Detecting Troll Tweets

Kemal Fidan, Nicholas Corbin, Shreyank Patel, Racheal Dylewski
December 4, 2020

Problem Setting

- Troll text typically concerns flaming or intentionally upsetting
- Pew Research Center found that 70% of 18 to 24 year olds who use the internet had experienced harassment
- No way of preventing someone on the internet to behave like a troll
- Identifying it can be helpful in hiding what people do not want to see

Dataset

- Borrowed from Kaggle
- Initially just contained Tweet and the label
- We defined troll text the way the dataset did: use of **aggression**

Tweet: “Get fucking real dude”

num_prof	prof	punc_cnt	compound	pos	neu	neg	len	annotation
1	1	0	0	0	1	0	22	1

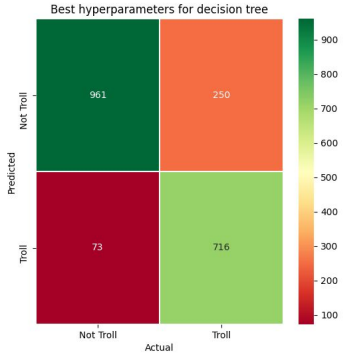
Approaches

- NLP with Doc2Vec:
  - Get paragraph embeddings from Doc2Vec
  - Try different models to learn vector values:
    - KNeighborsClassifier, DecisionTreeClassifier, GaussianNB, LogisticRegression
- Decision Trees:
  - Learn numeric values harvested from each Tweet
  - num_prof, prof, punc_cnt, compound, pos, etc

Evaluations/Observations

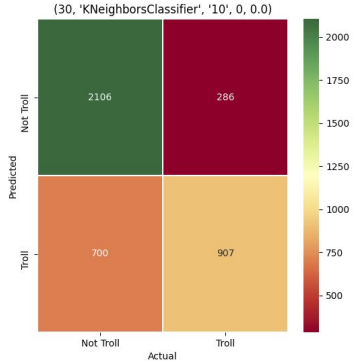
Best Performing Hyperparameters

Hyperparameters	Accuracy	Precision	Recall	F1 Score
Max_depth: 40, Min_impurity_decrease: 1e-6	0.85	0.75	0.91	0.82
No Threshold	0.85	0.75	0.94	0.85



Best Performing Hyperparameters

Hyperparameters	Accuracy	Precision	Recall	F1 Score
KNN K=5	0.75	0.76	0.72	0.73
SVC, poly, C=20, D=2	0.70	0.78	0.63	0.62
GaussianNB, s=1e-8	0.73	0.75	0.69	0.70



Human's agreement on sentiment is about 82%!

Future Work

- Improve Doc2Vec model:
  - Remove punctuation from Tweets
  - Replacing slang words or acronyms
- Neural nets for prediction:
  - Find complex relationships between features and label
  - Powerful tuning for over/underfitting