TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI
TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG
HaUI | SCHOOL OF INFORMATION & COMMUNICATIONS TECHNOLOGY

# Phân loại cảm xúc người dùng mạng X theo thời gian thực bằng Apache Spark MLlib

**Sinh viên: Phạm Đức Duy**
**CBHD: TS. Nguyễn Mạnh Cường**

## Giới thiệu

Trong bối cảnh dữ liệu mạng xã hội phát triển nhanh chóng, việc nắm bắt kịp thời cảm xúc người dùng là yếu tố quan trọng giúp doanh nghiệp quản lý truyền thông hiệu quả. Đề tài nghiên cứu sử dụng Apache Spark MLlib để xây dựng hệ thống phân loại cảm xúc người dùng mạng X theo thời gian thực, giúp các doanh nghiệp vừa và nhỏ dễ dàng tiếp cận công nghệ phân tích dữ liệu lớn với chi phí hợp lý.

## Phương pháp nghiên cứu

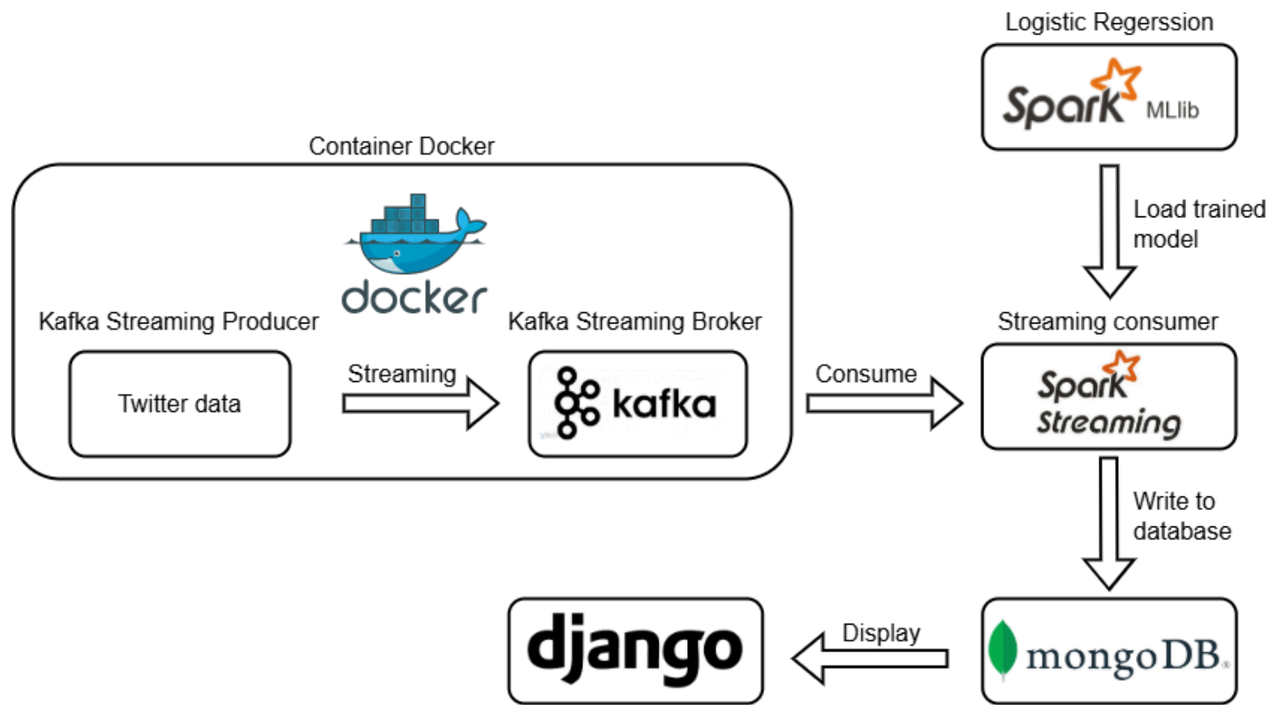
**Thu thập và tiền xử lý dữ liệu**

- Dữ liệu được lấy từ bộ Twitter Sentiment Analysis (74,000 tweet).
- Tiền xử lý dữ liệu: làm sạch văn bản, loại bỏ ký tự đặc biệt, URL, hashtag, mentions.

**Xây dựng mô hình phân loại**

- Áp dụng các mô hình Logistic Regression, Random Forest, Naive Bayes trên Apache Spark MLlib.
- Sử dụng Spark Streaming, Kafka để mô phỏng quá trình xử lý dữ liệu thời gian thực.
- Tối ưu hóa mô hình bằng Cross-validation và Grid Search.

## Kiến trúc hệ thống

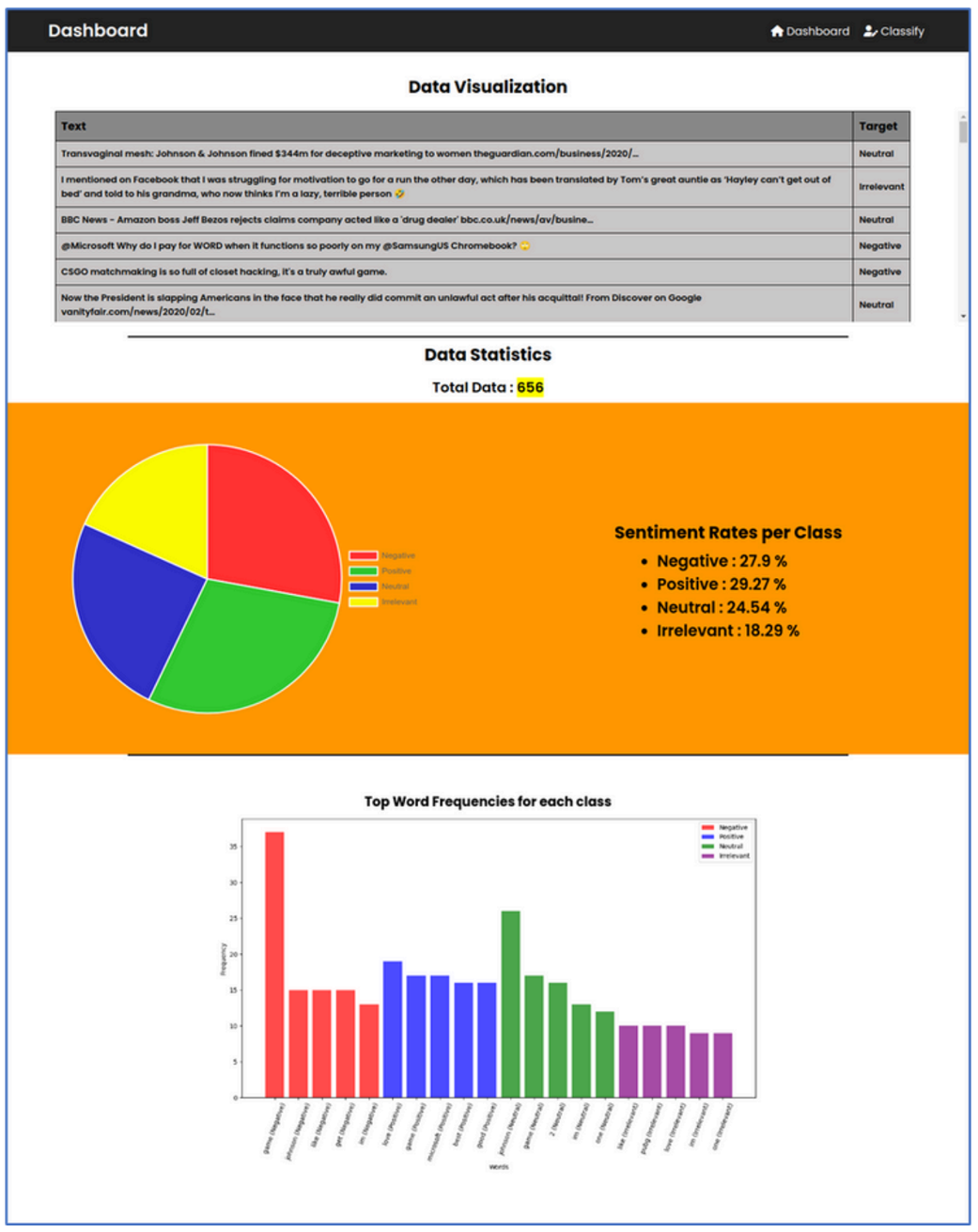



## Kết quả nghiên cứu

Mô hình Logistic Regression cho kết quả tốt nhất với độ chính xác đạt 87.8%, vượt trội so với Random Forest (45.7%) và Naive Bayes (74.0%). Ngoài ra, các chỉ số Precision, Recall và F1 Score của Logistic Regression đều trên 87%, cho thấy khả năng phân loại ổn định và cân bằng giữa các nhãn cảm xúc.

| Metrics / Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Logistic Regression | 0.878 | 0.881513 | 0.878 | 0.877403 |
| Random Forest | 0.457 | 0.662419 | 0.457 | 0.383252 |
| Naive Bayes | 0.740 | 0.747648 | 0.740 | 0.736290 |

## Giao diện hệ thống




## Hướng phát triển

Định hướng phát triển:

- Nâng cấp mô hình với các kỹ thuật học sâu như BERT, Transformer.
- Hỗ trợ đa ngôn ngữ, phân tích emoji, từ lóng.
- Cải thiện giao diện người dùng, hỗ trợ phân tích biểu đồ nâng cao.