

**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**  
**VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**



**BÁO CÁO MÔN HỌC PROJECT 3**

**SENTIMENT CLASSIFICATION USING MACHINE LEARNING**

Thu thập và Phân loại ý kiến người dùng trên Facebook về mức học phí đại học chính quy của năm học năm học 2020-2021 của trường Đại học Bách khoa Hà Nội

**LÊ CÔNG NGUYỄN**

MSSV: 20173290 - Chuyên ngành Khoa học máy tính

nguyen.lc173290@sis.hust.edu.vn

<b>Giáo viên hướng dẫn:</b>	TS. Phạm Đăng Hải
<b>Bộ môn:</b>	Khoa học máy tính
<b>Viện:</b>	Công nghệ thông tin và truyền thông

**Hà Nội, 11/2020**

## MỤC LỤC

<b>I. GIỚI THIỆU VÀ LỜI CẢM ƠN .....</b>	<b>3</b>
<b>II. PHƯƠNG PHÁP NGHIÊN CỨU .....</b>	<b>4</b>
<b>1. Thu thập dữ liệu và gán nhãn .....</b>	<b>5</b>
a. Thuật toán thu thập dữ liệu .....	5
b. Thư viện Selenium.....	9
c. Thu thập dữ liệu và Gán nhãn .....	10
<b>2. Tiền xử lý dữ liệu và thư viện Underthesea .....</b>	<b>16</b>
<b>3. Biểu diễn dữ liệu bằng TF-IDF .....</b>	<b>16</b>
<b>4. Phân loại ý kiến bằng giải thuật học máy SVM .....</b>	<b>18</b>
a. Quá trình train.....	19
b. Quá trình test .....	20
<b>III.KẾT QUẢ VÀ THẢO LUẬN.....</b>	<b>21</b>
<b>IV.KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN .....</b>	<b>22</b>
<b>V. TÀI LIỆU THAM KHẢO.....</b>	<b>23</b>

## DANH MỤC HÌNH

Figure 1: Sơ đồ phân loại ý kiến với giải thuật SVM.....	4
Figure 2: Sơ đồ thuật toán thu thập dữ liệu từ các bình luận.....	5
Figure 3: Giao diện đăng nhập của Facebook.....	6
Figure 4: Thông báo của Chrome .....	6
Figure 5: "Lúc khác" .....	7
Figure 6: Xem "Bình luận" .....	7
Figure 7: Xem thêm bình luận .....	8
Figure 8: “Xem thêm” nếu bình luận dài và các “Phản hồi” .....	8
Figure 9: Xem thêm phản hồi của một bình luận .....	8
Figure 10: Ví dụ tìm thông tin element “password”.....	9
Figure 11: Quá trình bấm "Lúc khác" và show "Bình luận" .....	10
Figure 12: Quá trình "Xem thêm các bình luận" .....	11
Figure 13: Quá trình xem các "Phản hồi" .....	11
Figure 14: Quá trình “Xem thêm phản hồi” .....	11
Figure 15: Quá trình "Xem thêm".....	12
Figure 16: Quá trình tìm và lấy phần text của các bình luận.....	12
Figure 17: Dữ liệu thô được thu thập.....	13
Figure 18: Dữ liệu sau khi được gán nhãn.....	14
Figure 19: Tỷ lệ tính chất của bình luận .....	15
Figure 20: Ví dụ tách từ bằng Underthesea .....	16
Figure 21: TF .....	17
Figure 22: IDF.....	17
Figure 23: TF-IDF.....	17
Figure 24: Ví dụ về giải thuật SVM.....	18
Figure 25: Kết quả test trên tập train .....	19
Figure 26: Kết quả phân loại.....	20
Figure 27: Trực quan tập test.....	21

## I. GIỚI THIỆU VÀ LỜI CẢM ƠN

Trong thời đại kỷ nguyên số như hiện nay, bên cạnh các nguồn cung dữ liệu như: E-commerce, Internet of things, Data-intensive experiments (bioinformatics, quantum physics, ect) thì Socical networks là một nguồn cung dữ liệu rất dồi dào và phong phú, trong khi đó: “Data is the new oil”.

Với sự phổ biến ngày càng tăng của các dịch vụ mạng xã hội như Facebook hay Twitter, chúng mang đến nhiều tiện ích, đáp ứng nhu cầu của người dùng, từ kết nối mọi người, cập nhật thông tin, công việc, học tập, giải trí hay kinh doanh. Dữ liệu mạng xã hội lại càng trở nên quan trọng để khám phá kiến thức về cộng đồng, vốn rất quan trọng trong tội phạm học, khủng bố, sức khỏe cộng đồng và nhiều ứng dụng khác.

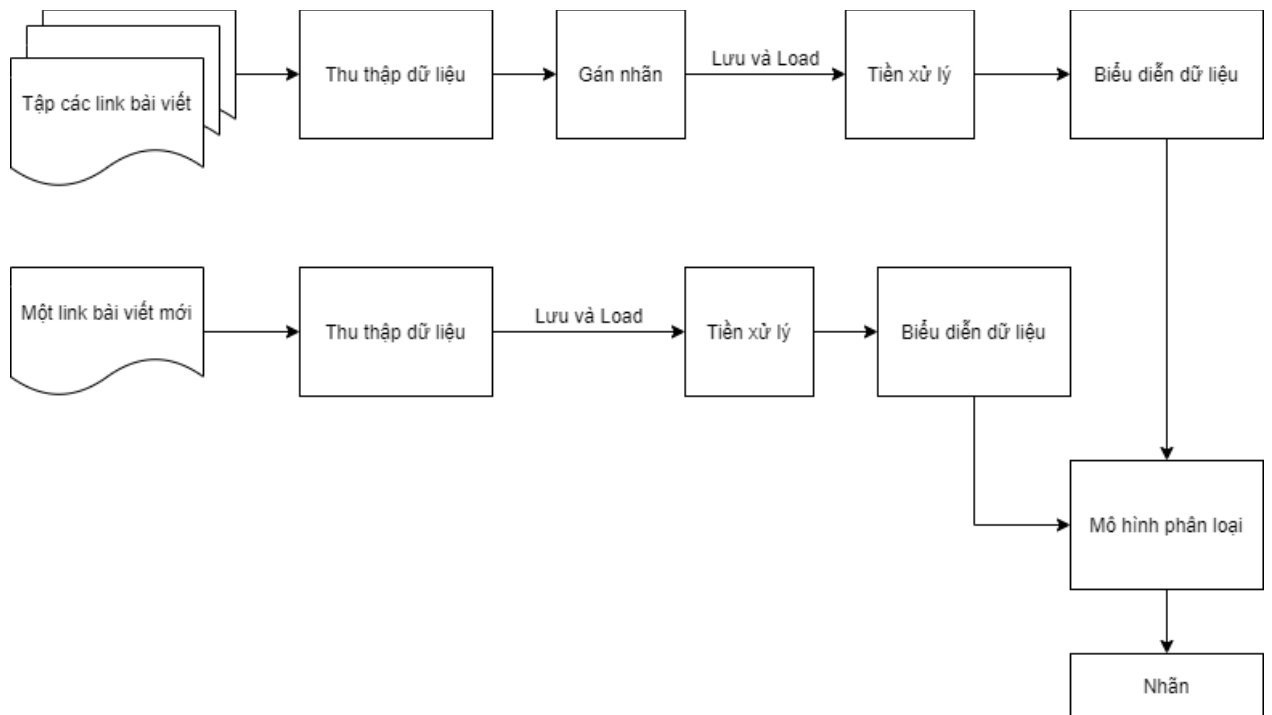
Môi trường của mạng xã hội khá tự do, nơi cảm xúc cá nhân được đề cao, là nơi có thể dễ dàng thu thập những ý kiến của người dùng về một chủ đề nào đó.

Bằng việc sử dụng framework “Selenium”, dữ liệu được thu thập trong đề tài là các ý kiến của người dùng về mức học phí đại học chính quy của năm học năm học 2020-2021 của trường Đại học Bách khoa Hà Nội trên Facebook từ các bình luận của các bài viết cá nhân, bài viết trên fanpage và bài viết trong group.

Dựa trên nội dung của các bình luận, từ đó xử lý và phân loại cho từng bình luận theo hướng đồng ý hay không đồng ý bằng giải thuật học máy “Support Vector Machine” (SVM).

Em xin chân thành cảm ơn sự hướng dẫn của thầy giáo **TS. Phạm Đăng Hải**, bộ môn Khoa học Máy tính, viện Công nghệ thông tin và truyền thông, trường Đại học Bách khoa Hà Nội đã giúp đỡ và tạo điều kiện thuận lợi cho em được nghiên cứu, tìm tòi, phát triển và hoàn thành đề tài nghiên cứu.

## II. PHƯƠNG PHÁP NGHIÊN CỨU



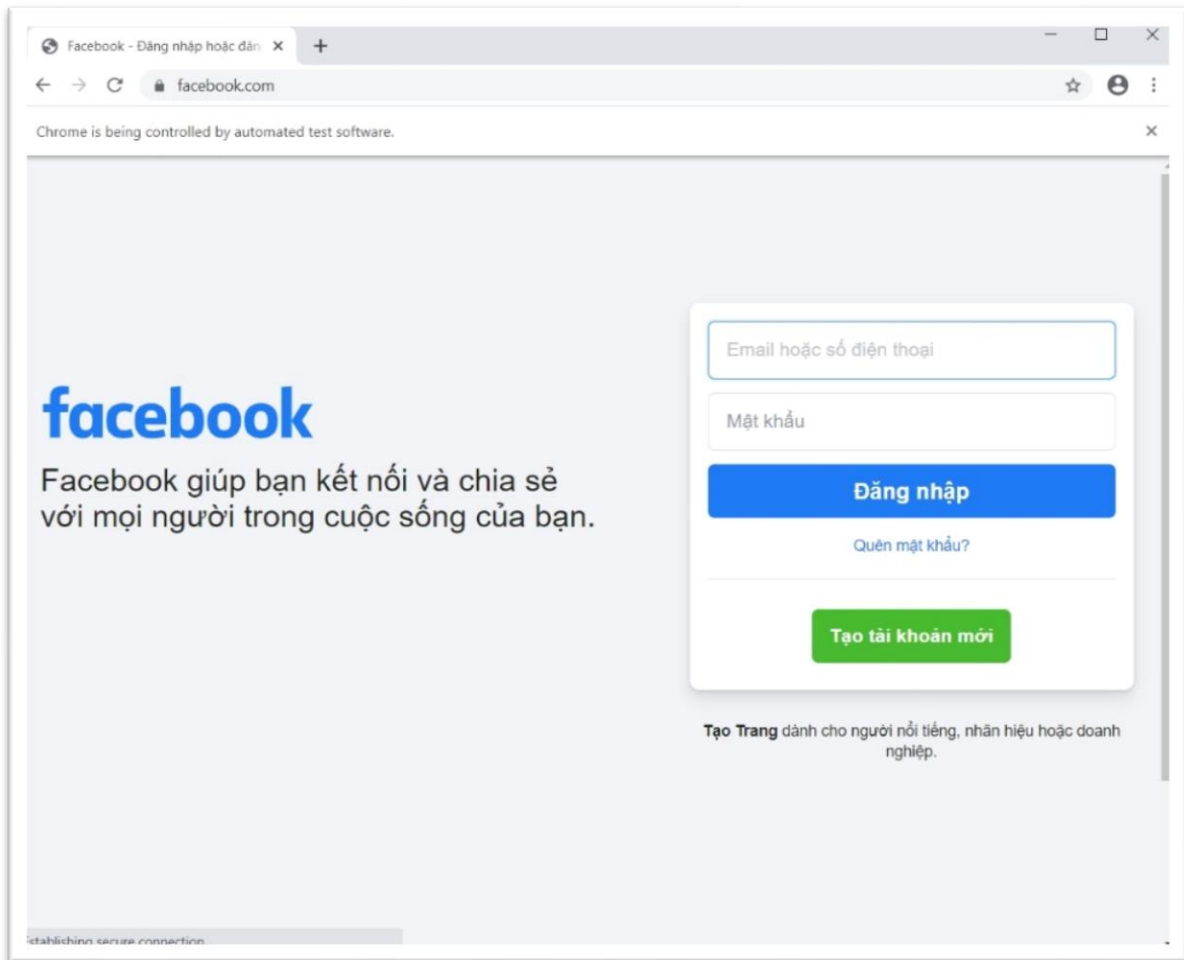
*Figure 1: Sơ đồ phân loại ý kiến với giải thuật SVM*

1. Thu thập dữ liệu và gán nhãn  
a. Thuật toán thu thập dữ liệu

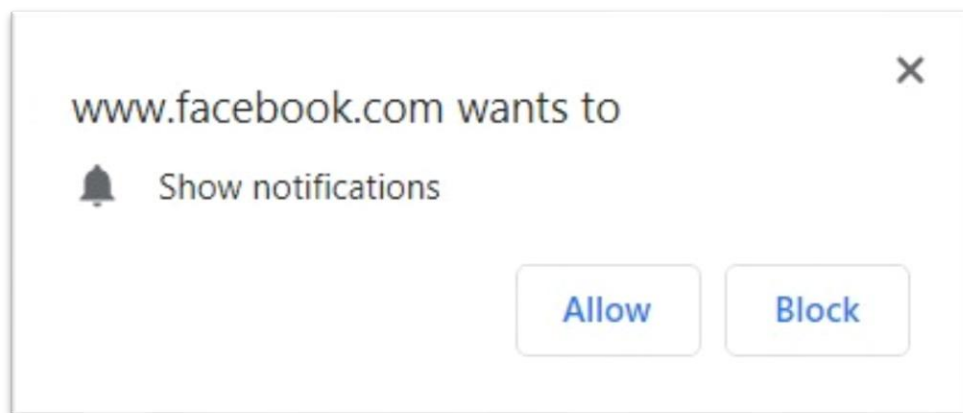
1. **Giả lập** trình duyệt Chrome
2. **Thu thập link** của các bài viết và lưu vào file txt
3. **Truy cập** từng bài viết đến khi hết các link ở file link vừa tạo:
  - a. Nếu là bài viết **cá nhân** hoặc trong **group**:
    - **Đăng nhập** vào Facebook:
    - Mở Facebook
    - Điền thông tin vào ô Username, Password và bấm Enter
    - Tắt thông báo của Chrome
  - b. Nếu là bài viết trên **fanpage**:
    - Bấm "**Lúc khác**"
    - Bấm "**Bình luận**"
  - c. Với mỗi bài viết:
    - Để hiển thị tất cả các bình luận, Bấm vào "**Xem thêm bình luận**" đến hết
    - Bấm "**phản hồi**" dưới mỗi bình luận nếu có
    - Nếu có nhiều phản hồi, bấm "**Xem thêm phản hồi**"
    - Với mỗi bình luận (chính và phản hồi):
      - Nếu bình luận dài, bấm "**Xem thêm**" để hiển thị hết
    - **Tìm** tất cả các bình luận, lấy phần **text**
    - **Ghi** bình luận vào file csv
4. **Đóng** trình duyệt

*Figure 2: Sơ đồ thuật toán thu thập dữ liệu từ các bình luận*

- **Giải thích:**



*Figure 3: Giao diện đăng nhập của Facebook*



*Figure 4: Thông báo của Chrome*

Xem thêm về Hội Sinh viên ĐH Bách khoa Hà Nội trên...

caoquocdat06@gmail.com

.....

**Đăng nhập**

Quên tài khoản?

hoặc

**Tạo tài khoản mới**

Lúc khác

Figure 5: "Lúc khác"

**GHI NHẬN PHẢN HỒI CỦA SINH VIÊN**

#HSVBKHN

👍 😄 ❤️ 1,7K

811 bình luận 77 lượt chia sẻ

🔗 Chia sẻ

Figure 6: Xem "Bình luận"



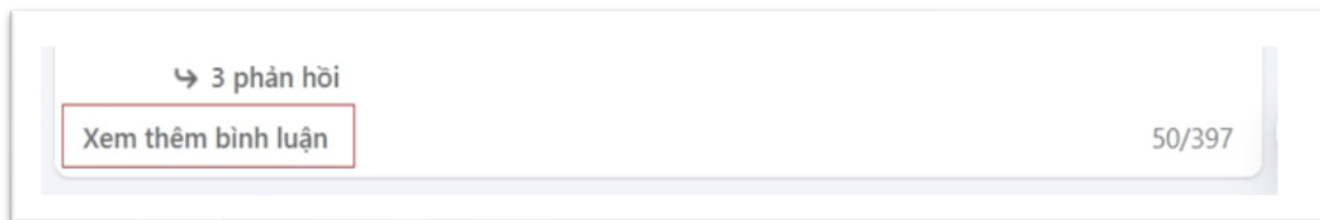


Figure 7: Xem thêm bình luận



Figure 8: “Xem thêm” nếu bình luận dài và các “Phản hồi”



Figure 9: Xem thêm phản hồi của một bình luận

## b. Thư viện Selenium

Selenium là một **Automates browsers** (trình duyệt web tự động), có khả năng như một web browser bình thường, nhưng được kết hợp thêm khả năng **thao tác và tùy chỉnh gần như mọi thứ** (gửi phím bấm của bàn phím, click chuột) với web browser này thông qua lớp có WebDriver kế thừa từ *Interface IWebDriver*.

Các WebDriver này có thể là: ChromeDriver, FirefoxDriver... Để tài sử dụng **ChromeDriver**.

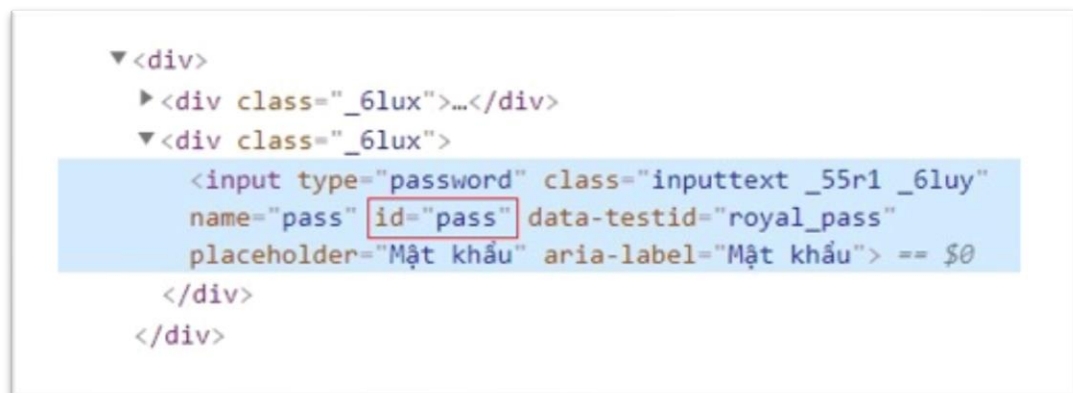
Hàm *sleep(random)* được sử dụng để dừng chương trình một thời gian để load thông tin và để Facebook không nghi ngờ rằng ta đang sử dụng “máy”.

Nhược điểm của *Facebook Graph API*:

- Không có tính mềm dẻo khi muốn thu thập dữ liệu từ nhiều trang web khác nhau.
- Phải tìm các thông tin như ID, Token... trong khi Selenium giống như việc chính bạn đang sử dụng facebook hằng ngày.

Cách lấy thông tin bằng Selenium:

- Tại giao diện của Facebook, ấn tổ hợp phím “Ctrl+Shift+I” hoặc ấn “right mouse” và chọn “Inspect”.
- Tìm các “Elements”, ví dụ như element “password” có *id=“pass”*:



```
▼ <div>
  ▶ <div class="_6lux">...</div>
  ▼ <div class="_6lux">
    <input type="password" class="inputtext _55r1 _6luy"
      name="pass" id="pass" data-testid="royal_pass"
      placeholder="Mật khẩu" aria-label="Mật khẩu"> == $0
  </div>
</div>
```

Figure 10: Ví dụ tìm thông tin element “password”

- Sau đó, sử dụng các hàm của Selenium: *find\_element\_by*, *click*, *send\_keys*, *text*...
  - *fill\_password = browser.find\_element\_by\_id('pass')*
  - *fill\_password.send\_keys('123@abc')*
  - *fill\_password.send\_keys(Keys.ENTER)*

### c. Thu thập dữ liệu và Gán nhãn

Em đã thu thập được 9 link để thu thập dữ liệu cho quá trình train và 1 link cho quá trình test. Tất cả thu thập được **gần 4500** bình luận.

Đối với bài viết cá nhân và trong group thì phải đăng nhập, do bài viết có thể không công khai. Và việc thu thập là hoàn toàn tương tự nhau, các đường dẫn *xpath* là hoàn toàn giống nhau.

Đối với các bài viết trên fanpage, sau khi đăng nhập, trang web không chỉ hiển thị bài viết chúng ta cần thu thập dữ liệu mà còn hiển thị các bài viết khác phía dưới. Bài viết chúng ta cần là bài viết đầu tiên. Các đường dẫn *xpath* sẽ khác với hai loại bài viết ở trên. Để dễ dàng thu thập hơn thì không cần đăng nhập nữa vì đa số bài viết trên fanpage là công khai, khi đó, trang web chỉ hiển thị bài viết mà chúng ta cần.

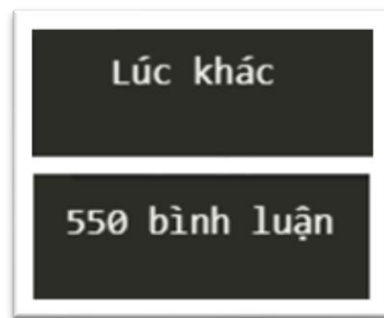
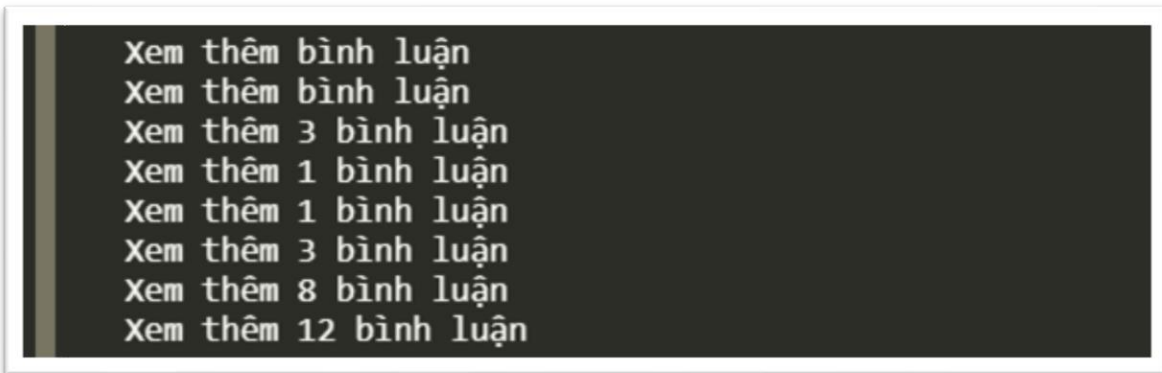
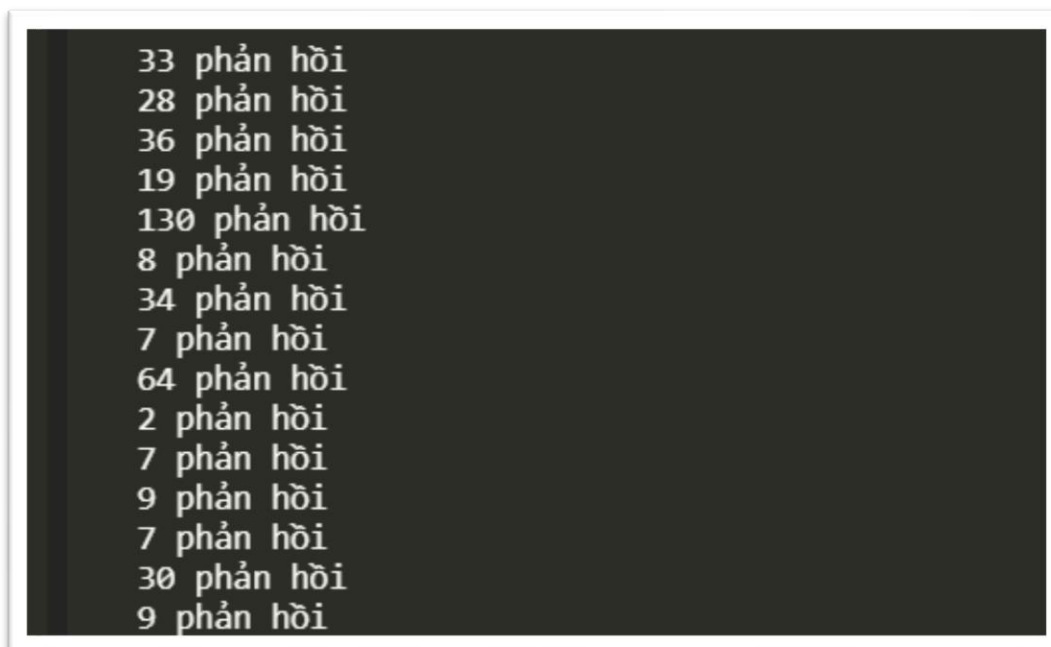


Figure 11: Quá trình bấm "Lúc khác" và show "Bình luận"



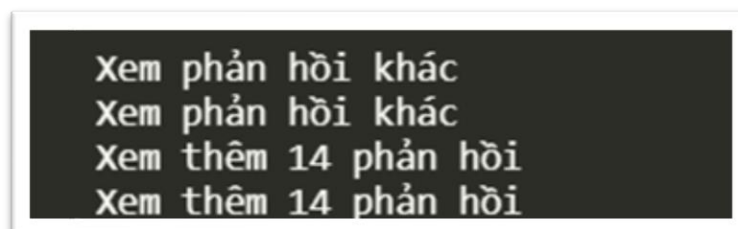
```
Xem thêm bình luận  
Xem thêm bình luận  
Xem thêm 3 bình luận  
Xem thêm 1 bình luận  
Xem thêm 1 bình luận  
Xem thêm 3 bình luận  
Xem thêm 8 bình luận  
Xem thêm 12 bình luận
```

*Figure 12: Quá trình "Xem thêm các bình luận"*



```
33 phản hồi  
28 phản hồi  
36 phản hồi  
19 phản hồi  
130 phản hồi  
8 phản hồi  
34 phản hồi  
7 phản hồi  
64 phản hồi  
2 phản hồi  
7 phản hồi  
9 phản hồi  
7 phản hồi  
30 phản hồi  
9 phản hồi
```

*Figure 13: Quá trình xem các "Phản hồi"*



```
Xem phản hồi khác  
Xem phản hồi khác  
Xem thêm 14 phản hồi  
Xem thêm 14 phản hồi
```

*Figure 14: Quá trình "Xem thêm phản hồi"*

```
Xem thêm
Xem thêm
Xem thêm
Xem thêm
Xem thêm
Xem thêm
Xem thêm
Xem thêm
Xem thêm
Xem thêm
```

Figure 15: Quá trình "Xem thêm"

[Output was trimmed for performance reasons.](#)

[To see the full output set the setting "python.dataScience.textOutputLimit" to 0.](#)

\*\*\*

nhất có quy định quy đổi như vậy, và liệu rằng quy đổi có tương xứng với thời lượng học, thời lượng sử dụng cơ sở vật chất và chất lượng tiếp thu kiến thức? Vấn đề thứ 2 là về cơ sở vật chất và đào tạo, việc học phí cao, mọi người cũng đang cho rằng là không tương xứng với cơ sở vật chất mà đang sử dụng, đặc biệt là những cơ sở vật chất liên quan đến chất lượng buổi học như phòng máy (thực hành và thi), điều hòa, phòng học đều không đảm bảo, em chỉ muốn lấy một ví dụ ở các bạn học IT chương trình Việt Nhật K62, bạn em có kể về việc phòng máy các bạn ấy học còn không dùng được, nhưng lại không thể xin phép dùng phòng máy, dù ng máy tính cá nhân, hay như các bạn thi Thuật toán ứng dụng đang thi thì máy tự nhiên tắt, hoặc câu chuyện điều hòa chả nước ở nhiều phòng D9?

Thứ 3 là về công bằng học phí, dù nhiều thành phần trong một lớp nhưng các bạn theo chương trình khác nhau thì đóng số tiền khác nhau, thậm chí người này gấp đôi người kia, liệu các chương trình đại trà, học chuyên ngành một lớp 120 người có đảm bảo chất lượng, các môn học không sử dụng cơ sở vật chất và nhiều môn tương tự lại có học phí đắt đỏ, không xứng đáng?

Thứ 4, em nghĩ là ức chế nhất, vậy sinh viên ức chế và thắc mắc nhiều năm trời thì giải quyết như thế nào, khi kênh yammer trở thành kênh truyền thông chính trong khi nhiều sinh viên còn chưa biết, bỏ các kênh facebook trong khi facebook có nhiều sv sử dụng và phổ biến hơn? Các câu hỏi trên Yammer thì không được hoặc chậm trả lời? Sinh viên giải đáp thắc mắc ở đâu khi một năm chỉ có 1 buổi shcd và buổi đó luôn làm sinh viên mệt mỏi với những câu trả lời không hề đúng ý nhưng P

Figure 16: Quá trình tìm và lấy phần text của các bình luận



Do bình luận mang tính không đồng tình chiếm đa số, vì vậy để rút ngắn thời gian gán nhãn thì ban đầu, em gán tất cả các bình luận với nhãn bằng 0.

Thực hiện gán nhãn: Nhãn **1** ứng với bình luận có tính **đồng ý**, nhãn **0** ứng với các bình luận **còn lại**.

Việc gán nhãn cho mỗi bình luận **không phụ thuộc vào ngữ cảnh**.

	A	B
1	Text	Sentiment
2	Tôi muốn viết vài ý rất quan trọng để các bạn sinh viên hiểu: Tất cả chúng ta đều h	0
3	Thừa độ ảo tưởng và thiếu độ thực tế. Ở Châu Âu: trung bình ~€ 2,000-3000, tức k	0
4	Nhiều giảng viên đi dạy còn ăn bớt giờ	0
5	Em thừa thầy, môn đồ án của em 3 tín học phần 6 tín học phí Nhưng tất cả vài vóc	0
6	Tran Van Top em thừa thầy nhưng thực tế cơ sở vật chất ở các phòng thí nghiệm,	0
7	Tran Van Top thầy ơi. Theo em cảm nhận, thứ nhà trường thiếu là tiền, gia đình sir	0
8	Đăng Phạm Mỹ <a href="https://www.usnews.com/.../paying-for-college-infographic">https://www.usnews.com/.../paying-for-college-infographic</a> Sing: t	0
9	Tran Van Top Thừa thầy. E muốn hỏi là nếu trường đã bảo tự chủ thì tại sao lại yê	0
10	Tran Van Top Em hiểu khi tự chủ chắc chắn học phí phải tăng nhưng đến 1 lúc nào	0
11	Hoàng Xuân Đích thôi thôi, bạn có đi học đâu =))	0
12	Thế này mà Nhà nước không đầu tư thì bao giờ mới sánh vai được với cường quốc	0
13	Tran Van Top Dạ em xin có ý kiến như này ạ. Mỗi năm tăng 8-10% mà năm nhất có	0
14	Nguyễn Xuân Dũng làm nghìn việc tốt chẳng ai nhớ Chỉ 1 lần sai đã thất thời	0
15	Em chỉ mong phòng học ở nhà d6 tài trợ quạt cho 4 dãy bàn cuối	0
16	Em nghĩ là học phí cao hay thấp không quan trọng, nhưng có những điều bất hợp l	0
17	Việt Hoàng :))	0
18	Hoang Chu hay quá bạn	0
19	Hoang Chu không biết bạn/anh lấy số liệu kia ở đâu nhưng mình/em khẳng định số	0
20	"Mọi thứ chỉ có GIÁ TRỊ khi nó đi với GIÁ CẢ phù hợp" -- Warren Buffet. E chưa ba	0

Figure 17: Dữ liệu thô được thu thập

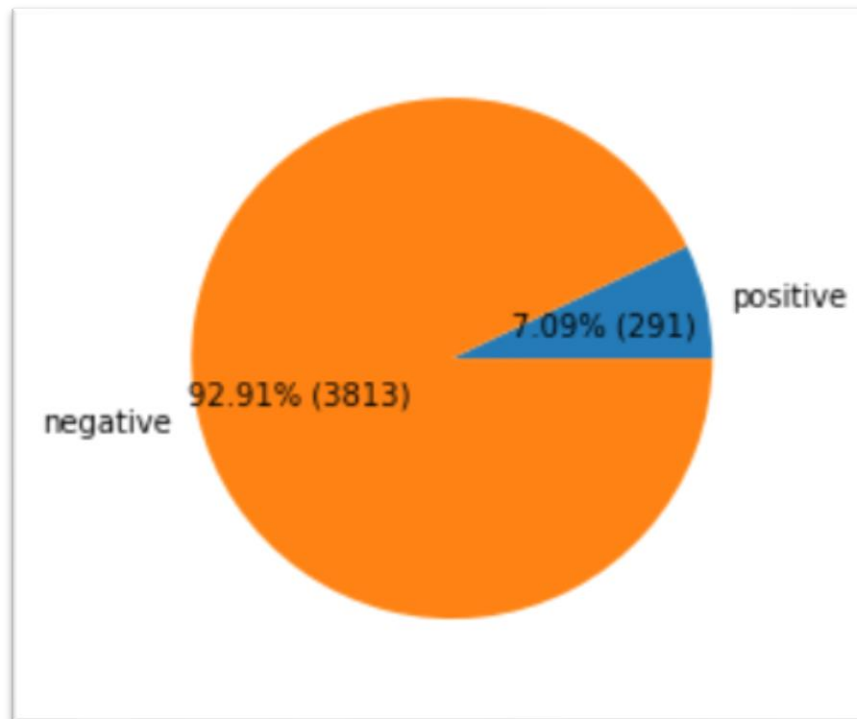
	A	B
1	Text	Sentiment
2	Tôi muốn viết vài ý rất quan trọng để các bạn sinh viên hiểu: Tất	1
3	Thừa độ ảo tưởng và thiếu độ thực tế. Ở Châu Âu: trung bình ~€ 2	0
4	Nhiều giảng viên đi dạy còn ăn bớt giờ	0
5	Em thưa thầy, môn đồ án của em 3 tín học phần 6 tín học phí Nh	0
6	Tran Van Top em thưa thầy nhưng thực tế cơ sở vật chất ở các p	0
7	Em mong tăng học phí sẽ đi đôi với tăng chất lượng mọi mặt ạ.	1
8	Tran Van Top thầy ơi. Theo em cảm nhận, thứ nhà trường thiếu l	0
9	Đăng Phạm Mỹ <a href="https://www.usnews.com/.../paying-for-college-">https://www.usnews.com/.../paying-for-college-</a>	0
10	Tran Van Top Thưa thầy. E muốn hỏi là nếu trường đã bảo tự ch	0
11	Tran Van Top Em hiểu khi tự chủ chắc chắn học phí phải tăng nhu	0
12	Thế này mà Nhà nước không đầu tư thì bao giờ mới sánh vai đượ	0
13	Tran Van Top Dạ em xin có ý kiến như này ạ. Mỗi năm tăng 8-10%	0
14	Cơ sở vật chất sẽ tốt lên thôi mà	1
15	Nguyễn Xuân Dũng làm nghìn việc tốt chẳng ai nhớ Chỉ 1 lần sai đ	0
16	Em chỉ mong phòng học ở nhà d6 tài trợ quạt cho 4 dãy bàn cuối	0
17	Em nghĩ là học phí cao hay thấp không quan trọng, nhưng có nhữ	0
18	Hoang Chu không biết bạn/anh lấy số liệu kia ở đâu nhưng mình/	0
19	"Mọi thứ chỉ có GIÁ TRỊ khi nó đi với GIÁ CẢ phù hợp" -- Warren	0
20	Thứ ta thiếu là gì thì em chưa biêts nhưng em biết thứ bố mẹ em	0

Figure 18: Dữ liệu sau khi được gán nhãn

Sau khi đọc, gán nhãn và xử lý các bình luận, em có các nhận xét sau:

- Có **201/4104** bình luận **đồng ý**, chiếm **7.09%**.
- Các bình luận mang tính một chiều, đa số là các bình luận không đồng ý.
- Có nhiều các bình luận không liên quan, bình luận bằng ảnh hay là chỉ *tag* tên bạn bè.

Các đặc điểm trên là một khó khăn của bài toán, ảnh hưởng lớn đến kết quả.



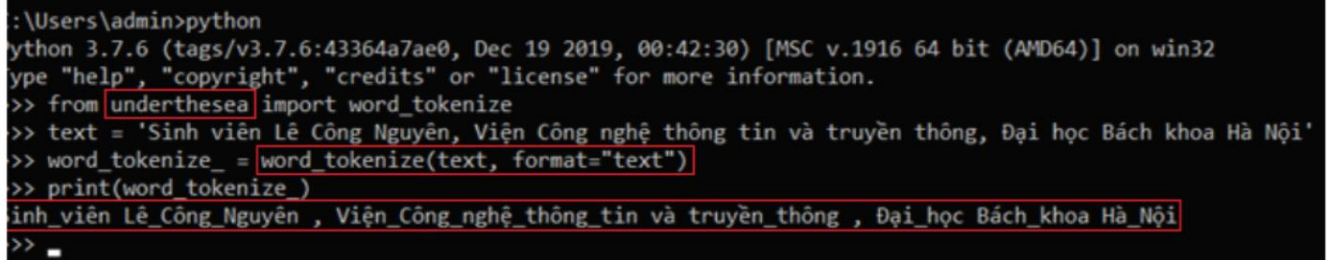
*Figure 19: Tỷ lệ tính chất của bình luận*



## 2. Tiền xử lý dữ liệu và thư viện Underthesea

Loại bỏ dấu câu và các ký tự đặc biệt như: dấu chấm, dấu phẩy, dấu chấm hỏi...

**Tokenize** văn bản bằng thư viện “Underthesea” - Vietnamese NLP Toolkit.



```
:\Users\admin>python
Python 3.7.6 (tags/v3.7.6:43364a7ae0, Dec 19 2019, 00:42:30) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>> from underthesea import word_tokenize
>> text = 'Sinh viên Lê Công Nguyễn, Viện Công nghệ thông tin và truyền thông, Đại học Bách khoa Hà Nội'
>> word_tokenize_ = word_tokenize(text, format="text")
>> print(word_tokenize_)
sinh_vien_Lê_Công_Nguyễn , Viện_Công_nghệ_thông_tin_và_truyền_thông , Đại_học_Bách_khoa_Hà_Nội
>> _
```

Figure 20: Ví dụ tách từ bằng Underthesea

## 3. Biểu diễn dữ liệu bằng TF-IDF

**Embedding** văn bản bằng kỹ thuật “Term Frequency - Inverse Document Frequency” (TF-IDF).

TF-IDF được sử dụng để đánh giá độ quan trọng của một từ hoặc cụm từ trong một văn bản, giá trị TF-IDF cao thể hiện độ quan trọng cao, là những từ xuất hiện nhiều trong văn bản đang được đánh giá, và xuất hiện ít trong các văn bản còn lại [1].

TF-IDF giúp loại bỏ những từ phổ biến (có thể là stopwords) và giữ lại những từ có giá trị cao (từ khóa của văn bản).

TF-IDF gồm hai thành phần là TF và IDF:

- TF: Tần suất xuất hiện của từ trong một văn bản.

$$TF_{ij} = \frac{f_{ij}}{n_j}$$

$f_{ij}$  : the frequency of term  $i$  in document  $j$

$n_j$  : the total number of words in document  $j$

*Figure 21: TF*

- IDF: Tần suất nghịch của từ trong tập các văn bản.

$$IDF_i = 1 + \log\left(\frac{N}{c_i}\right)$$

$N$  : the total number of documents

$c_i$  : the number of documents that contain word  $i$

*Figure 22: IDF*

- TF-IDF:

$$w_{ij} = TF_{ij} \times IDF_i$$

*Figure 23: TF-IDF*

#### 4. Phân loại ý kiến bằng giải thuật học máy SVM

SVM là một thuật toán học máy thuộc nhóm *Supervised Learning* (học có giám sát) dùng để *Classification* (phân loại) dữ liệu thành các nhóm riêng biệt bằng cách tìm một *Hyper Lane* (siêu phẳng).

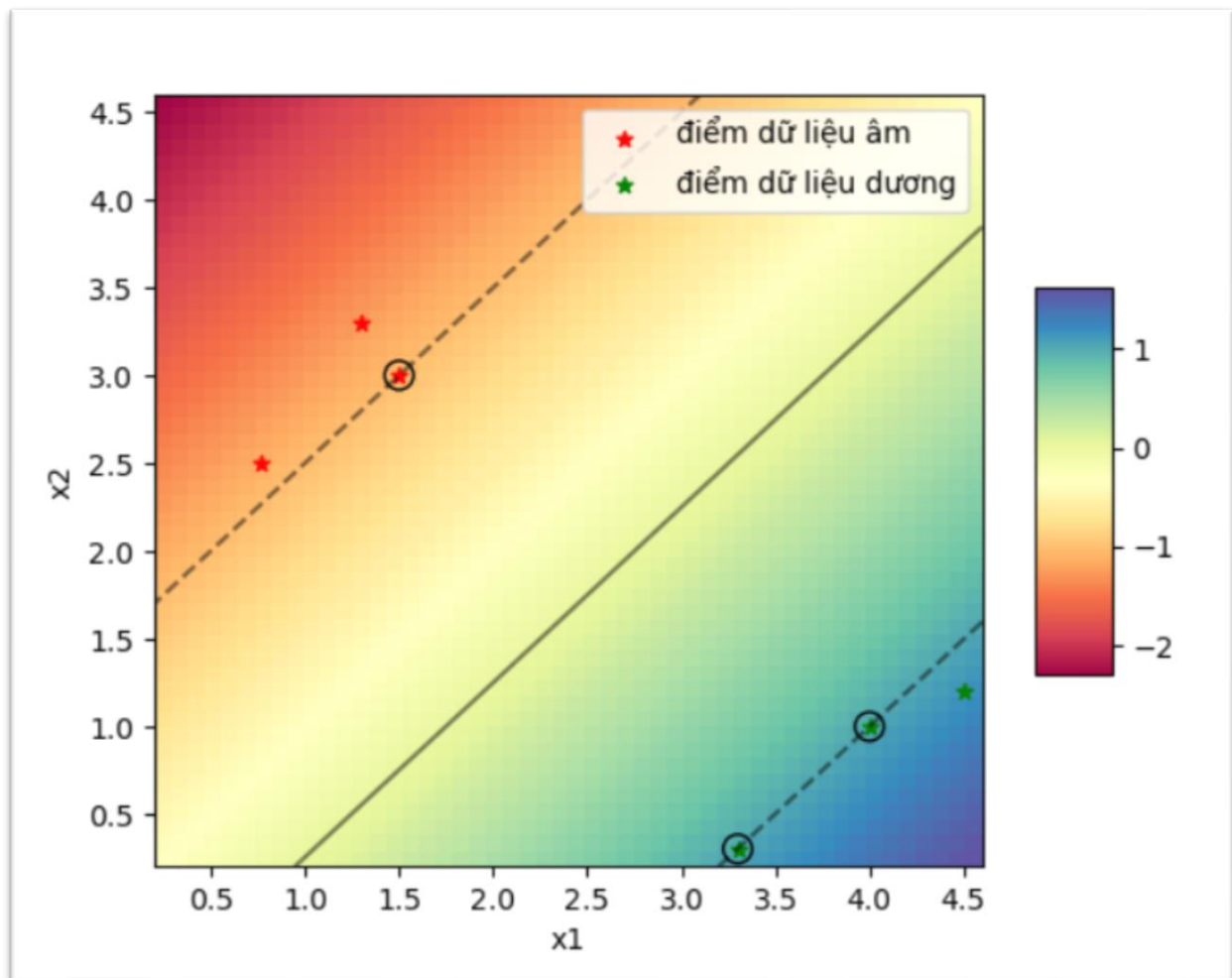


Figure 24: Ví dụ về giải thuật SVM

### a. Quá trình train

Bước 1: Crawl dữ liệu và gán nhãn

Bước 2: Tiền xử lý, Tokenize

Bước 3: Embedding văn bản

Bước 4: Train model SVM

- Input: Các vector embedding
- Output: Các nhãn đã gán

Bước 5: Lưu model vào file

Trong quá trình tạo Vocabulary, em chỉ xét các từ có tần suất xuất hiện trên **3** lần trong các văn bản và **dưới 80%** trong toàn bộ văn bản. Các từ trên 80% khả năng cao là các *stopwords* [2]. Sau khi embedding xong thì lưu ra file.

Em chia tập dataset thành 2 tập là tập train với tỉ lệ **80%** và tập test với tỉ lệ **20%**. Sau khi train xong thì lưu model ra file, đồng thời thu được kết quả như hình dưới.



```
Model Score = 0.9220462850182704
```

*Figure 25: Kết quả test trên tập train*

## b. Quá trình test

Bước 1: Đọc một link bài viết và crawl dữ liệu

Bước 2: Tiền xử lý, Tokenize

Bước 3: Embedding văn bản

Bước 4: Lần lượt đưa các bình luận vào model để phân loại

Bước 5: Liệt kê các bình luận theo nhãn

```
Số bình luận đồng ý: 2
Số bình luận không đồng ý: 120
```

1	Text	Sentiment
2	Nghĩ lại năm của mình 105k 1 tín_chỉ mà thấy ảm lòng Kỳ 2 năm nhất còn có 2tr6 Giờ nhìn các em đóng gấp 9 10 lần mà sợ	Không đồng ý
3	1 năm dao_động đến 40tr cho nhanh Với mức học_phí này gần bằng chương_trình quốc_tế ngày_xưa và sv đa_số học bk là	Không đồng ý
4	Lâm_Quyền lấy được học_bổng đủ chết r	Không đồng ý
5	Lâm_Quyền thấy mng cmt bảo thầy cô giáo đọc slide ĐÚNG Có_môn e học thầy đi dạy 3 15 hay 16 buổi k nhớ nữa Và thầy	Không đồng ý
6	Đề ra trường đúng_hạn tối_thiểu tầm 16 tín_học_phần một kỳ Nhưng đáng nói là tín_học_phí thường chẳng bằng tín_học_ph	Không đồng ý
7	À đợt mình tìm_hiểu là khoảng 17tr cho năm nhất nhé	Không đồng ý
8	Đỗ Thanh_Thủy e cũng vì điều này mà đẩy hust lên nv1 ạ	Không đồng ý
9	Cao Nam	Không đồng ý
10	Nguyễn_Thúy_Quỳnh năm 3 đc đi làm ùi k lo hehe	Không đồng ý
11	Lí_thuyết với thực_tế khác nhau nhiều đó	Không đồng ý
12	cái quan_trọng là học_phí tăng phải đi_đôi với chất_lượng đào_tạo các bạn nghĩ sao khi đến lớp_học có giảng_viên ngồi ch	Không đồng ý
13	học_phí kia là nhà_trường đăng_kí cho lúc vào trường tầm 12 14 tín_thời còn sau tự_đăng_kí hơn thì tiền nó lại khác bọ	Không đồng ý
14	Vũ_Văn_Thực 0 tín_học_phần là như nào ạ	Không đồng ý
15	tín_học_phần bằng 0 cần qua môn là được	Không đồng ý
16	Môn Qsc chẳng_hạn 5 tín_học_phí 0 tín_học_phần	Không đồng ý

1	Text	Sentiment
122	vẫn rẻ ý	Đồng ý
123	vẫn hợp lý	Đồng ý

Figure 26: Kết quả phân loại

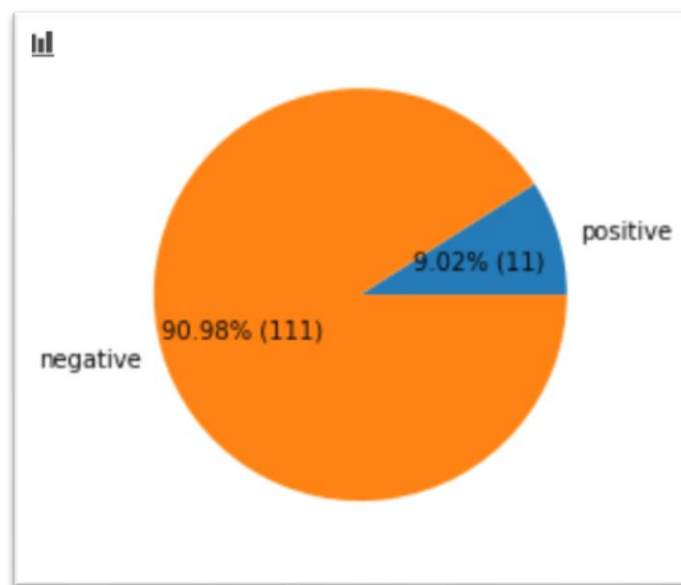
### III. KẾT QUẢ VÀ THẢO LUẬN

#### Kết quả:

Khi test trên tập train, độ chính xác khá cao, lên đến **92.2%**.

Với tập test, em thu thập được **122 bình luận**, và sau khi phân loại thu được kết quả là 120 bình luận không đồng ý và 2 bình luận đồng ý.

Để so sánh, em đã gán nhãn cho tập test và trực quan được như sau:



*Figure 27: Trực quan tập test*

Kết quả cho thấy, có **11/122** bình luận đồng ý, chiếm **9.02%**, tỉ lệ phân loại đúng khá cao **113/122** bình luận, tuy nhiên thì chỉ phân loại đúng được 2/11 bình luận đồng ý.

#### Thảo luận:

Mặc dù độ chính xác phân loại khá cao, nhưng những tính chất của bình luận thực sự có ảnh hưởng lớn đến kết quả này. Bên cạnh là các tham số đầu vào trong quá trình train.

**Link source code:** <https://github.com/nguyenlecong/Project-III>

## IV. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

### Kết luận:

Bằng việc sử dụng framework “Selenium”, dữ liệu được thu thập và lưu trữ là ý kiến ở các phần bình luận chính và các phần bình luận phản hồi của các bài viết về vấn đề học phí. Tiếp theo là tiền xử lý dữ liệu và sử dụng framework “Underthesea” để thực hiện Tokenize. Kết quả sẽ là đầu vào cho kỹ thuật embedding “TF-IDF”. Đầu ra của thuật toán là các vector embedding và là đầu vào của model phân loại SVM theo nhãn đồng ý hay không đồng ý. Kết quả cho thấy, hiệu quả mang lại hiệu quả khá cao, tuy nhiên một phần quan trọng đến từ dữ liệu thu thập được có tính chất đặc biệt, gây ảnh hưởng lên kết quả này.

### Khó khăn:

Vì đây là chủ đề ngắn hạn nên số lượng bài viết rất **hạn chế**. Một số bài viết bị xóa sau đó.

Các ý kiến mang tính **một chiều, không quá trọng tâm**.

### Trong tương lai:

Với bài toán này, em sẽ cố gắng thực hiện phân loại các nhãn đầu ra của các bình luận theo nhiều nhãn hơn nữa, chi tiết hơn nữa về cảm xúc của bình luận đó và phân loại được cả các bình luận không liên quan.

Hơn nữa là sẽ thử nghiệm với một số thuật toán biểu diễn văn bản và thuật toán học máy khác để tăng độ chính xác.

Bài toán là bước khởi đầu cho nhiều dự án có tính ứng dụng cao trong thực tiễn, là bước phát triển ban đầu cho những bài toán khác, không chỉ trên nền tảng mạng xã hội Facebook mà còn trên nhiều nền tảng mạng xã hội hay bất kỳ trang web nào khác.

## V. TÀI LIỆU THAM KHẢO

[1] *TF-IDF là gì?*

<https://nguyenvanhieu.vn/tf-idf-la-gi/>

[2] *[NLP Series #1] Thử làm hệ thống đánh giá sản phẩm Lazada*

<https://www.miai.vn/2020/05/04/nlp-series-1-thu-lam-he-thong-danh-gia-san-pham-lazada/>