USING TEXT MINING ALGORITHM TO DETECT GENDER DECEPTION BASED ON MALAYSIAN CHATROOM LINGO

BY

ASSOCIATE PROFESSOR DIANNE L.M. CHEONG
NUR ATIQAH SIA BINTI ABDULLAH@SIA SZE YIENG

OCTOBER 2006

UNIT PENYELIDIKAN PROTRAD
FAKULTI TEKNOLOGI MAKLUMAT DAN SAINS KUANTITATIF
UNIVERSITI TEKNOLOGI MARA
40450 SHAH ALAM
E-mail: protrad@tmsk.itm.edu.my
Homepage: http://protrad.tmsk.itm.edu.my

Tarikh : 29 September 2005
Surat kami: 100-PT/RU/3/6-(43/05)

Prof. Madya Dianne L. M. Cheong
Pensyarah (Ketua Projek Penyelidikan)
Fakulti Teknologi Maklumat dan Sains Kuantitatif
UiTM, Shah Alam

Puan,

CADANGAN PENYELIDIKAN:
USING TEXT MINING ALGORITHM TO DETECT GENDER DECEPTION BASED ON MALAYSIAN CHATROOM LINGO

Dengan segala hormatnya perkara di atas adalah dirujuk.

Sukacita dimaklumkan bahawa ProTRAD pada 21 September 2005 telah membuat keputusan:

i. Bersetuju meluluskan cadangan penyelidikan yang dikemukakan oleh Puan dan Nur Atiqah Sia Abdullah.

ii. Tempoh projek penyelidikan ini ialah **12 bulan**, iaitu mulai 1 November 2005 hingga 31 Oktober 2006.

iii. Kos yang diluluskan ialah sebanyak RM 8,000.00 sahaja.

iv. Penggunaan geran yang diluluskan hanya akan diproses setelah perjanjian ditandatangani.

v. Semua pembelian peralatan yang kosnya melebihi RM 500.00 perlu menggunakan Pesanan Jabatan Universiti Teknologi MARA (LO). Pihak Puan juga dikehendaki mematuhi peraturan penerimaan peralatan.

vi. Kertas kerja boleh dibentangkan dalam seminar setelah **75% draf awal Laporan Akhir projek** dihantar ke Unit Penyelidikan ProTRAD untuk semakan. Walau bagaimanapun, permohonan kepada Institut Penyelidikan, Pembangunan dan Pengkomersilan (IRDC) untuk tujuan pembentangan perlu dibuat terlebih dahulu.

vii. Pihak Puan dikehendaki mengemukakan Laporan Kemajuan Projek Penyelidikan bagi tempoh 4 bulan pertama, iaitu sehingga 28 Februari 2006, 4 bulan berikutnya iaitu sehingga 30 Jun 2006 dan 4 bulan berikutnya iaitu sehingga 31 Oktober 2006. Laporan akhir

ABSTRACT

E-mail is used for communication between strangers and friends. It can be a fantasy playground for identity experimentations where players take on an imaginary persona and interact with each other in the virtual world. In communication, knowing the identity of those whom you communicate is essential for understanding and evaluating an interaction. However, the presentation of self in the virtual world is often a conscious and deliberate endeavour. Therefore, gender deception is difficult and 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 designed text algorithm, a prototype in Visual Basic is developed. The prototype was tested with 16 documents; each consists of 5 e-mails exchanges of respective individuals. Out tests have shown that the prototype is at 81.3% of accuracy level. This is consistent with a human reader of the documents. The tested prototype will be a tool to assist interest 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 to detect gender, program to detect gender, gender deception.*

TABLE OF CONTENTS