# BÁO CÁO NGHIÊN CỨU

## PHÂN LOẠI CÂU HỎI VỀ CẦU THỦ, GIẢI ĐẤU VÀ ĐỘI BÓNG TRONG BÓNG ĐÁ

### 1. Giới thiệu

Trong lĩnh vực bóng đá, các hệ thống hỏi đáp tự động và chatbot thể thao ngày càng được ứng dụng rộng rãi. Một trong những bài toán quan trọng trong lĩnh vực này là phân loại câu hỏi theo các danh mục như cầu thủ, giải đấu và đội bóng. Việc tự động phân loại câu hỏi giúp nâng cao hiệu quả tìm kiếm thông tin, cải thiện trải nghiệm người dùng và tối ưu hóa các hệ thống trợ lý ảo.

Báo cáo này trình bày nghiên cứu về bài toán phân loại văn bản trong lĩnh vực bóng đá, áp dụng các phương pháp học máy để tự động phân loại câu hỏi. Phân tích cách xây dựng tập dữ liệu phù hợp và so sánh hiệu suất của các phương pháp khác nhau trên tập dữ liệu này nhằm tìm ra phương pháp tối ưu nhất.

### 2. Mục tiêu nghiên cứu

Mục tiêu chính của nghiên cứu này bao gồm:

* Xây dựng tập dữ liệu gồm các câu hỏi liên quan đến cầu thủ, giải đấu và đội bóng.
* Áp dụng các phương pháp học máy để phân loại câu hỏi theo danh mục.
* Đánh giá hiệu suất của từng phương pháp và đề xuất giải pháp tối ưu.

### 3. Lợi ích của tập dữ liệu

Việc xây dựng tập dữ liệu phù hợp cho bài toán phân loại câu hỏi trong bóng đá mang lại nhiều lợi ích quan trọng:

* **Tối ưu hóa hệ thống tìm kiếm**: Khi câu hỏi được phân loại chính xác, hệ thống có thể nhanh chóng cung cấp câu trả lời phù hợp, giúp người dùng tìm kiếm thông tin dễ dàng hơn.
* **Cải thiện chatbot thể thao**: Chatbot có thể hiểu ngữ cảnh câu hỏi tốt hơn, đưa ra phản hồi chính xác hơn và giúp nâng cao trải nghiệm người dùng.
* **Hỗ trợ huấn luyện mô hình AI**: Tập dữ liệu giúp đào tạo mô hình học máy để nhận diện chính xác chủ đề của câu hỏi, từ đó áp dụng vào nhiều ứng dụng khác như trợ lý ảo hay phân tích dữ liệu thể thao.
* **Phát triển ứng dụng phân tích dữ liệu thể thao**: Các hệ thống AI có thể sử dụng dữ liệu này để thống kê xu hướng tìm kiếm, đưa ra dự đoán về các chủ đề phổ biến trong cộng đồng bóng đá.

### 4. Xây dựng tập dữ liệu tổng hợp bằng Large Language Models

#### 4.1. Lý do sử dụng dữ liệu tổng hợp

Trong nhiều trường hợp, việc thu thập dữ liệu thực tế có thể gặp phải các vấn đề như hạn chế quyền truy cập, thiếu dữ liệu có nhãn hoặc dữ liệu không cân bằng giữa các danh mục. Do đó, việc sử dụng dữ liệu tổng hợp (synthetic data) được tạo ra bằng các mô hình ngôn ngữ lớn (Large Language Models - LLMs) là một giải pháp hiệu quả. Các LLM như GPT, BERT, hay LLaMA có thể sinh ra các câu hỏi tự nhiên, đa dạng và phù hợp với ngữ cảnh thực tế.

#### 4.2. Phương pháp tạo tập dữ liệu tổng hợp

Quy trình tạo tập dữ liệu tổng hợp sử dụng LLMs bao gồm các bước sau:

* **Xác định danh mục câu hỏi**: Chia tập dữ liệu thành ba danh mục chính: cầu thủ, giải đấu và đội bóng.
* **Thiết lập prompt phù hợp**: Sử dụng LLM để tạo ra các câu hỏi tương tự với dữ liệu thực tế bằng cách cung cấp prompt mô tả chi tiết.
* **Sinh dữ liệu tự động**: Chạy mô hình nhiều lần để tạo ra một tập câu hỏi phong phú, đảm bảo sự đa dạng về cấu trúc và nội dung.
* **Tiền xử lý và kiểm tra chất lượng**: Loại bỏ câu hỏi không phù hợp, sửa lỗi ngữ pháp và đảm bảo dữ liệu có nhãn chính xác.

#### 4.3. Ưu điểm của dữ liệu tổng hợp

* **Đảm bảo số lượng và độ cân bằng**: Có thể tạo ra số lượng câu hỏi lớn và đảm bảo sự cân bằng giữa các danh mục.
* **Tùy chỉnh dễ dàng**: Có thể điều chỉnh nội dung câu hỏi theo từng yêu cầu cụ thể.
* **Giảm chi phí thu thập dữ liệu thực tế**: Không cần thu thập thủ công hoặc xử lý dữ liệu từ nhiều nguồn khác nhau.
* **Tính khả dụng cao**: Có thể tái sử dụng và mở rộng tập dữ liệu cho nhiều mục đích khác nhau.

### 5. Các phương pháp biểu diễn văn bản thông thường

#### 5.1. Bag of Words (BoW)

Bag of Words (BoW) là một trong những phương pháp biểu diễn văn bản phổ biến và cơ bản nhất trong xử lý ngôn ngữ tự nhiên (NLP). Phương pháp này biểu diễn văn bản dưới dạng một tập hợp các từ xuất hiện trong tài liệu mà không quan tâm đến thứ tự của chúng. Cách tiếp cận này giúp chuyển đổi văn bản thành dạng số học để có thể sử dụng trong các thuật toán học máy.

##### Nguyên lý hoạt động

1. **Tiền xử lý văn bản**: Xóa dấu câu, chuyển đổi chữ hoa thành chữ thường, loại bỏ từ dừng (stopwords).
2. **Xây dựng từ điển (Vocabulary)**: Tạo danh sách tất cả các từ xuất hiện trong tập dữ liệu huấn luyện.
3. **Tạo vector đặc trưng**: Với mỗi tài liệu, đếm số lần xuất hiện của từng từ trong từ điển và lưu trữ dưới dạng vector.

##### Ưu điểm của BoW

* Đơn giản, dễ hiểu và dễ triển khai.
* Hiệu quả đối với các tập dữ liệu có kích thước nhỏ hoặc trung bình.
* Không yêu cầu quá nhiều tài nguyên tính toán.

##### Hạn chế của BoW

* Không giữ được thông tin về ngữ cảnh và thứ tự từ.
* Ma trận đặc trưng có thể trở nên rất lớn khi số lượng từ vựng tăng.
* Không thể xử lý các từ đồng nghĩa một cách hiệu quả.

### 6. Các phương pháp học máy truyền thống

#### 6.1. Naive Bayes

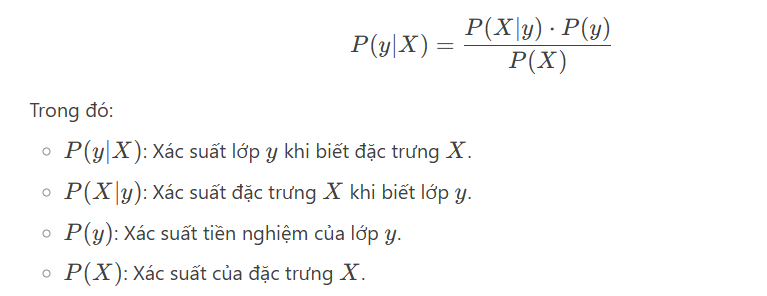
#### Lý thuyết:

Naive Bayes là một thuật toán phân loại dựa trên định lý Bayes với giả định "naive" rằng các đặc trưng (features) là độc lập với nhau.

Nó thường được sử dụng cho các bài toán phân loại văn bản, lọc spam, và các bài toán phân loại đơn giản.

#### Nguyên lý hoạt động:

Tính xác suất hậu nghiệm (posterior probability) của một lớp (class) dựa trên các đặc trưng đầu vào.

Sử dụng công thức Bayes:

#### Công thức tính toán:

Với giả định các đặc trưng độc lập:



Lớp dự đoán là lớp có xác suất hậu nghiệm lớn nhất:



#### Điểm mạnh:

Đơn giản, dễ triển khai và hiệu quả với dữ liệu có kích thước lớn.

Hiệu quả với dữ liệu có số chiều cao.

Hoạt động tốt với dữ liệu nhiễu.

#### Hạn chế:

Giả định độc lập giữa các đặc trưng có thể không đúng trong thực tế.

Kém hiệu quả với dữ liệu có mối quan hệ phức tạp giữa các đặc trưng.

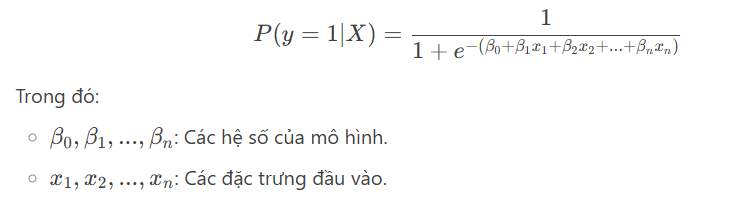
#### 6.2. Logistic Regression

#### Lý thuyết:

Logistic Regression là một mô hình phân loại tuyến tính, sử dụng hàm logistic (sigmoid) để dự đoán xác suất của một lớp.

Nó thường được sử dụng cho các bài toán phân loại nhị phân (binary classification).

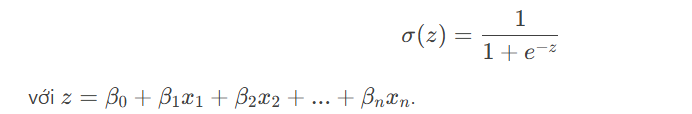
#### Nguyên lý hoạt động:

Tính toán xác suất của lớp dương (positive class) dựa trên tổ hợp tuyến tính của các đặc trưng đầu vào.

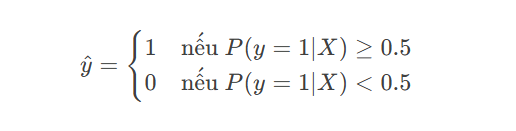
Sử dụng hàm sigmoid để chuyển đổi giá trị đầu ra thành xác suất:

#### Công thức tính toán:

Hàm sigmoid:



Lớp dự đoán:



#### Điểm mạnh:

Đơn giản, dễ hiểu và hiệu quả với dữ liệu có mối quan hệ tuyến tính.

Có thể mở rộng cho bài toán phân loại đa lớp (multinomial logistic regression).

#### Hạn chế:

Giả định mối quan hệ tuyến tính giữa đặc trưng và log-odds.

Kém hiệu quả với dữ liệu có mối quan hệ phi tuyến phức tạp.

#### 6.3. Decision Tree

#### Lý thuyết:

Decision Tree là một mô hình phân loại và hồi quy dựa trên cây quyết định, nơi mỗi nút trong cây đại diện cho một đặc trưng, mỗi nhánh đại diện cho một quyết định, và mỗi lá đại diện cho kết quả.

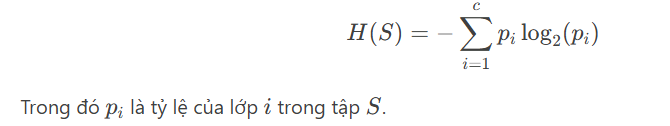
#### Nguyên lý hoạt động:

Phân chia dữ liệu thành các tập con dựa trên các đặc trưng và giá trị của chúng.

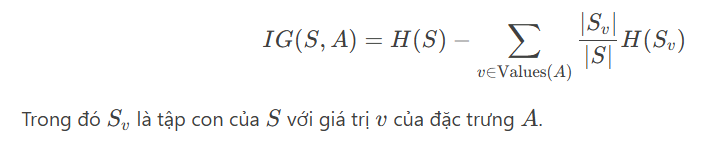
Sử dụng các chỉ số như Gini Index, Entropy, hoặc Information Gain để chọn đặc trưng tốt nhất để phân chia.

#### Công thức tính toán:

**Entropy**:



**Information Gain**:



#### Điểm mạnh:

Dễ hiểu và trực quan.

Không cần chuẩn hóa dữ liệu.

Xử lý tốt cả dữ liệu số và dữ liệu phân loại.

#### Hạn chế:

Dễ bị overfitting nếu cây quá phức tạp.

Nhạy cảm với sự thay đổi nhỏ trong dữ liệu.

Không hiệu quả với dữ liệu có mối quan hệ tuyến tính mạnh.

#### 6.4. ****Doc2Vec với Deep Learning****

**Doc2Vec** là một mô hình học sâu (deep learning) dùng để biểu diễn văn bản (document) dưới dạng các vector có số chiều cố định. Nó là một mở rộng của mô hình Word2Vec, không chỉ biểu diễn từ mà còn biểu diễn cả đoạn văn hoặc tài liệu.

##### Nguyên lý hoạt động:

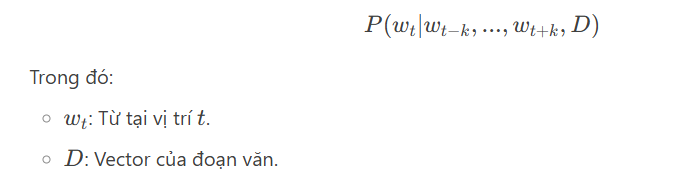
**Doc2Vec** sử dụng hai kiến trúc chính:

**PV-DM (Paragraph Vector - Distributed Memory)**: Mô hình này kết hợp vector của từ (word vectors) và vector của đoạn văn (paragraph vector) để dự đoán từ tiếp theo.

**PV-DBOW (Paragraph Vector - Distributed Bag of Words)**: Mô hình này sử dụng vector của đoạn văn để dự đoán các từ trong đoạn văn mà không quan tâm đến thứ tự của chúng.

##### Công thức:

Vector của đoạn văn D được học cùng với các vector của từ trong quá trình huấn luyện:



##### Điểm mạnh:

Biểu diễn được cả đoạn văn dưới dạng vector.

Dễ dàng tích hợp vào các mô hình học máy khác.

##### Hạn chế:

Cần một lượng lớn dữ liệu để huấn luyện hiệu quả.

Không thể tận dụng được các mô hình ngôn ngữ lớn (pretrained models) như BERT.

#### 2. ****Sử dụng Pretrain Models (BERT/PhoBERT)****

**Pretrain Models** là các mô hình ngôn ngữ lớn đã được huấn luyện trước trên một lượng lớn dữ liệu văn bản. Chúng có thể được sử dụng để trích xuất các biểu diễn vector cho từ, câu hoặc đoạn văn.

##### a. ****Không sử dụng Pretrain Models****

Quá trình huấn luyện bao gồm:

Khởi tạo các vector từ và vector đoạn văn ngẫu nhiên.

Cập nhật các vector này thông qua quá trình tối ưu hóa để dự đoán từ hoặc đoạn văn.

##### b. ****Sử dụng Pretrain Models (BERT/PhoBERT)****

**BERT (Bidirectional Encoder Representations from Transformers)** và **PhoBERT** (phiên bản tiếng Việt của BERT) là các mô hình ngôn ngữ lớn sử dụng kiến trúc Transformer.

Chúng có thể được sử dụng để trích xuất các biểu diễn vector cho văn bản.

###### Cách sử dụng:

**Trích xuất Embedding từ BERT/PhoBERT**:

Đưa văn bản vào mô hình BERT/PhoBERT.

Lấy các vector từ lớp ẩn cuối cùng hoặc các lớp ẩn trung gian để biểu diễn từ, câu hoặc đoạn văn.

Ví dụ: Sử dụng vector của token [CLS] để biểu diễn cả câu.

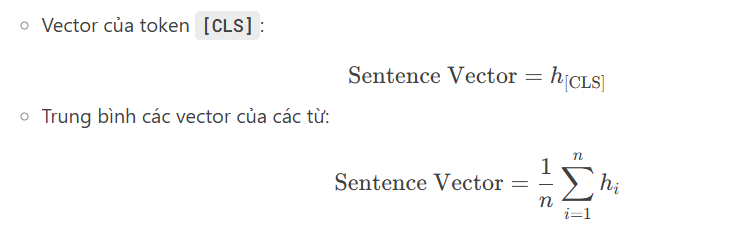
**Fine-tuning (tinh chỉnh)**:

Nếu có dữ liệu được gán nhãn, bạn có thể fine-tuning mô hình BERT/PhoBERT trên tập dữ liệu của mình để cải thiện hiệu suất.

###### Công thức:

Đầu ra của BERT/PhoBERT là một chuỗi các vector h1,h2,...,hn*h*1​,*h*2​,...,*hn*​, trong đó hi*hi*​ là vector biểu diễn của từ thứ i*i*.

Để biểu diễn cả câu, bạn có thể sử dụng:



##### Điểm mạnh:

Tận dụng được kiến thức từ các mô hình ngôn ngữ lớn đã được huấn luyện trước.

Biểu diễn được ngữ nghĩa sâu sắc của văn bản.

Hiệu quả cao trong các bài toán NLP phức tạp.

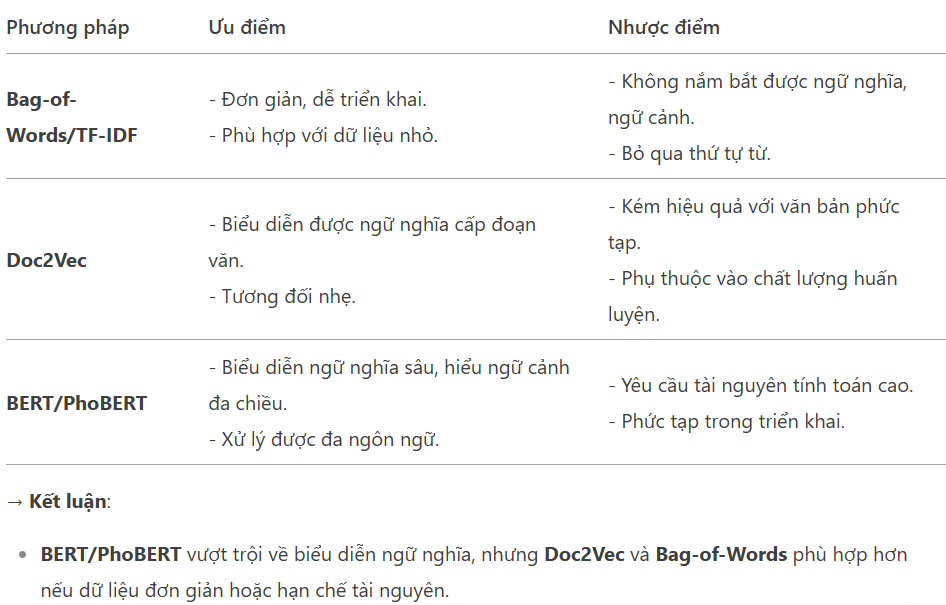
##### Hạn chế:

Yêu cầu tài nguyên tính toán lớn để huấn luyện và fine-tuning.

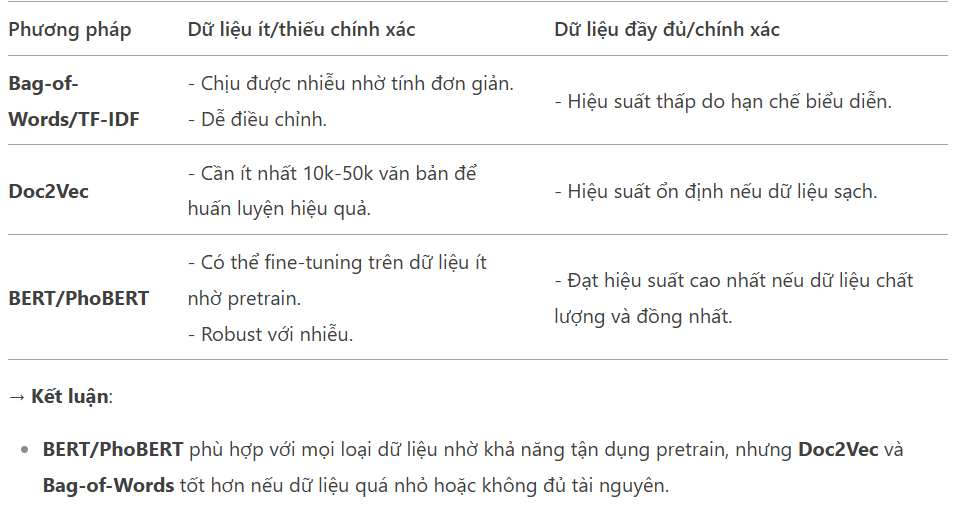
Cần hiểu biết về kiến trúc Transformer và cách sử dụng các thư viện như Hugging Face Transformers.

### ****So sánh các tiếp cận dữ liệu và phương pháp học máy cho bài toán phân loại văn bản****

#### ****Yếu tố 1: Biểu diễn văn bản****



#### ****Yếu tố 2: Tính đầy đủ và chính xác của dữ liệu****



#### ****Yếu tố 3: Phương pháp học máy****

