JUNE 2006
ISSN 1675-7017

SOCIAL AND MANAGEMENT RESEARCH JOURNAL

INSTITUTE OF RESEARCH, DEVELOPMENT AND COMMERCIALISATION

SOCIAL AND MANAGEMENT RESEARCH JOURNAL

Social and Management Research Journal is jointly published by Institute of Research, Development and Commercialisation (IRDC) and University Publication Centre (UPENA), Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia.

SOCIAL AND MANAGEMENT RESEARCH JOURNAL

Vol. 3 No. 1 **June 2006** **ISSN 1675-7017**

Using Text Mining Algorithm to Detect Gender Deception Based on Malaysian Chat Room Lingo

Dianne L.M. Cheong
Nur Atiqah Sia Abdullah @ Sia Sze Yieng
Faculty of Information Technology and Quantitative Sciences,
Universiti Teknologi MARA (UiTM), Malaysia
Email: dianne@tmsk.uitm.edu.my, atiqah@tmsk.uitm.edu.my

ABSTRACT

E-mail can be a fantasy playground for identity experimentations where players take on an imaginary persona and interact with each other in the virtual world. Therefore, gender deception is difficult, risky and it can be abandoned at will. Inference can be made both from writing style and from clues hidden in the posting data. A text-mining algorithm was designed to detect gender deception based on gender-preferential features at the word or clause level of Malaysian e-mail users. Based on this algorithm, a prototype in Visual Basic is developed. It was tested with 16 documents; each consists of 5 e-mails exchanges of respective individuals. The tests shown the prototype have 81.3% of accuracy level. This is consistent with a human reader of the documents. This prototype can be a tool to assist interested parties such as the Criminology and Forensic Department, e-mail users and virtual communities to successfully identify gender deception.

Keywords: *gender detection, gender of e-mail author, text-mining algorithm, program to detect gender, gender deception*

Introduction

E-mail is used for communication between strangers and friends. It can be a fantasy playground for identity experimentations where players take on an

ISSN 1675-7017

imaginary persona and interact with each other in the virtual world. Knowing the identity of those with whom you communicate is essential for understanding and evaluating an interaction. The presentation of self in the virtual world is often a conscious and deliberate endeavour. Gender deception is difficult and risky in the real world but a very common activity and can be abandoned at will in cyberspace.

According to a research conducted by Mind Share Media Guide 2000, the Internet users in Malaysia are dominated by males (64%) whilst females represent 36% of the Internet users. Of the overall Internet users, 24% are in the 15 – 19 age groups, 40% are in the 20 – 29 age groups, and 19% are in the 30 – 39 age groups. Approximately 16% of the Internet users are above 40 years of age. Those aged between 15 to 29 years old form the largest consumer group and they enjoy the most modern gadgets that make life convenient and interesting. At the forefront of these gadgets is the computer and with it the Internet and connectivity. This has created a generation of 'wired' youths.

The focus of this research is to develop a program to detect gender deception through chat room vocabulary used mainly by Malaysians. The basic premise is that the users are who they claim to be as identity cues are sparse. However, inference can be made both from writing style and from clues hidden in the posting data.

Statement of the Problem

The online social environment as accorded by the e-mail and IM provide the opportunity to 'pretend' to be someone else. A user might, for example, be a male but tell other users that he is female in order to get more attention (Bruckman, 1993). Friendship is secured online while deception and fraud are revealed offline.

Online communities are growing rapidly and their participants face this dilemma: Many of the basic cues about personality and social role we are accustomed to in the physical world are absent. Identity cues are sparse in cyber space but not non-existent. People become attuned to the nuances of e-mail addresses and signature styles. New phrases evolve that help mark their users as members of a chosen subculture. Virtual reputations are established and impugned.

Therefore, the presentation of self in the virtual world is often a conscious and deliberate endeavour. In the real world, gender deception is difficult and risky but it is a common activity in the virtual world. Ironically, gender deception can be abandoned at will. By looking closely at some identity cues, at how they work and when they fail, we can learn a great deal about online deception through the Malaysian chat room lingo.

Objectives of the Study

The focus of the research is to develop a program to detect gender deception through chat room vocabulary used mainly by Malaysians. The basic premise is that the users are who they claim to be as identity cues are sparse. However, inference can be made both from the writing style and from cues hidden in the posting data of the e-mail. The objectives of this research are mainly to:

- Design a text mining algorithm to detect gender deception based on gender-preferential features at the word or clause level of Malaysian e-mail users.
- Design a program in Visual Basic based on the designed text algorithm.
- Evaluate the Visual Basic program through test runs on the compiled e-mail exchanges.

Literature Review

A study (Cheong & Foo, 2006) was conducted to examine whether males and females can effectively convey a false gender identity in computer-mediated communication.

(CMC), and what aspects of their language changed from typical gender-preferential language in attempting to do so. A Malaysian chat room lingo was compiled for this study. The study showed that when an individual is attempting to create a false gender identity they will vary obvious, consciously controlled aspects of communications such as topic, length of text, number of questions, rather than gender-preferential linguistic features at the word or clause level. False gender identities were more extreme than real gender identities. Many aspects of one's own gender-preferential language are retained while attempting to create a false gender identity.

Some of the language features that may be used to provide readers with some clues to predict gender are:

i. The use of intensive adverbs (McMillan, Clifton, McGrath, & Gale, 1977; Mulac & Lundell, 1986; Mulac, Wiemann, Widenmann, & Gibson, 1988).
ii. The number of references to emotion (Mulac, Studley, &Blau, 1990).
iii. The number and use of modals and tag questions (McMillan et al., 1977).
iv. The frequency of compliments (Holmes, 1988).
v. The use of minimal responses (Carli, 1990).
vi. The use of personal pronouns, oppositions, subordinating conjunctions (Mulac & Lundell, 1986; Thomson & Murachver, 2000).
vii. The frequency of questions (Tannen, 1994).

Cheong and Foo (2006) discovered that on an average of five e-mail exchanges, Malaysians of age 23 show the following:

Table 1: Frequency of Features in 5 e-mail Exchanges by Malaysians

Feature	Gender	
	Male	Female
Exclamations	130	145
Questions	8	10
Reference to emotions	6	12
Request for information	3	3
Give personal information	3	4
Give opinion	3	4
Self-derogatory comments	0.33	0.16
Compliments	0.83	0.98
Apologies	1.25	0.91

Although there is substantial evidence of gender differences in language, these differences are gender-preferential rather than gender-exclusive (Fitzpatrick, Mulac, & Dindia, 1995; Thomson & Murachver, 2000). Some features might be more characteristic of one gender than the other and obtain small gender differences (Thomson & Murachver, 2000).

Theoretical Framework of the Study

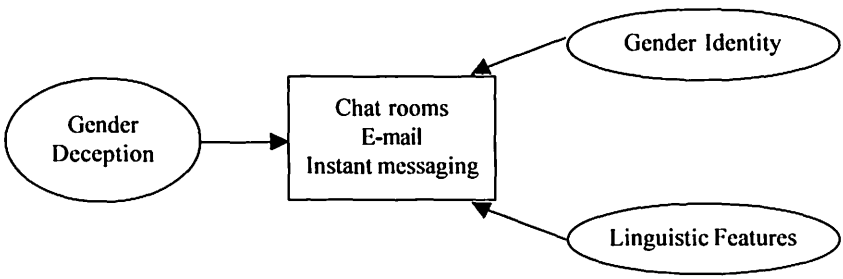

Figure 1: The Three Attributes of Computer-mediated Communication as Depicted by Past Research

Figure 1 shows a theoretical framework of the study. Computer-mediated communication such as chat rooms, e-mail, and instant messaging has three attributes. There are namely gender deception, gender identity and linguistic features.

Conceptual Framework of the Study

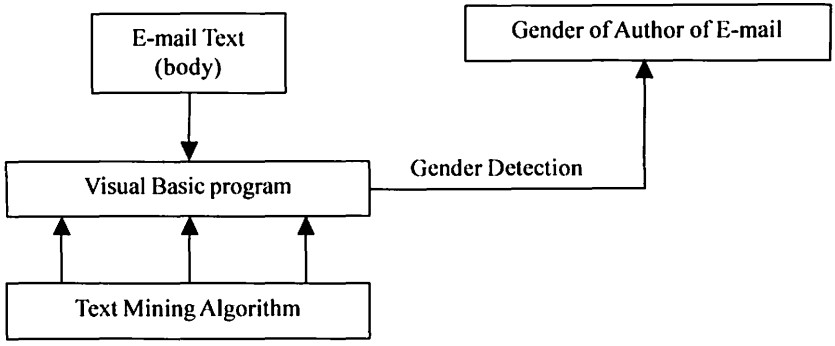

Figure 2: Proposed Conceptual Framework for the Study

Figure 2 shows a proposed conceptual framework for the study. An e-mail text is read by the program, which designed and developed from a text-mining algorithm. The text-mining algorithm is designed based on the features found in the study by Cheong and Foo (2006). The output of the program identifies the gender of the author of five e-mail exchanges to a specified percentage of accuracy.

Methodology

This study involves a qualitative design and development of a text mining algorithm that reads an e-mail text as its input, and gives a conclusion of its processing after five e-mail text exchanges from the same author. The gender of the author is determined to a specified degree of accuracy. The text-mining algorithm is implemented in a Visual Basic program that serves as a prototype for gender detection in e-mails.

Content Analysis by Features

Cheong and Foo (2006) have outlined ten categories of linguistic features in the e-mail communications of Malaysians of mean age 22. Table 2 shows the content analysis.

Table 2: Content Analysis by Feature in 5 e-mails Exchanges

Feature	List	Frequency by Gender	
		Male	Female
Word count	Isolated words, characters, symbols	360	435
Exclamation	... ! !! !!! ?? ??? !?!	26	29
Questions	? How What Where When Did you Are you Do you know Agree?	8	10
Request for Information	Give your Let me have Give me Can I have	3	3
Reference to Emotions	like love hate happy sad nice fine sweet tired bored lazy busy hope glad good sleepy shy excited relax worry stressed sorry laugh (including list of emoticons)	6	12
Personal Information	name nickname address my...phone number you your favourite	3	4
Opinion	I think In my opinion I guess I thought I feel I found I find I felt You should You must	3	4
Self-derogatory comments	I'm not good I'm not great	0.33	0.16
Compliments	That's great Wow This good A good one	0.83	0.98
Apology	Sorry Please forgive me	1.25	0.91

Text Mining Algorithm Design

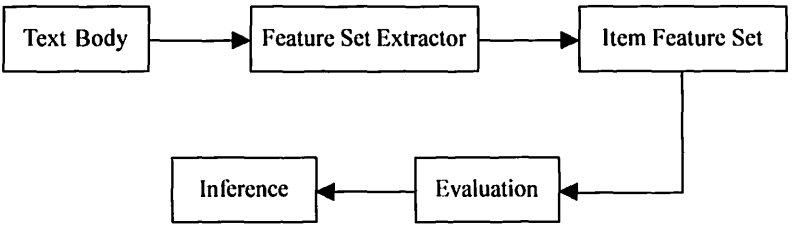

Figure 3: The Operation of Text Mining Algorithm

Text Body

Five successive e-mail exchanges from an individual subject are read into the Text Body. The Text Body highlights the text body of each e-mail before passing it to the Feature Set extractor. This process is repeated for the subsequent e-mail that belongs to the same individual.

Feature Set Extractor

For each e-mail that belongs to a set the Feature Set extractor scans for words or terms that match against the lists of each feature. The function of the Feature Set extractor is to recognise and classify significant vocabulary items in restricted natural language texts.

In general, our implementation of feature extraction relies on pattern matching together with a limited amount of lexical information, such as part-of-speech information. We neither use huge amount of lexicalised information, nor do we perform in-depth syntactic and semantic analyses of texts. This decision allows us to achieve two major goals:

- Very fast processing to be able to deal with mass data
- Domain-independence for general applicability

Item Feature Set

This tool helps to identify a feature set by listing terms or words that are common in the group. It provides a set of sample phrase or keyword to characterise each feature. This phase produces a statistical count, which is subsequently used to categorise the five text bodies as a document.

Evaluation

The evaluation tool returns a list of statistical count for the ten features in a document. It also returns the confidence level for each document being categorised. The document can be assigned to more than one category (gender).

Inference

If the confidence level is low, then typically the document would be put aside so that a human categoriser can make the final decision. Our tests have shown that, provided the set of defined features does not match the subject matter of incoming documents, the evaluation tool agrees with the human categorisers to the same degree as human categorisers agree with one another.

Program Design and Coding

There are five forms or interfaces in this prototype, which are used to implement the text mining algorithm. There are Flash Screen, Main Form, Statistic Form, Result Form and Help Form.

Flash Screen is used to show the name and version of the prototype whenever the executable file is activated. Main Form will be shown after the Flash Screen. The main source codes are contained in this form. It consists of several important parts; include the menu for the prototype (File, Edit, Tools, and Help) and all the source codes for detections and calculations for gender deception.

In the File menu, there are several options such as New, Open, Close, Save, Save As, Print and Exit. Each of these options is executable accordingly to its functions. In the Edit menu, this prototype only provides Copy and Paste options. These copy and paste options allow the user to copy from any document file to paste them inside the editor pad in this prototype. In Tools menu, there consists of Analyze and View Result options. The Analyze option allows the user to analyse the activated document according to the programmed text mining algorithm. The View Result option is able to display the entire statistic for the analysis that has been carried out by the Analyze option. In the Help option, it is only About option. This About option displays the brief description about this prototype.

Besides from the menu options, this prototype also provides short cuts to the common used options such as New, Open, Save, Print, Copy and Paste. These short cuts performs their tasks practically the same as the menu options. At the bottom of the form, this prototype provides status, date and time of execution.

The main source codes are contained in the Analyze option. These source codes are text mining process for finding the terms and symbols according to the criteria that are listed in the algorithm. The text mining process is done through a function called Find.

Statistic Form captures the figures that representing the male and female sender. These criteria are based on the aspect of linguistic and terms that are typical of each gender. The Result Form is used to display the resultant analysis on the gender of the document's author. Help Form displays a brief description about this prototype.

Program Implementation and Testing

Sixteen documents, which each consist of five e-mail exchanges of the respective individuals, are tested through the analyzer. For example, the text body of the e-mails are pasted in the text editor as shown in Figure 4.

There were ten criteria that were analysed by this prototype. These criteria include number of words, number of exclamation marks, number of questions, number of emotions, number of requests, number of personal information, number of opinions, number of self derogatory, number of compliment and number of apologies. The algorithm is used here to make the analysis automatic.

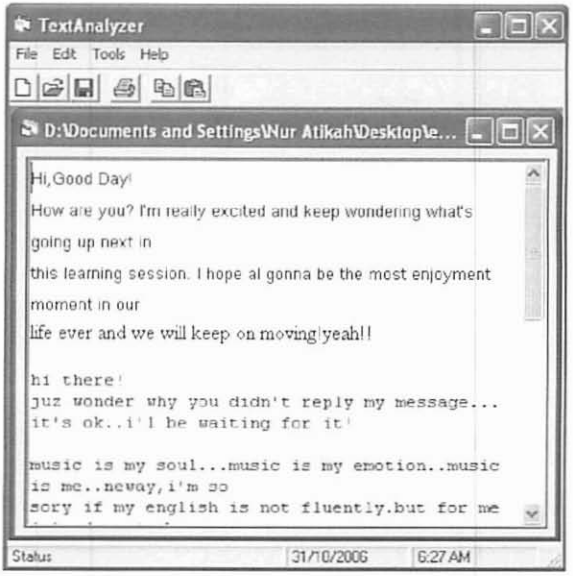

Figure 4: Text Body of Five e-mails

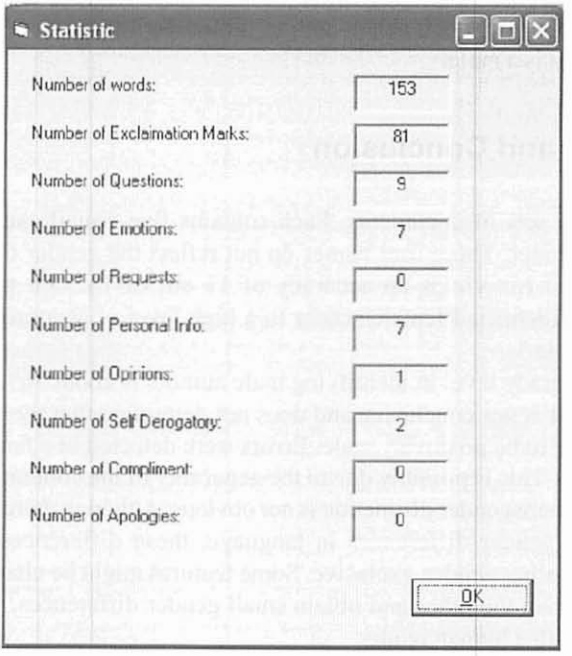

Figure 5: The Statistics after Clicking 'Analyze' Option

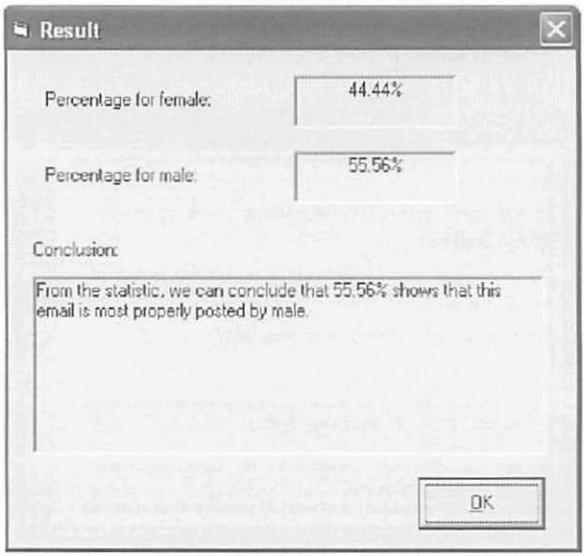

Figure 6: The Result and Conclusion

The result shows the percentages for female and male. taken as conclusion to the gender of the document's author. prototype detects that the sender of this document is a male.

Findings and Conclusion

There are 16 sets of documents. Each contains five e-mail exchanges of an individual sender. These user names do not reflect the gender of the account user. The test run yields an accuracy of 13 out of 16. Our prototype has successfully identified female sender to a high level of accuracy (> 78%) as shown in Table 3.

The accuracy level in identifying male authors is about 56%. This shows that the result is not conclusive and does not determine that the gender of the e-mail author to be positively male. Errors were detected in a female, and two male authors. This is possibly due to the generality of the content of the e-mail exchanges where gender distinction is not obvious. Although there is substantial evidence of gender differences in language, these differences are gender-preferential rather gender-exclusive. Some features might be characteristics of one gender than the other and obtain small gender differences. This result is consistent with a human reader.

Table 3: Results of the Test Run

E-mail Author	Gender	Accuracy percentage
loot_5	male	55.56
loot_28	female	77.78
loot_16	female	77.78
loot8248	female	66.67
creep10_10	female	88.89
creep2111	male	55.56
ger1007	female*	55.56
basm83	female	55.56
basm815	female	88.89
happy17_dragonfruit	female	77.78
happy022_022	male	55.56
pump104	male	55.56
sotong_123456789	female*	55.56
sotong_06	male*	55.56
sotong7_batusatu	female	55.56
sotong08	female	77.78

Note: *gender is wrongly identified by the program

Conclusion

In this research, we have described our notion of text mining and product for text mining application – detecting gender of e-mail author. As this application shows text mining to date can be used as an effective forensic tool that supports decision making by preparing and organizing unstructured textual data (e-mail) and by supporting the extraction of relevant information from a large amount of unstructured textual data through automatic pre-selection based on user-defined criteria (feature set). Using automatic mining processes to organize and scan huge repositories of textual data can significantly enhance both the efficiency and quality of a routine task while still leaving the more challenging and critical part of it to the one who can do it best, the human reader.

Future Research Works

Our implementation of feature extraction relies on pattern matching together with a limited amount of lexical information, such as part-of-speech information. We neither use huge amount of lexicalized information, nor do we perform in-

depth syntactic and semantic analyses of texts. The Item Feature Set requires an exhaustive database characterizing each of the ten features, which is used to categorize the five text bodies of a document. Future work can be done in these areas to further enhance the reliability and usability of the present work.

References

Bruckman, A.S. (1993). *Gender Swapping on the Internet.* Paper presented at The Internet Society, San Francisco, CA, August 1993. Available at: ftp.media.mit.edu/pub/asb/papers.

Carli, L. (1990). Gender, language, and influence. *Journal of personality and Social Psychology*, 5, 941-951.

Cheong, L. M. and Foo, F.L. (2006). *Detecting Gender Deception on the Internet through chatroom lingo: A Malaysian Perspective.* An IRDC project.

Fitzpatrick, M., Mulac, A. and Dindia, K. (1995). Gender-preferential language use in spouse and stranger interaction. *Journal of Language and Social Psychology*, 14, 18-39.

Holmes, J. (1988). Paying compliments: A sex-preferential politeness strategy. *Journal of Pragmatics*, 23, 445-465.

Holzman, L. E., Fisher, T. A., Galitsky, L. M., Kontostathis, A. and Pottenger, W. M. (2003). A software infrastruture for research in textual data mining. *International Journal on Artificial intelligence Tools*, 13 (4), 829-849.

Kontostathis, A., Galitsky, L. M., Pottenger, W. M., Roy, S. and Pheps, D. J. (2003). *A Comprehensive Survey of Text Mining.* Springer-Verlag.

McMillan, J. R., Clifton, A. K., McGrath, D. and Gale, W. S. (1977). Women's language: uncertainty or interpersonal sensitivity and emotionality. *Sex Roles*, 3, 545-559.

Mulac, A. and Lundell, T. L. (1986). Linguistic contributors to the gender- linked language effect. *Journal of Language and Social Psychology*, 5, 81-101.

Mulac, A., Studley, L. B. and Blau, S. (1990). The gender-linked language effect in primary and secondary students' impromptu essays. *Sex Roles*, 23, 439-469.

Mulac, A., Wiemann, J.M., Widenmann, S. J. and Gibson, T.W. (1988). Male/ female language differences and effects in same-sex and mixed-sex dyads: the gender-linked language effect. *Communication Monographs*, 55, 315-335.

Tannen, D. (1994). *Talking from 9 to 5*. London: Virago Press.

Thomson, R. and Murachver, T. (2000). Predicting gender from electronic discourse. In Weber, R. P. (1990). *Basic content analysis* (2nd Ed.). Newbury Park, CA: Sage.