D53 WISER: Automatically Classifying Written Statements as True or False

Carole E. Chaski, PhD, ALIAS Technology, LLC, Institute for Linguistic Evidence, 25100 Trinity Drive, Georgetown, DE 19947; Angela Almela, PhD, Universidad Catolica de Murcia, Campus de Los Jerónimos, 135, 30107 Guadalupe, Murcia , SPAIN; Gary Holness, PhD, Delaware State University, Computer Science Dept, 1200 N Dupont Highway, Dover, DE 19901; and Larry Barksdale, MA, University of Nebraska, Forensic Science Dept, 211 Entomology Hall, Lincoln, NE 68583*

After attending this presentation, attendees will better understand current approaches and results in computational classification of written statements as true or false. Attendees will be able to compare current and traditional approaches and evaluate methods on linguistic sophistication, psycholinguistic realism, accuracy, data requirements, and implementability.

This presentation will impact the forensic science community by providing measures of reliability to automated deception detection of written statements and an overall framework for investigators to evaluate different methods.

Witness Statement Evaluation Research (WISER) is a project using automated text analysis and statistical classifiers to determine the best protocol for computational classification of true and false statements in the forensic-investigative setting. In this research project, current and traditional approaches are being compared and tested and methods are being evaluated on linguistic sophistication, psycholinguistic realism, accuracy across different datasets, data quality requirements, and implementability. The most well-known and famous method for deception detection or veracity assessment is Sapir's Scientific Content Analysis (SCAN) described at a website and methods following it are generally known as statement analysis.[1] SCAN has been automated to a great degree with reported accuracy results of 71% by Fitzpatrick and Bachenko on a dataset that has not been made available to the WISER project.[2] Another approach is Pennebaker's Linguistic Inquiry and Word Count (LIWC), a word count and categorization program that has been applied to authorship identification, gender and age estimation as well as deception, even though its original purpose and design was to classify writing samples into personality types for psychological assessment.[3] A third approach that is popular among computer scientists working in text classification is the Bag Of Words (BOW) model, a term invented by Harris and developed by Salton and McGill in which each text is seen as a list of words and their frequencies without regard to any syntax, semantics, or grammatical relations.[4,5] Both LIWC and BOW have been tested as deception detection methods by Almela and Mihalcea and Strapparava.[6,7] A fourth approach is based in Text Analysis Toolkit Toward Linguistic Evidence Research (TATTLER), a text analysis tool; TATTLER combines linguistic analysis at the phonological, syntactic, and lexico-semantic levels and has been applied to deception detection classification on both experimental and high-stakes datasets.[8,9] The TATTLER results showed a remarkable difference between the experimental data, in which students were asked to write two narratives of a traumatic experience, one truthful and the other false, and high-stakes data, actual statements from real criminal investigations with non-linguistic evidence of their veracity or falsehood. Using the same TATTLER algorithm, the texts in the experimental data were correctly classified using leave-one-out cross-validation 71% of the time, while the texts in the high stakes data were correctly classified using leave-one-out cross-validation 93% of the time.

This result demonstrates that there is a real difference between lies told in an experimental setting (i.e., in a classroom as part of a class assignment) and lies told in a police investigation. Further, this result demonstrates that linguistic sophistication may be needed to develop highly reliable linguistic variables. Moreover, this result demonstrates that different statistical approaches which may work better with different linguistic variable sets (e.g., Support Vector Machine or Discriminant Function Analysis) may fit better with TATTLER variables than with LIWC or BOW. In this presentation, the results of testing the text-analytic linguistic variables from LIWC, BOW, and TATTLER using several different statistical approaches such as Discriminant Function Analysis, Support Vector Machine, and others, on three different datasets (experimental from the classroom setting on trauma narratives; experimental from an electronic venue for opinion mining; and, high-stakes data from real case investigations) are reported. Further, the methods on linguistic sophistication, psycholinguistic realism, accuracy over different datasets, data quality and quantity requirements, and implementability within the investigative, interview and interrogation workflow are evaluated.

References:

1. Lesce, T. 1990. SCAN: Deception Detection by Scientific Content Analysis. Law and Order Volume 38:8. Posted at http://www.lsiscan.com/id37.htm
2. Fitzpatrick, E. and Bachenko, J. 2013. "Detecting Deception across Linguistically Diverse Text Types." *Linguistic Society of America Annual Meeting*. Boston, MA.
3. Pennebaker, J. and Francis, M. 1999. Linguistic Inquiry and Word Count: LIWC. Erlbaum Publishers.
4. Harris, Z. 1954. "Distributional Structure." Word 10 (2/3): 146–62.
5. Salton, G. and McGill, M. 1983. Introduction to Modern Information Retrieval. McGraw-Hill.
6. Almela, A., Valencia-Garcia, R. and Cantos, P. 2013. "Seeing through Deception: A Computational Approach to Decite Detection in Spanish Written Communication." Linguistic Evidence in Security, Law and Intelligence Volume 1:1. http://www.lesli-journal.org/ojs/index.php/lesli/article/view/5
7. Mihalcea, R. and Strapparava, C. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics).
8. Chaski, C. E. 2013. "ILER: A Web-Accessible Resource for Research in Forensic Linguistics." Linguistic Evidence in Security, Law and Intelligence, Volume 1:1. http://lesli-journal.org
9. Chaski, C.E. Barksdale, L. and Reddington, M. 2014. "Collecting Forensic Linguistic Data: Police and Investigative Sources of Data for Deception Detection Research." *Linguistic Society of America Annual Meeting*. Minneapolis, MN.

Deception Detection, Text Classification, Witness Statements