**4. Đề xuất (Proposal)**

Đề xuất cần bao gồm các phần sau:

**4.1. Giới thiệu về bài toán và dữ liệu**

Trong thời đại kỹ thuật số ngày nay, dữ liệu không chỉ đóng vai trò như một nguồn tài nguyên quan trọng mà còn được xem là yếu tố quyết định để hiểu rõ hơn về xã hội và thị trường. Mạng xã hội, đặc biệt là Twitter, đã trở thành một trong những nền tảng giao tiếp và truyền tải thông tin phổ biến nhất trên toàn cầu. Với hàng triệu bài đăng mỗi ngày, việc duy trì nội dung hấp dẫn và phù hợp với xu hướng là một thách thức không nhỏ đối với các cá nhân, doanh nghiệp và tổ chức.

Việc sáng tạo nội dung thủ công đòi hỏi nhiều thời gian, công sức và sự sáng tạo, điều này đôi khi làm giảm hiệu suất hoặc không đáp ứng kịp thời những thay đổi liên tục của thông tin. Để khắc phục, sự phát triển của trí tuệ nhân tạo, đặc biệt là các mô hình Text Generation, đã mang lại một giải pháp đột phá trong việc tự động hóa sáng tạo nội dung. Các mô hình này sử dụng học sâu (Deep Learning) để phân tích dữ liệu văn bản và tạo ra những nội dung mới một cách tự nhiên, phù hợp với ngữ cảnh. Đây là cách tiếp cận hiệu quả để tạo ra các bài viết có chất lượng cao và đồng thời giảm thiểu chi phí về thời gian.

Dự án này tập trung vào ứng dụng Text Generation nhằm tạo ra nội dung bài viết trên Twitter dựa trên các mô hình học sâu hiện đại. Thay vì chỉ dừng lại ở việc xử lý dữ liệu, hệ thống này giúp tự động hóa toàn bộ quá trình sáng tạo nội dung từ khâu phân tích đến khâu xuất bản, với mục tiêu đảm bảo nội dung được tạo ra không chỉ sáng tạo mà còn mang tính cá nhân hóa và bắt kịp xu hướng.

Hệ thống Text Generation có thể được triển khai trong nhiều lĩnh vực thực tế như tiếp thị số (digital marketing), quản lý truyền thông xã hội, hoặc dự báo xu hướng. Chẳng hạn, các doanh nghiệp có thể sử dụng công nghệ này để xây dựng chiến lược truyền thông hiệu quả hơn, tạo các bài viết tương tác với khách hàng, hoặc phát hiện sớm những thay đổi trong hành vi người dùng.

Với sự phát triển không ngừng của trí tuệ nhân tạo và học sâu, dự án không chỉ dừng lại ở việc tạo ra những nội dung sáng tạo mà còn đặt nền móng cho những hệ thống thông minh hơn, có khả năng đáp ứng nhanh chóng và linh hoạt trong thế giới số hóa đang thay đổi từng ngày. Ứng dụng Text Generation trên Twitter chính là minh chứng cho tiềm năng vô hạn của công nghệ trong việc thay đổi cách chúng ta sáng tạo và giao tiếp trong kỷ nguyên 4.0.

**4.2. Kế hoạch phân tích dữ liệu**

**- Input (biến giải thích, X) và Output (biến phản hồi, Y):**

Biến giải thích (X):

* Nội dung tweet: trường "text".

Biến phản hồi (Y):

* Văn bản tạo ra từ mô hình Text Generation.

**- Độ đo và phương pháp đánh giá:**

Độ đo cho hiệu suất mô hình LSTM:

+Hàm mất mát (Loss):

* Qua quá trình huấn luyện, hàm mất mát giảm đều qua các epoch và đạt giá trị cuối cùng là 0.0274 tại epoch 72.
* Điều này cho thấy mô hình đã học tốt từ dữ liệu huấn luyện và hội tụ ổn định, với sự sai lệch giữa dự đoán và giá trị thực tế rất nhỏ. Mức loss thấp này chứng tỏ rằng mô hình có khả năng dự đoán chính xác và hiệu quả.

+ Perplexity:

* Giá trị perplexity của mô hình là 1.0278, chỉ ra rằng mô hình có khả năng dự đoán các chuỗi văn bản một cách rất chính xác.
* Perplexity thấp như vậy gần bằng 1, cho thấy mô hình không gặp khó khăn khi tạo ra các dự đoán tiếp theo trong chuỗi, đồng thời phản ánh khả năng hiểu ngữ cảnh và tính toán xác suất cho các phần tử tiếp theo trong chuỗi.

**- Các phương pháp mà nhóm tin rằng sẽ hữu ích trong việc trả lời các câu hỏi đặt ra:**

Tiền xử lý dữ liệu:

+ Nhận dữ liệu từ Kafka

* Kết nối đến Kafka topic nơi dữ liệu được truyền.
* Lấy dữ liệu dạng thô từ Kafka và chuyển thành DataFrame bằng Spark.

+ Làm sạch dữ liệu

* Loại bỏ các giá trị trống hoặc không hợp lệ.
* Chuyển đổi kiểu dữ liệu cho phù hợp với mô hình.
* Loại bỏ các ký tự đặc biệt, khoảng trắng thừa.

+ Chuẩn hóa dữ liệu

* Biến đổi văn bản thành chữ thường.
* Dùng thư viện NLP để xử lý văn bản: loại bỏ stopwords, stemming, lemmatization.

+ Tạo đặc trưng

* Sử dụng TF-IDF, Bag-of-Words hoặc embedding bằng mô hình pre-trained.
* Lưu dữ liệu đã xử lý vào file .txt để phục vụ cho quá trình huấn luyện mô hình.

Phân tích thống kê:

+ Thống kê mô tả

* Số lượng bản ghi sau khi làm sạch.
* Số lượng từ trung bình trong mỗi câu.

+ Phân tích tương quan

* Kiểm tra tương quan giữa các đặc trưng nếu có dữ liệu số.
* Xác định mối tương quan giữa các biến.

**- Kế hoạch thực hiện và phân công công việc:**

Giai đoạn 1: Thu thập và tiền xử lý dữ liệu

* Mục tiêu: Kết nối Kafka, thu thập dữ liệu, làm sạch và chuẩn hóa.

Giai đoạn 2: Phân tích thống kê

* Mục tiêu: Phân tích mô tả, thống kê và trực quan hóa dữ liệu.

Giai đoạn 3: Huấn luyện mô hình

* Mục tiêu: Xây dựng mô hình học máy, huấn luyện và đánh giá kết quả.

Giai đoạn 4: Viết báo cáo và trình bày

* Mục tiêu: Hoàn thiện báo cáo và trình bày kết quả.