# Analysis and Preprocessing for Youtube Channel Dataset Report

Team 22
Abdelrahman Mohamed Ahmed Abouelkheir 52-5388

March 2025

## 1 Introduction

**In this milestone, we focus on making dataset analysis and pre-processing for unsupervised text classification machine learning task. The goal is to automatically discover and categorize videos within a YouTube channel dataset according to its content without any predefined labels. This choice of task significantly influences what we aim to find during the data analysis phase. and consequently affects the preprocessing steps required to clean and prepare the data for unsupervised learning methods.**

**The data analysis phase aims to identify the structure of the dataset, understand key trends in the textual data, and highlight potential clusters or topics that could inform our unsupervised classification model and identify inconsistencies through the episodes text and transformation needed to the data. Following this analysis, we will perform preprocessing to ensure the dataset is in a suitable format for machine learning tasks. Preprocessing will include steps such as tokenization, stopword removal, and text normalization, all of which are critical for extracting meaningful features from the data.**

**Moreover, we also leverage the pretrained model ‘faisalq/bert-base-arabic-senpiece‘ from Hugging Face, use its tokenizer to just visualize and also examine how the data is fed to the BERT based models and see their tokenizer. We also wanted to examine the embeddings produced in the output of the model. We also take those emdeddings and extract classification related data and apply k-means clustering then we visualize the data. This provides an initial understanding of how well the data is grouped into meaningful categories. However, further work is needed in the including fine-tuning the model to improve its performance on this specific dataset. Model fine-tuning will allow us**

**to better adapt the pretrained model to the dataset and improve the quality of the unsupervised classification results.**

# 2 Dataset Analysis

In this section, we explore the key attributes of the YouTube channel dataset, For each video we have:

## 2.1 Video Metadata File

Video Metadata file includes attributes such as video titles, publish dates, list of categories, and list of keywords, video length, video url, author. Categories provided is very limited and generic and we want classification to the content of video.For the keywords, we cannot rely on it for the classification task for four reasons. First, It includes items the channel in general. Second, it includes items about the organization owning this channel. Third, it includes other channels that also belong to the same organization. These items in the keywords list is something common between all videos as they belong to the same channel. Fourth, if we decided to clean this data we will be left will small amount of keywords and some those will be generic for classification. I think this data may be used for evaluation of model.

## 2.2 Raw Data

Raw Data the text data for the content of video. We will rely on this file to achieve the task, thereby we need to make some analysis of this text to decide what we will do in the pre-processing phase. .

1. **Inconsistency in Writing:** The text contains several inconsistencies. For example, some text use ة when needed , while other text replace it with ه for the same words. Similarly, some text include أ , while others omit the ء, leading to inconsistency. This lack of uniformity needs to be addressed.

2. **Intro and Outro Phrases:**

   The author includes introductory and concluding phrases at the beginning and end of the text. These phrases are common across different videos.

3. **Incorrectly/incomplete Written Words:** Some words in the text are written incorrectly or incomplete.

4. **Egyptian Arabic with English Words:** The majority of the text is written in Egyptian Arabic, but there are small portions of English words, often used for company names, people's names, or scientific terms related to the topic. These English words are important as they provide context and information about the topic and should be preserved.

5. **Stop Words:** The text contains many stop words, such as لكن, هو, and هي, which do not contribute to the meaning of the content and can introduce noise. These stop words should be removed.

6. **Isolated Letters:** Some parts of the text contain isolated letters with no meaning, which should also be removed as they introduce unnecessary noise.

7. **Unwanted Unicode Characters:** There are certain Unicode characters present in the text that have no meaning and need to be removed.

8. **Bracketed Words:** The text contains words enclosed in brackets, such as [موسيقى] , which are usually references to sounds or actions rather than meaningful text. Since there is no mathematical content that justifies keeping these, all words between brackets should be removed to reduce noise.

9. **Newline Characters and Space Alignment:** The text may contain newline characters (‘\n‘) that break the flow of words but i doesn't act as a separator between sentences. These characters should be removed, and spaces between words should be aligned properly after removing them.

**Note: There are other files in the dataset but they are just for guiding.**

# 3 Preprocessing

To prepare the data for modeling, we perform the following preprocessing steps:

- **Tokenization**: Split text (e.g., video titles and descriptions) into individual tokens (words or subwords) using a tokenizer, such as the BERT tokenizer.

- **Stopword Removal**: Common words (e.g., "and," "the," "is") that do not contribute much meaning are removed to improve model performance.
- **Stemming/Lemmatization**: Words are reduced to their root forms (e.g., "running" becomes "run") to standardize the text.
- **Text Normalization**: Convert all text to lowercase, remove special characters and punctuations to standardize the input format.
- **Handling Missing Data**: Any missing or incomplete metadata (e.g., missing transcripts) is handled by either removing the entry or imputing the missing values.

## 4 Pretrained Model

We leverage pretrained models from the BERT family to fine-tune on our dataset:

- **Pretrained Model Selection**: We use 'faisalq/bert-base-arabic-senpiece', a pretrained model from Hugging Face designed for Arabic text, as the dataset contains Arabic video content.
- **Feature Extraction**: The CLS token embeddings from BERT are extracted to represent each video's textual content.
- **Fine-tuning**: The pretrained model is fine-tuned on the video titles and descriptions for tasks like classification or clustering.
- **Evaluation**: Performance is evaluated based on the model's ability to accurately cluster or classify videos into the appropriate categories.

## 5 Conclusion

In this milestone, we performed detailed data analysis and preprocessing of a YouTube channel dataset, preparing the data for NLP modeling. Using a pretrained BERT model, we extracted feature embeddings and explored their use in clustering and classification tasks. The next step would involve further model tuning, experimentation with additional models, and extending the analysis to broader applications like sentiment analysis and topic modeling.