Men Are Elected, Women Are Married: Events Gender Bias on Wikipedia

Jiao Sun[1,2] and **Nanyun Peng**[1,2,3]

[1]Computer Science Department, University of Southern California
[2]Information Sciences Institute, University of Southern California
[3]Computer Science Department, University of California, Los Angeles
jiaosun@usc.edu, violetpeng@ucla.cs.edu

Abstract

Human activities can be seen as sequences of events, which are crucial to understanding societies. Disproportional event distribution for different demographic groups can manifest and amplify social stereotypes, and potentially jeopardize the ability of members in some groups to pursue certain goals. In this paper, we present the first event-centric study of gender biases in a Wikipedia corpus. To facilitate the study, we curate a corpus of career and personal life descriptions with demographic information consisting of 7,854 fragments from 10,412 celebrities. Then we detect events with a state-of-the-art event detection model, calibrate the results using strategically generated templates, and extract events that have asymmetric associations with genders. Our study discovers that Wikipedia pages tend to intermingle personal life events with professional events for females but not for males, which calls for the awareness of the Wikipedia community to formalize guidelines and train the editors to mind the implicit biases that contributors carry. Our work also lays the foundation for future works on quantifying and discovering event biases at the corpus level.

1 Introduction

Researchers have been using NLP tools to analyze corpora for various tasks on online platforms. For example, Pei and Jurgens (2020) found that female-female interactions are more intimate than male-male interactions on Twitter and Reddit. Different from social media, open collaboration communities such as Wikipedia have slowly won the trust of public (Young et al., 2016). Wikipedia has been trusted by many, including professionals in work tasks such as scientific journals (Kousha and Thelwall, 2017) and public officials in powerful positions of authority such as court briefs (Gerken, 2010). Implicit biases in such knowledge sources

Name	Wikipedia Description
Loretta Young (F)	**Career**: In 1930, when she was 17, she `eloped` with 26-year-old actor <u>Grant Withers</u>; they were `married` in Yuma, Arizona. The `marriage` was `annulled` the next year, just as their second movie together (ironically entitled Too Young to Marry) was `released` .
Grant Withers (M)	**Personal Life**: In 1930, at 26, he `eloped` to Yuma, Arizona with 17-year-old actress Loretta Young. The `marriage` ended in `annulment` in 1931 just as their second movie together, titled Too Young to Marry, was `released` .

Table 1: The marriage events are under the *Career* section for the female on Wikipedia. However, the same marriage is in the *Personal Life* section for the male. `yellow` background highlights events in the passage.

could have a significant impact on audiences' perception of different groups, thus propagating and even amplifying societal biases. Therefore, analyzing potential biases in Wikipedia is imperative.

In particular, studying events in Wikipedia is important. An event is a specific occurrence under a certain time and location that involves participants (Yu et al., 2015); human activities are essentially sequences of events. Therefore, the distribution and perception of events shape the understanding of society. Rashkin et al. (2018) discovered implicit gender biases in film scripts using events as a lens. For example, they found that events with female agents are intended to be helpful to other people, while events with male agents are motivated by achievements. However, they focused on the intentions and reactions of events rather than events themselves.

In this work, we propose to use events as a lens to study gender biases and demonstrate that events are more efficient for understanding biases in corpora than raw texts. We define *gender bias* as the

asymmetric association of events with females and males,[1] which may lead to gender stereotypes. For example, females are more associated with domestic activities than males in many cultures (Leopold, 2018; Jolly et al., 2014).

To facilitate the study, we collect a corpus that contains demographic information, personal life description, and career description from Wikipedia.[2] We first detect events in the collected corpus using a state-of-the-art event extraction model (Han et al., 2019). Then, we extract gender-distinct events with a higher chance to occur for one group than the other. Next, we propose a calibration technique to offset the potential confounding of gender biases in the event extraction model, enabling us to focus on the gender biases at the corpus level. Our contributions are three-fold:

- We contribute a corpus of 7,854 fragments from 10,412 celebrities across 8 occupations including their demographic information and Wikipedia *Career* and *Personal Life* sections.

- We propose using events as a lens to study gender biases at the corpus level, discover a mixture of personal life and professional life for females but not for males, and demonstrate the efficiency of using events in comparison to directly analyzing the raw texts.

- We propose a generic framework to analyze event gender bias, including a calibration technique to offset the potential confounding of gender biases in the event extraction model.

2 Experimental Setup

In this section, we will introduce our collected corpus and the event extraction model in our study.

Dataset. Our collected corpus contains demographics information and description sections of celebrities from Wikipedia. Table 2 shows the statistics of the number of celebrities with *Career* or *Personal Life* sections in our corpora, together with all celebrities we collected. In this work, we only explored celebrities with *Career* or *Personal Life* sections, but there are more sections (e.g., *Politics* and *Background and Family*) in our collected

	Career		Personal Life		Collected	
Occ	F	M	F	M	F	M
Acting	464	469	464	469	464	469
Writer	455	611	319	347	1,372	2,466
Comedian	380	655	298	510	642	1,200
Artist	193	30	60	18	701	100
Chef	81	141	72	95	176	350
Dancer	334	167	286	127	812	465
Podcaster	87	183	83	182	149	361
Musician	39	136	21	78	136	549
All	4,425		3,429		10,412	

Table 2: Statistics showing the number of celebrities with *Career* section or *Personal Life* section, together with all celebrities we collected. Not all celebrities have *Career* or *Personal Life* sections.

corpus. We encourage interested researchers to further utilize our collected corpus and conduct studies from other perspectives. In each experiment, we select the same number of female and male celebrities from one occupation for a fair comparison.

Event Extraction. There are two definitions of events: one defines an event as the trigger word (usually a verb) (Pustejovsky et al., 2003b), the other defines an event as a complex structure including a trigger, arguments, time, and location (Ahn, 2006). The corpus following the former definition usually has much broader coverage, while the latter can provide richer information. For broader coverage, we choose a state-of-the-art event detection model that focuses on detecting event trigger words by Han et al. (2019).[3] We use the model trained on the TB-Dense dataset (Pustejovsky et al., 2003a) for two reasons: 1) the model performs better on the TB-Dense dataset; 2) the annotation of the TB-Dense dataset is from the news articles, and it is also where the most content of Wikipedia comes from.[4] We extract and lemmatize events e from the corpora and count their frequencies $|e|$. Then, we separately construct dictionaries $\mathcal{E}^m = \{e_1^m : |e_1^m|, ..., e_M^m : |e_M^m|\}$ and $\mathcal{E}^f = \{e_1^f : |e_1^f|, ..., e_F^f : |e_F^f|\}$ mapping events to their frequency for male and female respectively.

Event Extraction Quality. To check the model performance on our corpora, we manually annotated events in 10,508 sentences (female: 5,543,

---

[1]In our analysis, we limit to binary gender classes, which, while unrepresentative of the real-world diversity, allows us to focus on more depth in analysis.

[2]https://github.com/PlusLabNLP/ee-wiki-bias

[3]We use the code at https://github.com/rujunhan/EMNLP-2019 and reproduce the model trained on the TB-Dense dataset.

[4]According to Fetahu et al. (2015), more than 20% of the references are news articles on Wikipedia.

Metric	TB-D	S	S-F	S-M
Precision	89.2	93.5	95.3	93.4
Recall	92.6	89.8	87.1	89.8
F1	90.9	91.6	91.0	91.6

Table 3: The performance for off-the-shelf event extraction model in both common event extraction dataset TB-Dense (TB-D) and our corpus with manual annotation. `S` represents the sampled data from the corpus. `S-F` and `S-M` represent the sampled data for female career description and male career description separately.

male: 4,965) from the Wikipedia corpus. Table 3 shows that the model performs comparably on our corpora as on the TB-Dense test set.

3 Detecting Gender Biases in Events

Odds Ratio. After applying the event detection model, we get two dictionaries $\mathcal{E}^m$ and $\mathcal{E}^f$ that have events as keys and their corresponding occurrence frequencies as values. Among all events, we focus on those with distinct occurrences in males and females descriptions (e.g., `work` often occurs at a similar frequency for both females and males in *Career* sections, and we thus neglect it from our analysis). We use the Odds Ratio (OR) (Szumilas, 2010) to find the events with large frequency differences for females and males, which indicates that they might potentially manifest gender biases. For an event e_n, we calculate its odds ratio as the odds of having it in the male event list divided by the odds of having it in the female event list:

$$\frac{\mathcal{E}^m(e_n)}{\sum_{\substack{e_i^m \neq e_n \\ i \in [1,...,M]}}^i \mathcal{E}^m(e_i^m)} \bigg/ \frac{\mathcal{E}^f(e_n)}{\sum_{\substack{e_j^f \neq e_n \\ j \in [1,...,F]}}^j \mathcal{E}^f(e_j^f)} \quad (1)$$

The larger the OR is, the more likely an event will occur in male than female sections by Equation 1. After obtaining a list of events and their corresponding OR, we sort the events by OR in descending order. The top k events are more likely to appear for males and the last k events for females.

Calibration. The difference of event frequencies might come from the model bias, as shown in other tasks (e.g., gender bias in coreference resolution model (Zhao et al., 2018)). To offset the potential confounding that could be brought by the event extraction model and estimate the actual event frequency, we propose a calibration strategy by 1)

generating data that contains target events; 2) testing the model performance for females and males separately in the generated data, 3) and using the model performance to estimate real event occurrence frequencies.

We aim to calibrate the top 50 most skewed events in females' and males' *Career* and *Personal Life* descriptions after using the OR separately. First, we follow two steps to generate a synthetic dataset:

1. For each target event, we select all sentences where the model successfully detected the target event. For each sentence, we manually verify the correctness of the extracted event and discard the incorrect ones. For the rest, we use the verified sentences to create more ground truth; we call them *template sentences*.

2. For each template sentence, we find the celebrity's first name and mark it as a `Name Placeholder`, then we replace it with 50 female names and 50 male names that are sampled from the name list by Ribeiro et al. (2020). If the gender changes during the name replacement (e.g., Mike to Emily), we replace the corresponding pronouns (e.g., he to she) and gender attributes (Zhao et al., 2018) (e.g., Mr to Miss) in the template sentences. As a result, we get 100 data points for each template sentence with automatic annotations. If there is no first name in the sentence, we replace the pronouns and gender attributes.

After getting the synthetic data, we run the event extraction model again. We use the detection recall among the generated instances to calibrate the frequency $|e|$ for each target event and estimate the actual frequency $|e|^*$, following:

$$|e|^* = \frac{|e|}{TP(e)/(TP(e) + FP(e))} \quad (2)$$

Then, we replace $|e|$ with $|e|^*$ in Equation 1, and get k female and k male events by sorting OR as before. Note that we observe the model performances are mostly unbiased, and we have only calibrated events that have different performances for females and males over a threshold (i.e., 0.05).[6]

---

[5]We did not show the result for the `artists` and `musicians` due to the small data size.

[6]Calibration details and quantitative result in App. A.2.

Occupation	Events in Female Career Description	Events in Male Career Description	WEAT*	WEAT
Writer	♦ divorce, ♦ marriage, involve, organize, ♦ wedding	argue, ⊕ election, ▲ protest, rise, shoot	-0.17	1.51
Acting	♦ divorce, ♦ wedding, guest, name, commit	support, ▲ arrest, ▲ war, ■ sue, trial	-0.19	0.88
Comedian	♦ birth, eliminate, ♦ wedding, ♥ relocate, partner	enjoy, hear, cause, ● buy, conceive	-0.19	0.54
Podcaster	♥ land, interview, portray, ♦ married, report	direct, ask, provide, continue, bring	-0.24	0.53
Dancer	♦ married, ♦ marriage, ♥ depart, ♥ arrive, organize	drop, team, choreograph, explore break	-0.14	0.22
Artist	paint, exhibit, include, ♥ return, teach	start, found, feature, award, begin	-0.02	0.17
Chef	⊕ hire, △ meet, debut, eliminate, sign	include, focus, explore, award, ● raise	-0.13	-0.38
Musician	run, record, ♦ death, found, contribute	sign, direct, produce, premier, open	-0.19	–0.41
Annotations: ♦ Life ♥ Transportation ⊕ Personell ▲ Conflict ■ Justice ● Transaction △ Contact				

Table 4: Top 5 extracted events that occur more often for females and males in *Career* sections across 8 occupations. We predict event types by applying EventPlus (Ma et al., 2021) on sentences that contain target events and take the majority vote of the predicted types. The event types are from the ACE dataset.[5] We calculate WEAT scores with all tokens excluding stop words (WEAT* column) and only detected events (WEAT column) for *Career* sections.

(a) Male Writers (b) Female Writers (c) Actor (d) Actress
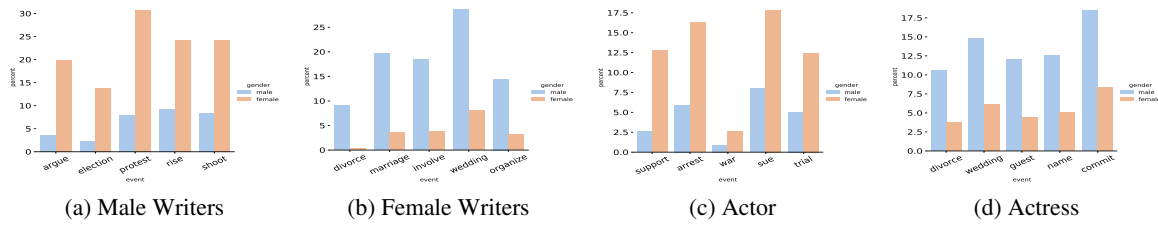
Figure 1: The percentile of extracted events among all detected events, sorted by their frequencies in descending order. The smaller the percentile is, the more frequent the event appears in the text. The extracted events are among the top 10% for the corresponding gender (e.g., extracted female events among all detected events for female writers) and within top 40% percent for the opposite gender (e.g., extracted female events among all detected events for male writers). The figure shows that we are not picking rarely-occurred events, and the result is significant.

WEAT score. We further check if the extracted events are associated with gender attributes (e.g., `she and her` for females, and `he and him` for males) in popular neural word embeddings like Glove (Pennington et al., 2014). We quantify this with the Word Embedding Association Test (WEAT) (Caliskan et al., 2017), a popular method for measuring biases in text. Intuitively, WEAT takes a list of tokens that represent a concept (in our case, *extracted events*) and verifies whether these tokens have a shorter distance towards female attributes or male attributes. A positive value of WEAT score indicates that female events are closer to female attributes, and male events are closer to male attributes in the word embedding, while a negative value indicates that female events are closer to male attributes and vice versa.[7]

To show the effectiveness of using events as a lens for gender bias analysis, we compute WEAT scores on the raw texts and detected events separately. For the former, we take all tokens excluding stop words.[8] Together with gender attributes from Caliskan et al. (2017), we calculate and show the WEAT scores under two settings as "WEAT*" for the raw texts and "WEAT" for the detected events.

4 Results

The Effectiveness of our Analysis Framework. Table 4 and Table 5 show the associations of both raw texts and the extracted events in *Career* and *Personal Life* sections for females and males across occupations after the calibration. The values in WEAT* columns in both tables indicate that there

---

[7]Details of WEAT score experiment in App. A.4.

[8]We use spaCy (`https://spacy.io/`) to tokenize the corpus and remove stop words.

Occupation	Events in Female Personal Life Description	Events in Male Personal Life Description	WEAT*	WEAT
Writer	bury, ♦ birth, attend, ▲ war, grow	know, report, come, ■ charge, publish	-0.05	0.31
Acting	♦ pregnant, practice, wedding, record, convert	accuse, ♥ trip, ♥ fly, ▲ assault, endorse	-0.14	0.54
Comedian	feel, ♦ birth, fall, open, decide	♥ visit, create, spend, propose, lawsuit	-0.07	0.07
Podcaster	date, describe, tell, life, come	play, write, ♦ born, release, claim	-0.13	0.57
Dancer	♦ marry, describe, diagnose, expect, speak	hold, involve, ● award, run, serve	-0.03	0.41
Chef	♦ death, serve, announce, describe, ♦ born	♦ birth, lose, ♦ divorce, speak, △ meet	-0.02	-0.80
Annotations: ♦ Life ♥ Transportation ⊕ Personell ▲ Conflict ■ Justice ● Transaction △ Contact				

Table 5: Top 5 events in *Personal Life* section across 6 occupations.[9] There are more `Life` events (e.g., "birth" and "marry") in females' personal life descriptions than males' for most occupations. While for males, although we see more life-related events than in the *Career* section, there are events like "awards" even in the *Personal Life* section. The findings further show our work is imperative and addresses the importance of not intermingling the professional career with personal life regardless of gender during the future editing on Wikipedia.

was only a weak association of words in raw texts with gender. In contrast, the extracted events are associated with gender for most occupations. It shows the effectiveness of the event extraction model and our analysis method.

The Significance of the Analysis Result. There is a possibility that our analysis, although it picks out distinct events for different genders, identifies the events that are infrequent for all genders and that the frequent events have similar distributions across genders. To verify, we sort all detected events from our corpus by frequencies in descending order. Then, we calculate the percentile of extracted events in the sorted list. The smaller the percentile is, the more frequent the event appears in the text. Figure 1 shows that we are not picking the events that rarely occur, which shows the significance of our result.[10] For example, Figure 1a and Figure 1b show the percentile of frequencies for selected male and female events among all events frequencies in the descending order for male and female writers, respectively. We can see that for the corresponding gender, event frequencies are among the top 10%. These events occur less frequently for the opposite gender but still among the top 40%.

Findings and Discussions. We find that there are more `Life` events for females than males in both *Career* and *Personal Life* sections. On the other hand, for males, there are events like "awards" even in their *Personal Life* section. The mixture of personal life with females' professional career events and career achievements with males' personal life events carries implicit gender bias and re-

inforces the gender stereotype. It potentially leads to career, marital, and parental status discrimination towards genders and jeopardizes gender equality in society. We recommend: 1) Wikipedia editors to restructure pages to ensure that personal life-related events (e.g., marriage and divorce) are written in the *Personal Life* section, and professional events (e.g., award) are written in *Career* sections regardless of gender; 2) future contributors should also be cautious and not intermingle *Personal Life* and *Career* when creating the Wikipedia pages from the start.

5 Conclusion

We conduct the first event-centric gender bias analysis at the corpus level and compose a corpus by scraping Wikipedia to facilitate the study. Our analysis discovers that the collected corpus has event gender biases. For example, personal life related events (e.g., `marriage`) are more likely to appear for females than males even in *Career* sections. We hope our work brings awareness of potential gender biases in knowledge sources such as Wikipedia, and urges Wikipedia editors and contributors to be cautious when contributing to the pages.

Acknowledgments

---

[10]See plots for all occupations in Appendix A.5.

Ethical Considerations

Our corpus is collected from Wikipedia. The content of personal life description, career description, and demographic information is all public to the general audience. Note that our collected corpus might be used for malicious purposes. For example, it can serve as a source by text generation tools to generate text highlighting gender stereotypes.

This work is subject to several limitations: First, it is important to understand and analyze the event gender bias for gender minorities, missing from our work because of scarce resources online. Future research can build upon our work, go beyond the binary gender and incorporate more analysis. Second, our study focuses on the Wikipedia pages for celebrities for two additional reasons besides the broad impact of Wikipedia: 1) celebrities' Wikipedia pages are more accessible than non-celebrities. Our collected Wikipedia pages span across 8 occupations to increase the representation of our study; 2) Wikipedia contributors have been extensively updating celebrities' Wikipedia pages every day. Wikipedia develops at a rate of over 1.9 edits every second, performed by editors from all over the world (wik, 2021). The celebrities' pages get more attention and edits, thus better present how the general audience perceives important information and largely reduce the potential biases that could be introduced in personal writings. Please note that although we try to make our study as representative as possible, it cannot represent certain groups or individuals' perceptions.

Our model is trained on TB-Dense, a public dataset coming from news articles. These do not contain any explicit detail that leaks information about a user's name, health, negative financial status, racial or ethnic origin, religious or philosophical affiliation or beliefs, trade union membership, alleged or actual crime commission.

References

2021. Wikipedia:statistics - wikipedia. https://en.wikipedia.org/wiki/Wikipedia:Statistics. (Accessed on 02/01/2021).

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8.

A. Caliskan, J. Bryson, and A. Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356:183 – 186.

Besnik Fetahu, Katja Markert, and Avishek Anand. 2015. Automated news suggestions for populating wikipedia entity pages. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 323–332.

Joseph L. Gerken. 2010. How courts use wikipedia. *The Journal of Appellate Practice and Process*, 11:191.

Rujun Han, Qiang Ning, and Nanyun Peng. 2019. Joint event and temporal relation extraction with shared representations and structured prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 434–444, Hong Kong, China. Association for Computational Linguistics.

S. Jolly, K. Griffith, R. Decastro, A. Stewart, P. Ubel, and R. Jagsi. 2014. Gender differences in time spent on parenting and domestic responsibilities by high-achieving young physician-researchers. *Annals of Internal Medicine*, 160:344–353.

K. Kousha and M. Thelwall. 2017. Are wikipedia citations important evidence of the impact of scholarly articles and books? *Journal of the Association for Information Science and Technology*, 68.

T. Leopold. 2018. Gender differences in the consequences of divorce: A study of multiple outcomes. *Demography*, 55:769 – 797.

Mingyu Derek Ma, Jiao Sun, Mu Yang, Kung-Hsiang Huang, Nuan Wen, Shikhar Singh, Rujun Han, and Nanyun Peng. 2021. Eventplus: A temporal event understanding pipeline. In *2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Demonstrations Track*.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5307–5326, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003a. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa

Ferro, et al. 2003b. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

M. Szumilas. 2010. Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie canadienne de psychiatrie de l'enfant et de l'adolescent*, 19 3:227–9.

A. Young, Ari D. Wigdor, and Gerald Kane. 2016. It's not what you think: Gender bias in information about fortune 1000 ceos on wikipedia. In *ICIS*.

Mo Yu, Matthew R. Gormley, and Mark Dredze. 2015. Combining word embeddings and feature embeddings for fine-grained relation extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1374–1379, Denver, Colorado. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

A Appendix

A.1 Quality Check: Event Detection Model

To test the performance of the event extraction model in our collected corpus from Wikipedia. We manually annotated events in 10,508 (female: 5,543, male: 4,965) sampled sentences from the *Career* section in our corpus. Our annotators are two volunteers who are not in the current project but have experience with event detection tasks. We asked annotators to annotate all event trigger words in the text. During annotation, we follow the definition of events from the ACE annotation guideline.[11] We use the manual annotation as the ground truth and compare it with the event detection model output to calculate the metrics (i.e., precision, recall and F1) in Table 3.

A.2 Calibration Details

To offset the potential confounding that could be brought by the event extraction model and estimate the actual event frequency of $|e|^*$, we use the recall for the event e to calibrate the event frequency $|e|$ for females and males separately. Figure 2 shows the calibration result for the 20 most frequent events in our corpus. Please note that Figure 2 (a)-(h) show the quantitative result for extracted events in the *Career* sections across 8 occupations, and Figure 2 (i)-(n) for the *Personal Life* sections.

Example Sentence Substitutions for Calibration. After checking the quality of selected sentences containing the target event trigger, we use 2 steps described in Section 3 *Calibration* to compose a synthetic dataset with word substitutions. Here is an example of using `Name Placeholder`: for target event trigger "married" in `Carole Baskin`'s *Career* section, we have:

> At the age of 17, Baskin worked at a Tampa department store. To make money, she began breeding show cats; she also began rescuing bobcats, and used llamas for a lawn trimming business. In January 1991, she married her second husband and joined his real estate business.

First, we mark the first name Baskin as `Name Placeholder` and find all gender attributes and

---
[11]https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf

pronouns which are consistent with the celebrity's gender. Then, we replace Baskin with 50 female names and 50 male names from Ribeiro et al. (2020). If the new name is a male name, we change the corresponding gender attributes (none in this case) and pronouns (e.g., she to he, her to his).

Another example is for the context containing the target event trigger "married" in `Indrani Rahman`'s *Career* section, where there is no first name:

> In 1952, although married, and with a child, she became the first Miss India, and went on to compete in the Miss Universe 1952 Pageant, held at Long Beach, California. Soon, she was travelling along with her mother and performing all over the world...

We replace all pronouns (she to he, her to his) and gender attributes (Miss to Mr).

Interpret the Quantitative Calibration Result. We use the calibration technique to calibrate potential gender biases from the model that could have complicated the analysis. In Figure 2, we can see that there is little gender bias at the model level: the model has the same performance for females and males among most events.

Besides, we notice that the model fails to detect and has a low recall for few events in the generated synthetic dataset. We speculate that this is because of the brittleness in event extraction models triggered by the word substitution. We will leave more fine-grained analysis at the model level for future work. We focus on events for which the model performs largely different for females and males during our calibration. Thus, we select and focus on the events that have different performance for females and males over a threshold, which we take 0.05 during our experiment, to calibrate the analysis result.

A.3 Top Ten Extracted Events

Table 6 and Table 7 show the top 10 events and serves as the supplement of top 5 events that we reported for *Career* and *Personal Life* sections.

A.4 Details for Calculating WEAT Score

The WEAT score is in the range of -2 to 2. A high positive score indicates that extracted events for females are more associated with female attributes in the embedding space. A high negative score means that extracted events for females are more

Occupation	Events in Female Career Description	Events in Male Career Description
Writer	divorce, marriage, involve, organize, wedding, donate, fill, pass, participate, document	argue, election, protest, rise, shoot, purchase, kill, host, close, land
Acting	divorce, wedding, guest, name, commit, attract, suggest, married, impressed, induct	support, arrest, war, sue, trial, vote, pull, team, insist, like
Comedian	birth, eliminate, wedding, relocate, partner, pursue, impersonate, audition, guest, achieve	enjoy, hear, cause, buy, conceive, enter, injury, allow, acquire, enter
Podcaster	land, interview, portray, married, report, earn, praise, talk, shoot, premier	direct, ask, provide, continue, bring, election, sell, meet, read, open
Dancer	married, marriage, depart, arrive, organize try, promote, train, divorce, state	drop, team, choreograph, explore, break, think, add, celebrate, injury, suffer
Artist	paint, exhibit, include, return, teach, publish, explore, draw, produce, write	start, found, feature, award, begin, appear, join, influence, work, create
Chef	hire, meet, debut, eliminate, sign, graduate, describe, train, begin, appear	include, focus, explore, award, raise, gain, spend, find, launch, hold
Musician	run, record, death, found, contribute, continue, perform, teach, appear, accord	sign, direct, produce, premier, open, announce, follow, star, act, write

Table 6: The top 10 extracted events in *Career* section.

Occupation	Events in Female Personal Life Description	Events in Male Personal Life Description
Writer	bury, birth, attend, war, grow, serve, appear, raise, begin, divorce	know, report, come, charge, publish, claim, suffer, return, state, describe
Acting	pregnant, practice, wedding, record, convert, honor, gain, retire, rap, bring	accuse, trip, fly, assault, endorse, meeting, donate, fight, arrest, found
Comedian	feel, birth, fall, open, decide, date, diagnose, tweet, study, turn	visit, create, spend, propose, lawsuit, accord, arrest, find, sell, admit
Podcaster	date, describe, tell, life, come, leave, engage, live, start, reside	play, write, bear, release, claim, birth, divorce, meet, announce, work
Dancer	marry, describe, diagnose, expect, speak, post, attend, come, play, reside	hold, involve, award, run, serve, adopt, charge, suit, struggle, perform
Chef	death, serve, announce, describe, born, die, life, state, marriage, live	birth, lose, divorce, speak, meet, work, diagnose, wedding, write, engage

Table 7: The top 10 extracted events in *Personal Life* section.

associated with male attributes. To calculate the WEAT score, we input two lists of extracted events for females E_f and males E_m, together with two lists of gender attributes A and B, then calculate:

$$S(E_f, E_m, A, B) = \sum_{e_f \in E_f} s(e_f, A, B) - \sum_{e_m \in E_m} s(e_m, A, B),$$
(3)

where

$$s(w, A, B) = \text{mean}_{a \in A} cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} cos(\vec{w}, \vec{b}).$$
(4)

Following Caliskan et al. (2017), we have "female, woman, girl, sister, she, her, hers, daughter" as female attribute list A and "male, man, boy, brother, he, him, his, son" as male attributes list B. To calculate WEAT*, we replace the input lists E_f and E_m with all non-stop words tokens in raw texts from either *Career* section or *Personal Life* section.

A.5 Extracted Events Frequency Distribution

We sort all detected events from our corpus by their frequencies in descending order according to Equation 1. Figure 3 (a)-(l) show the percentile of extracted events in the sorted list for another 6 occupations besides the 2 occupations reported in Figure 1 for *Career* section. The smaller the percentile is, the more frequent the event appears in the text. These figures indicate that we are not picking events that rarely occur and showcase the significance of our analysis result. Figure 3 (m)-(x) are for *Personal Life* sections across 6 occupations, which show the same trend as for *Career* sections.

(a) Writers-*c*   (b) Comedians-*c*   (c) Podcasters-*c*   (d) Dancers-*c*

(e) Artists-*c*   (f) Musicians-*c*   (g) Acting-*c*   (h) Chefs-*c*

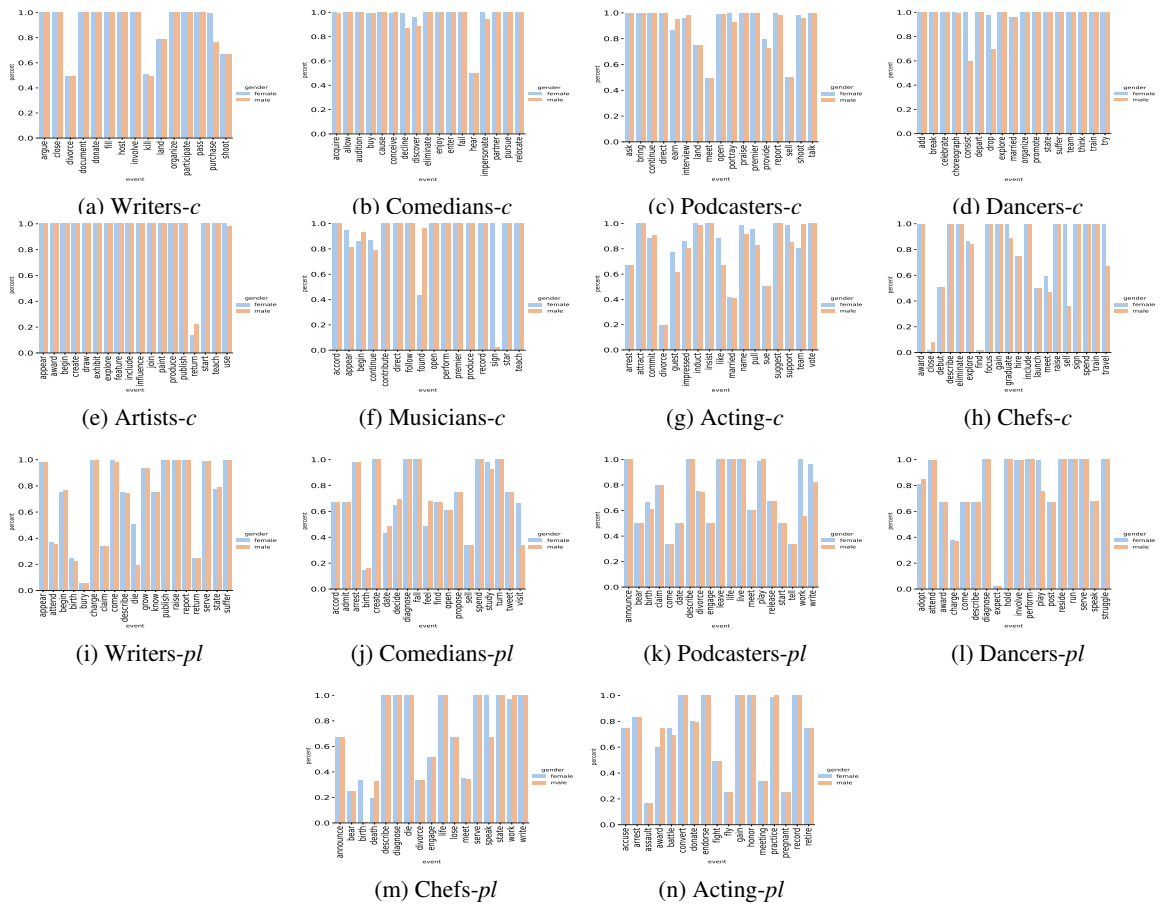(i) Writers-*pl*   (j) Comedians-*pl*   (k) Podcasters-*pl*   (l) Dancers-*pl*

(m) Chefs-*pl*   (n) Acting-*pl*

Figure 2: Detection recall on the strategically-generated data. (*c*: *Career* section, *pl*: *Personal Life* section)

(a) Male Comedians-*c*   (b) Female Comedians-*c*   (c) Male Dancers-*c*   (d) Female Dancers-*c*

(e) Male Podcasters-*c*   (f) Female Podcasters-*c*   (g) Male Chefs-*c*   (h) Female Chefs-*c*

(i) Male Artists-*c*   (j) Female Artists-*c*   (k) Male Musicians-*c*   (l) Female Musicians-*c*

(m) Male Comedians-*pl*   (n) Female Comedians-*pl*   (o) Male Dancers-*pl*   (p) Female Dancers-*pl*

(q) Male Podcasters-*pl*   (r) Female Podcasters-*pl*   (s) Male Chefs-*pl*   (t) Female Chefs-*pl*

(u) Male Writers-*pl*   (v) Female Writers-*pl*   (w) Actors-*pl*   (x) Actress-*pl*
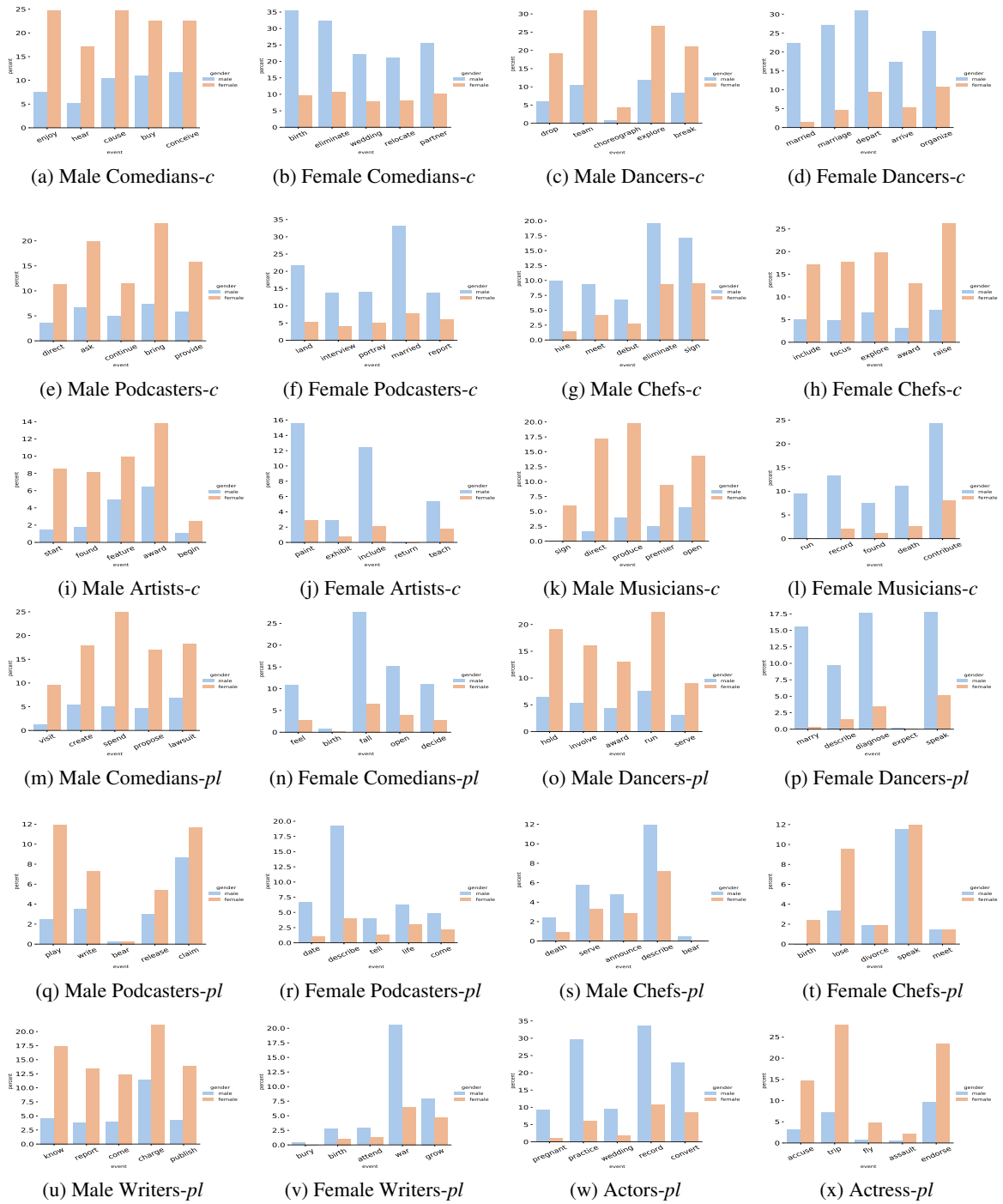
Figure 3: The percentile of extracted event frequencies. (*c*: *Career* section, *pl*: *Personal Life* section)