BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC CẦN THƠ
TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

NIÊN LUẬN
NGÀNH AN TOÀN THÔNG TIN

Đề tài:

NGHIÊN CỨU PHÁT HIỆN TIN GIẢ TRÊN MẠNG XÃ HỘI DỰA TRÊN ĐẶC TRƯNG CẢM XÚC KÉP

Sinh viên thực hiện:
LÊ THANH SANG
MSSV: B2203732
KHÓA 48

Cần Thơ, 12/2025

BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC CẦN THƠ
TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG
KHOA MẠNG MÁY TÍNH VÀ TRUYỀN THÔNG

NIÊN LUẬN
NGÀNH AN TOÀN THÔNG TIN

Đề tài:

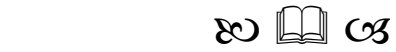
NGHIÊN CỨU PHÁT HIỆN TIN GIẢ TRÊN MẠNG XÃ HỘI DỰA TRÊN ĐẶC TRƯNG CẢM XÚC KÉP

Giảng viên hướng dẫn:
TS NGUYỄN HỮU VÂN LONG

Sinh viên thực hiện:
LÊ THANH SANG
MSSV: B2203732
KHÓA 48

Cần Thơ, 12/2025

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN



.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

MỤC LỤC

CHƯƠNG 1: GIỚI THIỆU 1
1.1. Lý do chọn đề tài. 1
1.2. Mục tiêu của đề tài. 2
1.3. Đối tượng nghiên cứu. 3
1.4. Phạm vi nghiên cứu 3
1.5. Phương pháp nghiên cứu. 4
1.6. Ý nghĩa khoa học và thực tiễn của đề tài. 4
1.7. Bố cục niên luận. 4
CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 6
2.1. Tổng quan về phát hiện tin giả trên mạng xã hội. 6
2.1.1. Khái niệm 6
2.1.2. Đặc điểm của tin giả trên mạng xã hội 7
2.1.3. Các phương pháp tiếp cận hiện nay 7
2.2. Tập dữ liệu PHEME. 8
2.2.1. Giới thiệu tập dữ liệu PHEME 8
2.2.2. Các sự kiện được thu thập 8
2.2.3. Cấu trúc dữ liệu của PHEME 9
2.3. Cảm xúc kép (Dual Emotion). 9
2.3.1. Khái niệm cảm xúc trong phân tích tin giả 9
2.3.2. Khái niệm cảm xúc kép (Dual Emotion) 10
2.3.3. Vai trò của cảm xúc kép trong phát hiện tin giả 10
2.3.4. Cảm xúc kép trong tập dữ liệu PHEME 11
2.3.5. Ý nghĩa của cảm xúc kép trong đề tài nghiên cứu 11
2.4. Các mô hình trích xuất đặc trưng cảm xúc 12
2.4.1. Khái quát về trích xuất đặc trưng cảm xúc 12
2.4.2. Trích xuất đặc trưng cảm xúc dựa trên từ điển 12
2.4.3. Trích xuất đặc trưng cảm xúc dựa trên biểu diễn vector văn bản 12
2.4.4. Trích xuất đặc trưng cảm xúc bằng Sentence Embedding 13
2.4.5. Kết hợp đặc trưng cảm xúc trong mô hình cảm xúc kép 13
2.5. Các mô hình học máy sử dụng trong đề tài. 14
2.5.1. Logistic Regression 14
2.5.2. Support Vector Machine (SVM) với kernel RBF 14
2.5.3. Random Forest 15

2.5.4. XGBoost 15
CHƯƠNG 3: PHƯƠNG PHÁP THỰC HIỆN 17
3.1. Tải và tiền xử lý dữ liệu 17
3.1.1. Thu thập dữ liệu từ tập PHEME 17
3.1.2. Tiền xử lý văn bản 18
3.2. Trích xuất cảm xúc văn bản 18
3.2.1. Mô hình phân tích cảm xúc sử dụng 18
3.2.2. Cảm xúc nguồn và phản hồi 19
3.3. Xây dựng đặc trưng cảm xúc kép 19
3.4. Huấn luyện mô hình và cân bằng dữ liệu 19
3.5. Huấn luyện và tối ưu mô hình học máy 20
3.5.1. Tối ưu 20
3.5.2. Chuẩn bị mô hình và tham số 20
3.5.3. Huấn luyện mô hình 21
3.6. Đánh giá mô hình 21
CHƯƠNG 4: KẾT QUẢ THỰC NGHIỆM 21
4.1. Thu thập dữ liệu. 21
4.2. Môi trường thực nghiệm. 22
4.3. Thư viện và công cụ được sử dụng: 22
4.4. Kết quả thực nghiệm. 23
4.4.1. Kết quả đánh giá: 23
4.4.2. Biểu đồ so sánh: 25
CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 26
5.1. Kết luận. 26
5.2. Hạn chế. 27
5.3. Hướng phát triển. 27
Tài liệu tham khảo 28

LỜI CẢM ƠN

Lời đầu tiên, tôi xin chân thành cảm ơn Trường Đại học Cần Thơ, Khoa Mạng máy tính và truyền thông dữ liệu đã tạo điều kiện cho tôi thực hiện đề tài này.

Tôi xin cảm ơn gia đình và bạn bè đã luôn ủng hộ và đồng hành cùng tôi trong suốt thời gian học tập. Tôi xin chân thành cảm ơn các anh chị và thầy cô đã giúp đỡ, đóng góp ý kiến và chỉ bảo tôi để tôi có thể hoàn thành nghiên cứu. Đặc biệt, tôi xin gửi lời cảm ơn sâu sắc đến thầy TS – Nguyễn Hữu Vân Long người đã tận tình bỏ thời gian hướng dẫn, chỉ bảo giúp tôi hoàn thiện đề tài: "Nghiên cứu phát hiện tin giả trên mạng xã hội dựa trên đặc trưng Cảm xúc kép".

Với điều kiện thời gian cũng như kinh nghiệm còn hạn chế, niên luận này không thể tránh được những thiếu sót. Tôi rất mong nhận được sự chỉ bảo, đóng góp ý kiến của các thầy cô để tôi có điều kiện bổ sung, nâng cao kiến thức của mình, phục vụ tốt hơn công tác thực tế sau này.

Tôi xin chân thành cảm ơn!

Lê Thanh Sang

Lớp An toàn thông tin A1 – K48

TÓM TẮT

Sự bùng nổ của mạng xã hội đã biến các nền tảng trực tuyến thành kênh thông tin chủ đạo, nhưng đồng thời cũng tạo điều kiện cho tin giả lan truyền nhanh chóng, gây ảnh hưởng tiêu cực đến nhận thức cộng đồng và an ninh thông tin. Các phương pháp phát hiện tin giả truyền thống hiện nay chủ yếu tập trung vào khai thác nội dung văn bản để tìm kiếm các đặc điểm ngôn ngữ bất thường. Tuy nhiên, kỹ thuật ngụy tạo tin giả ngày càng tinh vi, thường sử dụng văn phong bắt chước tin thật để đánh lừa các bộ lọc nội dung, khiến việc chỉ dựa vào phân tích văn bản đơn thuần trở nên thiếu hiệu quả và dễ bỏ sót các trường hợp tinh vi.

Để khắc phục hạn chế trên, niên luận này đề xuất một hướng tiếp cận mới dựa trên lý thuyết "Cảm xúc kép" (Dual Emotion Features), tập trung khai thác mối quan hệ tương tác tâm lý giữa người đưa tin và độc giả. Phương pháp này không chỉ phân tích cảm xúc chủ quan của người đăng (Source) mà còn đối chiếu với phản ứng cảm xúc thực tế của cộng đồng (Replies). Bằng cách xem xét sự mâu thuẫn hoặc cộng hưởng giữa hai luồng cảm xúc này, nghiên cứu mong muốn tìm ra những đặc trưng tiềm ẩn quan trọng, giúp nhận diện tin giả chính xác hơn ngay cả khi nội dung văn bản được che đậy khéo léo.

ABSTRACT

The rapid expansion of social media has transformed online platforms into primary information channels, yet it simultaneously facilitates the swift spread of misinformation, negatively impacting public perception and information security. Traditional detection methods predominantly focus on mining textual content to identify linguistic anomalies. However, fake news fabrication techniques have become increasingly sophisticated, often mimicking the style of legitimate news to bypass content filters, rendering text-only analysis ineffective and prone to overlooking subtle cases.

To address these limitations, this study proposes a novel approach based on "Dual Emotion Features" theory, focusing on exploiting the psychological interaction between publishers and readers. This method not only analyzes the subjective emotions of the publisher (Source) but also contrasts them with the actual emotional responses of the community (Replies). By examining the dissonance or resonance between these two emotional streams, the research aims to uncover latent features that enhance fake news detection accuracy, even when textual content is deceptively camouflaged.

CHƯƠNG 1: GIỚI THIỆU

1.1. Lý do chọn đề tài.

Trong thập kỷ qua, sự bùng nổ của Internet và các nền tảng mạng xã hội (như Twitter, Facebook, Reddit) đã thay đổi hoàn toàn cách con người tiếp cận và chia sẻ thông tin. Mạng xã hội đã trở thành kênh tin tức chủ đạo nhờ tốc độ lan truyền nhanh chóng và khả năng tương tác cao. Tuy nhiên, sự thuận tiện này cũng đi kèm với một hệ quả nghiêm trọng: sự gia tăng không kiểm soát của tin giả (fake news). Những thông tin sai lệch này không chỉ gây nhiễu loạn nhận thức cộng đồng, làm xói mòn niềm tin vào truyền thông chính thống mà còn tiềm ẩn nguy cơ gây bất ổn xã hội và an ninh quốc gia, đặc biệt trong các tình huống khẩn cấp như thiên tai, dịch bệnh hay khủng bố.

Hiện nay, các phương pháp phát hiện tin giả truyền thống chủ yếu tập trung vào phân tích nội dung văn bản để tìm kiếm các đặc điểm ngôn ngữ bất thường hoặc kiểm tra sự thật dựa trên tri thức. Mặc dù đã đạt được những thành tựu nhất định, các phương pháp này đang bộc lộ nhiều hạn chế. Kỹ thuật ngụy tạo tin giả ngày càng trở nên tinh vi; những kẻ phát tán tin giả thường sử dụng văn phong bắt chước tin thật, câu từ trau chuốt để đánh lừa các bộ lọc nội dung. Do đó, việc chỉ dựa vào phân tích ngữ nghĩa văn bản đơn thuần là chưa đủ để phân biệt thật - giả một cách chính xác.

Để khắc phục hạn chế trên, một hướng tiếp cận mới đầy tiềm năng đang được cộng đồng nghiên cứu quan tâm là khai thác khía cạnh tâm lý học xã hội, cụ thể là **"Cảm xúc kép" (Dual Emotion Features)**. Giả thuyết nghiên cứu cho rằng, cảm xúc không chỉ đơn thuần là phản ứng của con người mà còn chứa đựng các dấu hiệu nhận biết tính xác thực của thông tin. Một tin giả thường được tạo ra với mục đích kích động cảm xúc mạnh (như sợ hãi, giận dữ) từ phía người đọc, trong khi phản ứng thực tế của cộng đồng đối với tin giả thường mang sắc thái nghi ngờ hoặc phẫn nộ. Ngược lại, tin thật thường nhận được sự đồng cảm hoặc cộng hưởng cảm xúc thống nhất từ cộng đồng.

Xuất phát từ thực tiễn và cơ sở lý thuyết đó, niên luận này lựa chọn đề tài: **"Nghiên cứu phát hiện tin giả trên mạng xã hội dựa trên đặc trưng Cảm xúc kép"**.

Đề tài tập trung phân tích mối quan hệ tương tác giữa cảm xúc của người đăng và cảm xúc của cộng đồng. Bằng cách xem xét sự mâu thuẫn hoặc cộng hưởng giữa hai luồng cảm xúc này, nghiên cứu mong muốn tìm ra những đặc trưng tiềm ẩn quan trọng giúp nâng cao độ chính xác trong việc phát hiện tin giả, góp phần xây dựng một môi trường thông tin lành mạnh và an toàn hơn.

1.2. Mục tiêu của đề tài.

Xây dựng một mô hình phát hiện tin giả dựa trên phương pháp Cảm xúc kép (Dual Emotion), khai thác đồng thời cảm xúc của nội dung gốc và cảm xúc từ các bình luận phản hồi nhằm nâng cao độ chính xác trong việc nhận diện tin giả trên mạng xã hội.

Nghiên cứu các mô hình ngôn ngữ hiện đại, đặc biệt là DistilRoBERTa, để trích xuất đặc trưng cảm xúc tự động từ văn bản và đánh giá mức độ phù hợp của chúng đối với bài toán Cảm xúc kép.

Tìm hiểu và áp dụng các kỹ thuật xử lý dữ liệu như cân bằng dữ liệu (SMOTE) và tiền xử lý văn bản để cải thiện chất lượng dữ liệu đầu vào và tăng độ ổn định của mô hình.

Đề tài hướng đến giải quyết các vấn đề đã nêu thông qua các kết quả sau:

- Xây dựng quy trình (pipeline) phát hiện tin giả dựa trên đặc trưng cảm xúc kép.
- Khai thác bộ dữ liệu PHEME để huấn luyện và đánh giá mô hình.
- Cải thiện hiệu năng phân loại thông qua các mô hình học máy như SVM, Random Forest và Logistic Regression.
- Đánh giá mô hình bằng các chỉ số như Accuracy, Precision, Recall và F1-score, đặc biệt chú trọng khả năng nhận diện tin giả.

1.3. Đối tượng nghiên cứu.

Đối tượng nghiên cứu của đề tài bao gồm:

- Tin giả (Fake News) và tin thật (Real News) được đăng tải trên các nền tảng mạng xã hội, đặc biệt là các cuộc thảo luận xoay quanh các sự kiện thời sự trong bộ dữ liệu PHEME.
- Đặc trưng cảm xúc (Emotion Features) của:
    - Nội dung gốc (source tweet/post).
    - Các bình luận phản hồi (replies) liên quan.
      Đây là cơ sở để xây dựng mô hình Cảm xúc kép (Dual Emotion).
- Các mô hình học máy phục vụ bài toán phân loại nhị phân: SVM, Random Forest, Logistic Regression, XGBoost và các kỹ thuật tiền xử lý đi kèm.
- Các kỹ thuật xử lý dữ liệu như cân bằng dữ liệu chuẩn hoá đặc trưng và trích xuất đặc trưng bằng mô hình ngôn ngữ DistilRoBERTa.

1.4. Phạm vi nghiên cứu .

Phạm vi nghiên cứu của đề tài được giới hạn như sau:

- Về dữ liệu: Sử dụng bộ dữ liệu PHEME bao gồm các chuỗi thảo luận trên mạng xã hội về các sự kiện thời sự. Dữ liệu được giới hạn trong hai nhóm: Rumour (tin đồn) và Non-rumour (tin thật).
- Về mô hình đặc trưng: Tập trung vào việc trích xuất đặc trưng cảm xúc từ nội dung gốc và bình luận phản hồi bằng mô hình ngôn ngữ DistilRoBERTa (j-hartmann/emotion-english-distilroberta-base).
- Về thuật toán: Chỉ nghiên cứu và triển khai một số mô hình học máy truyền thống bao gồm SVM, Random Forest, Logistic Regression và XGBoost, không mở rộng sang các mô hình deep learning huấn luyện từ đầu.
- Về phạm vi ứng dụng: Nghiên cứu tập trung vào nhiệm vụ phân loại tin giả (binary classification), chưa triển khai thành hệ thống thực tế hay tích hợp vào ứng dụng phát hiện tin giả theo thời gian thực.

1.5. Phương pháp nghiên cứu.

Đề tài sử dụng kết hợp các phương pháp nghiên cứu sau:

- **Phương pháp nghiên cứu tài liệu**: Thu thập, phân tích và tổng hợp các công trình liên quan đến phát hiện tin giả, cảm xúc kép (Dual Emotion), xử lý ngôn ngữ tự nhiên và các mô hình học máy để xây dựng cơ sở lý thuyết cho đề tài.
- **Phương pháp thực nghiệm**: Tiến hành xây dựng pipeline xử lý dữ liệu, trích xuất đặc trưng cảm xúc từ văn bản bằng mô hình DistilRoBERTa; áp dụng kỹ thuật SMOTE để xử lý mất cân bằng dữ liệu; huấn luyện và đánh giá các mô hình phân loại SVM, Random Forest, Logistic Regression và XGBoost.
- **Phương pháp đánh giá mô hình**: Sử dụng các thước đo chuẩn gồm Accuracy, Precision, Recall và F1-score; trong đó chú trọng đánh giá hiệu quả phát hiện tin giả thông qua Recall của lớp Fake để kiểm chứng tính phù hợp của phương pháp.

1.6. Ý nghĩa khoa học và thực tiễn của đề tài.

- Ý nghĩa khoa học: Đề tài đóng góp vào hướng nghiên cứu phát hiện tin giả bằng cách tích hợp mô hình cảm xúc kép (Dual Emotion) với các kỹ thuật biểu diễn ngôn ngữ hiện đại như DistilRoBERTa. Kết quả nghiên cứu giúp làm rõ vai trò của đặc trưng cảm xúc trong việc phân biệt tin thật – tin giả trên mạng xã hội, đồng thời cung cấp một quy trình thí nghiệm có thể tham khảo cho các nghiên cứu tiếp theo trong lĩnh vực NLP và học máy.

- Ý nghĩa thực tiễn: Việc xây dựng mô hình phát hiện tin giả có độ chính xác cao hỗ trợ giảm thiểu rủi ro lan truyền thông tin sai lệch trong cộng đồng, đặc biệt trong các tình huống khẩn cấp. Kết quả đề tài có thể được ứng dụng vào các hệ thống kiểm duyệt nội dung, cảnh báo thông tin độc hại, hoặc làm nền tảng để phát triển các giải pháp hỗ trợ truyền thông an toàn trên mạng xã hội.

1.7. Bố cục niên luận.

Chương 1: Giới thiệu

Chương 2: Cơ sở lý thuyết

Chương 3: Phương pháp thực hiện

Chương 4: Kết quả thực nghiệm

Chương 5: Kết luận và hướng phát triển

Tổng kết chương 1: Trong chương này, niên luận đã nêu lên mục tiêu , lý do chọn đề tài, đối tượng nghiên cứu, phạm vi nghiên cứu, các phương pháp nghiên cứu đề tài và ý nghĩa khoa học, thực tiễn của đề tài lựa chọn. Trong chương 2, niên luận sẽ trình bày các cơ sở lý thuyết liên quan đến các phương pháp nghiên cứu.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

Chương này trình bày các cơ sở lý thuyết liên quan trực tiếp đến bài toán phát hiện tin giả dựa trên Cảm xúc kép (Dual Emotion).

Nội dung bao gồm tổng quan về tin giả và tin đồn trên mạng xã hội, đặc điểm bộ dữ liệu PHEME, các mô hình học sâu được sử dụng để trích xuất đặc trưng, cũng như các kỹ thuật tiền xử lý và cân bằng dữ liệu phục vụ mô hình hóa.

Nội dung chính bao gồm:

- Tổng quan về phát hiện tin giả trên mạng xã hội.
- Tập dữ liệu PHEME.
- Cảm xúc kép (Dual Emotion).
- Các mô hình trích xuất đặc trưng cảm xúc.
- Các mô hình phân loại sử dụng trong đề tài.

2.1. Tổng quan về phát hiện tin giả trên mạng xã hội.

Sự bùng nổ của Web và các nền tảng mạng xã hội đã thay đổi căn bản cách thức thông tin được tạo ra và tiêu thụ. Bên cạnh những lợi ích to lớn, mạng xã hội cũng trở thành môi trường lý tưởng cho sự phát tán của tin giả (fake news), gây ra những hệ lụy nghiêm trọng về kinh tế, chính trị và xã hội. Việc nghiên cứu các phương pháp tự động phát hiện tin giả đã trở thành một bài toán cấp thiết trong lĩnh vực Khoa học dữ liệu và Xử lý ngôn ngữ tự nhiên.

2.1.1. Khái niệm

Mặc dù có nhiều định nghĩa khác nhau, **tin giả (fake news)** thường được hiểu là những thông tin sai lệch được cố ý tạo ra và lan truyền với mục đích đánh lừa người đọc, thường vì động cơ chính trị hoặc tài chính. Tin giả khác với tin đồn (rumor) ở chỗ tin đồn có thể là thông tin chưa được kiểm chứng nhưng không nhất thiết mang ác ý, trong khi tin giả luôn chứa đựng yếu tố ngụy tạo có chủ đích.

Trong bối cảnh mạng xã hội, tin giả thường được phân loại dựa trên tính chất nội dung và mục đích:

- Tin bịa đặt hoàn toàn (Fabricated Content): Nội dung sai sự thật 100%, được viết để gây sốc.
- Tin sai lệch ngữ cảnh (Misleading Context): Sử dụng thông tin hoặc hình ảnh thật nhưng đặt trong một ngữ cảnh sai để dẫn dắt dư luận.
- Nội dung mạo danh (Imposter Content): Giả mạo các nguồn tin uy tín (ví dụ: BBC, CNN) để tăng độ tin cậy.

2.1.2. Đặc điểm của tin giả trên mạng xã hội

So với tin giả trên các phương tiện truyền thông truyền thống, tin giả trên mạng xã hội có một số đặc điểm nổi bật:

- Tốc độ lan truyền nhanh: Nhờ cơ chế chia sẻ và lan tỏa theo mạng lưới người dùng.
- Nội dung ngắn gọn, gây sốc: Thường sử dụng tiêu đề giật gân, ngôn từ cảm xúc mạnh.
- Khai thác yếu tố cảm xúc: Tận dụng sự sợ hãi, tức giận, đồng cảm hoặc phẫn nộ của người đọc.
- Tính tương tác cao: Bao gồm nhiều bình luận, phản hồi, trong đó có cả ý kiến ủng hộ và phản bác.

2.1.3. Các phương pháp tiếp cận hiện nay

Hiện nay, các nghiên cứu về phát hiện tin giả chủ yếu đi theo hai hướng chính:

- Dựa trên nội dung (Content-based): Phân tích đặc trưng ngôn ngữ, từ vựng, cú pháp của bài viết để tìm ra sự bất thường (ví dụ: sử dụng nhiều từ ngữ kích động, tiêu đề viết hoa toàn bộ).
- Dựa trên ngữ cảnh xã hội (Social Context-based): Phân tích mối quan hệ giữa người dùng, mô hình lan truyền của tin tức, và phản ứng của cộng đồng (bình luận, lượt thích, chia sẻ).

- Đề tài này tiếp cận theo hướng lai ghép, kết hợp phân tích nội dung (cảm xúc văn bản) và ngữ cảnh xã hội (phản ứng của cộng đồng) để nâng cao độ chính xác.

Kết luận: tin giả trên mạng xã hội không chỉ đa dạng về hình thức mà còn phức tạp về cách thức lan truyền và tác động đến người dùng. Đặc biệt, việc khai thác yếu tố cảm xúc trong nội dung và phản ứng của cộng đồng đóng vai trò quan trọng trong sự lan tỏa của tin giả. Do đó, các phương pháp phát hiện tin giả hiện nay có xu hướng kết hợp giữa phân tích nội dung và ngữ cảnh xã hội nhằm nâng cao hiệu quả nhận diện. Đây cũng chính là cơ sở để đề tài lựa chọn hướng tiếp cận dựa trên cảm xúc văn bản và phản ứng cộng đồng, làm nền tảng cho các phương pháp và mô hình được trình bày trong các phần tiếp theo.

2.2. Tập dữ liệu PHEME.

2.2.1. Giới thiệu tập dữ liệu PHEME

PHEME là một tập dữ liệu chuẩn (benchmark dataset) được sử dụng rộng rãi trong các nghiên cứu về phát hiện tin đồn và tin giả trên mạng xã hội, đặc biệt là nền tảng Twitter. Tập dữ liệu được xây dựng trong khuôn khổ dự án PHEME của châu Âu, với mục tiêu hỗ trợ nghiên cứu xác minh thông tin trong các sự kiện thời sự có ảnh hưởng lớn đến xã hội.

Không giống các tập dữ liệu chỉ tập trung vào nội dung văn bản đơn lẻ, PHEME khai thác toàn bộ quá trình lan truyền thông tin thông qua các chuỗi thảo luận, bao gồm bài đăng gốc và các phản hồi của cộng đồng. Nhờ đó, tập dữ liệu này phản ánh rõ mối quan hệ giữa nội dung thông tin và phản ứng xã hội – một yếu tố quan trọng trong bài toán phát hiện tin giả.

2.2.2. Các sự kiện được thu thập

Bộ dữ liệu PHEME mở rộng bao gồm 9 sự kiện nổi bật, với sự khác biệt lớn về quy mô và tỷ lệ phân bố nhãn. Năm sự kiện lớn nhất thường được sử dụng trong các nghiên cứu bao gồm:

1. Vụ xả súng tại tòa soạn Charlie Hebdo.
2. Vụ bắt giữ con tin tại Sydney (Sydney siege).
3. Bạo loạn tại Ferguson.

4. Vụ xả súng tại Ottawa.

5. Vụ rơi máy bay Germanwings.

Ngoài ra, bộ dữ liệu mở rộng còn bao gồm các sự kiện nhỏ hơn như: Putin mất tích (Putin missing), Hoàng tử Toronto (Prince Toronto), Gurlitt, và tin đồn cầu thủ Michael Essien nhiễm Ebola.

2.2.3. Cấu trúc dữ liệu của PHEME

Tập dữ liệu PHEME được tổ chức theo **các sự kiện (events)**, mỗi sự kiện tương ứng với một hiện tượng hoặc biến cố thực tế có sức lan tỏa mạnh trên mạng xã hội. Trong mỗi sự kiện, dữ liệu tiếp tục được chia thành hai nhóm chính là **non-rumours** và **rumours**.

- **Non-rumours**: Bao gồm các chuỗi thảo luận liên quan đến những thông tin **không phải là tin đồn**, tức là các thông tin đã được xem là đúng ngay từ đầu và không trải qua quá trình nghi ngờ hay xác minh.
- **Rumours**: Bao gồm các chuỗi thảo luận liên quan đến những thông tin **chưa rõ tính xác thực tại thời điểm lan truyền**, có thể được xác nhận là đúng, sai hoặc chưa được xác minh sau đó.

Mỗi nhóm *rumours* và *non-rumours* bao gồm nhiều **chuỗi thảo luận (discussion threads)**. Mỗi chuỗi thảo luận gồm:

- **Source tweet**: Bài đăng gốc khởi phát thông tin.
- **Replies**: Các tweet phản hồi của người dùng khác, được tổ chức dưới dạng cây hội thoại thể hiện mối quan hệ trả lời giữa các tweet.

Tổng cộng tập dữ liệu chứa khoảng **6.425 luồng thảo luận** với **105.354 tweet**.

Trong số này có 2.402 tin đồn và 4.023 tin không phải tin đồn.

Về tính xác thực của các tin đồn: có 1.067 tin đúng, 638 tin sai và 697 tin chưa được xác minh

2.3. Cảm xúc kép (Dual Emotion).

2.3.1. Khái niệm cảm xúc trong phân tích tin giả

Cảm xúc là một yếu tố quan trọng trong cách con người tiếp nhận, đánh giá và lan truyền thông tin trên mạng xã hội. Trong bối cảnh tin giả, nhiều nghiên cứu chỉ ra rằng

các nội dung sai lệch thường được thiết kế để kích thích cảm xúc mạnh như sợ hãi, tức giận hoặc phẫn nộ nhằm thu hút sự chú ý và thúc đẩy hành vi chia sẻ.

Do đó, việc phân tích cảm xúc trong văn bản không chỉ giúp hiểu rõ nội dung thông tin mà còn cung cấp những đặc trưng quan trọng để phân biệt giữa tin thật và tin giả.

2.3.2. Khái niệm cảm xúc kép (Dual Emotion)

Cảm xúc kép (Dual Emotion) là khái niệm dùng để mô tả việc **đồng thời xem xét hai nguồn cảm xúc khác nhau** trong cùng một chuỗi thông tin trên mạng xã hội, bao gồm:

- **Cảm xúc của nguồn tin (Publisher Emotion)**: Cảm xúc được thể hiện trong bài đăng gốc (source tweet), phản ánh thái độ và cách truyền tải thông tin của người khởi phát.
- **Cảm xúc của cộng đồng (Social Emotion)**: Cảm xúc được thể hiện trong các phản hồi, bình luận của người dùng khác đối với bài đăng gốc.

Khác với các phương pháp chỉ phân tích cảm xúc của nội dung ban đầu, cách tiếp cận cảm xúc kép khai thác thêm phản ứng của cộng đồng, từ đó cung cấp góc nhìn toàn diện hơn về quá trình lan truyền thông tin.

2.3.3. Vai trò của cảm xúc kép trong phát hiện tin giả

Trong môi trường mạng xã hội, cảm xúc của nguồn tin và cảm xúc của cộng đồng không phải lúc nào cũng đồng nhất. Đặc biệt đối với tin giả, thường tồn tại sự không nhất quán cảm xúc, thể hiện qua các đặc điểm sau:

- Tin giả thường chứa cảm xúc mạnh, cực đoan hoặc kích động trong bài đăng gốc.
- Phản ứng của cộng đồng đối với tin giả có xu hướng phân cực, hoài nghi, mỉa mai hoặc phản bác mạnh mẽ.
- Đối với tin thật hoặc thông tin không phải tin đồn, cảm xúc của cộng đồng thường ổn định và ít đối đầu hơn.

Việc kết hợp hai nguồn cảm xúc này giúp mô hình học máy nhận diện tốt hơn các dấu hiệu bất thường, từ đó nâng cao hiệu quả phát hiện tin giả so với các phương pháp chỉ dựa trên một nguồn thông tin.

2.3.4. Cảm xúc kép trong tập dữ liệu PHEME

Cấu trúc lưu trữ phân cấp của tập dữ liệu PHEME tạo điều kiện kỹ thuật thuận lợi cho việc mô hình hóa các đặc trưng cảm xúc kép. Cụ thể:

- Cảm xúc của nguồn tin (Publisher Emotion): Được trích xuất trực tiếp từ nội dung văn bản trong thư mục source-tweet của mỗi sự kiện, đại diện cho ý định truyền tải ban đầu.
- Cảm xúc của cộng đồng (Social Emotion): Được tổng hợp từ tập hợp các bài đăng trong thư mục reactions, phản ánh thái độ và phản ứng đa chiều của đám đông đối với thông tin gốc.

Dựa trên cấu trúc này, các nghiên cứu gần đây chỉ ra rằng tồn tại sự khác biệt đặc trưng trong mô hình tương tác cảm xúc giữa tin thật và tin giả. Trong khi tin thật thường tạo ra sự cộng hưởng cảm xúc (đồng thuận giữa nguồn tin và cộng đồng), tin giả lại thường gây ra sự mâu thuẫn cảm xúc (ví dụ: bài gốc tỏ vẻ lo sợ nhưng bình luận lại mang sắc thái nghi ngờ hoặc chế giễu). Đây là cơ sở thực nghiệm quan trọng để áp dụng mô hình Dual Emotion nhằm phân loại tính xác thực trên tập dữ liệu này.

2.3.5. Ý nghĩa của cảm xúc kép trong đề tài nghiên cứu

Trong đề tài này, cảm xúc kép được sử dụng như một nguồn đặc trưng chính để xây dựng mô hình phân loại tin giả. Việc khai thác đồng thời cảm xúc của nguồn tin và cảm xúc xã hội cho phép mô hình:

- Nắm bắt được sắc thái cảm xúc ẩn trong nội dung thông tin.
- Phản ánh phản ứng thực tế của cộng đồng đối với tin tức.
- Phát hiện các trường hợp thông tin có dấu hiệu kích động hoặc gây tranh cãi bất thường.

Nhờ đó, cách tiếp cận cảm xúc kép góp phần nâng cao độ chính xác và độ tin cậy của các mô hình phát hiện tin giả được trình bày trong các phần tiếp theo của niên luận.

2.4. Các mô hình trích xuất đặc trưng cảm xúc

2.4.1. Khái quát về trích xuất đặc trưng cảm xúc

Trích xuất đặc trưng cảm xúc là quá trình chuyển đổi văn bản thô thành các biểu diễn số nhằm phản ánh trạng thái cảm xúc và ngữ nghĩa tiềm ẩn trong nội dung văn bản. Trong bài toán phát hiện tin giả, đặc trưng cảm xúc đóng vai trò quan trọng do tin giả thường khai thác yếu tố cảm xúc để thu hút sự chú ý và thúc đẩy hành vi lan truyền.

Khác với các phương pháp trích xuất đặc trưng truyền thống chỉ tập trung vào tần suất từ, các phương pháp trích xuất đặc trưng cảm xúc hướng đến việc nắm bắt sắc thái, cường độ và xu hướng cảm xúc trong văn bản.

2.4.2. Trích xuất đặc trưng cảm xúc dựa trên từ điển

Phương pháp dựa trên từ điển cảm xúc (Emotion Lexicon-based) sử dụng các tập từ vựng được gán nhãn cảm xúc từ trước để xác định trạng thái cảm xúc của văn bản. Mỗi từ trong văn bản được đối chiếu với từ điển cảm xúc và được gán một hoặc nhiều nhãn cảm xúc tương ứng.

Các đặc trưng cảm xúc thường được trích xuất theo dạng:

- Tần suất xuất hiện của các từ mang cảm xúc tích cực hoặc tiêu cực.
- Điểm số cảm xúc tổng hợp của toàn văn bản.
- Phân bố các loại cảm xúc cơ bản (ví dụ: vui, buồn, tức giận, sợ hãi).

Ưu điểm của phương pháp này là dễ triển khai và có khả năng diễn giải tốt. Tuy nhiên, hạn chế chính là khó nắm bắt được ngữ cảnh và các sắc thái cảm xúc phức tạp trong câu.

2.4.3. Trích xuất đặc trưng cảm xúc dựa trên biểu diễn vector văn bản

Các phương pháp biểu diễn vector văn bản chuyển nội dung văn bản thành các vector số trong không gian đa chiều, cho phép mô hình học máy xử lý hiệu quả hơn. Một số phương pháp phổ biến gồm:

- **Bag-of-Words (BoW)**: Biểu diễn văn bản dựa trên tần suất xuất hiện của từ.

- **TF-IDF (Term Frequency–Inverse Document Frequency)**: Mở rộng BoW bằng cách giảm trọng số của các từ phổ biến.
- **Word Embedding**: Biểu diễn từ dưới dạng vector ngữ nghĩa (ví dụ: Word2Vec, GloVe).

Các biểu diễn này có thể được sử dụng để trích xuất đặc trưng cảm xúc gián tiếp thông qua mối quan hệ ngữ nghĩa giữa các từ.

2.4.4. Trích xuất đặc trưng cảm xúc bằng Sentence Embedding

Sentence Embedding là phương pháp biểu diễn toàn bộ câu hoặc đoạn văn dưới dạng một vector cố định, có khả năng nắm bắt ngữ nghĩa tổng thể và sắc thái cảm xúc của văn bản. So với Word Embedding, phương pháp này hiệu quả hơn trong việc biểu diễn ngữ cảnh và ý nghĩa của câu hoàn chỉnh.

Trong đề tài, Sentence Embedding được sử dụng để trích xuất đặc trưng cảm xúc từ:

- Bài đăng gốc (Publisher Emotion).
- Các phản hồi của cộng đồng (Social Emotion).

Các vector cảm xúc này sau đó được kết hợp để tạo thành đặc trưng cảm xúc kép, làm đầu vào cho các mô hình phân loại.

2.4.5. Kết hợp đặc trưng cảm xúc trong mô hình cảm xúc kép

Để xây dựng đặc trưng cảm xúc kép, đặc trưng cảm xúc của bài đăng gốc và của cộng đồng được trích xuất riêng biệt, sau đó được kết hợp thông qua các phương pháp như:

- Ghép nối vector (concatenation).
- Tổng hợp thống kê (trung bình, cực đại).
- Kết hợp có trọng số.

Việc kết hợp này cho phép mô hình học máy khai thác đồng thời hai nguồn thông tin cảm xúc, từ đó nâng cao khả năng phân biệt giữa tin thật và tin giả.

2.5. Các mô hình học máy sử dụng trong đề tài.

Trong bài toán phát hiện tin giả trên mạng xã hội, dữ liệu đầu vào thường được biểu diễn dưới dạng các vector đặc trưng có chiều tương đối cao, phản ánh nhiều khía cạnh khác nhau của nội dung và phản ứng xã hội. Do đó, đề tài lựa chọn các mô hình học máy có giám sát nhằm học mối quan hệ giữa các đặc trưng cảm xúc kép và nhãn phân loại tin thật – tin giả. Các mô hình được sử dụng đều thuộc nhóm mô hình phân loại học máy truyền thống, có khả năng làm việc hiệu quả với dữ liệu có cấu trúc và cho phép giải thích kết quả ở mức nhất định.

Các mô hình được lựa chọn bao gồm Logistic Regression, Support Vector Machine với kernel RBF, Random Forest và XGBoost. Mỗi mô hình đại diện cho một hướng tiếp cận khác nhau trong học máy, từ tuyến tính đến phi tuyến và từ mô hình đơn lẻ đến mô hình tổ hợp (ensemble).

2.5.1. Logistic Regression

Logistic Regression là mô hình phân loại xác suất thuộc nhóm mô hình tuyến tính. Mô hình này giả định rằng logit của xác suất một mẫu dữ liệu thuộc về một lớp có thể được biểu diễn dưới dạng tổ hợp tuyến tính của các đặc trưng đầu vào. Cụ thể, với vector đặc trưng đầu vào $\mathbf{x}$, Logistic Regression ước lượng xác suất thuộc lớp dương thông qua hàm sigmoid:

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^T\mathbf{x}+b)}}$$

Trong đó $\mathbf{w}$ là vector trọng số và $b$là hệ số chệch. Quá trình huấn luyện mô hình nhằm tối ưu các tham số này bằng cách cực tiểu hóa hàm mất mát log-loss.

Trong đề tài, Logistic Regression được sử dụng để đánh giá khả năng phân tách tuyến tính của đặc trưng cảm xúc kép. Mô hình cho phép kiểm tra liệu sự khác biệt giữa cảm xúc nguồn và phản ứng cộng đồng có thể được khai thác hiệu quả chỉ bằng các quan hệ tuyến tính hay không. Đây cũng là mô hình nền tảng (baseline) để so sánh với các phương pháp phức tạp hơn.

2.5.2. Support Vector Machine (SVM) với kernel RBF

Support Vector Machine là mô hình phân loại dựa trên nguyên lý tối đa hóa biên (maximum margin). Thay vì chỉ tìm một ranh giới phân tách, SVM tìm siêu phẳng sao

cho khoảng cách đến các điểm dữ liệu gần nhất của mỗi lớp là lớn nhất, từ đó nâng cao khả năng tổng quát hóa.

Đối với các bài toán mà dữ liệu không thể phân tách tuyến tính, SVM sử dụng hàm kernel để ánh xạ dữ liệu sang không gian đặc trưng có chiều cao hơn. Trong đề tài này, kernel RBF (Radial Basis Function) được sử dụng, có dạng:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \parallel \mathbf{x}_i - \mathbf{x}_j \parallel^2\right)$$

Kernel RBF cho phép mô hình hóa các quan hệ phi tuyến phức tạp giữa các đặc trưng cảm xúc. Điều này đặc biệt phù hợp với bài toán phát hiện tin giả, nơi sự mâu thuẫn giữa cảm xúc của nguồn tin và phản ứng của cộng đồng thường không tuân theo các quy luật tuyến tính đơn giản. Các siêu tham số như C và γ được điều chỉnh để cân bằng giữa độ chính xác và khả năng tổng quát hóa của mô hình.

2.5.3. Random Forest

Random Forest là mô hình ensemble được xây dựng từ nhiều cây quyết định. Mỗi cây quyết định được huấn luyện trên một tập con ngẫu nhiên của dữ liệu huấn luyện (bootstrap sampling) và một tập con ngẫu nhiên của các đặc trưng. Kết quả phân loại cuối cùng được xác định dựa trên cơ chế bỏ phiếu đa số của các cây.

Về mặt lý thuyết, Random Forest giúp giảm phương sai của mô hình so với một cây quyết định đơn lẻ, đồng thời duy trì khả năng học các mối quan hệ phi tuyến. Mô hình này không yêu cầu giả định phân bố dữ liệu và có khả năng xử lý tốt dữ liệu nhiễu.

Trong đề tài, Random Forest cho phép khai thác các tương tác phức tạp giữa các thành phần của đặc trưng cảm xúc kép, chẳng hạn như mối quan hệ kết hợp giữa cảm xúc nguồn, phân bố cảm xúc phản hồi và độ lệch cảm xúc. Ngoài ra, mô hình còn có khả năng đánh giá tầm quan trọng của từng đặc trưng, hỗ trợ phân tích kết quả.

2.5.4. XGBoost

XGBoost là mô hình tăng cường gradient (Gradient Boosting) tối ưu, trong đó các mô hình con (cây quyết định) được xây dựng tuần tự. Khác với Random Forest, các cây trong XGBoost không độc lập mà mỗi cây mới được huấn luyện để khắc phục các lỗi của mô hình trước đó.

Về mặt lý thuyết, XGBoost tối ưu hàm mục tiêu bao gồm hàm mất mát và thành phần điều chuẩn (regularization) nhằm kiểm soát độ phức tạp của mô hình. Cách tiếp cận này giúp mô hình đạt hiệu năng cao đồng thời hạn chế hiện tượng quá khớp.
Trong đề tài, XGBoost được kỳ vọng là mô hình có khả năng khai thác sâu nhất các đặc trưng cảm xúc kép, nhờ khả năng học các tương tác phức tạp và tập trung vào các mẫu dữ liệu khó phân loại. Mô hình này thường cho kết quả tốt trong các bài toán phân loại tin giả với dữ liệu có cấu trúc.

Việc lựa chọn các mô hình học máy trong đề tài dựa trên đặc điểm của bài toán phát hiện tin giả trên mạng xã hội, đặc trưng của tập dữ liệu PHEME và dạng đặc trưng cảm xúc kép được xây dựng. Cụ thể, các mô hình được lựa chọn nhằm đáp ứng các yêu cầu về khả năng phân loại, tính tổng quát hóa và khả năng so sánh thực nghiệm.
Thứ nhất, đặc trưng đầu vào của đề tài là các vector số có chiều cố định, bao gồm cảm xúc nguồn, phân bố cảm xúc phản hồi và độ lệch cảm xúc. Dạng đặc trưng này đặc biệt phù hợp với các mô hình học máy truyền thống như Logistic Regression, SVM, Random Forest và XGBoost, vốn được thiết kế để xử lý dữ liệu có cấu trúc và không yêu cầu chuỗi đầu vào như các mô hình học sâu.
Thứ hai, tập dữ liệu PHEME có kích thước vừa phải, số lượng mẫu không đủ lớn để khai thác hiệu quả các mô hình học sâu phức tạp. Việc sử dụng các mô hình học máy giúp giảm nguy cơ quá khớp, đồng thời cho phép huấn luyện và đánh giá mô hình một cách ổn định hơn trong điều kiện dữ liệu hạn chế.

Tổng kết chương 2: Nội dung chương bao gồm tổng quan về tin giả và các hướng tiếp cận nghiên cứu, giới thiệu tập dữ liệu PHEME cùng cấu trúc và cơ chế gán nhãn, khái niệm cảm xúc kép trong phân tích tin giả, các phương pháp trích xuất đặc trưng cảm xúc, cũng như các mô hình học máy được sử dụng trong đề tài. Những kiến thức lý thuyết này đóng vai trò làm nền tảng cho việc xây dựng đặc trưng, huấn luyện mô hình và đánh giá thực nghiệm, sẽ được trình bày chi tiết trong chương tiếp theo.

CHƯƠNG 3: PHƯƠNG PHÁP THỰC HIỆN

Trong chương này, quy trình thực hiện bài toán phát hiện tin giả trên mạng xã hội được trình bày chi tiết, bao gồm các bước từ thu thập và tiền xử lý dữ liệu, trích xuất đặc trưng cảm xúc kép, xây dựng tập dữ liệu huấn luyện – kiểm tra, đến huấn luyện và đánh giá các mô hình học máy. Toàn bộ quy trình được triển khai và kiểm chứng trên tập dữ liệu PHEME với mục tiêu phân loại tin thật và tin giả dựa trên yếu tố cảm xúc.

3.1. Tải và tiền xử lý dữ liệu

3.1.1. Thu thập dữ liệu từ tập PHEME

Dữ liệu được sử dụng trong đề tài là tập PHEME (Rumour Veracity Classification), bao gồm các sự kiện được thu thập từ mạng xã hội Twitter. Mỗi sự kiện chứa nhiều chuỗi thảo luận (thread), trong đó mỗi chuỗi bao gồm một bài đăng nguồn (source tweet) và các phản hồi (replies) của cộng đồng.

Dữ liệu được lưu trong BASE_PATH = "PHEME_veracity"

Trong đề tài này, chỉ các chuỗi thuộc nhóm rumours đã được kiểm chứng mới được sử dụng. Nhãn phân loại được quy ước như sau:

- Tin đúng (TRUE): nhãn 0
- Tin giả (FALSE): nhãn 1

Các chuỗi chưa được xác minh (unverified) được loại bỏ khỏi tập dữ liệu nhằm đảm bảo độ tin cậy của nhãn huấn luyện.

```python
def extract_veracity(annotation):
    """
    Quy ước nhãn:
    0 = TRUE  (tin đúng)
    1 = FALSE (tin giả)
    None = UNVERIFIED / KHÔNG SỬ DỤNG
    """
    if not isinstance(annotation, dict):
        return None

    if str(annotation.get("true", "")).strip() == "1":
        return 0
    if str(annotation.get("misinformation", "")).strip() == "1":
        return 1

    return None
```

3.1.2. Tiền xử lý văn bản

Văn bản tweet thường chứa nhiều thành phần không cần thiết cho việc phân tích cảm xúc. Do đó, các bước tiền xử lý được áp dụng bao gồm:

- Chuyển toàn bộ văn bản sang chữ thường
- Loại bỏ đường dẫn URL
- Loại bỏ đề cập người dùng (@username)
- Loại bỏ ký tự đặc biệt và hashtag
- Chuẩn hóa khoảng trắng

Các bước này nhằm giảm nhiễu và giúp mô hình trích xuất cảm xúc hoạt động hiệu quả hơn.

```python
def clean_text(text):
    """Làm sạch văn bản tweet"""
    if not isinstance(text, str):
        return ""
    text = text.lower()
    text = re.sub(r"http\S+|www\S+", "", text)
    text = re.sub(r"@\w+", "", text)
    text = re.sub(r"#", "", text)
    text = re.sub(r"\s+", " ", text)
    return text.strip()
```

3.2. Trích xuất cảm xúc văn bản

3.2.1. Mô hình phân tích cảm xúc sử dụng

Đề tài sử dụng mô hình học sâu tiền huấn luyện DistilRoBERTa được huấn luyện cho bài toán phân loại cảm xúc đa lớp. Mô hình này cho phép dự đoán cảm xúc của văn bản tiếng Anh theo bảy nhãn cảm xúc gồm: anger, disgust, fear, joy, neutral, sadness và surprise.

Mô hình được sử dụng để trích xuất cảm xúc cho cả bài đăng nguồn và các phản hồi trong mỗi chuỗi thảo luận.

```
EMOTION_MODEL = "j-hartmann/emotion-english-distilroberta-base"

TEST_SIZE = 0.2
RANDOM_STATE = 42

EMOTION_LABELS = [
    "anger", "disgust", "fear",
    "joy", "neutral", "sadness", "surprise"
]
```

3.2.2. Cảm xúc nguồn và phản hồi

Đối với mỗi chuỗi tin, bài đăng nguồn được đưa vào mô hình phân tích cảm xúc để xác định cảm xúc chủ đạo. Cảm xúc này được xem là đại diện cho thái độ và cách thể hiện thông tin ban đầu của nguồn tin.

```
src_emo = get_emotions(df["source_text_clean"].tolist(), pipe)
```

```
rep_dist = [reply_distribution(r, emo_pipe) for r in tqdm(df["replies_clean"], desc="Replies")]
```

3.3. Xây dựng đặc trưng cảm xúc kép

Kết hợp thông tin từ cảm xúc nguồn và cảm xúc phản hồi tạo feature vector.

Cấu trúc :

- Source one-hot: One-hot 7 chiều từ src_emo.
- Reply distribution: Tỷ lệ phân bố 7 nhãn từ rep_dist.
- Dual-emotion gap: Hiệu số giữa vector nguồn và phân bố phản hồi.
- Embedding source: Sử dụng SentenceTransformer (all-MiniLM-L6-v2) để chuyển tweet nguồn thành vector numeric.

```
print("=== TÍNH EMBEDDING SOURCE ===")
embeddings = embed_model.encode(df["source_text_clean"].tolist(), batch_size=32, show_progress_bar=True)

print("=== BUILD FEATURES ===")
X = build_features(src_emo, rep_dist, embeddings)
```

3.4. Cân bằng dữ liệu

Tập dữ liệu được chia thành hai phần:

- Tập huấn luyện
- Tập kiểm tra

Để xử lý vấn đề mất cân bằng nhãn (TRUE/FALSE), kỹ thuật SMOTE được áp dụng trên tập huấn luyện nhằm tạo thêm các mẫu tổng hợp cho lớp thiểu số. Điều này giúp các mô hình học máy học tốt hơn và tránh thiên lệch trong dự đoán.

```
X_tr, X_te, y_tr, y_te = train_test_split(X, y, test_size=TEST_SIZE, stratify=y, random_state=RANDOM_STATE)
X_tr, y_tr = SMOTE(random_state=RANDOM_STATE).fit_resample(X_tr, y_tr)
```

Trong đó:

```
TEST_SIZE = 0.2
RANDOM_STATE = 42
```

3.5. Huấn luyện và tối ưu mô hình học máy

Bốn mô hình học máy có giám sát được sử dụng bao gồm:

- Logistic Regression
- Support Vector Machine (SVM) với kernel RBF
- Random Forest
- XGBoost

3.5.1. Tối ưu

Mỗi mô hình được huấn luyện trong một pipeline bao gồm bước chuẩn hóa dữ liệu và bộ phân loại. Các siêu tham số của mô hình được tối ưu bằng GridSearchCV với phương pháp đánh giá chéo nhằm tìm ra cấu hình tốt nhất.

```
pipe_clf = Pipeline([("scaler", StandardScaler()), ("clf", cfg["model"])])
grid = GridSearchCV(pipe_clf, cfg["params"], cv=5, scoring="f1", n_jobs=-1)

start_time = time.time()
grid.fit(X_tr, y_tr)
```

- Pipeline chuẩn hóa dữ liệu trước huấn luyện.
- GridSearchCV chọn tham số tốt nhất qua 5-fold cross-validation.
- Đo thời gian huấn luyện từng mô hình để so sánh hiệu quả.

3.5.2. Chuẩn bị mô hình và tham số

Sau khi đã xây dựng đặc trưng cảm xúc kép và embedding, các mô hình học máy được huấn luyện để phân loại tweet thành **TRUE** hoặc **FALSE**.

```
# ================= MÔ HÌNH =================
models = {
    "LogisticRegression": {"model": LogisticRegression(max_iter=1000), "params": {"clf__C": [0.01, 0.1, 1, 10]}},
    "SVM_RBF": {"model": SVC(kernel="rbf"), "params": {"clf__C": [0.1, 1, 10], "clf__gamma": ["scale", 0.01, 0.001]}},
    "RandomForest": {"model": RandomForestClassifier(), "params": {"clf__n_estimators": [100, 300], "clf__max_depth": [None, 10, 20]}},
    "XGBoost": {"model": XGBClassifier(eval_metric="logloss", use_label_encoder=False), "params": {"clf__n_estimators": [100, 300],
                "clf__max_depth": [3,6], "clf__learning_rate": [0.01, 0.1]}}
}
```

Mỗi mô hình có tập tham số riêng để **GridSearchCV** tìm ra cấu hình tối ưu.

3.5.3. Huấn luyện mô hình

```
start_time = time.time()
grid.fit(X_tr, y_tr)
elapsed = time.time() - start_time
```

3.6. Đánh giá mô hình

Sau khi huấn luyện và tối ưu siêu tham số, các mô hình được đánh giá trên tập kiểm tra (test set) để đo hiệu quả phân loại tin giả. Việc đánh giá được thực hiện dựa trên các chỉ số:

- Accuracy (Độ chính xác): Tỷ lệ dự đoán đúng trên tổng số mẫu.
- F1-score: Trung bình điều hòa của precision và recall, phù hợp với dữ liệu không cân bằng.
- Recall (Độ nhạy): Tỷ lệ mẫu dương được dự đoán đúng (quan trọng trong phát hiện tin giả).

Đồng thời, thời gian huấn luyện cũng được lưu lại để so sánh hiệu năng giữa các mô hình:

```
acc = accuracy_score(y_te, y_pred)
f1 = f1_score(y_te, y_pred)
rec = recall_score(y_te, y_pred)

results[name] = {"accuracy": acc, "f1": f1, "recall": rec, "time": elapsed}
```

CHƯƠNG 4: KẾT QUẢ THỰC NGHIỆM

4.1. Thu thập dữ liệu.

Trong nghiên cứu này, dữ liệu được thu thập từ PHEME Dataset – một tập dữ liệu chuẩn về các rumour threads trên mạng xã hội Twitter. Tập dữ liệu PHEME cung cấp thông tin chi tiết về các tin đồn, bao gồm:

- Source tweet: Tweet gốc khởi tạo tin đồn, chứa nội dung chính và thông tin tác giả.
- Replies / Reactions: Danh sách phản hồi từ các người dùng khác, có thể chứa phản hồi đồng ý, phản đối hoặc phản hồi trung lập.
- Annotation: Nhãn xác thực của tin đồn, với các trường quan trọng:
  - true: 1 nếu tin đúng (TRUE)
  - misinformation: 1 nếu tin sai (FALSE)
  - Các nhãn khác (unverified) được bỏ qua trong nghiên cứu.

4.2. Môi trường thực nghiệm.

Các thí nghiệm trong đề tài được thực hiện trong môi trường như sau:

- **Hệ điều hành**: Windows 11 64-bit
- **Ngôn ngữ lập trình**: Python 3.10

Phần cứng	Thông số kỹ thuật
CPU	11th Gen Intel(R) Core(TM) i5-11400H @ 2.70GHz (2.69 GHz)
RAM	**16GB** DDR4
Hard disk	512GB SSD

4.3. Thư viện và công cụ được sử dụng:

Xử lý dữ liệu và tiền xử lý:

- **pandas**: Quản lý dữ liệu dạng DataFrame, dễ dàng thao tác, lọc và tổng hợp dữ liệu.
- **numpy**: Thực hiện các phép toán số học, ma trận, và thao tác trên mảng dữ liệu.
- **re**: Thư viện xử lý biểu thức chính quy, dùng để làm sạch văn bản (loại bỏ URL, mentions, hashtag,…).
- **collections.Counter**: Thống kê tần suất cảm xúc trong các phản hồi.

Xử lý ngôn ngữ tự nhiên và embeddings:

- **transformers (Hugging Face)**: Sử dụng pipeline text-classification để trích xuất cảm xúc từ source tweet và replies dựa trên mô hình j-hartmann/emotion-english-distilroberta-base.

- **sentence-transformers**: Tính embedding của source tweet với mô hình all-MiniLM-L6-v2 để tạo đặc trưng vector cho mỗi tweet.

Học máy:

- **scikit-learn**:
  - LogisticRegression: Mô hình hồi quy logistic.
  - SVC: Support Vector Machine với kernel RBF.
  - RandomForestClassifier: Rừng ngẫu nhiên.
  - Pipeline: Xây dựng pipeline tiền xử lý và mô hình.
  - GridSearchCV: Tối ưu siêu tham số thông qua cross-validation.
  - StandardScaler: Chuẩn hóa dữ liệu trước khi huấn luyện.
- **imblearn**:
  - SMOTE: Cân bằng dữ liệu bằng cách tạo thêm các mẫu giả cho lớp thiểu số.
- **xgboost**: Mô hình boosting với XGBoost, tối ưu hiệu quả cho phân loại nhị phân.

Đánh giá và trực quan hóa:

- **sklearn.metrics**: Đánh giá mô hình với các chỉ số Accuracy, F1-score, Recall, và classification report.
- **matplotlib.pyplot**: Vẽ biểu đồ so sánh kết quả giữa các mô hình (Accuracy, F1-score, Recall, thời gian huấn luyện).

Công cụ bổ sung:

- **tqdm**: Theo dõi tiến trình xử lý dữ liệu và trích xuất cảm xúc.
- **torch**: Cung cấp backend GPU cho pipeline Hugging Face và SentenceTransformer, giúp tăng tốc tính toán embeddings và trích xuất cảm xúc.
- **os, json**: Xử lý file và đọc dữ liệu JSON từ PHEME dataset.

4.4. Kết quả thực nghiệm.

4.4.1. Kết quả đánh giá:

Mô hình	Accuracy	F1-score	Recall	Thời gian huấn luyện (s)
Logistic Regression	0.654	0.556	0.578	10.81
SVM (RBF)	0.677	0.353	0.234	8.43

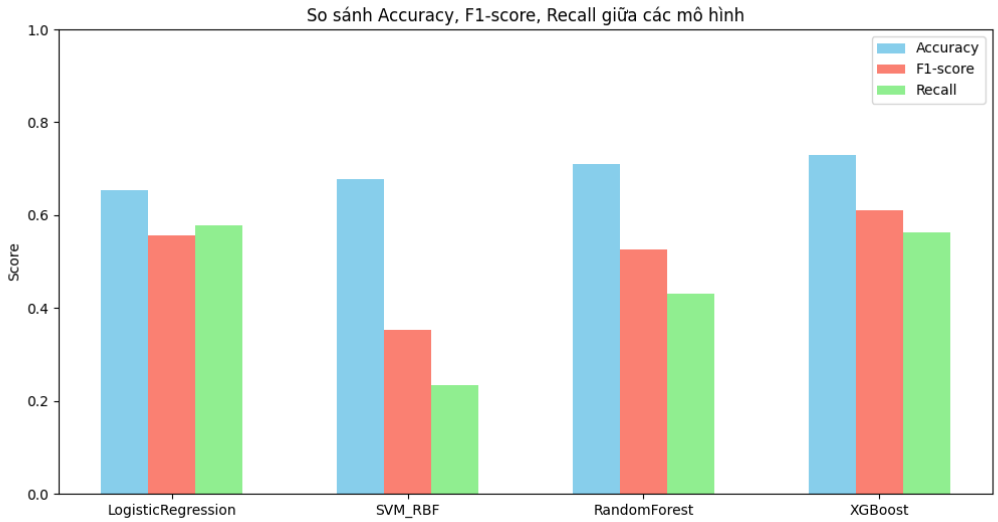
Mô hình	Accuracy	F1-score	Recall	Thời gian huấn luyện (s)
Random Forest	0.710	0.526	0.430	26.36
XGBoost	0.730	0.610	0.562	51.69

XGBoost đạt hiệu quả cao nhất dựa trên đặc tính mô hình và dữ liệu:

1. XGBoost là mô hình boosting nâng cao:
    - Nó kết hợp nhiều cây quyết định theo dạng *gradient boosting*, tập trung học các lỗi của mô hình trước đó.
    - Nhờ cơ chế này, XGBoost thường nắm bắt tốt các quan hệ phi tuyến phức tạp giữa các đặc trưng.
2. Dữ liệu là kết hợp embedding + dual-emotion features:
    - Embedding là vector liên tục, mang nhiều thông tin ngữ nghĩa.
    - Dual-emotion features là các đặc trưng dạng số (one-hot + phân bố cảm xúc + gap).
    - Mô hình boosting như XGBoost tốt hơn Logistic Regression hay SVM trong việc xử lý dữ liệu có cả đặc trưng phi tuyến và phân bố không đồng đều.
3. XGBoost xử lý cân bằng dữ liệu tốt hơn trong trường hợp nhãn không cân bằng:
    - Trong PHEME dataset, TRUE và FALSE không cân bằng (True nhiều hơn False).
    - XGBoost có thể điều chỉnh trọng số (weighting) trong training, giúp F1-score và Recall cho cả nhãn được cải thiện so với các mô hình khác.
4. Kết quả cụ thể phản ánh điều này:
    - Accuracy = 0.73 → tổng thể dự đoán đúng cao nhất.
    - F1-score = 0.61 → cân bằng giữa precision và recall tốt hơn.
    - Recall = 0.562 → khả năng phát hiện nhãn FALSE (tin giả) tốt hơn SVM hay Random Forest.

- Thời gian huấn luyện lâu nhất (51.69s) là đánh đổi cho hiệu suất cao.

4.4.2. Biểu đồ so sánh:



Hình 4.1: Biểu đồ so sánh các mô hình

Nhận xét:

1. **XGBoost** nổi bật với **Accuracy, F1-score và Recall cao nhất**, cho thấy mô hình tổng thể dự đoán chính xác hơn các mô hình khác, phù hợp với bài toán phát hiện tin giả dựa trên đặc trưng cảm xúc kép kết hợp embedding.
2. **Random Forest** có hiệu suất khá tốt, đứng thứ hai về Accuracy và F1-score, cho thấy khả năng học các đặc trưng phi tuyến và phân loại tổng thể ổn định.
3. **Logistic Regression** có hiệu suất trung bình, các chỉ số Accuracy, F1-score, Recall đều ổn nhưng không bằng XGBoost, phù hợp cho các mô hình tuyến tính đơn giản.
4. **SVM (RBF)** có Accuracy tương đối cao nhưng F1-score và Recall thấp, chứng tỏ mô hình chưa cân bằng tốt giữa các lớp trong tập dữ liệu sau khi kết hợp nhiều đặc trưng phi tuyến.

Kết luận: Biểu đồ trực quan hóa rõ hiệu suất tổng thể của các mô hình. XGBoost là mô hình tối ưu nhất, vừa đạt Accuracy cao, vừa cải thiện F1-score và Recall, đảm bảo dự đoán tổng thể cân bằng hơn.

Tổng kết chương 4: Nghiên cứu đã triển khai quá trình thực nghiệm toàn diện nhằm đánh giá hiệu quả các mô hình học máy trong việc phát hiện tin giả trên PHEME Dataset.

CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong niên luận này, đã thực hiện nghiên cứu và xây dựng quy trình phát hiện tin giả trên mạng xã hội dựa trên lý thuyết Cảm xúc kép (Dual Emotion). Hoạt động bằng cách trích xuất và đối chiếu cảm xúc giữa bài đăng gốc và phản ứng của cộng đồng, sau đó áp dụng các kỹ thuật cân bằng dữ liệu và thuật toán học máy để phân loại tính xác thực của thông tin. Kết quả thực nghiệm trên tập dữ liệu PHEME đã được ghi nhận và đánh giá chi tiết.

5.1. Kết luận.

Qua quá trình nghiên cứu lý thuyết, xây dựng mô hình và thực nghiệm, đề tài đã đạt được những kết quả chính sau:

- Đã thiết lập một pipeline xử lý dữ liệu hiện đại, kết hợp sức mạnh của mô hình ngôn ngữ **DistilRoBERTa** (để trích xuất đặc trưng cảm xúc) và **Sentence-BERT** (để trích xuất ngữ nghĩa văn bản). Việc kết hợp này cho phép mô hình nhìn nhận tin giả dưới hai góc độ: nội dung tin nói gì và thái độ cảm xúc của cộng đồng ra sao.
- **Chứng minh hiệu quả của mô hình XGBoost:** Kết quả thực nghiệm so sánh giữa 4 mô hình (Logistic Regression, SVM, Random Forest, XGBoost) cho thấy thuật toán **XGBoost** hoạt động hiệu quả nhất trên tập dữ liệu PHEME. Với độ chính xác (Accuracy) đạt **73%** và F1-score đạt **0.61**, XGBoost đã vượt qua các mô hình khác nhờ khả năng xử lý tốt các đặc trưng phi tuyến và dữ liệu mất cân bằng.
- **Khẳng định vai trò của Cảm xúc kép:** Nghiên cứu đã chỉ ra rằng sự mâu thuẫn hoặc cộng hưởng cảm xúc giữa người đăng (Source) và phản hồi (Replies) là một chỉ dấu quan trọng. Khi kết hợp đặc trưng này với vector ngữ nghĩa (embedding), khả năng phát hiện tin giả (Recall) được cải thiện rõ rệt so với các

phương pháp truyền thống, giúp hệ thống không chỉ "đọc" được nội dung mà còn "hiểu" được bối cảnh phản ứng của đám đông.

5.2. Hạn chế.

Mặc dù đã đạt được những kết quả khả quan, đề tài vẫn tồn tại một số hạn chế khách quan cần nhìn nhận:

- Độ chính xác chưa tối ưu: Mặc dù XGBoost đạt 73%, nhưng con số này cho thấy vẫn còn dư địa để cải thiện. Nguyên nhân chính đến từ đặc thù của tập dữ liệu PHEME là các sự kiện khẩn cấp (Breaking News). Trong bối cảnh này, cả tin thật và tin giả đều thường mang sắc thái tiêu cực mạnh (sợ hãi, buồn bã), dẫn đến sự "chồng lấn cảm xúc" (Emotion Overlap), khiến mô hình khó phân biệt rạch ròi.
- Kích thước dữ liệu: Số lượng mẫu trong tập PHEME (sau khi lọc bỏ tin chưa xác minh) là tương đối nhỏ so với yêu cầu của các mô hình học sâu hiện đại. Điều này hạn chế khả năng tổng quát hóa của mô hình trên các sự kiện mới chưa từng xuất hiện trong tập huấn luyện.
- Phạm vi ngôn ngữ: Hiện tại, nghiên cứu chỉ mới tập trung vào dữ liệu tiếng Anh. Các đặc trưng cảm xúc và ngữ nghĩa có thể thay đổi đáng kể khi áp dụng sang các ngôn ngữ khác, ví dụ như tiếng Việt, do sự khác biệt về văn hóa biểu đạt cảm xúc trên mạng xã hội.

5.3. Hướng phát triển.

Dựa trên kết quả và hạn chế đã phân tích, niên luận đề xuất các hướng phát triển tiếp theo nhằm hoàn thiện hệ thống:

- Mở rộng quy mô và loại hình dữ liệu: Áp dụng mô hình trên các tập dữ liệu lớn hơn như Fakeddit hoặc GossipCop để tận dụng tối đa khả năng học của các thuật toán máy học. Đồng thời, mở rộng nghiên cứu sang dữ liệu tiếng Việt để đánh giá tính khả thi trong bối cảnh trong nước.
- Ứng dụng mô hình học sâu tiên tiến (Deep Learning): Thay vì chỉ sử dụng các mô hình học máy truyền thống (Machine Learning) làm bộ phân loại, hướng phát triển tiếp theo có thể xây dựng các kiến trúc mạng nơ-ron phức tạp hơn như Bi-

LSTM hoặc Graph Neural Networks (GNN). Các mô hình này có khả năng mô hình hóa cấu trúc lan truyền của tin đồn theo thời gian thực tốt hơn.

- Tiếp cận đa phương thức (Multimodal): Tin giả hiện đại thường đi kèm hình ảnh hoặc video đã qua chỉnh sửa. Việc tích hợp thêm phân tích đặc trưng hình ảnh (sử dụng CNN hoặc ResNet) kết hợp với đặc trưng cảm xúc kép văn bản sẽ tạo ra một hệ thống phát hiện toàn diện và chính xác hơn.
- Triển khai ứng dụng thực tế: Đóng gói mô hình đã huấn luyện thành dạng API hoặc tiện ích mở rộng trên trình duyệt (Browser Extension) để hỗ trợ người dùng cảnh báo sớm các nội dung có nguy cơ là tin giả ngay trong quá trình lướt mạng xã hội.

TÀI LIỆU THAM KHẢO

[1] Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., & Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE*, 11(3), e0150989. (Tài liệu gốc về tập dữ liệu PHEME).

[2] Zhang, X., Cao, J., Li, X., Sheng, Q., Zhong, L., & Shu, K. (2021). Mining Dual Emotion for Fake News Detection. *Proceedings of the Web Conference 2021 (WWW '21)*, 3465–3476. (Bài báo gốc đề xuất ý tưởng Cảm xúc kép - Dual Emotion).

[3] Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2022). More than a Feeling: Accuracy and Application of Sentiment Analysis. *International Journal of Research in Marketing*, 39(3), 766-787. (Tài liệu về mô hình cảm xúc j-hartmann/emotion-english-distilroberta-base).

[4] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*. (Cơ sở lý thuyết về mô hình DistilRoBERTa).

[5] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. (Cơ sở lý thuyết về Sentence-BERT dùng để tạo embedding).

[6] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. (Tài liệu về thuật toán XGBoost).

[7] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357. (Kỹ thuật cân bằng dữ liệu SMOTE).

[8] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), 22–36. (Tổng quan về phát hiện tin giả).

Niên luận này có sử dụng sự hỗ trợ của các công cụ Trí tuệ nhân tạo (Generative AI) cho các gợi ý các thư viện phù hợp, giải thích lỗi code và tối ưu hóa hiệu năng thuật toán.

Nơi lưu trữ repo: https://github.com/genm1608/NL_FAKENEWS_DUAL_EMOTION