الكلية الجامعية
للعلوم التطبيقية
رائدة الإبداع

Report: Data Collection for Classifying Palestinian News (Real/Fake)

Natural Languages Processing - NLP Course

Data Preparation

June 2025

1. Introduction

The goal of this task is to prepare a dataset that can help us classify Palestinian news as **Real** or **Fake** using Natural Language Processing techniques. We focused on collecting news that is clearly related to Palestine and comes from different types of sources, including trusted media and social media platforms.

2. Data Collection

We collected around 400+ news articles about Palestine.

- Collected most news articles from January 2024, **with around 2 articles collected from dates before October 2023**.
- **Real news** (300 articles) came from trusted sources like **Al Jazeera**, **Misbar**, and **official Palestinian websites** (ministries, organizations, etc.).
- **Fake news** (about 100 articles) came from **social media platforms** such as **Twitter, Facebook, YouTube**, and some **suspicious websites**.

3. Data Extraction

- Used **web scraping** techniques (with the newspaper3k library) to extract titles and content from Al Jazeera articles and some other sources.
- Manually collected news from government websites by copying articles published in January 2024.
- Extracted publish dates from URLs using `urllib.parse`.

4. Cleaning the Data

- We removed unnecessary parts like **URLs**, **ads**, and **random symbols**.
- We **deleted articles not related to Palestine** to keep the data focused.

5. Data Organization

- Combine all collected data into a single dataset.
- Added an **ID column** to uniquely number each news article.
- Labeled all Al Jazeera and official government platform articles as **Real** because they are trusted sources. For **fake** news, we relied on verification platforms like Misbar, Tayqan, Tahaqaq, and Chayyek to identify and label the articles accordingly.

6. Saving the Dataset

- Saved the final cleaned and organized data as a CSV file named `GroupC_NLP_Task3.csv`.
- The final dataset includes **403 rows** (300 real, 103 fake) with 5 columns (ID, title, content, date, label).