



CANTHO UNIVERSITY

ĐẠI HỌC CẦN THƠ

TRƯỜNG CÔNG NGHỆ THÔNG TIN & TT

XÂY DỰNG MÔ HÌNH NHẬN DẠNG TWEETS LIÊN QUAN ĐẾN THẢM HỌA

Giảng viên hướng dẫn:

TS.Lưu Tiến Đạo

Sinh viên thực hiện:

Trần Nguyễn Nhật Huy B2113333

Lê Nhật Trọng B2106819

Nguyễn Phúc Hậu B2106790



CANTHO UNIVERSITY

NỘI DUNG TRÌNH BÀY

1. MÔ TẢ BÀI TOÁN
2. MÔ TẢ VÀ PHÂN TÍCH DỮ LIỆU
3. CÁC MÔ HÌNH HUẤN LUYỆN
4. KẾT QUẢ HUẤN LUYỆN VÀ ĐÁNH GIÁ MÔ HÌNH
5. TRIỂN KHAI HỆ THỐNG
6. TỔNG KẾT



















CANTHO UNIVERSITY

1. MÔ TẢ BÀI TOÁN



MÔ TẢ BÀI TOÁN

- Thực hiện xây dựng mô hình dự báo các Tweets trên nền tảng mạng xã hội Twitter có liên quan đến thảm họa.

Tweet	 WCUCOM Tornado @WCUCOM_Tornado · Apr 29 Be careful, storms are approaching the Hattiesburg area    
User Mention	 Paul @PaulC0071 · 2m @WCUCOM_Tornado tornado on Hardy near the USM. Take shelter! #Take Shelter 4:24 PM - 30 Nov 2014 · Details     Hide conversation
Retweet	 Paul retweeted  WCUCOM Tornado @WCUCOM_Tornado · Apr 29 Be careful, storms are approaching the Hattiesburg area   1  



- + Nội dung
- + Từ khóa
- + Thông tin phụ như: hashtag và vị trí địa lý (nếu có)





CANTHO UNIVERSITY

ĐẦU RA



- + 0: tweet không liên quan đến thảm họa
- + 1: tweet có liên quan đến thảm họa



CANTHO UNIVERSITY

2. MÔ TẢ VÀ PHÂN TÍCH DỮ LIỆU



MÔ TẢ DỮ LIỆU

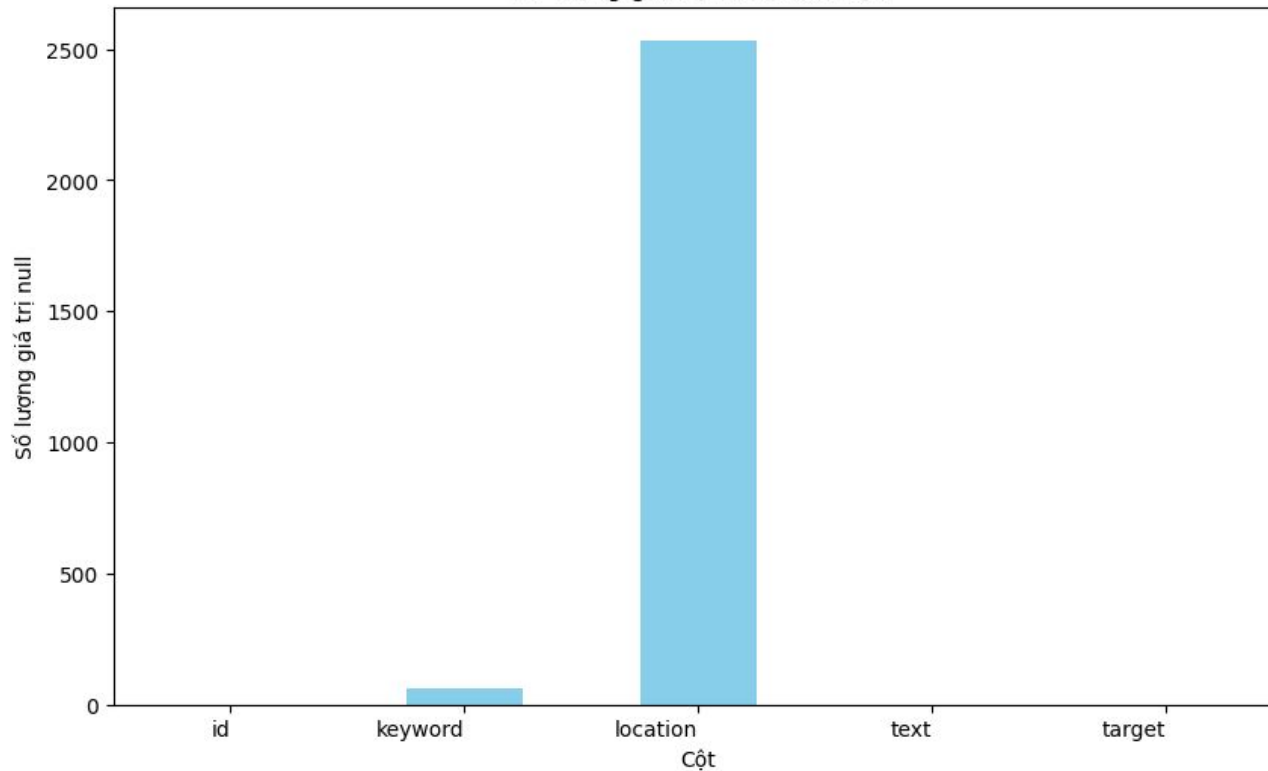
- Tập dữ liệu gồm 7613 mẫu, với 5 thuộc tính:
- + id: từ khóa chính của tập dữ liệu.
- + keyword: từ khóa có liên quan đến tweet.
- + location: vị trí có liên quan đến tweet.
- + text: thuộc tính quan trọng nhất, thể hiện nội dung của tweet, chứa thông tin liên quan thảm họa hoặc không.
- + target: có giá trị nhị phân, với 0 là không liên quan đến thảm họa, 1 là có liên quan đến thảm họa

	id	keyword	location	text	target
0	1	NaN	NaN	Our Deeds are the Reason of this #earthquake M...	1
1	4	NaN	NaN	Forest fire near La Ronge Sask. Canada	1
2	5	NaN	NaN	All residents asked to 'shelter in place' are ...	1
3	6	NaN	NaN	13,000 people receive #wildfires evacuation or...	1
4	7	NaN	NaN	Just got sent this photo from Ruby #Alaska as ...	1



CÁC GIÁ TRỊ NULL

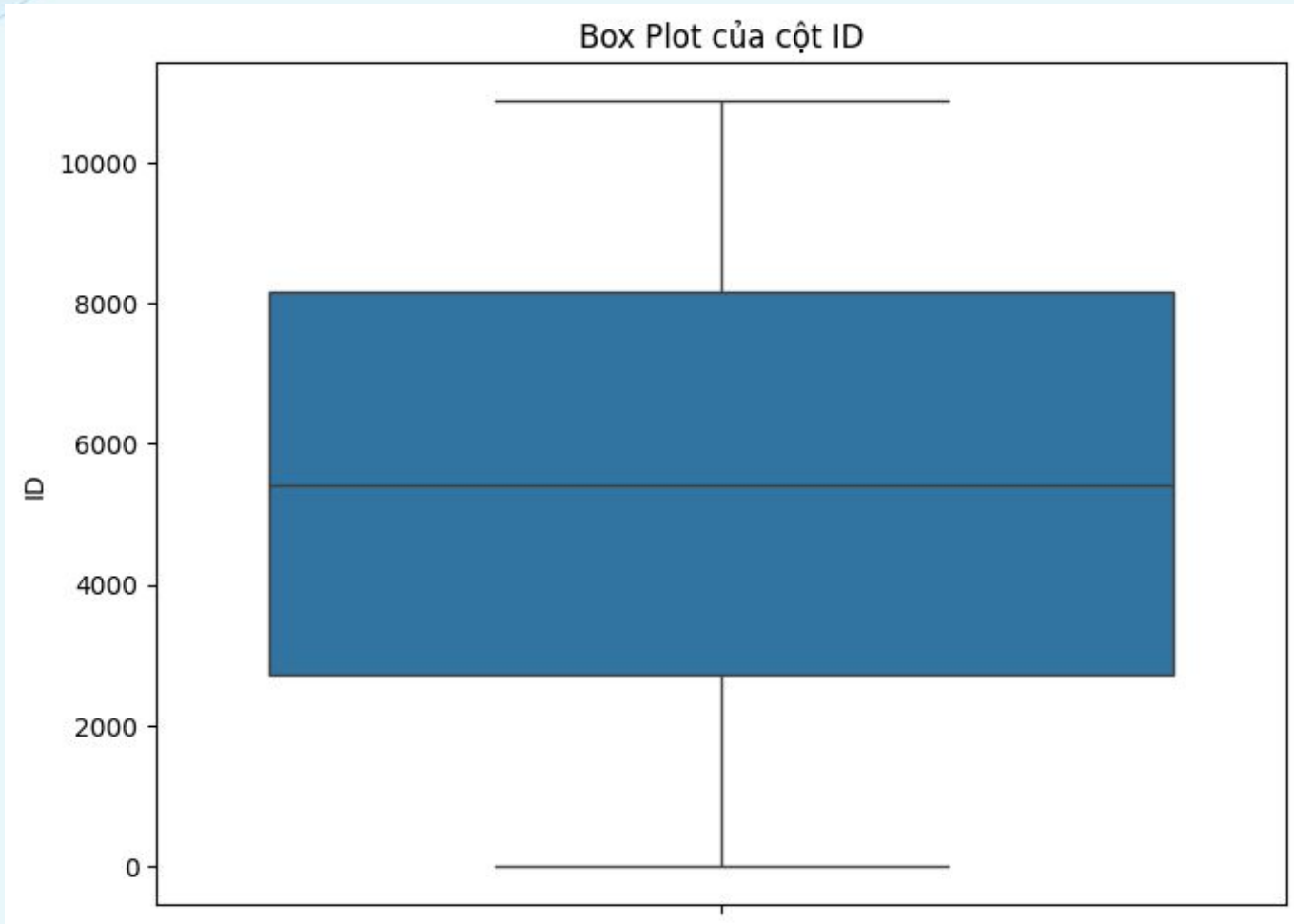
Số lượng giá trị null ở mỗi cột



	NullCount	Percentage
id	0	0.000000
keyword	61	0.801261
location	2534	33.285170
text	0	0.000000
target	0	0.000000



CỘT GIÁ TRỊ ID

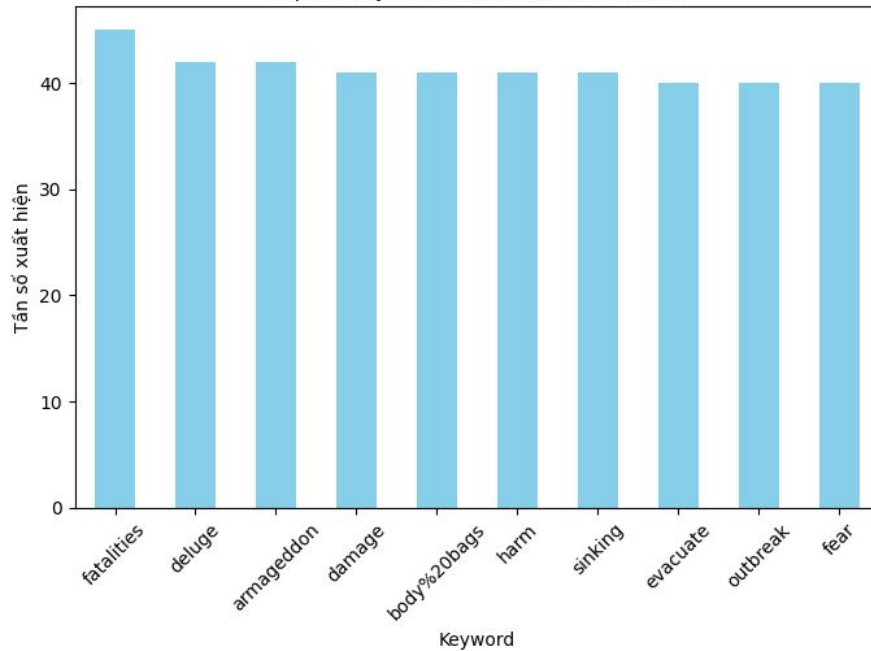




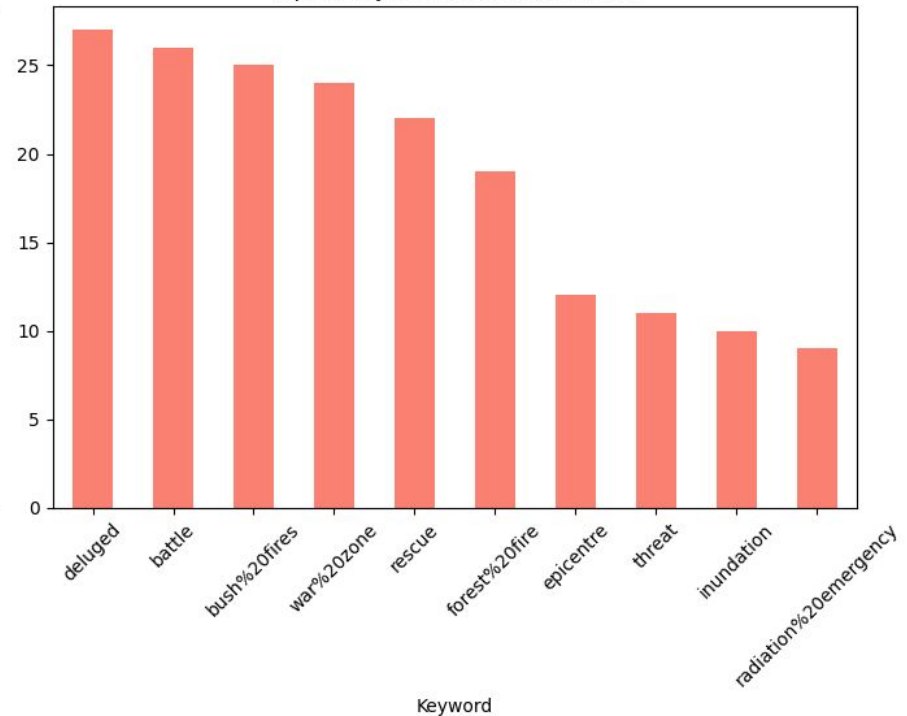
CỘT GIÁ TRỊ KEYWORD

Gồm các thuật ngữ như “derailment” (trật đường ray), “famine” (nạn đói), “earthquake” (động đất) và các sự kiện liên quan đến thảm họa tự nhiên và nhân tạo khác nhau

Top 10 Keywords xuất hiện nhiều nhất



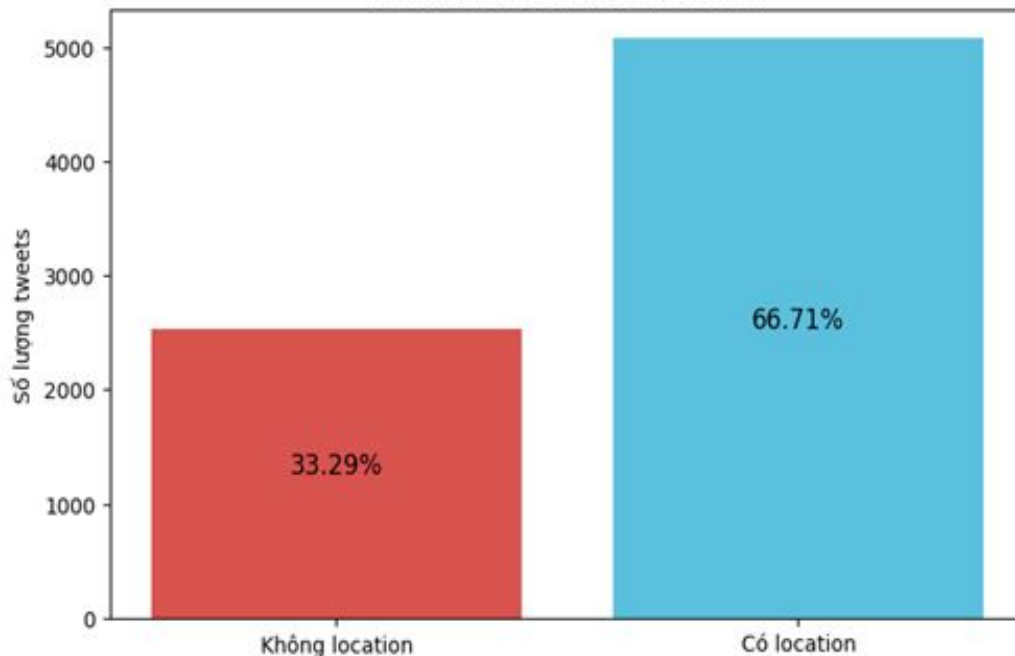
Top 10 Keywords xuất hiện ít nhất





CỘT GIÁ TRỊ LOCATION

Phân bố location vào các tweets



Các giá trị bị nhiễu như:

1324, <http://www.amazon.com/dp/B00HR>

1325, "New York, NY"

1326, [kate + they/them + infp-t]

107, "North Carolina"

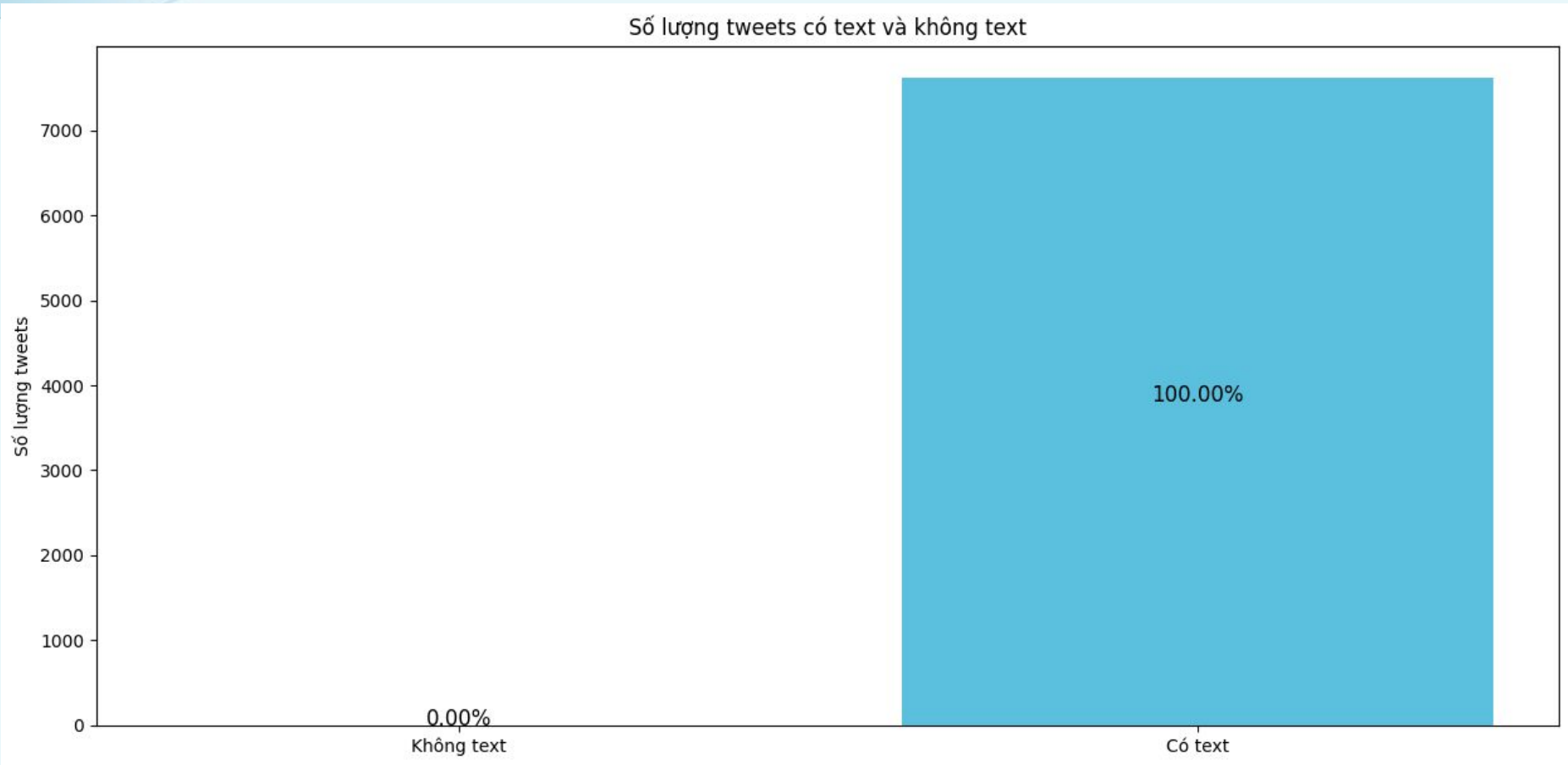
108, Instagram - @heyimginog

119, 304

136, "19.600858, -99.047821"



CỘT GIÁ TRỊ TEXT





CỘT GIÁ TRỊ TEXT

@bbcmtd Wholesale Markets ablaze <http://t.co/lHYXEOHY6C>

We always try to bring the heavy. #metal #RT <http://t.co/YAo1e0xngw>

#AFRICANBAZE: Breaking news:Nigeria flag set ablaze in Aba. <http://t.co/2nndBGwyEi>

Crying out for more! Set me ablaze

On plus side LOOK AT THE SKY LAST NIGHT IT WAS ABLAZE <http://t.co/qqsmsshaJ3N>

@PhDSquares #mufc they've built so much hype around new acquisitions but I doubt they will set the EPL ablaze this

INEC Office in Abia Set Ablaze - <http://t.co/3ImaomknnA>

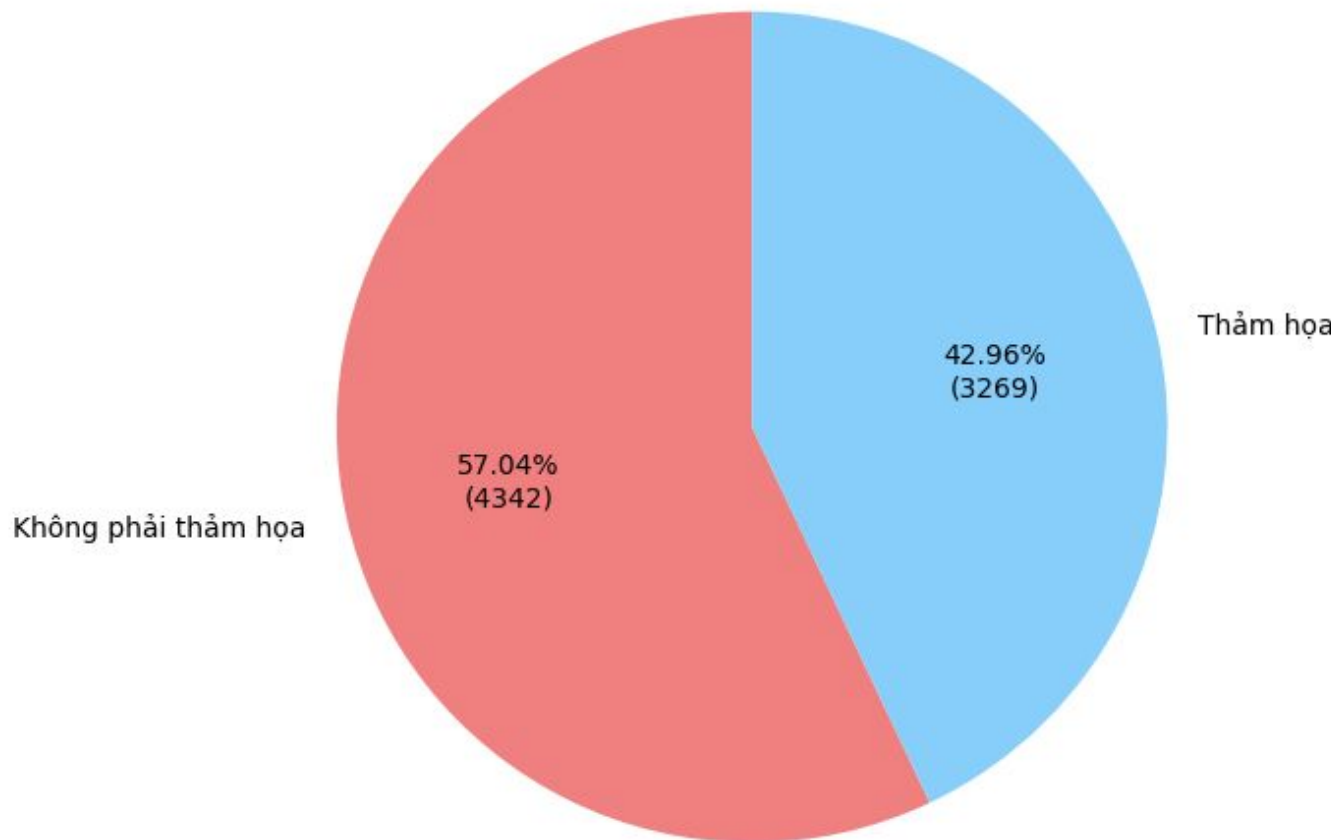
Barbados #Bridgetown JAMAICA 🇵🇸 Two cars set ablaze: SANTA CRUZ 🇵🇸 Head of the St Elizabeth Police Superintende..

Ablaze for you Lord :D



CỘT GIÁ TRỊ TARGET

Phân bố các tweets theo nhãn

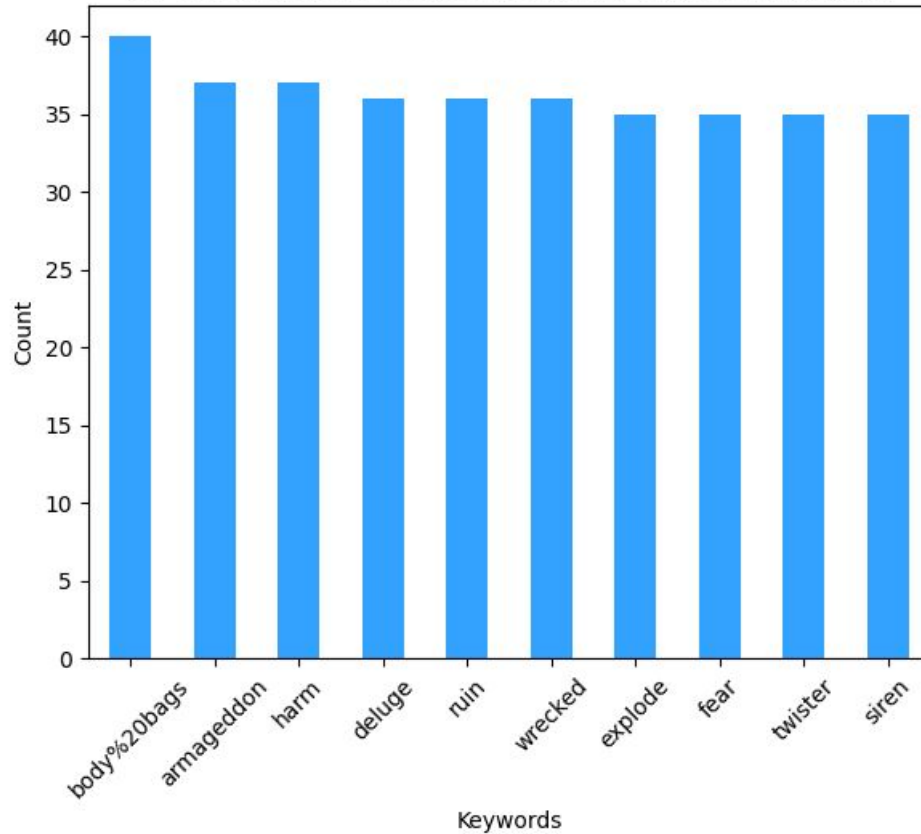




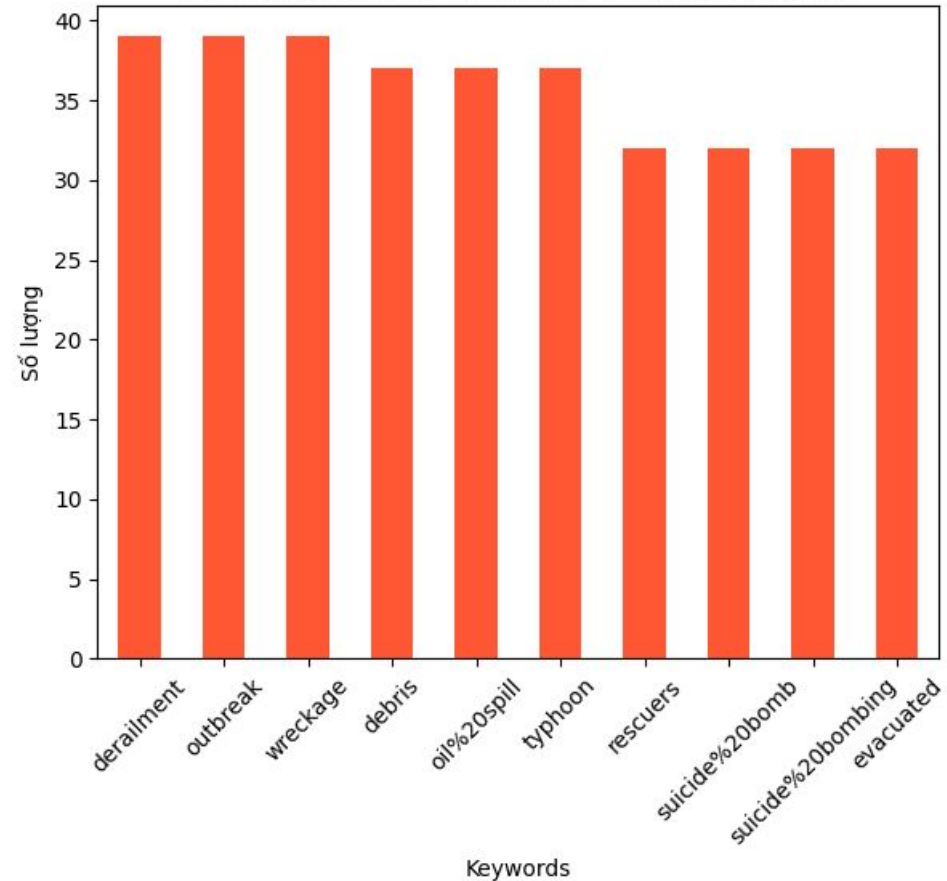
MỐI QUAN HỆ GIỮA KEYWORD VÀ TARGET

CANTHO UNIVERSITY

Top 10 keywords có tweets không liên quan đến thảm họa



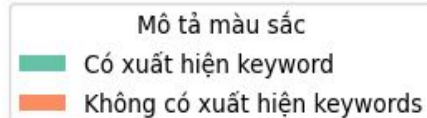
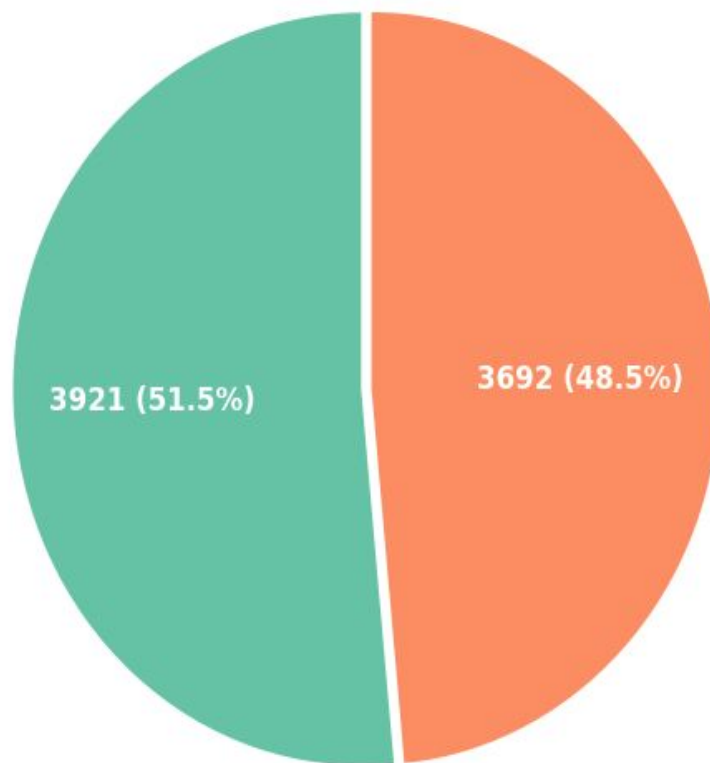
Top 10 keywords có tweets liên quan đến thảm họa





MỐI QUAN HỆ GIỮA KEYWORD VÀ TEXT

Tỉ lệ text có xuất hiện keyword



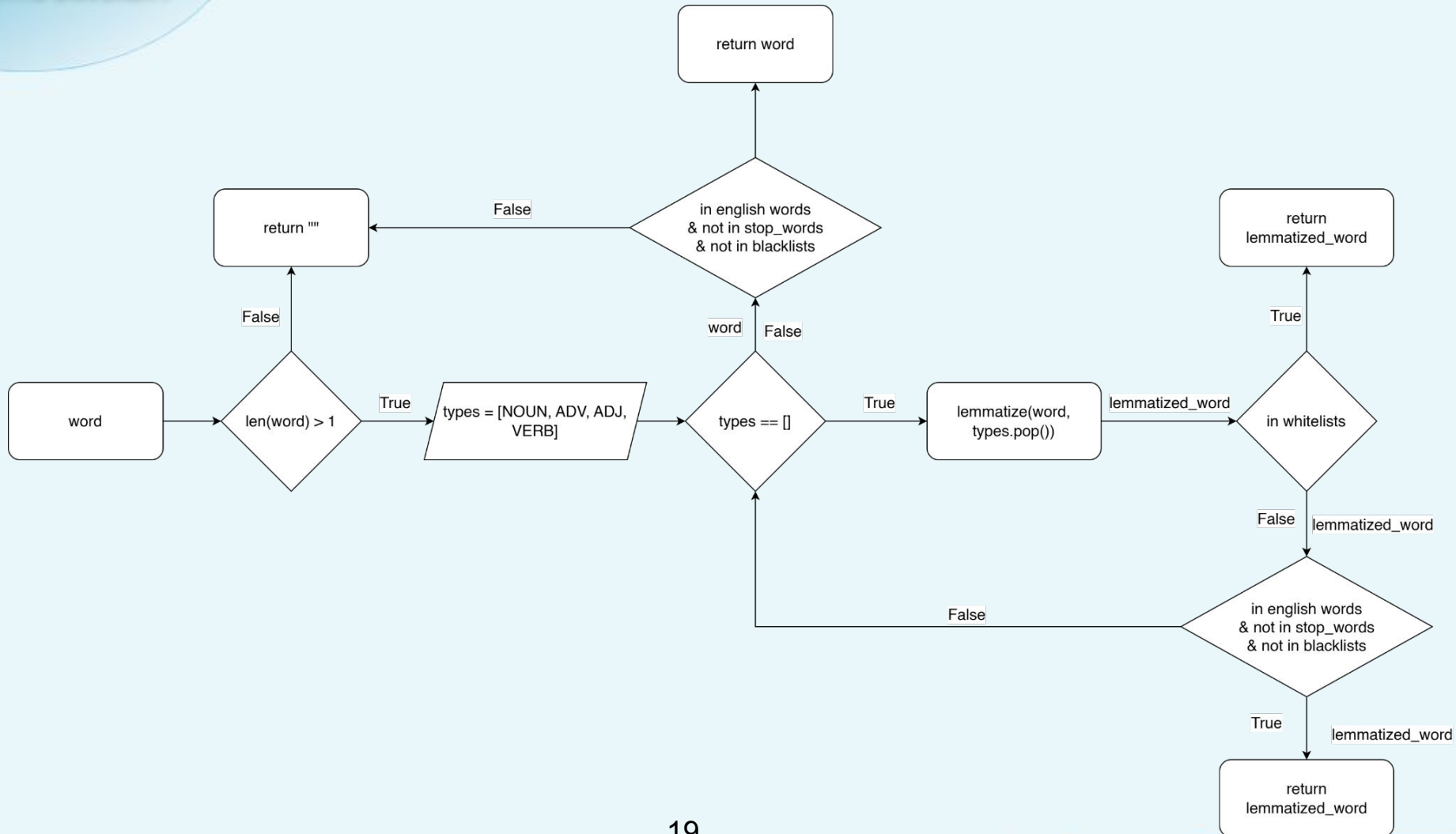


XỬ LÝ DỮ LIỆU

- Xử lý dữ liệu: bỏ một số cột giá trị để tránh ảnh hưởng đến quá trình huấn luyện và kết quả, cũng như giảm thiểu độ phức tạp của dữ liệu.
 - + **id**: các giá trị trong cột id không liên tục → bỏ hoàn toàn cột này
 - + **location**: số lượng tweets không có giá trị chiếm 33%, vị trí địa lý không ảnh hưởng ý nghĩa tweet → bỏ cột dữ liệu này
 - + **keyword**: có nhiều keyword không liên quan đến thảm họa sẽ làm nhiễu dữ liệu → không được dùng để huấn luyện mô hình
- Dữ liệu được đưa đi huấn luyện gồm 1 cột đặc trưng text và cột nhãn target

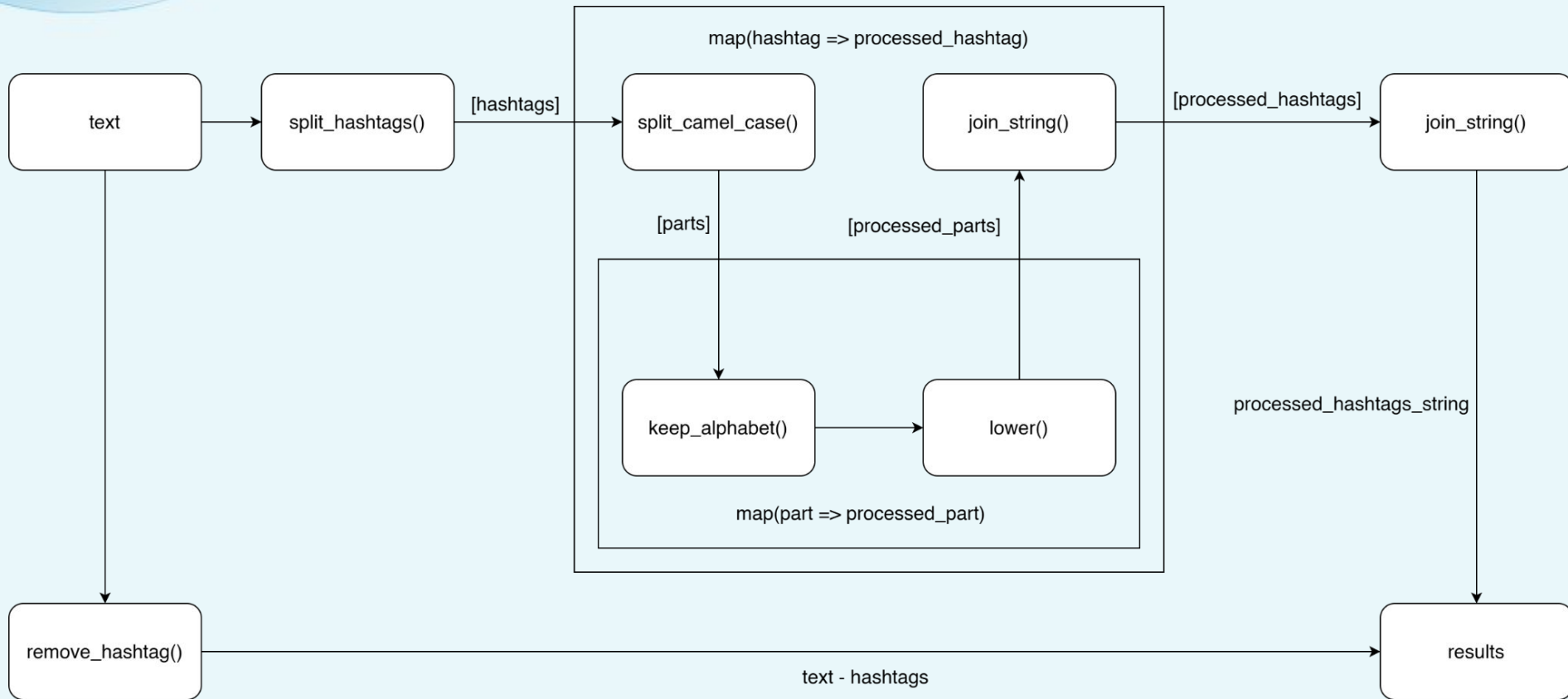


THIẾT KẾ HÀM LEMMATIZE_WORD



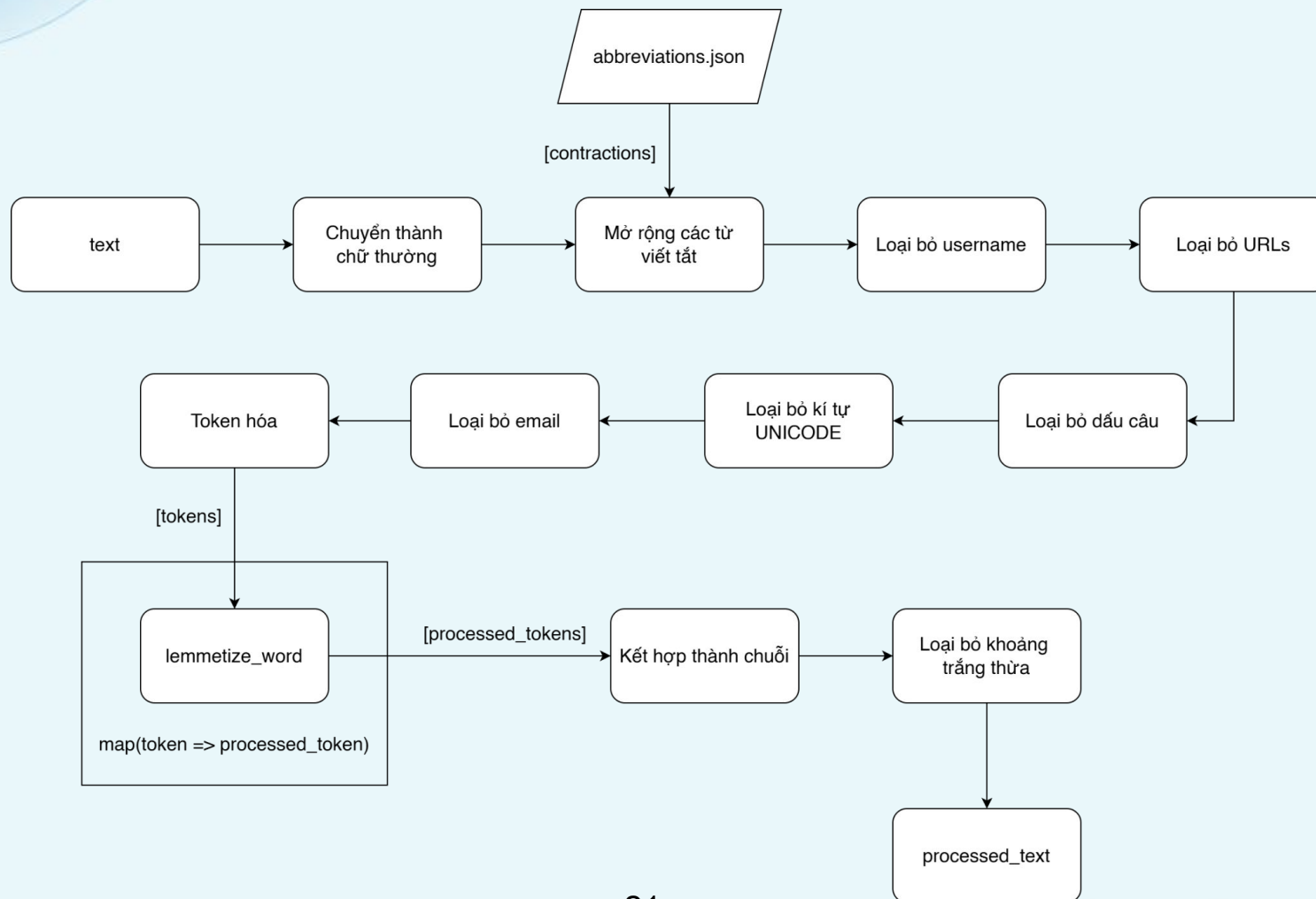


THIẾT KẾ HÀM PROCESS_HASHTAG



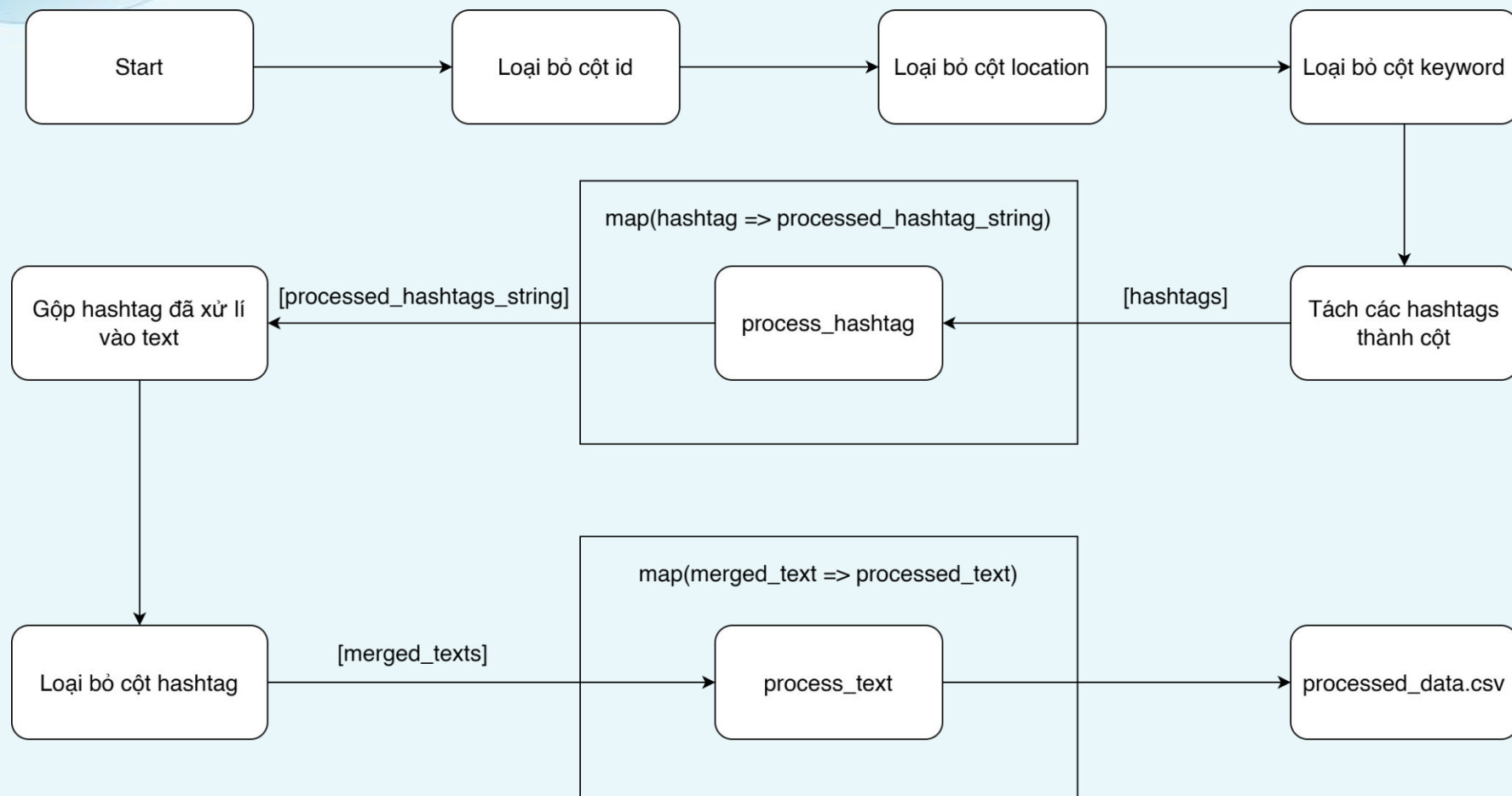


THIẾT KẾ HÀM PROCESS_TEXT





MÔ HÌNH TỔNG QUAN CÁC BƯỚC TIỀN XỬ LÝ DỮ LIỆU





DỮ LIỆU SAU XỬ LÝ

outside ablaze alive dead inside	0
awesome time visit head office site ablaze thank take care	0
pump ablaze	0
want set ablaze preach hotel	0
gain follower last week know grow	0
west burn thousand wildfire ablaze alone	1
build perfect life leave streets ablaze	0
check	0
first night retainer quite weird better get use wear every single night next year least	0
deputy man shoot home set ablaze	1



CANTHO UNIVERSITY

3. HUẤN LUYỆN MÔ HÌNH



TF - IDF (TERM FREQUENCY – INVERSE DOCUMENT FREQUENCY)

TF-IDF

$$tf \times idf$$

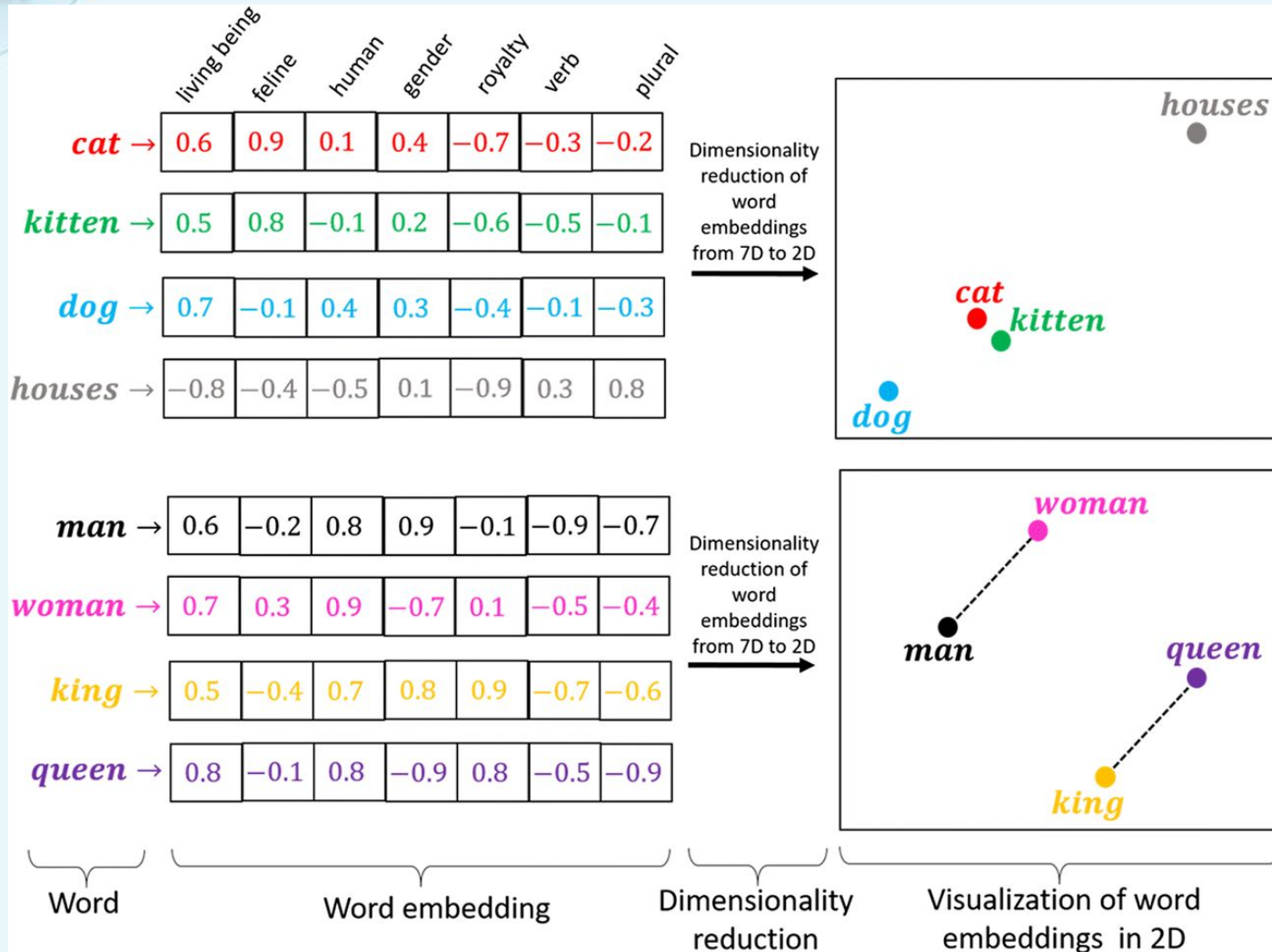
TF

IDF

$$tf(t,d) = count(t,d)$$

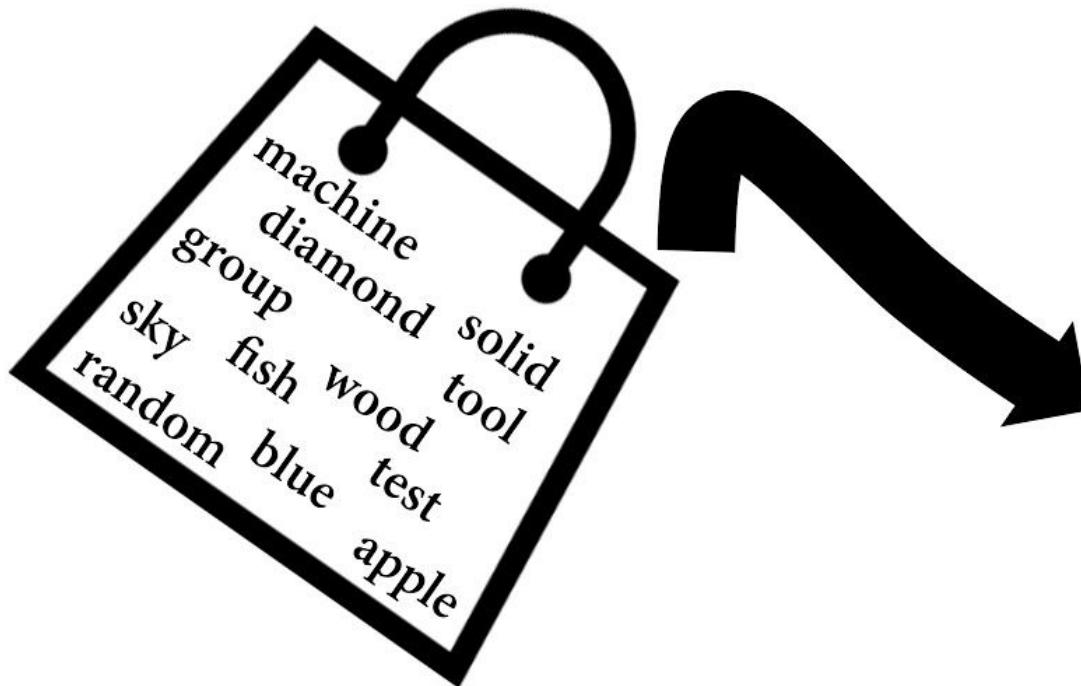
$$idf(t) = \log\left(\frac{1 + N_{documents}}{1 + df(t)}\right) + 1$$

WORD2VEC





BAGS OF WORDS



[1,1,0,0,1,0,2,3,0,1,0,2]
[0,0,1,2,0,0,1,1,0,3,0,0]
[2,1,1,0,0,0,0,0,0,0,1,1]
[0,0,2,1,2,1,0,0,1,0,0,0]
[1,1,0,0,1,1,0,0,1,0,0,0]
[0,2,0,1,0,0,1,0,3,1,1,1]
[1,1,0,1,1,1,0,1,0,0,0,1]



HUẤN LUYỆN MÔ HÌNH

Các mô hình huấn luyện:

- KNN
- Naïve Bayes
- Logistic Regression
- SVM Linear Kernel
- SVM Non-Linear Kernel
- Decision Tree
- Random Forest
- Mô hình học sâu:
 - + FNN với TF - IDF
 - + FNN với Bag of Words
 - + RNN với tầng Embedding



CÁC MÔ HÌNH CỔ ĐIỂN

```
"KNN": KNeighborsClassifier(n_neighbors=7),
```

```
"Bayes": MultinomialNB(),
```



CÁC MÔ HÌNH CỔ ĐIỂN

```
"Decision Tree": DecisionTreeClassifier(  
    max_depth=50, min_samples_split=4,  
    criterion='gini', random_state=42),
```

```
"Random Forest": RandomForestClassifier(  
    n_estimators=200, max_depth=100,  
    min_samples_split=4, criterion='entropy'),
```



CÁC MÔ HÌNH CỔ ĐIỂN

```
"Logistic Regression": LogisticRegression(  
    C=0.1, random_state=42, max_iter=1000),
```

```
"SVM Linear": SVC(kernel='linear',  
    C=1.0, random_state=42),
```

```
"SVM Non-linear": SVC(kernel='rbf',  
    C=1.0, gamma='scale',  
    random_state=42),
```



FNN với TF - IDF

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 64)	402,368
dropout (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 32)	2,080
dropout_1 (Dropout)	(None, 32)	0
dense_2 (Dense)	(None, 1)	33

Total params: 404,481 (1.54 MB)

Trainable params: 404,481 (1.54 MB)



FNN VỚI BAG OF WORDS

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 64)	402,432
dropout (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 32)	2,080
dropout_1 (Dropout)	(None, 32)	0
dense_2 (Dense)	(None, 1)	33

Total params: 404,545 (1.54 MB)

Trainable params: 404,545 (1.54 MB)



RNN VỚI TẦNG EMBEDDING

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 21, 64)	402,368
lstm (LSTM)	(None, 21, 64)	33,024
dropout (Dropout)	(None, 21, 64)	0
lstm_1 (LSTM)	(None, 64)	33,024
dropout_1 (Dropout)	(None, 64)	0
dense (Dense)	(None, 1)	65

Total params: 1,405,445 (5.36 MB)

Trainable params: 468,481 (1.79 MB)

Non-trainable params: 0 (0.00 B)

Optimizer params: 936,964 (3.57 MB)



4. KẾT QUẢ HUẤN LUYỆN VÀ ĐÁNH GIÁ MÔ HÌNH



ACCURACY

	KNN	Naives Bayes	Logistic Regression	SVM Linear	SVM Non Linear	Decision Tree	Random Forest
TF-IDF	0.7634	0.7906	0.7388	0.7941	0.7983	0.7501	0.7796
Word2Vec	0.6479	0.5411	0.5640	0.5718	0.5694	0.6260	0.7054
Bag of Words	0.7095	0.7883	0.7898	0.7735	0.8001	0.7507	0.7772



PRECISION

	KNN	Naives Bayes	Logistic Regression	SVM Linear	SVM Non Linear	Decision Tree	Random Forest
TF-IDF	0.7776	0.8170	0.9295	0.8124	0.8575	0.8038	0.8545
Word2Vec	0.6320	0.4772	0.4457	0.8583	0.0000	0.5664	0.7422
Bag of Words	0.8260	0.7743	0.8304	0.7588	0.8506	0.7940	0.8431



RECALL

CANTHO UNIVERSITY

	KNN	Naives Bayes	Logistic Regression	SVM Linear	SVM Non Linear	Decision Tree	Random Forest
TF-IDF	0.6347	0.6651	0.4292	0.6817	0.6406	0.5588	0.5912
Word2Vec	0.4370	0.6834	0.0478	0.0061	0.0000	0.5617	0.4839
Bag of Words	0.4175	0.7215	0.6472	0.6993	0.6532	0.5729	0.5964



F1-SCORE

	KNN	Naives Bayes	Logistic Regression	SVM Linear	SVM Non Linear	Decision Tree	Random Forest
TF-IDF	0.6987	0.7330	0.5868	0.7411	0.7330	0.6590	0.6987
Word2Vec	0.5165	0.5618	0.0858	0.0121	0.0000	0.5639	0.5857
Bag of Words	0.5541	0.7468	0.7272	0.7277	0.7387	0.6654	0.6985



CÁC MÔ HÌNH HỌC SÂU

	Accuracy	Precision	Recall	F1
FNN & TF-IDF	0.81	0.81	0.79	0.79
FNN & Bag of Words	0.80	0.80	0.79	0.79
RNN & Embedding	0.78	0.78	0.78	0.78



MÔ HÌNH BERT

	Accuracy	Precision	Recall	F1
BERT	0.81	0.78	0.78	0.78



ĐÁNH GIÁ MÔ HÌNH

Các mô hình đều có kết quả gần bằng nhau, trong đó:

- Tốt nhất: học sâu, TF - IDF

	Accuracy	Precision	Recall	F1
CNN & TF-IDF	0.81	0.81	0.79	0.79



ĐÁNH GIÁ MÔ HÌNH

Các mô hình đều có kết quả gần bằng nhau, trong đó:

- Xấu nhất: SVM, Word2vec

	KNN	Naives Bayes	Logistic Regression	SVM Linear	SVM Non Linear	Decision Tree	Random Forest
Word2Vec	0.5165	0.5618	0.0858	0.0121	0.0000	0.5639	0.5857



CANTHO UNIVERSITY

5. TRIỂN KHAI HỆ THỐNG



CANTHO UNIVERSITY

CÔNG NGHỆ



FastAPI

NEXT.JS



Render



CANTHO UNIVERSITY

DEMO CHƯƠNG TRÌNH



CANTHO UNIVERSITY

5. TỔNG KẾT

TỔNG KẾT

Kết quả đạt được:

- Xử lý dữ liệu dư thừa và bị thiếu thông tin.
- Xây dựng và đánh giá tổng quan dưới nhiều mô hình.
- Triển khai ứng dụng Web cho phép kiểm tra nội dung đoạn Tweet có liên quan đến thảm họa hay không.

Hướng phát triển:

- Mở rộng phạm vi nghiên cứu để phân loại chi tiết các thảm họa.
- Phát hiện thảm họa và kèm theo biện pháp phòng tránh.



CANTHO UNIVERSITY

THANKS FOR LISTENING