

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**Phương Pháp Nghiên Cứu Khoa Học**

**Đề Tài Nghiên Cứu Khoa học  
Nhận diện cảm xúc trong văn bản bằng mô hình máy học**



**Giáo viên hướng dẫn: Đỗ Như Tài**

**Sinh viên thực hiện:**

**Trịnh Bảo Quân - 3119410335   
 Nguyễn Nhựt Hào - 3122410097**

**Ngô Khánh Tâm - 3122410370**

***Thành phố Hồ Chí Minh, tháng 05/2025***

## 

# **Mục Lục**

**[I. Lý do nghiên cứu (Vì sao chọn đề tài này/Ai thụ hưởng kết quả):](#_mfellfukgs2v) 1**

**[2. Lịch sử nghiên cứu (Ai đã làm gì? Làm đến đâu ?):](#_jiwip9l1rn9) 2**

**[3. Mục tiêu nghiên cứu:](#_k2oowviwjm3c) 2**

**[4. Khách thể nghiên cứu:](#_txiyss9hcqji) 3**

**[5. Phạm vi nghiên cứu:](#_wa01cwyspkeo) 4**

**[6. Vấn đề nghiên cứu:](#_iftl3ga2cbqh) 4**

**[7. Giả thuyết nghiên cứu:](#_4iu4mz4iau4x) 5**

**[8. Nội dung nghiên cứu:](#_o91zk6wdwv) 5**

**[9. Phương pháp chứng minh giả thuyết:](#_b8cciel9a965) 7**

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# **Lời mở đầu:**

# Đề tài nghiên cứu **"Nhận diện cảm xúc trong văn bản tiếng Việt bằng mô hình máy học"** tập trung vào một lĩnh vực ngày càng có vai trò quan trọng trong bối cảnh sự phát triển mạnh mẽ của dữ liệu trực tuyến tiếng Việt. Nghiên cứu này hướng đến việc khám phá khả năng ứng dụng của các mô hình học máy cơ bản trong việc tự động phân loại các loại cảm xúc khác nhau được thể hiện trong các đoạn văn bản ngắn. Chúng tôi kỳ vọng rằng kết quả của nghiên cứu sẽ cung cấp những hiểu biết giá trị về hiệu quả của các phương pháp tiếp cận đơn giản đối với bài toán nhận diện cảm xúc trong ngôn ngữ tiếng Việt.

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# **I. Lý do nghiên cứu (Vì sao chọn đề tài này/Ai thụ hưởng kết quả):**

**Tính cấp thiết:** Trong kỷ nguyên số, mạng xã hội đã trở thành một kênh giao tiếp và chia sẻ thông tin rộng rãi, nơi người dùng bày tỏ ý kiến và cảm xúc của mình thông qua các bình luận. Dữ liệu bình luận này, dù được nhập trực tiếp hay trích xuất từ hình ảnh (ví dụ: ảnh chụp màn hình hoặc ảnh tải lên), chứa đựng những thông tin giá trị về thái độ và tình cảm của cộng đồng đối với các sự kiện, sản phẩm, dịch vụ hoặc các vấn đề xã hội. Việc tự động nhận diện cảm xúc trong các bình luận này có tiềm năng ứng dụng to lớn trong nhiều lĩnh vực như phân tích dư luận, quản lý khủng hoảng truyền thông, cá nhân hóa trải nghiệm người dùng và hỗ trợ các hệ thống tương tác thông minh.

Đề tài này tập trung vào việc xây dựng và đánh giá một mô hình máy học có khả năng phân loại cảm xúc từ văn bản bình luận, đồng thời mô tả một quy trình cơ bản để thu thập và xử lý dữ liệu văn bản từ các nguồn khác nhau, bao gồm cả việc trích xuất văn bản từ ảnh.

Việc thực hiện đề tài này mang lại cơ hội quý báu cho sinh viên trong việc nắm vững các khái niệm và kỹ thuật cơ bản của xử lý ngôn ngữ tự nhiên và học máy, đồng thời áp dụng chúng vào một bài toán cụ thể và có tính ứng dụng.

**Đối tượng thụ hưởng kết quả:** Kết quả của nghiên cứu này có thể mang lại lợi ích cho:

* **Sinh viên và những người học:** Cung cấp kiến thức nền tảng và kinh nghiệm thực hành trong lĩnh vực nhận diện cảm xúc và xử lý văn bản.
* **Nhà nghiên cứu:** Đóng góp vào việc khám phá các phương pháp và mô hình khác nhau cho bài toán nhận diện cảm xúc.
* **Các nhà phát triển ứng dụng:** Cung cấp một cái nhìn tổng quan về quy trình xây dựng một hệ thống nhận diện cảm xúc, có thể là tiền đề cho việc phát triển các ứng dụng thực tế trong tương lai.

# **2. Lịch sử nghiên cứu (Ai đã làm gì? Làm đến đâu ?):**

**Các nghiên cứu gần đây về nhận diện cảm xúc trong văn bản:** Trong những năm gần đây, lĩnh vực nhận diện cảm xúc trong văn bản đã chứng kiến nhiều tiến bộ đáng kể nhờ sự phát triển của học sâu. Một nghiên cứu tiêu biểu ở nước ngoài là công trình **"Attention Is All You Need" của Ashish Vaswani và cộng sự (2017)**, được công bố tại Advances in Neural Information Processing Systems 30 (NIPS 2017). Công trình này giới thiệu kiến trúc Transformer, một nền tảng quan trọng cho nhiều mô hình học sâu hiện đại trong xử lý ngôn ngữ tự nhiên, bao gồm cả nhận diện cảm xúc. Việc đề cập đến nó cho thấy sự nhận thức về những tiến bộ mới nhất trong lĩnh vực.

Tại Việt Nam, đề tài nhận diện cảm xúc trong văn bản tiếng Việt cũng thu hút sự quan tâm của nhiều nhà nghiên cứu. Đồ án tốt nghiệp **"Nhận diện cảm xúc trong văn bản tiếng Việt bằng mô hình máy học" của Nguyễn Thanh Huy (2022)** tại Học viện Công nghệ Bưu chính Viễn thông là một ví dụ. Đồ án này đã nghiên cứu và xây dựng mô hình để phân loại cảm xúc trong văn bản tiếng Việt, có thể đã khám phá các phương pháp học máy truyền thống hoặc các kỹ thuật xử lý ngôn ngữ tự nhiên cơ bản.

Mặc dù đã có nhiều nghiên cứu về nhận diện cảm xúc trong văn bản, cả với các mô hình học máy truyền thống và các kiến trúc học sâu tiên tiến, vẫn còn tồn tại những thách thức và cơ hội để khám phá, đặc biệt là trong bối cảnh cụ thể của tiếng Việt. Các mô hình phức tạp đôi khi đòi hỏi nguồn lực tính toán lớn và dữ liệu huấn luyện đồ sộ. Nghiên cứu này, được thực hiện ở cấp độ sinh viên, hướng đến việc tìm hiểu và ứng dụng một mô hình học máy cơ bản nhưng hiệu quả để giải quyết bài toán nhận diện cảm xúc trong văn bản tiếng Việt. Hướng tiếp cận này nhằm mục đích làm rõ các nguyên tắc cơ bản của phân loại văn bản và đánh giá khả năng của một mô hình đơn giản trong việc xử lý ngôn ngữ tiếng Việt cho nhiệm vụ này.

# **3. Mục tiêu nghiên cứu:**

* **Mục tiêu tổng quát:** Nghiên cứu và xây dựng một mô hình máy học để nhận diện cảm xúc trong văn bản tiếng Việt.
* **Mục tiêu cụ thể:**
  1. Tìm hiểu lý thuyết về các mô hình máy học cơ bản (ví dụ: Naive Bayes, Logistic Regression) và các kỹ thuật tiền xử lý ngôn ngữ tự nhiên cơ bản liên quan đến bài toán nhận diện cảm xúc.
  2. Xây dựng một tập dữ liệu văn bản tiếng Việt có gán nhãn cảm xúc (ví dụ: tích cực, tiêu cực, trung tính) với quy mô vừa phải (ví dụ: khoảng 500 - 1500 mẫu) phù hợp cho việc huấn luyện và đánh giá các mô hình cơ bản.
  3. Xây dựng, huấn luyện và đánh giá hiệu suất của một trong các mô hình máy học đơn giản (ví dụ: Naive Bayes hoặc Logistic Regression) trên tập dữ liệu đã xây dựng, kỳ vọng đạt được độ chính xác ở mức khởi đầu (ví dụ: 55% - 70%) do tính đơn giản của mô hình và quy mô dữ liệu hạn chế.
  4. Phân tích kết quả, xác định những trường hợp mô hình dự đoán đúng và sai, đồng thời thảo luận về những yếu tố có thể ảnh hưởng đến hiệu suất của mô hình trong việc nhận diện cảm xúc tiếng Việt.
* **Đối tượng nghiên cứu:** Các đoạn văn bản tiếng Việt chứa đựng thông tin về cảm xúc (tập trung vào ba nhãn cơ bản: tích cực, tiêu cực, trung tính).
* **Cây mục tiêu:**
  1. **Gốc:** Nhận diện cảm xúc trong văn bản tiếng Việt bằng mô hình máy học cơ bản.
     + **Nhánh 1:** Tìm hiểu về các mô hình cơ bản (ví dụ: NB, LR) và tiền xử lý cơ bản.
     + **Nhánh 2:** Xây dựng tập dữ liệu tiếng Việt gán nhãn (quy mô vừa phải, 3 nhãn cảm xúc).
     + **Nhánh 3:** Xây dựng, huấn luyện và đánh giá một mô hình cơ bản (kỳ vọng độ chính xác 55-70%).
     + **Nhánh 4:** Phân tích kết quả và thảo luận về các yếu tố ảnh hưởng.
* **Nhiệm vụ nghiên cứu:**
  1. Nghiên cứu lý thuyết về nhận diện cảm xúc và các mô hình máy học cơ bản (ví dụ: Naive Bayes hoặc Logistic Regression) cùng các kỹ thuật tiền xử lý văn bản đơn giản.
  2. Thu thập hoặc tạo và tiền xử lý một tập dữ liệu văn bản tiếng Việt (khoảng 500-1500 mẫu), gán nhãn cho mỗi mẫu thuộc một trong ba loại cảm xúc: tích cực, tiêu cực, trung tính.
  3. Chia tập dữ liệu thành tập huấn luyện và tập kiểm tra.
  4. Lựa chọn và huấn luyện một mô hình máy học cơ bản trên tập dữ liệu huấn luyện.
  5. Đánh giá hiệu suất của mô hình trên tập dữ liệu kiểm tra bằng các độ đo như độ chính xác (accuracy), và có thể thêm precision, recall, F1-score.
  6. Viết báo cáo về quá trình nghiên cứu, kết quả đạt được và những phân tích, thảo luận liên quan đến hiệu suất của mô hình.

# **4. Khách thể nghiên cứu:**

Khách thể nghiên cứu của đề tài này là **các đoạn văn bản tiếng Việt chứa đựng thông tin về cảm xúc**. Cụ thể hơn, chúng tôi sẽ tập trung vào các bình luận, đánh giá hoặc các đoạn văn ngắn được thu thập từ các nguồn trực tuyến khác nhau.

* **Loại hình văn bản:** Các bình luận trên mạng xã hội, đánh giá sản phẩm/dịch vụ, hoặc các đoạn văn ngắn từ các diễn đàn trực tuyến bằng tiếng Việt.
* **Đặc điểm văn bản:** Văn bản không chính quy, có thể chứa lỗi chính tả, từ lóng, biểu tượng cảm xúc (emoji) và các hình thức diễn đạt cảm xúc đa dạng.
* **Nhãn cảm xúc:** Các đoạn văn bản này sẽ được gán nhãn thuộc một trong các loại cảm xúc cơ bản, bao gồm:
  + Tích cực (Positive)
  + Tiêu cực (Negative)
  + Trung tính (Neutral)

Quá trình nghiên cứu sẽ tập trung vào việc xây dựng và đánh giá mô hình máy học trên tập dữ liệu này để tự động phân loại cảm xúc ẩn chứa trong nội dung của chúng.

# **5. Phạm vi nghiên cứu:**

* **Giới hạn về đối tượng:** Đề tài này tập trung vào việc nhận diện cảm xúc trong các đoạn văn bản tiếng Việt. Chúng tôi sẽ giới hạn loại hình văn bản chủ yếu là các bình luận trên mạng xã hội và đánh giá trực tuyến.
* **Giới hạn về thời gian:** Thời gian thực hiện nghiên cứu này dự kiến từ [1/3/2025] đến [**30/5/2025**]. Quá trình này bao gồm các giai đoạn: thu thập và tiền xử lý dữ liệu, xây dựng và huấn luyện mô hình, đánh giá kết quả và viết báo cáo.
* **Giới hạn về nội dung:** Nghiên cứu này sẽ tập trung vào việc phân loại cảm xúc của văn bản vào ba nhóm chính:  
  + **Tích cực:** Thể hiện sự hài lòng, yêu thích, đồng tình, vui vẻ, v.v.
  + **Tiêu cực:** Thể hiện sự không hài lòng, ghét bỏ, phản đối, buồn bã, v.v.
  + **Trung tính:** Thể hiện ý kiến khách quan, thông tin đơn thuần, không rõ ràng về cảm xúc.
* Chúng tôi sẽ sử dụng một mô hình học máy cơ bản, ví dụ như **[Lặp lại mô hình bạn đã chọn hoặc một trong các mô hình bạn cân nhắc, ví dụ: Naive Bayes]**, để thực hiện việc phân loại này. Các khía cạnh nâng cao như xử lý đa ngôn ngữ, nhận diện các sắc thái cảm xúc phức tạp (ví dụ: mỉa mai, сар сар), hoặc sử dụng các mô hình học sâu phức tạp sẽ nằm ngoài phạm vi của nghiên cứu này.

# **6. Vấn đề nghiên cứu:**

Trong bối cảnh số hóa và sự phát triển mạnh mẽ của các nền tảng trực tuyến tại Việt Nam, lượng lớn dữ liệu văn bản tiếng Việt chứa đựng ý kiến và cảm xúc của người dùng được tạo ra hàng ngày. Việc tự động phân tích và hiểu được những cảm xúc này có ý nghĩa quan trọng đối với nhiều ứng dụng, từ việc theo dõi phản hồi của khách hàng, đánh giá hiệu quả các chiến dịch truyền thông, đến việc phát hiện các xu hướng dư luận xã hội.

Tuy nhiên, việc nhận diện cảm xúc trong văn bản tiếng Việt mang những thách thức riêng do đặc điểm phức tạp của ngôn ngữ, bao gồm sự đa dạng trong cách diễn đạt, sử dụng từ lóng, tiếng địa phương, và các biểu tượng cảm xúc. Mặc dù đã có nhiều nghiên cứu về nhận diện cảm xúc trên các ngôn ngữ khác, việc áp dụng và đánh giá hiệu quả của các mô hình học máy cơ bản trên dữ liệu tiếng Việt vẫn là một lĩnh vực cần được khám phá.

Do đó, vấn đề nghiên cứu trung tâm mà đề tài này tập trung giải quyết là:

"Trong bối cảnh đặc trưng của tiếng Việt, liệu một mô hình học máy cơ bản (ví dụ: Naive Bayes hoặc Logistic Regression) có thể được xây dựng và ứng dụng hiệu quả như thế nào để tự động nhận diện các loại cảm xúc cơ bản (tích cực, tiêu cực, trung tính) trong các đoạn văn bản ngắn như bình luận trực tuyến?"

# **7. Giả thuyết nghiên cứu:**

Dựa trên tổng quan các nghiên cứu đã được công bố về lĩnh vực nhận diện cảm xúc trong văn bản và xem xét những đặc trưng riêng biệt của tiếng Việt, chúng tôi đưa ra giả thuyết trung tâm của nghiên cứu này như sau:

*"Các mô hình học máy cơ bản, ví dụ như Naive Bayes, Logistic Regression, hoặc Support Vector Machines (SVM), có khả năng đạt được độ chính xác ở mức tương đối, dao động trong khoảng 60% đến 70%, khi thực hiện phân loại các cảm xúc cơ bản (tích cực, tiêu cực, trung tính) trong các đoạn văn bản tiếng Việt ngắn, đặc biệt là các bình luận trực tuyến. Chúng tôi kỳ vọng rằng các mô hình này sẽ có thể học được mối liên hệ giữa sự xuất hiện của một số từ ngữ nhất định và các nhãn cảm xúc tương ứng, dựa trên tần suất và phân bố của chúng trong tập dữ liệu huấn luyện. Tuy nhiên, chúng tôi cũng nhận thức được rằng độ chính xác có thể bị hạn chế bởi nhiều yếu tố, bao gồm sự đa dạng trong cách diễn đạt cảm xúc của người Việt, việc sử dụng ngôn ngữ không chính quy, từ lóng, các hình thức biểu đạt ẩn dụ hoặc сар сар mà các mô hình tuyến tính hoặc dựa trên thống kê đơn giản có thể gặp khó khăn trong việc giải mã. Thêm vào đó, chất lượng và quy mô của tập dữ liệu huấn luyện tiếng Việt có gán nhãn cảm xúc sẽ có tác động trực tiếp đến khả năng khái quát hóa và hiệu suất cuối cùng của mô hình."*

# **8. Nội dung nghiên cứu:**

Để đạt được các mục tiêu đã đề ra và kiểm chứng giả thuyết nghiên cứu, đề tài này sẽ tập trung vào các nội dung chính sau:

1. **Nghiên cứu lý thuyết:**
   * Tìm hiểu sâu về cơ sở lý thuyết của bài toán nhận diện cảm xúc trong văn bản.
   * Nghiên cứu các mô hình học máy cơ bản thường được sử dụng cho tác vụ phân loại văn bản, ví dụ như **Naive Bayes**, **Logistic Regression**, và **Support Vector Machines (SVM)**.
   * Tìm hiểu các kỹ thuật tiền xử lý văn bản cơ bản và phổ biến áp dụng cho ngôn ngữ tiếng Việt, ví dụ như tách từ (word segmentation) sử dụng thư viện **Underscore**, loại bỏ stop words sử dụng danh sách stop words tiếng Việt có sẵn, chuẩn hóa văn bản về chữ thường, và các phương pháp biểu diễn văn bản đơn giản như **Bag-of-Words** (sử dụng CountVectorizer trong Python) và **TF-IDF** (sử dụng TfidfVectorizer trong Python).
   * Nghiên cứu các phương pháp đánh giá hiệu suất mô hình phân loại (ví dụ: độ chính xác - accuracy sử dụng hàm accuracy\_score trong Scikit-learn, độRecall - recall sử dụng recall\_score, độ chính xác - precision sử dụng precision\_score, F1-score sử dụng f1\_score trong Scikit-learn).
2. **Xây dựng tập dữ liệu:**
   * Thu thập dữ liệu văn bản tiếng Việt từ các nguồn trực tuyến (ví dụ: bình luận trên mạng xã hội Facebook, đánh giá sản phẩm trên Shopee).
   * Xây dựng quy trình gán nhãn cảm xúc cho các đoạn văn bản, phân loại chúng vào ba nhóm chính: tích cực, tiêu cực, và trung tính. Quá trình này có thể bao gồm việc gán nhãn thủ công bởi sinh viên thực hiện đề tài hoặc sử dụng các công cụ hỗ trợ gán nhãn.
   * Phân chia tập dữ liệu đã gán nhãn thành tập huấn luyện (training set) và tập kiểm tra (testing set) theo tỷ lệ 80:20 sử dụng hàm train\_test\_split trong Scikit-learn.
3. **Xây dựng và huấn luyện mô hình:**
   * Áp dụng các kỹ thuật tiền xử lý đã nghiên cứu (ví dụ: loại bỏ stop words, chuyển đổi TF-IDF) lên tập dữ liệu huấn luyện.
   * Xây dựng và huấn luyện một trong các mô hình học máy đã lựa chọn, ví dụ: sử dụng lớp NaiveBayes từ thư viện Scikit-learn hoặc lớp LogisticRegression từ Scikit-learn trên tập dữ liệu huấn luyện đã được tiền xử lý.
   * Thực hiện điều chỉnh các tham số của mô hình (nếu cần thiết) bằng cách sử dụng các kỹ thuật như Grid Search với GridSearchCV trong Scikit-learn để tìm ra cấu hình tham số tốt nhất trên một tập kiểm chứng
   * (validation set, nếu được tách ra từ tập huấn luyện).
4. **Đánh giá và phân tích kết quả:**
   * Sử dụng tập dữ liệu kiểm tra để đánh giá hiệu suất của mô hình đã huấn luyện bằng các độ đo đã nghiên cứu (accuracy, precision, recall, F1-score) sử dụng các hàm tương ứng từ Scikit-learn.
   * Phân tích kết quả đánh giá, ví dụ như xem xét ma trận nhầm lẫn (confusion matrix) được tạo bởi hàm confusion\_matrix trong Scikit-learn để hiểu rõ hơn về các loại lỗi mà mô hình mắc phải.
   * Thực hiện phân tích lỗi bằng cách xem xét các mẫu văn bản mà mô hình dự đoán sai để xác định các yếu tố ngôn ngữ có thể gây nhầm lẫn.
5. **Viết báo cáo:**
   * Tổng hợp toàn bộ quá trình nghiên cứu, kết quả đạt được, và các phân tích, thảo luận liên quan thành một báo cáo khoa học chi tiết.
   * Đưa ra kết luận về khả năng của mô hình (hoặc các mô hình) trong việc nhận diện cảm xúc tiếng Việt và đề xuất các hướng nghiên cứu tiếp theo, ví dụ như thử nghiệm với các mô hình học sâu đơn giản hoặc sử dụng các kỹ thuật biểu diễn văn bản phức tạp hơn như word embeddings.

# **9. Phương pháp chứng minh giả thuyết:**

Để kiểm chứng giả thuyết đã được đề xuất, chúng tôi sẽ tiến hành nghiên cứu theo các bước sau:

1. **Thu thập và chuẩn bị dữ liệu**: Chúng tôi dự kiến thu thập một tập dữ liệu gồm khoảng 800 - 1200 bình luận tiếng Việt từ các nền tảng mạng xã hội khác nhau. Quá trình chuẩn bị dữ liệu bao gồm việc gán nhãn cảm xúc cho mỗi bình luận vào một trong ba类别: tích cực, tiêu cực, hoặc trung tính. Tập dữ liệu sau đó sẽ được chia thành tập huấn luyện (khoảng 70%) và tập kiểm tra (khoảng 30%).
2. **Thử nghiệm với các mô hình học máy cơ bản:** Chúng tôi sẽ tập trung vào việc thử nghiệm với các mô hình học máy cơ bản sau:  
   * Naive Bayes: Đây là một mô hình phân loại xác suất dựa trên định lý Bayes, thường hiệu quả trong các tác vụ phân loại văn bản. Chúng tôi kỳ vọng mô hình này có thể đạt được độ chính xác trong khoảng 60% - 65% trên tập dữ liệu của nghiên cứu.
   * Logistic Regression: Là một mô hình hồi quy tuyến tính được sử dụng cho các bài toán phân loại nhị phân và đa lớp. Chúng tôi kỳ vọng Logistic Regression có thể mang lại độ chính xác cao hơn, ước tính trong khoảng 65% - 70%.
3. **Huấn luyện và đánh giá mô hình:** Các mô hình đã chọn sẽ được huấn luyện trên tập dữ liệu huấn luyện đã chuẩn bị. Sau quá trình huấn luyện, hiệu suất của mô hình sẽ được đánh giá trên tập dữ liệu kiểm tra bằng các chỉ số đo lường phổ biến trong bài toán phân loại, bao gồm độ chính xác (accuracy), cũng như các chỉ số khác như precision, recall, và F1-score để có cái nhìn toàn diện hơn về khả năng của mô hình.
4. **Phân tích kết quả:** Chúng tôi sẽ tiến hành phân tích các kết quả đánh giá để xác định hiệu suất của từng mô hình trên tập dữ liệu tiếng Việt. Việc so sánh hiệu suất giữa các mô hình (nếu thử nghiệm nhiều hơn một) sẽ được thực hiện. Đồng thời, chúng tôi sẽ phân tích các trường hợp dự đoán sai để xác định các yếu tố ngôn ngữ có thể gây ra lỗi và đánh giá mức độ phù hợp của các mô hình cơ bản cho bài toán nhận diện cảm xúc tiếng Việt. Cuối cùng, kết quả đạt được sẽ được đối chiếu với giả thuyết ban đầu của nghiên cứu.