# Kế hoạch thực hiện

# Bảng phân công công việc và đánh giá

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **MSSV** | **Thành viên** | **Danh sách công việc** | **Mức độ hoàn thành** | **Tổng hợp đánh giá chéo** |
| 16520724 | Bùi Nguyên Mão |  |  |  |
| 2 | Cao Văn Nam |  |  |  |
| 3 | Thừa Lê Thanh Chương |  |  |  |
| 4 |  |  |  |  |

# Nội dung

## Giới thiệu đề tài:

Công nghệ thông tin hiện nay đã mở ra một thời kỳ phát triển mạnh mẽ của xã hội tri thức, nơi mà tất cả mọi người đều có thể tiếp cận dễ dàng với tin tức bất kể không, thời gian. Tuy nhiên, mặt trái của nó đó là sự phát tán các tin tức nhảm, sai lệch vẫn luôn hiện diện như một căn bệnh từng bước đẩy con người vào bước đường lạc lối. Với số lượng tin bài được xuất hiện mỗi giây, rất khó có một tổ chức hay các nhân nào có thể kiểm tra, rà soát được sự đúng đắn hay nội dung vì vậy đây là một vấn đề nhức nhối mà cần phải có biện pháp để giải quyết.

Sự phát triển của machine learning và natural language processing như một vị cứu tinh cho đứa con chiên đang lạc lối. Với những công nghệ này, con người có thể giải quyết được một phần vấn đề trên.

Đề tài này chỉ là một bước tìm hiểu nhỏ nhưng sẽ mang lại một cái nhìn rõ hơn cho mọi người về cách mà NLP đã và đang tiếp tục hoạt động, phát triển để góp phần xây dựng một môi trường dữ liệu trong sạch.

*Bài toán đặt ra:*

kenh12 sau khi nhận đươc nhiều báo cáo vì số lượng tin bài giả đang gia tăng nghiêm trọng trên mạng xã hội của mình đã quyết đinh phải kiểm duyệt tất cả tin tức bài đăng, nhưng do số lượng tin bài nhận được mỗi ngày quá nhiều, bộ phận kiểm tra gồm 100 nhân viên làm ngày làm đêm vẫn không thể hoàn thành được nhiệm vụ. Một cuộc họp khẩn giữa 12 giờ đêm đã được diễn ra, cuối cùng cũng đưa ra một phương pháp có thể giải quyết được vấn đề trên.

*Giải pháp:*

* Sử dụng máy tính để làm việc kiểm duyệt thay cho con người.
* Xây dựng bộ dữ liệu lớn để máy phân loại được các bài viết theo tiêu chí các bài sử dụng các từ tương đồng nhau thường có mục đích như nhau.
* Để tối ưu chi phí và nguồn lực hiện có, tạm thời chỉ kiểm tra với tiêu đề của bài viết.

## Dữ liệu - Phân tích - Hướng giải quyết:

### Bộ dữ liệu

Đề tài sử dụng bộ dữ liệu tiêu đề tin tức ở nguồn: [*https://www.kaggle.com/rmisra/news-headlines-dataset-for-sarcasm-detection*](https://www.kaggle.com/rmisra/news-headlines-dataset-for-sarcasm-detection)

Thông tin bộ dữ liệu:

* Dữ liệu được thu thập ở 2 trang: *theonion* chuyên viết các bài châm biếm, mỉa mai và *huffingtonpost* chuyên viết tin chính thống.
* Gồm 26709 dòng dữ liệu kiểu JSON, có các thông tin: link bài viết, tiêu đề và nhãn (có phải tin bài mỉa mai)
* Các thông tin tiêu đề đều bằng tiếng Anh.

### Hướng giải quyết

Bài toán trở thành một bài toán phân lớp dữ liệu ngôn ngữ tự nhiên. Máy tính chỉ có thể xử lý với các con số, nên cần có bước đầu tiên là bước chuyển dữ liệu chữ sang số.

Có một số kỹ thuật để giải quyết vấn đề trên nhưng ở đây, chúng ta sẽ chọn cách đơn giản là Bag of Words. Kỹ thuật này sẽ xây dựng

## Tiền xử lý

## Huấn luyện với naive bayes

## Kết quả, đánh giá

## Hướng phát triển

## Bài toán:

# Tài liệu tham khảo