HTW BERLIN/NIKOLAS FAHLBUSCH

Plagiarism detectors are a crutch, and a problem

Academics and editors need to stop pretending that software always catches recycled text and start reading more carefully, says **Debora Weber-Wulff.**

When it comes to plagiarism, many academics seem to believe in magic numbers. Last month, a company offering plagiarism-detection software announced that it would be acquired for US$1.7 billion later this year. It is one of several firms offering software systems that apply 'black box' algorithms to produce a score that purports to show how well one text matches others. Because these systems do find some cases of plagiarism, there is a misguided belief that they duly document all cases.

Horsefeathers, my grandmother would have said. I have been testing plagiarism-detection software for the past 15 years. The results are often hard to interpret, difficult to navigate, and sometimes just wrong. Many systems report false positives for common phrases, long names of institutions or even reference information. Software also produces false negatives. A system might fail to find plagiarism if the source of the plagiarized text has not been digitized, contains spelling errors or is otherwise not available to the software system. Many cases of plagiarism slip through undetected when material is translated or taken from multiple sources. Assessments depend on both the algorithms used and on the corpus of work available for comparison. For systems that check random samples, repeating the test of the document minutes later can produce different results. I have also seen different systems rank a text as completely or partially plagiarized, or plagiarism-free.

Yet the number that these systems produce — variously known as 'originality score', 'non-unique content' or 'PlagLevel' — is usually taken at face value. A second opinion is seldom sought, although there are dozens of systems available. Actually reading the reports produced by the software can reveal correctly quoted material, such as a properly referenced methods section, marked as plagiarism.

But time-pressed editors, professors and administrators often focus on that simple number when making decisions that are crucial to scholars and scholarship. If the software reports a low number, the person assessing the paper might skip over obvious indicators of plagiarism such as style shifts, misspellings, font changes or underlined words that suggest the text has been pasted from Wikipedia. And, yes, I've seen this in dozens of doctoral dissertations and scientific publications.

If the software reports a high number, editors or professors might unjustly consider a submission as unequivocal plagiarism. Universities formally define 'acceptable' levels of plagiarism, evaluated by the software, for various degree levels. Teachers want the software to flag up the 'bad' papers, so they don't have to read them. But students, afraid of having accidentally plagiarized, use the same systems to rewrite their work, swapping words with synonyms and rearranging sentences until the number looks good, to the detriment of readability.

Journal editors use the numbers as a crutch to quickly filter out papers that they can reject outright, or that they can publish without worry if reviewers give a thumbs up. Some journals and conferences even publish their threshold online.

Duplicated and plagiarized texts do harm: they distort scholars' true academic output and make the literature even harder to navigate. It cannot be tolerated, but these dodgy numbers are not the solution. I have been corresponding with journal editors about problematic publications for years. Duplicate publications are those that have essentially the same text (or even data) and share at least one author. In some cases, the title and the abstract are different, and authors have been added, removed or shuffled. Plagiarized articles have no authors in common.

Some of the editors I contact are quite surprised. They use plagiarism-detection software, so they expect to be in the clear. But duplication evades detection for many reasons. Potential sources, such as doctoral theses, might be stored in a repository or behind a paywall and are not available for comparison. Texts that have been cleverly (or even algorithmically) reworded will also fall below thresholds.

PLAGIARISM **CANNOT BE TOLERATED,** BUT THESE DODGY NUMBERS ARE **NOT THE SOLUTION.**

This year, abstracts submitted to the World Conference on Research Integrity were analysed by software, with a text-overlap threshold set at 30%. And, indeed, 38 out of 449 submitted abstracts registered above this level. After investigating, 15 were considered to be plagiarism and 23 contained text from the author's previously published research. Most of the abstracts were rejected; in some of the instances in which authors had recycled their own text, the abstracts were demoted to posters. This amount of plagiarism and duplication is shocking, especially for a conference on academic integrity; it is also probably an underestimate.

Software cannot determine plagiarism; it can only point to some cases of matching text. The systems can be useful for flagging up problems, but not for discriminating between originality and plagiarism. That decision must be taken by a person. The most important method for finding plagiarism is reading a text and studying the references for inconsistencies. A spot check with an Internet search engine, using three to five words from a paragraph or a particularly nice turn of phrase can uncover copyists. Searching for a reference that looks odd might turn up a source that mangled the reference in the same manner. Only if a text is somehow off, and online searching does not help, should software systems be consulted. In those cases, it's best to use two or three systems, and to read the reports, not take the numbers at face value.

Academic integrity is a social problem; due diligence cannot be left to unknown algorithms. Keeping science honest depends on scientists willing to work hard to protect the literature. ■

Debora Weber-Wulff *is professor of media and computing at the HTW Berlin – University of Applied Sciences.*
e-mail: weberwu@htw-berlin.de