**Thu thập dataset và xử lý dữ liệu phù hợp với dự án.**

1. SOCC corpus (SFU Opinion and Comments Corpus)

Dataset là một file csv bao gồm các trường thuộc tính article\_id,comment\_counter,title,globe\_url,url,comment\_text,is\_constructive,is\_constructive:confidence,toxicity\_level,toxicity\_level:confidence,did\_you\_read\_the\_article,did\_you\_read\_the\_article:confidence,annotator\_comments,expert\_is\_constructive,expert\_toxicity\_level,expert\_comments.

Ta sẽ giữ lại các trường là comment\_text, toxicity\_level, toxicity\_level:confidence. Bởi lẽ, ta vốn không quan tâm đến việc comment đó có constructive hay không (vì ta không bàn đến chủ đề của chúng, nếu xét theo tính constructive sẽ vô cùng chủ quan), mà chỉ cần biết là mức độ toxic của nó:

1 không toxic.

2 hơi toxic.

3 toxic.

4 rất toxic.

Ta hợp nhất hai giá trị 1 và 2 thành “bình thường”, 3 và 4 thành “toxic”. Dĩ nhiên vẫn còn một rào cản, đó là toxicity\_level không đơn thuần chỉ bao gồm 1 giá trị mà có thể có hai giá trị cách nhau bởi ‘\n’. Do đó, ta sẽ tách những record có đa trị thành nhiều record có comment\_text giống nhau, nhưng khác toxicity\_level. Đồng thời, ta nhận thấy chỉ số toxicity\_level:confidence biểu diễn độ tin cậy vào giá trị rời rạc của toxicity\_level, và vì vậy, ta sẽ biến nó thành một đặc trưng của điểm dữ liệu (tương ứng với một comment). Từ đó ta sẽ có:

Comment có label toxic hoặc bình thường và có thêm một đặc trưng là confidence giúp nó có thể phân biệt với các comment toxic khác.

Sẽ có hai comment có nội dung giống nhau nhưng hai label khác nhau, và chúng có lẽ vẫn sẽ có giá trị confidence khác nhau, một cách tốt để phân biệt giữa chúng.

Nhưng vẫn còn một vấn đề khác, ta hợp nhất 1, 2 và 3, 4. Dĩ nhiên ta cần phải gộp chung hai giá trị confidence lại với nhau (nếu có), đối với trường hợp comment có các label khác loại, ta chỉ đơn giản là phân tách chúng thành 2 comments với các label và các giá trị confidence tương ứng. Còn trường hợp comment có hai label cùng chủng loại (1 và 2 hoặc 3 và 4), ta sẽ sử dụng hệ thức sau:

New\_confidence\_value = max(old\_confidence\_value1, old\_confidence\_value2)

Trong đó thì old\_confidence\_value1, old\_confidence\_value2 tương ứng với các confidence value của mỗi label cùng chủng loại.

Bây giờ, sẽ có một vấn đề nữa, đó là thứ tự dữ liệu sẽ thay đổi trong khi sử dụng hàm concat của pandas. Nhưng đó có vẻ không phải là vấn đề quan trọng lúc bấy giờ.

Above is the first method, the second method is: Thay vì hợp nhất 1 và 2 -> normal, 3 và 4 -> toxic, ta lại có cách khác là 1 -> normal, 2, 3 và 4 ->. Trong khi mà 2 cũng được xem là mức độ toxic thấp nhất. Tuy nhiên, qua thực nghiệm, ta thấy cách thứ hai này không hiệu quả, nó khiến cho model bị suy giảm độ chính xác (từ 89% ở method 1 về khoảng 60% - 70%). Cách 2 này không hiệu quả.

Trong cách thứ ba, ta sẽ thử nghiệm xóa chỉ số confidence đi, nghĩa là ta phải tìm cách nào đó convert từ dữ liệu multi-label sang multi-class. Ta chỉ việc chọn label có confidence cao nhất thay vì tách nó ra làm hai và thêm confidence như là một trường dữ liệu mới.

1. DETOX corpus (comments from English Wikipedia talk pages).

Dataset là một bộ gồm ba file tsv chứa các tập dữ liệu đã được gắn nhãn tương ứng với mỗi file. File thứ nhất được gắn nhãn aggression (hung hang). File thứ hai được gắn nhãn attack (công kích). File thứ ba là toxicity (độc hại). Tập dữ liệu gồm các trường sau: rev\_id, comment, year, logged\_in, ns, sample, split.

Ta giữ lại trường comment (đổi tên thành text). Sau đó, gộp ba file lại thành một dataset, trong đó, thêm một trường dữ liệu mới là label, label = aggression nếu dòng dữ liệu đó từng thuộc file thứ nhất, label = attack nếu dòng dữ liệu đó từng thuộc file thứ hai, label = toxicity nếu dòng dữ liệu đó từng thuộc file thứ ba. Ta không có gì đảm bảo dữ liệu này là multi-class, vì có thể có cùng một dòng comment ở trong hai file khác nhau, tuy nhiên đó thật sự chưa phải là vấn đề lớn, hơn nữa là phương pháp xử lý dữ liệu vẫn chưa thật sử kỹ lưỡng đến mức như vậy, nghĩa là ta phải tạm chấp nhận chuyện hi hữu này có thể xảy ra.

Sau khi xử lý xong, ta có một file duy nhất là detox.csv lưu trữ tất cả dữ liệu từ ba file. Dĩ nhiên có một vấn đề khác, đó là làm sao để biết được khi nào comment là bình thường (vì chỉ có ba class là aggression, attack và toxic). Thực ra ta có thể declare một threshold, khi giá trị dự đoán thấp hơn threshold, có nghĩa là comment đó là bình thường.

1. HASOC corpus (Hate Speech & Offensive content Identification in Indo-European Languages).

Dataset là một bộ gồm sáu file csv tương ứng với ba ngôn ngữ English, German và Hindi và tập train và test. Mỗi file csv gồm các trường sau:

Tweet\_id, text, task1, task2, ID.

Trong đó task1 dường như lưu trữ nhãn nhị phân của comment đó (có annotated không), task2 là nhãn cụ thể của comment (profanity – thô tục, none – không có gì, offensive – công kích, hateful – ngôn ngữ thù địch). Trong bài toán của chúng ta (vì thuật toán GFMM hỗ trợ đa lớp, nên ta chọn task2 thay vì task1), ta sẽ chọn các trường text và task2, trong đó, ta đổi tên task2 thành label. Dữ liệu đã được phân sẵn theo dạng multi-class, do đó, việc xử lý dễ dàng hơn rất nhiều so với SOCC. Ta đã có sẵn tên lớp và là tên lớp duy nhất của mỗi dòng dữ liệu.

Sau khi xử lý xong, ta có tổng cộng 6 file csv tương ứng. Ta sẽ dùng các file train cho việc train dữ liệu (bao hàm cả việc dùng để tuning tham số and so on), các file test sẽ được dùng làm tập validate cuối cùng.

1. Trawling corpus (samples of comment from Twitter, Reddit, Wikipedia talk pages).

Do một số vấn đề, dataset chỉ bao gồm dữ liệu từ Twitter. Nhưng không sao cả.

Trong tập dữ liệu của chúng ta có một file csv, bao gồm các trường: count, hate\_speech, offensive, neither, class và tweet.

Theo quan sát thì class=0 tương ứng voi81 hateful (thù ghét), class=1 tương ứng với offensive(công kích), class=2 tương ứng với neither(trung lập) hay normal. Các trường tương ứng là hate\_speech, offensive và neither tồn tại để bổ sung mức độ cho comment của chúng ta. Do những chỉ số bổ sung này sẽ gây ra một bài toán multi-label,trong khi mục đích của chúng ta là một bài toán multi-class, mà dữ liệu của chúng ta đã đươc xây dựng sẵn theo hướng multi-class rồi (từ trường class), cho nên ta tạm thời bỏ qua các giá trị bổ sung này mà focus on hai trường class và tweet (mà sẽ được đổi tên thành label và text, đồng thời đổi vị trí cho nhau cho dễ nhìn).

Sau khi xử lý xong ta sẽ bao gồm một file csv đầu ra.