Identifying Links Between Human Values and Persuasion

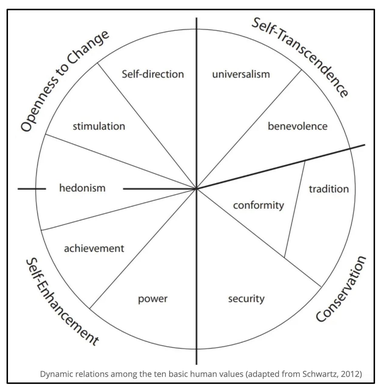
Jenna Donaldson, Anatolii Evdokimov, Na'Dyah Wynn, Jon Park
University of Richmond, VA 23173

We hypothesize that shared values play a key role in the outcome of persuasive discourse. More specifically, a persuader who makes arguments rooted in values the persuadee cares about may be more likely to succeed in pursuing them than those who do not. This project is an initial study to test this hypothesis using large-scale real-world conversations. We first present the Schwartz Value Theory, which consists of 10 major human values. Then, we describe the raw dataset crawled from the ChangeMyView subreddit, carefully filtered and structured to capture the link between human values and persuasion. Lastly, we discuss our initial attempt at annotating the real-world conversations using the Schwartz Value Theory.

I. Schwartz Value Theory [1]

The Schwartz Theory of Basic Human Values defines 10 personal values of people across all cultures that may inform how a person acts or forms beliefs. We chose the Schwartz Value Theory following the work of ValueNet [2]. Our annotations were based on the following values and keyword definitions:

- **self direction**: curious, independent, exploration, identity
- **stimulation**: daring, excitement, adventure, intense
- **hedonism:** pleasure, enjoy, amusement, satisfaction
- **achievement:** successful, intelligent, talented, completion
- **power:** authority, recognition, influence, force
- **security:** health, safety, public, welfare
- **conformity:** discipline, obedient, respectful, compliant
- **tradition:** humble, respect, religion, integrity
- **benevolence:** spiritual, friendship, responsible, kindness
- **universalism:** equality, unity, moral, understanding



Dynamic relations among the ten basic human values (adapted from Schwartz, 2012)

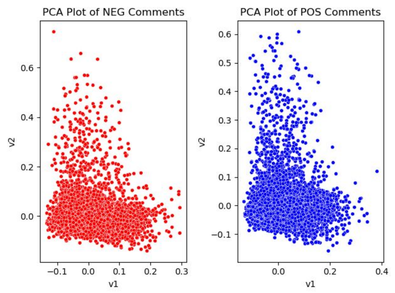
The above chart shows the relationships between various values:
- close values are more similar
- distant values are more in opposition to one another

II. Data

- Original dataset retrieved from the **Reddit API** for the **ChangeMyView** sub-reddit
- Dataset consisted of over **146-thousand** posts and comments
- Filtered into **2370 triplets,** where each triplet is **one post, one positive comment,** and **one negative comment**
  - positive comments are defined as comments that received a **delta (Δ)** from the original poster

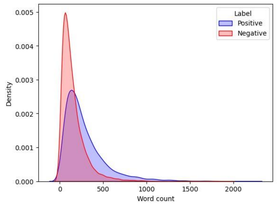
Average #	Positive	Negative
words	289.58	154.65
sentences	13.15	7.52
Median		
words	216	111
sentences	10	5
Total # Triplets	9318	

Table shows counts for triplets with no length restriction

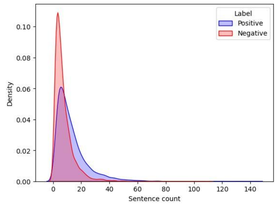


Positive and negative samples overlap, which posed difficulties with the annotation process.

The guideline for the filtering is as follows:
- **selftext**: does the post contain a text body
- **body:** does the comment contain a body that has 2-15 sentences
- **nest-level:** is the comment at nest level 1 and did OP respond
- **author**: does the comment have a listed author
- **delta**: does at least one comment in the thread receive a delta
- **upvote**: choose the highest score of both positive and negative
- **triplet**: does the entire thread contain 1 post, 1 neg, and 1 pos



Density plot of positive and negative samples based on average words



Density plot of positive and negative samples based on average sentences

Both plots have negative and positive samples that were unrestricted in length

III. Annotation

We ran 3 rounds of test annotations before preparing an experiment through Amazon's Mechanical Turk (unreleased):

1. Preliminary round for **comparing two value systems** where only relevant clauses were given a value (chose to follow ValueNet in using Schwartz Value Theory)
2. Data separated by sentence and not every sentence was given a value; one **post/comment can have more than one value**
3. Annotated sentences from **Value Net [2] to compare annotation** decisions/understanding of values

> Plenty of young women, such as myself, have been advanced on by male taxi/Uber drivers and as a result of this, would feel safer getting into a car with a female.

Example of a sentence labeled "security"

MTurk Preparation:
- Given low annotator agreement, we decided to move to labeling for values at the **post/comments level** and with regard to every value (select 0, 1, or -1) where one sentence had to be selected as **most exemplary** of the chosen value
- Edited one sentence of a selection of posts/comments to manufacture specific value annotations (to be used in verifying the credibility of MTurkers)

IV. Conclusion

The **initial annotation** of all the rounds showed the agreement score to be **relatively low**. The project ended after preparing the dataset for Mechanical Turk; however, it was **not launched** on MTurk due to the **varying agreement** on all the rounds. It is acknowledged that annotations would be better with **'higher level classes'** where the annotators would choose between **fewer values**.

References

[1] Shalom H. Schwartz. 2012, **An Overview of the Schwartz Theory of Basic Values**. *Online Readings in Psychology and Culture*
[2] Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, Song-Chun Zhu. 2022, **ValueNet: A New Dataset for Human Driven Dialogue** . *AAAI-22* pages 11183-11191.

Acknowledgements

We would like to thank the University of Richmond and the School of Arts and Sciences for their support in allowing us the opportunity to work on this project.