# PII Detection in Low-Resource Languages Using Explainable Deep Learning Techniques

Byamugisha Africano
byamugisha.africano@students.mak.ac.ug
Makerere University
Kampala, Uganda

Daudi Jjingo
daudi.jjingo@mak.ac.ug
Makerere University
Kampala, Uganda

Mpungu Gideon
mpungu.gideon@students.mak.ac.ug
Makerere University
Kampala, Uganda

Ggaliwango Marvin
ggaliwango.marvin@mak.ac.ug
Makerere University
Kampala, Uganda

## ABSTRACT

Safeguarding Personally Identifiable Information (PII) in an increasingly interconnected world presents intimidating challenges, particularly in low-resource languages like Luganda where computational resources for Natural Language Processing (NLP) are scarce. This research attempts to address these challenges, focusing on PII detection in Luganda, a low-resource language spoken in Uganda. The research leverages advanced deep learning methodologies, with attention mechanisms, to enhance PII detection efficacy amidst data scarcity. By directing models to key linguistic features and integrating Explainable AI (XAI) techniques, the study aims to improve both performance and transparency. Three distinct models were implemented i.e. luganda-ner-v6, DeBERTa-v3-Base, and afroxlmr-large-ner-masakhaner. Evaluation results demonstrate promising precision, recall, and F1 scores, while all models perform well, afroxlmr-large-ner- masakhaner consistently excels than the other models on all metrics. for instance he afroxlmr-large-ner-masakhaner model has the highest accuracy with 96.3%, followed closely by luganda-ner-v6 at 95.1%, and deberta-v3-base at 93.9%. The significance of this research extends beyond privacy protection, but also lies in contributing to the broader fields of NLP and privacy technology in low resource languages. This research suggest potential improvement areas including creating PII datasets for multilingual model training, transfer learning, explicit implementation of attention mechanisms, and domain-specific knowledge to advance PII detection and anonymisation in low resource languages.

## CCS CONCEPTS

• **Computing methodologies** → **Deep belief networks**; • **Security and privacy** → *Privacy protections.*



*IC3 2024, August 08–10, 2024, Noida, India*

ACM ISBN 979-8-4007-0972-2/24/08
https://doi.org/10.1145/3675888.3676036

## KEYWORDS

PII Detection, Low-Resource Languages, Deep Learning, Attention Mechanisms, Explainable AI, Privacy Protection

**ACM Reference Format:**
Byamugisha Africano, Daudi Jjingo, Mpungu Gideon, and Ggaliwango Marvin. 2024. PII Detection in Low-Resource Languages Using Explainable Deep Learning Techniques. In *2024 Sixteenth International Conference on Contemporary Computing (IC3-2024) (IC3 2024), August 08–10, 2024, Noida, India.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3675888.3676036

## 1 BACKGROUND AND INTRODUCTION

In today's era of information technology, the need to protect personally identifiable information has become a critical concern. PII breaches can have severe consequences, not only leading to privacy violations but also posing various security risks that can affect both individuals and organizations. Safeguarding PII is particularly challenging in a global context where data flows across borders and through multiple jurisdictions, processed by diverse systems. This dynamic environment requires robust detection and anonymization protocols to ensure the confidentiality and integrity of sensitive information. The task becomes even more difficult for low-resource languages like Luganda, which is predominantly spoken in Uganda. These languages often lack computational resources and specialized tools in the field of natural language m processing (NLP), making the task of developing effective privacy-preserving technologies for detecting PIIs and anonymizing them particularly challenging.

### 1.1 Problem Statement

This study aims to respond to the pressing demand for effective identification of personally identifiable information in languages with limited resources, focusing on Luganda as a specific example. The two main computational problems of focus are (a) the limited availability of annotated PII datasets in low-resource languages hinders the effectiveness of model training for PII detection and (b) the "black-box" nature of deep learning models often leads to a lack of transparency, making it difficult to interpret their decision-making processes limiting the trust and adoption of AI systems in privacy-sensitive applications.

Therefore, this research attempts to solve the above challenges by employing state-of-the-art deep learning techniques with attention mechanisms. These mechanisms have shown promise in capturing

the nuances of language, which is crucial for accurately identifying PII in texts as attention mechanism helps models focus on the important parts of input, hence achieving higher accuracy on small datasets and addressing the issue of data scarcity in PII detection.

Furthermore, to ensure the explainability and transparency of the models, Explainable AI techniques will be applied to provide insights into the decision-making process of the deep learning models. The goal is to develop a robust and interpretable PII detection system for low-resource languages like Luganda.

### 1.2 Computational Problem Description

Despite the growing concern for privacy and data security, particularly in low-resource languages, it has been evident that detecting Personally Identifiable Information (PII) in these languages is still not straight forward task as they present unique computational challenges. This study focused on the challanges of :-

a) Computational digital resources and tools are scarce, especially the annotated datasets and pretrained models which limits the development of more accurate PII identification and anonymisation models:

b) Furthermore, it was recognized that the importance of transparency in machine learning which has largely limited the adaption and trust in these models as their process of decision making is mostly hidden

### 1.3 Deep learning computational objectives

The main goal of this research is to explore and implement deep learning techniques for PII detection in the low-resource language of Luganda, with attention mechanisms and applying XAI. Specifically:

- To Implement various deep learning models with attention mechanisms to accurately identify PII in low resource languages texts like Luganda.
- To apply Explainable Artificial Intelligence methods to illustrate model decision-making processes in identifying PII in low resource languages texts like Luganda.

### 1.4 Research Contributions

The significance of this research lies in its potential to enhance privacy protections within digital ecosystems and to ensure compliance with international data protection regulations. Additionally,it contributes to explainable PII detection in the low-resource linguistic contexts by: (1) The implementation deep learning models for PII detection for low-resource languages, specifically focusing on Luganda, using state-of-the-art deep learning techniques with attention mechanisms, (2) The application of Explainable AI methods to provide transparency and interpretability in the decision-making processes of the deep learning models used for PII detection in low resource languages texts like Luganda, (3) Evaluatioon and discussion of the limitations of XAI and attention mechanisms for PII detection in low resourced languages.

## 2 LITERATURE REVIEW

The detection of personally identifiable information has been a growing area of interest in the field of natural language processing and information security. With the rapid proliferation of digital data and the increasing concerns over privacy protection, the accurate and efficient detection of PII has become a critical research area. Previous studies have primarily focused on high-resource languages, such as English, where abundant annotated datasets and NLP tools are available.

However, the effectiveness of these advancements are still tethered to the availability of extensive, high-quality corpora, presenting a formidable challenge for low-resource languages where such datasets are scarce or entirely absent. [2]. Additionally, the extension of these methods to low-resource languages has presented significant challenges also due to linguistic complexities [2] [20].

The task of privacy protection in low-resource languages has gained attention due to its unique set of challenges. Existing literature has highlighted the scarcity of annotated PII datasets in low-resource languages and the subsequent limitations in model training and evaluation [23]. Additionally, the linguistic nuances and specific characteristics of low-resource languages, such as Luganda, present additional complexities for accurate PII detection and anonymization [20]

The utilization of deep learning models with attention mechanisms has shown considerable promise in effectively capturing the nuances of language, thereby enabling accurate identification of PII in texts [24]. The attention mechanism allows the models to focus on the most relevant parts of the input, which is particularly advantageous in low-resource language scenarios [17]. Previous works have demonstrated the effectiveness of attention-based models in improving PII detection accuracy, especially when dealing with small, annotated datasets, as is often the case with low-resource languages [17].

With the advent of deep learning, a transformative shift occurred in the field. State-of-the-art models such as XLM-RoBERTa [13] and DeBERTa [7] demonstrated unprecedented capabilities in a wide array of language understanding tasks. These models, powered by intricate neural network architectures, set new benchmarks in the NLP community. Yet, their performance on low-resource languages was impeded by the limited training data available, which is a critical factor in the model's ability to learn and generalize [20].

In an attempt to bridge this gap, the Masakhane project emerged as a beacon of hope, particularly for African languages, which have traditionally been neglected in NLP research. The project's AfroXLM-R model was specifically tailored to accommodate the unique linguistic features of African languages, marking a commendable stride towards inclusivity [19]. Despite these efforts, the disparity in performance between high-resource and low-resource languages persists, underscoring the necessity for continued research and development in this area.

Futhermore, the "black-box" nature of deep learning models has been a significant obstacle in their application to privacy-sensitive tasks such as PII detection. To address this challenge, researchers have increasingly turned to Explainable AI techniques to provide insights into the decision-making processes of deep learning models [10] [14]. By employing XAI methodologies, researchers enhance the transparency and interpretability of the models, as explainability is essential for users to well trust, understand, and adapt to powerful AI applications [8].

It noteworthy to mention that in high resource languages, higher advances have been made including automating PII detection [22]

[16], classification and anonymisation [11] , development of open source libraries [21] and tools [6].

While such advances have ben made in high-resource languages, there is a noticeable gap in the literature regarding the application of advanced privacy-preserving technologies to low-resource languages. The current study seeks to bridge this gap by presenting a comprehensive case study of applying attention-based deep learning models and Explainable AI techniques to the specific low-resource language of Luganda, thereby contributing valuable insights and methodologies for privacy protection in similar linguistic environments.

## 3 LIMITATIONS OF XAI AND ATTENTION MECHANISMS FOR PII DETECTION IN LOW RESOURCED LANGUAGES

Overtime, XAI and attention mechanisms have become increasingly crucial in sensitive domains such as the detection of PII in text data. While substantial progress has been made in high-resource languages, there are still limitations the application of these advanced techniques to low-resourced languages.

### 3.1 XAI Limitations for PII Detection in Low Resourced Languages

XAI aims to make the predictions of complex models understandable to humans, which is essential for trust and accountability. However, evidently, there are disparities in the effectiveness of XAI when applied to low-resourced languages, including the scarcity of annotated data sets, linguistic tools, and pre-trained models for these languages severely hampers the interpretability [15]. That is, the absence of rich linguistic features, which are readily available for well-studied languages, leads to a diminished capacity for XAI models to provide meaningful interpretations in low-resourced linguistic contexts. This lack of nuanced linguistic understanding is a significant barrier, as PII detection often hinges on subtle contextual cues and language-specific idiosyncrasies [15][18].

### 3.2 Limitations of Attention Mechanisms for PII Detection in Low Resourced Languages

Attention mechanisms allow models to 'focus' on specific parts of the input data when making predictions, seemingly offering an improvement in the model's performance. The reliance of attention mechanisms on extensive training data sets, which are scarce for low-resourced languages, is a critical limitation [3].

Additionally, attention mechanisms typically use the hidden state of the recurrent neural network (RNN) to determine which parts of an input sequence are most important. Named entities are then located based on context information around high attention words which is not always the case, and a systematic way to locate the start and end of an entity is desired [12].

Furthermore, attention mechanisms rely heavily on the existence and appearance of named entities in a sentence. However, this assumption is not always true; entities such as dates, times, and alphanumeric IDs can be equally important as a named entity and must be located and classified in order to consider the document as de-identified [25]. In scenarios characterized by sparse data, such as those involving low-resourced languages, attention weights may reflect artifacts of overfitting or other biases rather than true explanatory factors [25].

### 3.3 Potential Approaches to PII Detection in Low Resourced Languages

Despite the challenges inherent in working with low-resourced languages, there have been promising efforts to bridge the gap in most low resourced language NLP tasks that PII detection could leverage. Various approaches have advanced low resourced language NLP including PII detection, including transfer learning adapts models from high-resourced to low-resourced languages, mitigating data scarcity [4], multilingual models to capture cross-lingual representations, aiding PII detection across diverse languages [26], data augmentation techniques, like back-translation, enrich training data, enhancing model robustness [5], semi-supervised and self-supervised learning leverage unlabeled data, reducing annotation needs [9], domain adaptation methods bridge distribution gaps between languages, and improving model generalization, and collaborative efforts share resources, including datasets and models.

## 4 METHODOLOGY

As shown in figure 1, the framework majorly has the layers of data input, data preprocessing, analyser, anonymiser and tokeniser Instances.

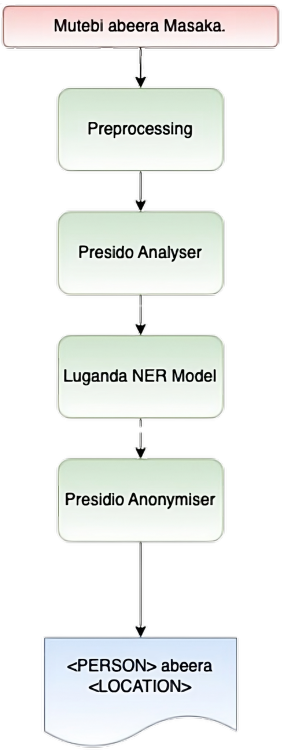


**Figure 1: Visualisation of the implementation flow**

## 4.1 Dataset Description

This research leverages the" Conrad747/lg-ner" [1] dataset, an annotated corpus comprising text data in the Luganda language. The dataset, rich with annotations tailored for named entity recognition (NER), is instrumental in identifying potential PII embedded within the texts. The dataset provides tokenised text along with corresponding named entity tags. It is comprised of a total of 2,979 samples, divided into training, validation, and test sets with 2085, 358 and 536 samples respectively. Each sample consists of text sequences annotated with Named Entity Recognition (NER) tags with main features of tokens (sequence of tokens representing the text) and ner tags (sequence of labels corresponding to the NER tags for each token). Each example consists of a sequence of tokens, where each token is associated with a label indicating the named entity type (e.g., person names, organizations, locations, etc.).

Figure 2 shows the most common tags in the dataset, the size of each word in the cloud corresponds to its frequency in the dataset. The word cloud correctly shows that words that are very common in Luganda vocabulary like *"mu, ku, nnga, ye, nti"* occur more frequently in the dataset since these are mostly joining words used in every sentence


**Figure 2: Common tokens in the dataset**

## 4.2 Data Preprocessing Steps

Data preprocessing involved several steps to prepare the datasets for model training. Initially, text normalization standardizes the text format across datasets. Following this, tokenization breaks down sentences into individual words or tokens. The preprocessing phase also includes removing irrelevant features, such as non-linguistic symbols, and encoding the data, particularly the NER tags, to be suitable for model training.

For instance, Text sequences were tokenized using the Auto Tokenizer with truncation and padding to ensure uniform length sequences and NER tags were aligned with the tokenized inputs to maintain consistency between tokens and labels as shown in table 1.

## 4.3 Exploratory Data Analysis visualizations

The Exploratory Data Analysis aimed to uncover the distribution, patterns and insights within the data, focusing on Distribution of PIIs, Linguistic features, Frequency and types of PII present in the dataset with the results presented in form of tables and charts.

| | | |
|---|---|---|
| "id2label": { | "8": "U-NORP", | "17": "B-ORG", |
| "0": "O", | "9": "B-DATE", | "18": "I-ORG", |
| "1": "B-PERSON", | "10": "I-DATE", | "19": "L-ORG", |
| "2": "I-PERSON", | "11": "L-DATE", | "20": "U-ORG", |
| "3": "L-PERSON", | "12": "U-DATE", | "21": "B-LOCATION", |
| "4": "U-PERSON", | "13": "B-USERID", | "22": "I-LOCATION", |
| "5": "B-NORP", | "14": "I-USERID", | "23": "L-LOCATION", |
| "6": "I-NORP", | "15": "L-USERID", | "24": "U-LOCATION" |
| "7": "L-NORP", | "16": "U-USERID", | } |

**Table 1: A table of aligned inputs tokenised**

Figure 3 shows the distribution of NER (Named Entity Recognition) tags across different entity types. Each NER tag corresponds to a specific entity type such as person, organization, location, etc. The countplot displays the frequency of each NER tag in the dataset. From the visualization, certain NER tags like a Person's name (*NER tag = [1,2,3,4]*) occur from frequently than a DATE (*NER tag = [8,9,10,11,12]*) for example which might affect how the model performs on that particular tag.

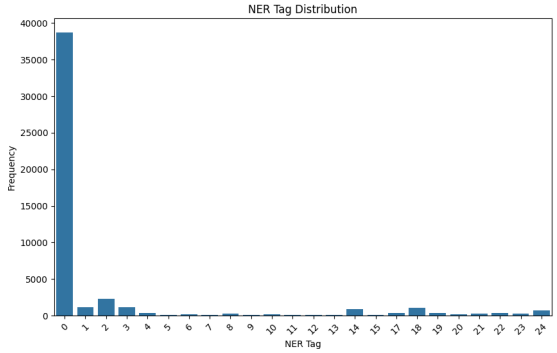


**Figure 3: Distribution of NER tags across different entity types**

As shown in figure 4, which displays the top 10 most common NER tags in the dataset along with their frequencies to identify the most prevalent entity types recognized by the NER model. From the plot, entity types like PERSON and LOCATION occur more frequently in the dataset.

As illustrated in figure 5, it is evident of insights into the complexity and density of entity mentions within sentences, which can influence model design and performance. The plot shows that most sentences have a good density of entity mentions which would contribute to better training of the model.

The scatter plot in figure 6 visualizes the relationship between sentence length and the number of NER tags. Each point represents a sentence, with its x-coordinate being the sentence length and y-coordinate being the number of NER tags. It helps in identifying any patterns or correlations between these two variables. From the

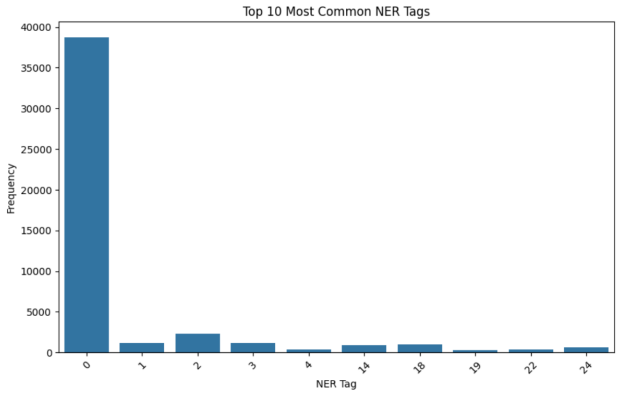


**Figure 4: The top 10 most common NER tags**

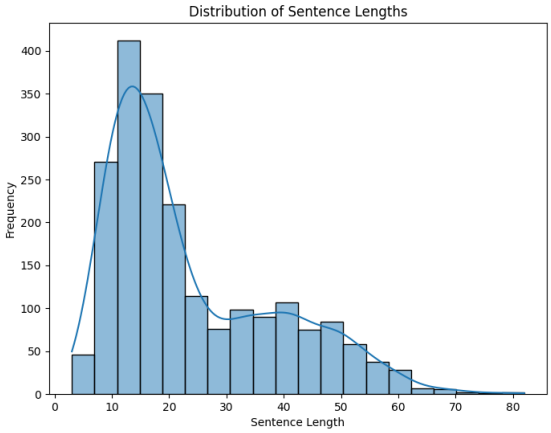


**Figure 5: The distribution of the number of NER tags per sentence**

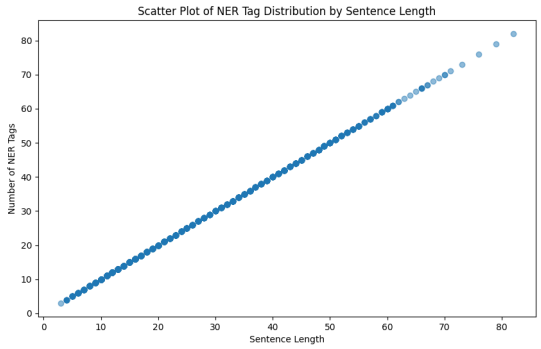


**Figure 6: The relationship between sentence length and the number of NER tags**

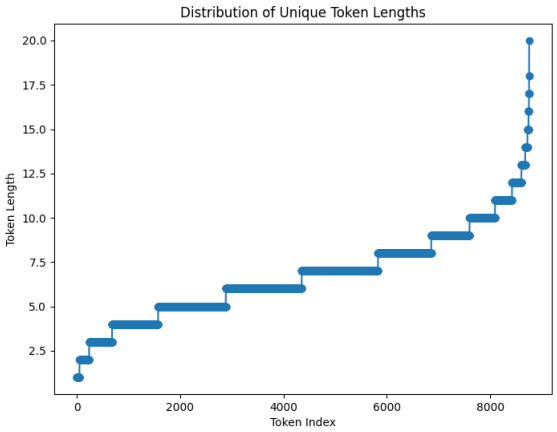


**Figure 7: Distribution of unique token length**

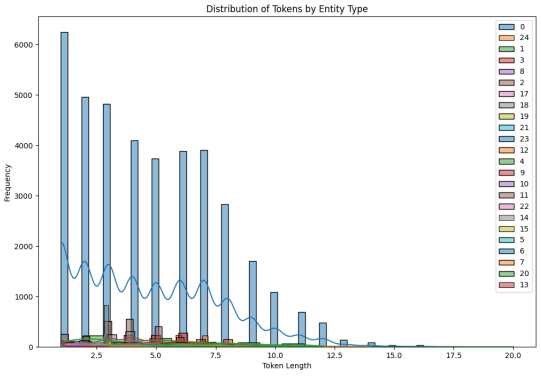


**Figure 8: Tokens categorized by their entity types**

plot, longer sentences tend to have a higher density of Ner tags which would imply that the longer the sentence, the more likely it is to find more PII in that sentence.

Figure 7 shows the distribution of unique token lengths with each point on the plot representing the length of a unique token, sorted by token index.

Figure 8, shows the distribution of tokens categorized by their entity types. Each entity type is represented by a different color in the histogram.

## 4.4 Deep Learning Models

Three distinct deep learnig models were implemented i.e. the luganda-ner-v6, deberta-v3-base, and afroxlmr-large-ner-masakhaner.

The Luganda-NER-v6 model is a specialized model designed for Named Entity Recognition (NER) in Luganda, a language spoken in Uganda. Based on the xlm-roberta-base architecture, this model leverages the Transformer's powerful self-attention mechanism to process text. It comprises about 277M parameters that enables the model to capture complex patterns and features specific to

Luganda named entities. It was fine-tuned on a dataset specifically annotated for NER tasks, which includes various entity types such as names, locations, and organizations. The primary role of luganda-ner-v6 is to identify and classify named entities within Luganda text, aiding in tasks like information extraction and data analysis. Its success is reflected in its high accuracy and F1 score, demonstrating the model's precision in understanding and categorizing Luganda entities. Was Selected as it has been fine-tuned on a dataset rich in Luganda entities, enabling it to discern and categorize PII with remarkable accuracy.

DeBERTa-v3-Base, is based on disentangled attention mechanism, which separately processes the content and position of each word in a sentence. This base model has 12-layer architecture and 768 hidden size, it contains 86M backbone parameters with a vocabulary containing 250K tokens which introduces 190M parameters in the Embedding layer. It trained on a massive 160GB dataset. The primary role of deberta-v3-base was machine translation. It has shown remarkable success, outperforming previous models like BERT and RoBERTa on a variety of natural language understanding benchmarks. This success is a testament to the model's ability to decode and interpret complex language patterns more effectively.

Afroxlmr-Large-NER-Masakhaner, is a multilingual NER model that addresses the need for language technology in African languages. Fine-tuned on the MasakhaNER dataset, which includes 20 African languages, this model is based on the Transformer architecture boasting about 559M parameters. The large number of parameters gives it ability to capture a wide range of linguistic features across multiple African languages such as 80.5% for Amharic in MasakhaNER 1.0 to 90.5% for Yorùbá in MasakhaNER 2.0, in dentifying entities of dates, locations, organizations, and persons. Its primary role is to facilitate NER tasks across a wide range of African languages, many of which have been historically underrepresented in NLP research. The model's success lies in its ability to accurately identify named entities across these diverse languages, contributing significantly to the inclusivity of language technologies.

### 4.5 Model Training and Evaluation Methods

Each of the 3 models were trained on the dataset training and evaluated on the validation and testing set. The training was monitored on epochs by model loss and accuracy gain as shown in figure 9. From figure 9, it is evident that DeBERTa-v3-Base model has the poorest initial training performance but had the highest learning gain, this suggest that further fine tuning would allow it to produce better results. Despite srting high, luganda-ner-v6 performance keeps flactuating suggesting sensitivity and the Afroxlmr-Large-NER-Masakhaner consistently maintaned an improvement and reached an optimal performance at an earlier epoch (7) than the other models. For each model, a the standard performance metrics such as precision, recall, and the F1 score were utilised to evaluate and compare the models. These metrics were calculated for each category of PII identified within the dataset, providing a assessment of the models' effectiveness. In parallel, XAI techniques were integrated to offer a window into the inner workings of the models. By applying LIME, the the models' predictions wer interpreted and shone a light on the influential features that underpin the detection of PII. This level of transparency is not merely a byproduct

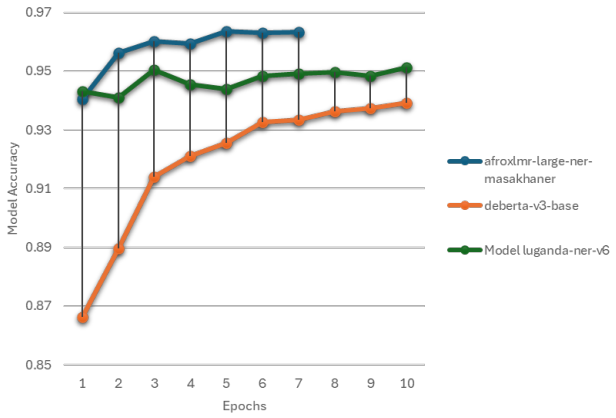


**Figure 9: Model Learning rate**

of the methodology but a deliberate effort to foster trust and understanding in AI systems while simultaneously enhancing the adaptability of the models to the unique linguistic nuances present in low resource languages texts like Luganda.

## 5 RESULTS & DISCUSSION

### 5.1 Model Output Results

To demonstrate the model's capabilities, below are some of the sample test cases:

**case 1:** text = *"Ssemakula yategese ekivvulu okutalaaga ebitundu omuli Buddu ne Bulemeezi."*

The model predicts the following entities: PERSON: *Ssemakula* LOCATION: *Buddu, Bulemeezi*

**case 2:** text = *"Katikiro Ssebugwaawo asisinkanye Minisita wa Kampala Minsa Kabanda"*

Model Detected PII entities:*['Katikiro', 'Ssebugwaawo','Minisita', 'Kampala', 'Minsa', 'Kabanda']*

which are indeed PII entries.

### 5.2 Model Evaluation Results and Comparison

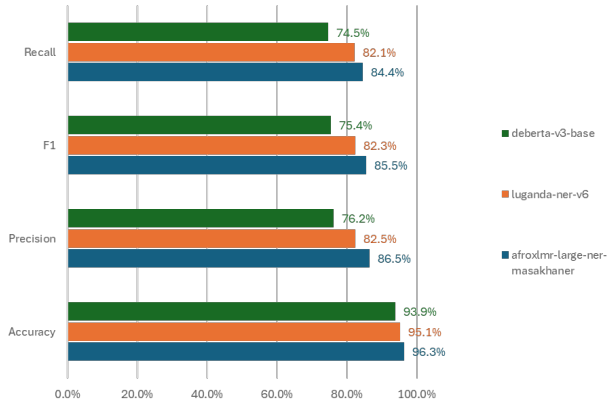


**Figure 10: Model performance comparison**

A shown in figure 10, the afroxlmr-large-ner-masakhaner model has the highest accuracy with 96.3%, followed closely by luganda-ner-v6 at 95.1%, and deberta-v3-base at 93.9%. Precision, crucial for minimizing false positives, favors afroxlmr-large-ner-masakhaner at 86.5%, luganda-ner-v6 at 82.5%, and deberta-v3-base at 76.2%. The F1 score, balancing precision and recall, highlights afroxlmr-large-ner-masakhaner's lead with 85.5%, luganda-ner-v6 at 82.3%, and deberta-v3-base at 75.4%. Similary, recall emphasizes afroxlmr-large-ner-masakhaner's dominance at 84.4%, followed by luganda-ner-v6 at 82.1%, and deberta-v3-base at 74.5%.

In conclusion, while all models perform well, afroxlmr-large-ner-masakhaner consistently excels, showcasing the need for tailored linguistic models in PII detection within diverse language contexts.

| Entity Type | Precision | Recall | F1 Score | Number |
|---|---|---|---|---|
| NORP | 94.6% | 78.4% | 85.7% | 111.0 |
| USERID | 81.2% | 76.5% | 78.8% | 17.0 |
| LOCATION | 79.1% | 83.8% | 81.4% | 357.0 |
| PERSON | 77.7% | 78.9% | 78.3% | 336.0 |
| ORG | 77.4% | 71.9% | 74.5% | 114.0 |
| DATE | 50.0% | 50.0% | 50.0% | 32.0 |
| Overall | 79.0% | 78.8% | 78.9% | - |

**Table 2: Model performance across different PII categories**

Table 2 shows the results of the PII detection in Luganda text across different entity types. The NORP (Nationalities or religious or political groups) category shows the highest precision at 0.946, indicating that when the model predicts an entity as NORP, it is correct 94.6% of the time. However, its recall is lower at 78.4%, suggesting that the model misses some NORP entities present in the text.

The USERID category has a good balance between precision and recall, with scores of 0.812 and 76.5% respectively, leading to an F1 score of 78.8% This indicates a relatively reliable performance in detecting user IDs within the text.

LOCATION entities are detected with a precision of 79.1% and a recall of 83.8%, resulting in an F1 score of 81.4%. This is noteworthy because it suggests the model is quite adept at identifying locations, which often have clear contextual indicators.

The detection of PERSON names has nearly equal precision and recall, both around 78%, showing that the model has a consistent performance in identifying individual names, although there is room for improvement.

ORG(Organizations) detection has a precision of 77.4% and a recall of 71.9%, with an F1 score of 74.5%. This indicates a moderate level of accuracy in identifying organizations, which can be challenging due to the varied ways organizations can be referred to in text.

The DATE category has the lowest performance with equal precision and recall at 5, leading to an F1 score of 0.5. This suggests significant difficulty in detecting dates, which could be due to the complexity of date formats and expressions in Luganda.

Overall, the model achieves an F1 score of 0.789, which is quite robust. However, the varying performance across different categories highlights the challenges in PII detection in low-resource languages. The results suggest that further refinement is needed, particularly in improving recall for NORP and precision for DATE entities. Additionally, the relatively low number of USERID and DATE entities in the dataset may have influenced the model's ability to learn from these examples effectively.

## 5.3 XAI Results

The applied XAI technique is LIME, which provides explanations for individual predictions made by the models by highlighting the words that contributed the most to the model's prediction. Typically, LIME visualisations (i.e. figures 11, 12, 13) were used to present results which contains the model's predictions,features contributions, and the actual prediction for each feature. The height of each bar indicates the probability assigned by the model to the corresponding NER tag label. The contributions associated with each word in the input sentence indicate how much they influence the model's prediction for specific NER tag labels. The actual prediction section shows the input sentence with varying shades representing the importance of each word in the model's decision-making process.Darker shades imply higher importance to the prediction for the corresponding NER tag labels, while lighter shades suggest less influence.

*5.3.1 XAI for luganda-ner-v6 .* As shown in figure 11, Each index corresponds to a specific NER tag as follows, *18: 'I-ORG', 19: 'L-ORG', 4: 'U-PERSON', 5: 'B-NORP' "other": represents all other NER tag labels not explicitly listed here, such as 'O' (non-entity)* or other entity types not included in the provided list 1. From the feature contribution graphs and actual senstence prediction for luganda-ner-v6 in figure 11, it can be seen that "*asisinkanye*" has a higher importance to the prediction of the NER tags for the Luganda NER model, followed by words like "*Katikiro*", "*Ssebugwawo*", "*Minsa*", then lesser importance for "*Minisita*", "*wa*", "*Kampala*", and "*Kabanda*" respectively.

*5.3.2 XAI for deberta-v3-base.* From figure 12, we can observe that for the indicies of 18, 19, 4, 5, and "other" which correspond to a specific NER tags of *0: 'O', 1: 'B-PERSON', 2: 'I-PERSON', 3: 'L-PERSON', "other": all other NER tag labels not explicitly listed here, such as 'B-ORG', 'I-ORG', 'L-ORG', 'U-ORG', etc.* It can be observed from feature contribution graphs and actual senstence prediction in figure 12, that all the words in the sentence *"Katikiro Ssebugwaawo asisinkanye Minisita wa Kampala Minsa Kabanda"* have equal importance to the prediction of the NER tags for this model.

*5.3.3 XAI for afroxlmr-ner-masakhaner.* As shown from figure 13, for the indices of 19, 18, 2, 3, and "other" which correspond to NER tag as *19: 'L-ORG', 18: 'I-ORG', 2: 'I-PERSON', 3: 'L-PERSON' "other": all other NER tag labels*, the words "*Ssebugwawo*", "*Kampala*" have the highest importance to the prediction of the NER tags for the afroxlmr-ner-masakhaner model, followed by "*Minista*", "*Katikiro*", "*asisinkanye*", and "*wa*", with "*Minsa*" and "*Kabanda*" as the least important.

## 5.4 Model and XAI Selection Justification

The models of Luganda-NER-v6, DeBERTa-v3-Base, and Afroxlmr-Large-NER-Masakhane, where each selected not arbitrary but as a calculated decision grounded in the unique attributes and proven efficacy of each model.

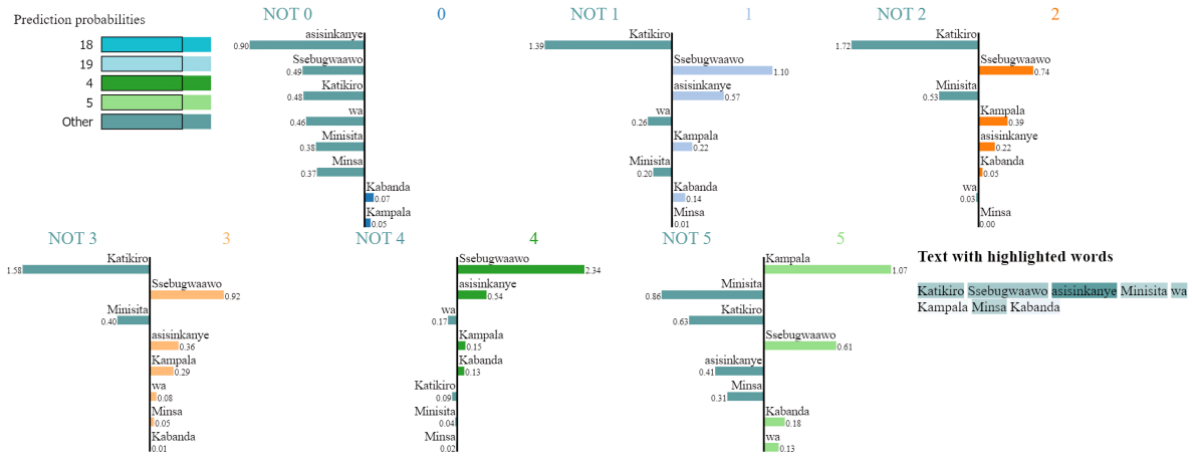


**Figure 11: luganda-ner-v6 LIME visualisation**

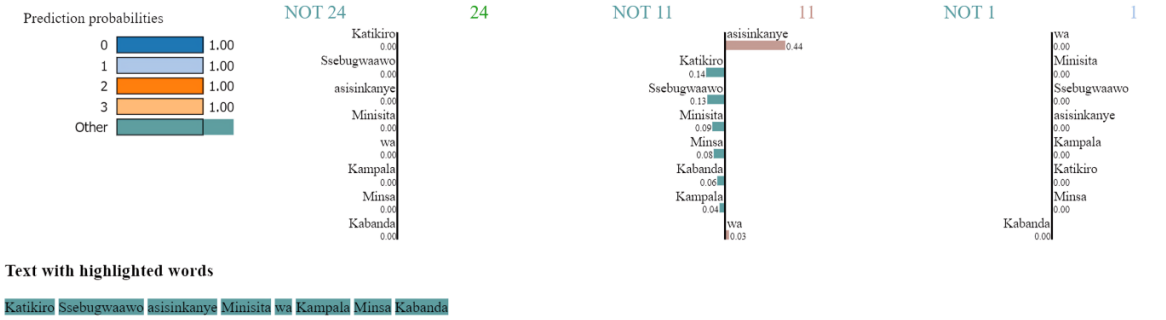


**Figure 12: deberta-v3-base LIME visualisation**

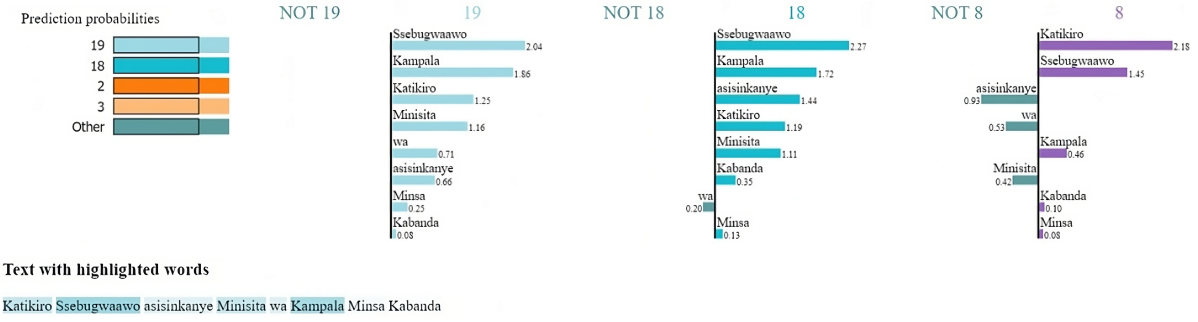


**Figure 13: afroxlmr-ner-masakhaner LIME visualisation**

Luganda-NER-v6 was selected for its specialization in the Luganda language, Its design and training tailored for Named Entity Recognition (NER) within the NER linguistic context of Luganda text.

The DeBERTa-v3-Base was chosen for its innovative disentangled attention mechanism which enables it to process content and positional information distinctly, achieving a nuanced comprehension of context. This is vital for the accurate detection of PII, where the subtleties of language play a significant role. The model's training on an expansive 160GB dataset equips it with an extensive knowledge base.

From the fact that the Afroxlmr-Large-NER-Masakhaner is a multilingual, having been trained across 20 African languages and in a low resource setting. This positions it as an invaluable tool for PII detection in low-resource languages, including Luganda. Its fine-tuning on the MasakhaNER dataset, which is specifically curated

for NER tasks across African languages, ensures its adeptness in pinpointing PII. Moreover, its role in fostering inclusivity within language technologies cannot be overstated.

Similarly, the choice of Local Interpretable Model-agnostic Explanations(LIME) as the XAI technique was deliberate based on model-agnostic Explanations, Feature Importance in Low-Resource Settings and visualisation. LIME offer methods to understand feature importance in low resource settings. Furthermore, the technique is model-agnostic hence flexibility to allow the seamless integration of XAI into the deep learning pipeline regardless of the final model architecture.

## 6 LIMITATIONS AND FUTURE WORK

This section presents limitations of this study and potential future improvement.

The scarcity of annotated datasets for PII in Ugandan languages stands as a formidable barrier. This limitation not only hampers the development of specialized multi-lingual models but also restricts the thorough evaluation of their efficacy. Furthermore, the comparative analysis of Explainable AI techniques was constrained primarily to LIME. A more extensive comparison could unveil deeper insights into the decision-making processes of the models. Another notable limitation is the imbalance in available Name Entity Recognition (NER) categories. This imbalance led to skewed model performance, potentially biasing the results, and limiting the generalizability of the findings. Additionally, the substantial computational resources required for training and evaluating transformer based deep learning models posed a significant challenge with requirement of high resources like RAM, GPUs and storage. The approach also assumes uniform characteristics across all domains, an assumption that may not always hold true. This could adversely affect the accuracy of the models when applied to domain-specific scenarios.

Looking ahead, (1) the development of a comprehensive PII dataset for Ugandan languages is imperative. Such a dataset would catalyze the training and evaluation of robust multi-lingual models. (2) An explicit implementation of attention mechanisms could further enhance model interpretability and performance, while leveraging transfer learning from high-resource languages promises to improve model robustness and adaptability. (3) The application of a broader range of XAI models, including SHAP and Eli5, will allow for a more nuanced understanding of model behavior. (4) Generating synthetic PII examples could also mitigate the issue of imbalances PII categories and improve model training. (5) Additionally, exploring the development of lightweight and efficient model architectures could increase adaptability and reduce computational demands. (6) Incorporating domain-specific knowledge and data into PII detection models holds the potential to significantly enhance their accuracy and applicability in real-world scenarios.

By addressing these limitations and pursuing these avenues of future work, further research will contribute meaningfully to the advancement of PII detection in low-resource language settings, paving the way for more secure and privacy-conscious computational linguistics research.

## 7 CONCLUSION

In this investigation into PII detection and anonymization in low-resource languages, particularly with Luganda as the primary focus, the deep learning models that were evaluated included xlm-roberta-base, microsoft/deberta-v3-base, and masakhane/afroxlmr-large-ner-masakhaner-1.02.0. The findings revealed that while xlm-roberta-base demonstrated commendable precision, recall, and F1 score, it was outperformed by masakhane/afroxlmr-large-ner-masakhaner-1.02.0, which exhibited superior performance across all evaluation metrics.

Additionally, the evaluation across distinct PII categories highlighted the nuanced nature of PII detection task, with varying performance observed for different types of information such as names, persons, and places. This underscores the importance of tailored approaches focusing on specific types of PII to enhance detection accuracy and mitigate privacy risks associated with PII exposure.

Furthermore, the incorporation of explainable AI (XAI) techniques provided invaluable insights into model decisions, enhancing transparency and interpretability. This transparency not only fosters trust in the models but also facilitates informed decision-making regarding their deployment and usage.

In conclusion, the study underscores the significance of leveraging advanced deep learning models and XAI techniques to improve PII detection and anonymization in low-resource languages. By refining model performance, enhancing transparency, and adopting tailored approaches, it can effectively mitigate privacy risks and bolster trust in PII detection systems, thereby contributing to the development of more secure and privacy-preserving digital ecosystems.

## ACKNOWLEDGMENTS

We acknowledge the contribution of the MSc. Computer Science (Data Science & AI major) class of 2023/24 at Makerere University particulary for the feedback in discussion sessions.

## REFERENCES

[1] [n. d.]. “Conrad747/lg-ner · Datasets at Hugging Face”. https://huggingface.co/datasets/Conrad747/lg-ner

[2] 2023. Building Text and Speech Datasets for Low Resourced Languages: A Case of Languages in East Africa. https://lanfrica.com/record/building-text-and-speech-datasets-for-low-resourced-languages-a-case-of-languages-in-east-africa

[3] Vishvajit Bakarola and Jitendra Nasriwala. 2021. Attention-based Sequence to Sequence Learning for Machine Translation of Low Resourced Indic Languages – A case of Sanskrit to Hindi. *International Journal of Engineering Trends and Technology* 69, 9 (Sept. 2021), 230–235. https://doi.org/10.14445/22315381/ijett-v69i9p227

[4] Michael Beukman and Manuel Fokam. 2023. Analysing Cross-Lingual Transfer in Low-Resourced African Named Entity Recognition. *arXiv preprint arXiv:2309.05311* (2023).

[5] Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440* (2017).

[6] Somchart Fugkeaw, Ananya Chaturasrivilai, Pitchayapa Tasungnoen, and Weerapat Techaudomthaworn. 2021. AP2I: Adaptive PII Scanning and Consent Discovery System. In *2021 13th International Conference on Knowledge and Smart Technology (KST)*. 231–236. https://doi.org/10.1109/KST51265.2021.9415803

[7] Gaia Gambarelli, Aldo Gangemi, and Rocco Tripodi. 2022. Is Your Model Sensitive? SPeDaC: A New Benchmark for Detecting and Classifying Sensitive Personal Data. *arXiv (Cornell University)* (01 2022). https://doi.org/10.48550/arxiv.2208.06216

[8] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Science Robotics* 4, 37 (2019), eaay7120. https://doi.org/10.1126/scirobotics.aay7120 arXiv:https://www.science.org/doi/pdf/10.1126/scirobotics.aay7120

[9] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. *Advances in neural information processing systems* 27 (2014).
[10] Rishika Kohli, Shreyas Chatterjee, Shaifu Gupta, and Manoj Singh Gaur. 2023. Tracking PII ex-filtration: Exploring decision tree and neural network with explainable AI. In *2023 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*. 183–188. https://doi.org/10.1109/ANTS59832.2023.10469568
[11] Poornima Kulkarni and NK Cauvery. 2021. Personally identifiable information (pii) detection in the unstructured large text corpus using natural language processing and unsupervised learning technique. *International Journal of Advanced Computer Science and Applications* 12, 9 (2021).
[12] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360* (2016).
[13] Bing Li, Yujie He, and Weiran Xu. 2021. Cross-Lingual Named Entity Recognition Using Parallel Corpus: A New Approach Using XLM-RoBERTa Alignment. *arXiv (Cornell University)* (01 2021). https://doi.org/10.48550/arxiv.2101.11112
[14] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2020. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* 23, 1 (12 2020), 18–18. https://doi.org/10.3390/e23010018
[15] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lecue, Gianclaudio Malgieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith, and Simone Stumpf. 2024. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion* 106 (2024), 102301. https://doi.org/10.1016/j.inffus.2024.102301
[16] Richard Marciano, William Underwood, Mohammad Hanaee, Connor Mullane, Aakanksha Singh, and Zayden Tethong. 2018. Automating the detection of personally identifiable information (PII) in Japanese-American WWII incarceration camp records. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2725–2732.
[17] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2020. Deep Learning Based Text Classification: A Comprehensive Review. https://doi.org/10.48550/arxiv.2004.03705
[18] Chinasa T. Okolo, Nicola Dell, and Aditya Vashistha. 2022. Making AI Explainable in the Global South: A Systematic Review. In *Proceedings of the 5th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies* (Seattle, WA, USA) *(COMPASS '22)*. Association for Computing Machinery, New York, NY, USA, 439–452. https://doi.org/10.1145/3530190.3534802
[19] Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Touré Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, Kevin Degila, Herman Kamper, Bonaventure F. P. Dossou, Chris Chinenye Emezue, Kelechi Ogueji, and Abdallah Bashir. 2020. Masakhane – Machine Translation For Africa. https://doi.org/10.48550/arxiv.2003.11529
[20] Wuraola Fisayo Oyewusi, Olubayo Adekanmbi, Ifeoma Okoh, Vitus Onuigwe, Mary Salami, Opeyemi Osakuade, Sharon Ibejih, and USMAN Musa. 2021. NaijaNER : Comprehensive Named Entity Recognition for 5 Nigerian Languages. https://doi.org/10.48550/arxiv.2105.00810
[21] Eidan J Rosado. 2022. PII-Codex: a Python library for PII detection. (2022).
[22] Md Hasan Shahriar, Abrar Hasin Kamal, and Anne V. D. M. Kayem. 2024. Discovering Personally Identifiable Information in Textual Data - A Case Study with Automated Concatenation of Embeddings. In *Advanced Information Networking and Applications*, Leonard Barolli (Ed.). Springer Nature Switzerland, Cham, 145–158.
[23] Samuel Sousa and Roman Kern. 2022. How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing. https://doi.org/10.1007/s10462-022-10204-6
[24] Amirsina Torfi, Rouzbeh A. Shirvani, Yaser Keneshloo, Nader Tavaf, and Edward A. Fox. 2020. Natural Language Processing Advancements By Deep Learning: A Survey. https://doi.org/10.48550/arxiv.2003.01200
[25] Hao Wang, Lekai Zhou, Jianyong Duan, and Li He. 2023. Cross-Lingual Named Entity Recognition Based on Attention and Adversarial Training. *Applied Sciences* 13, 4 (2023). https://doi.org/10.3390/app13042548
[26] Mesay Gemeda Yigezu, Michael Melese Woldeyohannis, and Atnafu Lambebo Tonja. 2021. Multilingual Neural Machine Translation for Low Resourced Languages: Ometo-English. In *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*. 89–94. https://doi.org/10.1109/ICT4DA53266.2021.9671270

# A APPENDIX

## A.1 Code Implementation

*A.1.1 link to the Github Repository.* Github Repository HERE

*A.1.2 Links to the the notebooks.*

- 01 Initial Deep Learning Notebook.ipynb: This notebook contains exploratory data analysis (EDA) and the exploration of the initial models.
- 02 xAI_PII.ipynb: the initial model's implementation and the application of Explainable AI (XAI) techniques are explained and demonstrated
- 03 xAI_LugandaNER.ipynb : The final notebook integrates additional models and XAI techniques for improved PII detection and transparency.