On the Feasibility of Author Identification in the Era of Big Data

Neil Zhenqiang Gong
EECS, UC Berkeley

Arvind Narayanan, Hristo Paskov (CS, Stanford), Neil Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin and Dawn Song (EECS, UC Berkeley)
Motivation

• Anonymous/pseudonymous contents are everywhere!
Motivation

• Anonymous contents:
  – Sensitive political topics
  – Sensitive personal psychological/health issues.

• Identifying authors = huge privacy attack!

• Possible via writing style at Large scale?
Notable Coups for Stylometric Author Identification

Shakespeare-Bacon controversy in 19\textsuperscript{th} century

Disputed Federalist Papers ~50 years ago
Author identification behaves qualitatively different at large scale.
Threat Model

Attacker: oppressive government, etc.

Authors are not protecting themselves

Use author ID as first step
Follow up with other methods: topic, viewpoints, location...
Problem Definition

• Given:
  • N authors
  • A set of labeled documents for each author.

• Target:
  • Identify the author of anonymous documents.
Approach

• Identification is a multi-class classification problem.
  – Classes: authors
  – Training examples: labeled documents
  – Test examples: anonymous documents
Roadmap

• Issues of large scale
• Dataset
• Experimental results
• Conclusion
• Future work
Machine Learning Framework

- Feature extraction
- Feature selection
- Training classifiers
- Classifying anonymous documents

Scale impacts every part
# Feature Extraction

## Writeprints Features

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>number of words/characters in post</td>
<td>2</td>
</tr>
<tr>
<td>Vocabulary richness</td>
<td>Yule’s $K^2$ and frequency of <em>hapax legomena</em>, <em>dis legomena</em>, etc.</td>
<td>11</td>
</tr>
<tr>
<td>Word shape</td>
<td>frequency of words with all upper-case letters, all lower-case, etc.</td>
<td>5</td>
</tr>
<tr>
<td>Word length</td>
<td>frequency of words that have 1–20 characters</td>
<td>20</td>
</tr>
<tr>
<td>Letters</td>
<td>frequency of a to z, ignoring case</td>
<td>26</td>
</tr>
<tr>
<td>Digits</td>
<td>frequency of 0 to 9</td>
<td>10</td>
</tr>
<tr>
<td>Punctuation</td>
<td>frequency of .?!;,,:();(-)&quot;'</td>
<td>11</td>
</tr>
<tr>
<td>Special characters</td>
<td>frequency of other special characters</td>
<td>21</td>
</tr>
<tr>
<td>~@#$%^&amp;*_+=[]{}/&lt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Function words</td>
<td>frequency of words like ‘the’, ‘of’, and ‘then’</td>
<td>293</td>
</tr>
<tr>
<td>Syntactic category pairs</td>
<td>frequency of every pair $(A, B)$, where $A$ is the parent of $B$ in the parse tree</td>
<td>789</td>
</tr>
</tbody>
</table>
Syntactic Features

Previous feature:

Our feature:

A sample parse tree produced by the Stanford Parser.

~1200 features!
Machine Learning Framework

Feature extraction

Feature selection

Training classifiers

Classifying anonymous documents
Feature Selection

- Information gain
- Document frequency

- Helpful for small scale
- Not helpful for large scale
Machine Learning Framework

1. Feature extraction
2. Feature selection
3. Training classifiers
4. Classifying anonymous documents
Classifiers

• Nearest neighbor (NN)
• Naïve Bayes (NB)
• Support vector machines (SVM)
• Regularized least square classifier (RLSC)
• Ensemble classifier
  – NN + RLSC
Regularized Least Square Classifier (RLSC)

• Comparable accuracy to SVM
• Much more scalable than SVM
• One-vs-all
  – Training binary classifier for each author
• Class imbalance
  • Subsampling a small number of negative examples
  • Cost sensitive learning.
  – Penalizing more for misclassifying positive examples
Dataset

ICWSM 2009 Dataset: ~94k blogs

Minimum 7,500 characters per blog (roughly 8 paragraphs)
Dataset: Google Profiles

<table>
<thead>
<tr>
<th>Posts</th>
<th>About</th>
<th>Photos</th>
<th>Videos</th>
<th>+1's</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>I'm a post-doctoral computer science researcher at Stanford and a CIS junior affiliate scholar. I study information privacy and security, and moonlight in tech policy. My doctoral research exposed the problems with data anonymization. My thesis, in a sentence, is that the level of anonymity that consumers expect—and companies claim to provide—in published or outsourced databases is fundamentally unrealizable.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bragging rights</td>
<td>Many, many years ago I was in the International Math Olympiad. Since then my math ability has steadily gotten worse. I've also forgotten half a dozen languages and I'm down to about 1.5.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

~6,000 blogs
~3,600 authors
Data Size

- ~6,000 blogs
- ~94,000 blogs
- 24 posts
- ~300 words
Experimental Design

• **Post-to-blog experiment**
  - Identifying a single or a few anonymous posts.
  - Test posts: random sample a few (e.g., 3) posts from each blog

• **Blog-to-blog experiment**
  - Identifying an entire blog
  - Test blogs: blogs crawled from URLs specified in Google profiles belong to the same author.
Post-to-blog

Three test posts (roughly 900 words) for each blog.

Improved to 80% later!!

20% precision
Blog-to-blog

Improved to 50% later!!

12% precision
Confidence estimation

• Mapping input/output pair of classifier to real values

• Gap Statistics
  • Similarity or distance difference between the best and second best match

• Output the prediction when ‘gap’ is bigger than some threshold
Confidence estimation

Post-to-blog: 80% precision, 50% recall

Blog-to-blog: 50% precision, 50% recall
Experiments Summary

• **Post-to-blog**
  • Best classifier: NN + RLSC.
  • Three test posts, exact match: 20% precision
  • More training/test data, exact match: 40-50%
  • Confidence estimation: 80% precision. 50% recall

• **Blog-to-blog**
  • Exact match: 12%
  • Confidence estimation: 50% precision. 50% recall
Conclusion

• We identified issues introduced by large scale author identification
• We introduced/discussed strategies to address them
• Large-scale author identification is possible!

• People should be informed
• Be careful when you post sensitive content
Future work

• Better understand what makes authors more/less fingerprintable

• Design better classifiers

• Automatically transform writing style while preserving document semantics
Thanks!