ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN

Tính toán đa phương tiện

Học kỳ II: 2019-2020

Lớp: CS232.K21

Đề tài: Xây dựng một hệ thống phân tích cảm xúc trong một câu bình luận

Thành viên: Nguyễn Lê Hoàng Hùng 18520285

Nguyễn Hữu Hoàng 18520283

Phan Phát Huy 18520287

Giảng viên hướng dẫn: Mai Tiến Dũng

Thành phố Hồ Chí Minh, ngày 05 , tháng 08 , Năm 2020

Mục lục

I. Lời mở đầu

1. Vấn đề đặt ra

Trong thời đại công nghệ thông tin tiên tiến hiện nay cùng với việc đa số mọi nhà đều có chiếc máy tính kết nối internet. Điều đó đã tạo ra cơ hội cho việc phát triển một thị trường online, ở trên đó bạn có thể tìm để mua mọi thứ dù nơi bán cách chỗ bạn nửa vòng trái đất. Nhưng đi kèm với việc buôn bán xa đó thì phản hồi của khách hàng về sản phẩm đó cũng được thực hiện trên mạng internet. Số lượng nhận xét của khách hàng là một con số rất lớn. Con người sẽ rất tốn thời gian lẫn sức lực để có thể xử lý tay hết tất cả dữ liệu nhận xét của khách hàng. Vì vậy nhóm em đã quyết định chọn vấn đề "Phân tích cảm xúc của nhận xét sản phẩm của khách hàng trên trang web của Thế Giới Di Động".

2. Mục tiêu

Đồ án tập trung vào việc tìm hiểu các bước cơ sở để xây dựng nên một hệ thống ứng dụng vào bài toán phân lớp quan điểm cho nhận xét của người dùng dựa trên dữ liệu thu nhập được. Trong đồ án, nhóm em đã lựa chọn bộ phân lớp Naïve Bayes để cài đặt và thử nghiệm.

II. Bài toán phân lớp cảm xúc và hướng tiếp cận

1. Bài toán phân tích cảm xúc

Phân tích cảm xúc (sentiment analysis) là một lĩnh vực khó khăn trong vấn đề xử lý ngôn ngữ tự nhiên, nó nghiên cứu về thái độ và cảm xúc của mọi người về một đối tượng nào đó. Các đối tượng có thể là người, vật, sự việc, một dịch vụ bất kì. Sentiment analysis (SA) là một lĩnh vực nóng bỏng, được nhiều người quan tâm bởi vì các yếu tố sau.

Một, là sự ứng dụng rộng rãi của nó vào nhiều lĩnh vực trong đời sống. Ví dụ như trong việc kinh doanh, phân tích và nắm bắt được cảm xúc quan điểm của khách hàng có thể tạo ra các góp ý cải thiện và phát triển sản phẩm để tốt hơn. Hoặc trong giáo dục, khi có một sự thay đổi mới nào thì việc phân tích để nắm bắt sự đồng ý hay bất đồng ý trong các

ý kiến của học sinh, giáo viên, phụ huynh sẽ giúp cho Sở giáo dục điều chỉnh lại hệ thống giáo dục để tốt hơn.

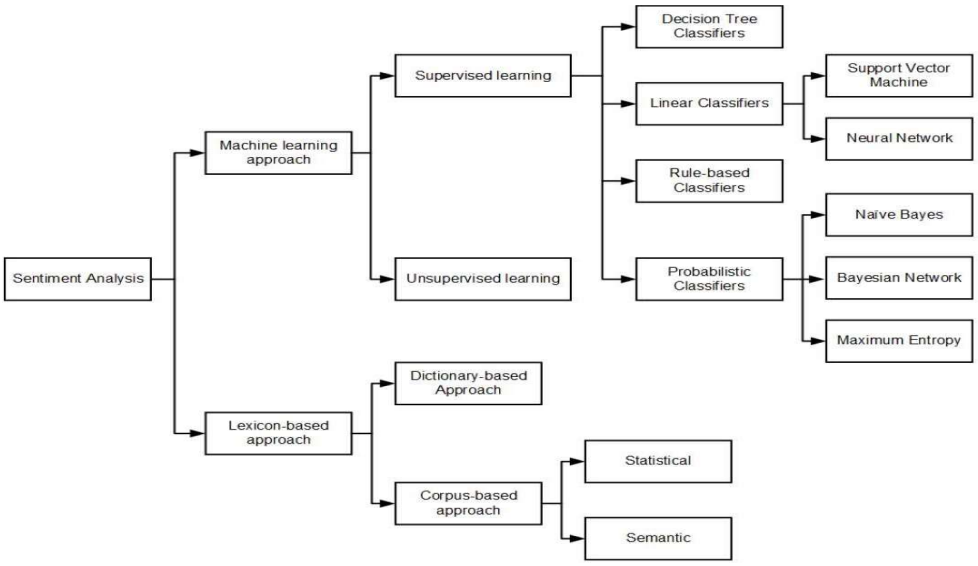
Hai, đó là sự thông dụng của internet và bùng nổ dữ liệu mạng xã hội. Trong quá trình phát triển của loài người, đây là thời điểm mà lượng dữ liệu, thông tin trên mạng đang ngày càng tăng trưởng với tốc độ chóng mặt. Chỉ với facebook mà đã có 33,86 triệu người dùng ở Việt Nam vào năm 2018, mỗi người một giờ đăng một bài viết lên tường của họ thì một ngày sẽ được hơn 8 tỷ bài viết.

Ba, đó là sự hấp dẫn của kiến thức. Sentiment Analysis có thể được chia nhỏ thành các bài toán con như phân lớp chủ quan và khách quan, phân lớp ý kiến trái chiều, phát hiện ý kiến rác, …

Cảm xúc mà bài toán nhóm em thực hiện là có hai loại: tích cực (positive) và tiêu cực (negative).

2. Hướng tiếp cận

Để giải quyết bài toán trên có rất nhiều kỹ thuật phân loại được phát triển và công bố cho mọi người, như hình sau:



Hình 1. Các kỹ thuật cho bài toán Sentiment Analysis

Trong đó, có hai hướng giải quyết chính là sử dụng thuật toán học có giám sát và không giám sát cộng với các kiến thức về từ vựng và ngữ nghĩa. Trong đó lại được chia thành các thuật toán nhỏ khác. Ngoài ra, với sự xuất hiện thành công của Deep Learning đã được ứng dụng vào bài toán phân tích cảm xúc.

Các thuật toán học có giám sát được áp dụng trong Sentiment Analysis là; Naïve Bayes, Maxinum Entropy và SV. Các thuật toán này được đánh giá cao về tính xác và hiệu quả trong giải quyết bài toán phân lớp quan điểm người dùng.

Tổng quát, các bước để xây dựng nên một bộ phân lớp gồm có:

- Bước 1: Tiền xử lý dữ liệu, trong bước này ta sẽ data cleaning và data normalization để làm đầu cho bước tiếp theo.
- Bước 2: Trích chọn đặc trưng và rút gọn đặc trưng (mục tiêu giảm độ phức tạp).
- Bước 3: Xây dựng mô hình học
- Bước 4: Phân lớp

Các thuật toán học máy có giám sát đều có sườn giải thuật chung như sau:

Đầu vào:

- d: tài liệu cần phân loại
- c: tập label. Trong bài toán của nhóm em là c= {tích cực, tiêu cực}
- Tập dữ liệu huấn luyện với các tài liệu được gán nhãn, phân loại thủ công.

Đầu ra: Bộ phân lớp đã học xong.

Sau khi suy nghĩ nhóm em quyết định chọn giải thuật Naïve Bayes để làm mô hình phân lớp cho bài toán vì nó sử dụng định luật thống kê Naïve Bayes chúng em mới vừa học để giải quyết vấn đề bài toán.

3. Mô hình phân lớp Naïve Bayes

Bộ phân lớp quan điểm Naïve Bayes được xây dựng dựa trên lý thuyết Bayes về xác suất có điều kiện và sử dụng mô hình "bag of words" để phân loại văn bản:

$$P(c|d) = P(c).\frac{P(d|c)}{P(d)}$$

Mục tiêu là tìm được phân lớp c* sao cho P(c*|d) là lớn nhất hay xác suất của tài liệu d thuộc lớp c* là lớn nhất.

Từ công thức trên ta có thể nhận thấy P(d) không đóng vai trò gì trong việc quyết định phân lớp c nên $P(c|d)_{max} \Leftrightarrow P(c).P(d|c)_{max}$.

Để có thể xấp xỉ giá trị của P(d|c), thuật toán Naïve Bayes giả sử rằng: các vector đặc trưng fi của một tài liệu khi đã biết phân lớp là độc lập với nhau. Từ đó ta có công thức:

$$P(c|d)\ max = argmax\ P(c).P(f1, f2,\ldots,fn|c)$$

$$\Leftrightarrow P(c|d)\ max = argmax\ P(c).\prod_{1SiSn} P(f_j|c)$$

$$\Leftrightarrow P(c|d)\ max = argmax(\log P(c) + \sum_{1SiSn} \log P(f_i|c))$$

Trong đó f là các vector đặc trưng cho tài liệu d.

Khi tiến hành huấn luyện, thuật toán sử dụng phương pháp xấp xỉ hợp lý cực đại MLE (Maximum Likelihood Estimation) để xấp xỉ P(c) và P(fi|c) cùng thuật toán làm mịn add-one (add-one smoothing). Ta có:

$$P(c) = \frac{N_c}{N}$$

Trong đó Nc là số văn bản được phân loại vào lớp c; N là tổng số văn bản trong tập huấn luyện.

$$P(f_i|c) = \frac{N_{cf_i}}{\sum_{f \in F} N_{cf}}$$

Trong đó N_{cf_i} là số lần xuất hiện của vector đặc trưng i trong tài liệu thuộc phân lớp c.

Đánh giá bộ phân lớp sử dụng thuật toán học máy Naive Bayes, ta nhận thấy phương pháp này các ưu điểm như: đơn giản, dễ cài đặt, bộ phân lớp chạy nhanh và cần ít bộ nhớ lưu trữ. Bộ phân lớp cũng không cần nhiều dữ liệu huấn luyện để xấp xỉ được bộ tham số. Tuy nhiên, bộ phân lớp này có nhược điểm là thiếu chính xác do giả thiết độc lập của các vector đặc trưng khi đã biết phân lớp là không có thực trong thực tế.

III. Xây dựng mô hình phân lớp

1. Tạo bộ dữ liệu

Bước 1: Thu nhập dữ liệu phản hồi đánh giá của khách hàng trên trang thegioididong.com. Kết quả thu được 11796 mẫu dữ liệu khác nhau.

Bước 2: Tiền xử lý dữ liệu bằng cách làm sạch dữ liệu và chuẩn hóa tất cả về dạng chữ thường.

Bước 3: Nhận dạng thủ công từ câu trong bộ dữ liệu và gán nhãn tiêu cực (0) hoặc tích cực (1) cho chúng.

Bước 4: Chia thành hai tập là training set và test set với tỷ lệ 8:2

2. Tiền xử lý dữ liệu

Sau khi thu nhập được dữ liệu ta thực hiện tách từ và gán nhãn từ loại cho chúng thông qua thư viện xử lý là Underthesea. Ví dụ như:

```
s = "Tôi mua một quyển vở và hai cái bút."
print (word_tokenize("tôi mua một chiếc xe hơi")) #Tách từ
print (pos_tag(s)) #Gán nhãn từ loại cho câu
```

```
['tôi', 'mua', 'một', 'chiếc', 'xe hơi']
[('Tôi', 'P'), ('mua', 'V'), ('một', 'M'), ('quyển', 'Nc'), ('vở', 'N'), ('và', 'C'), ('hai', 'M'), ('cái', 'Nc'), ('bút',
'N'), ('.', 'CH')]
```

3. Xây dựng model

a. Lựa chọn đặc trưng

Các đặc trưng cho mô hình phải là các từ thể hiện cảm xúc hoặc ý kiến đánh giá của khách hàng với sản phẩm, thông thường nó nằm ở hai dạng là tính từ hoặc trạng từ + tính từ. Ví dụ như sau:

```
print (pos_tag("Tốt"))
[('Tốt', 'A')]

print(pos_tag("không tốt"))
[('không', 'R'), ('tốt', 'A')]
```

Khi tách từ bằng underthesea, nó sẽ tách các dạng từ trạng từ và tính từ thành hai từ khác nhau nhưng nhược điểm của Naïve Bayes là không thể hiện mối liên hệ giữa các từ với nhau vì thế phải tạo trạng từ + tính từ thành một cụm tính từ mới.

```
words = []
license = ['A','R']
s = word_tokenize(sents,format='text').split(" ")
for i in range(len(s)):
    if ('A' == pos_tag(s[i])[0][1] and s[i] not in words):
            if (s[i] not in stopword):
                words.append(s[i])
    elif (i<len(s)-1 and 'R' == pos_tag(s[i])[0][1] and 'A' == pos_tag(s[i+1])[0][1] and (s[i]+'_'+s[i+1]) not in words):
            if (s[i]+' '+s[i+1]) not in stopword:
                words.append(s[i]+'_'+s[i+1])
return ' '.join(words)
```

b. Cài đặt thuật toán

Chúng em thực hiện cài đặt bộ phân lớp sử dụng hệ điều hành Window 10 và ngôn ngữ lập trình Python với công cục là Jupyte Notebook (Anaconda3)

4. Kết quả thử nghiệm

a. Các thông số đánh giá

Thực hiện đánh giá dựa trên bộ ba tiêu chí đánh giá sau:

- Precision = $\frac{\text{Số thực thể phân loại đúng}}{\text{Tổng cố thực thể đã phân loại}}$
- Recall = $\frac{\text{Số thực phân loại đúng}}{\text{Tổng cố thực thể đúng trong thực tế}}$
- F1-score = 2. $\frac{\text{Precicion} . \text{Recall}}{\text{Precicion} + \text{Recall}}$

b. Kết quả

```
Accuracy: 0.768
Precision_score: 0.750
Recall_score: 0.698
F1_score: 0.711
```

5. Đánh giá kết quả

Bộ phân Naïve bayes cho kết quả chưa cao chỉ tầm khoảng 71%. Kết quả này do một số lý do như sau:

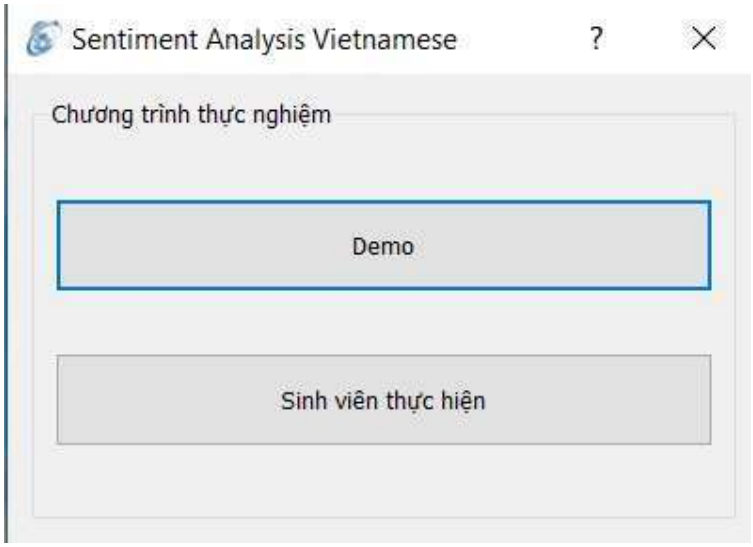
- Sự cách biệt của số lượng giữa hai nhãn trong dữ liệu cào về.
- Tập dữ liệu chưa được gán nhãn tay một cách chính xác.
- Chưa phân tích kĩ về phần cấu trúc mạng ngữ nghĩa để xác định từ vựng làm feature cho mô hình phân lớp tốt hơn.

IV. Xây dựng chương trình demo

1. Lựa chọn công cụ xây dựng

Sử dụng ngôn ngữ Python với thư viện hỗ trợ lập trình tạo GUI là PyQt5 để thiết kế giao diện demo cùng với sự hỗ trợ của công cụ Pycharm.

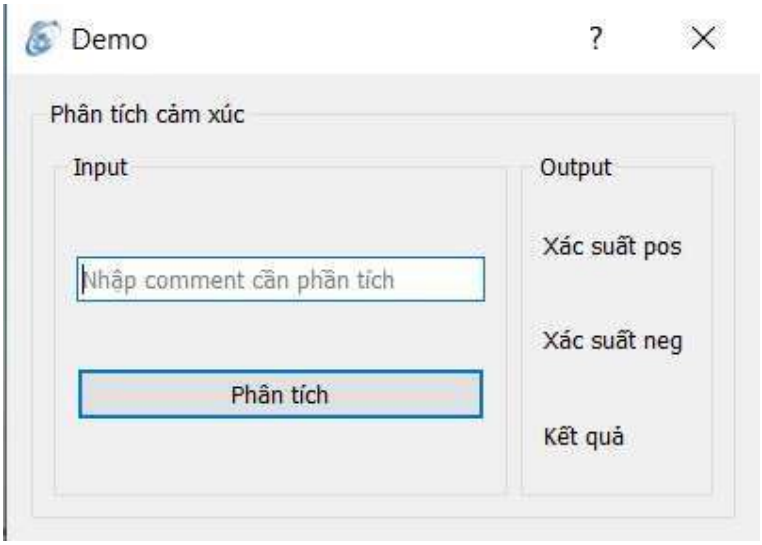
2. Giới thiệu khái quát



Đây là giao diện chính gồm có hai chứa năng là **Demo** và **Sinh viên thực hiện**.

a. Demo

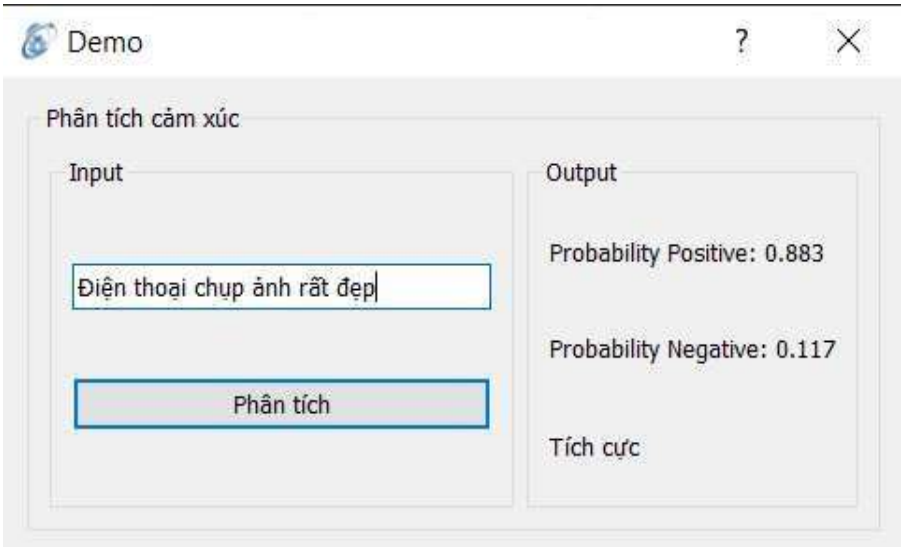
Đây là giao diện chương trình demo

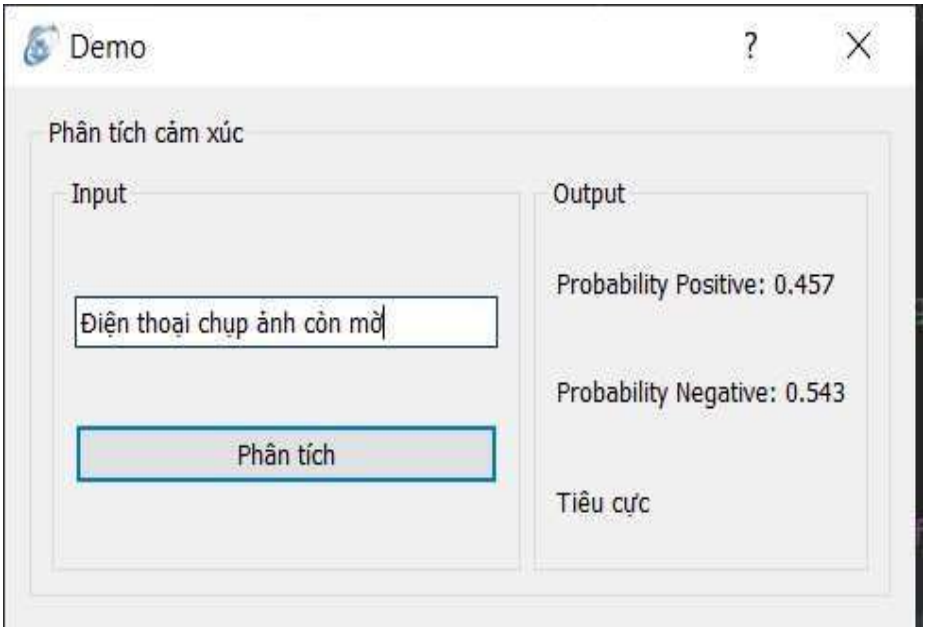


Để nhập dữ liệu cần phân tích ta nhập vào khu vực bên dưới, sau đó nhấn vào nút phân tích:

Sau đó kết quả sẽ trả về trong phần Output. Trường kết quả gồm có ba phần: Xác suất pos là xác suất dữ liệu thuộc vào lớp tích cực, tương tự xác suất neg là xác suất dữ liệu thuộc vào lớp tiêu cực. Kết quả sẽ thông báo cho ta biết dữ liệu ta phân tích thuộc lớp tích cực hay tiêu cực.

Ví dụ:





b. Sinh viên thực hiện