VIETNAM GENERAL CONFEDERATION OF LABOR

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**NATURAL LANGUAGE PROCESSING**

**MIDTERM PROJECT**

*Instructor*:  ****Assoc.Prof.PhD. Le Anh Cuong****

*Executor*: **NGUYỄN PHẠM QUANG HUY – 522H0138**

**CAO NGUYỄN THÁI THUẬN – 522H0092**

Class **: 22H50302**

Course  **: 26**

**HO CHI MINH CITY, YEAR 2024**

VIETNAM GENERAL CONFEDERATION OF LABOR

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**NATURAL LANGUAGE PROCESSING**

**MIDTERM REPORT**

*Instructor*:  ****Assoc.Prof.PhD. LE ANH CUONG****

*Executor*: **NGUYỄN PHẠM QUANG HUY – 522H0138**

**CAO NGUYỄN THÁI THUẬN – 522H0092**

Class **: 22H50302**

Course  **: 26**

**HO CHI MINH CITY, YEAR 2024**

ACKNOWLEDGEMENTS

"In order to complete this essay, I would like to express my sincere gratitude:

To the Board of Directors of Ton Duc Thang University for providing favorable conditions, a modern educational environment, and a wealth of information that made it easy to research.

I would like to thank the lecturer of the practical subject -  **Assoc.Prof.PhD. Le Anh Cuong** for his support, assistance, and explanations for difficult questions. His detailed and comprehensive lectures helped me quickly absorb the necessary knowledge.

As I do not have much experience in writing essays, there will certainly be shortcomings. I look forward to receiving comments, suggestions, and criticisms to achieve better results in future assignments.

Lastly, I would like to wish all teachers good health and abundant success."

**PROJECT COMPLETED**

**AT TON DUC THANG UNIVERSITY**

I hereby affirm that this is my own project product and was guided by Dr. Nguyen Van A. The research content and results in this topic are honest and have not been published in any form before. The data in the tables serving for analysis, comments, and evaluations were collected by the author from various sources clearly stated in the reference section.

In addition, the project also uses some comments, evaluations as well as data from other authors, other organizations all have citations and source annotations.

If any fraud is detected, I fully take responsibility for the content of my project. Ton Duc Thang University is not related to any copyright or intellectual property violations caused by me during the implementation (if any).

*Ho Chi Minh City, 14th April 2024*

*Author*

*(Sign and clearly state your full name)*

*Nguyễn Phạm Quang Huy*

*Cao Nguyễn Thái Thuận*

**INSTRUCTOR VERIFICATION AND EVALUATION SECTION**

**Confirmation from the instructor**

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

Ho Chi Minh City, dd/mm/yyyy

(Sign and clearly state your full name)

**The teacher's assessment section marks the test**

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

Ho Chi Minh City, dd/mm/yyyy

(Sign and clearly state your full name)

SUMMARY

Learn how to pose problems to solve problems using logical arguments, proficiently use truth tables and use Prolog well in programming.

CATEGORY

[ACKNOWLEDGEMENTS ii](#_Toc387692905)

[INSTRUCTOR VERIFICATION AND EVALUATION SECTION i](#_Toc387692906)v

[SUMMARY v](#_Toc387692907)

[CATEGORY 1](#_Toc387692908)

[LIST OF TABLES, DRAWING, CHARTS 2](#_Toc387692908)

[CHƯƠNG 1 – CATEGORIZE QUESTIONS ABOUT PLAYERS, LEAGUES AND TEAMS IN FOOTBALL............................................................................................ 4](#_Toc387692910)

[1.1 Introduction 4](#_Toc387692911)

[1.2 Research Objectives 4](#_Toc387692916)

[1.3 Benefits of Datasets 4](#_Toc387692916)

[CHƯƠNG 2 – Build synthetic datasets using Large Language Models](#_Toc387692917)

[5](#_Toc387692917)

[2.1 Reasons for using aggregated data 5](#_Toc387692918)

[2.2 Methods for creating synthetic datasets 5](#_Toc387692919)

[2.3 Advantages of synthetic data 6](#_Toc387692919)

[CHƯƠNG 3 – Common methods of text representation](#_Toc387692920) 6

[3.1 Bag of Words (BoW) 6](#_Toc387692921)

[3.2 Naive Bayes 7](#_Toc387692922)

[3.3 Logistic Regression](#_Toc387692923) 9

[3.4 Decision Tree 1](#_Toc387692923)0

[3.5 Using Pretrain Models (BERT/PhoBERT) 1](#_Toc387692923)1

[CHƯƠNG 4 – Comparison of data approaches and machine learning methods for text classification 1](#_Toc387692920)3

#### [4.1 Element 1: Machine learning methods](#_Toc387692922)

[1](#_Toc387692922)3

[4.2 Element 2: Completeness and accuracy of dat 14](#_Toc387692922)

#### [4.3 Element 3: Text representation](#_Toc387692922)

[1](#_Toc387692922)4

[BIBLIOGRAPHY](#_Toc387692920) 16

LIST OF TABLES, DRAWINGS, CHARTS

**LIST OF IMAGE**

[Image 3.2.1: Working Principle of Navie Bayes 8](#_Toc387689394)

[Image 3.2.2: Formula of feature independence of navie Bayes 8](#_Toc387689394)

[Image 3.2.3: Formula of highest posterior probability of Navie Bayes](#_Toc387689394) 8

[Image 3.3.1: Working Principle of Logistic Regression 9](#_Toc387689394)

[Image 3.3.2: Formula sigmoid of Logistic Regression 9](#_Toc387689394)

[Image 3.3.3: Formula predicted class of Logistic Regression](#_Toc387689394)

[1](#_Toc387689394)0

[Image 3.4.1: Formula Entropy of Decision Tree 1](#_Toc387689394)1

[Image 3.4.2:Formula information gain of Decision Tree 1](#_Toc387689394)1

[Image 3.5.1: Formula of Pretrained Models Doc2Vec 1](#_Toc387689394)2

**LIST OF TABLES**

[Table 4.1.1: Table Machine learning methods 1](#_Toc387689363)3

[Table 4.2.1: Table Completeness and accuracy](#_Toc387689363)

[1](#_Toc387689363)4

[Table 4.3.1: :Table Text representation 1](#_Toc387689363)4

CHAPTER 1 - CATEGORIZE QUESTIONS ABOUT PLAYERS, LEAGUES AND TEAMS IN FOOTBALL

1.1 Introduction

In the field of football, automated question-answering systems and sports chatbots are increasingly being widely applied. One of the critical challenges in this domain is classifying questions into categories such as **players**, **tournaments**, and **teams**. Automatically categorizing questions enhances information retrieval efficiency, improves user experience, and optimizes virtual assistant systems.

This report presents research on the text classification problem in football, employing machine learning methods to automatically categorize questions. It analyzes the construction of appropriate datasets and compares the performance of different approaches on these datasets to identify the most optimal method.

**1.2 Research Objectives**

**The main objectives of this study include:**

1. **Building a dataset** of questions related to **players**, **tournaments**, and **football teams**.
2. **Applying machine learning methods** to classify questions into predefined categories.
3. **Evaluating the performance** of each method and **proposing an optimal solution**.

1.3 Benefits of Datasets

Building a suitable dataset for classifying football-related questions offers several key benefits:

- Optimizing search systems: Accurately categorized questions allow systems to quickly provide relevant answers, enabling users to find information more easily.

- Improving sports chatbots: Chatbots can better understand question contexts, deliver more precise responses, and enhance user experience.

- Supporting AI model training: The dataset helps train machine learning models to accurately identify question topics, which can then be applied to other applications like virtual assistants or sports analytics.

- Developing sports data analytics applications: AI systems can leverage this data to track search trends, predict popular topics, and generate insights within the football community.

CHAPTER 2 - Build synthetic datasets using Large Language Models

**2.1 Reasons for using aggregated data**

In many cases, collecting real-world data may encounter challenges such as restricted access, a lack of labeled data, or imbalanced categories. Therefore, leveraging synthetic data generated by Large Language Models (LLMs) like GPT, BERT, or LLaMA serves as an effective solution. These LLMs can produce natural, diverse, and contextually relevant questions tailored to real-world scenarios.

**2.2 Methods for creating synthetic datasets**

The process of creating a synthetic dataset using LLMs includes the following steps:

* **Identify question categories**: Divide the dataset into three main categories: players, tournaments, and teams.
* **Design appropriate prompts**: Use LLMs to generate questions similar to real-world data by providing detailed prompt descriptions.
* **Automated data generation**: Run the model multiple times to produce a diverse set of questions, ensuring variation in structure and content.
* **Preprocessing and quality control**: Remove inappropriate questions, correct grammatical errors, and ensure accurately labeled data.

**2.3 Advantages of synthetic data**

* **Ensure quantity and balance**: Generate a large number of questions while maintaining balance across categories.
* **Easy customization**: Adjust question content to meet specific requirements.
* **Reduce the cost of real-world data collection**: Eliminate the need for manual collection or processing data from multiple sources.
* **High availability**: Reuse and expand the dataset for various purposes.

CHAPTER 3 - Common methods of text representation

**3.1 Bag of Words (BoW)**

Bag of Words (BoW) is one of the most common and fundamental text representation methods in Natural Language Processing (NLP). This method represents a text as a set of words appearing in the document without considering their order. This approach helps convert text into a numerical format, making it usable for machine learning algorithms.

### Working Principle

1. **Text Preprocessing**: Remove punctuation, convert uppercase letters to lowercase, and eliminate stopwords.
2. **Build Vocabulary**: Create a list of all unique words appearing in the training dataset.
3. **Generate Feature Vectors**: For each document, count the occurrences of each word from the vocabulary and store them as a vector.

### Advantages of BoW

* Simple, easy to understand, and easy to implement.
* Effective for small to medium-sized datasets.
* Requires minimal computational resources.

### Limitations of BoW

* Does not retain contextual information or word order.
* Feature matrices can become very large as vocabulary size increases.
* Cannot effectively handle synonyms.

#### **3.2** Naive Bayes

### Theory:

Naive Bayes is a classification algorithm based on Bayes' theorem, with the "naive" assumption that features are independent of each other.  
It is commonly used for text classification, spam filtering, and other simple classification tasks.

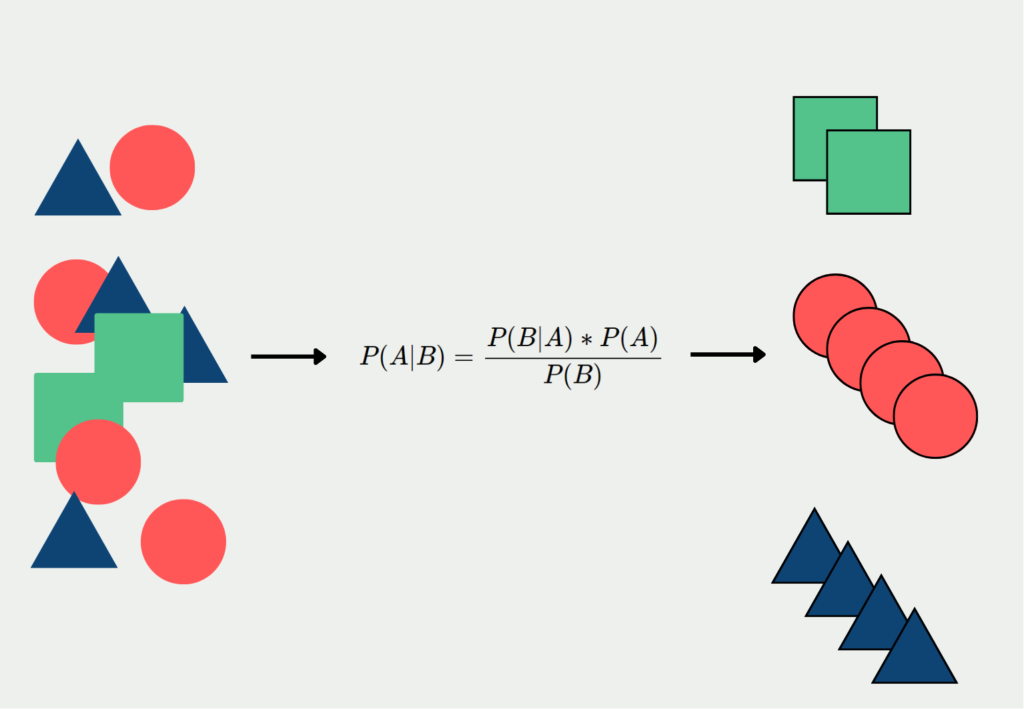
### Working Principle:

* Computes the **posterior probability** of a class based on the given features.
* Uses Bayes' theorem for calculation:



Where:

* P(y∣X): The probability of class yyy given the features X.
* P(X∣y): The probability of features X given class y.
* P(y): The prior probability of class y.
* P(X): The probability of features X.



*Image 3.2.1: Working Principle of Navie Bayes*

#### Calculation Formula:

Given the assumption of feature independence:



*Image 3.2.2: Formula of feature independence of navie Bayes*

The predicted class is the one with the highest posterior probability:



*Image 3.2.3: Formula of highest posterior probability of Navie Bayes*

### Strengths:

* Simple, easy to implement, and efficient for large datasets.
* Works well with high-dimensional data.
* Performs well with noisy data.

### Limitations:

* The independence assumption between features may not hold in real-world scenarios.
* Ineffective for datasets with complex feature dependencies.

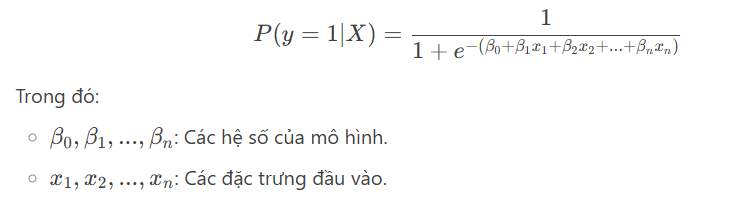
#### **3.3** Logistic Regression

### Theory:

Logistic Regression is a linear classification model that uses the logistic (sigmoid) function to predict the probability of a class.  
It is commonly used for binary classification problems.

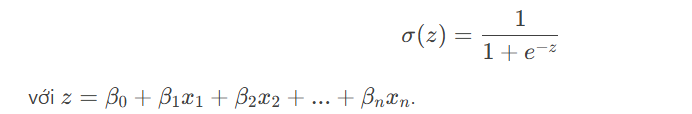
### Working Principle:

* Computes the probability of the positive class based on a linear combination of input features.
* Uses the **sigmoid function** to convert the output value into a probability:

Image 3.3.1: *Working Principle of Logistic Regression*

#### Calculation Formula:

**Sigmoid function:**



*Image 3.3.2: Formula sigmoid of Logistic Regression*

**Predicted class:**

### 

Image 3.3.3: *Formula predicted class of Logistic Regression*

### Strengths:

* Simple, easy to understand, and effective for data with a linear relationship.
* Can be extended to multi-class classification (multinomial logistic regression).

### Limitations:

* Assumes a **linear relationship** between features and log-odds.
* Ineffective for data with complex nonlinear relationships.

#### **3.4 Decision Tree**

### **Theory:**

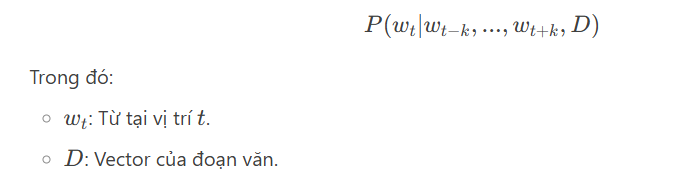
A Decision Tree is a classification and regression model based on a tree structure, where each node represents a feature, each branch represents a decision, and each leaf represents an outcome.

### **Working Principle:**

* Splits data into subsets based on features and their values.
* Uses metrics such as **Gini Index**, **Entropy**, or **Information Gain** to determine the best feature for splitting.

### **Calculation Formula:**

**Entropy:**



*Image 3.4.1: Formula Entropy of Decision Tree*

**Information Gain:**

### 

Image 3.4.2: *Formula information gain of Decision Tree*

### **Strengths:**

* Easy to understand and interpret.
* No need for data normalization.
* Works well with both numerical and categorical data.

### **Limitations:**

* Prone to **overfitting** if the tree is too complex.
* Sensitive to small changes in data.
* Inefficient for datasets with strong linear relationships.

#### **3.5** Using Pretrain Models (BERT/PhoBERT)

Pretrained Models are large language models that have been trained on vast amounts of text data. They can be used to extract vector representations for words, sentences, or paragraphs.

### **a. Without Pretrained Models**

The training process involves:

* Initializing word and sentence vectors randomly.
* Updating these vectors through optimization to predict words or sentences.

### **b. Using Pretrained Models (BERT/PhoBERT)**

**BERT (Bidirectional Encoder Representations from Transformers)** and **PhoBERT** (the Vietnamese version of BERT) are large language models utilizing the Transformer architecture.  
They can be used to extract vector representations for text.

### **Usage:**

#### **Extracting Embeddings from BERT/PhoBERT:**

1. Input text into the BERT/PhoBERT model.
2. Retrieve vectors from the final hidden layer or intermediate layers to represent words, sentences, or paragraphs.
   * Example: Use the vector of the [CLS] token to represent an entire sentence.

#### **Fine-tuning:**

* If labeled data is available, fine-tune the BERT/PhoBERT model on your dataset to enhance performance.

### **Formula:**

The output of BERT/PhoBERT is a sequence of vectors:

### 

*Image 3.5.1: Formula of Pretrained Models Doc2Vec*

### **Advantages:**

- Leverages knowledge from large pretrained language models.  
- Captures deep semantic meanings of text.  
- Highly effective for complex NLP tasks.

### **Limitations:**

- Requires substantial computational resources for training and fine-tuning.  
- Requires knowledge of Transformer architectures and libraries such as Hugging Face Transformers.

CHAPTER 4 - Comparison of data approaches and machine learning methods for text classification

#### **4.1** Element 1: Machine learning methods

|  |  |  |
| --- | --- | --- |
| **Model** | **Advantages** | **Disadvantages** |
| **Naive Bayes** | - Fast, stable with large data. - Suitable for TF-IDF. | - Assumes unrealistic independence. - Low performance on complex data. |
| **Logistic Regression** | - Flexible, handles multidimensional data. - Easy to interpret. | - Cannot handle non-linear relationships. |
| **Decision Tree** | - No need for data normalization. - Intuitive and easy to understand. | - Prone to overfitting. - Ineffective with vector-based data (BERT/Doc2Vec). |
| **BERT/PhoBERT** | - Highest accuracy. - Automatically extracts features. | - Requires GPU and long training time. |

*Table 4.1.1: Table Machine learning methods*

**Conclusion:**

* **Logistic Regression** and **BERT** are the best choices when combined with modern text representation methods.
* **Decision Tree** should only be used for simple data (e.g., keyword-based classification).

#### **4.2** Element 2: Completeness and accuracy of data

|  |  |  |
| --- | --- | --- |
| **Method** | **Low/Incomplete Data** | **High-Quality/Complete Data** |
| **Bag-of-Words/TF-IDF** | - Handles noise well due to simplicity. - Easy to adjust. | - Low performance due to representation limitations. |
| **Doc2Vec** | - Requires at least 10k-50k documents for effective training. | - Stable performance if data is clean. |
| **BERT/PhoBERT** | - Can be fine-tuned on small data thanks to pretraining. - Robust with various data types. | - Achieves highest performance with high-quality, consistent data. |

Table 4.2.1:*Table Completeness and accuracy*

**Conclusion:**

* **BERT/PhoBERT** is suitable for all types of data due to pretraining advantages, but **Doc2Vec** and **Bag-of-Words** are better if resources are limited or data is insufficient.

#### **4.3** Element 3: Text representation

| **Method** | **Advantages** | **Disadvantages** |
| --- | --- | --- |
| **Bag-of-Words/TF-IDF** | - Simple, easy to implement. - Suitable for small datasets. | - Cannot capture semantics or context. - Ignores word order. |
| **Doc2Vec** | - Captures semantics at the paragraph level. - Relatively lightweight. | - Less effective on complex text. - Dependent on training data quality. |
| **BERT/PhoBERT** | - Deep semantic representation, understands contextual meaning. - Handles multiple languages. | - Requires high computational resources. - Complex implementation. |

Table 4.3.1:*Table Text representation*

**Conclusion:**

* **BERT/PhoBERT** excels in capturing semantic meaning, but **Doc2Vec** and **Bag-of-Words** are better suited for simple data or limited resources.

**BIBLIOGRAPHY**

1. <https://www.greelane.com/vi/khoa-h%E1%BB%8Dc-c%C3%B4ng-ngh%E1%BB%87-to%C3%A1n/m%C3%B4n-to%C3%A1n/what-is-a-converse-error-3126461/>
2. <https://www.greelane.com/vi/khoa-h%E1%BB%8Dc-c%C3%B4ng-ngh%E1%BB%87-to%C3%A1n/m%C3%B4n-to%C3%A1n/difference-between-type-i-and-type-ii-errors-3126414/#google_vignette>
3. <https://philosophy.stackexchange.com/questions/65103/could-an-argument-with-false-premises-and-a-true-conclusion-be-logically-valid>
4. <https://www.csun.edu/~hcphi003/PHIL100/Supplements/1_validity.pdf>
5. <https://ahaslides.com/vi/blog/self-concept-examples/>
6. <https://www.greelane.com/vi/nh%C3%A2n-v%C4%83n/anh/circular-reasoning-petitio-principii-1689842/>
7. <https://wikiluat.com/2024/01/31/jump-to-conclusion-la-gi/>
8. <https://www.swi-prolog.org/pldoc/man?section=trace-example>
9. <https://www.researchgate.net/figure/Instrumentation-specification-for-the-SQLite-core-API_fig5_317687585>
10. <https://math.stackexchange.com/questions/3437198/fitting-an-activation-function-with-tanh>