**Bảng đề xuất đồ án**

I/ INPUT/OUTPUT

Input: một câu tiêu đề của một bài báo, link tin đó

Outpur: phân loại tiêu đề đó coa khả năng là tin giả hoặc thật.Nếu là tin giả gán nhãn là 1, ngược lại gán nhãn 0.

II/ NGỮ CẢNH

Ngày nay, với quyền tự do ngôn luận, việc dăng một bài báo không qua kiểm duyệt, hay kiểm duyệt không chặt chẽ dễ dẫn tới những bài báo với những tin không có thật nhầm câu like. Vì thế Fake news classification sẽ giúp lọc những bài báo có tin ảo nhầm giúp người đọc có thể xác định những tin chính thống và tăng chất lượng của những nhà báo.

II/ Cách xây dựng DATASET

DATASET bao gồm 4 cột: Url, Headline, Body, Label

DATASET được lấy được 2 loại trang web:

* Những trang web chính thống có độ uy tín cao sẽ là bộ dataset được gán nhãn là 0 với những thông tin thật.
* Những trang web còn lại là những trang web có nhiều thông tin ảo hay thường đưa những thông tin không thật.

Tất cả những trang web trên được nhóm tìm là chọn lọc trong danh sách những trang web chính thống và trang web tin ảo do Wikipedia liệt kê.

Sau khi chọn lọc trang web, nhóm sẽ vào từng bài báo của các trang đó và lấy link, headline, và thông tin trong bài báo. Với những thông tin lấy được từ trang web chính thống sẽ được nhãn là 0, ngược lại là 1.

III/ PIPELINE

Import Library

Data preprocessing

Text preprocessing

Building Model and result

IV/ CONCLUSION