

ĐẠI HỌC HUẾ

**KHOA KỸ THUẬT VÀ CÔNG NGHỆ**

🙠🙟🕮🙝🙢

**A picture containing symbol, logo, red, emblem

Description automatically generated**

**BÁO CÁO ĐỒ ÁN**

**Học kỳ I, năm học 2023 - 2024**

**Học phần:**

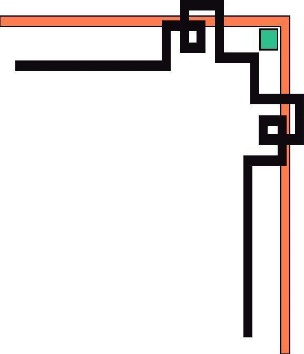
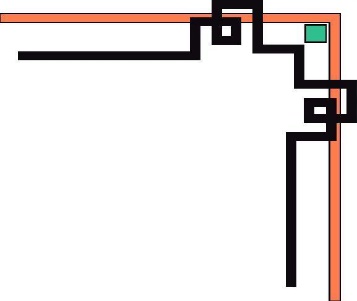
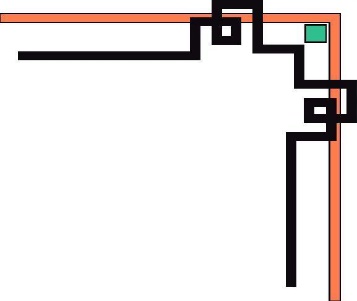
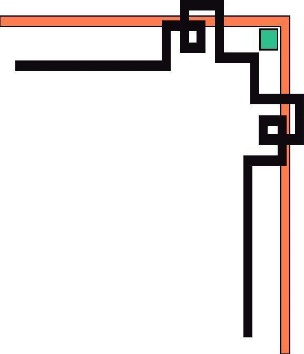
**Khai phá dữ liệu lớn**

**Đề tài: Ứng dụng phương pháp BoW kết hợp SVM, CNN cho Văn bản để dự đoán ngành nghề dựa trên mô tả của công ty**

|  |
| --- |
| **Số phách**  *(Do hội đồng chấm thi ghi)* |

**Thừa Thiên Huế, tháng 01 năm 2024**

ĐẠI HỌC HUẾ



**KHOA KỸ THUẬT VÀ CÔNG NGHỆ**

🙠🙟🕮🙝🙢

**A picture containing symbol, logo, red, emblem

Description automatically generated**

**(Bìa phụ 2)**

**BÁO CÁO ĐỒ ÁN**

**Học kỳ I, năm học 2023 - 2024**

**Học phần:**

**Khai phá dữ liệu lớn**

**Đề tài: Ứng dụng phương pháp BoW kết hợp SVM, CNN cho văn bản để dự đoán ngành nghề dựa trên mô tả của công ty**

**Giảng viên hướng dẫn: Hồ Quốc Dũng**

**Lớp: KHDL&TTNT – K1**

**Sinh viên thực hiện: Nguyễn Tiến Thịnh**

|  |
| --- |
| **Số phách**  *(Do hội đồng chấm thi ghi)* |

**Thừa Thiên Huế, tháng 01 năm 2024**

MỤC LỤC

[1. Ý TƯỞNG 3](#_Toc156292405)

[1.1. Giới thiệu tập dữ liệu 3](#_Toc156292406)

[1.2. Ý tưởng với dữ liệu 4](#_Toc156292407)

[2. CHUẨN BỊ DỮ LIỆU 5](#_Toc156292408)

[3. ĐỀ XUẤT MÔ HÌNH NGHIÊN CỨU 8](#_Toc156292409)

[4. ĐỀ XUẤT CÁC MÔ HÌNH HỌC MÁY 9](#_Toc156292410)

[4.1. Sử dụng phương pháp Bag-of-Words kết hợp SVM 9](#_Toc156292411)

[4.2. Sử dụng CNN cho Văn bản 9](#_Toc156292412)

[5. THỰC THI MÔ HÌNH 10](#_Toc156292413)

[5.1. Mô hình BoW kết hợp SVM 10](#_Toc156292414)

[5.1.1. Xây dựng mô hình 10](#_Toc156292415)

[5.1.2. Kết quả của mô hình 11](#_Toc156292416)

[5.1.3. Kết quả dự đoán của mô hình 13](#_Toc156292417)

[5.2. Mô hình CNN cho Văn bản 14](#_Toc156292418)

[5.2.1. Xây dựng mô hình 14](#_Toc156292419)

[5.2.2. Kết quả của mô hình 14](#_Toc156292420)

[5.2.3. Kết quả dự đoán của mô hình 15](#_Toc156292421)

[6. TỔNG KẾT 17](#_Toc156292422)

[TÀI LIỆU THAM KHẢO 18](#_Toc156292423)

DANH MỤC HÌNH ẢNH

[Hình 1. Dữ liệu đã crawl 3](#_Toc156292424)

[Hình 2. Dữ liệu dùng cho mô hình 5](#_Toc156292425)

[Hình 3. Các ngành nghề có số lượng lớn hơn 100 6](#_Toc156292426)

[Hình 4. Dữ liệu đã được làm sạch 7](#_Toc156292427)

[Hình 5. Kết quả của mô hình BoW - SVM 11](#_Toc156292428)

[Hình 6. Confusion Matrix của mô hình BoW-SVM 12](#_Toc156292429)

[Hình 7. Confusion Matrix của mô hình CNN 15](#_Toc156292430)

# Ý TƯỞNG

## 1.1. Giới thiệu tập dữ liệu

Sau khi crawl dữ liệu trên trang Glassdoor[1], ta đã có dữ liệu của tất cả công ty trên trang web. Dữ liệu được minh họa như sau:



Hình 1. Dữ liệu đã crawl

Dữ liệu trên có tất cả thông tin của các công ty trên trang Glassdoor, bao gồm 8485 công ty.

Trong đó:

|  |  |
| --- | --- |
| **Tên** | **Giá trị** |
| Name | Tên của công ty |
| Rate | Đánh giá (thang điểm 5) |
| Image\_url | Đường dẫn đến link ảnh logo của công ty |
| Glb\_company\_size | Số lượng nhân viên |
| Industry | Ngành nghề của công ty |
| Review | Số lượt đánh giá |
| Salary | Trung bình lương |
| Jobs | Số lượng công việc |
| Description | Mô tả về công ty |
| Url\_company | Đường dẫn đến trang Glassdoor của công ty |

## 1.2. Ý tưởng với dữ liệu

Trong dự án này, chúng ta sẽ thảo luận về một loạt ý tưởng sáng tạo và đa dạng về cách sử dụng phương pháp học máy để khám phá thông tin từ tập dữ liệu hiện có. Một số ứng dụng cụ thể bao gồm:

* Dự đoán Đánh giá Công ty: Phát triển một mô hình học máy để tự động dự đoán đánh giá của các công ty, sử dụng thông tin như ngành nghề, kích thước, lĩnh vực hoạt động, mức lương, và số lượng công việc.
* Phân loại Ngành Nghề Công ty: Áp dụng mô hình phân loại để xác định ngành nghề chính của một công ty dựa trên đặc điểm mô tả, hình ảnh, và tên của công ty.
* Dự đoán Mức lương Dự kiến: Xây dựng một mô hình để dự đoán mức lương dự kiến dựa trên các biến như đánh giá công ty, ngành nghề, kích thước công ty, và các yếu tố khác.
* Phân loại Kích thước Công ty: Sử dụng học máy để phân loại kích thước của công ty dựa trên thông tin về số lượng nhân sự, số lượng công việc và doanh thu.
* So sánh Văn bản Mô tả Công ty: Xây dựng một mô hình để so sánh mô tả văn bản của các công ty và đánh giá sự tương đồng giữa chúng.
* Phân loại Công ty Tốt với Xấu: Tạo mô hình để phân loại các công ty thành nhóm tốt và nhóm xấu dựa trên các đặc điểm như đánh giá, mức lương, và ngành nghề.
* Dự đoán Số Lượng Việc Làm: Sử dụng mô hình để dự đoán số lượng việc làm trong tương lai, dựa trên các yếu tố như đánh giá và kích thước công ty.
* Phân loại Công ty Theo Chiều sâu Ngành Nghề: Áp dụng mô hình để phân loại công ty vào các phân khúc chiều sâu của ngành nghề, sử dụng mô tả ngành nghề để hiểu rõ hơn về bản chất của chúng.

**Lựa chọn Chiến lược:**

Trong hướng tiếp cận dự án này, em sẽ chú trọng vào ý tưởng "Phân loại Ngành nghề Công ty dựa trên Đặc điểm Mô tả". Điều này giúp ta khai thác sâu vào mô tả của công ty để hiểu rõ hơn về ngành nghề chính của họ và cung cấp thông tin giá trị để hỗ trợ quyết định và phân loại.

# CHUẨN BỊ DỮ LIỆU

Trong quá trình xử lý dữ liệu, việc chọn lựa các đặc trưng quan trọng để huấn luyện mô hình là một bước quan trọng và có ảnh hưởng lớn đến hiệu suất của mô hình. Trong tập dữ liệu mà chúng ta đang xử lý, mục tiêu là dự đoán ngành nghề của các công ty dựa trên mô tả của chúng. Để thực hiện điều này, chúng ta đã quyết định tận dụng hai cột quan trọng nhất trong tập dữ liệu đó là "Industry" và "Description".

Cột "Industry" cho chúng ta thông tin về ngành nghề chính của mỗi công ty. Đây là một yếu tố quan trọng, vì ngành nghề sẽ ảnh hưởng lớn đến nhiều khía cạnh của doanh nghiệp, từ mô hình kinh doanh đến phong cách lãnh đạo. Việc tích hợp thông tin từ cột này vào mô hình giúp mô phỏng mối quan hệ giữa ngành nghề và mô tả công ty, giúp mô hình học được các đặc trưng quan trọng trong việc phân loại ngành nghề.

Cột "Description" cung cấp mô tả chi tiết về hoạt động và đặc điểm của mỗi công ty. Mô tả này có thể chứa những từ ngữ, cụm từ hay đặc điểm nổi bật mà mô hình có thể sử dụng để tìm ra các mẫu và mối quan hệ có thể giúp dự đoán ngành nghề của công ty. Việc sử dụng thông tin từ cột này giúp mô hình nắm bắt được sự đa dạng và sự phong phú trong mô tả của các công ty, từ đó làm tăng khả năng tổng quát hóa của mô hình.

.A screenshot of a website

Description automatically generated

Hình 2. Dữ liệu dùng cho mô hình

Quá trình làm sạch dữ liệu đóng vai trò quan trọng trong việc chuẩn bị dữ liệu cho việc huấn luyện mô hình. Để đảm bảo dữ liệu được xử lý một cách chính xác và hiệu quả, chúng ta đã thực hiện một số bước quan trọng để loại bỏ nhiễu và đảm bảo chất lượng dữ liệu sau cùng.

Đầu tiên, chúng ta đã xóa bỏ các dòng dữ liệu chứa giá trị NaN. Điều này là bước quan trọng để loại bỏ những giá trị thiếu thông tin và đảm bảo rằng dữ liệu không bị nhiễm bẩn bởi các giá trị không hợp lý. Sau khi thực hiện bước này, số lượng dòng dữ liệu đã giảm từ 8485 xuống còn 8475, làm tăng độ chính xác và tin cậy của tập dữ liệu.

A close-up of a computer code

Description automatically generated

Tiếp theo, để giảm thiểu ảnh hưởng của các ngành nghề có lượng dữ liệu ít, chúng ta đã quyết định chỉ giữ lại những ngành nghề có số lượng công ty lớn hơn 100. Việc này giúp đảm bảo rằng mô hình sẽ được huấn luyện trên những dữ liệu đại diện và tránh được ảnh hưởng quá mức từ các ngành nghề có dữ liệu quá ít.

A graph with blue and white text

Description automatically generated

Hình 3. Các ngành nghề có số lượng lớn hơn 100

Cuối cùng, để đảm bảo sự đồng đều và công bằng trong quá trình huấn luyện mô hình, chúng ta đã chia dữ liệu thành 28 ngành nghề, mỗi ngành nghề có đúng 100 mô tả khác nhau. Điều này giúp mô hình có cơ hội học từ đủ độ đa dạng và biểu diễn chính xác các đặc trưng của từng ngành nghề. Việc này không chỉ giúp mô hình trở nên hiệu quả hơn mà còn tạo ra một tập dữ liệu thích hợp cho quá trình kiểm thử và đánh giá mô hình.

A list of jobs in a computer

Description automatically generated

Hình 4. Dữ liệu đã được làm sạch

# ĐỀ XUẤT MÔ HÌNH NGHIÊN CỨU

Để phân loại ngành nghề của các công ty dựa trên mô tả về công ty, chúng ta có thể áp dụng một loạt các phương pháp học máy truyền thống và kỹ thuật học sâu để tận dụng tối đa thông tin từ mô tả. Dưới đây là một số phương pháp mà chúng ta có thể thử nghiệm để xây dựng mô hình phân loại hiệu quả:

* **Phương pháp Bag-of-Words (BoW):** Chuyển đổi mỗi mô tả công ty thành vector Bag-of-Words để biểu diễn tần suất xuất hiện của từng từ. Sử dụng mô hình học máy như Naive Bayes, Support Vector Machines (SVM), hoặc Logistic Regression để phân loại ngành nghề[2].
* **TF-IDF (Term Frequency-Inverse Document Frequency):** Sử dụng biểu diễn TF-IDF để đánh giá tầm quan trọng của từng từ trong mô tả công ty. Áp dụng các mô hình học máy như Naive Bayes, SVM, hoặc Logistic Regression[3].
* **Word Embeddings:** Sử dụng các mô hình như Word2Vec, GloVe hoặc FastText để nhúng từng từ trong mô tả công ty thành không gian vector có chiều sâu. Đưa vào mô hình học máy như Neural Network hoặc Recurrent Neural Network (RNN) để phân loại ngành nghề[4].
* **BERT (Bidirectional Encoder Representations from Transformers):** Sử dụng BERT hoặc các mô hình học sâu Transformer khác để tạo biểu diễn cho mô tả công ty. Đưa vào mô hình phân loại đa lớp (multiclass classification) để dự đoán ngành nghề[5].
* **CNN (Convolutional Neural Network) cho Văn bản:** Áp dụng mô hình CNN để học các đặc trưng cấp độ cao từ các cụm từ trong mô tả công ty. Sử dụng các lớp phân loại để dự đoán ngành nghề[6].
* **Học máy có giám sát với các thuật toán như Random Forest hoặc Gradient Boosting:** Sử dụng các thuật toán này để xây dựng mô hình phân loại dựa trên biểu diễn của mô tả công ty[7].

Qua việc thử nghiệm nhiều phương pháp, chúng ta có cơ hội tìm ra mô hình phù hợp nhất với dữ liệu cụ thể và đạt được độ chính xác cao trong việc dự đoán ngành nghề của các công ty dựa trên mô tả của chúng. Việc này không chỉ mang lại sự linh hoạt mà còn tăng khả năng áp dụng mô hình cho các tình huống thực tế.

# ĐỀ XUẤT CÁC MÔ HÌNH HỌC MÁY

## 4.1. Sử dụng phương pháp Bag-of-Words kết hợp SVM

Phương pháp Bag-of-Words (BoW)[2] là một khái niệm cơ bản và phổ biến trong xử lý ngôn ngữ tự nhiên (NLP) được sử dụng để biểu diễn văn bản. Ý tưởng chính của BoW là xem xét một đoạn văn bản như một "túi từ" (bag of words), nơi chúng ta bỏ qua thứ tự từ và chỉ quan tâm đến số lần xuất hiện của từ trong văn bản. Quá trình này bao gồm việc thu thập dữ liệu từ các nguồn như mô tả công ty, tiền xử lý văn bản để loại bỏ các stop words và chuyển đổi văn bản thành lowercase, sau đó xây dựng một từ điển bằng cách đếm số lần xuất hiện của từ trong toàn bộ tập dữ liệu. Mỗi văn bản sau đó được biểu diễn dưới dạng vector, với mỗi phần tử của vector tương ứng với số lần xuất hiện của từ trong từ điển. Phương pháp này thường được sử dụng trong các tác vụ như phân loại văn bản, phân tích ý kiến và dự đoán nhãn từ văn bản.

Support Vector Machine (SVM) là một thuật toán học máy được ứng dụng rộng rãi trong phân loại và hồi quy[8]. SVM tìm ra một siêu phẳng (hyperplane) tốt nhất để phân tách giữa các lớp dữ liệu. Thuật toán này sử dụng các điểm dữ liệu gọi là vector hỗ trợ để xác định siêu phẳng và có khả năng áp dụng các hàm kernel để chuyển dữ liệu vào không gian cao chiều, giúp tìm ra siêu phẳng phân tách hiệu quả hơn. SVM thường được áp dụng trong nhiều lĩnh vực, bao gồm phân loại văn bản, hình ảnh và nhiều ứng dụng khác.

Khi kết hợp BoW và SVM trong mô hình phân loại ngành nghề của công ty, chúng ta sẽ sử dụng BoW để biểu diễn mô tả công ty thành các vector. Sau đó, SVM được áp dụng để huấn luyện mô hình phân loại dựa trên các vector này. Quá trình này có thể bao gồm điều chỉnh tham số của SVM để tối ưu hóa hiệu suất mô hình, đảm bảo rằng mô hình có khả năng tổng quát hóa tốt trên tập dữ liệu và đưa ra dự đoán chính xác về ngành nghề của các công ty dựa trên mô tả của chúng.

## 4.2. Sử dụng CNN cho Văn bản

Sử dụng Convolutional Neural Network (CNN) cho xử lý văn bản là một phương pháp mạnh mẽ và đa dạng, bất chấp việc CNN thường được biết đến chủ yếu trong lĩnh vực xử lý hình ảnh. Trong bối cảnh này, CNN có thể được áp dụng để trích xuất đặc trưng quan trọng từ mô tả công ty, mở rộng khả năng ứng dụng của nó vào lĩnh vực xử lý ngôn ngữ tự nhiên[6].

Khái niệm cơ bản của CNN trong ngữ cảnh xử lý văn bản là sử dụng các lớp convolution để tìm ra các đặc trưng quan trọng từ dữ liệu. Các embedding layer được sử dụng để chuyển đổi từ văn bản thành các vector có chiều thấp, và sau đó, văn bản được chia thành các cửa sổ nhỏ để áp dụng lớp convolution. Qua các lớp convolution, CNN học được các đặc trưng cấp độ cao từ các cụm từ, giữ lại thông tin về thứ tự của chúng.

Max pooling và Fully Connected Layers được sử dụng sau lớp convolution để giảm kích thước của đầu ra và giữ lại đặc trưng quan trọng. Các lớp fully connected cuối cùng được sử dụng để thực hiện phân loại, đưa ra dự đoán về ngành nghề của công ty dựa trên mô tả của nó.

Để tránh overfitting, CNN thường sử dụng kỹ thuật dropout và các kỹ thuật regularization để kiểm soát trọng số của mô hình trong quá trình huấn luyện. Quá trình huấn luyện mô hình CNN thường đòi hỏi sự chú ý đặc biệt đối với các siêu tham số và quá trình tinh chỉnh chúng để đạt được hiệu suất tốt nhất.

Khi so sánh giữa Bag-of-Words (BoW) và CNN, BoW biểu diễn dựa trên số lượng từ xuất hiện trong văn bản và không giữ thông tin về thứ tự từ. Nó thích hợp cho các vấn đề đơn giản và dữ liệu lớn. Ngược lại, CNN học được các đặc trưng từ các cửa sổ nhỏ của văn bản, giữ lại thông tin về thứ tự của từ và thích hợp cho việc xử lý dữ liệu có cấu trúc phức tạp hơn. Sự lựa chọn giữa BoW và CNN phụ thuộc vào độ phức tạp của vấn đề và mức độ chi tiết thông tin cần thiết để đạt được hiệu suất cao.

# THỰC THI MÔ HÌNH

## 5.1. Mô hình BoW kết hợp SVM

### 5.1.1. Xây dựng mô hình

Quá trình huấn luyện mô hình SVM bằng phương pháp Bag-of-Words (BoW) để dự đoán ngành nghề của công ty dựa trên mô tả công ty là một quá trình chi tiết và cân nhắc cẩn thận với các bước nhất định nhằm đảm bảo hiệu suất và chất lượng của mô hình. Dưới đây là mô tả chi tiết về từng bước của quá trình này:

Chuẩn bị Dữ liệu:

- Sử dụng `CountVectorizer` để chuyển đổi mô tả công ty thành ma trận Bag-of-Words. Điều này bao gồm việc xây dựng một từ điển từ toàn bộ tập dữ liệu, sau đó biểu diễn mỗi mô tả công ty dưới dạng vector, với mỗi phần tử của vector là tần suất xuất hiện của từ trong từ điển.

- Ma trận Bag-of-Words này sẽ đóng vai trò quan trọng trong quá trình huấn luyện, vì nó bám sát vào sự xuất hiện của từng từ trong mô tả, giúp mô hình SVM hiểu rõ mối liên kết giữa các từ và ngành nghề.

Chuẩn bị Nhãn:

- Sử dụng `LabelEncoder` để chuyển đổi nhãn ngành nghề thành dạng số. Việc này giúp mô hình SVM làm việc hiệu quả hơn vì nó có thể xử lý dữ liệu số hóa.

Xây Dựng và Huấn Luyện Mô Hình SVM:

- Sử dụng máy học Support Vector Machine (SVM) với kernel tuyến tính. Lựa chọn kernel tuyến tính thường được ưu tiên trong các bài toán phân loại văn bản với Bag-of-Words vì tính đơn giản và hiệu quả của nó.

- Thực hiện huấn luyện trên ma trận Bag-of-Words của tập dữ liệu huấn luyện và nhãn đã được chuyển đổi sang dạng số. Quá trình này đào tạo mô hình SVM để tối ưu hóa sự chính xác trong việc phân loại dữ liệu, và tham số C được sử dụng để kiểm soát độ linh hoạt của mô hình. Điều này giúp mô hình đạt được sự cân bằng giữa việc tìm hiểu từ dữ liệu huấn luyện và khả năng tổng quát hóa đối với dữ liệu mới.

Quá trình này không chỉ giúp mô hình học được mối liên kết phức tạp giữa từng từ trong mô tả công ty và ngành nghề tương ứng, mà còn giúp đảm bảo rằng mô hình là một công cụ mạnh mẽ và linh hoạt cho việc dự đoán ngành nghề của các công ty dựa trên mô tả của chúng.

### 5.1.2. Kết quả của mô hình

A screenshot of a document

Description automatically generated

Hình 5. Kết quả của mô hình BoW - SVM

Mô hình hiện thực của chúng ta đã đạt được một độ chính xác đáng kể là 56.53%. Qua sự phân tích, chúng ta nhận thấy rằng mô hình thực sự xuất sắc trong việc dự đoán các ngành nghề như Colleges & Universities, Civic & Social Services, Restaurants & Cafes, National Agencies,..., nơi mà khả năng dự đoán chính xác là rất cao.

Tuy nhiên, như mọi mô hình dự đoán, vẫn còn những khía cạnh mà chúng ta cần cải thiện. Chúng ta nhận thấy sự không chính xác xuất hiện trong việc dự đoán một số ngành nghề cụ thể như Computer Hardware Development, Consumer Product Manufacturing, Enterprise Software & Network Solutions, Internet & Web Services,... Điều này đặt ra một thách thức và cơ hội để cải thiện mô hình, có thể thông qua việc tăng cường dữ liệu hoặc điều chỉnh các tham số của mô hình để nâng cao khả năng tổng quát hóa.

Dù vậy, thành công của mô hình trong việc dự đoán nhiều ngành nghề chứng tỏ sự tiến triển và tiềm năng trong việc áp dụng phương pháp học máy vào lĩnh vực dự đoán ngành nghề của công ty. Điều quan trọng là chúng ta tiếp tục nghiên cứu và tối ưu hóa để tạo ra một mô hình mạnh mẽ và đáng tin cậy hơn.

Confusion Matrix

A screenshot of a crossword puzzle

Description automatically generated

Hình 6. Confusion Matrix của mô hình BoW-SVM

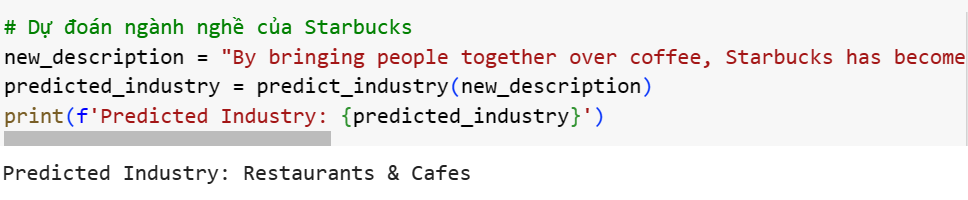
Qua việc phân tích confusion matrix, ta nhận thấy rằng đường chéo chính của nó khá đậm, cho thấy mô hình của chúng ta đã có những dự đoán chính xác và đáng tin cậy[9]. Tuy nhiên, đối diện với sự thành công đó, vẫn tồn tại những sự nhầm lẫn đáng kể, đặc biệt là trong việc dự đoán các ngành nghề như Computer Hardware Development, Internet & Web Services, và Enterprise Software & Network Solutions.

Điều này đưa ra một cơ hội quan trọng để cải thiện mô hình. Bằng cách tập trung vào những ngành nghề này và tìm hiểu kỹ hơn về đặc trưng cụ thể của chúng, ta có thể điều chỉnh mô hình để nâng cao khả năng dự đoán chính xác trong các trường hợp này. Có thể sẽ cần tăng cường dữ liệu cho các ngành nghề đặc biệt này hoặc điều chỉnh các tham số của mô hình để đạt được sự cân bằng tốt hơn giữa độ chính xác và tổng quát hóa.

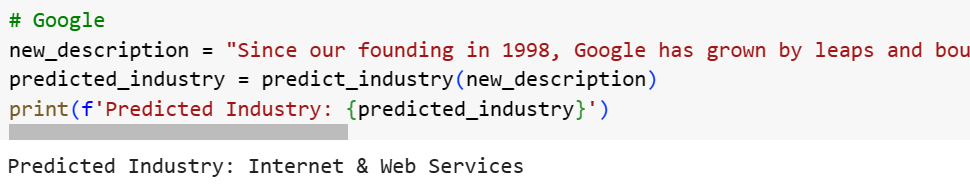
Mặc dù vẫn còn những thách thức, nhưng sự nhìn nhận sâu sắc vào confusion matrix là một công cụ hữu ích để chúng ta hiểu rõ hơn về hiệu suất của mô hình và hướng đi để cải thiện nó. Điều này cũng là một phần quan trọng trong quá trình liên tục tối ưu hóa và phát triển mô hình dự đoán ngành nghề công ty.

### 5.1.3. Kết quả dự đoán của mô hình

Dự đoán mô tả của Starbucks



Dự đoán mô tả của Google



Dự đoán mô tả của công ty FPT

A screenshot of a computer program

Description automatically generated

## 5.2. Mô hình CNN cho Văn bản

### 5.2.1. Xây dựng mô hình

Quá trình huấn luyện mô hình Convolutional Neural Network (CNN) để dự đoán ngành nghề của công ty dựa trên mô tả công ty là một quá trình phức tạp và linh hoạt, bao gồm các bước chi tiết sau đây:

Chuẩn Bị Dữ Liệu Cho Mô Hình:

- Chuyển đổi Mô Tả Công Ty thành Dãy Số: Sử dụng `Tokenizer` để ánh xạ từng từ trong mô tả công ty thành các số nguyên, tạo ra một biểu diễn dãy số của văn bản.

- Đảm Bảo Độ Dài Đồng Đều: Sử dụng `pad\_sequences` để điều chỉnh độ dài của các dãy số, đảm bảo rằng mỗi mô tả có độ dài như nhau, điều này quan trọng để mô hình có thể xử lý chúng hiệu quả.

Xây Dựng Mô Hình CNN:

- Mô Hình Tuần Tự: Sử dụng `Sequential()` để xây dựng mô hình, tạo ra một chuỗi các lớp được thêm vào theo thứ tự từ đầu vào đến đầu ra.

- Chuyển Đổi Từ Vựng: Áp dụng lớp `Embedding` để chuyển đổi từng từ trong từ vựng thành các vector có chiều thấp, giúp mô hình hiểu được mối quan hệ giữa các từ.

- Lớp Convolution (Conv1D): Thêm lớp convolution với 128 bộ lọc, kích thước cửa sổ là 5 và hàm kích hoạt là ReLU. Lớp này giúp mô hình học được các đặc trưng quan trọng từ dữ liệu.

- Lớp GlobalMaxPooling1D: Áp dụng lớp pooling để giảm chiều dài của đầu ra, giữ lại thông tin quan trọng nhất.

- Lớp Fully Connected (Dense): Thêm lớp fully connected với 64 neuron và hàm kích hoạt ReLU để học các mối quan hệ phức tạp giữa các đặc trưng đã học được.

- Lớp Dropout: Áp dụng lớp dropout để ngăn chặn hiện tượng overfitting bằng cách ngẫu nhiên loại bỏ một số neuron trong quá trình huấn luyện.

- Lớp Output: Cuối cùng, thêm lớp output với số neuron bằng số lớp đầu ra và hàm kích hoạt softmax để dự đoán xác suất của mỗi lớp.

Biên Dịch Mô Hình:

- Cấu Hình Mô Hình: Sử dụng `compile` để cấu hình mô hình với optimizer là 'adam', hàm loss là 'sparse\_categorical\_crossentropy', và độ đo hiệu suất là 'accuracy'. Cấu hình này giúp mô hình được huấn luyện hiệu quả để dự đoán ngành nghề của công ty dựa trên mô tả.

### 5.2.2. Kết quả của mô hình

A screenshot of a computer code

Description automatically generated

Sau 100 epoch, mô hình đã đạt được độ chính xác là 45.53%. Kết quả này là một biểu hiện của quá trình học của mô hình trên tập dữ liệu huấn luyện qua nhiều vòng lặp, và cung cấp một cái nhìn tổng quan về khả năng dự đoán của nó trên tập kiểm tra.

Mặc dù độ chính xác này có thể được xem xét là một kết quả khá, nhưng nó cũng đặt ra câu hỏi về cơ hội cải thiện. Trong quá trình tiếp theo, có thể tăng số lượng epoch, điều chỉnh tỷ lệ học (learning rate), hoặc thực hiện các điều chỉnh khác trên mô hình để kiểm soát và cải thiện hiệu suất của nó.

Confusion Matrix

A screenshot of a computer

Description automatically generated

Hình 7. Confusion Matrix của mô hình CNN

Dựa vào confusion matrix, có thể thấy mô hình đang dự đoán khá tốt. Mô hình hiển thị sự chính xác đáng kể trong việc dự đoán các ngành nghề như Advertising & Public Relations, Health Care Services & Hospitals, và Colleges & Universities. Điều này chứng tỏ khả năng ổn định và chính xác của mô hình đối với một số lĩnh vực chuyên sâu.

Tuy nhiên, vẫn còn một số sai lầm trong dự đoán, đặc biệt là trong việc phân biệt giữa Enterprise Software & Network Solutions và các ngành nghề có liên quan đến Công nghệ, cũng như giữa Machinery Manufacturing và Electronics Manufacturing. Điều này đặt ra một thách thức trong việc cải thiện khả năng phân loại của mô hình trong những ngành nghề có đặc trưng tương đồng.

Để nâng cao hiệu suất, có thể cân nhắc tăng cường dữ liệu trong những lĩnh vực đó, điều chỉnh trọng số của mô hình, hoặc thực hiện fine-tuning để tối ưu hóa dự đoán. Bằng cách này, mô hình có thể học được các đặc trưng cụ thể và đặc biệt hóa khả năng phân biệt giữa các ngành nghề tương đồng.

### 5.2.3. Kết quả dự đoán của mô hình

Dự đoán mô tả đại học Standford

A screen shot of a computer

Description automatically generated

Dự đoán mô tả của công ty Apple

A screenshot of a computer code

Description automatically generated

Dự đoán mô tả công ty phần mềm Bdata

A screenshot of a computer program

Description automatically generated

# TỔNG KẾT

Đồ án đã tận dụng một loạt các mô hình học máy để giải quyết vấn đề dự đoán ngành nghề của công ty dựa trên mô tả công ty. Trong quá trình này, hai mô hình chính được triển khai là mô hình Bag-of-Words kết hợp Support Vector Machine (BoW + SVM) và mô hình Convolutional Neural Network (CNN). Dưới đây là sự tổng kết về hiệu suất của cả hai mô hình:

|  |  |  |
| --- | --- | --- |
| Phương pháp | Mô tả | Độ chính xác |
| BoW kết hợp SVM | Sử dụng phương pháp Bag-of-Words để biểu diễn mô tả công ty thành vector. Áp dụng mô hình Support Vector Machine (SVM) với kernel tuyến tính. | 56.53% |
| CNN | Sử dụng Convolutional Neural Network để học đặc trưng từ mô tả công ty. | 45.53% |

**Nhận xét và Lựa Chọn Mô Hình Thích Hợp:**

- Mô hình BoW kết hợp SVM đã cho thấy hiệu suất tốt hơn so với mô hình CNN, với độ chính xác đạt 56.53%.

- Điều này đề xuất rằng mô hình BoW + SVM phản ánh tốt hơn sự phức tạp của dữ liệu mô tả công ty trong bài toán này.

- Tuy nhiên, đồ án cũng đã nhấn mạnh rằng cả hai mô hình đều phải đối mặt với thách thức của việc phân loại ngành nghề từ mô tả công ty, nơi thông tin có thể đa dạng và phức tạp.

**Hướng phát triển:**

- Để cải thiện hiệu suất, có thể xem xét thử nghiệm các phương pháp biểu diễn văn bản khác nhau, như Word Embeddings hoặc BERT, để xem chúng có thể nắm bắt thông tin phức tạp hơn từ mô tả công ty hay không.

- Cân nhắc thực hiện fine-tuning trên mô hình đã chọn để tối ưu hóa hiệu suất cho ngành nghề cụ thể mà bạn đang làm việc.

- Hợp nhất các mô hình hoặc triển khai mô hình kết hợp để tận dụng sức mạnh của từng loại mô hình.

# TÀI LIỆU THAM KHẢO

[1] “Companies & Reviews,” Glassdoor. Accessed: Jan. 16, 2024. [Online]. Available: https://www.glassdoor.com/Reviews/index.htm

[2] “Bag-of-words model,” *Wikipedia*. Dec. 13, 2023. Accessed: Jan. 16, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Bag-of-words\_model&oldid=1189645647

[3] “Understanding TF-IDF for Machine Learning,” Capital One. Accessed: Jan. 16, 2024. [Online]. Available: https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/

[4] “Word embeddings in NLP: A Complete Guide.” Accessed: Jan. 16, 2024. [Online]. Available: https://www.turing.com/kb/guide-on-word-embeddings-in-nlp

[5] “BERT (language model),” *Wikipedia*. Dec. 13, 2023. Accessed: Jan. 16, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=BERT\_(language\_model)&oldid=1189729206

[6] “Thuật toán CNN trong xử lý ngôn ngữ tự nhiên – Luyện Code.” Accessed: Jan. 16, 2024. [Online]. Available: https://blog.luyencode.net/thuat-toan-cnn-trong-nlp/

[7] “Random forest,” *Wikipedia*. Dec. 18, 2023. Accessed: Jan. 16, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Random\_forest&oldid=1190553008

[8] “Machine Learning cơ bản.” Accessed: Jan. 16, 2024. [Online]. Available: https://machinelearningcoban.com/2017/04/09/smv/

[9] “Tìm hiểu về Confusion matrix trong Machine Learning? - Viblo.” Accessed: Jan. 16, 2024. [Online]. Available: https://viblo.asia/p/tim-hieu-ve-confusion-matrix-trong-machine-learning-Az45bRpo5xY

ĐẠI HỌC HUẾ

**KHOA KỸ THUẬT VÀ CÔNG NGHỆ**

🙠🙟🕮🙝🙢

**PHIẾU ĐÁNH GIÁ ĐỒ ÁN**

**Học kỳ I, năm học 2023 - 2024**

|  |  |
| --- | --- |
| **Cán bộ chấm thi 1** | **Cán bộ chấm thi 2** |
| **Nhận xét:**                  **Điểm đánh giá của CBCT1:**  Bằng số:.................................................  Bằng chữ: ............................................. | **Nhận xét:**                  **Điểm đánh giá của CBCT2:**  Bằng số:..................................................  Bằng chữ:............................................... |

Điểm kết luận:..........................................................................................................................

Bằng số: ...................................................................................................................................

Bằng chữ: ................................................................................................................................

*Thừa Thiên Huế, ngày tháng năm 2024*

|  |  |
| --- | --- |
| **Cán bộ chấm thi 1**  *(Ký và ghi rõ họ và tên)* | **Cán bộ chấm thi 2**  *(Ký và ghi rõ họ và tên* |