Malicious language detection

Meeting notes

Meeting notes

Meeting with Sandra 1-28-2020

Sandra wants to work on getting baselines running for all the languages we're examining

- Still running into difficulty getting Petya access to the system since Brandi left

She thinks that using YASS for Arabic may be useful

- It may not be the most linguistically sound way to do things but it will be consistent and not be too agressive as a root identification system would (if we reduce the words to only their roots, maybe we loose too much information)
- Maybe for arabic and other languages it makes sense to use YASS to do splitting rather than stemming (e.g. keep the suffix that gets stripped)
-

Meeting notes

March 22-2019

absolute 2 way ig is not working well. The negative class always has higher absolute values.

Should probably stop running experiments soon, the deadline is in a month or so.

Is the difference between German and English due to differences in the dataset size? What if we tried to see if the classifiers are cluing into topics and not malicious language.

Look at the English data and get a sense of what we think people are bitching about.

Look at the ig features and see what winds up in there. to see if particular features are showing up.

Send danny an email with the number of features used.

Yue has high 70% for accuracy. She will update us once it it converges.

Meeting notes

April 20-2019

LDA in sklearn running over words only (unigrams). Cutoff of 3 2 topics pick number of top words it spits out.

Compare to XGBoost and SVM baselines. E.g. How often does the output from the SVM correspond to a particular topic from LDA? How often does the output from the XGBoost corresond to a particular topic from LDA?

Check out cases where topic modeler is not confident e.g. probabilities are close to 50. Maybe ones that are missed are on the borderline in the lDA model.

Could also run svm and get the coefficients for different word features.

Ken is running sampling experiments again. We're both running with 70,000 ig features 30,000 features and then 10,000 features

Ken is having some issues with cluster centroids. It has been running for 2 days and still hasn't finished downsampling

It appears to be using a sparse to dense conversion because it's using up to 50 gb of ram

Results dump

Results dump

Results BERT base allennlp

```
2019-10-22 00:15:51,049 - INFO - allennlp.models.archival - archiving weights and vocabulary
to /tmp/bert-mal-detect/model.tar.gz2019-10-22 00:16:20,067 - INFO - allennlp.common.util -
Metrics: {
  "best_epoch": 28,
  "peak_cpu_memory_MB": 3421.912,
  "peak_gpu_0_memory_MB": 12,
  "peak_gpu_1_memory_MB": 2528,
  "training_duration": "0:19:41.596926",
  "training_start_epoch": 0,
  "training_epochs": 29,
  "epoch": 29,
  "training_accuracy": 0.7378208300926253,
  "training_loss": 0.5308393001960953,
  "training_cpu_memory_MB": 3421.912,
  "training_gpu_0_memory_MB": 12,
  "training_gpu_1_memory_MB": 2528,
  "validation_accuracy": 0.7391857506361323,
  "validation_loss": 0.5321029019355774,
  "best_validation_accuracy": 0.7512722646310432,
  "best_validation_loss": 0.530970630645752
}
```

```
              precision    recall  f1-score   support

           1  0.68041237 0.39919355 0.50317662       496           2  0.76736924 0.91356877
0.83411116      1076

    accuracy                       0.75127226      1572   macro avg  0.72389081 0.65638116
0.66864389      1572
weighted avg  0.73993247 0.75127226 0.72969415      1572
```

Information gain

This houses investigation into 2 way information gain

Information gain

Two way IG feature selection

We're doing feature selection with two way IG like I did with mutual information for constructing my distantly supervised twitter sentiment lexicon in the sarcasm detection project (Comparison of two-side MI). However, in this case, we will just be trying the absolute version since that seemed to work best for sentiment.

One of the possibilities that we wanted to explore was whether you could bias against the majority class by doing feature selection that includes more information from the minority class.

Right now (03-17-2018) I am running balanced two way IG. After that, I will get results for 1.5 times more features from the minority class versus the majority class.

Revisiting 2-way IG

Sandra wants us to try out some different ways of calculating IG and then look at the top 100 results for each method to see what the quality of the results we're getting are.

I proposed the following methods to Danny

- IG(pos) - Abs(IG(neg)) grabbing from lowest and highest
- Norm(IG(pos)) - Norm(Abs(IG(neg))) grabbing from lowest and highest
- Abs(IG(pos) - Abs(IG(neg))) grabbing from highest
- Abs(Norm(IG(pos)) - Norm(Abs(IG(neg)))) grabbing from highest