TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KÌ**

**XỬ LÝ DỮ LIỆU LỚN**

**Phát hiện tin giả tiếng Việt**

Người hướng dẫn: **TS. BÙI THANH HÙNG**

Người thực hiện: **NGUYỄN HOÀNG MINH THƯ – 518H0061**

**MẠC THUẬN ĐẠT – 518H0606**

Nhóm: **17**

Khóa: **22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2022**

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Chúng tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được thực hiện dưới sự hướng dẫn của TS. Bùi Thanh Hùng. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào chúng tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do chúng tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Nguyễn Hoàng Minh Thư*

*Mạc Thuận Đạt*

# LỜI CẢM ƠN

Chúng em xin chân thành cảm ơn thầy Hùng đã tận tâm giảng dạy, chia sẻ các kiến thức, tài liệu tốt thầy thu thập được và đồng thời cũng đã hỗ trợ chúng em rất nhiều trong quá trình thực hành các bài tập trong bộ môn này.

# PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

TP. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

TP. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

# MỤC LỤC

[LỜI CẢM ƠN 3](#_Toc92319057)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN 4](#_Toc92319058)

[MỤC LỤC 5](#_Toc92319059)

[DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT 6](#_Toc92319060)

[DANH MỤC CÁC HÌNH VẼ 7](#_Toc92319061)

[DANH MỤC CÁC BẢNG 8](#_Toc92319062)

[1. GIỚI THIỆU VÀ PHÂN TÍCH BÀI TOÁN 9](#_Toc92319063)

[1.1. Giới thiệu về bài toán 9](#_Toc92319064)

[1.2. Phân tích yêu cầu của bài toán 10](#_Toc92319065)

[1.2.1. Yêu cầu của bài toán 10](#_Toc92319066)

[1.2.2. Các phương pháp giải quyết bài toán 10](#_Toc92319067)

[1.2.3. Phương pháp đề xuất giải quyết bài toán 10](#_Toc92319068)

[1.3. Phương pháp giải quyết bài toán 11](#_Toc92319069)

[1.3.1. Mô hình tổng quát 11](#_Toc92319070)

[1.3.2. Đặc trưng của mô hình đề xuất 11](#_Toc92319071)

[1.4. Thực nghiệm 12](#_Toc92319072)

[1.4.1. Dữ liệu 12](#_Toc92319073)

[1.4.2. Xử lý dữ liệu 12](#_Toc92319074)

[1.4.3. Công nghệ sử dụng 12](#_Toc92319075)

[1.4.4. Cách đánh giá 12](#_Toc92319076)

[1.5. Kết quả đạt được 12](#_Toc92319077)

[1.6. Kết luận 13](#_Toc92319078)

[TÀI LIỆU THAM KHẢO 14](#_Toc92319079)

[PHỤ LỤC 16](#_Toc92319080)

[HƯỚNG PHÁT TRIỂN 17](#_Toc92319081)

[TỰ ĐÁNH GIÁ 18](#_Toc92319082)

# DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

**CÁC KÝ HIỆU**

*f Tần số của dòng điện và điện áp (Hz)*

*p Mật độ điện tích khối (C/m3)*

**CÁC CHỮ VIẾT TẮT**

TF-IDF Term Frequency – Inverse Document Frequency

SVD Singular Value Decomposition

SVM Support Vector Machines

LSVM Linear Support Vector Machines

KNN K-Nearest Neigbor

DT Decision Tree

SGD Stochastic Gradient Descent

XGB eXtreme Gradient Boosting

LR Logistic Regression

AUC Area Under the Curve

LGBM Light Gradient Boosting Machine

# DANH MỤC CÁC HÌNH VẼ

[Hình 1. Cây cấu trúc HTML **Error! Bookmark not defined.**](#_Toc87559285)

[Hình 2. Biểu đồ thống kê số lượng câu nói nổi tiếng **Error! Bookmark not defined.**](#_Toc87559286)

[Hình 3. Thống kê độ tuổi tác giả **Error! Bookmark not defined.**](#_Toc87559287)

# DANH MỤC CÁC BẢNG

[Bảng 1. Mô tả về bộ dữ liệu **Error! Bookmark not defined.**](#_Toc87559298)

# PHÁT HIỆN TIN GIẢ TIẾNG VIỆT

## Giới thiệu về bài toán

Với các xu hướng công nghệ ngày càng phát triển, các thể loại báo điện tử, tin tức ngày càng được chia sẻ một cách rộng rãi và dễ dàng hơn bao giờ hết thông qua mạng xã hội hay các trang web. Việc phân biệt được giữa một trang báo chính thống và một trang báo “lá cải” đã là một nhiệm vụ không dễ gì đối với người đọc. Nhưng hơn đó, khả năng phân biệt được giữa một tin tức khách quan, báo cáo đúng sự thật và một tin tức giả được dàn dựng để lan truyền thông tin sai sự thật, gây hoang mang cho người đọc là một vấn đề khó hơn rất nhiều.

Do đó, việc phát hiện được tin tức giả là một nhiệm vụ quan trong xã hội ngày nay và đã được nghiên cứu trong nhiều lĩnh vực, chẳng hạn như trong các tài liệu, bài báo khoa học, tin tức và các trang mạng xã hội.

Một ví dụ với Facebook, mạng xã hội với hàng triệu người sử dụng, bất kỳ người dùng nào cũng đều có thể đăng nội dung không được kiểm duyệt lên trang cá nhân của mình và công khai chia sẻ chúng. Như trong thời gian dịch bệnh năm vừa qua tại Việt Nam, đã có vô số các thông tin sai sự thật về dịch bệnh COVID-19 được lan truyền với tốc độ nhanh chóng và rộng rãi đến mức khiến nhiều người tin rằng chính những thông tin giả đó mới là sự thật. Việc này gây ra nhiều sự nhầm lẫn, hoang mang và sai lệch kiến thức cho người đọc.

Vì vậy, một thuật toán có thể tự động nhận diện được thông tin giả chính xác trước khi nó được lan truyền đến người đọc sẽ góp phần rất lớn trong việc xây dựng một không gian mạng lành mạnh cùng với các nội dung xác thực.

## Phân tích yêu cầu của bài toán

### Yêu cầu của bài toán

Bài toán phát hiện tin giả tiếng Việt, ta sẽ cần bộ dữ liệu về tin tức bao gồm nội dung, thông số cùng với nhãn của các tin tức đó. Trong đó, nội dung của tin tức là thông tin chủ yếu cần phải có. Sau đó biến đổi dữ liệu đó thành dạng thích hợp và sử dụng các thuật toán Machine Learning hoặc Deep Learning để phân loại tin tức đó là thật hay giả.

### Các phương pháp giải quyết bài toán

Trong nghiên cứu [1], Admed et al. dùng TF-IDF kết hợp với n-gram để trích xuất đặc trưng của nội dung tin tức. Kết quả của nghiên cứu này cho thấy thuật toán LSVM cho độ chính xác (accuracy) cao nhất trong các mô hình được sử dụng với trích xuất đặc trưng TF-IDF cùng với chuỗi 1-gram và 2-gram.

Trong một nghiên cứu khác [2], Reis et al. sử dụng cách trích xuất đặc trưng về cú pháp, lexical[[1]](#footnote-1), ngữ nghĩa, thái độ, các thông số về tương tác và thông tin của bài viết. Thông qua hai độ đo là AUC và F1, kết quả tốt nhất đạt được từ thuật toán Random Forest và XGB.

Các thuật toán phân loại được sử dụng ở hai nghiên cứu trên là là Support Vector Machines (SVM), Linear Support Vector Machines (LSVM), Logistic Regression (LR), K-Nearest Neigbor (KNN), Decision Tree (DT), Random Forest (RF), Stochastic Gradient Descent (SGD), eXtreme Gradient Boosting (XGB) và Light Gradient Boosting Machine (LGBM).

### Phương pháp đề xuất giải quyết bài toán

Do các phương pháp nói trên đã được thực nghiệm và cho kết quả tốt. Chúng em sẽ thực hiện ba phương pháp trích xuất đặc trưng là:

* TF-IDF kết hợp với n-gram
* Các đặc trưng về cú pháp, lexical và các thông số khác của dữ liệu
* Kết hợp hai đặc trưng trên

Sử dụng các thuật phân loại: LR, KNN, LSVM, DT, RF, XGB, SGD, và LGBM.

## Phương pháp giải quyết bài toán

### Mô hình tổng quát

Vẽ được mô hình tổng quát giải quyết bài toán và trình bày tổng quát về từng phần trong mô hình này

### Đặc trưng của mô hình đề xuất

#### Term Frequency – Inverse Document Frequency (TF-IDF)

#### Singular Value Decomposition (SVD)

#### Logistic Regression (LR)

#### K-Nearest Neigbor (KNN)

#### Linear Support Vector Machines (LSVM)

#### Decision Tree (DT)

#### Stochastic Gradient Descent (SGD)

#### eXtreme Gradient Boosting (XGB)

#### Light Gradient Boosting Machine (LGBM)

Trình bày cụ thể chi tiết từng thành phần trong mô hình đề xuất trên.

## Thực nghiệm

### Dữ liệu

Dữ liệu được lấy ở đâu, từ nguồn nào

Dữ liệu gồm có những gì, mô tả chi tiết về dữ liệu

### Xử lý dữ liệu

Có phải tiền xử lý dữ liệu không? Tại sao

Xử lý dữ liệu như thế nào?

### Công nghệ sử dụng

Ngôn ngữ lập trình sử dụng: Python phiên bản 3.9.7.

(Phần này trình bày không quá 3-6 dòng)

Trình bày vắn tắt về tên của Ngôn ngữ lập trình sử dụng ,

Các thư viện sử dụng để hiện thực bài toán,

Các công cụ sử dụng giải quyết bài toán

### Cách đánh giá

Cách đánh giá bằng độ đo gì, độ đo đó được tính bằng công thức nào?

## Kết quả đạt được

Trình bày các tham số cụ thể hóa các thực nghiệm: ví dụ như số epoch, số tầng, số lớp, tham số x, …

Trình bày các kết quả đạt được theo các độ đo ở trên và trên các bộ dữ liệu hay các phương pháp học máy, học sâu khác nhau.

So sánh với ít nhất 1 phương pháp khác, Vẽ được biểu đồ so sánh giữa các phương pháp theo các độ đo đã trình bày ở trên ví dụ như: Accuracy, MSE, RMSE, MAP, …. (hãy lựa chọn ít nhất 2 độ đo trong các độ đo phổ biến để đánh giá bài toán trên)

Giải thích tại sao đạt được kết quả đó, và kết quả khác nhau giữa các phương pháp trên là nguyên nhân nào?

## Kết luận

Trình bày tóm tắt các kết quả đạt được của bài toán trên,

Trình bày được hạn chế của phương pháp giải quyết bài toán

Trình bày hướng phát triển trong tương lai

# TÀI LIỆU THAM KHẢO

1. Ahmed, H., Traore, I., & Saad, S. (2017). Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, 127–138. doi:10.1007/978-3-319-69155-8\_9

[Link](https://www.researchgate.net/publication/320300831_Detection_of_Online_Fake_News_Using_N-Gram_Analysis_and_Machine_Learning_Techniques)

1. Reis, J. C. S., Correia, A., Murai, F., Veloso, A., Benevenuto, F., & Cambria, E. (2019). Supervised Learning for Fake News Detection. IEEE Intelligent Systems, 34(2), 76–81. doi:10.1109/mis.2019.2899143

[Link](https://www.sentic.net/supervised-learning-for-fake-news-detection.pdf)



**Chú ý các tài liệu tham khảo phải được chỉ mục vào trong các Chương nếu có tham khảo, phải có ít nhất từ 7-15 tài liệu tham khảo chính thống: Sách, bài báo Hội nghị, tạp chí - journal, … và tài liệu này phải được đánh chỉ mục từ nhỏ tới lớn và gắn vào trong báo cáo.**

**Các hình ảnh tham khảo hay các phần tham khảo ít thì có thể đánh thành footnote ngay trong trang tham khảo**

1. Quách Ngọc Ân (1992), “Nhìn lại hai năm phát triển lúa lai”, *Di tuyền học ứng dụng*, 98(1), tr. 10-16.
2. Bộ nông nghiệp & PTNT (1996), *Báo cáo tổng kết 5 năm (1992-1996) phát triển lúa lai,* Hà Nội.
3. Nguyễn Hữu Đống, Đào Thanh Bằng, Lâm Quang Dụ, Phan Đức Trực (1997), *Đột biến –* *Cơ sở lý luận và ứng dụng,* Nhà xuất bản nông nghiệp, Viện khoa học kỹ thuật nông nghiệp Việt Nam, Hà Nội.
4. Nguyễn Thị Gấm (1996), *Phát hiện và đánh giá một số dòng bất dục đực cảm ứng nhiệt* *độ,* Luận văn thạc sĩ khoa học nông nghiệp, Viện khoa học kỹ thuật nông nghiệp Việt Nam, Hà Nội.

……….

1. Anderson J.E. (1985), The Relative Inefficiency of Quota, The Cheese Case, *American* *Economic Review*, 75(1), pp. 178-90.
2. Borkakati R. P.,Virmani S. S. (1997), Genetics of thermosensitive genic male sterility in Rice, *Euphytica* 88, pp. 1-7.
3. Boulding K.E. (1955), *Economics Analysis*, Hamish Hamilton, London.
4. Burton G. W. (1988), “Cytoplasmic male-sterility in pearl millet (penni-setum glaucum L.)”, *Agronomic Journal* 50, pp. 230-231.
5. Central Statistical Oraganisation (1995), *Statistical Year Book*, Beijing.
6. FAO (1971), *Agricultural Commodity Projections (1970-1980)*, Vol. II. Rome.
7. Institute of Economics (1988), *Analysis of Expenditure Pattern of Urban Households in* *Vietnam,* Departement pf Economics, Economic Research Report, Hanoi.

# PHỤ LỤC

Phần này bao gồm những nội dung cần thiết nhằm minh họa hoặc hỗ trợ cho nội dung đồ án như số liệu, biểu mẫu, tranh ảnh. . . . nếu sử dụng những câu trả lời cho một *bảng câu hỏi thì bảng câu hỏi mẫu này phải được đưa vào phần Phụ lục ở dạng nguyên bản* đã dùng để điều tra, thăm dò ý kiến; **không được tóm tắt hoặc sửa đổi**. Các tính toán mẫu trình bày tóm tắt trong các biểu mẫu cũng cần nêu trong Phụ lục của luận văn. Phụ lục không được dày hơn phần chính của đồ án

**MỘT SỐ CHÚ Ý KHI VIẾT BÁO CÁO**

1. Thống nhất kích cỡ chữ, kiểu chữ trong toàn bộ báo cáo. Không tô màu chữ, chỉ dùng màu đen
2. Các công thức phải tự gõ và đánh số theo Chương, ví dụ 1.1, 2.1, 2.2, 2.3
3. Các hình và Bảng phải đánh số theo chương, ví dụ Hình 1.1, Hình 2.1, Bảng 3.1, Bảng 3.2
4. Các hình nếu lấy ở ngoài phải đề footnote chú thích nguồn ở dưới
5. Hình mô hình tổng quát phải tự vẽ bằng Word, không dán hình
6. Các tài liệu tham khảo phải đính vào luận văn theo thứ tự từ nhỏ tới lớn, bắt đầu từ 1, ít nhất phải từ 7-15 tài liệu tham khảo, lựa chọn các tài liệu tham khảo mới
7. Tóm tắt trình bày được các nội dung sau: giới thiệu, phương pháp làm, kết quả, nhận xét (không dùng hình, bảng ở mục này)

PHÁT HIỆN TIN TỨC GIẢ TIẾNG VIỆT

BÀI TOÁN

Phân tích yêu cầu bài toán

Phân tích yêu cầu bài toán: Phân tích được yêu cầu của bài toán là gì (0.5 điểm)

Phương pháp giải quyết bài toán

Phương pháp giải quyết: Trình bày được các phương pháp giải quyết bài toán. Giải thích lý do tại sao chọn phương pháp này, Vẽ được sơ đồ tổng quát giải quyết bài toán (2 điểm)

THU THẬP DỮ LIỆU

XỬ LÝ DỮ LIỆU

MÔ HÌNH SỬ DỤNG

ĐÁNH GIÁ KẾT QUẢ

# HƯỚNG PHÁT TRIỂN

# TỰ ĐÁNH GIÁ

(Với nhóm có 2 thành viên)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Câu | Nội dung | Điểm chuẩn | Tự chấm | Ghi chú |
| 1  (8.5) | I.1. Giới thiệu về bài toán | 0.5 |  |  | |
| I.2. Phân tích yêu cầu của bài toán | 1.0 |  |  | |
| 1.3 Phương pháp giải quyết bài toán | 1.5 |  |  | |
| 1.4 Thực nghiệm | 4.0 |  |  | |
| **1.5 Kết quả đạt được** | 1.0 |  |  | |
| 1.6 Kết luận | 0.5 |  |  | |
| 2 | Điểm nhóm | 0.5 | 0.5 |  | |
| 3 | Báo cáo (chú ý các chú ý 2,3,4,6 ở trang trước, nếu sai sẽ bị trừ điểm nặng) | 1.0 |  |  | |
| Tổng điểm | | |  |  |

TỰ ĐÁNH GIÁ

(Với nhóm có 1 thành viên)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Câu | Nội dung | Điểm chuẩn | Tự chấm | Ghi chú |
| 1  (9 đ) | 1.1 Giới thiệu về bài toán | 0.5 |  |  |
| 1.2 Phân tích yêu cầu của bài toán | 1.0 |  |  |
| 1.3 Phương pháp giải quyết bài toán | 1.5 |  |  |
| 1.4 Thực nghiệm | 4.5 |  |  |
| **1.5 Kết quả đạt được** | 1 |  |  |
| 1.6 Kết luận | 0.5 |  |  |
| 2 | Báo cáo (chú ý các chú ý 2,3,4,6 ở trang trước, nếu sai sẽ bị trừ điểm nặng) | 1đ |  |  |
| Tổng điểm | | |  |  |

1. Lexical: các đặc trưng về từ vựng. Ví dụ như: số lượng từ, từ khác nhau, dấu câu… [↑](#footnote-ref-1)