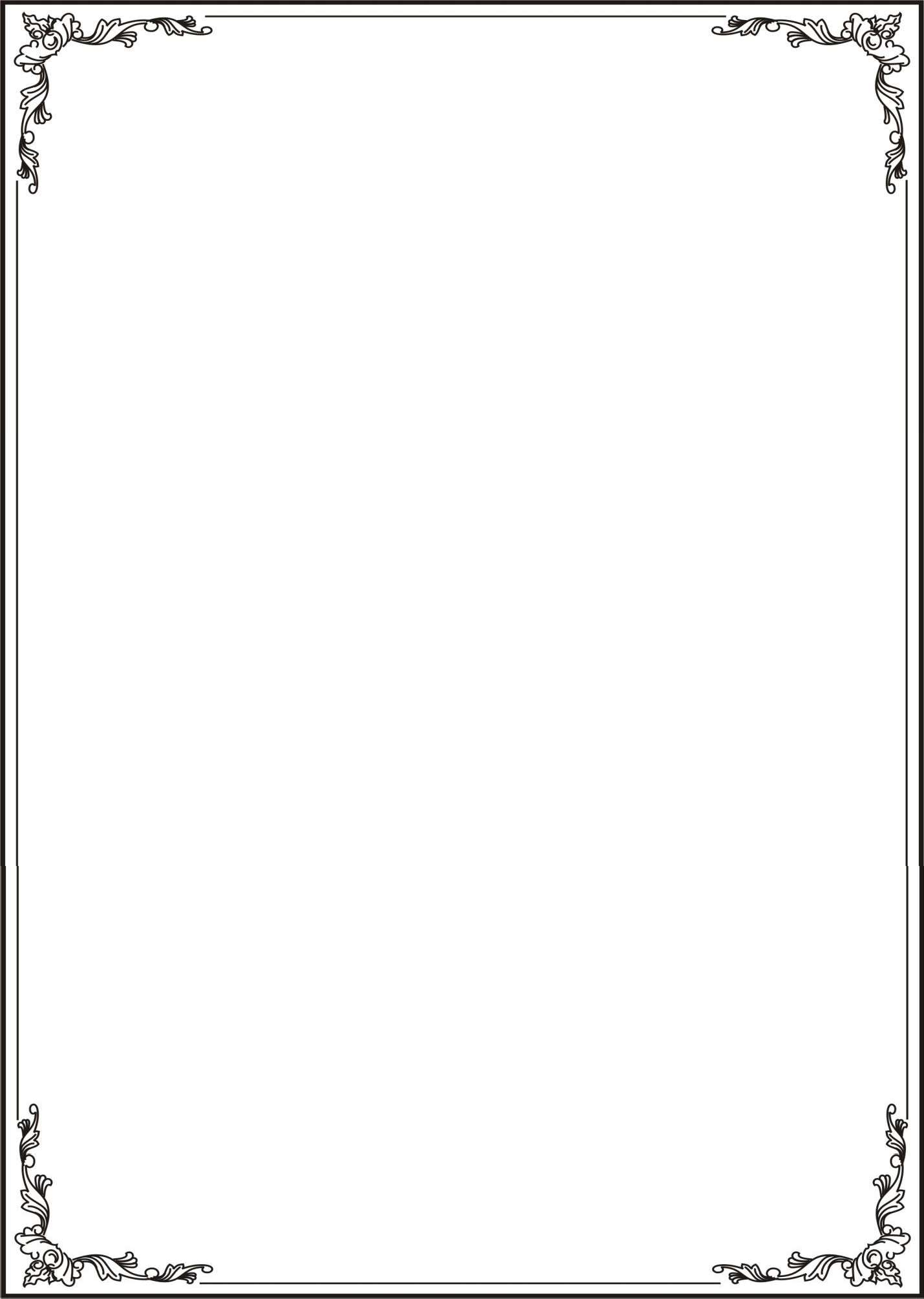
**BỘ GIÁO DỤC VÀ ĐÀO TẠO**



**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

**ĐỀ TÀI: TEXT-GENERATION TẠO BÀI POST ĐĂNG TRÊN TWITTER**

**Ứng dụng dữ liệu lớn: Học máy ở quy mô lớn**

**NHÓM SINH VIÊN THỰC HIỆN:**

|  |  |
| --- | --- |
| **Đinh Đức Nguyên Vũ** | **20128171** |
| **Nguyễn Trung Đức** | **21142261** |
| **Đỗ Tiến Đạt** | **21133024** |
| **Lương Tường Vy** | **21133093** |

**GIÁO VIÊN HƯỚNG DẪN:**

**ThS Quách Đình Hoàng**

**ĐH SƯ PHẠM KỸ THUẬT TP.HCM CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM**

**DANH SÁCH THÀNH VIÊN THAM GIA**

**THỰC HIỆN ĐỀ TÀI VÀ VIẾT BÁO CÁO**

**Môn: Ứng dụng dữ liệu lớn: Học máy ở quy mô lớn -** *HỌC KÌ II – NĂM HỌC 2023 – 2024*

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **HỌ VÀ TÊN** | **MSSV** | **TỶ LỆ ĐÓNG GÓP** |
| 1 | Đinh Đức Nguyên Vũ | 20128171 | 100% |
| 2 | Nguyễn Trung Đức | 21142261 | 100% |
| 3 | Đỗ Tiến Đạt | 21133024 | 100% |
| 4 | Lương Tường Vy | 21133093 | 100% |

**Nhận xét của giảng viên:**

**……………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………**

*Ngày … tháng 12 năm 2024*

*Giảng viên chấm điểm*

**LỜI CẢM ƠN**

Lời đầu tiên, nhóm chúng em xin gửi lời cảm ơn chân thành đến Trường Đại học Sư phạm Kỹ thuật TPHCM đã đưa môn học **Ứng dụng dữ liệu lớn: Học máy ở quy mô lớn** vào chương trình giảng dạy. Đặc biệt, chúng em xin gửi lời cảm ơn sâu sắc đến giảng viên bộ môn – Quách Đình Hoàng đã dạy dỗ, truyền đạt những kiến thức quý báu cho chúng em trong suốt thời gian học tập vừa qua. Trong thời gian tham gia lớp học **Ứng dụng dữ liệu lớn: Học máy ở quy mô lớn** của Thầy, nhóm em đã có thêm cho mình nhiều kiến thức bổ ích, tinh thần học tập hiệu quả, nghiêm túc. Đây chắc chắn sẽ là những kiến thức quý báu, là hành trang để chúng em có thể vững bước sau này.

**Môn học Ứng dụng dữ liệu lớn: Học máy ở quy mô lớn** là môn học thú vị, vô cùng bổ ích và có tính thực tế cao. Đảm bảo cung cấp đủ kiến thức, gắn liền với nhu cầu thực tiễn của sinh viên.

Mặc dù chúng em đã cố gắng hết sức nhưng chắc chắn bài đồ án này khó có thể tránh khỏi những thiếu sót và nhiều chỗ còn chưa chính xác, kính mong Thầy xem xét và góp ý để bài đồ án của chúng em được hoàn thiện hơn.

Nhóm chúng em xin chân thành cảm ơn!

**Mục Lục**

[1.Tóm tắt (abstract) 5](#_Toc185161385)

[2.Giới thiệu (introduction) 5](#_Toc185161386)

[3.Sơ đồ kiến trúc dự án 6](#_Toc185161387)

[4.Dữ liệu (data) 7](#_Toc185161388)

[4.1.Dữ liệu thu nhập từ Twitter 7](#_Toc185161389)

[4.2.Dữ liệu để training model 9](#_Toc185161390)

[5.Phương pháp (methods) 10](#_Toc185161391)

[6.Thực nghiệm, kết quả, và thảo luận 13](#_Toc185161392)

[6.1.Cấu hình thực nghiệm 13](#_Toc185161393)

[6.1.1.Các công cụ và cấu hình 14](#_Toc185161394)

[6.2.Phương pháp đánh giá 14](#_Toc185161395)

[Ví dụ: 15](#_Toc185161396)

[6.3.Kết quả thực nghiệm 15](#_Toc185161397)

[6.4.Phân tích kết quả 15](#_Toc185161398)

[6.4.1.Ưu điểm của hệ thống 15](#_Toc185161399)

[6.4.2.Hạn chế 15](#_Toc185161400)

[7.Kết luận (conclusion) - khoảng 1-2 đoạn 16](#_Toc185161401)

[8.Phụ lục (appendices) - tối đa 2 trang 16](#_Toc185161402)

[9.Đóng góp (contributions) 18](#_Toc185161403)

[10.Tham khảo 19](#_Toc185161404)

# 1.Tóm tắt (abstract)

Dự án này trình bày một giải pháp tự động hóa trong việc tạo nội dung quảng cáo liên quan đến sự kiện Black Friday và đăng tải trên mạng xã hội Twitter. Quy trình dựa trên một hệ thống pipeline được thiết kế với các công cụ tiên tiến như Apache NiFi, Apache Kafka, và Apache Spark. Dữ liệu được thu thập từ Twitter API thông qua Apache NiFi, sau đó được truyền tải và xử lý theo thời gian thực bằng Apache Kafka. Spark được sử dụng để thực hiện các bước tiền xử lý, chuyển đổi dữ liệu và chuẩn bị cho việc huấn luyện mô hình Text Generation. Mô hình này, dựa trên các kiến trúc học sâu như LSTM hoặc GPT, được huấn luyện để tạo ra các bài đăng có nội dung sáng tạo và phù hợp với chủ đề Black Friday. Hệ thống cũng tích hợp API Twitter để tự động hóa việc đăng bài, giúp tăng hiệu quả tiếp cận khách hàng. Kết quả thực nghiệm cho thấy hệ thống có khả năng tạo nội dung với chất lượng cao và tiết kiệm đáng kể thời gian cũng như công sức. Ngoài ra, còn cung cấp tính linh hoạt trong việc tùy chỉnh các chiến lược tiếp thị và có tiềm năng mở rộng sang nhiều ứng dụng khác trong lĩnh vực xử lý ngôn ngữ tự nhiên. Đây là một bước tiến quan trọng trong việc ứng dụng công nghệ AI để giải quyết các bài toán thực tế, đặc biệt trong lĩnh vực tiếp thị số.

# 2.Giới thiệu (introduction)

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.

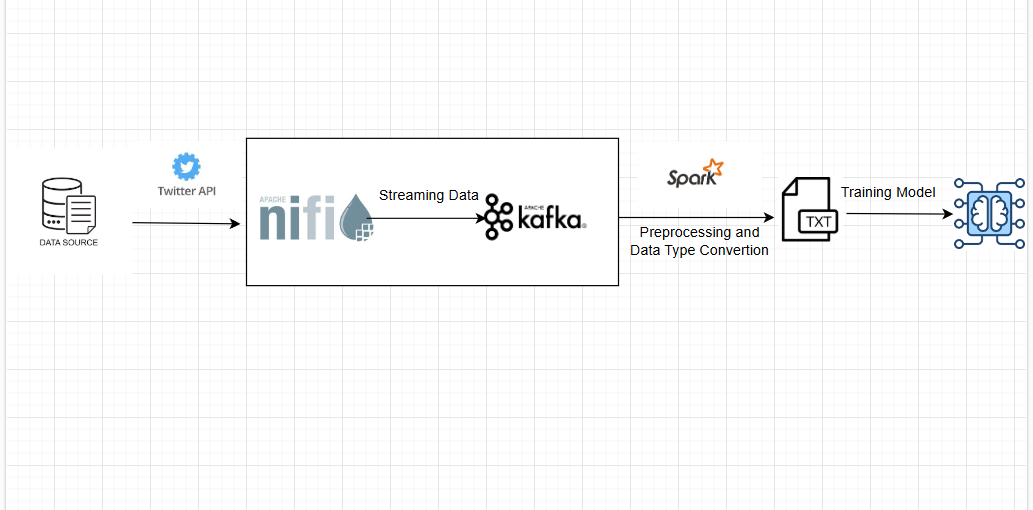
Việc sáng tạo nội dung thủ công đòi hỏi nhiều thời gian, công sức và sự sáng tạo, điều này đôi khi làm giảm hiệu suất hoặc không đáp ứng kịp thời những thay đổi liên tục của thông tin. Để khắc phục, sự phát triển của trí tuệ nhân tạo, đặc biệt là các mô hình Text Generation, đã mang lại một giải pháp 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, phân tích cảm xúc (sentiment analysis), 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.

# 3.Sơ đồ kiến trúc dự án



# 4.Dữ liệu (data)

## 4.1.Dữ liệu thu nhập từ Twitter

Trong chủ đề này chúng ta sẽ sử dụng **Twitter API** để truy cập và tương tác với dữ liệu từ Twitter. Dữ liệu này bao gồm các tweet, người dùng, xu hướng (trends) và nhiều thông tin khác liên quan đến hoạt động trên Twitter.

Dữ liệu nhận được từ Twitter là một JSON object chứa các thông tin chi tiết về một tweet bao gồm các siêu dữ liệu liên quan đến tweet, video và thông tin về người dùng.

Cấu trúc dữ liệu bao gồm:

**Thông tin chung về Tweet**:

"tweet\_id": ID duy nhất của tweet

"creation\_date": Ngày và giờ tạo tweet

"text": Nội dung của tweet #sport

"language": Ngôn ngữ của tweet

"favorite\_count": Số lượt thích tweet đã nhận

"retweet\_count": Số lượt retweet

"reply\_count": Số lượt trả lời

"quote\_count": Số lượt tweet đã được trích dẫn

"views": Số lượt xem của tweet

"expanded\_url": URL đầy đủ của tweet

"source": Nền tảng từ đó tweet được đăng tải

**Thông tin về Video và Media**:

"media\_url": Danh sách URL của hình ảnh đại diện của video

"video\_url": Danh sách các URL của các phiên bản video với các mức bitrate khác nhau:

URL m3u8 dành cho streaming

URL MP4 với các bitrate khác nhau như 632000, 950000, và 2176000.

"extended\_entities": Thông tin chi tiết về media trong tweet:

"media": Danh sách các media đính kèm trong tweet.

"type": Loại media (video).

"sizes": Các kích thước có sẵn của media (large, medium, small, thumb).

"video\_info": Thông tin chi tiết về video như tỷ lệ (aspect\_ratio), thời lượng (duration\_millis), và các phiên bản video (variants) với bitrate khác nhau.

**Thông tin về người dùng**:

"user\_id": ID của người dùng

"username": Tên đăng nhập của người dùng

"name": Tên hiển thị của người dùng

"follower\_count": Số người theo dõi

"following\_count": Số người dùng đang theo dõi

"favourites\_count": Số lượng tweet đã đánh dấu yêu thích

"is\_private": Kiểm tra tài khoản có ở chế độ riêng tư (null có nghĩa là không có thông tin).

"is\_verified": Người dùng có tick xanh xác thực không

"is\_blue\_verified": Người dùng có xác thực Blue Verified không

"location": Vị trí của người dùng

"profile\_pic\_url": URL ảnh đại diện của người dùng.

"profile\_banner\_url": URL banner của người dùng.

"description": Phần mô tả của người dùng.

"external\_url": URL bên ngoài liên kết với hồ sơ của người dùng.

"number\_of\_tweets": Số lượng tweet người dùng đã đăng

"bot": Xác định người dùng có phải bot không

"timestamp": Thời điểm tạo tài khoản (dưới dạng UNIX timestamp).

**Các trường khác**:

"retweet": Kiểm tra tweet có phải là retweet không

"in\_reply\_to\_status\_id": ID của tweet mà tweet này trả lời

"quoted\_status\_id": ID của tweet được trích dẫn

"binding\_values": Các giá trị liên kết

"conversation\_id": ID cuộc hội thoại chứa tweet này.

"retweet\_status": Trạng thái retweet

"quoted\_status": Trạng thái trích dẫn

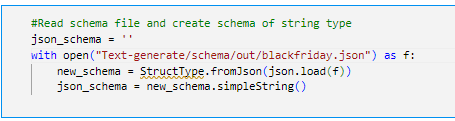
"bookmark\_count": Số lượt đánh dấu của tweet

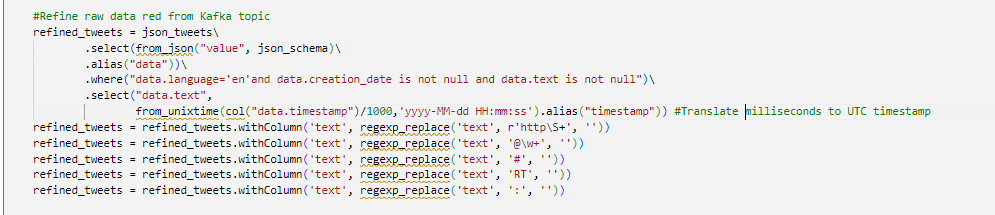
"community\_note": Ghi chú cộng đồn

## 4.2.Dữ liệu để training model

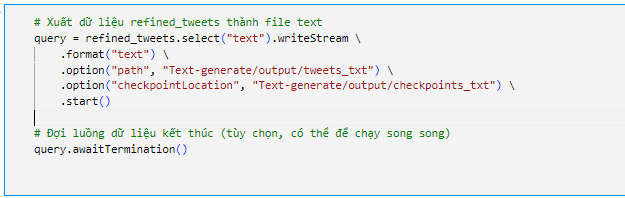
Sau khi thu thập được file JSON chứa dữ liệu Tweets, việc tiền xử lý là một bước quan trọng để chuẩn bị dữ liệu cho các bước tiếp theo. Trong trường hợp này, mục tiêu là **chỉ giữ lại nội dung của đoạn Tweets (thuộc tính “text”)** và chuyển đổi dữ liệu sang định dạng file .txt để dễ dàng sử dụng trong việc training mô hình.

**Các bước tiền xử lý:**

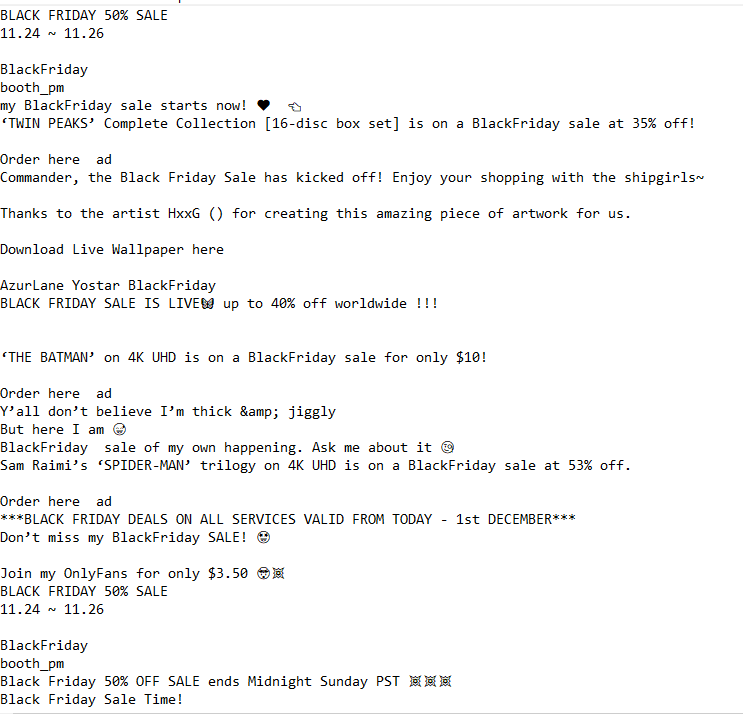
* **Đọc file JSON:**
* **Trích xuất nội dung Tweets:**Chỉ giữ lại trường text trong mỗi đối tượng Tweet.
* **Làm sạch dữ liệu văn bản:**
  + Loại bỏ các ký tự đặc biệt, liên kết (URLs), và hashtags.
  + Chuẩn hóa văn bản (ví dụ: chuyển toàn bộ chữ viết thường).



* **Lưu trữ dữ liệu vào file .txt:**Ghi toàn bộ nội dung văn bản (mỗi dòng là một Tweet) vào một file .txt để thuận tiện cho việc sử dụng với các mô hình NLP.



Và thu được kết quả như sau:



# 5.Phương pháp (methods)

Để giải quyết bài toán sinh văn bản dựa trên chuỗi ký tự, nhóm đã lựa chọn thuật toán **LSTM (Long Short-Term Memory)**, một biến thể nổi bật của mạng hồi quy tuần hoàn (**RNN - Recurrent Neural Network**). LSTM được thiết kế đặc biệt để xử lý các chuỗi dữ liệu dài bằng cách giải quyết vấn đề **vanishing gradient** – một hạn chế thường gặp trong các RNN truyền thống. Thông qua cơ chế lưu trữ và loại bỏ thông tin linh hoạt, LSTM có khả năng ghi nhớ các mối quan hệ dài hạn giữa các phần tử trong chuỗi, giúp tối ưu hóa hiệu quả khi xử lý dữ liệu tuần tự.

Thuật toán này đặc biệt phù hợp trong bài toán **xử lý ngôn ngữ tự nhiên (NLP)**, nơi ngữ cảnh đóng vai trò quan trọng để hiểu và tạo ra chuỗi văn bản có nghĩa. Trong trường hợp này, LSTM không chỉ học cách dự đoán ký tự tiếp theo mà còn đảm bảo tính liên kết giữa các ký tự để sinh ra một văn bản hoàn chỉnh, tự nhiên.

Một unit trong mạng LSTM được cấu trúc với ba cổng chính, mỗi cổng đóng vai trò kiểm soát luồng thông tin vào và ra, bao gồm:

**Cổng quên (Forget Gate)**:

Cổng này có nhiệm vụ lựa chọn những thông tin liên quan và loại bỏ những thông tin không liên quan. sau khi chọn thông tin liên quan, nó sẽ được chuyển qua cổng đầu vào.

Đầu tiên, thông tin từ trạng thái hiện tại và trạng thái ẩn trước đó được chuyển qua chức năng kích hoạt. đây sẽ là hàm kích hoạt sigmoid. Giá trị trả về của hàm kích hoạt sigmoid trong khoảng từ 0 đến 1.a gần bằng 0 nghĩa là thông tin hiện tại sẽ bị bỏ qua nếu không nó sẽ được chuyển qua cổng đầu vào.



Trong đó:

*Vector chỉ số quên, giá trị nằm trong khoảng [0, 1]*

: Trạn thái ẩn từ bước trước  
:  Đầu vào hiện tại

Cổng vào (Input gate)

Cổng này có nhiệm vụ thêm thông tin vào mô hình bằng cách sử dụng hàm kích hoạt sigmoid. Sử dụng hàm kích hoạt tanh sẽ tạo ra một mảng thông tin được truyền qua cổng đầu vào. mảng thông tin chứa các giá trị từ -1 đến 1 và hàm sigmoid lọc và duy trì thông tin nào sẽ được thêm vào mô hình và thông tin nào nên loại bỏ.

A group of black text

Description automatically generated

Trong đó:

: Vecto quyết định mức độ lưu giữ thông tin

tanh: Hàm kích hoạt hyperbolic tangent

: Trạng thái ô nhớ được đề xuất, biểu diễn thông tin mới

**Cổng ra (Output)**

Cổng đầu ra chịu trách nhiệm tạo ra các trạng thái ẩn tiếp theo cùng với các trạng thái ô được chuyển sang bước thời gian tiếp theo. Nó tạo ra một trạng thái ẩn bằng cách sử dụng hàm kích hoạt và giá trị của nó nằm trong khoảng từ -1 đến 1

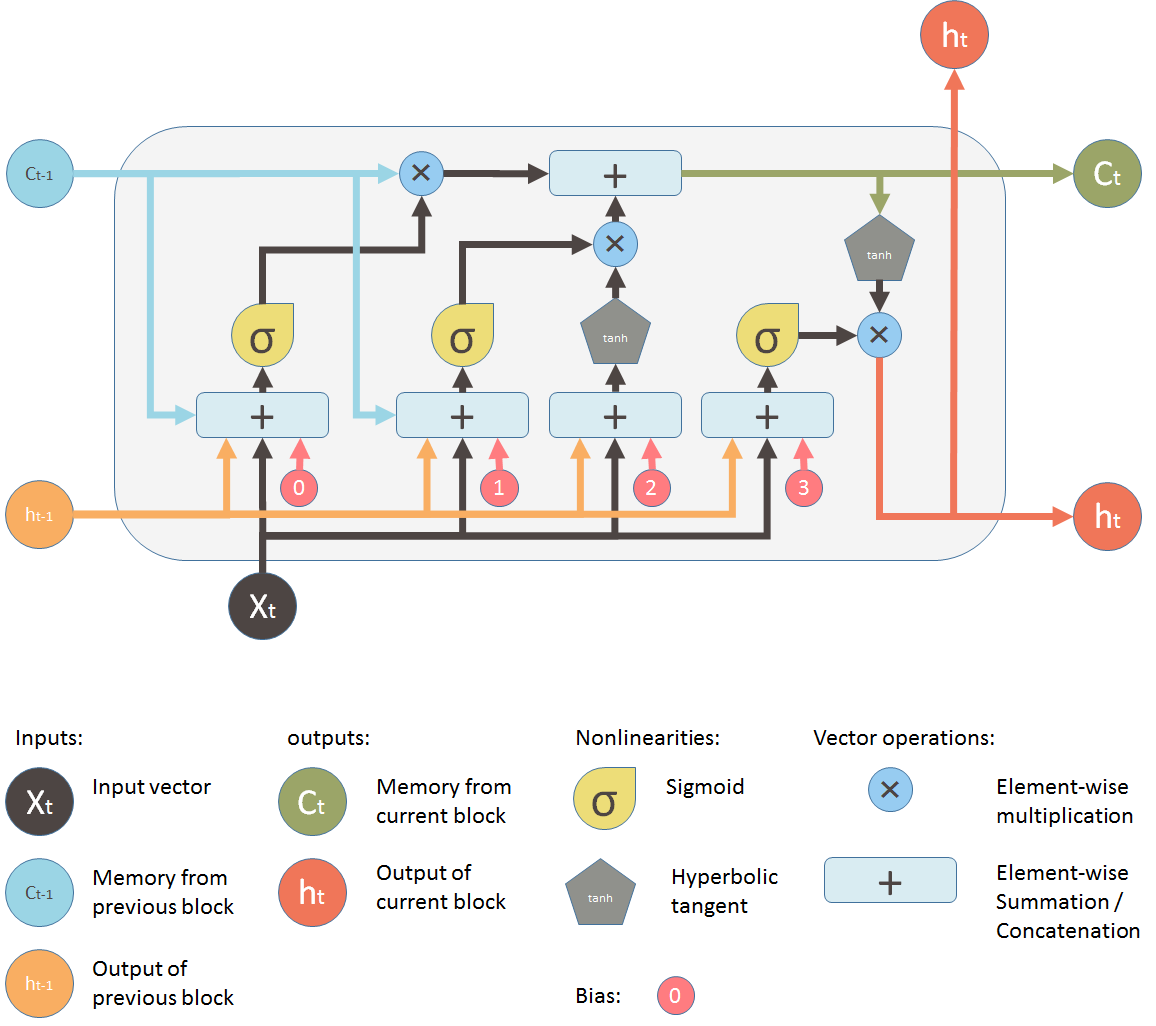
A math equations with black text

Description automatically generated with medium confidence

Trong đó:

: vecto kiểm soát đầu ra

ht: Trạn thái ẩn tại thời điểm hiện tại



**Hàm mục tiêu (Categorical Cross-Entropy)**

Dùng để đo lường sự khác biệt giữa phân phối xác suất thực tế và dự đoán

Trong đó:

yi: Vector nhãn thực tế (one-hot encoding).

: Xác suất dự đoán cho từng lớp.

**Thuật toán tối ưu hóa (Adam Optimizer**): Adam kết hợp giữa Momentum và RMSProp, giúp tăng tốc độ hội tụ và tối ưu hóa trọng số hiệu quả. Điều này làm giảm yêu cầu tinh chỉnh siêu tham số trong quá trình huấn luyện.

**Lợi ích của việc sử dụng LSTM**

1. **Ghi nhớ ngữ cảnh dài hạn:** LSTM có khả năng lưu giữ thông tin quan trọng trong các chuỗi dài, giúp tăng độ chính xác khi dự đoán chuỗi văn bản.
2. **Xử lý thông tin linh hoạt:** Các cơ chế cổng giúp mô hình lọc và lưu trữ thông tin cần thiết, giảm thiểu tình trạng quá tải dữ liệu.
3. **Hội tụ nhanh chóng:** Adam Optimizer giúp mô hình đạt hiệu suất cao mà không cần điều chỉnh siêu tham số phức tạp.

# 6.Thực nghiệm, kết quả, và thảo luận

## 6.1.Cấu hình thực nghiệm

**Quy trình tổng quát**

Hệ thống được thiết kế dựa trên pipeline tích hợp các công cụ hiện đại như Apache Kafka, Apache Spark, và mô hình LSTM. Pipeline này không chỉ giúp xử lý dữ liệu theo thời gian thực mà còn đảm bảo tính chính xác và hiệu quả trong việc sinh văn bản phù hợp với chiến dịch quảng cáo Black Friday.

* **Giai đoạn đầu tiên** bắt đầu bằng việc sử dụng Apache Kafka để thu thập dữ liệu từ Twitter API. Kafka được sử dụng như một trung tâm lưu trữ và xử lý dữ liệu streaming, cho phép dữ liệu luân chuyển liên tục từ nhiều nguồn khác nhau. Dữ liệu được truyền từ Kafka sang Apache Spark, nơi xử lý chính diễn ra.
* **Trong giai đoạn thứ hai**, Apache Spark đảm nhiệm nhiệm vụ xử lý dữ liệu thời gian thực bằng cách thực hiện các bước như chuẩn hóa, trích xuất schema, và chuyển đổi dữ liệu thô thành dạng có cấu trúc. Spark Streaming đảm bảo dữ liệu đầu vào được xử lý một cách liên tục và không bị gián đoạn, nhờ vào tính năng checkpoint và fault tolerance (khả năng phục hồi lỗi).
* **Giai đoạn cuối cùng** là sử dụng dữ liệu đã được chuẩn bị để huấn luyện mô hình LSTM. Với khả năng học tập từ các chuỗi ký tự và mối liên kết giữa chúng, mô hình LSTM sinh ra các bài viết sáng tạo, phù hợp với chiến lược tiếp thị Black Friday.

Nhờ thiết kế này, hệ thống đạt được sự tích hợp liền mạch giữa các công cụ, giảm thiểu sai sót và đảm bảo tính đồng bộ trong toàn bộ quy trình.

### 6.1.1.Các công cụ và cấu hình

**Apache Kafka**

Apache Kafka đóng vai trò như một trung tâm xử lý dữ liệu streaming. Nó cho phép dữ liệu được thu thập và truyền đi một cách nhanh chóng. Các cấu hình cụ thể như kafka.bootstrap.servers đảm bảo Kafka kết nối với Spark và các nguồn dữ liệu, trong đó dữ liệu được lấy từ topic "blackfriday". Đây là yếu tố quan trọng trong việc duy trì dữ liệu liên tục.

**Apache Spark**

Apache Spark đóng vai trò là công cụ xử lý chính trong pipeline. Spark không chỉ xử lý dữ liệu mà còn chuẩn bị các tệp văn bản và trích xuất schema, đảm bảo rằng dữ liệu đầu vào luôn ở trạng thái tốt nhất trước khi đưa vào mô hình LSTM. Một số bước quan trọng bao gồm:

* **Chuẩn hóa dữ liệu:** Đảm bảo tính đồng nhất của dữ liệu, loại bỏ các ký tự không cần thiết.
* **Trích xuất schema:** Lưu schema dưới dạng JSON để tái sử dụng và duy trì sự nhất quán trong các bước xử lý kế tiếp.

**Mô hình LSTM**

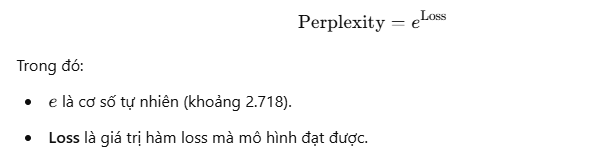
LSTM được sử dụng để huấn luyện trên dữ liệu đã qua xử lý, với các siêu tham số chính như:

* **Learning rate:** 0.001, giúp tốc độ hội tụ ổn định mà không gây dao động lớn.
* **Batch size:** 64, cân bằng giữa tốc độ tính toán và mức độ học của mô hình.
* **Dropout:** 0.2, giảm thiểu hiện tượng overfitting.
* **Epochs:** 80, đảm bảo mô hình học đầy đủ các mối liên hệ phức tạp trong dữ liệu.

## 6.2.Phương pháp đánh giá

Mô hình LSTM được đánh giá bằng:

* **Hàm Loss:** Là chỉ số đánh giá mức độ sai lệch giữa dự đoán của mô hình và giá trị thực tế. Trong trường hợp này, mô hình LSTM sử dụng hàm **categorical cross-entropy** để tính toán loss, giúp đo lường độ chính xác của dự đoán phân loại đa lớp.
* **Perplexity:** Là chỉ số đo lường khả năng dự đoán của mô hình ngôn ngữ, phản ánh mức độ "khó khăn" mà mô hình gặp phải khi dự đoán từ hoặc ký tự tiếp theo. Perplexity thấp hơn cho thấy mô hình dự đoán chính xác hơn, trong khi giá trị cao cho thấy mô hình đang gặp khó khăn trong việc hiểu ngữ cảnh của dữ liệu.



* **Perplexity thấp** (gần 1) nghĩa là mô hình đang dự đoán tốt và ít "ngạc nhiên" trước dữ liệu.
* **Perplexity cao** (lớn hơn 10, có thể đến 100 hoặc cao hơn) chỉ ra rằng mô hình không hiểu tốt ngữ cảnh hoặc dữ liệu đầu vào, và do đó dự đoán của nó kém chính xác hơn.

Ví dụ:

* Nếu **perplexity = 1**, mô hình dự đoán hoàn hảo, tức là nó không gặp phải khó khăn gì khi dự đoán từ tiếp theo.
* Nếu **perplexity = 10**, mô hình có độ chính xác trung bình nhưng vẫn chưa tối ưu hoàn toàn.
* Nếu **perplexity = 100**, mô hình gặp nhiều khó khăn trong việc dự đoán chính xác.

## 6.3.Kết quả thực nghiệm

**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.

## 6.4.Phân tích kết quả

### 6.4.1.Ưu điểm của hệ thống

* **Tích hợp liền mạch:**  
  Hệ thống kết hợp hiệu quả Apache Kafka và Spark Streaming, cho phép xử lý dữ liệu thời gian thực với độ trễ thấp. Tính năng fault tolerance của Spark đảm bảo dữ liệu không bị mất ngay cả khi có lỗi xảy ra.
* **Khả năng tự động hóa:**  
  Pipeline tự động hóa toàn bộ quy trình từ thu thập dữ liệu đến sinh văn bản, giúp tiết kiệm đáng kể thời gian và công sức so với quy trình thủ công. Hơn nữa, việc tích hợp với API Twitter cho phép đăng bài tự động, tăng hiệu quả chiến dịch tiếp thị.
* **Chất lượng văn bản:**  
  Các bài viết sinh ra có tính sáng tạo cao và phù hợp với chủ đề Black Friday. Điều này giúp nâng cao khả năng tiếp cận khách hàng và đạt được mục tiêu tiếp thị.

### 6.4.2.Hạn chế

* **Kích thước dữ liệu:**  
  Tập dữ liệu từ Twitter API có thể chưa đủ lớn để mô hình học được đầy đủ các ngữ cảnh phức tạp.
* **Phụ thuộc vào cấu hình:**  
  Pipeline phụ thuộc vào cấu hình Kafka và Spark. Bất kỳ sự cố nào trong cấu hình của một trong hai thành phần này đều có thể ảnh hưởng đến hiệu suất tổng thể.
* **Giới hạn của mô hình:**  
  Mô hình LSTM hiện tại chỉ xử lý tốt với chuỗi đầu vào cố định (100 ký tự). Điều này có thể làm giảm khả năng học các ngữ cảnh dài hơn hoặc phức tạp hơn.

# 7.Kết luận (conclusion)

* Dự án này đã chứng minh tiềm năng mạnh mẽ của công nghệ Text Generation trong việc tự động hóa quy trình tạo và đăng nội dung quảng cáo trên mạng xã hội. Bằng cách tích hợp các công cụ tiên tiến như Apache NiFi, Apache Kafka, Spark, và các mô hình học sâu, hệ thống không chỉ tiết kiệm thời gian mà còn tăng cường khả năng tiếp cận khách hàng một cách linh hoạt và hiệu quả. Kết quả thực nghiệm cho thấy mô hình có khả năng tạo ra các bài đăng với nội dung phù hợp, hỗ trợ tối ưu hóa chiến lược tiếp thị và nguồn lực.
* Tuy nhiên, dự án cũng gặp phải một số thách thức quan trọng, bao gồm việc nâng cao chất lượng nội dung tạo ra, đảm bảo tính chính xác và sáng tạo của các bài đăng, đồng thời giảm thiểu lỗi ngữ nghĩa và cải thiện khả năng diễn đạt tự nhiên. Một hướng nghiên cứu tiềm năng là tích hợp các mô hình ngôn ngữ tiên tiến như GPT-4 hoặc các kiến trúc Transformer hiện đại, giúp tăng cường hiệu suất và chất lượng đầu ra.
* Hơn nữa, hệ thống có thể được mở rộng ứng dụng sang các lĩnh vực khác như dịch vụ khách hàng tự động, hệ thống hỗ trợ nội dung, hoặc cá nhân hóa trải nghiệm người dùng. Việc kết hợp thêm các nguồn dữ liệu phong phú và cải thiện khả năng xử lý theo thời gian thực sẽ tiếp tục thúc đẩy tiềm năng ứng dụng của giải pháp Text Generation trong các bối cảnh công nghệ hiện đại. Với những cải tiến này, hệ thống hứa hẹn sẽ trở thành một công cụ hỗ trợ đắc lực cho nhiều doanh nghiệp trong bối cảnh nhu cầu về nội dung sáng tạo và tự động hóa ngày càng tăng cao.

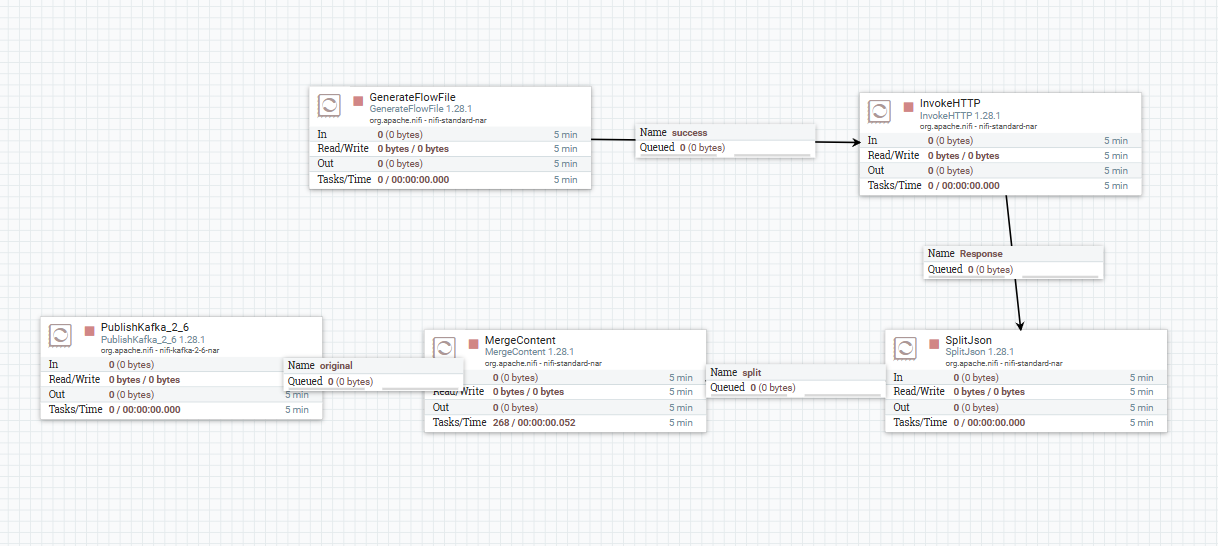
# 8.Phụ lục (appendices)

Do những thay đổi trong chính sách của Twitter, việc sử dụng API trực tiếp từ tài khoản cá nhân để thu thập dữ liệu đã trở nên khó khăn hơn. Cụ thể, người dùng phải trả thêm chi phí nếu muốn tiếp tục khai thác dữ liệu thông qua API của Twitter.

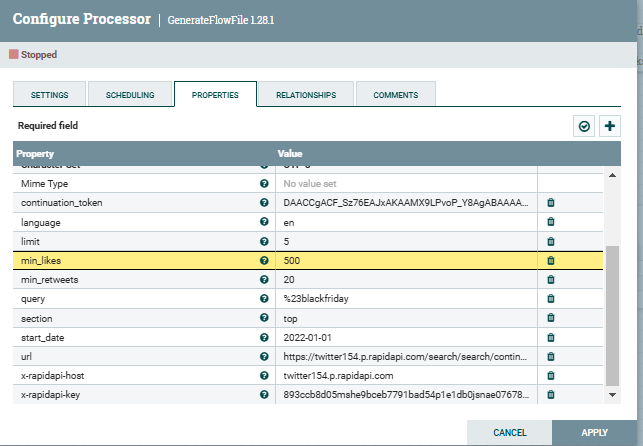
Để khắc phục vấn đề này, nhóm quyết định chuyển sang sử dụng **RapidAPI**, một nền tảng cung cấp API từ nhiều dịch vụ mạng xã hội khác nhau, trong đó bao gồm cả Twitter. RapidAPI mang lại lợi thế là dễ dàng truy cập vào nhiều API khác nhau thông qua một giao diện chung, đồng thời giúp giảm bớt các rào cản kỹ thuật liên quan đến việc cấu hình và tích hợp API riêng lẻ.

Song song đó, để tự động hóa quá trình thu thập và xử lý dữ liệu, nhóm sử dụng **Apache NiFi** – một công cụ mạnh mẽ trong việc xây dựng luồng xử lý dữ liệu (workflow). Apache NiFi cung cấp nhiều **processor** chuyên dụng, giúp hỗ trợ việc kết nối với các API trên RapidAPI, thu thập dữ liệu theo thời gian thực, và đẩy dữ liệu này vào các hệ thống lưu trữ hoặc phân tích. Với giao diện kéo-thả trực quan và khả năng mở rộng linh hoạt, NiFi giúp đơn giản hóa quá trình tích hợp, giảm thiểu sai sót và tiết kiệm thời gian triển khai.

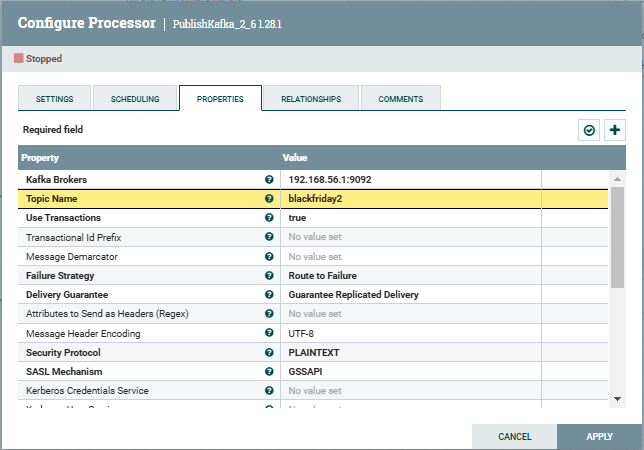
Đây là workflow được thiết lập trong Nifi



Tại generate FlowFile thiết lập filter chọn lọc dữ liệu



Sau khi dữ liệu được thu nhập sẽ được truyền lên Kafka thông qua processor PublishKafka tại đây sẽ tạo Producer

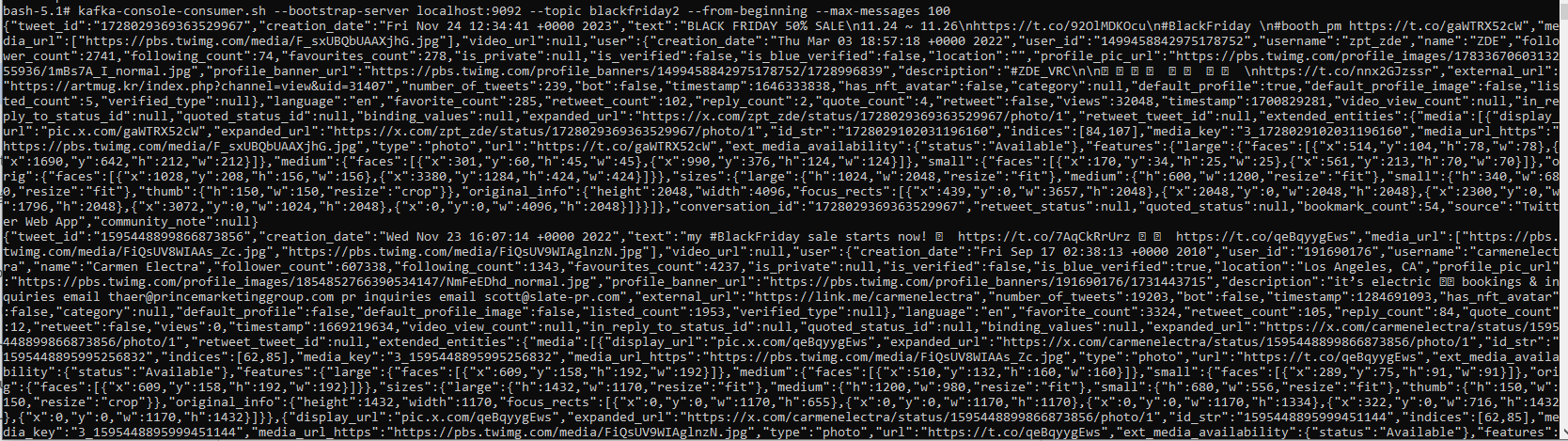


Và để nhận dữ liệu tại từ Producer ta sẽ tiến hành khởi tạo Consumer thông qua lệnh

“*kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic blackfriday2 --from-beginning --max-messages 100*”

\*Lưu ý: các topic phải trùng khớp với nhau và topic phải được tạo trước trong Kafka.

Ta thu được dữ liệu nhận được từ Consumer



Từ đây ta sẽ tiến hành tiền xử lý dữ liệu để đem dữ liệu thực hiện việc trainning model.

# 9.Đóng góp (contributions)

* Vũ: Thiết kế sơ đồ kiến trúc dự án, chuẩn bị dữ liệu (Dữ liệu thu nhập từ Twitter và Dữ liệu để training model), viết phụ lục.
* Đức: Xây dựng phương pháp, báo cáo thực nghiệm, kết quả và thảo luận.
* Đạt: Viết tóm tắt, giới thiệu, kết luận, xây dựng proposal
* Vy: Powerpoint, xây dựng milestone

# 10.Tham khảo

[1]. Twitter Developer. (n.d.). *Twitter API Documentation*. Retrieved

[2]. How To Use The Twitter API in 4 Easy Steps [Tutorial] By Team RapidAPI // <https://rapidapi.com/blog/how-to-use-the-twitter-api/>, April 20, 2021

[3]. Use NiFi to extract and parse data from HTTP endpoints,

<https://www.projectpro.io/recipes/use-nifi-extract-and-parse-data-from-http-endpoints-and-store-data-persistent-storage> , 23 Jan 2023

[4]. What is Text Generation? <https://www.datacamp.com/blog/what-is-text-generation?irclickid=zJVVmM35fxyKU4oxFsULOVy0UkCXRsUBbwZrVg0&irgwc=1&utm_medium=affiliate&utm_source=impact&utm_campaign=000000_1-390418_2-mix_3-all_4-na_5-na_6-na_7-mp_8-affl-ip_9-na_10-bau_11-Wildfire%20Systems&utm_content=ONLINE_TRACKING_LINK&utm_term=A358> , May 24, 2023

[5]. AIContentfy team, The Step-by-Step Text Generation Process Demystified, <https://aicontentfy.com/en/blog/step-by-step-text-generation-process-demystified>,

November 6, 2023

[6]. Vrunda Gadesha, Eda Kavlakoglu, What is text generation? <https://www.ibm.com/topics/text-generation>, 19 March 2024