Results on tf-idf embeddings of ASD behavioral phenotypes from ODFLab

April 2019

Tf-idf (*term frequency - inverse document frequency*) is a numerical statistic that represents how important is a word to a document in a collection.

In this work we represent each subject with ASD as a temporally ordered sequence of behavioral terms that represents their phenotype (i.e., the *document*). The whole dataset is the corpus. In particular, we consider 3 hierarchical levels when taking into account the instruments administered to the subjects. Level 1 selects the raw item scores, Level 2 considers the scores obtained through the composition of the raw item scores and Level 3 takes into account the total scores. Depending on the desired level, we obtain different datasets where each subject is represented by a sequence of terms (with different lengths) that depend on the selected hierarchical level. Each subscale/scale identifier has attached the corresponding score, see Example 0.1.

Example 0.1. Let's consider a subject from the dataset (subj_id), here the 3 level representations are displayed.

- Level 1:

  subj_id – ados-m1::a2::2, ados-m1::a8::1, ados::b1::2, ados-m1::b3::1, ados-m1::b4::1, ados-m1::b5::1, ados-m1::b9::2, ados-m1::b10::2, ados-m1::b11::0, ados-m1::b12::1, ados-m1::a3::0, ados::d1::0, ados::d2::0, ados-m1::d4::1

- Level 2:

  subj_id – ados::sa_tot::13, ados::rrb_tot::1

- Level 3:

  subj_id – ados::sarrb_tot::14

The subscales that different instruments have in common are considered as a general word and the instrument version/module is not specified.

1 Data statistics

Period span: 2010-01-27 – 2019-05-03

Instrument list:

Table 1: Statistics for number of assessments (i.e., number of administered instruments) and number of encounters (i.e., longitudinal assessments at different times)

	Assessments	Encounters
Mean	5.11	1.39
Median	5.0	1.0
Min - Max	$(1, 24)$	$(1, 5)$

- Cognitive assessment: wisc-iv, wppsi-iiifascia2.6-3.11, griffithsmentaldevelopmentscales, wppsi-iiifascia4.0-7.3, leiterinternationalperformancescalerevised, wais-iv, wisc-iii, wais-r, wppsi;
- Diagnostic tools: ados-2modulotoddler, ados-2modulo3, ados-2modulo1, ados-2modulo2, ados-2modulo4;
- Caregiver administered questionnaires: srs, psi-sf, vineland-ii.

Mean age of the subjects: 13.93 (sd=9.10)
N Female: 35 – N Male: 212
Total number of subjects: 247

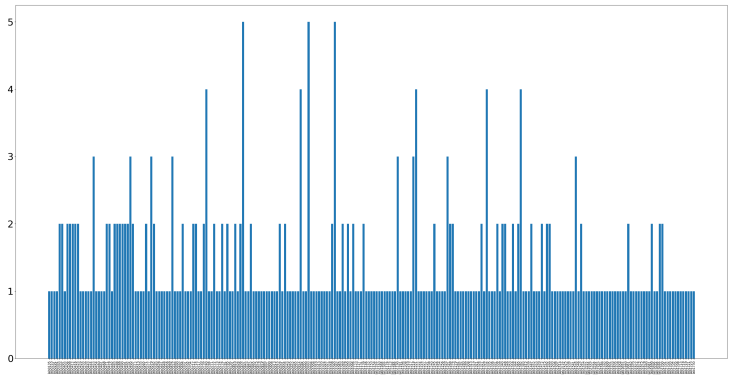

Figure 1: Distribution of the number of encounters for each patient to assess the dataset longitudinality.

2 Level 1 encodings

Vocabulary size: 1820 (repeated terms: 1264, terms with one occurrence: 556).

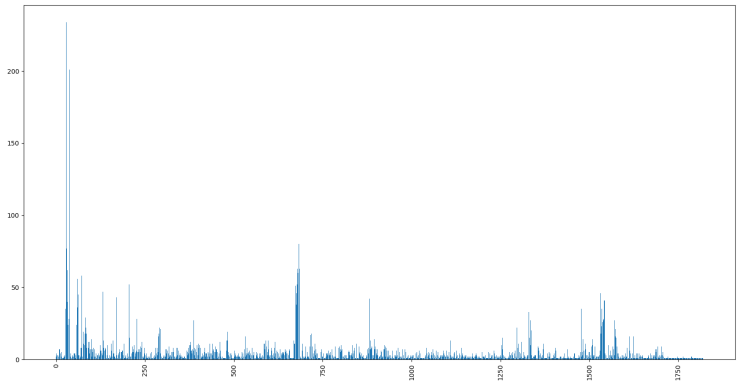

Figure 2: Distribution of the most frequent terms (frequency greater than 20).

Ados subscales **b1, d1, d2** with scores 2, 0, 0, respectively are the most frequent terms. In particular **b1::2** occurs 234 times, **d1::0** occurs 201 times and **d2::0** occurs 182 times throughout the whole dataset.

2.1 Hierarchical clustering of tf-idf subject encodings

The number of clusters is based on the best mean score of the MCC from Random Forest and the Silhouette Score.

Table 2: Patient subtyping

MCC	Silhouette	N clusters
0.98	0.30	4

Table 3: Subcluster numerosities

	N
Cluster 0	76
Cluster 1	54
Cluster 2	70
Cluster 3	47

2.2 Cluster biases inspection

We inspect possible confounders that can drive the clustering among: *sex, number of encounters, list of instrument, age* and we perform statistical tests (t-test

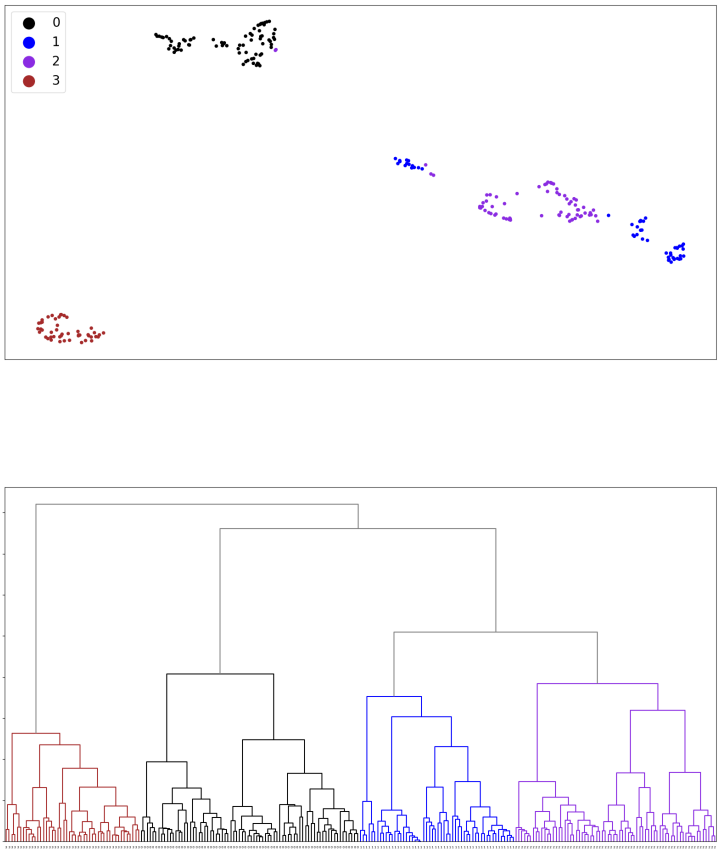

Figure 3: Clusters (Uniform Manifold Approximation and Projection visualization) and dendrogram.

and chi-square).

Running a t-test with Bonferroni correction on the mean ages we obtain that Cluster 0 mean age is significantly different from all the other cluster mean ages ($p < 0.05$) and that Cluster 3 mean age is statistically significant when compared to all the others ($p < 0.001$). Cluster 1 and Cluster 2 mean age difference is not significant.

The average number of encounters in Cluster 1 is significantly greater than the average number of encounters in Cluster 0 and Cluster 3 ($p < 0.01$).

Running a chi-square test for cluster independence from sex counts we obtained that no comparison in sex count was significant, although the number of females is smaller in each cluster than the number of males, reflecting the co-

Table 4: Confounder statistics. Mean score and Max/Min interval are reported for age and encounters. Counts are reported for sex.

	Age	Encounters	Sex
Cluster 0	13.16 (5.06, 22.77)	1.22 (1, 3)	(F=4, M=72)
Cluster 1	9.89 (1.37, 30.76)	1.67 (1, 5)	(F=12, M=42)
Cluster 2	9.24 (3.67, 21.51)	1.50 (1, 5)	(F=15, M=55)
Cluster 3	26.81 (1.08, 61.81)	1.17 (1, 3)	(F=4, M=43)

Table 5: Instruments in each cluster

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
ados-m1	x	x	x	
ados-m2	x	x	x	
ados-m3	x	x	x	x
ados-m4				x
ados-mT		x		
gmds	x	x	x	
leiter	x	x	x	x
psi	x	x	x	x
srs	x	x	x	x
vineland	x	x	x	
wais-iv	x	x		x
wais-r				x
wisc-iii	x	x	x	x
wisc-iv	x	x	x	x
wppsi		x		
wppsi-iii(2.6-3.11)		x	x	
wppsi-iii(4.0-7.3)	x	x	x	

hort numerosity and the male-to-female ratio found in ASD (4:1). See Table 4. The list of instrument whose terms were detected in each cluster is reported in Table 5.

Given the significant difference in age and the list of instrument in Cluster 3, it is straightforward to observe that the adults are mainly clustered together and that the fact that only adults have ados module 4 scores may have biased the patient encodings. Nevertheless, in order to investigate if ASD behavioral profiles can be found in these subclusters we decided to report the average scores on Levels 2 and 3 measures for each subgroup (see `inspection-l1-2.log` and `inspection-l1-3.log` files, respectively for the term list of each cluster with the corresponding average score). Profiles on these levels can also be visually investigated via the heatmaps `HMP-level-2.html` and
`HMP-level-3.html`. Moving the mouse over a box the corresponding mean score is highlighted. Grey boxes are missing values.

Mean scores from the two levels between clusters that include subjects with the same item from the same instrument are compared via t-tests with Bonferroni correction for multiple comparisons (see Tables 6-7).

Table 6: T-test between clusters on Level 2 measures. *ns* stands for non significant result. If empty, the item was not present in one of the clusters to compare. *vineland-MP* is vineland Madre Padre. Test scores not present in the table were not significant or only found in one cluster.

	0-1	0-2	0-3	1-2	1-3	2-3
ados-rrb_tot	< 0.001	< 0.001	ns	ns	ns	0.011
ados-sa_tot	ns	< 0.001	ns	< 0.01	ns	< 0.01
gmds-q_A	ns	ns	-	< 0.001	-	-
gmds-q_B	ns	ns	-	< 0.001	-	-
gmds-q_C	ns	0.044	-	< 0.001	-	-
gmds-q_D	ns	ns	-	< 0.001	-	-
gmds-q_E	ns	ns	-	< 0.001	-	-
leiter-scaled_fg	0.043	ns	ns	< 0.01	ns	ns
srs-Padre-raw_rirb	0.018	ns	ns	ns	ns	
vineland-MP-sum_CD	ns	< 0.001	-	ns	-	-
vineland-MP-sum_DLSD	ns	0.018	-	ns	-	-
vineland-MP-sum_SD	ns	ns	-	< 0.01	-	-
wechsler-P	0.03	ns	ns	ns	< 0.001	0.011
wechsler-V	ns	ns	< 0.02	ns	< 0.001	< 0.01
wechsler-VC	ns	ns	ns	ns	ns	0.033

Table 7: T-test between clusters on Level 3 measures. *ns* stands for non significant result. If empty, the item was not present in one of the clusters to compare. *vineland-MP* is vineland Madre Padre. Test scores not present in the table were not significant or only found in one cluster.

	0-1	0-2	0-3	1-2	1-3	2-3
ados-sarrb_tot	0.04	< 0.001	ns	< 0.01	ns	< 0.001
gmds-GQ	ns	ns	-	< 0.001	-	-
vineland-MP-standard_ABC	ns	0.032	-	0.012	-	-
wechsler-FS	ns	ns	ns	ns	< 0.001	< 0.01