Estimating the Strength of Authorship Evidence with a Deep-Learning-Based Approach

Shunichi Ishihara[1], Satoru Tsuge[2], Mitsuyuki Inaba[3], Wataru Zaitsu[4]
shunichi.ishihara@anu.edu.au, tsuge@daido-it.ac.jp, inabam@sps.ritsumei.ac.jp, w.zaitsu@mejiro.ac.jp

[1] Speech and Language Laboratory, The Australian National University, Canberra, Australia
[2] Department of Information Systems, Daido University, Aichi, Japan
[3] College of Policy Science, Ritsumeikan University, Kyoto, Japan
[4] Department of Psychological Counselling, Mejiro University, Tokyo, Japan

Abstract

This study is the first likelihood ratio (LR)-based forensic text comparison study in which each text is mapped onto an embedding vector using RoBERTa as the pre-trained model. The scores obtained with Cosine distance and probabilistic linear discriminant analysis (PLDA) were calibrated to LRs with logistic regression; the quality of the LRs was assessed by log LR cost (C_{llr}). Although the documents in the experiments were very short (maximum 100 words), the systems reached the C_{llr} values of 0.55595 and 0.71591 for the Cosine and PLDA systems, respectively. The effectiveness of deep-learning-based text representation is discussed by comparing the results of the current study to those of the previous studies of systems based on conventional feature engineering tested with longer documents.

1 Introduction

In forensic science, the likelihood ratio (LR) framework has long been considered the logically and legally correct approach to interpreting the analysis of forensic evidence (Aitken and Stoney, 1991; Aitken and Taroni, 2004; Morrison, 2022; Robertson et al., 2016). The LR framework is standard in DNA typing. The community of forensic text comparison (FTC), commonly known as forensic authorship verification, recently recognised the importance of this framework (Grant, 2022). Despite the importance of the LR framework in forensic science, LR-based studies on textual evidence are still conspicuously rare (Ishihara, 2017, 2021; Ishihara and Carne, 2022).

Many studies claim to be forensic but treat the problem as a usual authorship verification problem. However, there are important differences between conventional and forensic authorship verification. Conventional authorship verification aims to answer a verification problem. Forensic authorship verification aims to assist the fact finder in concluding the case, not answering the problem. Legally, giving an answer to a verification problem (even in a probabilistic term) equates to referring to the ultimate question of 'guilty vs. not guilty', which is only permitted for the fact finder. Logically, forensic scientists without all evidential information of the case cannot estimate the probability of a hypothesis from incomplete evidence. Thus, they cannot logically refer to the ultimate question. However, forensic scientists can logically and legally estimate the strength of evidence via LR (Aitken and Stoney, 1991; Aitken and Taroni, 2004; Robertson et al., 2016).

LR is given in Equation (1). LR is the ratio of two conditional probabilities; one is the probability of evidence (E) given the prosecution hypothesis (H_p) and the other is the probability of the same evidence given the defence hypothesis (H_d).

$$LR = \frac{P(E|H_p)}{P(E|H_d)} \tag{1}$$

The relative strength of the given evidence with respect to the competing hypotheses is reflected in the magnitude of the LR. The greater the LR value is than 1, the stronger support the evidence is considered to provide for the prosecution; the smaller the LR value is than 1, *mutatis mutandis*, for the defence hypothesis. It is very important to note that the LR is not a binary expression of truth.

With an LR estimated as the strength of evidence, the fact finder's belief regarding the hypotheses (quantified as prior odds) is raised to the posterior odds through the Bayesian theorem, as shown in Equation (2).

$$\underbrace{\frac{P(H_p|E)}{P(H_d|E)}}_{\text{posterior odds}} = \underbrace{\frac{P(H_p)}{P(H_d)}}_{\text{prior odds}} \times \underbrace{\frac{P(E|H_p)}{P(E|H_d)}}_{LR} \tag{2}$$

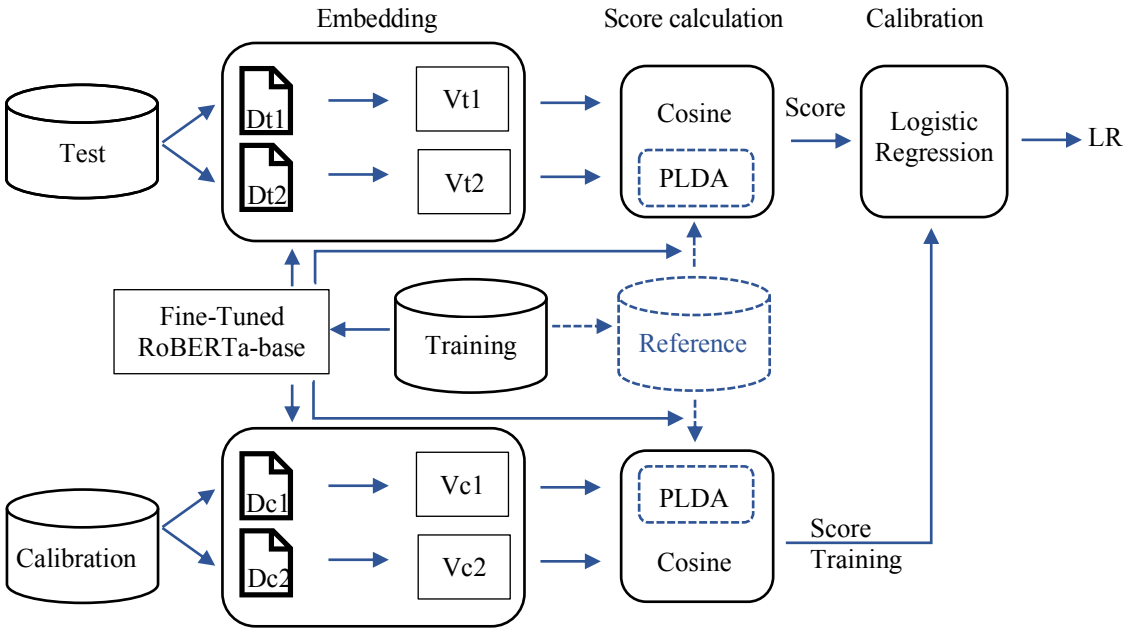

Figure 1: Process of estimating LRs. D{t,c} = (t)est or (c)alibration document; V{t,c} = vectorised (t)est or (c)alibration document; PLDA = probabilistic linear discrimination analysis.

The posterior odds are equivalent to the fact finder's belief regarding the hypotheses given the evidence

Despite the success of deep learning in many natural language processing tasks, a conventional machine learning approach with traditional feature engineering remains effective in authorship verification, particularly for small datasets (Kestemont et al., 2019; Kestemont et al., 2018). Nonetheless, deep-learning-based systems gradually started achieving better verification accuracy than conventional approaches, in particular with a large volume of data (Kestemont et al., 2021; Kestemont et al., 2020; Zhu and Jurgens, 2021). Despite of its clear presence, deep learning has no yet made inroads into the LR-based FTC. This preliminary study looks in the effectiveness of a deep-learning approach in LR-based FTC.

2 Methodology

2.1 Datasets

This study used the dataset of Amazon reviews prepared by Zhu and Jurgens (2021) with minor modifications. They filtered out reviews that are shorter than 50 tokens, and selected authors who contributed at least 5 reviews and at least in two product domains; there are 17 product domains. The text length did not exceed 100 tokens; i.e. $\text{max\_length} = 102$.

Table 1 shows the numbers of authors, same author (SA) and different author (DA) comparisons in each dataset. The former is the simulation of the H_p and the latter is that of the H_d. The training and development datasets were used as originally prepared by Zhu and Jurgens (2021). The original test dataset was evenly split into two: one half was used as the test, and the other was used as the calibration dataset.

Dataset	Author	SA	DA
Test	32,124	96,253	96,491
Training	51,398	148,845	149,389
Development	12,849	36,429	36,317
Calibration	32,124	96,253	96,491

Table 1: Numbers of authors and SA/DA comparisons for each dataset.

2.2 Embedding and Fine-Tuning

Stylistic embedding of each text was performed as described by Zhu and Jurgens (2021) and using their tools.[1] They demonstrated the superiority of their system to various deep-learning-based baseline systems.

Each text was mapped into an embedding vector (z) by merging the last hidden states ($= \{h_0, h_1, \cdots, h_n\}$) into a single embedding vector

[1] https://github.com/lingjzhu/idiolect

$(= h_o)$ by attention pooling. The underlying pre-trained model was RoBERTa (specifically `roberta-base` as the encoder) (Liu et al., 2019). The training was done using the proxy-anchor loss function (Kim et al., 2020) with $\alpha = 30$; $t_s = 0.6$; $t_d = 0.4$; $t_t = t_s + t_d/2$. It is a continuous approximation of the max-margin loss of which the additional parameter enables better control over the penalty magnitude for difficult comparisons. An embedding vector dimension is 768. The hyperparameter values for fine-tuning were set according to Zhu and Jurgens (2021). The batch size was set at 256. Adam optimiser was used with a learning rate of $1e^{-5}$. The models were set to train for five epochs, after which no further improvement in performance was observed.

2.3 Estimating Likelihood Ratios

Estimating LRs for a pair of documents in the form of an embedding vector is illustrated in Figure 1. It is a two-stage process comprising score calculation and calibration.

Two methods were tested for estimating a score for each comparison of documents. One method was based on Cosine distance and the other on probabilistic linear discriminant analysis (PLDA) (Prince and Elder, 2007). The PLDA model used in this study was a two-covariance model. Besides the information regarding the author's unique writing style (x), each embedding vector (h_o) carries some residual noise (ε); for example, noise caused by thematic variations. Thus, h_o can be represented as Equation (3):

$$h_o = x + \varepsilon \tag{3}$$

A Gaussian generative model was assumed for the probability density function for x and ε, which requires a within-author and between-author covariance matrix, respectively. Authors were randomly selected from the training dataset to train the matrices (N = 10,000). A PLDA score was calculated using Equation (4), where z_i and z_j are embedding vectors under comparison.

$$score = \frac{P(z_i, z_j|H_p)}{P(z_i|H_d)P(z_j|H_d)} \tag{4}$$

The scores of the test dataset calculated through the two methods were converted to LRs at the calibration stage using logistic regression, the most common calibration approach for LR-based systems (Morrison, 2013; Ramos and Gonzalez-Rodriguez, 2013). The scores obtained from the calibration dataset were used to train the logistic regression.

2.4 Evaluation

Evaluation metrics based on classification or identification accuracy are not appropriate for assessing the performance of LR-based systems. Such metrics are inappropriate because (1) the category-based classification accuracy does not properly assess the magnitude of LRs (which is continuous), and (2) they implicitly refer to the accuracy of decision making, guilty vs. not guilty; only the fact finders (not forensic scientists or FTC experts) are legally permitted to refer to this ultimate question. The standard evaluation metric for LR-based systems is the log LR cost (C_{llr}) expressed in Equation (5):

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{SA}}\sum_{i}^{N_{SA}} \log_2\left(1 + \frac{1}{LR_{SA_i}}\right) + \frac{1}{N_{DA}}\sum_{j}^{N_{DA}} \log_2\left(1 + LR_{DA_j}\right)\right) \tag{5}$$

In Equation (5), N_{SA} and N_{DA} are the numbers of SA and DA comparisons, respectively. 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 costs, which were calculated for all LRs. The closer to $C_{llr} = 0$, the better the performance. If $C_{llr} \geq 1$, it denotes that the evidence is not informative for inference. With the pool-adjacent-violators algorithm, C_{llr} can be decomposed into C_{llr}^{min} and C_{llr}^{cal}, which assess the discrimination and calibration performance of the system, respectively. Thus, C_{llr}=C_{llr}^{min}+C_{llr}^{cal}. *EER* is also given for reference. A Tippett plot was used to visualise the magnitude of the derived LRs.

3 Results

The experimental results for the C_{llr}-based metrics are shown in Table 2.

	C_{llr}	C_{llr}^{min}	C_{llr}^{cal}	*EER*
Cosine	0.55595	0.55487	0.00108	0.17263
PLDA	0.71591	0.67159	0.04432	0.21855

Table 2: Experimental results.

Table 2 shows that the Cosine system outperforms the PLDA system in all metrics. The C_{llr}^{cal} values are close to zero, indicating that the

derived LRs are well-calibrated for both systems. The PLDA model probabilistically considers the between- and within-author variabilities. Theoretically, the model is expected to suit the authorship verification task. Therefore, it was expected to outperform the Cosine system. The contrary result could be due to the amount of data available for each document—100 words maximum. This finding warrants further study. The Cosine system has been reported as robust against adverse conditions, including the scarcity of data (Ishihara, 2021; Ishihara and Carne, 2022). The derived LRs were plotted as Tippett plots to observe their magnitudes (see Figure 2).

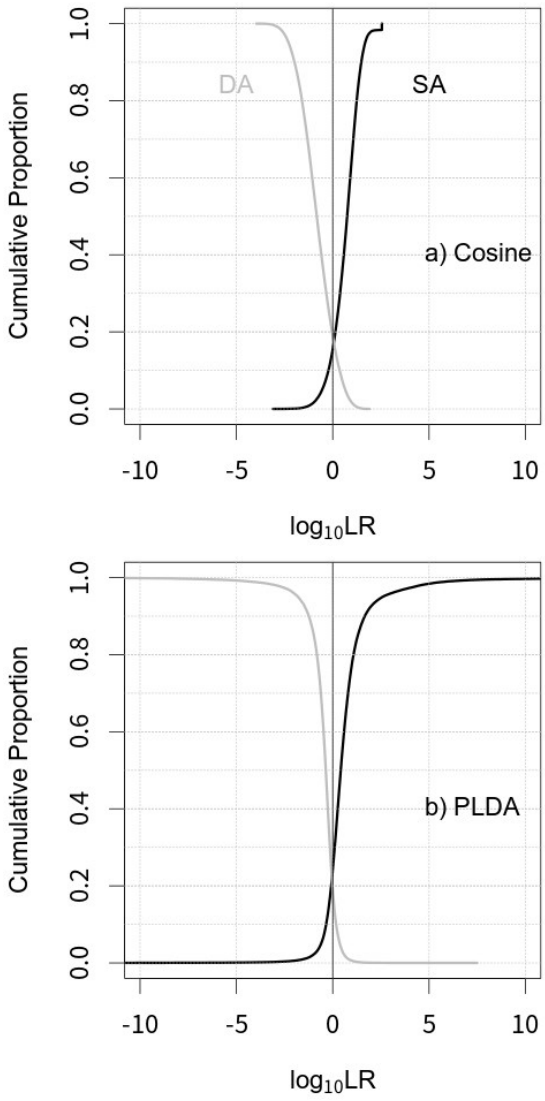

Figure 2: Tippett plots: Panel a) is for the Cosine system and Panel b) is for the PLDA system. The solid black curves = SA LRs and the solid grey curves = DA LRs.

The derived LRs from the Cosine system were conservative in magnitude; most LRs were within the range of the $\log_{10}$LR of ±2.5. Conversely, Figure 2b shows some excessively strong LRs of the PLDA system (e.g., greater than a $\log_{10}$LR of ±10). The strong contrary-to-fact LRs raise concerns. The excessively strong LR values both for the contrary-to-fact comparisons and the consistent-with-fact comparisons indicate the model's instability. Since each document only contains a maximum of 100 words, it is sensible not to have overly strong LRs.

Ishihara (2021) conducted LR-based FTC experiments by measuring the Cosine distance of documents modelled via word unigrams. The target documents were also product reviews for Amazon. Each document was approximately 4 kB in data (approximately 800 words in length)—considerably longer than the current study's (maximum 100 words). Ishihara reported a C_{llr} of 0.70640 as the optimal result. Ishihara's experiments were carried out with the test, reference and calibration datasets, each of which had 720 authors.

Despite the very short documents, the systems tested in this study achieved nearly the same level of performance (Cosine: C_{llr} = 0.55595; PLDA: C_{llr} = 0.71591) as Ishihara's (2021) system based on documents of approximately 800 words (C_{llr} = 0.70640). Although the experiments are not directly comparable, the effectiveness of the deep-learning-based text representation for estimating LRs can be conjectured.

4 Conclusions

In this study, the LRs were estimated by logistic regression calibrating the scores obtained through two systems: one based on Cosine distance and the other on the PLDA model. The documents were mapped on embedding vectors using RoBERTa as the pre-trained model, and the derived LRs were assessed with C_{llr}. Albeit the documents being very short, the systems reached the C_{llr} values of 0.55595 and 0.71591, respectively for the Cosine and PLDA systems. The effectiveness of the deep-learning-based text representation was discussed in comparison to the results of a previous study which was based on the system with conventional feature engineering and longer documents.

Acknowledgements

The authors thank the reviewers for their valuable comments.

References

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

C. G. G. Aitken F. Taroni. 2004. *Statistics and the Evaluation of Evidence for Forensic Scientists*. Chichester: John Wiley & Sons, Chichester, 2nd edition.

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

S. Ishihara. 2017. 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.

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.

M. Kestemont, E. Manjavacas, I. Markov, J. Bevendorff, M. Wiegmann, E. Stamatatos, … M. Potthast. 2021. Overview of the cross-domain authorship verification task at PAN 2021. In *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum*, pages 1–17.

M. Kestemont, E. Manjavacas, I. Markov, J. W. M. Bevendorff, E. Stamatatos, M. Potthast and B. Stein. 2020. Overview of the cross-domain authorship verification task at PAN 2020. In *Proceedings of the CLEF 2020 Conference and Labs of the Evaluation Forum*.

M. Kestemont, E. Stamatatos, E. Manjavacas, W. Daelemans, M. Potthast and B. Stein. 2019. Overview of the cross-domain authorship attribution task at PAN 2019. In *Proceedings of the CLEF 2019 Conference and Labs of the Evaluation Forum*, pages 1–15.

M. Kestemont, M. Tschuggnall, E. Stamatatos, W. Daelemans, G. Specht, B. Stein and M. Potthast. 2018. Overview of the author identification task at PAN-2018: Cross-domain authorship attribution and style change detection. In *Proceedings of the CLEF 2018 Conference and the Labs of the Evaluation Forum*, pages 1–25.

S. Kim, D. Kim, M. Cho and S. Kwak. 2020. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3238–3247.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi W. Chen, … Veselin Stoyanov. 2019. *RoBERTa: A robustly optimized BERT pretraining approach*. *Computing Research Repository*, arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692.

Geoffrey 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.

Geoffrey S. Morrison. 2022. Advancing a paradigm shift in evaluation of forensic evidence: The rise of forensic data science. *Forensic Science International: Synergy*, 5: 100270. https://doi.org/10.1016/j.fsisyn.2022.100270.

Simon J. D. Prince and James H. Elder. 2007. Probabilistic linear discriminant analysis for inferences about identity. In *Proceedings of the 2007 IEEE 11th International Conference on Computer Vision*, pages 1-8.

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*. Chichester: John Wiley & Sons, Chichester, 2nd edition.

Jian Zhu and David Jurgens. 2021. *Idiosyncratic but not arbitrary: Learning idiolects in online registers reveals distinctive yet consistent individual styles*. *Computing Research Repository*, arXiv:2109.03158. Version 3. https://arxiv.org/abs/2109.03158v3.