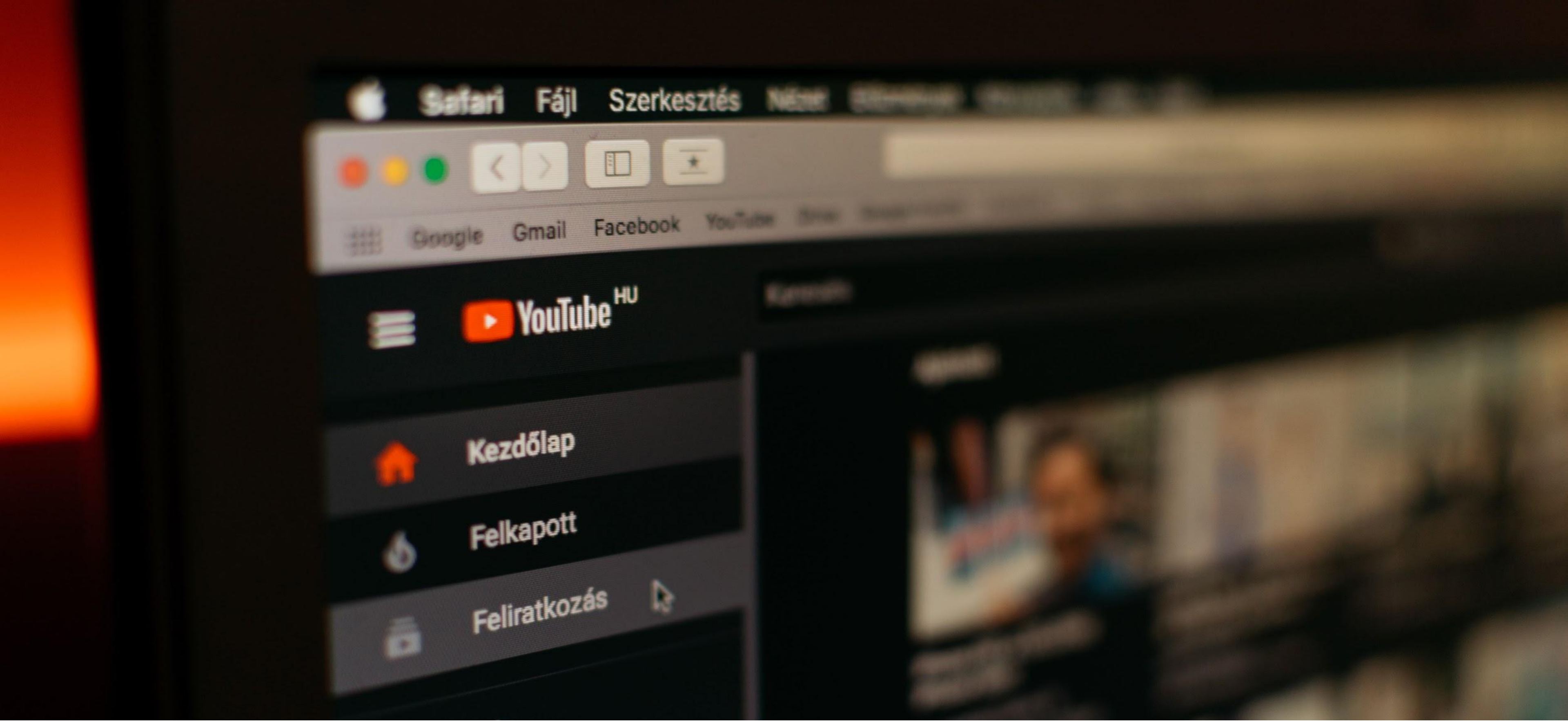


Topic Modeling and Text Classification

Scoping the challenge 1

YouTube

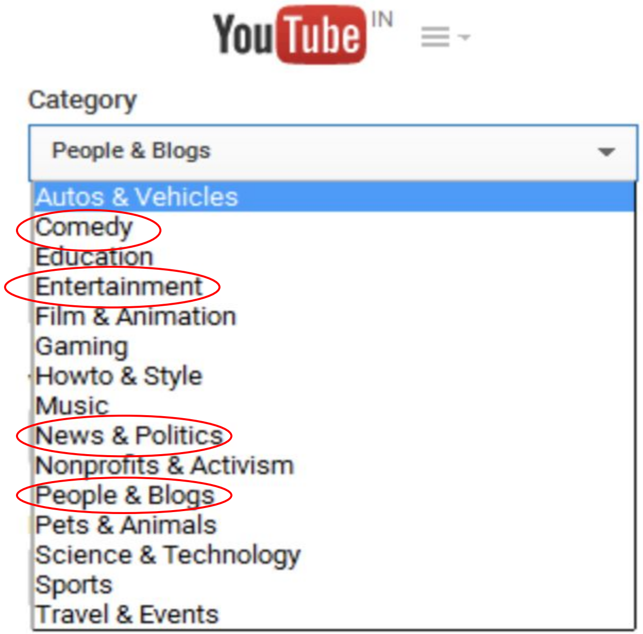
- The first step to building an **engaged** audience is using the **right keywords** to describe a video

- However, the choice of keywords used in the title is often based on trial and error, what is popular and trending rather than on what is most **relevant**

For content creators, *engagement* rather than *views* drives monetisation

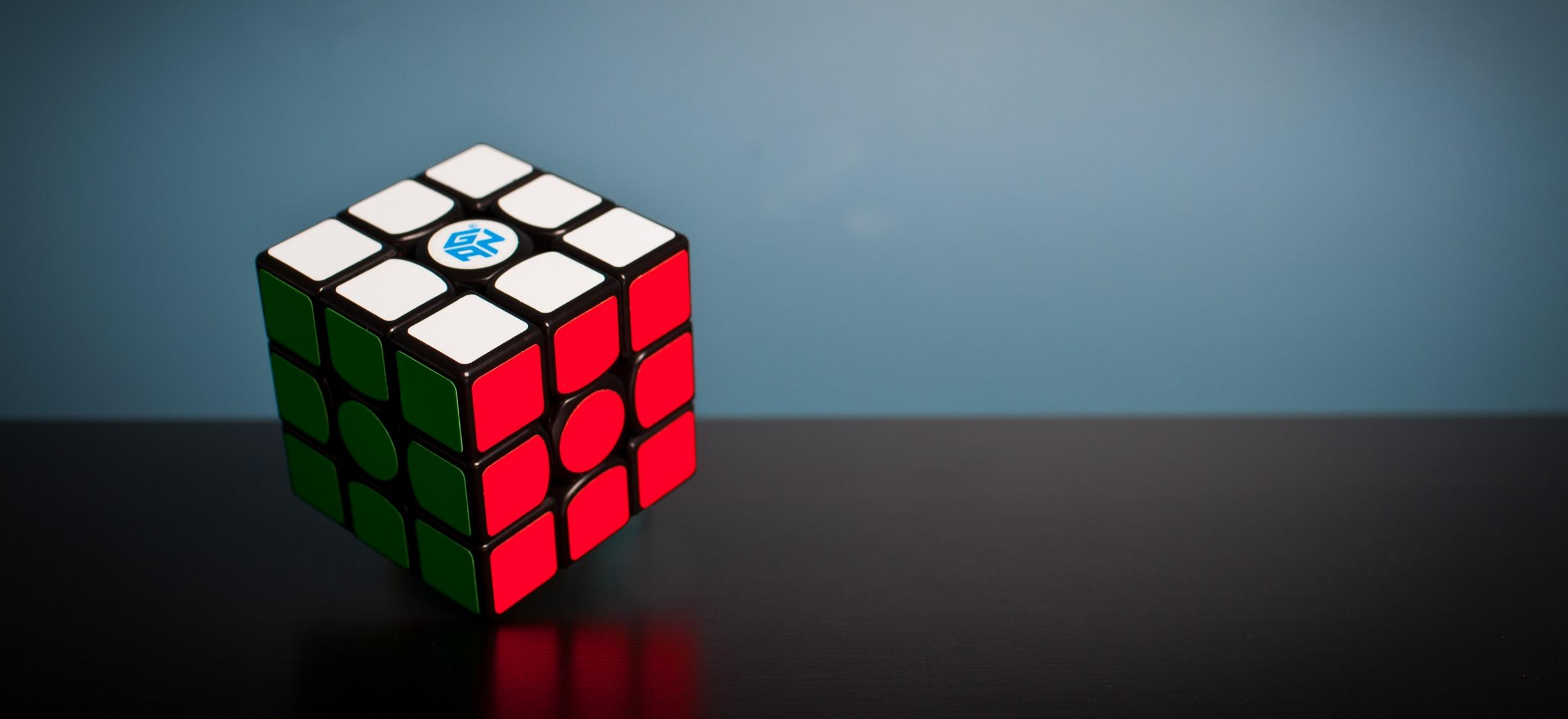
Scoping the challenge 2

- Each video can only be uploaded into one category
- With over 30 categories to choose from, a wrong classification of title with category can lead to videos not appearing during search



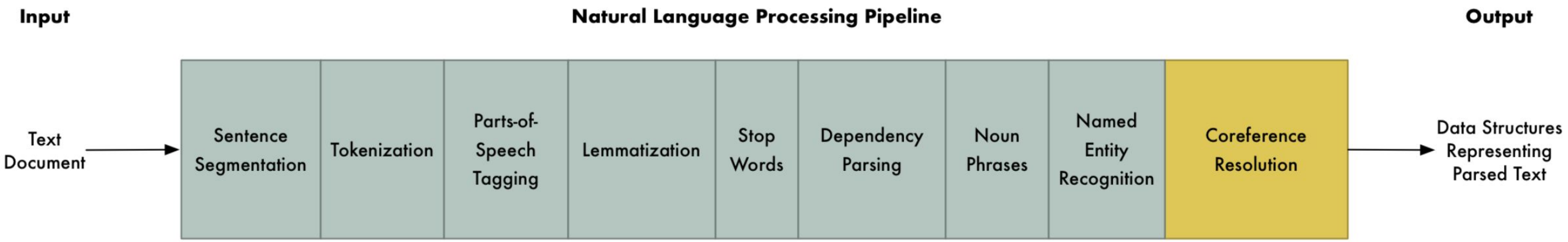
Which of these categories would Trump's latest gaffe fall into?

Problem-solving

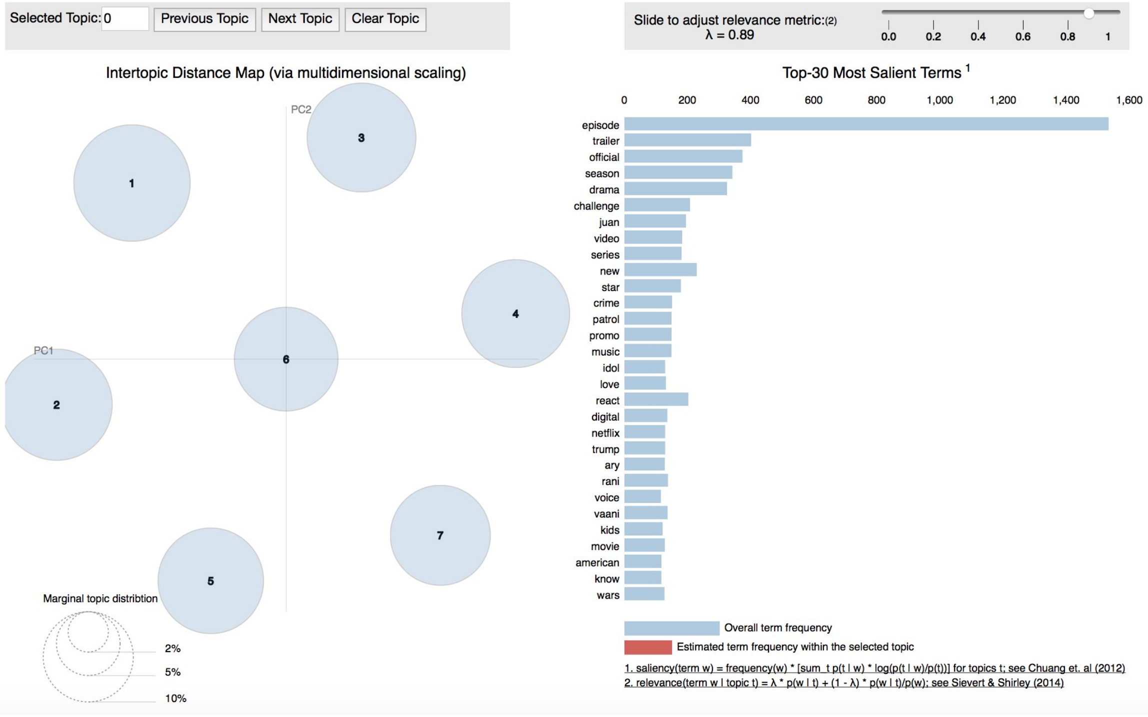


Optimising keyword search

- Since we are interested in optimising the use of keywords for video search, we decided to explore 120,000 daily trending videos in USA, Canada and GB between 2017 and 2018
- To understand the role of text in optimising, we used Natural Language Processing to process, extract and engineer the features in our text data



- Using a technique called Latent Dirichlet Allocation (LDA), we create lists of words that occur in statistically meaningful ways



- For each category, we can see the most salient words and the cluster of words that occur together

Occur less frequently together

Occur more frequently together

Dominant term

Dominant term

Insights

```
# 'Autos & Vehicles':
Most correlated unigrams:
------------------------------
. chevy
. drift
. mercedes
. bmw
. lamborghini
Most correlated bigrams:
------------------------------
. rb engine
. engine conversion
. snowtrax television
. cheap lamborghini
. know speed
```

```
# 'Comedy':
Most correlated unigrams:
------------------------------
. daily
. award
. closer
. yiay
. comment
Most correlated bigrams:
------------------------------
. daily trevor
. trevor noah
. hannah stock
. closer look
. comment award
```

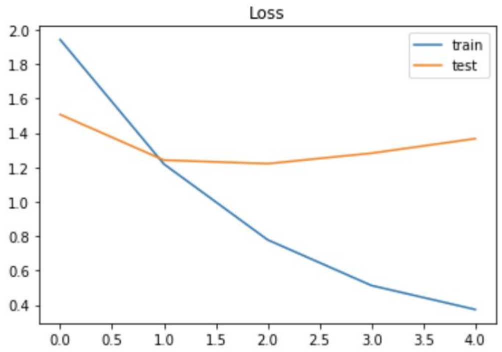
```
# 'Howto & Style':
Most correlated unigrams:
------------------------------
. refinery
. tutorial
. beauty
. hack
. makeup
Most correlated bigrams:
------------------------------
. wing hot
. life hack
. hot ones
. makeup tutorial
. food wish
```

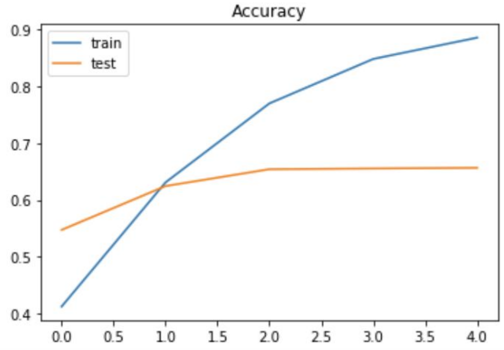
Using LDA we were able to find the most closely related words for the above categories

Predicting the category

- Using a recurrent neural network called LSTM, we tried to classify which category a video would belong to, based on its title

- With a large number of categories (16) our data was fairly imbalanced (more samples in a few categories) which often leads to over-fitting (when the framework captures too much noise and is unable to generalise on all of the data)

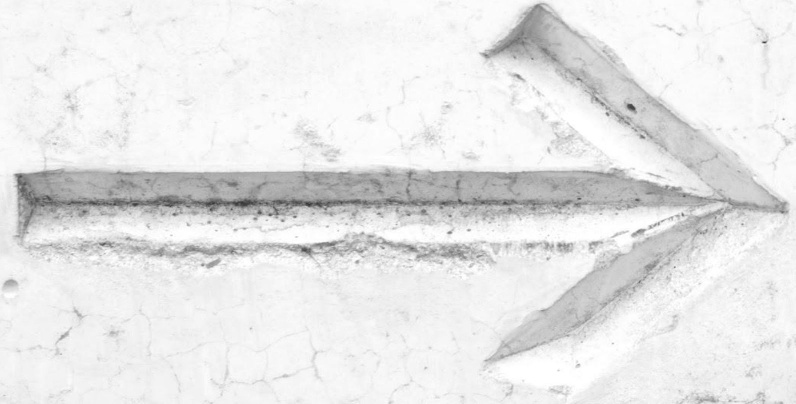




Insights

- However, LSTM is a very good indicator at predicting the more popular categories such as **Music** and **Sports**, which are also amongst the most searched categories on Youtube
- Categories such as **Education** and **Nonprofits & Activism** had a small presence in our data which explains why the framework produced such low precision and recall scores for them

	precision	recall	f1-score	support
Autos & Vehicles	0.63	0.45	0.53	80
Comedy	0.55	0.62	0.58	663
Education	0.55	0.33	0.41	214
Entertainment	0.73	0.75	0.74	2637
Film & Animation	0.65	0.67	0.66	391
Gaming	0.75	0.57	0.65	278
Howto & Style	0.75	0.72	0.73	545
Music	0.82	0.82	0.82	812
News & Politics	0.65	0.72	0.69	825
Nonprofits & Activism	1.00	0.21	0.35	14
People & Blogs	0.44	0.44	0.44	780
Pets & Animals	0.83	0.60	0.70	88
Science & Technology	0.57	0.52	0.54	268
Shows	0.57	0.73	0.64	22
Sports	0.84	0.80	0.82	641
Travel & Events	0.72	0.65	0.68	63
accuracy			0.68	8321
macro avg	0.69	0.60	0.62	8321
weighted avg	0.68	0.68	0.68	8321



- Classification of words based on audience viewing habits to accurately map what people are watching with the right content

- Exploring other techniques which can map tags with categories (the LDA model did not work very well for this)

Mahima Kaushiva

https://www.linkedin.com/in/mahima-kaushiva-340a59a/

Ref

http://bl.ocks.org/AlessandraSozzi/raw/ce1ace56e4aed6f2d614ae2243aab5a5/#topic=0&lambda=0.6&term=

https://www.researchgate.net/publication/261039001_A_LDA-based_method_for_automatic_tagging_of_Youtube_videos

https://analyticsindiamag.com/beginners-guide-to-latent-dirichlet-allocation/

https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/

http://users.umiacs.umd.edu/~jbg/docs/nips2009-rtl.pdf

https://www.machinelearningplus.com/nlp/topic-modeling-python-sklearn-examples/#20howtoclusterdocumentsthatsharesimilartopicsandplot

Appendix

Inferring the topic from keywords

story, new, season, scene, perform, reveal, interview, win, time, year

Movies / Entertain ment

trailer, **love**, want, kid, need, kiss, blood, break, girl, explain

Romance

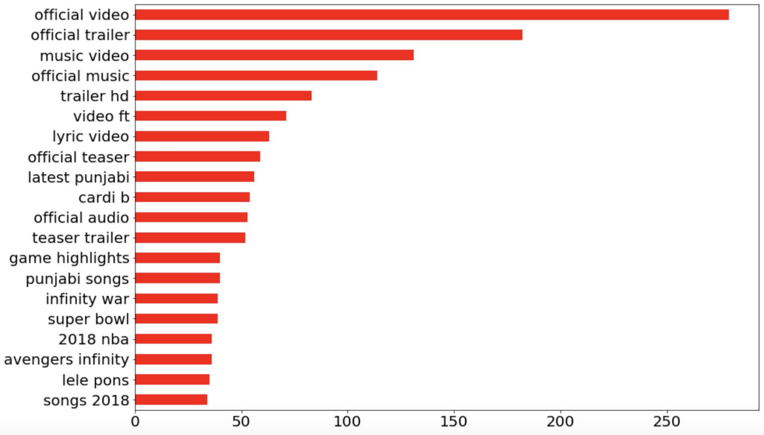
audio, talk, ft., good, performance, song, night, watch, come, cat

Music / Audio

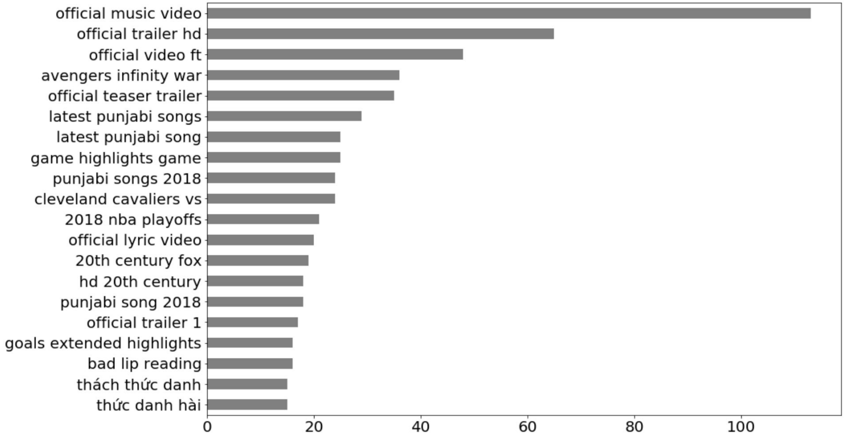
official, video, music, live, movie, lyric, teaser, eurovision, final, cover

Music / Entertain ment

Identifying popular words



Most popular bigrams

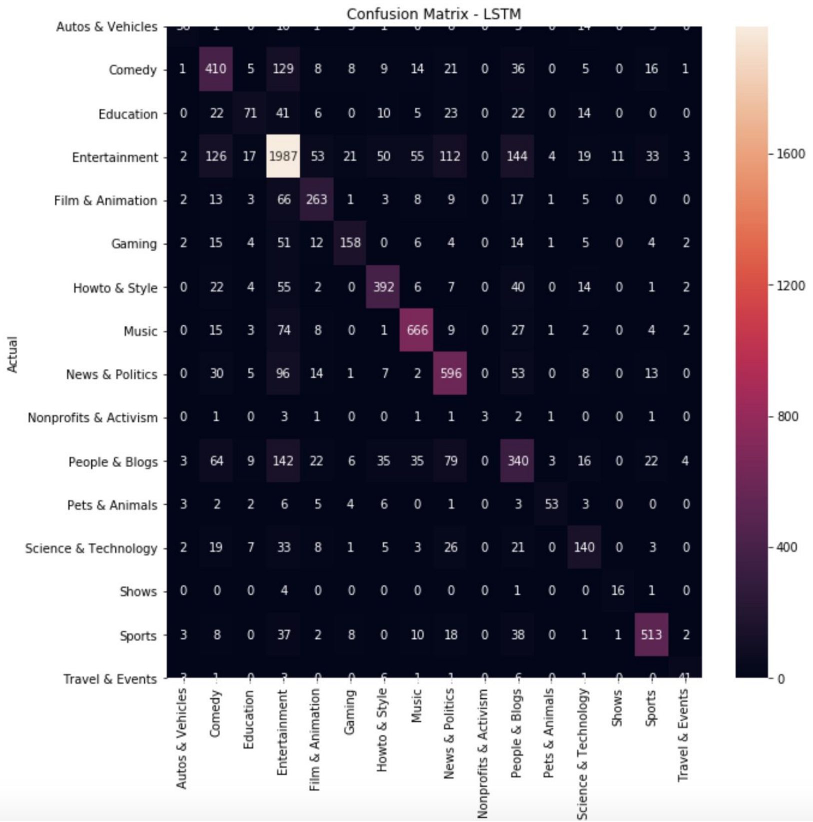


Most popular trigrams

Identifying dominant keywords

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	1.0	0.4480	story, new, season, scene, perform, reveal, interview, win, time, year	[miss, play]
1	1	0.0	0.2500	audio, talk, ft, good, performance, song, night, watch, come, cat	[]
2	2	1.0	0.8792	story, new, season, scene, perform, reveal, interview, win, time, year	[earthquake, deadly, tremor, hit, border, region]
3	3	2.0	0.8040	trailer, love, want, kid, need, kiss, blood, break, girl, explain	[boyfriend, net, worth]
4	4	2.0	0.4429	trailer, love, want, kid, need, kiss, blood, break, girl, explain	[people, awesome, collective, present, pet, awesome]
5	5	0.0	0.6249	audio, talk, ft, good, performance, song, night, watch, come, cat	[wow]
6	6	2.0	0.7472	trailer, love, want, kid, need, kiss, blood, break, girl, explain	[face, transplant, patient, meet, donor, family]
7	7	2.0	0.7259	trailer, love, want, kid, need, kiss, blood, break, girl, explain	[love, sight]
8	8	2.0	0.8301	trailer, love, want, kid, need, kiss, blood, break, girl, explain	[kid, use, celeb, connection]
9	9	3.0	0.4183	official, video, music, live, movie, lyric, teaser, eurovision, final, cover	[vertical, video]

LSTM accuracy and CM




```
score, acc = model.evaluate(X_test, Y_test, batch_size=batch_size, verbose=2)
print('Test accuracy:', acc)
```

Test accuracy: 0.6832111477851868