Statistical Analysis of Textual Data

- Statistical text analysis has a long history in literary analysis and in solving disputed authorship problems
- First (?) is Thomas C. Mendenhall in 1887

SCIENCE.-

FRIDAY, MARCH 11, 1887.

THE CHARACTERISTIC CURVES OF COMPOSITION.

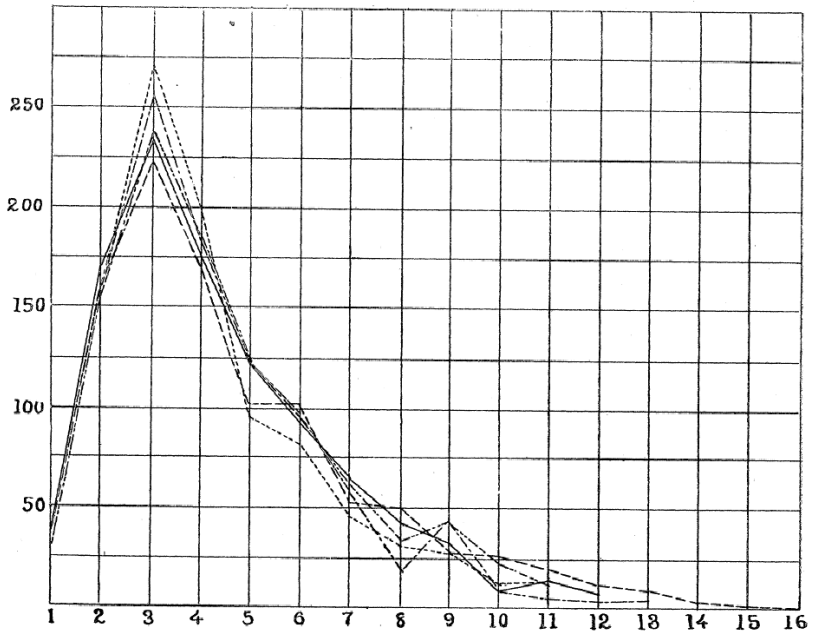


FIG. 2.—SHOWING FIVE GROUPS, OF ONE THOUSAND WORDS EACH, FROM 'OLIVER TWIST.'

Text categorization

- Automatic assignment of documents with respect to manually defined set of categories
- Applications automated indexing, spam filtering, content filters, medical coding, CRM, essay grading
- Dominant technology is supervised machine learning:
  - Manually classify some documents, then learn a classification rule from them (possibly with manual intervention)

Document Representation

- Documents usually represented as “bag of words:”

$$\mathbf{x_i} = \{x_{i1}, \ldots, x_{id}\}$$

- $\mathbf{x}_i$’s might be 0/1, counts, or weights (e.g. tf/idf, LSI)

- Many text processing choices: stopwords, stemming, phrases, synonyms, NLP, etc.

Classifier Representation

- For instance, linear classifier:

$$\text{IF } \sum_j \beta_j x_{ij} > \theta, \text{THEN } y_i = +1$$

$$\text{ELSE } y_i = -1$$

- $\mathbf{x_i}$'s derived from text of document
- y_i indicates whether to put document in category
- β_j are parameters chosen to give good classification effectiveness

Logistic Regression Model

- Linear model for log odds of category membership:

$$\ln \frac{P(y_i = +1 \mid \mathbf{x_i})}{P(y_i = -1 \mid \mathbf{x_i})} = \sum_j \beta_j x_{ij} = \hat{\mathbf{a}}\mathbf{x_i}$$

- Equivalent to

$$P(y_i = +1 \mid \mathbf{x_i}) = \frac{e^{\hat{\mathbf{a}}\mathbf{x_i}}}{1 + e^{\hat{\mathbf{a}}\mathbf{x_i}}}$$

- Conditional probability model

Logistic Regression as a Linear Classifier

- If estimated probability of category membership is greater than p, assign document to category:

$$\text{IF} \sum_j \beta_j x_{ij} > \ln \frac{p}{1-p}, \text{THEN } y_i = +1$$

- Choose *p* to optimize expected value of your effectiveness measure
- Can change measure w/o changing model

Polytomous Logistic Regression

- Sparse Bayesian (aka lasso) Logistic regression trivially generalizes to 1-of-k problems

- Laplace prior particularly appealing here:
  - Suppose 100 classes and a word that predicts class 17
  - Word gets used 100 times if build 100 binary models, or if use polytomous with Gaussian prior
  - With Laplace prior and polytomous it's used only once

1-of-K Sample Results: brittany-l

Feature Set	% errors	Number of Features
"Argamon" function words, raw tf	74.8	380
POS	75.1	44
1suff	64.2	121
1suff*POS	50.9	554
2suff	40.6	1849
2suff*POS	34.9	3655
3suff	28.7	8676
3suff*POS	27.9	12976
3suff+POS+3suff*POS+Argamon	27.6	22057
All words	23.9	52492

4.6 million parameters

89 authors with at least 50 postings. 10,076 training documents, 3,322 test documents.

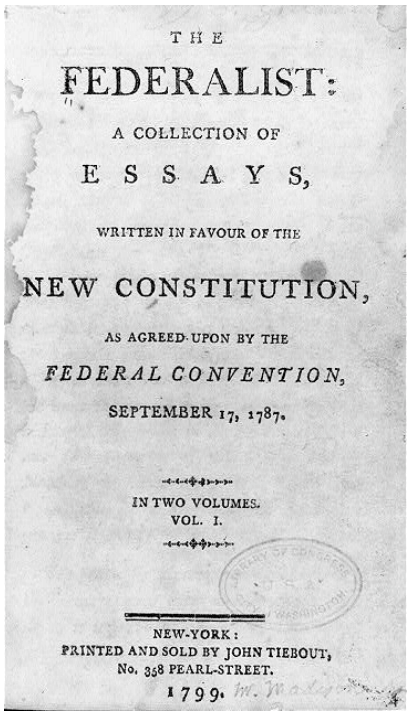
BMR-Laplace classification, default hyperparameter

The Federalist

- "The authorship of certain numbers of the 'Federalist' has fairly reached the dignity of a well-established historical controversy." (Henry Cabot Lodge, 1886)

- Historical evidence is muddled

http://www.gutenberg.org/dirs/etext91/feder16.txt


THE
FEDERALIST:
A COLLECTION OF
ESSAYS,
WRITTEN IN FAVOUR OF THE
NEW CONSTITUTION,
AS AGREED UPON BY THE
FEDERAL CONVENTION,
SEPTEMBER 17, 1787.
IN TWO VOLUMES.
VOL. I.
NEW-YORK:
PRINTED AND SOLD BY JOHN TIEBOUT,
No. 358 PEARL-STREET.
1799.

Paper Number	Author
1	Hamilton
2-5	Jay
6-9	Hamilton
10	Madison
11-13	Hamilton
14	Madison
15-17	Hamilton
18-20	Joint: Hamilton and Madison
21-36	Hamilton
37-48	Madison
49-58	**Disputed**
59-61	Hamilton
62-63	**Disputed**
64	Jay
65-85	Hamilton

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 302 JUNE, 1963 *Volume 58*

INFERENCE IN AN AUTHORSHIP PROBLEM[1,2]

A comparative study of discrimination methods applied to the authorship of the disputed *Federalist* papers

FREDERICK MOSTELLER
Harvard University
and
Center for Advanced Study in the Behavioral Sciences
AND
DAVID L. WALLACE
University of Chicago

- Used function words with Naïve Bayes with Poisson and Negative Binomial model
- Out-of-sample predictive performance

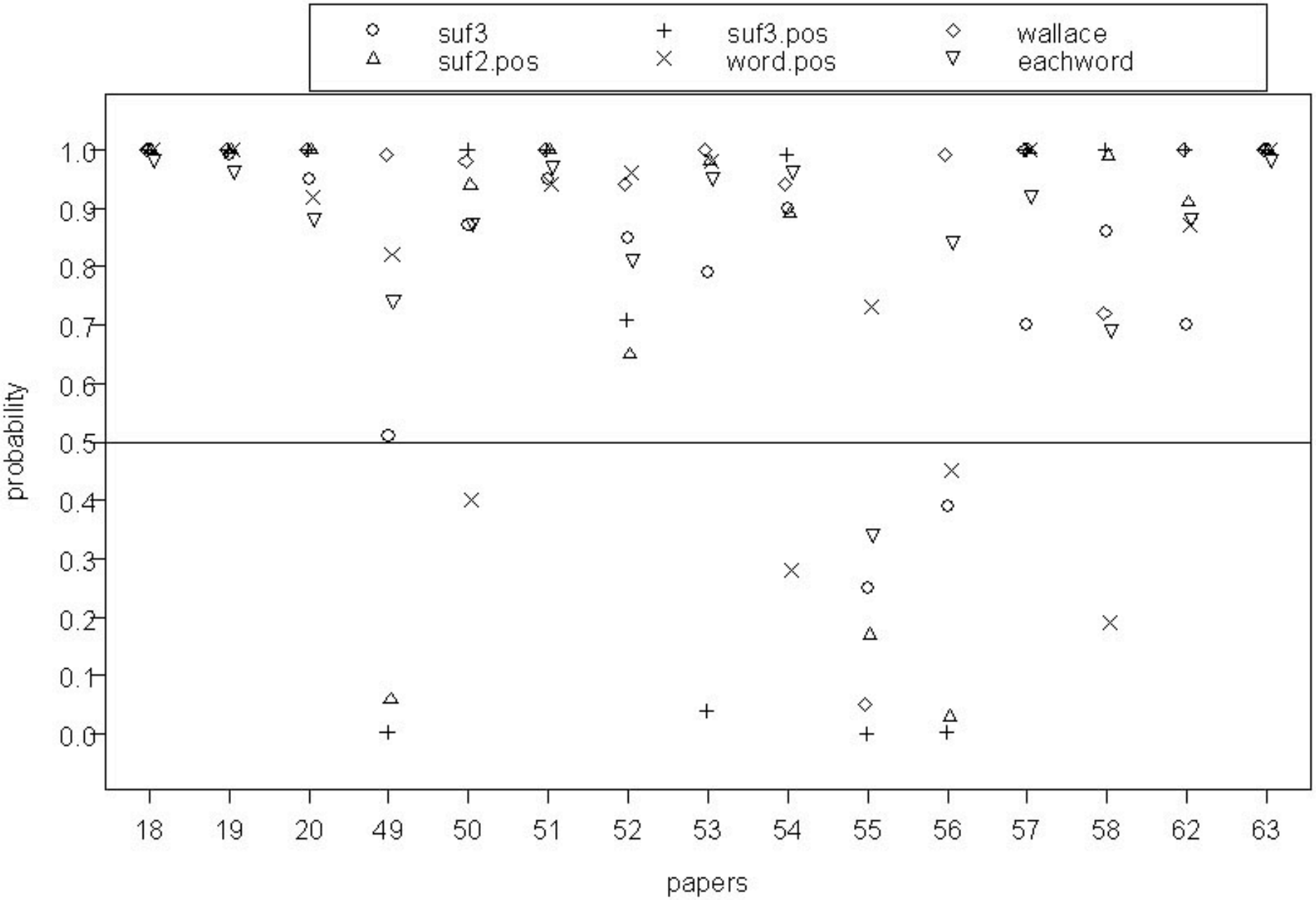
F. Summing up

In summary, the following points are clear:

1) Madison is the principal author. These data make it possible to say far more than ever before that the odds are enormously high that Madison wrote the 12 disputed papers. Weakest support is given for No. 55. Support for Nos. 62 and 63, most in doubt by current historians, is tremendous.

Feature Set	10-fold Error Rate
Charcount	0.21
POS	0.19
Suffix2	0.12
Suffix3	0.09
Words	0.10
Charcount+POS	0.12
Suffix2+POS	0.08
Suffix3+POS	0.04
Words+POS	0.08
484 features	0.05
Wallace features	0.05
Words (>=2)	0.05
Each Word	0.05

four papers to Hamilton


suf3
suf2.pos
suf3.pos
word.pos
wallace
eachword
probability
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
18
19
20
49
50
51
52
53
54
55
56
57
58
62
63
papers

Conclusion

- Authorship attribution needs to pay serious attention to predictive uncertainty deriving from representational issues.