University of Padova

Department of Mathematics

Master Thesis in Data Science

Analysis of frames and persuasion techniques in online news articles in a multilingual setup

Supervisor
Giovanni Da San Martino
University of Padova

Master Candidate
Dacia Braca

Academic Year
2023-2024

To those who were close to me in my time of need.

Abstract

Persuasion is the ability to influence people's behaviour and choices by exploiting words and emotions to achieve a set goal, such as convincing the interlocutor of a message or idea. The definition has no inherently negative meaning but recognising such techniques and the context in which they are used can be essential int understanding the dynamics behind communication.

Within this framework is the contribution of the mass media which, in order to quickly disseminate information or effectively send a message to as wide an audience as possible, have refined their communication strategies by using persuasion techniques, in such cases in an inflated manner. Combined with these, the skilful use of perspective framing in the description of an event and the narrative techniques that characterise a journalistic article on a narrative level allows the digital information system to be exploited for propaganda purposes.

Previous works focus separately on each of these aspects. On the other hand, in this project we intend to provide a complete overview of propaganda approaches, studying all its nuances and carrying out three different analyses of multilingual texts from online newspapers. Specifically, the main aim of this work is to investigate twenty-three persuasion techniques used within news articles, accompanied by a category label and written in six different languages, to cover a wide range of topics according to fourteen possible framings. Several variants of artificial intelligence models based on the XLM-RoBERTa architecture are proposed, allowing a given article, written into one among in six languages, to be classified by category, to detect prospective framing and persuasion techniques within it. For the latter the classification is assessed at the level of the whole article and at the fine level, not only to inform about the presence of a technique but also to label it and detect the fragment of text in which it appears.

Collecting textual data with a good level of information in this sense is a challenge not to be underestimated, as the supervision procedure is demanding, time-consuming and expensive in terms of resources. For this reason, Transformer-based architectures, such as mBART and ChatGPT, are used to perform data augmentation and acquire useful information to bridge language discrepancies and improve performance.

Contents

Listing of figures

Listing of tables

Listing of acronyms

NLP . . . . Natural Language Process

RNN . . . . Recurrent Neural Network

LSTM . . . . Long-Short Term Memory

BERT . . . . Bidirection Encoder Representation for Transformers

MLM . . . . Masked Language Model

NSP . . . . Next Sentence Prediction

RoBERTa . . . . . . Robustly Optimized BERT Approach

XLM-RoBERTa Cross-lingual Language Model RoBERTa

BART . . . . Bidirectional and Auto-Regressive Transformers

mBART . . . . Multilingual BART

GPT . . . . Generative Pre-trained Transformer

ChatGPT . . . . . . Chat Generative Pre-trained Transformer

1 Introduction

Over the past decades, the Internet has radically revolutionised the way people access and consume information, in particular, it has assumed a predominant role as a vehicle for daily news of even international interest, transforming the global media landscape. This transformation has been fuelled by a number of factors, including the advent of social media, the proliferation of online news sites and the development of streaming technologies. News and information have become accessible in real time and from anywhere in the world, allowing people to stay constantly updated on the latest events and developments. This rapid dissemination of information has had a significant impact on the way people understand and participate in public debates, also influencing the perception and construction of narratives about events and topics of global interest. In addition, the unprecedented access to information has created new challenges and opportunities for journalists, who must navigate between the need for speed and accuracy in news presentation.

In the context of contemporary journalism, online articles present themselves as a varied mosaic of styles, tones and objectives. There is no universal formula to describe a journalistic article, as they can adopt multiple forms and functions depending on the content covered and the objectives of the journalist or editor. In recent years, however, online propaganda news articles have caused increasing concern about the spread of disinformation and manipulation of public opinion. These are designed to influence the opinions and perceptions of readers through the use of targeted persuasion techniques and the presentation of issues of common interest in a framing directed towards particular targets. The persuasion techniques used in online propaganda articles are varied and often subtle. One of the most common methods is the use of emotive and value-laden language to elicit emotional reactions in the reader and lead them to develop a sympathy or antipathy towards certain subjects or topics. Other common techniques include the use of selective or anonymous testimonies to confirm the propaganda narrative. The goal of these persuasion techniques is to shape the public's opinions and perceptions, pushing them to support certain interests or points of view. Online propaganda articles can be used to promote political ideologies, support economic interests or promote a certain worldview, and their dissemination on digital platforms makes it easy to reach a wide audience and influence opinions on a global scale. Added to this is the complexity of the international landscape in terms of linguistic communication.

In this work, an in-depth analysis of the annotated multilingual dataset SemEval-2023, with online news articles in six languages (English, German, French, Italian, Polish and Russian), was therefore performed, with the first objective to see how the approach to information varies between different cultures, nationalities and languages. In addition, thirteen artificial intelligence models based on XLM-RoBERTa algorithm have been developed, with the aim of classifying a journalistic article according to the category among opinion, reporting and satire articles (four models), to 14 possible perspective framing with which an event can be presented to the reader (five models), and tracing among 23 persuasion techniques used for propaganda purposes at two different levels of granularity (3 models at whole text level and one at span level). Special attention was paid to the study and use of neural networks, suitably trained for Natural Language Procress tasks, based on the attention mechanism and Transformers architecture, with which it was not only possible to tackle multi-class and multi-label classification tasks but also to perform two different data augmentation techniques. The mBART model was employed to perform translations between the six languages considered, in order to improve performance on predictions and to bridge the different frequency distributions between the six language datasets. In addition, ChatGPT was also used to generate new items based on the definitions and information provided.

2
Theoretical Background

This section is devoted to the description of the theoretical backgroud required to tackle classification tasks at different levels of granularity in practice. The fundamental architectures and mechanisms are explored in detail, tracing the path followed in the acquisition of Natural Language Process skills. Not all the architectures proposed in the sections that follow are actually used, but their study was an integral part of the path that led to the choice of the model with which we mainly worked, namely XLM-RoBERTa. For each topic, there is a specific section in which the architecture will be introduced and whose constituent aspects will be explained, sometimes in a summarised manner, sometimes in greater detail as they are functional to the work actually carried out. Certain concepts that are common between the architectures will be recalled as required and to ensure that each section can be read and understood independently of the others. In particular, we start with a presentation of the state of the art before the advent of Transformers, on which the main focus of the chapter will then be on. We then proceed with the presentation of the Bidirectional Encoder Representation for Transformers model, progenitor of the RoBERTa, XLM-RoBERTa and mBART models.

2.1 Previous Architectures

Before the advent of Transformers, the models considered most efficient in solving sequence transduction tasks were based on convolutional and recursive neural networks [1].

When handling sequential data, not only text but also audio-visual material, whatever the task to be solved it is necessary to make efficient use of the contained information as there may be some correlations or fundamental causal relationship between individual words, sound fragments or framing images.

Let us suppose, for instance, we want to translate the title of a newspaper article from Italian into English:

> USA, Trump vince le primarie repubblicane in New Hampshire. Haley: "Mi congratulo, ma la corsa non è finita".
>
> ↓
>
> USA, Trump wins Republican primaries in New Hampshire. Haley: "I congratulate him but the race is not over yet".

This short text consists of two sentences. The first is self-contained as all the information necessary for its comprehension, and thus also for the translation, is already accessible. On the other hand, the second sentence rests on the first one, in fact otherwise we would not be able to tell who "Haley" is congratulating. If taken individually, the references and perhaps the context are missing. We then need a model able to establish dependencies and connections between the various sequential elements, keep track of previous information imitating the human memory mechanism.

2.1.1 Recurrent Neural Network

Recurrent Neural Networks, for instance, have a loop structure that allows information to be retained and passed through a series of *recurrent units* arranged in a sequence, one after the other. Each recurrent unit receives two types of input: one is the current input (e.g. an element of a data sequence) and the other is the output of the same unit in the previous time step. This allows the RNN to maintain an internal memory of the information processed up to that point in the sequence (see Figure 2.1).

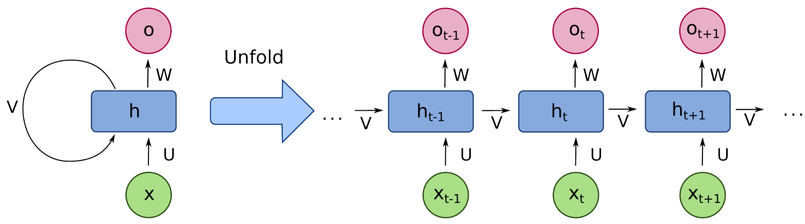


Figure 2.1: Recurrent Neural Network chain

The RNN's computational process can be described in more detail through three main steps. (i) *Initialisation*: At the beginning of the sequence $t = 1$, the recurrent unit receives an initial input x_1, e.g. the first element of the input sequence, and an initial memory h_0 usually initialised to zeros. (ii) *Information passing and iteration*: Let t be the current time, the recurrent unit combines the current input x_t with internal memory from the previous time step h_{t-1} to calculate a new internal state h_t. This new internal state is then used to generate current output o_t. The information passing process is repeated for each element in the sequence. Each recurring unit uses its updated internal state as input for the next time step (2.1). (iii) *Output*: At the end of the sequence, the output of the last recurring unit can be used.

$$\begin{cases} h_t = f\big(Ux_t + Vh_{t-1} + b\big) \\ o_t = f\big(Wh_t + c\big) \end{cases} \tag{2.1}$$

RNNs are particularly suitable for modelling long-term sequential data, as they can capture long-term dependencies between elements in the sequence through their recurrent connections. However, traditional RNNs can also suffer from vanishing gradient or exploding gradient problems during training, which can make it difficult to learn long-term dependencies in very long sequences.

To mitigate this problem, variants of RNNs such as Long Short-Term Memory (LSTMs) and Gated Recurrent Units (GRUs) have been developed, which include gating mechanisms to control the flow of information through the network and mitigate the gradient vanishing/exploding problem.

2.1.2 Long-Short Term Memory

The structure and functioning of Long-Short Term Memory networks (LSTM) include several key components. (i) *LSTM Cell* is the basic unit of an LSTM network. It consists of one or more layers of neurons connected recurrently. The LSTM cell contains elements such as the input gate, output gate, and forget gate, which regulate the flow of information within the cell itself. (ii) *Input Gate* regulates how much external data should be added to the new internal state of the LSTM cell. It takes into account the current input and the previous state to decide which information is relevant for updating the internal state. (iii) *Output Gate* decides which part of the LSTM cell's internal state should be exposed as output to the next time step. This gate regulates the output based on the cell state and the current input. (iv) *Forget Gate* determines which information in the LSTM cell should be forgotten or retained in the long term. This gate helps maintain the long-term memory of the network, allowing it to capture long-term temporal relationships in data sequences. (v) *Internal State* of the LSTM cell is an internal representation of the network's memory. It is updated at each time step based on the current input and the previous internal state, using the input, output, and forget gates.

Certainly this architecture proves to be more effective in preventing the vanishing/exploding gradient effect, however it does not completely solve the problems associated with handling long sequences. The sequentiality of the operations makes training times prohibitive without returning predictions with a satisfactory quality.

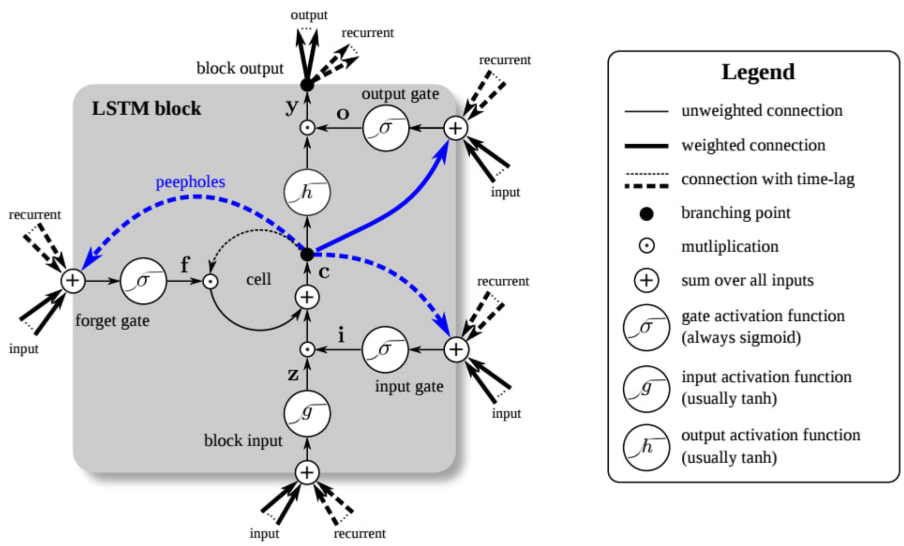
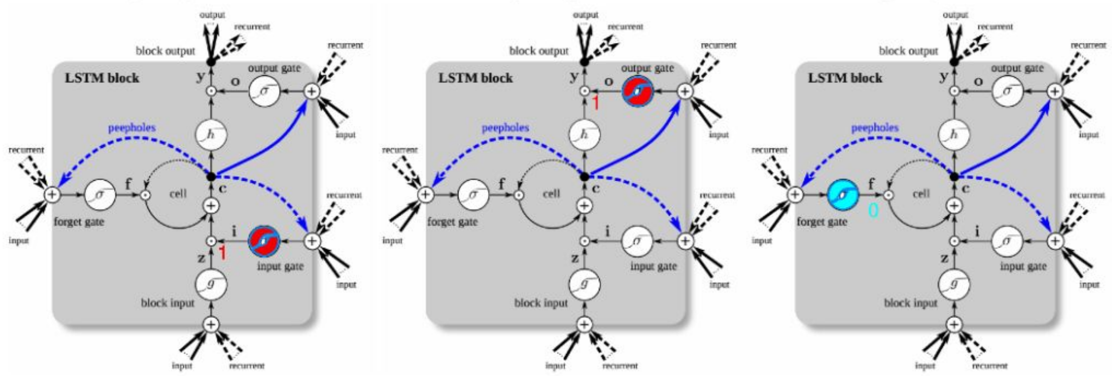


Figure 2.2: Long-Short Term Memory structure



(b) Input Gate on

(c) Output Gate on

(c) Forget Gate on

Figure 2.3: Long-Short Term Memory Gates

2.2 Transformers

The Transformers architecture, originally developed by a team of Google researchers in 2017 and presented to the public in the paper "Attention is All You Need" [2], has revolutionised the field of Natural Language Processing (NLP) and then transcended this boundaries finding applications in computer vision, audio and multi-modal processing. The public release of the *transformers library* came later, in 2019, thanks to the company Hugging Face*. This made models based on Transformer architectures more accessible and easier to use for a broad community of machine learning researchers and developers. Transformers architecture has also led to the development of highly successful pre-trained models, such as generative pre-trained transformers (GPT) and Bidirectional Encoder Representations from Transformers (BERT) [3].

The previous dominant sequence transduction models were based on complex recurrent or convolutional neural networks that include an encoder and a decoder, connected through an attention mechanism in the best performing models. The innovative aspect of the Transformers architecture lies in its structural simplicity, in fact it is based solely on attention mechanisms (see 2.2.2), dispensing with recurrence and convolutions entirely [2]. The absence of sequential structures, replaced precisely by a principle of parallelisation of computational operations, brings multiple advantages: more efficient management of information, optimisation of the training process with significant savings in time and resources, and an appreciable increase in

*https://huggingface.co/

the quality of predictions. For these reasons Transformers are widely used for NLP tasks: here a textual input sequence is firstly translated into a numerical representation called *token* and each token can be processed simultaneously by several computing units through multi-headed attention mechanisms. The bidirectionality of the processing allows effective contextualisation of each element within a context window, amplifying the signal for key tokens and diminishing the importance for less relevant ones. The Transformer's architectures is made up by two

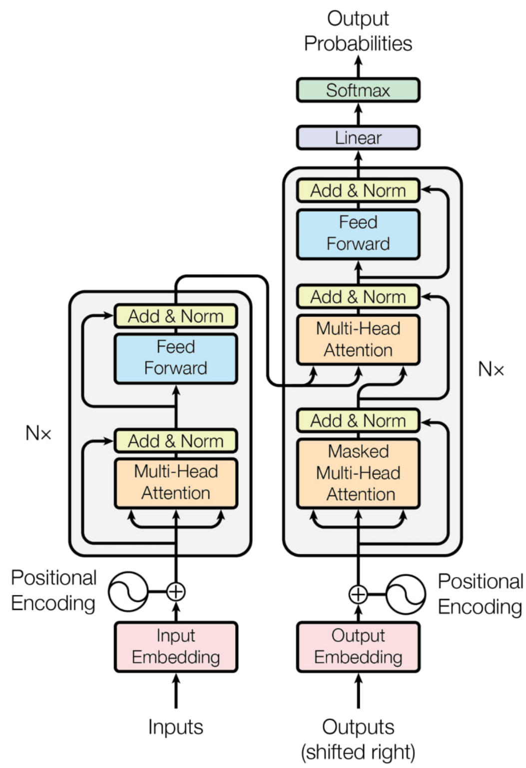


Figure 2.4: Transformer architecture

main components, each one presenting its own weight parameters: *encoder* and *decoder* block. The encoder is responsible for converting the input into a vector representation an it consists of a stack of multi-headed attention module followed by a fully connected (feed-forward) layer. Each multi-head attention module calculates the attention between all word pairs in the input, allowing the network to focus on specific parts of the text during processing. The attention layer's output is then combined to the initial embedded input and can been interpreted as a series of vectors representing the input in a contextual and information-rich manner. The decoder is fed with the vector representation generated by the encoder and produces the desired output. It also consists of a stack of multi-headed attention module and includes another attention block, called ”attention over the encoder outputs”, which allows the decoder to focus on the relevant parts of the input while generating the output. The decoder produces the output sequentially, using the information received from the encoder through the attention mechanism.

2.2.1 Tokenization and Embedding

Before diving into the details of how information is processed by the encoder and decoder of a transformer, it is essential to understand two crucial steps: *tokenization* and *embedding* of the input. Together, they provide the transformer with a structured and semantic representation of the input data, which can be efficiently processed by the Transformers blocks. This process allows the model to understand the context and relationship between words within the text, enabling accurate and robust translation, natural language analysis or other processing.

Tokenization procedure consists of transforming the text into a series of smaller units, called *tokens*, which represent the units of meaning within the text. There is not necessarily a univocal correspondence of a word with a single token, but each word could be mapped univocally to several tokens: a token may thus represent single words, parts of words or even special symbols and indicates the structure of the text depending on the tokenization scheme in use (WordPiece, Byte-Pair Encoding, etc.). Special tokens are then added to the tokenised text to provide additional information to the model. For example, a [CLS] token is added to the beginning of each sequence of text to indicate the start of a text instance, and a [SEP] token is added to the end of each sequence to separate phrases or sentences in the case of models that support such information. A [PAD] token is eventually added to bring the tokenised sentence length to a standard measure. For example, if we consider the simple sentence ”The jockey is riding the horse.” we will have the following deconstruction in tokens:

The jockey is riding the horse.

↓

[CLS, The, jockey, is, rid#, #ing, the, horse, . , SEP]

During the pre-training phase, a vocabulary with size M is created, containing all the possible tokens a model can use to represent the text. Machines do not understand language the way we represent it, so we need to convert the tokens into a comprehensible representation: each token is then converted into a sparse numeric vector called *one-hot encoding* vector, of size equal to the total number of tokens in the vocabulary M and where there is a single 1 corresponding to the assumed position m of the token within it. Such a representation, however, is neither efficient

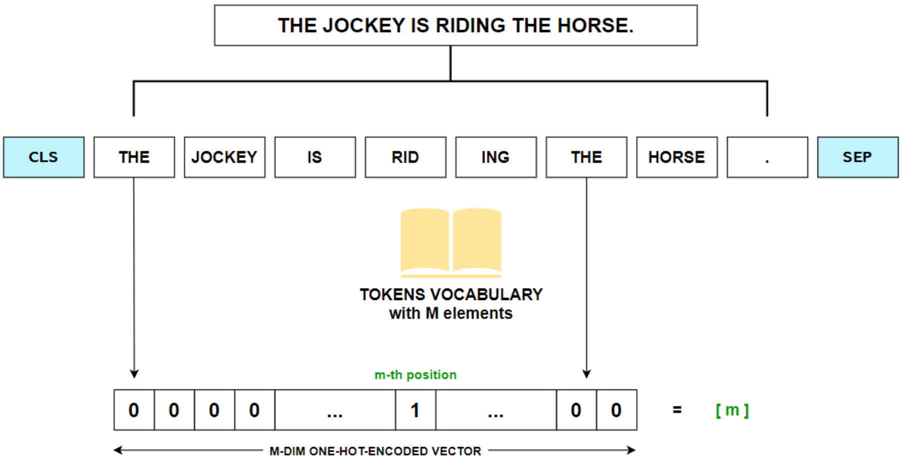


Figure 2.5: Sentence tokenization

nor optimal, because the number of elements for each encoded representation is exorbitant, as a vocabulary may have as many as hundreds of thousands of tokens. It is therefore necessary to subject these encoded vectors to a transformation that makes them more manageable and more information dense. There are many techniques that allow us to perform these operations, in this case we resort to *word embedding*.

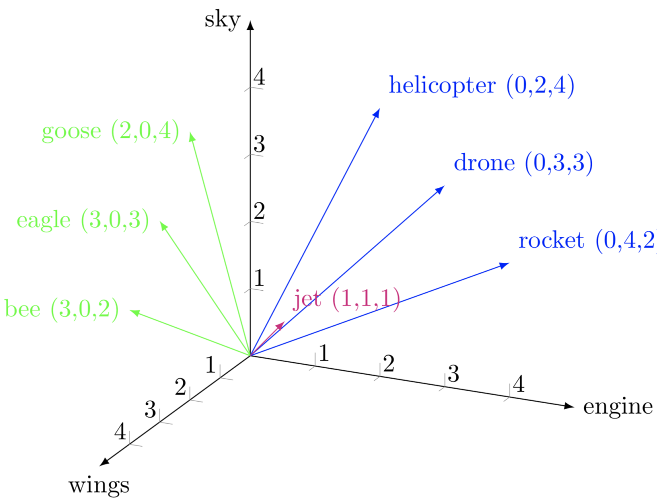


Figure 2.6: Words representation in the embedding space

Word Embedding captures the semantic meaning and context of tokens within the text: each token is mapped into a multidimensional vector space, where its position reflects its meaning and relationship to other tokens in the context of the text. An example, the tokens "goose" and "eagle" might have similar embeddings as they both refer to animals with wings, whereas the tokens "goose" and "helicopter" might have more distant embeddings as they represent semantically different concepts (see Figure 2.7).

But how exactly does this conversion take place? It starts with the sparse encoded representation of each token: each M-dimensional one-hot-encoded vector is passed as input to a pre-trained embedding layer which returns a new vectorial representation, resuming the token's semantic meaning. This computational component also participates in the training, so the weights are updated along with the parameters of the Transformers.

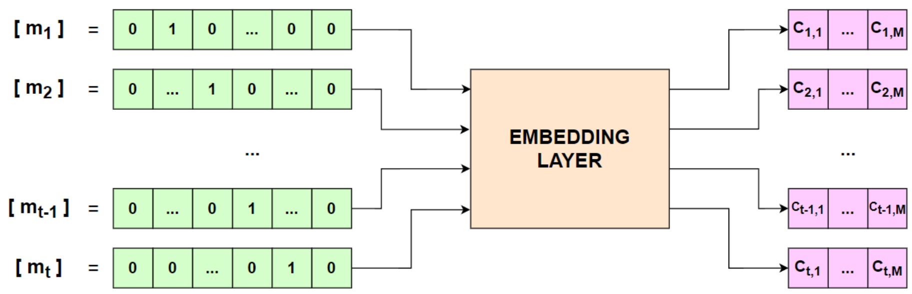


Figure 2.7: One-hot encoded vectors transformation into embedded vectors

Furthermore, we would like to be able to take into account not only the position of the token within the vocabulary (which provides a rather abstract definition of it) but above all the meaning it takes on within the sentence, its position and thus its connection with the other elements. Since Transformers models do not have a recursive structure, in order to understand the order of the words in the text and provide relative information on the position of the individual token within the entire sequence, a position vector called *positional encoding* is added. It is usually implemented as a position vector generated using sine and cosine functions.

When the input text is composed of several sentences, an additional information demand arises regarding the position each segment occupies in the corpus. This aspect is crucial in order to understand the relationship between elements on a lower granularity level and to keep track of the complexity of the entire text, perhaps in order to capture the intertwining of causality and semantic correlation between several periods. A further element must then be computed to obtain the final representation of the text: *segment embedding*. Each token is then associated with an identification code that defines its position in the text at sentence level, i.e. to which specific period it belongs. The combinations of embedded, positional and segment vectors associated with each token are concatenated to obtain the final embedded representation. If we set N as the upper limit of tokens in a sample text, the final representation is given by a matrix of size $N \times M$. Please refer to Figure 2.8 for a broad graphic representation of the process followed and the individual constituent elements.

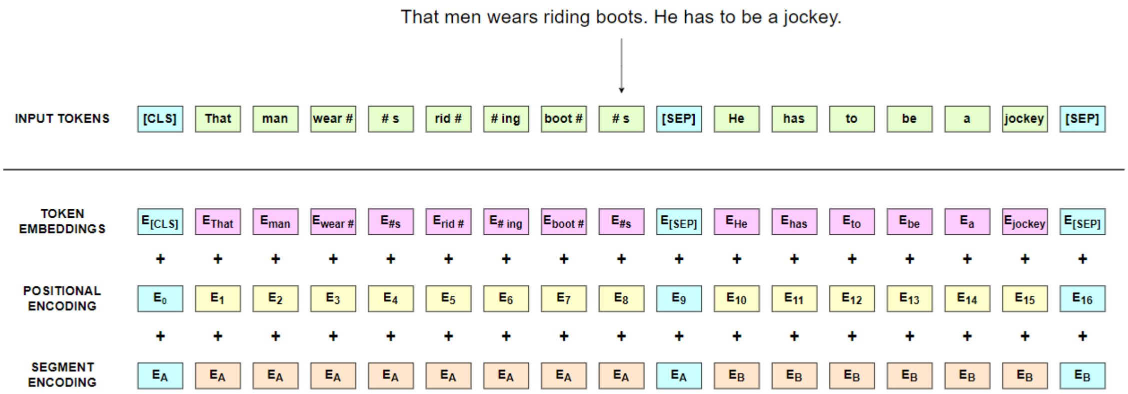


Figure 2.8: Final embedding

2.2.2 Attention Mechanism

Some of the limitations afflicting previous architectures have been addressed with the introduction of the *attention mechanism*. This is a key mechanism that has revolutionized the way information is handled as it allows the model to focus on specific parts of the input during the processing phase, ensuring greater precision and understanding of the context.

The operating principle is inspired by human cognitive processes related to attention, here defined as the ability to select and concentrate on certain stimuli or information, while ignoring or filtering out distractions. The main idea is that not all available information is useful in the same way, so we should only pay attention to the part we are currently interested in, ignoring everything else. A particular advantage of this mechanism is its versatility: the selection process can be dynamic and can vary depending on the context and the requirements of the task. Indeed, there are various variants of the attention mechanism, but the main idea is that the model assigns a weight to each part of the input based on its relevance to the current task. In practice, this mechanism amounts to assigning importance weights to each input or, in other words, applying a transformation matrix A in which higher number values correspond to information deemed worthy of attention. For instance, in the NLP tasks, it can focus on specific words or fragment of texts. During processing, the model calculates these attention weights and uses them to weigh the input before performing subsequent operations.

In particular, the use of the attention mechanism - in both the Transformer's encoder and decoder blocks - allows to manage multiple words in parallel and thus speed up the learning process considerably, compared to, for example, the aforementioned LSTMs.

An innovative aspect of the Transformers architecture that plays a decisive role is the presence of *multi-head attention* layers, but before going into details, it is useful to clarify the most basic concept of *self-attention*. In short, while self-attention focuses on the relationship between words within a single sequence while multi-headed attention expands on multiple representations of the sequence itself, allowing the model to capture richer and more complex relationships. But let us see them more in depth.

The *self-attention* mechanism, also known as intra-text attention, is a technique that allows each word (or token) within a sequence to directly influence all other words in the same sequence. In practice, this means that each token calculates a weight for each other token in the sequence, based on their semantic relationship. This weight determines how much the output of one token must "pay attention" to the content of the other tokens when calculating the final output. Self-attention is crucial in transformers as it allows the model to capture long-term relationships within an input sequence without depending on a recurrent network structure.

Multi-head attention is a generalisation of the self-attention mechanism. Instead of performing a single attention on one input, multi-head attention performs several attentions simultaneously on different input representations: the input is projected into different embedding spaces via projection matrices, called *heads*, and attention is computed separately for each head. The outputs of the different heads are then concatenated and again projected into a common space via another projection matrix. Multi-head attention enables the model to learn richer and more complex representations of the data, while allowing for greater parallelisation of the calculations.

In particular, three main characteristics with the corresponding linear transformation matrices are taken into account: W^Q for *query*, W^K for *key* and W^V for *value*. Given the input matrix X, the idea is to generate compressed information thus highlighting the correlations:

$$\begin{cases} X \cdot W_i^Q = Q_i \\ X \cdot W_i^K = K_i \\ X \cdot W_i^V = V_i \end{cases} \implies \quad \mathrm{MHA}\big(Q_i, K_i, V_i\big): \quad Z_i = \mathrm{Softmax}\Big(\frac{Q_i {K_i}^T}{\sqrt{d_{K_i}}}\Big) V_i \tag{2.2}$$

where $i = 1, \ldots n$. Instead of having just one matrix for each feature i, we can have more than one with the result of obtaining the corresponding number of Z where each attention Z_i focuses on one specific element:

$$Z = Z_1 Z_2 \ldots Z_n W^0. \tag{2.3}$$

What makes this conceptually so much more appealing than an LSTM Cell is that we can physically see a separation in tasks: the encoder learns what is a language, what is the relative grammar and more importantly what is context. The decoder learns how do starting language words relate to final language words. Both of these basic elements, even separately, have some underlying understanding of language and it is because of this understanding that we can pick up part this architecture and built systems that understand language.

If we line up many decoders one after another, we get the GPT transformer architecture; conversely if we stack just the encoders we obtain a Bidirectional Encoder Representation from Transformer (BERT).

2.3 Bidirectional Encoder Representative for Transformers (BERT)

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a pre-trained language model based on the Transformers architecture, first introduced by researchers at Google Research. The model was presented through a paper published in 2018 entitled "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" [3].

BERT has revolutionised the field of NLP (Natural Language Processing) through its ability to capture the bidirectional context of words in a text using an unsupervised learning approach. This model was pre-trained on huge amounts of unannotated text from different sources, such as Wikipedia, books and websites, in order to capture a wide range of linguistic knowledge. The main innovation introduced by BERT compared to previous models is its ability to consider the context on both the left and right of each word in a sentence during pre-training. This means that the model is better able to capture semantic and syntactic relationships between words, significantly improving performance on a range of NLP tasks, such as language comprehension, text classification and text generation. In addition, BERT introduced the idea of randomly masking some words in the input during pre-training, so that the model has to make predictions about missing words. This approach, known as token masking, allows BERT to gain a deeper understanding of the semantic relationships between words. Due to its effectiveness and ability to improve performance on a wide range of NLP tasks, BERT has become one of the benchmark models in the field of machine learning and NLP. It has inspired numerous variants and helped push forward the boundaries of research in this field. Its architecture is obtained by concatenating a variable number of Transformers' encoders: 12 for the basic variant (with about 110 million parameters) and 24 for the large variant (about 340 million parameters).

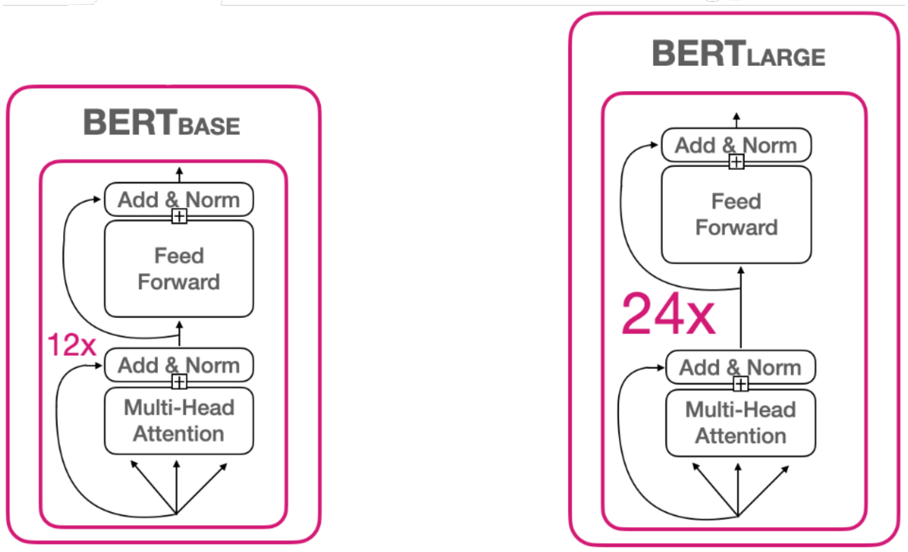


Figure 2.9: Bidirection Encoder Representation for Transformers architectures with different sizes

2.3.1 Fine Tuning

As it is a relatively complex neural network with a large number of parameters, it is not advisable to train it from scratch every time one wishes to approach a task with a specific dataset, as this would require an insane amount of time and resources. For this reason, BERT falls into that category of algorithms that lend themselves to the use of transfer learning. Transfer learning is a machine learning technique in which a pre-trained model on a dataset is further adapted or transferred to a different task or similar data domain. Rather than training a model from scratch for each specific task, transfer learning exploits knowledge learned from pre-trained models on large amounts of data to improve performance on new tasks or data domains. The key idea of transfer learning is that pre-trained models have already acquired general knowledge about the data and real-world characteristics, which can be used to help solve new tasks or similar data domains. By transferring this prior knowledge to a new task or data domain via the adaptation process, the model can achieve higher performance than training from scratch. There are several modes of transfer learning, in particular, *fine-tuning* is used to customise BERT training. In this mode, a pre-trained model is further adapted (fine-tuned) using data specific to the target task. This involves the continuous training of the model using labelled data related to the new task, allowing the model to adapt to task-specific details.

The steps to be followed to fine-tune a pre-trained model consist mainly in the preparation of the dataset (to be passed as input to the pre-trained neural network), the customisation of the architecture according to the specific task and the refinement of the parameters through training in the canonical sense. First of all, however, the pre-trained BERT model (or other architectures) must be chosen and loaded using prepared libraries, such as Hugging Face Transformers. Then the dataset must be adapted to the specific requirements of the model, possibly brought into the right size or format and finally be passed as input to the pre-trained neural network. It is usually necessary to customise the last layers of the network to adapt the model to the specific task, e.g. for classification. In this case, the size of the output layer will correspond to the number of classes in the target task. During training, the weights of the BERT model and the final classification layer will be updated using an optimisation algorithm to minimise the loss between model predictions and actual labels. If necessary, perform further fine-tuning iterations with different hyperparameter configurations or training strategies to further improve model performance.

2.3.2 BERT Pre-Training

The bulk of the learning process is concentrated in this phase. The model is instructed in the basic concepts needed to tackle NLP tasks: learning the basics of language, understanding what context is and how words relate to each other. It is structured in two simultaneous unsupervised procedures: *Masked Language Model* (MLM) and *Next Sentence Prediction* (NSP).

Before proceeding with the description of the two processes, it is necessary to mention that since BERT is a model based on the Transformers architecture and ideally designed to solve NLP tasks, the textual input has to be converted in the usual way so that it can be understood and handled by the compiler. Each sample text must then undergo tokenization, in this case using the *Word Piece Tokenization* technique, and each token obtained must be suitably transformed into a vector representation on the basis of semantic and positional information. The

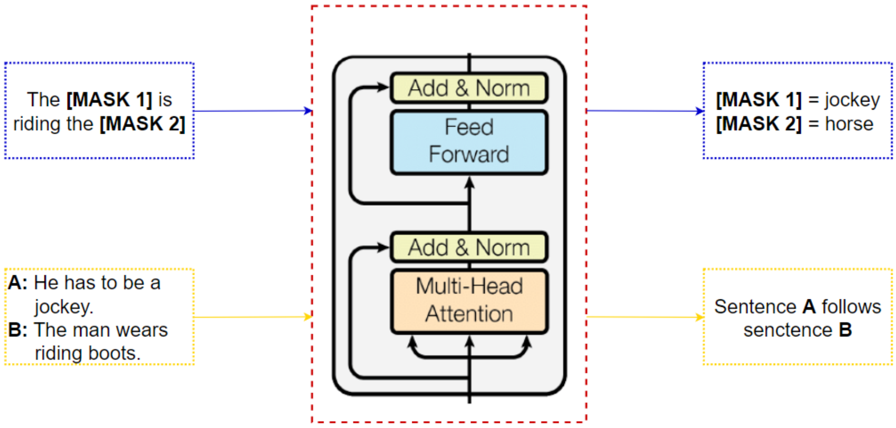


Figure 2.10: BERT pre-training steps: Masked Language Model (blue) and Next Sentence Prediction (yellow)

procedure follows the steps already described in subsection 2.2.1.

Masked Language Model   The MLM is a pre-training strategy that involves randomly masking certain tokens within a sequence of text and training the model to predict these masked tokens. This task is designed to help the model develop a deeper understanding of context and meaning of words within a sentence, being able to capture the linguistic dependencies between them. It helps generate more robust and contextually sensible language representations that can be used for a wide range of Natural Language Processing (NLP) tasks without the need for model-specific adaptation. The MLM model is exposed to a wide range of unannotated text from various sources, such as Wikipedia, books and web pages. Each text is then tokenized and divided into sequences of tokens that are used to train the model, some elements in the sequence are randomly masked and replaced with a special masking token (`[MASK]`). This is a kind of "fill the black" task, the model must then predict which words have been masked, based on the surrounding context provided by the other words in the sequence. For example, if we have the sentence "The jockey is riding the horse", a masked version of this sentence could appear as "The [MASK] is riding the [MASK]". The model must be able to predict which missing (masked) words best fit the context of the sentence, using information from surrounding words.

Next Sentence Prediction   The NSP is another pre-training task used to gain a deeper understanding of the context and structure of natural language, the training of which takes place *simultaneously* with the MLM. This is designed to teach the BERT model to interpret the relationship between two consecutive sentences within a text, perhaps by grasping a causal and correlation relationship. Here is a more detailed explanation of the Next Sentence Prediction task in BERT: S starts by considering pairs of consecutive sentences taken from the corpus of unannotated text, named sentence A and sentence B. For each pair, two versions of input are created that will constitute training examples for the model: in the first version, A and B are concatenated and used as input, in the second version sentences are concatenated with a probability of 50% and a random sentence is chosen as sentence B. Each training example is labelled with a binary label indicating whether sentence B actually follows sentence A in the original sequence: if sentence B is the direct subsequence of sentence A, the label will be 1, otherwise it will be 0. The model must therefore predict whether sentence B follows sentence A in the original sequence so that BERT will be able to understand and predict the relationship between two consecutive sentences, developing a deeper knowledge of the structure and context of natural language.

Figure 2.11 shows a schematic representation of what happens during the pre-training phase.

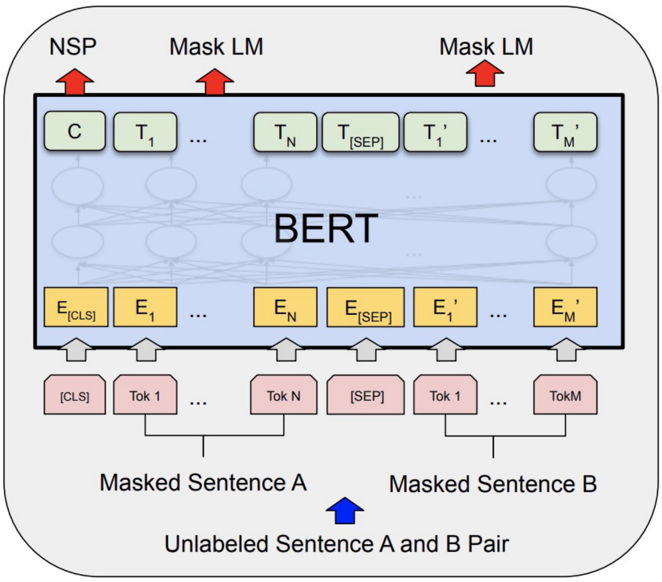


Figure 2.11: Masked Language Model and Next Sentence Prediction training

2.4 XLM-RoBERTa

XLM-RoBERTa is a Transformer-based model introduced in November 2019 in the paper "Unsupervised Cross-lingual Representation Learning at Scale" by Facebook AI [4]. As the name suggests, it is a variant of the RoBERTa architecture (Robustly optimised BERT approach), which in turn is an improved version of BERT (Bidirectional Encoder Representations from Transformers) developed by Google. The acronym "XLM" stands for "Cross-lingual Language Model", indicating its ability to handle multiple languages, and one of its greatest strengths lies in the fact that, unlike other XLM multilingual models, it does not require language tensors to understand which idiom is used and should be able to correctly determine it from the input ids. This is made possible by the use of segmentation tokens and specific pre-processing methods to handle multilingual texts in a single instance of the model. Indeed, XLM-RoBERTa is a large multi-lingual language model with trained on a massive information: about 2.5TB data in 100 different languages, taking its cue from RoBERTa's massive pre-training which involved text corpora of several languages from Wikipedia, Common Crawl and other online text sources. The training was performed using the technique of supervised learning by self-learning, where the model attempts to maximise the prediction of the next or hidden word within a sequence of text. This process is known as pre-training and can be performed

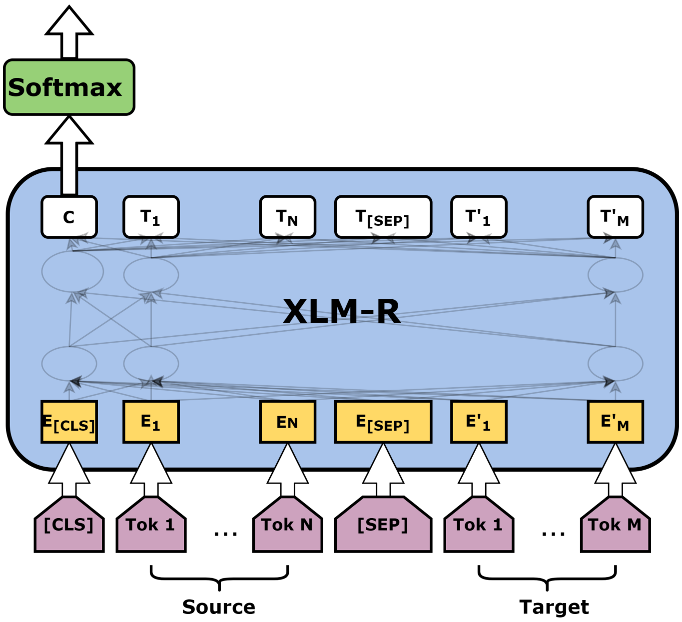


Figure 2.12: XLM-RoBERTa representative architecture scheme

on huge amounts of not annotated text. XLM-RoBERTa has demonstrated excellent performance in a range of multilingual NLP tasks, including named entity recognition, sentiment analysis, machine translation, text generation and more. Its performance has been tested on a wide range of languages, demonstrating good generalisation ability and solid cross-linguistic understanding.

Analogies and Differences with previous architectures A first interesting aspect to explore concerns the links that this architecture has with its predecessors, RoBERTa and BERT. XLM-RoBERTa in fact stands as an evolved version of the RoBERTa architecture, adapted for the precise purpose of addressing NLP tasks in a multitude of languages. Both have contributed significantly to the advancement of Natural Language Processing, offering superior performance on a wide range of tasks however, while sharing a common origin with BERT, they have distinctive differences that make them unique in their approach. In fact, RoBERTa focuses mainly on the English language while XLM-RoBERTa is specifically designed to support multilingual NLP by including texts in several languages in its pre-training. In this regard, RoBERTa uses segmentation tokens [SEP] to separate sentences, while XLM-RoBERTa uses specific segmentation tokens to handle multilingual texts, allowing the model to understand the relationships between sentences in different languages. The structural nature of the two neural networks, however, is the same: both are based on transformer architecture and use self-attention mechanism layers to capture long-range relationships between words in a text, thus improving contextual understanding (see Subection 2.2.2 to explore the theory behind this mechanism).

That said, it remains to be understood what the differences are with the common progenitor BERT, with which both architectures share many key features and at the same time have several key differences. A first distinction concerns the size of the corpus: BERT was trained on a large text corpus of English language texts from Wikipedia and digitised books, but RoBERTa-based models have generally undergone more extensive pre-training, using a larger text corpus with a variety of additional sources, including texts from Common Crawl. A second very important difference concerns pre-training. In the subsection 2.3 it was explained how BERT uses a training technique based on two simultaneous tasks: MLM (Masked Language Modeling) and NSP (Next Sentence Prediction), where the model tries to predict missing terms and whether two sentences are consecutive or not. RoBERTa, on the other hand, uses dynamic pre-training, including masking elimination (MLM) and two-way training (NSP) for the entire training process [5]. Using the entire two-way context for each token without masking some words improves the generalisation of the model. Finally, in order to optimise the training procedure and improve performance on a number of NLP tasks, RoBERTa modified some hyper-parameters compared to those of BERT, such as the number of batch sizes, the number of pre-training steps, and the learning rate.

Dynamic Masking In the architecture of BERT, we recall that during pretraining, BERT engages in language modeling by attempting to predict a certain proportion of masked tokens. However, a drawback of the original implementation lies in the repetitive masking of identical tokens across different batches within the training dataset. To elaborate, the training dataset is replicated tenfold, resulting in each sequence being masked in only ten distinct ways. Considering BERT undergoes 40 training epochs, sequences with identical masking are presented to BERT four times each. Researchers have discovered that employing dynamic masking, whereby masking is uniquely generated each time a sequence is fed into BERT, yields slightly superior results. This approach reduces the repetition of data during training, allowing the model to interact with a more diverse array of data and masking patterns.

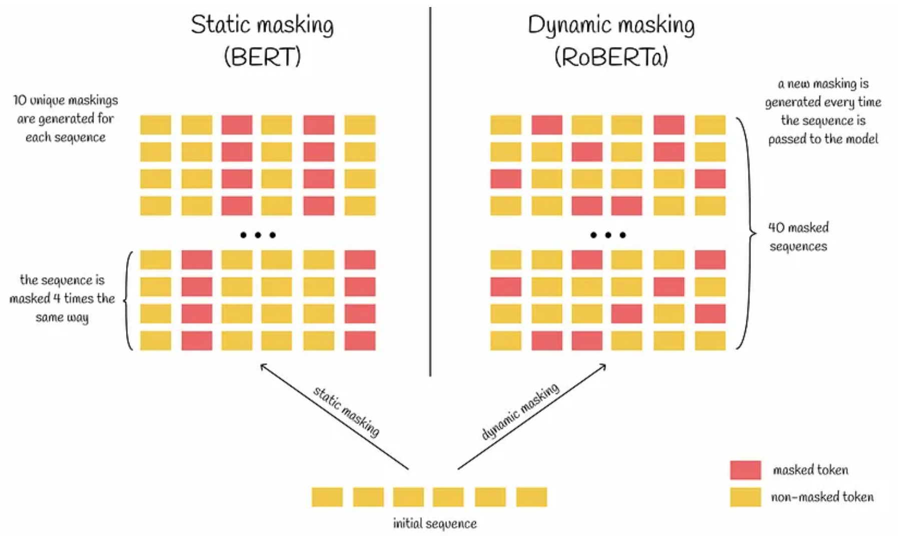


Figure 2.13: Static vs Dynamic Masking

Next Sentence Prediction According to the findings in the article "RoBERTa: A Robustly Optimized BERT Pretraining Approach," [5] removing the next sentence prediction loss yields slightly better performance. Conversely, inputting single natural sentences into BERT input detrimentally affects performance compared to sequences comprising multiple sentences. This phenomenon is likely due to the challenge for the model to learn long-range dependencies solely relying on individual sentences.

Constructing input sequences by sampling contiguous sentences from a single document proves more advantageous than sampling from multiple documents. Typically, sequences are built from contiguous full sentences of a single document, ensuring a maximum length of 512 tokens. However, when reaching the end of a document, researchers explored whether halting sentence sampling or additionally sampling the first several sentences of the next document (with a corresponding separator token) would be more beneficial. The study revealed that stopping at the document boundary is preferable.

In the final RoBERTa implementation, the authors retained the first two aspects and omitted the third. Despite the observed improvement from the third insight, researchers decided against its inclusion to maintain consistency in comparisons with previous implementations. Introducing document boundaries and stopping at them would result in input sequences containing fewer than 512 tokens. To maintain consistent batch sizes across all sequences, batch sizes would need augmentation, leading to variable batch sizes and complicating comparisons. Hence, the authors opted to maintain uniformity in batch sizes for simplicity in comparisons.

2.5 mBART

The mBART (Multilingual BERT for Question Answering) architecture is a multilingual language model developed by Facebook AI and presented in 2020 in the article 'Multilingual Denoising Pre-training for Neural Machine Translation'. [6].

It is based on the Transformers architecture and in particular on BERT (Bidirectional Encoder Representations from Transformers) and represents a breakthrough in multilingual natural language processing, especially in the question-answer task. This architecture uses layers of transformer encoders to capture complex semantic relations in textual data, focusing on bidirectionality of operations, which means that the model can consider both preceding and subsequent contexts during the training process. With BERT it also shares the organisation of pre-training (MLM and NSP described in 2.3) but one of the distinguishing features of mBART is its predisposition to multilingual tasks: in fact, it is trained on a large corpus of text that includes data in many different languages, allowing the model to understand and generate texts in several languages without the need for specific adaptations. The mBART model is specifically designed for the multilingual question-answer task but lends itself to customisation in use thanks to a fine-tuning process, during which it takes into account the peculiarities of the different languages supported. For example, it can be adapted for more specific tasks such as the automatic translation of texts from a source language and a target language, both of which have to be specified.

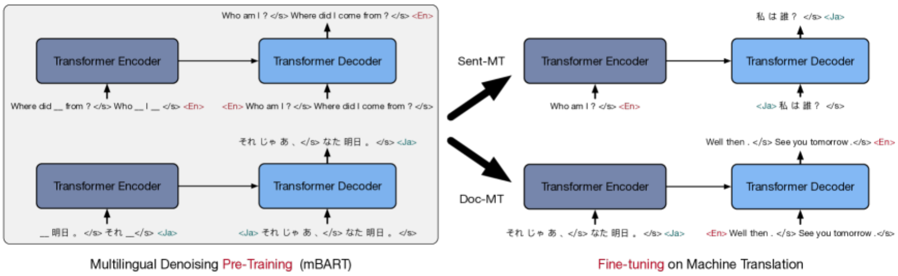


Figure 2.14: mBART pre-training and fine-tuning on question-answering taskscheme

mBART Denoised Training Unlike traditional translation models, mBART underwent a unique training method known as *denoised training*. This methodology was initially introduced in another language model called BART, often regarded as mBART's precursor (hence, mBART is also referred to as Multilingual BART). In denoised training, the model deviates from the conventional approach of directly translating sentences: it learns to reconstruct the original version of a sentence from a deliberately corrupted or noisy variant. This noise is introduced by perturbing the original sentence, such as removing certain tokens or substituting random characters. This technique enables the model to develop an internal representation of the language's structure and semantics. The primary divergence between mBART and BART lies in the scale of denoised training. While BART was trained on a single language, mBART's denoised training incorporates 25 diverse languages. Subsequently, the model undergoes fine-tuning using 24 bilingual corpora, each comprising English and one of the other 24 languages. This extensive training regimen transforms mBART into a potent translation model capable of seamlessly translating between multiple languages.

Translation Performances The performance of mBART was evaluated on various machine translation and natural language understanding tasks, showing promising results in many different languages. Should one wish to use this architecture in translation tasks, one should be aware of its merits and shortcomings.

The massive multilingual pre-training has given the model an excellent readiness for tasks of this type: thanks to multilingualism, the quality of translations is found to be more than good over a wide range of languages, surpassing other machine translation models in some cases. The mBART model provides a fairly complete language coverage, proving to work well even with less widely spoken languages and thus making itself useful for non-standard translations. Like other models based on the Transformers architecture, mBART is able to maintain context and semantic consistency during translations, which leads to more accurate and comprehensible results. It also retains a fair amount of flexibility: the model can be easily adapted to new language pairs by fine-tuning on bilingual corpora, making it suitable for an even wider range of translation contexts. Precisely due to its potential, mBART is a rather large and complex model, which may require significant computational resources for training and implementation: the use of mBART for real-time or large-scale translations may require considerable computational resources, especially in resource-limited environments, and if one wishes to customise it, additional good-quality data are required to achieve optimal results.

In summary, mBART is a powerful and versatile model for multilingual machine translation, with promising results in terms of translation quality and language coverage. However, it is also important to consider the computational costs and adaptation required to achieve the best results in different applications.

3

Classification Tasks Presentation

The ultimate goal of the work is to analyze the similarities and differences in the reporting of relevant topics of international interest in online news articles aimed at a readership of different nationalities, therefore in multilingual setup. To fulfill the purpose, it was necessary to develop robust information extraction models at different levels of granularity. Specifically, based on what was proposed in the SemEval-2023 competition[*], three different supervised tasks were addressed in this work.

1. *Document Categorization*: classify an article at document-level by choosing from 3 possible categories: opinion, reporting and satire (multi-class classification task).

2. *Framing Detection*: associate to a news article one or more framing labels appearing in the whole text at least one time, among 14 possible domain-independent perspective frames (multi-label text classification task).

3. *Persuasion Techniques Detection*: identify which persuasion techniques occur in a news online article at text level (multi-label classification task at text level) and extract the portion of text in which one or more labels were used among 23 communication strategies (multi-label token classification)[†].

In the following three subsections (one per macro task), each of the classes and labels considered will be discussed in detail individually, providing a definition and examples.

[*] https://semeval.github.io/SemEval2023/

[†] Often in the work we will use span token level interchangeably. This is because at classification level one works with tokens but the aim is to identify the phrases, i.e. the spans in which a technique occurs.

3.1 News Categorization

Categories are mutually exclusive and define the nature of the entire paper, providing the reader an introductory clue about the style and narrative approach adopted for a given topic's presentation. It follows that only one of the three categories considered can be assigned to each document.

3.1.1 Opinion

Opinion articles reflect the someone's personal point of views, thoughts or feelings regarding a specific topic. These sometimes may include analysis or quotations, usually partial and accompanied by comments or criticism that make the writer's position clear, relating to current common interest issues or events with the aim of corroborating a thesis, supporting or discrediting a side. Therefore, such articles are characterized by a more or less subjective narrative slant, typically influenced by ideological, political, religious etc orientation. An example:

> The Jewish groups said this anti-Semitic rhetoric "was used to encourage students to vote specifically against Noah Lew. It is under this context that the (general assembly) occurred, and the report fundamentally misunderstands this, which alters the entire findings of the report." They say Boudreau's report "insinuates that Jewish students who engage with mainstream Jewish community organizations are permitted to be precluded from holding political office."

3.1.2 Reporting

A news reporting provides detailed, objective information about events, facts, or issues of public interest. It usually follows a clear, direct and unbiased writing style with a linear and orderly narrative structure. The author provides a comprehensive overview of the situation, clarifying cause-and-effect relationships, sharing direct quotes or interviews with experts, witnesses, or people involved in the topic, without neglecting relevant data, statistics, or evidence because the goal is precisely to *report* to the reader everything there is to know to enable them to formulate their own opinion. The borders between *opinion* and *reporting* might be sometimes blurred, however, the categorization requirements of the latter are significantly more stringent since it is a piece of writing that must be as objective as possible. For this reason, articles that contain even a single sentence which is or could be interpreted as the author's opinion on the specific matter, a speech, an interview or a conference with the opinion of a single politician or expert without reporting the counterpart reply, should be labelled as opinion. An example:

> CLEVELAND — Police investigating domestic disputes had previously gone to the home where a man fatally shot two police officers over the weekend, but no arrests were ever made, police reports from the Columbus suburb of Westerville show. Westerville Officers Eric Joering, 39, and Anthony Morelli, 54, were killed shortly after noon Saturday in this normally quiet suburb while responding to a 911 hang-up call. The suspect, 30-year-old Quentin Smith, was shot and wounded by the officers and taken to Ohio State University Wexner Medical Center in critical condition Saturday.

3.1.3 Satire

A satirical journalistic article is a type of writing that uses humor and irony to comment on or criticize events, people, or situations, focusing more on entertainment through parody and caricature. The purpose is not to report but to provoke and point out shameful, corrupt or hypocritical behaviors, ridiculing otherwise serious topics or decision-making figures through a variety of expedients: exaggerations, absurdities or paradoxical situations. The style is definitely unmistakable as well as bursting by definition, so a simple satirical phrase or joke, perhaps in closing, is not enough to classify the entire document as satirical. An example:

> Bill Maher has said he doesn't need the Robert Mueller report to know President Donald Trump is a traitor "because he has a TV". Speaking on his Friday night show the Real Time host said: "I must say, I don't think it looks good." No further indictments, which means not Don Jr., even after the "I love it" memo, really? Not Jared, not Manafort or Stone for working with the Russians. "Did the Democrats put too must trust in the Mueller report? Because I don't need the Mueller report to know he's a traitor. I have a TV."

3.2 Framing Detection

Framing, in the world of communication, refers to the selection and presentation of information in a way that influences the public's perception of a specific topic. It is a process by which the media interpret and represent events, issues or themes in a way that promotes certain views, values or goals. This communicative principle works at the level of the whole document; in fact, to characterize the article, it is sufficient to present a news item in such a way that the reader's attention is focused on particular perspectives of interest (framing) even only once.

3.2.1 Capacity and Resources

This framing identifies parts of the articles referred to the availability of physical, human or financial resources, and capacity of current systems. An example:

> "Madagascar, typically like many African countries, doesn't have many doctors. There are around three-and-a-half thousand doctors for 22 million people. They only have around 6,000 hospital beds, so they aren't particularly well positioned to cope with these kind of events."

3.2.2 Crime and Punishment

This framing identifies parts of the articles referred to effectiveness and implications of laws and their enforcement. An example:

> Don Lemon wanted to defend the "peaceful" migrants who are headed toward the US border to illegally enter the country against our immigration laws, but said that the "biggest terror threat in this country is white men, most of them radicalized to the right." How can he contradict himself like that and not even bat an eyelid?

3.2.3 Cultural Identity

This framing aims to focus the reader's attention on traditions, customs or values embraced by a social group in relation to a policy issue. An example:

> And she was quoted as saying that Abedin "feels a deep responsibility to encourage more mutual understanding between her beliefs and culture and American culture."

3.2.4 Economic

The reader's attention is directed toward strictly economic issues like costs, benefits or other financial implications. An example:

> She may be given May 23, the day of EU elections, as a compromise but only if her deal passes the British parliament. A no-deal crash out on March 29 would create utter chaos for months. It would be catastrophic for Britain's economy.

3.2.5 External Regulation and Reputation

This type identifies parts of the articles in which international reputation or foreign policy are put in a spot. An example:

> Soros compared Orban unfavorably to both the Nazis and the Communists, saying his rule evoked dark tones from the 1930's — when Hungary was allied with Nazi Germany — and was more oppressive than Cold War Soviet occupation. Orban has tightened the screws on non-government organizations, particularly ones funded by Soros, and attempted to close a prominent Soros-founded university. Attributing to Soros a recent United Nations plan on creating a global blueprint to handle the migration crisis, Orban said he anticipated that powerful allies would help him prevent the U.N. from greasing the wheels of migration.

3.2.6 Fairness and equality

This type identifies parts of the articles referred to the balance or distribution of rights, responsibilities and resources. An example:

> In May, according to the Household Survey, total employment rose by 105,000. Non-Hispanics actually gained ground:
>
> - Total employment: up 105,000 (+0.08 percent)
> - There are Non-Hispanic employment: up 190,000 (+0.16 percent)
> - There are Hispanic employment: down 85,000 (-0.42 percent)

3.2.7 Health and Safety

This type identifies parts of the articles referred to health care, sanitation, public safety. An example:

> The World Health Organization (WHO) said on Tuesday that it was taking steps to help deal with a new outbreak of Ebola in the Democratic Republic of Congo's rural northwest, after two cases of the deadly virus were confirmed in the market town of Bikoro.
>
> Congo's Health Ministry said two of the five samples it sent to the National Institute of Biological Research in Kinshasa, came back positive for the disease. The samples were gathered after health officials in Equateur Province notified Kinshasa on May 3 of about 21 cases of a hemorrhagic fever in the Ikoko Impenge area, including 17 deaths, according to WHO and Congo's government. What is Ebola? Rare but deadly, the viral disease is most commonly affecting primates and humans. Initial symptoms can include fever, headache, joint and muscle aches, weakness, diarrhoea, vomiting, stomach pain lack of appetite and in some cases internal and external bleeding, according to WHO. Where did it originate from?

3.2.8 Legality, Constitutionality and Jurisprudence

All those references to rights, freedoms, and authority of individuals, corporations and government belong to this framing. An example:

> President Donald Trump's attorney, Rudy Giuliani, said he and the president's other lawyers are confident that there is no finding of collusion by the president.

3.2.9 Morality

This framing affects all articles presenting positions related to the moral sphere, with religious or ethical implications. An example:

> Morality really matters when it comes to serving in public office, and we are not going to send people that engage in sexually inappropriate behavior to Washington anymore."

3.2.10 Policy, Prescription and Evaluation

This framing is referring to all those articles where discussion of specific policies aimed at addressing problems. An example:

> San Francisco prosecutors, who had long ago deprioritized marijuana charges, dismissed the decades-old charge and released Garcia Zarate on April 15, 2015. Due to San Francisco's policy of limiting cooperation with federal immigration — which some refer to as a "sanctuary" policy — the city did not inform ICE when they released Garcia Zarate.

3.2.11 Political

This framing includes all those references to political context including lobbying, elections, and attempts to sway voters. An example:

> Trump tweeted last month that he had selected Haspel to replace Mike Pompeo, who is being considered to be secretary of state. But the president had not sent the formal paperwork to Capitol Hill.

3.2.12 Public Opinion

This framing identifies the articles referred to attitudes and opinions of the general public, including polling and demographics An example:

> A government spokesman said individuals whose presence "is not conducive to the public good" could be excluded by the home secretary. He added: "We condemn all those whose behaviours and views run counter to our shared values and will not stand for extremism in any form."

3.2.13 Quality of Life

THis framing identifies portions of articles referred to threats and opportunities for the individual's wealth, happiness, and well-being. An example:

> After Hurricane Maria barreled through Puerto Rico in September 2017, it left hundreds of thousands of people displaced and 80 to 90 percent of homes destroyed in some communities. But even before the hurricane, housing in the U.S. territory—where 43.5 percent of people live below the poverty line—was in crisis, and many homes on the island were built with salvaged fixtures and without permits, insurance or inspections.

3.2.14 Security and Defense

This framing focus the reader's attention on threats to welfare of the individual, community or nation. Examples: An example:

> Ambassador Haley accused the Iranian regime of continuing to "play" the Security Council. "Iran hides behind its assertion of technical compliance with the nuclear deal while it brazenly violates the other limits on its behavior. And we have allowed them to get away with it. This must stop." Ambassador Haley proceeded to list various violations by the Iranian regime of Security Council resolutions pertaining to the transfer of conventional weapons from Iran and the arming of terrorist groups, including the Houthi rebels in Yemen and Hezbollah. She also pointed to what she called the Iranian regime's "most threatening act" – its launch of ballistic missiles capable of carrying nuclear weapons. "When a rogue regime starts down the path of ballistic missiles, it tells us that we will soon have another North Korea on our hands," Ambassador Haley said. "If it is wrong for North Korea to do this, why doesn't that same mentality apply to Iran?"

3.3 Persuasion Techniques Detection

Persuasion is the art of changing the other's attitude or behavior through the tools of verbal communication, both oral and written. The calibrated use of language and the correct selection of words aim to influence the audience's thinking and actions in order to steer them in a specific direction, often by leveraging emotional and affectional involvement. Persuasion techniques are many, organized on multiple levels of complexity and depth, each characterized by a specific logical, syntactic and semantic structure. In this paper, 23 persuasion techniques were identified - in annotation phase at first and in classification task solving then - each one extracted from 6 high level macro-groups, which will be briefly defined below along with a quick introduction of persuasion techniques.

- *Call*: a statement devoid of argumentative structure or precise language employs words or phrases that are either non-neutral, confusing, exaggerated, loaded, etc., with the intent of influencing the reader, often on an emotional level. For this first macro-group, the following three techniques were taken into account:
  - Appeal to Time
  - Conversation Killer
  - Slogans

- *Manipulative wording*: a statement is deprived of argumentative structure or the language used employs words or phrases that are non-neutral, confusing, exaggerated, loaded with polarizing sentiments, etc., with the intent to influence the reader, often to playing on the emotional factor. For this second macro-group, four persuasion techniques were considered:
  - Exaggeration or Minimization
  - Loaded Language
  - Repetition
  - Obfuscation, Vagueness and Confusion

- *Attack on the reputation*: rather than dwelling properly on the actual subject of the discussion, the attention is decentralized to entirely unrelated topics typically with the aim of questioning or undermining a participant's credibility, for example by dwelling on his or her personality, experience, and actions. The subject of the argument may also refer to a group of individuals, an organization, or an activity. Five persuasion techniques were considered for this macro-section:
  - Appeal to Hypocrisy
  - Casting Doubt

  - Guilt by Association
  - Name Calling or Labelling
  - Questioning the Reputation

- *Distraction*: this macro-category includes all those devices aimed at distracting or redirecting the reader's attention. Here, three techniques were highlighted:

  - Red Herring
  - Strawman
  - Whataboutism

- *Simplification*: a statement, situation or event is presented in a simplified way, details or any cause-effect relationships are deliberately ignored. The partiality of information during exposition activates the emotional response of the reader, who is thus easier to convince, to polarize, to direct toward the persuader's thesis. Three calling techniques were considered:

  - Causal Oversimplification
  - Consequential Oversimplification
  - False Dilemma or No Choice

- *Justification*: as the name itself suggests, the idea is to provide a justification, not necessarily by resorting to logical arguments, to support a previous statement. For this last macro-section, five persuasion techniques were considered.

  - Appeal to Authority
  - Appeal to Fear and Prejudice
  - Appeal to Popularity
  - Appeal to Values
  - Flag Waving

In the following paragraphs, each of the 23 persuasion techniques mentioned will be described in detail and accompanied by a few examples of their application.

3.3.1 Appeal to Time

The strategy lies in generating a sense of urgency in the reader, sometimes distorting the perception of time and emphasizing the importance of immediate action ("Act Now!"). A couple of examples follow:

> All nations should contribute to the fight against climate change by reducing their emissions as quickly as possible .
>
> Wenn bis zum nächsten Monat der Prozentsatz älterer Menschen, die geimpft sind, 50 Prozent übersteigt, bestehen gute Chancen, dass die Belastung der Krankenhäuser abnimmt. [POLISH]

3.3.2 Conversation Killer

Words or phrases that deter critical thinking and meaningful discourse on a particular subject. This form of loaded language, often disguised as common wisdom, aims to prematurely conclude debates and alleviate cognitive dissonance.

> L'inflazione è fuori controllo, sarà una tragedia per la gente . [ITALIAN]
>
> Les ressources ne sont pas suffisantes et pour, les chercheurs ont une explication tout à fait valable: "On ne peut pas sauver tout le monde" .

[FRENCH]

3.3.3 Slogans

A concise and memorable statement often characterized by labeling and stereotyping. Slogans are designed to evoke emotions and can serve as powerful appeals to sentiment.

> начительную часть товаров, например, из Шанхая в Роттердам: [RUSSIAN]
>
> Нет льда — нет законов
>
> Il cancelliere tedesco, povera stella: toglietemi tutto ma non il gas russo! [ITALIAN]

3.3.4 Exaggeration or Minimization

This strategy is developed in two variants but they share the communicative principle: distort the impact and magnitude of an event. In the first case, the approach is based on exaggerating something excessively, amplifying its qualities or significance; in the second one, on diminishing its importance or scale, disregarding statements and accusations made by an opponent and downplaying their significance.

> Imigracja w Europie stała się nie do opanowania, w tym tempie w naszym kraju będą setki tysięcy nielegalnych imigrantów . [POLISH]
>
> It is an insignificant expense given the company's budget

3.3.5 Loaded Language

Utilizing precise language with potent emotional connotations, either positively or negatively, to sway and persuade the audience of an argument's validity or truth. This fallacy is alternatively referred to as euphemisms, appeal to emotion or argument from emotive language. It may happens this technique is accompanied by another, reinforcing its effectiveness.

> Le ridicole sanzioni contro la Russia hanno infatti rafforzato la cooperazione tra Russia, Bielorussia, Armenia, Tagikistan, Kirghizistan Repubblica Popolare Cinese. [ITALIAN]
>
> Armia najeźdźców, którą należy odeprzeć. Za pomocą broni. Zamiast być witanym przez byłych wojskowych w upale . [POLISH]

3.3.6 Repetition

A fallacy in which the speaker repetitively uses the same word, phrase, story, or imagery in the expectation that the audience will be more easily persuaded.

> Bien sûr! Ignorer, ignorer et ignorer encore les protestations des travailleurs honnêtes. [FRENCH]
>
> Stiamo mantenendo, ad esempio, migliaia di figli di albanesi che arrivano qui senza visto e poi si presentano ai comuni dichiarandosi 'non accompagnati'. Le coop fanno soldi. Tanti soldi . [ITALIAN]

3.3.7 Obfuscation, Vagueness and Confusion

Employing deliberately ambiguous language to allow the audience to form their own interpretations. This occurs when unclear phrases with multiple or vague definitions are used within the argument, ultimately failing to support the conclusion. Statements that intentionally lack precision or evade fully addressing the posed question also fall within this category.

> Voyons combien de temps ce texte restera publié, sans être retiré, comme le précédent sur ces programmes américains de financement des laboratoires biologiques militaires en Ukraine.
> [FRENCH]

> Стоит дважды подумать, прежде чем ввязываться в войну, естественным финалом которой станет смерть изгоя и посмертная демонстрация бутылки коньяка Hennessy бойцами групп "Альфа" или "Вымпел".
> [RUSSIAN]

3.3.8 Appeal to Hypocrisy

The technique targets the reputation of its subject by accusing them of hypocrisy or inconsistency. A well-known illustrative appellation with a Latin origin that easily sums up the strategy is "*tu quoque*", i.e. "also you", since it consists of pointing out that those who criticize you for something you have done have also behaved similarly in the past. This accusation may be made overtly by directly calling out hypocrisy, or it may be implied by highlighting contradictions between past positions or actions.

> Таковы обвинения, выдвигаемые Западом. Возможно, он забыл о своем поведении в последние годы.
> [RUSSIAN]

> Die Ukraine galt in Washington noch vor kurzem als eines der korruptesten Länder; jetzt wird sie mit Geld und Waffen überschwemmt.
> [POLISH]

3.3.9 Casting Doubt

Instead of presenting a relevant argument concerning the topic at hand, casting doubt on the character or personal attributes of an individual or entity is utilized to undermine their overall credibility or quality. This approach may involve discussing the target's professional history as a means to discredit their argument. Additionally, casting doubt can be achieved by referencing past or anticipated actions or events of an entity that have failed to achieve their intended objectives or appear likely to result in failure.

> L'insipienza della UE lascia esterrefatti . Le sanzioni saranno terrificanti per il popolo europeo. [ITALIAN]

> Le programme climatique de J. Biden était un échec programmé, notamment parce qu'il aurait rendu les États-Unis dangereusement dépendants des énergies renouvelables.
> [FRENCH]

3.3.10 Guilt by Association

It is also known as *Reductio ad Hitlerum* and it consists on targeting the opponent or an activity by linking it to another group, activity, or concept that carries strong negative connotations for the intended audience. It is crucial to note that this tactic is not confined solely to references to that particular group but, more precisely, this can involve asserting a connection or similarity between the subject of the technique and any individual, group, or event—whether present or historical—that is universally regarded in a negative light (such as being deemed a failure) or is portrayed as such.

> And anyone who is against this principle is nothing but a communist!

> Их заставляли носить евреев, как на территориях, контролируемых нацистской Германией во время Холокоста.
> [RUSSIAN]

3.3.11 Name Calling or Labelling

A technique of argumentation involving the use of loaded labels aimed at an individual or group, often in a derogatory or belittling manner. These labels portray the target object as either something the intended audience fears, despises, or conversely desires or admires. Belonging to *Pathos persuasion strategies*, this method relies on subjective and emotional judgment, disregarding factual evidence and focusing solely on the essence of the subject being labeled. It operates as a manipulative form of expression, functioning within the realm of nominal groups rather than presenting a fully developed argument with premises and conclusions. In political discourse, for example, this technique commonly employs adjectives and nouns as labels pertaining to political leanings, opinions, personal traits, organizational affiliations, as well as insults. What sets it apart from the Loaded Language technique is its exclusive focus on characterizing the subject, rather than manipulating the overall tone or emotional impact of the discourse.

> Zuerst hat man in Deutschland und Europa auf den Irrweg der "grünen" Energien gesetzt, der 2021 seine Schwächen gezeigt hat, als im Sommer der Wind für die Windkraftanlagen ausgeblieben ist, was zu der Gaskrise ab Oktober 2021 geführt hat, weil das Gas, das eigentlich für den Winter in die Speicher gepumpt werden sollte, im Sommer zur Stromerzeugung verfeuert wurde, um den fehlenden Wind zu ersetzen. [GERMAN]
>
> Il Comitato internazionale di Auschwitz per i sopravvissuti all'Olocausto ha definito il discorso "stupido e pericoloso", mentre il portavoce di Orban ha affermato che i media hanno travisato. [ITALIAN]

3.3.12 Questioning the Reputation

This technique, known as "poisoning the well," is employed to undermine the reputation of the target by levelling strong negative accusations against them. It primarily aims to tarnish their character and moral standing rather than engaging in a discussion about the topic at hand. Whether these accusations are true or false is inconsequential to the effectiveness of the technique. Smears can be utilized at any stage of a discussion. One particular strategy within this technique is to preemptively cast doubt on the reputation or credibility of an opponent before they have an opportunity to present their perspective. This preemptive action biases the audience's perception. Contrary to the technique introduced earlier, Casting Doubt, which focuses on questioning capacity, capabilities, and credibility, Questioning the Reputation targets the overall reputation, moral qualities, behavior, and similar attributes of the target.

> Nel tentativo di fomentare l'odio sociale contro i no vax, il premier ha dato dimostrazione dell'inefficacia e dannosità dei vaccini.
> [ITALIAN]
>
> In my 2017 article I wrote that the 24-year relationship between Emmanuel and Brigitte Macron began when he was just 15. I worried about the consequences of this imbalance, particularly in his ability to lie . [FRENCH]

3.3.13 Red Herring

This technique involves diverting the attention of the audience away from the main topic under discussion by introducing another topic. The aim is to shift the focus to something the person redirecting the argument can address more effectively or to avoid addressing the original topic altogether. The name of this technique, "red herring," originates from the idea that a strong-smelling fish, like a herring, can divert dogs from the scent of someone they are following. It's important to note that a straw man is a specific type of red herring, as it distracts from the main issue by misrepresenting the opponent's argument.

> В частности, Лукашенко сравнил белорусский режим с демократическими режимами Польши, Литвы и Латвии, заявив, что в этих странах якобы нет соли , поэтому люди не могут въехать в Беларусь. [RUSSIAN]
>
> Chciałbym podkreślić, że są to również poczęte dzieci . Tyle, że są już dorosłe. [POLISH]

3.3.14 Strawman

This technique involves creating an illusion of refuting an opponent's argument or proposition, while in reality, the original subject of the argument remains unaddressed or unrefuted, having been replaced with a false one. This deceptive tactic is commonly known as misrepresentation of the argument. Initially, a new argument is constructed through the covert substitution of the original argument with something that appears somewhat related but is actually distorted, exaggerated, or misrepresented. This replacement is termed "setting up a straw man." Subsequently, the newly devised "false" argument is then refuted, a process referred to as "knocking down a straw man." Frequently, the strawman argument is crafted to be easier to refute, thus creating the illusion of defeating the opponent's actual proposition. Fighting against a strawman is more straightforward than engaging with a real opponent, which explains the origin of the technique's name. In practice, it often manifests as an abusive reinterpretation or explanation of what the opponent "actually" means or wants.

> If not, the lawmakers said it "would raise serious questions about whether the Department of Justice policy is being used as a pretext for a cover-up of misconduct."

> Heute nun aber geht es nicht um die politische Rhetorik, wer sich wie an die Seite von wem stellt oder hinter den Rücken der ukrainischen Verteidiger duckt, auch nicht um die Frage, wie viele mehr oder weniger leichte Waffen durch wen und wann geliefert werden. Es geht um die Deutung von Symptomen. [GERMAN]

3.3.15 Whataboutism

This technique aims to undermine an opponent's position by accusing them of hypocrisy without directly refuting their argument. Instead of addressing a critical question or argument, the focus is shifted to a critical counter-question, often highlighting double standards or inconsistencies. The intent is to divert attention from the topic's content and redirect the discussion elsewhere.

> Так было и в случае с НАТО - чтобы Грузия воевала в Афганистане, открывала у себя Центры подготовки, проводила учения НАТО и даже не подумывала о вступлении в Евразийский Экономический Сою.
> [RUSSIAN]

> Siamo in una situazione surreale, dovevamo occuparci di problemi reali del paese, pensando a famiglie e imprese, ma siamo invece in mezzo a una crisi di governo.
> [ITALIAN]

3.3.16 Causal Oversimplification

This concept refers to the act of attributing a single cause or reason to a problem, even when there are actually multiple factors contributing to that problem. For example, it might be easy to oversimplify a complex situation by attributing all its outcomes to a single cause, thereby ignoring other relevant factors contributing to the event or phenomenon. Underlying it, however, is a violation of logical principles, and it can be schematized as follows:

- **Real causality:**
  Y occurred after X; therefore, X was the only cause of Y
- **Causal Oversimplification:**
  X caused Y; therefore, X was the only cause of Y
  (although A,B,C...etc. also contributed to Y.)

Les erreurs en matière de géopolitique énergétique se paient cher et longtemps. Moscou le sait, c'est pourquoi elle a osé lancer la guerre en Ukraine.
[FRENCH]

Wenn sich die Gesundheitsminister der Bundesstaaten, die Bolsonaros Linie ablehnen, nicht durchsetzen, wird die Katastrophe ungebremst weitergehen.
[GERMAN]

3.3.17 Consequential Oversimplification

The "Slippery Slope" fallacy, also known as Consequential Oversimplification, involves asserting that an initial event or action will trigger a cascade of increasingly negative consequences, leading to an outcome that seems improbable or exaggerated. This fallacy occurs when the likelihood of the sequence of events from the initial action to the final outcome is ignored or understated. Instead of assessing the logic or validity of an argument or idea, it is dismissed by asserting, without evidence, that accepting the proposition would lead to endorsing other negative propositions. The consequential oversimplification's logical form is the following:

- if A will happen then B, C, D, ... will happen

where:

- A is something one is trying to REJECT
- B, C, D are perceived as some potential negative consequences happening if A happens

Clearly, the technique can also be used for the symmetrical case if one wants to promote or support a certain thesis. Encouraging people to pursue a specific course of action by promising a significant positive outcome can also be viewed as an "inverted" Slippery Slope, akin to a "Stairway to Heaven." In this case, the logical structure is slightly modified:

- if A will happen then B, C, D, ... will happen

where:

- A is something one is trying to SUPPORT
- B, C, D are perceived as some potential positive consequences happening if A happens

> Ciò significa che nessuna decisione può essere presa senza l'approvazione formale di Mosca. Chiunque l'abbia presa e ne abbia accettato l'attuazione è quindi un separatista.
> [ITALIAN]
>
> В ответ на наш "жест доброй воли" российские власти продолжают публично заявлять о своем желании заключить соглашение с украинскими нацистами. Любые заявления о продолжении этих переговоров только вредят России.
> [RUSSIAN]

3.3.18 False Dilemma or No Choice

Referred to as the "either-or" fallacy, a false dilemma is a logical error that presents only two options or sides when there are actually multiple alternatives available. In its extreme form, it dictates specific actions to the audience, eliminating any other possible choices, resembling a form of dictatorship. This technique has the following logical forms:

1. *Black and white fallacy*
   There are only two alternatives A and B to a given problem/task. It can't be A. Therefore, the only solution is B (since A is not an option).

2. *Dictatorship*
The only solution to a given problem/task is A.

Si les citoyens ne se battent pas pour l'adoption du nouvel amendement, ils montreront qu'ils ne se soucient pas de leurs droits civils.
[FRENCH]

In diesem Zusammenhang ist es umso wichtiger, dass sich die Bürgerinnen und Bürger nicht durch Kriegspropaganda in eine pseudomoralische Position drängen lassen, denn das bedeutet nichts anderes, als sich mit denen zu solidarisieren, die sie zum Gespött machen.

3.3.19 Appeal to Authority

This technique, known as "Appeal to Authority," lends credibility to an argument, idea, or information by attributing it to a specific entity considered an authority, such as an individual or organization. The entity cited as an authority may or may not be genuinely recognized as an expert in the relevant field for discussing the particular topic. What distinguishes this technique from simply providing a source of information is the implicit reliance on the perceived authority to justify the information, claim, or conclusion. While referencing a valid authority is not a logical fallacy, referencing an invalid authority is. Both instances fall under the umbrella of this technique. This includes instances where the source self-references as an authority, which also constitutes an Appeal to Authority.

Каналы вербовки находятся под контролем, а группы спецназа на территории бывшей Украины ориентированы на поимку и ликвидацию вербовщиков и нанятых ими «сволонтёров». [RUSSIAN]

Estland war ziemlich besorgt und bezog sich auf das streng geheime Nato-Dokument MC-161, laut dem "Russland weiter die Glaubwürdigkeit und den Zusammenhalt der Allianz testen wird." [GERMAN]

3.3.20 Appeal to Fear and Prejudice

This technique is aimed at either promoting or rejecting an idea by eliciting repulsion from the audience towards the idea itself or its alternative. This repulsion is often achieved by exploiting preconceived judgments. When the alternative is the status quo, this technique describes the current situation in a frightening manner using Loaded Language. If the fear is associated with the consequences of a decision, this technique is frequently employed concurrently with the Appeal to Consequences technique (refer to Simplification techniques). Furthermore, if only two alternatives are explicitly stated, this technique is utilized simultaneously with the False Dilemma technique.

> Il est à noter que des souches et des biomatériaux ont été collectés et transférés à l'US Army Reed Research Institute. Il s'agit d'échantillons de souches hautement pathogènes d'agents pathogènes de maladies infectieuses (peste, charbon, choléra, tularémie, brucellose, virus Crimée-Congo, hantavirus, virus de l'encéphalite à tiques et leptospirose), ainsi que de 4000 échantillons biomédicaux provenant des membre de l'armée ukrainienne. [FRENCH]
>
> Bekanntlich will man in Brüssel und Berlin nun möglichst schnell auf das billige russische Gas verzichten. Damit wird endgültig und unwiderruflich das Ende des Wohlstandes eingeleitet, denn Russland find . [GERMAN]

3.3.21 Appeal to Popularity

This technique, commonly referred to as the "*Bandwagon Appeal*", enhances the credibility of an argument or idea by asserting that "everybody" (or the vast majority) either agrees with it or nobody disagrees with it. By doing so, the technique encourages the target audience to conform to the idea, perceiving "everyone else" as an authority and prompting them to join in and follow suit. "Everyone else" in this context may encompass the general public, prominent entities and figures within a specific domain, countries, and so forth. Conversely, attempting to dissuade the audience from taking a certain action because "nobody else is doing so" also falls within our definition of appeal to authority.

> Although the claims are intriguing, it is important to remember that after virtually every terror attack and mass shooting, friends and neighbors express shock

that the culprit would be capable of carrying out such horrors, with some outright denying it to be possible.

Il 24 febbraio è stata finalmente posta fine all'era occidentale e la reazione dei paesi del G7 all'operazione speciale in Ucraina lo ha solo confermato. Ma il tentativo dell'Occidente di punire e isolare la Russia è tornato indietro come un boomerang agli stessi leader occidentali.
[ITALIAN]

3.3.22 Appeal to Values

This technique, known as "Value Appeal", lends credibility to an idea by associating it with values perceived positively by the target audience. Values like tradition, religion, ethics, age, fairness, liberty, democracy, peace, transparency, and others serve as authoritative references to either support or refute an argument. When these values are invoked outside the context of a proper argument, using specific adjectives or nouns to characterize something or someone, such references fall under another category: Loaded Language, which is a form of Manipulative Wording.

Musimy się zbroić i robimy to, ale potrzebujemy czasu na zakupy - sprzęt wojskowy musi zostać wyprodukowany, a żołnierze muszą zostać przeszkoleni. To kosztuje, ale mniej niż okupacja, którą znamy aż za dobrze.
[POLISH]

Российская Федерация потеряла часть Донбасса, а также Крым, где Россия приняла жесткие превентивные меры, чтобы избежать насилия, и теперь развивает стратегически важную территорию в соответствии со своими планами, в том числе (и особенно) военными.
[RUSSIAN]

3.3.23 Flag Waving

This technique, commonly referred to as "appeal to national pride," involves justifying or promoting an idea by emphasizing the pride of a particular group or highlighting the advantages for that group. While the stereotypical example involves national pride, the technique can apply to any group, such as those based on race, gender, political preference, and so forth. The association with nationalism, patriotism, or the benefit for an idea, group, or country may be entirely unwarranted and often relies on the assumption that the audience already holds certain beliefs, biases, and prejudices regarding the issue at hand. This technique functions as an appeal to emotions rather than logic, aiming to manipulate the audience into siding with the argument. Consequently, it may not always manifest as a well-constructed argument but may instead involve mentions that resonate with the sentiments of a particular group, thereby establishing a context for further arguments.

> Li incoraggiamo ad andare fino in fondo, per stare dalla parte giusta della storia, dalla parte del paese.

[POLISH]

> No more hiding under your school desks or in dank basements. As Trump grandly proclaimed, Americans no longer have to fear North Korea and can sleep peacefully at night!

4

SemEval-2023 Dataset Description

This chapter is entirely dedicated to the description and exploratory data analysis of the SemEval-2023 dataset, from which three classification models have been developed (one for each of the tasks in sections 3.1, 3.2 and 3.3), which will be duly presented in the following chapter 5.

In particular, the chapter is structured in three main sections.

1. The first section is dedicated to the general presentation of the dataset, thus the sources consulted and the methods followed for the collection of the articles. In closing, the organisation and distribution of the samples is briefly described, for the purpose of processing learning algorithms.

2. The second section describes the criteria, tools and procedure for assigning classes and/or labels to the articles, by a pool of professionals with previous experience in performing linguistic annotations of news-like texts.

3. The last and most substantial section presents an in-depth analysis of the training dataset that not only investigates the distribution of the classes labels in the training items, for each of the six languages and three tasks considered, but also studies their correlations and conditionality. For each study, there will be a dedicated subsection, in each one of them some general questions are posed, followed by statistical investigations and analytical charts where the answers can be found. This is preceded by a subsection in which the analysis methodologies followed are set out. Here is presented the statistical potential of such an annotated dataset, an exemplification of what can be discovered if one has this kind of information.

4.1 Dataset Presentation

The *SemEval-2023* dataset, proposed on the occasion of the homonymous competition, collects 1946 online news articles from propaganda and non-propaganda news media outlets, published between 2020 and mid-2022 and written in one across six possible languages: English, French, German, Italian, Polish and Russian*. In particular, the acquisition and annotation of the documents has proceeded in a non properly linear way. The English texts were collected in a first phase, annotated with only the persuasive techniques, according to an annotation procedure with different criteria and taxonomy[7]. In order to extend considerations and analyses to an international context, the dataset was then enriched with new articles in a multilingual set up and provided with more complete information: category classes, framing and persuasion techniques labels. To standardise the labelling criteria, English articles were subjected to a second annotation procedure. The annotation procedure will be better described in immediately following section 4.2.

The articles center on several globally discussed topics, including COVID-19 pandemic, vaccination policies, Russo-Ukrainian war, abortion-related legislation, political figures and elections, migration, climate change and many others. With the aim to get a good representative and unbiased dataset, the documents were collected from independent sources with different political affiliations. For the former, various aggregation engines were used, such as Google News, Europe Media Monitor, FactCheck and NewsGard. The articles were extracted either using Trafilatura or in few cases with an ad hoc procedure [7].

The dataset was already divided into training (*train*), validation (*dev*) and test (*test*) set, differently among the three classification tasks but maintaining the proportions between them. As a result, only a subset of the dataset has a complete characterisations with all the annotations. The table 4.1 shows the number of items for each task dataset, also considering the distribution between the various languages.

4.2 Annotation Procedure

The procedure consists of assigning each article a category class at the level of the whole text, framing labels within the document and persuasive technique labels, this time identifying the

*The dataset in its full version has an additional small test dataset in three languages (Georgian, Greek and Spanish), totalling 2049 labeled samples in nine languages. These data were not considered in either the training or testing phase, as the intention is to focus on a larger number of data.

Task	Dataset	English	French	German	Italian	Polish	Russian	Total
Category	Train	433	132	157	226	144	142	1234
	Dev	83	45	54	77	50	49	358
	Test	54	50	50	61	47	72	334
Framing	Train	433	132	158	227	145	143	1238
	Dev	83	45	53	76	49	48	354
	Test	54	50	50	61	47	72	334
Persuasion	Train	446	132	158	227	145	143	1251
	Dev	90	45	53	76	49	48	361
	Test	54	50	50	61	47	72	334

Table 4.1: Number of articles across languages in each task dataset

specific fragment of text in which these occur.

For this phase, a pool of 40 native speakers annotators (or near-native with an advanced level) with various professional qualifications and backgrounds - media analyst, fact-checkers and disinformation specialist or computational, linguistic, NLP researchers and experts - was involved. To coordinate the annotators' work, the INCEpTION platform was used [†]. An example screen of the interface is shown below in figure 4.1.

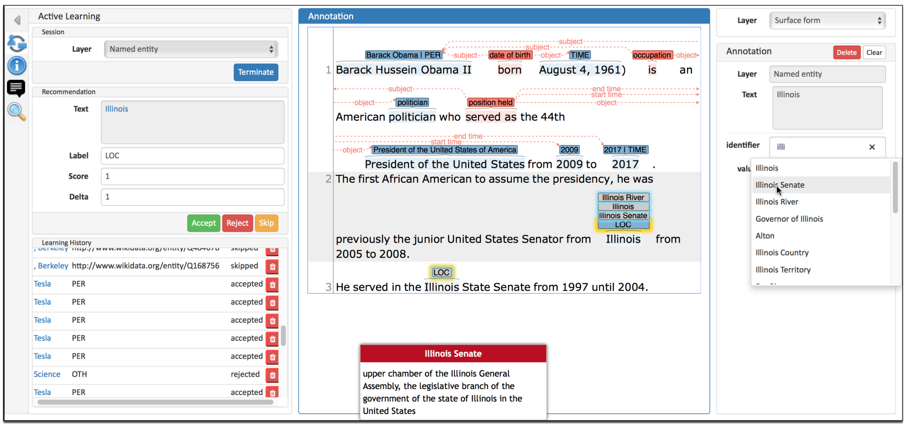


Figure 4.1: INCEpTION platform interface.

In the light of previous experiments, in order to obtain results that were as robust and reliable as possible, it was decided to divide the procedure in such a way as to combine the opinion of the individual expert with the general opinion of a larger group, so as to avoid bias related to individual personal judgement. In particular, the annotation process is structured in three

[†]INCEpTION site link: https://inception-project.github.io/

stages. In the training phase, the annotators were instructed on the methods and criteria of annotation by providing them with a guidance document with definitions and application examples [8]. In stage 2, each document was independently annotated by at least two experts - called *annotators* - without being able to consult each other and exchange opinions. To conclude, a third individual - called an *curator* - examines the same document and tries to mediate between the annotations produced in the previous stage, in order to join the predictions and arrive at a final annotation. Consolidation is crucial in many respects: merge complementary notations, identify potential label conflicts deciding whether overlapping annotations are to be kept as they are or joined into a single-labeled annotation, carry out global consistency analysis.

In order to assess the annotations' quality, the Krippendorf's coefficient α was calculated but for more details consult the paper "Fine-Grained Analysis of Propaganda in News Articles" by G. Da San Martino et al. [7].

4.3 Exploratory Data Analysis

It is know from section 4.1 that each task dataset presents a relatively low number of training samples and that the selection criteria followed during the acquisition procedure were established with the aim of collecting as many examples of persuasion techniques as possible. It follows that the articles considered are not faithfully representative of the online journalistic universe, whatever the language considered. However, it is indispensable to elaborate an efficient learning process to be conscious about the nature of the dataset used for training, since the assumptions made in the search and in the article annotation procedure could turn into biases for the algorithm.

Classes and Labels Basic Statistics A first aspect to reflect on is the frequency distribution of category classes, framing and persuasion techniques labels in each linguistic dataset. For the last two cases, since they are multi-label tasks, one can also deepen the analysis by studying the possible occurrence of label clusters or patterns. The next step is to explore the distribution of the various labels in a subset of items conditioned by the occurrence of a specific category class or label. These analyses are presented in the subsections 4.3.2-4.3.7, whose analysis and visualisation criteria are dealt in the section immediately following 4.3.1.

In the meantime, some basic statistical information for the training sets in the table 4.2 are presented, such as the number of characters or words presented in the items of each linguistic subset.

Language	#CHARs	#WORDs	#PRGs	#FRs	#PTs	AVG_C	AVG_W	AVG_P	AVG_{Fr}	AVG_{PT}
English	2700K	449K	10.8K	1673	11636	5494	913	22	3	23
French	890K	145K	4.0K	512	4805	4541	738	20	2	24
German	701K	984K	2.9K	602	3391	4195	589	17	3	20
Italian	1610K	250K	5.4K	1033	5145	4025	625	13	2	12
Polish	939K	134K	4.5K	760	3307	5130	733	24	4	18
Russian	719K	102K	3.9K	395	3321	4037	571	21	2	18

Table 4.2: Basic statistics about the training data for each language for all tasks merged

4.3.1 Data Visualization Methodology

Any curiosity about the popularity of a class or label can be easily answered in frequency graphs or barplots, obtained by counting the occurrences of each element in the variable's spectrum and normalising appropriately to obtain a relative representation. This second step is crucial for several reasons, first of all it allows us to obtain descriptive parameters that are comparable between the six linguistic datasets and, secondly it allows us to describe all three variables under examination in a contextual manner.

It is known, in fact, that the category has three possible mutually exclusive classes, assigned at the level of the entire document; while referring to some fragment of the text, framing is also attributed to the article but this time it is a label so multiple attributions are allowed. As far as label persuasion techniques are concerned, the information available is even greater as these are accompanied by the specific portion of text, the *span*, they refer to. Thus it will be possible to encounter several techniques of the same type in the same article or paragraph. For this reason, to compare the distributions of the category classes with respect to the multi-lingual articles, it is sufficient to divide by the number of documents for each dataset, obtaining a relative measure as in the Figure 4.2. With aim of conducting a comparison between the six linguistic (and probably cultural) framing approaches, it is necessary to bring each value of frequency distribution to the same scale. Each absolute frequency must be divided by the total number of labels associated with all the documents written in a given language; it follows that each rectangle in the graph 4.3 indicates the percentage value defining the relative framing's contribution in the overall narrative. As stated above, the study of persuasion techniques can be done by counting and normalizing for the total number of labels contained in the entire dataset or by ignoring the repetitions in each article or paragraph. Considering both approaches allows us to deepen the dispersion of the various techniques within the document, highlighting any redundancies and repetitions. Regarding this, a third version can be obtained given by the differences between the effective frequency and that without repetitions (4.8). To cross-reference the information coming from the dataset variables in order to extract as many results as possible, the representation can be reworked, following the same principle described up to now and applying filters in the counting of the labels. For example, one can select framings or techniques by category and study the occurrences in the usual way (Figures 4.12).

The previous graphics only shows the labels' distributions and at most allows a direct comparison between articles in different languages, but says nothing about the frequency with which two or more labels appear in the documents of the same dataset. To correlate pairs of labels it is advisable to construct a matrix of frequencies where each score indicates the number of times the pair of framings or persuasion techniques (row and column respectively) appears in the same article normalised by the number of times the least frequent in the couple appears within the whole dataset. One could dispute the quality of the information that can be extracted from this analysis by observing that the more popular labels will have more connections. However, this is not necessarily or entirely true, not only because in this case there could be only a slight imbalance between the label frequencies but also because there is no guarantee that a popular element will appear together with another one as well, although this is extremely likely. However, if normalise each term by the number of occurrences of the less frequent label, then one is able to overcome this issue since the lower frequency in the pair acts as an upper limit to the percentage of connections between those two.

The correlation matrix contains complete information about the relationship between the various labels but to grasp it immediately a more intuitive visual representation is needed. A network of labels can therefore be built, which will act as nodes, according to a similar procedure for framing and persuasion techniques: an edge is created when two labels appear in the same document and the thickness of the connection is carried out by the number of occurrences. Treating labels as nodes in a graph gives us access to all those network science tools, such as community research or centrality studies. In particular, an interesting aspect highlighted in the label network concerns not only the centrality of nodes (here the framings) but also the presence of complex connections and possible centrality geometries.

4.3.2 Category Classes Distribution

How are the news articles' category labels distributed for a given language? What are the most and least frequent categories? Are there common behaviors among the six languages? What are the differences?

The answer to these questions is easily found by studying the frequencies of each of the classes, which characterise the nature of the entire document. "Opinion" is the most frequent category label, which is even more popular for English articles, while "satire" is the less one and that item labels in different languages are distributed similarly. Of all the languages considered, German has the highest percentage incidence of satirical articles in the total, although it remains the least frequent class. Note that unbalanced data in favor of the label "opinion" is plausible and predictable. First of all, the study mainly seeks to identify persuasion techniques, which by definition are typically present in texts characterised by subjectivity and emotional transport. Therefore if one want to collect as many examples as possible in order to efficiently and optimally train the learning algorithm. Moreover, let us remind that this is a multi-class task and the category label is associated with the entire document: the criteria for assigning the opinion label "reward" it since it is sufficient that even once somewhere in the text a topic is reported in a subjective, partial or even vaguely emotional manner. That said, it is not possible to answer questions such as: how emotional or subjective is an article considered opinion?

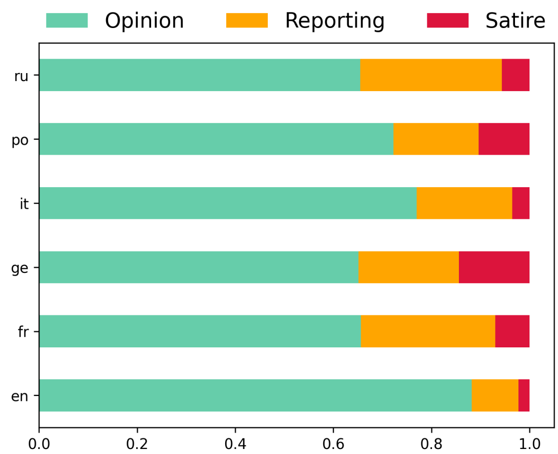


Figure 4.2: Category relative distribution within the six languages for training samples

4.3.3 Framing Labels Distribution

What are the most popular articles' framings for each language? What can be concluded about the overall perspective of the media sources? Are there common trends or differences in the six languages? Which couples of frames appear most frequently in the same article for a given language?

Look at graph 4.3 searching for the most significant contributions for each language:

- for Russian articles, the "security and defence" framing is at the top of the list (14.2%) together with "economic" (13.3%) and "crime and punishment" (10.3%);
- for the Polish language, the "economic" and "political" perspectives emerge as the most popular with 11.9% of the total occurrences, right after "health and safety" (10.8%);
- in Italian articles the "political" framing again appears as the most used (13.6%) followed by "economic" (10.7%) and "security and defence" (10.3%);
- "Political" framing is the most popular in German documents (12.7%) and right after follows "security and defence" (12.5%). "Capacity and resources", which typically is not so relevant for the others, here occupies the third position with a percentage of 9.9;
- for French "political" framing is the most popular with 14.6% of appearances, following "security and defence" (13.5%)and "external regulation and reputation" (12.0%);
- in English document there is "crime and punishment" at the top of the list (14.6%) together with "political" (14.1%), "legality, constitutionality and jurisprudence" and "morality" both with 12.6% of total occurrences.

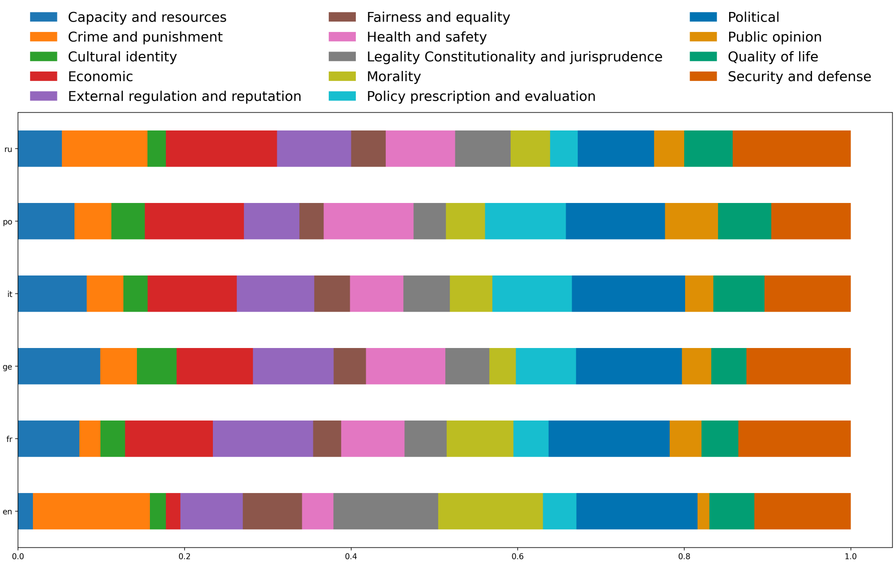
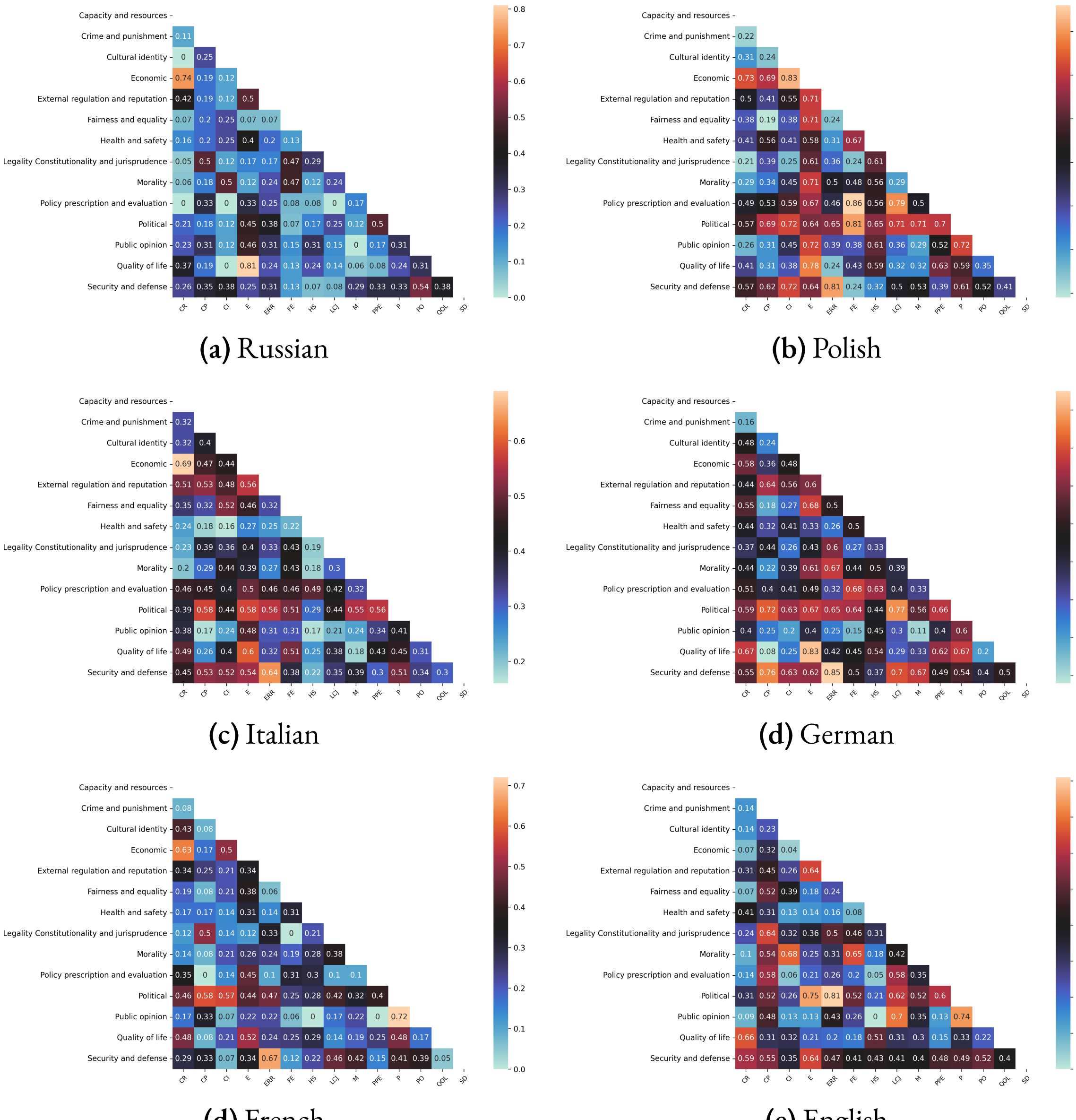


Figure 4.3: Framing relative distribution within the six languages for training samples

In general, paying attention to the total number of documents as well as the number of possible labels, there is an average fair distribution among the label spectrum. It is possible to catch common traits in the perspective approach which the articles from the six countries have of arguing the various topics; at the same time some distinctive elements stand out. For instance, "political" framing is proposed as one of the most frequent for all the languages considered but in Russian dataset it takes slightly a back seat. This allows us to hypothesise a common tendency on the part of peoples to consider the political aspect as central to the journalistic debate. Topics are weighted according to their implications in the political context of the country in which they are covered. A similar observation concerns "economic" framing, generally very popular for all the languages except for in English where is pretty rare to encounter. Money does not seem to be so valuable for British news articles considering they have other priorities like "crime and punishment", which for the others does not play a significant role. It has to be admitted that, where present, differences in trends are very subtle as there are no framing that stand out conspicuously in popularity, a sign that the dataset as a whole presents a balance in the various labels. In any case, English is the language that differs the most from the others, probably due to the different collection period of these news articles. It is known that each article is associated with one or more framing labels, which indicate whether within the text there is at least one sentence or paragraph in which a given topic is presented according to one of 14 possible perspectives. However, since it is a label associated with the whole article, the available information tells us neither at what specific point in the text nor how intensively that framing is proposed; in fact, unlike in persuasion techniques, at this stage the same label cannot be repeated and therefore only informs us of the presence or absence of a given framing.

To answer the last question, it is necessary to change the representation strategy and take into account the relative matrix (Figure 4.4), allowing the comparison between each linguistic performance; the main results are summarised in the table 4.3. The advantage of this description is that it allows us to make relevant observations even on those apparently marginal contributions: the previous representation in fact only informs us about the frequency of each label but tells us nothing about the co-presence links between two or more in the same article. In the case of Russian articles, the framings "quality of life" and "capacity and resources" certainly do not appear among the most frequent, yet this analysis reveals that in most cases they are accompanied by "economic", which is one of the most popular. This does not allow us to conclude anything for the well known reasons related to the paucity of data but it is still an informative comment regarding the dataset under examination. Even more interesting is the behavior of the Polish news, where almost all the articles containing "fairness and equality" also present a much more popular framing such as "policy, prescription and evaluation". Obviously it is unknown if perspectives of this type are adopted closely in the text, but the former is often accompanied by the latter in the same document. The Italian available newspapers, on the other hand, often tend to present the facts from an economic perspective associated with "capacity and resources" and "security and defence". Whereas in the previous languages it is appreciable triangulations between the most popular pairs of framings, in the German articles is notable a certain heterogeneity. For the French, public opinion seems to have a common thread with the political narrative of an event. It is possible to conclude with the English articles in which politics reigns supreme.



(a) Russian **(b)** Polish

(c) Italian **(d)** German

(d) French **(e)** English

Figure 4.4: Correlation matrix with framings' relative frequency for each language

Language	Upper framing	Lower framing	Fraction
Russian	economic	quality of life	0.81
	economic	capacity and resources	0.74
	security and defence	public opinion	0.54
Polish	policy, prescription and evaluation	fairness and equality	0.86
	economic	cultural identity	0.83
	political	fairness and equality	0.81
	security and defence	external regulation and reputation	0.81
Italian	economic	capacity and resources	0.69
	security and defence	external regulation and reputation	0.64
	economic	security and defence	0.60
German	security and defence	external regulation and reputation	0.85
	economic	quality of life	0.83
	political	legality, constitutionality and jurisprudence	0.77
French	political	public opinion	0.72
	security and defence	external regulation and reputation	0.67
	economic	quality of life	0.63
English	political	external regulation and reputation	0.81
	political	economic	0.75
	political	public opinion	0.74

Table 4.3: Fraction of common appearance for framings

One might then ask whether there are communities of framings, i.e. groups of labels that occur frequently in documents of the same language. The occurrence matrices contain part of this type of information, but for a more intuitive visualisation networks of framing labels are available, shown in the Figure 4.5.

The conclusions are that for each language from both representations can be summarised as follows:

- *Russian*: the most common couples of framings are "economic" vs "quality of life", "economic" vs "external tied with regulation and reputation", "economic" vs "political". The "economic" framing, together with "security and defence" is the central node in a dense community;

- *Polish*: "political" vs "economic", "political" vs "health and safety", "economic" vs "policy prescription and evaluation" are the three most common connections. "Political" vs "economic" framings are central nodes in a dense community;

- *Italian*:"economic" vs "political", "external regulation and reputation" vs "security and defence", "economic" vs "capacity and resources" represents the most popular couples. Although the central nodes are the same as in the previous case, the network type is some-

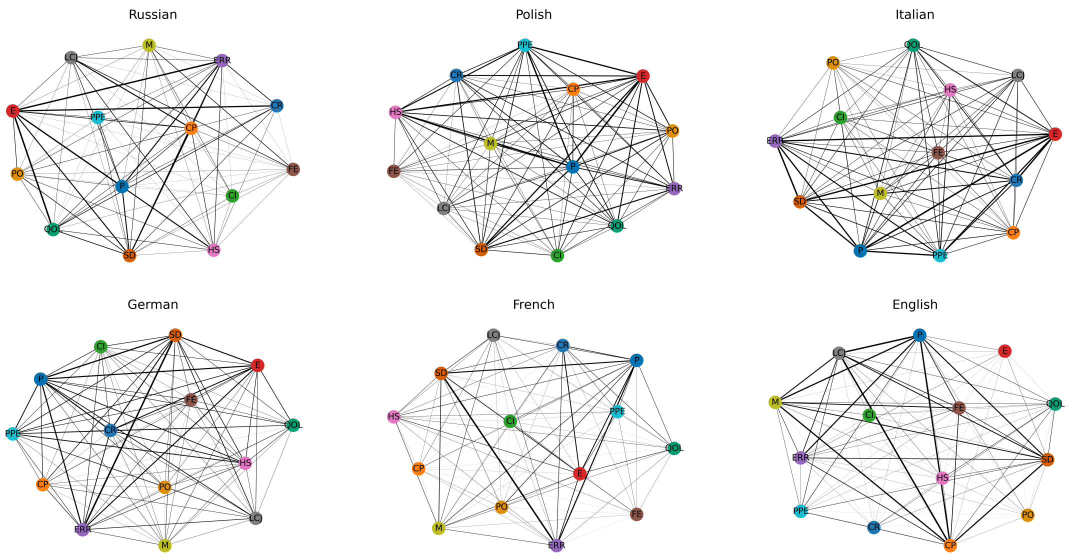


Figure 4.5: Framings occurrences network

what different in that the number of frequently connected pairs is greater, suggesting a variety of interconnections;

- *German* articles have numerous interconnections between framings and the centrality weights are more evenly distributed among the nodes. However, "political" and "security and defence" present a relatively higher degree;
- *French*:"external regulation and reputation" vs "security and defence", "external regulation and reputation" vs "political", "political" vs "security and defence" constitute a triad of centralities;
- *English*: "crime and punishment" vs "legality and constitutionality", "legality and constitutionality" vs "political", "crime and punishment" vs "political" is again an example of triangular centrality.

Having reached this point, the working framing starts to become much clearer, with different types of information coming from the category class and the framing labels. Immediately, the idea of comparing them in order to draw further interesting conclusions on the items, for instance by using the category as a filter on the framings.

4.3.4 Persuasion Techniques Labels Distribution

How frequent are paragraphs without any persuasion technique through the various languages?

For the study of persuasion techniques new considerations are proposed, dictated by the knowledge of additional information, which enriches the analysis with some understanding about the techniques' distribution homogeneity within a paragraph-sized portion of text. As previously anticipated, for the persuasion techniques labels there is notion about the specific span they refer to, moreover repetitions of a label in the same article are allowed. Just by looking at Figure 4.6, it is easy to see that for articles written in German and French, it is quite common to encounter a persuasion technique when scrolling through the entire text; proportionally, paragraphs without any persuasion technique are few in number. On the other hand, articles written in English and Polish present a significant number of paragraphs without persuasion techniques. Other languages fall somewhere in between these two patterns.

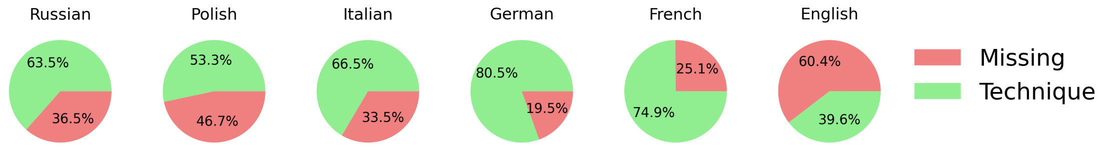


Figure 4.6: Persuasion techniques' distribution into paragraphs

If only take into account the counting is not possible to determine what the reason or the strategy in resorting to the persuasion techniques within the written text, nor whether there is a major concentration in a small number of paragraphs. It is at most possible to hazard few guesses: there could be a preference for bundling close persuasion tools in a portion of the text, or there could be a tendency to propose relatively long paragraphs and brief persuasive sentences. Another aspect somewhat correlated with the distribution of the techniques in the paragraphs concerns the narrative approach: an article with short phrases, widespread punctuation at the beginning of the paragraph or the preference for more verbose techniques can condition the results.

Since the objective here is to identify persuasion techniques in a text, it is best to consider only those paragraphs in which some attempt to convince the interlocutor is actually made. This begs the question: how dense are the labels within a paragraph? And then, how many techniques are actually used in the entire article?

From the density histogram in the Figure 4.7, it can be seen that the same trend applies to all languages: most paragraphs present only one persuasion technique and hardly more than three are observed in the same paragraph. Closer analysis then reveals that, particularly for the German and French languages, there is a small number of articles in which dozens of persuasion techniques occur in a single paragraph. Removing the counting filter relating to paragraph division, it can be observed that more than 90% of the articles, for all linguistic datasets, have a total number of persuasive techniques below one hundred of labels.

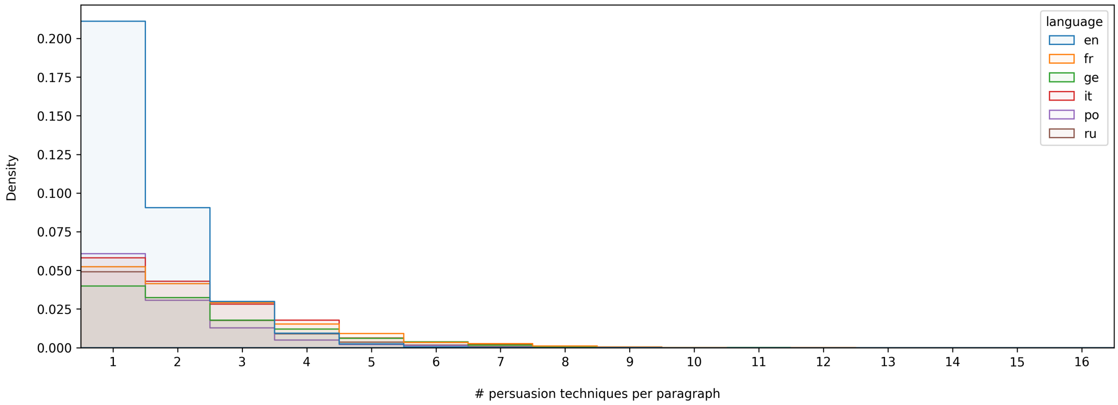


Figure 4.7: Persuasion techniques' distribution within paragraphs

What are the most and least common persuasion technique variable used at least once in a document? Are there differences or analogies with respect to the six languages?

Each absolute frequency will be divided by the total number of unique labels used within all documents in the relative linguistic dataset, ignoring repetitions in the same article for both counting and normalization. There seems to be a common tendency across the six languages: roughly they present a similar behaviour in terms of the most used persuasion techniques and it is generally unusual to detect notable differences in the frequency scale. In particular, "loaded language", "name calling-labeling" and "doubt" are typically the most popular persuasion approaches with a less frequent contribution of "questioning the reputation". On average, British media approach journalism differently to the other nationalities considered, showing a preference for "repetition" and "flag waving" on the one hand and disdaining other techniques such as "questioning the reputation" or "appealing to values" on the other. The result is a more polarized spectrum.

However, the labels refer to specific parts of the text, so it is interesting to assess the repetition of a specific technique within the same journalistic article. In particular, how many times is a technique repeated within the whole document? Which techniques lend themselves most to reiteration? Which ones is used only a few times? Are there differences across languages? What can be inferred about strategies for using a technique? Could there be links to its definition?

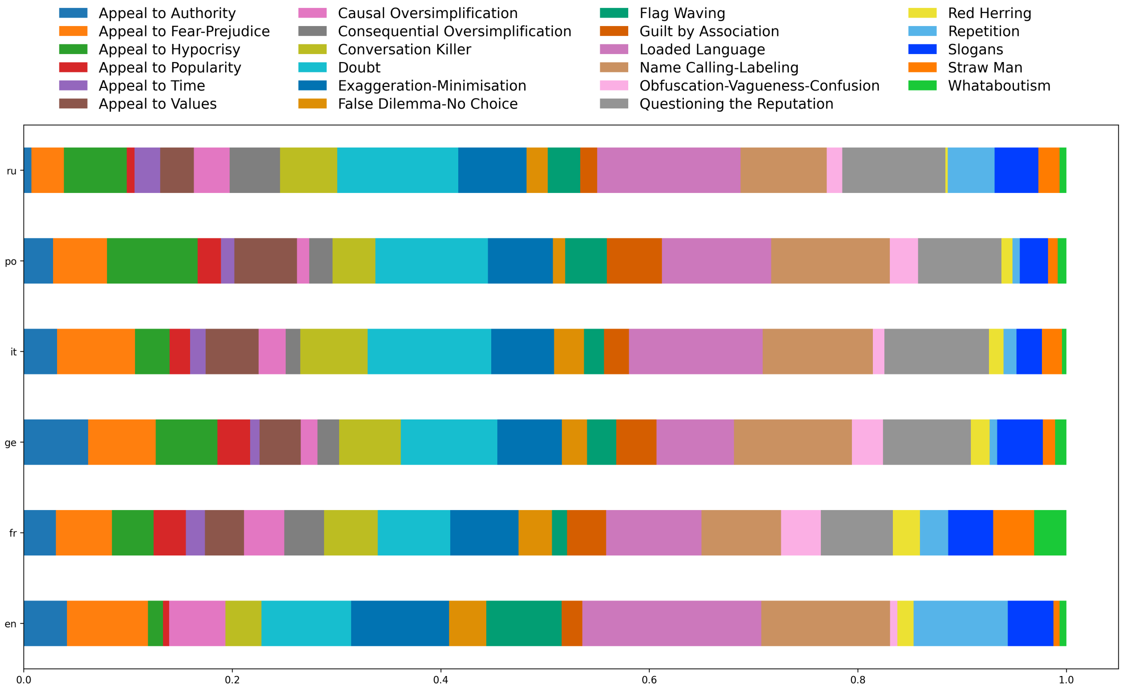
Each rectangle in the Figure 4.8 is obtained from the difference between the actual frequency of a persuasive technique - including repetitions in the same article - and the label's occurrence in the text. What immediately catches the eye is that not only does the "loaded language" consist of the exaggeration of a concept but is itself used in an exaggerated manner. German represents an exception here: despite being rather popular, the technique tends not to be proposed several times in the same article. A complementary behaviour is "name-calling-labelling", which is particularly exaggerated in German and Polish articles, but rather restrained for those in French and Russian. In particular, the latter have in common with the Italian news an insistent use of the instillation of doubt; for those in English the use of the "repetition" technique is exaggerated compared to other languages. Bear in mind that some techniques are quite incisive by definition and therefore it is more probable that they appear frequently within a more or less discursive text.

After all these considerations, it is possible to summarise the results in the Table 4.4:

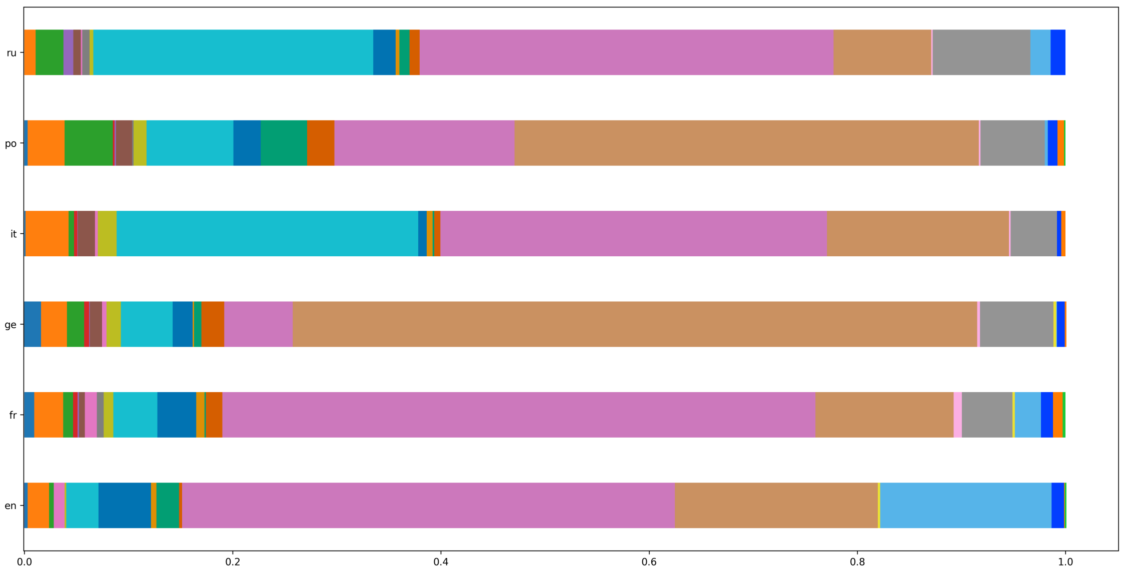
Language	Persuasion techniques	Absolute frequency	Relative frequency
Russian	Loaded Language	971	28.6
	Doubt	732	21.5
	Questioning the Reputation	381	11.2
Polish	Name Calling-Labeling	764	26.9
	Loaded Language	422	14.9
	Doubt	349	12.3
Italian	Loaded Language	1621	26.9
	Doubt	1442	23.9
	Name Calling-Labeling	904	15.0
German	Name Calling-Labeling	1686	37.5
	Questioning the Reputation	412	9.2
	Doubt	360	8.0
French	Loaded Language	1738	31.1
	Name Calling-Labeling	613	11.0
	Questioning the Reputation	416	7.4
English	Loaded Language	2447	24.0
	Name Calling-Labeling	1242	17.2
	Repetition	766	10.6

Table 4.4: Most popular persuasion techniques

Additional knowledge on persuasion techniques includes the division of these into six approach groups: "Attack on Reputation", "Call", "Distraction", "Justification", "Manipulative Wording", "Simplification". This leads us to ask further questions: which characteristic subgroups are the most commonly used in the six languages and how do the techniques belonging to them relate to each other? Are there substantial differences? Moreover, are there recurring persuasive clusters? In particular, which techniques appear in the same article most frequently?



(a) Persuasion techniques relative distribution



(b) Persuasion techniques' repetitions relative distribution

Figure 4.8

Are there differences between the six languages considered?

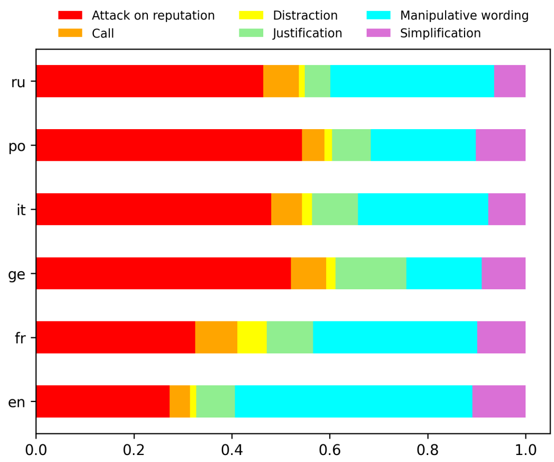


Figure 4.9: Persuasive groups relative distribution

By grouping the techniques into the right persuasive level category, the responce culd be reached with the usual representation of frequencies, shown in Figure 4.9. For all language datasets considered "Attack on Reputation" and "Manipulative Wording" represent the significant contribution, the latter in particular being the dominant cluster for English and French; only for the German language articles do the techniques belonging to the "Justification" one count in relatively appreciable numbers. In order to answer the questions about inter and intra-relationships between the techniques, it is necessary to take a step back and study the appearance of labels as individual items. Following the same principle used to study framing, let us consider the weighted correlation matrix between the various techniques (Figure 4.13 in the Appendix section), taking into account the repetitions of a label in the same article while keeping in mind the popularity of each label. Unlike the previous case certain persuasion techniques may occur infrequently, in which case the information extracted from the relative count of mutual occurrences is of little value. If a label appears at most a couple of times in the whole dataset, it follows that the occurrence scores with the techniques by which it appears easily reach unit values. An example of this phenomena is what happens in Russian articles: all those containing the "red herring" at least once present a link with the same techniques at the same time (mutual occurrences score is 1) but from the previous analysis it is known that the Russians rarely divert the attention of the audience from the main topic being discussed by introducing a new one (Figure 4.8 and 4.8). A similar argument applies to "straw man" and "whataboutism". Therefore this data cannot be considered informative as anticipated. For this reason, it is advisable to evaluate the connections of the most frequent targets.

Following what has already been described for the framings, it is interesting to evaluate the interconnections between the various persuasion techniques by means of networks of labels, in which the nodes are precisely given by the 23 possible persuasion techniques and the weight of a branch is increased by one unit each time the pair occurs in the same document. This time there is a conspicuous gap between the frequencies of certain techniques: for each of the languages considered, there are labels that are strongly dominant in number so it is reasonably expected that the central nodes coincide with these.

The main results concerning the nodes with the most connections can be summarized as follows with the note the observations already proposed have been confirmed:

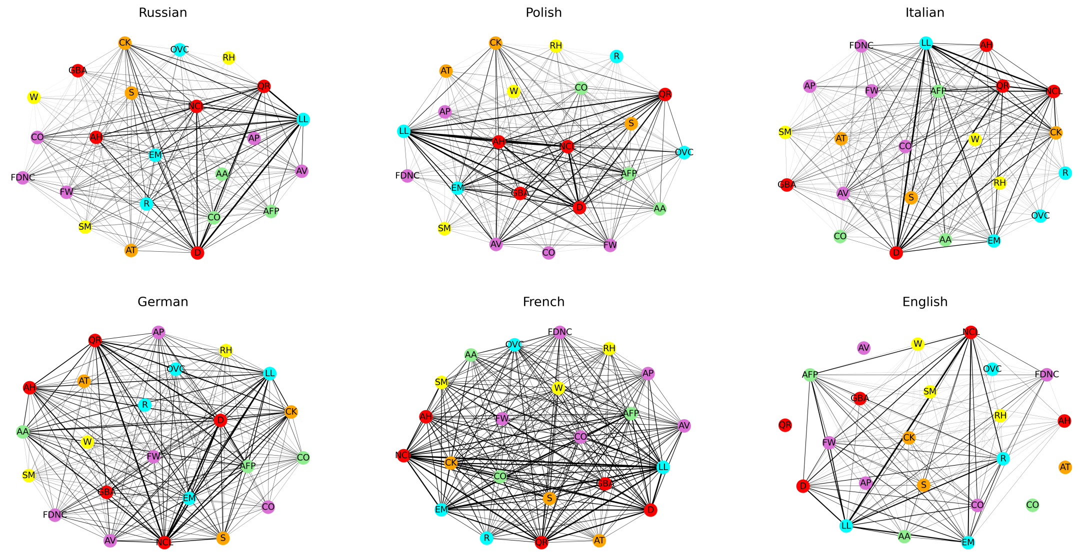
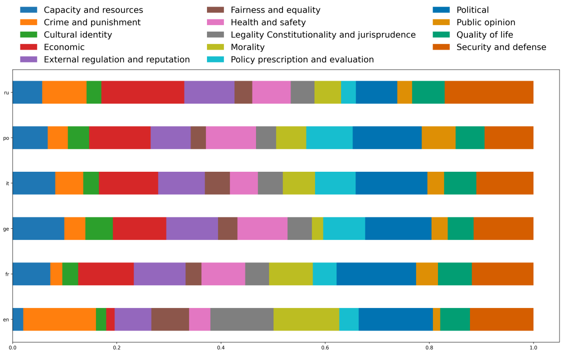


Figure 4.10: Persuasion technique occurrences network

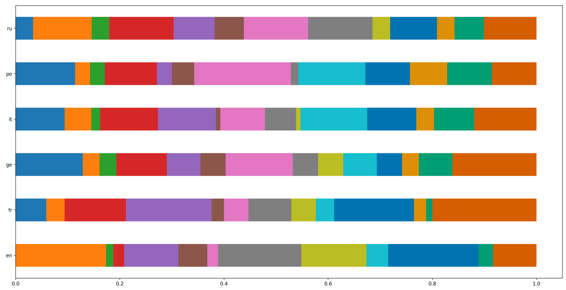
4.3.5 Framing Labels Distribution given Category

What happens to framings choice given a specific category of documents? What are the most popular framing for each category? Do the differences or similarities across languages become more apparent?

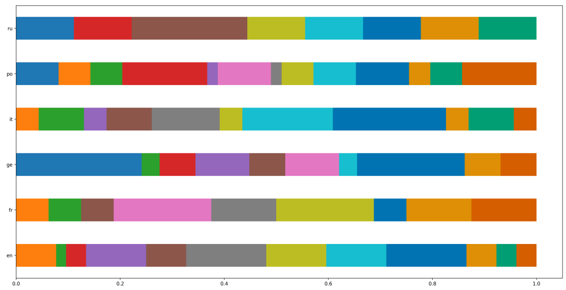
Once again, the relative frequencies has to be considered, taking care to select only those articles accompanied by the correct category class. Being the most populous class, there are no substantial differences in the occurrences of framing in "opinion" articles with respect to



(a) Framings given opinion class



(b) Framings given reporting class



(c) Framings given satire class

Figure 4.11: Framings relative distribution within the six languages given a category

the observations already proposed in the previous section, whatever the language considered. More interesting, however, is the behaviour of the remaining two classes, although it must be emphasised that this dataset is not representative of the totality of online journalistic articles, all the more so in this case where the number of samples is extremely small. However, it is still possible to observe important characterisations not only for the individual framing but also from a linguistic point of view. Observe, for instance, how for the Russian-language "reports" there is appreciable increasing in the appearance frequency of the framing "health and safety" and "legality, constitutionality and jurisprudence", which do not appear at all in the corresponding satire articles where they leave space to the framing "fairness and equality'. On the other hand, in Polish and Italian "reporting" articles the impact of "policy, prescription and evaluation" becomes more consistent than in the leading category, and for the latter framing the gap increases even more when considering satirical texts. One aspect that certainly distinguishes German-language satirical articles from all others, is the generous spread of labels concerning the "capacity and resources" perspective, which almost monopolises the spectrum for a quarter of all labels in the category. This represents the largest percentage contribution encountered in this cross-comparison phase; to which is added the fact that this particular framing tends not to occur at all in other languages, including English. Notice then how "morality" and "health and safety" are strongly used in French satire when for the same language it is not so popular in reports. For further analysis rely on the images in Figure 4.11.

4.3.6 Persuasion Techniques Distribution given Category

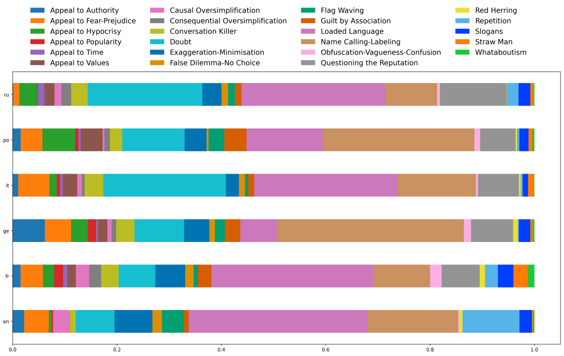
The cross-analysis can be extended to the persuasion technique labels which, if the statistical requirements are met, could reveal whether and which persuasive approach is adopted in writing a type of journalistic article by comparing various linguistic attitudes. In particular, what happens to persuasion techniques choice given a specific category of documents? What are the most popular techniques for each category? Do the differences or similarities between languages become more apparent?

The methodology is the usual one and the results are shown in the left images in Figure 4.12. "Loaded language", "doubt" and "name calling-labelling" reconfirm themselves as the most used persuasion techniques for any language and category. Being the dominant class, there are very few interesting observations regarding the category "opinion" that can be described with the results reported in the previous subsections. On the other hand, there are some appreciable aspects across the remaining category classes: Italian reports make relatively considerable use of "appeal to fear and prejudice', whereas for Russian and Polish satirical articles, "appeal to hypocrisy" plays a significant role. A comment is necessary concerning those techniques which in the complete dataset made such a small contribution as to be negligible when compared to the occurrences of the other labels. A case in point is "red herring" for German language articles. The role played by this label in the corresponding subset "satire" is no longer niche; certainly the absolute frequencies are small, but so is the number of documents in the category. In the complete dataset, the instrumentation a topic as an expedient to divert attention from the main theme could be a *modus operandi* typical of the satirical world. To corroborate this observation, further journalistic information needs to be gathered.

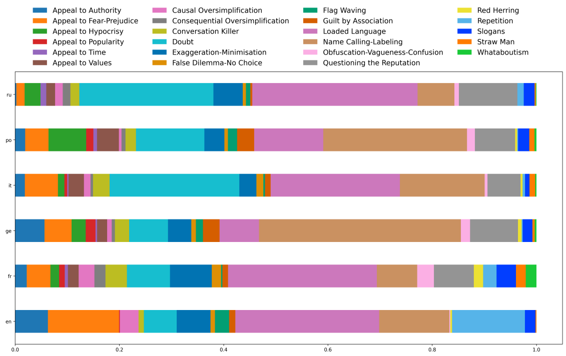
4.3.7 Persuasion Techniques Distribution given Framing

What happens to persuasion techniques choice given a specific framing? Do the differences or similarities between languages become more apparent?

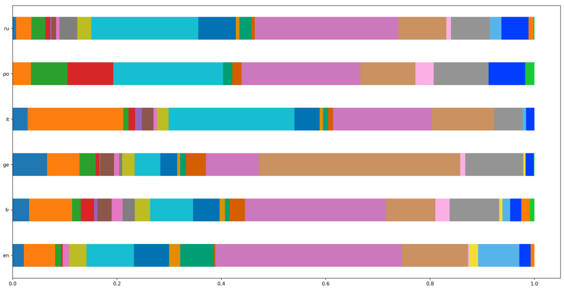
This time, the filter to be applied concerns the framing with which a given topic is approached and the procedure again consists of counting the relative frequencies. Clearly in this case, the number of combinations soars, with as many as 23 persuasion techniques and 14 framings. However, the information from such a representation is often not particularly relevant, keeping the behaviour almost unchanged or modifying it slightly, so for the sake of brevity only a few of the possible combinations will be reported. The first graph on the right in Figure 4.12, concerns the framing "health and safety" and, bearing in mind what was deduced from the analysis in Figure 4.8, it is possible to note that for the articles in Italian there is a slight increase in "appeal to authority", "appeal to fear-prejudice" and "repetition". If, on the other hand, from the graph concerning "external regulation and reputation" and in particular the documents written in French there is an appreciable decrease in the "loaded language" technique, although it remains the most popular. For the Polish language, on the other hand, a doubling of the relative contribution made by the "appeal to hypocrisy" can be observed, although this technique remains at the bottom of the ranking. Selecting only those articles in which the framing "fairness and equality" appears, then that both "loaded language" and "appeal to hypocrisy" are used more frequently by Russian journalists at the expense of "doubt", which instead suffers a slight decline. For Polish documents, on the other hand, "name calling-labelling" increases dramatically.



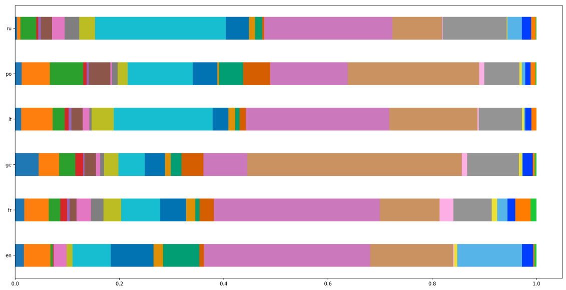
(a) Persuasion techniques given opinion class



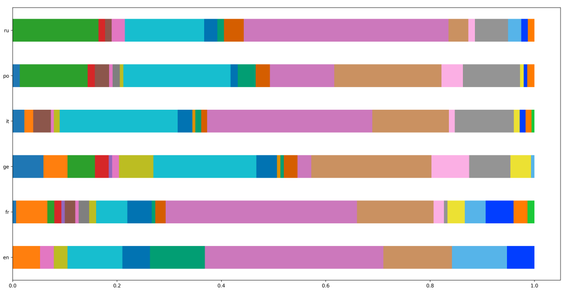
(d) Framings given Health and Safety framing



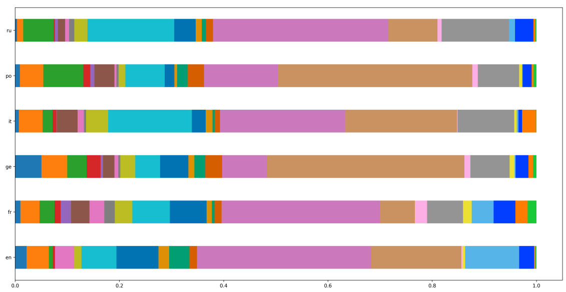
(b) Persuasion techniques given opinion class



(e) Framings given External Regulation and Reputation framing

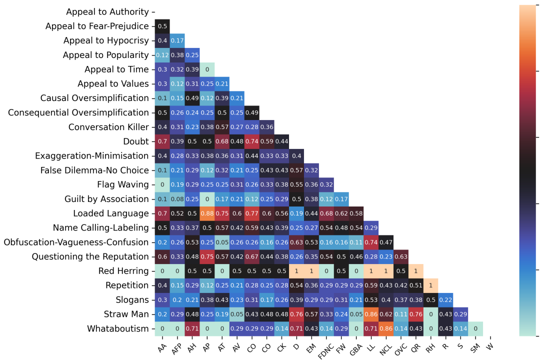


(c) Persuasion techniques given opinion class

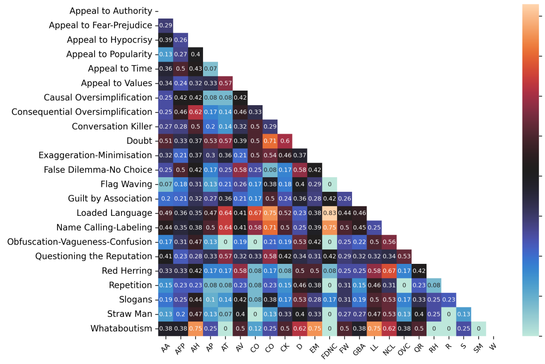


(f) Framings given Fairness and Equality framing

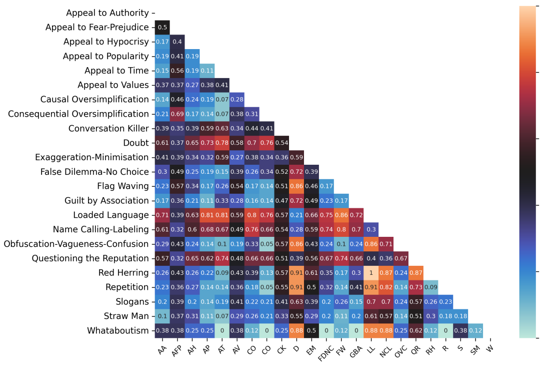
Figure 4.12: Persuasion techniques relative distribution across the six languages given a category or framing as filter



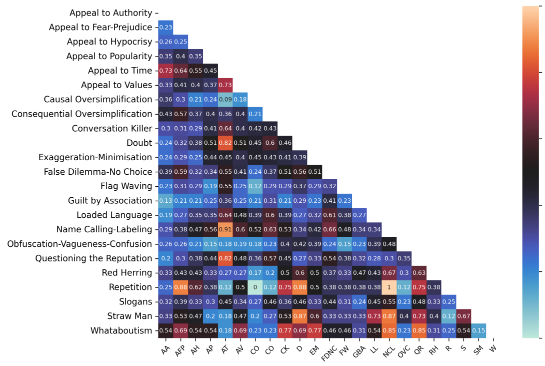
(a) Russian



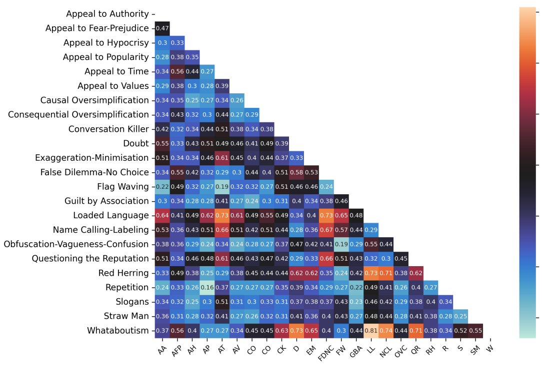
(b) Polish



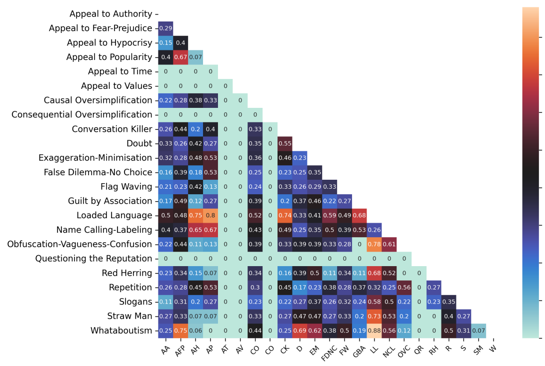
(c) Italian



(d) German



(d) French



(e) English

Figure 4.13: Correlation matrix with persuasion techniques' relative frequency for each language

5

Methodology and Experiments

In this chapter, the procedure followed for solving the classification tasks is presented step by step, in the same order as in the practical phase.

The first element of difficulty encountered concerns the handling of texts with a length greater than the selected model can handle, and was overcome by introducing a technique of dividing the text into smaller components. Where this technique was applied, it was done for all the texts available: training, validation and test set. In order to improve classification performance, two different data augmentation approaches were explored, having well presented the nature of the classes and labels to be attributed to each article, and taking care to synthesise datasets that were balanced in linguistic distribution. Lastly, the computational conditions under which all the experiments were carried out are presented, and all the tools used in both the training and evaluation phases are presented, dwelling on the aspects that motivated these choices. The last section is dedicated to presenting the results achieved during the training of the models, for each of the classification tasks considered, reserving a specific subsection for each of them.

5.1 Text Windowing

In tackling the tasks presented in the previous sections, one encounters an element of difficulty related to the size of the texts and the natural limitations of the model in use. In fact, the version of `XLM-RoBERTa-large` accessible from the HuggingFace platform, which is the algorithm used for the resolution of each of the three tasks under consideration, requires that each encoded input of the network be made up of 512 contiguous tokens (at most) plus special ones. However, already a quick analysis of the articles reveals that a substantial number of texts already have well over 512 words, which translate into even more tokens. Proceeding with the encoding of the textual input by asking not to include special tokens, it is obtained that more than 70% of all the articles at our disposal have a number of tokens exceeding the input size required by the network (Figure 5.1).

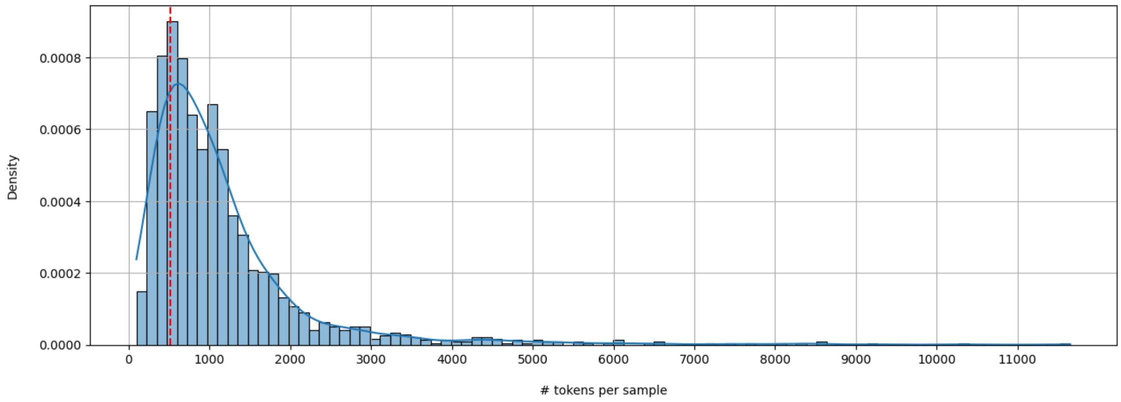


Figure 5.1: Tokens distribution in the whole dataset

If the type of task permits, and If it is known that relevant information for the task is in the first few lines of the text, one can proceed with a brute truncation of the sample to the number of tokens one have chosen. In this case, all special tokens will be added at the encoding stage. This approach may prove satisfactory in the resolution of a multi-class classification task such as the one described in section 3.1, i.e. the categorisation of online news articles. By definition, a journalistic article must, from the very first lines, fit into the characterisation of the category to which it belongs, displaying all the relevant narrative specifications. Added to this is the practical necessity of capturing the reader's attention right from the start, giving them an immediate idea of the type of article they are reading. For example, satirical articles base their effectiveness on a biting irony that does not hesitate to reveal itself while an opinion article will try to emotionally involve the reader without wasting time.

In the case of the remaining tasks, however, i.e. multi-label classifications with different levels of granularity, this approach is not to be considered a viable option for several reasons. In the case of framing detection, even though it is a classification at the level of the entire text, truncating the item by eliminating what does not fit within the granted limits would not only be a significant loss of information but also a conceptually incorrect operation. In fact, it is true that it is not known the exact portion of text to which a given framing refers but it is that each label *could* refer to any part of the article, even to the last sentences for instance. Therefore, deleting part of the text and keeping all the labels originally associated with the sample, could affect the correctness of attribution between features and target.

To get a clearer idea, consider this short excerpt from a newspaper article about the farmers' protests from first an economic and then a political perspective. Two framing labels are associated with the entire article: economic and political. Suppose one want to select only the first part of the text, excluding that portion highlighted in light blue. Since it is not known a priori the exact location of the two, it is possible to associate the same labels with the truncated text, thus making an attribution error because the part to which the framing actually referred has been eliminated.

> In recent weeks, tractor protests have captured the attention of the nation, highlighting the deep-seated economic grievances of farmers. These demonstrations, characterized by blockades and rallies, underscore the challenges facing agricultural communities amidst evolving government policies and economic pressures.
>
> At the heart of the protests lies frustration over issues such as declining farm incomes, increasing production costs, and inadequate government support. Farmers are demanding fairer prices for their produce, better access to credit and markets, and greater protection against volatile market forces. [ECONOMIC]
>
> The protests have also become a symbol of broader discontent with government policies perceived as favoring corporate interests over the welfare of small-scale farmers. As tensions escalate, the protests serve as a stark reminder of the urgent need for meaningful dialogue and policy reforms to address the underlying economic and political inequalities plaguing the agricultural sector.[POLITICAL]
>
> $\Longrightarrow$ [ECONOMIC, POLITICAL]

$\downarrow$

> In recent weeks, tractor protests have captured the attention of the nation, highlighting the deep-seated economic grievances of farmers. These demonstrations,

characterized by blockades and rallies, underscore the challenges facing agricultural communities amidst evolving government policies and economic pressures. At the heart of the protests lies frustration over issues such as declining farm incomes, increasing production costs, and inadequate government support. Farmers are demanding fairer prices for their produce, better access to credit and markets, and greater protection against volatile market forces. [ECONOMIC]

$\Longrightarrow$ [ECONOMIC, POLITICAL]

A similar argument can be made for the task of searching for persuasive techniques in the text, whatever level of granularity one decides to consider.

A rather widespread technique - as well as known for its versatility since it can be used on data of various kinds (text, images, audio, etc.) - is the Data Windowing. In the NLP context, this consists of dividing text into smaller portions of a specific size, called *windows* or *chunks*, and using them as processing units for analysis or model training. The objective of text windowing is to capture the local information contained in the text, allowing machine learning models to examine the surrounding context of the words or phrases of interest while not exceeding the threshold of an any given number of tokens. There are various approaches to implement this technique and generally when working with text, the individual elements to be grouped in a window are words. In this case, however, due to the need to comply with size requirements, it is better to work in terms of tokens. Each article will first undergo the encoding procedure - asking the `Tokenizer` not to include the special tokens which will be added manually after the windowing procedure - in order to obtain tensors of various sizes, whose elements are given by the token identification codes.

In the simplest version, the subdivisions of the text will be adjacent, as shown in the figure 5.2: assuming N the number of tokens in the entire original tensor and L_{chunk} the fixed length of the window (i.e. the number of tokens in the chunk), the number of extracted windows C is given by:

$$C = \left\lfloor \frac{N}{L_{chunk}} \right\rfloor + 1$$

in the latter case, the last chunk has a shorter length than required and so it is possible to proceed with padding, adding as many tokens `[PAD]` as necessary to reach L_{chunk}.

However, this selection criterion presents a criticality that cannot be underestimated, especially when handling textual data. Supposing a text is composed of two sentences S_1 and S_2, separated from each other by a dot, and one wish to subdivide this text into chunks of length $L_{chunk} = 30$. The entire text is subjected to the procedure of conversion into tokens and

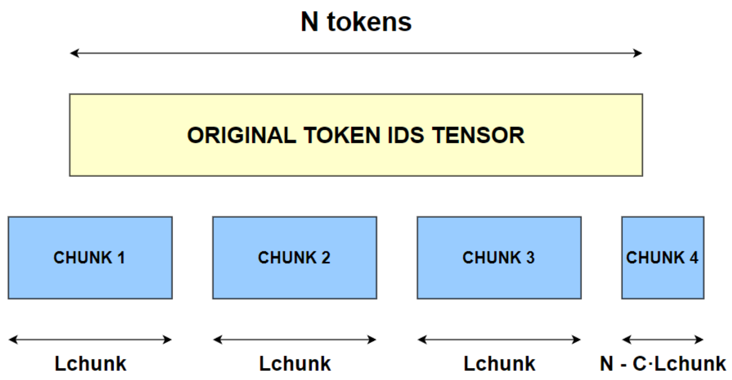


Figure 5.2: Consecutive Chunking procedure

then divided into as many windows as can be obtained given N and L_{chunk}. In almost all cases, the window will not cut the text precisely at the appropriate punctuation; it follows that the context as well as the general meaning of the sentences is not preserved within the windows.

[S$_1$] The migrant crisis in Poland has sparked debates on immigration policies and the country's response to humanitarian challenges. [S$_2$] As tensions escalate, there is a pressing need for coordinated efforts and compassionate solutions to address the plight of migrants.

$\downarrow$

581, 43017, 52028, 23, 164111, 1556, 131999, 297, 29865, 7, 98, 6, 199417, 102880, 136, 70, 23295, 25, 7, 57553, 47, 75757, 3378, 127125, 5, 1301, 63672, 7, 57826, 67, 4, 2685, 83, 10, 24234, 214, 3871, 100, 176866, 297, 79825, 136, 375, 126365, 2182, 51347, 47, 29823, 70, 915, 20016, 111, 43017, 7, 5]

$\downarrow$

[**Chunk**$_1$] [CLS] 581, 43017, 52028, 23, 164111, 1556, 131999, 297, 29865, 7, 98, 6, 199417, 102880, 136, 70, 23295, 25, 7, 57553, 47, 75757, 3378, 127125, 5, 1301, 63672, 7, 57826, 67 [SEP] [**Chunk**$_2$] [CLS] 4, 2685, 83, 10, 24234, 214, 3871,100, 176866,297,79825, 136, 375, 126365, 2182, 51347, 47, 29823, 70, 915, 20016, 111, 43017, 7, 5, 1, 1, 1, 1, 1 [SEP]

In order to limit this phenomenon and allow the algorithm to be clear about the context of each sentence (at least once) without sacrificing text coverage, a variant of the procedure described above, called *sliding windows*, was considered. In this strategy shown in figure 5.3, a sliding window of fixed size L_{chunk} is moved along the text document with a specified step L_{step}. Again, if the last window does not meet the standard size, missing elements are added. In questo caso il numero di finestra create è dato da:

$$C = \left\lceil \frac{N - L_{chunk}}{L_{step}} \right\rceil + 1$$

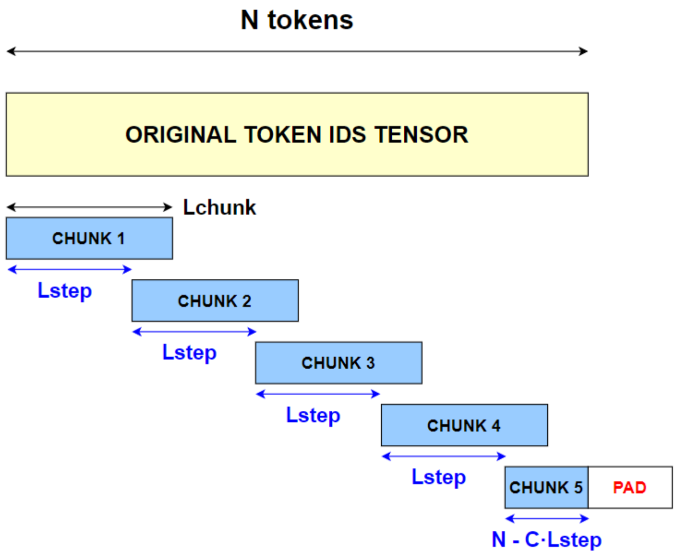


Figure 5.3: Chunking procedure with windows step

Here again, however, the problem of mismatching between chunk and target label persists. However, considering all the information provided by the text this problem could be mitigated, as at least one of the chunks associated with an article will certainly contain the framing persuasion technique considered problematic in this respect.

As anticipated, this windowing procedure was not followed to deal with the categorisation of articles, while the following parameters were set for the remaining tasks: $L_{chunk} = 510$ and $L_{step} = 255$.

5.2 Translation with mBART

Text augmentation via text translation is particularly useful when working with limited NLP datasets, as it allows for efficient dataset expansion and improved model performance on a variety of tasks, including text classification. Translation is an even more recommended technique in a multilingual set-up, being able to distribute information inter-language. The pre-trained model selected for this purpose is mBART, a multilingual version of the well-known BART algorithm, having been trained on a wide range of idioms and being able to translate to and from any pair of languages (among those supported).

On average, the journalistic articles contained in the SemEval-2023 dataset are medium-long texts so the translation of each element may take a fair amount of time. It should also be remembered that the articles assigned to the training dataset change in the three tasks under consideration, so the data augmentation procedure must be performed separately for each of them. It follows that it is strictly necessary to take steps to optimise the procedure with the idea of saving both time and resources.

Each article is accompanied by an identification code, so the first step was to identify those articles designated in the training set for each of the 3 tasks.

English	German	French	Italian	Polish	Russian	Total
426	97	119	170	106	107	1025

Table 5.1: Common training data across languages

Then it was necessary to determine how much new data to produce for each article and the translation criterion to follow, i.e. in which languages to translate each text. The first point is quickly resolved as it is generally not recommended to translate the same text into several languages both because the procedure would take an unreasonable amount of time and because it would become counterproductive. Increasing the data too much could lead to the dilution of useful information in the original data. If artificially generated data are not of high quality or do not adequately capture the variety and complexity of real data, they may not contribute significantly to improving model performance. Augmented data may introduce noise into the training dataset, as it may contain incorrect or misleading information. For example, in the case of machine translation, if the artificially translated data is not accurate, it may compromise the overall quality of the translation model. Finally, although the main goal of data augmentation is to prevent overfitting, there is a risk that overfitting may instead increase the risk. If artificially generated data do not accurately represent the distribution of real data, the model may learn spurious or irrelevant patterns that only occur in augmented data.

Therefore, it was decided to proceed as follows. (i) Translation will take place at a ratio of 1:1, one translated article for each original article. (ii) The criterion followed for the assignment of the target language does not take into account category class distributions or framing and persuasion techniques labels, but rather will aim to bridge the gap between the number of articles in the various languages. Then let L_s be the set of original samples written in the source language l_s. The articles were randomly divided into five L_{t_i} groups of equal size (each one with the respective target language $l_{t_i} \neq l_s$) and where samples remained unassigned, they were translated into the language with the fewest articles in the original dataset. As English is the language with the largest number of samples in the original dataset, it was decided to give priority to the less populated ones (typically German or Russian). One is aware of the fact that English is generally the language on which the pre-trained algorithms have been trained the most, but in this specific project, having a multi-language overview takes priority.

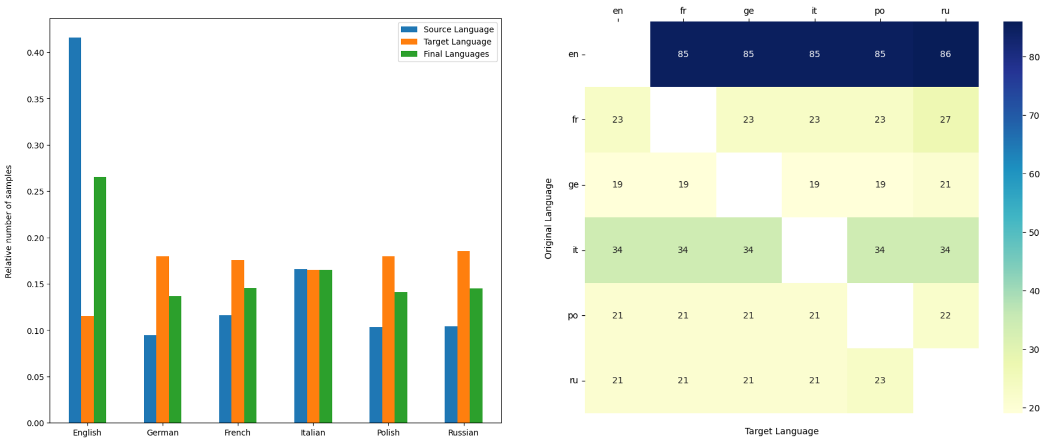


Figure 5.4: Common training data translation across languages

(iii) Translate the remaining training samples for each of the task datasets (respectivelly with 209, 213 and 226), maintaining the same criterion: give preference as targets to those languages which occurred in smaller quantities in the original dataset.

Once the subdivision of the samples and the assignment of the target language had been established, it is possible to proceed with the actual translation of the articles. Once again, resorted to a pre-trained model available on the HugginFace paittaform, specifically the model `mBART` for Condition Generation `facebook/mbart-large-50` (with matching tokenizer). Again, the problem of the maximum limit of input tokens the model can support arises, so it was decided to split each article into shorter sentences, corresponding to a new paragraph or a full stop. This was in light of the fact that the translator seemed to perform better when handling sentences of short to medium length.

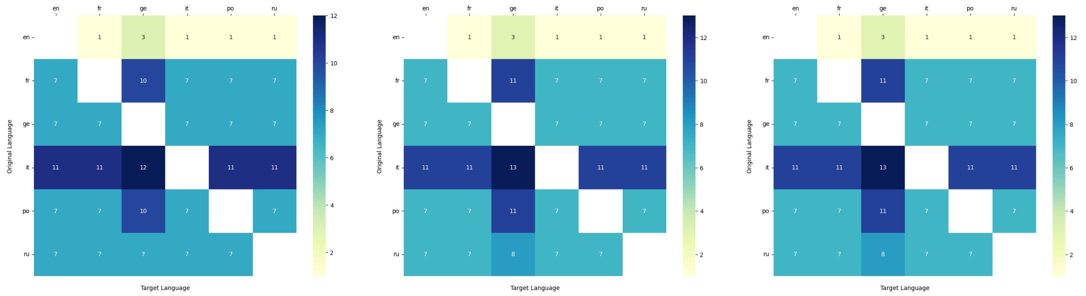


Figure 5.5: Tasks training data translation across languages

5.3 Text Generation with ChatGPT-3

Using the information contained in the dataset to produce new data is certainly helpful but, as already explained in the 5.2 section, there is a limit to the number of new items that can be introduced basing on those already present. Ideally, looking for new journalistic articles is a preferrable option but the problem of assigning the correct classes and labels would arise. In the first case, there is a wider margin of tolerance since the category of the document is an easy piece of information to obtain, whereas in the case of perspective framing and persuasion techniques, it is advisable (or even *condicio sine qua non*) to have a team of experts who can correctly identify them in the text. An alternative to data crawling are generative NLP artificial intelligence algorithms, which are able to produce new texts from existing training data or specific user-supplied input. The model chosen for the procedure is `ChatGPT`, an artificial intelligence developed by OpenAI based on the `GPT-3` (Generative Pre-trained Transformer 3) architecture trained on a large amount of text from the Internet.

Before even proceeding with the generation of new articles, it was necessary to ensure that the algorithm had the requested information and was aware about the correct definitions for each of the category classes, framing and persuasion technique labels. The input request was as follows:

> Generate articles in l language that meets all these queries:
>
> - has a minimum of 300 words and a maximum of 1000;
> - relates to a topic of common interest of choice among: political figures, climate change, war in Ukraine, immigration policy, abortion, civil rights, Olympics, conferences, etc.
> - belongs to only one of three journalistic categories: opinion piece, journalistic reporting and satirical article;

- exposes the topic from at least 3 of the following 14 framings: Capacity and resources , Crime and punishment , Cultural identity , Economic , External regulation and reputation , Fairness and equality , Health and safety , Legality Constitutionality and jurisprudence , Morality , Policy prescription and evaluation , Political , Public opinion, Quality of life , Security and defense;
- presents many persuasive techniques among these 23: Appeal to Authority, Appeal to Fear-Prejudice, Appeal to Hypocrisy, Appeal to Popularity, Appeal to Time, Appeal to Values, Causal Oversimplification, Consequential Oversimplification, Conversation Killer, Doubt, Exaggeration - Minimisation, False Dilemma-No Choice, Flag Waving, Guilt by Association, Loaded Language, Name Calling-Labeling, Obfuscation - Vagueness - Confusion, Questioning the Reputation, Red Herring, Repetition, Slogans, Straw Man, Whataboutism.

The procedure took more than 6 hours, plus the time needed to process the data and make it useful for each of the classification tasks algorithms. Furthermore, supervisor's interaction with the generating algorithm was constant and careful, as was the monitoring of the texts produced and the analysis of the distributions of category classes and labels. The aim was not only to generate quality data but to have a varied and complete information: if any framing or persuasion technique was not used sufficiently or in the correct manner, the supervisor intervened by pointing out the need to change the subject, to focus more on certain categories of text or to use a specific list of labels explicitly indicated. Sometimes examples were given on how to set up the narrative. Thus, 1102 new articles were generated, evenly distributed among the category classes and with a good variety of topics, framing and persuasion techniques. Again, it was decided to prioritise those languages that occurred less frequently in the original dataset.

English	French	German	Italian	Polish	Russian	Total
100	200	202	200	200	200	1102

Table 5.2: Common training data across languages

Compared to the original dataset, the synthetic one generated with `ChatGPT` presents a more homogeneous distribution of articles in the three category classes, whatever the language of interest. Indeed, there is a considerable amount of articles with a satirical slant, unlike the formerly frequent opinion text (see figure 5.6).

The first aspect to observe concerns the consistency between the framing distributions in the articles for the six language datasets generated with `ChatGPT`. With respect to the original

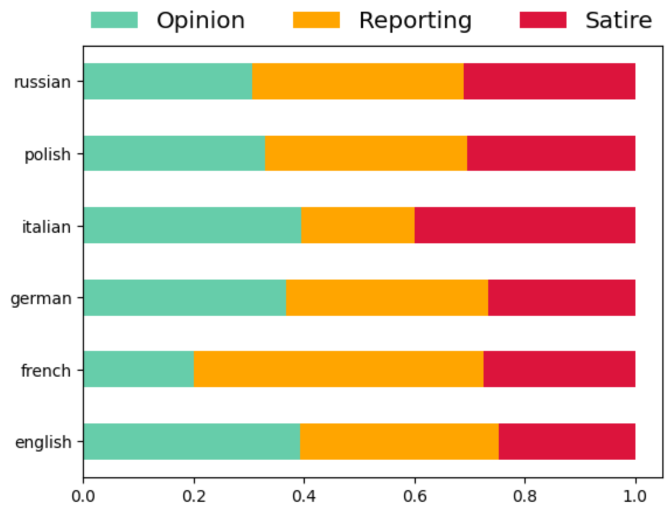


Figure 5.6: Category classes distribution among languages

frequencies, the various labels generally occur with the same relative frequency in the six generated language datasets, with a few minor exceptions such as “crime and punishment” occurring more frequently in English articles than in the other languages. If, on the other hand, it is informative to compare the incidences of a given label in the original vs. generated dataset pairs in a specific language, it is possible to note, for example, that “crime and punishment” is generally more frequent in the former (with the exception of Polish and French). Please consult the graph in figure 5.7 if one wish to further explore the comparison between datasets and between languages.

A similar argument can be made for the analysis of the persuasion techniques generated. Comparing the results obtained for each language dataset leads to the same conclusions, i.e. the persuasion techniques present in the generated articles respect the same proportions across languages. Generally speaking, the most frequent techniques are “Appeal to Authority”, “Loaded Language” and “Red Herring. The latter is definitely a valuable boost to increase the variety of the dataset on which to train the classification algorithm as it was little used in the original dataset. For further investigation rely on the graph in figure 5.8.

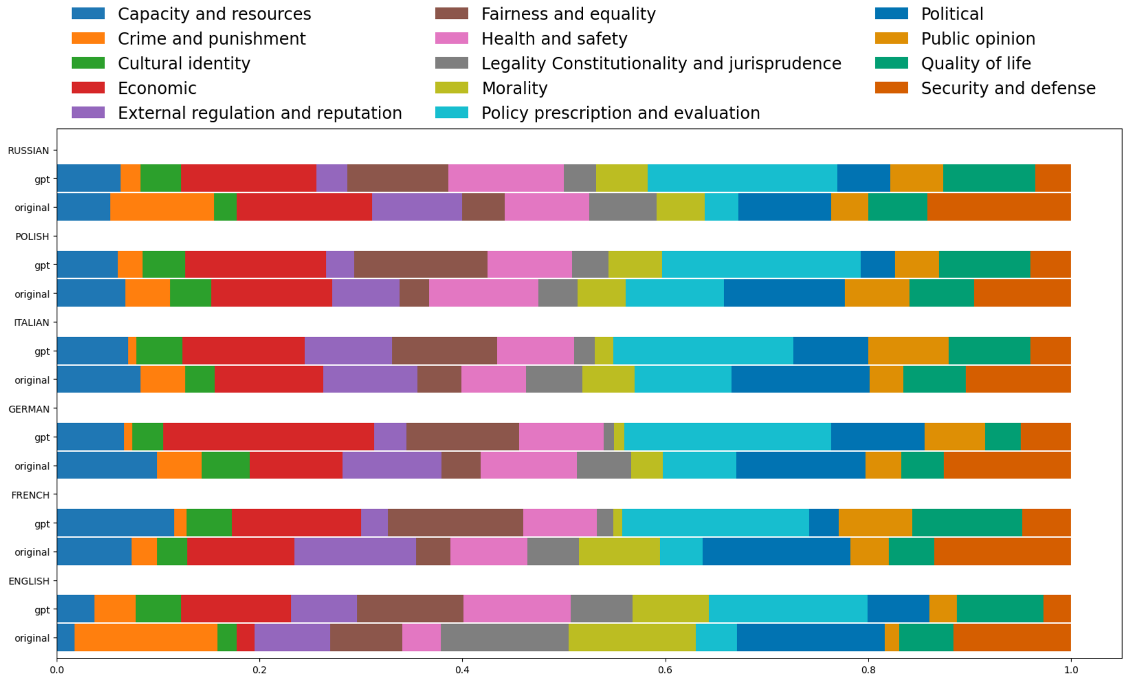


Figure 5.7: Framing labels distribution among languages

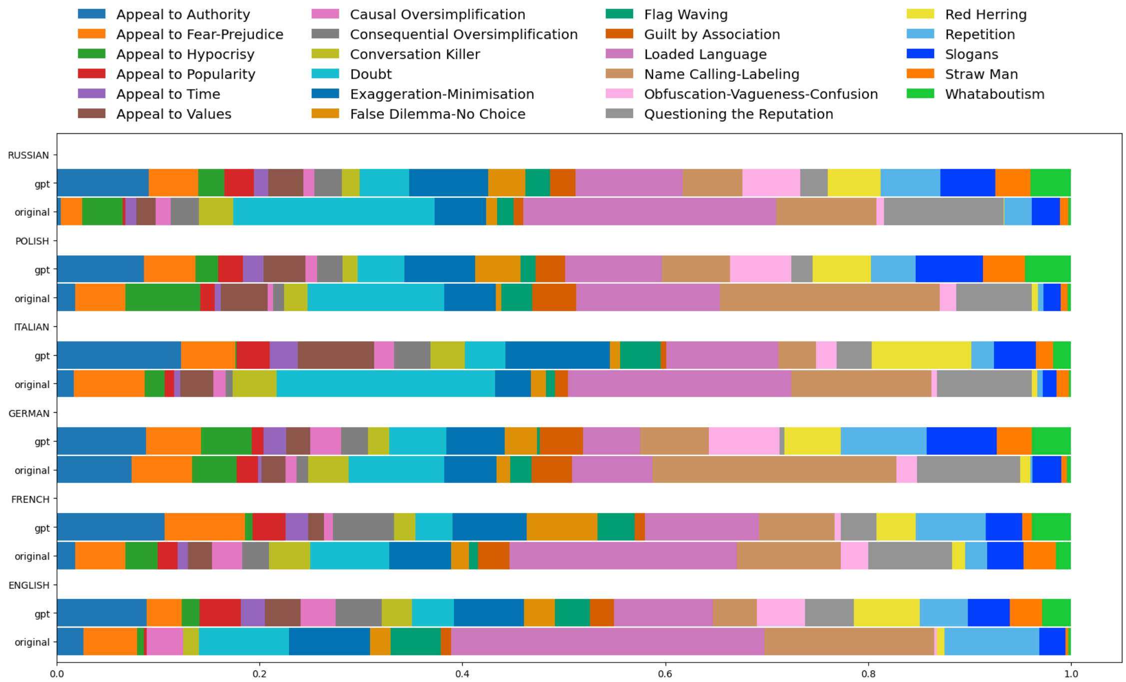


Figure 5.8: Persuasion technique labels distribution among languages

5.4 Training Setup

Architectures designed for solving NLP tasks generally present a marked complexity, especially if one wants to guarantee a satisfactory performance in predictions. They are extremely complex and layered neural networks, with a number of parameters ranging from tens of millions (e.g. `T5-small`) to the order of hundreds of billions (e.g. `GPT-3`). Training one of these models from scratch would require an unreasonable amount of resources and time, as well as being a decidedly unsustainable practice. It is then advisable to find an alternative to the bare-bones training of these heavy models: *fine-tuning* is one such technique that allows to adapt pre-trained neural networks for specific tasks or datasets. It is a form of transfer learning which entails making subtle adjustments to the model's internal parameters with the aim of achieving better model's performance on a new, related and specific task without necessitating training from scratch. This approach significantly saves time and computational resources while effectively adapting the model's capabilities to specific applications. Indeed, by starting with a model that has already learned many relevant features, it is possible to skip the initial stages of training and focus on customizing the model to the particular task at hand. All classification tasks in

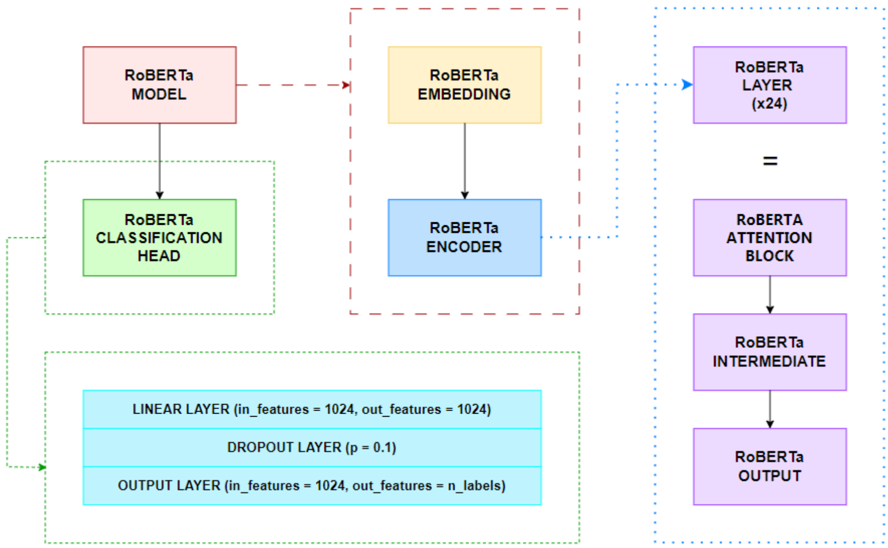


Figure 5.9: `XLM-RoBERTa-large` architecture

this project make use of the pre-trained `XLM-RoBERTa-large` model available on the Hugging Face platform, by making a change on the classification head depending on the specific task and fine-tuned on the related dataset. The number of output nodes in the last fully-connected layer will in fact be equal to the number of classes or labels: 3 for category classification task, 14 for the framing labels one and 23 for the persuasion techniques labels (indepentely from the granulary level). The main structure of the network is given by an initial embedding block (`RoBERTa Embedding Block`) and 24 `RoBERTa Layer`, each one consisting of three blocks that follow one another in the order: attention block mechanism, intermediate dense layer and a dense output layer followed by normalisation and dropout (see Figure 5.9).

Loss Functions Different loss functions were selected from `PyTorch` library for each classification task.

In the case of category multi-class classification task with number of classes $C = 3$, the adopted criterion is the `CrossEntropyLoss`*. The input is a tensor of size $(minibatch, C)$, or simply C if there are no minibatches, containing the unnormalized logits for each class without the request of being positive or sum to 1. The loss function is given by the formula:

$$l(x, y) = \sum_{n=1}^{N} \frac{l_n}{\sum_{n=1}^{N} w_{y_n} \cdot \mathbf{1}\{y_n \neq \text{ignore\_index}\}} \tag{5.1}$$

where x and y are respectively the input and the target, w is the weight parameter, N is the minibatch dimension and l_n is defined as:

$$l_n = -w_{y_n} \log \frac{\exp(x_{n,y_n})}{\sum_{c=1}^{C} \exp(x_{n,c})} \tag{5.2}$$

.

For the remaining multi-label classification tasks (both for framing and persuasion techniques detection) the `BCEWithLogitsLoss`†. This version of loss function combines a Sigmoid layer and the Binary Cross Entropy loss in one single class in a more stable way than the traditional approach (a plain Sigmoid followed by a BCELoss): by combining the operations into one layer, taking advantage of the log-sum-exp trick for numerical stability. The formula for the multi-label classification case:

$$l_c(x, y) = \text{mean}(L_c) = \text{mean}\big(\{l_{1,c}, \ldots, l_{N,c}\}\big)^{\top} \tag{5.3}$$

*https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html
†https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html

where x and y are input and target respectively, N the batch size and l_n is defined as follows:

$$l_{n,c} = -w_{n,c}\big[p_c y_{n,c} \cdot \log\big(\sigma(x_{n,c})\big) + (a - y_{n,c}) \cdot \log\big(1_\sigma(x_{n,c})\big)\big] \tag{5.4}$$

with c class number, n number of samples in the batch and p_c is the weight of the positive answer for the class c.

Metrics The calculation of the loss function was accompanied by the evaluation of several metrics, with the aim of monitoring the model's performance during training. In particular, the two versions of the `f1-score` metrics (`macro` and `micro`) were calculated, using the library's default function `scikit-learn`:

$$F1_{\text{macro}} = \frac{1}{C}\sum_{i=1}^{C} F1_i = \frac{1}{C}\sum_{i=1}^{C} \frac{2 \cdot TP_i}{2 \cdot TP_i + FP_i + FN_i} \tag{5.5}$$

$$F1_{\text{micro}} = \frac{2 \cdot \sum_{i=1}^{C} TP_i}{2 \cdot \sum_{i=1}^{C} TP_i + \sum_{i=1}^{C}(FP_i + FN_i)} \tag{5.6}$$

where C is the total number of classes or labels, TP_i true positives, FP_i false positives and FN_i false negatives for class or label i. These metrics, in addition to the evaluation of the results on the test, were considered for the selection of the best model: for each task, the checkpoint that returned the best performance for the chosen metric on the validation set was considered. In particular, for the multi-class classification task the metric `f1score-macro` was considered, while for the remaining multi-label classification tasks the `f1score-micro` was evaluated. However, for the sake of completeness, both will be calculated for all cases.

Finally, for the task of classifying persuasion techniques at the token level, resulting in the detection of the reference span, the F1-score micro metric was manually evaluated along with other informative parameters such as the number of correctly classified tokens. Working at the level of individual tokens and knowing the correctly associated labels, the elements of the confusion matrix were counted with which the metric was calculated in its canonical definition.

Computational Set Up From a computational point of view, several aspects need to be taken into account in order to successfully fine-tune the model, first of all the memory capacity of the GPU, necessary for training deep learning models as it allows for higher performance and shorter training times than a CPU. The graphics card used is a `Quadro M4000 NVIDIA` with 8GB of memory, the driver is `CUDA Version 11.4` and the library `python` used to implement the neural network is `PyTorch`, (version `1.12.1+cu102`). The pre-trained version of `XLM-RoBERTa-large` in use has about 355 million parameters, the architecture alone with the pre-trained weights occupies a space of about 3GB and proceeding with the training, the total space required exceeds the space available here and the script returns an `Out Of Memory Error`. Generally in these cases, it is advisable to adjust the parameters and hyper-parameters of the model on the basis of the tools at one's disposal, e.g. by choosing a lighter model or decreasing the training batch size. Fortunately, the Hugging Face `transformers` library provides alternative tools to optimise GPU usage and refine complex models without sacrificing optimal performance. In particular, three tricks were used to be implemented as topics in the `transformers trainer class`: (i) mixed precision training, (ii) gradient accumulation steps and (iii) gradient checkpoint. Moreover, the (iv) `adafactor` optimizer was chosen [‡].

(i) The idea of mixed precision training is that no all variables need to be stored in full (32-bit) floating point precision: if it is possible to reduce the precision, the variables and their computations are faster. The main advantage comes from saving the activation in half (16-bit) precision. Although the gradients are also computed in half precision they are converted back to full precision for the optimization step so no memory is saved here. Since the model is present on the GPU in both 16-bit and 32-bit precision this can use more GPU memory (1.5x the original model is on the GPU), especially for small batch sizes. In conclusion, some computations are performed in full and some in half precision and this is way the approach is also called mixed precision training.

(ii) The concept of gradient accumulation involves breaking down the computation of gradients for the entire batch into smaller data subsets, called *mini batch*. This is achieved by iteratively computing gradients over smaller batches, performing forward and backward passes through the model, and accumulating the gradients throughout this process. Once a sufficient number of gradients have been accumulated, the model's optimization step is executed. By adopting this approach, it becomes feasible to augment the overall batch size to levels that would exceed the memory capacity of the GPU. However, it's worth noting that the additional forward and backward passes may introduce a slight slowdown in the training process. In all classification experiments, the same parameters were selected: `per_device_train_batch_size = 3` and `gradient_accumulation_steps = 4`.

(iii) Even when setting the batch size to 1 and use gradient accumulation one can still run out of memory when working with large models. In order to compute the gradients during the backward pass all activations from the forward pass are normally saved. This can create a big

[‡]https://huggingface.co/docs/transformers/v4.18.0/en/performance

memory overhead. Alternatively, one could forget all activations during the forward pass and recompute them on demand during the backward pass. This would however add a significant computational overhead and slow down training. The process could be appreciated more in details with an simple example of feed-forward neural networks with n-layers.

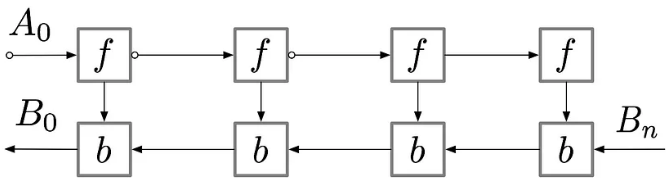


Figure 5.10: Simple feed-forward neural network

The activations of the neural network layers correspond to the nodes marked with an f: during the forward pass, all these nodes are evaluated sequentially. The gradient of the loss with respect to the activations and parameters of these layers is indicated by the nodes marked with a b: during the backward pass, all these nodes are evaluated in reverse order. The results obtained for the f nodes are necessary to compute the b nodes, and thus all f nodes are retained in memory after the forward pass. Only when back-propagation has progressed sufficiently to compute all dependencies of an f node, can it be removed from memory (see Figure 5.10). This implies that the memory required by simple back-propagation increases linearly with the n number of neural network layers.

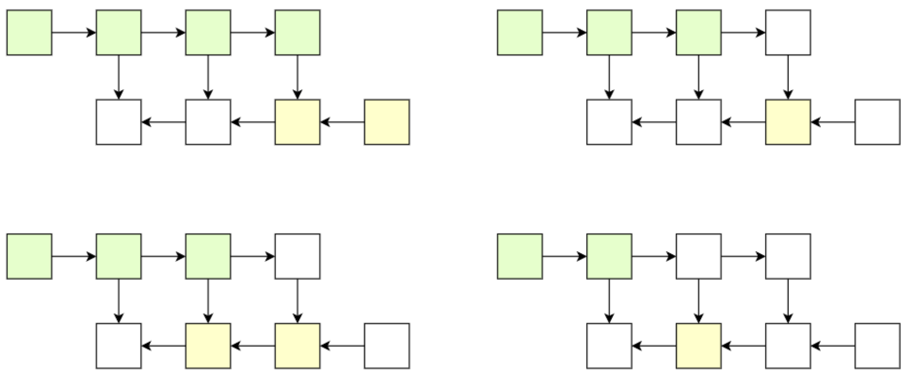


Figure 5.11: Traditional back-propagation computation steps

While simple back-propagation is computationally efficient as it computes each node only once, optimizing for memory conservation may lead us to consider recomputing nodes. For instance, it is possible to simply recompute every node from the forward pass whenever needed. This approach conserves memory optimally, but now the number of node evaluations scales quadratically, with n^2, compared to the previous linear scaling of n. Basically, each of the n nodes is recomputed around n times. However, the increased computational overhead renders this method impractical for deep learning tasks. To strike a balance between memory and computation, a strategy is needed thus allowing a not too frequent node recomputation. A common approach is to designate a subset of the neural network activations as checkpoint nodes.

In conclusion, gradient checkpointing strikes a compromise between the previous two approaches and saves strategically selected activations throughout the computational graph so only a fraction of the activations need to be re-computed for the gradients.

(iv) Last tool set up was the training optimizer: the `transformers` library manual suggests to choose `adafactor`. Instead of keeping the rolling average for each element in the weight matrices Adafactor only stores aggregated information (row and column-wise sums of the rolling averages) which reduces the footprint considerably.

Each model was fine-tuned with a learning rate of $3e{-5}$ and weight decay of 0.01. The metrics evaluation on validation set was performed at each 100 iterations and each checkpoint was saved at 100 or 200 iterations.

5.5 Experiments Results

In this section, the results of models' fine-tuning for each of the classifications considered will be separately analysed: a specific subsection is reserved for each task, where a pair of graphs is presented showing the performance of the metrics over the entire validation set every 100 iterations. The various techniques proposed in the previous sections, text windowing and data augmentation, were tested in different combinations. In particular, for the categorisation of articles, the impact of data augmentation was analysed, while the text windowing technique was also introduced for framings detection. As this was again a multi-label classification, for the detection of whole-text persuasion techniques, the text was split into smaller chunks and the original dataset was combined first with the data obtained via translation with `mBART` model and then with the articles generated with `ChatGPT`. For the high granularity variant, however, the experiment was performed only on the original dataset as it was difficult to determine with certainty the spans to which each technique referred. The number of iterations was automatically determined depending on the training progresses: if no relevant improvements were detected, the training process stopped itself to prevent side effects like a savage overfitting. The models that performed best on the entire validation set - considering the appropriate metrics

- will then be used to obtain predictions on the test set, differentiating the performances by linguistic subsets (see chapter 6).

5.5.1 Category Classification

Four models were fine-tuned for the articles' categorisation, all without the text windowing technique but different number of iterations: 2000 for the base version with the only original dataset (`base`), 1900 for the model with translated data (`mBART`), 2400 for that one with the addition of generated texts (`GPT`) and, lastly, 2200 iterations for model with a combination of the two data augmentation techniques. The metric used to evaluate the goodness of each model

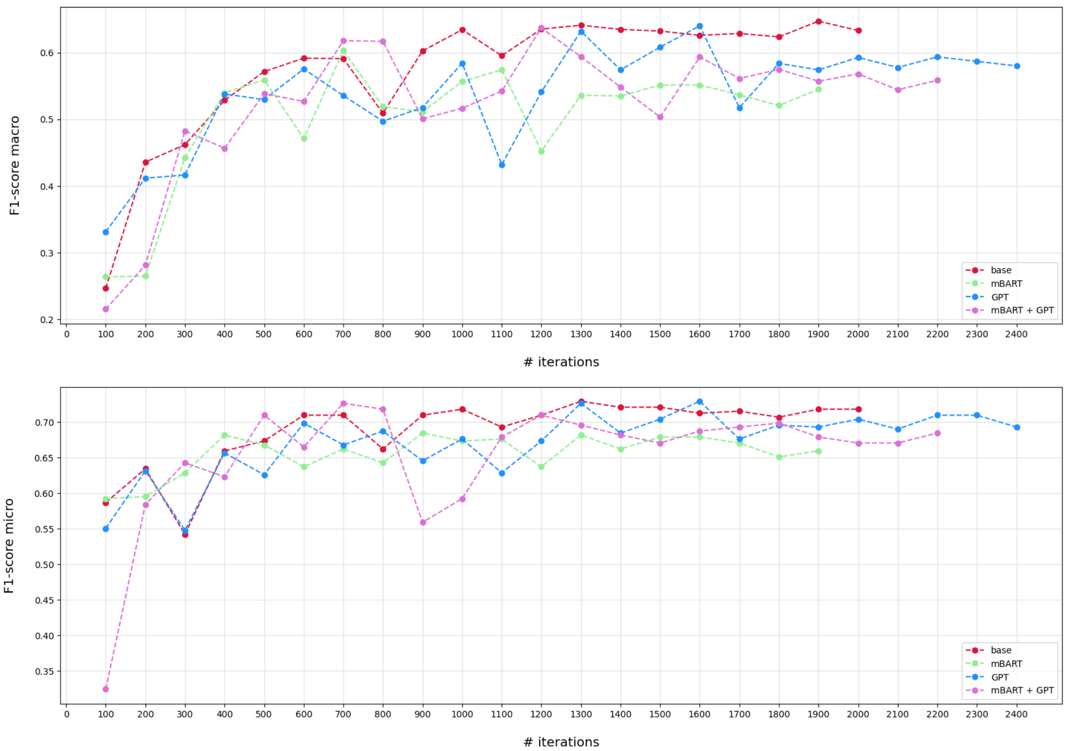


Figure 5.12: Category classification task: Macro and Micro F1-score on validation set

is the F1score-macro but the F1score-micro is shown too (look at Figure 5.12). The progression of the metrics over the entire validation set as the iterations vary for the base model (red) is on average more satisfactory: the curve already stabilises beyond 1000 iterations and is above the others at most checkpoints. However, the trends of the models with the addition of the synthesised articles (blue) and with the combined augmented dataset (pink) also return good results, sometimes managing to place themselves above the base model even if their behaviour is rather erratic. The addition of the translated data alone, on the other hand, fails to compete with the others (green). With validation performance in mind, four checkpoint models were therefore selected for each version to be used for the prediction on test set.

5.5.2 Framing Classification

Five models were fine-tuned for the framing labels detection, one trained without the text windowing technique but the remaining four with the text fragmentation: 2000 iterations were requested for the most basic version textttnochunk + base), 1700 for the corresponding version with the text windowing technique (`chunk + base`), 2700 iterations with the addition of translated texts (`chunk + mBART`), 2200 for the model with the generated dataset (`chunk + GPT`) and, lastly, 2700 iterations for model with a combination of the two data augmentation techniques. The metric used to evaluate the goodness of each model is the F1score-micro but

Figure 5.13: Framing labels classification task: Macro and Micro F1-score on validation set

the F1score-macro is shown too (look at Figure 5.12). An interesting consideration to be made here concerns the gap between the performance of the two basic models with and without the text windowing technique (red and gold curves): the results achieved by the latter on the validation set are decidedly more satisfactory, especially when looking at the F1score-macro. On the other hand, evaluating the benchmark metric, one can see that there is still an improved performance, but with a less obvious detachment. On the other hand, for this task, the data augmentation seems to have actually made some improvement: the translated and generated data introduced valid information that allowed for improved performance on the validation set, both separately (green and blue curves) and combined (pink curves) With validation performance in mind, four checkpoint models were therefore selected for each version to be used for the prediction on test set.

5.5.3 Persuasion Technique Classification

Two different levels of granularity were considered for this experiment: multi-label classification at the whole article level and at the token level, in order to identify the precise points in the text where a given technique occurs.

Text Level Classification In the first case, three models were fine-tuned, all using the text windowing technique: only with the original dataset, with the addition of only the translated data and of the only generated articles. Unfortunately, the introduction of data augmentation did not produce satisfactory results on the performance of the validation set, whatever metric one wishes to consider, especially in the case of the articles generated with `ChatGPT`. A possible explanation for this result lies in the quality of the data generated: the algorithm did not present a complete and articulate mastery in the use of persuasion techniques, especially in non-English language. A quick reading of some of the generated items reveals that there is some redundancy in the way the techniques are used, the same phrases with little variability and sometimes in the inappropriate way. For this reason, the last model was discarded and the best checkpoints from the first two models will be used to predict the results on the test set.

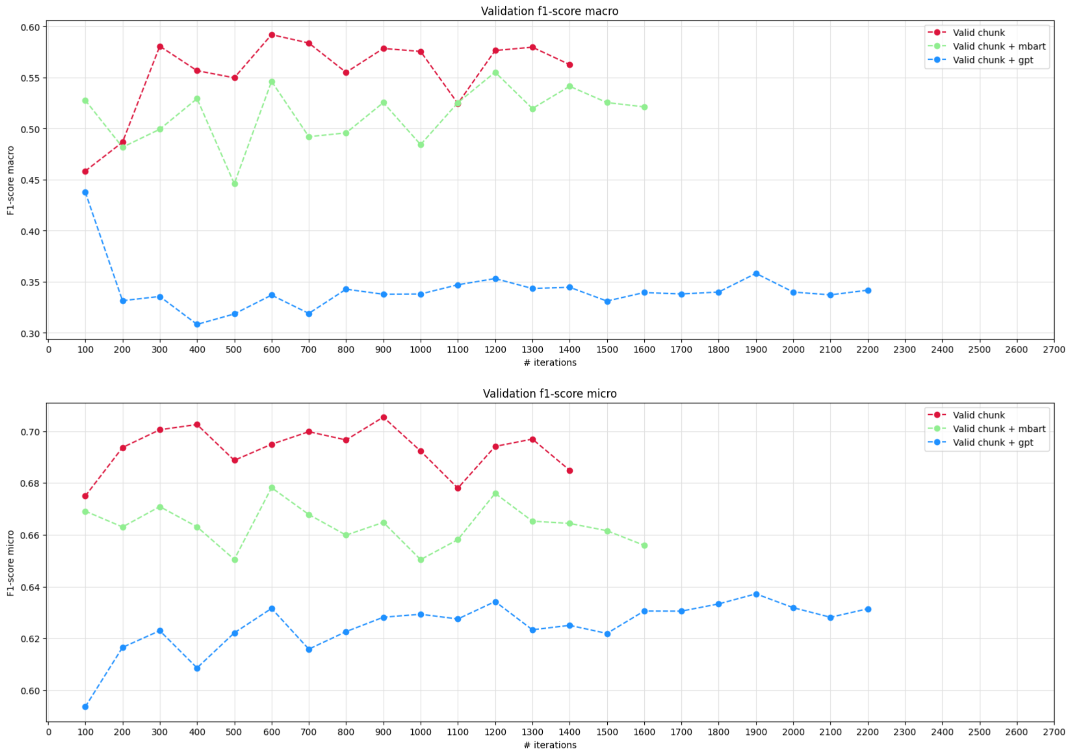


Figure 5.14: Persuasion Technique labels classification task at Text Level: Macro and Micro F1-score on validation set

Token Level Classification For the last of the tasks considered, and the one with the highest level of granularity, a single model was trained for 2000 iterations, using only the original data and applying the text windowing technique. In this case, only one metric was evaluated, which is accompanied by the graph showing the number of correctly identified tokens as a function of iterations. The curve concerning the performance on the validation set tends to stabilise after a few hundred iterations, varying in a range of values between 23% and 25% of F1-score micro. On the other hand, the number of correctly identified tokens exhibits a more pronounced instability, although it has an average increasing trend. As with the other models trained previously, the best checkpoint was selected as the model to be used to obtain the predictions on the test set.

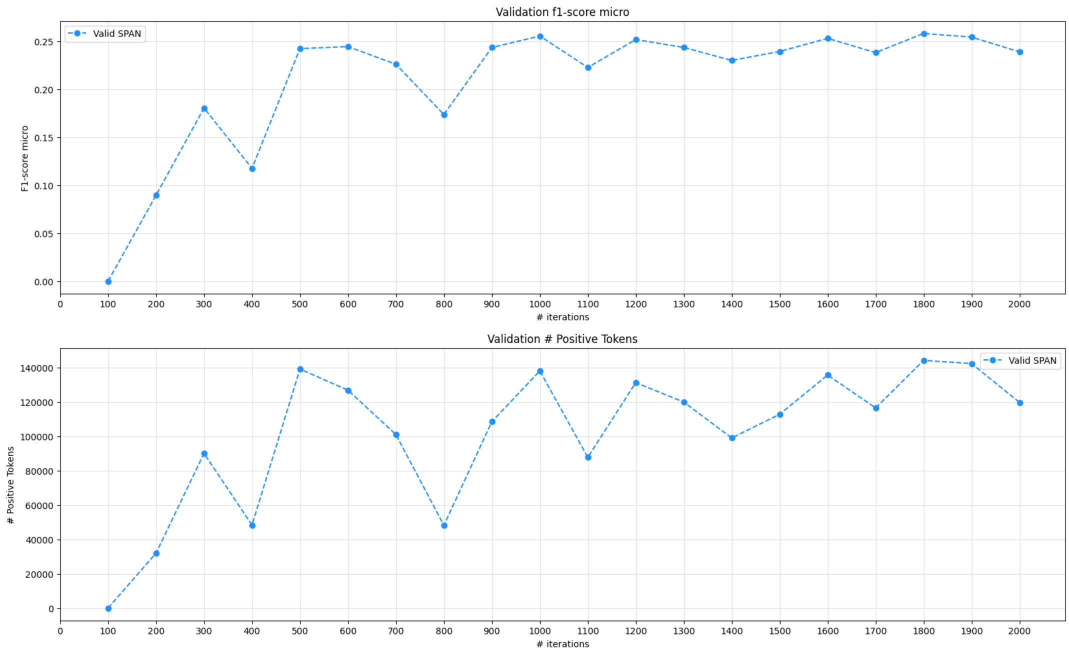


Figure 5.15: Persuasion Technique labels classification task at Span Level: Macro and Micro F1-score on validation set

6
Evaluation Results

This chapter presents the results and predictions that the selected best models returned on the test set, with a specific section devoted to each classification task. The performances on the metrics for each of the language subgroups (the one used for the evaluation procedure) are summarised in a table and the best ones are highlighted. This is followed by a comparative barplot in which the classes or labels predicted by each of the trained models are compared with the actual test data distributions, remembering that as in the previous sections, each barplot gives the relative distribution of each item over the whole belonging set. This graph gives us an initial overview of the incidence of each class or label in the entire set of predictions, however, it remains purely qualitative information. If one wishes to verify in actual fact how many elements have been correctly predicted so as to have a quantitative awareness, one must consult the confusion matrices shown immediately afterwards. In this case, results are shown only for those models that returned the best performance on the specific linguistic dataset. In the case of the framing and persuasion technique detection tasks, a small confusion matrix is reported for each label, again following the criterion just specified.

6.1 Category Classification

For the multi-class category classification task, the metric used for evaluation is the f1score-macro. Each of the checkpoints considered for the predictions had at least one of the best results in the six linguistic datasets, except for the model trained with only the addition of the translated data using the mBART architecture. The fine-tuned base model on the original dataset alone performed best on the German, French and Russian language datasets, while the model that also took into account the data generated with ChatGPT scored best for the English and Italian data. Particularly in the former case, it seems that the introduction of new data was rather profitable for the prediction of the category (40.0% vs 73.3%) and the same considerations can be made for the other data augmentation technique (40.0% vs 65.4%). Having enriched the Polish language dataset with both translated and generated articles led to an excellent result with a f1score-macro of 81.4%. The same cannot be said for German and Russian, whose categorisation seems to be rather affected by the introduction of new data.

Model	English	German	French	Italian	Polish	Russian
BASE	0.400	0.840	0.756	0.615	0.577	0.672
mBART	0.654	0.626	0.640	0.646	0.726	0.622
GPT	0.734	0.658	0.639	0.718	0.737	0.583
mBART + GPT	0.663	0.803	0.657	0.552	0.814	0.500

Table 6.1: Category classification task: macro F1 score on test set across six languages

Before analysing the actual predictions in the three classes of the six best models, we look at the graph 6.1 which gives us an initial description of the relative distributions. It immediately jumps to attention the massive presence of items predicted as "reporting" in the Russian dataset by the GPT model compared to the test dataset: this may already give us an idea of why the performance of this model is poorer than the base one. A similar observation can also be made for the English language dataset, with the difference that the other two classes cannot be said to be "crushed" by the reporting items. Here, in fact, it is the base model that performs worse and the graph points out that there were no predictions of satirical texts, which may have penalised the score overall. For the German language, there are no particular differences, so this graph cannot justify the discrepancy between the scores obtained with the four models. The distributions are probably correct for almost all of them but do not map the data correctly. The French dataset presents a distribution strongly in favour of opinion articles, which the mBART model massively predicts, but again the results on the metrics take us in another direction. The other two classes will have been predicted more carefully in the `base` model. Among the Italian language predictions proposed by the `GPT` model, there are few for the satire class, but still more than in the other models, which rather preferred the other two classes. There is a certain homogeneity in the Polish dataset, so it is plausible that this graph would not be able to pick up any major differences.

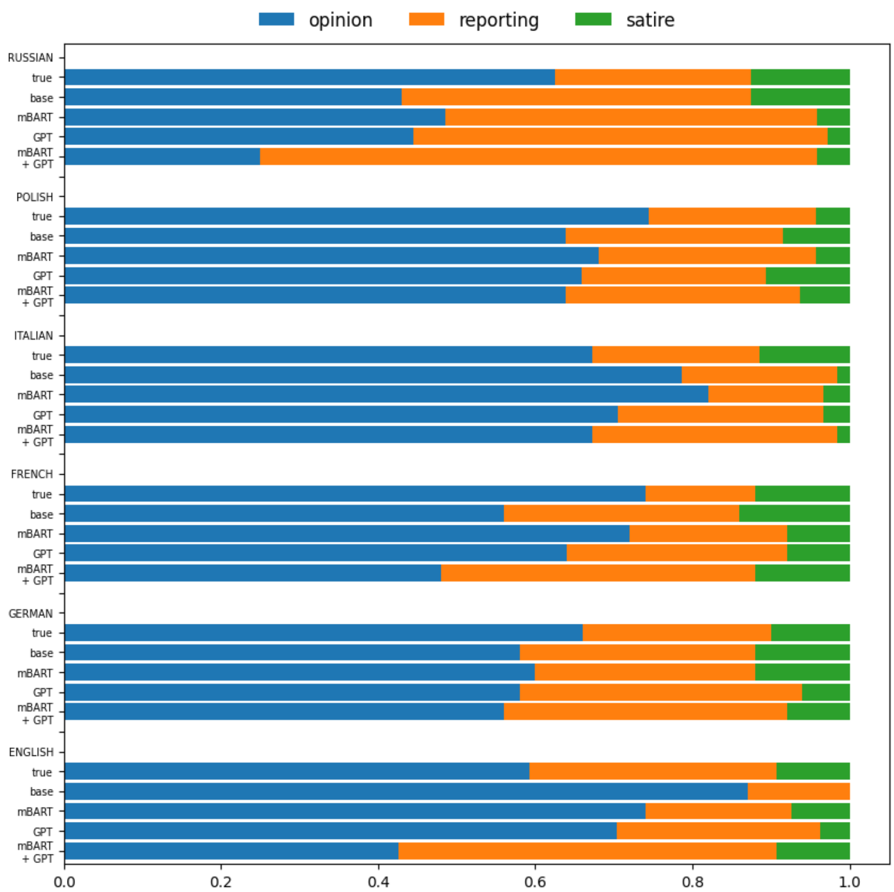
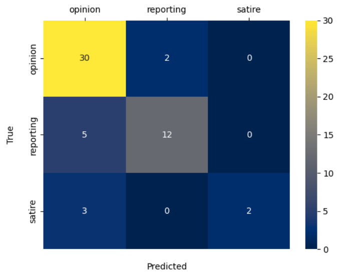
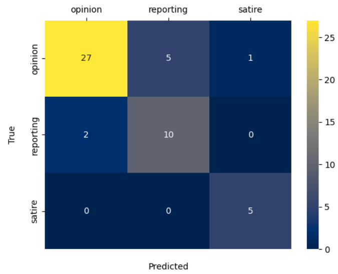


Figure 6.1: Category classification task: predictions distribution on test set across six languages

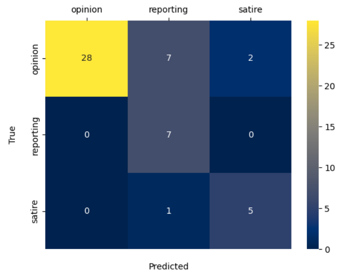
The confusion matrices provide us with more detailed quantitative information on the correctness of the predictions. For the English dataset, the `GPT` model performs very well in predicting the most frequent class while decreasing on satirical reports and articles. On the other hand, all German language "satire" classes are correctly predicted by the `base` model and we have more or less the same results for French articles, where the prediction of reports reaches the optimal level. Unfortunately, the `GPT` model does not perform well on the less populated class



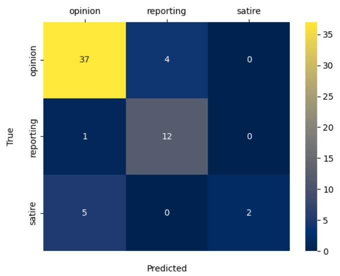
(a) Model with GPT augmentation for English test set



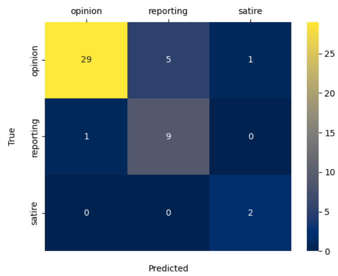
(b) Model with no data augmentation for German test set



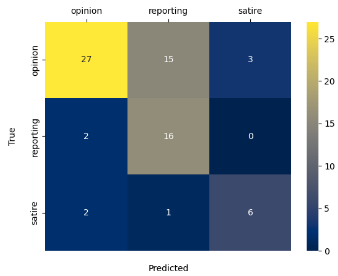
(c) Model with no data augmentation for French test set



(d) Model with GPT augmentation for Italian test set



(e) Model with mBART and GPT augmentation for Polish test set



(f) Model with no data augmentation for Russian test set

Figure 6.2: Framing label classification task: confusion matrices on test set for best model across six languages

while it achieves good results on the other two. The combination of the two data augmentation techniques produced excellent results on the Polish dataset, in all categories. Finally, opinion articles in Russian are often mistaken for reporting articles, so more work needs to be done on this dataset. The addition of text windowing might help.

6.2 Framing Classification

Four different models were trained to detect framing labels at the whole article level, all of which made use of the text windowing technique with the exception of the first basic model. The reference metric for this task is the F1-score micro for which all models returned average satisfactory results, whichever language subgroup is considered. In particular, the version enriched with the translated data in the English language resulted in a score of 63.1% followed by the model with the synthetic dataset generated with `ChatGPT`. The combination of the two data augmentation techniques ensured a better result for the German dataset than when considering them individually and slightly better than the version without new information (`chunk base`). Adding only the generated articles to the original French language articles proved to be the winning strategy, even though the chosen data augmentation techniques can be said to improve performance, whichever combination is considered. Similar considerations can be drawn from the Italian dataset, for which the model trained with the contribution of the translated data returned a score of 61.4% on the test set. The same model performed optimally on the Polish data while all the models considered generally performed worse on the Russian language dataset.

Model	English	German	French	Italian	Polish	Russian
NO CHUNK	0.509	0.628	0.510	0.577	0.628	0.444
CHUNK BASE	0.600	0.710	0.552	0.596	0.705	0.422
mBART	0.631	0.682	0.560	0.614	0.723	0.432
GPT	0.622	0.703	0.574	0.600	0.702	0.442
mBART + GPT	0.617	0.714	0.571	0.604	0.707	0.409

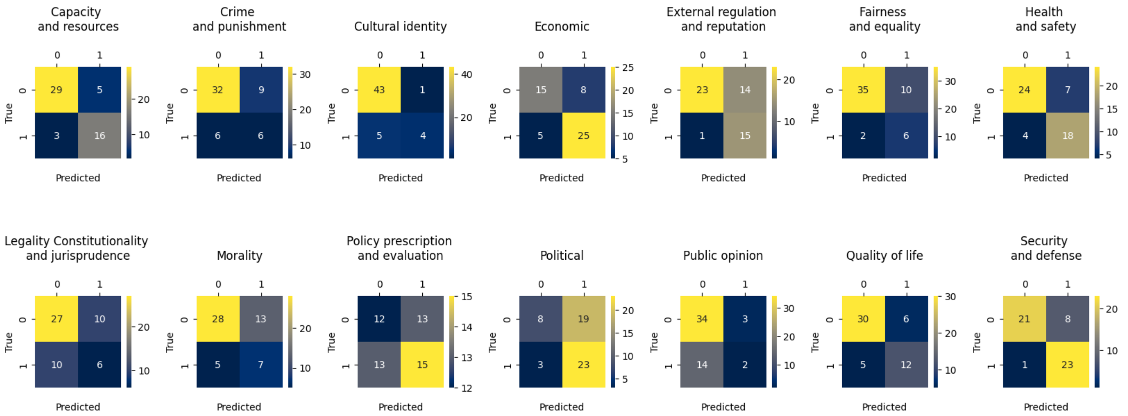
Table 6.2: Framing labels classification task: micro F1 score in test set across six languages

The following graph shows the relative distributions of the labels in the test set and in the predictions returned by the five models considered, all grouped by language subsets. The models trained solely on the original English language dataset were not able to detect any framing regarding cultural identity or public opinion while predictions for political framing and external regulation and reputation were more frequent. The addition of the articles obtained by translation seems to make the distributions more similar to those expected but only the analysis of the confusion matrices will be able to quantify the correctness of the predictions. For the German dataset there is a rather consistent behaviour between the various models, although once again cultural identity framing is hardly picked up (this is something we can see in all language subgroups). The predictions of the `GPT` model on French articles are rather consistent with the original distributions, similarly for Italian and Polish. The framing of the Russian dataset was more difficult to detect, due to the fact that there is an overwhelming prevalence of certain labels at the expense of others that may not occur at all (policy prescription and evaluation).

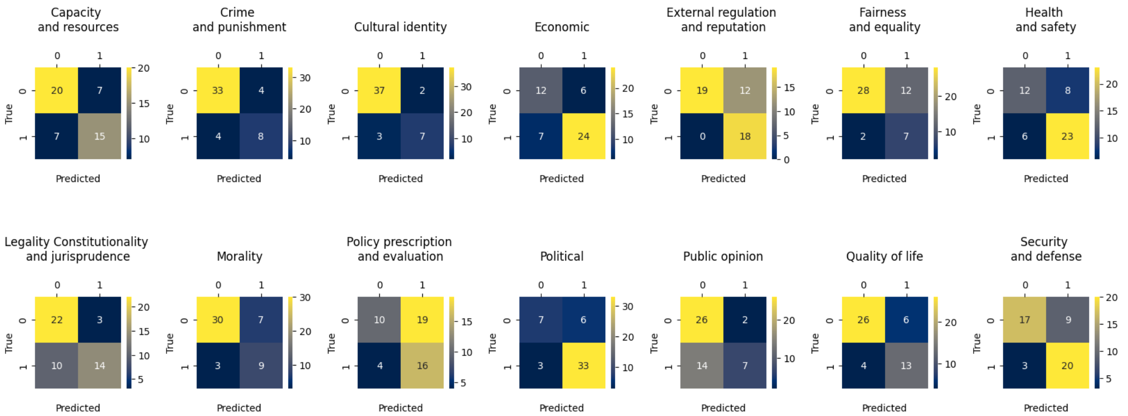
As this is a multi-label task, in order to better understand the spectrum of predictions, 14 confusion matrices were produced for each of the best performing models, one for each language. In the English language predictions, there does not seem to be one particular framing that was detected with precision and accuracy. There are some labels with a high number of true positives but this seems to be conditioned by a tendency to detect them even where there are none (many false positives). An example is political framing or even external regulation and reputation. This observation can also be extended to the case of German-language articles where, however, we are able to detect framing that has generally been detected where there

Figure 6.3: Framing labels classification task: predictions distribution on test set across six languages

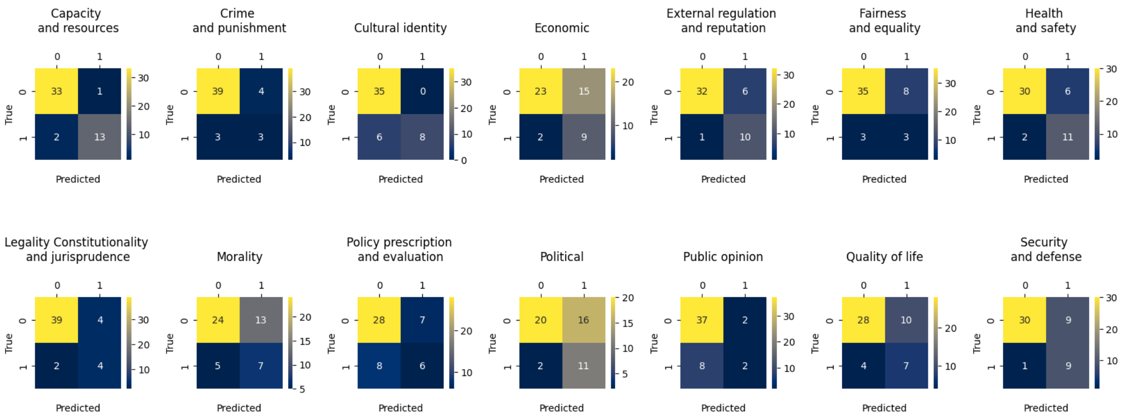
was (Economic). The presence of false positives is even higher on average in the case of the French-language articles, think of the framing Political, Morality and Economic. Good performances instead on Capacity and Resources. The same considerations apply to the Italian dataset, where the false positives often exceed the correct ones. All Polish articles with political labels were correctly identified and in general there is a tendency to detect framing in articles, noting, however, that sometimes this attribution is excessive, increasing the number of false positives. On the other hand, the base model struggles to detect framing in the Russian dataset, tending to be more conservative and less reckless in its attribution of a label.



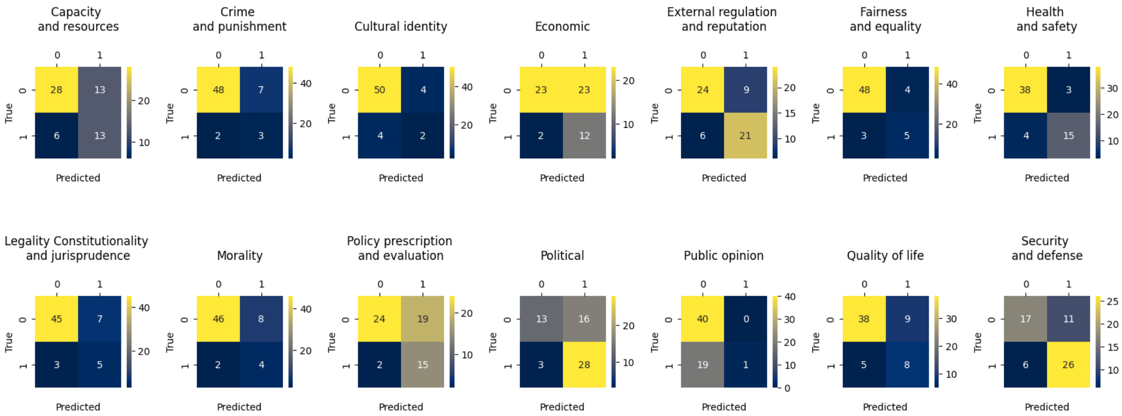
(a) Model with mBART augmentation for English test set



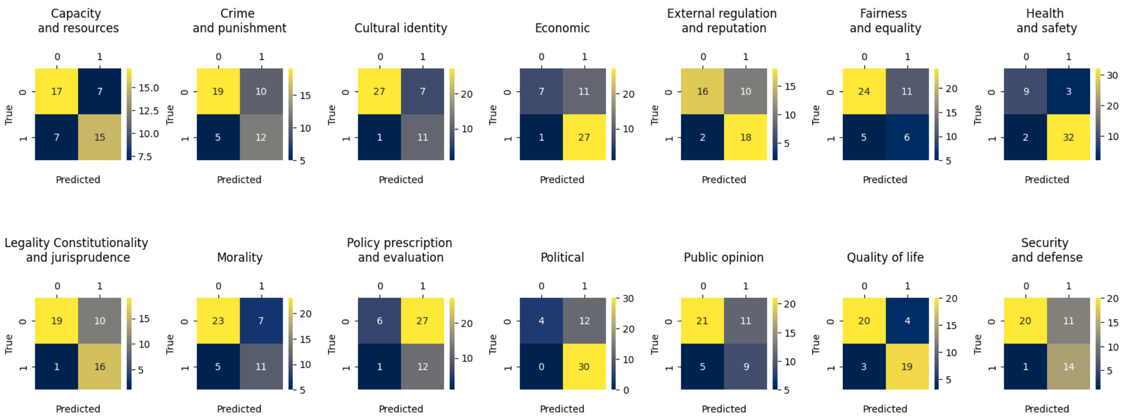
(b) Model with mBART and GPT augmentation for German test set



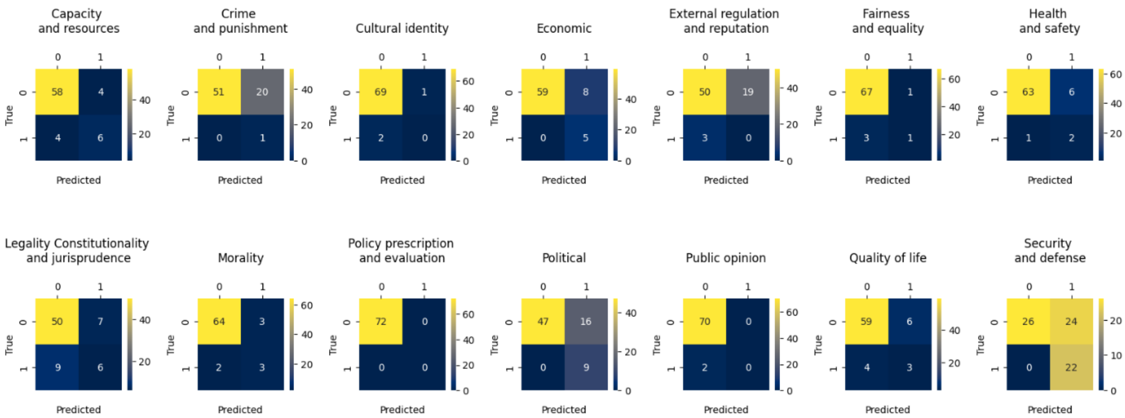
(c) Model with GPT augmentation for French test set



(d) Model with mBART augmentation for Italian test set



(e) Model with mBART augmentation for Polish test set



(f) Model with no text windowing nor data augmentation for Russian test set

Figure 6.4: Framing labels classification task: confusion matrices on test set for best model across six languages

6.3 Persuasion Technique Classification

The best checkpoints, selected on the basis of the F1-score micro performance on the validation set, were used to obtain the predictions on the test set by separating the six linguistic datasets.

Model	English	German	French	Italian	Polish	Russian
BASE	0.556	0.692	0.705	0.696	0.696	0.547
mBART	0.660	0.691	0.694	0.706	0.688	0.541
SPAN	0.159	0.314	0.283	0.269	0.190	0.188

Table 6.3: Persuasion Technique classification task: micro F1 score on test set across six languages

Text Level Classification We note that the two algorithms generally performed similarly, with slightly more success for the *base* model. We can therefore conclude that the data

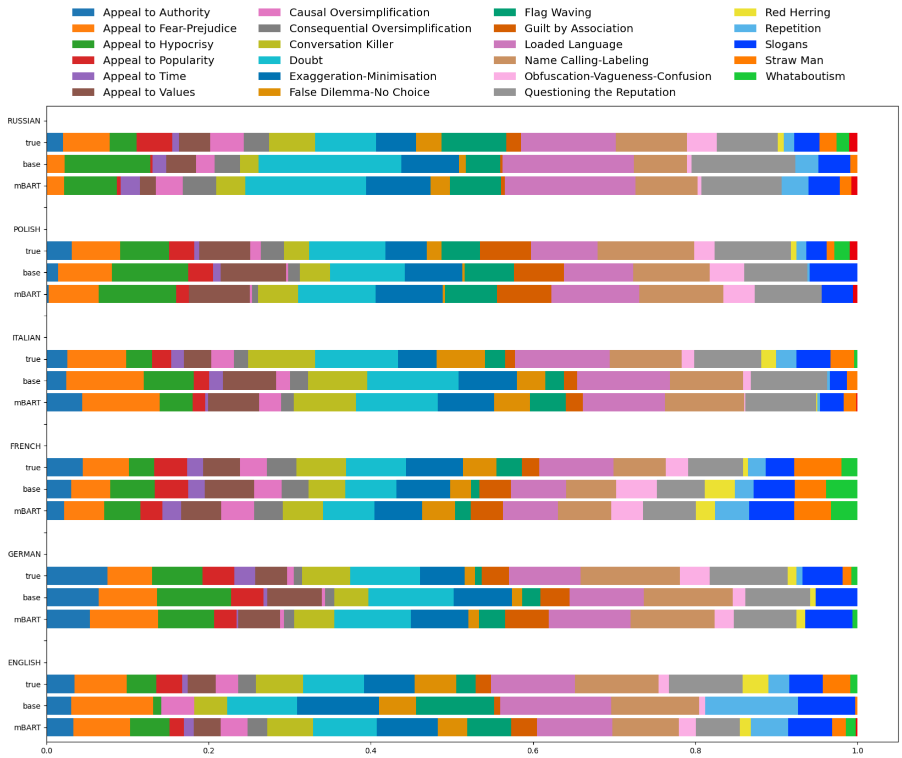


Figure 6.5: Framing labels classification task at text level: predictions distribution on test set across six languages

augmentation for this task did not produce any significant results that could motivate the translation procedure, which involves non-negligible time. The *mBART* model outperformed the simplest model in the case of English and Italian language data. With regard to the higher granularity version, we can observe that the model returned better results on the German and French datasets. Comparing the distributions of labels of the test dataset with those of the predictions, we note that: for the Russian articles there is an overestimation of the labels ”Appeal to Authority”, ”Doubt”, ”Questioning the reputation” which is on the other hand accompanied by a low prediction of techniques such as ”Appeal to Fear-Prejudice”. In other languages too, we can observe similar behaviour: think of ”Crime and Punishment” vs. ”Appeal to Hypocrisy” for English or ”Appeal to Values” for German.

Token Level Classification Since this is a single token detection, we are able to see where exactly the persuasion technique is located and can therefore assess any repetition in the text. From the Figure 6.6, it is possible to ascertain there are numerous overestimates and un-

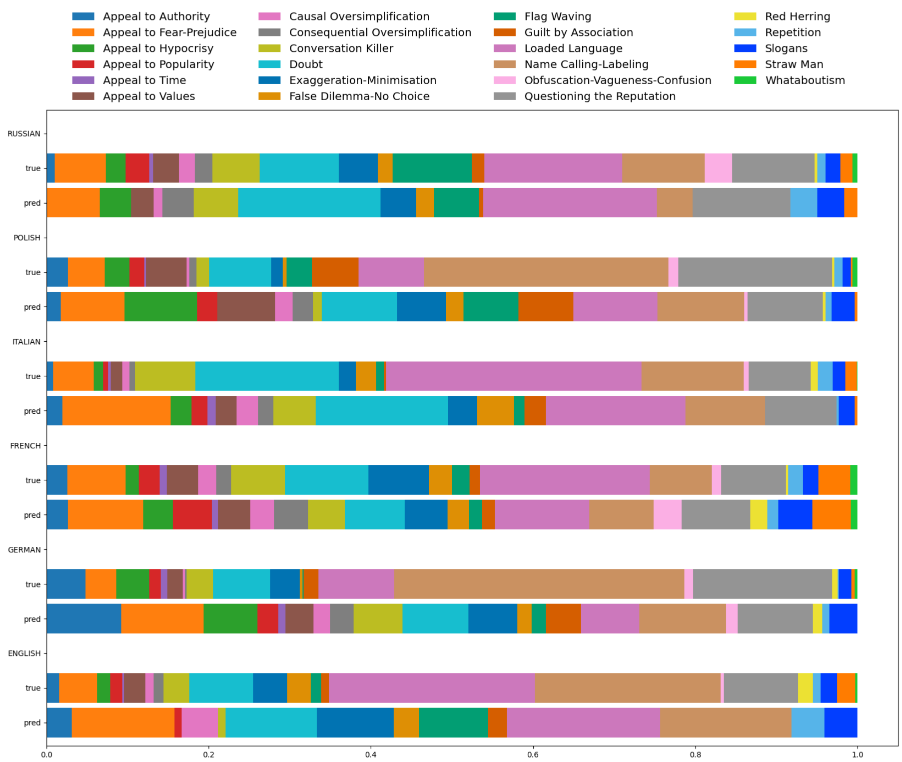


Figure 6.6: Framing labels classification task at span level: predictions distribution on test set across six languages

derestimates in all the language datasets. For example, in the original Polish articles, there is a substantial presence of ’Name Calling-Labeling’ technique that is, however, not revealed by the predictions. A similar behaviour is exhibited by the Italian dataset with the ’Loaded Language’ persuasion technique. On the other hand, there is a higher prediction of the Appeal to Authority in the English language articles than is actually present. This graph is not informative about the overlap of the phrases but only shows us the distributions so it is good to consider the number of tokens classified correctly by consulting the Table 6.4.

English	German	French	Italian	Polish	Russian
12096	35707	26246	26398	33808	15173

Table 6.4: Persuasion Technique labels classification at Span Level: Number of Positive Tokens across six languages

7 Conclusion

This work has demonstrated the potential of artificial intelligence in the management of NLP tasks, on different levels of specialisation and granularity, and in the creation of tools with interesting application value. In a world where information travels rapidly and extensively, and given the existence of techniques to convey propaganda messages, providing tools to assist the user in understanding journalistic material is a priority for a healthy communication system. Considering the international aspect in a multilingual set-up then allows not only to optimise and enhance these tools, but also to grasp important differences and similarities in communication approaches between the various nations of the world.

For these purposes, therefore, four multi-class classification models were implemented, with the aim of categorising multilingual journalistic articles between opinion texts, reporting and satire, five models to detect the perspective framing with which a given topic is proposed by the author and finally four models, one of which with high granularity, capable of identifying persuasion techniques among 23 possible ones. It was found that the use of techniques such as text windowing, capable of capturing all the information in the dataset and preparing it for appropriate implementation, can be a valuable aid in improving classification performance. At the same time, resorting to data augmentation techniques such as machine translation or text generation via interactive artificial intelligence algorithms has generally brought benefits in performance. However, it was also possible to identify limitations of these tools, which is a one more reason for further study and development. Finally, it can be concluded that the XLM-RoBERTa model lived up to expectations: endowed with excellent flexibility, it was used to solve four different NLP tasks and generally proved to provide excellent results for all the languages considered.

References

[1] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. [Online]. Available: https://aclanthology.org/D14-1179

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *CoRR*, vol. abs/1911.02116, 2019. [Online]. Available: http://arxiv.org/abs/1911.02116

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[6] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.47

[7] G. D. S. Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, and P. Nakov, "Fine-grained analysis of propaganda in news articles," 2019.

[8] J. Piskorski, N. Stefanovitch, N. Nikolaidis, G. Da San Martino, and P. Nakov, "Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 3001–3022. [Online]. Available: https://aclanthology.org/2023.acl-long.169

Acknowledgments

First of all, I would like to thank my advisor, Prof. Giovanni Da San Martino, who accompanied me through the process of writing the thesis. My thanks to the researchers at the Joint Research Centre who guided me in understanding the data and the correct tools. Finally, I want to thank my colleagues and all the people who have supported and sustained me during these years of study.