NLP Milestone 1 Report

- 52-3899 Karim Mohamed Gamaleldin T-01
- 52-1008 Aly Raafat AbdelFattah T-01
- 52-4509 Omar Ahmed Saad T-02

March 2025

1. Introduction

1.1 Project Overview

This report details the progress of Milestone 1 for the NLP Project. The primary focus is on data preprocessing and analysis to prepare the dataset for downstream NLP tasks such as Multi-label Tagging.

1.2 Goal Tasks

A. Multi-Label Tagging

The goal of this task is to classify transcripts with multiple tags using a deep learning model for multi-label classification. To build the training dataset, labels will be extracted from the transcripts using prompting with large language models (LLMs) and TF-IDF. These labeled transcripts will then be used to train a classifier that can predict multiple relevant tags for new transcripts.

B. Retrieval System with Clustering and Topic Modeling

To further improve retrieval efficiency, we incorporate clustering and topic modeling as essential subtasks within the system. These techniques help structure the data, making searches faster and more relevant in multiple ways:

- Enhancing Indexing – By grouping similar transcripts, the retrieval process becomes more efficient, allowing targeted searches within predefined clusters.
- Building Thematic Collections – Topic modeling ensures that documents are categorized into meaningful groups, enhancing relevance in search results.
- Enable Hybrid Search – Combining keyword-based retrieval (from extracted labels) with semantic similarity search (using embeddings).

1.3 Datasets Used

For this milestone, the following datasets were used:

- ❖ Da7ee7
- ❖ Al Mokhbir Al Eqtisadi
- ❖ Fi Al Hadaraa

These datasets consist of transcribed text from various sources and were collected as part of a research cluster.

---

2. Data Loading

- The `.txt` files were loaded into a pandas dataframe.
- Metadata files were used to extract additional information such as episode length and tags.

3. Data Integrity Check

- Checked for missing values, duplicates, and incorrect data types.
- Conducted basic statistical analysis using `.describe()`.

4. Data Understanding (EDA)

In this section, the data being used will be explored, so we can gain insights on the data that could help us in downstream tasks we will perform using this dataset.

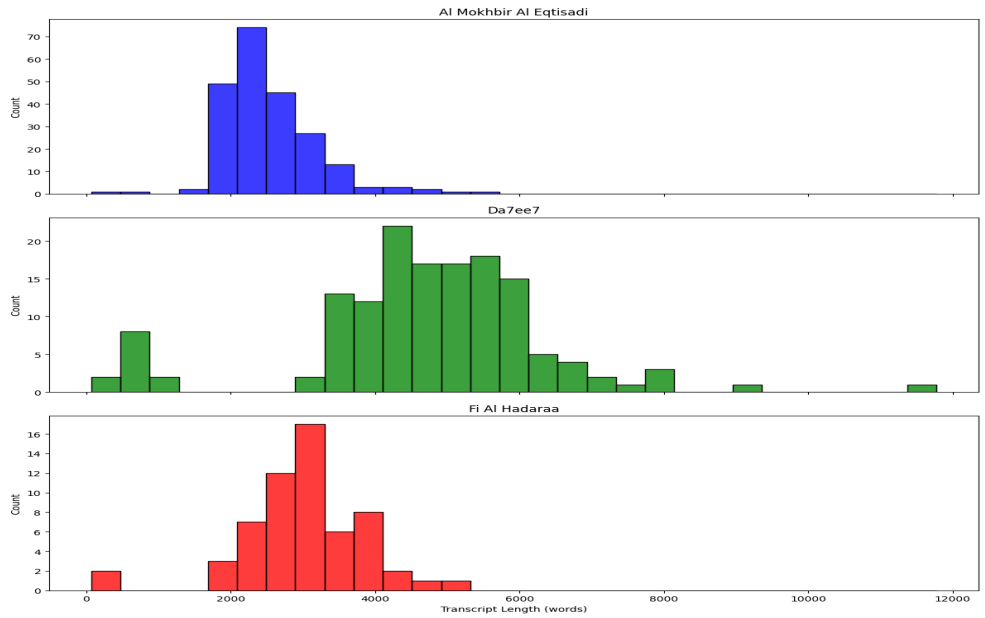
4.1 Length & Transcript Length Analysis

1. The mean number of words of transcript length per creator is being calculated to understand how long the transcripts we have on our dataset are on average for each creator.

Channel Name	Mean transcript length
Da7ee7	4694
Fi Al Hadaraa	2999
Al Mokhbir Al Eqtisadi	2511

- ❖ This tells us that on average, the transcript length of **Da7ee7** is longer than the other 2 creators.
- ❖ Also, Al Mokhbir Al Eqtisadi tends to have the lowest mean transcript length.

2. Plot the histogram to view the distribution of each transcript length per creator.



a. **Variation in Transcript Length Distribution**:
   - The three creators exhibit different distributions in their transcript lengths, indicating variations in content style and format.

b. **Al Mokhbir Al Eqtisadi (Top Plot - Blue)**:
   - The transcript lengths are relatively short and concentrated within a narrow range.
   - Most transcripts fall between **1000 to 3000 words**, with a peak around **2000 words**.
   - This suggests a structured and consistent format, likely adhering to a predefined script length.

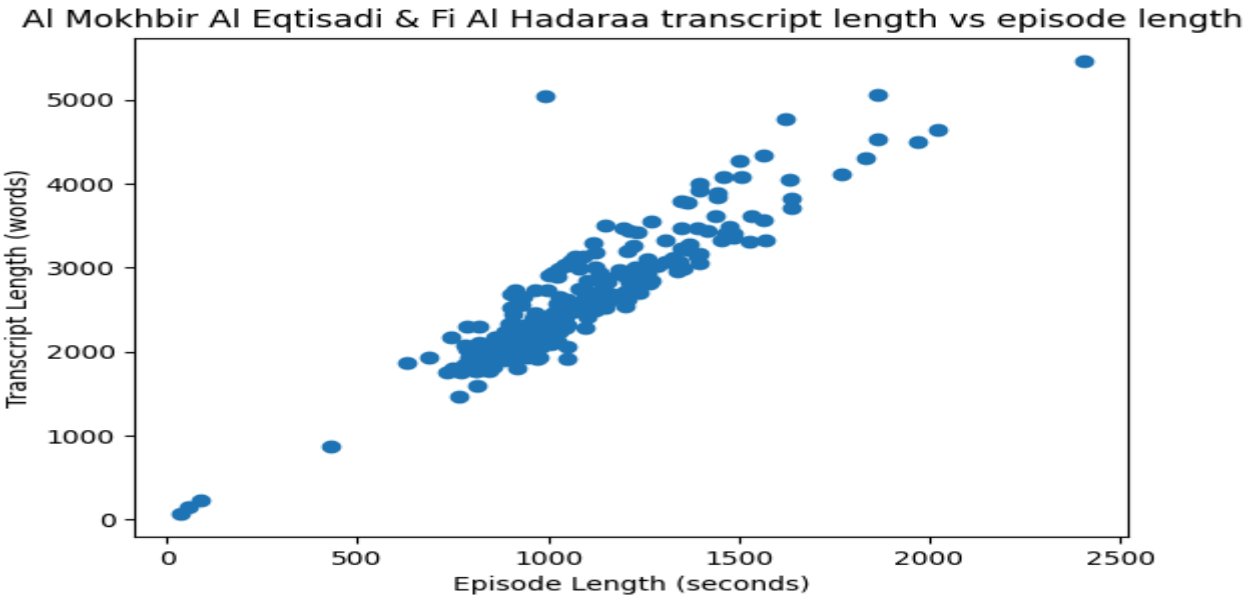
c. **Da7ee7 (Middle Plot - Green)**:
   - The transcript lengths are more widely spread compared to the other two creators.
   - The histogram shows a **bimodal** distribution, with two peaks: one around **2000–3000 words** and another around **4000–5000 words**.

- A few extreme outliers exceed **10,000 words**, suggesting occasional long-form content.
- This variability suggests that Da7ee7's content length fluctuates based on the topic complexity.

d. **Fi Al Hadaraa (Bottom Plot - Red)**:
   - The transcript lengths show a normal-like distribution with a clear peak around **3000–4000 words**.
   - There are fewer short transcripts compared to Da7ee7, indicating a more uniform content length.
   - There are some long transcripts, but the number decreases beyond **5000 words**, suggesting an upper limit for most episodes.

e. **Comparative Insights**:
   - **Al Mokhbir Al Eqtisadi** has the most consistent and shortest transcripts.
   - **Da7ee7** has the widest range of transcript lengths, suggesting flexibility in content delivery.
   - **Fi Al Hadaraa** falls between the two, with a relatively consistent but slightly longer format than Al Mokhbir Al Eqtisadi.

3. Analyzed mean length per creator.
4. Plotted histograms for length distribution per creator.
5. Assessed correlation between episode length and transcript length using scatter plots.



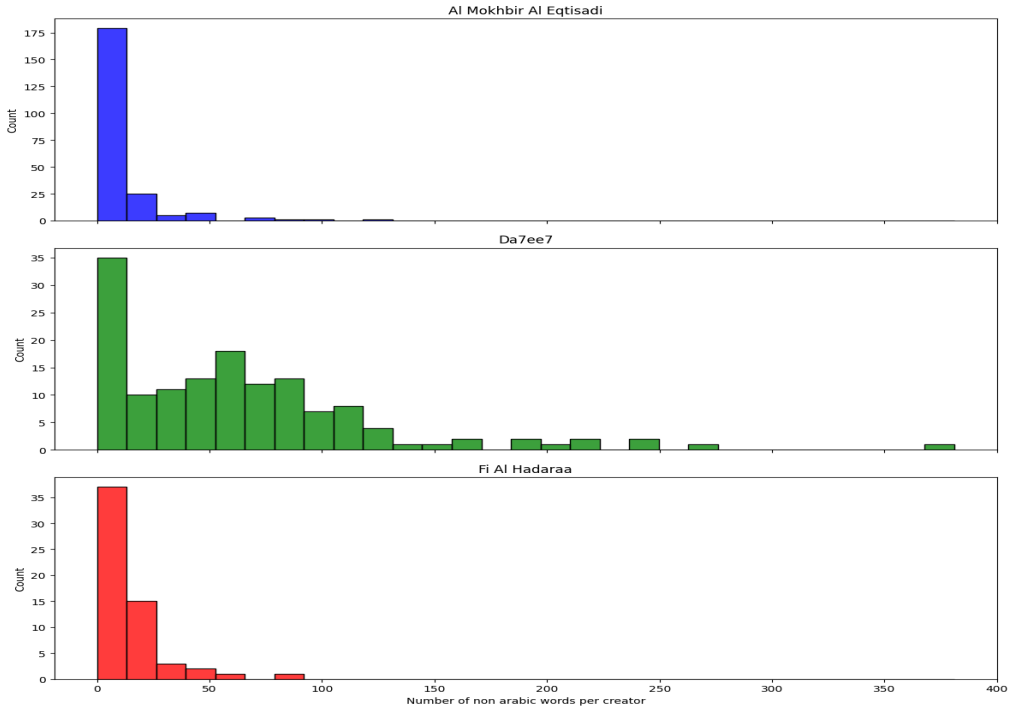
4.2 Non-Arabic Word Analysis

1. Using Regular Expressions identify words that contain non arabic letters “English letters”
2. The mean number of non arabic words per creator is calculated

Creator	Mean number of non arabic words
Da7ee7	63
Fi Al Hadaraa	13
Al Mokhbir Al Eqtisadi	8

- ❖ This tells us that Da7ee7 uses much more english words in his episodes than the other 2 creators.

3. Plotted histograms of non-Arabic word usage per creator to view the distributions



- **Al Mokhbir Al Eqtisadi (Blue - Top Chart)**:
    i. Uses very few non-Arabic words.
    ii. The majority of transcripts have close to **zero** non-Arabic words.
    iii. Suggests a strong preference for purely Arabic content.
- **Da7ee7 (Green - Middle Chart)**:
    i. Uses the most non-Arabic words among the three creators.
    ii. The distribution is more spread out, with many transcripts containing **50–150** non-Arabic words.
    iii. Some episodes have even higher counts, exceeding **200+ words**.
    iv. Indicates frequent use of foreign terms, possibly for scientific or technical discussions.
- **Fi Al Hadaraa (Red - Bottom Chart)**:
    i. Similar to Al Mokhbir Al Eqtisadi but slightly more non-Arabic words.
    ii. Most transcripts have **fewer than 50** non-Arabic words.
    iii. Still maintains a predominantly Arabic vocabulary.

4. Analysing the non arabic word cloud for each creator

❖ Da7ee7

➢ **Common Themes:**

- **Science & Technology:** Words like *AI, ratio, chip, transistors, penicillin, electrodes, unicode, TSMC* suggest discussions on scientific and technological topics.
- **Mathematics & Physics:** Words like *sinx* hint at mathematical or physics-related explanations.
- **Pop Culture & Entertainment:** *Marvel, superheroes, WrestleMania, Jerry* indicate references to pop culture.
- **Business & Economy:** Terms like *business, brand, premier, marketing* suggest economic discussions.

➢ Many of the non-Arabic words are technical in a lot of fields, which is logical as Da7ee7 covers a lot of themes in his episodes.

❖ Al Mokhbir Al Eqtisadi

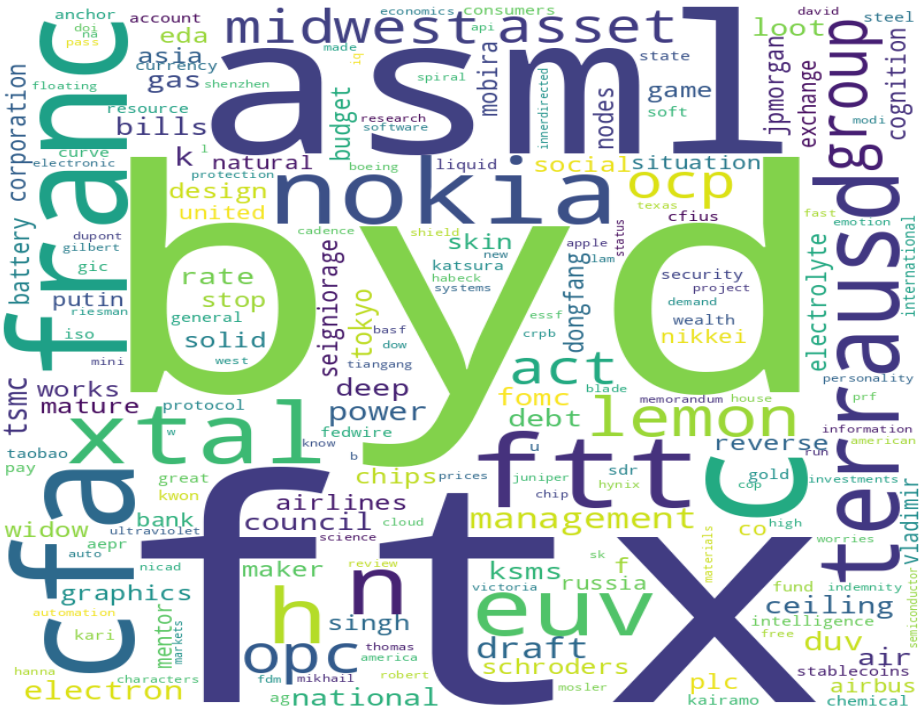
➢ **Common Themes:**

- **Finance & Economy –** Terms related to global markets, investments, debt, wealth, and monetary policies (*asset, debt, budget, currency, nikkei, FOMC*).
- **Technology & Semiconductors –** Focus on chip manufacturing, electronics, and AI advancements (*ASML, TSMC, chips, automation, AI*).

- **Corporate & Business News –** Mentions of major companies and industries (*Nokia, Boeing, Airbus, Schroders, JP Morgan*).

➢ AlMokhbir Al-Iqtisadi's content is heavily focused on economics, finance, global markets, and technology.

➢ Unlike Da7ee7, which had a mix of science, pop culture, and technology, this word cloud is much more finance-oriented.



❖ Fi Al Hadaraa

➢ **Common Themes:**

- **Digital & Social Media Culture** – Terms related to online presence, engagement, and digital influence (*influencers, followers, likes, post, profile, subscribe, online, search*).
- **Entertainment & Pop Culture** – References to gaming, TV series, and fictional worlds (*fantasy, thrones, game, league, bodyguard, zombies*).
- **Consumerism & Lifestyle** – Mentions of modern products and habits (*airpods, scooter, doritos, fried, chicken, gym, credit, cards, application*).

➢ Fi Al Hadara highlights a **blend of entertainment, modern consumer habits, work stress, and social trends**.

2.3 Arabic Word Analysis

- **Generated word clouds for arabic words per creator:**
  - We first cleaned the transcripts using the method of data preprocessing & cleaning mentioned later below.
  - **Al Mokhber Al Eqtisadi:**
    - Dominant Topics: Words like ", "العالم" , "النفط" , "الصين" , "الإسرائيلي" الأمريكي"" suggest a strong focus on geopolitics, global economy, and energy markets.
    - Economic Emphasis: Terms like "الدول" , "النفط" , "الغاز" , "السوق" indicate discussions related to trade, commodities, and international financial dynamics.
    - Political Context: The presence of "احتلال" , "أمريكي" , "الروسي" suggests coverage of political conflicts and international relations.

- Repetition of "الإسرائيليين" & "الإسرائيلي": This might indicate that a significant portion of the content is related to Israeli affairs, possibly in the context of political, economic, or strategic discussions.

- **Da7ee7:**
  - General and Philosophical Tone: Words like "الناس" , "العالم" , "كبير", "الحياة" , "أكثر"" suggest a broad, thought-provoking narrative, characteristic of Da7ee7's educational and storytelling approach.
  - Curiosity-Driven Discussions: The presence of "الحلقة" , "الموضوع", "الكلام"" indicates a focus on explaining various subjects in an engaging manner.
  - Diverse Topics: Unlike Al Mokhbir Al Eqtisadi, this word cloud does not emphasize geopolitics or economics but instead features words related to science, philosophy, and social themes.
  - Conversational Style: The frequent use of words like "بتقول", "تفكر" , "بيحصل"" aligns with Da7ee7's storytelling and explanatory format, which engages the audience in a casual yet informative manner.

- **Fi Al Hadaraa:**
  - Human-Centric Themes: The most prominent words include "الحياة" , "الناس" , "الإنسان" , indicating a strong focus on societal development
  - Unlike Al Mokhbir Al Eqtisadi, which focuses on economics and geopolitics, `Fi Al Hadara` leans toward human civilization, history, and philosophical discussions.

- **All Channels Together:**
  - The combined dataset captures a broader spectrum of topics than individual channels.
  - The mix of economic, historical, and philosophical keywords suggests a well-balanced dataset that covers both factual and analytical discussions.
  - Potential Application: This dataset could be useful for multi-label classification, as different transcripts might intersect across multiple themes (e.g., economy + history, philosophy + geopolitics)
  - Key Recurring Themes Across Channels:
    - Humanity & Society: Dominant words like ","الإنسان" , "الناس" الحياة"" highlight a strong focus on societal topics.
    - Global and Geopolitical Discussions: The presence of ", "العالم الأمريكي" , "الإسرائيلي" , "الروسي" , "الصين" , "الدول"" (American) suggests coverage of international relations and economic affairs

- **Constructed phrase clouds based on:**
    - Performed a simple text cleaning technique on the raw transcripts to see the original phrases
        - Change the text to lowercase
        - Remove punctuation
        - Split the text on white space
    - Then, we perform n-gram generation (bigrams and trigrams) and generate the phrase clouds
        - **Al Mokhber Al Eqtisadi**
            - Phrase clouds generated for bigrams
                - Common phrases like "من" , "على سبيل المثال" , "في نفس الوقت خلال" suggest that the content of ten involves comparisons, contextual explanations, and structured argumentation.
                - This aligns with the channel's analytical and explanatory nature, which aims to break down economic and political events.

- Phrases like "في السوق" , "على قطاع غزة" indicate that discussions revolve around market trends and geopolitical issues.

- Phrase clouds generated for trigrams
  - Phrases like "حلقات جديدة باذن " (new episodes, God willing), "في حلقه في" (in an episode in), "في حلقة النهاردة" (today’s episode) suggest that the most prominant phrases are the promotional or introductory ones.

- **Da7ee7**
  - Phrase clouds generated for bigrams
    - The most prominent phrase "يا عزيزي" is a signature phrase of El Da7ee7, reinforcing his casual, friendly, and engaging storytelling approach.
    - The repetition of "يا عزيزي كدا" and "يا عزيزي إن" indicates a strong use of rhetorical style and audience engagement
    - frequent use of nicknames used like "يا أبو حميد"

- Phrase clouds generated for trigrams
  - Same thing as the phrase cloud of ngrams=2 in which el Da7ee7 uses his signature phrases and catchphrases which is " أقولك يا عزيزي"

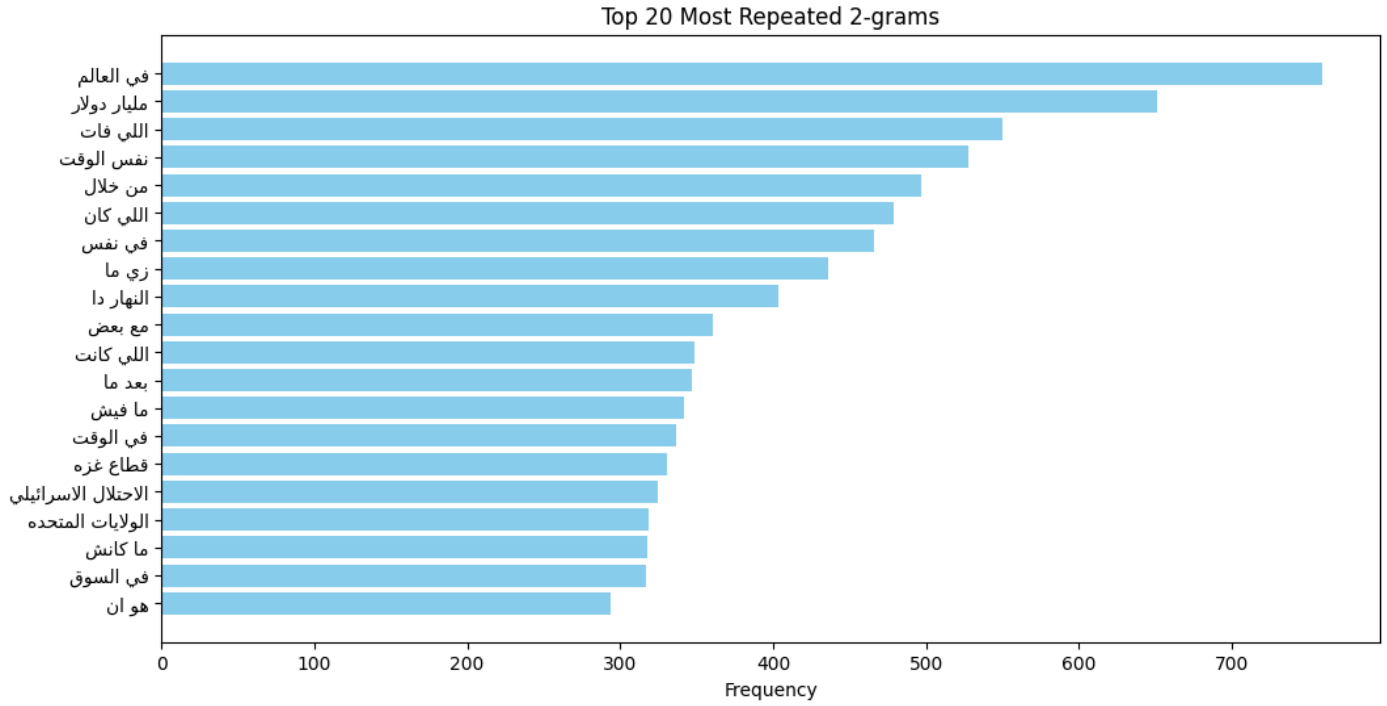
- **Fi Al Hadaraa**
  - Phrase clouds generated for bigrams
    - The phrase "يا عم شلبي" introduces an informal, engaging, and possibly humorous aspect, making the content more relatable and digestible for the audience
    - Frequent use of pronouns, temporal markers, and stopwords.

  - Phrase clouds generated for trigrams
    - Fi Al Hadara uses a direct and engaging storytelling style, frequently addressing the audience as "صديقي الانسان" to create a personal and reflective atmosphere
    - Lots of regularly-used phrases like "في الحالة دي", "ان شاء الله",

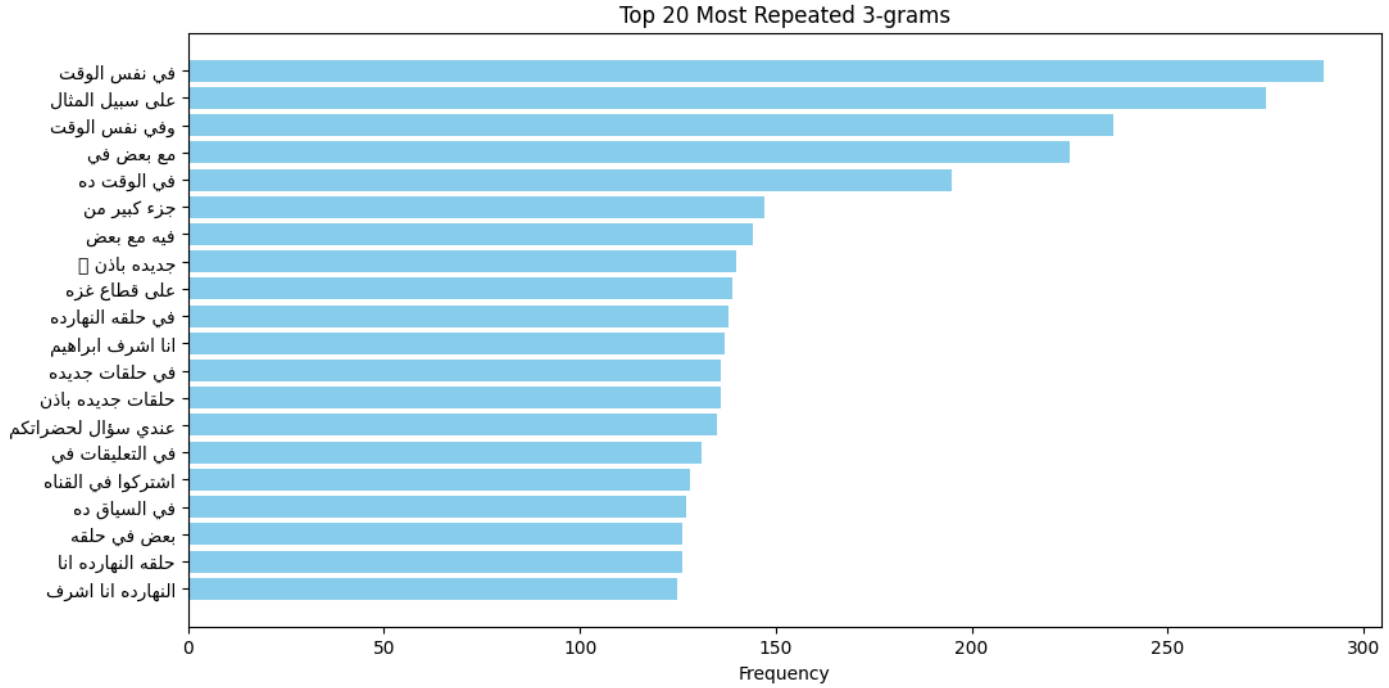
- **All Channels Together**
  - Phrase clouds generated for bigrams
    - Dominance of Conversational and Catchphrase Elements "يا عزيزي", "أبو حميد" coming from Da7ee7 channel and "نفس الوقت" coming from Fi el hadaraa and Al Mokhbir Al Eqtisadi suggests informal, friendly interactions.

- Phrase clouds generated for trigrams
  - The phrases mainly consist of introductory and catchy expressions from all channels, often used to engage the audience and set the context rather than convey deep content:
    - Strong Influence of Da7ee7's Conversational Style like "أقولك يا عزيزي" , "دي يا عزيزي", "هنا يا عزيزي" highlight Da7ee7's rhetorical and engaging storytelling.
    - Phrases from Al Mokhbir Al Eqtisadi like "في نفس الوقت" , " في حلقة النهاردة"

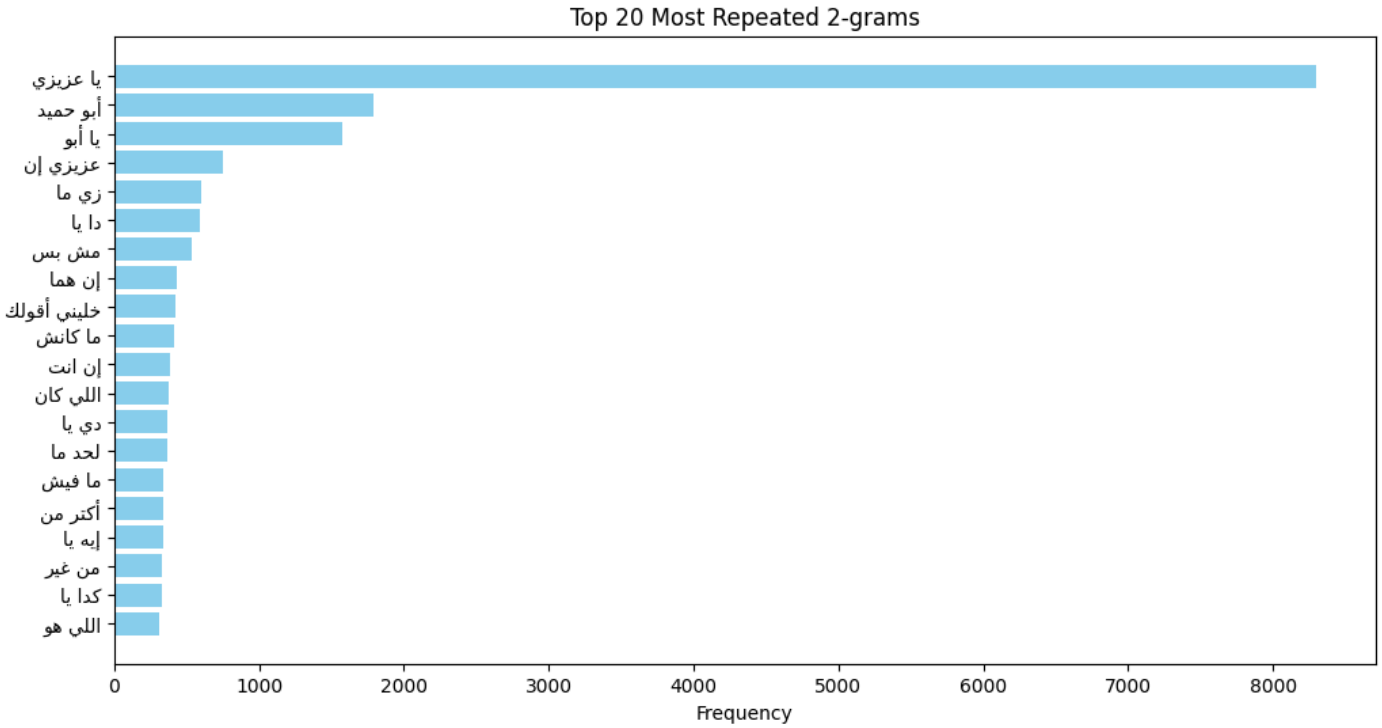
- Visualization of top frequent n-grams per creator:
  - Performed a simple text cleaning technique on the raw transcripts to see the original phrases
    - Change the text to lowercase
    - Remove punctuation
    - Split the text on white space
  - Then, we perform n-gram generation (bigrams and trigrams) and generate the bar charts to show us the most used phrases for each creator and among all creators as a whole
    - **Al Mokhber Al Eqtisadi**
      - Bar Chart generated for bigrams



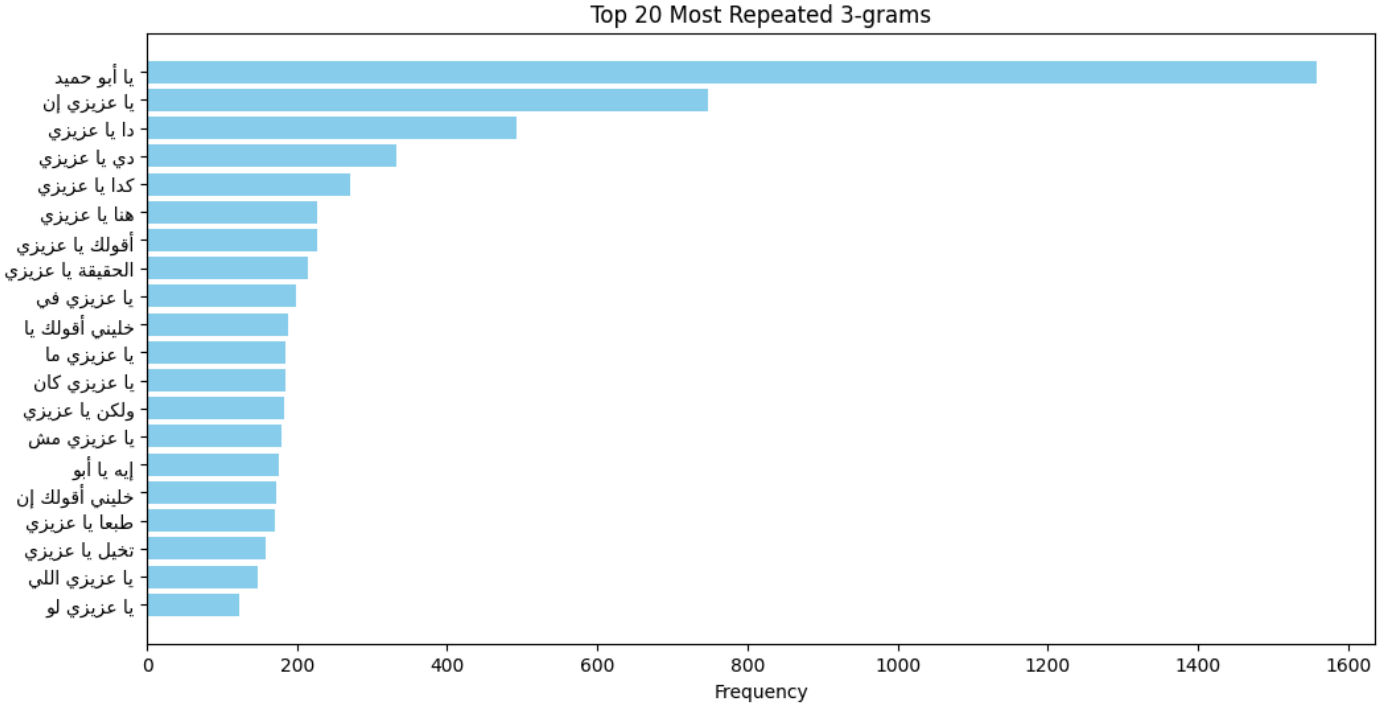
- Bar Chart generated for trigrams



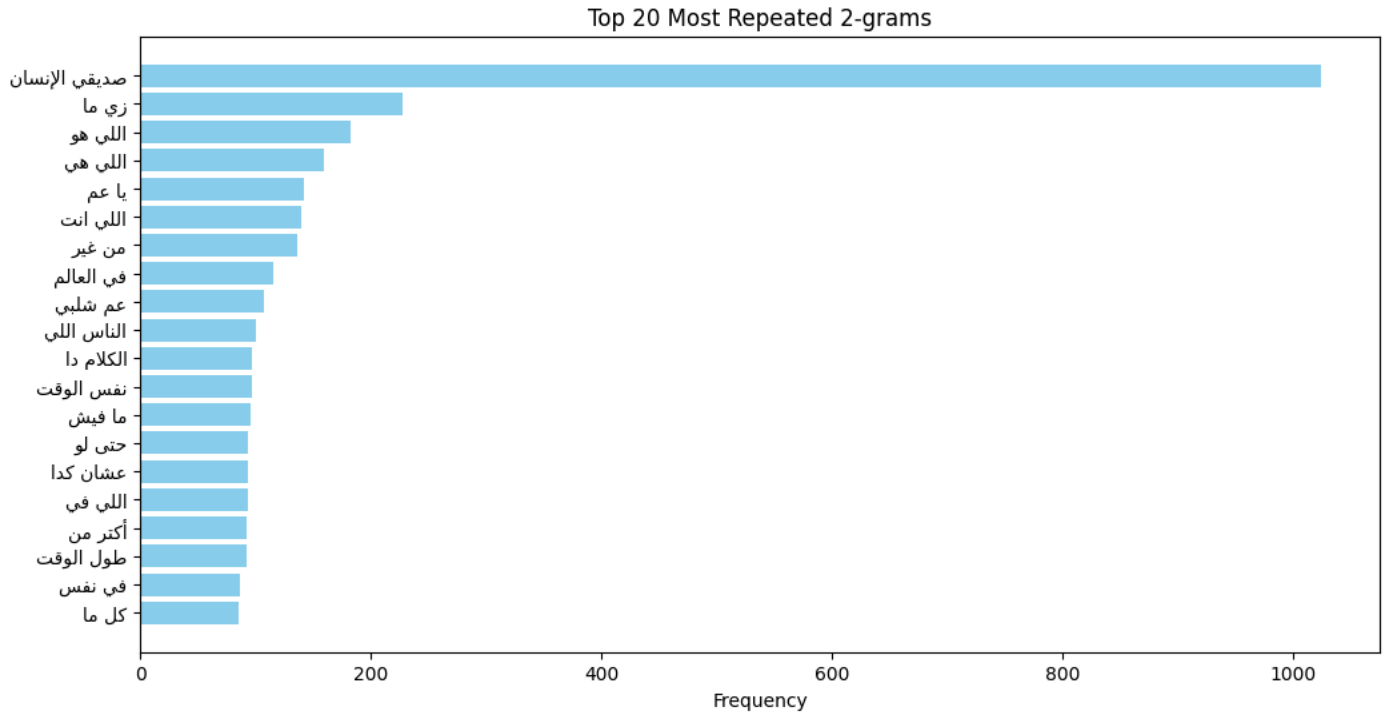
- **Da7ee7**
  - Bar Chart generated for bigrams



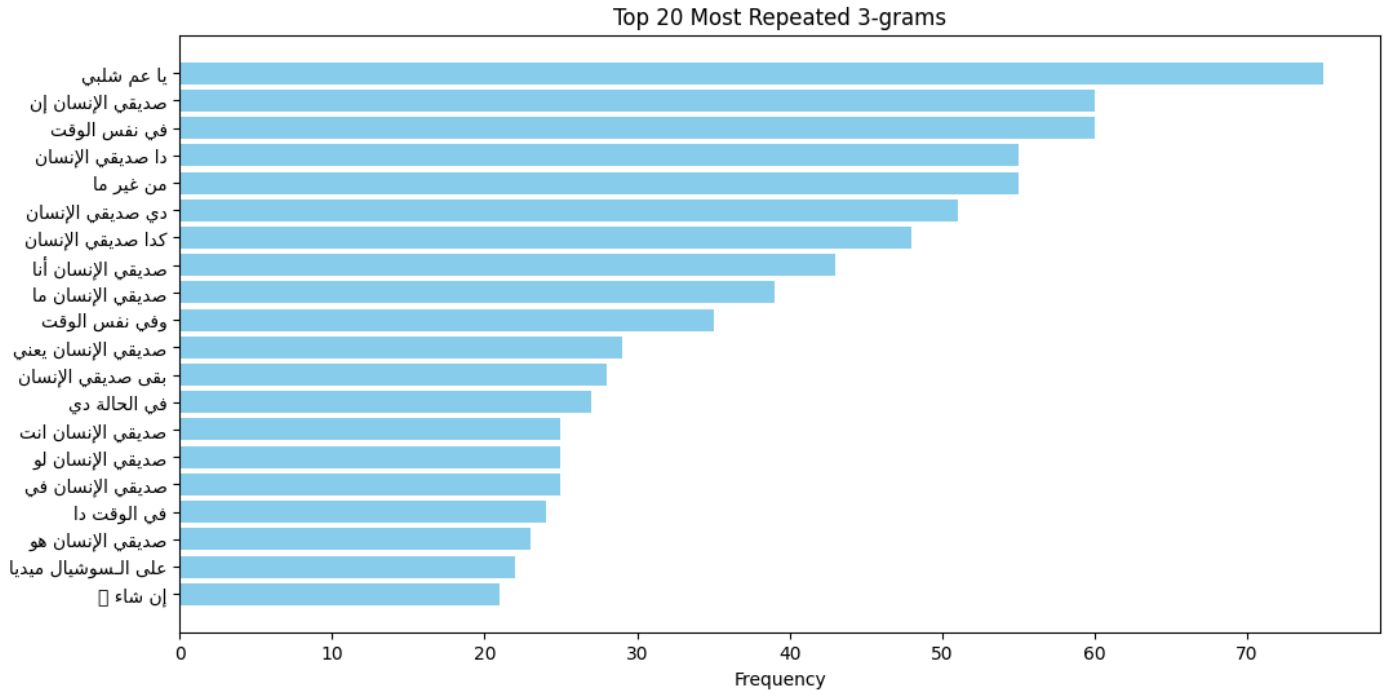
- Bar Chart generated for trigrams



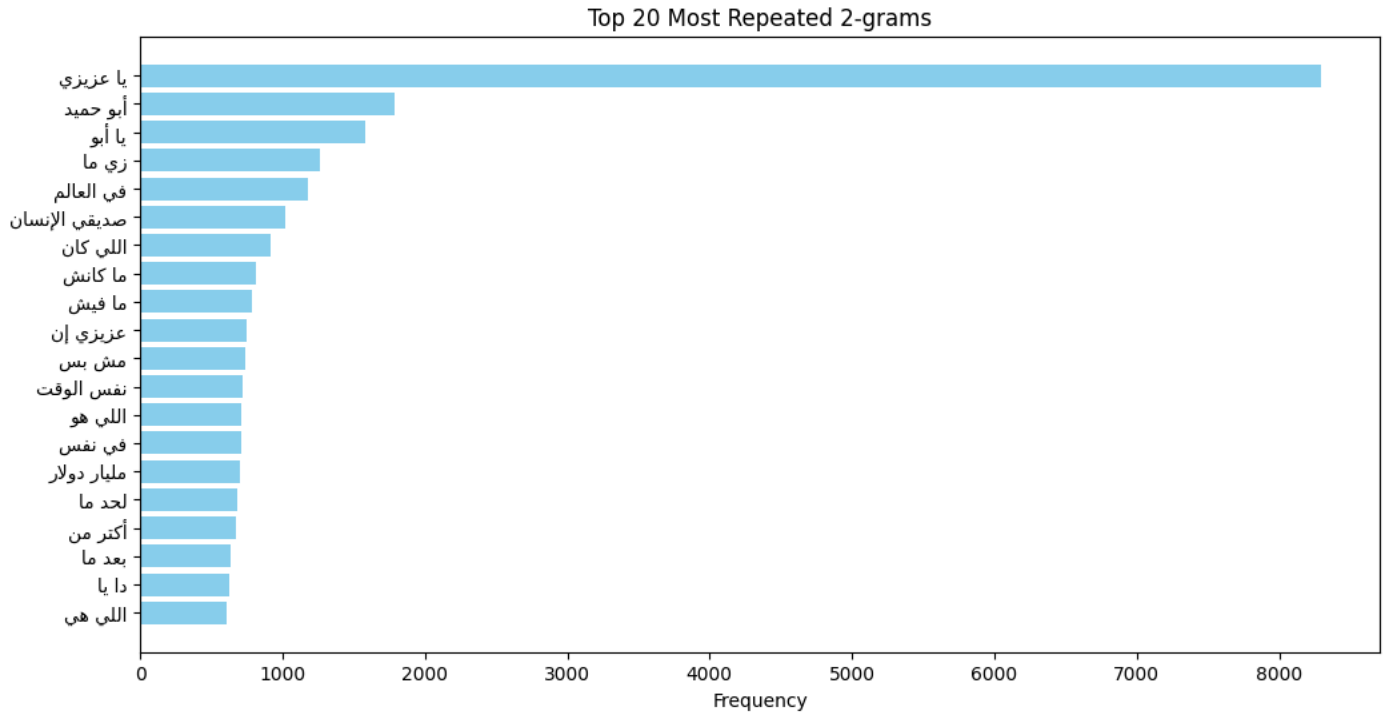
- **Fi Al Hadaraa**
  - Bar Chart generated for bigrams



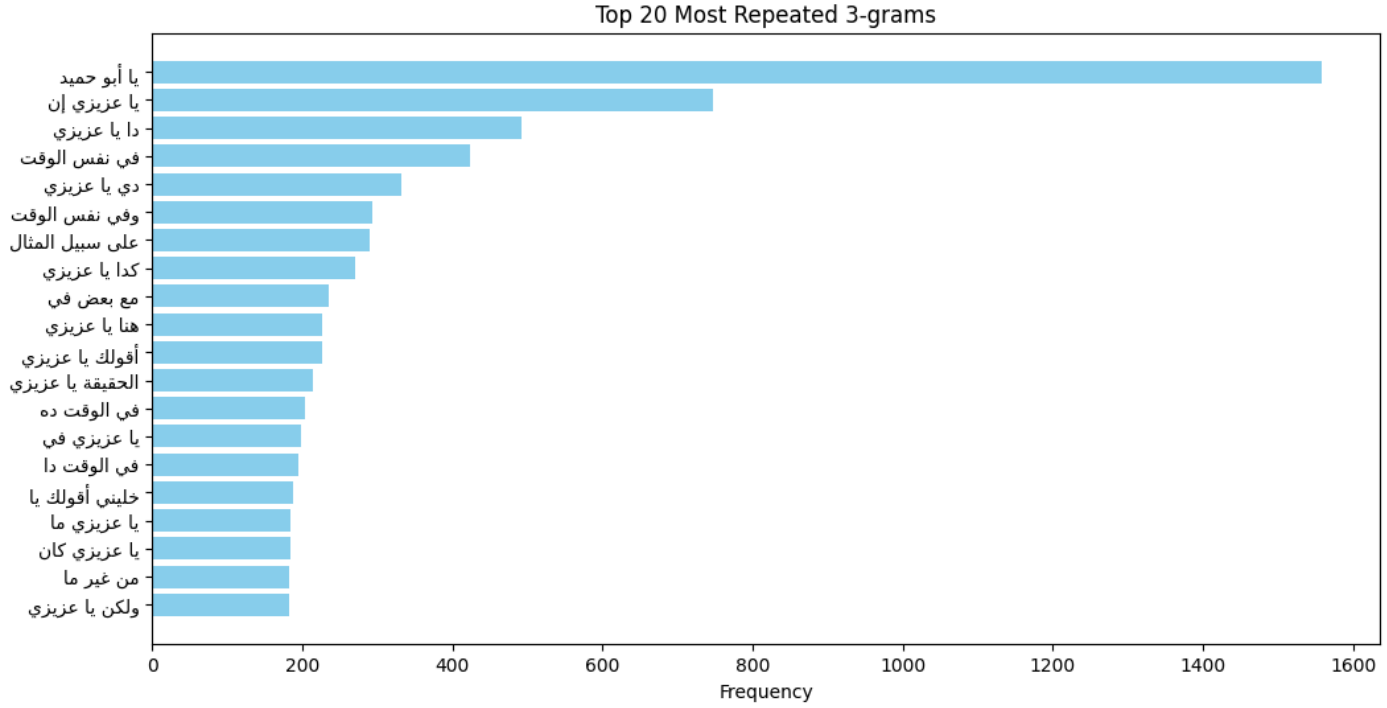
  - Bar Chart generated for trigrams



- **All Channels Together**
  - Bar Chart generated for bigrams



  - Bar Chart generated for trigrams

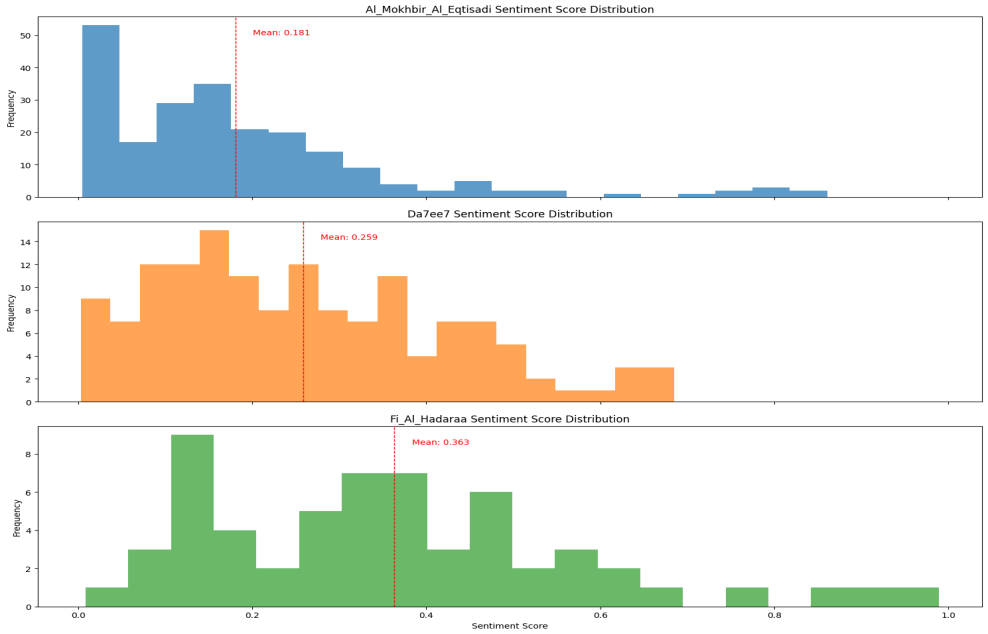


2.4 Sentiment Analysis

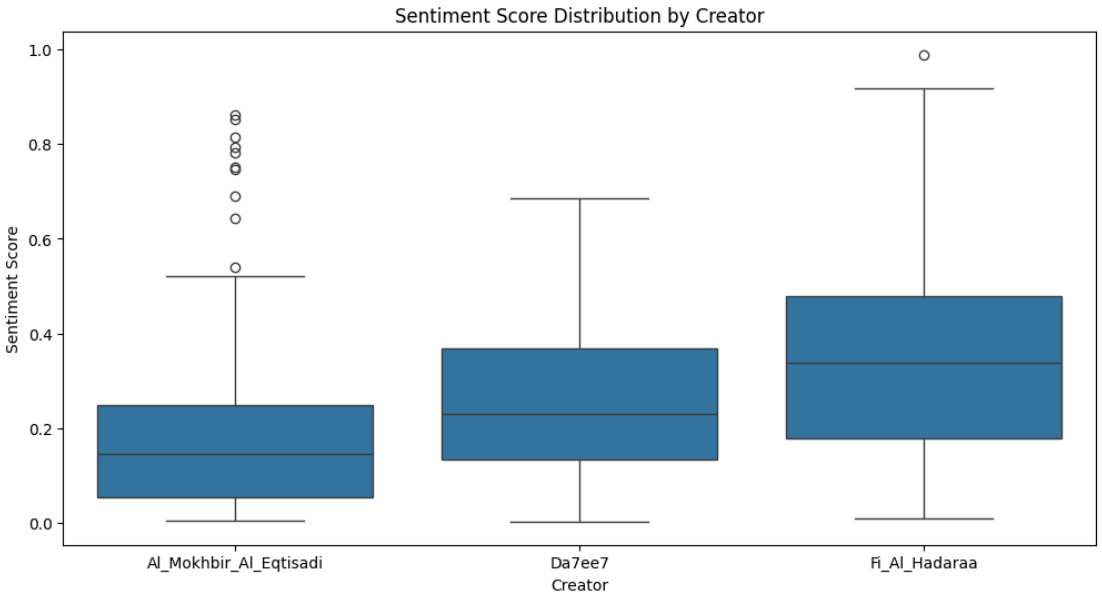
- ❖ The `"Walid-Ahmed/arabic-sentiment-model"` from hugging face has been used for the sentiment analysis in this project.
- ❖ Plotted sentiment distribution histograms per creator.

- **Sentiment Score Distribution Analysis**
  1. **Al Mokhbir Al Eqtisadi (Blue - Top Chart)**
     - The lowest average sentiment (**0.181**), indicating a **more negative tone** overall.
     - Most transcripts are **skewed toward lower sentiment values**, suggesting that the content is **more critical or serious**, likely due to its focus on economic and analytical discussions.
  2. **Da7ee7 (Orange - Middle Chart)**
     - A slightly higher average sentiment (**0.259**), meaning the tone is **less negative** than Al Mokhbir Al Eqtisadi but still leans toward the **neutral-to-negative range**.
     - The sentiment distribution is **more spread out**, implying that some episodes have **a mix of neutral and slightly positive content**, possibly depending on the topic.
  3. **Fi Al Hadaraa (Green - Bottom Chart)**
     - The highest average sentiment (**0.363**), indicating a **less negative and more neutral-to-positive tone**.
     - The wider distribution suggests that while some episodes are **neutral, others lean toward a more optimistic or positive perspective**.

  ❖ **Key Takeaways:**
  - ❖ **Al Mokhbir Al Eqtisadi** has the most **negative sentiment**, likely due to its focus on economic challenges and critical analysis.
  - ❖ **Da7ee7** is **less negative**, with a mix of neutral and slightly positive content, aligning with its educational yet engaging nature.
  - ❖ **Fi Al Hadaraa** is the **least negative and closest to neutral**, possibly reflecting a **more balanced or optimistic storytelling approach** in historical and cultural discussions.



❖ Plot the box plot to compare the min, max, distribution width and outliers



- ❖ **Higher Median Sentiment Score** – Fi Al Hadaraa has the highest median sentiment score among the three creators, suggesting overall more positive content.
- ❖ **Wider Sentiment Distribution** – The range of sentiment scores for Fi Al Hadaraa is broader than the others, with values extending up to 1.0, indicating a mix of both neutral and highly positive sentiments.

❖ **Number of outliers** - Al Mokhbir Al Eqtisadi has much more outliers than the other 2 creators. This can be seen from the white circles on the boxplot graph. After counting the number of outliers using the equations Q1 - 1.5 * IQR & Q3 + 1.5 * IQR, here are the outlier numbers

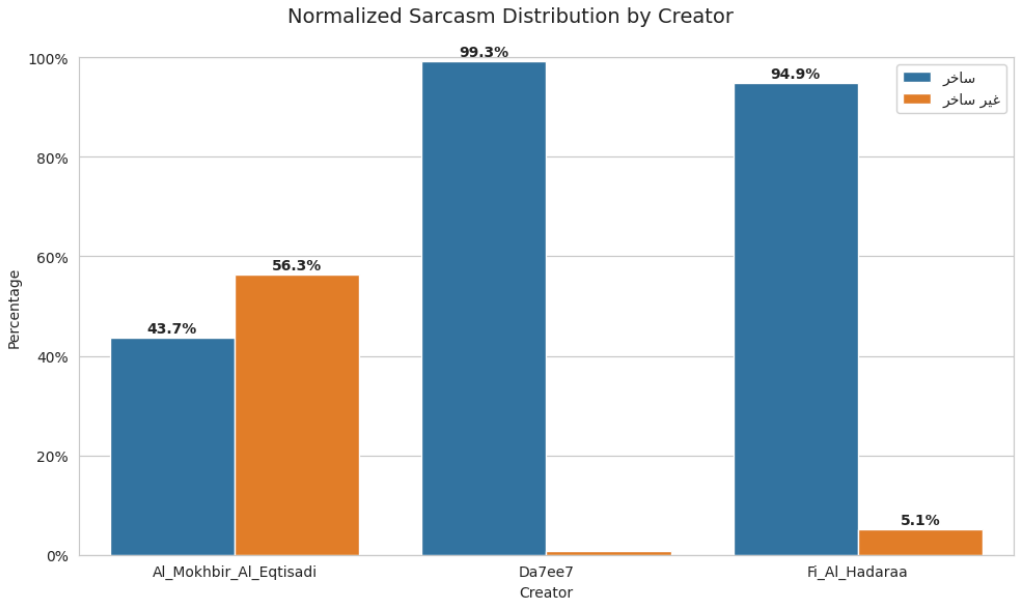
Creator	Number of Outliers
Da7ee7	0
Fi Al Hadaraa	1
Al Mokhbir Al Eqtisadi	10

❖ Conclusions:
- ➢ **Al Mokhbir Al Eqtisadi** has the lowest sentiment distribution, with many lower values and outliers.
- ➢ **Da7ee7** has a slightly higher median than Al Mokhbir Al Eqtisadi but remains below Fi Al Hadaraa.

❖ Limitations:
- ➢ **Weak Model Performance** – The sentiment analysis models used in this study struggle to accurately classify sentiment, frequently predicting neutral or negative sentiments, even when the actual tone of the content is more positive.
- ➢ **Bias Toward Negative Sentiment** – The models tend to **overpredict negativity**, which might not accurately reflect the true emotional impact of the content. This could be due to a lack of nuanced understanding of **Egyptian Arabic expressions**, sarcasm, or cultural context.
- ➢ **Challenges with Egyptian Dialect Arabic** – Many sentiment analysis models, including the ones we used, are not well-trained on **Egyptian Arabic**, leading to frequent misclassifications. Even after testing multiple models, we found that they **struggle with dialect-specific phrasing, slang, and informal writing styles**.

2.5 Sarcasm Analysis

- The `MohamedGalal/arabert-sarcasm-detector` was tried for sentiment analysis.
  - The model wasn't that good, as it predicted most of the text as not sarcastic
  - For example: "طبعا الجو جميل جدًا اليوم، ممطر وبارد وأنا أحب أن أتمشى تحت المطر بدون مظلة!" was labeled as not sarcastic
- Then we tried using gemini 2.0 flash to label each transcript as ساخر أو غير ساخر
- Plot the bar chart to view a normalized barplot for each of the creator where the orange bar represent the غير ساخر and the blue bar represent the ساخر category for each creator



1. Fi Al Hadaraa

- **Highly Sarcastic:** 94.9% of the content is sarcastic, making it one of the most sarcasm-heavy channels.
- **Minimal Serious Content:** Only 5.1% of posts are non-sarcastic.
- **Implication:** Since sarcasm often carries hidden meaning, sentiment analysis models might misinterpret the actual sentiment.

2. Da7ee7

- ❖ **Almost Entirely Sarcastic:** 99.3% of posts are sarcastic, the highest among the three channels.
- ❖ **Virtually No Non-Sarcastic Content:** Just 0.7% of posts are labeled as non-sarcastic.
- ❖ **Implication:** Given the extreme sarcasm, any sentiment analysis must account for this to avoid incorrect conclusions.

3. Al Mokhbir Al Eqtisadi

- ❖ **More Balanced Distribution:** 43.7% sarcastic, 56.3% non-sarcastic.
- ❖ **More Direct Communication:** Compared to the other two, this channel has a higher proportion of straightforward content.
- ❖ **Implication:** Sentiment analysis on this channel is likely to be more accurate since sarcasm is less dominant.

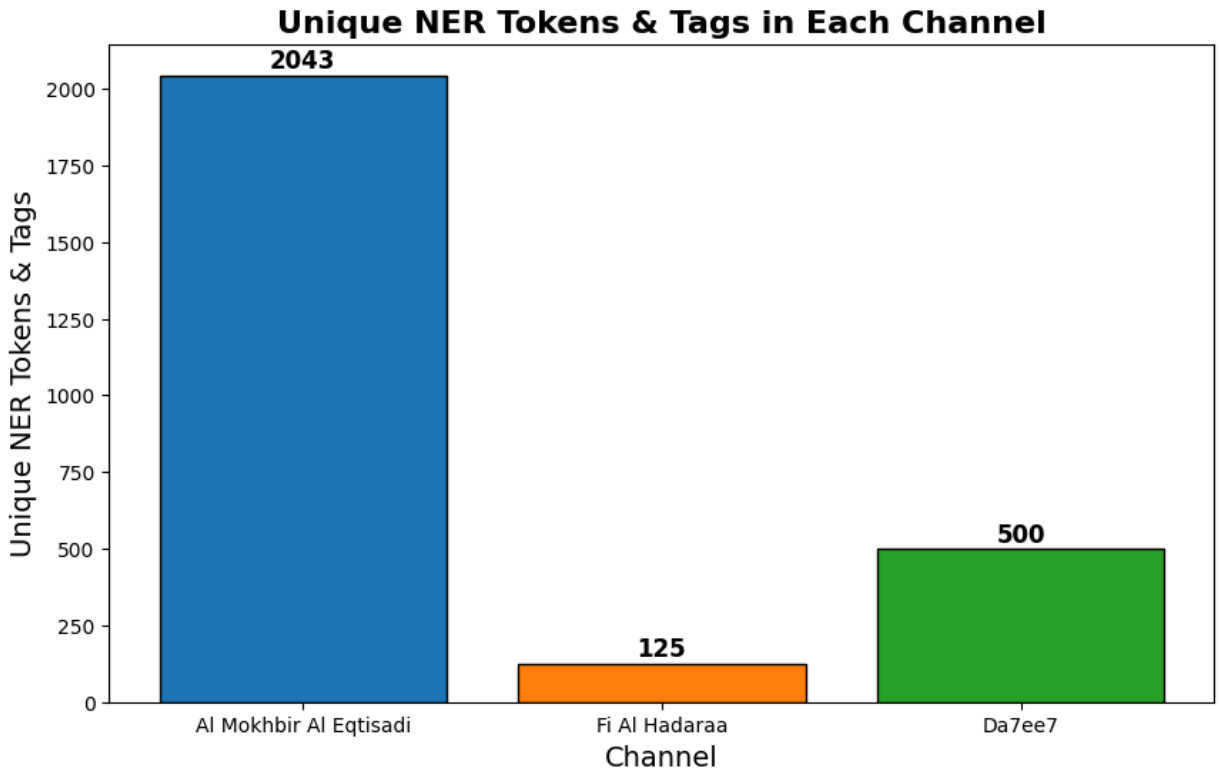
❖ Limitations:

- ➢ No Sarcasm Intensity Score
  - ■ LLMs can tell if something is sarcastic but can't measure how sarcastic it is. Some sarcasm is subtle, while other sarcasm is extreme, and the model treats them the same.
- ➢ Expensive API Calls
  - ■ Running LLMs on large datasets costs money, making it impractical for frequent or large-scale sarcasm detection.
- ➢ Misses Tone and Expressions
  - ■ Sarcasm isn't just about words—it's also in tone of voice and facial expressions. LLMs analyze only text, so they struggle with sarcasm in spoken language.

2.6 Named Entity Recognition (NER)

The objective of this analysis was to apply Named Entity Recognition (NER) to extract named entities from the transcripts and analyze their occurrences across different content creators. Specifically, we aimed to:

- Extract named entities from the transcripts using the `Marefa Arabic Named Entity Recognition` model from the Hugging Face library.
- Visualize the frequency of extracted entities per creator using bar plot :



- Examine the overlap of named entities across different creators using Venn diagrams.
- Analyze the intersection count of named entities using an additional Venn diagram representation.
- By performing this analysis, we aimed to uncover common entities among different creators and assess the effectiveness of the NER model in this context.

Conclusion :

After extracting named entities from the transcripts, we compared the entities identified across different channels to determine commonalities. However, our analysis found no direct intersection between the extracted entities.

Several factors may have contributed to this outcome:

- Some extracted terms were not actual named entities, affecting the accuracy of the results.
- Certain entities were not recognized correctly due to domain differences between the transcripts and the pre-trained NER model.
- While there was an apparent lack of intersection, a closer examination revealed semantic variations in entity representation (e.g., “السعودية” vs. “المملكة العربية السعودية”), suggesting potential overlaps that were not captured by exact matching.

Although our findings indicate no direct intersection, a purely visual inspection suggests that entities across different channels may still be related, albeit with variations in wording. Future improvements could involve entity normalization techniques to enhance accuracy and consistency in entity comparisons.

2.7 TF-IDF & Clustering

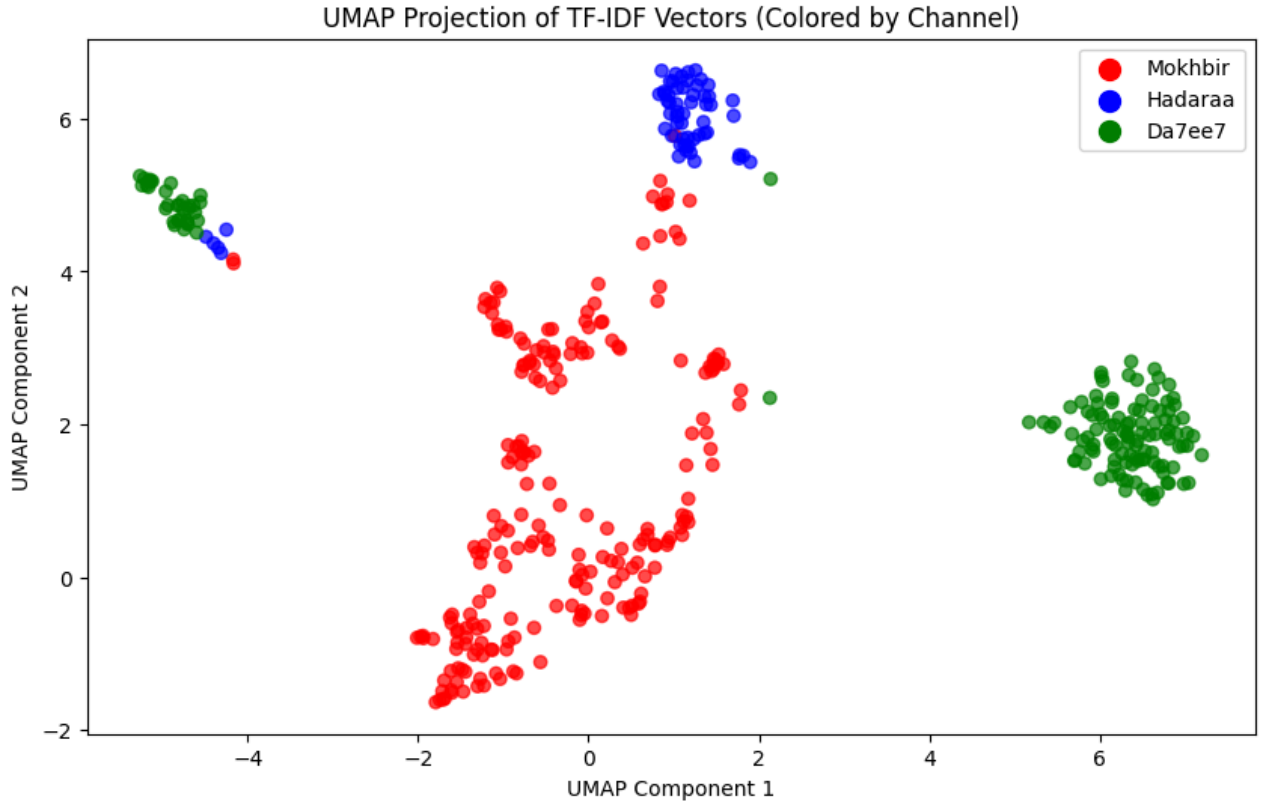
To analyze the relationships between different transcripts and assess their clustering potential, we transformed the transcripts into numerical representations using TF-IDF (Term Frequency-Inverse Document Frequency) vectors. The main objectives of this analysis were:

- Convert transcripts into TF-IDF vectors to represent them as frequency-based feature vectors.
- Reduce the dimensionality of the TF-IDF vectors using UMAP (Uniform Manifold Approximation and Projection) to 2D and 3D for visualization.
- Explore the spatial distribution of the transcripts to identify potential relationships between episodes from different channels.
- Perform K-Means clustering as an initial clustering approach to observe how transcripts group together based on their content.

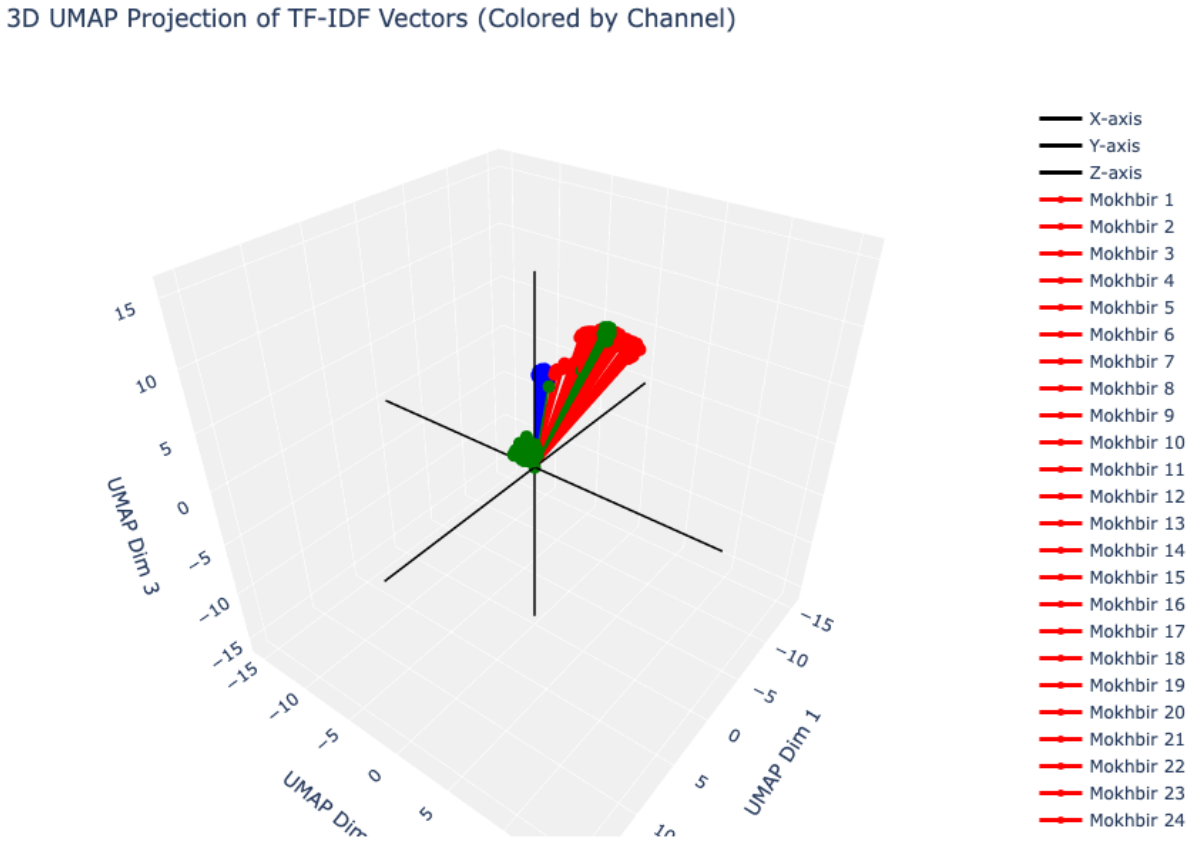
TF-IDF and Visualization

To better understand the structure of the data and its clustering potential, we applied UMAP to reduce the high-dimensional TF-IDF vectors to 2D and 3D representations:

- 2D Visualization: This provided an overview of how the transcripts are positioned relative to each other, helping to identify initial patterns and possible clusters.



- 3D Visualization: By increasing the dimensionality, we aimed to capture more complex relationships between the transcripts, allowing for a more detailed exploration of their distribution.



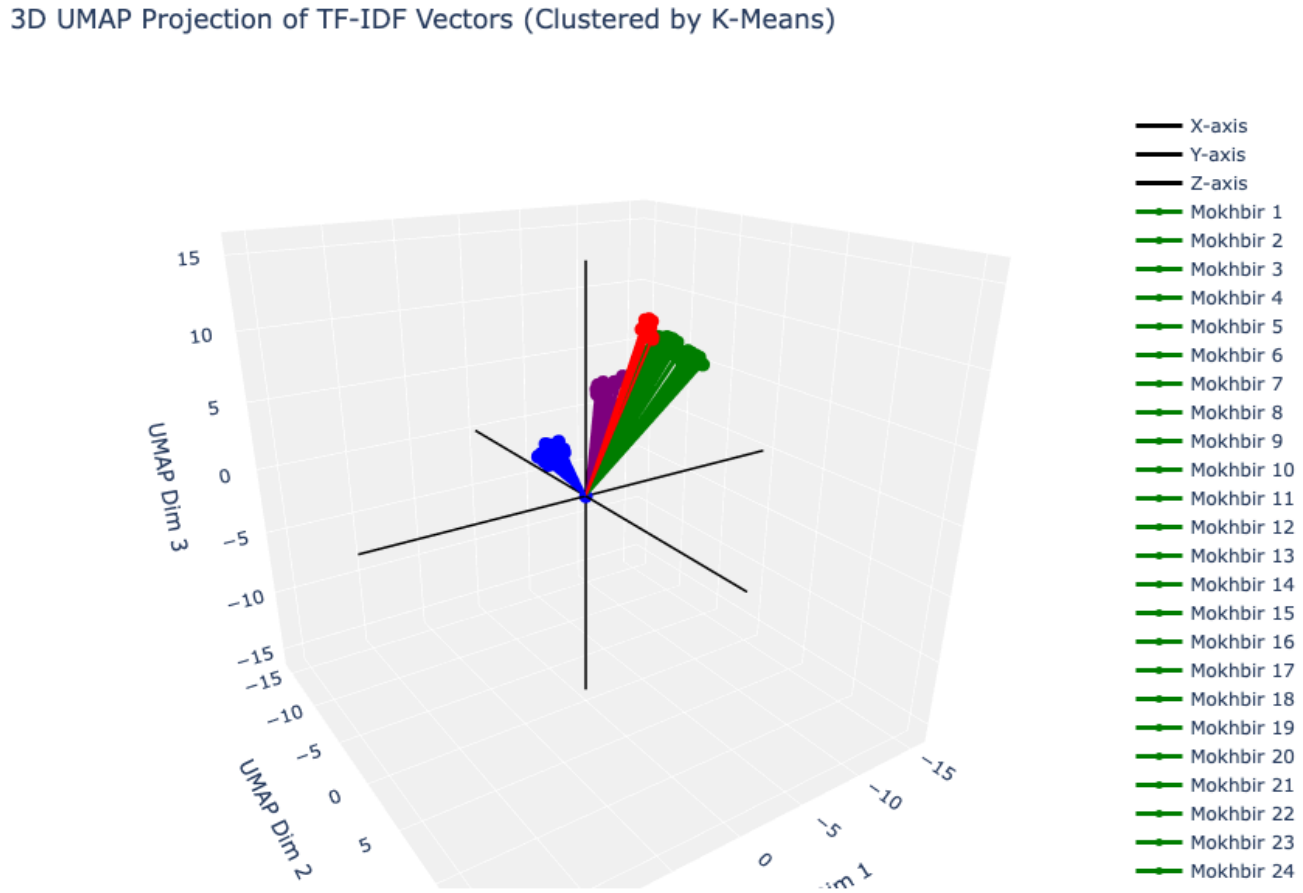
- These visualizations served as an initial step to demonstrate that the data exhibits meaningful relationships and could potentially be clustered.

Clustering with K-Means

Following the visual analysis, we applied K-Means clustering as an initial method to group the transcripts based on their textual similarities. The goal was to:

- Identify distinct clusters of episodes and analyze whether they align with different channels.

- Evaluate the effectiveness of K-Means in distinguishing content similarities and differences.



The results of this clustering process provided insight into the potential for more advanced clustering techniques and further analysis of content relationships across different creators.

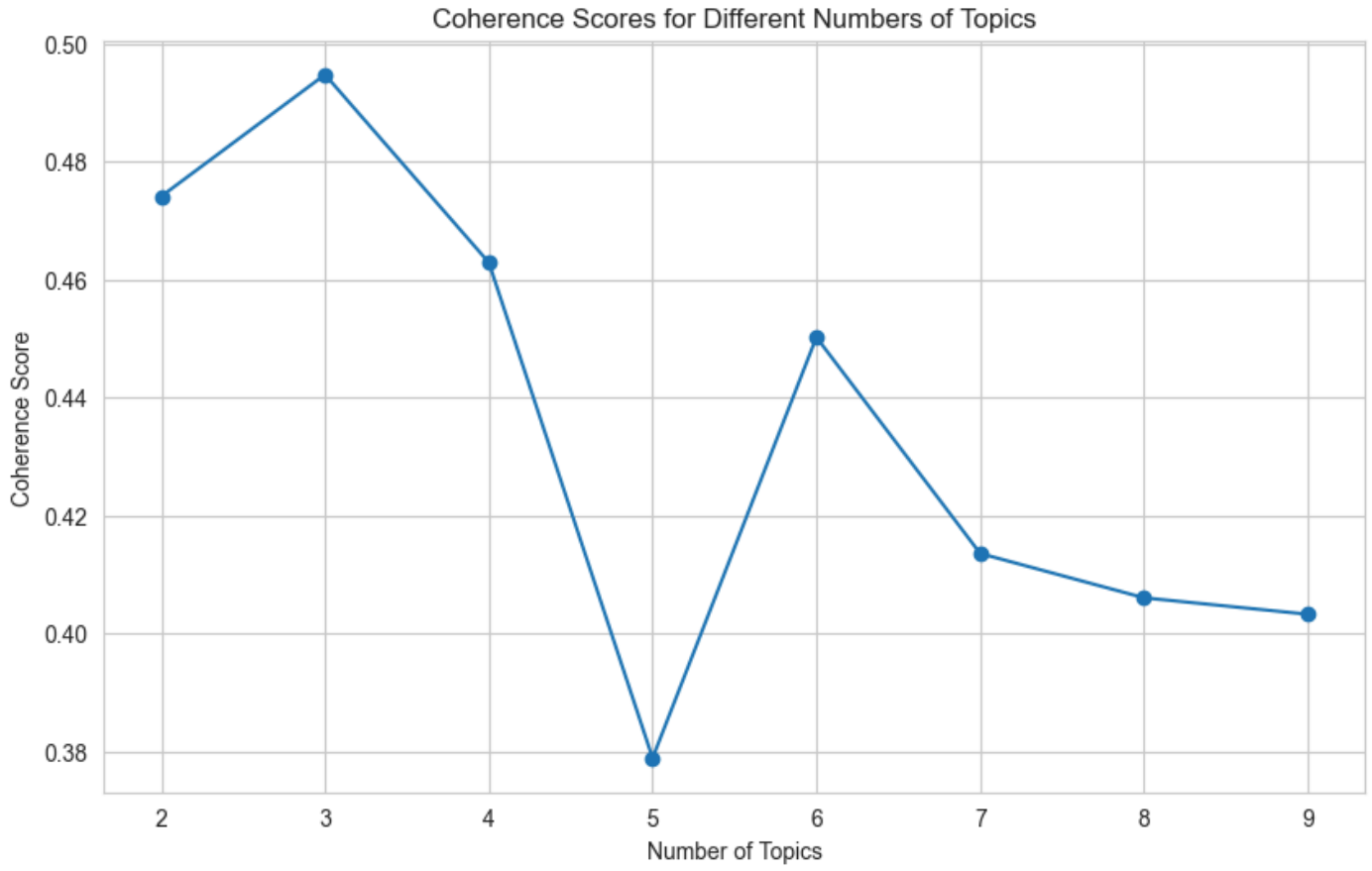
2.8 Topic Modeling

Objective

Topic modeling was conducted to uncover hidden themes in the dataset and analyze the primary discussion topics. We used Latent Dirichlet Allocation (LDA) to extract topics and evaluate their coherence to determine the optimal number of topics.

Coherence Score Analysis

The coherence score graph shows how well topics are grouped. A higher coherence score indicates more meaningful and interpretable topics. Based on the graph, the highest coherence score is achieved at **3 topics**, suggesting that three distinct themes best represent the dataset's structure.



Discovered Topics

Below are the extracted topics along with their most representative words:

- **Topic 0**: الاسرائيلي, الإسرائيليين, الامريكيه, العالم, الامريكي
- **Topic 1**: الإنسان, الناس, العالم, تاني, طبع
- **Topic 2**: العالم, أمريكا, الأمريكية, شركة, الصين

Each topic suggests a distinct theme:

- **Topic 0: Geopolitics & International Affairs**
  - Words related to Israel, the U.S., and global politics suggest discussions on international conflicts, diplomatic relations, and geopolitical strategies.
- **Topic 1: Society & Humanity**
  - Words like الإنسان and الناس indicate a general discussion on human nature, societal issues, and philosophical reflections on life.
- **Topic 2: Global Economy & Trade**
  - The presence of الصين ,أمريكا, and شركة suggests discussions about economic policies, corporate influence, and international trade.

Influence of "Al Mokhbir Al Eqtisadi" on Topics

One key insight is that the dominance of **Al Mokhbir Al Eqtisadi** in the dataset has shaped the extracted topics. His content mainly covers **economic trends and global politics**, which aligns with the identified themes.

This influence explains why:

- The topics focus heavily on **geopolitics and economy.**
- Countries like **the U.S. and China** appear frequently.

If the dataset included a wider range of sources, we might observe different topics. A potential improvement for future analysis is **balancing the dataset** by incorporating diverse speakers with varying perspectives.

Conclusion

The topic modeling results highlight that the dataset primarily discusses **global politics and economic affairs**, largely due to the influence of **Al Mokhbir Al Eqtisadi**. Also, another topic discussing general societal issues is found this could be of the influence of **Da7ee7** & **Fi Al Hadaraa.**

---

3. Preprocessing Steps

3.1 Data Cleaning

To ensure data quality, the following preprocessing steps were applied:

- Removed timestamps (e.g., [4.25]).
- Removed tags (e.g., [موسيقي]).
- Eliminated ellipses.
- Separated punctuation & quotations from words.
- Removed all punctuation.
- Removed English characters.
- Removed diacritics (Tashkeel).
- Removed elongation (Tatweel).
- Removed stopwords:
    - Created a comprehensive list of stopwords
        - NLTK arabic stopwords
        - Stopwords from this github repo: (https://github.com/mohataher/arabic-stop-words)
        - Stopwords we added by hand

Why didn't we use Lemmatization and Stemming?

Lemmatization and stemming aim to reduce words to their base forms, but in **Arabic**, they often cause **loss of meaning** due to the language's **complex structure**. So some of the words tend to lose their true meaning or even change their meaning entirely. Moreover, lots of the words in our dataset are written in **عامية** (colloquial Arabic) rather than **فصحى.** Since **عامية** often lacks standardized spelling and grammar rules, applying lemmatization or stemming can result in significant errors, causing words to either lose their true meaning or change entirely.

- **Example:** "عمان وقطر والبحرين والسعوديه" → "عمن قطر بحر سعد"
    - The system incorrectly stems **"بحرين"** to **"بحر"**, mistaking the country name for the generic word "sea." Additionally, **"عمان"** was incorrectly reduced to **"عمن"**, and **"السعودية"** was altered to **"سعد"**, completely changing the intended meaning. (nltk's ISRIStemmer)
- **Example:** "عمان وقطر والبحرين والسعوديه" → "عمان قطر بحرين سعودي"
    - The stemming process alters **"سعودي"** to **"سعوديه"**, potentially changing the intended reference. It becomes unclear whether the text refers to a

nationality, a country, or a general Saudi-related entity, reducing precision in text analysis. (FarasaStemmer).

---

4. Obtaining Labels

4.1 Using TF-IDF

- **Objective:** Instead of manually labeling the data, we use TF-IDF scores to identify the most relevant words in each transcript. These top-ranked words are then used as labels for multi-label classification.
- **Process:**
  - Converting the cleaned transcripts (using the data preprocessing technique mentioned before) into a numerical format where each document is represented as a vector of word importance scores
  - TF-IDF scores highlight words that appear frequently in a transcript but are not common across all transcripts, making them good candidates for labels.
  - Since each transcript may cover multiple topics, we extract the top **three** words with the highest TF-IDF scores as potential labels.
    - Converting each transcript's TF-IDF row into an array
    - Sorting the values to find the highest TF-IDF scores.
    - Retrieving the corresponding words from the vocabulary.
    - These words are assumed to represent the main topics of the transcript.
  - The extracted words serve as automatically generated labels for the transcript
  - The TF-IDF-based label extraction generated suboptimal labels, as many of the top-ranked words were too generic, irrelevant, or lacked contextual meaning so further refinement is needed through techniques like **n-grams, topic modeling, or threshold-based filtering** to ensure the extracted labels better reflect the main themes of each transcript.

4.2 Using Gemini API

To enhance the labeling process for multi-label classification, the **Gemini API** is used to generate contextually relevant Arabic labels for each transcript. This method ensures that the extracted labels align with the actual topics and themes present in the transcripts, overcoming the limitations of purely statistical approaches like TF-IDF.

- Instead of processing each transcript separately, transcripts are grouped into batches of 5 to minimize API requests and improve efficiency.
- The model is provided with clear instructions to extract short, relevant labels that describe the main topics in each transcript
- Previously identified labels are retained and reused, ensuring consistency across batches while allowing the system to discover new topics dynamically.
- Despite using the Gemini API for label extraction, a large number of unique labels were still generated, which suggests that the prompt may need further refinement to ensure more **consistent** label generation. However, this could also be attributed to the nature of the channels themselves, as they cover a wide range of topics. For example, El Da7ee7 explores a different subject in each episode, such as discussing food in one episode and the mafia in the next, leading to naturally diverse label outputs.

---

5. Conclusion

In this milestone, we focused on **data preprocessing and exploratory data analysis (EDA)** to prepare a robust dataset for downstream **multi-label classification and retrieval tasks**. Our work involved **cleaning and structuring transcripts**, analyzing linguistic patterns, and identifying key features that influence model performance.

Key takeaways from our analysis include:

1. **Dataset Composition** – The transcripts vary significantly in length, vocabulary, and content style across different creators, affecting both retrieval and classification challenges.

2. **Non-Arabic Word Usage** – Certain creators, such as **Da7ee7**, incorporate a substantial number of English words, particularly in scientific and technical discussions, which impacts preprocessing decisions.
3. **Sentiment and Sarcasm Analysis** – Sentiment trends vary by content type, with **Al Mokhbir Al Eqtisadi** exhibiting the most neutral-to-negative tone, while **Da7ee7 and Fi Al Hadaraa** contain heavy sarcasm, making sentiment classification more complex.
4. **Named Entity Recognition (NER) Limitations** – Extracted entities lacked a direct intersection, possibly due to dataset variability and the inability of pre-trained NER models to handle dialectal variations.
5. **TF-IDF & Topic Modeling** – Thematic clustering and topic modeling revealed that economic and geopolitical discussions dominated the dataset, largely influenced by **Al Mokhbir Al Eqtisadi**, while other creators contributed content related to science, history, and social topics.
6. **Preprocessing Challenges** – The use of **عامية (colloquial Arabic)** alongside **فصحى (Modern Standard Arabic)** posed difficulties in **lemmatization and stemming**, as informal words lacked standardized forms.

Through **TF-IDF and Gemini API-based labeling**, we explored methods to **automate tag extraction** for multi-label classification. However, improvements are still needed to refine label consistency and semantic understanding.

Moving forward, the insights gained from **EDA and preprocessing** will directly inform model training, retrieval system design, and overall performance improvements in subsequent milestones. Future work will focus on **optimizing label extraction, refining clustering techniques, and improving dialectal handling in NLP models** to ensure accurate and meaningful results.

---

6. References

List all references, including dataset sources, model documentation, and external tools used.

Data Manipulation and Analysis

- `pandas` - For data manipulation and analysis
- `numpy` - For numerical operations

Visualization

- `matplotlib.pyplot` - For creating plots and charts
- `seaborn` - For statistical data visualization
- `wordcloud` - For generating word clouds

Natural Language Processing

- `nltk` - Natural Language Toolkit for NLP tasks
- `pyarabic.araby` - For Arabic text processing
- `arabic_reshaper` - For reshaping Arabic text for display
- `bidi.algorithm` - For handling bidirectional text (Arabic)
- `gensim` - For topic modeling

Machine Learning

- `umap` - For dimensionality reduction and visualization

Large Language Models & APIs

- `langchain` - For using and integrating the Gemini API
- `Gemini Flash 2.0` - Used for generating multiple labels per transcript

Youtube's Dataset

- `Created by Hamza Gehad`