VerifyMe: A New Approach to Authorship Attribution in the Post-ChatGPT Era

O'Sullivan, Sean (School: Colaiste Chiarain)

OpenAI's ChatGPT has blurred the line between human and AI-written text, challenging academia's ability to verify authorship. Existing AI detection systems fail with paraphrased content or mimicked writing styles and misclassify human work. As we optimise for human-like writing, detecting generated text has become an increasingly distant target. I hypothesised that a stylometric approach focusing on individual authorship verification would be more effective. This research evolved through two iterations: an initial version for BTYSTE (Jan. ‘24) and an enhanced system for EUCYS (Sep. ‘24). VerifyMe analyses writing style rather than “detect AI”. The system captures an author's unique "stylistic fingerprint" by measuring 112 different aspects of their writing, creating a compact numerical representation of their style. When a new text is submitted, VerifyMe creates the same stylistic fingerprint and compares it to the given author's profile using a transformer encoder based Siamese neural network. The system was trained on a combined corpus of BAWE and filtered Project Gutenberg samples (2,034 authors total). When tested on unseen authors, VerifyMe achieved 85% accuracy on human-human verification from BAWE and 92% from Gutenberg. Against GPT-4o's mimicry attempts, while traditional detectors achieved only 0-2% accuracy, VerifyMe maintained 96% accuracy. The system outperformed commercial AI detectors by factors of 1.89× to 17.84×. Matthews Correlation Coefficient values reached 0.776 for BAWE and 0.829 for Gutenberg samples, with Area Under ROC Curve values of 0.956 and 0.972 respectively. These results demonstrate that focusing on verifying the author, not detecting AI, could provide a more reliable solution to authorhship verification in this post-ChatGPT era.

Awards Won: