Reading Between the Turns: Statistical Modeling for Identity Recognition and Verification in Chats

Giorgio Roffo [†] Cristina Segalin [†] Alessandro Vinciarelli [‡] Vittorio Murino [⊕ †]
Marco Cristani [† ⊕]

[†] Department of Computer Science, University of Verona (IT)
[⊕] Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia, Genova (IT)
[‡] School of Computing Science, University of Glasgow, (UK)

Abstract

Identity safekeeping has recently become an important problem for the social web: as a case study, we focus here on instant messaging platforms, proposing novel soft-biometric cues for user recognition and verification. Specifically, we design a set of features encoding effectively how a person converses: since chats are crossbreeds of written text and face-to-face verbal communication, the features inherit equally from textual authorship attribution and conversational analysis of speech. Importantly, our cues ignore completely the semantics of the chat, relying solely on non-verbal aspects, taking care of possible privacy and ethical issues. We apply our approach on a novel dataset of 94 different individuals, whose chat conversations have been recorded for an average period of five months; recognition rate, intended as normalized AUC on CMC curve, is 95.73%, while verification rate amounts to 95.66%, as normalized AUC on ROC curve.

1. Introduction

Protection from identity violation on social networks is a critical problem our society faces across both geographical and cultural boundaries. Essentially, there are two ways an identity can be violated; the first is by *identity theft* [13], where an impostor manages to access the personal account of someone else, mostly due to Trojan horse keystroke logging programs (as Dorkbots [6]), or by *social engineering* (*i.e.*, tricking individuals into disclosing login details or changing user passwords) [3]. The second consists in *faking an identity*, that is, acting as an invented person, or emulating another person [10].

Since communication through social networks, such as Facebook, Twitter, and Skype is rapidly growing [17], identity violation is becoming a primary threat to people's cultural attitudes and behaviours in social networking. The numbers are impressive: the Federal Trade Commission reported that 9.9 million (22% more than 2007) Americans suffered from identity theft in 2008 [7].

As a matter of fact, identity violation by theft is easier than one can think given the cyber-habits of average users: a survey conducted in 2008 in the U.S.A, Canada, and some European countries illustrates that 25% of Germans and 60% of Americans have shared their account passwords with a friend or family member; Furthermore, 50% of the Americans use as passwords important dates, nicknames, or pet and family member names.

The urgency of attacking the identity violation problem led several institutions (banks, enforcement agencies and judicial authorities) to produce algorithms capable of limiting or discovering as soon as possible this threat: for example, the Identity Theft Red Flags Rule, issued in 2007, requires creditors and financial institutions to implement identity theft prevention strategies. Such strategies should be triggered by patterns, practices, or specific activities − known as red flags − that could indicate identity theft [7].
In this paper, we follow this line by investigating technologies aimed at revealing the genuine identity of a person involved in instant messaging activities. In practice, we simply require that the user under analysis (from now on, the *probe* individual) engages in a conversation for a limited number of turns. This allows us to extract cues and provide statistical measures of how well the user matches the samples in a *gallery* of signatures. Subsequently, the match mesures can be employed for performing user recognition or verification.

It is worth noting the novelty of our contribution: chat texts are an intriguing type of data, representing crossbreeds of literary text and spoken conversations. Whereas Authorship Attribution (AA) of standard written text has a long history [1], recognizing the participants of a chat conversation is pretty novel, especially when it comes to taking into

account the similarity with spoken conversations.

The proposed features inherit equally from the two realms, actually exploiting the turn-taking dynamics while considering the turn as the most basic chunk of information to be analysed, rather than the entire conversation as in standard AA. In addition, our features are *non-verbal* in nature, meaning that the semantics of the words is completely discarded: we simply neglect the content of the conversations by substituting the letters with a generic symbol. In this way, the features are privacy preserving and make the usage of our system more plausible in commercial applications.

The experiments have been performed over a corpus of chat conversations involving 94 individuals in total. The chat conversations of each subject have being recorded for an average period of five months; in particular, for each individual we get around 13 hours of chatting activity, on average. The recognition rate, intended as normalized AUC on CMC curve, is 95.7%, while verification rate amounts to 95.7%, as normalized AUC on ROC curve.

The rest of the paper is organized as follows: Sec. 2 surveys previous approaches for authorship attribution on different kinds of texts; Sec. 3 presents the features proposed in this work and Sec. 4 shows how they have to be used for matching. Experiments on recognition and verification are discussed in Sec. 5, and, finally, Sec. 6 concludes the paper.

2. Related work

Authorship Attribution (AA) is the domain aimed at automatically recognizing the author of a given text sample, based on the analysis of *stylometric* cues. AA attempts date back to the 15th century [15]. Since then, many stylometric cues were designed, usually partitioned into five major groups: *lexical, syntactic, structural, content-specific and idiosyncratic* [1]. Table 1 is a synopsis of the features applied so far in the literature.

Typically, state-of-the-art approaches extract stylometric features from data and use discriminative classifiers to identify the author (each author corresponds to a class). The application of AA to chat conversations is recent (see [16] for a survey), with [1, 2, 11, 18] the most cited works. In [18], a framework for authorship attribution of online messages is developed to address the identity-tracing problem. Stylometric features are fed into SVM and neural networks on 20 subjects, validating the recognition accuracy on 30 random messages. PCA-like projection is applied in [2] for Authorship identification and similarity detection on 100 potential authors of e-mails, instant messages, feedback comments and program code. A unified data mining approach is presented in [11] to address the challenges of authorship attribution in anonymous online textual communication (email, blog, IM) for the purpose of cybercrime investigation.

In the last ten years, authorship attribution and forensic analysis have extended their research to IM communication[8]. In [14], four authors of IM conversations have been identified based on sentence structure as well as use of special characters, emoticons, and abbreviations.

The main limitation of the works above is that they do not process chat exchanges as conversations, but as normal texts. In practice, the feature extraction process is always applied to the entire conversation and individual turns, while being the basic blocks of the conversation, are never used as analysis unit. This work overcomes such limitation and introduces, among the others, a new class of features that account for the presence of turns (see below) in chat conversations. Furthermore, the proposed approach does not apply the feature extraction process to the entire conversation (like in all works above), but to individual turns.

3. A new set of features for AA on chats

In our work, we examine chats among pairs of people, *i.e.*, dyadic interactions. These conversations can be considered as sequences of *turns*, where each "*turn*" is a set of symbols typed consecutively by one subject without being interrupted by the other person. In addition, each turn is composed by one or more *sentences*: a sentence is a stream of symbols which is ended by a "return" character. Each sentence is labeled by a temporal ID, reporting the time of delivery.

For each person involved in a conversation, we examine the stream of turns (suppose T), completely ignoring the input of the other subject. This corresponds to assuming that the chat style (as modeled by our features) is independent of the interlocutor: this has been validated experimentally. From these data, a personal signature is extracted, that is composed by different cues: some of them could be associated to particular classes of the taxonomy of Table 1, while others need a new categorization, being tightly connected with the "conversational" nature of a chat. Therefore, we define a new class of "turn-taking" features.

In all the cases, it is very important to note that in standard AA approaches, the features are counted over entire conversations, obtaining a single quantity. In our case, we consider the turn as a basic analysis unit, obtaining T numbers for each feature. For ethical and privacy isues, we decided to discard all cues involving the content of the conversation. Even if this choice is very constraining, because it prunes out many features of Table 1, the results obtained are encouraging.

In the following, we list the proposed features: in cursive bold, we indicate the novel features[1], together with a brief explanation.

[1] In some sense, all the features are novel, since they are collected on turns instead of the whole text; still here we want to highlight "structurally" novel features.

Group	Description	Examples	References
Lexical	Word level	Total number of words (=M), # short words/M, # chars in words/C, # different words, chars per word, freq. of stop words	[2, 11, 14, 16, 18]
	Character level	Total number of characters (chars) (=C), # uppercase chars/C, # lowercase chars/C, # digit chars/C, freq. of letters, freq. of special chars	[2, 14, 16, 18]
	Character—Digit n-grams	Count of letter—digit n-gram (a, at, ath, 1 , 12 , 123)	[2, 16, 18]
	Word-length distribution	Histograms, average word length	[2, 11, 14, 16, 18]
	Vocabulary richness	Hapax legomena, dislegomena	[2, 11, 16, 18]
Syntactic	Function words	Frequency of function words (of, for, to)	[2, 11, 14, 16, 18]
	Punctuation	Occurrence of punctuation marks (!, ?, :), multiple !—?	[2, 11, 14, 16, 18]
	Emoticons—Acronym	:-), L8R, Msg, :(, LOL	[14, 16]
Structural	Message level	Has greetings, farewell, signature	[2, 11, 14, 16, 18]
Content-specific	Word n-grams	Bags of word, agreement (ok, yeah, wow), discourse markers—onomatopee (ohh), # stop words, # abbreviations, gender—age-based words, slang words	[2, 11, 14, 16, 18]
Idiosyncratic	Misspelled word	Belveier instead of believer	[2, 11, 14, 16]

Table 1. Synopsis of the state-of-the-art features for AA on chats. "#" stands for "number of".

Lexical Features

- Number of words, chars, mean word length, number of uppercase letters;
- **Number of Uppercase / Number of Chars**; usually, entire words written in capital letters indicate a strong emotional message. This feature records such communicative tendency.
- **n-order Length Transitions (noLT)**; These features resemble the n-grams of [9]; the strong difference here is in the fact that we consider solely the length of the words, and not their content. In practice, for a noLT of order $n = 1$ (1oLT), we build probability transition matrices that in the entry i, j, $1 \leq i, j \leq I$, exhibit the probability of moving from a word of length i to a word of length j. In our case, we set $I = 15$. noLT of order $n = 2$ (2oLT) are modeled by transition matrices of I^3. We did not take into account superior order, for sparsity issues.

Syntactic Features

- Number of ? and ! marks, three points (...), generic marks (",.:*;), rate of emoticons / words, rate of emoticons / chars;

Turn-taking Features

- **Turn duration**; the time spent to complete a turn (in seconds); this feature accounts for the rhythm of the conversation with faster exchanges typically corresponding to higher engagement. As shown in [5], turn duration modeling allows one to finely characterize dyads, highlighting for example the degree of the engagement, as soon as the age of the participants.
- **Writing speed**; number of typed characters -or words- per second (typing rate); these two features indicate whether the duration of a turn is simply due to the amount of information typed (higher typing rates) or to cognitive load (low typing rate), i.e. to the need of thinking about what to write
- **Emoticons Category *Positive*, *Negative*, *Other***; these features aim at individuating a particular mood expressed in a turn through emoticons. In particular, we partition 101 diverse emoticons in three classes, portraying positive emotions (happiness, love, intimacy, etc. $-$ 20 emot.), negative emotions (fear, anger, etc. $-$ 19 emot.), and neutral emoticons (portraying actions, objects etc. $-$ 62 emot.), counting their total number of occurrences. We are conscious that our partition is somewhat subjective: still, our attempt was to discover whether emotion-oriented classes of emoticons were more expressive than a unique class, reporting all the possible emoticons. Experimentally, our choice lead to higher recognition performance.
- **Mimicry**; ratio between number of chars -or words- in current turn and number of chars -or words- in previous turn; this feature models the tendency of a subject to follow the conversation style of the interlocutor (at least for what concerns the length of the turns). The mimicry accounts for the social attitude of the subjects.
- **Answer Time**; this feature is the time spent to answer a question in the previous turn of another interlocutor.

These quantities are extracted from each turn, as written above, with the exception of the mean word length, the noLT feature, the Emoticons Category: actually, in such cases, the turn does not offer sufficient statistics for a robust description. Therefore, for these cues, we consider all the turns of a subject as they were a unique corpus. Conversely, for all the other cues, we have T numbers; these numbers are then described employing histograms. On our data, we noted that most of the features extracted are strongly collapsed toward small numeric values: for this reason, we

adopt exponential histograms, where small-sized bin ranges are located toward zero, increasing their sizes while going to higher numbers. Experimentally, we get much better results than exploiting uniformly binned histograms over the whole range of the features.

4. Matching personal descriptions

Let us suppose to have collected the features for two subjects, A and B. We now have to exploit them for obtaining a single distance, describing the overall similarity between A and B. As first step, we derive a plausible distance for each feature separately: in the case we have histograms, we employ the Bhattacharyya distance. For the features represented by mean values, we adopt the Euclidean distance. In the case of the $noLT$ features, we consider the *diffusion distance* [12], which acts similarly to the Pyramid Matching Kernel [9]. In practice, the diffusion distance measures the linear distance among the matrices' entries, applying iteratively (L times) Gaussian kernels of increasing variance: this allows to include cross-entries relations in the final measure, thus alleviating sparsity problems as well as quantization effects. Briefly speaking, given M_A and M_B the $noLT$ matrices, the diffusion distance $K(M_A, M_B)$ is

$$K(M_A, M_B) = sum_{l=0}^{L} |d_l(\mathbf{x})| \quad (1)$$

where

$$d_0(\mathbf{x}) = M_A(\mathbf{x}) - M_B(\mathbf{x}) \quad (2)$$

$$d_l(\mathbf{x}) = [d_{l-1}(\mathbf{x}) * \phi(\mathbf{x}, \sigma)] \downarrow_2 \quad l = 1, \ldots, L \quad (3)$$

with $\mathbf{x}$ the elements of a matrix, with dimension $I \times I$; $\phi(\mathbf{x}, \sigma)$ is the 2D Gaussian filter of standard deviation σ; L indicates the number of levels employed, and $\downarrow_2$ denotes half size downsampling. The parameter $\sigma = 4$ and the level $L = 4$ have been set by crossvalidation.

Since the aim of this paper is explorative on the nature on the features, and not how to fuse them, we do not investigate how such features should be combined together. Therefore, in this paper, we adopt a simple average rule, *i.e.*, the final distance is obtained by averaging over the contribute of the single distances, opportunely normalized between 0 and 1.

5. Experiments

In the experiments, we evaluate the effectiveness of each feature in performing identity recognition; subsequently, we analyze how the compound of all the features does the same task; finally we consider the identity verification task.

5.1. The dataset

In this work, we examine a corpus of 94 dyadic italian chat conversations collected with Skype, performed by $N = 94$ different users. The conversations are spontaneous, i.e. they have been held by the subjects in their real life, collected over a time span of 5 months: in particular, for each individual we get around 13 hours of chatting activity. The number of turns per subject ranges between 200 and 1000. Hence, the experiments are performed over 110 turns for each person. The turns of each subject are split into *probe* and *gallery* set, each including 55 samples. In this way, any bias due to differences in the amount of available material should be avoided. When possible, we pick different turns selections (maintaining their chronological order) in order to generate different probe/gallery partitions.

ID	Name	Range	nAUC	Rank
1	#Words(W)	[0,1706]	75.6%	5
2	#Chars(C)	[0,15920]	77.3%	2
3	Mean Word Length	[0,11968]	74.2%	7
4	#Uppercase letters	[0,11968]	70.7%	14
5	***#Uppercase / C***	[0,1]	71.7%	12
6	***1o_LT***	[0,127]	76.1%	4
7	***2o_LT***	[0,127]	70.0%	15
8	# ? and ! marks	[0,21]	58.8%	21
9	#Three points (...)	[0,54]	71.4%	13
10	#Marks (",.:*;)	[0,1377]	83.1%	1
11	***#Emoticons / W***	[0,4]	77.0%	3
12	***#Emoticons / C***	[0,1]	75.0%	6
13	***Turn Duration***	[0,1800]	72.5%	11
14	***Word Writing Speed***	[0,562]	72.9%	9
15	***Char Writing Speed***	[0,5214]	72.9%	10
16	***#Emo. Pos.***	[0,48]	73.0%	8
17	***#Emo. Neg.***	[0,5]	62.8%	17
18	***#Emo. Oth.***	[0,20]	61.2%	19
19	***Imitation Rate / C***	[0,2611]	65.2%	16
20	***Imitation Rate / W***	[0,1128]	62.9%	18
21	***Answer Time***	[0,2393]	59.8%	20

Table 2. Stylometric features used in the experiments and recognition statistics.

5.2. Preliminary feature analysis

For the sake of clarity, the features are numbered in Table 2, reporting also their minimum and maximum values. The first part of the experiments aims at assessing each feature independently, as a simple ID signature. A particular feature of a single subject is extracted from the probe set, and matched against the corresponding gallery features of all subjects, employing a given metric (see Sec. 4). This happens for all the probe subjects, resulting in a $N \times N$ distance matrix. Ranking in ascending order the N distances for each probe element allows one to compute the *Cumulative Match Characteristic* (CMC) curve, i.e., the expectation of finding the correct match in the top n positions of the ranking.

The CMC is an effective performance measure for AA approaches [4], and in our case is a valid measure for evaluating the task of *identity recognition*: given a test sample, we want to discover its identity among a set of N subjects. In particular, the value of the CMC curve at position 1 is the probability that the probe ID signature of a subject is closer to the gallery ID signature of the same subject than to any other gallery ID signature; the value of the CMC curve at position n is the probability of finding the correct match in the first n ranked positions.

Given the CMC curve for each feature (obtained by averaging on 10 trials), the normalized Area Under Curve (nAUC) is calculated as a measure of accuracy. For the sake of clarity, the features are partitioned in two sets: those resembling the classical AA features, now calculated on turns (Fig. 1) and the novel ones (Fig. 2). As visible, all our features give performances above chance: in Table 2, last two columns, are reported the nAUC score and the rank built over the nAUC score. In order to understand the informa-

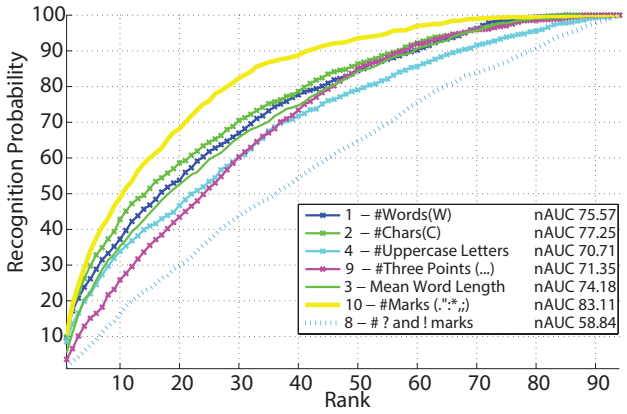

Figure 1. CMC curve for each "classical" feature.

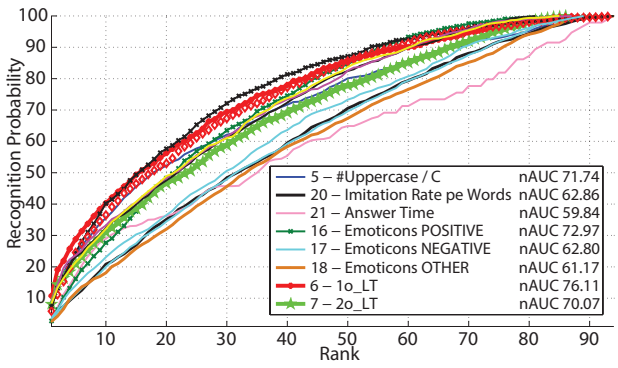

Figure 2. CMC curve for each novel feature.

tion contained in all the features, and how they are interrelated, we calculate the Spearman's rank correlation coefficient (see Fig. 3), highlighting in the upper triangular part statistically significant correlations with p-value$<1\%$, in the bottom triangular part those correlations significant at p-value$<5\%$. In general, a high level of correlation is ex-

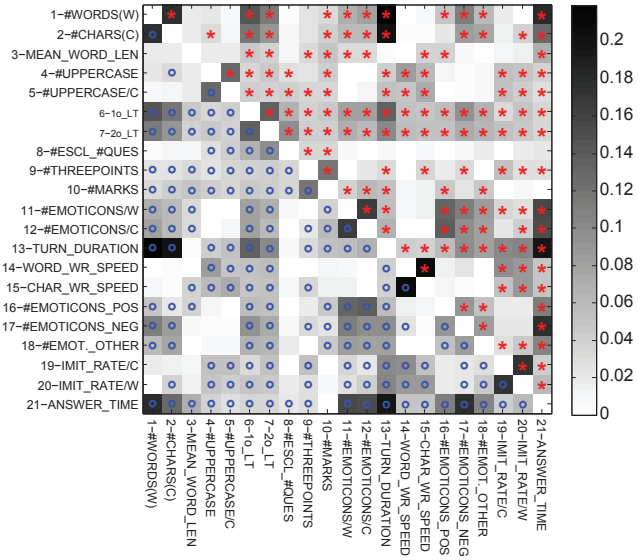

Figure 3. Correlation analysis between features: red asteriscs report correlations significant with p-value$<1\%$, blue dots correlations with p-value$<5\%$ (best viewed in colors).

istent between features. Quite interesting, $noLT$ features seem to be correlated with all the other cues.

5.3. Identity Recognition and Verification by all the features

In this section, we put together all the proposed features as described in Sec. 4. In Fig. 4 is reported the CMC curve obtained by joining all the distances, which gives the identity recognition performance. This performance is signif-

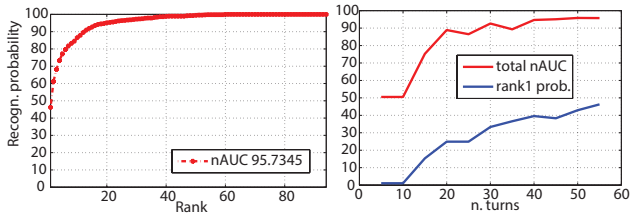

Figure 4. (left) Global CMC Curve; (right) nAUC and rank1 probability varying the number of turns.

icant higher than for any cue cosidered individually, realizing an nAUC of 0.9573. This witnesses that, even if the features are strongly correlated, they model complementary information. In fact, adopting standard feature selection strategies like, e.g., Forward Feature Selection, shows that all the features increase the recognition rate. It is worth noting that, the probability of guessing the correct user at rank 1 is slighty below 50% which is quite encouraging (actually, in standard people re-identification tasks, where the features

are the images of people, performances with a similar number of subject into play is quite inferior). To investigate how

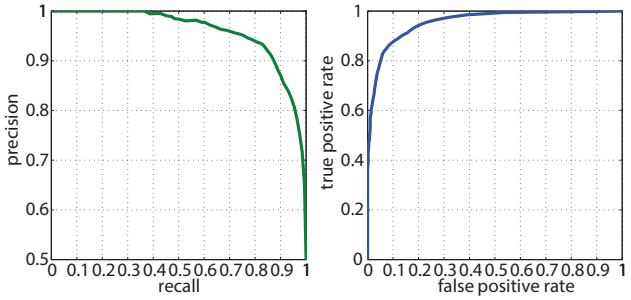

Figure 5. Precision, Recall and ROC Curve.

important is the number of turns taken into account for modeling the gallery and probe subjects, we show in Fig. 4 the nUAC of the CMC curve and the rank1 probability by varying the number of turns. Intuitively, the higher the number of turns, the higher the recognition rate.

Considering the verification task, we adopt the following strategy: given the signature of user i, if it matches with the right gallery signature with a matching distance which is ranked below the rank K, it is verified. Intuitively, there is a tradeoff in choosing K. A high K (for example, 50) gives a 100% of true positive rate (this is obvious by looking at the global CMC - Fig. 4), but it brings in a lot of potential false positives. Therefore, taking into account the number K as varying threshold, we can build ROC and precision/recall curves, portrayed in Fig.5. Considering the nAUC of both the curves, we get 0.9566 and 0.9351, respectively. The best compromise between precision and recall is obtained calculating the F1 value, which gives 0.88 for precision 0.90 and recall 0.87, corresponding to the value of $K = 45$.

6. Conclusions

In this paper, we proposed a new facet for biometrics, considering the chat content as personal blueprint. From tens of turns, we extracted heterogeneous features, which take from the Authorship Attribution and the Conversational Analysis background. On a test set of 94 people, we show that identification and verification can be performed definitely above chance; even if our performances are far from the level required for commercial systems, many improvements can be done. First of all, fusion strategies for collapsing intelligently the cues should be investigated. Secondly, learning policies should be taken into account, which are expected to boost the accuracy in a consistent fashion.

References

[1] A. Abbasi and H. Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions On Information Systems*, 26(2):1–29, 2008.

[2] A. Abbasi, H. Chen, and J. Nunamaker. Stylometric identification in electronic markets: Scalability and robustness. *Journal of Management Information Systems*, 25(1):49–78, 2008.

[3] R. J. Anderson. *Security Engineering: A Guide to Building Dependable Distributed Systems*. John Wiley & Sons, Inc., 2001.

[4] R. Bolle, J. Connell, S. Pankanti, N. Ratha, and A. Senior. *Guide to Biometrics*. Springer Verlag, 2003.

[5] M. Cristani, A. Pesarin, C. Drioli, A. Tavano, A. Perina, and V. Murino. Generative modeling and classification of dialogs by a low-level turn-taking feature. *Pattern Recognition*, 44(8):1785 – 1800, 2011.

[6] Z. Deng, D. Xu, X. Zhang, and X. Jiang. Introlib: Efficient and transparent library call introspection for malware forensics. In *Proceedings of Digital Forensic Research Workshop*, pages 13 – 23, 2012.

[7] K. M. Finklea. *Identity Theft: Trends and Issues*. DIANE Publishing, 2010.

[8] J. Gajadhar and J. Green. An analysis of nonverbal communication in an online chat group. Technical report, The Open Polytechnic of New Zealand, 2003.

[9] K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *Journal of Machine Learning Research*, 8:725–760, 2007.

[10] J. P. Harman, C. E. Hansen, M. E. Cochran, and C. R. Lindsey. Liar, liar: Internet faking but not frequency of use affects social skills, self-esteem, social anxiety, and aggression. *CyberPsychology & Behavior*, 8(1):1–6, 2005.

[11] F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi. A unified data mining solution for authorship analysis in anonymous textual communications. *Information Sciences*, 231:98–112, 2011.

[12] H. Ling and K. Okada. Diffusion distance for histogram comparison. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 246–253, 2006.

[13] R. C. Newman. Cybercrime, identity theft, and fraud: practicing safe internet - network security threats and vulnerabilities. In *InfoSecCD*, pages 68–78, 2006.

[14] A. Orebaugh and J. Allnutt. Classification of instant messaging communications for forensics analysis. *The International Journal of Forensic Computer Science*, 4(1):22–28, 2009.

[15] G. Shalhoub, R. Simon, R. Iyer, J. Tailor, and S. Westcott. Stylometry system–use cases and feasibility study. *Forensic Linguistics*, 1:8, 2010.

[16] E. Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.

[17] D. L. Williams, V. L. Crittenden, T. Keo, and P. McCarty. The use of social media: an exploratory study of usage among digital natives. *Journal of Public Affairs*, 12(2):127–136, 2012.

[18] R. Zheng, J. Li, H. Chen, and Z. Huang. 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, 2006.