Comparing Intrinsic and Extrinsic Evaluation of Sensitivity Classification

Mahmoud F. Sayed, Nishanth Mallekav, and Douglas W. Oard
University of Maryland, College Park

Introduction

- Goal: Make more content available for search
- Some is intermixed with sensitive content
  - Personal information, private conversations, etc.

- Manual segregation of sensitive content from that which can be shared in impractical
- Objective is to study the relation of sensitivity classification effectiveness on a search engine that seeks to protect sensitive content

Test Collections

- Avocado email research collection.
  - ~800,000 email messages
  - 65 topics judged for relevance & sensitivity
  - Sensitivity based on one of two personas:
    - John Snibert: Corporate engineer
    - Holly Palmer: University professor
  - Each topic has ~100 judged docs
- OHSUMED test collection
  - ~250,000 MEDLINE abstracts
  - 106 topics judged for relevance & sensitivty
  - Two simulated “sensitive” categories:
    - C12, C13 (Urogenital Diseases)
  - Topics have ~152 judged docs on average

Classification Effectiveness

- We build three sensitivity classifiers based on document text

a. Logistic Regression (LR)
b. DistilBERT
c. OR combination of LR and DistilBERT

	OHSUMED				
Classifier	Precision↑	Recall↑	F_1↑	F_2↑	Accuracy↑
(a) LR	76.72	73.29	74.96	73.95	94.01
(b) DistilBERT	**82.75**	80.08	**81.39**	80.60	**95.52**^{a,c}
(c) Combined	74.61	**83.81**	78.94	**81.8**	94.53^{a}

	Avocado: Holly Palmer				
Classifier	Precision↑	Recall↑	F_1↑	F_2↑	Accuracy↑
(a) LR	**72.29**	69.98	71.12	70.43	**90.34**^{b,c}
(b) DistilBERT	66.20	67.85	67.02	67.52	88.65
(c) Combined	64.15	**80.11**	**71.25**	**76.31**	89.02

	Avocado: John Snibert				
Classifier	Precision↑	Recall↑	F_1↑	F_2↑	Accuracy↑
(a) LR	**80.53**	84.85	**82.63**	83.95	**83.06**^{b,c}
(b) DistilBERT	72.87	87.00	79.31	83.75	78.44
(c) Combined	70.86	**93.73**	80.71	**88.05**	78.72

Search Among Sensitive Content

- We used normalized Cost Sensitive Discounted Cumulative Gain (nCS-DCG), which rewards finding relevant documents but penalizes revealing sensitive documents.

$$CS\text{-}DCG_k = \sum_1^k \left(\frac{g_i}{d_i} + c_i\right)$$

$$nCS\text{-}DCG = \frac{CS\text{-}DCG - CS\text{-}DCG_{worst}}{CS\text{-}DCG_{best} - CS\text{-}DCG_{worst}}$$

- We built our ranking models using the Coordinate Ascent ranking algorithm.
- We used two approaches for combining a ranking model and a sensitivity classifier.

a. A post-filter approach that uses the sensitivity classifier on the ranking model’s output to filter out any result that is predicted to be sensitive. The ranking model optimizes toward nDCG@10.
b. A joint approach which works by directly optimizing the ranking model toward nCS-DCG@10, which balances between relevance and sensitivity.

Sensitivity-Aware Ranking Effectiveness

- Jointly modeling relevance and sensitivity yields better results than post-filtering
- When training data is limited, F_2 might be a useful intrinsic measure with which to initially compare sensitivity classifiers when optimizing for measures such as nCS-DCG that penalize failures to detect sensitive content.

Collection: Topics	OHSUMED: 106		Holly Palmer: 35		John Snibert: 35	
Classifier	Post-filter	Joint	Post-filter	Joint	Post-filter	Joint
(a) LR	83.11	83.81	79.92	87.38	76.32	80.87
(b) DistilBERT	84.57^{a}	**85.95**^{a,c}	82.41	86.30	75.48	80.74
(c) Combined	**84.97**^{a}	84.44	**84.40**^{a}	**90.67**^{a}	**79.65**	**83.46**^{a}
Oracle	89.44	88.70	92.19	89.64	95.40	91.91

Acknowledgments

- This project is funded by NSF grant IIS1618695

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

COLLEGE OF INFORMATION STUDIES

UMIACS