ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

-oOo-

BÁO CÁO ĐỒ ÁN CUỐI KỲ

PHÂN TÍCH CẢM XÚC TỪ BÌNH LUẬN SẢN PHẨM SHOPEE

Môn học: Tư duy tính toán – CS117.L22.KHCL

Giảng viên lý thuyết: Ngô Đức Thành

Giảng viên thực hành:

Sinh viên thực hiện:

1. Lương Phạm Bảo 19521242
2. Nguyễn Phú Lộc 19520687
3. Đoàn Duy Ân 19521172

TP.Hồ Chí Minh, tháng 08 năm 2021

MỤC LỤC

I. Phân tích cảm xúc cho bình luận sản phẩm

1. Lý do chọn bài toán

Bán hàng online là xu thế công nghệ của ngày nay, gần như mọi gia đình đều sẽ mua ít nhất một món hàng online mỗi tuần. Tuy nhiên khó có thể mà kiểm định được chất lượng có đảm bảo hay không. Đặc biệt với các mặt hàng đắt tiền được bày bán nhan nhản ở khắp mọi nơi trên internet, và việc lựa chọn mua ở đâu, mua hãng gì cho tốt trở thành mối quan tâm lớn cho người dùng.

Một trong những cách để quyết định có nên mua hay không là dựa vào đánh giá từ những người đã mua trước, tuy nhiên số lượng đánh giá rất lớn, không có nhân lực để thống kê được hết. Vì thế áp dụng machine Learning nói riêng cũng như cách giải quyết vấn đề dựa trên máy tính nói chung trong việc phân loại đánh giá của khách hàng là một việc đơn giản và hiệu quả và tiết kiệm chi phí.

2. Xác định bài toán

Input là gì?

• Có 1 input duy nhất.

• Một câu comment có định dạng text về một bình luận về một sản phẩm(dữ liệu ở dưới dạng Tiếng Việt có dấu),1 câu comment có thể chứa nhiều câu .

• Chú ý:câu bình luận có độ dài không quá 200 từ (thông thường các câu có độ dài 200 từ thường là spam) .

Output là gì?

• Có 1 output duy nhất.

• Output có đinh dạng text.

• Giá trị của output gồm một trong hai loại (Positive,Negative) .

• Positive: Câu bình luận mang tính tích cực,đánh giá cao về sản phẩm VD: Shop giao hàng nhanh, đóng gói cẩn thận.

• Negative: Câu bình luận mang tính tiêu cực,chê bai,không hài lòng về sản phẩm VD: Khô gà ăn ỉu, cay nhiều chứ ko phải cay vừa. Ko đc ngon như lần chị m mua.

3. Evaluation

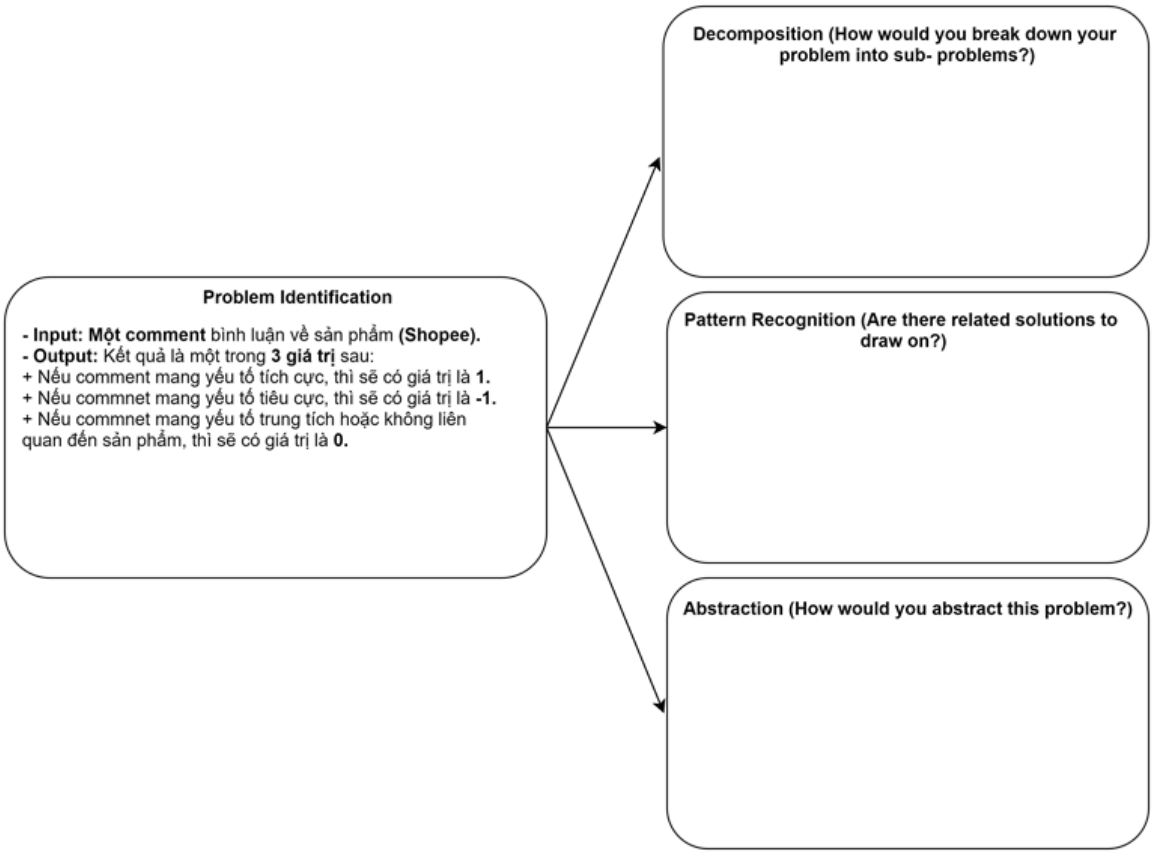
Cách thu thập dữ liệu: Data được thu thập thông qua 2 cách: crawl các comment về sản phẩm từ trang sản phẩm của shopee và lấy các tập dữ

liệu được thu thập từ các cuộc thi của shopee,dữ liệu được crawl sẽ được 2 người đánh nhãn độc lập(nhãn có tỉ lệ đồng thuận trên 70%) .

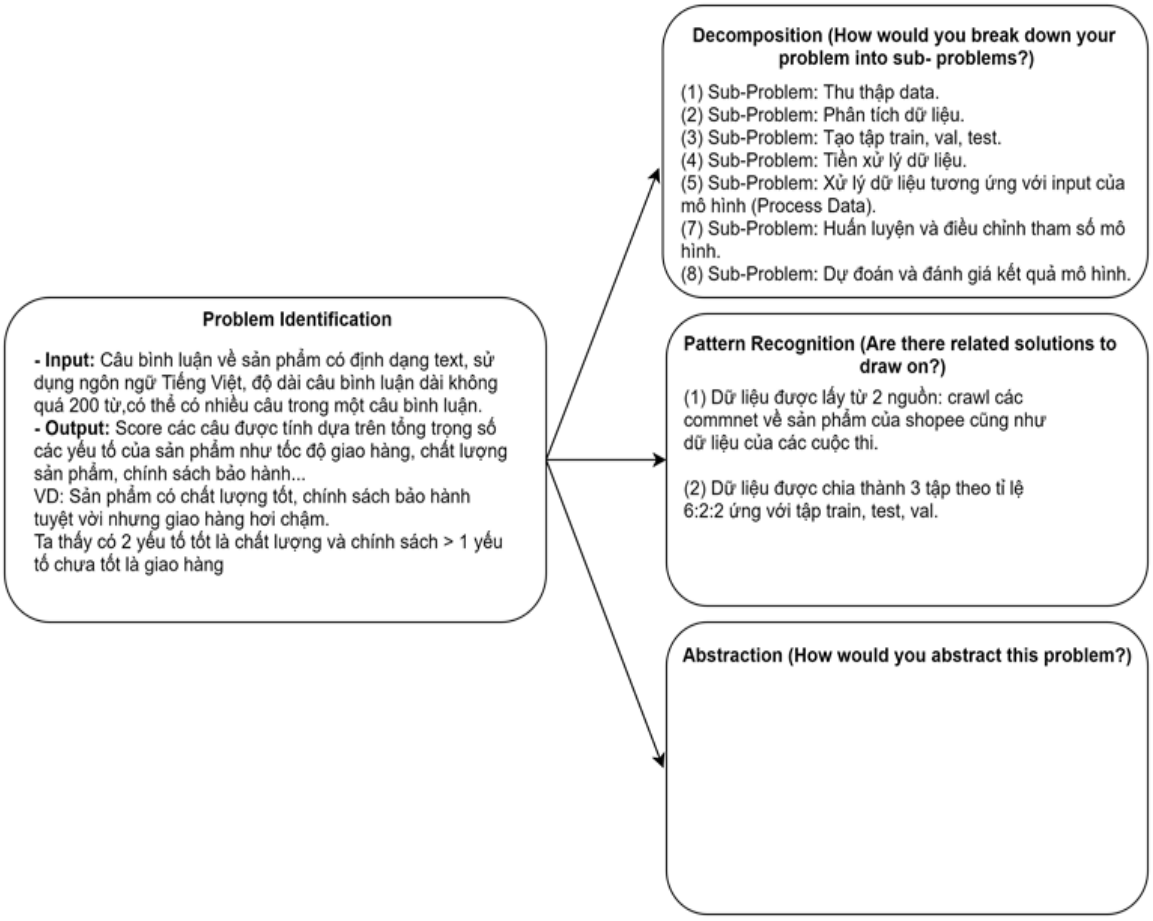
Tiêu chí đánh giá: Được đánh giá bằng kết quả dự đoán đúng trên những câu bình luận khác(các câu chưa có trong bộ dữ liệu, các câu comment mới cho một sản phẩm) ,=s ố lượng câu dự đoán đúng/(trên)số câu muốn dự đoán. Ngoài ra có thể sử dụng một đơn vị đo trong Máy học như F1 score,accuracy Dữ liệu thử nghiệm được tổng hợp lại và gán nhãn sẵn (tỉ lệ đồng thuận trên 85%) nhưng không đưa vào tập huấn luyện (cả 3 bạn đều đánh nhãn độc lập và có thống nhất lại để có tỉ lệ đồng nhất cao) .

II. Graphic Organizer

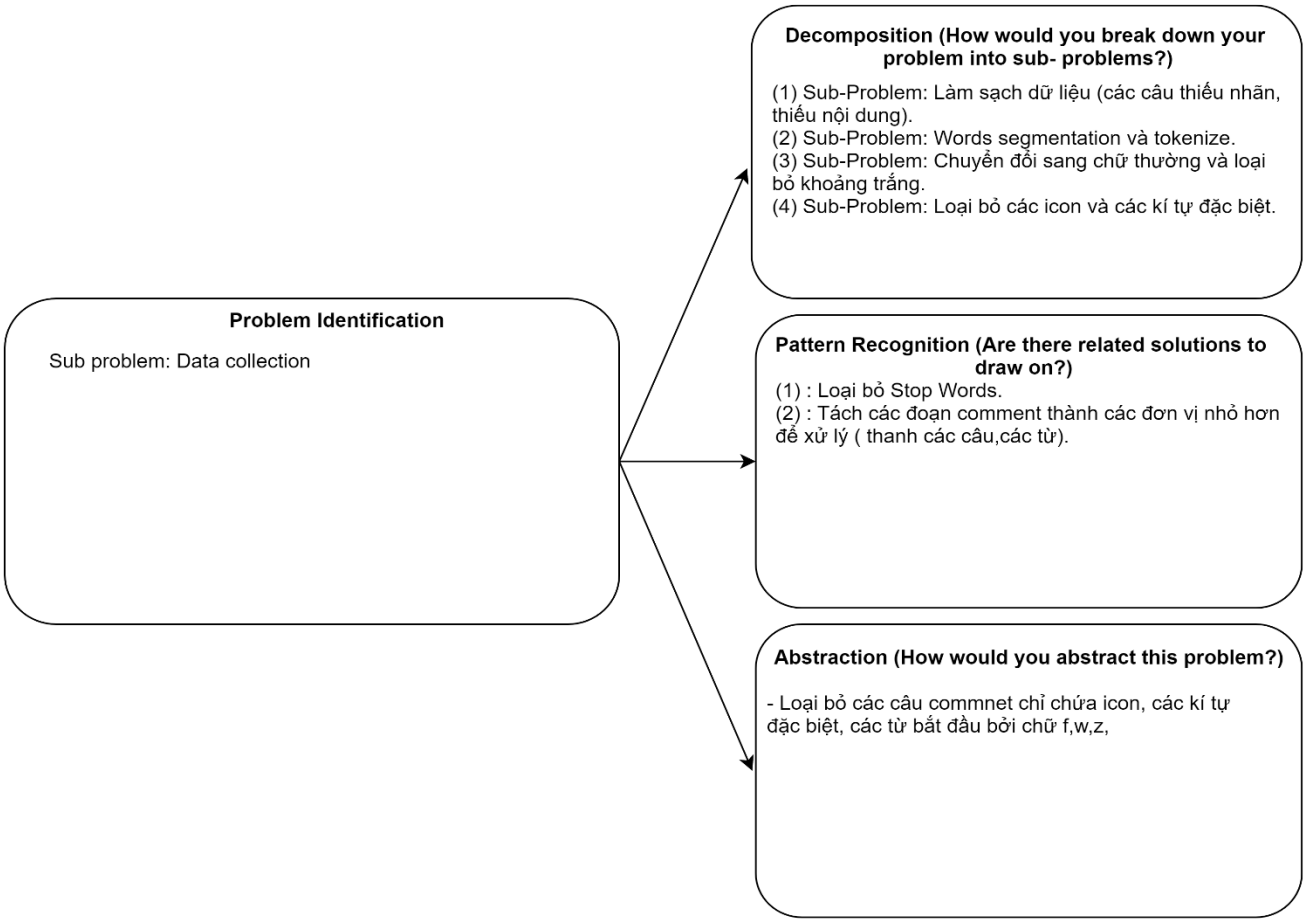
1. Iteration 1



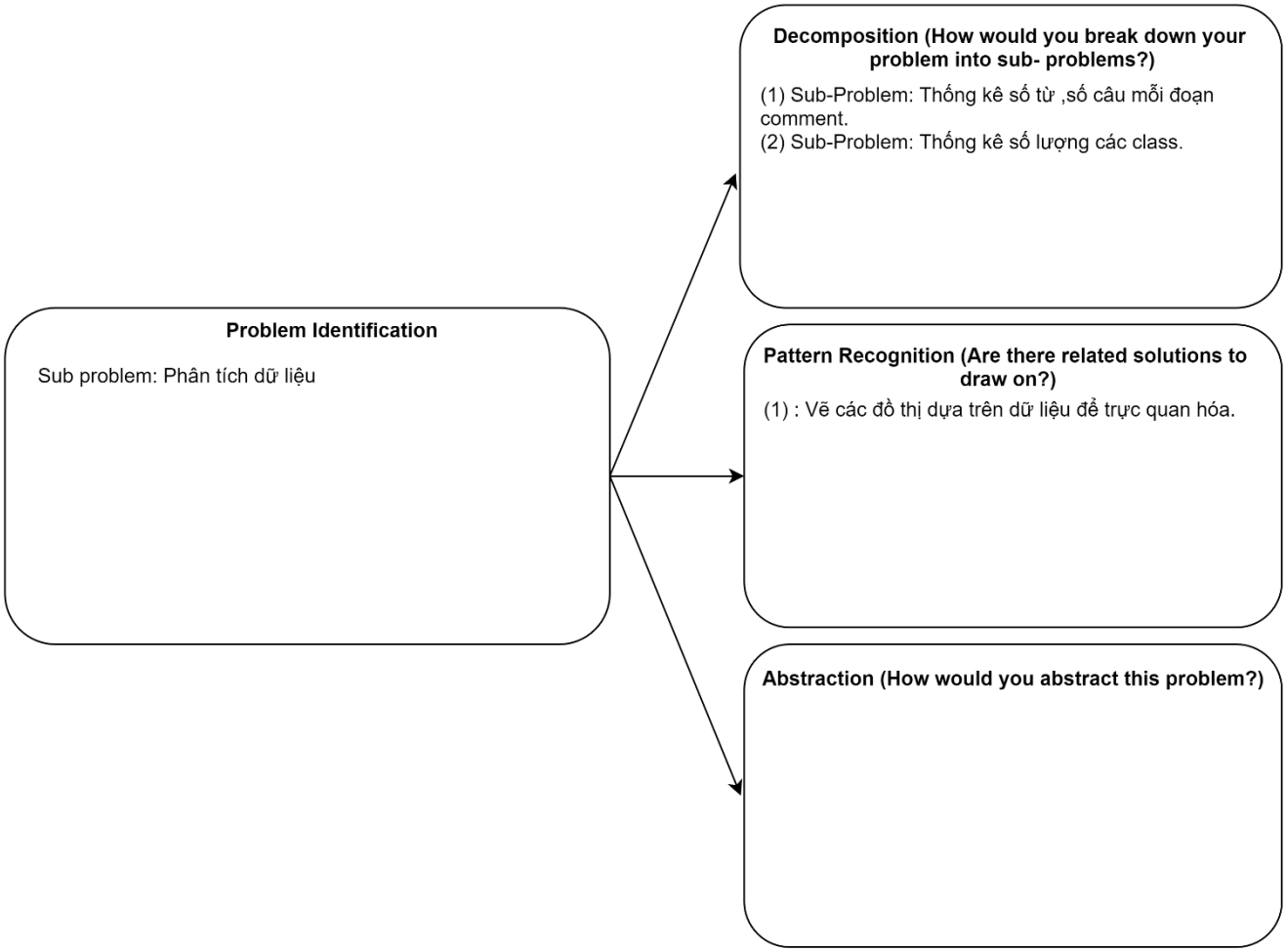
2. Iteration 2



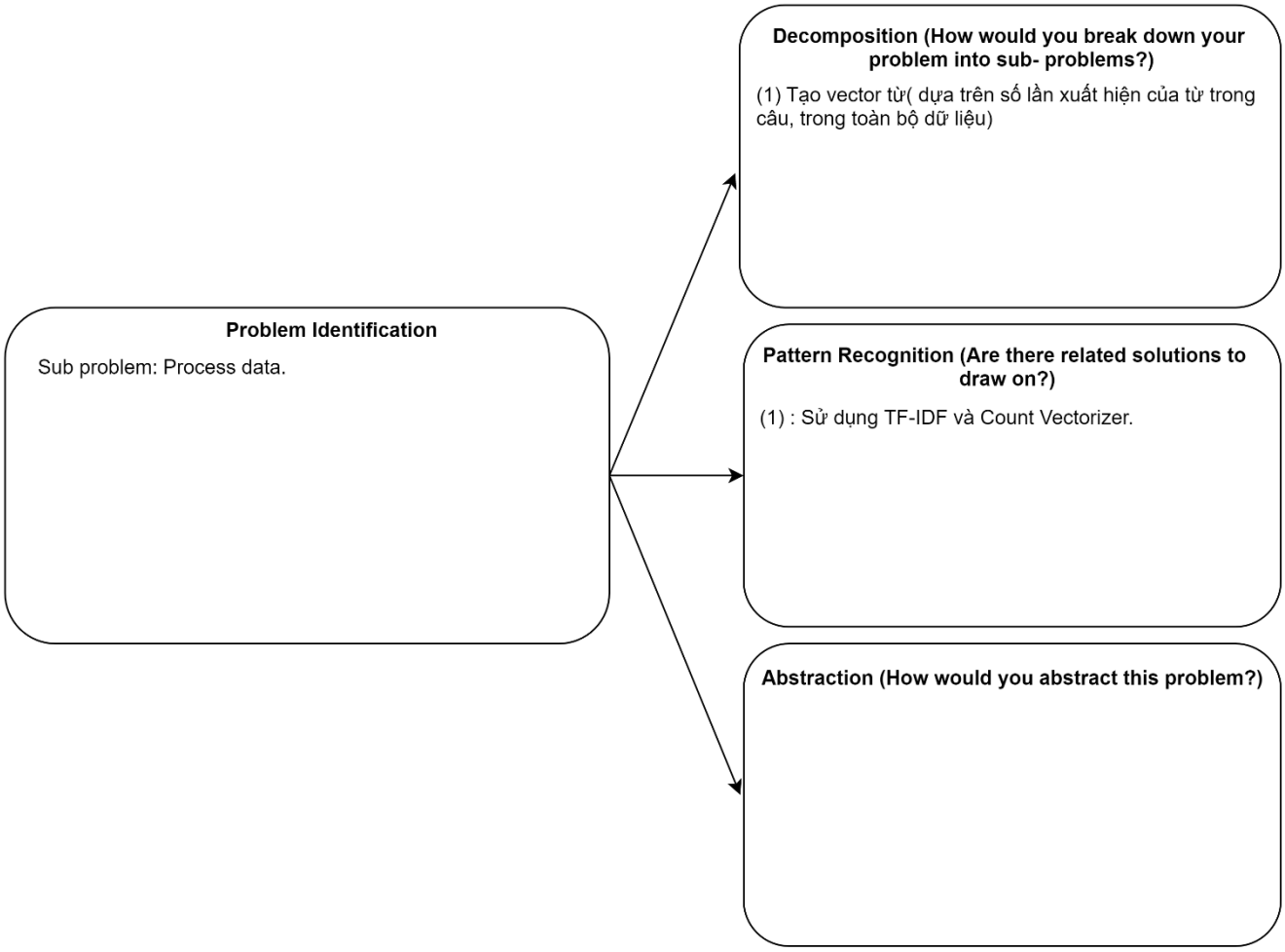
3. Iteration 3



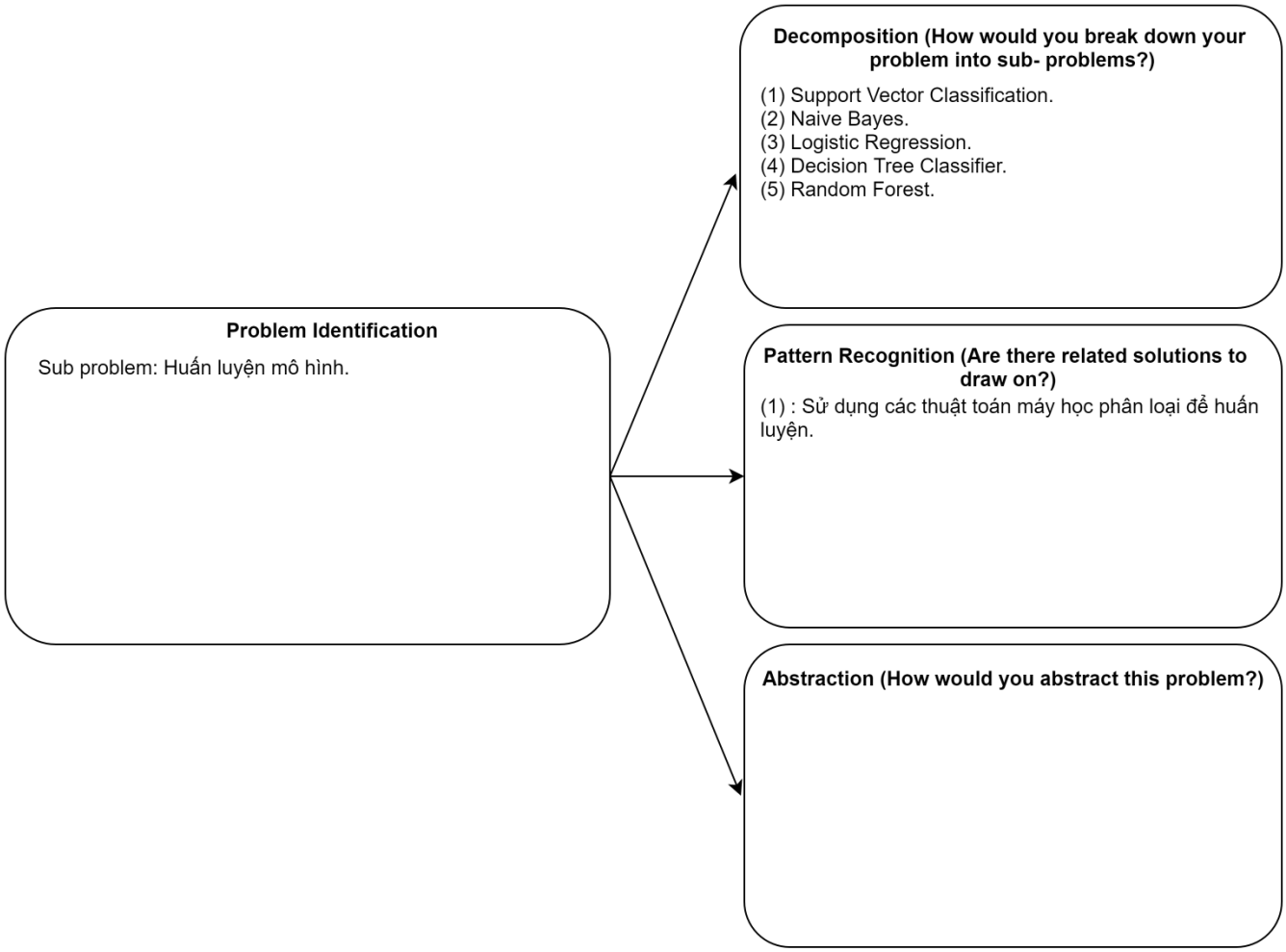
4. Iteration 4



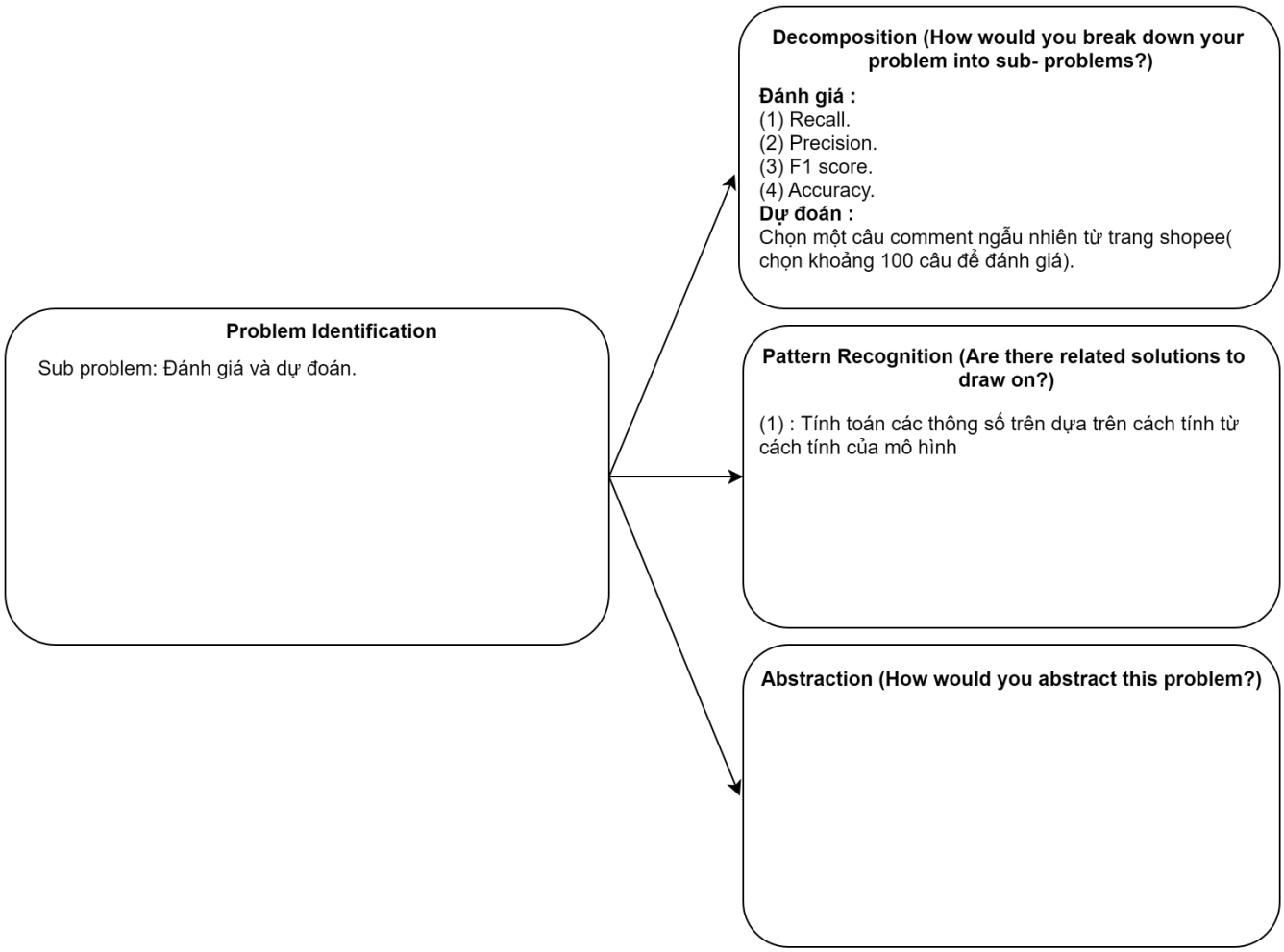
5. Iteration 5



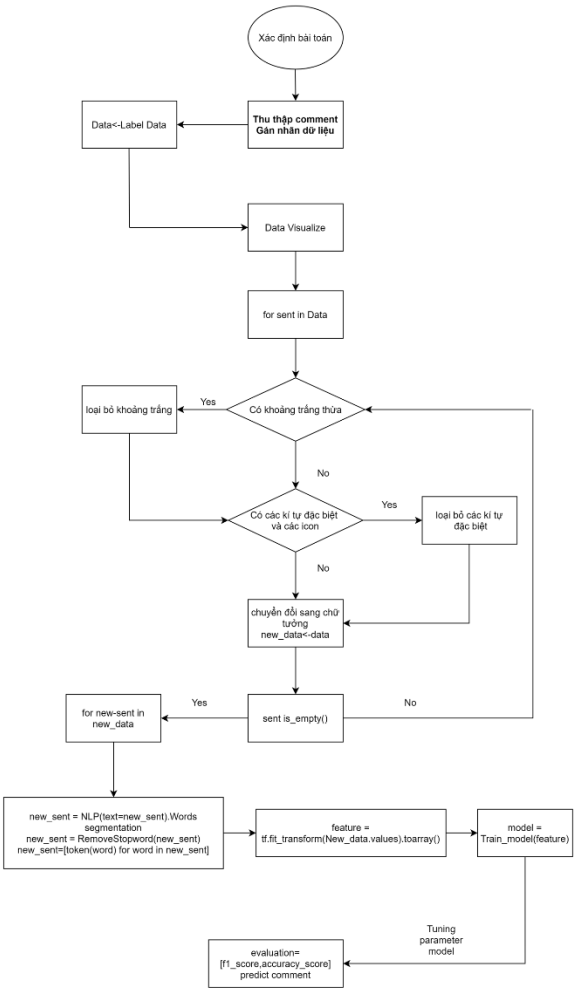
6. Iteration 6



7. Iteration 7



III. FlowChart



IV. Mô tả các kĩ thuật cho bài toán:

1. TF-IDF

ĐN: TF-IDF là trọng số của một từ trong văn bản thu được qua thống kê ,thể hiện mức độ quan trọng của từ này trong một văn bản, mà bản thân văn bản đang xét cũng nằm trong một tập hợp các văn bản.

Tf- term frequency: Dùng để ước lượng tần xuất xuất hiện của từ trong văn bản. Tuy nhiên với mỗi văn bản thì có độ dài khác nhau, vì thế số lần xuất hiện của từ có thể nhiều hơn . Vì vậy số lần xuất hiện của từ sẽ được chia độ dài của văn bản (tổng số từ trong văn bản đó).

TF(t, d) = (số lần từ t xuất hiện trong văn bản d) / (tổng số từ trong văn bản d)

IDF- Inverse Document Frequency: Dùng để ước lượng mức độ quan trọng của từ đó như thế nào. Khi tính tần số xuất hiện tf thì các từ đều được coi là quan trọng như nhau. Tuy nhiên có một số từ thường được được sử dụng nhiều nhưng không quan trọng để thể hiện ý nghĩa của đoạn văn , ví dụ:

Từ nối: và, nhưng, tuy nhiên, vì thế, vì vậy, …

Giới từ: ở, trong, trên, …

Từ chỉ định: ấy, đó, nhỉ, …

Vì vậy ta cần giảm đi mức độ quan trọng của những từ đó bằng cách sử dụng IDF :

IDF(t, D) = log_e(Tổng số văn bản trong tập mẫu D/ Số văn bản có chứa từ t)

Tf-idf có trọng số cao đối với các từ xuất hiện thường xuyên trong tài liệu hiện tại, nhưng hiếm khi xuất hiện trong tập tài liệu tổng thể (document collection), cho thấy rằng từ đó đặc biệt liên quan đến tài liệu này. Đối với mỗi từ xuất hiện trong câu, ta lấy tổng số lượng của từ đó trong tài liệu () nhân với nghịch đảo tần suất của tài liệu đó trên tập tài liệu tổng thể.

2. **Word Segmentation**

Word Segmentation là quá trình phân chia văn bản đã viết thành các đơn vị có nghĩa, chẳng hạn như từ, câu hoặc chủ đề. Thuật ngữ này áp dụng cho cả các quá trình tinh thần được sử dụng bởi con ngườikhi đọc văn bản và các quy trình nhân tạo được thực hiện trong các lĩnh xực xử lý ảnh, xử lý ngôn ngữ tự nhiên.

3. **Stop Word**

Với các bài toán NLP nói chung, ta cần phải rút trích các thông tin thông qua các đặc trưng từ các câu,các từ trong văn bản. Thông thường ta sẽ lấy các từ có số lần xuất hiện nhiều trong văn bản để đánh giá ,tuy nhiên có nhưng câu dù có số lần xuất hiện nhiều nhưng lại không mang lại thông tin nào cụ thể. Vì vậy trước khi tính tần số của các từ trong văn bản, ta cần loại bỏ bớt các từ như thế ra khỏi văn bản nhằm giúp cho kết quả huấn luyện và dự đoán của mô hình trở nên tốt hơn.

VD: **sản phẩm có mẫu mã đẹp, giao hàng nhanh nhưng sản phẩm có bị trầy**

=> sản phẩm mẫu mã đẹp, giao hàng nhanh sản phẩm trầy

V. Mô tả dữ liệu

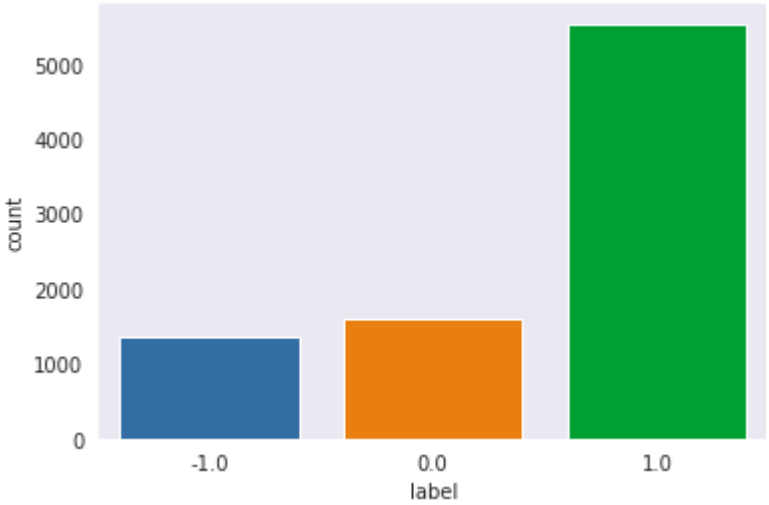
Nhóm chúng em sẽ sử dụng dữ liệu từ 2 nguồn:

craw data: nhóm em sẽ sử web crapping từ thư viện Beatiful Soup của python để craw các comment của shopee ở một số mặt hàng.

data thu thập: lựa chọn một dataset từ cuộc thi của kaggle của shopee.

Dữ liệu comment craw được khoảng 20000 comment. Do khó khăn về mặt thời gian nên chúng em sẽ lựa chọn 6000 câu tốt từ 20000 câu trên để label, sau đó lọc ra một số câu để được khoảng 5000 câu và gộp với dữ liệu với dữ liệu chúng em kiếm được để có được 8497 câu bình luận.

Phân phối nhãn:



1 : 5538

0 : 1598

-1: 1361

Nhóm em sẽ chia dữ liệu thành 2 tập train và test theo tỉ lệ 7: 3 để huấn luyện mô hình.

Ta thấy rằng dữ liệu bị mất cân bằng nên nhóm em sẽ cân nhắc sử dụng thêm F1 score để đánh giá bài toán một cách chính xác hơn.

VI.Kết quả mô hình

1. Kết quả mô hình Random Forest

```
Model RandomForestClassifier
Train score:  0.9880612073314277
Test score:  0.7450980392156863
F1 score:  0.6086471823913261
```

2. Kết quả mô hình Logistic Regression

```
Model LogisticRegression
Train score:  0.8824617454178577
Test score:  0.7964705882352942
F1 score:  0.7047796406957462
```

3. Kết quả mô hình SVM

```
Model SVC
Train score:  0.9043215066420044
Test score:  0.7898039215686274
F1 score:  0.6994389843763932
```

4. Kết quả mô hình Naive Bayes

```
Model MultinomialNB
Train score:  0.7921641163611906
Test score:  0.7533333333333333
F1 score:  0.5910709596263652
```

Như vậy mô hình tốt nhất là mô hình SVM(có sự cân bằng kết quả giữa tập train và test, giữa độ chính xác và F1 score) và có F1 score cao nhất là 69%.

Code demo ở link sau:
https://colab.research.google.com/drive/18ax_K_hr9Mnw-A30ZiGkNMzhJEekvti1?usp=sharing

VII. Hướng phát triển

Trong tương lại nhằm tăng tính ứng dụng thực tế của mô hình, chúng em sẽ mở rộng ra thêm nhiều class hơn để có thể đánh giá mức độ hài lòng,thõa mãn của người dùng một cách tốt hơn, cũng như tăng cường thêm dữ liệu, sử dụng các model deep learning hiện đại hơn để tăng tính hiệu quả của việc huấn luyện cũng như kết quả dự đoán của mô hình.

Có thể xây dựng các app đánh giá bình luận người dùng từ đó thu thập dữ liệu để cải thiện và thõa mãn tiêu chí tiện lợi với các người dùng.

VIII. Nguồn tham khảo

TF-IDF:

https://vi.wikipedia.org/wiki/Tf%E2%80%93idf

Sentiment Analysic:

https://en.wikipedia.org/wiki/Sentiment_analysis

Nhóm bạn Trọng Khánh:

https://github.com/trong-khanh-1109/CS117.L22.KHCL/blob/main/Final-Report.pdf

Computational Thinking:
https://www.coursera.org/learn/compthinking/supplement/4mnoE/introduction-to-human-trafficking-case-study

Code tham khảo:

https://github.com/Long-1234kfgkl/CS114.K21/blob/master/BaoCaoCuoiKy_CS114.K21/Main.ipynb

https://viblo.asia/p/phan-tich-phan-hoi-khach-hang-hieu-qua-voi-machine-learningvietnamese-sentiment-analysis-Eb85opXOK2G