Youtube video titles classification

Kunture Junuspayeva

Problem

The goal of the project is to classify videos using a video title. The idea is to help content makers choose the right title for their video depending on whether it is for children or not.

Objective

Create a model that takes a sentence and produces either 1 (indicating the sentence from kids channel) or a 0 (indicating the sentence from another channel).

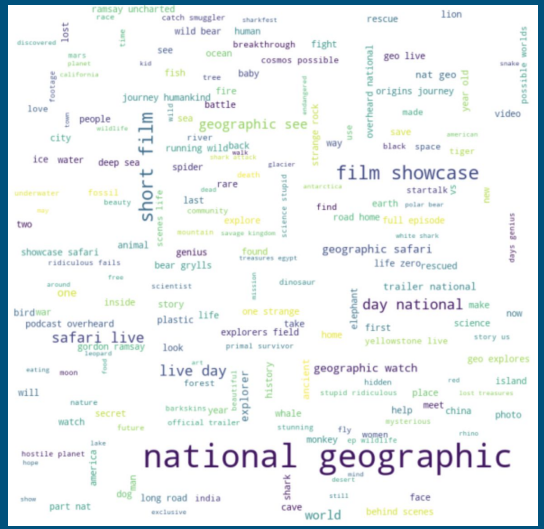
Data source

Videos from two channels on YouTube will be used, thus we get 2 target groups. One of the channels with children's content, and another without. First channel - “Ryan’s World”, second - “National Geographic”.

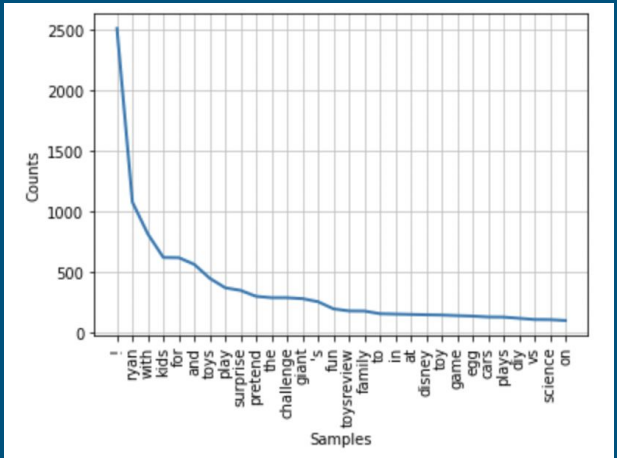
	title	channel_title	desc	date	tags	liked	disliked	views
0	Easy DIY Science Experiment for Kids Rainbow S...	Ryan's World	Easy DIY Science Experiment for Kids Rainbow S...	2021-01-24T13:00:02Z	["Ryan's World", 'Ryan ToysReview', 'science e...	1503	399	200970
1	Ryan play with Giant Soccer Ball and Learn abo...	Ryan's World	Ryan play with Giant Soccer Ball and Learn abo...	2021-01-23T13:00:32Z	["Ryan's World", 'force', 'force and motion', ...	2434	626	331667
2	Ryan hides the Golden Egg from King Collectors...	Ryan's World	Ryan hides the Golden Egg from King Collectors...	2021-01-22T13:00:22Z	["Ryan's World", 'Pretend play', 'Combo Panda'...	2469	605	332385

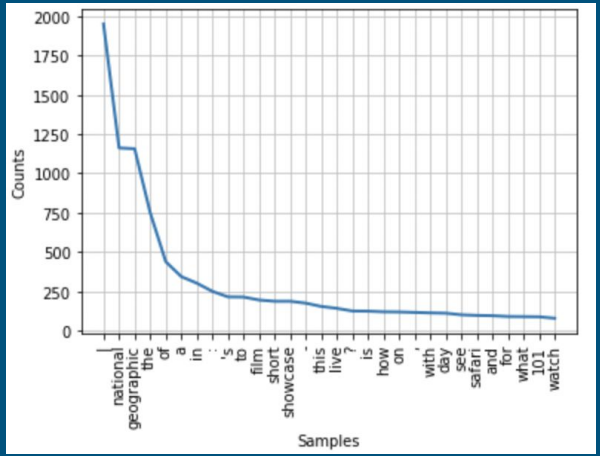
Exploratory data analysis

- Most used phrases in kids channel - ‘ryan toysreview’, ‘pretend play’, 'surprise toy’, ’toy’, ‘kid’
- Most used phrases in natgeo channel - ‘national geographic’, ‘film showcase’, ‘short film’



- Most used words/symbols in kids channel - ‘!’, ‘ryan’, 'with’, ‘kids’ etc.
- Most used words/symbols in natgeo channel - ‘|’, ‘national’, ‘geographic’ etc.





Model preprocessing

Tokenization

Break sentences into word and subwords in the format BERT is comfortable with.

+

Padding

Pad all lists to the same size, so we can represent the input as one 2-d array, rather than a list of lists (of different lengths).

+

Masking

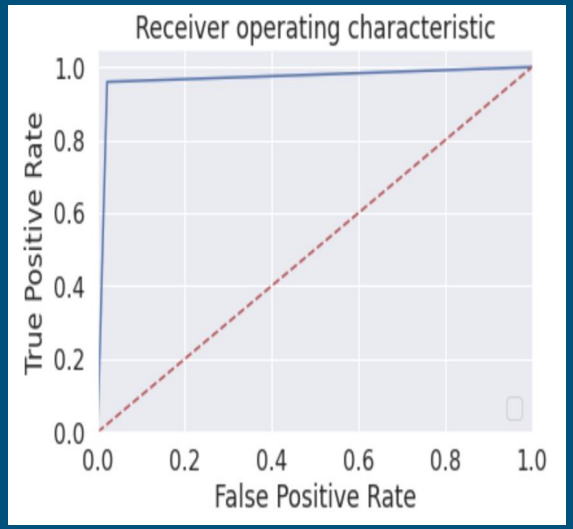
Ignore (mask) the padding we've added when it's processing its input.

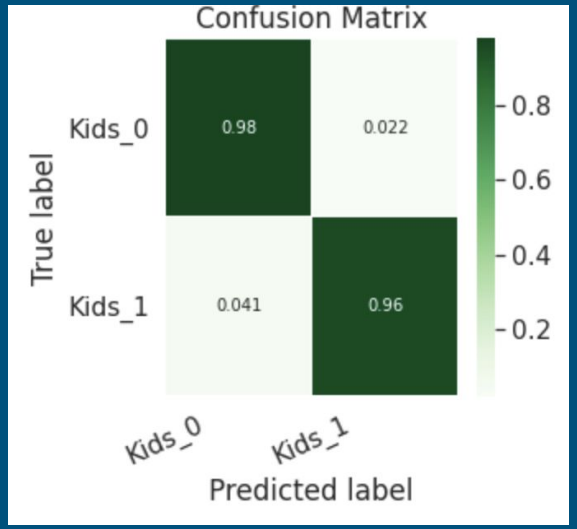
DistilBERT

processes the sentence and passes along some information it extracted from it onto the next model.

Data modeling

- Data are partitioned into train and test sets with a size ratio of 7/3.
- Data modeling using logistic regression
- Model prediction:
  on test set AUC = 0.996





Conclusion

Developed binary classification model with 0.996 AUC. Model can easily classify whether title of video is from kids channels or not.

Recommendation

for more complex model

- Use data of videos more than from two channels
- Pick channels with similar titles, but different content