Chapter 3

Humanities and Social Science Reading through Machine Learning

Marisa Plumb
San Jose State University

Introduction

The purposes of computational literary studies have evolved and diversified a great deal over the last half century. Within this dynamic and often contentious space, a set of fundamental questions deserve our collective attention: does the computation and digitization of language recast the ways we read, value, and receive words? In what ways can research and scholarship on literature become a more meaningful part of the future development of computer systems? As the theory and practice of computational literary studies evolve, their potential to play a direct role in revising historical narratives and framing new research questions poses cross-disciplinary implications.

It's worthwhile to anchor these questions in the origin stories that today's digital humanists tell, from the work of Josephine Miles at Berkeley in the 1930s (Buurma and Heffernan 2018) to Roberto Busa's work in the 1940s to work that links Structuralism and Russian Formalism at the turn of the 19th century (Algee-Hewitt 2015) to today's systemized explorations of texts. The sciences and humanities have a shared history in their desire to solve the patterns and systems that make language functional and impactful, and there have long been linguistic and computational tools that help advance this work. What's more challenging to unravel and articulate from these origin stories are the mathematical concepts behind the tools that humanists wield. Ideally one would navigate this historical landscape when assessing the fitness of any given computational technique for addressing a specific humanities research question, but often researchers choose

tools because they are powerful and popular, without a robust understanding of the conceptual assumptions they embody, which are defined by the mathematical and statistical principles they are based on. This can make it difficult to generate reproducible results that contribute to a tool's methodological development.

This is related to a set of issues that drive debates among computationally-minded scholars, which regularly appear in digital humanities forums. In 2019, for instance, Nan Da issued a harsh critique of humanists' implementation of statistical methods in their research.[1] Her claim is that computational methods are not a good match for literary research, and she systematically shows how the results from several computational humanities are not only difficult to reproduce, but can be easily skewed with minor changes to how an algorithm is implemented. Although this debate about digital methods points to a necessary evolution in the field (in which researchers become more accountable to the computational laws that they are utilizing), her essay's broader mission is to question the appropriateness of using computational tools to investigate literary objects and ideas.

Refutations to this claim were swift and abundant (Critical Inquiry 2019), and highlight a number of concepts central to my concern here with future intersections of machine learning and literary research. Respondents such as Mark Algee-Hewitt pointed out that literary scholars employ computational statistical models in order to reveal something about texts that human readers could not. In doing so, literary scholars are at liberty to note where computation reaches its useful limit[2] and take up more traditional forms of literary analysis (Algee-Hewitt 2019). Katherine Bode explores the promise and pitfalls of this hybrid "close and distant reading" approach in her 2020 article on the intersection of topic modeling and bias. Imperfect as the hybrid method is, stressing the value of familiar interpretive methods remains important, politically and practically, when bringing computation into humanities departments.

This essay extends the argument that computational tools do more than turn big data into novel close reading opportunities. Machine learning, and word embedding algorithms in particular, may have a unique ability to shift this conversation into new territory, where scholars begin to ask how historical research can contribute more sophisticated approaches to treating words as data. With historically-minded approaches to dataset creation for machine learning, issues emerge that engender new theoretical frameworks for evaluating the ability of statistical models of information to reveal cultural and artistic dimensions of language. I will first contextualize what they do, and then show a few of the mathematical concepts that have driven their development.

Of the many available machine learning algorithms, word embedding algorithms have shown particular promise in capturing contextual meanings (of words or other units of textual data) more accurately than previous techniques in natural language processing. Word embeddings encompass a set of language modeling techniques where words or phrases from a large set of texts (i.e., "corpus") are analyzed through the use of a neural network architecture. For each vocabulary term in the corpus, the neural network algorithm uses the term's proximity to other words to assign it values that become a vector of real numbers — one high-dimensional vector is generated for each word. (The term "embedding" refers to the mathematics that turns a space with many

[1] Da's critique of statistical model usage in computational humanities work sparked a forum of responses in *Critical Inquiry*.

[2] This limit typically exists for a combination of three reasons: computer programs can only generate models based on the data we give them, a tool isn't fully understood and so not robustly explored, and many algorithms and tools are being used in experimental ways.

dimensions per word into a continuous vector space with a much lower dimension.)[3] They raise three critical issues to this essay: How do word embeddings reflect the contexts of words in order to capture their relative meanings? If word embeddings approximate word meanings, do they also reflect culture? How can literary history and cultural studies inform how scholars use them?

Word embeddings are powerful because they calculate semantic similarities between words based on their distributional properties in large samples of language data. As computational linguist Jussi Karlgren puts it:

> Language is a general-purpose representation of human knowledge, and models to process it vary in the degree they are bound to some task or some specific usage. Currently, the trend is to learn regularities and representations with as little explicit knowledge-based linguistic processing as possible, and recent advances in such general models for end-to-end learning to address linguistics tasks have been quite successful. Most of those approaches make little use of information beyond the occurrence or co-occurrence of words in the linguistic signal and take the single word to be the atomary unit.

This is notable because it highlights the power of word embeddings to assign values to words in order to represent their relative meanings, simply based on unstructured language data, without a system of linguistic rules or a labelling system. It also highlights the fact that a word embedding model's success is based on the parameters of the task it is designed to address. So while the accuracy and power of word vector algorithms might be recognizable in general-purpose applications that improve with larger training corpora (for instance Google News and Wikipedia), they can be equally powerful representation learning systems for specific historical research tasks that use different benchmarks for success. Humanists using these machine learning methods are learning to think differently about corpora size, corpora content, and the utility of a successfully-trained model for analysis and interpretation.

No matter what the application, the success of machine learning applications is predicated on creating good datasets. As a recent paper in IEEE Transactions on Knowledge and Data Engineering notes, "the majority of the time for running machine learning end-to-end is spent on preparing the data, which includes collecting, cleaning, analyzing, visualizing, and feature engineering" (Roh et al. 2019, 1). Acknowledging this helps contextualize machine learning algorithms for text analysis tasks in the humanities, but also highlights data curation challenges that can be taken up in new ways by humanists. This naturally raises questions about how machine learning algorithms like word embeddings are implemented for text analysis, and how they should be modified for historical research—they require different computational priorities and frameworks.

In parallel to the corpora considerations that computational humanities scholars ponder, there is an abundance of work, across disciplines such as cognitive science and psychology (Griffiths et al. 2007), that attempts to refine the problems and limits of using large collections of text for training embeddings. These large collections tend to reflect the biases that exist in society and history, and in turn, systems based on these datasets can make troubling inferences, now well documented as algorithmic bias.[4] Computer science researchers need to evaluate the social dimensions of their applications in diverse societies and find ways to fairly represent all populations.

[3] See Koehrsen 2018 for a fuller explanation of the process.

[4] As investigated, for instance, in Noble 2018.

Digital humanities practices can implicitly help address these issues. Literary studies, as it evolves towards multivocality and canon expansion, makes explicit a link between methods of literary analysis and digital practices that are deliberately inclusive, less-biased, and diachronic (rather than ahistorical). Emerging literary scholarship uses computational methods to question hegemonic practices in the history of the field, through the now-familiar practice of data curation (Poole 2013). But this work can also help combat algorithmic bias more broadly, and expand beyond corpus development into algorithmic design. As digital literary scholarship continues to deepen its exchanges with Sociology, History, and Information Science, stronger methodologies for using fair and representative data will become pervasive throughout these disciplines, as well as in commercial applications. Interdisciplinary methodologies are foundational to future computational literary research that can make sophisticated contributions to text analysis.

The Bengal Annual: A Case Study

Complex relationships between words cannot be fully assessed with one flat application of a powerful tool to a set of texts. But this does not mean that the usefulness of machine learning for literature is limited: rather, scholars can wield it to control how machines learn sets of relationships between concepts. Choosing which texts to include in a corpus is coupled to decisions about whether and how to label its contents, and how to tune the parameters of the algorithms. For the purposes of literary analysis, these should be embraced as interpretive, biased acts—ones that deepen understanding of commonly-employed computational methods—and folded into emerging methodologies. Because humanities scholars are not generating models to serve applications with thousands of end-users who primarily expect accuracy, they can exploit the fallacies of machine learning in order to improve how dataset management and feature engineering are conducted. Working with big data in order to generate models isn't valuable because it reveals history's "true" cultural patterns, but because it demonstrates how machines already circulate those "truths." A scholar's deep knowledge of the historical content and formalities of language can determine how corpora are compared, how we experiment with known biases, and how we move towards a future landscape of literary analysis that is inclusive of marginalized texts and the latest cultural theory.

Roopika Risam, for instance, advocates for both a theoretical and practice-based decolonization of the digital humanities, noting ways that postcolonial digital archives can intervene in knowledge production in society (2018, 79). Corpora created from periods of revolution, then, might reveal especially useful vector relationships and lead to better understanding of semantic changes during those times. Those word embeddings might be useful for teaching computers racialized language over timelines, so that machine learning applications do not only "read" history as a flat set of relationships, and inevitably reflect the worst of its biases.

To begin to unpack this process, I will present a case study on the 1830 *Bengal Annual* and a corpus of similarly-situated texts. Our team, made up of students in Katherine D. Harris's graduate seminar on decolonizing Romantic Literature at San Jose State University, asked: can we operationalize questions that arise from close readings of texts to turn problematic quantitative evaluations of words into more complex methods of interpretation? A computer cannot interpret complex cultural concepts, but it can be instructed to weigh time period, narrative perspective, and publication venue, much as a literary scholar would.

With the explosion of print culture in England in the first half of the nineteenth century, publishers began introducing new forms of serialized print materials, which included serialized

publications known as literary annuals (Harris 2015). These multi-author texts were commonly produced as high-quality volumes that could be purchased as gifts in the months leading up to the holiday season. As a genre, the annual included poetry, prose, and engravings, among other varieties of content, very often from well-known authors. Literary annuals represent a significant shift in the economics surrounding the production of print materials for mass consumption—for instance, contributors were typically paid. And annuals, though a luxury item, were more affordable than books sold before the mechanization of the printing press (Harris 2015, 1-29).

Literary annuals and other periodicals are interesting sites of literary study because they can be read as reinforcing or resisting the British Empire. London-based periodicals were eventually distributed to all of Britain's colonial holdings, including India (Harris 2019). As *The Bengal Annual* was written in India and contains a small representation of Indian authors, our project investigates it as a variation on British-centric reading materials of the time, which perhaps offered a provisional voice to a wider community of writers (though not without claims of superiority over the colonized territory it exploits).

Some of the contents invoke themes that are affiliated with major Romantic writers such as William Wordsworth and Samuel T. Coleridge, but editor D.L. Richardson included short stories and fiction, which were not held in the same regard as poetry. He also employed local native Indian engravers and writers. To explore the thesis that the concepts and genres typically associated with British Romantic Literature are represented differently in a text that was written and produced in a different space with a set of contributors who were not exclusively British natives, we experimented with word embeddings on semantic similarity tasks, comparing the annual to texts like *Lyrical Ballads*. Such a task is within the scope of traditional literary analysis, but my agenda was to probe the idea that we need large-scale representations of marginalized voices in order to show real differences from the ideas of the dominant race, class, and gender.[5]

The project team first used statistical tools to find out if the Annual's poetry, non-fiction, and fiction contained interesting relationships between vocabularies about body parts, social class, and gender. We gathered information about terms that might reveal how different parts of the body were referenced depending on sex. These differences were validated by traditional close-reading knowledge about British Romantic Literature and its historical contexts,[6] and signaled the need to read and analyze the Annual's passages about body parts, especially ones by writers of different genders and social backgrounds. These simple methods allowed us to take a streamlined approach to confirming that an author's perspective indeed altered his or her word choices and other aspects of their references to male vs. female bodies.

Collecting and mapping those references, however, was not enough to build a larger argument about how discourse on bodies might be different in non-canonical British Romantic Literature. Based on the potential for word embeddings to model semantic spaces for different corpora and compare the distribution of terms, the next step was to build a corpus of non-canonical texts of similar scope to a corpus of canonical works, so that models for each could be legitimately compared. This work, currently in progress, faces challenges that are becoming more familiar to digital historians: the digitization of rare texts, the review of digitization processes for accuracy, and the cleaning of data.

The primary challenge is to find the correct works to include: this requires historical exper-

[5] Such textual repositories are important outside of literature departments, too. We need data to represent all voices in training machines to represent any social arena.

[6] Some of these findings are illustrated in the project's Scalar site: `http://scalar.usc.edu/works/the-bengal-annual/bodies-in-the-annual`.

tise, but also raises the question of how to uncover unknown authors. Manu Chander's *Brown Romantics* calls for a global assessment of Romantic Literature's impact by "calling attention to its genuinely unacknowledged legislators" (Chander 2017, 11). But he contends that even the authors he was able to study were already individuals who aspired to assimilate with British culture and ideologies in some ways, and perhaps don't represent political resistance or views entirely antithetical to the British Empire.

Guided by Chander's questions about how to locate dissent in contexts of colonization, we documented instances in the text that highlight the dynamics of colonialism, race, and nationalism, and compared them to a set of statistical explorations of the text's vocabulary (particularly terms related to national identity, gender, and bodies). Chander's call for a more globally-comprehensive study of Romanticism speaks to the politics of corpora curation discussed above, but also suggests that corpus comparison can benefit from formal methodological guidelines. Puzzling out how to best combine traditional close readings with quantitative inquiries, and then map that work to a machine-learning research framework, revealed several shortcomings in methodological standardization. It also revealed several opportunities for rethinking the way algorithms could be implemented, by adopting and systematizing familiar comparative research practices. Ideas about such methodologies are emerging in many disciplines, which I highlight later in this essay.

Disciplinary directions for word vector research

The potential of word embedding techniques for projects such as our *Bengal Annual* analysis can be seen in the new computational research directions that have emerged in humanities research.[7] Vector-space representations are based on high-dimensional vectors[8] of real numbers.[9] Those vectors' values are assigned using a word's relationship to the words near it in a text, based on the likelihood that a word will appear in proximity to other words it is told to "look" at. For example, this visualization demonstrates an embedding space for a historical corpus (1640-1699) using the values assigned to word vectors (figure 3.1).

In a visualized space (with reduced dimensions) such as the one in figure 3.1, distances among vectors can be assessed, for example, to articulate the forty words most similar to *wit*. This particular model (trained using the word2vec algorithm), published in the 2019 *Debates in the Digital Humanities*,[10] allowed the authors to visualize the term *wit* with synonyms on the left side, and terms related to argumentation on the right, such as *indeed, argues,* and *consequently.* This initial exploration prompted Gavin and his co-authors to look at a vector space model for a single author (John Dryden), in order to both validate the model against their subject matter expertise and explore the model's results. Although word vectors are often employed for machine translation tasks[11] or to project analogistic relationships between concepts,[12] they can also be used to

[7] See Kirschenbaum 2007 and Argamon and Olsen 2009.

[8] A word vector may have hundreds or even thousands of dimensions.

[9] Word embedding algorithms are modelled on the linguistic concept that context is a primary way that word meanings are produced. Their usefulness is dependent on the breadth and domain-relevance of the corpus they are trained on, meaning that a corpus of medical research vs. a corpus of 1980s television guides vs. a corpus of family law proceedings would generate models that show different relationships between words like "family," "health," "heart," etc.

[10] See Goldstone 2019.

[11] Software used to translate text or speech from one language to a target language. Machine translation is a subfield of computational linguistics that can now allow for domain-based (i.e. specialized subject matter) customizations of translations, making translated word choices more context-specific..

[12] Although word embeddings aren't explicitly trained to learn analogies, the vectors exhibit seemingly linear behavior (such as "woman is to queen as man is to king"), which approximately describe a parallelogram. This phenomenon is

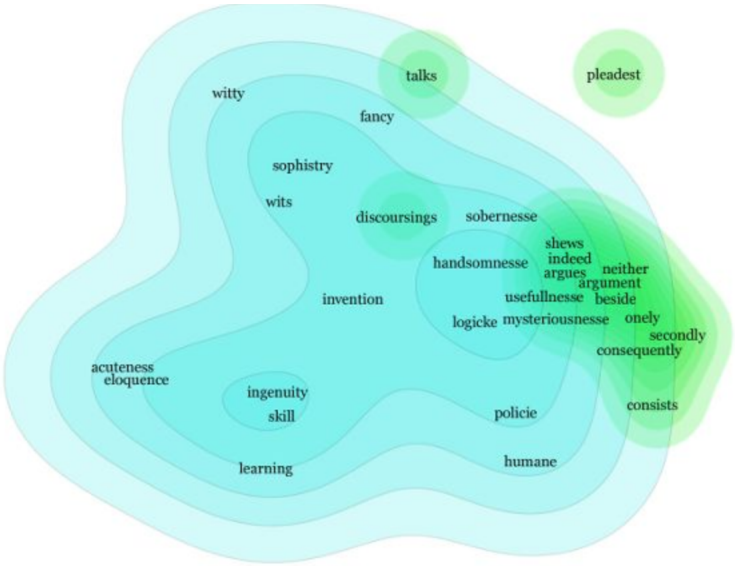

Figure 3.1: A visualized space with reduced dimensions of a neighborhood around *wit* (Gavin et al. 2019, Figure 21.2).

question concepts that are traditionally associated with particular literary periods and evaluate those associations with new kinds of evidence.

What this type of study suggests is that we can look at cultural concepts like *wit* in new ways. These results can also facilitate a comparison of historical models of *wit* to contemporary ones—to show how its meaning may have shifted, using its changing relationship to other words as evidence. This is a growing area of research in the social sciences, computational linguistics, and other disciplines (Kutuzov et al. 2019) In a survey paper on current work in diachronic word embeddings and semantic shifts, Kutuzov et al. note that the surge of interest points to its importance for natural language processing, but that it currently lacks "cohesion, common terminology and shared practices."

Some of this cohesion might be generated by putting the usefulness of word vectors in context of the history of information retrieval and the history of distributed representation. Word embeddings emerged in the 1960s, with data modeled as a matrix, and a user's query of a database represented as a vector. Simple vector operations could be used to locate relevant data or documents. Gerald Salton is generally credited as one of the first to do this, based on the idea that he could represent a document as a vector of keywords and use measures like cosine similarity and dimensionality reduction to compare documents.[13] Since the 1990s, vector space models have been

explored in Allen and Hospedales 2019.

[13] Algorithms like word2vec take as input the linguistic context of words in a given corpus of text, and output an N dimensional space of those words—each word is represented as a vector of dimension N in that Euclidean space. Word vectors with thousands of values are transformed to lower-dimensional spaces in which the directionality of two vectors can be measured using cosine similarity—words that exist in similar contexts would be expected to have a similar cosine measurement and map to like clusters in the distributed space.

used in distributional semantics. In a paper on the history of vector space models, which examines the trajectory of Gerald Salton's work, David Dubin notes that these mathematical models can be defined as "a consistent mathematical structure designed to correspond to some physical, biological, social, psychological, or conceptual entity" (2004). In the case of word vectors, word context and colocations give us quantifiable information about a word's meaning.

But research in cognitive science has long questioned the property of linguistic similarity in spatial representations because they don't align with important aspects of human semantic processing (Tversky 1977). Tversky shows, for example, that people's interpretation of semantic similarity does not always obey the triangle inequality, i.e., the words w1 and w3 are not necessarily similar when both pairs of (w1, w2) and (w2, w3) are similar. While "asteroid" is very similar to "belt" and "belt" is very similar to "buckle", "asteroid" and "buckle" are not similar (Griffiths et al. 2007). One reason this violation arises is because a word is represented as a single vector even when it has multiple meanings. This has led to research that attempts new methods to capture different senses of words in embedding applications. In a paper surveying techniques for differentiating words at the "sense" level, Jose Camacho-Collados and Mohammad Taher Pilehvar show that these efforts fall in two camps: "*Unsupervised* models directly learn word senses from text corpora, while *knowledge-based* techniques exploit the sense inventories of lexical resources as their main source for representing meanings" (2018, 744).

The first method, an unsupervised model, induces different meanings of a word — it is trained to analyze and represent each word sense based on statistical knowledge derived from the contexts within a corpus. The second method for disambiguation relies on information contained in other databases or sources. WordNet, for instance, associates multiple words with concepts, providing a sense inventory for terms. It is made up of synsets, which represent unique concepts that can be expressed through nouns, verbs, adjectives or adverbs. The synset of a concept such as "a business where patrons can purchase coffee and use WiFi" might be "cafe, coffeeshop, internet cafe" etc. Camacho-Collados and Pilehvar review different ways to process word embedding results using WordNet and similar resources, which essentially provide synonyms that share a common meaning.

There exists a relationship between work that addresses word disambiguation and work that addresses the biases that word vector algorithms produce. Just as researchers can modify general word embedding models to capture a word's multiple meanings, they can also modify them according to a word's usage over time. These evolving methods begin to account for the social, historical, and psychological dimensions of language. If one can show that applying word embedding algorithms to diachronic corpora or corpora of different domains produces different biases, this would suggest that nuanced shifts in vocabulary and word usage can be used to impact data curation practices that seek to isolate and remove historical bias from other word embedding models.

Biases, one might say, persist despite contextual changes. Or, one might say that the shortcomings of word embeddings don't account for changes in bias that are present in context. This is where the domain expertise of literary scholars also becomes essential. Historians' domain expertise and natural interest in comparative corpora (from different time periods or containing different types of documents) situates their ability to curate datasets that tend to both data ethics and computational innovation. Such work could have impact beyond historical research, and result in data-level corrections to biases that emerge in more general-purpose embedding applications. This could be more effective and reproducible than correcting them superficially (Gonen and Goldberg 2019). For instance, if novel cultural biases can be traced to an origin period, texts

from that period could constitute a sub-corpus. Embedding models specific to that corpus might be subtracted from the vectors generated from a broader dataset.

Examining a methodology's history is an essential way in which scholars can strengthen the validity of computationally-driven research and its integration into literary departments—this type of scholarship reconstitutes literary insights after the risky move of flattening literary texts with the rigor of machines. But as Lauren Klein (2019) and others reveal, scholars have begun to apply interpretation and imagination in both the computational and the "close reading" aspects of their research. This reinforces that computational shifts in the study of literature are more than just the adoption of useful tools for the sake of locating a novel pattern in data. An increasingly important branch of digital literary research demonstrates the efficacy of engaging the interdisciplinary complexity of computational tools in relation to the complexity of literary analysis.

New ideas for close readings and analysis can serve as windows into defining secondary *computational* research questions that emerge from an initial statistical exploration. As in the work reviewed by Camacho-Collados Pilehvar, outside knowledge of word senses can be used for post-processing word embeddings that address theoretical issues. Implementing this type of process for humanities research, one might begin with the question: can I generate word vector models that attend to both author gender and word context if I train them in innovative ways? Does this require a corpus of male authors and one of female authors? Or would this be better accomplished with an outside lexical source that has already associated word senses with genders?

Multi-disciplinary scholars are experimenting with a variety of methods to use word vector algorithms to track semantic complexities, and humanities researchers need an awareness of the technical innovations across a range of these disciplines because they are in a position to bring important domain knowledge to these efforts. Ideally, the questions that unite these disciplinary efforts might be: how do we make word contexts and distributional semantics more useful for both historians, who need reproducible results that lead to new interpretation, and technologists, who need historical interpretation to play a larger role in language generalization? Modeling language histories depends on how deeply humanists can understand word embedding models, so that they can augment their inherent shortcomings. Cross-disciplinary collaborations help scholars return to fundamental issues that arise when we treat words as data, and help bring more cohesive methodological standards to language modeling.

New directions in cross-disciplinary machine learning frameworks

Literary scholars set up computational inquiries with attention to cultural complexity, and seek out instances of language that convey historical context. So while they aren't likely to lead the charge in correcting fundamental shortcomings of language representation algorithms, they can increasingly impact social assessments of those algorithms, provide methodologies for those algorithms to locate anomalies in language usage, and assess whether those algorithms embody socially just practices (D'Ignazio and Klein 2020). Some literary scholars also critique the non-neutral ideologies that are in place in both computing and the humanities (Rhody 2017, 660). These efforts not only make the field of literary studies (and its history) more relevant to a digitally and computationally-driven future, but also help literary scholars create meaningful intersections between their computational tools and theoretical training. That training includes frameworks

for reading and analysis that computers cannot yet perform, but should aspire to—from close reading, Semiotic Criticism, and Formalism to Post-structuralism, Cultural Studies, and Feminist Theory. The varied systems literary scholars have developed for thinking about signs, words, and symbols should not be seen as irreconcilable with computational tools for text analysis. Instead, they should become the foundation for new methodologies that tackle the shortcomings of machine learning algorithms and project future directions for text analysis.

Linguists and scientists interested in natural language processing have often looked to the humanities for methods that assign rules to the production of meaning. Such methods exist within the history of literary criticism, some of which are being newly explored as concepts for language modeling algorithms. For instance, data curation takes inspiration from cultural studies, which empowers literary scholars to correct for bias and underrepresentation in literature by expanding the canon. Subsequent literary findings from that research need not only be literary ones: they have the potential to serve as models for best practices for computational tools and datasets more broadly. While the rift between society's most progressive ideas and its technological advancement is not unique to the rise of machine learning, practical opportunities exist to repair the rift with a blend of literary criticism and computational skills, and there are many recent examples[14] of the growing importance of combining rich technical explanations, interdisciplinary theories, and original computational work in corpus linguistics and beyond. A desire to wield social and computational concerns simultaneously is evident also in recent work in Linguistics,[15] Sociology,[16] and History.[17]

Studies in computational Sociology by Lauren Nelson, Austin C. Kozlowski, Matt Taddy, James A. Evans, Peter McMahan, and Kenneth Benoit contain important parallels for machine learning-driven text analysis. Nelson, for instance, calls for a new three-step methodology to computational sociology, one that "combines expert human knowledge and hermeneutic skills with the processing power and pattern recognition of computers, producing a more methodologically rigorous but interpretive approach to content analysis" (2020, 1). She describes a framework that can aid in reproducibility, which was noted as a problem by Da. Kozlowski, Taddy, and Evans, who study relationships between attention and knowledge, in a September 2019 paper on the "geometry of culture" use a vector space model to analyze a century of books. They show "that the markers of class continuously shifted amidst the economic transformations of the twentieth century, yet the basic cultural dimensions of class remained remarkably stable. The notable exception is education, which became tightly linked to affluence independent of its association with cultivated taste" (1). This implies that disciplinary expertise can be used to isolate sub-corpora for use in secondary word embedding research problems. Resulting word similarity findings could aid in both validating the initial research finding and defining domain-specific datasets that are reusable for future research.

The idea of using humanities methodologies to inform model architectures for machine learn-

[14] See Whitt 2018 for a state-of-the-art overview of the intersecting fields of corpus linguistics, historical linguistics, and genre-based studies of language usage.

[15] A special issue in the journal *Language* from the Linguistic Society of America published responses to a call to reconcile the unproductive rift between generative linguistics and neural network models. Christopher Potts's response (2019) advocates an imperative integration between deep learning and traditional linguistic semantics.

[16] Sociologist Laura K. Nelson (2020) calls for a three-step methodological framework called computational grounded theory.

[17] Another special issue, this one from *Isis*, a journal from the History of Science Society, suggests that "the history of knowledge can act as a bridge between the world of the humanities, with its tradition of close reading and detailed understanding of individual cases, and the world of big data and computational analysis" (Laubichler, Maienschein, and Renn 2019, 502).

ing is part of a wider history of computational scientists drawing inspiration from other fields to make AI systems better. Designing humanities research with novel word embedding models stands to widen the territory where machine learning engineers look for conceptual concepts to inspire strategies for improving the performance of artificial language understanding. Many computer scientists are investigating the figurative (Gagliano et al. 2019) and the metaphorical (Mao et al. 2018) in language. As machines get better at reading and interpreting texts, literary studies and theories will become more applicable to how those machines are programmed to look at multiple layers and dimensions of language. Ted Underwood, Andrew Piper, Katherine Bode, James Dobson, and others make connections between computational literary research and social dimensions of the history of vector space model research. Since vector models are based on the 1950s linguistic notion of similarity (Firth 1957), researchers working to show superior algorithmic performance focus on different aspects of why similarity is important than do researchers seeking cultural insights within their data. But Underwood points out that a word vector can also be seen as a way to quantitatively account for more aspects of meaning (2019). Already, cross-disciplinary scholarship draws on computational linguistics,[18] information science,[19] and semantic linguistics, and the imperative to understand concepts from all of these fields is growing. As better methods are developed for using word embeddings to better understand texts from different domains and time periods, more sophisticated tools and paradigms emerge that echo the complexity of traditional literary and historical interpretation.

Systematic data curation, combined with word embedding algorithms, represent a new interpretive system for literary scholars. The potential of machine learning methods for text analysis goes beyond historical literary text analysis, and the methods for literary text analysis using machine learning also go beyond literature departments. The corpora they model and the way they frame their research questions reframe the potential to use systems like word vectors to understand aspects of historical language and could have broader ramifications on how other applications model word meanings. Because such literary research generates novel frameworks for using machine learning to represent language, it's imperative to explore the question: Are there ways that humanities methodologies and research goals can exert greater influence in the computational sciences, make the history of literary studies more relevant in the evolution of machine learning techniques, and better serve our shared social values?

References

Algee-Hewitt, Mark. 2015. "The Order of Poetry: Information, Aesthetics and Jakobson's Theory of Literary Communication." Presented at the Russian Formalism & the Digital Humanities Conference, April 13, Stanford University, Palo Alto, CA. https://digitalhumanities.stanford.edu/russian-formalism-digital-humanities.

Algee-Hewitt, Mark. 2019. "Criticism, Augmented." *In the Moment* (blog). April 1, 2019. https://critinq.wordpress.com/2019/04/01/computational-literary-studies-participant-forum-responses/.

Allen, Carl, and Timothy Hospedales. 2019. "Analogies Explained: Towards Understanding Word Embeddings." In *International Conference on Machine Learning*, 223–31. PMLR. http://proceedings.mlr.press/v97/allen19a.html.

[18] Linguistics scholars are also adopting computational models to make progress with theories related to semantic similarity. For instance, see Potts 2019.

[19] See Lin 1998, for example.

Argamon, Shlomo and Mark Olsen. 2009. "Words, Patterns and Documents: Experiments in Machine Learning and Text Analysis." *Digital Humanities Quarterly* 3 (2). http://www.digitalhumanities.org/dhq/vol/3/2/000041/000041.html.

Bode, Katherine. 2020. "Why You Can't Model Away Bias." *Modern Language Quarterly* 81 (1): 95–124. https://doi.org/10.1215/00267929-7933102.

Buurma, Rachel Sagner, and Laura Heffernan. 2018. "Search and Replace: Josephine Miles and the Origins of Distant Reading." *Modernism / Modernity Print*+ 3, Cycle 1 (April). https://modernismmodernity.org/forums/posts/search-and-replace.

Camacho-Collados, Jose and Mohammad Taher Pilehvar. 2018. "From Word To Sense Embeddings: A Survey on Vector Representations of Meaning." *Journal of Artificial Intelligence Research* 63 (December): 743–88. https://doi.org/10.1613/jair.1.11259.

Chander, Manu Samriti. 2017. *Brown Romantics: Poetry and Nationalism in the Global Nineteenth Century*. Lewisburg, PA: Bucknell University Press.

Critical Inquiry. 2019. "Computational Literary Studies: A Critical Inquiry Online Forum." *In the Moment* (blog). March 31, 2019. https://critinq.wordpress.com/2019/03/31/computational-literary-studies-a-critical-inquiry-online-forum/.

Da, Nan Z. 2019. "The Computational Case against Computational Literary Studies." *Critical Inquiry* 45 (3): 601–39. https://doi.org/10.1086/702594.

D'Ignazio, Catherine, and Lauren Klein. 2020. *Data Feminism*. Cambridge: MIT Press.

Douglas, Samantha, Dan Dirilo, Taylor-Dawn Francis, Keith Giles, and Marisa Plumb. n.d. "The Bengal Annual: A Digital Exploration of Non-Canonical British Romantic Literature." https://scalar.usc.edu/works/the-bengal-annual/index.

Dubin, David. 2004. "The Most Influential Paper Gerard Salton Never Wrote." *Library Trends* 52 (4): 748–64. https://www.ideals.illinois.edu/bitstream/handle/2142/1697/Dubin748764.pdf?sequence=2.

Firth, J.R. 1957. "A Synopsis of Linguistic Theory." In *Studies in Linguistic Analysis*, 1–32. Oxford: Blackwell.

Gagliano, Andrea, Emily Paul, Kyle Booten, and Marti A. Hearst. 2019. "Intersecting Word Vectors to Take Figurative Language to New Heights." In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, 20-31. San Diego, California: Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-0203.

Gavin, Michael, Collin Jennings, Lauren Kersey, and Brad Pasanek. 2019. "Spaces of Meaning: Conceptual History, Vector Semantics, and Close Reading." In *Debates in the Digital Humanities 2019*, edited by Matthew K. Gold and Lauren F. Klein, 243–267. Minneapolis: University of Minnesota Press.

Goldstone, Andrew. 2019. "Teaching Quantitative Methods: What Makes It Hard (in Literary Studies)." In *Debates in the Digital Humanities 2019*, edited by Matthew K. Gold and Lauren F. Klein. Minneapolis: University of Minnesota Press. https://dhdebates.gc.cuny.edu/read/untitled-f2acf72c-a469-49d8-be35-67f9ac1e3a60/section/620caf9f-08a8-485e-a496-51400296ebcd#ch19.

Gonen, Hila and Yoav Goldberg. 2019. "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them." *ArXiv*:1903.03862, September. https://arxiv.org/abs/1903.03862.

Griffiths, Thomas L., Mark Steyvers, and Joshua B. Tenenbaum. 2007. "Topics in Semantic Representation." *Psychological Review* 114 (2): 211–44. https://doi.org/10.1037/

0033-295X.114.2.211.

Harris, Katherine D. 2015. *Forget Me Not: The Rise of the British Literary Annual, 1823 – 1835*. Athens: Ohio University Press.

Harris, Katherine D. 2019. "*The Bengal Annual* and #bigger6." *Keats-Shelley Journal* 68: 117–18. https://muse.jhu.edu/article/771132.

Kirschenbaum, Matthew. 2007. "The Remaking of Reading: Data Mining and the Digital Humanities." Presented at the National Science Foundation Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation, Baltimore, MD, October 11. https://www.csee.umbc.edu/~hillol/NGDM07/abstracts/talks/MKirschenbaum.pdf.

Klein, Lauren F. 2019. "What the New Computational Rigor Should Be." *In the Moment* (blog). April 1, 2019. https://critinq.wordpress.com/2019/04/01/computational-literary-studies-participant-forum-responses-5/.

Koehrsen, Will. 2018. "Neural Network Embeddings Explained." *Towards Data Science*, October 2, 2018. https://towardsdatascience.com/neural-network-embeddings-explained-4d028e6f0526.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings." *American Sociological Review* 84 (5): 905–949. https://doi.org/10.1177/0003122419877135.

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. "Diachronic word embeddings and semantic shifts: a survey." In *Proceedings of the 27th International Conference on Computational Linguistics*, 1384-1397. Santa Fe, New Mexico: Association for Computational Linguistics. https://www.aclweb.org/anthology/C18-1117.

Laubichler, Manfred D., Jane Maienschein, and Jürgen Renn. 2019. "Computational History of Knowledge: Challenges and Opportunities." *Isis* 110 (3): 502-512.

Lin, Dekang. 1998. "An Information-Theoretic Definition of Similarity." In *Proceedings of the Fifteenth International Conference on Machine Learning*, 296–304. San Francisco, California: Morgan Kaufmann Publishers Inc.

Mao, Rui, Chenghua Lin, and Frank Guerin. 2018. "Word Embedding and WordNet Based Metaphor Identification and Interpretation." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 1222–31. Melbourne, Australia: Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1113.

Nelson, Laura K. 2020. "Computational Grounded Theory: A Methodological Framework." *Sociological Methods & Research* 49 (1): 3–42. https://doi.org/10.1177/0049124117729703.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

Poole, Alex H. 2013. "Now Is the Future Now? The Urgency of Digital Curation in the Digital Humanities." *Digital Humanities Quarterly* 7 (2). http://www.digitalhumanities.org/dhq/vol/7/2/000163/000163.html.

Potts, Christopher. 2019. "A Case for Deep Learning in Semantics: Response to Pater." *Language* 95 (1): e115–24. https://doi.org/10.1353/lan.2019.0019.

Rhody, Lisa. 2017. "Beyond Darwinian Distance: Situating Distant Reading in a Feminist *Ut Pictura Poesis* Tradition." *PMLA* 132 (3): 659-667.

Risam, Roopika. 2018. "Decolonizing the Digital Humanities in Theory and Practice." In *The*

Routledge Companion to Media Studies and Digital Humanities, edited by Jentery Sayers, 78–86. New York: Routledge.

Roh, Yuji, Geon Heo, and Steven Euijong Whang. 2019. "A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective." *IEEE Transactions on Knowledge and Data Engineering* Early Access: 1–20. https://doi.org/10.1109/TKDE.2019.2946162.

Tversky, Amos. "Features of Similarity." *Psychological Review* 84 (4): 327–52. https://doi.org/10.1037/0033-295X.84.4.327.

Underwood, Ted. 2019. *Distant Horizons: Digital Evidence and Literary Change*. Chicago: University of Chicago Press.

Whitt, Richard J., ed. 2018. *Diachronic Corpora, Genre, and Language Change*. John Benjamins Publishing Company.