Embedding model for the Malagasy informal language

Francis Rakotomalala
Laboratory for Mathematical and Computer Applied to the Development Systems (LIMAD)
University of Fianarantsoa
Fianarantsoa, Madagascar
francis_rakotomalala@ymail.com

Aimé Richard Hajalalaina
Laboratory for Mathematical and Computer Applied to the Development Systems (LIMAD)
University of Fianarantsoa
Fianarantsoa, Madagascar
arhajalalaina@yahoo.fr

Ndaohialy Manda Vy Ravonimanantsoa
Engineering and Innovation Sciences and Techniques (STII)
University of Antananarivo
Antananarivo, Madagascar
ndaohialy@gmail.com

***Abstract*—Processing informal Malagasy language presents major challenges due to linguistic variations, abbreviations, and frequent code-switching in digital communication. This study proposes a text embedding model based on DistilBERT and XML-RoBERTa, specifically adapted to informal Malagasy. Through fine-tuning on custom corpora, we observe a gradual improvement in performance, with a significant reduction in loss function and lower perplexity, indicating a better understanding of linguistic structures. The evaluation shows that the generated embeddings effectively capture semantic similarities, even across varied formulations. DistilBERT outperforms XML-RoBERTa, demonstrating better generalization. These results highlight the importance of adapting language processing models to low-resource languages and open up new perspectives for applications in the automatic understanding of informal language.**

***Keywords*—BERT, Embedding model, Malagasy language, Informal language**

I. INTRODUCTION

With the rise of communication technologies, chat-based exchanges have become a preferred mode of interaction, both in personal and professional contexts [1]. These conversations are often characterized by informal language featuring abbreviations, emoticons, and linguistic variations [2]. While these features facilitate human communication, they present numerous challenges for NLP models, particularly for low-resource languages such as Malagasy [3].

Malagasy, primarily spoken in Madagascar, is a linguistically rich language with significant dialectal diversity, further complicating its automated processing. The lack of sufficient linguistic resources and annotated corpora exacerbates the difficulties in understanding informal Malagasy language. In this context, developing appropriate methods to effectively handle these specificities becomes essential to ensure the relevance of language understanding and text generation systems [4].

This article addresses this challenge by proposing a text embedding model tailored to informal Malagasy. Building on the architectures of pretrained models such as DistilBERT and XLM-RoBERTa [5], [6], our approach aims to capture the linguistic nuances found in informal conversations. Our contributions focus on:

- Understanding informal language by developing techniques capable of handling abbreviations, emoticons, and common language variations in chat conversations [7].
- Adaptation to low-resource languages through methods that take into account the specific features of Malagasy to improve the quality of text representations and the relevance of generated responses [8].

Through this research, we aim to lay the groundwork for better understanding of informal exchanges in Malagasy, while proposing solutions that can be generalized to other languages with similar characteristics [9].

II. RELATED WORK

Word Embedding is an NLP technique used for language modeling and feature learning. It can be unsupervised, supervised, or self-supervised, depending on the specific application. Before creating word embeddings, the text must first be segmented into individual words. Each word is then assigned an index value in a predefined vocabulary. Once segmented, we proceed to the embedding process, where words are converted into dense, continuous vectors of real numbers.

In the word embedding space, words with similar meanings are positioned closer to each other. For instance, the words *"king"* and *"queen"* would be placed nearby in the vector space, reflecting their related meanings. This ability helps AI models capture the semantic meanings of words based on their context and relationships across a large text corpus. Embeddings are often learned from large amounts of unlabeled data using unsupervised methods.

Cite as: F. Rakotomalala, A. R. Hajalalaina, and N. M. V. Ravonimanantsoa, "Embedding model for the Malagasy informal language", *Syst. Theor. Control Comput. J.* , vol. 5, no. 1, pp. 1–13, Dec. 2025.
DOI: 10.52846/stccj.2025.5.1.64

Historically, word embedding models were primarily based on techniques such as Word2Vec [10] and GloVe [11], which generated dense vectors for each word. However, these models had major limitations, particularly their inability to capture the broader context in which a word appears. For example, the word *"right"* in the phrases *"You have the right to remain silent"* and *"Turn right at the next intersection"* can have very different meanings depending on context, which these models failed to capture. Masked Language Models (MLMs), such as BERT [12], addressed this issue by adopting a bidirectional approach, where the model predicts a masked word using the context from both its left and right. This strategy allows for more nuanced and contextual word representations, which is especially crucial for processing complex and informal languages like Malagasy. Multilingual models such as multilingual DistilBERT and XLM-RoBERTa are modern extensions of earlier models and have greatly facilitated the processing of various languages, including so-called low-resource ones. DistilBERT, a lighter version of BERT, retains much of the original model's capabilities while being faster and less resource-intensive [5]. Its smaller size makes it suitable for languages that lack large labeled corpora, such as Malagasy. XLM-RoBERTa [6], in contrast, is pretrained on a large multilingual corpus covering 100 languages, including some African and Asian ones. This model is particularly suited to Malagasy, which suffers from a lack of annotated data necessary for training effective NLP systems. XLM-RoBERTa stands out for its ability to provide robust representations of multilingual texts, including informal language varieties.

Informal Malagasy presents particular challenges for NLP systems. It is heavily influenced by French, leading to a wide variety of hybrid words and expressions. Additionally, informal Malagasy writing is often inconsistent, with frequent use of slang, abbreviations, and orthographic variations. This makes building an appropriate NLP model especially complex. Models such as DistilBERT and XLM-RoBERTa prove valuable in this context, as their multilingual architecture enables them to capture the diversity and richness of linguistic constructions in informal texts. The use of multilingual models for languages like Malagasy holds great promise, but several challenges remain. Malagasy continues to be underrepresented in the training corpora of NLP models. While progress has been made in developing tools for low-resource languages, much work remains to adapt these multilingual models to the cultural and linguistic specificities of Malagasy. Another major challenge lies in data quality: despite XLM-RoBERTa being trained on a broad range of languages, the availability of high-quality Malagasy data remains a significant obstacle to achieving optimal performance in practical NLP applications.

III. MODEL DESCRIPTION

As part of textual embedding, we focus on natural language models that leverage the BERT architecture. BERT uses a deep learning component called transformers [13], which processes all the words in a sentence in parallel. This speeds up training and allows BERT to handle large datasets more efficiently. BERT is released under an open-source license, which allows to download and freely use it for various natural language processing tasks. These models enable the extraction of high-quality linguistic features from text or can be fine-tuned for specific NLP tasks to achieve the desired predictions.

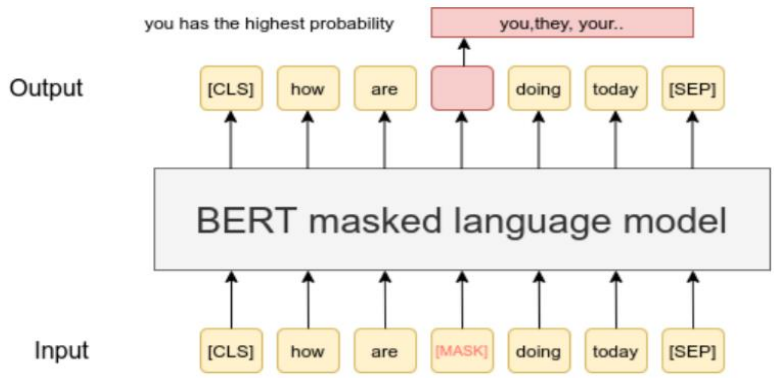

Fig. 1. Principe BERT MLM[1]

[1] https://www.sbert.net/examples/unsupervised_learning/MLM/README.html

A. BERT Model Formulation

1) Input representation

For an input sequence $X = (x_1, x_2, \dots, x_n)$, the initial embedding of token $x_i$is defined as

$$h_i^{(0)} = E_{\text{token}}(x_i) + E_{\text{position}}(i) + E_{\text{segment}}(s_i) \quad (1)$$

where E_{token}, E_{position}, and E_{segment}denote token, positional, and segment embeddings, respectively.

2) Self-attention mechanism

At layer l, the query, key, and value matrices are computed as

$$Q = H^{(l-1)}W_Q, K = H^{(l-1)}W_K, V = H^{(l-1)}W_V \quad (2)$$

The scaled dot-product attention is defined as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (3)$$

$$y = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (4)$$

3) Multi-head attention

$$\text{MultiHead}(H) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W_O \quad (5)$$

where each attention head is computed independently.

4) Layer representation

The contextual representation at layer $l$is obtained by

$$H^{(l)} = \text{LayerNorm}\left(H^{(l-1)} + \text{MultiHead}(H^{(l-1)})\right) \quad (6)$$

followed by a feed-forward network

$$H^{(l)} = \text{LayerNorm}\left(H^{(l)} + \text{FFN}(H^{(l)})\right) \quad (7)$$

5) Masked Language Modeling objective

For masked token positions M, BERT is trained to maximize

$$\mathcal{L}_{\text{MLM}} = -\sum_{i \in M} \log P(x_i \mid x_{\setminus M}) \quad (8)$$

$$\mathcal{L}_{\text{MLM}} = -\sum\nolimits_{i \in M} \log P(x_i \mid x_{\setminus M}) \quad (9)$$

We selected two models for this experiment: Multilingual DistilBERT and XLM-RoBERTa, both built on the BERT architecture and designed to handle multilingual contexts. Given that Malagasy is underrepresented in data corpora and NLP research, these models offer an interesting solution for our case. Their ability to generalize across different languages even those with limited resources makes them suitable for adaptation to languages like Malagasy.

B. Multilingual DistilBERT

1) Definition

In 2019, Hugging Face presented DistilBERT as a smaller, faster alternative to BERT, a powerful natural language processing model developed by Google. DistilBERT uses a distillation technique to train a smaller model to reproduce BERT's behavior, resulting in a model with fewer parameters while maintaining high accuracy and performance [5], [14], [15]. DistilBERT still uses the transformer architecture and is pre-trained using self-supervised learning on a large corpus of textual data.
Many researchers and practitioners in the NLP community have adopted DistilBERT because of its smaller size and faster training and inference times, without sacrificing much accuracy or performance.

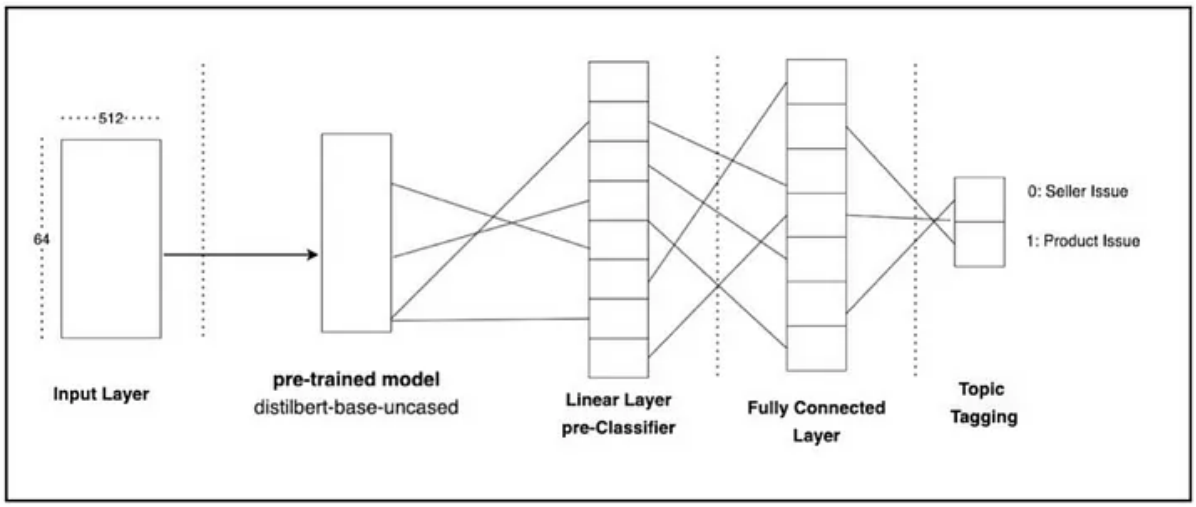

Fig. 2. The DistilBERT model is pre-trained on a large corpus of Amazon negative product review data and can be refined for various natural language processing tasks.[2]

The DistilBERT model is composed of several layers:

- Input layer: DistilBERT takes input IDs and attention masks. These inputs are encoded representations of the input text obtained using a tokenizer.
- Pre-trained layer: Pre-trained weights are used to initialize the DistilBERT model. It processes the input tensors to produce the hidden state output. The first token of the sequence is used to obtain a pooled representation of the input sequence from the hidden_state tensor.
- Linear pre-classifier: The linear pre-classifier is responsible for extracting features from the data by taking the output of the pre-trained layer and passing it to a fully connected layer.
- Fully connected layer: The fully connected layer is a linear layer that takes the output from the linear pre-classifier and maps it to the desired output dimensionality. In this case, the output dimension is two, corresponding to the topic tagging task.

During the forward pass, the input IDs and mask are passed through the pre-trained layer, generating a masked state sequence. Finally, the output from the dropout layer is passed through the linear output layer to produce the final output of the model a vector of size two representing the probabilities of the two classes in the topic tagging task.

2) Structural features

The DistilBERT Multilingual model is a lighter and faster version of BERT, designed to deliver similar performance with fewer parameters. Here is its detailed structure:

- Number of layers: 6 transformer layers.
- Number of attention heads: It contains 12 attention heads per layer, identical to the standard BERT version.
- Embedding size: The embedding vectors are 768-dimensional, the same as in BERT.
- Number of parameters: DistilBERT Multilingual contains approximately 66 million parameters.
- Vocabulary size: 119,547 tokens.
- Maximum input sequence: It can process sequences up to 512 tokens in length.

C. XLM-RoBERTa

1) Definition

XLM and XLM-RoBERTa are two multilingual language models developed by Facebook AI. They are designed to understand and generate text in multiple languages, addressing the challenges of multilingual comprehension.
XLM is based on the BERT architecture and learns to encode multilingual information by training on parallel data, consisting of sentence-aligned phrases in different languages. XLM-RoBERTa is an extension of XLM that uses the RoBERTa architecture, a variant of the Transformer model, for pre-training. It is based on the RoBERTa architecture, an optimized version of BERT. XLM-RoBERTa builds upon the strengths of XLM while incorporating advancements from RoBERTa to achieve even better performance in multilingual language understanding and downstream NLP tasks.
The key idea behind XLM-RoBERTa is to leverage RoBERTa's pre-training methodology, which involves large-scale pre-training and extensive hyperparameter tuning, to create a more powerful and efficient multilingual language model.

[2] https://medium.com/@kumari01priyanka/bert-and-distilbert-model-for-nlp-7352eb16915e

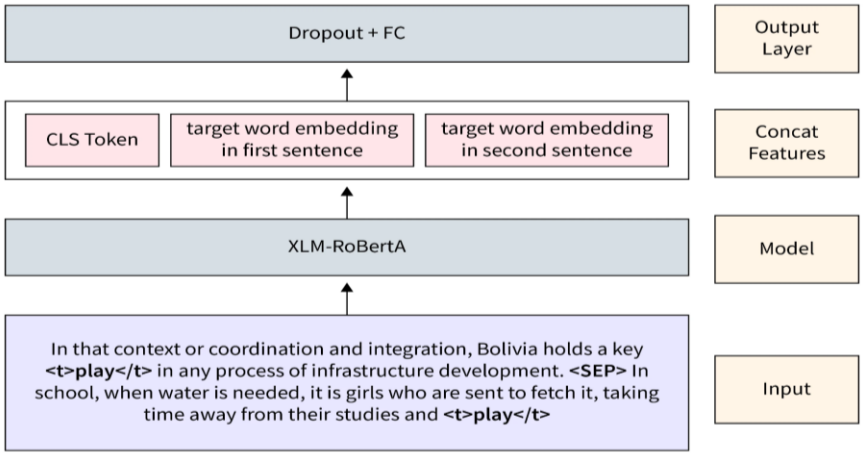

Fig. 3. Principe du modèle XLM-RoBERTa.[3]

The XLM-RoBERTa architecture is similar to other transformer-based models like BERT. It consists of embedding layers, transformer encoders, and a downstream structure. Here's a description of the architecture :

- Embedding Layers:

Like other transformer models, XLM-RoBERTa starts with embedding layers. These layers map input tokens to continuous vector representations, known as word embeddings. XLM-RoBERTa uses Byte Pair Encoding (BPE) to handle subword units, which allows it to manage out-of-vocabulary words and capture morphological information.

- Transformer Encoders:

The core of the XLM-RoBERTa architecture is the transformer encoder. It consists of several layers of self-attention mechanisms and feed-forward neural networks. Each encoder layer processes the input sequence in parallel, enabling the model to capture both local and global dependencies.

The self-attention mechanism allows the model to focus on different parts of the input sequence while encoding contextual information. It calculates attention weights for each input token, allowing the model to concentrate on relevant information during encoding.

The feed-forward neural networks within each transformer layer help capture complex relationships and non-linearities in the input sequence.

- Downstream Structure:

The output from the transformer encoder passes through a downstream structure, which can vary depending on the specific task for which the model is trained. This structure typically consists of additional layers (e.g., fully connected layers) that transform the encoded representations into task-specific outputs.

Overall, the XLM-RoBERTa architecture is designed to learn powerful sentence representations that can capture both semantic and syntactic information across different languages. By training on a large amount of multilingual data, XLM-RoBERTa can leverage shared information between languages to improve performance on various multilingual tasks.

2) Structural features

Here are its main structural characteristics:

- Number of Layers: XLM-R has 12 layers (for the base version), also called transformer blocks. The large version has 24 layers.
- Number of Attention Heads: Each XLM-R layer uses 12 attention heads in the base version and 16 heads in the large version.
- Embedding Size: The size of the embedding vectors is 768 dimensions for the base version and 1024 dimensions for the large version.
- Number of Parameters: The base version (XLM-R base) has approximately 270 million parameters. The large version (XLM-R large) has about 550 million.
- Vocabulary Size : 250,002 tokens.
- Maximum Input Sequence Length: The maximum input sequence length is 512 tokens.

[3] https://www.scaler.com/topics/nlp/xlm-roberta/

IV. DATASET USED

A. Description

1) FLICs [16]:

This dataset for informal Malagasy is a corpus designed to capture the nuances of informal Malagasy language use. The corpus mainly focuses on everyday conversations, social media exchanges, and informal texts. The goal of this dataset is to provide a broad range of texts reflecting the informal use of Malagasy, including features characteristic of informal communication, such as abbreviations, dialects, slang, emoticons, and keywords. Furthermore, it highlights the lesser-studied Malagasy language, used in an extremely informal manner and mixed with French through code-switching.

- Document Type : Informal texts (posts, comments, social media conversations).
- Language: Malagasy (informal variant).
- Tasks: Pre-training language models, analysis of informal language.
- Use: Training models to better understand the variations of the Malagasy language in informal contexts.

2) CC100-Malagasy Dataset

The CC100-Malagasy Dataset is part of the CC100 project, which has collected extensive multilingual corpora for a set of underrepresented languages in the context of pre-training multilingual models. The Malagasy subset of CC100 contains web data from various sources, primarily in the formal domain such as news articles, blogs, and other types of published texts in Malagasy.

- Document Type : Formal web texts (blogs, articles, websites).
- Langage : Formal Malagasy.
- Tasks: Linguistic modeling, text classification.
- Use: Training language models for understanding Malagasy in more formal contexts.

The combined use of the FLICs and CC100-Malagasy datasets in our work allows us to create a robust language model, capable of understanding and generating text in both formal and informal contexts. This approach provides comprehensive coverage of the variations of the Malagasy language, which is essential for addressing the needs of various tasks, ranging from informal language analysis to more traditional linguistic modeling in formal contexts.

B. Corpus preparation

1) Input formatting

Since our models are pre-trained models that expect input data in a specific format, we will need:

Special tokens used for specific tasks such as the beginning and end of sequences, word masking, or separating text segments.

The table 1 shows the main special tokens with their respective annotations.

TABLE 1. MAIN SPECIAL TOKENS OF DISTILBERT AND XML-ROBERTA.

Label	**DistilBERT**	**XLM-RoBERTa**
Start of a text sequence	[CLS]	<s>
End of sequence	[SEP]	</s>
Masking words	[MASK]	<mask>
Sequence filling to maintain uniform batch length	[PAD]	<pad>

2) Tokenization and addition of New Vocabularies

These models each provide their own tokenizer. However, the WordPiece tokenization algorithm used by our models is not always well-suited for underrepresented languages like Malagasy. To improve the model's understanding, new vocabularies have been added based on the particularities of the language:

- New words: Specific Malagasy words, as well as those from the Malagasy-French mix, have been directly integrated into the vocabulary. This helps the model better capture the nuances of the language and avoid excessive fragmentation of these terms during tokenization.
- Adapted subwords: The subwords used by the model have been adjusted to reflect the grammatical structures of Malagasy and to prevent frequently used words from being overly segmented.

Once the texts are tokenized using the tokenizer, each sentence in the corpus is converted into a series of IDs corresponding to the tokens in the model's vocabulary. This process transforms each word, or word fragment, into a

numerical representation that is understandable by the model.

3) Concatenating tokenized sequences

After tokenization, all sequences are concatenated into a single long sequence of IDs. This step allows the entire corpus to be treated as a single entity, maximizing continuity and context retention across the entire data set, even if they come from different sentences.

By concatenating sequences, we enable the model to retain contextual continuity even when sentences are short or fragmented.

4) Sequence truncation and segmentation

Once the tokens have been concatenated, we apply truncation to divide this long sequence into subsequences of fixed length, in our case 32. This standardizes the length of the sequences, ensuring that all sequences processed by the model have a consistent size.

This step ensures that all sequences have a uniform size of 32 tokens, facilitating batch processing during training.

What's more, a fixed-size sequence format enables more efficient use of GPU memory and simplifies training by better parallelizing computations.

5) Dynamic token masking in input sequences

In our experiment, the use of a data collator plays a central role in dynamically masking tokens during model training, particularly for tasks based on Masked Language Modeling (MLM).

a) Data Collator function

A data collator is a function that takes a batch of raw training data and prepares it to be sent to the model. When it comes to MLM, it's responsible for applying dynamic masking to a fraction of the tokens in each sequence. This allows different tokens to be masked each time a batch is passed to the model, making learning more robust and diverse.

b) Dynamic token masking

The data collator performs masking on the fly during training, unlike static masking where the positions of masked tokens are predefined in the dataset. Dynamic masking offers several advantages:

- Input variability: With each iteration, the model sees different masked versions of the input sequences, which improves generalization.
- Contextual adaptation: The model learns to predict missing tokens by taking into account the full context, as masked tokens constantly change from one iteration to the next.

c) Masking configuration

In our case, around 15% of the tokens in each input sequence are dynamically masked.

For example, we use a DataCollatorForLanguageModeling data collator, which generates masked sequences for each training batch. This approach is particularly effective for training the model on sparsely represented languages such as Malagasy, as it offers a wide range of contexts in which the model can observe missing tokens.

Here's an overview of the input and output after these different processes on the corpus:

“maro ny manohana sy mahatsapa hoe tena mila atosika ny tanora [SEP] [CLS] mpizaha tany antsasakà tapitrisa”

>>> maro ny manohana [MASK] mahatsapa hoe tena [MASK] atosika ny tanora [SEP] [CLS] mpizaha tany antsasakà [MASK]”

6) Final Dataset structure

- For training:

The training dataset contains 1,894,910 examples, each consisting of the following elements:

  - input_ids: The token IDs generated by the tokenizer for each text sequence.
  - attention_mask: An attention mask indicating which tokens the model should focus on (1) or ignore (0).
  - word_ids: Indicates the IDs of the original words before tokenization.
  - labels: The output labels for each token, used in the context of Masked Language Modeling (MLM).

- For validation:

The test dataset includes 474,230 examples, organized in the same way, and is used to evaluate the model's performance on unseen data.

V. RESULTS

A. DistilBERT fine-tuned

1) Loss function

a) Training data

The initial loss observed was between 4 and 5, indicating that the model started with a very limited understanding of

the task. This initially high level of loss shows that the model had a lot to learn, not least because of the complexity and specificities of the Malagasy language.

Over the course of training, with a total of 50 epochs, the loss gradually decreased to 1.48 on the training data. This steady decrease testifies to the model's ability to better capture semantic and syntactic relations in the language.

b) Validation data

On the validation data, the loss was measured at 1.69, slightly higher than on the training data, revealing that the model generalizes well on new data. This result indicates a good balance, suggesting that the model is not overlearning and is maintaining its performance on unseen examples.

The figure 4 shows the graph containing the loss curves of the DistilBERT model on training and test data.

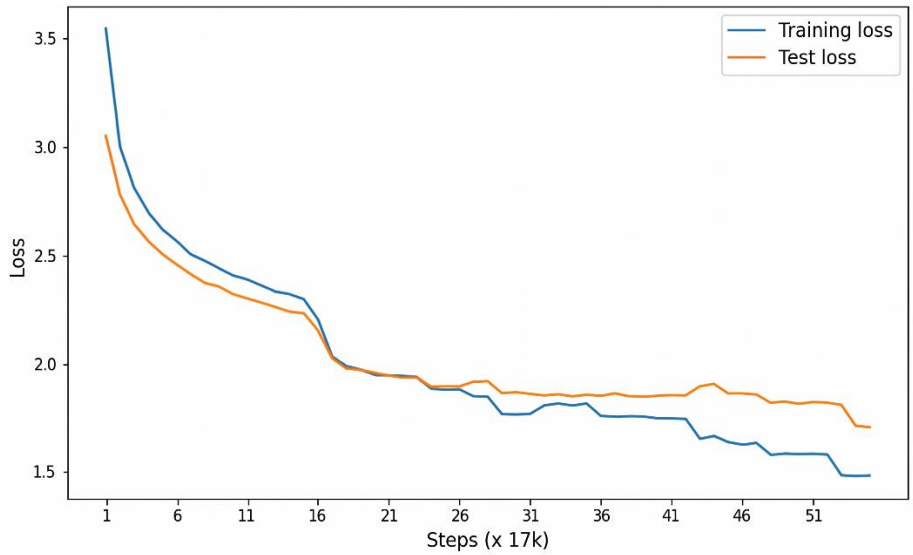

Fig. 4. Evolution of the loss function by 17k steps (DistilBERT).

2) Accuracy

The model achieved 71% accuracy on the test data. This result indicates that, in 7 cases out of 10, the model correctly classifies the tokens or sequences tested according to the assigned task. It is important to note that this metric evaluates the accuracy of overall predictions, but does not give detailed information on the distribution between classes or on finer aspects such as the model's ability to recognize complex grammatical structures.

3) Perplexity

Perplexity, a key metric for evaluating language models, showed significant improvement throughout training. Initially, perplexity was above 49, indicating that the model had difficulty predicting the correct chip sequences and showed high uncertainty in the face of the data.

After training, perplexity dropped to 5.41, an excellent indicator of the model's improved ability to understand and predict the syntactic and semantic structures of the Malagasy language. Lower perplexity means that the model is more confident in its predictions and produces more natural and coherent sequences.

4) Text embedding

a) Embedding extraction

After training, the model is able to generate vector representations (embeddings) for each text. These embeddings encapsulate the semantic features of the Malagasy language and enable text sequences to be transformed into numerical vectors that can be easily compared or used for other downstream tasks.

Example:

Let's imagine we want to extract the embeddings for the sentence: *"Mankasitraka ee. Ireo no taranaka hoavin'ny firenena"*.

The DistilBERT fine-tuned model will transform this sentence into a vector of fixed size, a list of 768 dimensions, representing the meaning and relationships between the words in the text. This vector could then be used in tasks such as text classification or clustering.

Here is an overview of the average vector for each token:

[6.2725e-02, -7.5908e-02, 6.3000e-01, 1.7048e-01, **…**, 8.0600e-02, 2.1993e-01, -1.5617e-01, -3.6516e-02]

b) Text similarity

Using extracted embeddings, it is possible to compare texts on the basis of their semantic similarity. Cosine similarity was used to quantify this similarity, and the results suggest that the model is able to effectively distinguish similar from dissimilar texts. This demonstrates that the model captures not only lexical information, but also the underlying relationships between sentences.

Example:

Consider the following three sentences:

Sentence 1: *"Akor, tamvin tam taony fir reine Elizabet maty ?"*

Sentence 2: *"Manao ahoana, taona inona no nodimandry ny mpanjaka Elisabeth ?"*

Sentence 3: *"Ts azk mits oe inn n tna blem anty fiar ty f simb matetka"*

After extracting the embeddings, we obtain the following cosine similarity scores:

- Similarity between Sentence 1 and Sentence 2: 0.70

This score indicates a high semantic similarity, suggesting that the two sentences essentially express the same idea, even though they are phrased differently. It is worth noting that Sentence 1 uses a more informal register.

- Similarity between Sentence 1 and Sentence 3: 0.40

This relatively low score shows that there is little semantic connection between the two sentences, highlighting that Sentence 3 addresses a completely different topic. Although both sentences have a similar structure, this does not necessarily indicate contextual alignment.

- Similarity between Sentence 2 and Sentence 3: 0.20

This very low score indicates weak similarity, confirming that Sentence 3 does not share any semantic context with the other two sentences. However, it shows some structural similarity with Sentence 1, despite the lack of contextual connection.

B. XLM-RoBERTa fine-tuned

1) Loss function

a) Training data

The initial loss observed for the refined XLM-RoBERTa model was 3.82, reflecting a very limited understanding of the task at the outset. This relatively high level of loss illustrates the challenge of learning the specifics of the Malagasy language for the model. Over the course of 50 training epochs, the loss gradually decreased to 1.50, reflecting a better grasp of the semantic and syntactic relations in this language.

b) Validation data

On the validation data, the loss was measured at 1.71, indicating that the model generalizes correctly while performing slightly as well on this set. The associated figure shows the loss curves for training and test data for the XLM-RoBERTa model.

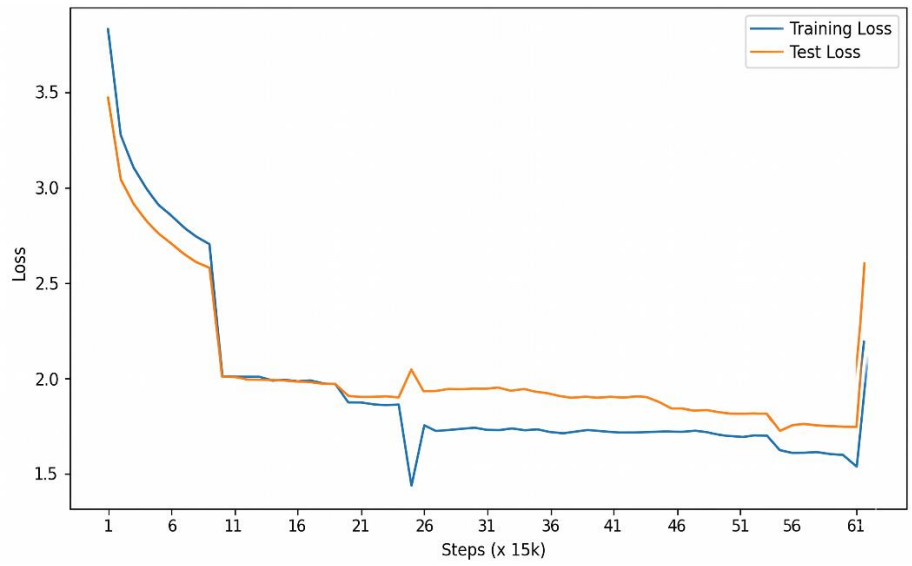

Fig. 5. Evolution of the loss function by 15k steps (XML-RoBERTa).

2) Accuracy

The model achieved 67% accuracy on the test data. This overall measure of predictive accuracy is positive.

3) Perplexity

Perplexity fell sharply throughout the training. It was initially above 40, indicating a high degree of uncertainty in the model's predictions. After training, perplexity dropped to 5.52, illustrating a significant advance in the model's ability to capture and predict the semantic and syntactic structures of the Malagasy language. This low level of perplexity reflects a significant improvement, revealing a good adaptation of the model to the particularities of the language.

4) Text embedding

a) Embedding extraction

Like DistilBERT, the XLM-RoBERTa fine-tuned model can generate vector representations for each text.

For example, if we wish to obtain the embeddings of the sentence *"F nga lasa pambotra gateau ndrai?"*, the model will convert this sentence into a fixed-size vector, composed of 768 dimensions, capturing the meaning and relationships between the words in the text.

Here's a look at the average embedding vector for each token:

[-1.1866e-01, 1.1575e-01, -1.8565e-01, -2.2054e-02, -1.7412e-01, **...**, 1.4938e-02, -6.9536e-02, 6.9673e-02, -9.2627e-02, 2.1428e-01]

b) Text similarity

Cosine similarity was also applied to quantify the similarity between sentences.

Example:

Consider the following sentences:

Sentence 1: *"lanjalanjao ny fijerena television , phon e , de miinana akondro."*

Sentence 2: *"ahena ny fijeren fahitalavitra dia misakaf tsr."*

Sentence 3: *"raha ireny adala mirenireny ireny no mba misy mi-kidnapper , mihena ny adala."*

The cosine similarity scores obtained are as follows:

- Similarity between Sentence 1 and Sentence 2: 0.91
- Similarity between Sentence 1 and Sentence 3: 0.65
- Similarity between Sentence 2 and Sentence 3: 0.51

These results show that the model has a high degree of confidence in the measured similarities, which may limit its ability to differentiate texts with no contextual connection, as evidenced by Sentence 3, which shares no common context with the other two despite moderate similarity scores. This is likely due to the model's complexity, which captures superficial similarities and may sometimes overestimate the connections between dissimilar sentences.

VI. DISCUSSIONS AND RECOMMENDATIONS

A. Global performance

The results obtained during the fine-tuning of the models show a continuous improvement in performance. The loss function, initially high, gradually decreased over the course of the 50 epochs, indicating that the model effectively learned to capture the semantic and syntactic relationships of the Malagasy language, with good generalization and no overfitting. Additionally, the models achieved promising accuracy values on the test data, which is encouraging, although further details on the class distribution are lacking. Perplexity, on the other hand, showed a marked improvement in the models' ability to predict more coherent and natural sequences. These results suggest that fine-tuning on the Malagasy language allowed the models to better capture linguistic specifics and informal registers, a notable improvement over the original model. However, improvements can still be made, particularly regarding the management of dialectal variations and the inclusion of more diverse data for broader language coverage.

B. Embedding quality

The use of extracted embeddings allows for the application of cosine similarity to compare texts. The results show that the model is not only effective at distinguishing similar texts from dissimilar ones, but it also captures lexical nuances and deeper semantic relationships between sentences. For example, the model is able to detect similarities despite different or informal formulations, demonstrating its ability to understand the subtleties of the Malagasy language, even in contexts where the underlying meaning remains the same. Furthermore, it successfully discerns semantic differences, even when grammatical structures or syntactic constructions are similar, proving its robustness in natural language analysis. However, improvements could still be made, particularly to refine its sensitivity to cultural contexts and dialectal variations, in order to further enhance the quality of semantic comparison.

C. Comparison of the two fine-tuned models

The table shows a comparison of performance obtained by training DistilBERT and XML-RoBERTa.

TABLE 2. COMPARATIVE TABLE OF THE PERFORMANCES OBTAINED

Label	**DistilBERT**	**XML-RoBERTa**
Training loss	1.48	1.50
Validation loss	1.69	1.71
Accuracy	71%	67%
Perplexity	5.41	5.52

DistilBERT proves to be slightly more effective than XLM-RoBERTa in this comparison. With a lower loss value, DistilBERT demonstrates better learning ability and slightly more efficient generalization than XLM-RoBERTa. In terms of accuracy, DistilBERT achieves 71% compared to 67% for XLM-RoBERTa, showing that it more often predicts the correct target classes. Finally, the slightly lower perplexity of DistilBERT confirms a higher confidence in its predictions. Overall, DistilBERT seems better suited for this study due to its accuracy and generalization ability.

D. Advantages

1) Enhanced Captured Features

The fine-tuned models demonstrated a significantly improved ability to capture features specific to the Malagasy language, particularly those related to the informal register. While the original models were limited to more formal or generic contexts, the fine-tuned models successfully incorporated more nuanced aspects, such as colloquial expressions and local variations in syntax. This is especially relevant in dialogues or informal exchanges, where formal language is not always used. This improvement is crucial for applications that require more natural interaction with native speakers.

2) Efficiency in Generalization

One of the major strengths obtained is the ability to generalize efficiently on unpublished data. The small deviations observed between loss on training and validation data suggest that the model learns robustly without over-adjusting to training data. This is all the more important in contexts where the language evolves or varies according to dialect. For practical applications such as translation, named entity recognition or recommendation systems, this ability to generalize well on new data means that the model can be used for a wide range of applications.

3) High Quality Vector Representation

Embedding extraction has also enhanced the quality of the vector representations of the texts. These embeddings encapsulate complex information, not only about the words but also about the relationships between them in their context, facilitating advanced tasks such as classification, clustering, and document similarity detection. The ability to capture deep semantic and syntactic information is a valuable asset for applications such as information retrieval, where the accuracy of the results heavily relies on the quality of the textual representations.

4) Flexibility and adaptability

The models have become more adaptable to the specific needs of the Malagasy language. This opens up various potential uses in specific cultural and linguistic contexts, where language characteristics may vary depending on the dialect or level of formality. This flexibility broadens their scope of application, making it possible to use them in sectors such as education, localized content creation, or multilingual virtual assistants.

5) Potential for Downstream Application Optimization

With these improvements, the models show great potential for other downstream NLP tasks. By better capturing the specifics of Malagasy, including variations in form and style, these models could contribute to better language understanding by machines, in sectors such as customer service or conversational interfaces, where linguistic precision is crucial.

E. Limits

1) Relative accuracy

Although the accuracy is promising, it still reveals a number of incorrect predictions, particularly in more complex or ambiguous cases. This suggests potential for improvement, especially in recognizing minority classes or less frequent sequences. The models may require further fine-tuning, particularly to better handle linguistic subgroups and syntax exceptions specific to the Malagasy language.

2) Complexity of the Malagasy language

The Malagasy language exhibits significant dialectal diversity and a widely used informal register, particularly in modern communication contexts. Although progress has been made, the models may still struggle to capture regional dialectal nuances or lexical variations that are not uniformly represented in the training data. This can lead to some imprecision in handling certain local or informal expressions.

3) Underfitting of certain linguistic aspects

Despite promising results, some more subtle aspects of the language, such as wordplay, double meanings, or non-

standard grammatical constructions, may not be fully understood by the model. This could limit its effectiveness in text comprehension tasks that require a more detailed analysis of intent or context.

4) Limited capacity for long sequences

As with many Transformer-based models, the ability to handle long sequences can be a challenge. The models may show a degradation in performance when faced with particularly long or complex texts, which could limit their usefulness in certain applications requiring extended context.

F. Recommendations

1) Training on diversified data

To improve the performance of the models, it is essential to train them on data covering a wider diversity of dialects and linguistic registers of the Malagasy language. Adding corpora from different regions and social contexts would enhance their generalization. Recent work on large language models suggests that data diversity is a key factor for performance in multilingual or dialectal environments. Furthermore, data augmentation techniques could also be considered to enrich the data.

2) Continuous improvement

A continuous fine-tuning strategy is recommended to strengthen the robustness of the model. This could include the use of new data collected over time, as well as feedback from the models' use in real-world applications. This iterative approach would help adapt the models to language evolution and changing linguistic habits, especially in dynamic contexts like social media or informal digital communications.

3) Multimodal approach

With the evolution of artificial intelligence techniques, the integration of multimodal approaches (text, audio, video) could help capture richer information, particularly for nuances in Malagasy pronunciation or specific cultural contexts. The development of systems combining text and voice could improve language understanding in interactive or conversational environments.

4) Exploiting other non-textual syntaxes

A particularly relevant aspect for better understanding informal language, especially in contexts like social media, is to leverage non-textual syntax forms such as emoticons, hashtags, emojis, and other visual communication elements. These symbols play a crucial role in expressing emotions, intentions, and context, especially in online interactions. By integrating these elements into the model's training process, it would be possible to capture finer nuances, thereby enriching semantic understanding and increasing the model's relevance in digital and social environments where language is constantly evolving.

VII. CONCLUSION

This study explored the adaptation of Transformer-based models for understanding informal Malagasy by fine-tuning DistilBERT and XLM-RoBERTa on specialized corpora. The evaluation of the models allowed us to measure their performance across several criteria, including loss function, accuracy, and perplexity, as well as their ability to produce high-quality embeddings for the semantic analysis of texts. The results analysis showed a significant improvement in performance after fine-tuning, particularly for DistilBERT, which demonstrated better accuracy and lower perplexity compared to XLM-RoBERTa. The effectiveness of the extracted embeddings was confirmed through text similarity tests, highlighting the models' ability to capture the linguistic and semantic nuances specific to informal Malagasy. However, several challenges remain, including the complexity of the Malagasy language, the variability of informal registers, and the models' limited ability to handle long sequences.

Regarding the advantages, the study highlighted the models' flexibility, their ability to generalize across diverse data, and the quality of the vector representations produced. These features open the door to various applications, particularly in machine translation, conversation analysis, and intelligent response generation. Nonetheless, some limitations need to be addressed, such as the relative accuracy of the models, the underfitting of certain linguistic aspects, and the need for training on even more diverse corpora.

Thus, we recommend enriching the training datasets with more diverse sources, exploring multimodal approaches incorporating other types of data, and optimizing architectures for better handling of informal syntax. Furthermore, we plan to conduct additional experiments to strengthen our results and ensure their robustness. These perspectives will further refine the modeling of the Malagasy language and improve the robustness of the models in various contexts.

REFERENCES

[1] D. Crystal, « Language and the Internet ». 2004.

[2] J. Eisenstein, « What to do about bad language on the internet », in *Proceedings of the 2013 conference of the North American Chapter of the association for computational linguistics: Human language technologies*, 2013, p. 359-369.

[3] P. Joshi, S. Santy, A. Budhiraja, K. Bali, et M. Choudhury, « The State and Fate of Linguistic Diversity and Inclusion in the NLP World », 27 janvier 2021, *arXiv*: arXiv:2004.09095. doi: 10.48550/arXiv.2004.09095.

[4] S. Bird, « Decolonising speech and language technology », in *28th International Conference on Computational Linguistics, COLING*

2020, Association for Computational Linguistics (ACL), 2020, p. 3504-3519.

[5] V. Sanh, L. Debut, J. Chaumond, et T. Wolf, « DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter », 1 mars 2020, *arXiv*: arXiv:1910.01108. doi: 10.48550/arXiv.1910.01108.

[6] A. Conneau *et al.*, « Unsupervised Cross-lingual Representation Learning at Scale », 8 avril 2020, *arXiv*: arXiv:1911.02116. doi: 10.48550/arXiv.1911.02116.

[7] T. Baldwin, P. Cook, M. Lui, A. MacKinlay, et L. Wang, « How noisy social media text, how diffrnt social media sources? », in *Proceedings of the sixth international joint conference on natural language processing*, 2013, p. 356-364.

[8] C. M. Keet, « Bootstrapping NLP tools across low-resourced African languages: an overview and prospects », 21 octobre 2022, *arXiv*: arXiv:2210.12027. doi: 10.48550/arXiv.2210.12027.

[9] N. Aepli, « There Is Plenty of Room at the Bottom: Challenges & Opportunities in Low-Resource Non-Standardized Language Varieties », PhD Thesis, University of Zurich, 2024. Consulté le: 7 avril 2025. [En ligne]. Disponible sur: https://www.zora.uzh.ch/id/eprint/262877/1/Aepli_Noemi_Dissertation.pdf

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, et J. Dean, « Distributed representations of words and phrases and their compositionality », *Advances in neural information processing systems*, vol. 26, 2013, Consulté le: 7 avril 2025.

[11] J. Pennington, R. Socher, et C. D. Manning, « Glove: Global vectors for word representation », in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, p. 1532-1543.

[12] J. Devlin, M.-W. Chang, K. Lee, et K. Toutanova, « BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding », 24 mai 2019, *arXiv*: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.

[13] A. Vaswani, « Attention is all you need », *Advances in Neural Information Processing Systems*, 2017.

[14] A. A. Mary, P. Acharya, R. Rakshinee, et S. Jeyaseelan, « Enhancing Question Answer Generation from PDFs: A Fusion of BERT, RAKE, T5 and DistilBERT with RQUGE », in *Proceedings of 5th International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2024, Volume 1*, Springer Nature, p. 249.

[15] A. F. Adoma, N.-M. Henry, et W. Chen, « Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition », in *2020 17th international computer conference on wavelet active media technology and information processing (ICCWAMTIP)*, IEEE, 2020, p. 117-121.

[16] F. Rakotomalala, A. R. Hajalalaina, M. V. Ravonimanantsoa Ndaohialy, A. Andriavelonera Alexandre, et A. H. Ranaivoson, « FLICs (Facebook Language Informal Corpus): a novel dataset for informal language », *Int J Data Sci Anal*, vol. 18, nº 4, p. 393-403, oct. 2024, doi: 10.1007/s41060-023-00460-2.