UEH UNIVERSITY
College of Technology and Design

Xây dựng mô hình phân tích cảm xúc người dùng

Chủ đề: Đại học UEH

Tài nguyên: Google Maps

- LÊ THỊ NGỌC ÁNH
- TRẦN PHẠM HẢI NAM
- LÝ MINH NGUYÊN

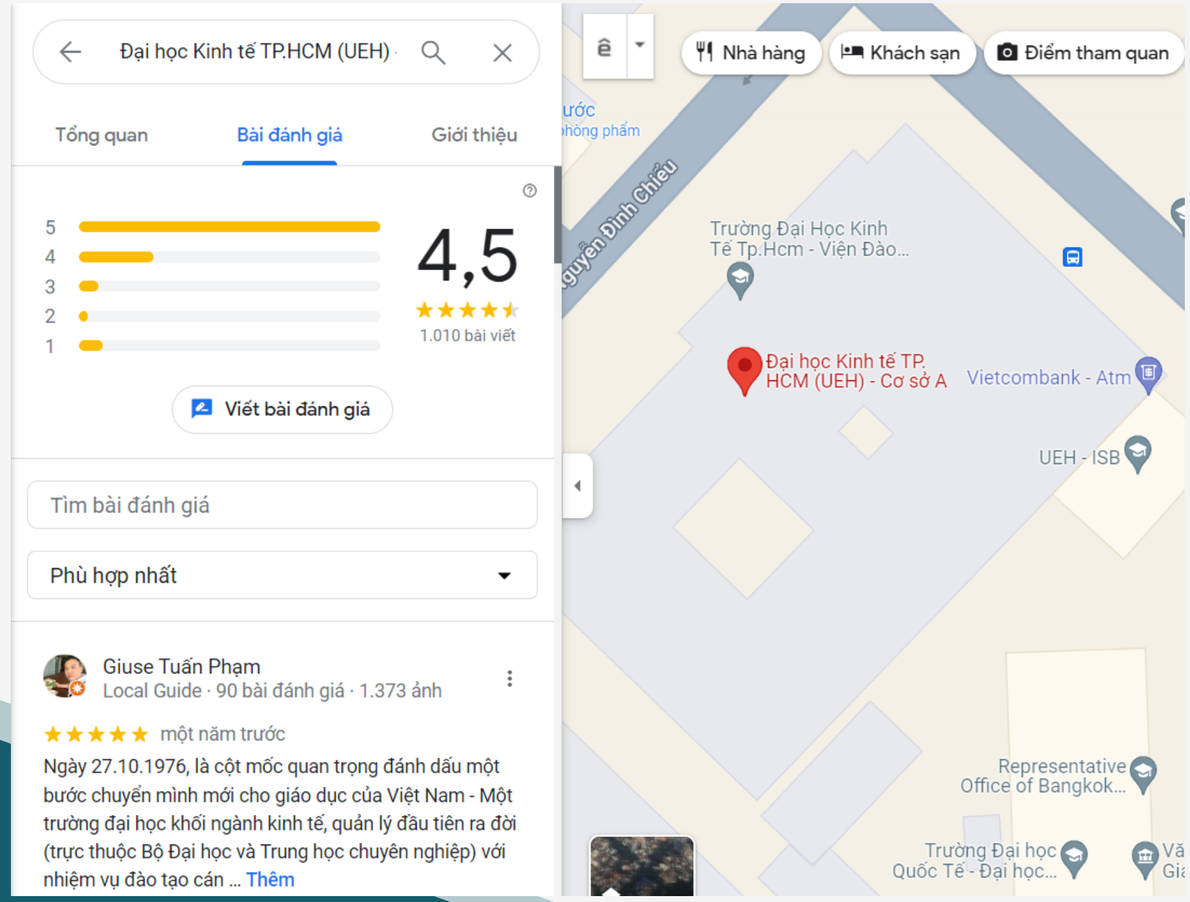
Học phần: Xử lý ngôn ngữ tự nhiên

UEH UNIVERSITY College of Technology and Design

Tính cấp thiết

- Áp dụng tri thức **giải quyết vấn đề** liên quan tới UEH;
- **Nắm được và đề xuất giải pháp** cho nhu cầu của người học;
- **Xây dựng mô hình dự đoán** cảm xúc người học, từ đó **cải thiện** những điểm chưa tốt của các đánh giá tiêu cực.

UEH UNIVERSITY
College of Technology and Design

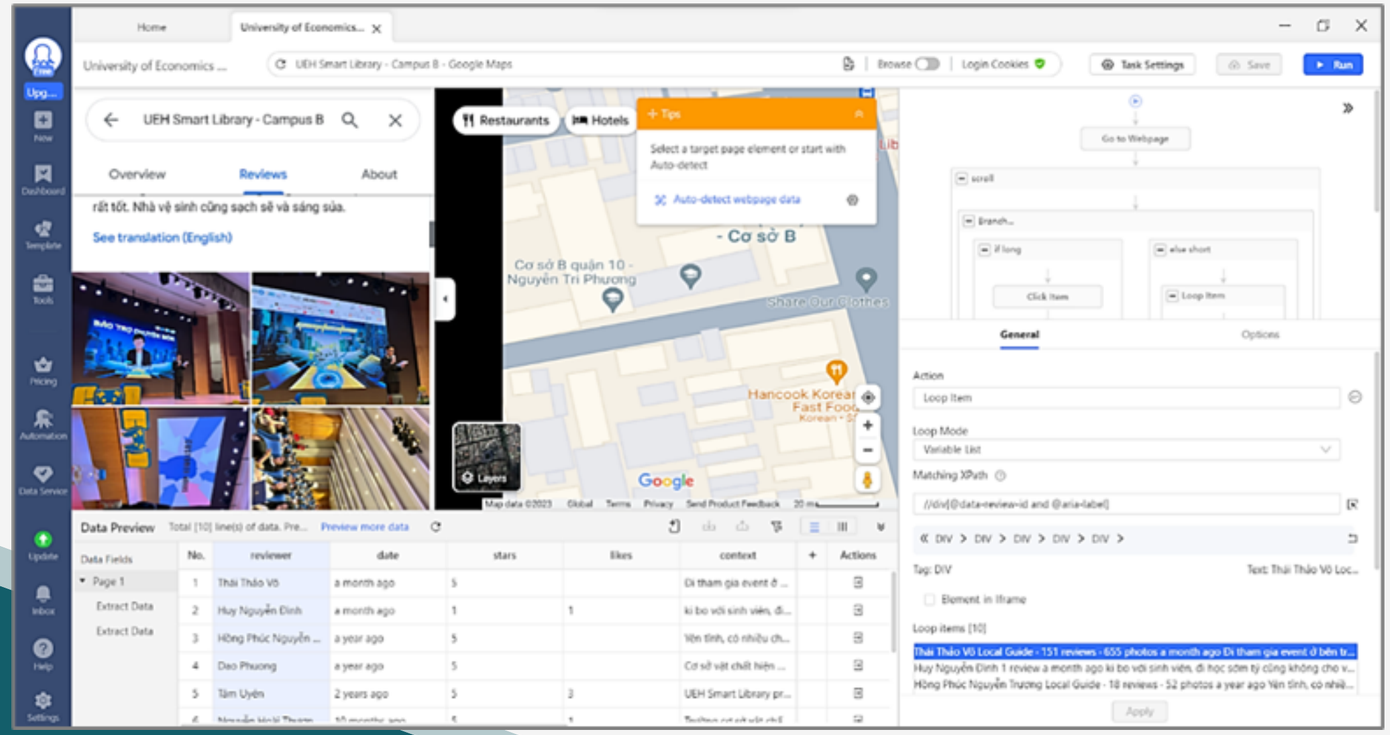


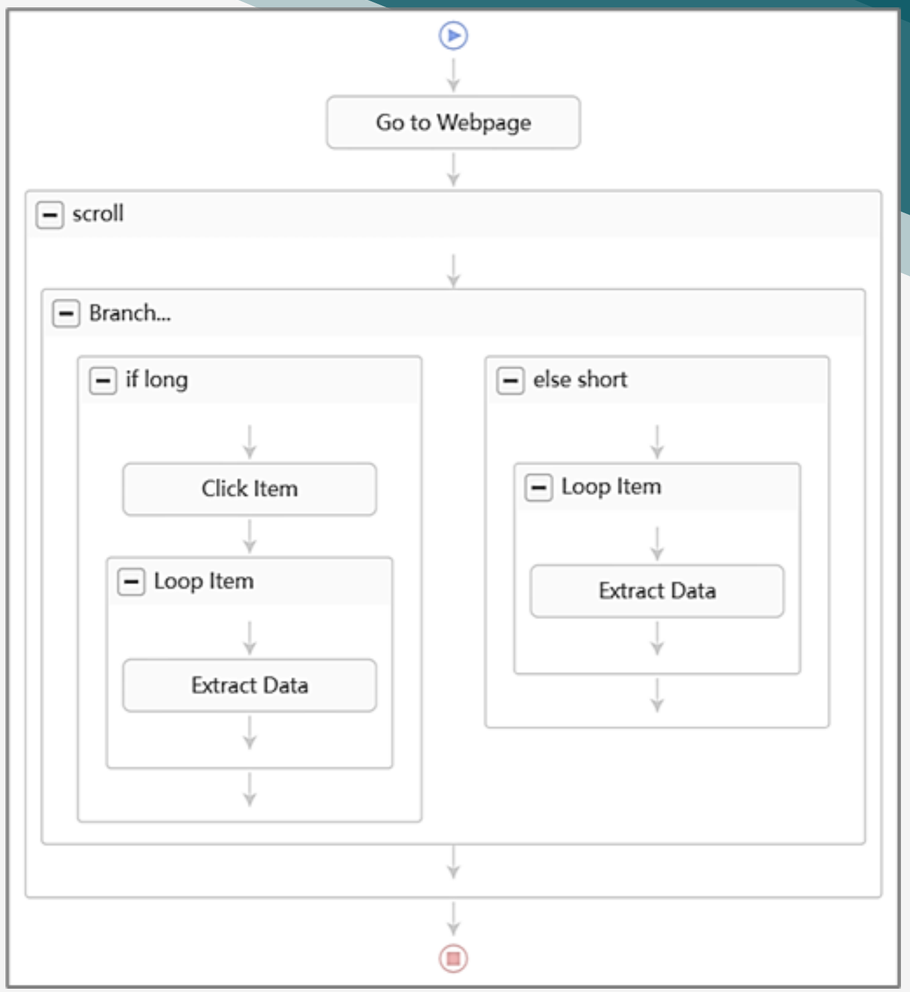
uehA.xlsx
uehB.xlsx
uehC.xlsx
uehD.xlsx
uehE.xlsx
uehH.xlsx
uehHOTEL.xlsx
uehI.xlsx
uehKTX135-1.xlsx
uehKTX135-2.xlsx
uehKTX4345.xlsx
uehLibB.xlsx
uehLibN.xlsx
uehN.xlsx
uehQ8.xlsx
uehV.xlsx
uehVL1.xlsx

UEH-reviews-cleaned.xlsx

UEH UNIVERSITY
College of Technology and Design

Data Crawling





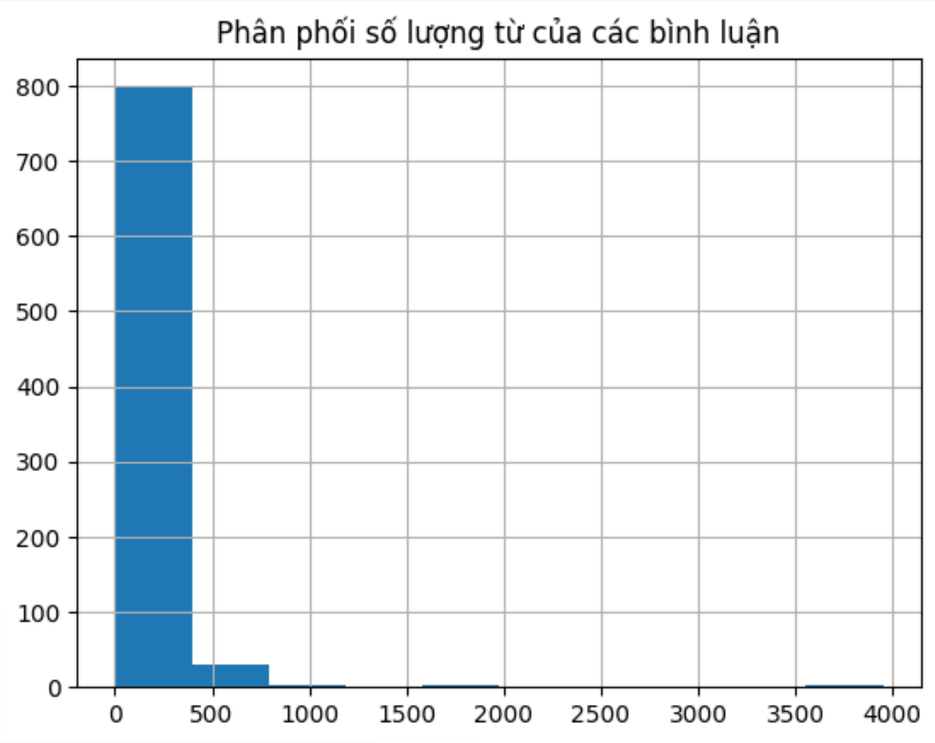
Octoparse

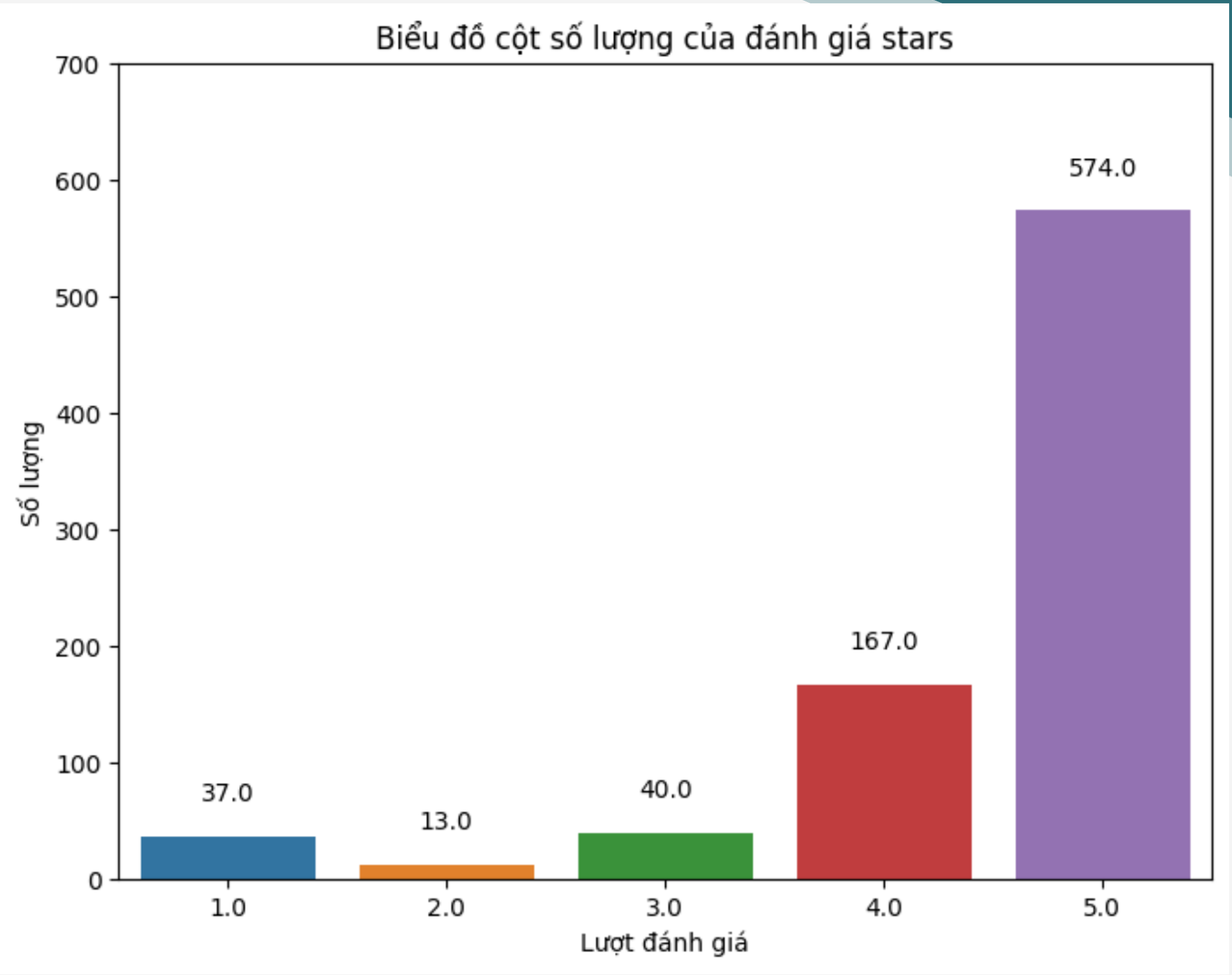
UEH UNIVERSITY College of Technology and Design

TÊN THUỘC TÍNH	MÔ TẢ	CHÚ THÍCH
reviewer	Tên của người đánh giá dựa theo tài khoản Google	Những tên này có thể không tuân theo quy luật bình thường
date	Thời điểm người đánh giá đăng bài đánh giá	Thời gian từ lúc bắt đầu đánh giá tới thời điểm ngày 10/11/2023
stars	Số lượng sao người đánh giá đăng trong bài đánh giá	Mức độ hài lòng đối với địa điểm đó
likes	Số lượt thích của đánh giá	Mức độ đồng cảm (độ tin cậy) đối với nội dung đánh giá
context	Nội dung bài đánh giá	Bao gồm tiếng Anh và tiếng Việt
location	Địa điểm đánh giá đề cập trong bài viết	Đây là một trong các cơ sở của UEH
sentiment	Cảm xúc của bài đánh giá	Có 2 giá trị là “positive” (tích cực) và “negative” (tiêu cực)

UEH UNIVERSITY College of Technology and Design

EDA





UEH UNIVERSITY College of Technology and Design

Preprocess

```python
def remove_punctuation(text):
    return ''.join([i for i in text if i not in string.punctuation])
def remove_emoji(text):
    emoji_pattern = re.compile('['
                               u'\U0001F600-\U0001F64F'  # emoticons
                               u'\U0001F300-\U0001F5FF'  # symbols & pictographs
                               u'\U0001F680-\U0001F6FF'  # transport & map symbols
                               u'\U0001F700-\U0001F77F'  # alchemical symbols
                               u'\U0001F780-\U0001F7FF'  # Geometric Shapes Extended
                               u'\U0001F800-\U0001F8FF'  # Supplemental Arrows-C
                               u'\U0001F900-\U0001F9FF'  # Supplemental Symbols and Pictographs
                               u'\U0001FA00-\U0001FA6F'  # Chess Symbols
                               u'\U0001FA70-\U0001FAFF'  # Symbols and Pictographs Extended-A
                               u'\U00002702-\U000027B0'  # Dingbats
                               u'\U000024C2-\U0001F251'
                               ']+', flags=re.UNICODE)
    return emoji_pattern.sub(r' ', text)
def preprocess(text):
    text = remove_punctuation(text)
    text = remove_emoji(text)
    text = text.lower()
    return text
```

UEH UNIVERSITY College of Technology and Design

```python
# Tạo cột để phân biệt 2 ngôn ngữ Anh-Việt
def detect_language(text):
    return langid.classify(text)[0]
```

from **vietnam_number** import **n2w**

```python
def convert_n2w_vn(text):
    if text.isdigit():
        return n2w(text)
    else:
        return text
```

tokens = ViTokenizer.tokenize(sentence).split()

num2words

```python
def convert_n2w_eng(text):
    if text.isdigit():
        return num2words(text)
    else:
        return text
```

tokens = word_tokenize(sentence)

UEH UNIVERSITY College of Technology and Design

```
# Xử lý stopwords cho tiếng Anh
eng_stopwords = nltk.corpus.stopwords.words('english')
eng_df['pre_context'] = eng_df['pre_context'].apply(lambda x: ' '
                        .join([word for word in x.split() if word.lower() not in eng_stopwords]))
# Xử lý stopwords cho tiếng Việt
vn_df['pre_context'] = vn_df['pre_context'].apply(lambda x: ' '
                      .join([word for word in x.split() if word.lower() not in vn_stopwords]))
```

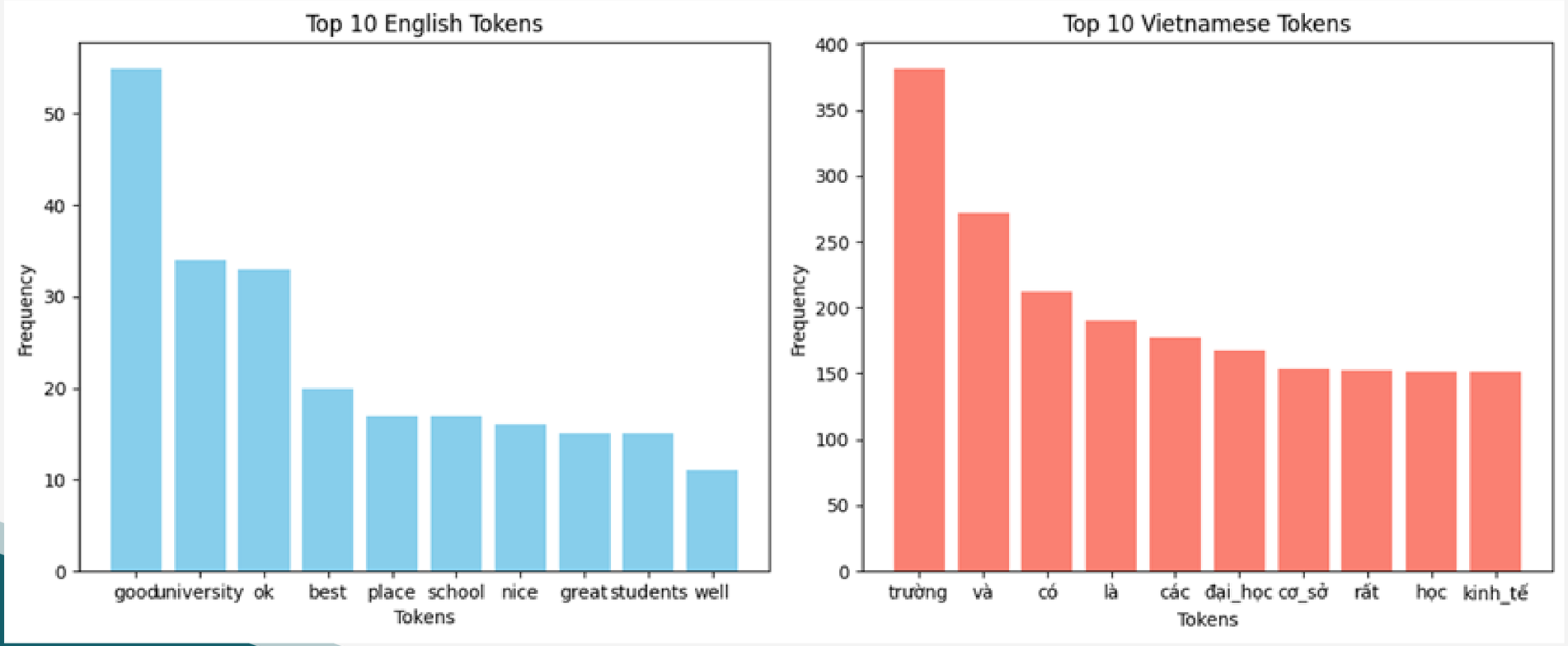
UEH UNIVERSITY College of Technology and Design

```python
def create_data_tuples(df):
    data_tuples = []
    for index, row in df.iterrows():
        # Extract the list of words
        word_list = row['tokens']

        # Create a feature dictionary
        feature_dict = {word: True for word in word_list}

        data_tuples.append((feature_dict, row['sentiment']))
    return data_tuples
```

```python
[ ]  def wt_predicted(sentence, lang):
        if lang == 'en':
            tokens = word_tokenize(sentence)
        else:
            tokens = ViTokenizer.tokenize(sentence).split()
        return dict([(token, True) for token in tokens])
```

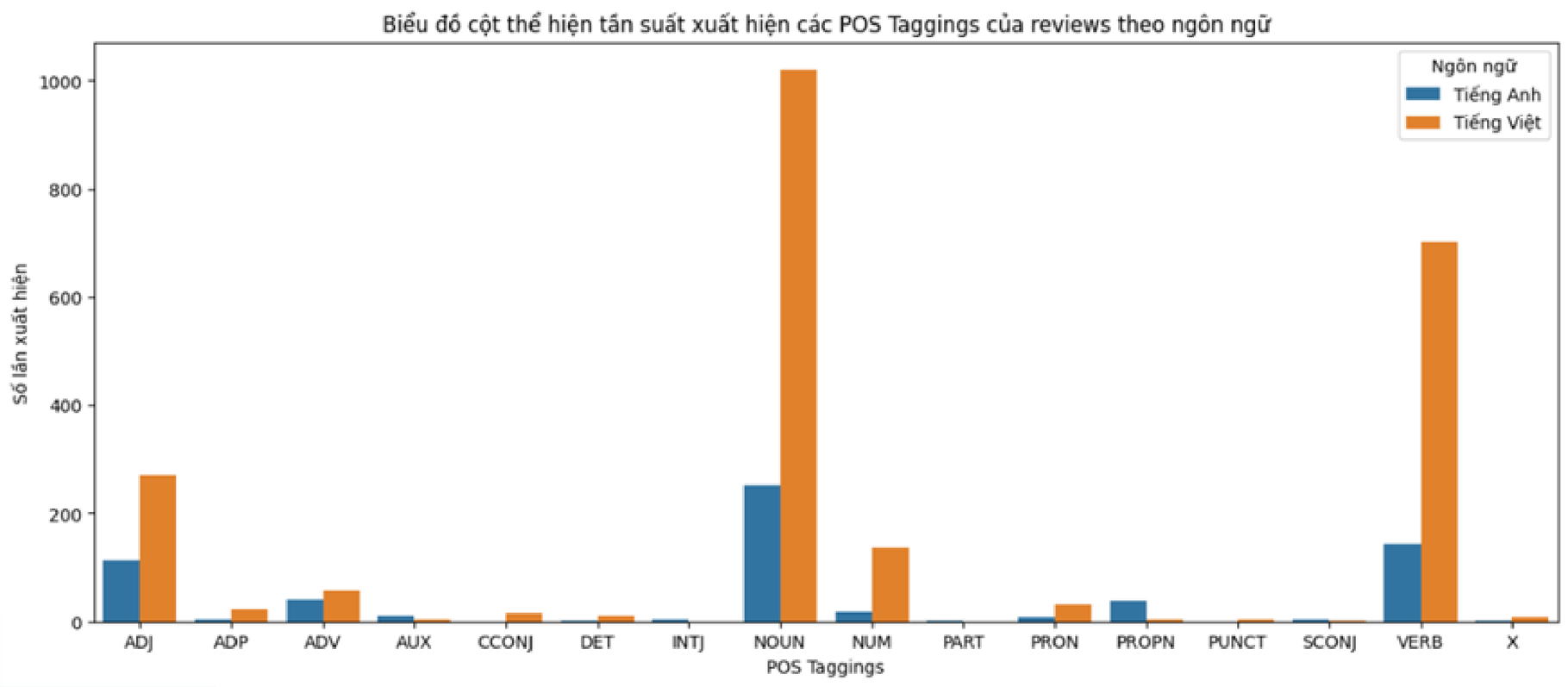
UEH UNIVERSITY
College of Technology and Design

Tokenize & POS tagging



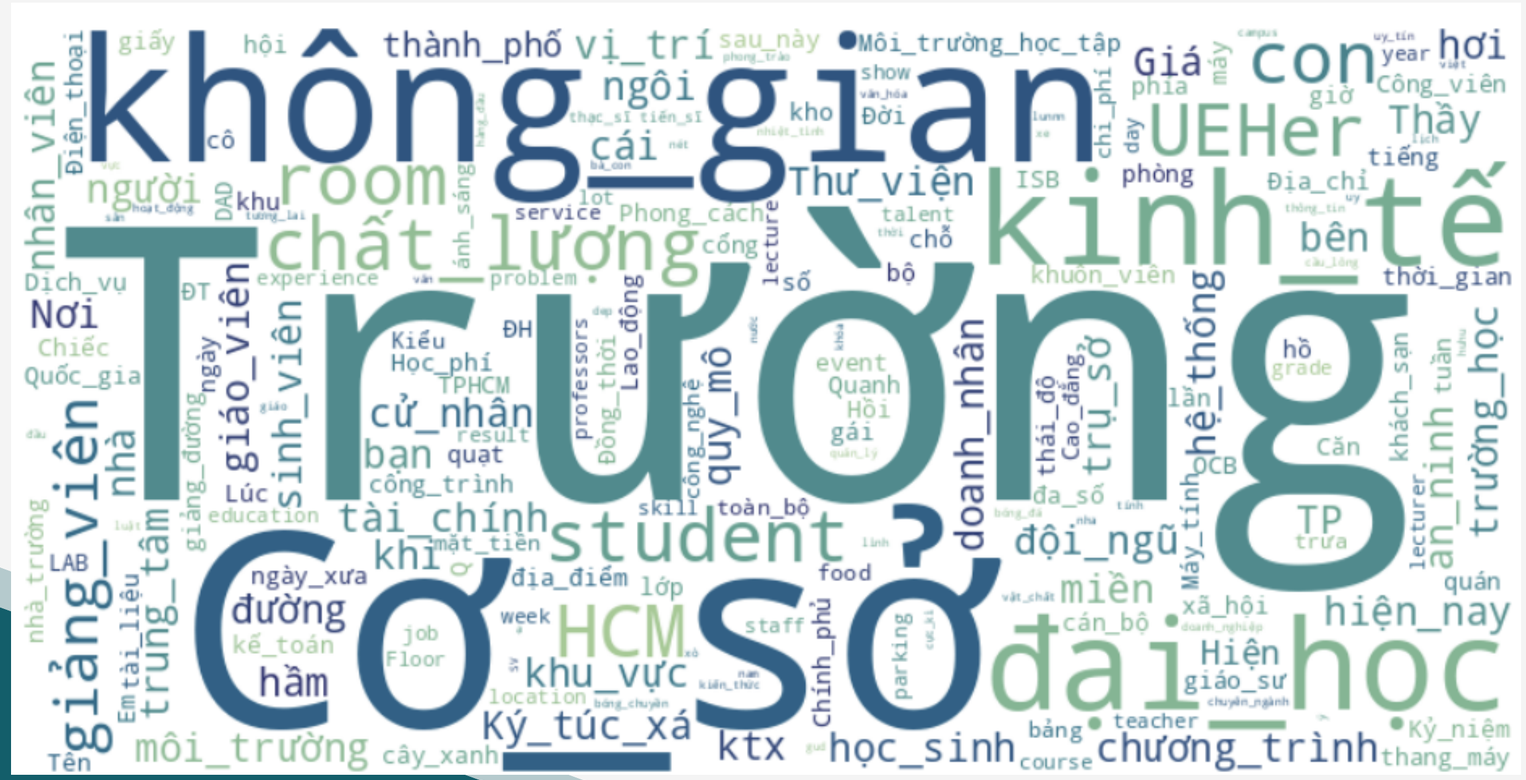
UEH UNIVERSITY
College of Technology and Design

Tokenize & POS tagging



UEH UNIVERSITY
College of Technology and Design

Tokenize & POS tagging



- Giảng viên
- Cơ sở vật chất
- Hoạt động sinh viên

UEH UNIVERSITY College of Technology and Design

Mô hình Naive Bayes (nltk.NaiveBayesClassifier)

```
[ ]  # Train Naive Bayes Classifier
     classifier = NaiveBayesClassifier.train(train_data)
```

```
     time: 68 ms (started: 2023-12-21 06:43:47 +00:00)
```

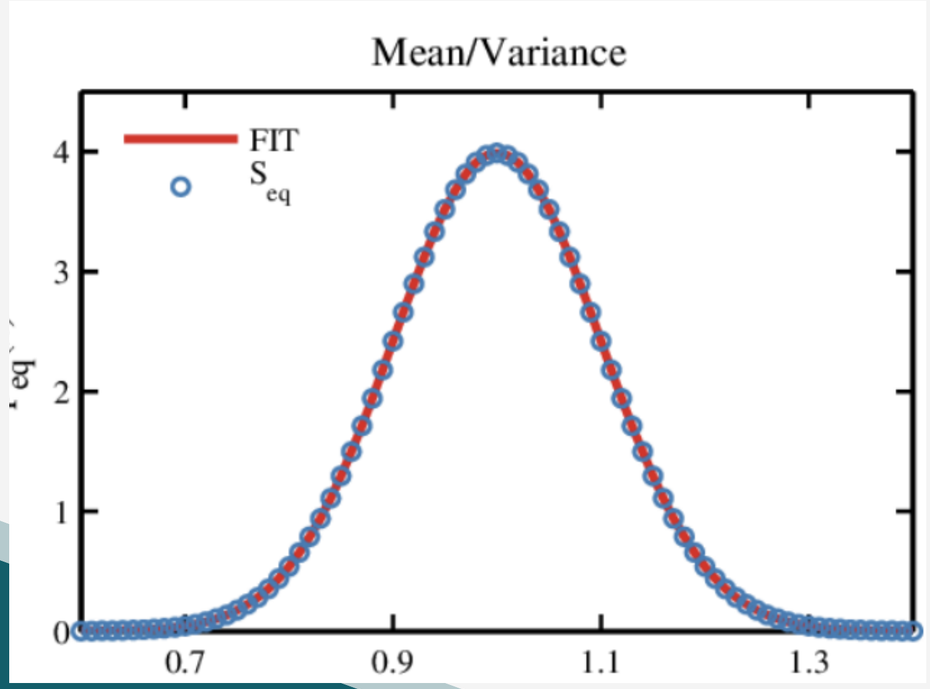
```
[ ]  predicted_labels = [classifier.classify(example[0]) for example in test_data]
     true_labels = [example[1] for example in test_data]
     accuracy = accuracy_score(true_labels, predicted_labels)
     print("Accuracy:", accuracy)
```

```
     Accuracy: 0.8047337278106509
     time: 14.8 ms (started: 2023-12-21 06:43:47 +00:00)
```

UEH UNIVERSITY College of Technology and Design

Mô hình Maxent

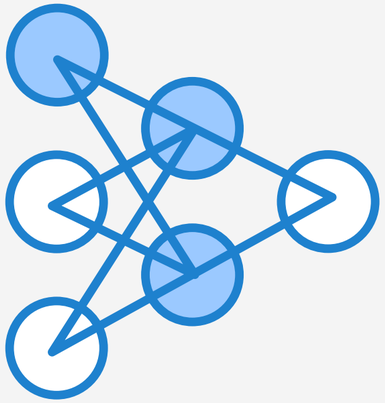
(nltk.MaxentClassifier)

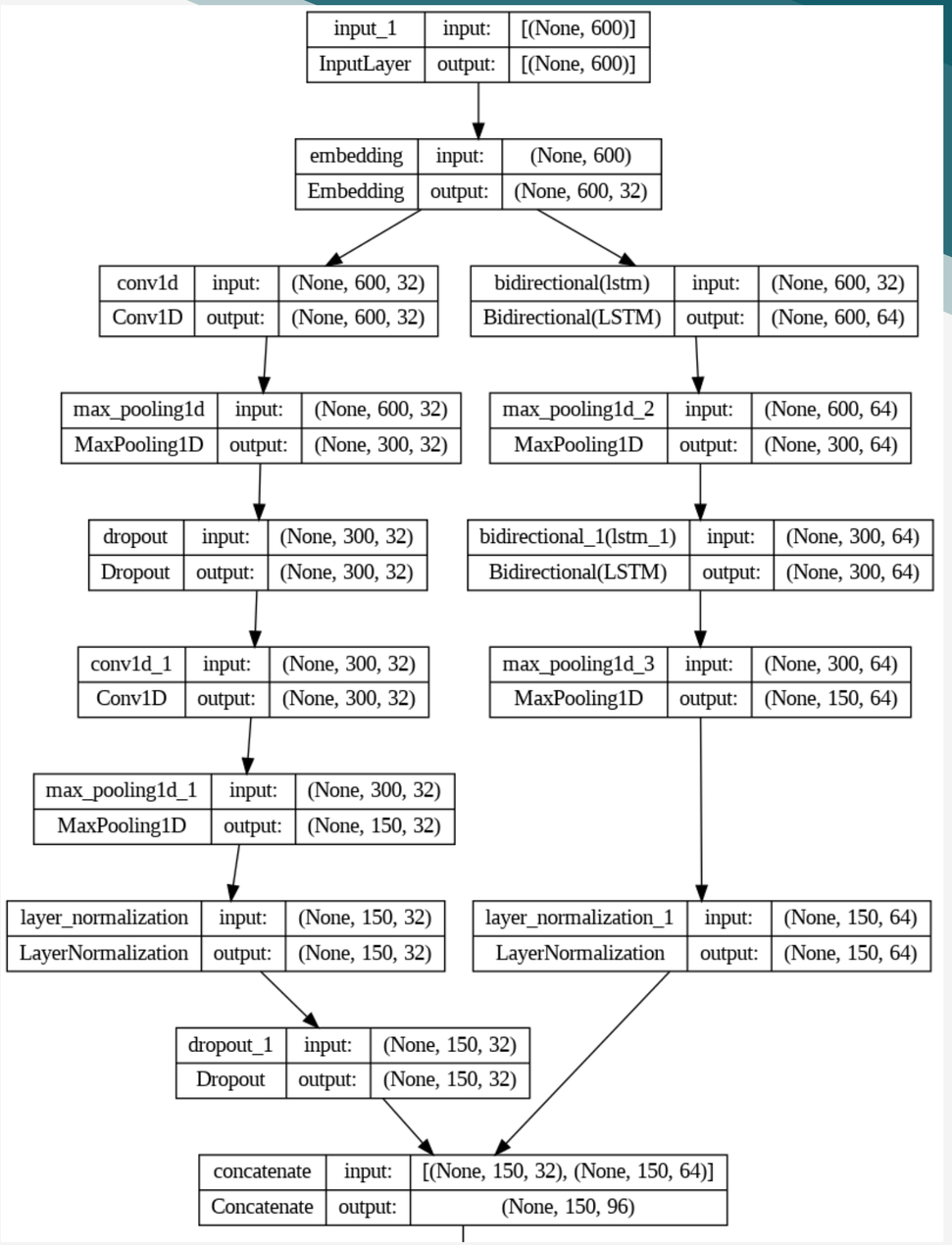



```
==> Training (20 iterations)

    Iteration    Log Likelihood    Accuracy
    ---------------------------------------
           1          -0.69315       0.866
           2          -0.34073       0.900
           3          -0.27419       0.943
           4          -0.24059       0.957
           5          -0.21844       0.967
           6          -0.20195       0.970
           7          -0.18884       0.976
           8          -0.17797       0.978
           9          -0.16871       0.982
          10          -0.16068       0.982
          11          -0.15362       0.982
          12          -0.14734       0.984
          13          -0.14170       0.987
          14          -0.13660       0.991
          15          -0.13195       0.993
          16          -0.12770       0.993
          17          -0.12379       0.993
          18          -0.12017       0.993
          19          -0.11681       0.993
       Final          -0.11369       0.993
```

UEH UNIVERSITY
College of Technology and Design

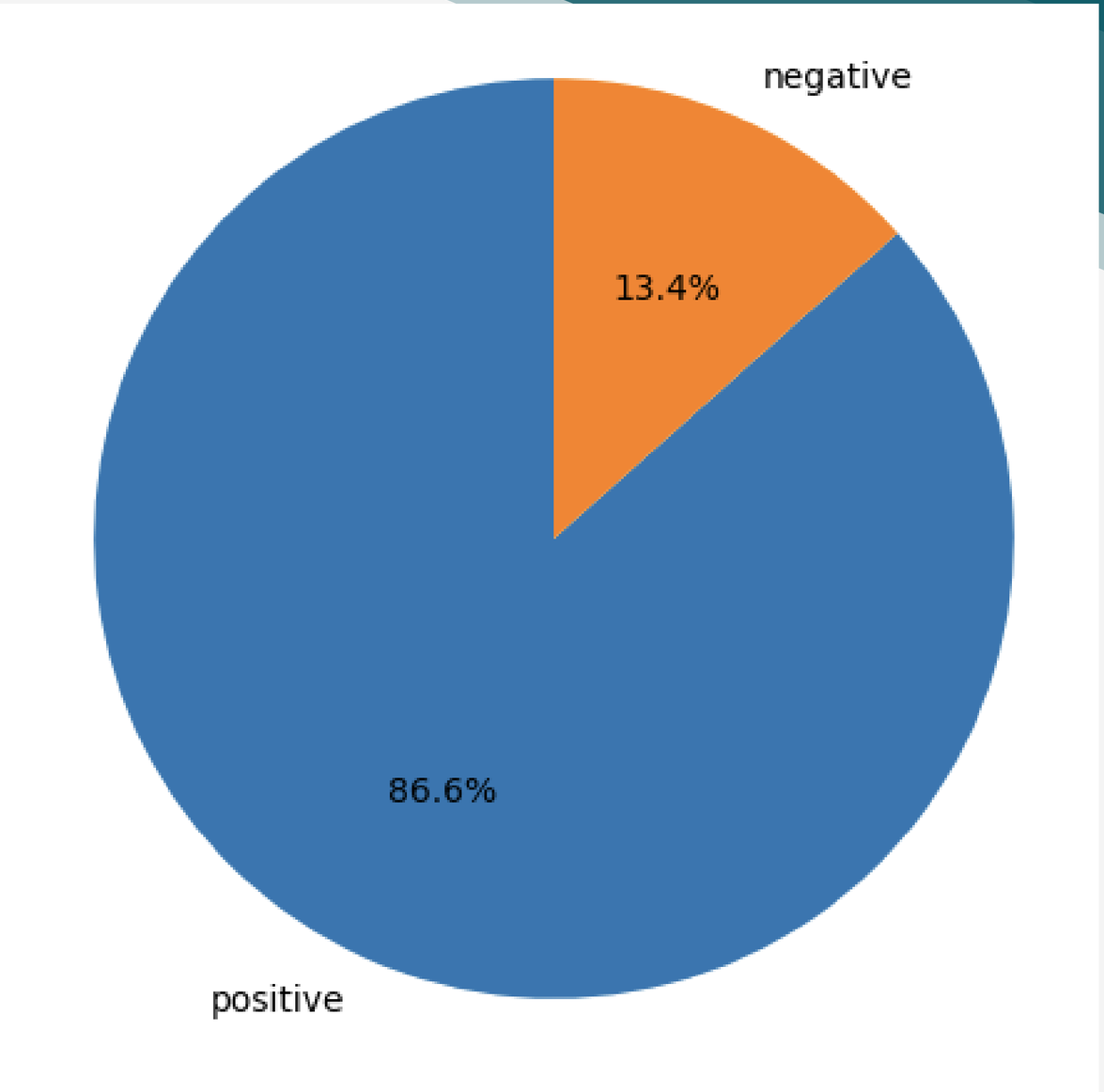
Mô hình CNN kết hợp BiLSTM





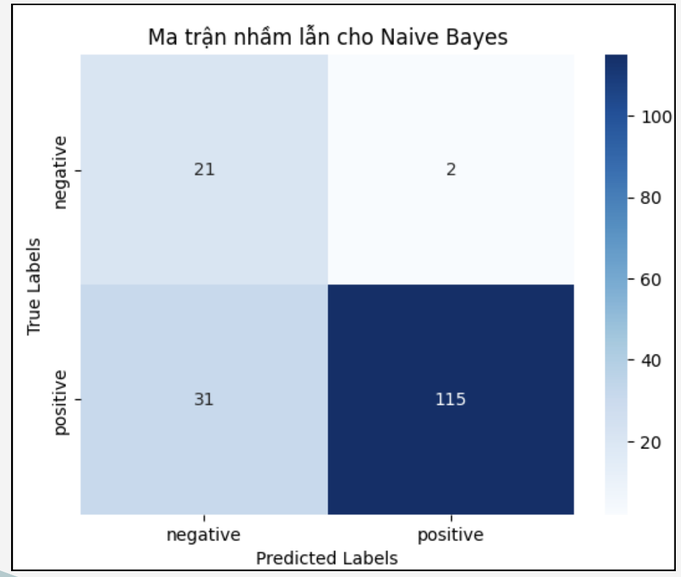
UEH UNIVERSITY
College of Technology and Design

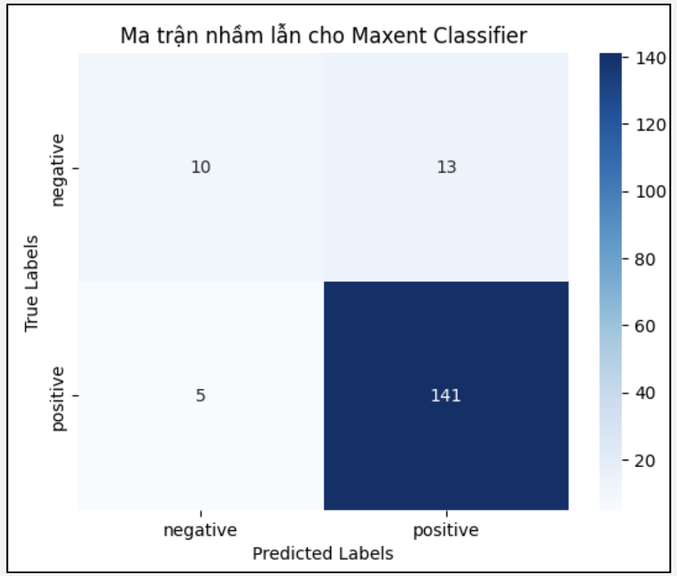
```
positive: 728
negative: 113
```

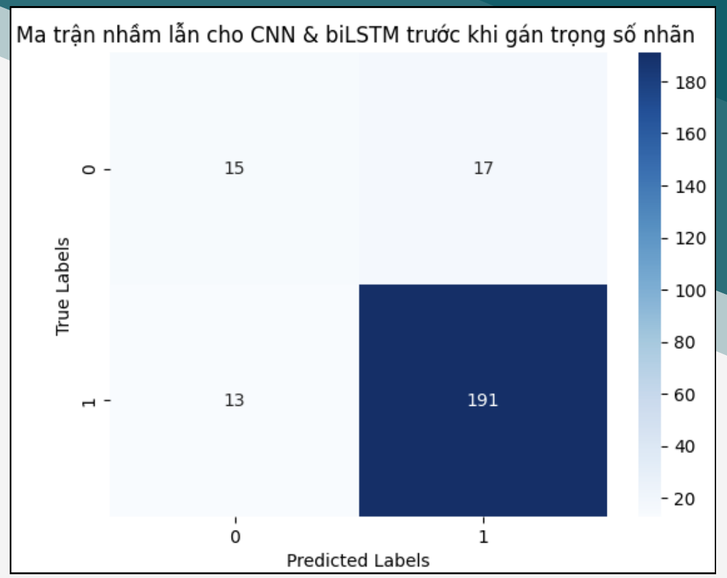


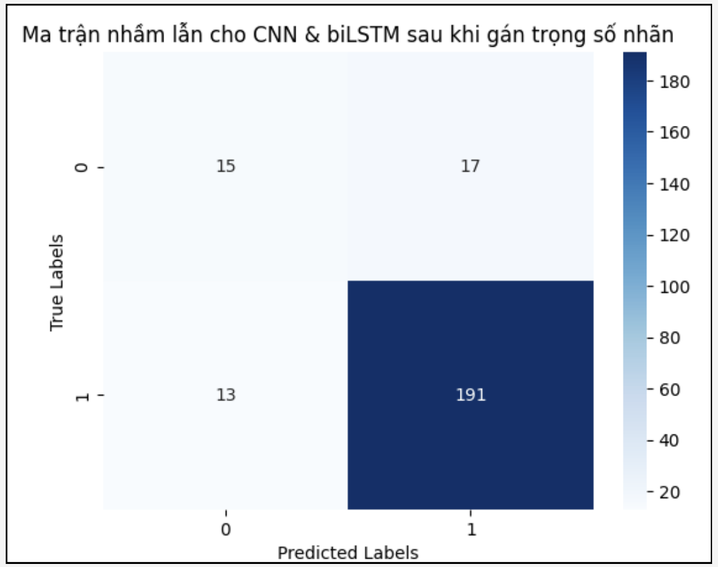
UEH UNIVERSITY
College of Technology and Design

Đánh giá

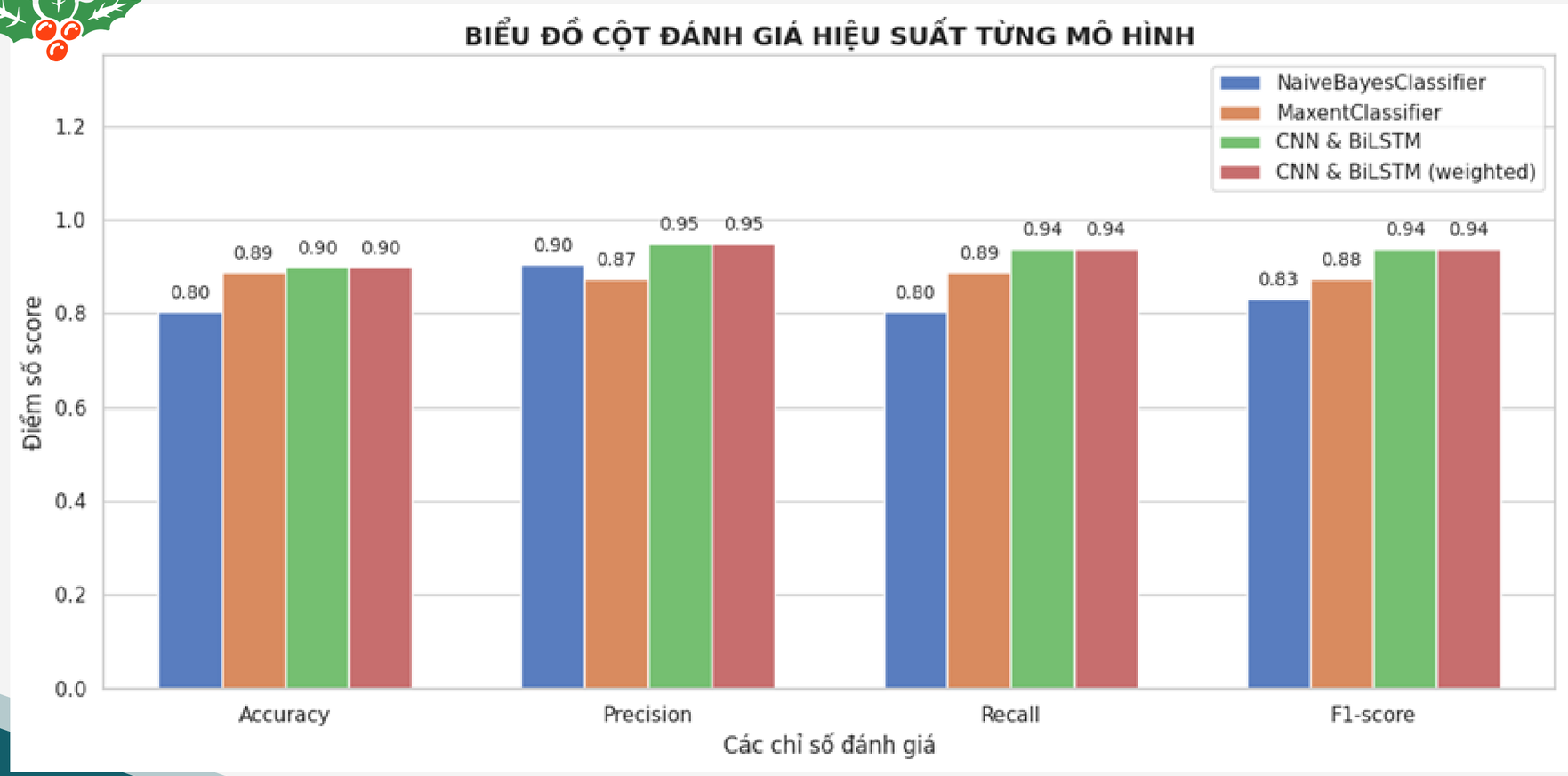








UEH UNIVERSITY
College of Technology and Design



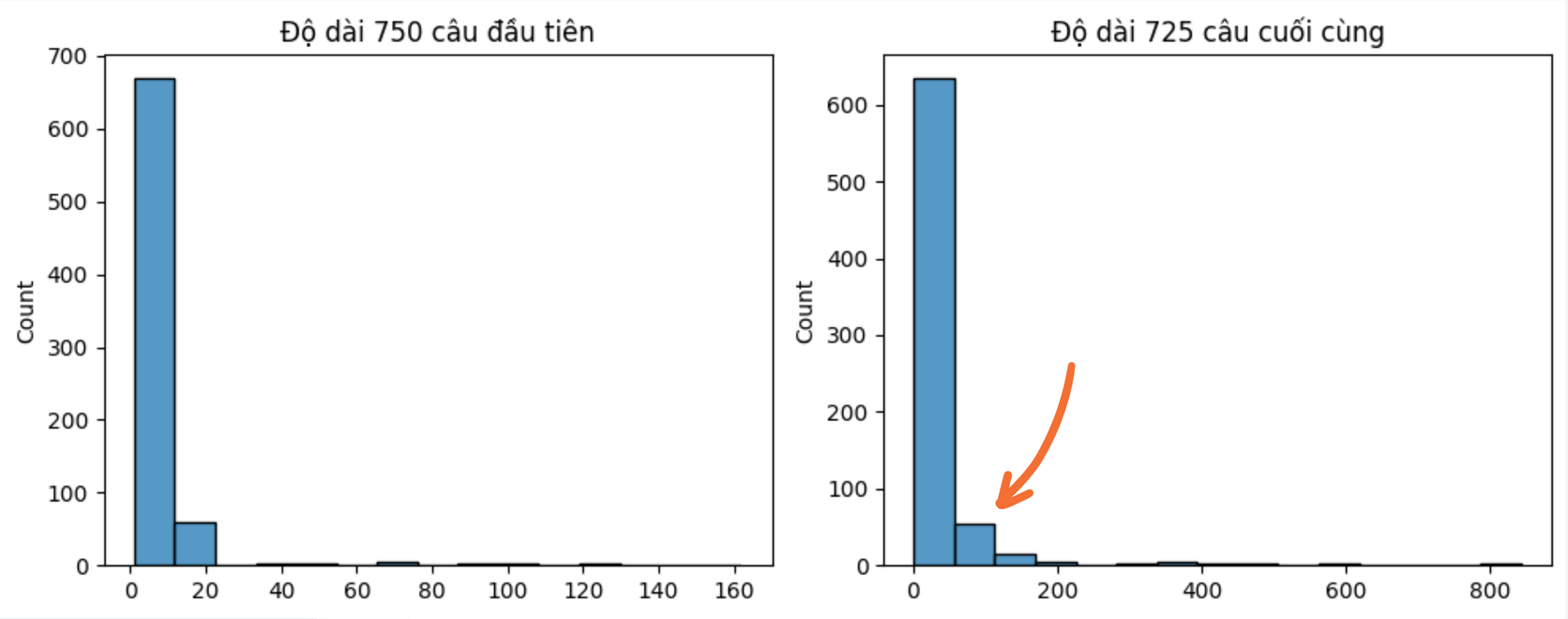
UEH UNIVERSITY College of Technology and Design

		Naïve Bayes	Maxent	CNN & BiLSTM	CNN & BiLSTM (weighted)
Accuracy		0.80	0.89	0.90	0.90
Execution time (giây)		0.068	54.9	268	262
Precision	*1*	0.90	0.87	0.95	0.95
	0	0.40	0.64	0.62	0.63
Recall	*1*	0.80	0.89	0.94	0.94
	0	0.91	0.39	0.66	0.69
F1-score	*1*	0.83	0.88	0.94	0.94
	0	0.56	0.49	0.64	0.66

UEH UNIVERSITY College of Technology and Design

Hạn chế & cải thiện

- Bộ dữ liệu ít quan sát → khó khắc phục imbalanced dataset
- Vectơ embedding chỉnh maxlen quá lớn (600)



UEH UNIVERSITY College of Technology and Design

Demo