PHÁT HIỆN GIAN LẬN CV SỬ DỤNG KỸ THUẬT XỬ LÝ NGÔN NGỮ TỰ NHIÊN (NLP)

Đinh Hồng Huỳnh - 230201044

Tóm tắt

- Lớp: CS2205.MAR2024
- Link Github: https://github.com/huynhdh18/CS2205.MAR2024/
- Link YouTube video: https://youtu.be/leGZonpbeVI
- Ảnh + Họ và Tên: Đinh Hồng Huỳnh

Giới thiệu

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.

Hiện nay, có một số phương pháp được sử dụng để phát hiện gian lận CV, bao gồm:

- **Kiểm tra thủ công:** Phương pháp này người tuyển dụng sẽ đọc và đánh giá CV để tìm kiếm các dấu hiệu gian lận
- **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
- **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.

Mục tiêu

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

1. Thu thập dữ liệu

   Thu thập một tập dữ liệu lớn gồm CV thật và giả. CV thật có thể được thu tập 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

   - Ký tự đặc biệt
   - Ký tự trắng
   - Chuyển đổi chữ hoa 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

Nội dung và Phương pháp

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.

- **Bag-of-words encoding**
- **Word embedding**

2.3 Phân tách tính năng

- 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

Nội dung và Phương pháp

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)
- Học tăng cường(Reinforcement learning)

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.

- 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

Nội dung và Phương pháp

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
    - Điều chỉnh mô hình sao cho phù hợp với các trường hợp cụ thể

Kết quả dự kiến

- 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

- [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.