8-30-2021

Reaffirming the Superiority of Human Attorneys in legal Document Review and Examining the Limitations of Algorithmic Approaches to Discovery': Not So Fast

Maura R. Grossman
University of Waterloo

Gordon V. Cormack
University of Waterloo

Follow this and additional works at: https://scholarship.richmond.edu/jolt

'Reaffirming the Superiority of Human Attorneys in Legal Document Review and Examining the Limitations of Algorithmic Approaches to Discovery': Not So Fast

Maura R. Grossman, J.D., Ph.D.* & Gordon V. Cormack, Ph.D.**

Cite as: Maura R. Grossman & Gordon V. Cormack, *'Reaffirming the Superiority of Human Attorneys in Legal Document Review and Examining the Limitations of Algorithmic Approaches to Discovery': Not So Fast*, 27 RICH. J.L. & TECH., no. 4, (2021).

[1]   In *Humans Against the Machines: Reaffirming the Superiority of Human Attorneys in Legal Document Review and Examining the Limitations of Algorithmic Approaches to Discovery*,[1] Robert Keeling et al. "challeng[e] the prevailing wisdom around what predictive coding purports to do, and argu[e] that machines are simply not what they are promoted to be, especially in the discovery process."[2] We choose not to address their erroneous claims that "the results of prior research on predictive coding . . . reveal flaws," or their asserted "correct[ion] of misunderstandings," except to say that the study cited in footnote 143 of their paper[3] corroborates the prior research findings and explicitly addresses the impact of manual review as a component of technology-assisted review ("TAR"), showing that manual screening of the results of "predictive coding"[4] increases precision at the expense of recall.

[2]   Instead, we focus our analysis on a material error in *Humans Against the Machines* that calls into question its results and conclusions.

---

[1] Robert Keeling et al., *Humans Against the Machines: Reaffirming the Superiority of Human Attorneys in Legal Document Review and Examining the Limitations of Algorithmic Approaches to Discovery*, 26 RICH. J. L. & TECH., 2020 at 1.

[2] *Id.* at 2.

[3] Gordon V. Cormack & Maura R. Grossman, *Navigating Imprecision in Relevance Assessments on the Road to Total Recall: Roger and Me,* SIGIR '17: PROC.'S OF THE 40TH INT'L ACM SIGIR CONF. ON RES. AND DEV. IN INFO. RETRIEVAL 5, 5–14 (2017), https://dl.acm.org/doi/pdf/10.1145/3077136.3080812 [https://perma.cc/4XH6-TMN9].

[4] *See* Maura R. Grossman & Gordon V. Cormack, *The Grossman-Cormack Glossary of Technology-Assisted Review*, 7 FED. CTS. L. REV. 1, 26 (2013), https://www.fclr.org/fclr/articles/html/2010/grossman.pdf (defining predictive coding). [https://perma.cc/B5US-XZWF].

[3]   Keeling et al. claim that their experimental results demonstrate that manual review of the results of predictive coding yields an improvement in precision from 80.69% to 96.4%, at a "small cost" of 2.67% in recall.[5] These values are incorrect, and the data necessary to calculate them correctly have been omitted from the raw numbers provided in a confusion matrix in Table 6 of their paper.[6] Nonetheless, it is possible to infer an estimate of these values, and hence more accurate precision and recall estimates, from statistics provided elsewhere in their article (see Table 1 below). The corrected estimates reveal that *post-hoc human review of the results of predictive coding increases precision from **80.69% to 89.31%**, at the cost of reducing recall from **75% to 71.69%***.

[4]   This result supports the unremarkable conclusion that post-predictive-coding human review trades recall for precision. *It does not affirm the superiority of human review*, especially in light of Keeling et al.'s assertion that "between the two measures of precision and recall, **recall** is more important to attorneys, regulators, and courts because it measures whether predictive coding is actually identifying the responsive documents,"[7] and because one case study, conducted on a single document collection, without statistical testing or the reporting of confidence intervals, cannot possibly "reaffirm[] the superiority" of anything.

[5]   Over and above the serious quantitative error noted above, *Humans Against the Machines* compares the use of a trained automatic classifier (*i.e.*, predictive coding) to the use of a trained automatic classifier followed by human review. Both are forms of technology-assisted review; the former is **not** a form of TAR evaluated in "The TREC Data Study,"[8] and the latter is certainly not "exhaustive manual review," as examined in

---

[5] Keeling et al., *supra* note 1, at 49–50.

[6] *Id.* at 50.

[7] *Id.* at 15 (emphasis added).

[8] *See id*. at 8 n. 18 (citing Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, 17 RICH. J. L. & TECH. (2011)).

the TREC Data Study. Therefore, even if the quantitative results were to show the latter to be superior—which they do not—they would still not affirm the superiority of human review.

	First Level Review	Subject Matter Expert: Responsive	Subject Matter Expert: Non-Responsive
Responsive	1,384	1,236	148
Non-responsive	218	57	161
Total	1,602	1,293	309

Table 1: Inferred confusion matrix. Total Subject Matter Expert: Responsive is determined from the predictive model's stated precision of 80.69%,[9] from which we can infer that 80.69% of the 1,602-document sample (1,293 documents) are Subject Matter Expert: Responsive. The remaining cells may be filled in by subtraction. From the stated recall of 75%,[10] we may infer that 1,293 is 75% of the total number of responsive documents, which is necessarily 1,724. Recall is therefore 1,236/1,724≈**71.69%** (not 72.33% as reported), and precision is 1,236/1,384≈**89.1%** (not 96.4% as reported).

---

[9] *Id.* at 49.

[10] *Id.*