Recovering Self-selected YouTube Video Categories

CAPP 30255: Advanced Machine Learning for Public Policy

Ta-Yun Yang & Patrick Lavallee Delgado*

8 June 2020

1 YouTube data

Videos by category

Category	All videos		With captions	
Entertainment	1619	25%	1149	28%
How-to & Style	595	9%	515	13%
Comedy	547	9%	444	11%
People & Blogs	498	8%	354	9%
News & Politics	505	8%	302	7%
Science & Technology	380	6%	300	7%
Music	799	13%	235	6%
Education	250	4%	228	6%
Film & Animation	318	5%	212	5%
Sports	451	7%	165	4%
Pets & Animals	138	2%	67	2%
Gaming	103	2%	51	1%
Autos & Vehicles	70	1%	37	1%
Travel & Events	60	1%	34	1%
Nonprofits & Activism	14	0%	9	0%
Shows	4	0%	4	0%
Sum	6351	100%	4106	100%

*Candidates, MS Computational Analysis and Public Policy, {tayuny, pld}@uchicago.edu.

2 Models

Results from average word embedding models

	WITH CAPTIONS			WITHOUT CAPTIONS		
News & Politics vs:	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Entertainment	0.77	—	0.00	0.70	—	0.13
How-to & Style	0.65	—	0.01	0.55	0.51	0.61
Comedy	0.55	—	0.00	0.53	0.52	0.57
People & Blogs	0.57	0.83	0.52	0.62	0.65	0.65
Science & Technology	0.50	0.40	0.28	0.54	0.55	0.95
Music	0.63	0.62	1.00	0.45	0.40	0.43
Education	0.59	0.59	1.00	0.72	0.72	1.00
Film & Animation	0.60	0.60	1.00	0.57	0.58	0.99
Sports	0.70	0.69	1.00	0.52	0.66	0.28
Corpus	0.92	—	0.00	0.87	—	0.02

Results from CNN models

	WITH CAPTIONS			WITHOUT CAPTIONS		
News & Politics vs:	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Entertainment	0.77	—	0.00	0.75	—	0.00
How-to & Style	0.65	—	0.00	0.47	0.47	0.98
Comedy	0.55	—	0.00	0.49	0.51	0.27
People & Blogs	0.49	—	0.00	0.51	0.56	0.23
Science & Technology	0.48	0.48	0.00	0.54	0.54	0.99
Music	0.63	0.62	1.00	0.57	0.50	0.48
Education	0.59	0.59	1.00	0.39	0.62	0.29
Film & Animation	0.60	0.60	1.00	0.59	0.58	1.00
Sports	0.39	0.71	0.21	0.49	0.49	1.00
Corpus	0.92	—	0.00	0.91	—	0.00

Results from LSTM models

	WITH CAPTIONS			WITHOUT CAPTIONS		
News & Politics vs:	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Entertainment	0.68	0.27	0.20	0.74	—	0.08
How-to & Style	0.40	0.32	0.58	0.45	0.45	0.83
Comedy	0.61	0.54	0.73	0.46	0.47	0.70
People & Blogs	0.59	0.70	0.39	0.55	0.56	0.84
Science & Technology	0.44	0.45	0.79	0.56	0.58	0.69
Music	0.56	0.63	0.67	0.58	0.54	0.39
Education	0.49	0.56	0.56	0.66	0.71	0.89
Film & Animation	0.47	0.55	0.54	0.48	0.56	0.51
Sports	0.63	0.69	0.82	0.48	0.44	0.24
Corpus	0.92	—	0.00	0.91	—	0.02

Results from GRU models

	WITH CAPTIONS			WITHOUT CAPTIONS		
News & Politics vs:	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Entertainment	0.63	0.28	0.37	0.72	0.33	0.11
How-to & Style	0.48	0.36	0.59	0.49	0.46	0.37
Comedy	0.50	0.43	0.35	0.49	0.55	0.20
People & Blogs	0.54	0.53	0.83	0.58	0.59	0.79
Science & Technology	0.59	0.63	0.35	0.46	0.51	0.40
Music	0.61	0.68	0.69	0.49	0.48	0.32
Education	0.53	0.65	0.41	0.64	0.71	0.84
Film & Animation	0.59	0.67	0.63	0.54	0.58	0.81
Sports	0.63	0.69	0.84	0.54	0.51	0.72
Corpus	0.92	—	0.00	0.92	—	0.00