# T.M. Scanlon at SemEval-2023 Task 4: Leveraging Pretrained Language Models for Human Value Argument Mining with Contrastive Learning

**Milad Molazadeh Oskuee**
Department of Computer Science
Iran University of Science and Technology
Tehran, Iran
molazadeh_milad@comp.iust.ac.ir

**Mostafa Rahgouy**
Department of Computer Science
Auburn University, Alabama, USA
mzr0108@auburn.edu

**Hamed Babaei Giglou**
TIB Leibniz Information Centre for
Science and Technology
Hannover, Germany
hamed.babaei@tib.eu

**Cheryl D Seals**
Department of Computer Science
Auburn University, Alabama, USA
sealscd@auburn.edu

## Abstract

Human values are of great concern to social sciences which refer to when people have different beliefs and priorities of what is generally worth striving for and how to do so. This paper presents an approach for human value argument mining using contrastive learning to leverage the isotropy of language models. We fine-tuned DeBERTa-Large in a multi-label classification fashion and achieved an F1 score of 49% for the task, resulting in a rank of 11. Our proposed model provides a valuable tool for analyzing arguments related to human values and highlights the significance of leveraging the isotropy of large language models for identifying human values.

## 1 Introduction

Argument mining is a promising area of research in natural language processing that aims to identify and analyze arguments from the text. Identifying human values in arguments is significant because different beliefs and priorities can cause conflicts or alignment between individuals or cultures, leading to disagreements on controversial issues. Understanding the values behind different perspectives can improve communication and decision-making. The automatic identification of values in written arguments is valuable for argument faceted search, value-based argument generation, and value-based personality profiling (Kiesel et al., 2023, 2022). The *ValueEval*[1], a task 4 of the shared task on SemEval 2023 is designed to automatically identify human values in arguments. The task has been defined as follows:

> Given a textual argument and a human value category, classify whether or not the argument draws on that category.

Touché23-ValueEval dataset (Mirzakhmedova et al., 2023) for this task consists of a valuable collection of argumentative texts, where each argument is labeled with one or more of the human values. Arguments are given as premise text, conclusion text, and binary stance of the premise to the conclusion whatever it is "in favor of" or "against" the premise text. The task uses a set of 20 value categories compiled from the social science literature. The dataset for this task consists of two different sets called *Main*, and *Supplementary* datasets, where in this research we focus on all 20 value categories of the *Main* dataset for identifying human values behind arguments in the English language.

Contrastive learning (CL) (Gao et al., 2021) has shown promising results in various natural language processing tasks, including text classification, sentiment analysis, and named entity recognition. CL is a technique used to learn representations that are invariant to changes in certain aspects of the input data while preserving important information. In the context of argument mining it is important to identify which part of the information is representative of arguments. So, we applied CL with additional information to language models to leverage the pre-trained weights for language understanding capabilities to be able to make differences between different arguments. The proposed method for this task is implemented in *Python* and published on *GitHub*[2] for the research community in this field.

The rest of the paper is organized as follows. Section 2 presents backgrounds. Section 3 describes

[1] https://touche.webis.de/semeval23/touche23-web/

[2] https://github.com/MiladMolazadeh/ValueEval

*Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 603–608
July 13-14, 2023 ©2023 Association for Computational Linguistics

the proposed system overview. Section 4 describes the experimental setups such as dataset, metrics, and training setups. Next, in section 5 we discussed results and analysis. Finally, section 6 presents our conclusions

# 2 Background

## 2.1 Human Value Argument Mining

For the first time, (Schwartz, 1994) established a theory of basic individual values in measuring human values in a survey in 1994 for investigating universal aspects in the structure and contents of human values. Since then this theory has gained wide acceptance in psychology and proposed human values are used widely in assessing human values in arguments. According to this research, fundamental human values provide a framework for understanding differences in motivations, attitudes, and behaviors among individuals and cultures. So, people prioritize certain values over others, and these priorities shape their behavior and decision-making. By understanding the values that are most important to individuals and groups, it is possible to understand their behavior better and develop more effective interventions and policies. More later (Schwartz et al., 2012) refined the theory by organizing the values to provide greater heuristic and explanatory power than the original theory. The theory defines values on the continuum based on their compatible and conflicting motivations, expression of self-protection versus growth, and personal versus social focus.

The work of (Alshomary and Wachsmuth, 2021) presents an innovative method for creating arguments that have the potential to enhance the effectiveness of persuasive communication by factoring in the audience's background and beliefs. They highlighted two research questions that could contribute toward building an audience-aware argumentation technology along with the potential challenges and possible opportunities in each area. The first question is how to model the audience's beliefs from a given representation, this representation being texts, preferences, etc. The second question is how to encode a model of beliefs into argumentative texts. In recent work, (Kiesel et al., 2022) aimed at identifying the values behind arguments computationally. They believed that values are connected specifically with the argument's premise and automatic models might still improve when incorporating the textual conclusion as context for the textual premise.

## 2.2 Isotropy of PLMs

Isotropy is a key geometric property of the semantic space of Pretrained Language Models (PLMs). Recent studies identify the anisotropy problem of PLMs, which is also known as the representation degeneration problem(Gao et al., 2021). Recent research in isotropization has shown that regularizing the feature space towards isotropy can further improve the performance of supervised pretraining.

The (Rajaee and Pilehvar, 2021) work demonstrates that isotropic embeddings have a significant improvement in downstream task performance, as they capture more semantic information and reduce noise. Moreover, (Xiao et al., 2022) exhibits that contrastive learning brings isotropy to sentence representation learning. Moreover, (Zhang et al., 2022) reveals that correlation regularizer penalizes the correlation between different features of pretrained language models and by reducing correlation, the feature space becomes more isotropic and the PLMs become more generalized.

# 3 System Overview

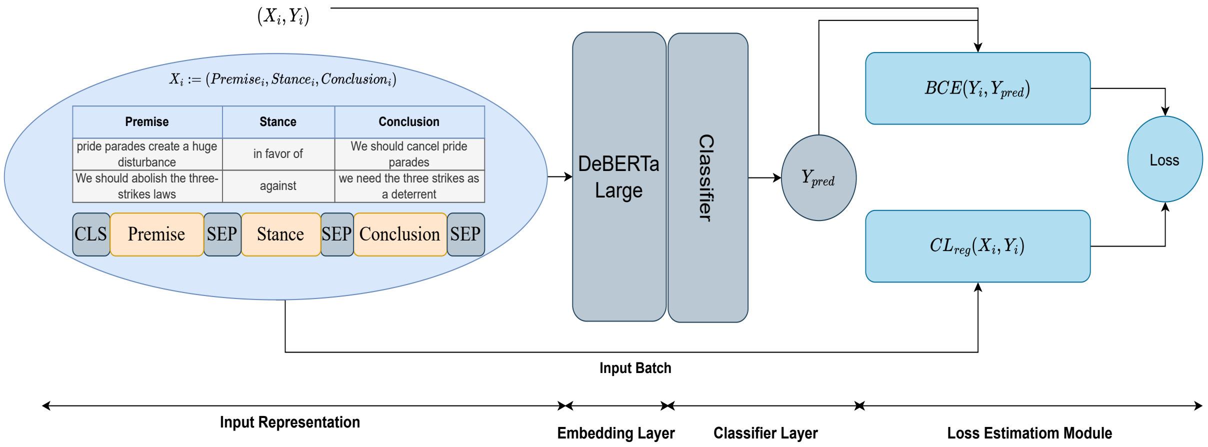
We propose two regularizers based on contrastive learning and binary cross entropy loss for achieving isotropy in the feature space for human values. In this section, we describe the details of our proposed deep learning model for human value argument mining. The proposed methodology is illustrated in Figure 1.

## 3.1 Input Representation

In order to fine-tune a transformer model for human value argument mining. Based on experimentations, we find out the optimal way of concatenating the premise text with the stance and conclusion texts. This input format allows the model to process the entire argument as a single sequence, while also providing information about the relationship between the premise, stance, and conclusion. Where $X_i := (Premise_i, Stance_i, Conclusion_i)$ represents concatenated input representations. In the end, inputs are truncated to a maximum acceptable length for the model.

## 3.2 Embedding Layer

We used DeBERTa-Large (He et al., 2021) language model for fine-tuning our task. DeBERTa-Large has 345 million parameters and has been



| Premise | Stance | Conclusion |
|---|---|---|
| pride parades create a huge disturbance | in favor of | We should cancel pride parades |
| We should abolish the three-strikes laws | against | we need the three strikes as a deterrent |

Figure 1: Architecture of Proposed Model

pre-trained on a massive amount of text data, making it a powerful choice for our task. To speed up the fine-tuning process and prevent overfitting, we used DeBERTa-Large in a partial fine-tuning fashion. This means we froze the first 8 layers of the model and fine-tuned the remaining layers. The early layers of the model learn low-level features, such as character and word embeddings, that are less likely to be useful in this task. However, the later layers of the model learn higher-level features, such as sentence and document representations, that are more specific to the task at hand. By fine-tuning these layers, we allowed the model to adapt to the specific patterns and structures while still leveraging the general knowledge learned during pre-training.

### 3.3 Classifier Layer

Since this task generally is a multi-label classification task, we used a sigmoid classifier to calculate the probability of each label being present in the input text. The input to the sigmoid classifier is a vector of hidden states from the last layer of the DeBERTa model, and the output is a set of probabilities, one for each label between 0 and 1. A threshold value of 0.5 is applied to the predicted probabilities to determine the labels for the input text.

### 3.4 Contrastive Learning

Inspired by (Zhang et al., 2022; Su et al., 2021) to mitigate the isotropy of the PLM fine-tuned by supervised pre-training, we used a loss estimation methodology that incorporates both contrastive learning and binary cross entropy (BCE) loss. Our loss formulation is given by:

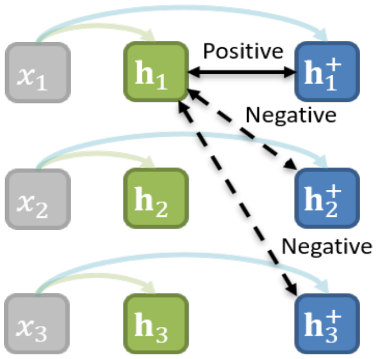


Figure 2: Illustration of Contrastive Learning Based Regularizer. Figure from (Zhang et al., 2022).

$$L = L_{BCE} + CL_{reg} * \lambda$$

Where $L_{BCE}$ is the binary cross-entropy loss, $CL_{reg}$ is the contrastive learning regularizer, and $\lambda$ is the weight parameter for the CL regularizer, which we set to $0.1$. The aim is to learn human value argument minings while maintaining an appropriate degree of isotropy. For $CL_{reg}$ we used (Yan et al., 2021) technique that maximizes the agreement between one representation and its corresponding version that is augmented from the same input while keeping it distant from other input representations in the same batch. Figure 2 shows an illustration of this technique, where positives are for maximizing agreements and negatives are for keeping distance from other inputs in the same batch. To obtain $CL_{reg}$, the $X_i$ is passed to the PLM twice to produce $h_i^+$ and $h_i$ for the following contrastive learning loss:

| Argument Source | Year | Arguments | | | | Unique Conclusions | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Train | Validation | Test | Σ | Train | Validation | Test | Σ |
| Main dataset | | | | | | | | | |
| IBM-ArgQ-Rank-30kArgs | 2019–20 | 4576 | 1526 | 1266 | 7368 | 46 | 15 | 10 | 71 |
| Conf. on the future of Europe | 2021-22 | 591 | 280 | 227 | 1098 | 232 | 119 | 80 | 431 |
| Group Discussion Ideas | 2021-22 | 226 | 90 | 83 | 399 | 54 | 23 | 16 | 93 |
| $\Sigma(main)$ | | 5393 | 1896 | 1576 | 8865 | 332 | 157 | 106 | 595 |

Table 1: Key statistics of the main dataset by argument source.

$$CL_{reg} = -\frac{1}{N} log \frac{exp(sim(h_i, h_i^+)/\tau)}{\mathop{}_{j=1, j\neq i}^{N} exp(sim(h_i, h_j^+)/\tau)}$$

where $sim(\cdot)$ indicates the cosine similarity function, $\tau$ is the temperature parameter which set to $0.05$, and $N$ is the batch size. The $(h_i, h_i^+)$ represents a positive pairs, similarly $(h_i, h_i^+)$ represents a negative pairs. By minimizing the contrastive loss, positive pairs are pulled together while negative pairs are pushed away, which in theory enforces an isotropic feature space (Gao et al., 2021) in a supervised manner.

## 4 Experimental Setup

**Dataset**: Table 1 presents the statistics of the *main* dataset by argument sources. The main dataset consists of 8865 arguments, and 595 unique conclusions from three different sources from 2019 to 2022. The dataset was divided into train, validation, and test sets (Mirzakhmedova et al., 2023).

**Evaluation Metrics**: Due to the multi-label nature of task models are evaluated based on F1-score using macro-precision and macro-recall, with averaged over all value categories and for each category individually.

**Training Setups**: Using train and validation sets we made hyperparameters tuning. For earlybird submission, we used a train set and for final submission, we combined train and validation sets for training models. In this study, we used a batch size of 8 and trained the model for 10 epochs. We employed the Adam optimizer with a learning rate of $1e-05$ and a weight decay of $0.01$.

**External Libraries**: We used *PyTorch* (Paszke et al., 2017) and *Transformers* (Wolf et al., 2020) libraries in our experimentations.

## 5 Results

**Main Quantitative Findings:** The table 2 presents the final results on the test set for submitted 5 different runs including early bird submission. The main quantitative findings are:

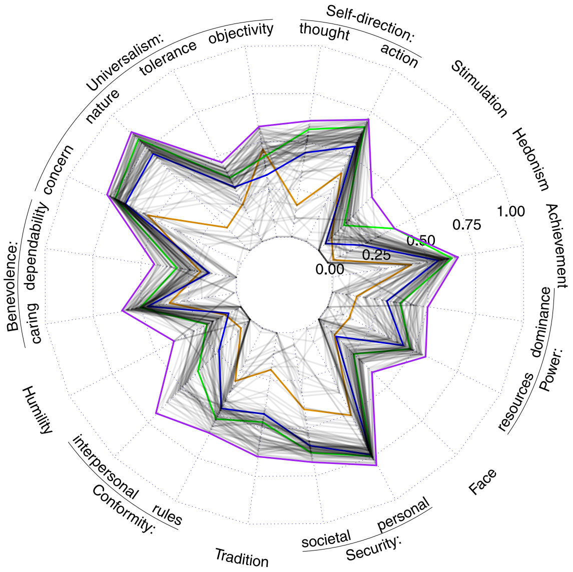


Figure 3: The 1-baseline is in orange, the BERT baseline is in blue, the Best per category is in purple, our best model is in green, and the rest of the runs from all submissions are in gray.

- The proposed model achieved **11th** place among 39 participants.
- In the **Hedonism** class our model performed better than early bird submission and ranked **1st** in the whole task runs (red colored f1 score in the table 2).
- The proposed model defeated baseline models by a large margin.
- According to the results, Large Language Models (LLMs) with more parameters are promising in this task. However, isotropy made a significant improvement in both base and large variants of PLMs.
- The *Earlybird Submission* which was trained on train set only, performed well in *Universalism: concern*, *Universalism: nature*, and

| **Test set / Approach** | **All** | **Self-direction: thought** | **Self-direction: action** | **Stimulation** | **Hedonism** | **Achievement** | **Power: dominance** | **Power: resources** | **Face** | **Security: personal** | **Security: societal** | **Tradition** | **Conformity: rules** | **Conformity: interpersonal** | **Humility** | **Benevolence: caring** | **Benevolence: dependability** | **Universalism: concern** | **Universalism: nature** | **Universalism: tolerance** | **Universalism: objectivity** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Main dataset* | | | | | | | | | | | | | | | | | | | | | |
| Best per category | .59 | .61 | .71 | .39 | .39 | .66 | .50 | .57 | .39 | .80 | .68 | .65 | .61 | .69 | .39 | .60 | .43 | .78 | .87 | .46 | .58 |
| Best approach | .56 | .57 | .71 | .32 | .25 | .66 | .47 | .53 | .38 | .76 | .64 | .63 | .60 | .65 | .32 | .57 | .43 | .73 | .82 | .46 | .52 |
| BERT | .42 | .44 | .55 | .05 | .20 | .56 | .29 | .44 | .13 | .74 | .59 | .43 | .47 | .23 | .07 | .46 | .14 | .67 | .71 | .32 | .33 |
| 1-Baseline | .26 | .17 | .40 | .09 | .03 | .41 | .13 | .12 | .12 | .51 | .40 | .19 | .31 | .07 | .09 | .35 | .19 | .54 | .17 | .22 | .46 |
| DeBERTa-Base + $L_{BCE}$ | .45 | .51 | .64 | .10 | .28 | .56 | .34 | .35 | .09 | .70 | .60 | .57 | .41 | .35 | .10 | .52 | .18 | .72 | .73 | .37 | .41 |
| DeBERTa-Large + $L_{BCE}$ | .44 | .49 | .57 | .09 | .19 | .60 | .27 | .41 | .14 | .74 | .60 | .45 | .49 | .41 | .10 | .53 | .16 | .70 | .72 | .41 | .40 |
| DeBERTa-Base + $L_{BCE}$ + $CL_{reg}$ | .47 | .55 | .67 | .12 | .19 | .57 | .40 | .36 | .17 | .73 | .64 | .55 | .51 | .39 | .26 | .49 | .32 | .75 | .79 | .37 | .44 |
| DeBERTa-Large + $L_{BCE}$ + $CL_{reg}$ | **.49** | .56 | .67 | .18 | .39 | .63 | .36 | .48 | .26 | .75 | .63 | .47 | .53 | .38 | .20 | .50 | .31 | .73 | .82 | .37 | .42 |
| *Earlybird Submission* | | | | | | | | | | | | | | | | | | | | | |
| DeBERTa-Large + $L_{BCE}$ + $CL_{reg}$ | **.49** | .52 | .69 | .07 | .29 | .60 | .35 | .46 | .23 | .74 | .65 | .57 | .52 | .20 | .18 | .55 | .30 | .74 | .84 | .35 | .46 |

Table 2: Achieved $F_1$-score of team t-m-scanlon per test dataset, from macro-precision and macro-recall (All) and for each of the 20 value categories. Approaches in gray are shown for comparison: an ensemble using the best participant approach for each individual category; the best participant approach; and the organizer's BERT and 1-Baseline. Red color represents our best result in best per category, blue represents our outperformance result regarding the Best approach.

*Security: societal* classes even better than *Best approach* (blue colored f1 score in the table 2).

- The DeBERTa+$L_{BCE}$+$CL_{reg}$ model in final submission achieved F1 score of **0.4939** and in early bird submission **0.4879**. Whereas in the final submission training and validation sets are combined for the training model. So, the data augmentation technique may not appropriately solve this task, since the LLMs are suffering from semantic overlaps in identifying human values behind arguments.

**Quantitative Analysis**: Figure 3 represents all the runs submitted to this task in gray color and best per category in purple color, BERT baseline in blue, 1-baseline in orange colors, and our Best model result in green colors. According to this figure, the quantitive analysis is presented as follows:

- Most of the time the proposed model successfully defeated baseline models by large margins, especially the 1-baseline model. Except in *Universalism: objectivity* class, the 1-baseline model performed better than our models in the final run. However, in the Earlybird submission, they both performed the same.
- According to this figure, our model behaves similarly to the BERT baseline most of the time, with a high score for individual classes. It reveals that a single model may not perform well in all of the individual classes at the same time and this task requires ensemble learning approaches.

## 6 Conclusion

This paper presented our approach for SemEval2023 Task 4: ValueEval: Identification of Human Values behind Arguments. We investigated this problem by leveraging PLMs using a contrastive learning technique. The proposed study in this paper shows that isotropization for human value detection is effective and requires more attention in this field. In the final, we achieved an average F1 score of 0.4939. Regarding our evaluations, our approach is promising for human value argument mining.

## References

Milad Alshomary and Henning Wachsmuth. 2021. Toward audience-aware argument generation. *Patterns*, 2(6):100253.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the Human Values behind Arguments. In *60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pages 4459–4471. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. 2023. Semeval-2023 task 4: Valueeval: Identification of human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Barriere Valentin, Doratossadat Dastgheib, Omid Ghahroodi, Mohammad Ali Sadraei, Ehsaneddin Asgari, Lea Kawaletz, Henning Wachsmuth, and Benno Stein. 2023. The Touché23-ValueEval Dataset for Identifying Human Values behind Arguments. *CoRR*, abs/2301.13771.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Sara Rajaee and Mohammad Taher Pilehvar. 2021. A cluster-based approach for improving isotropy in contextual embedding space. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 575–584, Online. Association for Computational Linguistics.

Shalom H Schwartz. 1994. Are there universal aspects in the structure and contents of human values? *Journal of social issues*, 50(4):19–45.

Shalom H Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, et al. 2012. Refining the theory of basic individual values. *Journal of personality and social psychology*, 103(4):663.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chenghao Xiao, Yang Long, and Noura Al Moubayed. 2022. On isotropy and learning dynamics of contrastive-based sentence representation learning. *arXiv preprint arXiv:2212.09170*.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. 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 5065–5075, Online. Association for Computational Linguistics.

Haode Zhang, Haowen Liang, Yuwei Zhang, Li-Ming Zhan, Xiao-Ming Wu, Xiaolei Lu, and Albert Lam. 2022. Fine-tuning pre-trained language models for few-shot intent detection: Supervised pre-training and isotropization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 532–542, Seattle, United States. Association for Computational Linguistics.