The Influence of Background Data Size on the Performance of a Score-Based Likelihood Ratio System: A Case of Forensic Text Comparison

Shunichi Ishihara
Speech and Language Laboratory
The Australian National University
shunichi.ishihara@anu.edu.au

Abstract

This study investigates the robustness and stability of a likelihood ratio–based (LR-based) forensic text comparison (FTC) system against the size of background population data. Focus is centred on a score-based approach for estimating authorship LRs. Each document is represented with a bag-of-words model, and the Cosine distance is used as the score-generating function. A set of population data that differed in the number of scores was synthesised 20 times using the Monte-Carol simulation technique. The FTC system's performance with different population sizes was evaluated by a gradient metric of the log–LR cost (C_{llr}). The experimental results revealed two outcomes: 1) that the score-based approach is rather robust against a small population size—in that, with the scores obtained from the 40~60 authors in the database, the stability and the performance of the system become fairly comparable to the system with a maximum number of authors (720); and 2) that poor performance in terms of C_{llr}, which occurred because of limited background population data, is largely due to poor calibration. The results also indicated that the score-based approach is more robust against data scarcity than the feature-based approach; however, this finding obliges further study.

1 Introduction: The Likelihood Ratio Framework and Forensic Text Comparison

The likelihood ratio (LR) conceptual framework has been studied for its effect on various types of forensic evidence; it was mathematically shown that, with some very reasonable assumptions, the LR is the only way of assessing the uncertainty inherited in evidential evaluation (Aitken, 2018; Aitken and Taroni, 2004; Good, 1991). It is becoming recognised as the logical and legally correct framework for both analysing forensic evidence and presenting it in court (Balding, 2005; Evett et al., 1998; Marquis et al., 2011; Morrison, 2009; Neumann et al., 2007). Yet, some argue that the LR is one possible tool for communicating to decision makers (Lund and Iyer, 2017: 1). Although forensic text comparison (FTC) currently lags behind other forensic sciences, some studies have demonstrated that linguistic text evidence can be properly analysed using the LR framework (Ishihara, 2014, 2017a, 2017b).

In the LR framework, instead of assessing the probabilities of two competing hypotheses given the evidence, the probabilities of observing the evidence (E) are assessed given the hypotheses: the prosecution hypothesis (H_p) against the defence hypothesis (H_d) (Aitken and Stoney, 1991; Aitken and Taroni, 2004; Robertson et al., 2016). Therefore, the LR can be defined as in Equation (1).

$$LR=\frac{p(E|H_p)}{p(E|H_d)} \quad (1)$$

In the case of FTC, the LR is the ratio between the two conditional probabilities of the measured difference (considered the evidence E) between the source-known texts (i.e., from the suspect) and the source-questioned texts (i.e., from the offender): one represents the probability of the evidence if they had been produced by the same author (H_p), and the other represents the probability of observing the same evidence if they had originated from different authors (H_d).

Thus, the evidence E, which is the measured difference between two texts (x,y) can be expressed as $\Delta(x,y)$. A bag-of-words model is used to represent each text in this study. Thus x and y stand for the vectors of relative word frequencies ($w_i^j,\ i \in \{1 \cdots N\}, j \in \{x, y\}$) of the texts to be compared ($x= \{w_1^x, w_2^x \cdots w_N^x\}$ and $y= \{w_1^y, w_2^y \cdots w_N^y\}$).

Thus, Equation (1) can be rewritten as Equation (2), where f denotes a probability density function.

$$LR = \frac{f(\Delta(x,y)|H_p)}{f(\Delta(x,y)|H_d)} = \frac{f(\Delta(\{w_1^x, w_2^x \cdots w_N^x\},\{w_1^y, w_2^y \cdots w_N^y\})|H_p)}{f(\Delta(\{w_1^x, w_2^x \cdots w_N^x\},\{w_1^y, w_2^y \cdots w_N^y\})|H_d)} \quad (2)$$

The probability density functions under H_p and H_d need to be trained from a data set of scores.

Once a forensic scientist has estimated the LR as the weight of the evidence, the LR is then interpreted as a multiplicative factor by which the Bayesian theorem is used to update the prior odds (the factfinder's prior beliefs about the hypotheses) to the posterior odds (the factfinder's beliefs after observing the evidence). The factfinder (e.g., jury or judge) is thus responsible for quantifying the prior odds of the hypotheses, and the forensic scientist is responsible for estimating the LR. That is, the ultimate decision of a case (i.e., guilty or not guilty) is determined by the factfinder, who must update the prior odds to the posterior odds with the LR.

In this study, LRs are estimated using a score-based approach that has been extensively studied with several evidence types (Bolck et al., 2015; Hepler et al., 2012; Ramos et al., 2017). An alternative to the score-based approach is the feature-based approach, which has been applied to authorship text evidence (Ishihara, 2014). In score-based approaches, the likelihood of the score—which is usually quantified as a similarity/difference or a distance between paired samples that can be represented in the form of feature vector—is assessed against the probabilistic distributions from the same-source and different-source scores. This process is called score-to-LR conversion. The conversion model must be constructed with relevant training data; naturally, the more the data, the more accurately the system can perform.

The types and conditions of the linguistic evidence used in criminal cases are all unique. It is often the case that relevant data for the case must be collected in a customised manner from scratch to train the score-to-LR conversion model. However, forensic scientists usually cannot afford to collect such a large number of data. Therefore, it is crucial that forensic scientists know how the FTC system's performance is influenced by the number of data. For this purpose, a series of experiments was conducted with the data that were synthesised by a Monte-Carlo simulation technique.

2 Experiment Design

Two sets of experiments were conducted, with the first set aiming to identify the conditions under which the FTC system optimally performs (see Section 3.1).

In the second set, with the best-performing conditions set, the FTC system's performance is assessed by altering the data number for training the score-to-LR conversion model (see Section 3.2). The database, pre-processing of data, logistic–regression calibration and assessment metrics are also discussed in this section.

2.1 Database

The current study used a portion of the Amazon Product Data Authorship Verification Corpus [1] (Halvani et al., 2017), which contained 21,534 product reviews from 3,228 reviewers. The review texts were equalised to be approximately 4kB in size, which corresponds to approximately 750 words in length. The reviewers contributed multiple product reviews for Amazon, but only those who produced six or more reviews were selected from the corpus, resulting in 2,160 reviewers. Only the first six reviews of each reviewer were selected for the two sets of experiments.

To compare a source-questioned (offender) sample and a source-known (suspect) sample, the six reviews were first separated into two groups: the first three and the last three, from which three documents that differed in word length (750, 1,500 and 2,250 words) were created by concatenating them. The first review text of each group was used as it originally appeared (i.e., as a document of 750 words). The first and second texts were also concatenated into a document of 1,500 words. All three texts were then combined into a document of 2,250 words. Documents of different word lengths were prepared for testing the correlation between the number of words and the system's performance.

2.2 Database Partition

The entire database was divided into the three mutually exclusive sub-databases of 'test', 'back-

[1] Available at http://bit.ly/1OjFRhJ.

ground' and 'development', each of which comprised documents from 720 authors (=2,160/3). The documents from the test database were used to assess the system's performance by generating same-author (SA) and different-author (DA) comparisons. From the 720 authors from the test database, each of whom had two documents for each word length, 720 SA comparisons and 517,680 DA comparisons (${}_{720}C_2 \times 2$) were possible for each word length.

The documents from the background database were used to obtain SA and DA scores, which were in turn used to train the score-to-LR conversion model. The composition of the background database was identical in quantity to that of the test database. That is, 720 SA scores and 517,680 DA scores could be obtained from the background database.

The resultant LRs after the score-to-LR conversion may not have been calibrated due to various reasons. In this case, the uncalibrated LRs had to be converted to interpretable LRs through a process of calibration. A typical and robust model for the calibration procedure is logistic regression (Morrison, 2013), and the development database was used to train the logistic regression. A more detailed explanation for logistic–regression calibration is provided in Section 2.4.

2.3 Tokenisation and a Bag-of-Words Model

Documents were tokenised with the tokens() function in the quanteda library (Benoit et al., 2018) of the R statistical package in the default setting. That is, all characters were changed to lower case and punctuation marks were *not* removed; the punctuation marks were thus considered single-word tokens. No stemming algorithm was applied.

The 420 most frequent words appearing in the entire dataset were selected as components for the bag-of-words model. The relative frequencies of the words in the model were then calculated for each document. These relative frequencies were used instead of word counts because the length of each document varied. The word frequencies of the bag-of-words vector were z-score normalised to equalise the amount of information across the words in the vector. If this step was not taken, then the information that was encoded in the frequently occurring words would substantially and unevenly influence the outcomes of the experiments, as word frequencies follow the distribution described by Zipf's law (Zipf, 1932).

2.4 Logistic–Regression Calibration

The LRs that are estimated using the score-based approach are usually well calibrated; they can thus be interpreted as the weight of evidence. As will be reported in Section 4, LRs become less calibrated when the background data are limited.

Figure 1 contains two Tippett plots which show the magnitude of the LRs derived from a simulation under a specific experimental condition (randomly generated scores from 10 authors for 2,250 words). Tippett plots show the cumulative proportion of the LRs of the SA comparisons, which are plotted rising from the left, as well as of the LRs of the DA comparisons, plotted rising from the right. For the Tippett plots, the cumulative proportion of trails is plotted on the Y-axis against the $\log_{10}$ LRs on the X-axis. The intersection of the two curves is the equal error rate (EER) which indicates the operating point at which the miss and false alarm rates are equal. As can be seen from Figure 1a, the intersection of the two curves is not aligned with $\log_{10}LR=0$. That means, the derived LRs are not well calibrated; thus they cannot be interpreted as the weight of evidence.

These uncalibrated LRs must be converted to calibrated LRs to be interpreted as the weight of evidence. A logistic–regression calibration (Brümmer and du Preez, 2006) is employed for this purpose. Logistic-regression calibration is operated by applying linear shifting and scaling to the uncalibrated LRs, in the log odds space, relative to a decision boundary; its aim is to minimise the magnitude and incidence of uncalibrated LRs that are known to misleadingly support the incorrect hypothesis, and also to maximise the values of uncalibrated LRs correctly supporting the hypotheses. A logistic-regression line, the weights of which are estimated on the basis of the LRs derived from a training database, is used to monotonically shift and scale the uncalibrated LRs to the calibrated LRs. By way of exemplification, assuming a logistic-regression line of the type $y=ax+b$ (where x is the uncalibrated LR and y is the calibrated LR, and the weights, a and b, are estimated on the basis of the (uncalibrated) LRs derived from the development database), the formula $y=ax+b$ is used to shift by the amount of b, and scale by the amount of a, the uncalibrated LRs to the calibrated LRs. The LRs presented in Figure 1b are the outcome of the application of logistic-regression calibration to the LRs given in Figure 1a.

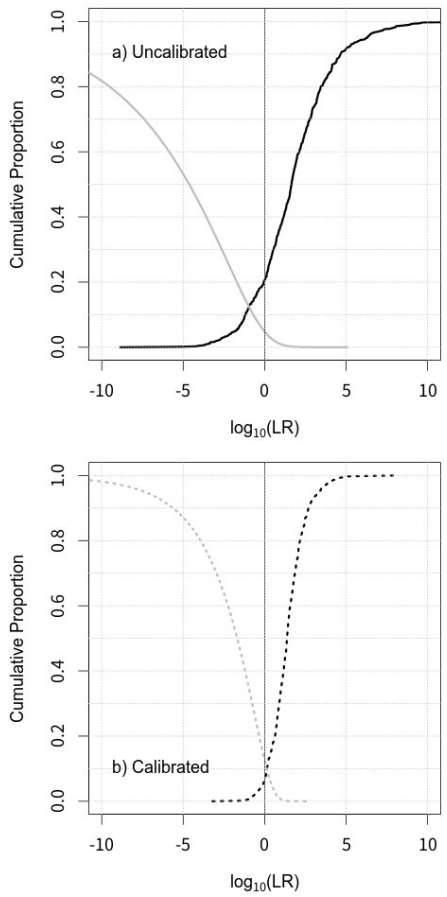

Figure 1: Example Tippett plots showing uncalibrated (Panel a) and calibrated (b) LRs. Black=SA LRs; Grey=DA LRs; Solid curves=uncalibrated LRs; Dotted curves=calibrated LRs.

2.5 Performance Evaluation

It is common to assess the performance of any identification or classification system based on its accuracy and error rates. However, accuracy and error rates are binary and categorical (e.g., correct or incorrect); this is not suitable for the nature of LR, which is gradient and continuous.

A more appropriate metric for assessing LR-based systems is arguably the log–LR cost (C_{llr}) (Brümmer and du Preez, 2006), which was originally developed for LR-based automatic speaker recognition systems. C_{llr} can be obtained through Equation (3).

$$C_{llr} = \frac{1}{2}\left(\begin{array}{c} \left[\frac{1}{N_{SA}}\sum_{i}^{N_{SA}} log_2\left(1+\frac{1}{LR_i}\right)\right] \\ +\left[\frac{1}{N_{DA}}\sum_{j}^{N_{DA}} log_2\left(1+LR_j\right)\right] \end{array} \right) \quad (3)$$

N_{SA} and N_{DA} refer to the number of SA and DA comparisons, respectively. LR_i and LR_j refer to the linear LRs that are derived from these SA and DA comparisons. In this metric, all LRs (except ±infinity) are attributed penalties in proportion to their magnitudes, with the LRs that support the counter-factual hypotheses being more severely penalised. The C_{llr} is based on information theory, and if the C_{llr} value is higher than one, then the system is performing worse than not utilising the evidence at all.

The C_{llr} is a metric that assesses a system's overall validity. It comprises two components: discrimination loss (C_{llr}^{min}) and calibration loss (C_{llr}^{cal}): $C_{llr}=C_{llr}^{min}+C_{llr}^{cal}$. The C_{llr}^{min} is a theoretical minimum C_{llr} value that can be obtained through pool adjacent violators algorithms (Brümmer and du Preez, 2006).

3 Experiments

3.1 Preparatory Experiments and Outcomes

A series of FTC experiments was conducted with a score-based LR approach to identify under what conditions the system would yield the best outcome. In these experiments, scores were measured with Cosine distance, with the bag-of-words model consisting of N most frequent words. The scores were then converted to their LRs based on the conversion model that was trained by the scores calculated from the SA and DA comparisons, which were compiled from the background database. The size (N) of the bag-of-words vector is incremented from N=20 to N=420 by 20 to identify the best-performing N. The Normal, Log-Normal, Weibull and Gamma models were tried as possible conversion models, but only the model that fit the data best in terms of the Akaike information criterion (AIC) (Akaike, 1974) was selected for each experiment (separately for the SA and DA models). Cosine distance was used because of its superior performance to other measures (Evert et al., 2017; Smith and Aldridge, 2011).

The C_{llr} values are plotted as a function of the feature number (N) in Figure 2, separately for 750, 1,500 and 2,250 words. Regardless of the word length, the system performed best with N=260. The overall trend for the C_{llr} trajectory is similar across the word lengths, revealing a relatively large improvement in performance as the N increased from 20 to 120 and the C_{llr} values started converging towards N=260. After N=260, the performance remained relatively unchanged, indicating that the inclusion of less-frequent words did not contribute to the improvement.

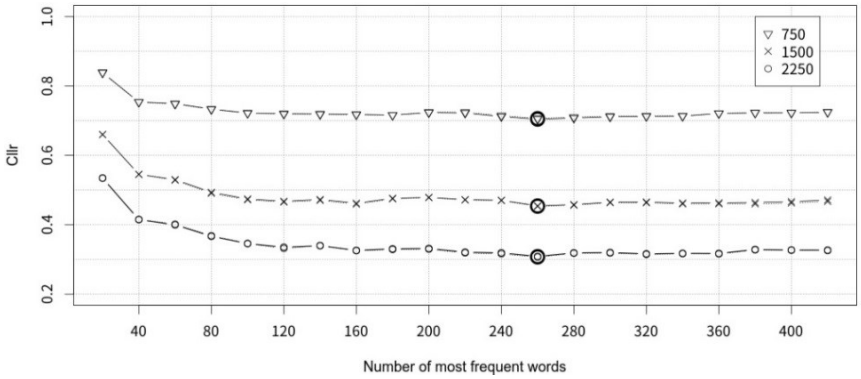

Figure 2: C_{llr} values plotted as a function of the number of features, separately for the word lengths of 750, 1,500 and 2,250. The large circles indicate the best C_{llr}.

The best-fitted models when N=260 are outlined in Table 1 and are used for the Monte-Carlo simulation.

	SA scores	DA scores
750	Weibull	Weibull
1500	Weibull	Normal
2250	Weibull	Normal

Table 1: Best-fitted parametric models for the SA and DA scores.

3.2 Experiments with the Monte-Carlo Simulation

In the preparatory experiment, the score-to-LR conversion models were trained with the data in the background database, which comprised texts written by 720 authors. Using the model as the basis, the scores of X number of authors (X=[5, 10, 20, 30, 40, 60, 80, …, 720]) were randomly generated 20 times to build the conversion models.

The Normal, Log-Normal, Weibull and Gamma parametric models were fitted to the scores that were randomly generated separately for SA and DA comparisons in the maximum likelihood estimation method. The best-fitted model was chosen according to its AIC values.

Figure 3 illustrates the simulation process for the length of 750 words. Out of the texts written by 720 authors from the background database, 720 SA and 517,680 DA scores were estimated. These scores are plotted as histograms: the white histogram represents SA and the grey histogram represents DA. Their fitted models (Weibull) are presented as solid red and blue curves, respectively. From these two models, the scores for the SA and DA comparisons—which are possible from 30 authors (i.e., 30 SA and 870 SA scores)—were randomly generated 20 times. Their models are represented by thin black curves. These models were used for the score-to-LR conversion.

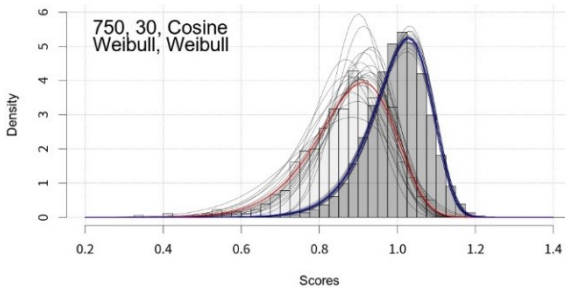

Figure 3: Illustration of a Monte-Carlo simulation with the base SA and DA scores, of which the histograms are white and grey, respectively. The red and blue curves are models of the SA and DA scores, respectively. The thin lines represent the models of the 20 sets of randomly generated scores from 30 authors.

4 Results and Discussions

The boxplots presented in Figure 4 reveal the degree of fluctuations in the C_{llr} values of the 20 simulations; they also indicate how the C_{llr} values converge as the number of authors increases.

Regardless of the word length, the FTC system's performance substantially fluctuates when the background database only comprises the text samples from 5~10 authors; that is, the performance is not stable. However, this instability quickly recovers if the text samples are collected from 20 or more authors. This is a positive finding in terms of FTC's practical application, as forensic scientists cannot afford the time and money required to collect a large number of data that are relevant to each case if they cannot find an already-existing dataset that is suitable to the case.

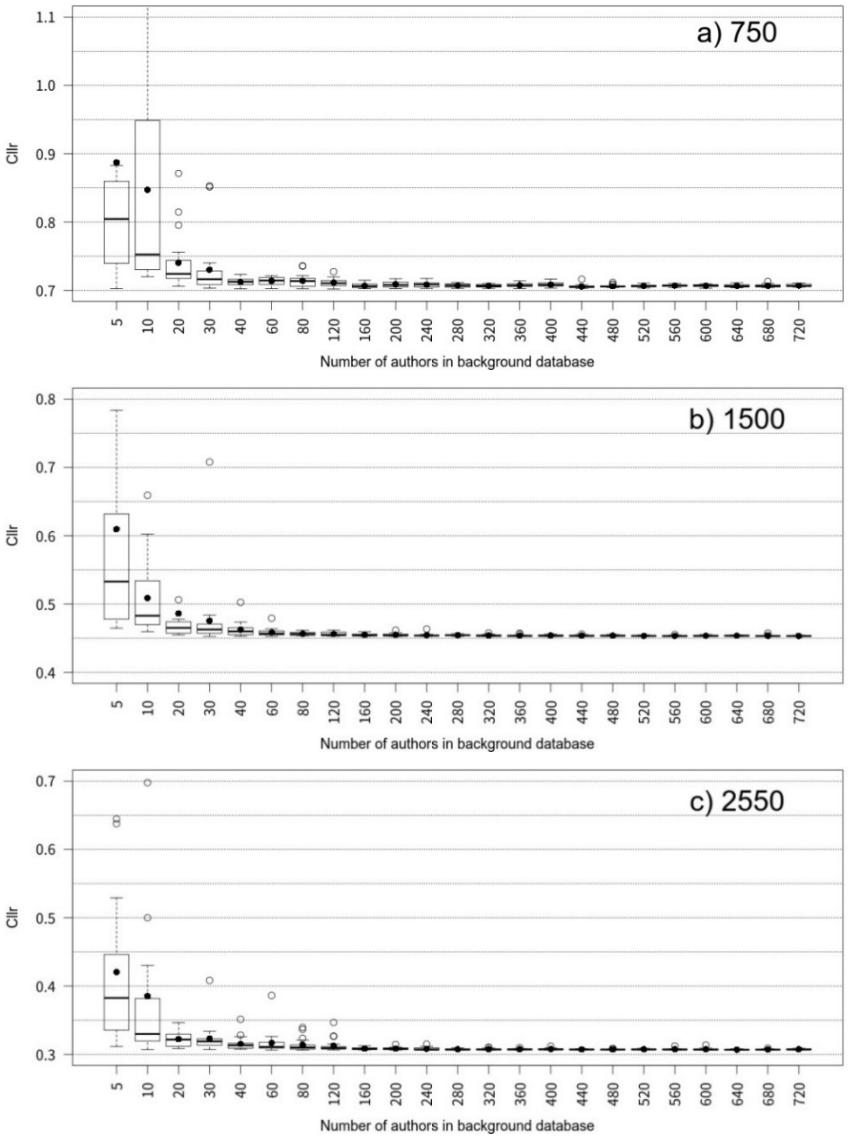

Figure 4: Boxplots displaying the degree of fluctuation in C_{llr} values as a function of the size of the background database. Black circles indicate the mean C_{llr} values for each size of the background database.

It is evident from Figure 4 (black circles) that the system's overall performance improves exponentially from N=5 to N=40, resulting in the outcome in which the performance with N=40 is nearly compatible with its performance with N=720.

To further investigate the reasons underlying the fluctuations in performance (especially with the small number of N), the C_{llr}^{min} and C_{llr}^{cal} values (discrimination loss and calibration loss, respectively) are plotted separately in Figures 5 and 6, respectively. They are presented in the same manner as Figure 4. As can be observed in Figure 5, being apart from the word length of 750, the system's discriminability is highly stable, even with small Ns. Specifically, regarding the word length of 2,250, Figure 5c reveals that the C_{llr}^{min} values are constant and far less fluctuated, as they are not affected by the number of authors in the background database. That is, in terms of discrimination performance, when many words (e.g., 1,500 and 2,550 words) are available, the system is robust and stable against a small background population size.

In contrast, Figure 6 indicates that the C_{llr}^{cal} values exhibit a highly similar trend to that of the C_{llr} values that are plotted in Figure 4—in that, a great variability in the C_{llr}^{cal} values is observed when the number of authors is small (e.g., N=5~10); however, this variability begins converging rapidly with more authors. This signifies that the C_{llr}^{cal} values also demonstrate a quick recovery with more authors. The observations drawn from Figures 5 and 6 reveal that the poor performance associated with a small number of authors (N=5~10), as indicated by the C_{llr} values from Figure 4, is not due to the system's poor discriminability, but due to poor calibration.

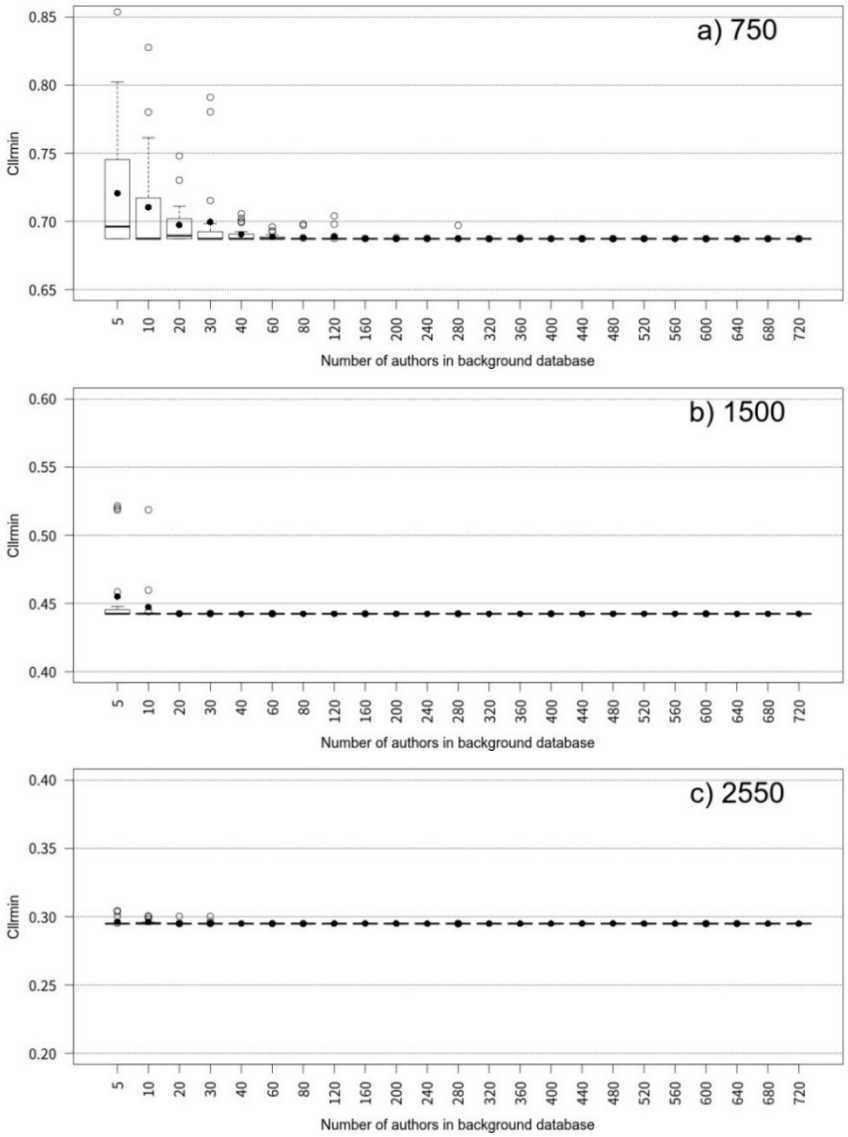

Figure 5: Boxplots displaying the degree of fluctuation in C_{llr}^{min} values as a function of the size of the background database. Black circles indicate the mean C_{llr}^{min} values for each size of the background database.

Following this interpretation, logistic–regression calibration was applied to all LRs, in which a gain in overall performance was expected. The C_{llr} values of the calibrated LRs are again plotted as boxplots in Figure 7. It is apparent from Figure 7 that the system's performance has noticeably improved in both stability and accuracy; the degree of fluctuations in the C_{llr} values is lessened and the mean C_{llr} values are lower, even with small *N*s.

Ishihara (2016) previously investigated how background population size affected the performance of an LR-based FTC system. In the experiments, the LRs were estimated using the multivariate kernel density (MVKD) LR formula (Aitken and Lucy, 2004), with two to eight stylometric features. Texts collected from 140 authors were used to extract necessary statistical information for a Monte-Carlo simulation, for which a mixture Gaussian model was used. The MVKD is a type of feature-based approach for estimating LRs. The population size was incremented by 10 from 10 authors to 140 authors.

Although a direct comparison between the current study and Ishihara's (2016) study cannot be validly made, some noticeable differences can still be highlighted. The number of features (2~8) used in Ishihara's study was far smaller than that of the current study (260), and Ishihara reported a great improvement in C_{llr} (from 10 to 50~60 authors), after which a small but continuous improvement could be observed with more authors. He also reported a relatively high variability in C_{llr}, even with a large number of authors (e.g., 140).

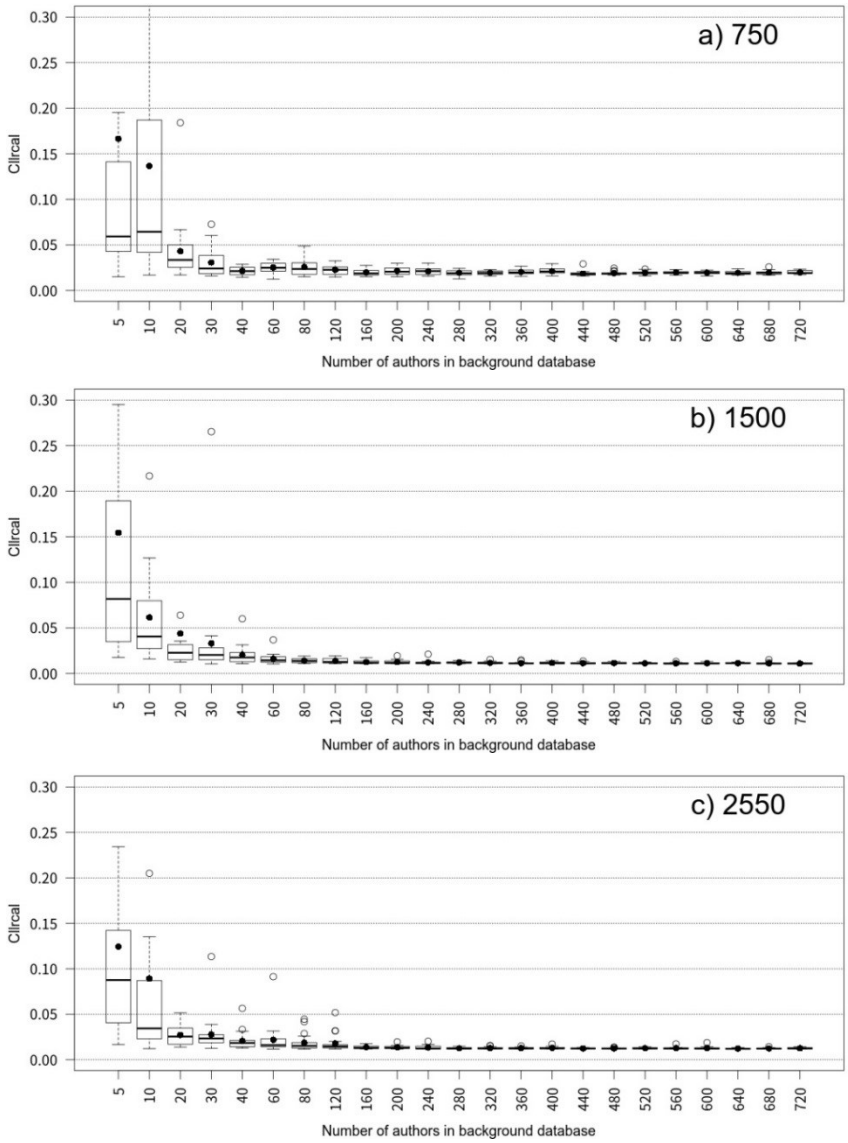

Figure 6: Boxplots showing the degree of fluctuation in C_{llr}^{cal} as a function of the size of the background database. Black circles indicate the mean C_{llr}^{cal} values for each size of the background database.

In light of these comparative observations, the FTC system's performance appears to reach its optimum with a smaller population size for the score-based approach rather than for the feature-based approach. Further, the fluctuation in performance also begins converging with a lesser number of background data for the score-based approach than for the feature-based approach. The relative robustness of the score-based approach that the current study revealed for linguistic text evidence aligns with the findings in previous studies regarding other types of evidence (Aitken, 2018; Bolck et al., 2015). However, the difference in performance between the score- and feature-based approaches must be further investigated under mutually comparable conditions.

Based on Figure 7, it can be concluded that logistic–regression calibration leads to an improvement in terms of the system's stability and validity. For training the logistic–regression weights, the development database that comprised the texts from 720 authors was employed. It is evident that the calibration performance also mainly relies on the quantity of the data in the development database. The positive outcome after applying the calibration is likely attributable to the amount of data in the development database. Therefore, it is pertinent to analyse how the development database's size influences the FTC system's performance, as the application of calibration appears to be essential when the background database is substantially small in number (e.g., 5~10 authors).

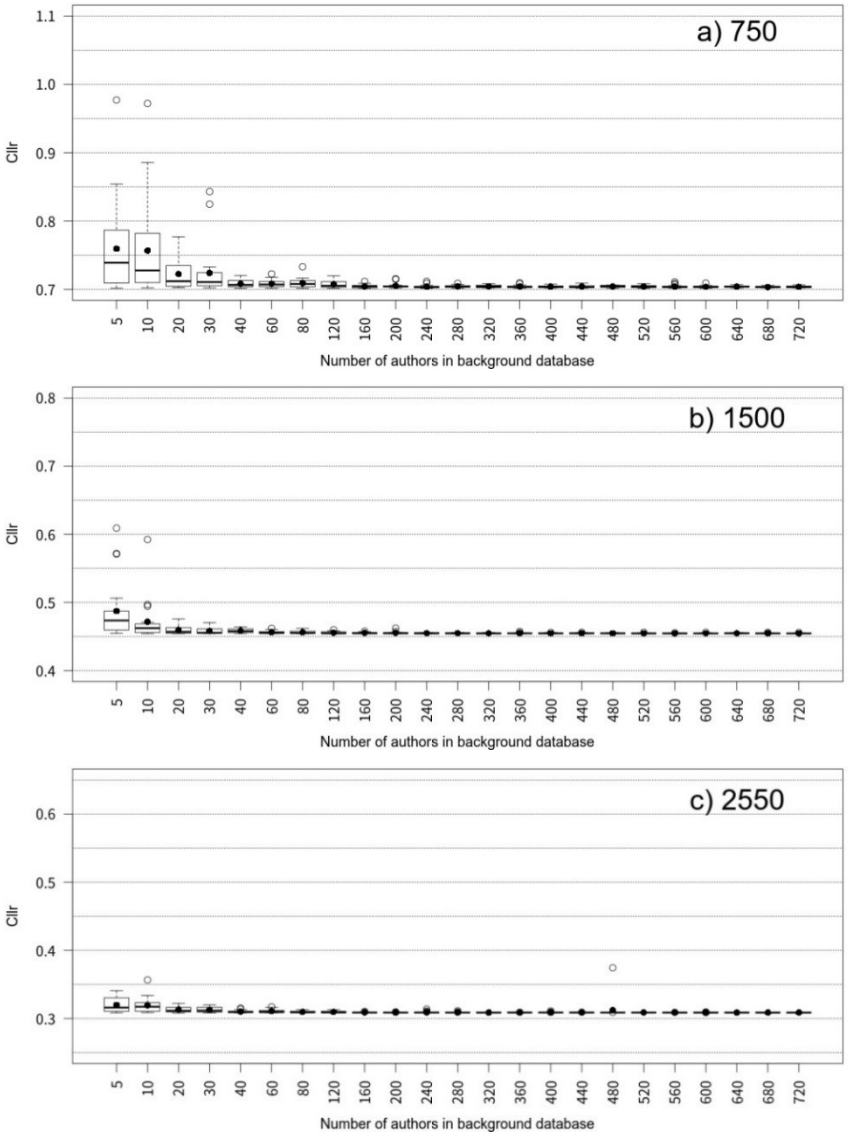

Figure 7: Boxplots revealing the fluctuation of C_{llr} after logistic–regression calibration.

5 Conclusion and Further Study

The robustness and stability of a score-based LR FTC system with a bag-of-words model were investigated with different numbers of background population data, which were synthesised by a Monte-Carol simulation. The experiments' results revealed that the score-based FTC system is fairly robust and stable in performance against the limited number of background population data. For example, with 40~60 authors, the performance is both nearly compatible and as stable as with 720 authors. This is a beneficial finding for FTC practitioners. Additionally, the instability and suboptimal performance observed in terms of C_{llr} with a small number of data (e.g., 5~20 authors) were mainly attributed to poor calibration (i.e., the derived LRs were not calibrated) rather than to the poor discriminability potential.

A comparison with the outcomes of previous studies indicates that the score-based approach may be more robust against a limited number of background population data than a feature-based approach; however, this point warrants further study.

Acknowledgements

The author thanks the reviewers for their valuable comments.

References

Aitken, C. G. G. (2018) Bayesian hierarchical random effects models in forensic science. *Frontier in Genetics* *9*(Article 126): 1-14. https://doi.org/10.3389/fgene.2018.00126

Aitken, C. G. G. and Lucy, D. (2004) Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* *53*(1): 109-122. https://dx.doi.org/10.1046/j.0035-9254.2003.05271.x

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

Aitken, C. G. G. and Taroni, F. (2004) *Statistics and the Evaluation of Evidence for Forensic Scientists*. Chichester: John Wiley & Sons.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* *19*(6): 716-723. https://dx.doi.org/10.1109/TAC.1974.1100705

Balding, D. J. (2005) *Weight-of-Evidence for Forensic DNA Profiles*. Hoboken: John Wiley & Sons.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S. and Matsuo, A. (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

Bolck, A., Ni, H. F. and Lopatka, M. (2015) Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: Applied to forensic MDMA comparison. *Law, Probability and Risk* *14*(3): 243-266. https://dx.doi.org/10.1093/lpr/mgv009

Brümmer, N. and du Preez, J. (2006) Application-independent evaluation of speaker detection. *Computer Speech and Language* *20*(2-3): 230-275. https://dx.doi.org/10.1016/j.csl.2005.08.001

Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C. and Vitt, T. (2017) Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities* *32*(suppl_2): ii4-ii16. https://doi.org/10.1093/llc/fqx023

Evett, I. W., Lambert, J. A. and Buckleton, J. S. (1998) A Bayesian approach to interpreting footwear marks in forensic casework. *Science & Justice* *38*(4): 241-247. https://dx.doi.org/10.1016/S1355-0306(98)72118-5

Good, I. J. (1991) Weight of evidence and the Bayesian likelihood ratio. In C. G. G. Aitken and D. A. Stoney (eds.), *The Use of Statistics in Forensic Science* 85-106. Chichester: Ellis Horwood.

Halvani, O., Winter, C. and Graner, L. (2017). Authorship verification based on compression-models. *arXiv preprint arXiv:1706.00516*. Retrieved on 25 June 2020 from http://arxiv.org/abs/1706.00516

Hepler, A. B., Saunders, C. P., Davis, L. J. and Buscaglia, J. (2012) Score-based likelihood ratios for handwriting evidence. *Forensic Science International* *219*(1-3): 129-140. http://dx.doi.org/10.1016/j.forsciint.2011.12.009

Ishihara, S. (2014) A likelihood ratio-based evaluation of strength of authorship attribution evidence in SMS messages using N-grams. *International Journal of Speech Language and the Law 21*(1): 23-50. http://dx.doi.org/10.1558/ijsll.v21i1.23

Ishihara, S. (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 T. Cohn (ed.), *Proceedings of Proceedings of the Australasian Language Technology Association Workshop 2016*: 113-121.

Ishihara, S. (2017a) Strength of forensic text comparison evidence from stylometric features: A multivariate likelihood ratio-based analysis. *The International Journal of Speech, Language and the Law* *24*(1): 67-98. https://doi.org/10.1558/ijsll.30305

Ishihara, S. (2017b) Strength of linguistic text evidence: A fused forensic text comparison system. *Forensic Science International 278*: 184-197. https://doi.org/10.1016/j.forsciint.2017.06.040

Lund, S. P. and Iyer, H. (2017) Likelihood ratio as weight of forensic evidence: A closer look. *Journal of Research of the National Institute of Standards and Technology* *122*(Article 27): 1-32. https://doi.org/10.6028/jres.122.027

Marquis, R., Bozza, S., Schmittbuhl, M. and Taroni, F. (2011) Handwriting evidence evaluation based on the shape of characters: Application of multivariate likelihood ratios. *Journal of Forensic Sciences* *56*(Suppl_1): S238-242. https://dx.doi.org/10.1111/j.1556-4029.2010.01602.x

Morrison, G. S. (2009) Forensic voice comparison and the paradigm shift. *Science & Justice 49*(4): 298-308. https://dx.doi.org/10.1016/j.scijus.2009.09.002

Morrison, G. S. (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

Neumann, C., Champod, C., Puch-Solis, R., Egli, N., Anthonioz, A. and Bromage-Griffiths, A. (2007) Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. *Journal of Forensic Science 52*(1): 54-64. https://dx.doi.org/10.1111/j.1556-4029.2006.00327.x

Ramos, D., Krish, R. P., Fierrez, J. and Meuwly, D. (2017) From biometric scores to forensic likelihood ratios. In M. Tistarelli and C. Champod (eds.), *Handbook of Biometrics for Forensic Science* 305-327. Cham: Springer.

Robertson, B., Vignaux, G. A. and Berger, C. E. H. (2016) *Interpreting Evidence: Evaluating Forensic Science in the Courtroom* (2nd ed.). Chichester: John Wiley and Sons, Inc.

Smith, P. W. H. and Aldridge, W. (2011) Improving authorship attribution: Optimizing Burrows' Delta method. *Journal of Quantitative Linguistics 18*(1): 63-88. https://dx.doi.org/10.1080/09296174.2011.533591

Zipf, G. K. (1932) *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge: Harvard University Press.