Table 3. Classification accuracies of 4900 arXiv **doc**uments (above), 3500 **sec**tions (middle), and 1400 **abs**tracts (below) into 14 subject classes using different classifier (columns), and text or math encodings (rows). The highest mean/maximum is highlighted in yellow/red. It is orange if an encoding or classifier yields both the highest mean and maximum value. The shortest runtime is marked in green.

Encoding/Classifier	LogReg	LinSVC	RbfSVC	kNN	MLP	DecTree	RandForest	GradBoost	Mean	Max
doc2vecText	64.5	60.3	75.2	18.6	72.8	31.7	45.6	64.6	54.2	75.2
docText_tfidf	80.6	82.8	72.8	78.1	82.6	57.0	51.8	75.8	72.7	82.8
doc2vecMath_op	37.7	37.4	38.1	22.8	31.8	16.1	21.3	33.8	29.9	38.1
docMath_op_tfidf	14.9	15.0	13.7	10.1	14.7	14.4	14.6	14.7	14.0	15.0
doc2vecMath_id	42.2	36.0	33.8	22.5	43.3	14.8	17.8	36.8	30.9	43.3
docMath_id_tfidf	23.0	22.8	16.4	15.2	23.0	18.0	20.9	22.1	20.2	23.0
doc2vecMath_opid	45.7	39.7	24.8	17.7	46.7	14.2	17.9	39.6	30.8	46.7
docMath_opid_tfidf	25.5	25.8	17.1	16.7	25.2	17.3	21.4	24.7	21.7	25.8
doc2vecMath_semantics	49.5	46.5	24.8	14.4	51.3	12.0	14.8	40.6	31.7	51.3
docMath_semantics_tfidf	63.2	63.4	43.7	7.2	64.7	37.1	44.0	56.4	47.5	64.7
doc2vecTextMath_opid	64.2	59.3	74.7	10.4	71.7	31.8	41.5	63.3	52.1	74.7
doc2vecTextMath_semantics	61.7	57.4	61.4	23.7	70.6	31.6	41.4	64.0	51.5	70.6
Mean	47.7	45.5	41.4	21.5	49.9	24.7	29.4	44.7	38.1	49.9
Max	80.6	82.8	75.2	78.1	82.6	57.0	51.8	75.8	73.0	82.8
Runtime [%]	1.1	1.6	4.6	0.1	100.0	0.4	0.1	46.3	19.3	100.0

Encoding/Classifier	LogReg	LinSVC	RbfSVC	kNN	MLP	DecTree	RandForest	GradBoost	Mean	Max
sec2vecText	51.7	51.9	51.8	58.0	61.9	51.8	33.5	51.7	51.5	61.9
secText_tfidf	68.9	69.0	69.4	70.1	77.5	69.2	49.3	69.1	67.8	77.5
sec2vecMath_op	29.9	30.2	29.9	21.2	40.2	30.2	18.5	29.9	28.8	40.2
secMath_op_tfidf	14.8	14.6	14.5	11.2	16.7	14.7	13.7	14.8	14.4	16.7
sec2vecMath_id	27.6	28.0	27.9	14.0	40.7	28.2	13.8	28.0	26.0	40.7
secMath_id_tfidf	23.8	23.9	23.7	16.1	24.1	23.9	20.3	23.9	22.5	24.1
sec2vecMath_opid	30.1	30.2	30.0	10.6	42.3	29.9	16.5	30.3	27.5	42.3
secMath_opid_tfidf	27.0	27.1	26.1	17.3	25.9	26.6	22.6	26.5	24.9	27.1
sec2vecMath_semantics	32.7	33.3	32.3	10.3	47.5	32.5	12.3	32.3	29.2	47.5
secMath_semantics_tfidf	54.6	54.5	55.0	8.1	61.0	55.3	41.9	54.7	48.1	61.0
sec2vecTextMath_opid	49.9	50.4	50.2	52.0	63.0	50.3	31.0	50.5	49.7	63.0
sec2vecTextMath_semantics	50.8	50.7	50.7	23.7	60.5	50.8	31.8	50.8	46.2	60.5
Mean	38.5	38.7	38.5	26.1	46.8	38.6	25.4	38.5	36.4	46.8
Max	68.9	69.0	69.4	70.1	77.5	69.2	49.3	69.1	67.8	77.5
Runtime [%]	100.0	100.0	95.6	0.2	78.5	100.0	0.2	100.0	71.8	100.0

Encoding/Classifier	LogReg	LinSVC	RbfSVC	kNN	MLP	DecTree	RandForest	GradBoost	Mean	Max
abs2vecText	42.6	38.5	50.2	25.6	47.1	17.1	21.4	38.1	35.1	50.2
absText_tfidf	58.9	61.1	49.1	50.9	61.6	33.1	37.4	46.4	49.8	61.6
abs2vecMath_opid	26.1	26.5	22.1	16.6	23.8	13.8	17.0	22.0	21.0	26.5
absMath_opid_tfidf	10.6	10.7	9.9	8.4	10.4	10.5	10.2	10.5	10.2	10.7
Mean	34.6	34.2	32.8	25.4	35.7	18.6	21.5	29.3	29.0	35.7
Max	58.9	61.1	50.2	50.9	61.6	33.1	37.4	46.4	50.0	61.6
Runtime [%]	1.8	8.7	3.5	0.2	100.0	1.6	0.4	55.0	21.4	100.0

Table 4. Clustering purities of 4900 arXiv **doc**uments (above), 3500 **sec**tions (middle), and 1400 **abs**tracts (below) with 14 subject classes using different clusterers (columns), and text or math encodings (rows). The highest mean/maximum is highlighted in yellow/red for the group of clusterers with specified cluster number (KMeans, Agglomerative, GaussianMixture) and unspecified (Affinity, MeanShift, HDBSCAN) respectively. It is orange if an encoding or clusterer yields both the highest mean and maximum. The shortest runtime is marked in green.

Encoding/Clusterer	KMeans	Affinity	Agglomerative	MeanShift	GaussianMixture	HDBSCAN	Mean	Max
doc2vecText	57.5	73.4	57.9	7.1	56.5	77.0	54.9	85.6
docText_tfidf	63.2	83.5	64.9	83.5	55.3	87.8	75.3	89.5
doc2vecMath_op	23.3	78.2	20.9	94.6	21.0	45.0	47.2	94.6
docMath_op_tfidf	34.5	95.5	35.8	82.9	45.5	46.1	56.7	82.9
doc2vecMath_id	41.5	81.6	29.0	99.8	68.7	43.9	60.8	99.8
docMath_id_tfidf	24.0	67.1	21.1	92.7	19.4	37.6	43.7	92.7
doc2vecMath_opid	51.0	46.5	36.4	97.9	42.0	49.1	53.8	97.9
docMath_opid_tfidf	18.0	76.9	18.7	7.1	25.8	34.3	30.1	57.8
doc2vecMath_semantics	62.7	96.5	33.9	99.9	69.4	7.1	61.6	99.9
docMath_semantics_tfidf	36.0	21.8	44.6	93.6	93.4	33.1	53.8	44.6
doc2vecTextMath_opid	55.9	67.5	58.5	7.1	52.5	44.1	47.6	67.5
doc2vecTextMath_semantics	59.0	94.4	52.8	99.4	61.0	43.9	68.4	99.4
Mean	43.9	73.6	39.5	72.1	50.9	45.8	54.3	68.9
Max	63.2	96.5	64.9	99.9	69.4	89.5	80.6	99.9
Runtime [%]	5.9	3.7	7.7	100.0	2.0	0.6	20.0	100.0

Encoding/Clusterer	KMeans	Affinity	Agglomerative	MeanShift	GaussianMixture	HDBSCAN	Mean	Max
sec2vecText	43.0	62.1	41.3	7.1	43.5	60.5	42.9	62.1
secText_tfidf	53.2	79.6	57.8	7.1	56.4	27.3	46.9	79.6
sec2vecMath_op	24.6	58.2	24.3	97.3	22.5	7.1	39.0	97.3
secMath_op_tfidf	19.6	94.0	21.3	82.9	26.2	27.3	45.2	94.0
sec2vecMath_id	27.7	42.3	28.7	53.6	81.7	7.1	40.1	81.7
secMath_id_tfidf	25.7	69.2	24.1	7.1	20.3	34.8	30.2	69.2
sec2vecMath_opid	33.2	41.9	32.5	53.6	57.5	7.1	37.6	57.5
secMath_opid_tfidf	18.6	54.3	17.4	7.1	26.8	34.7	26.5	54.3
sec2vecMath_semantics	51.7	61.6	30.6	98.2	61.2	7.1	51.7	98.2
secMath_semantics_tfidf	46.5	21.8	46.2	7.1	44.2	36.8	33.8	46.5
sec2vecTextMath_opid	43.4	68.7	41.8	7.1	41.4	40.9	40.6	68.7
sec2vecTextMath_semantics	47.0	62.8	40.3	86.7	65.0	7.1	51.5	86.7
Mean	36.2	59.7	33.9	42.9	45.6	24.8	40.5	59.7
Max	53.2	94.0	57.8	98.2	81.7	60.5	74.2	98.2
Runtime [%]	8.2	3.0	6.2	100.0	6.8	32.5	26.1	100.0

Encoding /Clusterer	KMeans	Affinity	Agglomerative	MeanShift	GaussianMixture	HDBSCAN	Mean	Max
abs2vecText	37.5	75.0	31.8	98.3	40.6	7.1	48.4	98.3
absText_tfidf	32.1	58.9	53.6	7.1	25,1	70.2	41,2	70.2
abs2vecMath_opid	35.8	90.3	23.0	98.9	63.8	35.4	57.9	98.9
absMath_opid_tfidf	35.8	81.5	35.7	90.3	42.2	34.9	53.4	90.3
Mean	35.3	76.4	36.0	73.7	42,9	36.9	50,2	76.4
Max	37.5	90.3	53.6	98.9	63.8	70.2	69.1	98.9
Runtime [%]	3.0	2.0	1.1	100.0	3.6	0.4	18.3	100.0