Natural Language Processing

Project Description - Clustering of YouTube News Channel Videos based on Title & Description

Jan Paul Kieffer
Chiying Zoe Lai
Max Wurthmann

1 Motivation

We are trying to create a overview of the content on news channels broadcasting in English on YouTube. By using natural language processing models, we want to cluster the videos by topic. Once that is done, we can compare the focus on topics by different news station. Arguably the length of a video might need to be considered, when it comes to making assumptions concerning the focus of news outlets. At this stage the number of views in relation to topics should give us an insight in consumer behaviour. Last but not least, we want to get familiar with using language models for future projects.

2 Goal(s)

- **Clustering of data by topics** (i.e. topics like "Ukraine War", "Economy", "Politics", ...)
  - Different methods of clustering
- **Comparison of clusters within single news station**
  - Distance between clusters
  - Size and popularity of clusters
  - Cluster development over time
  - Relation between emotionality of title and view count
  - Sentiments of news channel on identified topics
- **Comparing clusters of different news channels**
  - Differences and overlaps of clusters
  - Comparison of similar clusters (i.e. matching categories)

3 Data

We have collected the **metadata of uploaded videos** on the following Youtube news channels (DW-News, CNN-News, BBC-News, Al-Jazeera-English, Fox-News and CCTV-Video-News-Agency).

3.1 Data Collection

The collected information of the videos has the following attributes.

Title	Views	Video-Length	Description	Time-Upload	Time-Crawling	Channel
...	...	...	...	...	...	...

The data of the approximately 30 latest videos, of each news channel, are queried every 15 minutes since 25-03-2023.

3.2 Data Preprocessing

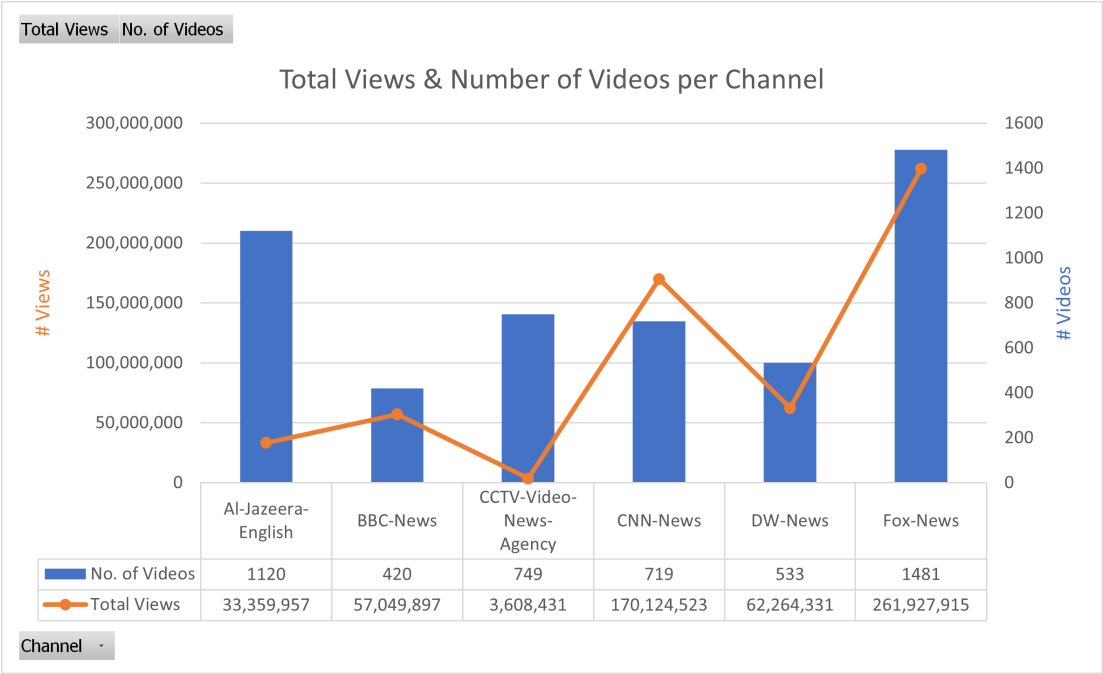
The data of each channel are collected separately and we have got a total of 6 csv files. To aid further processes, we combined all the data together into 1 consolidated data set in csv format.

There are some data cleaning processes required here. Firstly, we would like to conform formatting such as removing excessive white spaces in some cells, which may lead to potential interpretation of the cell values later with Python. Secondly, it is highly possible that the same video appeared in the data sets more than once, when the channel uploads at a frequency lower than 30 videos every 15 minutes. Hence we would remove duplicates of the same video, leaving just the latest record of it.

During the integration process, we added a new column to label the source of each video, so we are still able to identify the video channel in the consolidated data set. In addition, we added a "Video-ID"column so that the retrieving of details would be smoother later.

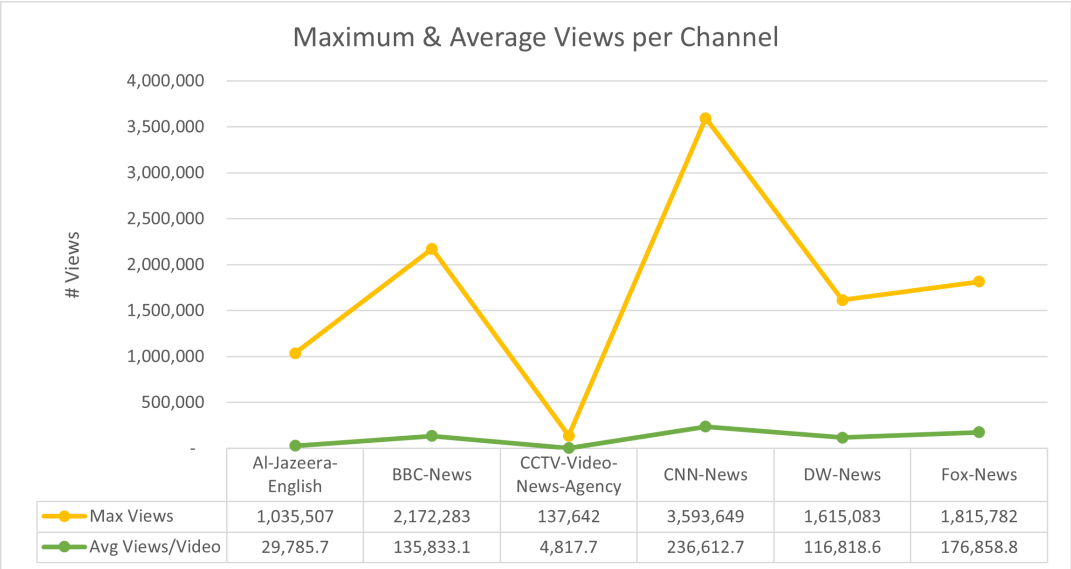
3.3 Exploratory Data Analysis

After cleaning our collected data, we have a total of 5,022 videos in our consolidated data set. Their total views range from 3.6 millions to 261.9 millions. A clear overview of the 6 channels is shown in below graph.



	Al-Jazeera-English	BBC-News	CCTV-Video-News-Agency	CNN-News	DW-News	Fox-News
No. of Videos	1120	420	749	719	533	1481
Total Views	33,359,957	57,049,897	3,608,431	170,124,523	62,264,331	261,927,915

When we look at individual video of each channel, CNN-News has a particular video of over 3.5 millions views. The average views per video, on the other hand, range from 4.8K to 236.6K.



	Al-Jazeera-English	BBC-News	CCTV-Video-News-Agency	CNN-News	DW-News	Fox-News
Max Views	1,035,507	2,172,283	137,642	3,593,649	1,615,083	1,815,782
Avg Views/Video	29,785.7	135,833.1	4,817.7	236,612.7	116,818.6	176,858.8

4 Models and Methods

4.1 Embeddings

- Doc2Vec
- Contextual sentence embeddings using a pretrained Model
  - commonly used available models:
    1. BERT variation like RoBERTa
    2. T5
  - possible inputs:
    1. concatenate title and description of video and use this composed document as input.
    2. input title and description separately and combine the resulting embedding with a weighted mean or a concatenation

4.2 Visalization

Ideas

- Wordcloud: After performing a clustering, display a word cloud for each cluster.
- SOM ? (self organizing maps)
- use common dimension reduction methods to reduce to 2 or 3 dimensions and then plot the data
  - PCA: PCA is good at preserving large distance relations but neglects small distances and neighborhoods
  - t-SNE: t-SNE is good at preserving neighborhoods but distorts large distance relations

4.3 Clustering

There are two potentially interesting approaches to clustering the data.

1. clustering on whole dataset: Better representation of News but distorted by confounding variable (channel)
2. separate data set on variable (channel), and then perform clustering on each subset (corresponding to the channels): Provides insights into differences between news channels

Unsupervised methods: These operate on previously extracted embeddings.

- K-Means Clustering with silhouette plot to determine k, the number of clusters.
- SOM ? (self organizing Maps)

4.4 Topic Inference

Path 1 (unsupervised):

1. perform clustering
2. for each cluster: try to extract an n-gram or tokens that are very frequent in this cluster but comparatively infrequent in other clusters

Path 2 (supervised):

1. annotate parts of the data by hand (e.g. 10%)
2. fine tune a pretrained model to predict the topic based on title and description on the annotated data

5 Tasks of each team member with description

- Data preprocessing (Zoe)
- Create model overview (Max/Paul)
- Problem formulation (Paul)
- Power Point presentation (Zoe)
- Model application & tweaking (all)

6 Your plan and schedule for the project steps

- Data preprocessing - until end of May
- Model selection - until mid June
- Model application - until end July
- Evaluation - until mid July

7 Potential problems

Since the scraper has not collected data for too long, the data set might not be sufficiently large enough to yield significant results, being analysed by our language models. Another disadvantage of the data set, is the incomplete video description. While scraping the data, only the first part of longer video descriptions is collected. During processing the scraped data, all commas and semicolons have been removed, this could have an influence of the effectiveness of the language models.