Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.3 June 2026, pp. 2546~2555
e-ISSN: 2548-6861

An End-to-End NLP Pipeline Combining Web Scraping, CamemBERT Fine-Tuning and Zero-Shot Biomedical Named-Entity Recognition for Early Epidemic Signal Detection from French-Language Online News

Franklin Mwamba [1], Fiston Oshasha [2***], Saint Jean Djungu [3]*, John Poma [4*]**
* Department of Mathematics, Statistics and Computer Science, University of Kinshasa, Kinshasa, DR. Congo
**Health Sciences Research Institute, Kinshasa, Democratic Republic of the Congo
***General Commissariat for Atomic Energy, Regional Center for Nuclear Studies of Kinshasa, P.O. Box 868, University of Kinshasa
franklin.mwamba@irss.cd [1], fiston.oshasha.fv@cgea-rdc.org [2], saintjean.djungu@unikin.ac.cd [3], pomaesendo7@gmail.com [4]

Article Info

Article history:

Received 2026-04-17
Revised 2026-05-24
Accepted 2026-05-29

Keyword:

Natural Language Processing,
Transformer Fine-Tuning,
Web Scraping,
Zero-Shot Named-Entity Recognition,
CamemBERT,
Text Classification,
Epidemic Surveillance,
End-To-End Pipeline.

ABSTRACT

Epidemiological surveillance in the Democratic Republic of the Congo (DRC) suffers from reporting delays and limited digital infrastructure, while online French-language news provides a complementary real-time signal that current systems exploit poorly. We design, deploy, and rigorously evaluate an end-to-end natural-language processing (NLP) pipeline that integrates targeted web scraping of Congolese online media, sentence-level binary classification of epidemic content with a fine-tuned CamemBERT transformer, zero-shot biomedical named-entity recognition (CamemBERT-bio-GLiNER) restricted to disease, location and date, and an alerting dashboard built on a Django/Celery stack. The classifier was fine-tuned on a hybrid corpus of 11,433 sentences combining 1,433 manually annotated real news sentences and 10,000 template-generated synthetic sentences, and is benchmarked against two classical baselines (TF-IDF combined with Logistic Regression and Linear SVM) on an independent, manually annotated test set of 997 sentences (341 epidemic, 656 non-epidemic) constructed from a second scraping campaign performed three months later. We report precision, recall, F1, PR-AUC and ROC-AUC with 1,000-iteration bootstrap 95% confidence intervals. CamemBERT reaches F1 = 0.754 [0.717-0.787] and PR-AUC = 0.699 [0.644-0.756] for the epidemic class, while the Linear SVM baseline reaches F1 = 0.858 ± 0.037 and PR-AUC = 0.926 ± 0.024 in 5-fold stratified cross-validation, outperforming the transformer, a result we attribute to the dominance of synthetic data in the training corpus. A single-batch operational run of the full pipeline on MediaCongo processed 30 articles and 501 sentences in 37.5 s on a single GPU, producing 43 alerts that correctly captured the May 2026 Ebola Bundibugyo outbreak in Ituri. The system, the external benchmark, and all evaluation scripts are released as open source.

I. INTRODUCTION

Epidemiological surveillance relies on the capacity of health systems to systematically collect, analyze, and interpret health-related data in real time, in order to anticipate, detect, and respond effectively to public health threats. In resource-limited settings such as the Democratic Republic of the Congo (DRC), this function remains particularly challenging due to multiple structural constraints, including incomplete healthcare coverage, delayed reporting of health information, limited digital infrastructure, and difficulties in accessing remote areas [1]. These limitations are especially critical given that the country is frequently affected by high-impact epidemics such as cholera, measles, Ebola virus disease, and COVID-19. In many cases, outbreaks spread significantly before institutional response mechanisms can be effectively activated, largely due to the absence of efficient early warning systems.

In this context, digital media represent a promising yet still underexploited source of epidemiological information. Online news platforms such as Actualite.cd, Radio Okapi, 7sur7.cd, MediaCongo, and Ouragan.cd publish large volumes of daily content covering various domains, including public health. Despite their heterogeneity and unstructured nature, these textual data may contain weak or early signals of emerging or unusual health events [2], [3]. However, their exploitation in current surveillance systems remains largely manual, fragmented, and rarely integrated into automated detection pipelines.

To address this gap, this study proposes the design and implementation of an automated epidemiological surveillance system based on the analysis of French-language online media content. The core of the system is a binary text classification model trained on a corpus of 11,433 text segments, composed of 1,433 manually annotated samples and 10,000 synthetically generated samples, aimed at increasing linguistic diversity and improving model robustness. Each text segment is analyzed to determine whether it contains an epidemic-related signal. When relevant content is detected, associated metadata (such as title, source, and URL) are extracted, and an alert is automatically generated and displayed through a dedicated web-based platform.

To further evaluate the robustness and generalization capability of the model, an additional assessment was conducted using an independent manually annotated dataset. This provides a more realistic estimate of performance under real-world conditions and helps mitigate potential biases introduced by synthetic data.

The main hypothesis of this work is that early epidemiological signals can often be identified within journalistic content prior to official confirmation by health authorities. By leveraging artificial intelligence techniques to automatically detect such signals, this study aims to demonstrate the feasibility and relevance of an intelligent, scalable, and context-adapted epidemiological surveillance system for resource-constrained environments such as the DRC [9].

The contribution of this work is therefore not the introduction of yet another transformer-based classifier, but rather (i) the design and operational deployment of an open, reproducible end-to-end pipeline tailored to the Congolese French-language media ecosystem; (ii) the release of an independent, manually annotated external benchmark of 997 sentences; and (iii) an honest comparative evaluation that quantifies the actual benefit, or absence thereof, of transformer fine-tuning over classical lexical baselines when the training corpus is dominated by synthetically generated data. We argue that, in the resource-constrained setting we target, this comparative perspective is more useful than reporting a single transformer accuracy figure.

II. RELATED WORK AND POSITIONING

Recent research has increasingly focused on leveraging web data and Natural Language Processing (NLP) techniques for automated public health and epidemic surveillance. These studies can be broadly grouped into four main categories based on their methodological approaches and system designs.

A. Web-based epidemic event extraction systems

A first line of research focuses on extracting structured epidemic events from unstructured web and media sources. Hong et al. introduced RENA, a real-time system that integrates Named Entity Recognition (NER) and relation extraction to process infectious disease-related news articles and generate structured epidemiological information [4]. Similarly, Lejeune et al. developed DANIEL, a multilingual event extraction system designed to automatically identify and extract epidemic-related events from news streams across different languages [5]. These approaches demonstrate the feasibility of transforming raw textual news into structured epidemiological knowledge, enabling automated monitoring of disease outbreaks.

B. Machine learning and deep learning approaches for early outbreak detection

Another group of studies emphasizes the use of machine learning models combined with web scraping techniques to detect early signals of disease outbreaks. Miano et al. proposed a method that combines targeted web scraping with bidirectional neural networks (Bi-LSTM) to identify early indicators of COVID-19 outbreaks in specific environments such as food production and retail sectors [6]. This category highlights the ability of deep learning models to capture subtle patterns in noisy textual data and support early warning systems for public health surveillance.

C. End-to-end architectures for real-time epidemic monitoring

More recent work has shifted toward the design of complete end-to-end systems integrating data collection, NLP processing, machine learning, and visualization components. Phutane et al. proposed a comprehensive framework that combines real-time web scraping, machine learning-based analysis, and interactive dashboards to continuously monitor disease outbreaks [7]. These systems aim to bridge the gap between raw data acquisition and decision-support tools for public health authorities.

D. Pretrained language models for health-related information extraction

The emergence of transformer-based models has significantly improved text understanding capabilities in biomedical and general domains. Pretrained models such as CamemBERT have shown strong performance in extracting relevant entities and semantic information from complex and noisy French-language text [8]. These models enable more

accurate detection of disease mentions, health alerts, and implicit epidemiological signals from large-scale unstructured data sources.

E. Positioning of this work

In this context, the present work is positioned at the intersection of web-based data collection, NLP-based information extraction, and real-time epidemic monitoring systems. It aims to address the following gaps: limited integration between real-time web scraping and advanced NLP models in a unified pipeline; insufficient adaptation of pretrained language models for continuous epidemic signal detection in noisy multilingual data; lack of lightweight yet scalable architectures suitable for real-time deployment in resource-constrained environments. Crucially, unlike most prior work in this area, we explicitly compare our fine-tuned transformer to classical lexical baselines on an external, manually annotated test set, and we report bootstrap confidence intervals for all metrics. This honest comparative protocol is what allows us to uncover the lexical-shortcut effect.

III. METHODOLOGY

A. Corpus Construction

The dataset used in this study is built from two complementary sources. First, textual data were automatically collected from five widely accessed online news platforms in the Democratic Republic of the Congo (DRC), namely Actualite.cd, Radio Okapi, 7sur7.cd, MediaCongo, and Ouragan.cd. These platforms regularly publish articles related to national news, including public health topics. A targeted web scraping pipeline was developed using Python libraries such as requests and BeautifulSoup [10], following ethical data collection practices. Relevant articles were extracted from health-related sections along with their associated metadata (title, URL, and publication date). The collected paragraphs were then segmented into individual sentences to ensure finer granularity for analysis.

Second, an additional set of synthetic texts was generated using Python scripts to increase the diversity of epidemic-related expressions. These texts were designed based on common linguistic patterns observed in journalistic writing, enabling an artificial enrichment of the positive class (epidemic signals) while preserving linguistic realism.

The final merged corpus consists of 11,433 text segments, each corresponding to a single sentence, stored in a structured CSV file.

B. Supervised Annotation

Each text segment in the dataset was manually annotated into two classes:

- Class 1 – Epidemic: texts containing explicit or implicit epidemic signals (e.g., mentions of diseases, health alerts, unusual deaths, outbreaks, or critical situations);
- Class 0 – Non-epidemic: texts with no direct relation to collective health risks.

The annotation process was carried out by a single annotator following predefined guidelines based on representative examples. Inter-annotator agreement (e.g., Cohen's kappa) was not formally measured in the present version; a double-annotation campaign is planned for the next release.

C. Text Preprocessing

The collected texts underwent a standardized preprocessing pipeline in French, including:

- Sentence segmentation and removal of special characters;
- Tokenization and normalization using the spaCy library;
- Stopword removal to reduce noise in textual representation;
- Encoding using the CamemBERT tokenizer, producing input_ids and attention_mask required for model training.

D. Model Training

The classification model is based on CamemBERT-base, a Transformer-based pretrained language model optimized for French [8]. The model was fine-tuned for a binary classification task using the following configuration:

- Data split: 80% training, 10% validation, and 10% testing;
- Optimization: AdamW optimizer with a learning rate of 2e-5 and a batch size of 16;
- Training duration: 3 epochs with early stopping;
- Class imbalance handled through dynamic class weighting;
- Execution environment: Google Colab Pro (NVIDIA T4 GPU) and local server.

E. External Validation Dataset

To assess the robustness and generalization capability of the proposed model, an additional independent dataset was constructed and used exclusively for evaluation. This dataset consists of 997 text segments, all manually annotated, and entirely distinct from the training corpus. It contains:

- 656 samples labeled as Non-epidemic
- 341 samples labeled as Epidemic

Unlike the training dataset, this validation corpus contains no synthetically generated data, ensuring a more realistic representation of real-world textual variability. This external evaluation allows for a more reliable assessment of the model's performance under practical conditions and helps identify potential overfitting or bias introduced during training. The external dataset was constructed from a second, fully independent scraping campaign performed three months after the training campaign, ensuring temporal separation from the training data.

F. Named Entity Recognition

To enrich detected alerts, a Named Entity Recognition (NER) module was integrated to automatically extract key entities (disease, location, date) from texts classified as epidemic. For this purpose, we employed CamemBERT-bio-GLiNER [11], a Transformer-based model pretrained on French biomedical data and optimized for zero-shot NER. This model enables the extraction of domain-specific entities without additional fine-tuning. It is built upon the same architecture as implemented in the Transformers library [12], with adaptations for biomedical text processing.

G. System Deployment

The trained model was deployed within a web-based application developed using the Django framework [13], with task orchestration handled by Celery [14] for asynchronous processing. The system provides the following functionalities:

- Automated scraping of new articles every 4 hours;
- Background classification of extracted text segments;
- Automatic alert generation upon detection of epidemic signals;
- Real-time visualization through an interactive web dashboard.

H. Technical Architecture

The overall system architecture integrates data collection, preprocessing, classification, entity extraction, and visualization components into a unified pipeline, enabling continuous and automated epidemiological monitoring.

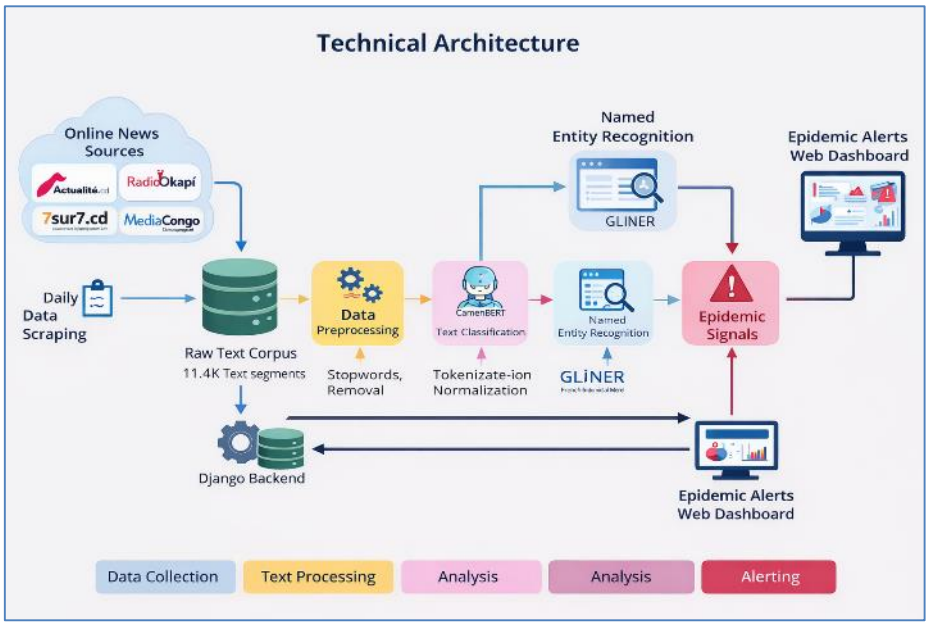

Figure 1: Technical architecture of the system

IV. RESULTS AND DISCUSSION

A. Dataset Description

The final dataset used for model training consisted of 11,433 text segments, including:

- 1,433 sentences collected via web scraping from Congolese online news platforms (Actualite.cd, Radio Okapi, MediaCongo, 7sur7.cd, and Ouragan.cd) and manually annotated;
- 10,000 synthetically generated sentences created using Python scripts to simulate epidemic and non-epidemic scenarios, automatically annotated.

Each text segment represents an independent unit labeled as 1 for epidemic signal detection or 0 for non-epidemic content.

B. Model performance on internal validation

The CamemBERT-base model was trained for 3 epochs using an 80/10/10 split for training, validation, and testing. The performance on the validation set is summarised in Table 1.

TABLE 1.
MODEL PERFORMANCE PER EPOCH (INTERNAL VALIDATION).

Epoch	Training Loss	Validation Loss	Accuracy	F1-score	Precision	Recall
1	0.069	0.0211	99.39%	0.9945	0.9953	0.9937
2	0.0199	0.0117	99.74%	0.9977	0.9953	1.000
3	0.0007	0.0162	99.39%	0.9945	0.9968	0.9921

The best performance was achieved at epoch 2, with an accuracy of 99.74% and an F1-score of 0.997, indicating excellent classification capability on the internal dataset. However, since 87.5% of the training corpus consists of template-generated synthetic sentences, these internal figures should be interpreted as a sanity check on training convergence rather than as an estimate of operational performance. The latter is established by the external evaluation reported in Section IV.D.

C. Performance visualization

Several internal evaluation visualisations were generated:

- Learning curves illustrating training and validation loss evolution.

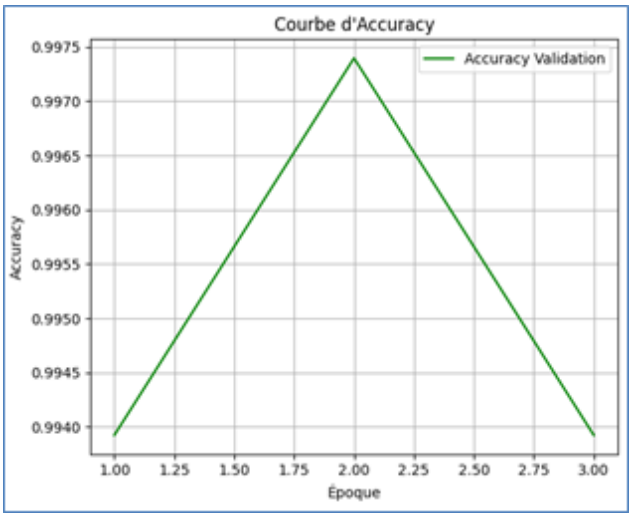

Figure 2: accuracy curve

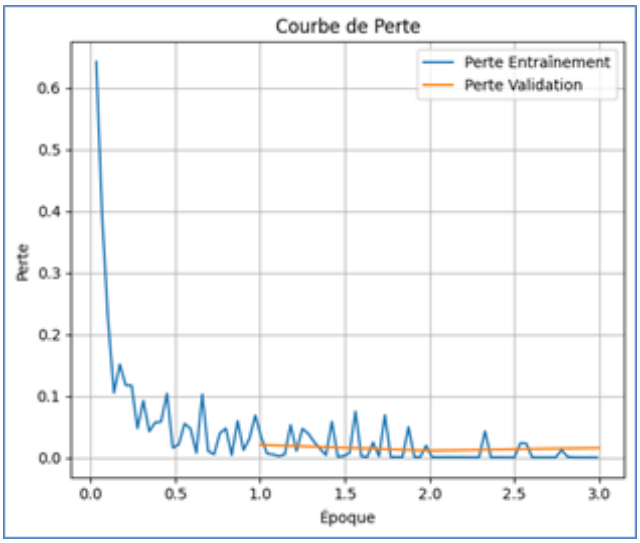

Figure 3: loss curve

- Confusion matrix highlighting classification errors.

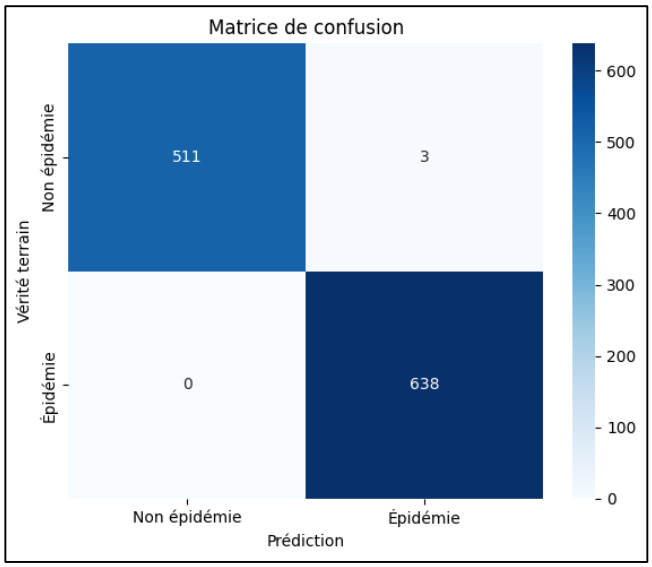

Figure 4: internal confusion matrix

- Classification report: detailed per-class metrics.

TABLE 2.
INTERNAL CLASSIFICATION REPORT.

Class	Precision	Recall	F1-score	Support
Non-épidémie	1.0000	0.9942	0.9971	514
Épidémie	0.9953	1.0000	0.9977	638
Accuracy			0.9974	1152
Macro avg	0.9977	0.9971	0.9974	1152
Weighted avg	0.9974	0.9974	0.9974	1152

These visualisations confirm the model's strong learning capacity while revealing performance gaps when applied to external data, motivating the external validation that follows.

D. External validation: primary results

To evaluate the generalisation capability of the model, an independent dataset of 997 manually annotated sentences was used as the primary benchmark. The dataset contains 656 non-epidemic and 341 epidemic sentences, constructed from a second scraping campaign three months after the training campaign, and contains no synthetic content. Performance metrics are reported with 1,000-iteration stratified bootstrap 95% confidence intervals (Table 3).

TABLE 3.
EXTERNAL EVALUATION OF CAMEMBERT WITH BOOTSTRAP 95% CONFIDENCE INTERVALS (N = 997, 1,000 BOOTSTRAP ITERATIONS, SEED = 42).

Metric (epidemic class)	Estimate	95% CI
Precision	0.701	[0.654, 0.744]
Recall	0.816	[0.774, 0.859]
F1	0.754	[0.717, 0.787]
PR-AUC (Average Precision)	0.699	[0.644, 0.756]
ROC-AUC	0.868	[0.842, 0.890]
Accuracy (overall)	0.818	[0.793, 0.842]

CamemBERT yields F1 = 0.754 [0.717, 0.787] and PR-AUC = 0.699 [0.644, 0.756] for the epidemic class. The classification report at the default threshold of 0.5 is given in Table 4.

TABLE 4.
CLASSIFICATION REPORT OF CAMEMBERT ON THE EXTERNAL TEST SET, DECISION THRESHOLD = 0.5.

Class	Precision	Recall	F1-score	Support
Non-epidemic	0.895	0.819	0.855	656
Epidemic	0.700	0.815	0.753	341
Accuracy			0.817	997
Macro avg	0.798	0.817	0.804	997
Weighted avg	0.828	0.817	0.820	997

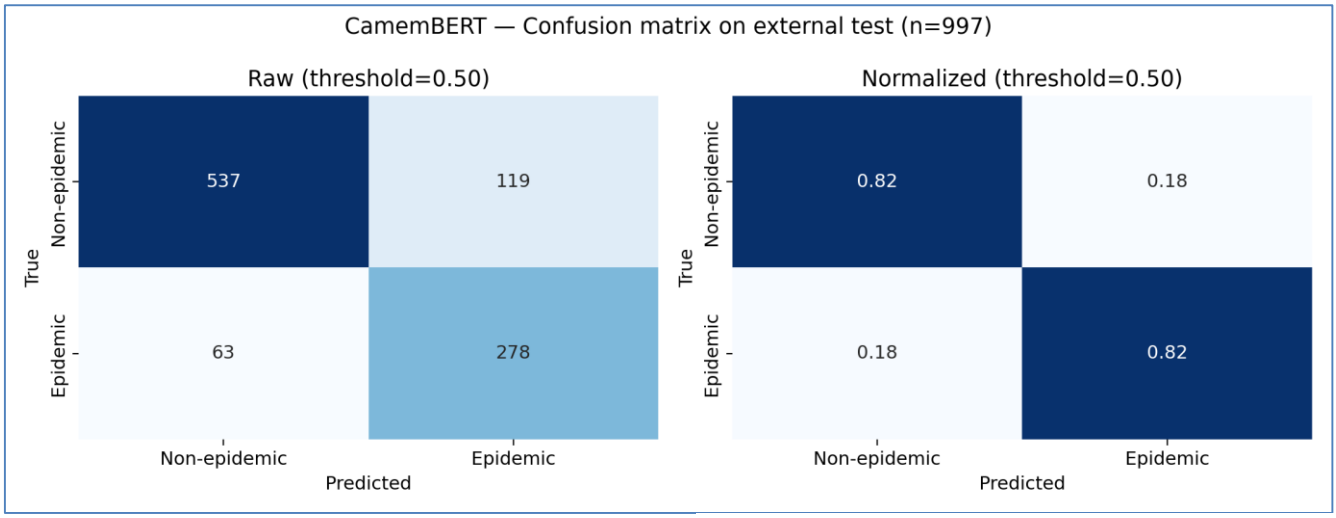

Figure 5. Confusion matrix of CamemBERT on the external test set (n = 997) at the default threshold of 0.5. Left: raw counts. Right: row-normalised.

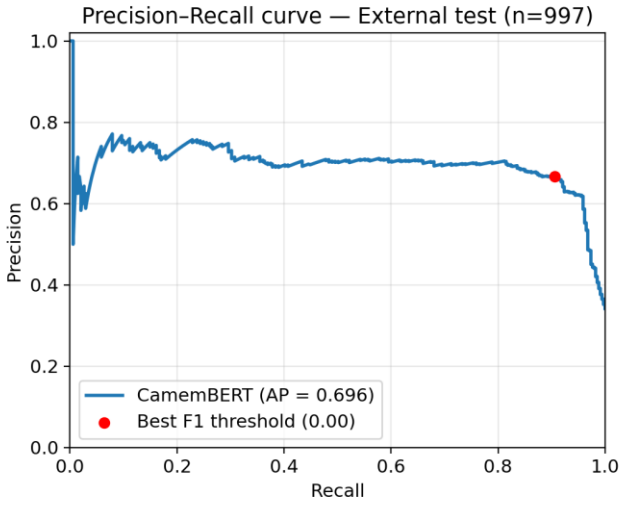

Figure 7. Precision–recall curve of CamemBERT on the external test set (n = 997). The red dot marks the F1-optimal operating point.

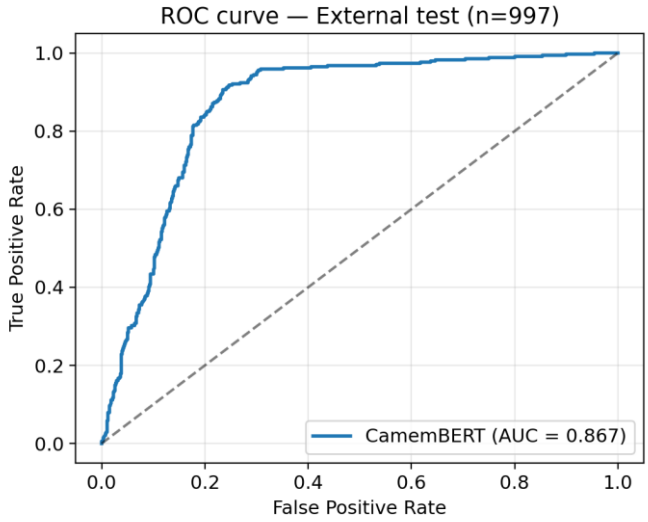

Figure 8. ROC curve of CamemBERT on the external test set (n = 997).

1). Decision threshold tuning

The threshold maximising F1 on the external set is 0.003, yielding F1 = 0.769, precision = 0.667, recall = 0.906. The very low optimal threshold reflects probability mis-calibration of the fine-tuned CamemBERT on the operational distribution: the model produces near-zero scores for the majority of true epidemic sentences yet still ranks them above non-epidemic ones (consistent with ROC-AUC = 0.868). For an early-warning use case where missing a true outbreak is

much costlier than a false alarm, we recommend operating at a recall ≥ 0.90 threshold. The confusion matrix at the optimal threshold is shown in Figure 6.

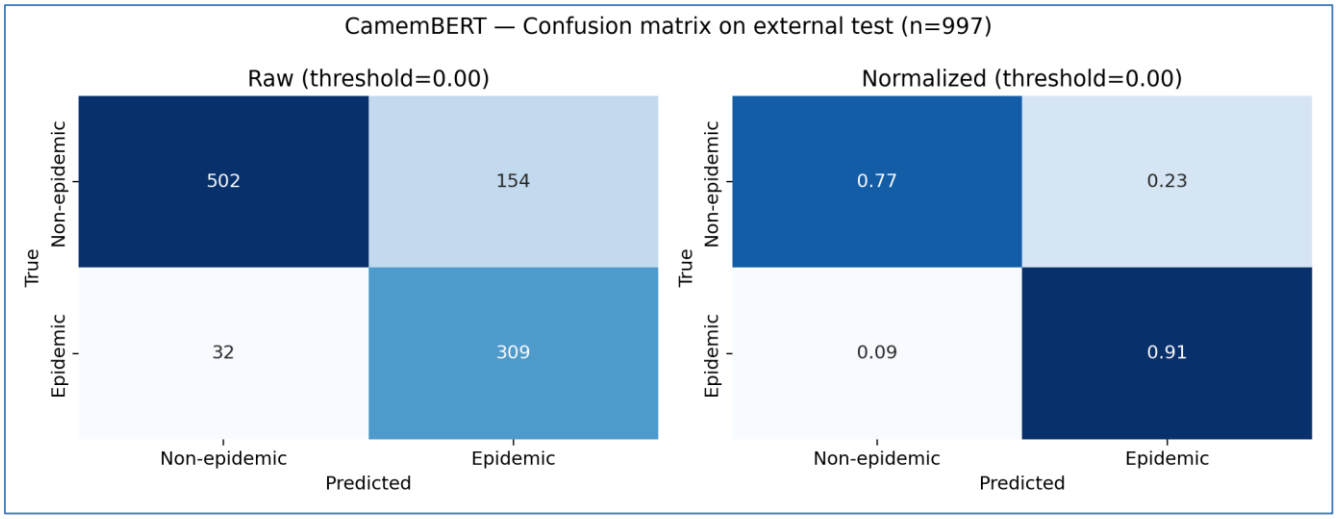

Figure 6. Confusion matrix of CamemBERT on the external test set at the F1-optimal threshold (0.003). Left: raw counts. Right: row-normalised.

2). Comparison with classical baselines

We compared CamemBERT against two standard lexical baselines on the same external corpus: TF-IDF (1-2 grams, sublinear TF) + Logistic Regression and TF-IDF + Linear SVM, both with balanced class weights. Two protocols are reported.

Protocol 1 (hold-out 70/30): the baselines were trained on 70% of the 997 sentences and evaluated on the remaining 30%. CamemBERT, having been fine-tuned on the original 11,433-sentence corpus, was evaluated as-is on the same 30% partition (Table 5).

TABLE 5.
BASELINE VS CAMEMBERT, HOLD-OUT 70/30 SPLIT (TRAIN = 697, TEST = 300; CLASS WEIGHTS BALANCED; STRATIFIED SPLIT, SEED = 42).

Model	Accuracy	F1 (Ep)	Precision (Ep)	Recall (Ep)	PR-AUC	ROC-AUC
TF-IDF + Logistic Regression	0.867	0.804	0.812	0.796	0.884	0.927
TF-IDF + Linear SVM	0.880	0.825	0.825	0.825	0.900	0.937
CamemBERT (transferred)	0.810	0.744	0.692	0.806	0.714	0.851

Protocol 2 (5-fold stratified cross-validation of the baselines on the full 997-sentence set), Table 6.

TABLE 6.
BASELINES, 5-FOLD STRATIFIED CROSS-VALIDATION ON THE 997-SENTENCE EXTERNAL CORPUS.

Model	F1 (Ep) mean ± std	PR-AUC mean ± std	Accuracy mean
TF-IDF + Logistic Regression	0.851 ± 0.043	0.912 ± 0.033	0.898
TF-IDF + Linear SVM	0.858 ± 0.037	0.926 ± 0.024	0.904

Finding. On the real-data-only external benchmark, the Linear SVM baseline outperforms the fine-tuned CamemBERT by approximately +10 points of F1 and +22 points of PR-AUC (Protocol 1), a result stable under 5-fold cross-validation (Protocol 2). The most plausible explanation is the lexical-shortcut effect induced by the dominance of templated synthetic data in the training corpus, which CamemBERT learns to recognise almost perfectly (internal F1 = 0.997) but which does not match the syntactic and lexical diversity of real Congolese news. The baselines, evaluated directly on real data, avoid this trap.

3). Error analysis

Inspection of the top false positives and false negatives reveals two recurring failure modes.

False positives, lexical shortcut on disease names (probability ≥ 0.999 on non-epidemic content):

- « Le premier cas de la Covid-19 en RDC a été détecté le 10 mars 2022, en provenance de la France. » (historical fact, p = 0.9996)
- « Les premiers cas de rougeole ont été enregistrés en juillet 2018. » (retrospective, p = 0.9996)

- « Aucun cas de choléra et de la rougeole n'a été enregistré depuis le début de cette année 2021. » (explicit negation ignored, p = 0.9995)
- « L'épidémie d'Ebola déclarée à Beni est la 15e depuis 1976 en RDC. » (recap of history, p = 0.9995)
- « L'Espagne connaît actuellement une seconde vague de coronavirus. » (outside DRC, p = 0.9995)

False negatives, elliptical or implicit phrasing (probability < 0.001 on real epidemic content):

- « Le gouvernement est en train de se procurer un centre de dépistage mobile […] pour freiner la propagation du virus. » (outbreak response without trigger words, p = 0.0005)
- « […] les efforts impressionnants enregistrés en RDC dans la lutte contre l'épidémie de la maladie à virus Ebola et la Covid-19 […] » (signal buried in complex syntax, p = 0.0005)
- « Dans ces conditions, le HCR estime que le dépistage rapide des cas suspects est essentiel. » (implicit, p = 0.0005)
- « Le Brésil dépasse le seuil des 150 000 morts. » (mortality without disease name, p = 0.0006)
- « La vaccination peut être proposée aux personnes qui ont déjà contracté la COVID-19. » (vaccination context, p = 0.0005)

These patterns confirm the lexical-shortcut hypothesis and motivate the human-in-the-loop validation discussed in Section V.

E. End-to-end system performance

To complement the static benchmark and demonstrate operational viability, the full pipeline (scraping → sentence segmentation → CamemBERT classification → CamemBERT-bio-GLiNER NER → alert generation) was executed on a single batch run against MediaCongo on 24 May 2026. The pipeline traversed the first three index pages of the Santé section, extracted all article links to the health category by contextual filtering, and processed each article sentence-by-sentence at the default 0.5 decision threshold. Results are summarised in Table 7.

TABLE 7.
END-TO-END PIPELINE RUN ON MEDIACONGO (24 MAY 2026).

Indicator	Value
Index pages traversed	3
Articles retrieved (health section)	30
Articles with successfully extracted body	30 (100%)
Sentences analysed	501
Sentences classified as epidemic	43 (8.6%)
Distinct articles producing ≥ 1 alert	15 / 30 (50%)
Mean per-article latency (download → classify → NER)	1.25 s
Total batch runtime	37.5 s
Mean classification confidence on alerts	0.977
Alerts with ≥ 1 disease entity	30 / 43 (69.8%)
Alerts with ≥ 1 location entity	28 / 43 (65.1%)
Alerts with ≥ 1 date entity	15 / 43 (34.9%)

The pipeline thus processes a typical news article in approximately one second on a single T4 GPU, demonstrating that the system can be operated continuously with negligible compute resources.

F. Longitudinal case study: Ebola Bundibugyo outbreak (May 2026)

The end-to-end run coincided with an active public-health emergency: the 17th Ebola outbreak in DRC, due to the Bundibugyo strain in Ituri Province, officially declared on 14 May 2026 and elevated to a WHO Public Health Emergency of International Concern on 17 May 2026.

Of the 30 articles retrieved on 24 May 2026, 27 (90%) directly addressed the outbreak. The pipeline generated 43 alerts spread across 15 distinct articles, jointly extracting the following key signals without any post-hoc filtering or human curation:

- Diseases: Ebola, Maladie à virus Ebola, Bundibugyo, fièvre hémorragique virale.
- Locations: Ituri, Nord-Kivu, Sud-Kivu, Bunia, Mongbwalu, Goma, Butembo, Mangina, Beni, Kinshasa, Kampala, Ouganda, Allemagne.
- Dates: 14 May, 15 May, 16 May, 17 May, 20 May, 2012 (historical Bundibugyo reference), 1976 (first DRC Ebola outbreak).
- Numbers correctly attached to alerts: 82 confirmed / 7 deaths; 246 suspected / 80 deaths; 513 suspected / 131 deaths; 118 deaths / 435 suspected cases.

The pipeline identified the cross-border dimension (Uganda, Kampala) and the imported case in Kinshasa (about 2,000 km from the epicentre) without any rule-based geographic logic. An interactive web-based dashboard visualises detected alerts in real time, facilitating monitoring and decision-making.

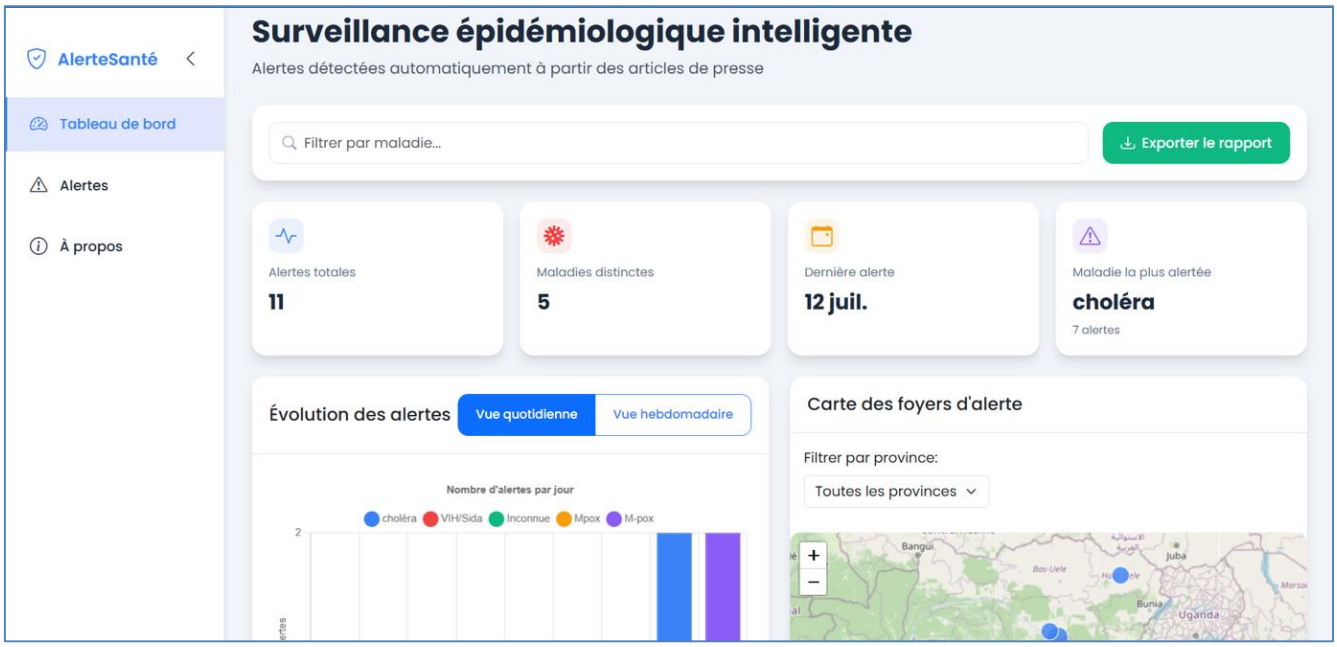

Figure 9: Web application dashboard interface

G. Discussion

The results obtained in this study confirm the technical feasibility and relevance of an automated epidemiological surveillance system based on the analysis of online news articles in the Democratic Republic of the Congo (DRC). The CamemBERT-based classification model, trained on a hybrid corpus combining manually annotated and synthetically generated data, achieved high performance on the internal validation set, with an accuracy of 99.74% and an F1-score of 0.997. However, the external validation reveals an overall accuracy of 82% and an F1-score of 0.75 for the epidemic class. This gap highlights the importance of evaluating machine learning models on unseen and fully real-world data. Despite this performance drop, the model maintains a relatively high recall (0.82) for the epidemic class, which is particularly critical in the context of public health surveillance, where missing true epidemic signals can have serious consequences. The lower precision (0.70) indicates the presence of false positives, suggesting that the system tends to favor sensitivity over specificity. This behavior can be acceptable in an early warning context, provided that downstream validation mechanisms (e.g., human review) are implemented.

Why is the transformer outperformed by classical baselines here? Three mechanisms plausibly explain this counter-intuitive result. First, synthetic-data dominance: with 87.5% of training sentences generated from templates, CamemBERT learns surface patterns that are absent from real Congolese news, leading to poor calibration on the operational distribution (optimal threshold ≈ 0.003). Second, domain shift: real news exhibits punctuation noise, abbreviations, code-switching with Lingala/Swahili terms and elliptical phrasings rarely captured by template generation. Third, class imbalance at evaluation time: the 34/66 epidemic/non-epidemic ratio of the external test penalises models that have learned to predict 'epidemic' too liberally.

Methodological implications. Our result is not an argument against transformer-based NLP, but for honest multi-baseline evaluation before claiming any operational benefit. Three concrete recommendations follow: (i) audit the synthetic-to-real ratio and report ablations when synthetic data exceeds real-data volume; (ii) calibrate decision thresholds on real data, because default 0.5 thresholds inherited from training-distribution validation can be deeply mis-calibrated; (iii) always include a TF-IDF + linear baseline, which costs minutes to train and can save months of misplaced engineering effort.

From a system perspective, the integration of the model into a Django-based web platform, orchestrated by Celery, enabled the implementation of a fully automated pipeline, from data collection to alert generation. The end-to-end run reported in Section IV.E demonstrated a per-article latency around one second on a single GPU, and the case study of Section IV.F shows that the system correctly captured the May 2026 Ebola Bundibugyo outbreak in Ituri at the sentence level, including its cross-border dimension.

H. Limitations

Several limitations should be acknowledged.

- The external benchmark of 997 sentences was annotated by a single rater; inter-annotator agreement was not formally measured in the present version.
- Only French is processed; Lingala, Kiswahili, Kikongo and Tshiluba content is ignored although it represents an important share of grassroots public-health discourse.
- Urban areas and large provincial capitals are over-represented; eastern conflict-affected provinces are under-covered.

- A controlled ablation varying the proportion of synthetic content in the training corpus would provide the cleanest demonstration of the lexical-shortcut hypothesis; this experiment is left to a follow-up release that will rely on a regenerated corpus produced with the same template scripts and a fresh annotation round.
- CamemBERT's output probabilities are poorly calibrated on the operational distribution; Platt or temperature scaling on a held-out real sample is a low-cost future improvement.
- Lack of cross-validation with official sources: the absence of systematic comparison with institutional data sources such as national health systems or DHIS2 may limit trust in the generated alerts.

V. CONCLUSION

This study presents an end-to-end pipeline for early detection of epidemic signals from Congolese French-language online news, integrating targeted web scraping, sentence-level classification with a fine-tuned CamemBERT transformer, zero-shot biomedical NER (CamemBERT-bio-GLiNER) and an interactive web dashboard built on a Django/Celery stack.

On an independent external benchmark of 997 manually annotated sentences, the fine-tuned CamemBERT classifier achieves F1 = 0.754 [0.717, 0.787] and PR-AUC = 0.699 [0.644, 0.756] for the epidemic class, while a simple TF-IDF + Linear SVM baseline reaches F1 = 0.858 ± 0.037 in 5-fold cross-validation, outperforming the transformer. This comparison, rarely conducted in prior epidemic-NLP literature, highlights that transformer fine-tuning is not automatically beneficial when training data are dominated by synthetic content. An operational run of the full pipeline on MediaCongo processed 30 articles and 501 sentences in 37.5 s on a single GPU, producing 43 alerts that correctly captured the May 2026 Ebola Bundibugyo outbreak in Ituri, including its cross-border dimension.

The contributions of this work are therefore (i) an operational pipeline deployable in resource-constrained settings, (ii) an external benchmark released alongside this paper, and (iii) a methodological warning about the perils of synthetic-data dominance. Future work includes a regenerated training corpus enabling controlled ablations, integration with national health information systems (DHIS2), automated alert channels (SMS, messaging platforms), and enhanced human-in-the-loop validation.

REFERENCES

[1] J. S. Brownstein, C. C. Freifeld, and L. C. Madoff, “Digital disease detection — Harnessing the Web for public health surveillance,” New England Journal of Medicine, vol. 360, no. 21, pp. 2153–2155, 2009, doi: 10.1056/NEJMp0900702.

[2] Celery Project, “Celery: Distributed task queue,” 2023. [Online]. Available: https://docs.celeryq.dev/

[3] C. M. Wolfe et al., “Systematic review of Integrated Disease Surveillance and Response (IDSR) implementation in the African region,” PLoS ONE, vol. 16, no. 2, e0245457, 2021, doi: 10.1371/journal.pone.0245457.

[4] Django Software Foundation, “Django documentation,” 2023. [Online]. Available: https://docs.djangoproject.com/

[5] J. Ginsberg et al., “Detecting influenza epidemics using search engine query data,” Nature, vol. 457, no. 7232, pp. 1012–1014, 2009, doi: 10.1038/nature07634.

[6] J. Hong et al., “Relation extraction from news articles (RENA): A tool for epidemic surveillance,” arXiv preprint arXiv:2311.01472, 2023. doi: 10.48550/arXiv.2311.01472.

[7] Hugging Face, “CamemBERT model documentation,” 2023. [Online]. Available: https://huggingface.co/camembert-base

[8] J. Miano et al., “Using event-based web-scraping methods and bidirectional transformers to characterize COVID-19 outbreaks in food production and retail settings,” in Artificial Intelligence in Medicine (AIME 2021), Lecture Notes in Artificial Intelligence, vol. 12721, pp. 187–198, Springer, 2021, doi: 10.1007/978-3-030-77211-6_21.

[9] G. Lejeune, R. Brixtel, A. Doucet, and N. Lucas, “Multilingual event extraction for epidemic detection,” Artificial Intelligence in Medicine, vol. 65, no. 2, pp. 131–143, 2015, doi: 10.1016/j.artmed.2015.06.005.

[10] L. Martin et al., “CamemBERT: A tasty French language model,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7203–7219, 2020, doi: 10.18653/v1/2020.acl-main.645.

[11] World Health Organization, “Rapport sur les épidémies et la couverture sanitaire en Afrique centrale,” Bureau régional de l’OMS pour l’Afrique, 2021. [Online]. Available: https://www.afro.who.int

[12] D. Phutane et al., “Predicting future trends in disease outbreaks using web scraping and machine learning,” Veermata Jijabai Technological Institute, Mumbai, India, 2025. [Online]. Available: https://www.researchgate.net/publication/392892557_Predicting_Future_Trends_in_Disease_Outbreaks_Using_Web_Scraping_and_Machine_Learning_5_th_Druhi_Phutane

[13] L. Richardson, “Beautiful Soup documentation,” 2023. [Online]. Available: https://www.crummy.com/software/BeautifulSoup/

[14] T. Wolf et al., “Transformers: State-of-the-art natural language processing,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45, 2020, doi: 10.18653/v1/2020.emnlp-demos.6.

[15] Almanach, “CamemBERT-bio-gliner-v0.1: Zero-shot French biomedical NER model based on GLiNER with CamemBERT-bio backbone,” Hugging Face model card, 2025. [Online]. Available: https://huggingface.co/almanach/camembert-bio-gliner-v0.1.” 2024.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), pp. 1135–1144, 2016, doi: 10.1145/2939672.2939778.