Phân loại bình luận đánh giá sản phẩm di động

Học máy thống kê - Nhóm 8

OUR TEAM

Dương Huy Hoàng

Nguyễn Vũ Hoàng Long

Phạm Quốc Anh Khoa

BÀI TOÁN

Phân loại bình luận thành các nhóm chính như tích cực, tiêu cực, hoặc trung tính để hiểu cảm nhận tổng thể của người dùng
Xác định các khía cạnh cụ thể của sản phẩm mà người dùng đề cập đến trong bình luận, ví dụ như chất lượng camera, hiệu suất, thiết kế, giá cả, v.v.

MỤC TIÊU

Giúp đỡ các shop bán lẻ quản lý chất lượng sản phẩm, kiểm soát bình luận

ĐỐI TƯỢNG

các bình luận tiếng Việt trên các trang web thương mại điện tử như Shopee, Tiki, Lazada, Sendo, etc

PHẠM VI

DATASET

NHÓM 8

TỔNG QUAN

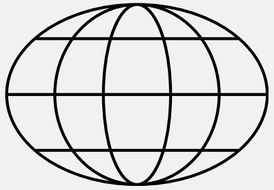
UIT-ViSFD thu thập phản hồi bằng văn bản từ khách hàng về điện thoại thông minh trên một trang thương mại điện tử lớn tại Việt Nam. Nhãn của tập dữ liệu là mười class (GENERAL, SCREEN, CAMERA, FEATURES, BATTERY, PERFORMANCE, STORAGE, DESIGN, PRICE, SER&ACC, OTHERS) và ba lớp (Positive, Negative và Neutral)

Train:	7.786 samples
Dev:	1.112 samples
Test:	2.224 samples

TIỀN XỬ LÝ DỮ LIỆU

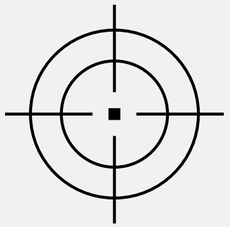
LOẠI BỎ CỘT KHÔNG CẦN THIẾT

Nhóm chú trọng phân tích cảm xúc phần bình luận dưới dạng text, không bao gồm số sao đánh giá sản phẩm và thông tin ngày một bình luận được người dùng đăng lên các trang thương mại điện tử. Do đó cột n_star, date_time, index cần được loại bỏ



TẠO THÊM CỘT TƯƠNG ỨNG VỚI SỐ LƯỢNG CÁC KHÍA CẠNH

Label gốc có dạng là chuỗi các khía cạnh và lớp tương ứng nối tiếp nhau dưới dạng string rất khó nhìn và khó thao tác với trường dữ liệu này. Do đó, nhóm sẽ tách các nhãn thành các cột riêng như hình bên dưới.



LABEL ENCODING

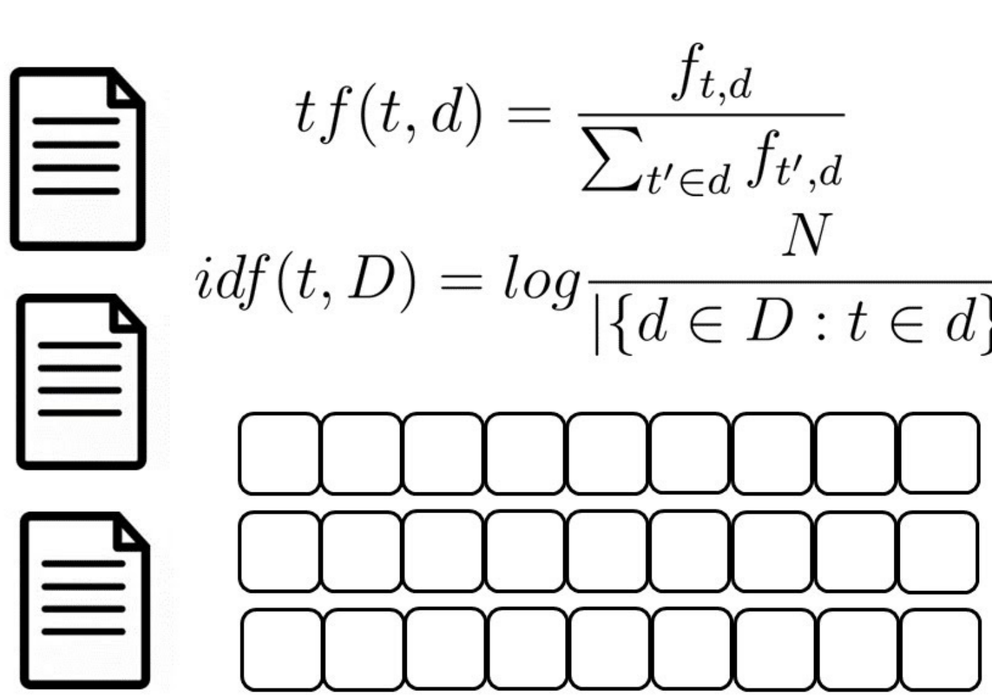
Phần lớn các nhãn có 3 lớp là ‘Positive’, ‘Negative’ và ‘Neutral’ đều ở dạng chữ, vì vậy nhóm đã sử dụng kỹ thuật label encoding để chuyển ‘Positive’ thành ‘1’, ‘Neutral’ thành ‘0’ và ‘Negative’ thành ‘-1’

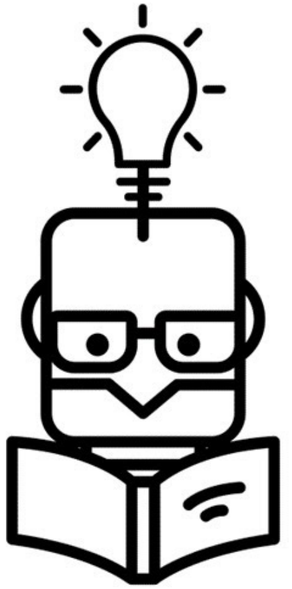
VECTOR HÓA

TF-IDF

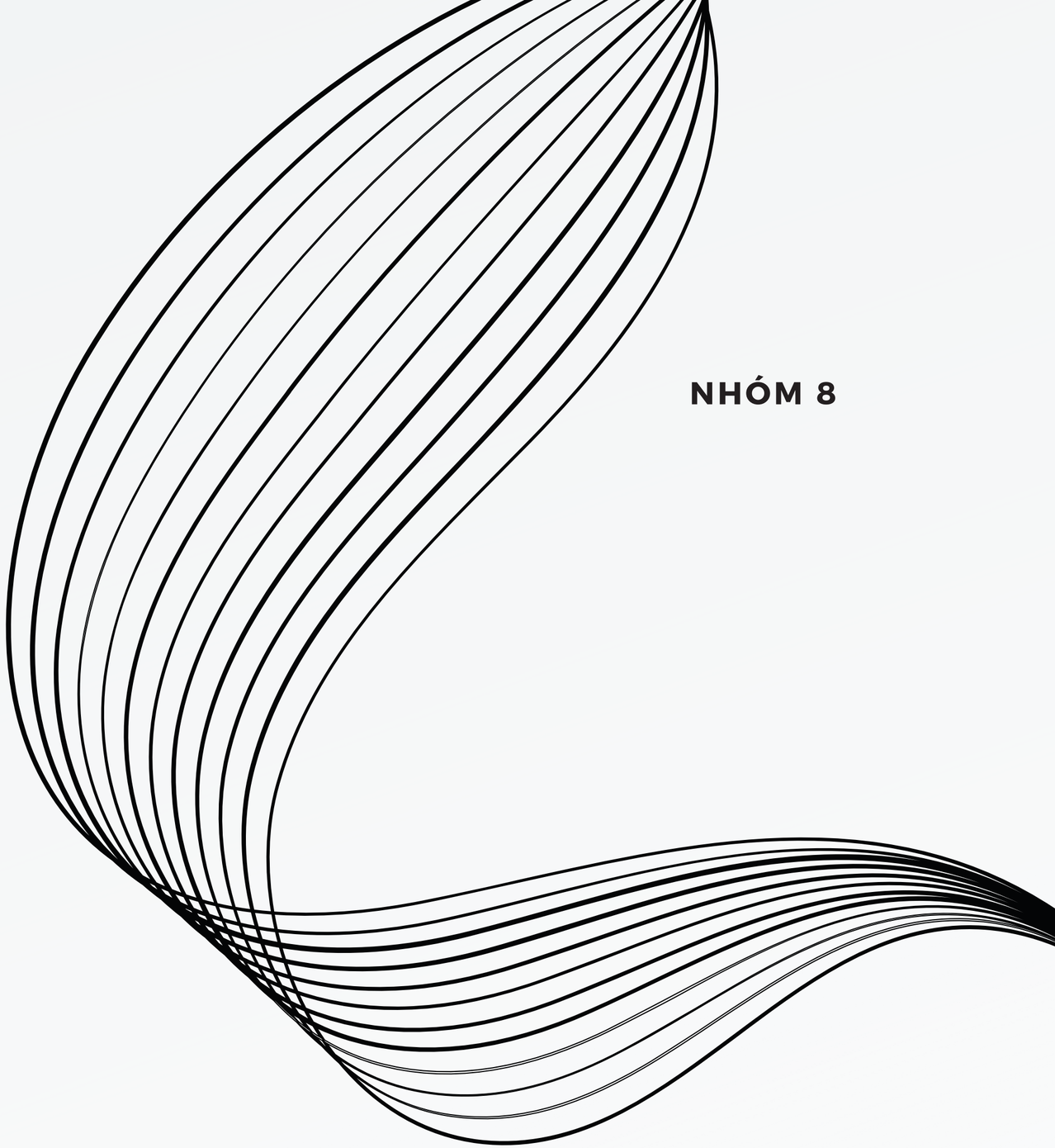
$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

$$idf(t,D) = log \frac{N}{|\{d \in D : t \in d\}|}$$

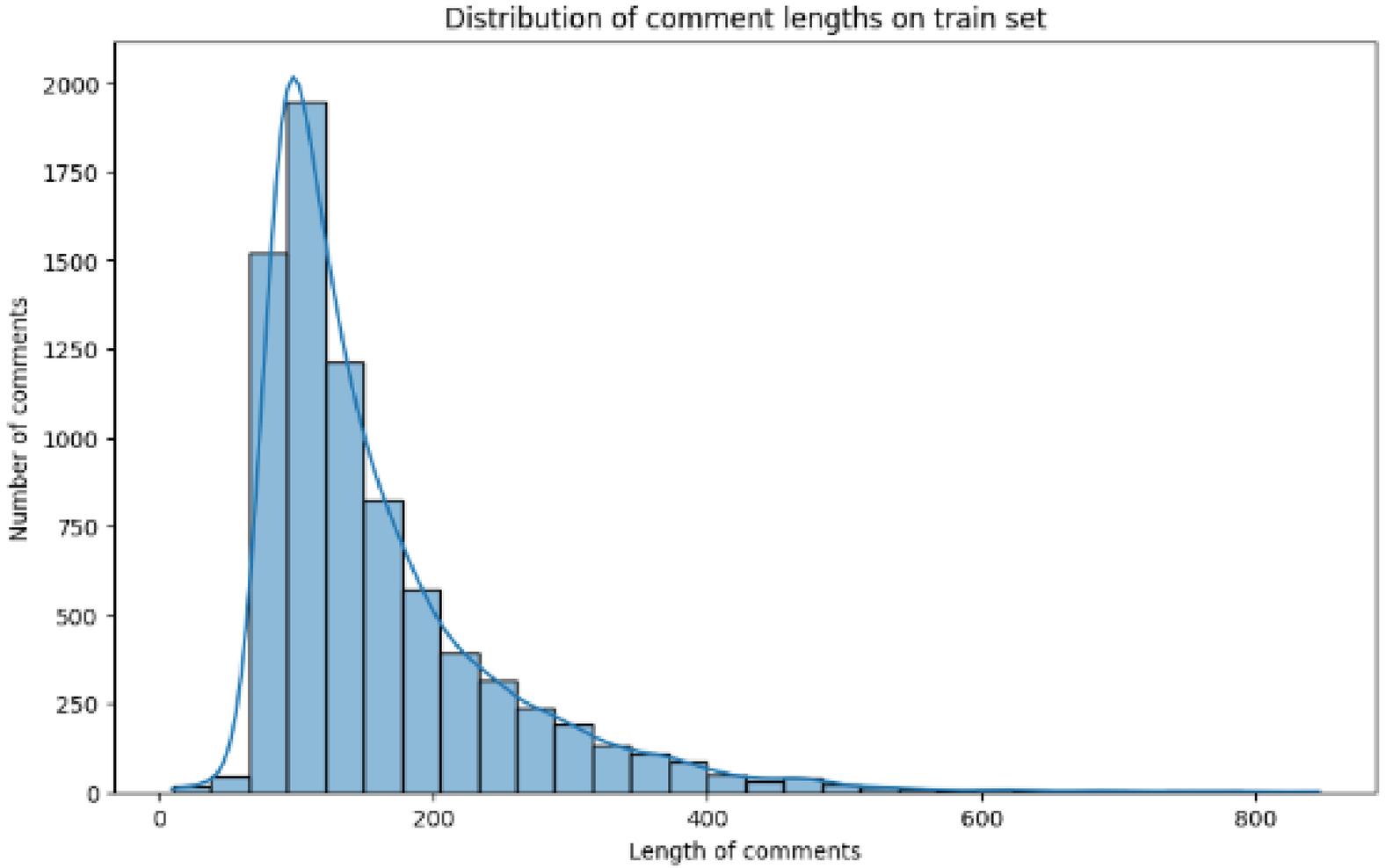


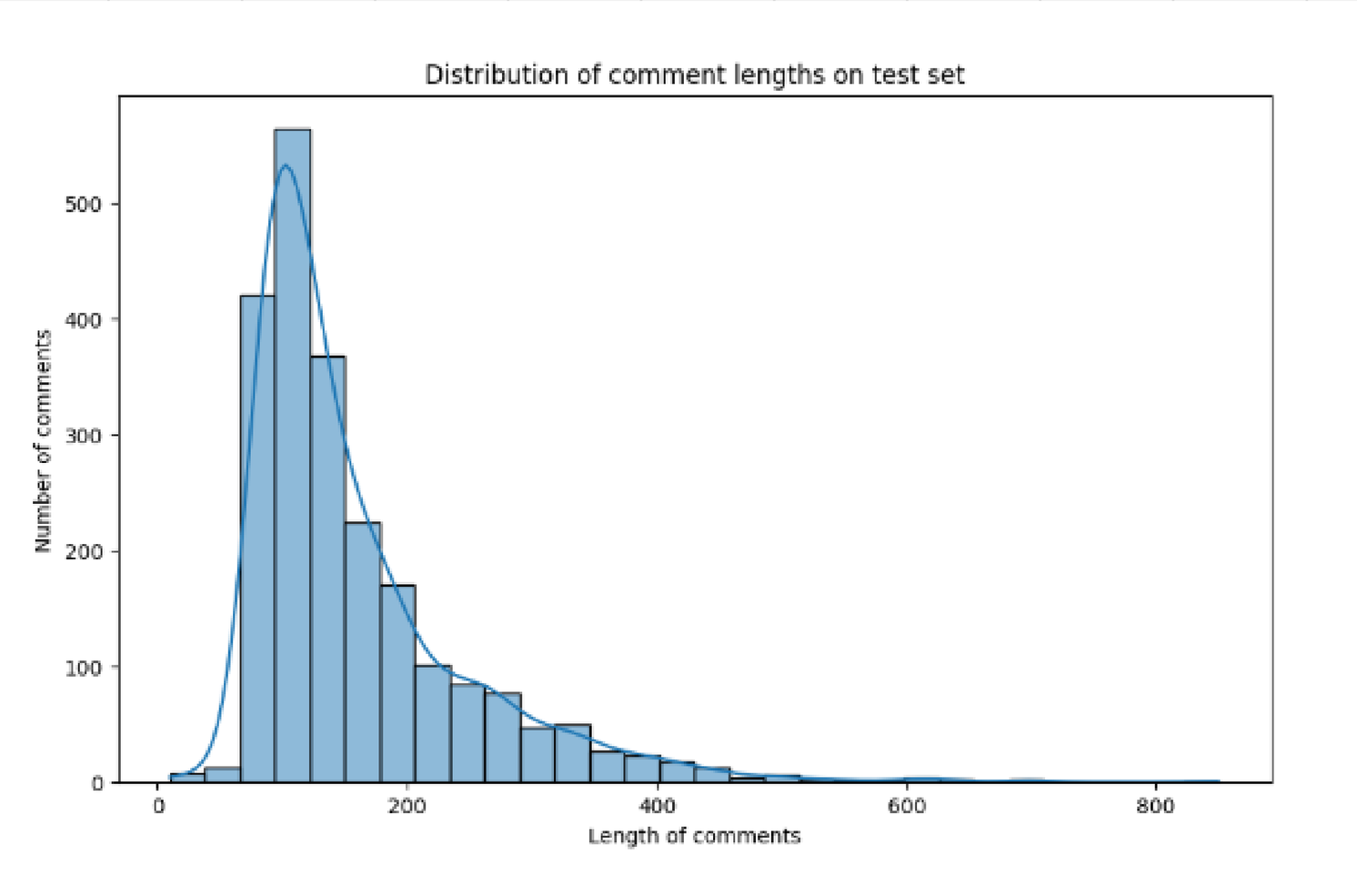


EXPLORATORY DATA ANALYSIS



NHÓM 8


Distribution of comment lengths on train set
Number of comments
Length of comments
0
250
500
750
1000
1250
1500
1750
2000
0
200
400
600
800


Distribution of comment lengths on test set
Number of comments
Length of comments
0
100
200
300
400
500
0
200
400
600
800

PHƯƠNG PHÁP

NHÓM 8

CÁC MÔ HÌNH

SVM

SVM là một mô hình mạnh mẽ có thể xử lý phân loại tuyến tính và phi tuyến tính. Nó hoạt động tốt với các tập dữ liệu có số lượng tính năng lớn.

Softmax regression

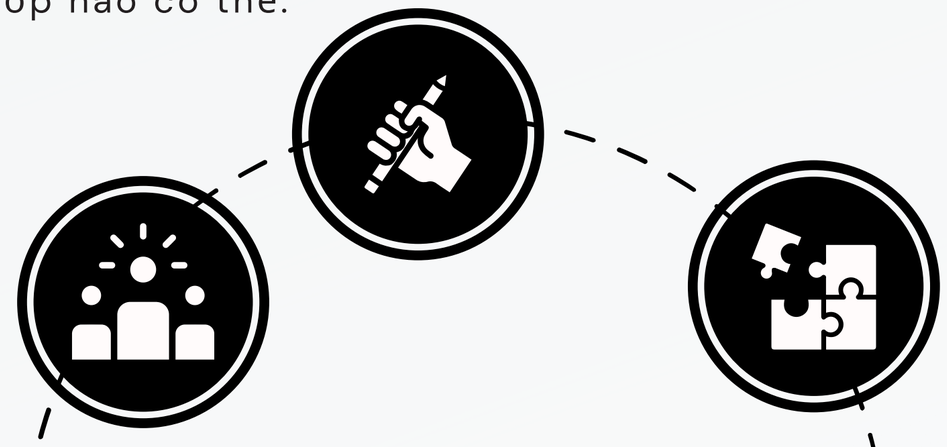
Hồi quy Softmax là một phương pháp trong học máy cho phép phân loại đầu vào thành các lớp rời rạc.Không giống như hồi quy logistic thường được sử dụng, chỉ có thể thực hiện phân loại nhị phân, softmax cho phép phân loại thành bất kỳ số lượng lớp nào có thể.

Random Forests

Random forest là một phương pháp thống kê mô hình hóa băng máy (machine learning statistic) dùng để phục vụ các mục đích phân loại, tính hồi quy và các nhiệm vụ khác băng cách xây dựng nhiều cây quyết định

Decision tree

Cây quyết định (Decision Tree) là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật. Các thuộc tính của đối tượngncó thể thuộc các kiểu dữ liệu khác nhau như Nhị phân (Binary) , Định danh (Nominal), Thứ tự (Ordinal), Số lượng (Quantitative) trong khi đó thuộc tính phân lớp phải có kiểu dữ liệu là Binary hoặc Ordinal.



KẾT QUẢ

NHÓM 8

	GENERAL	SCREEN	CAMERA	FEATURES	BATTERY	PERFORMANCE	STORAGE	DESIGN	PRICE	SER&ACC	OTHERS
Decision Tree	0.643000	0.898000	0.878000	0.795000	0.824000	0.732000	0.986000	0.865000	0.905000	0.841000	0.703000
Softmax	0.745000	0.920000	0.874000	0.834000	0.831000	0.768000	0.990000	0.889000	0.917000	0.876000	0.775000
SVM	0.755000	0.915000	0.878000	0.833000	0.839000	0.775000	0.989000	0.888000	0.914000	0.873000	0.776000
Random Forest	0.725000	0.893000	0.858000	0.817000	0.839000	0.766000	0.989000	0.882000	0.920000	0.871000	0.770000

REFERENCES

- HANG, DO THI THUY, ET AL. “HATE SPEECH DETECTION ON VIETNAMESE SOCIAL MEDIA TEXT USING THE BIDIRECTIONAL-LSTM MODEL.” 2019, P. 4.
- MOHAMMAD ERFAN MOWLAEI, ET AL. “ASPECT-BASED SENTIMENT ANALYSIS USING ADAPTIVE ASPECT-BASED LEXICONS.” VOL. 148, 2020.
- NAZIR, AMBREEN. ISSUES AND CHALLENGES OF ASPECT-BASED SENTIMENT ANALYSIS: A COMPREHENSIVE SURVEY. 2020.
- PHAN, LUONG LUC. “SA2SL: FROM ASPECT-BASED SENTIMENT ANALYSIS TO SOCIAL LISTENING SYSTEM FOR BUSINESS INTELLIGENCE.” 2021, P. 12.
- ZHANG, WENXUAN, ET AL. “ARXIV.” A SURVEY ON ASPECT-BASED SENTIMENT ANALYSIS: TASKS, METHODS, AND CHALLENGES, 2023, P. 21.