Aix Marseille université
Socialement engagée

amse
école d'économie d'aix-marseille
aix-marseille school of economics

Hackathon Airbus Helicopters

TEAM BrainWave

Text summarization using T5-Base

Authors:
Hicham EL MAKAOUI
Jean-baptiste GOMEZ
Nelly AGOSSOU
Oussama RHITI
Ulrich SEGODO

February 2024

Table of contents

I. General

II. Data used

III. Data processing

IV. Description of the model

V. Description of the learning methodology

VI. Separation between training and validation

General

Our project revolves around the implementation of a finetuning of the T5 transformer (Text-To-Text Transfer Transformer) for the text summarization task due to its ability to perform different natural language processing tasks at the same time. using a single architecture. Leveraging cutting-edge natural language processing (NLP) techniques, T5 offers a flexible, high-performance architecture that can be tailored to generate accurate text summaries.

T5 is pre-trained on large datasets, allowing it to capture general linguistic representations from various text domains. In our project, we will use it mainly to train the data from the Airbus database because it allowed us to have better performance and precise summaries for the target domain.

Data

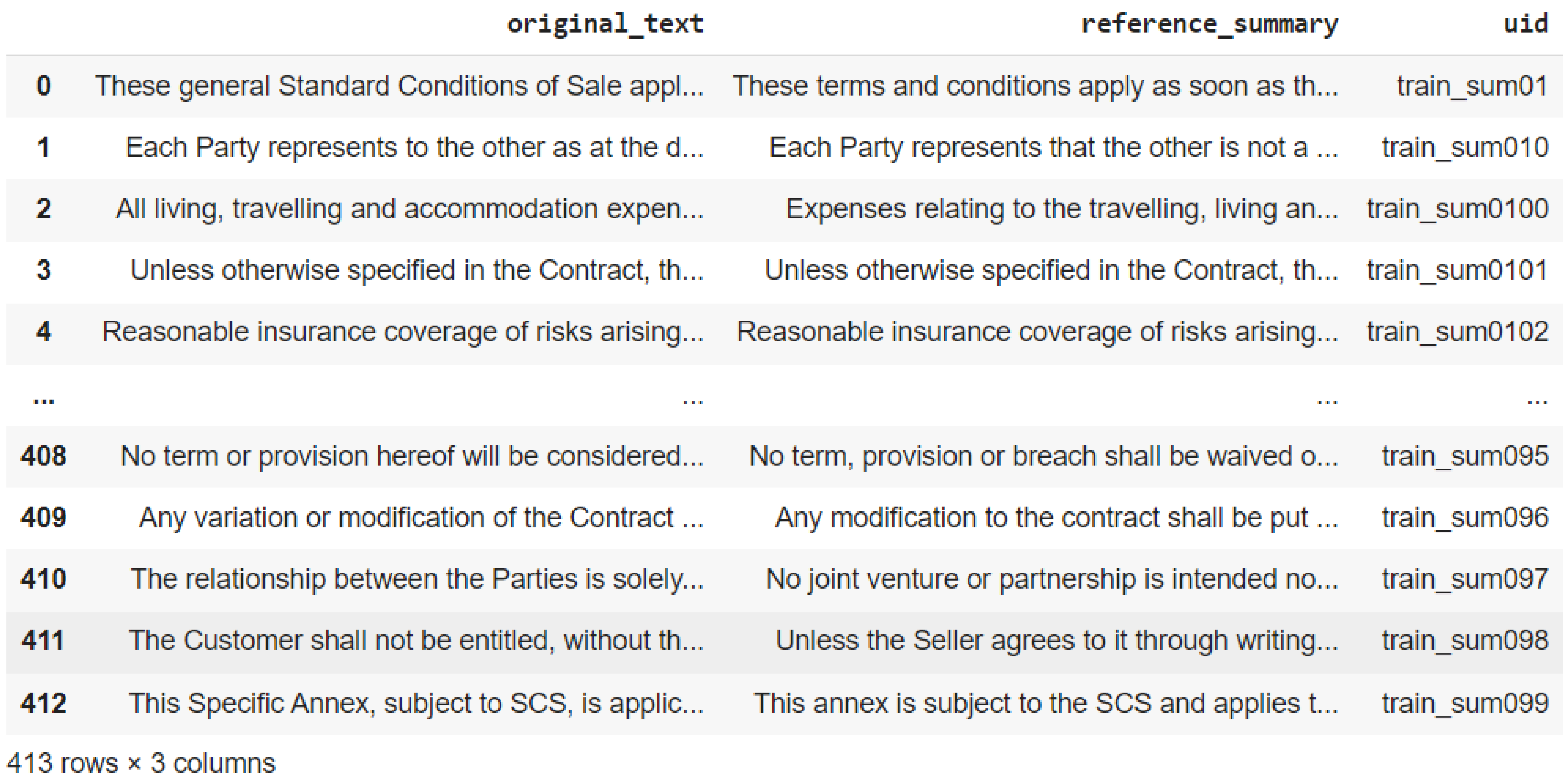
Data Sources

The primary dataset for our project is derived from Airbus, encapsulated in a JSON format. This dataset serves as the foundation for training and evaluating our text summarization model. Through thorough data exploration and manipulation, we prepare the information for further processing.

Number of Observations: 413 entries

Objective:

- Training and evaluation of automatic summarization models
- Development of domain-specific summarization systems

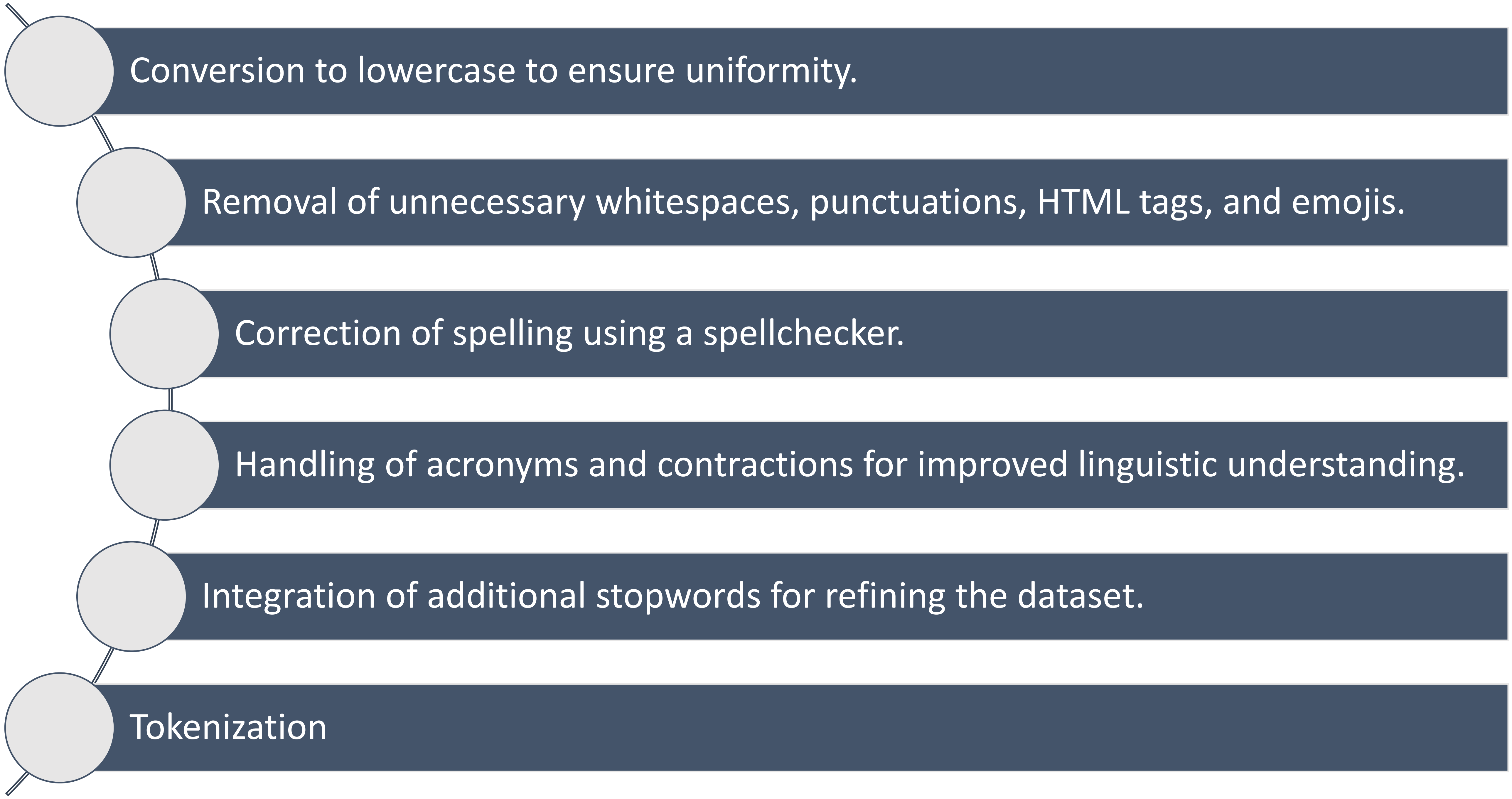


	original_text	reference_summary	uid
0	These general Standard Conditions of Sale appl...	These terms and conditions apply as soon as th...	train_sum01
1	Each Party represents to the other as at the d...	Each Party represents that the other is not a ...	train_sum010
2	All living, travelling and accommodation expen...	Expenses relating to the travelling, living an...	train_sum0100
3	Unless otherwise specified in the Contract, th...	Unless otherwise specified in the Contract, th...	train_sum0101
4	Reasonable insurance coverage of risks arising...	Reasonable insurance coverage of risks arising...	train_sum0102
...	...	...	...
408	No term or provision hereof will be considered...	No term, provision or breach shall be waived o...	train_sum095
409	Any variation or modification of the Contract ...	Any modification to the contract shall be put ...	train_sum096
410	The relationship between the Parties is solely...	No joint venture or partnership is intended no...	train_sum097
411	The Customer shall not be entitled, without th...	Unless the Seller agrees to it through writing...	train_sum098
412	This Specific Annex, subject to SCS, is applic...	This annex is subject to the SCS and applies t...	train_sum099

413 rows × 3 columns

Source: Provided by Airbus Helicopters

Data processing



- Conversion to lowercase to ensure uniformity.
- Removal of unnecessary whitespaces, punctuations, HTML tags, and emojis.
- Correction of spelling using a spellchecker.
- Handling of acronyms and contractions for improved linguistic understanding.
- Integration of additional stopwords for refining the dataset.
- Tokenization

Description of the model

We need to set some **basic configurations** for the training and dataset preparation pipeline.
We choose to fine-tune the **t5-base** model. The batch size is 4 and the number of processes used for parallel processing is 4 as well. We will train for 10 epochs, and the maximum context length of the original text will be 512. Taking into account the average word length of the original texts. Therefore, all original texts below 512 tokens will be padded and all original texts above 512 tokens will be truncated. This is the right size for this dataset.

Tokenizing the Dataset

Tokenizing means converting a word into a numerical value. Sometimes a single word may be broken down into multiple ones. First, the **T5 Tokenizer** is loaded followed by the process function.
Note that for each input original text, we again prepend the “summarize: ” text. This will act as the trigger token and the model will learn to summarize the original text that goes into the labels.
The tokenized datasets are stored in **tokenized_train** and **tokenized_valid** respectively.

Description of the model

Initializing the Model

We load the **T5 Base model** and move it to the computation device. The T5 Base model contains around 223 million parameters. It may look like a large model, but it works much better compared to the **T5 Small model.**

Training the Model

To train the text summarization model using T5, we need to define the training arguments and training API. The model will be evaluated every 200 steps. Do note that we pass the *preprocess_logits_for_metrics* and *compute_metrics* methods to the Trainer API.

Separation between training and validation

To ensure the model's performance and prevent overfitting, the dataset was divided into training and validation sets (20%). The model was trained on the training set, and its performance was evaluated on the validation set. This separation allows us to assess the model's generalization ability and make necessary adjustments to improve its performance.

The training process involves multiple epochs, each refining the model's understanding of the provided data.

Through this comprehensive approach, our T5 model demonstrates proficiency in summarizing complex textual information, showcasing its adaptability and effectiveness in the domain of automatic text summarization.