Masking Actor Information Leads to Fairer Political Claims Detection

Erenay Dayanik and **Sebastian Padó**
IMS, University of Stuttgart
Stuttgart, Germany
{erenay.dayanik,sebastian.pado}@ims.uni-stuttgart.de

Abstract

A central concern in Computational Social Sciences (CSS) is fairness: where the role of NLP is to scale up text analysis to large corpora, the quality of automatic analyses should be as independent as possible of textual properties. We analyze the performance of a state-of-the-art neural model on the task of political claims detection (i.e., the identification of forward-looking statements made by political actors) and identify a strong frequency bias: claims made by frequent actors are recognized better. We propose two simple debiasing methods which mask proper names and pronouns during training of the model, thus removing personal information bias. We find that (a) these methods significantly decrease frequency bias while keeping the overall performance stable; and (b) the resulting models improve when evaluated in an out-of-domain setting.

1 Introduction

In recent years, NLP methods have found increasing adoption in the social sciences as part of the movement towards *Computational Social Sciences* or CSS (Lazer et al., 2009). An important part of the appeal of CSS is the promise to scale up the amount of data under consideration: from what can be annotated manually to what can be analyzed automatically, typically an increase by several orders of magnitude, enabling a paradigm shift towards new research questions (Chang et al., 2014). However, this shift comes with new challenges: if the analyses are carried out by a machine, how can we trust that any outcomes really stem from the underlying data, rather than from processing artifacts?

Consequently, CSS must be crucially interested in the *algorithmic fairness* or (absence of) *bias* of the underlying machine learning methods (e.g., Binns, 2018; Canetti et al., 2019). However, work on this topic in NLP over the last years has found that more applications contain biases than not, including lexical semantics (Bolukbasi et al., 2016), emotion detection (Kiritchenko and Mohammad, 2018), coreference (Zhao et al., 2018), recommendation generation (Chakraborty et al., 2019) and textual inference (Rudinger et al., 2017). It is therefore surprising that, to our knowledge, the bias of NLP methods applied in the CSS domain have found little attention so far.

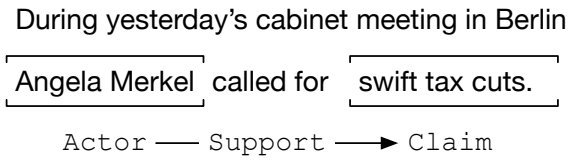

Figure 1: Political claims detection: Text (above), actor–polarity–claim structure (below)

In this paper, we consider the CSS task of *political claim analysis* (Koopmans and Statham, 2010), an entity and relation extraction task from the domain of argument(ation) mining (Cabrio and Villata, 2018). Its goal is to extract (Actor, Polarity, Claim) tuples from text, as illustrated in Figure 1. This is a structured prediction task with the goal of identifying actors, their claims, and polarities (support/opposition). We investigate neural models for the claim identification aspect of political claims analysis trained on a German dataset, MARDY (Padó et al., 2019), and find that these models exhibit a strong frequency bias: claims made by frequently occurring actors are retrieved with higher recall than claims by infrequently mentioned actors. This is worrying, because it means that actors who repeat their claims often will now receive 'preferential treatment' in the aggregated analysis and, arguably, be perceived as even more prominent than they are (Hovy and Spruit, 2016).

We interpret these patterns as overfitting of the claim detection model: it relies too much on *actor*

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4385–4391
July 5 - 10, 2020.

mentions (i.e., either proper names or pronouns) as indicators of claims. To debias the model, we propose three methods: (1) *mask* the actor information by anonymizing referential expressions in the texts, which masks actor information; (2) train claim detectors adversarially by actor frequency; (3) assign more weight to low-frequency training examples in the loss function. We find that actor masking leads to almost no loss in performance but greatly reduces the frequency bias, at the same time improving out-of-domain generalization.

2 Political Claims Detection

Task. For political science, the analysis of political debate provides a window into the process of decision making that is crucial for democracy (Leifeld, 2016). An influential framework in this area is *political claims analysis* (Koopmans and Statham, 2010) which is interested in the association between political *actors* and their *claims* (cf. Figure 1), where claims are statements about specific future actions that the actor endorses or rejects. Such actor-claim pairs can be aggregated into discourse networks and analyzed for aspects such as discourse coalitions or developments over time (Haunss, 2017; Wang and Wang, 2017).

From an NLP perspective, full political claims analysis is a relatively complex process (Padó et al., 2019) that involves recognizing entities (actors), opinions (claims), and the relations between them (actor–claim pairs). In this paper, we focus on the task of claims detection in a narrow sense, namely the identification of claim spans in running text (cf. the right-hand markable in Figure 1), a task that is structurally related to (shallow) argument mining (Swanson et al., 2015; Vilares and He, 2017).

Dataset. We use the MARDY dataset, a corpus of articles relevant to the German immigration debate of the year 2015 drawn from the major German newspaper *Die Tageszeitung (taz)* (Padó et al., 2019). The corpus consists of 959 articles with a total of 1841 claims with an average length of 20 tokens. Each claim is associated with an actor. For about half of the claims (879), the actor is local (i.e., inside the claim); for the rest, it is non-local (i.e., somewhere in the document context).

Model. We investigate a model inspired by the best claims detection model from Padó et al. (2019). Our claims detector is also a transformer based on BERT (Devlin et al., 2019) with a default pretraining objective. However, we make two changes: (a), instead of framing the task as token sequence labeling, we perform sentence-level classification by placing a Softmax classifier on top of BERT, using the final hidden state of the special `[CLS]` token as sentence meaning representation; (b), instead of using the Multilingual BERT model, which is known to have problems with finding sensible subword units for German, we use a BERT model trained solely on German corpora[1]. On the standard training/test split of the MARDY dataset, where Padó et al. (2019) report an Macro average F1 score of 65.5 (P=64.8, R=66.2). Using the same token-level evaluation, our model achieves an moderately improved F1 score of 67.6 (P=64.1, R=71.3), with a similar precision and a 5% increase in recall.

Actor freq. band	All	Low	Mid	High
Freq. range	1–48	1	2–3	>3
# unique actors	186	85	70	31
# claims	879	122	226	531
Model recall	77.1	74.5	77.0	78.0

Table 1: Properties of claims with local actors in MARDY (all and by frequency band) as well as recall of the STANDARD claim detector

3 Frequency Bias and Debiasing

We carry out an analysis of the predictions of our claim detector on the MARDY dataset with 10-fold CV to maximize the amount of data under consideration. We group the actors into three frequency bands using the gold standard actor annotation, as shown in Table 1. Almost half of the actors occur only once, indicating that actors follow a Zipfian distribution as typical for language data.

We now evaluate the performance of our model per actor frequency band. Since actor prediction is not part of the model, we only analyze recall at the claim (not token) level. We also restrict ourselves to the 879 claims with local actors, assuming that local actors influence claim detection. Indeed, as Table 1 shows, the prediction quality differs substantially across actor frequency bands: in particular claims made by *hapax legomena* actors (i.e., single-occurrence actors) show a worse recall (74.5%) than frequent actors (77–78%).

Note that the claim detection model should only pay attention to mentions of actor to the extent this

[1] https://deepset.ai/german-bert

helps in its task. Its sensitivity to actor frequency indicates that the presence of a previously seen actor name is a strong indicator for the presence of a claim. We nevertheless believe that this is an undesirable situation, since it means that the model extracts a *systematically biased* set of claims from the corpus: claims made by frequently mentioned actors (such as office holders or spokespersons) are reinforced, while claims made by infrequently mentioned actors are disregarded. This type of bias can lead to 'echo chambers' (Del Vicario et al., 2016) and confers overly high visibility onto frequent actors (Hovy and Spruit, 2016). To avoid exactly this type of bias, discourse analysis in social science generally factors out the 'newsworthiness' of claims by disregarding its number of mentions. We computationally debias our claims classifier.

3.1 Methods for Frequency Debiasing

Computational debiasing methods generally either modify the model objectives (e.g., Bolukbasi et al., 2016) or the input data (e.g., Zhao et al., 2018). We experiment with both approaches.

Actor Masking. Actor Masking is a data modification method where we mask all referential expressions referring to political actors by replacing the referential expressions with placeholders. We consider two variants:

MASKNAME This model masks the most frequent realization option of political actors, namely proper names of persons. We operationalize 'person name' as all phrases marked as `PER` by the SpaCy German Named Entity Recognizer (F-Score 83.0 on WikiNER).[2]

MASKNAMEPRON This model masks persons names as above. In addition, it masks all personal pronouns in MARDY, which can also provide actor information, even though in a more indirect and thus less informative way. It uses the same placeholder.

These masking procedures make it impossible for the claim detector to use information about the actor identity. The motivation is similar to using denoising autoencoders for text representation, which introduce perturbations in the input to encourage models to discover stable latent rather than surface text properties (Glorot et al., 2011).

Adversarial Debiasing. Adversarial debiasing forgoes changes in the dataset, preferring to use adversarial training to have the model learn representations of the input that do not exhibit biases (in our case, frequency biases) in any substantial way McHardy et al. (2019). Concretely, we train our model simultaneously to predict whether the given text contains any claim and to prevent the adversarial component from predicting how frequently the claim actor occurs (Figure 2): The adversarial and main components share the feature extractor whose parameters (θ_f) are therefore updated by the gradients coming through the objective functions of both model parts. Formally, let J_c and J_{fr} be the cross-entropy loss functions of the main (claim detector) and adversarial (frequency detector) components, let λ be the meta-parameter for the trade-off between the two losses.[3], and let η be the learning rate. Then the updates are defined as:

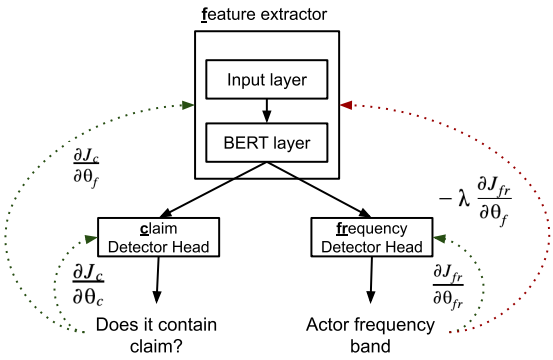

Figure 2: Visualization of adversarial debiasing.

$$\theta_c := \theta_c - \eta \frac{\partial J_c}{\partial \theta_c} \text{ and } \theta_{fr} := \theta_{fr} - \eta \frac{\partial J_{fr}}{\partial \theta_{fr}} \quad (1)$$

$$\theta_f := \theta_f - \eta \left(\frac{\partial J_c}{\partial \theta_f} - \lambda \frac{\partial J_{fr}}{\partial \theta_f} \right) \quad (2)$$

Eq. (2) causes the feature extractor to receive the opposite gradients from the two model components, maximizing the loss of the frequency detector.

Sample Weighting. Sample weighting aims to mitigate frequency bias by punishing model more for false negative predictions on claims by infrequent actors. Each training example is assigned to a weight which reflects the importance of the instance when computing the loss function. Concretely, we introduce three weights (γ_{low}, γ_{mid}, γ_{high}) for the three actor frequency bands from Table 1.[4] Parameter updates (i.e., back-propagation)

[2]Source for model and evaluation figures: https://spacy.io/models/de#de_core_news_sm.

[3]Following hyper-parameter search, we set λ to 1.0.

[4]Following hyperparameter search, we set $\gamma_{low} = 0.5$, $\gamma_{mid} = 0.3$ and $\gamma_{high} = 0.2$, and assign $\gamma = 0.1$ to negative instances (i.e. non-claims).

	Precision	Recall	F2-Score
STANDARD	40.1	74.7	63.7
MASKNAME	39.3	**75.6**	63.8
MASKNAMEPRON	39.8	**75.6**	**64.1**
ADVERSARIAL	**45.5**	69.1	62.6
SAMPLEWEIGHTING	42.3	73.5	**64.1**

Table 2: Exp. 1 (in-domain): Results for all claims.

Actor freq. band	Low	Mid	High
STANDARD	74.5	77.0	**78.0**
MASKNAME	80.3	80.1	77.4
MASKNAMEPRON	**81.4**	**82.7**	77.2
ADVERSARIAL	77.1	73.5	74.5
SAMPLEWEIGHTING	72.1	79.2	76.3

Table 3: Exp. 1 (in-domain): Recall on claims with local actors by actor frequency band.

are performed using scaled loss values.

4 Experiment 1: In-Domain Modeling

We first investigate the effect of frequency debasing in a standard in-domain setting, re-using the setup from Section 3 (10-fold cross-validation, claim-level evaluation) to train one standard and four debiased models. Table 2 shows results on all claims.[5]

We find that the two actor masking models show a slight increase in recall (around 1 point), accompanied by a similar drop in precision. Thus, the F2-Scores of the three models are more or less on par (the differences are not statistically significant): the debiased models perform as well as STANDARD despite the loss of information in the dataset. The two ML-focussed debiasing methods have a completely different impact on the claim detector: Both ADVERSARIAL and SAMPLEWEIGHTING improve the precision significantly, but suffer a decrease in recall. Thus, the data modification methods, in particular MASKNAMEPRON, appear competitive.

Next, we repeat the analysis by frequency band on the set of local claims from Section 3 for all five models. Table 3 shows the recall values. We find that actor masking leads to a slight decrease in recall (under 1 point) for actors from the High band: we believe that this is unproblematic, given the redundancy of newspaper reporting. At the same time, brings about substantial improvements in recall for both the Low (+7 points) and the Mid (+5 points) actor frequency bands – so claims advanced by infrequent actors have a substantially better chance of being recognized by the system. As for the representation-based methods, adversarial training does also, to some extent, lead to fairer claim detector: It mitigates the differences across low and high bands; however, it also leads to significant decrease in overall recall. SAMPLEWEIGHTING is the least effective debiasing method, performing rather badly on the low frequency band.

Regarding a more qualitative understanding of the actor masking methods, consider the following claim which was recognized by both debiased models but not STANDARD:

> Der Dresdner Superintendent Christian Behr ruft zu Nächstenliebe und Dialogbereitschaft auf. (1)
> (Dresden superintendent Christian Behr calls for charity and readiness for dialog.)

We also see improvements for actors realized as general noun phrases (which are almost guaranteed to occur infrequently):

> Anwohner und NPD-Politiker protestierten gegen die geplante Unterkunft. (2)
> (Local residents and NPD politicians protested against the planned accommodation facilities.)

Comparing the two actor masking methods, the improvements in MASKNAMEPRON surpass those of MASKNAME, which indicates that a more consistent treatment of referring expressions by replacing both proper names and pronouns is advantageous, maybe due to the fact that there is often a relatively free choice between pronouns and proper names.

5 Experiment 2: Out-Of-Domain Modeling

We now carry out a second experiment following the intuition that models relying on less specific features generalize better to out-of-domain data – which was also the original motivation for denoising autoencoders (Glorot et al., 2011). As out-of-domain dataset, we used the AKW (Haunss et al., 2013) corpus. This is another German corpus for the task of political claims identification, which covers the debate on the future of nuclear energy use in Germany in the four months after the nuclear

[5]The difference between the F-score reported here for STANDARD and the one from Section 2 is the difference between token-level F1-Score and claim-level F2-Score evaluation. We believe that claim-level evaluation provides a more meaningful evaluation of claim identification but have reported token-level evaluation above for comparison to previous work. Regarding the precise metric, weighting recall higher then precision provides a better match for a semi-automatic setup with manual post-correction (Haunss et al., 2020), which is arguably necessary at the present level of performance.

	Precision	Recall	F2-Score
STANDARD	19.8	40.4	33.4
MASKNAME	21.3	**43.2**	**35.8**
MASKNAMEPRON	20.5	42.2	34.8
ADVERSARIAL	**26.0**	33.0	31.3
SAMPLEWEIGHTING	22.8	40.0	34.8

Table 4: Exp. 2 (cross-domain): Results for all claims.

Actor freq. band	Low	Mid	High
STANDARD	44.9	49.2	52.5
MASKNAME	**48.5**	**53.4**	**54.3**
MASKNAMEPRON	46.2	47.0	51.9
ADVERSARIAL	35.3	40.1	42.2
SAMPLEWEIGHTING	44.0	49.5	52.9

Table 5: Exp. 2 (cross-domain): Recall on claims with local actor by actor frequency band.

disaster of Fukushima, Japan in March 2011. The dataset contains 828 articles and 934 claims, all associated with one of 348 unique actors. We re-use the frequency bands computed for MARDY, under the assumption that it is the frequency distribution in the training data that matters for performance. AKW differs from the MARDY corpus in the subject of the debate, the time span, and the newspapers (*Die Welt* and *Süddeutsche Zeitung*). We used AKW solely as test set for models trained on MARDY.

Table 4 shows the main results. The significant decrease in F-scores compared to Table 2 shows that current claim detection is substantially domain specific. Nevertheless, both MASKNAME (+2 points F-score), MASKNAMEPRON (+1 point F-score) and SAMPLEWEIGHTING (+1 point F-score) generalize somewhat better than STANDARD. MASKNAMEPRON and MASKNAME also beat STANDARD in both precision and recall. ADVERSARIAL, on the other hand, shows a 2.0 points decrease in F2-score as a result of the overall decrease in Recall compared to STANDARD.

Table 5 shows recall values for claims by author frequency bands. As in Exp. 1, this analysis is restricted to claims with locally realized actors.[6] We observe a similar pattern to Exp. 1 (cf. Table 3) for actor masking models: (1) The STANDARD model suffers from frequency bias in the form of worst performance on the Low band (-7 points compared to High); (2) both actor masking models improve performance for the Low band, thus decreasing frequency bias. The two representation-based models, on the other hand, show an overall low recall with no decrease in frequency bias, and particularly bad results on the Low band for ADVERSARIAL.

[6]We only consider actors that occur in MARDY, assuming that it is the frequency in the training set that matters.

6 Conclusion

This paper has discussed the task of political claims analysis as an example of Computational Social Science where NLP methods are finding adoption to scale analysis to large data sets. We have argued that this scenario must be aware of systematic biases in the output of the NLP methods.

The NLP community has mostly focused on biases grounded in extralinguistic reality, e.g., gender (Bolukbasi et al., 2016; Rudinger et al., 2018; Stanovsky et al., 2019), race (Kiritchenko and Mohammad, 2018), or age (Hovy and Søgaard, 2015). We identified *frequency* as a language-internal bias present in a current neural model in political claims analysis. It warrants the same kind of attention as other bias types: lower recall for infrequent actors is inherently unfair, hitting those who can afford least to have their contribution overlooked.

We compared two approaches to mitigating frequency bias in political claims detection and tested them on in-domain and out-of-domain settings. We found that a simple data modification strategy does as good as or better than modifying the model objective. Actor masking improves recall for infrequent actors without affecting overall performance, and, as a side benefit, also improves out-of-domain generalization. While we only evaluated the strategy on one model, we believe its benefits carry over to other model architectures and similar tasks.

Clearly, actor frequency is only one of a large number of potential frequency-related biases. Since frequency is known to be strongly correlated with performance in machine learning-based NLP, such biases should be investigated more systematically in areas building on NLP such as Computational Social Sciences. To remove these biases, however, presumably more sophisticated methods will be necessarily in the general case.

Acknowledgments

Funding was provided by Deutsche Forschungsgemeinschaft (DFG) through project MARDY in SPP RATIO. We would like to thank G. Lapesa, N. Blokker and S. Haunss for valuable comments.

References

Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 149–159, New York, NY, USA.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of NeurIPS*, pages 4349–4357.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *Proceedings of IJCAI*, pages 5427–5433, Stockholm, Sweden.

Ran Canetti, Aloni Cohen, Nishanth Dikkala, Govind Ramnarayan, Sarah Scheffler, and Adam Smith. 2019. From soft classifiers to hard decisions: How fair can we be? In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, page 309–318, Atlanta, GA.

Abhijnan Chakraborty, Gourab K Patro, Niloy Ganguly, Krishna P Gummadi, and Patrick Loiseau. 2019. Equality of voice: Towards fair representation in crowdsourced top-k recommendations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 129–138.

Ray M. Chang, Robert J. Kauffman, and YoungOk Kwon. 2014. Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems*, 63:67 – 80.

Michela Del Vicario, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2016. Echo chambers: Emotional contagion and group polarization on facebook. *Scientific Reports*, 6(1):37825.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186, Minneapolis, MN.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of ICML*, pages 513–520, Bellevue, WA.

Sebastian Haunss. 2017. (De-)legitimating discourse networks: Smoke without fire? In *Capitalism and Its Legitimacy in Times of Crisis*, pages 191–220. Palgrave Macmillan.

Sebastian Haunss, Matthias Dietz, and Frank Nullmeier. 2013. Der Ausstieg aus der Atomenergie. Diskursnetzwerkanalyse als Beitrag zur Erklärung einer radikalen Politikwende. *Zeitschrift für Diskursforschung*, 1(3):288–316.

Sebastian Haunss, Jonas Kuhn, Sebastian Padó, Andre Blessing, Nico Blokker, Erenay Dayanik, and Gabriella Lapesa. 2020. Integrating manual and automatic annotation for the creation of discourse network data sets. *Politics and Governance*, 8(2). To appear.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of ACL*, pages 483–488, Beijing, China.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of ACL*, pages 591–598, Berlin, Germany.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of STARSEM*, pages 43–53, New Orleans, LA.

Ruud Koopmans and Paul Statham. 2010. Theoretical Framework, Research Design, and Methods. In *The Making of a European Public Sphere*, pages 34–59. Cambridge University Press.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Computational social science. *Science*, 323(5915):721–723.

Philip Leifeld. 2016. *Policy Debates as Dynamic Networks: German Pension Politics and Privatization Discourse*. Campus Verlag, Frankfurt/New York.

Robert McHardy, Heike Adel, and Roman Klinger. 2019. Adversarial training for satire detection: Controlling for confounding variables. In *Proceedings of NAACL:HLT*, pages 660–665, Minneapolis, MN.

Sebastian Padó, André Blessing, Nico Blokker, Erenay Dayanik, Sebastian Haunss, and Jonas Kuhn. 2019. Who sides with whom? Towards computational construction of discourse networks for political debates. In *Proceedings of ACL*, page 2841–2847, Florence, Italy.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of NAACL*, pages 8–14, New Orleans, LA.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of SIGDIAL*, pages 217–226, Prague, Czech Republic.

David Vilares and Yulan He. 2017. Detecting perspectives in political debates. In *Proceedings of EMNLP*, pages 1573–1582, Copenhagen, Denmark.

Chengwei Wang and Luhao Wang. 2017. Unfolding policies for innovation intermediaries in China: A discourse network analysis. *Science and Public Policy*, 44(3):354–368.

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 NAACL*, pages 15–20, New Orleans, LA.