Aml Yasser El-Sayed 52-21214
Shahenda Maisara 52-23665

Milestone 1 Report

Podcast Topic Modeling & NLP Analysis

Project Overview

This project aims to preprocess a diverse set of 6 Egyptian Spotify podcasts, each belonging to a different category for a Topic Modeling task. The dataset consists of transcribed episodes from various genres, and the goal is to analyze text data, perform exploratory data analysis (EDA), preprocess text, and prepare the data for topic modeling.

Dataset

The dataset consists of 6 different podcasts, each belonging to a unique category:

- Food (16 episodes)
- Relationships (11 episodes)
- Self-Help (25 episodes)
- Educational (2 episodes)
- Comedy (22 episodes)
- TV & Film (43 episodes)

Each episode is stored as a text file, containing transcriptions of spoken dialogue.

The **horizontal** sampling approach ensures category richness and variety to support the training of a topic modeling task.

Exploratory Data Analysis (EDA)

Before preprocessing, EDA was performed to understand the dataset:

- **Metadata Enhancement:** The timestamps were used to calculate episode durations, as they were missing from the original data and saved in a meta data for each episode
- **Podcast Category extraction:** The category of each episode was extracted from its metadata files.
- **Word & Sentence Counts:** Analyzed word count per episode, sentence count, unique words, and most frequent words to understand the data better
- **Visualization**:
    - ## Word Clouds:

      -Generated for each episode to visualize common words.

      - We used matplotlib for visualization, word cloud to generate word clouds, arabic_reshaper to properly connect Arabic letters, and bidi.algorithm to ensure correct right-to-left text rendering. The process involved iterating through podcast directories, reading Arabic transcripts, applying text reshaping and bidirectional formatting, then generating and displaying word clouds to visualize frequently used words in each episode.

      #Insights from Word Clouds (Raw Transcripts)

      - The raw word clouds show many common stopwords (e.g., `"`,`"في"` ,`"انا"` `"على"`) that need removal.

      - "يعني" appears frequently as a filler word, similar to "well" in English.

      - The Egyptian dialect influences word variations (e.g., `"عايز"` instead of `"أريد"`).

      - Some category-specific words are visible but mixed with noise aka "stop words".

      - Therefore Preprocessing is needed (stopword removal, lemmatization, NER) to extract keywords that define each podcast category.

    - **Histograms & Bar Charts**: Compared statistics across podcasts.
- **Sentiment Analysis**: Attempted using pre-trained Arabic models, but results were inaccurate for Egyptian Arabic words.
  We experimented with multiple sentiment analysis models, including **EgyBERT ("elgeish/egybert-arabic-sentiment")** and **AraBERT ("aubmindlab/bert-base-arabertv02-twitter")**,
  using the **transformers** library. We loaded pre-trained models and tokenizers, processed podcast transcripts from raw data , and performed text classification. Since models have token limits (512 tokens), we implemented **text chunking** to analyze longer episodes. However, many models struggled with Egyptian Arabic, yielding **inconsistent or low-accuracy results**. This analysis was essential to determine whether existing sentiment analysis models could be integrated into our topic modeling workflow,
  but results showed the need for a **custom-trained model** tailored for Egyptian Arabic.
    - Pre-trained Models used

  - **EgyBERT (elgeish/egybert-arabic-sentiment)**– Worked but struggled with accuracy.
  - **AraBERT (aubmindlab/bert-base-arabertv02-twitter**– Could process text, but its understanding of Egyptian Arabic was weak.
  - **Other Transformer-based models** – Some didn't even run due to compatibility issues.

  Example output using the arabertv02 model

```
Some weights of BertForSequenceClassification were not initialized from the model checkpoint at aubmindlab/ber
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Text: وحش اوي | Sentiment: Negative | Score: 0.57
Text: حلو اوي | Sentiment: Negative | Score: 0.55
```

which yielded both as negative however the second text should have a positive sentiment.

- Arabertv02:

```
📢 Processing Podcast: كاروهات
📁 Podcast: كاروهات | Episode: كوكتيل منوعات _tokenized.json | Sentiment: LABEL_0 | Score: 0.54
📁 Podcast: كاروهات | Episode: الحلقة التيبس_tokenized.json | Sentiment: LABEL_1 | Score: 0.52
📁 Podcast: كاروهات | Episode: السنارة والعِدة الأرضي_tokenized.json | Sentiment: LABEL_1 | Score: 0.54
📁 Podcast: كاروهات | Episode: الانسان و الاسنان _tokenized.json | Sentiment: LABEL_1 | Score: 0.54
📁 Podcast: كاروهات | Episode: جايه اثير الجدل_tokenized.json | Sentiment: LABEL_0 | Score: 0.55
📁 Podcast: كاروهات | Episode: الأعلام الحمرا _tokenized.json | Sentiment: LABEL_1 | Score: 0.51
📁 Podcast: كاروهات | Episode: يلا ساحل ؟_tokenized.json | Sentiment: LABEL_0 | Score: 0.53
📁 Podcast: كاروهات | Episode: لما تروّحي كليني _tokenized.json | Sentiment: LABEL_1 | Score: 0.54
📁 Podcast: كاروهات | Episode: العاشر من كاروهات _tokenized.json | Sentiment: LABEL_1 | Score: 0.53
📁 Podcast: كاروهات | Episode: كيف بدأ القرف_tokenized.json | Sentiment: LABEL_0 | Score: 0.51
📁 Podcast: كاروهات | Episode: العذراء و الشعرَايه البيضاء_tokenized.json | Sentiment: LABEL_0 | Score: 0.55
📁 Podcast: كاروهات | Episode: خالد و دعاء_tokenized.json | Sentiment: LABEL_0 | Score: 0.58
📁 Podcast: كاروهات | Episode: راس السنه مع راس الافعى _tokenized.json | Sentiment: LABEL_1 | Score: 0.52
📁 Podcast: كاروهات | Episode: القشطوطه مش دايما موجوده _tokenized.json | Sentiment: LABEL_0 | Score: 0.55
📁 Podcast: كاروهات | Episode: تيبس و تريكس الرجوع ل الأكس _tokenized.json | Sentiment: LABEL_0 | Score: 0.53
📁 Podcast: كاروهات | Episode: فول ستوب _tokenized.json | Sentiment: LABEL_0 | Score: 0.51
📁 Podcast: كاروهات | Episode: كونفيوز _tokenized.json | Sentiment: LABEL_1 | Score: 0.56
📁 Podcast: كاروهات | Episode: هرشة جعفر العمده السابعه _tokenized.json | Sentiment: LABEL_0 | Score: 0.52
📁 Podcast: كاروهات | Episode: هابي عيد _tokenized.json | Sentiment: LABEL_0 | Score: 0.54
📁 Podcast: كاروهات | Episode: التوعيه مسؤليه_tokenized.json | Sentiment: LABEL_0 | Score: 0.52
📁 Podcast: كاروهات | Episode: أول يوم_tokenized.json | Sentiment: LABEL_0 | Score: 0.51
📁 Podcast: كاروهات | Episode: بوسطه _tokenized.json | Sentiment: LABEL_0 | Score: 0.51
```

- Camel Bert:

```
📢 Processing Podcast: كاروهات
📁 Podcast: كاروهات | Episode: كوكتيل منوعات _tokenized.json | Sentiment: LABEL_1 | Score: 0.00
model.safetensors: 100% 439M/439M [00:10<00:00, 44.3MB/s]
📁 Podcast: كاروهات | Episode: الحلقة التيبس_tokenized.json | Sentiment: LABEL_1 | Score: 0.00
📁 Podcast: كاروهات | Episode: السنارة والعِدة الأرضي_tokenized.json | Sentiment: LABEL_1 | Score: 0.00
📁 Podcast: كاروهات | Episode: الانسان و الاسنان _tokenized.json | Sentiment: LABEL_1 | Score: 0.00
📁 Podcast: كاروهات | Episode: جايه اثير الجدل_tokenized.json | Sentiment: LABEL_1 | Score: 0.00
📁 Podcast: كاروهات | Episode: الأعلام الحمرا _tokenized.json | Sentiment: LABEL_1 | Score: 0.00
📁 Podcast: كاروهات | Episode: يلا ساحل ؟_tokenized.json | Sentiment: LABEL_1 | Score: 0.00
📁 Podcast: كاروهات | Episode: لما تروّحي كليني _tokenized.json | Sentiment: LABEL_1 | Score: 0.00
📁 Podcast: كاروهات | Episode: العاشر من كاروهات _tokenized.json | Sentiment: LABEL_1 | Score: 0.00
📁 Podcast: كاروهات | Episode: كيف بدأ القرف_tokenized.json | Sentiment: LABEL_1 | Score: 0.00
📁 Podcast: كاروهات | Episode: العذراء و الشعرَايه البيضاء_tokenized.json | Sentiment: LABEL_1 | Score: 0.00
📁 Podcast: كاروهات | Episode: خالد و دعاء_tokenized.json | Sentiment: LABEL_1 | Score: 0.00
📁 Podcast: كاروهات | Episode: راس السنه مع راس الافعى _tokenized.json | Sentiment: LABEL_1 | Score: 0.00
📁 Podcast: كاروهات | Episode: القشطوطه مش دايما موجوده _tokenized.json | Sentiment: LABEL_1 | Score: 0.00
📁 Podcast: كاروهات | Episode: تيبس و تريكس الرجوع ل الأكس _tokenized.json | Sentiment: LABEL_1 | Score: 0.00
📁 Podcast: كاروهات | Episode: فول ستوب _tokenized.json | Sentiment: LABEL_1 | Score: 0.00
📁 Podcast: كاروهات | Episode: كونفيوز _tokenized.json | Sentiment: LABEL_1 | Score: 0.00
📁 Podcast: كاروهات | Episode: هرشة جعفر العمده السابعه _tokenized.json | Sentiment: LABEL_1 | Score: 0.00
📁 Podcast: كاروهات | Episode: هابي عيد _tokenized.json | Sentiment: LABEL_1 | Score: 0.00
📁 Podcast: كاروهات | Episode: التوعيه مسؤليه_tokenized.json | Sentiment: LABEL_1 | Score: 0.00
📁 Podcast: كاروهات | Episode: أول يوم_tokenized.json | Sentiment: LABEL_1 | Score: 0.00
📁 Podcast: كاروهات | Episode: بوسطه _tokenized.json | Sentiment: LABEL_1 | Score: 0.00
```

- **Named Entity Recognition (NER):** which faced similar challenges due to the lack of support for Egyptian Arabic words.
  Detailed explanation :

  We tried multiple models to identify names, places, and organizations in our podcast transcripts, but **not all of them worked well** with Egyptian Arabic.

Models used:
- **CamelTools NER**: didn't work .
- **hatmimoha/arabic-ner (Transformers-based)**: Worked better, recognizing words like *"محمد"* (Mohamed) and *"سيتي ستارز"* (city stars), but missed some entities like "القاهرة"
- **Farasa NER**: An older tool that wasn't great with Egyptian dialect.

Some models **recognized formal Arabic well** but got confused by **dialect words and casual speech**. Others **ignored important names** or classified them incorrectly.


```
  },
  {
      "entity": "PERSON",
      "word": "كريم اسماعيل",
      "score": 0.9834988117218018
  },
```



```
  {
      "entity": "PRODUCT",
      "word": "فيلم عسل اسود",
      "score": 0.8926796317100525
  },
```


This is a sample output using the arabic ner library in which it gave a good results but following is a snippet in which it mismatched the entity

classification :

Preprocessing Steps

To prepare the data for topic modeling, the following steps were performed:

- **Tokenization:** Splitting text into individual words using spacy
- **Text Cleaning:** Removed punctuation, numbers, extra letters, extra spaces. special characters, timestamps, and non-Arabic symbols to clean noise that might introduce irrelevant topics.
- **Normalization:** Unifying variations of Arabic letters.
  Example: أ to ا
  Hence we prevented duplicate representations of the same word.
- **Stopword Removal:**
  - Compiled stopwords from various sources: common Egyptian Arabic stopwords, standard Arabic stopwords, and frequently occurring words from each episode in the dataset.
  - Stored stopwords in a JSON file for consistency.
  - Removed all stopwords from the dataset.
  - We used the nltk library
- **Lemmatization:** The aim of trying lemmetization, was to reduce the number of stop words like (نبقى , تبقى , يبقى) we wanted them to be reduced to a simple word (بقى) to be easily removed and treated with one word but we faced challenges due to limited support for Egyptian Arabic in libraries such as NLTK, SpaCy, and Farasa.
  -we tried several libraries and the most appropriate one was tashaphyne.stemming so we used it for the whole dataset .

- **Word Segmentation**: Planned as much as supported by available libraries.
- Stemming: We did stemming to split the words to the root for and remove pronouns.

Insights Extraction

With the cleaned dataset, we performed additional analysis:

- **Sentence Length Analysis:** Compared sentence lengths across different podcast categories to identify trends. (the different between episode length and sentence length
- **Word Clouds**: Regenerated word clouds post-cleaning to visualize important words per podcast.
    - Observed that all the filler and stop words got eliminated and the main words in the word cloud highlighted the theme and content of each category as shown below for the "ايه المشكلة"podcast which is an islamic podcast so words like "سبحانه" and "شيخ" appeared clearly

- **Keyword Extraction:** Extracted key phrases and relevant words.
  After processing
- **Bi-gram Analysis:** Identified common phrases and expressions used in different categories and to define speaker style.
  Output analyze :

- For من غير مونتاج:

It's clear that the presenter speaks in a casual and engaging way, using a friendly and relatable Egyptian dialect. The presence of phrases like "بنت لذينة" reflects a playful and slightly sarcastic tone . You can tell he\she really engages with the audience with phrases like "أهلا بكم" and "مستمعين أهلا", which make listeners feel welcomed.

- But for البشمهندس :

The presenter's language in this cultural podcast is more structured and analytical, reflecting the nature of the topics discussed. Phrases like "الألفية الثالثة" and "أسباب رئيسية" suggest a formal tone, making the podcast feel more like an intellectual discourse rather than casual conversation.

- After Preprocessing we found that: stemming was actually unnecessary to do as it messed up with the important words, so it was better to keep the tokenized words as is and apply stop words removal on them, we used both combinations in our post-preprocessing analysis to show the difference.

Preparing Data for Topic Modeling

The final step involves structuring the data for effective topic modeling:

- **Creating a Structured Dataset:**
  - Compiled a DataFrame where each row represents an episode.
  - Columns include podcast name, episode name, category, and text content.
- **TF-IDF Vectorization:** Converted text into numerical form using Term Frequency-Inverse Document Frequency (TF-IDF).
- **Finalizing the Dataset:** Saved the structured dataset into a CSV file for future modeling tasks.

Challenges & Insights

- **Egyptian Arabic Complexity:** Existing NLP models for NER & Sentiment Analysis performed poorly due to the lack of Egyptian Arabic support.
- **Short Episode Transcripts:** Some podcasts (e.g., Educational with only 2 episodes) lacked sufficient text data.

Future Improvements

- Train a custom sentiment model for Egyptian Arabic.
- Improve entity recognition using fine-tuned models.
- Implement topic modeling using LDA (Latent Dirichlet Allocation) or another suitable model.
- Evaluate model performance and interpret discovered topics.
- Improve preprocessing to better handle Egyptian Arabic linguistic nuances.