NLP milestone 1 Report

- Youssef maged 52-13803
- Kirollos Luka 52-4363

This report will follow our process into cleaning and preprocessing.

Cleaning

First we extracted the texts from the txt files and removed the timestamps and the unnecessary new lines.

Preprocessing

We created a dictionary that will contain all the necessary data needed for our preprocessing.

This included:

- Tokens : These are the tokens extracted from the cleaned text.
- Filtered Tokens : These are the tokens without the stop words and punctuation.
- Full Text : This contains the cleaned text without any further processing.
- Sentiment : This contains the sentiment of the current podcast episode.
- Category : This contains the category of the episode, This was extracted from the JSON file provided in the dataset.

- Channel Name : This contains the channel name of the creator of the episode, This was also extracted from the JSON.

Stop words

First to decide what preprocessing can be done we created a word cloud, However when created on the pure text of the podcast we discovered that the NLTK stop words were not adequate enough for the Egyptian dialect.

We created new stop words that are adjusted to the Egyptian dialect and used them as extra stop words on top of the NLTK ones.

After this we created the word cloud again, and it showed improvement towards showing words that maintain the theme of the podcast episode.

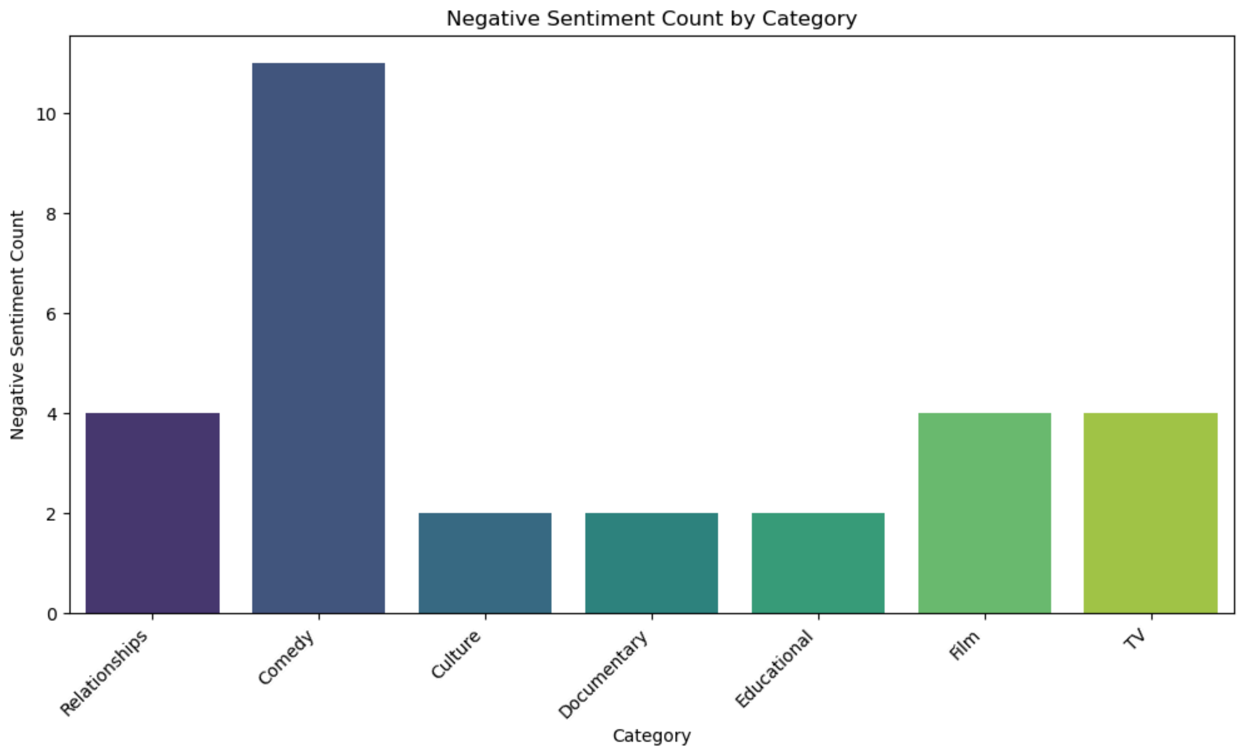
Sentiment Analysis

In this section we provided the podcast texts per each episode to a pretrained sentiment analysis model called camel. This will be used in later parts of the project in our data analysis.

Questions we explored

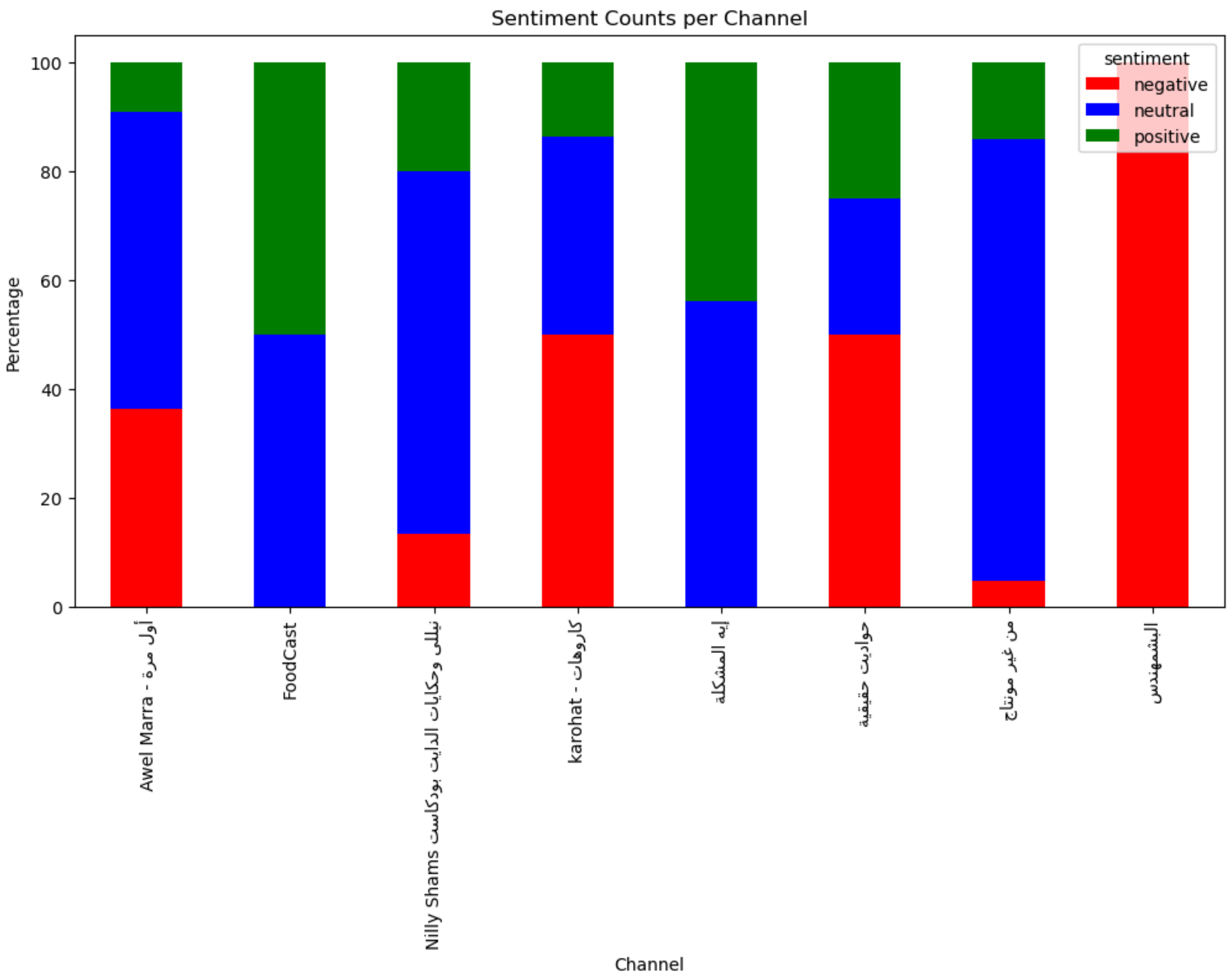
Most Negative Category

When we analyzed all the episodes' sentiments and found the most negative sentiment was the comedy episodes as shown in the below figure. Which is illogical because how all this comedy episodes are negative or sad. This gives us the hint that the pretrained model may be not accurate enough and has trouble detecting the difference between sarcasm and actual criticism



Consistently Negative Channels

We found that the podcast “البشمهندس” is 100% negative, which also has 2 episodes only! That doesn't give us the actual sentiment of the channel. But most of the channels has a reasonable number of negative episodes, for example “FoodCast” does not have any negative episodes which makes sense, how a food podcast contains negative episodes, for
Awel Marra podcast contains some negative episodes which also makes sense for a relationship podcast.



Semantic similarity Among Podcasts

We used two ways to answer this question, we used Tf-idf, and we used embeddings from library torch to find the top 10 pairs of episodes that are similar and if they share the same category.
For the Tf-idf, 9/10 of the pairs share the same category and all of them are ‘Food’ category, which makes sense giving recipes or discussing food have a same style in the podcasts.

For the embeddings approach, We had more variety for the categories this time, More categories other than food were in the top 10 most similar podcasts like Film and comedy. Having comedy in the top 10 shows that the embedding is able to detect similarity better than TF-IDF as its more difficult to detect the similarity between a more general topic like comedy than a specific topic like food.

Identify channels by Speech Style

We tried to find channels and see if it has a unique speech style by finding the most bigram for each channel, and we found that most of the channels has a logical and unique bigram, such as “ايه المشكلة” it is a religious podcast and the most common bigram is “عز و جل” and for البشمهندس addresses his audience by 'سيدي الفاضل' which may be a unique style for this channel, but for example ‘karohat’ the most common bigram is “والله العظيم” which is not a unique word it's a common statement that is said in everyday life of every one of us.

Most common word per sentiment

We analyzed the most common word and most common bigram per all the episodes for each sentiment, there are a lot of common words across each sentiment like “شاء الله”, and no unique words or bigram that is repeated in each sentiment. This approves that the pretrained model may be inaccurate in evaluating the sentiment of the episodes.

Most common word per category

We analyzed the most common words per category. This time we see more unique words per category and also connected to the category, as “مطعم” for the food category, “رجال” and “بنات” for the relationship category, “فيلم” for the Film and TV categories.

Task:

Podcast Category Classification

Based on the analysis conducted, it was observed that different podcast categories exhibit distinct linguistic patterns, common words, and recurring phrases. This suggests that a machine learning model could be trained to classify podcast episodes based on their textual content. As a result, podcast category classification was identified as a suitable task for further exploration.

Objective

The goal of this task is to develop a model capable of predicting the category of a given podcast episode using textual features extracted from the transcript. Automating this process would enhance the organization of podcast content and assist in verifying the accuracy of existing labels.

Feature Engineering

To prepare the data for classification, different textual features can be explored:

- **TF-IDF (Term Frequency-Inverse Document Frequency):** Captures the importance of words within each episode while reducing the impact of common words.
- **Word Embeddings:** Pre-trained word embeddings such as AraBERT and FastText for Arabic text can be used to capture deeper semantic relationships.
- **Bigram & Trigram Features:** Certain phrases are strongly associated with specific categories (e.g., "مطعم" in Food podcasts, "فيلم" in Film & TV podcasts), making n-gram features valuable for classification.

References:

- NLTK
- CAMeL
- AraBERT
- https://github.com/Curated-Awesome-Lists/awesome-arabic-nlp