Bauhaus-Universität Weimar

Fakultät Medien

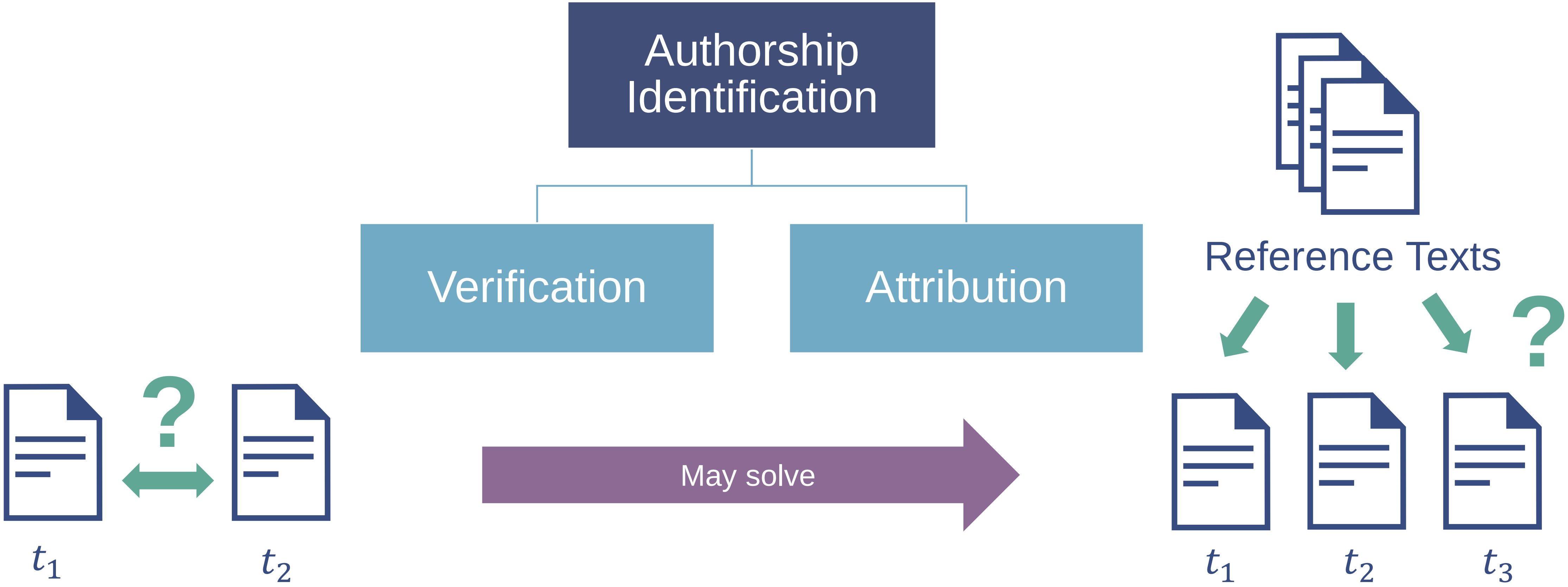
# Authorship Verification and Obfuscation Using Distributional Features

Bachelor’s Thesis Defense by
Janek Bevendorff

**Date:** 27. October 2016

**Referees:** Prof. Dr. Benno Stein
PD Dr. Andreas Jakoby

# What Is Authorship Verification?



Authorship Identification

Verification

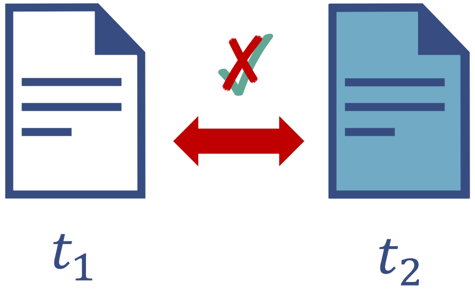
Attribution

Reference Texts

May solve

# What Is Authorship Obfuscation?

*“Given two documents by the same author, modify one of them so that forensic tools cannot classify it as being written by the same author anymore.”*



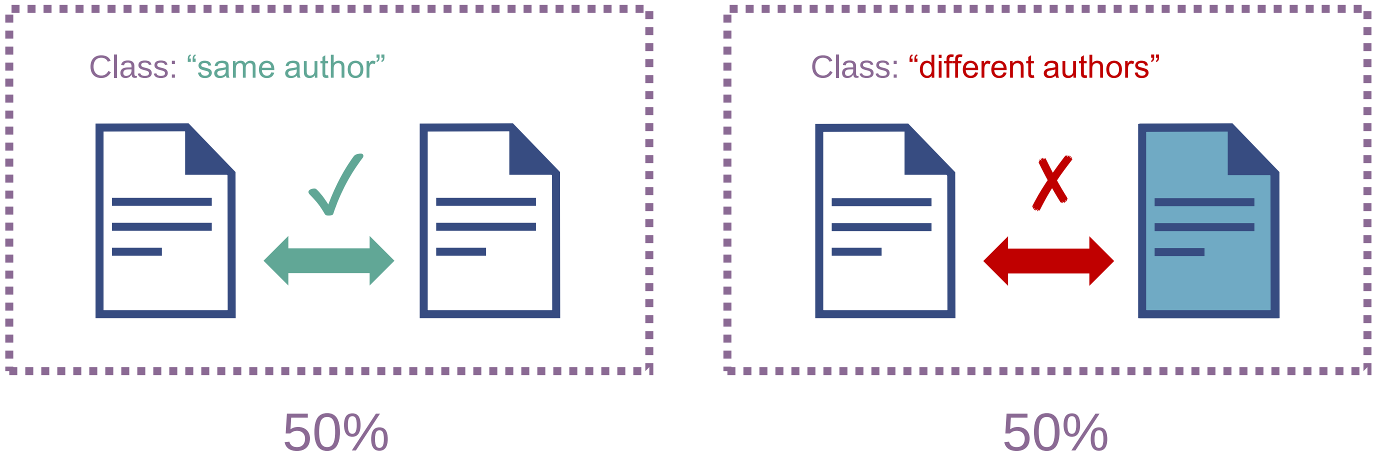
# Reasons for Obfuscating Authorship

- General privacy concerns
- Protection from prosecution
- Anonymity of single / double blind reviews
- Style imitation (writing contests)
- Impersonation (malicious intents)
- …

# Corpus Setup

Used corpus: PAN15 Corpus (English)

- Training / test: 100 / 500 cases
- Two classes with balanced number of cases
- Each case consists of two documents either by the same or different author(s)
- Test documents have 400-800 words on average



Class: “same author”

✓

50%

Class: “different authors”

✗

50%

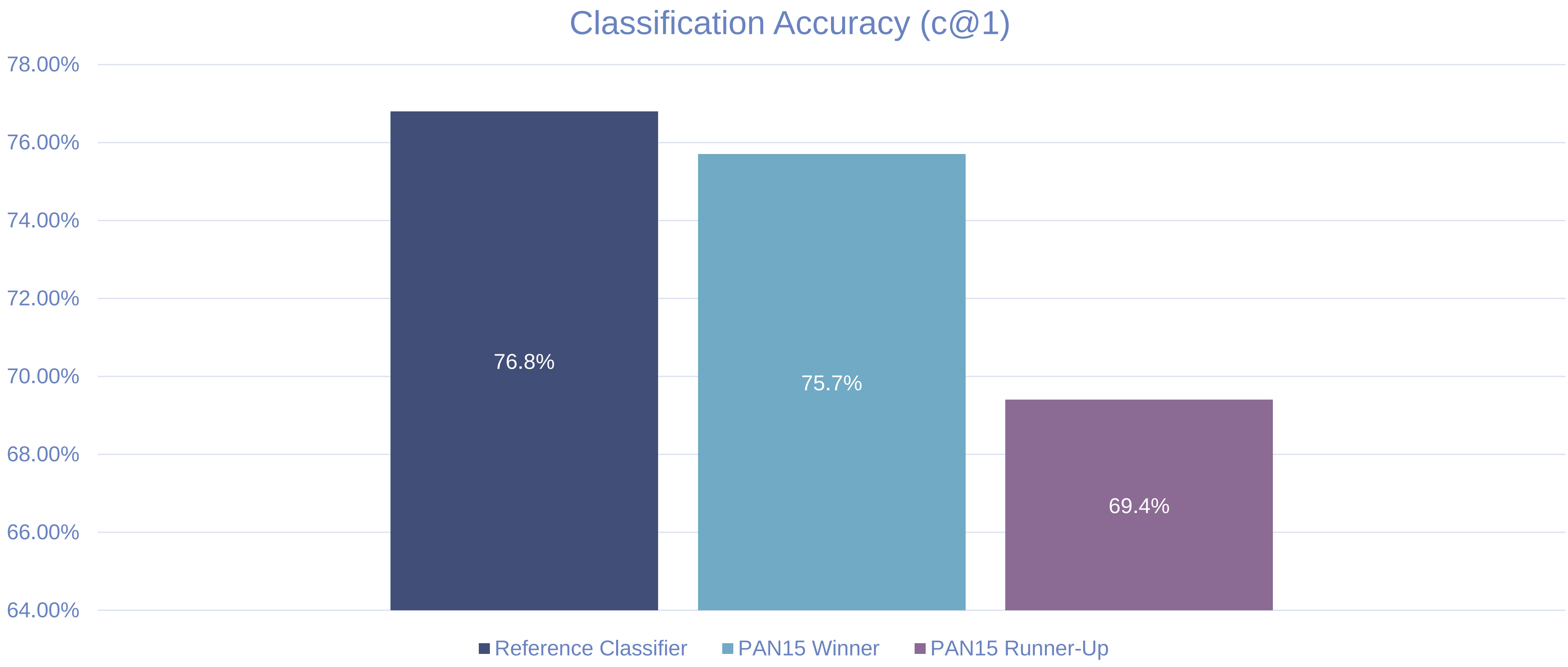
# Reference Classifier

*Decision tree* classifier with 8 features:

- **Kullback-Leibler divergence (KLD)**
- **Skew divergence (smoothed KLD)**
- **Jensen-Shannon divergence**
- **Hellinger distance**
- Cosine similarity with TF weights
- Cosine similarity with TF-IDF weights
- Ratio between shared n-gram set and total text mass
- Average sentence length difference in characters

The first 7 features use character 3-grams

# Classification Results



# Obfuscation Idea (1)

- Attack KLD as main feature
- Assumes other features not to be independent

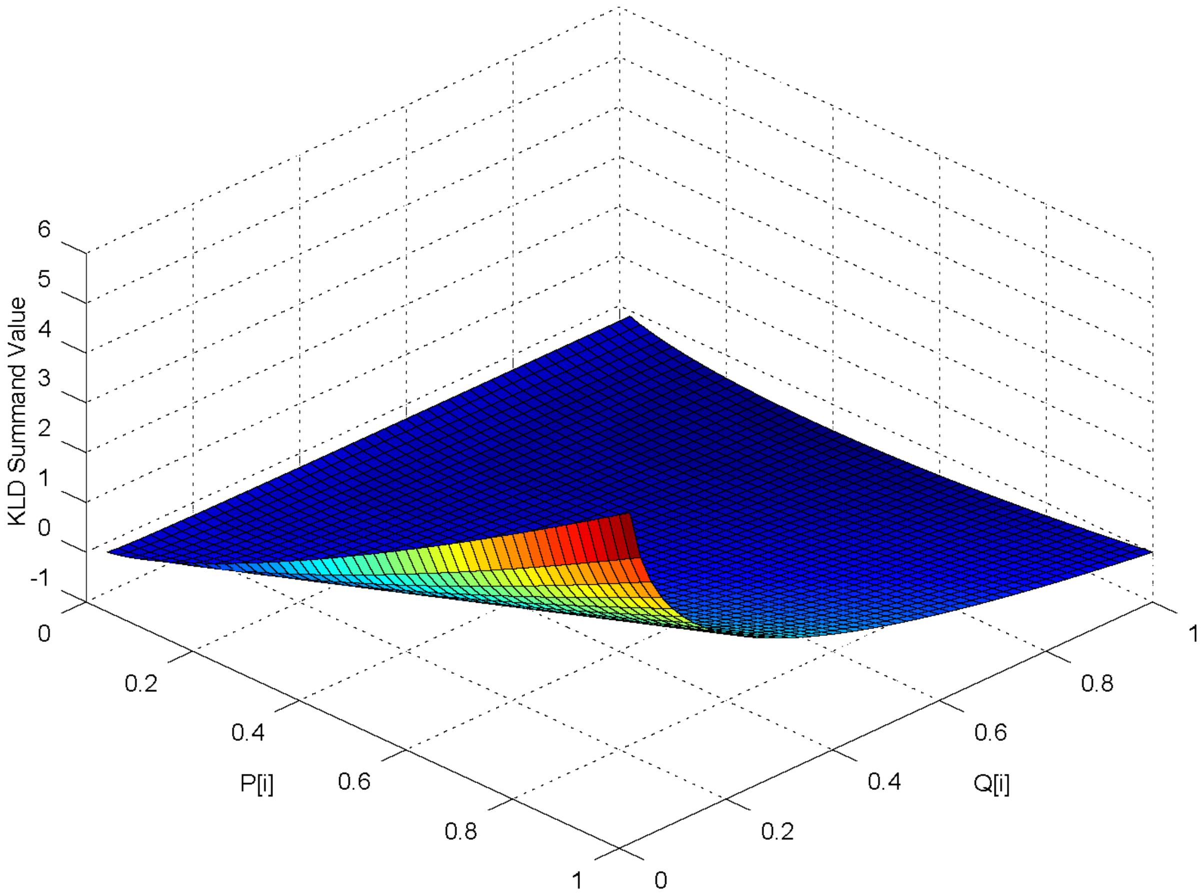
$$\mathrm{KLD}(P||Q) = \sum_{i} P[i] \log_2 \frac{P[i]}{Q[i]}$$

KLD Definition

Variables:

- $i$: n-gram appearing in both texts $t_1$ and $t_2$
- $P[i]$: relative frequency of n-gram $i$ in the portion of $t_1$ whose n-grams also appear in $t_2$
- $Q[i]$: analogous to $P[i]$

# KLD Properties



- ➢ KLD range: $[0, \infty)$
- ➢ KLD = 0 for identical texts
- ➢ **PAN15 corpus:** 0.27 < KLD < 0.91
- ➢ KLD only defined for n-grams where $Q[i] > 0$
- ➢ **PAN15 corpus:** at least 25% text coverage by only using n-grams that appear in both texts

# Obfuscation Idea (2)

**Idea:** obfuscate by increasing the KLD

- Assumption: not all n-grams are equally important for the KLD
- Only touch those with highest impact
- High-impact n-grams can be found by KLD summand derivative:

$$\frac{\partial}{\partial q}\left(p \log_2 \frac{p}{q}\right) = -\frac{p}{q \ln 2}$$

KLD Summand Derivative

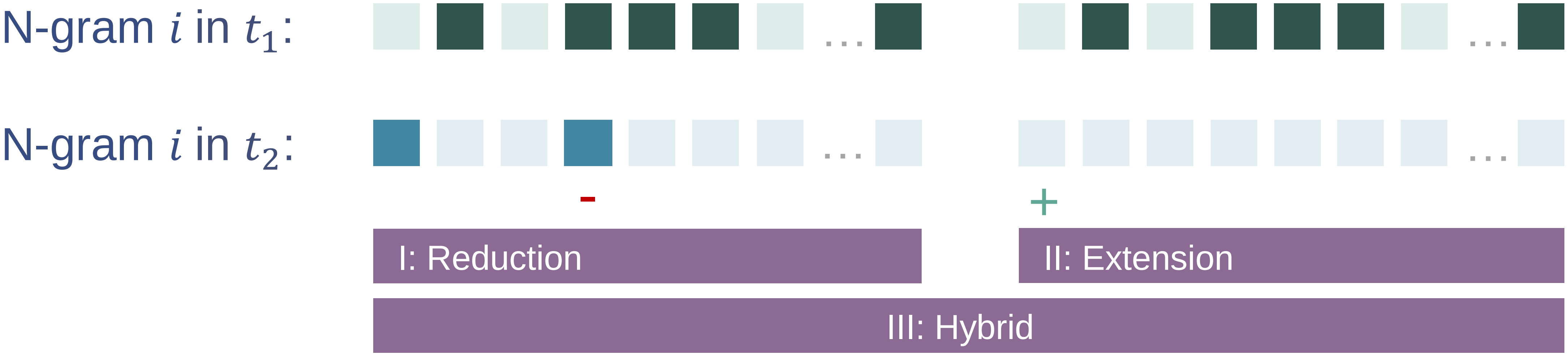
where $p$ and $q$ denote probabilities $P[i]$ and $Q[i]$ for any defined $i$

# Obfuscator Implementation

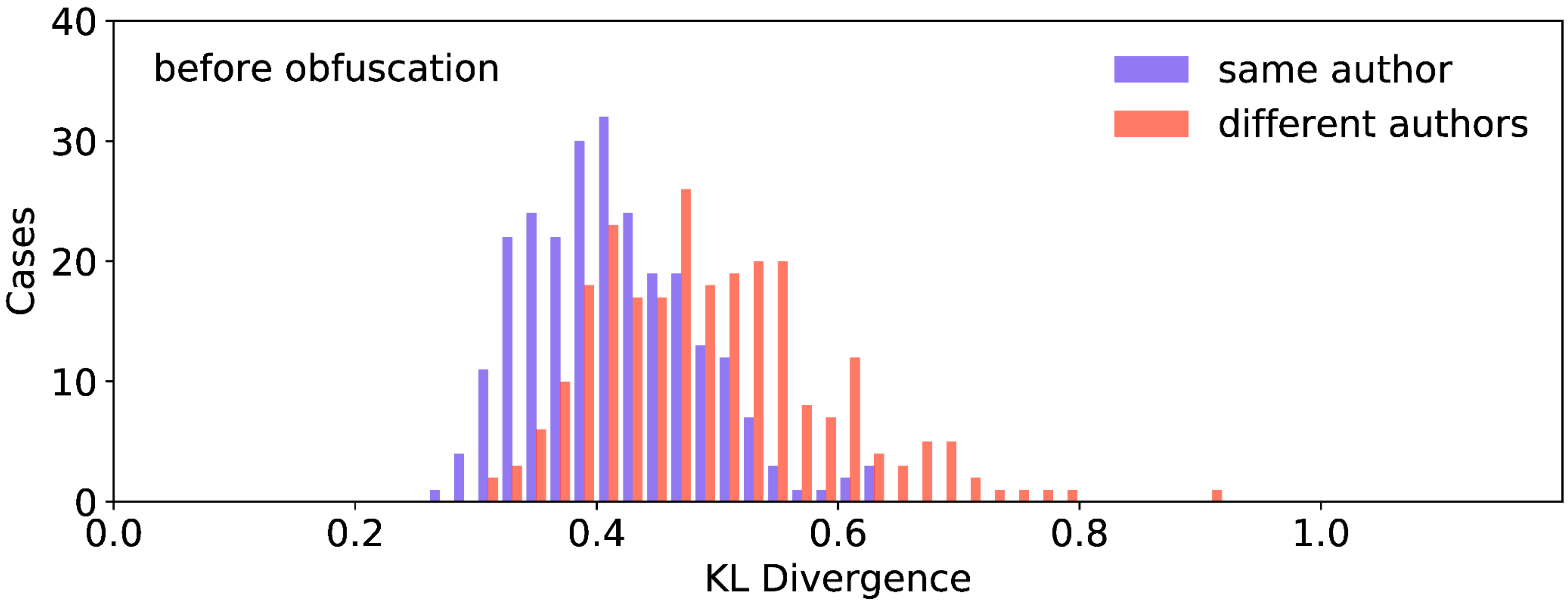
Only need to consider the (modifiable) n-gram $i$ that maximizes

$$\frac{P[i]}{Q[i]}$$

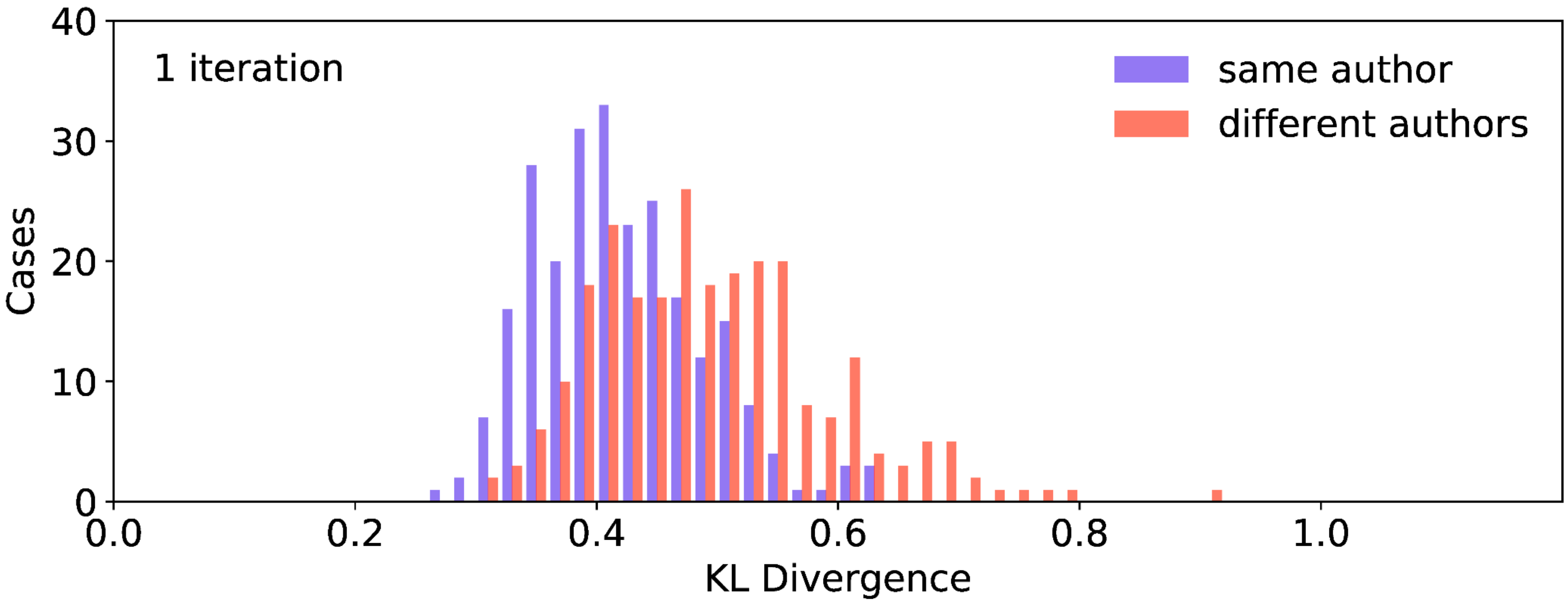
Three possible obfuscation strategies:



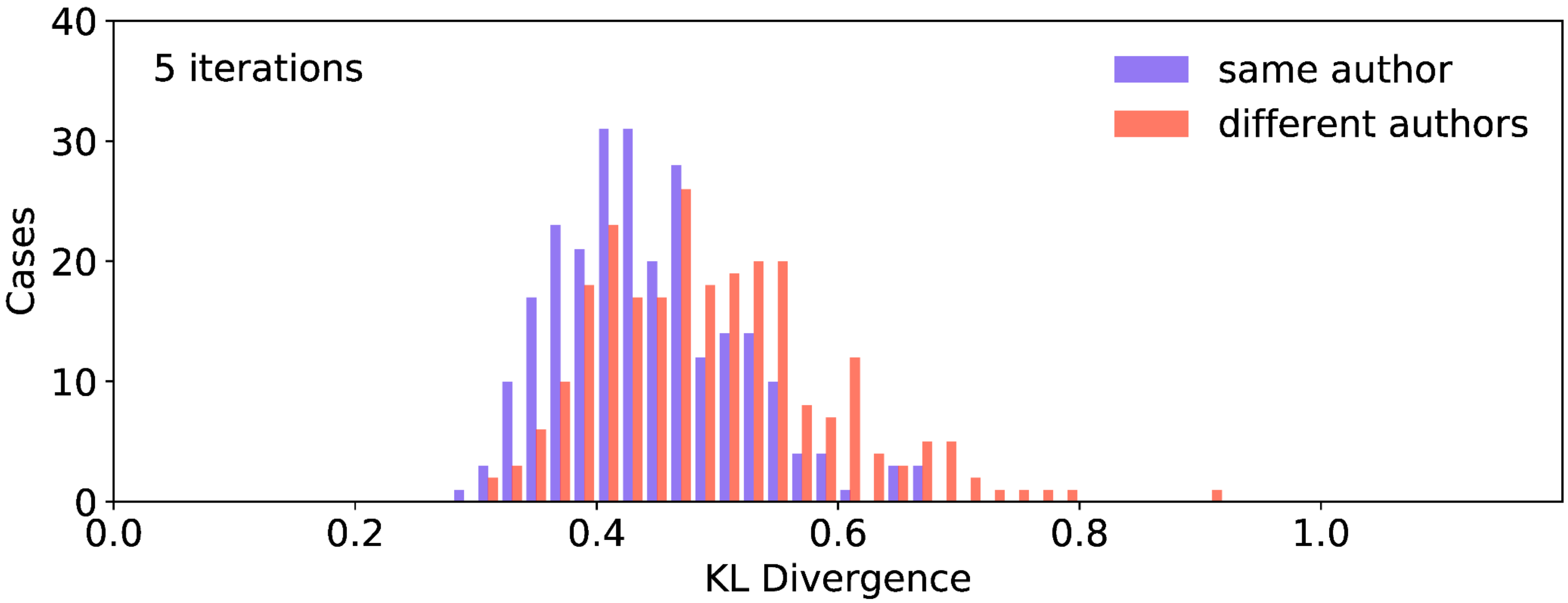
# Obfuscation Results



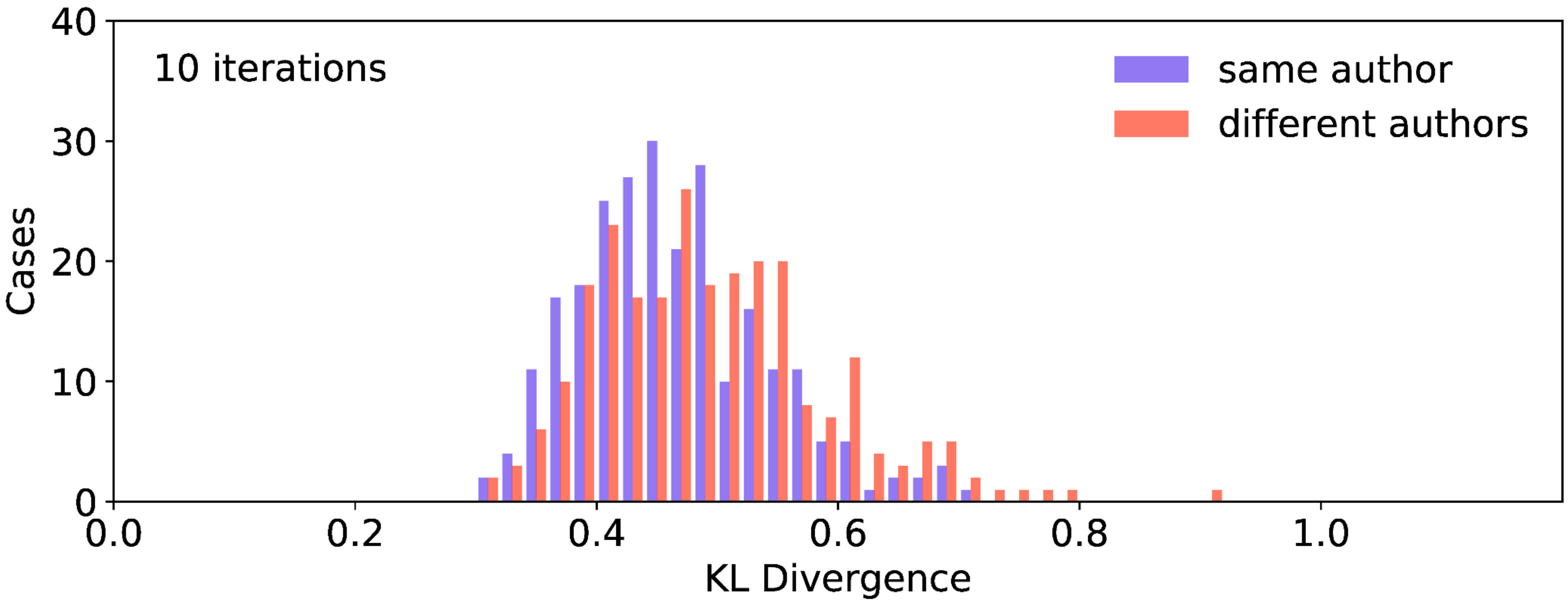
# Obfuscation Results



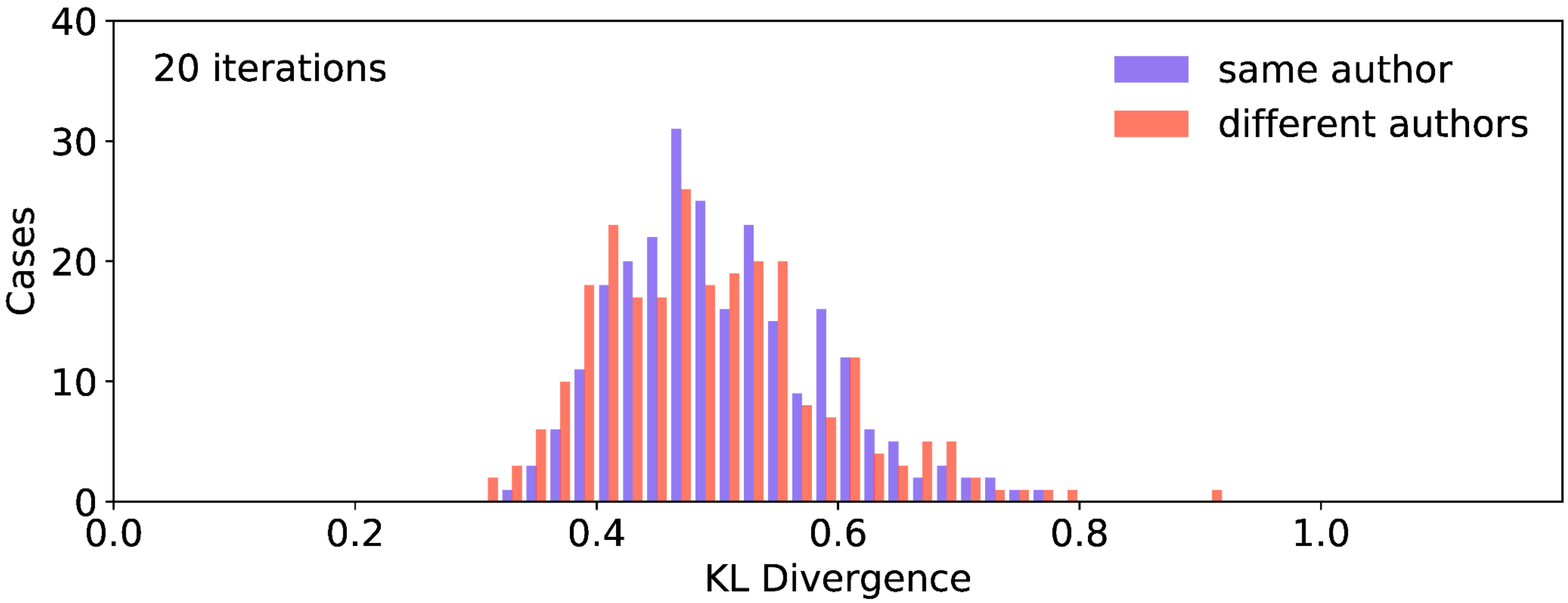
# Obfuscation Results



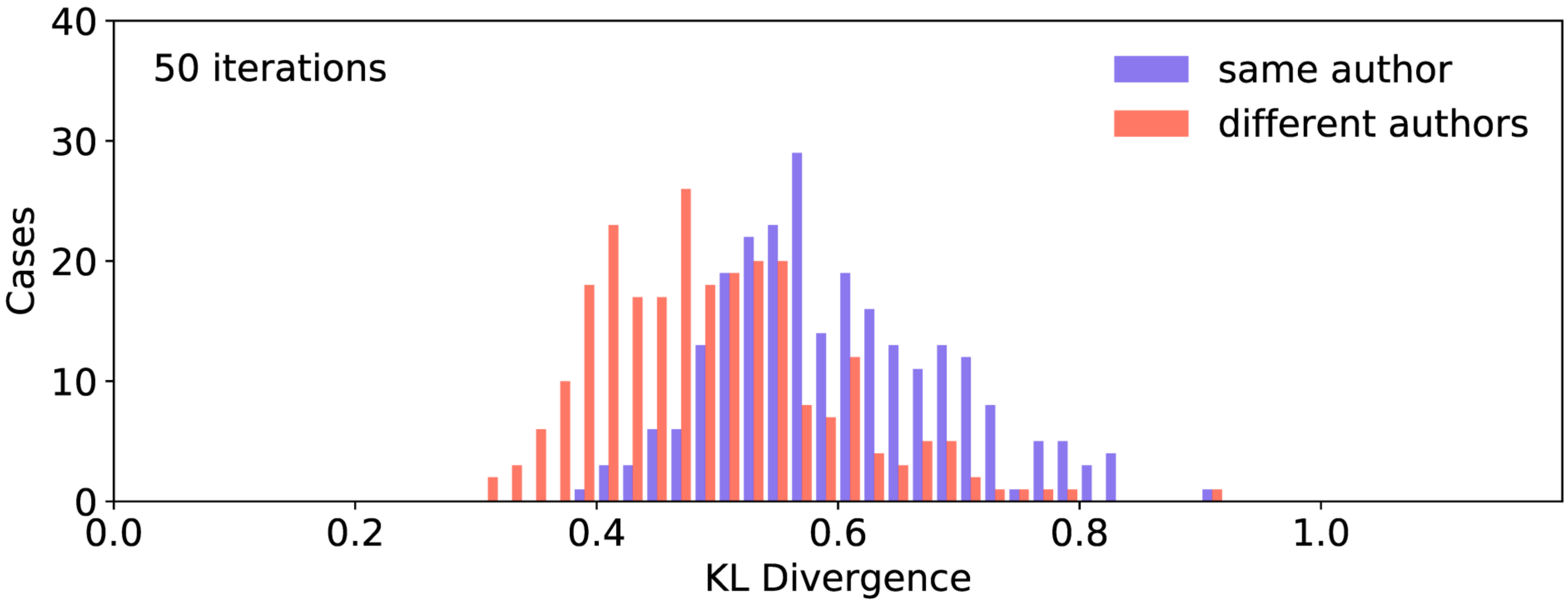
# Obfuscation Results



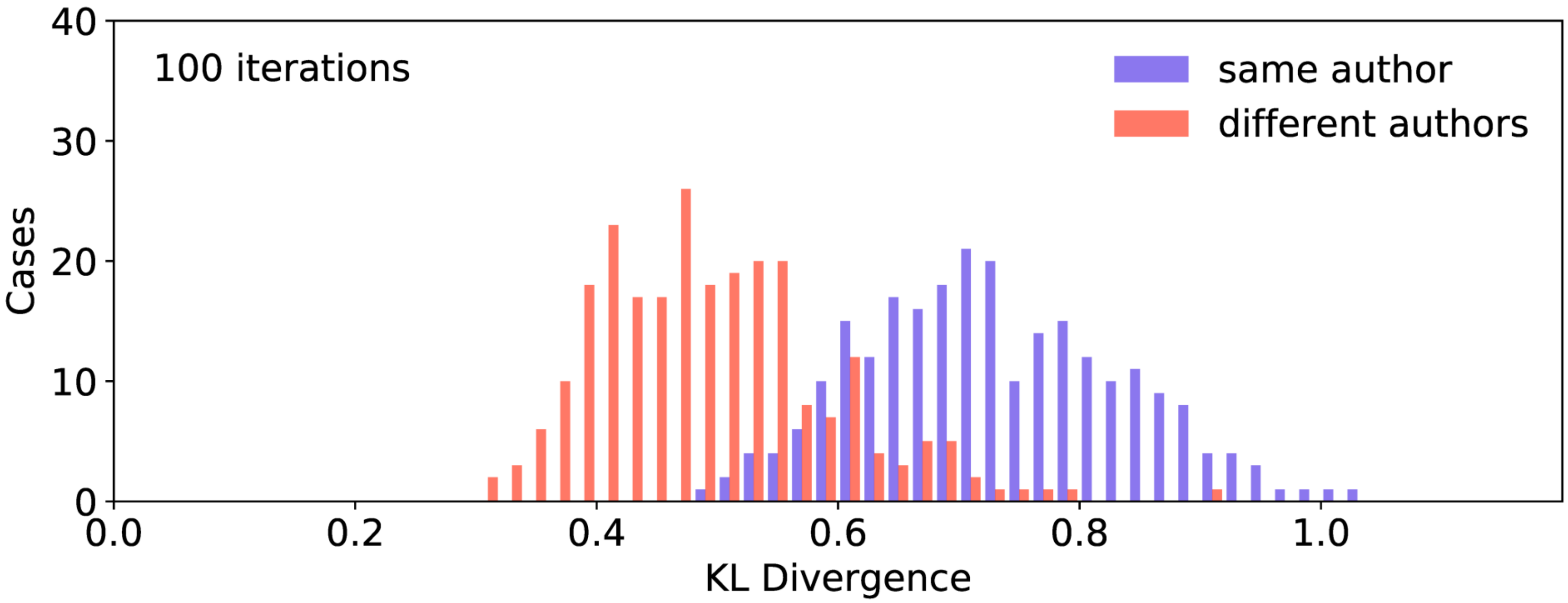
# Obfuscation Results



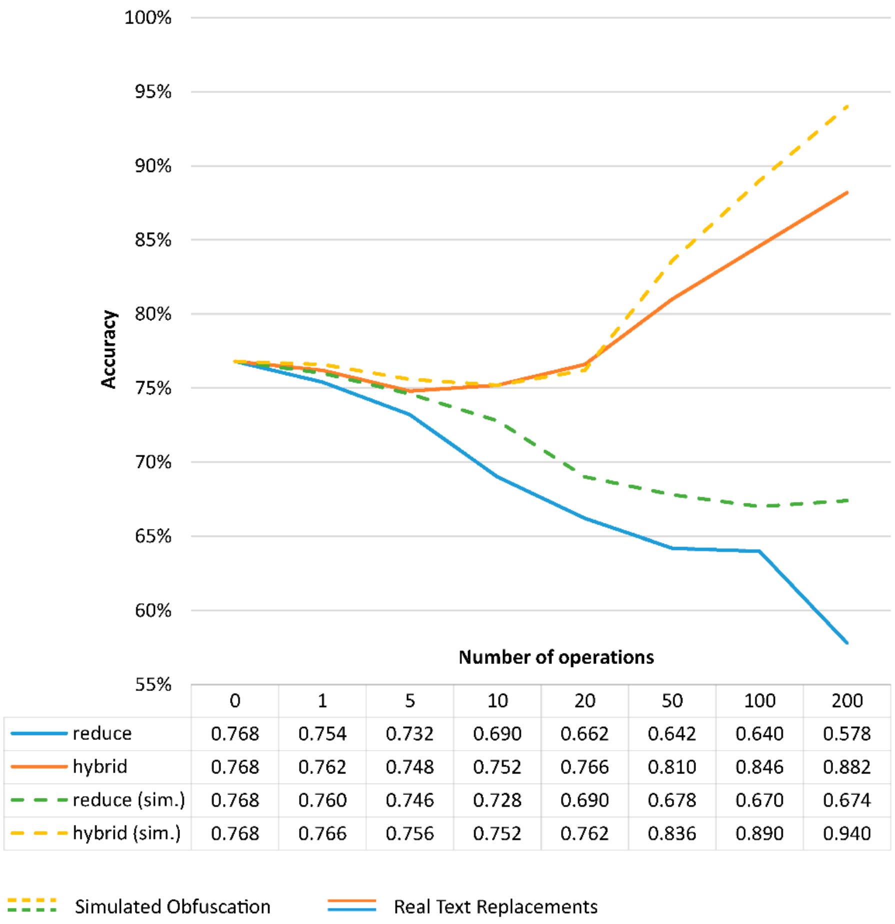
# Obfuscation Results



# Obfuscation Results



# Obfuscation Results



| | 0 | 1 | 5 | 10 | 20 | 50 | 100 | 200 |
|---|---|---|---|---|---|---|---|---|
| reduce | 0.768 | 0.754 | 0.732 | 0.690 | 0.662 | 0.642 | 0.640 | 0.578 |
| hybrid | 0.768 | 0.762 | 0.748 | 0.752 | 0.766 | 0.810 | 0.846 | 0.882 |
| reduce (sim.) | 0.768 | 0.760 | 0.746 | 0.728 | 0.690 | 0.678 | 0.670 | 0.674 |
| hybrid (sim.) | 0.768 | 0.766 | 0.756 | 0.752 | 0.762 | 0.836 | 0.890 | 0.940 |

**Observation Hybrid:** accuracy rises despite KLD increase

**Possible explanation:** adding n-grams improves other features.

Cross-validation with single features confirms explanation:

| | Baseline Accuracy | 20 Iterations |
|---|---|---|
| KLD | **67.2%** | 51.4% |
| TF-IDF | 74.4% | **82.2%** |

**Solution:** only use reductions

# Results Analysis

- Significant KLD increase possible with only few iterations
- KLD histograms fully overlap after 10-20 iterations (~2% of text modified)
- Overall classification accuracy down to ~66%
- Extensions are problematic for TF-IDF

# Corpus Flaws

Results promising, but corpus appears to be flawed

- Very short texts
- Test corpus much larger than training corpus
- Corpus-relative TF-IDF very strong feature (discrimination by topic)
- Only chunks of 15 different stage plays by 5 unique authors
- No proper text normalization

# Development of New Corpus

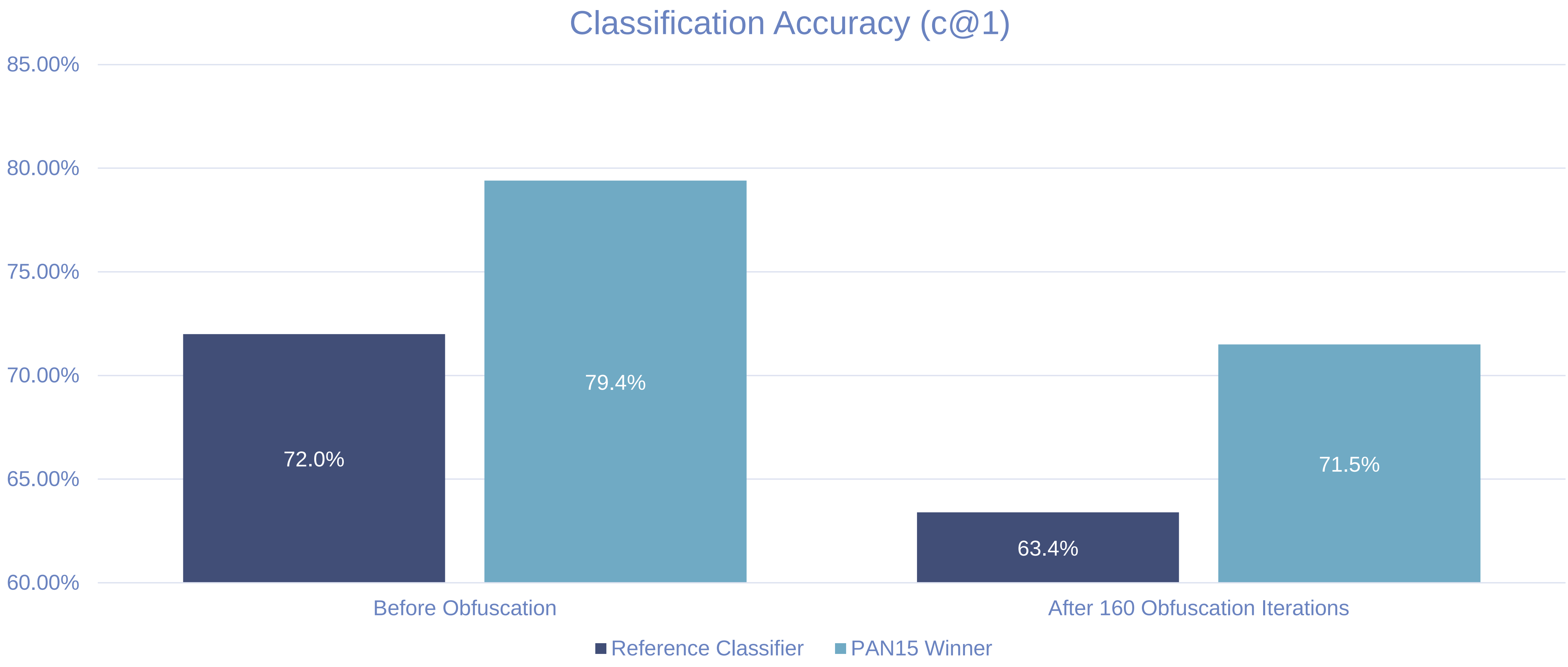
New corpus was developed with books from Project Gutenberg:

- 274 cases from three genres and two time periods
- Authors unique within genre / period
- Avg. text length of 4000 words (few exceptions)
- Proper text normalization
- 70 / 30 split into training / test (192 / 82 cases)

# Classifier Changes

Cosine similarity (TF and TF-IDF) features were removed to avoid accidental classification by topic

# Classification Results



# Summary

- Medium / high classification accuracy with only simple features
- Obfuscation possible by attacking main feature
- Results reproducible on more diverse corpus
- Obfuscation also works against other verification systems

# Future Work

- Improve classifier by
  - …adding more features
  - …integrating “Unmasking” by Koppel and Schler [2004]
- Attack more features
- Use paraphrasing
- Randomize obfuscation to harden against reversal

# Thank you

for your attention