Disarming Loaded Words: Addressing Gender Bias in Political Reporting

Irena Fischer-Hwang, Dylan Grosz, Xinlan Emily Hu, Anjini Karthik, Vivian Yang*
{ihwang,dgrosz,xehu,anjinik,vivianca}@stanford.edu

ABSTRACT

Gender bias has been a pervasive issue in U.S. political coverage since women gained the franchise in 1920. Women are routinely targeted for their appearance, likeability, and familial qualities—characteristics for which men are rarely scrutinized. As a way to help journalists and editors identify and correct gender-biased language in political reporting, we introduce "Disarming Loaded Words" (DLW) as a new solution to this issue. DLW is a computational tool, backed by both machine learning and human expert curation, that tags potentially biased words in a document and provides feedback and context for why a word may be problematic. DLW provides a minimally disruptive way to nudge journalists toward understanding unconscious biases. It was successfully prototyped as a standalone Google Docs add-on, and the code for its model, interface, and API are now open-source and available for use on other platforms.

CCS CONCEPTS

• **Applied computing** → **Text editing**.

KEYWORDS

politics, gender, machine learning, word embeddings

ACM Reference Format:
Irena Fischer-Hwang, Dylan Grosz, Xinlan Emily Hu, Anjini Karthik, Vivian Yang. 2020. Disarming Loaded Words: Addressing Gender Bias in Political Reporting. In *Computation + Journalism '20: Computation + Journalism Symposium, March 20–21, 2020, Boston, MA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.1122456

*All authors contributed equally to this research, and the order of the names is alphabetical.

Computation + Journalism '20, March 20–21, 2020, Boston, MA

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Over the past century, female politicians have soared to increasing levels of prominence [19]. Yet even as women have sought to make themselves heard in the political sphere, political coverage in journalism continues to exhibit gender bias. Since the 1980s, female politicians have been asked irrelevant queries, such as "Does she bake muffins?" [20]. In 2016, Hillary Clinton endured unnecessary commentary about her vocal range [18], comportment [24], "likeability" [17] and body language [21] from the media. More recently, Elizabeth Warren has also had her likeability scrutinized [23], and Amy Klobuchar has been cast as paranoid and demanding by the New York Times [22].

Despite persistent issues of gender bias, the current industry standard is to ask "Would you make this comment about a man?" [9].

We suggest that the field of journalism can benefit from a computational approach to "disarm" (that is, tag and provide explanations for) words that perpetuate gender bias in political reporting. We propose a tool that uses machine learning techniques to identify words and phrases that have historically been used in a gender-biased way — and perhaps shed light on new biases in political reporting.

2 METHODOLOGY

Identifying gender bias in text is a complex problem because word meanings can change over time and depend heavily on context. We view a word as biased if it is used to describe a woman and carries a negative connotation, but would not be used to describe a man in a similar political context. Such words are inevitably informed by historical use and contemporary social context (i.e., stereotypes). As a result, we first approached this problem by interviewing industry experts and reviewing previous work on the topic.

We conducted interviews with staff from PolitiFact, The Washington Post, The Poynter Institute for Media Studies, and the School of Journalism and Communication at the University of Oregon. We also consulted the Women's Media Center's *Media Guide to Gender Neutral Coverage of Women Candidates + Politicians* [2], and the Bem Sex-Role Inventory [4]. Finally, we asked journalists to identify problematic words and phrases in four articles that were about women involved in the 2008 [8], 2016 [3], and upcoming 2020 presidential election [16]. Through these sources, we developed

a corpus of 70 words that have been identified as problematic when used in a gendered context, especially in political reporting.

To account for the fact that many of the words in our corpus are frequently found in a non-gendered context, we further split the corpus into two lists: one "hard-coded" list that included exceedingly gendered words that should not be used in political reporting under any circumstance, and one list of contextually biased words. The first list currently includes only ten words: *bimbo, bitch, catty, cleavage, feisty, frump, matronly, pussy, vagina*, and *whore*. The second list, which we refer to as the list of "contextually biased" words, includes the remaining words in our corpus. These words include terms like "shrill," which may be used to describe both the sound produced by an inanimate object — e.g. "The fire alarm made a shrill noise" — or may be loaded with decades of sexism — e.g. "Her voice was shrill."

In tandem with developing and fine-tuning a word corpus, we also developed a tool that journalists may use to easily identify potentially problematic words in their work. We describe the tool in detail in the following subsection.

A Computational Tool

We introduce a computational tool, "Disarming Loaded Words" (DLW), designed to identify and tag problematic words in political articles. The tool both identifies historically problematic words and uncovers words that may be used in a biased manner as language evolves.

To achieve these goals, the tool receives as input a block of text (an article) and iterates through the text word-by-word. For each word, the tool performs the following analysis:

(1) Check if the word is in the hard-coded ("always biased") list. If it is, tag the word as biased.
(2) If the word is not in the hard-coded list, check the context of the word.
(3) If the word is being used to modify a woman, or the actions of a woman, check if the word is in the list of contextually biased words.
(4) If the word is in the list of contextually biased words, tag the word as biased.
(5) If the word is not in the list of contextually biased words, calculate a bias score for the word using a machine learning model. If the bias score is above a certain threshold t_b, tag the word as biased.

The core of DLW is a word2vec model [12] trained on a repository of articles from Google News. Our work draws from Bolukbasi et al.'s prior analysis of gender differences quantified in word embeddings (e.g., $\vec{man} - \vec{woman}$; $\vec{father} - \vec{mother}$) [5]. This work found, among other surprising results, that $\vec{man} - \vec{woman} = \vec{computer\ programmer} - \vec{homemaker}$. Instead of performing the analogy comparison task used by Bolukbasi et al., we condensed the gender bias of each individual word into a single score.

To create this score, our model defines "feminine" and "masculine" bias directions using a dimension-reduced vector or subspace of a word embedding set containing female basis words, such as "woman," "girl," "herself," etc. Each word passed into the model is then compared to the bias direction in order to determine its gender association. The word is projected onto the dimension-reduced vector or subspace, and the resulting projection coefficient is used to generate a bias score relative to each gender bias direction. Each word is then tagged with the respective gender of the bias score that is larger. Words whose bias scores exceeded some bias threshold t_b are considered suspect "loaded" terms.

Finally, to improve the model's recall, suspect words are cleaved off through analysis of the word's part of speech, the gender of nouns modified by the word, and by other, more sophisticated natural language sentence graphing techniques using a dependency parser [1]. Due to short-term dependency conflicts, the current version of the model tied to the front-end via an API does not include this cleaving, which is still currently under development.

Our process is summarized in the following figure.

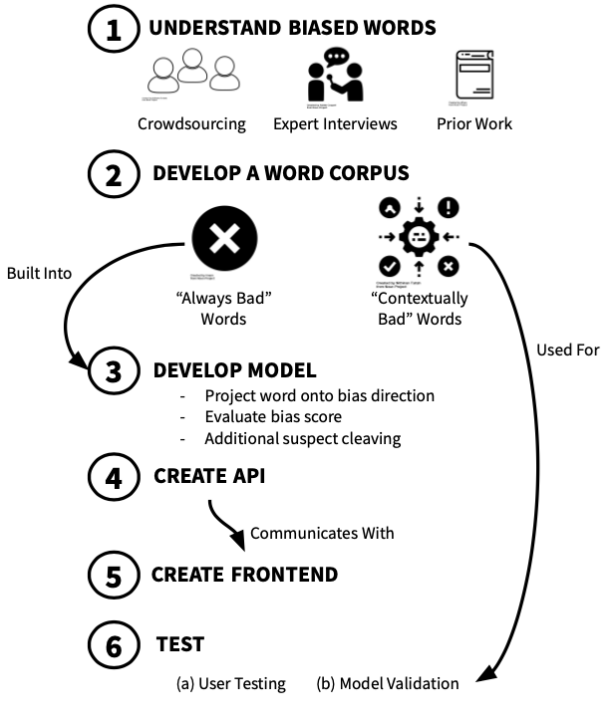

Figure 1: Summary of our methodology.

3 RESULTS

DLW has been packaged as a Google Docs add-on that enables journalists to identify potentially gender-biased words in a document as they compose the article. Our backend machine learning model, described in the previous section, is also available as an API package that enables integration with any other platform.

Crucially, our user interface provides not only a visual tagging of potentially problematic words, but also includes explanations and references that describe why the identified word may be considered problematic. Figure 1 illustrates the live tool in use on a biased article written about Clinton in 2013.

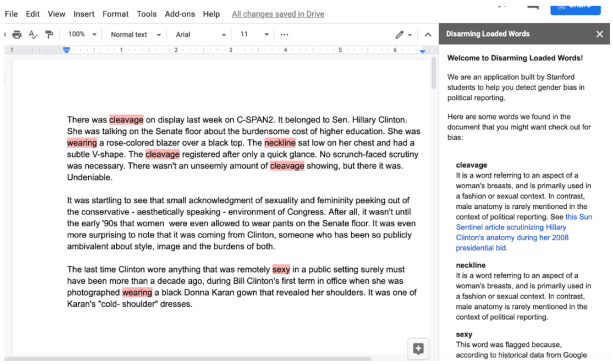

Figure 2: The user interface with flagging (red markup) and the insights panel.

User Interface Evaluation

We conducted three rounds of user testing, with 6 total participants, and used the RITE (Rapid Iterative Testing and Evaluation) Method to continuously improve our prototypes between participants. Participants were sourced through the Stanford Journalism Program. Preliminary results revealed positive feedback about the tool's niche, the ease of identifying flagged words, and the ease of integration into the journalist's workflow. User feedback also revealed several areas of improvement, including detection of quoted words, phrase detection, and development of actionable writer feedback.

Model Evaluation

We evaluated the model's effectiveness in two stages, with a preliminary qualitative stage and a secondary quantitative stage.

Qualitative Human Expert Analysis. In a preliminary analysis, we compared the outputs of the model with the the results of four journalism experts, who hand-tagged the same article for biased language. We learned that journalists identified words that generally agreed with our model output. However, human annotations tended to encompass the entire phrase containing problematic words, rather than tagging only individual words.

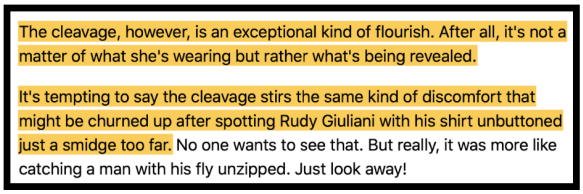

Figure 3: Expert taggers agreed with the outputs of our model, but tagged entire phrases rather than individual biased words.

Figure 4: Excerpts from expert tagging of articles.

Quantitative Precision/Recall Analysis. We then evaluated the model by comparing the words it tagged to those in a human-sourced gold standard set via the common performance metrics of precision, recall and F1-score. The performance metrics were calculated based on true positives ($T.P.$), false positives ($F.P.$) and false negatives ($F.N.$) identified relative to the gold standard. We briefly review the definitions of these classic performance metrics in our experimental context: $T.P.$ words are defined to be those identified by the model and also present in the gold standard set, $F.P.$ words are those identified by the model but not present in the gold standard set, and $F.N.$ words are those not identified in the model but present in the gold standard set. Recall (R) is the proportion ($T.P./(T.P. + F.N.)$) (i.e. the ratio of all words that are correctly tagged by the model, relative to the gold standard), precision (P) is the proportion of tagged words that are true ($T.P./(T.P. + F.P.)$), and F1-score ($F1$) is the harmonic mean of the recall and precision $(2(R \times P)/(R + P))$.

To create the gold standard set of biased words, we solicited human judgement of bias in 10 articles [3, 6–8, 10, 11, 13–16]. The articles were pulled from a diverse set of media sources, which are known to cover a range of political expression: Breitbart, Fox, POLITICO, The New York

Post, and The New York Times. For each article, four to five human labelers were asked to tag words that they believed to be biased. In order to mirror the behavior of the model, labelers were instructed to treat each article as a "bag of words," and to identify all unique biased words from the article text — including titles, secondary titles and subheadings, and excluding figure captions and author information.

For each article, we stemmed each labeler's tags, and then created three sets of gold standards: a simple union of all labelers' tags, a majority of all labelers' tags ($n \geq 3$), and a plurality of all labelers' tags ($n \geq 2$). The simple union gold standard set is self-explanatory. The majority gold standard set contains all word stems that appeared in the sets of three or more labelers. We chose this majority threshold to account for the fact that four of the 10 articles were labeled by four labelers, and the remaining six articles were labeled by five labelers. Finally, the plurality gold standard operated similarly to the majority standard, but with a lower threshold of only two labelers per included word. In theory, taking the intersection of labelers' tags is also a natural choice for gold standard set creation. However we found that for some articles, there were no words that were tagged by all labelers, resulting in an empty intersection gold standard set. For this reason, we limited our analysis to the union, majority, and plurality flavors of gold standard.

After establishing the gold standard sets with our diverse set of 10 articles, we ran our model on the same articles under different tool conditions. In the DLW tool, the bias score threshold is the only tunable parameter, so to explore the space of possible tool performance we ran the model at multiple bias score thresholds ($t_b = 0.25, 0.4, 0.5$). A low t_b corresponds to high tool sensitivity—i.e. DLW tends to tag more words as possibly biased—while DLW operating under a high t_b selects potentially biased words in a more parsimonious manner.

We then compared the words identified by DLW at different t_b to each of the gold standard sets. Our results are summarized in Table 1, which reports the gold standard and t_b conditions which resulted in maximum $F1$, as well as the corresponding P, R and $F1$ under those conditions.

Article	(GS, t_b)	P	R	$F1$
1	(U, 0.5)	1.0	0.5	0.667
2	(U, 0.25)	0.625	0.375	0.469
3	(U, 0.25)	0.061	0.4	0.105
4	(U/M/PL, 0.25)	0.053	1.0	0.1
5	(M/PL, 0.25)	0.017	1.0	0.034
6	(U, 0.25)	0.038	0.222	0.065
7	(U, 0.25)	0.048	0.444	0.087
8	(M/PL, 0.25)	0.111	0.4	0.174
9	(M, 0.5)	0.667	0.222	0.333
10	(M, 0.5)	0.5	0.2	0.286

Table 1: Results from quantitative precision/recall analysis. We report precision (P), recall (R) and F1-score ($F1$) for the gold standard (GS) and bias threshold (t_b) pair that resulted in the maximum $F1$. Gold standard flavors are: union (U), majority (M) or plurality (PL).

Analysis

The results of our model on the sample of 10 articles reveal insights about both trends in the political reporting of women, as well as limitations of current natural language processing techniques.

Best Performers. Our model performed best in articles that featured overt sexism, often paired with polarizing and gendered language. Articles 9 and 10 were both sourced from Breitbart News; Article 2 is an article titled "Clinton's Neckline Takes the Plunge," which refers to the then-senator in highly sexualized terms [8]. Article 1, which received the highest $F1$ score, is actually somewhat of an anomaly; the article had only two words in its gold standard, of which one was detected.

Worst Performers. Our model achieved a very low performance on articles 5 and 6, which received $F1$ scores of only 0.034 and 0.065, respectively. This poor performance is attributable in part to the length of the articles (8 pages and 25 pages, respectively). Longer articles have a greater probability of featuring more complex sentences, making it difficult for the model to precisely detect bias. Moreover, articles with poor performance tended to feature minimal, or only very subtle, amounts of gender-biased language. For example, Article 6 features a largely fair spotlight of Senator Kamala Harris [16]. While Harris is at times described in gendered terms, these cases were not overt, requiring contextual information that the model struggled to detect.

Precision and Recall. We observe that the best performers (Articles 1,2,9, and 10) have higher precision than recall, whereas poor performers (4-8) have higher recall than precision. This can be attributed to the fact that the model largely erred on the side of over- rather than under-labeling. As a result, the best performers, which were strongly biased, received higher precision scores. Conversely, poor performers, which were more subtly biased, received more false positives.

Trends in the Political Reporting of Women. Finally, we observed that political coverage of the upcoming 2020 election tended to feature less overt gender-bias. Even after examining news outlets known for polarizing content, we found few articles of the sort that, in 2016, called Hillary Clinton "Baby Jane" [13], or referred to Clinton by her genitalia [15].

We theorize that this could be due to increased awareness of bias since 2016; alternatively, it may be too early to pass judgement on the level of hostility against women, since, as of the writing of this article, the election remains in the early primary stage. We believe that it will be useful to analyze the language of political coverage as the election cycle progresses.

4 OPEN-SOURCED TOOL

Disarming Loaded Words is a proof-of-concept that demonstrates a novel application of NLP tools in the newsroom. However, we recognize that further updates and customizations will be required to fit the needs of individual journalists, as well as to improve the model in ways discussed in the next section. We therefore encourage others to continue building on this work.

Our tool is open-sourced on Github. The code for the language model and API can be found here. [url: github.com/dylangrosz/Disarming_Loaded_Words] The code for the Google Docs extension can be found here [url: github.com/vivianca/Disarming_Loaded_Words_App].

5 CONCLUSIONS AND FUTURE WORK

We have presented DLW, a user-friendly tool that helps journalists identify words in their writing that may be perpetuating gender stereotypes. Our preliminary results show that DLW is a tool that journalists could integrate into their writing pipeline, and that it succeeds at identifying potentially problematic words. However, several areas remain open for improvement.

We plan on continuing to improve the model, expanding to more complex types of sentence structures, as well as incorporating phrases and crowdsourced human insights. Because the bias of words depends heavily on social trends, we expect that the tool's hard-coded list of words and list of contextually biased words will also need to be continually updated. We also expect that the explanations supporting problematic words will need to be maintained, and anticipate developing a web-crawler that can search for articles that demonstrate gender-biased use of tagged words.

The cleaving process is another crucial area of further improvement. While our work employs a standard NLP dependency parsing over all sentences, it hardly accounts for the exhaustive linguistic and graph analysis necessary to figure out if a suspect word is truly dependent on a female noun (e.g. woman, Hillary, etc.); additional work is also required to confidently determine if a noun is female, male or neither.

If possible, given resource constraints, we also plan to launch more comprehensive validation tests of our tool. The tests would involve a much larger corpus of articles annotated by human readers — recruited through a platform like Amazon Mechanical Turk — in order to provide a thorough and rigorous evaluation of our tool.

We also aim to expand to additional types of bias, such as racial and religious bias. The core of the DLW model is flexible and can easily be customized to other contexts where bias is an issue.

6 ACKNOWLEDGEMENTS

The authors would like to thank R.B. Brenner, Maneesh Agrawala, and Krishna Bharat for their helpful mentorship, advice, and feedback throughout the course of this project. Thank you also to Kylie Jue for her dedication and inspiration, and to Jay Hamilton for feedback on an early draft. Finally, thank you to Lucy Bernholz, Argyri Panezi, and Toussaint Nothias for their draft feedback.

REFERENCES

[1] [n.d.]. *Neural Network Dependency Parser.* Retrieved December 11, 2019 from https://nlp.stanford.edu/software/nndep.html

[2] 2012. *Name It. Change It. The Women's Media Center Guide to Gender Neutral Coverage of Women Candidates + Politicians (2012).* Technical Report.

[3] Isaac Arnsdorf. 2016. How Secretary of State Hillary Clinton Cared for Democratic Donors. Retrieved February 12, 2020 from https://www.politico.com/story/2016/01/clinton-emails-fundraising-soros-spielberg-216535

[4] Sandra L Bem. 1974. The measurement of psychological androgyny. *Journal of consulting and clinical psychology* 42, 2 (1974), 155.

[5] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. arXiv:cs.CL/1607.06520

[6] Jack Crosbie. 2020. Pelosi Delivers Another Resistance Meme, and Nothing Else. Retrieved February 14, 2020 from https://www.rollingstone.com/politics/political-commentary/pelosi-state-of-the-union-trump-democratic-party-948016/

[7] Matt Flegenheimer and Sydney Ember. 2019. How Amy Klobuchar Treats Her Staff. Retrieved February 14, 2020 from https://www.nytimes.com/2019/02/22/us/politics/amy-klobuchar-staff.html

[8] Robin Givhan. 2007. Clinton's Neckline Takes the Plunge. Retrieved February 12, 2020 from https://www.sun-sentinel.com/news/fl-xpm-2007-07-23-0707200237-story.html?fbclid=IwAR1WyR5naw_pMk3m7kDoZZDsKtB8YXG5BmED9Ac8YFiAwyAP5qiC-R9LV3Y

[9] Senior Vice President of The Poynter Institute Kelly McBride. 2019. Personal communication.

[10] Ken Klukowski. 2016. Hillary Clinton: Christians are Bigoted and Deplorable, 'Religious Liberty Code for Discrimination. Retrieved February 14, 2020 from https://www.breitbart.com/politics/2016/09/16/hillary-clinton-christians-are-bigoted-and-deplorable-religious-liberty-code-for-discrimination/

[11] Jon Levine. 2020. AOC, boyfriend discuss 'combating racism as a white person' on Instagram. Retrieved February 14, 2020 from https://nypost.com/2020/02/08/aoc-boyfriend-talk-about-combating-racism-as-a-white-person-in-instagram-stories/

[12] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:cs.CL/1301.3781

[13] John Nolte. 2015. The Republican Party Won Last Nights Debate — and Hillary Lost. Retrieved February 14, 2020

from https://www.breitbart.com/politics/2015/08/07/the-republican-party-won-last-nights-debate-and-hillary-lost/

[14] The New York Times Opinion. 2020. Winners and Losers of the Democratic Debate. Retrieved February 14, 2020 from https://www.nytimes.com/interactive/2020/01/15/opinion/democratic-debate-who-won.html

[15] Ben Shapiro. 2016. Republicans Arent Fighting Hillary, Theyre Fighting the Hillary-Loving Media. Retrieved February 14, 2020 from https://www.breitbart.com/politics/2016/02/16/republicans-arent-fighting-hillary-theyre-fighting-the-hillary-loving-media/

[16] Elizabeth Weil. 2019. Kamala Harris Takes Her Shot. Retrieved February 12, 2020 from https://www.theatlantic.com/magazine/archive/2019/05/kamala-harris-2020-campaign/586033/

[17] www.huffpost.com. 2016. Hillary Clinton vs. Donald Trump: Unlikable vs. Unpredictable. Retrieved December 11, 2019 from https://www.huffpost.com/entry/hillary-clinton-vs-donald_b_10745814

[18] www.msnbc.com. 2016. Woodward: Clinton has to get off 'screaming stuff'. Retrieved December 11, 2019 from https://www.msnbc.com/morning-joe/watch/woodward-clinton-has-to-get-off-screaming-stuff-614955075537

[19] www.nbcnews.com. 2019. DNC names 20 candidates who will appear on stage for first Democratic debate. Retrieved December 11, 2019 from https://www.nbcnews.com/politics/2020-election/dnc-names-20-candidates-who-will-appear-stage-first-democratic-n1017316

[20] www.nytimes.com. 1984. Mississippi Farm Topic: Does She Bake Muffins? Retrieved December 11, 2019 from https://www.nytimes.com/1984/08/02/us/mississippi-farm-topic-does-she-bake-muffins.html

[21] www.nytimes.com. 2016. In Democratic Debate, Candidates Clash on Money's Role. Retrieved December 11, 2019 from https://www.nytimes.com/2016/02/05/us/politics/democratic-debate.html

[22] www.politico.com. 2018. How Amy Klobuchar Treats Her Staff. Retrieved December 11, 2019 from https://www.nytimes.com/2019/02/22/us/politics/amy-klobuchar-staff.html

[23] www.politico.com. 2018. Warren battles the ghosts of Hillary. Retrieved December 11, 2019 from https://www.politico.com/story/2018/12/31/elizabeth-warren-hillary-clinton-1077008

[24] www.twitter.com. 2016. I think this is who she sort of is. Retrieved December 11, 2019 from https://twitter.com/cillizzacnn/status/695441165974269952