Stability of Forensic Text Comparison System

Susan Brown[1], Shunichi Ishihara[2]
firstname.lastname@anu.edu.au

[1,2] Speech and Language Laboratory, The Australian National University, Canberra, Australia
[1] College of Arts and Social Science, The Australian National University, Canberra, Australia
[2] College of Asia and the Pacific, The Australian National University, Canberra, Australia

Abstract

This study investigates how the reliability of likelihood ratio (LR)-based forensic text comparison (FTC) systems is affected by the sampling variability regarding author numbers in databases. When 30–40 authors (each contributing two 4 kB documents) are included in each of the test, reference and calibration databases, the experimental results demonstrate: 1) the overall performance (validity) of the FTC system reaches the same level of performance as a system with 720 authors, and 2) the variability of the system performance (reliability) starts to converge. A similar trend can be observed regarding the magnitude of fluctuation in derived LRs. The variability of the overall system performance is mostly due to the large variability in calibration, not discrimination. Furthermore, FTC systems are more prone to instability when the dimension of the feature vector is high.

1 Introduction

Many studies on source-detection systems emphasise improving the system's overall performance or system validity. In data-driven forensic science, empirical testing of the system, demonstrating the system's validity and reliability, is essential for evidence to be accepted in court (President's Council of Advisors on Science and Technology [U.S.], 2016). However, studies of reliability are limited (Wang et al., 2022). The current study analyses the reliability of forensic text comparison (FTC) regarding the effect of sampling variability and sample size. Sample size is a well-known factor affecting the system's validity and reliability (Ishihara, 2016, 2020).

When reporting the system performance in court as an expert witness, an astute lawyer may question whether the system could achieve the same level of performance if it were tested with another set of samples from the same population, particularly when the sample size is small. Thus, forensic scientists must measure reliability to reduce the probability of a miscarriage of justice (Brümmer and Swart, 2014; Morrison, 2011, 2016).

FTC typically involves the analysis of two documents: the source-known (suspect) document and the source-questioned (offender) document. It is widely acknowledged that expert opinions should be expressed as the strength of evidence, quantified as a likelihood ratio (LR) (Robertson et al., 2016). The importance of the LR framework, long argued as the logically and legally correct framework (Aitken, 1995; Aitken and Stoney, 1991), is now recognised for FTC (Grant, 2022). However, FTC studies based on the LR framework are limited (cf. Ishihara, 2021; Ishihara and Carne, 2022).

The current study investigates the reliability and validity of the LR-based FTC system by conducting repeated random sampling (50 iterations) of a given number of authors from a large database. The experiments are conducted with two different dimensions of feature vectors (20 and 500), anticipating some different degrees of reliability. Logistic Regression calibration (Morrison, 2013) was employed to convert the estimated scores with the Dirichlet-Multinomial model (Bolck and Stamouli, 2017) to LRs. See Subsection 2.4 for the details of calibration as it is used in a difference sense from ML/NLP. Word unigrams are used to model each document.

2 Methodology

2.1 Database and Comparisons

The present study assessed a database of 4 kB-sized documents extracted from the dataset prepared by Ishihara (2021). This database is based on the Amazon Product Data Authorship Verification Corpus (Halvani et al., 2017) and

includes 4,320 documents (two documents each from 2,160 authors. The average document length is 830.47 words (standard deviation, 33.998 words). Ishihara (2021) provided justification for the use of product review texts for forensic studies.

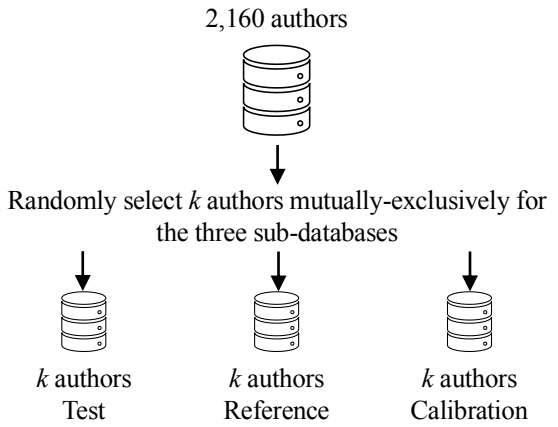

Figure 1: Random selections of authors

In order to test the reliability of the FTC system; in other words, the (in)stability of the system, arising from the sampling variability and the number of authors included in the experiments, k (={5,10,20,30,40,60,80,100,125,150,175,200,225, 250}) authors were randomly selected for each of the three sub-databases of test, reference and calibration 50 times (see Figure 1). Therefore, 50 random samplings of data for each experiment of k authors were conducted. For a small k, a high level of fluctuation in system performance across the 50 iterations of the experiment is predicted.

From k authors in the test sub-database, k same-author (SA) and $\binom{k}{2}$ different-author (DA) comparisons are possible. Note that more DA than SA comparisons can be made for the same number of authors.

The system is assumed to be unstable if the dimension of a feature vector is high because the amount of data for the statistical model to be appropriately trained exponentially increases as the feature dimension increases (Silverman, 1986). As such, two different feature numbers (20 and 500) are compared to investigate to what extent the feature vector dimension influences system reliability.

2.2 Tokenisation and Word Unigrams

The `tokens()` function of the quanteda R library (Benoit et al., 2018), which recognises punctuation marks and special characters as single words, was used to perform tokenisation. No stemming algorithm was used. Each document was modelled with word unigrams. From the entire database, the 500 most frequent words (term frequency) were identified, and those words, sorted in descending order of frequency, were used as the elements of a feature vector; i.e. a global feature selection was applied.

Figure 2 illustrates the process of calculating LRs. The LRs are calculated for SA and DA comparisons generated from the test sub-database. Estimating LRs is a two-stage process consisting of the score calculation stage, followed by the calibration stage. For the score calculation stage, the same processes are applied to the test and calibration sub-databases. However, the scores of the test sub-database were calibrated to LRs, while the score of the calibration sub-database were used to train the calibration model. The documents stored in the reference database are used to obtain statistical information for the typicality assessment of the documents being compared.

2.3 Score Calculation

When the LR interpretive framework is applied to FTC, textual evidence (E) is assessed under the two competing hypotheses; the SA (H_{SA}) and the DA (H_{DA}). These are generally called the prosecution and defence hypotheses, respectively. The evidence usually includes two types of text samples: the source-known text from the suspect (X) and the source-questioned text from the offender (Y). Thus, the score is expressed as given in Equation (1).

$$Score = \frac{f(E|H_{SA})}{f(E|H_{DA})} = \frac{f((X,Y)|H_{SA})}{f((X,Y)|H_{DA})} \quad (1)$$

Each piece of evidence (X and Y) are modelled with the counts of a given set of unigrams (m; maximum $m = 500$): $X = \{x_1, x_2, \cdots x_m\}andY = \{y_1, y_2, \cdots y_m\}$. The similarity between X and Y is assessed as the probability of X against the multinomial model given Y of which the parameter is $\Pi = \{\pi_1, \pi_2, \cdots \pi_m\}$. If a prior is assumed for the model parameter, it can be formulated by a Dirichlet distribution with a hyperparameter ($A = \{a_1, a_2, \cdots a_m\}$). With the multivariate Beta function ($B = (\Gamma(a_1) \cdots \Gamma(a_m))/(\Gamma(a_1 + \cdots a_m))$), Equation (1) can be rewritten as Equation (2).

$$Score = \frac{B(A)B(A+X+Y)}{B(A+X)B(A+Y)} \quad (2)$$

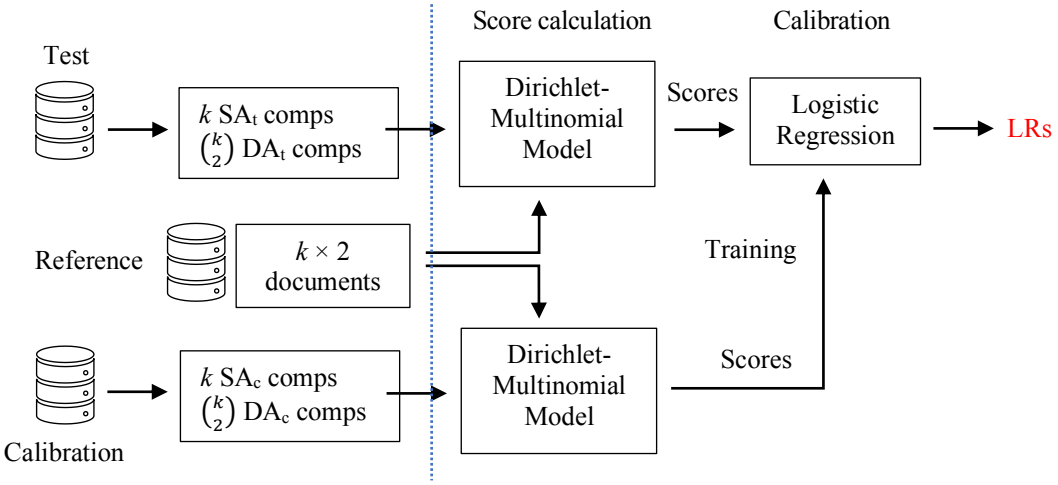

Figure 2: Process of calculating likelihood ratios (LRs). k = the number of authors in sub-database; SA = same-author; DA = different-author; t = test sub-database; comps = comparisons; c = calibration sub-database.

The maximum likelihood estimation was employed to obtain the parameter values of the Dirichlet model using the reference sub-database. Note that although the Dirichlet-Multinomial model follows Bayesian logic, the parameters of the Dirichlet model are fixed in this study instead of random variables. See Section 4 for the application of a Bayesian statistical approach as a future study. Refer to Bolck and Stamouli (2017) for a detailed derivational process from Equation (1) to (2).

2.4 Score to Likelihood Ratio Conversion

The calculated score for each comparison of the test sub-database must be converted to a LR, as the uncalibrated score alone cannot be interpreted as demonstrating the strength of the evidence. Logistic regression is most commonly used to calculate the LR (Morrison, 2013; Ramos and Gonzalez-Rodriguez, 2013). The calculated comparison scores from the calibration sub-database are used to train the logistic regression model.

2.5 System Evaluations

For the evaluation of a forensic system of which the outcome is used to assist the factfinders' legal decision, those evaluation metrics which are based on classification or identification accuracy are not appropriate. This is because 1) the category-based classification accuracy does not properly assess the magnitude of LRs, which is continuous and 2) it implicitly refers to the accuracy of the decision making: guilty vs not guilty; which is only permitted for the factfinders.

The standard evaluation metric for LR-based forensic systems is the log LR cost (C_{llr}), mathematically expressed in Equation (3).

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{SA}}\sum_{i}^{N_{SA}} log_2\left(1+\frac{1}{LR_{SA_i}}\right) + \frac{1}{N_{DA}}\sum_{j}^{N_{DA}} log_2\left(1+LR_{DA_j}\right)\right) \quad (3)$$

In Equation (3), N_{SA} and N_{DA} are the number of SA and DA comparisons, respectively, and LR_{SA_i} and LR_{DA_j} are the i_{th} SA and j_{th} DA linear LRs, respectively. The C_{llr} is the overall average of the pooled costs calculated for all LRs. A certain amount of cost is computed for each LR, but the cost is greater as the value is further away from unity (LR = 1), and contrary-to-fact LRs give rise to a far greater cost than consistent-with-fact LRs. The closer to C_{llr} = 0, the better the performance. A $C_{llr} \geq 1$ denotes that the evidence is not informative for inference. The C_{llr} can be decomposed into C_{llr}^{min} and C_{llr}^{cal} to assess the discrimination and calibration performance of the system, respectively; thus, $C_{llr} = C_{llr}^{min} + C_{llr}^{cal}$.

The variability or (in)stability of the performance across the 50 random samplings of k authors is quantified by the range of the C_{llr} values observed across the 50 iterations.

3 Results: System Performance

3.1 Reference Performance

The 2,160 authors of the entire database were evenly separated into three sub-databases, with 720 authors in each. With this maximum number of

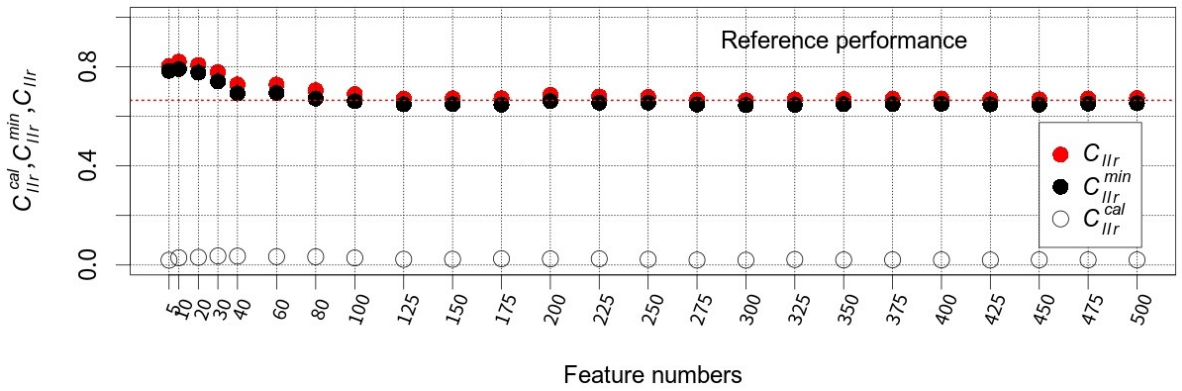

Figure 3: Reference performance of the forensic text comparison system with 720 authors in each sub-database. The red dotted horizontal line indicates the best C_{llr} value (0.66469), attained with 300 features.

authors (720) in each sub-database, a set of experiments was carried out by gradually increasing the number of features = {5,10,20,30,40,60,80,100,125,…500}) to understand how well the FTC system works with the full dataset, but with different feature numbers. The C_{llr}, C_{llr}^{min} and C_{llr}^{cal} were plotted as a function of the number of features in Figure 3.

Regardless of the feature numbers, the C_{llr}^{cal} values were all close to zero, indicating that the resultant LRs are very well calibrated. The overall performance of the system (C_{llr}) improves as the feature number increases to approximately 125 features, after which the C_{llr} value stays more or less unchanged even with the addition of more features. The system achieved the best performance for 300 features (C_{llr} = 0.66469). The C_{llr}^{min} values exhibit a very similar trend to the C_{llr} values.

3.2 Variability in Performance

The reliability and validity of the system caused by the random sampling of given numbers of authors for the sub-databases were analysed. For this, the mean and range of the C_{llr} values of the 50 iterations of experiments were plotted together according to the number of authors in Figure 4; Panel a) shows the data for 20 features, and Panel b) shows the data for 500 features.

For the mean C_{llr} values, the system does not require many authors to achieve the same level of performance as systems with the full number of authors. Figure 4a and 4b demonstrate that regardless of the feature numbers, systems with 10 authors averaged the same level of performance as the systems using the full number of authors. When the feature dimension is low (20 features) (see Figure 4a), the average system performance is similar for any number of authors. However, when the feature dimension is high (500 features) (Figure 4b), analyses using 5 authors substantially worsened the system performance. This indicates that system (in)stability is subject to the feature dimension.

As can be seen from Figure 4b, the range of the C_{llr} values was large for 50 iterations for 5 authors but narrowed with an increasing number of authors. Although the range appeared to converge with the inclusion of 30–40 authors, it continues to decrease in very small increments as the number of authors further increases. With only 5 authors, the range of the C_{llr} values is far wider for 500 features (116.292) than for 20 features (2.53864).

To visually compare the levels of (in)stability caused by the different feature numbers, the ranges of C_{llr} values for 20 and 500 features are plotted together in Figure 4c. A narrower scale (between 0 and 1) is used for the Y-axis of Figure 4c to make visual comparison easier. However, this scale reduction resulted in some C_{llr} range values being out of the plot; thus, the C_{llr} range values are given in Table 1 for 5, 10 and 20 authors.

Author number	Feature number	
	20	**500**
5	2.53864	116.292
10	1.25353	1.80968
20	0.59871	0.83678

Table 1: Ranges of the C_{llr} values with 20 and 500 features.

Figure 4c and Table 1 show that the C_{llr} range values are higher for 500 features than 20 features. However, the ranges are similar for author numbers ≥150. In contrast, for fewer authors (≤20), the

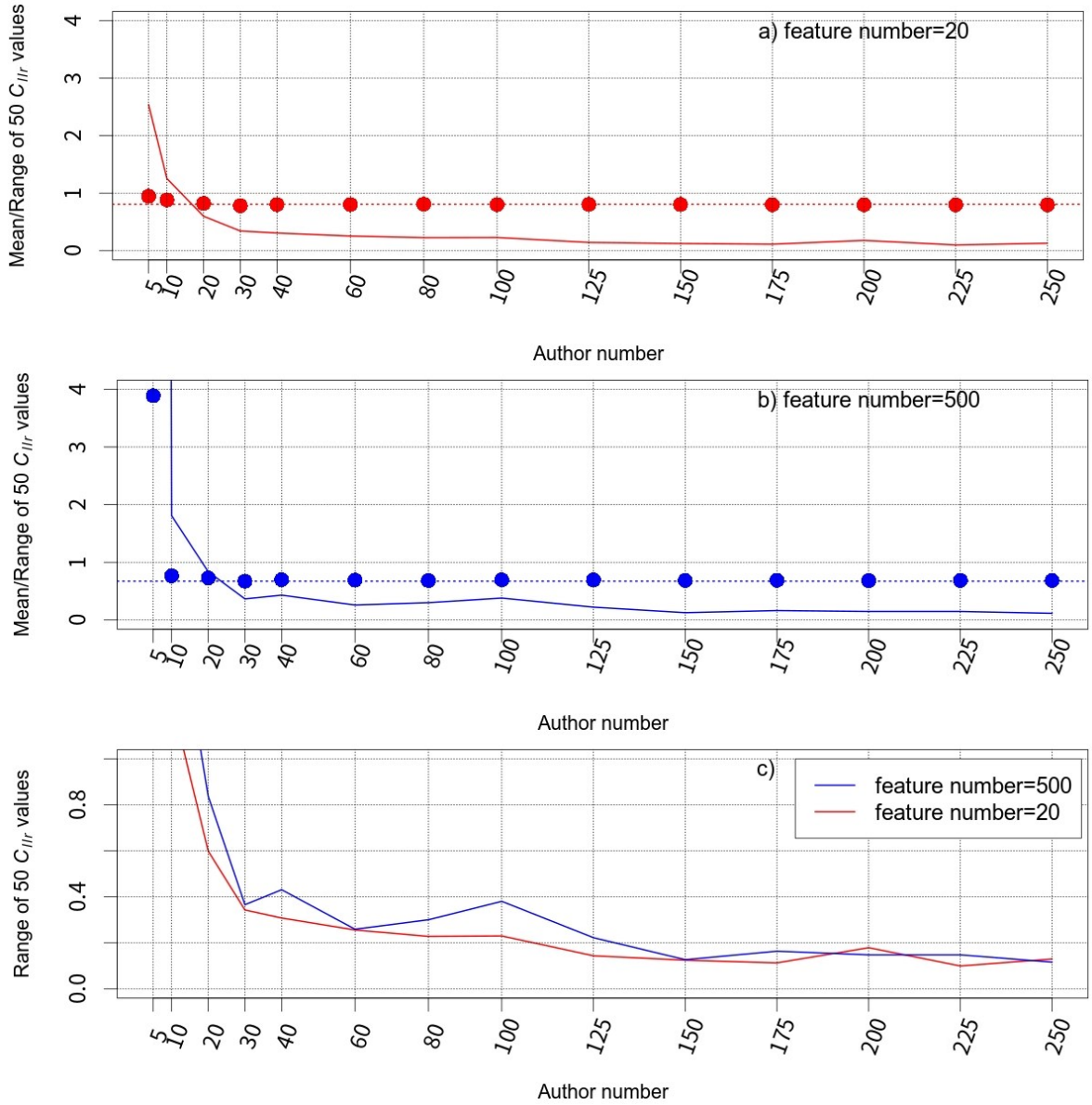

Figure 4: Mean C_{llr} values (circles), plotted as a function of the number of authors with the range of the C_{llr} values (solid curves). Panels a) and b) demonstrate the C_{llr} values for 20 and 500 features, respectively. The ranges of the C_{llr} values are plotted together in Panel c) for better visual comparison. The dotted horizontal lines of Panels a) and b) show the C_{llr} value for the maximum authors (720). Note that some values extend beyond the range of the Y-axis, which is narrower in Panel c).

difference in the C_{llr} range between 20 and 500 features is larger (113.75, 0.55615 and 0.23807, for 5, 10 and 20 authors, respectively) than for author numbers >20.

The experimental results presented in this subsection demonstrate that the performance instability caused by the sampling variability is evident in FTC. When the author number is very small (5 authors), the magnitude of the performance instability, measured in terms of the range of C_{llr} values, is large. Equally, the average performance is low compared to the systems with the full number of authors.

However, performance instability is quickly reduced as more authors are added. For example, with 30–40 authors, the range of C_{llr} values becomes substantially moderate and starts to converge. With 30–40 authors, the average performance of the system is as good as that of a system with the full number of authors.

It appears that the (in)stability of the system is interrelated with the number of features. That is, the system is prone to instability with a higher feature vector dimension. In particular, the instability is more sizeable with a small number of authors, but becomes negligible when many authors (≥150) are

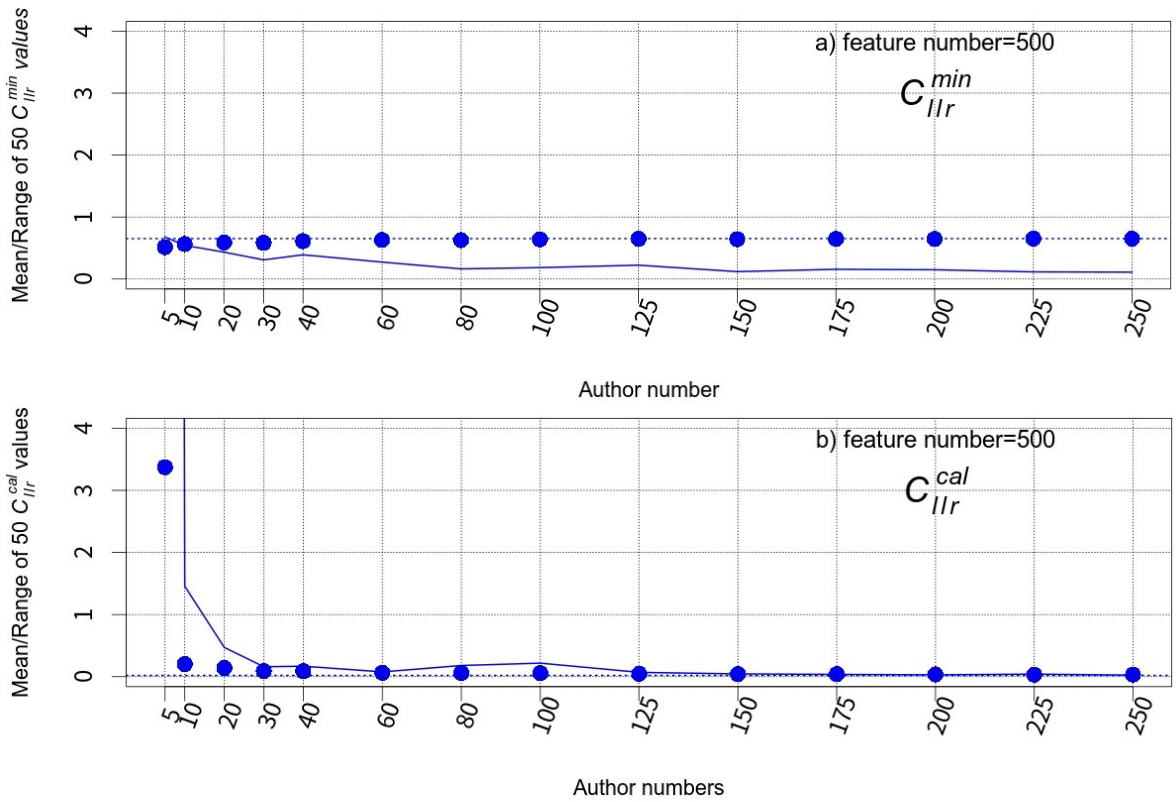

Figure 5: Mean C_{llr}^{min} (a) and C_{llr}^{cal} (b) values (circles), plotted as a function of author numbers, the curves show the range of the C_{llr}^{min} and C_{llr}^{cal} values (curves). The dotted horizontal lines in a) and b) show the best C_{llr}^{min} and C_{llr}^{cal} values, respectively, for the maximum authors (720). Note that some values extend beyond the range of the Y-axis.

included in each sub-database; therefore, there is improved stability when the statistical model is trained with an appropriate amount of data.

3.3 Cause of Variability

Subsection 3.2 investigated to what extent 50 random samplings of a given set of authors affect the reliability and validity of the system by assessing the C_{llr} values. However, as explained in Subsection 2.5, the C_{llr} is an assessment metric for the overall performance of a LR-based system, and consists of two components: discrimination (C_{llr}^{min}) and calibration (C_{llr}^{cal}). In this subsection, the previously observed variability is further investigated from the viewpoints of the discrimination and calibration performance.

Figure 5 shows how the mean and ranges of the C_{llr}^{min} and C_{llr}^{cal} values vary as a function of author numbers. Figure 5 shows this variation for 500 features, and the observation made for 20 features is uniform. As can be observed in Figure 5a, the mean C_{llr}^{min} value stays more or less the same regardless of the author numbers (even for 5 authors). This observation means that, on average, the discrimination ability of the system is not largely influenced by the number of included authors.

The range of discrimination ability, measured using C_{llr}^{min}, displays trifling fluctuations even with the small numbers of authors (≤40 authors), and the degree of fluctuation is far smaller than the ones observed for the C_{llr} (see Figure 4).

In contrast to the discrimination ability of the system, the changes in the mean and range of the C_{llr}^{cal} values display a similar trend as observed for the C_{llr} counterparts presented in Figure 4. Even with as few as 10 authors, a very similar level of mean calibration performance (C_{llr}^{cal} = 0.20338) is found in the case with the maximum number of authors (C_{llr}^{cal} = 0.01955). However, with 5 authors, the mean C_{llr}^{cal} value deviates (C_{llr}^{cal} = 3.37324) far from the calibration performance achieved with the maximum number of authors (C_{llr}^{cal} = 0.01955). Likewise, the range of the C_{llr}^{cal} values is large (116.06) with 5 authors. As can be observed in Figure 5b, the large range observed for 5 authors decreases as the author number increases, and the range becomes as narrow as 0.16045 with 30 authors.

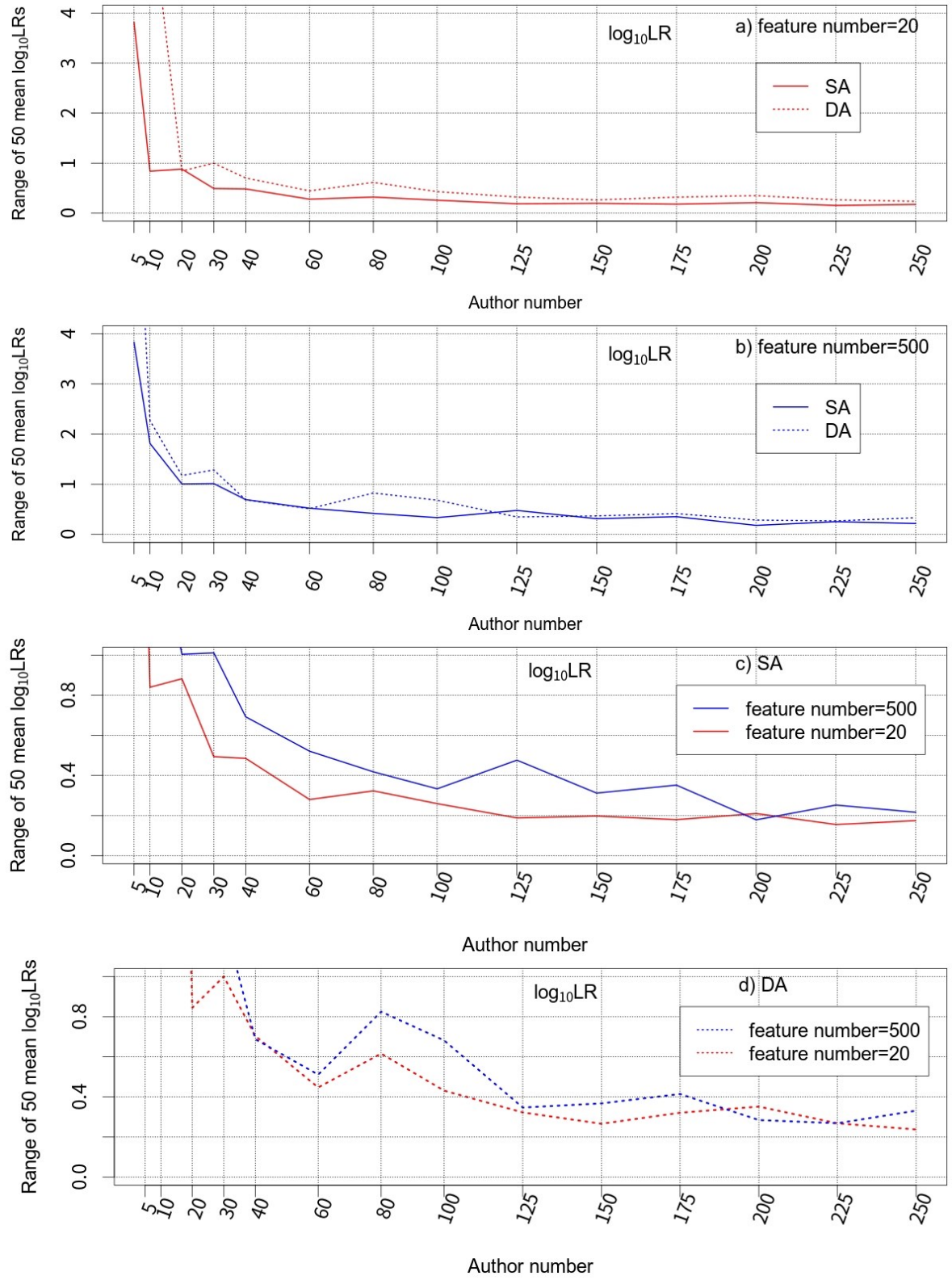

Figure 6: Range of the 50 mean $\log_{10}$LRs plotted separately for (a) 20 features, (b) 500 features, (c) SA LRs and (d) LRs. The Y-axis is narrower for Panels c) and d). Some values are beyond the range of the Y-axis.

The different characteristics displayed between Panels a) and b) of Figure 5 for discrimination ability and calibration, respectively, mean that the deterioration in mean overall performance and wide range of performance fluctuation shown for a small number of authors in Figure 4 are largely due to poor performance in calibration, not discrimination performance.

3.4 Variability in Likelihood Ratios

The variability in performance reported in Subsection 3.2 is fundamentally caused by variability in the derived LRs. Thus, this subsection investigates the characteristics of the derived SA and DS LRs. For each number of k authors, experiments were repeated 50 times by randomly sampling authors from the entire database for each sub-database. Therefore, for the same k, each iteration of the experiment should return k SA LRs and $\binom{k}{2}$ DA LRs. The mean values of the SA and DA LRs were calculated for each iteration. Wide variation in the mean LR was expected for a small k. The range of the mean SA and DA LRs was also calculated for each k to assess the degree of variability observed in LRs.

The ranges of the mean LRs for 20 and 500 features are plotted in Figure 6a and 6b, respectively. As for the variability in overall performance (see Figure 4), the range of the mean LRs observed with 5 authors quickly tapers and starts converging for 30–40 authors, regardless of the number of features and whether SA or DA comparisons are made.

In Figure 6c and 6d, the range of mean $\log_{10}$LR values are plotted against SA or DA LRs, respectively, to visually investigate any influences arising from the different number of features on the (in)stability of the derived LRs. A narrower Y-axis range was used in Figures 6c and 6d; the values beyond the Y-axis range are given in Table 2.

	Author number	Features number	
		20	**500**
SA	**5**	3.81995	3.83104
	10	0.84000	1.81251
	20	0.88233	1.00447
DA	**5**	4.35546	8.46658
	10	6.27607	2.26897
	20	0.84479	1.17478

Table 2: Ranges of the mean $\log_{10}$LR values with 20 and 500 features for 5, 10 and 20 authors.

Although the data in Figure 6c and 6d and Table 2 is not straightforwardly clear for the DA LRs, the derived LRs are susceptible to instability when the dimension of the feature vector is high. However, this difference is negated when 200 or more authors are included.

4 Conclusions

This study investigated the reliability and validity of a LR-based FTC system by varying the sampling number and sample size. When only 5 authors were included in the test, reference and calibration sub-databases (two 4 kB documents from each author), the reliability and validity of the system were considerably compromised. However, adding more authors to the database compensated for this deterioration in reliability and validity. When 30–40 authors were included, the mean performance (validity) of the system was nearly equivalent to that for as many as 720 authors. Likewise, when 30–40 authors were included, the fluctuation (reliability) of the system performance substantially decreases and starts to converge. A similar observation was made for the derived LRs; the wide range of the mean LR values across 50 iterations of experiments with 5 authors greatly diminishes if 30–40 authors are included in each sub-database.

The experimental results also show: 1) a system with a high dimension of feature vector (500 features) is more prone to instability than a system with fewer feature vectors (20 features), and 2) the low reliability and poor validity found when a small number of authors are included (e.g., 5 and 10 authors) are largely due to the poor calibration, not discrimination ability, of the system.

The approach that was employed in this study is rather primitive; e.g. the number and type of features, and there would be considerable potential to improve the model, consequently leading to a better performance. However, this may compromise the stability of the system due to the resultant even higher dimensionality of feature vector. This needs further investigation, while seeking the benefits of feature selection/reduction.

In the current study, the (in)stability and overall performance of the FTC system was measured. However, besides the quantification, the instability of the system ultimately needs to be minimised to prevent the misinterpretation of evidence and miscarriage of justice. As such, it is essential to apply a Bayesian statistical approach that considers the degree of uncertainty to the LRs (Morrison and Poh, 2018) with the outcome being Bayes factors. Obviously, the application of a Bayesian statistical approach to FTC is another step to take as an extension of the current study.

Acknowledgments

The authors thank the reviewers for their valuable comments.

References

Colin G. G. Aitken. 1995. *Statistics and the Evaluation of Evidence for Forensic Scientists*. John Wiley & Sons Ltd, Chichester.

C. G. G. Aitken and D. A. Stoney. 1991. *The Use of Statistics in Forensic Science*. Ellis Horwood, New York, NY.

K. Benoit, K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, A. and Matsuo. 2018. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774–776. https://doi.org/10.21105/joss.00774

A. Bolck and A. Stamouli. 2017. Likelihood ratios for categorical evidence: Comparison of LR models applied to gunshot residue data. *Law, Probability and Risk*, 16(2–3):71–90. https://dx.doi.org/10.1093/lpr/mgx005

N. Brümmer and A. Swart. 2014. Bayesian calibration for forensic evidence reporting. *Proceedings of Interspeech*, 2014:388–392.

T. Grant. 2022. *The Idea of Progress in Forensic Authorship Analysis*. Cambridge University Press, Cambridge.

O. Halvani, C. Winter and L. Graner. 2017. Authorship verification based on compression-models. *Computing Research Repository*. ArXiv:1706.00516. Version 1.

S. Ishihara. 2016. An effect of background population sample size on the performance of a likelihood ratio-based forensic text comparison system: A Monte Carlo simulation with Gaussian mixture model. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 113–121.

S. Ishihara. 2020. The influence of background data size on the performance of a score-based likelihood ratio system: A case of forensic text comparison. In *Proceedings of the 18th Workshop of the Australasian Language Technology Association*, pages 21–31.

S. Ishihara. 2021. Score-based likelihood ratios for linguistic text evidence with a bag-of-words model. *Forensic Science International*, 327:110980. https://doi.org/10.1016/j.forsciint.2021.110980

S. Ishihara and M. Carne. 2022. Likelihood ratio estimation for authorship text evidence: An empirical comparison of score- and feature-based methods. *Forensic Science International*, 334:111268. https://doi.org/10.1016/j.forsciint.2022.111268

G. S. Morrison. 2011. Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, 51(3):91–98. https://dx.doi.org/10.1016/j.scijus.2011.03.002

G. S. Morrison. 2013. Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, 45(2):173–197. https://dx.doi.org/10.1080/00450618.2012.733025

G. S. Morrison. 2016. Special issue on measuring and reporting the precision of forensic likelihood ratios: Introduction to the debate. *Science and Justice*, *56*(5):371–373. http://dx.doi.org/10.1016/j.scijus.2016.05.002

G. S. Morrison and N. Poh. 2018. Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios/Bayes factors. *Science & Justice*, 58(3):200–218. https://doi.org/10.1016/j.scijus.2017.12.005

President's Council of Advisors on Science and Technology (U.S.). 2016. *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. Retrieved on 29 December 2018, from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf

D. Ramos and J. Gonzalez-Rodriguez. 2013. Reliable support: Measuring calibration of likelihood ratios. *Forensic Science International*, 230(1–3):156–169. https//dx.doi.org/10.1016/j.forsciint.2013.04.014

B. Robertson, G. A. Vignaux and C. E. H. Berger. 2016. *Interpreting Evidence: Evaluating Forensic Science in the Courtroom* (second ed.). John Wiley & sons Ltd, Chichester.

B. W. Silverman. 1986. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, New York.

B. X. Wang, V. Hughes and P. Foulkes. 2022. The effect of sampling variability on systems and individual speakers in likelihood ratio-based forensic voice comparison. *Speech Communication*, 138:38–49. https://doi.org/10.1016/j.specom.2022.01.009