Emerald Information Management and Computer Security

Language independent gender identification through keystroke analysis

Journal:	*Information Management and Computer Security*
Manuscript ID:	Draft
Manuscript Type:	Original Article
Keywords:	Biometrics, Forensics

Language independent gender identification through keystroke analysis

Author list

Abstract

Purpose – In this work we investigate the feasibility of identifying the gender of an author by measuring the keystroke duration when typing a message.
Design/methodology/approach – Three classifiers were constructed and tested. We empirically evaluated the effectiveness of the classifiers by using empirical data. We used primary data as well as a publicly available dataset containing keystrokes from a different language to validate the language independence assumption.
Findings – The results of this work indicate that it is possible to identify the gender of an author by analyzing keystroke durations with a probability of success in the region of 70%.
Research limitations/implications – The proposed approach was validated with a limited number of participants and languages, yet the statistical tests show the significance of the results. However, this approach will be further tested with other languages.
Practical implications – Having the ability to identify the gender of an author of a certain piece of text has value in digital forensics, as the proposed method will be a source of circumstantial evidence for "putting fingers on keyboard" and for arbitrating cases where the true origin of a message needs to be identified.
Social implications – If the proposed method is included as part of a text composing system (such as email, and instant messaging applications) it could increase trust toward the applications that use it and may also work as a deterrent for crimes involving forgery.
Originality/value – The proposed approach combines and adapts techniques from the domains of biometric authentication and data classification.

Keywords: keystroke dynamics, keystroke duration, gender recognition, user classification.

1. Introduction

When investigating a computer crime, one of the core concepts and main challenges in digital forensics is to "put fingers on keyboard", that is to identify the human who is responsible for the generation or handling of digital data related to the respective criminal or offensive activity in general (Shavers, 2013). The increased complexity and large scope of the underlying problem space, makes circumstantial digital evidence critical when attempting to identify a user.

Textual communication between users dominates most other forms such as audio, video or image. For every second, there are about 2.4 million emails sent (internetlivestats.com). A typical incident relating to textual communication is a masquerade attack, where a malicious user assumes the identity of another legitimate user. To date, email spoofing detection is mainly focused to the examination of the email metadata found on the headers and its correlation with network information, such as IP addresses, geo-location, and so forth. Focus is not given to verifying the individual. However, there does exist a wealth of research literature on biometric based authentication of users based on the way a user types (Clarke *et. al*, 2003, Clarke & Furnell, 2007). However, current email applications do not record the biometric characteristics of the author

and therefore biometric based identification techniques are not applicable in such a context. Hence, in order to adopt biometric authentication approaches in identification of user characteristics, the authoring applications need to maintain biometric data recording capabilities. Adding such abilities to authoring applications would offer a number of benefits and protection to the users. Consider for example a system that can identify the gender of the author. Masquerading is a popular behavior in Internet applications such as social networks, associated to various offenses (phishing, pedophilia, etc.). A system that would have the ability to warn or inform the unsuspected recipient of a message about the gender of the sender of that message, would act as a deterrent towards offenses leveraging gender misrepresentation, it would increase the protection of the legitimate users and of course support investigations in case of a reported incident.

In this paper we propose the use of keystroke analysis in order to determine the gender of a user. We investigate the validation of such hypothesis in order to develop a business case for creating applications that include functionality for allowing gender identification of the creator of the text, while this text is being typed. We envisage that the proposed method will provide a forensic analyst with enough circumstantial evidence to support their investigation and perform e-discovery through informed decision making.

The rest of the paper is organised as follows. In the Section 2 the related work on gender recognition within written text is examined. The experimental design and methodology are presented in Section 3, with the accompanying results documented in Section 4. The paper concludes with Section 5 presenting the key findings and future work.

2. Related Work

The question of indentifying the gender of a writer intrigued researchers since texts were only in handwritten form. Trudgill (1972) highlights the differences in linguistic styles between male and female authors of a handwritten text. Similarly, there are plenty of works studying the differences that appear between males and females in speech and informal writing, like the work of Holmes (1988), for instance.

Gender identification is an interesting topic with a number of useful applications, such as better translation between languages as some languages employ different grammatical structures depending on the gender, the targeted advertising for Internet users, and of course forensic analysis.

The research community has done a significant amount of work on gender and user identification in general, by proposing a number of techniques with varying prediction success rates. In his work, Lai (2006) constructed datasets from books and blogs, and proposed a Naïve-Bayes classifier on a selection of words achieving a success rate between 64% and 91%, depending on the particular set of features.

De Vel et al. (2002) acknowledge the proliferation of e-mail as a modern means of communication, combining old and new characteristics of written speech; the new characteristics refer to the abbreviations – frequently found in text chatting – as well as intentionally misspelled words done by the author of a message in order to convey emotional information. They collected e-mails from an academic organization comprised of approximately 15,000 members and trained a system using a large number of features, such as, the number of characters, lines, paragraphs, alphabetic characters, upper-case characters, spaces, words ending with “able”, words ending with “al”, “sorry” words, and so forth. They achieved a success rate, depending on the size of e-mail, of between 56% and 71%.

In a similar manner, more recently, Cheng et al. (2011) used a dataset from e-mails and journalist’s articles and utilized three different classifiers to achieve a success rate between 55% and 85%. The features they selected included; the number of lines, paragraphs, sentences, some special characters, like “%”, “&”, the number of punctuations, articles, pronouns, auxiliary-verbs, and so forth, as well as the presence of some particular words. Argamon et al. (2003) used documents from the British National Corpus and checked the frequency of appearance of pronouns, common nouns, proper nouns, as well as the frequency of appearance of some particular words, both in fiction and nonfiction texts. Using a Bayesian technique they achieved a success rate of 80%. Based on this research a tool named “Gender Genie” was developed (Gender Genie, nd.). The tool allocates a number of points either on a “male” or on a “female” variable, every time a particular word appears. The highest variable of the two indicate the gender of the author. Another tool is the “Gender Guesser” (n.d.), which is based on “Gender Genie” but is more advanced in the sense that it can process informal writing often found in blogs and casual messages. However, in order to perform with an improved prediction rate the tool requires a minimum of 300 words. Similarly, another tool is “uClassify”

(n.d.) which has been trained on 11,000 blogs, half of which were written by male and the other half female authors.

A common characteristic of all of the research published to date is that they are language dependent. More specifically, they require that the language is specified – English in particular. Consequently, other languages such as Spanish, Chinese, Greek, German and so forth would require additional training with respective datasets. This requirement results in a significant limitation (Doyle and Keselj, 2005). Despite the acknowledgement of the language dependency limitation, the solution proposed by Doyle and Keselj is language dependent, but it is relatively simple to adapt it for other languages. In their work 500 student essays were used and the appearances of n-grams (a character, a digram, a trigram, etc.) for both male and female authors were obtained offering a success rate between 51% and 81%. Giot and Rosenberger (2012) claim to have introduced the first language independent gender recognition system. This was achieved using keystroke duration, digram latency and a vector which is the concatenation of the four previous timing values. They achieved a success rate of 90% and use it to improve user authentication. The main limitation of the work by Giot and Rosenberger (2012) is the focus on the protection of passwords. That is, the users were profiled based on their typing of two specific words (Giot *et al.,* 2009).

Our proposed work aims at providing continuous identification throughout a typing session. In addition, provided that our approach involves the collection of a larger volume of user data, we could reduce the number of required attributes, improving the efficiency of the gender identification system.

3. Keystroke-Based Gender Experiment

3.1. Experimental Methodology

In this work we attempt to perform gender recognition of the author of a text using a limited number of attributes and more particularly the parameters of keystroke dynamics. This suggests that the proposed recognition system is language independent, because the results are based on the way the user types rather than what they type. Furthermore, in order to have a system with increased user acceptance and reduced non-compliance risks due to possible privacy regulations, the keystroke features utilized are limited to the keystroke duration. Nevertheless, maintaining a dataset of keystroke duration information will allow the construction of other related attributes such as digram – or n-gram in general – based latencies, thus supporting further investigation and identification of the attribute of the highest accuracy.

Unfortunately, to date there seems to be a limited number of public datasets of keystroke dynamic, therefore it was necessary to design a data collection methodology and implement a data collection tool. A key logging application was developed in Visual Basic. The application upon recording the user with a suitable unique identifier required the user to type a fixed text of 850 characters containing letters, digits, and other symbols. Upon completion of the typing exercise, a comma-separated file was created named after the subject's username, with each line containing the character pressed, the keydown and keyup time in milliseconds. Whilst it is possible to capture keystroke characteristics with a greater resolution, prior research within this domain when applied to authentication have typically focused upon milliseconds (Joyce and Gupta, 1990).Each typing session is captured in a single file. That is, if a user for example participates in two different sessions (say by running the experiment twice – one on a laptop and one on a desktop), two files will be created. The times recorded are measured as the time elapsed from the execution of the keylogger application.

An initial team of 17 volunteers used the keylogger application both on desktops and laptops in order to cover a wide variety of typing cases. The recording period was undertaken from 11.10.2012 until 21.11.2012. Whilst the number of participants is not large, careful thought was given to their selection, to ensure an appropriate representation. This involved controlling for gender and left-right handedness.. The volunteers' characteristics with respect to the general population representation, are shown in **Table 1**. The number of participants who were male was almost equal to the number of female volunteers. The proportion of left-handed volunteers was about 11%, closely reflecting the proportion of the whole population. The educational level of the participants corresponds to the ratio of the level of education of a population with a Greek nationality.

user id	Gender	Handedness	Educational Level
a1	Male	Left-Handed	High School
a2	Male	Right-Handed	High School
a3	Male	Right-Handed	University
a4	Female	Right-Handed	University
a5	Male	Right-Handed	T.E.I.*
a6	Female	Right-Handed	High School
a7	Male	Right-Handed	T.E.I.
a8	Male	Right-Handed	High School
a9	Female	Right-Handed	High School
a10	Female	Left-Handed	High School
a11	Male	Right-Handed	High School
a12	Female	Right-Handed	T.E.I.
a13	Female	Right-Handed	T.E.I.
a14	Female	Right-Handed	High School
a15	Male	Right-Handed	University
a16	Female	Right-Handed	University
a17	Male	Right-Handed	University

* Technological Educational Institution.

Table 1. User profiles

3.2 Descriptive Statistics

After the recording period, 34 files, 18 corresponding to male authors and 16 to female authors, were available for processing. From these files the keystroke durations of every key that was pressed were extracted and all the records from "male" files were merged to produce a single dataset; the same procedure was followed for the "female" files – creating the two classes of interest.

The dataset was sanitized by removing all outliers that corresponded to values exceeding three times the mean value, a standard methodological approach utilized in keystroke analysis studies (Clarke & Furnell, 2007). Subsequently, the statistical features for each character were calculated to find if there are any differences in the way that type the members of the two classes. Although the sample was not too big, some important differences appeared (as illustrated in **Table 2).**

Character	Key Code	Appearances per User (mean)	"Female" file		"Male" file		Percentage Difference (%)	
			Mean (ms)	Standard Deviation (ms)	Mean (ms)	Standard Deviation (ms)		
(Space)	32	102	99.59	32.79	93.66	27.75	-5.96	-15.38
Alpha	65	45	92.00	27.73	99.14	30.80	7.76	11.08
Iota	73	50	89.11	27.42	88.95	22.21	-0.19	-19.01
Pi	80	24	83.34	31.24	92.45	27.21	10.94	-12.90
Tau	84	41	80.53	26.87	92.12	25.92	14.40	-3.52

Table 2. Differences in statistical features

It should be noted that significant differences are not limited to the characters of Table 2, but are encountered almost in half of the tested character set. Some characters display significant differences in their mean value, others in the standard deviation and some in both. As a general observation of these preliminary findings is that the males hold their fingers a little bit longer on the keys than the females, while the females are not as consistent as males are, due to the larger standard deviations experienced.

The normality and subsequent t-test results between the means of the two classes are shown in Table 3. As it can be seen from the normality P-values, both classes are normal, permitting us to run a paired t-test in order to establish whether the means of the two classes are different. The resulting probability was equal to 0.0003 indicating that the means of the two classes are significantly different and therefore the key latency attribute can be used to distinguish the gender of the author.

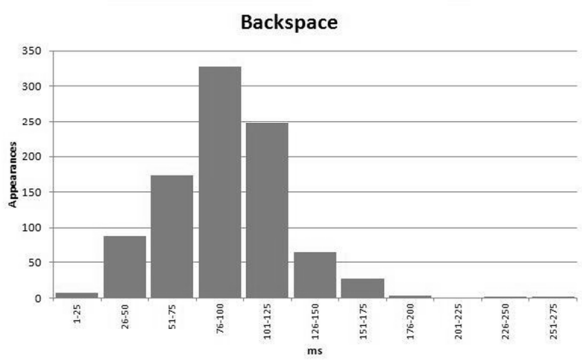
	Mean (ms)	Standard Deviation (ms)	Normality test		paired t-test [P]
			χ^2	P	
Male	88.1075	5.1144	1.8356	0.3993	[0.0003]
Female	91.8302	4.9292	2.6530	0.2654	

Table 3. Normality and t-test results

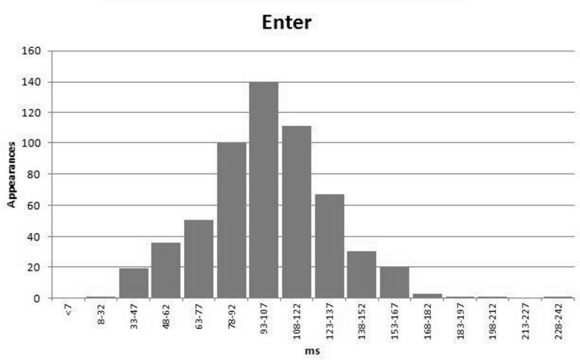
3.3 Classifier Design and Evaluation

In this work we constructed and evaluated three classifiers, namely a Naïve Bayes, and two classifiers based on the Manhattan and Euclidean distance respectively. This is in agreement with the current relevant literature, where a considerable volume of research displayed preference to these classifiers and corresponding methods in a variety of different problem domains – see for example the work by Phyu (2009), Khamar (2013), Kotsiantis (2007). In addition, the rationale for selecting these specific three candidate classifiers was the significant differences in the statistical descriptors as well as the distribution types of the keystroke duration histograms, making thus these classifiers appropriate for the current work More specifically, the Naïve Bayes classifier was a suitable classifier since our attribute follows a normal distribution (Bouckaert, 2004). The significant differences of the standard deviations between the male and female keystrokes warrants the use and evaluation of the Manhattan distance (Li *et al.* 2005). Finally, the statistically significant differences between the two means are explored and utilized through the Euclidean distance.

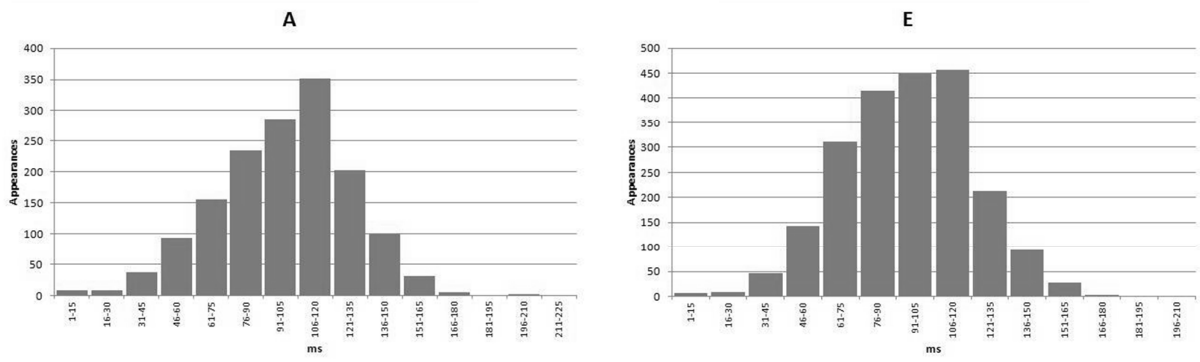
Firstly we examine the performance of a Naïve Bayes classifier. A histogram of keystroke duration appearances of a representative set of keys is shown in **Figure 1**.



(a)



(b)



(c) (d)

Figure 1. The distribution of keystroke durations

Since the keystroke duration attribute follows a normal distribution, the following probability can be attached to every observation:

$$p(x \mid g) = \frac{1}{\sqrt{2\pi\sigma_{x,g}^2}} e^{-\frac{(x-\mu_{x,g})^2}{2\sigma_{x,g}^2}} \quad (1)$$

where

$p(x|g)$ is the probability of "x" character being typed by user of gender "g",

$\mu_{x,g}$ and $\sigma_{x,g}$ is the mean value and the standard deviation of the data respectively, for the "x" character and for the "g" gender.

Upon calculating the probability for each character, they are binned into the two classes as follows:

$Male\_Accumulator = p(\text{'A'}|male) + p(\text{'B'}|male) + \ldots$
$Female\_Accumulator = p(\text{'A'}|female) + p(\text{'B'}|female) + \ldots$

The final probabilities are then obtained from:

$$p(male) = \frac{Male\_Accumulator}{Male\_Accumulator + Female\_Accumulator}$$

$$p(female) = \frac{Female\_Accumulator}{Male\_Accumulator + Female\_Accumulator}$$

It can be seen that $p(male)=1-p(female)$. The higher of the two is the final gender guess. The success rate of the proposed system was 61.76%.

As mentioned earlier, the observation that most of the characters exhibit significant differences in their standard deviations between "male" and "female" classes led us to evaluating the performance of the Manhattan distance. The Manhattan distance between the calculated standard deviation from that of "male" and "female" classes indicates the gender of the author of the text. In this experiment the correct predictions reached a success rate of 64.71%, but the results showed a bias towards the "male" class.

The third approach was the exploitation of the differences between the means. The mean values of keystroke durations of specific characters were calculated and these values were compared with the respective values of "male" and "female" classes. The shorter distance indicates the author's gender. Once again the test was performed over all available texts and for some characters the success rate was 67.65%. However, one

important drawback of this approach is the dependence upon a single character, reducing thus the reliability of its performance.

To remedy this limitation consideration was given to the average values for a selection of characters separately, rather than an aggregate value for all characters. After this the Euclidean Distances from the values of model are calculated and the shorter of these indicates the author's gender. More analytically, the following comparison was performed:

$$\sqrt{\sum_x \left(\frac{tot\_dur\_x}{count\_x} - \mu_{x,f} \right)^2} < \sqrt{\sum_x \left(\frac{tot\_dur\_x}{count\_x} - \mu_{x,m} \right)^2} \quad (2)$$

where

tot_dur_x is the sum of char "x" keystroke durations,
count_x is number of times char "x" was typed
$\mu_{x,f}$ and $\mu_{x,m}$ are the mean values of char "x"'s keystroke duration from the data that were created by females and males volunteers, respectively.

The success rate achieved was 64.71% and the system is not dependent on a single character, addressing the aforementioned limitation.

After a thorough study of the above metrics and the corresponding results, a scoring system based on the Manhattan distance was introduced. For each different character the following comparison is performed:

$$\left| \frac{tot\_dur\_x}{count\_x} - \mu_{x,f} \right| < \left| \frac{tot\_dur\_x}{count\_x} - \mu_{x,m} \right| \quad (3)$$

where tot_dur_x, count_x, $\mu_{x,f}$ and $\mu_{x,m}$ have the same meaning as the terms in expression (2).

The outcome of (3) dictates how the weighted scoring system is to be applied. If (3) holds, the points will be assigned to the probability denoting that the text belongs to a male author. In the opposite case, the points are assigned to the probability of the female author. The number of assigned points depends on the appearance frequency of the character "x" in the text and on the difference between the values $\mu_{x,f}$ and $\mu_{x,m}$ of the model.

4. Validation and results

The results of the latter scoring system with the volunteer dataset are shown in **Table 4**. The scoring system considered expression (3) above and included weights for every character, depending on the appearance frequency of the character in the text and its difference between the two classes (male and female). That is, a frequently appearing character with a significant statistical difference would be given the maximum weight (5 points in our system), as opposed to a character appearing a limited number of times and having low differences between the classes (1 point). Group 1 below involves the case of the volunteers that were used to train the system.

Group 1 (control group)								
Females			**Males**			**Total**		
Successes	Failures	Success Rate	Successes	Failures	Success Rate	Successes	Failures	Success Rate
11	5	68,75%	13	5	72,22%	24	10	70,59%

Table 4. Manhattan Distance based classifier

The validation of the constructed classifier was performed with two additional datasets. First, we invited a second group of volunteers. The usernames, gender of each member of the second group, as well as the device type that they were recorded on are shown in **Table 5**.

User id	Gender	Recorded on
b1	Female	Desktop
b2	Female	Desktop & Laptop
b3	Male	Desktop & Laptop
b4	Male	Desktop & Laptop
b5	Female	Desktop & Laptop
b6	Male	Desktop & Laptop
b7	Female	Desktop & Laptop
b8	Male	Desktop & Laptop
b9	Male	Desktop
b10	Female	Desktop & Laptop
b11	Female	Desktop & Laptop
b12	Male	Desktop & Laptop
b13	Male	Desktop & Laptop
b14	Male	Desktop & Laptop

Table 5. Second group of volunteers

This group was asked to type the same control text as the first group, producing a total of 26 files, as two of the volunteers did not provide laptop typing data. The results of the classifier are shown in **Table 6**.

Group 2 (same fixed text)								
Females			Males			Total		
Successes	Failures	Success Rate	Successes	Failures	Success Rate	Successes	Failures	Success Rate
7	4	63.64%	11	4	73.33%	18	8	69.23%

Table 6. Second group results

The performance of the classifier is almost the same as those of the first group. That is, the percentage of correct prediction is in the region of 70% with better results attributed to the "male" class.

A third dataset was used to validate the language independence claim. This is a dataset produced by Bello et al. (2010). The dataset contains the keystroke logging of some particular sentences in Spanish and some particular UNIX commands from a team of 40 male and 15 female users, producing thus 55 files in total.

The results of applying the proposed method are shown in **Table 7**.

Group 3 (Spanish users)									
Females			Males			Total		
Successes	Failures	Success Rate	Successes	Failures	Success Rate	Successes	Failures	Success Rate
10	5	66.67%	26	14	65.00%	36	19	65.45%

Table 7. Results from the team of third party dataset.

This result is close to the previous measurements, albeit the small number of characters included in the weighted calculation. As such, we conjecture that more characters will improve the accuracy and reliability of the proposed method. **Table 8** shows the results by aggregating all files from all datasets.

All of Files and Users								
Females			Males			Total		
Successes	Failures	Success Rate	Successes	Failures	Success Rate	Successes	Failures	Success Rate
28	14	66.67%	50	23	68.49%	78	37	67.83%

Table 8. Overall results

5. Conclusions and Future Work

Leveraging keystroke dynamics research to construct user profiles in the context of a digital investigation is a promising area of research and a domain with practical importance to electronic discovery. In this paper we focused on the feasibility of identifying whether a user typed a certain text is male or female, but a complete solution would need to consist of independent tests corresponding to a variety of characteristics or properties. Besides user authentication, keystroke dynamics may be useful to detect the emotional state of the user, or to identify his handedness, or to assess whether the user is typing in their native language or not.

Due to the preliminary yet promising results, the model will be extended to consider other user characteristics or properties in order to form a concise and concrete solution. The complete approach which is part of our ongoing research involves the identification of correlation of the user properties through latent variables in order to establish the mutual information between them and the construction of a formal evidence handling framework based on known evidence fusion constructs such as the Dempster-Shafer theory of evidence.

A limitation of the current research was the use of a fixed text to create the reference model, which departs from the realistic behavior of users. An improvement would be to use an agent that logs the user in a real working environment and we conjecture that this would increase the success rates. Another parameter increasing the prediction accuracy is a larger user sample. Finally, an improvement that will significantly raise the reliability of the system would be to create feedback mechanism, enhancing in this way the database that generated the equations for the possibilities export, and the weights of each character.

References

Argamon, S., Koppel, M., Fine, J. and Shimoni, A. R. (2003), "Gender, Genre and Writing Style in Formal Written Texts", *Text - Interdisciplinary Journal for the Study of Discourse*, Vol. 23, Issue 3, pp. 321-346.

Bello, L., Bertacchini, M., Benitez, C., Pizzoni, C.J., Cipriano, M. (2010), "Collection and Publication of a Fixed Text Keystroke Dynamics Dataset", in *XVI Congreso Argentino de Ciencias de la Computacio 2010*, pp. 822-831.

Bouckaert, R. (2005), "Naive Bayes Classifiers That Perform Well with Continuous Variables", in Webb G.I. and Yu X. (Eds), *AI 2004: Advances in Artificial Intelligence - Lecture Notes in Computer Science, Volume 3339*, Springer Berlin Heidelberg, pp. 1089-1094.

Chao-Yue L. (2006), Author Gender Analysis. I256 Applied Natural Language Processing. Available at: http://courses.ischool.berkeley.edu/i256/f09/Final%20Projects%20write-ups/LaiChaoyue_project_final.pdf

Cheng, N., Chandramouli R. and Subbalakshmi K.P. (2011), "Author Gender Identification from Text", *Digital Investigation*, Vol. 8, Issue 1, pp 78-88.

Clarke, N. and Furnell, S. (2007), "Authenticating Mobile Phone Users Using Keystroke Analysis", *International Journal of Information Security*, Vol. 6, Issue 1, pp. 1-14.

Clarke, N., Furnell, S., Lines, B. and Reynolds, P. (2003), "Keystroke Dynamics on a Mobile Handset: a Feasibility Study", *Information Management & Computer Security*, Vol. 11, Issue 4, pp.161-166.

Doyle, J. and Keselj V. (2005), "Automatic Categorization of Author Gender via N-Gram Analysis", in *6th Symposium on Natural Language Processing*, Chiang Rai – Thailand.

Gender Genie, available at http://www.hackerfactor.com/GenderGuesser.php (accessed 23 May 2014).

Giot, R. and Rosenberger, C. (2012) "A New Soft Biometric Approach for Keystroke Dynamics Based on Gender Recognition", *International Journal of Information Technology and Management*, Vol. 11, Issue 1/2, pp. 35-49.

Giot, R., El-Abed, M. and Rosenberger, C. (2009), "GREYC Keystroke: a Benchmark for Keystroke Dynamics Biometrics Systems", in *3rd IEEE International Conference on Biometrics: Theory, Applications and Systems, BTAS '09*, IEEE, pp. 1-6.

Holmes, J. (1988), "Paying Compliments: A Sex-Preferential Positive Politeness Strategy", *Journal of Pragmatics*, Vol. 12, Issue 3, pp. 445-465.

Kagstrom, A., Karlsson, A. and Kagstrom, E., "uClassify – GenderAnalyzer_v5", available at http://www.uclassify.com/browse/uClassify/GenderAnalyzer_v5 (accessed 20 August 2013).

Khamar, K. (2013), "Short Text Classification Using kNN Based on Distance Function", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol.2, Issue 4, pp. 1916-1919.

Kotsiantis, S. (2007), "Supervised Machine Learning: A Review of Classification Techniques", *Informatica*, Vol.31, pp. 249-268.

Krawetz, N. (2002), "Hacker Factor: Gender Guesser", available at http://www.hackerfactor.com/GenderGuesser.php (accessed 18 August 2013).

Li, W., Wang, K., Stolfo, S. and Herzog, B.(2005), "Fileprints: Identifying File Types by N-Gram Analysis", in *Sixth Annual IEEE SMC, Information Assurance Workshop, IAW '05*, IEEE, pp. 64-71.

Phyu, T.N. (2009), "Survey of Classification Techniques in Data Mining", in *International MultiConference of Engineers and Computer Scientists 2009 Vol. I, IMECS 2009*, Hong Kong, pp. 727-731.

Shavers, B. (2013), "Placing the Suspect Behind the Keyboard: Using Digital Forensics and Investigative Techniques to Identify Cybercrime Suspects", Elsevier/Syngress.

Trudgill, P. (1972), "Sex, Covert Prestige and Linguistic Change in the Urban British English of Norwich", *Language in Society*, Vol. 1, Issue 2, pp.179-195.

Vel, O.D., Corney, M., Anderson, A. and Mohay G. (2002), "Language and Gender Author Cohort Analysis of E-mail for Computer Forensics", in Second Digital Forensic Research Workshop, Syracuse, NY.