DramatVis Personae: Visual Text Analytics for Identifying Social Biases in Creative Writing

Md Naimul Hoque
nhoque@umd.edu
University of Maryland, College Park
College Park, MD, USA

Bhavya Ghai
bghai@cs.stonybrook.edu
Stony Brook University
Stony Brook, NY, USA

Niklas Elmqvist
elm@umd.edu
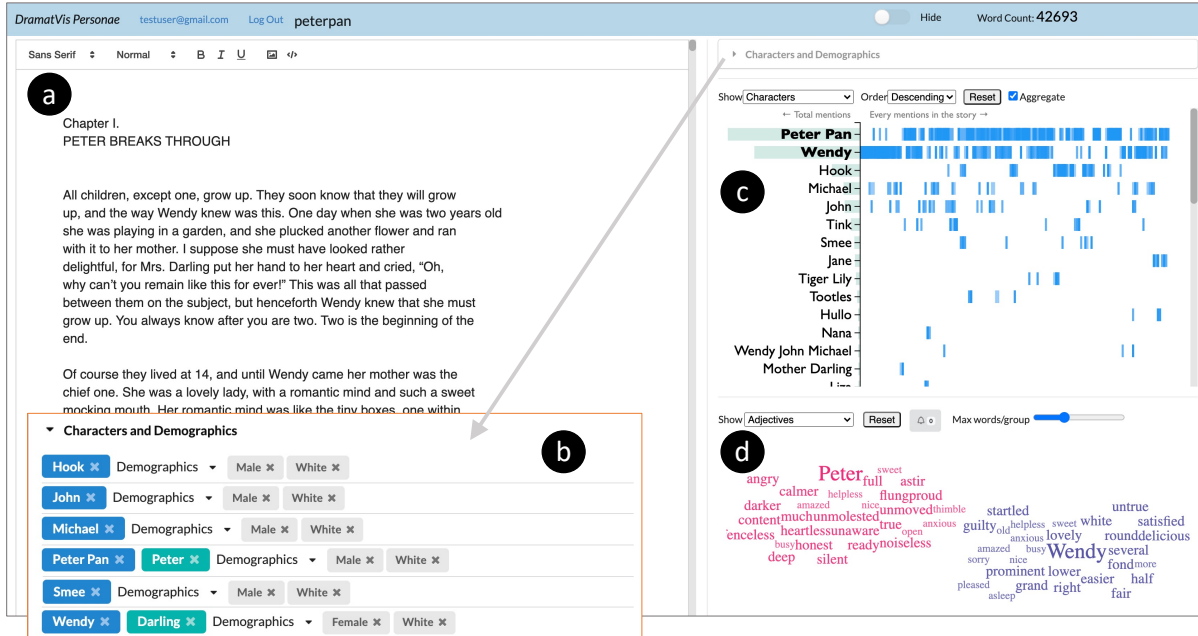University of Maryland, College Park
College Park, MD, USA

Figure 1: Overview of DramatVis Personae (DVP). (a) Rich text editor. (b) Characters and Demographics panel for listing characters identified by the system, merging aliases, and assigning social identities to characters. (c) Timeline representation of the story (*Peter Pan* by J. M. Barrie (1911)) showing every mentions of characters as well as the total number of mentions of characters. (d) Word zone [38] showing sample adjectives used for the selected characters (Peter Pan and Wendy).

ABSTRACT

Implicit biases and stereotypes are often pervasive in different forms of creative writing such as novels, screenplays, and children's books. To understand the kind of biases writers are concerned about and how they mitigate those in their writing, we conducted formative interviews with nine writers. The interviews suggested that despite a writer's best interest, tracking and managing implicit biases such as a lack of agency, supporting or submissive roles, or harmful language for characters representing marginalized groups is challenging as the story becomes longer and complicated. Based on the interviews, we developed DramatVis Personae (DVP), a visual analytics tool that allows writers to assign social identities to characters, and evaluate how characters and different intersectional social identities are represented in the story. To evaluate DVP, we first conducted think-aloud sessions with three writers and found that DVP is easy-to-use, naturally integrates into the writing process, and could potentially help writers in several critical bias identification tasks. We then conducted a follow-up user study with 11 writers and found that participants could answer questions related to bias detection more efficiently using DVP in comparison to a simple text editor.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**;
• **Social and professional topics** → **Gender**; **Race and ethnicity**.

KEYWORDS

Creative writing, bias, visualization, and NLP.

1 INTRODUCTION

Gandalf. Elizabeth Bennet. Hermione Granger and Ron Weasley. Atticus Finch. Anomander Rake, Lord of Moon's Spawn and Son of Darkness. Holly Golightly, Lisbeth Salander, and Hannibal Lecter: literature lives and dies by its characters. Heroes and anti-heroes, villains and bad guys, innocent bystanders or willing accomplices—fiction is arguably about conjuring more or less complete humans out of whole cloth and then providing audiences with the emotional release of *catharsis*, often by having these characters go through hell and high water. Therein lies also the secret of great literature: creating believable, nuanced, and multidimensional characters that spring out of the written page and come to life in the reader's mind. Achieving this is no mean feat, particularly when considering that truly great fiction often requires a diverse, inclusive, and just treatment of its cast of characters; its *dramatis personae*.

Creative writing or storytelling can also be seen as a reflection of our societal beliefs, while at the same time societal beliefs can be influenced by stories [15]. Thus, it is imperative for written stories to not promote biased representation of minority and marginalized groups. However, current literature is filled with tired, unoriginal, and sometimes harmful stereotypes related to race, gender, sexuality, ethnicity, and age, such as the trope of the angry African-American woman, the studious Asian person, or the helpless damsel in distress [6, 24, 41, 45, 50, 53, 67, 69].

Change is coming, with various organizations, institutes, writers, and the publishing community working continuously to raise awareness against biases in creative writing. For example, the Geena Davis Institute regularly publishes reports of gender and racial representation in Hollywood and foreign creative materials [67, 69]. Twitter hashtag "OwnVoices" promotes writers from marginalized groups who write about their community. Professional writers now seek feedback from "Sensibility Readers", a group of readers who especially look for harmful stereotypes before the material is published. On the computational front, Natural Language Processing (NLP) has been used to measure stereotypes in creative writing. Many of these studies have helped us understand how stereotypes operate in culture by analyzing corpus containing millions of books, a scale much larger than any previous analysis [24, 41, 66].

As a result of these efforts, writers are increasingly becoming aware of harmful biases and stereotypes, and we see reports of better representation in recent creative materials [44, 68, 70]. However, little is known about writers' current practices for addressing biases

in their writing. Given that biases often take nuanced, complicated, and intersectional forms that can be hard to detect, we speculate that computational support in this regard can help writers detect biases and write more inclusive and representative materials. Motivated by that, this paper seeks to understand creative writers' current practices for tackling biases and how computational tools can potentially help them in this regard. To inform our research, we conducted formative interviews with nine creative writers with published stories in their portfolios. The interviews revealed that writers are mostly concerned about two types of biases: (1) *Lack of agency* for minority characters (e.g., a female character introduced only to forward the plot for a male protagonist); (2) *Stereotypes* encoded in how characters are *described* and the *actions* they take in the story (e.g., a female character described as beautiful and homely). Writers mentioned that they actively look for such biases in their stories. However, these biases are often implicit and unconscious and difficult to wheedle out even for the best and most self-reflective of authors. The process is even more challenging for longer and complicated stories where many characters take intersectional identities.

Based on the findings of the formative interviews, we designed DramatVis Personae (DVP), a web-based visual analytics system to help writers identify stereotypes in creative writing. DVP is designed to integrate smoothly with the writer's own creative process, allowing them to analyze existing literature for research, upload their written content as it becomes available or even write in the tool itself, and then having its text analytics and visualizations update in real time. Using NLP methods such as entity recognition, co-reference resolution, and dependency parsing, DVP automatically detects characters in the text and collects data about them as the story progresses, including their aliases, mentions, and actions. The author can then furnish demographic information for each character, such as their age, ethnicity, gender, etc. The DVP dashboard uses this continually growing dataset to visualize the presence of characters and social identities over time.

After our initial design and implementation, we approached writers from the formative interviews and conducted think-aloud sessions using the tool. During a hands-on evaluation session conducted over videoconferencing, one writer was asked to use the tool to write a short story given a specific writing prompt. Other participants used the tool to evaluate their own existing stories. We observed their performance and then interviewed them with regard to their experience. All participants expressed positive sentiment about the DVP tool, claiming that it helped them get a better grip of their characters and their story arcs throughout the process. In particular, all participants appreciated that the tool managed and visualized character demographics, suggesting that the tool is helpful in writing a more nuanced and equitable story. We further conducted a user study with 11 participants to evaluate the effectiveness of DVP in detecting biases. The study revealed that participants could answer questions related to bias detection more efficiently using DVP in comparison to a simple text editor.

In sum, we claim the following contributions with this work: (1) findings on how to support the creative writing process by mitigating implicit bias, via an interactive interview session with nine fiction writers; (2) a visual analytics tool called DramatVis Personae (DVP) for supporting both online creative writing as

well as offline analysis of fiction; (3) results from three separate think-aloud sessions of deploying DVP with creative writers in both story generation as well as story analysis settings; and (4) results from a summative user study, outlining the effectiveness of DVP in detecting biases and stereotypes.

2 BACKGROUND AND RELATED WORK

In this paper, we focus on "creative writing," or the production of the written artifacts capturing the narrative, such as the book manuscript, fiction, or short story. Creative writing falls under the umbrella of "creative storytelling" since authors are essentially telling stories through writing. The rest of this section is designed to discuss research around creative writing, bias in creative writing, NLP, and text and literary visualization.

2.1 Bias in Creative Writing

Stereotypes in the form of art often mirror the problems, issues, thinking, and perception of different social groups in society [83, 91]. They can further reinforce biases and stereotypes against minority and marginalized groups in society [15]. The presence of biases and stereotypes, especially gender and racial bias, has been reported ubiquitously in different forms of creative writing. We provide a brief overview of research in this area below.

Many researchers have shown the prevalence of gender stereotypes in children's books, dating back to the early 1970s [26]. Since then, several studies have reported that males are often portrayed as active and dominating, while females are instead described as passive and soft [50, 65, 72]. Other studies have found the presence of racial bias [53], stereotypes against disability [6], and occupation [36] in children's books. Researchers have argued that the presence of such stereotypes in children's books is severely problematic as children are susceptible to inheriting stereotypes at an early age [50, 91]. While the situation is improving (i.e., females are portrayed with more active roles in recent children's books) due to increased social awareness, the improvement is not significant [74], and there are still reports of the prevalence of different stereotypes in children's books [2, 27, 36].

Another form of creative writing medium that has been heavily criticized for promoting stereotypes is movie scripts. The *Geena Davis Institute* regularly publishes reports of gender and racial representation in Hollywood movies and is a valuable resource for current representational problems. The institute has found underrepresentation and misrepresentation of females [67] and Black or African American females in Hollywood [69]. Beyond Hollywood, researchers have found similar sorts of biases in television shows and movies in other countries. Emons et al. [21] found stereotypes in gender roles of males and females in U.S.-produced Dutch TV shows, misrepresenting females in Dutch society. Madaan et al. [58] has shown the existence of gender biases in Bollywood movie scripts. Similar to movie scripts, many studies have shown how biases and stereotypes operate implicitly in news articles and how they adversely affect the audience [15, 22, 79, 92].

Finally, newer forms of writing such as blogs, online writeups, and social media posts are rife with harmful stereotypes. Fast et al. [24] analyzed fiction written by novice writers in the online community Wattpad and found it to be rampant with common gender stereotypes. Joseph et al. [45] analyzed forty-five thousand Twitter users who actively tweeted about the Michael Brown and Eric Garner tragedies. Their method can quantify semantic relations between social identities. Other work discussed the impact of stereotypes in Reddit [25], Facebook [60], and U.S. history books [57].

All the above-mentioned research has been instrumental in raising awareness among writers, directors, and the general audience, a critical step towards equality. As a result, there are reports of better representation and inclusivity in recent years [44, 68, 70]. Originating on Twitter in 2015, "OwnVoices" has become a campaign for promoting writers from diverse backgrounds writing about their experiences and cultures. We strongly believe that empowering writers from historically excluded groups is tremendously important for our society. At the same time, we believe writers from all backgrounds need to be cautious when writing creative materials that represent a culture or social identity and invest efforts to learn about relevant communities for correct portrayals. In fact, we believe our tool will be most useful for writers who want to ensure that they represent other communities correctly and do not propagate stereotypes. We intentionally avoided identifying a specific user group for our tool as we acknowledge that identities are intersectional, meaning a person who is considered privileged in one dimension of their identity, may well be a minority in another dimension. Rather, we strove to include writers from diverse backgrounds in the design of this tool. We believe our tool can assist in the production of more inclusive and representative writings that will have a positive impact on society.

2.2 Computational Support for Creative Writing

Beyond the everyday-use text processors such as Microsoft Word, there are several professional and open-source software available to writers for helping them in guiding character development. Scrivener [55] is a paid service for organizing stories. It allows flexible page breaking, adding synopsis and notes to each section, and easy merging or swapping between sections. It also has a distraction-free writing mode where everything else on the computer is tuned out. Granthika [34] is a similar sort of paid service that helps writers in tracking characters and events in a story. It lets users integrate knowledge into the system as they write, and then use that knowledge for tracking in a timeline. It also allows writers to apply causal constraints to the events and people in a story (e.g, "The inquest must happen after the murder"). Grammarly is used for checking grammatical errors.

Over the years, NLP research has developed and refined different techniques such as coreference resolution, named entity recognition (NER), dependency parsing, sentiment analysis, etc, which can be instrumental in analyzing stories. For example, NER helps extract named entities such as person names, organizations, locations, etc from the text. In our context, this can help identify different characters of a story [18]. Dependency parsing finds relationships between words. This can help identify different adjectives/verbs linked to a character in a story [64]. Similarly, coreference resolution can help track the representation of different characters and their demographic groups across the storyline [51]. Other studies in the NLP literature have specifically focused on analyzing stories. This

includes segmenting stories by predicting chapter boundaries [75], recognizing flow of time in a story [48], analyzing emotional arc of a story [80], extracting character networks from novels [52], etc. Using such methods, researchers have conducted large-scale analysis of books, and stories [24, 41, 45, 52, 66, 73].

Finally, the invent of powerful generative language models such as GPT-3 [9] have fueled research on human-AI collaboration for creative writing. Many tools in this area generate stories iteratively: a writer provides an initial prompt for the story; AI generates a section of the story automatically; the writer edits the generated story and provides further prompts and so on [12, 13, 54]. Recently, Chung et al. [11] proposed TaleBrush, an interactive tool where AI can generate a narrative arc based on a sketch drawn by a writer that outlines the expected changes in the fortunes of characters. Other works include IntroAssist [43], a web-based tool that helps entrepreneurs to write introductory help-requests to potential clients, investors, and stakeholders; INJECT [59], a tool for supporting journalists explore new creative angles for their stories under development; and an interactive tool proposed by Sterman et al. [87] that allows writers to interact with literary style of an article.

While these works provide necessary background for the technical design of our tool, none of these works have features to add social identities to characters and investigate potential biases. Additionally, prior works have primarily used textual descriptions for summarizing and communication. While that is helpful, we utilize data visualization, a visual communication medium, for gathering insights from the information extracted by the tool efficiently. .

2.3 Visualization for Text and Literature

Harking back to some of the original approaches to visualizing "non-visual" text documents [95], data visualization has long been proposed as an alternative to reading through large document corpora [3, 29, 85]. For example, the investigative analytics tool Jigsaw [86] is often styled as a "visual index" into a document collection; while it is not a replacement for reading, it provides a linked collection of entities and documents for easy overview and navigation. This is generally also true for document and text visualization as a whole; the goal is to be able to "see beyond" the raw text into content, structure, and semantics [3].

Beyond simplistic text visualization techniques such as word clouds, data visualization can become particularly powerful when applied to entire documents [29]. These ideas can also be used for literary analysis of fiction and poetry [16]. For example, Rohrer et al. [81] used implicit 3D surfaces to show similarities between documents, such as the work of William Shakespeare. Similarly, Keim and Oelke propose a visual fingerprinting method for performing comparative literature analysis [47]. McCurdy [61] present a organic linked visualization approach to scaffolding close reading of poetry. The literary tool Myopia [10] uses color-coded entities to show the literary attributes of a poem for readers. Abdul-Rahman et al. [1] also apply data visualization to poetry. Our work here is inspired by, if not the design and implementation, then at least the motivation of these tools; however, in comparison, our goal is to support the creative processing while focusing on identifying and mitigating implicit social bias.

Finally, visualization can also be applied to the stories themselves rather than the actual text. XKCD #657,[1] titled *Movie Narrative Charts*, shows temporal representations of plotlines in five movies, including the original *Star Wars* trilogy (1977–1983), *Jurassic Park* (1993), and the complete *Lord of the Rings* movie trilogy (2001–2003). Liu et al. [56] propose an automated approach to generating such storyline visualizations called StoryFlow. Tanahashi and Ma discuss design considerations for best utilizing storylines [88]. Some effort has also been directed towards minimizing crossings in storyline visualizations [35]. VizStory [42] takes a very literal approach to visualize stories by identifying segments and themes and then searching the web for appropriate representative images. TextFlow [17] and ThemeDelta [28] and related topic modeling visualizations can be used to automatically extract and visualize evolving themes in a document (or document collection) over time. StoryPrint [94] shows polar representations of movie scripts and screenplays in a fashion similar to Keim and Oelke's fingerprints; the authors note that this approach could also be used to support the creative writing process. StoryCurves [49] shows non-linear narratives in movies on a timeline representation. Finally, Story Analyzer [64] shows several visualizations representing summary statistics of a story. Common for all of these storylines and plot visualization tools is that they are designed mostly for retrospective analysis and not for online creative writing. None of these tools support bias identification in an interactive environment. Nevertheless, we draw on all of these tools in our design of DVP.

3 FORMATIVE STUDY

To understand creative writers' current practices and challenges for addressing biases, we conducted semi-structured interviews with 9 creative writers. The study was approved by our university's Institutional Review Board (IRB).

3.1 Participants

We recruited 9 creative writers through advertisements to social media such as Twitter and Facebook, local mailing lists, and university mailing lists. Our inclusion criteria included prior experiences in creative writing such as novels, short stories, screenplays, etc, and familiarity with writing in text editors. All participants had published materials in their portfolios. Table 1 presents participants' demographic information. Each participant received an Amazon gift card worth $15 for their participation.

3.2 Procedure

We conducted the interviews over Zoom. Each interview lasted around 1 hour and was divided into three parts. First, after gathering informed consent, we asked the writers to share their perspectives on bias in creative writing. In the second part, we asked the writers about the challenges they faced in addressing biases in their writing, and their current approach for overcoming biases. Finally, writers brainstormed with the study administrator for outlining the potentials and requirements for digital tools that might help them in managing biases and stereotypes. We provide the questionnaire for the interviews as a supplement.

---

[1]https://xkcd.com/657/

Table 1: Participant demographics.

Id	Gender	Race	Age	Expertise	Yrs Exp
W1	Male	Asian	30	Short stories, poems, and blogs	10
W2	Male	Asian	27	Short stories, poems, and critiques	12
W3	Female	White	26	Novels (fiction/non-fiction), short stories, screenplays, poems, blogs, critiques, and fanfiction	15
W4	Female	White	49	Novels (fiction/non-fiction), and roleplaying games	25
W5	Female	Black or African-American	44	Picture books and books for beginning readers	6
W6	Non-binary	Prefer not to respond	36	Novels (fiction/non-fiction), short stories, and poems	10
W7	Female	Asian	25	Short stories and poems	15
W8	Male	Asian	25	Screenplays, blogs, and critiques	12
W9	Female	White	34	Screenplays and poems	15

3.3 Analysis

We created anonymized transcript for each interview from the recorded audio. Two authors of this paper open-coded the transcripts independently. A code was generated by summarizing relevant phrases or sentences from the transcripts with a short descriptive text. Both coders then conducted a thematic analysis [8] to group related codes into themes. Throughout this process, the coders refined the themes and codes over multiple meetings by discussing disagreements and adjusting boundaries, scopes, and descriptions of the codes and themes. The open codes and themes were also regularly discussed with the full research group. We present the findings from the interviews next.

3.4 Findings

Our findings relate to the topics of implicit bias in creative writing, mitigating bias, and the potential for computational support for such issues.

3.4.1 Current State of Biases in Creative Writing. The publishing and writing community have progressed towards an inclusive and representative environment; however bias in creative writing is still a present and complicated problem (W1-W9). W3 mentioned that publishers now encourage writers to seek feedback from "Sensitivity Readers" for potential misrepresentation. W4 lauded the recent Twitter hashtag "#ownvoices" that started the discussion about the importance of writers from marginalized groups writing about their community. Despite these efforts, there is still a need for diverse writers while writers from all communities need to be conscious about how their writing represents different social groups (W1-W9). W5 said: *"One recent study[2] found that there were more characters that feature animals than there were of all of the minorities. I mean, it is horrendous, to be honest with you, in 2021... and it is really sad."* W7 emphasized the importance of research (e.g., reading literature, understanding culture) before writing about characters that represent marginalized groups.

3.4.2 Bias 1: Lack of agency for minority characters. A critical challenge for writers is to ensure that characters representing minority groups have impact in the story and are not sidelined. W3 said: *"I believe it is Gail Simone. She is a comic book writer and she started*

___
[2]https://www.theguardian.com/books/2020/nov/11/childrens-books-eight-times-as-likely-to-feature-animal-main-characters-than-bame-people

this website called "Women in Refrigerators" because there is a very famous storyline from a Green Lantern comic (vol. 3, #54) where the Green Lantern comes home to find that his girlfriend has been murdered and shoved in his refrigerator. And the only reason that the woman existed was to be killed. So it has become a name for this type of killing of female characters to advance the man's plot, which is unfortunately very common." W3 also referred to the *sexy lamp test*, which tests if a female character of a story is replaced with a sexy lamp whether the story still makes sense. W4 mentioned the *Bechdel* test, which asks whether a story features at least two women who talk to each other about something other than a man.

To address this type of bias, writers employ several self-evaluation techniques. They mentally track the presence of characters as well as their social identities in the story (**N1|W3-7, W9**). This frequently requires them to go back and forth between different parts of the story and read through them (**N2|W3-7, W9**). While reading, writers investigate interactions between characters (**N3|W3-7**), and how actions of the characters reflect on the social identities they represent (**N4|W3-5, W7**). However, this process becomes challenging and tiresome as the stories become longer and complicated (W3-7, W9). Moreover, the identities often take intersectional form (e.g., a Female African-American character) which makes it even more mentally demanding to track them. It is worth noting here that writers also apply these methods to analyze existing literature as part of their research; although often that process is not as thorough as correcting their own work and depends mostly on high-level subjective understanding (W2-4, W7-9). Such research helps them understand representation of different social identities as portrayed in existing literature (W2-4, W7-9).

3.4.3 Bias 2: Stereotypes in describing characters. Another form of bias is how characters are described in the story. For example, W1 mentioned that female characters are often described as homely, beautiful, and sacrificial. W7 said: *"One example would be Truman Capote's Breakfast at Tiffany's. I read the book as a teenager, and it's very beautiful and well-written overall. But one example of a stereotyped character is Holly's neighbor Mr. Yunioshi. He is a side character but is portrayed simply as the irritable neighbor who always tells Holly off for forgetting her keys. This depiction of East Asian characters as grumpy old people is very stereotypical and common in Western literature."*

To address this type of bias, writers remain careful while writing and examine the actions (**N4 from above| W1-4, W9**) and descriptors (**N5| W1-3, W7-9**) they use for the characters and social identities. This again requires them to go back and forth between different parts to critically read the story (**N2 from above| W1-4, W7-9**), and constantly check the actions and descriptors. Writers employ similar evaluation techniques for analyzing existing literature (W2-4, W9).

3.4.4 Potential and recommendations for an interactive tool. All participants were enthusiastic for an analytic tool for identifying potential stereotypes. Critique groups and sensitivity readers provide important feedback to writers; however, they are usually available at the advanced stage of a formal publication. During writing, a tool supporting their self-reflective process would be helpful. The tool will also be helpful to a writer to analyze existing literature as a part of their research (W2-4, W7-9).

Writers also provided a few recommendation for such a tool. First, writers suggested that the tool could support story writing since the various bias mitigation strategies discussed in Section 3.4.2 and 3.4.3 are closely integrated into the writing cycle (**N6| W2, W4, W8**). Second, the tool could have a "distraction-free" mode, similar to Scrivener, where writers could concentrate only on writing if they want to (W2-5, W8). Finally, writers suggested that the tool should **avoid suggesting interpretations of its own**; rather the analytic tool should support writers in **exploring their work** so that they can make informed decisions. Writers provided several reasons behind this suggestion. W1 said: *"Language around different social groups is fluid and continuously evolving. What is acceptable today, may not be acceptable a few months from now."* W6 said: *"Creative writing can be subjective to a writer's own experience, or imagination, and may elicit intended stereotypes for the plot. I want to write about discrimination against my community which if flagged by a tool would be disappointing"*

4 DESIGN GUIDELINES

Here we outline and discuss the design guidelines and decisions that originated from the literature and formative study.

DG1. Nature of the Support: Exploratory. The first design decision we made is the nature of the support. The formative study suggested several challenges that writers face in identifying biases and stereotypes. They apply several evaluation techniques (Table 2) that can be mentally demanding. All participants were enthusiastic about a tool that will support them in this process. However, based on the findings from Section 3.4.4, it was clear that the tool should avoid flagging any writing as inappropriate and suggesting how to reduce biases. Suggesting interventions for reducing biases without a proper understanding of a writer's goal could seriously lower user trust in the system and be counter-productive. Another potential concern is writers using such suggestions to write about cultures and identities outside their own identity without really investing efforts to learn about them. Considering all these, we decided that the primary goal of the tool would be to **help writers perform N1 to N6 easily, without any explicit recommendation**. We decided to use interactive visualization for this purpose which is an

Table 2: Design needs identified from the formative study, based on the evaluation techniques performed by writers to identify biases.

Design Need	Purpose	Participant
N1. Evaluate presence of characters and social identities.	Bias 1	W3-7, W9
N2. Move between different parts of the story for reading.	General	W1-9
N3. Evaluate interactions between characters-characters and social identities-social identities.	Bias 1	W3-7
N4. Evaluate and compare actions of characters and social identities.	Bias 1, Bias 2	W1-7, W9
N5. Evaluate and compare descriptions of characters and social identities.	Bias 2	W1-3, W7-9
N6. Support writing in the tool.	General	W2, W4, W8

effective way to summarize any form of abstract data and enable exploration in an interactive environment.

DG2: Help writers evaluate agency for characters and social identities. The formative interviews revealed that writers are concerned about the lack of agency for characters that represent minority groups. To ensure agency for characters, writers practice three evaluation techniques (**N1, N3, N4** from Table 2). Prior research has also used presence, and interactions between characters to quantify character agency [40, 70]. Thus, our tool should support writers in these tasks.

DG3: Help writers evaluate stereotypes in describing characters. Another important concern raised during the formative interviews was the stereotypes used for characters. Writers mentioned that they search for possible stereotypes in the action characters take in the story (verbs, **N4**), and the words that describe the characters (adjectives, **N5**). Previous research has also used verbs and adjectives to quantify biases and stereotypes [21, 24, 30, 41]. Thus, our tool should help writers in these tasks.

DG4: Support writing and critical reading. **N2** and **N6** are not directly related to bias identification. However, the formative study suggests support for these features is necessary to facilitate bias identification. They will also enable bias identification in different stages of writing a story. For example, during the pre-writing (e.g., analyzing existing literature) or post-writing stage, a writer can use the support for reading (**N2**) and other bias identification methods for critical reading. During the writing stage, writers will require support for both writing (**N6**) and reading (**N2**).

We also decided to include a "distraction-free" mode similar to Scrivener [55] that will allow writers to write or read without any distraction and use our bias identification support only when they want to. This will ensure that our tool does not obstruct the creative process for writers.

DG5: Inclusive design. Prior technological efforts in bias identification and mitigation often focused on a specific type of bias, or a binary view on identity (e.g., male, female as gender category) [30, 40, 41]. However, the formative interviews suggest that writers may use diverse and intersectional identities for characters that may not conform to any pre-defined categories. We anticipated that the tool would be demoralizing if a social identity that an author is writing about is not supported in the tool; especially since authors may themselves share that identity. Thus, our design should be inclusive, and support social identities of any kind.

DG6: Easy-to-understand visualization and scalable computation. To ensure accessibility for writers who are not familiar with complex data visualization paradigms, our tool should use easy-to-understand visualizations. Finally, the formative study suggests the tool would be most useful as the story becomes larger and complicated. Thus, our tool should support stories of large size and provide feedback to the writers in a short response time. Similarly, the visual components should be able to show information extracted from large textual data.

5 DRAMATVIS PERSONAE (DVP)

We present the visual components and analysis pipeline for the DVP tool below.

5.1 Visual Interface

In this section, we demonstrate four visual components (Figure 1) of the interface. We link the design guidelines from Section 4 wherever applicable. We also discuss design rationales and design alternatives considered during the development of the tool. The interface supports analysis for three different types of entity: (a) Characters, (b) Social Identities (e.g., Male, Female), and (c) Intersectional Social Identities (e.g., Muslim Males). We use the term "entity" or "entities" to refer to the three entities together whenever applicable for brevity. For demonstration purpose, we use several well-known western stories.

Text Editor. The central component of DVP is a text editor (**DG4**). The text editor is equipped with traditional formatting features such as selecting fonts, font sizes, font-weights, etc. We use QuillJS [78] as a rich text editor which has been widely used on the web including social media apps such as Slack, and Linkedin.

The visualizations are contained in a sidebar; a user can hide or show the sidebar by clicking the *"Hide"* toggle button (**DG4**). This ensures that writers can see the visualizations whenever they want, but also can concentrate on writing and reading by hiding the sidebar whenever they want.

Characters and Demographics Panel. The Characters and Demographics Panel lists all the named entities identified by our NLP pipeline. The panel supports several validation functionalities. First, a user can delete a character from the list in the case they do not wish to track that character, or if the character was wrongly identified by the NLP pipeline (e.g., an institute identified as a person). Second, a user can merge different names of the same character together (e.g., merging Peter Parker, and Spiderman as one character which might be identified as separate characters by the NLP pipeline). A user can use drag and drop for merging characters.

Note that these validation functions are not necessarily a part of a writer's work process, but are needed to use the tool reliably given the nature of current NLP toolkits.

Each character in this panel has a dropdown named *Demographics*. Using this dropdown, a user can add multiple social identities to each character (Figure 2a). We populate the dropdowns with commonly used identities. However, to support **DG5**, we made the dropdowns dynamically extendable. A user can add any number of new identities in these dropdowns. For example, Figure 2b and 2c show how a user can add Profession as a new identity and Doctor as a profession in the dropdowns.

Timeline. To support **DG2**, we designed a timeline representation of the story (see Fig.1 (C)). The timeline is divided into two parts. On the left side of the y-axis, we encode the total number of mentions for entities using bars encompassing the axis labels. The y-axis can be sorted either in descending or ascending order. A user can choose the sort order from the *Order* dropdown. The dropdown defaults to descending order for showing the prominent entities at the top. On the right side of the y-axis, we show individual mentions for each character. The x-axis represents a linear scale with a range $(1, S)$ where S is the total number of sentences. For a mention of an entity in sentence s, we draw a tile (rectangle) with width $(pos(s) + 0.5) - (pos(s) - 0.5)$ where $pos(s)$ represents the position of s in the x-axis. We used linear scale instead of ordinal scale to make the adjacent tiles connected and smoother.

For larger documents, due to space constraints, the tiles can become extremely small, making it difficult to interact with them. To scale the visualization to larger documents (**DG6**), we added an *Aggregate* option in the toolbar of the timeline (Figure 3a). In the case of a document with more than 500 sentences ($S > 500$), the option is automatically triggered, and the document is binned together to restrict the x-axis to range $(1, 500)$. At that point, the x-axis represents passages, instead of sentences. The user can disable the aggregate option and see the timeline in terms of a single sentence. As an alternative, we considered making the x-axis scrollable. The tile size would have remained the same for any documents in that design. However, we noticed that this design does not provide a full overview of the timeline together and it becomes difficult to compare different parts of the timeline as the user scrolls left and right in the timeline.

Figure 3a shows an aggregated version of the timeline for the children's book, *Sleeping Beauty*, retold by Arthur Quiller-Couch and Charles Perrault, freely available under Project Gutenberg [5]. The opacity of a tile represents the number of times an entity was mentioned in a particular passage. A user can hover over any tile and see the relevant passage, highlighted in the text editor (Figure 3b). This facilitates easily going back and forth between different parts of the story (**DG4**).

Based on the identities defined by a user in the Characters and Demographics Panel, the timeline can also show the presence of different identities in the timeline (Figure 3c). Using the *Show* dropdown in the timeline (Figure 3a), a user can choose to show the timeline for characters or demographics. A user can also choose to show the presence of intersectional identities in the timeline. For example, Figure 4a presents a timeline showing the presence of intersectional groups in the Movie *The Amazing Spiderman-2*. Note
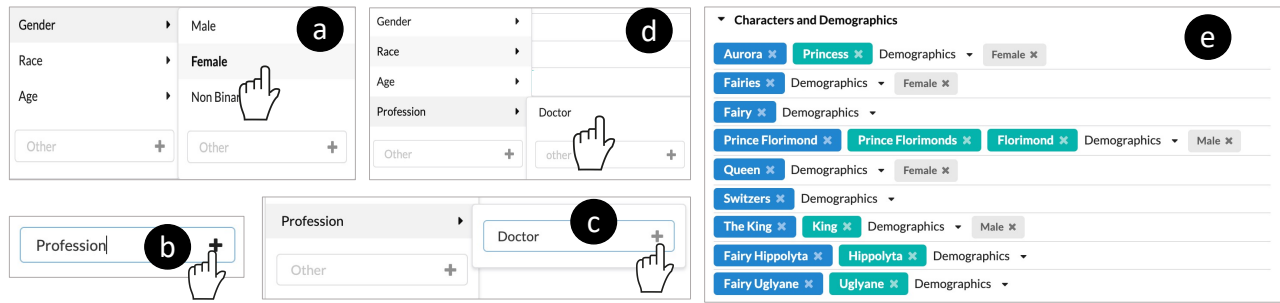
Figure 2: Characters and Demographics panel. (a) A dropdown menu with different social identities is available for each character in the Characters and Demographics Panel. (b) The dropdown can be extended dynamically. User can add a new identity (e.g., Profession). (c) User can then add categories under the newly added identity (e.g., Doctor). (d) The new Demographics dropdown with Doctor as a new profession. (e) An example demographics panel for the children's book, *Sleeping Beauty*, retold by Arthur Quiller-Couch and Charles Perrault, freely available under Project Gutenberg [5].
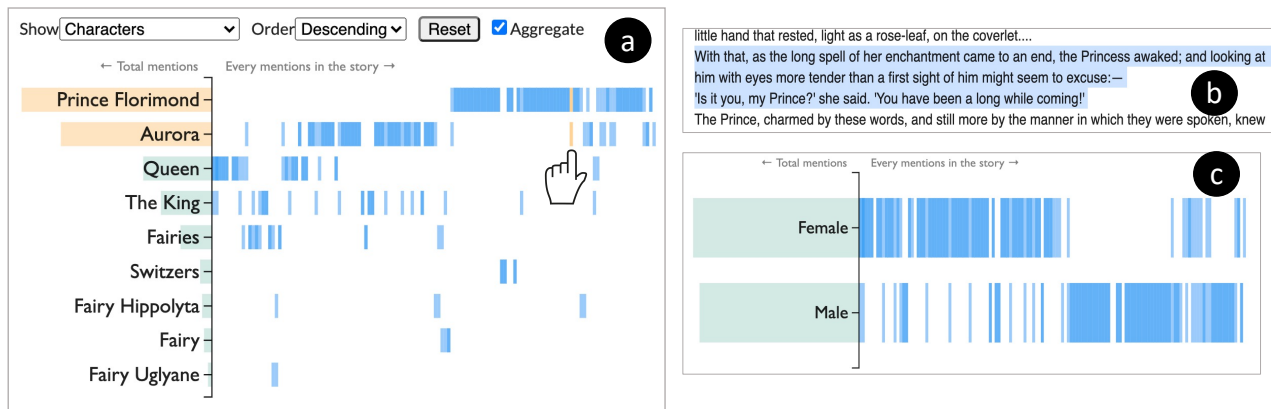

Figure 3: Timeline interface. (a) An example timeline based on the characters of *Sleeping Beauty*. Note that even though the story is about Aurora, the sleeping beauty, it was Prince Florimond who had the most number of mentions. Note also the lack of overlaps between the timelines of two characters. (b) Upon hovering over a tile from the timeline, a user can see the relevant passage in the editor. The relevant passage in this case is the first scene after the Princess wakes up. (c) A user can also visualize the presence of different social identities in the timeline. We only used Male and Female for this example. Note that when aggregated, the female presence in the story is slightly larger than the male presence.

that for brevity, we do not show the Characters and Demographics Panel for this movie. For the relevant Characters and Demographics Panel, please refer to the supplemental video. We also note that the race assigned for the characters of this movie is based on the perception of the authors and may be different in reality.

In summary, the goal of the timeline visualization is to facilitate writers investigating agency for characters and social identities (**DG2**). Since the task is subjective to a story and a writer's perspective, we intended to expose potential gaps in the story and let a writer explore and navigate the story easily (**DG1**). We decided that visualizing the mentions of the characters in a timeline can be a starting point for this task and will allow users to see gaps easily. The user can then use the timeline to further investigate any part of the story. To aid this process, other visual components are also connected to the timeline, as described below.

Impact Graph. The timeline visualization primarily shows the presence of entities in a story. While a user can identify the interactions between entities by observing overlaps in their presence in the x-axis, the timeline does not give a definitive answer to how an entity interacted with the other entities. Additionally, multiple entities can be mentioned in a passage of a story; however, that does not necessarily mean they interacted with each other.

To overcome this shortcoming, we introduced the *Impact Graph*, a force-directed network visualization that shows interactions between entities (**DG2**). We consider an interaction between two entities if they are mentioned together in a sentence. The impact graph for an entity is available whenever a user clicks on a y-axis label in the timeline.

Figure 4b-d show impact graphs for Peter Parker, Harry Osborne, Max Dillon, and Gwen Stacy from the movie *The Amazing Spider-man 2*. The selected entity is placed at the center of the network. The
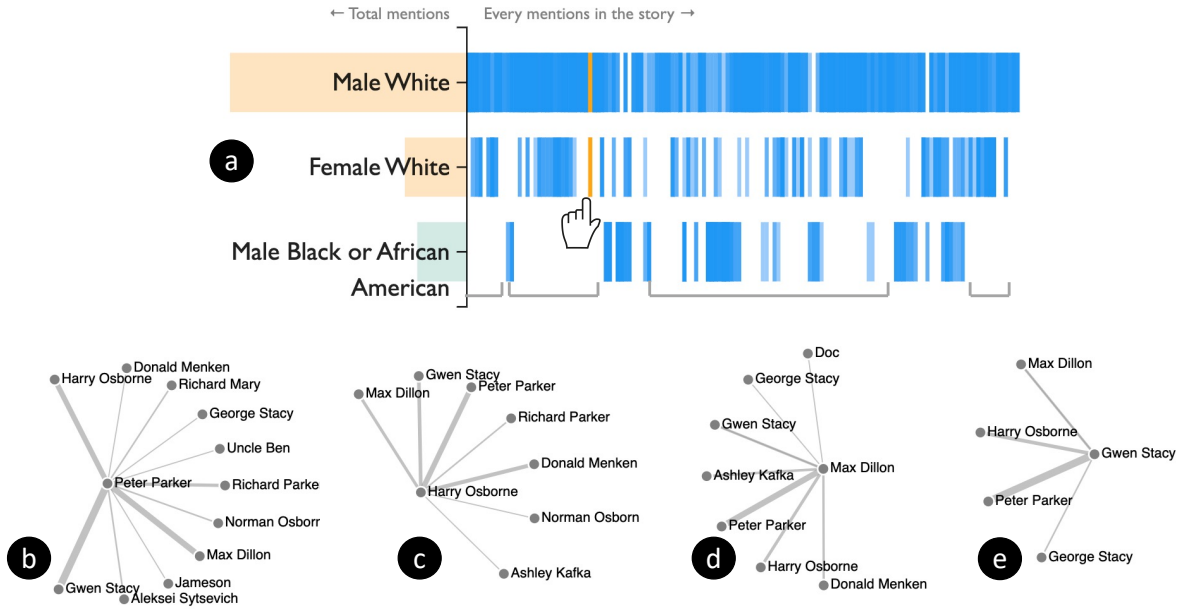
Figure 4: Timeline interface and impact graphs. (a) An example timeline showing the presence of three intersectional identities (*Male White, Male Black or African American,* and *Female White*) in the Movie *The Amazing Spiderman 2* (2014). Male White characters are present throughout the storyline. Further, there appears to be a lack of interactions (identified by the grey encompassing x-axis lines) between Male Black or African-American and Female White characters except for a few aberrant ones. However, both groups have interactions with Male White characters. The orange bars show one such interaction between Male White and Female White characters. (b-e) Example impact graphs for the characters Peter Parker, Harry Osborne, Max Dillon, and Gwen Stacy.

edge widths represent the strength of the interactions. For reducing clutter, only edges with at least five interactions are shown. We observe that the impact graph for Gwen Stacy is much smaller than the other three male characters. Further, Gwen Stacy has significant interactions with Peter Parker only.

Word Zones. Word cloud is a popular visual representation for showing a collection of words. They have aesthetic value to lay users and are fun, and engaging [38, 93]. In contrast, researchers have shown that they are not well-suited for analytic tasks such as finding a word and comparing the frequencies of words [38]. To balance the utility and aesthetic value of word clouds, Hearst et al. [38] proposed Word Zones, a variation of word clouds, where words are grouped based on predefined labels/categories. Since our users will most likely be non-experts in terms of visualization expertise, we decided to use Word Zones, thus opting for a visualization that is expected to be well-known to the writers as well as has better representation for analytic purposes (**DG6**).

A writer can add an entity for seeing words used with the entity in the word zone visualization whenever a user clicks on a y-axis label in the timeline (**DG3**). A writer can add as many entities as they want in the word zone. A user can control the number of words to show for each entity in the word zone using a slider. Users also have a dropdown to see relevant adjectives, verbs, or both. The

verbs and adjectives for characters are extracted using dependency parsing. They are then aggregated for the social identities. Overall, this mechanism helps in answering questions such as: *How the female characters were described in the story? What were the actions of female characters in the story?* (**DG3**).

We considered each entity (a character or identity) as a document and the full story as a corpus of documents (entities). Based on that, the weight of a word (w) for an entity (e) is calculated as:

$$weight(w, e) = tf(w, e) * (1/df(w)) \tag{1}$$

where $tf(w, e)$ is the frequency of w in e and $df(w)$ is the frequency of w in the whole story. It is essentially a normalized version of *tf-idf*, popularly used for filtering out common and stop words, finding words of interest. However, *tf-idf* is often applied on large corpora. In our case, the number of entities is limited to at most a few hundred. We also consider adjectives and verbs only. Thus, we opted for a simple normalization.

Figure 5 shows an example word zone showing adjective used for Dolly (a female character), and Vronsky (a male character) from *Anna Karenina* (1877) by Leo Tolstoy. Note the words such as "charming", "envious", "jealous", "helpless", "oblivious" and others in Dolly's word zone.

Figure 5: Word Zone representation for Dolly and Vronsky from *Anna Karenina* (1877) by Leo Tolstoy, publicly available under Project Gutenberg [89].

5.2 Information Extraction from Text

We used NeuralCoref [23] along with Spacy [39] for extracting named entities, and their mentions in a document (Coreference Resolution [46]). Both packages are considered state-of-the-art and widely used in different applications. However, coreference resolution for long documents such as a book can take a lot of time and system memory [90]. This could hinder the usability of the tool as our tool is targeted as a writing tool that provides instant feedback to writers as they write (**DG6**).

To increase the scalability of computation and reduce processing time, we adopted a divide-and-conquer method and run the coreference resolution model paragraph-wise, instead of the full document together. We noticed that it is unlikely that a character would be referenced with a pronoun without the actual noun of the character in a new paragraph. We reached out to the participants from the interview studies and confirmed that this is the usual case. Figure 6 presents an example scenario of our method. On detecting a new trigger symbol such as ".", "?", "!" or a copy/paste event, the UI sends the document and the current Delta [77] object to the server. The Delta object contains information about how many and what characters (symbols) are inserted, deleted, or retained since the last update.

Upon receiving the document, the server splits the document into paragraphs. The paragraphs are identified by double newlines or "\n\n". Based on the Delta object, the server then determines which paragraphs are retained (same as the previous update), and which paragraphs have new contents (insert or delete). In Figure 6, the server detects three new paragraphs (grey boxes). The server then runs entity recognition and coreference resolution models on the newly detected paragraphs. Although our current implementation processes the paragraphs sequentially, they can be processed in a parallel fashion since the paragraphs are mutually independent. The information extracted from the newly added paragraphs

is then aggregated together with the stored information of previously processed paragraphs. The client-side then performs another aggregation to combine aliases.

5.3 Implementation Notes

DVP is a web-based writing tool. We used Python as the back-end, JavaScript as the front-end language, and D3 [7] for interactive visualization. Semantic-UI and Bootstrap were used for styling various visual objects. We used Spacy [39] and NeuralCoref [23] for all the NLP tasks. The tool is available here: https://github.com/tonmoycsedu/DramatVis-Personae.

6 EVALUATION

We evaluated DVP in two parts. First, we conducted think-aloud sessions with three writers to understand the potential of DVP as a writing tool and identify potential usability issues. We then conducted a user study with 11 writers to identify the benefits of DVP in comparison to a simple text editor. We describe each study in turn below.

6.1 Think-aloud Sessions

For this study, we invited 3 writers to test our tool. Two of them participated in our formative study (W2 and W5 from Table 1). We refer to them as W1 and W2 in the following section. The other writer (W3) did not participate in the formative study. W3's self-reported demographics are male, white, 44 years of age, with more than 20 years of writing experience.

Before the start of the study sessions, we emailed each writer independently asking whether they would like to write a new short story using our tool, or they would like to analyze stories that have already been written by them. W1 and W2 wished to analyze previously written stories while W3 wished to write a story using our tool. We asked writers that their stories should have at least three significant (named) characters and any number of supporting characters, at least two different scenes, and some dialogue between the characters. Additionally, we encouraged W3 to ideate and take notes on their story and characters, but asked not to begin the actual writing process until the session.

One author of this paper administrated the sessions. After consent, the administrator demonstrated the tool with a sample story. We encouraged participants to ask questions at this stage. Once participants felt comfortable with the tool, we asked them to analyze a story (W1 and W2) or generate a new short story (W3). While using the interface, participants thought aloud and conversed with the administrator regularly. Each session ended with an exit interview. The sessions with W1 and W2 took around 1 hour while the session with W3 lasted around 2 hours. All sessions were conducted using Zoom video conferencing and were audio and video recorded for post-session analysis. At the end of the interview, each participant received a $15/hour worth Amazon gift card. The study administrator and another author of this paper independently analyzed the recorded videos, and study notes and then met together to sort the feedback from the writers and our observations into the following five thematic categories:
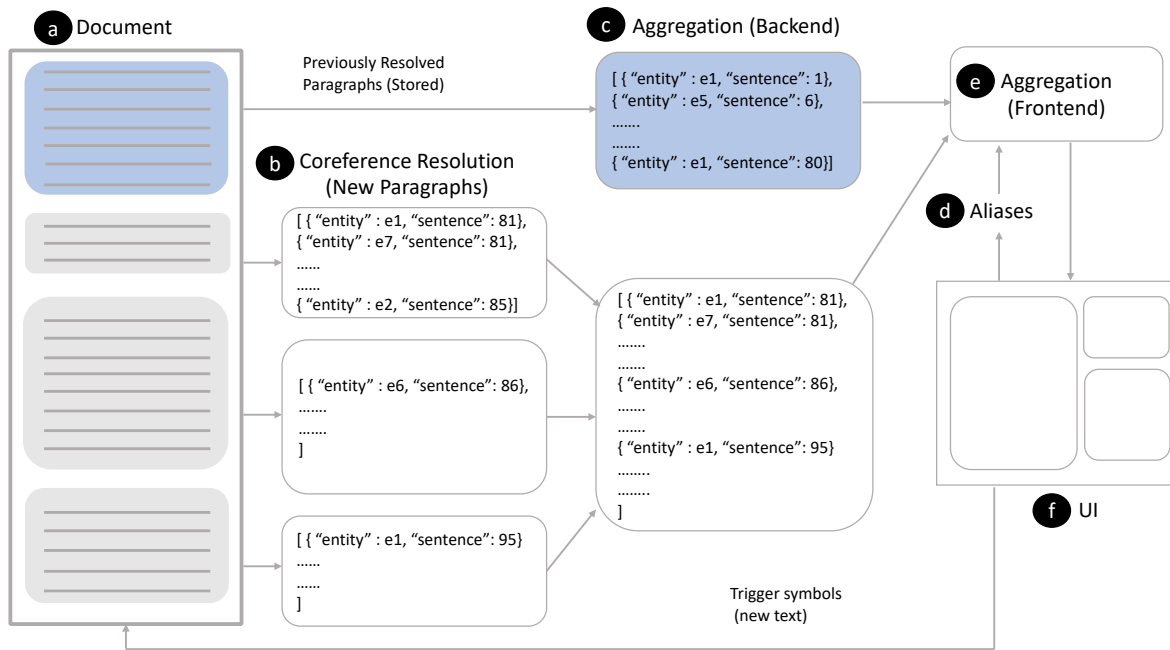
Figure 6: Example workflow of DVP. (a) The system detects three new paragraphs in the text editor. The blue box represents the previously written text while the grey boxes represent the newly added paragraphs. (b) The system runs co-reference resolution and entity extraction independently on the newly added paragraphs. (c) They are then aggregated together with the stored entity information of the previous text. (e) A second aggregation on the front end is applied based on the aliases (d) identified by the user. (f) Finally the JSON formatted data is fed to the UI. The UI listens for the trigger symbols such as ".", "?", "!" or copy/paste event in the text editor for updating the UI.

6.1.1 Intuitive and Easy-to-use. All participants found the interface to be easy and intuitive. Overall, the design, especially the incorporation of visualizations in a writing tool, intrigued participants. Initially, W2 and W3 were slightly confused about the timeline representation but quickly became accustomed to the timeline once they started interacting with the tiles. Participants appreciated that the visualization and text editor were connected, and that writers can easily use the timeline to go to any part of a story. We did not have to provide any explanation for the word zones; participants were already familiar with them. Finally, while it took participants a few minutes to validate the characters, merge aliases, and then assign social identities to the characters, all participants were enthusiastic about this process. W3 suggested a few stylistic changes such as auto-indenting a new paragraph and adding a thesaurus.

6.1.2 Effect of Assigning Social Identities to Characters. The mere presence of an interface where writers can assign social identities to characters had an effect on the writers. W3 said: *"This is an important feature for me as I usually plan very little before writing, and this helped me incorporate planning early into the writing and made me think how the characters should be represented in the story."* W2 mentioned that assigning social identities to characters could be a good practice for writers, especially when characters do not share identities with the writer. This will help them think about the characters from the beginning.

6.1.3 Usage Patterns. All participants mentioned that DVP does not obstruct their creative process. They appreciated that by using DVP they can explore biases when they want to and concentrate on writing or reading other times. For example, W3 used the "Hide" option to hide the visualizations while writing and only checked them periodically. W2 also suggested that they will probably check the visualizations after drafting a scene or section.

Both W1 and W2 extensively used the word zone by comparing words used for characters and social identities. W2 suggested that it would be useful to quickly find the characters that have significantly different adjectives or verbs.

6.1.4 Errors in Identifying Character Mentions. DVP is powered by NLP toolkits and their capabilities to extract character names and identify character mentions throughout the story. The pipeline may sometimes not recognize characters or their relevant mentions. We noticed such cases created confusion among writers. One of the animal characters from W2's story was not identified by our tool. W2 asked whether our tool can identify unusual names since their sci-fi stories often have names that are not common. Similarly, W1 noticed the tool missing out a few mentions of a character in the timeline. While we expect the mitigation of such problems as NLP techniques become increasingly powerful, an alternative would be to allow writers to highlight name entities in the text editor and manually add them for tracking.

6.1.5 Potentials and Impact. At exit interviews, participants discussed several use cases and the impacts of DVP. All participants suggested that DVP would be helpful to any creative writer. They admired its support for different forms of writing (e.g., screenplays, fiction, non-fiction) which makes it usable for a wide range of creative materials. W1 teaches novice journalists on how to write critiques, and was interested in using the tool to show examples of correct and biased representation to the students. Further, W1 thought novice journalists can use the DVP tool as a learning tool and editors can use it to quickly check drafts.

6.2 User Study

We conducted a user study to measure the effectiveness of DVP in identification of biases. We designed a repeated-measures within-subject study with two interfaces: **I1. Baseline**, presented a simplified version of DVP, featuring only the text editor (Figure 1a) and demographic information of the characters (Figure 1b); and **I2. DVP Tool**, presented our interface with all features. We provide a screenshot of the baseline in the supplement.

Before conducting this study, based on the feedback from the think-aloud sessions, we made the following three refinements to our prototype. First, we added a feature for users to manually highlight a character in the text editor and add it to the Characters and Demographics Panel for tracking. This ensures that any character missed by the NLP pipeline can be supported in the system. Second, the study session suggested that while word zone provides an easy way to investigate adjective and verbs, the search space to explore all the combination of entities can be large. To aid writers in this process, we introduced a word embedding based approach to find potential candidate pairs of entities for investigation. Let $\vec{g}1$ be the mean vector representing the words from entity $ent1$ and $\vec{g}2$ be the mean vector representing the words from entity $ent2$. Then, the cosine distance between $\vec{g}1$ and $\vec{g}2$ can indicate how different the words from $ent1$ and $ent2$ are. This method closely matches the *relative norm difference* [30, 32], previously used to quantify gender bias in word embeddings. The top 10 pairs of entities sorted by their cosine distance are fed to the interface as a notification in the word zone control panel. Finally, we added a few convenience features such as auto indent and word count in the system.

6.2.1 Participants. We recruited 11 participants through local mailing lists, university mailing lists, and public posts on Facebook groups and Twitter. Our inclusion criteria included prior experience with any forms of creative writing (fiction, non-fiction, short story, screenplays, etc). We did not require participants to have published articles in their portfolios. The participants varied in age from 21 to 45 (mean 26.5, s.d. 4.32), gender (5 males, 6 females), and race (6 Asian, 4 White, and 1 African American or Black). All participants had prior experience with writing creative materials and using a text editor. None of them participated in any of our prior studies. Participants received $15 per hour worth of Amazon gift cards for their participation.

6.2.2 Data: Short Stories. We used two newly written short stories for the study. One of the authors of this paper who had published fiction in their portfolio wrote one short story ("A Kiss of Ice") for the study. The other story ("The Prisoners") was written by a professional writer, hired from Fiverr.com. We compensated the professional writer with $50. As requirements, we asked both writers to write a short story that includes at least 4-5 characters and 3-4 different intersectional identities with a significant amount of interaction (having conversations, or participating in a scene together) between them. We also requested them to introduce a few stereotypes in the stories, hoping they will incite meaningful discussion in the exit interviews. The two stories and social identities of the characters are provided in the supplement.

6.2.3 Study Questions. DVP is designed to support the identification of two types of biases: (1) lack of agency for characters; and (2) stereotypes in describing characters. Thus, we wanted to test how effective DVP would be in detecting these biases. To do so, we designed 7 questions for each story that either fall into the first or second bias, or both. A few examples of questions are *"Which of the following characters has the least presence in the story? Which of the following social identities has the least presence in the story? Which of the following two characters have the least interactions (having conversations, or participating in a scene together) with each other in the story? How would you categorize how Female characters were portrayed in the story?"*

We provided multiple choices for each question. The questions were similar for both stories although they were slightly modified to adjust to the nature of the stories. Two authors of this paper separately and manually analyzed the stories to get the correct answers and verified them by discussing with each other. The correct answers were also validated by the writers of the stories. All questions are provided in the supplement. Note that DVP is capable of supporting more demanding tasks such as sorting presence of every character; however, we intentionally did not include these tasks as they might be exhausting and overbearing to perform in a plain text editor (our baseline interface).

6.2.4 Procedure. Similar to the think-aloud sessions, we conducted the study sessions via Zoom. The study participants were provided a url where DVP was hosted. They interacted with the system via their own web browser. A study session began with the participant signing a consent form. Following this, the participants were introduced to the assigned first condition. All participants had prior experience in writing and reading in a simple text editor (I1) so they did not require any training for this interface.

For DVP (I2), we provided a brief description of the interface and a demo showing its different features. The participants then interacted with the system (with a training story), during which they were encouraged to ask questions until they were comfortable.

Each participant was then given a short story for the assigned first interface and asked to read the story. On average, it took participants 10 minutes (min = 5 minutes, max = 18 minutes) to read the short story. After participants finished reading the story, we provided them with the task list (questions) for the first interface and asked them to answer the questions using the interface within 15 minutes. The questions were presented in random order to the participants. We provided them verbal clarification for the questions, if needed. At the end of the first interface, we administrated the NASA-TLX [37] questionnaire to measure the participants' perceived workload, and SUS questionnaire to measure user experience and usability of that interface. The same process was carried out
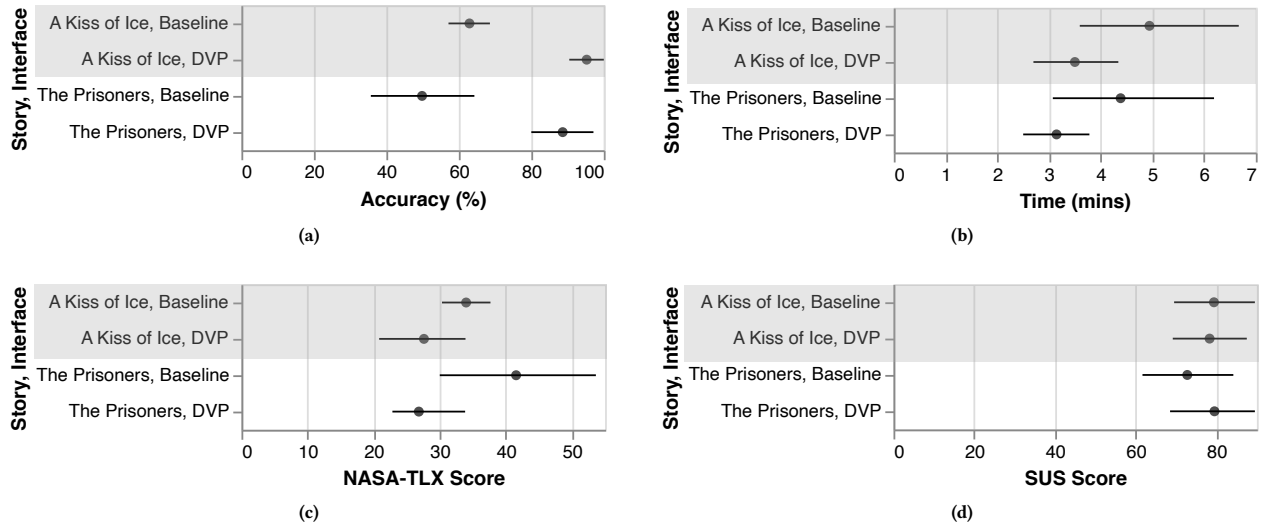
Figure 7: Study Results. (a) Participants' accuracy in answering the study questions; (b) Task completion time in minutes; (c) NASA-TLX (perceived workload) score; and (d) SUS score (usability ratings). Error bars show 95% confidence intervals (CIs).

for the second interface. Each session lasted around 1.5 hours and ended with an exit interview.

6.2.5 Study Design Rationale.
Our research objective from this study was to determine whether DVP can actually help writers identify biases. To evaluate that, we contemplated different study tasks and story types described as follows:

Task. The ideal scenario would perhaps be to ask writers to write a long and complicated story using DVP since that is the primary use case of DVP. However, given the sensitive nature of bias, it is unlikely that writers would be comfortable analyzing their work critically and answering questions related to bias in the presence of a study administrator. Additionally, writing a story, even a short one, takes significant planning and effort, and is likely to be a very long and involved process.

An alternative would be to ask writers to use our tool on their own and provide feedback periodically. While that might lead to insightful feedback, the process is uncertain and it will require a long-term commitment from a group of writers. We felt that this was impractical at this stage of the research project. Thus, we decided to avoid asking writers to write new stories in this study.

Story Type. Since we decided to avoid the writing task, we needed already written stories for the study. We considered using stories written by the writers themselves or freely available stories online. In both cases, participants may have already read and analyzed the stories. Thus, using them may not lead us to isolate the effect of DVP. Additionally, we needed to know the social identities of the characters, which are typically not explicitly provided in published articles. Considering these factors, we decided to use newly written stories for the study.

6.2.6 Results.
To measure performance, we calculated participants' accuracy in answering the questions. We also measured task completion time, perceived workload (NASA-TLX), and usability (SUS). Following recent guidelines for statistical analysis in HCI [19], we intentionally avoided traditional null-hypothesis statistical testing (NHST) in favor of estimation methods to derive 95% confidence intervals (CIs) for all measures. More specifically, we employed non-parametric bootstrapping [20] with $R = 1,000$ iterations. We also report mean differences as sample effect size and Cohen's d as a standardized measure of effect size [14].

Accuracy. Figure 7a presents participants' accuracy in answering the questions. Overall, participants were more accurate when answering the questions using DVP. For the "A Kiss of Ice" story, participants were more accurate on average by 32.49 (d=4.27) percentage points using DVP. Similarly, for "The Prisoners" story, participants were more accurate on average by 38.25 (d=2.31) percentage points using DVP. The very large effect sizes suggest strong practical difference. We noticed participants performed slightly better for the story "A Kiss of Ice" than "The Prisoners" using both interfaces. We anticipated this result as "The Prisoners" featured frequent interactions between different characters and social identities which raised the difficulty level for answering the questions.

Task Completion Time. Figure 7b presents participants' task completion time. The completion time was measured by calculating number of minutes participants took to answer the study questions, excluding the time to read the stories. Unsurprisingly, participants spent less time to answer the questions using DVP. For "A Kiss of Ice," participants spent on average 1.37 (d=0.88) minutes less using DVP. Similarly, for "The Prisoners," participants spent on average 1.24 (d=0.72) minutes less using DVP.

Perceived Workload (NASA-TLX).. Figure 7c presents NASA-TLX scores, average of the raw ratings provided by the participants on

six commonly used workload measures: mental demand, physical demand, temporal demand, effort, performance, and frustration [37]. DVP reduced workloads of participants in answering the questions. For "A Kiss of Ice," on average the perceived workload of the participants were 5.66 (d=0.84) points less using DVP. Similarly, for "The Prisoners" on average the perceived workload of the participants was 14.79 (d=1.08) points less using DVP.

Usability (SUS Score). Participants rated both condition similarly in terms of usability. Figure 7d presents SUS scores calculated from the 10 individual SUS questions. As observed in Figure 7d , DVP did not warrant any major usability issues.

Qualitative Feedback. Participants mentioned that DVP adds a new dimension to reading stories (P2, P4, P7-9). The demographic information of the characters and visualization created a situational awareness among them. For example, P4 said *"Normally, I do not think about the identities of the characters while reading, beyond what can be inferred from the names, and description of the characters. The visualization certainly created a new dimension to my reading and I can imagine the characters in new ways that I would not do usually."* We noticed several participants used the default search and find feature of browsers when answering the questions using the simple text editor. In the exit interviews, several participants mentioned that they used the search and find feature as a way to quickly find relevant information for answering the questions. However, it was not effective in answering the questions as they needed to mentally analyze the search results along with the demographic information which might not be directly encoded as key words in the story. In contrast, the visualization in DVP supported them in answering the questions by providing summarized information. Participants also saw value in DVP as a writing tool. They mentioned that the tool will be most useful for writing long stories with many inter-connected characters and subplots (P1-5, P7, P9). Participants mentioned several use cases where DVP may be helpful. For example, P1 said *"I think this tool has a clear application for writing adapted screenplays. When I write an adapted screenplay, I constantly try to evaluate my writing with the representation in the original book. The tool can be helpful in such scenarios."*

7 DISCUSSION, LIMITATIONS, AND FUTURE WORK

In this section, we discuss potential implications and limitations learned from the development and evaluation of DVP.

Long Term Deployment of DVP.. The think-aloud sessions and user study show promise for DVP as a new medium for writing and critical reading. Even for short stories that were used in the user study, DVP has made the bias identification process much easier for writers. Further, all participants were enthusiastic about the tool and thought it would be useful for them. To evaluate DVP to its full extent, our future work will concentrate on a long term deployment of DVP with writers (e.g., writing a full-length story using DVP).

Impact on Bias in Creative Writing. Bias in creative writing is ubiquitous, as described in Section 2.1. Despite our success with DVP, we believe it is important to acknowledge that our work will not solve systematic or infrastructural discrimination that exists in the publishing community. For example, there is still a lack of writers from marginalized groups in the current writer's community [71]. We consider DVP as a part of the larger movement against biases in creative writing, a probe for reducing biases, and a catalyst for future efforts in this direction. We hope its adaptability to different forms of writing (novels, children's books, short stories) will attract diverse writers.

Balancing Automation and Artistic Freedom/Agency. From the beginning of this project, we were vigilant about the artistic freedom of writers, and how we can balance agency with automation. We understand writers may want to intentionally write about discrimination and biases against a social group, and we do not want to see the tool preventing them in this process. That is one of the reasons why we did not use any automation in correcting biases and stereotypes.

It is also worth noting here that our intention was not to replace critique groups or sensitivity readers. We believe it is important to have subjective feedback about creative writing, especially from relevant marginalized groups. Our intention was to help writers during their writing process by offering another set of eyes—albeit electronic ones.

Design Implications for Human-centered AI.. AI systems can inherit harmful biases and stereotypes. These biases can impact social groups disparately, especially when used as a decision-making platform for critical resources [4, 33, 84]. These systems may also lack inclusivity (e.g., lack of supports for non-binary identities). These limitations motivated several design decisions of our system.

First, based on DG5, we made the *demographics* dropdowns unconstrained so that a writer could add any social identity required for the story. Second, we provided several functions in the interface for the writers to validate (e.g., merge, delete) the results returned by the NLP pipeline, a safety check against potential biases in coreference resolution [96], and dependency parsing [31]. Additionally, a writer can interactively add an entity for tracking in the case the entity was not recognized by the NLP pipeline. Thus, our work promotes human-centered AI for creative writing.

Limitations of the Underlying AI.. NLP methods employed in this paper are highly accurate; however, they are not error free. In particular, it is worth noting that different NLP models like coreference resolution, named entity recognition might exhibit biases based on social identities like gender [62, 96]. This might result in skewed performance against minority groups [62, 63]. Moreover, the NLP models used in this work were trained on a set of news articles, weblogs, etc. Such data might differ from literary texts like novels books, etc. as they might contain longer sentences, more sophisticated language, etc [82]. Although we did not observe any such issues in our experiments, future work might employ fairness-aware NLP models that are trained on domain-specific datasets like books, novels, etc. for better performance.

Finally, the design and development process of this tool has been influenced by English-speaking study participants and NLP models trained over English language corpus. Hence, our tool can currently only support the English language. Having said that, we know that social biases based on gender, race, etc, transcend societies and languages [76, 97]. We believe that ideas put forward in this

work will help facilitate the development of similar tools for other languages as well. Future work might support others languages like French, Chinese, etc. by incorporating NLP models trained over different language corpora.

8 CONCLUSION

In this paper, we have presented an in-depth case study on supporting creative writing and combating implicit bias in fiction using interactive technologies, data visualization, and NLP. In particular, we have reported on an interview study involving 9 creative writers where we asked them about their process, how they navigate harmful stereotypes, and how they think tool support could help in this work. Based on these interviews, we designed DramatVis Personae (DVP), a visual analytics tool using NLP to visualize characters, their demographics, and their story arcs in an effort to mitigate implicit bias. The tool can be used both in an online manner while writing a story, as well as offline during the analysis of an already written story.

We believe that our work here suggests many interesting future avenues of research. For one thing, while creative writing is a notoriously individual and idiosyncratic process, and while the human touch is vital to true art, our moderate success with DramatVis Personae points to possible ways to augment this human touch to improve even such famously crooked processes. In particular, we think that our work shows how automatic machine eyes, while certainly less keen and discerning than human ones, can be helpful for certain applications such as mitigating bias if only because they—unlike human eyes—remain unblinking. We are not so foolish as to believe in the idea of an "unbiased algorithm"—all algorithms are created by humans and thus intrinsically carry the biases of their creators—but we do believe in the virtue of training as many different lenses as possible on a creative artifact in the hope of uncovering yet another harmful stereotype or instance of implicit bias. Thus, we tend to think that our work here is in no way indicative of an end of art, but rather a new beginning.

ACKNOWLEDGMENTS

REFERENCES

[1] Alfie Abdul-Rahman, Julie Lein, Katherine Coles, Eamonn Maguire, Miriah D. Meyer, Martin Wynne, Chris R. Johnson, Anne E. Trefethen, and Min Chen. 2013. Rule-based Visual Mappings - with a Case Study on Poetry Visualization. *Computer Graphics Forum* 32, 3 (2013), 381–390. https://doi.org/10.1111/cgf.12125

[2] Anjali Adukia, Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha, and Teodora Szasz. 2021. *What we teach about race and gender: Representation in images and text of children's books.* Technical Report. National Bureau of Economic Research.

[3] Aretha Barbosa Alencar, Maria Cristina Ferreira de Oliveira, and Fernando Vieira Paulovich. 2012. Seeing beyond reading: a survey on visual text analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2, 6 (2012), 476–492. https://doi.org/10.1002/widm.1071

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23 (2016), 2016.

[5] Charles Perrault Arthur Quiller-Couch. 2016. The Sleeping Beauty and other fairy tales. https://www.gutenberg.org/files/51275/51275-h/51275-h.htm Accessed: 08/30/2021.

[6] Angharad Beckett, Nick Ellison, Sam Barrett, and Sonali Shah. 2010. 'Away with the fairies?' Disability within primary-age children's literature. *Disability &*

[7] *Society* 25, 3 (2010), 373–386.

[8] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D³ data-driven documents. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2301–2309.

[8] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.

[9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *CoRR* abs/2005.14165 (2020), 75. arXiv:2005.14165 https://arxiv.org/abs/2005.14165

[10] Manish Chaturvedi, Gerald C. Gannod, Laura Mandell, Helen Armstrong, and Eric Hodgson. 2012. Myopia: A Visualization Tool in Support of Close Reading. In *Proceedings of the Annual International Conference of the Alliance of Digital Humanities Organizations*, Jan Christoph Meister (Ed.). Hamburg University Press, Hamburg, Germany, 148–150.

[11] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching Stories with Generative Pretrained Language Models. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, to appear.

[12] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. 2018. Creative writing with a machine in the loop: Case studies on slogans and stories. In *Proceedings of the ACM Conference on Intelligent User Interfaces*. ACM, New York, NY, USA, 329–340.

[13] Andy Coenen, Luke Davis, Daphne Ippolito, Emily Reif, and Ann Yuan. 2021. Wordcraft: a Human-AI Collaborative Editor for Story Writing. *CoRR* abs/2107.07430 (2021), 7. arXiv:2107.07430 https://arxiv.org/abs/2107.07430

[14] Jacob Cohen. 1988. *Statistical Power Analysis for the Social Sciences*. Erlbaum Associates, Hillsdale, NJ, USA.

[15] Joshua Correll, Bernadette Park, Charles M. Judd, and Bernd Wittenbrink. 2007. The influence of stereotypes on decisions to shoot. *European Journal of Social Psychology* 37, 6 (2007), 1102–1117.

[16] Michael Correll, Michael Witmore, and Michael Gleicher. 2011. Exploring Collections of Tagged Text for Literary Scholarship. *Computer Graphics Forum* 30, 3 (2011), 731–740. https://doi.org/10.1111/j.1467-8659.2011.01922.x

[17] Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai Gao, Huamin Qu, and Xin Tong. 2011. TextFlow: Towards Better Understanding of Evolving Topics in Text. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2412–2421. https://doi.org/10.1109/TVCG.2011.239

[18] Niels Dekker, Tobias Kuhn, and Marieke van Erp. 2019. Evaluating named entity recognition tools for extracting social networks from novels. *PeerJ Computer Science* 5 (2019), e189.

[19] Pierre Dragicevic. 2016. Fair Statistical Communication in HCI. In *Modern Statistical Methods for HCI*. Springer Publishing, New York, NY, USA, 291–330.

[20] Bradley Efron. 1979. Bootstrap methods: another look at the jackknife. *The Annals of Statistics* 7, 1 (1979), 1–26. https://www.jstor.org/stable/2958830

[21] Pascale Emons, Fred Wester, and Peer Scheepers. 2010. "He Works Outside the Home; She Drinks Coffee and Does the Dishes" Gender Roles in Fiction Programs on Dutch Television. *Journal of Broadcasting & Electronic Media* 54, 1 (2010), 40–53.

[22] Robert M. Entman. 1992. Blacks in the news: Television, modern racism and cultural change. *Journalism Quarterly* 69, 2 (1992), 341–361.

[23] Hugging Face. 2021. NeuralCoref 4.0: Coreference Resolution in spaCy with Neural Networks. https://github.com/huggingface/neuralcoref Accessed: 2022-02-04.

[24] Ethan Fast, Tina Vachovsky, and Michael S. Bernstein. 2016. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Proceedings of the International AAAI Conference on Web and Social Media*. Association for Computational Linguistics, Florence, Italy, 112–120.

[25] Xavier Ferrer, Tom van Nuenen, Jose M Such, and Natalia Criado. 2021. Discovering and categorising language biases in Reddit. *Proceedings of the International AAAI Conference on Web and Social Media*, 140–151.

[26] Vicki C Flerx, Dorothy S Fidler, and Ronald W Rogers. 1976. Sex role stereotypes: Developmental aspects and early intervention. *Child Development* (1976), 998–1007.

[27] The Centre for Literacy in Primary Education. 2020. Reflecting Realities-Survey of Ethnic Representation within UK Children's Literature. https://clpe.org.uk/research/clpe-reflecting-realities-survey-ethnic-representation-within-uk-childrens-literature Accessed: 08/30/2021.

[28] Samah Gad, Waqas Javed, Sohaib Ghani, Niklas Elmqvist, E. Thomas Ewing, Keith N. Hampton, and Naren Ramakrishnan. 2015. ThemeDelta: Dynamic Segmentations over Temporal Topic Models. *IEEE Transactions on Visualization and Computer Graphics* 21, 5 (2015), 672–685. https://doi.org/10.1109/TVCG.2014.

2388208

[29] Qihong Gan, Min Zhu, Mingzhao Li, Ting Liang, Yu Cao, and Baoyao Zhou. 2014. Document visualization: an overview of current research. *WIREs Computational Statistics* 6, 1 (2014), 19–36. https://doi.org/10.1002/wics.1285

[30] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644.

[31] Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-Bias in Part-of-Speech Tagging and Dependency Parsing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3493–3498.

[32] Bhavya Ghai, Md Naimul Hoque, and Klaus Mueller. 2021. WordBias: An Interactive Visual Tool for Discovering Intersectional Biases Encoded in Word Embeddings. In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–7.

[33] Bhavya Ghai, Mihir Mishra, and Klaus Mueller. 2022. Cascaded Debiasing : Studying the Cumulative Effect of Multiple Fairness-Enhancing Interventions. , 12 pages. arXiv:2202.03734 https://arxiv.org/abs/2202.03734

[34] Granthika. 2021. Granthika Writing Tool. https://granthika.co Accessed: 2022-02-04.

[35] Martin Gronemann, Michael Jünger, Frauke Liers, and Francesco Mambelli. 2016. Crossing Minimization in Storyline Visualization. In *Proceedings of the International Symposium on Graph Drawing and Network Visualization (Lecture Notes in Computer Science, Vol. 9801)*. Springer, 367–381. https://doi.org/10.1007/978-3-319-50106-2_29

[36] Mykol C. Hamilton, David Anderson, Michelle Broaddus, and Kate Young. 2006. Gender stereotyping and under-representation of female characters in 200 popular children's picture books: A twenty-first century update. *Sex Roles* 55, 11 (2006), 757–765.

[37] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the human factors and ergonomics society annual meeting* 50, 9 (2006), 904–908.

[38] Marti A. Hearst, Emily Pedersen, Lekha Patil, Elsie Lee, Paul Laskowski, and Steven Franconeri. 2019. An evaluation of semantically grouped word cloud designs. *IEEE Transactions on Visualization and Computer Graphics* 26, 9 (2019), 2748–2761.

[39] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. *spaCy: Industrial-strength Natural Language Processing in Python*. https://doi.org/10.5281/zenodo.1212303

[40] Md Naimul Hoque, Nazmus Saquib, Syed Masum Billah, and Klaus Mueller. 2020. Toward Interactively Balancing the Screen Time of Actors Based on Observable Phenotypic Traits in Live Telecast. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–18.

[41] Alexander Miserlis Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, Isabelle Augenstein, and Ryan Cotterell. 2019. Unsupervised Discovery of Gendered Language through Latent-Variable Modeling. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1706–1716. https://doi.org/10.18653/v1/P19-1167

[42] Chieh-Jen Huang, Cheng-Te Li, and Man-Kwan Shan. 2013. VizStory: Visualization of Digital Narrative for Fairy Tales. In *Proceedings of the Conference on Technologies and Applications of Artificial Intelligence*. IEEE Computer Society, Piscataway, NJ, USA, 67–72. https://doi.org/10.1109/TAAI.2013.26

[43] Julie S Hui, Darren Gergle, and Elizabeth M Gerber. 2018. IntroAssist: A tool to support writing introductory help requests. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173596

[44] Darnell Hunt and Ana-Christina Ramón. 2020. Hollywood Diversity Report 2020. https://socialsciences.ucla.edu/wp-content/uploads/2020/02/UCLA-Hollywood-Diversity-Report-2020-Film-2-6-2020.pdf

[45] Kenneth Joseph, Wei Wei, and Kathleen M. Carley. 2017. Girls rule, boys drool: Extracting semantic and affective stereotypes from Twitter. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, New York, NY, USA, 1362–1374.

[46] Daniel Jurafsky and James H Martin. [n.d.]. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

[47] Daniel A. Keim and Daniela Oelke. 2007. Literature Fingerprinting: A New Method for Visual Literary Analysis. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*. IEEE, Piscataway, NJ, USA, 115–122. https://doi.org/10.1109/VAST.2007.4389004

[48] Allen Kim, Charuta Pethe, and Steven Skiena. 2020. What Time Is It? Temporal Analysis of Novels. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 9076–9086.

[49] Nam Wook Kim, Benjamin Bach, Hyejin Im, Sasha Schriber, Markus H. Gross, and Hanspeter Pfister. 2018. Visualizing Nonlinear Narratives with Story Curves. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 595–604.

[50] Richard Kolbe and Joseph C La Voie. 1981. Sex-role stereotyping in preschool children's picture books. *Social Psychology Quarterly* (1981), 369–374.

[51] Eve Kraicer and Andrew Piper. 2019. Social characters: the hierarchy of gender in contemporary English-language fiction. *Journal of Cultural Analytics* 1, 1 (2019), 11055.

[52] Vincent Labatut and Xavier Bost. 2019. Extraction and analysis of fictional character networks: A survey. *Comput. Surveys* 52, 5 (2019), 1–40.

[53] Heidi Layne and Amikeng A Alemanji. 2015. "Zebra world": The promotion of imperial stereotypes in a children's book. *Power and Education* 7, 2 (2015), 181–195.

[54] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. *procCHI* (2022).

[55] Literature and Latte. 2021. Scrivener. https://www.literatureandlatte.com/scrivener/overview/ Accessed: 2022-02-04.

[56] Shixia Liu, Yingcai Wu, Enxun Wei, Mengchen Liu, and Yang Liu. 2013. StoryFlow: Tracking the Evolution of Stories. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2436–2445. https://doi.org/10.1109/TVCG.2013.196

[57] Li Lucy, Dorottya Demszky, Patricia Bromley, and Dan Jurafsky. 2020. Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in Texas US history textbooks. *AERA Open* 6, 3 (2020), 2332858420940312.

[58] Nishtha Madaan, Sameep Mehta, Taneea Agrawaal, Vrinda Malhotra, Aditi Aggarwal, Yatin Gupta, and Mayank Saxena. 2018. Analyze, detect and remove gender stereotyping from bollywood movies. In *Conference on fairness, accountability and transparency*. PMLR, 92–105.

[59] Neil Maiden, Konstantinos Zachos, Amanda Brown, George Brock, Lars Nyre, Aleksander Nygård Tonheim, Dimitris Apsotolou, and Jeremy Evans. 2018. Making the news: Digital creativity support for journalists. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–11.

[60] Ariadna Matamoros-Fernández. 2017. Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society* 20, 6 (2017), 930–946.

[61] Nina McCurdy, Julie Lein, Katherine Coles, and Miriah D. Meyer. 2016. Poemage: Visualizing the Sonic Topology of a Poem. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 439–448. https://doi.org/10.1109/TVCG.2015.2467811

[62] Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2020. Man is to person as woman is to location: Measuring gender bias in named entity recognition. In *Proceedings of the ACM Conference on Hypertext and Social Media*. ACM, New York, NY, USA, 231–232.

[63] Shubhanshu Mishra, Sijun He, and Luca Belli. 2020. Assessing Demographic Bias in Named Entity Recognition. *arXiv preprint arXiv:2008.03415* (2020).

[64] Michel Mitri. 2020. Story analysis using natural language processing and interactive dashboards. *Journal of Computer Information Systems* (2020), 1–11.

[65] May M Narahara. 1998. Gender Stereotypes in Children's Picture Books. (1998).

[66] Cathrine Norberg. 2016. Naughty boys and sexy girls: the representation of young individuals in a web-based corpus of English. *Journal of English Linguistics* 44, 4 (2016), 291–317.

[67] Geena Davis Institue on Gender in Media. 2008. Gender Stereotypes: An Analysis of Popular Films and TV. https://seejane.org/wp-content/uploads/GDIGM_Gender_Stereotypes.pdf

[68] Geena Davis Institue on Gender in Media. 2019. Historic Gender Parity in Children's Television. https://seejane.org/research-informs-empowers/seejane-2019/

[69] Geena Davis Institue on Gender in Media. 2019. Representation Of Black Women in Hollywood. https://seejane.org/research-informs-empowers/representations-of-black-women-in-hollywood/

[70] Geena Davis Institue on Gender in Media. 2020. Historic Gender Parity in Family Films! https://seejane.org/2020-film-historic-gender-parity-in-family-films/

[71] Geena Davis Institue on Gender in Media. 2021. Behind the Scenes: the State of Inclusion and Equity In TV Writing. https://seejane.org/research-informs-empowers/behind-the-scenes-the-state-of-inclusion-and-equity-in-tv-writing/

[72] Kelly Crisp Paynter. 2011. *Gender stereotypes and representation of female characters in children's picture books*. Liberty University.

[73] Michael Pearce. 2008. Investigating the collocational behaviour of MAN and WOMAN in the BNC using Sketch Engine. *Corpora* 3, 1 (2008), 1–29.

[74] Sharyl Bender Peterson and Mary Alyce Lach. 1990. Gender stereotypes in children's books: Their prevalence and influence on cognitive and affective development. *Gender and education* 2, 2 (1990), 185–197.

[75] Charuta Pethe, Allen Kim, and Steven Skiena. 2020. Chapter Captor: Text Segmentation in Novels. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 8373–8383.

[76] Arun K Pujari, Ansh Mittal, Anshuman Padhi, Anshul Jain, Mukesh Jadon, and Vikas Kumar. 2019. Debiasing Gender biased Hindi Words with Word-embedding.

https://doi.org/10.1109/TVCG.2017.2744118

In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*. 450–456.

[77] QuillJS. 2021. Delta. https://github.com/quilljs/delta/ Accessed: 2022-02-04.

[78] QuillJS. 2021. QuillJS: a Rich Text Editor. https://quilljs.com Accessed: 2022-02-04.

[79] Srividya Ramasubramanian and Mary Beth Oliver. 2007. Activating and suppressing hostile and benevolent racism: Evidence for comparative media stereotyping. *Media psychology* 9, 3 (2007), 623–646.

[80] Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science* 5, 1 (2016), 1–12.

[81] Randall M. Rohrer, John L. Sibert, and David S. Ebert. 1998. The Shape of Shakespeare: Visualizing Text using Implicit Surfaces. In *Proceedings of the IEEE Symposium on Information Visualization*. IEEE, Piscataway, NJ, USA, 121–129. https://doi.org/10.1109/INFVIS.1998.729568

[82] Ina Rösiger, Sarah Schulz, and Nils Reiter. 2018. Towards coreference for literary text: Analyzing domain-specific phenomena. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. 129–138.

[83] Tara Ross. 2019. Media and stereotypes. *Journal: The Palgrave Handbook of Ethnicity* (2019), 1–17.

[84] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–33.

[85] Artur Ŝilic and Bojana Dalbelo Basic. 2010. Visualization of Text Streams: A Survey. In *Proceedings of the International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (Lecture Notes in Computer Science, Vol. 6277)*, Rossitza Setchi, Ivan Jordanov, Robert J. Howlett, and Lakhmi C. Jain (Eds.). Springer, 31–43. https://doi.org/10.1007/978-3-642-15390-7_4

[86] John T. Stasko, Carsten Görg, Zhicheng Liu, and Kanupriya Singhal. 2007. Jigsaw: Supporting Investigative Analysis through Interactive Visualization. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*. IEEE, Piscataway, NJ, USA, 131–138. https://doi.org/10.1109/VAST.2007.4389006

[87] Sarah Sterman, Evey Huang, Vivian Liu, and Eric Paulos. 2020. Interacting with Literary Style through Computational Tools. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12.

[88] Yuzuru Tanahashi and Kwan-Liu Ma. 2012. Design Considerations for Optimizing Storyline Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2679–2688. https://doi.org/10.1109/TVCG.2012.212

[89] Leo Tolstoy. 1998. Anna Karenina. https://www.gutenberg.org/files/1399/1399-0.txt Accessed: 08/30/2021.

[90] Shubham Toshniwal, Sam Wiseman, Allyson Ettinger, Karen Livescu, and Kevin Gimpel. 2020. Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 8519–8526.

[91] Ya-Lun Tsao. 2008. Gender issues in young children's literature. *Reading Improvement* 45, 3 (2008), 108–115.

[92] Nicholas A Valentino. 1999. Crime news and the priming of racial attitudes during evaluations of the president. *Public Opinion Quarterly* (1999), 293–320.

[93] Fernanda B Viegas, Martin Wattenberg, and Jonathan Feinberg. 2009. Participatory visualization with wordle. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 1137–1144.

[94] Katie Watson, Samuel S. Sohn, Sasha Schriber, Markus Gross, Carlos Manuel Muñiz, and Mubbasir Kapadia. 2019. StoryPrint: an interactive visualization of stories. In *Proceedings of the ACM Conference on Intelligent User Interfaces*. ACM, New York, NY, USA, 303–311. https://doi.org/10.1145/3301275.3302302

[95] James A. Wise, James J. Thomas, Kelly Pennock, David Lantrip, Marc Pottier, Anne Schur, and Vern Crow. 1995. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Proceedings of the IEEE Symposium on Information Visualization*. IEEE, Piscataway, NJ, USA, 51–58. https://doi.org/10.1109/INFVIS.1995.528686

[96] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876* (2018).

[97] Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining Gender Bias in Languages with Grammatical Gender. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 5279–5287.