A Forensic Authorship Classification in SMS Messages: A Likelihood Ratio Based Approach Using N-gram

Shunichi Ishihara
The Australian National University
School of Culture, History and Language
Department of Linguistics, Canberra ACT 0200 Australia
shunichi.ishihara@anu.edu.au

Abstract

Due to its convenience and low–cost, short message service (SMS) has been a very popular medium for communication for quite some time. Unfortunately, however, SMS messages are sometimes used in illicit acts, such as communication between drug dealers and buyers, extortion, fraud, scam, hoax, false reports of terrorist threats, and many more. This study is a forensic study on the authorship classification of SMS messages in the Likelihood Ratio (LR) framework with the N–gram modelling technique. The aims of this study are to investigate 1) how accurately it is possible to classify the authors of SMS messages; 2) what degree of strength of evidence (LR) can be obtained from SMS messages and 3) how the classification performance and the LRs are affected by the sample size for modelling. The resultant LRs are calibrated by means of the logistic regress calibration technique. The results of the classification tests will be rigorously assessed from different angles, using the techniques proposed for automatic speaker recognition and forensic voice comparison.

1 Introduction

We often come across news stories on so–called cyber crimes which take advantage of the high visual anonymity of, for example, email and SMS messages. In order to combat these cyber crimes, the Australian Government is currently trying to pass the Cybercrime Legislation Amendment Bill 2011 (hereafter, Cyber Law). This Cyber Law was introduced and read for the first time at the House of Representatives in June, 2011.[1] This legislation will enable police and intelligence agencies to instruct phone companies and internet carriers not to destroy sensitive information, such as text messages or emails from terrorists or criminals, that is important for investigations and prosecutions. This legislation also set the framework for Australia to join the Council of Europe Convention on Cybercrime, which more than 40 nations have joined or plan to join.

Needless to say, SMS messages, which are the focus point of the current study, hold a very important position in the above–mentioned legislation. As Grant (2007, p2) states "[o]ver recent years there has been considerable and growing interest in forensic authorship analysis", it is predicted that SMS messages will be increasingly used as evidence in Australian courts and in national and international security contexts (Coulthard and Johnson, 2007).[2] The fact that the use of mobile phones has been increasing exponentially and that the SMS is becoming a more and more common medium of communication, is apparently a strong driving force and motivation for the above–mentioned legislation and the conduct of fundamental research on SMS messages as scientific evidence.

Having said that, there is a large amount of research on authorship attribution in general (Thisted and Efron, 1987; Pennebaker and King, 1999; Doddington, 2001; Woolls, 2003; Slatcher et al., 2004)

[1] http://www.aph.gov.au/house/committee/jscc/cybercrime_bill/

[2] Some actual cases where authorship attribution was performed on SMS and email messages are given in Grant (2007) and Mohan et al. (2010).

Shunichi Ishihara. 2011. A Forensic Authorship Classification in SMS Messages: A Likelihood Ratio Based Approach Using N-gram. In *Proceedings of Australasian Language Technology Association Workshop*, pages 47−56

and on individual linguistic idiosyncrasies (Webber et al., 2002; Shriberg and Stolcke, 2008; Ishihara, 2010) whereas studies specifically focusing on the authorship of SMS messages in forensic contexts are conspicuously sparse (cf. Mohan et al. 2010).

A possible scenario in which SMS messages can be used as evidence of an incriminating act is as follows: the police authority obtained a set of incriminating messages written by a criminal while another set of messages were obtained from a suspect. The relevant parties would like to know whether these two sets of messages were actually written by the same author or different authors. We simulate this scenario in our study.

This study adopts the approach used in other forensic fields, such as DNA and speaker recognition, the Likelihood Ratio (LR)–based evidence evaluation (Aitken and Stoney, 1991; Aitken, 1995; Robertson and Vignaux, 1995; Aitken and Taroni, 2004). As we know, SMS messages are usually (very) short while the ways people write their messages are unique (e.g. the use of acronyms, shorthand, etc) (Tagg, 2009). However, to the best of our knowledge, there have not been any empirical studies on the authorship classification of SMS messages in the framework of the LR (cf. Grant, 2007; Mohan et al., 2010) . Thus, we cannot answer even some fundamental questions, such as "How well can we correctly identify two groups of messages that were written by the same author as being written by the same author, and *mutatis mutandis*, by different authors?" and "What is the degree of strength of evidence (= LR) that we are likely to obtain from SMS messages?". We attempt to provide some answers to these questions by conducting a series of simple authorship classification tests in the LR framework.

Thus, more precisely, the aims of this study are to investigate 1) how accurately it is possible to classify the authors of SMS messages; 2) what degree of strength of evidence (LR) can be obtained from SMS messages; and 3) how the performance of the authorship classification and the strength of evidence are influenced by the sample size for modelling. The resultant LRs are calibrated by means of the logistic regress calibration technique (Brümmer and du Preez, 2006). The results of the classification tests are evaluated by means of the techniques originally proposed for automatic speaker recognition and forensic voice comparison (Gonzalez–Rodriguez et al., 2007). The effect of the calibration on the LRs obtained from the SMS messages will also be discussed.

2 Likelihood Ratio–based Approach

2.1 Likelihood Ratio

In the Bayesian analysis of evidence, opinions about the hypotheses are expressed in the form of posterior probabilities (or the posterior odd which is the ratio of the two conditional probabilities) as shown in (1), where H_p = prosecution hypothesis; H_d = alternative or defence hypothesis; E = forensic evidence. In the context of the forensic authorship classification of SMS messages, E will be the similarities/differences between the offender and defendant SMS messages. Thus, the posterior odd is the ratio between the probability that the same author hypothesis (or the prosecution hypothesis) is true ($p(H_p|E)$) and the probability that the different author hypothesis (or the defence hypothesis) is true ($p(H_d|E)$), given the evidence (E).

$$\underbrace{\frac{p(H_p|E)}{p(H_d|E)}}_{\textit{posterior odds}} \tag{1}$$

The solution to (1) is Bayes' theorem as the posterior odds is the product of the prior odds (province of the court) and the likelihood ratio (province of the forensic scientist) as shown in (2).

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

It has been stressed that the task of the forensic expert is to provide the court with a strength–of–evidence statement by estimating the LR, and that they should NOT be asked their opinion about the probabilities given the evidence (= posterior odds) (Aitken and Stoney, 1991; Aitken, 1995; Robertson and Vignaux, 1995; Aitken and Taroni, 2004).

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 and Vignaux, 1995, p17). For forensic authorship classification, it will be the probability of observing the difference between the group of messages written by the offender and that written by the suspect if they have come from the same author (i.e. if the prosecution hypothesis is true) relative to the probability of observing the same evidence if they have been produced by different authors (i.e. if the defence hypothesis is true). Thus, LR can be expressed in (3).

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

The relative strength of the given evidence supporting the hypothesis is reflected in the magnitude of the LR. The more the LR deviates from unity (LR = 1; logLR = 0), the greater support for either the prosecution hypothesis (LR $>$ 1; logLR $>$ 0) or the defence hypothesis (LR $<$ 1; logLR $<$ 0). It is also common practice to express the LR logarithmically, in which case the neutral value is 0. Unless specifically expressed, $\log_{10}$LR values are used in this study.

Although the value of LR quantifies the strength of evidence, the value is not readily interpretable to the court. Thus, in order to aid the court to interpret LR values, some verbal interpretations of the ranges of LR values have been proposed. The one proposed by Champod and Evett (2000) is given in Table 1. In this study, whenever appropriate, we verbally express the strength of evidence based on Table 1.

2.2 Likelihood Ratios in Forensic Science

LRs show many advantages for evidence evaluation and presentation (Robertson and Vignaux, 1995, p21). Firstly, the majority of evidence submitted to the court is by nature only indicative, not determinative. For the indicative nature of evidence, which means, in other words, forensic evidence has an uncertain nature, probability is ideal to use in the inference process in a scientific way.

Another reason is that the role of forensic experts is clearly defined in the legal system, with the decision on whether or not the defendant is guilty not being left to the forensics experts: this is the job of juries (or judges in some judicial systems). Thus, as expressed in §2.1, the task of the forensic expert is NOT to provide their opinion about the source of evidence, but to estimate and give the court the strength of the evidence in the form of an LR.

Besides the appropriateness for the legal system as explained above, LRs have another advantage in evidence presentation: they allow evidence from different sources (e.g. voice, blood-stain) to be combined to give an overall LR in support of a hypothesis.

According to *Daubert*,[3] any scientific and technical evidence needs to satisfy certain criteria to be admitted in court. These criteria can be summarised as the issues of *transparency* and *testability*. It has been well discussed that the use of LR for evidence evaluation and presentation is appropriate from the view points of *transparency* and *testability* (Gonzalez–Rodriguez et al., 2007).

3 Authorship Classification Tests: Methodology

3.1 Database

In this study, we use the SMS corpus compiled by the National University of Singapore (the NUS SMS corpus).[4] A new version of the NUS SMS corpus has been released almost monthly, and we use *version 2011.05.11* which contains 38193 messages collected from 228 authors. The top three countries that contributed the most to the NUS SMS corpus by the number of messages are Singapore, India and the USA. 69% of the total messages were written by native speakers of English; 30% by non–native; 1% unknown. Male authors account for 71%; female for 16%; unknown for 13%. The average length of a message is 13.8 words with punctuations (sd = 13.5; max = 231; min = 1).

3.2 Selection of Messages

In authorship classification tests, two types of author pairs—same author pairs and different author pairs—are compared and evaluated using an LR as discriminant function. The former same author pairs are used for so–called *Same Author Comparison* (SA comparison) where two groups of messages produced by the same author need to be correctly identified as the same author whereas the latter different author pairs are for *mutatis mutandis*, *Different Author Comparison* (DA comparison). Thus, we

[3] *Daubert v. Merrel Dow Pharmaceuticals Inc., 509 US 593 (1993).*

[4] http://wing.comp.nus.edu.sg:8080/SMSCorpus/

LR	Log$_{10}$ equivalent	Possible verbal equivalent	
> 10000	> 4	Very strong . . .	support for the prosecution hypothesis
1000 to 10000	3 to 4	Strong . . .	
100 to 1000	2 to 3	Moderately strong . . .	
10 to 100	1 to 2	Moderate . . .	
1 to 10	0 to 1	Limited . . .	
1 to 0.1	0 to -1	Limited . . .	support for the defence hypothesis
0.1 to 0.01	-1 to -2	Moderate . . .	
0.01 to 0.001	-2 to -3	Moderately strong . . .	
0.001 to 0.0001	-3 to -4	Strong . . .	
< 0.0001	< -4	Very strong . . .	

Table 1: Verbal equivalents of LRs and Log$_{10}$LRs (Champod and Evett, 2000).

need two groups of messages from each of the authors in authorship classification tests.

As explained in §1, one of the aims of this study is to investigate how the performance of the authorship classification and the strength of evidence are influenced by the sample size, i.e. the number of message words used for modelling. It can be safely predicted that the more messages we can use, the better the performance will be. However, each SMS message is essentially short, and it is forensically unrealistic to conduct experiments using thousands of messages to model an author's attribution. Thus, as shown in Table 2, we created 15 different datasets (DS) in which the number of words appearing in each message group is different (N = 200, 400, . . . 2800, 3000 words).

For DS200, each message group contains a total of approximately 200 words. Since we cannot control the number of the words appearing in one message, it needs to be *approximately* 200 words. In order to compile a message group of about 200 words, we added one message by one message from the chronologically sorted messages to the group until the word number reached more than 200 words. As explained earlier, we need two groups of messages from the same author. For one message group, we started from the top of the chronologically sorted messages while for the other of the same author, from the bottom so that the two groups of messages from the same author are non–contemporaneous. Thus, the topics of the messages belonging to one group are likely to be different from those belonging to the other.

DS+N	auths.	SA	DA
DS200	85	85x2	14280x2
DS400	68	68x2	9112x2
DS600	56	56x2	6160x2
DS800	49	49x2	4704x2
DS1000	43	43x2	3612x2
DS1200	41	41x2	3280x2
DS1400	38	38x2	2812x2
DS1600	37	37x2	2664x2
DS1800	35	35x2	2380x2
DS2000	34	34x2	2244x2
DS2200	31	31x2	1848x2
DS2400	28	28x2	1512x2
DS2600	25	25x2	1200x2
DS2800	24	24x2	1104x2
DS3000	24	24x2	1104x2

Table 2: Dataset (DS) configurations: sample size (N) = the number of words included in each message group; auths. = the number of authors appearing in the DS; SA = number of SA comparisons; DA = number of DA comparisons.

It should be noted that the number of messages which were contributed by each author to the NUS SMS Corpus is not the same: some contributed hundreds of messages, but some just one. Thus, some authors may not have enough messages to create two groups of messages as specified by the sample size. The second column of Table 2 shows the number of authors included in each DS. According to the second column, the number of authors included in the DSs decreases as the sample size increases. For example, as for DS3000, two sets of 3000 word messages can be created only from 24 authors. For DS3000, 24x2 same author (SA) comparisons and

1104x2 different author (DA) comparisons are possible.

3.3 Tokenisation and N-grams

The SMS messages were tokenised using the *SimpleTokenizer* function of the *opennlp-tools version 1.5.0*[5] without any stemming algorithms. The *SimpleTokenizer* provides simple tokenisation based on space and punctuations.

In some cases, it is difficult to automatically locate a sentence boundary in SMS messages as the use of upper/lower cases, punctuation, space, etc do not always conform to the standard orthographic rules. Therefore, the words appearing in the same message were treated as a sequence of words, without parsing them into sentences in this study.

We use the *ngram-count* and *ngram* functions of the *Speech Technology and Research Laboratory Language Modelling Toolkit (SRLM)*[6] in this study. As explained in §3.2, we need to compare two groups of messages many times. The *ngram–count* function is used to build an N–gram language model for a group of messages (model group). The resultant N–gram language model should represent the characteristics of this particular group of messages. The *ngram* function is used to calculate log probabilities between the N–gram language model of a given group of messages (model group) and another given group of messages (test group). The log probabilities calculated by the *ngram* function show the degree of similarities/differences between the former group of messages which were modelled in the form of the N–grams (model group) and the latter group of messages (test group). The backoff technique was used for the calculation of log probabilities (Jurafsky and Martin, 2000).

An 'open–vocabulary' N–gram language model (N = 1,2,3) was built for each group of messages. The minimal count of N–grams was set as > 9, which is the default setting of the *SRLM toolkit*. Thus, all N–grams with frequency of < 9 was discounted to 0. This is based on the results of some test experiments, in which the classification performance did not significantly improve with the threshold being set as ≥ 5. The default Good–Turing discounting was used for smoothing.

3.4 Likelihood Ratio Calculation

There are some different formulae proposed for calculating LRs (Lindley, 1977; Doddington, 2001; Aitken and Lucy, 2004). In this study, a conventional $\log_{10}$LR was estimated using the formula given in (4) (Doddington, 2001).

$$LR_{i,j} = \frac{log_{10}\frac{\Lambda^{i}_{author}(j)}{\Lambda_{background}(j)}}{N_j} \tag{4}$$

Thus, the $LR_{i,j}$ of the test message group (j) against the model message group (i) is defined to be the log ratio of the similarity between the test message group (j) and the author model (Λ^{i}_{author}) of the model message group (i) to the typicality of the test message against the background author model ($\Lambda_{background}$), normalised by the number of words appearing in the test message group (N_j). The background author model was built in the cross–validated manner, using all messages appearing in the NUS SMS corpus, except those in comparison. The configurations of the N–grams for the background author model are the same as those used for the model message group.

The calculated raw LRs were calibrated using linear logistic regression using the *FoCal toolkit*[7]. Calibration is an affine transformation to a set of scores (e.g. LRs) 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 et al., 2011).

3.5 Evaluation

In this study, the results of the authorship classification tests are rigorously assessed using the equal error rate (*EER*), the Tippett plot, and the *log–likelihood–ratio cost* or C_{llr} matrices (Brümmer and du Preez, 2006). Using LR values as discriminant scores, we can measure the accuracy of the authorship classification systems in terms of *EER*. *EER* is a good indicator of the overall accuracy of a system, but does not refer to how *good* the LR values are. An LR is an estimate of the *degree* of support for a

[5] http://incubator.apache.org/opennlp/

[6] http://www.speech.sri.com/projects/srilm/

[7] http://www.dsp.sun.ac.za/ nbrummer/focal/

hypothesis against its alternative. Thus, the value of an LR itself is very important.

The Tippett plots show the distributions of the LRs given the prosecution hypothesis and the defence hypothesis, respectively together. Useful information that the Tippett plots can graphically provide is not only how strongly the LRs support the correct hypotheses but also how strongly the LRs support the incorrect hypotheses. More detailed explanations will be given about the Tippett plots when the results of the classification tests are presented in §4.

In short, the Tippett plots are graphical representations of the 'goodness of LRs' (Brümmer and du Preez, 2006). However, they do not give a scalar value of this goodness. The solution for this problem is the *log–likelihood–ratio cost function* or C_{llr} (5), which is a measure proposed in the area of automatic speaker recognition (Brümmer and du Preez, 2006),

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{H_p}} \sum_{i for H_p = true}^{N_{H_p}} \log_2\left(1 + \frac{1}{LR_i}\right) + \frac{1}{N_{H_d}} \sum_{j for H_d = true}^{N_{H_d}} \log_2\left(1 + LR_j\right)\right) \quad (5)$$

where N_{H_p} and N_{H_d} are the number of LR values in the evaluation set for the prosecution hypothesis H_p being true or the defence hypothesis H_d being true. As can be seen from (5), incorrect LR values (i.e. same author comparisons with LR $<$ 1; logLR $<$ 0 and different author comparisons with LR $>$ 1; logLR$<$ 0) will have a strong penalty (high C_{llr}) and *vice versa*. The lower the C_{llr} value is, the better the performance of the system is. C_{llr} can be split into a discrimination loss (C_{llr}^{min})—which is the value achievable after the application of a calibration procedure—and a calibration loss (C_{llr}^{cal}) ($C_{llr} = C_{llr}^{min} + C_{llr}^{cal}$). Thus, the C_{llr} can provide an overall evaluation of a system while the C_{llr}^{min} and C_{llr}^{cal} can specifically show how the discrimination loss and the calibration loss contributed to the overall performance of the system. The *FoCal toolkit* is used to calculate C_{llr} in this study.

4 Authorship Classification Tests: Results and Discussions

The results of the authorship classification tests with different sample sizes are given in Table 3 in terms of *EER*, C_{llr}, C_{llr}^{min} and C_{llr}^{cal}.

DS+N	*EER*	C_{llr}	C_{llr}^{min}	C_{llr}^{cal}
DS200	0.40	1.29	0.96	0.33
DS400	0.39	1.14	0.93	0.21
DS600	0.37	1.08	0.90	0.18
DS800	0.36	1.04	0.87	0.16
DS1000	0.32	0.99	0.84	0.14
DS1200	0.30	0.97	0.82	0.15
DS1400	0.30	0.94	0.78	0.15
DS1600	0.30	0.93	0.77	0.15
DS1800	0.28	0.90	0.78	0.12
DS2000	0.23	0.87	0.72	0.14
DS2200	0.20	0.86	0.68	0.17
DS2400	0.21	0.84	0.65	0.18
DS2600	0.20	0.81	0.67	0.14
DS2800	0.20	0.82	0.67	0.15
DS3000	0.20	0.80	0.62	0.17

Table 3: The results of the authorship classification tests are given in terms of *EER*, C_{llr}, C_{llr}^{min}, C_{llr}^{cal} with 15 different sample sizes (N).

With respect to *EER*, C_{llr} and C_{llr}^{min}, the results of the authorship classification tests improve as the sample size increases. However, the C_{llr}^{cal} values do not show much improvement after a sample size of 400. When the sample size is greater than 400, the C_{llr}^{cal} values fluctuate between 0.12 and 0.18. That is, the degree of calibration is more or less stable with a sample size of 600 or greater, and the C_{llr} values improve as the sample size increases because the discrimination (not calibration) performance improves. The C_{llr}^{cal} values given in Table 3 are fairly small, even with a sample size of 200. That means that the LRs obtained from SMS messages are well calibrated.

The accuracy of the authorship classification increases from c.a. 60% with a sample size of 200 words to c.a. 80% with a sample size between 2200∼3000 words. As can be judged from these accuracies, SMS messages carry some idiosyncratic information about the authors. The best result was achieved with a sample size of 3000 in terms of C_{llr}

(0.80).

To the best of our knowledge, Mohan et al. (2010) is the only study on authorship attributions of SMS messages, having an application to forensics in mind. They reported in their study, in which the NUS SMS corpus and an N–gram technique were also used, that the author of an SMS message could be correctly predicted with an accuracy of 65%∼70%. Their reported accuracy is comparable with that of the current study. However, what their study lacks is the reference to the strength of evidence (or LRs) as they did not employ the likelihood ratio based approach.

Figure 1 contains the Tippets plots of the LRs obtained with a sample size of 200 (panel 1), 1000 (2), 2000 (3) and 3000 (4). Figure 1 graphically shows how the 'goodness of the LRs' changes with the increase in sample size. The LRs, which are equal to or greater than the value indicated on the x-axis, are cumulatively plotted separately for the SA comparisons (black) and the DA comparisons (grey). In Figure 1, both uncalibrated (dotted curves) and calibrated (solid curves) LRs are included. The calibrated LRs were obtained by the logistic–regression calibration procedure which is a linear monotonic transformation, using the *FoCal toolkit*. Calibration aims to present the relevant information in such a way that the fact finder makes appropriate decisions (Ramos–Castro, 2007).

It can be observed from Figure 1 that before calibration, the crossing points of the SA and DA LRs (dotted curves) are slightly off from $\log_{10}$LR = 0, whereas, after calibration, the crossing points (solid curves) are right on Log_{10}LR = 0. Theoretically speaking, the crossing point of the SA and DA LRs should align with $\log_{10}$LR = 0 even before calibration.

The logistic–regression calibration brought different effects on the LR values. When the sample size is small (i.e. 200 and 1000), the calibration has resulted in a major reduction in LR values (both correct and incorrect LRs). This major reduction of the LRs resulted in the calibrated LRs being not very meaningful as evidence. The ranges of the calibrated LRs are from -0.220 to 0.439 for the SA comparisons and from -0.281 to 0.443 for the DA comparisons with a sample size of 200 (Figure 1–1). According to Table 1 in which the verbal interpretations of LR values are given, the LRs between 0 and 1 for the SA comparisons and those between -1 and 0 for the DA comparisons provide only "limited" support for the prosecution and defence hypothesis, respectively.

Even with a sample size of 1000 (Figure 1–2), almost all of the calibrated LRs fall in the range of between -1 and 1. That is, again, the calibrated LR values give only "limited" support for either hypothesis.

With a sample size of 3000 (Figure 1–4), the calibration leads to the enhancement of the LRs: the ranges of the calibrated LRs are 2.868 (from -0.657 to 2.211) and 4.711 (from -2.735 to 1.976) for the SA and DA comparisons, respectively, which are much larger than the ranges of the uncalibrated LRs: 1.606 (from -0.184 to 1.422) and 2.640 (-1.349 to 1.291) for the SA and DA comparisons, respectively. The strongest calibrated LR values are 2.211 and -2.735 for the SA and DA comparisons, respectively. These values can be quoted as showing "moderately strong" support for the same and different author hypothesis, respectively.

Approximately 10% of the same author LRs "moderately" or "moderately strongly" support the same author hypothesis and approximately 65% have only "limited" support for the same author hypothesis. Likewise, approximately 15% of the different author LRs have "moderate" or "moderately strong" support for the different author hypothesis and approximately 60% have only "limited" support for the different author hypothesis.

The downside of this enhancement in LR values with a large sample size (i.e. 3000 words) is that the misleading LRs also increased their values after calibration. For example, the most misleading uncalibrated LR value for the DA comparisons is LR 1.291, which is incorrectly in favour of the same author hypothesis. After calibration, this misleading LR was intensified to LR 1.976. This value could be presented in court by a forensic expert as "moderately" supporting the same author hypothesis. This is a grave concern.

Considering the fact that SMS messages are usually (very) short, it may not be forensically realistic to be able to use as many as 3000 words for SMS authorship classification. Please note that the average length of a message is 13.8 words in the NUS

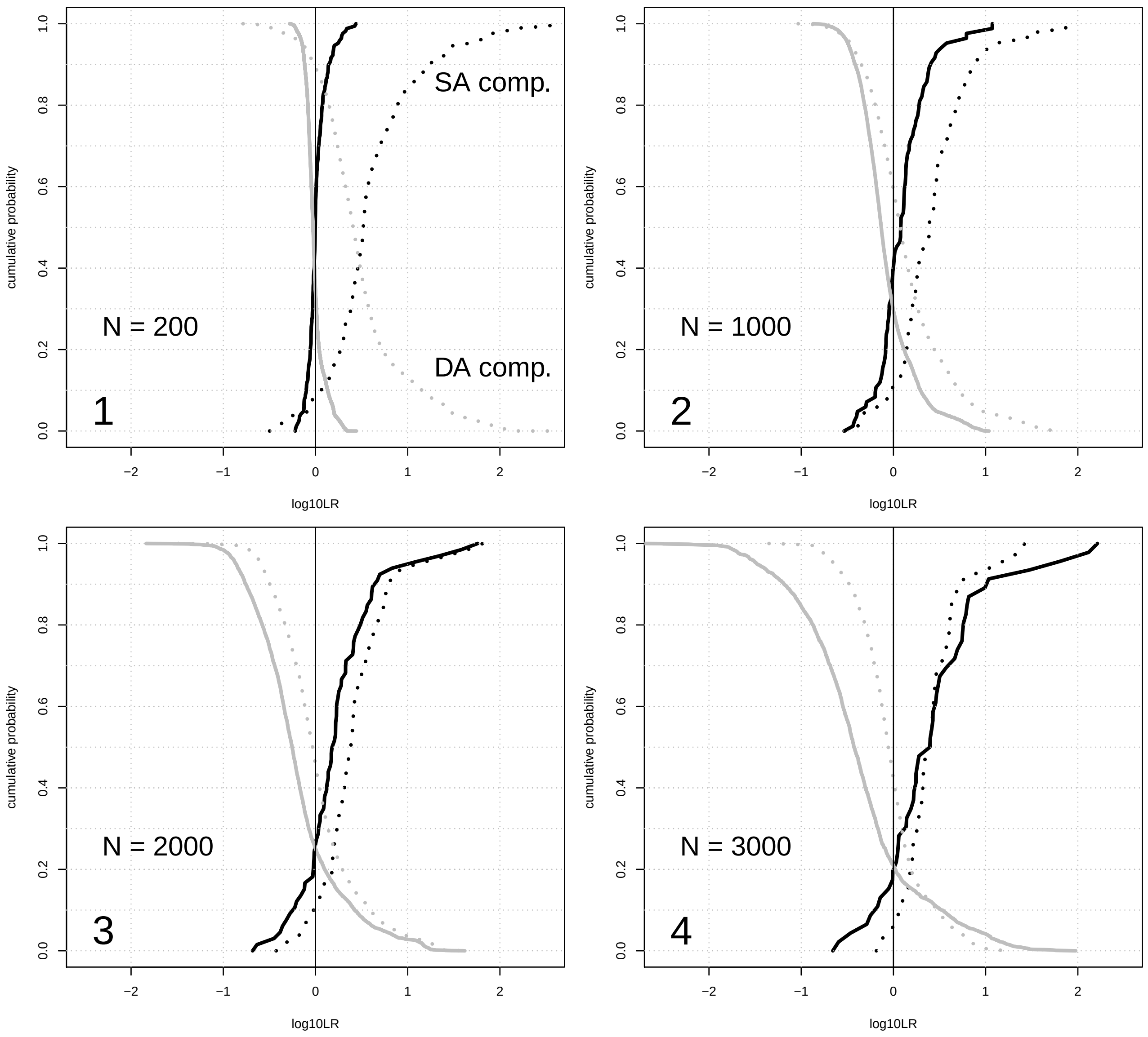

Figure 1: Tippet plots showing uncalibrated (solid curves and calibrated (dotted curves) LRs for the sample size (N) of 200 (panel 1); 1000 (2); 2000 (3) and 3000 (4). Grey = same author (SA) comparisons; black = different author (DA) comparisons.

SMS corpus, and therefore about 218 messages are required to be equivalent to 3000 words. However, our results demonstrated that if the sample size is small ($\leq$ 1000), having real cases in mind, the obtained LRs only give "limited" support for either hypothesis.

5 Conclusions

We found out that 1) the classification accuracy reaches c.a. 80% when we use a sample size of 2200 words or more; 2) the calibrated LR values are very weak, in particular when the sample size is small ($\leq$ 1000), in that the LR values provide only "limited" support for either hypothesis; 3) when we use a large sample size (i.e. 3000), the approximately 10$\sim$15% of the calibrated LR values provide "moderately strong" support for either of the correct hypotheses whereas the calibration undesirably increases the values of the misleading LRs as well.

6 Future Studies

The techniques we employed are rather simple and standard. Therefore, there is some room whereby the classification accuracy and the magnitude of the

LRs can improve even with a small set of messages if we apply different techniques. For this purpose, we should try different techniques at all different stages of the authorship classification (i.e. focus on specific words/expressions which are high in idiosyncrasy, pre–process of messages prior to modelling, different modelling techniques, different LR calculation techniques) to see how much we can improve the results of the authorship classification.

In order to estimate the strength of evidence as an LR, a background sample from the relevant population—in other words, the potential population of offenders—is essential. The SMS messages included in the NUS SMS Corpus are largely from Singaporeans. If we know that the offender is Singaporean, the SMS messages which were contributed by Singaporeans are appropriate as a background population data and desirable to estimate the accurate strength of evidence in LRs. However, if we know that the criminal is an Australian person, the use of this corpus is not suited in order to estimate the strength of evidence. Thus, in order to operate a forensic SMS authorship classification analysis in real cases, and calculate an LR as accurately as possible, the choice of appropriate population data is important. However, it goes without saying that this is difficult in many cases due to the lack of appropriate corpora. In the context of Australia, we lack a corpus of SMS messages written by Australians, which prevents forensic scientists from using SMS messages as evidence and limits the fundamental forensic studies on authorship classification in SMS messages. Thus, a compilation of a relevant corpus is an urgent task in Australia.

Acknowledgments

This study was financially supported by the Research School of Asia and the Pacifc, ANU. The author thanks anonymous reviewers for their valuable comments.

References

C. G. G. Aitken. 1995. *Statistics and the Evaluation of Evidence for Forensic Science*. UK, Chichester: Wiley.

C. G. G. Aitken and D. Lucy. 2004. Evaluation of trace evidence in the form of multivariate data, *Applied Statistics*. 53(4):109–122.

C. G. G. Aitken and D. A. Stoney. 1991. *The Use of Statistics in Forensic Science*. UK, Chichester: Ellis Horwood.

C. G. G. Aitken and F. Taroni. 2004. *Statistics and the Evaluation of Forensic Evidence for Forensic Scientist*. UK, Chichester: Wiley.

N. Brümmer and J. du Preez. 2006. Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2–3):230–275.

C. Champod and I. W. Evett. 2000. Commentary on A. P. A. Broeders (1999) 'Some observations on the use of probability scales in forensic identification, Forensic Linguistics 6(2):228–41', *Forensic Linguistics*. 7:238–243.

M. Coulthard and A. Johnson. 2007. *An Introduction to Forensic Linguistics: Language in Evidence*. London, New York: Routledge.

G. Doddington. 2001. Speaker recognition based on idiolectal differences between speakers, *Proceedings of the Eurospeech 2001*, 2521–2524.

J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano, J. Ortega-Garcia. 2007. Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition, *IEEE Transactions on Audio, Speech, and Language Processing*, 15:2104–2115.

T. Grant. 2007. Quantifying evidence in forensic authorship analysis. *The International Journal of Speech, Language and the Law*, 14(1):1–25.

S. Ishihara. 2010. Variability and consistency in the idiosyncratic selection of fillers in Japanese monologues: Gender differences, *Proceedings of the ALTA Workshop 2010*, 9–17.

D. Jurafsky and J. H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prince–Hall, Inc.

D. V. Lindley 1977. A problem in forensic science. *Biometrika*, 64:207–213.

A. Mohan, I. M. Baggili, M. K. Rogers. 2010. Authorship attribution of SMS messages using an N-grams approach. *CERIAS Tech Report 2011–11*.

G. S. Morrison, C. Zhang, P. Rose. 2011. An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. *Forensic Science International*. 208:59–65.

J. W. Pennebaker and L. A. King. 1999. Linguistics styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296–1312.

D. Ramos–Castro. 2007. *Forensic Evaluation of the Evidence Using Automatic Speaker Recognition Systems*. A PhD thesis, Universidad Politécnica de Madrid.

B. Robertson and G. A. Vignaux. 1995. *Interpreting Evidence*. UK, Chichester: Wiley.

E. Shriberg and A. Stolcke. 2008. The case for automatic higher–level features in forensic speaker recognition, *Proceedings of Interspeech 2008*, 1509–1512.

R. Slatcher, C. Chunga, J. Pennebaker and L. Stone. 2004. Winning words: Individual differences in linguistic style among U.S. presidential and vice presidential candidates. *Journal of Research in Personality*, 41(1):63–75.

C. Tagg. 2009. *A Corpus Linguistic Study of SMS Text Messaging*. A PhD thesis, the University of Birmingham.

R. Thisted and B. Efron. 1987. Did Shakespeare write a newly–discovered poem? *Biometrika*, 74(3):445–455.

F. Weber, L. Manganaro, B. Peskin and E. Shriberg. 2002. Using Prosodic and lexical information for speaker identification, *Proceedings of the ICASSP 2002*, 141–144.

D. Woolls. 2003. Better tools for the trade and how to use them. *Forensic Linguistics. The International Journal of Speech, Language and the Law*, 10(1):102–112.