Internship Report 2024
MSc. in Machine Learning and Data Mining (MLDM)
Knowledge Extraction from Textual Data: Applications to the Discover of Land-Use and Land Cover Change Rules

Grisel Quispe Aramayo[1 2] **Arthur Crespin-Boucaud**[1] **Sarah Valentin**[2] **Maguelonne Teisseire**[2]

Abstract

This research investigates Land Use and Land Cover Changes (LULCC) in West Africa by systematically extracting and analyzing textual data from scientific literature. We developed a dataset from scientific articles, using classification approaches SVM, BERT to identify and categorize relevant segments. Through our analysis, we identified important bias, particularly the transition from general context to more focused temporal and quantitative information. These findings aim to improve our understanding of land cover dynamics and to aid in making better-informed decisions for managing land use in regions facing changes in this region.

Keywords: Natural Language Processing, Supervised Learning, Entity Extraction, Land Use Land Cover, LULCC

1. Introduction

Global warming and rapid changes in land use are having a big impact on places like the Sahel in West Africa, where getting enough water is already a major challenge. These environmental changes are causing more frequent and severe events, like the devastating floods in Niamey back in 2020 (CECC Project, 2024). To manage these challenges effectively, it's really important to understand what's driving these land use and cover changes, whether it's climate shifts, population growth, or more farming.

As part of the larger CECC (Cycle de l'eau et changement climatique) project, which focuses on figuring out future changes in water resources because of global climate change, our study focuses on Land Use and Land Cover Change (LULCC) in West Africa. We did this by creating a dataset from textual information we extracted and labeled from scientific articles that talk about LULCC processes.

We used advanced techniques like hierarchical extraction, supervised methods, and entity extraction to classify and analyze this information. This let us dig into the key features of different classes, especially focusing on the ones that are crucial for understanding the context and scale of LULC changes.

Given how complex and important these issues are, our research is guided by the question: ”How can we effectively classify and analyze textual information related to Land Use and Land Cover Change (LULCC) in West Africa to better understand the temporal and quantitative dynamics driving these changes?”

Our aim is not just to better understand these changes but also to improve the methods we use for labeling and analyzing this kind of data. By doing this, we hope to provide valuable insights that can help guide better land use decisions in the region, which lines up with the broader goals of the CECC project.

1.1. State of the art

Automatically extracting information from scientific documents introduces distinct challenges compared to other machine learning tasks, such as image classification or structured data analysis. While both text and image classification share foundational principles, the intricacies involved in processing scientific texts are notably more complex. The preprocessing phase, in particular, demands careful handling due to the diverse range of domains—such as chemistry, biology, agronomy, and computer science—represented in these documents. This complexity necessitates domain-specific expertise to ensure accurate dataset construction and validation. Notably, there is a lack of prior research focused on extracting textual information automatically specifically related to land use and land cover. To address this gap, we have developed an innovative methodology, drawing on advanced techniques and addressing the unique challenges of this domain. The details of our approach are outlined in the following sections.

1.1.1. Supervised techniques

When a labelled dataset is available, supervised approaches can be used to extract information from scientific documentation. A common approach is to convert knowledge extraction into a classification task. In this context, supervised approaches are popular in this field. The SVM (Support Vector Machine) (Wang et al., 2005) is particulary well known for text classification (Deng et al., 2019). This approach requires a labelled dataset, in the work of (Kiritchenko et al., 2010), the authors use a corpus composed of clinical trial publications, from which they aim to automatically extract key trial characteristics, such as eligibility criteria, sample size, drug dosage, and primary outcomes. They perform an interesting preprocessing step in which they transform PDF files into HTML or XML formatted text to preserve document hierarchy and remove irrelevant sections. The authors use a manually labeled dataset containing the 21 features they wish to extract. Finally, they train an SVM model to classify sentences into a ranked list that contains the target information.

In line with the approach of working with the corpus in a hierarchical manner and employing classical machine learning techniques for classification is the work of (Agarwal et al., 2020). In this study, the authors aim to extract procedures from technical documentation, such as product manuals and user guides. By converting the documents to text, they preserve the structure and hierarchy of the content. For feature extraction, they use the ’Bag of Words’ technique (van der Veen, 2023), one of the earliest methods in NLP (Natural Language Processing). The sentences are then classified into categories such as ’Actionable,’ ’Goal-based,’ ’Relatedness,’ ’Structural,’ and ’Context-based’ using a LinearSVC (Mishu & Rafiuddin, 2016) classification model.

An approach closely related to the one we propose in this paper is presented in (Koptelov et al., 2023) where rules are extracted from urban planning documents. The authors utilize a dataset consisting of sentences labelled with categories such as Verifiable, Non-verifiable, Informative, and Non-pertinent. To classify the sentences, they employ an approach based on three rounds of classification, using the CamemBERT model (Martin et al., 2019), which is based on a version of BERT (Koroteev, 2021), a transformer-based pre-trained model that uses bidirectional attention mechanisms to analyze text from both directions simultaneously. This approach allows BERT to better understand word meanings in context and capture richer semantic information, making it a leading model in NLP tasks (Sabharwal & Agrawal, 2021).

1.1.2. Non-supervised techniques

When dealing with data that lacks labels, using an unsupervised learning approach becomes an option. This technique allows for discovering patterns and insights without relying on predefined categories or annotations. However, in the study by (Blanchy et al., 2023), the study utilizes methods such as Topic Modeling (Snyder, 2015) to identify discussed topics, Regular Expressions to construct a quantitative database, and k-means clustering (Zhao et al., 2018) to extract relevant relationships by measuring distances between target clusters and determining factors.

2. Materials and Methods

Important definitions

Land Cover refers to a physical material on the surface of the earth including vegetation (like forests, grasslands, and crops), water bodies (like rivers, lakes, and oceans), artificial structures (like buildings and roads), bare soil, and other natural features. It can be determined by analyzing satellite and aerial imagery (National Oceanic and Atmospheric Administration, 2024).

Land Use documents how people are using the land, whether for development, conservation, or mixed uses. The different types of land cover can be managed or used quite differently (National Oceanic and Atmospheric Administration, 2024).

Land Process Change Land cover change denotes a change in certain continuous characteristics of the land such as vegetation type, soil properties, and so on, whereas land-use change consists of an alteration in the way certain area of land is being used or managed by humans (Patel et al., 2019).

2.1. Data

2.1.1. Corpus

Since no existing corpus met our specific criteria, we needed to create one ourselves. To do this, we used Google Scholar to find 69 English-language scientific studies and focused on the abstracts to make sure they were relevant. We used the following keywords to select the papers: land use change, land cover change, LULCC, reforestation, deforestation, urbanization, savannization , and agriculture change in combination with West Africa. The titles of these papers are listed in Appendix A.

2.1.2. Nomenclature

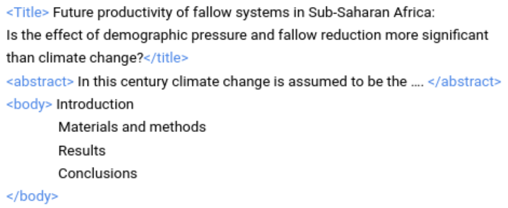
The nomenclature is a dictionary we created to compile terms related to land use, land cover, and the processes that cause changes in LULC in West Africa. This nomenclature is based on the JECAM nomenclature (JECAM, 2024), adapted to tropical environments. It was developed with the help of an expert in the field and includes nouns such as ’crop’, ’cereal’, and ’grain’, as well as processes like 'cultivation', 'savannization', and 'desertification'. It also features acronyms such as LULC (Land Use, Land Cover) and LULCC (Land Use Land Cover Change). To expand and refine the nomenclature, we searched for related terms in AGROVOC a database of the FAO, which we used to verify meanings and synonyms. The final nomenclature consists of 196 terms and is detailed further in Appendix B.

2.2. Methods

2.2.1. Preprocess

This process is essential for extracting relevant segments from the corpus of PDF documents, as conventional tools like PyPDF2 (PyPDF2, 2024) often introduce significant noise into the extracted data. Given that scientific documents include sections such as references, footnotes, Bibliography and more that are not intereset for us to analyze.

To address this issue, we adopted a methodology recognized in the literature that preserves the document's structural hierarchy during information extraction. We utilize GROBID (Lopez, 2009), an advanced tool for automatic bibliography extraction and structure recognition, which allows for more accurate data extraction. The process begins with converting the PDF document into a TEI file, which is then transformed into JSON to work exclusively with the title, abstract, and body text. Figure 1 presents a schematic of the resulting format after preprocessing.



```
<Title> Future productivity of fallow systems in Sub-Saharan Africa:
Is the effect of demographic pressure and fallow reduction more significant
than climate change?</title>
<abstract> In this century climate change is assumed to be the .... </abstract>
<body> Introduction
        Materials and methods
        Results
        Conclusions
</body>
```

Figure 1: Extraction Schema

2.2.2. Spliting sentences

To extract information effectively, it is necessary to divide the text into manageable segments. In this context, we have decided to segment the text into individual sentences, as a sentence represents the smallest unit of a paragraph. This level of segmentation was chosen because it allows for more precise and manageable manual analysis in later stages of the process.

For this segmentation, we utilized the Regular Expression library in Python (Regex)(Kuchling, 2023), using sequences of periods followed by spaces as delimiters. This approach ensures that each segment corresponds to a complete sentence.

To systematically organize and analyze the text segments, they are stored in a CSV file after being split. The first column of this file contains a 'key', which includes the name of the document from which the segment was extracted, along with the corresponding segment number. The second column stores the textual content of the segment.

The next step involves identifying candidate segments that may be relevant for information extraction. A segment is considered relevant if it contains at least one term from the nomenclature. Each segment undergoes a text analysis process using NLP techniques, including tokenization, stemming, and lemmatization (Hardeniya, 2016), with the SpaCy library (Malik, 2023). Regular expressions are subsequently used to ensure that the segments correspond to the terms in the nomenclature. Finally, the candidate segments from all documents in the corpus are randomized, yielding a dataset comprising 2,078 segments.

```
{
  "Key": "1-s2.eu-Seu3eu8521X12eueueu157
    -main.pdf.tei_13",
  "Segment": "In fact, a major reason
    why farmers grow cotton is to gain
    access to fertilisers for their
    subsistence food"
}
```

Figure 2: Example of Candidate Segment extracted.

Figure 2 illustrates an example of a segment that has been identified as a candidate for analysis. This segment contains a relevant nomenclature term, specifically the word 'cotton', and is associated with a key that identifies it within the dataset.

2.2.3. Dataset labelization

In the extracted dataset, segments were categorized based on their relevance: some contained valuable information, others exhibited significant noise (particularly those including tables), and some were deemed irrelevant. To assess the relevance of each segment, we conducted a manual labeling process.

To begin, we established three label categories: 0 for non-relevant segments, 1 for segments that could be relevant depending on the context, and 2 for segments definitively relevant. This labeling was independently carried out by three experts from different fields: Sarah Valentin, an expert in knowledge extraction; Arthur Crespin-Boucaud, specializing in rule-based modeling; and Grisel Quispe, who focuses on machine learning and data mining. We initially selected 50 segments at random, with each expert working independently. Afterward, we had meetings to resolve any discrepancies in the labeling. This iterative process was repeated three times, resulting in a total of 161 labeled segments. This approach allowed us to refine the category definitions and enhance labeling accuracy. Detailed definitions can be found in Table 1.

After developing and identifying the best-performing models based on this initial dataset, a new set of 160 segments was received for further testing. This additional dataset was labeled in a diferent way. Here, the labeling process was in charge of Arthur and Sarah each labeling 80 segments independently. This dataset was then used to test the performance of the previously trained models, allowing us to evaluate the models' generalization on newly labeled data.

2.2.4. Dataset Clasification

To classify the unlabeled segments in the dataset, we trained several models using an initial labeled set of 161 segments: 93 in class 0, 44 in class 1, and 24 in class 2. Due to the class imbalance (57.76% in class 0, 27.33% in class 1, and 14.91% in class 2), we used a round-based classification approach to improve accuracy.

In the first iteration, we merged categories 1 and 2 into a single class. This transformation resulted in a binary classification task, where class 0 contained 93 segments and the merged class (now labeled as class 1) contained 68 segments. This step allowed us to address the imbalance and create a more simplified classification task. In the second round, we further differentiated between classes 1 and 2, refining the classification to identify the specific categories. Figure 3 illustrates the schema.

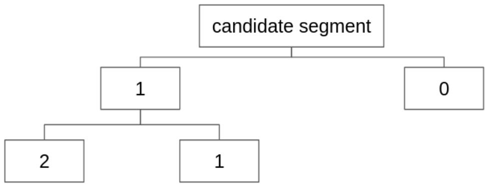


Figure 3: Classification schema

Since SVM (Deng et al., 2019) and BERT (Sabharwal & Agrawal, 2021)are known as top-performing methods for supervised text classification, we decided to use both to train our models and see how they compare. Previously, we performed a thorough cleaning of the textual segments to remove special characters and JSON or XML attributes, to retain only alphanumeric characters, spaces, dots, and percent symbols. To carry out this process, we used the Python Regex library (Kuchling, 2023) to ensure that the data was in a better format for further analysis.

Paths are designated for storing evaluation metrics, classification reports, and confusion matrices. The training configuration includes strategies for model saving and evaluation, ensuring that the model is initialized with the best-performing version identified during the training process.

To train the SVM model, we utilized the scikit-learn library (scikit-learn, 2024). The dataset is divided into training and evaluation sets, and to optimize results, we experiment with various splits (70/30, 80/20, 60/40). Additionally, we test different numbers of epochs, ranging from 10 to 25. To ensure reproducibility, a fixed random seed was applied. Additionally, stratification was employed during the dataset split, ensuring that the proportion of each class in both the training and testing sets mirrors the original distribution of labels. After this, the subsets were converted into a format compatible with NLP tasks within the Hugging Face dataset library environment (Hugging Face, 2024a). This conversion is crucial for ensuring that the model processes well-structured data, thereby enhancing the accuracy of model evaluation and overall performance.

Once the dataset was prepared, the text was directly converted into numerical representations using a TF-IDF (Term Frequency-Inverse Document Frequency) method. This approach assigns a weight to each term based on its frequency within a document and its rarity across the entire dataset, effectively highlighting the most important words for the classification task(Havrlant & Kreinovich, 2017).

To further optimize model performance, we employed a Grid Search with 3-fold cross-validation. This approach enabled us to determine the optimal hyperparameters for the classification model. The parameters evaluated during the search were as follows:

- **C**: [0.1, 1, 10, 100, 150, 200]
- **Kernel**: ['linear', 'rbf', 'poly']
- **Gamma**: [0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10]

For the other approach, we first prepare a text classification model using a pre-trained transformer model and tokenizer. We specifically use RoBERTa (Hugging Face, 2024b), which has been trained on extensive text corpora. We experiment with different data splits (70/30, 80/20) and test various numbers of epochs, ranging from 10 to 25. This pre-training helps the model use its existing knowledge, making it easier to adjust to our specific dataset. The RoBERTa model employs the AdamW optimizer (PyTorch, 2024) by default, a choice that benefits from its robustness in handling deep learning models.

2.2.5. Evaluation Metrics

There exist several metrics to evaluate the model performance. Here we define the ones that we are going to use; accuracy, precision, recall and F1 Score.

TP is the number of True Positives, **TN** is the number of True Negatives, **FP** is the number of False Positives, **FN** is the number of False negatives.

Accuracy: Measures the overall correctness of the model's predictions. It is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision measures the accuracy of positive predictions. It is defined as

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: Measures the proportion of actual positives that are correctly identified. This metric is important to us because it emphasizes the model's ability to capture as many relevant instances as possible. We prioritize recall because we prefer to have some false positives rather than risk losing important information. It is defined as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 Score: Is the harmonic mean of precision and recall. It provides a single metric that combines both precision (the accuracy of positive predictions) and recall (the ability to identify all relevant instances).

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

2.2.6. Information Extraction

We use Information Extraction (IE)(Singh, 2018) to transform unstructured text from our labeled dataset into a structured format, this will be useful to analyze the data effectively. A key part of this process is Named Entity Recognition (NER) (Sharnagat, 2014), a technique for automatically identify and categorize entities in a text, in our dataset the important information related to LULC dynamics. This includes extracting measurements, directions, dates, locations, and specialized vocabularies relevant to our study.

To implement NER, we rely on the SpaCy NLP library, along with tools like Regex and the Entity Ruler to refine the extraction process. These methods enable us to gather both quantitative data and qualitative insights, which are essential for understanding and analyzing LULC changes. The Table 2 shows the specific features we're extracting and the methods we're using to extract them.

Extracting and analyzing these entities, gives us both hard data and deeper insights into LULC changes. Quantitative entities, like measurements, numbers, percentages, and dates provide clear data for tracking trends and making comparisons. Qualitative entities, such as land cover terms and change processes, will be useful to understand the nature and causes of these changes.

3. Results

3.1. Labelization process

The experts' agreement on labeling was evaluated over three rounds, as shown in Table 3. In the first round, the agreement was low at 30%, revealing significant differences in how each expert understood the category definitions, suggesting that the initial definitions were too vague. After discussing these differences, the experts refined the definitions, leading to a much better agreement of 58% in the second round. This improvement showed that the discussions helped the experts align their interpretations.

In the third round, with 61 more segments added, the agreement slightly decreased to 56.6%. Although there was a small drop, the agreement remained fairly consistent, possibly due to the increased complexity of the new segments. Overall, despite some initial challenges, the ongoing process of discussion and refinement resulted in more consistent and accurate labeling of the dataset, with 161 segments labeled by the end.

Round	Match Agreement	Number of sentences
1	30%	50
2	58%	50
3	56.6%	61
Total		161

Table 3: Match Agreement by Round

3.2. Supervised Classification

This section presents the results of our supervised classification efforts, where we aimed to accurately categorize text segments into specific classes. We focused on enhancing recall to ensure that relevant segments were identified, even if it resulted in more false positives. The models were then evaluated on a newly labeled, unseen dataset to assess their generalization and effectiveness in handling class imbalances.

In this experiment, we used a Support Vector Machine (SVM) model with an 80/20 split between training and testing data. The dataset for this experiment consisted of three classes: Class 0: 60 segments, Class 1: 25 segments, and Class 2: 14 segments. This dataset was obtained after the second round of labeling.

To improve the model's performance, we combined Class 1 and Class 2 into a single class, resulting in a dataset with 60 segments for Class 0 and 39 segments for the newly merged Class 1. As shown in Table 4, this adjustment improved the model's ability to recognize what were previously Class 1 and Class 2, which had been poorly predicted when treated separately in an earlier experiment (see Appendix C Table 1 for details). The performance metrics, based on predictions in the testing data, now indicate a more balanced performance between the two classes, with Class 0 achieving a Precision of 0.727 and an F1-Score of 0.695, while the newly merged Class 1 achieved a Precision of 0.55 and an F1-Score of 0.58.

The improvement in recognizing the combined class suggests that the SVM model was better able to generalize when the distinctions between Class 1 and Class 2 were removed. This likely indicates that the model struggled to differentiate between these two classes due to their similarity or overlap in features. By merging them, the model could focus on broader distinctions, leading to improved performance across both classes in the testing data.

Class	Precision	Recall	F1-Score	Support
0	0.727	0.66	0.695	12
1	0.55	0.62	0.58	8

Table 4: Classification Report for the SVM Experiment on Testing Data

In our efforts to improve classification between Class 1 and Class 2, we conducted a cross-validation process to identify the optimal BERT hyperparameters. This cross-validation, detailed in Appendix B, utilized stratification to handle class imbalance across three folds. After determining the best-performing hyperparameters, we applied them to a new dataset that was received following the third round of labeling. This updated dataset included Class 0 with 93 segments, Class 1 with 44 segments, and Class 2 with 24 segments, totaling 161 segments.

We retrained the model using these optimal hyperparameters (C: 1, gamma: 1e-05, kernel: linear) with a new data split of 60/40 for training and testing, respectively to try to see if this new splitting would improve the model performance handling in a different way the imbalance. The performance results, summarized in Table 5, reflect the metrics specifically for Class 1 and Class 2 based on the test set.

The selected hyperparameters demonstrated reasonable performance for Class 1, achieving a Precision of 0.65, Recall of 0.83, and F1-Score of 0.73. However, the model struggled with Class 2, where it only achieved a Precision of 0.40, Recall of 0.20, and F1-Score of 0.27. The shift to a 60/40 split may have influenced these results by altering the data distribution and potentially affecting the model's ability to generalize effectively to the test set, particularly in distinguishing the more challenging Class 2.

Class	Precision	Recall	F1-Score	Support
1	0.65	0.83	0.73	18
2	0.40	0.20	0.27	10

Table 5: Classification Report for the Test Set Using Selected Hyperparameterst with SVM

After retraining our SVM model with the newly labeled data and encountering limitations in classifying Class 2, we explored the potential of BERT to enhance our classification results, with a specific focus on improving recall. To optimize the model's ability to correctly identify relevant segments, we experimented with different data splits and epochs. The details of these experiments, including various configurations, are provided in Appendix B. The key performance metrics from the most promising BERT experiments are summarized in Table 6.

We tested two main setups: the first with 10 epochs and an 80/20 split between training and testing data, and the second with 25 epochs and a 70/30 split. In the first setup, where we merged Class 1 and Class 2, the model achieved a Recall of 0.73 for Class 0 and an impressive 0.92 for the merged Class 1, indicating the model's strong capability to identify relevant segments in this configuration.

In the second setup, where Class 1 and Class 2 were treated separately, the recall was 0.71 for Class 1 and notably higher at 0.85 for Class 2. This improvement in recall for Class 2, from 0.20 in the previous SVM experiments to 0.85 with BERT, demonstrates a significant advancement in the model's ability to capture relevant segments within this challenging class.

These results highlight BERT's effectiveness in improving recall, particularly for Class 2, which was the most challenging to classify accurately in earlier experiments. The model's ability to achieve a recall of 0.85 for Class 2 with the 25-epoch, 70/30 split setup suggests that BERT is better at identifying relevant segments, even if it comes at the cost of generating some false positives. This aligns well with our goal of maximizing recall to ensure that fewer relevant segments are missed.

Following the final training phase, we evaluated the models trained in the previous step using BERT on a completely unseen dataset to validate their effectiveness. For this evaluation, we focused on two specific classification tasks: distinguishing between Classes 0 and 1, and distinguishing between Classes 1 and 2.

To assess these models, we obtained 160 additional segments that were newly labeled by two experts. The evaluation results on this unseen dataset are illustrated in the confusion matrix shown in Figure 4, and the detailed performance metrics are summarized in Table 7.

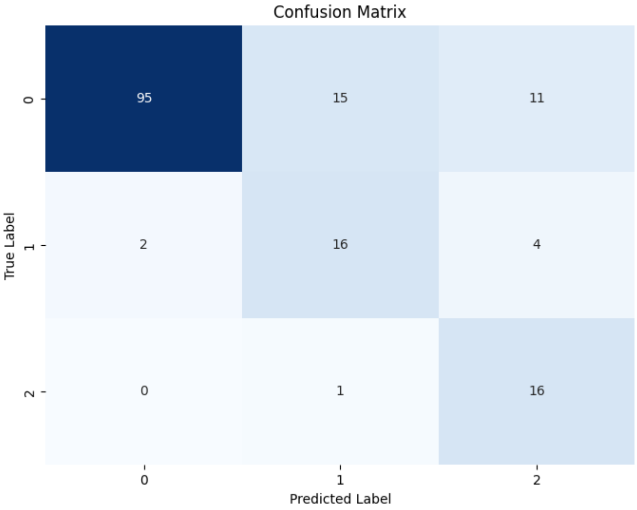


Figure 4: Confusion Matrix for Unseen Dataset

Class	Precision	Recall	F1-Score	Support
0	0.98	0.79	0.87	121
1	0.50	0.73	0.59	22
2	0.52	0.94	0.67	17

Table 7: Classification Report on Unseen Data with BERT

The confusion matrix and performance metrics reflect the models' ability to generalize to new data, particularly in identifying relevant segments in Class 2, which was previously challenging. The high recall for Class 2 suggests that our approach with BERT is effective in capturing a larger number of relevant segments, even if it leads to more false positives.

3.3. Information Extraction

In this section, we share the results of our information extraction process, beginning with an analysis of the top 10 segments with the most extracted entities. These segments were chosen from a dataset of 324 segments, divided into three classes: Class 0, Class 1, and Class 2. After reviewing these top segments, we also examine how entities are distributed across all segments in each class. This approach helps us understand both the most significant segments and the overall patterns in each class, giving us a clearer picture of the information our classification captured.

3.3.1. Analysis of Top 10 Segments

In Figure 5, a bar chart compares the total sum of various entities across the top 10 segments for each class.

Class 0 is mostly filled with noise-related data, especially CARDINAL entities, which likely come from tables. While there are also some DATE and LCover entities, these are less important because of the overall noisy nature of this class. The top-ranking segment in Class 0 has a total of 28 entities, 22 of which are cardinal data points from a table (see Appendix D, Table 1).

Class 1 shows a mix of different entities, with a strong focus on LProcess. This suggests that segments in this class are relevant to land cover processes, but they might need more context to be fully meaningful. The leading segment in Class 1 contains 9 entities, with 6 coming from land cover (LC) vocabulary. The main focus of this segment is the limited scientific evidence connecting NDVI (Normalized Difference Vegetation Index) trends to changes in cropland in the Sahel region (see Appendix D, Table 2).

Class 2 presents the most informative, with a high amount of LCover and PERCENT entities. Segments in this class provide the most detailed and relevant information for understanding changes in land cover, making Class 2 the most valuable in our analysis. The top segment in Class 2 has 16 entities, with 6 related to land cover vocabulary and 6 to percentages. It describes changes in land cover types between 1984 and 2001, providing percentage changes across various categories (see Appendix D, Table 3).

The bar chart clearly shows the differences in entity distribution across the three classes, highlighting Class 2's richness in information, the moderate significance of Class 1, and the noisy nature of Class 0.

3.3.2. Overall Entity Distribution by Class

The Figure 6 shows the distribution of different entity types across the three classes. In Class 0, it seems to cover a wide range of topics where terms like ”Land Cover Vocabulary (LCover)” and ”Land Cover Change Process (LProcess)” might appear, but these terms aren't really central to the main discussion. Mentions of ”Cardinal” numbers, ”Date”, and ”Verb Related to Change (VRC)” in Class 0 seem to act more as background information rather than key points of the study.

In Class 1, the main focus shifts to ”LCover,” which shows up more often than anything else, followed by ”LProcess.” There's a moderate amount of ”Cardinal” numbers and ”Date” entities, while ”VRC” entities are less common. ”Direction” and ”Percentage” entities are present too, but they're not really emphasized, suggesting that Class 1 is mostly about identifying and categorizing land cover types

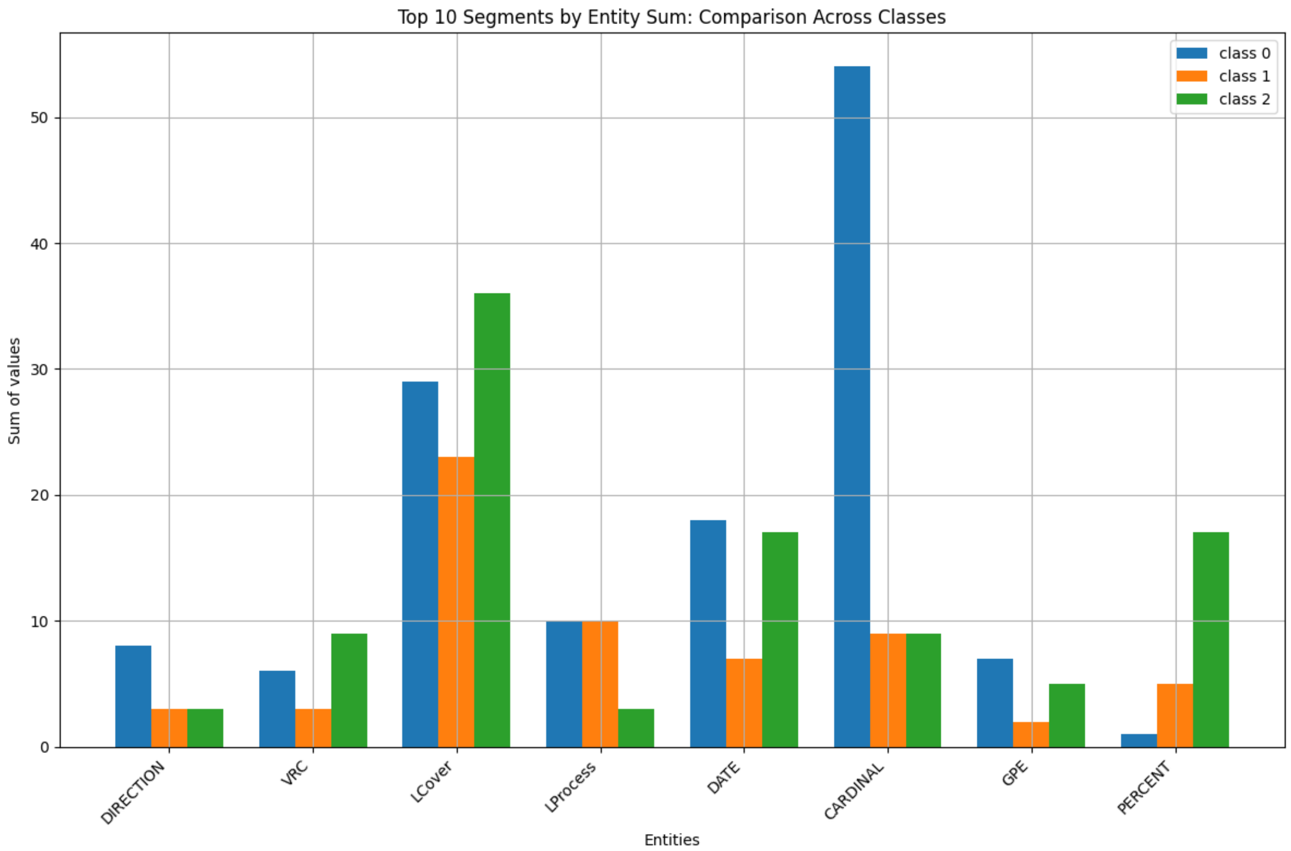


Figure 5: Ranking segments per class

and processes, without paying much attention to spatial orientation or proportional data.

In Class 2, ”LCover” continues to be the most important entity, though it’s a bit less dominant than in Class 1, with ”LProcess” still playing a key role. What’s interesting here is the increased focus on ”Percentage” and ”Direction” entities, indicating a shift towards analyzing the spatial and proportional aspects of land cover changes. While ”Cardinal” and ”Date” entities are less frequent compared to Class 1, the emergence of ”Geopolitical Entities (GPE)” suggests that the analysis in Class 2 is more detailed and geographically specific. This progression reflects a move from broader categorization in Class 1 to more refined, spatially-aware analysis in Class 2.

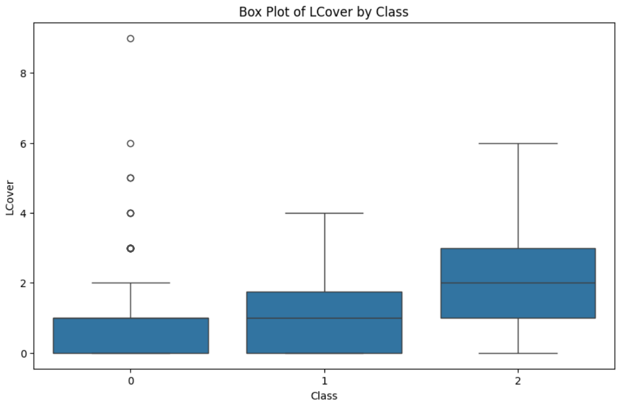


Figure 7: Distribution of Land Cover Vocabulary

The boxplot in Figure 7 shows the distribution of Land Cover entities across the classes. In Class 0, both the median and IQR (Inter Quartile Range)are pretty low, which suggests that specific land cover terms are not mentioned much in this class. However, there are a few outliers where these terms pop up more than usual, even though they are not the main focus of the discussion.

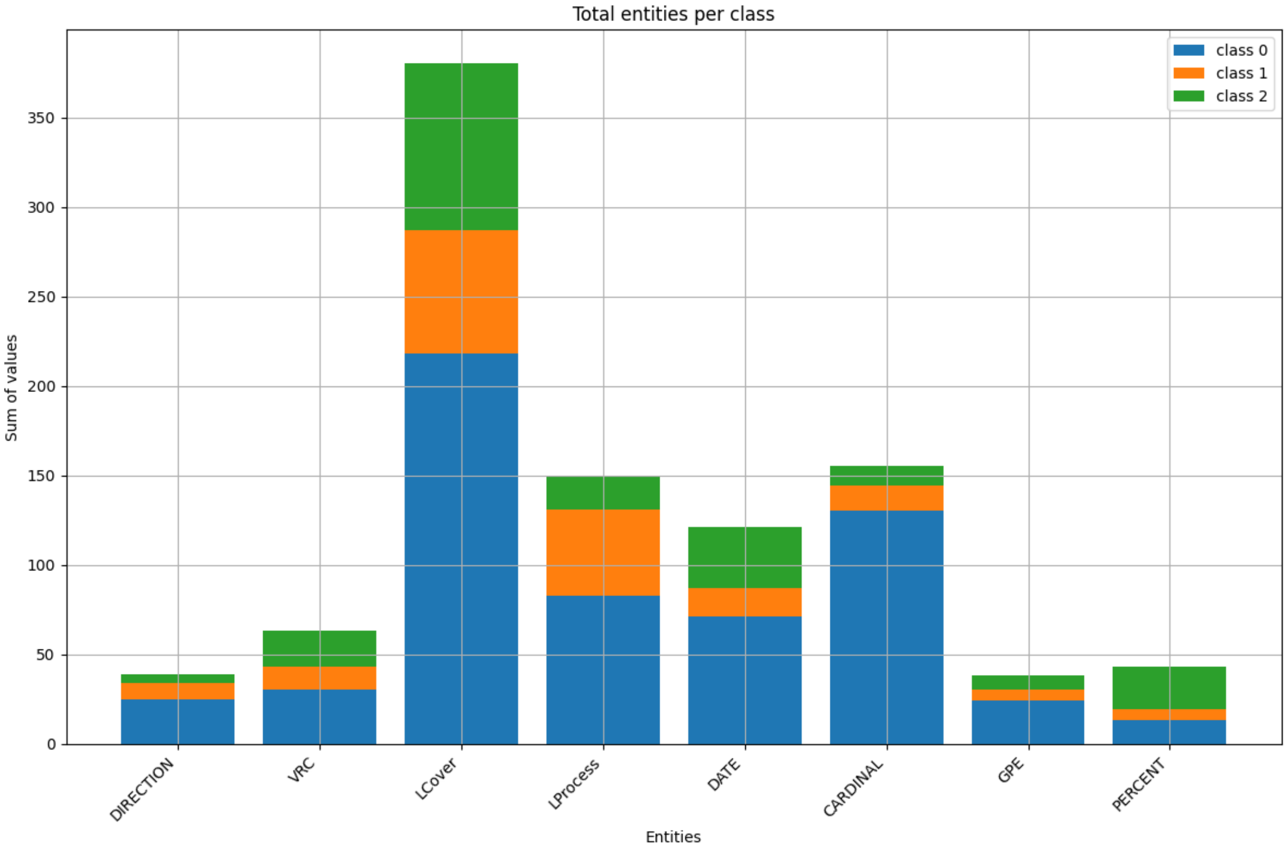


Figure 6: Entity distribution per class

In Class 1, the median is a bit higher, and the IQR is broader compared to Class 0. This tells us that in Class 1, there's a more consistent and deliberate use of land cover vocabulary, helping to make clear distinctions in land cover types. The fewer outliers in this class suggest that the discussions are fairly consistent in how they use land cover terms, with less variation.

In Class 2, both the median and IQR are at their highest, showing that land cover terms are central to the analysis and are mentioned frequently in different ways. The data indicates that these terms are a major part of the discussion in this class. There are not any extreme cases (outliers) where the terms are used a lot more or a lot less than usual, which suggests that the analysis in Class 2 is consistently focused on land cover vocabulary.

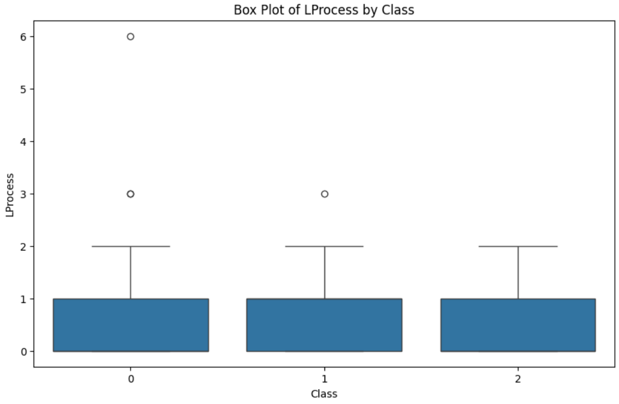


Figure 8: Distribution of Land Cover Change process

The Figure 8 shows that across all three classes, terms related to the land cover change process are mentioned consistently, but not very frequently. This indicates that while these terms are important, they might not be the central focus in any of the classes. The presence of outliers, especially in Classes 0 and 1, shows that sometimes these processes are discussed in more detail, even if it's not the norm. The

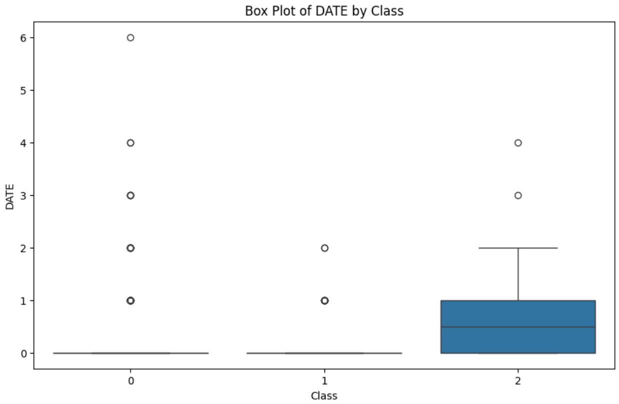


Figure 9: Distribution of Date

importance of temporal context (dates) in Figure 9 increases significantly since Class 0 until Class 2. While Classes 0 and 1 barely mention dates, Class 2 frequently references them, likely to connect land cover changes to specific time periods. This makes sense since Class 2 involves more detailed analysis, likely including time-series analysis to understand the progression or acceleration of land cover and land use changes (LULC). The few outliers suggest that sometimes dates are mentioned even more than usual, perhaps when discussing particularly significant changes over time.

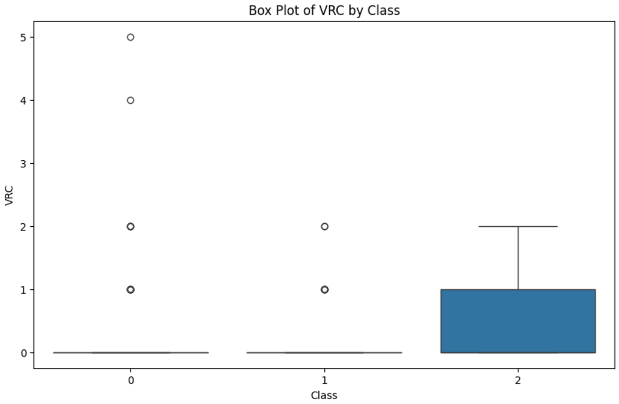


Figure 10: Distribution of VRC (Verb Related to Change)

The Figure 10 shows that for Class 0 and Class 1 the usage of verbs relate to change (VRC) is almost non-existent and the few outliers show that there are occasional instances where these verbs are used, but they are rare. In contrast in class 2 there is an increase, indicating that discussions here frequently involve verbs like ”increase” or ”decrease.” This aligns with the advanced analysis expected in Class 2, where understanding the dynamics and trends of land cover changes is crucial.

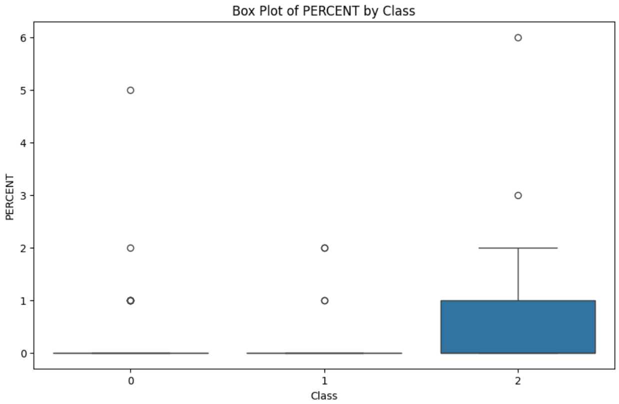


Figure 11: Distribution of Percent

For the entity Percent the Figure 11 Shows that for Class 0 and Class 1 mentions of percentages are almost non-existent. The presence of outliers indicates some discussions where percentages are used, but it’s not a typical focus, in contrast Class 2 shows a significant increase in the use of percentages. The box plot for this class is much more pronounced, indicating that percentages are frequently used to quantify the impact of LULC changes. This makes sense because Class 2 involves more advanced analysis, where understanding the exact proportion of land affected by changes is crucial.

Variable	**Correlation with Class**
CARDINAL	-0.090307
ORDINAL	-0.036146
DIRECTION	0.009042
SURFACE_UNIT	0.033922
GPE	0.044526
LProcess	0.093593
DATE	0.152048
VRC	0.181531
PERCENT	0.254219
LCover	0.267675
class	1.000000

Table 8: Correlation of Variables with Target Class

The Table 8 presents the correlation of various entities with the target class. Positive correlations, such as those with LCover and Percent, indicate a strong association with more informative classes like Class 2, while negative correlations, like with Cardinal, suggest a stronger association with noisier classes like Class 0.

4. Discussion

Particularly with regard to reaching high consensus among the specialists, the manual labeling process exposed some really important difficulties. With just 30% match agreement in the first round, we were particularly reminded of how arbitrary this process can be. But when we collaborated and refined the category definitions, we clearly noticed consistency in the next rounds. remarkably, agreement dropped slightly in the third round, implying that keeping consistency gets more difficult when we add more intricate sections. This emphasizes the requirement of having precise definitions and the need of constant improvement in joint labeling procedures.

To help lower subjectivity and increase consistency going forward, it might be wise to investigate more ordered or even automated labeling systems. Furthermore improving our criteria and adding more tagged segments could help the dataset to be even more dependable.

The models we developed on this dataset did really well, especially in terms of identifying pertinent segments, as seen by the enhanced recall for Class 2. Still, the challenge in precisely grouping the more intricate sections in Class 2 points to space for development. To properly address these difficulties, we might have to take into account more sophisticated methods or deal with a bigger training set.

Our investigation of the several factors across the three classes reveals a clear trend in how profoundly and intricately land cover and land use change (LULC). With the concentration and intensity of varied use scaling up from Class 0 to Class 2, each class seems to symbolize a different stage in the analytical process.

Class 0 is mostly about setting a general backdrop; concepts like LCover, LProcess, and VRC are referenced just sporadically. The sparse usage of dates and percentages also matches this class's function in providing general, background knowledge instead of delving into thorough study. Though these are rare exceptions rather than the norm, the few outliers we find in some variables point to times when these subjects are investigated more thoroughly.

Class 1 makes more methodically use of land cover vocabulary and techniques. This lesson aims to provide a basic knowledge of several land cover forms and their modifications. It's like laying the scene, organizing and outlining the main ideas, but not yet delving into the more intricate, deeper study.

Analysis moves much more finely tuned and sophisticated into Class 2. Terms like LCover, LProcess, DATE, VRC, and PERCENT are clearly more often used now. With an eye on not only what is changing but also when and by how much, this class is all about delving deeply into the mechanics of land cover changes. Class 2 is especially useful for practical purposes since the frequent use of dates and percentages in this class emphasizes their vital part in analyzing the timing and scale of LULC changes.

From broad, general conversations to thorough, data-driven analysis, the evolution from Class 0 to Class 2 generally shows a natural flow in research. This transformation is especially evident in the way Class 2's dates and percentages start to take front stage, therefore highlighting how crucial temporal and quantitative analysis is for comprehending LULC developments. The positive correlations for LCover and PERCENT with the target class confirm how these factors acquire significance as the study gets more advanced.

5. Conclusion

We used a dataset of 2,078 segments for this study. Of those, we labeled and analyzed 324 segments that were about Land Use and Land Cover Change (LULCC) in the Sahel region. We were able to successfully sort these segments into different groups and figure out what makes each group unique by using our method, which included hierarchical extraction, a BERT pre-trained model, and entity extraction.

Our research question asked how well we could classify and analyze textual information about LULCC to better understand what drives these changes. The differences we found between the classes helped us answer this question.

While Class 0 mostly gave general information, it didn't give many specifics about land cover or quantities. A broad outline was given, but not much detail had been provided. In Class 1, the approach was more methodical, utilizing land cover terms and processes to establish a fundamental comprehension of land cover types and variations.

Class 2, however, is where our approach really showed its strength. This class focused on detailed analysis, especially on the timing and scale of land cover changes. Being able to classify and understand Class 2 was crucial because it gave us deeper insights into the factors driving LULCC, particularly the temporal and quantitative aspects that are key to understanding these changes.

For future research, it would be helpful to refine the labeling process, especially by reducing noise from irrelevant numerical data. Expanding the dataset, especially in Class 2, and looking at larger text segments could improve the analysis even further and make our models more accurate. This would lead to a better understanding of LULCC dynamics and help inform land use decisions.

In summary, our findings show that using our approach is very effective for studying LULCC. This work lays a solid foundation for future research, providing valuable insights that are important for addressing the challenges of land use changes in this region.

Acknowledgments

I would like to thank my supervisors Arthur, Sarah, and Maguelonne for giving me the opportunity to work and learn here. I also extend my gratitude to HMS (Hydro Sciences Montpellier), the HEC team, UMR TETIS Lab, and the CECC Project for making this internship possible.

References

Agarwal, S., Atreja, S., and Agarwal, V. Extracting procedural knowledge from technical documents. *arXiv preprint arXiv:2010.10156*, 2020.

Blanchy, G., Albrecht, L., Koestel, J., and Garré, S. Potential of natural language processing for metadata extraction from environmental scientific publications. *Soil*, 9(1): 155–168, 2023.

CECC Project. Cecc project overview, 2024. URL http://www.projet-cecc.org/index.php/en/elementor-3775/. Accessed: 22-August-2024.

Deng, X., Li, Y., Weng, J., and Zhang, J. Feature selection for text classification: A review. *Multimedia Tools and Applications*, 78(3):3797–3816, 2019.

Hardeniya, N. *Natural Language Processing: Python and NLTK*, chapter 2 Text Wrangling and Cleansing. Packt Publishing, Birmingham, 2016. ISBN 978-1-78728-784-6.

Havrlant, L. and Kreinovich, V. A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation). *International Journal of General Systems*, 46(1): 27–36, 2017.

Hugging Face. Datasets documentation, 2024a. URL https://huggingface.co/docs/datasets/index. Accessed: 2024-08-20.

Hugging Face. Roberta model documentation, 2024b. URL https://huggingface.co/docs/transformers/model_doc/roberta. Accessed: 2024-08-20.

JECAM. Jecam - joint experiment for crop assessment and monitoring, 2024. URL https://jecam.org/. Accessed: 2024-08-20.

Kiritchenko, S., De Bruijn, B., Carini, S., Martin, J., and Sim, I. Exact: automatic extraction of clinical trial characteristics from journal publications. *BMC medical informatics and decision making*, 10:1–17, 2010.

Koptelov, M., Holveck, M., Cremilleux, B., Reynaud, J., Roche, M., and Teisseire, M. Towards a (semi-) automatic urban planning rule identification in the french language. In *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–10. IEEE, 2023.

Koroteev, M. V. Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*, 2021.

Kuchling, A. Regular expression howto, 2023. URL https://docs.python.org/3/howto/regex.html. Accessed: 2024-08-14.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lopez, P. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Research and Advanced Technology for Digital Libraries: 13th European Conference, ECDL 2009, Corfu, Greece, September 27-October 2, 2009. Proceedings 13*, pp. 473–474. Springer, 2009.

Malik, U. Python for nlp: Tokenization, stemming, and lemmatization with spacy library, 2023. URL https://stackabuse.com/python-for-nlp-tokenization-stemming-and-lemmatiz Accessed: 2024-08-14.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de La Clergerie, É. V., Seddah, D., and Sagot, B. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*, 2019.

Mishu, S. Z. and Rafiuddin, S. Performance analysis of supervised machine learning algorithms for text classification. In *2016 19th International Conference on Computer and Information Technology (ICCIT)*, pp. 409–413. IEEE, 2016.

National Oceanic and Atmospheric Administration. Land cover and land use, 2024. URL https://oceanservice.noaa.gov/facts/lclu.html. Accessed: 2024-08-20.

Patel, S., Verma, P., and Singh, G. Agricultural growth and land use land cover change in peri-urban india. *Environmental Monitoring and Assessment*, 191(9):600, 2019. doi: 10.1007/s10661-019-7591-4.

PyPDF2. Pypdf2 documentation, 2024. URL https://pypdf2.readthedocs.io/en/3.x/. Accessed: 2024-08-20.

PyTorch. torch.optim.adamw documentation, 2024. URL https://pytorch.org/docs/stable/

generated/torch.optim.AdamW.html. Accessed: 2024-08-20.

Sabharwal, N. and Agrawal, A. Bert algorithms explained. In Sabharwal, N. and Agrawal, A. (eds.), *Hands-On Question Answering Systems with BERT: Applications in Neural Networks and Natural Language Processing*, pp. 65 – 70. Apress L. P, Berkeley, CA, 1st edition, 2021.

scikit-learn. Feature extraction, 2024. URL https://scikit-learn.org/stable/modules/feature_extraction.html. Accessed: 2024-08-20.

Sharnagat, R. Named entity recognition: A literature survey. *Center For Indian Language Technology*, pp. 1–27, 2014.

Singh, S. Natural language processing for information extraction. *arXiv preprint arXiv:1807.02383*, 2018.

Snyder, R. M. An introduction to topic modeling as an unsupervised machine learning way to organize text information. *Association Supporting Computer Users in Education*, 2015.

van der Veen, A. M. Word-level machine translation for bag-of-words text analysis: Cheap, fast, and surprisingly good. *Computational Communication Research*, 5(2):1–, 2023. URL https://doi.org/10.1080/xyz123.

Wang, L. et al. *Support Vector Machines: Theory and Applications*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1st edition, 2005.

Zhao, W.-L., Deng, C.-H., and Ngo, C.-W. k-means: A revisit. *Neurocomputing*, 291:195–206, 2018. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2018.02.072. URL https://www.sciencedirect.com/science/article/pii/S092523121830239X.

Class	Definition	Example
0	The segment is not relevant at all.	The classification of land use was on visual interpretation only.
	The segment’s main topic is other than LULC.	”However agricultural fires should be distinguished from the other fires that are wildland fires and predominantly grass fires (savanna fires)”.
	The segment is too noisy to be understood.	Drivers of change in cropland.’, ’graphic’: {’@coords’: ’8,108.00,66.73,369.83,358.41’, ’@type’: ’bitmap’}}, {’@xmlns’: ’http://www.tei-c.org/ns/1.0’, ’@xml:id’: ’fig_3’, ’head’: ’Fig.
	The segment can contain more than one type of feature but is about another main topic (is not relevant in a thematic or technical way).	Mean maize yields are decreasing in land use scenario L1, L2 and L3 by 16, 24 and 18%, respectively.
1	The segment may be relevant depending on the context.	Other areas have experienced a decrease in cropland despite a population increase.
	The segment makes a general reference to Land Use and Land Cover (LULC), suggesting that the preceding or following sentences provide or will provide clarification.	One of the fundamental factors has been demographic growth and the growth in urban food demands
	The segment contains more than one type of feature and discusses a driver of Land Use and Land Cover (LULC).	One explanation could be the indistinct influence of indirect drivers on LULCC, for example, changes in religious patterns
	There is relevant information, but it is not sufficient to be classified as category 2.	Ineffective land use planning can contribute to an increase in land fragmentation resulting in loss of land for food provision
2	The segment is relevant and could provide more than one feature. It is independent of temporality.	About 26 and 31% of the total area of natural vegetation forest, woodland and tree savanna will be converted to farmland and degraded savanna by 2050 if we assume the dynamics recorded from 1975 to 1990 and 1990-2010, respectively
	The segment’s main topic explicitly addresses Land Use and Land Cover (LULC), either around a specific type of feature or LULC in general.	Agriculture in this region has been dominated over the past three decades by [cotton] production
	We have an explicit understanding that the sentence is about either an agricultural practice or a Land Use and Land Cover (LULC) change, and it is set within a specific spatial (geographic) and temporal context.	After the droughts in the 1970s and 1980s, the observed loss of woody vegetation cover was often considered as irreversible desertification and large parts of the Sahel were designated as degraded land

Table 1: Definition of classes, including examples

Type of Feature	**Definition**	**Example**	**Extraction Method**
SURFACE UNIT	Extracting measurements like kilometers (km) or hectares (ha) allows us to quantify the extent of land cover changes.	Kilometer	Regex + Entity Ruler
Direction	Directions such as North, South, East, and West provide spatial context to LULC changes. Understanding the orientation of land cover transformations can reveal patterns	South	Regex + Entity Ruler
Verb Relate to Change (VRC)	Verbs indicating changes, like ”increase” or ”decrease,” help us understand the dynamics of land cover transitions. These verbs are key to identifying trends over time, offering both qualitative insights (e.g., nature of change) and quantitative data (e.g., frequency of change).	Increasing	Lemma + Regex + Entity Ruler
Land Cover Vocabulary	This entity includes specific terms related to physical land cover, such as ”forest” or ”savanna”. Extracting this vocabulary enables us to categorize different types of land cover and track how they are referenced across different studies in West Africa	Forest	Lemma + Entity Ruler
Land Cover Change Process	Vocabulary from the nomenclature about transition or transformation. Identifying these processes helps us to track the qualitative nature of transitions.	Deforestation	Lemma + Entity Ruler
Date	Dates provide temporal context, allowing us to link land cover changes to specific time periods. This feature offers quantitative data for time-series analysis and helps in understanding the progression or acceleration of LULC changes.	1864	Spacy
Cardinal	Extracting cardinal numbers is crucial for capturing quantitative information related to land cover, such as the amount of land affected by a particular change. This data is essential for measuring and comparing the magnitude of different LULC events.	55	Spacy
GPE	Identifying locations like countries or cities helps in associating land cover changes with specific regions. This geographic context is critical for both qualitative analyses (e.g., regional patterns) and quantitative comparisons across different areas.	Maurirania	Spacy
Ordinal	Ordinal numbers often indicate sequences or ranks in land cover changes. This feature provides both qualitative (e.g., order of changes) and quantitative (e.g., number of changes) information, helping us to understand the hierarchy or progression of LULC processes	first	Spacy
Percentage	Percentages are used to express the proportion of land affected by changes, providing clear quantitative data that is essential for understanding the impact of LULC transitions.	50%	Spacy

Table 2: Entities, Definitions, Examples, and Extraction Methods

Epochs	Split Data	Class	Precision	Recall	F1-Score	Support
10	80/20	0	0.93	0.73	0.82	19
		1 (merged)	0.72	0.92	0.81	14
25	70/30	1	0.90	0.71	0.80	14
		2	0.60	0.85	0.70	7

Table 6: Performance metrics for BERT models across different splits and classes