Mining Email Content for Author Identification Forensics

O. de Vel et al. 2001

Why E-Mails?

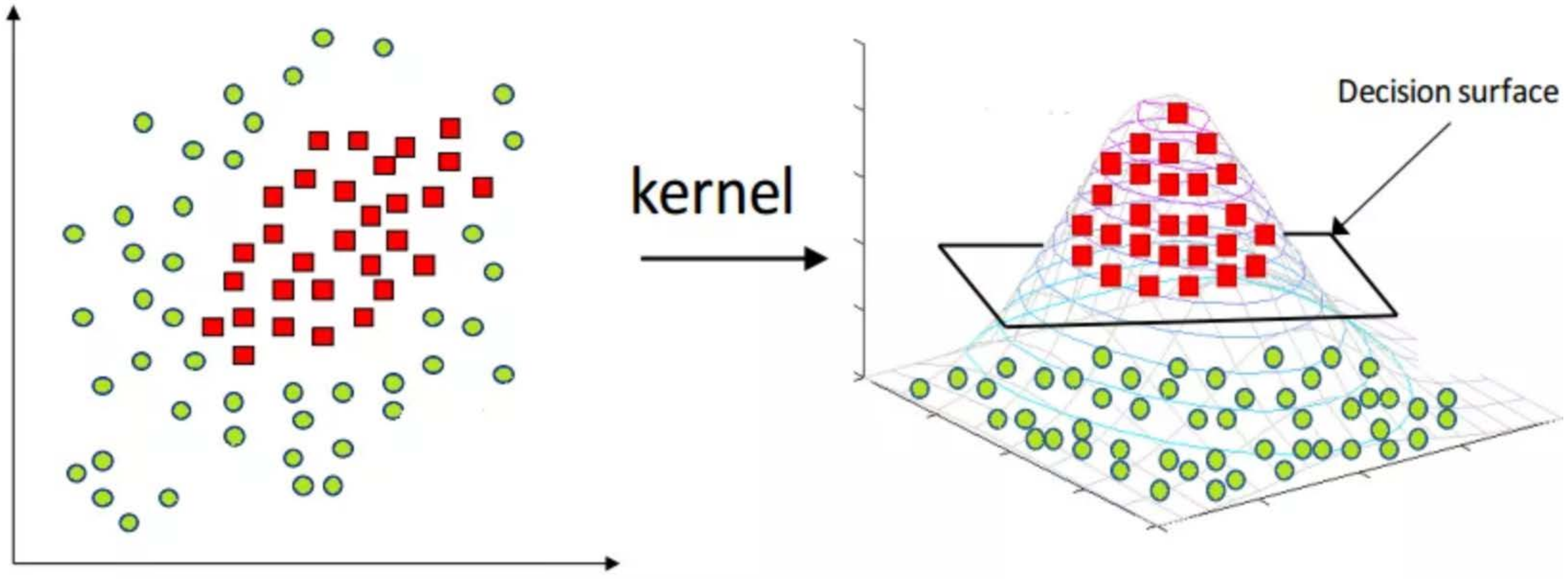
- Exchange with Solène, Arkel & Hervé (officers) from French ministry of Justice + Finance
  - Specialists in Text Forensics/E-Mail Forensics
- “Macron-Leaks”

Structure of the paper (I)

- Introduction:
  - Basic outline of relevance | published in: 2001(!)
  - …
- Authorship „Categorisation“
- Specificities of E-Mail Authorship Categorisation

Methodology: Support Vector Machine Classifier

- Methodology: Support Vector Machine Classifier
  - Structural risk minimisation (minimum generalisation error)



Data: „E-Mail-Corpus“

- Data: „E-Mail-Corpus“
  - Not further specified („private and ethical considerations“)
  - Argument against public E-Mail datasets (authors are from another era, to be fair)
  - 156 Documents, 12000 words per author for three topics (movies, food, travel)

Experimental Methodology (I) - 170 style marker attributes

- Number of blank lines/total number of lines (yet to better capture "line structure")
- Average sentence length
- Average word length (number of characters)
- Vocabulary richness i.e., V=M
- Total number of function words/M (lacking a clear definition of "all-purpose function words)
- Function word frequency distribution (122 features) (used 122 most frequent words, is this ok?)
- Total number of short words/M
- Count of hapax legomena/M
- Count of hapax legomena/V

- Total number of characters in words/C
- Total number of alphabetic characters in words/C
- Total number of upper-case characters in words/C
- Total number of digit characters in words/C
- Total number of white-space characters/C
- Total number of space characters/C (difference to white-space?)
- Total number of space characters/number white-space characters
- Total number of tab spaces/C
- Total number of tab spaces/number white-space characters
- Total number of punctuations/C
- Word length frequency distribution/M (30 features) (Computer too slow for large dataset with >6000 emails)

Experimental Methodology (II) – 21 structure marker attributes

- Has a greeting acknowledgment
- Uses a farewell acknowledgment (both primitively implemented by hand)
- Contains signature text
- Number of attachments
- Position of requoted text within e-mail body
- HTML tag frequency distribution/total number of HTML tags (16 features) (depends on data format)

See pdf

Experimental Methodology (III) – SVM classifier

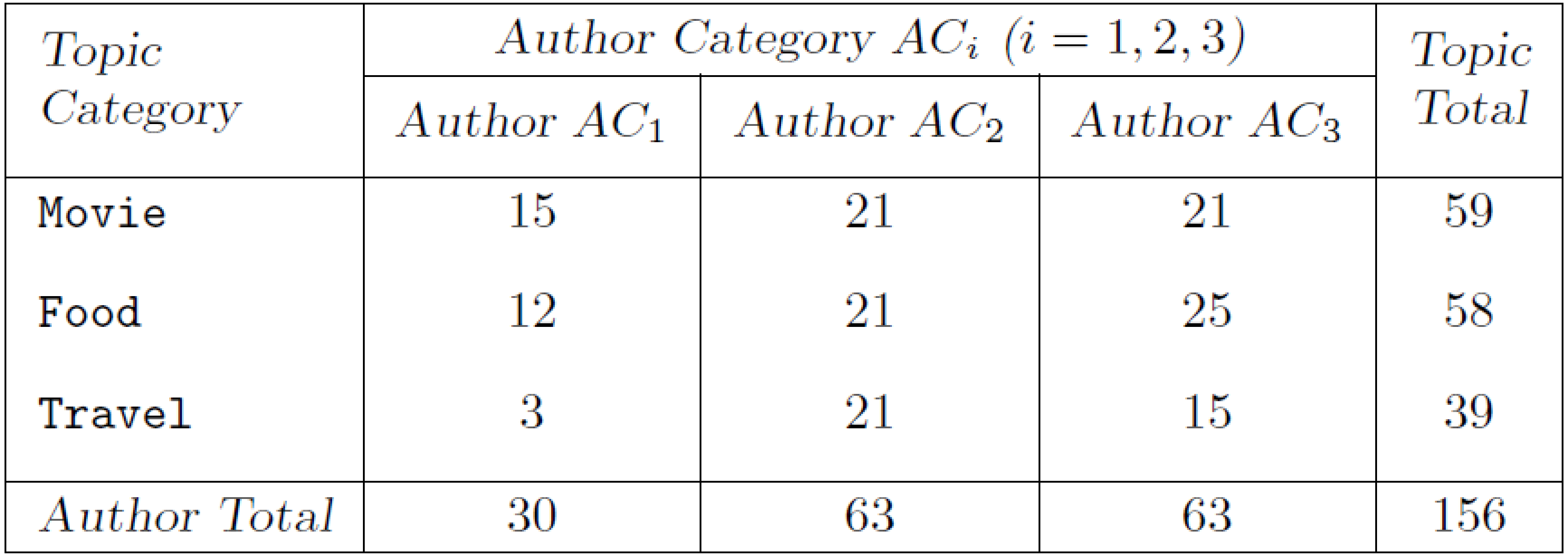
- SVM(light)-Classifier used (implementation of Vapnik‘s support VM)
- Exploration with several kernels maximal results with polynomial
- LOQO-Optimiser used (no reference, what is this?)
- Q two-way classification-models with Q-two-way classification matrices

Experimental Methodology (III) – SVM classifier

- SVM(light)-Classifier used (implementation of Vapnik‘s support VM)
- Exploration with several kernels maximal results with polynomial
  - I had much better results with radial kernel, tho
- LOQO-Optimiser used (no reference, what is this?)
- Q two-way classification-models with Q-two-way classification matrices

Evaluation

$$F_1 = \frac{2RP}{(R + P)}$$



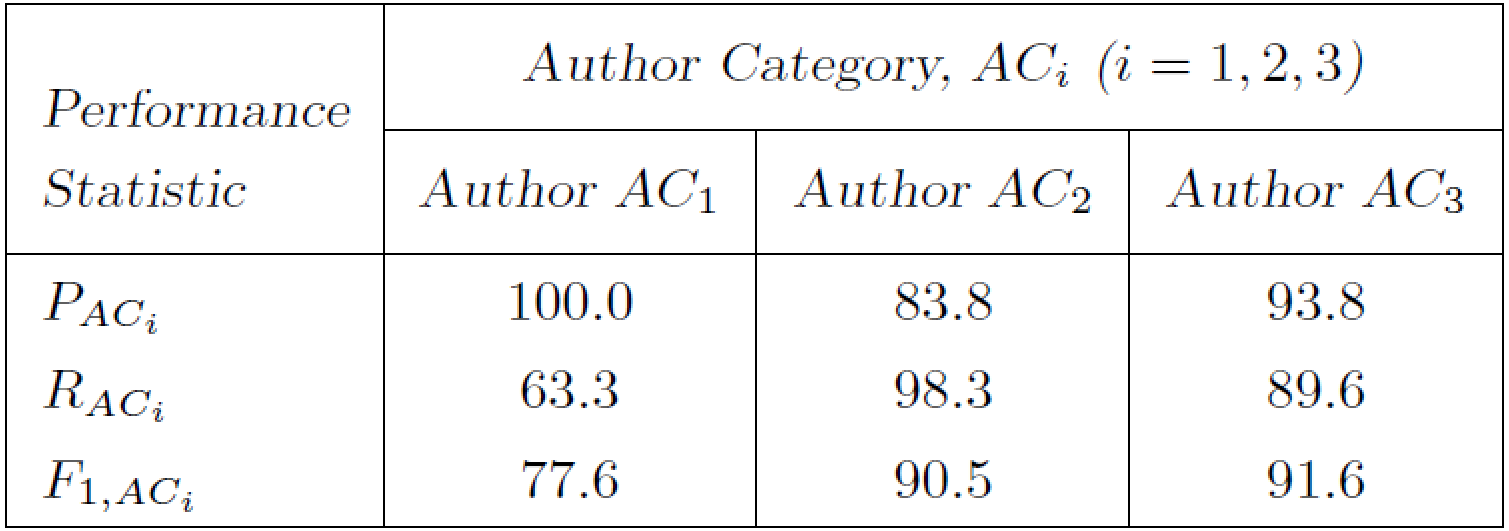
Topic Category	*Author Category AC_i $(i = 1, 2, 3)$*			*Topic Total*
	Author AC_1	*Author AC_2*	*Author AC_3*	
Movie	15	21	21	59
Food	12	21	25	58
Travel	3	21	15	39
Author Total	30	63	63	156

$$F_1^{(M)} = \frac{\sum_{i=1}^{N_{AC}} F_{1,AC_i}}{N_{AC}}$$

$$F_{1,AC_i} = \frac{2R_{AC_i} P_{AC_i}}{(R_{AC_i} + P_{AC_i})}$$

3 experiments

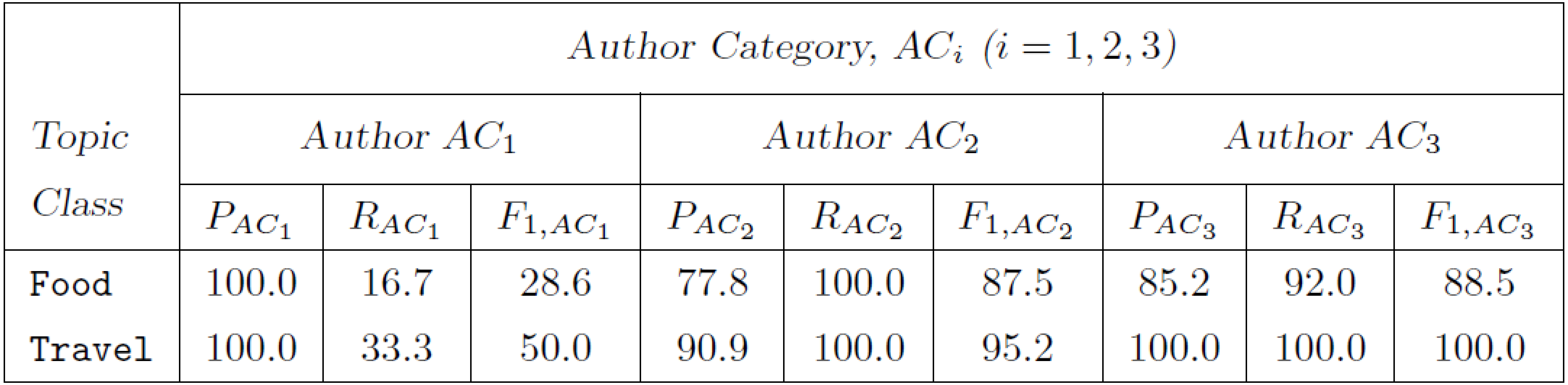
- 1: aggregated topic class (single-class)



Performance Statistic	Author Category, AC_i $(i = 1, 2, 3)$		
	Author AC_1	Author AC_2	Author AC_3
P_{AC_i}	100.0	83.8	93.8
R_{AC_i}	63.3	98.3	89.6
F_{1,AC_i}	77.6	90.5	91.6

3 experiments

- 2: Seperate Topic class (trained on different topic)



Topic Class	*Author Category,* AC_i $(i = 1, 2, 3)$								
	Author AC_1			*Author* AC_2			*Author* AC_3		
	P_{AC_1}	R_{AC_1}	F_{1,AC_1}	P_{AC_2}	R_{AC_2}	F_{1,AC_2}	P_{AC_3}	R_{AC_3}	F_{1,AC_3}
`Food`	100.0	16.7	28.6	77.8	100.0	87.5	85.2	92.0	88.5
`Travel`	100.0	33.3	50.0	90.9	100.0	95.2	100.0	100.0	100.0

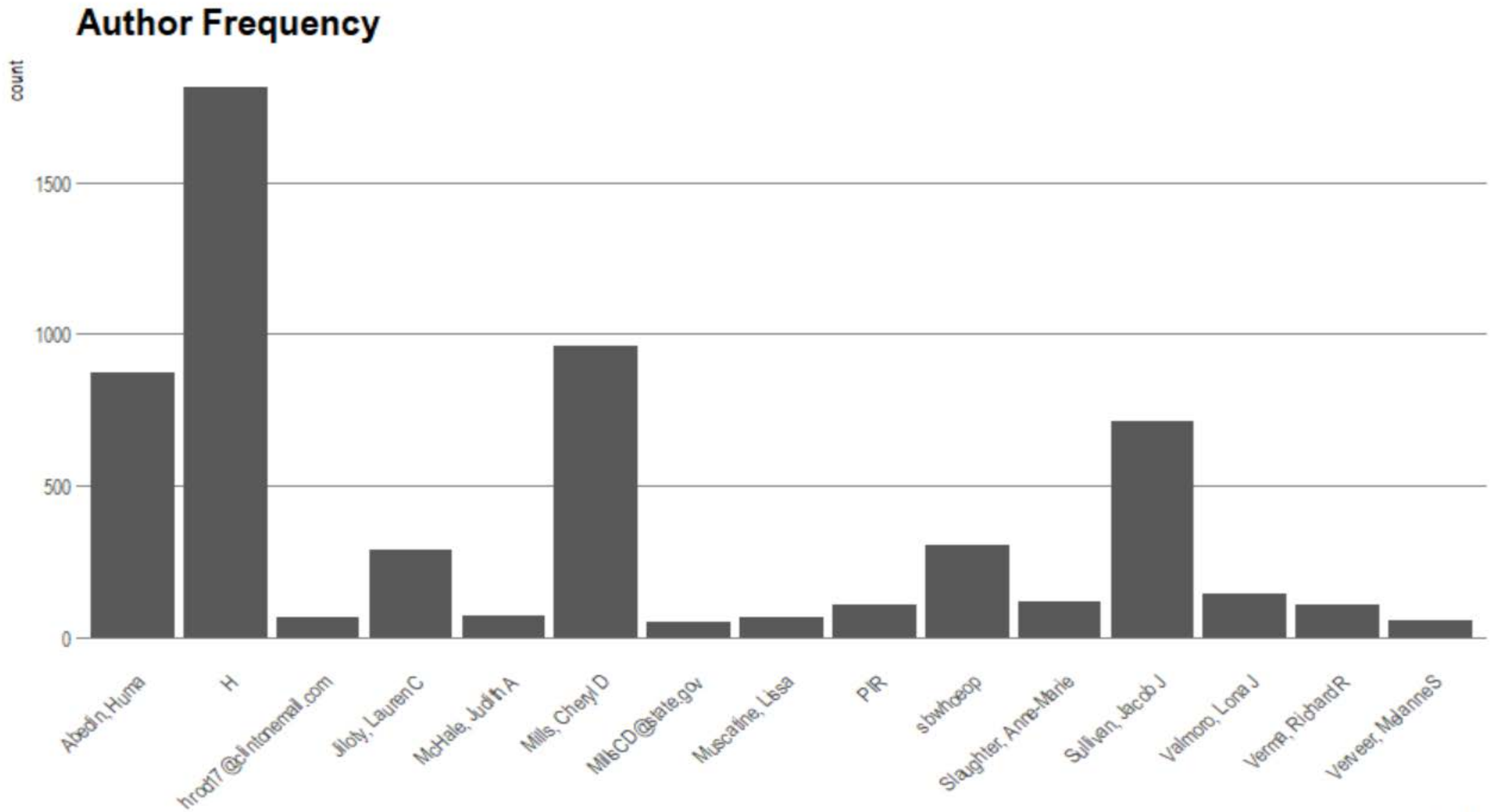
3 experiments

- 3: Function Word Type and Dimensionality
  - Some random, barely described additional experiments
  - Function word list increased from 122 to 320
  - Sets split in „parts-of-speech“ words (adverbs/auxiliaries..) and others (numbers etc.)
  - All did not improve results or deteriorated them (no concreteresults specified)

An own implementation

Hillary‘s Mails

- ~6000 non-empty mails from 216 total authors
- Topics: mostly foreign policy such as plans to invade Lybia, how to frame it, etc.



Descriptive Statistics of selected covariates

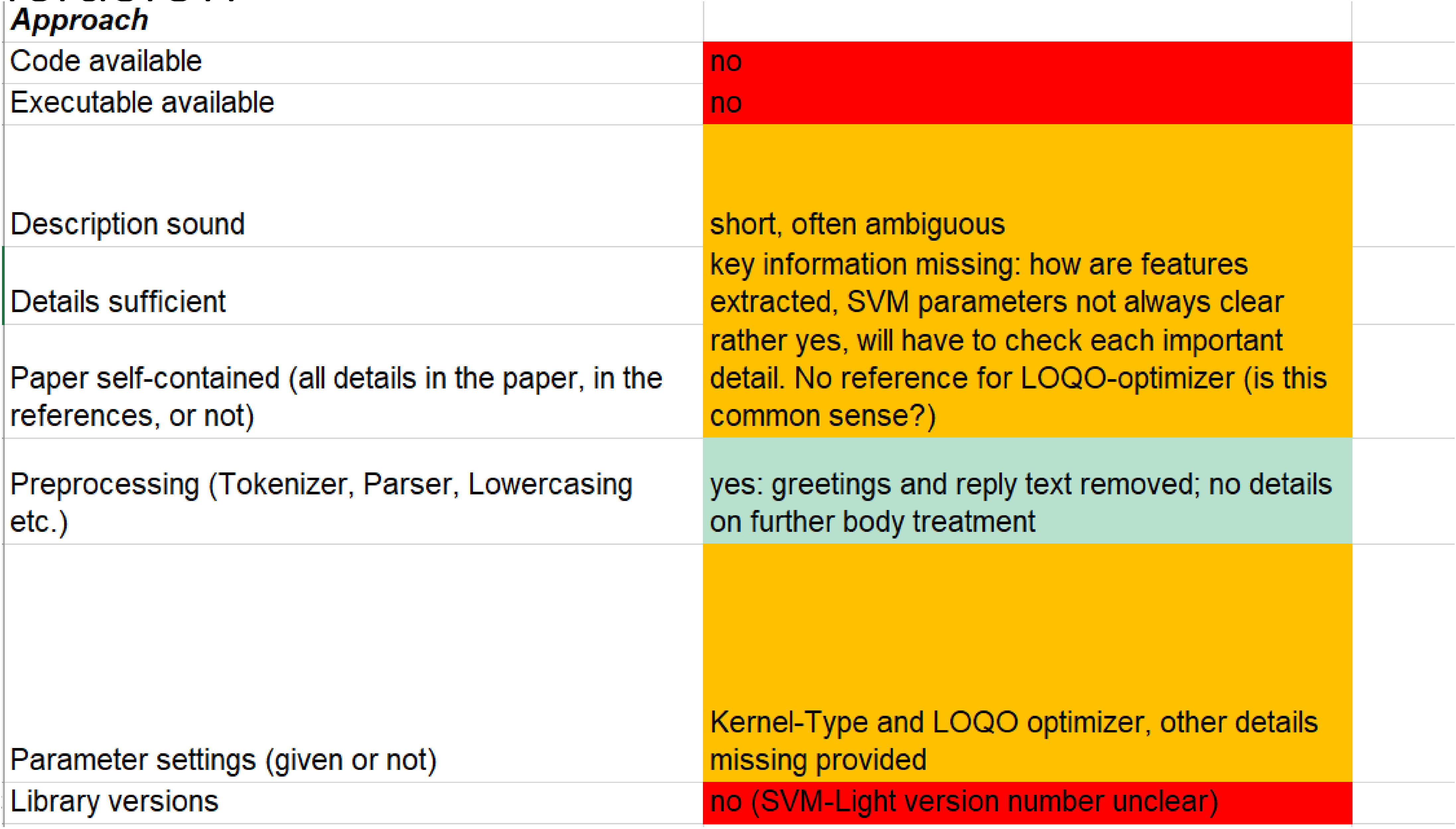
Look at different triples of authors – set 1

Observation Inequality - A Decisive Predictor!
- try out more equal triples

Also: due to computational restraints, model not trained for every triple but once globally.

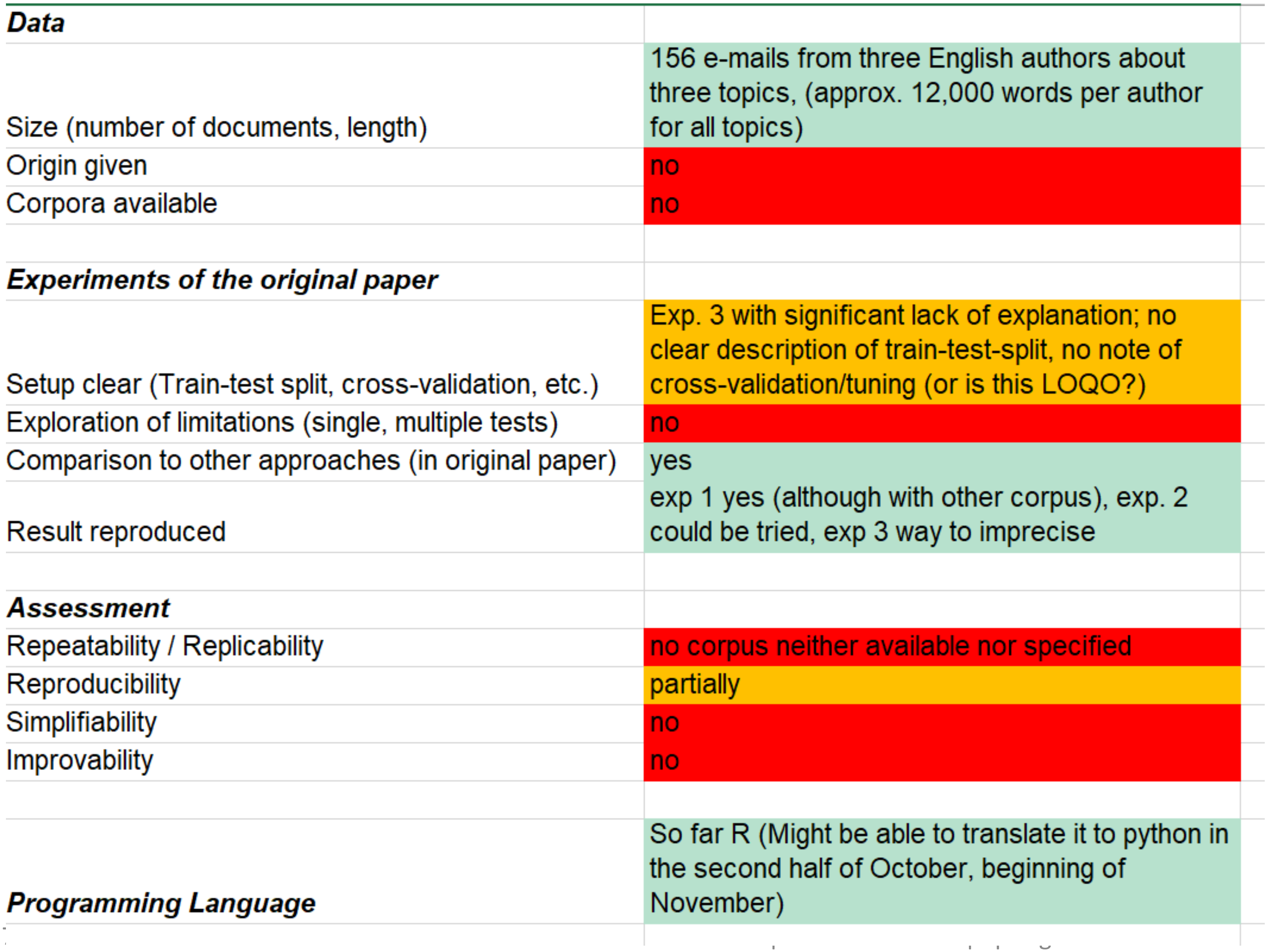
Conclusion

zeppelin universität



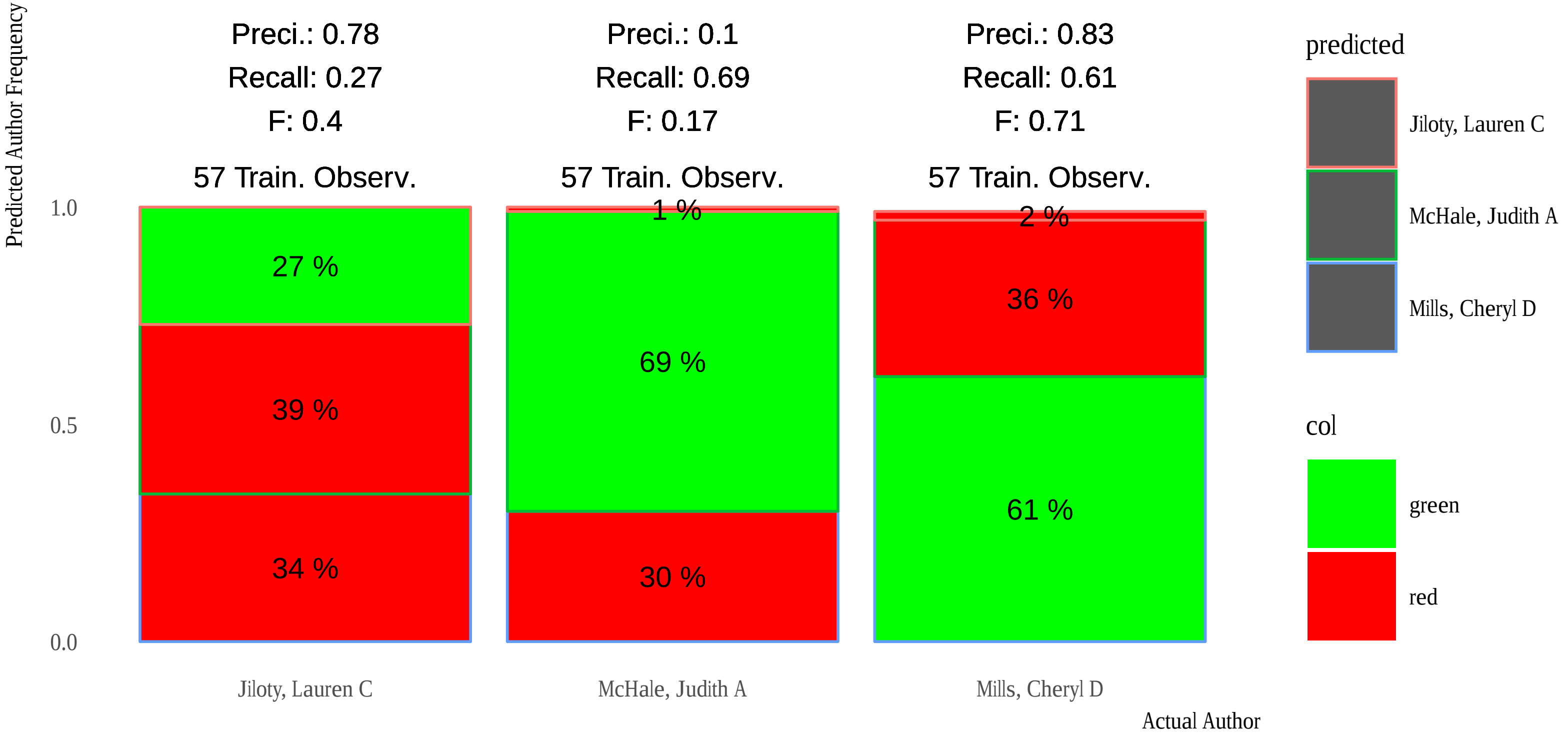
Approach	
Code available	no
Executable available	no
Description sound	short, often ambiguous
Details sufficient	key information missing: how are features extracted, SVM parameters not always clear
Paper self-contained (all details in the paper, in the references, or not)	rather yes, will have to check each important detail. No reference for LOQO-optimizer (is this common sense?)
Preprocessing (Tokenizer, Parser, Lowercasing etc.)	yes: greetings and reply text removed; no details on further body treatment
Parameter settings (given or not)	Kernel-Type and LOQO optimizer, other details missing provided
Library versions	no (SVM-Light version number unclear)

Conclusion

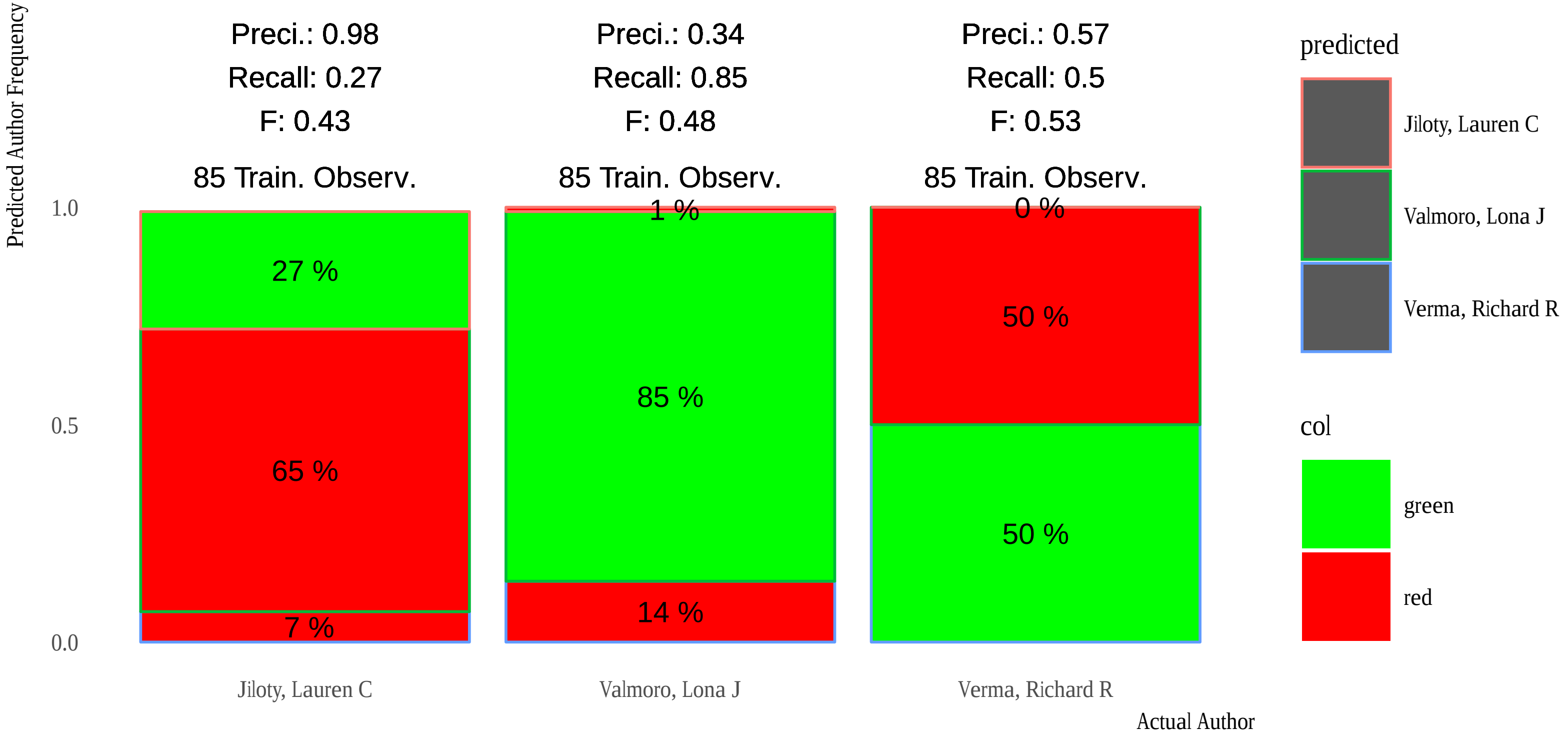


Data	
Size (number of documents, length)	156 e-mails from three English authors about three topics, (approx. 12,000 words per author for all topics)
Origin given	no
Corpora available	no
Experiments of the original paper	
Setup clear (Train-test split, cross-validation, etc.)	Exp. 3 with significant lack of explanation; no clear description of train-test-split, no note of cross-validation/tuning (or is this LOQO?)
Exploration of limitations (single, multiple tests)	no
Comparison to other approaches (in original paper)	yes
Result reproduced	exp 1 yes (although with other corpus), exp. 2 could be tried, exp 3 way to imprecise
Assessment	
Repeatability / Replicability	no corpus neither available nor specified
Reproducibility	partially
Simplifiability	no
Improvability	no
Programming Language	So far R (Might be able to translate it to python in the second half of October, beginning of November)

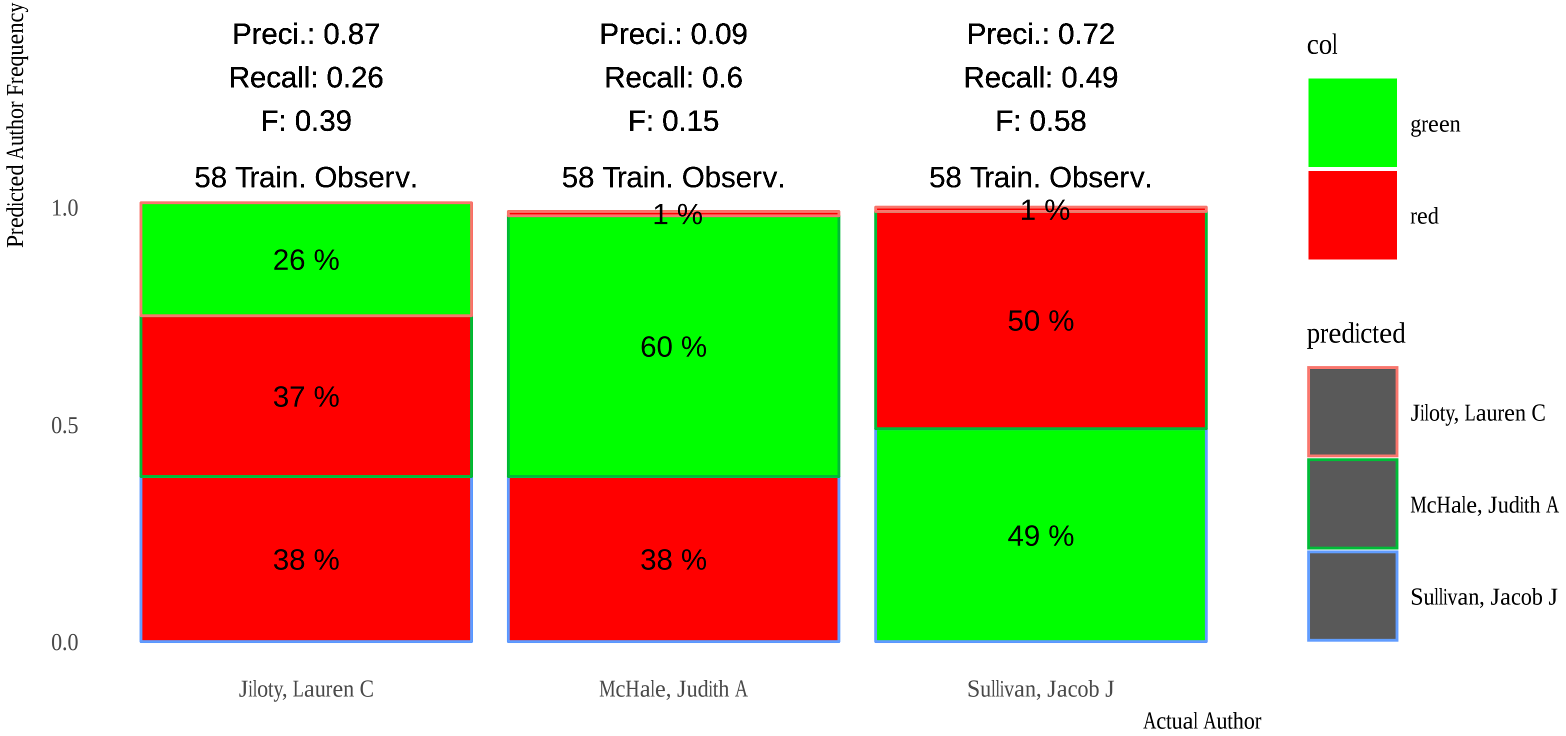
Author Sample 1



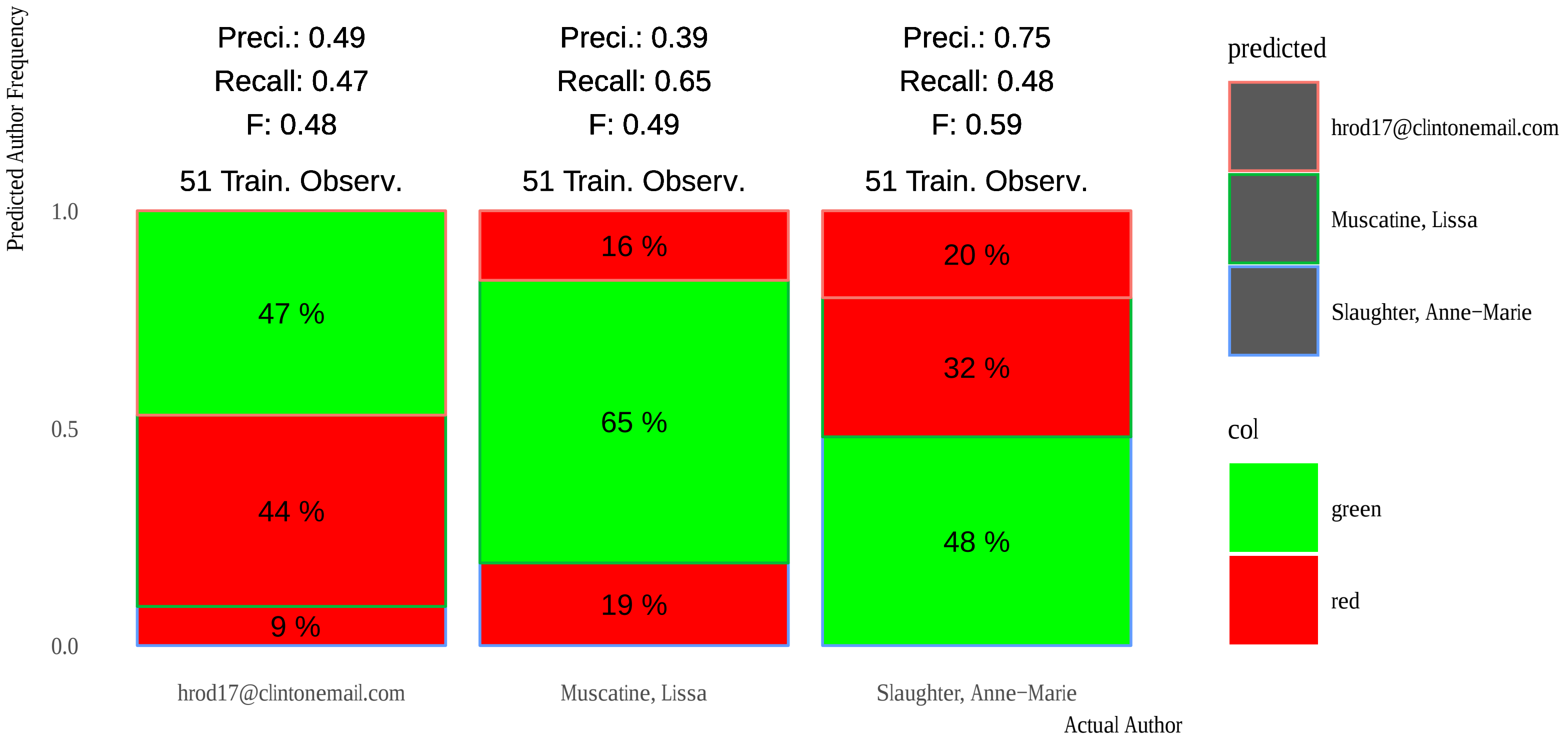
Author Sample 2



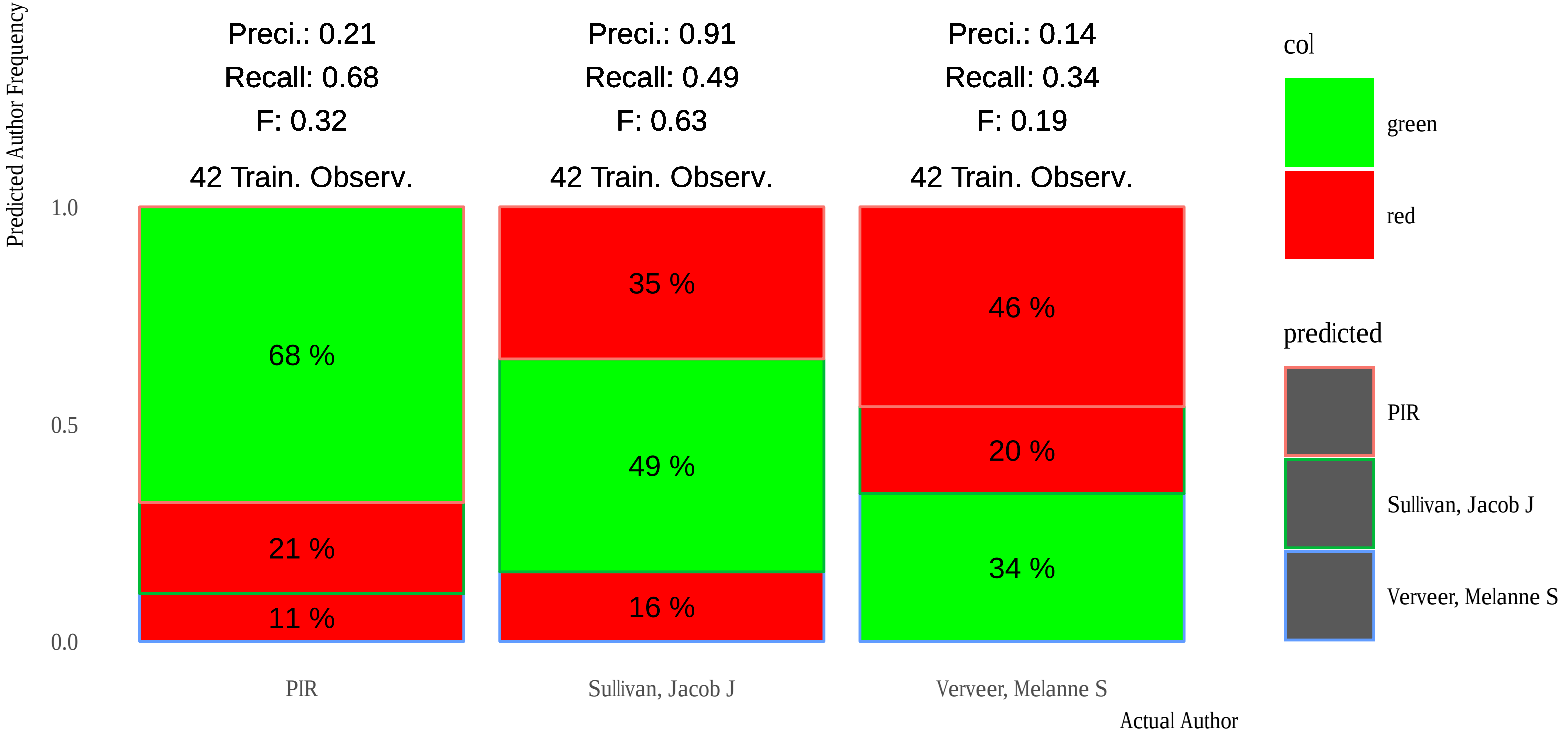
Author Sample 3



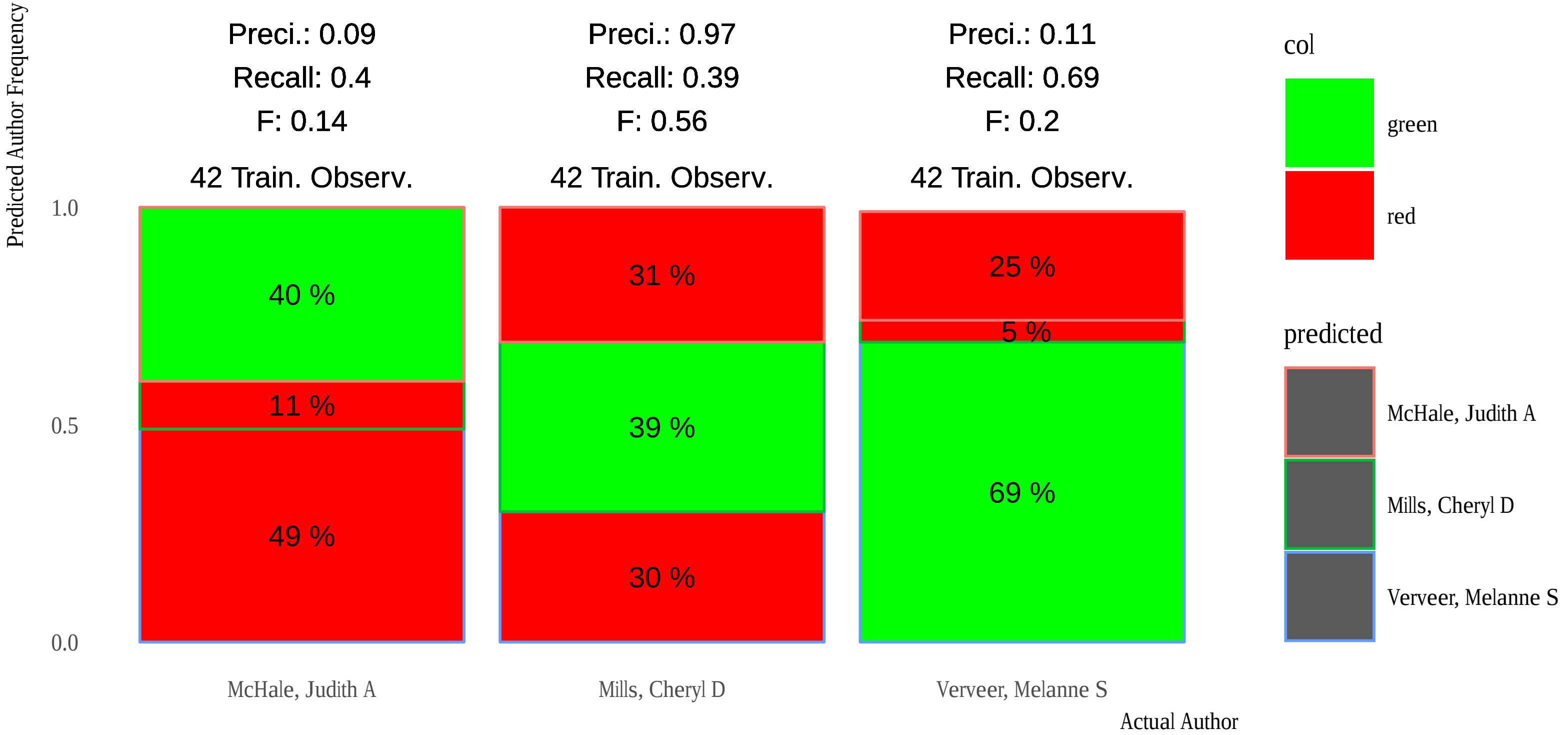
Author Sample 4



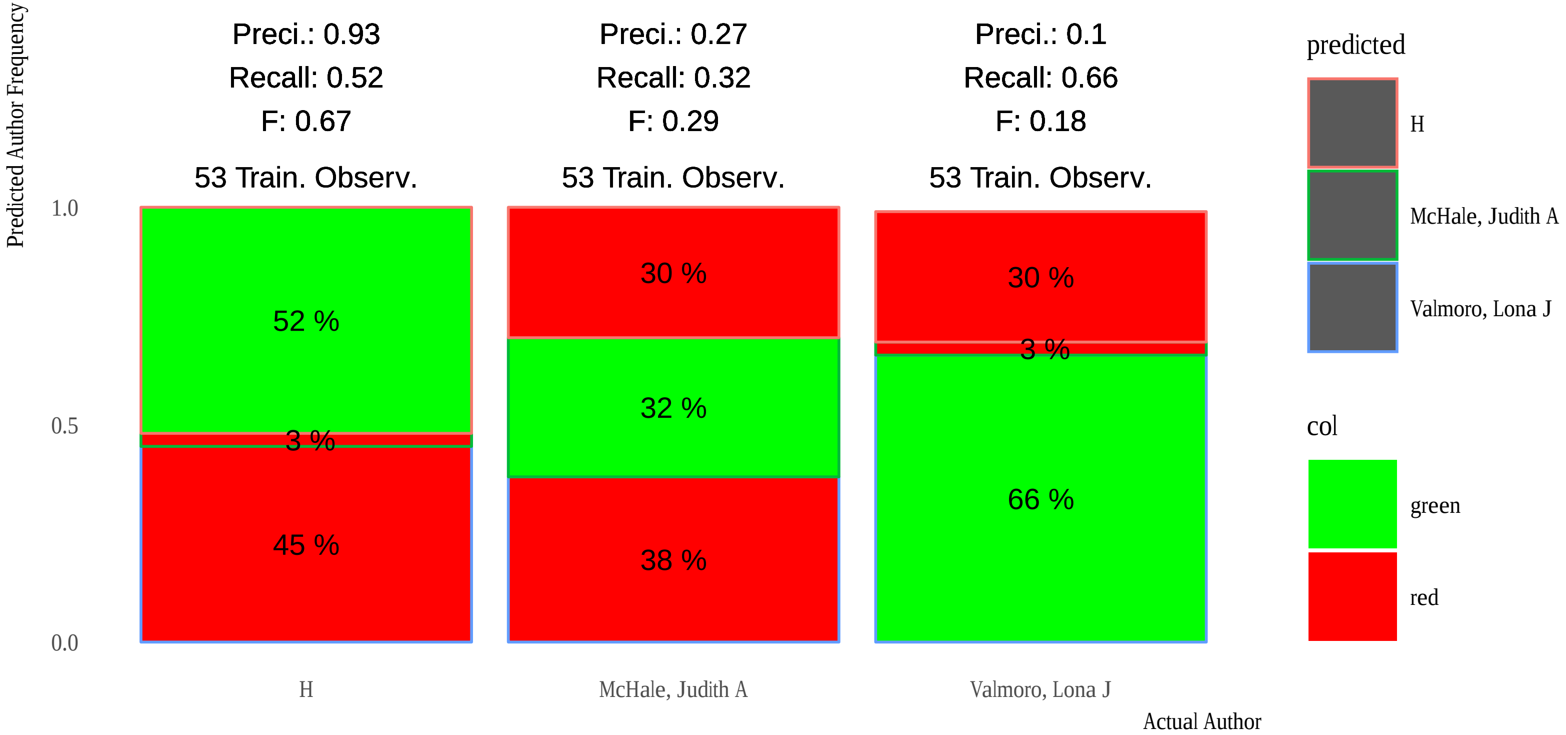
Author Sample 5



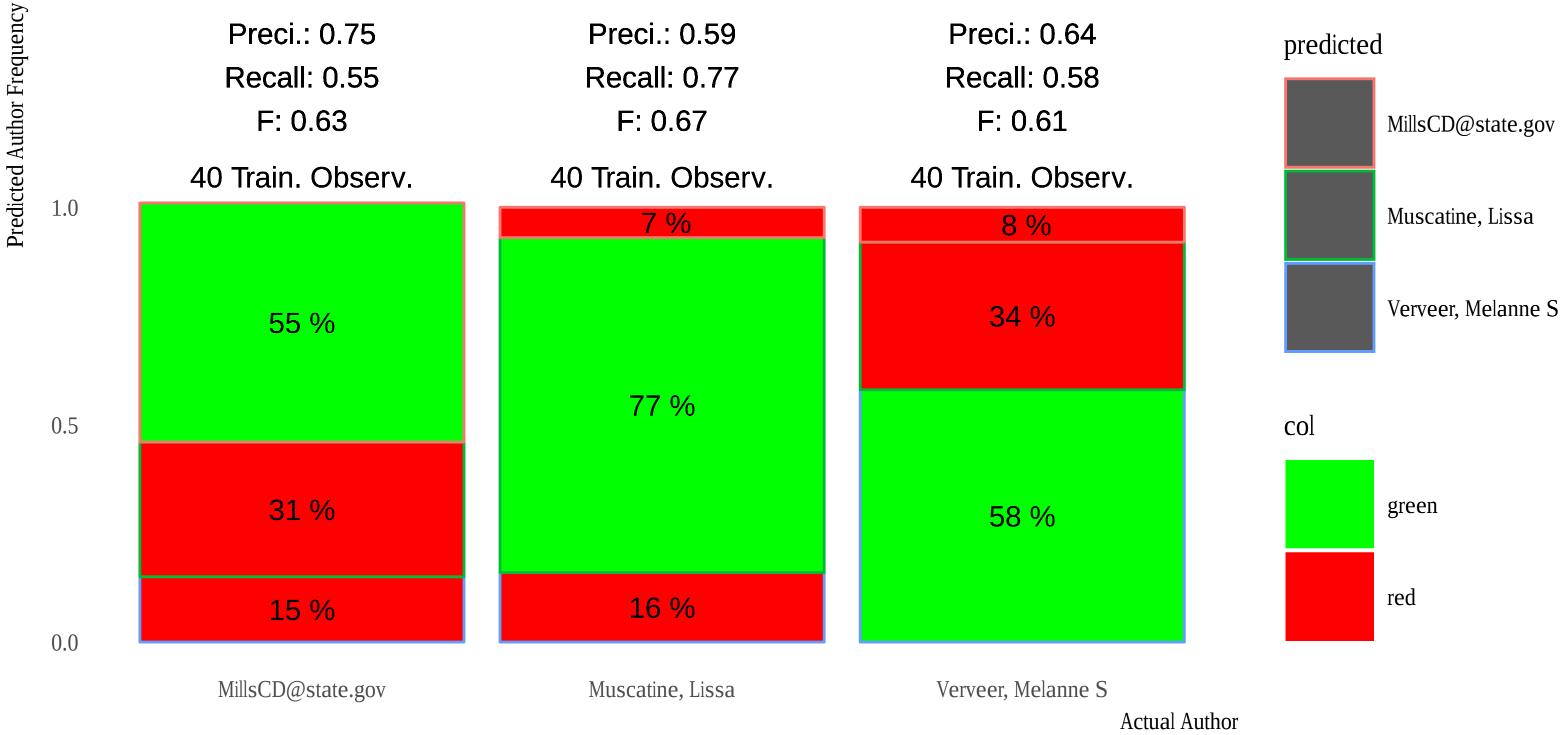
Author Sample 6



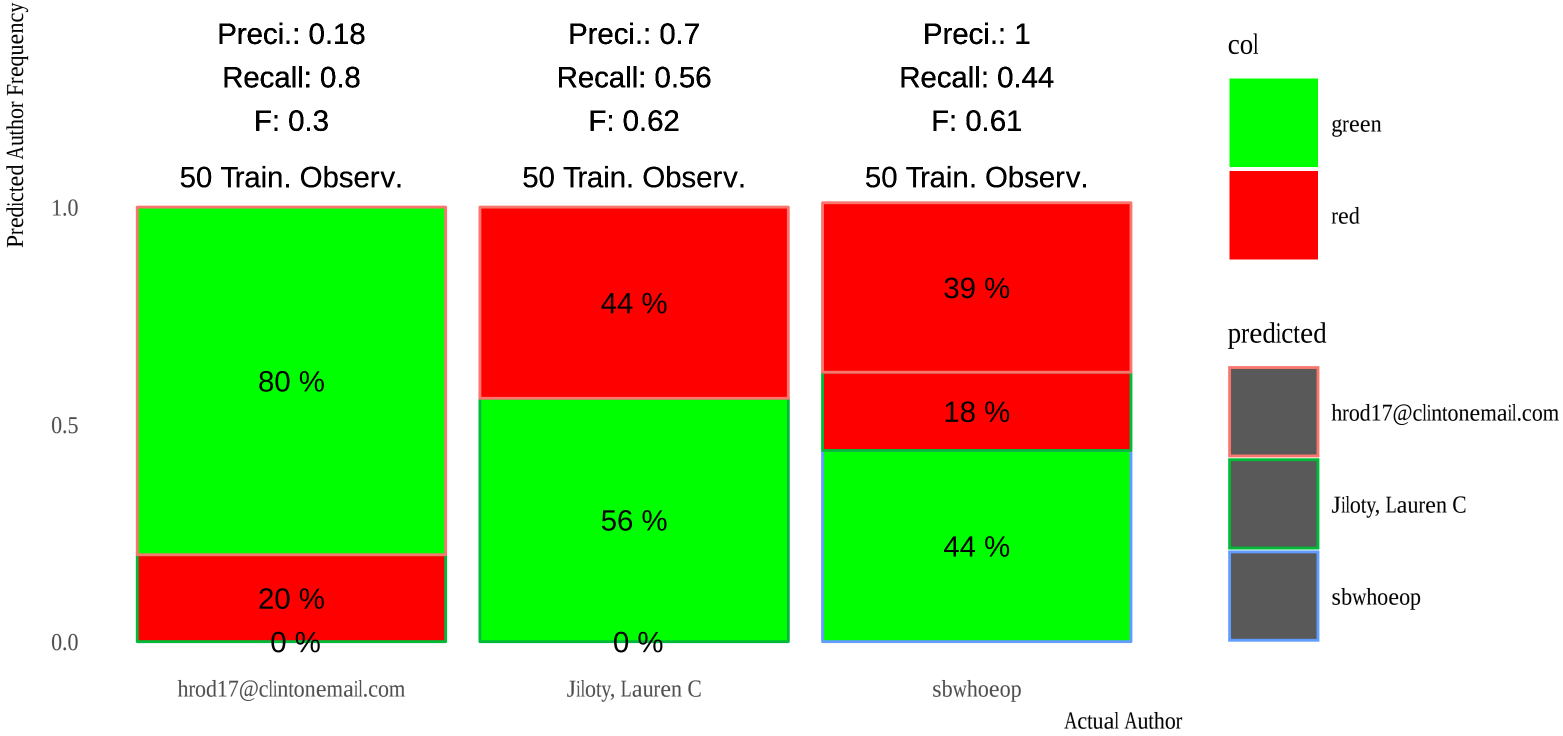
Author Sample 7



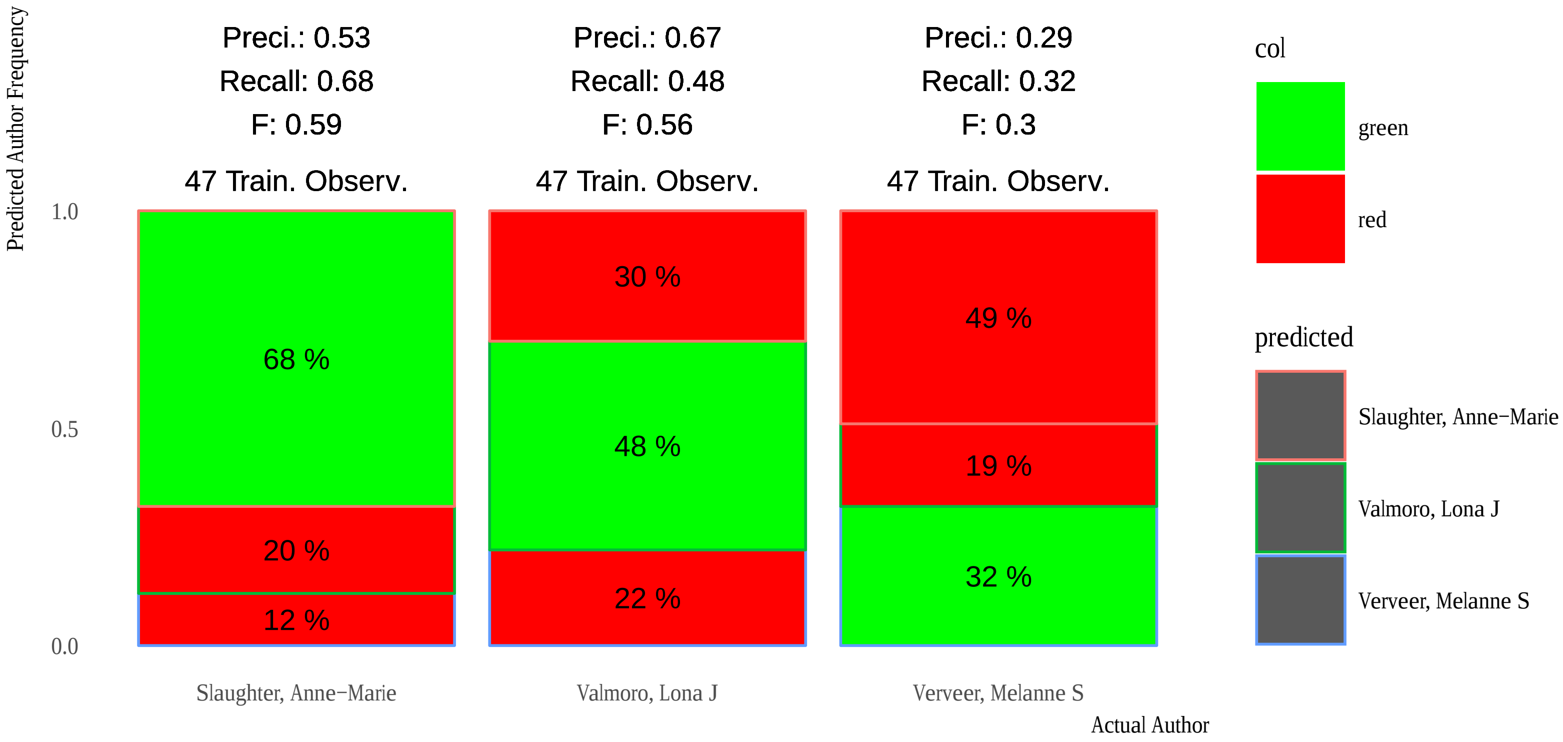
Author Sample 8



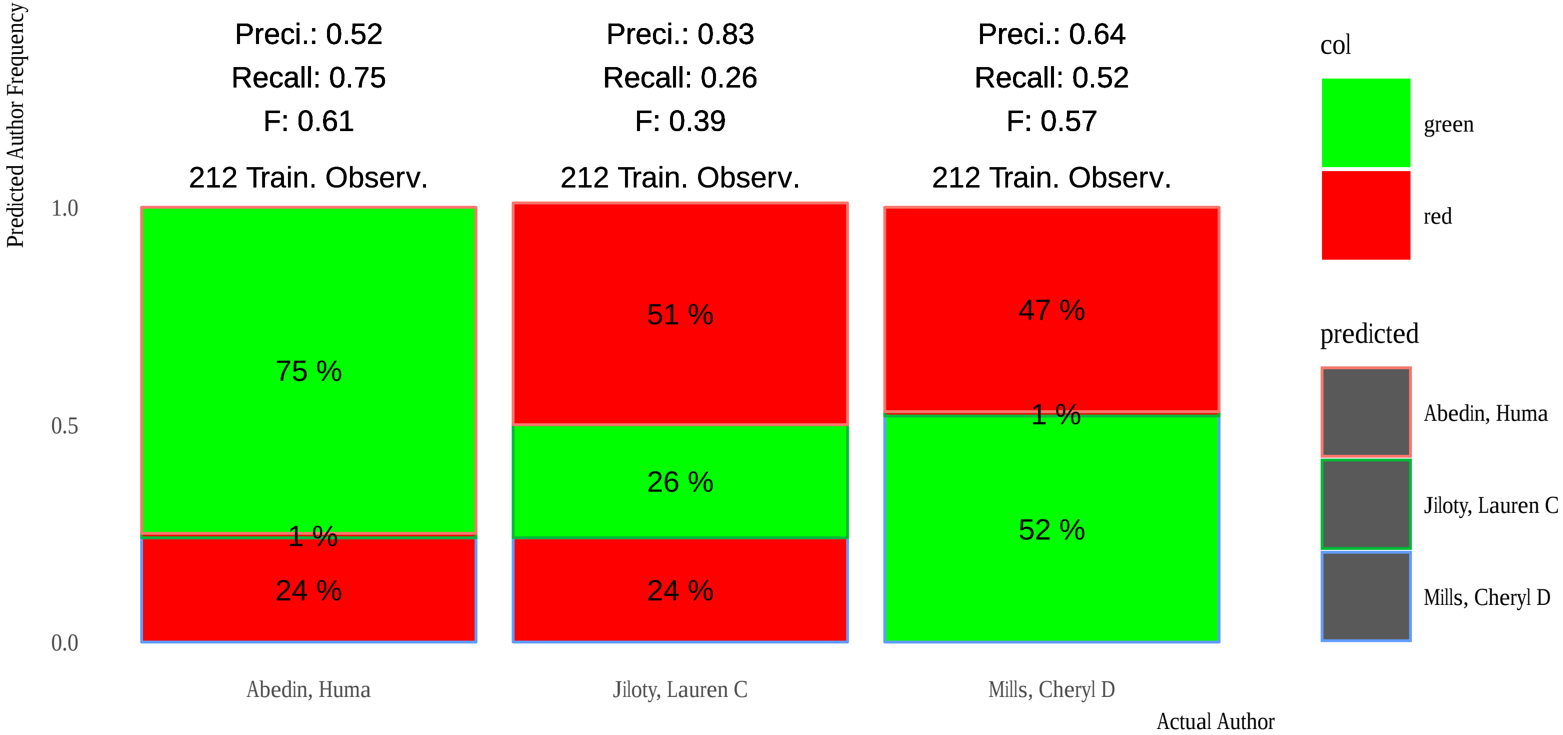
Author Sample 9



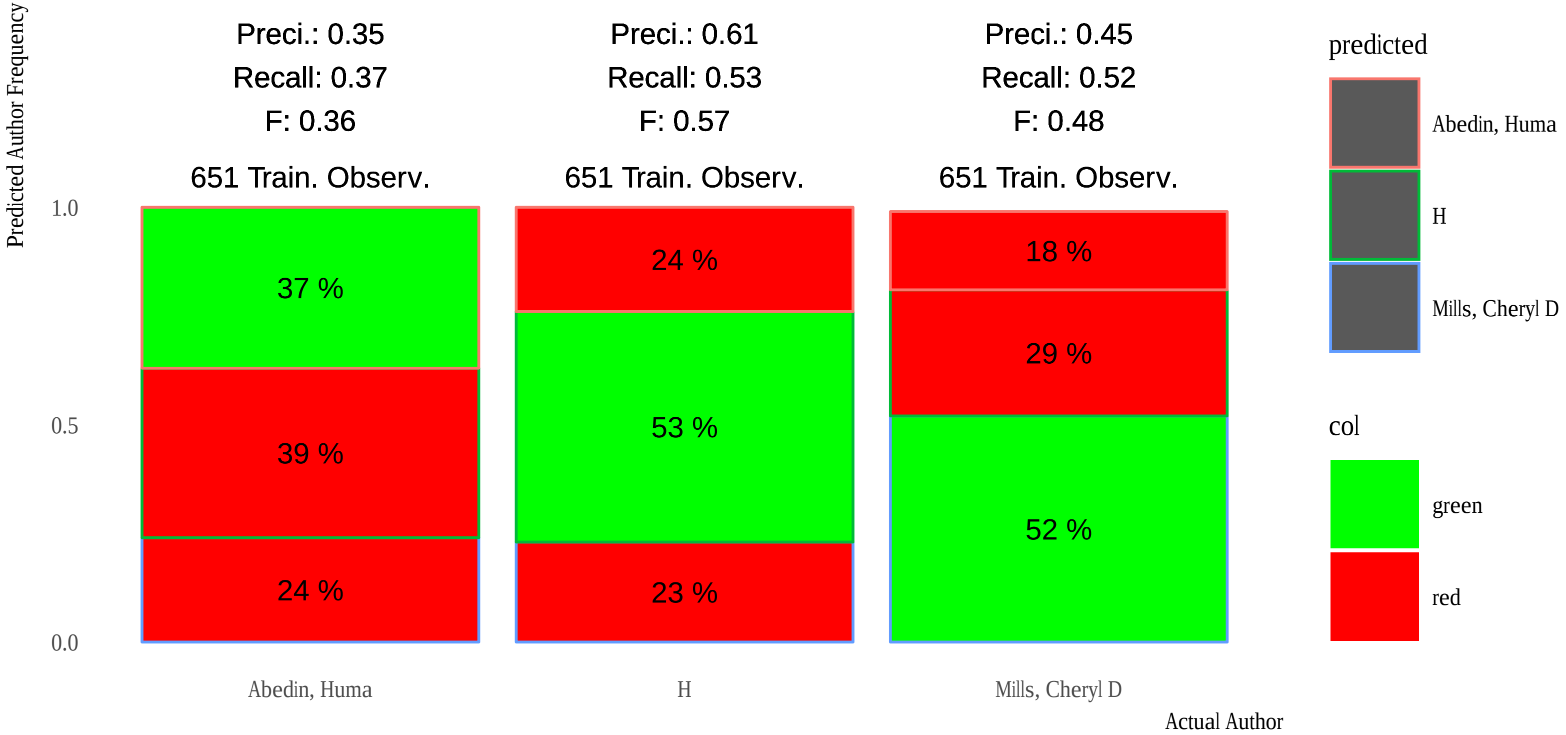
Author Sample 10



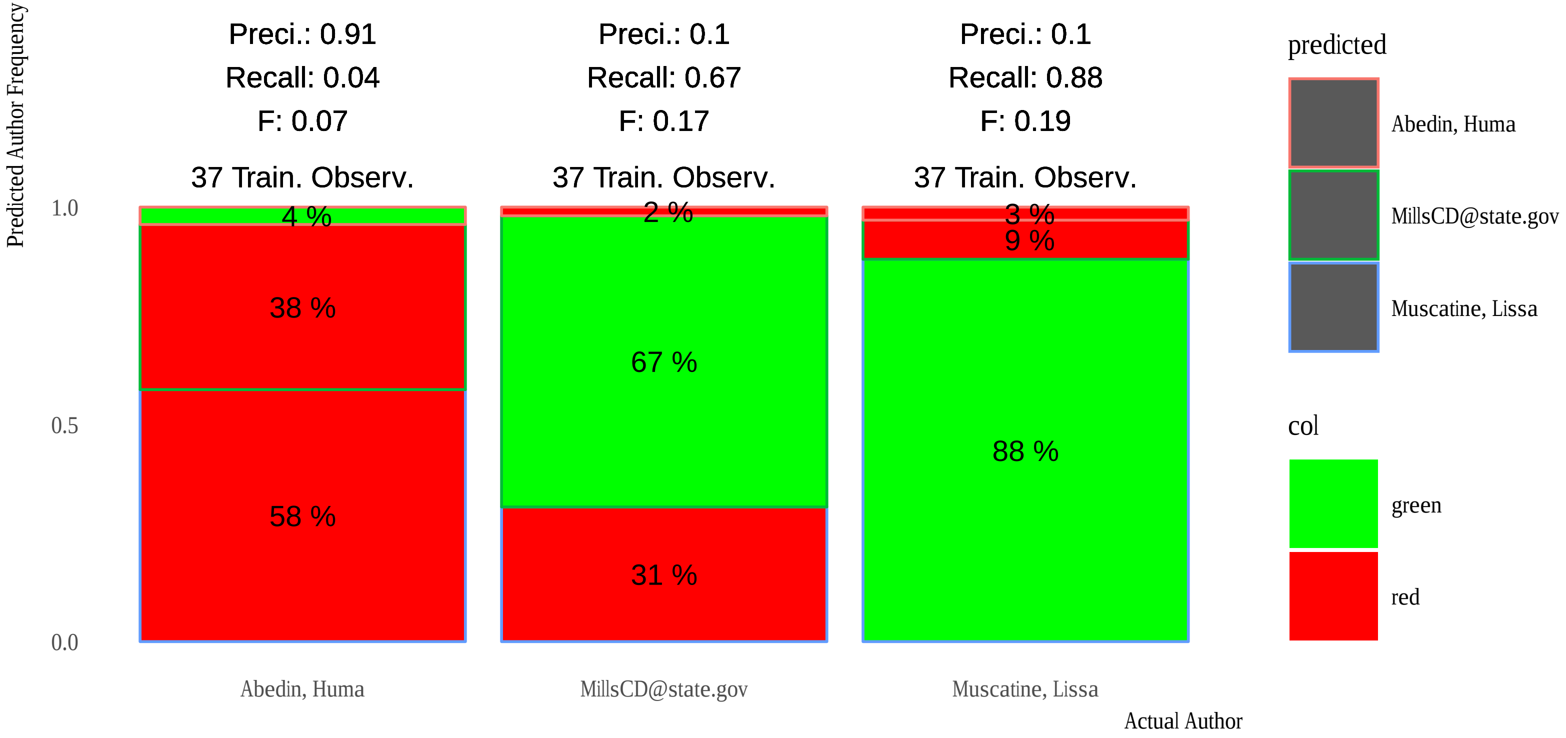
Author Sample 11

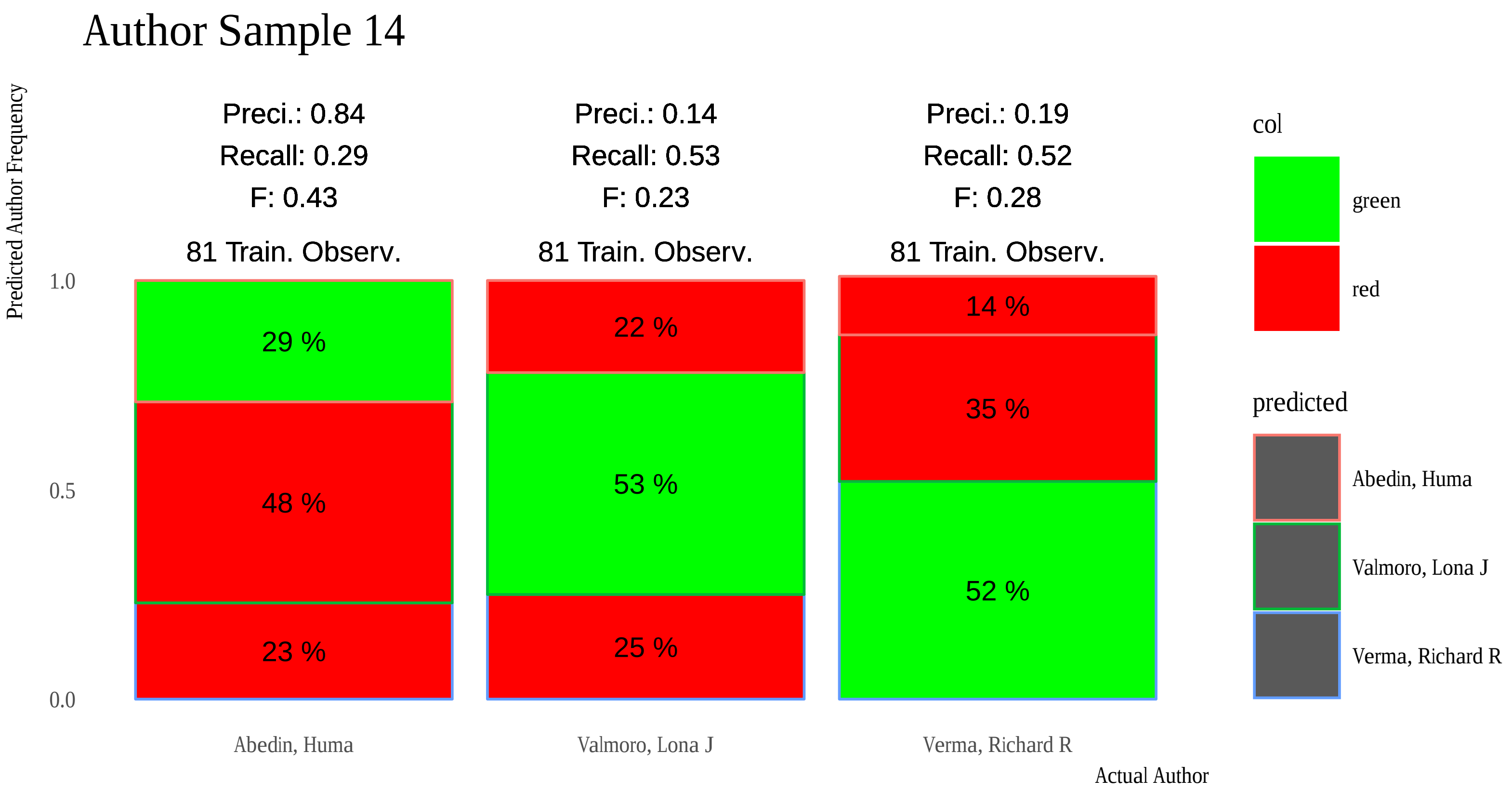


Author Sample 12

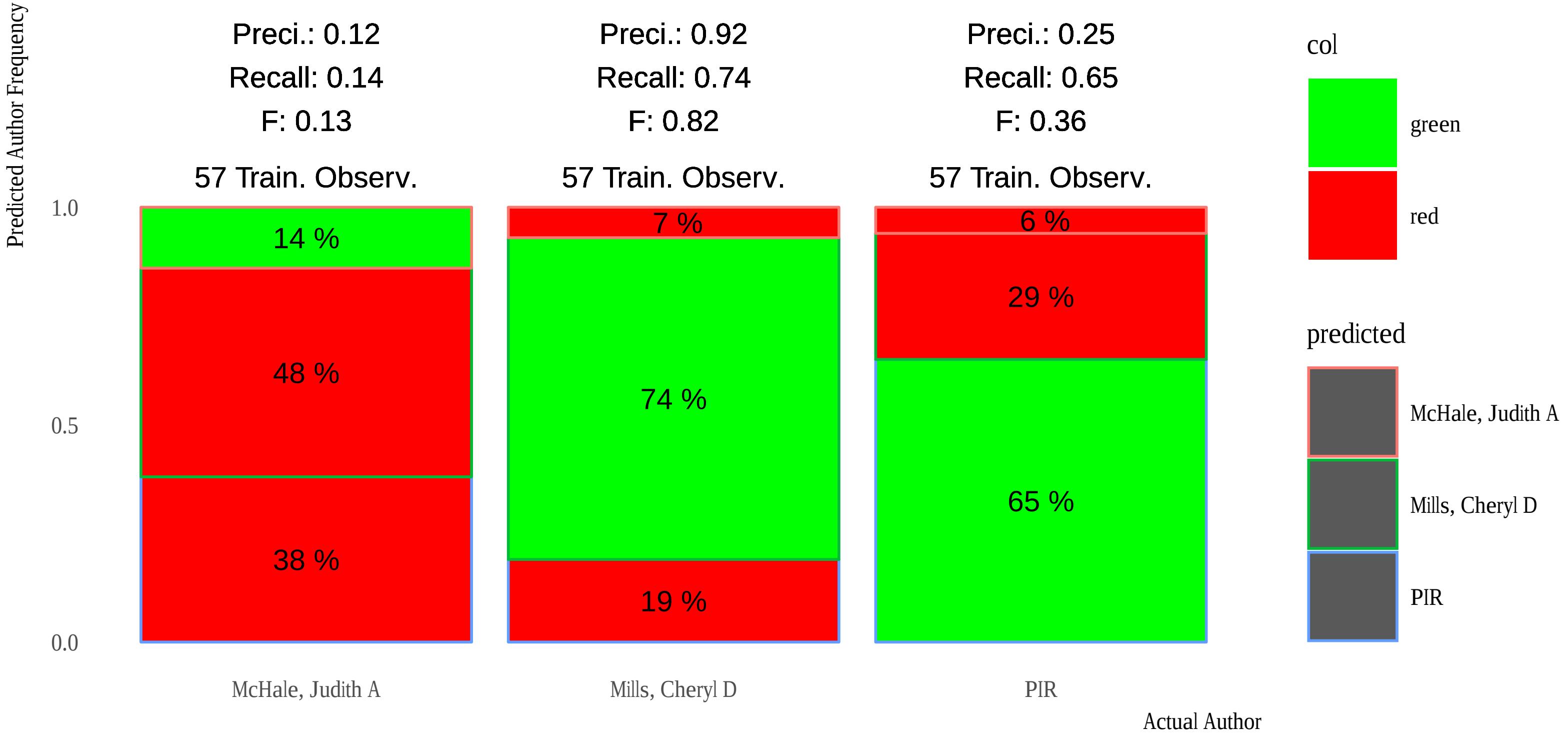


Author Sample 13

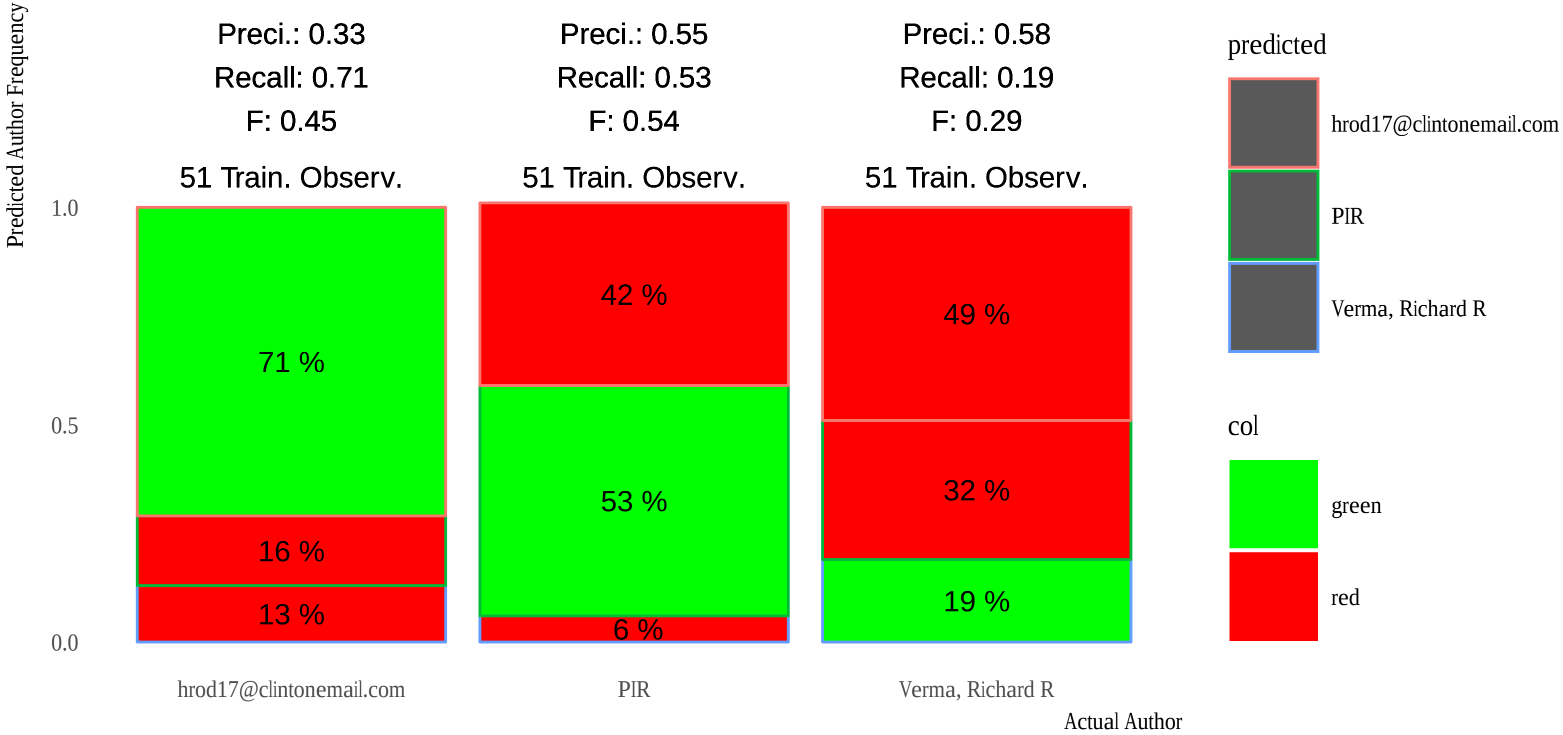



Author Sample 14
Predicted Author Frequency
1.0
0.5
0.0
Preci.: 0.84
Recall: 0.29
F: 0.43
81 Train. Observ.
29 %
48 %
23 %
Preci.: 0.14
Recall: 0.53
F: 0.23
81 Train. Observ.
22 %
53 %
25 %
Preci.: 0.19
Recall: 0.52
F: 0.28
81 Train. Observ.
14 %
35 %
52 %
Abedin, Huma
Valmoro, Lona J
Verma, Richard R
Actual Author
col
green
red
predicted
Abedin, Huma
Valmoro, Lona J
Verma, Richard R

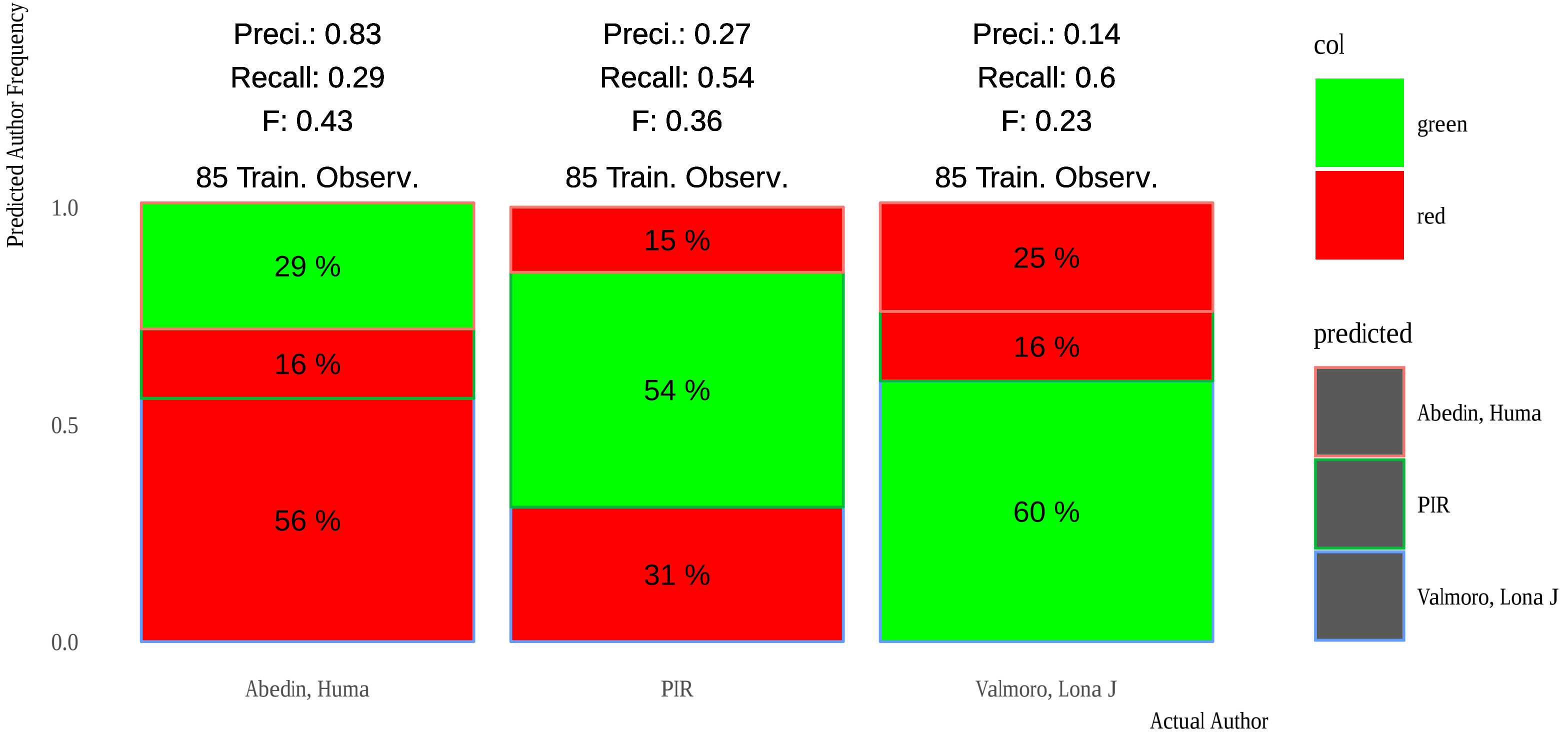
Author Sample 15



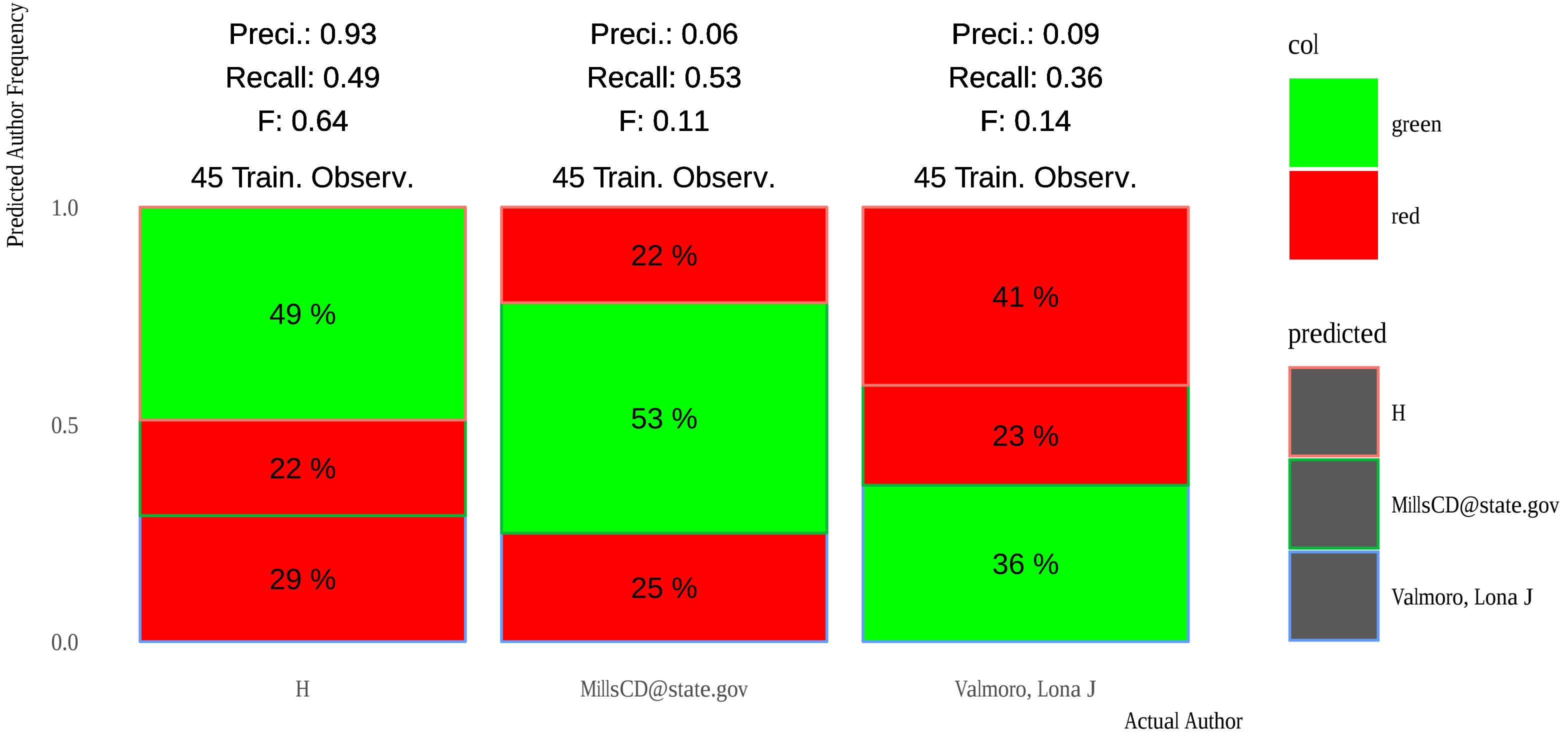
Author Sample 16

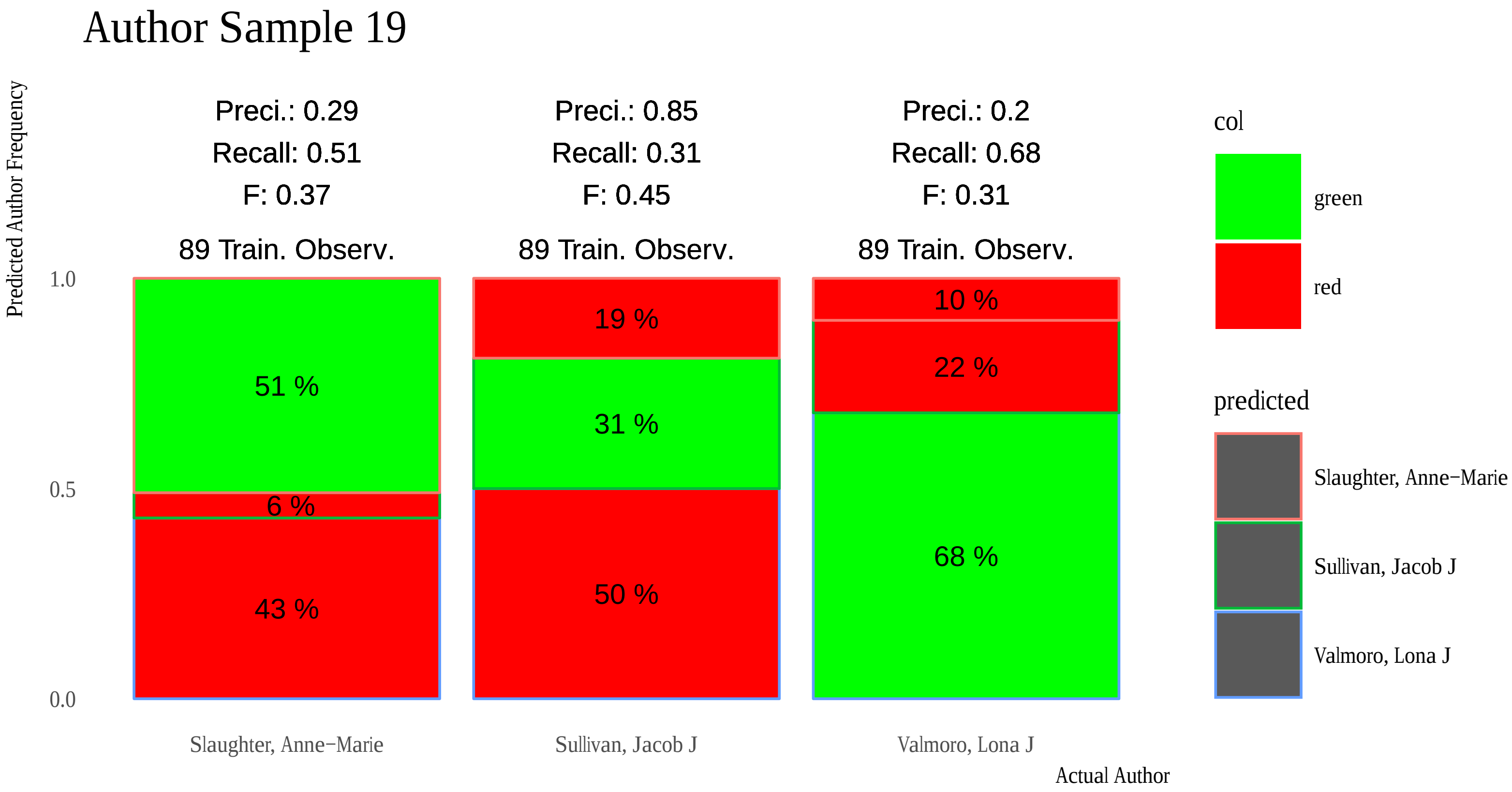


Author Sample 17

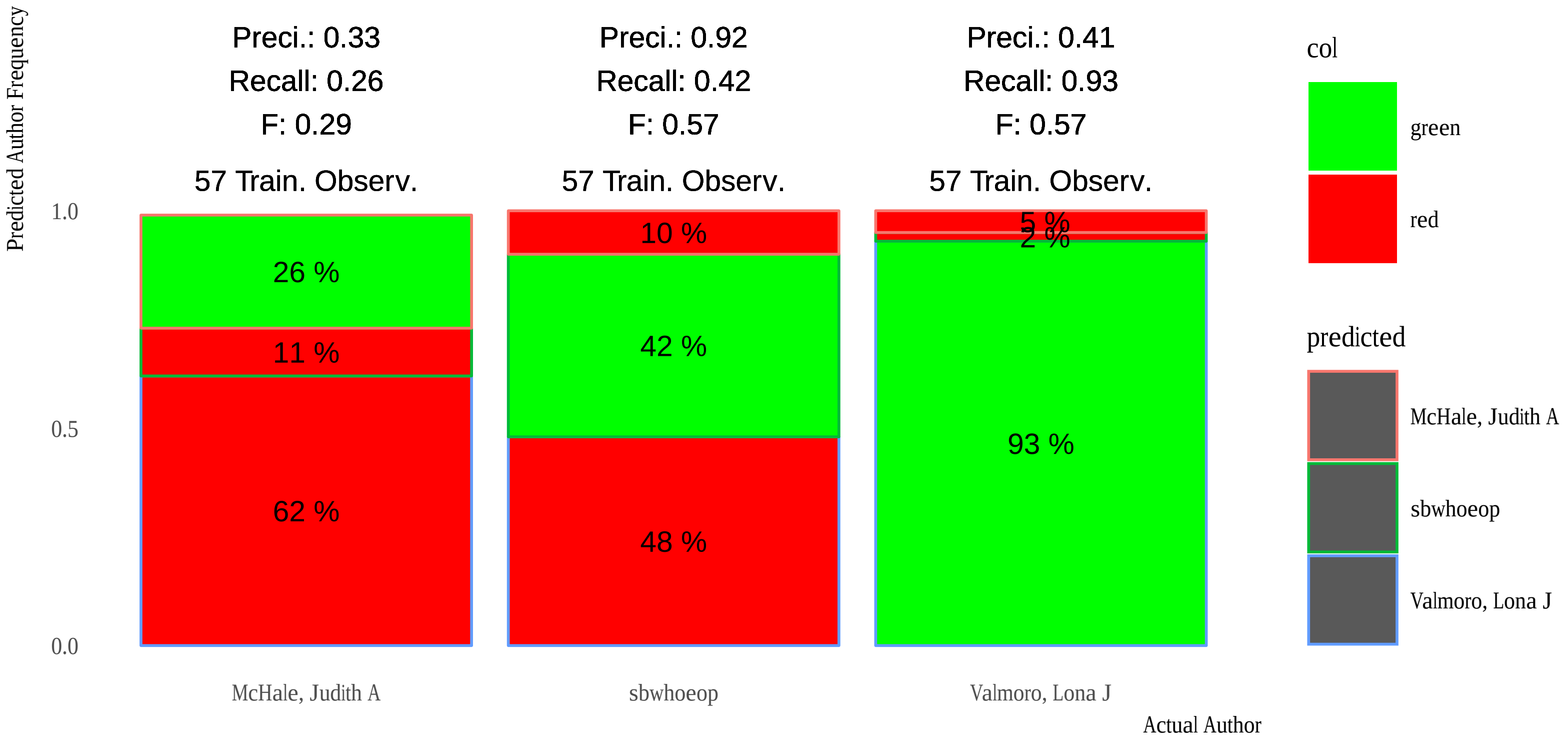


Author Sample 18

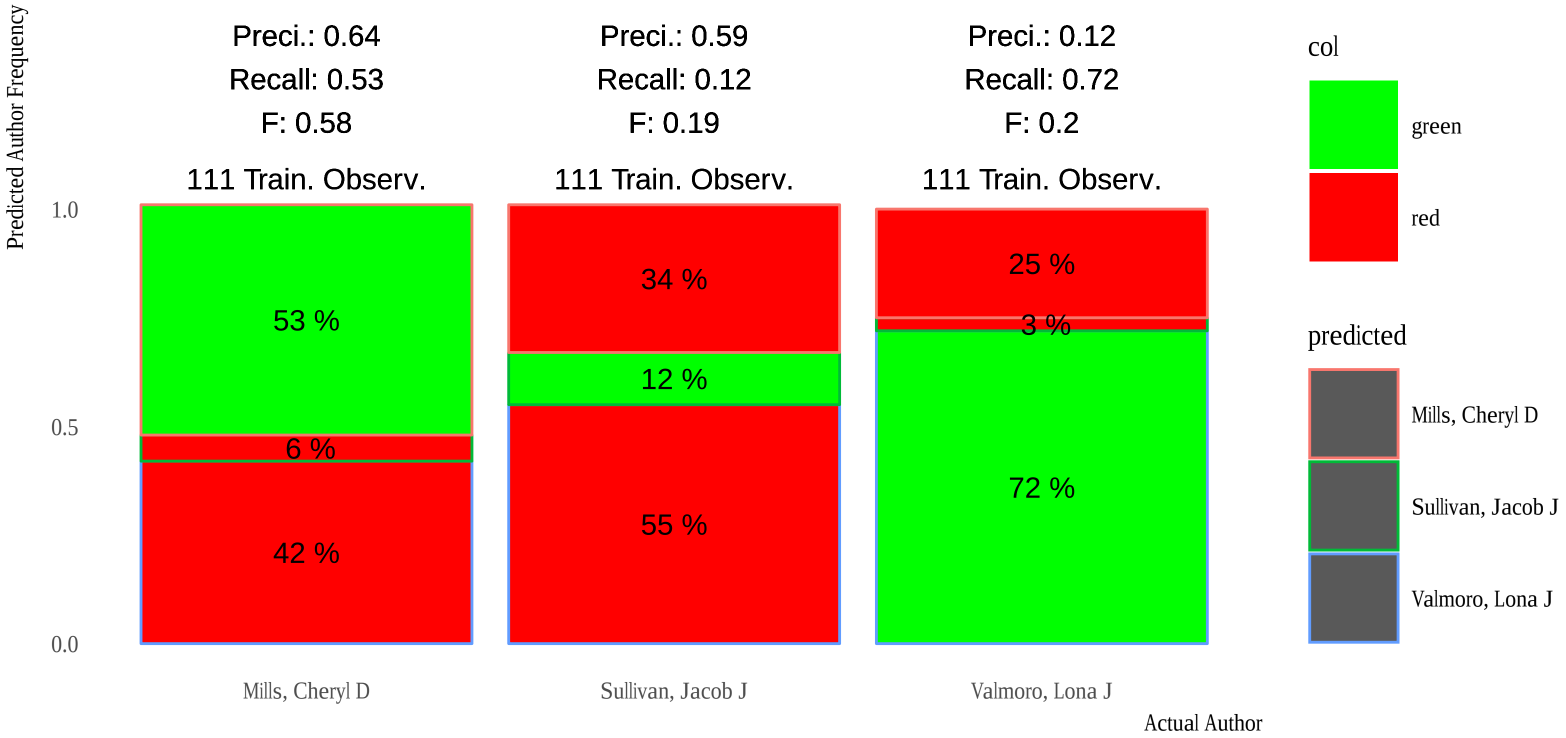



Author Sample 19
Predicted Author Frequency
1.0
0.5
0.0
Preci.: 0.29
Recall: 0.51
F: 0.37
89 Train. Observ.
51 %
6 %
43 %
Preci.: 0.85
Recall: 0.31
F: 0.45
89 Train. Observ.
19 %
31 %
50 %
Preci.: 0.2
Recall: 0.68
F: 0.31
89 Train. Observ.
10 %
22 %
68 %
Slaughter, Anne−Marie
Sullivan, Jacob J
Valmoro, Lona J
Actual Author
col
green
red
predicted
Slaughter, Anne−Marie
Sullivan, Jacob J
Valmoro, Lona J

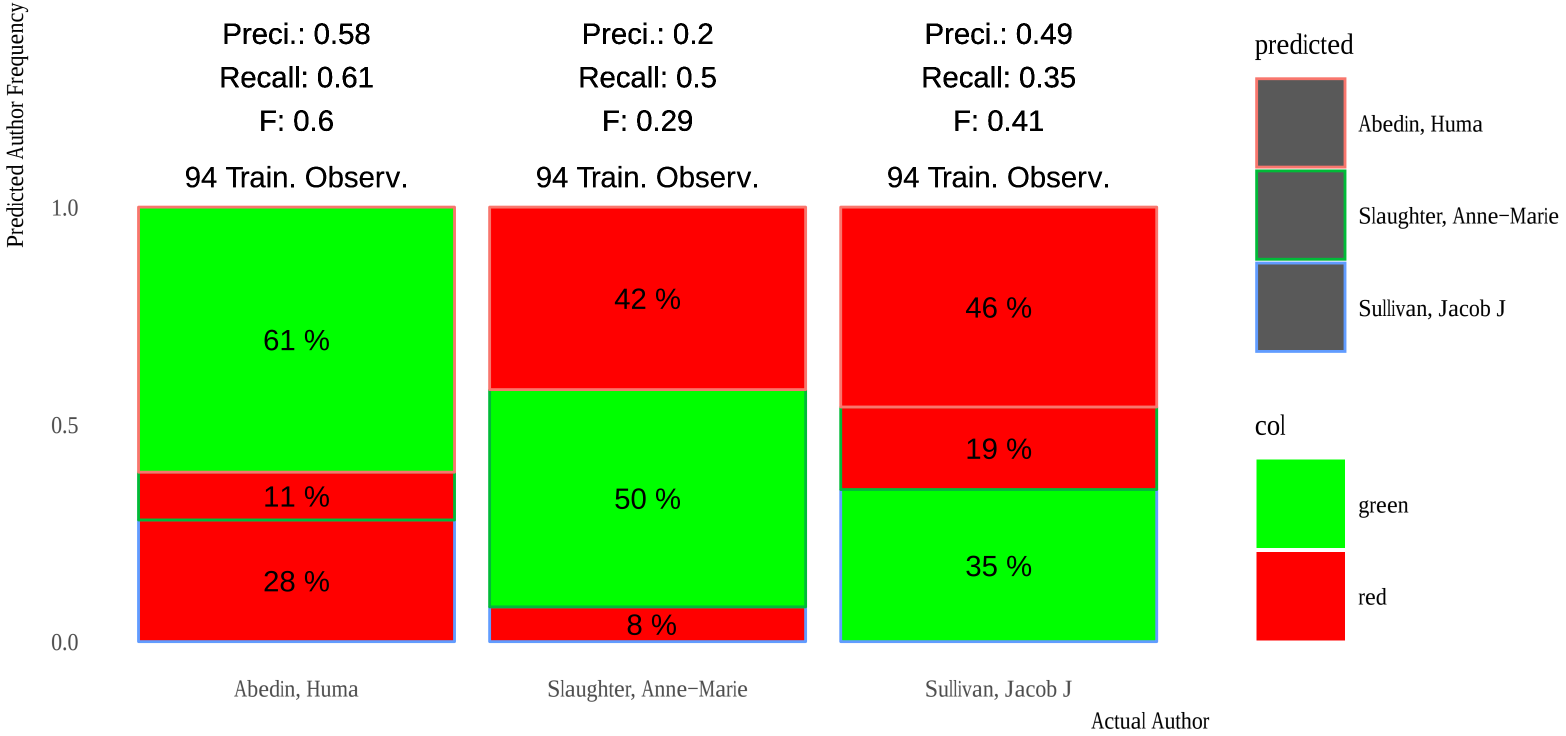
Author Sample 20



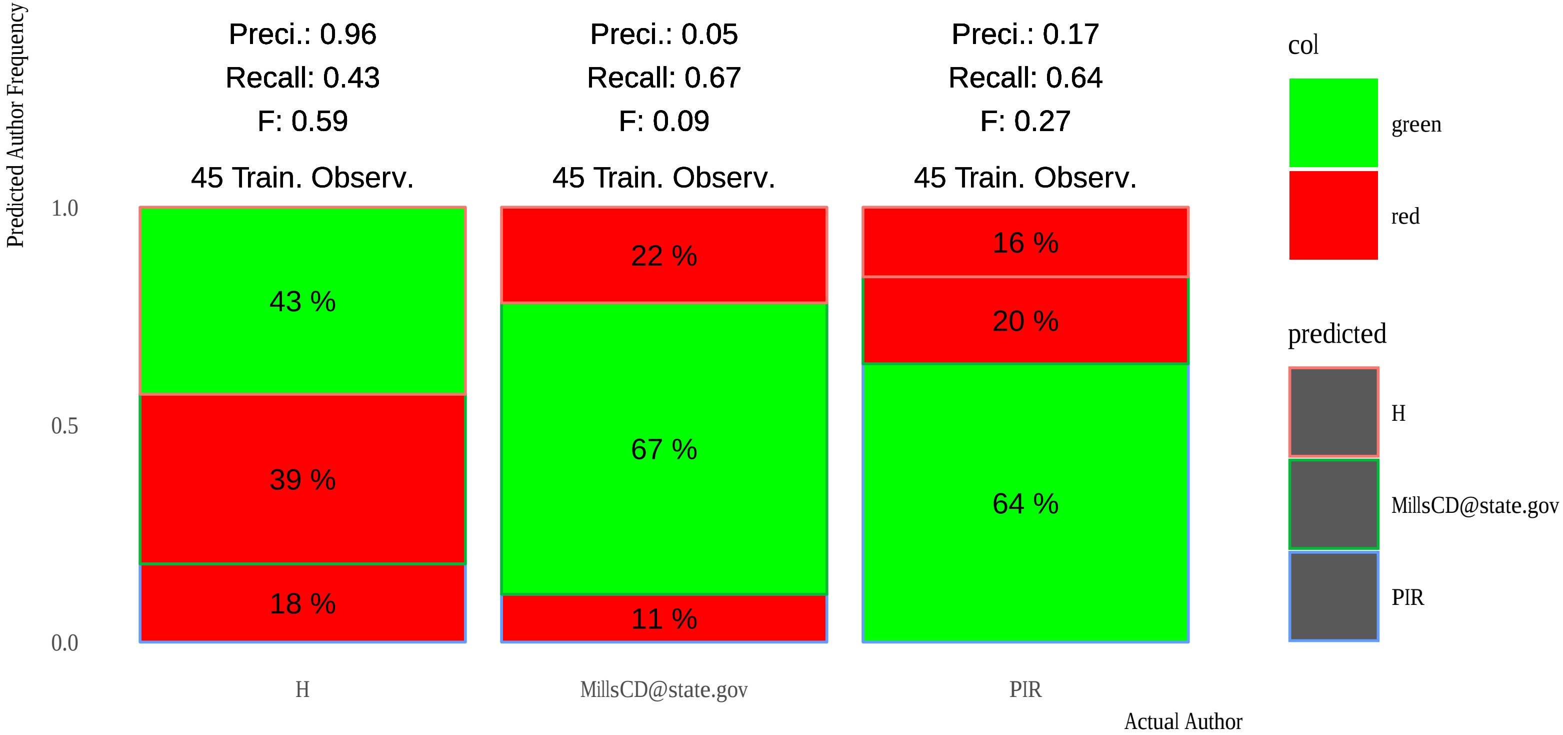
Author Sample 21



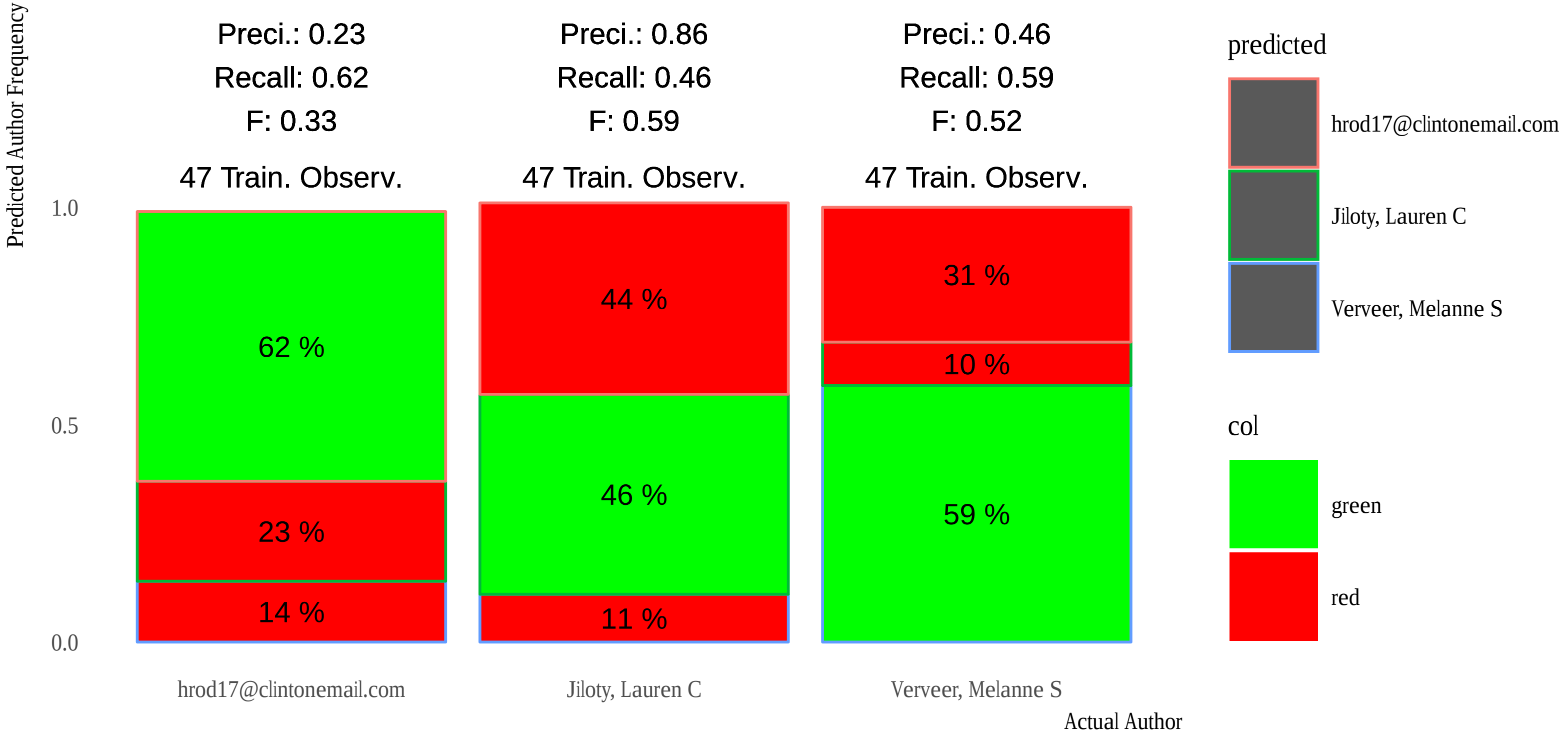
Author Sample 22



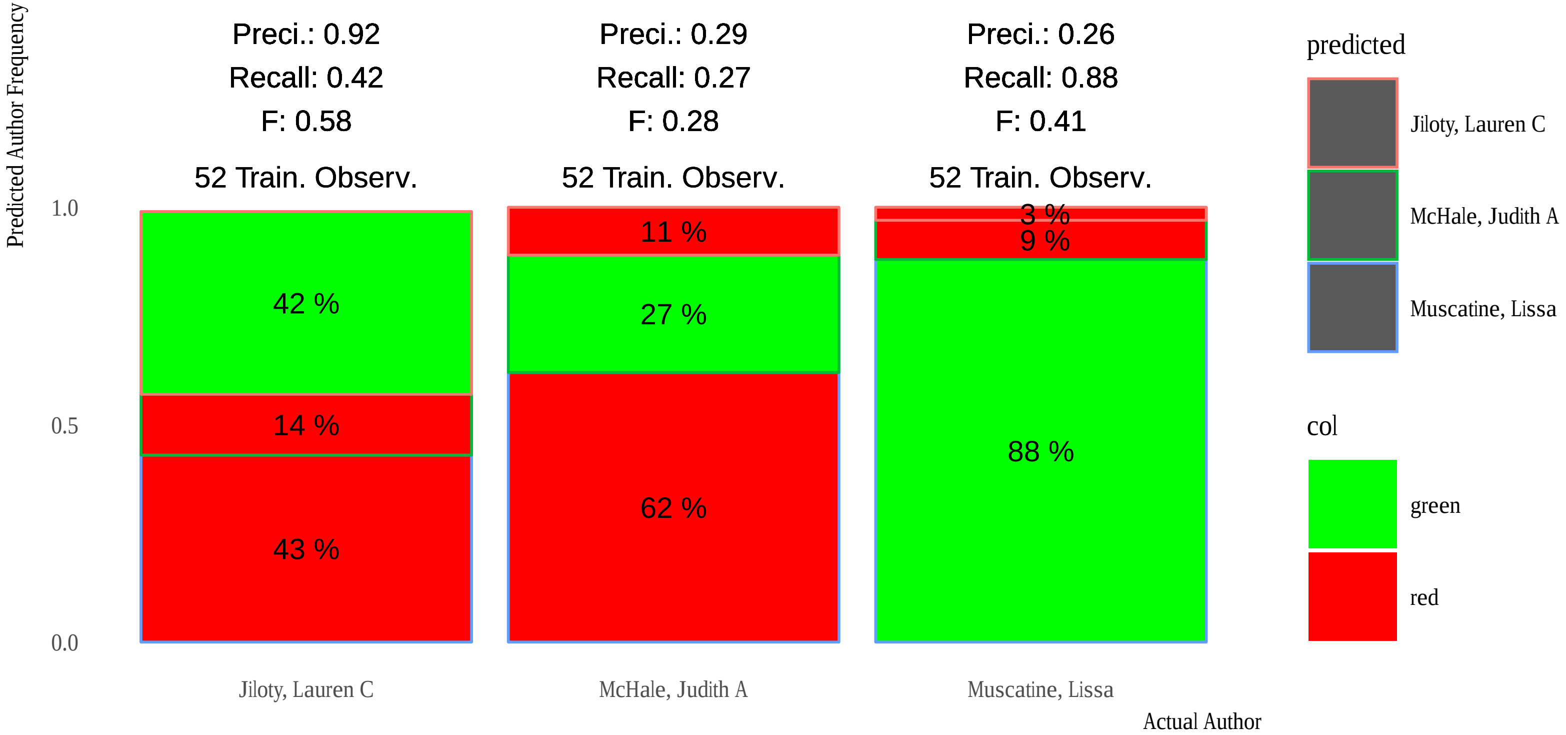
Author Sample 23



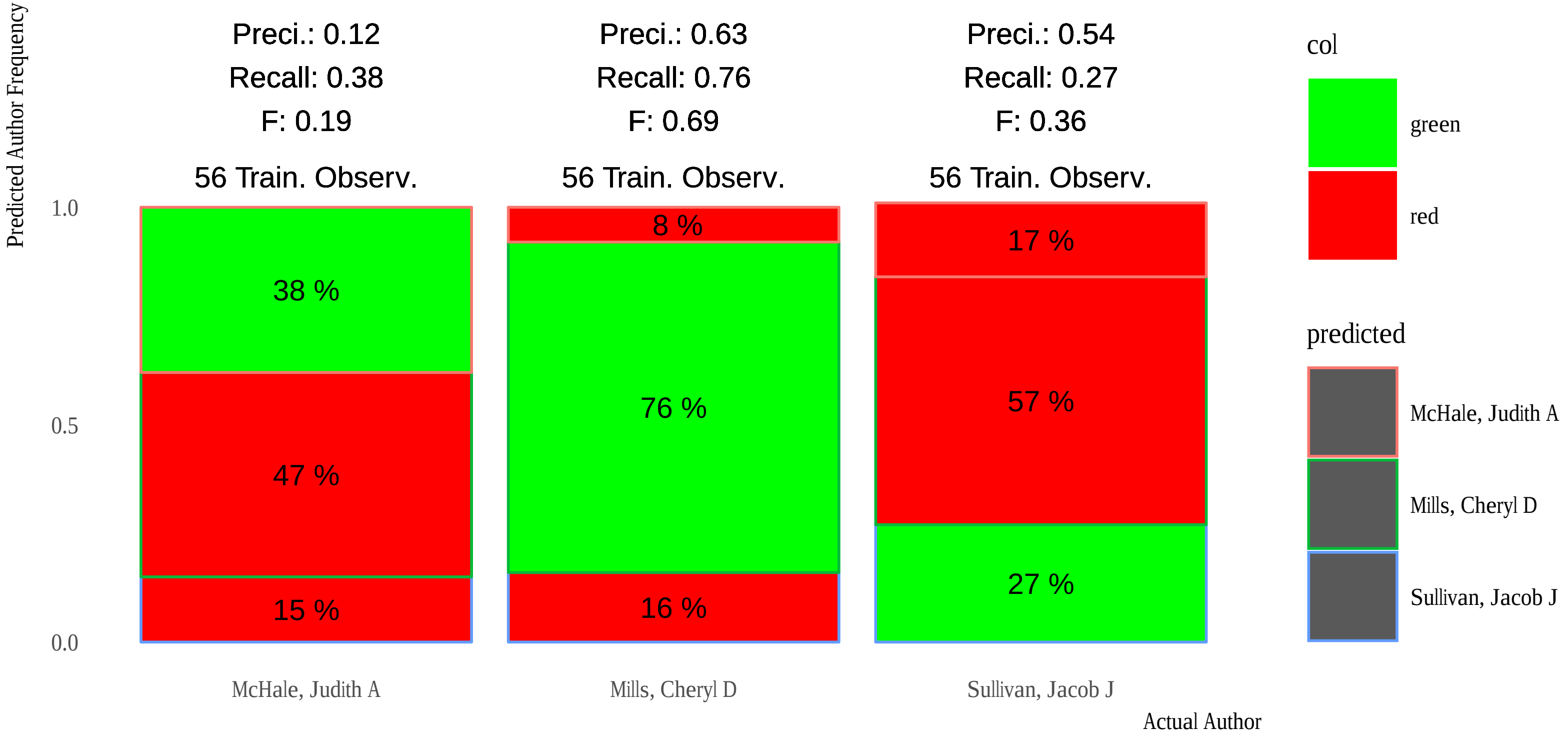
Author Sample 24



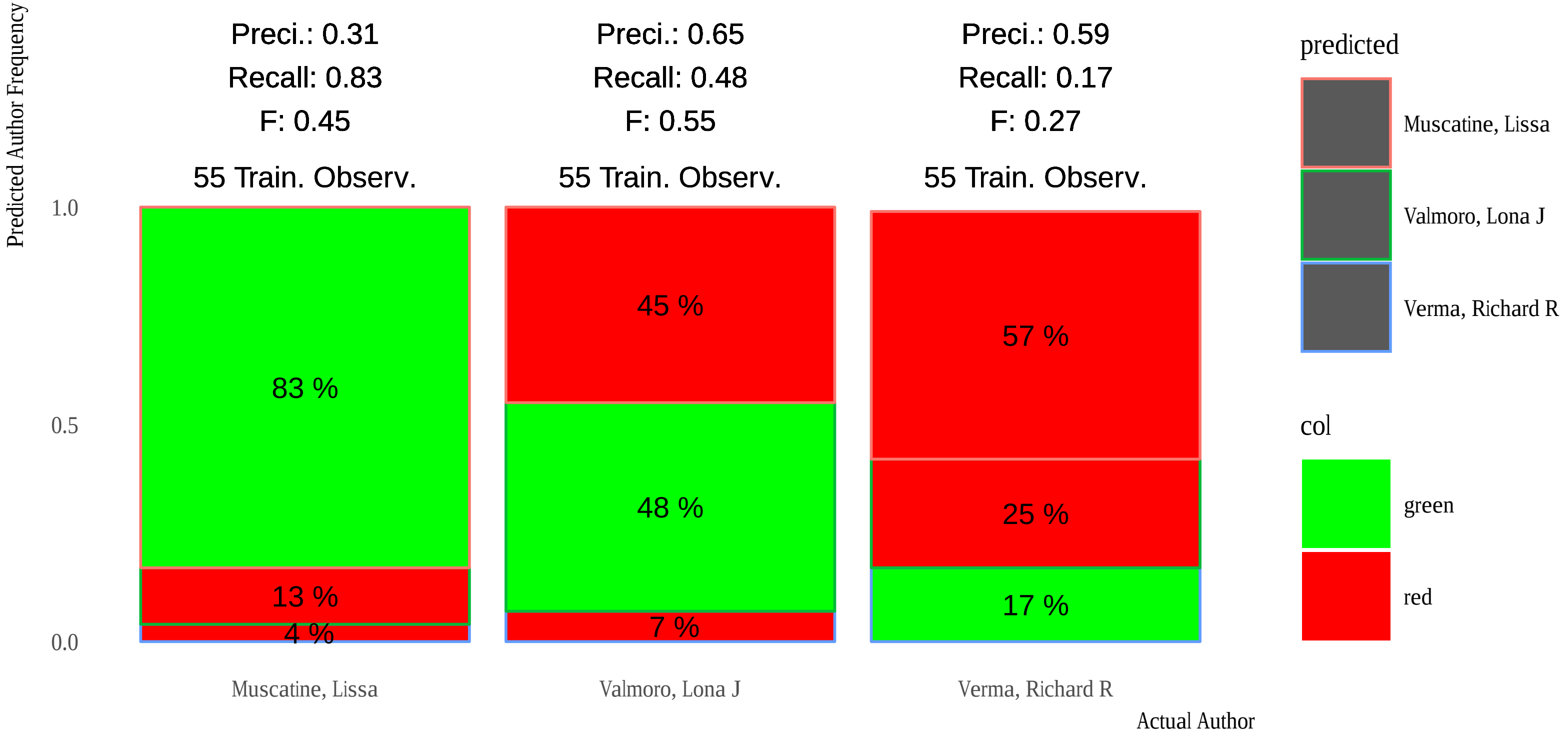
Author Sample 25



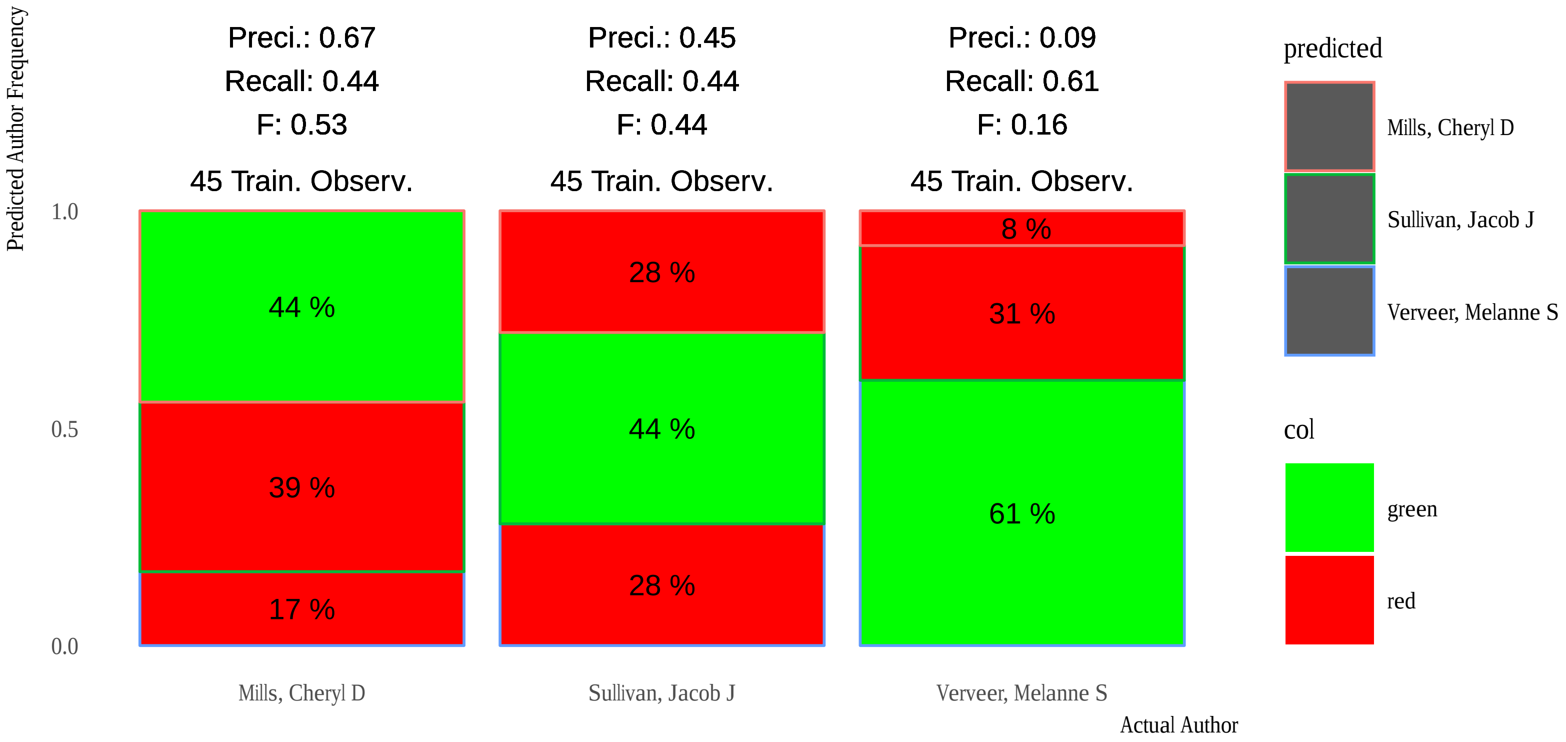
Author Sample 26



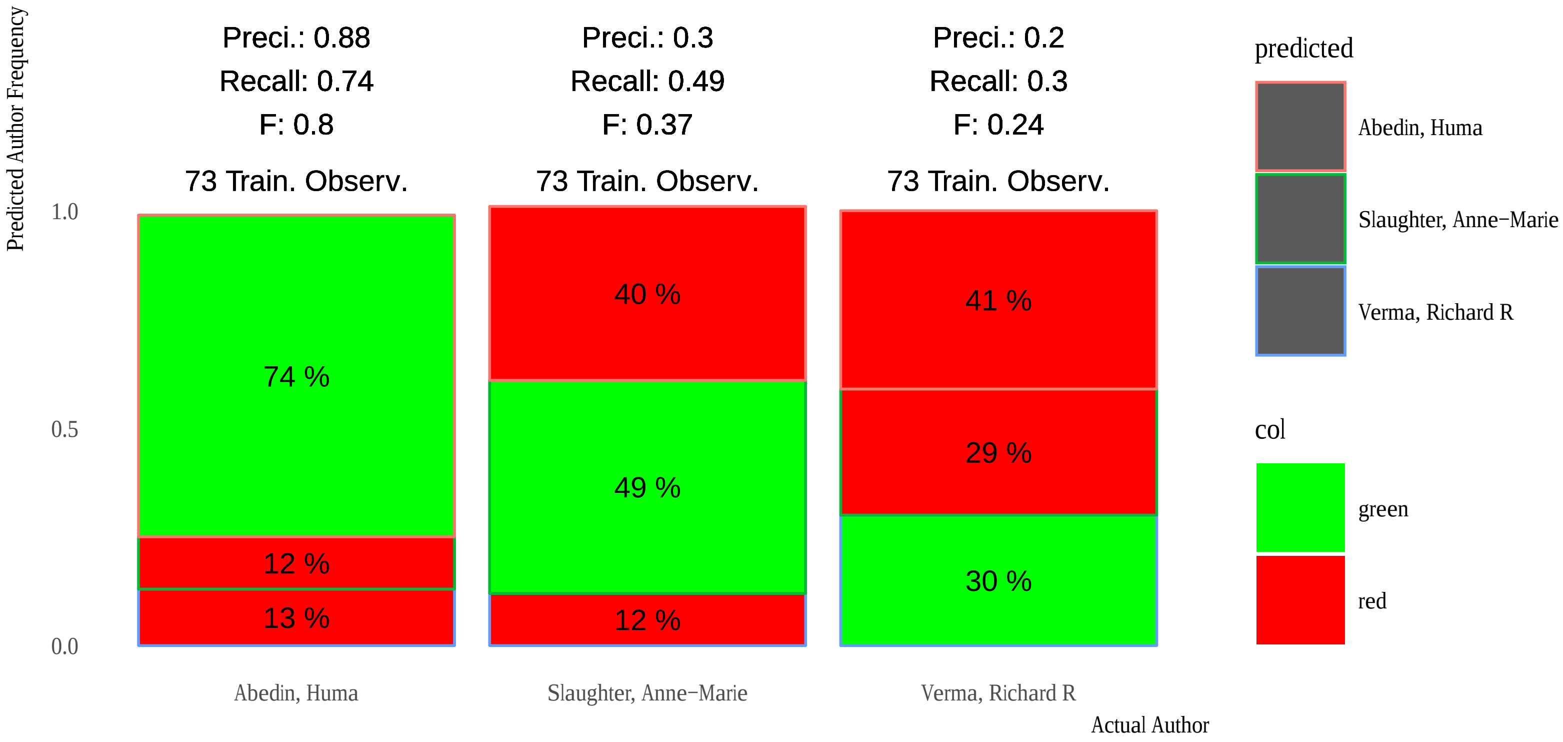
Author Sample 27



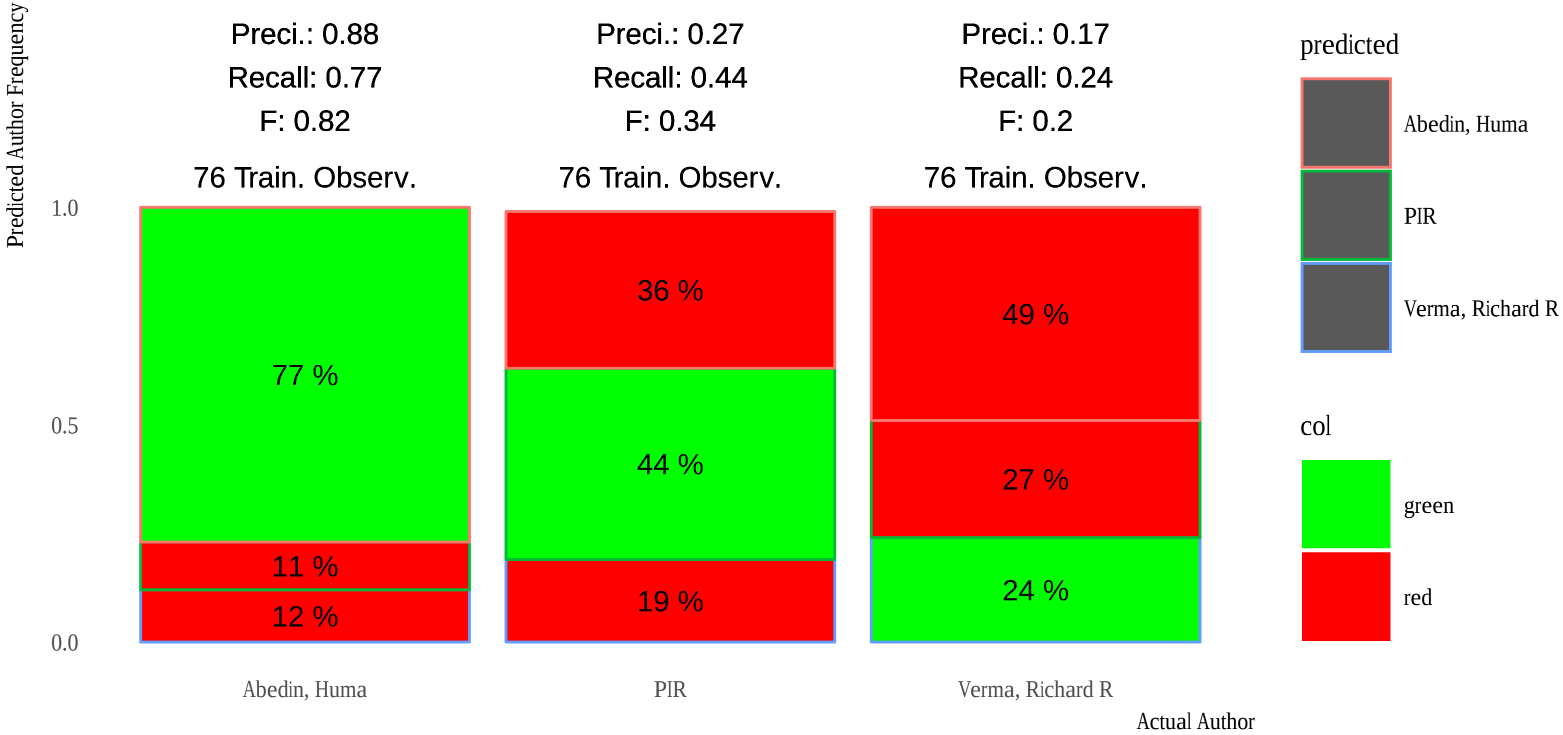
Author Sample 28



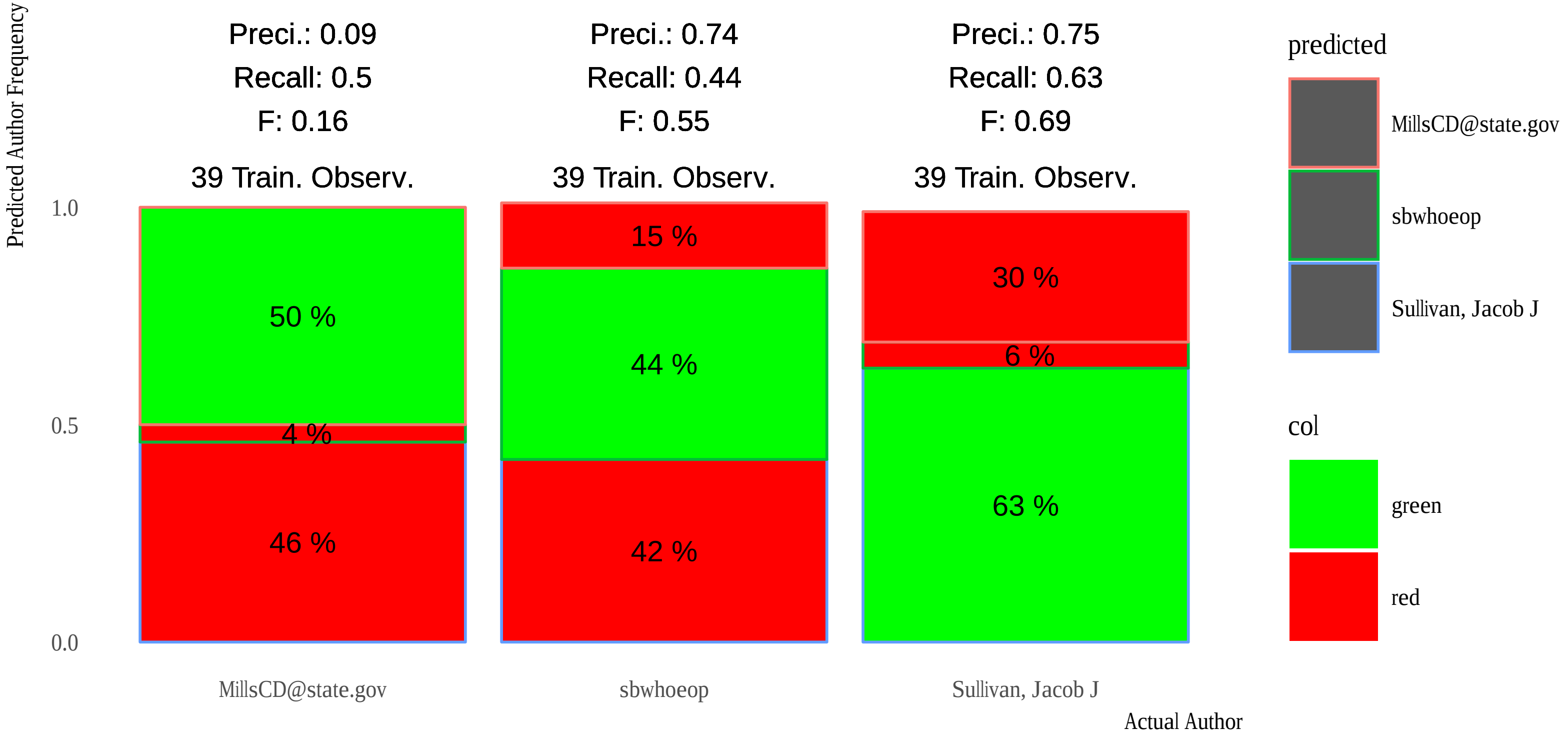
Author Sample 29



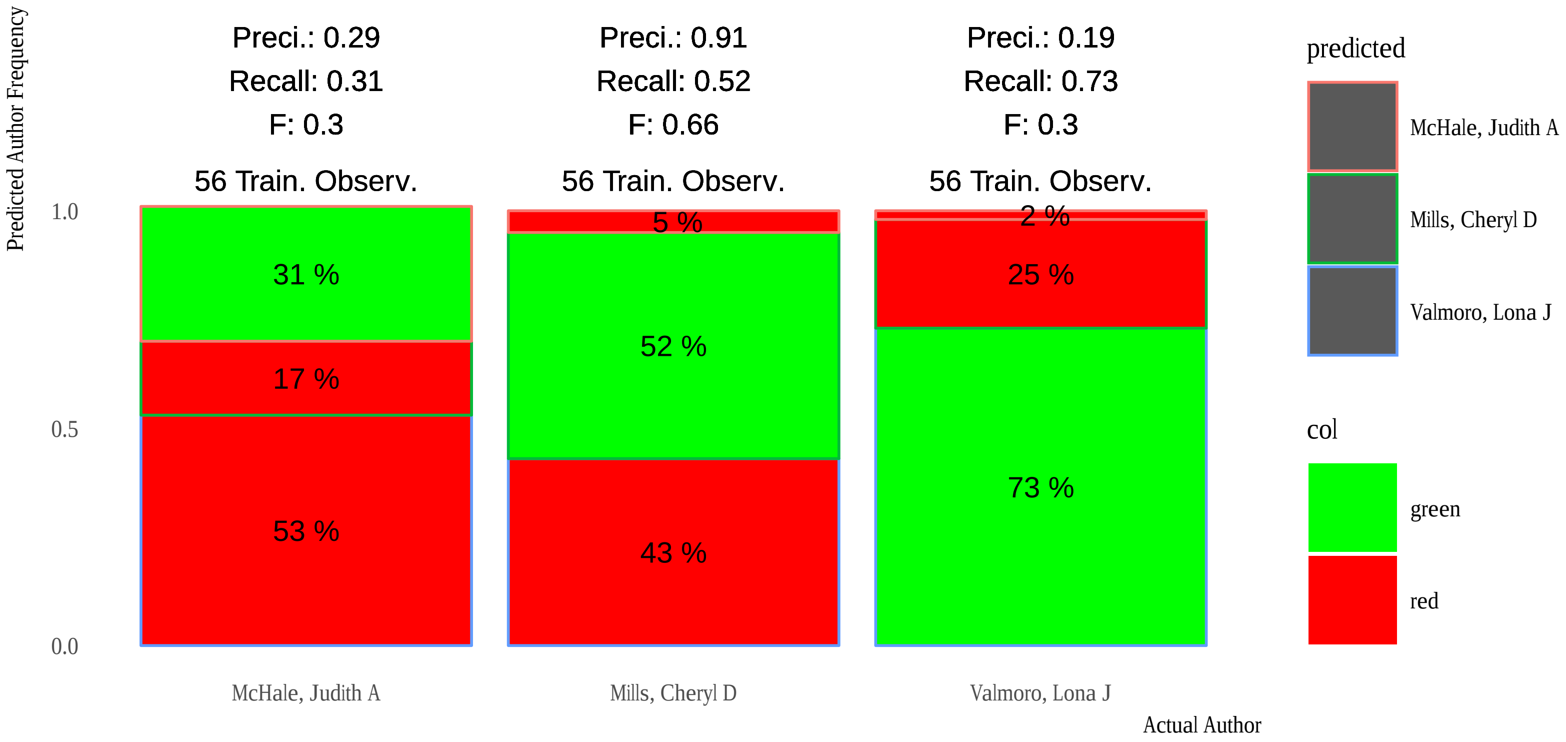
Author Sample 30

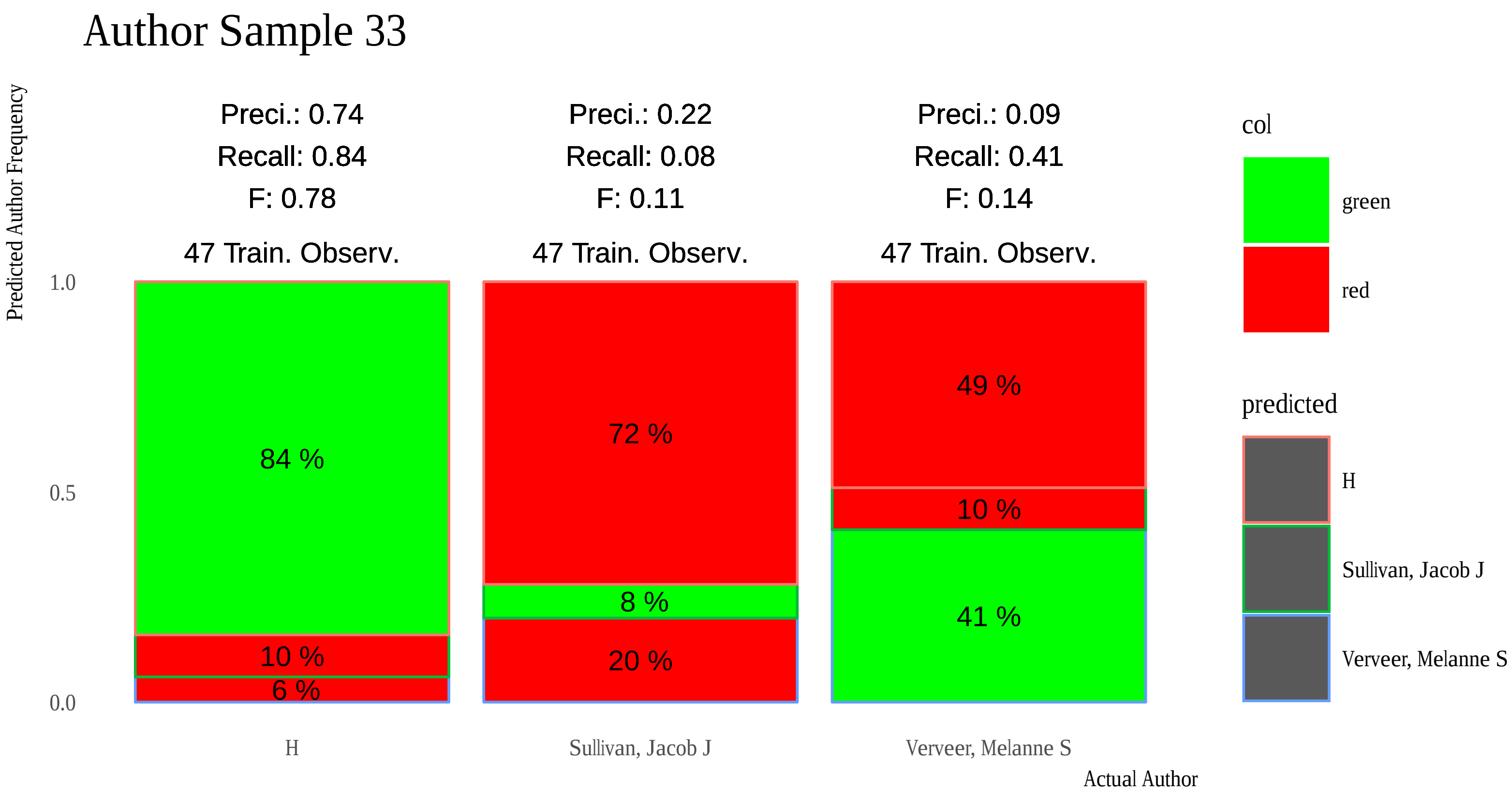


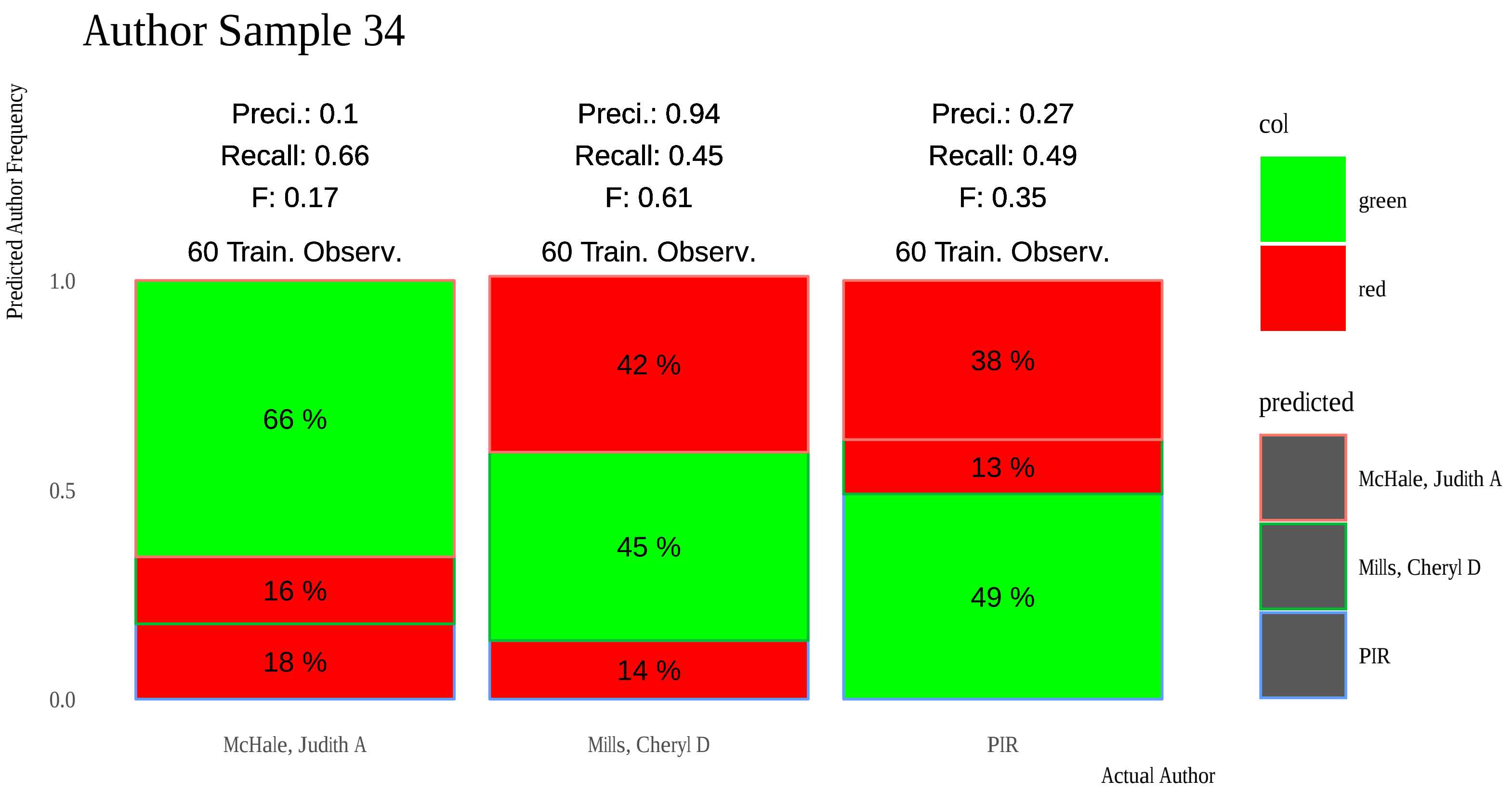
Author Sample 31



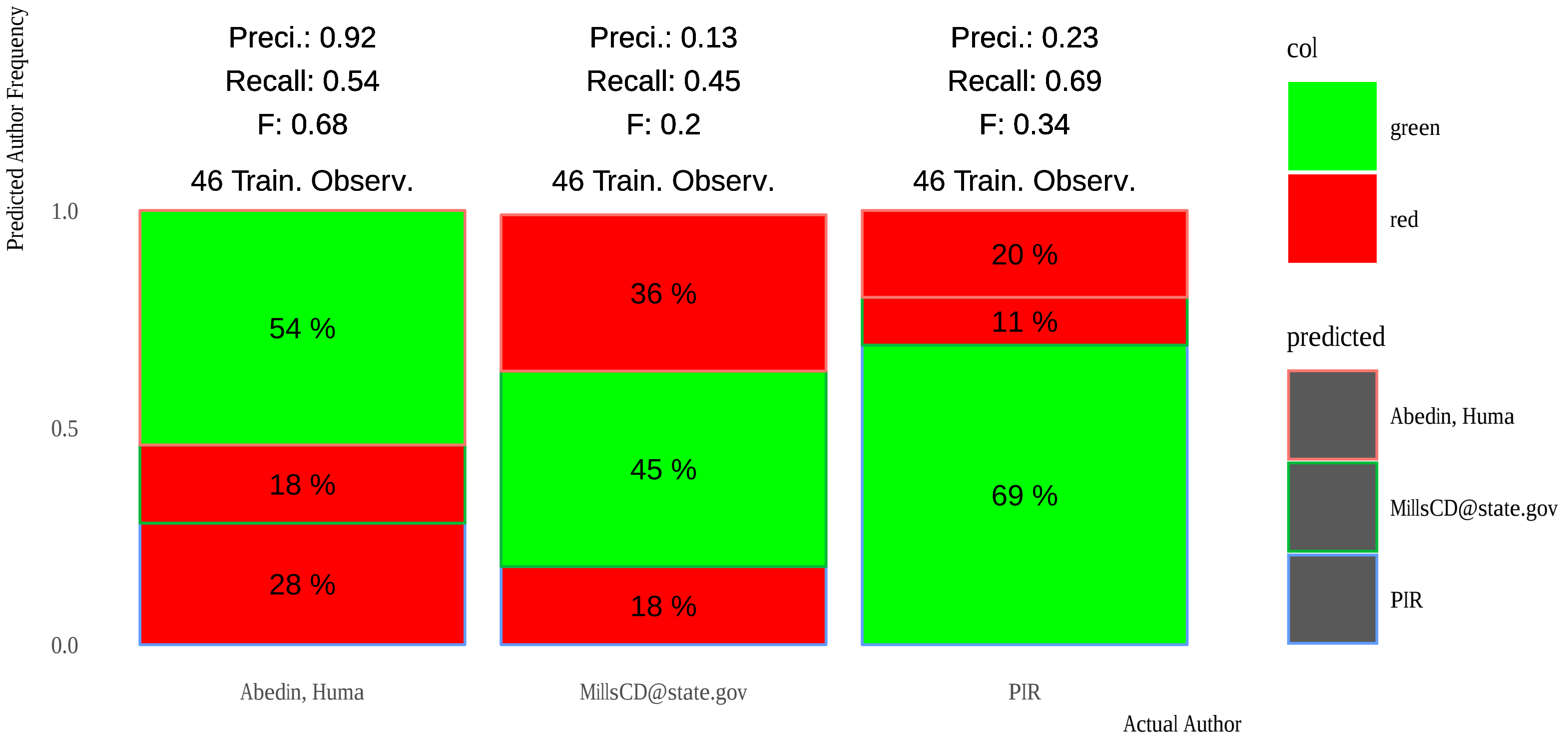
Author Sample 32



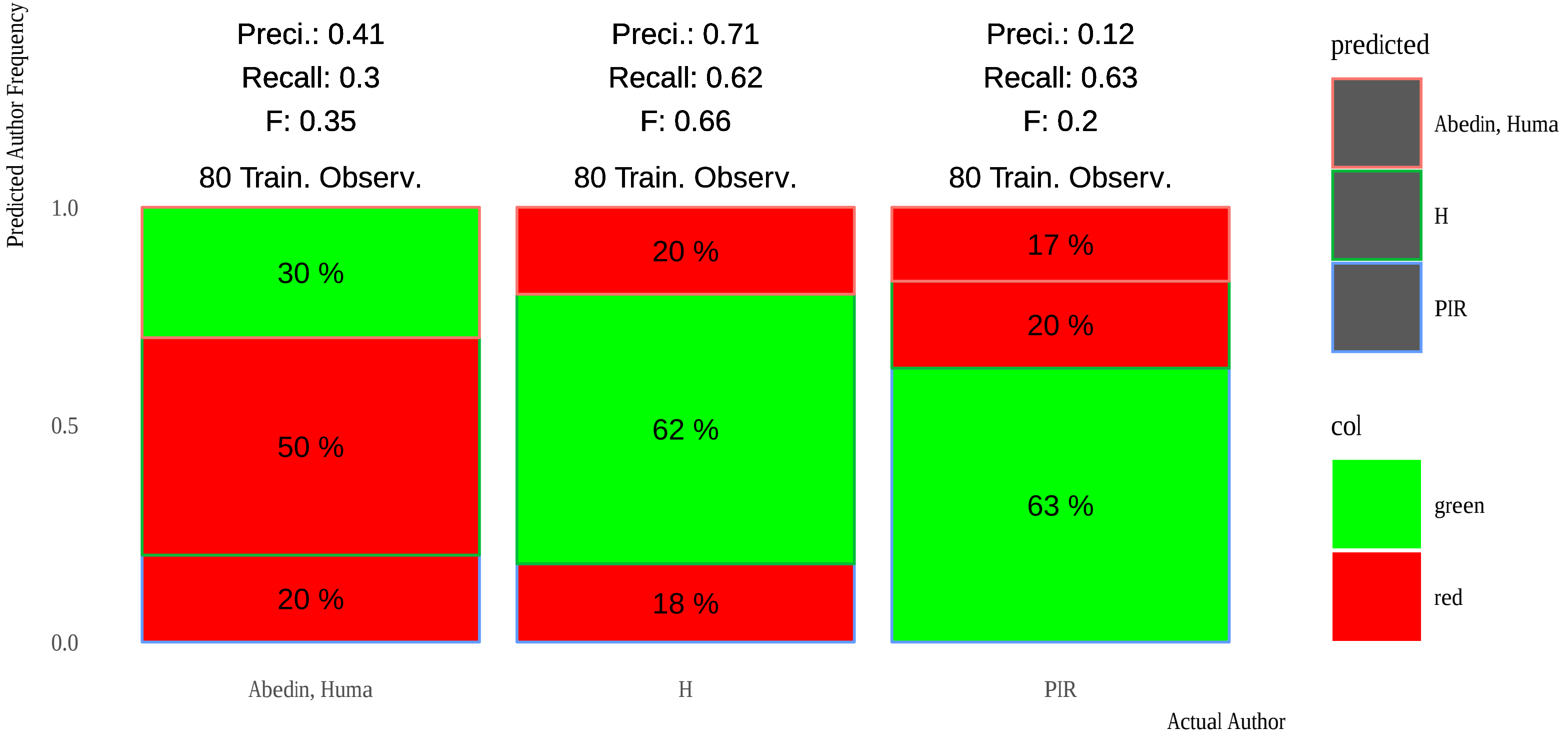

Author Sample 33
Predicted Author Frequency
1.0
0.5
0.0
Preci.: 0.74
Recall: 0.84
F: 0.78
47 Train. Observ.
84 %
10 %
6 %
Preci.: 0.22
Recall: 0.08
F: 0.11
47 Train. Observ.
72 %
8 %
20 %
Preci.: 0.09
Recall: 0.41
F: 0.14
47 Train. Observ.
49 %
10 %
41 %
H
Sullivan, Jacob J
Verveer, Melanne S
Actual Author
col
green
red
predicted
H
Sullivan, Jacob J
Verveer, Melanne S


Author Sample 34
Predicted Author Frequency
1.0
0.5
0.0
Preci.: 0.1
Recall: 0.66
F: 0.17
60 Train. Observ.
Preci.: 0.94
Recall: 0.45
F: 0.61
60 Train. Observ.
Preci.: 0.27
Recall: 0.49
F: 0.35
60 Train. Observ.
66 %
16 %
18 %
42 %
45 %
14 %
38 %
13 %
49 %
McHale, Judith A
Mills, Cheryl D
PIR
Actual Author
col
green
red
predicted
McHale, Judith A
Mills, Cheryl D
PIR

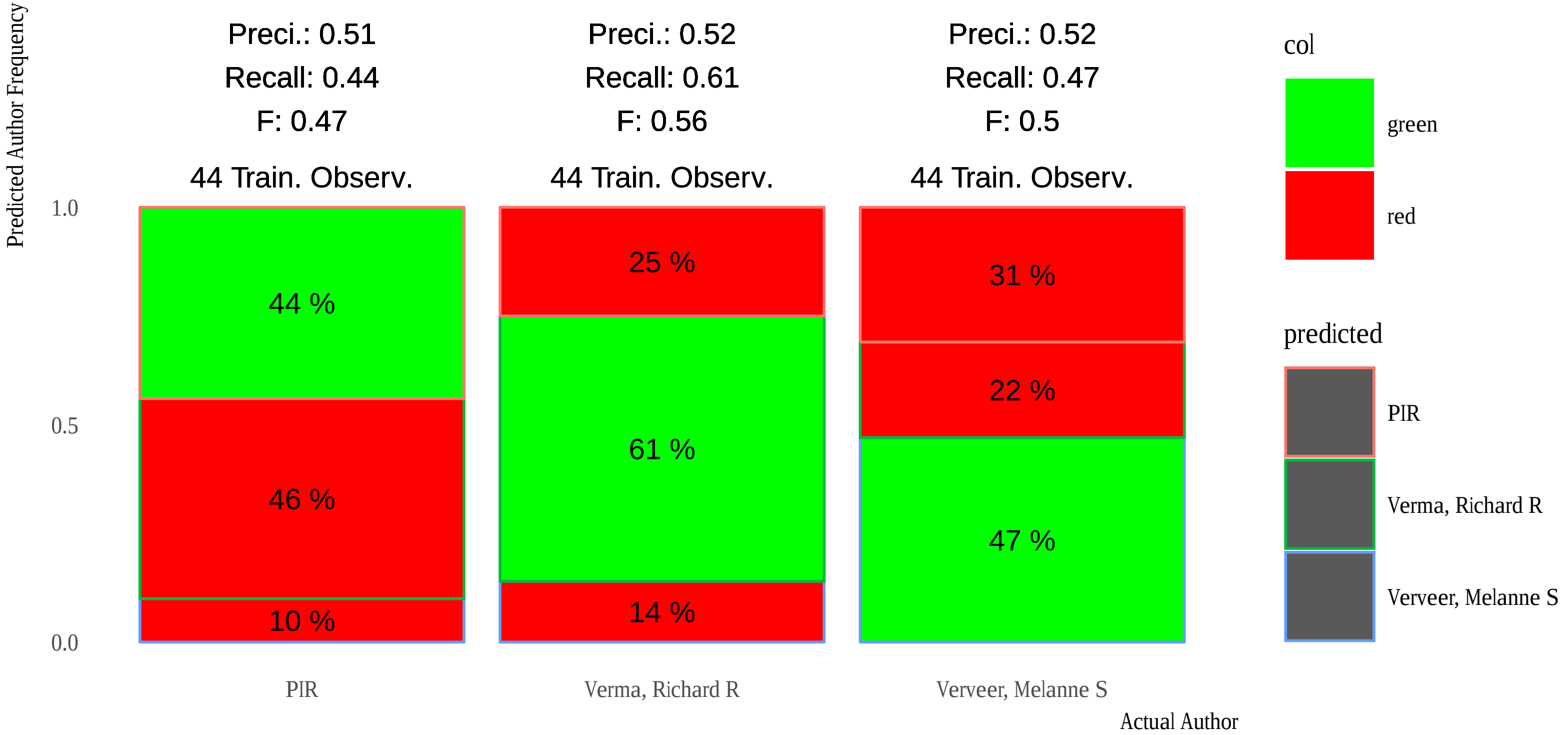
Author Sample 35



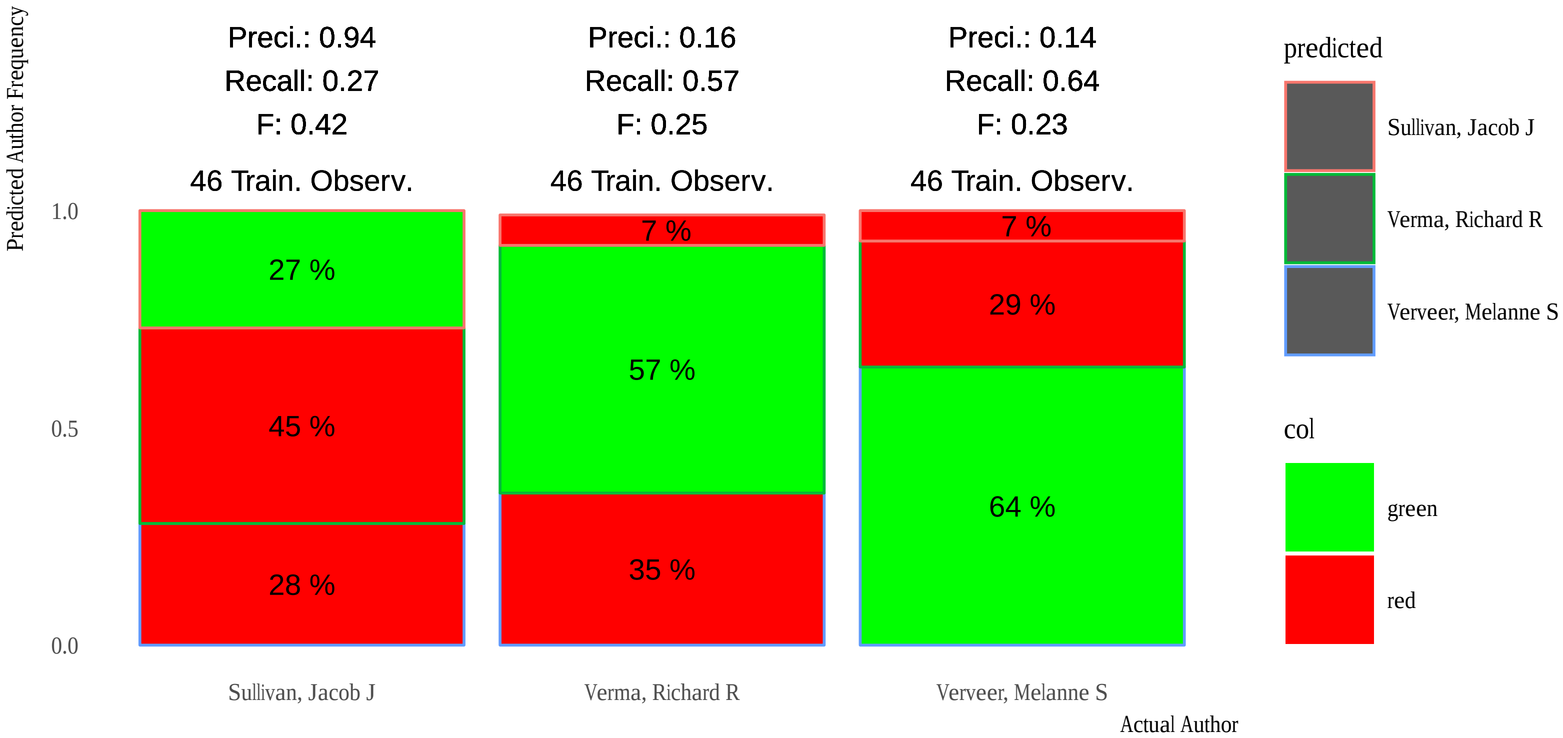
Author Sample 36



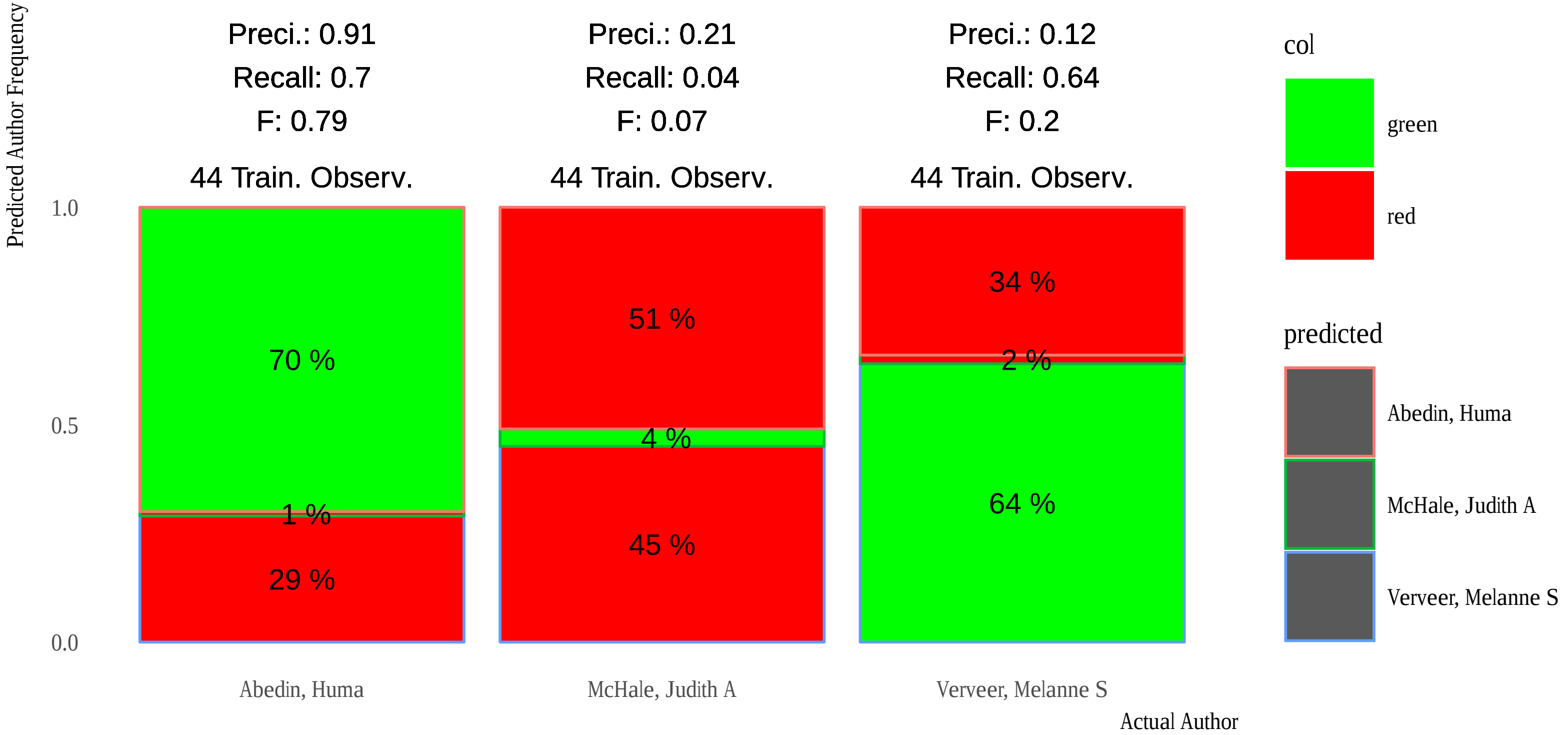
Author Sample 37

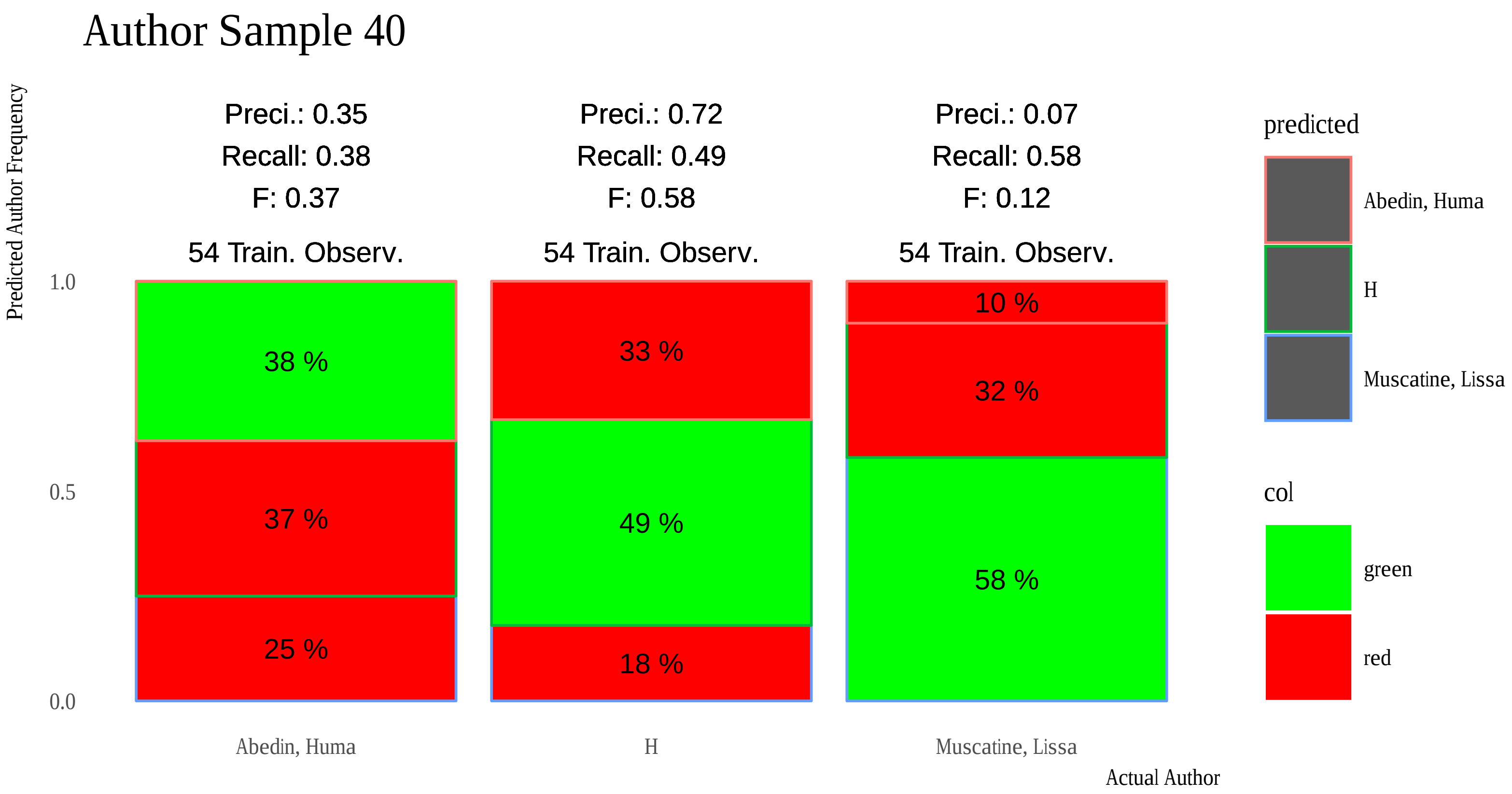


Author Sample 38



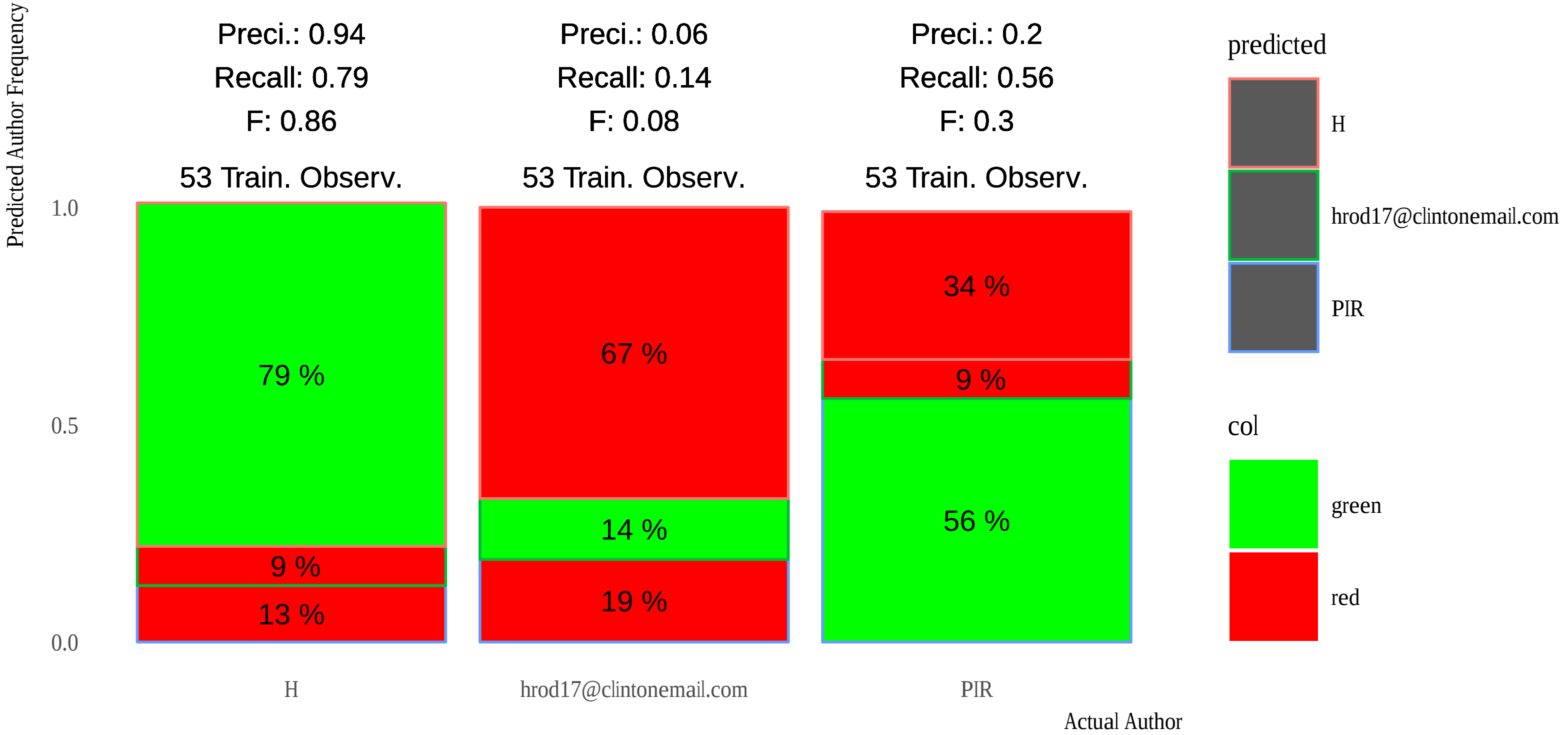
Author Sample 39



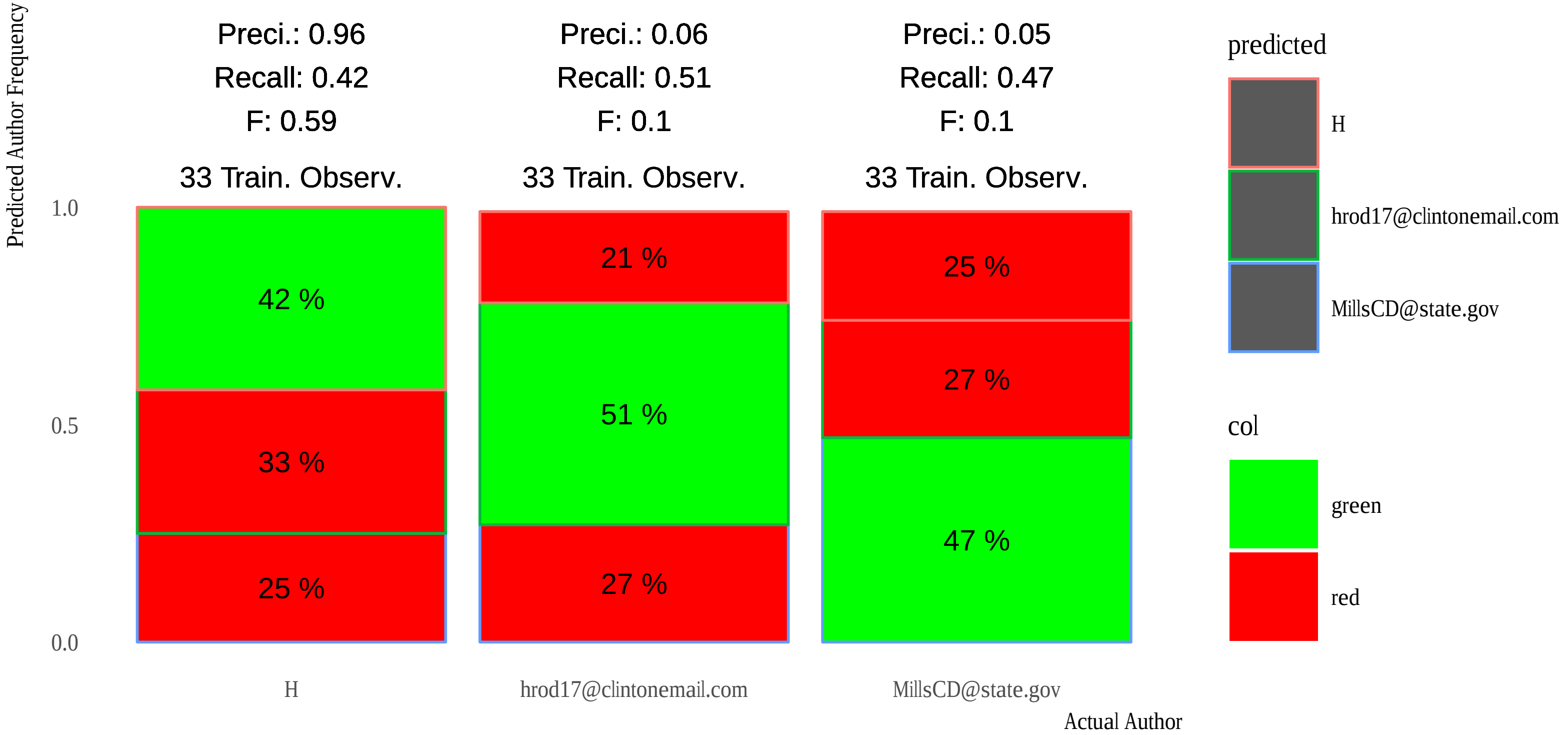


Author Sample 40

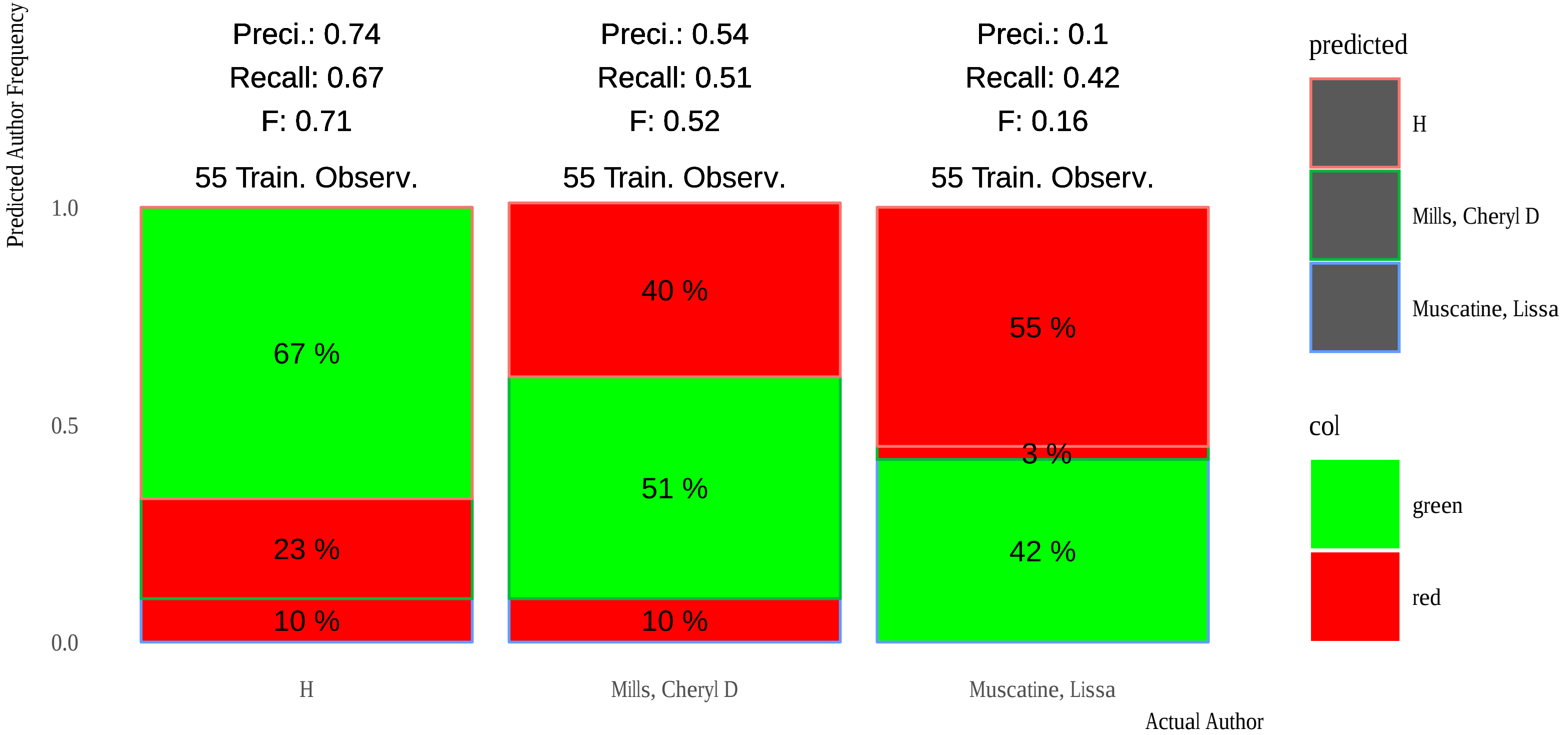
Author Sample 41



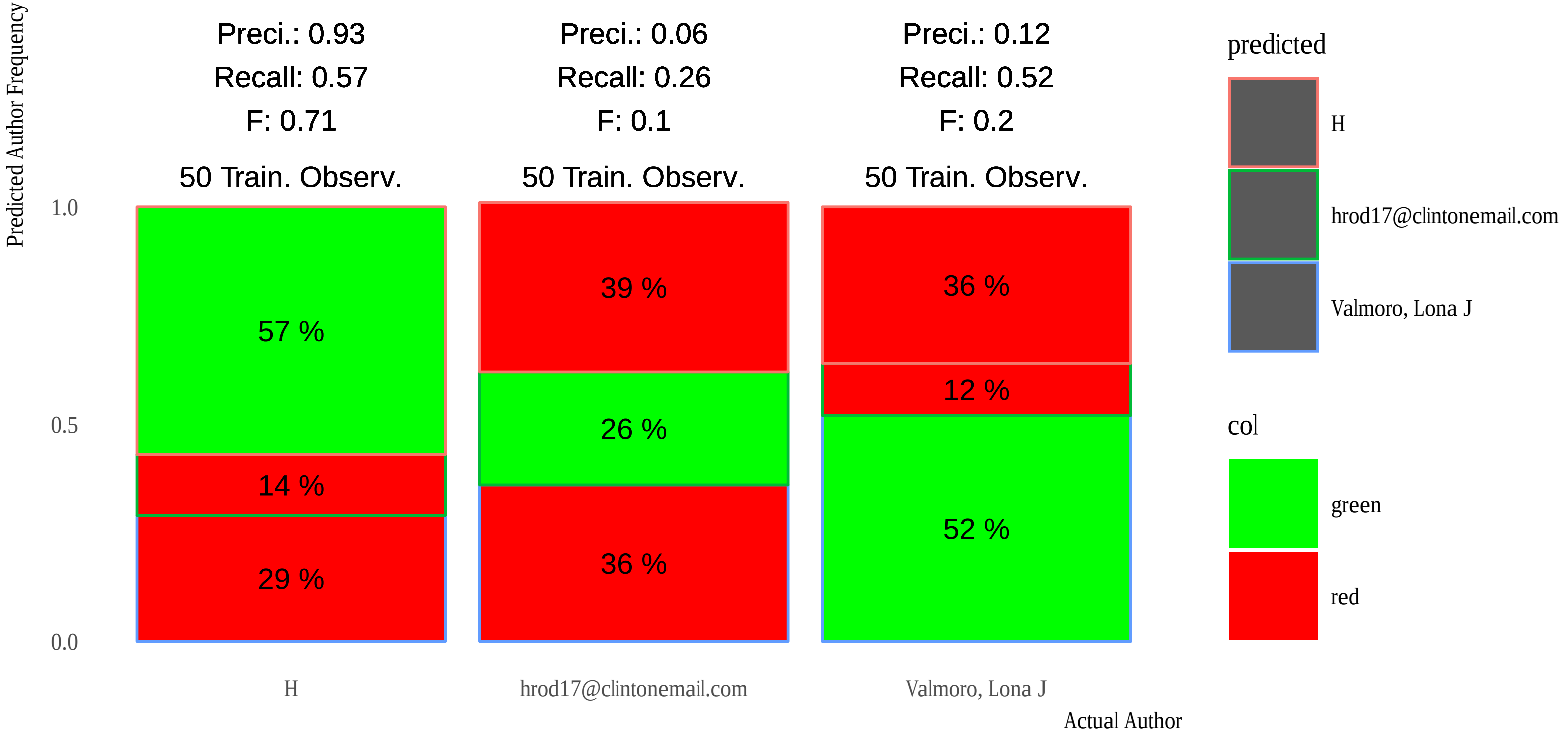
Author Sample 42



Author Sample 43



Author Sample 44



Preci.: 0.93
Recall: 0.57
F: 0.71

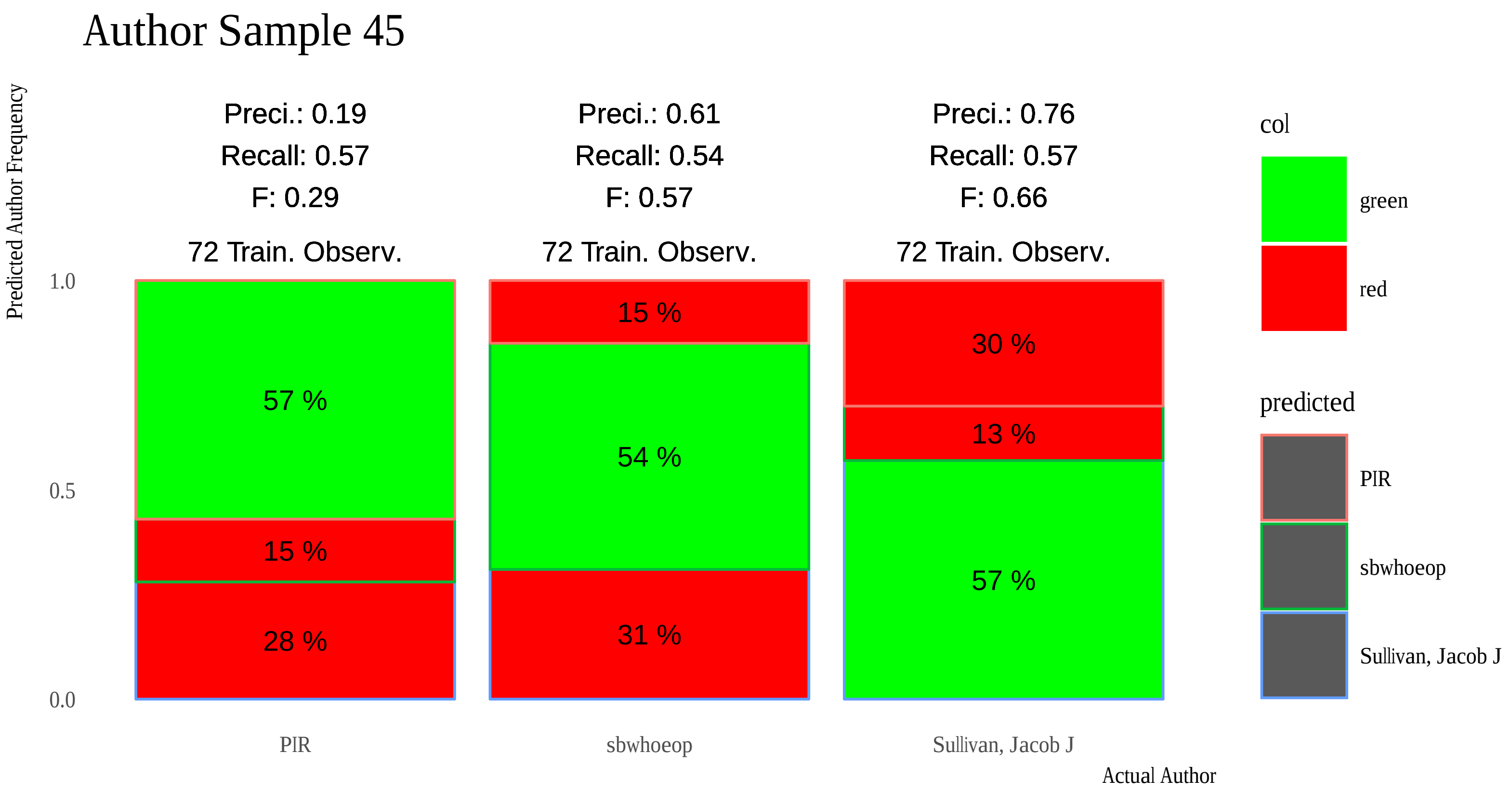
50 Train. Observ.

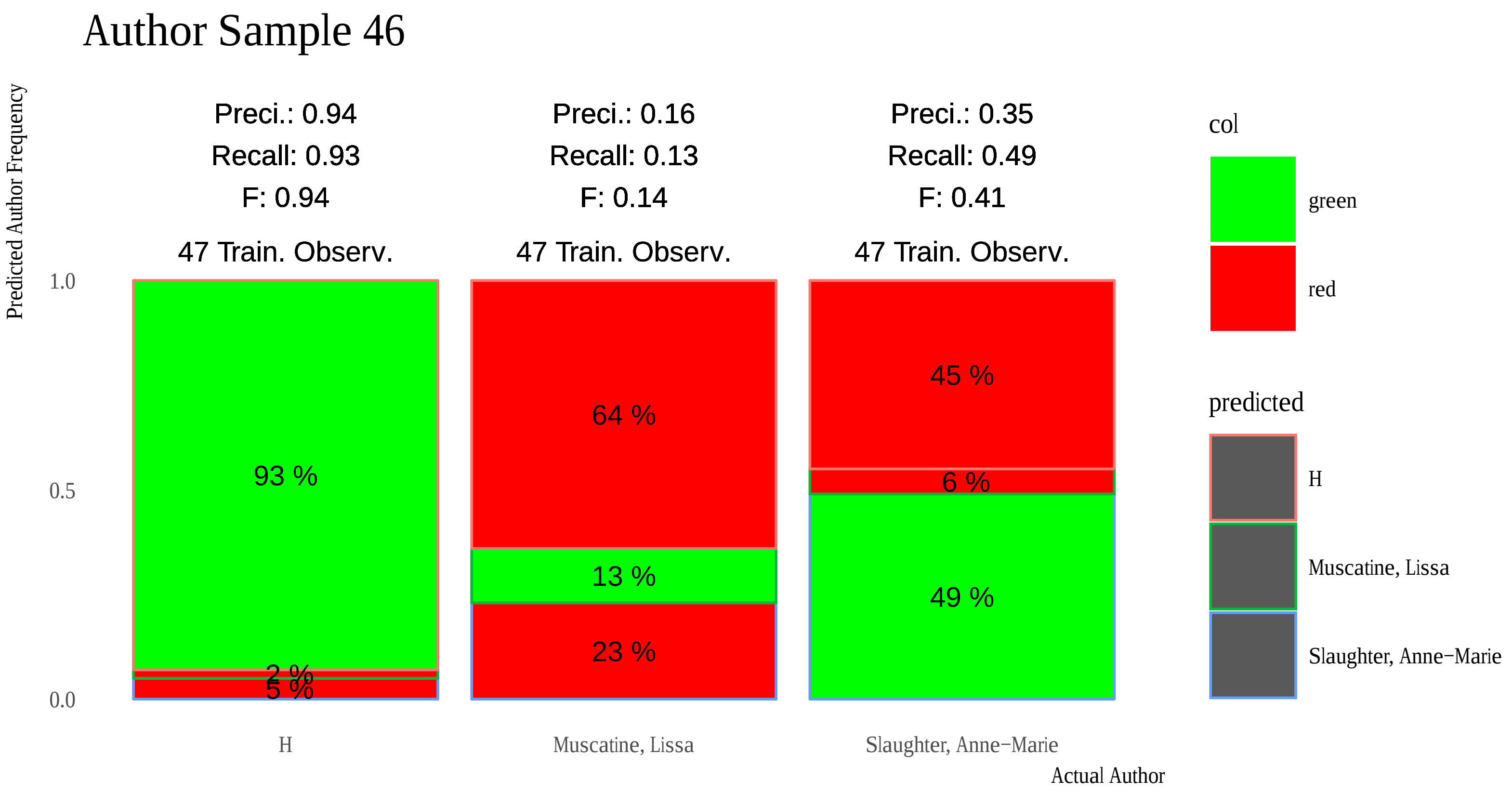
Preci.: 0.06
Recall: 0.26
F: 0.1

50 Train. Observ.

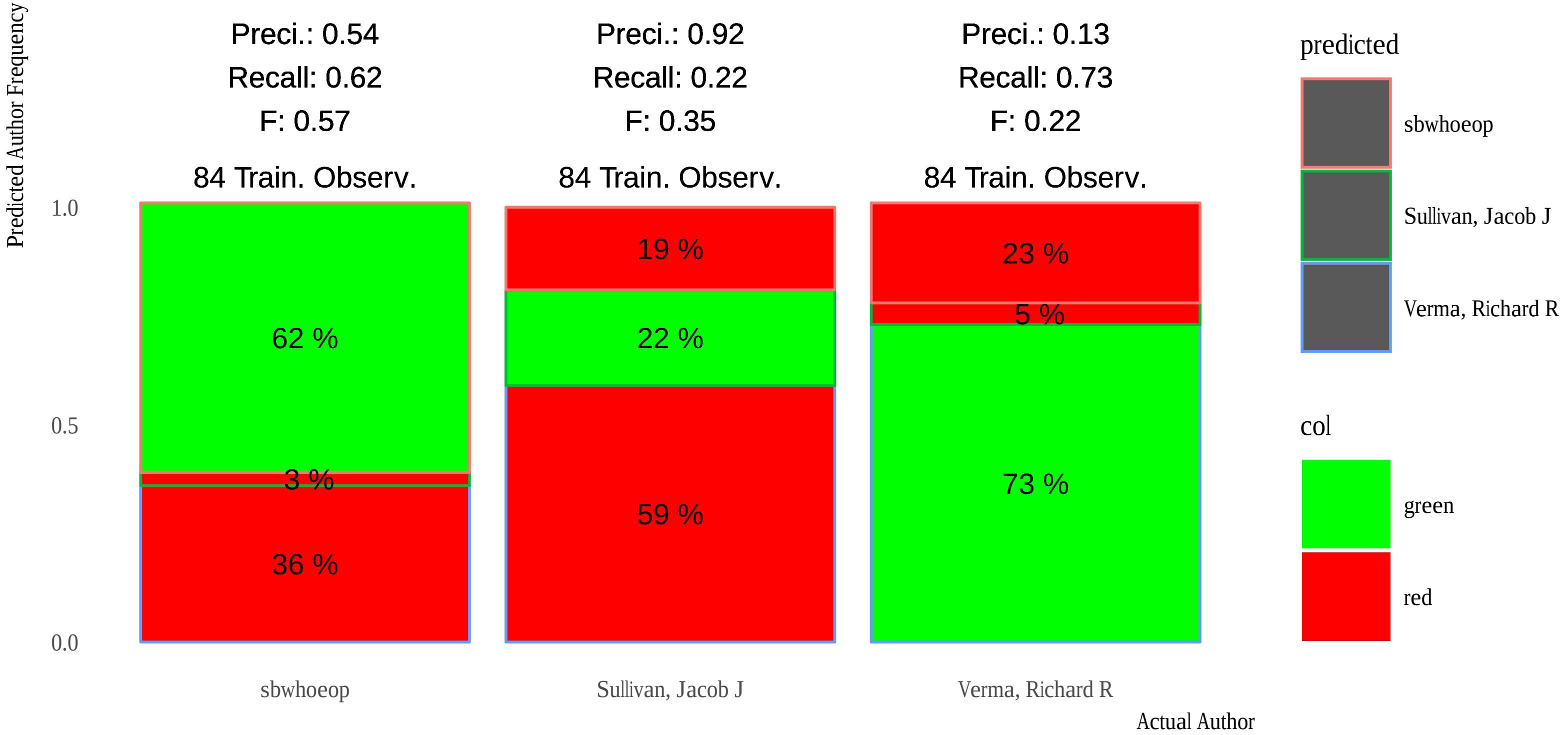
Preci.: 0.12
Recall: 0.52
F: 0.2

50 Train. Observ.

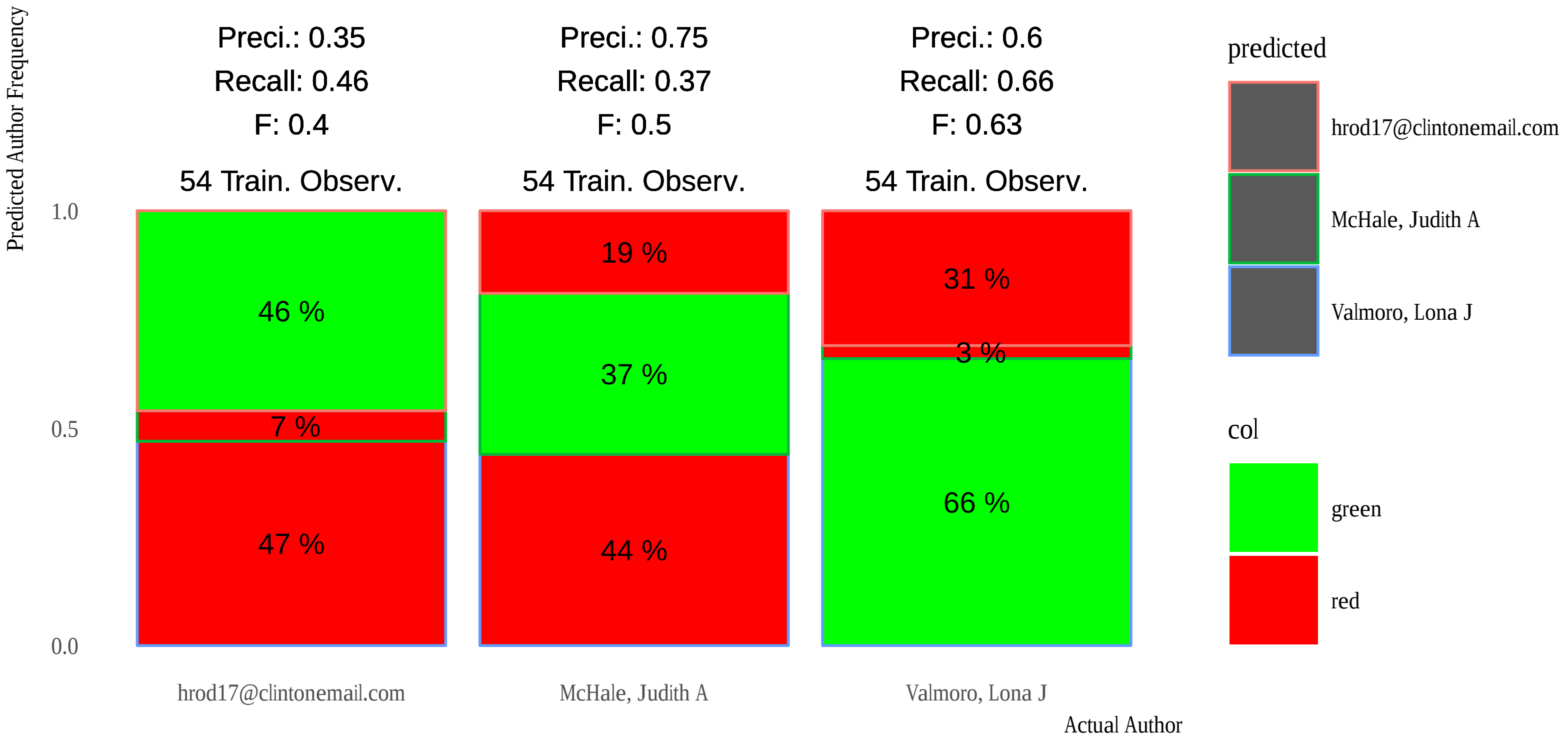

Author Sample 45
Predicted Author Frequency
Preci.: 0.19
Recall: 0.57
F: 0.29
72 Train. Observ.
Preci.: 0.61
Recall: 0.54
F: 0.57
72 Train. Observ.
Preci.: 0.76
Recall: 0.57
F: 0.66
72 Train. Observ.
1.0
0.5
0.0
57 %
15 %
28 %
15 %
54 %
31 %
30 %
13 %
57 %
PIR
sbwhoeop
Sullivan, Jacob J
Actual Author
col
green
red
predicted
PIR
sbwhoeop
Sullivan, Jacob J


Author Sample 46
Preci.: 0.94
Recall: 0.93
F: 0.94
47 Train. Observ.
Preci.: 0.16
Recall: 0.13
F: 0.14
47 Train. Observ.
Preci.: 0.35
Recall: 0.49
F: 0.41
47 Train. Observ.
Predicted Author Frequency
1.0
0.5
0.0
93 %
2 %
5 %
64 %
13 %
23 %
45 %
6 %
49 %
H
Muscatine, Lissa
Slaughter, Anne−Marie
Actual Author
col
green
red
predicted
H
Muscatine, Lissa
Slaughter, Anne−Marie

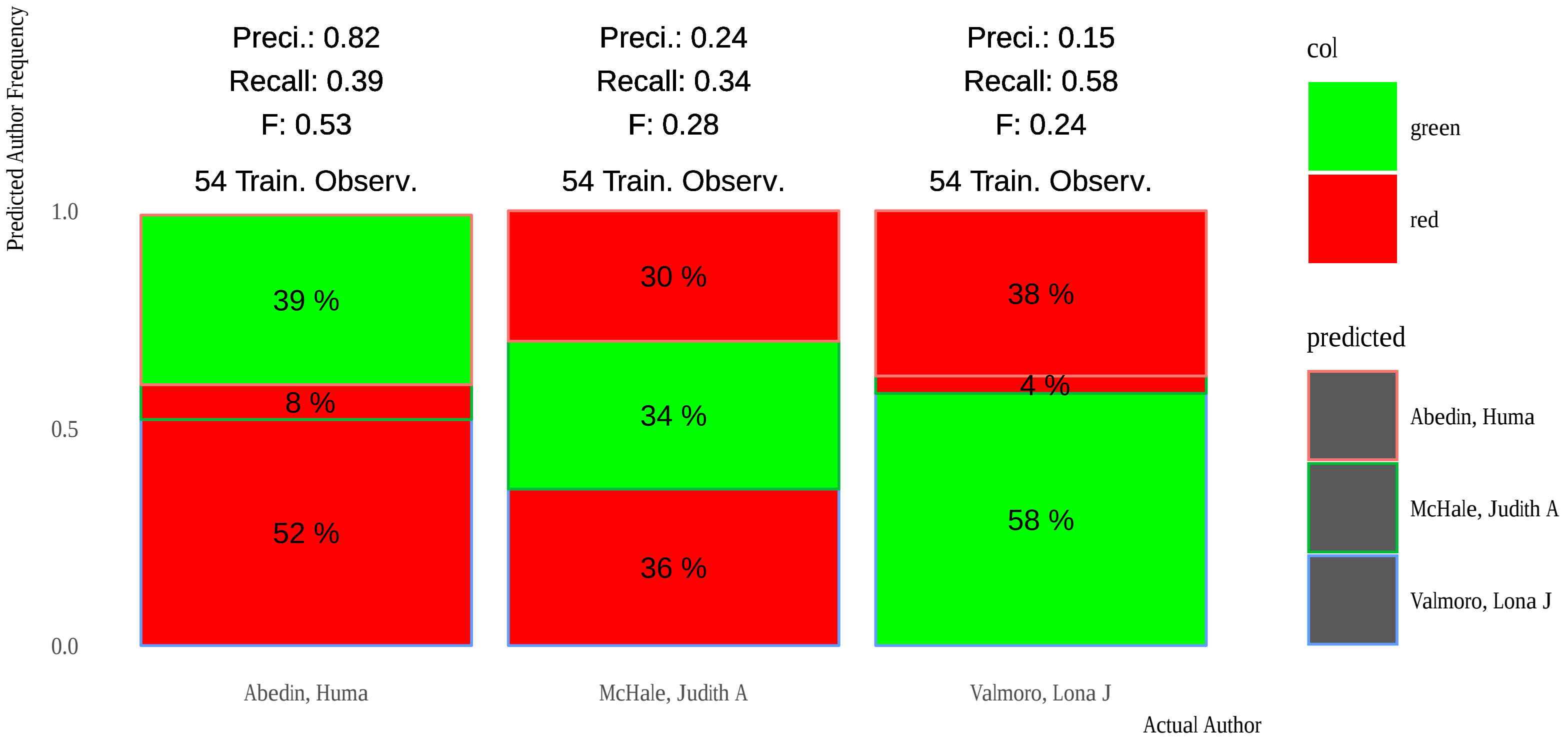
Author Sample 47

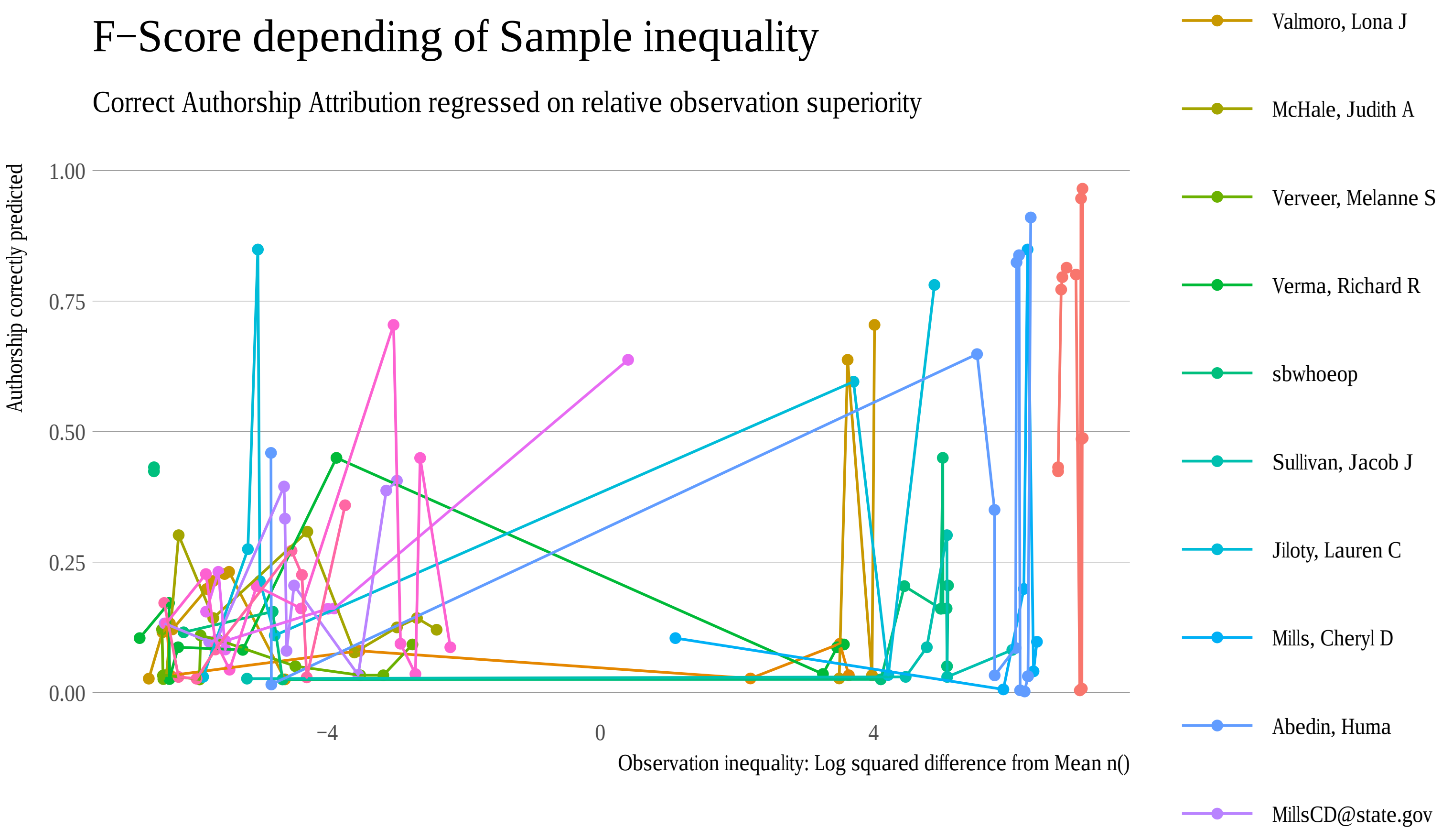


Author Sample 48



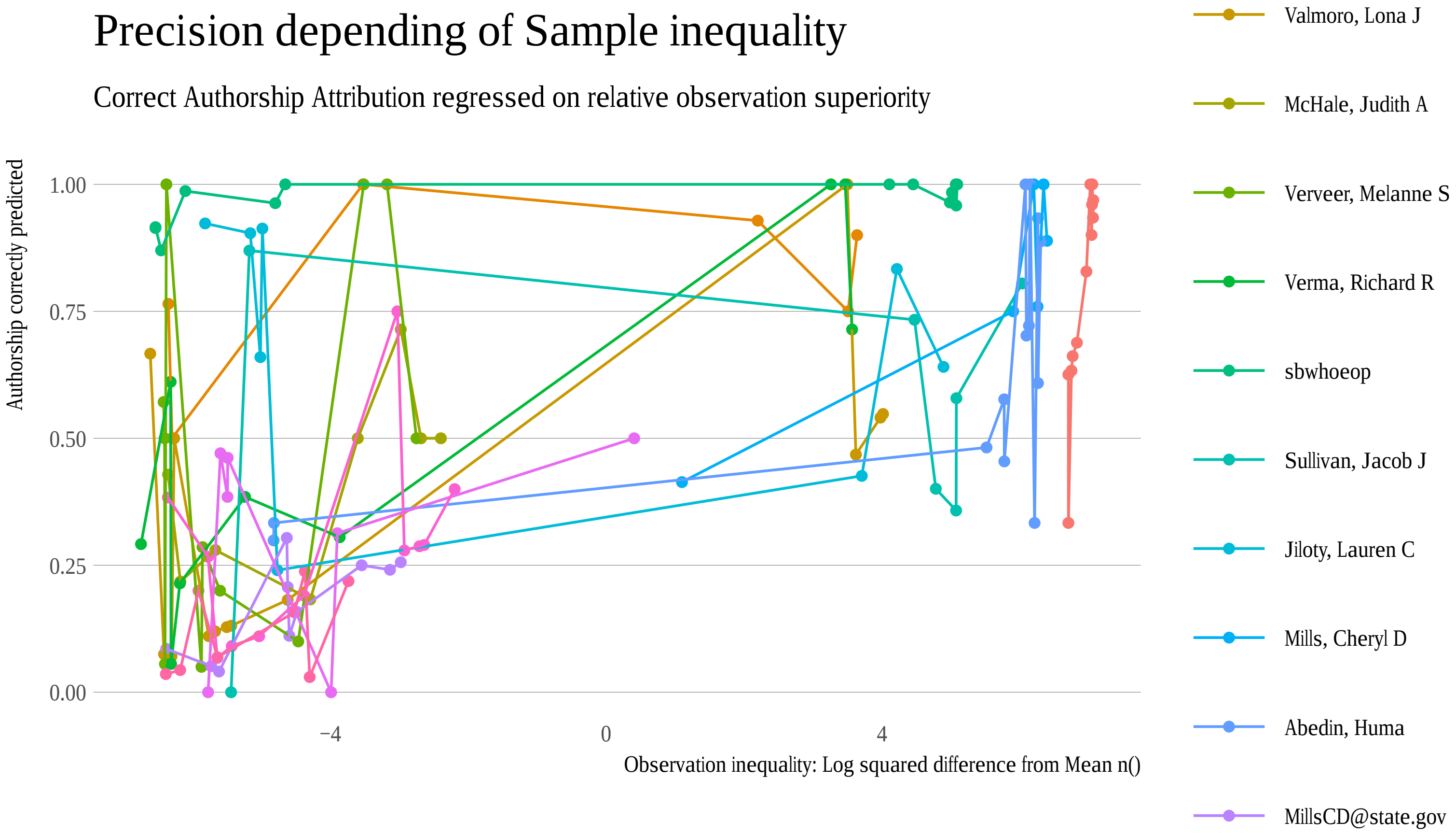
Author Sample 49





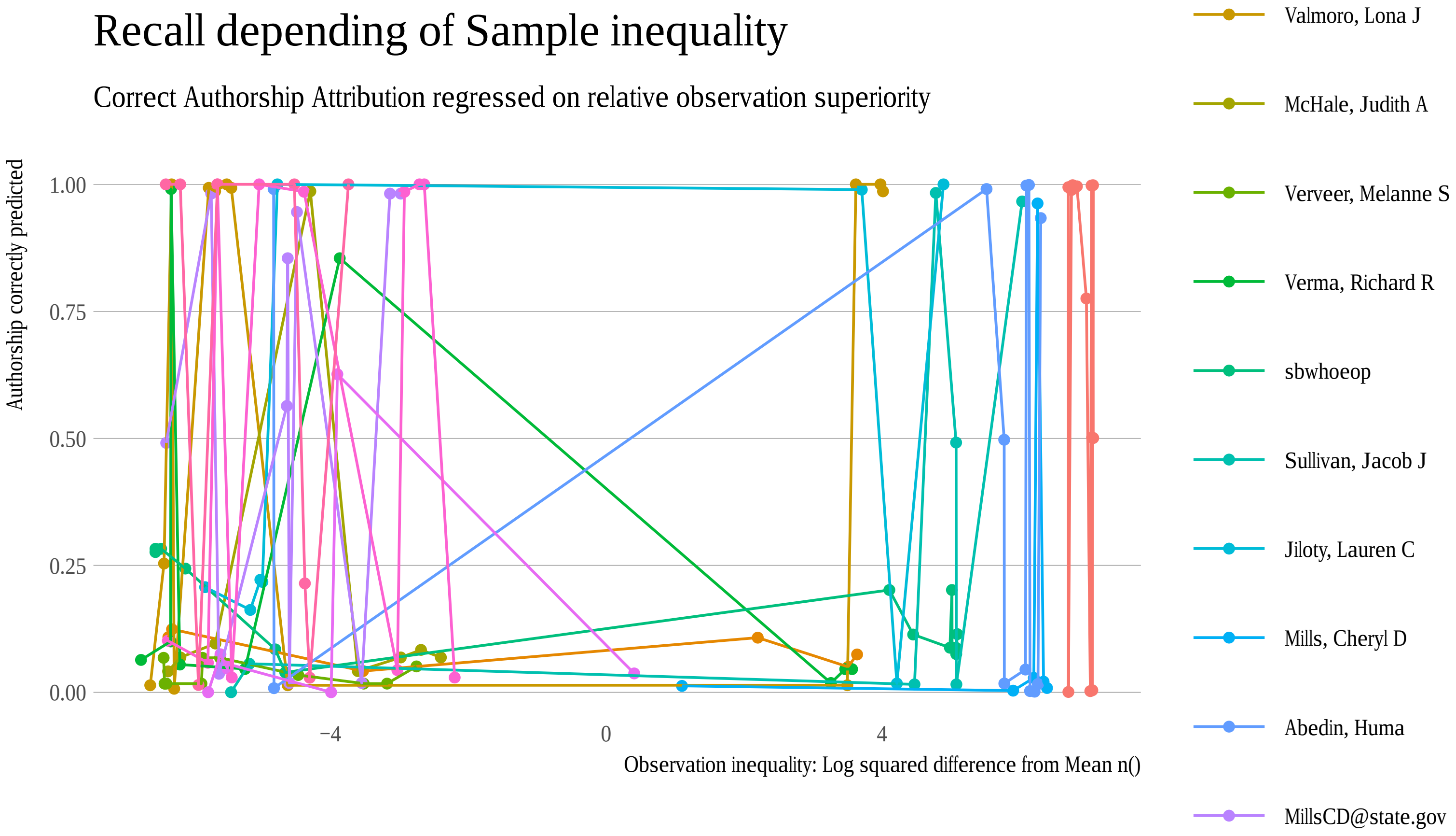
F−Score depending of Sample inequality

Correct Authorship Attribution regressed on relative observation superiority



Precision depending of Sample inequality

Correct Authorship Attribution regressed on relative observation superiority



Recall depending of Sample inequality

Correct Authorship Attribution regressed on relative observation superiority