VIETNAM LABOR UNION GENERAL

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**FINAL PROJECTS INTRODUCE THE MACHINE LEARNING SUBJECT**

**Present machine learning issues by building machine learning models for the classification problem.**

*Instructor:* **LÊ ANH CƯỜNG**

*Moderators:* **NGUYỄN THÀNH KHANG – 518H0372**

**MẠC THUẬN ĐẠT -- 518H**

Class**: 18H50303**

Course**: 22**

**HO CHI MINH CITY, 2020**

VIETNAM LABOR UNION GENERAL

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**FINAL PROJECTS INTRODUCE THE MACHINE LEARNING SUBJECT**

**Present machine learning issues by building machine learning models for the classification problem.**

*Instructor:* **LÊ ANH CƯỜNG**

*Moderators: Nguyễn Thành Khang-518H0372*

*Mạc Thuật Đạt – 518H*

Class**:518H0372**

Course**: 22**

**HO CHI MINH CITY, 2020**

# ACKNOWLEDGMENTS.

I would like to thank you teacher for guiding me during the time project time.

**THE SUBJECTS ARE COMPLETED**

**AT TON DUC THANG UNIVERSITY**

I hereby undertake that this is my own project / our project and under the guidance of Le Anh Cuong.; The research contents and results in this topic are truthful and have not been published in any form before. The data in the tables for analysis, comments and evaluation collected by the author from different sources are clearly stated in the references.

In addition, the project also uses a number of comments, assessments as well as data of other authors, other organizations and organizations with citations and origin notes.

If I find out there is any fraud I take full responsibility for the content of my project. Ton Duc Thang University is not related to the copyright and copyright violations caused by me in the implementation process (if any).

*TP. Ho Chi Minh City, 2020 TP. Ho Chi Minh City, 2020*

*Author. Author.*

*(sign and write full name) (sign and write full name)*

*Nguyen Thanh Tung. Phu phuong Tin.*

# TEACHER'S CONFIRMATION AND REVIEW SECTION

**Verification of the instructor**

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

City. Ho Chi Minh City, date ,month ,year.

(sign and write full name)

**The teacher evaluation section marks**

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

City. Ho Chi Minh City, date ,month ,year.

(sign and write full name)

# SUMMARY

Trình bày tóm tắt vấn đề nghiên cứu, các hướng tiếp cận, cách giải quyết vấn đề và một số kết quả đạt được, những phát hiện cơ bản trong vòng 1 -2 trang.

# TABLE OF CONTENTS

# LIST OF SIGNS AND ABBREVIATIONS

## SIGNAL

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

*p Mật độ điện tích khối (C/m3)*

## ABBREVIATIONS

CSTD Công suất tác dụng

MF Máy phát điện

BER Tỷ lệ bít lỗi

# LIST OF TABLES, FIGURES, GRAPHICS

## LIST OF FIGURES

Hình 2.1: Kiến trúc FTP 1

## LIST OF TABLES

Bảng 3.1 Ví dụ cho chèn bảng 1

# CHAPTER 1: OVERVIEW.

# CHAPTER 2: THEORETICAL BASIS / EXPERIMENTAL RESEARCH.

**1.1 Trình bày công thức toán học**

Table of Contents

**1.2 Trình bày một hình vẽ, sơ đồ**

# CHAPTER 3: RESULT EVALUATION.

**TÀI LIỆU THAM KHẢO**

**Chú ý các tài liệu tham khảo phải được chỉ mục vào trong các Chương nếu có tham khảo, phải có ít nhất từ 10-15 tài liệu tham khảo chính thống: Sách, bài báo Hội nghị, tạp chí - journal, …**

* Quách Ngọc Ân (1992), “Nhìn lại hai năm phát triển lúa lai”, *Di tuyền học ứng dụng*, 98(1), tr. 10-16.
* Bộ nông nghiệp & PTNT (1996), *Báo cáo tổng kết 5 năm (1992-1996) phát triển lúa lai,* Hà Nội.
* 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.
* 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.

……….

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

**PHẦN NỘI DUNG CỤ THỂ THEO MỤC LỤC SAU**

**PHẦN I**

**CHƯƠNG 1**

**PHÂN TÍCH BÀI TOÁN**

**1.1 Giới thiệu về bài toán**

Nêu được bài toán cần giải quyết

**1.2 Phân tích yêu cầu của bài toán**

Trình bày được các yêu cầu của bài toàn

**1.3 Hướng giải quyết**

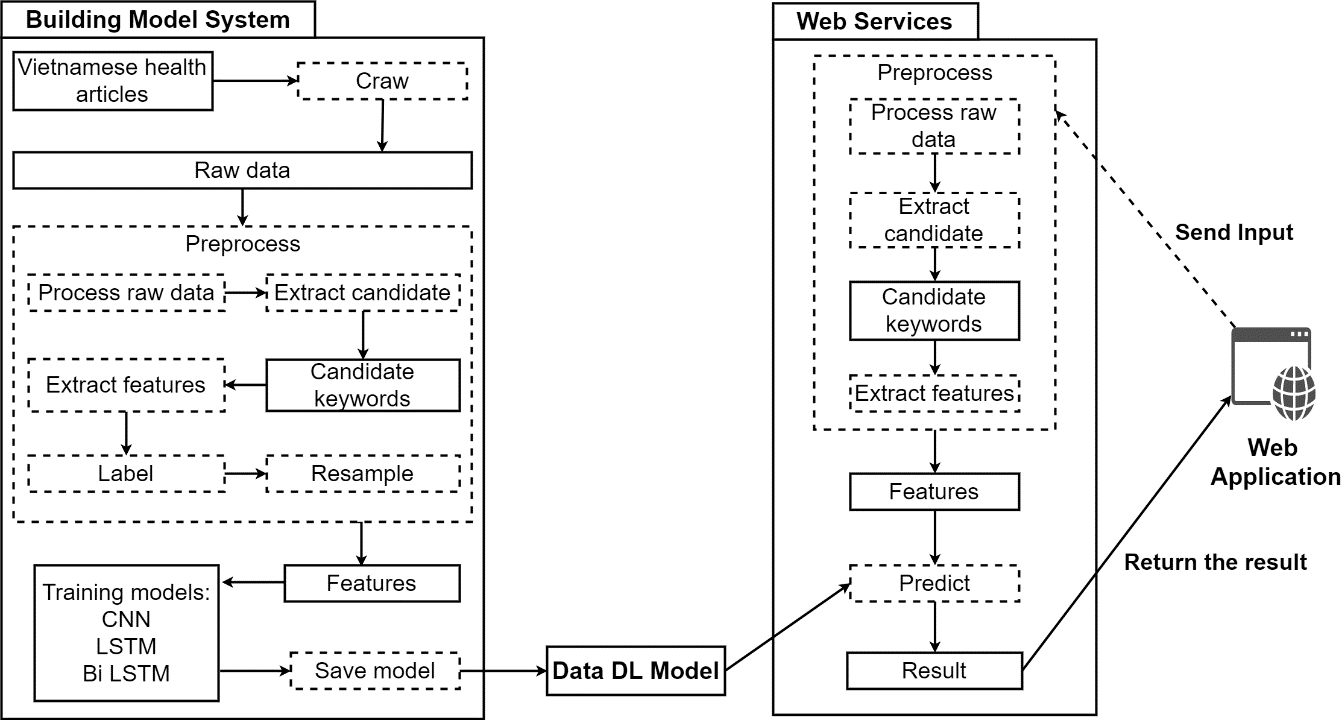
Đưa ra hướng tổng quát giải quyết bài toán

**CHƯƠNG 2**

**2.1 Giới thiệu chung về các phương pháp giải quyết bài toán**

Trong phần này phải vẽ được sơ đồ tổng quát giải quyết bài toán

Ví dụ: Mô hình tổng quát giải quyết bài toán Keyword Extraction



**2.2 Đặc trưng của mô hình đề xuất**

**2.2.1 Feature**

Ví dụ tham khảo của mô hình trên

Our team will extract features for candidate keywords from five concepts:

* Word Embedding with Pre-trained Word2Vec
* Named-Entity Recognition
* Frequency-based
* Position
* Phrase Length and Word Length

**Word Embedding**

Every word has reflected the structure of the word regarding the semantical/morphological/context/hierarchical/etc. information. The idea of Word Embedding is to capture with them as much as possible and convert it to vectors. We applied Word Embedding to represent a keyword – can be combined by 2 or more words – as a vector by plus vectors to each word in the dictionary we build.

Ex: Vec “Bệnh viện Trung Ương” = Vec”Bệnh viện”+ Vec”Trung Ương”

Vec “Đại học FPT thành phố Hồ Chí Minh” = Vec “Đại học” + Vec “FPT” + Vec “thành phố” + Vec “Hồ Chí Minh”

**Named-entity recognition**

In the research time, we realized that a noun phrase such as “Tp. Hồ Chí Minh”, “Đại học FPT”, “Bệnh viện Từ Dũ” can be a keyword. A noun phrase containing a named-entity that is considered a keyword is higher than others.

Number of words in Noun Phrase: nNP

Number of named-entity words in Noun Phrase: nNER

We combine both single value and the following formula to have a single feature value (Fner):

For example: Consider the keyword “Đại học FPT thành phố Hồ Chí Minh”.

Words: “Đại học”, “FPT”, “thành phố”, “Hồ Chí Minh”

nNP = 4

nNER = 2 (“FPT”, “Hồ Chí Minh”)

**Frequency-based**

If a noun phrase is occurring more frequently in a document, the phrase is assumed to more important in the document. [3]

Number of times a noun phrase occurs independently in a document: Phrase Frequency (PF).

Number of times a noun phrase appears in full as a part of other noun phrases: Phrase Link Count (PLC)

Two features, PF and PLC, are combined to have a single feature value using the following measure:

In the above formula, the frequency of a noun phrase (PF) is squared only to give it more important than the phrase link count (PLC). The value 1/2 has been used to moderate the value. We explain below about this formula with an example:

Assume that we have 3 keywords and PF and PLC for 2 keywords:

K1 = “sữa mẹ tốt cho trẻ em”, K2 = “sữa mẹ”, K3 = “sữa”

|  |  |  |  |
| --- | --- | --- | --- |
|  | **PF** | **PLC** | **PF+PLC** |
| **K2** | 10 | 5 | 15 |
| **K3** | 1 | 14 | 15 |

So, for these two cases, the simple addition of PF and PLC do not make any difference in assigning weights to the noun phrases although the independent occurrence of noun phrase K2 is more than that of the noun phrase K3. However, the independent existence of a phrase should get higher importance while deciding whether a noun phrase is keyword worthy or not.

For another more general case about PF and PLC with the same K2 and K3:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **PF** | **PLC** | **PF+PLC** |
| **K2** | 14 | 1 | 15 |
| **K3** | 1 | 19 | 20 |

In this situation, the simple addition of PF and PLC will favor the K3 “sữa”, but our formula will give a higher score to the K2 “sữa mẹ” because it occurs more independently than the K3 “sữa”.

Inverse document frequency (IDF) is a useful measure to determine the commonness of a term in a corpus.

N = total number of documents in a corpus.

df (document frequency) = the number of documents in which a term occurs.

A term with a lower df value means the term is less frequent in the corpus and hence idf value becomes higher. So, if idf value of a term is higher, the term is relatively rare in the corpus. In this way, idf value is a measure for determining the rarity of a term in a corpus.

In TF-IDF, where TF indicates the frequency of a term in a document, TF\*IDF measure favors a relatively rare term which is more frequent in a document. We combine and IDF in the following formula to have a variant of Edmundsonithe an thematic feature:

The value of this feature is normalized by dividing the value by the maximum score in a colletion of scores obtained by the phrases corresponding to a document.

**Position**

If a noun phrase occurs in the title or abstract of a document, it should be given more score. So, we consider the position of the first occurrence of a noun phrase in a document as a feature. Unlike the previous approach[3] that assumes the position of a noun phrase occurred first in the sentence I in the document and computed as a feature, in our work, we assume the position of a noun phrase occurred first in the document as I and computed using the following formula from the previous approach.

For example, assume that we have the sentence “Cúm A/H1N1 trở lại tại thành phố Hồ Chí Minh.” and the keyword “Cúm A/H1N1”:

I of “Cúm A/H1N1” = 1:

By this feature, we are easy to predict a keyword in the title of a document where is near to 1 and in summary of a document where is near to 0.

**Phrase Length and Word Length**

These two features can be considered as the structural features of a noun phrase. Noun phrase length becomes an important feature in the keyword extraction task because the length of the noun phrase usually varies from 1 word to 5 words. We find that noun phrase consisting of 6 or more words are relatively rare in our corpus.

Length of the words in a noun phrase can be considered as a feature. According to Zipf’s Law, shorter words occur more frequently than the larger ones. For example, articles occur more frequently in text. So, the word length can be an indication for the rarity of a word. In the previous approach[3] that considered a length of the longest word in noun phrase as a feature, we realized that a maximum length of single words in Vietnamese is 7 and 80% words in Vietnamese is a compound word, that is not the same as English. We decided to choose the length of characters in the noun phrase as a feature.

Length of a noun phrase: PL

Length of characters in the noun phrase: WL

We combine both single value and the following formula to create a feature[3]

For example, assume that we have the keyword “Đại học FPT thành phố Hồ Chí Minh”:

Keyword after tokenize: “Đại\_học FPT thành\_phố Hồ\_Chí\_Minh”

PL = 4, WL = 33

The value of this feature is normalized by dividing the value by the maximum value of the feature in the collection of phrases corresponding to a document.

**2.2.2 Extracting candidate keywords**

To extract the candidate keywords, we used Part-of-Speech (POS) Tagging and Chunking to extract a noun phrase.

POS Tagging is the process of marking a word in text (corpus) as corresponding to a particular part of speech based on both its definition and its context. A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc. In Vietnamese, we used Pyvi library from python to pos tagging Vietnamese words.

For example: “Hôm nay trời đẹp quá”

“ N N A R “ with N (noun), A(Adjective), R(Adverb)

Chunking or Noun Phrase Chunking (NP-chunking) is a natural language process that separates and segments a sentence into its sub constituents, such as noun, verb, and noun phrase. One of the most useful sources of information for NP-chunking is part-of-speech tags. In order to create a noun phrase, we first defined a chunk grammar, consisting of rules that indicate how sentences should be chunked. In our problem, we defined a grammar with a single regular-expression rule.

(<N.\*>+ <A>\* <E>)? <N.\*>+ <A>\*

<N.\*>+: One or more of any type nouns

<A>\*: Any number of adjectives

<E>: One or zero of preposition

So, this grammar meaning is a noun phrase can be combined by 2 others noun phrase with preposition word between.

In NP-chunking, we used nltk library from python to chunk a sentence. Using this above grammar and nltk chunking, we create a result tree, from which we can extract a noun phrase.

For example:

“Nguyên là một sinh viên giỏi của trường đại học FPT ở thành phố Hồ Chí Minh”

Calling a noun phrase we need is candidate keyword (CK)

Result tree after chunking:

Figure 30: Example of chunking

S(Sentence), V(Verb), Np(Proper Noun), M(Numeral), Ny(Noun abbreviation)

After all, we get 3 noun phrase: “Nguyên”, “sinh viên giỏi của trường đại học FPT”, “thành phố Hồ Chí Minh”. Sometimes, NP-Chunking by nltk get the noun phrase like “anh ấy”, we call this noun phrase is stop-word and remove it out of list noun phrase. In the experimental process, we found that noun phrase consisting of 6 or more words are relatively rare in our corpus. So we filter out all noun phrase that has 6 or more word in extracting process.

**2.2 Phương pháp 1**

**CNN**

Trình bày cụ thể về giải thuật này

**2.3 Phương pháp 2**

LSTM

Trình bày cụ thể về giải thuật này

**2.4 Phương pháp 3** (đây là phương pháp lựa chọn giải quyết bài toán)

Bi-LSTM

Mỗi phương pháp đều trình bày được sơ đồ giải thuật

Trình bày được ưu nhược điểm của mỗi phương pháp

Với phương pháp 3, giải thích lý do tại sao lựa chọn phương pháp này

**CHƯƠNG 3**

**THỰC NGHIỆM**

**3.1 Dữ liệu**

Dữ liệu được lấy ở đâu, từ nguồn nào

Dữ liệu gồm có những gì

**3.2 Xử lý dữ liệu**

Có phải tiền xử lý dữ liệu không? Tại sao

Xử lý dữ liệu như thế nào?

**3.3 Công nghệ sử dụng**

Trình bày vềNgôn ngữ lập trình,

Các thư viện,

Các công cụ

Ví dụ tham khảo:

* Collecting data: Selenium 3.12.0
* Tokenization & POS Tagging: Pyvi 0.0.9.1 (Tran Viet Trung 2016)
* Named-Entity Recognition: underthesea 1.1.8
* Chunking: nltk 3.3
* Word2vector pre-train: streetcodevn (Hung Le 2018)
* Word2vector train: gensim 3.4.0

**3.4 Cách đánh giá**

We evaluate our proposed model base on Precision, Recall and F1-score with test datasets.

All the parameter can be calculated by using leftmost four parameters.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Predict class** | | |
| **Actual class** |  | *Class = Yes* | *Class = No* |
| *Class = Yes* | True Positive | False Positive |
| *Class = No* | False Negative | True Negative |

*True Positives (TP)* - These are the correctly predicted positive values which mean that the value of actual class is yes and the value of predicted class is also yes.

*True Negatives (TN)* - These are the correctly predicted negative values which mean that the value of actual class is no and value of predicted class is also no.

*False Positives (FP)* – When actual class is no and predicted class is yes.

*False Negatives (FN)* – When actual class is yes but predicted class in no.

* **Precision**

Precision is the ratio of correctly predicted positive values to the total predicted positive values. This metric highlights the correct positive predictions out of all the positive predictions. High precision indicates low false positive rate.

* **Recall**

The recall is the ratio of correctly predicted positive values to the actual positive values. Recall highlights the sensitivity of the algorithm, i.e. out of all the actual positives how many were caught by the program.

* **F1- Score**

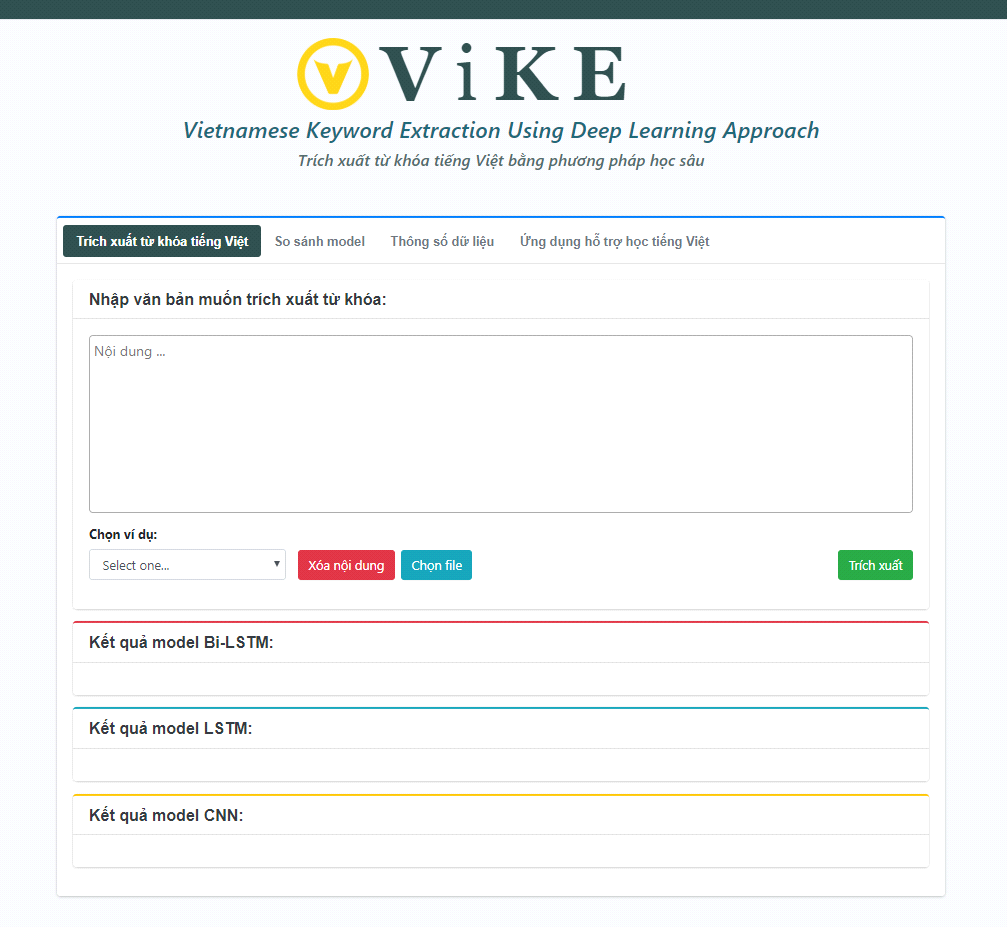
The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is:

**3.5 Kết quả đạt được**

Trình bày các kết quả đạt được

Cụ thể hóa các thực nghiệm: tham số, lớp, … và các kết quả đạt được

So sánh với ít nhất 1 phương pháp khác, Vẽ được biểu đồ so sánh giữa các phương pháp theo các độ đo 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)



**CHƯƠNG 4**

**KẾT LUẬN**

**4.1 Kết luận**

Trình bày tóm tắt các kết quả đạt được

Trình bày được hạn chế của đề tài

**4.2 Hướng phát triển**

Trình bày hướng phát triển trong tương lai

**TỰ CHẤM ĐIỂM**

|  |  |  |  |
| --- | --- | --- | --- |
| Điểm Đồ án | | | |
| Nội dung | Điểm chuẩn | Tự chấm | Ghi chú |
| Chương 1 | 1.5 đ |  |  |
| Chương 2 | 3.0 đ |  |  |
| Chương 3 | 4.0 đ |  |  |
| Chương 4 | 0.5 đ |  |  |
| Tổng điểm | |  |  |
| Điểm nhóm | | | |
| Điểm nhóm | 1đ |  |  |
| Tổng điểm | |  |  |