**TF-IDF** is a widely used method in Natural Language Processing (NLP) and text mining [[17], [18]](#re17). It helps assess how significant a word is within a document, based on both the frequency of that word in the document and the frequency of the word in the entire set of documents. In other words, TF-IDF allows us to determine which words are more prominent in a document compared to other documents in the same set.

In this study, we use TF-IDF as a **method to supplement information about the frequency of word occurrences, which helps to enhance the ability to distinguish between important and unimportant words. This contributes to improving the accuracy of fake news classification without interrupting the context learning process of Transformer models.**

**Specifically, TF-IDF is integrated as an additional layer into the features of the [CLS] token generated by Transformer models. This [CLS] token is the result of the hidden state from the final layer and represents the entire sentence, containing context information synthesized from all the words in the sentence. This approach will keep the process of learning context and the relationships between words in the sentence completely handled by the Transformer without being affected by TF-IDF.**

**TF-IDF** là một phương pháp được sử dụng rộng rãi trong xử lý ngôn ngữ tự nhiên (NLP) và khai thác văn bản [17], [18]. Nó giúp đánh giá mức độ quan trọng của một từ trong một tài liệu, dựa trên tần suất của từ đó trong tài liệu và tần suất của từ trong toàn bộ tập hợp tài liệu. Nói cách khác, TF-IDF cho phép chúng ta xác định từ nào nổi bật hơn trong một tài liệu so với các tài liệu khác trong cùng tập hợp

Trong nghiên cứu này, chúng tôi sử dụng TF-IDF như một phương pháp để **bổ sung các thông tin về tần suất xuất hiện của các từ, giúp tăng cường khả năng phân biệt giữa các từ quan trọng và không quan trọng. Việc này góp phần tăng độ chính xác trong việc phân loại tin giả mà không làm gián đoạn quá trình học ngữ cảnh của các mô hình Transformer.**

**Cụ thể, TF-IDF được tích hợp như một tầng bổ sung vào các đặc trưng của token [CLS] đã được các mô hình Transformer tạo ra. Token [CLS] này là kết quả từ hidden state của layer cuối cùng và đại diện cho toàn bộ câu, chứa thông tin ngữ cảnh tổng hợp từ tất cả các từ trong câu. Cách tiếp cận này sẽ giữ cho quá trình học ngữ cảnh và quan hệ giữa các từ trong câu được xử lý hoàn toàn bởi Transformer mà không bị ảnh hưởng bởi TF-IDF**

**Prolem 1:**

"Tác giả trình bày một phương pháp sử dụng các mô hình Transformer để phát hiện tin giả trên các nền tảng mạng xã hội, với ứng dụng cho tiếng Việt. Họ đã xử lý dữ liệu mẫu không cân bằng, sau đó tích hợp kỹ thuật tiền xử lý dữ liệu TF-IDF, và so sánh nhiều mô hình khác nhau.

Tại sao bạn lại chọn TF-IDF trong bước tiền xử lý? Bạn có điều tra và so sánh các phương pháp khác để xử lý tập dữ liệu không?"

=> Chúng tôi chọn TF-IDF vì phương pháp này giúp bổ sung thông tin về tần suất xuất hiện của các từ, giúp tăng cường khả năng phân biệt giữa các từ quan trọng và không quan trọng. Việc này góp phần tăng độ chính xác trong việc phân loại tin giả mà không làm gián đoạn quá trình học ngữ cảnh của các mô hình Transformer.

Chúng tôi cũng đã thêm thông tin này vào mục “2.2.6 TF-IDF” để làm rõ lý do và cách kết hợp TF-IDF cho mô hình của mình.

Sau khi nộp bản thảo, chúng tôi cũng đã triển khai thêm kết hợp với một số phương pháp khác như Word2Vec và đã mang lại kết quả tốt, chúng tôi cũng đã bổ sung thêm phương pháp Word2Vec vào bản thảo của mình để làm đa dạng hơn các kỹ thuật được kết hợp cùng Transformer thay vì chỉ dùng TF-IDF

Chúng tôi đã thêm mục “2.2.7 Word2Vec” gồm các kiến thức về Word2Vec, đồng cũng thời thêm kết quả đánh giá của PhoBERT + Word2Vec trong “Bảng 1” và các nhận xét tương ứng trong mục “4.3 Results”.

**Data processing**

**Sau bước thu thập chúng tôi đã thu được 1100 tin tức, trong đó có 812 tin đã được gán nhãn dựa trên nguồn thu thập từ các trang uy tín của quốc gia hoặc là các thông tin, kiến thức hoặc sự kiện thực tế đã được chứng minh là đúng hoặc sai. Còn lại 288 tin đã được chúng tôi gán nhãn thủ công dựa trên suy luận từ các kiến thức có sẵn, hoặc dựa theo quan điểm của cộng đồng tại thời điểm thông tin được đăng tải. Ngoài ra, các tin tức về mê tín dị đoan hay tuyên truyền chống phá nhà nước, xã hội đều được mặc định là tin giả trong quá trình gán nhãn. Cuối cùng chúng tôi thu được 703 tin thật và 397 tin giả cho bộ dữ liệu của mình.**

**Sau đó,** chúng tôi thực hiện làm sạch dữ liệu qua các bước sau: loại bỏ các mục trống, không hợp lệ hoặc bị trùng lặp, chuyển tất cả văn bản thành chữ thường và loại bỏ các ký tự đặc biệt và URL. Sau đó, chúng tôi chọn các trường dữ liệu sẽ được sử dụng và dữ liệu còn lại sẽ được hiển thị như trong Hình 3. Ở giai đoạn này, chúng tôi chỉ tập trung vào nội dung của các bài đăng và nhãn phân loại của chúng, nhưng chúng tôi dự định mở rộng nghiên cứu để bao gồm phân tích các bình luận trong tương lai.

Rõ ràng, số lượng mẫu tin giả thấp hơn nhiều so với tin thật (được thể hiện trong biểu đồ ở Hình 4), điều này có thể dẫn đến sự thiên lệch trong việc huấn luyện mô hình và kết quả không chính xác. Để giải quyết vấn đề này, chúng tôi đã thực hiện hai giải pháp:

● Thêm Dữ liệu từ VFND: Chúng tôi đã thêm các bài báo từ bộ dữ liệu VFND, được mô tả trong luận văn của Ho Quang Thanh, "VNFD - Vietnamese Fake News Datasets" [19]. Tuy nhiên, vì bộ dữ liệu này được thu thập từ năm 2019-2020, chúng tôi chỉ chọn những tin tức không thay đổi theo thời gian, chẳng hạn như kiến thức khoa học đã bị chứng minh là sai, và những tin tức không thay đổi theo thời gian như tin đồn về mê tín, hoặc lối sống lệch chuẩn. Dữ liệu bổ sung này chiếm không quá 20% tổng số tin giả mà chúng tôi hiện đang thu thập.

● Sử dụng Kỹ thuật Random Oversampling: Chúng tôi đã sử dụng kỹ thuật Random Oversampling từ thư viện "imbalanced-learn" để cân bằng lại dữ liệu. Kỹ thuật này sao chép ngẫu nhiên các mẫu hiện có của nhãn có số lượng ít hơn cho đến khi số lượng mẫu của các nhãn được cân bằng. **Tuy nhiên chúng tôi không sao chép những mẫu tin được chúng tôi gán nhãn thủ công.**

**Prolem 3:**

Chúng tôi đã chỉnh sửa các câu và cách diễn đạt sai được bạn nhắc đến trong bản thảo, chúng tôi rất biết ơn những góp ý của bạn đã giúp bản thảo của chúng tôi mượt mà hơn.

Chúng tôi cũng đã chỉnh sửa lại số lượng mẫu trong bộ dữ liệu là 1406 mẫu ở mục “4.1 Corpus” và cung cấp URL để truy cập đến bộ dữ liệu của chúng tôi. Đồng thời chúng tôi cũng bổ sung thêm thông tin trong mục “3.3 Data processing” để làm rõ trong bộ dữ liệu có bao nhiêu mẫu là được chúng tôi gán nhãn thủ công và cách mà chúng tôi làm điều đó như thế nào.

Cuối cùng chúng tôi cũng đã thay đổi three numbers after the decimal, instead six numbers ở trong bảng 1.

### ****Word2Vec****

Word2Vec (Word to Vector) là một kỹ thuật biểu diễn từ vựng thành các vector số, thường được xây dựng dựa trên dữ liệu là một tập văn bản có kích thước lớn và trả về kết quả là một không gian vector, có thể lên đến hàng trăm chiều. Trong không gian này, mỗi từ trong tập dữ liệu (corpus) sẽ được đại diện bởi một vector, các từ từ có ngữ cảnh, ngữ nghĩa tương tự, thường xuất hiện cùng nhau sẽ có các vectơ với vị trí gần nhau.

Chúng tôi muốn kết hợp các mô hình Transformer cùng Word2Vec để nâng cao hiệu quả trong các bài toán như phân loại tin giả nhờ khả năng chuyển các từ thành các vector ngữ nghĩa, từ đó giúp mô hình hiểu và phân tích các từ một cách chính xác hơn, phản ánh các mối quan hệ ngữ nghĩa và ngữ cảnh. Việc tích hợp Word2Vec giúp mô hình có được thông tin ngữ nghĩa sâu sắc và cải thiện khả năng xử lý dữ liệu văn bản.

Cụ thể trong nghiên cứu này, chúng tôi sử dụng word2vec để khởi tạo lớp embedding của Transformer, cung cấp các vector từ với mối quan hệ ngữ nghĩa làm điểm khởi đầu trước khi fine-tune trên dữ liệu huấn luyện. Trong suốt quá trình huấn luyện, lớp embedding này sẽ được điều chỉnh để học các đặc trưng ngữ cảnh từ dữ liệu huấn luyện, đồng thời vẫn duy trì được thông tin ngữ nghĩa từ word2vec. Phương pháp này giúp giảm tải chi phí tính toán, cải thiện khả năng học ngữ cảnh của Transformer và tối ưu hiệu suất cho các bài toán xử lý ngôn ngữ tự nhiên.

**Nhận xét**

● Roberta: Hiệu suất phân loại của Roberta khá kém, với độ chính xác (Accuracy) là 0.741 và AUC là 0.835. Mặc dù Precision đạt 0.845, nhưng Recall chỉ đạt 0.604, cho thấy mô hình bỏ lỡ nhiều trường hợp tin giả. Điểm F1 là 0.704 cho thấy mặc dù mô hình hoạt động ở mức hợp lý, nhưng không hiệu quả bằng các mô hình khác.

● BERT: Hoạt động tốt hơn Roberta trong phân loại, với độ chính xác là 0.787 và AUC là 0.858. Precision là 0.850 và Recall là 0.708, cho thấy hiệu suất cân bằng giữa việc phát hiện tin giả và tin thật. Điểm F1 là 0.773 cho thấy BERT là một mô hình mạnh, nhưng vẫn chưa phải là tốt nhất trong số các mô hình đã thử.

● PhoBERT: Đạt hiệu suất cao với độ chính xác là 0.872 và AUC là 0.948, cho thấy khả năng phân biệt giữa tin thật và tin giả **tốt**. **Precision đạt 0.851, Recall đạt 0.910 và Điểm F1 là 0.879 đã cho thấy PhoBERT đạt hiệu suất phân loại dành cho Tiếng Việt tốt hơn RoBERTa và BERT,**

● PhoBERT + TF-IDF: **Đạt độ chính xác cao hơn PhoBERT với Accuracy đạt 0.897. Điểm F1-Score (0.891) và AUC (0.958) đều đạt mức tốt**. **Việc kết hợp cùng TF-IDF đã làm giúp PhoBERT ít gán nhầm tin thật thành tin giả, với Precision đạt giá trị cao nhất trong các mô hình là 0.922. Tuy nhiên Recall chỉ đạt 0.862, đã giảm đi khá nhiều so với khi chỉ dùng PhoBERT riêng biệt, điều này mang lại rủi ro nhiều hơn khi mô hình đã bỏ sót khá nhiều tin tức giả trong tổng thể.**

● PhoBERT + Word2Vec: **Đây là mô hình đạt độ chính xác cao nhất trong số các mô hình, với Accuracy lên đến 0.904, cùng với F1-Score (0.905) và AUC (0.961) - đây đều là ba giá trị cao nhất so với các mô hình đã thử nghiệm, khiến nó trở thành mô hình hiệu quả nhất cho nhiệm vụ này. Mặc dù Precision của mô hình này thấp hơn so với PhoBERT + TF-IDF nhưng Recall của mô hình này đạt giá trị cao nhất là 0.927. Điều này cho thấy rằng mô hình này đã cẩn thận hơn trong việc phân loại tin tức, mặc dù đã gán nhầm nhiều tin thật là tin giả nhưng lại phát hiện được nhiều tin giả hơn trong tổng thể - ít mang lại rủi ro hơn cho xã hội.**

**Thảo luận**

Kết quả thực nghiệm chỉ ra rằng **PhoBERT và các phiên bản kết hợp cùng TF-IDF và Word2Vec đều là các mô hình hiệu quả cao trong việc phân tích và phân loại tin giả**. **PhoBERT + Word2Vec đạt được hiệu suất tốt nhất, với các chỉ số Accuracy, Recall và F1 cao nhất, cho thấy khả năng phân loại xuất sắc của mô hình này.** **PhoBERT + TF-IDF cũng có hiệu suất rất tốt, với các chỉ số Accuracy, Precision cao.**

Điều quan trọng cần lưu ý là mặc dù **PhoBERT + Word2Vec có Precision thấp hơn một chút so với PhoBERT + TF-IDF**, **nhưng lại đạt giá trị Recall cao**. Điều này cho thấy **PhoBERT + Word2Vec**  có thể bảo thủ hơn, có thể bỏ sót một số tin tức hợp pháp nhưng lại đưa ra các dự đoán chính xác hơn về tin giả trong tổng thể.

Bảng 2 trình bày một số trường hợp được trích xuất từ bộ dữ liệu kiểm tra **mà đa số các mô hình phân loại chính xác, đặc biệt là PhoBERT và các phiên bản kết hợp cùng TF-IDF và Word2Vec**. Trong hầu hết các trường hợp đơn giản của tin thật và tin giả, chẳng hạn như tin đầu tiên **“Metro Nhon Ga Ha Noi line will operate on August 9, 2024” là một thông tin thật đơn giản, và tin thứ 3 “A horrific scene unfolded as a truck attempted to cross the railway tracks, causing a train to derail. At least 100 people were injured or killed, and passengers were left in a state of panic” là một tin giả điển hình, có ngữ nghĩa mang tính bịa đặt, chứa nhiều từ ngữ thu hút sự chú ý và lan truyền. Đối với những tin tức này tất cả mô hình đều phân loại đúng.**

Tuy nhiên, đối với các trường hợp tin tức có nhiều thông tin hơn, các mô hình như BERT và RoBERTa gặp phải nhiều lỗi phân loại, khiến hai mô hình này có hiệu suất thấp và trở nên không đáng tin cậy. **Ví dụ như tin số 2 “The Olympic organizers canceled the second triathlon training session due to the water quality in the Seine River” là một tin thật nhưng RoBERTa đã phân loại là giả. Ngược lại tin tức số 4 “Shocking, unprecedented phenomenon. Thousands of people waiting in line for diphtheria testing” là một tin giả nhưng RoBERTa và BERT đã phân loại là thật.**

**Mặc dù PhoBERT, PhoBERT + TF-IDF và PhoBERT + Word2Vec** cho thấy độ chính xác cao trong dự đoán, vẫn có một số trường hợp ngoại lệ, đặc biệt là khi tin tức chứa thông tin đúng và sai đan xen **như những trường hợp trong Bảng 3**. **Ví dụ trong trường hợp đầu tiên: “A female employee at Samsung infected 16 people with HIV”, đây là một thông tin được lan truyền rộng rãi gần đây trên mạng xã hội Việt Nam dạo gần đây, và kết quả chỉ có PhoBERT + TF-IDF nhận diện đúng đây là tin giả, trong khi 4 mô hình còn lại đã nhận diện sai rằng đây là một tin thật.**

**Một tin tức giả khác cũng khá phức tạp như trường hợp 2:** “A particularly severe mining accident in Quảng Ninh killed 5 workers,... the puppet party and trade unions have never cared about workers' conditions” chứa thông tin đúng là “A particularly severe mining accident in Quảng Ninh killed 5 workers” nhưng phần bổ sung “the puppet party and trade unions have never cared about workers' conditions” lại không chính xác và không được xác thực. Trong trường hợp này, PhoBERT + TF-IDF **và PhoBERT + Word2Vec** đã phân loại đúng là tin giả, trong khi PhoBERT và các mô hình khác bị nhầm lẫn bởi phần thông tin đúng trong bài viết. PhoBERT + TF-IDF đã thể hiện khả năng chú trọng các từ khóa quan trọng và giảm ảnh hưởng của các từ phổ biến nhưng ít mang tính thông tin. **Trong khi đó Word2Vec cũng đã hỗ trợ PhoBERT trong việc hiểu toàn bộ ngữ nghĩa của câu một cách hoàn chỉnh hơn, chỉ ra rằng** phần thông tin bổ sung thiếu tính xác thực, **mang tính chủ quan** và không nên được coi là thật, từ đó cải thiện độ chính xác của phân loại.

**Ở trường hợp số 3 “A train derailment at Hai Van Nam station caused traffic congestion, forcing hundreds of passengers to transfer by buses”. Đây là một thông tin thật và nhưng chỉ có PhoBERT + Word2Vec diện đúng, các mô hình còn lại kể cả PhoBERT + TF-IDF đã cho kết quả dự đoán sai. Nguyên nhân có thể do TF-IDF đã nhận diện một số cụm từ khóa thường xuất hiện trong tin giả từ các tin tức trong tập huấn luyện như những cụm từ *“hàng trăm”* và *“nhiều giờ liền”* nên đã dẫn đến kết quả phân loại sai. Việc dựa trên tần suất xuất hiện của các từ thông qua TF-IDF đôi khi đã ảnh hưởng đến khả năng hiểu ngữ cảnh của các mô hình và dẫn đến kết quả dự đoán đây là tin giả, dù đây là một tin tức thật.**

**Đặc biệt ở trường hợp cuối cùng “Breaking news from Binh Thuan: A serious passenger bus accident resulted in the driver's death and 11 people being hospitalized for emergency treatment”. Đây là một thông tin thật nhưng tất cả mô hình đã dự đoán sai đây là tin giả. Nguyên nhân có thể là do tin tức này được trình bày với ngữ nghĩa và cấu trúc giống như các tin tức giả giựt gân khác đã được mô hình học trước đó. Các cụm từ như “Tin nóng”, “nghiêm trọng” đã mang sắc thái cường điệu và thu hút sự chú ý, làm sai lệch khả năng phân loại của các mô hình. Những từ này dù có mặt trong nhiều tin tức giả nhưng trong các trường hợp khẩn cấp hoặc thật sự nghiêm trọng cũng xuất hiện trong các tin thật.**

**Trường hợp này cho thấy mặc dù các mô hình Transformer như PhoBERT có tiềm năng, chúng vẫn có thể bị đánh lừa bởi những tin tức chính thống nhưng có ngữ nghĩa cường điệu.**

**Để cải thiện hiệu suất, cần điều chỉnh các mô hình một cách hiệu quả, chẳng hạn như tối ưu hóa tham số và điều chỉnh trọng số của các từ khóa. Việc tinh chỉnh này sẽ giúp mô hình phân biệt tốt hơn giữa tin thật và tin giả, ngay cả khi chúng chứa đựng những ngữ cảnh cường điệu hoặc thông tin đan xen giữa đúng và sai.**

Chúng tôi đã tạo thêm Bảng 3 về các trường hợp lỗi trong phần thảo luận, đồng thời điều chỉnh Bảng 2 để phù hợp với Bảng 3 và nội dung thảo luận thay đổi một chút so với ban đầu sau khi thêm mô hình PhoBERT + Word2Vec