# Unmasking Pseudonymous Authors

Koppel, Schler Bonchek-Dokow

Sebastian Wilhelm

- **We have**: examples ofthe writing of a single author

- **Task**: determine if given texts were or were not written by this author

- We do not lack negative examples
  - Just because text is more similar to A does not mean it was authored by A rather than by B
- Chunking the text so we have multiple examples (if text is long)
  - Given two example sets -> determine if sets were generated in a single generation process

- Authorship Verification: Naive Approaches
  - Lining up impostors:
    - Model A vs. Not-A
    - X -> chuked -> A or not-A
    - Not-A => not author (true)
    - A => author (not true)

- Authorship Verification: Naive Approaches
  - One class learning:
    - Circumscribes all positive examples of A
    - Conclude: X is authored A if a sufficient number of chuks of X lie inside boundry

- Authorship Verification: Naive Approaches
  - Comparing A directly to X:
    - Learn a model for A vs. X
    - Assess the extent of difference between A and X using cross-validation
    - Easy to distinguish => high accuracy in cross-validation => A did not write X

- New Approach: Unmasking
  - **Idea**: small number of features can distinguish between texts (e.g. he vs. she)
  - **Solution**: determining not only if A is distinguishable from X but also <u>how great is the difference between A and X</u>

- New Approach: Unmasking

  - => unmasking:
    - Iteratively remove those features that are most useful for distinguishing between A and X
    - Gauge the speed with which cross-validation accuracy degrades as more features are removed

  - A and X by same author => differences between them will be reflected in only a small number of features

- Unmasking Applied:
  - n words with highest average frequency in Ax and X as initial feature
    - 1. Determine the accuracy results of a ten-fold cross-validation experiment for Ax against X
    - 2. Eliminate the k most strongly weighted positive and negative features
    - 3. Go to step 1

# => Degeneration curves for each pair <Ax,X>

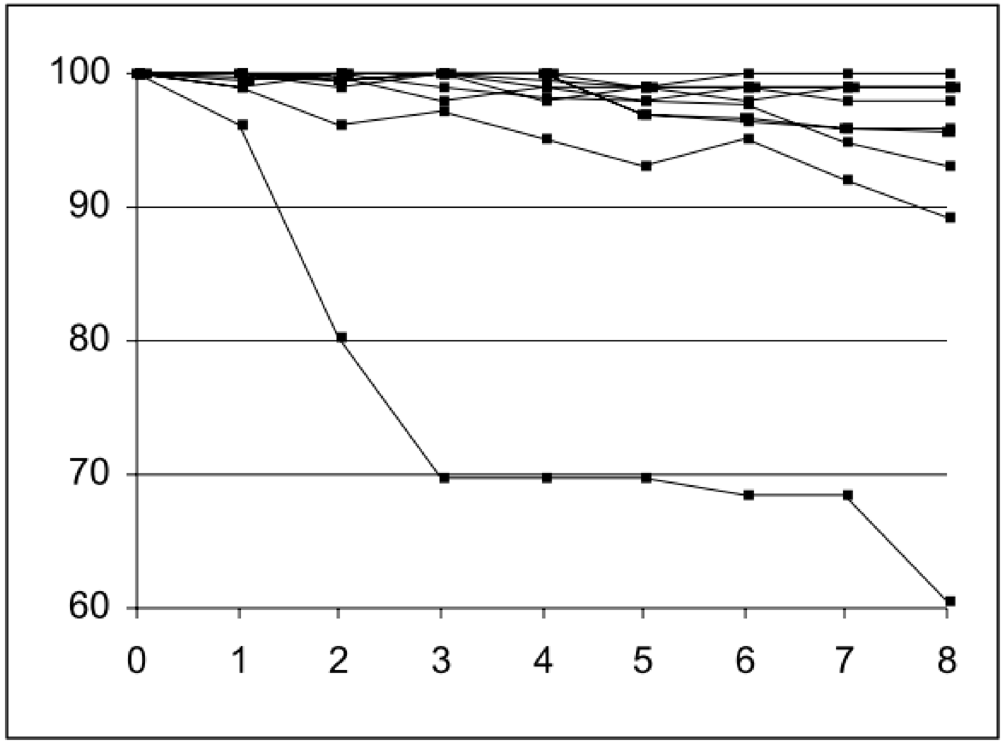


Figure 2. Unmasking *An Ideal Husband* against each of the ten authors (n=250, k=3). The curve below all the authors is that of Oscar Wilde, the actual author. (Several curves are indistinguishable.)

- Meta-learning: Identifying Same-Author Curves

  - Quantify the difference between same-author and different-author curves
  - Each curve as a numerical vector in terms of its essential features:

    - Accuracy after i elimination rounds
    - Accuracy difference between round i and i+1
    - Accuracy difference between round i and i+2
    - Highest accuracy drop in one iteration
    - Highest accuracy drop in two iterations

- Meta-learning:

  - Sort vectors in two subsets:
    - Ax, X = same author
    - Ax, X = different author

  - For all same-author curves:

    - Accuracy after 6 elimination rounds is lower than 89%
    - AND the second highest accuracy drop in two iterations is greater than 16%

| Features (n) | Features eliminated (k) | Iterations (m) | Correctly classified *same-author* (out of 20) | Correctly classified *different-author* (out of 189) | F1 (*macro average*) |
|---|---|---|---|---|---|
| 250 | 3 | 5 | 16 | 183 | 0.868 |
| | | 10 | 19 | 181 | 0.892 |
| | | 20 | 20 | 180 | 0.896 |
| | 6 | 5 | 20 | 182 | 0.916 |
| | | 10 | 20 | 180 | 0.896 |
| | | 20 | 20 | 181 | 0.906 |
| | 10 | 5 | 20 | 180 | 0.896 |
| | | 10 | 20 | 179 | 0.886 |
| 500 | 3 | 5 | 14 | 189 | 0.904 |
| | | 10 | 12 | 186 | 0.828 |
| | | 20 | 16 | 180 | 0.838 |
| | 6 | 5 | 13 | 184 | 0.826 |
| | | 10 | 18 | 180 | 0.868 |
| | | 20 | 19 | 179 | 0.873 |
| | 10 | 5 | 16 | 181 | 0.848 |
| | | 10 | 18 | 180 | 0.868 |
| | | 20 | 20 | 177 | 0.868 |
| 1000 | 3 | 5 | 11 | 189 | 0.843 |
| | | 10 | 11 | 188 | 0.831 |
| | | 20 | 12 | 183 | 0.797 |
| | 6 | 5 | 12 | 188 | 0.852 |
| | | 10 | 14 | 184 | 0.844 |
| | | 20 | 17 | 181 | 0.863 |
| | 10 | 5 | 15 | 184 | 0.862 |
| | | 10 | 16 | 182 | 0.857 |
| | | 20 | 16 | 177 | 0.812 |

Table 2 Accuracy results on the 21 book experiment for a variety of parameter setting

- Extension: Using Negative Examples

  - Learn model of A vs. Not A
  - Test each example of X (assigned to A or not-A?)
  - If many are assigned not A => X is not the author
  - BUT not true for the opposite conclusion

- Extension: Using Negative Examples

  - For each author A choose impostors A1…An (as not-A class)
    - Learn A vs. Not A
    - Learn models for each Ai vs. Not Ai
    - Test all examples in X against each other of these models

  - A(X) = percentage of examples of X classed as A
  - Ai(X)= percentage of examples of X classed as Ai
  - A(X) < Ai(X) for all i => A is not by author of X
  - Otherwise A may be by author of X

- Conclued that A is t the author of X if both methods indicate it

```
Given: anonymous book X, works of suspect author A,
       (optionally)impostors {A1,…,An}

Step 1 – Impostors method(optional)

if impostors {A1,…,An} are given then
{
    Build model M for classifying A vs. all impostors
    Test each chunk of X with built model M
    foreach impostor Ai
    {
         Build model Mi for classifying Ai vs. {A ∪ all other
                                                  impostors}
         Test each chunk of X with built model Mi
     }
     If for some Ai number of chunks assigned to Ai > number of
                                         chunks assigned to A
     then
            return different-author
}

Step 2 - Unmasking
Build degradation curve <A,X>
Represent degradation curve as feature vector (see text)
Test degradation curve vector (see text)
    if test result positive
        return same-author
    else
        return different-author

Method Build Degradation Curve:

Use 10 fold cross validation for A against X
foreach fold
{
      Do m iterations
      {
            Build a model for A against X
            Evaluate accuracy results
            Add accuracy number to degradation curve <A,X>
            Remove k top contributing features (in each
                                  direction) from data
       }
}
```

- Alternative: Measure of Depth of Difference

  - Check number of features with significant information gain between authors

  - Not as good as unmasking

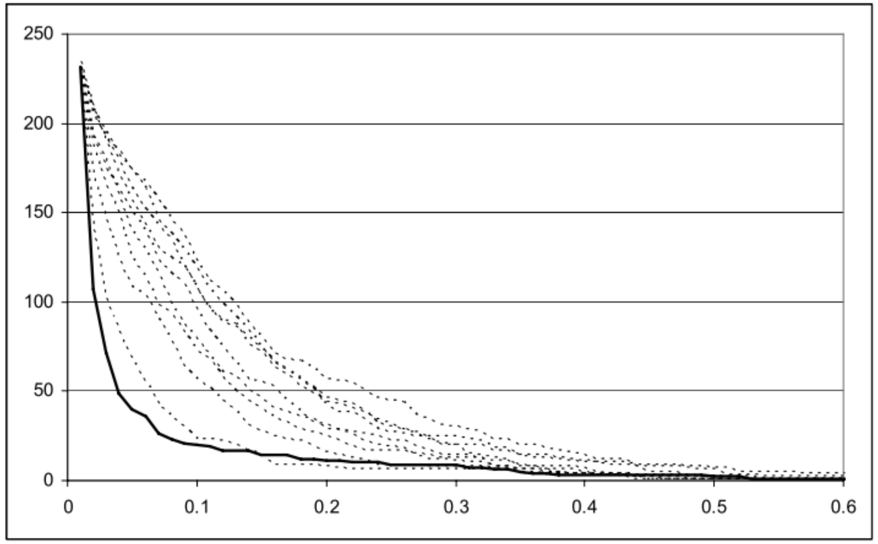


Figure 6. Information-gain curves for *An Ideal Husband* versus ten authors. The dark line is Oscar Wilde, the actual author.

- Conclusion
    - High accuracy
    - Even better with additional negative data
    - Language, period and genre independent