Post45
Issue 7: Post45 x Journal of Cultural Analytics

The Goodreads “Classics”: A Computational Study of Readers, Amazon, and Crowdsourced Amateur Criticism

Melanie Walsh and Maria Antoniak

04.21.21

What is a classic? This is "not a new question," as T.S. Eliot acknowledged more than seventy-five years ago.[1] More than simply "not new," this question now feels decidedly old, hashed out, and even passé. Perhaps most glaringly outdated is the word "classic." Literary scholars don't often use the term any-

more, at least not as a serious label for literature "of the highest rank or importance." [2] In 1991, John Guillory declared that the term classic was "all but retired." [3] The label, according to Guillory, signified not only a "relatively uncritical regard for the great works of Western literature" but the "precritical era of criticism itself." [4] Instead, in academic conversations, the ardent language of the "classics" has largely been displaced by the more critical vocabulary of the "canon," which frames literary significance more carefully as a product of cultural selection.

Yet the question — what is a classic? — remains surprisingly powerful in the twenty-first century because the classics are alive and thriving on the internet, in the marketplace, and among readers, even if not in universities or among academics. Contemporary readers not only use the term "classic," they use it a lot and often have strong feelings about it. "**To all the snobs on goodreads*:* I read a CLASSIC," one reader heralded in a Goodreads review of George Orwell's *1984*, brandishing the term six more times before concluding: "***I read a CLASSIC. wohooooooo.***" Another Goodreads user, Bren, rated Vladimir Nabokov's *Lolita* two out of five stars and explained: "I get that this a classic and book snobs who read this will sig[h] in indignation but I do not care…Lolita is a classic but it just is not MY classic." [5] Yet another disgruntled reader panned J.D. Salinger's *The Catcher in the Rye* and ventured that "anybody who tells me this is a classic or that I 'don't understand it' can kiss the whitest, fattest part of my ass." Why do contemporary readers use the term "classic" so frequently and so passionately? The classics, it turns out, are at the heart of some of the most significant developments in contemporary literary history, including the rise of digital literary culture, online amateur criticism, and internet corporations with bookish investments like Goodreads and Amazon. There are few places more instructive for understanding these developments and the contemporary classics than Goodreads, the focal point of this essay.

With more than 120 million members, Goodreads is the largest social networking site for readers on the internet and a subsidiary of Amazon, one of the wealthiest and most influential corporations in

the world. On Goodreads, internet users can categorize any book as a "classic" and publish their own responses to it — gushing praise, mean takedowns, critical analyses, snarky parodies, personal narratives, and more. Among thousands of literary categories on Goodreads, "classics" is one of the top ten most popular and includes some of the most rated and reviewed books across the entire site. The frequently tagged "classic" *To Kill a Mockingbird* (1969), for example, has been rated by Goodreads users more than 4 million times, a level of engagement only surpassed by three other books: J.K. Rowling's *Harry Potter and the Sorcerer's Stone* (1997), Suzanne Collins's *The Hunger Games* (2008), and *Stephenie Meyer's Twilight* (2005).

This flood of Goodreads classics content represents an excitingly large archive of amateur criticism and reader responses, an opportunity for scholars to hear nonacademic readers' voices in ways that were difficult if not impossible before the internet. For example, understanding how readers felt about classics in the Victorian period is difficult because there is little first-hand evidence from Victorian readers, as Richard D. Altick once explained: "The great majority of the boys and girls and men and women into whose hands fell copies of cheap classic reprints did not leave any printed record of their pleasure. Only occasionally did the mute, inglorious common reader take pen in hand." [6] Far from this "mute, inglorious" Victorian common reader, the twenty-first-century readers of Goodreads regularly publish records of their readerly pleasure and displeasure on the internet. Beyond providing a rich archive of reader responses, Goodreads also raises questions about whether its social network might enable a democratization of the classics. The classics, after all, have historically been defined by those in power and excluded "the interests and accomplishments of minorities, popular and demotic culture, or non-European civilizations," as Ankhi Mukherjee describes. [7] To what extent are millions of Goodreads users from around the globe now remedying or replicating such historical exclusions?

Though Goodreads data is a boon for literary criticism and a potentially transformative development for literary culture, it is also a

boon for corporations. Amazon's looming shadow over Goodreads data helps bring into focus a more financially minded definition of a classic, perhaps best summarized by poet and literary critic Mark Van Doren. A classic, Van Doren said, is simply "a book that remains in print." [8] For the twenty-first century, we might update Van Doren's definition and say that a classic is simply a book that continues to make money in whatever form it takes, whether as a print book, audiobook, e-book, screen adaptation, or as the subject of millions of online book reviews. In fact, it is clear, based on the Goodreads reviews that we analyze in this essay, that industries such as film, television, publishing, e-commerce, and tech not only profit from the classics but profit from each other in a circular loop, benefiting from the reinforcement of works as classics in other mediums and domains. Considered together, this *classics industry*, as we call it — a formulation inspired by and indebted to Simone Murray's "adaptation industry" as well as Pierre Bourdieu's theories of cultural production [9] — proves to be one of the strongest influences on Goodreads users' perception of the classics.

The tensions between democratic potential and corporate exploitation that we observe in the Goodreads classics are characteristic of many social networks and Web 2.0 platforms, which fundamentally rely on user-created content. These dynamics have been studied extensively by scholars of fandom, new media, and digital culture, among other fields. [10] Yet how social network dynamics and the internet economy are reshaping literary culture, in particular, is still a relatively new conversation, led by critics such as Murray, Aarthi Vadde, Lisa Nakamura, and Mark McGurl. [11] By examining Goodreads reviews in this essay, we hope to contribute to this emerging conversation. We also hope to add a quantitative, data-driven perspective to the discussion by curating a collection of more than 120,000 Goodreads reviews and by using computational methods to study some of the most salient trends. We believe that digital literary culture is an area that especially rewards the convergence of digital humanities methods and contemporary literary criticism. The massive number of Goodreads ratings and reviews is part of what makes the platform worthy of study and financially lucrative, but also what makes Goodreads difficult to understand in

broad strokes. Digital humanities and cultural analytics scholars have demonstrated, however, that computational methods can help us better understand cultural phenomena at scale. By employing these methods on Goodreads data in particular, we build on previous Goodreads-related DH research by Karen Bourrier and Mike Thelwall, J.D. Porter, Alexander Manshel, and Laura McGrath, James F. English, Scott Enderle, and Rahul Dhakecha, Allison Hegel, and Andrew Piper and Richard So, among others. [12]

Scale is not our only motivation for using computational methods. The contemporary book world, including but not limited to Goodreads and Amazon, is increasingly governed by algorithms and data, which presents a number of challenges for contemporary literary scholars. "Clearly the leviathan that is Amazon exerts immense influence on the global book trade," as Simone Murray contends, "but how are scholars to document, much less critique, algorithmic culture's self-reinforcing effects on cultural selection if denied access to the workings of the algorithm's engine-room?" [13] To provide one answer to Murray's urgent question, we believe that computational methods can supply a way of documenting, understanding, and critiquing algorithmic culture and its effects. By collecting and analyzing Goodreads data with computational methods, we are able to see, for example, that Goodreads only publicly displays a small fraction of its data. We are also able to detail some of the specific social effects of the platform's default sorting algorithm, which prioritizes the most liked and most commented on reviews. This digital infrastructure produces a feedback loop among Goodreads reviews, in which reviews that receive attention continue to receive more attention ad nauseam.

This feedback loop is a fitting metaphor for the consecration of the classics on Goodreads more broadly. Though Goodreads users technically define the classics for themselves, their definitions are clearly shaped by a reciprocal system of reinforcing influences — old institutions like high schools, universities, and publishing houses as well as new institutions like Amazon. The result is a reader-produced vision of the classics that is surprisingly less diverse, in terms of authors' race and ethnicity, than those represented by U.S.

literature syllabi, though more diverse in terms of genre, including more genre fiction, young adult fiction, and adapted fiction. Though Goodreads users seem strongly influenced by traditional institutions and the capitalist marketplace, they nevertheless demonstrate enormous creativity in finding ways to make this critical conversation their own — parodying and panning different literary styles, reliving and reimagining high school English classes, pushing back against the perceived arbiters of literary authority, and publicly changing their minds.

To close this introduction, we foreground our own approach to Goodreads data, since the exploitation of user data is a central subject of this essay. We have chosen not to publicly share our dataset of Goodreads reviews, though we have shared the code that we used to collect data from the Goodreads website, and we have obtained explicit permission from each Goodreads user who is directly quoted in this essay. [14] We believe that ethical approaches to user data will continue to be one of the most important conversations for digital humanities and cultural analytics research, and we expand on our choices further in the Appendix.

The Classics "Shelf": Genre, Hashtag, Advertising Keyword

This essay understands Goodreads users to be readers as well as "amateur critics," a framing that we draw from Aarthi Vadde, Melanie Micir, and Saikat Majumdar, among others. [15] As Vadde explains, "The ease and ubiquity of digital publishing have enabled the 'mass amateurization' of the critical, creative, and communicative arts, allowing amateurs to bypass the gatekeeping practices of specific institutions…and to perform acts of photography, journalism, or authorship without necessarily identifying with a specialized guild or benefitting from its resources." [16] The digital platform of Goodreads similarly allows amateurs to perform acts of literary criticism, to publish their own analyses and judgements of literature, without formal training and without access to traditional publishing venues. The three main ways that Goodreads users act as amateur critics are by rating books between one and five stars, by

reviewing books in 15,000 characters or less, and by "shelving" books into categories. We begin with an extended discussion of Goodreads "shelves" because they are one of the primary ways that users define the classics and that Amazon profits from the classics.

The first telling clue about these shelves is that the Goodreads website fluidly refers to them as "shelves," "genres," and "tags." This slippery relationship points to a significant evolution of genre among readers and amateur critics in the twenty-first century: genre is being subsumed and reshaped by *tagging*. Tagging is a common system for classifying and organizing content on the internet, in which users tag digital content with their own free-form descriptions, keywords, and metadata (think hashtags on Twitter). The shelf system on Goodreads is a *social* or *collaborative* tagging system because users can apply different tags to the same content, essentially crowdsourcing book categorization. Prior computational social science and natural language processing research has explored how these collaborative tagging systems produce "folk taxonomies" or *folksonomies*, classification systems built by communities from the ground up. [17] Literary genre, in the hands of internet users equipped with tagging systems, has similarly blossomed into a grassroots taxonomy that incorporates conventional genres but also splinters into new genres, microgenres, publishing industry categories, reception metadata, hashtags, and more. [18] For example, a Goodreads user named Candace tagged Margaret Atwood's *The Handmaid's Tale* (1985) as "classics" and six other distinct categories (**figure 1**): "wtf-did-i-just-read," "kindle-unlimited," "dark-themes," "favorites," "listened-to-audio-version," and "age-difference." [19] Fellow Goodreads users tagged *The Handmaid's Tale* as "science-fiction," "fantasy-sf," "man-booker-shortlist-longlist," "tv-series," "re-read," and "feminism," among many other tags. As these examples demonstrate, Goodreads users mold conventional genre to better represent their tastes, values, and cataloguing needs. Allison Hegel argues that Goodreads shelves may also help readers "articulate their identities to others and connect with larger communities." [20] According to Jeremy Rosen, most literary critics today understand genre "not as a rigid category that texts 'belong to' or a set of rules that one must abide by, but as a

flexible set of techniques that can be adapted *according to the needs of its users*." [21] While Rosen's "users" are mostly authors, who mold genre to create their own literary works, the ambiguous term suggests that others can use genre, too, including readers and amateur critics. Thus, "classics" emerges as an important contemporary genre for readers in addition to a label of literary value and publishing category.

Figure 1: This screenshot displays a Goodreads review of Margaret Atwood's The Handmaid's Tale written by a user named Candace. The red annotations highlight the "Shelves" section of the review, where Goodreads users can categorize books with their own personal shelves/genres/tags. Notably, Candace has shelved The Handmaid's Tale in "classics." The number of likes and comments that this particular review received is also underscored because it is the basis by which Goodreads sorts reviews by default, which we discuss in more detail in the section "The Goodreads Algorithmic Echo Chamber."

Though "classics" is just one Goodreads shelf among thousands, it is one of the most important and foundational. In the website's earliest days, the company used "classics" as their first anchoring example to introduce and explain the shelving system: "You can create your own personal bookshelves. From classics to canadabooks, to childrenslit and geek, you can create any category that suits your personal taste." [22] Ten years later, the classics remained Goodreads go-to example: "Shelf names range from classics and coffee-table-books to childrens-lit and sci-fi — you can create any category that suits your personal taste." [23] Because "classics" sup-

posedly represents the oldest and most traditional literary category, it serves as an effective foil for any unconventional literary category a Goodreads user might dream up, and it also invites a mass of readers and amateur critics to participate in a seemingly elite conversation. The classics thus make the entire shelf system legible and appealing.

Shelves are also financially lucrative for Goodreads and Amazon, the classics shelf particularly so. Each time a Goodreads user shelves a book in their personal library, that user simultaneously shelves the same book in the platform's massive library of more than two billion books. [24] "Goodreads turns the reader into a worker," as Lisa Nakamura observes, and through shelves, the company crowdsources the enormous work of organizing two billion books to the masses. [25] By shelving books, Goodreads users also (more unsettlingly) organize themselves into coherent audience categories that can be effectively targeted by advertisers. The same shelves that Goodreads users invent are sold as advertising target keywords, as Goodreads' informational material for advertisers shows in **figure 2**. These shelves represent not only books but also *people*: the Goodreads users who form communities around genres and subject areas, who read and discuss the books shelved into these categories. Browsing through the list of advertising "target values" reveals that some of these shelves are fascinatingly niche like *space-opera*, *mermaids*, and *reformation-history*. Yet other target values like *mental-illness* and *abuse* seem more serious and sensitive, raising the concerning possibility that vulnerable groups might be targeted by advertisers. Goodreads flags the "classics" as one of their top 10 most "prominent" genres for advertisers, putting it in the same company as "contemporary," "historical-fiction," "fantasy," "fiction," "manga," "mystery," "romance," "non-fiction," and "young-adult." Looking at the top 10 most rated books across the entire Goodreads website offers one clear picture of this prominence: five of the top 10 are classics (Table 1).

goodreads

Genre List for Advertisers

This is the master set of genres currently available to be used as target values for your ads on Goodreads.
Please work with your Account Manager to ensure that your campaign has a sufficient set of targets to achieve desired reach.
Contact your account manager, or advertising@goodreads.com with any questions.

Top 10 most prominent genres are in **bold**.

16th-century	ancient-history	british-literature	comic-books
17th-century	angels	buddhism	comics
18th-century	animal-fiction	bulgarian	comics-manga
19th-century	animals	business	coming-of-age
1st-grade	anime	canada	comix
20th-century	anthology	canadian-literature	communication
21st-century	anthropology	canon	computer-science
2nd-grade	apocalyptic	category-romance	computers
abuse	archaeology	catholic	**contemporary**
academia	architecture	cats	contemporary-romance
academic	art	chapter-books	cookbooks
action	art-history	chemistry	cooking
activism	arthurian	chick-lit	cozy-mystery
adaptations	asia	childrens	crafts
adolescence	asian-literature	childrens-classics	crime
adult	astronomy	china	criticism
adult-fiction	atheism	christian	culinary
adventure	australia	christian-fiction	cult-classics
africa	autobiography	christian-non-fiction	cultural
african-american	banned-books	christian-romance	cultural-studies
aliens	baseball	christmas	cyberpunk
alternate-history	batman	church	czech
alternate-universe	bd	church-history	danish
amazon	bdsm	cities	dark
american	biblical	civil-war	dark-fantasy
american-classics	biography	class	dc-comics
american-fiction	biography-memoir	classic-literature	death
american-history	biology	classical-studies	demons
american-novels	birds	**classics**	design
americana	bl	clean-romance	detective
amish	bookclub	collections	diary
amish-historical-romance-fiction	books-about-books	college	disability
ancient	brain	comedy	disabled-communities

Figure 2: The first page of a four-page document titled "Genre List for Advertisers," described as "the master set of genres currently available to be used as target values for your ads on Goodreads." The "classics" is bolded as one of the top 10 most prominent genres near the bottom-center of the page. This "Genre List for Advertisers" document can be found under "Target Advertising" on the "Advertise with Us" section of the Goodreads website.

Title	Author	Ratings	Publication Year	Top "Classics"
Harry Potter and the Sorcerer's Stone	J.K. Rowling	7.2m	1997	No
The Hunger Games	Suzanne Collins	6.5m	2008	No
Twilight	Stephenie Meyer	5m	2005	No
To Kill a Mockingbird*	**Harper Lee**	**4.6m**	**1960**	**Yes**
The Great Gatsby *	**F. Scott Fitzgerald**	**3.8m**	**1925**	**Yes**
The Fault in Our Stars	John Green	3.6m	2012	No
1984*	**George Orwell**	**3.2m**	**1949**	**Yes**
Pride and Prejudice*	**Jane Austen**	**3m**	**1813**	**Yes**
Divergent	Veronica Roth	2.9m	2011	No
The Hobbit*	**J.R.R. Tolkien**	**2.9m**	**1937**	**Yes**

Top 10 Most Rated Books on Goodreads (March 2021)

To fully grasp the significance of Goodreads users' shelving labor, it is helpful to compare Goodreads to Netflix, the world's largest video streaming service. Like Goodreads, Netflix has a massive microgenre system for its video content, featuring hyper-specific genres like **"Deep Sea Horror Movies"** and **"Romantic Dramas Based on Classic Literature."** To assemble these 70,000+ "altgenres," Netflix "paid people to watch films and tag them with all kinds of metadata," as Alexis Madrigal reported in 2014. [26] "When these tags are combined with millions of users' viewing habits, they become Netflix's competitive advantage," Madrigal argues. "The data can't tell them *how* to make a TV show, but it can tell them *what* they should be making." [27] By tagging books with their own extremely detailed metadata, Goodreads' 90 million users perform a similar service for Goodreads and Amazon, but they do it for free. [28]

The Classics According to Goodreads Users

When Goodreads users shelve books, they supposedly classify books on their own terms without direct intervention from the academy, the publishing industry, or Amazon. Technically, any of the two billion books in the Goodreads library could become a classic in users' hands. Yet when we collate the books that Goodreads users have collaboratively consecrated as classics, we find the strong influence of school curricula and what we call the *classics industry*, the interrelated network of businesses that generate and profit from the classics — such as publishing, film and television, and internet corporations like Goodreads itself. To identify this list of Goodreads classics, we first selected the top 100 literary works tagged as a classic the greatest number of times by Goodreads users throughout the site's history (2006-2019). We then added the top 100 literary works that were tagged as a classic and most read by Goodreads users in the first week of September 2019 (the week when we collected our data). The homepage for popular shelves like the classics prominently features books that were "Most Read This Week," displaying them even above the most tagged books in the genre. We decided to include this second group of books because they are conspicuously promoted by Goodreads and provide a slightly different perspective on the Goodreads classics — not only what users have tagged as classics but also which classics users actually seem to be reading. Many of the 100 most read classics overlap with the 100 most shelved classics, and in total the list includes 144 unique titles. [29]

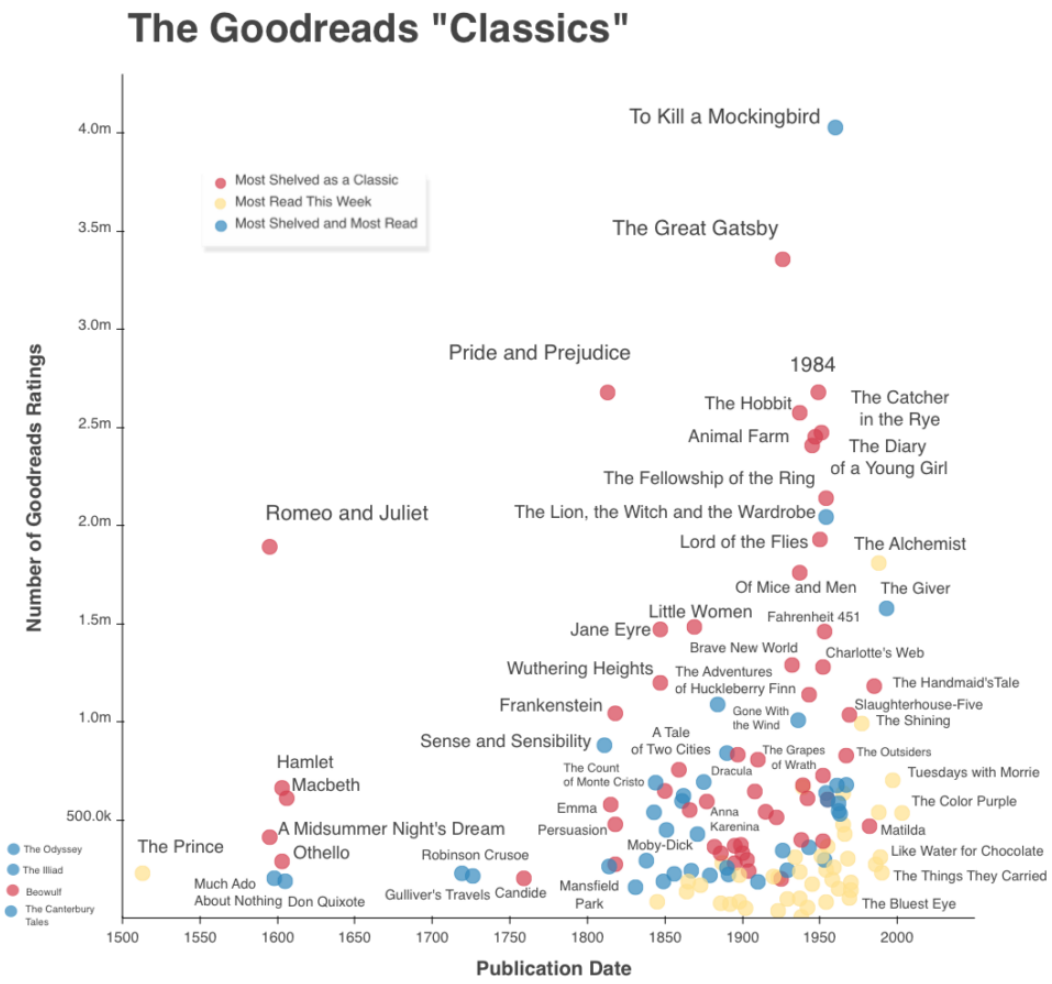

Figure 3: This figure displays the distribution of 144 Goodreads classics by publication date and total number of Goodreads ratings received as of September 2019. The same plot can be explored in more detail as an interactive data visualization. Blue points represent literary works most shelved as a classic throughout the history of Goodreads. Yellow points represent classics most read in September 2019. Red points represent the overlapping titles most shelved as a classic and most read in September 2019.

The makeup of the Goodreads classics (figure 3) confirms what Lisa Nakamura observes about the platform: "Reader tastes reflect the traditional literary canon more closely than one might expect." [30] These Goodreads classics include canonical mainstays such as Homer's *The Odyssey* (~700 BC) and Shakespeare's *Hamlet* (1603), Nathaniel Hawthorne's *The Scarlet Letter* (1850) and Charlotte Brontë's *Jane Eyre* (1847), F. Scott Fitzgerald's *The Great Gatsby* (1926) and Virginia Woolf's *Mrs. Dalloway* (1925), J.D.

Salinger's *The Catcher in the Rye* (1951) and Vladimir Nabokov's *Lolita* (1955). Their publication dates noticeably skew toward the late-nineteenth and twentieth centuries. More than a third were published after 1945. While the dominant form of literature in these Goodreads classics is fiction, there is also a small amount of poetry, drama, and non-fiction, such as Kahlil Gibran's book of prose poetry, *The Prophet* (1923), Oscar Wilde's play *The Importance of Being Earnest* (1895), and Anne Frank's *The Diary of a Young Girl* (1947).

Many texts labeled as classics by Goodreads users seem to overlap with English literature curricula from U.S. grade schools, high schools, and colleges. Though the Goodreads platform has an increasingly global audience — with notable emerging userbases in India and the UK — most of its users have historically hailed from the U.S. and still make up an estimated 40% of sitewide traffic. [31] For two rough estimates of how much the Goodreads classics overlap with school syllabi, we consulted a recommended reading list from the Advanced Placement (AP) English program — a common literature curricula in U.S. high schools — as well as a compilation of college-level English literature syllabi from the Open Syllabus Project, which draws on syllabi from many countries but predominantly from the U.S. [32] More than a third of the Goodreads classics authors are specifically recommended by the AP English program, and about half rank within the top 200 most assigned college-level authors.

Yet the Goodreads classics depart from these school-sanctioned lists in two particularly striking ways. First, the Goodreads classics are considerably less diverse in terms of the race and ethnicity of their authors. Race is extremely complex and difficult to reduce to data, especially because racial categories differ across different societies. However, if we acknowledge this reduction and use racial categories from the U.S. to reflect the perspective of the majority of Goodreads users, [33] then almost 94% of the Goodreads classics authors are white, which makes them whiter than both the AP recommended authors (70%) and the Open Syllabus authors (86%). The Goodreads classics include works by six Black writers: Alexandre Dumas, Frederick Douglass, Chinua Achebe, Zora Neale

Hurston, Toni Morrison, and Alice Walker. There are also works by Laura Esquivel and Gabriel García Márquez, who would likely be read as Latinx from the perspective of U.S. racial logics, and Kahlil Gibran, who would likely be read as Middle Eastern or North African from the same perspective. There are no works by Asian, Asian American, or Indigenous authors. Further, there are few texts written by authors beyond North America and Europe, with notable exceptions including Achebe's *Things Fall Apart*, Gibran's *The Prophet*, and Márquez's *One Hundred Years of Solitude*. This lack of racial and geographic diversity in the Goodreads classics is not entirely surprising when one considers the user demographics of Goodreads. Beyond the platform's U.S.-centrism, the racial demographics of its user base skew overwhelmingly white—at least according to Quantcast, one of the ad industry's leaders for measuring online traffic and user demographics. As of June 2020, according to Quantcast, Goodreads users were 77% Caucasian, 9% Hispanic, 7% African American, 6% Asian, and 1% other. [34] It is crucial to note, however, that Quantcast uses statistical modeling techniques to predict demographic characteristics such as gender, age, ethnicity, and income, and, as sociologist Ruha Benjamin argues, companies that "create racial-ethnic data to be sold to others" deserve intense scrutiny. [35] Quantcast data is nevertheless used by many companies, including Goodreads, which makes it important to consider. [36] With these purported user demographics in mind, the predominantly white Western makeup of these reader-produced classics is not shocking but it is nevertheless startling, and it cautions any outsized faith in crowdsourced technologies as necessarily or predictably democratizing tools.

The second significant departure from school curricula is the presence of genre fiction, young adult fiction, and film- and television-adapted fiction. For example, among the Goodreads classics, we find science fiction and mystery in Frank Herbert's *Dune* (1965) and Agatha Christie's *And Then There Were None* (1939); children's novels in E.B. White's *Charlotte's Web* (1952) and Frances Hodgson Burnett's *The Secret Garden* (1911); and the source material for iconic film adaptations in L. Frank Baum's *The Wonderful Wizard of Oz* (1900), Truman Capote's *Breakfast at Tiffany's* (1958), and

Stephen King's *The Shining* (1977). Because most of this genre fiction entered our Goodreads classics list from the "Most Read This Week" list, it perhaps points to readers' actual, or at least more typical, reading habits and tastes. [37]

The Goodreads Algorithmic Echo Chamber

Goodreads users have not, on the whole, disrupted or remade the traditional canon of classics in any clearly radical ways via their crowdsourced shelving practices — save perhaps for the incorporation of genre fiction. From the perspective of race and ethnicity, Goodreads users in fact seem to be reinforcing an even whiter and less diverse canon of classics than one would find in a typical high school or college classroom today. By analyzing Goodreads reviews in addition to shelf classifications, we hoped to better understand the forces and influences shaping this perception of the "classics" — who and what "is responsible for maintaining them in their preeminent position," as Jane Tompkins once put it. [38] When we turned to collect and analyze Goodreads users' reviews, we recognized one clear answer: Goodreads and Amazon. In this section, we briefly discuss the challenges that we faced while collecting Goodreads reviews, which we hope will be informative for others who wish to work with Goodreads reviews in the future. But more importantly these challenges reveal key insights about Goodreads/Amazon's proprietary algorithms and management of user data.

The first key insight is that Goodreads purposely conceals and obfuscates its data from the public. The company does not provide programmatic (API) access to the full text of its reviews, as some websites and social media platforms do. To collect reviews, we thus needed to use a technique called "web scraping," where one extracts data from the web, specifically from the part of a web page that users can see, as opposed to retrieving it from an internal source. [39] The Goodreads web interface makes it difficult to scrape large amounts of review data, however. It's not just difficult for researchers to collect Goodreads reviews. It's difficult for *anyone* to interact with Goodreads reviews. Though more than 90 million re-

views have been published on Goodreads in the site's history, one can only view 300 reviews for any given book in any given sort setting, a restriction that was implemented in 2016. Previously, Goodreads users could read through thousands of reviews for any given book. Because there are a handful of ways to sort Goodreads reviews (e.g., by publication date or by language), it is technically possible to read through 300 reviews in each of these sort settings. But even when accounting for all possible sort setting permutations, the number of visible and accessible Goodreads reviews is still only a tiny fraction of total Goodreads reviews. This throttling has been a source of frustration both for Goodreads users and for researchers.

	Oldest	Newest	Default	All
Number of Reviews	42,311 reviews	42,657 reviews	42,884 reviews	127,855 reviews
Mean Length of Reviews	54.6 words	91.8 words	261.2 words	136.3 words
Number of Unique Users	24,163 users	33,486 users	17,362 users	69,342 users
Mean Number of Reviews per User	1.75 reviews/user	1.27 reviews/user	2.47 reviews/user	1.84 reviews/user

Summary Statistics for Goodreads Classics Reviews

Working within these constraints, we collected approximately 900 unique reviews for each classic book—300 default sorted reviews, 300 newest reviews, and 300 oldest reviews—for a total of 127,855 Goodreads reviews. We collected these reviews regardless of whether the user explicitly shelved the book as a "classic" or not. We also explicitly filtered for English language reviews. Despite this filtering, a small number of non-English and multi-language reviews are included in the dataset, and they show up as outliers in some of our later results. Compared to the archives of most readership and reception studies, this dataset is large and presents exciting possibilities for studying reception at scale. But it is important to note that this dataset is not large or random enough to be a statistically representative sample of the "true" distribution of classics reviews on Goodreads. We believe our results provide valuable insight into Goodreads and the classics nonetheless.

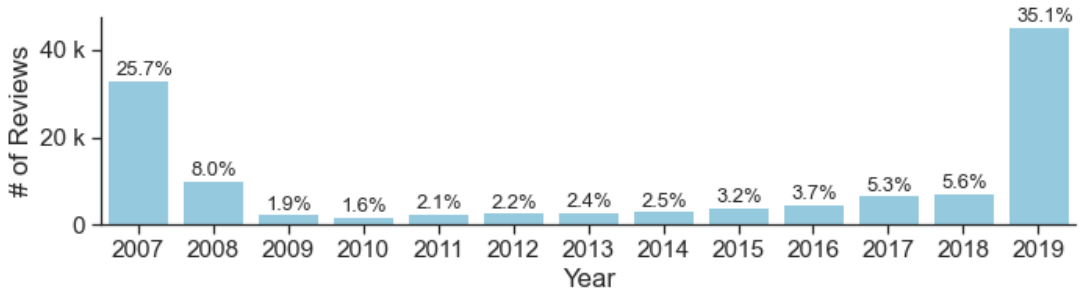

Figure 4: This figure shows the distribution of Goodreads classics reviews by year. The high number of reviews in 2007 and 2019 reflect the fact that, in addition to collecting default-sored reviews, we specifically collected the "oldest" reviews, most of which were published in 2007, and the "newest" reviews, most of which were published in 2019.

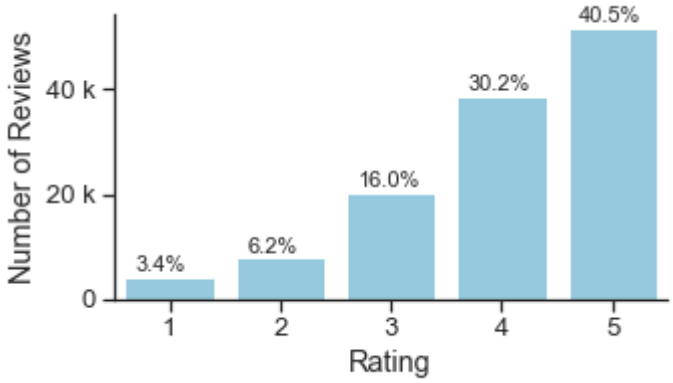

Figure 5: This figure shows the distribution of Goodreads classics reviews by star rating of the review.

Though the constraints of the Goodreads platform distort our dataset in certain ways, we tried to use this distortion to better scrutinize the influence of the web interface on Goodreads users. For example, the company never makes clear how it sorts reviews by default, but we found that reviews with a combination of more likes and more comments almost always appear above those with fewer — except in certain cases when there is, perhaps, another invisible social engagement metric such as the number of clicks, views, or shares that a review has received. Since we collected data in multiple sort settings, we are able to go further than this basic observation and investigate how exactly this default sorting algorithm shapes Goodreads users' behavior, social interactions, and

perceptions of the classics. Based on our analysis, we found that the first 300 default visible reviews for any given book develop into an echo chamber. Once a Goodreads review appears in the default sorting, in other words, it is more likely to be liked and commented on, and more likely to stay there (**figure 6**). Meanwhile the majority of reviews quickly age beyond "newest" status and become hidden from public view. These liking patterns reveal that Goodreads users reinforce certain kinds of reviews, such as longer reviews (**figure 7**), reviews that include a "spoiler alert" (**figure 9**), and reviews written by a small set of Goodreads users who likely have many followers (Table 2). If a review is prominently displayed by the default sorting algorithm, its author may be more likely to go back and modify this review. More default-sorted reviews included the words "update" or "updated" than oldest or newest reviews (**figure 8**). In one especially interesting updated review, a Goodreads user raised her rating of Toni Morrison's *The Bluest Eye* and apologized for the way that her original, more negative review offended others and reflected her white privilege, which other Goodreads users had pointed out.

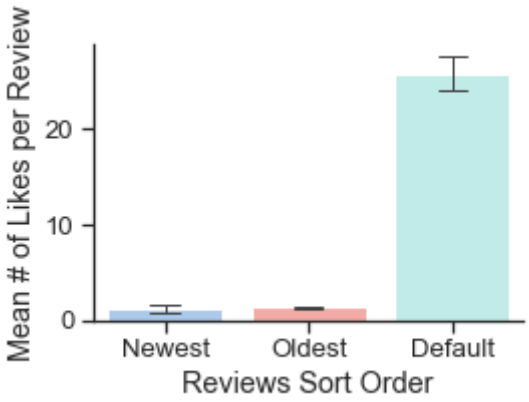

Figure 6: This figure shows the number of average likes per review, broken down by Goodreads main review sort orders. The error bars indicate the standard deviation across 20 bootstrapped samples of the books, providing a measure of instability when a particular book is included or excluded in the dataset.

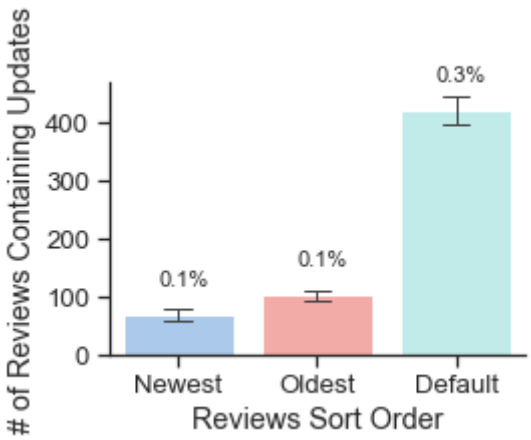

Figure 8: This figure shows the number of reviews that included the word "update" or "updated," Goodreads main review sort orders. The error bars indicate the standard deviation across 20 bootstrapped samples of the books, providing a measure of instability when a particular book is included or excluded in the dataset.

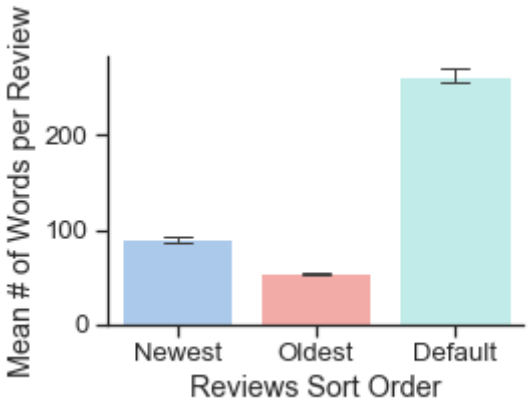

Figure 7: This figure shows the average length of reviews, broken down by Goodreads main review sort orders. The error bars indicate the standard deviation across 20 bootstrapped samples of the books, providing a measure of instability when a particular book is included or excluded in the dataset.

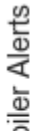

Figure 9: This figure shows the number of reviews that included a "spoiler" tag, broken down by Goodreads main review sort orders. The error bars indicate the standard deviation across 20 bootstrapped samples of the books, providing a measure of instability when a particular book is included or excluded in the dataset.

Topic Modeling Goodreads Reviews

Looking at the list of most popular Goodreads classics and analyzing liking patterns can only tell us so much about how Goodreads users perceive, define, and discuss the classics. To know more, we needed to listen to readers' own critical voices. To understand the most consistent conversations and overarching themes in Goodreads classics reviews, we analyzed the reviews with topic modeling, specifically a latent Dirichlet allocation (LDA) topic model: an unsupervised machine learning algorithm that essentially tries to guess the main themes of a collection of texts. [40] We preprocessed our reviews with Laure Thompson and David Mimno's "Authorless Topic Model" package to capture the most cross-cutting themes. [41] This package helps to remedy a common problem that occurs when topic modeling a collection of texts by multiple authors — or, in our case, a collection of reviews *about* texts by multiple authors —which is that the resulting topics often pick up on language specific to individual authors, such as words unique to Shakespeare plays or to Jane Austen novels. Author-specific topics can be desirable in some cases, but we wanted to reduce the signal of individual authors in order to amplify readers' collective voices across the reviews. The final 30 topics produced by the topic model help us pull out some of the major threads in the Goodreads clas-

sics reviews, which we manually labeled and split into four categories: "The Classics Industry," "Literary Themes," "Literary Qualities," and "Linguistic Styles." "The Classics Industry" includes topics such as "Adaptations & Audiobooks" and "Editions & Translations" (**figure 10**). The "Literary Themes" and "Literary Qualities" categories point to thematic or stylistic elements that readers' commonly discuss in their reviews, including topics such as "War & Adventure" or "Length & Pace" (**figure 11**, **figure 12**) Finally, the "Linguistic Styles" category captures both Goodreads users' writing styles and literary authors' writing styles, which commonly appear in the form of quotations. Sometimes the topics even pick up on a fascinating blend of readers' and authors' styles combined. For example, the "Conversational & Slangy" topic sometimes identifies the quoted voice of Holden Caulfield, *The Catcher in the Rye*'s angsty protagonist, but other times it identifies Goodreads users writing in a satirical Holden Caulfield-style voice (**figure 13**).

Topic Label	Top Words	Top Classics	Sample Review
School	school, high, time, class, first, remember, years, year, english, still, think, college, grade, since, teacher	The Scarlet Letter The Canterbury Tales The Outsiders Romeo and Juliet Lord of the Flies The Great Gatsby To Kill a Mockingbird	*This was the **first** Toni Morrison I read for 10th **grade** **English** while I was in **high** **school**. I couldn't get into at the **time**...I was more prepared for the novel this **time** around. —Goodreads User*
Editions & Translations	language, english, translation, words, text, edition, understand, original, version, word, time, first, modern, work	The Canterbury Tales Beowulf The Odyssey The Iliad The Prince War and Peace Don Quixote	*The Barnes and Noble **edition** places the full old **English** on one page with the **modern** **translation** on the other. You can read the old **English** to gain a sense of the rhythm and poetry and the **translation** for the content. —Matt*
Adaptations & Audiobooks	movie, audio, version, film, seen, listened, better, watched, movies, well, listening, adaptation, great, different, good	Breakfast at Tiffany's The Wonderful Wizard of Oz A Christmas Carol The Phantom of the Opera The Shining Peter Pan I Am Legend	*3 delicious hours of **audio** read by Mr. Michael C. Hall aka Dexter!!! What a wonderful performance of Truman Capote's novella! I saw the **movie** years ago but I've never read the book! -Jennifer Masterson*
Goodreads User Criticism	stars, didn, give, think, rating, star, maybe, get, writing, enjoy, feel, time, say	The Catcher in the Rye The Fellowship of the Ring The Old Man and the Sea Gullivers Travels Lolita Heart of Darkness Catch-22	*This review is inspired by some of my GR friends whose fearlessness about giving low **stars** to books they do not like has inspired me to change my **rating** of Lolita from three **stars** to two **stars** as that is what I really **feel** —Bren*
Review Culture & Meta-Review Discourse	review, first, years, com, edition, copy, new, goodreads, year, published, find, reviews, www, time, https,	A Christmas Carol Address Unknown Alice's Adventures in Wonderland The Mysterious Affair at Styles Fahrenheit 451 And Then There Were None	*For those new to me or my **reviews**...First the book **review** goes on **Goodreads**, and then I send it on over to my WordPress blog at **https://thisismytruthnow.com**, where you'll also find TV & Film **reviews** —James*

Figure 10: These are five of the 30 topics produced by our topic model (based on 120,000+ Goodreads reviews of "classic" texts), which we labeled The Classics Industry. The table displays our hand label for the topic; the most probable words for the topic; the texts that are most probable for the topic (when we aggregate all the reviews for that text); and a sample review that ranked highly for the topic, with top words bolded. For readability, we remove a set of common stopwords from the most probable words.

Topic Label	Top Words	Top Classics	Sample Review
Gender & Sexuality	women, men, woman, society, time, way, female, man, think, male, life, sex, mind, could	A Room of One's Own The Yellow Wall-Paper The Bell Jar The Color Purple The Handmaid's Tale Tess of the d'Urbervilles Their Eyes Were Watching God	*It is hard hitting to see what's written about **women** by **male** writers over the centuries. But what makes this book so beautiful is her ability to describe the **life** of a writer and how different it was for **men** and **women**. —Priya Sankar*
Race	novel, people, white,	Uncle Tom's Cabin	*I was floored to find that **white** **chracters** in*

	american, black, author, time, character, however, society, characters, still, narrative, reader, social, racism, written	Narrative of the Life of Frederick Douglass Things Fall Apart The Adventures of Huckleberry Finn Heart of Darkness To Kill a Mockingbird	*this book expressed a diverse range of opinions/thoughts/rationalizations for their position on slavery...EXACTLY the ones I continue to hear any time I talk to other **white people** about various **social**/political issues today like **racism**, poverty, education, immigration, etc. —Katie*
Family	family, life, man, father, people, mother, young, old, girl, away, good, boy, gets, get, home, children, lives, live	The Pearl The Grapes of Wrath A Tree Grows in Brooklyn Hatchet A Little Princess Of Mice and Men The Outsiders	*...this Pulitzer Prize-winning novel follows the Joad **family** who have been forced from their home in Oklahoma to travel west...I really felt the hardship of this poor **family** who simply want to **live** a **life** with dignity. —ebooksclassics*
Life & Death	life, human, man, self, nature, world, death, sense, god, yet, experience, without, perhaps, lives, humanity, may, mind, meaning, reader	The Death of Ivan Ilych As a Man Thinketh Siddhartha The Brothers Karamazov The Stranger The Picture of Dorian Gray	*Deals with the ephemeral way people usually live their **lives** focusing on the outward forms devoid of inner substance. Materialism, vanity, greed lust fill the **lives** and give it the illusion of **meaning** only to be shattered by the truth of **death**, Nicely written, brought tears to my eyes. —Yasir Sadiq*
War & Adventure	war, world, man, back, great, time, king, men, journey, first, adventure, end, two, long, island, battle	The Return of the King The Two Towers The Iliad A Farewell to Arms Prince Caspian The Old Man and the Sea Treasure Island	*Merry has joined with the renewed **King** of Rohan, who has joined Legolas and Gimli in their **battle** to save Gondor and fight for **men** against the amassed armies of Sauron... The Return of the **King**, the last leg of the **journey** of the One Ring is hopeful, rewarding, and highly informative... —Cupcake Book Lady (Amanda Leitch)*
Murder & Revenge	man, evil, death, good, murder, love, father, play, king, two, wife, end, crime, revenge, god, see	Macbeth Othello Hamlet In Cold Blood Crime and Punishment The Crucible The Count of Monte Cristo	*Hamlet’s ghost **father** demanding **revenge**, pretend insanity, **death**, real insanity, everyone plotting against each other, **death**, **play** within a **play**, more **death**, all wrapped up with insanely good poetry. —Tadiana Night Owl*
The Future (Dystopias)	world, people, society, human, war, future, new, power, history, political, science, fiction, thought, today, humans	1984 The Man in the High Castle Animal Farm Brave New World Do Androids Dream of Electric Sheep? Fahrenheit 451 The Handmaid’s Tale	*This was typical Phillip K. Dick fare, clever philosophical **science fiction** contemplating ideas about religion, **society** and in many ways what it is to be **human**...I felt that it provided clever parallels with the daily grind of **today’s** modern **world** —Jonathan*
Marriage	novel, young, first, mrs, woman, love, miss, two, family, friend, old, marriage, sister, years, well, wife, jane, character, man	Persuasion Mansfield Park Northanger Abbey The Mysterious Affair at Styles Emma Sense and Sensibility The Murder on the Links Rebecca	*Twenty-**two years old** and she is considered the most beautiful **young** lady in the neighborhood...Elizabeth’s rejection of Mr Collins’s **marriage** proposal is welcomed by her father... —Goodreads User*
Comedy	characters, play, work, sense, makes, often, little, seems, though, yet, perhaps, well, humor, funny, plot, almost	The Importance of Being Earnest Northanger Abbey Much Ado About Nothing A Midsummer Night’s Dream Don Quixote Catch-22 Candide	*A deliciously **funny** read...whose clever **plot** serves up some of Wilde’s funniest and must cutting wit...He repeatedly suspends his drama so that his **characters** can deliver impudent and hilarious asides... —Goodreads User*
Mystery & Suspense	spoiler, think, view, though, know, way, things, actually, time, hide, end, didn	The Murder on the Links The Turn of the Screw Death on the Nile The Strange Case of Dr. Jekyll and Mr. Hyde Murder on the Orient Express I Am Legend The Mysterious Affair at Styles	*A pretty solid mystery. **Though** the death **didn**’t happen until pretty far into the book, which was a bit surprising...(**view spoiler**)[I **didn’t really think** one of the murderer’s motivations was all that compelling **though**...(**hide spoiler**) —Monica*

Children's Literature	children, child, little, world, young, old, adult, kids, age, adventure, classic, boy, childhood, tale	The Lion the Witch and the Wardrobe The Magicians Nephew The Secret Garden Peter Pan A Little Princess Prince Caspian The Wonderful Wizard of Oz Anne of Green Gables	*An endearing **classic** and beautiful **tale** of friendship, animals and life. My 9 year **old** daughter and I really enjoyed reading this. I can't wait to read it again with my younger **child** —Sharen*

Figure 11: These are 11 of the 30 topics produced by our topic model (based on 120,000+ Goodreads reviews of "classic" texts), which we labeled Literary Themes. The table displays our hand label for the topic; the most probable words for the topic; the texts that are most probable for the topic (when we aggregate all the reviews for that text); and a sample review that ranked highly for the topic, with top words bolded. For readability, we remove a set of common stopwords from the most probable words.

Before fully diving into these topics, we want to briefly elaborate on the topic model to clarify this method and provoke a thought experiment. How might Goodreads and Amazon be extracting value from this data using computational methods? By demonstrating the kinds of patterns that our topic model can detect, we might better understand what's happening in Amazon's "engine-room," as Simone Murray puts it. [42] Because the topic model algorithm is "unsupervised," we do not specify in advance which topics to look for, only the number of topics to return. The number of topics that we decided on was a significant and subjective decision. The topic model is not an objective magic wand but an interpretive tool. We chose 30 topics because we experimented with different numbers and ultimately found that 30 topics produced the most coherent and compelling results.

Each topic consists of all the words in every recorded Goodreads review, ranked by their likelihood of appearing in a Goodreads review assigned to a particular topic. The most probable words for each topic typically represent a common theme, discourse, or linguistic style across the Goodreads reviews, such as "women," "men," "woman," "would," and "society," the five most probable words for the topic that we eventually hand labeled "Gender & Sexuality" (all topics were similarly hand labeled by us). These topic words may seem, at first glance, simplistic (e.g., "men" and

"women") or even arbitrary (e.g., "eyes," "upon," and "long"). Yet when we read through the individual Goodreads reviews that rank highly for each topic, we can start to understand their significance and critical utility. Simple words, it turns out, can help detect complex discussions of gender and race, and seemingly random groups of words can be the unexpected trademarks of particular linguistic styles. The topic containing the words "eyes," "upon," "long," "light," "man," "heart," and "world," for example, ranks highly in Goodreads reviews that include a quotation from the book being reviewed (**figure 13**). These basic words indeed identify the presence of literary language in a Goodreads review with remarkable regularity and accuracy, even across a wide range of source texts — from Fitzgerald's *The Great Gatsby* ("And as I sat there, brooding on the old unknown **world**, I thought of Gatsby's wonder when he first picked out the green **light**") to Morrison's *The Bluest Eye* ("God was a nice old white **man**, with **long** white hair, flowing white beard, and little blue **eyes**") to Shakespeare's *Macbeth* ("cleanse the stuffed bosom of that perilous stuff which weighs **upon** her **heart**"). These results bolster our confidence that the model is picking up on significant threads even when the assemblages of topic words do not seem immediately coherent. This ability to find significant threads playing out in individual Goodreads reviews is one of the major assets of the topic model for humanistic interpretation. We use the topic model not only to identify broad patterns in the collection but also to draw specific and noteworthy examples to the surface and to our critical attention.

Topic Label	Top Words	Top Classics	Sample Review
Critical Status	novel, fiction, literature, great, novels, work, time, classic, written, first, century, writing, works, modern, literary, best	The Time Machine The War of the Worlds Dracula A Room of One's Own Don Quixote Journey to the Center of the Earth Frankenstein	*A **classic** tale from a book considered as a **first modern novel written** in 1605 … considered by many the **best novel written**, I will say it's one of the funniest I've read ever! -Kaushlendra*
Plot & Characters	characters, character, novel, plot, reader, first, main, interesting, two, end, well, part, however, different, events, writing	Death on the Nile A Study in Scarlet The Man in the High Castle The Mysterious Affair at Styles The Murder on the Links The Two Towers Murder on the Orient Express	*A **well**-told **story**, this book had a great **plot** and some stable **character** development. The phantom's **character** was **well**-established and very **well** done, as he is an extremely complex **character** -Benjamin Hollon*
Unlikeable Characters	characters, character, didn, felt, could, found, writing, plot, good, boring, feel, get, main, nothing	Mansfield Park A Farewell to Arms Madame Bovary Emma Wuthering Heights The Sun Also Rises The Catcher in the Rye Sense and Sensibility Northanger Abbey	*A flat, disappointing story populated with one-dimensional **characters** that I grew to dislike. The narrator is a whiny young man prone to fainting and with very little backbone … it's also, frankly, **boring** —Rachel (Kalanadi)*
Beautiful Writing	writing, written, novel, beautiful, tale, characters, short, well, yet, way, human, reader, truly, prose, mind, powerful, beautifully	Chess Story The Yellow Wall-Paper The Things They Carried In Cold Blood The Pearl The Bluest Eye Address Unknown Lolita	*This was Toni Morrison's first **novel** and is a gritty and surreal **tale** of coming of age **written** in a **beautiful** flowing **prose**. So happy to have this book and should be read by all. —Luke Forsyth*
Length & Pace	time, first, pages, get, long, last, took, got, finally, years, started, end, finished, page, could, finish	Don Quixote One Hundred Years of Solitude Catch-22 War and Peace The Return of the King The Count of Monte Cristo The Brothers Karamazov	*This book literally took me **years** to **finish** … In a funny sense though this made me feel that I too was stuck on that island for **years** with Crusoe which didn't really harm the novel. —Nour*
Thought-Provoking	life, people, think, things, way, feel, world, time, different, person, see, made	Tuesdays with Morrie The Five People You Meet in Heaven The Diary of a Young Girl The Death of Ivan Ilych As a Man Thinketh The Alchemist Flowers for Algernon	*It had me thinking about my own **life**, the mistakes I have **made**, the **things** I have done well and the **things** I want to do better. The **way** I want to **live** my **life** and the ebb and flow of **life**. —Goodreads User*
Enjoyable & Interesting	enjoyed, interesting, good, liked, thought, bit, lot, found, well, little, didn, quite	The Strange Case of Dr. Jekyll and Mr. Hyde Dracula Around the World in Eighty Days The Time Machine The Phantom of the Opera Murder on the Orient Express Journey to the Center of the Earth	*I **didn**'t **enjoy** the writing as much as i **thought** i would ! I heard so many great things about this book so i decided to pick it up…There was a **lot** of scenes that i **didn**'t really care about … (but would recommend it to you if you love classics) —Alya Abu Khazna*
Re-Readable	time, favorite, ever, times, best, still, love, first, years, every, loved, great, written, amazing, definitely, good	A Christmas Carol Anne of Green Gables To Kill a Mockingbird Pride and Prejudice As a Man Thinketh The Importance of Being Earnest And Then There Were None	*5/5 Stars I'm not sure how many **times** I have read and re-read this book, but I always find something new **every time**. Just a wonderful **time** re-visiting this world! Recommended for everyone. —Amber Hyde*

Figure 12: These are eight of the 30 topics produced by our topic model (based on 120,000+ Goodreads reviews of "classic" texts), which we labeled Literary Qualities. The table displays our hand label for the topic; the most probable words for the topic; the texts that are most probable for the topic (when we aggregate all the reviews for that text); and a sample review that ranked highly for the topic, with top words bolded. For readability, we remove a set of common stopwords from the most probable words.

Topic Label	Top Words	Top Classics	Sample Review
Literary Language (Quotations)	man, heart, eyes, never, could, upon, every, light, mind, life, dark, men, time, long, world, words, ever	The Prophet Macbeth Moby-Dick or the Whale Mrs. Dalloway The Return of the King The Fellowship of the Ring The Iliad	*“and there quivered and felt the **world** come closer…**Life** itself, **every** moment of it, **every** drop of it, here, this instant, now, in the sun, in Regent’s Park…a freedom which the attached can **never** know.” —Mrs. Dalloway*
Conversational & Slangy Language	know, get, good, think, people, thing, want, going, mean, pretty, say, got, kind, guy, didn, right, lot	The Catcher in the Rye Romeo and Juliet Much Ado About Nothing The Shining Dracula Hamlet The Outsiders	*GHOST/DAD: Hamlet, your uncle killed me and married your mom. I **want** vengeance, so best **get** to murdering, plzthnx … GERTRUDE: Yum, poisoned wine. *dies* CLAUDIUS: Whoops, my bad. HAMLET: I KEEL YOU! —Madeline*
Description & Dialogue (Quotations)	said, little, could, time, back, day, went, got, around, night, right, old, get, going, two, home, know	The Sun Also Rises And Then There Were None Hatchet The Shining The Adventures of Tom Sawyer Alice’s Adventures in Wonderland Murder on the Orient Express	*“Five **little** Indian boys **going** in for law; One **got** in Chancery and then there were four.Four **little** Indian boys **going** out to sea; A red herring swallowed one and then there were three” —And Then There Were None*
Gushing & Loving Language	love, loved, characters, heart, character, beautiful, life, favorite, always, end, made, never, happy, every, ever,	Persuasion Anne of Green Gables Romeo and Juliet Gone with the Wind Pride and Prejudice Jane Eyre Sense and Sensibility	*I think I met one more favorite author through this book. I **loved** it and I **loved** how it made me fall in **love** with every tiny detail of it. —Nada Elshabrawy*
Talking & Speaking	know, say, people, think, good, things, said, could, something, never, want, thing, make, world, way, see	The Prince Tuesdays with Morrie The Little Prince As a Man Thinketh The Alchemist Siddhartha Jonathan Livingston Seagull To Kill a Mockingbird	*I don’t think I can **say** anything about this book that hasn’t been **said** a thousand times before, so I’ll **say** nothing except, exceptional. —Barry*
Non-English Reviews	que, non, una, con, che, come, los, tea, per, del, cup, por, para, era, libro, como, las, final, pero	Hamlet Great Expectations The Odyssey The Giver The Outsiders Fahrenheit 451 The Little Prince	*Shakespeare, Oh meu querido Shakespeare, **Que** tão bem sabias **como** arrebatar nossas emoções …“Hamlet” recomenda-se, mais ainda **para** quem aprecia a arte de contar histórias —Nelson Zagalo*

Figure 13: These are six of the 30 topics produced by our topic model (based on 120,000+ Goodreads reviews of "classic" texts), which we labeled Linguistic Styles and Non-English Reviews. The table displays our hand label for the topic; the most probable words for the

topic; the texts that are most probable for the topic (when we aggregate all the reviews for that text); and a sample review that ranked highly for the topic, with top words bolded. For readability, we remove a set of common stopwords from the most probable words.

By aggregating all ~900 reviews for each classic book, we can also identify the topics most associated with every book and, conversely, the books most associated with every topic. The classics that rank highest for the topic we have labeled "Gender & Sexuality" — which includes words like "women," "men," woman," and "society" — are literary works that explore subjects related to women's writing, feminism, misogyny, reproductive rights, and lesbian desire: Virginia Woolf's "A Room of One's Own" (1929), Charlotte Perkins Gilman's "The Yellow Wallpaper" (1892), Sylvia Plath's *The Bell Jar* (1963), Alice Walker's *The Color Purple* (1982), and Margaret Atwood's *The Handmaid's Tale* (1984) (**figure 11**). The classics that rank highest for the topic we have labeled "Race" — which includes words like "white," "black," "society," and "racism" — revolve around issues such as American slavery, the effects of anti-Black racism, and the colonization of Africa: Harriet Beecher Stowe's *Uncle Tom's Cabin* (1852), Frederick Douglass's *Narrative of the Life of Frederick Douglass* (1845), Chinua Achebe's *Things Fall Apart* (1958), Mark Twain's *The Adventures of Huckleberry Finn* (1884), Toni Morrison's *The Bluest Eye* (1970), and Joseph Conrad's *Heart of Darkness* (1902) (**figure 11**). These coherent clusters of literary works, grouped from within the broader 144 classic titles, are surprisingly intuitive classifications for an unsupervised algorithm trained on readers' responses alone, with no access to the texts themselves or to any external metadata about author, publication, or reception. Further, these clusters paint impressionistic pictures of the collective reader response to each book. For the hand-selected group of texts in **figure 14**, we can see which books generated more discussion of classrooms and school and which books generated more discussion of life and death, which books were more likely to be quoted from and which books were more likely to inspire gushing declarations of love. By incorporating rating information, we can also identify which topics corresponded to more positive ratings, like "Beautiful Writing," and which to more negative ratings, like "Unlikeable Characters" (**figure 15**). Using computa-

tional methods on Goodreads data, it is thus possible to learn a lot of information about readers — the kind of information that is ironically valuable both to literary critics and to corporations like Amazon.

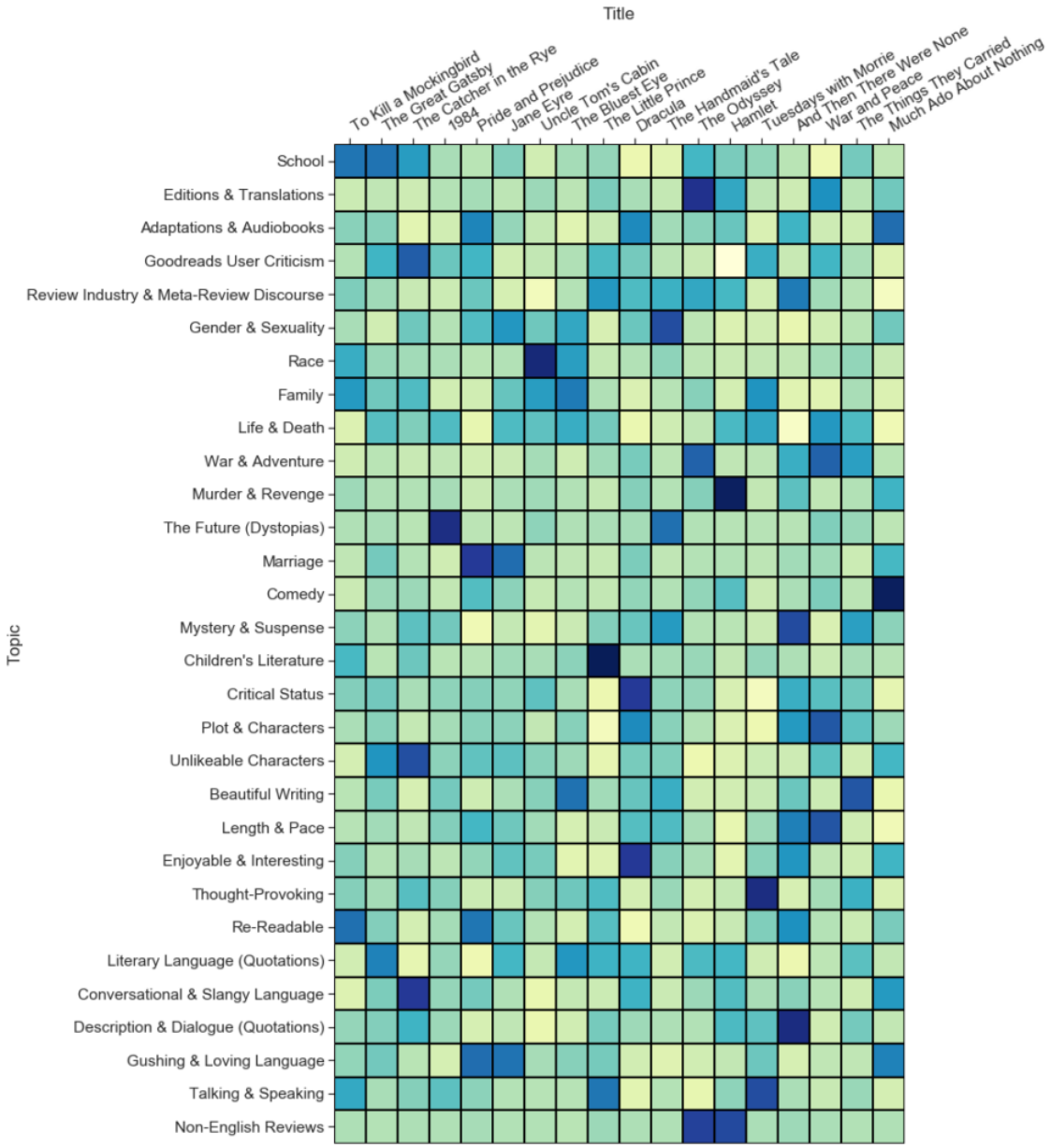

Figure 14: This heatmap represents the probability that Goodreads reviews for a given book would contain one of the 30 topics on the left. It can also be explored as an interactive data visualization. Darker tiles indicate a higher probability of containing the topic. Scanning left-to-right for the "School" topic, for example, reveals that To Kill a Mocking Bird, The Great Gatsby, and The Catcher in the Rye have the darkest tiles in this row, which indicates that reviews of these books are most likely to discuss school-related

subjects. Scanning top-to-bottom for Pride and Prejudice, to take another example, reveals darker tiles for the topics "Audiobooks & Adaptations," "Marriage," "Re-Readable," and "Gushing & Loving Language." The heatmap rows have been normalized to highlight differences between the books. We check the significance of these results via 95% bootstrapped confidence intervals, and the majority of visible differences are significant.

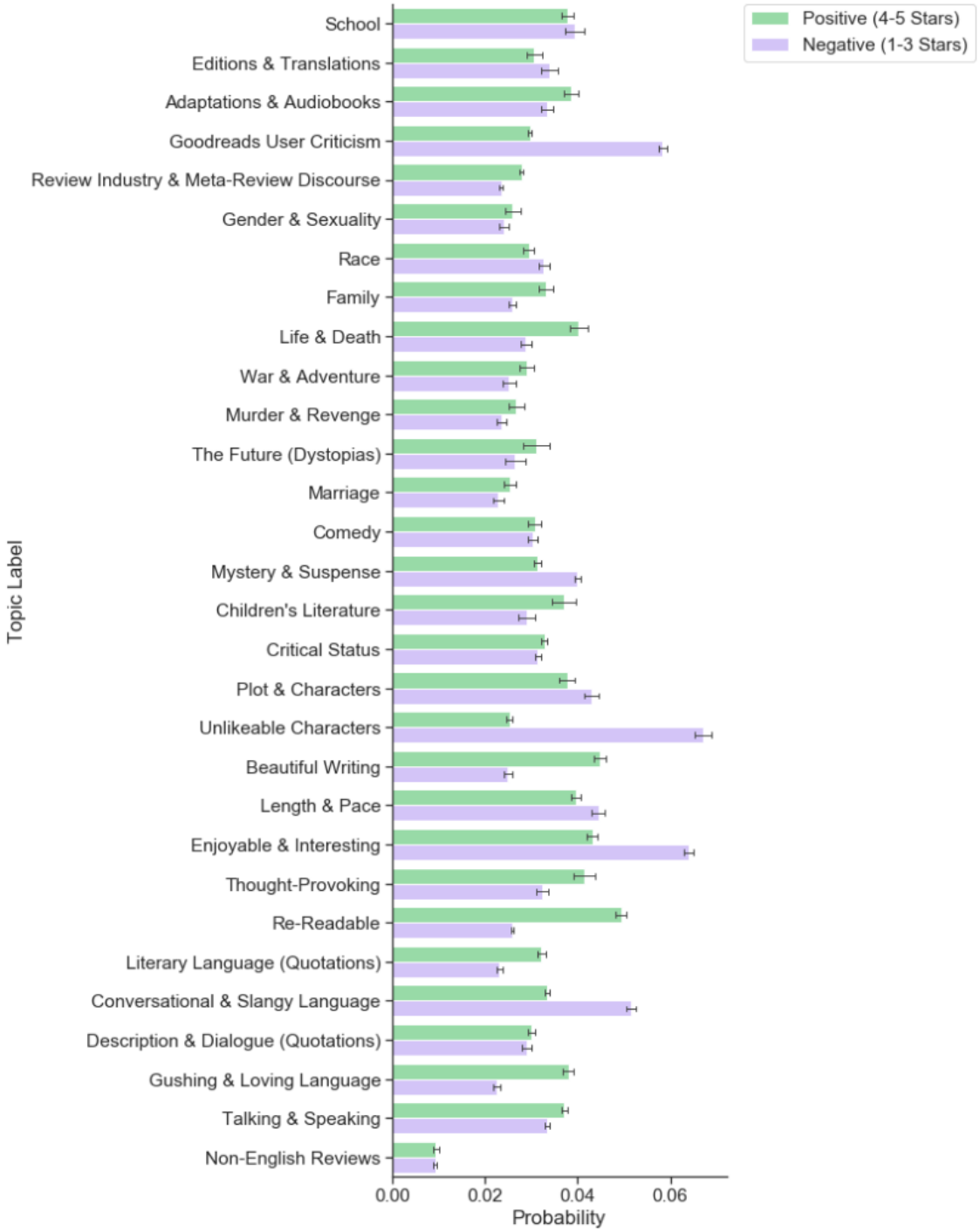

Figure 15: This figure shows whether Goodreads users were more likely to rate books positively (4-5 stars) or negatively (1-3 stars) when their reviews were likely to contain a certain topic. When Goodreads users published reviews that were likely to contain the "Unlikeable Characters" topic, for example, they tended to rate the text in question negatively. Perhaps counterintuitively, when Goodreads users published reviews likely to contain the "Enjoyable & Interesting" topic, they were also more likely to rate the text negatively, because reviewers often discussed not enjoying a book and not finding it interesting. These results are based on the full set of Goodreads reviews — all books in all three sort orderings. The error bars indicate the standard deviation across 20 bootstrapped samples of the books, providing a measure of instability when a particular book is included or excluded in the dataset.

The Classics Industry

The rest of this essay focuses on the category that we have labeled "The Classics Industry," the set of topics that help point to some of the institutions and phenomena most responsible for reinforcing the classics in the twenty-first century. This formulation is partly inspired by Murray's sociological account of the "adaptation industry," in which she maps "the industrial structures, interdependent networks of agents, commercial contexts, and legal and policy regimes within which adaptations come to be," mostly focusing on book-to-screen adaptations. [43] Though Goodreads users often allude to the academy and professional literary critics in their reviews, the prevalence of the term "classic" itself points to the shaping influence of forces beyond the academy. To put this prevalence in concrete numbers, more than 15,000 Goodreads reviews explicitly mentioned the words "classic" or "classics," while just under 400 reviews mentioned the words "canon" or "canonical." This simple metric reveals a clear fault line in literary critical discourse between scholars and readers. It also indexes the power of the classics as a marketing brand. We detail how this brand functions in the sections below, and we also call attention to the ways that Amazon specifically influences and profits from this branding.

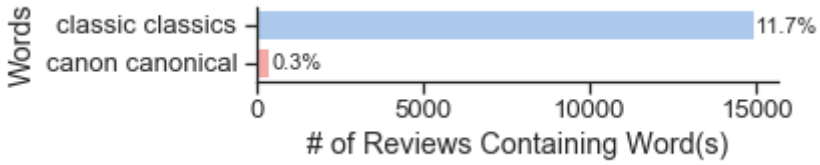

Figure 16: This figure displays the number of Goodreads reviews that explicitly mentioned "classic" or "classics" vs. "canon" or "canonical."

The Classics Industry: School

Though Goodreads users rarely discuss the canon and scholars today rarely discuss the classics, the academy remains an important engine for the classics industry. The topic that we have labeled "School, "which includes words like "school," "high," "time," "class," "first," "remember," "years," and "english," identifies the clear influence of school systems on Goodreads users' conceptions of the classics, aligning with theories of cultural production and canon formation proposed by scholars like John Guillory and Pierre Bourdieu. [44] The Goodreads reviews that rank highly for this topic reveal a few key patterns. While some Goodreads users talk about recent experiences in English literature classes, many more discuss literature classroom experiences from the past or refer to more generalized conceptions of "required reading." "This was the first Toni Morrison I read for 10th grade English while I was in high school," one Goodreads user reflected about Morrison's *The Bluest Eye* (1970), which she shelved under "classics." "I couldn't get into [it] at the time - and I think a good chunk of that had to do with how the story and it[s] difficult subjects were handled in a classroom setting. Now that I can say I've read it again for Book Riot's 2018 Read Harder Challenge (an assigned book you hated or never finished), I could definitely appreciate it more." [45] When users catalogue their reading histories, high school and college reading often figures as an essential part of a fully comprehensive account. Classics consumed from one's school days serve as something like a starter pack for a Goodreads catalogue, providing an easy way to rate and review a number of books immediately. Even Goodreads users who have wildly disparate genre inclinations will likely share these schoolbooks in common if they share common backgrounds.

Because of these common shared experiences, schoolbooks foster social interactions between users, and communities commonly form around and through them — whether to read a classic for the first time or to reread a previously hated classic á la Book Riot's **Read Harder Challenge**. Popular conceptions of school syllabi and required reading shape readers' habits long after their school days, and readers even self-assign books in order to belatedly join these communities. "Somehow I was never assigned to read this in high school, so I'm reading it now!" Goodreads user Edward Rathke exclaimed about *The Grapes of Wrath*. [46] "I had been planning to read '1984' for a long time," explained another Goodreads user named Andrew. "It's one of those books that you are supposed to read in high school. My high school AP Lit teacher had us read Aldous Huxley's 'Brave New World' instead." [47] These reviews may also explain why the Goodreads classics are less racially diverse than contemporary literary syllabi, since readers are clearly influenced by historical and imagined literary curricula more than contemporary literary curricula.

The Classics Industry: Publishing

School syllabi feed the classics industry. They are undoubtedly one of the reasons, if not the primary reason, that the classics are a prominent advertising target value on Goodreads. But they also feed another major node in the classics industry: the publishing industry. High school and college syllabi, as Rebecca Rego Barry writes, "are profitable to the classics publisher because they have a known market. These titles are thus doubly promoted for entrance to the canon, in classrooms and bookstores, and it is interesting to note that professors and publishers are symbiotic in this respect." [48] The topic that we labeled "Editions & Translations," which includes words like "translation," "edition," "original," and "version," picks up on discussions about which *edition* or *version* of a particular classic Goodreads users have read, purchased, or borrowed. These are the Penguin Classics, the Signet Classics, and the Modern Library Classics, which make up a significant part of the literary market. "The classics market is huge," *The Guardian* reported in 2016. "There's been a noticeable upswing in the number of publishers doing the classics." [49] Though comprehensive book sales

data is hard to come by, according to Publishers Weekly and NPD BookScan, the "classics" sold almost 3.6 million units in the first half of 2018 — making it the fifth-best selling literary category behind "General Fiction," "Suspense/Thrillers," "Romance," and "Mystery/Detective," and ahead of genre fiction heavyweights like "Fantasy" and "Science Fiction." [50] Even Amazon has now developed its own line of classics: *AmazonClassics*. In fact, almost every Goodreads classic currently in the public domain now has an *AmazonClassics* Kindle e-book for sale. As the series title *AmazonClassics* confirms, the publishing industry is one of the major forces that contributes to the gulf between "classic" and "canon" in readers' critical vocabularies. [51]

The Classics Industry: Adaptation & Audiobooks

Various classics editions from disparate publishers reestablish the classics in concert. They are solidified, as Barry puts it, not through any one edition but "through the continuous promotion of a given title in more than one imprint, certified by more than one set of arbiters over a longer period of time." [52] This "continuous promotion" is not limited to print publishing. The proliferation of literary texts into other mediums further reinscribes certain books as classics, as Sarah Cardwell argues and as our analysis confirms. [53] The topic that we have labeled "Audiobooks & Adaptations," which includes words like "movie," "audio," "version," "seen," and "listened," captures how Goodreads users' sense of the classics is shaped by adaptations. The cluster of classics that rank highest for this topic — Truman Capote's *Breakfast at Tiffany's*, L. Frank Baum's *The Wonderful Wizard of Oz*, Charles Dickens's *A Christmas Carol*, and Stephen King's *The Shining* — all have major decades-old Hollywood film adaptations. But many of the high-ranking Goodreads reviews in this topic also discuss audiobooks, which share a surprisingly strong relationship with film and television adaptations and with Amazon. Audible, the world's largest producer of audiobooks, is yet another subsidiary of Amazon. In the last 10 years, Audible has invested in a series of classic literature audiobooks, "Audible Signature Classics," narrated by famous film and television actors. Most of these classics are the same popular Goodreads classics

that we have already identified, paired with a performance by a high profile celebrity: *The Great Gatsby* narrated by Jake Gyllenhaal (2013), *Alice's Adventures in Wonderland* narrated by Scarlett Johansson (2016), *The Things They Carried* narrated by Bryan Cranston (2013), *Anne of Green Gables* narrated by Rachel McAdams (2016). [54] This catalogue represents what Cardwell refers to as "circular affirmation," when a certain selection of books are reinforced as classics by being adapted and confirmed "across various areas of the public sphere" — confirmed, in this case, not only through audiobook adaptation but also through association with Hollywood celebrities. [55] Based on our collection of Goodreads reviews, we find that this circle of affirmation sometimes marginalizes the print text itself. For example, in one review of Truman Capote's *Breakfast at Tiffany's*, a Goodreads user named Jennifer Masterson shelved the novella under "classics" and gushed:

> 3 delicious hours of audio read by Mr. Michael C. Hall aka Dexter!!! What a wonderful performance of Truman Capote's novella! I saw the movie years ago but I've never read the book! I'm so happy to have listened to this edition of the audio! 5+++++Stars for the narrator! 5 Stars for the story! Highly highly recommended!!! [56]

This review of an Audible Original audiobook, narrated by a television star, *Dexter*'s Michael C. Hall, was inspired by watching a Hollywood film. And though this Hollywood film was originally based on a novella, Jennifer, this particular Goodreads user, never read the novella and did not need to in order to review the book on Goodreads and perpetuate the classics industry. This review also demonstrates that Amazon-affiliated audiobooks inspire users to visit and rate books on Goodreads, to bounce from one Amazon subsidiary to another. We speculate that Amazon may also use Goodreads data to help determine which audiobooks, television shows, and films to invest in. One of the earliest Amazon Studios television series was an adaptation of Philip K. Dick's *The Man in the High Castle* (1962), a popular Goodreads classic, and one of Amazon Studios' biggest investments is a television series based on the

Goodreads classics with the highest average ratings in our dataset, J.R.R. Tolkien's *The Lord of the Rings* trilogy, the rights for which Amazon purchased for $200 million. [57] We are not claiming that Goodreads reviews and ratings directly motivated this decision. But it is important to recognize that Goodreads data is controlled by Amazon, a corporation that is making some of the most expensive and high profile literary investments of our time.

The Classics Industry: Goodreads Users

The classics are clearly perpetuated by many powerful institutions as well as the market economy. When Goodreads users shelve, rate, and review classics, they contribute to this system and help sustain it. Making this point forcefully, Murray argues:

> The [Goodreads] website's beguiling abundance of actual reader responses to books has obscured for scholars the limited extent to which users either understand or can influence its algorithmic operations, leading to overblown claims of readerly empowerment. Compelling evidence of reading's contemporary resilience and freely available research archive though it may be, Goodreads is above all else a node in platform capitalism. [58]

Goodreads is indeed "a node in platform capitalism," but we believe it is important to engage with how "beguiling" Goodreads reviews are and how empowering the platform can feel for some Goodreads users. In Aarthi Vadde's study of "amateur creativity" on the internet, she argues that it is not possible to "make a blanket case for or against the emancipatory potential of participatory culture on the Internet." [59] Instead Vadde suggests thinking of the public sphere as "an always already commercialized, industrialized, and pluralized space." [60] We believe this framing is helpful for teasing out how Goodreads users sometimes explicitly resist Goodreads and produce remarkably interesting amateur criticism all while being exploited by Goodreads.

One of the most fitting metaphorical representations of this ironic tension manifests when Goodreads users bash the classics, be-

cause, in doing so, they simultaneously reject and reinforce books as classics in the same stroke. The topic that we have labeled "Goodreads User Criticism," which includes words like "stars," "give," and "rating," picks up on a common rhetorical trope — the justification of a user's rating for a given text — and it includes a significant amount of classics bashing. We find that Goodreads reviews that rank highly for this topic are, overall, more likely to rate a text negatively (**figure 15**). Negative ratings seem to demand lengthy, reflexive justifications in their accompanying reviews. For example, a Goodreads user named Bren, mentioned in the introduction of this essay, shelved Nabokov's *Lolita* as a classic but rated the novel only three out of five stars. Though three stars was already a low rating (particularly within the Goodreads community), she later returned to the review and lowered the rating still further. In her updated review, Bren explained that she originally gave *Lolita* a higher rating "in deference of its classic status." [61] But as she watched other Goodreads users openly panning books, including *Lolita*, she gained new confidence to dissent from *Lolita*'s perceived reputation and from its imagined community of fans, whom she dubbed "book snobs." This retroactive rating is a triumphant moment that Bren jokingly compares to winning an Academy Award:

> This review is inspired by some of my GR [Goodreads] friends whose fearlessness about giving low stars to books they do not like has inspired me to change my rating of Lolita from three stars to two stars as that is what I really feel . . . I get that this a classic and book snobs who read this will sig[h] in indignation but I do not care. I just did not get it and still don't. I'd like to thank anti book snobs everywhere for giving me the courage to rate Lolita two stars. I will never forget you. Wow..is this what an Oscar speech feels like? [62]

Many Goodreads users like Bren seem to feel liberated when they reject the classics and express honest negative opinions about exalted books. When we reached out to Bren to seek her permission to publish this review, she further elaborated about what the

Goodreads community means to her and even alluded to its special significance during the COVID-19 pandemic: "There is something about speaking against a Classic that can be very intimidating. People on here are fearless and, at least for me, I never feel judged . . . When I first joined I was too shy to talk to people but years later, I have connected with wonderful people and it has become a wonderful source of comfort to me, especially in trying times like these." [63] For Bren, the Goodreads community is sincerely meaningful, and the ability to speak out against a classic is genuinely empowering.

Another Goodreads user, Peter Derk, reflected about the joys of publishing "really nasty review[s]" of the classics, but his joy, unlike Bren's, was premised on the perceived powerlessness of his Goodreads review in the face of a classic:

> Every so often I'll get into a classic. I guess because I feel like writing a really nasty review. Classics are great fodder for nasty reviews because 1. The people who made them are LONG dead . . . Saying bad stuff about a classic novel doesn't hurt the creator's feelings . . . 2. Classics have such a pedestal in the literary world already that the opinion of one lone weirdo . . . is pretty irrelevant. It's not like bashing on this book is suddenly going to render it a Not A Classic or affect its sales. Frankly, I think that about everything I read, but with classics, it's a pretty rock solid premise. [64]

Rather than an emboldened community taking on *Lolita*'s classic reputation, as Bren framed herself and her "GR friends," Derk describes himself as "one lone weirdo" who couldn't possibly make a dent in a classic book's reputation. Far from being able to hurt a classic's sales, as Derk acknowledges, his colorful, vehement 2000-word takedown of *The Phantom of the Opera* likely only contributes to its contemporary value by contributing to its continued discussion. This paradox is one of the reasons that the classics remain so powerful. Love them or hate them, the classics sustain themselves by staying in print, remaining a topic of conversation, and enduring as a commodity.

Conclusion

So what is a "classic" in the twenty-first century? Based on our analysis of 144 Goodreads classics and 120,000 accompanying reviews, there are at least a few clear answers. For Goodreads and Amazon, a classic is a prominent advertising target value, a marketing tool, and a source of lucrative adaptation material. For Goodreads users, a classic is a book read in high school, a book that inspired a TV show, or a book that other Goodreads users have tagged as a classic. As we have shown, the classics industry — the collaborative forces of publishing, film, television, Amazon, and more — defines the status of popular classics to a large extent. Yet for Goodreads users, a classic is also an invitation to become amateur critics and creative writers, a chance to reflect on their lives and relationships to power, a conduit for connecting to others, and an opportunity to enter a critical conversation that has long excluded them. Literary history lives both profitably and vibrantly in the world under the moniker of the classics. To recognize the significance of the term is to recognize some of the places where literary criticism is most alive, relevant, and valuable.

Beyond the classics, this essay also points to major trends in contemporary literary culture that pose data-related challenges for literary critics — trends such as the rise of reader social networks, online amateur criticism, and Amazon. We believe that computational methods like the ones used in this essay can play a significant role in facing these challenges. When combined, computational methods and internet data can help literary critics simultaneously capture the creative explosion of reader responses as well as critique algorithmic culture.

Appendix

User Ethics

Like professional book reviewers, many Goodreads users take pride in their reviews and craft them carefully. If we think of Goodreads reviewers as creative artists or amateur critics, as the authors

themselves seem to do do, then anonymizing their reviews (removing their names and/or paraphrasing the review text) would deprive them of proper creative credit. [65] However, prior work has shown that even when internet users post on public platforms, they have an expectation of privacy. [66] For these reasons, we have chosen not to publicly share our dataset, though we have shared the code that we used to collect data from the Goodreads website: **https://github.com/maria-antoniak/goodreads-scraper**. For Goodreads reviews directly quoted in this essay, we have obtained explicit permission from each reviewer. We messaged each of these selected reviewers on Goodreads, disclosed our affiliations and the project goals and structure, and asked for consent to publish parts of their review in this article. We offered users the option of being quoted in this essay and attributed by their Goodreads username or the option of being quoted in this essay but remaining anonymous. The Goodreads users who chose to be quoted but remain anonymous are simply referred to as "Goodreads user" throughout the essay.

The Goodreads Classics

This table includes the 144 Goodreads "classics" examined in this essay as well as Goodreads reception statistics from 2019. You can also explore a **searchable, sortable version of this table** with up-to-date Goodreads statistics.

Author	Title	Year	# Ratings	# Reviews	Classics Category
Homer	The Illiad	~750 BC	798k	11k	Most Shelved
Homer	The Odyssey	~700 BC	321k	6k	Most Shelved
Unknown	Beowulf	975	211k	6k	Most Shelved and Most Read
Geoffrey Chaucer	The Canterbury Tales	1390	176k	3k	Most Shelved
Niccolò Machiavelli	The Prince	1513	228k	7k	Most Read
William Shakespeare	Romeo and Juliet	1595	2M	18k	Most Shelved and Most Read
William Shakespeare	A Midsummer Night's Dream	1595	412k	7k	Most Shelved and Most Read
William Shakespeare	Much Ado About Nothing	1598	202k	3k	Most Shelved
William Shakespeare	Hamlet	1603	662k	11k	Most Shelved and Most Read
William Shakespeare	Othello	1603	288k	6k	Most Shelved and Most Read
Miguel de Cervantes Saavedra	Don Quixote	1605	187k	7k	Most Shelved

William Shakespeare	Macbeth	1606	610k	10k	Most Shelved and Most Read
Daniel Defoe	Robinson Crusoe	1719	228k	6k	Most Shelved
Jonathan Swift	Gulliver's Travels	1726	214k	5k	Most Shelved
Voltaire	Candide	1759	202k	7k	Most Shelved and Most Read
Jane Austen	Sense and Sensibility	1811	880k	14k	Most Shelved
Jane Austen	Pride and Prejudice	1813	3M	58k	Most Shelved and Most Read
Jane Austen	Mansfield Park	1814	263k	9k	Most Shelved
Jane Austen	Emma	1815	578k	15k	Most Shelved and Most Read
Mary Wollstonecraft Shelley	Frankenstein	1818	1M	27k	Most Shelved and Most Read
Jane Austen	Persuasion	1818	477k	16k	Most Shelved and Most Read
Jane Austen	Northanger Abbey	1818	274k	11k	Most Shelved

					and Most Read
Victor Hugo	The Hunchback of Notre-Dame	1831	157k	4k	Most Shelved
Charles Dickens	Oliver Twist	1838	292k	7k	Most Shelved
Charles Dickens	A Christmas Carol	1843	538k	15k	Most Shelved
Alexandre Dumas	The Count of Monte Cristo	1844	690k	19k	Most Shelved
Alexandre Dumas	The Three Musketeers	1844	275k	7k	Most Shelved
Frederick Douglass	Narrative of the Life of Frederick Douglass	1845	83k	4k	Most Read
Charlotte Brontë	Jane Eyre	1847	1M	37k	Most Shelved and Most Read
Emily Brontë	Wuthering Heights	1847	1M	32k	Most Shelved and Most Read
Charles Dickens	David Copperfield	1849	186k	6k	Most Shelved
Nathaniel Hawthorne	The Scarlet Letter	1850	647k	14k	Most Shelved and Most Read
Herman Melville	Moby-Dick, or, the Whale	1851	449k	14k	Most Shelved
Harriet Beecher Stowe	Uncle Tom's Cabin	1852	182k	7k	Most Shelved
Gustave Flaubert	Madame Bovary	1856	225k	9k	Most Shelved

Charles Dickens	A Tale of Two Cities	1859	755k	16k	Most Shelved and Most Read
Charles Dickens	Great Expectations	1861	596k	15k	Most Shelved
Victor Hugo	Les Misérables	1862	622k	15k	Most Shelved
Jules Verne	Journey to the Center of the Earth	1864	134k	5k	Most Read
Lewis Carroll	Alice's Adventures in Wonderland	1865	186k	8k	Most Read
Fyodor Dostoevsky	Crime and Punishment	1866	550k	16k	Most Shelved and Most Read
Leo Tolstoy	War and Peace	1867	241k	10k	Most Shelved
Louisa May Alcott	Little Women	1869	1M	20k	Most Shelved and Most Read
Lewis Carroll	Alice's Adventures in Wonderland & Through the Looking-Glass	1871	427k	10k	Most Shelved
Jules Verne	Around the World in Eighty Days	1873	167k	6k	Most Read
Mark Twain	The Adventures of Tom Sawyer	1875	693k	9k	Most Shelved
Leo Tolstoy	Anna Karenina	1877	593k	21k	Most Shelved and Most Read
Fyodor Dostoevsky	The Brothers Karamazov	1879	217k	10k	Most Shelved

Robert Louis Stevenson	Treasure Island	1882	363k	10k	Most Shelved and Most Read
Mark Twain	The Adventures of Huckleberry Finn	1884	1M	14k	Most Shelved
Robert Louis Stevenson	The Strange Case of Dr. Jekyll and Mr. Hyde	1886	330k	11k	Most Shelved and Most Read
Leo Tolstoy	The Death of Ivan Ilych	1886	75k	4k	Most Read
Arthur Conan Doyle	A Study in Scarlet	1887	281k	9k	Most Read
Oscar Wilde	The Picture of Dorian Gray	1890	840k	25k	Most Shelved
Frances Hodgson Burnett	A Little Princess	1890	256k	6k	Most Shelved
Thomas Hardy	Tess of the D'Urbervilles	1891	222k	8k	Most Shelved
Charlotte Perkins Gilman	The Yellow Wall-Paper	1892	71k	3k	Most Read
H.G. Wells	The Time Machine	1895	368k	10k	Most Shelved and Most Read
Oscar Wilde	The Importance of Being Earnest	1895	280k	8k	Most Shelved and Most Read
Bram Stoker	Dracula	1897	832k	22k	Most Shelved and Most Read

Henry James	The Turn of the Screw	1898	83k	6k	Most Read
H.G. Wells	The War of the Worlds	1898	218k	7k	Most Read
Joseph Conrad	Heart of Darkness	1899	373k	12k	Most Shelved and Most Read
L. Frank Baum	The Wonderful Wizard of Oz	1900	330k	11k	Most Shelved and Most Read
James Allen	As a Man Thinketh	1902	50k	3k	Most Read
Jack London	The Call of the Wild	1903	296k	9k	Most Shelved and Most Read
J.M. Barrie	Peter Pan	1904	239k	9k	Most Shelved and Most Read
L.M. Montgomery	Anne of Green Gables	1908	645k	19k	Most Shelved and Most Read
Frances Hodgson Burnett	The Secret Garden	1910	807k	16k	Most Shelved and Most Read
Gaston Leroux	The Phantom of the Opera	1910	184k	6k	Most Shelved
Franz Kafka	The Metamorphosis	1915	541k	14k	Most Shelved

					and Most Read
Agatha Christie	The Mysterious Affair at Styles	1920	211k	6k	Most Read
Hermann Hesse	Siddhartha	1922	513k	15k	Most Shelved and Most Read
Agatha Christie	The Murder on the Links	1923	36k	2k	Most Read
Kahlil Gibran	The Prophet	1923	207k	8k	Most Read
Virginia Woolf	Mrs. Dalloway	1925	199k	8k	Most Shelved and Most Read
F. Scott Fitzgerald	The Great Gatsby	1926	3M	61k	Most Shelved and Most Read
Ernest Hemingway	The Sun Also Rises	1926	344k	12k	Most Shelved
Virginia Woolf	A Room of One's Own	1929	97k	6k	Most Read
Ernest Hemingway	A Farewell to Arms	1929	244k	9k	Most Shelved
Aldous Huxley	Brave New World	1932	1M	26k	Most Shelved and Most Read
Agatha Christie	Murder on the Orient Express	1934	310k	20k	Most Read
Margaret Mitchell	Gone with the Wind	1936	1M	18k	Most Shelved
John	Of Mice and Men	1937	2M	31k	Most

Steinbeck					Shelved and Most Read
J.R.R. Tolkien	The Hobbit or There and Back Again	1937	3M	43k	Most Shelved and Most Read
Agatha Christie	Death on the Nile	1937	100k	4k	Most Read
Zora Neale Hurston	Their Eyes Were Watching God	1937	235k	12k	Most Read
Daphne du Maurier	Rebecca	1938	398k	20k	Most Shelved and Most Read
Kathrine Kressmann Taylor	Address Unknown	1938	6k	833	Most Read
Agatha Christie	And Then There Were None	1939	667k	26k	Most Read
John Steinbeck	The Grapes of Wrath	1939	677k	15k	Most Shelved and Most Read
Albert Camus	The Stranger	1942	610k	20k	Most Shelved and Most Read
Stefan Zweig	Chess Story	1942	53k	4k	Most Read
Antoine de Saint-Exupéry	The Little Prince	1943	1M	33k	Most Shelved and Most Read
Betty Smith	A Tree Grows in Brooklyn	1943	360k	19k	Most

George Orwell	Animal Farm	1945	2M	47k	Most Shelved and Most Read
John Steinbeck	The Pearl	1945	173k	8k	Most Read
Anne Frank	The Diary of a Young Girl	1947	2M	26k	Most Shelved and Most Read
George Orwell	1984	1949	3M	60k	Most Shelved and Most Read
C.S. Lewis	The Lion, the Witch and the Wardrobe	1950	2M	19k	Most Shelved and Most Read
J.D. Salinger	The Catcher in the Rye	1951	2M	52k	Most Shelved and Most Read
C.S. Lewis	Prince Caspian	1951	315k	6k	Most Read
Ernest Hemingway	The Old Man and the Sea	1952	726k	22k	Most Shelved and Most Read
E.B. White	Charlotte's Web	1952	1M	16k	Most Shelved and Most Read
John Steinbeck	East of Eden	1952	391k	17k	Most Shelved

					and Most Read
Ray Bradbury	Fahrenheit 451	1953	1M	42k	Most Shelved and Most Read
Arthur Miller	The Crucible	1953	296k	7k	Most Shelved
J.R.R. Tolkien	The Fellowship of the Ring	1954	2M	19k	Most Shelved and Most Read
C.S. Lewis	The Horse and His Boy	1954	245k	6k	Most Read
Richard Matheson	I Am Legend	1954	81k	5k	Most Read
William Golding	Lord of the Flies	1954	2M	33k	Most Shelved
J.R.R. Tolkien	The Two Towers	1954	637k	9k	Most Shelved
C.S. Lewis	The Magician's Nephew	1955	364k	11k	Most Read
J.R.R. Tolkien	The Return of the King	1955	604k	8k	Most Shelved and Most Read
Vladimir Nabokov	Lolita	1955	603k	22k	Most Shelved
Truman Capote	Breakfast at Tiffany's	1958	196k	9k	Most Read
Chinua Achebe	Things Fall Apart	1959	258k	12k	Most Read
Harper Lee	To Kill a Mockingbird	1960	4M	84k	Most Shelved
Joseph Heller	Catch-22	1961	675k	17k	Most

Philip K. Dick	The Man in the High Castle	1962	149k	10k	Most Shelved and Most Read
Ken Kesey	One Flew Over the Cuckoo's Nest	1962	584k	10k	Most Shelved
Anthony Burgess	A Clockwork Orange	1962	546k	12k	Most Shelved
Sylvia Plath	The Bell Jar	1963	531k	20k	Most Shelved
Frank Herbert	Dune	1965	636k	17k	Most Read
Truman Capote	In Cold Blood	1965	476k	15k	Most Read
Daniel Keyes	Flowers for Algernon	1966	428k	16k	Most Read
S.E. Hinton	The Outsiders	1967	827k	30k	Most Shelved and Most Read
Gabriel García Márquez	One Hundred Years of Solitude	1967	678k	27k	Most Shelved
Philip K. Dick	Do Androids Dream of Electric Sheep?	1968	303k	12k	Most Read
Kurt Vonnegut	Slaughterhouse-Five	1969	1M	24k	Most Shelved and Most Read
Ursula K. Le Guin	The Left Hand of Darkness	1969	102k	8k	Most Read
Toni Morrison	The Bluest Eye	1970	142k	7k	Most Read
Richard Bach	Jonathan Livingston Seagull	1970	181k	7k	Most Read
Stephen King	The Shining	1977	991k	20k	Most Read

Alice Walker	The Color Purple	1982	468k	13k	Most Shelved and Most Read
Margaret Atwood	The Handmaid's Tale	1985	1M	60k	Most Shelved and Most Read
Gary Paulsen	Hatchet	1986	273k	13k	Most Read
Paulo Coelho	The Alchemist	1988	2M	70k	Most Read
Roald Dahl	Matilda	1988	538k	14k	Most Read
Laura Esquivel	Like Water for Chocolate	1989	310k	8k	Most Read
Tim O'Brien	The Things They Carried	1990	231k	13k	Most Read
Lois Lowry	The Giver	1993	2M	62k	Most Shelved
Mitch Albom	Tuesdays with Morrie	1997	702k	23k	Most Read
Mitch Albom	The Five People You Meet in Heaven	2003	535k	19k	Most Read

---

Melanie Walsh is a Postdoctoral Associate and Visiting Lecturer in Information Science at Cornell University. Her research uses computational methods to study culture, literary history, reception, and readers. Her work has appeared in *American Quarterly*, *Los Angeles Review of Books*, and *Chicago Reader.*

Maria Antoniak is a PhD candidate in Information Science at Cornell University. Her research focuses on unsupervised natural language processing methods, computational social science, online communities, affect, and subjectivity.

We would like to thank the participants of the "DH Approaches to the Arts of the Present" seminar at ASAP 2019 for early feedback on this research. We are also grateful to our reviewers and to

Richard So for their guidance. We thank Arthur Wang and Kristen Carlson for editorial work, which made this essay clearer and better. Lastly, we would like to thank the Goodreads users who are featured in our research, particularly those who agreed to be quoted in this essay. Thank you for sharing your voices.

References

1. T. S. Eliot, *What Is a Classic?: An Address Delivered Before the Virgil Society on the 16th of October, 1944* (London: Faber & Faber, 1945). [↑]
2. "Classic, Adj. And N." in *OED Online* (Oxford University Press), accessed May 19, 2020. [↑]
3. John Guillory, "Canon, Syllabus, List: A Note on the Pedagogic Imaginary," *Transition* no. 52 (1991): 36-54. See also Guillory, *Cultural Capital: The Problem of Literary Canon Formation* (Chicago: University of Chicago Press, 2013), 6. [↑]
4. Ibid. [↑]
5. Bren, "Bren's Review of Lolita," Goodreads, April 11, 2018. [↑]
6. Richard D. Altick, "From Aldine to Everyman: Cheap Reprint Series of the English Classics 1830-1906," *Studies in Bibliography* 11 (1958): 23. [↑]
7. Ankhi Mukherjee, *What Is a Classic?: Postcolonial Rewriting and Invention of the Canon* (Stanford: Stanford University Press, 2013), 9. [↑]
8. Quoted in Robert Giroux, "Introduction," in Thomas Merton, *The Seven Storey Mountain* (Boston: Houghton Mifflin Harcourt, 1998), xviii. [↑]
9. Simone Murray, *The Adaptation Industry: The Cultural Economy of Contemporary Literary Adaptation*, (New York: Routledge, 2013); Pierre Bourdieu, *The Field of Cultural Production*, ed. Randal Johnson, (New York: Columbia University Press, 1993). [↑]
10. See Mel Stanfill, *Exploiting Fandom: How the Media Industry Seeks to Manipulate Fans*, (Iowa City: University Of Iowa Press, 2019); José van Dijck, *The Culture of Connectivity: A Critical History of Social Media*, (Oxford: Oxford University Press, 2013); Jodi Dean, "Why the Net Is Not a Public Sphere," *Constellations* 10, no. 1 (2003): 95-112. [↑]
11. Aarthi Vadde, "Amateur Creativity: Contemporary Literature and the Digital Publishing Scene," *New Literary History* 48, no. 1 (2017): 27-51; Simone Murray, "Secret Agents: Algorithmic Culture, Goodreads and Datafication of the Contemporary Book World," *European Journal of Cultural Studies*, December 5, 2019; Lisa Nakamura, "'Words with Friends': Socially Networked Reading on Goodreads," *PMLA* 128, no. 1 (January 1, 2013): 238-43; Mark McGurl, "Everything and Less: Fiction in the Age of Amazon," *Modern Language Quarterly* 77, no. 3 (September 1, 2016): 447-71. [↑]
12. Karen Bourrier and Mike Thelwall, "The Social Lives of Books: Reading Victorian Literature on Goodreads," *Journal of Cultural Analytics* (February 20, 2020): 1-34; J. D. Porter, "Popularity/Prestige," *Stanford Literary Lab* Pam-

phlet 17 (2018); Alexander Manshel, Laura McGrath, and J.D. Porter, **"Who Cares about Literary Prizes?,"** *Public Books*, September 3, 2019; James F. English, Scott Enderle, and Rahul Dhakecha, **"Mining Goodreads: Literary Reception Studies at Scale,"** accessed March 5, 2019; Allison Hegel, **"Social Reading in the Digital Age"** (Ph.D. dissertation, University of California, Los Angeles, 2018); Andrew Piper and Richard Jean So, **"Study Shows Books Can Bring Republicans and Democrats Together,"** *The Guardian*, October 12, 2016. [↥]

13. Murray, "Secret Agents," 7. [↥]
14. Maria Antoniak and Melanie Walsh, ***Goodreads Scraper***, Python, 2020. [↥]
15. Vadde, "Amateur Creativity"; Melanie Micir and Aarthi Vadde, **"Obliterature: Toward an Amateur Criticism,"** *Modernism/Modernity* 25, no. 3 (2018): 517-49; Saikat Majumdar and Aarthi Vadde, eds., *The Critic as Amateur* (New York: Bloomsbury Academic, 2019). [↥]
16. Vadde, "Amateur Creativity," 27. [↥]
17. Thomas Vander Wal, **"Folksonomy,"** February 2, 2007; Arkaitz Zubiaga, Christian Körner, and Markus Strohmaier, **"Tags Vs Shelves: From Social Tagging to Social Classification,"** in *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia - HT '11* (the 22nd ACM conference, Eindhoven, The Netherlands: ACM Press, 2011), 93; Shilad Sen et al., **"Tagging, Communities, Vocabulary, Evolution,"** in *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, CSCW '06 (Banff, Alberta, Canada: Association for Computing Machinery, 2006), 181-90. [↥]
18. Allison Hegel writes illuminatingly about the evolution of genre on Goodreads in her dissertation, **"Social Reading in the Digital Age."** [↥]
19. Candace, **"Candace's Review of The Handmaid's Tale,"** Goodreads, April 17, 2017. [↥]
20. Hegel, "Social Reading," 36. [↥]
21. Jeremy Rosen, **"Literary Fiction and the Genres of Genre Fiction,"** *Post45*, August 9, 2018, our emphasis. [↥]
22. **"How It Works,"** Goodreads, February 4, 2007. [↥]
23. **"How It Works,"** Goodreads, July 24, 2017. [↥]
24. **"About Goodreads,"** Goodreads, May 2020. [↥]
25. Nakamura, "'Words with Friends'," 241. [↥]
26. Alexis C. Madrigal, **"How Netflix Reverse-Engineered Hollywood,"** *The Atlantic*, January 2, 2014. [↥]
27. Ibid. [↥]
28. According to Lisa Nakamura, Goodreads users may in fact be the ones paying: "We pay with our attention and our readerly capital, our LOLs, rankings, conversations, and insights into narrative, character, and literary tradition." Nakamura, "'Words with Friends,'" 241. [↥]
29. See Porter, "Popularity/Prestige" for other uses of Goodreads reception data. See Mark Algee-Hewitt and Mark McGurl, **"Between Canon and Corpus: Six Perspectives on 20th-Century Novels,"** January 1, 2015 for other reader-produced definitions of the canon. [↥]

30. Nakamura, "'Words with Friends,'" 240. [⤒]
31. **"Audience Insights - Goodreads,"** Quantcast, accessed May 27, 2020. [⤒]
32. We compiled data on AP English recommended authors from The Princeton Review's *Cracking the AP English Literature & Composition Exam, 2020 Edition: Practice Tests & Prep for the NEW 2020 Exam* (New York: The Princeton Review, 2019). We compiled data on college syllabi by scraping **The Open Syllabus Project**'s top texts and authors for English Literature. [⤒]
33. We use a slightly expanded version of the racial categories presented in the U.S. census: White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander, Latinx, and Middle Eastern or North African (MENA). While the U.S. census currently treats Hispanic/Latino/Spanish origin as a question of ethnicity and not race, and it currently considers the MENA population as white, we include them as separate racial categories based on advocacy from groups such as the Arab American Institute and research from the U.S. Census Bureau that suggests that incorporating Latinx and MENA might lead to more reflective racial representation. We recognize, however, that racial categories from the U.S. census, even in an expanded form, are flawed and subject to criticism. For more on Latinx and MENA as expanded racial categories, as well as the flaws and history of racial categories in the U.S. census, see Hephzibah V. Strmic-Pawl, Brandon A. Jackson, and Steve Garner, **"Race Counts: Racial and Ethnic Data on the U.S. Census and the Implications for Tracking Inequality,"** *Sociology of Race and Ethnicity* 4, no. 1 (2018): 1-13. See also The United States Census Bureau, **"About Race"**; **"2015 National Content Test: Race and Ethnicity Analysis Report."** [⤒]
34. "Audience Insights - Goodreads." [⤒]
35. Quantcast, **"Understanding Digital Audience Measurement,"** February 2, 2019, 15-16; Ruha Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code*, (Medford: Polity, 2019), 91. [⤒]
36. In promotional materials for advertisers, such as the "Book Discovery Information Kit" **sent to potential advertisers upon request**, Goodreads draws on Quantcast data to offer demographic percentages related to users' gender, age, income, and education. [⤒]
37. In constructing the Stanford Lit Lab's Twentieth-20th-Century Corpus, Algee-Hewitt and McGurl similarly found that best-of lists voted on by readers included more genre fiction than their elite counterparts — a finding corroborated by Porter in his work on Goodreads as well. Algee-Hewitt and McGurl, "Between Canon and Corpus," 6; Porter, "Popularity/Prestige." [⤒]
38. Jane Tompkins, "**Masterpiece Theater: The Politics of Hawthorne's Literary Reputation**," *American Quarterly* 36, no. 5 (1984): 618. [⤒]
39. Antoniak and Walsh, ***Goodreads Scraper***. [⤒]
40. Andrew Kachites McCallum, "**Mallet: A Machine Learning for Language Toolkit**," 2002; David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research* 3 (January 2003): 993-1022. [⤒]

41. Laure Thompson and David Mimno, **"Authorless Topic Models: Biasing Models Away from Known Structure,"** in *Proceedings of the 27th International Conference on Computational Linguistics* (COLING 2018, Santa Fe: Association for Computational Linguistics, 2018), 3903-14. [⇡]
42. Murray, "Secret Agents," 7. [⇡]
43. Murray, *The Adaption Industry*, 6. [⇡]
44. As John Guillory writes, "It is only by understanding the social function and institutional protocols of the school that we will understand how works are preserved, reproduced, and disseminated over successive generations and centuries," *Cultural Capital: The Problem of Literary Canon Formation* (University of Chicago Press, 2013), vii. [⇡]
45. Goodreads User, **"Goodreads Review of The Bluest Eye,"** Goodreads, August 26, 2012. [⇡]
46. Edward Rathke, **"Edward Rathke's Review of The Grapes of Wrath,"** Goodreads, December 13, 2018. [⇡]
47. Andrew, **"Andrew's Review of 1984,"** Goodreads, May 2, 2007. [⇡]
48. Rebecca Rego Barry, **"The Neo-Classics: (Re)Publishing the 'Great Books' in the United States in the 1990s,"** *Book History* 6, no. 1 (2003): 264. [⇡]
49. John Walsh, **"Old Book, New Look: Why the Classics Are Flying Off the Shelves,"** *The Guardian*, September 17, 2016, sec. Books. [⇡]
50. **"Cooking and Sci-Fi Are the Hot Print Segments This Year so Far,"** PublishersWeekly.com, accessed June 7, 2020. [⇡]
51. Rebecca Barry reports that the "avoidance of that incendiary word" — *canon* — "has become a theme, especially in modern classics publishing," in "The Neo-Classics," 259. [⇡]
52. Ibid., 262. [⇡]
53. Sarah Cardwell, *Adaptation Revisited: Television and the Classic Novel* (Manchester: Manchester University Press, 2002); Simone Murray observes a similar reinforcement of symbolic capital with book-to-screen adaptions, though she purposely excludes "classic" film and television adaptations from her study because "the much longer cultural histories of such texts cause them to enter the contemporary economy already freighted with approbation and/or notoriety," *The Adaptation Industry*, 21. [⇡]
54. For more Audible classics narrated by celebrities, see **https://www.audible.com/ep/BucketListListens**. [⇡]
55. Sarah Cardwell, *Adaptation Revisited*, 2. [⇡]
56. Jennifer Masterson, **"Jennifer Masterson's Review of Breakfast at Tiffany's,"** July 8, 2016. [⇡]
57. As of September 2019, J.R.R. Tolkien's *The Return of the King* (1955) and *The Two Towers* (1954) had the two highest average Goodreads ratings in the dataset with scores of 4.52/5 and 4.44/5 stars, respectively. *The Fellowship of the Ring* (1954) takes fourth place with a score of 4.35/5 stars. Nellie Andreeva, **"Amazon Sets 'the Lord of the Rings' TV Series in Mega Deal with Multi-Season Commitment,"** *Deadline* (blog), November 13, 2017. [⇡]
58. Murray, "Secret Agents," 3. [⇡]
59. Vadde, "Amateur Creativity," 28. [⇡]

60. Ibid., 29. [↑]
61. Bren, **"Bren's Review of Lolita,"** Goodreads, April 11, 2018. [↑]
62. Ibid. [↑]
63. Melanie Walsh and Bren, "Goodreads Direct Message," May 9, 2020. [↑]
64. Peter Derk, **"Peter Derk's Review of The Phantom of the Opera,"** accessed October 28, 2019. [↑]
65. Amy Bruckman, **"Studying the Amateur Artist: A Perspective on Disguising Data Collected in Human Subjects Research on the Internet,"** *Ethics and Information Technology* 4, no. 3 (September 1, 2002): 217-31. [↑]
66. Casey Fiesler Proferes Nicholas, **"'Participant' Perceptions of Twitter Research Ethics,"** *Social Media + Society*, March 10, 2018. [↑]