The Indonesian Journal of Computer Science
www.ijcs.net
Volume 15, Issue 5, October 2026
https://doi.org/10.33022/ijcs.v15i5.5231

HIVMID: A Multilingual and Multimodal Dataset for HIV/AIDS Misinformation Detection in English and isiZulu

Rose Skhosana[1], Bassey Isong[2]
32289340@mynwu.ac.za[1], bassey.isong@nwu.ac.za[2]
[1,2]Computer Science Department, North-West University, Potchefstroom, South Africa

Article Information

Received : 16 Aug 2026
Revised : 21 Sep 2026
Accepted : 23 Okt 2026

Keywords

HIV/AIDS misinformation; dataset construction; low-resource languages; isiZulu; multimodal misinformation detection; health misinformation.

Abstract

Digital health misinformation can affect health outcomes and informed decision-making, yet resources for African low-resource languages (LRLs) and HIV/AIDS misinformation remain scarce. Existing health misinformation benchmarks also focus largely on English and COVID-19-related content. To address this gap, we developed a multilingual, multimodal dataset covering English and isiZulu. The dataset was constructed through a six-stage methodology informed by the Datasheets for Datasets and Data Statements for Natural Language Processing frameworks and contains 1,147 records (673 factual and 474 misinformation) from 92 normalised sources. It covers five thematic categories, two languages, and both text-only and text-plus-image modalities. Six classical machine learning classifiers were evaluated using TF-IDF representations, stratified cross-validation, and an independent test set. The best-performing model, Random Forest with word-level TF-IDF, achieved 93.45% accuracy, a 93.49% weighted F1-score, and a Matthews Correlation Coefficient of 0.874, substantially outperforming a majority-class baseline while maintaining comparable performance across both language subsets. The paper introduces the Health Misinformation Dataset Construction Framework (HMD-CF), which generalises the dataset construction process into a transferable methodology for multilingual health misinformation resources. The proposed dataset provides one of the first benchmark resources for multilingual HIV/AIDS misinformation detection(MID) in an African LRL setting and supports future research on health MID, multimodal learning, and LRL technologies.

Attribution-ShareAlike 4.0 International License

A. Introduction

Misinformation on digital platforms is widely recognised as a threat to public trust, informed decision-making, and, in the health domain, to human life [5], [6]. The World Health Organisation (WHO) has described the overabundance of accurate and false information accompanying major health events as an "infodemic" [28], and false claims about prevention, transmission, and treatment continue to circulate through social media, blogs, and interpersonal networks in both high-resource and low-resource-language (LRL) settings. HIV/AIDS misinformation is a particularly consequential case. Testing, treatment compliance, and preventive behaviour are all directly undermined by myths about HIV transmission, unproven treatments, and stigmatising statements about those living with HIV. These effects are amplified in communities where health information is not available in the local language.

Despite this risk, dataset resources for misinformation detection (MID) remain concentrated in English and in a narrow set of health topics, principally COVID-19 and vaccine-related claims. A recent systematic review of 51 peer-reviewed MID studies published between 2021 and 2026, Isong and Linah [19] applied the PRISMA 2020 protocol across LRL, health-domain, multimodal, and explainability dimensions, found that new corpora have been constructed for languages including Amharic, Bangla, Bengali, Luganda, Sepedi, Sesotho, Xitsonga, isiZulu, Kurdish Sorani, and several Arabic dialects [19]. The same review reports that health-domain coverage is narrow: research concentrates overwhelmingly on COVID-19 and vaccine misinformation, while HIV, malaria, and reproductive health appear in none of the 51 reviewed studies [19]. This absence is consistent with a broader pattern identified in surveys of LRL MID, which note that annotated datasets and detection tools for African low-resource languages (LRLs) remain scarce relative to the scale of the problem [11].

Dataset construction has therefore become an essential precursor to modelling work in this space, particularly for languages where publicly available annotated resources remain scarce [1], [2]. Prior South African efforts have demonstrated that web-scraped and translation-augmented corpora can support misinformation and hate-speech detection in Tshivenda, Sepedi, Sesotho, Xitsonga, and isiZulu [7], [20], [25], [26], [27], but none of these efforts addresses HIV/AIDS content, and none combines text with accompanying imagery to support multimodal detection research. This paper addresses that combined gap.

This study makes five contributions. First, it presents a six-stage methodology for constructing a multilingual and multimodal HIV/AIDS misinformation dataset for English and isiZulu, grounded in the Datasheets for Datasets [2] and Data Statements for Natural Language Processing (NLP) [4] frameworks. Second, it provides a statistical characterisation of the resulting 1,147-record dataset across label, category, modality, language, source, and text-length dimensions. Third, it establishes imbalance-aware, hyperparameter-tuned classical machine learning baseline benchmarks using word- and character-level TF-IDF representations across six classifiers, evaluated through cross-validation and an independent test set, checked against a majority-class floor baseline and paired significance tests. Fourth, it positions the dataset explicitly against the gap identified in recent LRL and health-domain MID literature, providing, to the

authors' knowledge, one of the first multilingual and multimodal HIV/AIDS misinformation resources for African LRLs. Fifth, it generalises the six-stage methodology into the Health Misinformation Dataset Construction Framework (HMD-CF), a reusable reference process, with an explicit transferable design decision at each stage, that a researcher could apply to a different disease, language, or region.

The remainder of the paper is structured as follows. Section 2 reviews related work on LRL and health-domain MID and on dataset documentation frameworks. Section 3 describes the dataset construction methodology and the baseline classification protocol. Section 4 presents the resulting dataset statistics and baseline classification results. Section 5 discusses the findings in relation to the broader literature and outlines limitations and future work. Section 6 proposes a reusable construction framework and demonstrates its application to a different ailment and language. Section 7 concludes the paper.

B. Related Works

B.1. MID in LRLs and health domains

MID has evolved from traditional machine learning classifiers like Naïve Bayes, Support Vector Machines, and Random Forest to deep learning and transformer-based architectures that capture contextual and semantic links beyond simple lexical features. However, more than two-thirds of published MID studies continue to focus on English-language content, leaving speakers of African and other LRLs significantly underserved [19]-[21]. Where LRL corpora exist, they have typically been constructed through one of two routes: native web scraping and manual annotation, as in the Tshivenda misinformation dataset built by Mukwevho et al. [7], or machine translation of existing high-resource corpora into the target language, as applied to the COVID-19 datasets of Sepedi [25], Sesotho [26], and Xitsonga [27]. Translation-based techniques typically identify translation quality as a bottleneck, with discrepancies between translation engines (e.g., Google Translate and Helsinki-NLP's Opus-MT) influencing downstream classifier performance [19], [26].

For isiZulu specifically, prior South African work has concentrated on adjacent tasks rather than misinformation directly. Sethosa et al. [20] translated an English hate-speech dataset into isiZulu and refined it with native speakers, training TF-IDF-based classifiers without an explainability component. No prior published dataset, to the authors' knowledge, provides isiZulu-language annotations specifically for HIV/AIDS misinformation, or pairs isiZulu text with a parallel English subset to support cross-lingual transfer experiments.

Health-domain MID research is dominated by COVID-19 and vaccine misinformation. Purpose-built datasets such as CoAID, FakeHealth, and MMCovid19 emerged during the pandemic [19], and dedicated vaccine-misinformation corpora such as ANTi-Vax [22] and the Constraint@AAAI2021 dataset used to fine-tune COVID-Twitter-BERT [23] have become standard benchmarks. A smaller body of work addresses non-COVID-19 health misinformation, including diabetes, maternal and neonatal health, and autism-related forum content, but these studies remain isolated rather than systematic, and a recent review of health MID identifies data scarcity and the near-absence of

explainable AI as the field's two primary unresolved gaps [24]. Despite its documented public health consequences, HIV/AIDS misinformation has not previously been the subject of a dedicated, annotated, multilingual dataset.

B.2. Dataset documentation and construction frameworks

Because LRL and health-domain corpora are often built from scratch, transparent documentation of provenance, collection procedure, and intended use has become a central methodological concern. In order to make dataset limits obvious to downstream users, the Datasheets for Datasets architecture [2] suggests a standard set of questions addressing rationale, composition, collection process, and suggested uses. Detailed reporting of language resources, curator demographics, and annotation guidelines is central to the Data Statements for NLP framework [4], which links these practices to reducing system bias and improving reproducibility. Since then, related guidelines have been expanded to domain-specific contexts, such as digital cultural heritage databases, where versioning, provenance, and responsible reuse are considered essential requirements rather than optional extras [3]. These principles informed the dataset development approach used in this study. Structured metadata that facilitates transparent reuse and replication is included with the final dataset.

B.3. Multimodal and multilingual MID

Text-only detection remains the dominant paradigm in the literature. Text-only detection is still the dominant paradigm in the literature; recent taxonomies of the field show that text-only approaches account for the vast majority of evaluated research, with multimodal (text-plus-image) and audio/video approaches accounting for a small minority [19]. Where multimodal fusion has been examined, it has consistently outperformed unimodal baselines in detection performance, especially on platforms like Facebook, Instagram, and TikTok, where disinformation is routinely disseminated through image-text combinations [13], [15], [16]. Multilingual transformer designs, such as mBERT and XLM-RoBERTa, have shown good cross-lingual transfer in comparable LRL misinformation tasks and are seen as a natural next step once a labelled multilingual corpus is available [11, 17]. The dataset presented in this paper is explicitly designed to support both directions: roughly half of its records pair text with an associated image, and its parallel English-isiZulu structure is intended to enable future cross-lingual and multimodal transformer experiments of the same type.

The reviewed literature indicates a consistent structural gap. Existing MID datasets do not address HIV/AIDS disinformation, there are no isiZulu-specific misinformation resources, and multimodal multilingual corpora for African LRLs are rare. The dataset construction process and baseline results reported in Sections 3 and 4 directly address this gap.

C. Research Method

This study developed a multilingual and multimodal HIV/AIDS misinformation dataset for LRLs. Dataset building has become an important part of NLP research, especially for LRLs where publicly available annotated datasets are rare [1], [2]. This procedure was influenced by two well-known NLP dataset creation

frameworks. While the Data Statements framework for NLP [4] pushes further into linguistic detail, requiring precise reporting on language resources so that results can be repeated and AI research is conducted responsibly, Datasheets for Datasets [2] asks researchers to fully document a dataset's features, collection methods, intended use, and limitations. Later work extended both frameworks into domain-specific territory. Provenance, versioning, and reuse now receive the same documentation treatment outside NLP proper, in datasets built for other fields entirely [3]. Six sequential stages make up the construction process used here, shown in Figure 1.

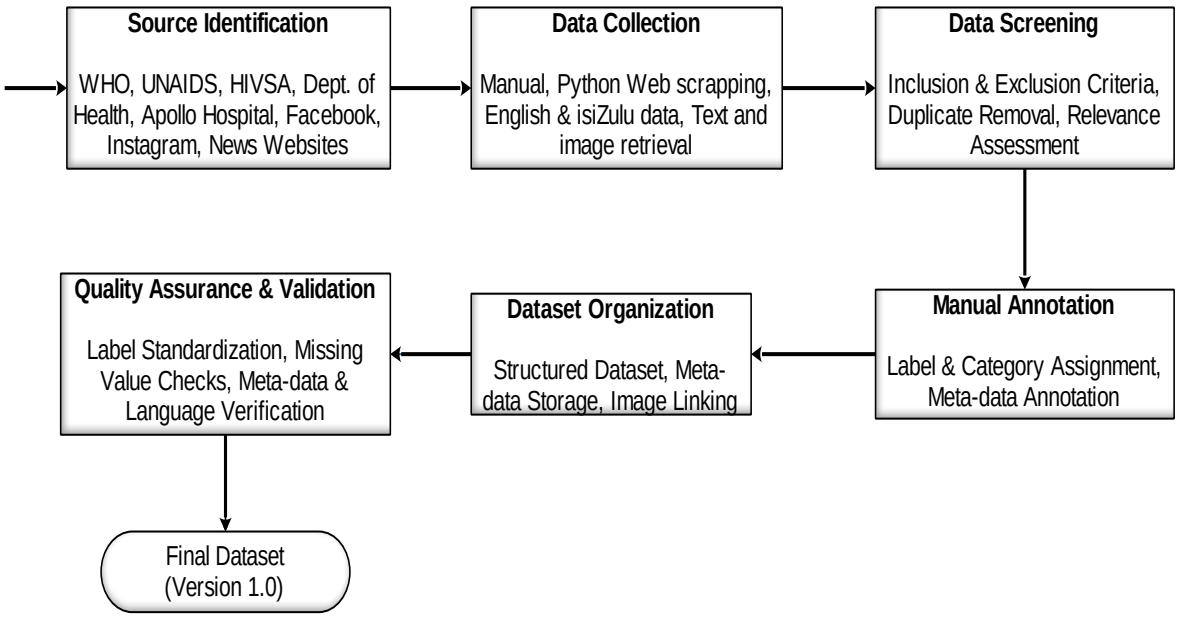


Figure 1. Overall dataset construction methodology

C.1. Source identification

The initial step was to identify reliable sources. Due to the rapid dissemination of false information via digital media, the search had to encompass both the public discourse where false information genuinely circulates as well as health information from reputable sources. Source diversity is important because, as previous research on disinformation collection has shown, it is necessary to draw from a variety of sources in order to capture how misinformation genuinely varies across low-resource and health-related sectors. In particular, social media speeds this up after public health events [5]. Public health communication in general has more challenges as a result of the spread [6].

We selected internationally recognised organisations as sources of factual, evidence-based content aligned with established medical guidelines. The factual sources comprised the World Health Organisation, UNAIDS, the South African National Department of Health, HIVSA, and Apollo Hospitals. Misinformation was instead drawn from Facebook and Instagram posts, blogs, online discussion boards, and news articles containing HIV/AIDS-related claims. Including these different source types allowed the dataset to reflect the range of writing styles,

communication formats, and misinformation patterns found in online environments.

C.2. Data collection

The goal of the data collection method was to compile information regarding HIV/AIDS that was both written and visual. The study addressed the limited availability of disinformation datasets for African LRLs by using a multilingual method, collecting data in both English and isiZulu.Data collection was performed through a combination of manual collection and web-based retrieval techniques implemented in Python. Similar approaches have been adopted in previous studies that constructed misinformation datasets for South African LRLs using web crawling and web scraping techniques [7]. Relevant HIV/AIDS information was identified using topic-specific search queries related to prevention, treatment, education, myths, stigma, HIV testing, antiretroviral therapy (ART), mother-to-child transmission, pre-exposure prophylaxis (PrEP), post-exposure prophylaxis (PEP), and HIV transmission.

To aid future multimodal MID studies, images that accompanied textual content were downloaded when accessible and connected to the related textual data. To provide traceability during the annotation and validation process, each gathered record was given a unique identification. The data collection pipeline is shown in Figure 2.

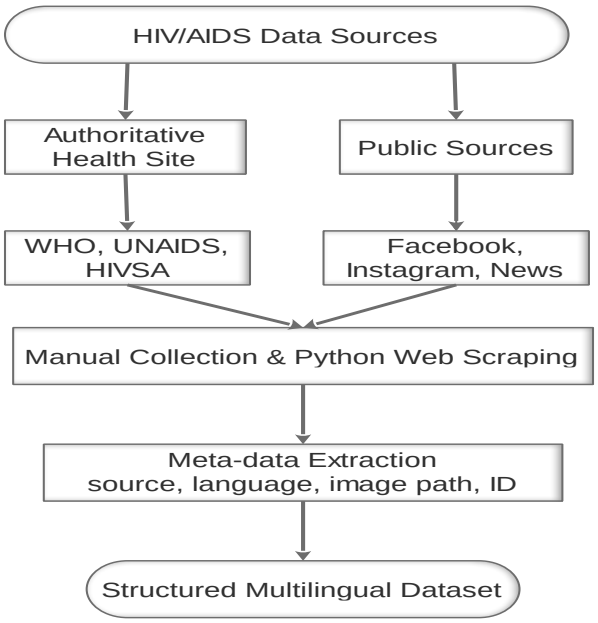


Figure 2. Data collection pipeline

C.3. Data screening

Before annotation, all collected records underwent a screening process to improve dataset quality. As in previous misinformation dataset studies [7], [8], inclusion and exclusion criteria were established to improve dataset consistency. Records were included if they contained HIV/AIDS-related information and were written in English or isiZulu. They also had to represent either factual information

or misinformation, originate from publicly accessible online sources, and contain textual information. Accompanying images were included where available.

Records were excluded if they were duplicated, contained incomplete or unreadable information, or were unrelated to HIV/AIDS. Records that consisted primarily of advertisements or promotional material were also excluded. Records that lacked crucial metadata needed for annotation or contained private or personally identifiable information were also eliminated. By using these standards, the dataset's consistency was enhanced, and only pertinent records were included.

C.4. Annotation guidelines

After screening, each record was manually annotated using predetermined guidelines. Manual annotation remains a reliable approach for producing high-quality supervised datasets in specialised fields such as healthcare [9], and reviews of low-resource MID similarly report that trustworthy detection models depend on manually annotated data [17].

Each record received two classifications. First, it was assigned a binary label: Factual, referring to evidence from authoritative health agencies, or Misinformation, referring to false, deceptive, or scientifically unfounded claims. Second, it was assigned to one of five thematic categories identified from the HIV/AIDS misinformation literature: Prevention (HIV prevention methods), Education (general HIV/AIDS knowledge and awareness), Treatment (antiretroviral therapy (ART), medication, viral load, CD4 count, and HIV management), Myth (false medical claims about transmission, testing, pregnancy, treatment, or prevention), and Stigma (claims promoting stereotypes, discrimination, or social exclusion of people living with HIV). These categories were selected to represent the misinformation themes observed in the collected data.

Figure 3 presents the resulting workflow. Language, modality, source, and image path were also recorded as metadata for each entry.

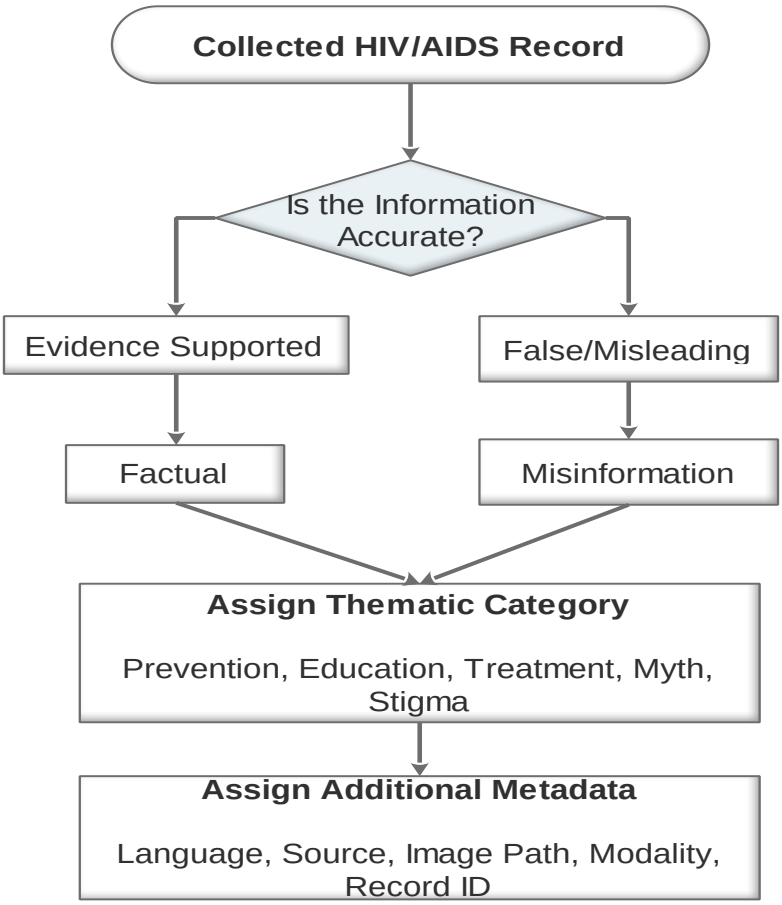


Figure 3. Annotation workflow

C.5. Dataset organisation

After annotation, the dataset was transported to an organised Excel table. Each record contains eight attributes, which are summarised in Table 1, and were particularly chosen to facilitate machine learning experimentation and reproducibility. Structured metadata of this nature is not coincidental. Current dataset documenting frameworks view it as critical to transparency, repeatability, and future reuse [3].

Table 1. Dataset organisation.

Attribute	Description
ID	Unique record identifier
Text	HIV/AIDS-related textual content
Image Path	File path to the associated image
Label	Factual or Misinformation
Category	Prevention, Education, Treatment, Myth, or Stigma
Modality	Text or Text + Image
Language	English or isiZulu
Source	Original source of the collected record

C.6. Quality assurance and dataset validation

Before the dataset was finalised, its consistency, completeness, and usability were examined. Repeated text inputs were found by manual examination and eliminated to stop duplicate records from influencing evaluation or training. The labels were then standardised: similar category names, such as Myth_Busting/Myth and Stigma_Awareness/Stigma, were combined into the five standardised categories, while capitalisation variations, such as "factual"/Factual and "misinformation"/Misinformation, were mapped to two binary classes.

The remaining checks addressed incomplete metadata, picture links, language assignments, and missing values. Language assignments were confirmed as either isiZulu or English, and records with unrecoverable missing values were eliminated instead of inferred. In order to facilitate traceability, picture paths for multimodal records were compared to the matching text, and the source, modality, category, and unique identifier fields were validated for completeness. Following these verifications, the dataset was frozen as Version 1.0 for the subsequent baseline studies and descriptive statistics. This release meets the requirements for responsible dataset publication, including the recording of provenance, quality, and version history [3]. Inter-annotator agreement remains pending. In a later iteration, a representative subset will be labelled by a second independent annotator, with Cohen's Kappa [9], [10] used to measure agreement beyond chance and provide an additional check on annotation consistency.

C.7. Ethical considerations

This study solely gathered publicly available HIV/AIDS-related content from online sources. The dataset did not include any private accounts, password-protected data, or personally identifying information. To ensure user privacy, identifiers were eliminated wherever possible. The dataset is only for research purposes and is accompanied by documentation indicating its origin, collection procedure, intended use, and limits in order to promote transparency and responsible reuse, in accordance with current recommendations for responsible dataset publication [2, 3].

C.8. Baseline classification

Before the findings presented in Section 4.2, a baseline classification algorithm was created and used to show that the dataset supports supervised MID. In order to report results without re-deriving the experimental design alongside them, this section details the entire methodology, including the computational environment, model selection, feature representation, class-imbalance handling, the train/test and cross-validation protocol, hyperparameter tuning, evaluation metrics, and the statistical testing protocol.

C.8.1 Experimental setup

The studies were carried out on a Windows 11 64-bit machine with an AMD Ryzen 8-core processor (4.10 GHz), 32 GB of RAM, and a 466 GB hard drive. A fixed random seed of 42 was used for data splitting, cross-validation, and classifier initialisation in all experiments, which were conducted in Python 3.14 in a different virtual environment. XGBoost (3.2.0), SciPy (1.17.1), joblib, NumPy

(2.4.4), pandas (3.0.2), scikit-learn (1.8.0), and Matplotlib were utilised in the analysis.

TF-IDF vectorisation, pipeline building, classifier implementation, GridSearchCV-based hyperparameter tuning, class-imbalance utilities, and evaluation metrics were all handled by Scikit-learn. Logistic Regression, Linear SVM, Random Forest, Gradient Boosting, Complement Naïve Bayes, and the majority-class dummy baseline were among the classifiers. The gradient-boosted tree classifier was supplied by XGBoost, while the paired t-test, Wilcoxon signed-rank test, and McNemar's exact test were supported by SciPy. The best-performing pipeline and label encoder were persisted using Joblib, while data manipulation, results storage, and figure creation were supported by NumPy, pandas, and Matplotlib. The reproducibility documentation that comes with the analysis code documents the exact package versions, the whole hyperparameter search area, and the winning configuration for each classifier.

C.8.2 Feature representation and model selection

Word-level and character-level TF-IDF representations were employed in the studies. Word TF-IDF included both unigrams and bigrams, with sublinear term-frequency scaling and a minimum document frequency of two. Using the same frequency and scaling parameters, Character TF-IDF employed 3-5 character n-grams inside word boundaries. Given isiZulu's wide morphological variety as an agglutinative language, the character representation was added because it can capture prefixes, suffixes, and incomplete word patterns that word-level features could overlook.

Six classification algorithms were selected to span the two families of approach most commonly used as baselines in the MID literature: linear/probabilistic classifiers (Naïve Bayes, Logistic Regression, Linear Support Vector Machine) and ensemble tree-based classifiers (Random Forest, Gradient Boosting, Extreme Gradient Boosting). This spread was chosen to establish whether the dataset's signal is exploitable by markedly different inductive biases, rather than being an artefact learnable only by one model family, and to keep the baseline evaluation reproducible on standard hardware without requiring a GPU; transformer-based architectures are treated as a separate, subsequent evaluation stage and are discussed under Future Work (Section 5.2) rather than included here.

C.8.3 Class imbalance handling

Given the label distribution reported in Section 4.1.1 (58.7% Factual, 41.3% Misinformation), an imbalance-handling mechanism was applied to every classifier rather than a subset of them. Logistic Regression, Linear SVM, and Random Forest used balanced class weights. Gradient Boosting, which has no native class-weighting parameter, used per-sample weights computed via balanced class weighting and passed at fit time. XGBoost used a positive-class scale factor computed from the training split's class ratio. Naïve Bayes was implemented as Complement Naïve Bayes rather than the standard multinomial form, since multinomial Naïve Bayes has no imbalance-handling mechanism and Complement

Naïve Bayes is the variant generally recommended for imbalanced text classification.

C.8.4 Train/test split and cross-validation protocol

Records with missing text or an unrecognised label were dropped before splitting, leaving the 1,144-record usable subset. This subset was partitioned into training and test sets using an 80/20 split stratified jointly by language and label (English-Factual, English-Misinformation, isiZulu-Factual, isiZulu-Misinformation), with a fixed random seed, yielding 915 training records and 229 test records. Five-fold stratified cross-validation was performed on the training partition only. The test set was held out and untouched until the single final evaluation reported in Section 4.2.3, so that no test-set information influenced model or hyperparameter selection.

However, one limitation of this split should be stated explicitly rather than left implicit. The split stratifies by language and label but does not group records by underlying content, so an English record and its isiZulu translation could in principle fall on opposite sides of the train/test boundary. The current schema (Table 1) does not carry a field linking a translated record to its English source record, so this cannot be verified or corrected retroactively for the results reported here. The near-identical accuracy observed across English and isiZulu in Section 4.2.6 should therefore be treated as provisional pending re-evaluation on a content-grouped split. A translation-pair identifier is planned for the next dataset version specifically to enable that check, and is treated as a required fix, not an optional one, before the cross-lingual result is relied upon in follow-up work.

C.8.5 Hyperparameter tuning protocol

Each classifier was tuned via grid search over a small, pre-specified grid rather than run with fixed default values. Logistic Regression and Linear SVM were tuned on regularisation strength (C). Random Forest on tree count and maximum depth; Gradient Boosting and XGBoost on tree count, learning rate, and maximum depth; and Naïve Bayes on the smoothing parameter (α). Each grid was scored using the same five-fold cross-validation described. The winning configuration was then refit on the full training partition before test-set evaluation. The search was not performed within an outer cross-validation loop. This was a deliberate design decision for the baseline benchmark study, rather than an error. A more conservative approach would have nested the search and fine-tuning within an outer cross-validation loop, but at substantially higher computational cost. The trade-off is stated explicitly rather than left implicit in the reported results. The same fixed random seed was used for each stochastic component. Exact package versions, complete hyperparameter grids, and the optimal configuration for each classifier are recorded in the reproducibility manifest exported with the analysis code.

C.8.6 Evaluation metrics

Classification performance was evaluated using six metrics. Accuracy, precision, recall, and weighted F1-score provide the conventional summaries, while macro F1-score gives equal weight to both classes regardless of their

frequency. ROC-AUC evaluates discrimination across all classification thresholds rather than at a single operating point. MCC was included because of the class imbalance and treats the four confusion-matrix cells symmetrically, reducing the inflation that weighted F1-score and accuracy may exhibit under imbalance [30]. Precision-Recall AUC, anchored on the Misinformation class, provides a more targeted measure of how well a model ranks misinformation above factual content. This is important because Section 4.2.4 identifies missed misinformation as the more costly error than a false alarm.

For the same reason that the ROC curve shows instead of the AUC value alone, a single figure obscures how performance varies between thresholds, the analysis code also creates the analogous curve for the best-performing model. All of this is grounded by another comparison. The modified classifiers were evaluated alongside a majority-class dummy baseline that predicted the more common label regardless of text content. This provided a performance floor against which any improvement could be interpreted as a genuine gain rather than an artefact of class imbalance.

C.8.7 Statistical significance testing protocol

Several configurations in Section 4.2 achieved very similar point-estimate performance. As a result, ranking models by mean score alone could overstate differences that are not distinguishable from fold-to-fold or sample-to-sample variation. Two complementary significance tests were therefore applied to the top-2 cross-validation configurations by mean weighted F1. First, a paired t-test and a Wilcoxon signed-rank test were computed across the five matched cross-validation folds (matched because both configurations were scored on the identical fold partition, given the fixed random seed described in Section 3.8.4), testing whether the two configurations' per-fold scores differ systematically rather than by chance, at a significance threshold of $\alpha = 0.05$. Second, McNemar's exact test was computed on the held-out test-set predictions of the same two configurations, testing whether their disagreements (cases where one configuration is correct and the other is not) are asymmetric beyond what chance would produce. Together, these tests distinguish a configuration that is reliably better from one that merely ranks higher on a single point estimate.

D. Results

D.1 Dataset Analysis

The developed multimodal HIV/AIDS misinformation dataset consists of 1,147 samples collected from multiple credible and publicly available online sources, of which 673 are labelled factual and 474 are labelled misinformation. The dataset contains both text-only and text-image samples, making it suitable for evaluating both unimodal and multimodal detection models. The dataset grew from 746 to 1,147 samples through the addition of 401 isiZulu translations of existing entries (203 factual and 198 misinformation); the English subset remained unchanged from the original 746-sample dataset. Every isiZulu row is text-only, so all 562 text-plus-image samples in the dataset remain English.

Table 2. Dataset overview

Metric	Full Dataset	English	isiZulu
Total samples	1,147	746	401
Factual samples	673	470	203
Misinformation samples	474	276	198
Text-only samples	585	184	401
Text + image samples	562	562	0
Number of categories	5	5	5
Unique sources (normalised)	92	-	-
Number of languages	2 (English, isiZulu)	-	-

Few multilingual, multimodal HIV/AIDS misinformation resources exist for English and isiZulu together; this dataset is among the first. Annotated multilingual data remains scarce across African LRL settings [11], a gap this resource is built to help close by giving researchers a benchmark for developing and assessing MID models in exactly that context.

D.1.1 Label distribution

Here, factual records exceed misleading records by 673 to 474, yielding a 58.7%/41.3% split across the entire dataset. English alone skews even more, at 63.0%/37.0%, while isiZulu brings the total result back into balance, with 203 true and 198 misinformation, virtually even at 50.6%/49.4%. The overall dataset ends up somewhat more balanced than English alone, then, and while some imbalance remains, it sits within a range standard class-weighting or resampling can handle during training. That balance is quite important. It reduces the likelihood of a classifier defaulting to the majority class and provides a solid platform for subsequent supervised experiments [12].

D.1.2 Category distribution

In Figure 4, we present the factual/misinformation distribution across the five thematic categories. Rather than concentrating on a single topic, the dataset covers multiple aspects of HIV/AIDS discourse, providing thematic diversity needed for benchmark misinformation datasets to support generalisation across different misinformation narratives [11], [13].

Two categories account for the largest shares of the dataset. Prevention content covering condoms, PrEP, and testing is the largest category, with 361 records (31.5%), followed by Myth with 266 records (23.2%). However, category size does not fully reflect the class distribution. Myth is strongly dominated by misinformation, with 210 of its 266 records, whereas Education and Prevention are predominantly factual. Stigma is the smallest category, with 106 records (9.2%), indicating that stigma-related content is comparatively under-represented in the dataset.

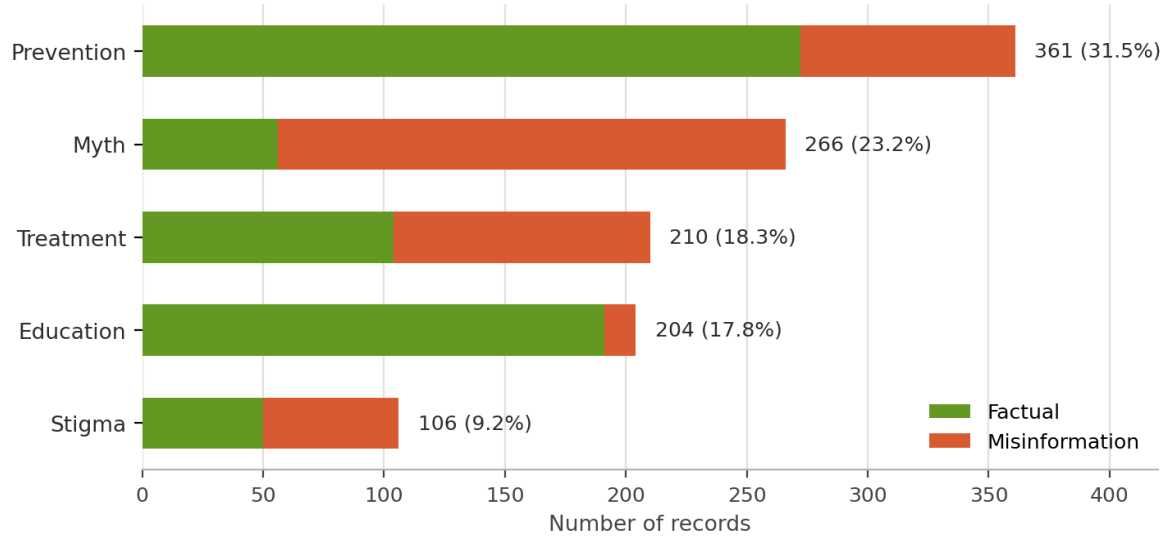


Figure 4. Factual and misinformation record counts within each thematic category

D.1.3 Modality distribution

The dataset contains a near-equal distribution of text-only (585 records, 51.0%) and text-plus-image (562 records, 49.0%) samples. This distribution differs from the English-only subset, where text-plus-image samples accounted for 75.3% of the 746 records. The near-equal split supports the evaluation of both unimodal and multimodal MID approaches, consistent with recent research showing that multimodal datasets can better capture the complexity of misinformation shared across online platforms [13], [15], [16].

D.1.4 Language and source distribution

IsiZulu makes up 35% of the dataset and is close to label-balanced, whereas English is skewed toward factual content (63.0%). Because the isiZulu rows are translations of a subset of the English entries rather than independently sourced, cross-lingual samples are paired by content, which is useful for cross-lingual transfer experiments, but the isiZulu subset alone should not be treated as an independently diverse sample of Zulu-language HIV/AIDS discourse. Including isiZulu addresses the scarcity of annotated datasets for African LRLs while simultaneously enabling cross-lingual learning experiments, a challenge consistently identified in recent multilingual MID literature [11], [17].

Table 3. Language distribution

Language	Count	% of Dataset	Factual	Misinfo.	Text-only	Text+Image
English	746	65.0%	470	276	184	562
isiZulu	401	35.0%	203	198	401	0

A total of 1,124 of 1,147 rows carry a source (23 missing, all isiZulu; see Section 4.1.5). After normalising URLs to their domain and collapsing platform name variants, the dataset draws on 92 distinct sources spanning social media, government and public health bodies, healthcare providers, educational institutions, fact-checkers, and news outlets, summarised in Table 4. Figure 5

shows the ten most frequent individual sources. Social media platforms (Facebook, Instagram, Pinterest, X) together account for roughly half of all sourced samples, which matches the project's focus on the channels through which HIV/AIDS misinformation circulates, while government and public health sources and healthcare providers supply most of the factual/educational counterweight. Since models trained on diverse sources typically generalise better when used in real-world disinformation situations, the diversity of sources is meant to support ecological validity [13, 18].

Table 4. Source categories.

Source Type	Examples
Government & Public Health	CDC, HIV.gov, UNAIDS, WHO
Healthcare Organisations	Continental Hospitals, ViiV Healthcare, HealthZA
Educational Institutions	Schulich, UC Davis
Fact-checking Organisations	Africa Check, Snopes
Social Media	Facebook, Instagram, X, Pinterest, LinkedIn
News & Media	BBC, Metro
Blogs & Health Websites	Prep2Me, DrPrem, Lal PathLabs, Labtoo

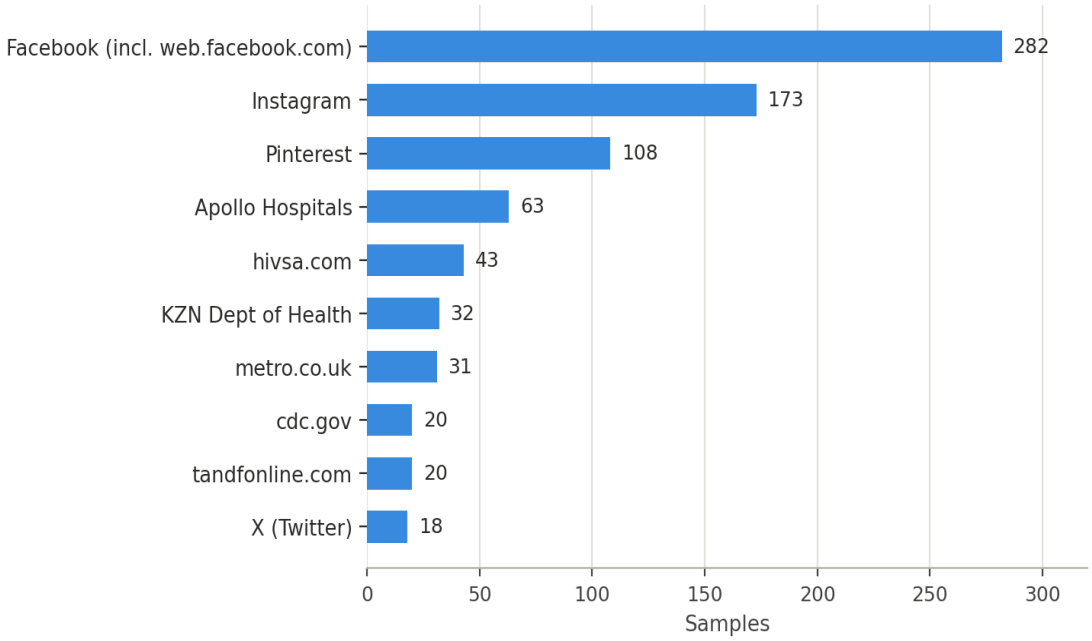


Figure 5. Ten most frequent individual sources by sample count

D.1.5 Text statistics and data quality

As shown in Table 5, nearly all records carry text: 1,144 of 1,147, with the remaining three being image-only posts. Length varies enormously, from three words at the shortest to 2,455 at the longest, and English shows this most sharply. Its mean of 58.78 words sits far above its median of 16, a skew traceable to a handful of long educational documents sitting among mostly short, social-media-length claims. IsiZulu texts are substantially shorter and more consistent,

averaging 8.42 words with a standard deviation of 2.64 and a maximum of 19 words. Across both languages, the combined mean is 41.13 words. The English 95th percentile length is 225.4 words, which remains within the processing limits of transformer models such as BERT, RoBERTa, and XLM-RoBERTa [14], [17]. Therefore, truncation or chunking would affect only a small proportion of unusually long texts in future transformer-based work.

Table 5. Text length statistics.

Statistic	Full Dataset	English	isiZulu
Samples with text	1,144 / 1,147	743 / 746	401 / 401
Mean length (words)	41.13	58.78	8.42
Median length (words)	12	16	8
Std. deviation	137.71	168.28	2.64
Min length	3	3	3
Max length	2,455	2,455	19
IQR (Q1-Q3)	8–21	11–39	6–10
95th percentile	130.7	225.4	13.0

The analysis script discovered three rows of missing content and 13 rows of identical text within the same language and label. The 13 repeated records must be manually reviewed to determine whether they are duplicates or valid repeated claims from several sources. Neither issue affects overall dataset usability, but both have been reported for transparency and will be resolved in the future version.

D.1.6 Misinformation theme analysis

The 276 English misinformation samples were examined using exploratory keyword analysis because the annotation schema did not include a topic label. IsiZulu was not included, as English keyword patterns do not transfer directly to isiZulu text. Figure 6 summarises the resulting classification. These categories are heuristic rather than verified ground truth and should be confirmed through manual review or a small labelled validation sample before use in model training.

The largest category, "Other/Stigma & Definitional Myths," accounted for 57.2% of the samples and included claims such as conflating HIV with AIDS or judging HIV status by appearance. Treatment myths (13.0%) and herbal or faith-based cure claims (9.8%) followed, consistent with the concentration of misinformation observed in the Treatment category (Section 4.1.2). Conspiracy-theory content was rare, with only one sample (0.4%), suggesting that the dataset contains more everyday misconceptions than organised disinformation. Exploratory keyword analysis provides an initial characterisation of the dataset, but manual annotation should be used to confirm these categories where possible [18].

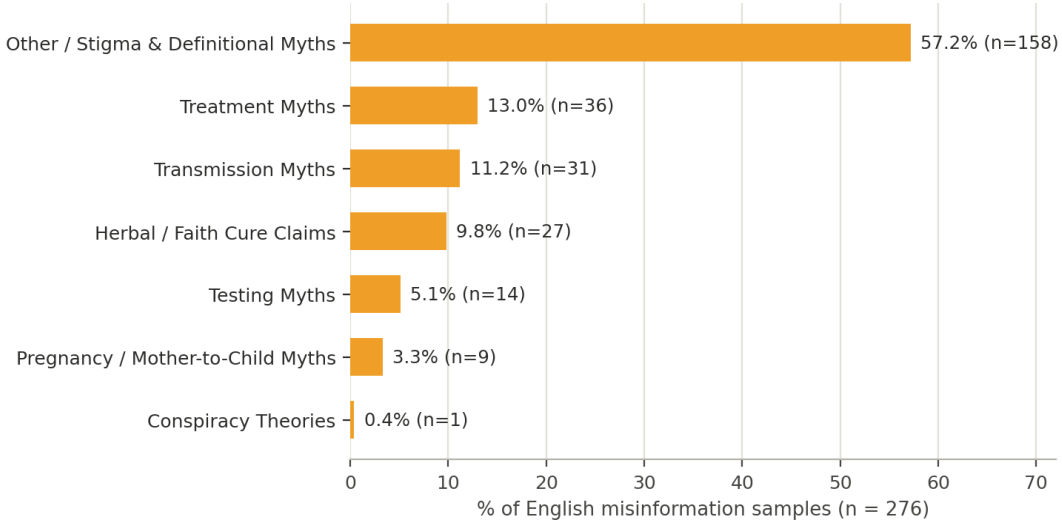


Figure 6. Exploratory theme distribution of English misinformation samples

The descriptive analysis indicates that the proposed dataset provides a diverse and well-structured resource for multilingual and multimodal HIV/AIDS MID, containing balanced factual and misinformation classes, multiple thematic categories, two languages, two modalities, and data drawn from 92 distinct online sources. These characteristics make the dataset suitable for evaluating both traditional machine learning and transformer-based MID models [13], [15].

D.2 Baseline Classification Results

To demonstrate the utility of the proposed dataset for MID, a series of baseline machine learning models were evaluated. These experiments establish benchmark performance for future studies using the dataset.

D.2.1 Classical machine learning baseline evaluation

Section 3.8 specifies the full baseline classification methodology, feature representations, model selection, class-imbalance handling, the train/test and cross-validation protocol, hyperparameter tuning, evaluation metrics, and the statistical testing protocol. This section reports the resulting cross-validation performance (Section 4.2.2), test-set performance (Section 4.2.3), the best-performing configuration's classification behaviour (Section 4.2.4), its ROC characteristics (Section 4.2.5), and its performance by language (Section 4.2.6).

D.2.2 Cross-validation performance

Table 6 presents the cross-validation performance of all tuned baseline configurations. The results demonstrate that the proposed dataset remains highly learnable after imbalance correction and hyperparameter tuning, with most models achieving weighted F1-scores at or above 0.85. Character TF-IDF combined with Logistic Regression and Character TF-IDF combined with Linear SVM produced the strongest cross-validation performance, both reaching a weighted F1-score of 0.888 and an MCC of 0.770. Random Forest and Gradient Boosting under character-level TF-IDF followed closely, while Naïve Bayes, even after

switching to the imbalance-appropriate Complement variant and tuning its smoothing parameter, remained the weakest configuration under character-level TF-IDF specifically.

Table 6. Cross-validation performance of tuned, imbalance-aware classical machine learning models

Feature Rep.	Model	Best Hyperparameters	CV F1	CV MCC
Character TF-IDF	Logistic Regression	C=10.0	0.888	0.770
Character TF-IDF	Linear SVM	C=1.0	0.888	0.770
Word TF-IDF	Linear SVM	C=1.0	0.884	0.761
Character TF-IDF	Random Forest	n_estimators=500, max_depth=None	0.883	0.761
Character TF-IDF	Gradient Boosting	n_estimators=300, lr=0.1, max_depth=3	0.882	0.760
Word TF-IDF	Logistic Regression	C=10.0	0.880	0.752
Character TF-IDF	XGBoost	n_estimators=200, max_depth=6, lr=0.1	0.873	0.741
Word TF-IDF	Naïve Bayes	alpha=0.1	0.866	0.724
Word TF-IDF	Random Forest	n_estimators=200, max_depth=None	0.850	0.701
Word TF-IDF	Gradient Boosting	n_estimators=300, lr=0.1, max_depth=3	0.835	0.675
Word TF-IDF	XGBoost	n_estimators=500, max_depth=4, lr=0.05	0.812	0.633
Character TF-IDF	Naïve Bayes	alpha=0.1	0.767	0.525

Best hyperparameters per grid search; per-model precision/recall/accuracy breakdowns are available in the accompanying results files.

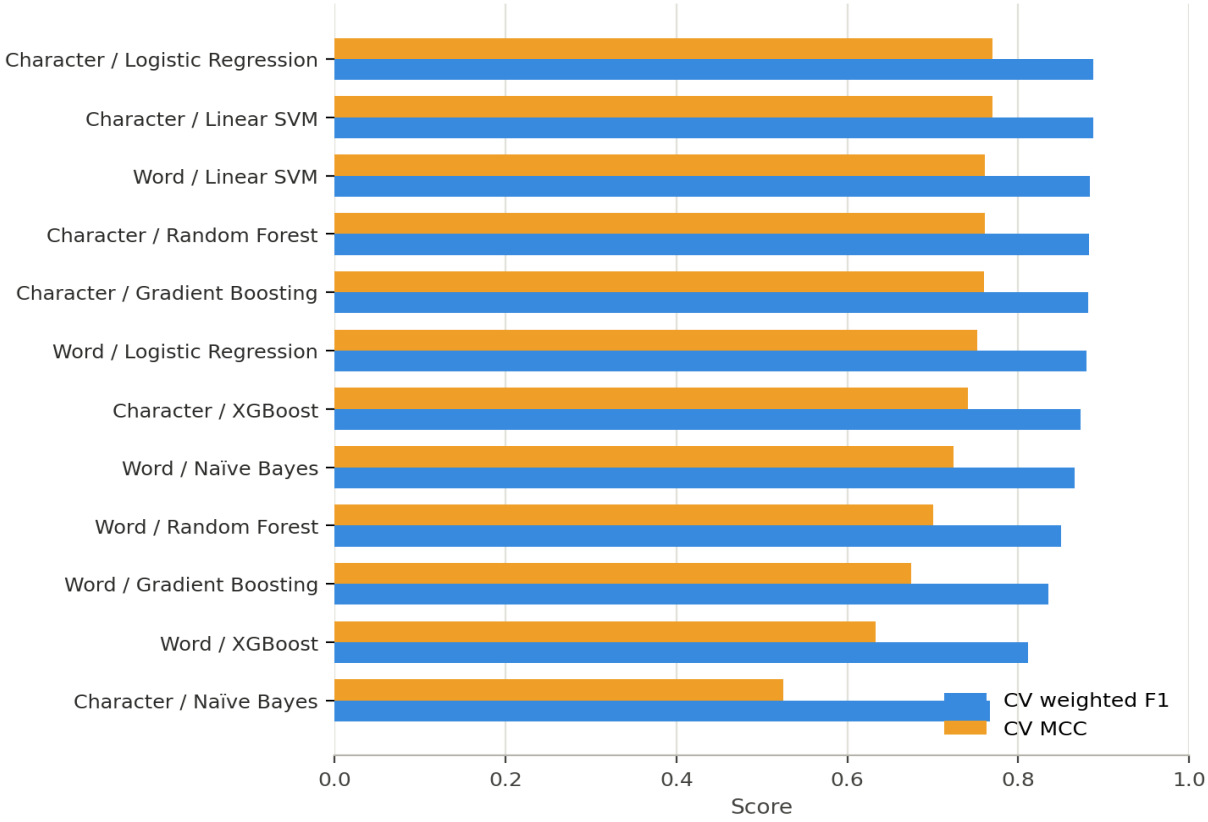


Figure 7. Cross-validation weighted F1 and MCC for all twelve model-vectorizer configurations, sorted by F1 descending.

Figure 7 shows a slight change in the leading configuration when MCC is used instead of weighted F1. Weighted F1 ranks Character TF-IDF with Logistic Regression first, whereas MCC ranks Character TF-IDF with a Linear Support Vector Machine first. The two configurations differ by less than 0.001 on either metric, making the ranking difference marginal. The paired t-test ($p = 0.994$) and the Wilcoxon signed-rank test ($p = 1.000$) indicated no statistically significant difference between the five matched folds. McNemar's exact test yielded identical results on the test-set predictions ($p = 1.000$), with each setup correctly identifying two records that the other did not.

Across the five folds shown in Figure 8, no configuration maintained a consistent lead. Their scores fluctuated from fold to fold, indicating variability rather than a stable performance advantage. On this dataset, Logistic Regression and Linear SVM with character-level TF-IDF should be considered statistically indistinguishable, rather than treating the nominal top row of Table 6 as the decisive winner. Their disagreements on individual test records are consistent with sampling variation rather than a systematic difference in modelling capacity.

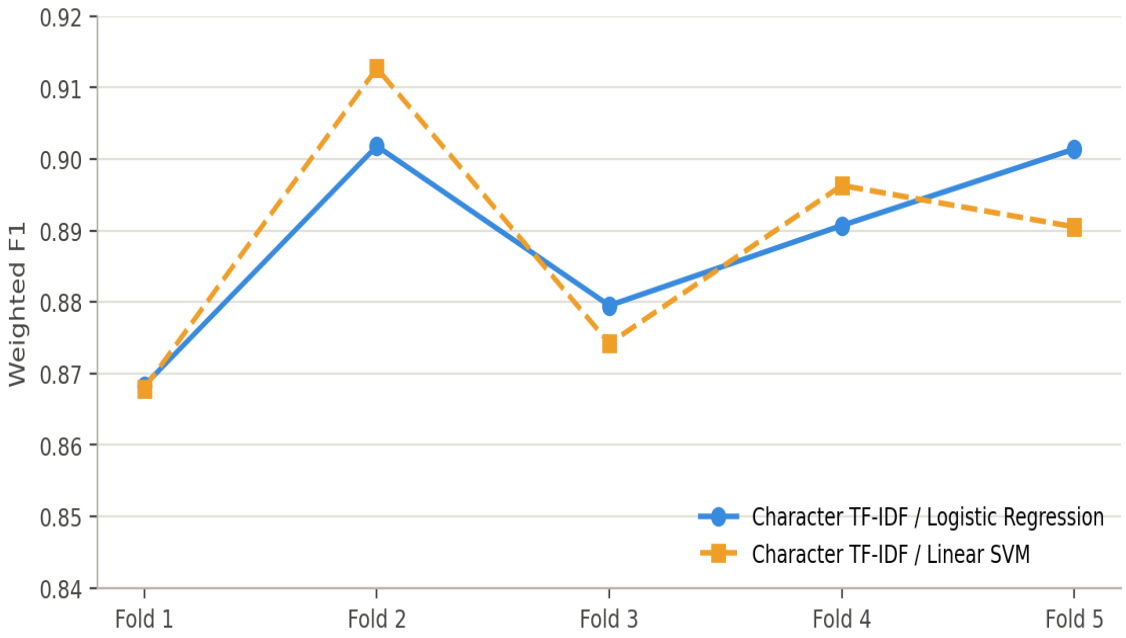


Figure 8. Per-fold weighted F1 for the top 2 cross-validation configurations

Character-level TF-IDF consistently outperformed word-level TF-IDF across both linear and ensemble classifiers. Character n-grams capture prefixes, suffixes, and partial word patterns that word-level features may miss, making them useful for isiZulu, an agglutinative language with extensive morphological variation. They can also reduce the effects of spelling variation and vocabulary sparsity by representing sub-word patterns rather than relying only on fixed lexical tokens. This result is consistent with multilingual MID research showing better generalisation of character-based representations across LRLs [11], [14].

All tuned configurations substantially outperformed the majority-class dummy baseline (58.5% accuracy, weighted F1 0.432, MCC 0.000, and PR-AUC 0.415). This difference indicates that the classifiers captured lexical and semantic information beyond the label distribution described in Section 4.1.1.

The effect of imbalance handling and tuning was assessed against an earlier comparable baseline. Figure 9 shows the change in cross-validation weighted F1 across the twelve configurations. Ten configurations changed by no more than 0.024 in either direction, whereas Naïve Bayes showed larger gains of 0.073 with character-level TF-IDF and 0.055 with word-level TF-IDF. Standard multinomial Naïve Bayes was the only classifier in the original pass without an imbalance-handling mechanism; the other models either already used one or received partial handling through balanced class weights, sample weighting, or a positive-class scale factor. The larger improvement for Naïve Bayes therefore corresponds to the classifier for which imbalance handling had previously been absent.

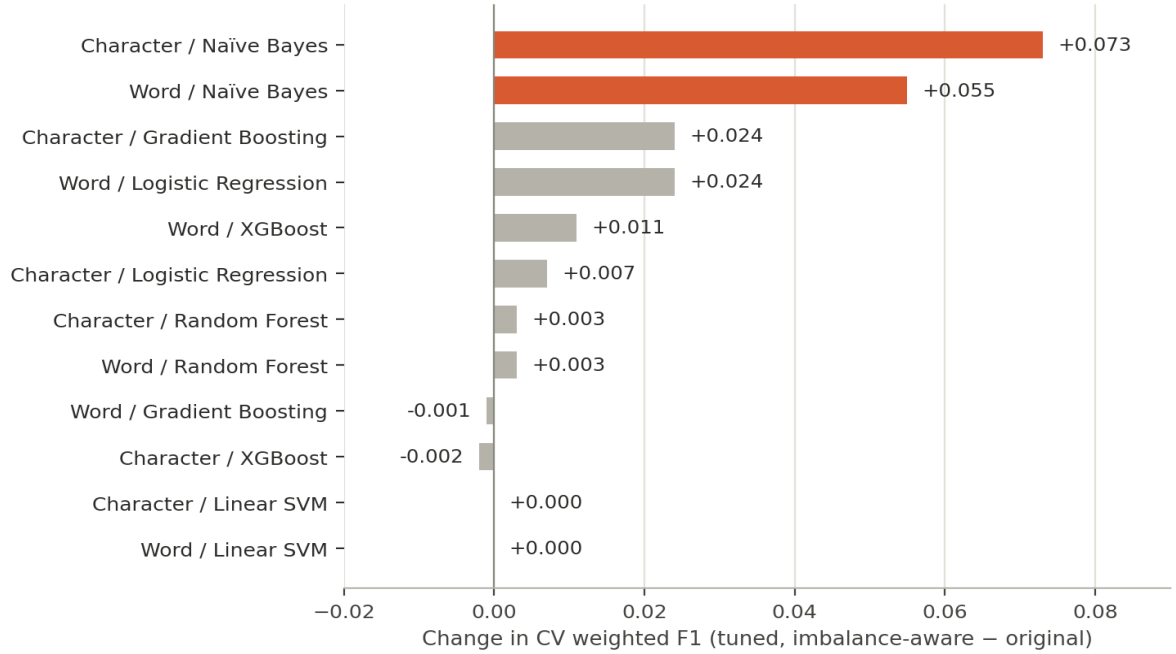


Figure 9. Change in cross-validation weighted F1 after imbalance handling and hyperparameter tuning, by configuration.

D.2.3 Test Set performance

Table 7. Held-out test-set performance of tuned classical machine learning models

Feature Rep.	Model	Accuracy	Weighted F1	MCC
Word TF-IDF	Random Forest	0.935	0.935	0.874
Character TF-IDF	Logistic Regression	0.926	0.926	0.847
Character TF-IDF	Linear SVM	0.926	0.926	0.847
Word TF-IDF	Logistic Regression	0.917	0.917	0.830
Character TF-IDF	Random Forest	0.917	0.917	0.829
Word TF-IDF	Linear SVM	0.913	0.913	0.821
Character TF-IDF	XGBoost	0.900	0.900	0.797
Word TF-IDF	Naïve Bayes	0.900	0.900	0.794
Word TF-IDF	Gradient Boosting	0.895	0.896	0.803
Character TF-IDF	Gradient Boosting	0.887	0.887	0.775
Word TF-IDF	XGBoost	0.825	0.826	0.675
Character TF-IDF	Naïve Bayes	0.799	0.800	0.595
N/A	Dummy (Most Frequent)	0.585	0.432	0.000

including the majority-class floor baseline (per-model precision/recall/ROC-AUC/PR-AUC breakdowns are available in the accompanying results files)

The final evaluation was conducted on the held-out test set to assess each model's ability to generalise to previously unseen data. Unlike the cross-validation ranking, the Random Forest classifier combined with word-level TF-IDF (200 trees, unrestricted depth, selected by grid search) emerged as the best-performing configuration on the test set, achieving 93.45% accuracy, a weighted F1-score of 93.49%, a macro F1-score of 93.36%, an MCC of 0.874, a Precision-Recall AUC of 0.973, and a ROC-AUC of 0.983. Character TF-IDF combined with Logistic

Regression and Linear SVM tied for second on the test set at 92.6% accuracy, consistent with the cross-validation finding in Section 4.2.2 that these two configurations are statistically indistinguishable from one another. Every tuned configuration substantially outperformed the majority-class floor of 58.5% accuracy and 0.432 weighted F1.

The fact that Random Forest ranked mid-table on cross-validation (Table 6, ninth of twelve by F1) but first on the independent test set is worth noting explicitly rather than smoothing over. This is consistent with the significance-testing result in Section 4.2.2: several configurations sit close enough together that which one comes out ahead can depend on the specific split, and a single held-out test set is one draw from that variability, not a guaranteed ranking. Random Forest's test-set lead should therefore be read as the outcome on this particular split rather than as decisive evidence that it is intrinsically superior to the character-level linear models it narrowly beat.

Word-level TF-IDF produced the highest-performing model on the independent test set despite character-level TF-IDF's advantage during cross-validation. This aligns with the split-level variability discussed earlier rather than contradicting it. The lexical patterns that separate factual information from misinformation appear stable enough for both representations to perform well. The specific ranking, however, is sensitive to the choice of split and the hyperparameters applied.

D.2.4 Classification Performance of the best baseline model

The best-performing baseline model, which is Random Forest with word-level TF-IDF, tuned to 200 trees with unrestricted depth, correctly identified 120 of 134 factual samples and 94 of 95 misinformation samples in the held-out test set. This resulted in 15 misclassified instances, with 14 factual samples labelled as misinformation and one misinformation sample labelled as factual. These findings differ from an earlier untuned run with the identical algorithm and test split, which missed no disinformation data. The distinction is that there is only one test record. The tuned results are shown here since using the more favourable untuned result would misrepresent the sensitivity of the discovery to hyperparameter selection.

From a public health standpoint, false negatives are the more consequential error because misinformation classified as factual can pass through undetected. False positives, by contrast, require only additional verification of factual content. The tuned model demonstrated good performance against this benchmark. It achieved a recall of 98.9% by properly identifying 94 out of 95 instances of false information. In the direction of the more important error, this suggests a decision boundary that is strongly oriented toward care. However, the former claim of zero missed cases for this setup should be retired because only one instance passed across; it should not reappear without re-verification. Precision yielded class-specific F1-scores of 0.94 and 0.93, with truthful content scoring 0.99 and disinformation scoring 0.87. Across both classes, a weighted F1 of 0.935 and an MCC of 0.874 point to strong performance overall, short of flawless.

D.2.5 Receiver Operating Characteristic Analysis

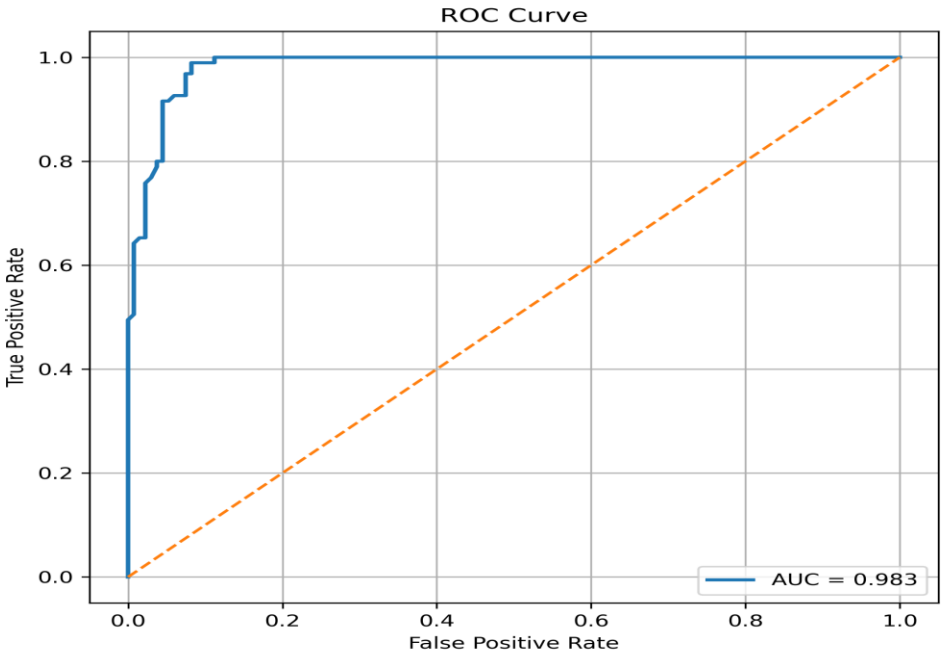


Figure 10. ROC curve of the best baseline model

Figure 10 presents the ROC curve for the best-performing baseline model involving Random Forest and word-level TF-IDF. The ROC-AUC of 0.983 indicates that the model discriminates effectively between factual information and misinformation across a wide range of classification thresholds, rather than at a single threshold. The difference is significant since accuracy evaluates performance at a specific decision point, whereas ROC-AUC captures the model's overall capability to discriminate the two classes regardless of where that threshold falls. An AUC this high points to something specific about the dataset itself: the lexical signal it carries is discriminative enough to support MID reliably, which in turn backs the case for using this dataset as a benchmark for both classical and deep learning approaches going forward.

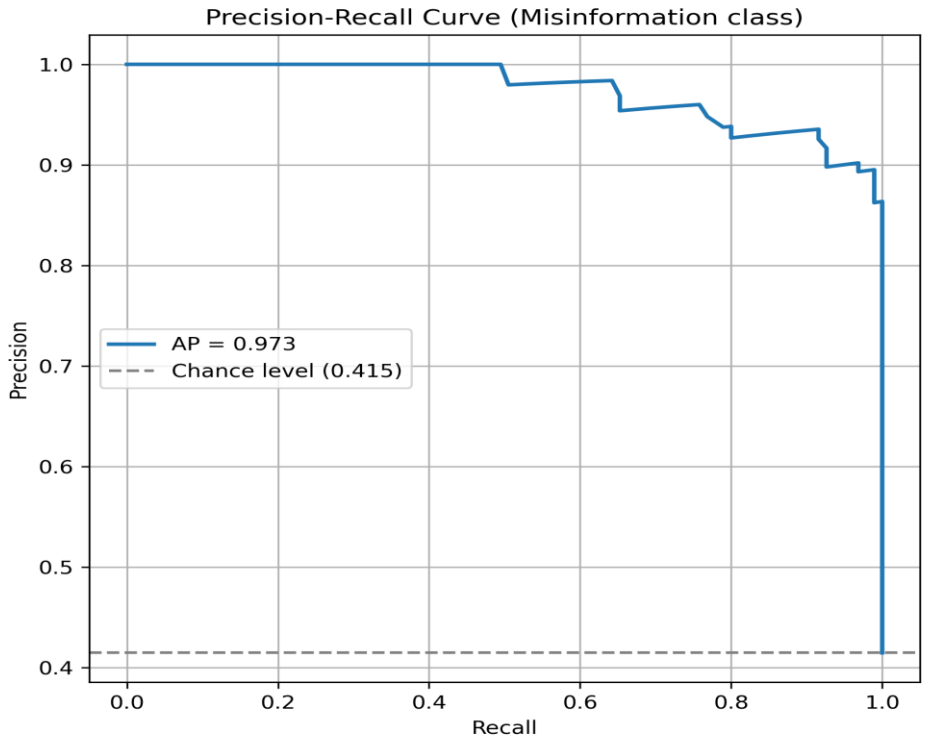


Figure 11. Precision-Recall curve of the best baseline model

Figure 11 presents the Precision-Recall curve for the same model, with Misinformation as the positive class. It complements the ROC curve in Figure 10 because ROC-AUC treats both classes symmetrically, which can obscure poor minority-class ranking under imbalance. PR-AUC addresses this limitation: precision remains above 0.95 until recall reaches about 0.65, then declines as recall approaches 1.0. The average precision is 0.973, well above the 0.415 baseline established by the Misinformation class's predominance in the test set and reported for the dummy baseline in Section 4.2.2. This pattern is also consistent with the single missed misinformation case discussed in Section 4.2.4. Thus, PR-AUC and ROC-AUC agree for this model, although such agreement cannot be assumed under class imbalance.

D.2.6 Language-specific performance

Multilingual support for English and isiZulu is a main dataset objective, making language-specific performance important. The best-performing baseline achieved 93.24% accuracy and a weighted F1-score of 93.32% on English, compared with 93.83% and 93.81%, respectively, on isiZulu. The difference is less than one percentage point, indicating broadly consistent performance across the two subsets. The isiZulu results match those obtained after hyperparameter tuning, whereas the English results are slightly lower because of the single additional misclassification discussed in Section 4.2.4.

Overall, the near-identical performance suggests that the translated isiZulu subset retained the semantic features needed for effective misinformation classification. Although the isiZulu corpus is substantially smaller than the English

subset, the comparable performance indicates that the multilingual extension did not negatively affect classification accuracy. Instead, the findings offer preliminary evidence that the suggested dataset can detect cross-lingual misrepresentation. The dataset may also serve as a valuable baseline for future multilingual transformer models aimed at African LRLs, where recent literature has indicated a scarcity of annotated resources [11], [17].

E. Discussion

This paper presents a multilingual, multimodal HIV/AIDS misinformation dataset for English and isiZulu, addressing a significant gap in benchmark resources for African LRLs. The dataset was developed using a structured six-stage methodology informed by the Datasheets for Datasets and Data Statements for NLP frameworks. evaluate its effectiveness, baseline experiments were conducted on the proposed multilingual HIV/AIDS misinformation dataset. The results reveal that the dataset contains enough informative lexical and semantic patterns to permit accurate MID using traditional machine learning approaches when class imbalance is handled and hyperparameters are modified rather than fixed. All tested classifiers exceeded the majority-class baseline described in Section 4.2.2, showing that the dataset is well-structured, internally coherent, and suitable for supervised learning. At the same time, the significance testing in Section 4.2.2 shows that several of the leading configurations are not reliably distinguishable from one another. As a result, the identity of the "best" model should be interpreted as one reasonable outcome of the evaluation rather than a definitive conclusion.

The strong performance observed across both English and isiZulu provides additional evidence that the multilingual extension preserves semantic consistency while expanding linguistic coverage. These results should be interpreted in the context of the gap identified in Section 2. Recent systematic reviews of LRL and health-domain MID reported that HIV, malaria, and reproductive health were absent from existing MID benchmarks [19]. They also reported that dedicated isiZulu misinformation resources did not yet exist. The dataset presented here addresses both gaps simultaneously. It also responds to the broader finding that text-only detection dominates the field with 86% of reviewed studies in [19], by including a near-equal split of text-only and text-plus-image records suitable for multimodal experimentation.

Although classical machine learning models achieved excellent predictive performance, these approaches rely primarily on sparse lexical representations and therefore cannot fully exploit contextual semantic relationships or multimodal information. Recent advances in transformer-based language models have shown substantial improvements in contextual language understanding through self-attention mechanisms and deep contextual embeddings. Consequently, while the classical baselines established here provide strong benchmark performance for the proposed dataset, further improvements may be expected from transformer-based architectures such as BERT, RoBERTa, and XLM-RoBERTa. These models are particularly effective when modelling contextual meaning, multilingual representations, and complex semantic relationships within HIV/AIDS misinformation.

E.1 Limitations

Several limitations in this study should be considered when interpreting these results and when reusing the dataset. First, the isiZulu subset consists of translations of a subset of the English entries rather than independently sourced isiZulu-language content; it should therefore not be treated as an independently diverse sample of Zulu-language HIV/AIDS discourse, and results on the isiZulu subset should be interpreted with this dependency in mind. Second, the train/test split does not group records by underlying content, so translation pairs may straddle the split boundary; the cross-lingual performance parity reported in Section 4.2.6 should be treated as provisional until re-evaluated on a content-grouped split.

Third, the misinformation theme analysis presented in Section 4.1.6 is a keyword-based heuristic rather than a manually verified ground truth, and would benefit from manual validation before being used for model training or evaluation. Fourth, inter-annotator agreement has not yet been formally computed; this is planned as part of future dataset releases, following the Cohen's Kappa protocol described in Section 3.6. Fifth, a small number of data quality issues, three missing-text rows and 13 rows sharing identical text within the same language and label, remain open and should be resolved or explicitly accepted before large-scale model training. Finally, the current baseline evaluation is restricted to classical machine learning classifiers using TF-IDF features; a transformer fine-tuning pipeline has been implemented but not yet evaluated, as described in Section 5.2, and multimodal models remain to be evaluated on this dataset.

E.2 Future Work

Building on the baseline results presented in this paper, future work will (i) add a translation-pair identifier to the dataset schema and re-run both the classical and transformer baselines on a content-grouped split, to close the leakage risk identified in Section 4.2.1 before the cross-lingual result is relied upon further; (ii) complete evaluation of the XLM-RoBERTa fine-tuning pipeline already implemented for this dataset will require a separate validation partition carved out of the training data. Model selection should not be performed against the test set, so that the held-out test set remains untouched until final reporting; (iii) develop and evaluate multimodal fusion models that combine the textual and image modalities present in approximately half of the dataset; (iv) compute formal inter-annotator agreement statistics on an expanded, independently annotated subset; (v) expand the isiZulu subset with natively sourced, rather than translated, content to improve linguistic diversity; and (vi) subject the dataset and resulting models to adversarial and out-of-distribution evaluation, an area identified as absent across the 51 studies reviewed in [19]. These directions are intended to move the dataset from an initial benchmark toward a resource capable of supporting deployment-relevant evaluation in African low-resource-language public health settings.

F. A Reusable Framework for Constructing Health Misinformation Datasets

Section 3 described six stages used to build this specific dataset. Across the wider literature, dataset construction is reported as the dominant research activity in low-resource-language and health-domain MID, yet each study describes its

own construction process narratively, with no shared structure that a subsequent study could follow directly [19]. Multilingual, multimodal resources such as MMIFND/MMCFND [29] demonstrate that scale is achievable across many languages, but their construction process, like most others in the field, is reported as a one-off account rather than a reusable procedure. This section generalises the six stages used in Section 3 into the Health Misinformation Dataset Construction Framework (HMD-CF): a disease- and language-agnostic reference process, with an explicit design decision attached to each stage, that a researcher could apply to a different disease, language, or region.

F.1 Framework Stages

HMD-CF keeps the six stages in the same order used in Section 3, but separates what is generalisable (the stage and its purpose) from what must be re-decided for a new context (the design decision). Figure 12 shows the six stages alongside their associated decision, and Table 8 gives the same structure with the decision each study must resolve set against how this study resolved it for HIV/AIDS, English, and isiZulu.

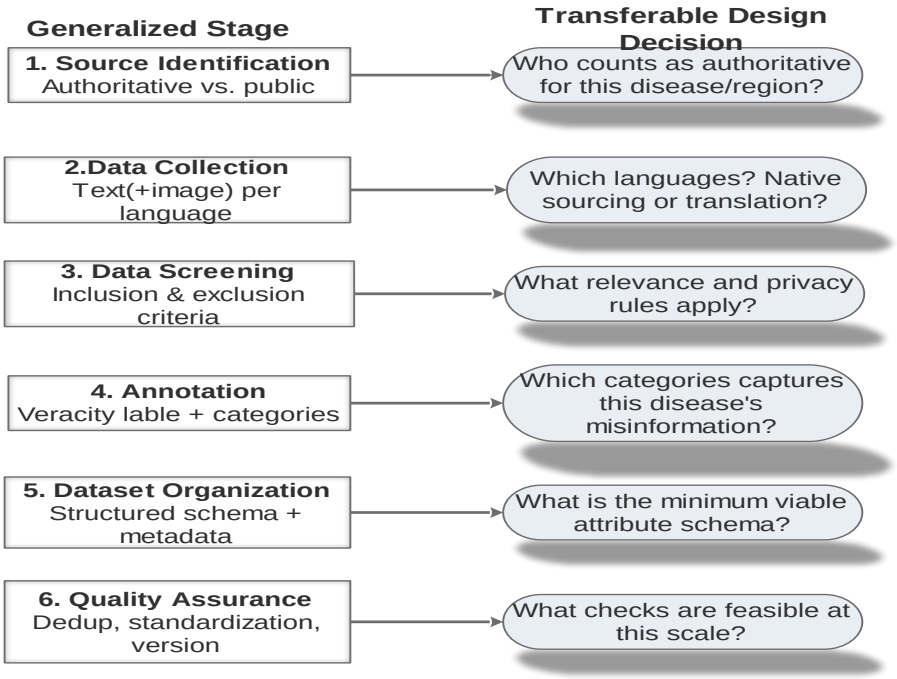


Figure 12. HMD-CF generalised construction stages paired with the design decision

Table 8. HMD-CF stage definitions, transferable design decisions, and this study's instantiation

Stage	Purpose	Transferable decision	This study's instantiation
1. Source identification	Locate authoritative and public sources for the target health topic	Which institutions are recognised as authoritative for this disease and region;	WHO, UNAIDS, SA Dept. of Health, HIVSA, Apollo Hospitals; Facebook, Instagram, blogs, news

		which public platforms carry relevant discourse	
2. Data collection	Acquire text and paired imagery where available, in each target language	Which languages to cover; native sourcing versus translation from a high-resource pivot language	Manual collection and Python-based scraping in English and isiZulu; images linked where present
3. Data screening	Apply inclusion/exclusion criteria before annotation	Domain-relevance rules; language rules; privacy and PII exclusion rules	HIV/AIDS relevance, English/isiZulu, factual-or-misinformation, public source, PII removed
4. Annotation	Assign a veracity label and domain-specific thematic categories	Which categories capture this disease's misinformation patterns, established through a disease-specific literature review	Binary Factual/Misinformation label plus Prevention, Education, Treatment, Myth, Stigma categories
5. Dataset organisation	Encode records into a structured, reproducible schema	Minimum viable attribute schema for the target modality and reuse case	Eight-attribute schema: ID, text, image path, label, category, modality, language, source (Table 1)
6. Quality assurance	Verify consistency, remove duplicates, and version the release	Which checks are feasible at the target scale; when inter-annotator agreement is computed	Deduplication, label standardisation, missing-value/language/metadata checks, Version 1.0 freeze; IAA planned

F.2 Transferability: A Worked Example

To illustrate how HMD-CF generalises beyond HIV/AIDS, English, and isiZulu, Table 9 walks through how each stage's decision would be resolved differently for a hypothetical malaria-misinformation dataset in Sesotho, a case not implemented in this paper but consistent with existing Sesotho-language dataset work [26]. The stage sequence and its purpose stay fixed; only the instantiation changes.

Table 9. Illustrative transfer of HMD-CF to a different disease and language, shown for comparison against this study's instantiation.

Stage	**This study: HIV/AIDS, English + isiZulu**	**Illustrative alternative: malaria, Sesotho**
1. Source identification	WHO, UNAIDS, SA Dept. of Health, HIVSA, social media	WHO, national malaria control programme, regional clinics, social media
2. Data collection	Manual + scraped, English and isiZulu, text + image	Manual + scraped, Sesotho, likely text-only given platform availability
3. Data screening	HIV/AIDS relevance, two named languages, public source	Malaria relevance, Sesotho only, public source
4. Annotation	Prevention, Education, Treatment, Myth, Stigma	Prevention, Diagnosis, Treatment, Myth, Seasonal/Climate claims
5. Dataset organisation	Eight-attribute schema with image path	Same schema, image-path field likely sparsely populated
6. Quality assurance	Dedup, standardise, Version 1.0, IAA planned	Same checks; IAA prioritised earlier given a single-language corpus

The categories in Stage 4 change because they are disease-specific by design. HIV/AIDS misinformation clusters around transmission and stigma, while malaria misinformation in many settings clusters around seasonal and environmental claims. What transfers is the requirement to derive categories from a disease-specific literature review before annotation begins, not the categories themselves. Stage 6 differs in emphasis rather than in content. A single-language corpus removes the cross-lingual verification step used in Section 3.6, making inter-annotator agreement comparatively cheaper to prioritise earlier in the process.

F.3 Relationship to Existing Documentation Frameworks

HMD-CF does not replace the Datasheets for Datasets [2] or Data Statements for NLP [4] frameworks used throughout Section 3. However, it operationalises them into an ordered, stage-by-stage process specific to health-domain, low-resource-language misinformation datasets. Datasheets and Data Statements specify what should be documented about a finished dataset. HMD-CF specifies the order in which decisions are made while the dataset is still being built, with each stage's decision feeding the next: the categories chosen in Stage 4 depend on the languages and sources fixed in Stages 1 and 2, and the quality-assurance checks feasible in Stage 6 depend on the schema fixed in Stage 5. Positioned in this way, HMD-CF addresses the gap identified in [19]. Dataset construction dominates research activity in the field but is rarely reported as a process that another study, whether in a different disease or language, could follow directly rather than reinvent.

G. Conclusion

This paper presented the construction, characterisation, and baseline evaluation of a multilingual and multimodal HIV/AIDS misinformation dataset for English and isiZulu. It also presented a reusable framework for constructing comparable datasets for other diseases and languages. The dataset was built through a six-stage methodology, including source identification, data collection, screening, annotation, organisation, and quality assurance, grounded in the Datasheets for Datasets and Data Statements for NLP frameworks. The resulting Version 1.0 release contains 1,147 records spanning two languages, two modalities, five thematic categories, and 92 normalised sources. Classical machine learning baselines, evaluated across six imbalance-aware, hyperparameter-tuned classifiers and two TF-IDF feature representations, and checked against a majority-class floor baseline and paired significance tests, achieved strong and consistent performance, with the best model, Random Forest using word-level TF-IDF, reaching 93.45% test accuracy, a 93.49% weighted F1-score, an MCC of 0.874, and a ROC-AUC of 0.983, with near-identical performance across English and isiZulu subsets. Generalising this methodology, the proposed HMD-CF framework separates six fixed stages from the design decisions attached to each stage, which must be re-resolved for a new disease or language. We also illustrated this transfer with a hypothetical malaria-misinformation dataset in Sesotho.

The dataset provides a learnable, internally coherent, and linguistically balanced benchmark for HIV/AIDS MID. More broadly, it addresses a documented gap in the MID literature: the complete absence of HIV-domain resources within

existing low-resource-language and health-domain benchmarks. By combining multilingual coverage, multimodal structure, and transparent documentation, the dataset is designed to support transformer-based and multimodal modelling. It also enables cross-lingual transfer experiments and adversarial robustness evaluation in a domain where undetected misinformation carries severe consequences, disproportionately affecting low-resource-language communities.

Data and Code Availability

Given the health-sensitive nature of its content, HIVMID is not deposited in an open public repository. It will be made available on a restricted, request-access basis upon completion of the broader research programme of which this study forms part. Researchers may contact the corresponding author to request access, subject to a data use agreement restricting use to non-commercial research purposes and, where applicable, evidence of institutional ethics approval for the requesting researcher's own study. The dataset is distributed as HIVMID.csv, following the eight-attribute schema described in Table 1. The data-cleaning, descriptive-statistics, and classical-baseline code used to produce Sections 4.1 and 4.2, including exact classifier and vectoriser configurations, will be made available alongside the dataset under the same access terms. A transformer fine-tuning pipeline (XLM-RoBERTa-base) has been implemented for this dataset but had not been evaluated at the time of writing; results will be reported in a subsequent revision.

Acknowledgment

This research was supported by the Unit for Data Science and Computing (UDSC) and the Department of Computer Science at the North-West University, Mafikeng campus, South Africa.

References

[1] Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. 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.

[2] T. Gebru et al., “Datasheets for datasets,” Communications of the ACM, vol. 64, no. 12, pp. 86–92, Dec. 2021, doi: 10.1145/3458723.

[3] Alkemade, H., Been, M., Boot, P., et al., “Datasheets for Digital Cultural Heritage Datasets,” Journal of Open Humanities Data, vol. 9, p. 23, Sep. 2023, doi: 10.5334/johd.124.

[4] E. M. Bender and B. Friedman, “Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science,” Transactions of the Association for Computational Linguistics, vol. 6, pp. 587–604, Dec. 2018, doi: 10.1162/tacl_a_00041.

[5] P. Meel and D. Vishwakarma, “Fake news, rumour, information pollution in social media and web: A contemporary survey of state-of-the-art, challenges and opportunities,” Expert Systems with Applications, vol. 153, p. 112986, Oct. 2020, doi: 10.1016/j.eswa.2019.112986.

[6] E. Aïmeur, S. Amri, and G. Brassard, “Fake news, disinformation and misinformation in social media: a review,” Social Network Analysis and Mining, vol. 13, no. 1, p. 30, Dec. 2023, doi: 10.1007/s13278-023-01028-5.

[7] Mukwevho, M., Rananga, S., Mbooi, M. S., Isong, B., & Marivate, V. Building a Dataset for Misinformation Detection in the Low-Resource Language. In 2024 IST-Africa Conference (IST-Africa), Dublin, Ireland, 2024, pp. 01-09.

[8] T. G. Gereme, W. M. Demissie, and G. A. Ayele, “Combating fake news in “low-resource” languages: Amharic fake news detection accompanied by resource crafting,” Information, vol. 12, no. 1, p. 20, Jan. 2021, doi: 10.3390/info12010020.

[9] R. Artstein and M. Poesio, “Inter-Coder Agreement for Computational Linguistics,” Computational Linguistics, vol. 34, no. 4, pp. 555–596, Dec. 2008, doi: 10.1162/coli.07-034-R2.

[10] J. Cohen, “A coefficient of agreement for nominal scales,” Educational and Psychological Measurement, vol. 20, no. 1, pp. 37–46, Jan. 1960, doi: 10.1177/001316446002000104.

[11] Wang, X., Zhang, W., & Rajtmajer, S. Monolingual and multilingual misinformation detection for low-resource languages: A comprehensive survey. arXiv preprint arXiv:2410.18390, 2024.

[12] R. Chalehchaleh and R. Farahbakhsh, “Addressing data scarcity in multilingual fake news detection: an LLM-based dataset augmentation approach,” Social Network Analysis and Mining, vol. 15, no. 1, p. 103, Aug. 2025, doi: 10.1007/s13278-025-01505-z.

[13] S. Harris, V. T. Ta, M. Trovati, G. Nakhla, F. Latif, and Y. Korkontzelos, “Multimodal misinformation detection across diverse languages using RAG and LLMs,” Journal of Intelligent Information Systems, vol. 64, no. 4, pp. 1719–1746, Aug. 2026, doi: 10.1007/s10844-026-01042-x.

[14] K. LekshmiAmmal and S. Madasamy, “A reasoning based explainable multimodal fake news detection for low resource language using large language models and transformers,” Journal of Big Data, vol. 12, no. 1, p. 46, Dec. 2025, doi: 10.1186/s40537-025-01093-x.

[15] R. Jadhav, V. Meshram, A. Bhosle, K. Patil, S. Dash, and S. Jadhav, “Explainable multilingual and multimodal fake-news detection: toward robust and trustworthy AI for combating misinformation,” Frontiers in Artificial Intelligence, vol. 8, article 1690616, Dec. 2025, doi: 10.3389/frai.2025.1690616.

[16] V. Gupta, R. Kumari, N. Ashok, T. Ghosal, and A. Ekbal, “MMM: An Emotion and Novelty-aware Approach for Multilingual Multimodal Misinformation Detection,” in Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, online, Nov. 2022, pp. 315–326.

[17] H. Noman and S. Ahmed examine misinformation detection in low-resource languages using multilingual transformers and explainable classification in Computers and Education Letters, vol. 3, no. 1, pp. 31-42, Jun. 2026, doi: 10.64038/cel.01202652

[18] Z. M. Hüsünbeyi, V. Mouilleron, L. Uhling, D. Foppe, T. Scheffler, and D. Seddah, "Multilingual, Multimodal Pipeline for Creating Authentic and Structured Fact-Checked Claim Dataset," arXiv preprint arXiv:2601.07985, 2026.

[19] B. Isong and R. Linah. Misinformation Detection in Low-Resource Languages and Health Domains: Review and Evaluation Framework. Indonesian Journal of Computer Science, in press, 2026.

[20] Sethosa, M. R., Rananga, S., & Mbooi, M. S. Hate Speech Detection in isiZulu: A Semantic Approach. In 2025 IST-Africa Conference (IST-Africa), Nairobi, Kenya, 2025, pp. 1-9.

[21] Rananga, S., Isong, B., Modupe, A., & Marivate, V. Misinformation Detection: A Review for High and Low-Resource Languages. Journal of Information Systems and Informatics, 6(4), 2892-2922, 2024.

[22] K. Hayawi, S. Shahriar, M. A. Serhani, I. Taleb, and S. S. Mathew, "ANTi-Vax: A novel Twitter dataset for COVID-19 vaccine misinformation detection," Public Health, vol. 203, pp. 23–30, Feb. 2022, doi: 10.1016/j.puhe.2021.11.021.

[23] J. Alghamdi, Y. Lin, and S. Luo, "Towards COVID-19 fake news detection using transformer-based models," Knowledge-Based Systems, vol. 274, p. 110642, Aug. 2023, doi: 10.1016/j.knosys.2023.110642.

[24] Schlicht, I.B., Fernandez, E., Chulvi, B. et al. Automatic detection of health misinformation: a systematic review. J Ambient Intell Human Comput 15, 2009-2021 (2024). https://doi.org/10.1007/s12652-023-04619-4

[25] Manaswe, M., Rananga, S., & Sindane, T. Detecting COVID-19 Misinformation in Sepedi Short Texts Using Data Augmentation Techniques. In 2025 IST-Africa Conference (IST-Africa), Nairobi, Kenya, 2025, pp. 1-7.

[26] Mokoena, L., Rananga, S., & Mbooi, M. Misinformation Detection in COVID-19 News for Low-Resource Languages: A Sesotho Case Study. In 2025 IST-Africa Conference (IST-Africa), Nairobi, Kenya, 2025, pp. 1-8.

[27] Baloyi, M. G., Rananga, S., & Mbooi, M. S. Fine-Tuning a Machine Learning Model to Detect Covid-19 Misinformation in Xitsonga. In 2025 IST-Africa Conference (IST-Africa), Nairobi, Kenya, 2025, pp. 1-10.

[28] World Health Organisation, "Infodemic," World Health Organisation, 2026. [Online]. Available: https://www.who.int/health-topics/infodemic [Accessed: Aug. 6, 2026].

[29] S. Bansal, N. S. Singh, S. S. Dar, and N. Kumar, "MMCFND: Multimodal Multilingual Caption-aware Fake News Detection for Low-resource Indic Languages," arXiv preprint arXiv:2410.10407, 2024.

[30] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, no. 1, article 6, Jan. 2020, doi: 10.1186/s12864-019-6413-7.