Janek Bevendorff, Matthias Hagen, Benno Stein, Martin Potthast

Bias Analysis and Mitigation in the Evaluation of Authorship Verification

Overview

The PAN series of shared tasks is well known for its continuous and high-quality research in digital text forensics.

We review the authorship verification task and conclude that the underlying experiment design cannot guarantee pushing forward the state of the art—in fact, it allows for top benchmarking with a surprisingly straightforward approach.

In this regard, we present a **"Basic and Fairly Flawed"** (BAFF) authorship verifier that is on a par with the best approaches submitted so far, and that **illustrates sources of bias that should be eliminated.**

BAFF, a "baffling" authorship verifier:

	Acc.	ROC
BAFF on the PAN15 corpus		
BAFF	**0.768**	0.746
Bagnall (best appr.)	0.757	**0.811**
BAFF on the PAN14 Novels corpus		
Modaresi, Gross (best appr.)	**0.715**	0.711
BAFF	0.651	0.715
Zamani et al. (runner-up)	0.650	**0.733**
BAFF on the PAN14 Essays corpus		
BAFF	**0.722**	**0.761**
Fréry et al. (best appr.)	0.710	0.723

Features:

- Cosine similarity (TF)
- Cosine similarity (TFIDF)
- Kullback-Leibler divergence (KLD)
- Skew divergence (balanced KLD)
- Jensen-Shannon divergence
- Hellinger distance
- Avg. logarithmic sentence length difference

I: Model Bias

B1: Corpus-relative features

Global features such as IDF derived from small samples of a population allow the classifier to reverse-engineer part of the ground truth.

B2: Feature scaling

Scaling features to an interval based on corpus statistics also exploits biases in the corpus when the size of the data is small (similar to Bias B1).

II: Data Bias

B3: Plain text heterogeneity

Non-homogenized texts preserve publication-specific formatting and provide inadvertent features, particularly in the presence of Bias B4.

B4: Population homogeneity

Splitting and reusing longer texts is a quick way to increase the number of author pairs, but it leads to over- and underrepresentation of authors and therefore poor generalization.

III: Evaluation Bias

B6: Test conflation

Providing verifiers with the full test corpus is a rather unrealistic scenario and allows exploitation of corpus information. Verifiers should only see one test case at a time and must not base decisions on previously seen data.

B5: Accidental text overlap

The strong performance of TFIDF hints at topic overlap between texts in same-author pairs. The classfier identifies authors by what they write about and not by their style. This is a result of Bias B4.

BAFF Dismantled: The Webis Authorship Verification Corpus

Our new Webis Authorship Verification Corpus avoids Biases B1–B5:

- 192 pairs training / 82 pairs test
- 3,900 words per text on average
- all texts from the same time period (19th / 20th century)
- texts have similar genre (adventure and sci-fi only)
- no author appears twice in any genre / period combination
- punctuation and special characters normalized
- all formatting and ASCII art removed

	Acc.	ROC
PAN15 corpus: IDF sources		
Corpus IDF (feat. scaled)	**0.768**	**0.746**
Corpus IDF (feat. unscaled)	0.622	0.639
Brown IDF (feat. scaled)	0.598	0.598
Brown IDF (feat. unscaled)	0.590	0.590
TFIDF single-feature performance		
PAN15 corpus (feat. scaled)	**0.742**	**0.769**
Webis corpus (feat. scaled)	0.570	0.599

Corpus and data:

github.com/webis-de/acl-19