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.
- **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.
- **Visualization**:
    - **Word Clouds:** Generated for each episode to visualize common words.

      Steps: we used `matplotlib` for visualization, `wordcloud` 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., `"انا"`,`"على"` ,`"في"`) 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 "stop words".
      - **Preprocessing is needed** (stopword removal, lemmatization, NER) to extract key words 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 `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 evaluation 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.

```
- **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 textblob library:
```

```
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
```

```
 Here is a sample output using arabertv02 model
which should lead to a negative number but it lead to zero;
```

- **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.

### What We Found
- **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.

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:

- **Text Cleaning:** Removed punctuation, numbers, extra letters, extra spaces. special characters, timestamps, and non-Arabic symbols.
- **Tokenization:** Splitting text into individual words.
- **Normalization:** Unifying variations of Arabic letters.
- **Stopword Removal:**
    - Compiled stopwords from various sources: common Egyptian Arabic stopwords, standard Arabic stopwords, and frequently occurring words from the dataset.
    - Stored stopwords in a JSON file for consistency.
    - Removed all stopwords from the dataset.
- **Lemmatization:** Attempted but faced challenges due to limited support for Egyptian Arabic in libraries such as NLTK, SpaCy, and Farasa.
- **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..
- **Keyword Extraction:** Extracted key phrases and relevant words.
- **Bi-gram Analysis:** Identified common phrases and expressions used in different categories and to define speaker style.
  Output analyze :
  For من غير مونتاج: i
  `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 engage with the audience with phrases like "أهلا بكم" and "مستمعين أهلا", which make listeners feel welcomed.`

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.