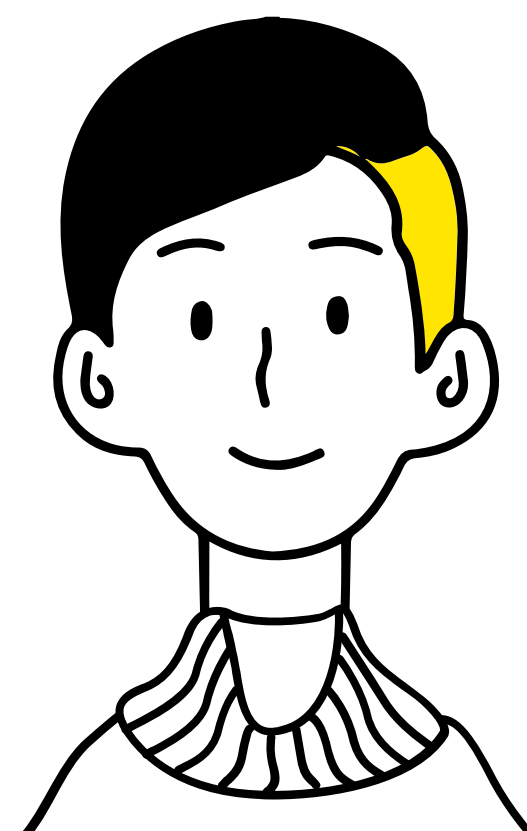
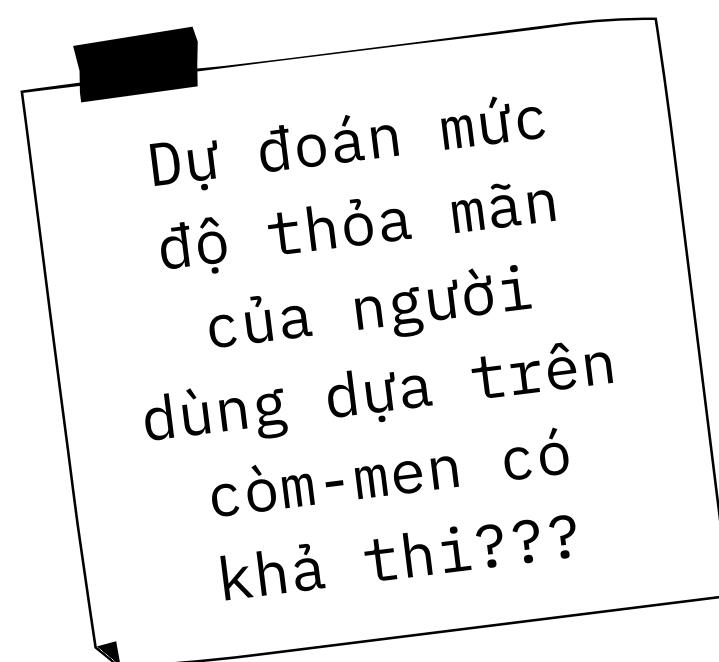


machine_learning_developer

Đồ án tốt nghiệp



Tổng quan

1

Xác định vấn đề, chủ đề

2

Thách thức, phạm vi đề án

3

Giải pháp, hướng tiếp cận

4

Cải thiện, cải tiến

1 Xác định vấn đề, chủ đề đồ án

Sự hữu ích của các phản hồi về sản phẩm? Tại sao chọn chủ đề này???

GÓC NHÌN DOANH NGHIỆP

- Cải thiện chất lượng dịch vụ
- Giảm chi phí
-

NGƯỜI DÙNG

- Tiết kiệm thời gian
- Trải nghiệm dịch vụ, sản phẩm tăng
- ...

Tại sao dùng mô hình AI?

2 Thách thức, phạm vi đồ án

mất cân bằng dữ liệu

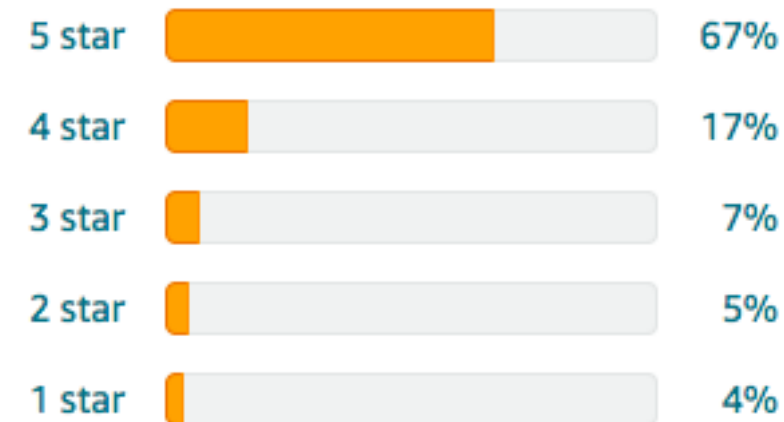
học có giám sát

bắt đầu với điều nhỏ nhất

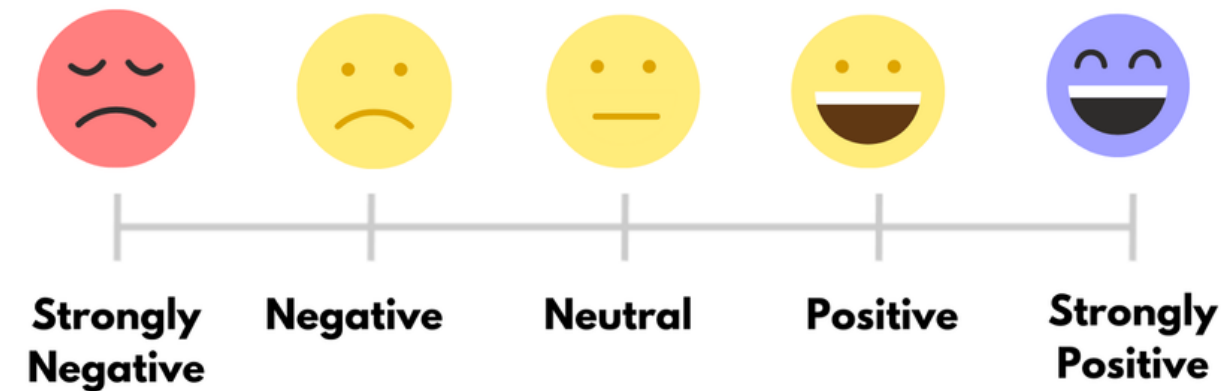
Customer reviews

★★★★☆ 4.4 out of 5

176 global ratings

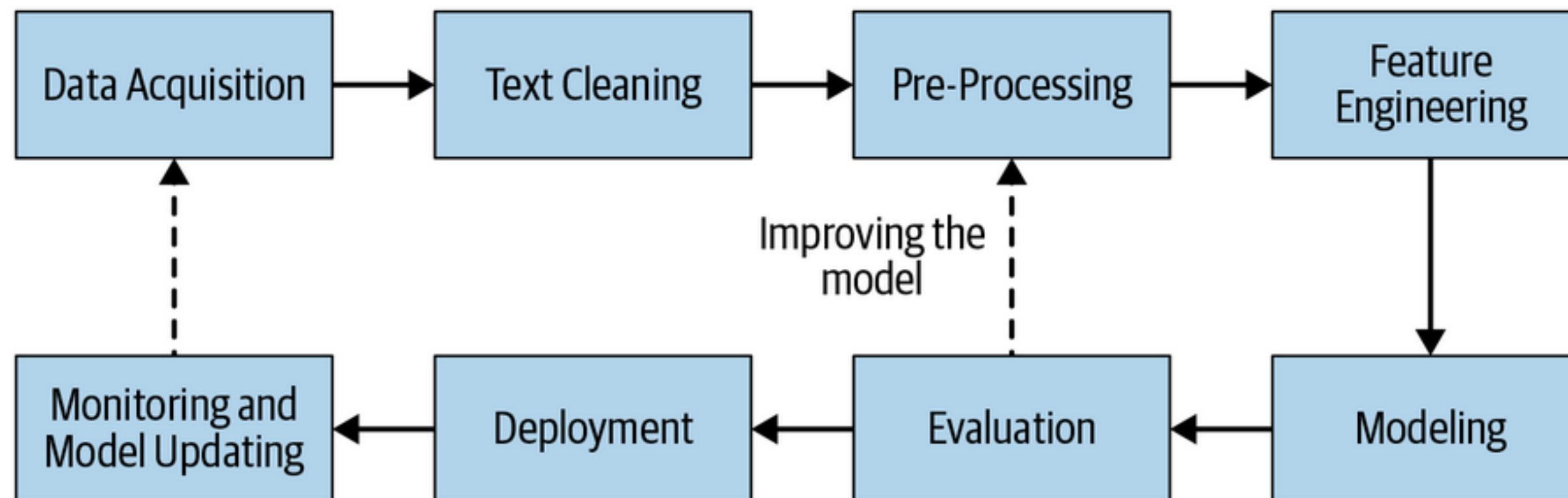


✓ How are ratings calculated?



3 Hướng tiếp cận

- Thu thập dữ liệu
- Tiền xử lý dữ liệu (+làm sạch dữ liệu)
- Feature Engineering
- Triển khai thuật toán
- Đánh giá
- Cải thiện
- Đưa vào production



Dữ liệu

1

Fashion and Beauty
Reviews kaggle
challenge

Detail

Compact

Column

# rating	A review	A summary	A product	A reviewer
<div><div></div></div> <div>1321k</div>	<div><div>[null]62%</div><div>however0%</div><div>Other (433454)38%</div></div>	<div><div>[null]80%</div><div>however0%</div><div>Other (223637)20%</div></div>	<div><div>[null]89%</div><div>however0%</div><div>Other (120535)11%</div></div>	<div><div>[null]94%</div><div>however0%</div><div>Other (69209)6%</div></div>
4.0	I received this cream about 8 days ago and have been using it and it seemed good. It doesn't have a ...	Thought this cream was making me sick...	B00DKEQYJY	A3IQA3VVDHGAK1
5.0	Beautiful pieces	Five Stars	B00L4JJKH0	A1QWCVZMSYG1N2
5.0	Really impressive tree. It goes together quickly and easily. Had a friend (who always gets a real tr...	Highly recommended	1620213982	A2QRLRHMFDJ25E
4.0	good	Four Stars	B01D2J35BG	A2MM01P2ZNEUNF

2

Amzon Review Data 3 bộ
(All Beauty, Clothing Shoes
and Jewelry, Luxury Beauty)

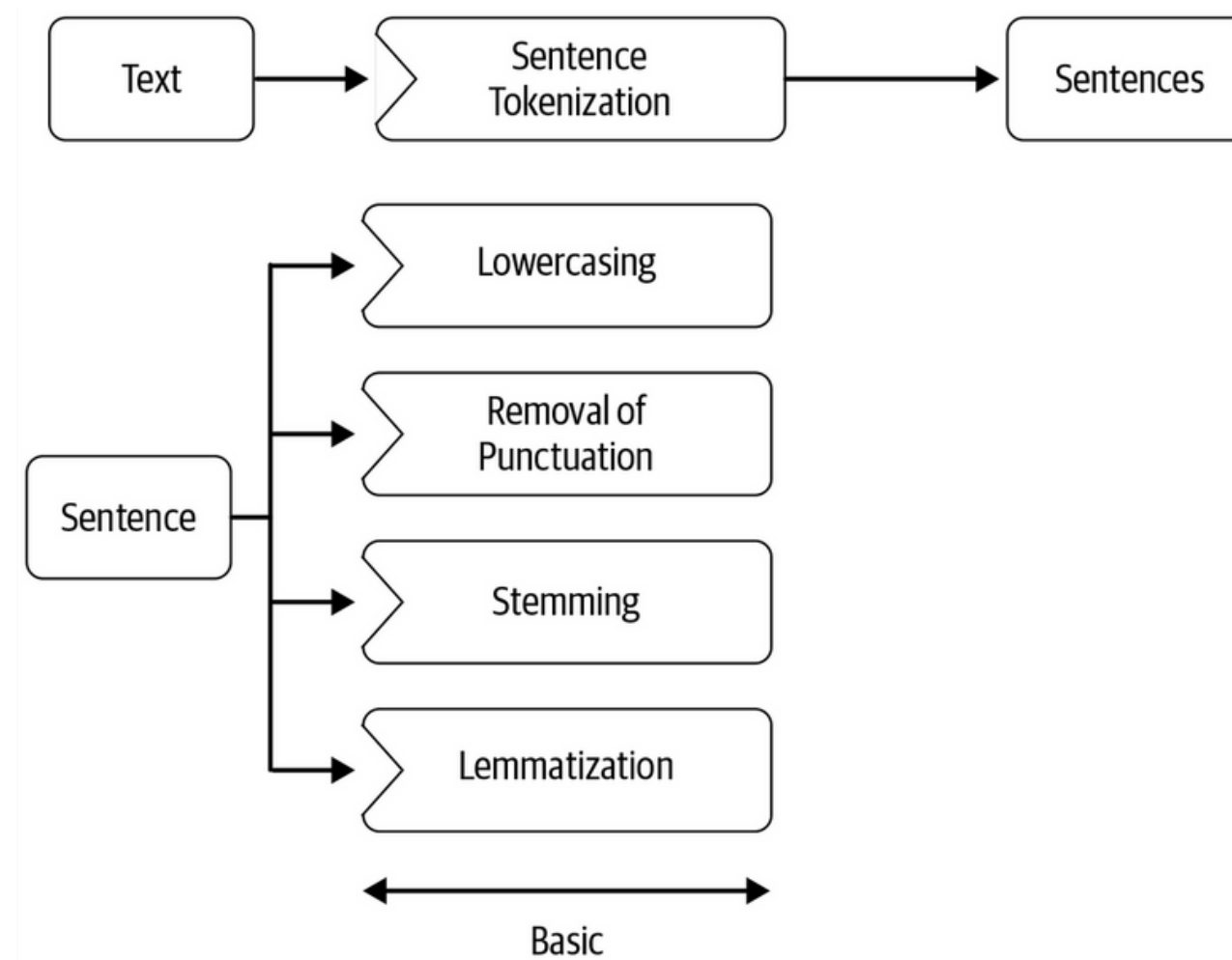
Sample review:

```
{
  "image": ["https://images-na.ssl-images-amazon.com/images/I/71eG75FTJJL._SY88.jpg"],
  "overall": 5.0,
  "vote": "2",
  "verified": True,
  "reviewTime": "01 1, 2018",
  "reviewerID": "AUI6WTTT0QZYS",
  "asin": "5120053084",
  "style": {
    "Size": "Large",
    "Color": "Charcoal"
  },
  "reviewerName": "Abbey",
  "reviewText": "I now have 4 of the 5 available colors of this shirt... ",
  "summary": "Comfy, flattering, discreet--highly recommended!",
  "unixReviewTime": 1514764800
}

{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "vote": 5,
  "style": {
    "Format": "Hardcover"
  },
  "reviewText": "I bought this for my husband who plays the piano. He is having a wonderful time playing these old hymns. The music is at times hard to read because we think the book was published for singing from more than playing from. Great purchase though!",
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```

Tiền xử lý + làm sạch dữ liệu

Xóa những giá trị hoặc trường thông tin không liên quan đến vấn đề giải quyết -> chuyển dữ liệu text thành những vector mô hình có thể xử lý.

[illegible]

Tiền xử lý + làm sạch dữ liệu

#ghi chú

- Nên xóa thẻ html trước khi xóa các dấu ngoặc.
- Làm sạch giúp giảm thời gian huấn luyện, tăng hiệu quả.
- Cẩn thận với 'trùng lặp'

GloVe Embeddings cover 25.77% of vocabulary and 92.70% of text in Training Set

```
]:  
train_glove_oov, train_glove_vocab_coverage, train_glove_text_coverage = check_embeddings(  
    print('GloVe Embeddings cover {:.2%} of vocabulary and {:.2%} of text in Cleaned data')
```

GloVe Embeddings cover 50.96% of vocabulary and 99.49% of text in Cleaned data

```
]:  
test_glove_oov, test_glove_vocab_coverage, test_glove_text_coverage = check_embeddings(  
    print('Crawl 300d2m cover {:.2%} of vocabulary and {:.2%} of text in Original Set')
```

Crawl 300d2m cover 28.62% of vocabulary and 94.04% of text in Original Set

+ Code

+ Markdown

```
test_glove_oov, test_glove_vocab_coverage, test_glove_text_coverage = check_embeddings(  
    print('Crawl 300d2m cover {:.2%} of vocabulary and {:.2%} of text in Cleaned data')
```

Crawl 300d2m cover 51.53% of vocabulary and 99.50% of text in Cleaned data

Feature Engineering

Lựa chọn và trích xuất đặc trưng từ dữ liệu.

1

Chọn cả 4 nếu thiếu dữ liệu

2

Chọn 3 cột 'review' 'summary' 'product' nếu thêm dữ liệu

3

Tổng số hàng được dùng == 2.296.984

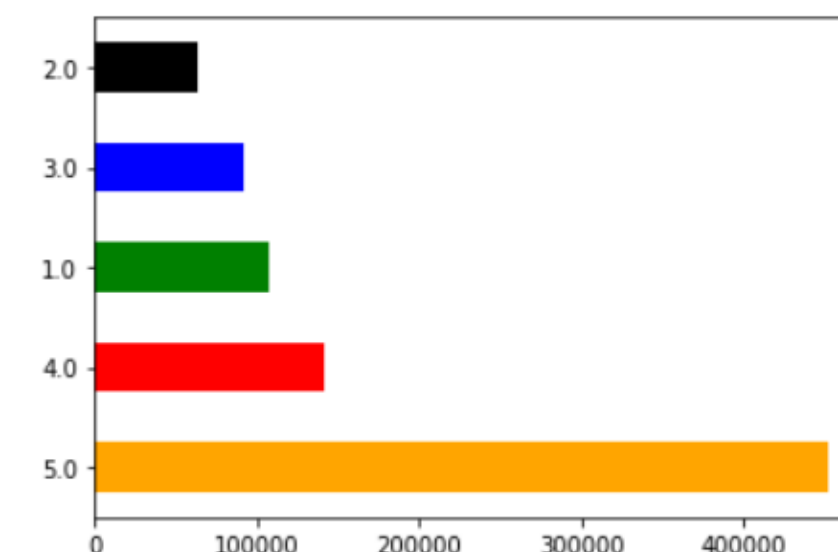
4

Câu có độ dài trên 55 là 587.912
và câu dài 55~100 là 381.352

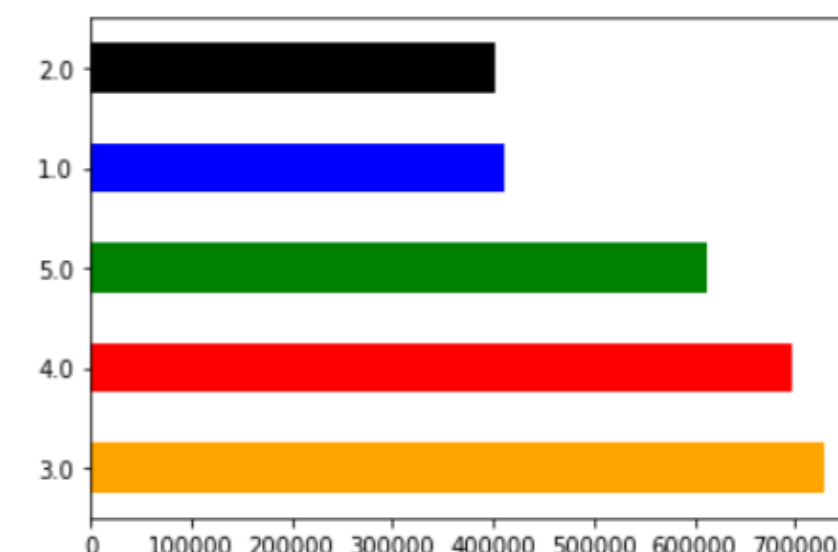
5

Số từ riêng biệt trong dữ liệu
train ~220k (có tính cả 'product')

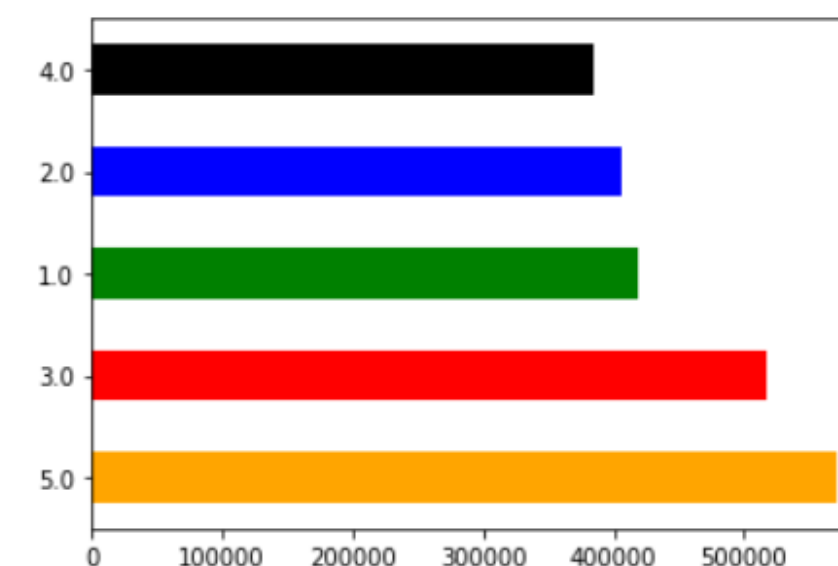
gốc



sau khi bổ sung

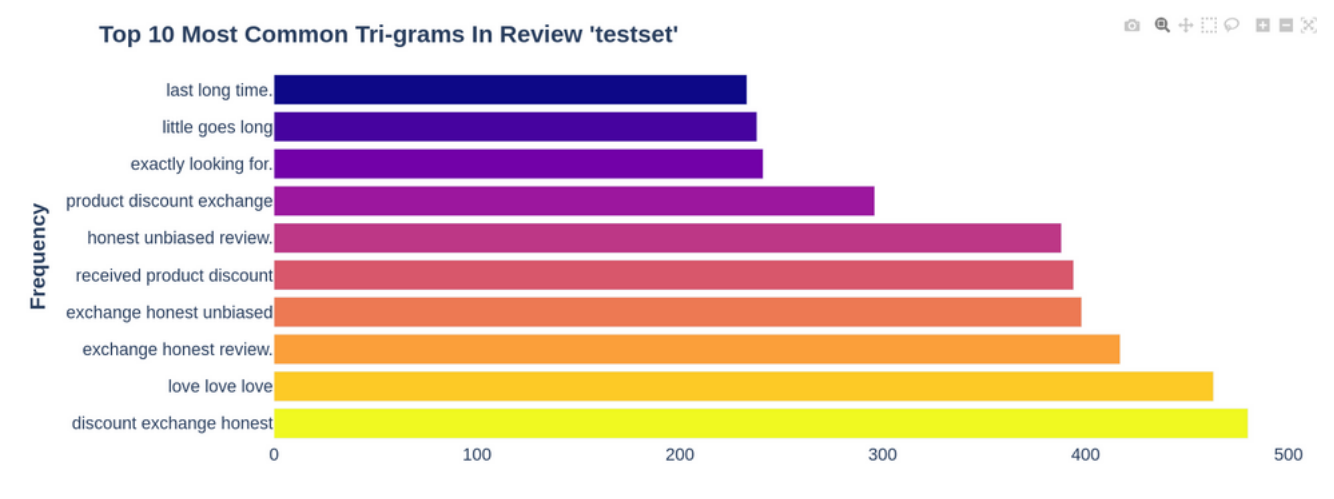
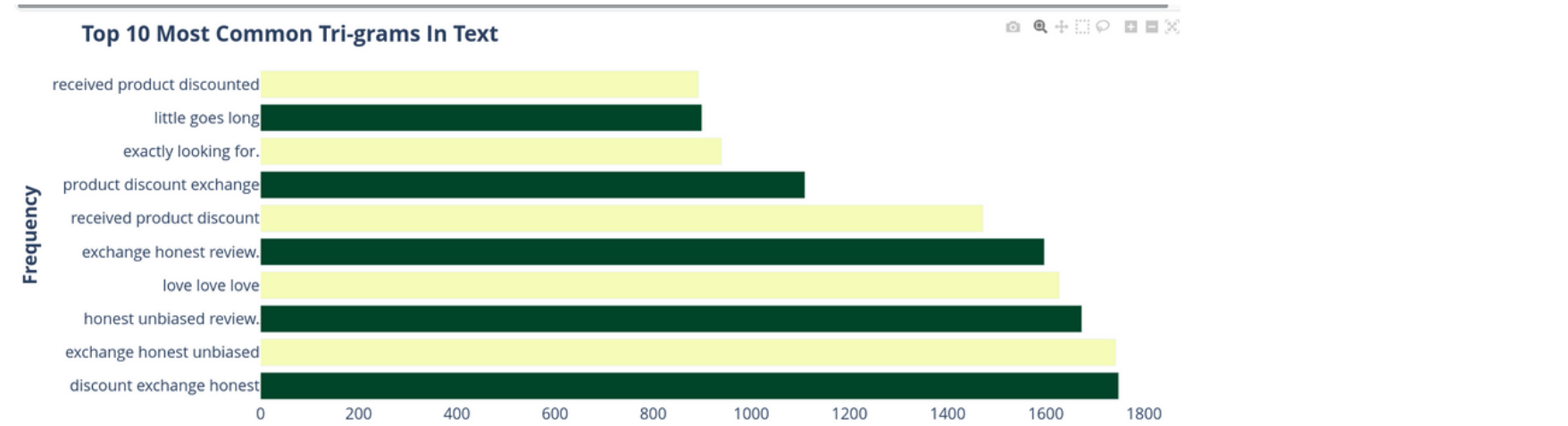
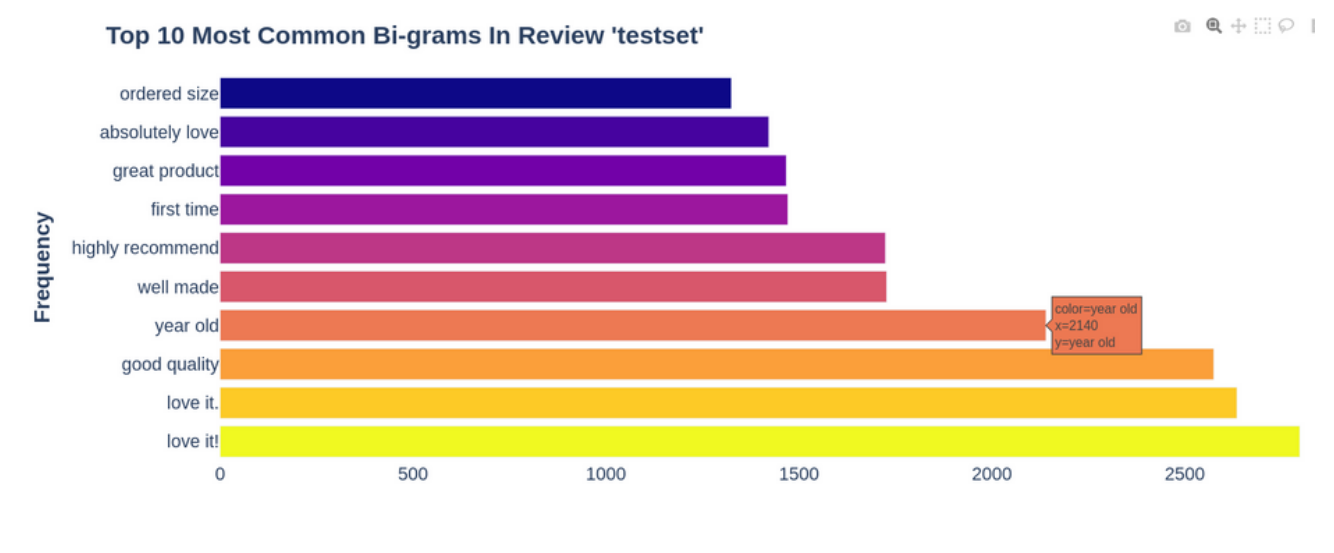
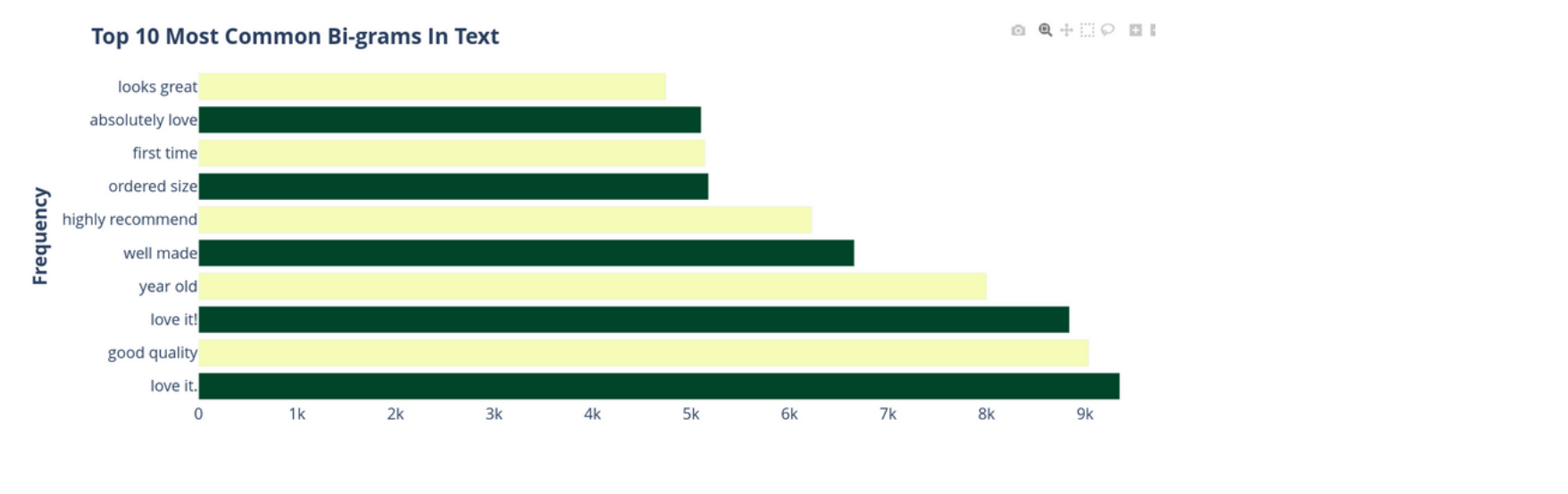
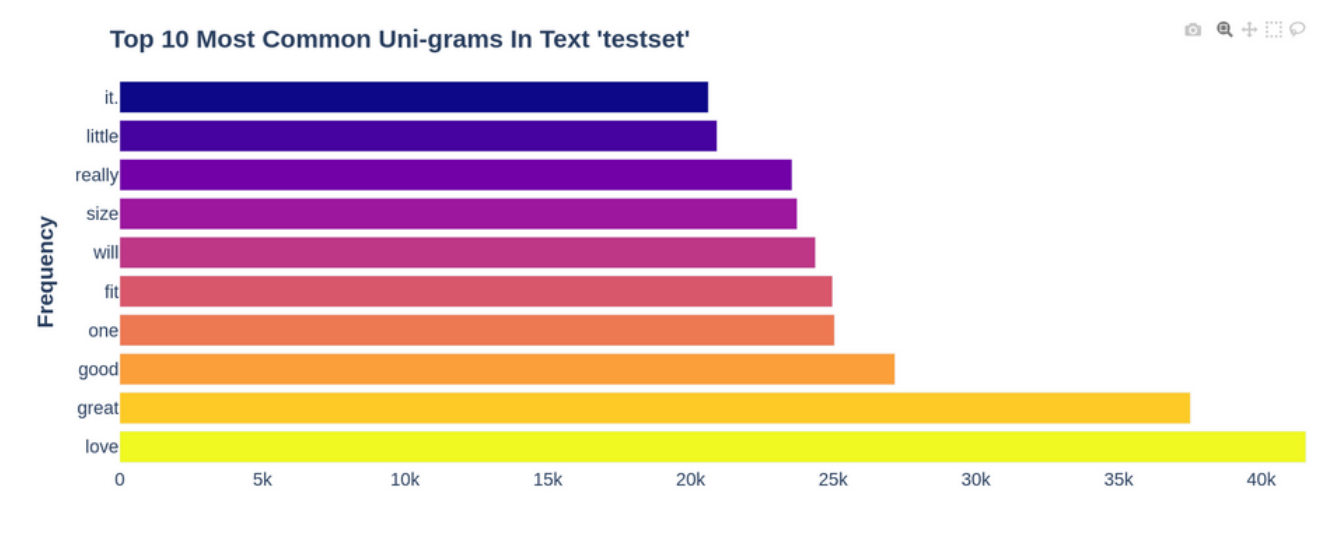
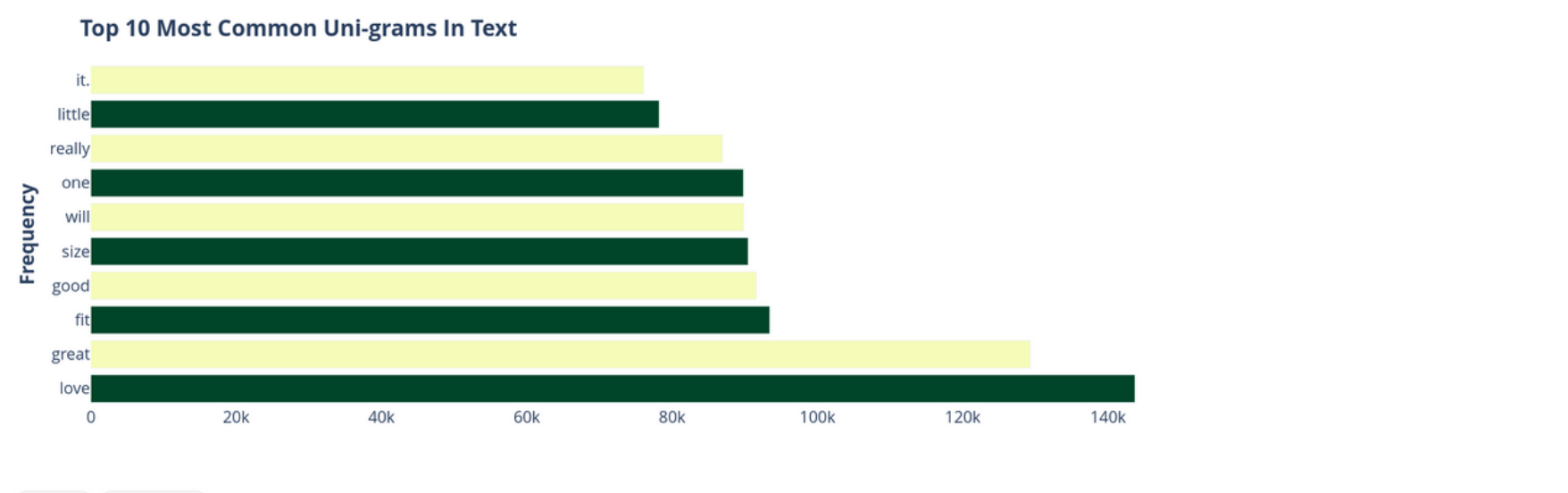


lấy ngẫu nhiên



Feature Engineering

top 10 uni, bi, tri-gram xuất hiện nhiều nhất



Triển khai thuật toán

Key takeaways

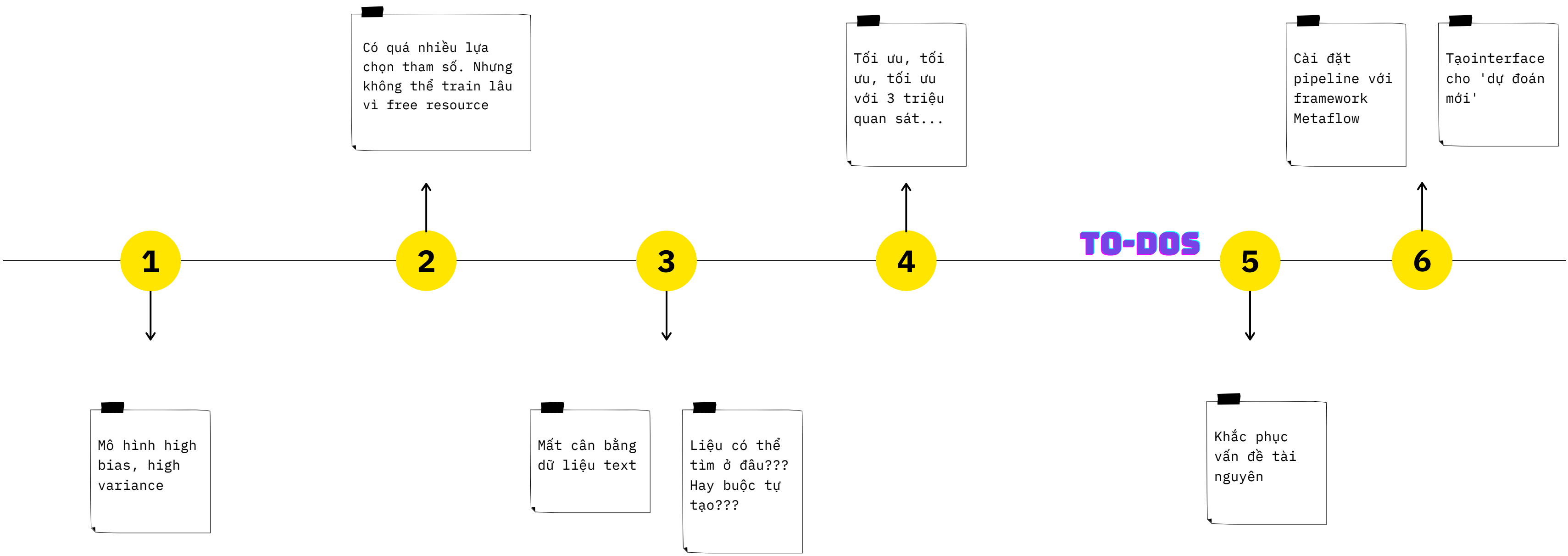
thuật toán	metrics
Đoán mò dựa trên lớp đa số	Classification accuracy: F1-Score, RMSE()
XGBoost	Speed: Training
Glove_840B_300d_2.2m	Disk: Tokenize, Model size
FastText_crawl_300d_2m	
DistilBert embedding	
Finetune DistilBert	

so sánh

thuật toán	classification_accuracy(RMSE)	speed(gpu P100)+memory	public_score (RMSE)
Đoán mò (tất cả 4)	3.057	⚡ ⚡	1.39928
XGBoost	3.006	⚡	1.73469
Glove_600B_50	0.2608	~2m5 params, 6 epochs, 3p50s/1epoch	0.95119
FastText_crawl_300d_2m	0.2207	~70m params, 9 epochs, 6p/1epoch	0.57324
DistilBert embedding	0.4158	~24m params, 10 epochs, (gpu K80) 9p/1epoch	0.57460
Finetune DistilBert	0.3895	~60 params, 2 epochs, 2h3p/1epoch	0.56306

Quá trình :-)

Thực hiện đồ án thế nào? Từng bước giải quyết vấn đề phát sinh.



4 Cải thiện

Đề án giải quyết vấn đề rất thực tế và thú vị. Mong muốn cải thiện nhiều hơn nữa!!!

Tối ưu

Cài đặt
End-to-
End
pipeline

Kết quả
tốt
nhất
với mô
hình
nhỏ
nhất có
thể

Bình
tĩnh,
phân
tích
vấn đề
hơn

Impleme
nt các
ý tưởng
nhanh,
hợp lý
hơn

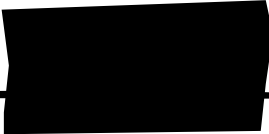
Đọc các
bài báo
khoa
học,
chứ
không
chỉ
dùng ở
sách và
blog...

xin cảm ơn!



tham khảo

Nhờ rất nhiều nguồn tham khảo, bản thân học được rất nhiều điều mới :-D

- 
- Practical Natural Language Processing
 - Natural Language Processing with Transformers
 - Real-World Natural Language Processing
 - <https://nijianmo.github.io/amazon/index.html#code>
 - Kaggle discussion
 - Medium blog
 - còn nữa...