CERIAS Tech Report 2010-11

Authorship attribution of SMS messages using an N-grams approach

by Ashwin Mohan, Ibrahim M. Baggili, Marcus K. Rogers
Center for Education and Research
Information Assurance and Security
Purdue University, West Lafayette, IN 47907-2086

Authorship attribution of SMS messages using an N-grams approach

Ashwin Mohan, Ibrahim M. Baggili, and Marcus K. Rogers

***Abstract*—The pervasive use of SMS is increasing the amount of digital evidence available on cellular phones. Consequently it has become important to detect SMS authors, as a post-hoc analysis technique deemed useful in criminal persecution cases. This paper investigates an N-grams based approach for determining the authorship of SMS messages. Despite the scarcity of words in SMS messages and the differences between SMS language and natural language characteristics, the chosen method shows encouraging results in identification of authors. In this paper the effects of the gram size and the similarity scoring technique on the prediction of SMS message authors are also examined.**

***Index Terms*— SMS, Authorship, Digital forensics, N-grams**

I. Introduction

ON the Internet, people turn to some form of textual communication to maintain their relationships (i.e., social networks) in a relatively safe environment. For wireless communications on cellular phones, one such textual environment is SMS messaging. Research has shown that environments of this kind permit higher levels of visual anonymity when compared to face-to-face communication [1].

Manuscript received January 17, 2010.

Ashwin Mohan is with the Centre for Education and Research in Information Assurance, Purdue University, 646 Oval Drive, West Lafayette, IN 47907 USA (e-mail: mohana@ purdue.edu).

Ibrahim M. Baggili is with the College of Information Technology, Zayed University, Abu Dhabi, UAE (phone: +9712 5993848, fax: +9712 4434847, e-mail: Ibrahim.Baggili@zu.ac.ae).

Marcus K. Rogers is with the Department of Computer and Information Technology, Purdue University, 401 N. Grant Ave , West Lafayette, IN 47907 USA (e-mail: rogersmk@purdue.edu).

However, the visual anonymity can be misused and exploited by criminals. One way of ensuring the integrity of visual anonymity is through linguistic analysis of text being communicated, focused on identifying the author of a particular text. We look at two real life cases where authorship attribution was performed. The first is the Danielle Jones case [2]. Danielle Jones disappeared on the 18th of June in 2001 and around the same time some text messages were received from her phone. There was a suspicion among Law Enforcement officials that the messages were not actually written by her. In the Jones case, linguistic analysis concluded that the messages were more likely written by her uncle. The second is the Jenny Nicholl case. In the Nicholl case, linguistic analysis showed that the text messages sent from her cellular phone were most likely written by her ex lover [3].

Authorship attribution is a growing scientific field. Over the years, as there has been a shift in textual environments, going from paper to digital, authorship attribution studies that have been undertaken have ranged from being able to identify Shakespearian literature [4] to being able to identify authors of online forum postings [5] and e-mails [6]. However, sparse attention has been given to the authorship attribution of SMS messages on cellular phones. The importance of scientifically studying the authorship of SMS messages is growing as the Cellular Telephone Industries Association (CTIA) reports that the number of monthly SMS messages has reached 75 billion and the number of annual SMS messages has reached 600.5 billion as of Jun 2008 [7].

II. The N-gram approach

1. An overview of N-grams

The concept of N-grams was first introduced in Shannon's seminal paper on Information Theory [8]. An N-gram is a token consisting of a series of characters or words. A token is generated by moving a sliding window across a corpus of text [9] where the size of the window depends on the size of the token (N) and its displacement is done in stages, each stage corresponds to either a word or a character. Based on the different types of displacements, N-grams can be classified into two categories: 1) character based and 2) word based [10]. In this paper, character based N-grams are used. The normal value of N used when performing information retrieval on an English corpus is 2 or 3 since few words in English share the same bigrams and trigrams [13]. Research literature indicates the same for other dictionary-based languages as well [14]. To illustrate the concept of character N-grams, let us take the word “Information”. If we attempt to acquire the character bigrams and trigrams for that word, we get the following:

Bigrams: IN, NF, FO, OR, RM, MA, AT, TI, IO, ON

Trigrams: INF, NFO, FOR, ORM, RMA, MAT, ATI, TIO, ION

Shannon used character based N-grams for analyzing and predicting printed English [11]. Since then, his approach with character based N-grams has been applied to other areas like spelling and error correction, text compression, language identification and text search and retrieval [12].

The N-gram based similarity between two words is measured using one of the many similarity measures available for token-based systems like Dice's coefficient, Euclidean Distance etc [15].

2. The Need for N-grams in SMS authorship

A considerable amount of research has been done for authorship attribution. These systems work on the principle that an author rather unintentionally uses semantic and syntactic devices at every level of written text and the context provided by them is strong enough to correctly determine authorship [16].

Many algorithms have been designed to this end. The most common approach involves stylistic analysis. It has two steps: First, the style features in the text are extracted using tagging, parsing or other similar Natural Language Processing (NLP) techniques [17]. Second, classification is done on the feature set obtained to select the relevant features using goodness of fit tests (e.g., Chi-Square or Kolmogrov- Smirnov) or information of correlation within the feature set [17] features, which are ultimately chosen, give the best representation of the author's vocabulary and the frequency of the words that they use most often. This approach has its own problems. First, feature extraction based on stylistic markers is not language independent, since different languages have different styles. As a result, the parser that can find features in a text of English will be of no use when working with a language that has a completely different structure, say Japanese or Chinese [18]. Second, although feature selection is an inevitable part of the analysis, it can remove insignificant features that may provide little impact individually, however, they may still exhibit a higher cumulative effect [19].

Other than the aforementioned problems, there are even more issues when standard approaches are used with SMS messages as the text source in digital forensics. First, most standard approaches assume a large text to work with in order to extract features and perform classification. However, in cellular phone forensics, data for a message and the number of messages for a user may be very limited. Concurrently the probable author set may be large which makes classification difficult. Secondly, SMS messages do not conform to a fixed syntactical structure (as most languages do) and vary from user to user. This renders syntactical features ineffective and reduces the relevant feature set for classification. Finally, in forensic analysis of SMS there is a need for high processing speed. Investigators are bound by extreme constraints when working on cases and they cannot expend too much time or other resources to provide them with a probable solution [20].

3. Advantages of N-grams

N-grams based authorship attribution is a simple method. The author profile is the N-grams that are used frequently and their corresponding frequency in a SMS text. The two important operations are:

1. Generating the optimized set of N-grams to be included in the author profile.
2. Calculating the similarity between two author profiles.

Some of the advantages that this approach has over traditional systems when using SMS data are:

1. N-grams are independent of the language under consideration. They are able to find the roots of common words in the messages during the generation process [9]. N-grams have been used for non-language systems such as DNA [21] and Music [22].
2. There is no need for segmentation of the text in words when using character-based N-grams [14], which is unlike other methods that require lexicons for gender, plurality etc [9]. This is especially useful when working with messages from Asian Languages where borders between words are not strongly marked [9].
3. There is a high tolerance for spelling mistakes and deformations since an N-gram approach creates a character based lexicon for an author and does not use a natural language based lexicon [9]. This is useful when working with SMS since its syntactical nature is based on changing word forms (e.g., The number 2 used instead of the word to) and varies from one person to another.
4. The generation process can be computationally intensive and depends on the amount of data provided. This would be an issue when working on N-gram generation of books, articles or any other large documents. However, SMS messages are small, and the computational overhead is considerably reduced when compared to other traditional systems. Similarity calculation is a rather simple process on the token sets obtained, which makes N-gram generation extremely feasible in real life applications for SMS message authorship attribution.

III. Methodology

Our methodology consists of three major steps:

1. Selection of the working data set (corpus).
2. Implementation of the N-gram generation and similarity scoring algorithms.
3. Variation of the parameters of the experimental setup.
4. Integration of steps 2 and 3 into a robust testing system.

1. Selection of the working data set (corpus)

In order to acquire SMS messages for the study, an online, freely available SMS corpus was used. This corpus was created as part of research at the National University of Singapore [23] and is available to the public. The SMS corpus includes 10,118 SMS messages, 4635 messages of which did not have an author identification assigned to them. The messages tagged with user ID numbers could be grouped into 132 users. Many authors had a small number of messages. For the testing environment of our authorship attribution system, the corpus was parsed to retrieve the authors with at least 50 messages. This yielded a total of 28 authors. The author IDs and their respective messages together constituted the working corpus for the experiment. The final working corpus had 1400 messages (28x50).

2. Implementation of the N-gram generation and similarity scoring algorithms

The C# programming language was used for implementing the N-gram generation and similarity scoring algorithms. This was done since the authors were more experienced in working with the C# language making the programming tasks more time efficient. Furthermore, there are resources on N-gram generation and similarity scoring in C# available on the Internet.

The N-gram tokenization technique was programmatically implemented by the authors and generalized to make it possible to work with varying gram sizes. In most systems using N-grams (documents, e-mail) capitalization and word spaces are removed. With SMS, the authors make the assumption that this is not feasible since the data available is limited in a single SMS message and that these could be stylistic features that are unique to an author. Besides these two features, the tokenization also uses punctuations like colon, semi-colon or ellipsis and whole or floating-point numbers as stylistic markers.

The University of Sheffield's NLP group in the United Kingdom had an open source package [15], which was used for the similarity matching. Their source code for the scoring techniques was modified to fit our tokenization algorithm. Seven token-based similarity-scoring algorithms were included to be implemented and tested in our system: Cosine similarity, Jaccard similarity, Dice's coefficient, Block distance, Euclidean distance, Overlay coefficient, Matching coefficient. To discuss more about similarity scoring is beyond the scope of this paper. Nonetheless, to get a more comprehensive understanding of these scoring techniques – one can refer to [15].

3. Variation of the parameters of the experimental setup

The sample set size (SS), the gram size (GS) and the similarity scoring techniques (SST) together form the experimental parameter set E_p.

$$E_p = \{SS, GS, SST\}$$

When each of the Parameters from E_p is assigned a value, we get a Value set V_p.

$$V_p = \{SS = a, GS = b, SST = c\}$$

Where a, b and c are the values. The V_p is the input for the experimental setup. The researchers programmatically generated value sets and passed them to the setup. A total of 1750 such sets were created. The parameters were varied as:

*SS***:** The sample set is the total number of messages in the system taken from the working set per user. The size of the SS was varied in the range R ϵ (2...10). This range was chosen to ensure that there was no bias in the experimental setup because of the order of the messages in the working set.

*GS***:** Research in literature shows that the best performance for systems is usually at an N-gram size of 2 or 3[13]. The researchers decided to work on gram sizes between 1 and 5 and observe whether the same holds true for SMS despite being a pseudo (partial attributes of a language based on dictionary or a grammar structure) language system.

SST: This has been discussed in section 3.2.

4. The Testing System

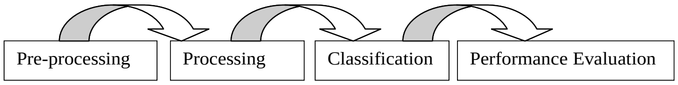

Fig. 1. The subunits of the Testing System

*Pre-processing***:** The testing process was initiated by loading the working corpus. The SS parameter was used to decide on the size of the sample sets (different sets for different users), which were then extracted from the corpus. While the samples were extracted, they were randomly selected using the Fisher Yates Algorithm [24]. Each user contributed the same number of messages to the training set (let this number be a) and the testing set (let this number be b).

Given that the experimental corpus had 28 users and the SS for each user had a size of a + b, then the training would be 28a messages and the testing set would be 28b messages. For e.g. When SS = 10, the numbers for the training and testing set would be 252 and 28 respectively.

*Processing***:** The first message from the testing set was selected and the N-gram tokenization technique was applied using the GS parameter (which indicates the gram size). The next step was to apply the N-gram tokenization technique for all the messages of a selected user in the training set using the same GS parameter. The third step was to generate a similarity score between the test message and the selected user (This score was normalized across all the messages for that user). The third step was repeated for all the users in the training set. This resulted in a predicted author list, which was ranked by the normalized similarity score. All these steps were then repeated for all the test messages.

Classification: The ranked user list for each test message was then used for classification. This is a classification problem with more than 2 classes (28 to be exact). In multi-class classification one of the classes is the positive class and the others together make the negative class. This distinction is done for two cases: for the actual result set and the predicted result set by the system. In our experiment, for the actual result set the positive class is the one to which the test message belongs to and the rest serve as the negative class. For the predicted result set the positive class is the one that occupies the first rank in the ranked user set for the message and the rest constitute the negative class.

The result for each message classification can be set as correct or incorrect by looking at its positive class in the actual result set with respect to the positive and negative classes in the predicted result [25]. The classification is correct if the predicted positive class is the same as the actual positive class or True Positive (TP) and incorrect if the actual positive class belongs to the predicted negative class or False Negative (FN). There are some results that can be obtained by using the actual negative class. These are the False Positive (FP) and the True Negative (TN). The two-way confusion matrix as shown in Table 1 is a good way to understand these four different results of classification.

TABLE I

TWO WAY CONFUSION MATRIX

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Performance Evaluation: A set of statistics was calculated for each author. These include the error rate (E_{AC}) [25] and the F-score (F_{AC}) [25] where AC is the author class as defined in [25]. To find the overall success of the classification experiment, the macro averaged error rate and F-score statistics (in percentages) were calculated across all authors and multiple runs.

$$F_{ACMavg} = (\Sigma^{Y}{}_{k=1} (\Sigma^{X}{}_{i=1} F_{ACi})/ X)/ Y$$

$$E_{AC\ Mavg} = (\Sigma^{Y}{}_{k=1} (\Sigma^{X}{}_{i=1} E_{ACi})/ X)/ Y$$

Here X was the total number of authors and Y the total number of runs.

IV. RESULTS

Table 2 and 3 represent the top results (FACMavg) in the experiment with the corresponding value of the parameters of the experimental setup. We now evaluate the effect of these parameters within the results provided.

SS or number of messages per author:

TABLE II

TOP TEN RESULTS FOR MESSAGE SET 2-5

Rank	No. of Messages per Author	Ratio between Training &Testing messages	Similarity Type	Gram Size	F_{AC}^{Mavg}	E_{AC}^{Mavg}
1	3	2:1	Block Distance	3	70.94	1.26
2	3	2:1	Dice Similarity	3	70.94	1.26
3	3	2:1	Euclid Distance	3	70.7	1.27
4	3	2:1	Cosine Similarity	3	69.68	1.24
5	3	2:1	Jaccard similarity	3	69.31	1.32
6	2	1:1	Overlap Coefficient	3	69.01	1.35
7	5	4:1	Matching Coefficient	5	68.16	1.48
8	2	1:1	Cosine Similarity	4	67.31	4.98
9	5	4:1	Overlap Coefficient	3	67.03	1.66
10	5	4:1	Matching Coefficient	4	66.07	1.46

TABLE III

TOP TEN RESULTS FOR MESSAGE SET 6-10

Rank	No. of Messages per Author	Ratio between Training and Testing messages	Similarity Type	Gram Size	F_{AC}^{Mavg}	E_{AC}^{Mavg}
1	7	6:1	Euclid Distance	5	72.81	1.11
2	7	6:1	Dice similarity	5	72.56	1.13
3	7	6:1	Block Distance	5	72.56	1.13
4	8	7:1	Euclid Distance	4	72.43	1.25
5	8	7:1	Dice Similarity	4	72.27	1.25
6	8	7:1	Block Distance	4	72.27	1.23
7	8	7:1	Jaccard Similarity	4	72.19	1.28
8	8	7:1	Cosine Similarity	4	71.87	1.25
9	7	6:1	Jaccard Similarity	5	71.77	1.14
10	7	6:1	Cosine Similarity	5	71.06	1.23

The variation in F-scores is less with more messages in the SS. In the top ten ranks, the change for group 6-10 is 2.4 % and for the group 2-5 is 6.8 %.

GS or Number of grams: From Tables 2 and 3, we see that the best results for smaller message sets are for gram size 3. However as the number of messages per user increases, there is a trend towards the gram size 5.

*SST or Similarity scoring technique***:** From Tables 2 and 3, the best results are seen for Euclidean and Block Distances among the ranked scores. However, the similarity scoring techniques have a lesser effect towards change in accuracy on the group with 6-10 messages.

V. CONCLUSIONS

The authors were able to use their testing system to determine that an N-gram approach for an SMS corpus can predict the author with an accuracy of 65-72%. The best gram size to use for tokenization is 3 for smaller message sets and 5 for larger message sets. The Euclidean similarity-scoring technique should be used with larger message sets, while for smaller message set scoring should be done by block distance.

Typically in a digital forensic case, the number of SMS messages obtained from cell phones may be limited. It is possible that different people may have written the various SMS messages than the owner of the cellular phone. Without any preset message tagging, it is difficult to identify the correct author of the message. Our results show promising application in the prediction of an author of an SMS message with an accuracy of 65%-70%, when the sample of SMS messages are small and the number of authors is comparably large.

This method can be used in two scenarios: a) For cases in which suspects have been identified or b) For cases in which suspects have not been identified. In the first, the suspects are limited to the forensic case under investigation. In the second, the investigator may have a database that contains a number of SMS messages with their respective N-grams, in which the examiner may try to find a correlation between the messages and an author.

VI. FUTURE WORK

The authors contemplate the importance of testing this system using other English SMS corpora in order to validate their results. A formal approach would involve using mobile devices to obtain SMS data and standardizing the acquired datasets. Future research would involve a fine-tuned analysis of the datasets, the effects of message lengths, and variations to changing individual parameters in the hope of improving the prediction for authorship attribution.

REFERENCES

[1] K. McKenna, A. Green and M. Gleason, "Relationship formation on the Internet: What's the big attraction?" *Journal of Social Issues*, vol. 58, no. 1, 2002, pp. 9-31.

[2] Tim Grant, "How text-messaging slips can help catch murderers".[Online]. Available: http://www.independent.co.uk/opinion/commentators/dr-tim-grant-how-textmessaging-slips-can-help-catch-murderers-923503.html. [Accessed: Nov. 29, 2008].

[3] Cellular-news, "SMS as a tool in murder investigations". [Online]. Available: http://www.cellular-news.com/story/18775.php. [Accessed: Nov. 24, 2008].

[4] T. Merriam, "Heterogeneous authorship in early Shakespeare and the problem of Henry V," *Literary and Linguistic Computing*, vol. 13, no. 1, 1998, pp. 15-28.

[5] A. Abbasi and H. Chen, "Applying authorship analysis to extremist-group web forum messages," *IEEE Intelligent Systems*, vol. 20, no. 5, 2005, pp. 67-75.

[6] B. Allison and L. Guthrie, " Authorship attribution of e-mail: comparing classifiers over a new corpus of evaluation," in *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, May 28-30, 2008, Marrakech, Morocco.

[7] CTIA, "Wireless Quick Facts".[Online]. Available: www.ctia.org/advocacy/research/index.cfm/AID/10323. [Accessed: Nov. 24, 2008].

[8] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, 1948, pp. 379–423, 623–656.

[9] Z.Elberrichi and B.Aljohar, "N-grams in Texts Categorization," *Scientific Journal of King Faisal University (Basic and Applied Sciences)*, vol. 8, no. 2, 1428H, 2007.

[10] V. Kešelj, F. Peng, N. Cercone and C. Thomas, "N-gram-based author profiles for authorship attribution," *in Pacific Association for Computational LINGuistics(PACLING'03)*, August 22-25, 2003, Nova Scotia, Canada, pp. 255-264.

[11]C. E. Shannon, "Prediction and entropy of printed English," *Bell System Technical Journal*, vol. 30, pp. 50-64.

[12] P Majumder, M. Mitra and B.B. Chaudhuri, "N-gram: a language independent approach to IR and NLP," *International Conference for Universal Knowledge*, Nov 25-29, 2002, Goa, India.

[13] A. El-Nasan, S. Veeramachaneni and G. Nagy, "Word discrimination based on bigram co-occurrences," *Proceedings of the 6th International Conference on Document Analysis and Recognition*, Sep 10-13, 2001, Seattle, Washington, pp. 149-153.

[14] N. Jian-yun, G. Jiangfeng, J. Zhang et al," On the Use of Words and N-gram for Chinese Information Retrieval,“ *Proceedings of the fifth international workshop on Information retrieval with Asian languages*, New York, USA, 2000, pp.141-148.

[15] SimMetrics: Open Source Similarity Measurement Library. [Online]. Available: http://www.dcs.shef.ac.uk/~sam/simmetrics.html. [Accessed: Oct. 12, 2008].

[16] M. Ephratt," Authorship attribution - the case of lexical innovations," Proceedings of *ACH-ALLC-97*, Jun 2-7, 1997, Kingston, Ontario.

[17] E. Stamatatos, N. Fakotakis and G. Kokkinakis, "Computer-based Authorship attribution without lexical measures," *Computers and the Humanities*, vol. 35, no. 2, pp. 193–214.

[18] S. Scott, and S. Matwin," Feature engineering for text classification," *In Proceedings of The Sixteenth International Conference on Machine Learning,* June 27-30, 1999, Bled, Slovenia.

[19] A. Aizawa," Linguistic techniques to improve the performance of automatic text Categorization," *In Proceedings of 6th Natural Language Processing Pacific Rim Symposium,* Nov 27-30, 2001, Tokyo, Japan.

[20] M.Rogers, R.Mislan, T.Wedge and J.Goldman, "Computer Forensics Process Triage Model," *ADFSL Conference on Digital Forensics, Security and Law,* April 20-21, 2006, Las Vegas, Nevada.

[21] J.Y.Kim and J.S. Taylor,, "Fast String Matching using an N-gram Algorithm," *Software Practice and Experience*, vol. 24, no.1, 1994, pp. 79-88.

[22] Y.C.Lap and B.Kao , “A study on N-gram indexing of musical features,” *IEEE International Conference on Multimedia*, May 10-14, 2000, Hangzhou, China, vol.2, pp. 869-872**.**

[23] NUS SMS Corpus. [Online]. Available: http://www.comp.nus.edu.sg/~rpnlpir/downloads/corpora/smsCorpus/. [Accessed: Nov. 24, 2008].

[24] R.A.Fisher and F.Yates, *Statistical tables for biological, agricultural and medical research, 3rd ed*., London,Oliver & Boyd,1948, pp. 26–27.

[25] M.W. Corney, A.M. Anderson, G.M. Mohay, and O. de Vel, “Identifying the Authors of Suspect Email,” 2008.