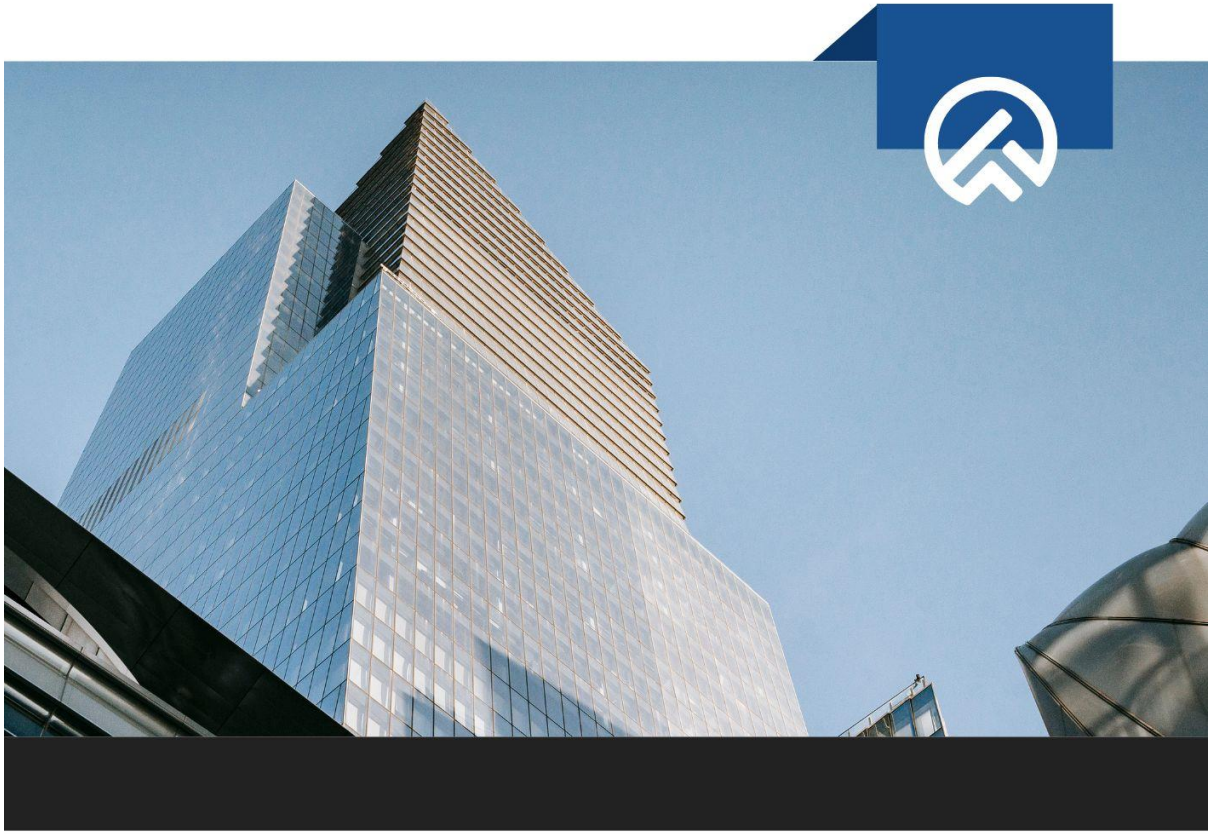


SENTIMENT ANALYSIS REPORT

Contents

CHAPTER 1: INTRODUCTION

1.1. Overview

There are many comments and programs that visitors watch every day on the FPT Play application, however, most of them are not collected for analysis. So, this project will analyze this pile of data and draw insights so we can understand a lot more about customers and their experiences when using services by Machine Learning and AI

1.2. Benefits of analyzing sentiment from customers

- ***Understand customers better:***
- Sentiment analysis helps businesses collect and better understand customers' opinions and feelings toward products, services, and brands.
- Thereby, businesses can identify the strengths and weaknesses of products and services, as well as the problems that customers are facing.
- From there, businesses can make appropriate improvements to improve product quality, service, and customer experience.
- ***Enhance customer service:***
- Sentiment analysis helps businesses listen and understand customers' wants and needs.
- Thanks to that, businesses can provide better customer service, quickly and effectively solve problems that customers encounter.
- This will help increase customer satisfaction and loyalty to the brand.

CHAPTER 2: DATA FOR ANALYSIS

2.1. Identify analytics data.

- Data is retrieved from MongoDB (NoSQL)
- After successfully extracting data, proceed to clean data step.
  - ***Use python to clean data.***

There are a total of 22 columns, however only 7 columns are necessary, so unnecessary columns need to be removed.

Convert column 'ip' to column 'city' using Geolocation DB in Python to get information about ip zone used.

For example: 116.105.57.67 => Ho Chi Minh City

For the 'timestamp' column, we will use that column to add and convert the 'date' column for analysis.

For example: 1628960400 => 15/08/2021

```
Unnamed: 0        int64
_id              object
ranking           int64
layer             int64
user_id           int64
like              int64
user_email       object
comment_on       object
ip               object
user_fullname    object
publish_status    int64
object_id        object
content          object
reviewer         object
report_count      int64
review_status     int64
timestamp         int64
device           object
report_reason    object
dislike           int64
comment_status    int64
device_id        object
dtype: object
```

(Initial data)

```
user_id                  int64
tag                     object
object_id               object
comment                 object
device                  object
datetime     datetime64[ns]
city                    object
dtype: object
```

(*After processing)*

2.2. Classification of sentiment

Method: Machine Learning combined with AI

Emotion classification is based on the **Phobert-Base-Vietnamese-Sentiment** model, which is a language model developed by VinAI, based on the RoBERTa architecture, and refined on a Vietnamese data set with emotion annotations.

Results after processing 10 thousand comments:

```
sentiment
Positive    7800
Negative    2200
Name: count, dtype: int64
```

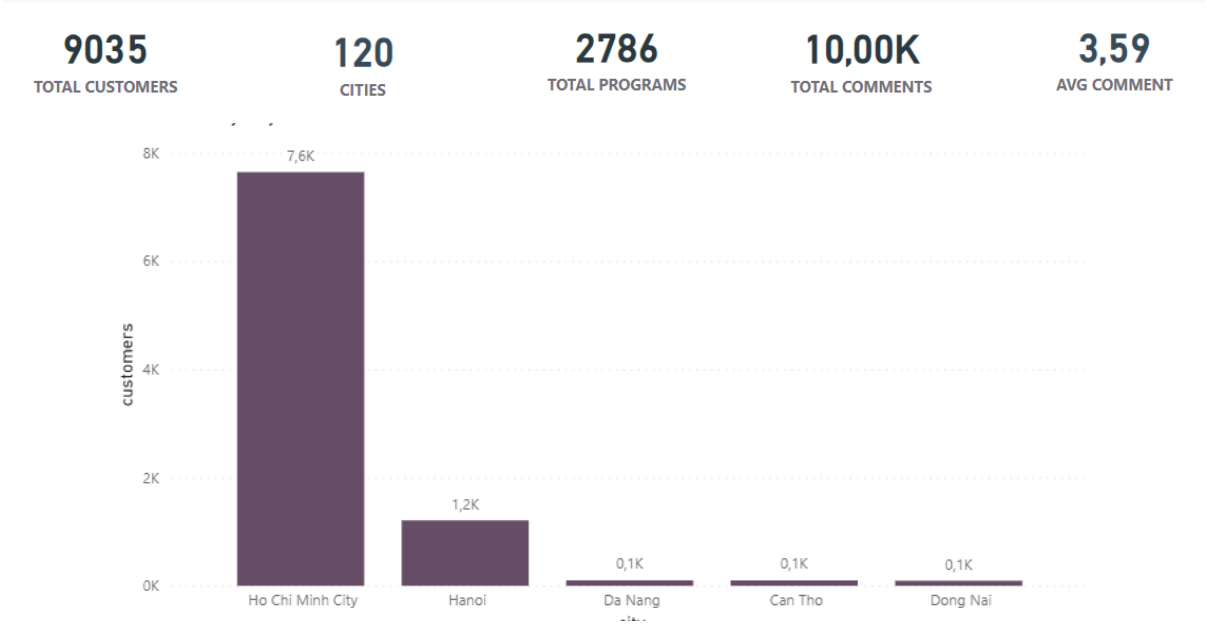
Dataset after completing the processing process

Column	Description
user_id	ID of the customer registered when using the service
tag	Channel or program customers commented
Object_id	Programs on FPT app
Comment	Show or channel on which viewers leave comments
Device	Devices customers use on FPT app
Datetime	Time customers leave comments
sentiment	Classify customer sentiment based on comments
city	Where the customer lives

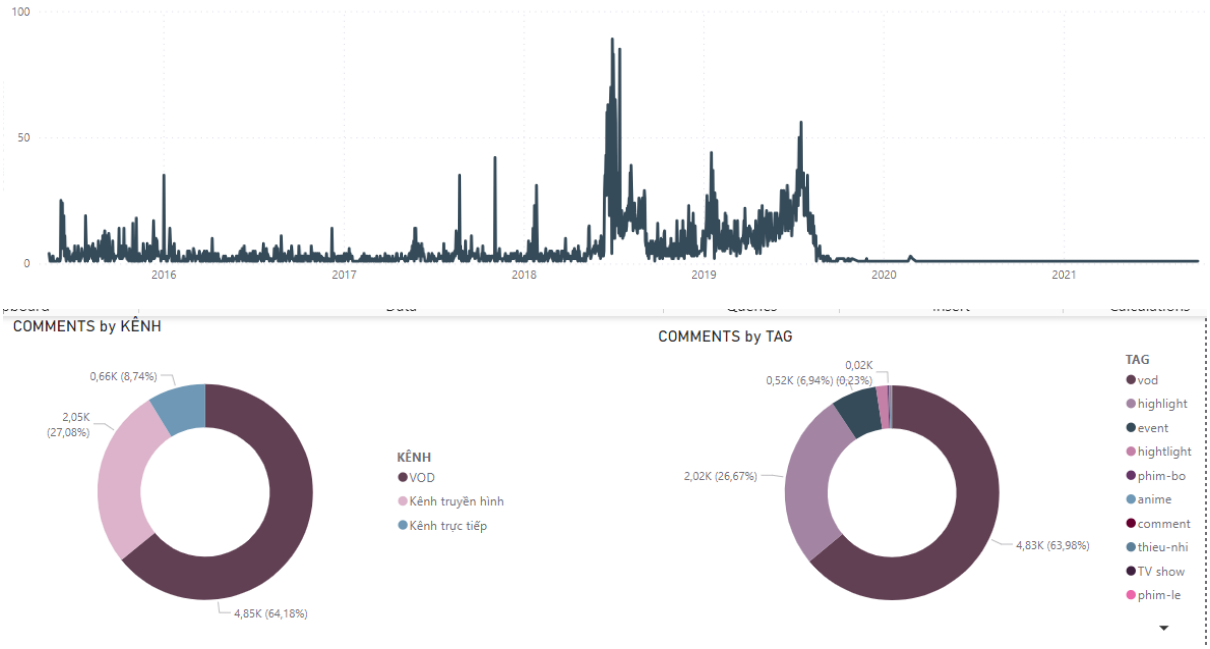
CHAPTER 3: ASSESSING THE SATISFACTION LEVEL OF CUSTOMERS USING FPT SERVICES THROUGH SENTIMENT ANALYSIS

3.1. TRENDS IN CUSTOMERS' SERVICE USE

The data below were recorded from 2015 to 2021.



- 97% of customers using FPT apps are mainly concentrated in big cities such as Ho Chi Minh and Hanoi
- 10% of interactive customers leave at least 2 comments.
- Each show has an average of 3.59 comments, which may indicate that each show receives some attention, but there may also be large differences between shows in terms of popularity or controversy.



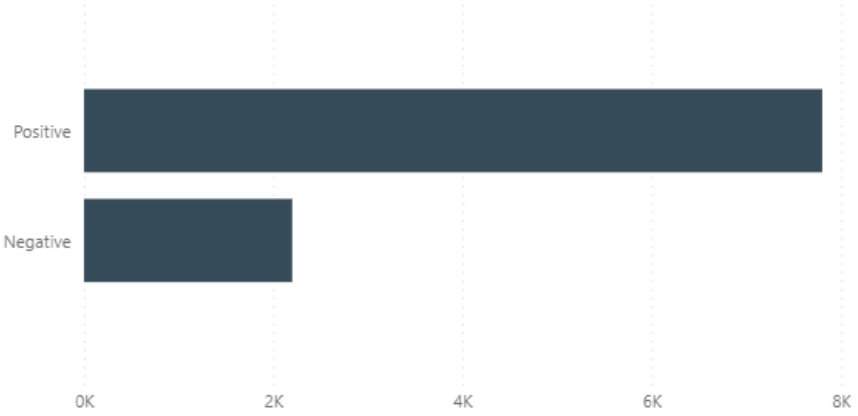
- During June and July, 2018, the amount of interaction was quite high because this was the time when the World Cup took place, which explains why there was a sudden increase in the amount of interaction during this period. Many VOD programs and highlight videos of football matches receive a large amount of interaction.
- VOD programs, TV channels and live channels may have provided content related to the World Cup, thereby attracting many comments and interactions from fans.

⇨ ***This high increase in engagement indicates that the content provided accurately reflects the needs and desires of customers in that moment.***

3.2. CUSTOMER EMOTION ANALYSIS (SENTIMENT ANALYSIS)

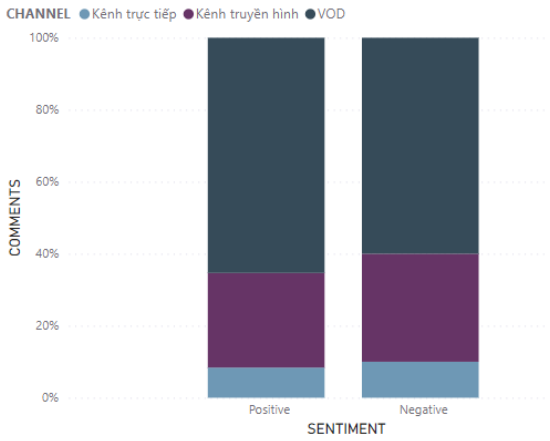
Based on the comments customers leave after using the service, we apply the pre-trained model and the Phobert-Base-Vietnamese-Sentiment -VinAi natural language processing algorithm to classify customer emotions. row

After analysis, we get the following general results:



- Most customers leave positive comments about the company's service. It shows ***that for every negative review, there are three positive reviews***, which is a good sign of satisfaction with the service.

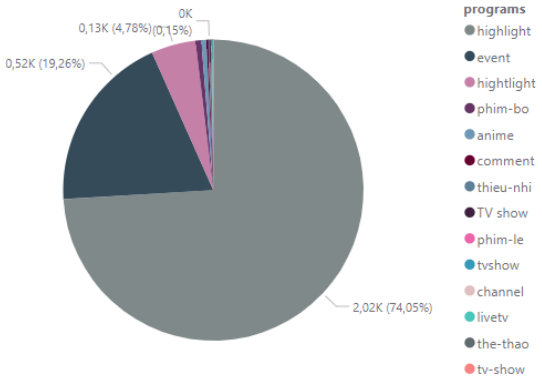
3.2.1. SENTIMENT ANALYSIS BY CHANNEL



Kênh	Tích Cực	Tiêu Cực
VOD	3858	990
Kênh truyền hình	1551	495
Kênh trực tiếp	495	165
Total	**5904**	**1650**

⇨ Data shows that the VOD channel has a strong attraction because it accounts for 55.18% of positive comments, but the rate of negative reviews is also quite high, accounting for 60% of the total number of negative comments, so it is also necessary to pay attention to managing and improving service quality to maintain and develop customer satisfaction

3.2.2. SENTIMENT ANALYSIS BY PROGRAMS

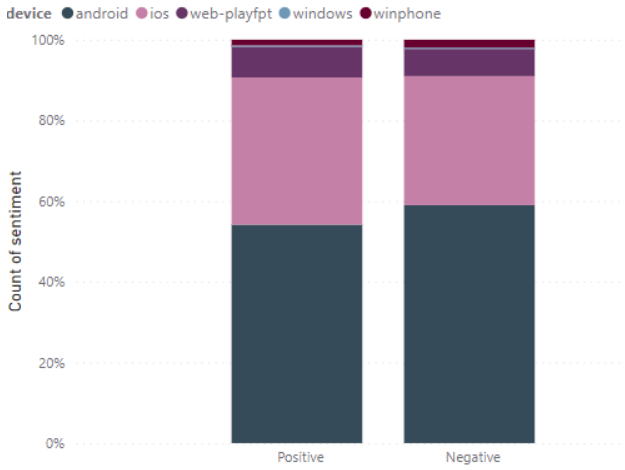


Programs	Tích Cực	Tiêu Cực
highlight	1525	490
event	388	136
hightlight	104	26
phim-bo	17	
anime	13	2
comment	3	2
thieu-nhi	2	2
TV show	2	1
tvshow	2	
phim-le	1	1
the-thao	1	
tv-show	1	
channel		1
livetv		1
Total	**2059**	**662**

- High interest in sports: A high number of positive comments shows that users have strong interest and excitement in sports content, especially highlights from matches.
- For every negative comment about match highlights, there are 3 positive comments

⇨ Data shows that sports content, especially highlights from matches, always attracts a lot of interaction from users.

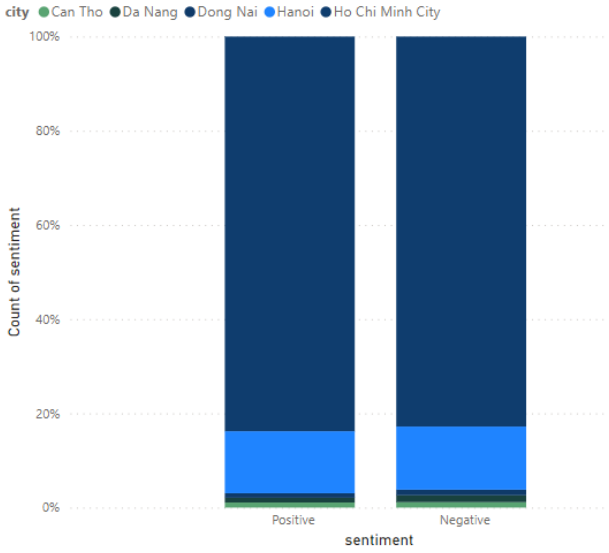
3.2.3. SENTIMENT ANALYSIS BY USER DEVICES



Device	Tích Cực	Tiêu Cực
ios	2845	704
android	2395	758
web-playfpt	530	138
winphone	110	43
windows	24	7
Total	**5904**	**1650**

- It seems that iOS users tend to leave more positive comments than Android users, based on the percentage of total positive and negative comments.
- The number of negative comments from Android users may indicate that the quality of service on this platform is not good.

3.2.4. SENTIMENT ANALYSIS BY AREA



city	Positive	Negative
Ho Chi Minh City	5988	1655
Hanoi	941	268
Da Nang	72	29
Dong Nai	69	24
Can Tho	77	23
Thanh Pho Thai Nguyen	51	17

- Customers are mainly concentrated in Ho Chi Minh City. Customer satisfaction in Ho Chi Minh can be understood through the high positive comment rate, 76.8%. This shows that your service meets their needs and expectations well. However, the rate of negative comments is also high, 75.2%, so it is necessary to consider improving and resolving specific issues to maintain and enhance this satisfaction.

⇨ Service Insight: Despite the high number of positive interactions, the rate of negative feedback is also significant. This could be an opportunity to review and improve aspects of service or customer support, especially in Ho Chi Minh.

CODE

Step 1: Get data from Mongodb

```
import pymongo
from pymongo import MongoClient
CONNECTION_STRING = "mongodb+srv://…. "
client = MongoClient(CONNECTION_STRING)
db = client['…']
collection = db["Comments"]
cursor = collection.find()
list_cur = list(cursor)
```

Step 2: Data Clean

```
import pandas as pd
import os
#loại bỏ những cột không dùng
df = df.drop(['Unnamed: 0',
'ranking','layer','like','user_email','user_fullname','reviewer','report_count
','review_status','report_reason','dislike','comment_status','device_id',
'_id']
        , axis=1)
# Chuyển đổi cột timestamp từ dạng chuỗi sang số (numeric type)
df['timestamp'] = pd.to_numeric(df['timestamp'])
# Sau đó chuyển đổi sang datetime
df['datetime'] = pd.to_datetime(df['timestamp'], unit='s')
df['date']= df['datetime'].dt.date
#Tìm khu vực khách sử dụng thông qua địa chi Ip
import requests
import json
from concurrent.futures import ThreadPoolExecutor
#Hàm để lấy thông tin địa lý từ địa chỉ IP sử dụng Geolocation DB
def get_geo_info(ip):
    try:
        request_url = f'https://geolocation-db.com/jsonp/{ip}'
        response = requests.get(request_url)
        result = response.content.decode()
        result = result.split("(")[1].strip(")")
        return json.loads(result)
    except Exception as e:
        print(f"Error getting geolocation for IP {ip}: {str(e)}")
        return None
    # Hàm để áp dụng xử lý song song
def get_geo_info_concurrent(ip_list):
    with ThreadPoolExecutor(max_workers=10) as executor:
        results = list(executor.map(get_geo_info, ip_list))
    return results
# Lấy thông tin địa lý cho mỗi địa chỉ IP
geo_info_list = get_geo_info_concurrent(df['ip'].tolist())
```

```
# Chuyển danh sách kết quả thành DataFrame
geo_df = pd.DataFrame(geo_info_list)
# Kết hợp thông tin địa lý với DataFrame gốc
df = df.join(geo_df)
```

Step 3: use Machine Learning to classify comments

```
!pip install transformers
!pip install torch
from transformers import AutoModelForSequenceClassification, AutoTokenizer
import torch
# Initialize tokenizer and model
tokenizer = AutoTokenizer.from_pretrained('vinai/phobert-base')
model = AutoModelForSequenceClassification.from_pretrained('vinai/phobert-
base')
# Function to predict sentiment
def predict_sentiment(text):
    inputs = tokenizer(text, return_tensors='pt', padding=True,
truncation=True, max_length=256)
    outputs = model(**inputs)
    probs = torch.nn.functional.softmax(outputs.logits, dim=-1)
    # Assuming that the model has three output neurons corresponding to the
classes
    sentiment_classes = ['Negative', 'Positive']
    return sentiment_classes[torch.argmax(probs)]
# Apply the function to the 'content' column
df['sentiment'] = df['content'].apply(predict_sentiment
```