SENTIMENT ANALYSIS

😊 Positive
😐 Neutral
😢 Negative

TRONG, NGUYENTHANH

Visit our application:
Sentiment Application

1. Business Understanding

Vision

- Based on the history of customer reviews provided earlier
- Data is collected from customer comments and reviews on Sendo and Tiki.

Target

- Building a predictive model to help sellers quickly understand customer feedback on their products or services (positive, negative, or neutral),
- This assists sellers in grasping their business situation, understanding customer opinions, thereby enabling them to improve their services and products.

2. Data Understanding

- The data is provided in the file Sendo_reviews.csv.
- In which, each sample contains the following information: product_id, customer_id, full_name, created_time, rating, content

Scrape more data from Tiki.

9576 user comments data from the e-commerce site Tiki along with Sendo data to build the model.

Scrap data of products and comments.

Scrap data of products to create an interface similar to Tiki, allowing for searching and predicting the outcomes of direct comments on products.

Allow adding comments.

The interface allows adding comments on products and predicts forecasts.

3. Data Preparation

Data Consolidate

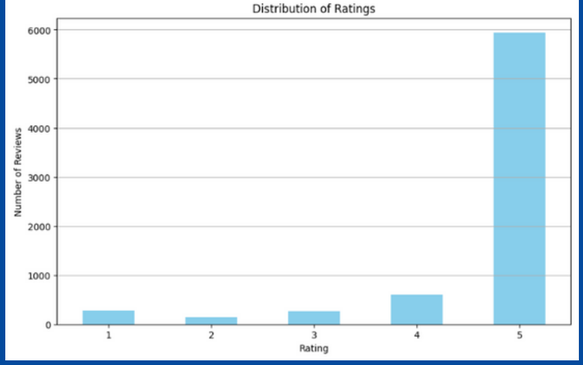
Data imbalance

=> Scrap more data from tiki

	rating	content
0	5	NaN
1	4	Shop phục vụ khá tốt.
2	3	Sản phẩm gần giống mô tả
3	5	NaN
4	5	NaN
...	...	...
7219	4	NaN
7220	4	NaN
7221	4	NaN
7222	4	NaN
7223	5	Đẹp như mô tả.

7224 rows × 2 columns

```
Distribution of ratings:
1      274
2      141
3      263
4      604
5     5942
Name: rating, dtype: int64
```



Data Sendo

	content	rating
0	Sản phẩm được giao nhanh chóng, kỹ lưỡng. Cây ...	5
1	Cây lau nhà lỗi ko thể dùng được. Lắp vào ko q...	1
2	Hàng méo cong vênh, chốt xoay gãy, mỏng yếu, y...	1
3	tốt	5
4	tốt dễ dàng sử dụng	5
...	...	...
9571	Hang đong goi ok , sản phẩm chưa sủ dụng , cho...	4
9572	Giao hàng nhanh hơn dự kiến, đúng mẫu. Pin ok ...	5
9573	Tikinow giao quá nhanh 😊😊😊	5
9574	Giao hàng nhanh, pin chính hãng nguyên đai ngu...	5
9575	mọi người nên mua vì shop hổ trợ rất nhiệt tình	5

9576 rows × 2 columns

```
5    5942
4     604
1     274
3     263
2     141
Name: rating, dtype: int64
```

```
5    82.253599
4     8.361019
1     3.792913
3     3.640642
2     1.951827
Name: rating, dtype:
```

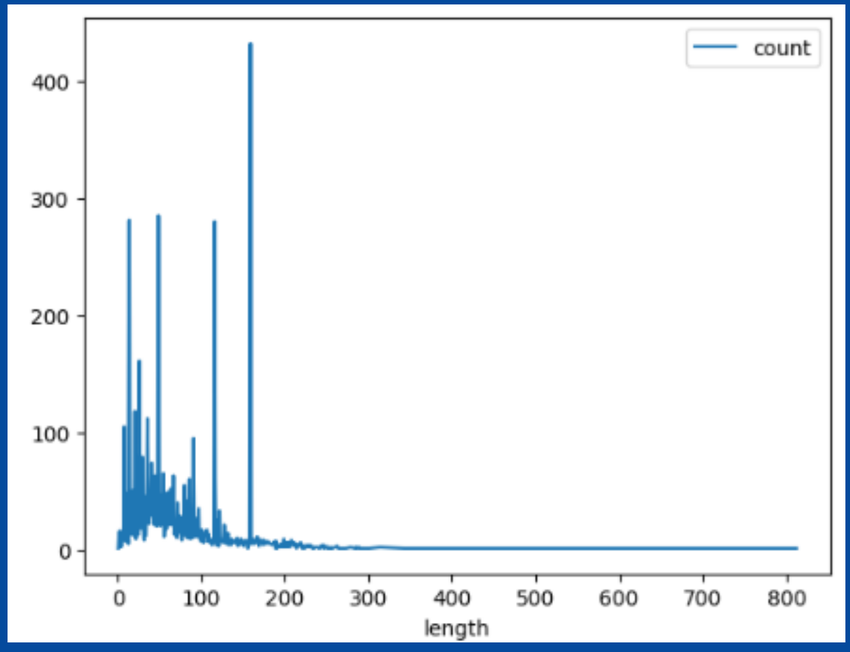
Data Tiki

Data

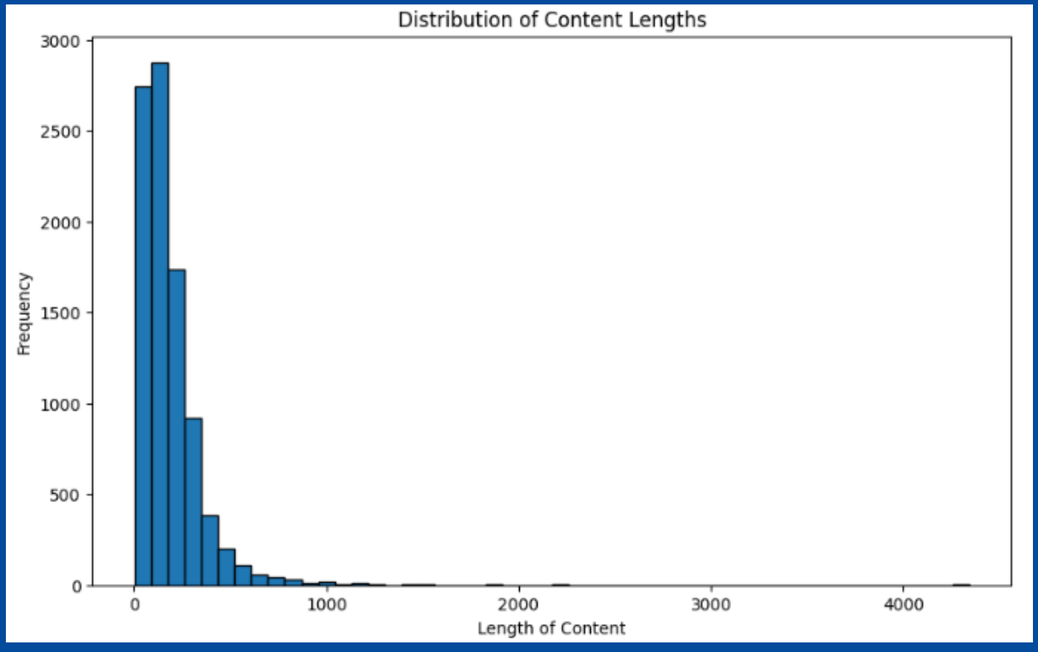
3. Data Preparation

Data Consolidate

Data Cleaning



Data Sendo



Data Tiki

Data

3. Data Preparation

Data Consolidate

Data Cleaning

Data Reduction

Remove unnecessary columns, only retain content and rating

```
# Remove duplicate
data = data.drop_duplicates()

# Column keep
columns_to_keep = ['content', 'rating']

# DF wtih column keep
data = data[columns_to_keep]
data_tiki = data_tiki[columns_to_keep]

#Append tiki data and sendo data
data = data.append(data_tiki, ignore_index=True)
```

	content	rating
0	NaN	5
1	Shop phục vụ khá tốt.	4
2	Sản phẩm gần giống mô tả	3
3	NaN	5
4	NaN	5
...	...	...
16795	Hang đong goi ok , sản phẩm chưa sủ dụng , cho...	4
16796	Giao hàng nhanh hơn dự kiến, đúng mẫu. Pin ok ...	5
16797	Tikinow giao quá nhanh 😊 😊 😊	5
16798	Giao hàng nhanh, pin chính hãng nguyên đai ngu...	5
16799	mọi người nên mua vì shop hỗ trợ rất nhiệt tình	5

16800 rows × 2 columns

3. Data Preparation

Data Consolidate

Data Cleaning

Data Reduction

Remove NaN

```
data = data.dropna(subset=['content']).copy()
data = data.dropna(subset=['rating']).copy()
```

3. Data Preparation

Data Consolidate

Data Cleaning

Data Reduction

Data Tranformation

```
data = data.copy()
data['sentiment'] = data['rating'].apply(lambda x: 'positive' if x > 3 else ('neutral' if x == 3 else 'negative'))
```

	content	rating	sentiment
1	Shop phục vụ khá tốt.	4	positive
2	Sản phẩm gần giống mô tả	3	neutral
5	Giao hàng nhanh bất ngờ.	5	positive
6	Sản phẩm/dịch vụ giống mô tả.	4	positive
7	Sản phẩm/dịch vụ giống mô tả và tốt hơn mong đ...	5	positive
...	...	...	...
16795	Hang đong goi ok , sản phẩm chưa sủ dụng , cho...	4	positive
16796	Giao hàng nhanh hơn dự kiến, đúng mẫu. Pin ok ...	5	positive
16797	Tikinow giao quá nhanh 😊😊😊	5	positive
16798	Giao hàng nhanh, pin chính hãng nguyên đai ngu...	5	positive
16799	mọi người nên mua vì shop hổ trợ rất nhiệt tình	5	positive

14490 rows × 3 columns

Based on the rating, divide into three groups

😍 **Positive**

😐 **Neutral**

😥 **Negative**

3. Data Preparation

Data Consolidate | **Data Cleaning** | **Data Reduction** | **Data Tranformation**

process_text

- Convert text to lowercase
- Remove special characters
- Convert emojis and teencode to words
- Remove punctuation and numbers
- Remove unwanted words from the wrong_list

emoji_dict

🤍	yêu
👍	thích
😍	cười
😊	cười
😘	cười
😂	cười
😁	cười
💧	thích
⭐	thích
🌟	thích
💜	yêu

teen_dict

ctrai	con trai
khôg	không
bme	bố mẹ
cta	chúng ta
mih	mình
mqh	mối quan hệ
cgai	con gái
nhữg	những
mng	mọi người
svtn	sinh viên tình nguyện
qtam	quan tâm
thươg	thương
qtâm	quan tâm
chug	chung
trườg	trường
thoy	thôi
đki	đăng ký

wrong_lst

aud
se
gd
tz
voi
hoa
cca
bay
son
vtd
tau
xd

3. Data Preparation

Data Consolidate | **Data Cleaning** | **Data Reduction** | **Data Tranformation**

covert_unicode

- Conver Unicode

translate_text

- Translate English to Vietnamese

Date features

- Add feature: count of negative, neutral, positive word/emoji

english	vnese
district	huyện
pebble	đá cuội
wedding	lễ cưới
comment	bình luận
thought	nghĩ
disrespected	không tôn trọng
pool	hồ bơi
ngoc	ngọc
muck	vấy bẩn
class	lớp học
hundred	hàng trăm
stranger	người lạ
get	được
much	nhiều
level	cấp độ

negative_words
kém
tệ
đau
xấu
bị
rè
ồn
buồn
rối
thô
lâu
sai
hư

neutral_words
chấp nhận được
trung bình
bình thường
tạm ổn
trung lập
có thể
không nổi bật
đủ ổn
đủ tốt
có thể chấp nhận
bình thường
thường xuyên
tương đối

positive_words
thích
tốt
xuất sắc
đúng
tuyệt vời
tuyệt hảo
đẹp
ổn
hài lòng
ưng ý
hoàn hảo
chất lượng
thú vị

```
negative_emojis = [
    "😑", "😔", "🙁", "☹️", "😟",
    "😢", "😭", "😩", "😣", "😖",
    "😠", "😡", "🤬", "😤", "😰",
```

```
positive_emojis = [
    "😄", "😃", "😀", "😁", "😆",
    "😅", "🤣", "😂", "🙂", "🙃",
    "😉", "😊", "😇", "🥰", "😍",
```

3. Data Preparation

Data Consolidate **Data Cleaning** **Data Reduction** **Data Tranformation**

After Pre-Processing

	content	rating	sentiment	new_content	neutral_words_count	negative_words_count	positive_words_count	negation_words_count	positive_emoji_count	negative_emoji_count
1	Shop phục vụ khá tốt.	4	positive	phục_vụ tốt .	0	0	1	0	0	0
2	Sản phẩm gần giống mô tả	3	neutral	gần giống .	0	0	0	0	0	0
5	Giao hàng nhanh bất ngờ.	5	positive	nhanh_chóng .	0	0	0	0	0	0
6	Sản phẩm/dịch vụ giống mô tả.	4	positive	giống .	0	0	0	0	0	0
7	Sản phẩm/dịch vụ giống mô tả và tốt hơn mong đ...	5	positive	giống tốt hơn . chất_lượng tốt . . tư_vấn hướn...	0	1	3	0	0	0
8	Shop tư vấn, hướng dẫn cực kỳ nhiệt tình.	5	positive	tư_vấn hướng_dẫn cực_kỳ nhiệt_tình .	0	1	0	0	0	0
10	Sản phẩm tốt vượt mong đợi	5	positive	tốt vượt .	0	0	1	0	0	0
11	Sản phẩm đúng mô tả	4	positive	.	0	0	1	0	0	0
12	Hài lòng	5	positive	hài_lòng .	0	0	0	0	0	0

3. Data Preparation

Data Consolidate

Data Cleaning

Data Reduction

Data Tranformation

Function word tokenizer

Use Work_Tokenizer from library Underthesea
Use ViTokenizer from pyvi to precess with Vietnamese

remove_stopword

```
1                                          phục_vụ tốt
2                                            gần giống
5                                          nhanh_chóng
6                                                giống
7     giống tốt hơn  chất_lượng tốt   tư_vấn hướng_d...
8                  tư_vấn hướng_dẫn cực_kỳ nhiệt_tình
10                                           tốt vượt
11
12                                           hài_lòng
13    nhanh_chóng_chóng  đóng_gói kĩ  giá_cả tốt  ủn...
14    chuyên_nghiệp thân_thiện  đóng_gói kỹ_lưỡng  c...
16                                       giống tốt hơn
18                                           tốt vượt
19                                           tốt vượt
20
21                                           tốt vượt
22                                  chất_lượng tuyệt_vời
24                                           tốt vượt
25    giống tốt hơn  chất_lượng tốt   lưu_ý khách_hàng
26    giống tốt hơn  chất_lượng tốt   tư_vấn hướng_d...
Name: new_content, dtype: object
```

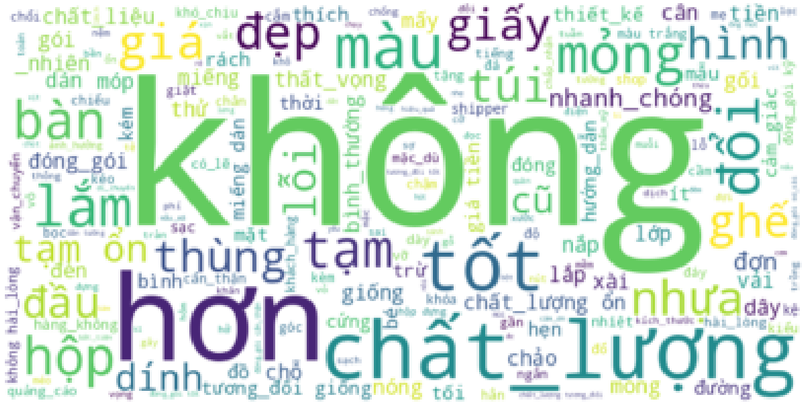
3. Data Preparation

Data Consolidate | **Data Cleaning** | **Data Reduction** | **Data Tranformation**

Workcloud

Word Cloud - Positive Sentiment

Word Cloud - Neutral Sentiment

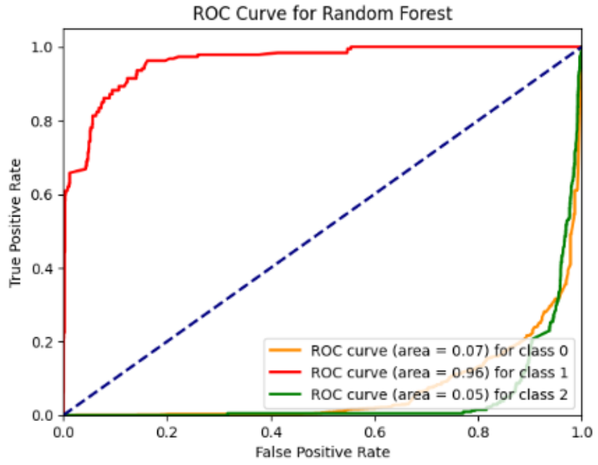


Word Cloud - Negative Sentiment

4. Modeling & Evaluation/Analyze

	Model	Accuracy	F1 Score	Positive Recall	Neutral Recall	Negative Recall	Positive Precision	Neutral Precision	Negative Precision	Confusion Matrix
0	Logistic Regression	0.869550	0.842605	0.976967	0.054054	0.631188	0.884691	0.555556	0.770393	[[2248, 6, 47], [146, 10, 29], [147, 2, 255]]
1	Decision Tree	0.924221	0.924529	0.957410	0.772973	0.804455	0.960331	0.725888	0.814536	[[2203, 39, 59], [27, 143, 15], [64, 15, 325]]
2	Random Forest	0.941522	0.940038	0.980443	0.745946	0.809406	0.955123	0.878981	0.881402	[[2256, 15, 30], [33, 138, 14], [73, 4, 327]]
3	Gradient Boosting	0.898270	0.884387	0.983485	0.264865	0.702970	0.910664	0.890909	0.811429	[[2263, 6, 32], [102, 49, 34], [120, 0, 284]]
4	LinearSVC	0.921453	0.916891	0.977836	0.513514	0.787129	0.939457	0.840708	0.832461	[[2250, 13, 38], [64, 95, 26], [81, 5, 318]]
5	K-Nearest Neighbors	0.804152	0.756753	0.971751	0.140541	0.153465	0.823270	0.433333	0.543860	[[2236, 26, 39], [146, 26, 13], [334, 8, 62]]
6	Multinomial Naive Bayes	0.829758	0.777298	0.995654	0.016216	0.257426	0.826479	0.750000	0.912281	[[2291, 1, 9], [181, 3, 1], [300, 0, 104]]

- The training and prediction time for the model is insignificant. The model choice with the highest accuracy is **Random Forest.**



Try adding the following columns:

- positive_words_count
- neutral_words_count
- negative_words_count
- positive_emoji_count
- negative_emoji_count

See if the model becomes more effective?

4. Modeling & Evaluation/Analyze

Before

	Model	Accuracy	F1 Score	Positive Recall	Neutral Recall	Negative Recall	Positive Precision	Neutral Precision	Negative Precision	Confusion Matrix
0	Logistic Regression	0.869550	0.842605	0.976967	0.054054	0.631188	0.884691	0.555556	0.770393	[[2248, 6, 47], [146, 10, 29], [147, 2, 255]]
1	Decision Tree	0.924221	0.924529	0.957410	0.772973	0.804455	0.960331	0.725888	0.814536	[[2203, 39, 59], [27, 143, 15], [64, 15, 325]]
2	Random Forest	0.941522	0.940038	0.980443	0.745946	0.809406	0.955123	0.878981	0.881402	[[2256, 15, 30], [33, 138, 14], [73, 4, 327]]
3	Gradient Boosting	0.898270	0.884387	0.983485	0.264865	0.702970	0.910664	0.890909	0.811429	[[2263, 6, 32], [102, 49, 34], [120, 0, 284]]
4	LinearSVC	0.921453	0.916891	0.977836	0.513514	0.787129	0.939457	0.840708	0.832461	[[2250, 13, 38], [64, 95, 26], [81, 5, 318]]
5	K-Nearest Neighbors	0.804152	0.756753	0.971751	0.140541	0.153465	0.823270	0.433333	0.543860	[[2236, 26, 39], [146, 26, 13], [334, 8, 62]]
6	Multinomial Naive Bayes	0.829758	0.777298	0.995654	0.016216	0.257426	0.826479	0.750000	0.912281	[[2291, 1, 9], [181, 3, 1], [300, 0, 104]]

After

	Model	Accuracy	F1 Score	Positive Recall	Neutral Recall	Negative Recall	Positive Precision	Neutral Precision	Negative Precision	Confusion Matrix
0	Logistic Regression	0.890657	0.870779	0.979140	0.135135	0.732673	0.904819	0.781250	0.804348	[[2253, 6, 42], [130, 25, 30], [107, 1, 296]]
1	Decision Tree	0.929066	0.930169	0.955671	0.816216	0.829208	0.965744	0.683258	0.854592	[[2199, 55, 47], [24, 151, 10], [54, 15, 335]]
2	Random Forest	0.951903	0.949920	0.993916	0.745946	0.806931	0.949751	0.938776	0.973134	[[2287, 9, 5], [43, 138, 4], [78, 0, 326]]
3	Gradient Boosting	0.947405	0.944719	0.989135	0.627027	0.856436	0.954298	0.950820	0.903394	[[2276, 6, 19], [51, 116, 18], [58, 0, 346]]
4	LinearSVC	0.923183	0.918105	0.977836	0.475676	0.816832	0.941029	0.822430	0.841837	[[2250, 11, 40], [75, 88, 22], [66, 8, 330]]
5	K-Nearest Neighbors	0.824221	0.817433	0.916993	0.227027	0.569307	0.893689	0.350000	0.562347	[[2110, 55, 136], [100, 42, 43], [151, 23, 230]]
6	Multinomial Naive Bayes	0.829758	0.808836	0.923946	0.021622	0.663366	0.894781	0.500000	0.529644	[[2126, 4, 171], [114, 4, 67], [136, 0, 268]]

4. Modeling & Evaluation/Analyze

	Model	Accuracy	F1 Score	Positive Recall	Neutral Recall	Negative Recall	Positive Precision	Neutral Precision	Negative Precision	Confusion Matrix
0	Logistic Regression	0.890657	0.870779	0.979140	0.135135	0.732673	0.904819	0.781250	0.804348	[[2253, 6, 42], [130, 25, 30], [107, 1, 296]]
1	Decision Tree	0.929066	0.930169	0.955671	0.816216	0.829208	0.965744	0.683258	0.854592	[[2199, 55, 47], [24, 151, 10], [54, 15, 335]]
2	Random Forest	0.951903	0.949920	0.993916	0.745946	0.806931	0.949751	0.938776	0.973134	[[2287, 9, 5], [43, 138, 4], [78, 0, 326]]
3	Gradient Boosting	0.947405	0.944719	0.989135	0.627027	0.856436	0.954298	0.950820	0.903394	[[2276, 6, 19], [51, 116, 18], [58, 0, 346]]
4	LinearSVC	0.923183	0.918105	0.977836	0.475676	0.816832	0.941029	0.822430	0.841837	[[2250, 11, 40], [75, 88, 22], [66, 8, 330]]
5	K-Nearest Neighbors	0.824221	0.817433	0.916993	0.227027	0.569307	0.893689	0.350000	0.562347	[[2110, 55, 136], [100, 42, 43], [151, 23, 230]]
6	Multinomial Naive Bayes	0.829758	0.808836	0.923946	0.021622	0.663366	0.894781	0.500000	0.529644	[[2126, 4, 171], [114, 4, 67], [136, 0, 268]]

Decided to opt for the Random Forest model as it offers better balance and is capable of accurately predicting all three groups: Positive, Neutral, and Negative.

Comment:

- Best Choice for Overall Performance: Based on the accuracy and F1 Score, we can observe that both the "Random Forest" and "Gradient Boosting" models deliver high performance. Both models have high accuracy and F1 Score, as well as good predictive performance for "positive," "neutral," and "negative" sentiments. Given the feedback, the "Decision Tree" model could potentially be the best choice as it provides balanced performance across all sentiment categories and has a high ROC AUC for all three categories.
- Best Choice for Predicting "Neutral" and "Negative": The "Decision Tree" model also emerges as the best choice when prioritizing the prediction of "neutral" and "negative" cases, as it exhibits the best performance for both of these sentiment categories, with a Neutral Recall of 0.816216 and a Negative Recall of 0.829208.

THANK YOU

FOR YOUR ATTENTION

SENTIMENT ANALYSIS

TRONG, NGUYENTHANH

Visit our application:
Sentiment Application