PHÂN LOẠI ĐÁNH GIÁ SẢN PHẨM

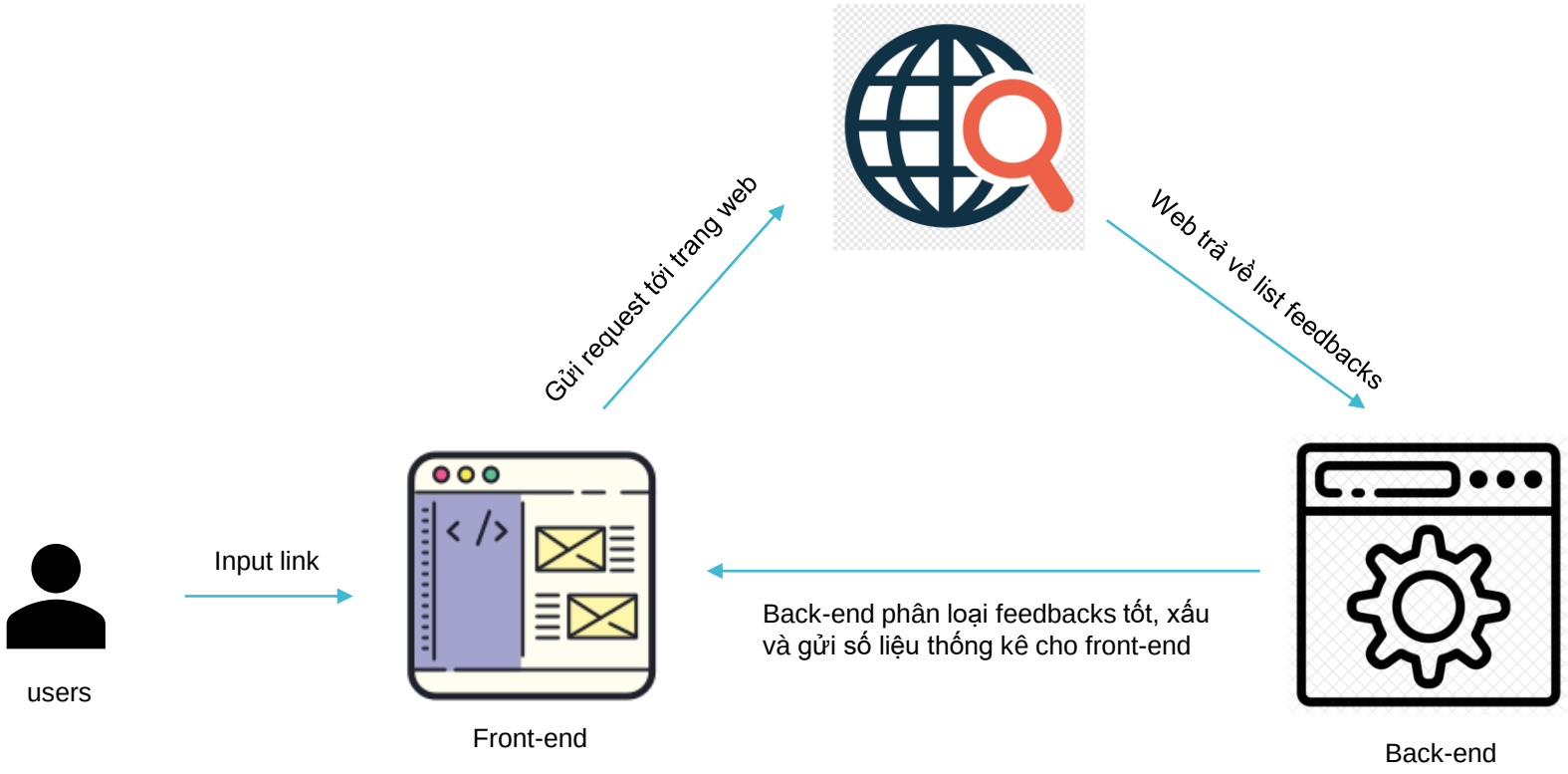
GVHD:
Thầy Phạm Nguyễn Tường An
Thầy Lê Đình Duy

SVTH:
Nguyễn Đình Vinh - 16521582
Phan Đăng Lâm – 16521710

Đặt vấn đề

- Bán hàng online là xu thế công nghệ của ngày nay, tuy nhiên khó có thể mà kiểm định được chất lượng có đảm bảo hay không. Đặc biệt với các mặt hàng đắt tiền như đtdđ thì có nhan nhản ở khăp mọi nơi, 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 feedback từ những người đã mua trước, tuy nhiên số feedback lớn và không có nhân lực để thống kê được hết nên áp dụng machine learning trong việc phân loại feedback của khách hàng là một việc đơn giản và hiệu quả.

Sơ Đồ Hệ Thống



Tổng quan

- **I) Thu thập và mô tả dữ liệu**
- **II) Tiền xử lý dữ liệu**
- **III) Feature Engineering**
- **IV) Training và thực nghiệm**
- **V) Đánh giá mô hình và tinh chỉnh Hyperparameter**
- **VI) Chạy demo thử nghiệm**

I. Thu thập & mô tả dữ liệu

Thu thập các review của khách hàng khi mua điện thoại trên Thế giới di động bằng Javascript.

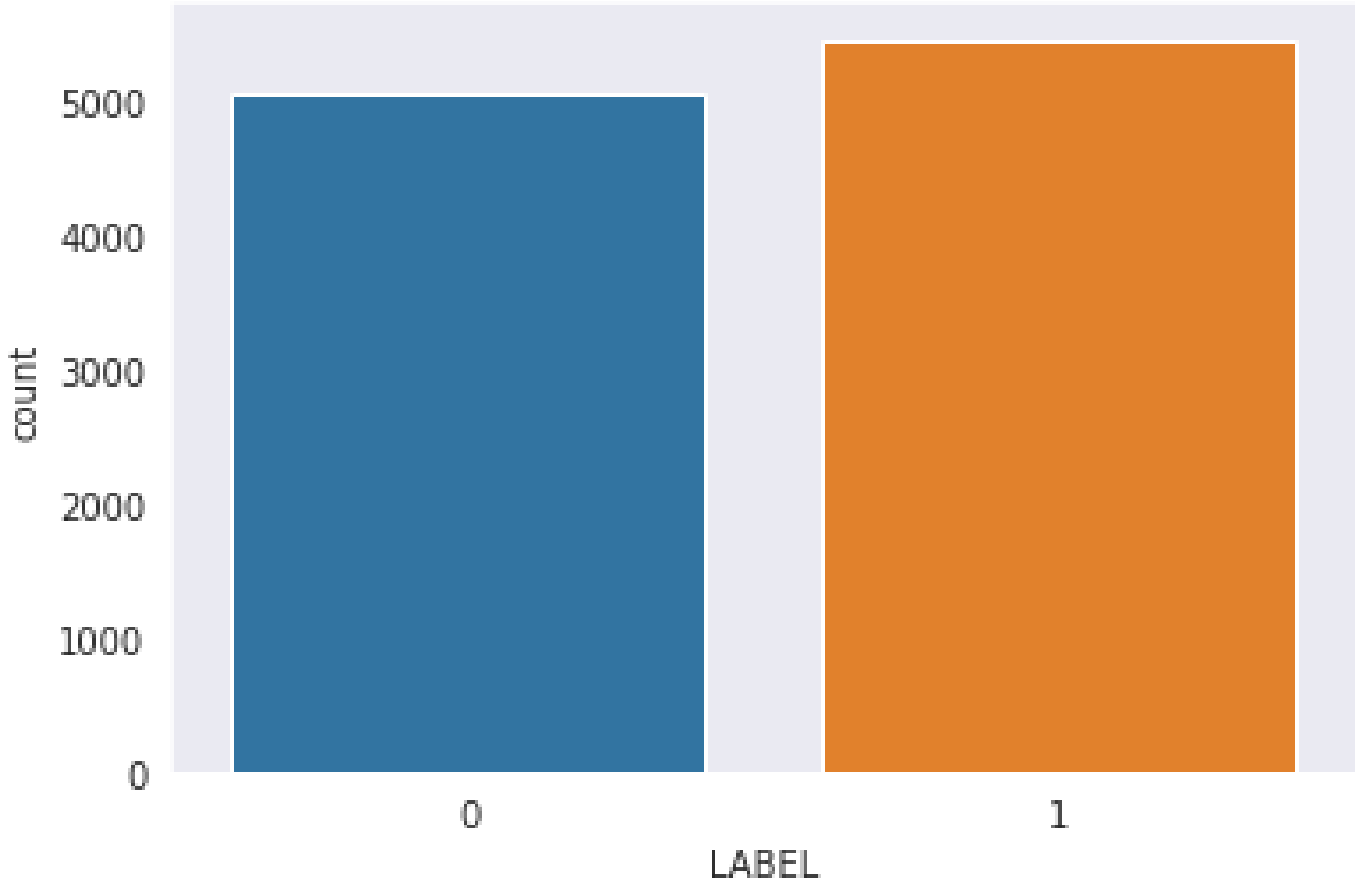
- Kích thước dữ liệu: 10545 mẫu, dưới 3 sao gán label 0, trên 3 sao gán label = 1

- Gồm các thuộc tính: CONTENT, STAR, LABEL

	CONTENT	STAR	LABEL
0	Vừa dùng được 3 ngày. Nói chung là khá thất vọ...	3	0
1	Nói tóm lại là xiaomi dạo này dùng chán camera...	1	0
2	Giá con này hơi cao nếu giá này nên chọn s10 l...	1	0
3	toàn chip đời thấp, chip 865 k thấy bán, giá v...	2	0
4	Máy dùng ok, mình dùng được có 13 ngày máy lỗi...	3	0

I. Mô tả dữ liệu

Số lượng mẫu giữa 2 class 0 và 1 (negative và positive) khá đồng đều, không bị mất cân bằng



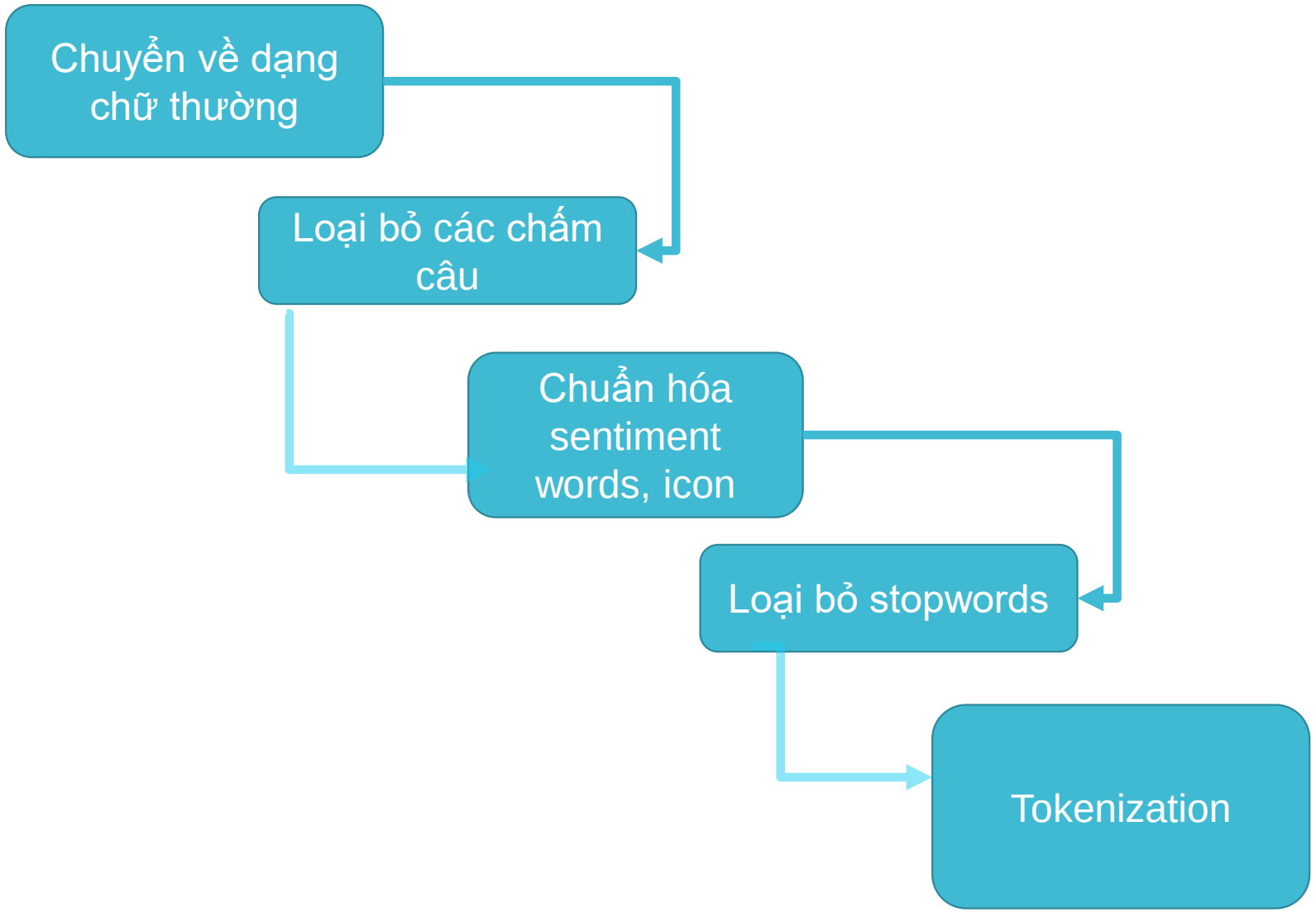
I. Thu thập & mô tả dữ liệu

Kết luận chung về dữ liệu

- Về chất lượng dữ liệu: dựa trên nội dung và số sao mà khách hàng đánh giá, nội dung đánh giá đôi khi không hợp với số sao, nên chỉ có tính chất tương đối
- Vì là feedback của mọi người nên không có quy chuẩn về mặt ngữ pháp, viết tắt nhiều, sai chính tả nhiều, teencode

17 Dt mik mua đk 1 năm rồy.. xài rất ổn .. rất mượt.. pin trâu.. mà tự nhiên mấy ngày nay lên androi 10 thấy sao sao á.. xạc pin chậm hơn.. lướt fb hay chơi game hay bị đứng đứng màn hình.. lắc lắc kg đk mượt như trước nữa.. wifi thì bắt kém... giờ mik mún đổi phềm mềm lại ar9 đk hk ạk.. hay là có cách nào để máy xài mượt như lúc đầu hk ạk...

II. Tiền Xử lý dữ liệu



II. Tiền Xử lý dữ liệu

Chuyển về dạng chữ thường

Original texts:
Máy ngon trong tầm giá. Phù hợp cho học sinh.
Nhân viên bán hàng tư vấn ok. Pin trâuuuu 👍

After cleansed:
máy ngon trong tầm giá. phù hợp cho học sinh.
nhân viên bán hàng tư vấn ok. pin trâuuuu 👍

II. Tiền Xử lý dữ liệu

Loại bỏ các dấu chấm câu, kí tự đặc biệt

```
Original texts:
Máy ngon trong tầm giá. Phù hợp cho học sinh.
Nhân viên bán hàng tư vấn ok. Pin trâuuuu 👍

After cleansed:
máy ngon trong tầm giá phù hợp cho học sinh
nhân viên bán hàng tư vấn ok pin trâuuuu 👍
```

II. Tiền Xử lý dữ liệu

Chuẩn hóa sentiment words, icon

```
#Quy các icon về 2 loại emoj: Tích cực hoặc tiêu cực
"👹": "nagative", "🐶": "positive", "💃": "positive",'👊': ' positive ', '👍': ' positive ',
"💄": "positive", "💎": "positive", "💩": "positive","😕": "nagative", "😱": "nagative", "😻": "positive",
"🦊": "nagative", "🚫": "nagative", "🤬": "nagative","🤷": "positive", "❤️": "positive",'🐶':' positive ',
'👎': ' nagative ', '😖': ' nagative ','✨': ' positive ', '❣': ' positive ','☀': ' positive ',
'♥': ' positive ', '😍': ' positive ', 'like': ' positive ', '📩': ' positive ',
'🤣': ' positive ', '🖤': ' positive ', '😋': ' positive ', ':(': ' nagative ', '😦': ' nagative ',
'❤️': ' positive ', '🤩': ' positive ', '😘': ' positive ', '😪': ' nagative ', '😊': ' positive ',
'?': ' ? ', '😁': ' positive ', '💔': ' positive ', '😐': ' nagative ', '😭': ' nagative ',
'💯': ' positive ', '❤': ' positive ', '♡': ' positive ', '💜': ' positive ', '🤗': ' positive ',
'^^': ' positive ', '😨': ' nagative ', '☺': ' positive ', '💋': ' positive ', '👌': ' positive ',
'😣': ' nagative ', '😀': ' positive ', ':((': ' nagative ', '😡': ' nagative ', '😠': ' nagative ',
'😒': ' nagative ', '🙂': ' positive ', '😉': ' nagative ', '😝': ' positive ', '😅': ' positive ',
'😊': ' positive ', '😷': ' nagative ', '😎': ' positive ', '😆': ' positive ', '💚': ' positive ',
'✌': ' positive ', '❤️': ' positive ', '😔': ' nagative ', '😓': ' nagative ', '🆗': ' positive ',
'😉': ' positive ', '😂': ' positive ', ':v': ' positive ', '=))': ' positive ', '😋': ' positive ',
'❤️': ' positive ', '😐': ' nagative ', ':3': ' positive ', '😫': ' nagative ', '😥': ' nagative ',
'😃': ' positive ', '😄': ' 😄 ', '😌': ' 😌 ', '💛': ' positive ', '🤝': ' positive ', '🎈': ' positive ',
'😗': ' positive ', '🤔': ' nagative ', '😑': ' nagative ', '🔥': ' nagative ', '🙏': ' nagative ',
'🆗': ' positive ', '😻': ' positive ', '💙': ' positive ', '💯': ' positive ',
'😁': ' positive ', '❌': ' nagative ', '👏': ' positive ', ';)': ' positive ', '<3': ' positive ',
'🙂': ' positive ', '🌷': ' positive ', '🌸': ' positive ', '🌺': ' positive ',
'🌼': ' positive ', '🍓': ' positive ', '🐅': ' positive ', '🐾': ' positive ', '👉': ' positive ',
'🌻': ' positive ', '🌹': ' positive ', '💥': ' positive ', '💪': ' positive ',
'💰': ' positive ', '😇': ' positive ', '😛': ' positive ', '😜': ' positive ',
'🙃': ' positive ', '🤢': ' positive ', '🤪': ' positive ','☹': ' nagative ', '💀': ' nagative ',
'😓': ' nagative ', '😟': ' nagative ', '😩': ' nagative ', '😰': ' nagative ', '😳': ' nagative ',
'😮': ' nagative ', '😶': ' nagative ', '😕': ' nagative ',
#Chuẩn hóa 1 số sentiment words/English words:
```

Cách quy chuẩn trên được tham khảo tại: https://forum.machinelearningcoban.com/t/chia-se-model-sentiment-analysis-aivivn-com-top-5/4537

II. Tiền Xử lý dữ liệu

Chuẩn hóa sentiment words, icon

```
#Chuẩn hóa 1 số sentiment words/English words
':))': '  positive ', ':)': ' positive ', 'ô kêi': ' ok ', 'okie': ' ok ', ' o kê ': ' ok ',
'okey': ' ok ', 'ôkê': ' ok ', 'oki': ' ok ', ' oke ':  ' ok ',' okay':' ok ','okê':' ok ',
' tks ': u' cám ơn ', 'thks': u' cám ơn ', 'thanks': u' cám ơn ', 'ths': u' cám ơn ', 'thank': u' cám ơn ',
'⭐': 'star ', '*': 'star ', '🌟': 'star ', '🎉': u' positive ',
'kg ': u' không ','not': u' không ', u' kg ': u' không ', '"k ': u' không ',' kh ':u' không ','kô':u' không ','hok':u' không ','
 không phải ',u' kô ': u' không ', '"ko ': u' không ', u' ko ': u' không ', u' k ': u' không ', 'khong': u' không ', u' hok ': u'

'he he': ' positive ','hehe': ' positive ','hihi': ' positive ', 'haha': ' positive ', 'hjhj': ' positive ',
' lol ': ' nagative ',' cc ': ' nagative ','cute': u' dễ thương ','huhu': ' nagative ', ' vs ': u' với ', 'wa': ' quá ', 'wá': u'
j ': u' gì ', '“': ' ',
' sz ': u' cỡ ', 'size': u' cỡ ', u' đx ': u' được ', 'dk': u' được ', 'dc': u' được ', 'đk': u' được ',
'đc': u' được ','authentic': u' chuẩn chính hãng ',u' aut ': u' chuẩn chính hãng ', u' auth ': u' chuẩn chính hãng ', 'thick': u'
 ', 'store': u' cửa hàng ',
'shop': u' cửa hàng ', 'sp': u' sản phẩm ', 'gud': u' tốt ','god': u' tốt ','wel done':' tốt ', 'good': u' tốt ', 'gút': u' tốt

'sấu': u' xấu ','gut': u' tốt ', u' tot ': u' tốt ', u' nice ': u' tốt ', 'perfect': 'rất tốt', 'bt': u' bình thường ',
'time': u' thời gian ', 'qá': u' quá ', u' ship ': u' giao hàng ', u' m ': u' mình ', u' mik ': u' mình ',
'ể': 'ể', 'product': 'sản phẩm', 'quality': 'chất lượng','chat':' chất ', 'excelent': 'hoàn hảo', 'bad': 'tệ','fresh': ' tươi
```

Cách quy chuẩn trên được tham khảo tại: https://forum.machinelearningcoban.com/t/chia-se-model-sentiment-analysis-aivivn-com-top-5/4537

II. Tiền Xử lý dữ liệu

Loại bỏ stopwords

Before:

mình đã mua a70 sau 1 tuần mình xin nhận xét như sau cấu hình máy khoẻ mượt đa nhiệm tốt pin 4k5 dùng lâu hết nhận diện khuôn mặt rất nhanh có điều là vân tay chưa được nhanh

After:

mình mua a70 1 tuần mình nhận xét cấu hình máy khoẻ mượt đa nhiệm tốt pin 4k5 dùng lâu hết nhận diện khuôn mặt rất nhanh vân tay chưa được nhanh

Nguồn stopwords: https://github.com/stopwords/vietnamese-stopwords

II. Tiền Xử lý dữ liệu

Tách từ (tokenization)

Sử dụng thư viện underthesea để tách từ

```
Before:
mình mua a70 1 tuần mình nhận xét cấu hình máy khoẻ mượt đa nhiệm tốt pin 4k5
dùng lâu hết nhận diện khuôn mặt rất nhanh vân tay chưa nhanh

After:
['mình', 'mua', 'a70', '1', 'tuần', 'mình', 'nhận_xét', 'cấu_hình', 'máy',
'khoẻ', 'mượt', 'đa_nhiệm', 'tốt', 'pin', '4k5', 'dùng', 'lâu', 'hết',
'nhận_diện', 'khuôn_mặt', 'nhanh', 'vân', 'tay', 'chưa', 'nhanh']
```

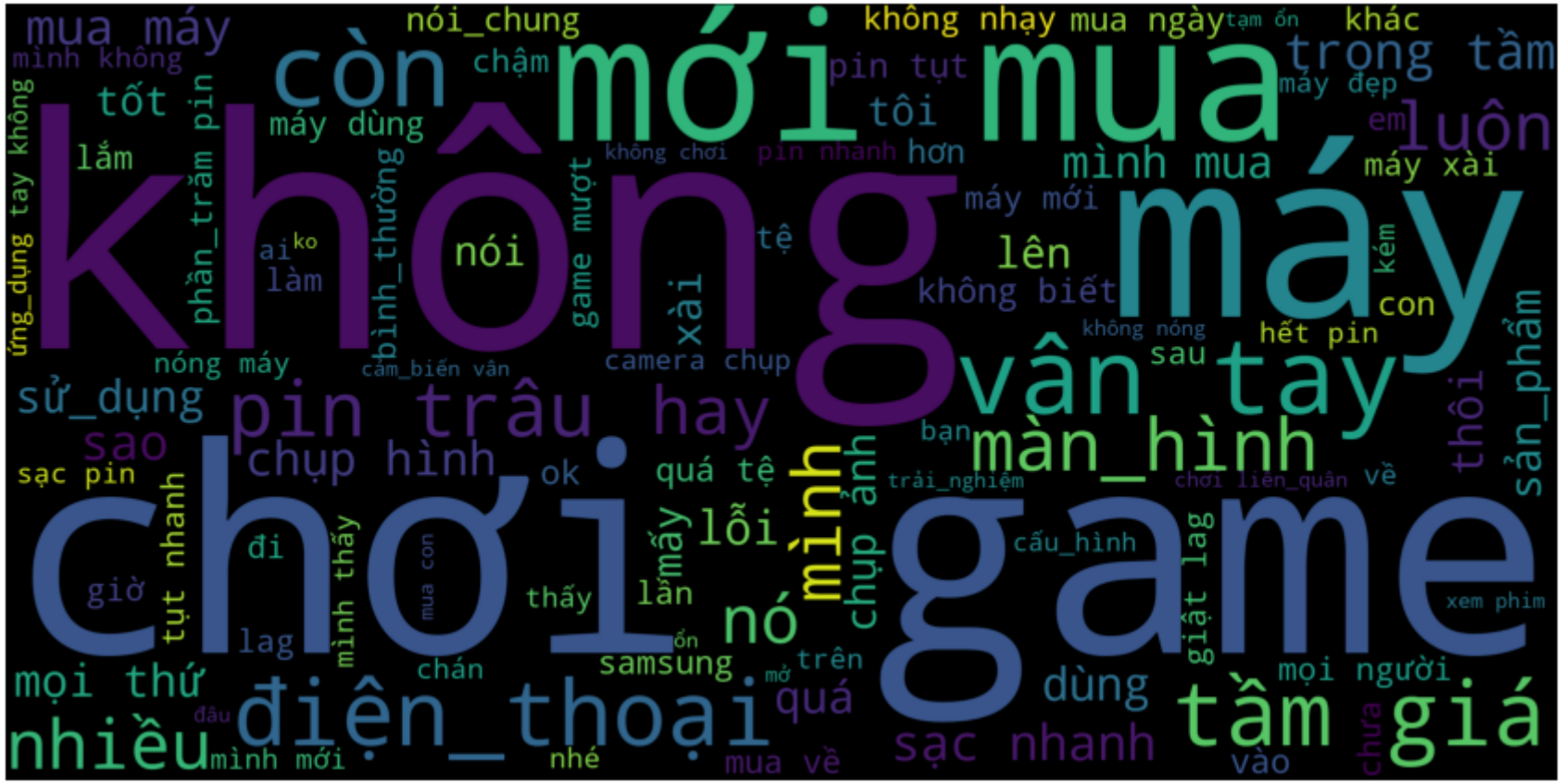
III. Feature Engineering

Xây dựng bộ từ vựng theo cấu trúc Bag-of-Word

```
Number of words: 10274
Most common words: [('không', 9763), ('máy', 8410), ('pin', 6412), ('mua', 6104),
<matplotlib.image.AxesImage at 0x7f3c183a7e80>
```

III. Feature Engineering

wordcloud



III. Feature Engineering

Vectorization

- **Nhóm sử dụng 3 model và mỗi model yêu cầu input đầu vào khác nhau nên việc vectorization cũng khác nhau**

Chia làm 2 nhóm:

- **Vectorization cho input mô hình Neural Network đơn giản**

- **Vectorization cho input mô hình SVM và Naïve bayes**

III.Feature Engineering

Nhóm 1: Vectorization cho input mô hình NN

Tập từ vựng

```
1 check
2 i
3 numbr
4 video
5 song
6 like
7 please
8 subscribe
9 love
10 youtube
11 channel
12 you
13 webaddress
14 get
...
```

```
Before vectorization
 ['i', 'like', 'song', 'i', 'like', 'you']

After vectorization
 [[2], [6], [5], [2], [6], [12]]
```

Tiến hành padding

```
Before vectorization & padding
 ['i', 'like', 'song', 'i', 'like', 'you']

After vectorization & padding
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  2  6  5  2  6 12]
```

IV. Vectorization

Nhóm 2: Vectorization bằng TF- IDF cho SVM classifer và NB classifer

TF-IDF *(Term Frequency – Inverse Document Frequency)* là 1 kỹ thuật sử dụng trong khai phá dữ liệu văn bản. Trọng số này được sử dụng để đánh giá tầm quan trọng của một từ trong một văn bản. Giá trị cao thể hiện độ quan trọng cao và nó phụ thuộc vào số lần từ xuất hiện trong văn bản nhưng bù lại bởi tần suất của từ đó trong tập dữ liệu

IV. Vectorization

Nhóm 2: Vectorization bằng TF- IDF cho SVM classifer và NB classifer

Tập từ vựng

```
233 lmfao
234 lol
235 look
236 lose
237 lot
238 love
...
389 song
390 soon
...
```

Trước khi vectorization

```
token:  ['love', 'song', 'love', 'Taylor', 'Taylor', 'Taylor', 'Taylor', 'Taylor']
```

sau khi vectorization

```
Input vectorization by TF-IDF:
  (0, 389)      0.4621859485949343
  (0, 238)      0.8867830337356488
```

TRAINING VÀ THỰC NGHIỆM

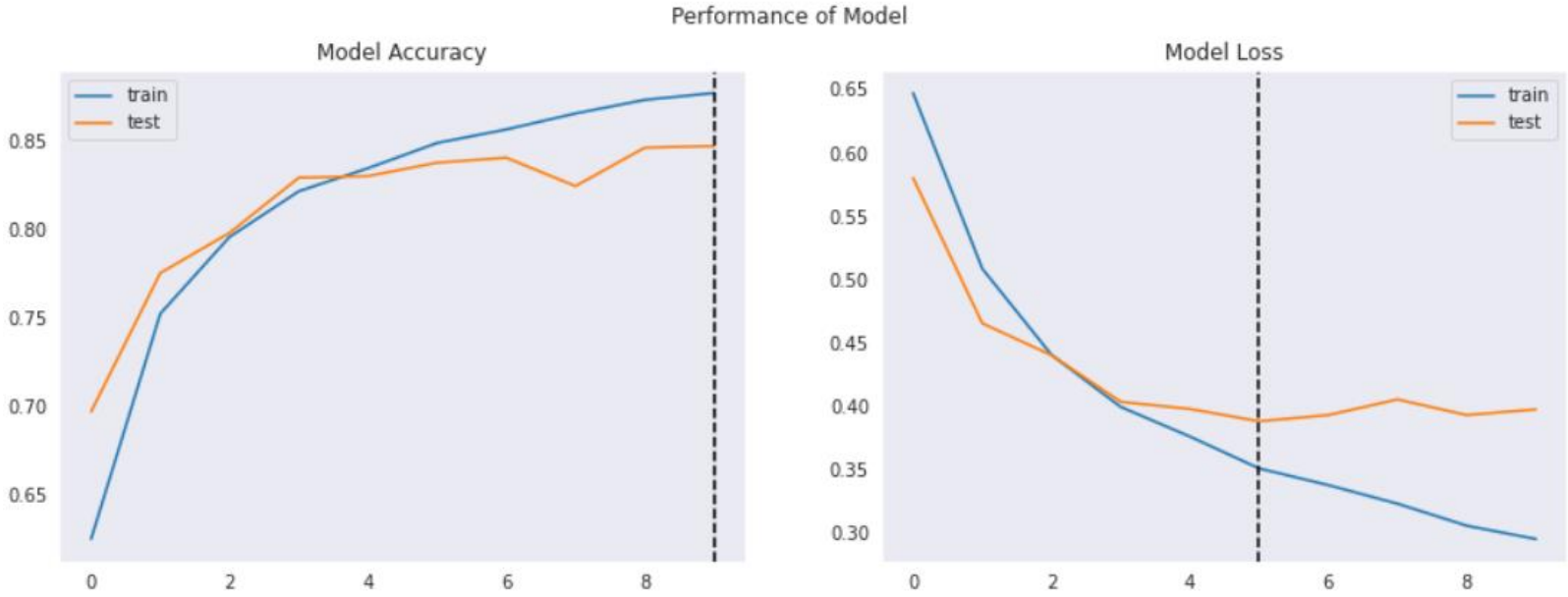
Neural Network đơn giản

```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_2 (Embedding)      (None, 455, 64)           192000
_________________________________________________________________
lstm_2 (LSTM)                (None, 96)                61824
_________________________________________________________________
dense_3 (Dense)              (None, 150)               14550
_________________________________________________________________
batch_normalization_2 (Batch (None, 150)               600
_________________________________________________________________
dropout_2 (Dropout)          (None, 150)               0
_________________________________________________________________
dense_4 (Dense)              (None, 1)                 151
_________________________________________________________________
activation_2 (Activation)    (None, 1)                 0
=================================================================
Total params: 269,125
Trainable params: 268,825
Non-trainable params: 300
_________________________________________________________________
```

Cách build model được tham khảo từ: https://www.kaggle.com/wflazuardy/sarcasm-detection-with-keras-preprocessing

TRAINING VÀ THỰC NGHIỆM

Neural Network đơn giản

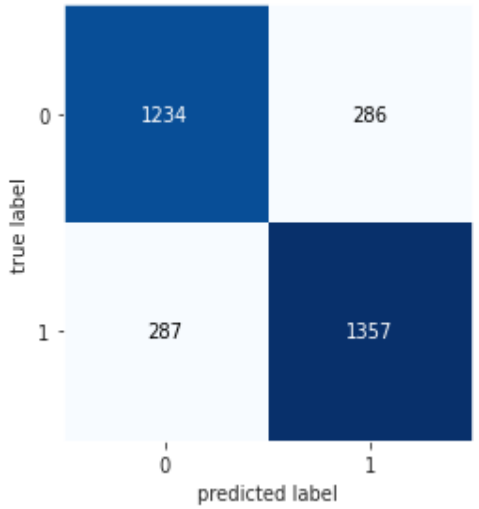


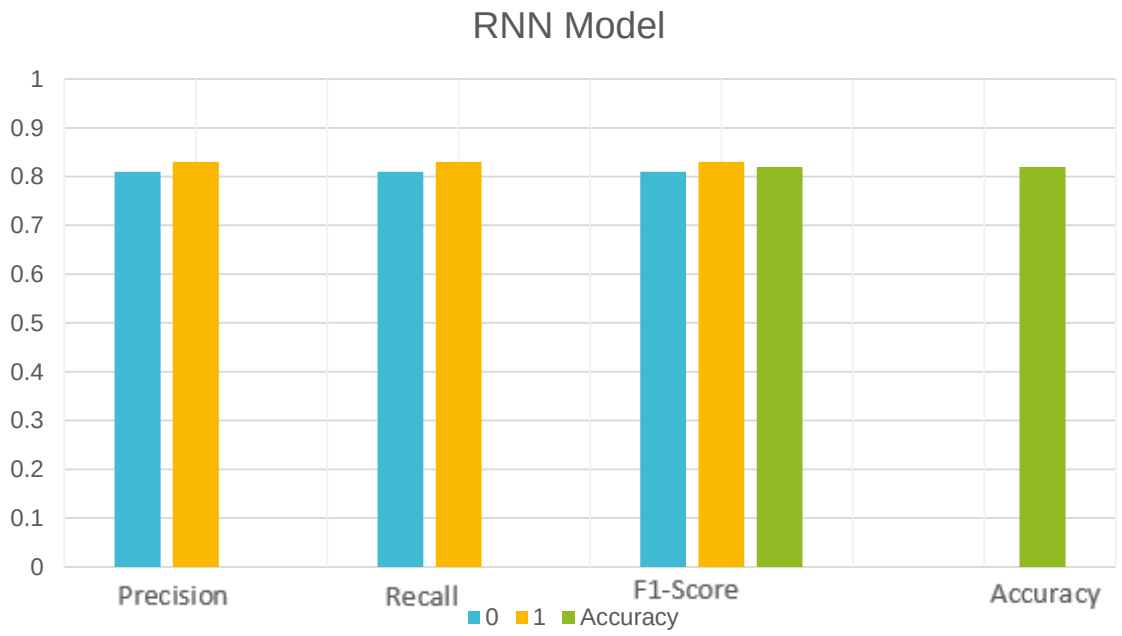
Epochs: 10
Batch_size=32
lr=1e-4
Split_train_test=70/30

TRAINING VÀ THỰC NGHIỆM

Neural Network đơn giản

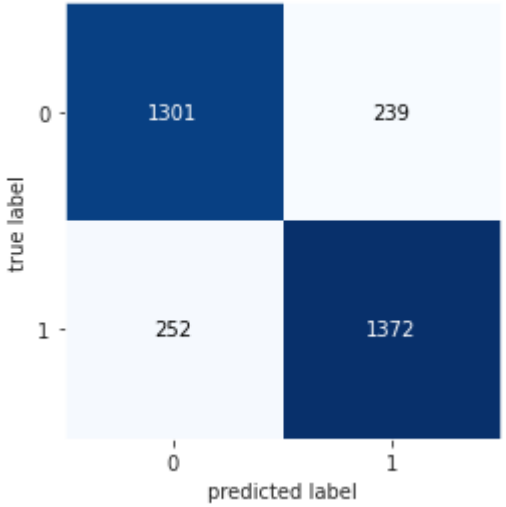
	precision	recall	f1-score	support
0	0.81	0.81	0.81	1520
1	0.83	0.83	0.83	1644
accuracy			0.82	3164
macro avg	0.82	0.82	0.82	3164
weighted avg	0.82	0.82	0.82	3164



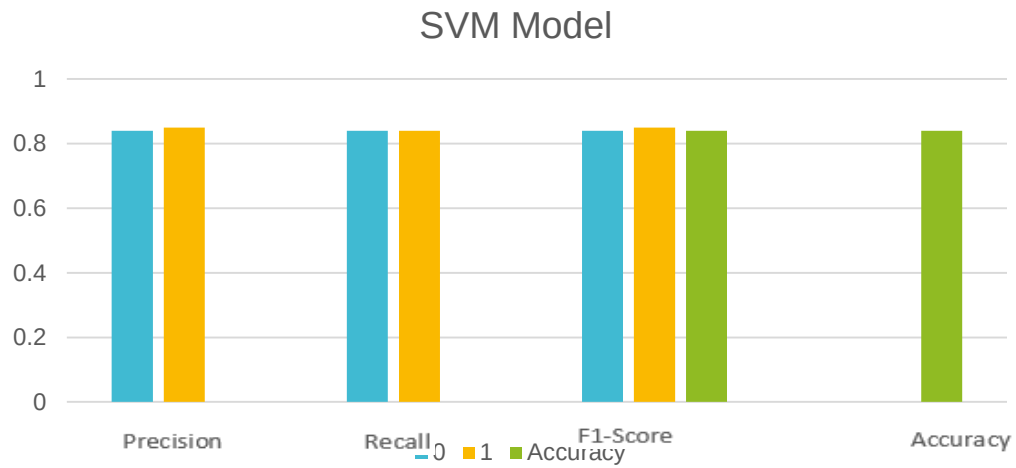


TRAINING VÀ THỰC NGHIỆM

SVM Model

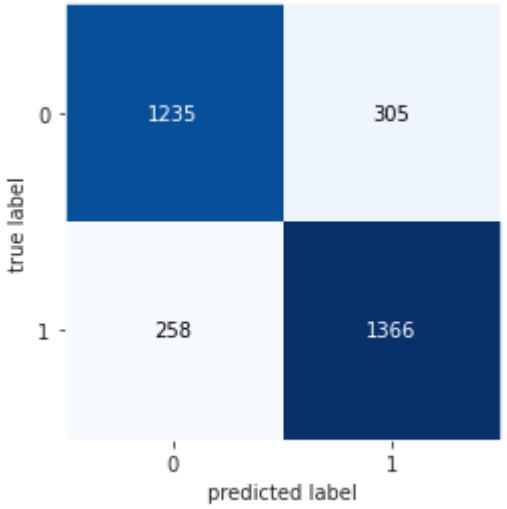


	precision	recall	f1-score	support
0	0.84	0.84	0.84	1540
1	0.85	0.84	0.85	1624
accuracy			0.84	3164
macro avg	0.84	0.84	0.84	3164
weighted avg	0.84	0.84	0.84	3164

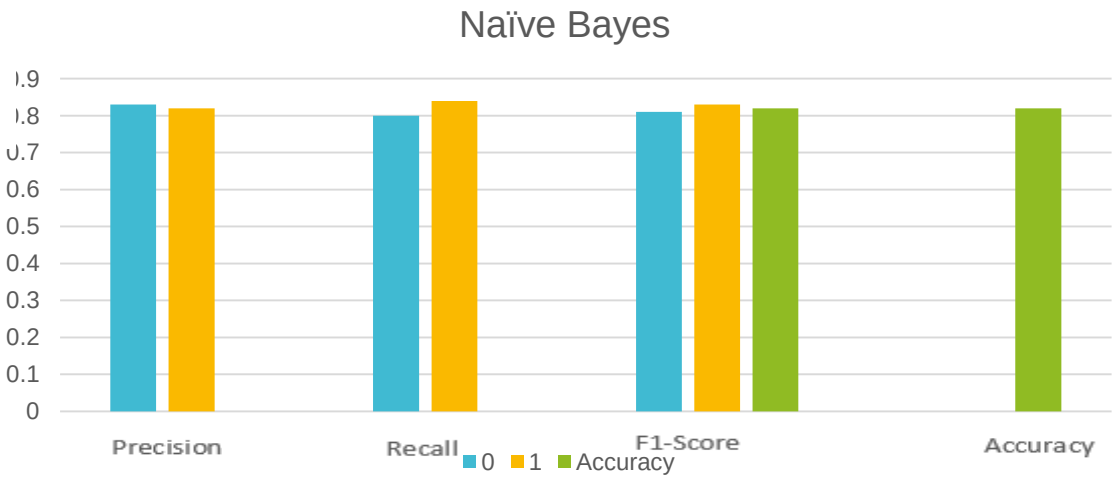


TRAINING VÀ THỰC NGHIỆM

Naïve bayes Model

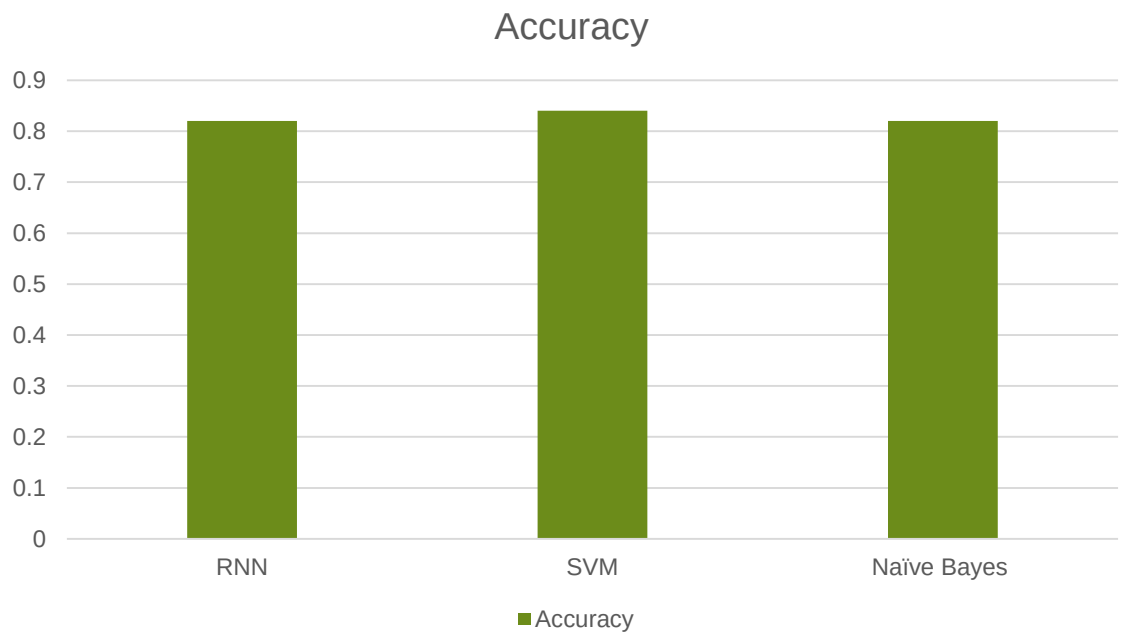


	precision	recall	f1-score	support
0	0.83	0.80	0.81	1540
1	0.82	0.84	0.83	1624
accuracy			0.82	3164
macro avg	0.82	0.82	0.82	3164
weighted avg	0.82	0.82	0.82	3164



SO SÁNH MÔ HÌNH

Accuracy



Tài liệu tham khảo

1. Machine Learning — Word Embedding & Sentiment Classification using Keras
2. Vietnamese Sentiment Analysis
3. Bag-of-words model
4. TF-IDF


thank
you