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

$t_1$, $t_2$, $t_3$
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
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

Classification Accuracy (c@1)

- Reference Classifier: 76.8%
- PAN15 Winner: 75.7%
- PAN15 Runner-Up: 69.4%

27. October 2016
Obfuscation Idea (1)

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

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

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: $0.27 < \text{KLD} < 0.91$
- 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:

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

where $p$ and $q$ denote probabilities $P[i]$ and $Q[i]$ for any defined $i$
Only need to consider the (modifiable) n-gram $i$ that maximizes

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

Three possible obfuscation strategies:

I: Reduction

II: Extension

III: Hybrid
Obfuscation Results
Obfuscation Results

5 iterations

- same author
- different authors

Cases

KL Divergence

27. October 2016
Obfuscation Results

10 iterations

Cases

KL Divergence

same author
different authors
Obfuscation Results

![Diagram showing KL Divergence with 20 iterations]

- Cases
- 0.0 to 1.0
- KL Divergence
- Same author
- Different authors
Obfuscation Results

50 iterations

Cases

KL Divergence

same author
different authors

27. October 2016
Obfuscation Results
Obfuscation Results

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

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

Cross-validation with single features confirms explanation:

<table>
<thead>
<tr>
<th></th>
<th>Baseline Accuracy</th>
<th>20 Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>KLD</td>
<td>67.2%</td>
<td>51.4%</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>74.4%</td>
<td>82.2%</td>
</tr>
</tbody>
</table>

**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
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

Classification Accuracy (c@1)

Before Obfuscation
- Reference Classifier: 72.0%
- PAN15 Winner: 79.4%

After 160 Obfuscation Iterations
- Reference Classifier: 63.4%
- PAN15 Winner: 71.5%
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