Cultural Value Resonance in Folktales: A Transformer-based Analysis with the World Value Corpus

Noam Benkler, Scott Friedman, Sonja Schmer-Galunder, Drisana Mosaphir, Vasanth Sarathy, Pavan Kantharaju, Matthew D. McLure, Robert P. Goldman

Smart Information Flow Technologies 319 N. 1st Avenue, Suite 400 Minneapolis, MN 55401 `[first initial][last name]@sift.net` https://www.sift.net

Abstract. Although implicit cultural values are reflected in human narrative texts, few robust computational solutions exist to recognize values that resonate within these texts. In other words, given a statement text and a value text, the task is to predict the label that resonates, conflicts or is neutral with respect to the value. In this paper, we present a novel, annotated dataset and transformer-based model for *Recognizing Value Resonance* (RVR). We created the World Values Corpus (WVC): a labeled collection of [statement, value] pairs of text based on the World Values Survey (WVS), which is a well-validated, comprehensive survey for assessing values across cultures. Each pair expresses whether the value resonates with, conflicts with, or is neutral to the statement. The 384 values in the WVC are derived from the WVS to assure the WVC's cross-cultural relevance. The statement pairs for each value were generated by a pool of six annotators across genders and cultural backgrounds. We demonstrate that off-the-shelf *Recognizing Textual Entailment* (RTE) models perform unfavorably on the RVR task. However, RTE models trained on the WVC achieve substantially higher accuracy on RVR, serving as a strong, replicable baseline for future RVR work, advancing the study of cultural values using computational NLP approaches. We also present results of applying our baseline model on the "World of Tales" corpus, an online repository of international folktales. The results suggest that such a model can provide useful anthropological insights, which in turn is an important step towards facilitating automated ethnographic modeling.

Keywords: Recognizing Values Resonance · World Values Corpus · Folktales Value Analysis

1 Introduction

The study of implicit values embedded in language has been a long-standing research focus of linguists, political scientists, social scientists, and cultural anthropologists. Graeber defines value as the importance of social action through which people demonstrate their belief in what is the good life [7]. Although values are dynamic and changing over time, cultures create different standards of value, and cross-cultural analysis have shown distinct patterns of values relating to well-being [5], trust [10], political support [2], gender gaps and biases [6], and other factors. Despite recent advances in NLP and computational social science, we lack widespread computational methods for recognizing cultural values that resonate in cultural texts. By *resonate*, we mean the degree to which texts amplify, exemplify or embody certain values. For example, we can understand the concept of value resonance on analogy to the sentence "the siren resonated across the harbor." Our approach automatically predicts whether "cultural value X resonates in this text." Importantly, this approach identifies values not explicitly mentioned in the text. It is an open question whether human readers also engage in value resonance reading. To date, cultural values are assessed at different scales - quantitatively (e.g. survey statistics) and qualitatively (e.g. fieldwork, interviews, etc.). However, new methodological approaches combining qualitative and quantitative methods allow us to better understand implicit values in text across languages and cultures.

This paper introduces the NLP problem of *Recognizing Value Resonance* (RVR): given a *statement* text (e.g., sentence(s) from a narrative or utterance) and a *value* text, predict the *label* for whether the value (a) resonates with, (b) conflicts with, or (c) is neutral with respect to the statement. In support of RVR, we present the World Value Corpus (WVC) dataset that includes labeled ⟨*statement, value, label*⟩ examples, and show a proof-of-concept application of RVR to a novel domain, using RVR to depict values that resonate or conflict with sentences of three folktales from different cultural origins. We formulate this problem analogous to the Recognizing Textual Entailment (RTE) problem [13], though our results (Section 3) show that state-of-the-art RTE models cannot predict RVR problems with higher than 60% accuracy (chance is 33%). Our RVR-trained model exceeds 90% accuracy, and we believe that future work can improve upon these results. We continue with a brief overview of background materials, describing the World Value Survey (WVS) in Section 2.1 and RTE models in Section 2.2. We then describe our approach in Section 2, including the problem statement, our labeled World Value Corpus (WVC) for RVR, and an initial, specialized RTE model for RVR. We present encouraging RVR results in Section 3 using a pre-trained RTE model with additional WVC training. We then present a preliminary application (Section 4). This provides early evidence for the applicability of RVR to a new domain (non-Western folktales) without training in the new domain, and shows plots relating to various values across cultures. These plots can be viewed as a distance metric across cultural narratives. We conclude with a discussion (Section 5) of limitations and future work. Our WVC dataset is publicly available at `omitted for blind review`.

2 Approach

Our approach to *Recognizing Value Resonance* involves the creation of a novel corpus based on the WVS, recognizing RVR as a *Recognizing Textual Entail-*

ment-proximal task re-targeted from factual implication to sentimental implication, and tuning a baseline RTE model on WVC data.

2.1 Background: World Value Survey

The World Value Survey [8, 14] is a global research project focused on people's values and beliefs, how they change over time, and how values impact social and political life in different countries of the world. The WVS's central research instrument is a representative comparative survey conducted every 5 years and consists of 290 standard questions. These questions cover 14 distinct thematic domains: social values, attitudes and stereotypes, societal well-being, social capital and trust, economic values, corruption, migration, post-materialistic values, science and technology, religion, security, ethics and norms, politics, and demography.[1] The breadth and depth of value coverage, extensive geographical scope, and open data accessibility make the WVS one of the most authoritative, widely-used cross-national surveys in social science research today.[2]

Previous work has used the WVS to explore determinants of well-being [5], social trust [10], support for democracy [2], gender gaps [6], and more. This paper extends the existing literature to provide a dataset—and preliminary NLP model—to help identify WVS values that resonate within narrative or ethnographic text.

2.2 Problem Statement: Recognizing Value Resonance

The NLP task of *Recognizing Textual Entailment* is: given a premise (statement) and a hypothesis (value), predict whether the facts in the premise necessarily imply the facts in the hypothesis. We define *Recognizing Value Resonance* as the task of, given a statement and a value, predicting whether the value resonates with, conflicts with, or is neutral to the statement. This is a distinctly different, yet proximal task to that of *Recognizing Textual Entailment*. Where RTE focuses on factual implication of a value given a statement, RVR expands this definition to cover endorsement or rejection of a value given a statement. While RVR and RTE are similar in nature we believe RTE models will not perform well when directly applied to the task of RVR and present a few reasons why.

First, RTE (and related NLI) notions cover concepts of entailment and contradiction. The associated datasets (SNLI, SICK, MNLI, etc.) generally contain examples that are inherently propositional in nature. For example, "two dogs are running through the field" from SNLI, is derived from a scene description of an image. The propositional nature of this example allows models to determine if another propositional sentence, say "there are animals outdoors" is true (entailed), might be true (neutral), or is definitely false (contradicted) given the original sentence. By contrast, in RVR, neither the statement nor the value are

[1] In certain nations, additional questions or value domains are covered under the World Values Survey, such as gender norms and family planning.

[2] https://www.worldvaluessurvey.org/WVSContents.jsp

Table 1. Value Resonance Examples Difficult for RTE models, suggesting "resonance" may possess distinct characteristics, different from textual entailment. RVR labels are coded as R (Resonant), N (Neutral) and C (Conflicted). RTE labels are coded as E (Entailment), N (Neutral) and C (Contradict)

Statement (Premise)	Value (Hypothesis)	True RVR Label	Primary RTE Label
It's foul that no one is allowed to perform abortions in this town!	Women should have access to free and safe abortions.	R	N
We live in a monarchy.	Democracy is very important to me.	N	C
We, the faithful will not be harmed by the incursion of science into our lives for science is but the pursuit of God's truth, as is religion.	We depend too much on science and not enough on faith.	C	N

generally propositional. For example, the sentence "we live in a monarchy" in Table 1 is at best "epistemically" propositional in that "the speaker believes they live in a monarchy." This suggests the examples in RVR might have a different distribution from those in RTE-related datasets.

Second, the notion of resonance is inherently different from that of entailment. One might be able to argue that a narrative text epistemically entails a value; but, one might just as easily argue that a value is an implicit social concept presupposed by the speaker in their narrative texts. We believe (and demonstrate) that entailment models, although not sufficient alone, can serve as useful starting point for learning and recognizing value resonance.

2.3 Dataset: The World Value Corpus

We constructed the World Values Corpus as a comprehensive dataset covering all questions in the World Values Survey and its Gender module that met two central criteria: a) the question was not inherently numeric [3] b) the question was easily framed to allow for binary restatement/negation annotation.[4] 335 WVS questions met both of these criteria.[5] Of the 335 questions included in the WVC, 27 required segmentation to ensure comprehensive WVS coverage under the strict binary (restatement/negation) WVC annotation framework. Each question was then converted to a directed value statement while staying as true to the underlying question as possible. These statements acted as parent prompts for annotation. The final list of WVC prompts consists of 384 prompts.

Following prompt generation, 6 annotators reviewed each prompt in the WVC and submitted annotations in one of 4 categories: direct restatement, direct negation, narrative restatement, and narrative negation. Restatements

[3] e.g. "What is your age?", "What is your income level?"

[4] This includes questions easily split into sub-questions that comprehensively cover the response spectrum.

[5] 266 questions from the standard WVS questionnaire and 69 questions from the WVS Gender module.

reaffirm the central message of the parent prompt and negations oppose the central message of the parent prompt. Direct annotations simply reword the prompt (or negate the prompt) whereas narrative annotations are episodic statements that convey the underlying parent prompt (or the negation of that prompt). For complete, single-level coverage we ensure every parent prompt in the WVC is matched to at least one narrative restatement.

Finally, all annotations and prompts were collected in [statement, value, label] sets with annotations as statements, parent prompts as values, and the relation[6] between the two as labels. Each label was re-coded such that restatements became 'resonant' and negations became 'conflicted.' Next, we randomly paired each unique statement with two unrelated values and each unique value with at least two unrelated statements. Our researchers examined each of these [statement, value] pairs to ensure neutrality with respect to the value. This process was repeated until each statement and value was part of at least two true neutral pairs. These pairs were appended to the WVC RVR dataset as sets coded: [statement, value, 'neutral']. Before modeling, we split the WVC RVR dataset into a training set (1114 examples), validation set (275 examples), and testing set (275 examples). The distribution of labels under each split was not significantly different from the underlying WVC according to a 2-sample Kolmogorov-Smirnov Test ($p - val > 0.99$).

2.4 Preliminary Model for Recognizing Value Resonance

We selected an RTE model to use as a baseline for text embedding and finetuning through a preliminary performance evaluation of the top 5 performing RTE models from AllenNLP [4, 12, 15, 17] at the task of RVR. Model performance was evaluated by mean overall accuracy averaged over 5 random splits of our WVC training set, and we chose the transformers [16] implementation of RoBERTa MNLI [12], roberta-large-mnli, as our baseline for finetuning a RVR model. We trained Resonance Tuned RoBERTa, our RVR tuned model, over the WVC training set, evaluating every epoch, and tuning hyperparameters to maximize accuracy over the WVC validation set. We tuned the following hyperparameters: learning rate (1.4e-05), alpha (0.708), momentum (2.16e-2), # training epochs (4), seed (87), batch size (8)[7]. This method was implemented using transformers [16] for training and ray-tune [11] for hyperparameter optimization. Each hyperparameter setting was run on a single machine with population based hyperparameter refinement and resource allocation using a Population Based Training scheduler [9] and stochastic gradient descent for optimization.

[6] Relation meaning: direct restatement, direct negation, narrative restatement, narrative negation.

[7] Final optimal values in parentheses.

3 Results: Evaluation on WVC

Here we describe the results of applying multiple RTE NLP models, including our value-resonance trained model (Section 2.4) on the problem of recognizing value resonance in statements and narratives.

Table 2 displays the comparative performance results of each RTE and RVR model explored in this paper against our WVC derived test set. We find our Resonance-Tuned RoBERTa outperforms all the top performing RTE models from AllenNLP in all performance metrics covered. Resonance Tuned RoBERTa outperformed all RTE baseline models by an accuracy margin of 0.914 over 0.672 and an F1 margin of 0.912 over 0.666.

Table 2. Comparative Model Performance on RVR, evaluated against WVC test-set.[9]

Model	Accuracy	Precision	Recall	F1
Resonance Tuned RoBERTa	0.914	0.92	0.914	0.912
roberta-large-mnli [13]	0.672	0.721	0.672	0.666
RoBERTa MNLI [12]	0.664	0.741	0.664	0.661
RoBERTa SNLI [12]	0.477	0.488	0.477	0.465
Adversarial Binary Gender Bias-Mitigated RoBERTa SNLI [17]	0.477	0.502	0.477	0.476
Binary Gender Bias-Mitigated RoBERTa SNLI [4]	0.445	0.447	0.445	0.431
ELMo-based Decomposable Attention [15]	0.438	0.445	0.438	0.42

Table 3. Breakdown of accuracy with respect to individual class in RVR <Resonant, Neutral, Conflicted>.

Model	Entailment	Neutral	Contradiction
Resonance Tuned RoBERTa	0.938	0.969	0.922
Adversarial Binary Gender Bias-Mitigated RoBERTa SNLI [17]	0.836	0.555	0.562
roberta-large-mnli [12]	0.828	0.695	0.82
RoBERTa MNLI [12]	0.82	0.672	0.836
Binary Gender Bias-Mitigated RoBERTa SNLI [4]	0.805	0.531	0.555
RoBERTa SNLI [12]	0.797	0.57	0.586
ELMo-based Decomposable Attention [15]	0.633	0.672	0.57

Table 3 shows the accuracy for each of the models tested by individual class. Per-class results from table 3 shows Resonance Tuned RoBERTA continues to outperform all the tested competitors on a class by class basis. Confusion matrices (not shown here) indicate that three of the RTE models made fewer False Entailment mistakes than Resonance Tuned RoBERTA, though all three performed worse on scoring True Entailments. Roberta-large-mnli outperformed Resonance

[9] Precision, Recall, and F1 scores are calculated as a weighted average by support for each label <resonant, neutral, conflicted>

Tuned RoBERTA at the task of scoring True Contradictions, though also made many more False Contradiction mistakes than Resonance Tuned RoBERTA.

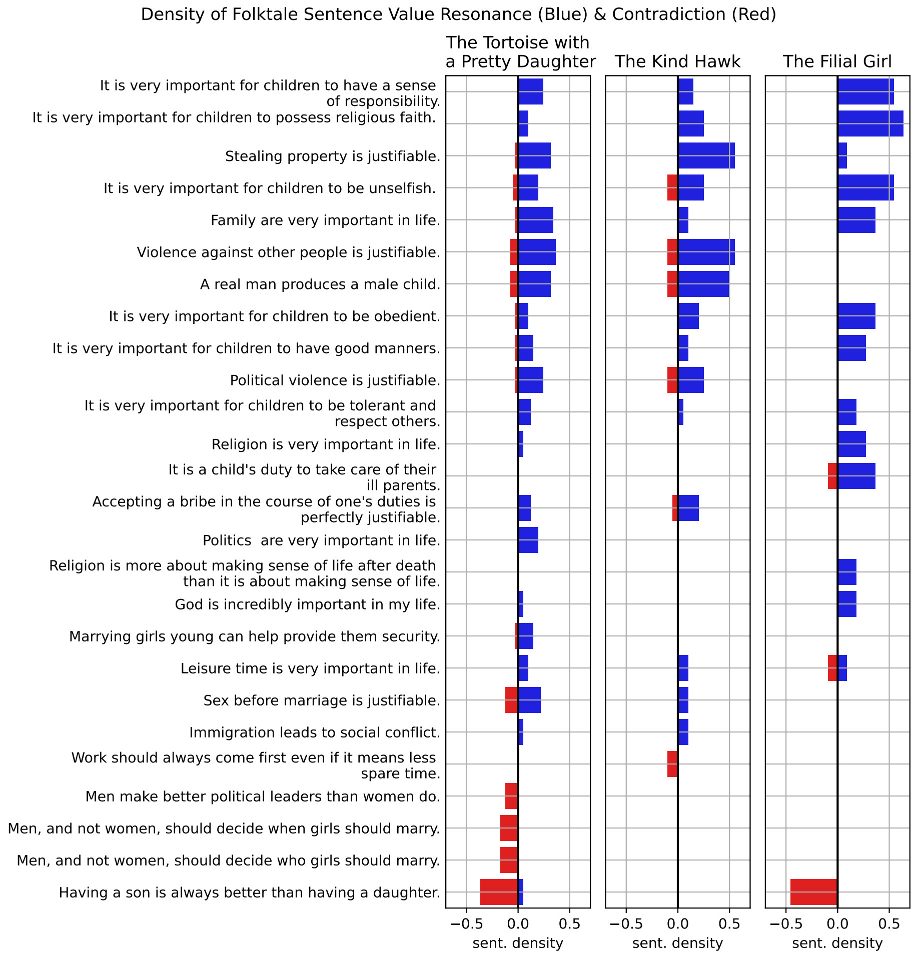

Fig. 1. WVS values resonating (blue, to right of axis) and conflicting (red, to left of axis) with sentences of three folktales, predicted by a RVR NLP model, from 242 possible WVS values.

4 Preliminary Results: Resonance in Folktales

We next present a proof-of-concept application of RVR in a novel domain, to provide early evidence of RVR NLP applicability without any training in the novel domain. For this experiment, we selected folktales from the "World of Tales" online repository of international folktales [1] as a novel domain:

1. *"The Tortoise with a Pretty Daughter"* describes a poor tortoise who had an attractive daughter that is married by a prince and thereby brings wealth to her family. This story reportedly originates from Nigeria.
2. *"The Kind Hawk"* describes a boy who is kidnapped by another tribe, and a hawk then rescues the boy and then steals the kidnappers' belongings for him. This reportedly originates from Native American tradition.
3. *"The Filial Girl"* describes a girl who cares for her parents and worships a mirror after her mother passes away, as the mirror projects her image, which resembles her mother's image. This reportedly originates from Japan.

All three folktales are in English, and we do not have access to the tales in their native languages, so we do not advise attributing the value-based results of these folktales to their reported cultural origins.

For this analysis, we remove a subset of the WVS values pertaining to (1) proper nouns such as the "World Health Organization" and "AIDS," (2) procedures or technologies such as "euthanasia" and "modern contraceptives," and (3) values involving self-reporting such as "I'm often..." and "I have..." and "I feel..." and "My spouse..." since these are less relevant to the folktale domain. After removing the above values, 242 WVS values remain for this folktale-based RVR analysis.

To prepare the folktale data, we sentencized each folktale with spaCy and used RVR to compute the proportion of each folktale's sentences that resonate with, contradict with, and are neutral toward each of the 242 WVS values. Results are shown in Figure 1, plotting each of the three folktales against all WVS values above a 10% sentence threshold, such that at least 10% of any story's sentences must either resonate with or contradict the value. Note that some values in Figure 1 have *both* resonance and contradiction from sentences in a single folktale. This is because, e.g., violence may sometimes be justified and sometimes unjustified, and so-forth.

The distinctive elements of each value plot illustrates the narratives' distinctive values. The "Tortoise" describes a marriage and so includes valence about having and marrying daughters (bottom of Figure 1), including the distinctive "marrying girls young can provide them security" (middle/bottom). The "Hawk" tells of retribution and theft from the kidnappers, and understandably surpasses the other narratives in resonant values of justifying stealing and violence. Compared to the other two, "The Filial Girl" distinctively resonates with values concerning children caring for their parents and concerning life after death.

Some of the common value resonance detected by RVR was unexpected. The possession of religious faith and the importance of religion were both detected in multiple narratives, though religion itself was not mentioned in any of the three folktales. Further, some of the higher-density values detected by RVR in Figure 1 (e.g., responsible children, unselfish children, obedient children) may be prominent in many folktales about children (as are these three folktales), or they may be areas where RVR is over-sensitive. We expect that training RVR in this domain will substantially improve its performance, but these early results are encouraging in a novel domain.

5 Conclusions

This paper has presented a NLP problem of Recognizing Value Resonance (RVR) in text, trained on a World Value Corpus (WVC) dataset derived from human contributions and annotations over the World Value Survey.

We demonstrated that training large RTE language models on RVR data improves their RVR performance over untrained RTE models, suggesting that RTE is related to, and can be repurposed for, the problem of RVR, but RTE alone is not sufficient for RVR.

We present preliminary out-of-domain results applying RVR on international folktales in a small comparative analysis to help characterize distinctive values and shared values. The folktale values detected by RVR cohered with central plot elements and character relationships in each folktale, but importantly, folktales about violence and theft do not necessarily *condone* or *promote* said violence and theft; consequently, applying RVR to folktales requires that we interpret it as a summary of the relationships and issues addressed by (or relevant to) the narrative.

We believe that automation of value recognition in cultural texts and statements will benefit anthropology, sociology, political science, and the social science community. However, as shown in our initial out-of-domain results, RVR models may misidentify cultural values and may contain suboptimal biases, so we do not believe this system is yet ready for large scale deployment for predicting or characterizing human cultures or values.

The WVC includes sentences comprising restatements, negations, narrative restatements and narrative negations of the WVS values, but these are not themselves cultural texts. Consequently, the results reported in this paper—even the folktale results—should not be interpreted as RVR performance on real cultural texts or as values held by human cultures.

Important ethical issues of consent from cultures to having their language and texts digitized, or analyzed in this way should be addressed on a case by case basis, and in equitable consultation with these cultures, e.g., [3].

While our initial results are encouraging, important work remains. Extending the WVC to new domains–folktales, ethnography, and social media–will improve its coverage, reduce error rates and increase its applicability.

As RVR can produce a value vector over a text, story, or corpus (see Figure 1), it might be usable as a value-based distance metric over texts or corpora.

Finally, as with previous work on RTE, RVR models and RVR datasets should be extensible, including (1) new (non-WVS) value statements and (2) new narrative sentences that are assessed with respect to WVS and non-WVS values. The present work uses the WVS as a source of vetted cross-cultural values, but we do not believe the WVS to be complete over all cultures or over time with respect to any single culture.

References

1. Andonov, V.: World of tales - stories for children, folktales, fairy tales and fables from around the world!, https://www.worldoftales.com/
2. Ariely, G., Davidov, E.: Can we rate public support for democracy in a comparable way? cross-national equivalence of democratic attitudes in the world value survey. Social Indicators Research **104**(2), 271–286 (2011)
3. Cormack, D., Kukutai, T.: Indigenous peoples, data, and the coloniality of surveillance. In: New Perspectives in Critical Data Studies, pp. 121–141. Springer (2022)
4. Dev, S., Li, T., Phillips, J.M., Srikumar, V.: On measuring and mitigating biased inferences of word embeddings. Proceedings of the AAAI Conference on Artificial Intelligence **34**(05), 7659–7666 (2020). https://doi.org/10.1609/aaai.v34i05.6267
5. Fleche, S., Smith, C., Sorsa, P.: Exploring determinants of subjective wellbeing in oecd countries: Evidence from the world values survey. Organization for Economic Co-Operation and Development (2012)
6. Friedman, S., Schmer-Galunder, S., Chen, A., Goldman, R., Rye, J.: Relating linguistic gender bias, gender values, and gender gaps: An international analysis. In: BRiMS Conference, Washington, DC, July (2019)
7. Graeber, D.: Toward an anthropological theory of value: The false coin of our own dreams. Springer (2001)
8. Inglehart, R., Basanez, M., Diez-Medrano, J., Halman, L., Luijkx, R.: World values surveys and european values surveys, 1981-1984, 1990-1993, and 1995-1997. Ann Arbor-Michigan, Institute for Social Research, ICPSR version (2000)
9. Jaderberg, M., Dalibard, V., Osindero, S., Czarnecki, W.M., Donahue, J., Razavi, A., Vinyals, O., Green, T., Dunning, I., Simonyan, K., Fernando, C., Kavukcuoglu, K.: Population based training of neural networks (2017)
10. Jen, M.H., Sund, E.R., Johnston, R., Jones, K.: Trustful societies, trustful individuals, and health: An analysis of self-rated health and social trust using the world value survey. Health & place **16**(5), 1022–1029 (2010)
11. Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J.E., Stoica, I.: Tune: A research platform for distributed model selection and training. arXiv preprint arXiv:1807.05118 (2018)
12. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. ArXiv **abs/1907.11692** (2019)
13. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
14. Minkov, M.: What makes us different and similar: A new interpretation of the World Values Survey and other cross-cultural data. Klasika i Stil Publishing House Sofia, Bulgaria (2007)
15. Parikh, A.P., T"ackstr"om, O., Das, D., Uszkoreit, J.: A decomposable attention model for natural language inference. ArXiv **abs/1606.01933** (2016)
16. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Huggingface's transformers: State-of-the-art natural language processing (2020)
17. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (2018)