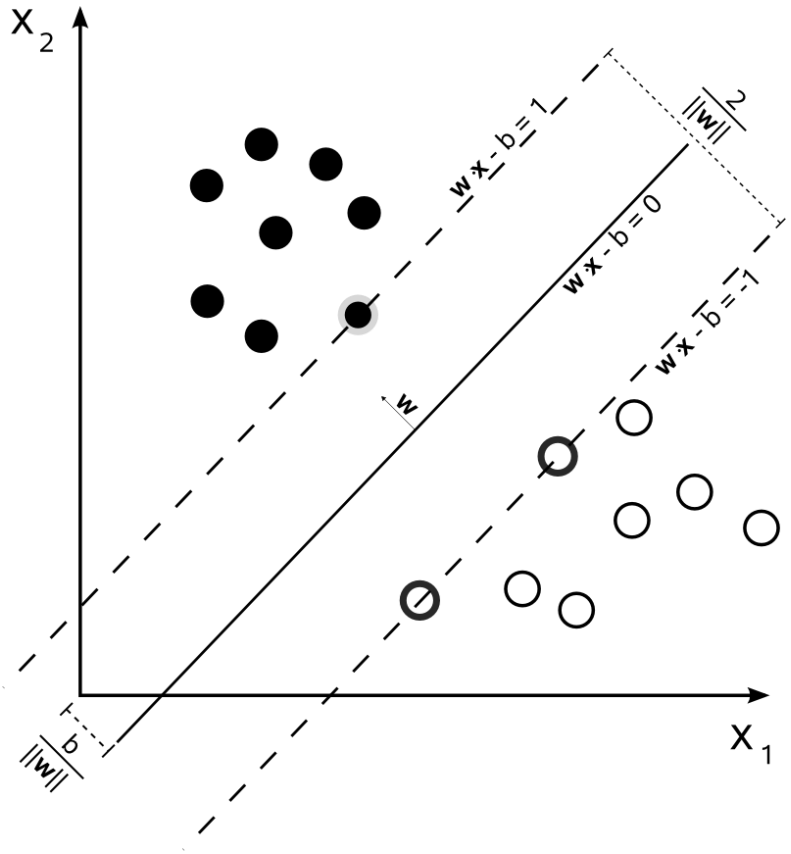
(Ảnh: https://machinelearningcoban.com/2017/04/09/smv/)

Ta có thể thấy được từ hình trên, khoảng cách từ các điểm dữ liệu gần nhau nhất tới hyperplane là như nhau. Giữa đường nét đứt màu xanh lục và đường nét liền màu đen, thì đường màu đen có margin rộng hơn. Margin rộng hơn sẽ mang lại hiệu quả phân tách tốt hơn, rạch ròi hơn. Vậy thì, đường màu đen sẽ là hyperplane tốt nhất, gọi là optimal hyperplane (hyperplane tối ưu).

Nếu bộ dữ liệu có khả năng phân biệt tuyến tính, ta có thể tìm ra hai hyperplane biên song song trên các vectơ hỗ trợ để tách hai lớp dữ liệu, sao cho khoảng cách giữa 2 hyperplane biên là lớn nhất. Khoảng không gian giới hạn bởi hai hyperplane biên đó được gọi là vùng biên, và optimal hyperplane chính là hyperplane song song và cách đều hai hyperplane biên.

2.1.3. Bài toán tối ưu cho SVM

Giả sử ta có một tập dữ liệu gồm N điểm là (x_1,y_1), (x_2,y_2), ..., (x_N,y_N) với vector $x_i \in R^d$ thể hiện dữ liệu đầu vào của dữ liệu và y_i nhận giá trị 1, −1 là nhãn của dữ liệu đó. Ta cần tìm ra hyperplane có margin lớn nhất mà phân tách các điểm x_i có $y_i = 1$ và các điểm x_j có $y_j = -1$ thành hai lớp.



Hình 2.4. Phân tích bài toán SVM

(Ảnh: https://en.wikipedia.org/wiki/File:Svm_max_sep_hyperplane_with_margin.png)

Bất kỳ một siêu mặt phẳng đều có thể được thể hiện dưới dạng tập hợp x thoả mãn điều kiện: $w^Tx - b = 0$ với w là vectơ pháp tuyến của hyperplane đó và tham số $\frac{b}{\|w\|}$ biểu diễn độ lệch của siêu mặt phẳng trên vector pháp tuyến tính từ gốc vector pháp tuyến đó.

Với một tập dữ liệu đã chuẩn hoá tiêu chuẩn hoá ta có thể xác định cặp hyperplane biên bằng hai phương trình sau:

$$w^Tx - b = 1$$

$$w^Tx - b = -1$$

Theo toán học, khoảng cách giữa 2 hyperplane biên này là $\frac{2}{\|w\|}$, nên nếu muốn tối đa hoá khoảng cách giữa hyperplane, ta phải cực tiểu hoá giá trị ||w||. Đồng thời, ta cũng

cần phải ngăn các điểm dữ liệu không cho vào vùng biên. Để làm thế, ta cần thêm các điều kiện sau:

Với mỗi i, ta có:

$w^T x_i - b \leq -1$ (những điểm dữ liệu nằm trên đường này thuộc nhãn 1)

$w^T x - b \geq 1$ (những điểm dữ liệu nằm dưới đường này thuộc nhãn −1)

Có thể viết gọn lại như sau: $y_i (w^T x_i - b) \geq 1 \; \forall i \in \{1, ..., n\}$.

Tóm lại, ta có bài toán tối ưu của SVM như sau:

$$\underset{w,\, b}{\text{minimize}} \quad \|w\|_2^2$$

subject to $y_i (w^T x - b) \geq 1 \; \forall i \in \{1, ..., n\}$.

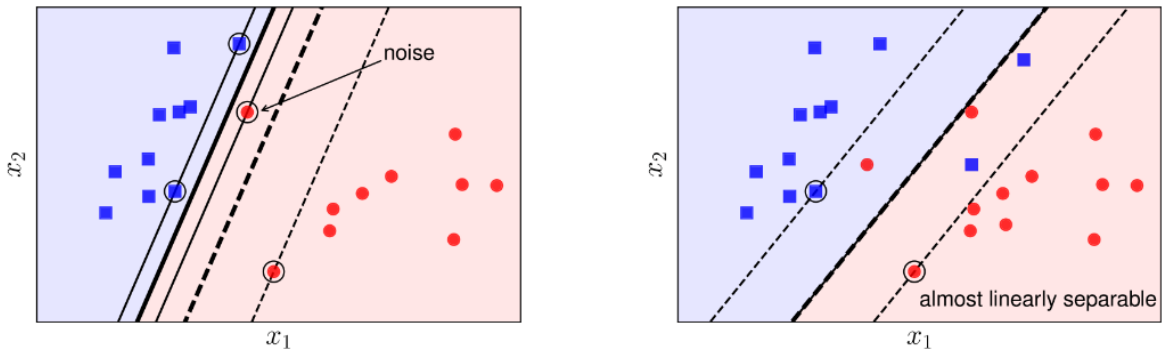
2.1.4. Các biến thể của SVM

SVM vừa trình bày bên trên hoạt động hiệu quả và có hiệu suất ổn trong trường hợp dữ liệu có thể phân biệt tuyến tính, nhưng nếu dữ liệu không thể phân biệt tuyến tính hoặc có dữ liệu nhiễu thì sao. Có 2 giải pháp để xử lý vấn đề dữ liệu nêu trên là:

- **Soft Margin SVM (SVM biên mềm)**

Soft Margin SVM là một biến thể của thuật toán SVM truyền thống để xử lý những trường hợp dữ liệu chứa nhiễu, gây khó khăn trong việc tìm ra optimal hyperplane. Thuật toán này cho phép SVM có sai số nhất định và giữ cho lề càng rộng càng tốt để các điểm khác vẫn có thể được phân loại chính xác. Nói một cách khác, nó cân bằng giữa việc phân loại sai và tối đa hóa lề.

Xét 2 ví dụ sau:



Hình 2.5. Soft Margin SVM

(Ảnh: https://machinelearningcoban.com/2017/04/13/softmarginsmv/)

- Trường hợp 1: Hình bên trái dữ liệu vẫn phân biệt tuyến tính nhưng có một điểm nhiễu màu đỏ nằm quá gần với lớp điểm màu xanh, Trong trường hợp này, nếu ta sử dụng SVM truyền thống thì margin được tạo ra sẽ rất nhỏ, hyperplane sẽ nằm quá gần lớp màu xanh và xa lớp màu đỏ. Ngược lại nếu như ta bỏ điểm nhiễu này đi thì sẽ có được một margin tốt hơn nhiều.
- Trường hợp 2: Hình bên phải dữ liệu chỉ gần như là phân biệt tuyến tính nên không thể sử dụng SVM truyền thống. Tuy nhiên, nếu như ta bỏ những điểm nằm ở gần biên của hai lớp, thì vẫn có thể tạo được một hyperplane khá tốt như đường nét đứt đậm. Các đường nét đứt mảnh vẫn tạo được một margin lớn. Với mỗi điểm nằm sang phía bên kia của các đường biên tương ứng, ta gọi điểm đó rơi vào vùng không an toàn.

Trong cả hai trường hợp trên, margin tạo bởi đường phân chia và đường nét đứt mảnh còn được gọi là soft margin. Theo cách gọi này, SVM truyền thống còn được gọi là Hard Margin SVM (SVM biên cứng). Phương pháp này sử dụng các biến slack ζ dùng để đo độ sai lệch

$$y_i (w^T x_i - b) \geq 1 - \zeta_i \quad \forall i \in \{1, ..., n\}$$

Ta sẽ có bài toán tối ưu ở dạng chuẩn cho Soft Margin SVM

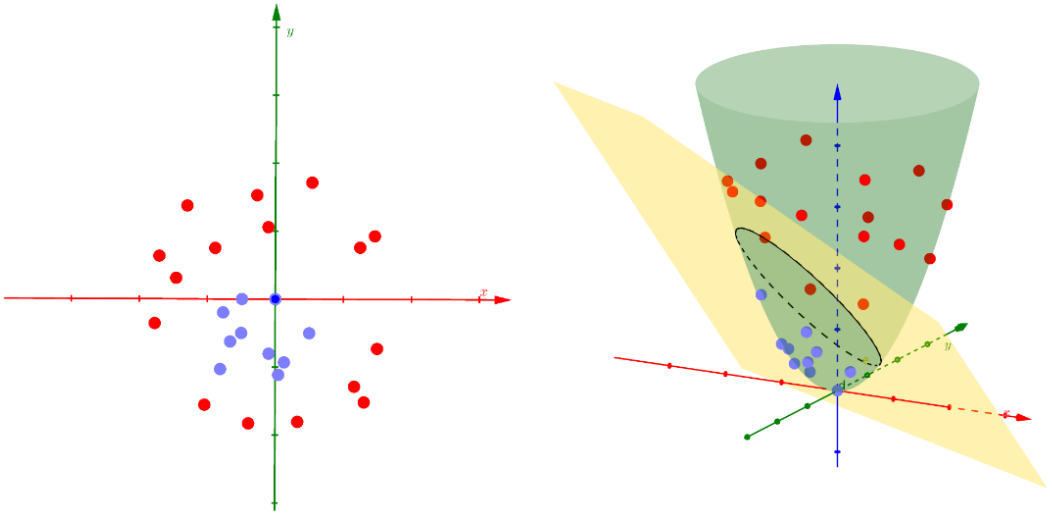
$$\underset{w,\, b}{\text{minimize}} \quad ||w||_2^2 + C \sum_{i=1}^{N} \zeta_i$$

$$\text{subject to} \quad y_i (w^T x - b) \geq 1 - \zeta_i \ , \zeta_i \geq 0 \quad \forall i \in \{1, ..., n\}.$$

Hằng số C được dùng để điều chỉnh tầm quan trọng giữa margin và sự hy sinh. Hằng số này được xác định từ trước bởi người lập trình.

- **Kernel SVM**

Ý tưởng cơ bản của Kernel SVM là tìm một phép biến đổi sao cho dữ liệu ban đầu là không phân biệt tuyến tính được sang không gian mới. Ở không gian mới này, dữ liệu sẽ trở nên phân biệt tuyến tính. Ví dụ như hình dưới đây.



Hình 2.6. Ví dụ về Kernel SVM

(Ảnh: https://machinelearningcoban.com/2017/04/22/kernelsmv/)

Hình bên trái, có thể thấy dữ liệu của hai lớp là không phân biệt tuyến tính trong không gian hai chiều. Nếu coi thêm chiều thứ ba là một hàm số của hai chiều còn lại z = $x^2 + y^2$, các điểm dữ liệu sẽ được phân bố trên 1 parabolic và đã trở nên phân biệt tuyến tính. Mặt phẳng màu vàng là hyperplane, có thể tìm được bởi Hard/Soft Margin.

Tóm lại, Kernel SVM là việc đi tìm một hàm số biến đổi dữ liệu x từ không gian ban đầu thành dữ liệu trong không gian mới bằng hàm:

$$\phi(x) = \phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{bmatrix}$$

Kết quả của phép biến đổi là một ma trận 3 chiều. Giờ chúng ta sẽ làm một ví dụ biến đổi 2 vectơ thực hiện nhân vectơ tạo thành.

$$\phi(a)^T \cdot \phi(b) = \begin{bmatrix} a_1^2 \\ \sqrt{2}a_1a_2 \\ a_2^2 \end{bmatrix} \cdot \begin{bmatrix} b_1^2 \\ \sqrt{2}b_1b_2 \\ b_2^2 \end{bmatrix} = a_1^2b_1^2 + 2a_1b1_1a_2b_2 = (a_1b1_1 + a_2b_2)^2$$

$$= \left(\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}^T \cdot \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}\right)^2 = (a^T \cdot b)^2$$

Hàm K(a, b) = $(a^T \cdot b)^2$ được gọi là kernel đa thức bậc 2. Trong học máy, kernel là hàm có khả năng tính được tích vô hướng mà chỉ dựa vào vectơ gốc.

Một số hàm kernel thông dụng:

- Linear: $K(a, b) = a^T \cdot b$

- Polynomial: $K(a, b) = (\gamma a^T \cdot b + r)^d$

- Gaussian RBF: $K(a, b) = \exp(-\gamma \|a^T - b\|^2)$

- Sigmoid: $K(a, b) = \tanh(\gamma a^T \cdot b + r)$

2.1.5. Ưu nhược điểm của SVM

- **Ưu điểm**

- Độ chính xác cao, đặc biệt là trong trường hợp dữ liệu có nhiều chiều.

- Hiệu quả với dataset nhỏ, vì dataset nhỏ sẽ yêu cầu số lượng nhỏ vectơ hỗ trợ.

- Có thể xử lý dữ liệu không phân biệt tuyến tính bằng cách sử dụng hàm kernel biến đổi dữ liệu

- Có thể sử dụng cho cả bài toán phân loại và hồi quy

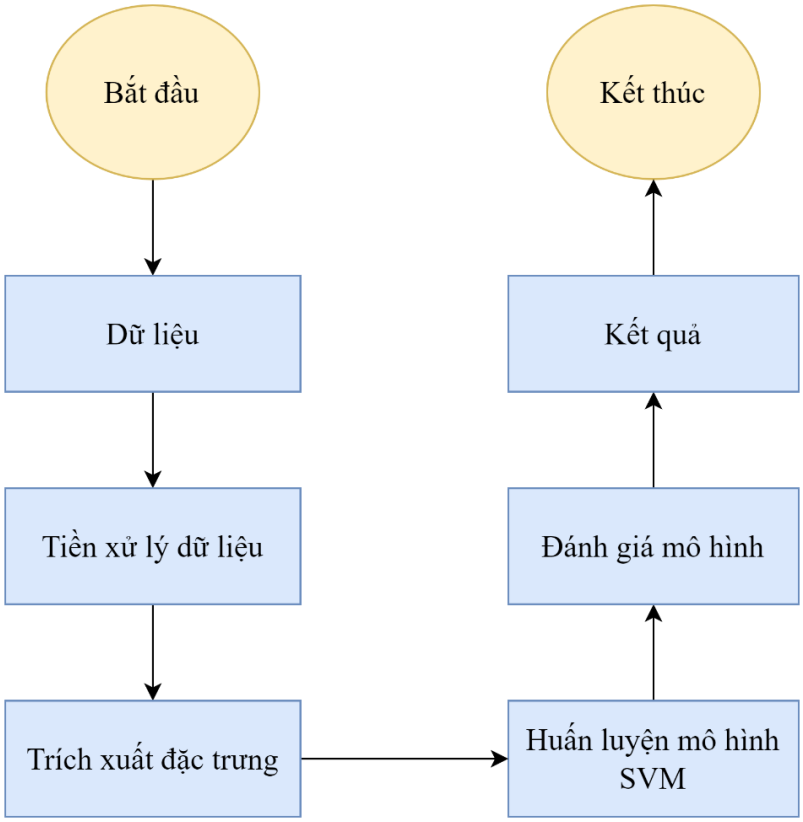
- **Nhược điểm**

-

- Việc xác định hàm kernel nào nên được sử dụng cũng là một khó khăn.

2.2. Xây dựng thuật giải cho bài toán nhận diện tin giả

2.2.1. Mô hình tổng quát



Hình 2.7. Mô hình tổng quát của bài toán

Mô hình tổng quát của bài toán nhận diện tin giả với thuật toán SVM gồm các bước như sau:

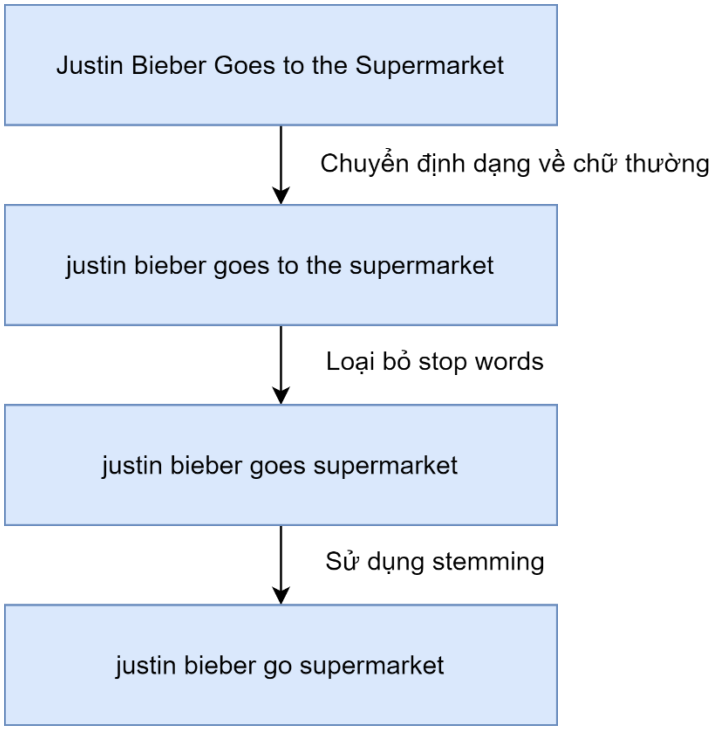
- Dữ liệu: Thu thập dữ liệu từ các trang web hoặc kiếm dữ liệu từ các bài báo, dữ liệu có label.
- Tiền xử lý dữ liệu: Thay đổi định dạng, loại bỏ stop words, tách từ, stemming để đưa từ về dạng cơ bản.
- Trích xuất đặc trưng: Trích lọc những đặc trưng ngôn ngữ cần thiết để phục vụ cho việc phân loại, nhận dạng nội dung.
- Huấn luyện mô hình SVM: Sử dụng các dữ liệu sau khi xử lý để huấn luyện mô hình SVM nhận diện tin giả.
- Đánh giá mô hình: Sử dụng tập dữ liệu test để kiểm tra đánh giá độ chính xác của mô hình, điều chỉnh tham số để đạt độ chính xác tốt nhất.

2.2.2. Tiền xử lý

Quá trình tiền xử lý dữ liệu bao gồm các bước sau:

- Thay đổi định dạng dữ liệu: Thay đổi định dạng của câu về chữ thường.

- Loại bỏ dấu câu/các ký tự không phải chữ số: Thay thế các ký tự không phải chữ số thành một khoảng trắng (spacebar).
- Tách token của từ: Tách input thành một tập hợp các token riêng biệt
- Loại bỏ stop words: Loại bỏ các từ không mang nhiều ý nghĩa nhưng thường xuất hiện như “i”, “me”, “am”, “is”, “are”, ...
- Chuẩn hoá từ sử dụng Stemming: Đưa từ về dạng cơ bản, ví dụ như “goes” thành “go”, “sitting” thành “sit”, ...
- Chia dữ liệu thành các tập con: Tách tập dữ liệu thành các tập training, test để sử dụng cho mô hình SVM



Hình 2.8. Tiền xử lý dữ liệu

2.2.3. Trích xuất đặc trưng

Sử dụng kỹ thuật TF-IDF để xác định đặc trưng. TF-IDF (Term Frequency - Inverse Document Frequency) là một phương thức thống kê được dùng để xác định độ quan trọng của một từ trong đoạn văn bản trong một tập nhiều văn bản khác nhau. Nó thường được sử dụng như một trọng số trong việc khai phá dữ liệu văn bản. TF-IDF chuyển đổi dạng biểu diễn văn bản thành dạng không gian vector

TF (Term Frequency): tần số xuất hiện của 1 từ trong 1 văn bản, tính như sau:

$$\text{tf(t, d)} = \frac{\text{count of t in d}}{\text{number of words in d}}$$

IDF (Inverse Document Frequency): dùng để đánh giá múc độ quan trọng của 1 từ trong băn bản. Khi tính tf mức độ quan trọng của các từ coi là như nhau. Tuy nhiên trong văn bản thường xuất hiện nhiều từ không quan trọng xuất hiện với tần suất cao như “is”, “a”, “this”, ... Do đó chúng ta cần giảm mức độ quan trọng của những từ đó bằng IDF.

$$idf(t, D) = log \frac{|D|}{|d \in D: t \in d|}$$

Với D là tổng số văn bản trong corpus D (gồm nhiều văn bản) và $|d \in D:t \in d|$ là số văn bản chứa từ t trong corpus D

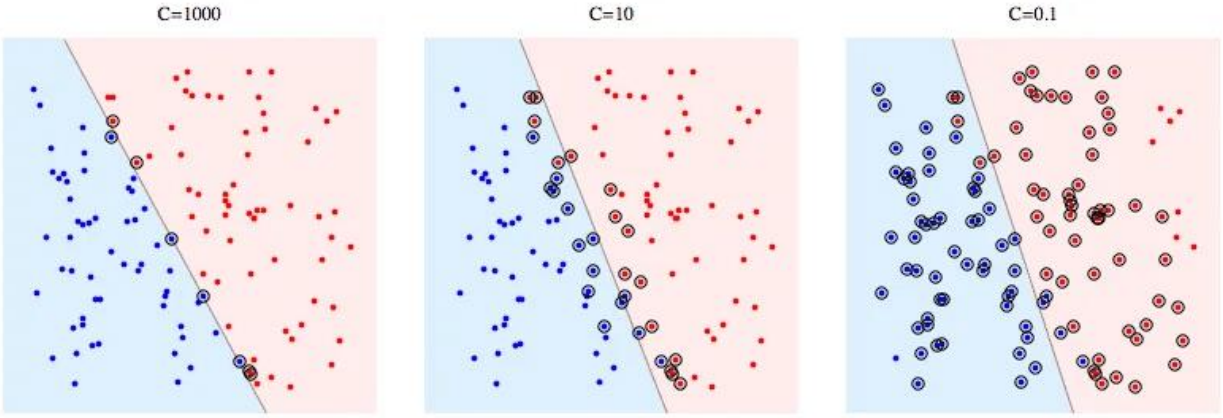
TF-IDF được tính như sau:

$$\text{TF-IDF} = \text{TF} \times \text{IDF}$$

2.2.4. Điều chỉnh thông số cho mô hình SVM

Các thông số quan trọng của mô hình SVM gồm:

- Kernel: phương pháp biến đổi dữ liệu đầu vào để đưa lên một chiều không gian cao hơn hiện tại. Các tham số kernel phổ biến của SVM như là: linear, polynominal, Gaussian RBF, Sigmoid.
- C: là hằng số điều chỉnh mức độ sai sót của thuật toán SVM. Nếu giá trị C càng lớn thì thuật toán càng tập trung vào độ chính xác khi phân loại dữ liệu, nhưng margin sẽ nhỏ và có nguy cơ bị overfitting. Giá trị C càng nhỏ thì SVM chấp nhận độ sai sót cao hơn, sẽ giúp mô hình tổng quát hóa tốt hơn, margin rộng hơn nhưng có thể bị underfitting.



Hình 2.9. Các giá trị của C

(Ảnh: https://trituenhantao.io/kien-thuc/svm-qua-kho-hieu-hay-doc-bai-nay/)

- Gamma: Tham số khi sử dụng kernel = RBF, Giá trị gamma càng lớn, mô hình sẽ phụ thuộc nhiều vào các điểm gần nhau, tăng độ nhạy cảm khi huấn luyện, có nguy cơ bị overfitting. Gamma thấp thì sẽ có độ nhạy cảm thấp hơn, mô hình sẽ tổng quát hơn.
- Probability: Có cho phép tính toán xác xuất đầu ra hay không, đặt probability=True, mô hình sẽ tính xác suất dự đoán.

2.3. Tóm tắt chương 2

Chương 2 giới thiệu về tổng quan về thuật toán SVM, bài toán tối ưu của thuật toán và các biến thể của thuật toán như Soft Margin SVM và Kernel SVM, cuối cùng là nói về ưu nhược điểm của thuật toán. Sau đó, tiến hành xây dựng thuật toán để giải bài toán nhận diện tin giả. Đầu tiên là bước thu thập dữ liệu, sau đó tới bước tiền xử lý dữ liệu, bước này thay đổi định dạng về chữ thường, loại bỏ stopwords, stemming và chia dữ liệu thành tập training và test. Còn phần trích xuất đặc trưng dữ liệu sẽ sử dụng kỹ thuật TF-IDF. Cuối cùng là huấn luyện cho mô hình SVM và sử dụng mô hình cho bài toán nhận diện tin giả.

CHƯƠNG 3. THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

3.1. Bộ dữ liệu thực nghiệm

Bộ dữ liệu cho thực nghiệm là bộ dữ liệu "COVID-19 Fake News Dataset" được P. Patwa và cộng sự giới thiệu trong bài nghiên cứu [18]. Dữ liệu được lấy về từ trang github của tác giả (https://github.com/parthpatwa/covid19-fake-news-detection). Dữ liệu gồm các cột chính là: id, tweet và label. Dữ liệu bao gồm 10700 tin, với 5600 tin thật và 5100 tin giả.

	A	B	C
1	id	tweet	label
2	1	The CDC currently reports 99031 deaths. In general the discrepancies in death	real
3	2	States reported 1121 deaths a small rise from last Tuesday. Southern states re	real
4	3	Politically Correct Woman (Almost) Uses Pandemic as Excuse Not to Reuse Plas	fake
5	4	#IndiaFightsCorona: We have 1524 #COVID testing laboratories in India and as	real
6	5	Populous states can generate large case counts but if you look at the new case	real
7	6	Covid Act Now found "on average each person in Illinois with COVID-19 is infe	real
8	7	If you tested positive for #COVID19 and have no symptoms stay home and aw	real
9	8	Obama Calls Trumpâ€™s Coronavirus Response A Chaotic Disaster https://t.cc	fake
10	9	???Clearly, the Obama administration did not leave any kind of game plan for	fake
11	10	Retractionâ€"Hydroxychloroquine or chloroquine with or without a macrolide	fake
12	11	Take simple daily precautions to help prevent the spread of respiratory illnes	real
13	12	The NBA is poised to restart this month. In March we reported on how the Uta	fake
14	13	We just announced that the first participants in each age cohort have been do	real
15	14	#CoronaVirusUpdates #IndiaFightsCorona More than 6 lakh tests done for 3rd	real
16	15	Protect yourself and others from #COVID19 when using public transportation.	real
17	16	As of 18 August 2020 8AM till now there have been a total of 4687 #COVID19 p	real
18	17	Because of Donald Trump's negligence and incompetence:	fake
19	18	#IndiaFightsCorona India continues to scale new peaks in #COVID19 tests! Mo	real
20	19	We just announced that we have shipped vials of mRNA-1273 the Companyâ€	real
21	20	Multiple Facebook posts claim that â€œAussies will be fined if they are found	fake
22	21	No Nobel Prize laureate Tasuku Honjo didn't say the coronavirus is "not natura	fake
23	22	The NZ COVID Tracker app will remain important and useful at Alert level 1. Pe	real
24	23	BREAKING NEWS# The president Cryill Ramaphosa has asked all foreign nation	fake
25	24	We are delighted that 78 high- and upper-middle countries and economies ha	real
26	25	Very intriguing possibility that mood changes and anxiety may be a function of	real
27	26	Elon Musck To New Baby; Get A Job Kid! https://t.co/bc8Re0Ai3Y #christmas #	fake
28	27	_A new alcohol-free sanitizer has been developed by the Dedan Kimathi Unive	fake
29	28	Just Appendix B gathering all the state orders on testing was a huge chunk of v	real
30	29	Yesterday our laboratories completed 2899 tests of those 726 were testing of p	real
31	30	Reusing #N95 masks? NIH found vaporized hydrogen peroxide UV light and he	real
32	31	.@realDonaldTrump has shifted his focus at different moments in the #Coron	fake
33	32	Doctored image of President Donald Trump shared claiming he asked people t	fake
34	33	This #FourthOfJuly weekend if you choose to spend time outdoors at an event	real
35	34	CDC Recommends Mothers Stop Breastfeeding To Boost Vaccine Efficacy	fake

Hình 3.1. Dữ liệu thực nghiệm

3.2. Các thư viện được sử dụng

Phần thực nghiệm này sẽ sử dụng một số thư viện như trong hình 3.2 sau đây.

```
import pandas as pd
import re # for removing punctuation
import nltk
from nltk.corpus import stopwords # stop words
nltk.download('stopwords')
from nltk.tokenize import word_tokenize # tokenizing word
from nltk.stem import PorterStemmer # stemming
from sklearn.feature_extraction.text import TfidfVectorizer # TF-IDF
from sklearn.model_selection import train_test_split # train/test split dataset
# SVM model
from sklearn.svm import SVC
from sklearn import metrics
from sklearn.metrics import accuracy_score, f1_score, precision_score, recall_score, confusion_matrix
# Graph
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

Hình 3.2. Các thư viện được sử dụng trong thực nghiệm

- Thư viện pandas được sử dụng để xử lý, phân tích dữ liệu trong python.
- Thư viện re, nltk được sử dụng cho bước tiền xử lý, cụ thể là để loại bỏ các dấu câu, tách token từ, loại bỏ stop words và cho kỹ thuật stemming.
- Thư viện sklearn gồm nhiều module khác nhau, module sklearn.feature_extraction.text để sử dụng kỹ thuật TF-IDF, module sklearn.model_selection dùng để tách bộ dữ liệu ra thành 2 tập train và test, module sklearn.svm để sử dụng mô hình SVM và cuối cùng là module svm.metrics để tính toán vẽ confusion matrix và tính toán các số liệu như accuracy, precision, recall và f1 score.
- Thư viện matplotlib, seaborn và numpy dùng để vẽ các biểu đồ về hiệu suất của model.

3.3. Tiền xử lý dữ liệu

Các bước xử lý cũng như sử dụng mô hình SVM để giải bài toán nhận diện tin giả sẽ được thực hiện trên máy tính cá nhân window 11, python phiên bản 3.12.2 cùng với framework jupyter trên IDE là Visual Studio Code.

Bước này sẽ xử lý dữ liệu trước khi đưa vào huấn luyện cho mô hình, đã được nêu ở mục 2.2.2 của chương 2. Cụ thể ở mục này gồm các bước như làm sạch dữ liệu bằng cách đưa dữ liệu về định dạng chữ thường, loại bỏ những ký tự không phải chữ số, tách token từ, loại bỏ stopwords và cuối cùng là áp dụng kỹ thuật Stemming. Sau đó loại bỏ bớt các cột đặc trưng không cần thiết, ở đây là cột id và cột tweet. Kết quả cuối cùng là cột final_tweet ở hình 3.3 bên dưới.

	label	final_tweet
0	real	cdc current report 99031 death gener discrep d...
1	real	state report 1121 death small rise last tuesda...
2	fake	polit correct woman almost use pandem excus re...
3	real	indiafightscorona 1524 covid test laboratori i...
4	real	popul state gener larg case count look new cas...
5	real	covid act found averag person illinoi covid 19...
6	real	test posit covid19 symptom stay home away peop...
7	fake	obama call trump coronaviru respons chaotic di...
8	fake	clearli obama administr leav kind game plan so...
9	fake	retract hydroxychloroquin chloroquin without m...
10	real	take simpl daili precaut help prevent spread r...
11	fake	nba pois restart month march report utah jazz ...
12	real	announc first particip age cohort dose phase 2...
13	real	coronavirusupd indiafightscorona 6 lakh test d...
14	real	protect other covid19 use public transport pra...
15	real	18 august 2020 8am till total 4687 covid19 pos...
16	fake	donald trump neglig incompet 110 000 peopl die...
17	real	indiafightscorona india continu scale new peak...
18	real	announc ship vial mrna 1273 compani vaccin nov...
19	fake	multipl facebook post claim aussi fine found t...

Hình 3.3. Dữ liệu cột “final_tweet” là dữ liệu sau bước tiền xử lý

3.4. Chia dữ liệu thành các tập con

Tập dữ liệu sau bước này sẽ được chia thành 2 tập con là tập train và tập test với tỉ lệ như sau:

- Tập train: 70%
- Tập test: 30%

Thư viện sklearn cung cấp một hàm để ta có thể tách tập dữ liệu thành tập train và tập test là hàm train_test_split() nhận vào các tham số cơ bản như:

- Test_size: Giá trị trong khoảng 0.0 tới 1.0, cho biết tỉ lệ dữ liệu sẽ được chia cho tập test. Ở phần thực nghiệm này test_size = 0.3 là 30%.
- Train_size: Giá trị trong khoảng 0.0 tới 1.0, cho biết tỉ lệ dữ liệu sẽ được chia cho tập train. Ở phần thực nghiệm này train_size = 0.7 là 70%.
- Random_state: Tham số này nhận vào một giá trị số nguyên, tham số này đặt một hạt giống (seed) cho hàm, mục đích là mỗi lần chạy lại hàm thì vẫn cho ra cùng một kết quả giống nhau.

3.5. Trích xuất đặc trưng sử dụng TF-IDF

Sử dụng thư viện sklearn để import lớp "TfidfVectorizer" và khởi tạo một đối tượng tên là "vectorizer". Sau đó sử dụng phương thức fit_transform() để áp dụng kỹ thuật TF-IDF lên 2 tập dữ liệu train và test.

Các tham số cơ bản của TfidfVectorizer là:

- ngram_range: Tham số này xác định kích thước của n-gram được sử dụng để tạo vectơ. N_gram là chuỗi liên tiếp nhiều từ trong văn bản, mặc định tham số này là ngram_range = (1,1), là chỉ lấy từ đơn lẻ (unigram).
- min_df: Tham số này đặt mức tần suất tối thiểu mà một từ phải xuất hiện trong dữ liệu để được bao gồm trong vectơ. Mặc định là min_df = 1.
- max_df: Tham số này đặt mức tần suất tối đa mà một từ có thể xuất hiện trong dữ liệu để được bao gồm trong vectơ. Mặc định là max_df = 1.0.

```
(0, 2105)     0.2515856344556419   (0, 3266)     0.21177509084381874
(0, 4645)     0.1656897186875163   (0, 4117)     0.24191216482516759
(0, 4742)     0.14050458501624274  (0, 4833)     0.06990431928656247
(0, 4833)     0.11236656160947148  (0, 5061)     0.30363487841655684
(0, 5089)     0.08262054021033971  (0, 5089)     0.10279806625328004
(0, 6624)     0.19548619175719884  (0, 5100)     0.19195162539420774
(0, 6936)     0.25452931260608097  (0, 5168)     0.21367224550468017
(0, 7623)     0.2698204018764472   (0, 5326)     0.20119648050361236
(0, 8006)     0.11231411728652743  (0, 5976)     0.24322768213203996
(0, 8206)     0.3122025719514993   (0, 6759)     0.23939839544851382
(0, 10377)    0.265302907079033    (0, 7719)     0.1611277706314055
(0, 10717)    0.15606862835596616  (0, 8006)     0.06987169316858452
(0, 10763)    0.3122025719514993   (0, 8483)     0.26234870734845867
(0, 11433)    0.1364260276046991   (0, 9449)     0.2691896374129386
(0, 11963)    0.15831908669757466  (0, 10167)    0.2875135366072167
(0, 12453)    0.1441867050997481   (0, 11095)    0.24744050794079578
(0, 13058)    0.1615077402628641   (0, 11235)    0.23012777166504422
(0, 13467)    0.2229207370039809   (0, 11433)    0.1697438892153911
(0, 14579)    0.25452931260608097  (0, 14642)    0.2353310697077915
(0, 14668)    0.14478128906786575  (0, 14668)    0.1801396663341008
(0, 15818)    0.23107948016984778  (0, 15189)    0.16606187911395368
(0, 16122)    0.335509242550581    (0, 15801)    0.19059052808529578
(1, 604)      0.08369169652653052  (1, 604)      0.14515690452324317
(1, 4287)     0.2993604902606114   (1, 3745)     0.5613841818462654
(1, 4833)     0.11649451024700029  (1, 4117)     0.3496105163545271
..                                 ...
(7489, 13824) 0.1937754785023075   (3209, 14455) 0.08444130168010781
(7489, 13926) 0.3481394700739914   (3209, 14969) 0.12544216840093508
(7489, 14111) 0.21097655765871848  (3209, 15801) 0.09322139392205446
(7489, 15801) 0.17866303854036217  (3209, 15822) 0.10787006186483888
```

Hình 3.4. Kết quả sau khi áp dụng TF-IDF lên tập dữ liệu train và test

3.6. Tham số của hàm SVM

Bước này đã được nêu ở phần 2.2.4 của chương 2, thư viện sklearn cung cấp lớp "SVC" để sử dụng model SVM, lớp đó sẽ nhận vào các tham số cơ bản như sau:

- Kernel = 'linear': Chỉ định sử dụng loại kernel tuyến tính, nghĩa là sẽ sử dụng mô hình SVM tuyến tính.
- C = 1: Đặt giá trị tham số C bằng 1. Tham số này điều chỉnh mức độ sai sót của mô hình SVM, nếu tham số C quá cao sẽ dễ bị overfitting, quá nhỏ thì sẽ dễ bị underfitting.
- Probability = True: Đặt tham số bằng true để bật hỗ trợ ước tính xác suất cho mô hình SVM

3.7. Huấn luyện mô hình SVM

Bắt đầu thực hiện huấn luyện mô hình SVM, sử dụng trên dữ liệu train (Đặt là X_train, Y_train). X_train là ma trận đặc trưng, mỗi hàng của ma trận ứng với một hàng dữ liệu và Y_train là vectơ nhãn ứng với dữ liệu của X_train.

Trước tiên phải khởi tạo một đối tượng "svm_model" từ lớp "SVC" của thư viện "sklearn". Sau đó, gọi phương thức svm_model.fit(X_train, Y_train) để bắt đầu huấn luyện, mô hình SVM sẽ được điều chỉnh cho phù hợp với dữ liệu huấn luyện.

```
svm_model = SVC(kernel='linear', C=1, probability=True)
svm_model.fit(X_train, Y_train)
```

Hình 3.5. Khởi tạo mô hình SVM và sử dụng phương thức fit()

Sau khi quá trình huấn luyện hoàn tất có thể sử dụng phương thức predict() để dự đoán cho dữ liệu test (X_test, Y_test) hoặc dữ liệu test khác.

```
prediction = svm_model.predict(X_test)
```

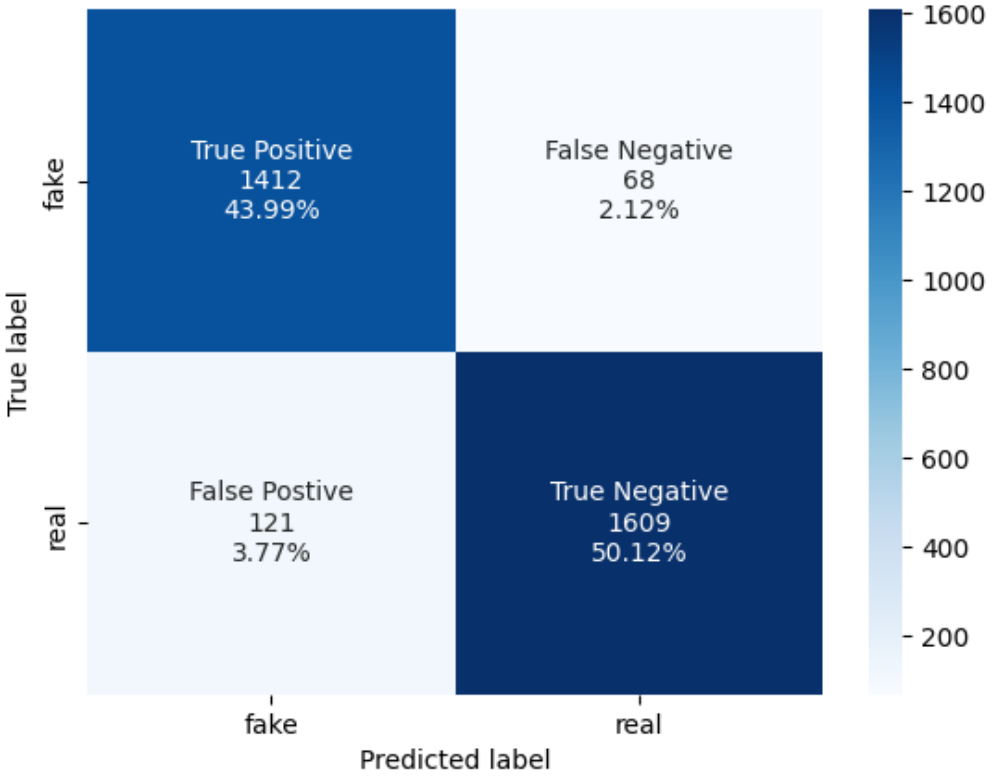
Hình 3.6. Sử dụng phương thức predict để dự đoán

Thời gian huấn luyện mô hình với bộ dữ liệu COVID-19 là 11s với độ chính xác trên tập train là xấp xỉ 99,1%, còn trên tập test là xấp xỉ 94,1%.

3.8. Đánh giá kết quả

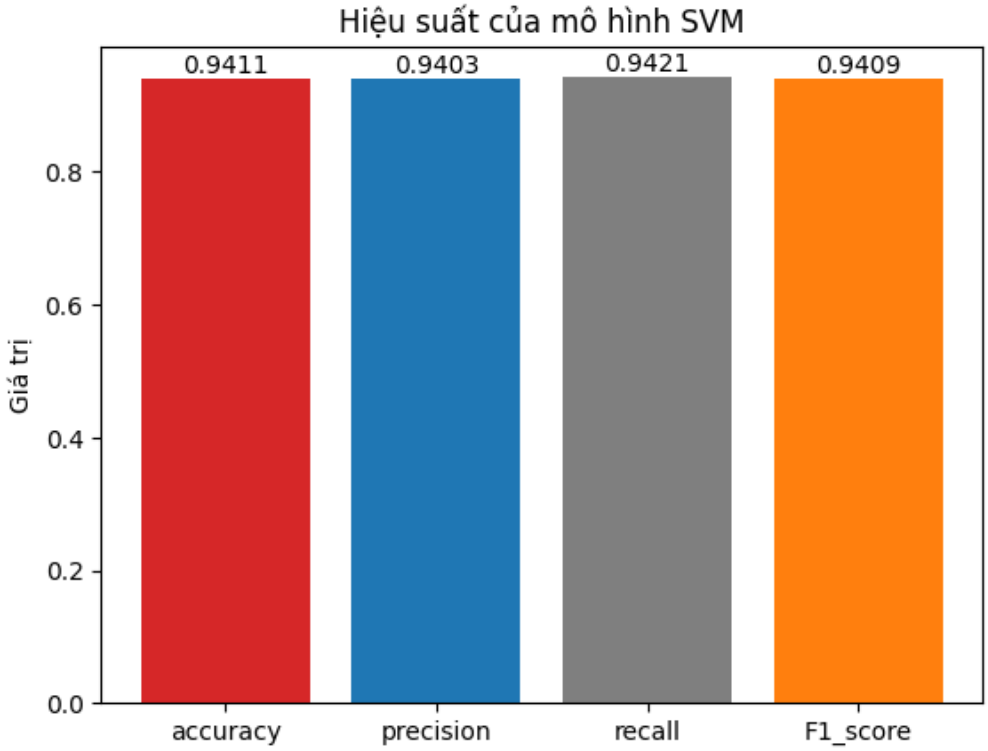
Sau khi huấn luyện mô hình SVM và dự đoán trên dữ liệu test ta thu được kết quả các số liệu đã được nêu ở mục 1.6 chương 1 như sau:

- Sử dụng confusion matrix cho các giá trị để thể hiện các giá trị TP, TN, FP, FN



Hình 3.7. Confusion Matrix cho các giá trị TP, TN, FP, FN

- Các số liệu đánh giá như accuracy, precision, recall, F1_score.



Hình 3.8. Biểu đồ hiệu suất của mô hình SVM

Kết quả đánh giá cho thấy mô hình SVM đạt được hiệu suất tốt với bộ dữ liệu COVID-19, với độ chính xác là xấp xỉ 94,1%, cho thấy được mô hình có độ chính xác cao trong việc phân loại tin giả.

Ngoài độ chính xác ra còn các số liệu khác như precision, recall, F1_score được tính toán để đánh giá hiệu suất. Chỉ số precision là tỉ lệ số lượng tin giả trong tổng số những tin được dự đoán là tin giả. Chỉ số recall là tỉ lệ số lượng tin giả trong tổng số những tin trên thực tế là tin giả. Chỉ số F1_score là số bình quân điều hoà (harmonic mean) của precision và recall.

Để thấy được rõ hơn sự hiệu quả của mô hình SVM, tôi sẽ so sánh mô hình với các thuật toán khác sử dụng cùng một bộ dữ liệu COVID-19 của các tác giả khác như:

- Trong bài nghiên cứu của tác giả Patwa cùng các cộng sự [18], độ chính xác của các thuật toán DT, LR và GBDT lần lượt là 85,2%, 92,8% và 86,8%
- Trong bài nghiên cứu của tác giả Felber [20], sử dụng các thuật toán LR, RF và NB có độ chính xác lần lượt là 95,4%, 90,8% và 93,3%
- Trong bài nghiên cứu của tác giả Rawat và Kanojia [21], sử dụng thuật toán học máy là LR có chỉ số precision là 95,3%. Và còn một số thuật toán học sâu như BERT, RoBERTa, XLNet với chỉ số precision lần lượt là 96,1%, 99,1% và 99,2%.
- Một bài nghiên cứu khác của tác giả Ayoub, Yang và Zhou [22], họ sử dụng những thuật toán học máy là LR và RF với độ chính xác là 93% và 91%. Các thuật toán học sâu như BERT và DistilBERT (Logistic) với độ chính xác là 99,3% và

Tóm lại, qua đánh giá trên, có thể thấy được mô hình SVM đã đạt được hiệu suất tốt với bài toán nhận diện tin giả, với các chỉ số vượt hơn đa số các thuật toán học máy khác. Tuy nhiên, mô hình SVM vẫn thua kém khi so sánh độ chính xác với các thuật toán thuộc phương pháp học sâu. Nhưng vẫn có thể thấy được mô hình SVM có tiềm năng lớn khi áp dụng vào giải bài toán nhận diện tin giả là một cách hiệu quả khi sử dụng để ngăn chặn, xử lý tin giả trong hiện nay.

3.9. Tóm tắt chương 3

Chương 3 nói về quá trình thực nghiệm trên bộ dữ liệu COVID-19 được tạo ra bởi tác giả Patwa cùng các cộng sự. Quá trình thực nghiệm bao gồm các bước tiền xử lý, chia tập dữ liệu thành 2 tập train và test, áp dụng kỹ thuật TF-IDF để trích xuất đặc trưng, sau đó là huấn luyện mô hình SVM cho bài toán nhận diện tin giả và cuối cùng là đánh giá kết quả thực nghiệm thu được, sau đó đem kết quả so sánh với một số thuật toán khác cũng sử dụng cùng bộ dữ liệu COVID-19. Nhìn từ kết quả thực nghiệm cũng như khi so sánh với các thuật toán khác, có thể thấy được Support Vector Machine (SVM) là thuật toán có tiềm năng lớn trong việc giải bài toán nhận diện tin giả.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Sau quá trình thực nghiệm, kết quả của việc áp dụng mô hình SVM để giải bài toán nhận diện tin giả cho thấy được hiệu suất, tiềm năng trong vấn đề này là rất cao. Từ kết quả thực nghiệm, có thể thấy được việc dùng mô hình SVM mang lại độ chính xác lên tới xấp xỉ 94,4% với bộ dữ liệu COVID-19 cùng với các chỉ số precision, recall và F1_score để đánh giá tổng quan hơn về mô hình.

Nhưng việc sử dụng mô hình SVM trong bài toán nhận diện tin giả vẫn còn gặp phải nhiều khó khăn và bài nghiên cứu vẫn có thể phát triển thêm trong tương lai. Hướng nghiên cứu kế tiếp có thể là việc mở rộng bộ dữ liệu thực nghiệm ra nhiều hơn nữa, như là tăng số lượng tin tức thu thập được, hay là ngoài việc phân tích nội dung tin thì có thể phân tích cả thông tin khác như tác giả bài đăng, lượt thích, lượt chia sẻ, lượt bình luận, … Việc thu thập thêm nhiều dữ liệu sẽ giúp cho mô hình trở nên toàn diện hơn, phân tích sẽ chính xác hơn trong nhiều trường hợp, ngôn ngữ khác nhau.

Tóm lại, bài toán nhận diện tin giả trong thời đại ngày nay là một vấn đề cần thiết và cấp bách. Sử dụng mô hình SVM trên sẽ mở ra nhiều hướng phát triển hơn cho trong lĩnh vực của bài toán, đồng thời cũng sẽ giúp cải thiện chất lượng thông tin, đóng góp vào việc bảo vệ an ninh mạng, hỗ trợ người dùng tiếp cận được những thông tin chất lượng và hạn chế sự lan truyền của tin giả trên các phương tiện truyền thông.

TÀI LIỆU THAM KHẢO

[1]. K. S. Amorim, M. Miranda, "Misinformation, disinformation, and malinformation: clarifying the definitions and examples in disinfodemic times", *Encontros Bibli Revista Eletrônica de Biblioteconomia e Ciência da Informação*, 2021.

[2]. V. T. Hùng, N. K. Chi, T. A. Kiệt, "Phát hiện tự động tin giả: Thành tựu và thách thức", *Tạp chí Khoa học và Công nghệ - Đại học Đà Nẵng*, Tập 20, Số 3, 2022

[3]. Bộ Thông tin và Truyền thông, "Cẩm nang phòng chống tin giả, tin sai sự thật trên không gian mạng", 2022

[4]. J. Brainard, P. R. Hunter, "Misinformation making a disease outbreak worse: outcomes compared for influenza, monkeypox, and norovirus", *Simulation: Transactions of the Society for Modeling and Simulation International*, Vol. 96, 2020

[5]. K. Shu, A. Silva, S. Wang, J. Tang, H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective", *ACM SIGKDD Explorations Newsletter*, Vol. 19, No. 1, 2017.

[6]. X. Zhou, R. Zafarani, "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities", *ACM Computing Surveys*, Vol. 53, No. 5, 2020.

[7]. X. Zhou, A. Jain, V. Phoha, R. Zafarani, "Fake News Early Detection: An Interdisciplinary Study", *Digital Threats: Research and Practice*, Vol. 1, 2020.

[8]. M. Tajrian, A. Rahman, M. A. Kabir, MD. R. Islam, "A Review of Methodologies for Fake News Analysis", *IEEE Access*, Vol. 15, 2023

[9]. V. P. Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea, "Automatic Detection of Fake News", *Proceedings of the 27th International Conference on Computational Linguistics*, 2018.

[10]. Y. Yao, B. Viswanath, J. Cryan, H. Zheng, B Y. Zhao, "Automated Crowdturfing Attacks and Defenses in Online Review Systems", *Proceedings*

of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017.

[11]. J. A. Nasir, O. S. Khan, I. Varlamis, "Fake news detection: A hybrid CNN-RNN based deep learning approach", *International Journal of Information Management Data Insights*, Vol. 1, No. 1, 2021.

[12]. T. Murayama, "Dataset of Fake News Detection and Fact Verification: A Survey, 2021.

[13]. A. Vlachos, S. Riedel, "Fact Checking: Task definition and dataset construction", *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 2014.

[14]. B. D. Horne, S. Adah, "This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News", *Proceedings of the International AAAI Conference on Web and Social Media*, Tập 11, 2017

[15]. W. Y. Wang, "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection", *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vol. 2, 2017.

[16]. K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, "FakeNewsNet: A Data Repository with News Content, Social Context and Spatiotemporal Information for Studying Fake News on Social Media", *Big data*, Vol. 8, No. 3, 2019.

[17]. G. C. Santia, J. R. Williams, "BuzzFace: A News Veracity Dataset withFacebook User Commentary and Egos", *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*, 2018.

[18]. P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, Md. S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, "Fighting an Infodemic: COVID-19 Fake News Dataset", *Communications in Computer and Information Science*, Vol. 1402, 2021

[19]. B. E. Boser, I. M. Guyon, V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers", *Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT'92)*, 1992.

[20]. T. Felber, “Constraint 2021: Machine Learning Models for COVID-19 Fake News Detection Shared Task”, 2021

[21]. M. Rawat, D. Kanojia, “Automated Evidence Collection for Fake News Detection”, *Proceedings of the 18th International Conference on Natural Language Processing*, 2021

[22]. J. Ayoub, X. J. Yang, F. Zhou, “Combat COVID-19 Infodemic Using Explainable Natural Language Processing Models”, 2021