News Articles Title Generation INF 582 - Data Challenge 2024

Abasse DABERE abasse.dabere@polytechnique.edu
Cynthia Yacel FUERTES PANIZO cynthia-yacel.fuertes-panizo@polytechnique.edu

***Abstract*—Automated news article title generation plays a crucial role in enhancing user engagement and facilitating efficient information consumption in today's digital era. In this challenge, we explore several Natural Language Processing (NLP) techniques to generate titles from news articles. We analyze the dataset, consisting of text and titles, to understand the distribution of token lengths and identify prevalent themes. Our investigation reveals a positive correlation between the lengths of the text and titles, with political news dominating the dataset. Additionally, through experimentation with various pre-trained language models and fine-tuning strategies, we identify the mT5_multilingual_XLSum model as the most effective, particularly when paired with a dataset refined for grammatical accuracy using language_tool_python. Our extensive training and evaluation efforts result in a notable improvement in Rouge-L F-scores, with a final achievement of 0.23819, showcasing the effectiveness of our approach in automated news title generation.**

I. Introduction

The challenge focuses on developing NLP models to automatically generate news article titles. In today's fast-paced digital age, the abundance of daily news articles can overwhelm readers, highlighting the importance of impactful headlines for reader engagement. Automated title generation streamlines content creation, enabling readers to swiftly grasp the essence of an article, thereby enhancing user experience and promoting efficient information consumption. This challenge underscores the significance of advancing news dissemination by employing innovative methods for automatic headline generation.

The main goal is to use pre-trained language models and summarization techniques to create concise titles from news articles. Besides, fine-tuning on diverse datasets is essential for ensuring the adaptability of the models. In this challenge, model performance will be evaluated using the ROUGE-L F-Score metric, which quantifies the overlap between system-generated and reference sentences. This challenge fosters the exploration of innovative approaches to propel automated news title generation forward.

II. Background

A. Text Summarization

Text summarization is a vital task in Natural Language Processing (NLP) aimed at condensing large bodies of text into shorter, coherent representations while preserving key information. Efficient summarization techniques have become increasingly essential for various applications, including information retrieval, document clustering, and content generation.

In NLP, summarization methods are broadly classified into extractive and abstractive approaches. Extractive summarization selects key sentences or phrases from the original text, while abstractive summarization generates concise summaries by paraphrasing the text. These techniques aid in efficient information retrieval and comprehension, driving innovation across diverse fields such as journalism, academia, and content curation.

B. ROUGE-L Evaluation Metric

This challenge uses the ROUGE-L F-Score metric to assess the model performance. This metric gauges the alignment between system-generated sentences and reference sentences. ROUGE is based on the proportion of n-gram overlap between the system-generated sentence and one or more reference sentences.

ROUGE-n Recall is computed as:

$$ROUGE\text{-}n_{Recall} = \frac{\sum_{s\in S}\sum_{gram_n\in s} C_{match}(gram_n, s, g)}{\sum_{s\in S}\sum_{gram_n\in s} C(gram_n, s)} \tag{1}$$

$$ROUGE\text{-}n_{Precision} = \frac{\sum_{s\in S}\sum_{gram_n\in g} C_{match}(gram_n, s, g)}{\sum_{g\in G}\sum_{gram_n\in g} C(gram_n, g)} \tag{2}$$

Where:

- ***S*:** a set of of reference sentences.
- ***g*:** generated sentence.
- ***n*:** the length of *n-gram.*
- ***C(x, y)*:** the number of occurrences.
- ***C***

ROUGE-L is distinguished among the variants of ROUGE for its emphasis on the Longest Common Subsequence (LCS) between reference and generated summaries. This method assesses similarity by considering the longest sequence of shared words, prioritizing the preservation of word sequence.

$$ROUGE\text{-}L_{Recall} = \frac{\sum_{s\in S} LCS(s, g)}{\sum_{s\in S} |s|} \tag{3}$$

$$ROUGE\text{-}L_{Precision} = \frac{\sum_{s\in S} LCS(s, g)}{\sum_{g\in G} |g|} \tag{4}$$

III. Data Analysis

Before working with the models to generate the titles, we are going to analyze the training data of the *text* and *titles* of the news. The training dataset comprises 21,401 instances. This is done in *Data_Analysis.ipynb*.

A. Text Length Analysis

First, we are going to get some statistics on the size of *text* and *titles*.

In table I, we observe the statistical analysis of the length of both the *text* and *titles* measured by both *words* and *tokens*. Here, a *word* is defined as a sequence separated by spaces. Instances, where the counts of tokens and words differ, are often due to the presence of special characters (?, -, !, etc.) in the titles. For example, the title ’- *Qui a quoi ? -*’ contains 3 tokens (‘qui‘, ‘a‘, ‘qui‘) and 6 words (‘-‘, ‘qui‘, ‘a‘, ‘qui‘, ‘¿, ‘-‘).

Statistics of text and titles of news				
	Tokens		**Words**	
Statistic	**Text**	**Title**	**Text**	**Title**
Mean	322.46	29.87	350.34	32.10
Std	156.75	10.80	170.75	11.48
Min	28	3	33	6
25%	213	22	231	24
50%	288	30	313	32
75%	390	37	423	39
Max	2,894	139	3,163	155

TABLE I
Statistics of tokens and words in the text and titles of news.

We observe a close similarity between the counts of tokens and words. For our analysis, we will focus on tokens. A token serves as a fundamental unit of lexical representation in natural language processing (NLP), typically corresponding to a word.

In the models developed for the challenge, we will limit our analysis to the first 512 tokens of the text in order to generate titles. We chose 512 because it is the nearest power of 2 to: 546 tokens = 390 tokens (the 75th percentile) + 156 tokens (one standard deviation). We work with the 75th percentile since longer lengths may be outliers. The entire histogram can be seen in Fig. 1.

As we can see in Fig. 2, the correlation matrix indicates a positive correlation, although it’s not a strong correlation, of approximately 0.24 between the lengths of the text and titles in the dataset. This suggests that as the length of the text increases, there tends to be a slightly larger length in the corresponding titles.

B. Unigrams

In the next step, we’ll analyze tokens as Unigrams. Unigrams are individual tokens in text, useful for frequency analysis, text classification, and language modeling. This will help us identify the most popular Unigram tokens appearing in the news, both in the *text* and *titles*.

To begin with, we filter the stop words and create word clouds to visualize the most common tokens in *text* and *titles*. As shown in Fig. 3, it’s evident that the predominant words

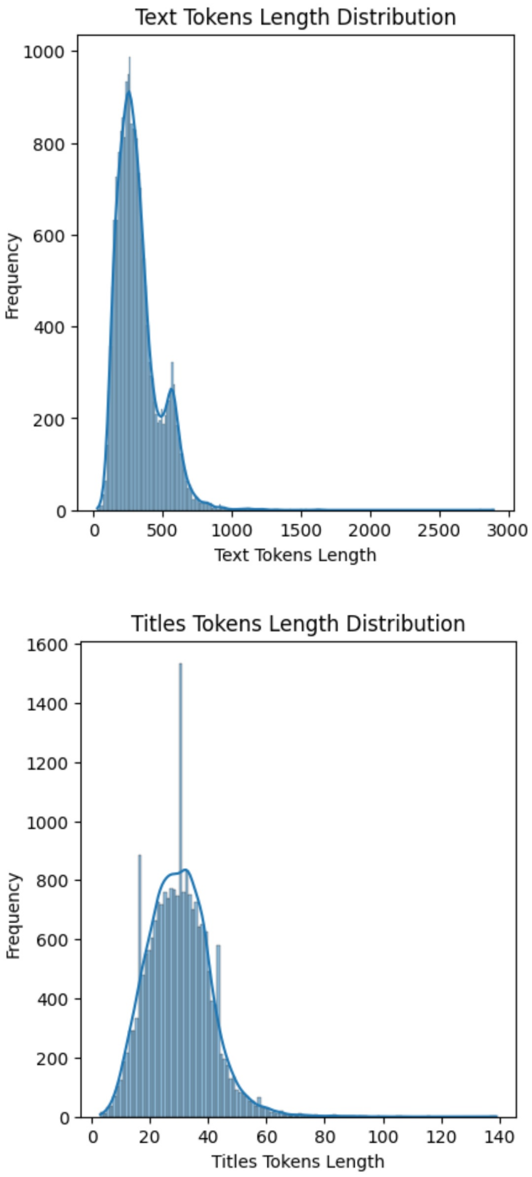

Fig. 1. Length histogram of text tokens and titles tokens.

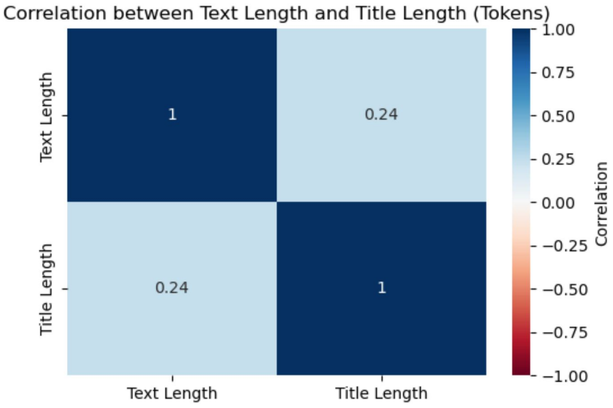

Fig. 2. There is a positive correlation between the number of tokens in the text and the number of tokens in the title.

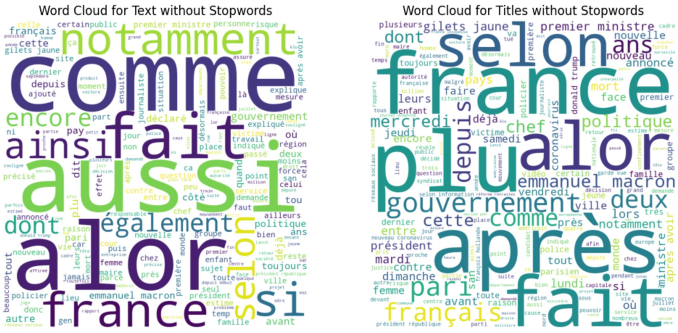

Fig. 3. Although we've removed stopwords from our token analysis, it's evident that the word cloud for unigrams in both text and titles still contains certain prepositions, adverbs, and similar words.

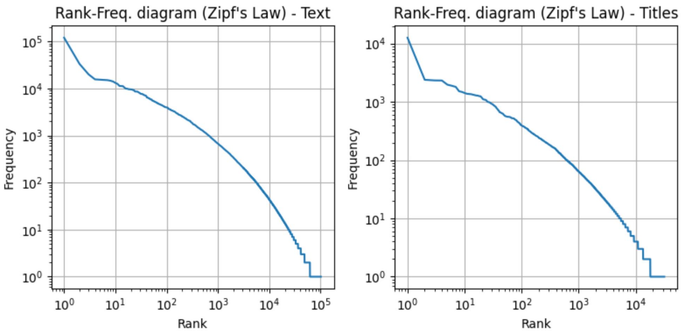

Fig. 4. Rank-frequency diagram that illustrates the relationship between word frequency and its rank.

mainly comprise prepositions and adverbs, but the word cloud plot is not ideal for getting additional insights, so we are going to elaborate additional plots.

A rank-frequency diagram illustrates the relationship between word frequency and its rank within a dataset, adhering to Zipf's Law, which states that a word's frequency is inversely proportional to its rank. This results in a logarithmic distribution of word occurrences. In Fig. 4, we observe this pattern, highlighting prominent words in news titles and text. In this context, "rank" denotes the ordinal position of a word in a sorted list based on its frequency, with higher ranks corresponding to less frequent words. This means that words with lower ranks are more commonly found in the dataset, while those with higher ranks are less frequent.

As depicted in Fig. 4, certain words emerge as predominant in news. Therefore, we opt to visualize some of the most common tokens (the top 30) in Fig. 5. This image is constructed by simply counting how many times each token repeats. After filtering out conjunctions, prepositions, adverbs, and similar terms, we observe that prominent words include 'france', 'français', 'ans', 'années', 'ministre', 'président', 'lundi', 'mardi', 'mercredi', 'jeudi', 'vendredi', 'samedi', 'dimanche', and 'coronavirus'.

Nevertheless, a more effective method for identifying the most relevant words is by utilizing TF-IDF (Term Frequency-Inverse Document Frequency), which is a technique in NLP that measures the importance of a word in a document relative to a collection of documents. It combines two metrics: term frequency (TF), which measures how often a word appears in a document, and inverse document frequency (IDF), which measures how unique a word is across a collection of documents. TF-IDF is calculated by multiplying TF and IDF, providing a numerical representation of word importance in a document. Therefore, TF-IDF offers a more refined approach compared to our previous method. In Fig. 6, the most frequent tokens are more clearly delineated. While many relevant tokens align with our previous analysis, TF-IDF provides a direct analysis without the need for additional manual efforts as the ones needed with the previous approach.

C. Bigrams

Bigrams are consecutive pairs of tokens used for co-occurrence analysis, language modeling, and pattern detection in text. Transitioning from unigram to bigram analysis, we'll now delve into visualizing TF-IDF scores, which often reveal more pertinent information. For instance, it unveils specific names like the French president or former U.S. presidents, as well as terms related to traffic news such as "gilets jaunes". We can see more detail in Fig. 7.

D. Trigrams

In the final part of our analysis, we'll examine Trigrams, which are sets of three consecutive tokens in the text. Similarly to bigrams, we'll exclusively delve into TF-IDF analysis, as it yields the most pertinent insights. Similarly to previous analyses, stopwords are excluded. Examining the trigrams in Fig. 8, reveals associations with political topics, suggesting a prevalence of political news in the training dataset. Moreover, the significance of days like "vendredi", "samedi", and "dimanche" is reiterated.

IV. Data processing

In this section, we will outline the data preprocessing that we performed before running the models.

Approach	Result
import re	*Parisest lacapitale de la France*: Using regex is not an efficient method for properly separating the words.
import spacy	*Parisest lacapitale de laFrance*: It's ineffective as it fails to insert spaces in our example.
import wordninja	*Paris est la capitale de la France*: Initially, it seems effectively. But, when applied to the training dataset, it fails, e.g., transforming "La mesure est décriée par" to "La mesure est d cri e par".
import nltk	*Parisest lacapitale de laFrance*: It's ineffective as it fails to insert spaces.
import language _tool _python	*Paris est la capitale de la France*: This approach appears effective in our example, and it seems to have good results when it's applied across the training dataset. Thus, **we choose this approach**.

TABLE II
Testing many approaches to try to correct "Parisest lacapitale de laFrance".

During our exploratory analysis, we discovered instances of sentences within the text lacking spaces in the training dataset.

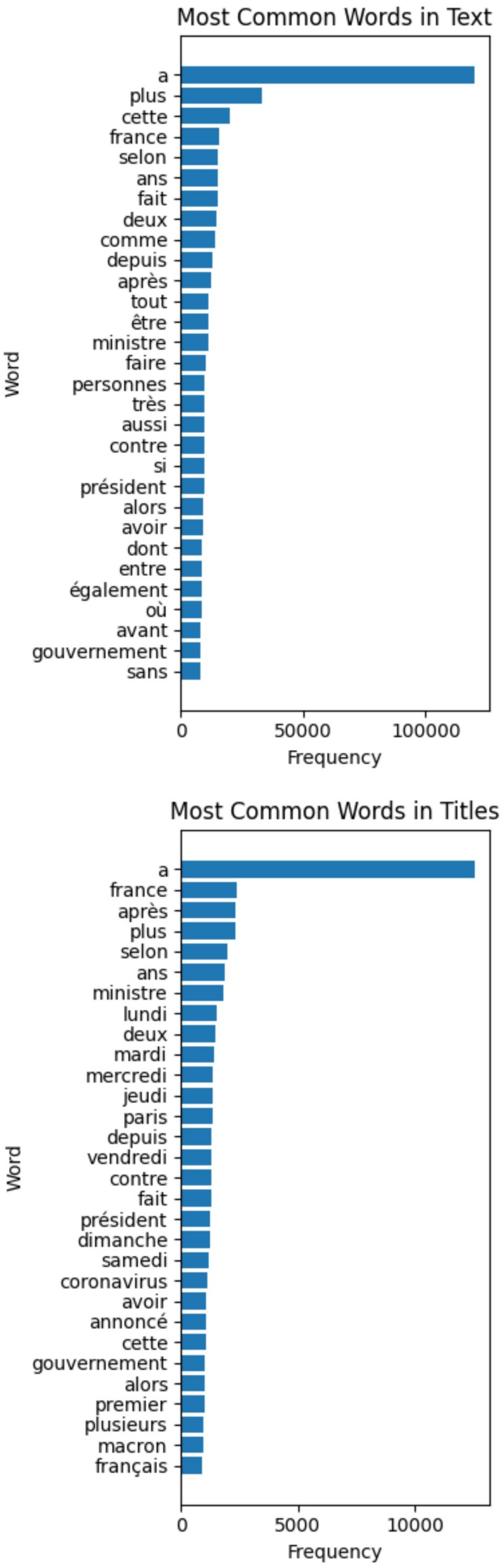

Fig. 5. Unigrams: Excluding prepositions, adverbs, and similar words, we observe that some of the most frequent words in the titles include 'france', 'français', 'ans', 'années', 'ministre', 'président', 'lundi', 'mardi', 'mercredi', 'jeudi', 'vendredi', 'samedi', 'dimanche', and 'coronavirus'.This bar plot is constructed by simply counting how many times each token repeats.

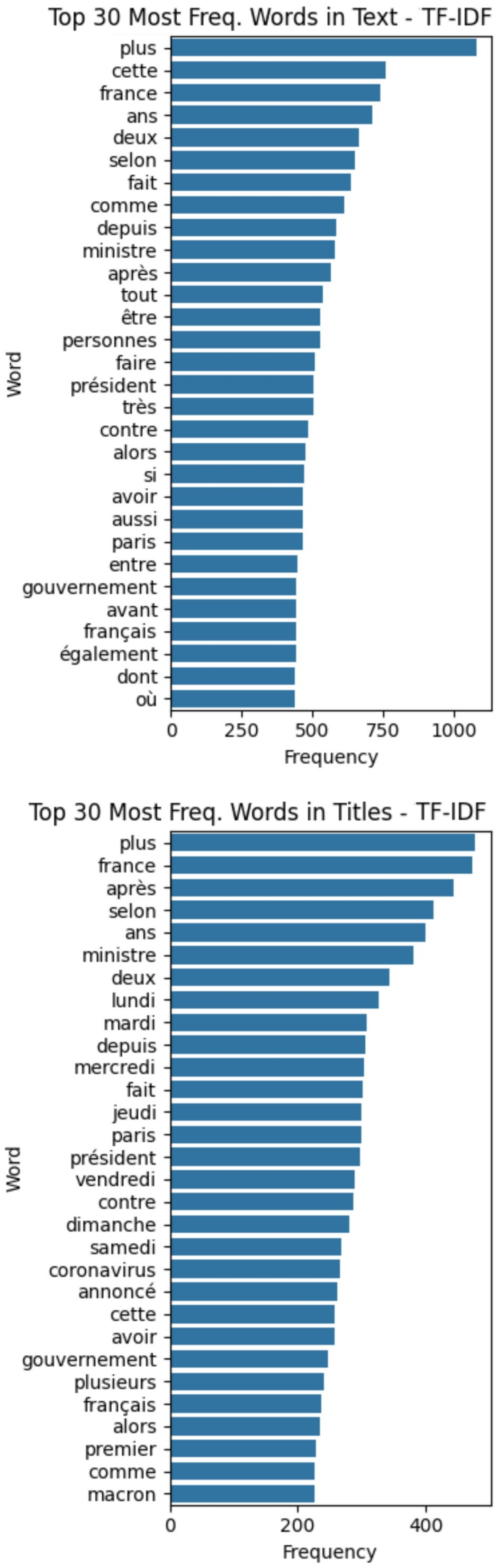

Fig. 6. TF-IDF offers a refined approach for identifying relevant words, providing direct analysis without additional manual efforts. We can see clearly that some of the most frequent tokens are: 'france', 'gouvernement', 'ministre', 'président', 'lundi', 'mardi', 'mercredi', 'jeudi', 'vendredi', 'samedi', 'dimanche','coronavirus', and so on.

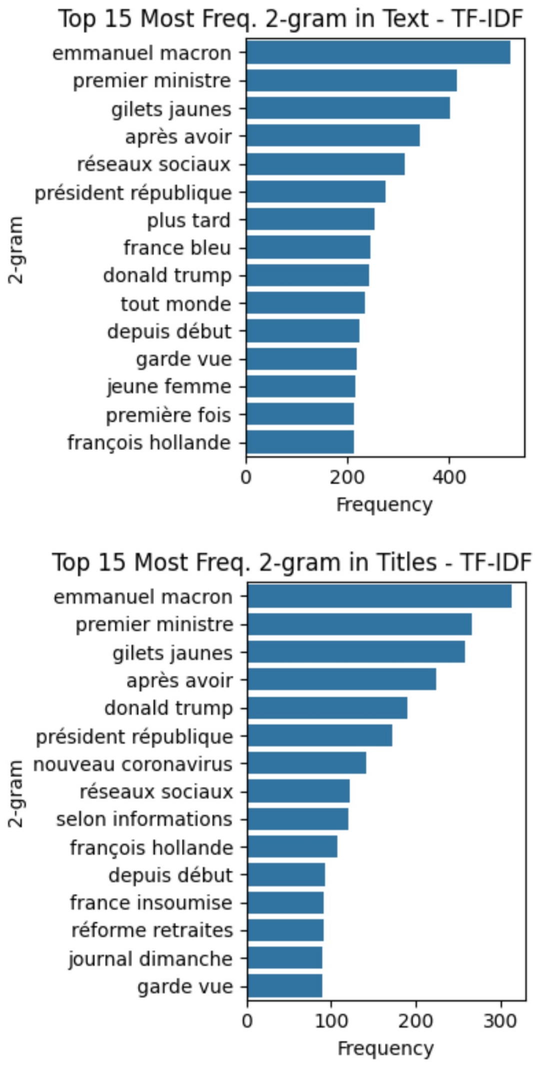

Fig. 7. TF-IDF for bigrams reveals previously unnoticed information, such as names of political figures like the President of France or former Presidents of the USA, as well as terms related to traffic news like ”gilets jaunes”.

For instance, in the phrase ”Un membre incontournableEt ce n’est autre,” we observed a missing space between ”incontournable” and ”Et,” which should be ”incontournable Et.” Consequently, we aim to enhance the grammar. Our progress and various approaches will be detailed in the notebook *Grammar_correction.ipynb*.

Prior to achieving precise grammar correction, we explored various approaches. The most successful method involved utilizing `import language_tool_python`. Table II outlines the five different approaches we analyzed. Within this table, we demonstrate the results of attempting to correct the grammar for the phrase `Parisest lacapitale de laFrance`. Besides, the table highlights two approaches that yielded promising results, both of which were applied to the training set. Notably, `import language_tool_python` demonstrated superior performance.

Using `import language_tool_python` yields effective results in grammar correction for the training set. Notable outcomes include:

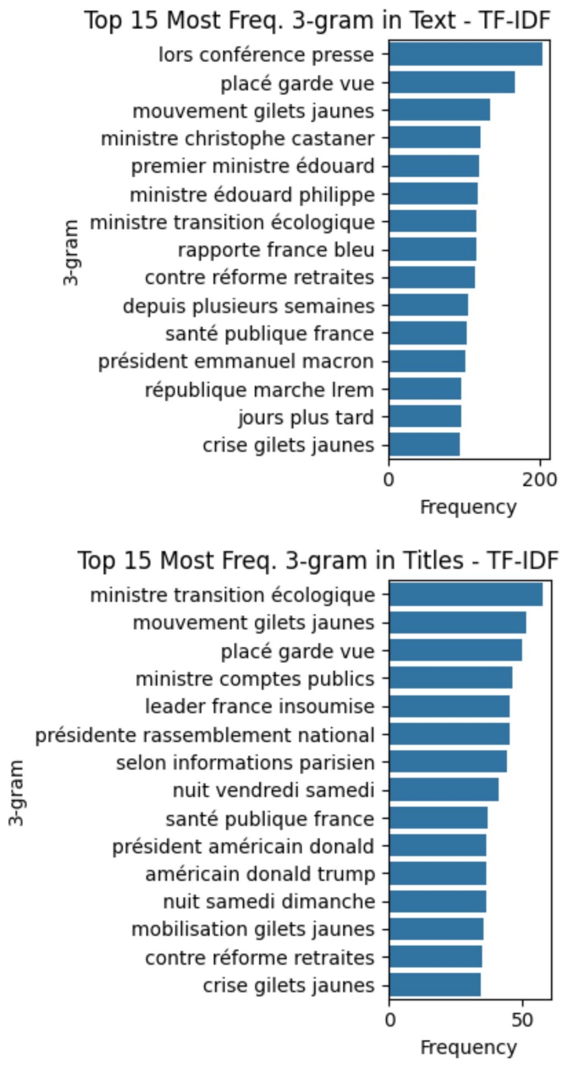

Fig. 8. TF-IDF for trigrams suggests a prevalence of political news in the training dataset. Moreover, the significance of days like ”vendredi”, ”samedi”, and ”dimanche” is reiterated.

- **data i = 10 - add space**
  Before: Un membre incontournableEt ce n’est autre
  After: Un membre incontournable Et ce n’est autre

- **data i = 50 - change words but maintains the meaning**
  Before: jusqu’à ce que les Etats remplissent certaines conditions en matière notamment de respect des droits humains.
  After: jusqu’à ce que les États remplissent certaines conditions en matière particulièrement de respect des droits humains.

- **data i = 2580 - a little wrong, e.g. it transform ”du” to ”de la”**
  Before: Les trois personnes atteintes du Covid-19 sont prises en charge au centre hospitalier d’Ajaccio
  After: Les trois personnes atteintes de la Covid-19 sont prises en charge au centre hospitalier d’Ajaccio

- **data i = 10580 - change to synonyms**
  Before: Leur analyse sur l'échec du parti aux élections législatives est également requis
  After: Leur analyse sur l'échec du parti aux élections législatives est aussi requis

When we utilize the dataset with grammatical correction from `import language_tool_python`, we notice that the score of our best model increases from `0.23212` to `0.23819`.

V. Models and Results

A. Models Utilized

Our project leveraged three distinct pre-trained language models from Hugging Face:

- lincoln_mbart_mlsum_automatic_summarization
- plguillou_t5_base_fr_sum_cnndm
- csebuetnlp_mT5_multilingual_XLSum

These models were selected for their proficiency in generating high Rouge-L F-scores, a critical metric in our evaluation process. The csebuetnlp_mT5_multilingual_XLSum, in particular, stood out for its notable performance in initial tests.

For model training and evaluation, we used both original data and versions with corrected grammar. This approach aimed to gauge the adaptability of the models to varying linguistic qualities in news articles.

B. Training Process

Each model underwent fine-tuning on an NVIDIA RTX A4000 GPU with 16086 MB memory. The Low-Rank Adaptation (LORA) technique was employed with varying ranks (2, 4, 8, 16, 20, 32, 64), to identify the optimal configuration. As an example, fine-tuning csebuetnlp_mT5_multilingual_XLSum with a LORA rank of 64 and a batch size of 10 took approximately 16 hours. Due to GPU availability constraints, we occasionally adjusted the LORA rank and batch size, which had a practical impact on the results.

C. Results

The csebuetnlp_mT5_multilingual_XLSum model exhibited superior performance from the early stages of fine-tuning, leading us to focus further efforts on this model.

Model	ROUGE-L F-Score
mbart_mlsum_automatic_summarization	0.1996
t5_base_fr_sum_cnndm	0.211
mT5_multilingual_XLSum	0.23819

TABLE III
The ROUGE-L F-Score of the models developed.

D. Observations and Insights

Our observations revealed that a higher LORA rank generally correlated with improved model performance. Additionally, utilizing data with corrected grammar consistently yielded better Rouge-L F-scores. These insights were instrumental in steering the fine-tuning process and achieving the noted results.

VI. Conclusions

- After analyzing the token lengths in *Text* and *Titles*, we notice that the most relevant length is to evaluate the first 512 tokens in the *Text* because it is the nearest power of 2 to the total number of tokens: 546 tokens = 390 tokens (which is the 75th percentile) + 156 tokens (one standard deviation).
- There is a positive correlation of 0.24 between the lengths of the *Text* and *Titles* in the dataset. This suggests that as the length of the *Text* increases, there tends to be a slightly larger length in the corresponding *Titles*.
- After analyzing the unigrams, bigrams, and trigrams, we observe that political news seems to prevail in the training dataset. Some key words for the *Titles* include 'france', 'ministre', 'président', 'lundi', 'mardi', 'mercredi', 'jeudi', 'vendredi', 'samedi', 'dimanche', and 'coronavirus'.
- After conducting several tests on different models, we found that the best score is obtained with the model `mT5_multilingual` and with the grammatical correction dataset created using `import language_tool_python`.

References

[1] Alexey Bukhtiyarov, Ilya Gusev. *Advances of Transformer-Based Models for News Headline Generation*. arXiv preprint arXiv:2007.05044, 2020.