Feature-Based Forensic Text Comparison Using a Poisson Model for Likelihood Ratio Estimation

Michael Carne and Shunichi Ishihara
Speech and Language Laboratory
The Australian National University
michael.carne@anu.edu.au, shunichi.ishihara@anu.edu.au

Abstract

Score- and feature-based methods are the two main ones for estimating a forensic likelihood ratio (LR) quantifying the strength of evidence. In this forensic text comparison (FTC) study, a score-based method using the Cosine distance is compared with a feature-based method built on a Poisson model with texts collected from 2,157 authors. Distance measures (e.g. Burrows's Delta, Cosine distance) are a standard tool in authorship attribution studies. Thus, the implementation of a score-based method using a distance measure is naturally the first step for estimating LRs for textual evidence. However, textual data often violates the statistical assumptions underlying distance-based models. Furthermore, such models only assess the similarity, not the typicality, of the objects (i.e. documents) under comparison. A Poisson model is theoretically more appropriate than distance-based measures for authorship attribution, but it has never been tested with linguistic text evidence within the LR framework. The log-LR cost (C_{llr}) was used to assess the performance of the two methods. This study demonstrates that: (1) the feature-based method outperforms the score-based method by a C_{llr} value of ca. 0.09 under the best-performing settings and; (2) the performance of the feature-based method can be further improved by feature selection.

1 Introduction

The essential part of any source-detection task is to assess the similarity or difference between the objects or items under comparison. For this purpose, in stylometric studies too, various distance measures have been devised and tested, particularly in studies concerned with the authorship of text sources (Argamon, 2008; Burrows, 2002; Hoover, 2004a; Smith and Aldridge, 2011). Burrows's Delta (Burrows, 2002) is probably the most studied distance measure in stylometric studies, and its effectiveness and robustness have been demonstrated for a variety of texts from different genres and languages (AbdulRazzaq and Mustafa, 2014; Hoover, 2004b; Rybicki and Eder, 2011; Þorgeirsson, 2018). Since Burrows (2002), several variants, including, for example, those based on Euclidian distance, Cosine similarity and Mahalanobis distance have been proposed to better deal with the unique characteristics of linguistic texts, expecting to result in a better identification and discrimination performance (Argamon, 2008; Eder, 2015; Hoover, 2004b; Smith and Aldridge, 2011).

Similarity- and distance-based measures make some assumptions about the distribution of the underlying data. For example, a Laplace distribution is assumed by Burrows's Delta, which itself is based on Manhattan distance, and a normal distribution by the Euclidean and cosine distances. However, it is well known that stylometric features do not always conform to, for example, a normal distribution (Argamon, 2008; Jannidis et al., 2015). Moreover, a normal distribution is not theoretically appropriate for discrete count data (e.g. occurrences of function words) Figure 1 shows the distributions of the counts of three words ('a', 'not' and 'they'), sampled from the database used in the current study. Frequently-occurring words, such as 'a' (Figure 1a), tend to be normally distributed. However the distribution starts skewing positively for less-frequently-occurring words, such as 'not' (Figure 1b) and 'they' (Figure 1c). In order to fill this gap between the theoretical assumption arising from distance measures and the nature of textual data, a one-level Poisson model is used in this study.

In the 1990s, the success of DNA analysis and some important United States court rulings, estab-

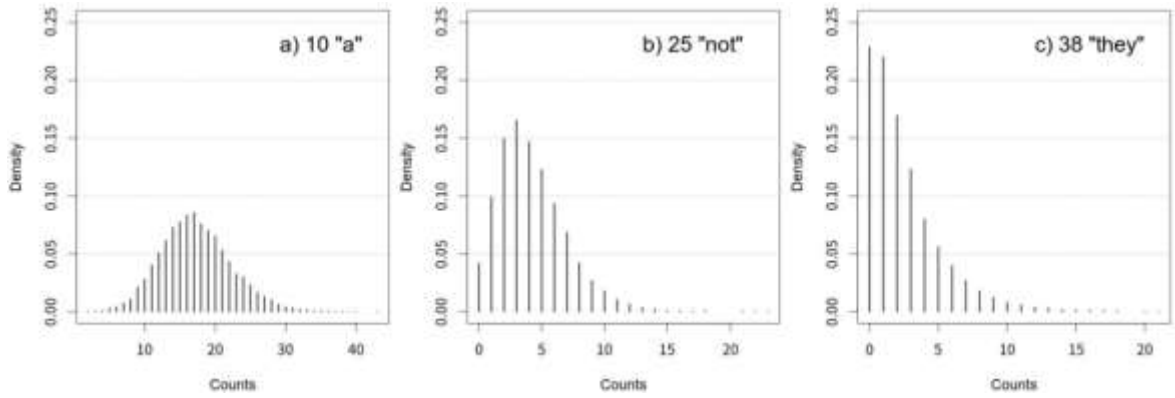

Figure 1: Histograms showing the distributional patterns of the counts of three words from the database; 'a', 'not' and 'they' for Panel a), b) and c), respectively. They are the 10th, 25th and 38th most frequently-occurring words in the database.

lishing the standard for expert evidence to be admitted in court, promoted the likelihood ratio (LR)-based approach as the standard for evaluating and presenting forensic evidence in court (Association of Forensic Science Providers, 2009). Although it is far less extensively studied than other areas of forensic science, it has been demonstrated that the LR framework can be applied successfully to linguistic textual evidence (Ishihara, 2014, 2017a, 2017b).

1.1 Previous Studies

There are two methods for deriving an LR model for forensic data, score- and feature-based. Each method has different strengths and shortcomings. The use of score-based methods is prevalent across different types of forensic evidence due to its robustness and ease of implementation relative to feature-based methods. The advantages and disadvantages of the methods are explained in §3.3 and §3.4.

Almost all previous LR studies, both feature- and score-based, use continuous data for LR estimation. Studies using feature-based LR models derived from probability distributions appropriate for discrete (or categorical) forensic features are rare.

To the best of our knowledge, Aitken and Gold (2013) and Bolck and Stamouli (2017) are the only two existing studies of this kind within the LR framework. Aitken and Gold (2013) propose a univariate discrete model for estimating LRs. They conducted only a small-scale experiment using limited data and features, which were used mainly for explanatory purposes.

Bolck and Stamouli (2017) investigate discrete multivariate models for estimating LRs using categorical data from gunshot residue. This study however uses a relatively low-dimensional feature space (only 12 features), and its modelling approach assumes independence between features. Text evidence however usually involves high-dimensional vector spaces and independence cannot be assumed, given correlation between features. The present study seeks to investigate these challenges in LR-based forensic text comparison (FTC) using discrete textual data in the form of counts of the *N* most frequently occurring words. It implements a feature-based LR model derived from the Poisson distribution, with logistic-regression fusion and calibration used as a means for dealing with correlation between features. This approach is compared to a score-based method using the cosine distance. To the best of our knowledge, this is the first FTC study to trial a feature-based method with a Poisson model in the LR framework.

2 Likelihood Ratio Framework

The LR framework has been proposed as a means of quantifying the weight of evidence for a variety of forensic evidence, including DNA (Evett and Weir, 1998), voice (Morrison et al., 2018; Rose, 2002), finger prints (Neumann et al., 2007), handwriting (Chen et al., 2018; Hepler et al., 2012), hair strands (Hoffmann, 1991), MDMA tablets (Bolck et al., 2009), evaporated gasoline residual (Vergeer et al., 2014) and earmarks (Champod et al., 2001). Collected forensic items from known- (e.g. a suspect's known text samples) and questioned-source (e.g. text samples from the offender) can be evaluated by estimating the LR under two competing hypotheses. One specifying the prosecution (or the same-author) hypothesis (H_p), and the other the defence (or the different-author) hypothesis (H_d).

These are expressed as a ratio of condition probabilities as shown in Equation 1).

$$LR = \frac{f(x, y|H_p)}{f(x, y|H_d)} \quad 1)$$

where x and y are feature values obtained from the known-source and questioned-source respectively. The relative strength of the evidence with respect to the competing hypotheses is reflected in the magnitude of the LR. The more the LR deviates from unity (LR = 1), the greater support for either the H_p (LR > 1) or the H_d (LR < 1).

The LR is concerned with the probability of evidence, given the hypothesis (either prosecution or defence), which is in concordance with the role of an expert witness in court, leaving the trier-of-fact to be concerned with the probability of either hypothesis, given the evidence.

3 Experiments

The two main approaches for estimating LRs, namely the score- and feature-based methods, will be implemented and their performance compared. After the database (§3.1) and the pre-processing and modelling techniques (§3.2) are introduced, the two methods are explained in §3.3 and §3.4, respectively, along with their pros and cons. Fusion/calibration techniques and performance metrics are described in §3.5 and §3.6, respectively.

3.1 Database

Data for the experiments were systematically selected from the Amazon Product Data Authorship Verification Corpus[1] (Halvani et al., 2017), which contains 21,534 product reviews posted by 3,228 reviewers on Amazon. Many of the reviewers contributed six or more reviews on different topics. Sizes of review texts are equalised to ca. 4 kB, which corresponds to approximately 700 words in length. From the corpus, the authors (= reviewers) who contributed more than six reviews longer than 700 words, were selected as the database for simulating offender vs. suspect comparisons. We decided on six reviews to maximise the number of same-author comparisons possible from the database. This resulted in 2,157 reviewers and a database containing a total of 12,942 review texts. Each review was further equalised to 700 words. The first three reviews of each author were grouped as source-known documents (i.e. suspect documents) and the second three reviews were grouped as source-unknown documents (i.e. offender documents). The total number of word tokens in each group was 2,100, which constitutes a realistic sample size for forensic studies in our casework experience. The database was evenly divided into three mutually exclusive test, background and development sub-databases, each consisting of documents from 719 authors.

The documents stored in the test database were used for assessing the FTC system performance by simulating same-author (SA) and different-author (DA) comparisons. From the 719 authors in the test database, 719 SA comparisons and 516,242 (= ${}_{719}C_2 \times 2$) DA comparisons can be simulated.

The documents stored in the background database were used differently depending on the method. For the score-based method, they were used to train the score-to-LR conversion model, and in the feature-based method, they were used to assess the typicality of the documents under comparison.

For various reasons, including violation of modelling assumptions and data scarcity, the estimated LRs may not be well calibrated, in which case they cannot be interpreted as the strength of evidence (Morrison, 2013). A development database is typically used to calibrate the raw LRs via logistic-regression. However, in this study it was found that the LRs derived from the score-based method were well calibrated to begin with; thus logistic-regression calibration was not required. The development database was only used to fuse and calibrate the LRs derived from the feature-based method in this study. A more detailed explanation on logistic regression fusion/calibration is given in §3.5.

The type of communication that the current study focuses on is the one-to-many type of communication. Although the selected database is designed specifically for authorship verification tests, it is not a forensic database. To the best of our knowledge, there are no databases available of real forensic messages, nor any specifically designed with forensic conditions in mind. Nevertheless, the database used in this study was judged to be the most appropriate of existing databases to simulate a forensic scenario involving one-to-many communication. The product reviews were written as personal opinions and assessments of a given product addressing a public audience, and the review

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

messages have a clear purpose; conveying one's views to others. So, the content of the messages is focused and topic specific, like the malicious use of the one-to-many type of communication platforms (e.g. the spread of fake news, malicious intent and the defamation of individuals/organisations).

3.2 Tokenisation and Bag of Words Model

The `tokens()` function from the `quanteda` library (Benoit et al., 2018) in `R` (R Core Team, 2017) was used to tokenise the texts with the default settings. That is, all characters were converted to lower case without punctuation marks being removed; punctuation marks are treated as single word tokens. In order to preserve individuating information in author's morpho-syntactic choices (HaCohen-Kerner et al., 2018; Omar and Hamouda, 2020), no stemming algorithm was applied.

The 400 most frequent occurring words in the entire dataset were selected as components for a bag-of-words model. The occurrences of these words were then counted for each document. More specifically, the documents (x, y) under comparison were modelled as the vectors ($x = \{w_1^x, w_2^x \cdots w_N^x\}$ and y = $\{w_1^y, w_2^y \cdots w_N^y\}$) with the word counts ($w_i^j, i \in \{1 \cdots N\}, j \in \{x, y\}$).

In the experiments, the size (N) of the bag-of-words vector is incremented by 5 from $N = 5$ to $N = 20$, and then by 20 until $N = 400$. The 400 most frequent words are sorted according to their frequencies in a descending order. $N = 400$ was chosen as the cap of the experiments because the experimental results showed the performance ceiling before $N = 400$.

3.3 Score-based Method with Distance Measure (Baseline Model)

Estimating LRs using score-based methods is common in the forensic sciences (Bolck et al., 2015; Chen et al., 2018; Garton et al., 2020; Morrison and Enzinger, 2018). For score-based methods, the evidence consists of scores, $\Delta(x, y)$, which are often measured as the distance between the suspect and offender samples. In this case, the LR can be estimated as the ratio of the two probability densities of the scores under the two competing hypothesis as given in Equation 2).

$$LR = \frac{f(x, y|H_p)}{f(x, y|H_d)} = \frac{f(\Delta(x, y)|H_p)}{f(\Delta(x, y)|H_d)} \qquad 2)$$

The probability densities are trained on the scores obtained from the SA and DA comparisons generated from a background database. That is, the probability densities are used as a score-to-LR conversion model. The Cosine distance was used as a baseline in the current study as its superior performance has been previously reported in authorship attribution studies (Evert et al., 2017; Smith and Aldridge, 2011). The three documents from each group were concatenated as a document of 2,100 words for the score-based method. The count of each word was *z*-score normalised in order to avoid the most frequent words biasing the estimation of the LRs. The *z*-score normalised values were used to represent each document in the bag-of-words model described in §3.2.

Score-based methods project the complex, multivariate feature vector into a univariate score space (Morrison and Enzinger, 2018: 47). Its robustness and ease of implementation for various types of forensic evidence have been reported as benefits (Bolck et al., 2015). However, information loss is inevitable due to the reduction in dimensionality. Another shortcoming is that score-based methods do not account for the typicality of the evidence. Because of these shortcomings, it is reported that the magnitude of the derived LRs is generally weak (Bolck et al., 2015; Morrison and Enzinger, 2018). Nevertheless, the approach has been widely studied across a variety of forensic evidence.

3.4 Feature-based Method with Poisson Model

Feature-based methods maintain the multivariate structure of the data through estimation of the LR directly from the feature values (Bolck et al., 2015). This has the potential to prevent information loss but comes at the cost of added model complexity and reduced computational efficiency. Feature-based methods allow the typicality, not only the similarity, of forensic data to be assessed. In feature-based methods, the LR is estimated as a ratio of two conditional probabilities, which express the similarity and typicality of the samples under comparison. These correspond respectively to the numerator and denominator of Equation 1). Similarity, in this context, refers to how similar/different the source-known and source-questioned documents are with respect to their measured properties, and typicality means how typical/atypical they are in the relevant population. In this study a Poisson distribution was used to construct the LR

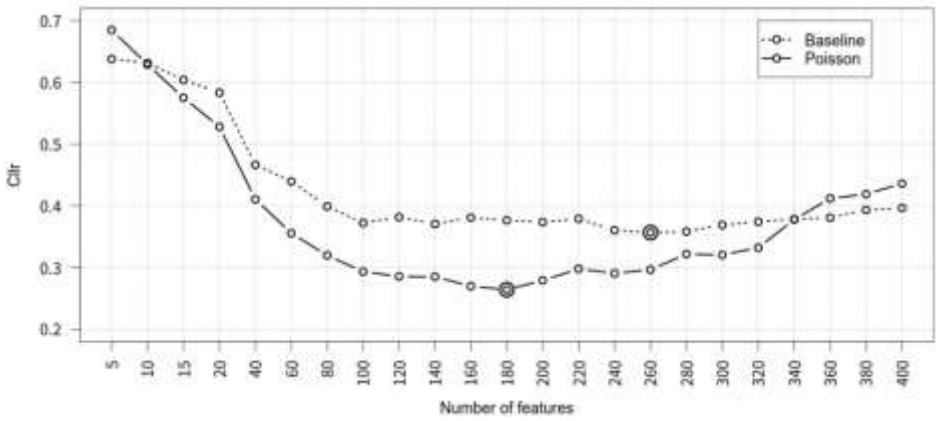

Figure 2: The C_{llr} values of the LRs with the N number of features indicated in the y-axis are plotted separately for the Baseline and the Poisson models. The features are sorted according to the frequencies of the words. The large circles indicate the best C_{llr} values for the models.

model. The probability mass function for the Poisson distribution is given in Equation 3) and the LR model in Equation 4).

$$p(k;\lambda) = e^{-\lambda}\frac{\lambda^k}{k!} \qquad 3)$$

In Equation 3), λ is the shape parameter which indicates the average number of events in the given time interval or space. That is, letting $x = (x_1, \cdots x_k)$ and $y = (y_1, \cdots y_k)$ be the counts of a given word for the suspect and offender documents, an LR for the pair of documents is estimated for the word by Equation 4).

$$LR = \frac{f(x,y|H_p)}{f(x,y|H_d)} = \frac{f(y|x,H_p)}{f(y|H_d)} = \frac{f(y|\lambda_x)}{f(y|\lambda_B)} = \frac{\prod_{i=1}^{k}(y_i|\lambda_x)}{\prod_{i=1}^{k}(y_i|\lambda_B)} = \frac{\prod_{i=1}^{k} e^{-\lambda_x}\frac{\lambda_x^{y_i}}{y_i!}}{\prod_{i=1}^{k} e^{-\lambda_B}\frac{\lambda_B^{y_i}}{y_i!}} \qquad 4)$$

where the λ_x is the mean of x and the λ_B is the overall mean λ of the background database. Both the suspect and offender documents consist of three texts; thus $k = 3$. The second fraction of Equation 4) can be reduced to the third fraction by assuming that the probability of the feature values x is independent of whether x comes from the same source as y or not, and that x and y are independent if H_d is true. LRs were estimated separately for each of the 400 features.

3.5 Logistic-Regression Fusion and Calibration

If the LRs derived separately for the 400 features were independent of one another, they could be multiplied in a naïve Bayesian manner for an overall LR. However, it is known empirically that independence cannot be assumed (Argamon, 2008; Evert et al., 2017). This means, they need to be fused instead, taking the correlations into consideration. Fusion enables us to combine and calibrate multiple parallel sets of LRs from different sets of features/models or even different forensic detection systems, with the output being calibrated LRs. Logistic-regression fusion/calibration (Brümmer and du Preez, 2006) is a commonly used method for LR-based systems. A logistic-regression weight needs to be calculated for each set of LRs, as shown in Equation 5).

$$Fused\ LR = a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_nx_n + b \qquad 5)$$

where, $x_1, x_2, x_3 \ldots x_n$ are the LRs of the first through nth set, and $a_1, a_2, a_3 \ldots a_n$ are the corresponding logistic-regression weights for scaling. The logistic-regression weight for shifting is b. The weights are obtained from the LRs estimated for the SA and DA comparisons from documents in the development database. The number (N) of features to be fused were incremented by 5 from $N = 5$ to $N = 20$, and then by 20 until $N = 400$.

The same technique can be applied to a single set of LRs, in which case, logistic-regression is used only for calibration. However, it was not applied to the LRs derived with the score-based method as they were well-calibrated to start with.

3.6 Evaluation Metrics: Log-LR Cost

The log-LR cost (C_{llr}), which is a gradient metric based on LR, was used to assess the performance of the FTC systems for the two different models (Baseline and Poisson).

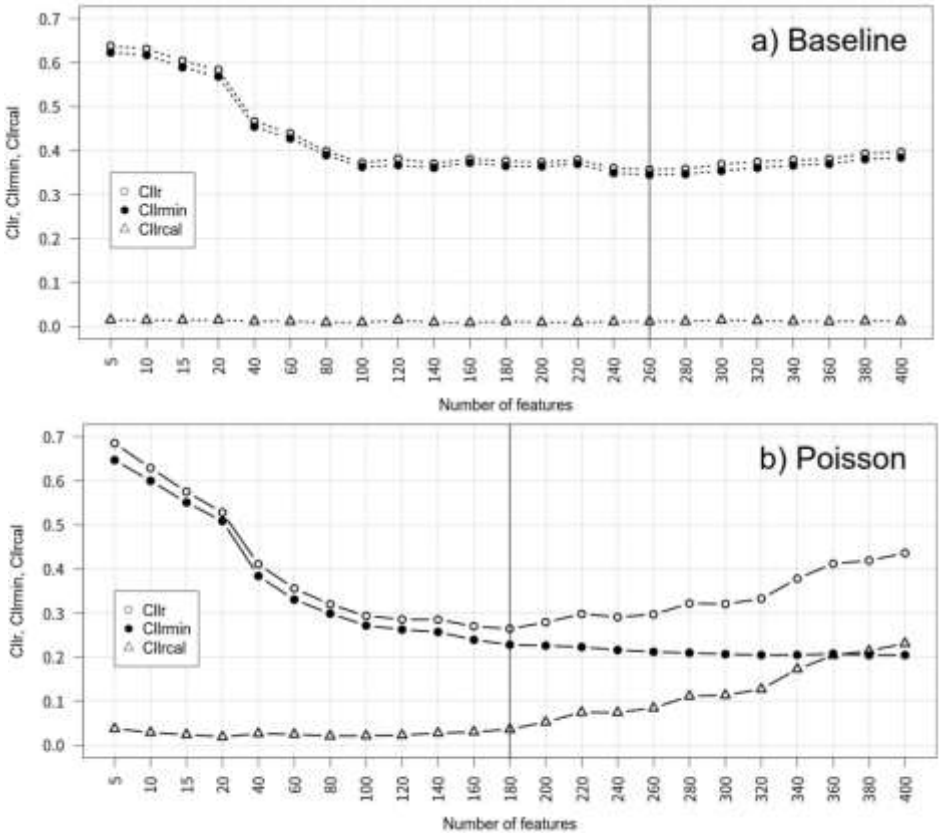

Figure 3: The C_{llr}, C_{llr}^{min} *and* C_{llr}^{cal} values of the LRs, with the *N* number of features indicated in the *y*-axis, are plotted separately for the Baseline (Panel a) and the Poisson (Panel b) models. The features are sorted according to word frequency. The vertical solid line indicates where the best C_{llr} value was obtained.

The calculation of C_{llr} is given in Equation 6) (Brümmer and du Preez, 2006).

$$C_{llr} = \frac{1}{2}\left(\begin{array}{l} \left[\frac{1}{N_{SA}}\sum_{i}^{N_{SA}} log_2\left(1+\frac{1}{LR_i}\right)\right] \\ +\left[\frac{1}{N_{DA}}\sum_{j}^{N_{DA}} log_2(1+LR_j)\right] \end{array} \right) \qquad 6)$$

In Equation 6), N_{SA} and N_{DA} are the number of SA and DA comparisons, and LR_i and LR_j are the LRs derived from the SA and DA comparisons, respectively. C_{llr} takes into account the magnitude of the LR values, and assigns them appropriate penalties. In C_{llr}, LRs that support the counter-factual hypotheses or, in other words, contrary-to-fact LRs (LR < 1 for SA comparisons and LR > 1 for DA comparisons) are heavily penalised and the magnitude of the penalty is proportional to how much the LRs deviate from unity. Optimum performance is achieved when C_{llr} = 0 and decreases as C_{llr} approaches and exceeds 1. Thus, the lower the C_{llr} value, the better the performance.

The C_{llr} measures the overall performance of a system in terms of validity based on a cost function in which there are two main components of loss: discrimination loss (C_{llr}^{min}) and calibration loss (C_{llr}^{cal}) (Brümmer and du Preez, 2006). The former is obtained after the application of the pooled-adjacent-violators (PAV) transformation – an optimal non-parametric calibration procedure. The latter is obtained by subtracting the former from the C_{llr}. In this study, besides C_{llr}, C_{llr}^{min} and C_{llr}^{cal} are also referred to.

The magnitude of the LRs derived from the comparisons are visually presented using Tippett plots. Details on how to read a Tippett plot are given in §5 when the plots are presented.

4 Results and Discussion

The C_{llr} values are plotted as a function of the number of features, separately for the Baseline model and the Poisson model in Figure 2. The number of the features is incremented by 5 from $N = 5$ to $N = 20$, and then by 20 from $N = 20$ to $N = 400$. For example, $N = 5$ means that the overall LRs were obtained by fusing the LRs derived with the five most-frequently occurring words for the feature-based method. Whereas the scores, which are to be converted to the LRs, were measured based on the vector of the five most-frequent words for the score-based method.

As can be observed from Figure 2, the performance of both models improves *en masse* as the N increases until a certain N, after which the performance remains relatively unchanged (or falls slightly). The Baseline model's performance stays relatively stable for a higher number of N, while the performance of the Poisson model begins to decline after 180 features. Due to the deterioration with a large number of feature numbers, although the Poisson model outperforms the Baseline model overall, the Baseline model does better with $N >$ 340.

The best performance, however, was observed for the Poisson with a lower number of features (C_{llr} = 0.26439; N = 180) relative to the Baseline model (C_{llr} = 0.35682; N = 260). The superior performance of the feature-based method (Poisson model) relative to the score-based method (Baseline model) conforms to the reports of previous studies on other types of evidence (Bolck et al., 2015; Morrison and Enzinger, 2018).

As described earlier, the Baseline and the Poisson models exhibit different performance characteristics in terms of the number of features required for optimal C_{llr} and the effect of increasing N. The performance of the Baseline model stays relatively unchanged with more features, while the performance of the Poisson model continuously declines with more features. In order to further investigate this performance difference, the C_{llr}, C_{llr}^{min} and C_{llr}^{cal} values are plotted separately for the two models in Figure 3.

For the Baseline model, it can be seen from Figure 3a that 1) the C_{llr}^{cal} values consistently remain close to 0, meaning the LRs are very well calibrated regardless of the number of features, and also that 2) the C_{llr}^{min} values display an almost identical trend as the C_{llr} values, meaning that like the C_{llr} values, the discriminability potential remains relatively constant even with an increase in the feature number after the best-performing point. In contrast, the three metrics plotted in Figure 3b reveal some notably different characteristics of the Poisson model. The C_{llr}^{cal} values stay low only until N = 140~160, after which the C_{llr}^{cal} values start increasing at a constant rate with an increase in the feature number; that is, the LRs become less well calibrated as N increases beyond 140~160 features. Unlike the calibration loss (and the Baseline model), the discriminability potential, quantified by C_{llr}^{min}, continues to improve at a small but constant rate, even after N = 180, where the best C_{llr} was observed. Thus, it is clear from Figure 3 that the deterioration of the Poisson model in performance after N = 180 is not due to a poor discrimination performance but due to a poor calibration performance. As explained in §3.5, logistic-regression fusion/calibration should theoretically yield well calibrated LRs. The poor calibration performance observed for the Poisson model for large feature numbers may be due to the interaction between the dimensions of the LRs to be fused and the amount of the training data for the fusion/calibration weights. This seems to be a typical example of the phenomenon known as the 'curse of dimensionality' (Bellman, 1961: p. 97), but further analysis is warranted. Nevertheless, it is clear that the use of a Poisson-based model, which theoretically better suits the distributional pattern of textual data and allows the rarity/typicality of evidence to be considered for LR estimation, can offer performance gains.

5 Feature Selection

For the Poisson model, LRs were first estimated separately for each of the 400 feature words. The resulting LRs were fused by gradually increasing the number of LRs included in the fusion set. LRs were arranged according to word frequency in the experiments reported in §4. Yet, the performance of a given feature (i.e. word) did not always correspond to the frequency of its occurrence. This is illustrated in Table 1, which lists the ten most frequently occurring words and the ten words with the highest discriminability (i.e. C_{llr}^{min}).

By word frequency		By C_{llr}^{min}	
Frequency	Words	Frequency	Words
1	'.'	3	','
2	'the'	1	'.'
3	','	41	'it's'
4	'and'	35	'!'
5	'i'	31	'-'
6	'a'	28	'('
7	'to'	27	')'
8	'it'	5	'i'
9	'of'	84	'i'm'
10	'is'	4	'and'

Table 1: Ten most-frequent (left) and lowest-C_{llr}^{min} (right) words

Thus, in this section, the words were first sorted according to their performance in terms of the C_{llr}^{min} values, and then the LRs were fused/calibrated based on the sorted words. The C_{llr} values of

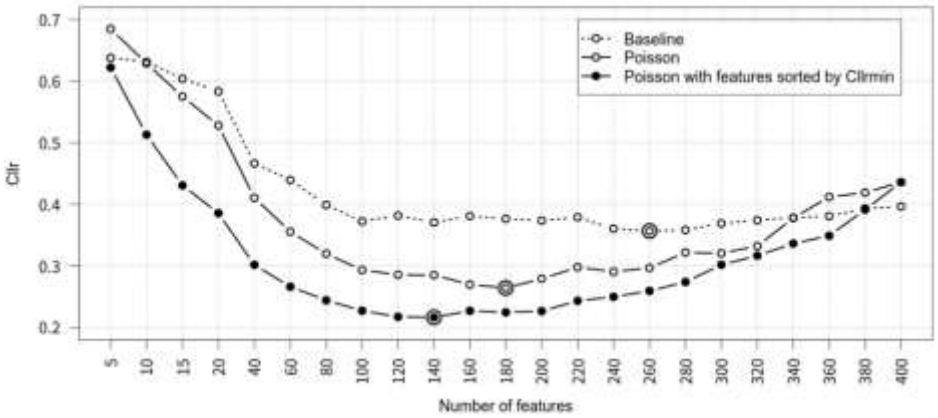

Figure 4: The C_{llr} values of the (fused) LRs with the N number of C_{llr}^{min}-sorted features indicated in the y-axis are plotted together with the results presented in Figure 2 for comparisons. The large circles indicate the best C_{llr} values for the models.

the experiments are plotted in Figure 4 including the results presented in Figure 2 for comparison.

It is clear from Figure 4 that selecting the features according to their C_{llr}^{min} values contributes to an improvement in performance for all numbers of features. As a result, the C_{llr} is lower (0.21664) with less features ($N = 140$) compared to the results with the unsorted features.

This feature selection approach was only possible because the LRs are estimated separately for the each of the 400 different words. This is possibly an advantage for the Poisson model.

The magnitude of the LRs with the best-performing settings are shown on Tippett plots, separately for the Baseline model, the original Poisson model, and the Poisson model with C_{llr}^{min}-sorted features in Figure 5. Tippet plots show the cumulative proportion of LRs from the SA comparisons (SALRs), which are plotted rising from the left, as well as of the LRs of the DA comparisons (DALRs), plotted rising from the right. For all 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 the low C_{llr}^{cal} values indicate, it can also be observed from Figure 5 that the LRs are very well calibrated. However, comparing Figure 5a and Figure 5bc we see that the magnitude of the LRs are weaker overall in the Baseline model compared to the two Poisson models; the Tippet lines are further from unity ($\log_{10}$ LR = 0) for the Poisson models than the Baseline models. Although the overall magnitude of LRs is greater for the Poisson models, unlike the Baseline model, they evince some very strong contrary-to-fact DALRs (which are indicated by arrows in Figure 5). This is a concern, and the reason for this needs to be further investigated.

6 Conclusions and Future Studies

A feature-based approach for estimating forensic LRs was implemented with a Poisson model for the first time in LR-based FTC. The results of the experiments showed that the feature-based FTC system outperforms the score-based FTC system with the Cosine distance. It has also been demonstrated that the performance of the feature-based system can be further improved by selecting the sets of LRs to be fused according to their C_{llr}^{min} values. It was observed that the discrimination loss in the feature-based FTC system reduces as the number of features increases, but becomes less well calibrated with a large number of features. It has been argued that this is a typical case of the 'curse of dimensionality' (Bellman, 1961: p. 97), but further investigation is required.

A simple one-level Poisson LR model shows good performance. However, it has been reported that word counts are often modelled poorly by standard parametric models such as the Binomial and Poisson models, and some alternatives have been proposed, such as the negative Binomial and the zero-inflated Poisson (Jansche, 2003; Pawitan, 2001). Alternatively, a two-level Poisson model might be implemented based if the prior distributions of λ is assumed (Aitken and Gold, 2013; Bolck and Stamouli, 2017). These alternatives should be tested to see if any improvements in performance are achievable.

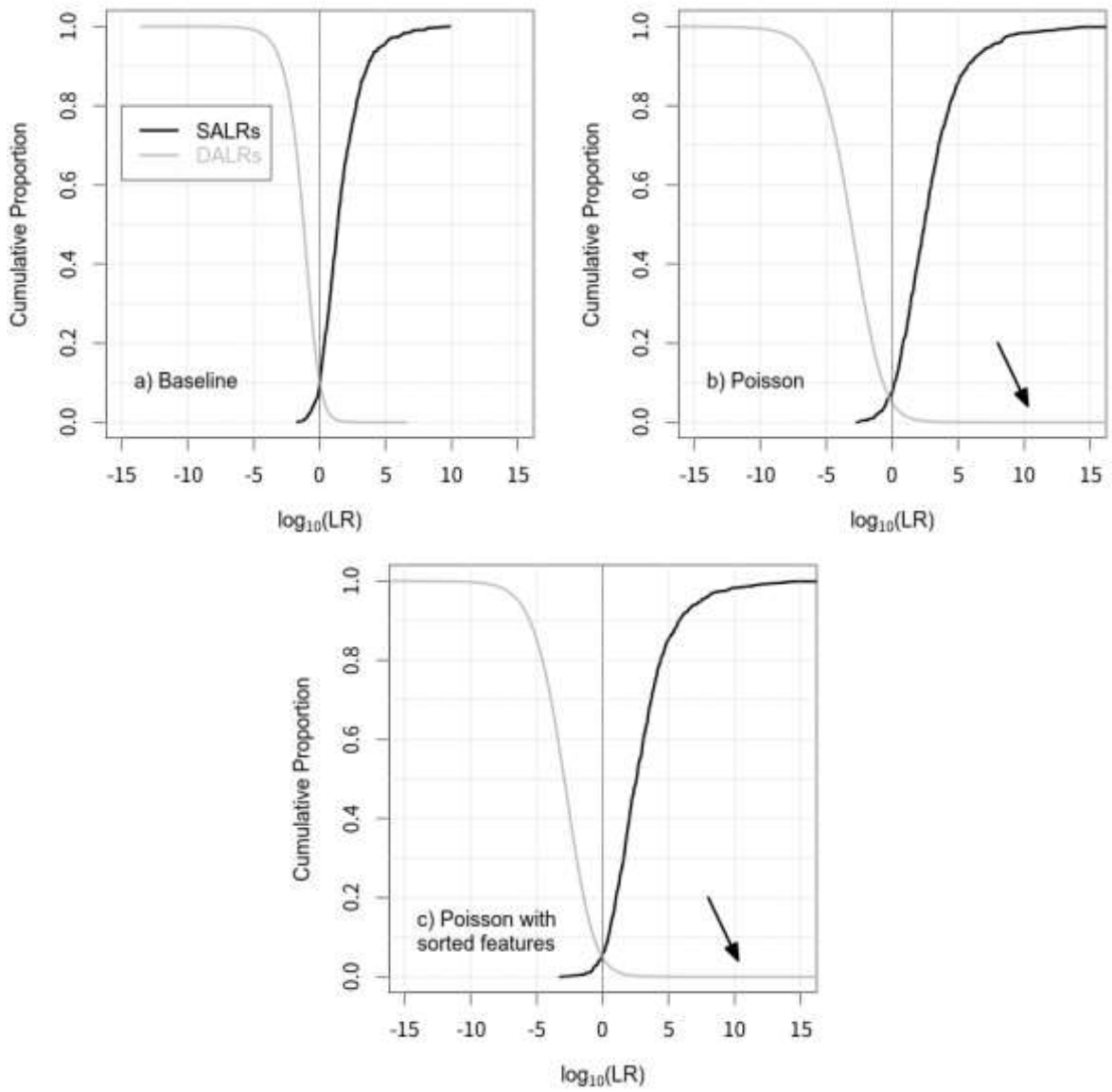

Figure 5: Tippett plots showing the magnitude of the derived LRs. Panel a) = Best-performing Baseline model; Panel b) = Best-performing original Poisson model; Panel c) = Best-performing Poisson model with sorted features according to their C_{llr}^{min} values. Note that some LRs extend beyond ±15 of the *y*-axis. Arrows indicate very strong contrary-to-fact DALRs.

The set of features tested in the current study is only one type of many potential authorship attribution features (according to Rudman (1997), over 1,000 different feature types have so far been proposed in the literature). While the purpose of the present study was to compare modelling approaches, rather than the relative performance of different feature types, an interesting future task would be to explore a richer feature set and the effect of different pre-processing techniques (e.g. stop word removal).

The LRs derived using the score-based method were well-calibrated, and therefore logistic-regression calibration was not necessary). This was not the case for LRs using the feature-based method where logistic-regression fusion/calibration was required. This procedure necessitates an extra set of data, namely a development database, and is another shortcoming of the feature-based method applied in this study

Acknowledgements

The authors thank the reviewers for their valuable comments. The first author's research is supported by an Australian Government Research Training Program Scholarship.

References

AbdulRazzaq, A. A. and Mustafa, T. K. (2014) Burrows-Delta method fitness for Arabic text authorship Stylometric detection. *International Journal of Computer Science and Mobile Computing 3*(6): 69-78.

Aitken, C. G. G. and Gold, E. (2013) Evidence evaluation for discrete data. *Forensic Science International 230*(1-3): 147-155. https://dx.doi.org/10.1016/j.forsciint.2013.02.042

Argamon, S. (2008) Interpreting Burrows's Delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing 23*(2): 131-147. https://dx.doi.org/10.1093/llc/fqn003

Association of Forensic Science Providers. (2009) Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice 49*(3): 161-164. https://doi.org/10.1016/j.scijus.2009.07.004

Bellman, R. E. (1961) *Adaptive Control Processes: A Guided Tour*. Princeton: Princeton University Press.

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

Bolck, A. and Stamouli, A. (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

Bolck, A., Weyermann, C., Dujourdy, L., Esseiva, P. and van den Berg, J. (2009) Different likelihood ratio approaches to evaluate the strength of evidence of MDMA tablet comparisons. *Forensic Science International 191*(1-3): 42-51. https://dx.doi.org/10.1016/j.forsciint.2009.06.006

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

Burrows, J. F. (2002) 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing 17*(3): 267-287. https://dx.doi.org/10.1093/llc/17.3.267

Champod, C., Evett, I. W. and Kuchler, B. (2001) Earmarks as evidence: A critical review. *Journal of Forensic Sciences 46*(6): 1275-1284. http://dx.doi.org/10.1520/JFS15146J

Chen, X. H., Champod, C., Yang, X., Shi, S. P., Luo, Y. W., Wang, N., . . . Lu, Q. M. (2018) Assessment of signature handwriting evidence via score-based likelihood ratio based on comparative measurement of relevant dynamic features. *Forensic Science International 282*: 101-110. https://dx.doi.org/10.1016/j.forsciint.2017.11.022

Eder, M. (2015) Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities 30*(2): 167-182. https://doi.org/10.1093/llc/fqt066

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. and Weir, B. S. (1998) *Interpreting DNA Evidence : Statistical Genetics for Forensic Scientists*. Sunderland, Mass.: Sinauer Associates.

Garton, N., Ommen, D., Niemi, J. and Carriquiry, A. (2020). Score-based likelihood ratios to evaluate forensic pattern evidence. *arXiv preprint arXiv:2002.09470*. Retrieved on July 20 2020 from https://arxiv.org/abs/2002.09470

HaCohen-Kerner, Y., Miller, D., Yigal, Y. and Shayovitz, E. (2018) Cross-domain authorship attribution: Author identification using char sequences, word unigrams, and POS-tags features. *Proceedings of Notebook for PAN at CLEF 2018*: 1-14.

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

Hoffmann, K. (1991) Statistical evaluation of the evidential value of human hairs possibly coming from multiple sources. *Journal of Forensic Sciences 36*(4): 1053-1058. https://dx.doi.org/10.1520/JFS13120J

Hoover, D. L. (2004a) Delta prime? *Literary and Linguistic Computing 19*(4): 477-495. https://dx.doi.org/10.1093/llc/19.4.477

Hoover, D. L. (2004b) Testing Burrows's Delta. *Literary and Linguistic Computing 19*(4): 453-475. https://dx.doi.org/10.1093/llc/19.4.453

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. (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

Jannidis, F., Pielström, S., Schöch, C. and Vitt, T. (2015) Improving Burrows' Delta. An empirical evaluation of text distance measures. *Proceedings of Digital Humanities 2015*: 1-10.

Jansche, M. (2003) Parametric models of linguistic count data. *Proceedings of Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*: 288-295.

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

Morrison, G. S. and Enzinger, E. (2018) Score based procedures for the calculation of forensic likelihood ratios - Scores should take account of both similarity and typicality. *Science & Justice 58*(1): 47-58. https://dx.doi.org/10.1016/j.scijus.2017.06.005

Morrison, G. S., Enzinger, E. and Zhang, C. (2018) Forensic speech science. In I. Freckelton and H. Selby (eds.), *Expert Evidence*. Sydney, Australia: Thomson Reuters.

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 Sciences 52*(1): 54-64. https://dx.doi.org/10.1111/j.1556-4029.2006.00327.x

Omar, A. and Hamouda, W. (2020) The effectiveness of stemming in the stylometric authorship attribution in Arabic. *International Journal of Advanced Computer Science and Applications 11*(1): 116-121. https://dx.doi.org/10.14569/IJACSA.2020.0110114

Pawitan, Y. (2001) *In All Likelihood : Statistical Modelling and Inference Using Likelihood*. Oxford: Oxford University Press.

Rose, P. (2002) *Forensic Speaker Identification*. London: Taylor & Francis.

Rudman, J. (1997) The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities 31*(4): 351-365. https://dx.doi.org/10.1023/A:1001018624850

Rybicki, J. and Eder, M. (2011) Deeper Delta across genres and languages: Do we really need the most frequent words? *Literary and Linguistic Computing 26*(3): 315-321. https://dx.doi.org/0.1093/llc/fqr031

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

Þorgeirsson, H. (2018) How similar are Heimskringla and Egils saga? An application of Burrows' delta to Icelandic texts. *European Journal of Scandinavian Studies 48*(1): 1-18. https://doi.org/10.1515/ejss-2018-0001

Vergeer, P., Bolck, A., Peschier, L. J. C., Berger, C. E. H. and Hendrikse, J. N. (2014) Likelihood ratio methods for forensic comparison of evaporated gasoline residues. *Science & Justice 54*(6): 401-411. https://dx.doi.org/10.1016/j.scijus.2014.04.008