PHÂN LOẠI CHỦ ĐỀ BÀI VIẾT TRONG UIT FACEBOOK GROUP

Vũ Quí San - 18520143 - CS114.K21.KHTN

Link Github:

https://github.com/santheipman/CS114.K21.KHTN

Tóm tắt

- Tên đề tài: Phân loại chủ đề bài viết trong UIT Facebook group.
- Tóm tắt về đồ án và kết quả đạt được: Đánh giá phương pháp Bag-of-words và phương pháp Tf-idf cho việc biểu diễn văn bản dưới dạng vector, đánh giá một số model cho bài toán phân loại văn bản trên tập dữ liệu mà nhóm thu thập và gán nhãn.
- Ảnh các thành viên của nhóm:

Vũ Quí San

Đặt vấn đề

Hiện nay sinh viên UIT có hai nơi chính để trao đổi, thảo luận, giải đáp thắc mắc về trường: một là forum UIT và hai là UIT Facebook group. Đa phần sinh viên thường sử dụng group hơn vì thói quen sử dụng Facebook. Tuy nhiên, trong khi các bài đăng tại forum được phân theo các mục (ví dụ như mục góc học tập, mục hỗ trợ đời sống, mục khu vực điều hành,...) thì các bài đăng tại group trên Facebook lại không được phân loại theo chủ đề hay theo các mục nào.

Do đó trong đồ án này nhóm sẽ ứng dụng Machine Learning để giải quyết bài toán phân loại chủ đề bài viết trong UIT Facebook group.

Mô tả bài toán

Phân loại chủ đề bài viết trong UIT Facebook group

- Input: nội dung bài viết
- Ouput: bài viết đó thuộc chủ đề nào trong **4** chủ đề sau: *thông báo*, *tìm đồ thất lạc*, *hỏi đáp*, *khác*.

Xây dựng tập dữ liệu

- Dữ liệu được thu thập từ **5** group **UIT K10 -> K14**.
- Sử dụng thư viện Selenium để đọc html và lấy ra nội dung các bài viết.
- Tổng số lượng bài viết thu thập được là **13154**.
- Tuy nhiên cùng một bài viết có thể đăng trong nhiều group nên khi gộp bài viết của các group lại thì sẽ xảy ra hiện tượng **trùng bài viết**. Do đó nhóm thực hiện thao tác loại bỏ trùng **trước khi** tiến hành gán nhãn.

	0
0	[TSSĐH]-THÔNG BÁO TỔ CHỨC LỚP ÔN TẬP CHUẨN BỊ ...
1	💗💗 Giao lưu cùng Nhà báo- Nhà thơ Nguyễn P...
2	Phòng Công tác Sinh viên thông tin đến sinh vi...
3	🔬 Các chuyên ngành đào tạo trường Đại học Kans...
4	👉 Bật mí với các bạn một chương trình vô cùng ...
...	...
13149	Bạn này vừa đến Trường nộp Giấy kết quả thi TH...
13150	Các em nhỏ nhớ đi MHX2020 nha\nChị share để ki...
13151	# chào mình đỗ ktpm clc ... bác nào chung khoa...
13152	Chào các bạn, phòng Công tác Sinh viên đã đăng...
13153	Những điều Thí sinh cần lưu ý!

13154 rows × 1 columns

	0
4812	Sau một thời gian dài vắng bóng thì, đây: Hoạt động đầu năm của Ban học tập Công nghệ Thông tin nè các bạn.\nSắp xếp thời gian để tuần sau chúng ta cùng nghe PSG. TS. Đỗ Phúc nói gì về lĩnh vực Khoa học Dữ liệu nhé 😬😬😬
8438	Sau một thời gian dài vắng bóng thì, đây: Hoạt động đầu năm của Ban học tập Công nghệ Thông tin nè các bạn.\nSắp xếp thời gian để tuần sau chúng ta cùng nghe PSG. TS. Đỗ Phúc nói gì về lĩnh vực Khoa học Dữ liệu nhé 😬😬😬
1019	Sau một thời gian dài vắng bóng thì, đây: Hoạt động đầu năm của Ban học tập Công nghệ Thông tin nè các bạn.\nSắp xếp thời gian để tuần sau chúng ta cùng nghe PSG. TS. Đỗ Phúc nói gì về lĩnh vực Khoa học Dữ liệu nhé 😬😬😬

Xây dựng tập dữ liệu

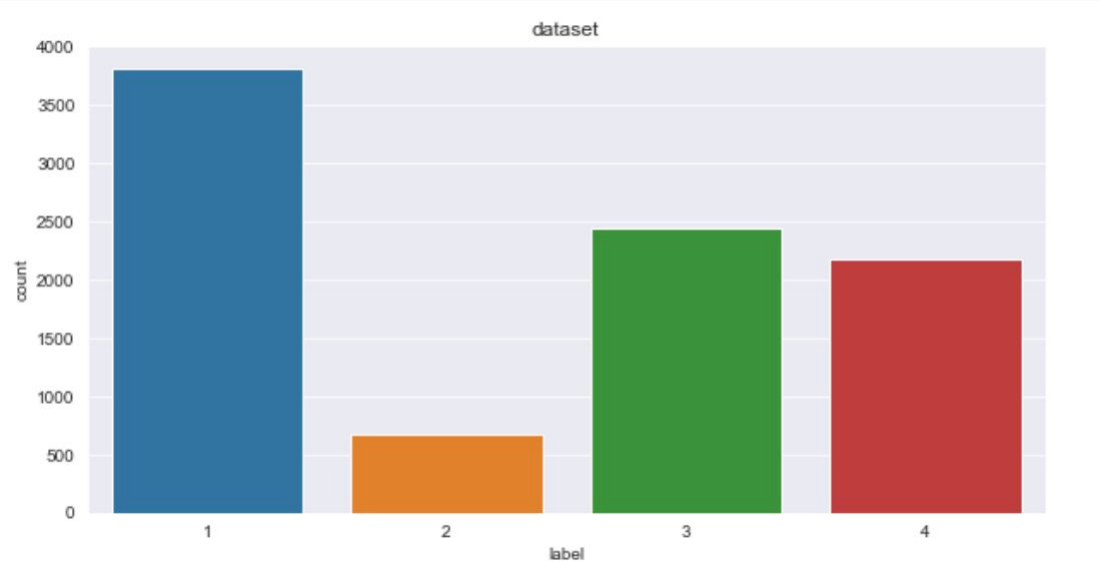
- Sau khi loại bỏ trùng thì tập dữ liệu còn lại **9094** mẫu.
- Tiến hành gán nhãn cho các mẫu. Nhóm đưa ra quy ước về các lớp như sau:
  - **Thông báo**: các thông báo của trường, của khoa, của các câu lạc bộ, của lớp (ví dụ: thông báo mở lớp); thông báo về học bổng.
  - **Tìm đồ thất lạc**: viết về việc thất lạc đồ hoặc tìm thấy đồ thất lạc.
  - **Hỏi đáp** (chủ yếu liên quan đến vấn đề cá nhân): hỏi về các hoạt động ở trường, nhờ điền form đăng ký mở lớp, bán lại đồ dùng,...
  - **Khác**: các hoạt động của cơ quan, tổ chức bên ngoài như tuyển dụng, quảng cáo khóa học, chương trình bốc thăm, tham quan doanh nghiệp,...; các bài không rõ chủ đề hoặc mang tính giải trí.

Tập dữ liệu

	label	text
3655	3	Mình cần đổi lớp tư tưởng thứ 4 sang thứ 5 :((có bạn nào có thể đổi giúp mình với được không 😘😘
2757	3	[CẦU CỨU 500AE]\nMọi người ơi giúp em bình chọn cho sdb 052 Nguyễn Thị Chiều Hoang với ạ!\nTội chị em đi thi bán kết mà anh diễn chung bỏ thi nên chị catwalk 1 mình và bị trừ điểm nhiều nên không vào chung kết được. Giờ chỉ còn chờ vào điểm vote cứu thí sinh thôi ạ! Chị đàn hay, hát giỏi, dễ thương, vui tánh lắm o.o Huhu... Chị em bị bỏ lại ở thứ 2 r.. MN giúp em với!\nEm cảm ơn rất nhiều! Và xin lỗi nếu có làm phiền ạ! ♡♡♡
2464	3	Bạn nào có nguyện vọng nhà trường mở lớp AV1 thì làm khảo sát này nhé. OK
7395	3	Có bạn nào nhường slot it002.j21.atcl hong ? Mình sẽ hậu tạ.Mình cảm ơn nhiều
3645	3	Cho hỏi đăng kí môn đã thành công,giờ vô check thì bị mất lớp,thì phải làm sao mọi người...cảm ơn nhiều 😭😭😭
4065	3	làm sao để khắc phục lỗi này đây các bạn. Thử lớp nào cũng vậy ???
414	3	[GÓC CHIA SẺ]\nMọi người muốn có trải nghiệm hoặc "xõa" sau thời gian dài dí đầu với deadline thì hãy đọc thử bài này nhé, bản thân mình thấy hay lắm.

Ví dụ về các mẫu thuộc lớp *hỏi đáp*

Tập dữ liệu

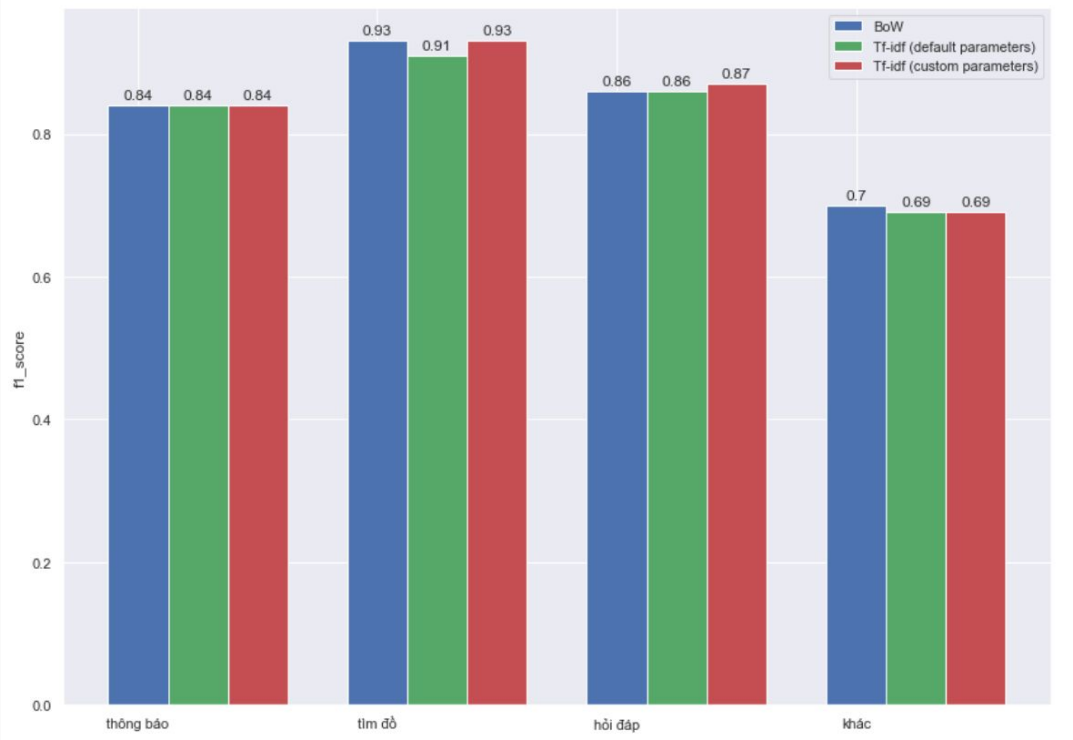


1. Thông báo 2. Tìm đồ thất lạc 3. Hỏi đáp 4. Khác

- Đồ thị bên cho biết số lượng mẫu của từng lớp.
- Tập dữ liệu **không cân bằng** nên nhóm sử dụng **F1 score** làm độ đo để đánh giá các phương pháp.

Trích xuất đặc trưng

- Tiền xử lý:
  - Chuẩn hóa một số từ (ví dụ: tks, thanks, cám ơn -> cảm ơn)
  - Chuyển thành chữ thường
  - Loại bỏ đường link, hashtag, kí tự đặc biệt.
  - Tách từ (tokenize)
- Chuyển văn bản thành vector: So sánh hai phương pháp là **Bag-of-words** (Bow) và **Tf-idf.** Lần lượt sử dụng từng phương pháp để train model Logistic Regression rồi so sánh f1 score.



Trích xuất đặc trưng

Tf-idf với default parameters cho kết quả thấp nhất
Tf-idf với custom parameters cho kết quả **tương đương** với BoW nhưng cho vector có **kích thước nhỏ hơn** (1918 so với 9893, xem hình dưới). Điều này xảy ra bởi vì Tf-idf với custom parameters có tham số min_df (những từ có tần suất xuất hiện nhỏ hơn min_df thì bị bỏ qua trong quá trình chuyển văn bản thành vector) được set bằng 10, cao hơn giá trị mặc định min_df=1. Giá trị min_df lớn hơn 1 (lớn hơn tham số mặc định) giúp loại bỏ một số danh từ riêng không mang lại lợi ích trong việc phân loại chủ đề bài viết.

```
BoW
shape = (9094, 9893)
```

```
Tf-idf (default parameters)
shape = (9094, 9893)
```

```
Tf-idf (custom parameters)
shape = (9094, 1918)
```

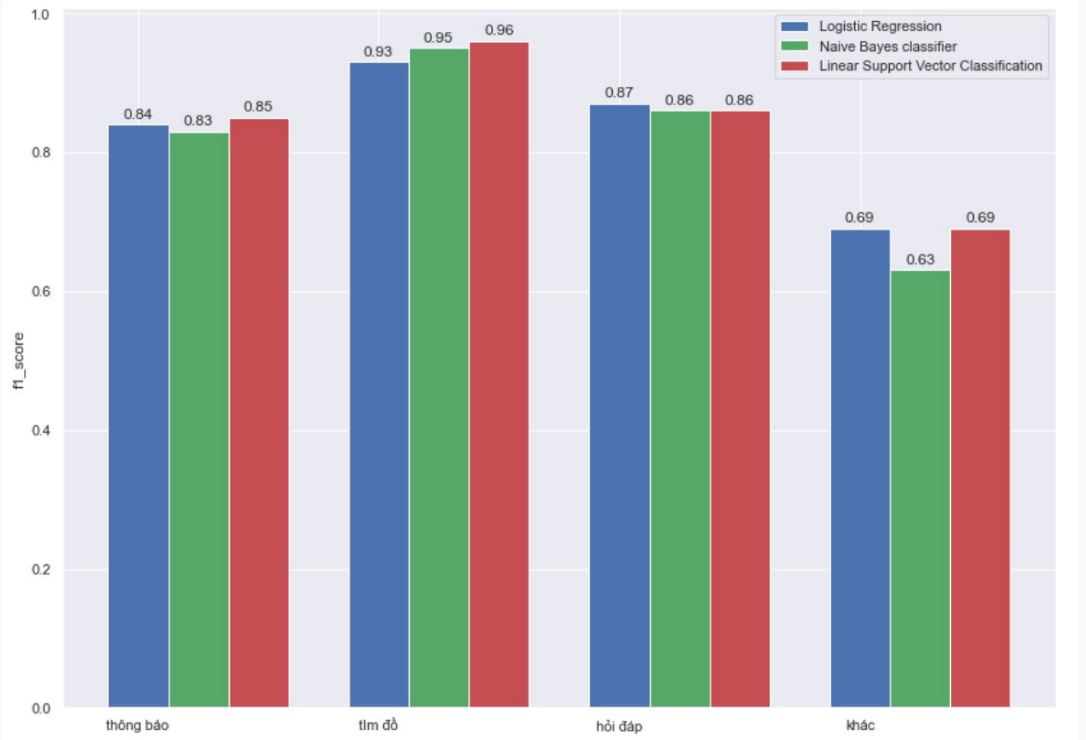
Qua kết quả trên nhóm **chọn Tf-idf với custom parameter** để vector hóa văn bản.

Chọn model

So sánh 3 model: Logistic Regression, Naive Bayes Classifier, Linear Support Vector Machine.

- Chia tập dữ liệu thành 80% train + 20% test
- Train từng model và thu được f1 score như đồ thị bên

Linear Support Vector Machine cho kết quả tốt nhất trong số 3 model. Do đó nhóm chọn model này cho bài toán.



Tinh chỉnh Hyperparameter

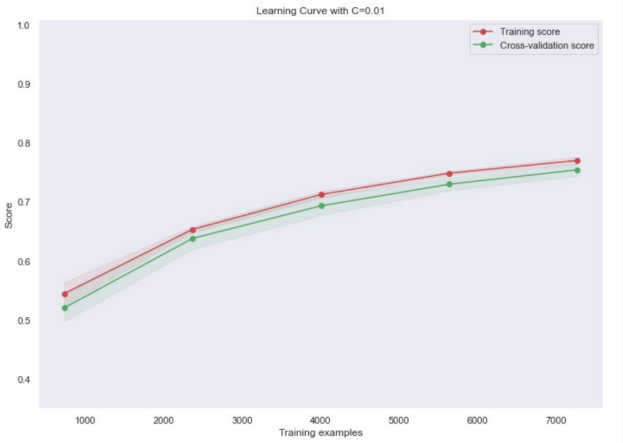
Đối với model Linear SVC từ thư viện sklearn thì hyperparameter quan trọng nhất là C, nhóm tiến hành sử dụng hai phương pháp để tinh chỉnh C là: GridSearchCV và đánh giá dựa trên learning curve.

1. GridSearchCV: cho một tập các giá trị C, hàm này trả về giá trị C cho score cao nhất. Ở đây hàm tìm được **C=0.1**

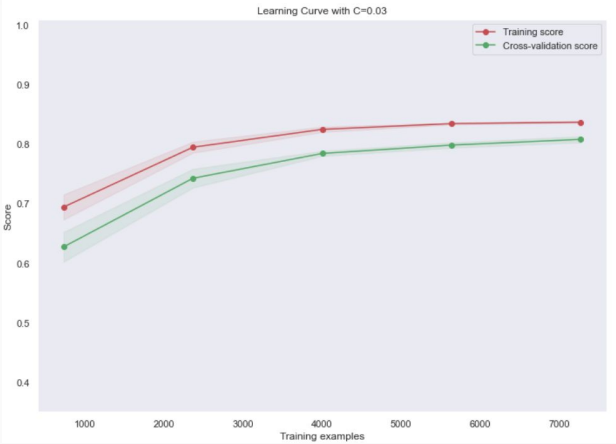
```
{'C': 0.1}
LinearSVC(C=0.1, class_weight=None, dual=True, fit_intercept=True,
     intercept_scaling=1, loss='squared_hinge', max_iter=1000,
     multi_class='ovr', penalty='l2', random_state=None, tol=0.0001,
     verbose=0)
```

2. Learning curve: Nhìn vào các đường learning curve có thể giúp ta nhận biết model **có khả năng** bị overfitting hoặc underfitting hay không

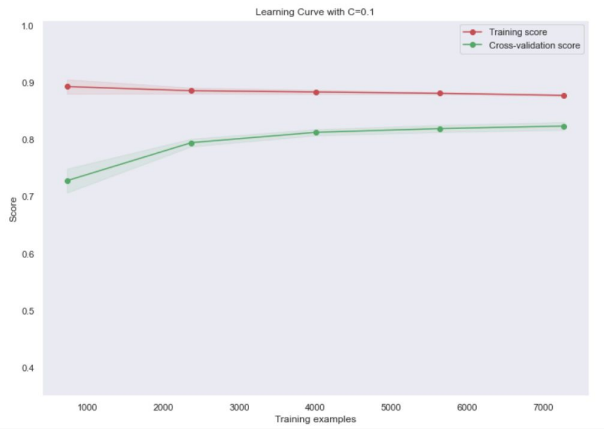
Tinh chỉnh Hyperparameter



C=0.01



C=0.03



C=0.1

Theo nhóm thì với C=0.01 thì model nhiều khả năng bị underfit, với C=0.1 (GridSearchCV chọn đây là hyperparameter tốt nhất) thì model nhiều khả năng bị overfit. Trường hợp C=0.03 thì model cân bằng giữa underfit và overfit hơn so với hai trường hợp còn lại và bên cạnh đó validation score có xu hướng còn tăng khi ta thêm data, vì vậy model có khả năng sẽ tổng quát hóa trên dữ liệu mới (generalize to unseen data) tốt hơn hai trường hợp còn lại. Do đó nhóm chọn **C=0.03** là kết quả của việc tinh chỉnh hyperparameter cho model.

Kết luận và hướng phát triển

- Kết luận:
  - F1 score trên lớp ***khác*** (lớp thứ tư) còn thấp bởi vì nội dung của lớp này rất đa dạng và còn có thể dễ nhầm lẫn với lớp khác, ví dụ thông báo từ sự kiện của trường (thuộc lớp *thông báo*) và thông báo từ sự kiện của tổ chức bên ngoài có thể có hình thức giống nhau.
  - Việc quy định các lớp là tùy thuộc vào nhu cầu của ta và chia lớp hợp lý có thể model dễ học được pattern của dữ liệu hơn và giúp ứng dụng có ích hơn.
  - Việc tinh chỉnh hyperparameter tự động bằng GridSearchCV có thể chọn ra hyperparameter làm cho model bị overfitting.
- Hướng phát triển:
  - Dựa vào learning curve tại bước tinh chỉnh hyperparameter, ta thấy việc thêm data có thể giúp tăng performance của model.
  - Chia lại lớp, gán lại label cho hợp lý hơn.

Xin cảm ơn quý vị đã quan tâm!