Is this the real life? The language of Reality Monitoring.

Keywords: Reality Monitoring, Natural Language Processing, Neural networks, Memory, Metacognition.

Extended Abstract

Reality monitoring is the process of attributing the source of a memory to an internal or external source, allowing to separate imagination from memory for real events[1,2]. Reality monitoring relies on cues from the different experiences of recollection vs. imagination[3].

Our research explores reality monitoring using Natural Language Processing (NLProc). This project involves training a language-based classification model for real and imaginary texts.
To this end, we collected texts of participants (N=1,054) recounting real and recombined events.
We trained classification models using several different approaches: 1) A radial SVM model for binary classification using 92 linguistic features derived from the texts, (e.g., LIWC dictionaries[4], psychological norms[5]). 2) A Transformer based model (e.g., BERT[6,7], GPT-3) fine-tuned for binary classification. 3) Human Based Direct Classification – A new sample of participants (N=820) read and rated 10 essays each on a scale of 1-100, with the goal of attaining at least 3 classifications per essay, the classifications were then aggregated per essay to form a human based classification, serving as a benchmark for model comparison. 4) Indirect human classification – In the same study used to attain the direct human rating model, participants also answered a series of questions about the stories they read (e.g., how likeable do you find the author? How vivid was the description in the story?), these human derived features of the text were modelled using logistic regression to provide an indirect human-based model, examining whether additional information regarding the essays was registered by human raters. Classification accuracy scores for the training set and held-out test set are shown in Table 1.

The transformer model outperformed the feature-based model and human classification. We further used the feature based model to explore the phenomenal differences between imagination and recollection. Several linguistic features (e.g., Number of episodic/non-episodic details) consistently (i.e., across train & test set) differentiated between real and imaginary events.

Our results demonstrate the feasibility of using NLProc in cognitive research, the differences between human raters and language models, and the degree of overlap/distinction between these separate approaches. Our language based reality-monitoring model can serve as a useful tool for research of state and trait reality monitoring via text analysis.

This corpus will soon be publicly shared, presenting the corpus and classification task as a challenge inviting other research groups to attempt to achieve higher classification accuracies on the test set, to maximize reliability of such a tool for studying individual differences in reality monitoring.

References

1. Johnson, M. K. & Raye, C. L. Reality monitoring. *Psychol. Rev.* **88**, 67 (1981).
2. Johnson, M. K. Memory and reality. *Am. Psychol.* (2006) doi:10.1037/0003-066X.61.8.760.
3. Johnson, M. K., Foley, M. A., Suengas, A. G. & Raye, C. L. Phenomenal characteristics of memories for perceived and imagined autobiographical events. *J. Exp. Psychol. Gen.* **117**, 371–376 (1988).
4. Boyd, R. L., Ashokkumar, A., Seraj, S. & Pennebaker, J. W. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin* 1–47 (2022).
5. Muraki, E. J., Abdalla, S., Brysbaert, M. & Pexman, P. M. Concreteness ratings for 62,000 English multiword expressions. *Behav. Res. Methods* (2022) doi:10.3758/s13428-022-01912-6.
6. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv [cs.CL]* (2018).
7. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv [cs.CL]* (2019).

Model	**Dataset**	**Accuracy**	**Precision**	**Recall**	**F1**	**AUC**
Transformer (DistilBERT + GPT)	Train (n=1,686)	0.82	0.82	0.81	0.82	0.85
	Test (n=422)	0.65	0.66	0.63	0.64	0.68
Linguistic Features (Radial SVM)	Train	0.67	0.67	0.67	0.67	0.74
	Test	0.59	0.58	0.62	0.60	0.60
Human Direct	Train	0.56	0.59	0.43	0.49	0.58
	Test	0.59	0.61	0.47	0.53	0.62
Human Indirect	Train	0.55	0.56	0.50	0.53	0.58
	Test	0.60	0.61	0.56	0.58	0.62
Bag of Words (Naïve Bayes)	Train	0.76	0.77	0.76	0.76	0.82
	Test	0.57	0.57	0.56	0.57	0.58

Table 1. Classification Accuracies in Reality Monitoring Corpus.