**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**PHÂN HIỆU TẠI TP. HỒ CHÍ MINH**

**BỘ MÔN CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO BÀI TẬP LỚN**

**MÔN: HỌC MÁY CƠ BẢN**

**ĐỀ TÀI: DỰ ĐOÁN KẾT QUẢ TRẬN ĐẤU BÓNG ĐÁ**

Giảng viên hướng dẫn:TS. LÊ NGỌC HIẾU

Sinh viên thực hiện:  Cao Hoàng Gia Khang

Phạm Phú Tân

Lớp :  CQ.62.CNTT

Khoá :**62**

Tp. Hồ Chí Minh, năm 2024

**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**PHÂN HIỆU TẠI TP. HỒ CHÍ MINH**

**BỘ MÔN CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO BÀI TẬP LỚN**

**MÔN: HỌC MÁY CƠ BẢN**

**ĐỀ TÀI: DỰ ĐOÁN KẾT QUẢ TRẬN ĐẤU BÓNG ĐÁ**

Giảng viên hướng dẫn:TS. LÊ NGỌC HIẾU

Sinh viên thực hiện:  Cao Hoàng Gia Khang

Phạm Phú Tân

Lớp :  CQ.62.CNTT

Khoá :**62**

Tp. Hồ Chí Minh, năm 2024

**LỜI CẢM ƠN**

Lời đầu tiên nhóm em xin gửi lời cảm ơn chân thành đến thầy Lê Ngọc Hiếu giảng viên Bộ môn Công nghệ thông tin – Phân hiệu Trường Đại học Giao thông vận tải. Người đã truyền dạy, trang bị cho nhóm em nhiều kiến thức để hoàn thành bài tập cho môn học Học máy cơ bản.

Nhờ các kiến thức thầy đã giảng dạy, truyền đạt đã giúp nhóm em hoàn thành bài tập lớn với đề tài “Mô hình dự đoán kết quả trận đấu bóng đá”. Tuy đã cố gắng trong quá trình nghiên cứu tìm hiểu nhưng do kiến thức còn hạn chế nên vẫn còn tồn tại nhiều thiếu sót. Vì vậy, em rất mong nhận được sự đóng góp ý kiến từ thầy để đề tài của nhóm em hoàn thiện hơn.

Sau cùng, chúng em xin kính chúc thầy lời chúc sức khỏe, luôn hạnh phúc và thành công hơn nữa trong công việc cũng như trong cuộc sống.

Em xin chân thành cảm ơn!

**NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**Tp. Hồ Chí Minh, ngày ….… tháng ….…  năm ….…**

**Giảng viên hướng dẫn**

    **TS. Lê Ngọc Hiếu**

Mục lục

[Abstract 3](#_Toc185195858)

[Danh mục 4](#_Toc185195859)

[**Từ viết tắt:** 4](#_Toc185195860)

[**Bảng:** 4](#_Toc185195861)

[**Hình:** 4](#_Toc185195862)

[Chương 1: Tổng quan đề tài nghiên cứu 5](#_Toc185195863)

[**1.1** **Giới thiệu đề tài** 5](#_Toc185195864)

[**1.2** **Lý do chọn đề tài** 5](#_Toc185195865)

[**1.3** **Mục tiêu nghiên cứu** 5](#_Toc185195866)

[**1.4** **Phạm vi nghiên cứu** 5](#_Toc185195867)

[**1.5** **Giới thiệu nhóm tác giả** 5](#_Toc185195868)

[**1.6** **Phương pháp nghiên cứu** 6](#_Toc185195869)

[Chương 2: Tổng quan lý thuyết 7](#_Toc185195870)

[**2.1** **Các thuật toán được sử dụng trong bài** 7](#_Toc185195871)

[2.1.1 Random Forest: 7](#_Toc185195872)

[2.1.2 Naïve Bayes: 9](#_Toc185195873)

[2.1.3 Logistic Regression: 9](#_Toc185195874)

[2.1.4 Support Vector Machine (SVM): 9](#_Toc185195875)

[2.1.5 Convolution Neural Network (CNN): 9](#_Toc185195876)

[2.1.6 Recurrent Neural Network (RNN): 9](#_Toc185195877)

[2.1.7 Long Term – Short Term (LSTM): 9](#_Toc185195878)

[**2.2** **Các nghiên cứu trước đây liên quan đến chủ đề này** 9](#_Toc185195879)

[Chương 3: Giới thiệu bộ dữ liệu 10](#_Toc185195880)

[Chương 4: Thực nghiệm và kết quả thu được 11](#_Toc185195881)

[**4.1** **Trình bày kết quả đạt được** 11](#_Toc185195882)

[**4.2** **Đánh giá hiệu suất của thuật toán** 11](#_Toc185195883)

[**4.3** **Đánh giá độ chính xác của thuật toán** 11](#_Toc185195884)

[Chương 5: Giải thích và bàn luận 12](#_Toc185195885)

[**5.1** **Ảnh hưởng của tiền xử lí dữ liệu** 12](#_Toc185195886)

[**5.2** **Phân tích kết quả mô hình** 12](#_Toc185195887)

[**5.3** **Hạn chế của nghiên cứu** 12](#_Toc185195888)

[**5.4** **Khả năng mở rộng và cải thiện** 12](#_Toc185195889)

[**5.5** **So sánh với các nghiên cứu khác** 12](#_Toc185195890)

[Chương 6: Kết luận 13](#_Toc185195891)

[Tài liệu tham khảo 14](#_Toc185195892)

# **Abstract**

Phát hiện tin giả đang trở thành một lĩnh vực nghiên cứu quan trọng do ảnh hưởng của nó đến dư luận và xã hội. Trong khi các phương pháp kiểm tra truyền thống như kiểm chứng thủ công gặp rất nhiều vấn đề về nguồn lực và thời gian, thì các công nghệ hiện đại như trí tuệ nhân tạo đang rất phát triển. Vì thế các nghiên cứu về việc áp dụng các phương pháp học máy, học sâu cho phân loại tin giả đang rất được quan tâm. Kết quả thực nghiệm cho thấy độ chính xác cao và hiệu suất tốt hơn so với các phương pháp truyền thống. Nhấn mạnh vai trò của học máy trong việc đối phó với các thông tin sai lệch, đồng thời đưa ra những gợi ý cải tiến các hệ thống phát hiện tin giả tự động trong tương lai.

Keywords – *Recurrent Neural Network (RNN), Long Term – Short Term (LSTM), Naïve Bayes, Natural Language Processing, Support Vector Machine (SVM), Convolution Neural Network NLP (CNN-NLP), Random Forest, Logistic Regression, Supervised Learning.*

# **Danh mục**

## **Từ viết tắt:**

ML: Machine Learning

DL: Deep Learning

NLP: Natural Language Processing

AI: Artificial Intelligence

LSTM: Long Term – Short Term

RNN: Recurrent Neural Network

SVM: Support Vector Machine

## **Bảng:**

## **Hình:**

# **Chương 1: Tổng quan đề tài nghiên cứu**

* 1. **Giới thiệu đề tài**
* Bằng các phương pháp học máy – học sâu cho bộ dữ liệu phân loại tin giả, qua đó tạo ra các mô hình có thể áp dụng để tự động phân biệt tin giả.
  1. **Lý do chọn đề tài**
* Tin giả là một trong những vấn đề nhức nhối hiện nay, ảnh hưởng đến toàn cầu. Với các phương pháp kiểm tra truyền thống như Fact-Checking (kiểm tra xác thực), Rumor Detection (Phát hiện tin đồn), … Tiêu tốn quá nhiều tài nguyên về nguồn lực và thời gian.
* Cùng với sự phát triển của các công nghệ hiện đại như trí tuệ nhân tạo (AI) hoặc các phương pháp như học máy (ML) hay học sâu (DL), ta có thể tự động việc phát hiện tin giả, giúp giảm đi nguồn lực và thời gian.
* Mặc khác, xử lí ngôn ngữ tự nhiên (NLP) đang rất phổ biến nên có khá nhiều tài liệu, phương pháp có thể ứng dụng.
  1. **Mục tiêu nghiên cứu**
* Thông qua quá trình áp dụng các phương pháp học máy – học sâu cho bộ dữ liệu NLP, tạo ra mô hình có thể áp dụng việc phân loại và dự đoán tin giả trên mạng xã hội. Qua đó giảm số lượng tin giả tiếp cận đến người đọc.
* Đánh giá mô hình cho bộ dữ liệu phân loại tin giả nói chung và bộ dữ liệu NLP nói riêng.
* Phát hiện hạn chế của mô hình, tìm cách cải tiến trong tương lai.
* Xem xét tính thực tiễn của mô hình.
  1. **Phạm vi nghiên cứu**
* Nghiên cứu trong 7 mô hình gồm 4 mô hình học máy và 3 mô hình học sâu.
* Bộ dữ liệu NLP, phân loại tin giả với hơn 21000 mẫu.
* Các thuật toán được áp dụng là các thuật toán thiên về phân loại nhị phân hoặc thiên về xử lí ngôn ngữ tự nhiên, xử lí tuần tự.
  1. **Giới thiệu nhóm tác giả**
* Cao Hoàng Gia Khang:
  + Sinh viên trường Đại học Giao Thông Vận Tải Phân Hiệu tại Thành Phố Hồ Chí Minh.
  + Mã số sinh viên: 6251071045.
  + Mã lớp: CQ.62.CNTT .
  + Chịu trách nhiệm nghiên cứu, ứng dụng thuật toán, tiền xử lí dữ liệu cho mô hình Naïve Bayes, RNN, LSTM.
* Phạm Phú Tân:
  + Sinh viên trường Đại học Giao Thông Vận Tải Phân Hiệu tại Thành Phố Hồ Chí Minh.
  + Mã số sinh viên: 6251071088.
  + Mã lớp: CQ.62.CNTT .
  + Chịu trách nhiệm nghiên cứu, ứng dụng thuật toán, tiền xử lí dữ liệu cho mô hình Random Forest, SVM, Logistic Regression, CNN.
  1. **Phương pháp nghiên cứu**
* Phương pháp chọn thuật toán:

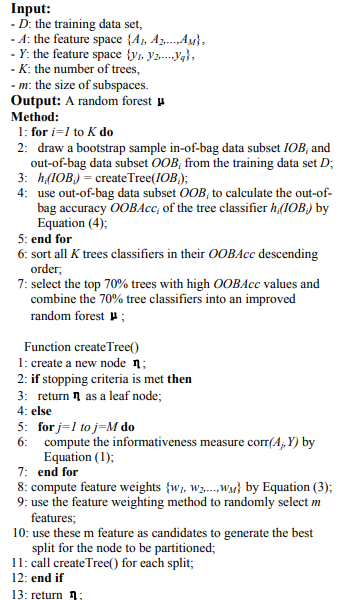
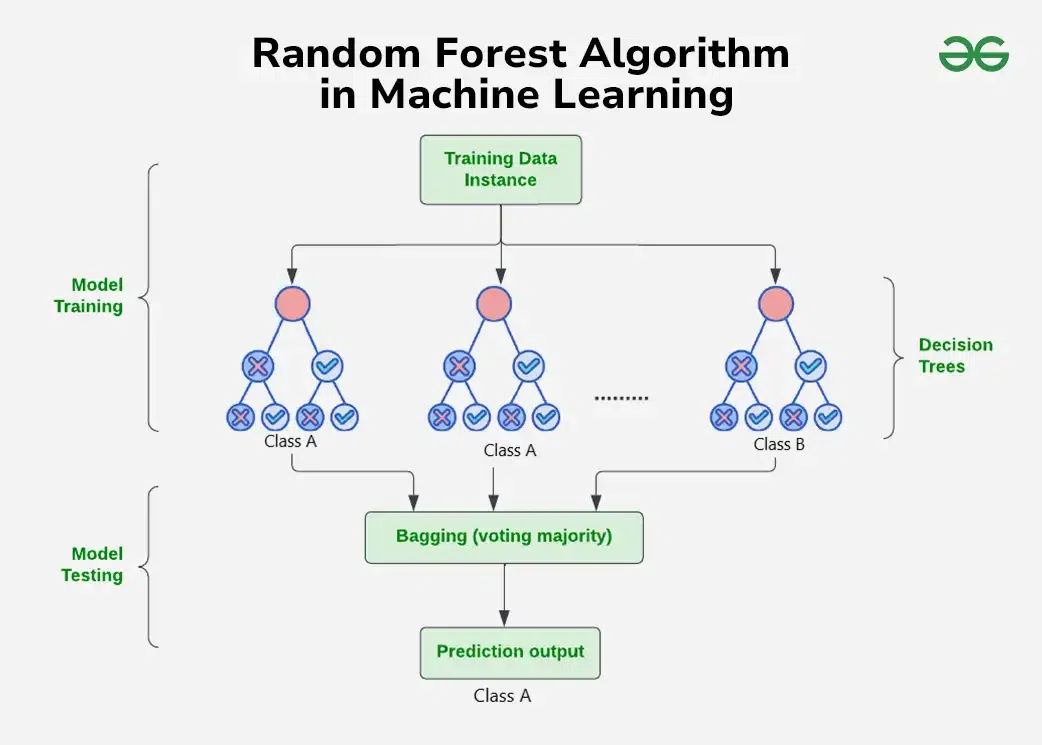
1. Chọn thuật toán phù hợp với bộ dữ liệu NLP: để mô hình có thể tận dụng toàn bộ khả năng mà dữ liệu có thể đem lại, thuật toán được lựa chọn phải có những đặc tính phù hợp như xử lí tuần tự, trích xuất đặc trưng, …
2. Chọn thuật toán phù hợp với mục tiêu của bộ dữ liệu: bộ dữ liệu có mục tiêu là phân loại đặc biệt là phân loại nhị phân, nên chọn những thuật toán phù hợp với phân loại đa lớp hay phân loại nhị phân.

* Phương pháp tiền xử lí dữ liệu:
* Đối với các mô hình học máy truyền thống, sử dụng các phương pháp vectorize như TF – IDF, One – Hot, Embedding, … để chuẩn hóa dữ liệu.
* Đối với các mô hình học sâu, sử dụng các phương pháp Tokenization như Word – based Tokenization để tiền xử lí dữ liệu.
* Phương pháp huấn luyện:
* Khởi tạo mô hình: mỗi mô hình sẽ được khởi tạo với một giá trị phù hợp với mô hình (thông thường giá trị khởi tạo bằng 0).
* Các mô hình học máy sẽ học tập dựa trên việc huấn luyện các vector đặc trưng là dữ liệu được vectorize sau tiền xử lí.
* Các mô hình học sâu sẽ học tập dựa trên quá trình feed forward, backward pass, cập nhật trọng số nhằm nắm bắt ngữ cảnh, ý nghĩa của từ.
* Kỹ thuật tối ưu:
* Khởi tạo trọng số: một số mô hình có cách khởi tạo trọng số giúp tối ưu mô hình như Glorot/Xavier (RNN), He, …
* Chuẩn hóa dữ liệu đầu vào, loại bỏ một số nơ-ron không cần thiết, …
* Early Stopping: một cách dừng sớm khi mô hình không thể học được nữa.[1]
* Model Parameter Optimization: Dropout [2]
* Đánh giá mô hình:
* Sau khi huấn luyện, mô hình sẽ được đánh giá dựa trên các thông số như accuracy, precision, recall, …
* Đánh giá mô hình dựa trên quá trình huấn luyện, vẽ các plot theo dõi quá trình huấn luyện.

# **Chương 2: Tổng quan lý thuyết**

## **Các thuật toán được sử dụng trong bài**

1. Random Forest:
2. Tổng quan mô hình:

* Quá trình Bootstrap Samping tạo ra các tập dữ liệu con ngẫu nhiên.
* Mỗi tập dữ liệu con tạo ra một cây với các đặc trưng ngẫu nhiên (một vài đặc trưng không phải toàn bộ).
* Mỗi cây huấn luyện trên tập dữ liệu con.

Random Forest 1 Model

Random Forest 2 Algorithm

* Sắp xếp tất cả các cây và chọn 70% cây với điểm out – of – bag accuracy (OOBAcc) cao (vì đây là bài toán phân loại).
* Kết hợp các cây đó và tạo ra mô hình Random Forest hoàn chỉnh. [3][4][5]

1. Quá trình xây dựng và huấn luyện mô hình hoàn chỉnh:

* Tiền xử lí dữ liệu: Dữ liệu được tiền xử lí và vectorize theo phương pháp TF – IDF.
* Bootstrap Sampling tạo ra các tập dữ liệu con ngẫu nhiên.
* Tạo ra các cây dựa trên các bộ dữ liệu con chứa đặc trưng ngẫu nhiên (lưu ý không trùng nhau).
* Mô hình huấn luyện dựa trên bộ dữ liệu, dùng cơ chế voting để tạo rừng hoàn chỉnh.
* Đánh giá mô hình.

1. Ưu điểm và nhược điểm:
   * + Ưu điểm:
   * Hiệu suất cao, ổn định.
   * Ít bị overfitting.
   * Xử lí được dữ liệu nhiều chiều.
   * Đánh giá tầm quan trọng các đặc trưng.
   * Xử lí tốt các giá trị nhiễu và thiếu
     + Nhược điểm:
   * Tốn tài nguyên tính toán.
   * Thời gian dự đoán chậm với dữ liệu lớn.
2. Tham số tối ưu:

* N\_Limit: giới hạn cây trong rừng.
* Max\_dept: giới hạn độ sâu của một cây, tránh bị overfitting và tràn tài nguyên nếu dữ liệu quá lớn.
* Range sample split: đặt khoảng giới hạn chia dữ liệu để tạo cây.

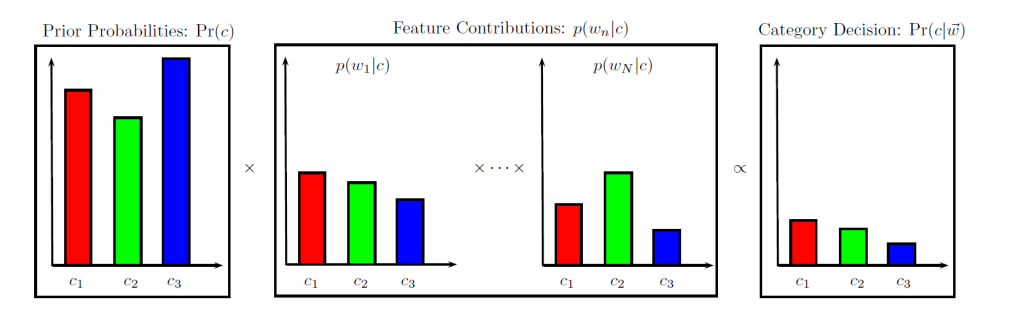
1. Naïve Bayes:
2. Tổng quan mô hình:

* Mô hình được phát triển dựa trên định lý Naïve Bayes.
* Nguyên lý cơ bản của mô hình:
  + - Định lý Bayes:

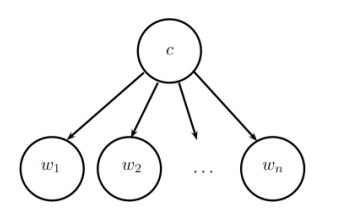
P(y|x) – xác xuất điểm x nằm trong lớp y.

* + - Giả định độc lập:

P(x1, x2, …, xn) = P(x1|y)\*P(x2|y)\*…\*P(xn|y) – giả định các đặc trưng độc lập với nhau. [7][8]

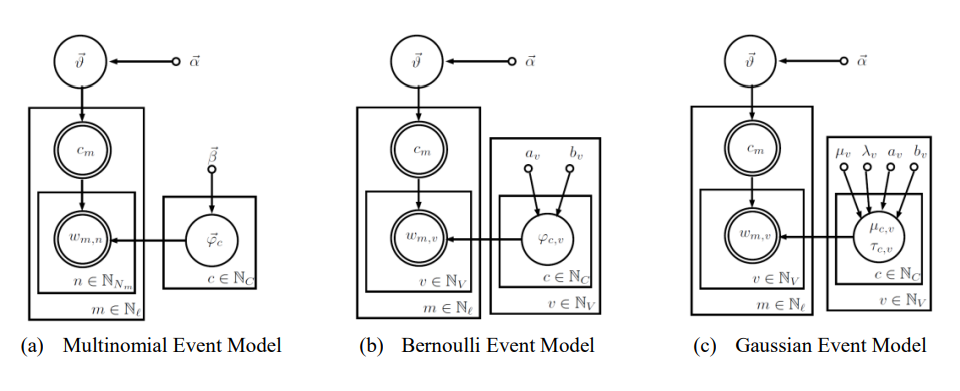


Naive Bayes 1 Minh họa giả định độc lập

* Thực tế, có thể xem phân loại Naïve Bayes như là một tiến trình sinh. Để tạo ra một tài liệu, phân loại Naïve Bayes chọn danh mục cho nó, sau đó tạo ra từng tính năng độc lập theo phân phối cụ thể theo danh mục.

Naive Bayes Minh họa tiến trình sinh

* **Multinomal Naïve Bayes:** Một phiên bản của Naïve Bayes biểu thị rằng mỗi tài liệu được đại diện bởi một tập các từ. Nghĩa là, thứ tự các từ không quan trọng. Nó tạo ra một túi từ ngữ (bag of words) đại diện cho các tài liệu.
* **Bernoulli Naïve Bayes:** Trong mô hình này, mỗi tài liệu là vector của đặc trưng nhị phân biểu thị khả năng xảy ra (không xảy ra) trong tài liệu.
* **Gaussian Naïve Bayes:** Trong phân loại văn bản, thật bình thường nếu tài liệu được đại diện như là TF x IDF vector. Bởi vì giá trị TF x IDF tăng theo tỉ lệ số lần xuất hiện của từ trong tài liệu, nhưng bị bù trừ bởi tần suất trong tài liệu. Khi làm việc với dữ liệu liên tục (continuous data), một giả định điển hình là các giá trị liên tục liên quan đến mỗi lớp được phân phối theo phân phối Gaussian. Một kỹ thuật phổ biến khác là sử dụng kỹ thuật phân loại để có được một tập hợp mới có các đặc điểm Bernoulli (trên thực tế, việc phân loại có thể loại bỏ một số thông tin phân biệt).



Naive Bayes 3 Event Model

1. Quy trình xây dựng và huấn luyện mô hình:

* Tiền xử lí dữ liệu: dữ liệu được tiển xử lí theo phương pháp TF x IDF, BoW tùy theo dạng bài toán.
* Xây dựng mô hình: lựa chọn mô hình Naïve Bayes muốn xây dựng Multinomal (đa thức), Bernoull (Nhị phân), Gaussian (phân phối Gaussian). [9]
* Huấn luyện và đánh giá mô hình.

1. Ưu điểm và nhược điểm:

* Ưu điểm:
  + Đơn giản, dễ triển khai.
  + Huấn luyện nhanh.
  + Hoạt động tốt với dữ liệu nhiều chiều.
  + Phù hợp dạng bài phân loại cảm xúc và văn bản.
* Nhược điểm:
  + Giả định thường không đúng với thực tế.
  + Có thể kém hiệu quả khi các đặc trưng có tương quan mạnh.
  + Nhạy cảm với dữ liệu nhiễu.

1. Cải tiến mô hình:

* Laplace smoothing: kỹ thuật giải quyết vấn đề xác xuất bằng 0 trong mô hình Naïve Bayes.
* Tiền xử lí dữ liệu: loại bỏ stopwords, nhiễu. Chuẩn hóa text. Stemming/Lemmatization.
* Lựa chọn đặc trưng: Chi-square test, Information Gain, Mutual Information, Feature Engineering, N-gram.
* Balanced Data: oversampling, undersampling, prior propability.
* Cross-validation và tuning: Grid Search, Essemble mô hình khác,…

1. Logistic Regression:
2. Tổng quan mô hình:
3. Quá trình xây dựng và huấn luyện mô hình:
4. Ưu điểm và nhược điểm:
5. Cải tiến mô hình:
6. Support Vector Machine (SVM):
7. Convolution Neural Network (CNN):
8. Recurrent Neural Network (RNN):
9. Long Term – Short Term (LSTM):

## **Các nghiên cứu trước đây liên quan đến chủ đề này**

# **Chương 3: Giới thiệu bộ dữ liệu**

# **Chương 4: Thực nghiệm và kết quả thu được**

1. **Trình bày kết quả đạt được**
2. **Đánh giá hiệu suất của thuật toán**
3. **Đánh giá độ chính xác của thuật toán**

# **Chương 5: Giải thích và bàn luận**

1. **Ảnh hưởng của tiền xử lí dữ liệu**
2. **Phân tích kết quả mô hình**
3. **Hạn chế của nghiên cứu**
4. **Khả năng mở rộng và cải thiện**
5. **So sánh với các nghiên cứu khác**

# **Chương 6: Kết luận**

# **Tài liệu tham khảo**

1. <https://www.bright-journal.org/Journal/index.php/JADS/article/view/312/219> .
2. <https://www.isca-archive.org/interspeech_2017/cheng17_interspeech.pdf> .
3. <https://www.jcomputers.us/vol7/jcp0712-09.pdf> .
4. <https://www.sciencedirect.com/science/article/pii/S1319157822000969> .
5. <https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/> .
6. <https://www.researchgate.net/profile/Sang-Bum-Kim-3/publication/3297622_Some_Effective_Techniques_for_Naive_Bayes_Text_Classification/links/56bc297a08ae7be8798bec38/Some-Effective-Techniques-for-Naive-Bayes-Text-Classification.pdf> .
7. <https://machinelearningcoban.com/2017/08/08/nbc/> .
8. <http://54xushuo.net/wiki/lib/exe/fetch.php?media=xushuo:papers:xu17.pdf> .
9. <https://www.kaggle.com/code/sklasfeld/tensorflow-naive-bayes-classifier> .