Additional Study: Convergence on the 20-newsgroup dataset

Materials and methods

In this study we used a subset of the 20-newsgroup dataset obtained from the scikit-learn[1] Python package (Pedregosa et al., 2011) to investigate the convergence between BERTAgent and other tools and measures used to quantify agency in textual data. This dataset contains messages pertaining to a variety of topics divided into categories based on online communities. It provides a diverse sample of natural language used in online communities, and it also is a well-established general-purpose benchmarking dataset previously used for reference or training in a variety of NLP applications, such as text classification and clustering (e.g., Gliozzo et al., 2005; Hu et al., 2021; Ruff et al., 2019).

Motivated by computational power and memory constraints, we randomly selected nine (out of 20) thematic newsgroups. From the initial dataset (*N*=8,079), we kept only messages that contained a nonempty message body text (*N*=7,850) covering several topics including sport, religion and political issues (details provided in SOM).

We contrasted agency quantification results obtained from BERTAgent with those obtained from other agency quantification indices using two sets of correlation-based comparisons, namely:

- comparison of BERTAgent agency-positive measures (BAPos, BATot and BAAbs) with other agency-positive measures (agency-positive convergent validity; Table 9);
- comparison of the BERTAgent agency-negative measure (BANeg) with other agency-negative measures (agency-negative convergent validity; Table 10);

[1] https://scikit-learn.org

Results and discussion

Our expectation of observing positive correlations in the two comparisons were confirmed. In the first comparison (see Table 9), we found that BAPos, BATot and BAAbs converge with all other agency-positive measures, although with small correlation values, and with, predominantly, small/medium effect sizes (see SOM for the full correlations table). In light of the results obtained in Studies 1, 2 and 3, low correlations values are expected, since we already demonstrated that BERTAgent outperforms other tools in terms of agreement with human ratings, i.e., it is the closest to the "ground truth". This low degree of concordance suggests a low reliability of at least some of the existing DWC-based tools.

In the second comparison (see Table 10), as expected, we found BANeg to be positively correlated with all agency-negative measures. Here we also obtained predominantly small and medium effect sizes, which also confirms the above discussion.

Table 9

Expected positive correlations for BERTAgent's indices BAPos, BATot and BAAbs and other agency-positive measures.

Variable	*M*	*SD*	1	2	3
1. BAPos	0.11	0.06			
2. BATot	0.07	0.09	.89** [.88, .89]		
3. BAAbs	0.14	0.06	.76** [.75, .77]	.38** [.36, .39]	
4. PietA	0.05	0.04	.11** [.09, .14]	.07** [.05, .09]	.13** [.11, .15]
5. PietB	0.01	0.02	.18** [.16, .20]	.13** [.11, .15]	.17** [.15, .19]
6. PietC	0.04	0.03	.15** [.13, .17]	.10** [.08, .13]	.16** [.14, .18]
7. NicoPos	0.02	0.02	.15** [.13, .17]	.13** [.11, .15]	.12** [.09, .14]
8. NicoCom	0.02	0.02	.16** [.14, .18]	.15** [.13, .18]	.11** [.08, .13]
9. NicoAbilityPos	0.04	0.03	.10** [.08, .13]	.08** [.06, .10]	.10** [.08, .12]
10. NicoStatusPos	0.02	0.02	.19** [.17, .21]	.16** [.14, .18]	.16** [.14, .18]

Note. This table uses the same notation of Table 4, to which we refer. In addition - **NicoAbilityPos**: ability-positive dictionary from Nicolas et al., (2021); **NicoStatusPos**: status-positive dictionary from Nicolas et al., (2021).

Table 10

Expected positive correlations for BERTAgent's index BANeg and other agency-negative measures

Variable	*M*	*SD*	1
1. BANeg	0.03	0.04	
2. NicoNeg	0.00	0.01	.09** [.07, .12]
3. NicoAbilityNeg	0.00	0.01	.19** [.17, .21]
4. NicoStatusNeg	0.01	0.01	.11** [.09, .13]

Note. This table uses the same notation of Table 4, to which we refer. In addition - **NicoAbilityNeg**: ability-negative dictionary from Nicolas et al., (2021); **NicoStatusNeg**: status-negative dictionary from Nicolas et al., (2021).