Accuracy Manual Log

# Purpose

The purpose of file is to log change made to classifiers and their effects. The features (vectorizers in case of scikit) are also described since they play an important role in accuracy and other metrics.

# Scikit Learn Naive Bayes Classifier

Trained from: “April 2014 Guardian News Articles” Set.

Tested on: “First 10 Days of July” Set.

Date: 15 Aug. 2014

## Original: (0.1)

The classifier starts with the following properties:

* On instantiation: **fit\_prior = False**
* Uses a **TfidfVectorizer** with (ngram\_range=(1,3), stop\_words=*'english'*)
* Accuracy: **86.3%**

### Changes attempted and reverted:

* On instantiation: **fit\_prior = True**,
  + Accuracy **reduced by 3%** to 83%.
  + Reverted.
* TfidfVectorizer option: **stop\_words=’english’ removed**.
  + Accuracy reduced by **9%** to **77%**.
  + Most important words per category were mostly stop words. Very bad.
  + Reverted.

## (0.2):

The classifier starts with the following properties:

* On instantiation: **fit\_prior = False**
* Uses a **TfidfVectorizer** with (ngram\_range=(1,1), stop\_words=*'english'*)
* Accuracy: **87.5% (+1.2%)**

### Changes from previous version:

* Uses unigrams only
* MUCH *faster* and *lighter* than using trigrams,