SENTIMENT ANALYTICS PROJECT

Table of Contents

I. ĐẶT VẤN ĐỀ

Hiện tại có rất nhiều bình luận trên ứng dụng trên FPT play, tuy nhiên những bình luận này đều không có thang điểm đánh giá cho người dùng.

Bài báo cáo này tôi sẽ thu thập dữ liệu về và sử dụng những phương pháp phân tích phù hợp để có thể hiểu được về khách hàng cũng như chất lượng sản phẩm mà chúng ta cung cấp cho khách hàng. Từ đó có thể đưa ra những chiến lược chăm sóc khách hàng phù hợp cũng như cải thiện dịch vụ.

II. PHƯƠNG PHÁP PHÂN TÍCH

2.1 Sentiment Analysis

- Phân tích cảm nghĩ (Sentiment Analysis) là hoạt động diễn dịch và phân loại tự động các cảm xúc (tích cực, tiêu cực hoặc trung tính) từ dữ liệu văn bản như các bài đánh giá bằng chữ, các bài đăng trên mạng xã hội.
- Áp dụng Machine Learning vào việc xử lý ngôn ngữ tự nhiên (Natural Language Proccessing)

2.2 Mô hình xử lý ngôn ngữ Tiếng Việt_PhoBERT

- Dựa trên kiến trúc Transformer và được huấn luyện trên khối lượng lớn dữ liệu văn bản tiếng Việt, có khả năng hiểu sâu sắc ngữ cảnh của từng từ trong câu, kể cả khi từ đó năm trong các cấu trúc phức tạp. Điều này đặc biệt hữu ích trong việc phân tích cảm xúc, nơi mà ý nghĩa của một câu không chỉ phụ thuộc vào từng từ đơn lẻ mà còn vào cách chúng tương tác với nhau trong ngữ cảnh.

III. CÁC BƯỚC THỰC HIỆN

- Dùng Python kết nối đến MongoDB
- Sử dụng code Python để lấy dữ liệu từ Mongo DB về (Mongo DB là một nơi chứa dự liệu lớn, băng dạng văn bản)
- Lọc ra những cột cần thiết để phân tích
- Sử dụng Python để làm sạch và biến đổi dữ liệu
- Ứng dụng Mô hình Pho_BERT để chấm điểm cho từng content
- Dùng Power BI để trực quan hóa dữ liệu

IV. DỮ LIỆU

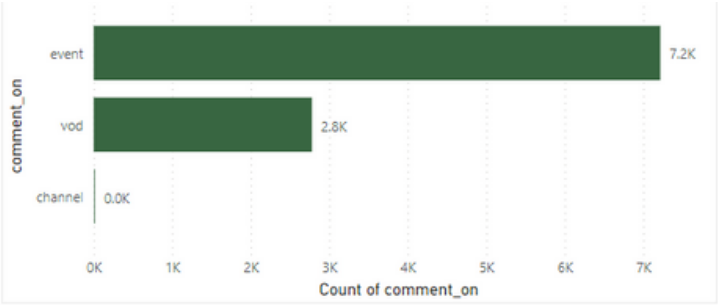
Sau khi dữ liệu đã được làm sạch và biến đổi ta có bảng sau:

_id	**mã của mỗi content**
user_id	mã người dùng
comment_on	thể loại của nội dung được comment
content	nội dung comment
date	ngày comment
hour	giờ comment
sentiment	điểm của từng comment (0 là tiêu cực, 1 là tích cực, 2 là trung tính)

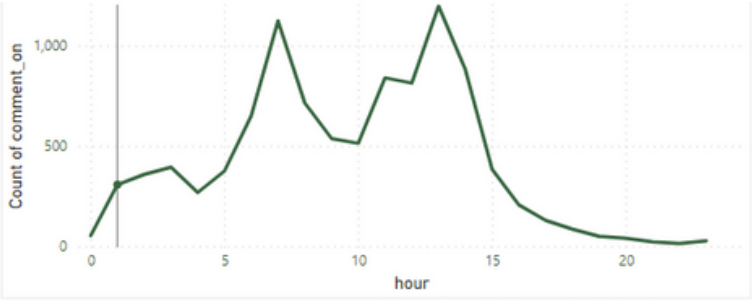
V. TRỰC QUAN HÓA DỮ LIỆU

5.1 Tổng quan

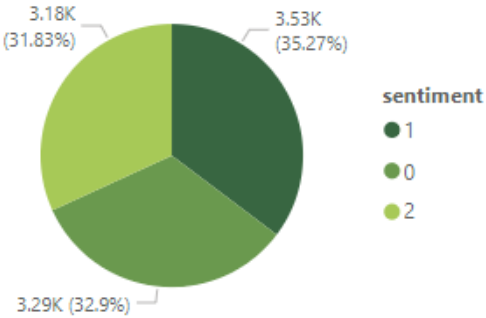
Trong khoảng thời gian là 1 tháng từ 13/5/2015 - 14/6/2015, FPT Play thu hút được 5893 khách hàng với 10.000 bình luận trên 3 hệ thống: event, vod, channel



Trên hệ thống event sẽ thu hút nhiều lượng bình luận nhất với 7200 lượt bình luận, 2800 lượt bình luận đến từ hệ thống vod, channel hầu như không có lượt bình luận nào. Người dùng chủ yếu bình luận vào khung giờ 7h sáng và 13h chiều.

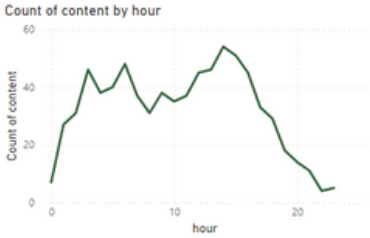


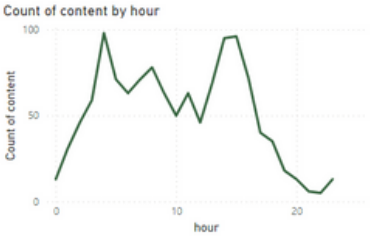
Trong số 10k lượt comments trên các nền tảng của FPT Play thì 35,27% là những comments tích cực, 32,9% là những comments tiêu cực



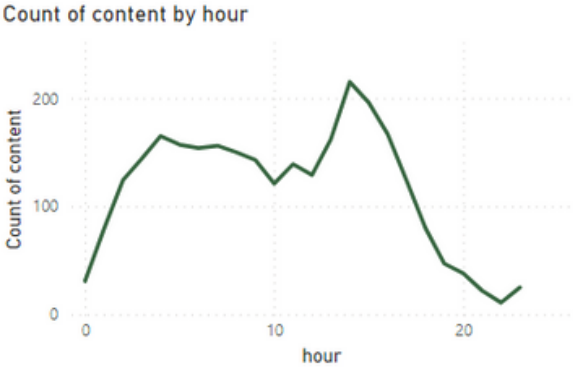
5.2 Phân tích theo thời gian

Nền tảng VOD: (Video on Demand) là một hệ thống truyền thông tương tác cho phép người dùng lựa chọn, xem và tương tác với nội dung video trên nhiều thiết bị thông qua kết nối Internet, không phụ thuộc vào lịch trình phát sóng cố định.





Những comments tiêu cực thường xuất hiện vào 3h, 6h, 12h-13h, 14h; những comments tích cực thường xuất hiện vào 4h, 8h, 14h-15



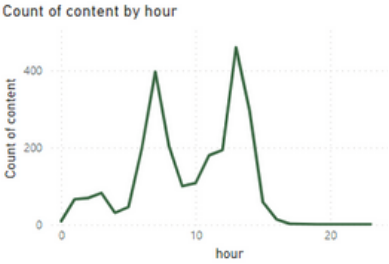
Người dùng thường sử dụng nền tảng này vào lúc 4h và 14h.

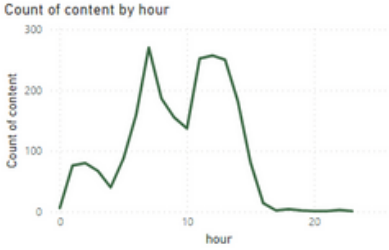
=> Những thời gian comments tích cực và tiêu cực ở trên cũng không thể đánh giá được do nền tảng VOD không phản ánh đúng thời gian thực

content
zxscvcjcjkcjk
yo quá fdi mădt
Yi An đừng thích Eun Bi mà. Ah thật là đần -_-
ý nghĩa quá
xóa hết cmnr
xmen9313 là bài Radioactive + Imagine Dragon
xen trên tv bị lỗi không lên hình.
xem xong mới biết nghiệp vụ cảnh sát hàn quốc quá kém :))))
xem ức chế vãi đợi 2 ngày rồi chưa ra tập nào
xem trên app thì tắt chặn quảng cáo kiểu gì? :))
xem thứ 2 nhảm vãi
xem phim ức quá mà phải bình luận. chưa bjio xem phim nào tồi tệ ntn. nội dung, tình tiết, võ thuật đều như shit
xem phim ức chế thêm .có 1 thằng đó mà ko giet dc
xem phim này chỉ vì có a Kiều Chấn Vũ . Bản liêu trai này tệ nhất không hấp dẫn
xem mà thấy tay kia sến quá =]]]
Xem lại tập 1 phần 1 mới nhận ra winterfell từng nóng như thế nào @_@. Chiều thứ 3 mới có
xem ko dc..
xem ko dc, pi lỗi rồi
xem hoài k dc toàn bị dật
xem dở thúi mà cũng coi đúng là trẻ con hzaha
xem đéo đk
xem đến tập 21 lại trở về 20 bực quá.

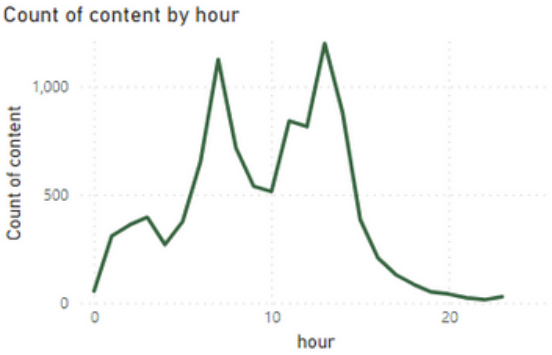
=> Những comments tiêu cực này hầu hết nói về lỗi không xem được. Vì vậy đội ngũ kỹ thuật cần xem lại về những lỗi trên và có biện pháp khắc phục kịp thời để nâng cao trải nghiệm người dùng.

Nền tảng event





Những comments tiêu cực thường xuất hiện vào 7h, 13h; những comments tích cực thường xuất hiện vào 7h, 11h-13h



Người dùng thường sử dụng nền tảng này vào lúc 7h và 13h.

content	Count of content
Thua roi	7
Chiến thuật gì thế này?! Đội hình BjQuery1111002120250533334911_1433944287387 Không sợ thua thái mãi thành vết bớt à	6
Thua rồi	6
co len viet nam	5
cố lên việt nam ơi	4
ko xem duoc	4
lại quảng cáo	4
việt nam cố lên	4
Buồn quá	3
Chán quá	3
Đá chán lắm	3
Đẳng cấp quá chênh lệch, tường thuật làm gì chứjQuery1111074418211214388886_143381717829477	3
https://www.youtube.com/watch?v=tzeltmUNYC0	3
sao cứ chiếu lại quài vậy..không có cái nào trực tiếp hết hả jQuery1111008705526590347 29_1434018531944????????	3
thua chac	3
thua nua roi du me	3
thua rui	3
Tiếp lại xem chán vãi	3
mạng lag quá	2
Total	**2518**

=> Những comments tiêu cực này hầu hết nói lên cảm xúc khi xem một trận đấu

CODE

IMPORT DATA FROM MONGO DB

```
!pip install pymongo
import pymongo
from pymongo import MongoClient
CONNECTION_STRING =
"mongodb+srv://longth:asdQWE123@cluster0.tehe2.azure.mongodb.net/?
retryWrites=true&w=majority"
client = MongoClient(CONNECTION_STRING)
db = client['Data_Engineer']
collection = db["Comments"]
cursor = collection.find()
list_cur = list(cursor)
import pandas as pd
# biến đổi dữ liệu thành dạng dataframe
comments_df = pd.DataFrame(list_cur)
comments_df
```

CLEAN DATA

```
df = comments_df[["_id", "user_id", "comment_on","ip", "object_id", "content", "timestamp",
"device"]]
df
print(df.dtypes)
from datetime import datetime
import pandas as pd
df['datetime'] = pd.to_datetime(df['timestamp'], unit='s')
df["timestamp"] = pd.to_datetime(df["timestamp"], unit='s')
df['date']= df['datetime'].dt.date
df['hour']= df['datetime'].dt.hour
df
```

USING PRE-TRAINED PHOBERT MODEL

```
!pip install transformers
import numpy as np
import pandas as pd
import torch
from transformers import AutoModel, AutoTokenizer, RobertaForSequenceClassification
import requests
```

```
# hàm apply để đánh giá từng comment; 0 là tiêu cực, 1 là tích cực, 2 là trung tính hoặc ko
thể xác định
def sentiment(sentence):
  try:
    input_ids = torch.tensor([tokenizer.encode(sentence)])
  except:
    return 2

  if input_ids.shape[1] > tokenizer.model_max_length:
    return 2

  with torch.no_grad():
      out = model(input_ids)
      return np.argmax(out.logits.softmax(dim=-1).tolist()[0])
import time
start_time = time.time()
# test 10000 first comments
test = df[:10000]
test['sentiment'] = test['content'].apply(sentiment)

print("--- %s seconds ---" % (time.time() - start_time))
data_10000=test[['user_id','ip','comment_on','content','sentiment', 'date', 'hour']]
data_10000
data_10000.to_csv('data_10000.csv')
```