



**VIỆN TRÍ TUỆ NHÂN TẠO**  
**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**



# Phân tích đánh giá lĩnh vực điện thoại di động

## Data Mining & Analysis Project

---

**Nhóm: 6**

**22022602 - Bùi Đức Mạnh**

**22022522 - Đàm Thái Ninh**

**22022666 - Lê Việt Hùng**

**22022569 - Trần Nam Anh**

# Mục lục



**01**

**Bài toán**

**02**

**Module**

**03**

**Dữ liệu**

**04**

**Phương  
pháp**

**05**

**Kết quả**

**06**

**Phân công**

**07**

**Hạn chế**

**08**

**Demo**

**09**

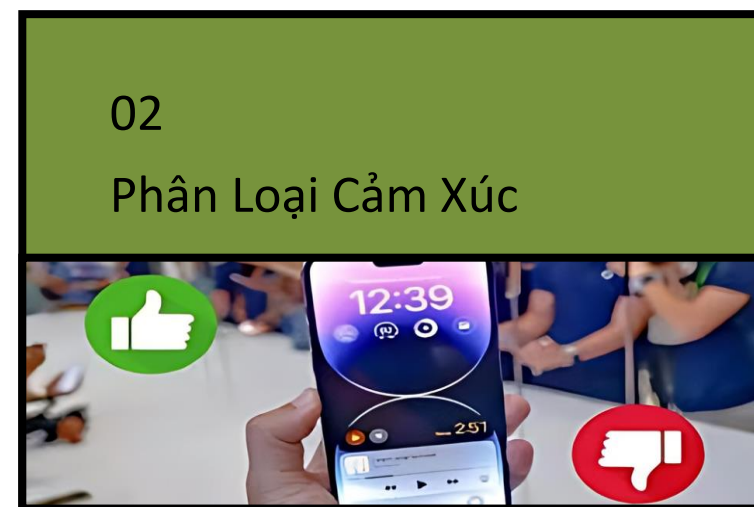
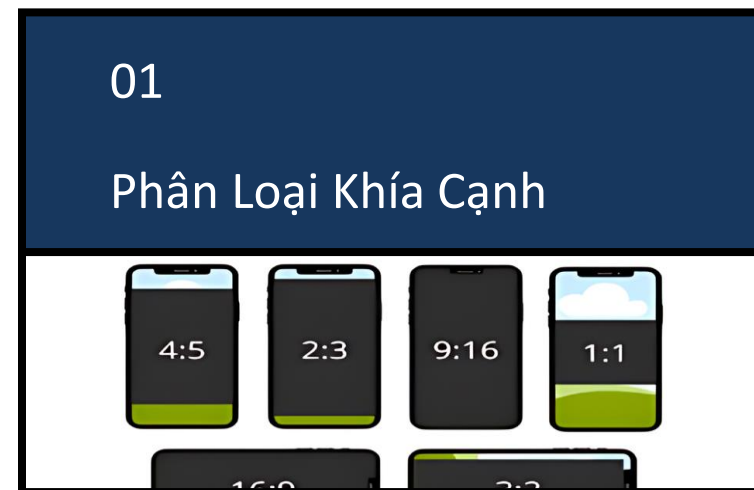
**Future  
work**

**10**

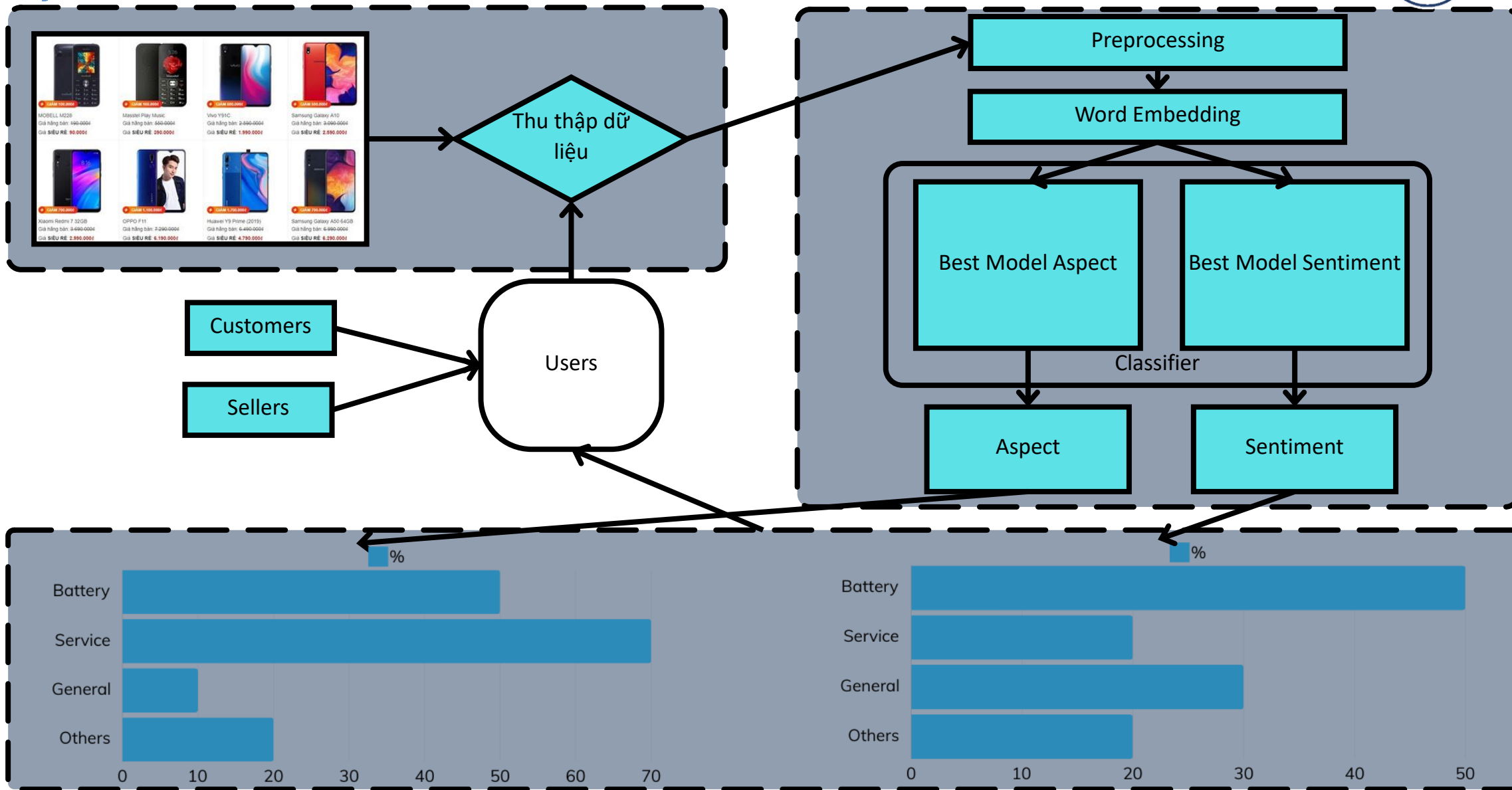
**Tổng kết**

# 01 | Bài toán:

- **Bài toán:** Aspect-based sentiment analysis (ABSA)
- **Lĩnh vực:** Điện thoại thông minh trên các trang web thương mại điện tử
- **Mục tiêu:** Giải quyết được 2 nhiệm vụ trong Phân tích cảm xúc dựa trên khía cạnh: **Phát hiện danh mục khía cạnh (Aspect Category Detection)** và **Phân loại phân cực cảm xúc (Sentiment Polarity Classification)**
  - **Phân loại được khía cạnh (aspect):** có/không
  - **Phân loại được cảm xúc (sentiment):** negative, neutral, positive



## 02 | Module:



## 03 | Dữ liệu:


- Thu thập đánh giá bằng văn bản từ khách hàng về điện thoại thông minh trên các trang thương mại điện tử lớn tại Việt Nam.

- Crawl bằng Selenium

Thảm  Đã mua tại TGDĐ



Dùng đc 2,5 tháng pin bị tụt 1%, mỗi lần sạc rất nóng dù sạc zin. Máy chị mình cũng ip 11 lại không bị gì

 Hữu ích | Đã dùng khoảng 2 tháng

```
with open('comments.csv', 'a', newline='', encoding='utf-8-sig') as csvfile:
    writer = csv.writer(csvfile)
    driver = webdriver.Chrome() # Open Chrome webdriver
    driver.get("https://www.thegioididong.com/dtdd") # Open the web
    sleep(3)

    # Click on the product
    driver.find_element(By.XPATH, "//*[@id='categoryPage']/div[3]/ul/li["+ str(index_of_products) +"]").click()
    sleep(3)

    try:
        # Click to see others comments
        driver.find_element(By.XPATH, "//a[@class='c-btn-rate btn-view-all']").click()
        sleep(3)
    except:
        check = False
        pass

    try:
        # Crawl comments and its star
        comments = driver.find_element(By.XPATH, "//div[@class='rt-list']").find_elements(By.CLASS_NAME, "cmt-txt")
        comments_stars = driver.find_element(By.XPATH, "//div[@class='rt-list']").find_elements(By.CLASS_NAME, "cmt-top-star")
        for i in range(len(comments)):
            txt = comments[i].text
            n_stars = len(comments_stars[i].find_elements(By.CLASS_NAME, "iconcmt-starbuy"))
            if (txt != "" or txt.strip() or (len(txt) >= 75 and len(txt) <= 300)):
                writer.writerow([txt, n_stars, ""])
    except:
        check = False
        pass
```

## 03 | Dữ liệu:



- Thu thập đánh giá bằng văn bản từ khách hàng về điện thoại thông minh trên các trang thương mại điện tử lớn tại Việt Nam.

Thẩm Đã mua tại TGDĐ

★★★★★

Dùng đc 2,5 tháng pin bị tụt 1%, mỗi lần sạc rất nóng dù sạc zin. Máy chị mình cũng ip 11 lại không bị gì

Hữu ích

Đã dùng khoảng 2 tháng

- Crawl bằng Selenium
- Tổng cộng: 1600 samples

Dataset	Reviews	Aspects	AvgLength	VocabSize
Train	1280	2363	81.592969	7095
Val	320	589	85.759375	3025

- Test data: UIT-ViSFD: A Vietnamese Smartphone Feedback Dataset for Aspect-Based Sentiment Analysis

## 03 | Dữ liệu:

- Đếm common words, chọn ra 7 khía cạnh có tỉ lệ xuất hiện nhiều
- Label 7 aspects với 500 samples
- Chọn ra 4 thuộc tính xuất hiện nhiều nhất, còn lại gộp vào Others
- Tiếp tục label phần sentiment

Comments	n_stars	len_cmt	Pin	Performance	Camera	Màn hình	Ngoại hình	Service	Others
Cuộc tình du dương này sai thì xem p	5	129	0	0	0	0	0	0	1
Thời gian trước hay mua quán lẻ ở i	5	126	0	0	0	0	0	1	0
Sản phẩm dùng sau gần 3 tháng kh	5	125	1	1	1	0	0	0	0
Máy bị lỗi màn hình hên là còn bảo l	2	187	0	0	0	0	0	1	1
Úi qua vừa đi mua xong phải nói là c	5	124	0	0	0	0	1	1	1
Cái đầu tiên là màu gấc xink nha, đư	5	130	0	0	0	0	1	0	1
Mình xài được gần 4 tháng nhưng tì	3	154	1	0	0	0	0	0	0
Mình mua xài được 2 thang mình xài	2	159	0	0	0	0	1	0	0
Rõ lần đầu chuyển sang iPhone mà	2	129	0	0	0	0	1	0	0
Máy dùng ổn, mượt mà, pin khá lâu,	5	198	1	1	0	0	0	1	0
			208	141	57	64	71	157	174
			0,4126984127	0,2797619048	0,1130952381	0,126984127	0,1408730159	0,3115079365	0,3452380952

## 03 | Dữ liệu:



### Data sau khi gán nhãn

Comments	n_stars	len_cmt	Pin	Service	General	Others	SGeneral	SPin	SSer	SOth
E mới mua iphone 11. Và trong máy chỉ có 1 dây sạc chứ ko có củ sạc	5	203	0	1	0	0	2	2.0	-1.0	2.0
Mới mua cách đây gần một tháng. Mọi thứ đều rất tốt nhưng pin em xê	5	108	1	0	1	0	1	-1.0	2.0	2.0
Máy em mua ngày 12/10 đến nay ngày 3/11 chưa đầy 1 tháng mà pin c	4	109	1	0	0	0	2	-1.0	2.0	2.0
Mình mới xài được 7 tháng xuống 7% pin. Chả hiểu máy mới kiểu gì n	1	110	1	0	0	0	2	-1.0	2.0	2.0
Đã mua, rất mượt, pin rất trâu, học online liên tục cả buổi chiều hết ~3:	5	237	1	0	0	1	2	1.0	2.0	-1.0
Máy rung khá êm nên nhiều lúc có cuộc gọi không biết! Vì lí do công vi	3	126	0	0	0	1	2	2.0	2.0	-1.0
Về mới thấy màn hình thì hơi vàng tí, chụp hình thì camera bị nhoè nh	2	83	0	0	0	1	2	2.0	2.0	-1.0
Vừa mua đôi với bố khá ok đẹp ,chụp hình ở mức ổn thôi , chơi game	5	256	0	1	0	1	2	2.0	0.0	1.0
Mình mới mua đt được 3 ngày mọi thứ đều ổn, nhưng gọi video bên z	4	127	0	0	1	1	1	2.0	2.0	0.0
Sử dụng mới 5 tháng chưa hết bảo hành bị tróc sơn khung viền lấy tay	2	243	0	1	0	1	2	2.0	-1.0	-1.0
em mới vừa mua hôm quá, máy xài rất ok nhưng sao lúc sạc pin thì m	4	88	0	0	1	1	1	2.0	2.0	-1.0
Mới mua máy đc 2 tháng máy tụt 99%, dù dùng rất cẩn thận, theo kiểu	3	245	1	0	0	0	2	-1.0	2.0	2.0
Mình mua máy đợt vừa rồi có cài cảm ứng chạm quả táo sau máy mà	3	236	0	0	0	1	2	2.0	2.0	-1.0
Sao lúc bật mạng 4g mà nó nhấp nháy mấy cái chỗ bật mạng vậy.ngày	3	168	0	0	0	1	2	2.0	2.0	-1.0
Máy mới mua gần 3 tháng khi call bị tè như tiếng rào rào ề ề trong khi	3	142	0	0	0	1	2	2.0	2.0	-1.0
Nói chung máy tốt pin xài trong ngày ok có đều lâu lâu máy bị đơ phải	4	87	1	0	1	1	1	1.0	2.0	-1.0
Mình mua ip11 xong được tặng 1 cái thẻ viettel mà nạp lại báo là thuê	5	110	0	0	0	1	2	2.0	2.0	0.0
May rất ok. Nhưng hài lòng nhất là nhân viên của dmx.rất nhiệt tình vớ	5	81	0	1	1	0	1	2.0	1.0	2.0
mới mua được 3 tháng mà tình trạng bin tụt còn 96%. mặc dù không v	4	81	1	0	0	0	2	-1.0	2.0	2.0
Mình mua máy đổi trả ở tgdd mà lúc đi cầm người ta báo là máy đã bị	3	108	0	1	0	0	2	2.0	-1.0	2.0



### Các loại nhãn

1. **Battery (Pin):** liên quan đến dung lượng, chất lượng, tốc độ sạc:
  - 0: không có aspect này
  - 1: aspect này được đề cập trong phản hồi của khách hàng
2. **Service (Dịch vụ):** liên quan đến dịch vụ bán hàng, giá cả, bảo hành, khuyến mãi, phụ kiện kèm theo:
  - 0: không có aspect này
  - 1: aspect này được đề cập trong phản hồi của khách hàng
3. **General:** Những đánh giá chỉ nói chung chung mà không nói cụ thể về thành phần nào:
  - 0: không có aspect này
  - 1: aspect này được đề cập trong phản hồi của khách hàng
4. **Others (Khác):** liên quan đến các vấn đề không được đề cập ở trên, lỗi phần cứng/phần mềm, bảo mật, nội dung liên quan đến Performance, Screen, Camera, Design, chỉ có từ cảm thán:
  - 0: không có aspect này
  - 1: aspect này được đề cập trong phản hồi của khách hàng
5. **SGeneral:** Sentiment liên quan đến General:
  - **Positive (1):** tốt, hài lòng
  - **Negative (-1):** tệ, chán, kinh khủng
  - **Neutral (0):** bình thường, không có gì nổi bật

### Annotation guildlines

### 03 | Dữ liệu:



Aspect	Fleiss Kappa
Pin	0,9333333333
Service	0,9006752562
General	0,9092592593
Others	0,8674348104

Sentiment	Fleiss Kappa
SPin	0,9331941545
SSer	0,9253297776
SGen	0,8674348104
SOth	0,9006752432

Annotation Agreement

### 03 | Dữ liệu:

**Training**

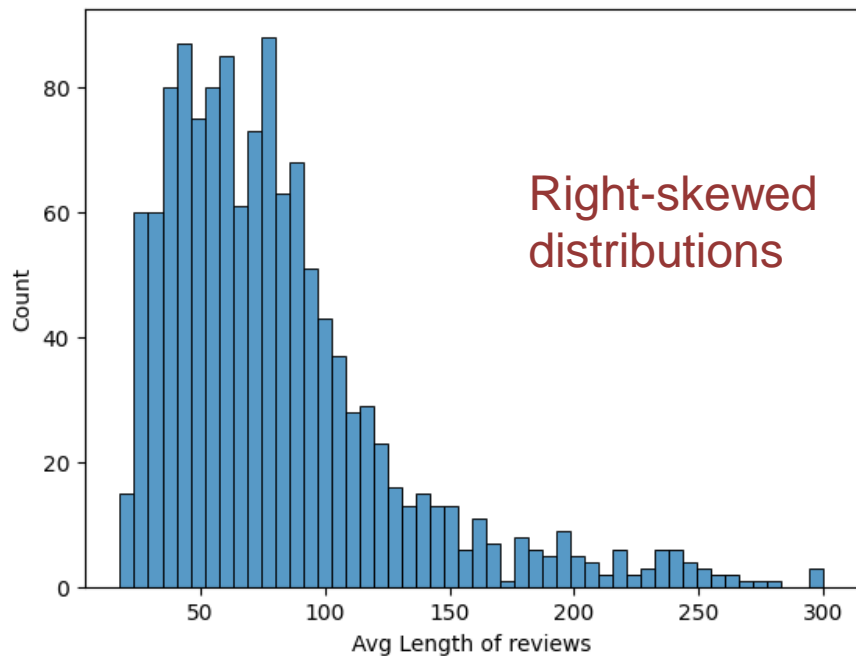
1280 samples

**Valid**

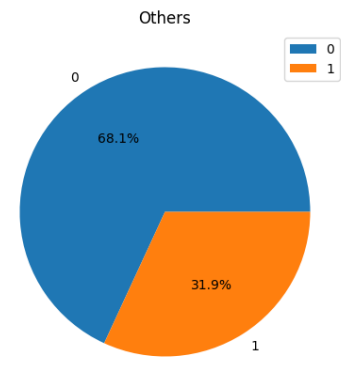
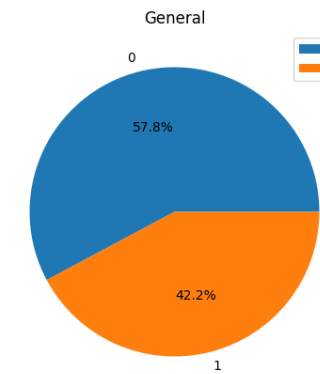
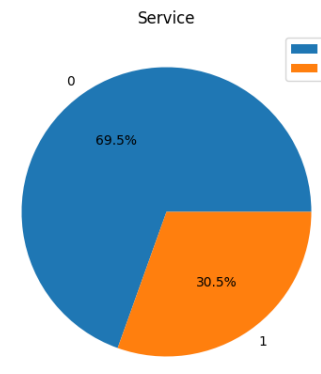
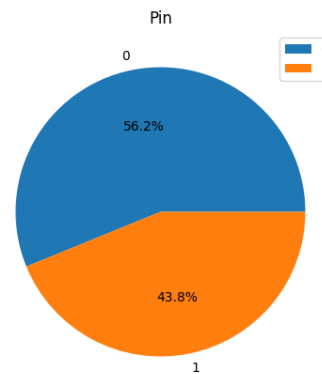
320 samples

**Test(UIT-ViSFD)**

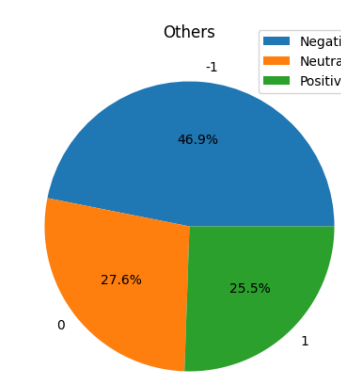
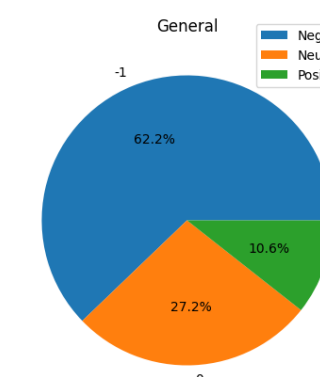
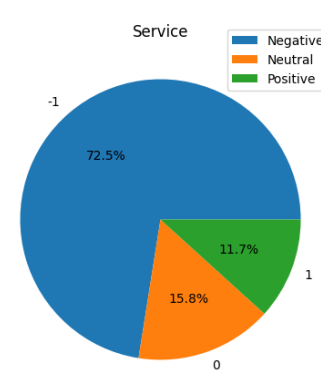
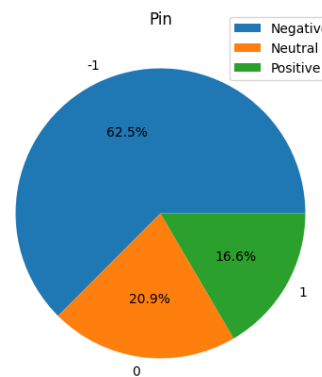
2224 samples



Phân phối của AvgLength



Phân phối của nhãn aspect



Phân phối của nhãn sentiment

## 03 | Dữ liệu:



train

	Comments	n_stars	len_cmt	Pin	Service	General	Others	SGeneral	SPin	SSer	SOth
0	Cho em hỏi sạc em dùng củ sạc chính hãng khi c...	3	88	1	0	0	1	0	0.0	0.0	0.0
1	Mình mua được khoảng 2 tháng và thấy cảm ứng n...	3	73	1	1	1	1	1	1.0	1.0	1.0
2	sp thiết kế nhỏ gọn, mỏng , đẹp màu đen nhám s...	2	41	0	1	0	1	0	0.0	1.0	0.0
3	Mới mua được 2 ngày\...\nDịch vụ tốt, tư vấn nhi...	4	91	1	1	1	1	1	1.0	1.0	1.0
4	Mình vừa test máy mới, sao máy này game liên q...	4	82	0	0	1	1	0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...
1275	Máy lỗi nghe nhạc bluetooth . Ra bảo hành thì ...	1	77	0	1	0	1	0	0.0	-1.0	-1.0
1276	Máy sài rất êm chụp hình rất đẹp không chê vào...	5	75	0	0	1	1	1	0.0	0.0	1.0
1277	Điện thoại đẹp, pin dùng rất khỏe, camera tốt....	5	82	1	0	1	1	1	1.0	0.0	1.0
1278	Tựa là mới phát hiện cái loa bị rè ..ko biết l...	1	45	0	0	0	1	0	0.0	0.0	1.0
1279	Mới mua 2 ngày cảm thấy khá hài lòng nhưng pin...	4	44	1	0	1	0	1	-1.0	0.0	0.0

1280 rows × 11 columns

## 04 | Phương pháp:



- Tiền xử lí:

```
# overall preprocessing
def text_preprocess(document, pos_list, nag_list, not_list):
    #đưa về lower
    document = document.lower()
    # xóa html code
    document = remove_html(document)
    # chuẩn hóa unicode
    document = convert_unicode(document)

    # chuẩn hóa các ký tự đặc biệt
    document = normalize_money(document)
    document = normalize_hastag(document)
    document = normalize_website(document)
    document = nomalize_emoji(document)
    document = normalize_elongate(document)
    document = normalize_acronyms(document)
    document = remove_numbers(document)

    # chuẩn hóa cách gõ dấu tiếng việt
    document = standardize_sentence_typing(document)
    # tách từ
    document = word_tokenize(document, format="text")
    # đưa về lower
    document = document.lower()
    # xóa các ký tự không cần thiết
    document = remove_unnecessary(document)
    # xử lý vấn đề phủ định
    document = add_sentiment_features(document, pos_list, nag_list, not_list)
    return document.translate(str.maketrans(string.punctuation, ' ' * len(string.punctuation))).replace(' '*4, ' ').replace(' '*3, ' ').replace(' '*2, ' ').strip()
```

## 04 | Phương pháp:



- Tokenizer: Mean Embedding kết hợp PhoW2V: Pre-trained word embeddings for Vietnamese

$$MEV = \frac{1}{n} \sum_{i=1}^n vec(word_i)$$

- SVM:
  - 8 mô hình riêng lẻ, mỗi mô hình dự đoán 1 aspect/sentiment
  - Input: vector 300 chiều
  - Output: nhãn phân loại 0/1 (aspect)  
-1/0/1 (sentiment)

## 04 | Phương pháp:



- Tuning sử dụng Optuna

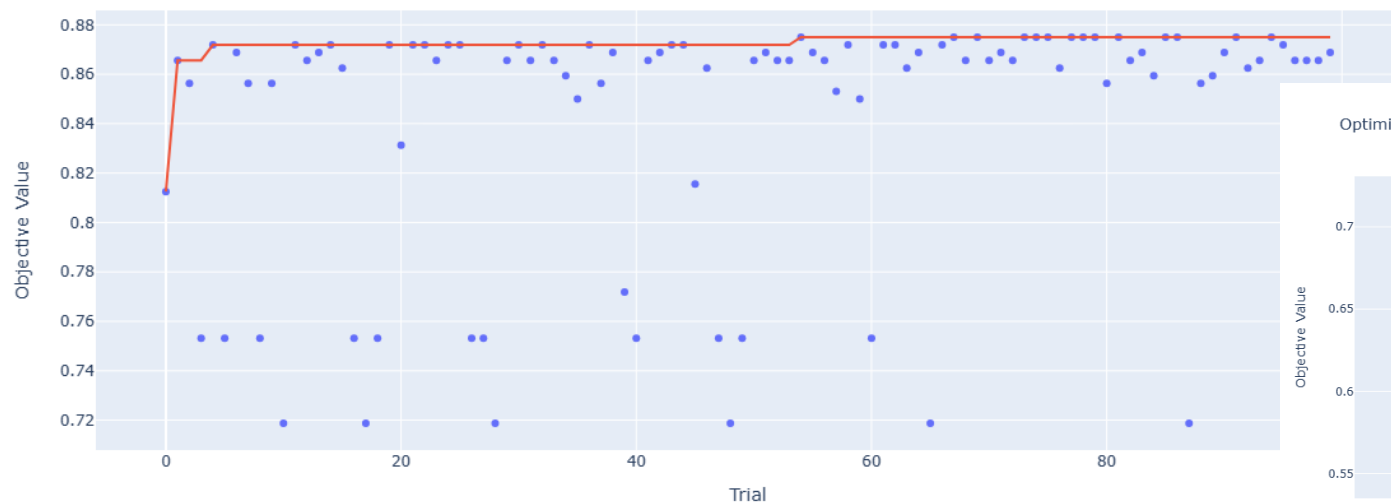
```
def objective(self, trial):  
    kernel = trial.suggest_categorical('kernel', ['linear', 'poly', 'rbf', 'sigmoid'])  
    C = trial.suggest_float('C', 1e-5, 100)  
    gamma = trial.suggest_categorical('gamma', ['scale', 'auto'])  
  
    svc_model = SVCModel(kernel=kernel, C=C, gamma=gamma, attribute=self.attribute)  
    svc_model.fit(self.X_train, self.y_train)  
  
    acc = svc_model.calculate_accuracy_score(self.X_val, self.y_val)  
  
    return acc
```

## 04 Phương pháp:

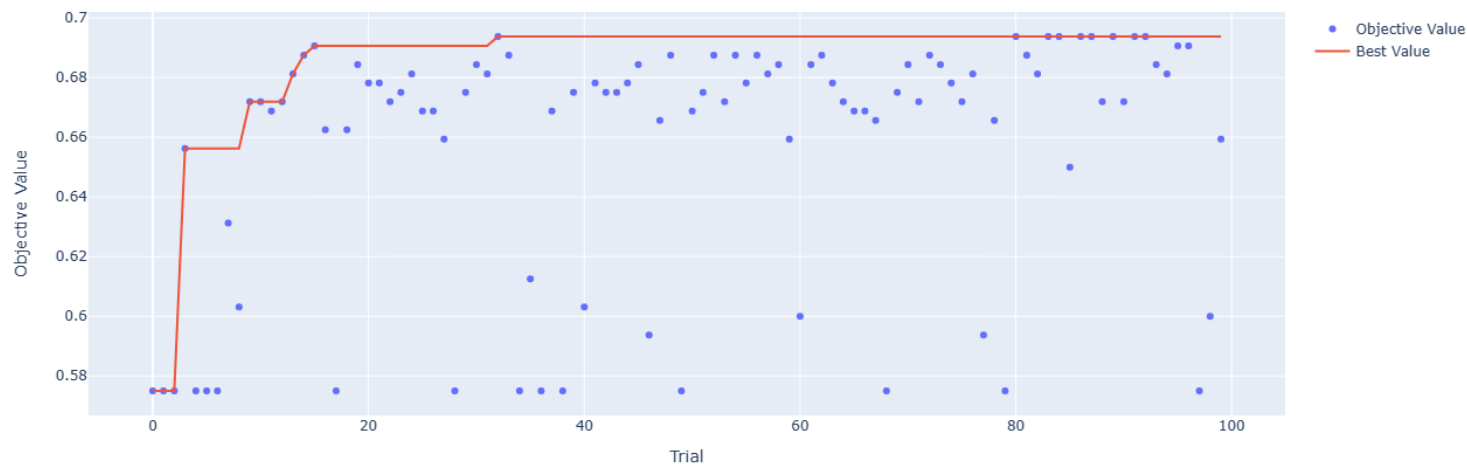
- Tuning sử dụng Optuna

```
def objective(self, trial):  
    kernel = trial.suggest_categorical('C',  
    C = trial.suggest_float('C', 1e-5, 1  
    gamma = trial.suggest_categorical('g  
  
    svc_model = SVCModel(kernel=kernel,  
    svc_model.fit(self.X_train, self.y_t  
  
    acc = svc_model.calculate_accuracy_s
```

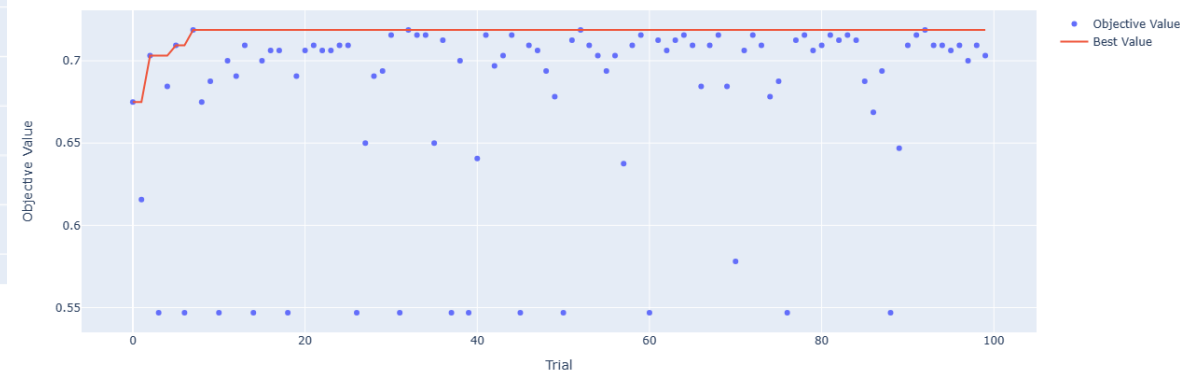
Optimization History Plot



Optimization History Plot



Optimization History Plot





## 05 | Kết quả:



	Base				Best			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
<b>Pin</b>	0.88	<b>0.94</b>	0.8	0.86	<b>0.92</b>	0.93	<b>0.88</b>	<b>0.91</b>
<b>Service</b>	<b>0.83</b>	<b>0.96</b>	0.38	0.54	0.81	0.63	<b>0.72</b>	<b>0.67</b>
<b>General</b>	0.59	<b>0.88</b>	0.39	0.54	<b>0.68</b>	0.82	<b>0.63</b>	<b>0.71</b>
<b>Others</b>	<b>0.84</b>	0.88	<b>0.94</b>	<b>0.91</b>	0.82	<b>0.91</b>	0.87	0.89
<b>SPin</b>	0.65	0.65	0.65	0.65	<b>0.78</b>	<b>0.78</b>	<b>0.78</b>	<b>0.78</b>
<b>Sser</b>	0.82	0.82	0.82	0.82	<b>0.85</b>	<b>0.85</b>	<b>0.85</b>	<b>0.85</b>
<b>SGeneral</b>	0.45	0.45	0.45	0.45	<b>0.65</b>	<b>0.65</b>	<b>0.65</b>	<b>0.65</b>
<b>SOthers</b>	0.61	0.61	0.61	0.61	<b>0.68</b>	<b>0.68</b>	<b>0.68</b>	<b>0.68</b>
<b>Mean</b>	0.70875	0,77375	0,63	0,6725	<b>0.77375</b>	<b>0,78125</b>	<b>0,7575</b>	<b>0,7675</b>

05 | Kết quả:

Comment

sao pin con này bị sao vậy shop

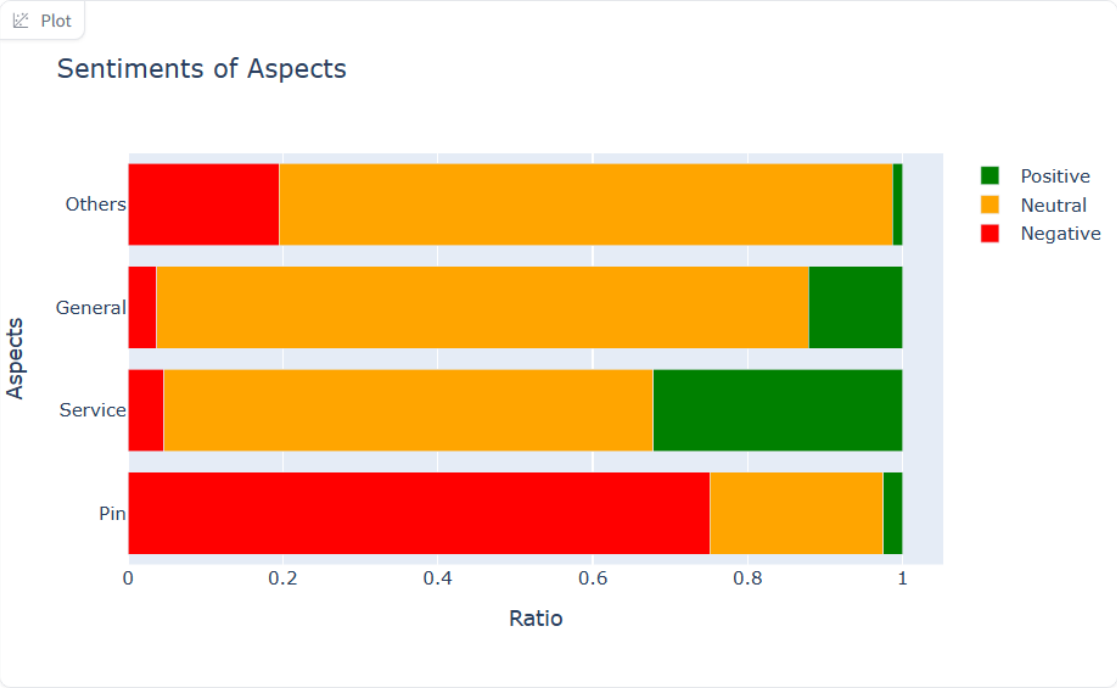
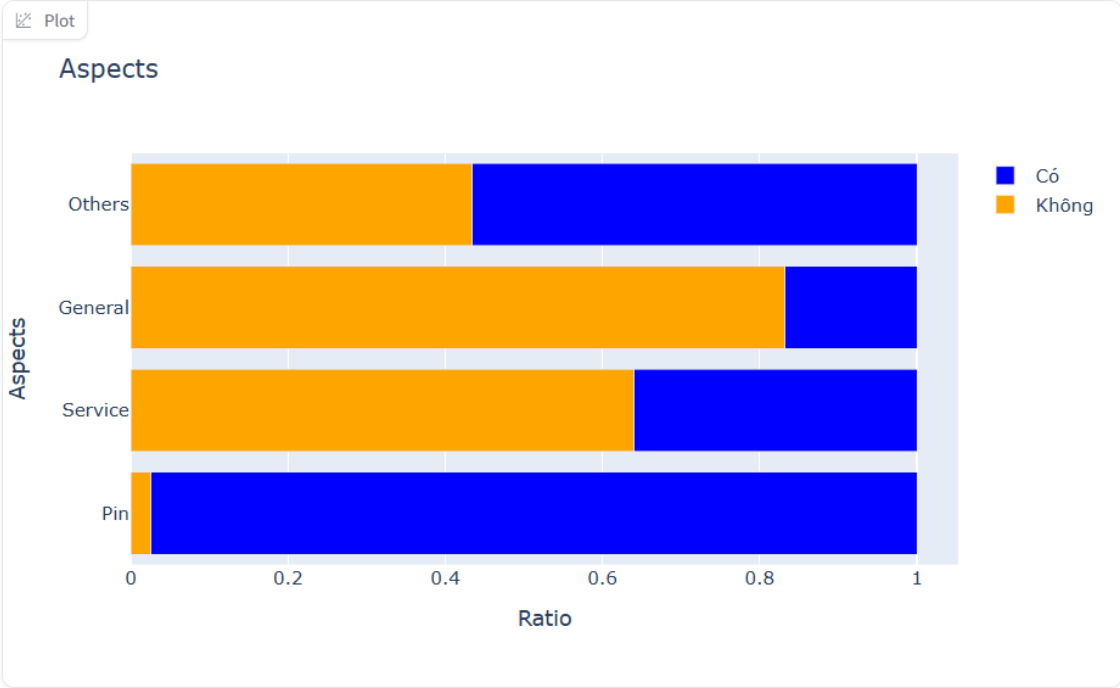
Submit

Choose Plot Type

Bar

Choose Plot Type

Bar



Dự đoán chính xác

05 | Kết quả:

Comment

mới chưa được tháng mà chạy tgdd trả lời giúp ae ai cùng cảnh ngộ xin cách xử lý không xạc qua đêm không vừa dùng vừa xạc nha có ai đổi được máy khác không

Submit

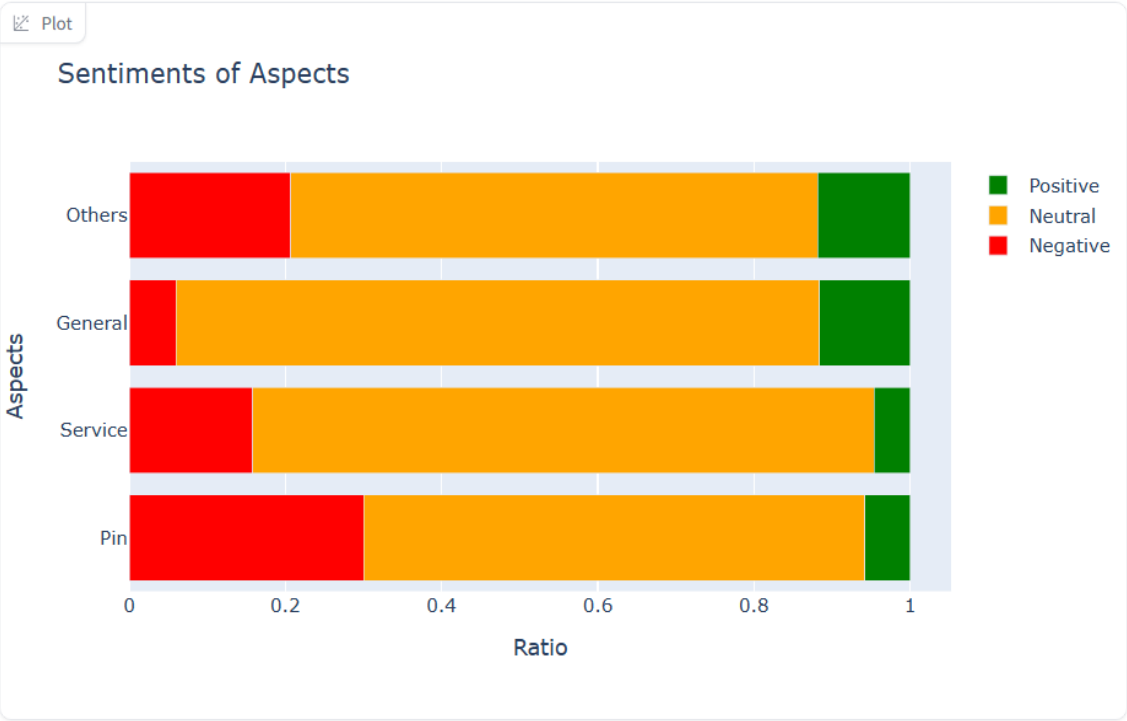
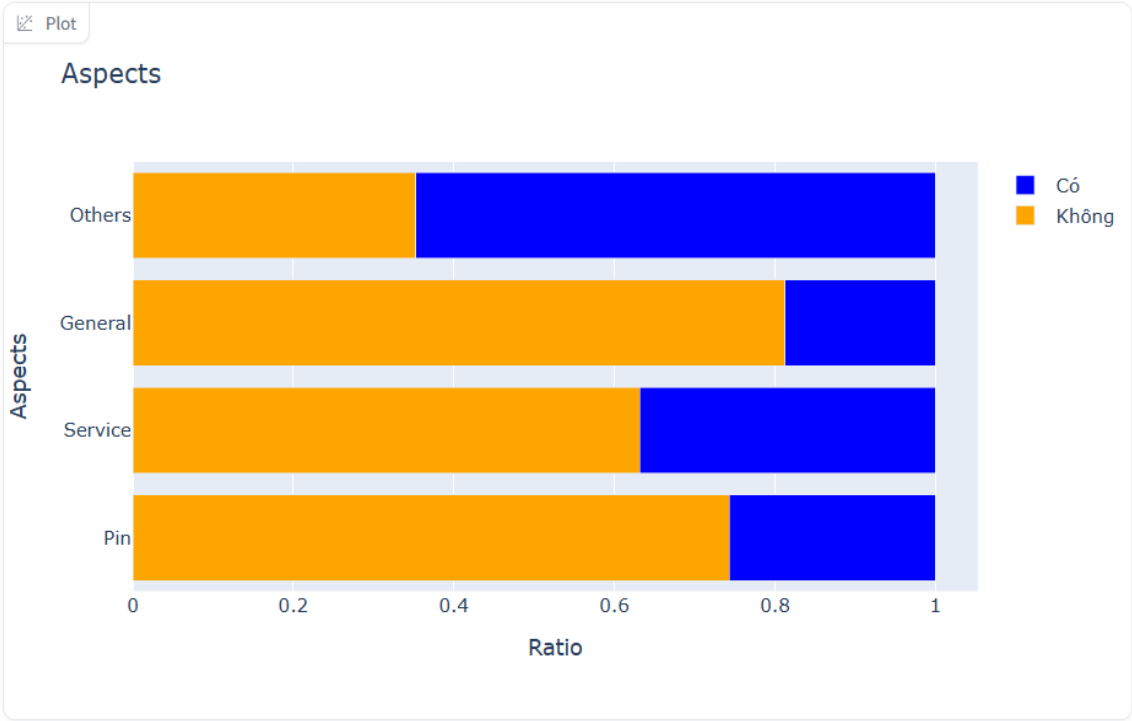
Choose Plot Type

Bar

Không nhắc cụ thể đến pin nhưng hàm ý là pin bị chai

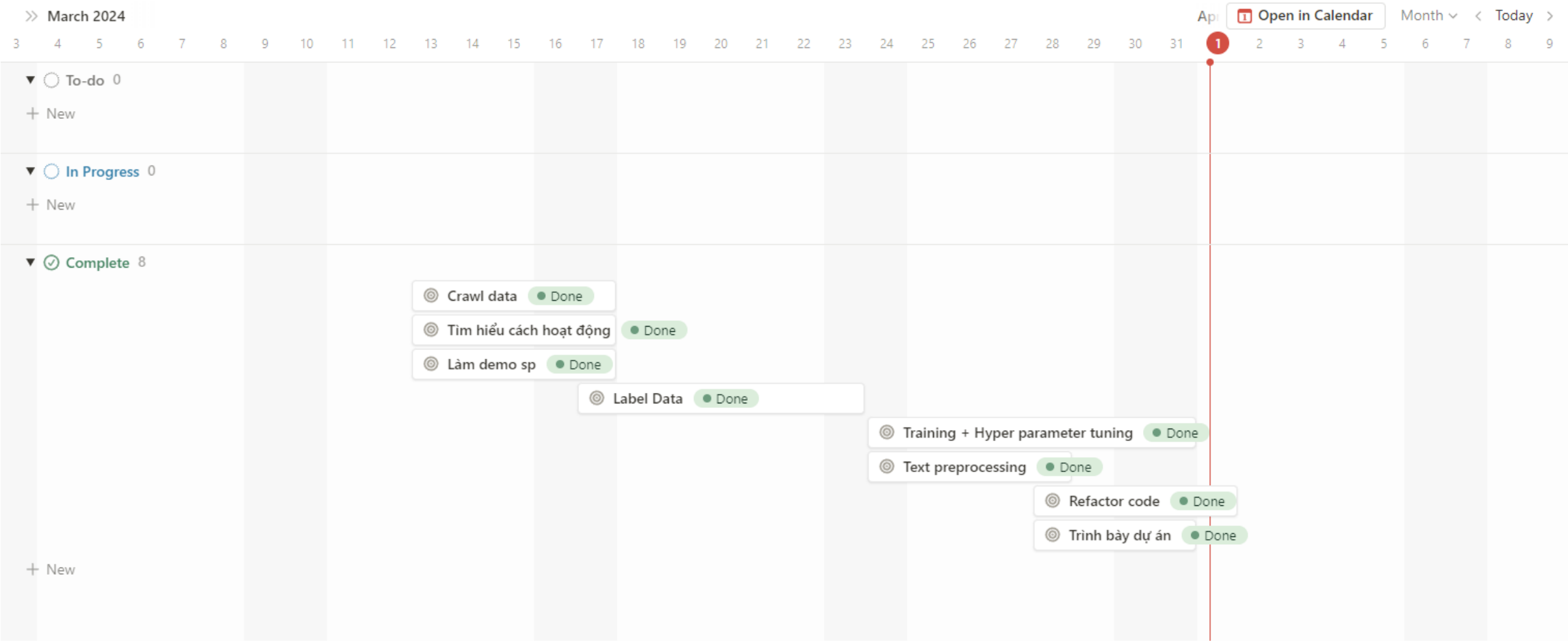
Choose Plot Type

Bar



Dự đoán sai

# 06 | Phân công:



>> March 2024

3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

▼ To-do 0

+ New

🎯

Tìm hiểu cách hoạt động

Status

Done

Owner

Empty

Dates

March 13, 2024 → March 17, 2024

Blocked By

Empty

Is Blocking

Empty

Launch date

March 13, 2024

Priority

Medium

Teams

Đàm Thái Ninh Trần Nam Anh

🎯

Crawl data

Status

Done

Owner

Empty

Dates

March 13, 2024 → March 17, 2024

Blocked By

Empty

Is Blocking

Empty

Launch date

March 13, 2024

Priority

High

Teams

Lê Việt Hùng

🎯 Refactor code

Done

🎯 Trình bày dự án

Done

06 | Phân công:



>> March 2024

3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

▼ To-do 0

+ New

🎯 Tìm hiểu cách hoạt động

🌟 Status

● Done

👤 Owner

Empty

📅 Dates

March 13, 2024 → March 17, 2024

🚫 Blocked By

Empty

🔗 Is Blocking

Empty

📅 Launch date

March 13, 2024

📌 Priority

Medium

👥 Teams

Đàm Thái Ninh

Trần Nam Anh

🎯 Label Data

🌟 Status

● Done

👤 Owner

Empty

📅 Dates

March 17, 2024 → March 23, 2024

🚫 Blocked By

Empty

🔗 Is Blocking

Empty

📅 Launch date

March 17, 2024

📌 Priority

Empty

👥 Teams

Lê Việt Hùng

Đàm Thái Ninh

Trần Nam Anh

Bùi Đức Mạnh

🎯 Crawl data

Month > < Today >

6 7 8 9

● Done

● Done

## 07 | Hạn chế:

- Không sử dụng các mô hình học sâu NLP tiên tiến.
- Quá ít đặc trưng khía cạnh, không đủ để phân tích các khía cạnh cụ thể của sản phẩm như camera, màn hình, chất lượng, giá cả, v.v.
- Phụ thuộc vào Trích xuất Đặc trưng:
  - Hiệu suất của các mô hình SVM phụ thuộc vào chất lượng và tính phù hợp của các đặc trưng được trích xuất từ dữ liệu văn bản.
- Mean embedding xử lý tất cả các từ như nhau, điều này có thể không lý tưởng. Ví dụ: các từ phủ định ("không", "không bao giờ") có thể thay đổi tình cảm, nhưng nếu chỉ tính trung bình thì có thể không nắm bắt được sắc thái đó.
- Mean embedding không xử lý hiệu quả các câu dài trong đó trật tự từ và ngữ cảnh rất quan trọng để hiểu được cảm xúc.
- SVM thường xử lý mỗi đặc trưng một cách độc lập, bỏ qua các phụ thuộc ngữ cảnh trong các câu hoặc tài liệu.

Comment

minh mua sản phẩm tại tgd lê đại hành phường đà lạt nhân viên tư vấn tốt nhiệt tình mình nhu cầu chơi game và xem youtube sản phẩm đáp ứng tốt chiến mạnh pin trâu

Submit

Choose Plot Type

Bar

Plot

Aspects



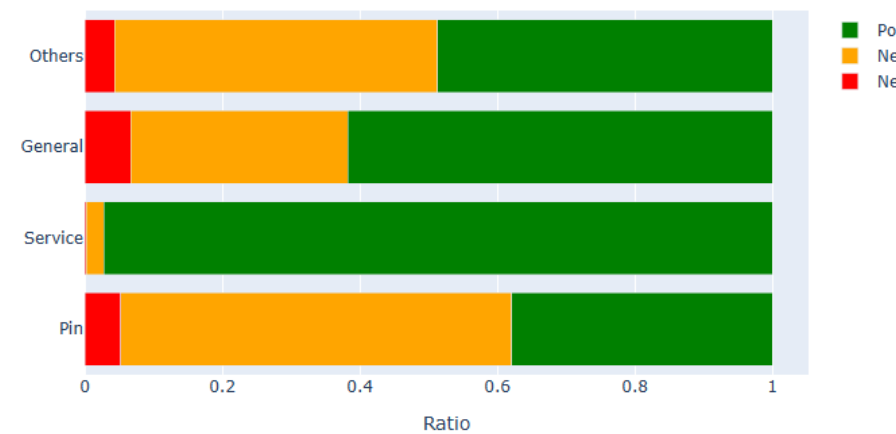
Aspects	Có (Blue)	Không (Orange)
Others	0.6	0.4
General	0.65	0.35
Service	0.88	0.12
Pin	0.3	0.7

Choose Plot Type

Bar

Plot

Sentiments of Aspects



Aspects	Positive (Green)	Neutral (Orange)	Negative (Red)
Others	0.5	0.45	0.05
General	0.62	0.32	0.06
Service	0.98	0.02	0.0
Pin	0.38	0.58	0.04

[Link](#)



## 09 | Future work:



- Nghiên cứu và triển khai các mô hình học sâu NLP tiên tiến như BERT, GPT để cải thiện khả năng phân tích và hiểu biết ngôn ngữ tự nhiên, từ đó cải thiện độ chính xác của việc phân tích cảm xúc và các khía cạnh của sản phẩm.
- Phát triển các phương pháp và mô hình để phân tích các khía cạnh cụ thể của sản phẩm như chất lượng, giá cả, dịch vụ khách hàng, v.v., để cung cấp cho người dùng thông tin chi tiết và đáng tin cậy.
- Cải tiến thành một trang web phân tích bán hàng, nơi người dùng chỉ cần dán đường link về một sản phẩm, hệ thống sẽ tự động thu thập và phân tích dựa trên các đánh giá về sản phẩm đó
- Thêm tính năng cho phép người dùng tương tác với hệ thống bằng cách thêm đánh giá của riêng họ, bình luận và đánh giá sản phẩm, từ đó tăng cường tính tương tác và tính đa dạng của dữ liệu.

## 10 | Tổng kết:



- Bài toán: Phân tích đánh giá lĩnh vực điện thoại di động
  - Input: đánh giá dạng text
  - Output: Aspect, Sentiment
- Đạt được mục tiêu đề ra
- Độ chính xác: 8 mô hình SVM cho độ chính xác trung bình 77% trên bộ tập test UIT-ViSFD
- Khó khăn:
  - Thời gian dành cho project bị hạn chế
  - Hạn chế về phần cứng
  - Model w2v tốn rất nhiều thời gian mỗi lần chạy