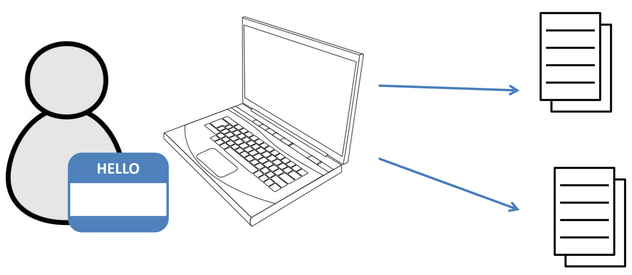
CERIAS

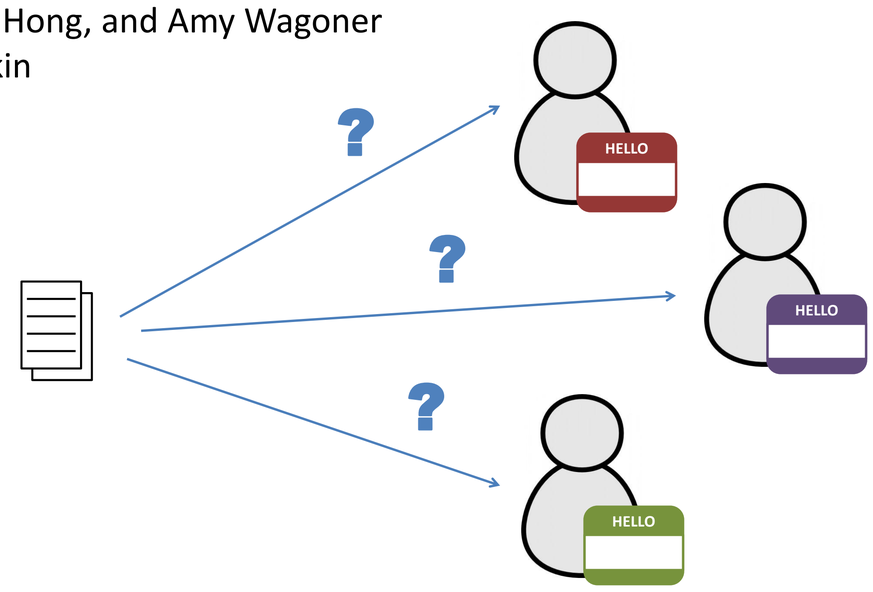
The Center for Education and Research in Information Assurance and Security

Determining Authorship with Style

Students: Lauren M. Stuart, Saltanat Tazhibayeva, Ji Hyeon Hong, and Amy Wagoner
Advisors: Julia M. Taylor, Victor Raskin



Authors write texts and publish them, with or without bylines. How do we attribute an anonymous or pseudonymous text to a known author, discriminate between two unknown authors, or verify the authorship of a disputed text? The author may have (perhaps unknowingly) left some clues.

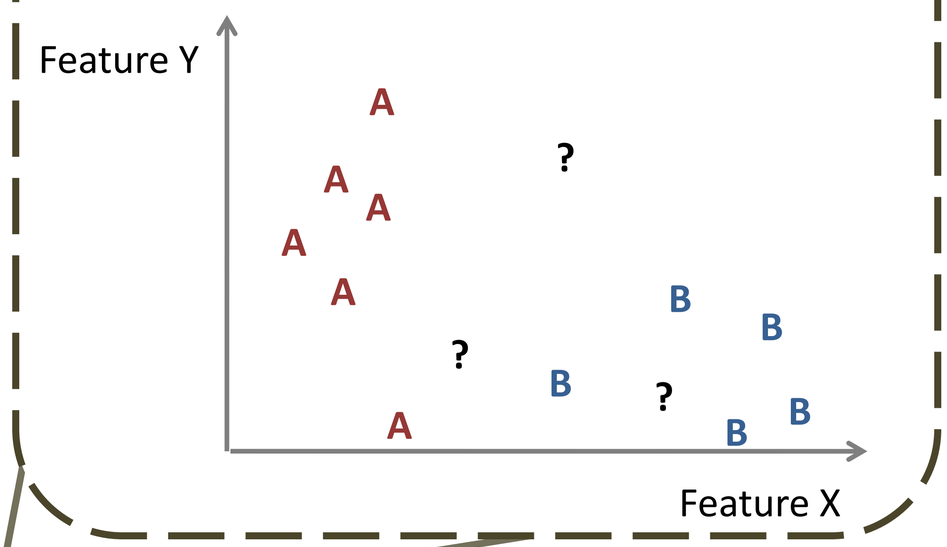


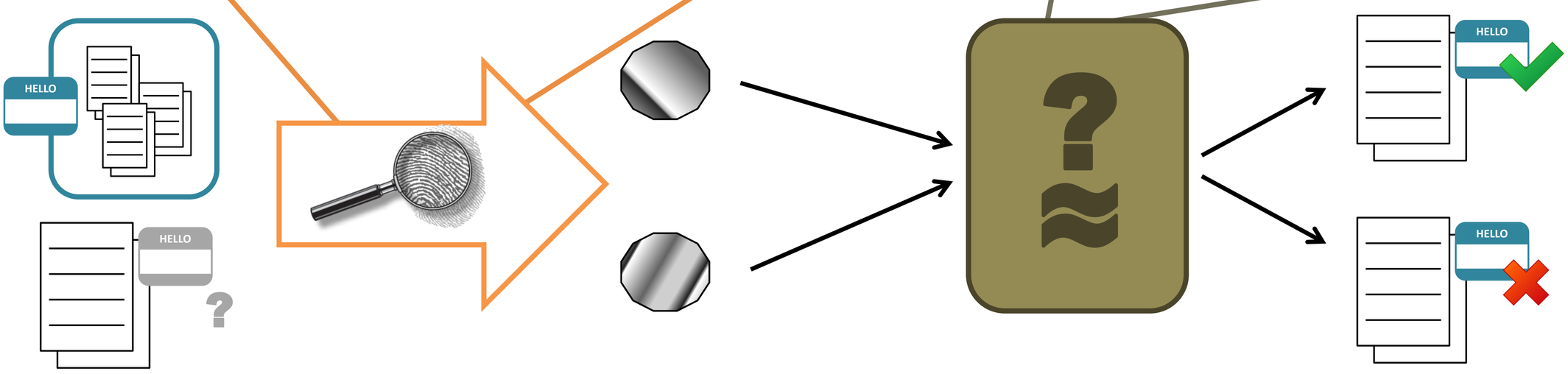
Stylometry is the characterization of a text's author by capturing the style of writing as measurable features expressed in the text[6]. There are three major applications for stylometry: authorship verification, authorship attribution, and deception detection. In authorship verification[3,3], the styles of a text and the body of work of the supposed author are matched. In authorship attribution[1,2], several authors' works are evaluated for style and compared to the unknown text. Deception detection[4] attempts to find instances in which an author has tried to conceal his or her style[6,7], perhaps by imitating another author. Most stylometric studies focus on punctuation marks (!,;"|) and other special characters, sentence and word length, word frequencies, and part-of-speech frequencies. Some look at sequences of words or part-of-speech tags, raw keystroke data, or capitalization.

Common stylometrics

Total words: 123
Total sentences: 8
Average words per sentence: 16
Word frequencies:
  (content) author: 4; 0.0325%
  (function) of: 6; 0.0488%
Frequent word bigrams: authorship verification
Frequent POS bigrams: {CC} {NN}
Unique words: 83
Numbers
Readability measures
Misspelled words
Letter n-grams
Topic words

Example comparison on two features





Research Questions

1. Can we use current tools[7] and techniques on a wider range of corpora, such as non-fiction, Twitter, academic papers?
2. How can we include syntactic and semantic structures as stylometrics?
3. Do some stylometrics vary more between topics[5] than between authors?
4. With current tools, what are the tradeoffs between accuracy and resources such as time, operations, and width of datasets?

1. Luyckx, K., & Daelemans, W. (2011). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, *26*(1), 35-55. *The Use of Stylometry for Email Author Identification: A Feasibility Study*
2. Gallagher, Gary W. (1986, July) A Widow and Her Soldier: Lasalle Corbell Pickett as Author of the George E. Pickett Letters**.** *The Virginia Magazine of History and Biography*, Vol. 94, No. 3, pp. 329-343.
3. Koppel, M., & Schler, J. (2004, July). Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning* (p. 62). ACM.
4. Pearl, L., & Steyvers, M. (2012). Detecting authorship deception: a supervised machine learning approach using author writeprints. *Literary and linguistic computing*, *27*(2), 183-196.
5. Van Peer, W. (1989). Quantitative studies of literature. A critique and an outlook. *Computers and the Humanities*, *23*(4), 301-307.
6. Kacmarcik, G., & Gamon, M. (2006, July). Obfuscating document stylometry to preserve author anonymity. In *Proceedings of the COLING/ACL on Main conference poster sessions* (pp. 444-451).
7. Brennan, M., & Greenstadt, R. (2009, April). Practical attacks against authorship recognition techniques. In *Proceedings of the Twenty-First Conference on Innovative Applications of Artificial Intelligence (IAAI), Pasadena, CA*.

CERIAS®

PURDUE UNIVERSITY