NLP-23-Assignment-2

Shea Durgin https://github.com/sheaDurgin/NLP23/tree/main/Assignment_2

February 2023

1 Introduction

A Naive Bayes classifier is a probabilistic algorithm based on Bayes' theorem, which uses a "naive" approach to predict a class given a document. It can be used for tasks such as, text and image classification, which includes sentiment analysis, fraud detection, and topic classification. However, the focus for this report will be on using it for topic classification.

2 Data and process

Taking the 20 most popular tags from the law forum on stack exchange that have posts before and after 2021, we can predict what tags certain forum post titles and questions come from. The text can be used in three ways, predicting with just the titles, just the question body's, or combining the title with its corresponding question. A training and test set can be created by training on all posts before 2021, and testing on all posts after 2021. After preprocessing the text to just get the important tokens, the Naive Bayes classifier can be trained using the train set. After it is trained, the classifier can attempt to predict what tag corresponds to each text it is given. This is where we get our final results from.

class	#training	#test
criminal-law	948	78
copyright	2016	181
united-states	5668	863
united-kingdom	1195	271
employment	238	36
international	316	43
canada	382	35
intellectual-property	301	29
england-and-wales	165	138
european-union	219	30
licensing	241	29
california	391	41
internet	416	39
business	171	7
rental-property	158	20
software	292	33
contract-law	1065	111
privacy	351	23
constitutional-law	177	21
gdpr	435	63

Figure 1: Data-set for Naive Bayes classifier

3 Results and analysis

The question the results are meant to help answer, are whether including both the title and body when training and testing together, leads to better results. As opposed to just training and testing on them separately. To discuss whether combining the title and body leads to more accurate results from a Naive Bayes classifier, let's reference the metrics gathered like accuracy, precision, recall, and the micro and macro f1 scores.

```
title
correct predictions 1052
wrong predictions 1039
total predictions 2091
united states predictions 1505
accuracy 0.5031085604973696 precision 0.31583656274229893 recall 0.15145062446851007
f1_micro 0.5031085604973696 f1_macro 0.14926116449813084
body
correct predictions 1054
wrong predictions 1037
total predictions 2091
united states predictions 1342
accuracy 0.5040650406504065 precision 0.21828506714459972 recall 0.16495297192092373
f1_micro 0.5040650406504065 f1_macro 0.1508106803257781
title+body
correct predictions 1076
wrong predictions 1015
total predictions 2091
united states predictions 1268
accuracy 0.5145863223338115 precision 0.27745889755923175 recall 0.18120504021974343
f1_micro 0.5145863223338115 f1_macro 0.16668924695895376
```

Figure 2: Results/Metrics for Naive Bayes classifier

What we see here is the results for the Naive Bayes classifier for just the title, just the body, and the title and body's bag of words combined. The first important metric is that they all have pretty similar correct predictions, all in the 10xx range. At first look, most would assume that they are all more or less the same in their classifying abilities. However, there is a statistic for the amount of united states predictions for a reason. As stated earlier, the data-set is made up of 20 classes that the classifier is made to predict based on their text. The numbers represent the amount of separate text input for that class. The united-states class represents over a third of the training data. This poses an issue when it comes to Naive Bayes classifiers, as it can cause them to over-predict a very common tag. This is why the amount of united states guesses is being tracked in the metrics. What does this metric tell us about the performance between the 3 tests? The title + body test set guesses a little more accurately than the others, all while guessing the most common tag less than the other 2. You can conclude from this, that it is getting more correct predictions from other tags that aren't the united states class.

4 OpenAI answer compared to accepted answer

To see how well OpenAI's GPT-3 language model compares to an accepted answer on the stack exchange forums, we can use one of the questions used in the data-set. The question is "can you film the police while they search your premises?". The response OpenAI produced was "It depends on the situation. If you are being detained or arrested, the police may confiscate your phone. If you are not being detained or arrested, you can film the police, but you may want to keep your distance to avoid being accused of interfering.". While the accepted answer was "Yes. The right to film police carrying out their duties is constitutionally protected. See, e.g, ACLU v. Alvarez, 679 F.3d 583 (7th Cir. 2012) and Glik v. Cunniffe, 655 F.3d 78 (1st Cir. 2011). Of course, one could always imagine some fact pattern in which some other consideration overrode this constitutional right (e.g. filming would have had the effect of destroying evidence that was light sensitive, so the police were using night vision equipment). Related: Boulder, Colorado pays $95,000 to settle the claim of a man arrested for filming the police." This displays one of the common tendencies of OpenAI responses, a lot of times it will start of its response with a wishy-washy statement like it does here, "It depends on the situation". I find this happens when it is giving an answer surrounding serious topics like, interacting with law enforcement. It doesn't want to give an answer like in an absolute sense so you do more research before doing something potentially harmful to yourself. Another difference is that the accepted answer uses sources since humans have access to searching the web for such things. While GPT-3 is just a language model, it is simply predicting the most likely words to come next given the question, it can't browse google and find a research paper to reference. I think both responses have merit, the human answer gives sources to back up their absolute claim, while OpenAI gives difference perspectives on the scenario. If you are being detained the police can confiscate your belongings for evidence and such.

5 Conclusion

In the end, Naive Bayes has its limitations. It is a probabilistic algorithm that considers each feature independently of others, whereas sometimes, this is not the case. The blame isn't all on the classifier though, as the distribution of classes may be the true culprit. A lot of the classes have a low amount of text to train on whereas the "united-states" tag gets a third of the entire training set. With more data, classifiers tend to perform better, but the data should be more equal between classes for a more nuanced Naive Bayes classifier.