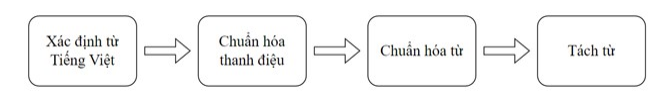
**Tiền xử lí**

Trong quá trình xử lý ngôn ngữ tự nhiên, tiền xử lí dữ liệu đóng vai trò vô cùng quan trọng. Đặc biệt, trong trường hợp ngôn ngữ Tiếng Việt, tiền xử lí dữ liệu trở nên càng quan trọng hơn do tồn tại nhiều biến thể của cùng một từ, đôi khi người dùng viết sai chính tả hoặc viết tắt.

Việc tiền xử lí tốt có thể giúp cải thiện đáng kể hiệu suất của mô hình. Đối với bài toán này, quá trình tiền xử lí phải đảm bảo đúng vị trí từng nhãn từ loại, do đó chúng tôi mô tả quá trình tiền xử lí trong Hình 2.



Hình 2: Các bước tiền xử lí dữ liệu

Đầu tiên, chúng tôi tiến hành kiểm tra xem từ đó có phải là từ Tiếng Việt hay không. Nếu đó là từ Tiếng Việt, chúng tôi sẽ chuẩn hóa các thanh điệu về dạng chuẩn theo các quy tắc về nguyên âm và phụ âm của Tiếng Việt (ví dụ: "maý" => "máy"). Tiếp theo, chúng tôi chuyển các câu bình luận sang kiểu viết thường, rút gọn các từ kéo dài ("xịnnnn" => "xịn"), sau đó chuẩn hóa các từ viết tắt thông dụng. Cuối cùng, chúng tôi thực hiện tách từ để tạo đầu vào cho các mô hình được huấn luyện trước trên Tiếng Việt. Bảng 2 mô tả đầu vào và đầu ra của quá trình tiền xử lí.

**Mô hình**

PhoBert:

PhoBERT là một mô hình ngôn ngữ dựa trên kiến trúc Transformer, được đào tạo đặc biệt cho tiếng Việt. Được xây dựng trên cơ sở của RoBERTa (A Robustly Optimized BERT Pretraining Approach), PhoBERT có khả năng hiểu và biểu diễn các ngữ cảnh ngôn ngữ tiếng Việt phức tạp.

Mô hình PhoBERT được huấn luyện trước (pretrain) trên một lượng lớn dữ liệu tiếng Việt từ các nguồn khác nhau trên Internet, bao gồm các trang web, bài viết, sách và văn bản trên mạng xã hội. Quá trình pretrain này giúp PhoBERT nắm bắt được nhiều thông tin ngôn ngữ phong phú, từ đó tạo ra các biểu diễn từ và câu có khả năng áp dụng cho nhiều tác vụ ngôn ngữ khác nhau.

PhoBERTv2:

PhoBERTv2 là một mô hình ngôn ngữ tiếng Việt cải tiến dựa trên kiến trúc Transformer. Nó có số lượng lớp mạng tăng, kích thước embedding và lớp ẩn được tăng lên. Qua quá trình huấn luyện trước trên dữ liệu lớn từ nhiều nguồn khác nhau, PhoBERTv2 có khả năng hiểu và biểu diễn ngôn ngữ tiếng Việt phức tạp một cách hiệu quả hơn.

PhoBERTv2 là phiên bản cải tiến của PhoBERT, nâng cao khả năng biểu diễn ngôn ngữ tiếng Việt. Nó sử dụng tăng số lượng lớp mạng, kích thước embedding và lớp ẩn để cải thiện hiệu suất. Quá trình pretrain của PhoBERTv2 trên dữ liệu lớn từ nhiều nguồn giúp nắm bắt ngữ cảnh và thông tin quan trọng, tạo ra các biểu diễn đặc trưng cho ngôn ngữ tiếng Việt.

XLM-RoBERTa:

XLM-RoBERTa là một mô hình ngôn ngữ tự nhiên đa ngôn ngữ, được huấn luyện trên cơ sở dữ liệu lớn từ nhiều ngôn ngữ khác nhau. Mô hình này là một phiên bản mở rộng từ RoBERTa (một biến thể của BERT) và có khả năng biểu diễn ngôn ngữ đa ngôn ngữ một cách hiệu quả. XLM-RoBERTa đã được áp dụng rộng rãi trong các nhiệm vụ xử lý ngôn ngữ tự nhiên, bao gồm cả nhận diện và trích xuất cảm xúc trong đoạn văn bản Tiếng Việt. Mô hình này cung cấp một khả năng đa ngôn ngữ linh hoạt và hiệu suất tốt trong việc hiểu và biểu diễn ngôn ngữ từ các nguồn dữ liệu đa ngôn ngữ.