Data Augmentation Based on Selective Masking of Language Models for One Health Context

Youssef Mahdoubi[1,3], Najlae Idrissi[1], Mathieu Roche[2,3], Sarah Valentin[2,3]
[1]IDC Team, Data4Earth Lab, Faculty of Science and Technology, Sultan Moulay Slimane University, Beni-Mellal, Morocco
[2]CIRAD, UMR TETIS, Montpellier, F-34398, France
[3]TETIS, Université de Montpellier, AgroParisTech, CIRAD, INRAE, Montpellier, France
mahdoubi.youssef@usms.ac.ma, n.idrissi@usms.ma
{mathieu.roche, sarah.valentin}@cirad.fr

Abstract

This study focuses on improving the performance of language models for two critical applications within the One Health context, specifically in epidemiological monitoring using textual data: (i) thematic classification across syndromic surveillance, biomedical and plant health domains, and (ii) detection of epidemic misinformation. A key challenge in these tasks is the limited availability of labeled textual data, which constrains the effectiveness of supervised learning methods. To overcome this limitation, we introduce two families of selective masking–based data augmentation strategies: lexical and non-lexical. Each family is implemented in a standard variant (Aug-SM-Lex and Aug-SM-NonLex), and a TF-IDF-weighted variant (Aug-SM-Lex-TFIDF and Aug-SM-NonLex-TFIDF). We perform two complementary experiments: the first determines the optimal masking rate, while the second evaluates the proposed strategies against LLM-based text reformulation. Experimental results indicate that selective masking-based augmentation outperformed both LLM-based reformulation (Mistral-7B and GPT-Neo-1.3B) and baseline models trained on original data alone across three of the five evaluated datasets, with the best performance achieved at a masking rate of 20%. This suggests that selective masking is a promising approach, potentially more effective than computationally expensive LLM-based reformulation.

1 Introduction

The emergence of pre-trained language models, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), has significantly improved various natural language processing (NLP) tasks by enabling the generation of rich contextual representations that capture subtle semantic nuances in large text corpora. Among these tasks, automatic text classification remains a critical task due to its central role in enabling domain-specific applications, including healthcare, agriculture, and social sciences. The One Health framework (Prata et al., 2022) is a transdisciplinary approach dedicated to protecting human, animal, and environmental health through coordinated surveillance, prevention, and mitigation measures. A key subdomain of One Health surveillance is the automatic collection and extraction of epidemiological information from textual documents, such as online news articles or scientific publications, where text classification plays a central role in most essential applications of this subdomain.

However, One Health applications face the challenge of scarcity of labeled data, which limits the performance of language models on classification tasks that require large volumes of labeled data to learn relevant textual representations. Although data manually labeled by domain experts is a possible solution and is used in several One Health applications, it remains costly in terms of time and resources, and it is difficult to annotate large amounts of data (Alzubaidi et al., 2023). Automatic textual data augmentation is a practical alternative, but it can introduce generation biases that lead to overfitting when fine-tuning language models for classification.

In line with the automatic data augmentation paradigm, our work focuses on improving the One Health subdomain of epidemiological monitoring based on the use of textual data. We address two key applications of this subdomain, including thematic classification in syndromic surveillance, biomedical and plant health, and the classification for the detection of epidemiological misinformation. To this end, we propose two families of data augmentation strategies: lexical and non-lexical, each available in a simple version and a version weighted by the TF-IDF (Aizawa, 2003) score. These strategies rely on selective masking, which consists of masking a controlled proportion

Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026), pages 266–276

of words in a targeted manner and predicting them using language models to generate new labeled texts. We evaluate their effectiveness through two experiments: (i) an analysis of the optimal masking rate that balances lexical diversity and contextual preservation, and (ii) a comparison with reformulation-based augmentation using large language models (LLMs).

The remainder of this paper is organized as follows. In Section 2, we provide a detailed review of related work. Section 3 describes the various proposed data augmentation strategies. Section 4 presents the experimental protocol, including datasets, implementation details, and evaluation metrics. In Section 5, we present and discuss our results. Finally, we conclude the paper and outline directions for future work in Section 6.

2 Related Work

The scarcity of labeled textual data is a widely recognized challenge in NLP applications, as it limits the performance of language models. This issue is particularly evident in classification tasks, where language models require large amounts of labeled data to learn relevant textual representations.

To address this problem, several studies have focused on textual data augmentation approaches, which involve creating artificial examples based on original texts using various methods. Easy Data Augmentation (EDA), proposed by Wei and Zou (2019), is one of the simplest augmentation methods, which combines four basic operations: synonym replacement, random word insertion, random deletion, and random word swap. Despite the simplicity of EDA, it demonstrates effective performance on small-scale datasets. Nevertheless, it has certain limitations, notably that it does not consider the overall context of the text and may generate grammatically incorrect outputs.

More advanced approaches exist, such as substitution-based augmentation methods. Within this category, Zhang et al. (2015) proposed an approach that consists of randomly replacing certain words in the text with synonyms using a thesaurus to generate new variants, which are then used to train a deep convolutional network on character-level representations. In the same substitution paradigm, Wang and Yang (2015) presents a data augmentation approach that consists of using pre-trained word representations (e.g. Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014)) to find, in the embedding space, the words closest to the word to be replaced. They used this approach to augment a dataset of tweets, enabling automatic classification of tweets based on the type of irritating behaviors or common annoyances they express, such as relationship issues, disrespectful actions, and hygiene concerns. This embedding-based augmentation approach showed a significant performance improvement, with a maximum gain of 6.1% in the F1-score. Another widely used technique is back-translation, presented by Edunov et al. (2018), which involves taking a sentence (e.g in English) and translating it into another language, then translating the new text back into the original language, generating a new instance. They demonstrated that back-translation, when applied on a large scale, significantly improves the performance of translation models, with the generated data sometimes proving to be as effective as real data. Beyond back-translation, other translation-based approaches have been proposed, including the method introduced by Ishigaki et al. (2023), which is based on translating existing texts in a given domain into the target language and using them directly for pre-training language models.

Finally, a recent approach of data augmentation by reformulation using LLMs, represented by Zhao et al. (2024), consists of using LLMs such as GPT (Brown et al., 2020) or T5 (Raffel et al., 2020) to rewrite existing texts while retaining their original labels, yielding generated labeled data. The results obtained by Zhao et al. (2024) show that this approach significantly improves the performance of language models such as BERT on classification tasks, particularly on limited or unbalanced datasets.

This study builds on previous research by focusing on selective masking–based data augmentation to increase labeled data in the One Health subdomain of epidemiological monitoring from textual data. Masking is the fundamental principle behind the masked language modeling (MLM) task, which is used during BERT's pre-training on large-scale unlabeled data (Devlin et al., 2019). This task involves randomly masking 15% of the tokens in a text and training the model to predict them, thereby enabling it to acquire general linguistic representations. Building on this principle, selective masking (SM) improves this approach by selectively masking specific terms and training the model to predict them, allowing it to acquire knowledge related to the target domain. The

choice of masking criteria varies depending on the context and objectives. Among the best known is masking based on a pre-collected domain-related lexicon, used by Pergola et al. (2021) to improve language models for question-answering tasks in biomedicine and by Borovikova et al. (2023) to improve the named entity extraction task. Other studies are based on a numeric statistic, notably the TF-IDF score, which reflects the importance of a term in a specific text, as proposed by Belfathi et al. (2024), and has shown good adaptation to tasks based on textual themes. Our work is based on using selective masking for data augmentation, by masking certain terms and letting the language model predict them, which generates new textual examples. To this end, we propose two families of selective masking-based augmentation strategies: lexical and non-lexical—each implemented in a simple version and a TF-IDF–weighted version.

3 Methodology

We used the two masking criteria proposed by Pergola et al. (2021), Borovikova et al. (2023), and Belfathi et al. (2024) (see Section 2) to create two families of augmentation approaches: lexical and non-lexical, each available in a simple version (Aug-SM-Lex and Aug-SM-NonLex) (see Fig. 1) and a version weighted by the TF-IDF score (Aug-SM-Lex-TFIDF and Aug-SM-NonLex-TFIDF) (see Fig. 2). Lexical masking targets domain-specific terms to preserve the overall context, whereas non-lexical masking maintains domain-specific information by masking other terms. Based on these principles, the proposed strategies generate a single augmented text per instance by selectively masking terms and predicting replacements.

3.1 Aug-SM-Lex

The Aug-SM-Lex strategy is based on masking terms in the target domain lexicon (i.e., 1-gram, 2-grams and 3-grams), which preserves the overall context of the text presented by terms outside the domain-specific lexicon, while introducing controlled lexical variability into the texts generated. The strategy is implemented as follows (see Fig. 1).

1. Mask terms that belong to the domain-specific lexicon.

2. If the $N\%$ (i.e., masking rate) is not reached, randomly mask additional terms until it is achieved.

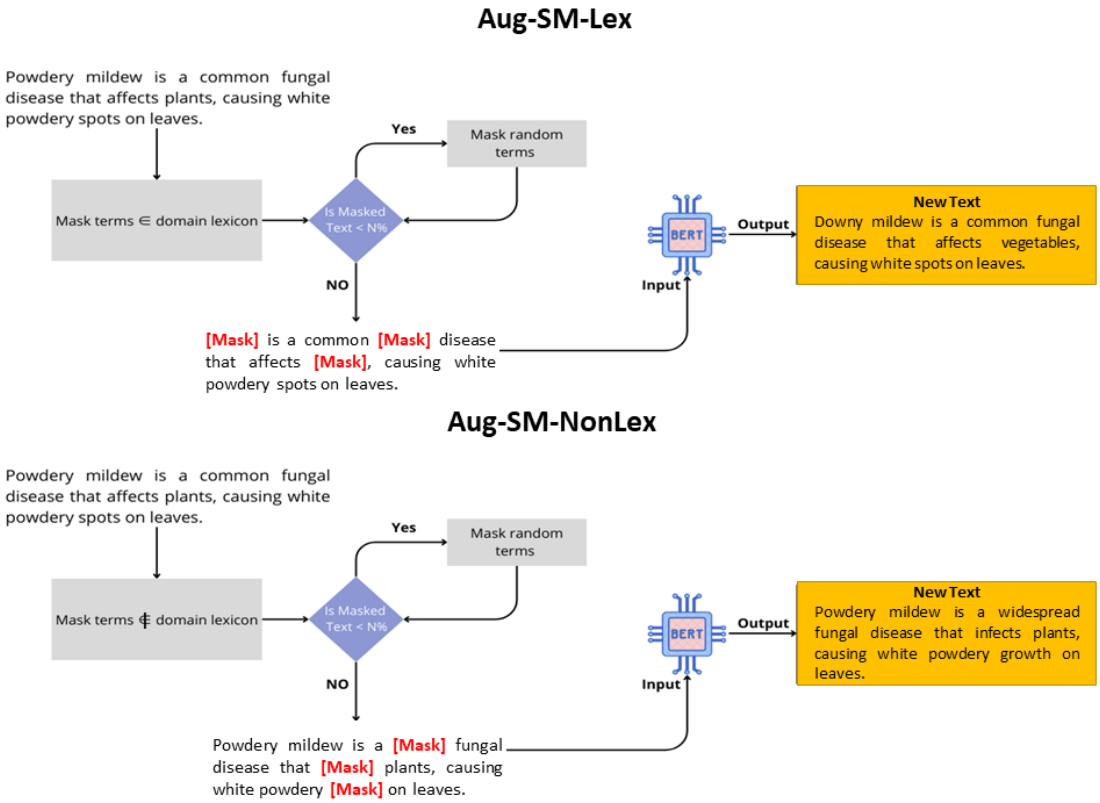

Figure 1: Overview of Aug-SM-Lex and Aug-SM-NonLex approaches.

3.2 Aug-SM-NonLex

Unlike the first augmentation strategy, Aug-SM-NonLex relies on masking terms that are not part of the target domain lexicon (i.e., 1-gram). This preserves the essential domain-specific context in the generated textual examples while introducing great lexical diversity, since masking is not limited to lexicon terms. The process for this strategy is as follows (see Fig. 1).

1. Mask terms that do not belong to the domain-specific lexicon.
2. If the $N\%$ is not reached, randomly mask additional terms until it is achieved.

3.3 Aug-SM-NonLex-TFIDF

The Aug-SM-NonLex-TFIDF strategy is a TF-IDF score-weighted version of the Aug-SM-NonLex approach, in which we add a condition to mask only terms with a low TF-IDF score (i.e., 1-gram). This choice reduces the masking interval by limiting the operation to terms outside the specialized lexicon that are considered uninformative based on their TF-IDF score, thereby preserving more context. The strategy is implemented as follows (see Fig. 2).

1. Select terms outside the domain-specific lexicon.
2. Compute the TF-IDF score for each term.
3. Rank the terms by ascending TF-IDF score.
4. Select the top $N\%$ of terms for masking.

3.4 Aug-SM-Lex-TFIDF

In line with the previous strategy, we weighted the Aug-SM-Lex strategy with the TF-IDF score, adding the condition that masked terms must belong to the lexicon (i.e., 1-gram, 2-gram and 3-gram) or, if they do not, have a low TF-IDF value. This approach allows only lexicon terms and less informative terms to be masked in order to generate lexical diversity while preserving the overall context. The strategy is implemented as follows (see Fig. 2).

1. Compute the TF-IDF score for all terms outside the domain-specific lexicon.
2. Assign low TF-IDF scores to terms belonging to the domain-specific (i.e., between 0.0 and 0.2).
3. Rank the terms by ascending TF-IDF score.
4. Select the top $N\%$ of terms for masking.

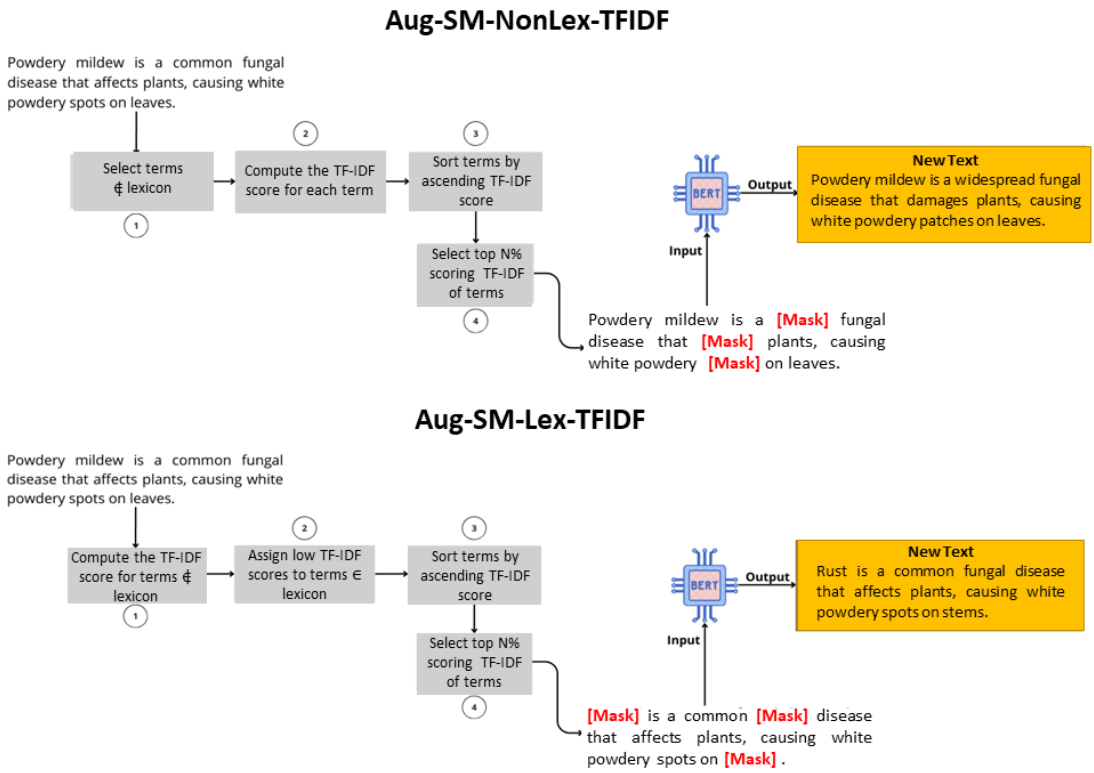

Figure 2: Overview of Aug-SM-Lex-TFIDF and Aug-SM-NonLex-TFIDF approaches

In the following sections, the selective masking-based augmentations presented in this section are compared to LLM-based reformulation to evaluate their ability to produce augmented texts that are both diverse and informative.

4 Experiments

4.1 Experimental Setup

All experiments carried out as part of this work were performed using the Python programming language, primarily with the PyTorch (Paszke et al., 2019) and Transformers (Wolf et al., 2020) libraries. We used $\text{BERT}_{\text{Base}}$ to apply our different selective masking-based data augmentation strategies, employing greedy decoding (top-1) to generate new replacement terms for each mask. We compared our approach to LLM-based reformulation, specifically Mistral-7b (Jiang et al., 2023) and GPT-Neo-1.3B (Black et al., 2021). The prompt used for these LLMs was designed to preserve the original meaning while introducing lexical variation.

> "Rewrite the following text, keeping the meaning but changing the wording: `[INPUT]`"

For the fine-tuning of $\text{BERT}_{\text{Base}}$ on the classification task, we used the following hyperparameters: a batch size of 16 and a learning rate of 2e-5 over 4 epochs. All training was performed on a T4 GPU.

4.2 Dataset Description

The various experiments conducted in this work focus on two applications within the One Health subdomain of epidemiological monitoring using textual data: (1) thematic classification in syndromic surveillance, biomedicine, plant health, and (2) detection of epidemic misinformation. To this end, we collected multiple labeled English datasets covering these domains (see Section 4.2.1 and Table 1). In addition, four domain-specific English lexicons were compiled (see Section 4.2.2 and Table 2) to be used with our different proposed augmentation strategies based on selective masking.

4.2.1 Labelled Datasets

- **Medical Text – Cancer**: 997 scientific articles and abstracts on the biomedical domain, specifically human cancers, extracted from the Medical Text Dataset - Cancer Doc Classification Dataset (FalgunPatel19, 2022). This corpus is divided into three classes (Thyroid Cancer: 283, Colon Cancer: 261, Lung Cancer: 453).

- **PADI-web Plant Health**: 748 news articles on the plant health domain, specifically on *Xylella fastidiosa* (i.e., plant disease), collected from PADI-web (Valentin et al., 2021) and manually classified by experts into two classes: relevant (317 articles, i.e., documents related to a new, suspected, or unknown outbreak) or not relevant (431 articles).

- **PADI-web Syndromic**: 769 online news articles on the syndromic surveillance domain collected from PADI-web, divided into two classes: positive, with 311 news articles dealing with unknown diseases, and negative, with 458 news articles where a pathogenic cause is identified.

- **CoAID**: 252 news articles and Facebook posts on the COVID-19 epidemic, extracted from the largest CoAID dataset (Cui and Lee, 2020). This corpus is divided into two classes: fake, with 126 fake news articles/posts, and true, with 126 real news articles/posts. We selected

Dataset	Domain	#Docs	#Classes
Medical Text – Cancer	Biomedical	997	3
PADI-web Plant Health	Plant health	748	2
PADI-web Syndromic	Syndromic Surveillance	769	2
CoAID	Epidemic misinformation	252	2
COVID-19 Fake News	Epidemic misinformation	6,120	2

Table 1: Overview of the labelled datasets

Lexicon	Domain	#Terms
Plant-Health-Lexicon	Plant Health	1,787
Biomedical-Lexicon	Biomedical	4,702
Syndromic-Surveillance-Lexicon	Syndromic Surveillance	465
Misinformation-Lexicon	Epidemic misinformation	11,653

Table 2: Overview of domain-specific lexicons

this amount of data to have a balanced corpus, because the original CoAID contains 3,565 true articles and posts and 204 fake articles and posts.

- **COVID-19 Fake News**: 6120 social media posts, such as Twitter and Facebook, on COVID-19, extracted from the COVID-19 Fake News Dataset (Patwa et al., 2020). This corpus is divided into two classes (fake: 3060, real: 3060).

4.2.2 Domain-Specific Lexicons

- **Biomedical-Lexicon**: 4,702 biomedical terms, extracted from a biomedicine dictionary published in Oxford Reference (Oxford Reference, 2010), and from another glossary published by RxList[1].

- **Plant-Health-Lexicon**: 1,787 plant health terms, including plant names, scientific names, diseases, disease categories, symptoms, pests, and general terminology used in plant health. We collected this lexicon from several sources, such as BSPP[2] and APS.

- **Syndromic-Surveillance-Lexicon**: 465 terms dedicated to the syndromic surveillance domain, including symptoms, clinical signs, terms used to describe unknown diseases (e.g., mysterious, unexpected), and general terminology used in this domain. We extracted this lexicon from a glossary of clinical signs in laboratory animals proposed by the University of Zurich[3] and from a list of keywords for the PADI-web set up in the Indian Ocean (Rasamoelina et al., 2023).

- **Misinformation-Lexicon**: 11,653 terms dedicated to the domain of epidemic misinformation. It includes the biomedical lexicon used in epidemiology, the most words used to describe news from Word Raiders [4], reporting verbs from Newcastle University[5], analysis verbs from HelpfulProfessor[6], adjectives to describe news, and words used to describe sentiments from the Bing sentiment dataset collected in the work of Hu and Liu (2004).

[1] https://www.rxlist.com/drug-medical-dictionary

[2] https://www.bspp.org.uk/glossary/

[3] https://www.tierschutz.uzh.ch

[4] https://wordraiders.com/guides/most-common-words-used-in-news/

[5] https://www.ncl.ac.uk/academic-skills-kit/writing/

[6] https://helpfulprofessor.com/analysis-verbs/

4.3 Data Preparation

In order to fine-tune our language models on the classification tasks, we considered two data distribution scenarios.

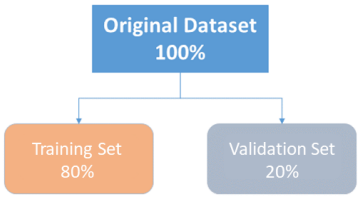

Figure 3: Overview of the data distribution for the original datasets

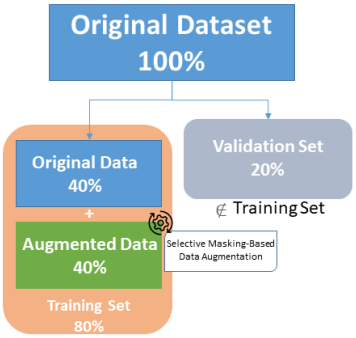

Figure 4: Overview of the data distribution with augmented data

The first scenario relies exclusively on the original datasets: each dataset is divided into training and validation sets according to an 80/20 ratio (see Fig. 3). The second scenario incorporates augmented data by retaining only 50% of the original training set (i.e., corresponding to half of the 80% training portion presented in Fig. 3), while the other half is replaced with data generated using data augmentation approaches (see Fig. 4). This scenario allows us to build a training set that combines original and augmented data, while maintaining the same overall size as the original dataset distribution (Scenario 1). Keeping the size of the dataset consistent ensures fair and balanced comparisons between the different data augmentation approaches, as any variation observed in the performance of the language model can be attributed to the quality of the augmented data, rather than to an increase in the size of the data compared to the original dataset. In both scenarios, the validation set consists solely of original labeled examples not included in the training set, ensuring a reliable evaluation while

limiting the risk of overfitting. The data distribution with augmented data from the various labeled datasets collected for each domain is summarized in Table 3.

Corpus	Train Orig.	Train Aug.	Val.	Total
Medical Text – Cancer	398	398	201	997
PADI-web Plant Health	299	299	150	748
PADI-web Syndromic	307	307	155	769
CoAID	101	101	50	252
COVID-19 Fake News	2448	2448	1224	6120

Table 3: Data distribution after augmentation

4.4 Evaluation Metrics

We adopted a two-part evaluation protocol. The first part concerns the evaluation of the quality of the augmented data and the choice of the optimal masking rate. To do this, we used BERTScore (Zhang et al., 2020) to evaluate the ability of the new texts to preserve context compared to the originals. This metric works by comparing texts at the semantic level, using contextual representations from BERT-type models, by computing the cosine similarity between token embeddings to estimate their similarity. We also used the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004) metric, which evaluates the lexical diversity of the new texts generated compared to the originals by measuring lexical overlap (i.e, ROUGE-1: 1-grams, ROUGE-2: 2-grams, or ROUGE-L: longest common subsequences, computed over all matching subsequences between reference and generated text). The objective of using these two metrics is to find a balanced combination between the conservation of the original context and ensuring lexical diversity. The second part is reserved for evaluating the impact of augmentation on the performance of the language model in the classification task. To do this, we used overall accuracy as a general metric, accompanied by the weighted F1-score, precision, and recall, to obtain a reliable and comprehensive estimate of performance.

5 Results and Discussion

In order to evaluate our different proposed augmentation strategies, we conducted two essential experiments. The first was to choose the optimal rate of terms to mask. The second compared our selective masking–based augmentation strategies with LLM-based reformulation approaches on the classification task, specifically for thematic classification and epidemic misinformation detection.

5.1 Experiment 1: Impact of the masking rate

In this experiment, we evaluated several masking rates, specifically 10%, 20%, and 30%, using our four proposed augmentation strategies: Aug-SM-Lex, Aug-SM-NonLex, Aug-SM-Lex-TFIDF, and Aug-SM-NonLex-TFIDF, to identify the optimal masking rate, i.e., the one that provides a balance between context preservation and significant lexical diversity. For this, we used BERTScore and ROUGE-1 as the main metrics for evaluation. The original BERT pre-training masking rate of 15% was not tested explicitly because this study focuses on data augmentation rather than pre-training; thus, masking rates were assessed in 10% increments to systematically examine the trade-off between lexical diversity and context preservation across a broader range. Masking rates greater than 30% were excluded due to excessive context loss resulting from the high proportion of modified text.

Metric / Masking rate	10%	20%	30%
BERTScore			
Aug-SM-Lex	0.78	0.71	0.67
Aug-SM-NonLex	0.79	0.74	0.69
Aug-SM-NonLex-TFIDF	0.78	0.70	0.67
Aug-SM-Lex-TFIDF	0.78	0.73	0.68
ROUGE-1			
Aug-SM-Lex	0.77	0.73	0.68
Aug-SM-NonLex	0.78	0.74	0.69
Aug-SM-NonLex-TFIDF	0.78	0.73	0.68
Aug-SM-Lex-TFIDF	0.78	0.73	0.69

Table 4: BERTScore and ROUGE-1 results for the PADI-web Plant Health Dataset, using different masking rates.

Metric / Masking rate	10%	20%	30%
BERTScore			
Aug-SM-Lex	0.75	0.71	0.67
Aug-SM-NonLex	0.74	0.64	0.57
Aug-SM-NonLex-TFIDF	0.75	0.70	0.55
Aug-SM-Lex-TFIDF	0.75	0.65	0.57
ROUGE-1			
Aug-SM-Lex	0.92	0.85	0.79
Aug-SM-NonLex	0.90	0.74	0.69
Aug-SM-NonLex-TFIDF	0.91	0.83	0.65
Aug-SM-Lex-TFIDF	0.92	0.78	0.69

Table 5: BERTScore and ROUGE-1 results for the COVID-19 Fake News Dataset, using different masking rates.

Tables 4 and 5 show the BERTScore and ROUGE-1 scores for the PADI-web Plant Health and COVID-19 Fake News Dataset, which are presented here as representative examples. The results

indicate that at a masking rate of 10%, the context is largely preserved (i.e., BERTScore remains high), but the diversity introduced is low, as indicated by the higher ROUGE-1 values, meaning that the generated texts remain very close to the originals. On the other hand, the results obtained using the masking rate of 30% show a greater lexical diversity (i.e., lower ROUGE-1), but at the cost of a lower BERTScore, showing that the texts begin to stray too far from the original content and lose contextual fidelity. Between these two extremes, the masking rate of 20% offers the best balance between lexical diversity and preservation of meaning.

5.2 Experiment 2: Selective masking-based augmentation vs. LLM-based reformulation augmentation

Building on the optimal 20% masking rate determined in Experiment 1, we conducted this experiment to verify the performance of our proposed augmentation strategies. To do this, we fine-tuned $\text{BERT}_{\text{Base}}$ for classification tasks on augmented datasets using our different selective masking-based strategies, following the data distribution defined in Section 4.3. These strategies were then compared with LLM-based reformulation augmentation and the baseline (i.e., the original data). Other conventional augmentation methods (e.g., EDA or embedding-based substitution) were not considered, as the specialized context and vocabulary of our datasets limit their effectiveness in the One Health domain, making LLM-based reformulation a more suitable benchmark.

The proposed selective masking-based data augmentation strategies offer better performance than LLM-based reformulation augmentation for most domains (see Table 6). Specifically, the Aug-SM-Lex-TFIDF strategy achieves the best performance in the syndromic surveillance domain (i.e., PADI-web Syndromic Dataset), with an accuracy of 78.57% and an F1-score of 78.52%, outperforming both the baseline and the LLM reformulation approaches. Similarly, in the domain of epidemic misinformation detection (i.e., COVID-19 Fake News Dataset), this strategy yields strong performance, with both accuracy and the F1-score reaching 97.38%. In the biomedical domain (i.e., Medical Text – Cancer), Aug-SM-NonLex achieves the highest performance, with an accuracy of 76.28% and an F1-score of 74.67%. However, reformulation using Mistral-7B provides the best results for the plant health domain (i.e., PADI-web Plant Health Dataset), with an accuracy of 86.66% and an F1-score of 86.58%. For the detection of epidemic misinformation (CoAID Dataset), none of the proposed augmentation strategies, including those based on LLM reformulation, exceeded the baseline. However, Aug-SM-NonLex-TFIDF and GPT-Neo-1.3B give strong results close to the base-

Datasets	Metric	Baseline	GPT-Neo-1.3B	Mistral-7B	Aug-SM-Lex	Aug-SM-NonLex	Aug-SM-NonLex-TFIDF	Aug-SM-Lex-TFIDF
Medical Text – Cancer	Accuracy	70.79	73.88	74.57	73.19	**76.28**	72.85	73.19
	Precision	70.80	74.14	74.79	74.23	**79.58**	72.93	73.29
	Recall	70.79	73.88	74.57	73.19	**76.28**	72.85	73.19
	F1-Score	70.77	73.15	74.39	72.42	**74.67**	72.97	73.09
PADI-web Plant Health	Accuracy	82.66	82.66	**86.66**	83.33	86.00	84.66	84.00
	Precision	82.75	83.04	**87.63**	83.34	86.31	84.82	84.64
	Recall	82.66	82.66	**86.66**	83.33	86.00	84.66	84.00
	F1-Score	82.65	82.61	**86.58**	83.31	85.96	84.64	83.92
PADI-web Syndromic	Accuracy	74.67	72.72	77.92	75.32	77.92	73.37	**78.57**
	Precision	79.44	72.72	77.99	76.19	79.18	75.66	**78.80**
	Recall	74.67	72.72	77.92	75.32	77.92	73.37	**78.57**
	F1-Score	73.60	72.72	77.90	75.11	77.68	72.76	**78.52**
CoAID	Accuracy	**94.00**	92.00	90.00	84.00	84.00	92.00	86.00
	Precision	**94.64**	92.27	90.00	84.21	87.87	93.10	89.06
	Recall	**94.00**	92.00	90.00	84.00	84.00	92.00	86.00
	F1-Score	**93.99**	91.98	89.99	83.97	83.57	91.98	85.72
COVID-19 Fake News Dataset	Accuracy	97.30	96.73	96.81	96.89	96.40	97.22	**97.38**
	Precision	97.37	96.75	96.81	96.89	96.45	97.24	**97.39**
	Recall	97.30	96.73	96.81	96.89	96.40	97.22	**97.38**
	F1-Score	97.30	96.73	96.81	96.89	96.40	97.22	**97.38**

Table 6: Comparison of $\text{BERT}_{\text{Base}}$ performance on classification task using selective masking-based augmentation vs. LLM-based reformulation augmentation (GPT-Neo-1.3B and Mistral-7B). The baseline represents fine-tuning $\text{BERT}_{\text{Base}}$ on the original data. All values are percentages (%).

line, with an accuracy of 92.00% and an F1-score of 91.98%, respectively.

Overall, these results demonstrate that the proposed selective masking–based data augmentation approaches generally outperform LLM-based reformulation using Mistral-7B and GPT-Neo-1.3B as well as the baseline for thematic classification tasks in the biomedical and syndromic surveillance domains. In the plant health domain, they achieve performance comparable to LLM-based reformulation and superior to baseline. Furthermore, for the detection of epidemic misinformation, our proposed approaches remain competitive relative to the baseline. These findings highlight the effectiveness of selective masking–based data augmentation strategies for thematic and misinformation classification tasks within the One Health context, supporting more effective epidemiological monitoring from textual data.

6 Conclusion

In this paper, we proposed two families of data augmentation approaches based on selective masking: lexical and non-lexical, each available in a simple version and a TF-IDF-weighted version. We applied these approaches to two types of applications representative of the One Health context, focusing on the subdomain of epidemiological monitoring from textual data: (1) thematic classification in biomedical, plant health, and syndromic surveillance, and (2) classification for the detection of epidemic misinformation. To evaluate our strategies, we conducted two experiments: the first to determine the optimal masking rate, and the second to compare our augmentation approaches with a reformulation method based on LLMs. The results indicate that, at a masking rate of 20%, selective masking–based strategies generally outperform LLM-based reformulation using Mistral-7B and GPT-Neo-1.3B.

This work is a first step toward designing robust data augmentation approaches for classification tasks in the One Health context. Looking ahead, our goal is to incorporate more detailed analyses and domain-expert evaluations. In addition, we plan to extend our experiments by comparing our approach with other LLMs and exploring alternative pre-trained language models, such as RoBERTa (Liu et al., 2019) and BioBERT (Lee et al., 2019), to further assess the generalizability of the proposed augmentation strategies. Finally, we intend to integrate LLMs for automatic domain-lexicon generation and to predict generated masks directly, potentially replacing conventional pre-trained models. These extensions should improve the effectiveness of the proposed augmentation strategies for thematic classification and detection of epidemic misinformation.

7 Limitations

The experiments are conducted exclusively on English-language data, which limits the generalizability of the proposed selective masking strategies to other languages with different linguistic structures. In addition, our approach relies on domain-specific lexicons to guide masking, which may reduce its effectiveness in domains where such lexicons are sparse, incomplete, or noisy. Finally, the evaluation is limited to a single classification task and the observed gains may not be directly transferred to other downstream NLP tasks.

8 Ethical Considerations

In this work, we used only corpora and lexicons collected from publicly available data that do not contain personal or sensitive information. Nevertheless, our proposed augmentation methods may unintentionally introduce several risks, such as amplifying biases present in the original data, causing semantic drift from the source texts, or introducing label leakage if the data resulting from the augmentation is not properly separated between the training and evaluation phases. These challenges could compromise the reliability of the classification results produced by the language model. To mitigate them, augmentation is applied only to training data, the masking rate is controlled, and measures such as BERTScore and ROUGE are used to evaluate semantic fidelity and lexical diversity. We also recommend human validation before any use in sensitive applications, ensuring responsible use of our methods.

Acknowledgements

This work was supported by the National Center for Scientific and Technical Research (CNRST) under the PhD-Associate Scholarship (PASS) program, in line with the Morocco 2030 National Strategy. This study was partially funded by French General Directorate for Food (DGAL) and by EU grant 874850 MOOD.

References

Akiko Aizawa. 2003. An information-theoretic perspective of tf-idf measures. *Information Processing & Management*, 39(1):45–65.

Laith Alzubaidi, Jinshuai Bai, Aiman Al-Sabaawi, Jose Santamaría, A. S. Albahri, Bashar Sami Nayyef Al-dabbagh, Mohammed A. Fadhel, Mohamed Manoufali, Jinglan Zhang, Ali H. Al-Timemy, and 1 others. 2023. A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, 10(1):46.

Anas Belfathi, Ygor Gallina, Nicolas Hernandez, Laura Monceaux, and Richard Dufour. 2024. Adaptation des modèles de langue à des domaines de spécialité par un masquage sélectif fondé sur le genre et les caractéristiques thématiques. In *Actes de la 31ème Conf. sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 283–294. ATALA and AFPC.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow. Zenodo.

Mariya Borovikova, Arnaud Ferré, Robert Bossy, Mathieu Roche, and Claire Nédellec. 2023. Could keyword masking strategy improve language model? In *Natural Language Processing and Information Systems*, pages 271–284. Springer, Cham.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. arXiv preprint.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*, pages 4171–4186.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. ACL.

FalgunPatel19. 2022. Medical text dataset - cancer document classification. Kaggle. Retrieved from `https://www.kaggle.com/datasets/falgunipatel19/biomedical-text-publication-classification`. Accessed September 26, 2025.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proc. of the Tenth ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining (KDD'04)*, Seattle, WA, USA.

Tatsuya Ishigaki, Yui Uehara, Goran Topić, and Hiroya Takamura. 2023. Pretraining language- and domain-specific BERT on automatically translated text. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 548–555. INCOMA Ltd., Shoumen, Bulgaria.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. ACL.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conf. on Learning Representations, Workshop Track Proc.*, Scottsdale, Arizona, USA.

Oxford Reference. 2010. *A dictionary of biomedicine*. Retrieved from `https://www.oxfordreference.com/display/10.1093/acref/9780199549351.001.0001/acref-9780199549351`. Accessed September 24, 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Parth Patwa, Shivam Sharma, Srinivas PYKL, Vineeth Guptha, Gitanjali Kumari, Md. Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Fighting an infodemic: COVID-19 fake news dataset. *CoRR*, abs/2011.03327.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. ACL.

Gabriele Pergola, Elena Kochkina, Lin Gui, Maria Liakata, and Yulan He. 2021. Boosting low-resource biomedical qa via entity-aware masking strategies. In *Proc. of the 16th Conf. of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1977–1985.

Joana C. Prata, Ana Isabel Ribeiro, and Teresa Rocha-Santos. 2022. Chapter 1—an introduction to the concept of one health. In *One Health*, pages 1–31. Academic Press.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

H. Rasamoelina, L. P. Veerapa-Mangroo, S. A. Bedja, and M. Roche. 2023. Mots-clés pour padi-web mis en place dans l'océan indien. CIRAD Dataverse.

Sarah Valentin, Elena Arsevska, Julien Rabatel, Sylvain Falala, Alizé Mercier, Renaud Lancelot, and Mathieu Roche. 2021. Padi-web 3.0: A new framework for extracting and disseminating fine-grained information from the news for animal disease surveillance. *One Health*, 13:100357.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing*, pages 2557–2563, Lisbon, Portugal. ACL.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In *Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th International Joint Conf. on Natural Language Processing, EMNLP-IJCNLP*, pages 6381–6387. ACL.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proc. of EMNLP 2020: System Demonstrations*, pages 38–45.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conf. on Learning Representations*, Addis Ababa, Ethiopia. OpenReview.net.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proc. of the 29th International Conf. on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

Huanhuan Zhao, Haihua Chen, Thomas A. Ruggles, Yunhe Feng, Debjani Singh, and Hong-Jun Yoon. 2024. Improving text classification with large language model-based data augmentation. *Electronics*, 13(13).