Digital Studies /
Le champ numérique

Research

How to Cite: Land, Kaylin. 2020. "Predicting Author Gender Using Machine Learning Algorithms: Looking Beyond the Binary." *Digital Studies/Le champ numérique* 10(1): 8, pp. 1–12. DOI: https://doi.org/10.16995/dscn.362

Published: 12 October 2020

Peer Review:

This is a peer-reviewed article in *Digital Studies/Le champ numérique*, a journal published by the Open Library of Humanities.

Copyright:

Open Access:

Digital Preservation:

Digital Studies /
Le champ numérique

Land, Kaylin. 2020. "Predicting Author Gender Using Machine Learning Algorithms: Looking Beyond the Binary." *Digital Studies/Le champ numérique* 10(1): 8, pp. 1–12. DOI: https://doi.org/10.16995/dscn.362

RESEARCH

Predicting Author Gender Using Machine Learning Algorithms: Looking Beyond the Binary

Kaylin Land
McGill University, CA
kaylin.land@mail.mcgill.ca

This paper explores the relationship between digital humanities studies that utilize computer algorithms to identify author gender and feminist and queer literary theory. I argue that utilizing computer algorithms to sort literature into the categories "authored by a male" or "authored by a female" is too reductive in its treatment of gender as binary. However, I suggest computer algorithms could be utilized to explore the performative aspects of author gender and to ask larger questions about algorithmic criticism, the author as a subject, and the relationship between morphological and cultural properties of texts.

Keywords: machine learning algorithms; feminist theory; queer theory; author gender; gender performance; algorithmic criticism

Cet article explore la relation entre les études des humanités numériques, qui se servent d'algorithmes informatisés afin d'identifier le genre d'un auteur, et la théorie féministe et homosexuelle. Je soutiens que l'usage d'algorithmes informatisés pour catégoriser la littérature comme « écrit par un homme » ou « écrit par une femme » est trop réducteur par rapport à son traitement binaire de genre. Cependant, je suggère que les algorithmes informatisés peuvent être employés dans le but d'explorer les aspects performatifs du genre de l'auteur, ainsi que dans le but de soulever de plus grandes questions sur la critique d'algorithme, sur l'auteur en tant que sujet et sur le lien entre les caractéristiques de textes morphologiques et culturels.

Mots-clés: algorithmes d'apprentissage automatique; théorie féministe; théorie queer; genre d'auteur; expression de genre; critique algorithmique

1. Introduction

In the ongoing discussion of the role of algorithms in creating or confirming bias, digital humanities projects that rely on algorithms should naturally also come under consideration. This article examines studies that use machine learning algorithms to identify the gender of the author of fictional texts. While such studies provide an illuminating look at language usage, the binary approach of dividing texts by author gender not only confirms gender stereotypes but reinforces them by supporting the existence of a distinct "women's writing" style. The underlying methodology of many of these studies is part of a larger tradition of assuming writing by men as the default mode, thus, characterizing writing by women as automatically deviant. Instead of describing writing by women, these studies end up characterizing the category "women" through writing. This definition is reductive and problematic.

This article is not meant to be an exhaustive examination of machine learning algorithms and author gender studies. Nor is it fair to conflate all efforts to identify elements of female writing as utilizing the same methodology. Rather, the aim is to question the utility of continuing to perform tests of authorship based on gender using computer algorithms, and to suggest that such tests performed in isolation of theory serve to reinforce outdated understandings of gender in literature. The majority of these projects ask questions of gender through computer algorithms that largely ignore advances made in feminist and queer literary studies. The usage of algorithms to identify author gender assumes a binary approach that rests upon an outdated understanding of gender as either male or female. Furthermore, this methodology reinforces what Argamon et al. call "a least common denominator approach" that all too easily creates stereotypical lists of characteristics that define writing written by a woman (Argamon et al. 2009, 4).

Naturally, an algorithmic examination of literature written by men and women must operate within the realm of binaries. As Ben Verhoeven and Walter Daelemans (2018) recognize, "[u]sing non-binary gender is currently unfeasible for NLP [Natural Language Processing] research due to lack of data" (11). Argamon et al. (2009) confirm that a computer will find differences between texts because that is what computer algorithms are designed to do. However, I challenge the assumption that

this binary approach is beneficial. More importantly I question the interpretation of the results computer algorithm approaches produce. It is too reductive to argue that a list of terms used more frequently by women than by men proves that women have a unique writing style. Instead, further questioning of the results of algorithmic criticism informed by queer and feminist literary theory is needed to examine how language usage is exemplary (or not) of the performative nature of gender. I refer here to Judith Butler's definition of gender performativity as the result of "sustained social performances" within the context of compulsory heterosexuality (Jagger 2008, 27).

Stephen Ramsay (2011) discusses algorithmic criticism and the challenges of utilizing the scientific method in tandem with digital humanities projects. The author's work provides crucial clarification for this argument. Ramsay argues that algorithmic criticism, "criticism derived from algorithmic manipulation of text," exists only in its nascent form (2011, 2). For Ramsay, the role of text analysis is not merely to fact-check assumptions but rather to "assist the critic in the unfolding of interpretative possibilities" (9). Projects that utilize machine learning algorithms to identify author gender use the scientific method to present a problem (was a text written by a man or a woman?), identify a hypothesis, test results, present analysis, and form conclusions. The problem with this methodology is that it operates under the fallacy that there is, as Ramsay identifies, one answer to the problem (18). Texts are either written by men or written by women and once that answer has been determined, there is nowhere to go with the analysis. Worse still, the analysis often reinforces understandings of women's writing that confirm long-held assumptions of how women view and write about the world. Understandably, researchers are eager to identify the computational power of machine learning algorithms and find satisfaction in predicting author gender with high degrees of accuracy. However, deeper questions should be asked about the relationship between an author and the text they produce, as well as how an author enacts gender. In so doing, digital humanists can expand their analysis to produce multiple answers that lead to further inquiry.

2. Assumptions of algorithm author gender studies

By utilizing algorithms to identify whether or not a piece of writing is written by a man, or a woman several assumptions are made. Such an analysis is predicated on

the presumed existence of an identifiable genre called women's literature. Efforts to identify elements of women's literature using digital humanities tools in some ways mirror the methods early literary scholarship used to discuss writing by women. For example, George Henry Lewes (1852) identified feminine literary traits as sentiment and observation; William L. Courtney (1904, xiii) claimed the female author was "at once self-conscious and didactic"; Bernard Bergonzi (1965) saw literature written by women as contained within a narrow focus (quoted in Showalter 1995, 5). Lists of content features for female authors from one study certainly support the sentimental view of women's literature with words like *cute, love, boyfriend, mom* and *feel* given as characteristic of women's writing (Argamon et. al 2009, 121). In the 1970s, the term women's literature arose as part of feminist literary critics' efforts to identify and prove the existence of a specific women's literary tradition. Feminist critics, such as Elaine Showalter, Sandra Gilbert and Susan Gubar (2000) sought to characterize women's literature as an identifiable genre with its own poetics and classifiable traits. Such first and second wave feminists reclaimed a place for literature written by women in the traditional literary canon. However, this understanding of women's literature still operates within a binary framework that has been criticized for its essentialist understanding of female writing.

In using computer algorithms to isolate elements of a feminine style, scholars accept a binary understanding of gender and support the idea that women write differently than men. These assumptions mirror earlier attempts to identify women's literature. However, the critical difference between outdated views of women's literature and contemporary scholarship that relies on machine learning algorithms is that the latter claims to provide quantitative evidence of the existence of a women's literary style that, consciously or not, adheres to certain poetics. Such studies equate women's literature as a genre with work written by someone who identifies themselves (or is identified by scholars) as a woman in a binary system. Indeed, most of the studies refer equally to "women" and "female" writers with little explanation of how they are using these terms.

Lesbian and queer feminists have problematized this binary understanding of gender. In her well-known essay "Gender Trouble," Judith Butler (2018, 2490)

claims gender does not have a "natural" (biological) origin but is rather the effect of "institutions, practices and discourses with multiple and diffuse points of origin" on society. Butler was influenced by Monique Wittig's definition of "woman" and "man" as political and economic categories and her assertion that there is no natural "women" group. While Butler and Wittig disagreed about the materiality of gender and the ways in which heterosexual oppression should be opposed, both discussed gender performance (Jagger 2008). Butler understood gender as an "enforced cultural performance" that is itself performative (Ibid, 21). However, Butler viewed the performance of gender as a result of "the materiality of signs and signification" within a poststructuralist model that challenges the humanist understanding of the subject as a rational being capable of escaping both nature and culture to make objective statements (Ibid, 30). In other words, Butler argued against the idea that gender is performed in a theatrical sense and claimed rather that this performance be understood as a "speech act model based on a poststructuralist understanding of subjectivity" (Ibid, 21). The question of author agency and the ability of an individual author to create work independently of the material conditions under which they write is one area of inquiry that deserves attention in connection with algorithmic criticism of author gender. According to Butler, gender performance is not a conscious choice individual writers make but rather the result of a heteronormative, phallogocentric world order. Questions examining the extent to which individual authors choose to write "like a woman" and how gender is consciously or unconsciously performed could successfully be paired with machine learning algorithms to develop questions of author agency.

Indeed, one of the largest assumptions made in computer algorithm studies of author gender is that there is a direct relationship between the author and the text. This understanding of the author goes against poststructuralist understandings of the author as dead. As Nancy Miller (1988, 104) recognizes, if the author is dead, it does not matter *who* writes. Computer algorithms that identify author gender reassert the importance of the author. Such algorithms could be used to explore the relationship between the author and the text further by comparing the differences between texts written by female authors under male pseudonyms and works written

by those same authors under female names as, for example, J.K. Rowling and James Galbraith and James T. Tiptree and Alice Sheldon.

3. Interpretative schemata of author gender studies

Algorithmic studies of literature use digital tools to identify elements of an "*écriture féminine*" in literary works. Scholars utilize machine learning algorithms to process a large number of texts written by men and women in the attempt to identify word usage (content analysis) and syntax patterns (linguistic analysis) that distinguish "women's literature" from "men's literature," or simply literature. Such an approach assumes men's writing as the norm against which women's writing deviates. As Sara Mills argues in her work on the gendered sentence, these definitions of women's writing keep us "trapped within the notion of women's language being deviant, powerless and submissive, and male language being normal" (1995, 36). The digital humanities studies considered for this article often present their results within this paradigm. Mills identifies the importance of "interpretative schemata" in understanding the underlying biases present in studies of author gender in writing (1995, 39). That is to say, the interpretative elements chosen to identify women's literature are arguably more important than the simple assertion that women write differently than men. Critically, the majority of digital humanities studies favour content analysis over linguistic analysis. As such, they point to the topics women write about as indicative of a "women's literature" rather than identifying a feminine style through linguistic features.

For example, Argamon et al. (2009) claimed a machine learning algorithm could predict whether a given text was written by a man or a woman with 90% accuracy. Using content analysis, they identify, for example, that "male authors…use religious terminology rooted in the church, while female authors use secular language to discuss spirituality" (Argamon et al. 2009, 1). Burrows (2003) examined a corpus of English-language prose fiction and found that connector words, auxiliary verbs and the subjunctive could be used to identify female writing. However, Burrows also makes significant use of content analysis, claiming that "[f]emale authors, it seems, had more to say of females: both male and female authors had much to say of males" (2003, 344). Olsen's work (2005) on French language materials examines literary

texts from the 16th through the 20th centuries. Pronouns, possessives and content words that express emotional states marked the works Olsen examined as written by a woman. Olsen explains the higher percentage of first person singular and second person plural pronouns among women writers as indicative of "more personalized and interactional" texts than those written by men (2005, 154). More recently, Sean Weidman and James O'Sullivan (2018) used a corpus of English-language fiction texts separated into Victorian, Modernist and Contemporary periods to perform statistical analysis. The authors identify a unique female style that becomes more pronounced over time and argue that each era of "women's literature" has its own distinctive features that distinguish it from previous and subsequent periods. In particular, they find that [f]emales generally tend toward the language of place in the private, or micro-sense—'home', 'kitchen', 'church', 'hallway', and 'school'—whereas males throughout all periods tend toward greater spaces—'country', 'earth', 'city', 'town', and 'world' (2018, 384). The above studies all rely on content words to argue that women write differently than men, which is problematic as it creates an assumption that all women write about similar topics and ignores the historical and social realities that influenced who was able to write about what.

Some of the studies examined do use linguistic analysis. For example, Ben Verhoeven and Walter Daelemans (2018) used the Europarl corpus to examine texts across four different European languages for use in machine learning experiments to identify that discourse aspects of text (connector words such as *but*, *moreover*, and *so*) contribute to the prediction of the author's gender.

Similarly, Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni (2002) analyzed a corpus of written and spoken English ranging from the 12th century to the present and found that connective words such as *for, with, not* and *in* are characteristic of female writing. However, such analysis still relies on an interpretative scheme that equates writing written by women with women's writing as a universal group.

Performed Gender and Machine Learning Algorithms

In her work on femininity in early Canadian fiction, Misao Dean (1998) discusses the way in which female authors enact femininity as a practice in order to make themselves recognized within a larger male literary tradition. Dean asserts that efforts

to equate femininity in literature with some kind of "feminine essence" or universal gendered self that "chooses to act as a member of one or other gender" is false (1998, 7). Rather than positioning women's writing as either a conscious feminist act to reject the female voice or an essentialist essence that informs every woman's writing, Dean cites Judith Butler's understanding of femininity as "a norm which [women] occupy, reverse, [and] resignify to the extent that the norm fails to define us completely" (1998, 8). This occupation of femininity in writing must be considered when discussing the ways in which women have historically used language.

Dean's understanding of women's enactment of femininity provides a helpful point of comparison for analyzing author gender studies that rely on algorithms. These studies either tacitly or unintentionally argue that women's writing as a category exists because there are both content and linguistic elements common to writing written by women. Dean's and Butler's arguments can be reconciled with the results of author gender analysis. Dean (1998, 9) argues that "[g]ender cannot... be escaped or thrown off" because it is within the ideological system of gender that literature written by women is created. Stated differently, women may write differently not due to some essentialist "essence" or shared female experiences but rather because women have had to claim femininity in order to be recognized as individual voices. Author gender studies that find women write differently than men could be combined with Butler's understanding of the performativity of gender to examine the role of the author in creating works of fiction.

Weidman and O'Sullivan (2018) begin to unpack some of these issues in their recent article on word usage in author gender attribution. In examining works of fiction from three different historical periods they find that "stereotypical stylistic differences between men and women fiction writers" exist across all three periods examined (2018, 383). Their analysis showed that works written by women clustered with other works written by women (and vice versa for male authors) (2018, 379). They highlight female author's tendency to be more object-oriented than males, referring to positioning of the body with words such as *hair, fingers, skin, eyes, heart, face, cheeks, dress,* and *gown* as opposed to male authors' use of directional language

such as *east, west, south,* and *north* (2018, 383). As explored above, this kind of content analysis reinforces stereotypical views of what women can and should write about.

Weidman and O'Sullivan recognize, however, that relying on content words removes their context and that "just because we see that men and women tend to use words that indicate a particular topic is no barometer of how they are treating a said topic" (2018, 384). Furthermore, they recognize the potential that contemporary female authors make use of recognized feminine tropes with the intention of tearing down those tropes and "rehabilitating their relevance" (2018, 385). For example, the authors cite the opening scene of Margaret Atwood's 2009 novel *Year of the Flood* as presenting an "ironic not-so-picturesque rooftop sunrise" that contains content words associated with female authors such as *window, crying, birds, flower, garden, gate* and *balcony* (2018, 385).

In her work on irony and femininity, Lydia Rainford (2005) discusses the role of the term "ironic mode" as it was used by Judith Butler and Linda Hutcheon. Rainford argues such thinkers claimed irony as a tool for the feminist to "use her alterity to her advantage, by using it to negate the terms of the prevailing hierarchy" (2005, 4). In this understanding of the ironic, women use their secondary position in the gender hierarchy as a form of "negative freedom" in order to question the validity of the very existence of gender. While Rainford's work goes on to question the political and disruptive powers of irony in feminist writing, it provides an important counterpoint to author gender studies. By considering that women may use language differently than men in order to subvert their place in the gendered hierarchy, such studies are necessarily complicated. As Weidman and O'Sullivan suggest, it is critical to ask questions that refute the easy conclusion that women are inherently different than men and thus must write differently.

4. Conclusion

In closing, and perhaps most importantly, it is critical for scholars to consider that computer algorithmic approaches to author gender reinforce understandings of gender as a binary and can ignore individual author's own gender identification.

Marian Posner (2016, 32) claims digital humanities currently face a "meaningful opportunity" to address issues of race, gender, and structures of power. She argues that "most of the data and data models we have inherited deal with structures of power, like gender…with a crudeness that would never pass muster in a peer-reviewed humanities publication" (2016, 33). This point is particularly salient when considering the digital humanities projects based on algorithms for identifying gendered authorship. The majority of these projects treat gender as a binary category and leave no room for exploring gender not as a given but, to paraphrase Posner, as a construction that is actively created from time to time and place to place (2016, 35). Algorithmic criticism should be used to open more interpretative possibilities, not reinforce stereotypes. Author gender studies need to perform more nuanced, less reductive analysis or risk continuing to marginalize and devalue writing that is not authored by a man.

Competing Interest

The author has no competing interests to declare.

Author Contributors

Special Congress 2019 Issue editor: Barbara Bordalejo, University of Saskatchewan, Canada.

Section/Copy editor: Shahina Parvin, University of Lethbridge Journal Incubator.

References

Argamon, Shlomo, Charles Cooney, Russell Horton, Mark Olsen, Sterling Stein, and **Robert Voyer.** 2009. "Gender, Race, and Nationality in Black Drama, 1950–2006: Mining Differences in Language Use in Authors and Their Characters." *Digital Humanities Quarterly* 3(2). Accessed May 15, 2020. https://digitalhumanities.org/dhq/vol/3/2/000043/000043.html.

Argamon, Shlomo, Moshe Koppel, James W. Pennebaker, and **Jonathan Schler.** 2009. "Automatically Profiling the Author of an Anonymous Text." *Communications of the ACM* 52(2): 119–123. DOI: https://doi.org/10.1145/1461928.1461959

Bergonzi, Bernard. 1965. "Mixed Company." *The New York Review of Books*, June 3, 1965.

Burrows, John. 2003. "Textual Analysis." In *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 323–347. London: Blackwell Publishing Ltd. DOI: https://doi.org/10.4135/9780857020017

Butler, Judith. 2018. "Gender Trouble." In *The Norton Anthology of Theory and Criticism*, edited by V. B. Leitch, Third ed., 2375–2377. New York: W. W. Norton and Company.

Courtney, William Lewes. 1904. *The Lady Novelists*. London: Chapman and Hall.

Dean, Misao. 1998. "Introduction: Practising Femininity." In *Practising Femininity: Domestic Realism and the Performance of Gender in Early Canadian Fiction*, 3–15. Toronto: University of Toronto Press. DOI: https://doi.org/10.3138/9781442678712-002

Gilbert, Sandra M., and **Susan Gubar.** 2000. *The Madwoman in the Attic: The Woman Writer and the Nineteenth-Century Literary Imagination*. 2nd ed. New Haven: Yale University Press.

Jagger, Gill. 2008. "Gender as Performance and Performative." In *Sexual Politics, Social Change and the Power of the Performative*, 17–49. London: Routledge.

Koppel, Moshe, Shlomo Argamon, and **Anat Rachel Shimoni.** 2002. "Automatically Categorizing Written Texts by Author Gender." *Literary and Linguistic Computing* 17(4): 401–12. DOI: https://doi.org/10.1093/llc/17.4.401

Mill, John Stuart, and **Harriet Taylor Mill.** 1970. "The Subjection of Women." In *Essays on Sex Equality*, edited by Alice S. Rossi, 123–242. Chicago: University of Chicago Press.

Miller, Nancy K. 1988. "Introduction." In *Subject to Change: Reading Feminist Writing*, 1–21. New York: Columbia University Press. DOI: https://doi.org/10.7312/mill93000

Mills, Sara. 1995. "The Gendered Sentence." In *Feminist Stylistics*, 33–49. London: Routledge.

Olsen, Mark. 2005. "Écriture Féminine: Searching for an Indefinable Practice?" *Literary and Linguistic Computing* 20(Suppl Issue): 147–64. DOI: https://doi.org/10.1093/llc/fqi020

Posner, Miriam. 2016. "What's Next: The Radical, Unrealized Potential of Digital Humanities." In *Debates in the Digital Humanities 2016*, edited by Matthew K. Gold and Lauren F. Klein, 32–42. Minneapolis: University of Minnesota Press. DOI: https://doi.org/10.5749/j.ctt1cn6thb.6

Rainford, Lydia. 2005. *She Changes by Intrigue: Irony, Femininity and Feminism*. New York: Rodopi.

Ramsay, Stephen. 2011. "An Algorithmic Criticism." In *Reading Machines: Toward an Algorithmic Criticism*, 1–20. Illinois Scholarship Online. DOI: https://doi.org/10.5406/illinois/9780252036415.001.0001

Showalter, Elaine. 1999. *A Literature of Their Own*. Expanded E. Princeton, NJ: Princeton University Press.

Verhoeven, Ben, and **Walter Daelemans.** 2018. "Discourse Lexicon Induction for Multiple Languages and Its Use for Gender Profiling." *Digital Scholarship in the Humanities* 34(1): 208–220. DOI: https://doi.org/10.1093/llc/fqy025

Weidman, G., and **James O'Sullivan.** 2018. "The Limits of Distinctive Words: Re-Evaluating Literature's Gender Marker Debate." *Digital Scholarship in the Humanities* 33(2): 347–90. DOI: https://doi.org/10.1093/llc/fqx017

How to cite this article: Land, Kaylin. 2020. "Predicting Author Gender Using Machine Learning Algorithms: Looking Beyond the Binary." *Digital Studies/Le champ numérique* 10(1): 8, pp. 1–12. DOI: https://doi.org/10.16995/dscn.362

Submitted: 30 September 2019 **Accepted:** 24 March 2020 **Published:** 12 October 2020

Digital Studies/Le champ numérique is a peer-reviewed open access journal published by Open Library of Humanities. OPEN ACCESS