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](#_Toc92372477)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN 4](#_Toc92372478)

[MỤC LỤC 5](#_Toc92372479)

[DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT 6](#_Toc92372480)

[DANH MỤC CÁC HÌNH VẼ 7](#_Toc92372481)

[DANH MỤC CÁC BIỂU ĐỒ 8](#_Toc92372482)

[1. PHÁT HIỆN TIN GIẢ TIẾNG VIỆT 9](#_Toc92372483)

[1.1. Giới thiệu về bài toán 9](#_Toc92372484)

[1.2. Phân tích yêu cầu của bài toán 9](#_Toc92372485)

[1.2.1. Yêu cầu của bài toán 9](#_Toc92372486)

[1.2.2. Các phương pháp giải quyết bài toán 10](#_Toc92372487)

[1.2.3. Phương pháp đề xuất giải quyết bài toán 10](#_Toc92372488)

[1.3. Phương pháp giải quyết bài toán 11](#_Toc92372489)

[1.3.1. Mô hình tổng quát 11](#_Toc92372490)

[1.3.2. Đặc trưng của mô hình đề xuất 12](#_Toc92372491)

[1.4. Thực nghiệm 13](#_Toc92372492)

[1.4.1. Dữ liệu 13](#_Toc92372493)

[1.4.2. Xử lý dữ liệu 16](#_Toc92372494)

[1.4.3. Công nghệ sử dụng 17](#_Toc92372495)

[1.4.4. Cách đánh giá 17](#_Toc92372496)

[1.5. Kết quả đạt được 18](#_Toc92372497)

[1.6. Kết luận 18](#_Toc92372498)

[TÀI LIỆU THAM KHẢO 20](#_Toc92372499)

[PHỤ LỤC 22](#_Toc92372500)

[HƯỚNG PHÁT TRIỂN 23](#_Toc92372501)

[TỰ ĐÁNH GIÁ 24](#_Toc92372502)

# 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

AUC Area Under ROC Curve

# DANH MỤC CÁC HÌNH VẼ

[Hình 1. Mô hình tổng quát qui trình giải quyết bài toán phát hiện tin giả 11](#_Toc92361894)

[Hình 2. Truy xuất 5 dòng đầu của bộ dữ liệu 13](#_Toc92361895)

[Hình 3. Wordcloud nội dung tin thật 14](#_Toc92361896)

[Hình 4. Wordcloud nội dung tin giả 15](#_Toc92361897)

[Hình 5. Ví dụ trích xuất đặc trưng tự chọn 16](#_Toc92361898)

# DANH MỤC CÁC BIỂU ĐỒ

[Biểu đồ 1. Số lượng nhãn tin thật và giả 13](#_Toc92361902)

[Biểu đồ 2. Thời gian đăng bài của tin thật và giả 14](#_Toc92361903)

# 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 với chuỗi 1-gram và 2-gram 10000 đến 50000 giá trị đặc trưng[[1]](#footnote-1).

Trong một nghiên cứu khác [2], Reis et al. sử dụng cách trích xuất đặc trưng về cú pháp, lexical[[2]](#footnote-2), 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 từ vựng, cú pháp, lexical và các thông số khác
* Kết hợp hai đặc trưng trên

Sử dụng 8 thuật toán phân loại là: LR, KNN, LSVM, DT, RF, XGB, SGD, và LGBM.

## Phương pháp giải quyết bài toán

### Mô hình tổng quát

Dữ liệu

TF-IDF + SVD

Trích xuất đặc trưng tự chọn

Tiền xử lý

(word segment, bỏ dấu câu, bỏ stopwords)

Chuẩn hóa

Đặc trưng kết hợp

Thuật toán phân loại

Dữ liệu mới

Phân loại

Đánh giá

Chọn phương pháp trích xuất đặc trưng và thuật toán tốt nhất

Hình 1. Mô hình tổng quát qui trình giải quyết bài toán phát hiện tin giả

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)

TF-IDF là một thuật toán thường được sử dụng để xử lý dữ liệu văn bản trong các bài toán xử lý ngôn ngữ tự nhiên.

Thông thường, trọng số TF-IDF được tính bởi hai giá trị là tần suất xuất hiện của từ () và tần suất xuất hiện của từ trong bộ văn bản nghịch đảo ().

Trong đó, được tính như sau:

và :

Ký hiệu:

* là từ đang xét
* là một văn bản đang xét
* là số lần xuất hiện của từ đang xét trên toàn bộ các văn bản

#### Singular Value Decomposition (SVD)

Mục đích của SVD được sử dụng trong bài này là để làm giảm chiều dữ liệu của ma trận TF-IDF [3]. Việc phân tích nhân tử ma trận dẫn đến việc phân rã ma trận cho trước thành một tập các giá trị tương quan.

#### Logistic Regression (LR)

#### K-Nearest Neigbor (KNN)

#### Linear Support Vector Machines (LSVM)

#### Decision Tree (DT)

#### Random Forest (RF)

#### 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

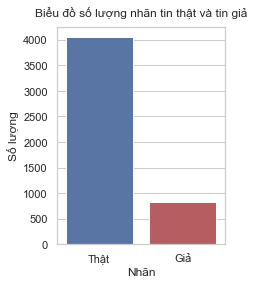
### Dữ liệu

Bài làm sử dụng dữ liệu được chia sẻ qua workshop VLSP2020[[3]](#footnote-3) thông qua cuộc thi ReINTEL[[4]](#footnote-4). Bộ dữ liệu bao gồm 4372 bản ghi là thông tin bài viết được đăng trên các mạng xã hội người Việt thường dùng như Facebook, Zalo và Lotus. Mỗi bản ghi bao gồm thông tin về người đăng, thời gian, nội dung, lượng tương tác của bài viết và nhãn tương ứng.



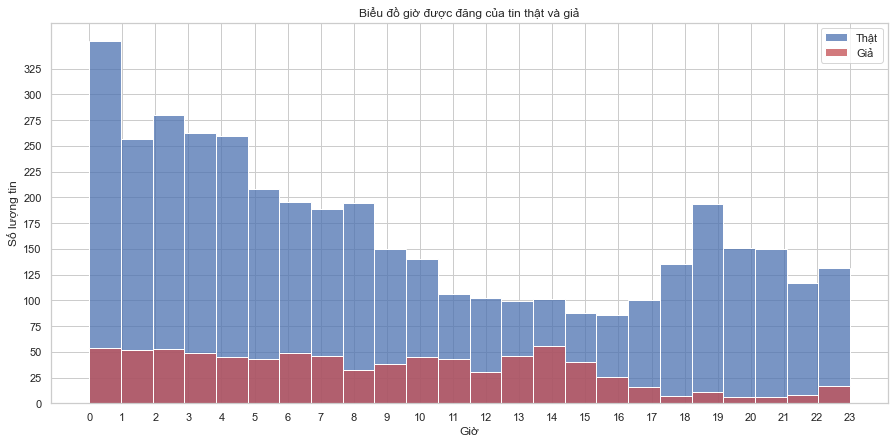
Hình 2. Truy xuất 5 dòng đầu của bộ dữ liệu

Trong số 4868 dữ liệu sau khi được tiền xử lý cơ bản (xem ở phần 1.4.2), có 4050 dữ liệu với nhãn tin thật và 818 dữ liệu nhãn tin giả, dễ dàng thấy là dữ liệu này khá không cân bằng.



Biểu đồ 1. Số lượng nhãn tin thật và giả

Thông qua Biểu đồ 2 dưới, ta thấy được thời gian nhiều tin giả được đăng là từ khoảng 0 giờ đến 15 giờ, còn tin thật thì thường được đăng ở hai khoảng là 0 giờ đến 8 giờ và 19 giờ đến 21 giờ.



Biểu đồ 2. Thời gian đăng bài của tin thật và giả

Trong bài toán phát hiện tin giả thì nội dung của nó là một thông tin quan trọng. Wordcloud sẽ biểu diễn các từ thông dụng được dùng trong nội dung của các dữ liệu. Do dữ liệu được thu thập vào năm 2020, đa số các từ thông dụng được dùng cả trong tin giả và thật đều liên quan đến dịch bệnh Covid-19.



Hình 3. Wordcloud nội dung tin thật



Hình 4. Wordcloud nội dung tin giả

### Xử lý dữ liệu

* Đối với trích xuất đặc trưng TF-IDF

Thực hiện tiền xử lý dữ liệu nội dung của tin bằng cách tokenize[[5]](#footnote-5), word segment[[6]](#footnote-6), bỏ tất cả dấu câu và bỏ các token là stopword[[7]](#footnote-7).

# remove stopwords in news text

def clean\_stopwords(text\_list):

''' sent\_list: already tokenized text in list type

'''

clean\_text = []

for word in text\_list:

if word not in stopwords:

clean\_text.append(word.lower())

return clean\_text

def tokenize(text):

text = text.translate(str.maketrans('', '', string.punctuation)) # clean punctuation

text = word\_tokenize(text.lower()) #tokenize, word segment

return clean\_stopwords(text)

Sử dụng TF-IDF để trích xuất đặc trưng của nội dung tin. Thực hiện giảm chiều dữ liệu với kỹ thuật SVD để giảm khối lượng tính toán. Cuối cùng thực hiện bước chuẩn hóa dữ liệu với MinMaxScaler.

# get tfidf vectors with n\_gram = 2

vectorizer = TfidfVectorizer(tokenizer=tokenize, ngram\_range=(1, 2), max\_features=50000)

X\_tfidf = vectorizer.fit\_transform(df['post\_message'])

# reduce dimensionality of tfidf vector to 100 dimensions as recommended for tfidf

svd = TruncatedSVD(n\_components=100, algorithm='arpack')

X\_tfidf\_svd = svd.fit\_transform(X\_tfidf)

# scale data to (0, 1) range as we've performed dimension reduction

scaler1 = MinMaxScaler()

X1 = scaler1.fit\_transform(X\_tfidf\_svd)

* Đối với trích xuất đặc trưng tự chọn

Chỉ thực hiện bước tiền xử lý là tokenize, word segment và chuẩn hóa dữ liệu. Bài làm trích xuất 19 các đặc trưng là: số lượng từ, dấu câu, token có 1 từ, token có 2 từ, token có >3 từ, chữ số, stopword, từ viết hoa, từ có chứa 1 chữ cái viết hoa, từ khác nhau, hashtag, lượt thích, bình luận, chia sẻ, giờ, thứ, ngày, tháng và số lượng các đường link có tag là url.

# extract custom features

X\_cf = np.array([extract\_feature(df.iloc[i, :]) for i in range(df.shape[0])])

scaler2 = MinMaxScaler()

X2 = scaler2.fit\_transform(X\_cf)

Ví dụ với một tin có nội dung: *Trong giờ học Thể dụ‌c do thầy giáo Nguyễn Văn Quân phụ trác‌h, em D. đã bị thầy Quân “đi đường quyền” lên người dẫn đến việc bị ngấ‌t tại trường và sau đó đã được Ban giám hiệu nhà trường đưa đi cấp cứ‌u để điều trị.*

Các đặc trưng được trích xuất sẽ là:



Hình 5. Ví dụ trích xuất đặc trưng tự chọn

* Đối với đặc trưng kết hợp

Kết hợp dữ liệu trước khi chuẩn hóa của hai đặc trưng kể trên và cũng chuẩn hóa sử dụng một scaler khác.

# combine tf-idf and custom features

X\_cb = np.append(X\_tfidf\_svd, X\_cf, axis=1)

scaler3 = MinMaxScaler()

X3 = scaler3.fit\_transform(X\_cb)

### Công nghệ sử dụng

* Ngôn ngữ lập trình sử dụng: Python phiên bản 3.9.7.
* Các thư viện sử dụng: numpy, pandas, underthesea và sklearn.
* Công cụ sử dụng: Jupyter Notebook.

### Cách đánh giá

Kết quả phân loại sẽ được đánh giá bằng hai độ đo là Accuracy và AUC (Area Under ROC Curve).

Accuracy là giá trị thể hiện độ chính xác với số trường hợp dữ liệu được phân loại đúng nhãn trên tổng số dữ liệu.

AUC có giá trị từ khoảng 0 đến 1 và nó cho biết khả năng của thuật toán có thể phân loại giữa các nhãn. Điểm AUC càng cao thì thuật toán có khả năng phân biệt tốt. Với điểm AUC là 0.5, nó thể hiện rằng thuật toán không thể phân biệt được sự khác nhau giữa hai nhãn. Và khi AUC là 0 thì thuật toán phân biệt tất cả các nhãn thật thành giả hoặc tất cả các nhãn giả thành thật.

## Kết quả đạt được

Sử dụng 3 phương pháp trích xuất đặc trưng và chia mỗi đặc trưng thành 2 tập train và test với tỉ lệ 8/2.

X1\_train, X1\_test, y\_train, y\_test = train\_test\_split(X1, y, test\_size=0.2, random\_state=42)

X2\_train, X2\_test, y\_train, y\_test = train\_test\_split(X2, y, test\_size=0.2, random\_state=42)

X3\_train, X3\_test, y\_train, y\_test = train\_test\_split(X3, y, test\_size=0.2, random\_state=42)

Trình bày các tham số cụ thể shóa 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)

1. Kadhim, Ammar & Cheah, Yu-N & Hieder, Inaam & Ali, Rawaa. (2017). Improving TF-IDF with Singular Value Decomposition (SVD) for Feature Extraction on Twitter. 10.23918/iec2017.16.

[Link](Kadhim,%20Ammar%20&%20Cheah,%20Yu-N%20&%20Hieder,%20Inaam%20&%20Ali,%20Rawaa.%20(2017).%20Improving%20TF-IDF%20with%20Singular%20Value%20Decomposition%20(SVD)%20for%20Feature%20Extraction%20on%20Twitter.%2010.23918/iec2017.16.)

# 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. Max feature. [↑](#footnote-ref-1)
2. Lexical: các đặc trưng về từ vựng. Ví dụ như: số lượng từ, từ khác nhau, dấu câu… [↑](#footnote-ref-2)
3. Vietnamese Language and Speech Processing. [↑](#footnote-ref-3)
4. Reliable Intelligence Identification on Vietnamese SNSs. [↑](#footnote-ref-4)
5. Tách từ. [↑](#footnote-ref-5)
6. Phân đoạn từ. [↑](#footnote-ref-6)
7. Từ dừng: là những từ xuất hiện nhiều trong ngôn ngữ tự nhiên. [↑](#footnote-ref-7)