PhoBERT là một biến thể của BERT được phát triển nhằm giải quyết những hạn chế của BERT khi xử lý tiếng Việt. Mặc dù BERT rất mạnh trong việc xử lý ngôn ngữ tự nhiên, nhưng vì nó chủ yếu được huấn luyện trên dữ liệu tiếng Anh, nên hiệu quả khi áp dụng cho các ngôn ngữ khác không cao. Do được đào tạo hoàn toàn trên dữ liệu văn bản tiếng Việt, PhoBERT nắm bắt rất tốt các đặc điểm ngữ nghĩa và cú pháp cụ thể của ngôn ngữ này.

PhoBERT tích hợp các cải tiến từ RoBERTa, chẳng hạn như loại bỏ nhiệm vụ Dự đoán Câu Kế Tiếp (NSP) và chỉ sử dụng Mô hình Ngôn ngữ Ẩn (MLM) – một nhiệm vụ giúp mô hình hiểu ngữ cảnh của các từ trong câu bằng cách dự đoán từ bị ẩn, đồng thời được đào tạo trên tập dữ liệu quy mô lớn.

PhoBERT được huấn luyện trên một tập dữ liệu đa dạng khoảng 20 GB, bao gồm khoảng 1 GB được trích xuất từ Wikipedia tiếng Việt và khoảng 19 GB từ các bài báo tiếng Việt. Điều này đảm bảo rằng mô hình có thể học từ nhiều nguồn khác nhau, bao gồm cả ngôn ngữ chính quy và ngôn ngữ báo chí, giúp nó có khả năng hiểu tốt các bối cảnh ngôn ngữ khác nhau trong thực tế.

Trước khi dữ liệu được đưa vào bộ mã hóa Byte-Pair Encoding (BPE), PhoBERT sử dụng công cụ RDRSegmenter từ VnCoreNLP để tách từ. Điều này giúp mô hình xử lý tốt hơn các vấn đề về ngữ nghĩa và cú pháp trong tiếng Việt, vì tiếng Việt là ngôn ngữ dính kết (agglutinative language), nghĩa là một từ có thể chứa nhiều thành phần khác nhau, làm phức tạp việc tách từ.

Nhờ vào việc huấn luyện và tinh chỉnh trên một tập dữ liệu tiếng Việt lớn và đa dạng, PhoBERT có khả năng xử lý tốt hơn các cấu trúc phức tạp của tiếng Việt so với BERT và RoBERTa. Từ đó PhoBERT có thể được sử dụng trong nhiều bài toán xử lý ngôn ngữ tự nhiên dành cho Tiếng Việt như: phân loại văn bản, trả lời câu hỏi, nhận diện thực thể có tên, dịch máy, tóm tắt văn bản,…

Cụ thể trong nhiệm vụ phân loại tin tức thật và giả bằng tiếng Việt, PhoBERT đóng vai trò quan trọng trong việc phân tích và hiểu ngữ nghĩa của văn bản Tiếng Việt. Bằng cách học từ một tập dữ liệu lớn và đa dạng, PhoBERT có thể phân biệt giữa tin tức thật và giả thông qua sự khác biệt về ngữ cảnh, ngữ pháp và phong cách ngôn ngữ. Từ đó giúp nâng cao độ chính xác của PhoBERT trong việc phát hiện tin giả Tiếng Việt so với các mô hình khác.