

# A likelihood ratio-based forensic text comparison in predatory chatlog messages

*Shunichi Ishihara*

*Australian National University*

*shunichi.ishihara@anu.edu.au*

An experiment in Forensic Text Comparison (FTC) within the Likelihood Ratio (LR) framework is described, which determines the strength of authorship attribution evidence from chatlog messages using so-called lexical features. More specifically, in this study I will investigate 1) the degree of evidential strength (or LR) that can be obtained from chatlog messages and 2) how the performance of the FTC system and the magnitudes of the LRs are influenced by the sample size for modelling. The performance of the system is assessed using the log-LR cost ( $C_{llr}$ ) and the magnitudes of the obtained LRs are visually presented as Tippett plots. It is demonstrated in this study that you can use the lexical features within the LR framework to discriminate same-author and different-author chatlog messages.

**Keywords:** likelihood ratio, forensic text comparison, chatlog messages, multivariate kernel density, the log likelihood ratio cost

## 1. Forensic Text Comparison and Bayes' Theorem

In a typical case where forensic text comparison (FTC) is performed, the texts (e.g. messages) belonging to one party (e.g. the suspect, defendant or accused) are compared with those written by an individual who is alleged to have committed an offence, in order to assist the trier-of-fact to conclude whether the suspect wrote the incriminating text or not. In the process of making a final verdict, the trier-of-fact needs to consider two conditional probabilities; namely the

*Selected Papers from the 44<sup>th</sup> Conference of the Australian Linguistic Society, 2013*

edited by Lauren Gawne and Jill Vaughan

probability of the suspect and the offender being the same person and the probability of the suspect and the offender being different people, given the evidence. The ratio of these two probabilities is called the posterior odds in Bayes' Theorem, and can be mathematically represented as follows:

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

The numerator is the probability ( $p$ ) of the prosecution hypothesis ( $H_p$ ), given ( $|$ ) the evidence ( $E$ ) (= the probability that the suspect and the offender are the same person given the evidence), and the denominator is the probability of the defence hypothesis ( $H_d$ ) given the same evidence (= the probability that the suspect and the offender are different people given the same evidence). If  $p(H_p|E)$  is 90% and  $p(H_d|E)$  is 10%, for example, the ratio between them, or the posterior odds, is 9. It means that given the evidence, it is 9 times more likely that the suspect and the offender are the same individual than they are different individuals. The interpretation of the value; whether it is strong enough or not strong enough to conclude that the suspect is guilty, is a privilege belonging only to the trier-of-fact.

As the posterior odds value increases relative to unity, it becomes more likely that the suspect and the offender are the same person. Conversely, as the posterior odds value decreases relative to unity, it becomes more likely that they are different people. But, how can we estimate these posterior probabilities in the first place? The solution is Bayes' Theorem, which is given below.

$$2) \quad \underbrace{\frac{p(H_p|E)}{p(H_d|E)}}_{\text{posterior odds}} = \underbrace{\frac{p(H_p)}{p(H_d)}}_{\text{prior odds}} * \underbrace{\frac{p(E|H_p)}{p(E|H_d)}}_{\text{likelihood ratio}}$$

As can be seen in the above formula, the posterior odds is the product of the prior odds and the likelihood ratio. That is, posterior probabilities cannot be estimated without prior probabilities, but the forensic expert is not usually privy

to these probabilities. That is, it is not logically possible for the forensic expert, who cannot have access to the prior probabilities<sup>1</sup>, to estimate the posterior probabilities. Since the decision of whether the suspect is guilty or not guilty is only assigned to the trier-of-fact, it is not legally appropriate for the forensic expert to refer to the posterior probabilities, which may well impinge on considerations of the ultimate issue: guilty vs. not guilty.

Thus, the task of the forensic scientist is to estimate the strength of evidence, which is technically called Likelihood Ratio (LR).

### 1.1 Likelihood ratio

The likelihood ratio (LR) is the probability that the evidence would occur if an assertion is true, relative to the probability that the evidence would occur if the assertion is not true (Robertson & Vignaux 1995). Thus, the LR can be expressed in 3).

$$3) \quad LR = \frac{p(E|H_p)}{p(E|H_d)}$$

For FTC, it will be the probability of observing the difference (referred to as the evidence;  $E$ ) between the texts written by the offender and those written by the suspect if they had come from the same author ( $H_p$ ) (i.e. if the prosecution hypothesis is true) relative to the probability of observing the same evidence ( $E$ ) if they had been produced by different authors ( $H_d$ ) (i.e. if the defence hypothesis is true). The relative strength of the given evidence with respect to the competing hypotheses ( $H_p$  vs.  $H_d$ ) is reflected in the magnitude of the LR. The more the LR deviates from unity ( $LR = 1$ ), the greater support for either the prosecution hypothesis ( $LR > 1$ ) or the defence hypothesis ( $LR < 1$ ). The important point is that the LR is concerned with the probability of the evidence, given the hypothesis (either prosecution or defence), which is the province of forensic

---

<sup>1</sup> Refer to Rose (2002: 63-64) and Morrison (2009a: 158-159) in which they provide more detailed explanation for the role of prior odds in Bayes' Theorem and also discuss why the forensic scientist should not and cannot calculate or know the prior odds.

scientists, while the trier-of-fact is ultimately concerned with the probability of the hypothesis (either prosecution or defence), given the evidence. Therefore, for example, an LR of 10 means that the evidence is 10 times more likely to occur if the offender and the suspect had been the same individual than if they had been different individuals. Note that an LR value of 10 does *not* mean that the offender and the suspect are 10 times more likely to be the same person than different people, given the evidence.

In the context of FTC, the task of the forensic scientist is to provide the court with a strength-of-evidence statement in answer to the question: How much more likely are the observed differences/similarities between text samples written by the offender and those written by the suspect to arise under the hypothesis that they have been written by the same author than the hypothesis that they have been written by different authors?<sup>2</sup>

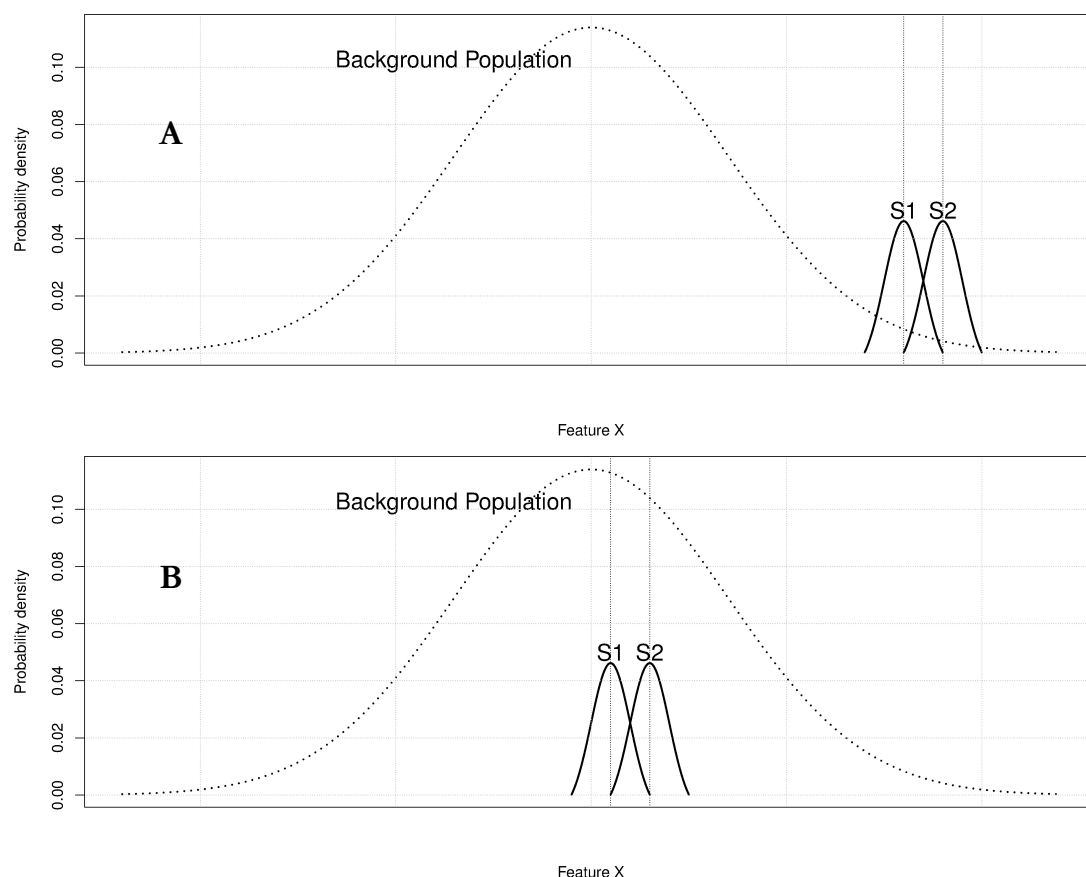
In order to calculate an LR, we need three sets of samples (e.g. messages): a set of questioned samples (offender's samples); a set of known samples (defendant's samples); and the background or reference samples. This is because an LR is a ratio of similarity to typicality, which quantifies how similar/different the questioned and the known samples are, and then evaluates that similarity/difference in terms of typicality/atypicality against the relevant background population (i.e. reference samples). If the questioned and known samples are more different or more typical, it will produce lower LRs than the LRs produced when the same samples are more similar or less typical.

This point is illustrated in Figure 1. The similarity/difference between S1 (the questioned sample) and S2 (the known sample) is identical both in Figure 1A and Figure 1B in that the distance between the mean of the S1 samples and the mean of the S2 samples is the same both for Figure 1A and Figure 1B whereas they are more typical against the background population in Figure 1B (the S1 and S2 samples are located close to the mode of the background population) than Figure 1A (the S1 and S2 samples are located at the outskirts of the background population). That is, for Figure 1B, there is a higher probability of observing that degree of similarity/difference between S1 and S2 by randomly selecting authors

---

<sup>2</sup> In real casework, the two conditional probabilities, the denominator and numerator of the LR formula, can be quoted in the report. However, the important point is that they need to be treated as one set. This is because both similarity and typicality need to be considered for evidence to be assessed.

from the relevant population than for Figure 1A. Thus, the S1 and S2 pair will produce a higher LR value for Figure 1A than for Figure 1B.



**Figure 1.** The relationship between similarity and typicality for LR values. S1 and S2 are questioned and known samples, respectively.

## 1.2 Paradigm shift and the role of forensic scientists

Having been largely motivated by the U.S. Supreme Court *Daubert* ruling (*Daubert v. Merrell Dow Pharmaceuticals Inc.*, 1993) on the admissibility of scientific evidence, the forensic sciences are experiencing a paradigm shift in the evaluation and presentation of evidence (Saks & Koehler 2005). According to Morrison (2009b: 299), one of the important components of this paradigm shift is “the adoption of the *likelihood-ratio framework* for the evaluation of evidence”.

There are a large number of authorship analysis studies claiming to be forensic, particularly in the fields of computational linguistics and natural language processing (Grant 2007, Lambers & Veenman 2009, Iqbal et al. 2013, Corney et al. 2001, Khan et al. 2012, Teng et al. 2004, Zheng et al. 2003, Mohan et al. 2010,

Grant 2010, Layton et al. 2010). Yet, many of them consider the problem as a classification problem; for example, whether a system correctly identifies texts written by the same author as being from the same author. Giving an answer to the classification problem is equivalent to referring to the posterior probabilities. However, as explained in §1, it is not logically possible for the forensic expert, who is usually not in the position of knowing the prior probabilities, to refer to the posterior probabilities. It is also legally inappropriate for the forensic expert to give an answer to the classification problem because it is the same as stating whether the suspect is guilty or not, and the forensic expert would violate the province of the trier-of-fact by doing so.

As repeatedly emphasised, the role of the forensic scientist is not to give an answer to a classification problem – the task of the trier-of-fact – but to assist the trier-of-fact to make a decision, by providing them with the strength of evidence (or LR) (Aitken & Stoney 1991, Aitken & Taroni 2004, Robertson & Vignaux 1995).

### 1.3 *Aims*

Emulating DNA forensic science, many fields of forensic sciences, such as fingerprint (Neumann et al. 2007), handwriting (Bozza et al. 2008) and voice (Morrison 2009b), started adopting the Likelihood Ratio (LR) framework to quantify evidential strength. However, despite this trend, and the fact that the use of the LR framework has been advocated as the logically and legally correct way of analysing and presenting forensic evidence in the major textbooks on the evaluation of forensic evidence (e.g. Robertson & Vignaux 1995), and by forensic statisticians (e.g. Aitken & Stoney 1991, Aitken & Taroni 2004), LR-based studies on forensic authorship analysis<sup>3</sup> are conspicuous in their rarity.

Thus, the aim of the current study is 1) to investigate the degree of evidential strength (or LR) that can be obtained from chatlog messages. Moreover, chatlog messages are essentially short. Thus, I will also investigate 2) how the

---

<sup>3</sup> Instead of the term ‘forensic authorship analysis’, I use the term ‘forensic text comparison’ in this study in order to emphasise that the task of the forensic scientist is to compare the offender’s and the suspect’s text samples as evidence, and estimate the strength of that evidence.

performance of the system and the magnitudes of the LRs are influenced by the sample size for modelling.

## 2. Testing design

### 2.1 Database and selection of messages

In this study, I used an archive of chatlog messages<sup>4</sup> which is a collection of real pieces of chatlog evidence used to prosecute paedophiles. Despite the wide recognition that the misuse of chat rooms by online child sex offenders is a serious problem, forensic studies specifically targeting chatlog messages are quite sparse (cf. Kucukyilmaz et al. 2008). This is my motivation to work on chatlog messages in this study. As of May 2013, the archive contains messages from more than 500 criminals. Out of the archive, 115 authors, whose chatlog messages were reformatted for the FTC experiments, were used for the current study.

The following are two chatlog communications extracted from the archive, which are given as examples to show how differently authors write their messages. The first one was an exchange between predator 1 (P1) and the undercover police officer (UP), and the second one is between P2 and UP. Only the first 12 messages were extracted from the original message set. The predators' messages are highlighted in bold face.

Chatlog 1: between P1 and UP

**P1: I'm from Portland...**

**P1: :)**

UP: cool me 2

**P1: I'm Pedro...**

UP: im brooke

**P1: I have an apartment on Forest ave...**

**P1: How old r u?**

UP: im good

UP: is that ur pic?

**P1: Yes...would u like 2 c another?**

UP: ok

**P1: have u got one u wish 2 share??**

---

<sup>4</sup> <http://pifi.org/>

Chatlog 2: between P2 and UP

**P2:** *hu*

**P2:** *hi*

UP: *hi*

**P2:** *?*

**P2:** *i said hi*

**P2:** *where in ga you from*

UP: *sw*

UP: *u*

**P2:** *sw*

**P2:** *bainbridge here*

UP: *near columbus*

**P2:** *cool what town if you dont mind me asking*

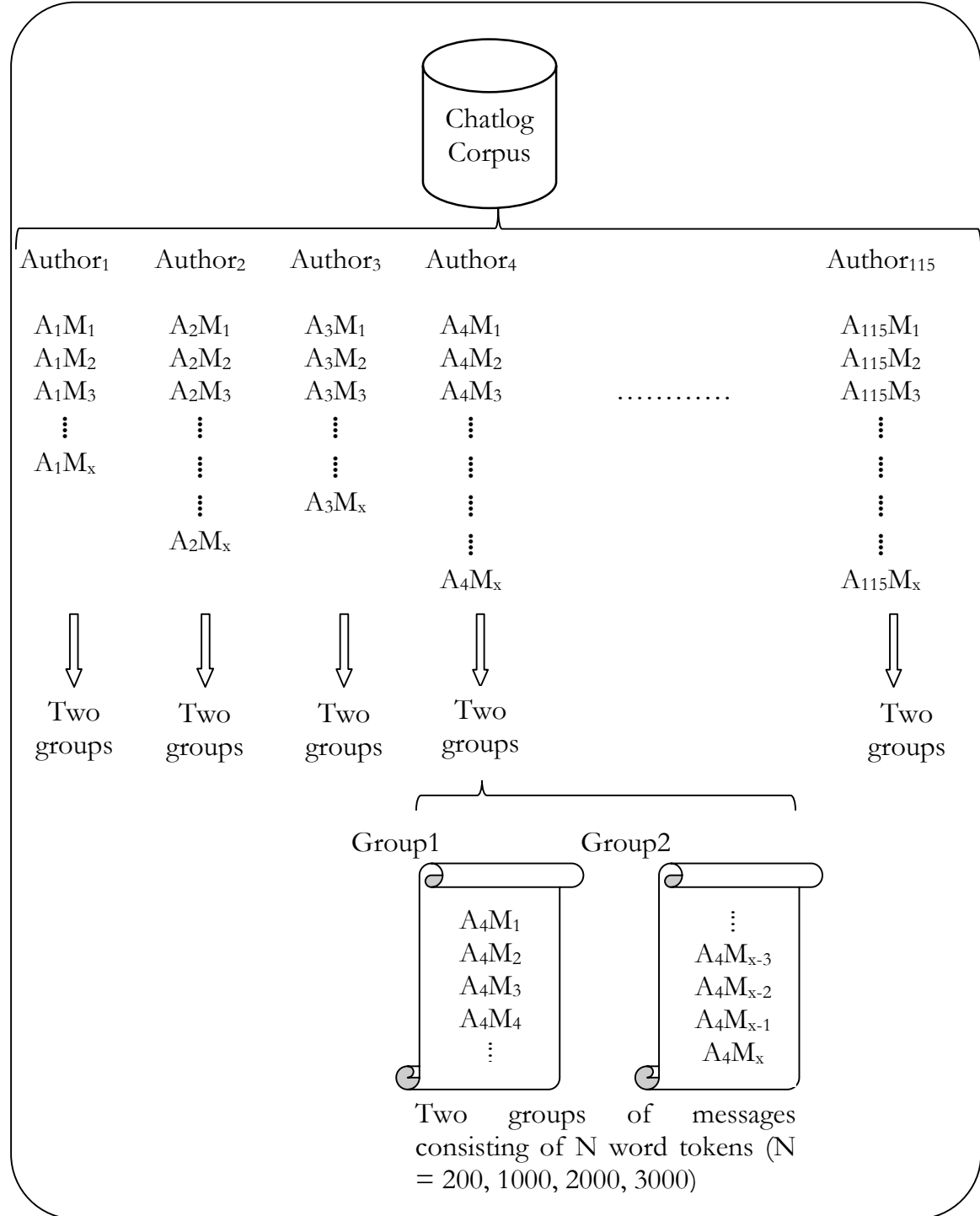
A few clear differences can be identified between P1 and P2 in their writing attributes as represented in these examples. The use of the ‘...’ and ‘??’ appears to be unique to P1 as P2 does not use the period and the question mark at all in the sentence-final position. Another difference is that P1 tends to capitalise the initial letter of the sentences (e.g. ‘I’m’, ‘How’, ‘Yes’) while P2 consistently uses lowercase in all messages (e.g. ‘i said hi’, ‘where in ga you from’).

Two types of comparisons are necessary to assess the validity of an FTC system: one type is same-author (SA) comparisons and the other type is different-author (DA) comparisons. In SA comparisons, two groups of texts produced by the same author will be compared and evaluated with the derived LR. Given the same origin, it is expected that the derived LR will be higher than 1. In DA comparisons, *mutatis mutandis*, they are expected to receive an LR lower than 1 given that they are from different-origins.

In order to set up SA and DA comparisons, we need two non-contemporaneous groups of messages from each of the authors. They need to be non-contemporaneous because offender and suspect samples are usually non-contemporaneous in nature. To achieve this, I added messages one by one from the chronologically sorted messages to the group. For one message group, I started from the top of the chronologically sorted messages, while for the other group of the same author, I started from the bottom, and then the two groups of messages were checked to see if they were truly non-contemporaneous. I created three different message groups differing in the number of word tokens included: 500, 1500 and 2500, so that I could investigate the second research aim.



The process of the above-described message selection and grouping is illustrated in Figure 2.



**Figure 2:** A schematic illustration of the selection of messages and the compilation of datasets differing in sample size.  $x$  = the number of messages belonging to an author ( $x$  can be different depending on the author). Messages are all chronologically sorted for each author.

After the process illustrated in Figure 2, each of the 115 authors has two non-contemporaneous message groups, which enables us to carry out 115 SA

comparisons. As for DA comparisons, four different DA comparisons are possible (e.g. (A)uthor<sub>1</sub>(G)roup<sub>1</sub> vs. A<sub>2</sub>G<sub>1</sub>; A<sub>1</sub>G<sub>1</sub> vs. A<sub>2</sub>G<sub>2</sub>; A<sub>1</sub>G<sub>2</sub> vs. A<sub>2</sub>G<sub>1</sub>; A<sub>1</sub>G<sub>2</sub> vs. A<sub>2</sub>G<sub>2</sub>) for each pair of different authors (e.g. A<sub>1</sub> vs A<sub>2</sub>). Since the number of non-overlapping different author pairs is 6555 ( $= {}^{115}C_2$ ) and four different DA comparisons are possible for each different pair, as a whole, 26220 DA comparisons ( $= 6555*4$ ) are possible ( ${}^{115}C_2*4$ ).

## 2.2 Tokenisation and feature extraction

The chatlog messages were tokenised<sup>5</sup> into words using the whitespace tokeniser of the Natural Language Toolkit.<sup>6</sup>

Following the results of previous authorship studies (Iqbal et al. 2010a, De Vel et al. 2001, Zheng et al. 2006, Ishihara 2012), the features listed in Table 1 were used in the current study in order to model each message group. These features are, in a broad sense, lexical features.

Type		Features
vocabulary richness	1.	<i>Yule's I</i> (the inverse of <i>Yule's K</i> )
	2.	Type-token ratio ( <i>TTR</i> )
	3.	<i>Honoré's R</i>
lexical: token-based	4.	Average token number per message line
	5.	SD of the token number per message line
lexical: character based	6.	Average character number per message line
	7.	SD of the character number per message line
	8.	Upper case character ratio
	9.	Digit character ratio
	10.	Average character number per token
	11.	Punctuation character ratio ( , . ? ! ; : ' " )
	12.	Special character ratio ( < > %   [ ] { } \ / @ # ~ + - * \$ ^ & = )

**Table 1.** List of features.

All of the features given in Table 1 are features based on words (types or tokens) and characters. They can be further sub-classified into vocabulary richness (1~3), token-based (4~5) and character-based (6~12) features. The latter two feature

<sup>5</sup> Tokenisation is the process of splitting text into tokens (i.e., individual words and punctuation).

<sup>6</sup> <http://nltk.org>

types (4~12) are self-explanatory. The formulae for the features of vocabulary richness, namely *Yule's I* (1), *TTR* (2) and *Honoré's R* (3) are given in 4), 5) and 6), respectively (Baayen 2001). All of these vocabulary richness indices attempt to capture the lexical repeat rate of a document.

$$4) \quad Yule's I = (M1 * M1) / ((M2 - M1),$$

where  $M1$  is the number of word types a text consists of, and  $M2$  is the sum of the products of each observed frequency to the power of two and the number of word types observed with that frequency.

Suppose that you have a very short text of 'this is my pen, but that is your pen' in which five words ('this', 'my', 'but', 'that', 'your') occur once and two words ('is', 'pen') twice. That is, this sentence has 7 word types and 9 tokens. For this text,  $M1$  is 7, and  $M2$  is 13 ( $= (5*1^2) + (2*2^2)$ ). *Yule's I* value for this text is ca. 8.1666 ( $= 7*7/(13-7)$ ).

$$5) \quad TTR = V/N,$$

where  $V$  and  $N$  are the number of different word types and the number of all tokens appearing in a given document.

Using the same example as the one given above,  $V$  is 7 and  $N$  is 9. Thus, the *TTR* value is ca. 0.7777 ( $= 7/9$ ).

$$6) \quad Honoré's R = 100 \log_{10} N / (1 - (V_1/N)),$$

where  $N$  in this equation is also the number of all tokens appearing in a given document.  $V_1$  is the number of the word types which appear only once in the document. For the same example as above,  $N$  is 9 and  $V_1$  is 5. Thus, *Honoré's R* is ca. 214.7406 ( $= 100 * \log_{10}(9) / (1 - (5/9))$ ).

If you use all of the features given in Table 1, each group of messages is modelled (= represented) as a feature vector of 12 dimensions. FTC tests were repeatedly carried out by changing feature sets with different dimensions to investigate which combination yields the best result. Since testing all possible permutations of these features with various dimensions is time-consuming, I systematically selected only some possible combinations. First of all, I tried all possible combinations of two features  $[f_1, f_2]$ , and selected the five best performing bi-features. Using these five best performing bi-features as bases, I tested the performance of the tri-features  $[f_1, f_2, f_3]$  by adding one of the remaining features one by one to these bases. I repeated this process for feature vectors of higher dimensions.

### 2.3 *Likelihood ratio calculation*

As explained in §2.2, FTC experiments were carried out maximally using the 12 different features given in Table 1. One must be cautious of the likelihood of correlation between some, many, or all of these 12 features. For example, the three features of vocabulary richness are obviously correlated with each other as they are designed to capture the same characteristic of a text; namely vocabulary richness. If one estimates the overall LR based on these three vocabulary richness features in a naïve manner – the multiplication of the LR of each of these features – without considering the correlation between them, the overall LR will be overestimated.

Aitken & Lucy (2004) addressed the problem of estimating LRs from correlated variables by deriving the multivariate kernel density LR (MVLRL) formula. Following the initial application of the formula to the data from glass fragments, it has been successfully applied to various types of forensic evidence, including fingerprint (Neumann et al. 2007), handwriting (Bozza et al. 2008) and voice (Morrison 2009b). The MVLRL formula considers the correlations of multi-features, and returns a single LR by discounting the correlations between them (refer to Aitken & Lucy (2004) for their exposition of the formula). An LR was calculated using the MVLRL formula for each of the 115 SA and 26220 DA comparisons in a cross-validated manner, which means that the message groups not used for the comparison were used together as the background data for estimating the typicality. For example, in the case of the SA comparison between

$A_1G_1$  (the first message group of the first author) and  $A_1G_2$ , all other samples of the other authors ( $A_2 \sim A_{115}$ ) were used as the background samples.

Logistic-regression calibration was also applied to the derived LRs from the MVLR formula (Brümmer & du Preez, 2006). Calibration is an affine transformation to a set of scores which involves a linear monotonic shifting and scaling to the scores relative to a decision boundary in order to minimise the magnitude and incidence of scores which are known to misleadingly support the incorrect hypothesis (Morrison 2013).

#### 2.4 Presentation of results and evaluation of performance

As pointed out in §1.2, many previous studies have treated forensic authorship analysis as a two-way classification problem. Consequently, the validity of the methodology has usually been assessed in terms of classification accuracy, such as precision, recall, equal error rate, etc. (De Vel et al. 2001, Iqbal et al. 2010b, Orebaugh & Allnutt 2009, Zheng et al. 2003). Morrison (2011: 93) argues that classification-accuracy/classification-error rates, such as equal error rate (EER), precision and recall, are inappropriate for use within the LR framework because they implicitly refer to posterior probabilities – which is the province of the trier of fact – rather than likelihood ratios – which is the province of forensic scientists – and “they are based on a categorical thresholding, error versus non-error, rather than a gradient strength of evidence ... An appropriate metric ... is the log-likelihood-ratio cost ( $C_{llr}$ )”, which is a gradient metric based on LRs. The formula for calculating  $C_{llr}$  is given in 7). It calculates the mean of two hypothesis-dependent logarithmic functions ( $i$  is for SA comparisons;  $j$  is for DA comparisons) (Brümmer & du Preez 2006).

7)

$$C_{llr} = \frac{1}{2} \left( \left[ \frac{1}{N_{H_p}} \sum_i^{N_{H_p}} \log_2 \left( 1 + \frac{1}{LR_i} \right) \right] + \left[ \frac{1}{N_{H_d}} \sum_j^{N_{H_d}} \log_2 (1 + LR_j) \right] \right)$$

In 7),  $N_{Hp}$  and  $N_{Hd}$  are the numbers of SA and of DA comparisons, and  $LR_i$  and  $LR_j$  are the LRs derived from the SA and DA comparisons, respectively. If the system is working ideally, all the SA comparisons should produce LRs greater than 1, and the DA comparisons should produce LRs less than 1, and different strengths of evidence are reflected in the magnitude of the LRs. In this approach, LRs which support counter-factual hypotheses are given a penalty. The size of this penalty is determined according to how significantly the LRs deviate from unity. That is, an LR supporting a counter-factual hypothesis with greater strength will be penalised more heavily than ones which have magnitudes closer to unity, because they are less misleading. For example, LR values of 10 and 100 for DA comparisons – which are counter-factual LRs indicating that the difference between the offender and suspect samples would be 10 and 100 times more likely to occur, respectively had they come from the same author than different authors – result in  $C_{llr}$  values of 3.459 ( $\approx \log_2(1+10)$ ) and 6.658 ( $\approx \log_2(1+100)$ ), respectively. It is clear that the latter LR value (100), which is more misleading than the former (10), is penalised more, resulting in a higher  $C_{llr}$  value. This is how  $C_{llr}$  works.  $C_{llr}$  is based on information theory; any value less than 1 means that the system is giving you information. The lower the  $C_{llr}$  value is, the better the performance of the system.

In the current study, the performance of features is ranked in terms of  $C_{llr}$ . The magnitudes of the derived LRs of the best performing features are shown using Tippett plots, which is the conventional way of showing the results of an LR-based comparison (I will explain how to read Tippett plots in §3 in which the results of the current study are given).

### 3. Results and discussions

The best performing feature set for each of the three different sample sizes is given in Table 2, along with the  $C_{llr}$  values. Table 2 also contains the system performance with all twelve features. The test results given in Table 2 show that it is not necessary to have all features included to obtain the best result. All of the different sample sizes (500, 1500 and 2500) achieved the best result with as few as four or five features (out of 12). Good features to be included regardless of the sample size are ‘average character number’ (6), ‘punctuation character ratio’ (11)

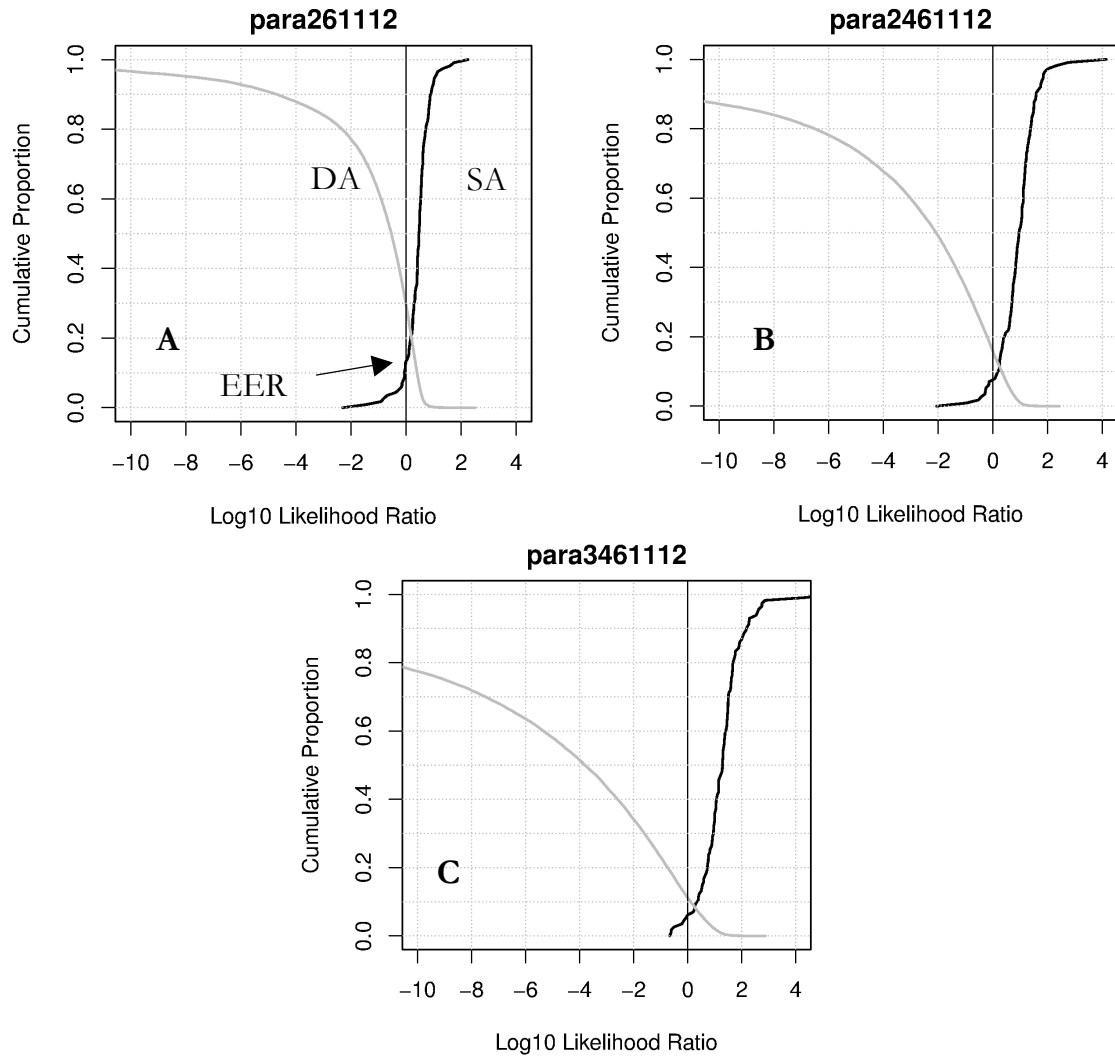
and ‘special character ratio’ (12). Features relating to vocabulary richness, such as ‘TTR’ (2) and ‘Honoré’s R’ (3), also appear to be good features. It is not surprising that, as shown in the  $C_{llr}$  values of Table 2, the performance of the system improves as a function of the sample number. The sample size of 2500 performs best with a  $C_{llr}$  value of 0.259.

Size	Features	$C_{llr}$	All Feature $C_{llr}$
500	2,6,11,12	0.604	0.637
1500	2,4,6,11,12	0.363	0.430
2500	3,4,6,11,12	0.259	0.326

**Table 2:** Performance evaluation with  $C_{llr}$ . *Size* = the number of word tokens included in each message group; *Features* = best performing feature sets. *All Feature  $C_{llr}$*  = the results using all twelve features.

The LR of the best performing features are graphically presented for the sample sizes of 500, 1500 and 2500 as Tippett plots in Figure 3. Please note that  $\log_{10}$ LR is used in Figure 3 in which case the neutral point is 0.

In the Tippett plots, the LR, which is equal to or greater than the value indicated on the x-axis, are cumulatively plotted separately for the SA (the black curve rising to the right) and DA (the grey curve to the left) comparisons. That is, for the SA comparisons, starting from the smallest LR value out of the 115 SA LRs, they are cumulatively plotted towards the highest SA LR, whereas for the DA comparison, the 26220 DA LRs are cumulatively plotted from the highest to the lowest. Tippett plots show how strongly the derived LRs not only support the correct hypothesis but also misleadingly support the contrary-to-fact hypothesis. It can be read from Figure 3A, for example, that ca. 24% of the DA comparisons have LRs greater than -2, and also that the greatest contrary-to-fact SA LR is an LR of ca. -2.3.



**Figure 3.** Tippet plots showing calibrated LR curves for the sample size of 500 (panel a); 1500 (b) and 2500 (c). Black = SA comparisons; grey = DA comparisons. Log<sub>10</sub> scale is used for LR, in which case the neutral value is 0. The crossing point of the two curves is equal error rate (EER).<sup>7</sup>

The Tippet plots also show the discriminability of a system. The crossing point between the SA and DA curves is the so-called equal error rate (EER)<sup>8</sup>, at which the error rate of the SA comparisons and that of the DA comparisons are the same. Thus, EER is commonly used as a metric showing the overall performance

<sup>7</sup> Please note that these Tippet plots are truncated at the Log<sub>10</sub> LR of -10 for stylistic reasons. That is, the actual cumulative DS LR curves extend further to the left and their cumulative proportions reach 1.0. Please also note that, for example, para261112 of panel a means that features 2, 6, 11 and 12 (refer to Table 1) were used in this experiment.

<sup>8</sup> Note that EER is used here in order not to assess and compare the performance of different systems, but to show that Tippet plots also show the discriminability of a system.



of a classification system. From Figure 3, it can be observed that the EER of the best performing feature set is ca. 20%, 11% and 8% for the sample size of 500, 1500 and 2500, respectively. That is, although it is to be expected, the discriminability of the system improves as the sample number increases.

When one compares the three Tippett plots given in Figure 3, there is a clear difference in the magnitude of the LRs in that the SA and DA curves spread further away from  $LR = 0$  with more sample numbers. Therefore, with the increase of sample size, the magnitude of the LRs that are consistent with reality become greater. A closer observation of Figure 3 further shows that the magnitude of the counter-factual SA LRs also became smaller (= less misleading LRs). The improvement in discriminability and degree of misleading LRs that were brought about by the larger sample size is the main contributor to the lower  $C_{llr}$  values for the larger sample sizes.

What can be observed consistently across the Tippett plots given in Figure 3 is that the magnitude of the SA LRs is generally weaker than the magnitude of the DA LRs. This is partly due to the nature of ‘being different and being similar’; that is, one can be different from another person in essentially unlimited ways, whereas one is so constrained when he or she tries to become similar to another person.

## 4. Conclusions

In this paper, I have attempted to concisely present the essence of FTC using chatlog messages as forensic evidence. In doing so, I have demonstrated that same-author messages can be rather well discriminated from different-author samples on the basis of the LRs estimated from lexical features, using the multivariate kernel density LR formula. I have also addressed how the performance of the FTC system and the magnitude of the LRs are influenced by the sample size, by assessing the performance with  $C_{llr}$ , and presenting the derived LRs in Tippett plots.

The modelling techniques and features used in the current study are relatively basic. Thus, I would like to try other techniques and features in order to achieve better performance, particularly in the cases involving smaller sample sizes. LR-

based FTC research is still in its early stages. A lot of fundamental research still needs to be done.

## Acknowledgements

The author would like to thank two anonymous reviewers for their valuable comments. All of their comments are reflected in the current paper. I also would like to express my gratitude to Professor Philip Rose who has continued to provide insightful comments to my forensic research. This research is partly supported financially by the ANU Research School of Asia and the Pacific.

## References

- Aitken CGG & D Lucy 2004 'Evaluation of trace evidence in the form of multivariate data' *Journal of the Royal Statistical Society Series C-Applied Statistics* 53: 109-122.
- Aitken CGG & DA Stoney 1991 *The Use of statistics in forensic science* New York; London: Ellis Horwood.
- Aitken CGG & F Taroni 2004 *Statistics and the evaluation of evidence for forensic scientists* Chichester: Wiley.
- Baayen RH 2001 *Word frequency distributions* Dordrecht; London: Kluwer Academic.
- Bozza S, F Taroni R Marquis & M Schmittbuhl 2008 'Probabilistic evaluation of handwriting evidence: Likelihood ratio for authorship' *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 57(3): 329-341.
- Brümmer N & J du Preez 2006 'Application-independent evaluation of speaker detection' *Computer Speech and Language* 20(2-3): 230-275.
- Corney MW, AM Anderson GM Mohay & O De Vel 2001 'Identifying the authors of suspect email' *Computers and Security*.
- De Vel O, A Anderson M Corney & G Mohay 2001 'Mining e-mail content for author identification forensics' *ACM Sigmod Record* 30(4): 55-64.
- Grant T 2007 'Quantifying evidence in forensic authorship analysis' *International Journal of Speech Language and the Law* 14(1): 1-25.
- Grant T 2010 'Text messaging forensics: txt 4n6: Idiolect free authorship analysis?' in AJ Malcolm Coulthard (ed.) *The Routledge handbook of forensic linguistics* New York: Routledge. pp. 508-522.
- Iqbal F, H Binsalleeh B Fung & M Debbabi 2010a 'Mining writeprints from anonymous e-mails for forensic investigation' *Digital Investigation* 7(1): 56-64.
- Iqbal F, H Binsalleeh BCM Fung & M Debbabi 2013 'A unified data mining solution for authorship analysis in anonymous textual communications' *Information Sciences*: 98-112.
- Iqbal F, LA Khan BCM Fung & M Debbabi 2010b 'E-mail authorship verification for forensic investigation' *Proceedings of the 2010 ACM Symposium on Applied Computing*: 1591-1598.
- Ishihara S 2012 'Probabilistic evaluation of SMS messages as forensic evidence: Likelihood ration based approach with lexical features' *International Journal of Digital Crime and Forensics* 4(3): 47-57.
- Khan SR, SM Nirakhi & RV Dharaskar 2012 *Author identification for e-mail forensic* Paper presented at National Conference on Recent Trends in Computing NCRTC.

- Kucukyilmaz T, BB Cambazoglu C Aykanat & F Can 2008 'Chat mining: Predicting user and message attributes in computer-mediated communication' *Information Processing & Management* 44(4): 1448-1466.
- Lambers M & CJ Veenman 2009 'Forensic authorship attribution using compression distances to prototypes' in Z Geradts KY Franke & CJ Veenman (eds) *Computational forensics* Springer Link. pp. 13-24.
- Layton R, P Watters & R Dazeley 2010 *Authorship attribution for twitter in 140 characters or less* Paper presented at the 2nd Cybercrime and Trustworthy Computing Workshop (CTC).
- Mohan A, IM Baggili & MK Rogers 2010 *Authorship attribution of SMS messages using an N-grams approach*. Paper presented at CERIAS Tech Report 2010-11, Center for Education and Research Information Assurance and Security Purdue University, USA.
- Morrison GS 2009a 'Comments on Coulthard & Johnson's (2007) portrayal of the likelihood-ratio framework' *Australian Journal of Forensic Sciences* 41(2): 155-161.
- Morrison GS 2009b 'Forensic voice comparison and the paradigm shift' *Science & Justice* 49(4): 298-308.
- Morrison GS 2011 'Measuring the validity and reliability of forensic likelihood-ratio systems' *Science & Justice* 51(3): 91-98.
- Morrison GS 2013 'Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio' *Australian Journal of Forensic Sciences* 45(2): 173-197.
- Neumann C, C Champod R Puch-Solis N Egli A Anthonioz & A Bromage-Griffiths 2007 'Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae' *Journal of forensic sciences* 52(1): 54-64.
- Orebaugh A & J Allnutt 2009 'Classification of instant messaging communications for forensic analysis' *The International Journal of Forensic Computer Science* 1: 22-28.
- Robertson B & GA Vignaux 1995 *Interpreting evidence: Evaluating forensic science in the courtroom* Chichester: Wiley.
- Rose P 2002 *Forensic speaker identification* London: Taylor & Francis.
- Saks MJ & JJ Koehler 2005 'The coming paradigm shift in forensic identification science' *Science* 309(5736): 892-895.
- Teng GF, MS Lai JB Ma & Y Li 2004 *Authorship mining for Chinese e-mail documents* Paper presented at 8th World Multi-Conference on Systemics, Cybernetics and Informatics, Vol II, Proceedings: Computing Techniques.
- Zheng R, JX Li HC Chen & Z Huang 2006 'A framework for authorship identification of online messages: Writing-style features and classification techniques' *Journal of the American Society for Information Science and Technology* 57(3): 378-393.
- Zheng R, Y Qin Z Huang & HC Chen 2003 'Authorship analysis in cybercrime investigation' *Proceedings of the 1st NSF/NIJ Conference on Intelligence and Security Informatics* 2665: 59-73.



Minerva Access is the Institutional Repository of The University of Melbourne

**Author/s:**

Ishihara, Shunichi

**Title:**

A likelihood ratio-based forensic text comparison in predatory chatlog messages

**Date:**

2014

**Publication Status:**

Published

**Persistent Link:**

<http://hdl.handle.net/11343/40956>

**File Description:**

A likelihood ratio-based forensic text comparison in predatory chatlog messages