Chia sẻ từ bạn **Long Lưu**:

Chào mọi người. Mình muốn chia sẻ những ý tưởng về bài toán này mà mình đã thử. Mình xin nói lẫn lộn Anh với Việt, vì một số từ mình không biết dịch ra ntn cho hợp lí.

Tổng quan:

- Tập data rất noisy. Lẫn lộn cả tiếng anh, tiếng việt, tiếng việt không dấu, HTML và cả teen code. Ví dụ từ "khôn- g" thì nó có cả "ko", "khg", "khong", "k".

- Rất nhiều từ bị lỗi dấu. Đơn giản như từ "lỗi" thì nó có thể thành "lô~ i".

- Bài toán này balance về class (số class 0 và số class 1 ngang nhau). Nếu CV tốt thì điểm LB sẽ tốt.

- Tập data rất ít. Tuy nhiên mình coi đây như là một thử thách nên mình không dùng thêm data. Ngoài ra tập data trên link nguồn không có phân sẵn ID như tập data đã được chuẩn bị. Để công bằng thì mình nên loại bỏ tập test ra khỏi tập data thêm kia, nhưng mình chưa thử.

Ý tưởng đã thử:

- Những thuật toán ML cơ bản như BoW và Tfidf đi kèm với Naive Bayes, Logistics Regression, SVM đã dễ dàng cho CV tới 86-89.

- Pretrain bộ Embedding sau đó bỏ vào model chạy tốt hơn hẳn việc vừa train Embedding vừa classify. Phương pháp 1 giúp CV mình lên tới 95 trong khi phương pháp 2 chỉ có 85-89. Tuy nhiên PP1 score lại không correlate với LB.

- Preprocessing giúp CV mình tăng từ 86 lên 90. Mình loại bỏ emoji, punctuation, HTML, dấu `\_`, etc.

- Không hiểu sao dùng Tokenizer => pad\_sequence => Embedding lại cho kết quả tốt hơn TextVectorization => Embedding, mặc dù theo lý thuyết 2 cách trên như nhau.

- CNN, LSTM, GRU kể cả Bidi đều cho kết quả xấp xỉ 90, chỉ bằng với các ML cơ bản.

- fastText rất tốt. Điểm cao nhất của mình hiện tại là dùng model fastText. Model này phải preprocess, nếu không có thể crash kernel.

- Mình chưa thử dùng bộ Embedding đã có sẳn bỏ vào model, tuy nhiên trong notebook tutorial mình chỉ đạt CV khoảng 87- 89.

Một số vướng mắc:

Mặc dù F1 là metric khá phổ biến trong Sentiment Analysis, tuy nhiên mình phải tune một hyperparam gần như...ít ai tune đó là cái Threshold mình cho nó là 1, còn lại là 0. Tune param này rất hên xui, nên mình nghĩ mình thích metric Log Loss hơn.

Các bạn có câu hỏi gì cứ mention mình nhé!

Chào mọi người. Mình chia sẻ bộ data của bài Foody này mà mình đã preprocess sơ bộ và đã augment thêm phần data bị lỗi.

Link notebook: https://www.kaggle.com/aeryss/vietai-nlp01-foody-augmentation

Link dataset: https://www.kaggle.com/aeryss/foody-preprocessed

Như mọi người thấy bộ data nó bị lỗi lúc nhập (hoặc lúc crawl), ví dụ từ bố bị ghi thành bô' . Mình tạo bằng tay một bộ mapper nó về tiếng việt thông thường (ở dưới comment).

Ngoài ra mình có dùng langid để phân nó ra 2 DF tiếng Việt và không phải tiếng Việt (trong đây hình như có cả tiếng Pháp, Nga và Anh). Mình giữ nguyên index và lưu nó dưới dạng pickle để load cho dễ, và nếu mọi người train riêng 2 model có thể concat nó và sort lại cho đúng format file submission.