Data Science Project

Bank App Sentiment Analysis

2022

Project Overview

Business Objective/Problem

- Apple Appstore & Google Play Store là hai chợ nổi tiếng nhất dành cho các ứng dụng di động. Apple và Android cũng là hai hệ điều hành phổ biến nhất hiện nay.
- Trên chợ người dùng có thể đưa ra các đánh giá về các app dành cho các nhà phát triển
- Các nhà phát triển app ngân hàng mong muốn có thể phân loại được các đánh giá này để từ đó cải thiện chất lượng của app
- Mục tiêu của dự án là phân loại được các đánh giá của các app ngân hàng tại Việt Nam

EDA

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18205 entries, 0 to 18204
Data columns (total 3 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   Bank          18205 non-null  object
 1   review_text   18157 non-null  object
 2   review_score  18205 non-null  int64 
dtypes: int64(1), object(2)
memory usage: 426.8+ KB
```

```
data = data.drop_duplicates()
data.shape
```

```
(17579, 3)
```

```
data.isnull().T.any().sum()
```

```
46
```

```
data.shape
```

```
(17533, 3)
```

```
len(data.Bank.unique())
```

```
29
```

- Dữ liệu thu thập trên Apple Appstore & Google Play Store
- Dữ liệu bị trùng/NaN/null
- Sau khi loại bỏ trùng/NaN, null, dữ liệu còn lại gồm 17.533 records (reviews) cho 29 app ngân hang tại Việt Nam.

EDA

- Hiển thị một vài dòng dữ liệu:

```
data.head()
```

	Bank	review_text	review_score
2	HDBank	Stupid bank app without help on how to generat...	1
3	HDBank	High fees. And money sometimes just disappear ...	1
4	HDBank	Unable to login by password and unable to logi...	4
5	HDBank	Whenever I choose transfer money option, the a...	3
6	HDBank	I can not run this app recently. It crashes al...	3

```
data.tail()
```

	Bank	review_text	review_score
18199	TPBank Mobile	App rất tốt, cảm ơn TPBank	5
18200	TPBank Mobile	Mình rất thích dùng app tpbank tiện lợi mà chu...	5
18202	TPBank Mobile	Giao diện đẹp, giao dịch nhanh , dễ dùng vote ...	5
18203	TPBank Mobile	Rất tiện và nhanh, lại không tốn phí.	5
18204	TPBank Mobile	App tuyệt vời, chuyển khoản nhanh. \nGiao diện...	5

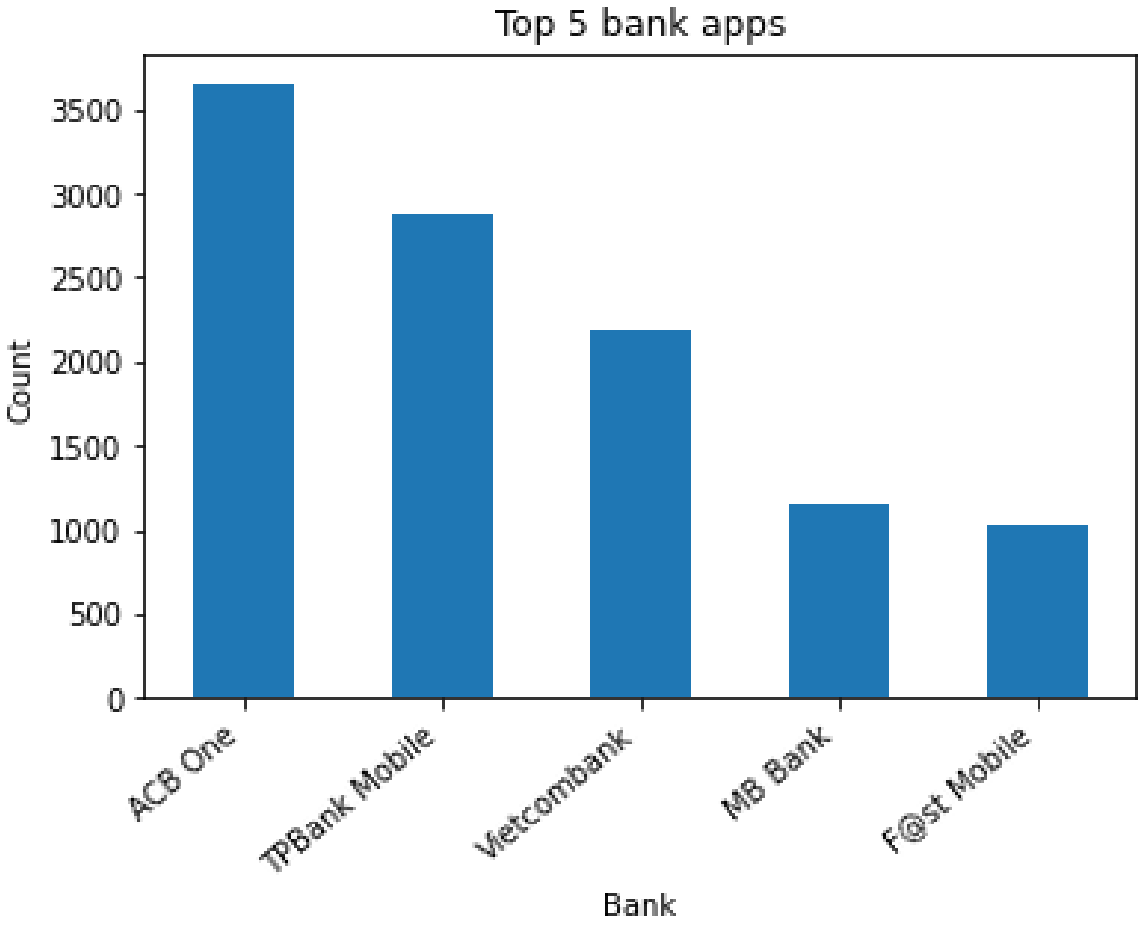
EDA

- Số lượng reviews các apps:

ACB One	3653	MSB mBank	198
TPBank Mobile	2887	OCTO by CIMB	158
Vietcombank	2190	SCB Mobile Banking	131
MB Bank	1143	SHB Mobile Banking	100
F@st Mobile	1029	TPBank QuickPay	93
VPBank NEO	980	BIZ MBBANK	76
Techcombank Mobile	847	TNEX - Ngân hàng số thế hệ mới	61
BIDV SmartBanking	804	AB Ditizen	55
Agribank E-Mobile Banking	728	PV Mobile Banking	47
VietinBank iPay	725	HDBank	45
SC Mobile Vietnam	387	Co-opBank Mobile Banking	9
MyVIB	339	Vietbank Digital	9
Sacombank Pay	315	Techcombank Business	8
Sacombank mBanking	283	VPBank NEOBiz	4
OCB OMNI - Digital Bank	229		

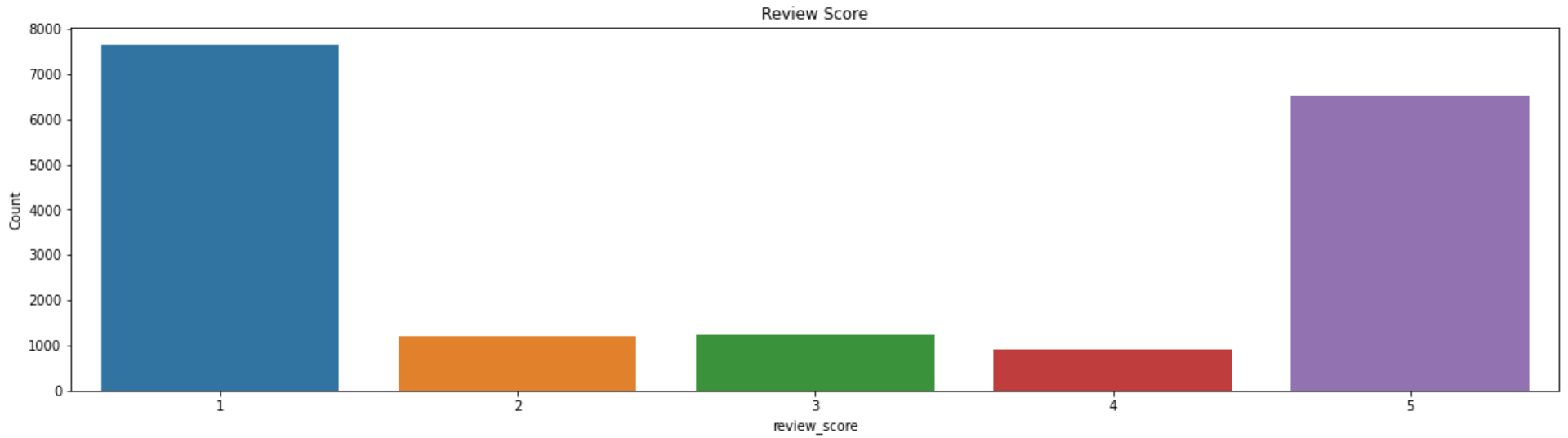
EDA

- Top 5 bank apps có nhiều reviews nhất:



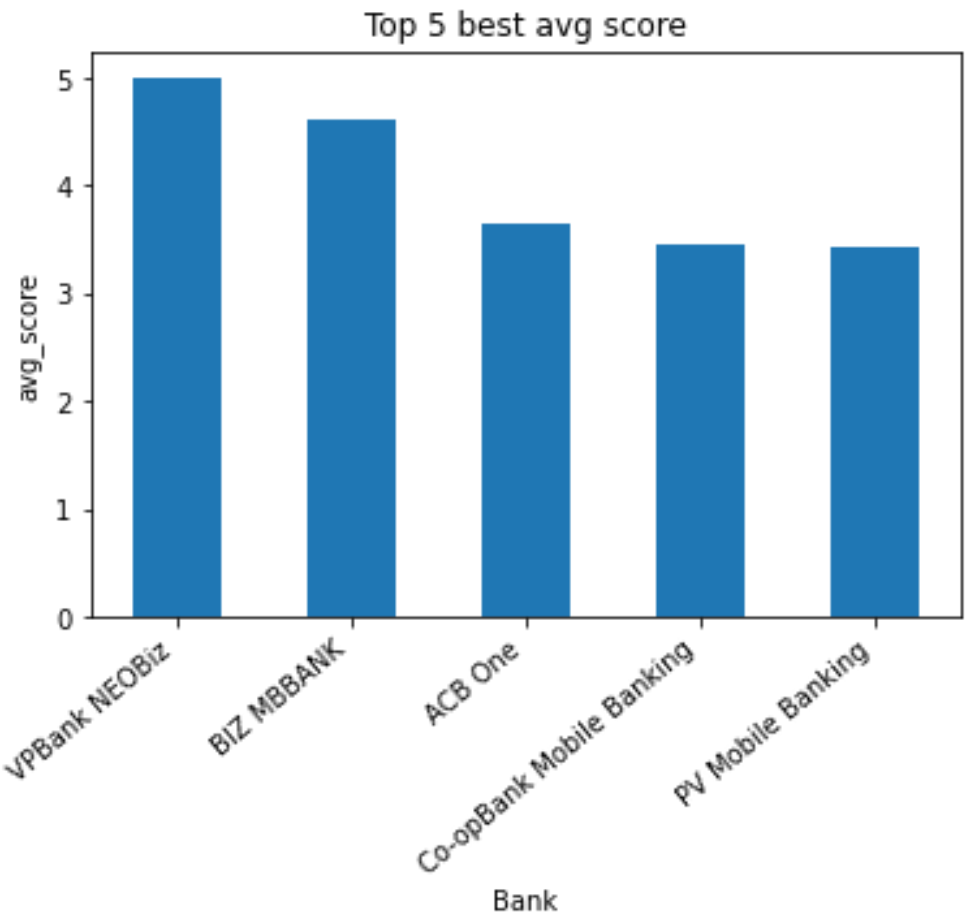
EDA

- Rating = 1 chiếm nhiều nhất với 7645 reviews, số lượng lớn thứ 2 là rating = 5 với 6524 reviews, các điểm rating từ 2-4 chiếm số lượng xấp xỉ nhau khoảng 1000 reviews



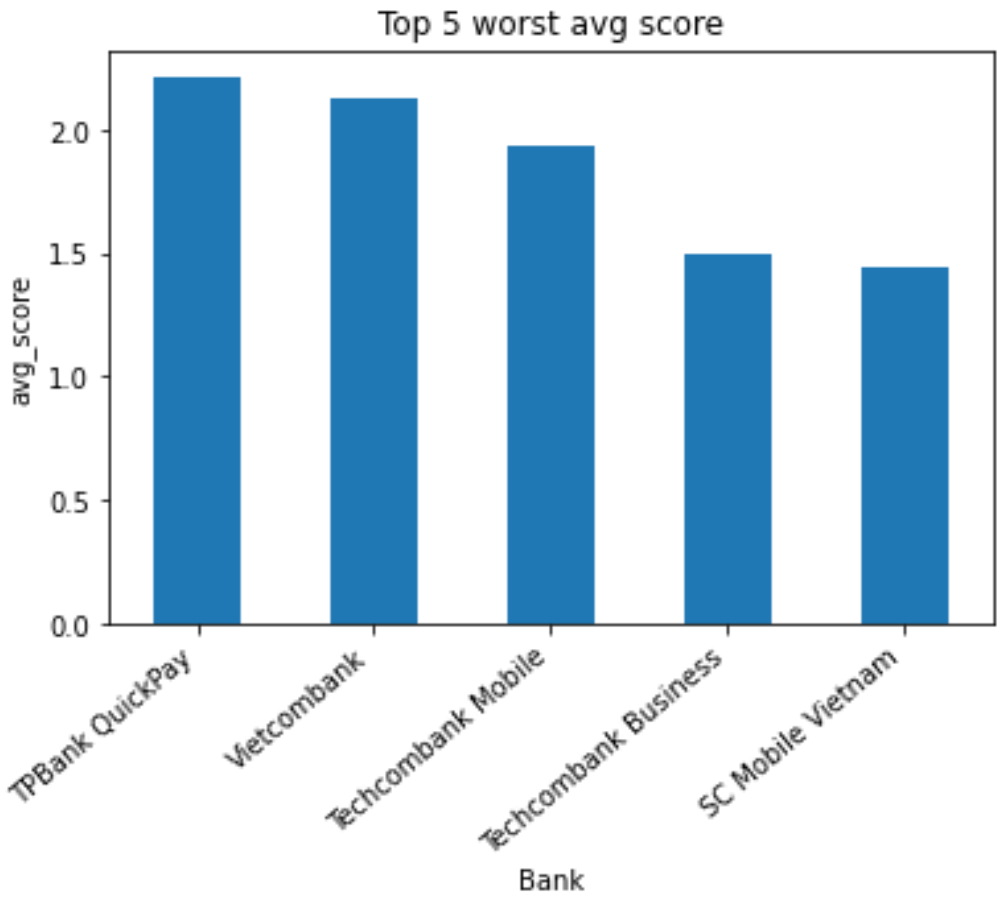
EDA

- Top 5 apps có điểm trung bình cao nhất:



EDA

- Top 5 apps có điểm trung bình thấp nhất:



Xử lý dữ liệu

- Hầu hết reviews sử dụng Tiếng Việt (14.837).
- Có nhiều reviews sử dụng Tiếng Anh (2.696).
- Tách làm 2 files dữ liệu để xây dựng model riêng
- Phân loại reviews của khách hàng theo 2 nhóm:
  - Target 0 (Dislike) với review_score < 4
  - Target 1 (Like) với review_score >= 4

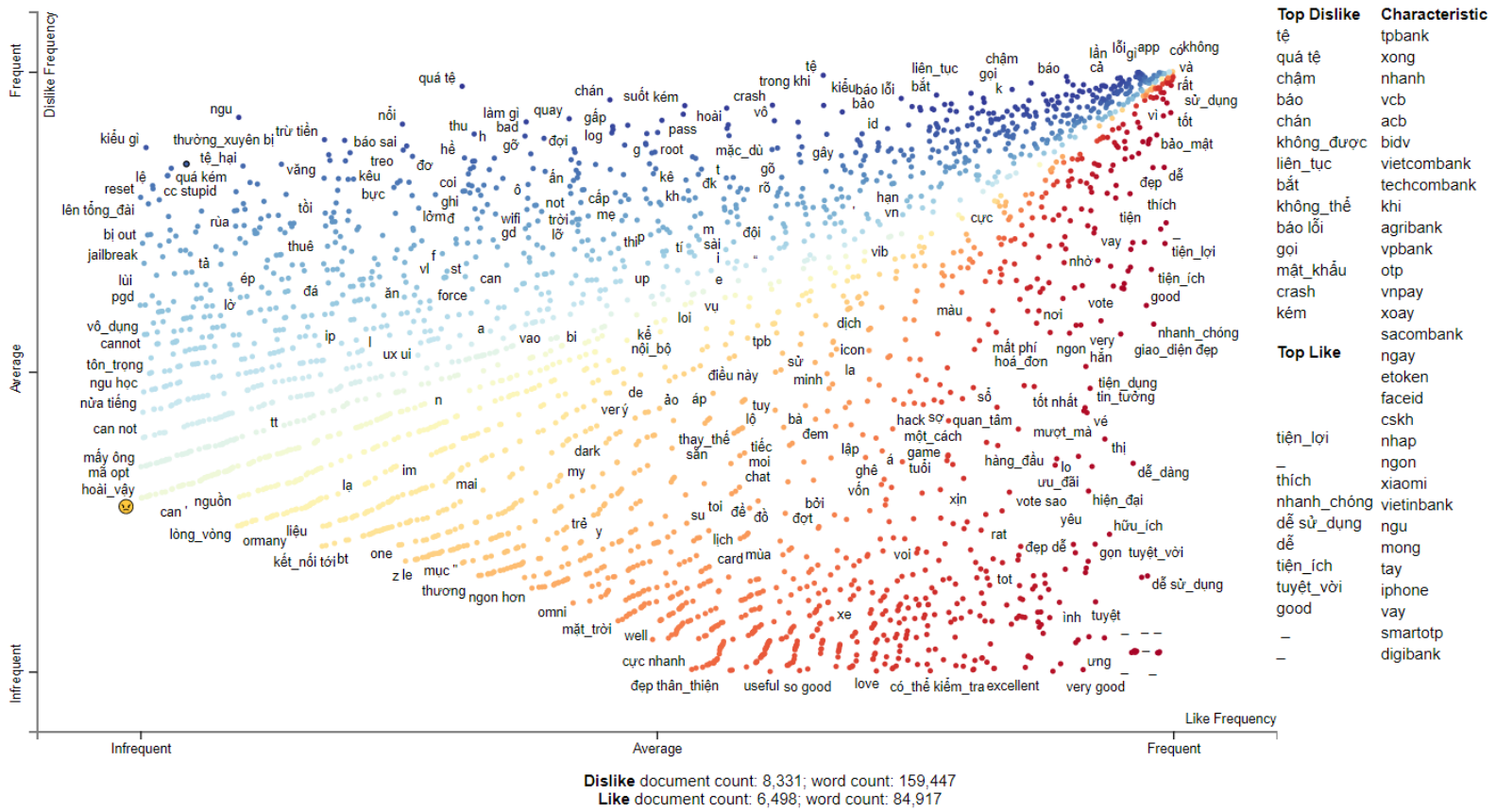
Xử lý dữ liệu Tiếng Việt

- Xử lý chuyển đổi emoticon, emoji, viết tắt/viết kiểu đặc biệt, dấu câu, … Tokenize review_text tạo column review_text_wt:

	review_text	review_text_wt
10	so far so good	so far so good
12	quá tệ liên tục bị lỗi kết nối tới server bị g...	quá tệ liên_tục bị lỗi kết_nối tới server bị g...
13	app của một ngân hàng có tham vọng mà để xảy r...	app của một ngân_hàng có tham_vọng mà để xảy r...
14	app ngân hàng thuộc top cùi nhất thị trg mỗi l...	app ngân_hàng thuộc top cùi nhất thị trg mỗi l...
15	mình dùng app để thanh toán vé máy bay và vé x...	mình dùng app để thanh_toán vé máy_bay và vé x...
16	sao vì dịch vụ chuyển tiền qua app chưa miễn p...	sao vì dịch_vụ chuyển tiền qua app chưa miễn_p...
17	đề nghị coi lại quy trình đặt lại mật khẩu mới...	đề_nghị coi lại quy_trình đặt lại mật_khẩu mới...
18	trải nghiệm tệ nhất với credit card của hd ban...	trải_nghiệm tệ nhất với credit card của hd ban...
19	ok	ok
20	rất tiện lợi	rất tiện_lợi

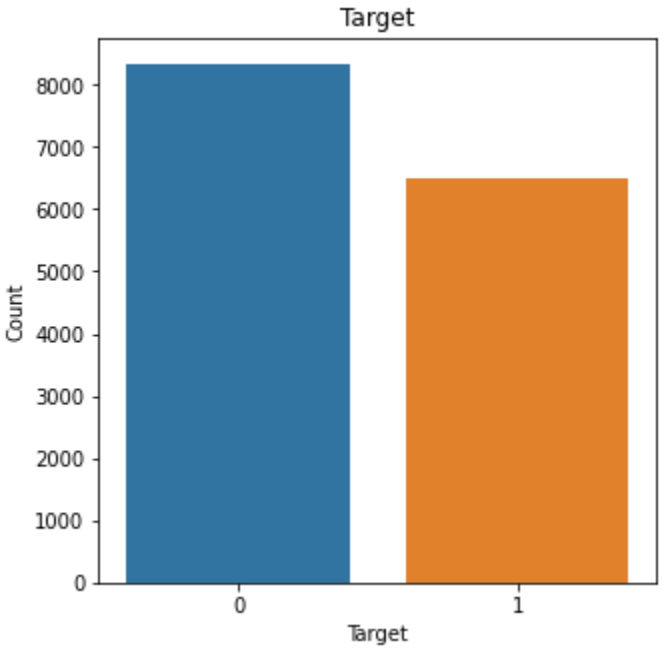
- Loại bỏ dữ liệu vô nghĩa. Dữ liệu còn lại 14831 dòng.

Spacy Tiếng Việt



- Class Dislike thường xuyên xuất hiện các từ: chậm, lỗi, tệ, tệ hại, quá_tệ, ngu, bị, chán, thường xuyên bị, quá kém, trừ tiền, …
- Class Like thường xuyên xuất hiện các từ: hữu ích, hiện đại, dễ dàng, tốt, bảo mật, dễ, đẹp, thích, tiện, tiện lợi, nhanh chóng, …

Vietnamese Sentiment Analysis – LR



- Dữ liệu giữa 2 target khá cân bằng.
- Sử dụng TfidfVectorizer kết hợp stopwords để chuyển text về dạng array.
- Sử dụng chi2 để lấy số lượng nhỏ các từ có mối quan hệ đáng kể với biến target, kết quả được 877 từ quan trọng nhất.
- Chia dữ liệu thành tập train, test với tỷ lệ 0.7:0.3

Vietnamese Sentiment Analysis – LR

- Đo hiệu quả của một số mô hình với dữ liệu.
- Lựa chọn Logistic Regression (LR) vì có Test Accuracy, F1 Score, AUC-ROC khá cao. SVC() cũng có Test Accuracy, F1 Score, AUC-ROC xấp xỉ LR nhưng thời gian thực thi quá lâu.

	Model	Test Accuracy	F1 Score	AUC-ROC	Time Fit
0	LogisticRegression	0.901798	0.886817	0.950009	0.625306
1	MultinomialNB	0.888989	0.864880	0.949610	0.049541
2	BernoulliNB	0.879326	0.865716	0.941513	0.114061
3	KNeighborsClassifier	0.799551	0.780512	0.884926	0.009263
4	DecisionTreeClassifier	0.844270	0.828253	0.855731	4.473930
5	RandomForestClassifier	0.886292	0.870456	0.943042	9.194668
6	XGBClassifier	0.835506	0.789534	0.908207	18.316882
7	SVC	0.901798	0.885692	0.945822	223.414538

Vietnamese Sentiment Analysis – LR

```
lr.score(x_train, y_train)
```

```
0.9115692129852615
```

```
lr.score(x_test, y_test)
```

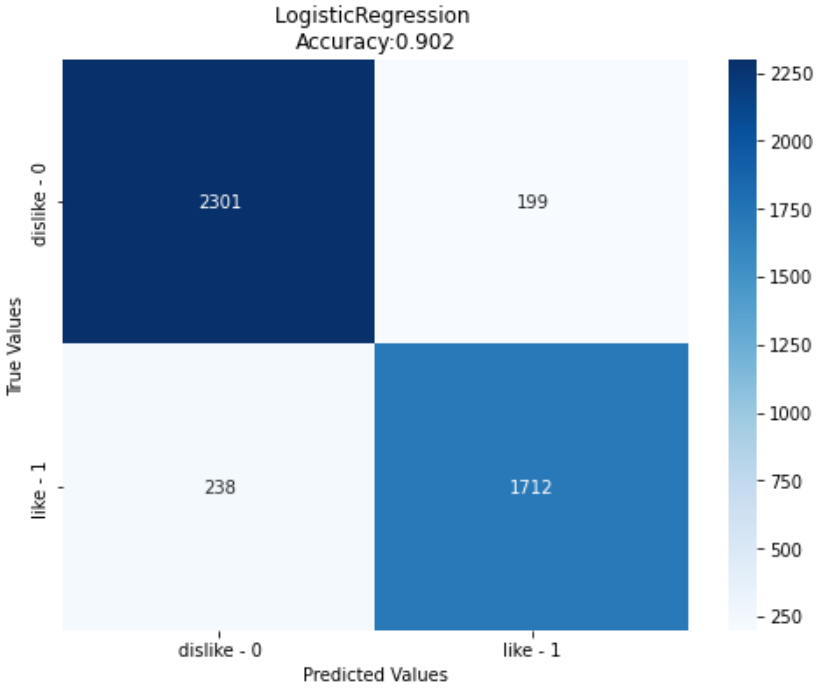
```
0.9017977528089888
```

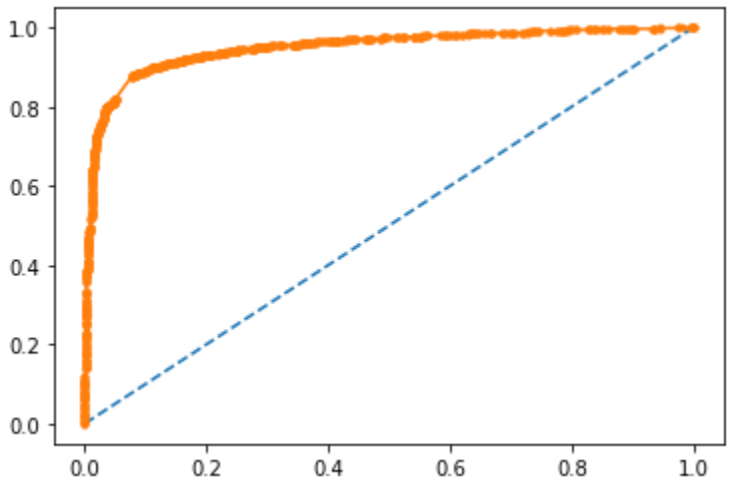
```
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.91	0.92	0.91	2500
1	0.90	0.88	0.89	1950
accuracy			0.90	4450
macro avg	0.90	0.90	0.90	4450
weighted avg	0.90	0.90	0.90	4450

```
cm = confusion_matrix(y_test, y_pred)
cm
```

```
array([[2301,  199],
       [ 238, 1712]])
```





Vietnamese Sentiment Analysis – LR

- LogisticRegression() mặc định cho kết quả khá tốt.
- Áp dụng tuning parameter cho LogisticRegression nhưng kết quả không tốt hơn.
- Lưu model để tiến hành dự đoán mới

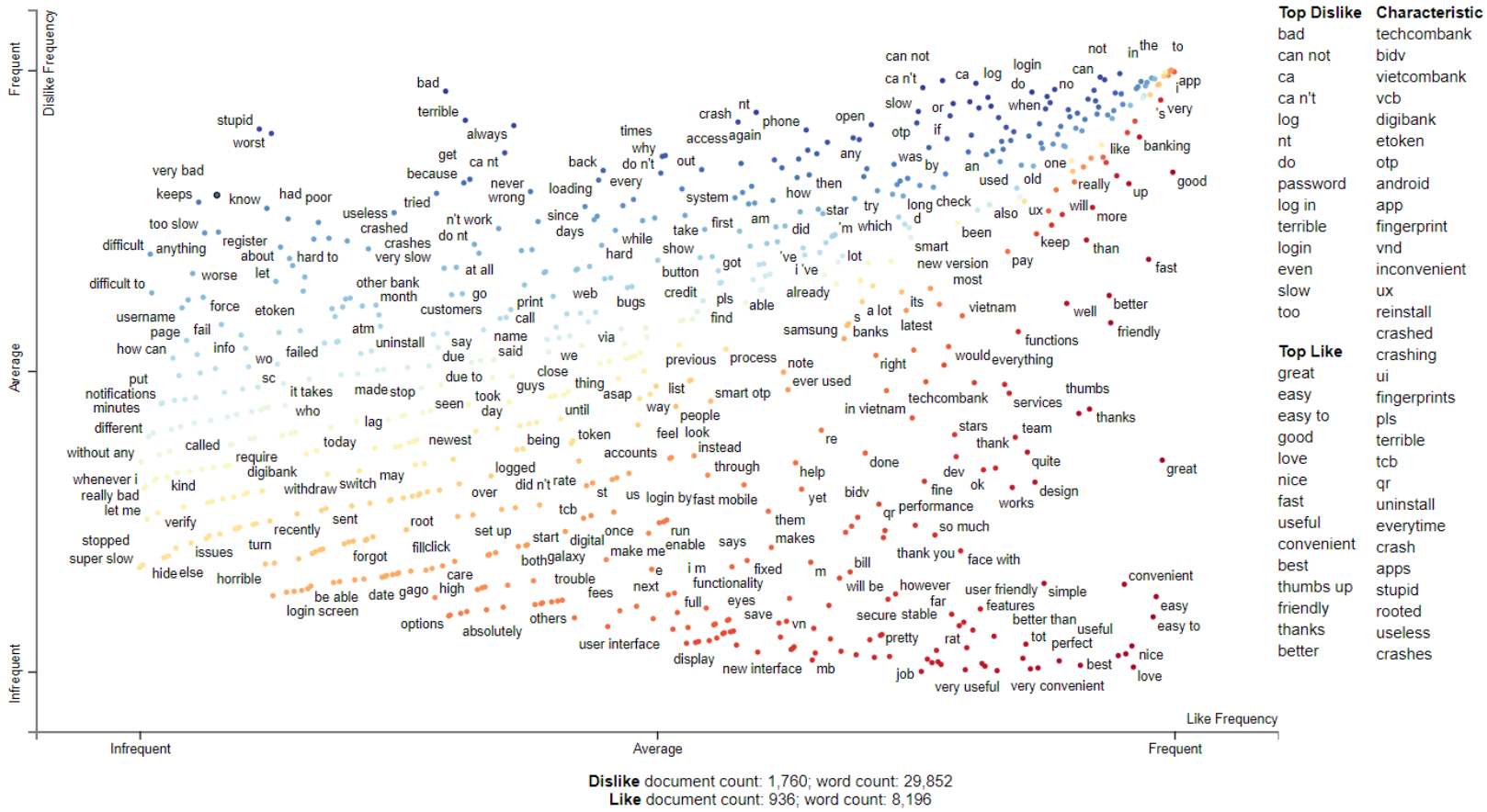
Xử lý dữ liệu Tiếng Anh

- Xử lý chuyển đổi emoticon, emoji, dấu câu, …

	Bank	review_text	review_score	english	target
0	HDBank	stupid bank app without help on how to generat...	1	1	0
1	HDBank	high fees and money sometimes just disappear f...	1	1	0
2	HDBank	unable to login by password and unable to logi...	4	1	1
3	HDBank	whenever i choose transfer money option the ap...	3	1	0
4	HDBank	i can not run this app recently it crashes all...	3	1	0
5	HDBank	can login the internet banking using laptop bu...	1	1	0
6	HDBank	nice and easy to use	5	1	1
7	HDBank	the app is much better now it works well	4	1	1
8	HDBank	app login crash	1	1	0
9	HDBank	dislike these colour	3	1	0

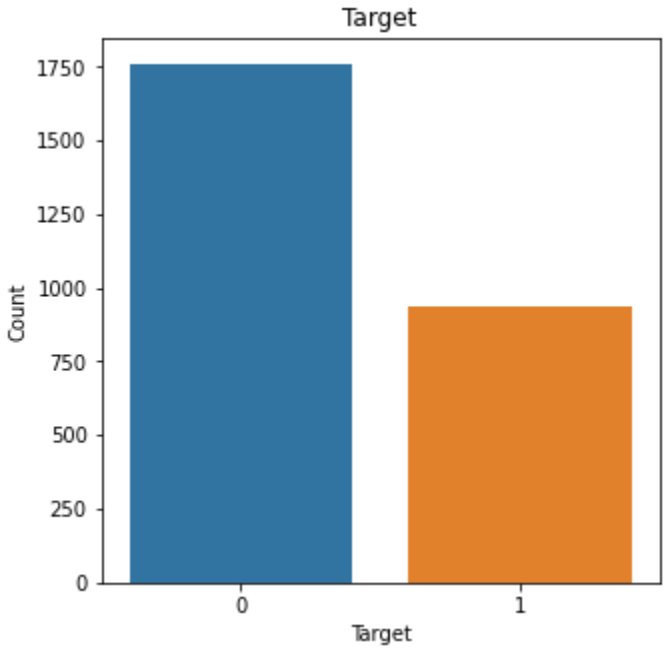
- Dữ liệu có 2.696 dòng.

Spacy Tiếng Anh



- Class Dislike thường xuyên xuất hiện các từ: bad, stupid, worst, very bad, too slow, difficult, terrible, crash, can not, login, …
- Class Like thường xuyên xuất hiện các từ: great, easy, good, love, nice, convenient, simple, usefull, perfect, fast, …

English Sentiment Analysis – LR



- Dữ liệu giữa 2 target mất cân bằng, có thể cần resample.
- Sử dụng TfidfVectorizer kết hợp stopwords để chuyển text về dạng array.
- Sử dụng chi2 để lấy số lượng nhỏ các từ có mối quan hệ đáng kể với biến target, kết quả được 181 từ quan trọng nhất.
- Chia dữ liệu thành tập train, test với tỷ lệ 0.7:0.3

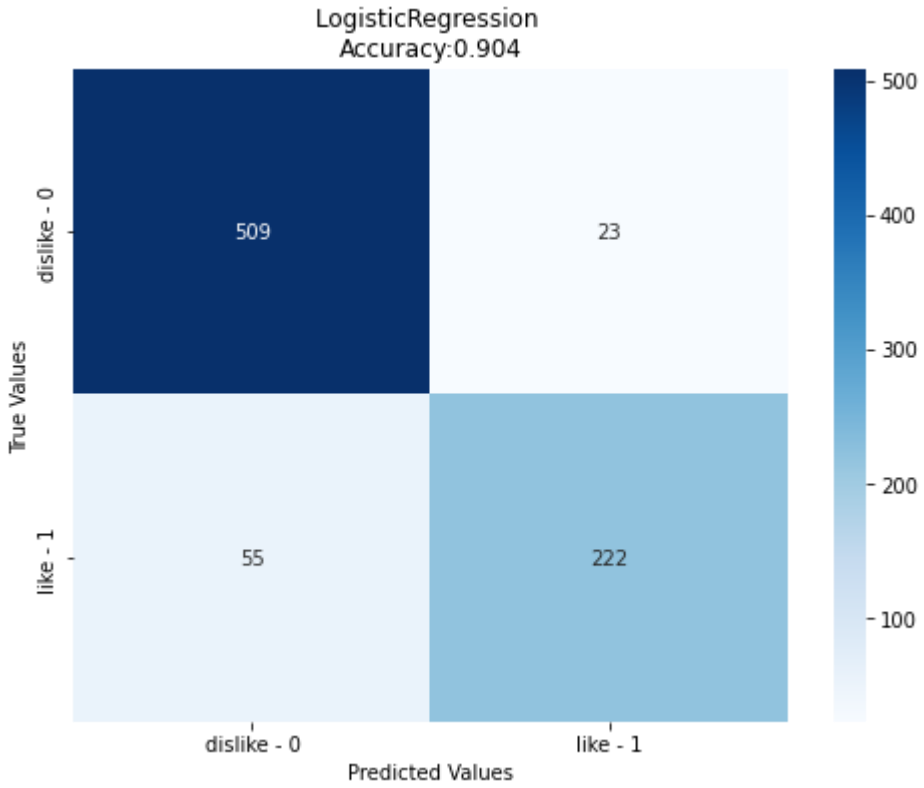
English Sentiment Analysis – LR

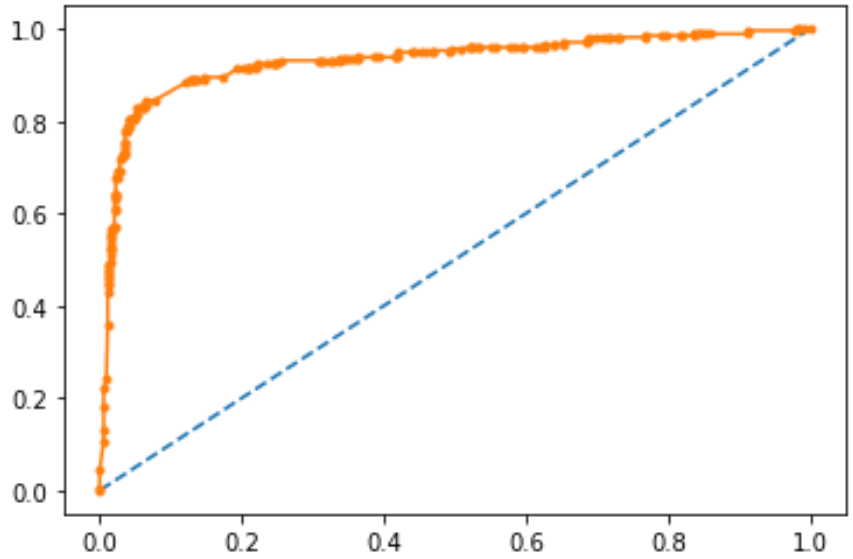
- Đo hiệu quả của một số mô hình với dữ liệu.
- Lựa chọn Logistic Regression (LR) và MultinomialNB vì có Test Accuracy, F1 Score, AUC-ROC cao nhất.

	Model	Test Accuracy	F1 Score	AUC-ROC	Time Fit
0	LogisticRegression	0.903585	0.850575	0.930678	0.037667
1	MultinomialNB	0.906057	0.856061	0.927856	0.002649
2	BernoulliNB	0.898640	0.848148	0.926658	0.005863
3	KNeighborsClassifier	0.871446	0.810219	0.915637	0.001064
4	DecisionTreeClassifier	0.873918	0.817204	0.879821	0.121992
5	RandomForestClassifier	0.891224	0.836431	0.930668	0.581580
6	XGBClassifier	0.875155	0.795960	0.918162	0.908142
7	SVC	0.899876	0.849162	0.920968	1.297471

English Sentiment Analysis – LR

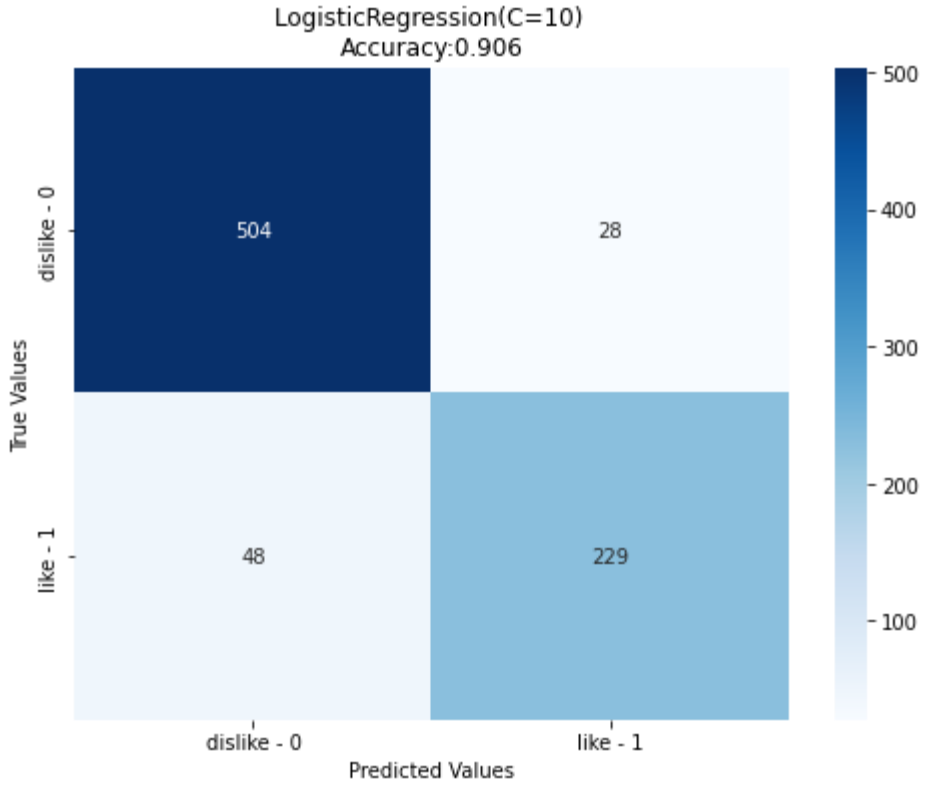
- LogisticRegression: Model dự đoán rất tốt ở class dislike.

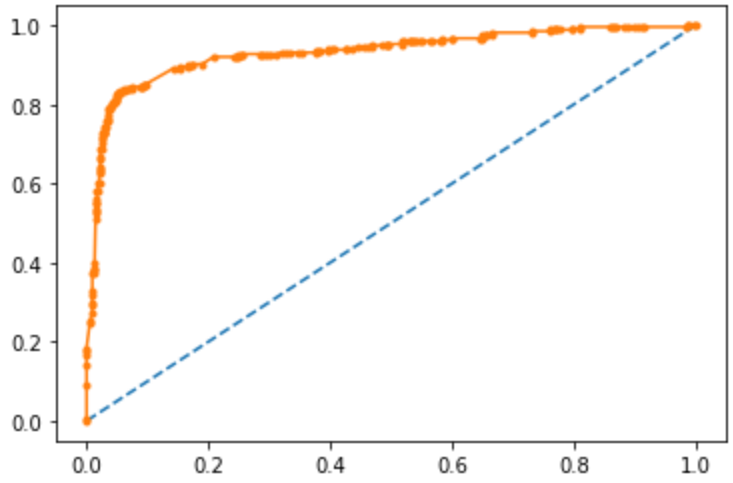




English Sentiment Analysis – LR

- Tuning parameter với Logistic Regression
- LogisticRegression(C=10): Kết quả có cải thiện nhỏ so với LogisticRegression mặc định.

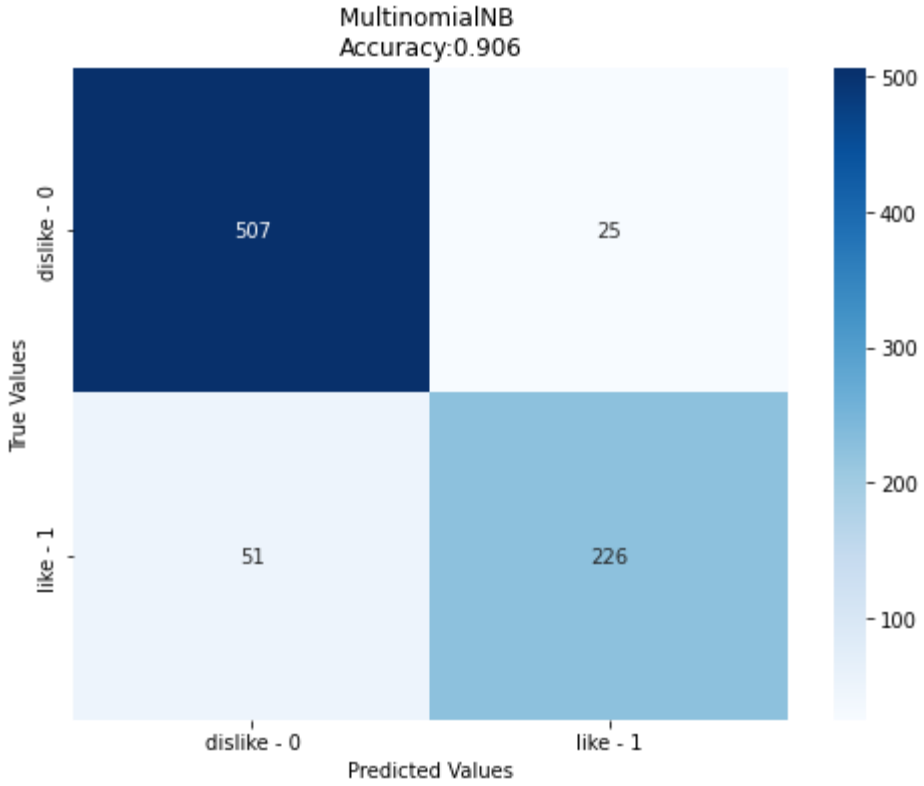


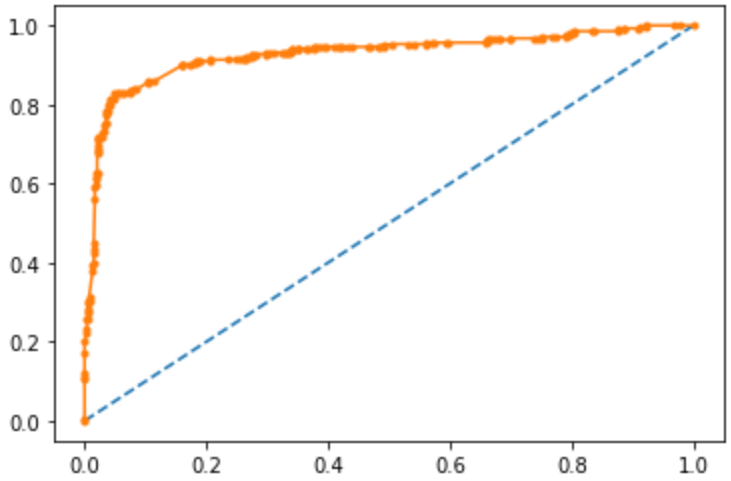


	precision	recall	f1-score	support
0	0.91	0.95	0.93	532
1	0.89	0.83	0.86	277
accuracy			0.91	809
macro avg	0.90	0.89	0.89	809
weighted avg	0.91	0.91	0.91	809

English Sentiment Analysis – LR

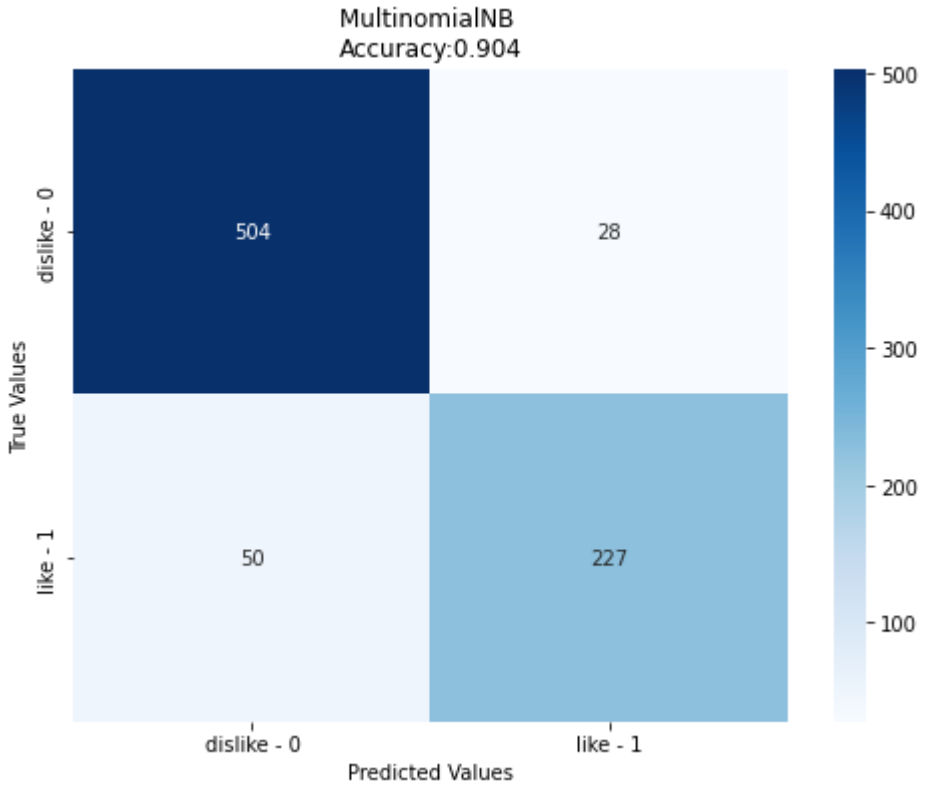
- MultinomialNB: Model dự đoán khá tốt.

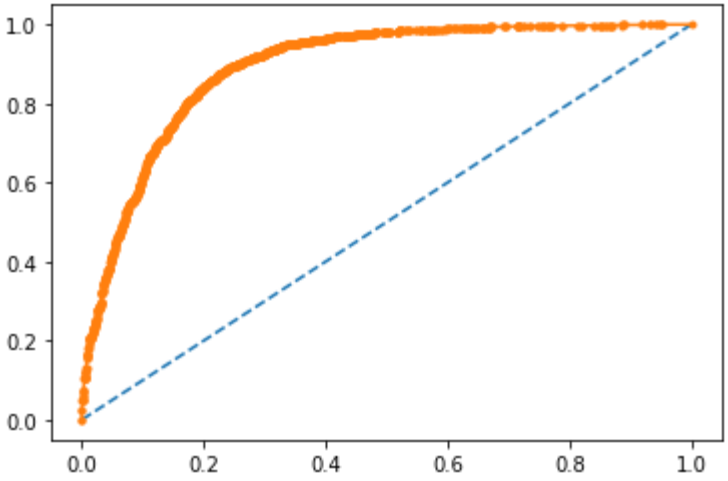




English Sentiment Analysis – LR

- Tuning parameter với MultinomialNB
- MultinomialNB(alpha = 0.5): Model dự đoán khá tốt





English Sentiment Analysis – LR

```
data_train_new.groupby('target').review_text.count()
```

```
target
0    1228
1    1228
```

- Đo hiệu quả của một số mô hình với dữ liệu đã resample.
- Dựa vào Test Accuracy, F1 Score, AUC-ROC, Time Fit, em lựa chọn LogisticRegression để build model:

	Model	Test Accuracy	F1 Score	AUC-ROC	Time Fit
0	LogisticRegression	0.894932	0.842884	0.931449	0.099306
1	MultinomialNB	0.885043	0.834813	0.929070	0.010628
2	BernoulliNB	0.867738	0.822554	0.925246	0.012331
3	KNeighborsClassifier	0.844252	0.782759	0.881019	0.001168
4	DecisionTreeClassifier	0.859085	0.794224	0.871563	0.127579
5	RandomForestClassifier	0.881335	0.823529	0.926777	0.919052
6	XGBClassifier	0.876391	0.801587	0.915889	1.912579
7	SVC	0.893696	0.840741	0.921965	2.497465

English Sentiment Analysis – LR

- Kết quả không tốt hơn khi build model với dữ liệu gốc.

```
lr_re.score(x_train, y_train)
```

```
0.9153094462540716
```

```
lr_re.score(x_test, y_test)
```

```
0.8949320148331273
```

```
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.91      0.93      0.92       532
           1       0.86      0.82      0.84       277

    accuracy                           0.89       809
   macro avg       0.89      0.88      0.88       809
weighted avg       0.89      0.89      0.89       809
```

```
cm = confusion_matrix(y_test, y_pred)
cm
```

```
array([[496,  36],
       [ 49, 228]])
```

English Sentiment Analysis – LR

- Lựa chọn model LogisticRegression (C = 10) được huấn luyện bởi dữ liệu gốc để dự đoán vì có kết quả dự đoán khá tốt
- Lưu model để tiến hành dự đoán mới

Prediction

```
predict_reviews('Cứ đăng nhập là bắt cập nhật,trong khi đó đã cập nhật rồi???')
```

```
'Không thích/ Dislike'
```

```
predict_reviews("Thích nhất gửi tiền tiết kiệm online")
```

```
'Thích/ Like'
```

```
predict_reviews("Love the app. It's similar to the banking app I have in Australia, which is very useful and easy to navigate
```

```
'Thích/ Like'
```

```
predict_reviews("Terrible app for a bank as it can only check balance... nothing else is works((( Its a pity that a bank with
```

```
'Không thích/ Dislike'
```

```
predict_reviews("good service, easy to use UX")
```

```
'Thích/ Like'
```

```
predict_reviews("Moved from iOs 12 to iOs 14, then the app always crash when opened, I had to come back to Sms OTP! Very bad
```

```
'Không thích/ Dislike'
```

Prediction

- Link: https://bank-apps-streamlit.herokuapp.com/