Intent Classification for FAP Chatbot

Dự án "Bạn Cóc"

Hồ Viết Hoàng - HE205149
Hoàng Hồng Quân - HE195052
Nguyễn Minh Giang - HE195096
Nguyễn Hải Anh - HE172327

I. GIỚI THIỆU

FPT Academic Portal đóng vai trò trung tâm trong đời sống học thuật của sinh viên, cung cấp các thông tin quan trọng như thời khóa biểu, lịch thi, điểm số và các thông báo từ nhà trường. Mặc dù là một công cụ mạnh mẽ, giao diện truyền thống của FAP đòi hỏi người dùng phải thực hiện nhiều bước để tìm kiếm thông tin cụ thể, dẫn đến tốn thời gian và đôi khi gây bất tiện. Sinh viên thường có những câu hỏi lặp đi lặp lại và mong muốn có một phương thức truy cập thông tin nhanh chóng, tự nhiên hơn.

Để giải quyết vấn đề này, dự án "Bạn Cóc" được đề xuất nhằm xây dựng một chatbot có khả năng hiểu ngôn ngữ tự nhiên với tiếng Việt và tự động phân loại yêu cầu của sinh viên thành các ý định tương ứng với các chức năng trên FAP.

II. PHƯƠNG PHÁP LUẬN

Pipeline của dự án được xây dựng theo các bước chuẩn của một bài toán xử lý ngôn ngữ tự nhiên (NLP), từ xây dựng dataset, tiền xử lý, trích xuất đặc trưng đến huấn luyện và đánh giá mô hình.

A. Dataset Curation

Do không có sẵn bộ dữ liệu công khai cho tác vụ này, dữ liệu sẽ được tạo ra dưới dạng synthetic data. Bộ dữ liệu này gồm các câu hỏi do mô

phỏng, phản ánh các tình huống thực tế mà sinh viên có thể hỏi chatbot, và được gán nhãn với 5 loại ý định chính:

- `lich_hoc`: Các câu hỏi liên quan đến thời khóa biểu hàng tuần (thời gian, địa điểm, giảng viên).
- `lich_thi`: Các câu hỏi về lịch thi (ngày thi, phòng thi, hình thức thi).
- `diem_danh`: Các câu hỏi về tình trạng chuyên cần, số buổi vắng mặt.
- `diem_so`: Các câu hỏi về điểm thi, điểm thành phần, điểm tổng kết và GPA.
- `hoc_phí`: Các câu hỏi về học phí môn học.

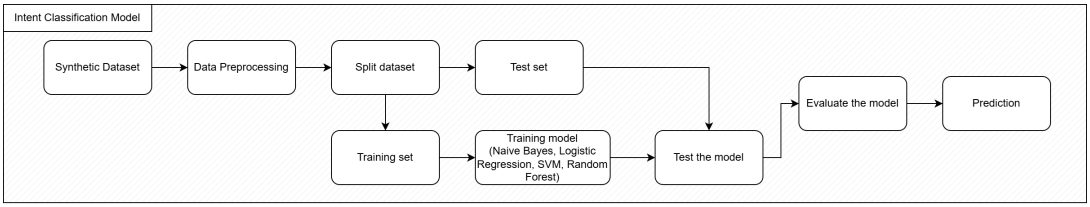
```
First few rows of the dataset:
                              sentence    intent
0       Lịch học hôm nay của tôi là gì?  lich_hoc
1  Cho tôi xem thời khóa biểu tuần này.  lich_hoc
2   Ngày mai tôi có tiết học nào không?  lich_hoc
3     TKB của tui tuần sau như thế nào?  lich_hoc
4      Hôm nay tôi học môn gì và ở đâu?  lich_hoc

Information about the dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 602 entries, 0 to 601
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   sentence  602 non-null    object
 1   intent    602 non-null    object
dtypes: object(2)
memory usage: 9.5+ KB
None

Description of the dataset:
                              sentence   intent
count                              602      602
unique                             602        5
top     Lịch học hôm nay của tôi là gì?  diem_so
freq                                 1      167
```

Thống kê mô tả bộ dữ liệu

B. Pipeline



Sơ đồ Pipeline của mô hình

1) **Preprocessing:** Nhằm chuẩn hóa và làm sạch dữ liệu văn bản thô.
   - *(Các bước chuẩn hóa Unicode, chuyển về chữ thường, loại bỏ dấu, loại bỏ stopwords và word segmentation được thực hiện trong hàm* `text_cleaning` *và* `apply(lambda s: s.split())` *trong notebook.)*
2) **Feature Extraction:** Các câu đã được tiền xử lý được chuyển đổi thành vector số học bằng hai phương pháp:
   - **TF-IDF (Term Frequency-Inverse Document Frequency):** Đánh giá tầm quan trọng của một từ trong một câu dựa trên tần suất xuất hiện của nó trong câu đó và tần suất nghịch đảo của nó trong toàn bộ tập dữ liệu.
   - **Word Embeddings (Word2Vec):** Huấn luyện một mô hình Word2Vec (với `embedding_size=100`) trên tập dữ liệu để học biểu diễn vector cho mỗi từ. Vector của mỗi câu được tính bằng cách lấy trung bình vector của các từ trong câu đó.
3) **Target Variable Encoding:**
   - Biến mục tiêu `intent` sẽ được mã hoá bằng `LabelEncoder` để đảm bảo yêu cầu đầu ra của mô hình Machine Learning.
4) **Training & Classification:**
   - Dữ liệu được chia thành tập huấn luyện (80%) và tập kiểm thử (20%) bằng phương pháp chia `stratify`.
   - Các mô hình Machine Learning được huấn luyện trên cả hai bộ đặc trưng (TF-IDF và Word2Vec) để so sánh hiệu suất:
     - Mô hình cơ sở (Baseline): `DummyClassifier`.
     - Mô hình cổ điển: `Multinomial Naive Bayes`, `Logistic Regression`, `Support Vector Machine (SVM)`, và `Random Forest`.
   - Kỹ thuật `GridSearchCV` được áp dụng để tìm ra các siêu tham số (hyperparameter) tốt nhất cho các mô hình (trừ Naive Bayes và Baseline).

III. KẾT QUẢ VÀ PHÂN TÍCH

A. Kết quả hiệu suất

Bảng dưới đây tổng hợp kết quả (Precision, Recall, F1-Score, Accuracy) của tất cả các mô hình trên tập kiểm thử. Các mô hình "Tuned"là kết quả sau khi áp dụng GridSearchCV.

Bảng 1: Tổng hợp kết quả hiệu suất

Category	Model	Precision	Precision Tuned	Recall	Recall Tuned	F1-Score	F1-Score Tuned	Accuracy	Accuracy Tuned
Baseline	Most Frequent Baseline	0.079	None	0.281	None	0.123	None	0.281	None
Baseline	Stratified Baseline	0.192	None	0.182	None	0.185	None	0.182	None
Trained Model	Multinomial Naive Bayes	0.922	None	0.917	None	0.916	None	0.917	None
Trained Model	Logistic Regression	0.953	**0.976 Δ+0.023**	0.950	**0.975 Δ+0.025**	0.951	**0.975 Δ+0.024**	0.950	**0.975 Δ+0.025**
Trained Model	Support Vector Machine	0.960	0.969 Δ+0.009	0.959	0.967 Δ+0.008	0.959	0.967 Δ+0.008	0.959	0.967 Δ+0.008
Trained Model	Random Forest	0.944	0.952 Δ+0.008	0.942	0.95 Δ+0.008	0.942	0.95 Δ+0.008	0.942	0.95 Δ+0.008
Embedding Model	Logistic Regression (Embeddings)	0.967	None	0.967	None	0.967	None	0.967	None
Embedding Model	Support Vector Machine (Embeddings)	0.967	None	0.967	None	0.967	None	0.967	None
Embedding Model	Random Forest (Embeddings)	0.945	None	0.942	None	0.942	None	0.942	None

B. Phân tích kết quả

Mô hình tốt nhất là **Logistic Regression** (sử dụng TF-IDF và đã được tinh chỉnh) với độ chính xác **0.975**.

Tuy nhiên, độ chính xác của các mô hình khác cũng rất cao. Đáng chú ý, các mô hình **Logistic Regression (Embeddings)** và **Support Vector Machine (Embeddings)** sử dụng Word2Vec cũng đạt hiệu suất xuất sắc với độ chính xác **0.967**, ngang bằng với mô hình SVM (TF-IDF) đã được tinh chỉnh.

Điều này cho thấy cả hai phương pháp trích xuất đặc trưng (TF-IDF và Word2Vec) đều rất hiệu quả cho bộ dữ liệu này. Mô hình Logistic Regression (TF-IDF) được chọn làm mô hình tốt nhất để triển khai.

IV. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

A. Kết luận

Dự án đã thành công trong việc xây dựng một pipeline hoàn chỉnh để phân loại ý định người dùng cho chatbot FAP. Với việc sử dụng các kỹ thuật NLP cơ bản, so sánh hai phương pháp đặc trưng hóa (TF-IDF và Word2Vec), và tinh chỉnh mô hình, dự án đã đạt được độ chính xác rất cao (97.5%) trong phân loại ý định.

B. Hướng phát triển trong tương lai

Để tiếp tục cải thiện và mở rộng hệ thống, các hướng phát triển sau đây được đề xuất:

- **Mở rộng và cân bằng Dataset:** Bổ sung thêm nhiều mẫu câu cho mỗi ý định, đặc biệt là các câu có cấu trúc phức tạp hoặc dễ gây nhầm lẫn. Sử dụng các kỹ thuật tăng cường dữ liệu (Data Augmentation) như thay thế từ đồng nghĩa hoặc back-translation.
- **Cải thiện biểu diễn văn bản:** Thử nghiệm với các mô hình Transformer chuyên biệt cho tiếng Việt như **PhoBERT** để cho ra embedding tốt hơn theo ngữ cảnh, thay vì chỉ dùng Word2Vec tính trung bình câu.