ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO

Nhóm: 13

Thành viên:

1. 20120040 – Nguyễn Quang Gia Bảo
2. 20120136 – Huỳnh Tuấn Nam
3. 20120136 – Lê Thành Nam
4. 20120140 – Nguyễn Đăng Nam
5. 20120146 – Nguyễn Thị Châu Ngọc
6. 20120158 – Trần Hoàng Anh Phi

Lớp: 20_21

Học phần: Phân tích dữ liệu thông minh

Thành phố Hồ Chí Minh, tháng 7 năm 2023

MỤC LỤC

I. Giới thiệu chung 1

1. Thông tin nhóm: 1

2. Khái quát về đồ án và bộ dữ liệu 1

a. Khái quát về đồ án: 1

b. Khái quát về bộ dữ liệu: 1

3. Phân công nhiệm vụ 2

4. Đánh mức độ hoàn thành 2

II. Khám phá – tiền xử lý dữ liệu – tạo đặc trưng 2

III. Xây dựng mô hình – Thử nghiệm – Đánh giá. 3

IV. Kết luận và hướng phát triển: 3

1. ***Kết quả trên leaderboard của Kaggle:*** 3

DistilBert (kết quả khi so sánh với perfect submission): 3

[174]:
```
print("accuracy = ", (perfect_score['target'] == yhat.reshape(-1)).sum() / len(yhat))
```
```
accuracy =  0.7805700275819798
```
..... 3

Bert Classiification model: 3

TÀI LIỆU THAM KHẢO 3

I. Giới thiệu chung

1. *Thông tin nhóm:*

Họ và tên	**MSSV**	**Email**
Nguyễn Quang Gia Bảo	20120040	20120040@student.hcmus.edu.vn
Huỳnh Tuấn Nam	20120136	20120136@student.hcmus.edu.vn
Lê Thành Nam	20120138	20120138@student.hcmus.edu.vn
Nguyễn Đăng Nam	20120140	20120140@student.hcmus.edu.vn
Nguyễn Thị Châu Ngọc	20120146	20120146@student.hcmus.edu.vn
Trần Hoàng Anh Phi	20120158	20120158@student.hcmus.edu.vn

2. *Khái quát về đồ án và bộ dữ liệu*

a. *Khái quát về đồ án:*

Đồ án này sẽ dựa trên cuộc thi **Natural Language Processing with Disaster Tweets** trên Kaggle.

Mô tả:

Twitter đã trở thành một kênh liên lạc quan trọng trong trường hợp khẩn cấp. Sự phổ biến của điện thoại thông minh cho phép mọi người thông báo các trường hợp khẩn cấp mà họ đang quan sát thấy trong thời gian thực. Do đó, nhiều tổ chức quan tâm đến việc xây dựng một chương trình tự động theo dõi Twitter nhằm phát hiện các tin khẩn cấp được người dùng đăng lên Twitter (chẳng hạn như các tổ chức cứu trợ thảm họa và hãng thông tấn báo chí). Tuy nhiên, không phải lúc nào bài đăng của người dùng cũng có thể xác định rõ đó có thực sự thông báo về một thảm họa hay không.

Trong cuộc thi này, các nhóm cần xây dựng một mô hình học máy dự đoán Tweet nào nói về thảm họa thực sự và Tweet nào không. Bạn sẽ có quyền truy cập vào tập dữ liệu gồm 10.000 tweet đã được phân loại thủ công.

b. *Khái quát về bộ dữ liệu:*

*Note: Phần **Khái quát về dữ liệu** được thực hiện chi tiết trong file: **data_collecting.ipynb** nên chúng em không viết lại trong báo cáo này. Thầy/cô giúp nhóm chấm bài trong file: **data_collection.ipynb**.

3. Phân công nhiệm vụ

Họ và tên	**MSSV**	**Phân công công việc**
Nguyễn Quang Gia Bảo	20120040	Tiền xử lý dữ liệu Xây dựng mô hình Viết báo cáo
Huỳnh Tuấn Nam	20120136	Tiền xử lý dữ liệu Thử nghiệm và đánh giá mô hình Viết báo cáo
Lê Thành Nam	20120138	Tiền xử lý dữ liệu Thử nghiệm và đánh giá mô hình Viết báo cáo
Nguyễn Đăng Nam	20120140	Khám phá dữ liệu Thử nghiệm và đánh giá mô hình Viết báo cáo
Nguyễn Thị Châu Ngọc	20120146	Khám phá dữ liệu Xây dựng mô hình Viết báo cáo
Trần Hoàng Anh Phi	20120158	Khám phá dữ liệu Xây dựng mô hình Viết báo cáo

4. Đánh mức độ hoàn thành

- Tất cả các thành viên đều hoàn thành đúng hạn nhiệm vụ được phân công.
- Kết quả: Nhóm đã hoàn thành tất cả nội dung yêu cầu của đồ án
- Đánh giá mức độ hoàn thành mỗi cá nhân: 100%/100%.

II. Khám phá – tiền xử lý dữ liệu – tạo đặc trưng

*Note: **Khám phá dữ liệu (thường đan xen với pha tiền xử lý dữ liệu kèm theo quá trình tạo đặc trưng)** theo yêu cầu của đề được thực hiện rất chi tiết trong file:

data_preprocessing.ipynb và phần đầu của natural-language-precessing-with-disaster-tweets.ipynb nên chúng em không viết lại trong báo cáo này.

III. Xây dựng mô hình – Thử nghiệm – Đánh giá.

*Note: **Xây dựng mô hình – thử nghiệm – đánh giá** theo yêu cầu của đề được thực hiện rất chi tiết trong file: **data modeling.ipynb và phần sau của natural-language-precessing-with-disaster-tweets.ipynb** nên chúng em không viết lại trong báo cáo này.

IV. Kết luận và hướng phát triển:

1. ***Kết quả trên leaderboard của Kaggle:***

   DistilBert (kết quả khi so sánh với perfect submission):

   Bert Classiification model:

TÀI LIỆU THAM KHẢO

Tham khảo cuối mỗi file .ipynb.