# Sequential Testing in Classifier Evaluation Yields Biased Estimates of Effectiveness

20

20

50 100 200 500 2000 5000

Training documents

STIVERSITL **Mossaab Bagdouri David D. Lewis Douglas W. Oard** William Webber University of Maryland University of Maryland David D. Lewis Consulting University of Maryland College Park, USA College Park, MD, USA Chicago, IL, USA College Park, MD, USA wew@umd.edu mossaab@umd.edu sigir2013pap@DavidDLewis.com oard@umd.edu Supported in part by NSF IIS-1065250 Introduction Experiments **20** Runs 20 Runs x 29 Topics 1 Run x 50 Confidence Levels x 29 Topics • Goal: Economical assured effectiveness Q Topic = M132, Frequency = 3.33% Build a good classifier 80 Certify that this classifier is good Stop Fail o Use nearly minimal total annotations Criterion 9.0 F₁≥τ 53.58% 2 • Common practice (sequential training): θ≥τ 8.13% 5.00% Select a fixed "certification" test set Sequential Trainin Desired 8 Obser ន Add some training instances Fixed (1067) Test Test whether effectiveness target reached 20 o Repeat add-and-test as needed 20 50 100 200 500 2000 5000 50 60 70 ลก 20 100 Training documents Intended confidence level (%) Sequential training introduces bias Topic = C31, Frequency = 5.04% Sequential testing introduces bias 80 Both together introduce bias Stop Fail 970 Criterion  $F_1 \ge \tau$ 100.0% Design 40 θ≥τ 31.55% Sequential Desired 5.00% Test 0.2 8 Fixed (2000) Training 80 Reuters newswire stories (RCV1-v2)  $\circ$  29 topics with  $\geq$  25,000 positive examples 20 200 500 2000 70 90 100 50 100 5000 50 60 80 **Training documents** Intended confidence level (%) • Random sampling for training and test 2 580 randomized runs (20 per topic) Topic = GPOL, Frequency = 7.07% 8 Stop Fail • F<sub>1</sub>: "True" effectiveness (on 700K documents) Criterion 0.6  $\circ$   $\mathbf{F}_1$ : Point estimate F₁≥τ 68.38% 5 θ≥τ  $\circ$   $\theta$  : Lower limit of one-sided 95% conf. int. for F<sub>1</sub> 9.40% S Desired 5.00%  $\circ$   $\tau$ : Target for  $F_1$ Sequential Training 8 8 (50%/50%) and Test

#### Confidence Level:

- Intended: Desired % of time  $\theta \ge \tau$  when we stop
- $\circ~$  Observed: Fraction of 580 runs that exceed  $\tau$



#### • Key results:

#### Test Collection

## • Passive Learning

### • Notation:

