

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	With captions		Without captions	
Entertainment	1149	28%	1619	25%
How-to & Style	515	13%	595	9%
Comedy	444	11%	547	9%
People & Blogs	354	9%	498	8%
News & Politics	302	7%	505	8%
Science & Technology	300	7%	380	6%
Music	235	6%	799	13%
Education	228	6%	250	4%
Film & Animation	212	5%	318	5%
Sports	165	4%	451	7%
Pets & Animals	67	2%	138	2%
Gaming	51	1%	103	2%
Autos & Vehicles	37	1%	70	1%
Travel & Events	34	1%	60	1%
Nonprofits & Activism	9	0%	14	0%
Shows	4	0%	4	0%
Sum	6351	100%	4106	100%

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

Vocabulary by category

News & Politics vs:	WITH CAPTIONS		WITHOUT CAPTIONS	
	Size	Shared	Size	Shared
Self	22650	100%	8278	100%
Entertainment	25002	50%	20509	55%
How-to & Style	25002	44%	12119	37%
Comedy	20800	41%	8517	35%
People & Blogs	23392	45%	10187	38%
Science & Technology	21897	43%	8579	35%
Music	12385	28%	11270	34%
Education	20491	42%	7853	31%
Film & Animation	18525	40%	8056	32%
Sports	12707	31%	5951	27%
Pets & Animals	6781	19%	3512	19%
Gaming	8685	24%	2528	14%
Autos & Vehicles	5363	16%	2246	13%
Travel & Events	5786	15%	2230	13%
Nonprofits & Activism	3837	13%	908	6%
Shows	950	3%	221	2%
Corpus	25002	53%	25002	62%

Note: vocabulary capped at 25,000 most frequent words, excluding stop words, plus tokens for unknown words and padding.

```
{
  video_id: 1ZAPwftrtAFY,

  title: 'The Trump Presidency: Last Week Tonight with John Oliver (HBO)',

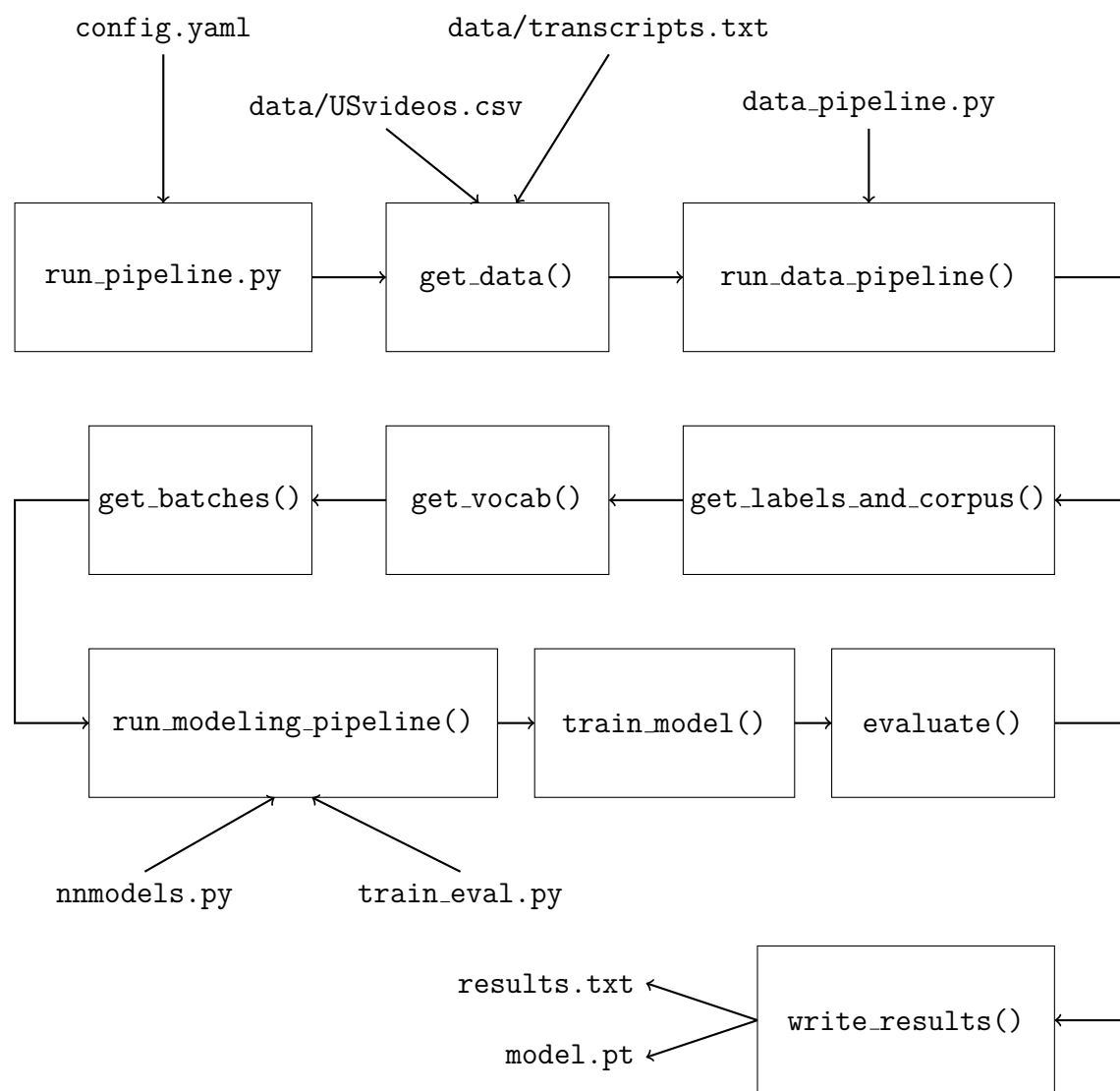
  category_id: 24, // Entertainment

  tags: 'last week tonight trump presidency|"last week tonight donald trump"
        "john oliver trump"|"donald trump"',

  description: "One year after the presidential election, John Oliver
discusses what we've learned so far and enlists our catheter cowboy to
teach Donald Trump what he hasn't.\n\nConnect with Last Week Tonight...",

  caption: 'The presidency of Donald Trump. The man voted "Least Edible" by
Cannibal Magazine -six years in a row. -(AUDIENCE LAUGHING) -A-- And I know,
I honestly know that the prospect of talking about Trump yet again feels...'
}
```

2 Pipeline



```

data:
  col_labels: 'category_id'
  col_corpus: ['title', 'tags', 'description', 'caption']
  label_target: 'News & Politics'
  label_others: ['Entertainment']
  splits: [0.4, 0.4, 0.2]
  ngram_size: 1
  vocab_size: 25000
  batch_size: 64
models:
  - model: CNN
    embedding_dim: 100
    n_filters: 200
    filter_sizes: [3, 4, 5]
    output_dim: 1
    dropout: 0.5
  - model: LSTM
    embedding_dim: 100
    hidden_dim: 50
    output_dim: 1
    n_layers: 4
    bidirectional: True
    dropout: 0.5
decision_metric: 'loss'
out_directory: 'politics_entertainment/'

```

3 Models

Results from average word embedding models

News & Politics vs:	WITH CAPTIONS			WITHOUT CAPTIONS		
	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

News & Politics vs:	WITH CAPTIONS			WITHOUT CAPTIONS		
	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

News & Politics vs:	WITH CAPTIONS			WITHOUT CAPTIONS		
	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

News & Politics vs:	WITH CAPTIONS			WITHOUT CAPTIONS		
	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