ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN

MÔN HỌC: TƯ DUY TÍNH TOÁN

ĐỀ TÀI:

UIT SMART CAMPUS:
PHÂN LOẠI FEEDBACK NGƯỜI HỌC

Giảng viên hướng dẫn: Ngô Đức Thành

Sinh viên thực hiện: Hoàng Xuân Vũ – 19522531

Nguyễn Văn Tài – 19522154

Nguyễn Đạt Huy Hoàng - 19521536

Lớp: CS117.L21

Thành phố Hồ Chí Minh, ngày 10 tháng 7 năm 2021

Mục Lục

1. Giới thiệu .

- Hiện nay, ở trường ĐH Công Nghệ Thông tin sau mỗi kỳ học sẽ có hoạt động khảo sát ý kiến của sinh viên về hoạt động giảng dạy. Những phản hồi của sinh viên là những tài nguyên có giá trị vô cùng to lớn và đóng góp thiết thực vào những cải thiện chất lượng giảng dạy của giảng viên ngày càng hoàn thiện hơn và chất lượng đào tạo ngày càng tốt hơn. Tuy nhiên với số lượng phản hồi nhiều việc đọc hết tất cả những phản hồi rất mất thời gian và khó khăn cho giảng viên. Việc phân loại các phản hồi trong hệ thống sẽ giúp cho việc lựa chọn đọc các phản hồi theo từng mức độ của giảng viên sẽ thuận tiện hơn.
- Vì vậy, mục tiêu của nhóm chúng em trong đồ án này là áp dụng các kĩ thuật của tư duy tính toán để xây dựng một mô hình máy học phân loại các phản hồi của sinh viên (Multi classification) dựa trên tập dữ liệu giảng viên cung cấp và nhóm đã gán nhãn.
- Để xây dựng hệ thống, nhóm chúng em đã thử nghiệm một số mô hình phổ biến trong bài toán phân loại, điển hình như Logistic Regression, SVM (Support Vector Machine), Decision Tree, và cả mô hình kết hợp như RandomForest trên 1 tập dữ liệu. Bên cạnh đó, nhóm cũng tiến hành vận dụng 2 phương pháp tinh chỉnh mô hình Grid Search và Ensemble Learning (mô hình máy học kết hợp) để phát triển mô hình có thể đạt hiệu quả tốt nhất.

2. Xác định yêu cầu bài toán và dữ liệu.

a. Xác định yêu cầu bài toán.

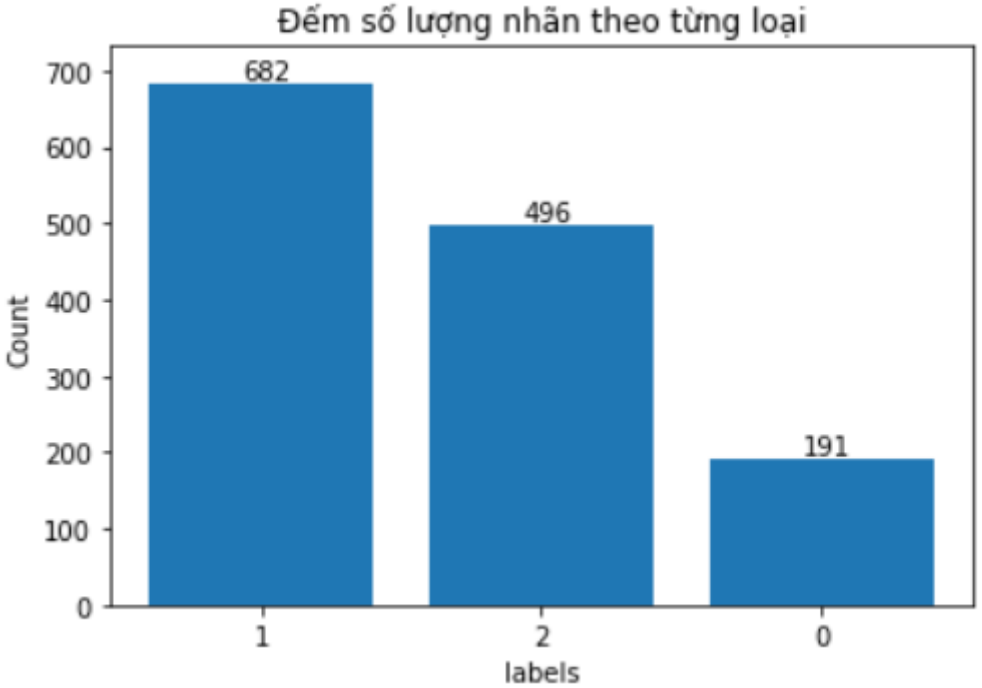
- Input: 1 đoạn feedback của sinh viên sử dụng ngôn ngữ tiếng Việt có dấu có độ dài không quá 256 từ.
- Output: Là 1 trong 3 số nguyên [0,1,2]
- Trong đó:
    - [0] : Tương ứng với Negative là những feedback đánh giá, phàn nàn về giảng viên về cách giảng dạy, cho điểm hay tác phong trong quá trình giảng dạy như: “Giảng viên trình bày không rõ, không hỗ trợ sinh viên”, “không cung cấp đủ giáo trình, giáo viên vào lớp không đúng giờ”,….
    - [1]: Tương ứng với Positive là những feedback đánh giá tốt về giảng viên như: “Giảng viên dạy dễ hiểu”, “truyền đạt tốt, hỗ trợ sinh viên nhiệt tình” , “slide đẹp nhiệt tình giải đáp các thắc mắt của sinh viên”, ….
    - [2]: Tương ứng với Neutral là những feedback đánh giá chung chung không cụ thể hoặc vừa khen vừa chê như: “Cách dạy khó tiếp thu nhưng thầy nhiệt tình”, “không có gì để nhận xét ạ”,….

Ví dụ:

Input	Output
- Thầy dạy rất hay, hỗ trợ sinh viên nhiệt tình, giải đáp các thắc mắc của sinh viên kịp thời.	1

b. Dữ liệu

- Bộ dữ liệu nhóm sử dụng là những phản hồi của sinh viên cho học kỳ I năm học 2020 và đã được nhóm tiến hành gán nhãn.
- Bộ dữ liệu gồm 1369 đánh giá.
- Số lượng nhãn trong bộ dữ liệu gồm 682 nhãn 1, 496 nhãn 2 và 191 nhãn 0.



- Để xây dựng mô hình có tính khách quan, nhóm tiến hành chia bộ dữ liệu thành 2 tập train và test với tỉ lệ 9:1. Nhóm sẽ tiến hành trực quan hóa dữ liệu, làm sạch dữ liệu, thử nghiệm huấn luyện mô hình và chỉnh sửa tham số ở tập train, rồi từ đó sử dụng tập test để tiến hành dự đoán kết quả của mô hình.

3. Graphic Organizer and Project Justification

- Để giải quyết bài toán và xây dựng được mô hình nhóm đã thực hiện phân tích bài toán và tiến hành theo từng bước đã phân tích. (Sử dụng format của Final project của khoá học Problem Solving Using Computational Thinking by University of Michigan trên Coursera).

Project Justification

After you complete the graphic organizer below, use this project justification document to explain how you used computational thinking in your project.

Problem Identification. For each iteration of your problem, please explain how you arrived at your identified problem.

Ở Interation 1, nhóm em đã đưa ra một câu hỏi có hàm ý bao quát khá lớn, thông qua đó đã những vấn đề cần xác định rõ khi giải quyết bài toán này: Input đầu vào là gì? Input có những ràng buộc nào? Output là gì? Căn cứ vào những đâu để chọn ra phương pháp phù hợp nhất để giải quyết bài toán?
Từ Interation 2 trở đi là những vấn đề được mở rộng ra khi tìm hiểu sâu về bài toán và cả phương pháp để giải nó, qua đó chia nhỏ vấn đề thành những vấn đề nhỏ hơn có thể giải quyết được.
Interation này có được là nhờ vào quá trình nhận dạng mẫu chung được xác định ở Interation 1

Decomposition. For each iteration where you decomposed an identified problem, please explain how this decomposition helped you solve your identified problem.

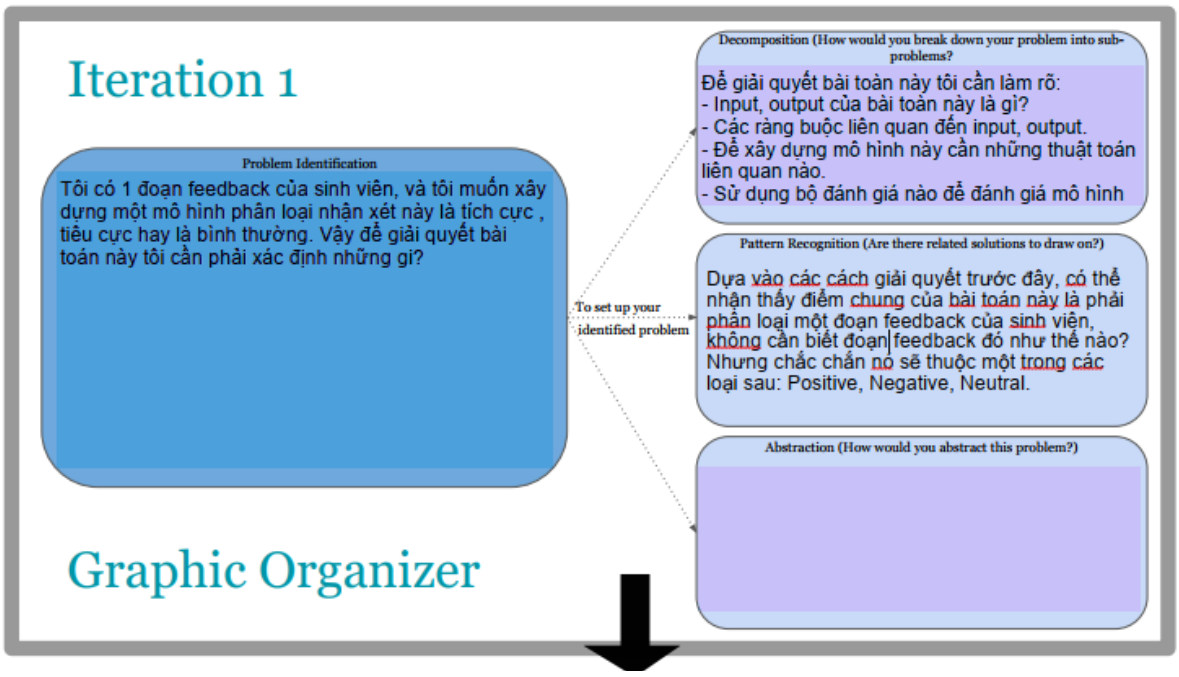
Các giai đoạn Decomposition rất quan trọng trong cả 5 lần Interation. Trong Interation 1 nhóm em decomposition thành 4 vấn đề nhỏ hơn, lần lượt trong Interation 2 là 3, trong Interation 3 là 3, trong Interation 4 là 3 , trong Interation 5 là 2. Việc Decomposion các vấn đề đã làm cho bài toán cụ thể hơn. Việc giải quyết các vấn đề nhỏ sẽ giúp hoàn thiện vấn đề chính và tối ưu các vấn đề nhỏ sẽ hoàn thiện bài toán hơn

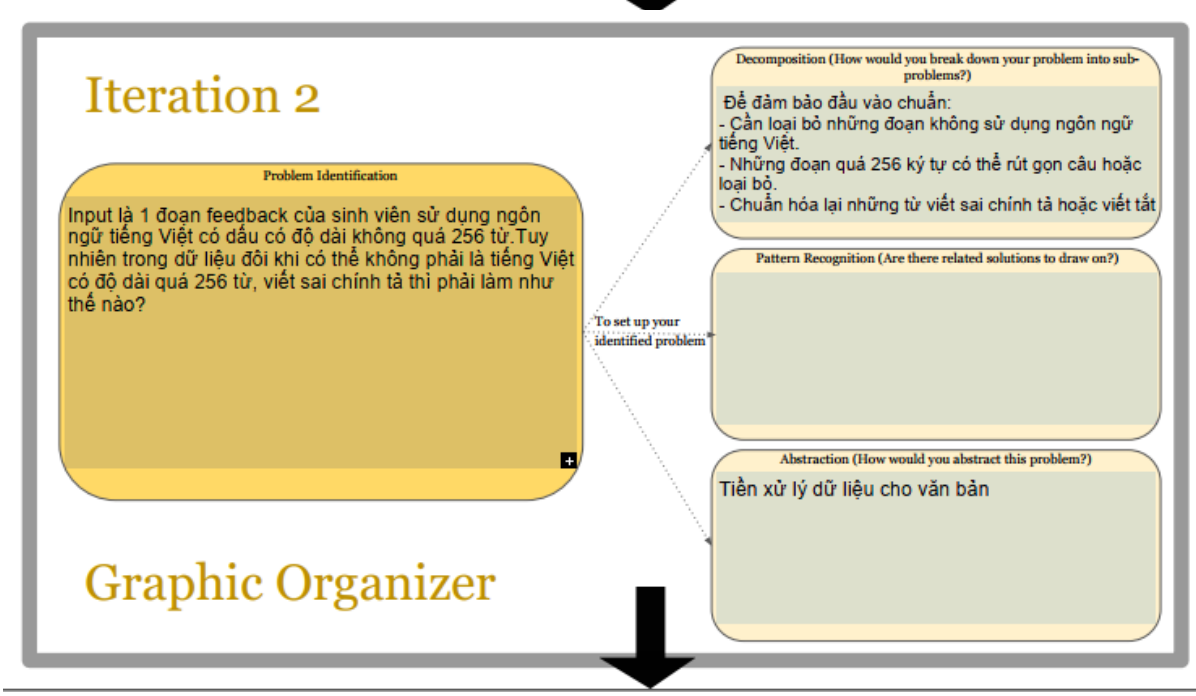
Pattern Recognition. For each iteration where you recognized patterns in data, please explain how these patterns helped you solve your identified problem.

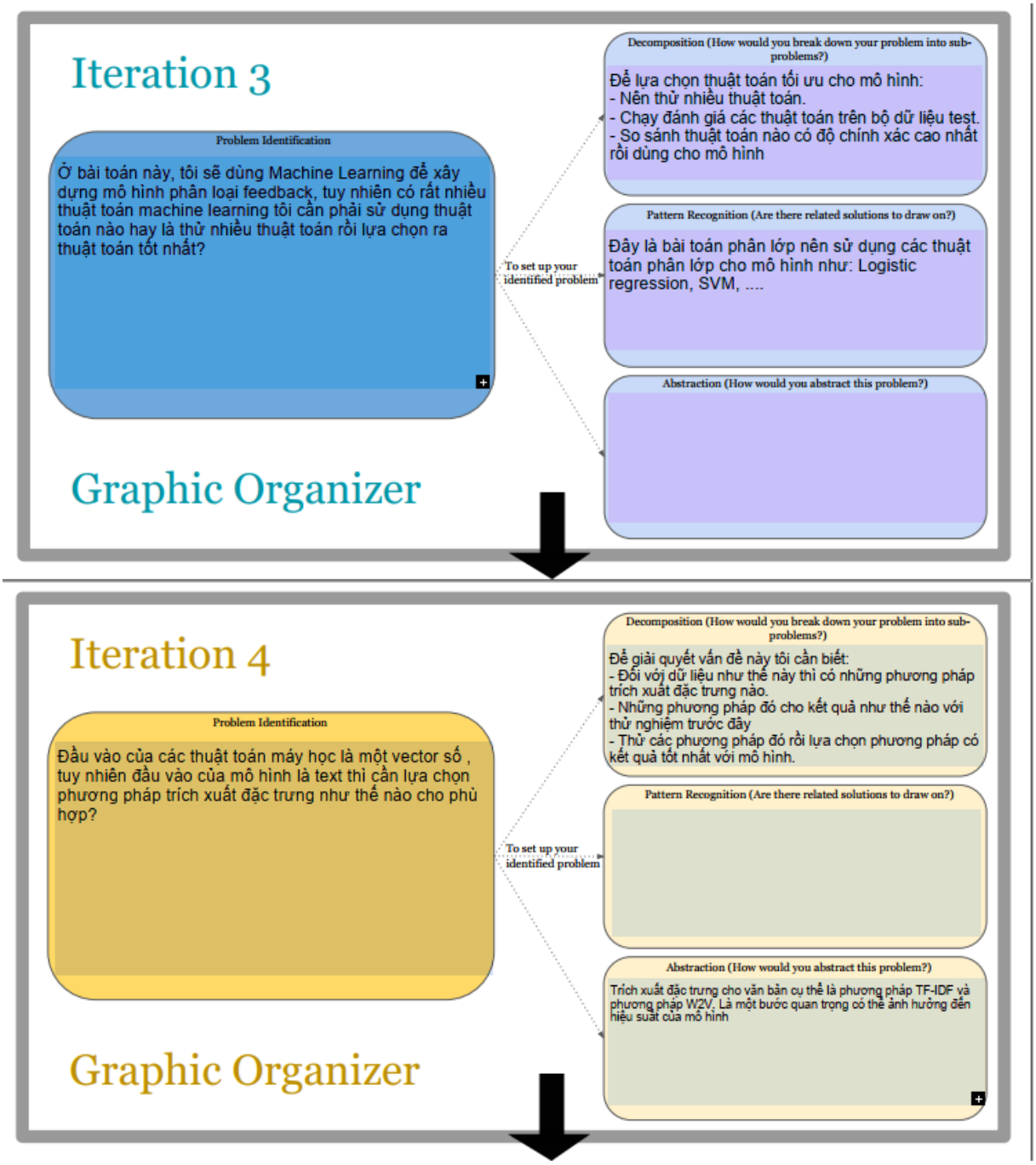
Đối với đồ án này, nhóm em đã sử dụng giai đoạn nhận dạng mẫu ở Interation 1. Đây là giai đoạn quan trọng giúp chúng em làm rõ được vấn đề, xác định được các vấn đề chính cần giải quyết, từ đó phân rã ra thành các vấn đề con nhỏ hơn, sau đó nhóm em chỉ việc giải quyết được các vấn đề con này thì bài toán lớn cũng đồng thời được giải quyết.

Abstraction. For each iteration where you abstracted information, please explain how abstraction allowed you to solve your identified problem.

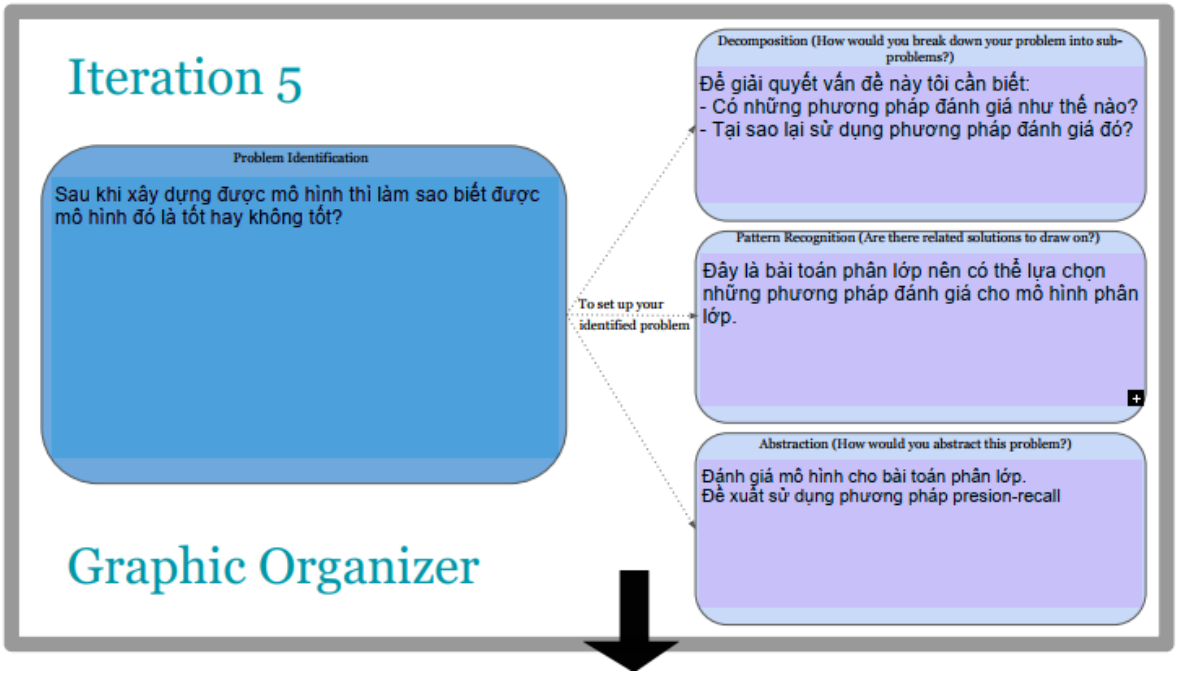
Các giai đoạn Abstraction được sử dụng trong tất cả các lần Interation. Abstraction giúp nhóm làm việc đúng mục tiêu. Cụ thể hóa việc cần làm và phương pháp làm ví dụ như Interation 2, 4,5 nhóm đã nêu cụ thể tập trung vào dữ liệu có liên quan và tập trung vào dữ liệu quan trọng,và bỏ qua các chi tiết không cần thiết để tránh lan man, mất thời gian.


Iteration 1
Problem Identification
Tôi có 1 đoạn feedback của sinh viên, và tôi muốn xây dựng một mô hình phân loại nhận xét này là tích cực , tiêu cực hay là bình thường. Vậy để giải quyết bài toán này tôi cần phải xác định những gì?
To set up your identified problem
Decomposition (How would you break down your problem into sub-problems?
Để giải quyết bài toàn này tôi cần làm rõ:
- Input, output của bài toàn này là gì?
- Các ràng buộc liên quan đến input, output.
- Để xây dựng mô hình này cần những thuật toán liên quan nào.
- Sử dụng bộ đánh giá nào để đánh giá mô hình
Pattern Recognition (Are there related solutions to draw on?)
Dựa vào các cách giải quyết trước đây, có thể nhận thấy điểm chung của bài toán này là phải phân loại một đoạn feedback của sinh viên, không cần biết đoạn feedback đó như thế nào? Nhưng chắc chắn nó sẽ thuộc một trong các loại sau: Positive, Negative, Neutral.
Abstraction (How would you abstract this problem?)
Graphic Organizer


Iteration 2
Problem Identification
Input là 1 đoạn feedback của sinh viên sử dụng ngôn ngữ tiếng Việt có dấu có độ dài không quá 256 từ.Tuy nhiên trong dữ liệu đôi khi có thể không phải là tiếng Việt có độ dài quá 256 từ, viết sai chính tả thì phải làm như thế nào?
To set up your identified problem
Decomposition (How would you break down your problem into sub-problems?)
Để đảm bảo đầu vào chuẩn:
- Cần loại bỏ những đoạn không sử dụng ngôn ngữ tiếng Việt.
- Những đoạn quá 256 ký tự có thể rút gọn câu hoặc loại bỏ.
- Chuẩn hóa lại những từ viết sai chính tả hoặc viết tắt
Pattern Recognition (Are there related solutions to draw on?)
Abstraction (How would you abstract this problem?)
Tiền xử lý dữ liệu cho văn bản
Graphic Organizer

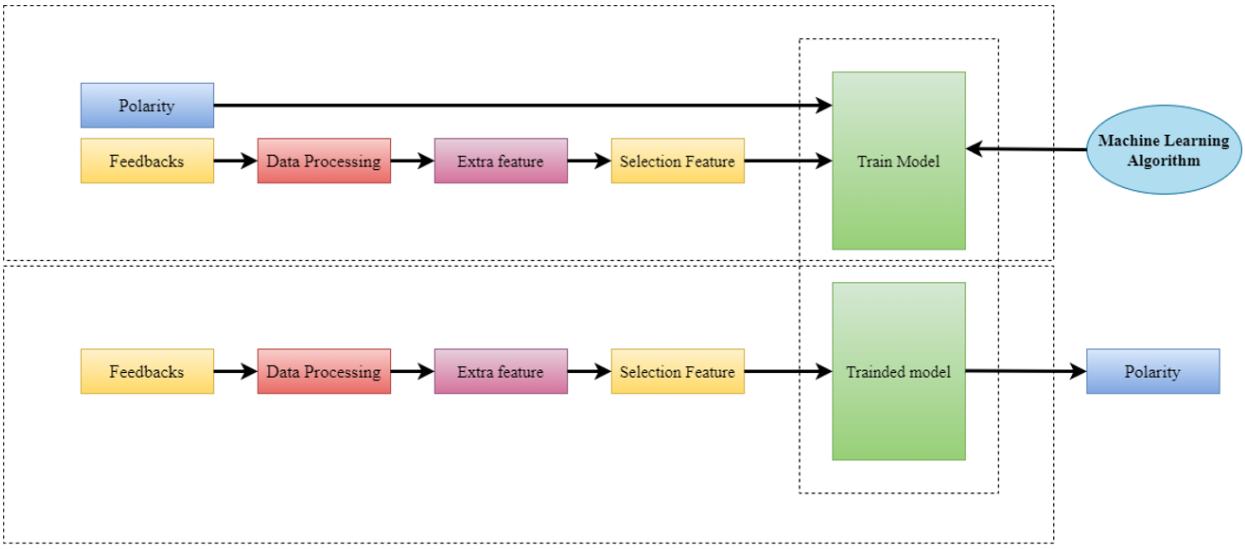

Iteration 3
Problem Identification
Ở bài toán này, tôi sẽ dùng Machine Learning để xây dựng mô hình phân loại feedback, tuy nhiên có rất nhiều thuật toán machine learning tôi cần phải sử dụng thuật toán nào hay là thử nhiều thuật toán rồi lựa chọn ra thuật toán tốt nhất?
To set up your identified problem
Decomposition (How would you break down your problem into sub-problems?)
Để lựa chọn thuật toán tối ưu cho mô hình:
- Nên thử nhiều thuật toán.
- Chạy đánh giá các thuật toán trên bộ dữ liệu test.
- So sánh thuật toán nào có độ chính xác cao nhất rồi dùng cho mô hình
Pattern Recognition (Are there related solutions to draw on?)
Đây là bài toán phân lớp nên sử dụng các thuật toán phân lớp cho mô hình như: Logistic regression, SVM, ...
Abstraction (How would you abstract this problem?)
Graphic Organizer

Iteration 4
Problem Identification
Đầu vào của các thuật toán máy học là một vector số , tuy nhiên đầu vào của mô hình là text thì cần lựa chọn phương pháp trích xuất đặc trưng như thế nào cho phù hợp?
To set up your identified problem
Decomposition (How would you break down your problem into sub-problems?)
Để giải quyết vấn đề này tôi cần biết:
- Đối với dữ liệu như thế này thì có những phương pháp trích xuất đặc trưng nào.
- Những phương pháp đó cho kết quả như thế nào với thử nghiệm trước đây
- Thử các phương pháp đó rồi lựa chọn phương pháp có kết quả tốt nhất với mô hình.
Pattern Recognition (Are there related solutions to draw on?)
Abstraction (How would you abstract this problem?)
Trích xuất đặc trưng cho văn bản cụ thể là phương pháp TF-IDF và phương pháp W2V. Là một bước quan trọng có thể ảnh hưởng đến hiệu suất của mô hình
Graphic Organizer



4. Mô tả thuật toán.

- Sau khi phân tích và đặt ra các câu hỏi cho bài toán nhóm lần lượt tìm lời giải và đáp án để giải quyết vấn đề.
- Nhóm đã đề xuất sử dụng 2 phương pháp trích xuất đặc trưng cho dữ liệu text là: Bag Of Word và TF-IDF.
- Mô hình máy học là : Logistic Regression, Support Vector Machine và mô hình Voting Ensemble kết hợp giữa Logictic Regression và Support Vector Machine.



Hình 1. Mô hình tổng quan

- Tiền xử lý dữ liệu: Tiền xử lý dữ liệu là một trong những thành phần quan trọng trong các mô hình và bài toán liên quan liên quan đến ngôn ngữ. Việc tiền xử lý dữ liệu tốt sẽ giúp mô hình giảm số lượng từ vựng, giảm trọng số huấn luyện và tăng độ chính xác dự đoán cho mô hình. Tiền xử lý dữ liệu gồm các bước như sau: Làm sạch, tách từ trong câu, chuẩn hóa từ, loại bỏ stopwords.
- Trích xuất đặc trưng: Là quá trình chọn ra một số từ từ dữ liệu văn bản sau đó chuyển đổi chúng thành bộ đặc trưng để phân lớp. Trong đề tài sẽ sử dụng một số phương pháp rút trích đặc trưng như: Bag of word, TF-IDF,….
- Lựa chọn đặc trưng: Là việc lựa chọn một tập hợp các đặc trưng đầu vào để đưa ra một tập nhỏ các đặc trưng có ý nghĩa nhất.
- Mô hình máy học: Để xây dựng mô hình phân lớp, đề tài sẽ lựa chọn một số thuật toán máy học cổ điển như Logistic Regression, SVM, Voting Ensemble.

a. Thuật toán Logistic Regression.

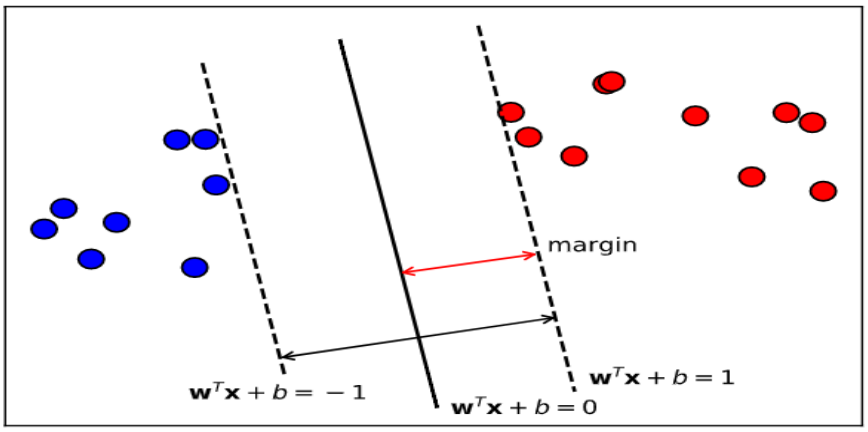
- Thuật toán Logistic Regression là một thuật toán điển hình của bài toán phân lớp.
- Logistic Regression là một loại thuật toán supervised learning tính toán mối quan hệ giữa các feature trong input và output dựa trên hàm logistic/sigmoid. Mặc dù gọi là Logistic Regression nhưng thuật toán này không dự đoán ra giá trị thực như các thuật toán Regression khác, Logistic Regression được dùng để dự đoán ra một kết quả nhị phân (với giá trị 0/1 hay -1/1 hay True/False) dựa vào input của nó. Nhưng Logistic Regression cũng có một chút giống với Linear Regression trong quá trình xây dựng model.

b. **Thuật toán Support Vector Machine**

- Thuật toán Support Vector Machine là một thuật toán thuộc nhóm Supervised Learning (học có giám sát) được nghiên cứu và áp dụng

trong nhiều bài toán khác nhau của lĩnh vực máy học nói chung và lĩnh vực Xử Lý Ngôn Ngữ Tự Nhiên nói riêng. Thuật toán này không chỉ hoạt động tốt cho các bài toán phân loại tuyến tính mà còn đạt hiệu quả với các dữ liệu phi tuyến.

- Ý tưởng:
    - Thu thập dữ liệu.
    - Chọn hai siêu mặt phẳng để phân chia tập dữ liệu mà bên trong đó không chứa điểm dữ liệu nào.
    - Tối ưu khoảng cách giữa chúng.



Hình 2. Ví dụ Margin trong thuật toán SVM

- Giả sử chúng ta xét bài toán phân lớp nhị phân tuyến tính với hai nhãn dữ liệu được mô tả trong một không gian hai chiều như hình 1. Mục tiêu của chúng ta cần xác định mặt phân cách $w^Tx + b = 0$ để phân biệt hai lớp dữ liệu. Khi đó hàm ước lượng :

$$H = x \rightarrow \text{sgn}(w^Tx + b = 0) ; w \epsilon R^N ; b \epsilon R$$

- Với hình 1 chắc chắn sẽ có nhiều mặt phẳng nằm giữa thỏa mãn điều kiện phân tách hai lớp. Mục tiêu của thuật toán là tìm được mặt phẳng tối ưu nhất nằm xa các lớp dữ liệu nhất thì đây là mặt phẳng tốt nhất để phân đều khoảng cách giữa hai lớp dữ liệu.

- Với mặt phần chia như trên thì margin là khoảng cách gần nhất từ một điểm tới mặt phẳng đó (bất kể điểm nào trong hai class).

$$margin = min_n \frac{y_{n(}w^T x_n + b)}{\|W\|_2}$$

c. Thuật toán Voting ensemble

- Đây là thuật toán kết hợp giữa các thuật toán máy học lại với nhau.
- Trong biểu quyết cứng (**hard voting**), mỗi lớp phân loại cá nhân bỏ phiếu cho một lớp và đa số chiến thắng.
- Trong biểu quyết mềm **(soft voting)**, mỗi bộ phân loại riêng lẻ cung cấp một giá trị xác suất mà một điểm dữ liệu cụ thể thuộc về một lớp mục tiêu cụ thể. Các dự đoán được tính theo mức độ quan trọng của trình phân loại và được tổng hợp lại. Sau đó, nhãn mục tiêu có tổng xác suất có trọng số lớn nhất sẽ thắng cuộc bỏ phiếu.

5. Kết quả và demo thực hiện.

- Phương pháp đánh giá:
  - Có nhiều phương pháp đánh giá khác nhau nhưng nhóm đã chọn phương pháp đánh giá Precision Recall. Vì dữ liệu giữa các lớp có sự chênh lệch .
  - Precision được định nghĩa là tỉ lệ số điểm true positive trong số những điểm được phân loại là positive (TP + FP).
  - Recall được định nghĩa là tỉ lệ số điểm true positive trong số những điểm thực sự là positive (TP + FN).
  - Một cách toán học, Precison và Recall là hai phân số có tử số bằng nhau nhưng mẫu số khác nhau:

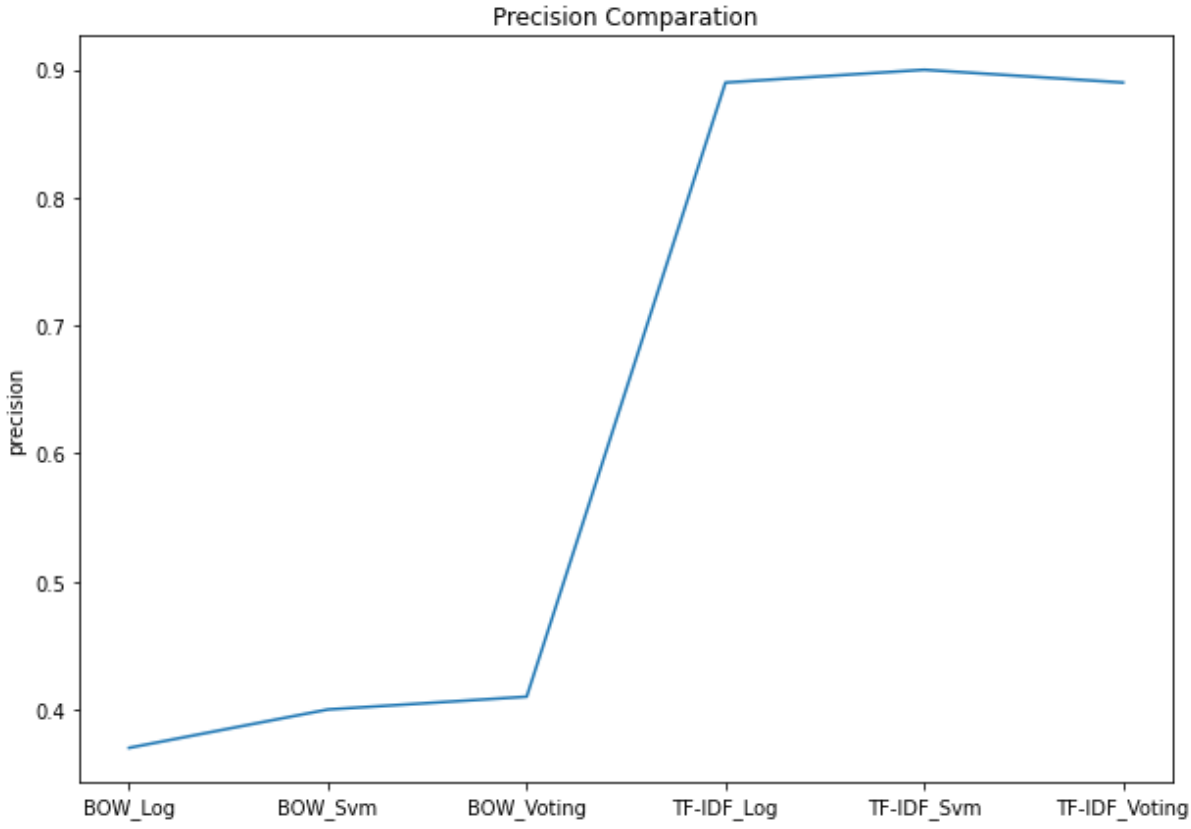
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

  - Tuy nhiên, chỉ có Precision hay chỉ có Recall thì không đánh giá được chất lượng mô hình.

- Chỉ dùng Precision, mô hình chỉ đưa ra dự đoán cho một điểm mà nó chắc chắn nhất. Khi đó Precision = 1, tuy nhiên ta không thể nói là mô hình này tốt.
- Chỉ dùng Recall, nếu mô hình dự đoán tất cả các điểm đều là positive. Khi đó Recall = 1, tuy nhiên ta cũng không thể nói đây là mô hình tốt.
- Khi đó F1-score được sử dụng. F1-score là trung bình điều hòa (harmonic mean) của precision và recall (giả sử hai đại lượng này khác 0).

  - Nhóm sẽ chú trọng vào chỉ số weighted avg F1-score (F1-score là trung bình điều hòa (harmonic mean) của precision và recall)
  - Tại vì số lượng các nhãn trong data không cân bằng nên weighted average sẽ tổng hợp các đóng góp của tất cả các lớp để tính toán số liệu trung bình. Chỉ số này sẽ đánh giá một cách khách quan cho mô hình.

- Sau khi chạy thử nghiệm nhóm đã thu được kết quả:



- Với sự kết hợp giữa phương pháp trích xuất đặc trưng TF-IDF và mô hình SVM đã cho kết quả tốt nhất :

	precision	recall	f1-score	support
class 0	0.79	0.79	0.79	29
class 1	0.92	0.92	0.92	66
class 2	0.93	0.93	0.93	42
accuracy			0.90	137
macro avg	0.88	0.88	0.88	137
weighted avg	0.90	0.90	0.90	137

- Mặc dù mô hình cho kết quả khá là tốt tuy nhiên với dữ liệu khá là ít và độ phong phú của ngôn ngữ tiếng Việt hay nhưng Feedback không rõ ràng, ghi sai chính tả thì mô hình vẫn chưa phân loại chính xác cho lắm. Nếu có thêm dữ liệu hoặc cải thiện bằng các thuật toán học sâu thì mô hình có thể hoạt động hiệu quả hơn.
- Link github trang web Demo:
  https://github.com/19522531/CS117.L21/tree/main/Web_demo
- Link Github Final_Project của nhóm:
  https://github.com/19522531/CS117.L21/

6. Tài liệu tham khảo.

1. **https://www.analyticsvidhya.com/blog/2020/12/understanding-text-classification-in-nlp-with-movie-review-example-example/**
2. **https://www.coursera.org/learn/compthinking/home/welcome**
3. **https://scikit-learn.org/stable/**
4. **https://nguyenvanhieu.vn/tf-idf-la-gi/**
5. **https://machinelearningcoban.com/2017/04/09/smv/**