♠ SPADE: A Big Five-Mturk Dataset of Argumentative Speech Enriched with Socio-Demographics for Personality Detection

Elma Kerz[1], Yu Qiao[1], Sourabh Zanwar[1] Daniel Wiechmann[2],

[1] RWTH Aachen University, Germany, [2] University of Amsterdam, Netherlands

elma.kerz@ifaar.rwth-aachen.de, yu.qiao@rwth-aachen.de, sourabh.zanwar@rwth-aachen.de, d.wiechmann@uva.nl

Abstract

In recent years, there has been increasing interest in automatic personality detection based on language. Progress in this area is highly contingent upon the availability of datasets and benchmark corpora. However, publicly available datasets for modeling and predicting personality traits are still scarce. While recent efforts to create such datasets from social media (Twitter, Reddit) are to be applauded, they often do not include continuous and contextualized language use. In this paper, we introduce ♠ SPADE, the first dataset with continuous samples of argumentative speech labeled with the Big Five personality traits and enriched with socio-demographic data (age, gender, education level, language background). We provide benchmark models for this dataset to facilitate further research and conduct extensive experiments. Our models leverage 436 (psycho)linguistic features extracted from transcribed speech and speaker-level metainformation with transformers. We conduct feature ablation experiments to investigate which types of features contribute to the prediction of individual personality traits.

The ♠ SPADE Dataset

- Crowdsourced dataset consisting of 333 speech samples from 214 speakers (20 hours of speech), covering three topics:
  1. Climate change is the greatest threat facing humanity today
  2. People should be legally required to get vaccinated
  3. The development of artificial intelligence will help humanity
- 849k words of transcribed speech; mean length of the speech transcripts: 2493 words (SD = 752.1 words)

Age	Gender	Education	Language Status
Min. :18.00	Male: 129	BA : 138	Mono: 122
1st Qu.:25.00	Female: 88	HS : 40	Biling: 57
Mean :32.27	Diverse: 3	MA : 40	L2 Eng: 38
3rd Qu.:35.00		PHD: 1	NA: 3
Max. :78.00		NA: 1	

Dimension	Mean	SD	Min	Max
Openness	0.107	0.343	-1.239	1.268
Conscientiousness	-0.104	0.461	-2.053	1.968
Extraversion	-0.090	0.437	-1.953	1.051
Agreeableness	0.161	0.577	-3.045	1.721
Neuroticism	-0.049	0.477	-2.177	1.363

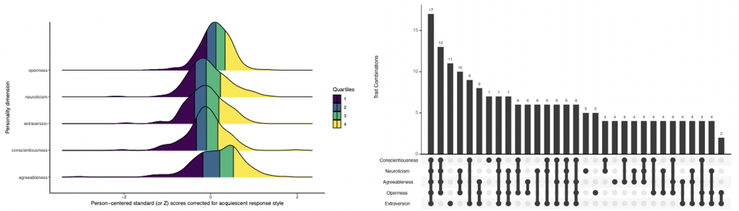

Fig. 1: Frequency of personality trait combinations. For each of the five personality dimensions, a trait was considered present when an individual's BFI score was greater than the group median on a given dimension.

Experimental setup: Benchmarking

Extraction of linguistic features

- The speech transcripts were automatically analyzed using CoCoGen a computational tool that implements a sliding window technique to calculate sentence-level measurements that capture the within-text distributions of scores for a given language feature.
- We extract a total of 436 features that fall into nine categories:
  1. syntactic complexity (N=16)
  2. lexical richness (N=15)
  3. information theoretic (N =3)
  4. register-based n-gram frequency (N=25)
  5. readability (N=14)
  6. psycholinguistic (N=37)
  7. LIWC-style (N=61)
  8. sentiment-related (N=209)
  9. emotion-related (N=56)
- Tokenization, sentence splitting, part-of-speech tagging, lemmatization and syntactic PCFG parsing were performed using Stanford CoreNLP

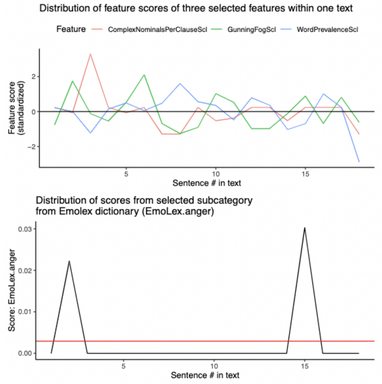

Fig. 2: Text contours for selected features of a single text. Top panel: Distribution of three z-standardized language feature scores from three different feature groups (red:syntactic, green: psycholinguistic, blue: readability). Bottom panel: Text contour of an individual feature of a closed-vocabulary feature (anger words from EmoLex dictionary).

Six benchmark models for the introduced dataset:

1. BERT-BLSTM: a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model
2. BLSTM-MBLF: a bidirectional LSTM neural network classifiers trained on text-average (means-based) language features
3. BLSTM-CBLF: a bidirectional LSTM neural network classifiers trained on sentence-level (contour-based) language features
4. BLSTM-CBLF+BERT: a hybrid model integrating BERT predictions with the language features
5. BLSTM-CBLF+BG: a hybrid model combining language features and sociodemographic features
6. BLSTM-CBLF+BG+BERT: a full model integrating language features and sociodemographic features with BERT predictions

Personality Detection Results

Measure	Model	E	A	C	N	O	Avg.
precision	BLSTM-CBLF	62.79	63.88	57.67	58.63	61.81	60.95
	BLSTM-MBLF	58.74	64.74	52.28	53.12	60.55	57.89
	BERT-FullyConnected	59.62	74.13	60.31	54.32	56.74	61.02
	BERT-BLSTM	47.79	46.01	53.3	29.62	30.2	41.38
	BLSTM-CBLF+BERT	62.32	66.43	57.39	57.89	62.22	61.25
	BLSTM-CBLF+BG	59.48	66.70	60.81	59.66	66.63	62.66
	BLSTM-CBLF+BG+BERT	59.19	67.51	59.68	60.59	65.66	62.53
recall	BLSTM-CBLF	48.02	57.76	46.20	49.67	58.24	51.98
	BLSTM-MBLF	49.67	57.76	48.04	51.79	58.35	53.12
	BERT-FullyConnected	52.14	59.72	49.33	50.04	57.53	53.75
	BERT-BLSTM	58.25	59.31	58.8	49.66	64.44	58.09
	BLSTM-CBLF+BERT	50.16	57.71	55.71	54.40	60.00	55.60
	BLSTM-CBLF+BG	50.66	64.58	55.33	57.23	63.30	58.22
	BLSTM-CBLF+BG+BERT	51.32	64.95	55.98	59.24	63.35	58.97
F1	BLSTM-CBLF	54.42	60.67	51.30	53.78	59.97	56.03
	BLSTM-MBLF	53.83	61.05	50.07	52.45	59.43	55.37
	BERT-FullyConnected	54.64	64.36	52.81	49.38	50.68	54.37
	BERT-BLSTM	49.13	49.65	53.46	32.82	36.76	44.36
	BLSTM-CBLF+BERT	55.59	61.76	56.54	56.09	61.09	58.21
	BLSTM-CBLF+BG	54.72	65.63	57.94	58.42	64.92	60.32
	BLSTM-CBLF+BG+BERT	54.97	66.21	57.77	59.91	64.49	60.67
accuracy	Majority class baseline	49.93	49.85	49.93	49.88	49.69	49.86
	BLSTM-CBLF	61.06	61.76	57.07	58.22	62.37	60.10
	BLSTM-MBLF	58.75	62.37	53.11	54.04	61.44	57.94
	BERT-FullyConnected	52.53	61.42	54.08	51.05	55.84	54.98
	BERT-BLSTM	54.21	59.45	57.84	50.55	58.5	56.11
	BLSTM-CBLF+BERT	**61.20**	63.51	58.09	58.32	63.01	60.82
	BLSTM-CBLF+BG	59.41	65.45	**60.69**	60.13	**66.89**	62.52
	BLSTM-CBLF+BG+BERT	59.31	**66.14**	59.95	**61.20**	66.22	**62.56**

Table 1: Evaluation results of the six benchmark models. Numbers represent classification accuracy (%) micro-averaged across 20 times 10-fold cv. In the 'Avg.' column, the macro-averaged classification accuracy (%) across 5 traits are presented

Ablation Study

Feature ablation studies to assess the informativeness of a feature group using Submodular Pick Lime (SP-LIME).

E		A		C		N		O	
Group	I	Group	I	Group	I	Group	I	Group	I
Sentiment	5.10	Sentiment	5.07	Sentiment	4.89	Sentiment	4.82	Sentiment	5.00
LIWC	3.47	LIWC	3.68	LIWC	3.49	LIWC	3.45	LIWC	3.62
Psycholing	3.36	Psycholing	3.32	Psycholing	3.09	Emotion	3.05	Psycholing	3.19
Readability	3.12	Syntactic	3.16	Ngram	3.07	Ngram	2.95	Ngram	3.10
Emotion	3.04	Emotion	3.14	Emotion	3.04	Psycholing	2.94	Emotion	3.10
Ngram	2.99	Ngram	3.03	Syntactic	2.99	Syntactic	2.90	Syntactic	3.01
Syntactic	2.85	Readability	2.81	Readability	2.96	Readability	2.71	Readability	2.80
Lexical	2.64	Lexical	2.66	Lexical	2.65	Lexical	2.52	Lexical	2.64
InfTheo	1.33	InfTheo	1.48	InfTheo	1.36	InfTheo	1.35	InfTheo	1.40

Table 2: Results of the feature ablation experiment: Feature importance (Model: BLSTM-CBLF) macro-averaged across 200 model instances. (20 $\times$ 10-fold CV)