RIVETER
Measuring Power and Social Dynamics Between Entities

Maria Antoniak[♣] **Anjalie Field**[†] **Jimin Mun**[♡] **Melanie Walsh**[◇]
Lauren F. Klein[♠] **Maarten Sap**[♡♣]

[♣]Allen Institute for AI [†]Johns Hopkins University [♡]Carnegie Mellon University
[◇]University of Washington [♠]Emory University

http://github.com/maartensap/riveter-nlp

Abstract

RIVETER provides a complete easy-to-use pipeline for analyzing verb connotations associated with entities in text corpora. We prepopulate the package with connotation frames of sentiment, power, and agency, which have demonstrated usefulness for capturing social phenomena, such as gender bias, in a broad range of corpora. For decades, lexical frameworks have been foundational tools in computational social science, digital humanities, and natural language processing, facilitating multifaceted analysis of text corpora. Working with verb-centric lexica specifically requires language processing skills, reducing their accessibility to other researchers. By conducting the language processing pipeline, providing complete lexicon scores and visualizations for all entities in a corpus, and providing functionality for users to target specific research questions, RIVETER greatly improves the accessibility of verb lexicons and can facilitate a broad range of future research.

Video: https://youtu.be/Uftyd8eCmFw

1 Introduction

Language often encodes social dynamics between people, such as perspectives, biases, and power differentials (Fiske, 1993). When writing, authors choose how to *portray* or *frame* each person in a text, highlighting certain features (Entman, 1993) to form larger arguments (Fairhurst, 2005). For example, in the screenplay for the 2009 film *Sherlock Holmes*, the authors dramatize a sudden reversal of power by playing on gender stereotypes. First, they describe how "the man with the roses **beckons** Irene forward" (Figure 1), which portrays the character Irene Adler as being lured by the man. After she is trapped, "she **slices** upward with a razor-sharp knife," reversing the power dynamic. Here, specific word choices shape and then challenge the viewers' expectations about how the interaction is presumed to unfold. More broadly,

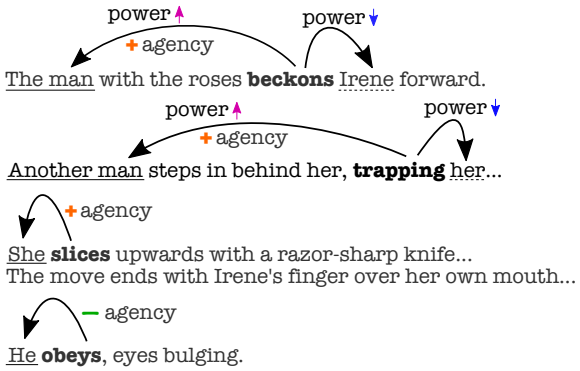

Figure 1: Figure from Sap et al. (2017) illustrating power and agency connotation frames extracted on an excerpt from the *Sherlock Holmes* (2009) film screenplay. Each connotation frame pertains to a **verb predicate** and its agent and theme.

such word choices can not only communicate important details about a character or narrative but they can also reveal larger social attitudes and biases (Blackstone, 2003; Cikara and Fiske, 2009), shape readers' opinions and beliefs about social groups (Behm-Morawitz and Mastro, 2008), as well as act as powerful mechanisms to persuade or to induce empathy (Smith and Petty, 1996; Keller et al., 2003). Studying these word choices across large datasets of political speeches, newspaper articles, novels, or other texts can illuminate domain-specific patterns of interest to scholars in the humanities and social sciences (e.g., in cultural analytics or computational social science).

Examining verbs—*who* does *what* to *whom*?—is one established approached for measuring the textual portrayal of people and groups of people. Each verb carries connotations that can indicate the social dynamics at play between the subject and object of the verb. Connotation frames (Rashkin et al., 2016; Sap et al., 2017) capture these dynamics by coding verbs with directional scores, indicating, e.g., who holds power and who lacks power, who is portrayed with positive and negative sentiment. In

the *Sherlock Holmes* scene description, "Another man steps in behind her, **trapping her**," the verb *to trap* implies that "the man" has more power over "her" (Figure 1; Sap et al., 2017). These verb lexica have been used successfully to study portrayals in many diverse contexts including films (Sap et al., 2017), online forums (Antoniak et al., 2019), text books (Lucy et al., 2020), Wikipedia (Park et al., 2021), novels, and news articles (Field et al., 2019).

Lexica in general are extremely popular among social science and digital humanities scholars. They are interpretable and intuitive (Grimmer and Stewart, 2013), especially when compared with black-box classification models, and continue to be a go-to resource (e.g., LIWC; Pennebaker et al., 2015). Verb-based lexica, however, pose specific technical hurdles for those less experienced in software engineering and natural language processing (NLP). They require core NLP skills such as traversing parse trees, identifying named entities and references, and lemmatizing verbs to identify matches. But the larger research questions motivating the usage of these lexica require deep domain expertise from the social sciences and humanities.

To meet this need, we introduce RIVETER,[1] which includes tools to use, evaluate, visualize, and create verb lexica, enabling researchers to measure power and other social dynamics between entities in text. This package includes a pipeline system for importing a lexicon, parsing a dataset, identifying people or entities, resolving coreferences, and measuring patterns across those entities. It also includes evaluation and visualization methods to promote grounded analyses within a targeted dataset. We release the package as a Python package, along with Jupyter notebook demonstrations and extensive documentation aimed at social science and humanities researchers.

To showcase the usefulness of this package, we describe two case studies: (1) power differentials and biases in GPT-3 generated stories and (2) gender-based patterns in the novel *Pride and Prejudice*. The first study provides a proof-of-concept; dyads with predetermined power differentials are used to generate stories, and we are able to detect these distribution shifts using RIVETER. The second study zooms in on a particular author, text, and social setting, examining a 19th century novelist both portrayed and subverted gender roles. These case studies highlight the diverse contexts and research questions for which this package can be used across human and machine-generated text, and across the social sciences and the humanities.

[1]The name Riveter is inspired by "Rosie the Riveter," the allegorical figure who came to represent American women working in factories and at other industrial jobs during World War II. Rosie the Riveter has become an iconic symbol of power and shifting gender roles—subjects that the Riveter package aims to help users measure and explore.

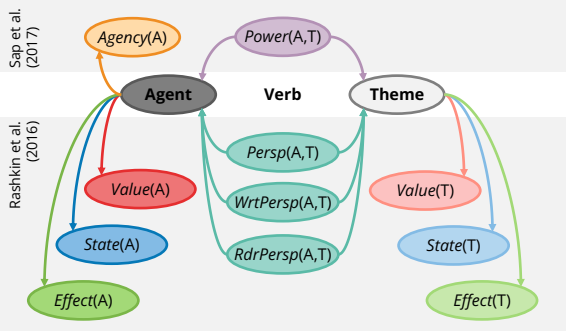

Figure 2: A verb predicate can connote various sentiment, power, and agency levels for its agent and theme; connotation frames distill these into six relation types (described in §2.3).

2 Background: Verb Lexica & Connotation Frames

2.1 Verb Predicates & Frame Semantics

Understanding the events and states described in sentences, i.e., *who did what to whom*, has been a central question of linguistics since first conceptualized by Indian grammarian Pāṇini between the 4th and 6th century BCE. Today, verb predicates and their relation to other words in a sentence are still a key focus of linguistic analyses, e.g., dependency parsing (Tesnière, 2015; Nivre, 2010) and semantic role labeling (Gildea and Jurafsky, 2002).

To model how one understands and interprets the events in a sentence, Fillmore (1976) introduced *frame semantics*, arguing that understanding a sentence involves knowledge that is evoked by the concepts in the sentence. This theory inspired the task of *semantic role labeling* (Gildea and Jurafsky, 2002), which categorizes how words in a sentence relate to the main verb predicate via frame semantics. This task defines *thematic roles* with respect to an EVENT (i.e., the verb predicate): the AGENT that causes the EVENT (loosely, the subject of the verb), and the THEME that is most directly affected by the EVENT (loosely, the object of the verb).

Work	Usage
Rashkin et al. (2016)	Analyzing political leaning and bias in news articles.
Sap et al. (2017)	Analyzing gender bias in portrayal of characters in movie scripts.
Rashkin et al. (2017a)	Analyzing public sentiment (and multilingual extension of Rashkin et al. (2016))
Volkova and Jang (2018)	Improving the detection of fake news & propaganda.
Ding and Riloff (2018)	Detecting affective events in personal stories.
Field et al. (2019)	Analyzing power dynamics of news portrayals in #MeToo stories.
Park et al. (2021)	Comparing affect in multilingual Wikipedia pages about LGBT people
Antoniak et al. (2019)	Analyzing the power dynamics in birthing stories online.
Lucy et al. (2020)	Analyzing the portrayal of minority groups in textbooks.
Mendelsohn et al. (2020)	Analyzing the portrayal of LGBTQ people in the New York Times.
Ma et al. (2020)	Text rewriting for mitigating agency gender bias in movies.
Lucy and Bamman (2021)	Analyzing gender biases in GPT3-generated stories.
Gong et al. (2022)	Quantifying gender biases and power differentials in Japanese light novels
Saxena et al. (2022)	Examining latent power structures in child welfare case notes
Borchers et al. (2022)	Measuring biases in job advertisements and mitigating them with GPT-3
Stahl et al. (2022)	Joint power-and-agency rewriting to debias sentences.
Wiegand et al. (2022)	Identifying implied prejudice and social biases about minority groups
Giorgi et al. (2023)	Examining the portrayal of narrators in moral and social dilemmas

Table 1: Examples of usages of connotation frames in NLP and CSS literature.

2.2 Connotation Frames of Sentiment, Power, and Agency

While frame semantics was originally meant to capture broad meaning that arises from interpreting words in the context of what is known (Fillmore, 1976), many linguistic theories have focused solely on denotational meaning (Baker et al., 1998; Palmer et al., 2005), i.e., examining only what is present in the sentence. In contrast, the *implied or connoted meaning* has received less attention, despite being crucial to interpreting sentences.

Connotation frames, introduced by Rashkin et al. (2016), were the first to model the connotations of verb predicates with respect to an AGENT and THEME's value, sentiment, and effects (henceforth, sentiment connotation frames). Shortly thereafter, Sap et al. (2017) introduced the *power and agency connotation frames* (Figure 2), which model the power differential between the AGENT and the THEME, as well as the general agency that is attributed to the AGENT of the verb predicate.[2]

For both sets of connotation frames, the authors released a lexicon of verbs with their scores. Verbs were selected based on their high usage in corpora of choice: frequently occurring verbs from a corpus of New York Times articles (Sandhaus, 2008) for sentiment connotation frames, and frequently occurring verbs in a movie script corpus (Gorinski and Lapata, 2015) for the power and agency frames. Each verb was annotated for each dimension by crowdworkers with AGENT and THEME placeholders ("*X implored Y*").

Since their release, connotation frames have been of increasing interest to the computational social science (CSS) and digital humanities (DH) communities; subsequent work has used them to analyze various types of portrayals in texts, ranging from news articles and textbooks to movie scripts and personal blog posts (see Table 1). Additionally, the frames have been incorporated into the 2023 edition of the textbook *Speech and Language Processing* (Jurafsky and Martin, 2023).

2.3 Connotation Frame Dimensions

Given a predicate verb v describing an EVENT and its AGENT a and THEME t, connotation frames capture several implied relations along sentiment, power, and agency dimensions. Each of these relations has either a positive ($+$), neutral ($=$), or negative ($-$) polarity.

Effect denotes whether the event described by v has a positive or negative effect on the agent a or the theme t. For example, in Figure 1, another man *trapping* Irene has a negative effect on her ($\textit{Effect}(t) = -$).

Value indicates whether the agent or theme are presupposed to be of value by the predicate v. For example, when someone *guards* an object, this presupposes that the object has high value ($\textit{Value}(t) = +$).

State captures whether the likely mental state of the AGENT or THEME as a result of the EVENT.

[2]While the value, sentiment, effects, and power relations require a verb to be transitive, the agency dimension is present with intransitive verbs as well.

For example, someone *suffering* likely indicates a negative mental state (*State*$(a) = -$).

Perspective is a set of relations that describe the sentiment of the AGENT towards the THEME and vice versa (*Persp*(a, t)). It also describes how the writer perceives the AGENT and THEME (*WrtPersp*(a, t)), as well as the reader likely feels towards them *RdrPersp*(a, t).

Power distills the power differential between the AGENT and THEME of the EVENT. For example, when a man *traps* Irene, he has power over Irene (*Power*$(a > t)$).

Agency denotes whether the AGENT of the EVENT has agency, i.e., is decisive and can act upon their environment. For example, Irene *slicing* connotes high agency (*Agency*$(a) = +$), whereas the man *obeying* connotes low agency (*Agency*$(a) = -$).

3 RIVETER Design & Implementation

3.1 Challenges Addressed

Unlike lexica that require only string matching, verb lexica also require parsing, lemmatization, named entity recognition, and coreference resolution. These are standard pieces of NLP pipelines, but each piece requires background knowledge in linguistics, NLP, and algorithms that inform library choices and merging their outputs.

Even as the output of a lexical analysis is easily interpretable and yields results that are readily incorporated back into qualitative textual analyses, it can be difficult for humanities and social science researchers to assemble and link together the various pieces of the required processing pipeline. RIVETER substantially lowers the implementation burden and text-processing knowledge required for using verb lexica by addressing three challenges:

Familiarity with Using NLP Tools The increasing availability of NLP packages has resulted in numerous existing packages for core NLP pipelines. We reviewed the performance (considering accuracy, speed, and ease of installation) of available tools and pre-selected optimal text processing pipelines for RIVETER, eliminating the need for users to be familiar with and decide between available text processing tools. We also provide documentation on incorporated packages and extensive demonstrations.

Interfacing between NLP Tools Even if one is familiar with individual tools, like parsers or entity recognizers, connecting outputs from one tool to another tool requires an additional engineering skill set. Traversing a parse tree to find semantic triples and then matching these triples to clusters from a coreference resolution engine is not a straightforward process for a researcher with domain expertise in humanities or social science topics but less expertise in programming and software engineering. To address this challenge, we (a) provide a system that connects these pipeline pieces for the user while also (b) providing functionality to explore the outputs of each individual system.

Interpreting Results Lexical methods can offer flexibility and interpretability not found in other NLP methods, but even so, validating and exploring lexical results can be challenging. Proper validation is not consistent even in NLP research using lexicon-based methods (Antoniak and Mimno, 2021). To address this challenge, we provide methods to explore the results numerically and visually.

3.2 System Description

Illustrated in Figure 3, RIVETER takes in a set of documents as input, and returns a set of scores for each person or entity in the documents. Under the hood, it first parses the documents, resolves coreference clusters, finds people mentions, extracts agent-verb-theme triples, and computes lexicon scores. We verify implementation through hand-constructed unit tests and testing of large corpora. We describe each of these components below.

Named Entity Recognition and Coreference Resolution Our package first parses a given document to find clusters of mentions that relate to people or entities. We extract general coreference clusters, which we cross-reference with mentions of people or entities labeled by a named entity recognition (NER) system, as well as a list of general people referents (containing pronouns, professions, etc). Coreference cluster mentions that overlap with a mention of a person or entity are then passed to the verb and relation identification module.

In our implementation, we use spaCy for NER, and the spaCy add-on *neuralcoref*[3] library for coreference resolution. These libraries are well-supported, have fast run-times relative to similar

[3] https://github.com/huggingface/neuralcoref

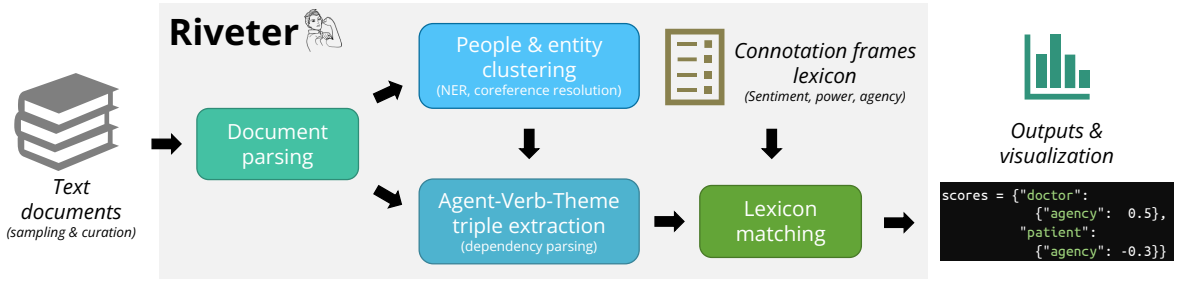

Figure 3: A visualization of the RIVETER pipeline components and their connections.

systems, and do not require GPU access, which are not accesible to many social science researchers.

Lexicon Loading We include two lexica by default: connotation frames from Rashkin et al. (2016) and frames of power and agency from Sap et al. (2017). These lexica come included in the package, and the user can select between the lexica and their dimensions (see §2.3 for dimension descriptions). For each lexicon, we convert the lexicon labels to numerical scores; each verb has a score of either $+1$ or -1 for each of [*agent*, *theme*].

Verb Identification and Entity Relation We extract the lemmas of all verbs and match these to the lexicon verbs. After identifying semantic triples (the subject (*nsubj*) and direct object (*dobj*) of each verb) using the spaCy dependency parser, we search for matches to the NER spans identified in §3.2. We track the frequencies of these for the canonical entity, using the converted scores.

Exploration and Visualization We provide functionality for users to easily view lexicon scores for entities in their input text. To maximize utility, we focus on facilitating analyses established in prior work (e.g., Table 1). This functionality includes:

- Retrieving the overall verb lexica scores for all entities identified in the entire input corpora (`get_score_totals`)
- Retrieving the overall verb lexica scores for all entities identified in a specific document (`get_scores_for_doc`)
- Generating bar plots of scores for entities in the corpora (e.g., filtering for the top-scored entities) (`plot_scores`) or in a specific document (`plot_scores_for_doc`)

We additionally provide functionality to reduce the opacity of lexicon scores and allow users to examine specific findings in more depth or conduct error checking. These functions include:

- Generating heat map plots for the specific verbs used with a user-specified entity (`plot_verbs_for_persona`)
- Retrieve all mentions associated with a specified entity, after co-reference resolution (`get_persona_cluster`)
- Retrieving various additional counts, including number of lexica verbs in a document, all entity-verb pairs in a document, number of identified entities, etc.

4 Case Studies Across Cultural Settings

4.1 Machine Stories with Power Dyads

As our first case study, we examine lexicon-based power differentials in machine-generated stories about two characters with a predetermined power asymmetry (e.g., *"doctor, patient"*, *"teacher, student"*, and *"boss, employee"*). By generating stories about entities with predetermined power asymmetries, this serves as a proof-of-concept for RIVETER; we expect to measure power scores in the predetermined directions.

Given a set of 32 pairs of roles, we obtain 85 short stories from GPT-3.5 (Ouyang et al., 2021) (see Appendix A for details). We then scored each of the characters with assigned roles using RIVETER and aggregated the scores for higher power roles and lower power roles using their names given by GPT-3.5.

Results As seen in Figure 4, higher power roles have a distribution shifted toward greater power scores than lower power roles. The mean score of the higher power roles was 0.265 and that of lower power roles was 0.0364. A t-test also shows statistical significance ($p < 0.05$). From these results we can conclude that the stories generated by

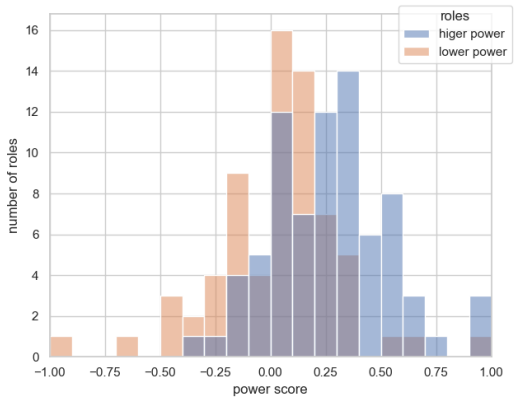

Figure 4: Power scores of higher and lower power roles

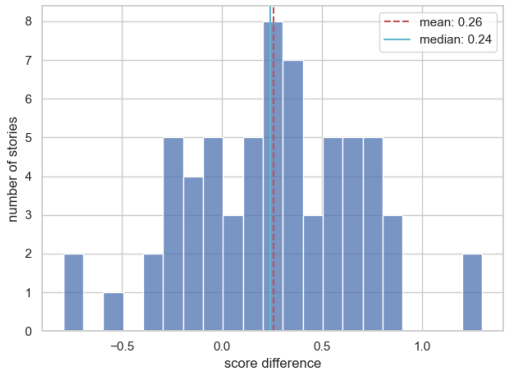

Figure 5: Power score differences in generated stories.

GPT-3.5 reflect the power dynamics in the relations given in the prompt, and our framework captures this expected phenomena.

The differences in power scores show similar results in Figure 5. These differences were calculated only for the stories where both higher and lower power figures had been scored. The mean and median were both positive, 0.26 and 0.24.

Analyzing the stories with both positive and negative score differences (see Table 3 in the Appendix) further confirms the results of our framework. For example, in the third story of the table, despite the assigned role as an interviewer, Emily shows greater power. In the sentence "Paul thanked Emily for her time and wished her luck," both *thank* and *wish* from our lexicon give Emily greater power than Paul. We are thus able to analyze stories more accurately looking beyond just assigned roles.

4.2 Gender Differences in *Pride and Prejudice*

Jane Austen's 1813 novel *Pride and Prejudice* is famous for its depiction of gender and class relations in 19th century England. By examining the power associations with third person pronouns (feminine, masculine, and plural), we can trace how gender hierarchies are enacted and subverted by Austen, through the actions of her characters.

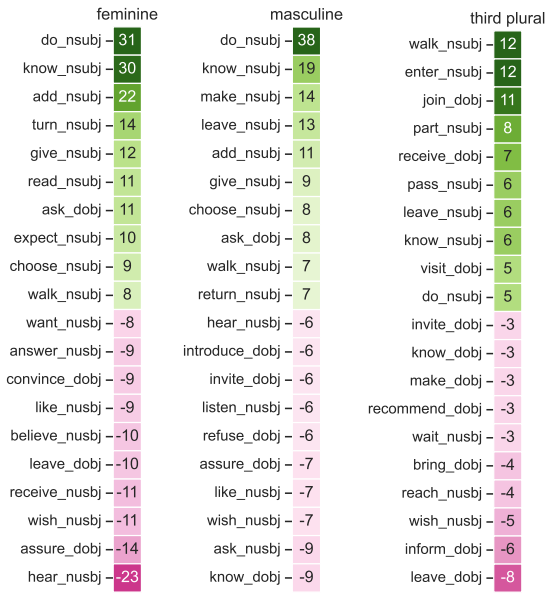

Figure 6: Verb counts for pronoun groups in *Pride and Prejudice*, using our package's visualization functions.

Results We find that, overall, feminine pronouns occur more frequently with low-power verbs (normalized power score of 0.032) in the novel than masculine pronouns (0.05) or third person plural pronouns (0.099) Our package allows us to interrogate these results further; figure 6 shows the verbs from Sap et al. (2017) contributing most frequently to each entity's power score. For example, we see that feminine pronouns are frequently used as subjects of the verb *hear*—emphasizing women's low-power role in this novel of waiting to hear news—while masculine pronouns are used with *refuse*—emphasizing men's lower-power role in their marriage proposals being refused.

We also observe that while feminine pronouns are often used in power relationships to verbs at rates similar to masculine pronouns, they have higher frequencies for low power relationships to verbs. In other words, in Austen's world, masculine and feminine entities both engage in high-power actions, but feminine entities engage in more lower-power actions, driving down their overall power scores. Arguably, though, some of the low-power positions are used by the feminine entities to obtain power, e.g., by *hearing* news or eavesdropping on others, the feminine entities can learn information that informs their future decisions and strategies.

5 Ethics and Broader Impact

RIVETER comes with some risks and limitations. This package is targeted only at English-language texts; we have not included non-English lexicons in the tool nor do we expect the parsing, named entity recognition, and coreference resolution to directly translate to other languages. While related lexica exist for non-English languages (e.g., Klenner et al. (2017) (German), Rashkin et al. (2017b) (extension to 10 European languages)), the generalizability of RIVETER is limited to English-language settings.

The results of RIVETER are only as reliable as the corpora and lexica used as input (and their relationships to one another). We have emphasized interpretability in designing this package, encouraging users to examine their results at different levels of granularity. However, there are still dangers of biases "baked-in" to the data, via its sampling and curation, or to the lexica, in the choice of terms and their annotations by human workers. Lexica that are useful in one setting are not always useful in other settings, and we encourage researchers to validate their lexica on their target corpora.

Drawing from a framework describing the roles for computational tools in social change (Abebe et al., 2020), we believe that RIVETER can fill multiple important roles. First, it can act as a *diagnostic*, measuring social biases and hierarchies across large corpora, as in Mendelsohn et al. (2020) where dehumanization of LGBTQ+ people was measured across news datasets and time. RIVETER can also act as a *formalizer*, allowing researchers to examine the specific words used by authors, adding concrete and fine-grained evidence to the constructions of broader patterns, as in Antoniak et al. (2019) where the supporting words were used to characterize healthcare roles during childbirth. Finally, RIVETER can act as a *synecdoche* by bringing new attention and perspectives to long-recognized problems, as in Sap et al. (2017) where renewed attention was given to the problem of gender biases in film.

References

Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. 2020. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 252–260, New York, NY, USA. Association for Computing Machinery.

Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1889–1904, Online. Association for Computational Linguistics.

Maria Antoniak, David Mimno, and Karen Levy. 2019. Narrative paths and negotiation of power in birth stories. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Elizabeth Behm-Morawitz and Dana E Mastro. 2008. Mean girls? the influence of gender portrayals in teen movies on emerging adults' gender-based attitudes and beliefs. *Journalism & Mass Communication Quarterly*, 85(1):131–146.

Amy M Blackstone. 2003. Gender roles and society. page 335.

Conrad Borchers, Dalia Gala, Benjamin Gilburt, Eduard Oravkin, Wilfried Bounsi, Yuki M Asano, and Hannah Kirk. 2022. Looking for a handsome carpenter! debiasing gpt-3 job advertisements. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 212–224.

Mina Cikara and Susan T Fiske. 2009. *Warmth, competence, and ambivalent sexism: Vertical assault and collateral damage*, volume 21. American Psychological Association, xvii, Washington, DC, US.

Haibo Ding and Ellen Riloff. 2018. Weakly supervised induction of affective events by optimizing semantic consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4):51–58.

Gail T. Fairhurst. 2005. Reframing the art of framing: Problems and prospects for leadership. *Leadership*, 1:165 – 185.

Anjalie Field, Gayatri Bhat, and Yulia Tsvetkov. 2019. Contextual affective analysis: A case study of people portrayals in online #metoo stories. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 158–169.

Charles J. Fillmore. 1976. Frame semantics and the nature of language*. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Susan T Fiske. 1993. Controlling other people. the impact of power on stereotyping. *American psychologist*, 48(6):621–628.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Salvatore Giorgi, Ke Zhao, Alexander H Feng, and Lara J Martin. 2023. Author as character and narrator: Deconstructing personal narratives from the r/amitheasshole reddit community. *arXiv preprint arXiv:2301.08104*.

Xiaoyun Gong, Yuxi Lin, Ye Ding, and Lauren Klein. 2022. Gender and power in japanese light novels. *Proceedings http://ceur-ws. org ISSN*, 1613:0073.

Philip Gorinski and Mirella Lapata. 2015. Movie script summarization as graph-based scene extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1066–1076.

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

Daniel Jurafsky and James Martin. 2023. *Speech and Language Processing, 3rd Edition*, 3 edition. Prentice Hall.

Punam Anand Keller, Isaac M Lipkus, and Barbara K Rimer. 2003. Affect, framing, and persuasion. *J. Mark. Res.*, 40(1):54–64.

Manfred Klenner, Don Tuggener, and Simon Clematide. 2017. Stance detection in Facebook posts of a German right-wing party. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 31–40, Valencia, Spain. Association for Computational Linguistics.

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

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):2332858420940312.

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. Powertransformer: Unsupervised controllable revision for biased language correction. In *EMNLP*.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in artificial intelligence*, 3:55.

Joakim Nivre. 2010. Dependency parsing. *Language and Linguistics Compass*, 4(3):138–152.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2021. Training language models to follow instructions with human feedback.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Chan Young Park, Xinru Yan, Anjalie Field, and Yulia Tsvetkov. 2021. Multilingual contextual affective analysis of lgbt people portrayals in wikipedia. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):479–490.

James W. Pennebaker, Roger J. Booth, Ryan L. Boyd, and Martha E. Francis. 2015. Linguistic inquiry and word count: LIWC 2015.

Hannah Rashkin, Eric Bell, Yejin Choi, and Svitlana Volkova. 2017a. Multilingual connotation frames: a case study on social media for targeted sentiment analysis and forecast. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 459–464.

Hannah Rashkin, Eric Bell, Yejin Choi, and Svitlana Volkova. 2017b. Multilingual connotation frames: A case study on social media for targeted sentiment analysis and forecast. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 459–464, Vancouver, Canada. Association for Computational Linguistics.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–321, Berlin, Germany. Association for Computational Linguistics.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Maarten Sap, Marcella Cindy Prasetio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *EMNLP*.

Devansh Saxena, Seh Young Moon, Dahlia Shehata, and Shion Guha. 2022. Unpacking invisible work practices, constraints, and latent power relationships in child welfare through casenote analysis. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–22.

Stephen M Smith and Richard E Petty. 1996. Message framing and persuasion: A message processing analysis. *Pers. Soc. Psychol. Bull.*, 22(3):257–268.

Maja Stahl, Maximilian Spliethöver, and Henning Wachsmuth. 2022. To prefer or to choose? generating agency and power counterfactuals jointly for gender bias mitigation. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+ CSS)*, pages 39–51.

Lucien Tesnière. 2015. *Elements of structural syntax*. John Benjamins Publishing Company.

Svitlana Volkova and Jin Yea Jang. 2018. Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In *Companion Proceedings of the The Web Conference 2018*, pages 575–583.

Michael Wiegand, Elisabeth Eder, and Josef Ruppenhofer. 2022. Identifying implicitly abusive remarks about identity groups using a linguistically informed approach. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5600–5612.

A Appendix: GPT-3.5 Generation Setup

Here we further detail our steps to evaluate our framework as discussed in Section 4.1. To generate characters with clear roles, names, and power differences, we used 32 dyadic relation pairs with explicit power asymmetry in our prompt. The full list of relations are shown in Table 2. The following is an example of the prompt used to generate such stories:

> Tell me a short story about a doctor and a patient, and give them names.
>
> doctor's name:

Using text-davinci-003 model, we generated 3 stories per pair with temperature set to 0.7 and max tokens set to 256. After cleaning ill-formatted results, we analyzed a total of 85 stories. A few examples of the generated stories along with the power scores of the characters are shown in Table 3.

(doctor, patient)	(teacher, student)
(interviewer, interviewee)	(parent, child)
(employer, employee)	(boss, subordinate)
(manager, worker)	(landlord, tenant)
(judge, defendant)	(supervisor, intern)
(therapist, client)	(owner, customer)
(mentor, mentee)	(politician, voter)
(rich, poor)	(elder, younger)
(artist, critic)	(host, guest)
(preacher, parishioner)	(expert, novice)
(counselor, advisee)	(coach, athlete)
(lender, borrower)	(king, subject)
(seller, buyer)	(umpire, player)
(printer, writer)	(sponsor, sponsored)
(producer, actor)	(legislator, citizen)
(commander, soldier)	(executor, beneficiary)

Table 2: Full list of dyadic roles

GPT3.5 Generated Story Dyadic Roles	Power Scores	Power Score Difference
Alan was a successful CEO of a large company. He had a sharp eye for detail and was known for his strict yet fair leadership style. One day, Alan noticed that one of his subordinates, Zara, was not performing up to the standards he expected. He called her into his office and asked her why her work wasn't up to par. Zara admitted that she was feeling overwhelmed with her workload and was having trouble keeping up. She asked Alan if he could help her prioritize her tasks and provide her with additional resources to help her complete her work. Alan was impressed by Zara's initiative and agreed to help her. He rearranged her tasks and assigned her additional help from other staff members. He also gave her some helpful tips on how to manage her workload more effectively. Zara was grateful for Alan's help and was able to complete her work on time and to a high standard. Alan was pleased with her performance and decided to reward her with a promotion. From then on, Zara and Alan had an excellent working relationship. Zara's hard work and resilience was an example to her colleagues, and Alan's willingness to help her out was a (boss, subordinate)	Alan: 0.4, Zara: -0.07	0.47
Mr. Jones was a landlord who owned an apartment building in a small town. He had been a landlord for many years, and he was known for his strict rules and no-nonsense attitude. Ms. Lopez was a single mother who had recently moved into the building. She was looking for a safe place to raise her young daughter, and Mr. Jones' building seemed to fit the bill. She was a bit intimidated by his strict rules, but she was willing to abide by them if it meant living in a safe and secure place. At first, things went smoothly between Mr. Jones and Ms. Lopez. He made sure the building was well maintained, and she was a model tenant, paying her rent on time every month. However, one day Ms. Lopez noticed that her hot water heater had stopped working. She immediately contacted Mr. Jones to let him know, but he refused to repair it. She was frustrated and angry, but she didn't want to risk getting evicted, so she kept quiet. After a few weeks, the hot water heater finally broke down completely and Ms. Lopez had no choice but to contact Mr. Jones again. (landlord, tenant)	Mr. Jones: 0.42, Ms. Lopez: 0.09	0.33
Paul was an experienced interviewer, but Emily was his most challenging interviewee yet. He had seen her resume online and was impressed with her qualifications. He had high hopes for her. When Emily arrived for the interview, Paul immediately noticed her confidence. She was wearing a suit and had a bright smile on her face. She was clearly prepared and ready for the interview. Paul asked Emily several questions about her past experience and qualifications. She answered each one with poise and detail. She was articulate and could back up her answers with examples. Paul was very impressed with Emily's performance. He could tell she was a strong candidate and he was confident she would do well in the job. At the end of the interview, Paul thanked Emily for her time and wished her luck. He was sure she would be a great addition to the company. Emily left the interview feeling pleased with how it went. She was confident she had done her best and was sure she had impressed Paul. The next day, Paul gave Emily an offer for the job. She happily accepted and started a few weeks later. Paul and Emily had a successful working (interviewer, interviewee)	Paul: 0.07, Emily: 0.25	-0.17

Table 3: Examples of GPT3.5 generated stories and the power scores of the characters