A Forensic Text Comparison in SMS Messages: A Likelihood Ratio Approach with Lexical Features

S. Ishihara

The Australian National University, Department of Linguistics, Australia
e-mail: shunichi.ishihara@anu.edu.au

Abstract

Due to its convenience and low cost, short message service (SMS) has been a very popular medium of communication for quite some time. Unfortunately, however, SMS messages are sometimes used for reprehensible purposes, e.g. communication between drug dealers and buyers, or in illicit acts such as extortion, fraud, scams, hoaxes, and false reports of terrorist threats. In this study, we perform a likelihood-ratio-based forensic text comparison of SMS messages focusing on lexical features. The likelihood ratios (LRs) are calculated in Aitken and Lucy's (2004) multivariate kernel density procedure, and are calibrated. The validity of the system is assessed based on the magnitude of the LRs using the log-likelihood-ratio cost (C_{llr}). The strength of the derived LRs is graphically presented in Tippett plots. The results of the current study are compared with those of previous studies.

Keywords

SMS messages, forensic text comparison, likelihood ratio, multivariate kernel density, log-likelihood-ratio cost, calibration

1 Introduction

Due to a continuous increase in the use of mobile phones, the short message service (SMS) is more and more becoming a common medium of communication. Unfortunately, its convenience, low cost and high visual anonymity can be exploited, with SMS messages sometimes used in, for example, communication between drug dealers and buyers, or illicit acts such as, extortion, fraud, scams, hoaxes, false reports of terrorist threats, and many more. SMS messages have been reportedly used as evidence in some legal cases (Cellular-news, 2006; Grant, 2007), and it is not difficult to predict that the use of SMS messages as evidence will increase.

That being said, there is a large amount of research on forensic authorship analysis in other electronically-generated texts, such as emails (De Vel *et al.*, 2001; Iqbal *et al.*, 2008), whereas forensic authorship analysis studies specifically focusing on SMS messages are conspicuously sparse (cf. Ishihara, 2011; Mohan *et al.*, 2010).

The forensic sciences are experiencing a paradigm shift in the evaluation and presentation of evidence (Saks and Koehler, 2005). This paradigm shift has already happened in forensic DNA comparison. Saks and Koehler (2005) fervently suggest that other forensic comparison sciences should follow forensic DNA comparison,

which adopts the likelihood-ratio framework for the evaluation of evidence. The use of the likelihood-ratio framework has been advocated in the main textbooks on the evaluation of forensic evidence (e.g. Robertson and Vignaux, 1995) and by forensic statisticians (e.g. Aitken and Stoney, 1991; Aitken and Taroni, 2004).

Thus, emulating forensic DNA comparison, the current study is a forensic comparison of SMS messages using the likelihood-ratio framework. Focusing on the lexical features of SMS messages, we test a forensic text comparison system. The validity of the system is assessed using the log-likelihood-ratio-cost function (C_{llr}) which was originally developed for use in automatic speaker recognition systems (Brümmer and du Preez, 2006), and subsequently adopted in forensic voice comparison (Morrison, 2011). The strength of likelihood ratios (= strength of evidence) obtained from SMS messages is graphically presented using Tippett plots.

2 Forensic Authorship Analysis

2.1 Profiling, Identification and Verification

Forensic authorship analysis can be broadly classified into the subfields of *authorship profiling*, *authorship identification* and *authorship verification*. Commonly-held descriptions of the tasks of these subfields are summarised below:

- *Authorship profiling* summarises the sociolinguistic characteristics, such as gender, age, occupation, educational and cultural background, of the unknown author (offender) of the (illicit) document in question (Stamatatos, 2009).
- The task of (forensic) *authorship identification* is to identify the most likely author (suspect) of a given (incriminating) document from a group of candidate authors (suspects) (Iqbal *et al.*, In Press).
- The task of (forensic) *authorship verification* is to determine or verify if a target author (suspect) did or did not write a specific (incriminating) document (Halteren, 2007).

Using the conventional terminology, the current study is one of forensic authorship *verification*.

2.2 Role of Forensic Expert

Commonly-held views about forensic authorship analysis have been summarised above. However, it is important to explicitly state here that the forensic scientist as a witness is NOT in a position, either legally or logically, to identify, confirm, decide or even verify if two samples (one associated with the offender and the other with a suspect) are from the same person or different people (Robertson and Vignaux, 1995). This is the task of the trier of fact, who can be the judge, the panel of judges, or the jury, depending on the legal system of a country. That is, the ultimate decision as to, for example, whether the author of a document in question is a suspect or not, does not lie with the forensic expert, but with the court. When a forensic scientist

presents evidence, it is important that he/she should not violate the province of the trier of fact, and he/she should not even be asked his/her opinion on the likelihood that, for example, it is the suspect who wrote the text in question (Doheny, 1996).

So, what is the role of the forensic scientist? Aitken and Stoney (1991), Aitken and Taroni (2004) and Robertson and Vignaux (1995) state that the role of forensic scientist is to estimate the strength of evidence. That is to say,

> "… the task of 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 the known and questioned samples to arise under the hypothesis that they have the same origin than under the hypothesis that they have different origins?" (Morrison, 2009).

The strength of evidence which is the main concern of forensic scientists is technically termed as likelihood ratio (LR).

3 Likelihood-Ratio Approach

The task of the forensic expert is to provide the court with a strength-of-evidence statement by estimating the likelihood ratio. What, then, is the likelihood ratio?

3.1 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 and Vignaux, 1995). Thus, the LR can be expressed in (5).

$$LR = \frac{p(E|H_p)}{p(E|H_d)} \qquad (5)$$

For forensic authorship analysis, it will be the probability of observing the difference (referred to as the evidence, E) between the group of texts written by the offender and that written by the suspect if they have 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 have 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; logLR = 0), the greater support for either the prosecution hypothesis (LR > 1; logLR > 0) or the defence hypothesis (LR < 1; logLR < 0).

3.2 Related Studies

To the best of our knowledge, Grant (2007) and Ishihara (2011) are the only studies on forensic authorship analysis based on the LR framework: the former is on email and the latter on SMS. The results of the present study will be compared with those of Ishihara (2011) in which the same dataset and evaluation procedures were utilised, but author attribution was modelled by N-grams.

4 Testing

4.1 Scenario

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. Needless to say, the task of the forensic expert is to provide the court with a strength-of-evidence statement (in other words, LR) so as to assist the trier of fact to make a decision as to whether the suspect is guilty or not.

4.2 Database

In this study, we use the SMS corpus compiled by the National University of Singapore (the NUS SMS corpus) (http://wing.comp.nus.edu.sg:8080/SMSCorpus). A new version of the NUS SMS corpus has been released almost monthly, and we use *version 2011.05.11* which contains 38193 messages from 228 authors. 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 (sd = 13.5; max = 231; min = 1).

4.3 Selection of Messages

Two message types of author pairs – same-author pairs and different-author pairs – are necessary to assess a forensic text comparison system. The same author pairs are used for so-called *Same Author Comparison* (SA comparison) where two groups of messages produced by the same author are expected to receive the desired LR value given the same-origin, whereas the different author pairs are for *mutatis mutandis*, *Different Author Comparison* (DA comparison). Thus, we need two groups of messages from each of the authors.

This study also investigates how the performance of the system and the strength of evidence (= LR) 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 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 41, we created 4 different

datasets (DS) in which the number of words appearing in each message group is different (N = 200, 1000, 2000 and 3000 words). For DS200, each message group contains a total of approximately 200 words. Since we cannot perfectly control the number of words appearing in one message, it needs to be *approximately* 200 words.

DS+N	auths.	SA	DA
DS200	85	85	14280
DS1000	43	43	3612
DS2000	34	34	2244
DS3000	24	24	1104

Table 4: Dataset (DS) configuration: 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.

In order to compile a message group of about 200 words, we added messages one by one 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.

4.4 Features

Following the results of previous authorship studies (De Vel *et al.*, 2001; Iqbal *et al.*, 2010; Zheng *et al.*, 2006), and given the general characteristics of SMS messages (Tagg, 2009), the features listed in Table 52 are used in the current study.

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

Table 5: List of features.

All features listed in Table 52 are, in a broad sense, lexical features. They can be further sub-classified into the features of *vocabulary richness*, *word-based lexical* and *character-based lexical features*. All feature values are normalised. Features related to sentences and paragraphs are not used in this study as in many cases it is difficult to automatically locate a sentence or a paragraph boundary in SMS messages since the use of upper/lower cases, punctuation, space, etc. does not always conform to standard orthographical rules.

Different combinations of features listed in Table 52 are tested to see what combination yields the best results. However, since testing all possible permutations of these features with various dimensions of a feature vector is time-consuming, we systematically selected only some possible combinations. First of all, we 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, we 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. We repeated this process for feature vectors of higher dimensions.

4.5 Likelihood Ratio Calculation

It is straightforward to combine multiple LRs from different evidence types or features by applying Bayes' Theorem, providing they are not correlated. This is a significant feature of the LR approach as most cases involve many different types of evidence. However, it is obvious that the features listed in Table 52 are correlated in one way or another, thus a simple combination is inappropriate. Aitken and Lucy (2004) addressed the problem of estimating LRs from correlated variables by deriving the multivariate kernel density LR (MVLR) formulae. Following the initial application of the formulae to the data from glass fragments, it has been successfully applied to forensic voice comparison, in particular with acoustic-phonetic features. Please refer to Aitken and Lucy (2004) for the exposition of the MVLR formulae.

A logistic-regression calibration was applied to the derived LRs from the MVLR formulae (Brümmer and du Preez, 2006). Given two sets of LRs derived from the SA and DA comparison pairs and a decision boundary, calibration is a normalisation procedure involving linear monotonic shifting and scaling of the LRs relative to the decision boundary so as to minimise a cost function (see §4.6).

4.6 Evaluation of Performance

Morrison (2011) argues that classification-accuracy/classification-error rates, such as equal error rate, 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 threshholding, 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. See (6) for calculating C_{llr} (Brümmer and du Preez, 2006). In (6), 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 producing good quality LRs, all the SA comparisons should produce LRs greater than 1, and the DA comparisons should produce LRs less than 1. 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 the neutral point. That is, an LR supporting a counter-factual hypothesis with greater strength will be penalised more heavily than the ones which have the strength close to the unity, because they are less misleading. The lower the C_{llr} value is, the better the performance is.

$$C_{llr}=\frac{1}{2}\left(\begin{array}{l}\frac{1}{N_{H_p}}\sum_{i\ for H_p=true}^{N_{H_p}}\log_2\left(1+\frac{1}{LR_i}\right)+\\ \frac{1}{N_{H_d}}\sum_{j\ for H_d=true}^{N_{H_d}}\log_2\left(1+LR_j\right)\end{array}\right) \qquad (6)$$

C_{llr} can be split into a discrimination loss (C_{llr_min}) – which is the value achievable after the application of a calibration procedure (see §4.5) – 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* (http://www.dsp.sun.ac.za/~nbrummer/focal/) is used to calculate C_{llr} in this study. Since C_{llr_min} is the theoretically best C_{llr} value of an optimally calibrated system, the performance of the system was assessed based on the C_{llr_min} values.

C_{llr} provides a scale value which shows the overall performance of a system. A Tippett plot is a graphical presentation which provides more detailed information about the derived LRs. A more detailed explanation of Tippett plots is given in §5.

5 Results and Discussions

The test results given in Table 63 show that it is not necessary to have all features included to obtain the best result. All of the DSs achieved the best result with as few as four or five features (out of 12). The features of vocabulary richness, in particular 'Yule's K' (1) and 'TTR' (2), are good features to be included regardless of the sample size. Other robust features are 'digit ratio' (8), 'average character number' (10) and 'punctuation ratio' (11).

DS+N	features	C_{llr}	C_{llr_min}	C_{llr_cal}
DS200	1,2,10,11	0.94	**0.85**	0.08
DS1000	2,8,10,11	0.72	**0.61**	0.10
DS2000	2,10,11,12	1.36	**0.54**	0.81
DS3000	1,2,4,8,11	1.29	**0.46**	0.83

Table 6: Performance evaluation. DS = Dataset; sample size (N) = the number of words included in each message group; features = best performing feature sets.

It is not surprising, as shown in the C_{llr_min} values of Table 63, that the performance of the system improves as a function of the sample number. DS3000 performs best with a C_{llr_min} value of 0.46.

The results of the current study outperform those of Ishihara (2011) in which datasets identical to those in the current study were assessed in terms of the LRs based on N-gram modelling (the C_{llr_min} values of DS200, DS1000, DS2000 and DS3000 are 0.96, 0.84, 0.72, and 0.62, respectively).

The LRs (uncalibrated and calibrated) of the best performing features are graphically presented as Tippett plots inFigure 1, in which the LRs, which are equal to or greater than the value indicated on the x-axis, are cumulatively plotted separately for the SA and DA comparisons. In Figure 1, a logarithmic (base 10) scale is used, in which case the neutral value is 0. Tippett plots show how strongly the derived LRs not only support the correct hypothesis but also misleadingly support the contrary-to-fact hypothesis.

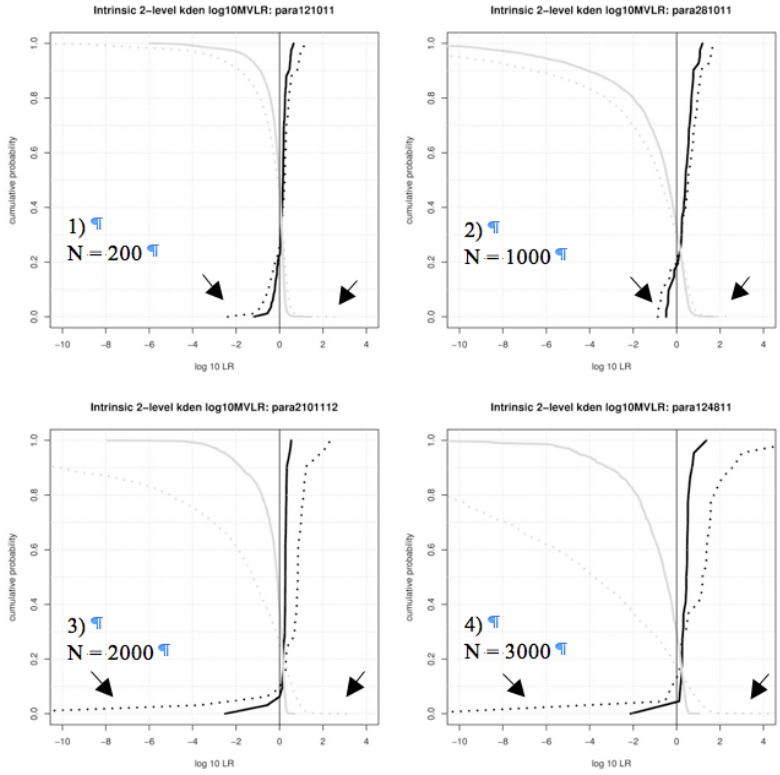

Figure 1: Tippett plots showing uncalibrated (dotted curves) and calibrated (solid curves) LRs for the sample size of 200 (panel 1); 1000 (2); 2000 (3) and 3000 (4). Black = SA comparisons; grey = DA comparison.

The C_{llr_cal} values of Table 63 indicate that DS200 (0.08) and DS1000 (0.10) are better calibrated than DS2000 (0.81) and DS3000 (0.83). This point is clear from Figure 1 (refer to the arrows) in that the uncalibrated LRs which incorrectly support the contrary-to-fact hypothesis are greater in values for the latter than for the former. The application of a calibration favourably results in a reduction in the magnitude of these misleading LRs. It also makes the magnitude of the correct LRs more conservative.

Only for reference, the equal error rates of the four best-performing systems are c.a. 34% (DS200), 24% (DS1000), 17% (DS2000) and 15% (DS3000), which are not bad. Overall, however, the LRs obtained are fairly weak. Using the verbal equivalents of LRs proposed by Champod and Evett (2000), regardless of the sample size, almost all of the calibrated LRs derived for the SA comparisons are between 1 and -1 in their strength, providing, correct or not, only limited support for either hypothesis (in other words, not very useful as evidence). Even for the best-performing result (DS3000), as many as 65% of the calibrated LRs of the DA comparisons are between -1 and 1, again providing only limited support.

6 Conclusions

We performed a likelihood-ratio-based forensic text comparison of SMS messages focusing on lexical features. The LRs were calculated in the multivariate kernel density LR formulae, and calibrated. The validity of the system was assessed based on the magnitude of the LRs using the log-likelihood-ratio-cost (C_{llr}). We demonstrated that the system with lexical features performed better than the one with N-grams. However, we pointed out that many of the derived LRs (calibrated) are weak in their strength as evidence, providing only limited support for either hypothesis.

7 Acknowledgements

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

8 References

Aitken, C.G.G. and Lucy, D. (2004), "Evaluation of trace evidence in the form of multivariate data", *Journal of the Royal Statistical Society Series C-Applied Statistics*, Vol. 53, pp109-122.

Aitken, C.G.G. and Stoney, D.A. (1991), *The Use of Statistics in Forensic Science*, Ellis Horwood, New York; London, ISBN: 0139337482

Aitken, C.G.G. and Taroni, F. (2004), *Statistics and the Evaluation of Evidence for Forensic Scientists*, Wiley, Chichester, ISBN: 0470843675.

Brümmer, N. and du Preez, J. (2006), "Application-independent evaluation of speaker detection", *Computer Speech and Language*, Vol. 20, No. 2-3, pp230-275.

Cellular-news (2006), "SMS as a tool in murder investigations", *Cellular-news*, http://www.cellular-news.com/story/18775.php, (Accessed 12 January 2012).

Champod, C. and Evett, I.W. (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", *International Journal of Speech Language and the Law*, Vol. 7, No. 2, pp238-243.

De Vel, O., Anderson, A., Corney, M. and Mohay, G. (2001), "Mining e-mail content for author identification forensics", *ACM Sigmod Record*, Vol. 30, No. 4, pp55-64.

Doheny (1996), *R v Doheny. Court of Appeal Criminal Division. No. 95/5297/Y2*.

Grant, T. (2007), "Quantifying evidence in forensic authorship analysis", *International Journal of Speech Language and the Law*, Vol. 14, No. 1, pp1-25.

Halteren, H.V. (2007), "Author verification by linguistic profiling: An exploration of the parameter space", *Proceedings of the ACM Transactions on Speech and Language (TSLP)*, Vol. 4, No. 1, pp1-17.

Iqbal, F., Binsalleeh, H., Fung, B. and Debbabi, M. (2010), "Mining writeprints from anonymous e-mails for forensic investigation", *Digital Investigation*, Vol. 7, No. 1, pp56-64.

Iqbal, F., Binsalleeh, H., Fung, B.C.M. and Debbabi, M. (In Press), "A unified data mining solution for authorship analysis in anonymous textual communications", *Information Sciences*.

Iqbal, F., Hadjidj, R., Fung, B. and Debbabi, M. (2008), "A novel approach of mining write-prints for authorship attribution in e-mail forensics", *Digital Investigation*, Vol. 5, No. Supplement, ppS42-S51.

Ishihara, S. (2011), "A forensic authorship classification in SMS messages: A likelihood ratio based approach using N-gram", *Proceedings of the Australasian Language Technology Workshop 2011*, pp47-56.

Mohan, A., Baggili, I.M. and Rogers, M.K. (2010), *Authorship attribution of SMS messages using an N-grams approach*, CERIAS Tech Report 2010-11, Center for Education and Research Information Assurance and Security Purdue University, USA.

Morrison, G.S. (2009), "Forensic voice comparison and the paradigm shift", *Science & Justice*, Vol. 49, No. 4, pp298-308.

Morrison, G.S. (2011), "Measuring the validity and reliability of forensic likelihood-ratio systems", *Science & Justice*, Vol. 51, No. 3, pp91-98.

Robertson, B. and Vignaux, G.A. (1995), *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*, Wiley, Chichester, ISBN: 0471960268.

Saks, M.J. and Koehler, J.J. (2005), "The coming paradigm shift in forensic identification science", *Science*, Vol. 309, No. 5736, pp892-895.

Stamatatos, E. (2009), "A survey of modern authorship attribution methods", *Journal of the American Society for Information Science and Technology*, Vol. 60, No. 3, pp538-556.

Tagg, C. (2009), *A Corpus Linguistics Study of SMS Text Messaging*, PhD thesis, The University of Birmingham.

Zheng, R., Li, J.X., Chen, H.C. and Huang, Z. (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*, Vol. 57, No. 3, pp378-393.