Fraud detection using machine learning

Nhóm 77 …

Slide 2: Các điểm thảo luận

* Agenda
* Tổng quan dữ liệu
* Phân tích đơn biến
* Phân tích theo loại email
* Phân tính theo nội dung email
* Một số câu hỏi quan trọng
* Kết luận

Slide 3: Agenda  
Chúng ta sẽ thảo luận về phân tích dữ liệu liên quan đến việc nhận biết lừa đảo qua email.

Chúng ta sẽ tiến hành phân tích tập dữ liệu theo các cách sau:

* **Phân tích đơn biến** (Univariate Analysis): Xem xét từng đặc trưng riêng lẻ của email để tìm hiểu sự khác biệt giữa email lừa đảo và email hợp lệ.
* **Phân tích theo loại email** (Email Type Analysis): Phân tích sự khác nhau giữa email lừa đảo, email spam và email hợp lệ.
* **Phân tích theo nội dung email** (Content-Based Analysis): Đánh giá các từ khóa, câu chữ, liên kết trong email để nhận diện các dấu hiệu lừa đảo.

Thông qua phân tích này, chúng ta sẽ tìm ra những yếu tố chính giúp nhận diện email lừa đảo, hỗ trợ phát triển hệ thống cảnh báo và ngăn chặn các hành vi gian lận qua email.

Slide 4: Data Summary

Dữ liệu này chứa các cột quan trọng để phân loại email thành **Spam** hoặc **Ham**. Một số cột chính bao gồm:

* **Category**: Nhãn phân loại email, gồm hai giá trị chính:
  + Ham: Email hợp lệ, không phải thư rác.
  + Spam: Email rác, có thể chứa nội dung lừa đảo hoặc quảng cáo không mong muốn.
* **Message**: Nội dung của email, bao gồm văn bản mà người gửi đã viết. Đây là dữ liệu đầu vào để mô hình phân loại tin nhắn có phải là spam hay không.

A diagram of a data flow

Description automatically generated

Slide 5:

Trong khi thực hiện phân tích đơn biến của tập dữ liệu Nhận biết lừa đảo Email đã cho, chúng tôi đã trả lời các câu hỏi sau:

 Email có chủ đề nào thường xuất hiện nhiều nhất trong các email lừa đảo?

 Loại từ khóa nào phổ biến nhất trong các email lừa đảo?

 Những mẫu email nào thường bị đánh dấu là spam nhất?

Slide 6:

* Phân phối số từ trong email spam vs hamA blue bar graph with white text

  Description automatically generated

Nhận xet

 Số lượng email **Ham** (email hợp lệ) chiếm đa số, khoảng **5000 email**.

 Số lượng email **Spam** (email rác) ít hơn rất nhiều, khoảng **gần 1000 email**.

* Tần suất các từ khóa phổ biến trong email spamA graph with blue bars

  Description automatically generated

Nhận xét:

*  Dữ liệu tin nhắn có xu hướng **ngắn gọn**, phần lớn tập trung trong khoảng **dưới 200 ký tự**.
*  Các tin nhắn dài có thể là **tin spam** hoặc tin nhắn chứa nhiều nội dung quan trọng.

Slide 7: Phân tính theo nội dung email

A close up of words

Description automatically generated

 **Từ khóa nổi bật**:

* Các từ lớn nhất như **"go"**, **"know"**, **"got"**, **"ok"**, **"one"**, **"free"**, **"need"**, **"love"** là những từ xuất hiện nhiều nhất trong tập dữ liệu email.
* Những từ như **"free"**, **"win"**, **"prize"**, **"urgent"** có thể là dấu hiệu của email spam, do chúng thường xuất hiện trong các email lừa đảo.

 **Xu hướng nội dung**:

* Một số từ có vẻ thiên về hội thoại thông thường như **"good"**, **"home"**, **"hey"**, **"friend"**, **"love"** → có thể là từ email hợp lệ (ham).
* Các từ mang tính cấp bách như **"urgent"**, **"prize"**, **"free entry"**, **"150p"** (giá tiền) thường liên quan đến email spam.

A graph with purple lines

Description automatically generated

**Biểu đồ phân bố độ dài tin nhắn:**

1. **Phân bố lệch phải (Right-Skewed Distribution)**:
   * Hầu hết các email có độ dài ngắn, tập trung nhiều ở khoảng **0 - 200 ký tự**.
   * Một số ít email có độ dài vượt **200 ký tự**, và rất hiếm email có độ dài trên **600 - 800 ký tự**.
2. **Đỉnh cao nhất ở khoảng 0 - 50 ký tự**:
   * Điều này cho thấy phần lớn email trong tập dữ liệu là **ngắn gọn**.
   * Những email này có thể là hội thoại thông thường (ham) hoặc tin nhắn quảng cáo ngắn.
3. **Hai đỉnh phụ quanh 100 - 150 ký tự và 180 - 200 ký tự**:
   * Đây có thể là nhóm email có nội dung dài hơn, có khả năng chứa nội dung quảng cáo hoặc thông báo quan trọng.
4. **Có một số email rất dài (outliers)**:
   * Các email có độ dài trên **400 ký tự** rất hiếm, có thể là email chi tiết hoặc chứa nội dung spam có nhiều thông tin quảng cáo.

A graph of a number of lines

Description automatically generated with medium confidence

**biểu đồ hộp (Boxplot) so sánh độ dài tin nhắn giữa Ham & Spam:**

1. **Email Spam có độ dài trung bình lớn hơn Email Ham**:
   * Hộp của **Email Spam (1)** dài hơn và nằm ở mức cao hơn so với Email Ham (0).
   * Điều này cho thấy **email spam thường có nội dung dài hơn** so với email thông thường.
2. **Email Ham có sự phân tán rộng hơn**:
   * Độ dài tin nhắn của email ham rất đa dạng, trải dài từ **dưới 50 ký tự đến hơn 200 ký tự**.
   * Có nhiều **outliers (điểm ngoại lai)** trong nhóm email ham, có một số tin nhắn rất dài **(trên 800 ký tự)**.
3. **Email Spam có phân bố tập trung hơn**:
   * Phần lớn email spam có độ dài từ **100 đến 200 ký tự**.
   * Có một số ít email spam có độ dài vượt 200 ký tự, nhưng không có nhiều điểm ngoại lai như email ham.
4. **Khả năng nhận diện spam qua độ dài tin nhắn**:
   * Nếu chỉ dựa vào độ dài tin nhắn, có thể thấy rằng **email spam có xu hướng dài hơn**.
   * Tuy nhiên, vẫn có sự chồng lấn giữa hai nhóm, vì vậy cần thêm các yếu tố khác như **từ khóa, số lượng ký tự đặc biệt, hoặc tỷ lệ từ khóa spam** để phân loại chính xác hơn.

Slide cuối: Kết luận  
 **Phát hiện email lừa đảo đóng vai trò quan trọng** trong việc bảo vệ người dùng khỏi các cuộc tấn công mạng, đặc biệt là phishing và spam có mục đích xấu.

 **Từ khóa phổ biến trong email** cho thấy rằng các email spam thường chứa các cụm từ như "free", "urgent", "prize", "win", v.v., trong khi email hợp lệ (ham) có nội dung tự nhiên hơn.

 **Độ dài tin nhắn là một yếu tố phân biệt**: Email lừa đảo (spam) thường có độ dài ngắn hơn nhưng chứa nhiều từ ngữ quảng cáo hoặc mang tính thúc giục hành động ngay lập tức.

 **Tỷ lệ hủy đơn của email spam cao**: Người dùng thường không phản hồi hoặc tương tác với email lừa đảo, trong khi email hợp lệ thường có tỉ lệ phản hồi cao hơn.

 **Các yếu tố quan trọng trong phân loại email**: Từ khóa, độ dài nội dung, số lượng liên kết trong email và tỷ lệ xuất hiện các từ khóa spam đều là những đặc điểm quan trọng trong việc phân loại email.

 **Hiệu suất mô hình phân loại email**: Mô hình học máy sử dụng thuật toán như Naïve Bayes, Random Forest hoặc LSTM có thể đạt độ chính xác cao trong việc phát hiện email lừa đảo.

 **Ứng dụng thực tế**: Kết quả nghiên cứu này có thể được sử dụng để cải thiện bộ lọc spam, giúp hệ thống email tự động nhận diện và chặn các email độc hại.