ScaDS.AI
DRESDEN LEIPZIG
CENTER FOR SCALABLE DATA ANALYTICS AND ARTIFICIAL INTELLIGENCE

Investigating and Mitigating Topic Bias in Authorship Attribution

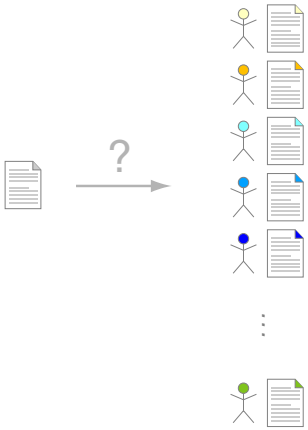
Niklas Deckers
Text Mining and Retrieval Group, Leipzig University

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

SACHSEN
Diese Maßnahme wird gefördert durch die Bundesregierung aufgrund eines Beschlusses des Deutschen Bundestages.
Diese Maßnahme wird mitfinanziert durch Steuermittel auf der Grundlage des von den Abgeordneten des Sächsischen Landtags beschlossenen Haushaltes.

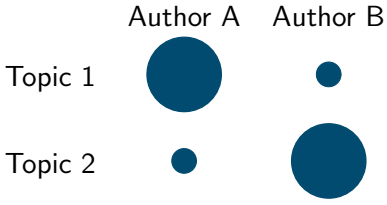
Authorship Attribution

- Classification task
- Based on the writing style of text



Topic Bias

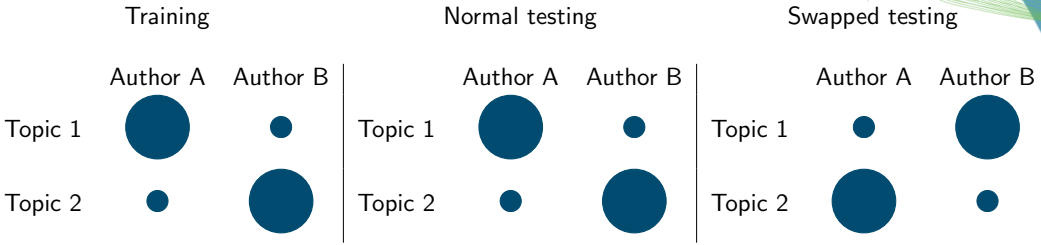
- Real-life scenarios usually have datasets with a skewed distribution of topics
- Traditional approaches fall for this
  had never been noticed because test data was biased the same way



Compiling Datasets to Control Topic Bias

- Thus require texts
  - with perfectly balanced topics (hard to find in natural data)
  - or at least topic labels
- Using Fanfiction data
  - Author labels are clear
  - Fandoms as topic labels
  - Many authors write in many fandoms

Detecting Bias Susceptibility in Existing Approaches: Swapping Experiment

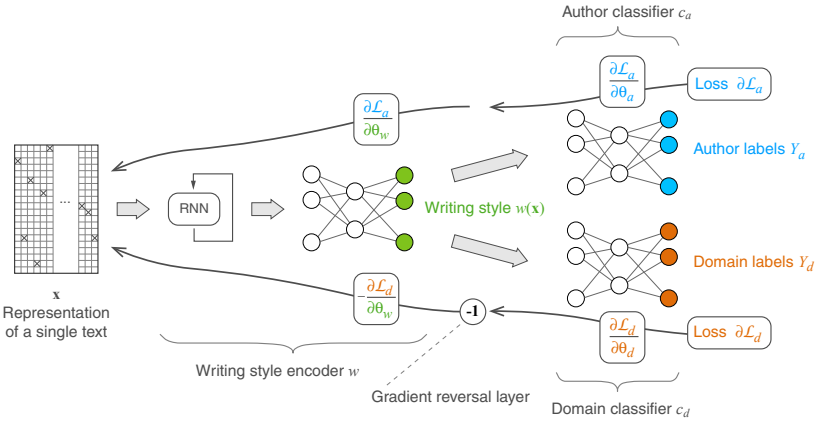


Mitigation Approaches

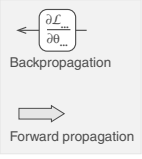
- Traditional approach: Masking to remove features

Original text	The cars, slightly smaller than the Ford Taurus and expected to be priced in the $15,000-$17,000 range, could help GM regain a sizeable piece of the mid size car market, a segment it once dominated.
DV-SA text distortion	The * , * * than the * * and expected to be * in the $# , # -$# , # range , could help * * a * * of the * size car market , a * it once * .

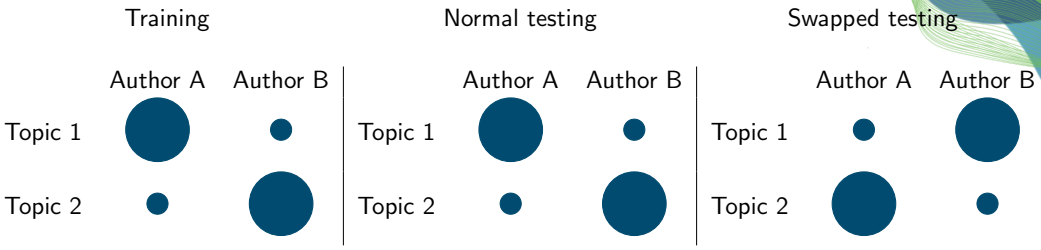
Bias Mitigation Using Gradient Reversal



Legend:



Results



- Traditional character trigram SVM approach: Drop of 37.7% accuracy
- Gradient reversal approach: Drop of 9.8% accuracy

Conclusion

- Using labeled datasets to detect bias susceptibility
- Training methods to mitigate topic bias

- Future work: Extracting universal writing style embeddings
- Mitigating bias in unlabeled scenarios