Human values behind arguments
a NLP-based approach for stance prediction

NLP Course Project

Lorenzo Molfetta, Memoona Shah Vito Parisi and **Alessandro Rizzo**
Master's Degree in Artificial Intelligence, University of Bologna
{lorenzo.molfetta, vito.parisi2, alessandro.rizzo14, memona.shah}@studio.unibo.it

Abstract

The ability to sense the characteristic features of interlocutors behind their arguments is a trending research field for social network analysis which is useful for profiling users and detecting opinion polarization. In this project, we report a general comparative analysis of different BERT models to check which model best performs on this human value detection dataset. Secondly, an observation based on input combination is made to check which input format is better for the multi-label classification task. In the last part, we propose CAESAR, an NLP-based approach to tackle the stance prediction task that aims to predict human opinion features behind arguments. Aware of the shortcomings of encoder models in dealing with short sentences without contextual information, we try to devise a model that supports the prediction with a tone sense classification. Finally, we compare the proposed model with baseline solutions and test their performances with different data distributions.

1 Introduction

In the last years, Natural language models have been widely used for a vast variety of application fields. Past prominent approaches like Recurrent Neural Network (RNN) (Hochreiter and Schmidhuber) and Gated Recurrent Unit (GRU) (Cho et al.) have shortly fallen into disuse after the introduction of attention mechanisms and Transformer models (Vaswani et al.). Such architectures proved to be more effective in encoding the inner features of textual data while the former have remained a valid alternative for temporal analysis. The breakthrough of bi-directional encoders like BERT (Devlin et al.) has further provided researchers with a baseline architecture to fine-tune for a wide range of tasks. The baseline models we are going to consider for comparison purposes are all based on classification models. Although other approaches, like Support Vector Machines (Hearst et al.), have also proved to be successfully applicable to this task, we've decided to consider only BERT-based models due to their great ability to encode textual information, This is indeed a desirable feature for a classification task like the one here addressed which is solely based on this type of data.

Understanding human values given some arguments has become a crucial task when we are dealing with a large amount of information coming from social networks. Analyzing such data allows us to derive cultural information about the evolution of specific groups during the years due to social or political changes. In addition to such a more historiographical approach, one could exploit the analysis of arguments in a social network to derive information about how much the debate has become polarized. (Zhang et al.) addresses this latter task by conducting connected behavioral analysis in which "connected behavior" is the property for which users who share similar political views are likely to agree on other topics, too. They prove how training encoder models with features showing a high level of correlation can help the prediction task. Such external information serves indeed as contextual support for the classification as the propensity toward one value may positively bias the accordance with another one.

To test our model we used the "Touché23-Human-Value-Detection" dataset presented in (Kiesel et al., 2022) which proposes a multi-level taxonomy of 54 human values from different countries in line with psychological research. As neither contextual nor user-related information is provided, the connected behavior approach in (Zhang et al.) is not applicable. However, a correlation between features widely spread in the same source country can be used to aid the classification. We have ex-

plored and compared the performance of different BERT-based models on a multi-class and multi-label problem. The models used in this project are BERT BASE, BERT LARGE, ROBERTA BASE, ROBERTA LARGE, and DISTILBERT. Our goal was to determine which of these models performs best on the given problem. To ensure fair comparisons, we used the same parameters for all models since most of them converge within four epochs. The dataset input contains three features - conclusion, stance, and premise - and the labels are multi-label data with binary values of 0 or 1. Through our analysis, we hope to provide insights into the effectiveness of different BERT-based models for multi-class and multi-label problems. We also propose CAESAR as our final model, an encoder-based model which deploys a traditional classification network for the feature predictions with an additional branch that enhances the input data representation by sensing the premise's tone so as to further support the stance evaluation.

Moreover, in order to extend this study, we chose to implement other neural architectures in order to compare the results with the original ones. First of all, we used a different Bert-based model, that is DeBERTa, as well as its different version which were introduced in different papers. Afterwards, we proceeded to implement XLNet, a model that is the extension of the transformer-XL model pre-trained using an autoregressive method to learn bidirectional contexts by maximizing the expected likelihood overl all permutations of the input sequence factorization order. We applied Transformer Layers Reinitialization, this idea is motivated by computer vision transfer learning results, in fact existing methods using transformers show that using the complete network is not always the most effective choice and usually slows down training and hurts performance. In the end, we tackled a different approach, Zero-Shot classification, implemented using BART LARGE.
Finally, we compared the results obtained by the aforementioned techniques through an error analysis section, putting some focus on the best performing model.

2 Background

Stance detection identifies whether an opinion is in favor of an idea or opposes it. Due to its tight connection with sentiment analysis, several implementations of models for this latter task have been applied to our problem. As such, the main issues of sentiment analysis are shared among the two fields, too.

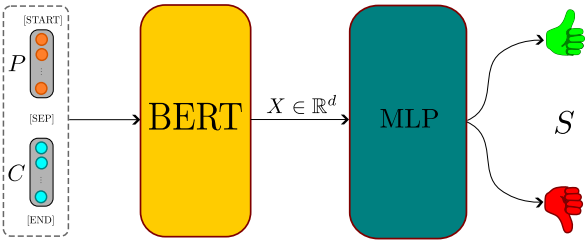

Figure 1: Baseline architecture

Although bidirectional encoders have led a significant advancement in text processing and understanding, if not properly fine-tuned such models lack syntactic consistency. While these encoders manage to successfully model contextual information, they are brittle to short pieces of text when it comes to determining the tone of sentences. These more syntactical features that include the detection of negations are crucial for stance prediction tasks as they are the cornerstones for a good understanding of the given user's intentions. Such a shortcoming of encoder-based methods can be hardly neglected when, like in our situation, one deals with very short sentences. Starting from the general structure of encoder-based models as shown in Figure 1, the past approaches that have tried to overcome the issues of this type of architecture have mainly focused on enhancing the input representation.

As mentioned in the introduction 1 and as exhaustively addressed in (Zhang et al.), non-topic-related information can be used to support the classification task because of their ability to depict a more general contextual frame. When data can be directly linked with specific owners, it is possible to exploit information from similar users and their values to support the predictions. Following this idea, (Darwish et al.) defines a graph of entities representing the interaction between users and their shared opinions. Leveraging such connections their model is able to classify users in a user-similarity feature space.

Other models that cannot resort to personalized information, like in our case, enhance the semantic representation of data by exploiting the strength of encoder models. (Levine et al.) combine a

masked-word prediction pre-training activity for the definition of a context-coherent embedding space with weakly supervised super-sense prediction task that fine-tunes such representation which makes the mapping operation able to cope with "good noise" in the input data. Specifically, applying a WordNet Lemmatizer on each input word w, the model is trained to maximize the probability that the predicted sense is in the set of allowed super-senses of the masked word w. This is beneficial for a stance prediction task as it improves the model's encoding abilities. Hence, only operating on the representation layer of the baseline architecture, they proved how one can increase the performance of the classifier without intervening in the non-linear mapping part of the model.

(Hosseinia et al.) leverages sentiment and emotion information separately with BERT representations obtained from the last BERT-base layer to form the input of a shallow recurrent neural network. Using the VADER model (Hutto and Gilbert) to derive a sentiment score for the input sentence ranging from -1 to 1, they concatenate such sentiment embeddings with the original ones to form the input $x_i = [h_i^{BERT}; h_i^{sent}]$ for a neural architecture comprising a bidirectional GRU layer and a final dense mapping. Instead of working directly on the representation of each input token, (Popat et al.) proposes to enhance the model by augmenting it with a novel consistency constraint to capture agreement between the premise and the conclusion. This is done by assuming that their representations should be dissimilar if the premise opposes the conclusion whereas they should be similar if the conclusion is supported by the perspective. To incorporate such consistency into the model they use a cosine embedding loss that is similar to a metric learning algorithm map in nearby regions of the embedding space - by lowering their cosine similarity - the representations of the premise and the conclusion of the former support the latter and viceversa.

Reinitializing Transformer Layers is an interesting approach where instead of using the pre-trained weights for all layers, one can re-initialize the pooler layers and the top Transformer blocks using the original Transformer initialization. The layers reinitialized results in destruction of gained pre-trained knowledge for those specific blocks. The idea behind this technique is motivated by computer vision transfer learning results where we know that lower pre-trained layers learn more general features while higher layers closer to the output specialize more to the pre-training tasks. Existing methods using Transformer show that using the complete network is not alwyas the most effective choice and usually slows down training and hurts performance.

3 System description

We propose two different approaches for the detection of human values. The first one aims at determining whether an individual supports a specific thesis based on some premises and some conclusions. This is encoded as a Stance Prediction task where the goal of the model is to learn to return positive or negative feedback based on the input statement. The second case study we analyze is the detection of the human values behind a thought. Leveraging the taxonomy described in section 4, we propose an approach for the prediction of values from different levels of the hierarchy starting from the premise, stance, and conclusion. In section 5 we carry out an ablative analysis reporting the performances of different fine-tuned encoder models employed to encode the input data.

Stance Prediction

Inspired by the solution shown in Figure 1, we propose CAESAR, a Bert-based model for the prediction of stances that includes an additional branch supporting the computation with sentiment analysis. This additional processing of the input text endows the model with the ability to more effectively encode the syntactic properties of the premises by predicting whether the input sentence has a positive or negative meaning. The predictions made by the two main branches of the architecture are then merged to return a unique and comprehensive outcome. Optimal results can be obtained by fine-tuning pre-trained models like (Hartmann et al., 2023) that are trained to classify the sentiment of tweets. Because of the nature of the short pieces of information they are able to deal with, these models are particularly suitable to be included in our model. However, due to the limitation of computational resources at our disposal, we defined a custom architecture that emulates the standard structure of the aforementioned network which usually

deploys a classification layer working on encoded information coming from Bert-based networks. As shown in Figure 3, we propose a "*Sentiment Model*" including a GRU layer (Cho et al.) whose inner representation is merged with the output of the encoder model.

Specifically, given a premise P, a conclusion C and their embedded representations $X^{\mathcal{E}} \in \mathbb{R}^H$, let us denote with $\mathcal{B}$ and $\mathcal{E}$ the BERT and Sentiment Model's embedding layers, respectively. The output of the former is used to set the hidden state of the GRU layer as follows:

$$X^{\mathcal{E}} = \mathcal{E}([P;C]),\ X^{\mathcal{B}} = \mathcal{B}([P;C])$$
$$g_{out} = GRU\left(X^{\mathcal{E}}, \mathcal{B}_{pooler\_output}\right)$$
$$X = [\mathcal{B}_{pooler\_output}; GRU_{pooler\_output}]$$

where ";" is the concatenation operator and $X \in \mathcal{R}^d$ is the combined representation of the input from the two sub-modules. The final stance prediction S is then the weighted sum of the two predictions with respect to the hyper-parameter λ:

$$x_{sentiment} = MLP_\theta(g_{out})$$
$$x_{stance} = MLP_\psi(X)$$
$$S = \lambda \cdot x_{sentiment} + (1-\lambda) \cdot x_{stance}$$

Human Values Classification

As far as the prediction of human values is concerned, we propose a different solution that adopts the architecture in Figure 1 for a classification task. In this scenario, the output is a multi-hot encoded vector that returns the probability for each cultural value to be conferred to individuals given their views.

We have explored and compared the performance of different BERT-based models: BERT BASE, BERT LARGE, ROBERTA BASE, ROBERTA LARGE, and DISTILBERT. The BERT-based neural network is followed by a dropout layer and a classification layer, added for the purpose of regularization and classification respectively. The fine-tuning was done on the entire architecture, which allows the model to adapt to the specific task at hand and results in improved performance. After that, we have analyzed the impact of different inputs using the architecture implemented with BERT-base. The first input is simply the concatenation of Conclusion and Sentence, while the second one is a concatenation of Conclusion, Stance, and Premise, with the additional use of the prompting technique.

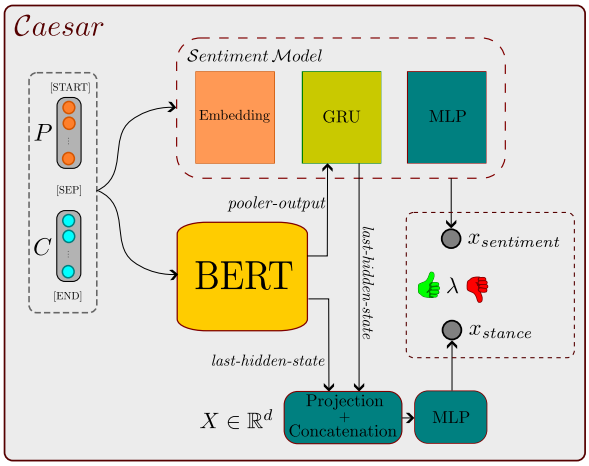

Figure 2: CAESAR architecture

For the additional part, as mentioned in the Introduction, we implemented three different models. For both the DeBERTa and XL-net architectures, we implemented the base and large models. For the former, each output vector produced by the architecture will have a dimension of 768, while this will be 1024 for the latter. Afterwards, we perform a mean-pooling of the output vectors a single vector with a dimension respectively of 768 or 1024. This vector will be the input of the fully-connected layer that will predict the 20 human value categories.

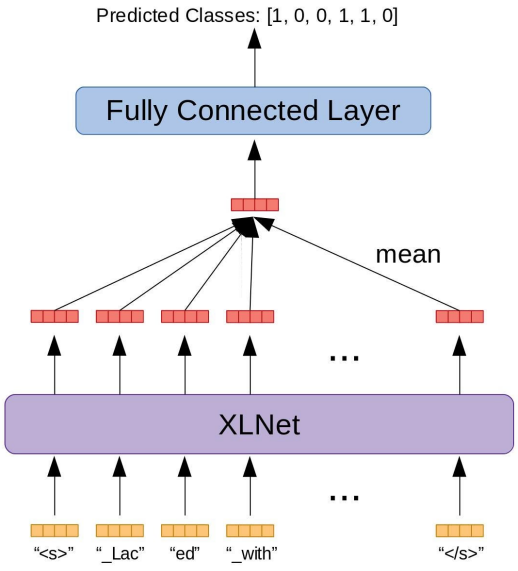

Figure 3: CAESAR architecture

4 Data

As briefly mentioned in Section 1, the dataset used for this project is part of the SemEval 2023 shared task on Human Values Detection (Kiesel et al., 2022). It consists of 5270 arguments annotated by three crowdworkers for all 54 values (represented as black dots) shown in Figure 5 with different levels of abstraction. For the sake of this project, we take into account only the 20 value categories represented in Level 2. The dataset is composed of four different parts: Africa, China, India, and the USA. Existing argument datasets are exclusively from a Western background therefore this dataset is heavily unbalanced towards the USA section, as shown in Figure 4.

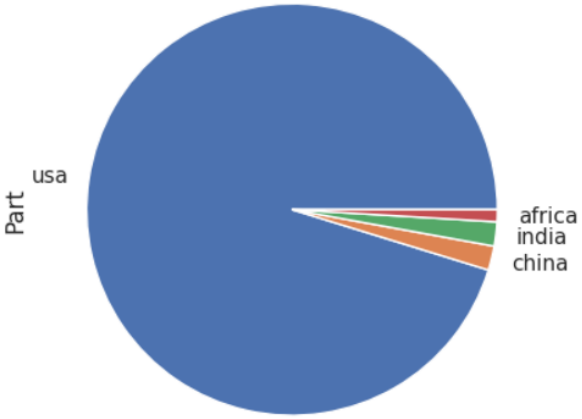

Figure 4: Distribution of the regions over the full dataset.

Hence, due to the difficulty of collecting datasets from various cultures, the data have been taken from different sources:

- **Africa**: 50 arguments have been extracted from editorials of the debating ideas section of a pan-African news platform, "African Arguments"[1].
- **China**: 100 arguments have been extracted from the recommendation and hotlist section of a Chinese question-answering website, "Zhihu".[2]
- **India**: 100 arguments have been extracted from the debate topics 2021 section of "Group Discussion Ideas".[3]
- **USA**: 5020 arguments have been taken with a manual argument quality rating from the arguments of the IBM-ArgQ-Rank-30kArgs dataset (Gretz et al., 2020).

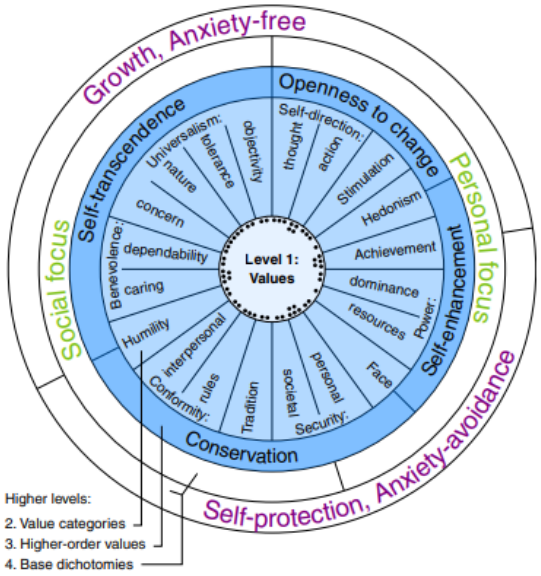

Figure 5: The taxonomy of 54 values that are categorized on the more abstract levels 2-4. The illustration is adapted from (Schwartz et al., 2012).

The dataset consists of four different levels, the higher the level the more abstract the value categories. Given level 3, which contains four Higher-order values, it can be noticed in Figure 6 how the USA part takes up most of the data available. Each argument in the dataset consists of an ID

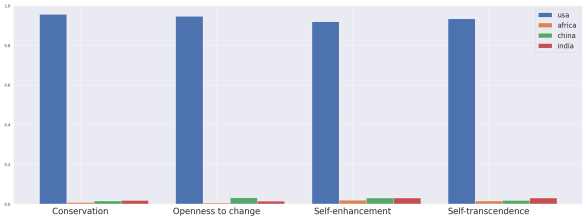

Figure 6: Histograms plotting the distribution of the Higher-order values in Level 3: Conservation, Openness to Change, Self-enhancement, Self-transcendence.

that uniquely identifies the argument, one premise, one conclusion, and a stance attribute indicating whether the premise is in favour of (pro) or against (con) the conclusion. A small example of the taxonomy of the dataset can be seen in Figure 7.

5 Experimental setup and results

As for the human classification tasks, the highest f1 macro average was obtained from Roberta Large

[1] africanarguments.org
[2] zhihu.com
[3] groupdiscussionideas.com

	Argument ID	Part	Usage	Conclusion	Stance	Premise
0	A01001	usa	train	Entrapment should be legalized	in favor of	if entrapment can serve to more easily capture...
1	A01002	usa	train	We should ban human cloning	in favor of	we should ban human cloning as it will only ca...
2	A01003	usa	train	We should abandon marriage	against	marriage is the ultimate commitment to someone...
3	A01004	usa	train	We should ban naturopathy	against	it provides a useful income for some people
4	A01005	usa	train	We should ban fast food	in favor of	fast food should be banned because it is reall...

Figure 7: Dataset sample

Model	f1-macro
Bert base	0.34
Bert large	0.34
Distilled Bert	0.35
Roberta base	0.36
Roberta large	0.41

Table 1: Human Values Prediction models performances

among the five models. For all the BERT models used for the comparison of the performances, it has been used as optimizer Adam, a learning rate equal to 2e-05, a number of epochs equal to 3 and batch size equal to 8. In the input evaluation phase where it has been used a BERT-base model, the hyperparameters remained unchanged except for the number of epochs, which is 10. As for the Stance Prediction models instead, we used 1e-3 as learning rate and trained each network for 15 epochs with batch size 32.

BCE (Binary Cross-Entropy) loss function was used to evaluate the probability of each category individually, rather than compared to other categories. This is because the goal was multi-label classification, not multi-class classification. The sigmoid activation function was used instead of the Softmax activation function, in line with the use of the BCE loss function. Accuracy metrics and F1 scores were calculated using the Scikit-learn package, rather than directly comparing the expected vs. predicted values.

In addition to the concatenation of premise and conclusion, we've devised a simple prompting technique following the intuitions from (Wei et al.), which should elicit the ability to derive a coherent and more meaningful contextual-representation in the BERT-encoder.

Table 2: Comparative results

Input Format	f1-macro
Conclusion + Stance	0.035
Conclusion + Stance + Premise (Prompting)	0.034

Model format	f1-score
Baseline with unbalanced dataset	0.5510
CAESAR with unbalanced dataset	0.5510
Baseline with balanced dataset	0.5714
CAESAR with balanced dataset	0.5510

Table 3: Stance prediction models

6 Discussion

Part - 1 Model Evaluation: This study aimed to assess the performance of various models, input formats, and dataset balancing techniques on a particular dataset through comparative analysis. To accomplish this objective, we evaluated five BERT-based models as they are known for their efficacy in text classification tasks. The primary evaluation metric used to assess the performance of these models was the f1 macro score. Notably, the Roberta large model outperformed the other models slightly, although the other models exhibited commendable scores ranging from 0.3 to 0.4. It is worth noting that the choice of BERT-based models for this study was based on their reputation as state-of-the-art models for text classification tasks. By utilizing these models, we were able to leverage their advanced architecture and pre-trained language models to achieve the best possible results. Our findings suggest that, while the Roberta large model may have a slight advantage over other models, the performance of the other models is also noteworthy, indicating the efficacy of the BERT-based approach in text classification tasks.

Part - 2 Input format evaluations: Apart from that, we chose two input formats to evaluate the performances. Firstly, we used the Conclusion + Stance format in which the focus was primarily on the stance. Secondly, the triple format input, Conclusion + Stance + Premise was used along with a prompting technique to observe the relationships among the three with respect to the dataset. This kind of input format can be useful to gain insights into the author's perspective. In the end, both input formats had similar f1-scores. The reason why performances have remained the same is partly due to the input data being short and concise. Using a better prompting format and batching more classification instances together may actually result in better global performances.

Part - 3 Balancing the Dataset and Stance Predictor models: The dataset analysis has shown the presence of a bias due to the training instances being only from the USA. This can affect the results of the text classification tasks, which is why we attempted to balance the datasets. In view of the cultural differences between Eastern and Western regions - which are even more marked when social and political aspects are concerned - we tried to balance the distribution of human values to endow the networks with better generalization capabilities. As the construction of the training set doesn't allow acting directly on the origin of instances, we opted for a balancing of the human values sets. By doing so, namely by reducing the presence of those sets containing the same classes, we tried to obtain a more evenly distributed representation of human values. Indeed, by lessening the redundancy of data, we hope to have at least limited the cultural values characterizing the US social tissue with respect to all the other social contexts. In our opinion, balancing the dataset allowed the baseline model to learn from more representative data and improved their ability to classify the text.

Since there's no way to tell whether the obtained distribution faithfully mirrors the diversity of human values regardless of the socio-political contexts, we rely on an empirical analysis. As can be seen in table 3, the results of the dataset balancing show that although the CEASAR did not show any significant improvement after balancing the dataset, the baseline model did show substantial progress with its f1-macro score increasing from 0.5510 to 0.5714. This makes us believe that this pre-elaboration has actually made the network capable of a better generalization. Furthermore, we outline how the results obtained from the proposed models are in line with the expectation, especially considering the problem deriving from the format of the input data discussed in section 2.

7 Conclusion

In conclusion, our research aimed to detect human values in social network arguments and to identify opinion polarization. Our study compared various BERT models to select the most effective one for detecting human values. We also investigated different input combinations for multi-label classification, ultimately finding that input 2 (stance conclusion and premise) yielded more promising results. In the model evaluation part, it was observed that the Roberta large model performs the best among the many. we addressed the dataset imbalance by re-balancing it, leading to further improved results. Finally, to tackle stance prediction, we proposed a CAESAR approach that utilized tone sense classification to overcome encoder model limitations. Our study provides insights and solutions for effectively detecting human values in social network arguments.

These proposals can be the object of further discussions and optimizations. The usage of pre-trained models for sentiment analysis and the employment of more extensive prompting techniques may help to improve the performances of the CAESAR model for Stance Prediction and the encoder models for human classification, respectively.

References

Kyunghyun Cho, prefix=van useprefix=true family=Merrienboer, given=Bart, Dzmitry Bahdanau, and Yoshua Bengio. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches.

Kareem Darwish, Walid Magdy, and Tahar Zanouda. Improved Stance Prediction in a User Similarity Feature Space. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17, pages 145–148. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7805–7813.

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.

M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. 13(4):18–28.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. 9(8):1735–1780.

Marjan Hosseinia, Eduard Dragut, and Arjun Mukherjee. Stance Prediction for Contemporary Issues: Data and Experiments.

C. Hutto and Eric Gilbert. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. 8(1):216–225.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. SenseBERT: Driving Some Sense into BERT.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. STANCY: Stance Classification Based on Consistency Cues. 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 6413–6418. Association for Computational Linguistics.

Shalom H. Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, and et al. 2012. Refining the theory of basic individual values. *Journal of Personality and Social Psychology*, 103(4):663–688.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.

Hong Zhang, Haewoon Kwak, Wei Gao, and Jisun An. Wearing Masks Implies Refuting Trump?: Towards Target-specific User Stance Prediction across Events in COVID-19 and US Election 2020.