Exploratory Data Analysis Report

Hespress Arabic Stories Data

WIDEBOT - 2023 Internship Program

Prepared by:

Abdalrhman Morsi

1. Introduction

The dataset contains 11,000 stories from Hespress and holds significant value for the development of Arabic Natural Language Processing (NLP) techniques. With this dataset, we can explore the characteristics of Arabic text and gain insights into its linguistic patterns and topics. The data was sourced from Hespress, a prominent news platform, making it relevant and representative of Arabic language usage.

The objective of this Exploratory Data Analysis (EDA) is to examine the dataset, understand its structure, and uncover valuable insights that can enhance NLP applications in the Arabic language.

2. Data Overview

The dataset consists of 11,000 stories, each associated with several attributes:

- Title: The title of the news story, providing a brief summary of its content.
- Author: The name of the author who wrote the news story.
- Publishing Date: The date when the story was published on Hespress.
- Topic: The category or topic to which the story belongs.

The data types include text (Title, Author, Publishing Date, and Topic). This dataset holds potential for understanding Arabic language usage and exploring the diverse topics covered in Hespress news stories. We will now proceed with data preprocessing to prepare the dataset for analysis.

3. Data Preprocessing

Before conducting the analysis, some preprocessing steps were applied to ensure data consistency and accuracy.

3.1 Concatenation of Topics

Initially, the data was distributed across multiple sheets, with each sheet representing a specific topic or category. To facilitate comprehensive analysis, all sheets were concatenated into a single dataset. This process allowed us to work with the entire collection of 11,000 stories, encompassing diverse topics covered by Hespress.

3.2 Date Conversion

The publishing dates in the original dataset were written in Arabic. To conduct temporal analysis efficiently, a two-step approach was employed to convert the dates into the English language and then to the datetime data type.

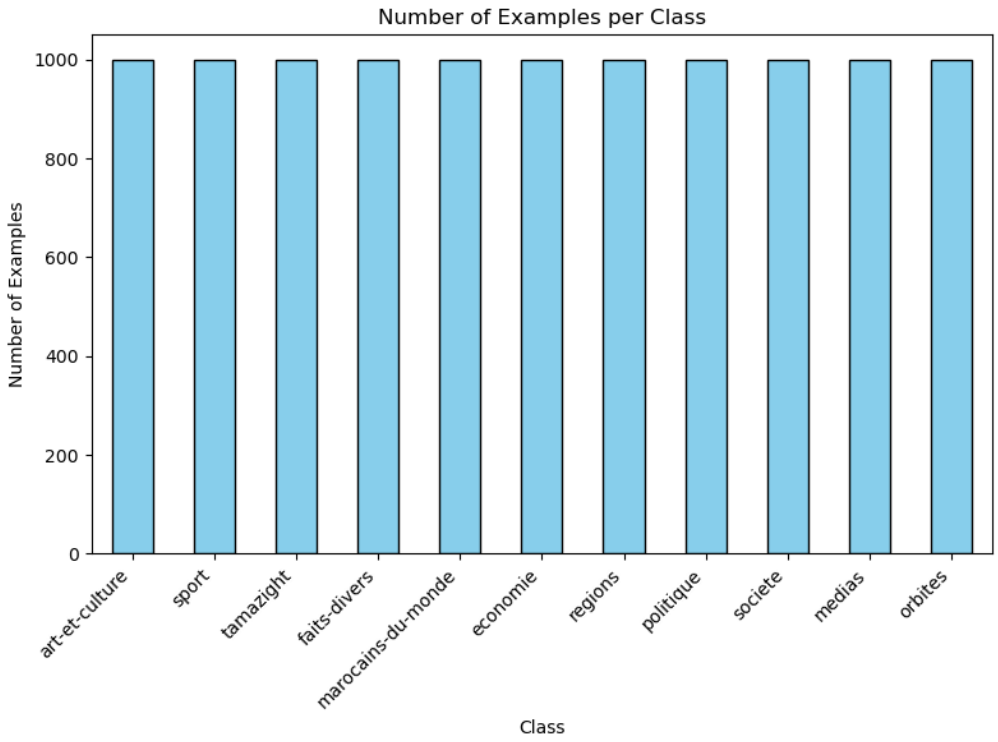
First, Arabic-to-English date conversion was performed, transforming the Arabic date strings into English equivalents. Next, the datetime data type was applied to the converted English dates. This conversion facilitated chronological ordering of the stories and enabled us to analyze temporal patterns over time.

With the data preprocessing steps complete, we are now ready to delve into the insights derived from the dataset.

4. Insights

Insight 1: Count the Number of Examples per Class

The dataset consists of 11,000 stories categorized into different topics. Each topic contains 1,000 stories, resulting in a balanced distribution across the dataset. This equal representation allows for comprehensive analysis and exploration of Arabic text data in diverse contexts.

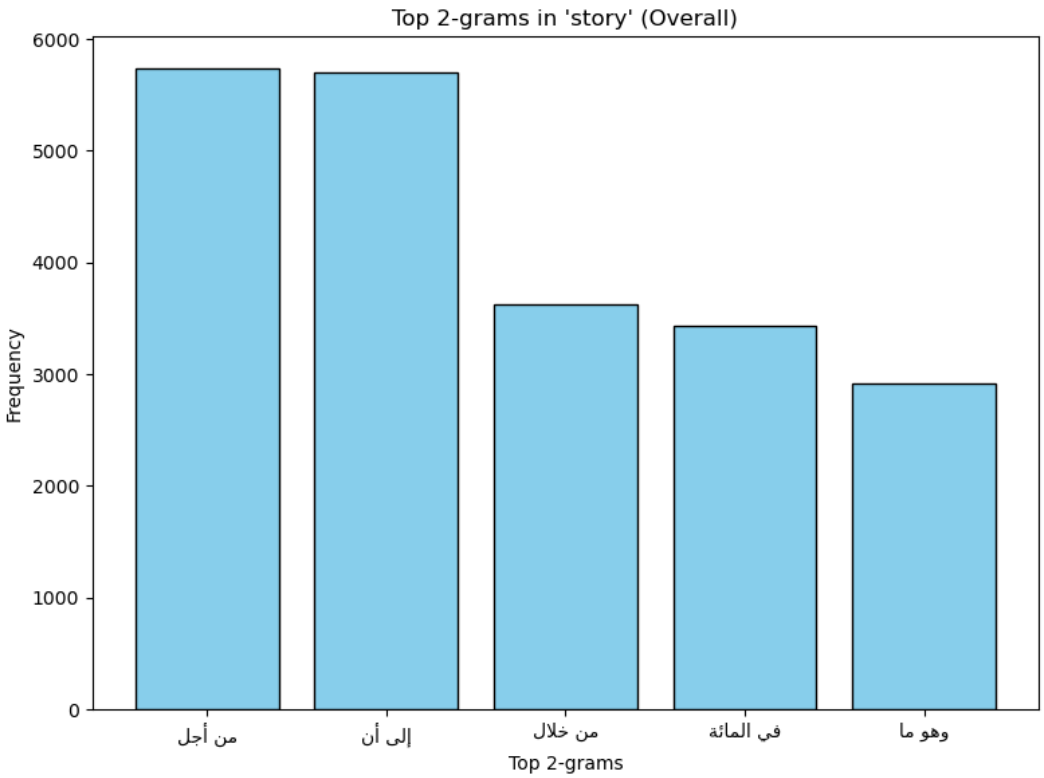


Insight 2: Top Frequent 2-grams Generally and per Class

To gain insights into the textual patterns within the dataset, we explored the most frequent 2-grams (pairs of consecutive words) both across the entire dataset and within each class (topic). By identifying the top frequent 2-grams, we aimed to uncover recurring word combinations that could be indicative of common themes or language usage.

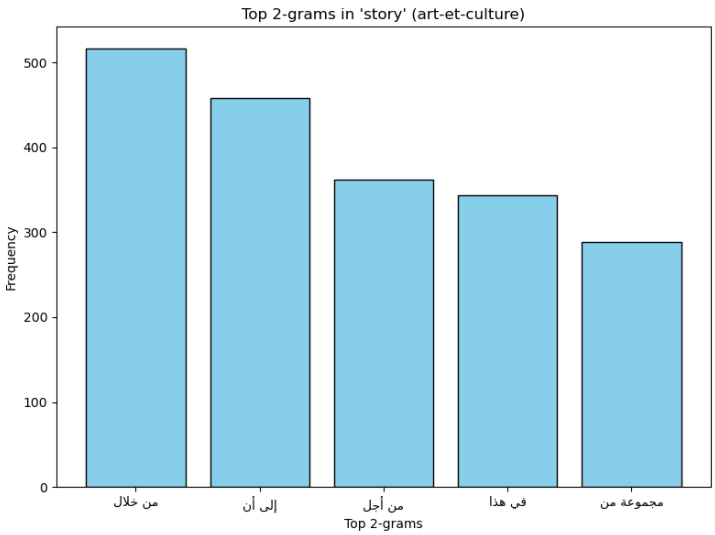
- Top 5 Frequent 2-grams Across the Dataset:

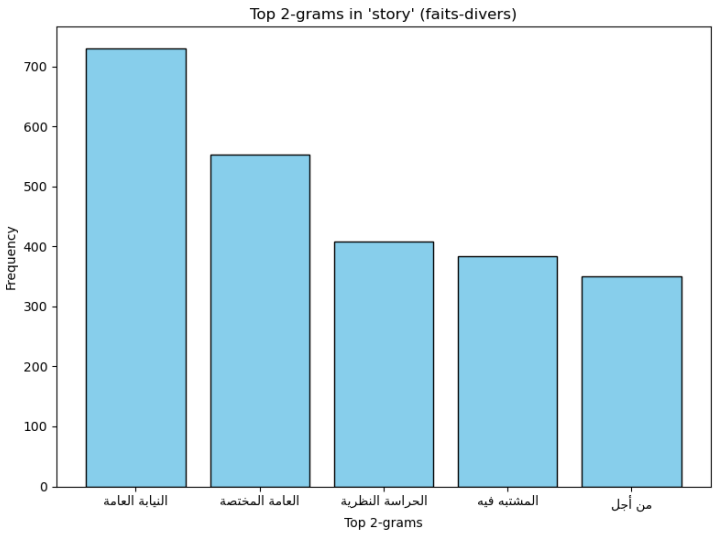
The bar chart below showcases the five most frequent 2-grams observed in the entire dataset. These combinations provide valuable insights into the prevalent language patterns found throughout the Hespress stories.

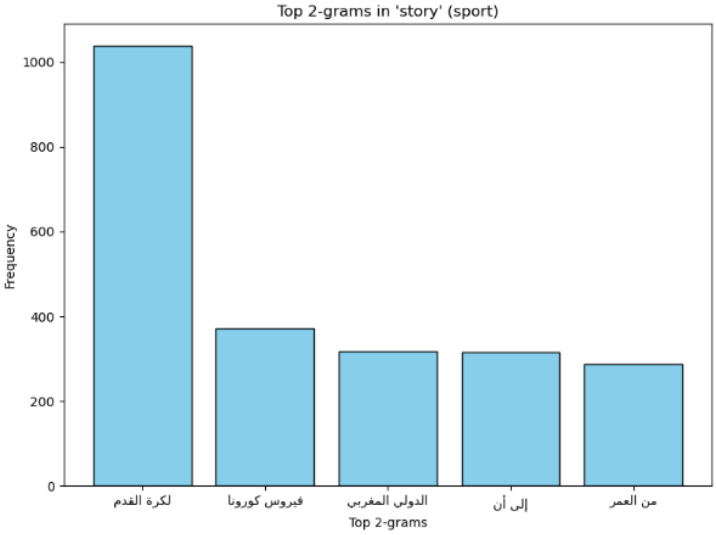


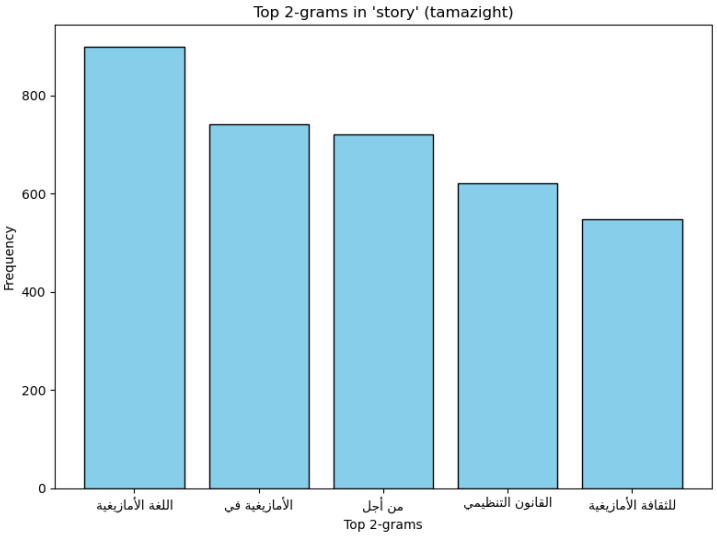
- Top 5 Frequent 2-grams per Class:

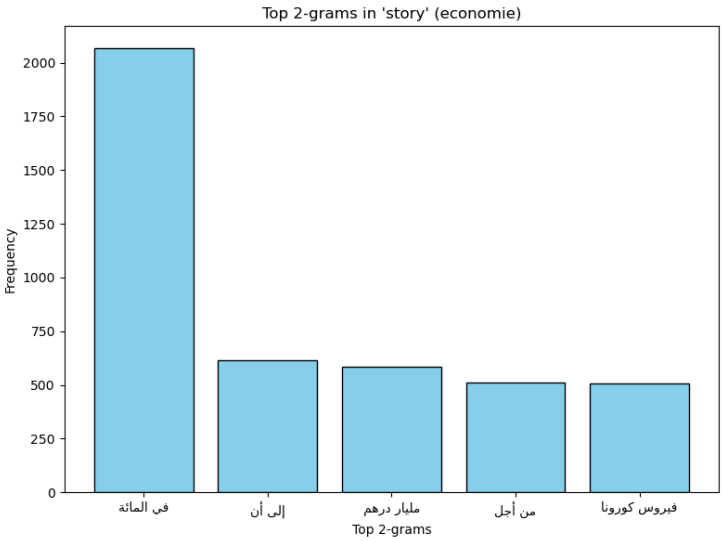
We further analyzed the most frequent 2-grams for each class (topic) individually. The following 11 bar charts display the top five 2-grams for each class. This analysis allows us to discern topic-specific language usage patterns and similarities.

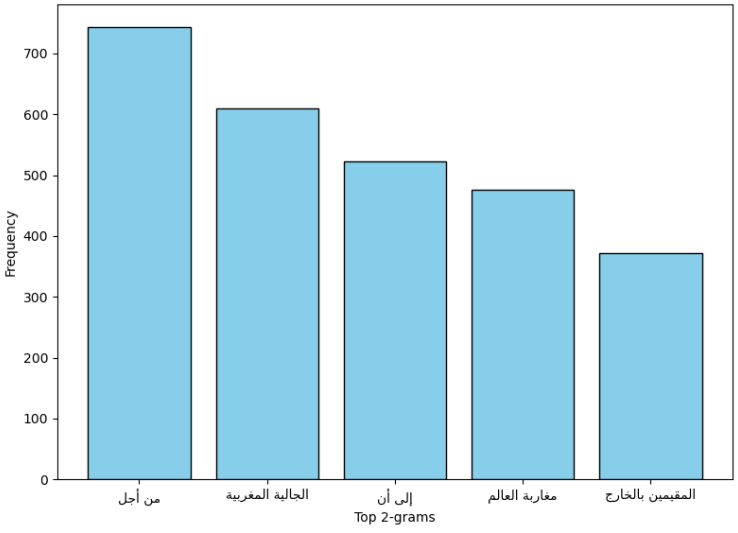

Top 2-grams in 'story' (art-et-culture)
Frequency
Top 2-grams
من خلال
إلى أن
من أجل
في هذا
مجموعة من

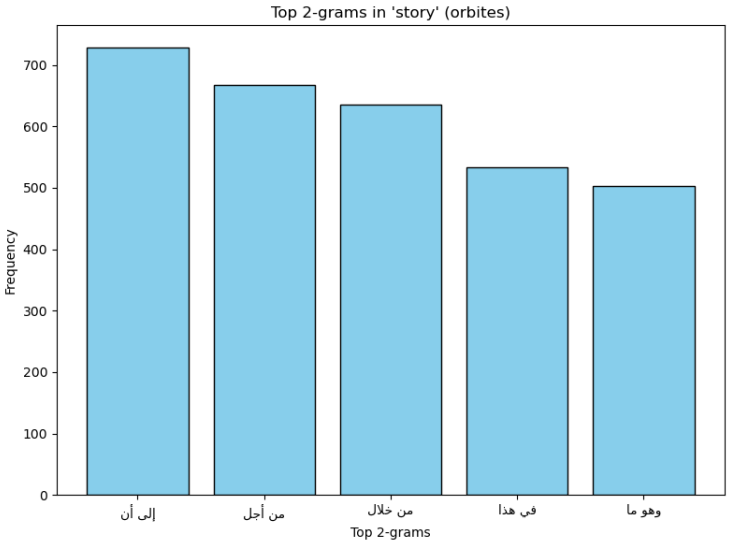

Top 2-grams in 'story' (faits-divers)
Frequency
Top 2-grams
النيابة العامة
العامة المختصة
الحراسة النظرية
المشتبه فيه
من أجل

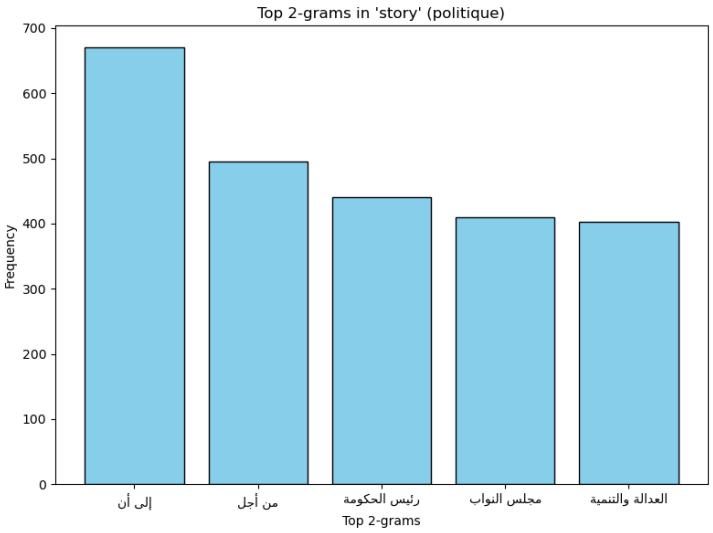

Top 2-grams in 'story' (sport)
Frequency
Top 2-grams
لكرة القدم
فيروس كورونا
الدولي المغربي
إلى أن
من العصر

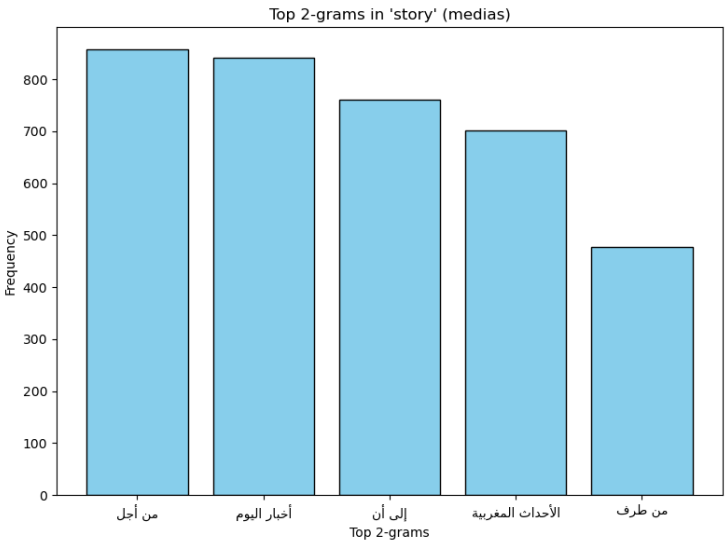

Top 2-grams in 'story' (tamazight)
Frequency
Top 2-grams
اللغة الأمازيغية
الأمازيغية في
من أجل
القانون التنظيمي
للثقافة الأمازيغية

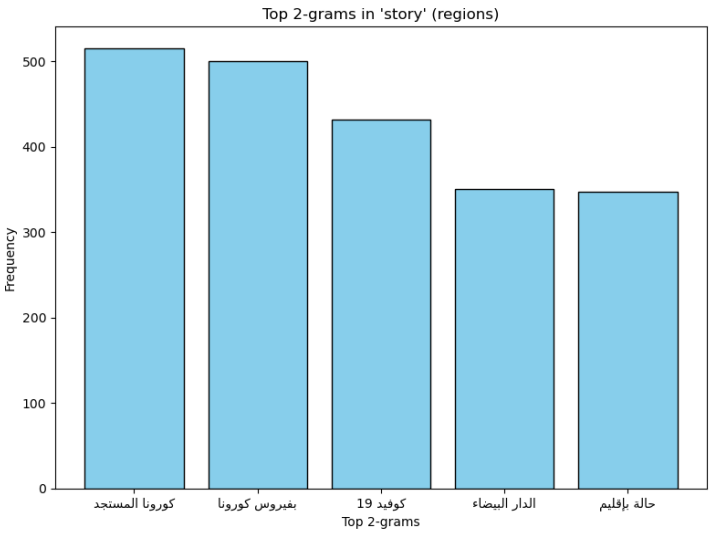

Top 2-grams in 'story' (economie)
Frequency
Top 2-grams
في المائة
إلى أن
مليار درهم
من أجل
فيروس كورونا

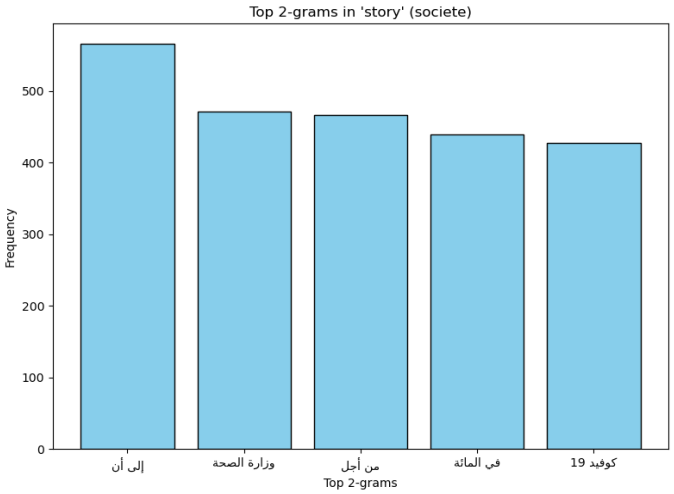

Top 2-grams in 'story' (marocains-du-monde)
Frequency
Top 2-grams
من أجل
الجالية المغربية
إلى أن
مغاربة العالم
المقيمين بالخارج









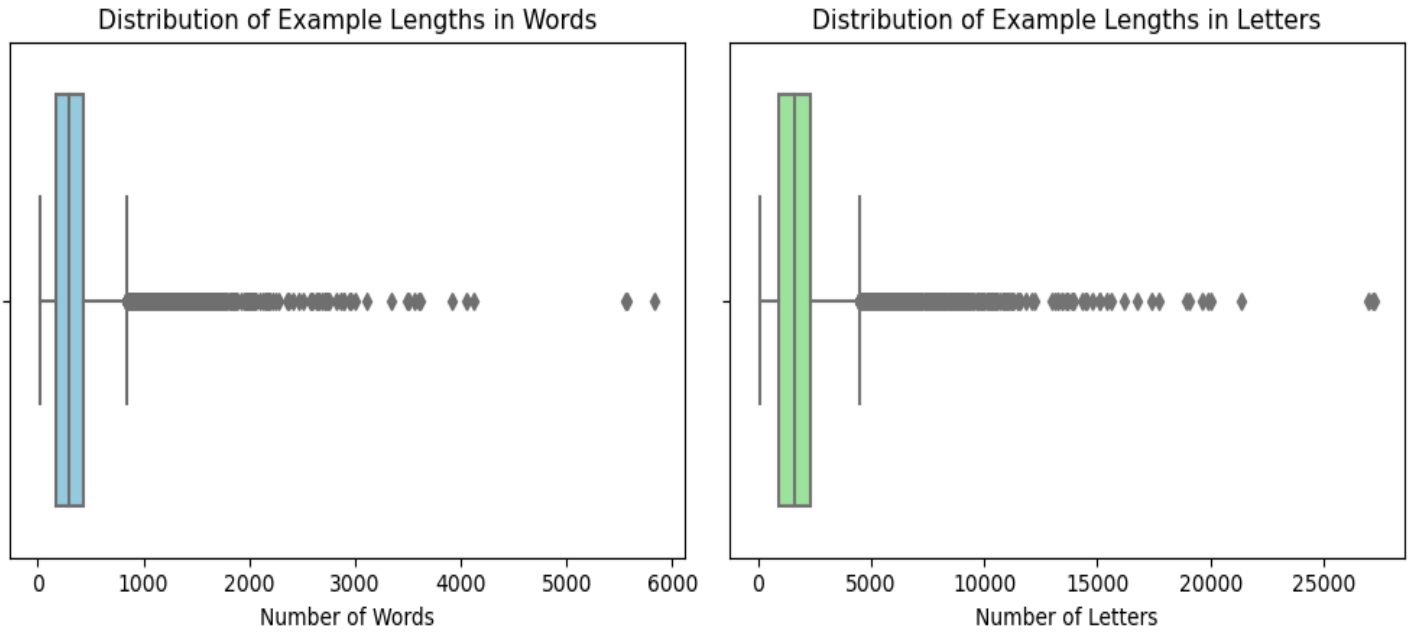


By examining the most frequent 2-grams both generally and within each class, we gain valuable linguistic insights that can aid in Arabic NLP development and topic categorization.

Insight 3: Lengths of Examples in Words and Letters

To better understand the distribution of story lengths in the dataset, we analyzed the number of words and letters per story. This analysis provides valuable insights into the variability and range of story lengths.

The box plots below illustrate the distribution of the number of words and letters per story. Most stories contain between 0 and 900 words and between 0 and 5,000 letters, with this range representing the most common length. However, some stories serve as outliers with word counts as high as 6,000 and 27,000 letters.



By analyzing the lengths of examples in words and letters, we gain valuable insights into the textual diversity present in the dataset. These observations can be significant for language processing tasks, such as text summarization or readability analysis.

Insight 4: Most Prolific Authors

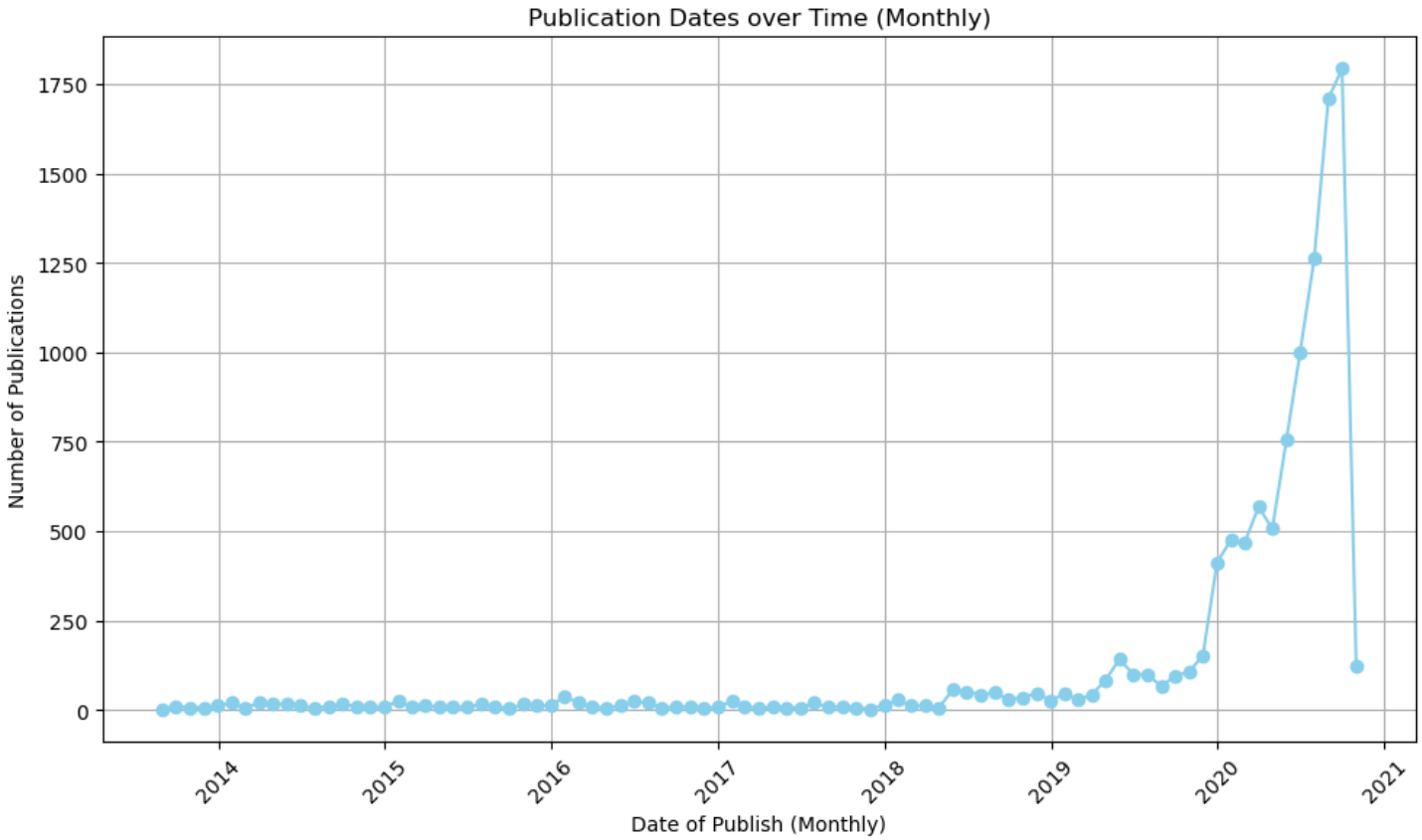
To identify the authors with the highest number of published stories, we analyzed the dataset to determine the most prolific writers. By quantifying the number of stories attributed to each author, we gain insights into the contribution and impact of individual writers in the collection.

The bar plot below highlights the top 50 authors who have contributed stories to the dataset. The first-ranked author holds a remarkable position, with over 400 stories to their credit. In contrast, the author in the second position has fewer than 100 stories published.

This analysis allows us to recognize the significant influence of certain authors and their contributions to the diverse content available in Hespress.

Insight 5: Dates of Publication

The timeline graph below illustrates the number of stories published per month from 2014 to 2020. During the years 2014 to 2019, the monthly story count remained relatively low, with less than 125 stories per month being published. However, a notable shift in publishing activity occurred in 2020. The timeline graph demonstrates a significant increase in the number of stories published per month, soaring to more than 1750 stories per month.



This observation indicates a substantial growth in story publications on Hespress during the year 2020, potentially signaling increased interest in the platform or changes in news reporting practices.

By examining the monthly distribution of stories, we can identify trends and fluctuations in publishing activity, and by analyzing the dates of publication over the entire timeline from 2014 to 2021, we gain valuable insights into the dynamics of news dissemination.