

INTRO TO IT

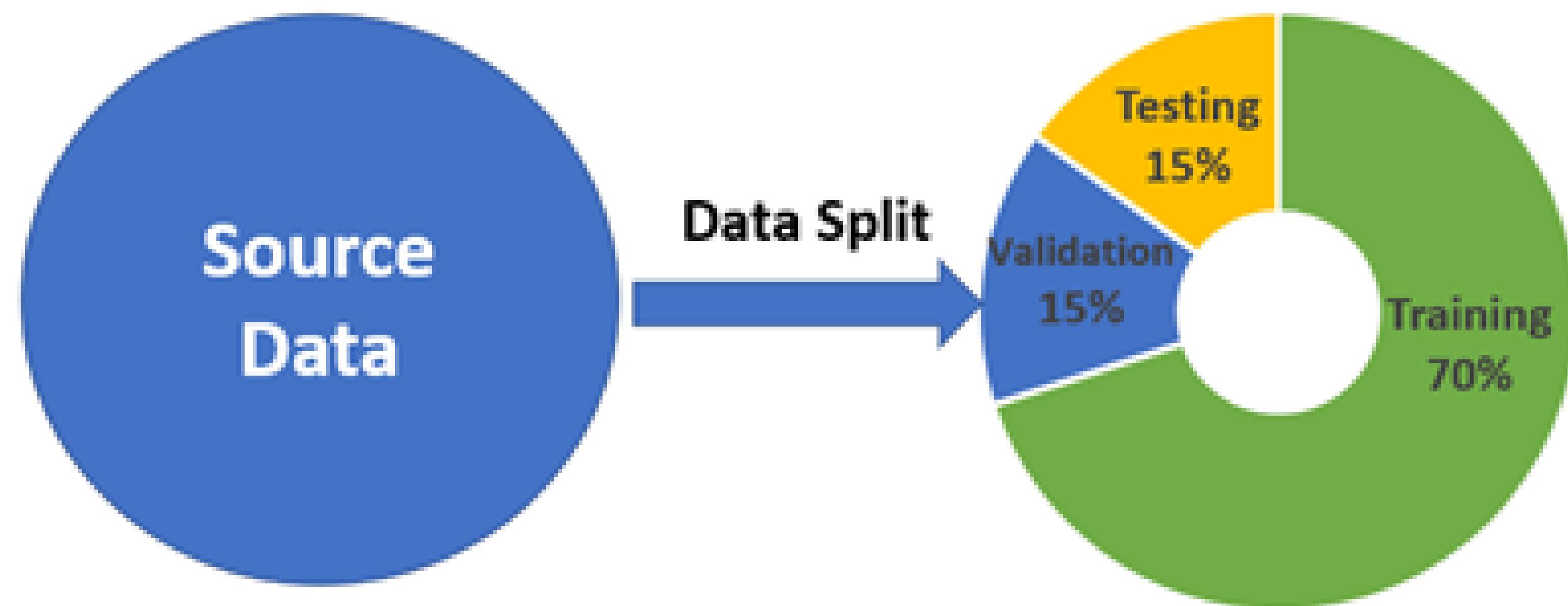
HATESPEECH CLASSIFICATION

BY TRẦN PHÚC HẢI

HateXplain là dataset tiêu chuẩn đầu tiên về hate speech*

THÔNG TIN VỀ DATASET

- **Targets:** African, Islam, Jewish, Homosexual, Women, Refugee, Arab, Caucasian, Hispanic, and Asian.
- Bao gồm 20148 dòng dữ liệu*



QUÁ TRÌNH HUẤN LUYỆN

- Sử dụng mô hình BERT đã huấn luyện sẵn*
- Huấn luyện
 - 1. Dùng MLM để học domain-specific patterns*
 - 2. Dùng sequence classification để học dự đoán phân loại hate speech

MLM

- Original sentence: "The cat is an animal."
- Masked sentence: "The **cat** is an [MASK]."
- Mô hình sẽ học cách dự đoán [MASK] dựa vào ngữ cảnh (các từ xung quanh)
 - > Mô hình sẽ hiểu ngữ cảnh, độ liên quan của các từ (cụ thể là đối với văn bản chứa hate speech)
- MLM ở cấp từ, cấp câu -> Kiến thức học được trong quá trình MLM sẽ được sử dụng cho Sequence classification để phân loại toàn bộ đoạn văn bản

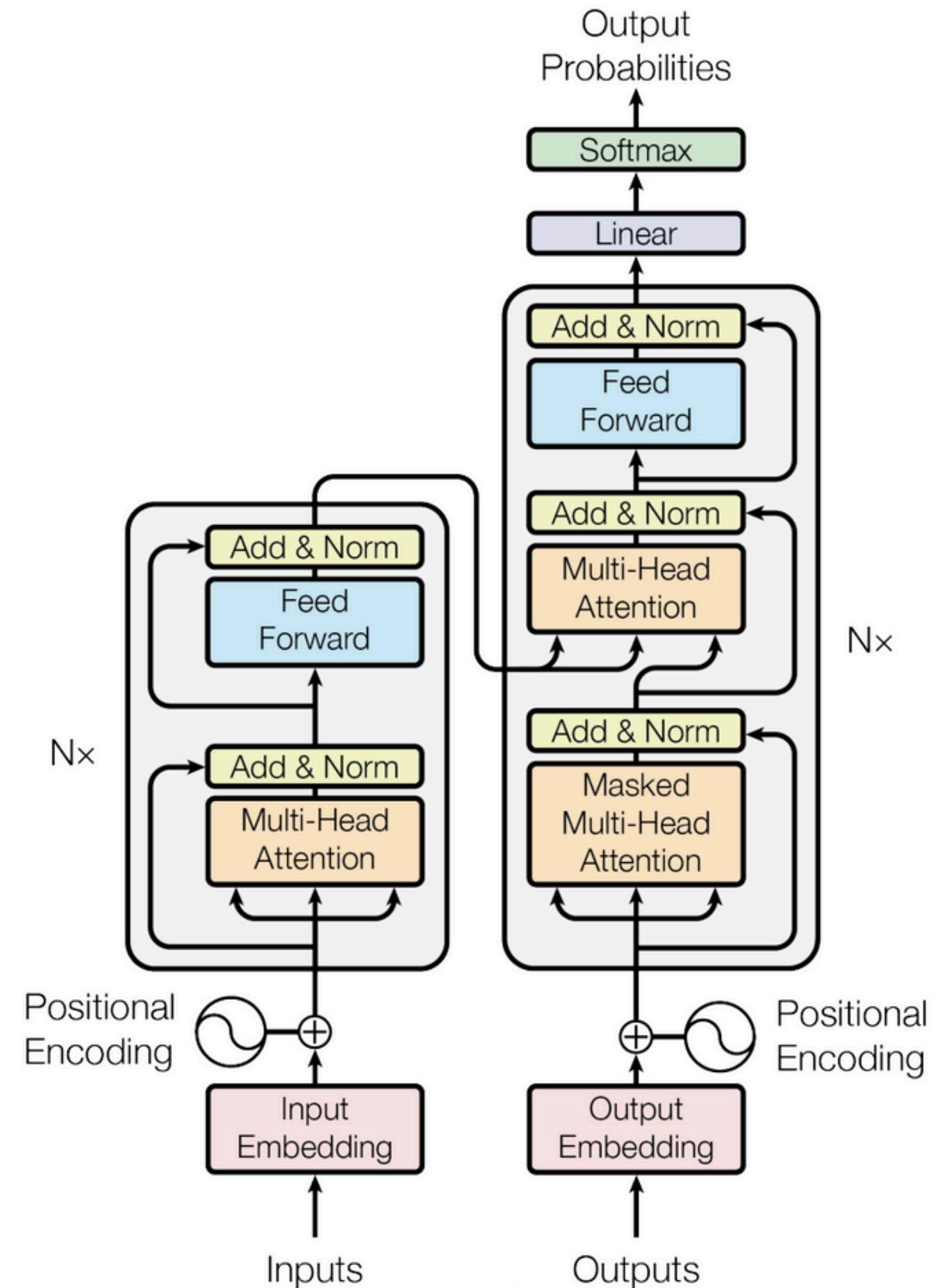
SEQUENCE CLASSIFICATION

- Dùng để gán nhãn (phân loại) cho một đoạn văn bản (bài twitter post có thể chứa hate speech)
- Sử dụng kiến trúc transformer đang hiện đại nhất bây giờ cho NLP (cũng chung cho ChatGPT)

TRANSFORMERS ARCHITECTURE

BERT

Encoder



GPT

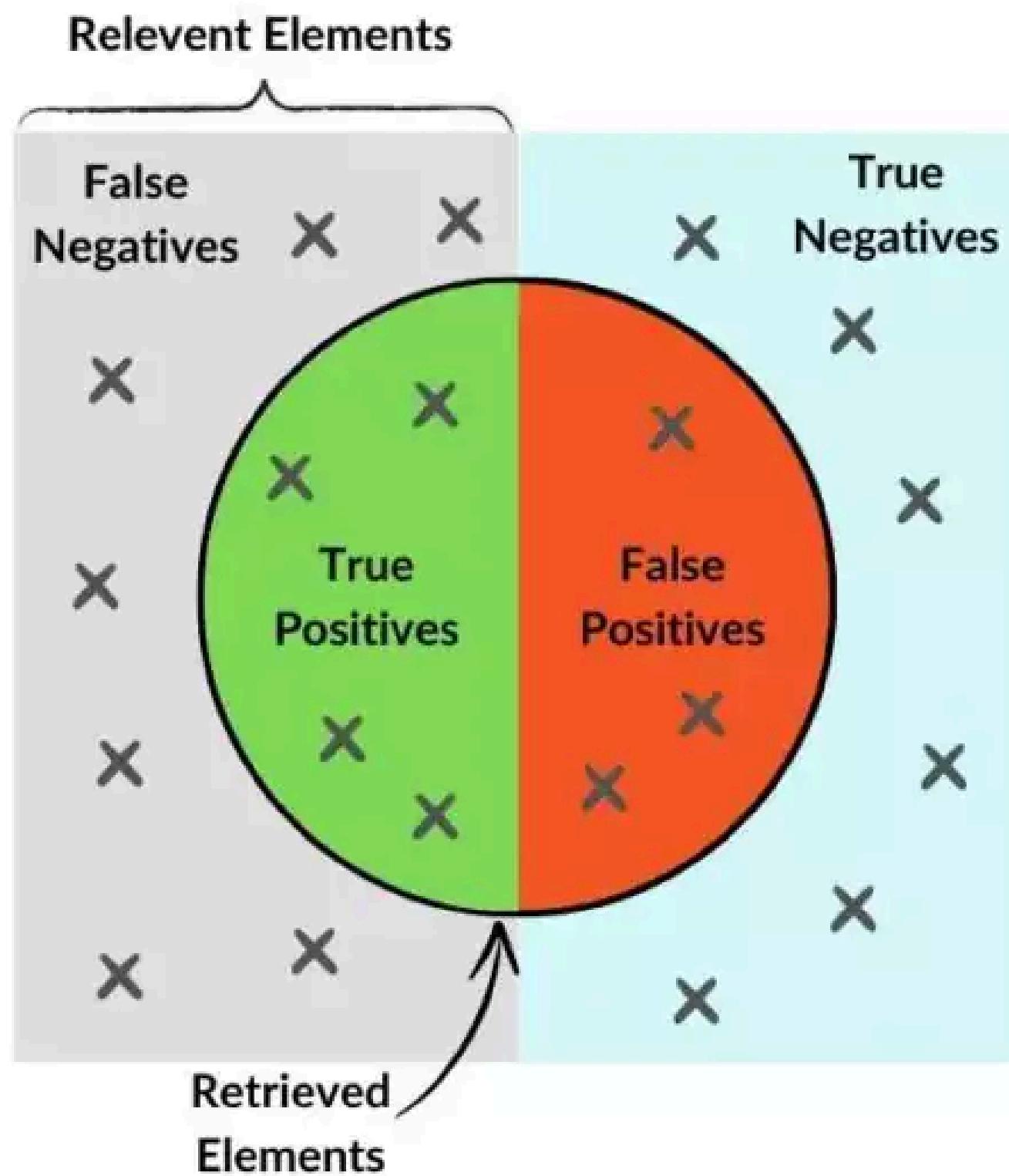
Decoder

ĐÁNH GIÁ KẾT QUẢ

AUROC: 0.858129956052606

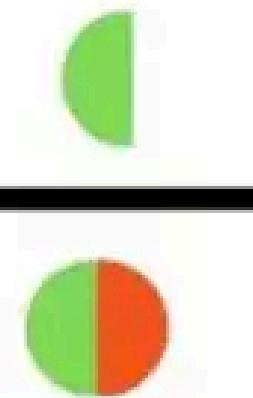
	precision	recall	f1-score	support
hatespeech	0.730	0.830	0.777	1187
normal	0.730	0.758	0.744	1563
offensive	0.600	0.478	0.532	1096
accuracy			0.700	3846
macro avg	0.687	0.689	0.684	3846
weighted avg	0.693	0.700	0.694	3846

ĐÁNH GIÁ KẾT QUẢ



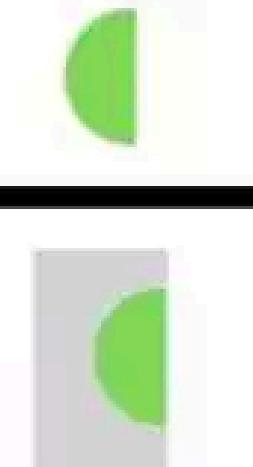
How many retrieved elements are relevant?

$$\text{Precision} = \frac{\text{True Positives}}{\text{Retrieved Elements}}$$



How many relevant elements are retrieved?

$$\text{Recall} = \frac{\text{True Positives}}{\text{Relevant Elements}}$$

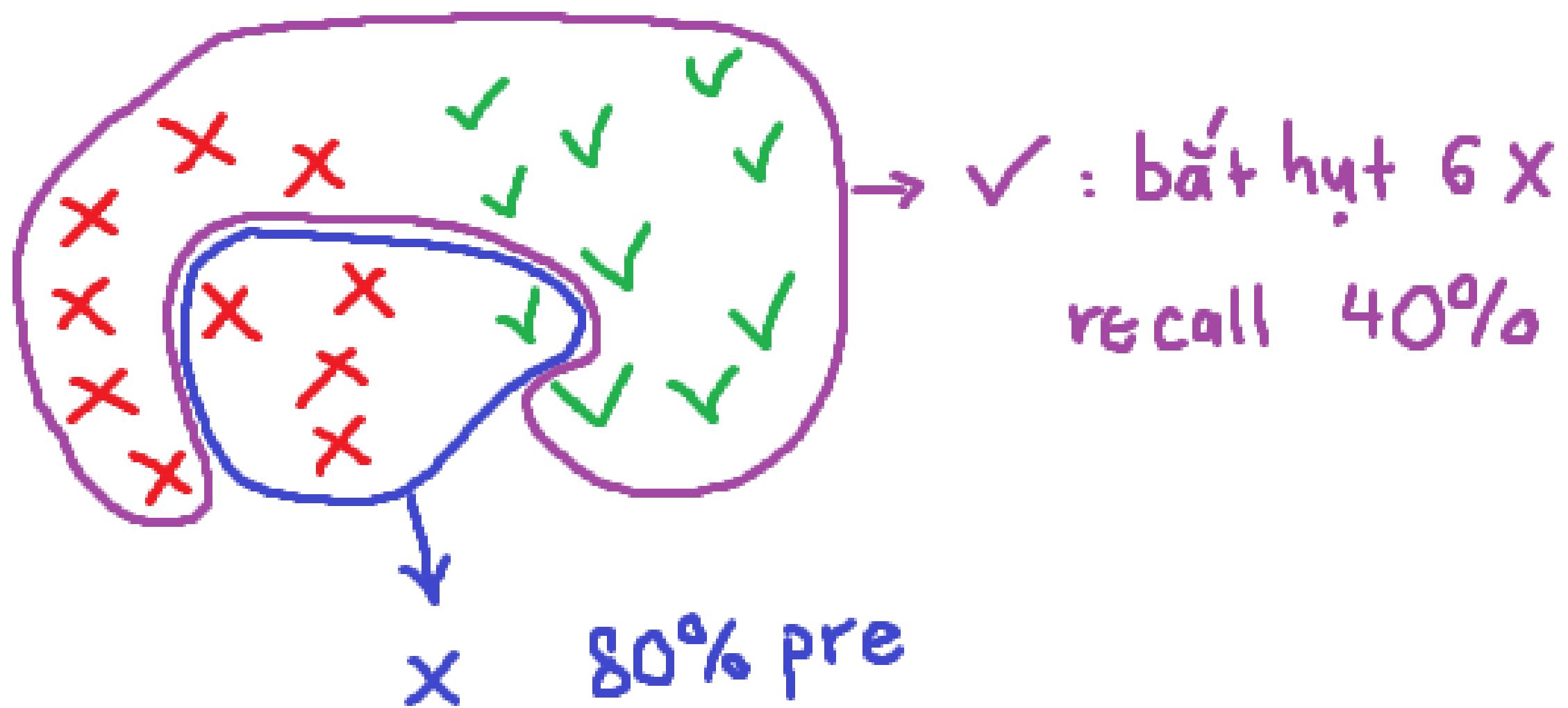


- 2 trường hợp dự đoán đúng với thực tế:
 - **True Positive (TP)**: Mô hình dự đoán **Yes*** và thực tế chính xác là **Yes**
 - **True Negative (TN)**: Mô hình dự đoán **No** và thực tế chính xác là **No**
- 2 trường hợp dự đoán sai với thực tế:
 - **False Positive (FP)**: Mô hình dự đoán **Yes** trong khi thực tế là **No**
 - **False Negative (FN)**: Mô hình dự đoán **No** trong khi thực tế là **Yes**
- Về trước về sau*

ĐÁNH GIÁ KẾT QUẢ

- **Precision** cho biết tỉ lệ dự đoán chính xác đối với các dự đoán là Yes
 - Ví dụ: Email spam detection
- **Recall** cho biết tỉ lệ dự đoán chính xác đối với các trường hợp thực tế là Yes
 - Ví dụ: Tỉ lệ phát hiện bệnh trên các bệnh nhân

ĐÁNH GIÁ KẾT QUẢ



-> Sử dụng F-score để có sự cân bằng:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

AUROC: 0.858129956052606

	precision	recall	f1-score	support
hatespeech	0.730	0.830	0.777	1187
normal	0.730	0.758	0.744	1563
offensive	0.600	0.478	0.532	1096
accuracy			0.700	3846
macro avg	0.687	0.689	0.684	3846
weighted avg	0.693	0.700	0.694	3846

- **Mong muốn:** cả 3 cột đều cao cho từng loại phân loại (hatespeech, normal, offensive)
- AUROC*

SO SÁNH CÁC MÔ HÌNH KHÁC

Model	AUROC	Accuracy	Macro F1
BiRNN-HateXplain [Attn]	0.805	0.629	0.629
BiRNN-HateXplain [LIME]	0.843	0.629	0.629
BERT-HateXplain [Attn]	0.851	0.698	0.687
BERT-HateXplain [LIME]	0.851	0.698	0.687
BertMLM + Bert Classifier (LIME)	0.858	0.700	0.684

TÀI LIỆU THAM KHẢO

- R. Mathew et al., "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection," <https://arxiv.org/pdf/2012.10289.pdf>
- J. Kim, B. Lee, and K.-A. Sohn, "Why Is It Hate Speech? Masked Rationale Prediction for Explainable Hate Speech Detection," <https://arxiv.org/pdf/2211.00243v1.pdf>
- Papers with code - HateXplain: A benchmark dataset for explainable hate speech detection HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection | Papers With Code. Available at: <https://paperswithcode.com/paper/hatexplain-a-benchmark-dataset-for-explainable-hate-speech-detection>
- Harshmeetsingh Chandhok, “Stopping The Hate,” [Stopping The Hate](#) | Devpost

**THANKS FOR
LISTENING**

