Developing Datasets from Limited or Scarce Data Sources

Tawanda Madzidzise[1], Monika Gondo[2]

M Tech in Software Engineering Student, Tawanda Madzidzise
Information Science and Technology, Harare Institute of Technology, Harare, Zimbabwe
Email: *awandamadzidzise[at]gmail.com*

M Tech Coordinator Software Engineering, Data Science and Analytics
Information Science and Technology, Harare Institute of Technology, Harare, Zimbabwe
Email: *mgondo[at]hit.ac.zw*

Abstract: ***Artificial intelligence(AI) is one of the best avenues of business success since there is a lot that the approach can achieve from data analysis, projections, insights and many more, this approach is likely to boost the world economy by $15.7 trillion by 2030, ("AI to Boost World Economy by over 15 Trillion Dollars in Seven Years" n.d.).In a scenario where one identifies a gap or a problem, and will need to embark on the solution in Artificial Intelligence, the very question will there be a dataset(s) that will be used for developing, training and testing the models or solution. This is one of the major limitations in developing solutions in the health sector in Zimbabwe mainly in the subsection of mental health where there is limited datasets which can be used in coming up with models to cater for the problems in mental health which include early detection of depression, anxiety and other problems which are in mental health.This should not be a major issue since we can have solutions which can provide solutions for scenarios like these, the research will identify possible drawbacks in developing models in scenarios of limited/scarce data targeting the case of mental health in Zimbabwe. There is a chance to develop model that speak to Mental Health in Zimbabwe, thus through finding ways to do so in scarce data environment. This will be was the major focus of the research and it was found to be achievable, then the researcher went on to develop a dataset with a limited number of data which was then used to train a model through different algorithms and it proved to be feasible at last. The result was great but there were challenges which were faced and some proposed solutions listed and further research proposals so as to attain better results and improve on the issue of bias.***

Keywords: model, AI, Mental health, data

1. Definition of terms

Data scarcity is defined or referred to scenarios in which there is limited or lack of data which is needed to build, train and test a model, (Joshi 2021).

Artificial Intelligence is referred to as a field, which combines computer science and robust datasets, to enable problem-solving.

Mental Health Gap Action Programme(mhGAP) aims at scaling up services for mental, neurological and substance use disorders for countries especially with low and middle income.

2. Background

The Ministry of Health And Childcare of Zimbabwe is overseeing the mental health initiatives in the country with the reports estimating a population prevalence of 0.1 % for schizophrenia, 0.5 % for bipolar disorder, 1.5 % for major depressive disorder (MDD), 0.3 % for alcohol use disorders, and 0.7 % for drug use disorders, suicide accounts for 1,8 % of all deaths, with this statistics we can identify a gap in the country, with 16 psychiatrists and 9 public on a population of 15.99 million by 2021 narrates to a major challenge, thus mental health in Zimbabwe is out of reach, ("Zim Only Has 16 Psychiatrists – Zimbabwe Situation" n.d.). Solutions have been proposed by the ministry and other stakeholders including the partnering with Non-Government Organisation to cover the gap, the mhGAP was one of the major initiatives in trying to capacitate the nurses to fill the gap("mhGAP Intervention Guide - Version 2.0" n.d.). These solutions are making a difference but the question is how many people are able to access the facility or carders?, another solution is which has been put forward in other countries mostly those first world countries is the use of artificial intelligence to detect, and provide solution in mental health through 3 branches which include Machine Learning and Deep learning, Computer Vision, Natural language processing, though the use of Computer-Aided Behavioural Therapy (CBT), the outcomes were as good as the standard CBT on depression, panic and phobias in 2006, ("AI in Mental Health - Examples, Benefits & Trends" 2022), this was successful in England, but due to other barriers the same approaches which were successful in other countries can have challenges in countries like Zimbabwe. In Zimbabwe there is scarce data source which can be utilised for model development, the Ministry of Health and ChildCare through the DHIS2 platform collects summarised data rather than client level data, and they are the custodian of the major source of data which can be used for Health Data, this will mean a major blow in coming up with the model that will be a solution to some mental health issues in Zimbabwe, but this must not limit initiatives or solutions to be developed, the solution to develop models in scarce data scenarios should be mechanised and custom.

3. Problem statement

How to develop datasets from limited data projects ?

Objectives

a) To identify methods which have been used in developing models in limited or scarce data scenarios
b) To identify problems associated with developing models in limited or scarce data.
c) To define strategies to develop datasets for models from these scarce data sources.

4. Literature Review

Scarce, Limited or incomplete data is a form of data imperfection that affects the ability of algorithms, machine learned models or human engineers to extract and induce knowledge, (Holst and Lohweg 2022), this is so because the modern methods for data analytics not only assume big data but also they require it for efficiency and effectiveness.(Wang et al. 2021) mentioned that there are two ways that data can be scarce, thus one with small samples and one with sparsely and irregularly observed time series covariates.The two have almost the same effects thus, (Duca 2022):

- Outliers – with the cases of mental health were we have sensitive data and the data is not always available, there are chances that at one point there are some areas that one can get the data from more than other areas, this will define the structure of the data that the model will learn, for example we can have more data in depression and anxiety more than other areas, this might mean that the results from the model will reflect much of these areas than others.
- Missing parameters – there are different models which are being used to help people with mental health issues, this might mean that some parameters which are common will be easy to collect than those which are unique leading to some model specific parameters being left out.

The area of Artificial Intelligence is broad, so as to formulate focus, the researcher came up with a clustering approach to organise, ("Cluster Analysis in Data Mining: Applications, Methods & Requirements | upGrad Blog" n.d.), to findings patterns of data which are in line with the objectives, thus the research was centred in finding data linked to the objectives. To do this 10 journals were taken into account, the research was centred on the following list:

- Causes, Characteristics and Implication of scarce data
- Time series data
- Pre-training framework to support various kinds of conversation
- Developing a Conversational Question Answering system
- Comparison of federated and classical machine learning.
- Discussion on different approaches of Machine Learning in line with Mental Health
- Discussion on Machine learning and deep learning approaches in mental health diagnosis
- An insight on the concept and applications of federated machine learning
- Communication-efficient learning of deep networks from decentralized data
- Discussion on the Federated Learning based on machine learning and deep learning

In addition to the above mentioned sources, the researcher also used published pieces of work on the world wide web, to search for relevant information the following searches were done:

- Problems related to machine learning
- Problems related to deep learning
- Scarcity in data for model development
- Artificial Intelligence in Mental Health
- Federated learning
- Data fusion

5. Findings

- The research reflected results from the study of mental health problems, according to ("Mental Disorders" n.d.)"*1 in every 8 people in the world live with a mental disorder*"**,** which include schizophrenia, anxiety and depression, bipolar disorder, post-traumatic Stress Disorder(PSTD) and Mental Health amongchildren.
- Machine learning approaches were utilised to predict the prevalence of these mental health disorders, the approaches include supervised learning, unsupervised learning, ensemble learning , neural networks and deep learning in most of the areas where the initiative was tried
- Most of the areas did not have proper or valuable data since there is less data patterning to Mental Health on the world so the same will apply to the scenario of Zimbabwe.
- The algorithms will work differently in accordance to the area it is applied, thus one algorithm might work well in predicting one mental disorder and will have adverse results on the other mental health issue.
- Small data can be used develop a dataset if you are willing to incur the adaptive learning techniques such as the federated learning, ("Federated Learning Explained" 2023), which is a decentralised approach to machine learning, this will allow the iterative upgrade of the dataset since new insights can be acquired from the new trainings then an updated is sent to the central repository.
- The dataset can be updated when the model has been developed from one with limited coverage.

The table below shows the actual findings from predicting the outcomes on five different of mental disorders

Table 1: Best model for a mental health scenario

Mental health disorder	Algorithm with the best prediction
Anxiety and depression	Bayesian network
Bipolar disorder	Gaussian process classification
Post-traumatic stress disorder	Random forest
Mental health problems among children	Multilayer perceptron

6. Research review/ Results

The results from the study gave green light to the researcher that there is a chance to develop a mental health model, with the motive that the researcher have access to data which is in line with the major area of study. The researcher extended his curiosity into trying the initiative of the dataset for a mental health dataset which will focus on mimicking the Friendship Bench counselling strategy, the Problem Solving Therapy, ("PROBLEM SOLVING THERAPY | Friendshipbench" n.d.), which is a counselling model for non-clinical conditions with sessions that can be done in the community, clinic or online.

To determine the feasibility of coming up with a dataset in data scarce environment, the researcher had to do the following:

- Define the problem statement which was to be addressed by one coming up with a dataset, in this case was having a dataset that can be used to train a model that will mimic a real delivering agent.
- Determine the structure of the dataset which was to be developed
- Define the sample framework and sample that has to be chosen and the characteristics of the data to be used.
- Train and test a model using this model and check if this model can be used for prediction in the session.

a) Structure of the dataset

The initial step was to come up with a template of what the dataset would look like, this was through the use of available data and the PST model steps to define what should be available in the dataset and how. In addition to the steps the researcher was able to collect sample data which is explained below and do data analysis on the data using Nvivo since it is qualitative data so as to come up with other features of the dataset. The following features were included in the template as a result from the analysis and using PST guidelines. A .csv file with the following columns was developed:

- Session code – reference to the chat number
- Session number – client visit number
- PST quality – defined by the ability of the counselor to utilise the PST steps well
- Utterance id – row number of the utterance text in .csv file
- Interlocutor- the person who communicated that time, thus either the agent or client
- Topic – in which category under the PST-model does the utterance text fall
- Features – within the categories, what exactly is that which was communicated
- Question type- if it is a question then it will be either open or closed else Not applicable (N/A)
- Reflection exists – what was the reaction from the interlocutor, is it a question, response or update.
- Question exists – question exists within the PST-Model script thus yes, no else N/A
- Agent input exists – either the input was from the Agent or not
- Client response – either the input was from the Client or not
- Priority – this will reflect on the impact of the input
- Utterance text – this was the actual input from the interlocutor

b) Sampling Framework

The targeted population, which is the group which was intended to be included in the study, were those who had done their sessions on line using WhatsApp, this allows the research to have the actual features of a real session and they were done in English also. In total there were more than 100 samples but not all did qualify to be part of the study because of some reasons which include, lack of consent, mixed languages, the session was not completed and many others. The WhatsApp conversation were exported as .txt files to google drive into one folder for easy manipulation using pandas and access.

c) Sample selected for dataset

The above mentioned scenarios were taken into consideration and also with the intention to check the feasibility of the research, a sample of 20 chats were considered after random selection and considering the following:

- Completed at least one session
- Ssq14 was administered and PST preceding steps done
- Depression or anxiety issues were part of discussion
- Psychoeducation was part of the session
- Could have a red flag or above cut off point

d) Tools

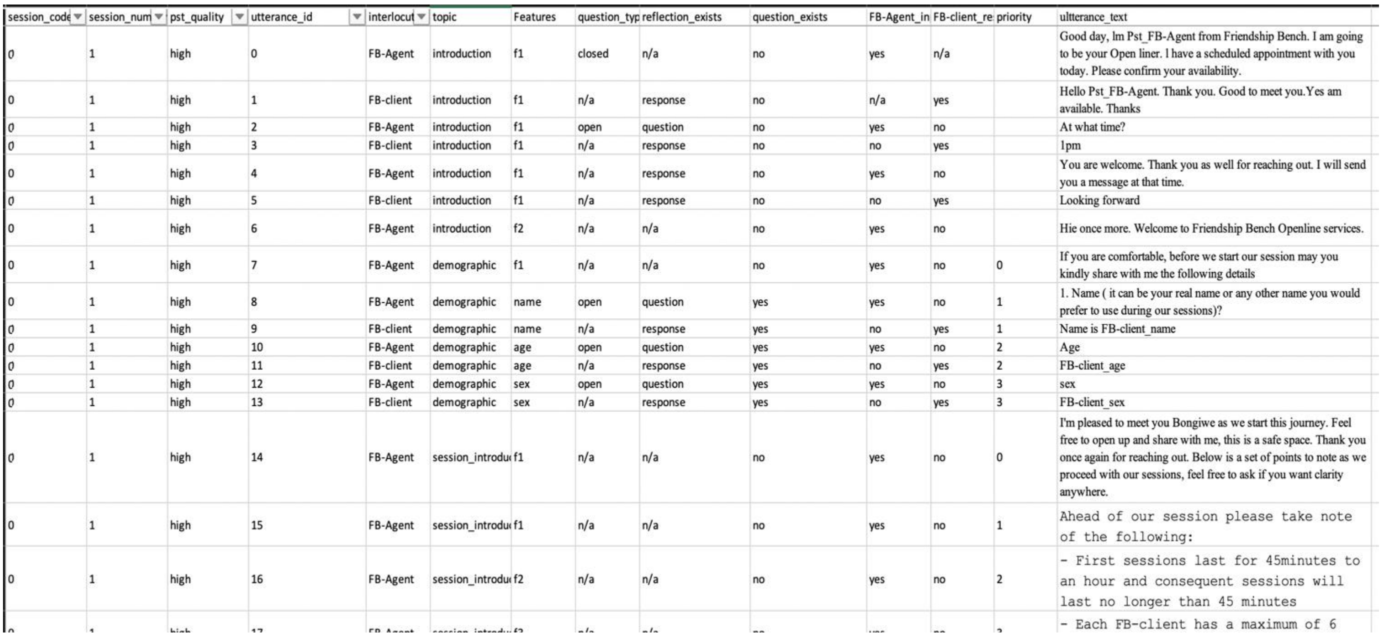
The tools were categorised into two, thus the sampling tools and data manipulation tools, thus under sampling tools we had the following:

- Nvivo for qualitative data analysis
- Raosoft for sample size
- Excel for random sampling.

For dataset development, the researcher developed a program using python in Google Colab (https://colab.research.google.com/drive/1I8Dpq9Zh2r11AzhySWNETAeBtzqcP5j1#scrollTo=qEcj1gtBJ6We&uniqifier=3), the following packages were uses:

- Regex – for pattern matching
- NumPy – array manipulation
- Pandas – data list manipulation
- Openpyxl – read and write to excel files
- Json – data formatting

After this process the dataset with the structure below was successfully generated:



session_code	session_num	pst_quality	utterance_id	interlocut	topic	Features	question_typ	reflection_exists	question_exists	FB-Agent_in	FB-client_re	priority	ultterance_text
0	1	high	0	FB-Agent	introduction	f1	closed	n/a	no	yes	n/a		Good day, Im Pst_FB-Agent from Friendship Bench. I am going to be your Open liner. l have a scheduled appointment with you today. Please confirm your availability.
0	1	high	1	FB-client	introduction	f1	n/a	response	no	n/a	yes		Hello Pst_FB-Agent. Thank you. Good to meet you.Yes am available. Thanks
0	1	high	2	FB-Agent	introduction	f1	open	question	no	yes	no		At what time?
0	1	high	3	FB-client	introduction	f1	n/a	response	no	no	yes		1pm
0	1	high	4	FB-Agent	introduction	f1	n/a	response	no	yes	no		You are welcome. Thank you as well for reaching out. I will send you a message at that time.
0	1	high	5	FB-client	introduction	f1	n/a	response	no	no	yes		Looking forward
0	1	high	6	FB-Agent	introduction	f2	n/a	n/a	no	yes	no		Hie once more. Welcome to Friendship Bench Openline services.
0	1	high	7	FB-Agent	demographic	f1	n/a	n/a	no	yes	no	0	If you are comfortable, before we start our session may you kindly share with me the following details
0	1	high	8	FB-Agent	demographic	name	open	question	yes	yes	no	1	1. Name (it can be your real name or any other name you would prefer to use during our sessions)?
0	1	high	9	FB-client	demographic	name	n/a	response	yes	no	yes	1	Name is FB-client_name
0	1	high	10	FB-Agent	demographic	age	open	question	yes	yes	no	2	Age
0	1	high	11	FB-client	demographic	age	n/a	response	yes	no	yes	2	FB-client_age
0	1	high	12	FB-Agent	demographic	sex	open	question	yes	yes	no	3	sex
0	1	high	13	FB-client	demographic	sex	n/a	response	yes	no	yes	3	FB-client_sex
0	1	high	14	FB-Agent	session_introdu	f1	n/a	n/a	no	yes	no	0	I'm pleased to meet you Bongiwe as we start this journey. Feel free to open up and share with me, this is a safe space. Thank you once again for reaching out. Below is a set of points to note as we proceed with our sessions, feel free to ask if you want clarity anywhere.
0	1	high	15	FB-Agent	session_introdu	f1	n/a	n/a	no	yes	no	1	Ahead of our session please take note of the following:
0	1	high	16	FB-Agent	session_introdu	f2	n/a	n/a	no	yes	no	2	- First sessions last for 45minutes to an hour and consequent sessions will last no longer than 45 minutes
0	1	high	17	FB-Agent	session_introdu	[illegible]	n/a	n/a	no	yes	no	3	- Each FB-client has a maximum of 6

This file was developed in .csv format and the data reflecting Personal Identification Information (PII) were replaced as follows:

- Agent name : FB-Agent
- Client name : FB-Client
- Age : FB-client_age
- Gender : FB-client_sex

e) Model training

The PST dataset was used to train a model using the following packages using Google Colab (https://colab.research.google.com/drive/1cgbYFcZlLKpmPpzOcBmwJj4om4s2fl0K#scrollTo=GLciTJ9SIGyM):

- Scikit – learn
- Natural language toolkit
- Regex
- Pandas
- NumPy

Algorithm	Precision	Recall	f1-score	support	accuracy
Naïve bayes	0.35	0.48	0.46	5545	0.59
support vector machines	0.42	0.58	0.51	5545	0.65
Nearest Neighbours	0.37	0.51	0.32	5545	0.48
Decision trees	0.45	0.23	0.4	5545	0.45
random Forests	0.54	0.76	0.6	5545	0.76
Average score	0.426	0.512	0.458	5545	0.586

The following results were achieved,

The above results have shown the feasibility of developing a dataset from the scarce data environment, this have made it possible to train a model from these datasets. Comparing the results which were attained, we could see that the random forest can be used to predict with an accuracy of F1-score : 0.6 and accuracy: 0.76, this result is good for a start but there in need to do a research on the best practices so as to achieve better results. Averages across the score categories were as follows:

Precision : 0.426
Recall : 0.512
F1-score : 0.458
Accuracy : 0.586

f) Achievements

The following achievements were achieved:

- A dataset was developed
- A model was trained with using 5 algorithms with better average results

g) Challenges

The following challenges were faced

- Lack of resources with a large computing power for better data manipulation, and human capital for better data wrangling
- The data was only focusing on one area of mental health thus depression and anxiety
- Reflections were not clear in most of the conversations.

h) Further research

The outcome from this research was a mini- dataset for mental health focusing mainly on depression and anxiety, for further study, the researcher will look more into onboarding adaptive training and federated training so as to develop a full scope dataset which will update the current impact.

7. Conclusion

The development of dataset from scarce data, where we have limited data can be possible if there is a leeway for one to adjust the dataset through other means which might include having other contributions after the initial dataset, and also having the dataset to be updated after finding other insights or relevant variables. The initiative will need more resources in terms of computing power and human capital among other requirement. The mini-dataset for depression and anxiety which was developed can be used for model building but there is need for the researcher to have additions to the dataset and training of the model so as to come up with a good results of an average of 85% accuracy and an F1-score around 75%. If these high level results are achieved then the next research will be on finding the best approaches to develop a delivering agent which is capable of conducting a session with at least 2 languages in one conversation.

References

[1] "AI in Mental Health - Examples, Benefits & Trends." 2022. *ITRex* (blog). December 13, 2022. https://itrexgroup.com/blog/ai-mental-health-examples-trends/.

[2] "AI to Boost World Economy by over 15 Trillion Dollars in Seven Years." n.d. The Jerusalem Post | JPost.Com. Accessed June 7, 2023. https://www.jpost.com/business-and-innovation/all-news/article-738795.

[3] Asad, Muhammad, Ahmed Moustafa, and Takayuki Ito. 2021. "Federated Learning Versus Classical Machine Learning: A Convergence Comparison." arXiv. http://arxiv.org/abs/2107.10976.

[4] Bao, Siqi, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. "PLATO: Pre-Trained Dialogue Generation Model with Discrete Latent Variable." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 85–96. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.9.

[5] Chung, Jetli, and Jason Teo. 2022. "Mental Health Prediction Using Machine Learning: Taxonomy, Applications, and Challenges." Edited by Aniello Minutolo. *Applied Computational Intelligence and Soft Computing* 2022 (January): 1–19. https://doi.org/10.1155/2022/9970363.

[6] "Cluster Analysis in Data Mining: Applications, Methods & Requirements | upGrad Blog." n.d. Accessed May 31, 2023. https://www.upgrad.com/blog/cluster-analysis-data-mining/.

[7] Duca, Angelica Lo. 2022. "Is a Small Dataset Risky?" Medium. February 19, 2022. https://towardsdatascience.com/is-a-small-dataset-risky-b664b8569a21.

[8] "Federated Learning Explained." 2023. AltexSoft. June 15, 2023. https://www.altexsoft.com/blog/federated-learning/.

[9] "Federated-Learning.Pdf." n.d. Accessed June 7, 2023. https://inst.eecs.berkeley.edu/~cs294-163/fa19/slides/federated-learning.pdf.

[10] Holst, Christoph-Alexander, and Volker Lohweg. 2022. "Scarce Data in Intelligent Technical Systems: Causes, Characteristics, and Implications." *Sci* 4 (4): 49. https://doi.org/10.3390/sci4040049.

[11] Hu, Kai, Yaogen Li, Min Xia, Jiasheng Wu, Meixia Lu, Shuai Zhang, and Liguo Weng. 2021. "Federated Learning: A Distributed Shared Machine Learning Method." *Complexity* 2021 (August): 1–20. https://doi.org/10.1155/2021/8261663.

[12] Iyortsuun, Ngumimi Karen, Soo-Hyung Kim, Min Jhon, Hyung-Jeong Yang, and Sudarshan Pant. 2023. "A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis." *Healthcare* 11 (3): 285. https://doi.org/10.3390/healthcare11030285.

[13] Joshi, Preetam. 2021. "Handling Data Scarcity While Building Machine Learning Applications." Medium. March 29, 2021. https://towardsdatascience.com/handling-data-scarcity-while-building-machine-learning-applications-e6c243b284b0.

[14] "Mental Disorders." n.d. Accessed June 7, 2023. https://www.who.int/news-room/fact-sheets/detail/mental-disorders.

[15] "mhGAP Intervention Guide - Version 2.0." n.d. Accessed June 7, 2023. https://www.who.int/publications-detail-redirect/9789241549790.

[16] "PROBLEM SOLVING THERAPY | Friendshipbench." n.d. Accessed October 27, 2023. https://www.friendshipbenchzimbabwe.org/problemsolvingtherapy.

[17] Reddy, Siva, Danqi Chen, and Christopher D. Manning. 2019. "CoQA: A Conversational Question Answering Challenge." *Transactions of the Association for Computational Linguistics* 7: 249–66. https://doi.org/10.1162/tacl_a_00266.

[18] Wang, Qiyao, Ahmed Farahat, Chetan Gupta, and Shuai Zheng. 2021. "Deep Time Series Models for Scarce Data." *Neurocomputing* 456 (October): 504–18. https://doi.org/10.1016/j.neucom.2020.12.132.

[19] Yang, Qiang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. "Federated Machine Learning: Concept and Applications." arXiv. http://arxiv.org/abs/1902.04885.

[20] "Zim Only Has 16 Psychiatrists – Zimbabwe Situation." n.d. Accessed June 7, 2023. https://www.zimbabwesituation.com/news/zim-only-has-16-psychiatrists/.