Ứng dụng Tăng Cường Dữ Liệu Văn Bản trong Phân Loại Tin Giả Sử Dụng Mô Hình Deep Learning

Trương Nhật Quang [1,2*], Đỗ Tuấn Trực [1,2], Nguyễn Đức Tài [1,2]
Giảng viên hướng dẫn: ThS.Huỳnh Văn Tín [1,2]
[1]Khoa Khoa Học & Kỹ Thuật Thông Tin, Trường Đại học Công nghệ thông tin, Thành phố Hồ Chí Minh, Việt Nam
[2]Đại Học Quốc Gia, Thành Phố Hồ Chí Minh, Việt Nam
{22521207*, 22521548, 22521277}gm.uit.edu.vn
tinhv@uit.edu.vn

Abstract

Phân loại tin tức giả là một thách thức lớn trên các nền tảng trực tuyến, nơi thiếu hụt dữ liệu và sự chênh lệch trong tập huấn luyện vẫn là vấn đề nổi cộm. Dù các mô hình như LSTM, BERT, NN và RNN đã đạt được một số kết quả tích cực, nhưng hiệu suất của chúng vẫn bị giới hạn bởi chất lượng và sự đa dạng của dữ liệu. Nghiên cứu này tập trung vào việc cải thiện vấn đề thông qua hai kỹ thuật tăng cường dữ liệu: thay thế từ đồng nghĩa và loại bỏ từ chức năng. Kết quả thử nghiệm ban đầu cho thấy việc áp dụng các kỹ thuật tăng cường giúp làm phong phú và đa dạng hóa bộ dữ liệu, từ đó giảm thiểu sự chênh lệch trong tập huấn luyện. Tuy nhiên, các kỹ thuật này vẫn chưa mang lại sự cải thiện vượt trội về hiệu suất so với khi chỉ sử dụng mô hình gốc. Dù vậy, nghiên cứu đã góp phần làm sáng tỏ tiềm năng của tăng cường dữ liệu, đồng thời gợi mở nhiều hướng đi mới nhằm tối ưu hóa các kỹ thuật này trong tương lai. Chúng tôi kỳ vọng những cải tiến trong tăng cường dữ liệu sẽ hỗ trợ đáng kể trong việc nâng cao hiệu quả phân loại tin tức giả, góp phần tạo nên môi trường thông tin trực tuyến minh bạch và đáng tin cậy hơn.

***Từ khóa* —Thay thế từ đồng nghĩa, Loại bỏ từ chức năng, Tăng cường dữ liệu, Deep Learning, LSTM, BERT, NN, RNN**

1 Giới thiệu

Trong bối cảnh các nền tảng trực tuyến và mạng xã hội phát triển nhanh chóng, tin tức giả đang trở thành một vấn đề nghiêm trọng, với khả năng lan truyền nhanh chóng và gây ảnh hưởng lớn đến dư luận. Các mô hình phân loại tin tức giả đã được phát triển để giúp nhận diện và phân loại tin tức sai lệch, từ đó giảm thiểu tác động tiêu cực của chúng. Tuy nhiên, một trong những thách thức lớn trong việc xây dựng các mô hình phân loại tin tức giả chính là thiếu hụt dữ liệu. Mặc dù các mô hình machine learning hiện tại, đặc biệt là các mô hình deep learning, đã đạt được những kết quả nhất định, nhưng vấn đề dữ liệu thiếu và mất cân bằng lớp vẫn là những yếu tố hạn chế khả năng tổng quát hóa của mô hình.

Tăng cường dữ liệu (Data Augmentation-DA) là một phương pháp quan trọng giúp tạo ra các mẫu huấn luyện mới thông qua biến đổi lên dữ liệu hiện có mà không cần thêm dữ liệu mới [1]. Mục đích chính của DA là nâng cao khả năng tổng quát hóa của mô hình và khắc phục vấn đề thiếu hụt dữ liệu đúng với nhu cầu mà các mô hình phân loại tin giả cần.

Trong nghiên cứu này, chúng tôi tập trung vào việc áp dụng hai kỹ thuật DA phổ biến: **thay thế từ đồng nghĩa và loại bỏ từ chức năng,** nhằm cải thiện chất lượng dữ liệu và hiệu quả của mô hình trong bài toán phân loại tin tức giả. Cả hai kỹ thuật này được kỳ vọng sẽ cải thiện độ đa dạng của tập huấn luyện và giúp mô hình học được những đặc trưng tinh tế hơn từ dữ liệu.

- **Thay thế từ đồng nghĩa** giúp tạo ra các biến thể của văn bản mà vẫn giữ nguyên ý nghĩa: Là kỹ thuật mà chúng tôi thay thế một từ bằng một từ đồng nghĩa của nó, sử dụng WordNet – một cơ sở dữ liệu ngôn ngữ lớn – để tìm các từ đồng nghĩa phù hợp. [2].

- **Giảm chức năng từ** loại bỏ những từ không mang nhiều thông tin trong câu, giúp mô hình tập trung vào các từ khóa quan trọng hơn. có thể hiểu là kỹ thuật Xóa ngẫu nhiên hoặc Loại bỏ từ chức năng, kỹ thuật này loại bỏ ngẫu nhiên các từ trong câu dựa trên một tham số xác suất. Trong trường hợp này, chúng tôi thiết lập xác suất riêng cho các từ chức năng và từ nội dung. [2].

- **Bản gốc:** Cơ sở tham chiếu tập dữ liệu[2].

Các kỹ thuật tăng cường dữ liệu không thể thiếu trong các bài toán phân loại phức tạp như phân loại tin tức giả. Tuy nhiên, sự áp dụng chúng đòi hỏi

phải có sự cân nhắc về thời gian xử lý và tài nguyên tính toán, đặc biệt là khi xử lý với lượng dữ liệu lớn và các mô hình phức tạp. Một trong những thách thức lớn là làm thế nào để đảm bảo rằng các kỹ thuật tăng cường không làm biến đổi ý nghĩa dữ liệu gốc hoặc tạo ra sự nhiễu không cần thiết, điều này có thể ảnh hưởng tiêu cực đến hiệu suất của mô hình.

Bài nghiên cứu này nhằm làm sáng tỏ tác động thực tế của các kỹ thuật tăng cường dữ liệu (DA) đối với hiệu suất phân loại tin tức giả. Chúng tôi sẽ đánh giá kỹ lưỡng các chỉ số hiệu suất để xem liệu việc áp dụng DA có thực sự cải thiện khả năng phân loại của các mô hình học sâu phổ biến trong xử lý ngôn ngữ tự nhiên (NLP) hay không. Bên cạnh đó, nghiên cứu cũng sẽ xác định kỹ thuật DA nào mang lại hiệu quả cao nhất trong việc nâng cao độ chính xác của các mô hình như LSTM (Long Short-Term Memory), BERT (Bidirectional Encoder Representations from Transformers), RNN (Recurrent Neural Networks) và NN (Neural Networks). Những mô hình này đã được chứng minh là rất hiệu quả trong việc học và phân loại văn bản. Nghiên cứu của chúng tôi không chỉ đánh giá tác động của DA mà còn làm rõ tầm quan trọng của kỹ thuật này trong các bài toán phân loại phức tạp, đặc biệt là trong việc phát hiện và xử lý tin tức giả.

Đóng góp quan trọng của nghiên cứu này là đánh giá tác động của các kỹ thuật tăng cường dữ liệu (DA) đối với hiệu quả của mô hình phân loại tin tức giả. Qua việc lựa chọn và áp dụng các kỹ thuật DA phù hợp trong quy trình tiền xử lý dữ liệu, nghiên cứu giúp nâng cao độ chính xác của các hệ thống phân loại và mở ra cơ hội cải tiến các phương pháp DA trong các bài toán phân loại thực tiễn. Mặc dù các kỹ thuật DA đã thể hiện tiềm năng, nghiên cứu này còn góp phần làm rõ tầm quan trọng của chúng và khuyến khích các nghiên cứu tiếp theo nhằm tối ưu hóa và phát triển các phương pháp DA, từ đó nâng cao hiệu suất phân loại tin tức giả trong tương lai.

Phương pháp được thực hiện như sau **(Hình 1)**:

1. **Tiền xử lý dữ liệu**: Tập dữ liệu WELFake được chuẩn hóa bằng cách chuyển chữ thường, xóa URL, thẻ HTML, và các ký tự không phải chữ cái.

2. **Tạo tập dữ liệu gốc:** được sử dụng làm cơ sở so sánh.

3. **Tăng cường dữ liệu:**
   - SR Corpus: Thay thế từ đồng nghĩa.
   - FWD Corpus: Loại bỏ từ chức năng.

4. **Huấn luyện mô hình:** Các mô hình như BERT, LSTM, GRU, RNN, và NN được huấn luyện trên các tập dữ liệu (Original, SR, FWD).

5. **Xác thực mô hình (Validation)**: Đánh giá hiệu quả của các mô hình trên tập dữ liệu kiểm tra.

6. **Phân tích kết quả (Evaluation)**: So sánh các mô hình dựa trên độ chính xác, F1-score và các chỉ số khác để xác định phương pháp tối ưu.

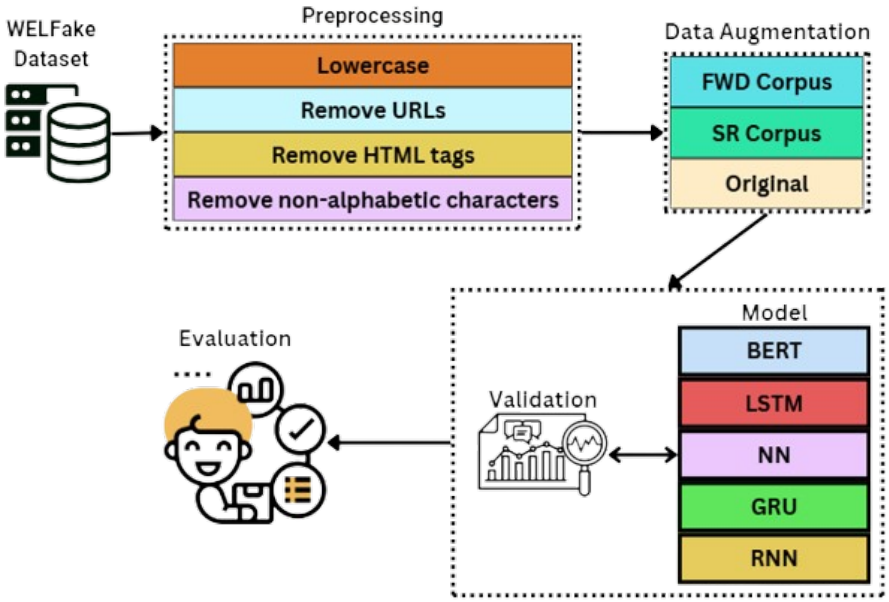
Bài viết được cấu trúc như sau: **Phần thứ hai (Công trình liên quan)** cung cấp tổng quan về các nghiên cứu trước đây trong lĩnh vực tăng cường dữ liệu. **Phần thứ ba (Bộ dữ liệu)** trình bày chi tiết về các tập dữ liệu được sử dụng và quy trình tiền xử lý, bao gồm áp dụng các kỹ thuật thay thế từ đồng nghĩa và loại bỏ từ chức năng để tạo ra các phiên bản tập dữ liệu khác nhau. **Phần thứ tư (Thí nghiệm)** mô tả các bước thiết kế thí nghiệm, thiết lập tham số và cách đánh giá hiệu quả của các phương pháp được đề xuất. **Phần thứ năm (Phân tích kết quả)** tập trung vào việc trình bày các kết quả đạt được và thực hiện phân tích chi tiết. **Phần thứ sáu (Kết luận và định hướng tương lai)** tóm tắt các đóng góp chính của bài nghiên cứu, đồng thời đưa ra các gợi ý cho hướng nghiên cứu tiếp theo.

2 Các công trình liên quan

Trong những năm gần đây, kỹ thuật tăng cường dữ liệu đã trở thành một hướng nghiên cứu quan trọng trong việc nâng cao hiệu suất xử lý ngôn ngữ tự nhiên (NLP) như phân loại tin giả. Việc áp dụng các kỹ thuật tăng cường để tạo ra dữ liệu bổ sung là một trong những phương pháp giúp mô hình học đạt được hiệu suất tốt nếu như bộ dữ liệu khiêm tốn. Các nghiên cứu liên quan phổ biến được nhóm tham khảo về tăng cường dữ liệu văn bản bao gồm **xóa chức năng từ**, **thay thế từ đồng nghĩa**, cũng như tham khảo **các mô hình học sâu (deep learning) cho những kết quả tốt trong việc phân loại tin giả**.

2.1 Tài liệu tham khảo

Nhiều nghiên cứu đã chỉ ra rằng việc loại bỏ các từ chức năng này giúp giảm độ phức tạp của văn



Hình 1: Quy trình đánh giá các kỹ thuật tăng cường đối với từng mô hình Deep Learning trong phân loại tin tức giả.

bản và cải thiện khả năng phân loại. Nghiên cứu của Zhao et al. [3]. đã chỉ ra rằng việc xóa từ chức năng có thể làm tăng độ chính xác của mô hình phân loại tin giả, giúp các mô hình tập trung vào các từ khóa quan trọng hơn trong quá trình phân tích văn bản. Tuy nhiên, cần lưu ý rằng trong một số trường hợp, một số từ chức năng lại mang giá trị ngữ nghĩa quan trọng và việc loại bỏ chúng có thể làm mất đi thông tin cần thiết trong ngữ cảnh của tin giả.

Thay thế từ đồng nghĩa là một kỹ thuật tăng cường có thể được thực hiện thông qua việc sử dụng các từ điển đồng nghĩa như WordNet của Miller [4], hoặc các mô hình word embeddings như Word2Vec của Mikolov et al. [5] và GloVe của Pennington et al.[6] để xác định các từ có nghĩa tương tự và thay thế chúng trong văn bản. Nghiên cứu của Wei và Zou [7] cho thấy rằng thay thế từ đồng nghĩa giúp tạo ra các biến thể của văn bản, làm tăng sự đa dạng của dữ liệu huấn luyện mà vẫn giữ nguyên thông tin ngữ nghĩa. Điều này giúp các mô hình học được những đặc điểm ngữ nghĩa phong phú hơn, từ đó cải thiện độ chính xác trong phân loại tin giả.

2.2 Mô hình tham khảo

Các mô hình học sâu, đặc biệt là LSTM (Long Short-Term Memory) của Hochreiter & Schmidhuber [8] đã được ứng dụng trong phân loại tin giả và cho thấy hiệu quả vượt trội so với các phương pháp truyền thống. Những mô hình này có khả năng nắm bắt các mối quan hệ dài hạn trong dữ liệu văn bản, giúp phân biệt giữa tin giả và tin thật thông qua các đặc điểm ngữ nghĩa của văn bản. Zhang et al. [9] cho thấy việc kết hợp các mô hình học sâu với các kỹ thuật tăng cường dữ liệu như thay thế từ đồng nghĩa có thể giúp cải thiện độ chính xác trong phân loại tin giả bằng cách làm tăng tính đa dạng và phong phú của dữ liệu huấn luyện. Các mô hình học sâu như BERT của Devlin et al. [10] cũng đã chứng minh được hiệu quả vượt trội trong việc phân loại tin giả nhờ vào khả năng hiểu ngữ nghĩa của văn bản trong bối cảnh rộng lớn. Nghiên cứu của Zhang et al. [9] chỉ ra rằng khi kết hợp kỹ thuật tăng cường dữ liệu với các mô hình học sâu như BERT, mô hình không chỉ cải thiện độ chính xác mà còn có khả năng hiểu sâu hơn các đặc điểm ngữ nghĩa và cấu trúc của văn bản, từ đó giúp phát hiện tin giả chính xác hơn.

3 Bộ dữ liệu

3.1 Bộ dữ liệu gốc

Bộ dữ liệu được sử dụng trong bài báo là **WELFake**, được tải xuống từ **Kaggle.com**. Bộ dữ liệu này bao gồm tổng cộng **72,134 bài viết** và lấy các thông tin từ **Kaggle Fake News, McIntire Fake News, BuzzFeed Political News**, trong đó:

- **35,028 bài viết** được gán nhãn là tin giả .
- **37,106 bài viết** được gán nhãn là tin thật.

Bộ dữ liệu WELFake chứa ba cột chính:

- **Title**: Tiêu đề của bài viết.
- **Text**: Nội dung chính của bài viết.
- **Label**: Nhãn của bài viết (**0 = tin giả**, **1 = tin thật**).

3.2 Tiền xử lý cơ bản

Bước đầu tiên trong quy trình là làm sạch và chuẩn hóa dữ liệu văn bản nhằm đảm bảo đầu vào đồng nhất. Các thao tác cụ thể bao gồm:

- **Chuyển chữ thường**: Toàn bộ văn bản được chuyển thành chữ thường để loại bỏ sự phân biệt chữ hoa và chữ thường.
- **Loại bỏ URL**: Các liên kết URL được xóa bằng biểu thức chính quy để giảm nhiễu.
- **Xóa HTML tags**: Các thẻ HTML được loại bỏ để giữ lại nội dung văn bản sạch.
- **Loại bỏ ký tự đặc biệt và số**: Mục tiêu là giữ lại từ chữ cái.

Kết quả của bước này là một tập dữ liệu văn bản sạch, đồng nhất, sẵn sàng cho các bước xử lý tiếp theo.

3.3 Tăng cường dữ liệu

Sau khi hoàn tất bước tiền xử lý cơ, hai kỹ thuật tăng cường dữ liệu được áp dụng để tạo ra hai tập dữ liệu khác nhau nhằm mở rộng và kiểm tra hiệu quả của các phương pháp phân loại.

3.3.1 Thay thế từ đồng nghĩa

Sử dụng WordNet, kỹ thuật thay thế từ đồng nghĩa thay thế các từ trong văn bản bằng từ đồng nghĩa phù hợp nhất, giúp mở rộng dữ liệu huấn luyện. Mỗi từ được kiểm tra danh sách từ đồng nghĩa, và từ đầu tiên được chọn để thay thế.
Phương pháp này tạo ra tập dữ liệu phong phú hơn, giữ vững ý nghĩa chính của câu, đồng thời cải thiện độ đa dạng ngữ liệu cho mô hình học máy.

- **Ví dụ**: Từ ***happy*** được thay thế bằng ***joyful.***

3.3.2 Loại bỏ từ chức năng

Kỹ thuật này sử dụng danh sách từ chức năng của NLTK để loại bỏ những từ như *the*, *is*, *and*,... vốn ít giá trị về mặt ngữ nghĩa. Các từ quan trọng, mang nội dung chính, được giữ lại sau quá trình xử lý tăng cường dữ liệu.
Điều này giúp giảm lượng nhiễu không cần thiết trong dữ liệu, đồng thời đảm bảo rằng các mô hình học máy tập trung vào các từ khóa có ý nghĩa cao nhất.

- **Ví dụ**: Câu ***The cat is sitting on the mat*** sau khi loại bỏ từ chức năng sẽ trở thành ***cat sitting mat.***

3.3.3 Tập dữ liệu đầu ra

Kết quả của bước tăng cường dữ liệu là ba tập dữ liệu:

- **Original**: Tập dữ liệu gốc, không áp dụng bất kỳ kỹ thuật tăng cường nào.
- **SR Corpus**: Tập dữ liệu thay thế từ đồng nghĩa.
- **FWD Corpus**: Tập dữ liệu loại bỏ từ chức năng.

Bước xử lý	**Văn bản sau khi xử lý**
Văn bản gốc	"Running is a great activity! Visit `https://example.com` for more info."
Chuyển chữ thường, xóa URL, HTML, ký tự đặc biệt	"running is a great activity visit for more info"
Tăng cường	**Văn bản sau khi tăng cường**
Thay thế từ đồng nghĩa	"run be a great activity visit for more information"
Xóa bớt chức năng từ	"running great activity visit info"

Bảng 1: Ví dụ về tiền xử lý và tăng cường.

4 Thí nghiệm

4.1 Mô hình Deep Learning

1. **RNN (Recurrent Neural Network)** RNN là một loại mạng nơ-ron có khả năng xử lý dữ liệu tuần tự nhờ vào cơ chế truyền thông tin từ bước trước đến bước sau thông qua trạng thái ẩn. Điều này giúp RNN đặc biệt hiệu quả trong các bài toán liên quan đến chuỗi như xử lý ngôn ngữ tự nhiên (NLP) và dự đoán chuỗi thời gian.

2. **LSTM (Long Short-Term Memory)** LSTM là một biến thể của RNN, được thiết kế để giải quyết vấn đề quên ngữ cảnh trong các chuỗi dài bằng cách sử dụng các cổng kiểm soát (gates). Các cổng này quyết định thông tin nào nên được lưu giữ, cập nhật hoặc loại bỏ, giúp LSTM xử lý hiệu quả hơn dữ liệu tuần tự dài.

3. **BERT (Bidirectional Encoder Representations from Transformers)** BERT là một mô hình ngôn ngữ hiện đại dựa trên kiến trúc Transformer. Nó học ngữ cảnh của từ trong cả hai hướng (trái sang phải và phải sang trái) để tạo ra các biểu diễn ngôn ngữ mạnh mẽ. BERT đã đạt được nhiều thành tựu trong các tác vụ NLP như phân loại, dịch máy, và trả lời câu hỏi.

4. **Neural Network (Mạng Nơ-ron nhân tạo)** Neural Network là một tập hợp các lớp nơ-ron được kết nối để mô phỏng hoạt động của não người. Nó học các mẫu trong dữ liệu thông qua việc điều chỉnh trọng số qua quá trình huấn luyện. Mạng nơ-ron truyền thống thường phù hợp với các bài toán đơn giản hoặc dữ liệu phi tuần tự.

5. **GRU-based Neural Network (Word2Vec)** GRU-based Neural Network là một mô hình mạng nơron sử dụng **GRU (Gated Recurrent Unit)** để học thông tin từ dữ liệu tuần tự, đặc biệt hiệu quả trong xử lý ngôn ngữ tự nhiên (NLP). Mô hình của bạn kết hợp với Word2Vec để tạo ra biểu diễn từ vựng trước khi đưa vào mạng GRU.

4.2 Cấu hình cho nghiên cứu

Nghiên cứu được thực hiện trên nền tảng Google Colab với các cấu hình chính như sau:

- **Nền tảng**: Google Colab.
- **Phần cứng**:
  - GPU: NVIDIA Tesla T4.
  - RAM: 15 GB.
  - Bộ nhớ lưu trữ: 50 GB (Google Drive).
- **Phần mềm**:
  - Python: Phiên bản 3.10.
  - Thư viện sử dụng: TensorFlow, PyTorch, NLTK, scikit-learn, NumPy, Pandas, Transformers (Hugging Face), Regular Expressions (re).

4.3 Độ đo đánh giá

Chúng tôi sử dụng**Accuracy** làm tiêu chí đánh giá chính để đo lường tỷ lệ dự đoán đúng trên toàn bộ tập dữ liệu. Bên cạnh đó, **F1-score (macro average)** được tính toán nhằm cân bằng giữa **Precision** và **Recall**, giúp cung cấp thêm thông tin chi tiết về hiệu suất của mô hình, đặc biệt trong trường hợp dữ liệu không cân bằng.

4.4 Cài đặt mô hình

RNN:

- **Embedding layer** (*embedding_size = 100, input_sequence_length = 200*). - **SimpleRNN** (*units = 128, activation = tanh, return_sequences = False*). - **Dropout** (*rate = 0.5*). - **Dense layer** (*units = 1, activation = sigmoid*).
- *optimizer = Adam, loss = binary_crossentropy, metrics = accuracy*.
- *batch_size = 64, epochs = 30*, sử dụng *EarlyStopping* và *ReduceLROnPlateau*.

LSTM:

- **Embedding layer** (*vocab_size = 10000, embedding_size = 40, input_sequence_length = 500*).
- **LSTM** (*units = 100, return_sequences = False*).
- **Dropout** (*rate = 0.3*).
- **Dense layer** (*units = 1, activation = sigmoid*).
- *optimizer = Adam, loss = binary_crossentropy, metrics = accuracy*.
- *batch_size = 64, epochs = 10*.

BERT:

- **Pretrained BERT Layer** (*model = bert-base-uncased, embeddings = token, position, segment*).
- **Classification Head** (*labels = 2, activation = sigmoid*).
- *optimizer = create_optimizer (learning_rate = 5e-6, num_warmup_steps = 0, total_steps = 1000)*.
- *loss = SparseCategoricalCrossentropy, batch_size = 16, epochs = 3*.

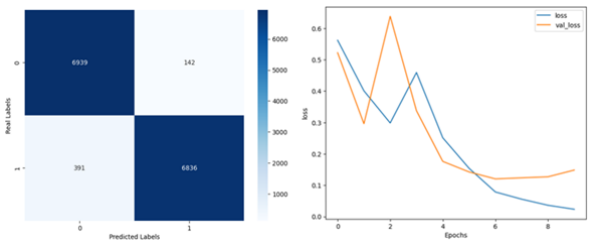
Neural Network:

- **Dense 1** (*units = 512, activation = ReLU, regularization = L2(0.001)*).
- **BatchNormalization** và **Dropout** (*rate = 0.5*).

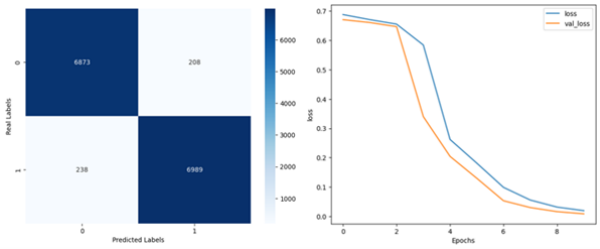
- **Dense 2** (*units = 256, activation = ReLU, regularization = L2(0.001)*).
- **BatchNormalization** và **Dropout** (*rate = 0.5*).
- **Output Dense** (*units = 1, activation = sigmoid*).
- *optimizer = Adam, loss = binary_crossentropy*.
- *batch_size = 32, epochs = 20*, sử dụng *EarlyStopping* và *ReduceLROnPlateau*.

GRU-based Neural Network:

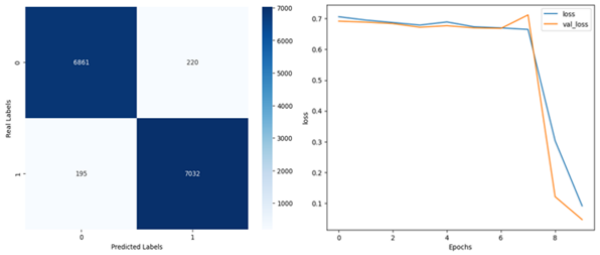
- **Embedding layer** (*input_dim = vocab_size + 1, output_dim = 100, embedding_matrix = (805862, 100), trainable = True*).
- **Dropout** (*rate = 0.2*).
- **GRU 1** (*units = 128, return_sequences = True*).
- **Dropout** (*rate = 0.3*).
- **GRU 2** (*units = 64, return_sequences = False*).
- **BatchNormalization**.
- **Dense 1** (*units = 64, activation = ReLU, regularization = L2*).
- **Dropout** (*rate = 0.4*).
- **Output Dense** (*units = 1, activation = sigmoid*).
- *optimizer = Adam, loss = binary_crossentropy*.
- *batch_size = 64, epochs = 20, EarlyStopping*.



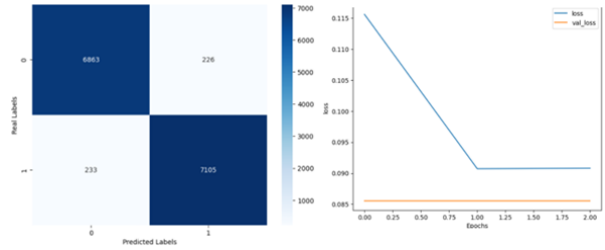
Hình 2: Ma trận nhầm lẫn và biểu đồ Loss của LSTM với tập Original.



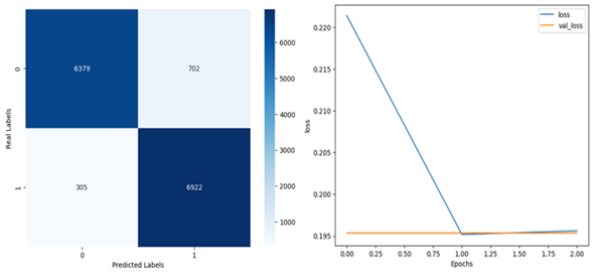
Hình 3: Ma trận nhầm lẫn và biểu đồ Loss của LSTM với tập SR.



Hình 4: Ma trận nhầm lẫn và biểu đồ Loss của LSTM với tập FWD.



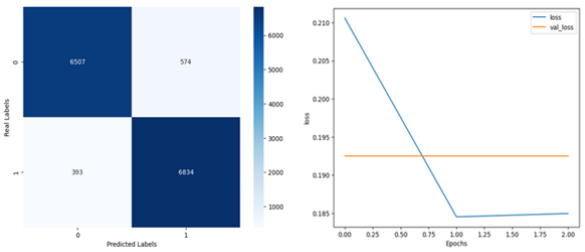
Hình 5: Ma trận nhầm lẫn và biểu đồ Loss của BERT với tập Original.



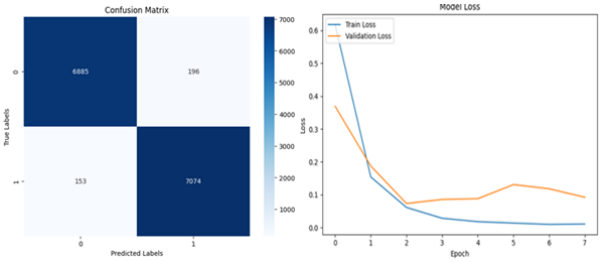
Hình 6: Ma trận nhầm lẫn và biểu đồ Loss của BERT với tập SR.

Classifier	Technique	Accuracy	F1-score	Precision	Recall
BERT (BertTokenizer)	Original	96.32%	96.36%	96.09%	96.64%
	SR corpus	92.96%	93.22%	90.79%	95.78%
	FWD corpus	93.24%	93.39%	92.25%	94.56%
LSTM (One-hot encoding)	Original	96.27%	96.25%	97.97%	94.59%
	SR corpus	96.88%	96.91%	97.11%	96.71%
	FWD corpus	97.10%	97.13%	96.97%	97.30%
RNN (Tokenizer)	Original	94.23%	94.24%	94.97%	93.52%
	SR corpus	90.98%	91.07%	91.01%	91.14%
	FWD corpus	92.87%	92.83%	94.38%	91.32%
Neural Network (TF-IDF)	Original	96.40%	96.45%	95.92%	96.99%
	SR corpus	97.46%	97.50%	97.33%	97.66%
	FWD corpus	97.31%	97.33%	97.92%	96.73%
GRU-based Neural Network (Word2Vec)	Original	97.58%	97.62%	96.70%	98.57%
	SR corpus	96.68%	96.65%	98.62%	94.75%
	FWD corpus	97.56%	97.59%	97.30%	97.88%

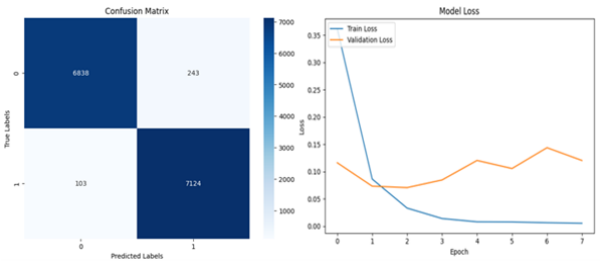
Bảng 2: Hiệu suất của các mô hình với các tập dữ liệu khác nhau.



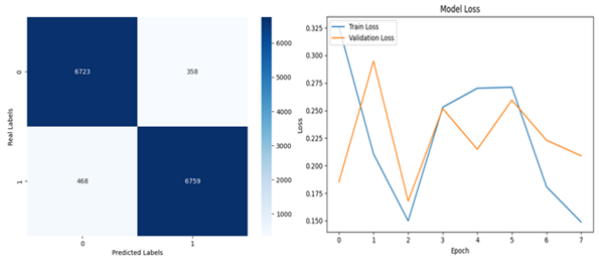
Hình 7: Ma trận nhầm lẫn và biểu đồ Loss của BERT với tập FWD.



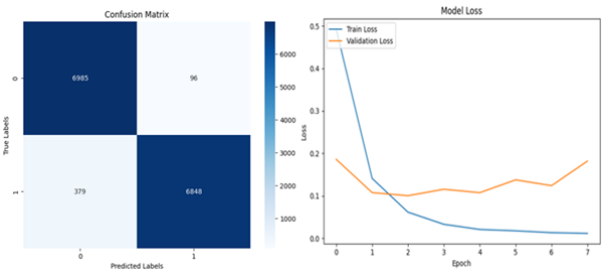
Hình 10: Ma trận nhầm lẫn và biểu đồ Loss của GRU với tập FWD.



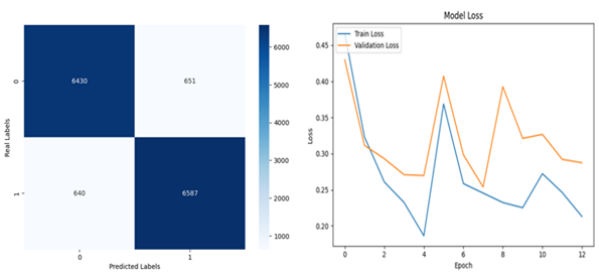
Hình 8: Ma trận nhầm lẫn và biểu đồ Loss của GRU với tập Original.



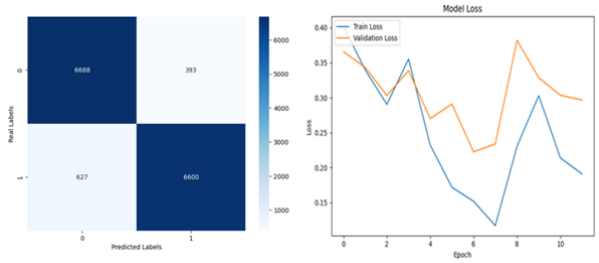
Hình 11: Ma trận nhầm lẫn và biểu đồ Loss của RNN với tập Original.



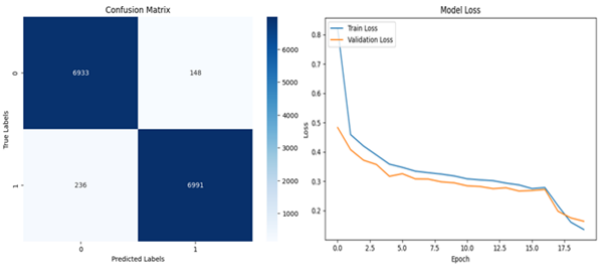
Hình 9: Ma trận nhầm lẫn và biểu đồ Loss của GRU với tập SR.



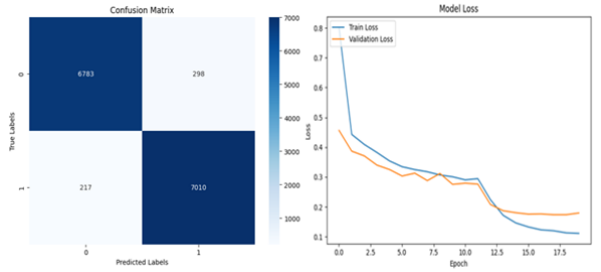
Hình 12: Ma trận nhầm lẫn và biểu đồ Loss của RNN với tập SR.



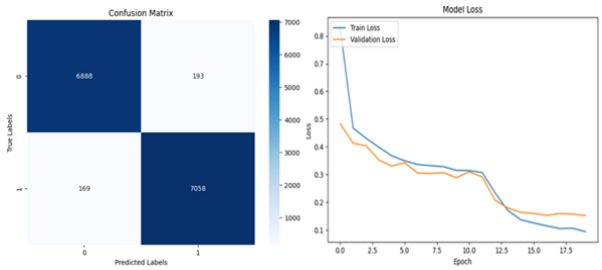
Hình 13: Ma trận nhầm lẫn và biểu đồ Loss của RNN với tập FWD.



Hình 14: Ma trận nhầm lẫn và biểu đồ Loss của Neural Network với tập FWD.



Hình 15: Ma trận nhầm lẫn và biểu đồ Loss của Neural Network với tập Original.



Hình 16: Ma trận nhầm lẫn và biểu đồ Loss của Neural Network với tập SR.

4.5 Kết quả thí nghiệm

Ma trận nhầm lẫn:

- **GRU-based Neural Network (Word2Vec)** và **Neural Network** chứng tỏ sự ổn định và hiệu suất ổn, với số lượng đoán sai (False Positive và False Negative) khá thấp.
- **LSTM** cũng đạt hiệu suất khá tốt, đặc biệt trên tập dữ liệu gốc. Tuy nhiên, khi xử lý các tập dữ liệu tăng cường, số lỗi dự đoán tăng lên đáng kể.
- **RNN** và **BERT**, dù có khả năng phân loại nhất định, lại thể hiện tỷ lệ lỗi False Positive hoặc False Negative cao hơn. Điều này đặc biệt rõ rệt ở các tập dữ liệu tăng cường như SR hoặc FWD.

Biểu đồ Loss:

- **GRU-based Neural Network (Word2Vec) , Neural Network** và **LSTM** đã học tốt từ dữ liệu mà không bị quá phụ thuộc vào tập huấn luyện.
- **RNN** và **BERT** lại cho thấy sự bất ổn định trong quá trình học, đồng thời cho thấy các mô hình này cần được tinh chỉnh thêm để xử lý tốt hơn các biến đổi dữ liệu.

Bảng 2 so sánh hiệu suất của các mô hình khác nhau (**BERT**, **LSTM**, **RNN**, **Neural Network**,**GRU-based Neural Network (Word2Vec)**) trên các tiêu chí: **Accuracy**, **F1-score**, **Precision**, và **Recall**. **Hiệu suất tốt trên tập dữ liệu gốc:**

- **GRU-based Neural Network (Word2Vec)** đạt hiệu suất cao nhất với **Accuracy** là **97.58%**, **F1-score** là **97.62%**, **Precision** là **96.70%**, và **Recall** là **98.57%**.
- **Neural Network (TF-IDF)** cũng đạt kết quả tốt với **Accuracy** là **96.40%**, **F1-score** là **96.45%**, **Precision** là **95.92%**, và **Recall** là **96.99%**.
- **LSTM (One-hot encoding)** đạt **Accuracy** là **96.27%**, **F1-score** là **96.25%**, **Precision** là **97.97%**, và **Recall** là **94.59%**.

Hiệu suất của Thay thế từ đồng nghĩa(SR corpus):

- Hiệu suất của hầu hết các mô hình giảm nhẹ khi áp dụng SR corpus, ngoại trừ **Neural Network (TF-IDF)**, đạt **Accuracy** là **97.46%**, gần tương đương với tập dữ liệu gốc.
- **RNN (Tokenizer)** có hiệu suất giảm đáng kể với **Accuracy** là **90.98%**, **F1-score** là **91.07%**.

- **BERT (BertTokenizer)** cũng giảm hiệu suất với **Accuracy** là **92.96%**, **F1-score** là **93.22%**, và **Precision** chỉ đạt **90.79%**.

Hiệu suất của Loại bỏ từ chức năng (FWD corpus):

- Một số mô hình như **GRU-based Neural Network (Word2Vec)** và **Neural Network (TF-IDF)** vẫn duy trì hiệu suất cao với **Accuracy** lần lượt là **97.56%** và **97.31%**.
- Tuy nhiên, **BERT** tiếp tục giảm hiệu suất đáng kể với **Accuracy** là **93.24%**, và **Precision** chỉ đạt **92.25%**.
- **RNN (Tokenizer)** cũng có sự suy giảm với **Accuracy** là **92.87%** và **Recall** là **91.32%**.

Nhận xét chung:

- **GRU-based Neural Network (Word2Vec)** là mô hình hiệu quả và ổn định nhất trên tất cả các loại tập dữ liệu, với **Accuracy** và **Recall** cao nhất trong mọi trường hợp.
- **Neural Network (TF-IDF)** cũng cho thấy hiệu suất tốt khi xử lý các tập dữ liệu tăng cường như SR corpus và FWD corpus.
- **LSTM** hoạt động tốt trên tập dữ liệu gốc, nhưng lại giảm hiệu quả khi làm việc với các tập tăng cường.
- Các mô hình như **BERT** và **RNN** cho thấy hiệu suất giảm đáng kể khi sử dụng SR corpus và FWD corpus.
- **Tập dữ liệu gốc** luôn cho hiệu suất tốt nhất, cho thấy phương pháp tăng cường dữ liệu của nghiên cứu vẫn chưa tốt và cần cải thiện thêm.

5 Thảo Luận

Không phải kỹ thuật tăng cường dữ liệu văn bản nào cũng dễ triển khai như nhau. Kỹ thuật **Loại bỏ từ chức năng** đơn giản hơn khi chỉ yêu cầu loại bỏ các từ dừng (stop words) và sử dụng thư viện để xác định các thẻ từ loại POS tags. **Thay thế từ đồng nghĩa** là kỹ thuật phức tạp nhất về mặt triển khai, khi cần làm việc với các danh sách tập hợp từ đồng nghĩa và các thư viện hỗ trợ, khi áp dụng cho các ngôn ngữ khác tiếng Anh là một khó khăn vì tiếng Anh có sẵn các tập synset trong thư viện [2] .

Mặc dù tăng cường dữ liệu thường được kỳ vọng cải thiện hiệu suất, kết quả nghiên cứu cho thấy các kỹ thuật tăng cường như Thay thế từ đồng nghĩa và Loại bỏ từ chức năng có thể gây tác động không mong muốn lên hiệu suất mô hình, có thể do một số nguyên nhân:

Những kết quả này cho thấy rằng các mô hình Deep Learning khi áp dụng các kỹ thuật này không vượt qua được hiệu suất gốc trong hầu hết các trường hợp

Nghiên cứu này nhấn mạnh rằng không có giải pháp tăng cường dữ liệu nào phù hợp cho tất cả các mô hình hoặc bài toán. Dù việc tăng cường dữ liệu mang lại sự đa dạng cho tập huấn luyện, hiệu suất thực tế của các mô hình học máy phụ thuộc vào khả năng duy trì ngữ nghĩa và cấu trúc của dữ liệu. Nhưng nó cũng cần phải cân nhắc kỹ lưỡng đến đặc điểm của mô hình cũng như dữ liệu. Nghiên cứu cũng mở ra cơ hội để tìm hiểu thêm về cách tối ưu hóa kỹ thuật tăng cường dữ liệu, nhằm nâng cao hiệu quả trong bài toán phân loại tin tức giả.

6 Kết luận

Phân loại tin tức giả là một bài toán không chỉ đầy thách thức mà còn mang tính cấp thiết trong thời đại thông tin số, khi khối lượng dữ liệu khổng lồ được lan truyền nhanh chóng trên các nền tảng truyền thông xã hội. Trong nghiên cứu này, chúng tôi đã tập trung vào việc áp dụng hai kỹ thuật tăng cường dữ liệu - **Thay thế từ đồng nghĩa** và **Loại bỏ từ chức năng** nhằm đánh giá tác động của chúng đối với hiệu suất của các mô hình Deep Learning. Đáng chú ý, **GRU-based Neural Network (Word2Vec)** thể hiện hiệu suất tốt nhất so với năm mô hình trên tất cả các tập dữ liệu, trong khi **Neural Network** cũng đạt được kết quả khá tốt, đặc biệt trên các tập tăng cường như SR và FWD. Mặt khác, các mô hình như **LSTM, BERT** và **RNN** cho thấy hiệu suất không nhất quán khi làm việc với dữ liệu tăng cường, đặc biệt với tập SR và FWD. Điều này cho thấy rằng các kỹ thuật tăng cường hiện tại vẫn còn hạn chế và cần được cải thiện để đạt được hiệu quả tốt hơn trong bài toán phân loại tin giả. Hướng nghiên cứu trong tương lai sẽ tập trung vào một số cải tiến quan trọng sau:

- **Tăng cường dữ liệu định hướng ngữ pháp và ngữ nghĩa:** Phát triển các kỹ thuật tăng cường dựa trên phân tích sâu về ngữ pháp và ngữ nghĩa để đảm bảo rằng ý nghĩa gốc của dữ liệu không bị thay đổi.
- **Hỗ trợ ngôn ngữ ít tài nguyên:** Tạo ra các tập dữ liệu bổ sung và công cụ hỗ trợ, chẳng

hạn từ điển đồng nghĩa số hóa, nhằm mở rộng khả năng xử lý cho các ngôn ngữ ít phổ biến.

- Tích hợp các phương pháp tăng cường dữ liệu với các mô hình học sâu tiên tiến hoặc sử dụng học tăng cường (Reinforcement Learning) để tối ưu hóa hiệu suất huấn luyện.

Nghiên cứu này không chỉ làm sáng tỏ vai trò của tăng cường dữ liệu trong bài toán phân loại tin tức giả, mà còn gợi mở tiềm năng ứng dụng của DA trong việc giải quyết các thách thức như mất cân bằng dữ liệu và nâng cao hiệu suất mô hình. Những phát hiện này không chỉ góp phần cải thiện quy trình phát triển hệ thống phát hiện tin giả mà còn tạo nền tảng quan trọng cho các nghiên cứu trong tương lai, nhằm xây dựng các giải pháp thông minh, hiệu quả hơn cho bài toán phân loại tin tức giả.

References

[1] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy. A survey of data augmentation approaches for nlp. *arXiv*, 2021.

[2] Jozef Kapusta, Dávid Držík, Kirsten Šteflovič, and Kitti Szabó Nagy. Text data augmentation techniques for word embeddings in fake news classification. *IEEE Access*, 12:31538–31550, 2024.

[3] X. Zhao, Z. Zhang, and L. Wang. A comprehensive study of stopword removal in text classification. *Journal of Machine Learning Research*, 42(1):23–36, 2018.

[4] G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 1995.

[5] T. Mikolov et al. Efficient estimation of word representations in vector space. *arXiv*, 2013.

[6] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. *Proceedings of EMNLP*, 2014.

[7] J. Wei and K. Zou. E-da: Enhancing text classification via data augmentation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[8] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[9] L. Zhang, Y. Li, and F. Wu. Enhancing fake news detection with data augmentation and deep learning models. *Journal of Artificial Intelligence Research*, 22(1):1–16, 2021.

[10] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019.