Exploration into uses of large unlabeled datasets

Sage Hahn (UG)

Binary text classification

Used the ‘Twenty Newsgroup Data Set’ of 20,000 labeled new articles to test different approaches.

Created an unlabeled ‘Corpus’ of documents from varied sources, Reddit posts, Tweets, Blog posts and a different set of news articles.

For preprocessing, I reduced each document to a lowercase sentence, and Stemmed each word using the “Snowball Stemmer”.

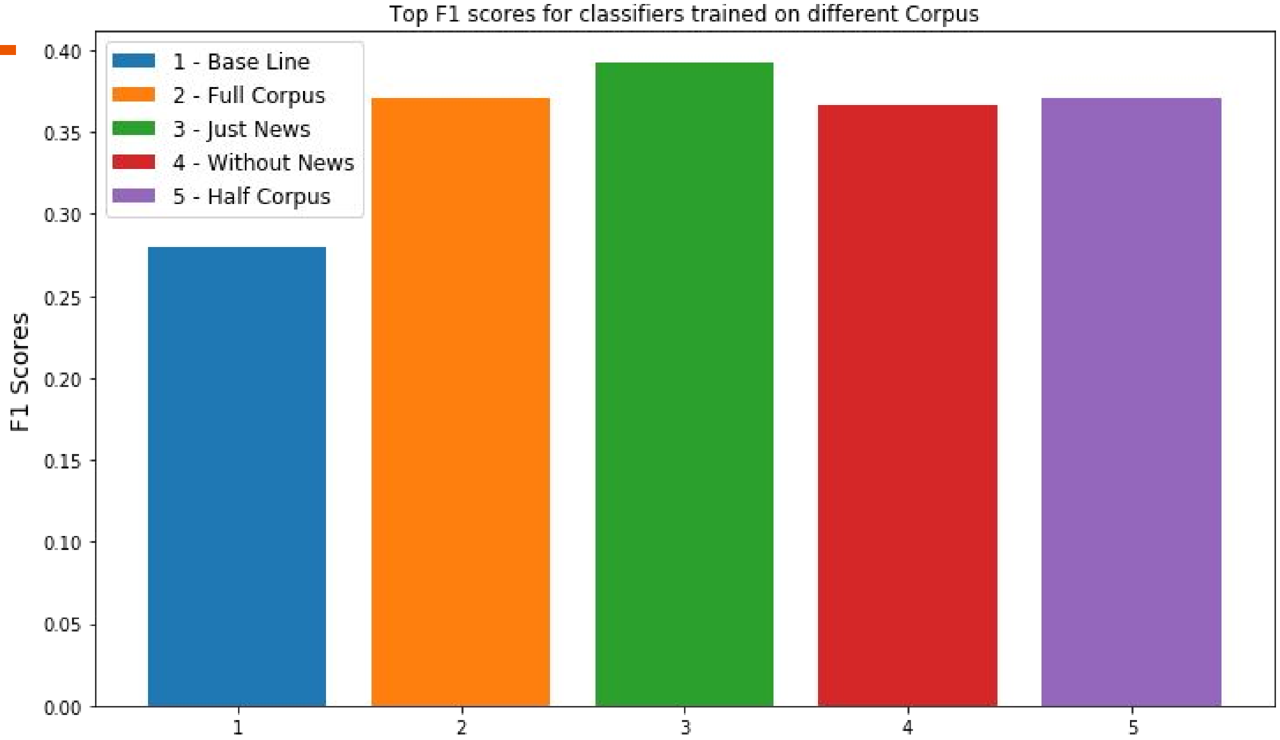
Classifier from scratch

Aimed to create an easy to use and implement binary text classifier from only user generated keywords.

Idea is to search instead through the unlabeled Corpus, using the sentences found as training data, along with randomly chosen examples to act as not about the subject.

Then train a SGD classifier from sklearn, and finally test accuracy on the Twenty Newsgroup data.

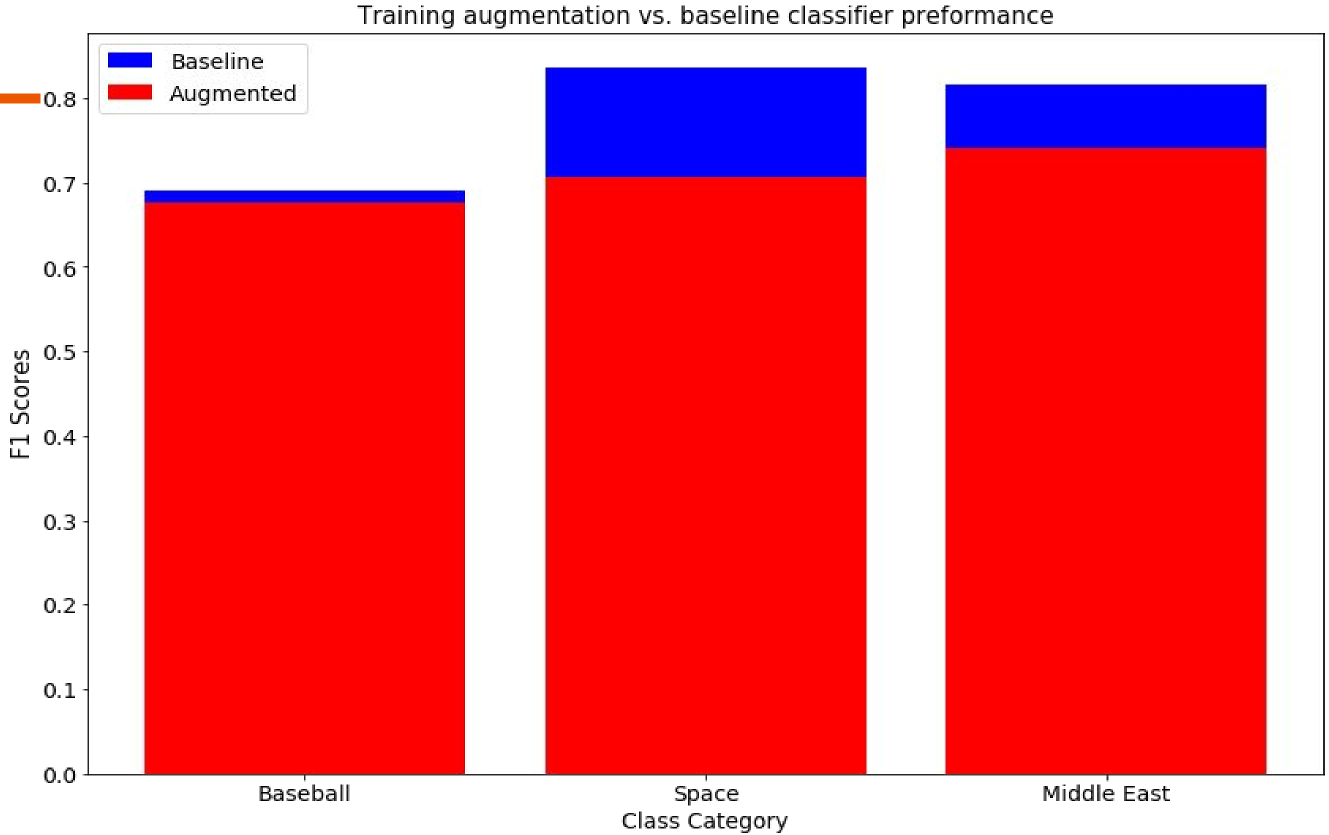
Tested the approach with keywords about Computer Graphics, e.g. “graphics card”, “texture mapping”


Top F1 scores for classifiers trained on different Corpus
F1 Scores
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
1 - Base Line
2 - Full Corpus
3 - Just News
4 - Without News
5 - Half Corpus
1
2
3
4
5

Improving supervised training performance

-Modified the Twenty Newsgroup Dataset to be a binary text classification problem

-Attempted to augment performance using similar technique as before - Search the Corpus for keywords, and add sentences found to the training data, then retrain an SGD classifier.


Training augmentation vs. baseline classifier preformance
Baseline
Augmented
F1 Scores
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Baseball
Space
Middle East
Class Category

Future Work

-Explore use of pre-trained word embeddings.

-Using ‘gloVe’ pre-trained word embeddings and a convolutional neural network and was able to get, for the ‘Middle East News’ category, an f1 score of .94. Compared to best performing SGD classifier with f1 score of .82.