ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN

BÁO CÁO ĐỒ ÁN MÔN HỌC
KHAI THÁC DỮ LIỆU TRUYỀN THÔNG XÃ HỘI

Đề tài:

Phân loại bình luận từ bài đăng trên VOZ

GVHD: ThS. Nguyễn Thành Luân

Nhóm sinh viên thực hiện:

1. Trần Duy Tân	MSSV: 22550020
2. Nguyễn Trường Thọ	MSSV: 22550022
3. Lý Thanh Dương	MSSV: 22550003
4. Phạm Đình Thắng	MSSV: 22550021

Tp. Hồ Chí Minh, 08/04/2024

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

……, ngày……tháng……năm 20…

Người nhận xét

(Ký tên và ghi rõ họ tên)

BẢNG PHÂN CÔNG, ĐÁNH GIÁ THÀNH VIÊN

Bảng phân công công việc, đánh giá thành viên

Họ và tên	MSSV	Công việc thực hiện	Đánh giá
Nguyễn Trường Thọ	22550022	- Viết Word bài thu hoạch báo cáo kết quả thực hiện đồ án. - Thu thập dữ liệu trên diễn đàng VOZ - Dán nhãn cho dữ liệu	Hoàn thành tốt
Trần Duy Tân	22550020	- Viết Word bài thu hoạch báo cáo kết quả thực hiện đồ án. - Dán nhãn cho dữ liệu - Tiền xử lý dữ liệu	Hoàn thành tốt
Lý Thanh Dương	22550003	- Viết Word bài thu hoạch báo cáo kết quả thực hiện đồ án. - Dán nhãn cho dữ liệu - Phân tích thống kê dữ liệu	Hoàn thành tốt
Phạm Đình Thắng	22550021	- Viết Word bài thu hoạch báo cáo kết quả thực hiện đồ án. - Lên ý tưởng góp ý định nghĩa bài toán. - Viết tài liệu cho bài báo. - Thực hiện huấn luyện mô hình.	Hoàn thành tốt

MỤC LỤC

LỜI CẢM ƠN

Lời đầu tiên, chúng em xin gửi lời cảm ơn chân thành đến Trường Đại học Công nghệ Thông tin – ĐHQG TP. Hồ Chí Minh đã tạo cơ hội cho chúng em được học tập tại một ngôi trường có cơ sở vật chất hàng đầu, với chất lượng giảng dạy vô cùng chuyên nghiệp và chiều sâu kiến thức chuyên ngành, và quan trọng nhất là trên tinh thần giảng dạy kiến thức thực tiễn cho chúng em.

Trên hết, em xin gửi lời cảm ơn chân thành nhất đến ThS. Nguyễn Thành Luân người đã dành thời gian và kiến thức sâu rộng để hướng dẫn và hỗ trợ tôi trong quá trình nghiên cứu và phát triển dự án. Sự tận tâm và tư duy sáng tạo của ThS. Nguyễn Thành Luân là nguồn động viên không ngừng giúp tôi vượt qua những thách thức và phát triển những ý tưởng mới.

Tuy nhiên, vì còn nhiều hạn chế về quỹ thời gian và kinh nghiệm thực hành nên kết quả báo cáo đồ án này sẽ không thể tránh được những thiếu sót. Chúng em rất mong nhận được sự cảm thông, nhận xét đóng góp từ các Quý Thầy Cô để chúng em có điều kiện bổ sung, điều chỉnh và nâng cao kiến thức chuyên môn để phục vụ tốt hơn cho công tác thực tế sau này.

Chúng em xin chân thành cảm ơn!

Sau đây, nhóm chúng em sẽ trình bày tổng quan báo cáo kết quả thực hiện đồ án kết thúc môn học Khai thác dữ liệu tuyền thông với đề tài phân loại bình luận trên bài đăng VOZ .Khi điểm qua các phần trong bố cục các chương nội dung sau:

1. Giới thiệu

Với sự phát triển mạnh mẽ về công nghệ thông tin các trang mạng xã hội, diễn đàn về đời sống xã hội như VOZ, TinhTe…Đối với những trang mạng xã hội đó được nhiều người dùng và tin dùng nên lượng dữ liệu trên các trang khá là lớn và phong phú. Trong bài báo cáo này chúng tôi sẽ xây dựng các mô hình phân loại bình luận từ đó phân tích các lời nói này nhằm mong muốn có thể tìm được những góp ý mang tính xây dựng và ẩn đi bớt nhưng ý kiến không xây dựng để có thể có một môi trường tốt hơn trên các nền tảng. Chúng tôi sử dụng dữ liệu chủ yếu thu thập trên VOZ và các chủ về nói về sách báo, thiết bị điện tử.

2. Nghiêm cứu mô hình

Trong phần này, chúng tôi trình bày về bài toán phân loại và các phương pháp học máy sử dụng cho bài toán cần giải quyết.

Các mô hình:

1. **vinai/phobert-base**

   PhoBERT: Mô hình ngôn ngữ được đào tạo sẵn cho tiếng Việt:

   Các mô hình PhoBERT được đào tạo trước là mô hình ngôn ngữ tiên tiến dành cho người Việt Nam (Phở, tức là "Phở", là một món ăn phổ biến ở Việt Nam):

   Hai phiên bản PhoBERT “base” và “large” là mô hình ngôn ngữ đơn ngữ quy mô lớn đầu tiên được đào tạo trước cho tiếng Việt. Phương pháp đào tạo trước PhoBERT dựa trên RoBERTa giúp tối ưu hóa quy trình đào tạo trước BERT để có hiệu suất mạnh mẽ hơn.

   PhoBERT vượt trội hơn các phương pháp tiếp cận đơn ngữ và đa ngôn ngữ trước đây, đạt được hiệu suất tiên tiến mới trên bốn nhiệm vụ NLP tiếng Việt ở hạ lưu là gắn thẻ Một phần giọng nói, Phân tích cú pháp phụ thuộc, Nhận dạng thực thể được đặt tên và Suy luận ngôn ngữ tự nhiên.

2. **FPTAI/vibert-base-cased**

   FPTAI/vibert-base-cased là một mô hình ngôn ngữ tự nhiên (NLP) được phát triển bởi FPT.AI, một công ty công nghệ ở Việt Nam. Mô hình này dựa trên kiến trúc "ViBERT" (Vietnamese BERT) và sử dụng tiền xử lý dựa trên mô hình BERT (Bidirectional Encoder Representations from Transformers) của Google. Mô hình

này được huấn luyện trên dữ liệu tiếng Việt và có thể được sử dụng cho nhiều tác vụ NLP khác nhau như phân loại văn bản, dịch máy, hoặc sinh văn bản. "cased" trong tên của mô hình chỉ ra rằng các ký tự viết hoa và viết thường được phân biệt trong quá trình huấn luyện, giúp cải thiện khả năng xử lý ngôn ngữ tự nhiên trong các trường hợp đòi hỏi sự nhạy cảm với kiểu chữ cá nhân.

3. **uitnlp/visobert**

uitnlp/visobert là một mô hình ngôn ngữ tự nhiên (NLP) được phát triển bởi UIT NLP Research Group, một nhóm nghiên cứu tại Đại học Công nghệ Thông tin (UIT) - Đại học Quốc gia Thành phố Hồ Chí Minh, Việt Nam. Mô hình này là một biến thể của BERT (Bidirectional Encoder Representations from Transformers) được huấn luyện trên dữ liệu tiếng Việt và được tinh chỉnh cho các tác vụ NLP cụ thể như phân loại văn bản, dịch máy, hoặc sinh văn bản.

So sánh

1. **Giống nhau**

Kiến trúc cơ bản: Cả ba mô hình đều dựa trên kiến trúc BERT, sử dụng cơ chế self-attention để hiểu ngữ cảnh hai chiều trong văn bản.

Mục đích sử dụng: Được huấn luyện để hiểu và xử lý ngôn ngữ tiếng Việt, hỗ trợ cho các nhiệm vụ xử lý ngôn ngữ tự nhiên (NLP) như phân loại văn bản, phân tích cảm xúc, nhận diện thực thể đặt tên, dịch máy, v.v.

Dữ liệu huấn luyện: Tất cả mô hình đều được huấn luyện trên các bộ dữ liệu tiếng Việt lớn để nắm bắt ngữ cảnh và ngữ pháp tiếng Việt.

Kích thước mô hình: viBERT (base), ViSoBERT (base), PhoBERT base là tương đương.

2. **Khác nhau**

Nguồn gốc và tổ chức phát triển:

PhoBERT: Được phát triển bởi Vingroup AI (VinAI), với hai phiên bản là base và large, và sau đó là base-v2.

viBERT: Được phát triển bởi FPT AI.

ViSoBERT: Được phát triển bởi UIT NLP Lab (Đại học Công Nghệ Thông Tin,

ĐHQG Việt Nam), với một phiên bản base được công bố.

Đặc điểm đào tạo và tối ưu hóa:

PhoBERT: Sử dụng các kỹ thuật token hóa đặc biệt (dựa trên byte pair encoding - BPE) phù hợp với đặc điểm của ngôn ngữ.

viBERT: Sử dụng các phương pháp tiền xử lý và token hóa khác biệt so với PhoBERT.

ViSoBERT: Có thêm tính năng "social", được tối ưu cho việc hiểu và xử lý ngôn ngữ trong các ngữ cảnh xã hội trực tuyến, có thể bao gồm việc nhận diện và xử lý ngôn ngữ tự nhiên trong các mạng xã hội hoặc các dạng văn bản không chính thống.

Ứng dụng:

PhoBERT và **viBERT** đều có thể được ứng dụng rộng rãi trong các tác vụ NLP tiếng Việt, từ phân loại văn bản, nhận diện thực thể, đến dịch máy và sinh văn bản.

ViSoBERT với đặc điểm được tối ưu cho ngữ cảnh xã hội, có thể được ưu tiên sử dụng trong các tác vụ liên quan đến xử lý dữ liệu từ mạng xã hội, như phân tích cảm xúc, nhận diện ý kiến, hoặc phát hiện tin tức giả mạo.

Định nghĩa bài toán

Thu thập các nhận xét, bình luận của cá nhân trên VOZ về các chủ đề trong xã hội và công nghệ. Để đảm bảo các thông tin thu được đa dạng và phong phú, chúng em đã giữ lại toàn bộ các bình luận mà chúng em thu thập được.

Và đánh giá theo 2 tiêu chí sau:

Mang tính xây dựng: các bình luận cung cấp phản hồi, ý kiến hoặc góp ý một cách lịch sự, mang tính xây dựng, có ích cho cộng đồng.

Không mang tính xây dựng: bao gồm các bình luận tiêu cực, sử dụng ngôn từ không phù hợp, gây mất đoàn kết hoặc thông tin cung cấp không có giá trị.

3. Bộ dữ liệu

Thu thập data:

Sử dụng công nghệ Python với các thư viện hỗ trợ lấy dữ liệu từ trang web: Requests, BeautifulSoup, Pandas, Regular Expressions. Thu thập các nhận xét, bình luận của cá nhân trên VOZ về các chủ đề trong xã hội. Để đảm bảo các thông tin thu được đa dạng và phong phú, chúng em đã giữ lại toàn bộ các bình luận mà chúng em thu thập được.

Và đánh giá theo 2 tiêu chí sau: Phân loại bình luận thành 2 nhãn:

Mang tính xây dựng là các bình luận cung cấp phản hồi, ý kiến hoặc góp ý một

cách lịch sự, mang tính xây dựng, có ích cho cộng đồng.

Không mang tính xây dựng bao gồm các bình luận tiêu cực, sử dụng ngôn từ không phù hợp, gây mất đoàn kết hoặc không cung cấp giá trị thực sự nào.

Dataset:

Trong bài báo này chúng em sử dụng bộ dataset chúng em tự thu nhập từ các bình luận người dùng về sách báo, thiết bị điện tử trên diễn đàn VOZ.Lựa chọn phải có tính tương tác cao và không có lọc các bình luận mang tính chất đóng góp hay không mang tính đóng góp.

Bảng 1. Chú thích bộ dataset.

Nhãn	Mô tả	Ví dụ
Không mang tính xây dựng (0)	Bình luận có thông tin nhưng không mang tính đóng góp.	Bình luân: là dùng 4G thoải mái không bác ơi
Mang tính xây dựng (1)	Bình luận có thông tin mang tính đóng góp.	Bình luận: Mình xài màn 32" FullHD, đúng hơn là TV, từ năm 2009 rồi, nên đủ biết to nhỏ ra sao

4. Phương pháp

Tiền xử lý dữ liệu

STOPWORDS[3] là những từ xuất hiện nhiều trong ngôn ngữ tự nhiên, tuy nhiên lại không mang nhiều ý nghĩa. Ở tiếng việt StopWords là những từ như: để, này, kia... Tiếng anh là những từ như: is, that, this... Tham khảo thêm tại danh sách stopwords[2] trong tiếng việt. Có rất nhiều cách để loại bỏ StopWords nhưng có 2 cách chính là:

– Dùng từ điển: cách này đơn giản nhất, chúng ta tiến hành filter văn bản, loại bỏ những từ xuất hiện trong từ điển StopWord.

– Dựa theo tần suất xuất hiện của từ: với cách này, chúng ta tiến hành đếm số lần xuất hiện của từng từ trong data sau đó sẽ loại bỏ những từ xuất hiện nhiều lần (cũng có thể là ít lần). Khoa học đã chứng minh những từ xuất hiện nhiều nhất thường là những từ không mang nhiều ý nghĩa.

Tokenizer[3] là quá trình lấy văn bản (chẳng hạn như một câu) và chia nó thành các thuật ngữ riêng lẻ (thường là các từ). Một lớp Tokenizer đơn giản cung cấp chức năng này. Ví dụ dưới đây cho thấy cách chia câu thành các chuỗi từ. Ví dụ: Phiên trần / luận tội → Phiên / điều trần / luận tội RegexTokenizer cho phép mã hóa nâng cao hơn dựa trên đối sánh biểu thức chính quy (regex). Theo mặc định, tham số "pattern" (regex, default: " s +") được sử dụng làm dấu phân tách để tách văn bản đầu vào. Ngoài ra, người dùng có thể đặt tham số "khoảng trống" thành false cho biết "mẫu" regex biểu thị "mã thông báo" thay vì chia nhỏ khoảng trống và tìm tất cả các lần xuất hiện phù hợp dưới dạng kết quả mã hóa.

5. Thực hiện

1. Pre-trained model

No.	Model name	Link
1	vinai/phobert-base	https://huggingface.co/vinai/phobert-base
2	FPTAI/vibert-base-cased	https://huggingface.co/FPTAI/vibert-base-cased
3	uitnlp/visobert	https://huggingface.co/uitnlp/visobert

Thông số	**Value**	**Note**
Số lượng records	4954	
Labels	[1 0]	**1** mang tính xây dựng **0** không mang tính xây dựng
batch_size	16	
train_size	90%	Tập huấn luyện
test_size	10%	Tập kiểm tra sau khi model được train xong
Max_length	256	Giới hạn của 3 model vinai/phobert-base, FPTAI/vibert-base-cased, uitnlp/visobert là 512

		token.
Learning rate	lr=5e-5	0.00005

2. Huấn luyện mô hình

2.1 Revision 1 (FPTAI/vibert-base-cased 1 epochs)

Params training model

No.	Hyperparams	Value
1	epoch	1
2	threshold	0.5

Evaluation Results

No.	Metrics	Value
1	Accuracy	0.6230
2	Precision	0.5600
3	Recall	0.6452
4	F1 Score	0.5996

2.2 Revision 2 (FPTAI/vibert-base-cased 2 epochs)

Params training model

No.	Hyperparams	Value
1	epoch	2
2	threshold	0.5

Evaluation Results

No.	Metrics	Value

1	Accuracy	0.6371
2	Precision	0.5595
3	Recall	0.4700
4	F1 Score	0.5109

2.3 Revision 3 (**FPTAI/vibert-base-cased 3 epochs)**

Params training model

No.	**Hyperparams**	**Value**
1	epoch	3
2	threshold	0.5

Evaluation Results

No.	**Metrics**	**Value**
1	Accuracy	0.6452
2	Precision	0.5874
3	Recall	0.6093
4	F1 Score	0.5982

2.4 Revision **1 (uitnlp/visobert 1 epochs)**

Params training model

No.	**Hyperparams**	**Value**
1	epoch	1
2	threshold	0.5

Evaluation Results

No.	**Metrics**	**Value**
1	Accuracy	0.6371
2	Precision	0.7333

3	Recall	0.2115
4	F1 Score	0.3284

2.5 Revision 2 (uitnlp/visobert 2 epochs)

Params training model

No.	Hyperparams	Value
1	epoch	2
2	threshold	0.5

Evaluation Results

No.	Metrics	Value
1	Accuracy	0.6976
2	Precision	0.6667
3	Recall	0.6486
4	F1 Score	0.6575

2.6 Revision 3 (uitnlp/visobert 3 epochs)

Params training model

No.	Hyperparams	Value
1	epoch	3
2	threshold	0.5

Evaluation Results

No.	Metrics	Value
1	Accuracy	0.6815
2	Precision	0.6034
3	Recall	0.6796
4	F1 Score	0.6393

2.7 Revision 1 (vinai/phobert-base 1 epochs)

Params training model

No.	Hyperparams	Value

1	epoch	1
2	threshold	0.3

Evaluation Results

No.	Metrics	Value
1	Accuracy	0.4476
2	Precision	0.4258
3	Recall	0.9466
4	F1 Score	0.5873

2.8 Revision 2 (vinai/phobert-base 2 epochs)

Params training model

No.	Hyperparams	Value
1	epoch	1
2	threshold	0.3

Evaluation Results

No.	Metrics	Value
1	Accuracy	0.6411
2	Precision	0.5640
3	Recall	0.5805
4	F1 Score	0.5721

2.9 Revision 3 (vinai/phobert-base 3 epochs)

Params training model

No.	Hyperparams	Value
1	epoch	1
2	threshold	0.3

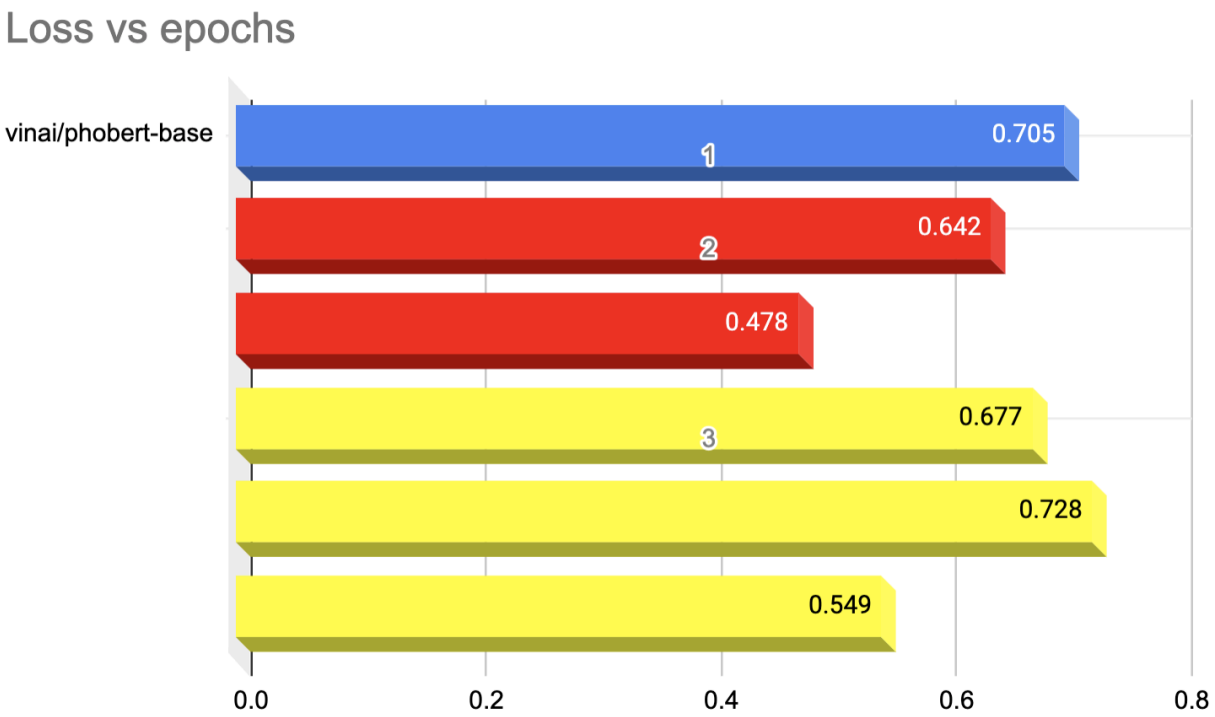
Evaluation Results

No.	Metrics	Value
1	Accuracy	0.5827

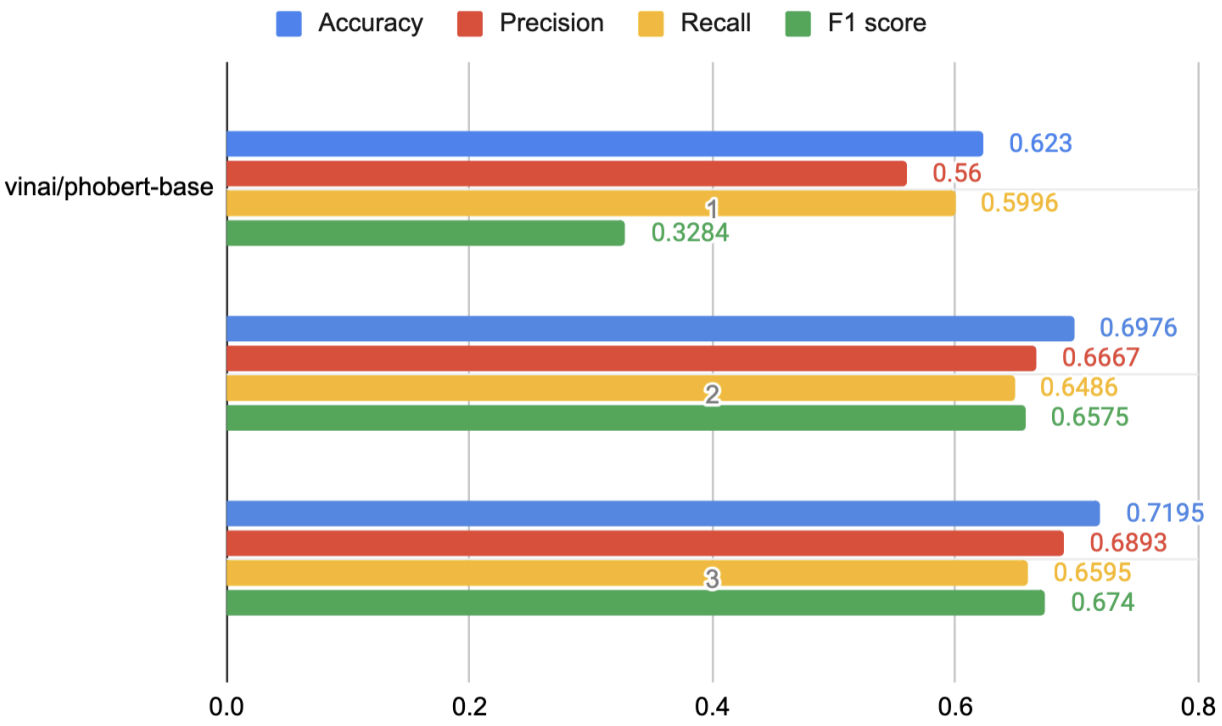
2	Precision	0.0000
3	Recall	0.0000
4	F1 Score	0.0000

6 Kết quả và đánh giá mô hình

1. Kết quả và đánh giá mô hình (vinai/phobert-base)

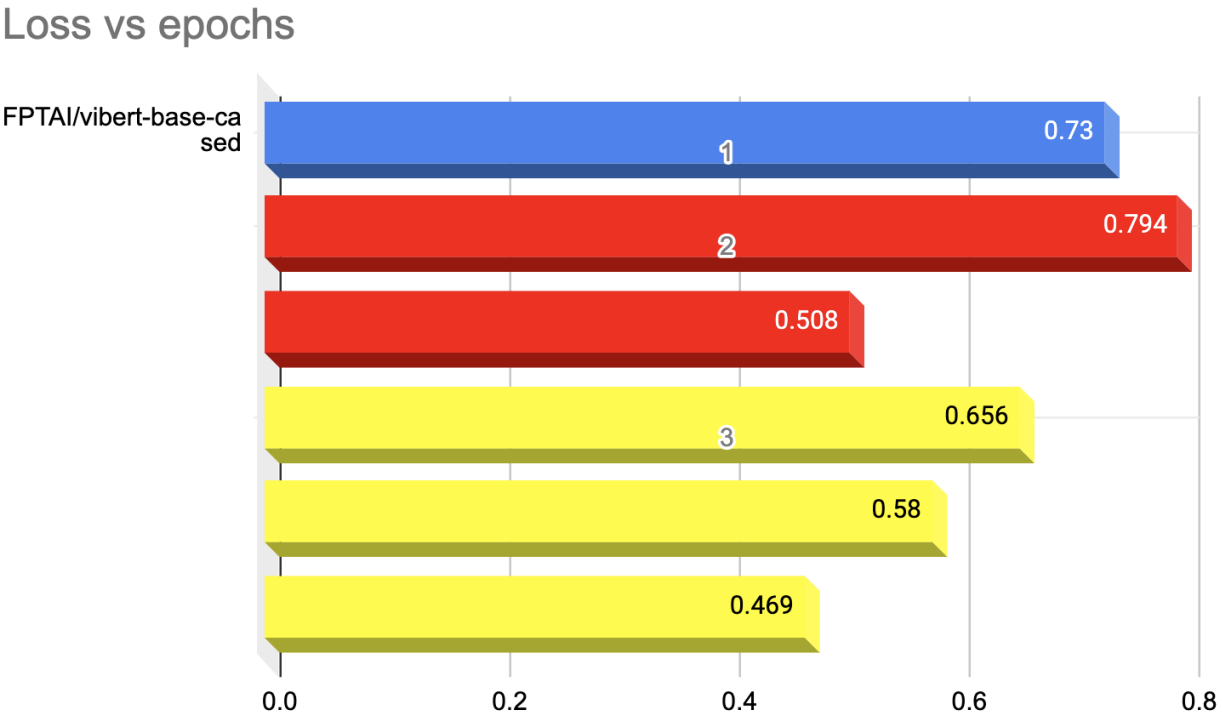


Hình 2 : Kết quả mô hình của vimai/phobert-base

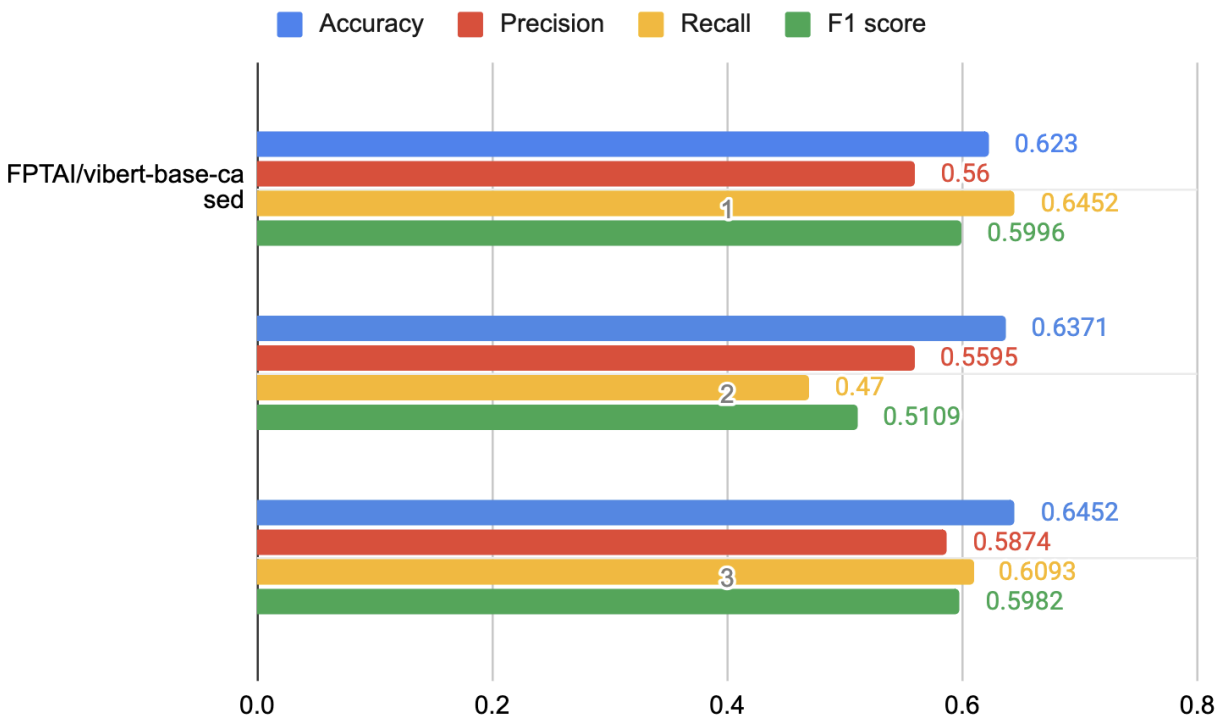


Hình 3 : Đánh giá mô hình của vimai/phobert-base

2. Kết quả và đánh giá mô hình (FPTAI/vibert-base-cased)

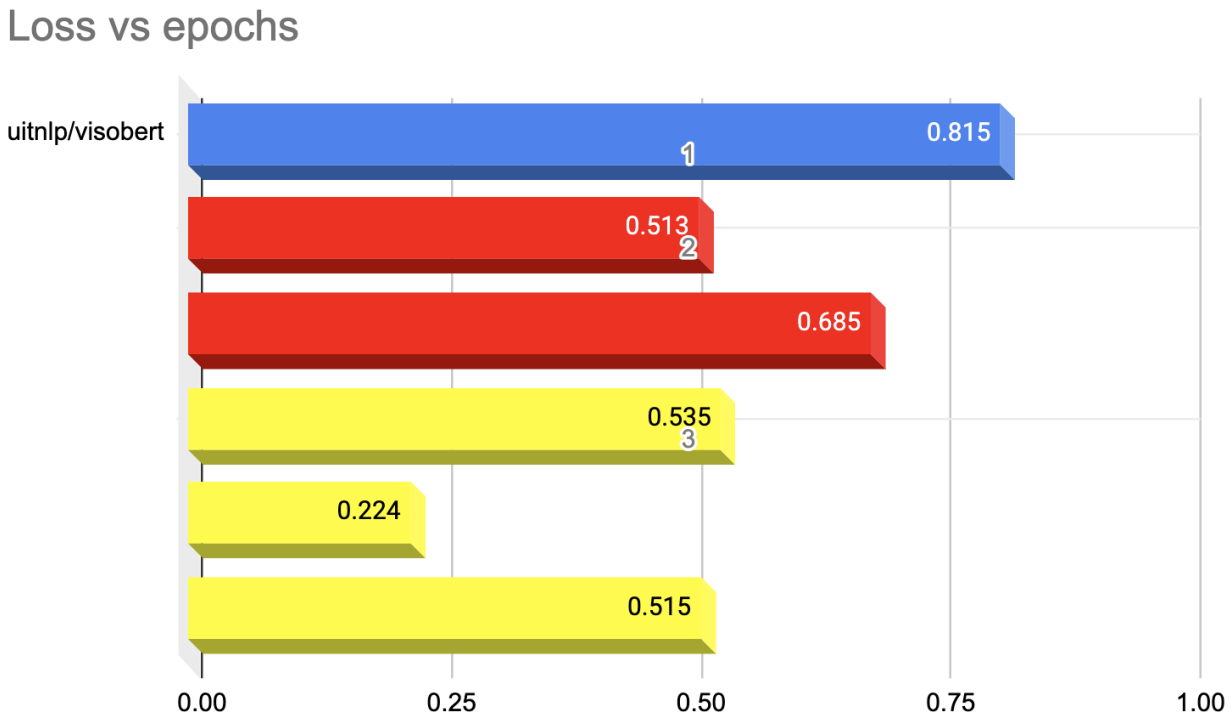


Hình 4 : Kết quả mô hình của FPTAI/vibert-base-cased

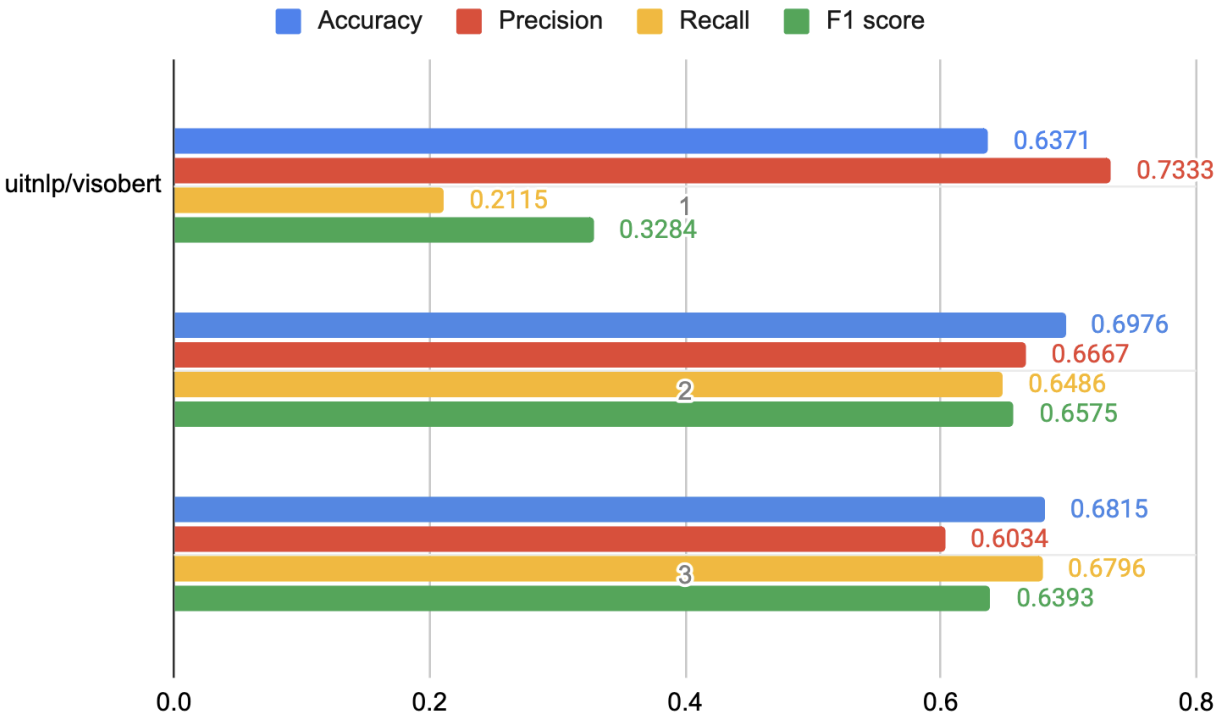


Hình 5 : Đánh giá mô hình của FPTAI/vibert-base-cased

3.Kết quả và đánh giá mô hình (uitnlp/visobert)



Hình 6: Kết quả của mô hình của uitnlp/visobert



Hình 7: Đánh giá mô hình của uitnlp/visobert

6. So sánh chung giữa các model

vinai/phobert-base

Epoch cuối (2): Loss: 0.642, Accuracy: 0.6976, Precision: 0.6667, Recall: 0.6486, F1 Score: 0.6575. Mô hình đạt được sự cân bằng khá tốt giữa Precision và Recall, dẫn đến F1 Score khá cao.

FPTAI/vibert-base-cased

Epoch cuối (3): Loss: 0.656, Accuracy: 0.6452, Precision: 0.5874, Recall: 0.6093, F1 Score: 0.5982. Mô hình có chỉ số F1 Score thấp hơn PhoBERT một chút, cho thấy có sự cân bằng giữa Precision và Recall nhưng không hiệu quả như PhoBERT.

uitnlp/visobert

Epoch cuối (3): Loss: 0.535, Accuracy: 0.6815, Precision: 0.6034, Recall: 0.6796, F1 Score: 0.6393. ViSoBERT có F1 Score cao hơn viBERT nhưng thấp hơn PhoBERT một chút. Tuy nhiên, nó có Recall cao nhất trong số ba mô hình, cho thấy mô hình có khả năng phát hiện lớp mục tiêu tốt hơn nhưng với sự chính xác thấp hơn một chút so với PhoBERT.

7. Kết luận

vinai/phobert-base là mô hình cung cấp sự cân bằng tốt nhất giữa Precision và Recall, dẫn đến F1 Score cao nhất trong ba mô hình được so sánh. Điều này cho thấy nó có thể là lựa chọn tốt nhất cho các ứng dụng cần đến cả hai yếu tố này.

FPTAI/vibert-base-cased không nổi bật so với hai mô hình còn lại về các chỉ số tổng thể.

uitnlp/visobert có F1 Score không cao nhất, nhưng nổi bật với Recall cao, có thể phù hợp với các tình huống mà việc giảm bỏ sót là quan trọng.