Internet Safety Technical Task Force
Technology Submission – Text Attribution Tool

Appen Speech and Language Technology Inc
http://www.appen.com.au

JULY 21, 2008

ABSTRACT
Appen's Text Attribution Tool (TAT) was developed under US government funding and sponsorship to meet identified needs of intelligence and law-enforcement organizations. The TAT has the capability to analyze text data and to produce a demographic and psychometric profile of its author. This capability can be applied to a number of problems within the realm of internet usage as identified by the Internet Safety Technical Task Force to provide a potential solution. These problems include: age verification/detection of age discrepancies on social network sites; detection of potentially predatory users/user-behaviors; and detection of potentially at-risk users/user-behaviors.

Keywords
profiling, filtering, searching.

Functional Goals
Please indicate the functional goals of the submitted technology by checking the relevant box(es):
☒ Limit harmful contact between adults and minors
☒ Limit harmful contact between minors
☐ Limit/prevent minors from accessing inappropriate content on the Internet
☐ Limit/prevent minors from creating inappropriate content on the Internet
☐ Limit the availability of illegal content on the Internet
☐ Prevent minors from accessing particular sites without parental consent
☒ Prevent harassment, unwanted solicitation, and bullying of minors on the Internet
☒ Psychometric profiling

PROBLEM INTRODUCTION

Age verification and detection of age discrepancies on social network sites
Although age restrictions are often in place on social network sites and many other forms of computer mediated communication (CMC) these restrictions typically rely upon the honesty of the individual submitting the information, and be easily circumvented by registering an account with a false date of birth. Individuals seeking to misrepresent their age include underage users of social network sites, and adult predators attempting to establish contact with young people.

Individuals seeking to misrepresent their age might use prototypical features of the communicative style of the age-identity they are attempting to assume (most problematically "teen"). These surface communicative features/strategies might include elements of slang/jargon, specialized vocabulary, or context specific language varieties such as *txt spk* (SMS text speak). The strength of the TAT in detecting individuals misrepresenting their age is that it does not focus on features which are surface or proto-typical (and are, arguably, for this reason easier to modify). Rather, when assigning demographic values, the TAT uses advanced linguistic processing to focus on covert features hidden from the individual writers.

Identification of potentially predatory users and user-behaviors
Bullying, and sexually predatory behaviors have been identified in the mass media as issues for social networking sites and for internet users more broadly. Using text content which has been identified as having been produced by known predators, or which displays overtly predatory characteristics, the psychometric profiling functions of the TAT can be tuned to identify these predispositions and behaviors.

Identification of potentially at-risk individuals and behaviors
A further problem to which the TAT might be applied is the detection of at-risk individuals. This may be minors or others who have been subject to harassment, unwanted solicitation and/or bullying. Detection of these individuals may lead to identification of perpetrators of predatory behavior, and might aid prevention of self-harm.

PROPOSED SOLUTION

The Text Attribution Tool operates using a similar underlying function when acting as a solution to each of the problems identified in the preceding section. The common aspects of the TAT's underlying functionality will be elaborated in detail with respect to Age Detection; with respect to the application of psychometric profiling to the identification of potentially predatory individuals and at-risk internet users, the same core functionality should be assumed.

Problem – Age Detection
The need to detect the age of users of social networking sites arises because for various reasons users desire to pretend to be an age they are not. In order to do this convincingly, they may adopt language traits in order to

sound younger or more grown up. However, the most readily adopted traits will be surface traits, leaving the underlying and covert language characteristics unchanged.

Studies show that there are many linguistic markers of age [2]. The TAT uses analyses at a number of different levels in order to generate feature sets for each document passed to it. These documents can be profiled on dimensions including gender, native language, and level of education, (see following section for further elaboration) but age is of greatest salience here. The feature sets extracted from a text include, but are not limited to, character level (e.g. word length, punctuation use), lexical (e.g. part-of-speech, use of function words) and structural (e.g. use of paragraphs).

The TAT determines author age by passing a document's features to a machine classifier (an SVM model; SMO as implemented in WEKA [6]). By using features other than surface level, the TAT is able to identify constructs that reveal an author's true age.

Problem – Psychometric Profiling

In the same manner that the TAT predicts age, it can also predict other demographics (age, gender, native language, level of education and country of origin) as well as psychometric traits. As currently trained/tuned: the English language model of the TAT analyses for the commonly accepted 'Big 5' model consisting of *neuroticism, extraversion, openness to experience, agreeableness and conscientiousness*; and the Arabic language version of the TAT analyses for a customized version of the Eysenck Personality Questionnaire Revised (EPQR-S) traits consisting of *extraversion, lie/social-desirability, neuroticism/emotionality, and psychoticism/tough-mindedness* [1]. Different combinations of the features extracted are presented to classifiers, one per trait, and predictions for each document are made.

For deployment in a social-network site (SNS) scenario the TAT would store its trait predictions alongside each document in the native SNS database. The TAT interface would then be used to retrieve records matching a specific profile.

Predatory and At-risk Behavior Identification

The principles behind the TAT are generalizable and its functionality modular. This means that given training data consisting of documents that are known to represent some factor of interest (eg. Depressed authors, texts aimed at grooming), enables the TAT to determine the linguistics features best suited to diagnosing these behaviors.

In addition to the relevant psychometric traits covered by the current version of the TAT, the English language model can be re-trained/tuned carry out psychometric profiling analysis on traits including psychoticism, depression and other traits which might be indicative of predatory and/or at-risk individuals and behaviors[5].

Additional information

Technical attributes – features and functionality.

There are three main tasks that a user is able to perform with the TAT:

- Predict traits (including age) of the author of a given document
- Retrieve documents from a database fitting a given author profile.
- Retrieve documents from a database whose author profile is similar to that of a given document

Solutions/Limitations

The TAT is intended to support human analysis by indentifying candidate material for more detailed assessment. It is not intended to provide definitive analysis. In its law-enforcement and intelligence configurations, the user brief was to provide an investigative profiling tool rather than evidentiary tool.

Strengths:

It is a strength that the TAT is not reliant on content words – this makes it possible to generalize TAT functionality to the analysis of diverse phenomena in multiple languages.

The TAT's reliance on training data is also a strength – this ensures classifications are based on empirically driven statistically derived data, rather than subjective observations and theory

Weaknesses:

The TAT is a fully operational prototype now undergoing final development prior to release.

Hardware/software requirements

The TAT is a server based tool. End-users access the tool through a web-browser.

End-user aptitude

Low to mid level computer skills only, would be required. The tool uses a simple web interface, and it was a requirement of the development sponsors that the interface be user friendly and oriented to the skill set of a main-stream law enforcement officer (i.e. not to the higher skill levels of officers working in a high tech crime environment).

Standards

The TAT and its contexts of use are novel, and as far as Appen is aware there are no relevant standards, existing or planned.

Reliance upon, and use of law and policy

For analysis of open source information, we do not believe there is any reliance upon law and/or policy. Analysis of email data (per Appen's intelligence and law enforcement brief) warrants would be required.

Viability of the technology in the US and Internationally
The TAT has been successfully prototyped in both English and Arabic, and was designed to be adaptable and modular in order to allow for generalization to new traits where needed, and to allow for extension to additional languages.

APPEN OVERVIEW
Appen develops and markets sophisticated computer-based speech and language technology products and services for major international information and communication companies and government organizations. Appen is recognized as a global leader in the quality and quantity of its products and services.

Appen's products cover a range of text applications such as Natural Language Processing ("NLP"), as well as speech recognition, text-to-speech (speech synthesis), phonetic search, machine translation applications. These are developed from the high-end fusion of Appen's computer science and linguistics capabilities. Appen works in over 80 languages. It is headquartered in Sydney, Australia, with a US subsidiary Appen Speech & Language Technology Inc.

Appen has been operating for 12 years and is privately owned. The company has shown consistent growth since its creation. It is profitable and has a strong balance sheet.

Major customers are most of the world's large ICT companies, including Microsoft, IBM, Google, Motorola, Siemens, Toshiba and Nokia. Appen is also active in the US government sector in the defense and homeland security sectors. Appen has a number of strategic relationships with its major clients but is independent of any other organization.

Appen maintains an active R&D program with a growing portfolio of IP and patents. Developmental work includes:

- Tools for processing speech materials, especially to support the development of speech recognition and speech synthesis systems, and tools such as lemmatisers, tagging tools & spelling standardization or less commonly taught languages.
- Tools for content extraction from text materials, in particular the extraction of entity related information inn nominated categories
- Licensable software products for use by government intelligence and law enforcement agencies for forensic authorship profiling and tracking of persons of interest on-line.

Appen has a professional staff of approximately 50 people, most with post-graduate qualifications in computer science and/or linguistics. Key development people include

Dr. Julie Vonwiller (Project Director and Appen founder)
Dr Julie Vonwiller is a Director and co-owner of Appen. She has worked in the field of speech science and technology for more than 15 years.

Dr. Vonwiller has worked both in private industry and research, and has skills in Project Management of software-oriented projects, successful commercialization of R&D, management of multi-disciplinary research and development teams and in-depth familiarity with leading edge speech technology techniques. Her experiences ranges widely across lexical and text processing in many languages.

Julie has led several successful development projects for the US Government, including the Text Attribution Tool software development and the Data Stream Profiling software development.

Dr Scott Nowson (Computational Linguist)
Dr Scott Nowson is an experienced Computational Linguist with deep experience in the linguistic study of online diaries and weblogs. His research experience encompasses personality theory, language analysis tools, language with personality and gender, and language in computer mediated communication. Scott has extensive experience with the collection and processing of text corpora and sociobiographic data, and with data annotation for machine learning.

Scott was educated at the University of Edinburgh, where he completed a B.Sc. degree with Honors in 1999. Subsequently he was awarded a M.Sc. in Cognitive Science (2001) and a Ph.D. in Informatics (2005), also from Edinburgh University. Prior to joining Appen in 2007, Scott worked as a Research Fellow at Macquarie University investigating the application of natural language technology within the financial domains.

Stephen Norris (Software Development Manager)
Stephen Norris has provided IT consulting and contracting services to Appen since the company was first established. In June 2005, Stephen joined Appen as the IT group manager.

Stephen has extensive experience in software development and IT management, as well as being a highly skilled programmer and system designer himself. He has worked on projects ranging from high-performance, high-availability financial information systems through web-based stock trading, banking and transaction processing systems to document processing and management systems used for on-line billing and related systems.

Dr. Judith Bishop (Linguist)
Judith Bishop has worked as a Linguist and Project Manager in Appen since December 2004. Her experience in Appen has included research in phonetic analysis of speech, project management of large scale speech databases

and language analysis in unusual languages. Judith has been a member of the Appen team working on the Text Attribution Tool development for the US Government.

Judith was awarded a PhD in Linguistics from the University of Melbourne (Australia) and completed graduate studies at Washington University in St Louis (USA) and University of Cambridge (UK).

Phil Hall (Advisor – Forensic Linguist)
Phil has a BA (Honors First Class) in Linguistics and the University Medal from Macquarie University and has received a number of scholarships, including a Scholarship from the Linguistic Society of America. He is completing a PhD in Linguistics focusing on forensic linguistics from Macquarie University. Phil has taught at university level and worked in a management capacity before joining Appen. Phil has acted as a consultant forensic linguist for the New South Wales Police in a number of extortion and stalking cases.

BUSINESS MODEL OVERVIEW
The Appen technology referred to in this submission can be made available under different business models. Typical users are envisaged to be

- Law enforcement organizations (federal, state and local
- Government agencies, especially in intelligence processing
- Relatively large commercial organizations who wish to provide screening in regard to their on-line user populations envisaged to be

Typical business models will be:

a) A system license in the form of a Software Development Kit (SDK) which will allow the system operator to incorporate the Appen technology into its own systems
b) A usage licensing agreement, with Appen providing the necessary servers and support, and users paying according to their level of usage.

Appen is in discussion with partners in the US market in relation to establishing all necessary support for deployment for the technology.

The technology has been delivered to the US Government as working prototypes, and is now undergoing final development ready for commercial deployment. Specific pricing has not yet been released.

MORE INFORMATION
Feel free to provide a URL that readers can go to for more information. This may include videos, detailed specs, or anything else that might be relevant. Indicate in this document what the readers might find if they go to the URL. This is a great place for information you would like to include that does not otherwise fit the structure of this document.

http://www.appen.com.au/files/products/TAT.pdf

CONTACT INFORMATION
Contact: Phil Hall (phall@appen.com.au)
Telephone: +61 2 9468 6321
Nth Tower, Level 6, 1 Railway Street
Chatswood, NSW 2067, Australia

CERTIFICATION
I certify that I have read and agree to the terms of the Internet Safety Technical Task Force Intellectual Property Policy."

REFERENCES
1. Estival, D., Gaustaad, T., Pham, S.B., Radford, W., and Hutchinson, B. Author Profiling for English Emails. In Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING). 2007.
2. Koppel, M., Schler, J., Argamon, S., and Pennebaker, J. Effects of Age and Gender on Blogging. In AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs. 2006.
3. Mairesse, F., Walker, M., Mehl, M., and Moore, R. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. Journal of Artificial Intelligence Research, 30, pages 457-500, 2007.
4. Oberlander, J., and Nowson, S. Whose thumb is it anyway? Classifying author personality frm weblog text. In Proceedings of the 44th Annual Meeting of the Association of Computational Linguistics (ACL) 2006.
5. Rude, S.S., Gortner, E.M., and Pennebaker, J.W. Language use of depressed and depression-vulnerable college students. Cognition and Emotion, 18, 1121-1133. 2004.
6. Witten, I.H., and Eibe, F. Data Mining: Practical machine learning tools and techniques. Morgan Kaufman, San Francisco, CA. 2005.