CP2 Report - Classify youtube videos whether it is for kids or not using NLP

Problem Statement

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.

We want to create a model that takes a sentence (just like the ones in our dataset) and produces either 1 (indicating the sentence from kids channel) or a 0 (indicating the sentence from another channel).

Under the hood, the model is actually made up of two models.

- DistilBERT processes the sentence and passes along some information it extracted from it onto the next model. DistilBERT is a smaller version of BERT developed and open sourced by the team at HuggingFace. It's a lighter and faster version of BERT that roughly matches its performance.
- The next model, a basic Logistic Regression model from scikit learn will take in the result of DistilBERT's processing, and classify the sentence as either positive or negative (1 or 0, respectively).

Data Source

We will load the data using the YouTube API. Videos from two channels on YouTube will be used, thus we get 2 target groups. One of the channels will be with children's content, and the other will not. First channel - "Ryan's World", second - "National Geographic".

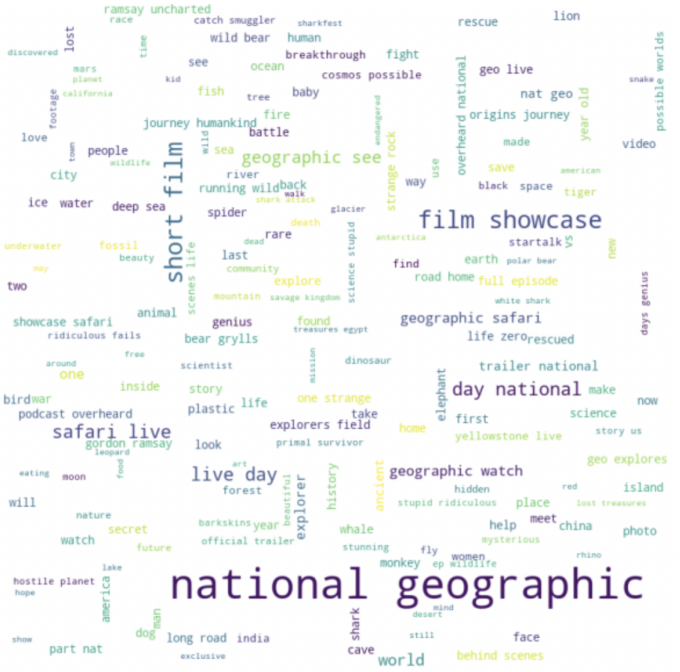
Dataset head:

	title	channel_title	desc	date	tags	liked	disliked	views
0	U.S. men's basketball surges in second half to...	NBC Sports	In a potential Gold Medal Match preview, the U...	2021-07-19T05:00:26Z	[Olympic Sports, NFL, NBA, MLB, NHL, PGA, Golf...	94	3	3753
1	U.S. women's basketball defeats Nigeria in fin...	NBC Sports	The U.S. women's basketball team earns a domin...	2021-07-19T00:45:18Z	[Olympic Sports, NFL, NBA, MLB, NHL, PGA, Golf...	126	7	5066
2	Tour de France 2021: Stage 21 extended highlig...	NBC Sports	The 2021 Tour de France concludes at Stage 21,...	2021-07-18T23:53:35Z	[tour de france, Olympics, cycling, cycling ra...	860	17	59507
3	2021 Tour de France crash compilation | Cyclin...	NBC Sports	From an unaware fan to tumbling down a mountai...	2021-07-18T23:13:34Z	[tour de france, Olympics, cycling, cycling ra...	186	19	18192
4	Haskell Stakes 2021 ends in jockey fall, disqu...	NBC Sports	Watch the 2021 Haskell Stakes at Monmouth Park...	2021-07-17T23:44:19Z	[Horse racing, monmouth park, monmouth 2021, f...	633	35	96102

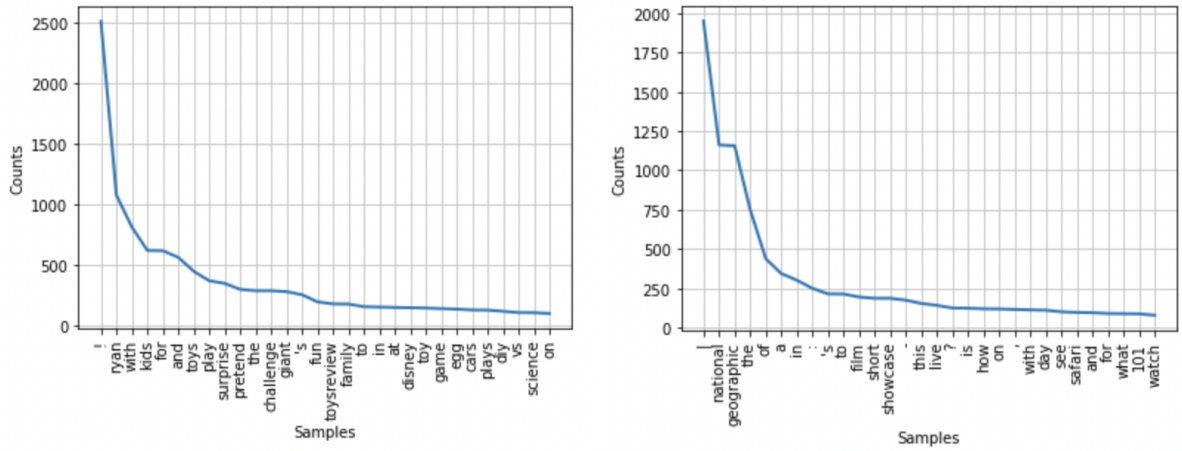
Exploratory Analysis

This project uses the column ‘title’ for the classification model, therefore all exploration was around this column.

We see that following phrases were highly used in these two datasets: ‘ryan toysreview’, ‘pretend play’, 'surprise toy', ’toy’, ‘kid’, ‘national geographic’.



Also when we look for highly frequent words, we can see the same picture.



As observed we have words in our column which are highly frequent, to have more balanced data it was decided to remove them.

Our final dataset head:

	title	channel_title	desc	date	tags	liked	disliked	views	filtered_title
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	Easy Science Experiment for s Rainbow Snowsto...
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	with Soccer Ball and Learn about Force and ...
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	hides the Golden Egg from King Collectors wit...

Model pre-processing

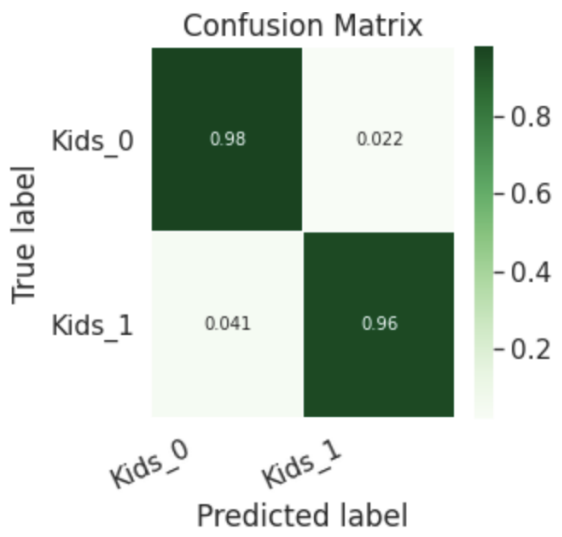
We will use a pre-trained deep learning model to process some text. We will then use the output of that model to classify the text.

Before we can hand our sentences to BERT, we need to do some minimal processing to put them in the format it requires.
Our first step is to tokenize the sentences -- break them up into word and subwords in the format BERT is comfortable with.
After tokenization, tokenized is a list of sentences -- each sentence is represented as a list of tokens. We want BERT to process our examples all at once (as one batch). It's just faster that way. For that reason, we need to 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).
If we directly send padded to BERT, that would slightly confuse it. We need to create another variable to tell it to ignore (mask) the padding we've added when it's processing its input.
Now we have our model and inputs ready.

Data modeling

Data are partitioned into train and test sets with a size ratio of 7/3. Data modeling is performed on a train set with a logistic regression model.

To evaluate the performance of the model in detail, confusion matrix and classification reports are generated. The model predicted the kids related title with a probability of 0.96(recall) and not related with 0.98. Thus, the model is efficient in classifying the kids related titles.



Conclusions and Discussions:

As we can see above our model can easily classify our title sentences whether it's kids related or not.
Possible reason for this can be that we used two Youtube channels, to have a more complex model we need more than two dataset.
Another possible reason is that our data sources from distinct channels(Nat geographic and Kids channels), maybe we should use two channels with similar words in the title but with different content.