Mapping Sentiments: A Journey into Low-Resource Luxembourgish Analysis

Nina Hosseini-Kivanani[a,*], **Julien Kühn**[a] **and Christoph Schommer**[a]

[a]Department of Computer Science, University of Luxembourg

Abstract.

Sentiment analysis (SA) plays a vital role in interpreting human opinions across different languages, especially in contexts like social media, product reviews, and other user-generated content. This study focuses on Luxembourgish, a low-resource language critical to Luxembourg's identity, utilizing advanced deep learning models such as BERT, RoBERTa, LuxemBERTand LuxGPT-2. These models were enhanced with transfer learning, active learning strategies, and context-aware embeddings, enabling effective Luxembourgish processing. These models further improved with context-aware embeddings and were able to accurately detect sentiments, categorizing news comments into positive, negative, and neutral sentiments. Our approach highlights the significant role of human-in-the-loop (HITL) methodologies, which refine model accuracy by aligning automated analyses with human judgment. The findings indicate that LuxemBERT, especially when enhanced with the HITL method involving feedback from 500 and 1000 annotated sentences, outperforms other models in both binary (positive vs. negative) and multi-class (positive, neutral, and negative) classification tasks. The HITL approach not only refined model accuracy but also provided substantial improvements in understanding and processing sentiments and sarcasm, often challenging for automated systems. This study establishes the basis for future research to extend these methodologies to other under-resourced languages, promising improvements in Natural Language Processing (NLP) applications across diverse linguistic landscapes.

Keywords: Human-in-the-loop, Low-resource languages, Luxembourgish, Sentiment analysis, Transfer learning

1 Introduction

Sentiment analysis (SA), a key branch of NLP, automates the extraction of opinions, emotions, and attitudes from texts concerning various entities such as products and organizations [28]. Emerging in the early 2000s and also referred to as opinion mining or sentiment mining, this field primarily aims to classify texts as positive, negative, or neutral [6]. Specialized forms detect whether texts are hateful or offensive [12, 42], and address social issues such as racism using data from social media [17]. Given the significant influence of social media on political elections and marketing, SA has become critically important for businesses and governments [2].

In the area of SA, researchers have explored a range of methods, including both supervised and unsupervised techniques, all showing promising results (e.g., [23]). Early studies indicate that unsupervised models, which use sentiment dictionaries, grammatical analysis, and sentence structure patterns with manually created rules, can perform just as well as traditional supervised methods, such as Support Vector Machines (SVMs) and Naïve Bayes classifiers [34]. This exploration sets the stage for addressing the impact of data scarcity in low-resource languages.

The significant challenges of applying advanced SA techniques become evident in the context of low-resource languages like Luxembourgish. Despite the superior performance of DL models over traditional methods, their dependency on extensive labeled datasets remains a major hurdle, given the high costs and time required for data annotation [18]. To mitigate these challenges, methods such as transfer learning and active learning have been introduced to offer viable solutions (e.g., [1]).

Consider Luxembourg, a trilingual nation of over 590,000 residents, home to the largest population of Luxembourgish speakers. Recognized as the national language in 1984, Luxembourgish is integral to the nation's identity and essential for communication within the country. Originally a Central Franconian dialect, Luxembourgish has evolved into an independent language, becoming essential for all forms of communication within the country. In contexts where all participants are fluent, switching to French or German is generally avoided [14]. The term "low-resource language" refers to languages that lack substantial annotated or digital data [32, 4]. In the NLP field, there has been a significant increase in methods targeting these low-resource languages, broadening the applicability of language models to a more diverse set of languages.

Traditional static word embeddings do not capture contextual variations that influence meaning, which is essential for accurately understanding and analyzing language. To address this limitation, context-aware embeddings like BERT (Bidirectional Encoder Representations from Transformers) [8], Robustly Optimized BERT Approach (RoBERTa) [29], and GPT have been developed. These models offer dynamic, contextual representations that adapt based on the surrounding text, thus providing a more precise reflection of subtle sentiment expressions within texts [24]. By understanding the specific context in which words are used, these advanced models significantly enhance the accuracy of SA, making them indispensable in extracting true sentiment from complex language constructions, such as irony or sarcasm, commonly found in social media and other digital communications. Further, human-in-the-loop (HITL) is particularly valuable as it allows systems to adjust to real-world variables and user-specific needs that may not be fully anticipated at the time of a model's initial training. For instance, in SA, HITL can be instrumental in refining the understanding and classification of language used in different contexts, such as irony or cultural-specific expressions that automated

* Corresponding Author. Email: nina.hosseinikivanani@uni.lu.

systems might misinterpret [44]. This approach not only enhances the performance and trustworthiness of AI systems but also enables them to become more aligned with human values and ethics, a critical consideration as AI becomes more pervasive in everyday life [38].

1.1 Contribution

The main contribution of this paper is the application of state-of-the-art frameworks and the integration of human-in-the-loop components [44] for SA of Luxembourgish. We specifically focus on developing an SA model that classifies Luxembourgish news comments into positive, negative, and neutral sentiment classes at the sentence level. The structure of the paper is as follows:

- The 'Related Works' section provides an overview of the research in this field, highlighting previous approaches to SA in multilingual and low-resource contexts.
- 'Materials and Methodology' describes the proposed models in detail.
  - The 'Dataset' section outlines the description and sourcing of the datasets used.
- Comprehensive discussions of the findings are presented in the 'Evaluation' section, including comparisons with existing models and discussion on the effectiveness of different methodologies.
- The paper concludes with final remarks in the 'Conclusion and Future Work' section, summarizing the implications and potential future directions for SA research in low-resource languages.

2 Related works

SA is a crucial aspect of understanding human opinions across various languages, particularly in the context of social media, product reviews, and other user-generated content. DL methods have shown significant promise in improving SA, especially for low-resourced languages like Luxembourgish. Recent advancements in DL architectures, particularly Transformer-based language models, have led to breakthroughs in SA tasks. These models use pre-trained knowledge to enhance performance on downstream tasks, a method that is particularly effective in contexts where annotated data is scarce [20].

In the area of SA for low-resource languages, the challenges and potential solutions are diverse and multifaceted. For example, translating datasets from resource-rich languages to those with fewer resources, such as Urdu, can often change the meaning of sentiment and cause performance degradation due to polarity shift [13]. This shift can make sentiment classification systems work poorly. This challenge is further compounded in domain adaptation scenarios, such as with Danish, where dramatic performance drops occur when switching domains [9].

Several approaches have proven effective in dealing with these issues. Using methods like transfer learning [22], unsupervised learning, semi-supervised learning, and active learning can significantly improve SA for these languages. Sentiment classification approaches are broadly categorized into supervised [36], semi-supervised [16], and unsupervised [19]. Although most studies use supervised methods, a major challenge remains the lack of well-organized datasets.

Traditionally, SA research has primarily focused on well-resource languages such as English, German, and Chinese. However, the focus has shifted towards investigating SA in low-resource languages in recent years, promoting greater linguistic inclusivity in NLP tools [11, 27]. A study by Pang et al. [36] demonstrated that ML techniques for sentiment classification significantly surpass human-generated benchmarks. They applied three ML models—Naïve Bayes, Support Vector Machine, and Maximum Entropy—to a dataset of movie reviews. Using a 3-fold cross-validation method, they compared the effects of feature presence versus feature frequency. They found that feature presence, which indicates the binary occurrence of a feature, was more effective than feature frequency, which measures how often a feature appears. Of the three classifiers, SVM yielded the best performance.

The application of ML techniques to comments in Bangla from the entertainment sector has shown promising results, with accuracy rates exceeding 75% for sentiment classification [39]. This indicates that even low-resource languages can achieve significant performance in SA tasks with the right methodologies. Similarly, adaptive pretraining and careful selection of source language have been shown to improve SA for African languages, leading to improvements of over 10% F1 score points [40].

The challenges of SA in low-resource languages are substantial but can be overcome, as demonstrated by various studies proposing cutting-edge strategies to enhance precision (e.g., [15]). A systematic review of multilingual SA techniques reveals a growing interest in developing models for such languages, with DL methods particularly recommended [31]. In detail, DL models, particularly those that incorporate attention mechanisms, have been successfully applied to SA in Albanian social media comments, achieving an F1 score of 72.09% [21]. Furthermore, transformer-based models have demonstrated their potential to improve SA for low-resource African languages such as Nigerian Pidgin and Yoruba, achieving top rankings in SemEval-2023 Task 12 [20].

The application of Pre-trained language models like Bidirectional Encoder Representations from Transformers (BERT) and multilingual BERT (mBERT) has also been noteworthy (e.g., [43]). These models can enhance SA tasks without extensive fine-tuning, thus reducing training time and resource consumption while maintaining or even improving accuracy [25]. Transfer learning techniques using pre-trained multilingual models often outperform language-specific models in low-resource settings, showing further improvements after fine-tuning even with a small number of samples [35].

Fawzy et al. [10] address the challenge of SA in Arabic, a low-resource language with diverse dialects and complex linguistic features. They propose an approach that combines a BERT model with a Convolutional Neural Network (CNN) to improve SA accuracy. The model, BERT-CNN, fine-tunes only the last four layers of a pre-trained BERT model, reducing computational requirements while leveraging the CNN as a classification head for enhanced feature extraction. Tested on three Arabic Twitter datasets, the BERT-CNN model not only outperforms existing state-of-the-art models but does so with 50% smaller batch sizes, fewer training layers, and approximately 20% fewer epochs on the datasets.

Human-in-the-loop (HITL) approaches have also shown promise. Human-in-the-loop linguistic Expressions with Deep Learning (HEIDL), a prototype HITL machine learning system, enables higher-level interaction between humans and machines, improving productivity and generalizing models to unseen data [37]. HITL NLP frameworks integrate human feedback to improve NLP models, with promising future studies in integrating human feedback in the development loop [41]. HITL can achieve comparable or better performance than unsupervised domain adaptation (UDA) in person re-identification scenarios when unlabeled target data is infeasible [7].

While manually annotated datasets are essential for training and evaluating NLP models, recent studies have highlighted that even

widely used benchmark datasets often contain many incorrect annotations. This reveals additional challenges in SA for low-resource languages, where data scarcity is compounded by quality concerns [26].

Building on these findings, recent efforts have extended these techniques to low-resource languages, where traditional feature extraction methods face challenges due to sparse data availability. Innovations in transfer learning and unsupervised learning methods are beginning to show promise in overcoming these barriers, enabling more effective SA across a broader spectrum of languages [5]. These developments underline the growing necessity and potential for applying advanced ML techniques to enhance linguistic inclusivity in NLP applications.

In summary, while SA for low-resource languages presents unique challenges, primarily due to the scarcity of sufficient annotated data and linguistic resources, research indicates that these can be effectively addressed with innovative techniques such as adaptive pretraining, DL, cross-lingual techniques, and data augmentation strategies. Emerging methods like mBERT and adversarial learning are proving effective in enhancing the precision and generalizability of SA models for these languages.

3 Materials and Methodolgy

3.1 Dataset

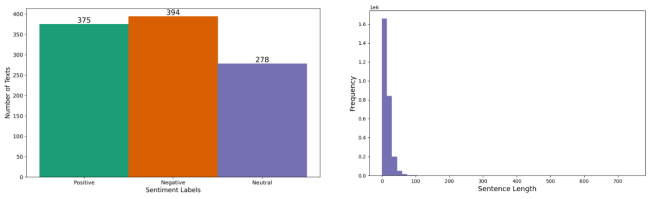

Figure 1: Sentiment Distribution Across the Dataset and Comment Lengths: The leftmost bar graph shows the distribution of sentiments, while the rightmost plot illustrates the distribution of sentence lengths.

The analysis in this study uses a corpus comprising user comments on news articles collected between 2009 and 2018 [1], with the longest sentence consisting of 744 words. The dataset, generously provided by RTL Luxembourg, is invaluable for SA as it encompasses responses to a wide range of topics authored by Luxembourgish speakers from diverse backgrounds and with varying personal histories. Figure 1 displays the number of annotations grouped by sentiment label (1047 comments in total), showing that *negative* comments were more frequent than *positive* or *neurtal* comments.

```
<sentence id="3bs">
  <w id="11" pos="" sen="3">pierre</w>
  <w id="12" pos="" sen="3">je</w>
  <w id="13" pos="" sen="3">suis</w>
  <w id="14" pos="" sen="3">surtout</w>
  <w id="15" pos="" sen="3">persuadé</w>
  <w id="16" pos="" sen="3">que</w>
  <w id="17" pos="" sen="3">L</w>
  <c id="18" pos="" sen="3">.</c>
</sentence>
```

Figure 2: Example of XML sentence structure from the dataset. *"pos"* refers to POS-tagging, which is not provided in this part of the dataset.

The data was provided in a simple XML file (see Figure 2), with each file containing a subset of the sentences. Prior to training the model, the XML file required preprocessing steps on user comments. This process involved transforming the XML data into a Pandas DataFrame [33], which is well-suited for handling such data operations in Python.

The preprocessing steps included:

- Parsing the XML Structure: Identifying and extracting key elements and attributes within the XML structure that contain the relevant information, such as sentence IDs and words.
- Cleaning the Data: Removing any extraneous tags and normalizing text to ensure consistency in sentiment classification.
- Annotating Sentiments: Ensuring the sentiment labels provided by annotators were correctly categorized into negative, neutral, and positive sentiments.

The sentiment labels were initially provided by human annotators and manually categorized. These annotations are critical as they form the basis for training and evaluating the SA models (see Figure 1).

3.2 Models

BERT: BERT is a transformative model in the field of NLP. Developed by Google, BERT has revolutionized how machines understand human language. It is based on the transformer architecture, which relies on attention mechanisms rather than sequence-aligned recurrent processing. This design allows for a more flexible interpretation of sentence structures.

The primary innovation of BERT is its approach to pre-training on a large corpus using only unlabeled data, followed by fine-tuning on smaller specific tasks. Unlike previous models that processed text in a single direction, either from left to right or right to left, BERT processes text bi-directionally. This bidirectional training is fundamental to its success, as it enables the model to capture the context of a word based by considering all surrounding text, both preceding and following. This capability allows BERT to understand the meaning of words within their specific sentence structures, which is a significant advance over traditional methods that often depend on labor-intensive feature engineering.

BERT's versatility is demonstrated in its ability to be fine-tuned with just an additional output layer to produce state-of-the-art results for a range of tasks, including question answering, language inference, and SA. In SA, BERT's ability to analyze the complete context of words makes it exceptionally effective in accurately classifying sentiments. This is particularly useful not just at the sentence level but also for more detailed aspect-level analysis, where the sentiments regarding specific aspects of a product or service are assessed.

The "bert-base-multilingual-cased" model is a variation of BERT designed to handle multiple languages. It retains BERT's powerful bidirectional context analysis while supporting text in various languages. This multilingual capability is particularly valuable for SA in multilingual settings, where it can interpret sentiments across different languages without needing separate models for each language. This feature extends BERT's versatility, allowing for consistent performance and ease of use in global applications.

BERT has consistently outperformed earlier models that relied on embeddings generated from simpler neural networks. Its ability to integrate and understand target-specific information within a text further enhances its performance, enabling more accurate sentiment discernment in complex scenarios. Research indicates that BERT's deep contextual understanding significantly improves performance across various NLP benchmarks, making it an essential tool for researchers and practitioners working with language data [8].

[1] www.rtl.lu.

RoBERTa: RoBERTa, or Robustly Optimized BERT Pretraining Approach, builds upon the foundational concepts of BERT by incorporating several key modifications that significantly improve its effectiveness. Unlike BERT, which is trained for a fixed amount of time on a set dataset size, RoBERTa benefits from training on larger datasets and for longer periods. This extended training allows RoBERTa to develop a more profound understanding of language details and complexities.

A critical enhancement in RoBERTa is the dynamic adjustment of the masking pattern during the pre-training phase. Whereas the masked language model (MLM) task in BERT randomly masks 15% of the tokens once at the beginning of training, which remains the same for every training epoch. In contrast, RoBERTa recalculates and randomizes the masks throughout the training process. This dynamic masking prevents the model from merely memorizing the masked positions, instead fostering more robust and generalizable language representations.

RoBERTa also adopts byte-level Byte-Pair Encoding (BPE) as its tokenization method, enhancing its ability to handle a more compact and efficient vocabulary. This approach is particularly beneficial for processing languages with rich morphology or those that use compound words, as it can decompose words into more frequently occurring subwords or bytes. By simplifying the vocabulary size and complexity, RoBERTa can process text data more quickly and with fewer resources than BERT.

Moreover, RoBERTa removes the next-sentence prediction (NSP) task, which BERT originally used. This decision is based on evidence that NSP does not significantly contribute to model performance on downstream tasks. Instead, RoBERTa focuses on optimizing the MLM objective, which has been shown to improve outcomes directly across a wide range of NLP benchmarks. This focused approach particularly benefits tasks requiring deep contextual understanding, such as SA, question answering, and natural language inference.

RoBERTa's performance demonstrates the importance of iterative improvements and optimizations in model pre-training strategies. It has outperformed BERT model and its other variants across various NLP benchmarks, establishing RoBERTa as one of the most potent models for tackling complex language processing challenges [29].

XLM-RoBERTa extends RoBERTa's capabilities to a multilingual setting. XLM-RoBERTa is designed to handle multiple languages simultaneously while supporting cross-lingual tasks. This model is especially valuable for SA in multilingual environments, where it can interpret and classify sentiments across different languages without the need for separate models for each language. XLM-RoBERTa retains RoBERTa's dynamic masking and BPE tokenization advantages, ensuring robust performance across diverse linguistic contexts.

LuxemBERT: LuxemBERT [30] is a state-of-the-art transformer-based language model specifically designed for the Luxembourgish language. It builds on the architecture of BERT (Bidirectional Encoder Representations from Transformers), which is renowned for its powerful bidirectional context understanding. This capability is particularly advantageous for addressing the challenges associated with low-resource languages like Luxembourgish.

LuxemBERT adopts BERT's robust framework, which features multiple layers of transformer encoders. These encoders use self-attention mechanisms to process input text sequences, allowing the model to discern complex dependencies and contextual relationships within the text. The bidirectional nature of LuxemBERT allows it to consider the context of each word from both preceding and following texts, thereby enhancing the accuracy of language representations.

A key strength of LuxemBERT is its adaptation to the Luxembourgish context through pre-training on language-specific corpora. Additionally, it uses transfer learning approaches, fine-tuned to specific downstream tasks relevant to Luxembourgish, such as SA, named entity recognition (NER), and text classification. This dual approach of pre-training and fine-tuning ensures that LuxemBERT can effectively generalize from limited data, maintaining high performance across diverse NLP applications.

LuxGPT-2[2]: LuxGPT-2 is an advanced transformer-based language model, developed using the GPT-2 (Generative Pre-trained Transformer 2) architecture, and specifically adapted for text generation in the Luxembourgish. Unlike traditional models that process text linearly, Lux-GPT-2 understands and generates text bi-directionally, where it predicts the next word in a sequence considering the entire context provided by all preceding words. This bidirectional approach is particularly effective in handling the subtleties of language that are critical for realistic and coherent text generation.

LuxGPT-2 was pre-trained on a substantial and varied corpus consisting of 711 MB of Luxembourgish text, which includes diverse sources such as RTL.lu news articles, parliamentary speeches, Wikipedia entries, and various web crawls. This extensive training set provides a rich linguistic foundation, allowing LuxGPT-2 to learn and reproduce the unique syntactic and semantic patterns of the Luxembourgish language.

The model's training process involved transfer learning techniques, where the model was initially conditioned on an English-based model to establish a broad understanding of linguistic structures. It then received further training (fine-tuning) to adapt these structures to Luxembourgish specifics. This phase included gradual layer freezing, a technique where lower layers of the model are incrementally locked as they stabilize, allowing the training focus to shift toward the upper layers that are responsible for capturing more complex and abstract language features.

Following its pre-training, LuxGPT-2 demonstrates remarkable versatility by being adaptable for fine-tuning on smaller, task-specific datasets. This flexibility enables LuxGPT-2 to excel in various NLP tasks, including SA, text summarization, and question-answering. Its ability to generate contextually rich, grammatically correct, and semantically detailed Luxembourgish text positions it as an essential resource for applications demanding high-quality Luxembourgish text output.

HITL: The concept of human-in-the-loop (HITL) has become increasingly relevant in various domains, particularly in fields like ML and artificial intelligence (AI). HITL methodologies are designed to integrate human judgment into the loop of automated systems, facilitating a dynamic interaction where humans provide real-time corrections and feedback.

HITL is instrumental in enhancing the reliability of AI systems. By incorporating human judgment, these systems can perform complex decision-making tasks with greater precision. Human intervention helps to refine AI responses by correcting errors that ML models may not identify on their own due to limitations in training data or inherent biases in their algorithms.

Involving human judgment in AI systems helps mitigate the risk associated with biases that are often present in the training data. Humans can identify and correct biased AI decisions in real-time, enhancing the fairness and impartiality of automated decisions. This real-time corrective feedback not only improves the model's current accuracy in the short term but also influences its learning trajectory, promoting better generalization and reliability in subsequent applications.

[2] https://huggingface.co/

Table 1: Performance metrics (accuracy, precision, recall) across BERT, RoBETa, LuxemBERT, and LuxGPT-2 models across two classification scenarios: Binary classification with positive and negative labels, and Multi-class classification with positive, neutral, and negative labels).

Models	Accuracy	Precision	Recall
BERT - Binary classification	57	60	62
RoBERTa - Binary classification	59	66	65
LuxemBERT - Binary classification	55	62	66
LuxGPT-2 - Binary classification	49	59	64
BERT - Multi-class classification	60	69	73
RoBERTa - Multi-class classification	64	71	72
LuxemBERT - Multi classification	67	73	72
LuxGPT-2 - Multi classification	35	49	52
HITL-multi (LuxemBERT-after training 500 sentences)	70	77	73
HITL-multi (LuxemBERT-after training 1000 sentences)	**75**	**78**	**77**

3.3 Training

To train our SA model, we used various Python libraries that support data manipulation, visualization, and ML. Key libraries included:

- NumPy: For numerical operations.
- Pandas: For managing data frames, allowing for seamless data manipulation and preparation.
- Matplotlib and Seaborn: For creating informative visualizations to analyze data trends and model performance.
- regex: For handling regular expressions.
- xml.etree.ElementTree: For parsing XML files.

Our model was trained using the TensorFlow framework. The dataset was divided into training (70%), validation (10%), and test (20%) sets through *stratified sampling* function from Scikit-learn [3]. This approach ensured that the distribution of data across each set matched that found in the original dataset, which is particularly important in studies with imbalanced classes. To further address the class imbalance, we incorporated class weights into the models. By assigning higher weights to the minority class, we ensured that the model penalized misclassifications of the minority class more heavily, thereby improving the overall balance and performance of the model across all classes.

3.4 Evaluation Metrics and their Implications

To evaluate the effectiveness of the classification techniques, specific metrics such as accuracy, precision, and recall were used:

- **Terminology for All Metrics:**
  - **TP (True Positives)** is the number of positive instances that the model correctly identifies.
  - **TN (True Negatives)** is the number of negative instances that the model correctly identifies.
  - **FP (False Positives)** are instances that the model incorrectly predicted as positive.
  - **FN (False Negatives)** are positive instances that the model fails to identify.
- **Accuracy** is a commonly used metric for evaluating classification models. It signifies the ratio of correctly classified instances to the total number of instances within the dataset. Specifically in text classification, accuracy measures the proportion of texts that are accurately categorized. A higher accuracy value indicates a more precise model. The formula for accuracy is expressed as:

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

- **Precision** assesses the proportion of true positive predictions (correctly classified instances) among all positive predictions made by the model. It is a crucial metric that measures the model's ability to minimize false positives. It is determined by dividing the number of true positives by the total of true positives and false positives. A higher precision value suggests that the model makes fewer false positive errors, which is particularly valuable in scenarios where the cost of a false positive is high.

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

- **Recall**, also referred to as sensitivity or the true positive rate, measures the proportion of true positives that are correctly identified out of the total sum of true positives and false negatives. Essentially, recall quantifies how effectively the model identifies all positive samples within the dataset. A higher recall value indicates a smaller number of false negatives. This metric evaluates the model's capability to detect all relevant instances without missing positive ones. In essence, recall measures how well the model captures all positive samples in the dataset.

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

These metrics not only provide a comprehensive overview of the model's accuracy but also help understand its performance in terms of specificity and sensitivity, guiding further refinements.

4 Results and Discussion

Table 1 presents the performance metrics—accuracy, precision, and recall—of selected language models in binary (negative vs. positive) and multi-class (positive, neutral, and negative) classification scenarios. The evaluated models include BERT, RoBERTa, LuxemBERT, and LuxGPT-2, along with outcomes from a HITL setup involving iterative feedback loops on 500 and 1000 sentences, specifically designed to refine and enhance LuxemBERT's learning algorithms. This

[3] https://scikit-learn.org/stable/

setup enabled targeted improvements in the model's understanding of nuanced sentiments, particularly in complex linguistic contexts.

In binary classification tasks, RoBERTa outperforms BERT, achieving higher accuracy (59% vs. 57%), precision (66% vs. 60%), and recall (65% vs. 62%). This reflects RoBERTa's ability to efficiently differentiate between positive and negative classes. LuxemBERT, while showing comparable recall, experiences a slight drop in precision and accuracy, indicating sensitivity due to its architectural differences. LuxGPT-2, with significantly lower scores in accuracy (49%) and precision (59%), suggests potential configuration deficiencies for binary tasks.

The multi-class classification results reveal that LuxemBERT excels, particularly in a multilingual setting, achieving the highest metrics across accuracy (67%), precision (73%), and recall (72%). This indicates its ability to handle more complex label distinctions. RoBERTa maintains robust performance, although it slightly lags behind LuxemBERT, highlighting the subtle differences in its processing capabilities. In contrast, LuxGPT-2 shows considerable underperformance in this scenario, indicating that enhancements may be necessary to improve its adaptability to multi-class contexts.

The Iterative training process of LuxemBERT with human annotations under the HITL methodology demonstrates significant improvements, with accuracy increasing to 70% and 75% with 500 and then 1000 sentences, respectively. This gradual improvement underscores the value of integrating human feedback into the training process, enhancing both the accuracy and reliability of the model. Following the incorporation of human feedback, significant performance improvements were observed across the models. Detailed metrics such as accuracy, precision, and recall for each model iteration are summarized in Table 1.

As detailed in Table 1, the LuxemBERT model showed superior performance in handling complex expressions after training with human-annotated data. One noteworthy example of HITL's impact is its correction of the misinterpretation of the Luxembourgish expression "*Dat war awer eppes!*". Thanks to the expert annotations, the model could adjust its training algorithms to understand that, despite the presence of "*awer*" (but), the expression conveyed a positive sentiment. This correction was a direct result of the iterative training and feedback process unique to our HITL methodology. Compared to a baseline LuxemBERT model trained on the same initial dataset but without human feedback, our HITL-enhanced LuxemBERT model showed 8% improvement in detecting complex sentiments such as irony, which are often misunderstood by traditional SA tools. Human annotators played a crucial role, particularly in correcting sentiments related to cultural idioms like "*Ech si gréng hannert den Oueren.*" Initially labeled as neutral by the model, annotators clarified that this typically expresses a negative sentiment about one's lack of experience. Incorporating these corrections reduced the model's error rate in similar contexts. The HITL approach led to substantial improvements in LuxGPT2-s ability to discern sarcasm, a sentiment often misinterpreted by automated systems without localized training inputs.

One of the primary challenges was the scalability of human annotations, as recruiting enough native speakers trained in linguists was time-consuming and costly. Despite efforts to mitigate bias, the dominance of certain dialects within the annotated data occasionally skewed the model's performance on regional variations of Luxembourgish. Our findings emphasize the role of HITL methodologies in enhancing the performance of SA models for low-resource languages, with potential applications in content moderation, customer service bots, and social media analytics for improved accuracy and cultural relevance [3].

Integrating HITL-enhanced models with multilingual platforms like Google Translate or customer relationship management (CRM) systems could further enhance adaptability and accuracy across different languages and dialects. This approach not only improves technological inclusivity for low-resource language speakers but also underscores the need for ethical considerations in handling sensitive sentiment data.

The broader impact of our research extends to enhancing the digital inclusion of minority language speakers by developing technology that accurately understands and processes their language. However, as we collect and use sensitive sentiment data, it is imperative to adhere to stringent data privacy laws and ethical guidelines to protect individual privacy and prevent biases that could inadvertently arise from data misinterpretation. Despite BERT's good cross-lingual performance on high-resource languages, it struggles with low-resource languages, indicating a need for more efficient pretraining techniques or more data [43].

5 Limitations and Challenges

A limitation of our study is the dependency on a sufficient number of trained human annotators, which poses scalability challenges. This dependency could limit the application of our methods in larger-scale environments or in cases where such resources are scarce. Furthermore, the iterative training process, while effective, requires substantial computational resources, which may not be feasible in all application scenarios. Future research should, therefore, aim to optimize these processes to balance accuracy with operational efficiency better.

6 Conclusion and Future Work

This study introduces a method that can be useful for low-resource languages by adopting an existing model to a new context. Training a new model for a specific language can be expensive and time-consuming, taking days or even weeks, depending on the available computing power.

In our research, we focused on Luxembourgish, a language considered resource-scarce, comparing various BERT-based models alongside a HITL strategy. To address the scarcity of data, we implemented HITL strategy, which demonstrated improvements in SA of news comments. While not all models showed statistical significance, the HITL approach on LuxemBERT model consistently outperforms the BERT-based alternatives.

Future work should consider integrating semi-supervised learning techniques to better use unlabeled data. This approach could potentially expand the model's training dataset without requiring extensive human annotation. Additionally, applying the HITL methodology to other under-resourced languages could provide insights into the generalizability of this method across diverse linguistic landscapes.

Acknowlegments

We would like to thank Christoph Purschke (**Faculty of Humanities, Education and Social Sciences**, University of Luxembourg) for sharing the data and Aria Nourbakhsh (**Faculty of Science, Technology, and Medicine**, University of Luxembourg) for his invaluable assistance with brainstorming for the paper.

References

[1] S. A. A. Asli, B. Sabeti, Z. Majdabadi, P. Golazizian, R. Fahmi, and O. Momenzadeh. Optimizing annotation effort using active learning strategies: A sentiment analysis case study in persian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2855–2861, 2020.

[2] M. Birjali, M. Kasri, and A. Beni-Hssane. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134, 2021.

[3] A. Chaudhary, J. Xie, Z. Sheikh, G. Neubig, and J. G. Carbonell. A little annotation does a lot of good: A study in bootstrapping low-resource named entity recognizers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5164–5174, 2019.

[4] C. Cieri, M. Maxwell, S. Strassel, and J. Tracey. Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549, 2016.

[5] F. Daneshfar. Enhancing low-resource sentiment analysis: A transfer learning approach. *Passer Journal of Basic and Applied Sciences*, 6(2): 265–274, 2024.

[6] N. C. Dang, M. N. Moreno-García, and F. De la Prieta. Sentiment analysis based on deep learning: A comparative study. *Electronics*, 9(3): 483, 2020.

[7] R. Delussu, L. Putzu, G. Fumera, and F. Roli. Online domain adaptation for person re-identification with a human in the loop. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3829–3836. IEEE, 2021.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] J. Elming, B. Plank, and D. Hovy. Robust cross-domain sentiment analysis for low-resource languages. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–7, 2014.

[10] M. Fawzy, M. W. Fakhr, and M. A. Rizka. Sentiment analysis for arabic low resource data using bert-cnn. In *2022 20th International Conference on Language Engineering (ESOLEC)*, volume 20, pages 24–26. IEEE, 2022.

[11] R. R. R. Gangula and R. Mamidi. Resource creation towards automated sentiment analysis in telugu (a low resource language) and integrating multiple domain sources to enhance sentiment prediction. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.

[12] A. Gaydhani, V. Doma, S. Kendre, and L. Bhagwat. Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. *arXiv preprint arXiv:1809.08651*, 2018.

[13] A. Ghafoor, A. S. Imran, S. M. Daudpota, Z. Kastrati, . Abdullah, R. Batra, and M. A. Wani. The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing. *IEEE Access*, 9:124478–124490, 2021. doi: 10.1109/ACCESS.2021.3110285.

[14] P. Gilles and C. Moulin. Luxembourgish. *Germanic standardizations: Past to present*, pages 303–329, 2003.

[15] V. Girija, T. Sudha, and R. Cheriyan. Analysis of sentiments in low resource languages: Challenges and solutions. In *2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE)*, pages 1–6. IEEE, 2023.

[16] A. B. Goldberg and X. Zhu. Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In *Proceedings of TextGraphs: The first workshop on graph based methods for natural language processing*, pages 45–52, 2006.

[17] E. Greevy and A. F. Smeaton. Classifying racist texts using a support vector machine. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 468–469, 2004.

[18] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, 2021.

[19] X. Hu, J. Tang, H. Gao, and H. Liu. Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on World Wide Web*, pages 607–618, 2013.

[20] N. Hughes, K. Baker, A. Singh, A. Singh, T. Dauda, and S. Bhattacharya. Bhattacharya_lab at semeval-2023 task 12: A transformer-based language model for sentiment classification for low resource african languages: Nigerian pidgin and yoruba. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1502–1507, 2023.

[21] Z. Kastrati, L. Ahmedi, A. Kurti, F. Kadriu, D. Murtezaj, and F. Gashi. A deep learning sentiment analyser for social media comments in low-resource languages. *Electronics*, 10(10):1133, 2021.

[22] Z. Kastrati, L. Ahmedi, A. Kurti, F. Kadriu, D. Murtezaj, and F. Gashi. A deep learning sentiment analyser for social media comments in low-resource languages. *Electronics*, 10(10):1133, 2021.

[23] G. Kaur and A. Sharma. A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis. *Journal of big data*, 10(1):5, 2023.

[24] S. Khan and T. Yairi. A review on the application of deep learning in system health management. *Mechanical Systems and Signal Processing*, 107:241–265, 2018.

[25] Y. Kit and M. M. Mokji. Sentiment analysis using pre-trained language model with no fine-tuning and less resource. *IEEE Access*, 10:107056–107065, 2022.

[26] M. Laurer, W. Van Atteveldt, A. Casas, and K. Welbers. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100, 2024.

[27] T. A. Le, D. Moeljadi, Y. Miura, and T. Ohkuma. Sentiment analysis for low resource languages: A study on informal indonesian tweets. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 123–131, 2016.

[28] B. Liu. *Sentiment analysis and opinion mining*. Springer Nature, 2022.

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

[30] C. Lothritz, B. Lebichot, K. Allix, L. Veiber, T. F. D. A. Bissyande, J. Klein, A. Boytsov, A. Goujon, and C. Lefebvre. Luxembert: Simple and practical data augmentation in language model pre-training for luxembourgish. In *Proceedings of the Language Resources and Evaluation Conference, 2022*, pages 5080–5089, 2022.

[31] K. R. Mabokela, T. Celik, and M. Raborife. Multilingual sentiment analysis for under-resourced languages: a systematic review of the landscape. *IEEE Access*, 11:15996–16020, 2022.

[32] A. Magueresse, V. Carles, and E. Heetderks. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*, 2020.

[33] W. McKinney et al. pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9, 2011.

[34] S. Momtazi et al. Fine-grained german sentiment analysis on social media. In *LREC*, volume 12, pages 1215–1220, 2012.

[35] A. Nugumanova, Y. Baiburin, and Y. Alimzhanov. Sentiment analysis of reviews in kazakh with transfer learning techniques. In *2022 International Conference on Smart Information Systems and Technologies (SIST)*, pages 1–6. IEEE, 2022.

[36] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*, 2002.

[37] P. Sen, Y. Li, E. Kandogan, Y. Yang, and W. Lasecki. Heidl: Learning linguistic expressions with deep learning and human-in-the-loop. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 135–140, 2019.

[38] B. Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020.

[39] N. Sultana, R. Sultana, R. I. Rasel, and M. M. Hoque. Aspect-based sentiment analysis of bangla comments on entertainment domain. In *2022 25th International Conference on Computer and Information Technology (ICCIT)*, pages 953–958. IEEE, 2022.

[40] M. Wang, H. Adel, L. Lange, J. Strotgen, and H. Schütze. Nlnde at semeval-2023 task 12: Adaptive pretraining and source language selection for low-resource multilingual sentiment analysis. *ArXiv*, abs/2305.00090, 2023. doi: 10.48550/arXiv.2305.00090.

[41] Z. J. Wang, D. Choi, S. Xu, and D. Yang. Putting humans in the natural language processing loop: A survey. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 47–52, 2021.

[42] H. Watanabe, M. Bouazizi, and T. Ohtsuki. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE access*, 6:13825–13835, 2018.

[43] S. Wu and M. Dredze. Are all languages created equal in multilingual bert? In *5th Workshop on Representation Learning for NLP, RepL4NLP*

2020 at the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, pages 120–130. Association for Computational Linguistics (ACL), 2020.

[44] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381, 2022.