Report Capstone Project
Introduction to Data Science

Project: Lazada recomendation base on reviews

Group 5: Giap Thi Thuy Van (leader) - 20176905
Ta Dinh Son - 20176862
Tran Thi Hang - 20176748
Duong Thi Hue - 20176772

Table of content

Abstract

The exigency of online shopping is increasing day by day. Following a survey of Q&Me in 2019 , the ambiguous quality of products take fifty percent of reasons that customers concern not to shopping online and sixty six percent in dissatisfy product qualification with online experience. Therefore,to improve online experience our project give the advices for customers about specific items base on reviews. We rate each item in three level: good, netral and bad and base on our rate and recomment the customer will have better decision and online experience. We chose Lazada for E-commerce Exchange because it is in top 10 E-commerce websites by monthly visits ranked by Alexa.

1. List of tasks and project distribution

- Collecting data

  Giap Thi Thuy Van : 50%

  Ta Dinh Son : 50%

- Process data: labelling

  Duong Thi Hue : 50%

  Tran Thi Hang : 50%

- Programming
  - Scrap data for test : Ta Dinh Son
  - Pre-train Bert: Giap Thi Thuy Van
  - K-nearest neighbor: Duong Thi Hue
  - Naive Bayes : Tran Thi Hang
- Write report

  Duong Thi Hue

- Presentation slide

  Tran Thi Hang

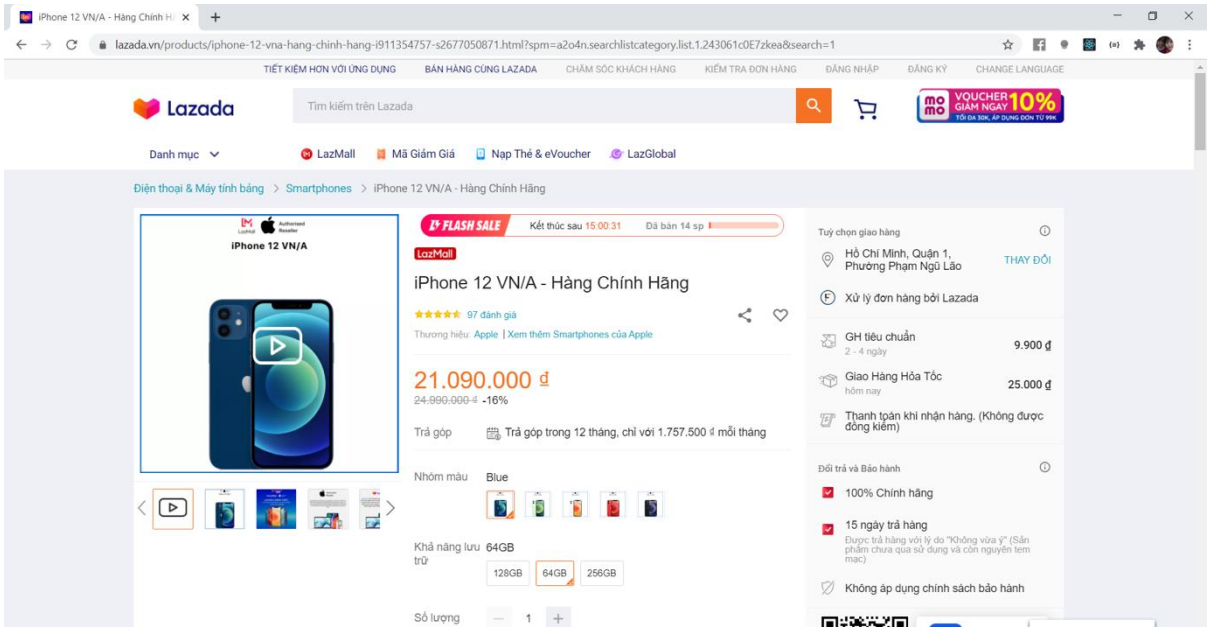
2. Formulate the problem

- **Problem**

From a link of product in Lazada, we will predict that the item is good, neutral or bad, base on comments of reviewers. Afterthat, we will response the advise whether the customer should buy this item.

- **Input** : link of product that you want to buy in Lazada

For example: https://www.lazada.vn/products/iphone-12-vna-hang-chinh-hang-i911354757-s2677050871.html?spm=a2o4n.searchlistcategory.list.1.243061c0E7zkea&search=1



- **Output:**

Label for the product in range 0 to 2 and recommendation for customer to buy or not

```
No of bad comments =  9
No of neutral comments =  7
No of good comments =  37
Good! You can buy it!
```

3. Data preparation

3.1. Data scrapy

For training dataset, we take from resource.

For test dataset, we scrap data from Lazada by using below piece of codes.

```
def load_url_selenium (url):
    driver=webdriver.Chrome(executable_path='/usr/bin/chromedriver')
    print("Loading url=", url)
    driver.get(url)
    review_csv=[]
    while True:
        #Get the review details here
        WebDriverWait(driver,10).until(EC.visibility_of_all_elements_located((By.CSS_SELECTOR,"div.item")))
        product_reviews = driver.find_elements_by_css_selector("[class='item']")
        # Get product review
        for product in product_reviews:
            review = product.find_element_by_css_selector("[class='content']").text
            if (review != "" or review.strip()):
                print(review, "\n")
                review_csv.append(review)
            # else:
            #     print(review)
            #     review_csv.append("No comments/review is an image")
        #Check for button next-pagination-item have disable attribute then jump from loop else click on the next button
        if len(driver.find_elements_by_css_selector("button.next-pagination-item.next[disabled]"))>0:
            break
        else:
            button_next=WebDriverWait(driver, 10).until(EC.visibility_of_element_located((By.CSS_SELECTOR, "button.next-pagination-item.next")))
            driver.execute_script("arguments[0].click();", button_next)
            print("next page")
            time.sleep(2)
    driver.close()
    print(review_csv)
    return review_csv
```

The dataset is an excel file includes comments with label which is from zero to two. The higher label number, the higher recommendation for buying this item .Therefore, we have to label the collected data in 3 levels “0”, ”1”, and “2” corresponding to “bad”, ”neutral” and “good” respectively. This step takes a lot of time because of the huge dataset.

	A	B
1	stars	comment
2	2	ok
3	0	máy mới sạc chưa sài mà đã hết pin .cảm ứng rất chậm
4	2	Máy bền, bé lâm rơi nhiều lần mà không bị hỏng.
5	1	Nhưng khi chơi game nhẹ thì OK game nặng giật tung đít với các chip 8lõi mà yếu hơn cả chip 4 lõi Snapdragon 425 ....ae chơi game không nên mua mặc dù thông số là thật đã kiểm tra bản phần mềm untutu 68k điểm nhưng liên quân low seeting vẫn tụt fps...
6	1	Tôi dùng con này hơn 1 năm rồi, rớt mấy lần mà bây giờ vẫn xài tốt, hồi trước mua 4 triệu 9 luôn :(Nhược điểm là bộ nhớ trong quá ít 4GB mà hệ thống chiếm hết hơn 1/2 còn trống có 2gb là cao, hệ điều hành ko update gì hết, 4.1 là ko up lên được nữa
7	0	nhận máy xong thật sự thất vọng về chất lượng. Kiểm tra xong thấy máy bị hở ở mép màn hình, Tem bảo hành cũng rách luôn. Shop trả lời là lỗi do người sử dụng do nhận máy đã 5 ngày rồi. Thôi thì lỗi của mình ham giảm giá và mở hàng ra chậm trễ. Đăng lên để người mua sau biết mà tránh mua hàng củ qua mạng. Nên đi mua trực tiếp cho yên tâm về chất lượng.
8	0	Chay con thua cai may cui chip 4 nhan 1.3g nữa. thấy chíp 8 nhân 2g ram mua về chơi lien quân cho đã. Mua zìa máy nhu lo . Thua cái máy cui bap chip 4 nhân cùi t. Hic hic... tình hình là sản phẩm hàng nhái mà không đc đổi trả. hiểu anh em làm ăn luôn. Bye bye lazada .
9	0	máy chất lượng kém mua dược 2ngay loa đã rè bin nhanh hết
10	0	hiện tại máy mình dùng thử thấy hay bị lỗi kết nối mạng WiFi! đang online thì bị offline rồi một lúc sau thì lại vào được! WiFi nhà m ổn định, đã kiểm tra trên các thiết bị khác nhưng không bị mất kết nối. shop check giúp!
11	0	Mình mua hàng của lazada cũng nhiều nhưng dạo này hàng xấu và lỗi ko, mua điện thoại thì giao điện thoại cũ. Hàng này thì ko xóa dc....
12	2	hang tuongdoi tot nhung phanmem shop nam o dau vayshopco lua gatphai khong
13	1	♡♡♡♡😄
14	1	chup hình hơi xấu máy chậm thua samsung lâu hết pin

3.2. Data cleaning, Data pre-processing, Data integrating

Firstly, from dataset, for each comment we will clear all special characters. Then we split each sentences into separate words based on white space and punctuation. After that, we embedded it to numeric vector.

```
# Load data from crawler file
def load_data():
    df = pd.read_csv("data_crawler.csv")
    return df

# Standardize text data
def standardize_data(row):
    # Xóa dấu chấm, phẩy, hỏi ở cuối câu
    row = re.sub(r"[\.,\?]+$-", "", row)
    # Xóa tất cả dấu chấm, phẩy, chấm phẩy, chấm than, ... trong câu
    row = row.replace(",", " ").replace(".", " ") \
        .replace(";", " ").replace("“", " ") \
        .replace(":", " ").replace("”", " ") \
        .replace('"', " ").replace("'", " ") \
        .replace("!", " ").replace("?", " ") \
        .replace("-", " ").replace("?", " ")
    row = row.strip()
    return row

# Tokenizer
#split into word
def tokenizer(row):
    return word_tokenize(row, format="text")
```

We use pre-trained deep learning model to process clear comments. We classified each sentence as either speaking "positively" about its subject of "negatively". We will then use the output of that model to classify the text.

By using BERT, the results of the processing will be returned into last_hidden_states. The output includes token CLS at the begining of every sentence, so we need to slice only the part of out put we need.

4. Machine Learning techniques

To solve our problem, we choose these Machine Learning techniques

4.1. Classification

In our problem, we define that the response variable is categorical , the model is supervised and the objective is predict. This technique classify the training dataset as correctly as possible and best predict the class of new sample.Therefore, this technique is the most suitable for us to solve problem and requirements of capstone project.

We found that ***K nearest-neigbour*** and ***Naive Bayes*** might solve our problem

4.1.1. K nearest-neighbour

We chose k = 5 and p = 1 means for each test data to find label base on major of 5 nearest neighbour and Manhattan distance.

Firstly, we define the parameter k. Then we compute Manhattan distance of test vector to planes that contain training dataset. After that, place them in ascending order. Finally, find the major class top k neigbours. We use neighbors library:

```
lr_clf = neighbors.KNeighborsClassifier(n_neighbors = 5, p = 1)
lr_clf.fit(train_features, train_labels)
```

4.1.2. Bernoulli Naive Bayes

Firstly, we have to train base on training dataset, calculate $P(C_i)$ and $P(X_k|C_{ik})$. To classify $X = \{x_1, x_2, x_3, ...\}$, we have to calculate probability of each class when know X, X is labeled on class have maximum probability following

$$\max_{C_i \in C} \left(P(C_i) \prod_{k=1}^{n} P(x_k|C_i) \right)$$

We applied BernoulliNB library

```
lr_clf = BernoulliNB(binarize = .5)
lr_clf.fit(train_features, train_labels)
```

4.2. Natural Language Processing

Our project bases on comments of reviewers to give user the recommendation, while this technique is widely used to prepare text for Machine Learning. Tons of comments in a variety of formats and most of them will be full of typos, missing characters and other words that needed to be filtered out.

This technique map text into numerical representation to compute frequency of each word in the text. Firstly, we create a matrix of integers where each row represents a text document and each column represents a word. This representation matrix of the word frequencies is commonly called ***Term Frequency Matrix*** (TFM). From there, we can create another popular representation matrix of a text document by dividing each entry on the matrix by a weight of how important each word is within the entire corpus of documents. We call this method ***Term Frequency Inverse Document Frequency*** (**TFIDF**) and it typically works better for machine learning tasks.

Therefore, this technique is so useful and indispensable for us to handle comments from dataset.

4.3. Word embeddings

Word embeddings can capture the context of a word in a document. With the word context, embeddings can quantify the similarity between words, which in turn allows us to do arithmetic with words.

Word2Vec is a method based on neural nets that map this word in a corpus to a numerical vector. We can then use these vectors to find synonyms by computing the cosine similarity between the vectors, perform arithmetic operations with words, or to represent text documents.

Therefore, this technique help us on pre-step to applying a machine learning algorithm.

5. Model evaluation and selection

After tuning their parameters, we have

	K-nn	Naive Bayes
Accurency	0.85	0.72

The complexity of trainning K-nn is O(1), predict time complexity is O(k*n*d). However, when k is small and noise, the result may not right. It has to save all data in memory that effect accurency.

The testing and training time of Naive Bayes is fast because of independent element. However,The performace evaluation of it is lower than K nn

Therefore the K nearest neigbour is better than Naive Bayes.

6. References

- EVBN- Report-E-commerce industray in VietNam

- https://medium.com/m/global-identity?redirectUrl=https%3A%2F%2Ftowardsdatascience.com%2F10-machine-learning-methods-that-every-data-scientist-should-know-3cc96e0eeee9
- https://machinelearningcoban.com/2017/01/08/knn/
- https://machinelearningcoban.com/2017/08/08/nbc/