**CÁC NGHIÊN CỨU CÓ LIÊN QUAN**

"Text Classification Algorithms: A Survey" (2018) - Tác giả: Ahmad Majid Tavana, Masoud Mohammadian. Nghiên cứu này tóm tắt các thuật toán phân loại văn bản phổ biến và so sánh hiệu suất của chúng trên các tập dữ liệu khác nhau.

"Convolutional Neural Networks for Sentence Classification" (2014) - Tác giả: Yoon Kim. Nghiên cứu này đề xuất sử dụng mạng nơ-ron tích chập (CNN) để phân loại câu văn dựa trên cấu trúc cú pháp của chúng.

"A Survey on Neural Network Language Models" (2018) - Tác giả: Yongwu Miao, Phil Blunsom. Nghiên cứu này tổng quan các mô hình ngôn ngữ dựa trên mạng nơ-ron, bao gồm cả phân loại văn bản theo chủ đề.

"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" (2018) - Tác giả: Jacob Devlin và cộng sự tại Google AI Language. Nghiên cứu này giới thiệu mô hình BERT, một mô hình pre-trained sâu bidirectional transformer được sử dụng rộng rãi cho nhiều tác vụ NLP, bao gồm phân loại văn bản theo chủ đề.

"Hierarchical Attention Networks for Document Classification" (2016) - Tác giả: Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, Eduard Hovy. Nghiên cứu này giới thiệu một kiến trúc mạng nơ-ron có tên là Hierarchical Attention Networks (HAN) để phân loại văn bản theo chủ đề, với khả năng tập trung vào các phần quan trọng của văn bản.

"Efficient and Robust Automated Machine Learning on Text Classification" (2019) - Tác giả: Hanyang Peng, Joey Tianyi Zhou, Zhi Ang Eu, Shuo Zhang, Kaixiang Lin. Nghiên cứu này tập trung vào việc tối ưu và làm cho mạnh mẽ các phương pháp tự động hóa trong việc phân loại văn bản theo chủ đề.

"Bertopic: Hierarchical Topic Modeling with BERT" (2021) - Tác giả: Thomas Bosc, Thomas Wolf. Nghiên cứu này kết hợp mô hình BERT với một kiến trúc hierarchical topic modeling để phân loại văn bản theo chủ đề.

**LỜI CẢM ƠN**

Đầu tiên, em xin gửi lời cảm ơn chân thành đến Trường Đại Học Kỹ Thuật - Công Nghệ Cần Thơ đã tạo điều kiện đưa các nghành công nghệ vào trương trình giảng dạy. Đặc biệt, em xin gửi lời cảm ơn sâu sắc đến giảng viên bộ môn cô Nguyễn Nhựt Quỳnh đã giúp đỡ, truyền đạt những kiến thức quý báu cho em trong suốt thời gian học tập vừa qua. Trong thời gian tham gia học tập chúng em đã được tiếp thu những kiến thức quan trọng, mới mẻ của nghành công nghệ, tinh thần học tập hiệu quả, nghiêm túc. Đây chắc chắn sẽ là những kiến thức quý báu, là hành trang để em có thể vững bước sau này.

Đề tài “Xây dựng hệ thống phân loại văn bản theo chủ đề ” là đề tài vô cùng bổ ích và có tính thực tế cao. Đảm bảo cung cấp đủ kiến thức, gắn liền với nhu cầu thực tiễn của sinh viên. Tuy nhiên, do vốn kiến thức còn nhiều hạn chế và khả năng tiếp thu thực tế còn nhiều bỡ ngỡ. Mặc dù em đã cố gắng hết sức nhưng chắc chắn bài tiểu luận khó có thể tránh khỏi những thiếu sót và nhiều chỗ còn chưa chính xác, kính mong cô xem xét và góp ý để bài tiểu luận của em được hoàn thiện hơn.

**TÓM TẮT**

Trong những năm gần đây các hệ thống giúp phân loại văn bản ngày càng phát triển để phục vụ nhu cầu cần thiết cho người dùng, với độ chính xác cao cùng với tính ổn định nó đã giúp cho chúng ta cũng như các doanh nghiệp tiết kiệm được thời gian rất lớn cho việc quản lí và phân loại dữ liệu phục vụ cho công việc.

Hệ thống phân loại văn bản theo chủ đề là một công nghệ được thiết kế để tự động phân loại văn bản thành các chủ đề khác nhau dựa trên nội dung giúp người dùng có được dữ liệu cần thiết mà không tốn quá nhiều thời gian.

Hệ thống phân loại văn bản theo chủ đề có vai trò quan trọng trong việc tổ chức, tìm kiếm và trích xuất thông tin từ các tài liệu văn bản lớn giúp tối ưu quá trình làm việc và quản lí dữ liệu trong nhiêu lĩnh vực khác nhau. Nhận thấy đây là một đề tài ứng dụng thực tiễn cao cùng với mong muốn tìm tòi học hỏi nên nhóm chúng em quyết định chọn đề tài “Xây dựng hệ thống phân loại văn bản theo chủ đề” cho đồ án lần này.

**LỜI CAM ĐOAN**

Nhóm chúng em xin giới thiệu với cô và mọi người đề tài “Xây dựng hệ thống phân loại văn bản theo chủ đề”. Chúng em chọn đề tài này vì nó thiết thực và bổ ích cho các bạn sinh viên trong học tập cũng như trong môi trường làm việc với máy tính trở nên thông dụng như hiện nay.

Trong quá trình thực hiện đề tài này còn có nhiều thiếu sót do kiến thức còn sơ sài nhưng những nội dung trình bày trong quyển báo cáo này là những biểu hiện kết quả của chúng em đạt được dưới sự hướng dẫn của Cô Nguyễn Nhựt Quỳnh.

Chúng em xin cam đoan rằng: Những nội dung trình bày trong quyển báo cáo tiểu luận này không phải là bản sao chép từ bất kì tiểu luận nào có trước. Nếu không đúng sự thật, chúng em xin chịu mọi trách nhiệm trước hội đồng cố vấn.

Giảng viên hướng dẫn Sinh viên thực hiện

(Ký tên)  (Ký tên)

CHƯƠNG 1 GIỚI THIỆU

11 GIỚI THIỆU TỔNG QUAN VÀ LÝ DO CHỌN ĐỀ TÀI

Với xu hướng phát triển của thế giới các ngành công nghệ đã có những thay đổi những bước tiến vượt bật và là một trong những lĩnh vực được ưu tiên phát triển hàng đầu trong đó trí tuệ nhân tạo đóng góp một phần rất lớn tạo nên cuộc thành công về chuyển đổi số hóa trong xã hội giúp con người dễ dàng tiếp cận với những thông tin trong đời sống, ngoài ra trong doanh nghiệp, y học, giáo dục, truyền thông.... các công nghệ dược ứng dụng một cách rất tiến bộ.

Trong thời đại internet ngày nay, dữ liệu ngày càng trở nên phong phú và đa dạng, đặc biệt trong bối cảnh của cuộc cách mạng công nghiệp 4.0. Cùng với sự bùng nổ thông tin, nhu cầu phân loại và khai thác thông tin trở nên ngày càng quan trọng trong mọi tổ chức và quy mô hoạt động. Khi dữ liệu văn bản được tạo ra và chia sẻ một cách rộng rãi. Các công ty, tổ chức và cá nhân thường phải đối mặt với một lượng lớn các văn bản, bao gồm email, tài liệu, bài báo, tin tức, và nhiều loại tài liệu khác. Việc phân loại tự động các văn bản có thể giúp tối ưu hóa quy trình làm việc và quản lý thông tin.

Phân loại văn bản theo chủ đề đóng vai trò quan trọng trong việc tổ chức và xử lý thông tin trong thế giới số ngày nay. Trong một thế giới mà thông tin phát triển không ngừng và dữ liệu văn bản ngày càng lớn, việc tự động gán nhãn cho các văn bản dựa trên nội dung của chúng là một yếu tố không thể phủ nhận. Qua quá trình này, chúng ta có thể tổ chức và tìm kiếm thông tin một cách hiệu quả hơn, từ việc phân loại email, quản lý tài liệu đến việc tổ chức dữ liệu trên internet.

Trong lĩnh vực thương mại, việc phân loại tin tức, phản hồi khách hàng và phân tích thị trường dựa trên nội dung văn bản có thể giúp các doanh nghiệp hiểu rõ hơn về nguyên nhân và kỳ vọng của khách hàng. Trong lĩnh vực y tế, phân loại văn bản có thể được sử dụng để phát hiện các triệu chứng, phân loại bệnh lý và dự đoán xu hướng dịch bệnh. Trong lĩnh vực tin tức và truyền thông, việc phân loại tin tức theo chủ đề giúp người đọc dễ dàng truy cập và tiếp cận thông tin một cách nhanh chóng và thuận tiện.

Nhận thấy đây là một chủ đề thực tế và có tính ứng dụng cao nên chúng em quyết định tìm hiểu về đề tài “XÂY DỰNG HỆ THỐNG PHÂN LOẠI VĂN BẢN THEO CHỦ ĐỀ” và sử dụng mô hình Naive Bayes để hổ trợ cho việc phân loại văn bản một cách nhanh chóng cũng như độ ổn định cao.

12 CÁC NGHIÊN CỨU CÓ LIÊN QUAN

13 MỤC TIÊU VÀ PHẠM VI NGHIÊN CỨU

Mục tiêu nghiên cứu là xây dựng được hệ thống phân loại văn bản theo chủ đề sử dụng mô hình Naïve Bayes.

Phạm vi nghiên cứu là xây dựng được hệ thống phân loại được một số chủ đề cho trước như thể thao, doanh nghiệp,…

14 ĐỐI TƯỢNG NGHIÊN CỨU

Các lí thuyết cơ bản về Machine Learning, NLP, Naïve Bayes, TF-IDF.

Quy trình máy học áp dụng vào hệ thống phân loại văn bản theo chủ đề.

Cách xây dựng mô hình cũng như các hoạt động của Naïve Bayes.

15 PHƯƠNG PHÁP NGHIÊN CỨU

Các phương pháp nghiên cứu được áp dụng trong hệ thống phân loại văn bản theo chủ đề:

* Thu thập tài liệu từ các nguồn chính thống để tìm hiểu về Machine Learning, NLP, TF-IDF,…
* Mô hình xử lý trong máy học cũng như mô hình Naïve Bayes.
* Phân tích, xử lí dữ liệu và xử lí ngôn ngữ tự nhiên.
* Phương pháp đánh giá mô hình đã xây dựng.

16 BỐ CỤC ĐỒ ÁN

Chương 1: Giới thiệu

Giới thiệu tổng quan và lý do chọn đề tài, mục tiêu và phạm vi nghiên cứu, đối tượng và phương pháp nghiên cứu, bố cục của đồ án.

Chương 2: Cơ sở lý thuyết

Giới thiệu về NLP, Machine Learning, Naïve Bayes, TF-IDF.

Chương 3: Phương pháp thực hiện

Thu thập dữ liệu, xử lý dữ liệu, huấn luyện , đánh giá mô hình.

Chương 4: Kết quả thực nghiệm và đánh giá giải pháp

Tổng quan kết quả đạt được và kết quả thực tế.

Chương 5: Kết luận và hướng phát triển

Kết luận và hướng phát triển

# CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

## 5.1 KẾT LUẬN

Sau khi tìm hiểu và thực hiện đề tài “XÂY DỰNG HỆ THỐNG PHÂN LOẠI VĂN BẢN THEO CHỦ ĐỀ” chúng em đã hoàn thành một hệ thống phân loại văn bản theo một số chủ đề cho trước như thể thao, doanh nghiệp,...

Hiểu được những lý thuyết cơ bản trong quá trình thực hiện như:

* Nắm được kiến thức cơ bản về hệ thống phân loại.
* Kiến thức về NLP, Machine Learning, TF-IDF, mô hình Naïve Bayes.
* Cách sử lý dữ liệu cũng như huấn luyện mô hình phân loại.

## 5.2 HƯỚNG PHÁT TRIỂN

Trong phạm vi đồ án lần này chúng em đã trình bày được những thành phần cơ bản nhất về hệ thống phân loại văn bản theo chủ đề. Tuy việc nghiên cứu chưa ứng dụng vào thực tiễn một cách chuyên sâu và bài bản nhưng cơ bản hệ thống có thể phân loại được dữ liệu với độ chính xác khá ổn định theo một số chủ đề đã được huấn luyện từ trước.

So với kết quả hiện tại thì hệ thống phân loại văn bản của chúng em vẫn còn nhiều hạn chế nhưng với việc tìm tòi học hỏi thì chương trình này vẫn có thể phát triển ổn định hơn thông qua các hướng phát triển như sau:

* Mở rộng thêm dữ liệu trong bộ dữ liệu huấn luyện để tăng thêm độ chính xác trong quá trình phân loại.
* Thêm nhiều chủ đề ở nhiều lĩnh vực khác nhau làm đa dạng nguồn dữ liệu hơn.
* Xây dựng thêm giao diện thân thiện với người dùng để dễ dàng thao tác trong quá trình sử dụng.