NLP Milestone 1 Report

Task at hand

Our data analysis aimed to uncover potential patterns between podcast categories and various linguistic and sentiment-based features, with the goal of using these insights to label the data based on the patterns we identified. These labeled datasets would then be used to train a model that could potentially predict the category of a podcast based on its content.

Why did we choose this task?

We chose this task because we realized that it could be highly valuable to identify the category of a podcast purely based on its text. Being able to classify podcasts automatically without relying on metadata or manual labeling would streamline content organization and improve searchability. By reading and analyzing the data, we quickly observed that the results varied significantly depending on the podcast category, reinforcing the idea that linguistic patterns, complexity, and sentiment are strong indicators of a podcast's genre.

Data Preprocessing

Preprocessing:

- **Tokenization**: The text of each episode was separated into an array of words (tokens).
- **Stop-Word removal**: The most common words that were irrelevant to the meaning of the text were removed from the array of tokens such as (ماشي, اه, انت,).
- **Punctuation removal**: All arabic punctuation was removed from the array of tokens (”`÷×–“…”!|+¦~{}',.؟":/؞-][%^&*()<>؛'").
- **Time Stamp removal:** The podcast transcript had time stamps which we removed from the tokens.

Why did we avoid stemming and lemmatization?

We found that stemming and lemmatization changed some words' meanings drastically. For lemmatization many words share the same root but have very different meaning, while for stemming it cuts parts of the words which may lead to meaningless words and ruin our semantic precision.

Data Analysis

Our data analysis involved multiple techniques, each providing different insights into our data. These techniques include:

It is important to note that our decision to use multiple data analysis techniques was based on the belief that each technique represents a unique feature, and no single method should be the sole determinant in categorizing a podcast.

Word Cloud

It is a visual representation of text data where words are displayed in different sizes based on their frequency in the text. The more frequently a word appears, the larger and bolder it is in the cloud. This information contributes to the labeling process, as it helps confirm whether the textual content aligns with its assigned category. The extracted keywords from the word cloud can then be used as features in the model, aiding in the prediction of a podcast's category based on its content.

Ngrams

It is a sequence of N consecutive words in a text. Analyzing Ngrams helps in understanding common word pairings and contextual relationships within a dataset. We used it to see the most common phrases and try to relate it with the podcast category. By examining these frequent word combinations, we can better understand how language is used in different categories, aiding in the labeling process. These identified phrases can then serve as important features for

training the model, improving its ability to predict a podcast's category based on its textual content.

Lexical Diversity

It measures the variety of words used in a text. It is typically calculated as the ratio of unique words to the total word count. A high lexical diversity score suggests a rich and varied vocabulary, while a low score indicates repetitive language. This metric is useful in assessing how different categories can have varying diversity.

Readability complexity

It refers to how difficult a text is to understand. We measured it the Gunning Fog Index, which considers factors such as sentence length and word difficulty. We used readability analysis to see how complex different categories can be. We would then use this as a feature to train our model.

Sentiment Analysis

It is the process of determining the emotional tone of a text. It classifies text as positive, negative, or neutral based on word choices and context. This technique is widely used in social media monitoring, customer feedback analysis, and market research.

How did we run the dataset through our preprocessing and analyzing techniques ?

We began by collecting the transcripts of all episodes from each podcast. These transcripts were then concatenated into a single text for each podcast. Next, we processed and analyzed the combined text using our code. We were trying to

find the relation between the average sentiment of the episodes of a podcast, the category of the podcast (from the Ngrams extracted) and the complexity/lexical diversity of the language used. The results presented below show the output we got from all the episodes of each podcast.

Readability Score Measure:

Score	**Difficulty Level**
6-8	Very Easy
8-10	Easy
10-12	Moderate
12-15	Difficult
15+	Very Difficult

Lexical Diversity Measure:

TTR Score	**Diversity Level**
0.3 - 0.4	Low Diversity
0.4 - 0.5	Moderate Diversity
0.5 - 0.6	High Diversity
0.6+	Very High Diversity

Summary About Results:

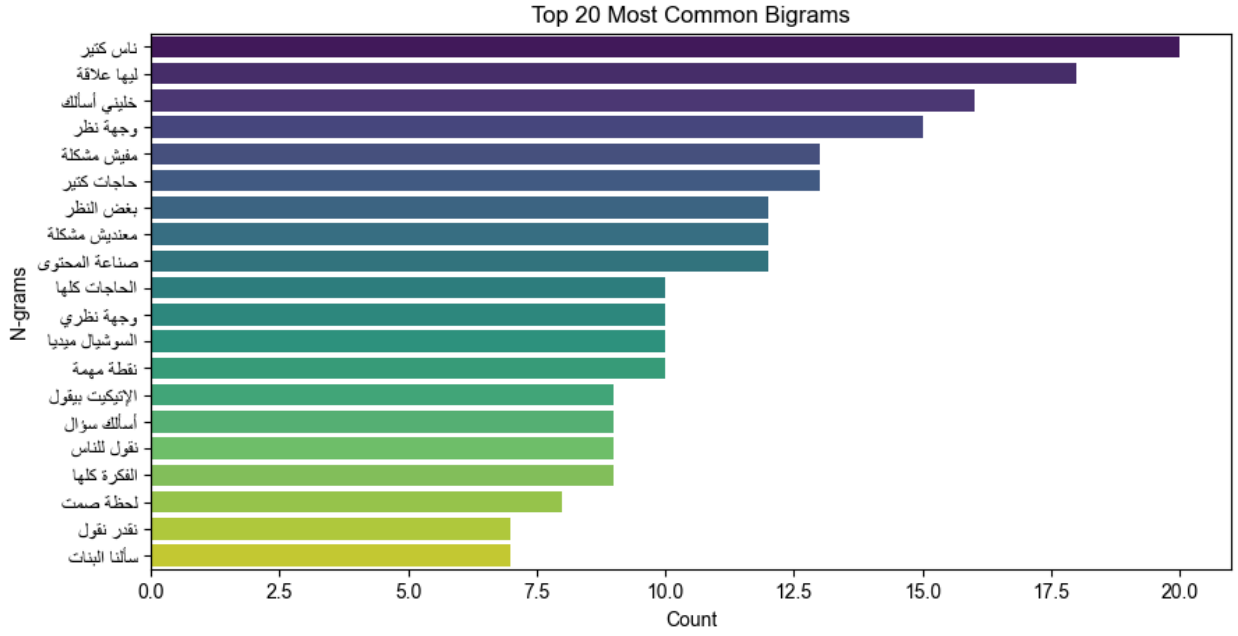
Podcast	Sentiment	Readability Complexity	Lexical Diversity	Category based on most common phrases
Awl Marra	Negative	9.4	0.32	Relationships
Foodcast	Neutral	7.73	0.21	Food/Social Media
Karrohat	Negative	10.84	0.39	Unknown
Nilly Shams	Neutral	10.79	0.26	Health and diet
Eih el moshkela	Neutral	7.52	0.26	Religious
El bashmohandes	Negative	13.25	0.73	Technical
Hawadeet ha2ee2ya	Neutral	9.89	0.66	Unknown
Men gher montaj	Neutral	14.23	0.21	Movies

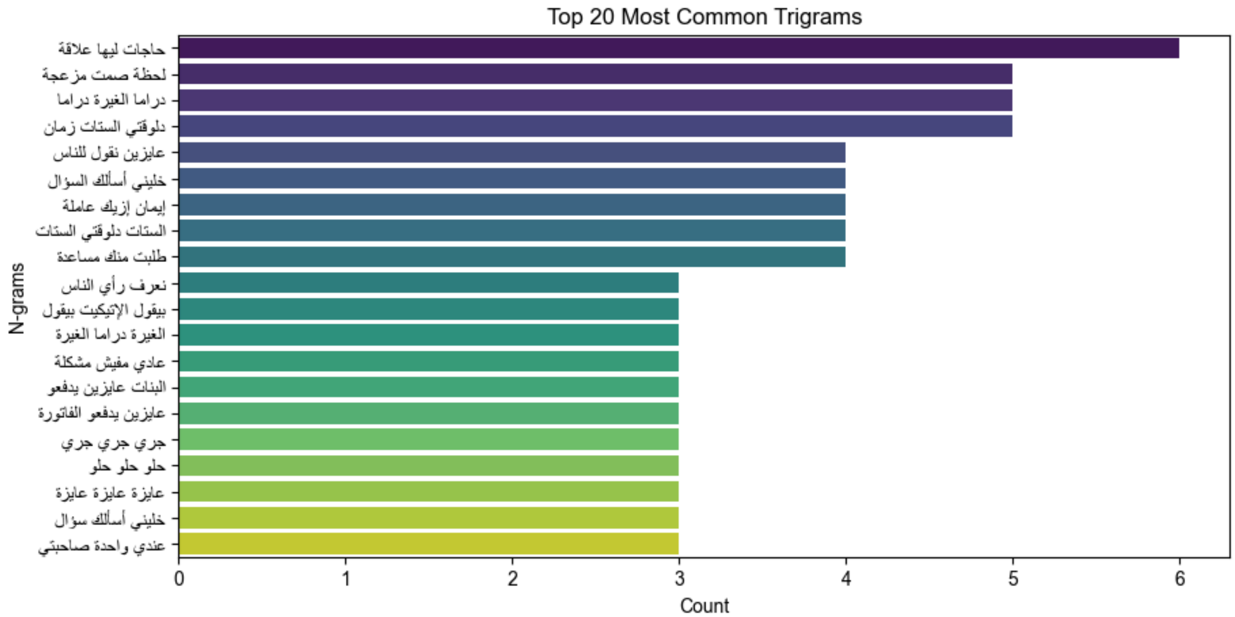
From the results above, we found that some podcasts, like *Elbashmohandes*, had high complexity and were lexically diverse, making it difficult to identify their category based on the most common phrases. On the other hand, podcasts like *Foodcast* were easier to understand and had lower lexical diversity, allowing their category to be easily determined based on the most common phrases. This does not mean that these features are solely dependent on each other, but it is an interesting observation to keep in mind.

*****Comment about sentiment**

When we used sentiment analyzer we applied it on episodes and then saw the most frequent sentiment across episodes for an individual podcast from the above results there is no positive podcast which can mean that our sentiment analyzer is predicting wrongly which can be due to a couple of reasons one of them can be because that we give it a big chunk of text to predict on (an entire episode) or that the egyptian dialect's sentiment can be hard to categorize.

Awl Mara

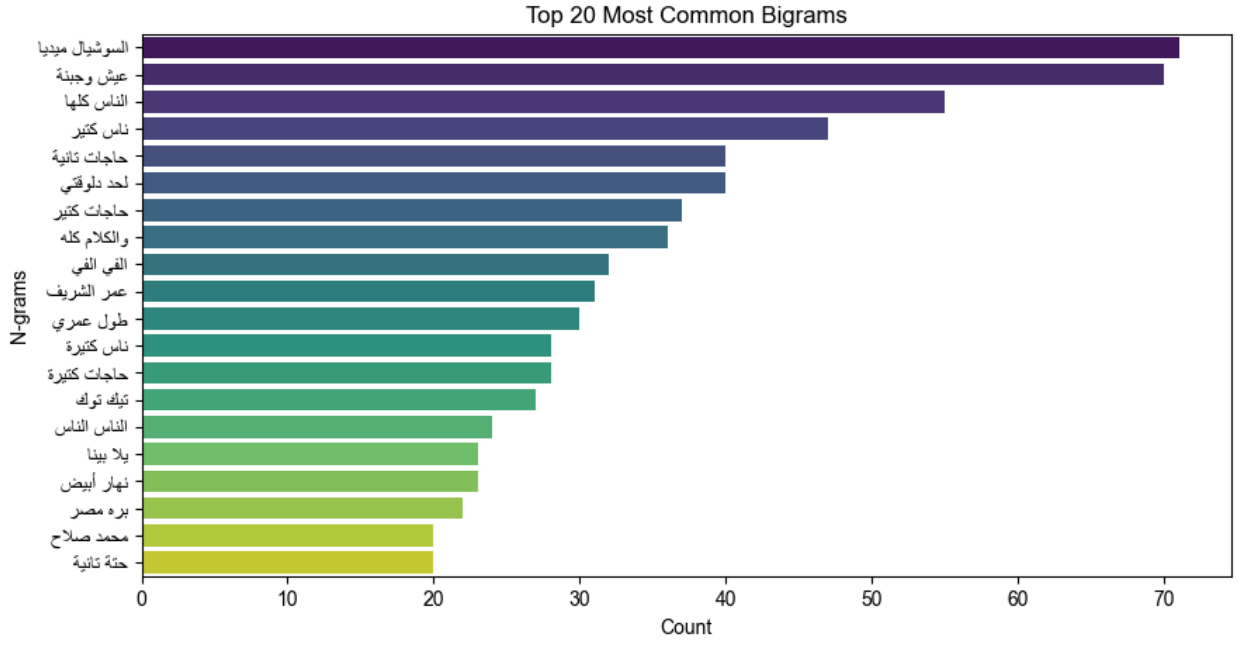


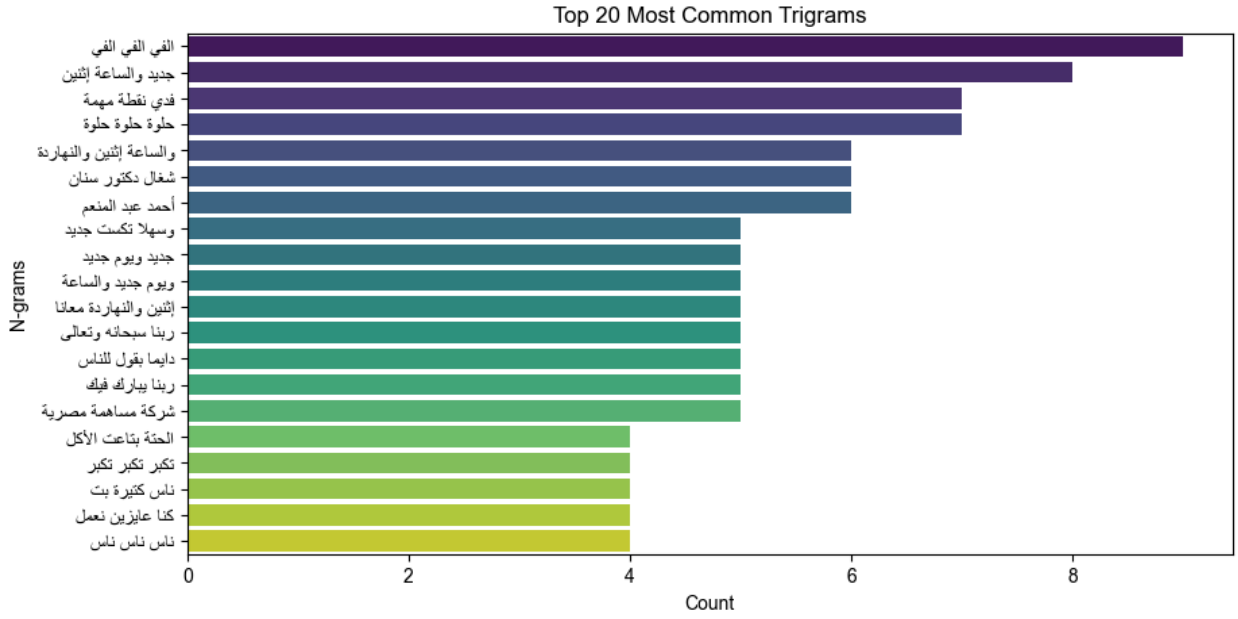


As you can see in the top common bigrams/trigrams and word cloud of the **Awl Mara** podcast we find some phrases that give us an idea on what this podcast’s main category is. Phrases like

(دراما الغيرة ,دلوقتي الستات ,البنات عايزين يدفعوا ,الراجل ,البنات, العلاقة) shows us that this podcast might be speaking about relationships.

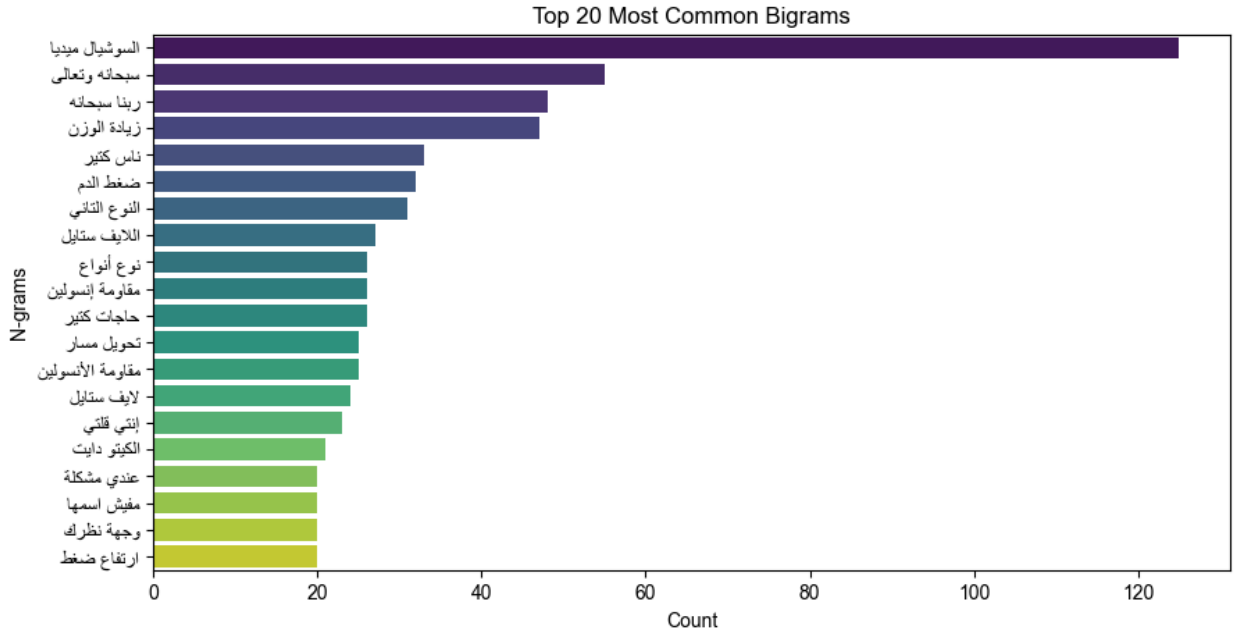
Foodcast

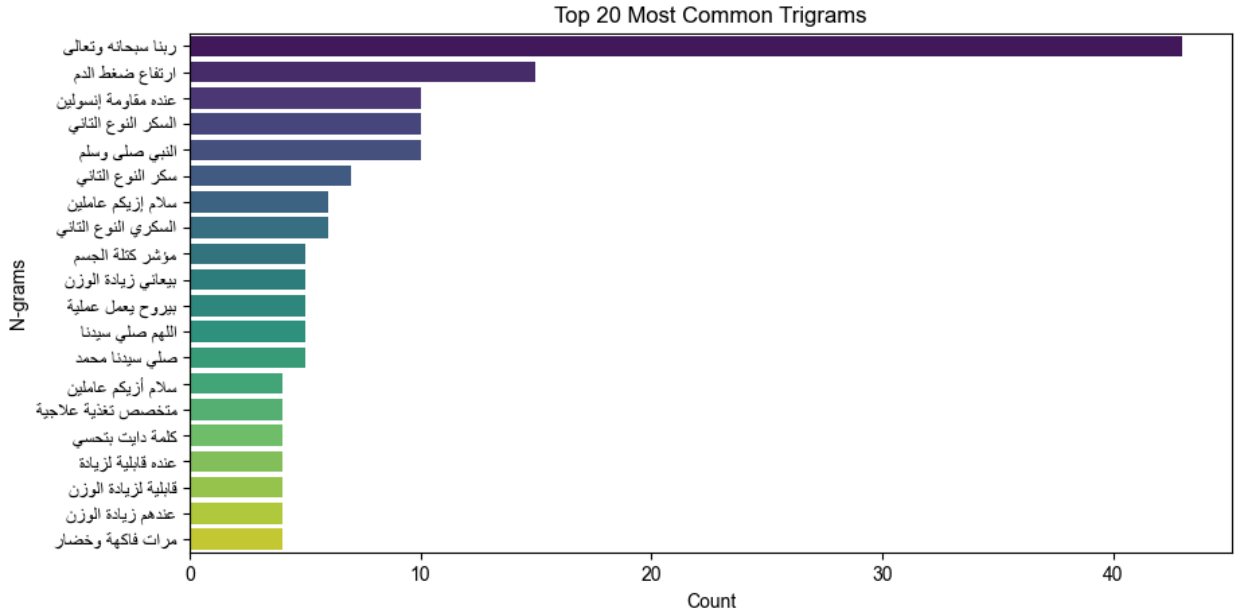




In the podcast **Foodcast** we found some phrases from the ngrams and the world cloud that imply that the category of the podcast is Food like (عيش و جبنة, الحتة بتاعت الأكل, المطعم, الاكل). However there were alot of other common phrases that imply something else such as (السوشيال ميديا, التيك توك, شغال دكتور أسنان). This made it harder for us to categorize this podcast.

Nilly shams


Top 20 Most Common Bigrams
السوشيال ميديا
سبحانه وتعالى
ربنا سبحانه
زيادة الوزن
ناس كتير
ضغط الدم
النوع التاني
اللايف ستايل
نوع أنواع
مقاومة إنسولين
حاجات كتير
تحويل مسار
مقاومة الأنسولين
لايف ستايل
إنتي قلتي
الكيتو دايت
عندي مشكلة
مفيش اسمها
وجهة نظرك
ارتفاع ضغط
N-grams
0
20
40
60
80
100
120
Count


Top 20 Most Common Trigrams
ربنا سبحانه وتعالى
ارتفاع ضغط الدم
عنده مقاومة إنسولين
السكر النوع التاني
النبي صلى وسلم
سكر النوع التاني
سلام إزيكم عاملين
السكري النوع التاني
مؤشر كتلة الجسم
بيعاني زيادة الوزن
بيروح يعمل عملية
اللهم صلي سيدنا
صلي سيدنا محمد
سلام أزيكم عاملين
متخصص تغذية علاجية
كلمة دايت بتحسي
عنده قابلية لزيادة
قابلية لزيادة الوزن
عندهم زيادة الوزن
مرات فاكهة وخضار
N-grams
0
10
20
30
40
Count

As shown above some of the most common phrases and words extracted from Nilly Shams podcast tells that this podcast speaks on health and diet. Some of these phrases/words are
(زيادة الوزن, ضغط الدم, اللايف ستايل, الكيتو دايت, السكر النوع التانى, متخصص تغذية).

Summary of Our Data Analysis and Findings

In this milestone we explored various linguistic and sentiment-based techniques to uncover patterns in podcast transcripts and determine whether these patterns could be used to predict podcast categories. We utilized word clouds and n-gram analysis to identify the most frequently used words and phrases in each category, examined lexical diversity and readability complexity to assess language complexity and difficulty, and used sentiment analysis to evaluate the emotional tone of different podcasts.

Our results showed that while some podcasts had low lexical diversity and repetitive patterns that made their categories easy to determine such as *Nilly Shams*. Podcasts like *Elbashmohandes*, were more complex and lexically diverse, making it harder to classify them based solely on common phrases. This suggests that while our data analysis approach is effective, it may require some additional enhancements to improve accuracy, especially for more complex and diverse podcasts.

One key challenge we faced was handling Egyptian Arabic dialect variations, as existing stopword removal techniques did not always work effectively for this context. Finding stopwords that properly fit the Egyptian dialect was difficult, which sometimes led to irrelevant words being included in the analysis. Improving stopword filtering by creating more dialect-specific stopword lists could help refine the analysis and improve classification accuracy.

Additionally, achieving better accuracy would likely require more advanced Arabic language processing tools, as current libraries may not fully capture the nuances of Arabic text. Using Arabic language context-aware models could enhance our ability to distinguish between categories more effectively and improve the data analysis process, making the classification process more accurate.

References

- ChatGPT
- CAMeLBERT-DA SA
- MohaTaher
- Gasser Ali