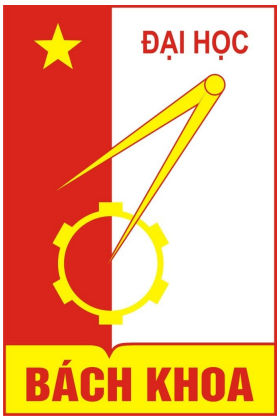
Đại học Bách Khoa Hà Nội

Viện Công nghệ thông tin và Truyền thông



Báo cáo project 1

Phân biệt thông tin đáng tin cậy trên mạng xã hội

Giảng viên hướng dẫn:
PGS.TS. Lê Thanh Hương

Họ tên sinh viên: Phan Việt Hoàng,
MSSV: 20180086
Lớp: KSTN-CNTT-K63

Hà Nội - Ngày 17 tháng 1 năm 2021

Mục lục

1 Lời mở đầu

Tin giả , còn được gọi là tin rác hoặc tin tức giả mạo, là một loại hình báo chí hoặc tuyên truyền bao gồm các thông tin cố ý hoặc trò lừa bịp lan truyền qua phương tiện truyền thông tin tức truyền thống (in và phát sóng) hoặc phương tiện truyền thông xã hội trực tuyến. Thông tin sai lệch thường được các phóng viên trả tiền cho các trang đăng tin để được đăng các tin tức này, một thực tế phi đạo đức được gọi là báo chí trả tiền. Tin tức kỹ thuật số đã mang lại và tăng việc sử dụng tin tức giả, hoặc báo chí màu vàng (yellow journalism). Tin tức sau đó thường được nhắc lại là thông tin sai trên phương tiện truyền thông xã hội nhưng đôi khi cũng tìm được đường đến những phương tiện truyền thông chính thống.

Tin giả được viết và xuất bản thường là với mục đích đánh lừa nhằm gây thiệt hại cho một cơ quan, thực thể hoặc người, và/hoặc đạt được về mặt tài chính hoặc chính trị, nó thường sử dụng lối viết giật gân, không trung thực hoặc dùng các tiêu đề bịa đặt để tăng lượng độc giả. Tương tự, các câu chuyện và tiêu đề bẫy để nhấn chuột vào kiếm doanh thu quảng cáo từ hoạt động này.

Trên cơ sở đó, cuộc thi REINTEL nhằm xác định các mẩu thông tin được chia sẻ trên các trang mạng xã hội (MXH), là đáng tin cậy hay không đáng tin cậy. Với sự bùng nổ nhanh chóng của MXH, ví dụ như Facebook, Zalo hoặc Lotus, có khoảng 65 triệu người dùng Việt Nam với mức tăng trưởng hàng năm là 2,7 triệu trong năm gần đây, theo báo cáo của Digital 2020. Mạng xã hội trở thành phương tiện thiết yếu để người dùng không chỉ kết nối bạn bè mà còn tự do sáng tạo, chia sẻ thông tin đa dạng như tin tức. Mặc dù vậy, một số người dùng có xu hướng phát tán thông tin không đáng tin cậy cho mục đích cá nhân của họ ảnh hưởng đến xã hội trực tuyến. Việc phát hiện xem tin tức lan truyền trong SNS là đáng tin cậy hay không đáng tin cậy đã thu hút được sự chú ý đáng kể gần đây. Do đó, cuộc thi này nhắm đến việc xác định thông tin được chia sẻ trên các nền tảng MXH của Việt Nam. Nó tạo cơ hội cho những người tham gia quan tâm đến vấn đề này, đóng góp kiến thức của mình để cải thiện xã hội trực tuyến vì lợi ích xã hội.

2 Tổng quan về bài toán REINTEL

2.1 Bộ dữ liệu

Mỗi mẫu dữ liệu bao gồm 6 thuộc tính chính như sau:

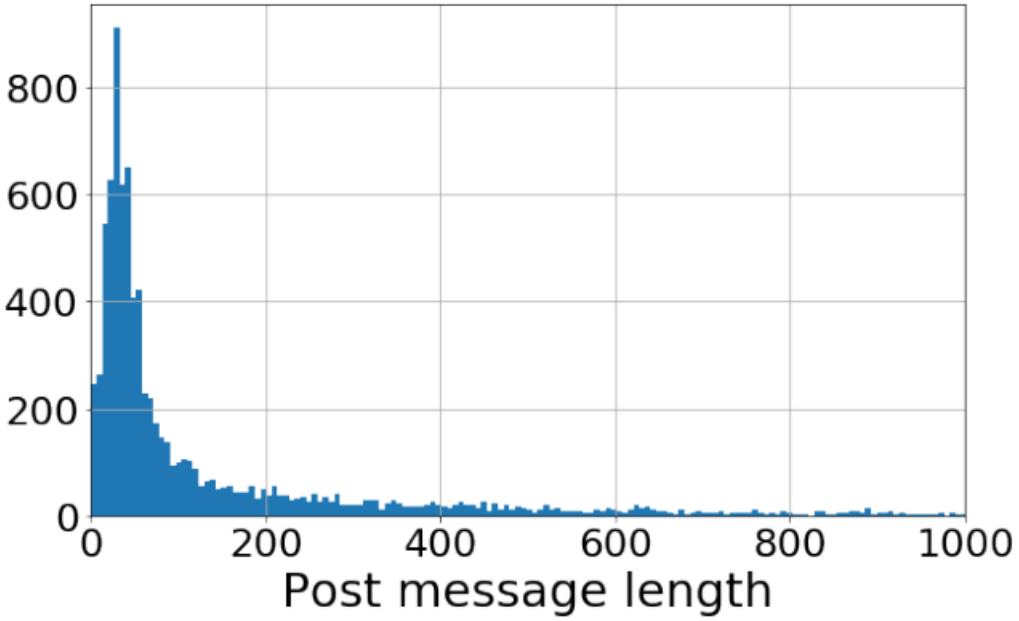
- id: id duy nhất cho một bài đăng trên MXH. Lưu ý rằng hình ảnh của mỗi bài đăng (nếu có) có thể truy cập được thông qua id này.
- user_name: id ẩn danh của chủ sở hữu
- post_message: nội dung văn bản của tin tức
- timestamp_post: thời điểm tin tức được đăng
- num_like_post: số lượt thích tin tức nhận được
- num_comment_post: số lượng bình luận mà tin tức nhận được
- num_share_post: số lượt chia sẻ mà tin tức nhận được
- label: nhãn, thể hiện tin tức là có thể đáng tin cậy hay không
  - 1: không đáng tin cậy
  - 1: 0: đáng tin cậy

Với tính chất của một cuộc thi về học máy, cuộc thi cung cấp 3 bộ dữ liệu với tổng cộng 6000 mẫu cho training và 2000 cho testing trong suốt cuộc thi, được trích nguồn sau đây:

- Warmup Phase:
  Train set
  Test set
- Public Phase:
  Train set
  Train images
  Test set
  Test images
- Private Phase:
  Test images
  Test set

2.2 Một số thống kê khác

	train	public test	private test
number of examples	5165	1642	1646
average of posts length	164	148	164
number of posts have images	1287	494	508
number of duplicated posts	313	31	34
number of duplicated users	1464	497	367



2.3 Nhiệm vụ

Xác định một mẫu tin tức đề ra (cùng với các thông tin bổ trợ và hình ảnh đính kèm) là có thể tin cậy hay không
Do đó bài toán được quy về phân loại văn bản với 2 nhãn 0-1

2.4 Đánh giá

Người tham gia phải gửi tệp .zip có chứa results.csv. Tệp results.csv phải chứa kết quả theo thứ tự như bộ thử nghiệm ở định dạng sau:

id1	xác suất nhãn 1
id2	xác suất nhãn 2
id3	xác suất nhãn 3
...	...

Bài nạp gửi sẽ được đánh giá bằng cách sử dụng chỉ số ROC-AUC

3 Đề xuất giải pháp

Bidirectional Encoder Representations from Transformers là một mô hình học máy xử lý ngôn ngữ tự nhiên do Google phát triển. BERT được tạo và xuất bản vào năm 2018 bởi Jacob Devlin và các đồng nghiệp của ông từ Google. Nó có thể được sử dụng trong nhiều bài toán NLP như:

- Phân loại văn bản
- Question answering
- Dịch máy
- Nhận dạng thực thể có tên
- Tóm tắt văn bản
- ...

Không nghi ngờ gì nữa, BERT là một bước đột phá trong việc sử dụng học máy trong xử lý ngôn ngữ. Hơn thế nữa, nó là mã nguồn mở và cho phép tinh chỉnh nhanh chóng, phạm vi ứng dụng thực tế là rất lớn.

Do đó cách tiếp cận được đề ra trong phạm vi project 1 này sẽ được xây dựng xung quanh cốt lõi là một mô hình pretrained transformer, mà cụ thể là mô hình PhoBERT, bên cạnh đó là các kĩ thuật pretraining (domain-adaptation) cho dữ liệu báo chí và trích xuất, chọn lọc đặc trưng từ meta-data

4 Triển khai

4.1 Tiền xử lý dữ liệu

Bộ dữ liệu training sẽ được chia thành 5-folds để thuận tiện cho việc so sánh giữa các lần chạy thực nghiệm

4.1.1 Dữ liệu dạng văn bản

Sử dụng bộ tách chuyên dụng cho PhoBERT là VNCoreNLP để tách từ trước khi đưa vào mô hình BERT

Dữ liệu sau khi được tách sẽ tiếp tục được xử lý, loại bỏ hoặc thay thế các biểu tượng emoji, link html, các dấu câu thừa ...

4.1.2 Các trường dữ liệu khác

Khác với dữ liệu văn bản, các trường còn lại thường xuyên có các giá trị là nhiễu, NULL hoặc có giá trị UNKNOWN, một cách đơn giản để giải quyết vấn đề này là sử dụng imputer, điền các giá trị này là trung bình của các mẫu dữ liệu khác

Các dữ liệu từ trường khác này sẽ được sử dụng như là các thông tin bổ trợ để đưa vào mô hình cùng với các đoạn văn bản

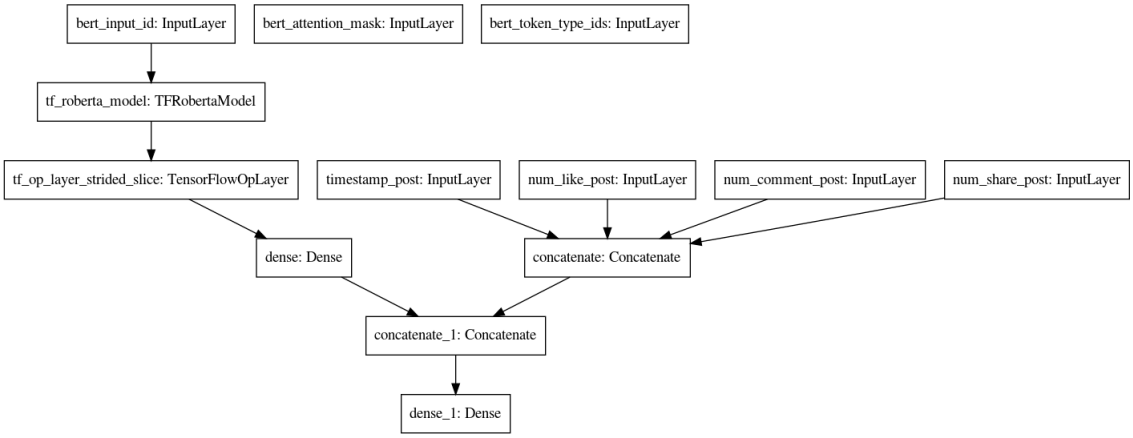
4.2 Mô hình

Tổng quan, dữ liệu đầu vào được chia làm 2 phần: văn bản và các trường phụ trợ, cách trích xuất đặc trưng để đạt được các vector biểu diễn là tương đối khác nhau

Sau cùng, các biểu diễn ẩn này được ghép lại và đưa qua mạng fully connected (một mô hình với Bert+CNN được sử dụng tuy nhiên lại cho kết quả tồi hơn) với activation là sigmoid

Sử dụng kĩ thuật domain-adaptaion, mô hình PhoBert được pretraining sử dụng Masked Language Model cho dữ liệu được crawl từ các trang mạng xã hội xoay quanh chủ đề covid và được review lại với mục đích cung cấp cho mô hình PhoBert có lượng dữ liệu lớn hơn (so với 6000 sample được cung cấp). Output của pre-training sẽ được dùng như là model Bert để encode cho văn bản

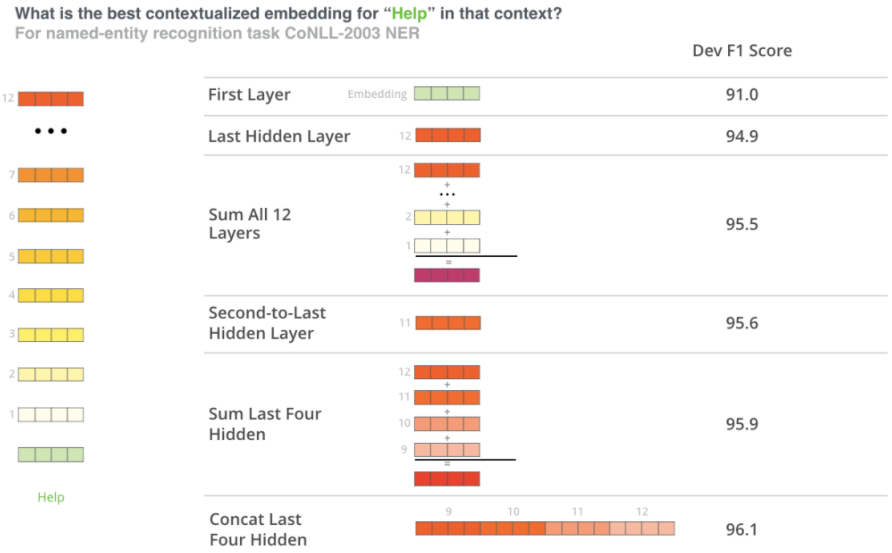
Kiến trúc tổng quan của mô hình như sau:



4.2.1 Dữ liệu dạng văn bản

Văn bản ban đầu, sau khi được tách thành các BPE tokens (token_id) truncate và padding nhờ vào bộ tách tử PhoBert sẽ được đưa vào mô hình cùng với attention_mask và token_type_id

Lấy ý tưởng từ paper gốc của paper BERT và kết quả của Bert-base trên share task CoNLL-2003 NER. Mô hình được đề xuất có thử nghiệm cả việc lấy embedding của token [CLS] và ghép 4 trạng thái ẩn cuối cùng của PhoBERT model. Kết quả thực nghiệm cho thấy việc lấy 4 trạng thái ẩn cho kết quả cao hơn trên nhiều seed khác nhau, do đó đây là cài đặt được sử dụng cho mô hình cuối cùng



Do đó sau khi đưa qua mô hình, văn bản ban đầu đã được encode thành vector biểu diễn. Trước khi đi tiếp vào huấn luyện, hãy đi qua cách encode được dùng cho các trường dữ liệu meta-data

4.2.2 Các trường dữ liệu khác

Các trường dữ liệu ví dụ như (số lượt thích, số bình luận, chia sẻ...) sau khi được tiền xử lý đã trờ thành các số thực, tuy nhiên việc đưa thẳng các số (hầu hết là các số nguyên dương có giá trị tương đối cao) này sẽ khiến cho mô hình khó hội tụ. Thậm chí sẽ ảnh hưởng đến kết quả nếu có sự chênh lệch lớn giữa 2 tập train-test

Do đó, trước khi đưa vào mô hình, các trường này được chuẩn hóa về gần giá trị 0

Biểu diễn ẩn của các trường dữ liệu này đạt được nhờ mạng encoder đơn giản bao gồm 1 lớp fully-connected

4.2.3 Huấn luyện mô hình

Tương tự như bài toán phân loại văn bản thông thưởng, huấn luyện mô hình cho bài toán REINTEL được thực hiện tương tự

Do kiến trúc khá lớn nên huấn luyện Bert được thực hiện song song trên nhiều GPU, cụ thể các thông số của mô hình như sau

Số fold	5
Độ dài văn bản	256
Batch size	24 (x 2 GPUS)
Epochs	5

GPU sử dụng là NVIDIA T4 với thời gian chạy 3 phút/epochs

5 Kết quả thực nghiệm

Dựa vào các mô hình được huấn luyện, kĩ thuật cuối cùng được sử dụng cho bài toán REINTEL là ensemble

5 mô hình được huấn luyện trên các folds khác nhau sẽ đưa ra dự đoán xác suất cho tập test, kết quả cuối cùng sẽ được tính dựa trên trung bình các dự đoán này

Kết quả khi submit trên Public Leaderboard của cuộc thi Reintel là 0.9405, xếp thứ 3 chung cuộc cho vòng này

Tuy nhiên do một số việc bận nên cá nhân em không nạp bài cho Private Test, dẫn đến không có kết quả cuối cùng

Toàn bộ source code cho cách tiếp cận này có thể được tìm thấy tại đây

6 Tổng kết

Bài toán phát hiện tin giả là một bài toán tương thú vị và mang nhiều ý nghĩa thực tiễn. VLSP 2020 cũng như BTC cuộc thi REINTEL đã mang đến một sân chơi vô cùng bổ ích, đây cũng là lần đầu em tham gia một cuộc thi NLP cho tiếng Việt

Trong suốt quá trình làm project 1 dưới sự hướng dẫn của cô Lê Thanh Hương, em đã thực sự học ra được nhiều kiến thức mới, cơ chế hoạt động của các mô hình State-of-the-art, cách áp dụng chúng cho bài toán của mình và tìm ra được những key-point chủ chốt giúp tăng được hiệu năng mô hình.

Bên cạnh đó, những sự góp ý của cô về các vấn đề technical khác ngoài việc xây dựng và tuning model đã giúp em có một cái nhìn vô cùng mới mẻ đối với các bài toán xử lý ngôn ngữ nói riêng và khoa học máy tính nói chung. Đó cũng chính là những hành trang vô cùng quan trọng cho những sinh viên ngành công nghệ thông tin.

Em xin chân thành cảm ơn cô đã cố vấn, giúp đỡ trong suốt quá trình hoàn thành đề tài môn học project 1 và một số bài toán khác trong học kì vừa qua!

7 Tài liệu tham khảo

1. PhoBERT
2. Transformers
3. BERT Word Embeddings Tutorial
4. K-Fold Cross Validation. Evaluating a Machine Learning
5. 6 Different Ways to Compensate for Missing Values In a Dataset
6. BERT
7. Universal Language Model Fine-tuning for Text Classification