PHÁT HIỆN BÌNH LUẬN THÔ TỤC

Trương Đức Vũ

18520194

Trường Đại Học Công Nghệ Thông Tin

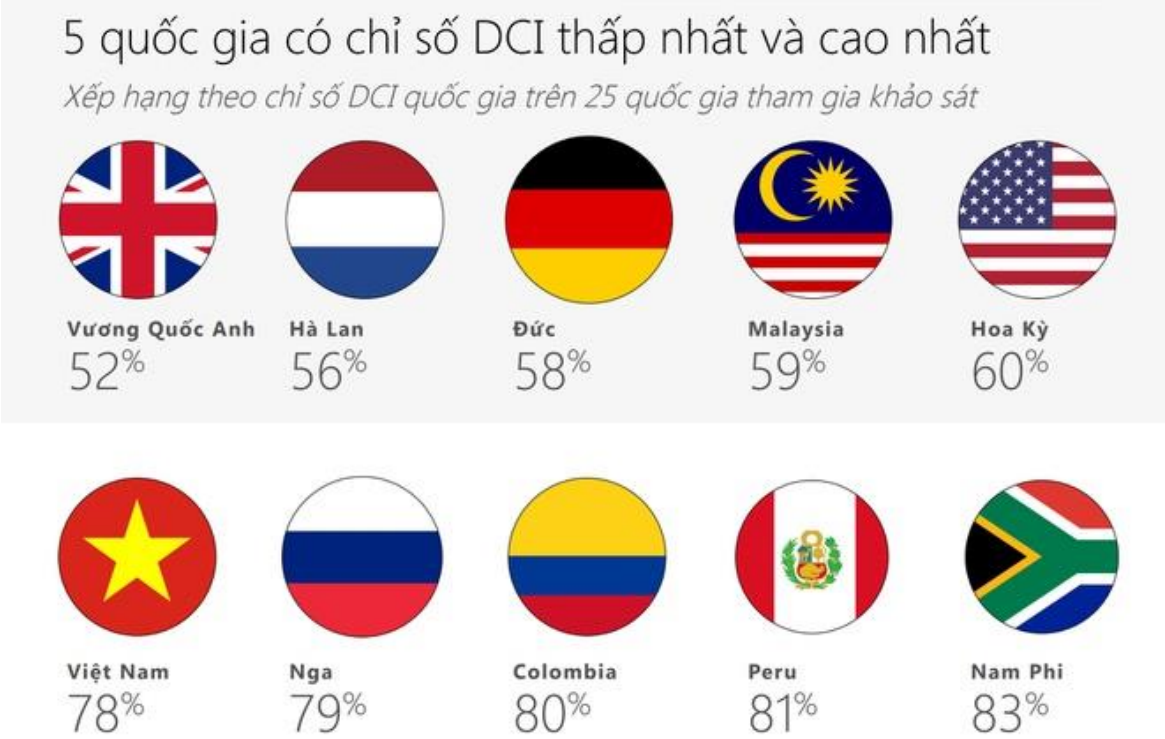
12/08/2020

-----o0o-----

1. Mô tả bài toán

Theo khảo sát mới được công bố của Microsoft, Việt Nam nằm trong top 5 quốc gia có chỉ số mức độ văn minh thấp nhất trên không gian mạng (DCI).

Mặc dù kết quả khảo sát này bất lợi cho hình ảnh cộng đồng mạng Việt Nam nhưng nó không tạo nên làn sóng phản đối từ người dùng. Theo khảo sát trên Zing.vn, 87% bạn đọc đồng tình với việc Microsoft xếp Việt Nam vào top 5 những nước hành xử kém văn minh trên Internet.



ảnh Zingnews.vn

Tuy vậy, vẫn có nhiều người phản đối kết quả của Microsoft. Trong đó, bình luận chủ tài khoản Facebook có tên Koba Yashi đến từ Việt Nam được chia sẻ rộng rãi nhất như một minh chứng rõ cho kết quả khảo sát.

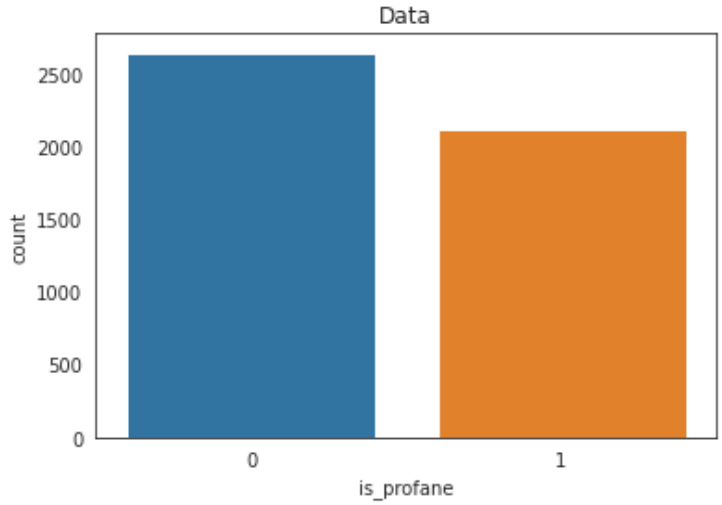
"Thấp cái *** ***, căn cứ vào đâu để đánh giá chứ", tài khoản Koba Yashi bình luận bên dưới liên kết của một đài truyền hình lớn ở Việt Nam. Tài khoản này sử dụng tên giả, từ ngữ bình luận thô tục và không có bất kỳ lập luận cụ thể nào cho ý kiến trái chiều của mình. (theo Zingnews.vn)

Từ đó em đã lựa chọn bài toàn phát hiện các bình luận thô tục, xúc phạm từ các bình luận. Bài toán phát hiện các bình luận có ý nghĩa thô tục, xúc phạm trên hầu hết các nền tảng mạng xã hội phổ biến hiện nay như Facebook, Youtube,...

Input là một bình luận bất kì (có thể chứa các icon và kí tự đặc biệt). Output là xuất ra bình luận đó có mang tính xúc phạm hay thô tục hay không.

2. Mô tả dữ liệu

- Cách thu thập bộ dữ liệu
  - Lấy tất cả các bình luận trong các bài viết gây xôn xao cộng đồng mạng
  - Lấy dữ liệu là comment trên facebook bằng cách crawl dữ liệu (cách crawl data và source code em để trên github)
- Số lượng (4764 bình luận – Từ đầu gần 10000 bình luận nhưng em đã lọc bớt các bình luận lặp, bình luận chỉ chứa icon, bình luận không có ý nghĩa gì,…)
  - 2121 bình luận mang tính thô tục, xúc phạm
  - Còn lại là bình luận bình thường.



- Phân chia
  - 75% dùng để làm training data, 25% còn lại dùng để làm test data, dev data được thu thập thêm ở ngoài sau khi training xong

- **Các thao tác tiền xử lý dữ liệu:**

- Cleaning data: là bước làm sạch dữ liệu trước khi bắt đầu bất kì xử lý nào trên tập dữ liệu của chúng ta, nó bao gồm một số khái niệm của xử lý ngôn ngữ tự nhiên như loại bỏ Stop Words (Năng cao: kiểm tra chính tả chẳng hạn)
- Words segmentation hay còn gọi là tách từ, là bước cực kỳ quan trọng và phức tạp đặc biệt là đối với Tiếng Việt.
- Chi tiết tiền xử lý dữ liệu:
    - Các comment thường viết hoa viết thường rất lung tung, nên em sẽ chuyển tất các các bình luận về dạng lower case.
    - Chuyển các chữ cái các bạn trẻ hiện nay thường kéo dài về dạng nguyên mẫu (vd: nguuuuuuuuuuuuuu --> ngu)
    - Loại bỏ dấu câu và các ký tự đặc biệt
    - Loại bỏ các icon không có ý nghĩa
    - Xử lý mất cân bằng dữ liệu:

      Under sampling Under sampling là việc ta giảm số lượng các quan sát của nhóm đa số để nó trở nên cân bằng với số quan sát của nhóm thiểu số. Ưu điểm của under sampling là làm cân bằng mẫu một cách nhanh chóng, dễ dàng tiến hành thực hiện mà không cần đến thuật toán giả lập mẫu.

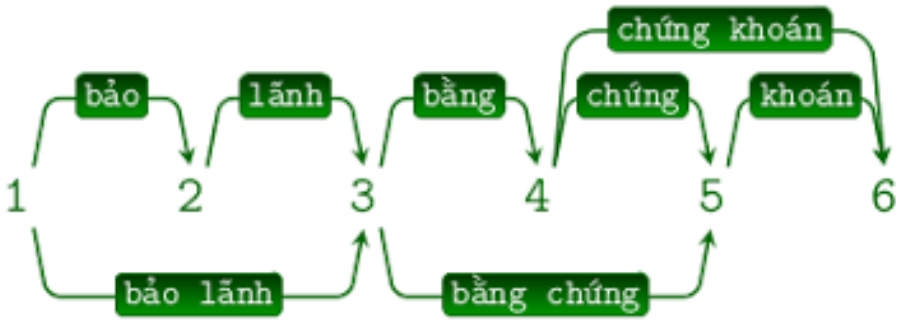
      Tuy nhiên nhược điểm của nó là kích thước mẫu sẽ bị giảm đáng kể. Gỉa sử nhóm thiểu số có kích thước là 500, như vậy để tạo ra sự cân bằng mẫu giữa nhóm đa số và thiểu số sẽ cần giảm kích thước mẫu của nhóm đa số từ 10000 về 500. Tổng kích thước tập huấn luyện sau under sampling là 1000 và chiếm gần 1/10 kích thước tập huấn luyện ban đầu. Tập huấn luyện mới khá nhỏ, không đại diện cho phân phối của toàn bộ tập dữ liệu và thường dễ dẫn tới hiện tượng overfitting. --> không sử dụng cách này

Over sampling Là các phương pháp giúp giải quyết hiện tượng mất cân bằng mẫu bằng cách gia tăng kích thước mẫu thuộc nhóm thiểu số bằng các kĩ thuật khác nhau.

TF-IDF (Term Frequency – Inverse Document Frequency)

Là 1 kĩ thuật sử dụng trong khai phá dữ liệu văn bản. Trọng số này được sử dụng để đánh giá tầm quan trọng của một từ trong một văn bản. Giá trị cao thể hiện độ quan trọng cao và nó phụ thuộc vào số lần từ xuất hiện trong văn bản nhưng bù lại bởi tần suất của từ đó trong tập dữ liệu. Một vài biến thể của tf-idf thường được sử dụng trong các hệ thống tìm kiếm như một công cụ chính để đánh giá và sắp xếp văn bản dựa vào truy vấn của người dùng. TF-IDF cũng được sử dụng để lọc những từ stopwords trong các bài toán như tóm tắt văn bản và phân loại văn bản.

Tách từ (Words segmentation)



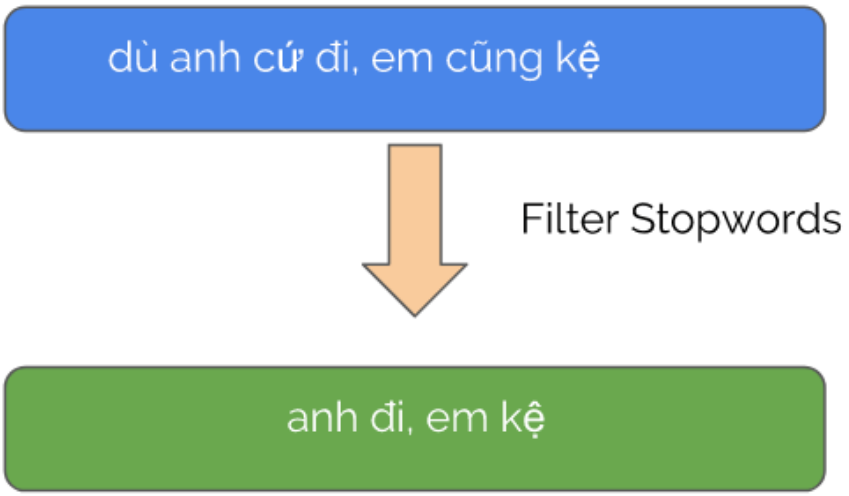
Là một bước quan trọng bậc nhất trong xử lý ngôn ngữ tự nhiên. Tiếng Việt không đơn giản như tiếng anh vì nó có thêm các từ ghép. Trong tiếng anh, thường thi mỗi từ sẽ được ngăn cách bởi dấu cách với các từ khác, nhưng trong tiếng Việt thì lại khác. Ví dụ: từ "Đất nước" nếu không sử dụng tách từ hợp lý thì nó sẽ biết thành 2 từ, nhưng dùng tách từ thì nó sẽ là "Đất_nươc".

Ví dụ: temp = "Chào các bạn tôi là một sinh viên công nghệ thông tin"

print NLP(text=temp).segmentation()

->>>Chào các bạn tô là một sinh_viên công_nghệ_thông_tin

Xóa bỏ stopword



Hãy tưởng tượng rằng ngôn ngữ của chúng ta giống như một đống gạo bị lẫn với thóc. Việc cần làm của trích chọn đặc trưng đó chính là chọn ra các hạt gạo chất lượng tốt nhất từ đống thóc đó. Những hạt thóc đó được gọi là stop words tức là những từ không có ý nghĩa lắm đối với việc phân loại của chúng ta. và có bổ sung thêm nhiều từ nữa để tối ưu bài toán em đang giải quyết.

3. Kết quả đạt được

Hình phía dưới là thống kê thành quả khi áp dụng 4 model: SVC (SVM), Naive
Bayes, Logistic Regression, Decision Tree Classifier. Với Accuracy lớn nhất thuộc về thuật toán Logistic Regression với 85%.

```
[INFO] Used TfidfVectorizer ...
   [INFO] evaluating SVM...
              precision    recall  f1-score   support

           0       0.83      0.91      0.87      1293
           1       0.88      0.78      0.82      1089

    accuracy                           0.85      2382
   macro avg       0.85      0.84      0.85      2382
weighted avg       0.85      0.85      0.85      2382

   [INFO] evaluating Naive Bayes...
              precision    recall  f1-score   support

           0       0.76      0.93      0.84      1293
           1       0.89      0.66      0.76      1089

    accuracy                           0.81      2382
   macro avg       0.83      0.79      0.80      2382
weighted avg       0.82      0.81      0.80      2382

   [INFO] evaluating Logistic Regression...
              precision    recall  f1-score   support

           0       0.78      0.95      0.85      1293
           1       0.91      0.67      0.78      1089

    accuracy                           0.82      2382
   macro avg       0.84      0.81      0.81      2382
weighted avg       0.84      0.82      0.82      2382

   [INFO] evaluating Decision Tree Classifier...
              precision    recall  f1-score   support

           0       0.84      0.86      0.85      1293
           1       0.82      0.80      0.81      1089

    accuracy                           0.83      2382
   macro avg       0.83      0.83      0.83      2382
weighted avg       0.83      0.83      0.83      2382
```

4. So sánh

Ở đây em so sánh với mô hình "Profanity Protection" với bộ dữ liệu toàn tiếng anh được thu thập trên mạng xã hội Twitter, Wiki(data set)

Dataset	Not Offensive	Offensive	Total
Tweets	4,163 (16.8%)	20,620 (83.2%)	24,783 (100%)
Wikipedia	143,346 (89.8%)	16,225 (10.2%)	159,571 (100%)
Combined	147,509 (80%)	36,845 (20%)	184,354 (100%)

Và kết quả đạt được

Package	Test Accuracy	Balanced Test Accuracy	Precision	Recall	F1 Score
profanity-check	95.0%	93.0%	86.1%	89.6%	0.88
profanity-filter	91.8%	83.6%	85.4%	70.2%	0.77
profanity	85.6%	65.1%	91.7%	30.8%	0.46

So với mô hình của em thì sẽ có kết quả cao hơn tại vì tiếng anh đơn giản hơn tiếng việt khá nhiều. Trong tiếng Việt, nhiều từ bị viết sai chính tả, viết tắt và viết sai phong cách chữ rất nhiều trong các câu bình luận. (four = phò, loz, đm, ...)

5. Khó khăn

Khó khăn lớn nhất chủ yếu tập trung vào việc thu thập dữ liệu. Dữ liệu phải thu thập bằng cách crawl rất nhiều lần và tốn rất nhiều thời gian, bên cạnh đó còn phải đánh label cho từng câu bình luận. Ngoài ra, còn thu thập dữ liệu từ nhiều chủ đề khác nhau để cho mô hình tránh được overfitting.
Khó khăn thứ hai đến từ việc xử lý dữ liệu, các câu bình luận khi crawl về rất tạp nham, phải chuẩn hóa, xóa bớt,... mới được thêm vào dataset

6. Kết luận

Tuy accuracy khá cao nhưng model đoán sai trong một số trường hợp ngoại lệ.

Ví dụ ngoại lệ:

- Mẹ: Yes

  Nhưng trong các bình luận, nếu ghi tự mẹ một mình như thế này nhiều người có thể hiểu đó là một câu chửi tục.

Ví dụ:

- Mấy cái loại súc vật này chỉ đáng bỏ đi :): Yes
- Vkl làm rk ai chơi, đi chết đi: Yes
- Mày chỉ đáng làm con four: Yes
- Nhà nó ở đâu vậy anh em: No
- Như một con đĩ: Yes
- Chị mua điểm đại học à?: No

Hướng phát triển

- Training thêm một model tự sửa lỗi chính tả tiếng Việt với dữ liệu được lấy từ các bài báo, bài văn tiếng việt.
- Khắc phục các trường hợp ngoại lệ.
- Thu thập thêm dữ liệu từ các thể loại khác để làm phong phú hơn bộ dữ liệu.

7. Tham khảo

https://towardsdatascience.com/building-a-better-profanity-detection-library-with-scikit-learn-3638b2f2c4c2

https://kipalog.com/posts/Machine-Learning---NLP--Text-Classification-su-dung-scikit-learn----python

https://viblo.asia/p/phan-loai-van-ban-tieng-viet-tu-dong-phan-1-yMnKM3bal7P

https://github.com/langmaninternet/VietnameseTextNormalizer?fbclid=IwAR3EN_3JNG16ZhBRYw2x4HHUqNTybyFBZ9xpkm4ABVCDUBzRj0elLm5YYqo

https://github.com/stopwords/vietnamese-stopwords