



Milestone 1 Report

CSEN1076: Natural Language Processing and Information Retrieval

Team Members:

- Mazen Soliman (52-2735)
- Mohamed Shamekh (52-0989)

Supervised By: Mayar Osama

1 Introduction

The purpose of this milestone was to put us in a situation where we have to practically implement a NLP preprocessing pipeline for Arabic text that makes the raw data suitable for performing a number of analysis tasks. Over the course of this report, we will provide an overview of the analysis tasks that we performed and what the purpose of the said analysis task might be. Then, we will cover some of the analysis tasks we performed or attempted to perform. We will discuss if the results of said analysis tasks give us any useful insights, or do we need different circumstances to get more use out of these tasks, or if the results should be discarded altogether.

2 Overview of Our Analysis

The Arabic language is fascinating when studied from a linguistic perspective. It is very different from most other languages and especially English. Therefore, when preprocessing Arabic text, we have to take these differences into account and act accordingly. This can manifest in some preprocessing steps that are not performed in other languages such as dediacritization, as well as taking different dialects into account, especially if we are dealing with transcribed spoken language. In this section, we will explore these steps in deeper detail.

Goal



The goal of this analysis is to preprocess Arabic language transcript dataset of famous YouTuber ElDa7ee7 to be able to perform various analysis or regression tasks such as visualizing the most common expressions used by the YouTuber (through N-Grams) or regression tasks such as predicting the number of views or likes a video will get based on the transcript.

Our analysis tasks consists of two main stages which are:

1. Preprocessing
2. Visualization


2.1 Preprocessing & Some Analysis Tasks

In the preprocessing stage, we tackled multiple approaches to analyse the corpus and find key features that could be beneficial while some of them did not work as we hope yet they provides a good overall analysis.

Cleaning the text is one of the most crucial step in our pipeline preprocessing, cleaning consists of multiple phases. First, it envolves removing arabic stopwords that does not add meaning we used popular stopwords dataset from NLTK , TASHAPHYNE  and also defined some others.

Also, we consider handling some other noises such as removing hashtags, punctuations, emojis which does not exist in our corpus but helped us with finetuning some of the models we had and could help us in latter milestones.

2.1.1 Tidying Up The Text

The first thing we need to do is make sure that our text is ready to be pre-processed. This involved doing several things at first, including removing the timestamps of the YouTube transcripts we had collected. We removed them because they weren't of any use for any analysis task we were considering. In addition, we also wanted to correct spelling inconsistencies, and we planned to do that using the Ghalatawi  library. However, the library did not give us the results we expected.

2.1.2 Splitting the Text into Sentences

Now, we can finally work on the raw text in our dataset. The first thing we wanted to do was separate this text into sentences. This step can be useful if we want to perform sentiment analysis on portions of the episode text, as performing it on an entire episode can yield inaccurate results. The thing is that YouTube transcripts do not really have proper sentence separations. Thus, in order to separate the sentences, we first use an SBERT model to give embeddings to the sentences, which are separated based on their semantic meanings. The cosine similarity was then calculated using the scikit-learn [🔗](#) library for these embeddings. After identifying breakpoints (based on a threshold of 0.5), we split the text into segments and remove any segments containing only spaces.

2.1.3 Translating Arabic to English

This approach aimed to convert Arabic text into English, process both languages using different models, and compare the results. However, it was unsuccessful because accurately translating Egyptian Arabic while preserving its meaning proved challenging. Furthermore, the translator we used did not consistently produce English translations for all Arabic statements.

2.1.4 Sentiment Analysis

Sentiment analysis is the process of determining whether a text expresses a positive, negative, or neutral sentiment.

We used the Hugging Face transformer model bert-base-arabic-camel-msa-sentiment, trained by Camel Lab [🔗](#), to classify text into these categories.

Although the model performed well on sample sentences, it produced a large number of negative sentiment classifications when tested on a corpus of Da7ee7 content as shown in

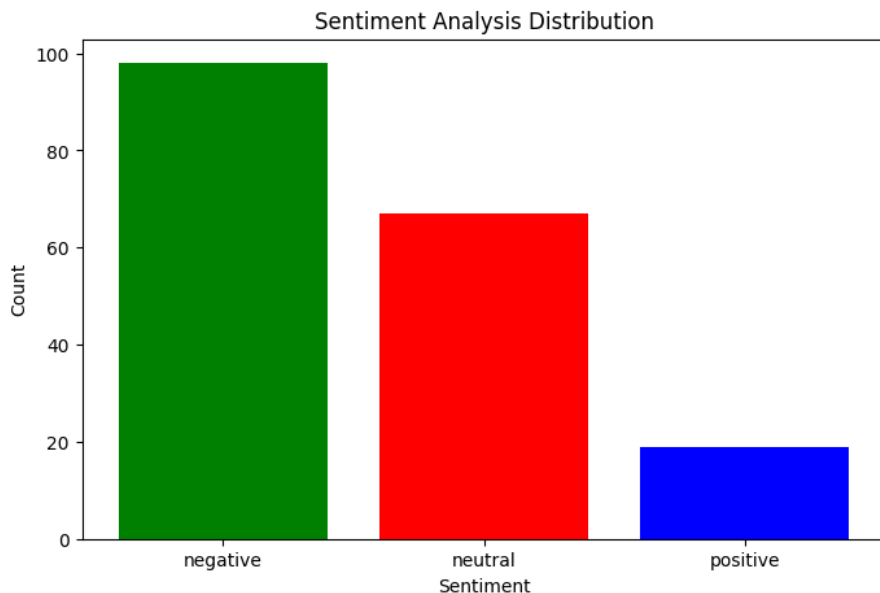


Figure 1: Sentiment Analysis of Da7ee7 corpus

Upon investigation, we found that Egyptian Arabic often contains words with negative connotations that do not necessarily imply a negative sentiment. However, the model tended to classify such sentences as negative due to the presence of these words.

2.1.5 Text Classification

Text classification is the process of determining the topic of a given text, such as culture, finance, medicine, politics, religion, sports or technology.

We initially used bert-base-arabertv2, a model trained by Aubmind Lab [↗](#), fine-tuned on a dataset available on Kaggle [↗](#).

The model performed well when tested on sample sentences, providing highly accurate classifications. However, when applied to our actual dataset, where each document was divided into sentences, it frequently classified sentences as religious, even when the overall topic was not related to religion. This was likely due to the frequent use of religious terms in everyday language.

Filtering out sentences classified as religious improved the model's topic predictions. However, due to the small number of remaining classifications, some gaps remained in the overall categorization.

2.1.6 Sarcasm Detection

Given the significant use of sarcasm in Da7ee7 content, we explored whether sarcasm detection was possible and if there was a correlation between sarcasm and video popularity (e.g., viral quotes from Da7ee7 episodes).

We fine-tuned bert-base-arabertv2 on a labeled data set from Twitter. However, the model's predictions remained inaccurate, primarily due to the small training dataset, which limited its ability to generalize. Additionally, we found that sarcasm varies across different Arabic dialects, making it even more challenging for the model to generalize effectively.

2.1.7 Summarizing Text

We explored the use of the mbert2mbert-arabic-text-summarization model from Hugging Face for text summarization, aiming to enhance video classification by removing unnecessary words. While the model performed well on small Arabic corpora, it struggled with larger texts, producing hallucinated summaries, especially when applied to transcribed Da7ee7 videos.

2.2 Visualization

2.2.1 Basic statistics

We explored the dataset starting from performing basic statistics such as determining the word count of a video, the unique word count, the average word count per video, number of sentences and average word length. We added this data to the dataframe that contains other relevant information, such as the name of the episode, its text, and the information present in the annotations. We also performed other analysis tasks such as calculating tf-idf which, in fairness, we were not going to use for any of our analysis tasks.

2.2.2 Word Cloud

The previous statistics were just numbers present in a dataframe, and we wanted to have something that is more visual. Thus, we thought of creating a word cloud of the YouTube transcripts of ElDa7ee7. We did this using the WordCloud [🔗](#) Python library. But, to pass our text to it we first had to give it to the arabic-reshaper [🔗](#) and python-bidi [🔗](#) libraries we

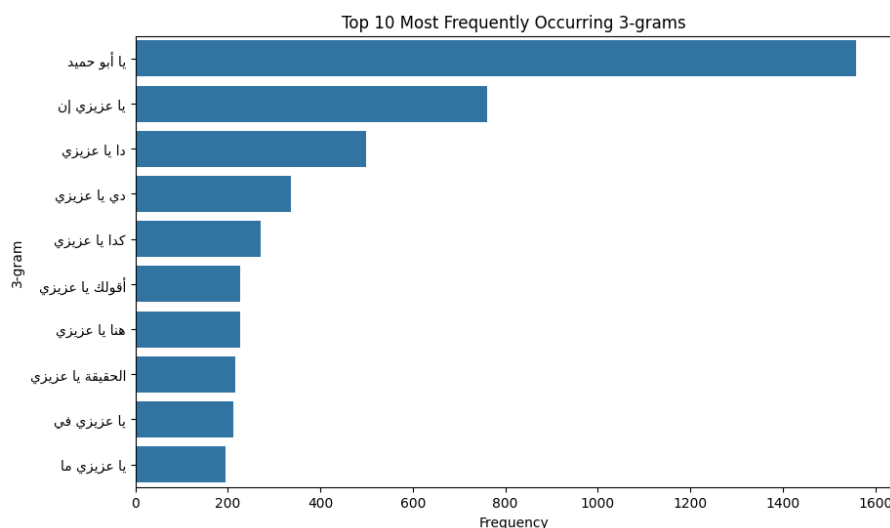
mentioned previously to make sure the WordCloud library will work with our Arabic text. The output can be seen below.



Because we gave the word cloud text without removing the stopwords, it is a bit obvious why **في** is very prominent in the word cloud. But, and also unsurprisingly, the word **عزيزي** is also very prominent in the word cloud, and this is because, as all ElDa7ee7 fans know, he uses this word quite frequently over the course of a single episode.

2.2.3 Most Frequent N-grams

Another interesting way in which we can analyze ElDa7ee7's speech patterns is by determining which are the most frequent N-grams in the transcripts of his YouTube videos. So, we explore the most repeated 3-words (3-gram) which, unsurprisingly, were يا أبو حيد and يا عزيزي, the visualization of the most frequent 3-grams can be seen in the graph below.



But the general analysis performed on the dataset was not very helpful for our goal due to the following reasons that even after cleaning the text Da7ee7 uses lots of repetitive words that does not indeed need to be stopwords but could act as outlier as they do not add a new meaning to the video or help in the analysis task. For instance, the word عزيزي does not really provide additional meaning to the text, but is used as a filler and as a way to move from one topic to another.

2.2.5 Videos Similarity

We used cosine similarity to explore the potential overlap among Da7ee7 videos and identify any repeated content. As shown in the similarity matrix (Figure 4), most video pairs exhibited low similarity. However, there were a few noteworthy exceptions—such as **الحب عن بعد** and **غدر الصحاب**—that showed moderate overlap.

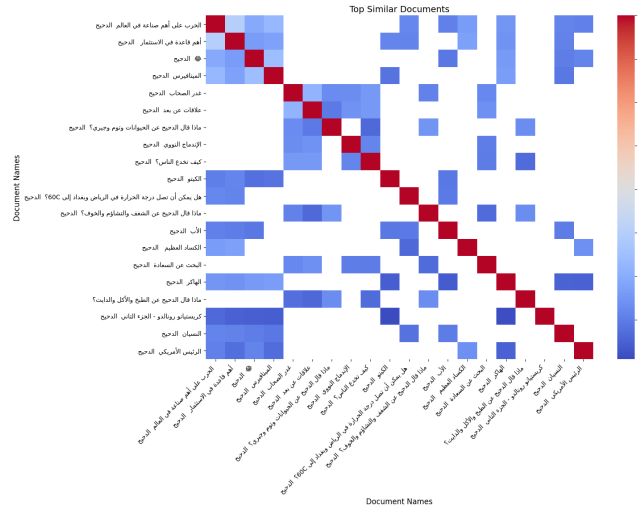


Figure 4: A cosine similarity between Da7ee7 videos

This matrix visualizes how often certain topics or phrases might recur across different videos. The diagonal (in red) represents each video’s self-similarity, while off-diagonal cells reflect varying degrees of content overlap. The results suggest that, overall, Da7ee7 covers a broad range of subjects with limited repetition. Nonetheless, these occasional similarities could indicate shared themes, references, or narratives in certain episodes.

2.2.6 Relation between Video Likes and Other Factors

1. **Category:** We analyzed which video categories tend to receive the most likes.

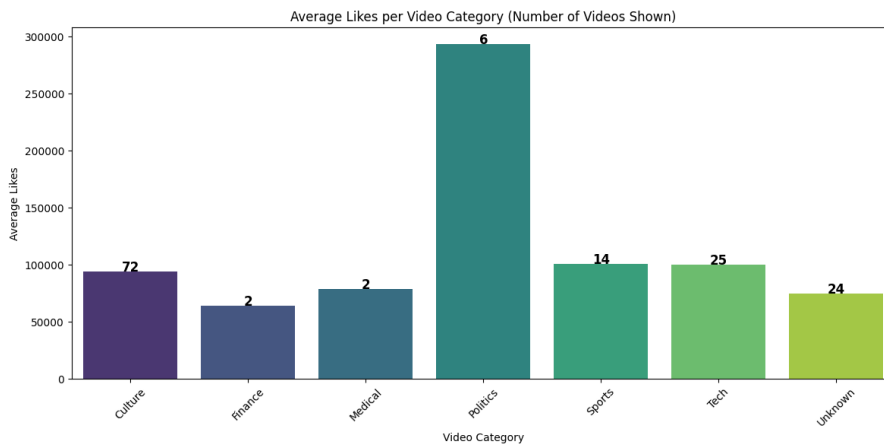


Figure 5: Average Number of Likes by Video Category

Observations:

- **Politics (6 videos):** Despite the smaller sample size, the Politics category stands out with an average of approximately 300,000 likes. This may indicate strong audience engagement driven by the topical or controversial nature of political content.
 - **Culture (72 videos) and Tech (25 videos):**
 - *Culture* maintains consistently high engagement, as indicated by its larger sample size.
 - *Tech* videos also attract a respectable number of likes, likely fueled by popular or trending technology discussions.
 - **Smaller Categories (Finance & Medical, 2 videos each):**
 - These categories show comparatively lower average likes; however, the very small sample size makes it difficult to draw definitive conclusions.
 - **Sports (14 videos) and Unknown (24 videos):**
 - The Sports category, with a moderate sample size, achieves a respectable like count that may vary depending on the popularity of specific sports or events.
2. **Length:** We also explored whether the length of a video's description correlates with the number of likes.

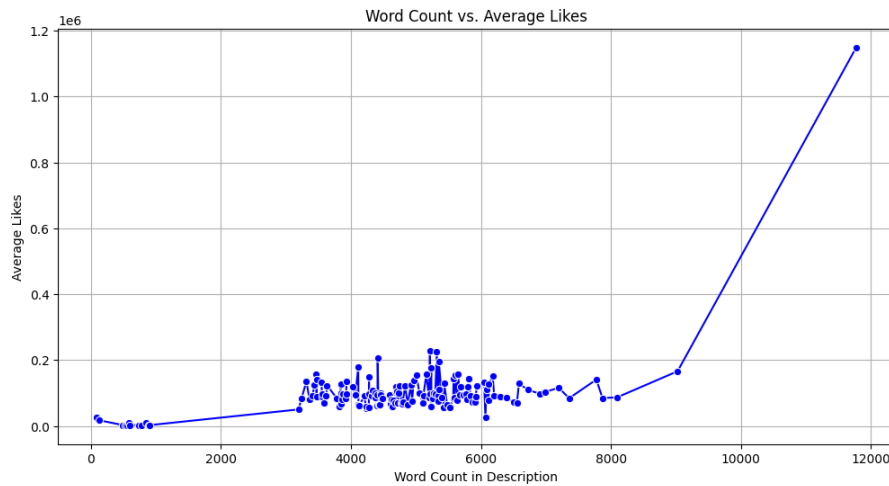


Figure 6: Average Likes versus Description Word Count

Observation: There appears to be a slight linear correlation with longer descriptions between the description's word count and the number of likes, suggesting that these videos may be favored by the audience.

2.2.7 Relation between Video Popularity (Views) and Other Factors

1. **Category:** Da7ee7 covers a broad spectrum of topics (e.g., science, history, psychology, culture). Our analysis aimed to determine if specific categories consistently attract higher view counts.

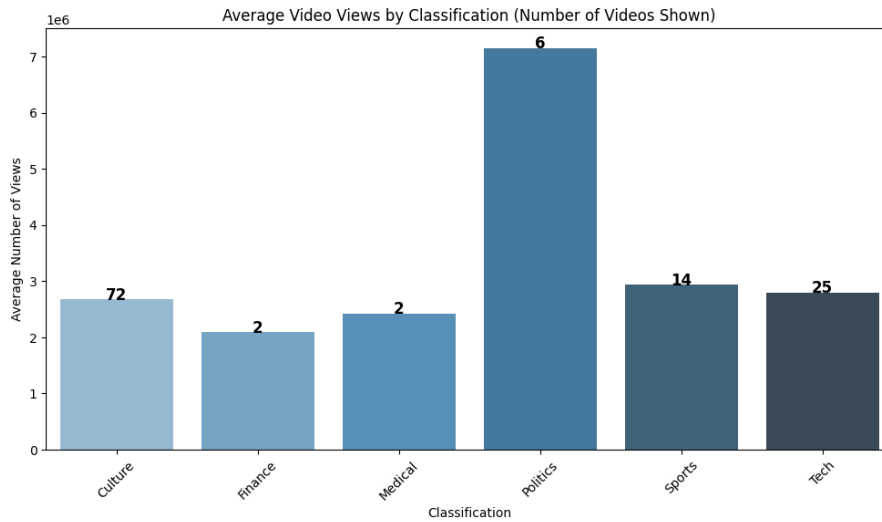


Figure 7: Average Video Views by Classification

Observations:

- Although some categories exhibit higher average views, the channel's *broad subject matter* and *mixed presentation style* tend to dilute any single dominant trend.
- As shown in Figure 7, videos related to politics often receive more views, possibly due to significant current events.
- The majority of videos are classified under the Culture category.

2. **Length:** We examined whether the length of a video influences its view count.

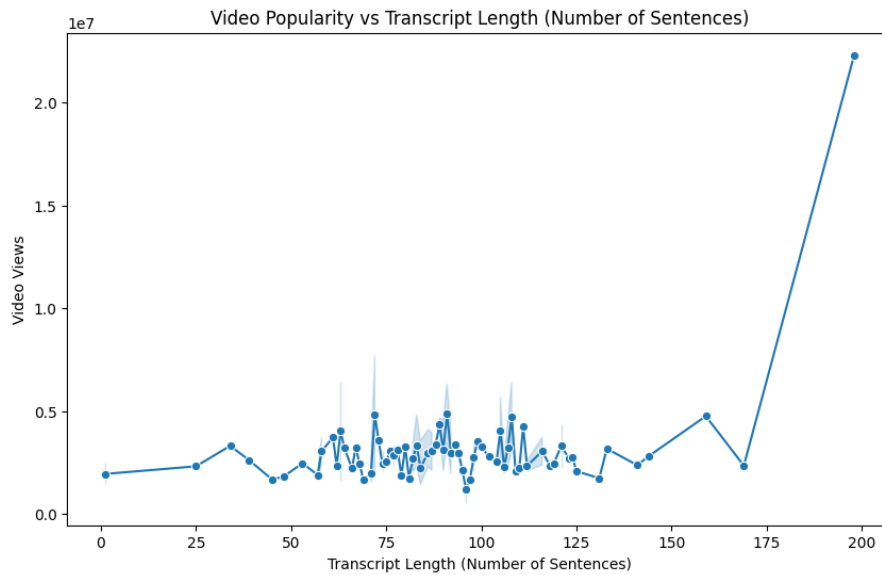


Figure 8: View Counts versus Number of Sentences in the Transcript

Observation: As shown in Figure 8, we see that there is almost no change in number of likes with respect to the number of words in the description except for videos with exceptionally high word count in the description. In this case, we see a noticeable increase in the average length.

3. **Sarcasm:** We also explored whether the amount of sarcasm in a video relates to its popularity.

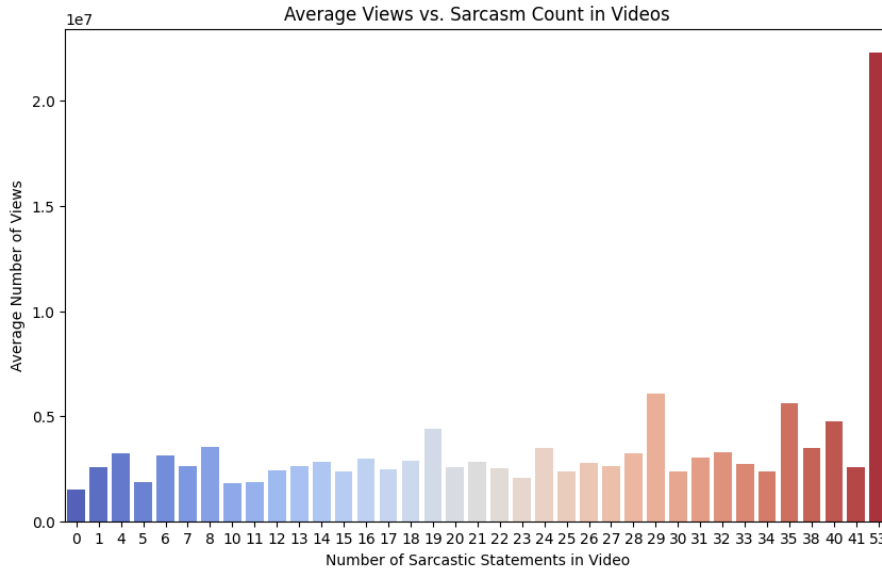


Figure 9: View Counts versus Sarcasm Frequency

Observation: At looking at the graph in Figure 9, we can notice that the number of video views in relation to the number of sarcastic comments in a video is relatively stable except for the videos with an exceptionally high number of sarcastic comments. In this case we see a spike in the number of views relative to the other ones.

2.3 Possible Tasks to Perform on this Data

One task we can perform is to do a regression on the number of views or likes a video will get based on the text.

3 Conclusion

This analysis demonstrates both the potential and the challenges of implementing an NLP preprocessing pipeline tailored to Arabic text, particularly for YouTube transcripts from ElDa7ee7. Our comprehensive approach—from text tidying, sentence segmentation, and cleaning, to applying advanced models for sentiment analysis, text classification, and sarcasm detection—has yielded valuable insights into the content and structure of the dataset.

Key takeaways include:

- **Correlation Insights:** We observed a strong linear correlation between long transcript length and video views, suggesting that more detailed videos tend to engage a larger audience. Additionally, specific categories, such as politics, show higher engagement, albeit from a smaller sample size.
- **Model Limitations:** While models for sentiment analysis and sarcasm detection performed adequately on controlled samples, they struggled with the nuances of Egyptian Arabic in real-world data. Similarly, the attempted translation of dialectal expressions proved inconsistent.

Although some techniques, like text summarization and translation, did not yield the expected improvements, the overall process comes handy in identifying properties of documents that could be beneficial for a neural network as to predict what category would Da7ee7 post about or the views & likes of his upcoming videos.