


THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút):
<https://youtu.be/leGZonpbeVI>
- Link slides (dạng .pdf đặt trên Github):
<https://github.com/huynhhdh18/CS2205.MAR2024/blob/main/Huy%CC%80nh%20%C4%90inh%20Ho%CC%82%CC%80ng%20-%20xCS2205.DeCuong.FinalReport.Template.Poster.pdf>
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

<ul style="list-style-type: none"> Họ và Tên: Đinh Hồng Huỳnh MSSV: 230201044 	<ul style="list-style-type: none"> Lớp: CS2205.MAR2024 Tự đánh giá (điểm tổng kết môn): 8.5/10 Số buổi vắng: 2 Số câu hỏi QT cá nhân: 3 Link Github: https://github.com/huynhhdh18/CS2205.MAR2024
---	---

ĐỀ CƯƠNG NGHIÊN CỨU

<p>TÊN ĐỀ TÀI (IN HOA)</p> <p>PHÁT HIỆN GIAN LẬN CV SỬ DỤNG KỸ THUẬT XỬ LÝ NGÔN NGỮ TỰ NHIÊN (NLP)</p>
<p>TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)</p> <p>CV FRAUD DETECTION USING NATURAL LANGUAGE PROCESSING (NLP) TECHNIQUES</p>
<p>TÓM TẮT <i>(Tối đa 400 từ)</i></p> <p>Hiện nay, gian lận trong CV là một vấn đề nhức nhối trong thị trường tuyển dụng. Việc ứng viên gian lận thông tin trong CV có thể gây ra nhiều hậu quả nghiêm trọng cho cả nhà tuyển dụng và ứng viên. Do đó, bài nghiên cứu này đề xuất một phương pháp mới để phát hiện gian lận trong CV sử dụng kỹ thuật Xử lý ngôn ngữ tự nhiên (NLP). Phương pháp này sẽ sử dụng các mô hình học máy được đào tạo trên một tập dữ liệu lớn gồm CV thật và giả để xác định các mẫu ngôn ngữ liên quan đến gian lận.</p>
<p>GIỚI THIỆU <i>(Tối đa 1 trang A4)</i></p> <p>1. Bối cảnh và tầm quan trọng của vấn đề</p> <p>Gian lận CV là hành vi khai gian thông tin trong CV của ứng viên nhằm mục đích đánh lừa nhà tuyển dụng. Việc gian lận CV có thể gây ra nhiều hậu quả nghiêm trọng cho cả nhà tuyển dụng và ứng viên. Đối với nhà tuyển dụng, việc tuyển dụng nhầm người gian lận có thể dẫn đến tổn thất về tài chính, năng suất và uy tín. Đối với ứng viên, việc bị phát hiện gian lận có thể ảnh hưởng đến danh tiếng và cơ hội nghề nghiệp trong tương lai.</p> <p>2. Các phương pháp phát hiện gian lận CV hiện tại</p> <p>Hiện nay, có một số phương pháp được sử dụng để phát hiện gian lận CV, bao gồm:</p> <ul style="list-style-type: none">- Kiểm tra thủ công: Phương pháp này người tuyển dụng sẽ đọc và đánh giá CV để tìm kiếm các dấu hiệu gian lận, Tuy nhiên, phương pháp này

tốn thời gian, tốn kém về tài chính và có thể không chính xác

- **Sử dụng phần mềm so sánh:** Các phần mềm so sánh CV của ứng viên với các CV khác trong cơ sở dữ liệu để tìm kiếm các trường hợp trùng lặp hoặc đáng ngờ. Tuy nhiên, các phần mềm này có thể bị đánh lừa bởi các CV được thiết kế một cách tinh vi.
- **Sử dụng các kỹ thuật học máy:** Các kỹ thuật học máy có thể được sử dụng để phân tích CV và xác định các mẫu ngôn ngữ liên quan đến gian lận. Tuy nhiên, hiệu quả của các phương pháp này phụ thuộc vào chất lượng của dữ liệu đào tạo.

3. Đóng góp của nghiên cứu

Nghiên cứu này đề xuất một phương pháp mới để phát hiện gian lận CV sử dụng kỹ thuật NLP. Phương pháp này sử dụng các mô hình máy học được đào tạo trên một tập dữ liệu lớn gồm CV thật giả để xác định các mẫu ngôn ngữ liên quan đến gian lận. Ưu điểm của phương pháp này là:

- **Tự động hóa:** Phương pháp này có thể tự động hóa quy trình phát hiện gian lận CV, giúp tiết kiệm thời gian và chi phí
- **Độ chính xác cao:** Phương pháp này có thể đạt được độ chính xác cao hơn các phương pháp truyền thống.
- **Khả năng thích ứng:** Phương pháp này có thể được điều chỉnh để phù hợp với các yêu cầu cụ thể của từng tổ chức.

MỤC TIÊU

(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)

Mục tiêu của nghiên cứu này là:

1. Phát triển một mô hình học máy để tự động phát hiện gian lận CV sử dụng kỹ thuật NLP
2. Đánh giá hiệu quả của mô hình học máy trên một tập dữ liệu lớn gồm CV thật và giả
3. Phân tích các yếu tố ảnh hưởng đến hiệu quả của mô hình học máy.

NỘI DUNG VÀ PHƯƠNG PHÁP

(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)

1. Thu thập dữ liệu

Thu thập một tập dữ liệu lớn gồm CV thật và giả. CV thật có thể được thu thập từ các nguồn như trang web tuyển dụng, cơ sở dữ liệu của công ty hoặc các tổ chức phi lợi nhuận. CV giả có thể được tạo ra bằng cách tổng hợp hoặc sửa đổi CV thật.

2. Tiền xử lý dữ liệu

2.1. Loại bỏ nhiễu

Loại bỏ nhiễu trong dữ liệu CV là bước đầu tiên và quan trọng trong quá trình tiền xử lý dữ liệu.

Nhiều trong dữ liệu CV có thể bao gồm:

- Ký tự đặc biệt: Loại bỏ các ký tự đặc biệt như dấu chấm phẩy, dấu hai chấm, dấu ngoặc đơn, dấu ngoặc kép, v.v.
- Ký tự trắng: Loại bỏ các ký tự trắng thừa như khoảng trắng, tab, v.v.
- Chữ hoa: Chuyển đổi các chữ thành chữ thường.
- Loại bỏ các thông tin khác như: HTML tag, URL, Email, Số điện thoại, v.v.

Việc loại bỏ nhiễu giúp dữ liệu CV trở nên sạch hơn và đồng nhất hơn, từ đó giúp cải thiện hiệu quả của mô hình

2.2. Chuyển đổi văn bản thành dạng số

Văn bản trong CV cần được chuyển đổi thành dạng số để mô hình học máy có thể hiểu và xử lý được. Có hai kỹ thuật chính được sử dụng để chuyển đổi văn bản thành dạng số:

- Bag-of-words encoding: Kỹ thuật này biểu diễn mỗi CV dưới dạng một Vector, trong đó mỗi phần tử trong vector biểu thị tần suất xuất hiện của một từ khóa cụ thể trong CV
- Word embedding: Kỹ thuật này biểu diễn mỗi từ khóa dưới dạng một Vector, trong đó mỗi phần tử trong Vector biểu diễn ý nghĩa ngữ

nghĩa của từ khóa

2.3. Phân tách tính năng

Phân tách tính năng là quá trình xác định các tính năng từ dữ liệu CV có thể được sử dụng để đoán gian lận. Các tính năng này có thể bao gồm:

- Độ dài CV: Chiều dài trung bình của CV.
- Mật độ từ khóa: Tần suất xuất hiện của các từ khóa liên quan đến gian lận trong CV.
- Sự không nhất quán: Mức độ mâu thuẫn giữa các thông tin trong CV
- Kinh nghiệm làm việc: Số năm kinh nghiệm làm việc của ứng viên
- Trình độ học vấn: Trình độ học vấn cao nhất của ứng viên
- Kỹ năng: Các kỹ năng được liệt kê trong CV của ứng viên

3. Đào tạo mô hình

Sau khi tiền xử lý dữ liệu, ta có thể tiến hành đào tạo mô hình để dự đoán gian lận CV. Có hai loại mô hình học máy chính được sử dụng:

- Học có giám sát(Supervised learning):

Trong học có giám sát, mô hình sẽ được cung cấp một tập dữ liệu gồm các CV được dán nhãn là thật hoặc giả. Mô hình học máy sẽ học cách phân biệt giữa các CV thật và CV giả dựa trên các tính năng được trích xuất từ dữ liệu

- Học tăng cường(Reinforcement learning):

Trong học tăng cường, mô hình học máy tương tác với môi trường và nhận phần thưởng hoặc hình phạt dựa trên hành động của nó. Mục tiêu của mô hình học máy này là học các thực hiện các hành động tối đa hóa phần thưởng.

4. Đánh giá mô hình

Sau khi đào tạo mô hình học máy, cần đánh giá hiệu quả của mô hình trên một tập dữ liệu kiểm tra riêng biệt. Các chỉ số đánh giá hiệu quả được sử dụng bao gồm:

- Độ chính xác: Tỷ lệ phần trăm các trường hợp được dự đoán đúng
- Độ nhạy: Tỷ lệ phần trăm các trường hợp gian lận được dự đoán đúng
- Độ đặc hiệu: Tỷ lệ phần trăm các trường hợp không gian lận được dự đoán đúng
- Diện tích dưới đường cong ROC (AUC): Đo lường khả năng phân biệt giữa các trường hợp thật và giả của mô hình

Ngoài các chỉ số định lượng, cũng cần đánh giá hiệu quả mô hình một cách định tính bằng cách phân tích các trường hợp được dự đoán sai.

Việc phân tích này có thể giúp xác định được những hạn chế của mô hình và đưa ra các giải pháp cải thiện

5. Phân tích kết quả

- Xác định các tính năng quan trọng: Xác định những tính năng đóng góp nhiều nhất vào khả năng dự đoán của mô hình
- Phân tích các trường hợp được dự đoán sai: Phân tích các trường hợp được dự đoán sai để hiểu rõ hơn về những hạn chế của mô hình
- So sánh hiệu quả của mô hình với các phương pháp khác: so sánh hiệu quả của mô hình học máy với các phương pháp phát hiện gian lận CV truyền thống

6. Đề xuất giải pháp

- Thu thập thêm dữ liệu: thu thập thêm dữ liệu CV để đào tạo mô hình, đặc biệt là dữ liệu về các trường hợp gian lận
- Sử dụng các kỹ thuật học máy tiên tiến: Sử dụng các kỹ thuật học máy tiên tiến hơn như học sâu (Deep learning) để cải thiện hiệu quả của mô hình
- Kết hợp các phương pháp khác: Kết hợp mô hình học máy với các phương pháp phát hiện gian lận CV truyền thống để tăng cường hiệu quả
- Điều chỉnh mô hình sao cho phù hợp với các trường hợp cụ thể: Điều chỉnh mô hình sao cho phù hợp với các yêu cầu cụ thể của từng tổ chức hoặc theo ngành nghề

KẾT QUẢ MONG ĐỢI

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)

- Phát triển một mô hình học máy có thể tự động phát hiện gian lận CV với độ chính xác cao
- Xác định các yếu tố ảnh hưởng đến hiệu quả của mô hình học máy
- Đề xuất các giải pháp để cải thiện hiệu quả của mô hình học máy

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

- [1]. Baraneetharan, E. (2022). Detection of fake job advertisements using machine learning algorithms. *Journal of Artificial Intelligence and Capsule Networks*, 4(3), 200-210.
- [2]. Zhang, H., Wang, M., Wang, Y., Li, Y., Gu, D., & Zhu, Y. (2023, October). ORFPPrediction: Machine Learning Based Online Recruitment Fraud Probability Prediction. In *2023 International Conference on the Cognitive Computing and Complex Data (ICCD)* (pp. 139-144). IEEE.
- [3]. Bhatia, T., & Meena, J. (2022, December). Detection of Fake Online Recruitment Using Machine Learning Techniques. In *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)* (pp. 300-304). IEEE.
- [4]. Mahbub, S., Pardede, E., & Kayes, A. S. M. (2022). Online recruitment fraud detection: A study on contextual features in Australian job industries. *IEEE Access*, 10, 82776-82787.