Stylometric Report – Heartland Institute Memo

Patrick Juola, Ph.D.

March 13, 2012

Summary

1 As an expert in computational and forensic linguistics, I have reviewed the alleged Heartland memo to determine who the primary author of the report is, and more specifically whether the primary author was Peter Gleick or Joseph Bast. I conclude, based on a computational analysis, that the author is more likely to be Gleick than Bast.

Qualifications

2 I am a tenured Associate Professor of Computer Science at Duquesne University, Pittsburgh, PA. I am also Director of the Evaluating Variations in Language Laboratory, also at Duquesne.

3 I received a Ph.D. and M.S. in computer science at the University of Colorado and an M.S.-level "certificate" in cognitive science from the University of Colorado, as well as a B.S. in electrical engineering at the Johns Hopkins University, Baltimore. I received post-doctoral training in experimental psychology as a postdoc at the University of Oxford, UK. I have published numerous journal articles and book chapters on the computational inference of document authorship via the statistical analysis of linguistic features.

4 I am a frequent ad-hoc reviewer on subjects pertaining to authorship attribution, stylometry, digital humanities, and text analysis for a number of journals, including LLC (formerly Literary and Linguistic Computing), JASIST (Journal of the American Society for Information Systems Technology), and SPE (Software Practices and Experiments). I am a member and Certified Forensic Consulant of the American College for Forensic Examiners Institute and a founding member of the Research Practitioner Track of The Association for Linguistic Evidence (TALE).

5 I am the primary architect and designer of the JGAAP (Java Graphical Authorship Attribution Program) authorship analysis system. This system, NSF-funded for nearly $2 million, is one of the most powerful and widely used systems for non-traditional (statistical) authorship attribution now available.

6 I have qualified as an expert in forensic linguistics in the Elizabeth, New Jersey branch of the Executive Office for Immigration Review.

7 I am also Director of Research for J Computing, Inc. (dba Juola & Associates, also J&A), a Pennsylvania corporation specializing in text and authorship analysis. Pursuant to that job, I have been asked to analyze the Heartland Institute memo as detailed in the following sections. J&A is performing this analysis *pro bono* based on a perceived great public interest in the outcome; I personally am receiving no compensation other than my normal agreement with J Computing. Neither the company's compensation nor mine depend upon the outcome of this matter.

Background and Assignment

8 On or about February 14, 2012, various documents were released purporting to come from the Heartland Institute. One key document in particular, entitled "Confidential Memo: 2012 Heartland Climate Strategy" purported to be (as the title suggests) a confidential memo detailing the 2012 strategy regarding policy and funding strategies regarding "the supposed risks of global warming"

and climate change. These releases were unauthorized; the leaker was initially identified only as "Heartland Insider" but has since been tentatively identified as Dr. Peter Gleick, an environmental scientist working at the Pacific Institute.

9 While Heartland confirmed the legitimacy of many of the released documents (inter alia, in a press release dated February 15, 2012[1]), it specifically denies the legitimacy of the confidential memo. From that release: "[The confidential memo] is a total fake apparently intended to defame and discredit The Heartland Institute." Perhaps needless to say, the leakers have not admitted to faking this document and hence its authenticity is in dispute.

10 Despite, or because of, this denial, these documents have become widely distributed and discussed and there is widespread public interest on the actual authorship of the confidential Heartland memo.

11 On February 23, 2012, Juola & Associates was approached by Anthony Watts via email and asked "to determine the likely author of the document based on other available writing samples."[2] Watts also provided J&A with several sample writings by Gleick which were used in part in the analysis.

Process

12 The two logical candidates for authorship of the disputed document (and the only two named seriously in the public discussion) are Joseph Bast, President of the Heartland Institute, and Peter Gleick. We therefore assume that the author is one of those two (and explicitly neglect less likely scenarios, such as an unnamed Heartland staff member writing the memo based on previous work by Best).

13 One potential issue with client-supplied samples in general is that a deceptive client may cherry-pick unrepresentative samples in order to obtain the best possible result. With no disrespect intended to Mr. Watts, we supplemented his Gleick documents with other writings obtained from the Internet as a possibly cleaner sample of Gleick's style.

14 We therefore collected samples of other published work by both Bast and Gleick from the Internet. From Best, we collected approximately 8000 words of roughly contemporaneous text from documents entitled "Are We Doomed?" "Heartland Replies To Media Matters," "Heartland Replies To Nature," "Heartland Replies To Science," "Is Jon Huntsman Stupid," and "You Call This Consensus?" From Gleick's published writings on the Huffington Post, we collected approximately 4000 words of documents including "A Cost of Denying Climate Change: Accelerating Climate Disruptions Death And Destruction," "Another Cost of Bottled Water: Environmental Injustice and Inequity," "Fox's Latest Assault On Climate Science: Attack *Sponge Bob*," "It's Hotter Than It Used to Be; It's Not as Hot as It's Going to Be," "More Climate BS At *Forbes*: Hiding the Energy Imbalance of The Planet," "What Do You Know? Water Conservation And Efficiency Actually Work," and "When Climate Changes Come and Water Policies Fail. Pray For Rain?" From Watts we received approximately 4000 words of Gleick-authored documents: "20100510 Gleick At Lucia," "20111016 Gleick Review Of Laframboise," "Comment On James Taylor," "Peter Gleick Responds," "Remarkable Editorial Bias On Climate Science At The Wall Street Journal," and "The 2011 Climate BS of the Year Awards."

15 Using the JGAAP stylometric analysis system[3], we performed several analyses to determine whether Gleick or Bast was more likely to have written the disputed memo.

16 The simplest method of analysis would be to use the the harvested set of undisputed Gleick and undisputed Bast documents as "training documents," allow the computer to "learn" their respective style, and then determine which style more closely matched that of the disputed memo.

[1] `http://heartland.org/press-releases/2012/02/15/heartland-institute-responds-stolen-and-fake-documents`
[2] Watts, personal communication
[3] `http://www.evllabs.com/jgaap/w/index.php/Main_Page`

17 Unfortunately and as discussed below, analysis of this document proved particularly problematic in several ways. Initial analyses uniformly returned unclear or ambigous results. We therefore found it necessary to perform additional (calibration) testing to find analytic methods that are/more likely to deliver accurate results under the conditions specific to this problem.

18 For this calibration, we used a testing technique called "leave-one-out validation." In essence, this method treated each *known* document in sequence as a temporarily unknown document and attempted to infer the author of that document based on the remaining training set. For example, we would test "Are We Doomed?" against the remaining five Bast-authored documents as as well as the six Gleick-authored documents (obtained from Watts), looking for methods that would accurately identify "Are We Doomed?' as Bast-authored. We repeated this for the other eleven documents, looking for methods that would correctly attribute all documents. When analyzing Gleick documents, the test documents were chosen from the Watts-delivered documents while the training documents were the ones we had harvested, thus doubly avoiding any training-on-test problems.

19 Our most accurate method used a combination of the following settings, all available as part of the most recent version of JGAAP:

- Canonicizers: Normalize Whitespace, Punctuation Separator, Unify Case
- Event Set : POS (Part-of-Speech) 5-grams
- Analysis Method : Weka SMO

This method correctly attributed all twelve (12/12) training documents using leave-one-out validation and thus we consider this to be an appropriate method for this specific problem.

20 These canonicizers do the following : first, any "whitespace" characters such as spaces, tabs, and carriage returns are normalized so that every word has a single space (" ") separating it from the next word. Second, all punctuation characters are separated from other (nonblank) characers by a single whitespace, thus breaking up blocks of characters. Finally, all case variations are neutralized by converting all upper case characters to lower case.

21 The documents were tagged for part of speech ("POS") and taken in overlapping groups of five. For example the following sentence would be treated as shown

	The	quick	brown	fox	jumped	over	the	lazy	dog
POS	ART	ADJ	ADJ	NOUN	VERB	PREP	ART	ADJ	NOUN
1st 5gram	ART	ADJ	ADJ	NOUN	VERB				
2nd 5gram		ADJ	ADJ	NOUN	VERB	PREP			
3rd 5gram			ADJ	NOUN	VERB	PREP	ART		
4th 5gram				NOUN	VERB	PREP	ART	ADJ	

22 The WEKA SMO classifier[4] is a standard classifier distributed as part of the WEKA project that implements Platt's "sequential minimal optimization algorithm for training a support vector classifier." As such, it is a fairly standard state-of-the-art classifier used in many areas and on many problems. As implemented in WEKA, it has many different parameters and setting; we used only the standard (default) settings.

23 As before, our set of methods was able to accurately classify all twelve of our baseline documents, so we consider it sufficiently accurate to use for the disputed document.

Analysis

24 This task is challenging for several reasons, some technical and some linguistic.

25 First, the Heartland memo as published contains a great many quotations taken from other sources. As originally published, the memo contains approximately 717 words, but at least 266 of those words have been identified as belonging to phrases (or paraphrases of phrases) found elsewhere

[4] `http://weka.sourceforge.net/doc/weka/classifiers/functions/SMO.html`

in the stolen documents). [N.b. this identification was done by the Heartland Institute[5], who admit that these 266 words are "paraphrases [of] text appearing in one of the stolen documuments." As paraphrases, they may nor may not reflect the style of the original authors, and they also may or may not reflect the style of the alleged forger. For this reason, we analyzed both the full document as well as the 451-word redacted document with the controversial passages removed.

26 Second, even the full-length document is rather short for an accurate analysis. Most authorship attribution experts recommend larger samples if possible. (E.g., Eder recommends[6] 3500 words per sample, noting that results obtained from fewer than 3000 words "are simply disastrous.")

27 Thirdly, perhaps as a result of the previous factors, we have observed that Bast and Gleick appear to have extremely similar writing styles.

Results

28 Despite this difficulty, we were able to identify and calibrate an appropriate analysis method. Using this method, we analyzed both the complete Heartland memo and the selections from the Heartland memo that had been identified as not copied from other stolen documents. In both analyses, the JGAAP system identified the author as Peter Gleick.

29 In particular, the JGAAP system identified the author of the *complete* (unredacted) memo as Peter Gleick, despite the large amount of text that even Bast[7] admits is largely taken from genuine writings of the Heartland Institute. We justify this result by observing, first, that much of the quotation is actual paraphrase, and the amount of undisputed writing is still nearly 2/3 of the full memo.

Conclusions

30 In response to the question of who wrote the disputed Heartland strategy memo, it is difficult to deliver an answer with complete certainty. The writing styles are similar and the sample is extremely small, both of which act to reduce the accuracy of our analysis. Our procedure by assumption excluded every possible author but Bast and Gleick. Nevertheless, the analytic method that correctly and reliably identified twelve of twelve authors in calibration testing also selected Gleick as the author of the disputed document. Having examined these documents and their results, I therefore consider it more likely than not that Gleick is in fact the author/compiler of the document entitled "Confidential Memo: 2012 Heartland Climate Strategy," and further that the document does not represent a genuine strategy memo from the Heartland Institute.

/s/
Patrick Juola, Ph.D.
Director of Research

[5] `http://heartland.org/media-library/pdfs/FORGED%20HEARTLAND%20MEMO.pdf`

[6] "Does Size Matter? Authorship Attribution, Small Samples, Big Problem," *Proc. Digital Humanities 2010* `http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-744.html`

[7] `http://heartland.org/media-library/pdfs/FORGED%20HEARTLAND%20MEMO.pdf`