Recovering Self-selected YouTube Video Categories

Ta-Yun Yang & Patrick Lavallee Delgado

Objectives

- Use the language with which users describes their content to identify the category selected.
- Distinguish news and politics from other YouTube content.

YouTube data

- Most popular videos in a week.
- Category, title, description, tags for each; captions for most.

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%

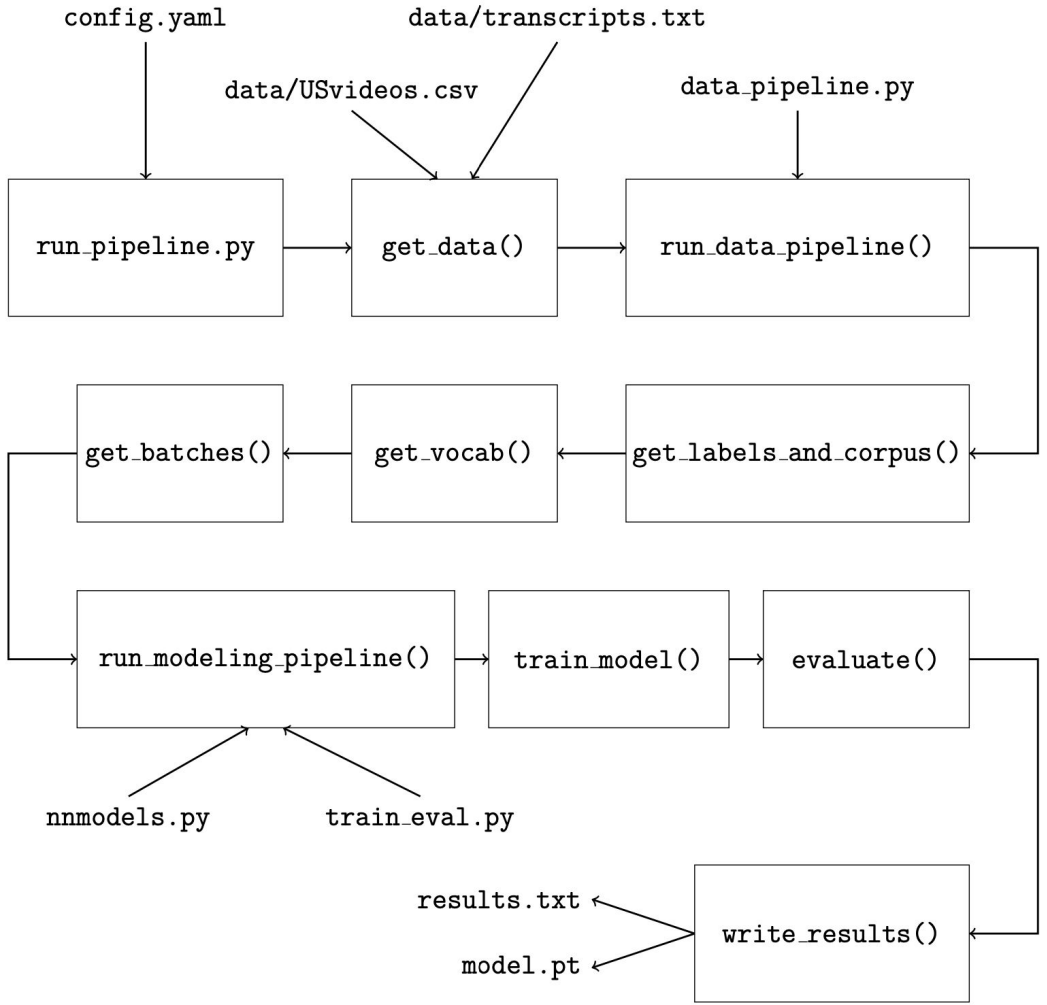
```
{
  video_id: 1ZAPwfrtAFY,

  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...'
}
```

Vocabulary by category

	WITH CAPTIONS		WITHOUT CAPTIONS	
News & Politics vs:	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.

```
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/'
```



Models

- Baseline with random forests and logistic regressions.
- Semantic relationships with word embeddings.
- Predictive word and phrase extraction with CNNs.
- Long-range dependencies with bidirectional LSTMs & GRUs.

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

A closer look

- Counterintuitive results with caption data.
- TF-IDF and cosine similarity between categories.
- Problem for unbalanced data
- Hidden information in positions and orders

Cosine Similarities

- **Method Description**
  - Bag of Words and TF-IDF matrix (# documents) * (# unique words)
  - Average TF-IDF values of documents within each category
  - Calculate Cosine Similarities of representing vectors of each category

Cosine Similarities (Using Captions)	
Film & Animation	0.923
Autos & Vehicles	0.869
Music	0.805
Pets & Animals	0.853
Sports	0.891
Travel & Events	0.839
Gaming	0.893
People & Blogs	0.912
Entertainment	0.905
Howto & Style	0.868
Education	0.942
Science & Technology	0.915
Nonprofits & Activism	0.907
Shows	0.684

Cosine Similarities (Using Tags /Titles/Descriptions)	
Film & Animation	0.512
Autos & Vehicles	0.462
Music	0.426
Pets & Animals	0.393
Sports	0.393
Travel & Events	0.467
Gaming	0.368
People & Blogs	0.55
Entertainment	0.622
Howto & Style	0.467
Education	0.512
Science & Technology	0.548
Nonprofits & Activism	0.52
Shows	0.19

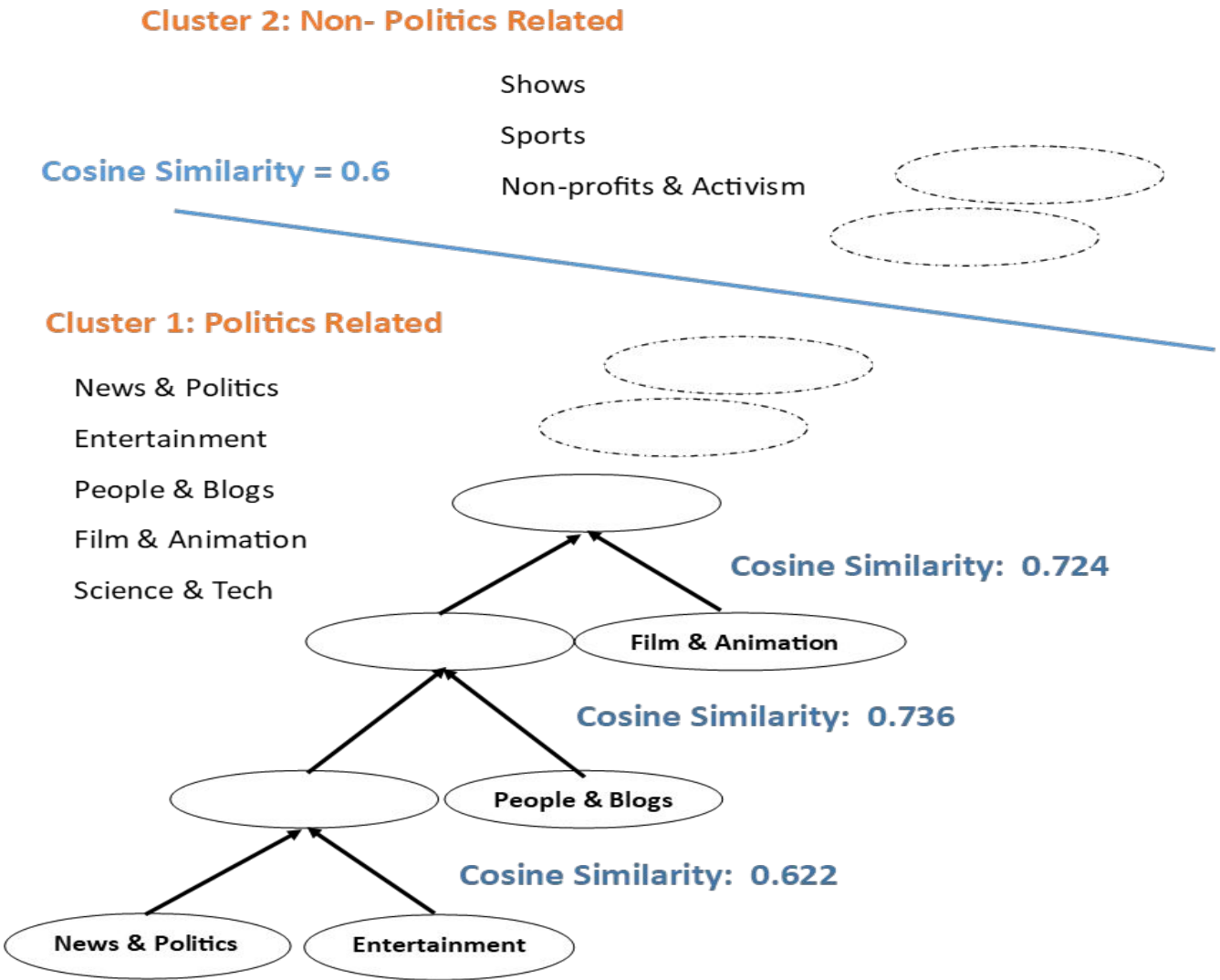
Cosine Similarities

- **Method Description**
  - Bag of Words and TF-IDF matrix (# documents) * (# unique words)
  - Average TF-IDF values of documents within each category
  - Calculate Cosine Similarities of representing vectors of each category
- **Bottom-Up Clustering**
  - Calculate cosine similarities for each category v.s. politics
  - Merge politics with the category that contains the highest similarity
  - Calculate cosine similarities for each category v.s. new created category
  - Operation continues until all categories are merge into one

Bottom-Up Clustering

Start Node:
News & Politics

Selected Data:
Titles /
Tags /
Descriptions



Top 100 Words with Highest TFIDF

- Use Titles / Tags / Descriptions
- Merge all documents into one representing document for each category
- Avoid IDF weighting within each category
- New TFIDF matrix has size (# categories) * (# unique words)
- Calculate cosine similarities between representing categories and politics
- Derive the 100 words with highest score of each category

New & Politics

'north', 'larry', 'fbi', 'iraq', 'kurdistan', 'press', 'online', 'kilauea', 'harry', 'catalog', 'reports', 'house', 'stories', 'wedding', 'local', 'ap', 'access', 'nthe', 'government', 'entertainment', 'day', 'check', 'happening', 'jones', 'videos', 'nwatch', 'nytvideo', 'cnb', 'cx', 'euronews', 'science', 'markle', 'bbc', 'washington', 'year', 'coverage', 'nightly', 'events', 'nget', 'people', 'kim', 'technology', 'guardian', 'list', 'eruption', 'york', 'olympics', '0bsajo', 'shooting', 'senate', 'meghan', 'korea', 'izonye', 'u2g06o', 'nvox', 'school', '2018', 'prince', 'playlist', 'royal', 'washingtonpost', 'cnn', 'google', 'xfrz5h', 'headlines', 'reporting', 'times', 'iran', 'donald', 'hoda', 'volcano', 'health', 'amtrak', 'breaking', 'sexual', 'channel', 'business', 'live', 'nassar', 'nhttps', 'ws', 'politics', 'president', 'cnbc', 'cbc', 'latest', 'world', 'goo', 'gl', 'morning', 'cbsn', 'time', 'trump', 'nbcnews', 'video', 'msnbc', 'nbc', 'today', 'vox', 'cbs'

Entertainment

'collider', 'nnbc', 'amzn', '10', 'nabout', 'complex', 'st', 'makeup', 'anytime', 'jennifer', 'bravotv', 'interview', 'smith', 'kids', 'time', 'marvel', 'rhettandlink', 'like', 'po', 'nthe', 'rhett', '2017', 'idol', 'website', 'film', 'hellthyjunkfood', 'box', 'fine', 'john', 'best', 'shows', 'comedian', 'nail', 'vanity', 'love', 'challenge', 'kardashian', 'online', 'stephen', 'degeneres', 'day', 'nfacebook', 'nbcsnl', 'gmm', 'jedi', 'ninstagram', 'original', 'ntwitter', 'celebrities', 'black', 'television', 'hollywood', 'episode', 'movies', 'talk', 'wwhl', 'series', 'youtu', 'james', 'wars', 'episodes', 'world', 'official', 'nlike', 'nwatch', 'snl', '2018', 'tumblr', 'react', 'mythical', 'celebrity', 'entertainment', 'trailer', 'nbc', 'comedy', 'season', 'channel', 'movie', 'star', 'gl', 'goo', 'corden', 'voice', 'tmz', 'colbert', 'night', 'bravo', 'fbe', 'nhttp', 'funny', 'video', 'videos', 'jimmy', 'cbs', 'ellen', 'nhttps', 'live', 'netflix', 'late', 'kimmel'

Final Improvements

- Solve the low precision in unbalanced data (politics v.s. non-politics)
- Run the models without Captions
- Limit the range of stop words, include phrases in url

Training on Balanced v.s. Unbalanced Data

Politics v.s. Non-Politics

Model	Accuracy	Precision	Recall
Linear NN	97.56%	NaN	69.61%
Simple RNN	92.24%	NaN	0%
BLSTM	92.94%	NaN	15.42%
BGRU	94.27%	NaN	35.97%
CNN	96.72%	NaN	0%
TFIDF ML	96.67%	92.5%	41.57%
BOW ngrams	97.01%	82.60%	77.55%

Unbalanced

7% vs 93%

Entertainment v.s. Non-Entertainment

Model	Accuracy	Precision	Recall
Linear NN	86.91%	73.77%	68.97%
Simple RNN	75.03%	NaN	4.17%
BLSTM	79.36%	71.24%	27.99%
BGRU	79.42%	61.36%	45.94%
CNN	86.01%	72.28%	67.94%
TFIDF ML	84.89%	90.56%	44.8%
BOW ngrams	86.62%	75%	70.31%

Balanced

24% vs 76%

Run Neural Network Models on Data with Binary Categories

- Selected Features: Titles / Tags / Descriptions
- Binary Categories
  - Politics v.s. Non-Politics (7.36% : 92.64%)
  - Politics v.s. Entertainment (23.79% : 76.21%)
  - Politics v.s People & Blogs (50.4% : 49.6%)
  - Politics v.s. Science & Technology (57.1% : 42.9%)
  - Politics v.s. Film & Animation (61.23% : 38.77%)

Politics v.s. Non-Politics

Model	Accuracy	Precision	Recall
Linear NN	97.56%	NaN	69.61%
Simple RNN	92.24%	NaN	0%
BLSTM	92.94%	NaN	15.42%
BGRU	94.27%	NaN	35.97%
CNN	96.72%	NaN	0%
TFIDF ML	96.67%	92.5%	41.57%
BOW ngrams	97.01%	82.60%	77.55%

Using Titles / Tags / Descriptions Data
Bad Precision in NN models (politics v.s. non-politics)
Balanced Data (binary categories)
Unbalanced Data (politics v.s. non-politics)

Politics v.s. Entertainment

Model	Accuracy	Precision	Recall
Linear NN	88.82%	69.81%	96.20%
Simple RNN	75.97%	NaN	NaN
BLSTM	86.26%	76.85%	62.88%
BGRU	86.48%	87.62%	48.38%
CNN	93.86%	87.44%	85.65%

Politics v.s. People & Blogs

Model	Accuracy	Precision	Recall
Linear NN	88.98%	85.96%	92.57%
Simple RNN	44.36%	40.51%	37.30%
BLSTM	76.91%	78.96%	74.62%
BGRU	73.35%	80.44%	63.18%
CNN	81.21%	95.45%	61.13%

If we did not drop URL in the descriptions?

Politics v.s. Entertainment

Model	Accuracy	Precision	Recall
Linear NN	75.0%	NaN	13.0%
BLSTM	74.0%	NaN	8.0%
BGRU	72.0%	33.0%	11.0%
CNN	75.0%	NaN	0%

Model	Accuracy	Precision	Recall
Linear NN	88.82%	69.81%	96.20%
Simple RNN	75.97%	NaN	NaN
BLSTM	86.26%	76.85%	62.88%
BGRU	86.48%	87.62%	48.38%
CNN	93.86%	87.44%	85.65%

Politics v.s. People & Blogs

Model	Accuracy	Precision	Recall
Linear NN	62.0%	65.0%	65.0%
BLSTM	56.0%	58.0%	69.0%
BGRU	58.0%	59.0%	79.0%
CNN	54.0%	54.0%	99.0%

Model	Accuracy	Precision	Recall
Linear NN	88.98%	85.96%	92.57%
Simple RNN	44.36%	40.51%	37.30%
BLSTM	76.91%	78.96%	74.62%
BGRU	73.35%	80.44%	63.18%
CNN	81.21%	95.45%	61.13%

Excluding URL

Including URL

Conclusions

- With Captions / Without Captions
- Cosine Similarities
- Balanced / Unbalanced
- Position and Order matters
- URL contains some information to classify similar categories

Conclusion

- **Caption data are mostly noise**
  - High cosine similarities between categories
  - Not significantly improve model performance when captions are included
  - Critical sentences and phrases are not frequent enough in the scripts
  - Entertainment-oriented nature of YouTube video
- **News and politics not obviously distinguishable from others**
  - High cosine similarities in bottom-up clustering (cosine similarities > 0.6)
  - Similar to entertainment, blogs, and sciences

Conclusion

- **Focus on using Titles / Tags / Descriptions**
- **NN models and ML models successfully classify politics category**
  - Traditional models work well even with unbalanced data
  - NN models work better when the data is more balanced
- **NN models work well in Politics v.s Entertainment and Politics v.s. Science & Technology, when we include URL**
  - The Positions and the orders of the words matter
  - Linnear NN, CNN, ML model performs the best (critical words are effective)
  - LSTM, GRU also provide valid results (rely on the memory of the sequence)

Discussions and Practical Notes

- Long runtime and huge amount of required memory for captions
  - Takes 1 - 2 hours for a single RNN round on the subset data
  - Requires at least 15 GB of memory
- Adjustment of batch size
  - Memory requirement increases dramatically when more layers are added to RNN and CNN
  - Change batch size from 64 to 32 reduces the information stored in each neuron
- Titles / Tags / Descriptions are sufficiently informative
- Although the order of the sequence contains some information, critical words are effective. Linear NN, CNN and ML are performing better.