Scientific Research Publishing

Open Journal of Applied Sciences, 2024, 14, 511-532
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Smart Approaches to Efficient Text Mining for Categorizing Sexual Reproductive Health Short Messages into Key Themes

Tobias Makai, Mayumbo Nyirenda

Department of Computer Science, University of Zambia, Lusaka, Zambia
Email: tobias.makai@cs.unza.zm, mayumbo.nyirenda@cs.unza.zm

How to cite this paper: Makai, T. and Nyirenda, M. (2024) Smart Approaches to Efficient Text Mining for Categorizing Sexual Reproductive Health Short Messages into Key Themes. *Open Journal of Applied Sciences*, **14**, 511-532.
https://doi.org/10.4236/ojapps.2024.142037

Received: January 2, 2024
Accepted: February 26, 2024
Published: February 29, 2024

Abstract

To promote behavioral change among adolescents in Zambia, the National HIV/AIDS/STI/TB Council, in collaboration with UNICEF, developed the Zambia U-Report platform. This platform provides young people with improved access to information on various Sexual Reproductive Health topics through Short Messaging Service (SMS) messages. Over the years, the platform has accumulated millions of incoming and outgoing messages, which need to be categorized into key thematic areas for better tracking of sexual reproductive health knowledge gaps among young people. The current manual categorization process of these text messages is inefficient and time-consuming and this study aims to automate the process for improved analysis using text-mining techniques. Firstly, the study investigates the current text message categorization process and identifies a list of categories adopted by counselors over time which are then used to build and train a categorization model. Secondly, the study presents a proof of concept tool that automates the categorization of U-report messages into key thematic areas using the developed categorization model. Finally, it compares the performance and effectiveness of the developed proof of concept tool against the manual system. The study used a dataset comprising 206,625 text messages. The current process would take roughly 2.82 years to categorise this dataset whereas the trained SVM model would require only 6.4 minutes while achieving an accuracy of 70.4% demonstrating that the automated method is significantly faster, more scalable, and consistent when compared to the current manual categorization. These advantages make the SVM model a more efficient and effective tool for categorizing large unstructured text datasets. These results and the proof-of-concept tool developed demonstrate the potential for enhancing the efficiency and accuracy of message categorization on the Zambia U-report platform and other similar text messages-based platforms.

Keywords

Knowledge Discovery in Text (KDT), Sexual Reproductive Health (SRH), Text Categorization, Text Classification, Text Extraction, Text Mining, Feature Extraction, Automated Classification Process, Performance, Stemming and Lemmatization, Natural Language Processing (NLP)

1. Introduction

Adolescence is a critical period of transition from childhood to adulthood, presenting various challenges for young people in African countries, including Zambia. Among these challenges are high HIV prevalence rates, sexual abuse, peer pressure, early and unprotected sex, early marriages, and unwanted pregnancies all made worse by insufficient knowledge of Sexual Reproductive Health (SRH). Many parents in Zambia are reluctant to discuss crucial issues such as sexuality and sexual reproductive health with their children [1] [2] [3], leading adolescents to seek information from social media and peers, which are unreliable sources. To address this issue, the National HIV/AIDS/STI/TB Council (NAC) in collaboration with UNICEF implemented the Zambia U-report, an interactive system for sharing SRH information via Short Message Service (SMS) with subscribers. The platform aims to promote behavioural change by providing young people with access to vital information on various SRH topics and enabling them to ask questions on specific issues through SMS using their mobile phones.

Despite the success of the Zambia U-report platform in reaching over 200,000 subscribers, the large volume of textual data generated presents a challenge for NAC and UNICEF in categorizing the data into relevant thematic areas for decision-making and impact assessment. The current manual categorization process is time-consuming and inefficient, prompting the need for an automated solution.

This study seeks to provide a proof of concept for an automated categorization model for incoming SMS text messages on the Zambia U-report platform using text mining techniques and machine learning, to improve the efficiency of the analysis process and enable NAC and UNICEF to better track knowledge gaps and emerging issues related to SRH among young people in Zambia by applying appropriate technologies and techniques in the fields of Computer Science, including Artificial Intelligence (AI), Text Mining (TM), Machine Learning (ML), and Natural Language Processing (NLP), the study seeks to employ computers for tasks such as information extraction, categorization, summarization, and topic tracing. This approach will uncover patterns in digitally available textual data, ultimately informing decision-making and improving the efficiency of the U-report platform's textual data analysis process. A theoretical review of the aforementioned technologies detailing how they relate to our problem is presented in this section.

2. Related Technologies

This study aims to automate the categorization of incoming SMS text messages from Zambia U-report subscribers into key thematic areas for better analysis and data-driven decision-making using text-mining techniques. In this section, we highlight artificial intelligence and machine learning techniques that are relevant to this study. For a detailed discussion, the reader is encouraged to explore the cited literature.

2.1. Artificial Intelligence

Artificial Intelligence (AI) is the field that studies the synthesis and analysis of computational agents that act intelligently. A computational agent is an agent whose decisions about its actions can be explained in terms of computation [4]. An agent is something that acts in a given environment. Artificial Intelligence can also be defined as the ability of a machine to learn from experience, adjust to new inputs and perform human-like tasks [5]. The human-like tasks for our assignment involve reading through a humongous pool of SMS text messages, analyzing them one after another while allocating an appropriate category to each. AI techniques such as Text Mining, Machine Learning, and Natural Language Processing are employed to automate the categorization of the Zambia U-report SMS text messages.

2.2. Text Mining

Text Mining (TM) refers to the process of extracting interesting and non-trivial information and knowledge from unstructured text [6]. The Text mining field has gained a great deal of attention in recent years due to the tremendous amount of text data generated from various sources such as social networks, patient records, health care insurance data, news outlets and so on [7]. Today, the web is the main source for the text (documents), the amount of textual information available to us is consistently increasing. Approximately 80% of the information of most organizations is stored in unstructured format (reports, email, views and news etc.). This shows that approximately 90% of the world's data is held in unstructured formats [8] including over 5.5 million SMS text messages available on the Zambia U-report platform [9]. Being unstructured, an overload of textual data is significantly hard to process for decision-making. Therefore, text mining undertakes to aid automatic information extraction from a pool of text data. Other terms such as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT) can be used to also refer to text mining [6].

Text mining encompasses multiple aspects including information retrieval, information extraction, text categorization, text summarization, text clustering and visualization. Information Retrieval (IR) involves finding unstructured material that satisfies an information need within large collections, with web search engines being a common application [10] [11] [12]. Information Extraction (IE) scans text for relevant information, including entities, relations, and events, and

serves as an initial step in analyzing unstructured textual data [13] [14] [15], while Text Categorization, also known as Text Classification (TC), assigns text to predefined categories based on content, supporting tasks like email sorting and topic identification [16] [17].

The text categorization process encompasses several stages which include text acquisition and extraction [18], text analysis and labeling [19], feature extraction, construction, and weighting [20], feature selection and projection [20] [21], training of classification model [22] and solution evaluation [18] [22] [23].

2.3. Machine Learning

Machine learning is the study of algorithms that automatically improve their performance with experience [24]. It is an artificial learning approach that involves developing programs that learn from past data, and, as such, is a branch of data processing and artificial intelligence. Machine learning involves the use of computing to design systems that can learn from data in a manner of being trained. The systems might learn and improve with experience, and with time, refine a model that can be used to predict outcomes of questions based on previous learning [25].

There are four variants to Machine learning, namely; supervised, unsupervised, semi-supervised and reinforcement machine learning. Supervised learning algorithms model relationships and dependencies between the target prediction output and the input features such that users can predict the output values for new data based on those relationships learned from the previous data sets. In Unsupervised learning, the computer is trained with unlabeled data by learning patterns in the data. Semi-supervised learning falls in between supervised and unsupervised learning. Semi-supervised algorithms are the best candidates for building models where labels are absent in the majority of the observations but present in few. The reinforcement learning method aims at using observations gathered from the interaction with the environment to take actions that would maximize a reward or minimize risk and maximize performance [26].

Our focus in this study is to employ the use of supervised machine learning algorithms to build an automatic text classification model. In supervised learning, the algorithm is given an input and an output and the goal is to find a mapping between the two which generalizes well to new input [27]. There have been many text classification algorithms that researchers have used to solve text classification problems. Some algorithms explored in this study are; Naive Bayes [28], Support Vector Machines [29], Regression-Based Classifiers [30] [31], Decision Trees [32] [33], and K-nearest Neighbors among others [34] [35]. Various studies have used machine learning to successfully solve problems in disparate disciplines [36] [37] [38]

2.4. Natural Language Processing

Natural Language Processing (NLP) is an interdisciplinary field of linguistics,

computer science, and artificial intelligence that studies the use of computers to automatically analyze, understand, and generate human language in spoken or written form [39]. Natural Language Processing involves translating natural language into data that a computer can use to learn about the world [40] *i.e.* automatic processing of text that is written in a natural language such as English or Swahili. NLP encompasses a wide range of tasks, from low-level tasks, such as segmenting text into sentences and words, to high-level complex applications such as semantic annotation and opinion mining [41]. Automatic Text categorization can be achieved by implementing a series of Natural Language Processing tasks.

There are two broad categories of Natural Language Processing, namely Natural Language Understanding (NLU) and Natural Language Generation (NLG). Natural Language Understanding refers to the identification of the desired semantic from various possible semantics derived from a natural language expression [42] whereas Natural Language Generation is the process of transforming structured data into natural language. Computers using regular none Artificially Intelligent algorithms are not very capable of processing natural language as new algorithms would have to be developed with every new document or set of words introduced [43]. Therefore, to achieve the various Natural Language Processing tasks of extracting and understanding information from natural language, Machine Learning (ML) techniques are usually employed. Machine learning techniques among other things enable the creation and implementation of rules to decipher any new text. Natural language processing for text categorization is achieved through two main sub-processes. These include; cleaning and preprocessing the text and using Artificial Intelligence (AI) to understand and generate language [44].

Among the key aspects of text cleaning and preprocessing are; text segmentation and tokenization [43] [45] [46], stemming [44] [47], lemmatization [44] [47], tagging part-of-speech [43] [44], dependency parsing [44] [48], named entity recognition [49] [50] and topic modeling [44] [51].

3. Related Works

The problem that this study addresses is similar to many text classification problems such as; language detection, topic categorization, sentiment analysis, profanity and abuse detection, opinion mining and so on. Approaches used in the aforementioned tasks can be adopted for our quest to achieve the automatic categorization of sexual reproductive health short message texts into thematic areas.

Rosa and Ellen [52] used machine learning classification methods to experiment with "micro-text" chat entries in military chat rooms using four different classifiers namely; Support Vector Machine (SVM), k-Nearest Neighbor (k-NN), Rocchio and Naïve Bayes (NB). Their research solves a problem similar to the one presented in this study. The study used a dataset of thousands of individual

lines of text from a synthetic military chat. The pool of 19,898 posts with 9995 used for training and 9903 for testing was categorized into five categories namely; filler, binary, numeric, class-value and text, described in [52]. Results of this study suggest that k-NN and SVM are well suited for categorizing synthetic military chat data and therefore show that the classifiers can be used to solve similar problems in the text classification.

Balabantaray *et al.* [53] designed an emotion classifier model trained using a Support Vector Machine (SVM) to classify text extracted from Twitter into emotional categories such as sad, anger, disgust, happy, fear and surprise to assess the possible emotions of the persons behind the individual. Our work is similar because it also focuses on classification of textual data using Machine Learning methods. Balabantaray *et al.* set out to extend an existing emotion classifier that classified text into only three categories namely positive, negative and neutral to classify into the aforementioned six basic emotion classes. A dataset of 8150 tweets was used in the experiment. The automatic emotion classification problem, which is also a classic text classification problem, and utilized the SVM algorithm in identifying emotion-bearing words in sentences was solved with a 73.24% accuracy.

In a quest to categorize the flow of conversation in counseling, Y. Hayashida *et al.* [54] applied Support Vector Machine (SVM) to category classification of counselling text data. They used a dataset comprised of conversations between clients and beginner counselors. The main goal for the study was to automate the process that normally involves a supervisor taking a look at direct transcripts of verbatim records of a given counseling session and coaching someone for beginner counselor. Y. Hayashida *et al.* implemented a category classification model based in SVM that achieved an accuracy rate of 63.5%.

C. Poulin *et al.* [55] developed text classification models using supervised machine learning to detect the risk of suicide from unstructured clinical notes taken from a national sample of U.S. Veterans Administration (VA) medical records. The study successfully determined useful text-based signals of suicide at accuracies above 60% for ensemble averages of 100 models.

B. Koopman *et al.* [56] adopted supervised machine learning and rule-based approaches to automate the classification of diseases from free-text death certificates into four diseases of interest; including diabetes, influenza, pneumonia and HIV for real-time surveillance.

4. Methodology

To meet the objectives of the study, the following key steps were undertaken:

1) Understanding the methods currently used to categorize text messages on the platform into key thematic areas.

2) Extracting and understanding the textual data.

3) Data preparation and pre-processing.

4) Text classification modeling.

5) Evaluation and deployment of the model.

An exploratory research method was used and the steps were met through semi-structured interviews, observation and document analysis.

To understand the methods currently used to categorize information on the Zambia U-report platform into key thematic areas, we interviewed counselors at the National HIV/AIDS/STI/TB Council (NAC) responsible for the task based on a semi-structured interview guide that was developed. Two counselors were interviewed independently to ensure that the data was unbiased. The tools used to categorize the textual data were closely observed.

To extract and understand the textual data, a copy of the Zambia U-report database, which contained all Zambia U-report data, including over 5,987,040 incoming and outgoing SMS text messages from the year 2012 to 2019 was obtained as a raw Structured Query Language (SQL) from the National HIV/AIDS Council (NAC). The database dump was then queried further to access the messages and extract a subset for pre-processing.

After the dataset was successfully extracted as a CSV file, the data preparation process began. Data preparation is a vital stage in text classification and involves converting the raw textual data into a format that can be used for machine learning. The preprocessing step involved data sampling and labeling, text cleaning and feature extraction and data splitting.

4.1. Data Sampling and Labeling

This step involved assigning categories to each text message so that supervised learning algorithms could learn patterns and make predictions. From the CSV dataset, a sub-dataset of 50,877 incoming text messages was extracted and manually assigned categories with the assistance of domain experts. This helped researchers obtain training on labeling messages, and as a result, 1,770 messages were labeled by researchers to improve their labeling skills. Two domain experts labeled the same dataset, from which an initial model was built using this labeled data. The prediction accuracy of the model was measured. We labeled the same dataset to assess our ability to label text messages consistently with domain experts. Accuracy was measured using the preliminary machine learning classification model that was trained on the initial model.

An independent dataset of 4567 expert-labeled U-report SMS text messages generated from 1st July to 31st September 2020 was used to verify the trained model and establish that the researchers can label like domain experts. We then moved on to labelling a total of 50,877 U-report SMS text messages with the help of domain experts to create the training dataset.

4.2. Text Cleaning

This stage aimed at improving data quality by removing unnecessary words and characters, otherwise referred to as noise. The text messages contained various types of noise, such as punctuations, and stop words like “a”, “the”, “is”, and

"are" which were removed to ensure the model accurately identified relevant information. The cleaning process also involved lemmatization, which reduced the dimensionality of text messages by grouping words with similar meanings and converting them to their root forms [57]. This process improved the accuracy and quality of our textual dataset and ultimately enhanced our text classification model's accuracy by subjecting the machine learning process to fewer unique words.

4.3. Feature Extraction and Data Splitting

Feature extraction entails converting the text into numerical data that can be understood by machine learning algorithms. To convert the cleaned text into numerical feature representation, the 'TfidfVectorizer' class from the scikit-learn library was used to transform the text data into a numeric matrix. It computed term frequency-inverse document frequency (TF-IDF) scores for each word, reflecting the importance of terms within a text relative to their importance across the dataset.

The final step in data preparation was splitting the data into training and test datasets to prevent overfitting and ensure accurate generalization to new data. Using the train_test_split method from the scikit-learn library, the dataset was shuffled, split into input (text message data) and target (label column) variables, and then divided into training and test subsets with a test size of 0.2 (20%). The training data was used to fit the TfidfVectorizer, which transformed the textual data into numerical form for machine learning models. The transformed training and test data were stored and used as input for machine learning algorithms.

Following data preparation, a text classification model was developed using Python and machine learning algorithms from the Scikit-learn library. The creation of the model involved selecting suitable algorithms and training them with the preprocessed dataset. Multiple algorithms, including K-nearest Neighbor, Decision Tree, Multinomial Naive Bayes, Support Vector Classification (SVC), Support Vector Machines (SVM), Random Forest Classifier, and Stochastic Gradient Descent (SGD) Classifier, were used due to their effectiveness in text classification tasks.

These algorithms were trained with the labeled training dataset, and their performance in predicting labels was evaluated. This process was iterated several times to find the most accurate model, which was eventually saved as the final model. The dataset of 50,877 labeled U-report SMS text messages was split into training and testing sub datasets in an 80:20 ratio. The training dataset was used to train the machine learning algorithm while the test dataset was used to evaluate the model's performance. After an acceptable level of accuracy was achieved, the model was used to classify 206,625 U-report messages previously unseen.

The final stage of the modeling process involved assessing the performance of the trained classification model. The multiple machine learning algorithms trained were evaluated on accuracy, precision, recall, and F1 scores, collectively

determining the overall performance of each model. Accuracy measures the ratio of correct predictions over the total number of instances evaluated [58]. Precision represents the proportion of true positive predictions out of all positive predictions, while recall represents the proportion of true positive predictions out of all actual positive instances. The F1-score is the harmonic mean of precision and recall.

The final stage of the modeling process involves assessing the performance of the trained classification model. The multiple machine learning algorithms trained were evaluated on accuracy, precision, recall, and F1 scores, collectively determining the overall performance of each model. The best-performing model is what was successfully applied on the unseen messages. A comparative analysis was then conducted between the automated process and the current manual classification method to determine which process was better.

5. Results and Discussion

In a quest to automate the classification of the Zambia U-report SMS text messages into key thematic areas, the study assessed the current manual process and its limitations, evaluated the researchers' ability to label text messages accurately and consistently like domain experts, compared the previously manually labeled dataset by the experts against the dataset predicted by the machine learning model that was trained using a dataset that was trained by researchers after assessing their labeling consistency with experts on the initial dataset. The research also involved training and assessing multiple machine learning algorithms to arrive at the final model. A categorization report and confusion matrix were also generated to illustrate the model's performance. Finally, the benefits of using the machine learning model over the manual process, in terms of time and efficiency, were analyzed and discussed in the study.

5.1. Current Categorization Model

The current text classification process is handled manually by counsellors employed by the National HIV/AIDS/STI/TB Council (NAC) and involves categorizing messages using a tally sheet based on each message's core theme. Counsellors assess and count each message received on the U-report platform under a specific category immediately after responding to it. This tally sheet (shown in **Figure 1**), consisting of various categories and tallies for each, is generated to record counts for incoming SMS messages determined to belong to a topic. Tools used to create tallies are notepads and pens. The process solely relies on the counsellors' expert assessments hence the need for a more efficient and automated solution.

Monthly reports are compiled by aggregating all daily tallies and are presented to the U-report Core Group, a special committee of stakeholders chaired by NAC to aid decision-making. If a notepad is lost or damaged, accurately reporting becomes difficult as experts have to retrieve the messages from the platform. A task so difficult as the platform lacks filters that could be used to find old

messages. This cements the need for a more efficient system. Figure 2 is an example of the monthly aggregate for April 2019, representing aggregated tallies from all three counsellors.

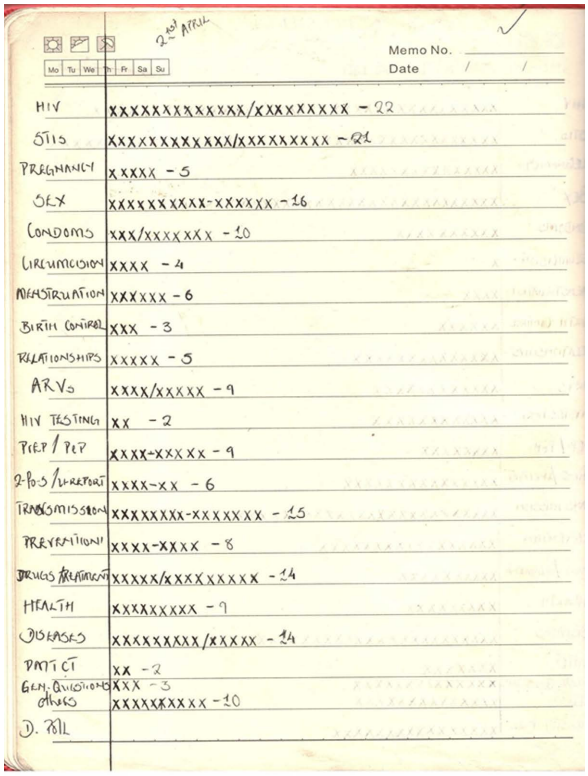

Figure 1. Primary topic tally sheet for U-Report messages created on 21st April 2019.

	1st - 7th April	7th - 20th April	21st - 28th April
HIV	17	37	22
STIs	14	42	21
Pregnancy	4	15	5
Sex	11	34	16
Condoms	5	10	10
Circumcission	2	1	4
Menstruation	2	4	6
Birth Control	2	6	3
Relationships	3	14	5
ARVs	7	12	9
HIV Testing	4	12	2
PrEP/PeP	1	8	9
2-for-5/U-Report	12	15	6
Transmission	15	28	15
Prevention	6	17	8
Drugs/Treatment	1	10	14
Health	6	9	9
Diseases	9	24	14
PMTCT	3	7	2
Gen Questions	13	14	3
Others	4	14	10
Masturbation	1	16	

Figure 2. Sample monthly aggregate for April 2019 resulting from the manual categorization of the text messages.

It was observed that each counselor requires approximately 15 minutes to finish their duties on the initial message and 5 minutes on succeeding messages. The breakdown of the duration is as follows: approximately 10 minutes to jot down the categories on a notepad, 3 minutes to reply to the message, and 2 minutes to categorize the message. The counselor must first analyze the message before assigning it to a category. In some cases, due to uncertainties arising from shorthand text, the use of the local language, environmental disruptions, and platform downtimes, the duration may increase. These durations are presented in Table 1.

The described flow is summarized in Figure 3.

Table 1. Table showing the average time spent by each counsellor responding to the first message.

Activity duration	Counsellor 1	Counsellor 2	Average
Listing topics	10.5 mins	9.8 mins	10.15 mins
Responding to text message	2.7 mins	3.4 mins	3.05 mins
Categorizing text message	1.7 mins	2.4 mins	2.05 mins
Total			**15.25 mins**

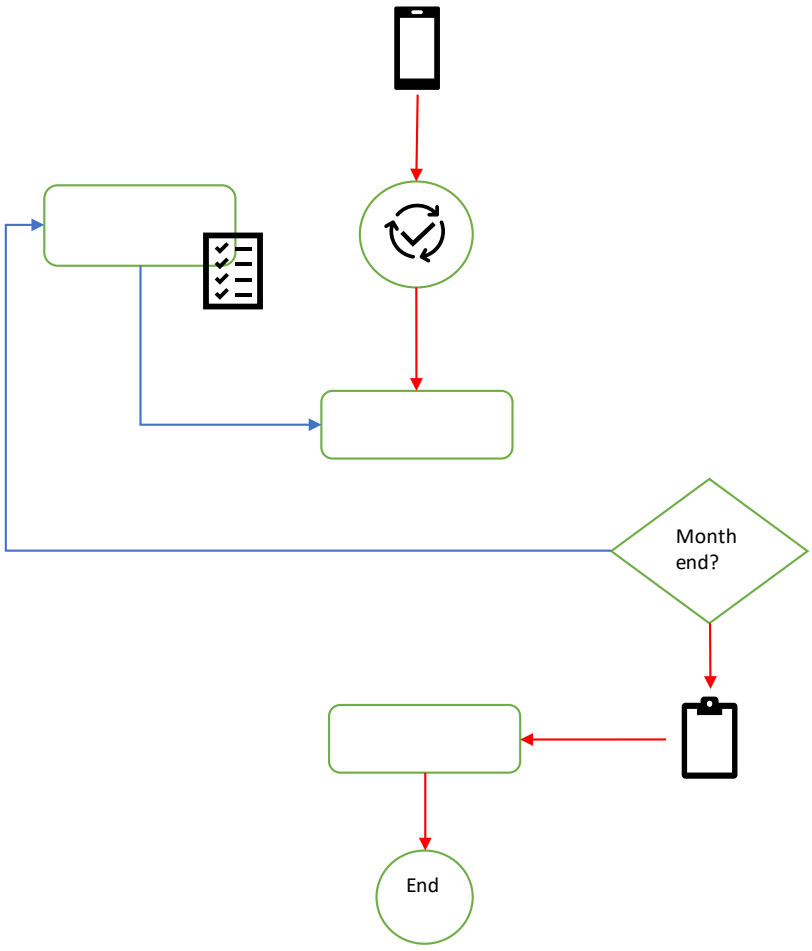

Figure 3. Model for the current manual categorization of U-Report messages.

The study revealed that categorizing 5,987,040 SMS text messages using the current manual process would take approximately 82 years assuming the counsellors worked based on International Labour Standards (ILO) [58] of 48 hours per week focusing only on categorizing the messages. This amount of time is impractical, thereby justifying the need for more efficient methods, such as the use of machine learning algorithms, to significantly improve the speed and accuracy of categorizing U-report SMS text messages.

5.2. Data Sampling and Labeling Outcomes

During the data preparation stage of the model development, we were trained on how to label U-report messages by counsellors (domain experts) after which our capacity to accurately and consistently annotate the messages was compared with them. A dataset of 1770 messages was labelled by the experts and researchers, respectively. For each labelled dataset, a model was trained and its performance measured. The findings presented in **Table 2**, reveal that the researchers' annotation was as consistent as that of the domain experts, indicating that the training had a considerable impact on our ability to label more U-report text messages to use during the development process of the machine learning model.

The machine learning model (SVM) was evaluated using a dataset manually labeled by specialists for the National HIV/AIDS/STI/TB Committee in 2020. The assessment aimed to further determine if the model could predict message categories comparably to domain experts. The dataset, covering July 1st to September 31st, 2020, included an additional topic, "Covid-19," not present in the trained model. As shown in **Figure 4**, the manual categorization by experts closely aligned with the models' predictions for topics like condoms, testing, prevention, STIs, and others. This consistency is further illustrated in **Figure 5**.

Results indicate the effectiveness of training non-expert annotators to assist in labeling text messages with a consistency comparable to domain experts. Trained

Table 2. Measuring the researchers' ability to label text messages to achieve the same accuracy and consistency as domain experts after being trained.

ALGORITHM	ACCURACY		
	EXPERT 1	EXPERT 2	RESEARCHERS
K-Nearest Neighbor	0.487759	0.487759	0.463277
Decision Tree	0.525424	0.521657	0.527307
Multinomial Naive Bayes	0.514124	0.506591	0.483992
Support Vector Classification (SVC)	0.563089	0.568738	0.553672
Support Vector Machines (SVM)	0.581921	0.587571	0.580038
Random Forest Classifier	0.548023	0.572505	0.542373
Stochastic Gradient Descent (SGD) Classifier	0.574388	0.566855	0.563089
K-Nearest Neighbor	0.487759	0.487759	0.463277

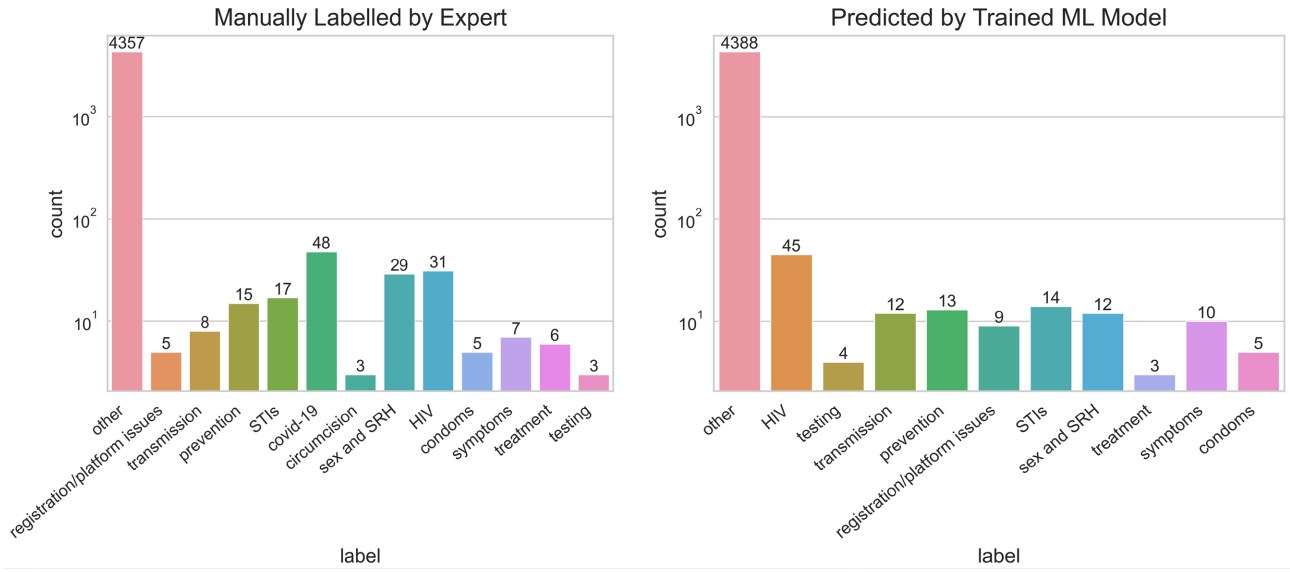

Figure 4. Model for the current manual categorization of U-Report messages.

Message	Predictions	expert label
thank	0	other
hiv spread	3	HIV
question covid going prevent infiniteve disease	6	prevention
went first covid said didn't know epectmy turm finaly came medical person got smear noseyou imagine supprosedand ocurri went chelstone clinic something ever covid test mandatoryi found interetingwill touch	11	covid-19
yet circumcised se one use comdoms contract hiv aids	12	transmission
thanks bt u stop invitus workshop u report zambia need learn a lot hiv aids stis gbv se reproductive health	8	sex SRH
yearscan male se	1	registration/platform issues
afternoon causes tiny pimple like bubbles around dick head weeks circumcisionnote	9	STIs
signs siferis sti	5	symptoms
allowed use two condoms sex	2	condoms

Legend
Cervical cancer - 0
Circumcision -1
Condoms - 2
HIV - 3
Masturbation - 4
Other - 5
Prevention - 6
Registration/Platform issues - 7
Sex and SRH - 8
STIs - 9
Symptoms - 10
Testing - 11
Transmission - 12
Treatment - 13

Figure 5. Sample of expert labels vs machine prediction labels. Model was trained from a researcher-labeled dataset.

non-experts are a viable alternative to support domain experts in labeling tasks as they can enhance efficiency and reduce costs in the development process of text classification machine learning models

5.3. Model Training and Algorithm Selection

Presented in this section, are evaluation results of the prediction accuracy of

various algorithms used in the process of determining the final model for our text classification tasks. The model creation involved selecting multiple suitable algorithms and training them on a preprocessed dataset. The final model training and selection process, illustrated in Figure 6, utilized a training dataset of 50,877 U-report incoming SMS text messages labeled by two domain experts and the researcher. Figure 7 presents the results of the dataset preparation step.

The above steps were aimed at creating a model that accurately categorizes U-report SMS text messages into identified key thematic areas. The prediction accuracies of each algorithm were analyzed and the best model was chosen as the final model for further analysis. The evaluation of prediction accuracies for each algorithm trained to determine the final model is presented in Table 3.

The evaluation of prediction accuracies for each algorithm trained to determine the final model are presented in Table 3.

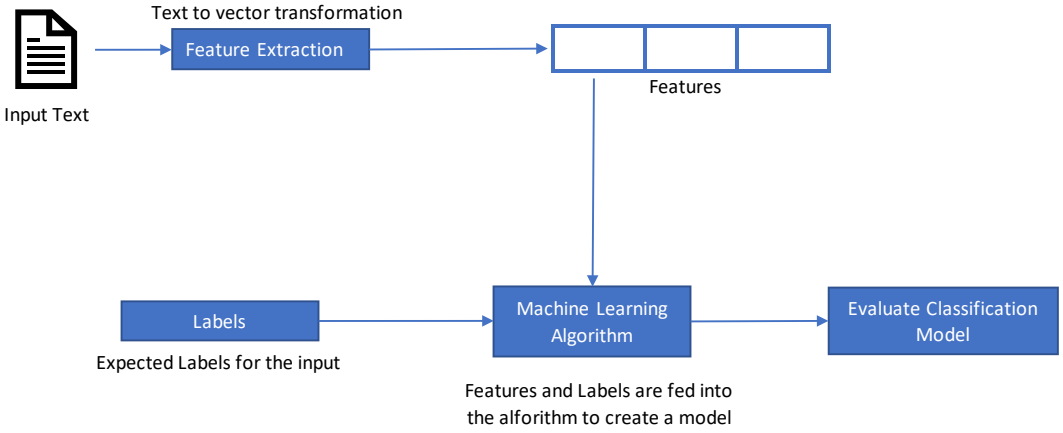

Figure 6. Model training and selection process.

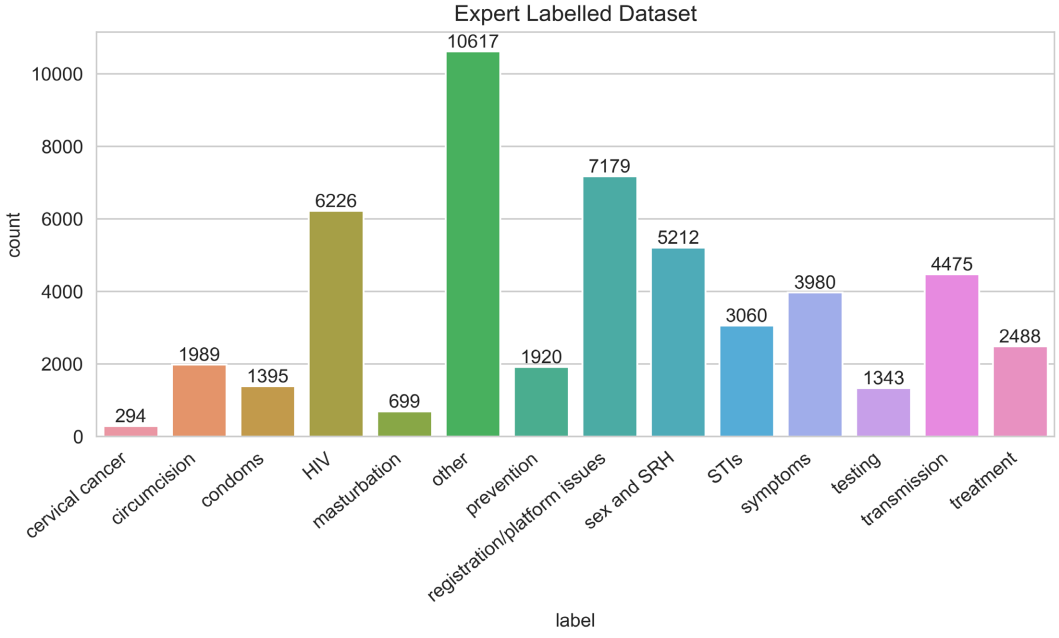

Figure 7. Final dataset used for the final model training and selection.

Table 3. Prediction accuracies of machine learning algorithms used for final model selection.

ALGORITHM	ACCURACY
K-Nearest Neighbor	51.2%
Decision Tree	63.7%
Multinomial Naive Bayes	64.8%
Support Vector Classification (SVC)	70.1%
Support Vector Machines (SVM)	70.4%
Random Forest Classifier	66.5%
Stochastic Gradient Descent (SGD) Classifier	69.8%

Among the machine learning algorithms assessed, the SVM algorithm emerged as the most accurate, with a 70.4% accuracy in classifying U-report SMS text messages into various categories. The decision tree and multinomial naive Bayes algorithms also performed relatively well, achieving accuracies of 63.7% and 64.8%, respectively. However, the K-nearest neighbor algorithm, with a 51.2% accuracy rate, proved less suitable for this task. The SVM algorithm was selected as the final machine learning model for further analysis.

5.4. Performance Analysis of the Final Machine Learning Model

The performance of the final machine learning model in categorizing U-report SMS text messages was analyzed and results were used to produce a categorization report showing precision, recall, F1 score, and support for each category, as well as overall accuracy, macro-average, and weighted-average. The report is presented as Table 4 and demonstrates the effectiveness of the SVM model in classifying messages into various categories, such as HIV, transmission, symptoms, and prevention, among others. The precision, recall, and F1 score for each category provide valuable insights into the model's performance, proving its potential to enhance the categorization process for U-report SMS text messages. Also established, are benefits of using the machine learning model over the manual process in terms of time and efficiency.

The report reveals the SVM model's overall accuracy at 70%, correctly classifying text messages into their respective categories. With macro-average and weighted-average F1-scores at 68% and 70%, the model demonstrates strong performance across all categories. The highest precision and recall scores were observed in STIs (94%) and sex and SRH (82%) categories, respectively. A confusion matrix provides a visual representation of the model's performance, depicting true positive, true negative, false positive, and false negative predictions for each category. The confusion matrix, shown in Figure 8, enables a deeper analysis of the model's strengths and weaknesses, contributing to the calculation of performance metrics like precision, recall, and F1-score. With an overall accuracy of 70%, the model performs well across all categories. To further assess the SVM model, it was applied to a new, unseen dataset of 206,625 text messages

Table 4. Categorization report for the best model (SVM).

CATEGORY	PRECISION	RECALL	F1-SCORE	SUPPORT
HIV	0.51	0.62	0.56	60
Transmission	0.80	0.73	0.76	384
Symptoms	0.63	0.79	0.70	256
Registration/Platform Issues	0.66	0.72	0.69	1237
Circumcision	0.81	0.61	0.69	149
Sex and SRH	0.67	0.82	0.74	2167
Other	0.65	0.58	0.61	345
STIs	0.94	0.81	0.87	1497
Prevention	0.66	0.59	0.62	1031
Condoms	0.63	0.59	0.61	643
Cervical Cancer	0.66	0.62	0.64	793
Treatment	0.68	0.65	0.66	266
Masturbation	0.66	0.62	0.64	848
Testing	0.72	0.61	0.66	500
Accuracy			0.70	10176
Macro Average	0.69	0.67	0.68	10176

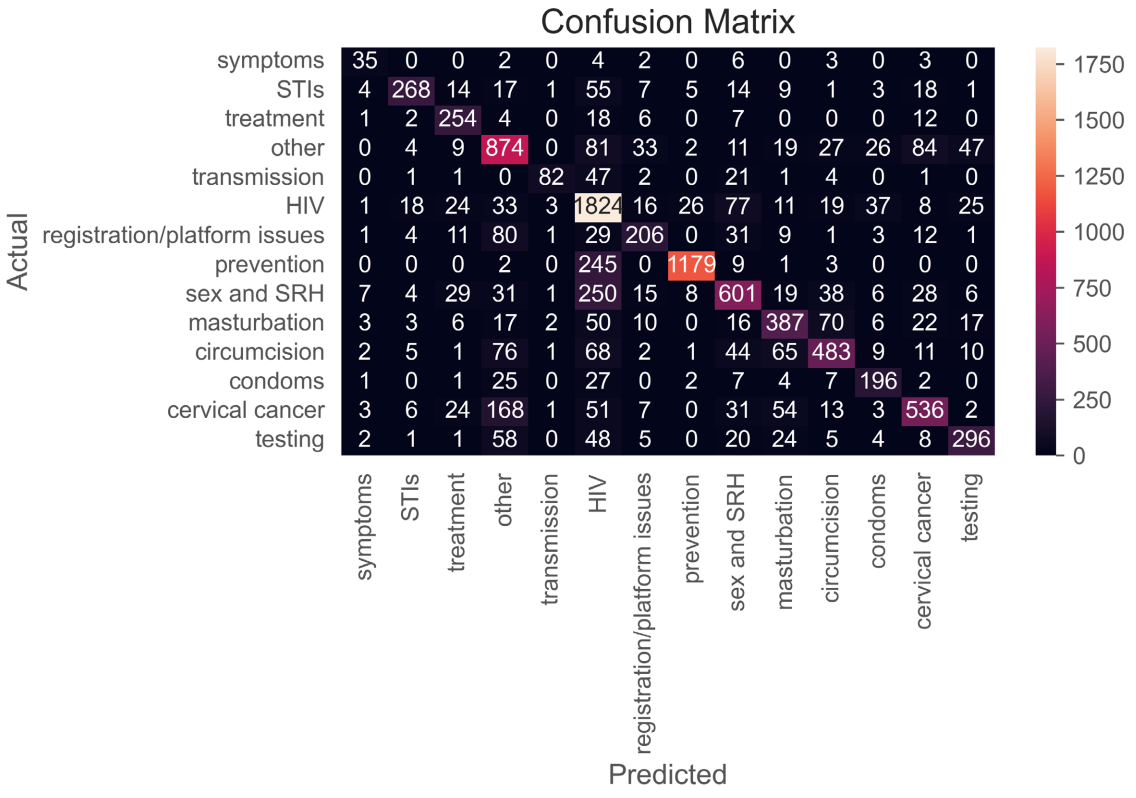

Figure 8. Confusion matrix.

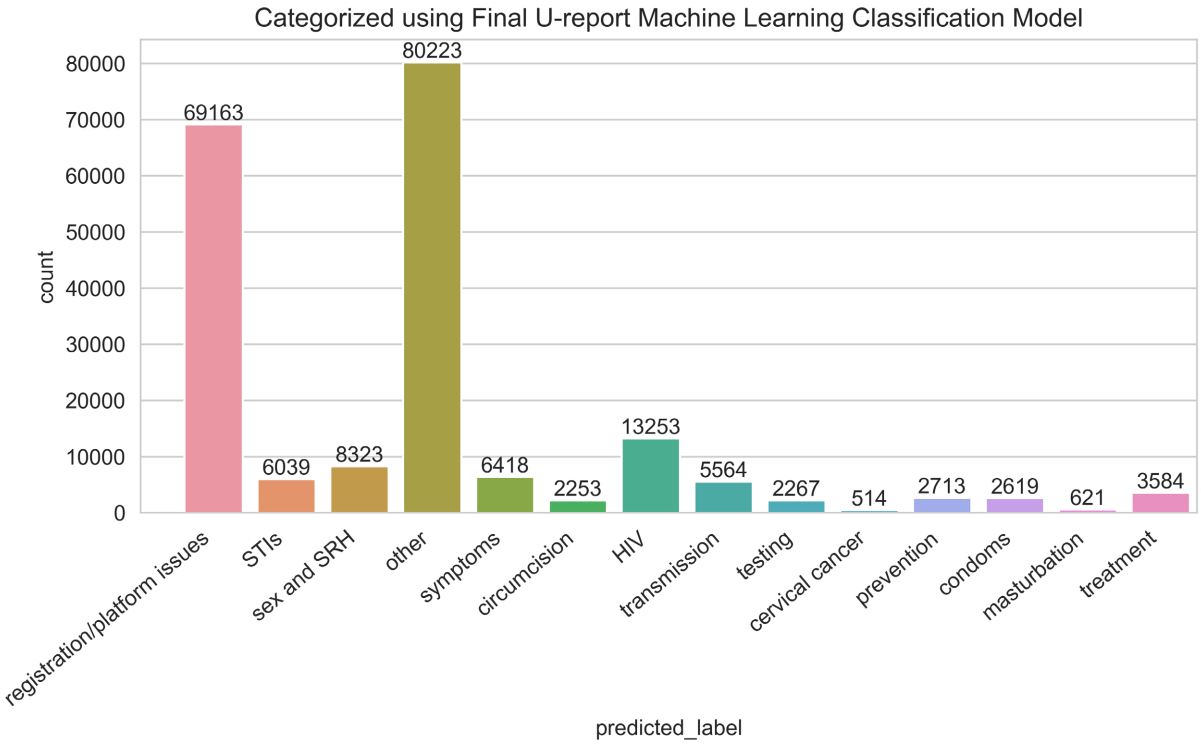

Figure 9. 206,625 Zambia U-report SMS messages categorized using the final model.

from November 2012 to July 2019. Categories were distributed as shown in Figure 9.

Figure 10 groups the resulting categories by year, showing the distribution of messages across categories for each year during the period specified.

The study revealed that manually categorizing the 206,625 SMS text messages dataset would take roughly 2.82 years, based on International Labour Standards on working hours. Using the trained SVM model would require only 6.4 minutes to categorize the messages, amounting to approximately 18.6 milliseconds per message. If applied to 5,987,040 messages, the automated model would take just 3.1 hours. Demonstrating that the automated method is significantly faster and more scalable compared to the manual categorization. Unlike manual categorization which can suffer from classification inconsistencies, the trained model produces more consistent results. These advantages make the SVM model a more efficient and effective tool for categorizing large unstructured text datasets.

6. Conclusions and Recommendations

This research shows that it is feasible to use supervised machine learning to automate the categorization of SMS text messages on the Zambia U-report platform, achieving 70.4% accuracy and significantly reducing categorization time. The study presents a proof of concept, highlighting the potential benefits of employing machine learning for text classification tasks. By reducing the time spent on categorization, National HIV/AIDS/STI/TB Council (NAC) staff can focus on more crucial tasks and make quicker, data-driven decisions.

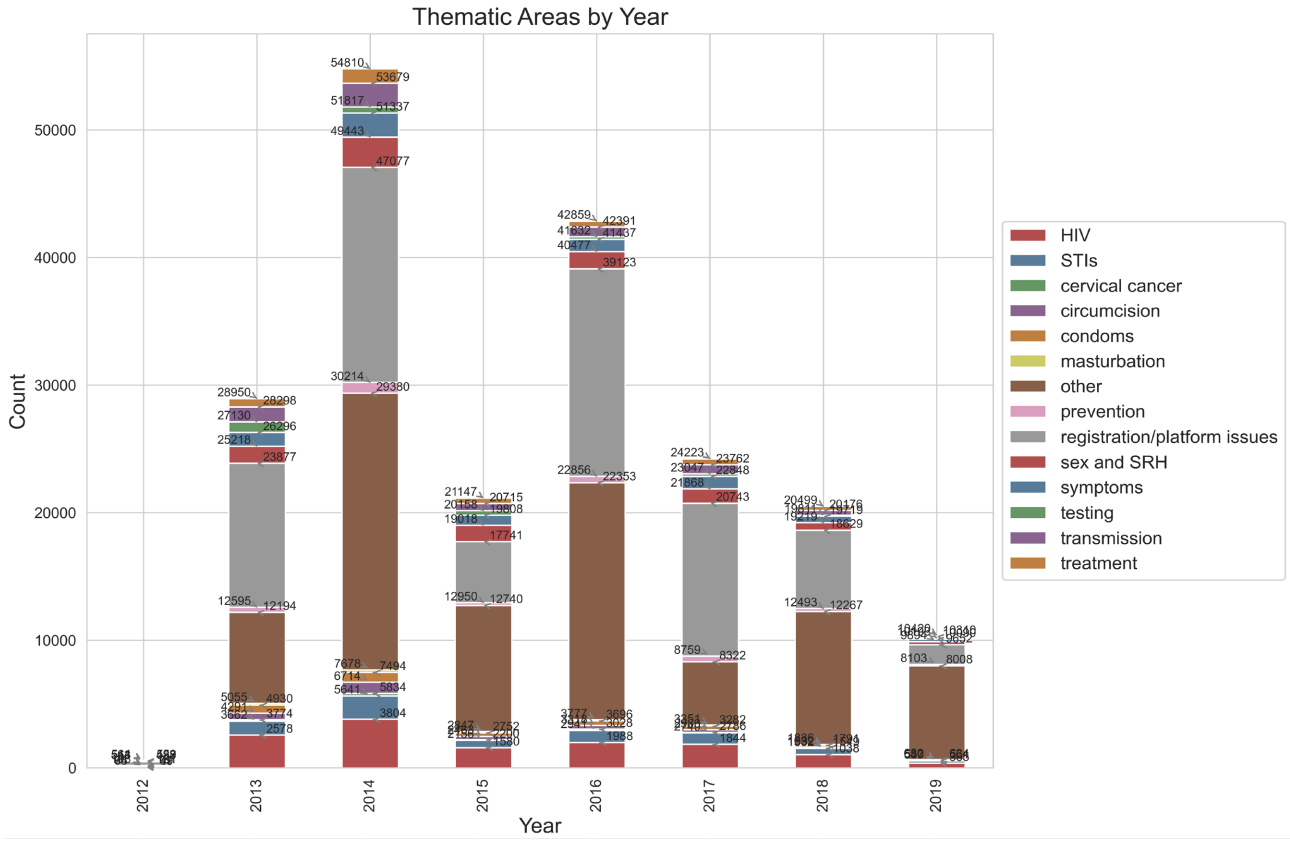

Figure 10. Key thematic areas for Zambia U-report messages disaggregated by year.

To enhance the model's accuracy, continuous training and expansion of the model's training dataset size are recommended. The study demonstrates the potential of leveraging machine learning to strengthen data analysis capabilities and facilitate more informed decision-making.

Acknowledgments

We are profoundly grateful to the Zambia National HIV/AIDS/STI/TB Council (NAC) and its technical partners for allowing us to pursue this study and providing us with all the required Zambia U-report information.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Planned Parenthood Association of Zambia (2018) Passport to Health.

[2] World Health Organization Africa (2015) Report on the Regional Meeting to Take Stock of the Progress Made in Adolescent Sexual and Reproductive Health and Rights, in the 20 Years since the International Conference on Population and Development, and on the Opportunities and Challenges in Moving the Agenda Forward. Visualizing the Problems and Generating Solutions for Adolescent Health in

the African Region, Congo Brazzaville.
https://www.afro.who.int/sites/default/files/2018-05/ASRH-%20AFRO%20-%20AH%20workshop%20report.pdf

[3] Ministry of Health (2011) Adolescent Health Strategic Plan 2011 to 2015, Lusaka Zambia.
https://zambia.unfpa.org/sites/default/files/pub-pdf/ZambiaAdolescentHealthStrategicPlan2011-2015.pdf

[4] Poole, D.L. and Mackworth, A.K. (2017) What Is Artificial Intelligence? In: Poole, D.L. and Mackworth, A.K., Eds., *Artificial Intelligence*: *Foundations of Computational Agents*, 2nd Edition, Cambridge University Press, Cambridge, 9.
https://doi.org/10.1017/9781108164085

[5] Duan, Y., Edward, J.S. and Dwivedi, Y.K. (2019) Artificial Intelligence for Decision Making in the Era of Big Data-Evolution, Challenges and Research Agenda. *International Journal of Information Management*, **48**, 63-71.
https://doi.org/10.1016/j.ijinfomgt.2019.01.021

[6] Gupta, V. and Lehal, G.S. (2009) A Survey of Text Mining Techniques and Applications. *Journal of Emerging Technologies in Web Intelligence*, **1**, 60-76.
https://doi.org/10.4304/jetwi.1.1.60-76

[7] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D.T., Gutierrez, J.B. and Kochut, K. (2017) A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *Proceedings of KDD Bigdas*, Halifax, August 2017, 1-13.
https://arXiv:1707.02919v2

[8] Dang, S. and Ahmad, P. H. (2015) A Review of Text Mining Techniques Associated with Various Application Areas. *International Journal of Science and Research*, **4**, 2461-2466.

[9] Zambia U-Report Database December 2012 to July 2019.
https://www.unicef.org/esa/documents/zambia-u-reports

[10] Manning, C.D., Raghavan, P. and Schutze, H. (2008) Boolean Retrieval in Introduction to Information Retrieval. Cambridge University Press, New York, 1-17.

[11] What Is Information Retrieval? What Does Information Retrieval Mean?
https://www.youtube.com/watch?v=kVD54hmeTV8

[12] Ceri, S., Bozzon, A., Brambilla, M., Valle, E.D., Fraternali, P. and Quarteroni, S. (2013) The Information Retrieval Process. In: Ceri, S., Bozzon, A., Brambilla, M., Valle, E.D., Fraternali, P. and Quarteroni, S., Eds., *Web Information Retrieval*, Springer, Berlin, 13-26. https://doi.org/10.1007/978-3-642-39314-3

[13] Hobbs, J.R. and Rilo, E. (2010) Information Extraction. In: *Handbook of Natural Language Processing*, Chapman & Hall, Boca Raton, 511-532.

[14] Wimalasuriya, D.C. and Dou, D. (2010) Ontology-Based Information Extraction: An Introduction and a Survey of Current Approaches. *Journal of Information Science*, **36**, 306-323. https://doi.org/10.1177/0165551509360123

[15] Russell, S.J. and Norvig, P. (2016) Information Extraction. In: *Artificial Intelligence*: *A Modern Approach*, 3rd Edition, Pearson Education Limited, Harlow, 873-882.

[16] Korde, V. and Mahender, C.N. (2012) Text Classifications and Classifiers—A Survey. *International Journal of Artificial Intelligence & Applications*, **3**, 85-99.
https://doi.org/10.5121/ijaia.2012.3208

[17] Wei, G., Gao, X. and Wu, S. (2010) Study of Text Classification Methods for Data Sets with Huge Features. *Proceedings* 2*nd International Conference on Industrial and Information Systems*, Dalian, 10-11 July 2010, 433-436.

https://doi.org/10.1109/INDUSIS.2010.5565817

[18] Mironczuk, M.M. and Protasiewicz, J. (2018) A Recent Overview of the State-of-the-Art Elements of Text Classification. *Expert Systems with Applications*, **106**, 36-54. https://doi.org/10.1016/j.eswa.2018.03.058

[19] Altexsoft, Labelling Approaches, 29 Mar. 2018. https://www.altexsoft.com/blog/datascience/how-to-organize-data-labeling-for-machine-learning-approaches-and-tools/

[20] Lillywhite, K., Lee, D., Tippetts, B. and Archibald, J. (2013) A Feature Construction Method for General Object Recognition. *Pattern Recognition*, **46**, 3300-3314. https://doi.org/10.1016/j.patcog.2013.06.002

[21] Pandaya, D., Amorimb, R.C. and Lanea, P. (2018) Feature Weighting as a Tool for Unsupervised Feature Selection. *Information Processing Letters*, **129**, 44-52. https://doi.org/10.1016/j.ipl.2017.09.005

[22] Zafra, M.F. (2019, June 16) Text Classification in Python. Towards Data Science. https://towardsdatascience.com/text-classification-in-python-dd95d264c802

[23] Montejo-Raez, A. (2005) Automatic Text Categorization of Documents in the High Energy Physics Domain. MS Thesis, Universidad de Granada, Granada. https://hera.ugr.es/tesisugr/15903837.pdf

[24] Hall, M.A. (1999) Correlation-Based Feature Selection for Machine Learning. PhD Dissertation, Dept. of Computer Sc., Univ. of Waikato, Hamilton. https://www.lri.fr/~pierres/donn%E9es/save/these/articles/lpr-queue/hall99correlationbased.pdf

[25] Bell, J. (2015) What Is Machine Learning? In: *Machine Learning*: *Hands-On for Developers and Technical Professionals*, John Wiley & Sons, Inc., Hoboken, 1-16.

[26] Fumo, D. (2017, June 15) Types of Machine Learning Algorithms You Should Know. https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861

[27] Lundborg, A. (2017) Text Classification of Short Messages. M.S. Thesis, Dept. of Computer Sc., Lund University, Lund. https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=8928009&fileOId=8928011

[28] Bansal, S. (2013) A Comprehensive Guide to Understand and Implement Text Classification in Python. https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/

[29] Campbell, C. and Ying, Y. (2011) Learning with Support Vector Machines. Springer, Berlin, 1-21. https://doi.org/10.1007/978-3-031-01552-6_1

[30] Mood, C. (2009) Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do about It. *European Sociological Review*, **26**, 67-82. https://doi.org/10.1093/esr/jcp006

[31] Banerjee, M., Filson, C., Xia, R. and Miller, D.C. (2014) Logic Regression for Provider Effects on Kidney Cancer Treatment Delivery. *Computational and Mathematical Methods in Medicine*, **2014**, Article ID: 316935. https://doi.org/10.1155/2014/316935

[32] Scikit-Learn, Decision Trees. https://scikit-learn.org/stable/modules/tree.html

[33] Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A. and Brown, S.D. (2004) An Introduction to Decision Tree Modeling. *Journal of Chemometrics*, **18**, 275-285. https://doi.org/10.1002/cem.873

[34] Chen, Y. and Hao, Y. (2017) A Feature Weighted Support Vector Machine and K-Nearest Neighbor Algorithm for Stock Market Indices Prediction. *Expert Systems with Applications*, **80**, 340-355. https://doi.org/10.1016/j.eswa.2017.02.044

[35] Teixeira, L.A. and Inácio de Oliveira, A.L. (2010) A Method for Automatic Stock Trading Combining Technical Analysis and Nearest Neighbor Classification. *Expert Systems with Applications*, **37**, 6885-6890. https://doi.org/10.1016/j.eswa.2010.03.033

[36] Knox, K., Nyirenda, M. and Kabemba, M. (2019) Data Mining for Fraud Detection in Large Scale Financial Transactions. *Proceedings of the International Conference in ICT* (*ICICT*2019), Lusaka, 20-21 November 2019, 172-177.

[37] Chiwamba, S.H., Phiri, J., Nkunika, P.O.Y., Nyirenda, M., Kabemba, M.M. and Sohati, P.H. (2019) Machine Learning Algorithms for Automated Image Capture and Identification of Fall Armyworm (FAW) Moths. *Zambia ICT Journal*, **3**, 1-4. https://doi.org/10.33260/zictjournal.v3i1.69

[38] Chulu, F., Phiri, J., Nyirenda, M., Kabemba, M.M., Nkunika, P. and Chiwamba, S. (2019) Developing an Automatic Identification and Early Warning and Monitoring Web Based System of Fall Army Worm Based on Machine Learning in Developing Countries. *Zambia ICT Journal*, **3**, 13-20. https://doi.org/10.33260/zictjournal.v3i1.71

[39] Lu, X. (2018) Natural Language Processing and Intelligent Computer-Assisted Language Learning (ICALL). In: Liontas, J.I., Ed., *The TESOL Encyclopedia of English Language Teaching*, John Wiley & Sons, Inc., Hoboken, 1-6. https://doi.org/10.1002/9781118784235.eelt0422

[40] Lane, H., Howard, C. and Hapke, H.M. (2019) Natural Language vs. Programming Language. In: *Natural Language Processing*: *Understanding*, *Analyzing*, *and Generating Text with Python*, Manning Publications Co., Shelter Island, 3-30.

[41] Maynard, D., Bontcheva, K. and Augenstein, I. (2017) Introduction. In: Maynard, D., Bontcheva, K. and Augenstein, I., Eds., *Natural Language Processing for the Semantic Web*, Springer, Berlin, 1-8. https://doi.org/10.1007/978-3-031-79474-2_1

[42] Priyadarshini, S.B.B., Bagjadab, A.B. and Mishra, B.K. (2020) A Brief Overview of Natural Language Processing and Artificial Intelligence. In: Mishra, B.K. and Kumar, R., Eds., *Natural Language Processing in Artificial Intelligence*, AAP Inc., Palm Bay, 211-224. https://doi.org/10.1201/9780367808495-8

[43] Editorial Team (2019) A Quick Guide to Natural Language Processing (NLP), AI and Intelligent Automation. https://www.intelligentautomation.network/learning-ml/articles/a-basic-guide-to-natural-language-processing-nlp

[44] Taulli, T. (2019) Natural Language Processing (NLP): How Computers Talk. In: Taulli, T., Ed., *Artificial Intelligence Basics*: *A Non-Technical Introduction*, Apress, Monrovia, 103-124. https://doi.org/10.1007/978-1-4842-5028-0_6

[45] Koshorek, O., Cohen, A., Mor, N., Rotman, M. and Berant, J. (2018) Text Segmentation as a Supervised Learning Task. *Proceedigns of the* 2018 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Vol. 2, 469-473. https://doi.org/10.18653/v1/N18-2075

[46] Chakravarthy, S. (2020) Tokenization for Natural Language Processing. Towards Data Science. https://towardsdatascience.com/tokenization-for-natural-language-processing-a179a891bad4

[47] Jivani, A.G. (2011) A Comparative Study of Stemming Algorithms. *International*

Journal of Circuit Theory and Applications, **2**, 1930-1938.

[48] Kübler, S., McDonald, R. and Nivre, J. (2009) Dependency Parsing. Synthesis Lectures on Human Language Technologies, Vol. 2, Springer, Berlin, 1-19. https://doi.org/10.2200/S00169ED1V01Y200901HLT002

[49] Thanaki, J. (2017) Feature Engineering and NLP Algorithms. In: *Python Natural Language Processing: Explore NLP with Machine Learning and Deep Learning Techniques*, Packt Publishing, Birmingham, 102-148.

[50] Marshall, C. (2020) What Is Named Entity Recognition (NER) and How Can I Use It? Super.AI-AI & Human Data Labeling, AI Model Training & Deployment, 2019. https://medium.com/mysuperai/what-is-named-entity-recognition-ner-and-how-can-i-use-it-2b68cf6f545d

[51] Dwivedi, P. (2018) NLP: Extracting the Main Topics from Your Dataset Using LDA in Minutes. Towards Data Science. https://towardsdatascience.com/nlp-extracting-the-main-topics-from-your-dataset-using-lda-in-minutes-21486f5aa925

[52] Rosa, K.D. and Ellen, J. (2009) Text Classification Methodologies Applied to Micro-Text in Military Chat. *Proceedings International Conference on Machine Learning and Applications*, Miami, 13-15 December 2009, 710-714. https://doi.org/10.1109/ICMLA.2009.49

[53] Balabantaray, R.C., Mohammad, M. and Sharma, N. (2012) Multi-Class Twitter Emotion Classification: A New Approach. *International Journal of Accounting Information Systems*, **4**, 48-53. https://doi.org/10.5120/ijais12-450651

[54] Hayashida, Y., Uetsuji, T., Ebara, Y. and Koyamada, K. (2017) Category Classification of Text Data with Machine Learning Technique for Visualizing Flow of Conversation in Counseling. 2017 *Nicograph International* (*NicoInt*), Kyoto, 2-3 June 2017, 37-40. https://doi.org/10.1109/NICOInt.2017.35

[55] Poulin, C., Shiner, B., Thompson, P., Vepstas, L., Young-Xu, Y., Goertzel, B., Watts, B., Flashman, L. and McAllister, T. (2014) Predicting the Risk of Suicide by Analyzing the Text of Clinical Notes. *PLOS ONE*, **9**, e0085733. https://doi.org/10.1371/journal.pone.0085733

[56] Koopman, B., Karimi, S., Nguyen, A., McGuire, R., Muscatello, D., Kemp, M., Truran, D., Zhang, M. and Thackway, S. (2015) Automatic Classification of Diseases from Free-Text Death Certificates for Real-Time Surveillance. *BMC Medical Informatics and Decision Making*, **15**, Article No. 53. https://doi.org/10.1186/s12911-015-0174-2

[57] Lovins, J.B. (1968) A Comparative Study of Stemming Algorithms for Information Retrieval. *ACM Computing Surveys*, **4**, 61-73.

[58] International Labour Organization (1930, June 30) Convention Concerning the Reduction of Hours of Work to Forty per Week. https://www.ilo.org/dyn/normlex/en/f?p=NORMLEXPUB:12100:0::NO::P12100_INSTRUMENT_ID:312175