

Social Orientation: A New Feature for Dialogue Analysis

Todd Morrill*, Zhaoyuan Deng*, Yanda Chen*, Amith Ananthram*, Colin Wayne Leach†, Kathleen McKeown*
Columbia University* Barnard College, Columbia University†

May 2024

Summary

Main Idea: Use social orientation features (e.g. Warm-Agreeable) to help predict and explain success/failure of a dialogue (e.g., CGA)

Sample Conversation

(1) **Speaker A:** Tranquilizer is supposed to be the first option. The gorilla did not demonstrate threatening behavior. My honest opinion: The gorilla is an endangered species. The human child is not. Either the parents did not educate their child about the dangerous nature of animals, or the child was willfully disobedient. Either way, the parents were obviously not watching their child. I don't think a critically endangered animal should suffer the consequences.

(2) **Speaker B:** But, to be clear, you do think that a naturally curious human child should suffer the consequences of someone else's mistake?

(3) **Speaker C:** >;you do think that a naturally curious human child should suffer the consequences of someone else's mistake? Flipping that around: Do you think a naturally acting animal should die for the curiosity of a child?

(4) **Speaker D:** simple decision. Yes. And anybody that puts the natural inner-species solidarity for survival of humans in question should get kicked out of the group. Go to your gorilla friends and see how they treat you.

Social Orientation Predictions

→ **Assured-Dominant**

→ **Unassured-Submissive**

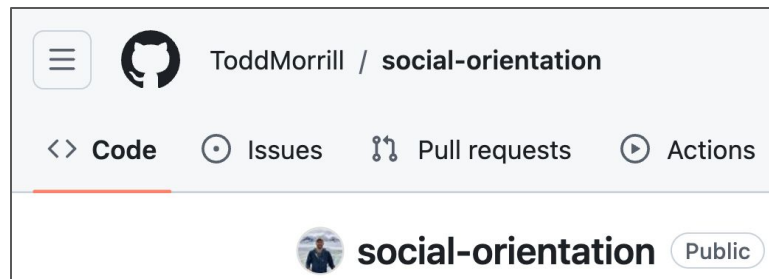
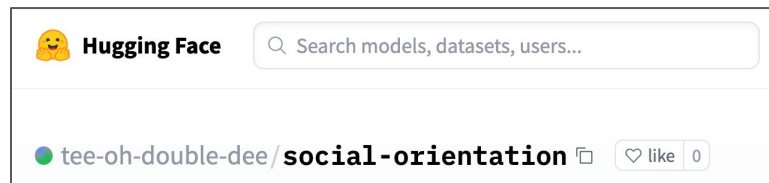
→ **Gregarious-Extraverted**

→ **Cold**

Predicted Conversation Outcome: Failure

Contributions

1. **New data set** of dialogue utterances labeled with social orientation tags and a **distilled model** trained to predict these tags (see **HuggingFace & Github**)
2. **SOTA performance** on 2 English dialogue outcome prediction data sets by using social orientation features
3. We construct a **new Chinese** dialogue outcome prediction **data set** and show that applying social orientation features increases task performance in a second language.
4. We demonstrate that including social orientation features in neural models increases **explainability** for dialogue outcome prediction tasks.
5. We show that in **low-resource** settings, social orientation features are more effective than text-only neural models.



HuggingFace 

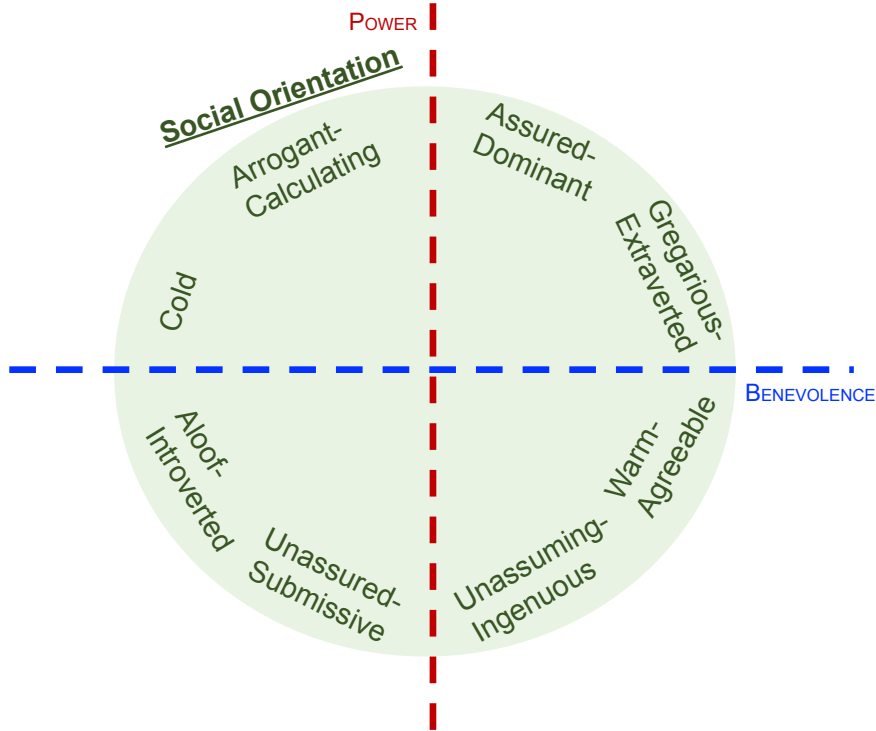


Github 

Introduction

Dialogue Outcome Predictions & Explanations Circumplex

Interpersonal Circumplex



- Customer service interactions, business negotiations, diplomatic discussions, and online forums are often contentious
 - **Wikipedia** page edit discussions
 - **Reddit** r/ChangeMyView
- We'd like to **predict** the outcomes of these interactions ahead of time
- AND **explain** why the outcome occurred
- Circumplex theory and **social orientation** tags attempts to model interactions between conversations participants

Related Work

Dialogue Outcome Prediction Modeling

- Our work most notably builds on top of the **Conversations Gone Awry** data sets (Zhang et al., 2018; Chang and Danescu-Niculescu-Mizil, 2019).
- Other works have predicted **negotiation** outcomes (Lewis et al., 2017), **debate** outcomes (Zhang et al., 2016), etc.
- Other works emphasize the use of **interpretable features**, e.g., age, sex, and social network features (Saveski et al., 2021)
- Previous works have used variants of circumplex theory for e.g., instant message conversations (Vaasen et al., 2012)

A1: Why there's no mention of it here? Namely, an altercation with a foreign intelligence group? True, by the standards of sources some require it wouldn't even come close, not to mention having some really weak points, but it doesn't mean that it doesn't exist.

A2: So what you're saying is we should put a bad source in the article because it exists?

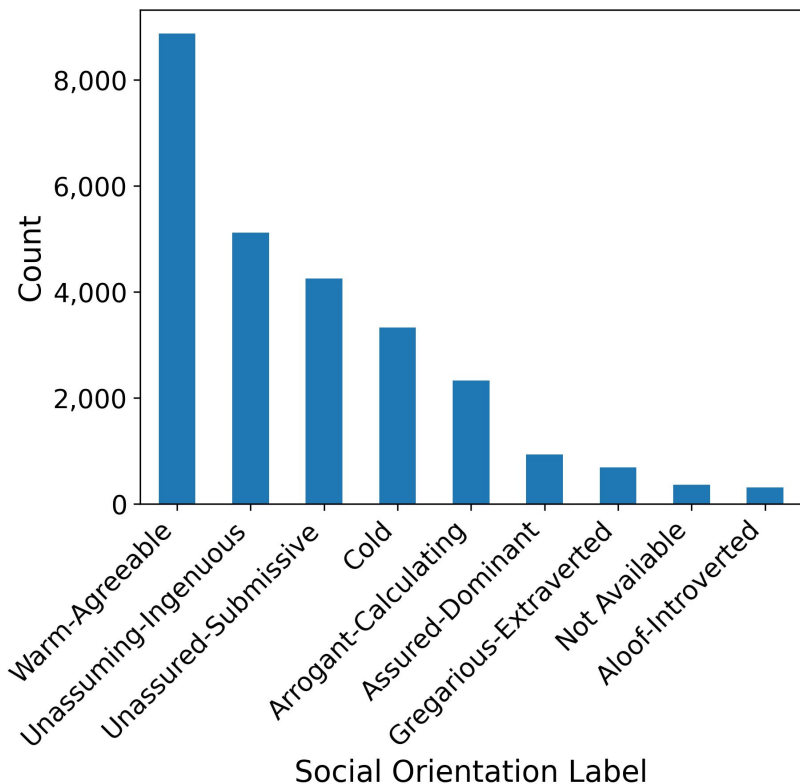
B1: Is the St. Petersburg Times considered a reliable source by wikipedia? It seems that the bulk of this article is coming from that one article, which speculates about missile launches and UFOs. I'm going to go through and try and find corroborating sources and maybe do a rewrite of the article. I don't think this article should rely on one so-so source.

B2: I would assume that it's as reliable as any other mainstream news source.

Figure 1: Two examples of initial exchanges from conversations concerning disagreements between editors working on the Wikipedia article about the Dyatlov Pass Incident. Only one of the conversations will eventually turn awry, with an interlocutor launching into a personal attack.

Collecting Social Orientation Tags

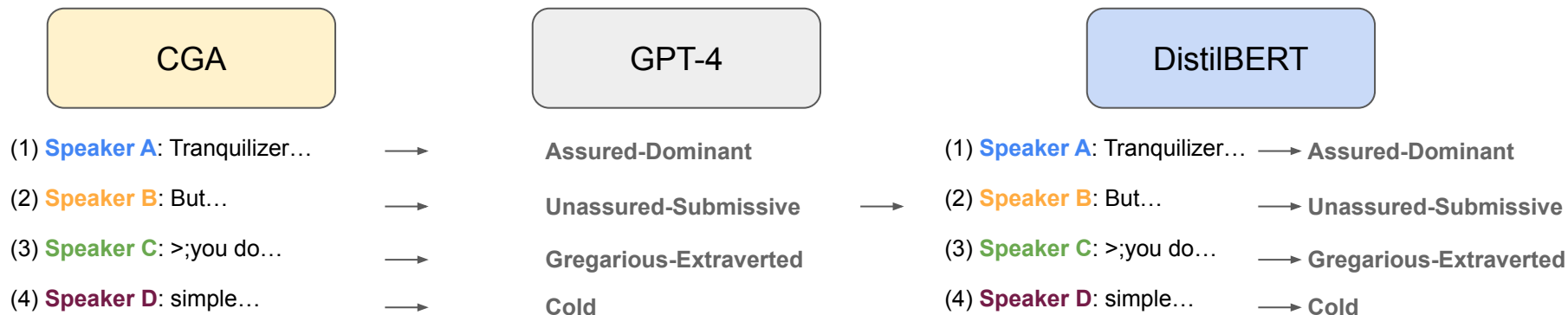
Collecting the Social Orientation Labels



- We collect social orientation tags for all 30,021 utterances in the Wikipedia portion of the Conversations Gone Awry (CGA) corpus using GPT-4
 - CGA conversations were manually reviewed for dialogue outcome
- 3 human annotators hand labeled 423 dialogue utterances resulting in a Fleiss' Kappa of 0.42, indicating moderate agreement
- Agreement with GPT-4 ranged from 20-30%, with most disagreements among neighboring tags
 - Cold versus Arrogant-Calculating
 - Humans used Assured-Dominant more
- Tags are nonetheless useful in downstream tasks, as our results show

Methods

Making Use of Social Orientation Tags



Downstream Dialogue Outcome Prediction Features

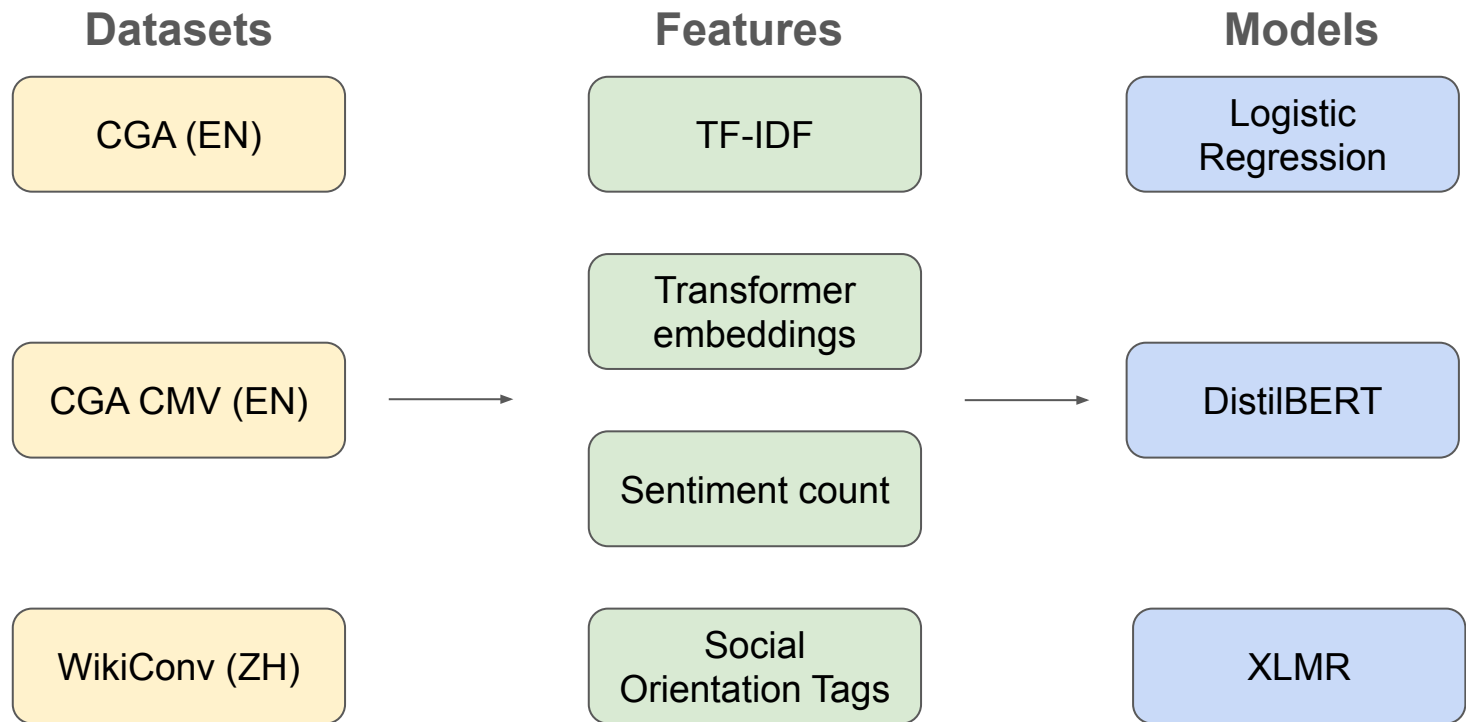
SpeakerA (Warm-Agreeable): That sounds like a good plan.

or

{Warm-Agreeable: 2, Unassured-Submissive: 2, Cold: 1, ...}

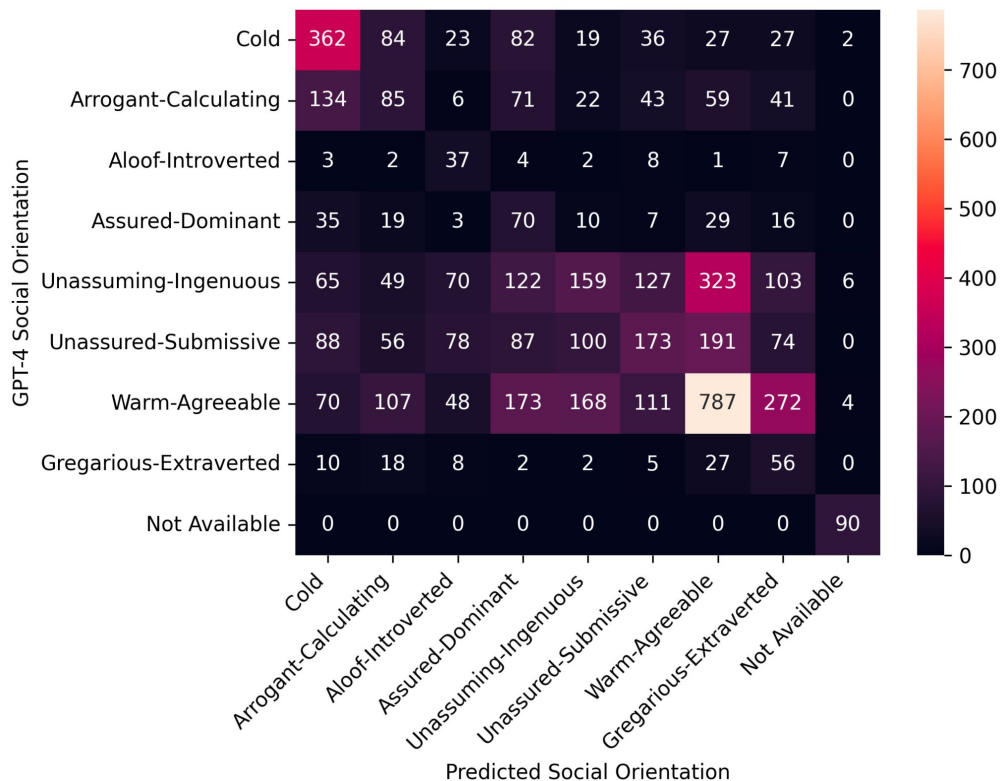
Experiments

Modeling



Results

Social Orientation Tagger



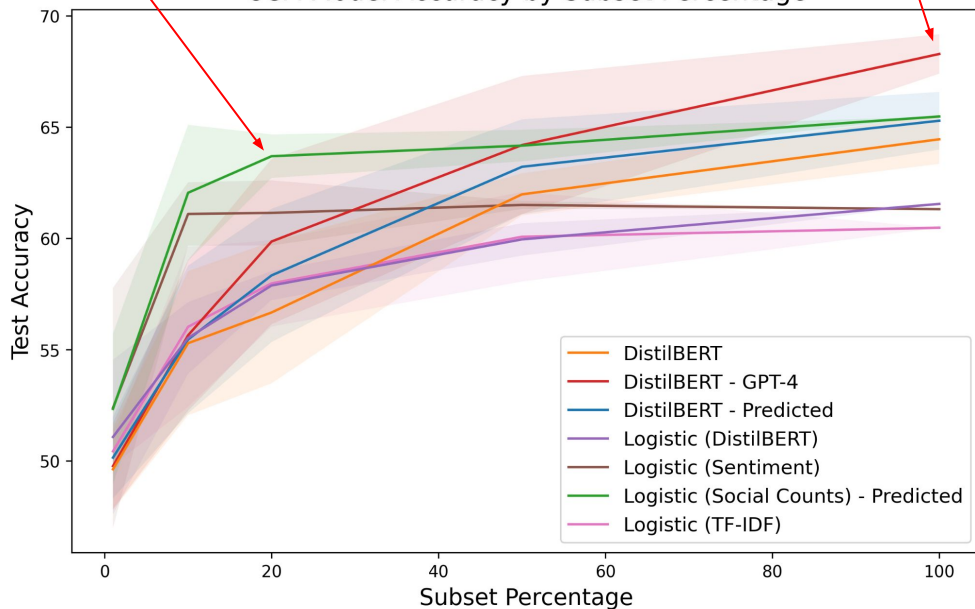
- 35% accuracy on this 8 way social orientation tagging task
- Using a weighted loss function to get more explanatory power in predictions
- 45%+ accuracy achievable without loss weighting
- Most misclassifications cluster around neighboring labels

Social Orientation Tags Increase Accuracy

Low-resource performance

SOTA performance

CGA Model Accuracy by Subset Percentage



- Social orientation features help deep learning models achieve state-of-the-art results
 - 68.29% accuracy on CGA (DistilBERT + GPT-4)
 - 65.01% accuracy on CGA CMV (DistilBERT - Predicted)
- Social orientation features outperform text-only models in low-resource settings

Explainability

- Is the model using social orientation tags as predicted by the theory?
- Make interventions on social orientation tags and observe change in model behavior.

SpeakerA (Cold): I don't know what you were thinking when you made that edit.

→

SpeakerA (Unassuming-Ingenuous): I don't know what you were thinking when you made that edit.

Intervention	Pos2Neg	Neg2Pos	Same
(1) Random	81	194	1,093
(2) (Unassuming-Ingenuous, Unassured-Submissive)	0	154	333
(3) (Arrogant-Calculating, Cold)	15	0	64
(4) (Assured-Dominant, Assured-Dominant)	27	0	214

Conclusion

Limitations, Conclusion

Limitations

- Important to have some dynamic range in the conversations and outcomes

Conclusion

- Created a new data set of dialogue utterances labeled with social orientation tags
- Demonstrated that social orientation tags are effective for predicting dialogue outcomes:
 - When used in deep learning models
 - And when used in low-resource settings (i.e., little data, logistic regression)
- Showed explainability of dialogue outcome predictions

Sources

- Justine Zhang, Jonathan Chang, Cristian DanescuNiculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.
- Jonathan P. Chang and Cristian DanescuNiculescu-Mizil. 2019. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. 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 4743–4754, Hong Kong, China. Association for Computational Linguistics.
- Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2443–2453, Copenhagen, Denmark. Association for Computational Linguistics.
- Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in Oxford-style debates. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 136–141, San Diego, California. Association for Computational Linguistics.
- Martin Saveski, Brandon Roy, and Deb Roy. 2021. The structure of toxic conversations on twitter. In Proceedings of the Web Conference 2021, WWW '21, pages 1086–1097, New York, NY, USA. Association for Computing Machinery.
- F. Vaasen, J. Wauters, F. Van Broeckhoven, M. Van Overveldt, W. Daelemans, K. Eneman, and P. Felicia. 2012. delearyous: Training interpersonal communication skills using unconstrained text input. In Proceedings of ECGBL 2012, The 6th European Conference on Games Based Learning, volume 6, pages 505–513. ACAD CONFERENCES LTD.

Questions

tm3229@columbia.edu