Data in Brief 60 (2025) 111475

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Swahili questions and answers dataset for aflatoxin knowledge domain

Pamela Chogo*, Elizabeth Mkoba, Neema Kassim

Nelson Mandela African Institution of Science and Technology, P.O.Box 447, Tengeru, Arusha, Tanzania

ARTICLE INFO

Article history:
Received 4 September 2024
Revised 11 March 2025
Accepted 12 March 2025
Available online 20 March 2025

Dataset link: Aflatoxin Question and Answer Swahili Dataset (AflaQnASD) (Original data)

Keywords:
Natural language processing
NLP-based chatbot
Knowledge sharing
Food security
Aflatoxin dataset

ABSTRACT

Aflatoxin contamination is a challenge facing food security, health, and trade in Tanzania and other parts of the world. This contamination affects maize, groundnuts, and other crops and animal products. Once contamination occurs, the contaminated crops and animal products become toxic causing illness or death to humans and animals who consume them. Lack of awareness and knowledge of the contamination is seen to be one of the reasons for its continued occurrence. Various awareness-creation and knowledge-sharing techniques have been used but the situation is still not appealing. For this case, the use of a Natural Language Processing (NLP) chatbot in sharing aflatoxin knowledge is proposed. This is because NLP chatbots have been successful in knowledge sharing in various contexts. This data article presents a Swahili text-based aflatoxin knowledge questions and answers dataset. Data were collected through 7 focus group discussion (FGD) sessions conducted in Arusha, Dodoma, Mtwara, Tabora, Morogoro, and Iringa regions in Tanzania. Respondents for the study were farmers, traders, and consumers of maize and groundnuts. The collected data were processed and analyzed using R qualitative data analysis tool. This allowed the identification of 6 themes with respective questions under each theme. The questions were shared with experts through 9 interview sessions and the experts gave answers to the questions. The set of questions and answers were then translated into Swahili language using google translate and manual verification. Finally, an aflatoxin knowledge dataset containing 221 paired questions and an-

* Corresponding author.
E-mail address: chogop@nm-aist.ac.tz (P. Chogo).

https://doi.org/10.1016/j.dib.2025.111475

swers organized into 6 knowledge areas Swahili dataset was developed. With this dataset, an NLP-based chatbot that uses Swahili language can be developed. This will be beneficial to farmers, traders, consumers, researchers, and policymakers. They can use it to learn more about aflatoxin and be able to make informed decisions. Moreover, the dataset can be adopted and modified to create NLP chatbots that can share aflatoxin knowledge in other languages apart from Swahili. The dataset also contributes to the availability of Swahili language datasets.

Specifications Table

Subject	Artificial Intelligence
Specific subject area	Swahili questions and answers dataset for aflatoxin knowledge for the maize and groundnut value chain to be used in developing a Natural Language Processing-based chatbot
Type of data	Text
Data collection	Data were collected through 7 focused group discussion sessions conducted in 6 regions of Tanzania. The respondents were randomly selected from a list of farmers, traders, and consumers provided by community leaders. Data from the FGD were transcribed, translated, coded, and analyzed into themes. This provided a list of aflatoxin questions answered through 9 interview sessions with experts and researchers. The questions and answers were combined and translated into the Swahili language, creating a Swahili aflatoxin knowledge dataset [1].
Data source location	The questions were collected through FGD sessions conducted in 6 regions in Tanzania. The regions are Arusha, Dodoma, Tabora, Mtwara, Iringa and Morogoro. The regions were selected based on their popularity in Maize and Groundnut farming. Furthermore, the selected regions were randomly selected from the 7 agriculture zones in Tanzania [2] and purposive sampling was used to select specific data collection sites from the regions. Furthermore, answers to the questions were collected from experts from the International Institute of Tropical Agriculture (IITA), the Tanzania Bureau of Standards (TBS), the Economic and Social Research Foundation (ESRF), the Tanzania Initiative for Preventing Aflatoxin Contamination (TANIPAC), and the Nelson Mandela African Institution of Science and Technology (NM-AIST)
Data accessibility	Repository name: Mendeley Data Data identification number: DOI:10.17632/zhpxzpnjfs.1 Direct URL to data: https://data.mendeley.com/datasets/zhpxzpnjfs/1
Related research article	P. Chogo, N. Kassim, E. Mkoba, Identification of Aflatoxin Knowledge Areas for Developing Natural Language Processing Chatbot Datasets. 2023 First International Conference on the Advancements of Artificial Intelligence in African Context (AAIAC). IEEE Xplore (2024), https://doi.org/10.1109/AAIAC60008.2023.10465309

1. Value of the Data

- Domain-specific data: The developed dataset provides an accurate aflatoxin knowledge questions and answers dataset useful for creating a domain-specific chatbot using Natural Language Processing.
- Improving agricultural practices: The dataset can help raise aflatoxin contamination awareness by creating advisory tools for best agricultural practices for preventing aflatoxin contamination to be integrated into other agriculture platforms.
- Unique dataset: To the best of our knowledge this is the only dataset specifically for aflatoxin contamination questions and answers.

- Resource for low-resource language: The dataset is useful for training chatbot models in Swahili language. Currently, the availability of Swahili language datasets is limited.
- Dataset adaptability: The dataset can be adopted and modified to create NLP chatbots that can share aflatoxin knowledge in other languages apart from Swahili and English. The dataset can also be used in Swahili translation models related to aflatoxin.
- Benefit to researchers: The dataset will be beneficial to researchers in low-resource languages such as Swahili.

2. Background

Aflatoxin contamination of maize, groundnuts, and other crops has had significant effects on public health, trade, and food security [3]. This dataset provides aflatoxin knowledge in the form of questions and answers. Preparations of the dataset adopted methodologies used in preparing other chatbot datasets in general agriculture. Whereby, stakeholders are involved to ensure relevant questions are obtained from the users and accurate responses are gathered from experts [4,5]. Primary data collection must be conducted with domain-specific datasets such as the aflatoxin knowledge dataset [6]. With its uniqueness, this dataset contributes to the availability of domain-specific datasets which are usually limited [7]. Furthermore, the dataset contributes to the availability of Swahili language datasets as the Swahili language is one of the African languages that are disadvantaged and classified as a low-resource language because of inadequate data for NLP [8,9]. For this case, a multilingual (Swahili and English) NLP-based chatbot for aflatoxin knowledge will be developed using this dataset. This will enable farmers, traders, consumers, policymakers, and researchers to learn about it and adopt best practices that will enhance contamination reduction hence minimizing the associated risks.

3. Data Description

The published dataset provides questions and answers on aflatoxin contamination. This dataset was published in an open access repository on the 11th July, 2024 [10]. A total of 9 files were stored in the Mendeley data repository, these files included the dataset files and other related documents stored in a folder named AflaQnASD and other sub folders named Datasets and other documents as seen on Fig 1. Data files are stored in .xlsx format while research clearance letter, consent forms, FGD and interview guides are stored in PDF format. Furthermore, the first dataset file AflaQnAD-Swahili.xlsx contains 221 aflatoxin contamination-related paired questions and answers in the Swahili language. These questions and answers are classified into knowledge areas as shown in Table 1. The dataset is organized in four columns titled Sno, knowledge area, questions and answers as shown in Figs. 2 & 3. On the other hand, the second data file AflaQnAD-English.xlsx contains a translated set of the same questions and answers in English language.

Table 1
Number of questions on knowledge areas.

Sno	Knowledge areas	QnA Pairs
1	Meaning of aflatoxin	59
2	Effects of aflatoxin	33
3	Management of contaminated crops	5
4	Laws and regulations	10
5	Prevention	58
6	Awareness of aflatoxin	56
7	Total pairs	221

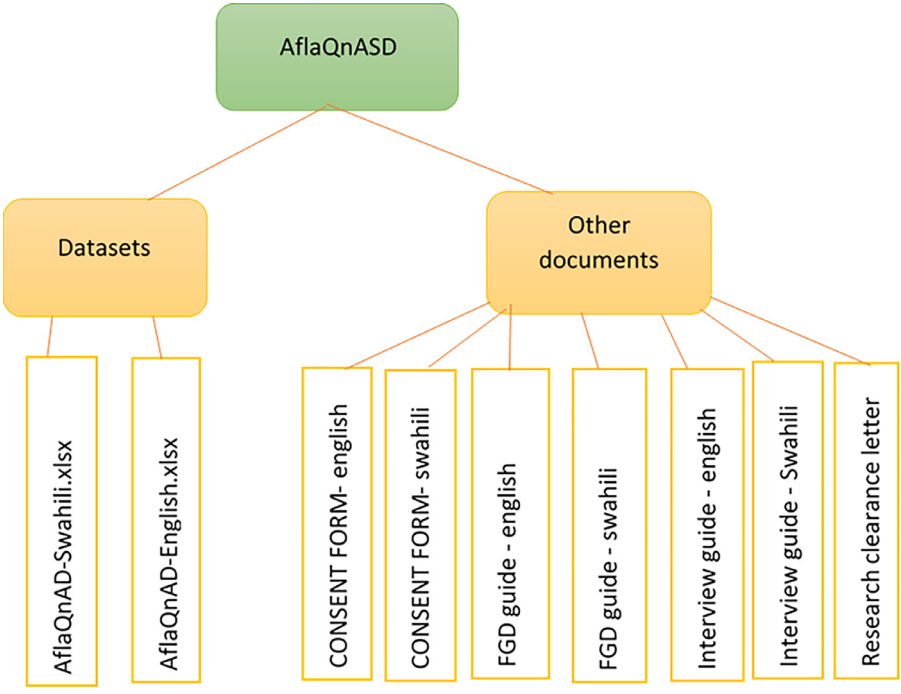
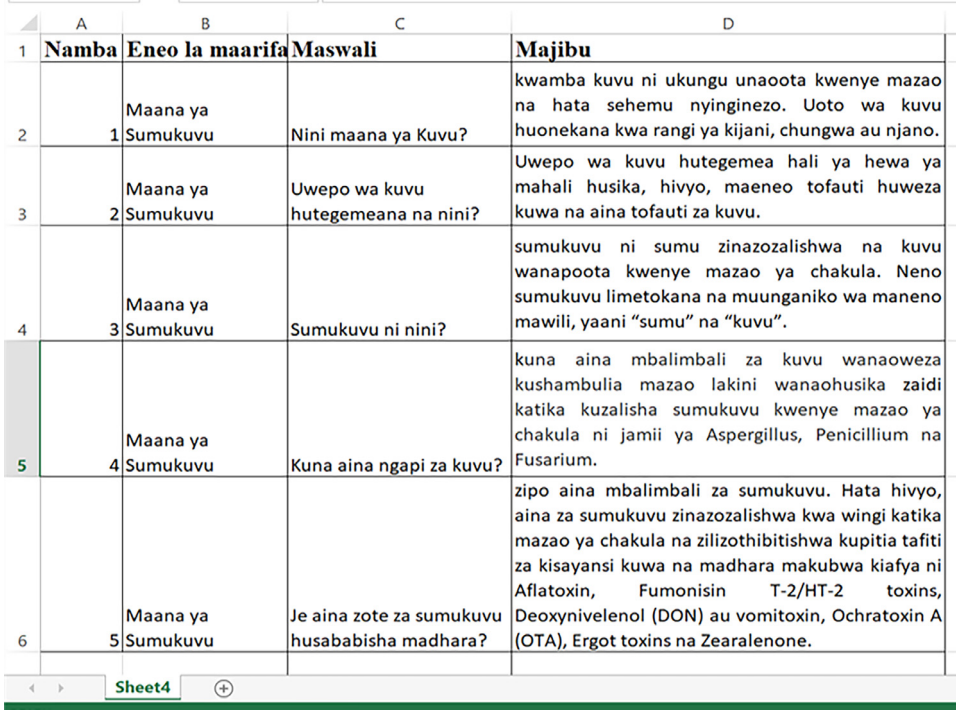


Fig. 1. Files and folders.

	A	B	C	D
1	Sno	Knowledge Area	Question	Answer
2	1	Meaning of Aflatoxin	What is fungus?	Fungus is a fog that grows on crops and even other places. Fungal growths appear green, orange or yellow
3	2	Meaning of Aflatoxin	What do you think determines the presence of fungus?	The presence of fungi depends on the climate of the place, therefore, different areas may have different types of fungi
4	3	Meaning of Aflatoxin	What is aflatoxin?	Aflatoxins are toxins produced by fungi when they grow on food crops. The word aflatoxin is derived from the combination of two words, namely "poison" and "fungus".
5	4	Meaning of Aflatoxin	How many types of fungi causing aflatoxin are there?	There are various types of fungi that can attack crops, but the ones most involved in producing aflatoxin are Aspergillus, Penicillium and Fusarium.
6	5	Meaning of Aflatoxin	Do all types of aflatoxin cause harm?	There are different types of aflatoxin. However, the types of toxins produced in large quantities in food crops and proven through scientific studies to have serious health effects are Aflatoxin, Fumonisin T-2/HT-2 toxins, Deoxynivelenol (DON) or vomitoxin, Ochratoxin A (OTA), Ergot toxins and Zearalenone.

Sheet4

Fig. 2. Sample English dataset rows and columns.



	A	B	C	D
1	Namba	Eneo la maarifa	Maswali	Majibu
2	1	Maana ya Sumukuvu	Nini maana ya Kuvu?	kwamba kuvu ni ukungu unaoota kwenye mazao na hata sehemu nyinginezo. Uoto wa kuvu huonekana kwa rangi ya kijani, chungwa au njano.
3	2	Maana ya Sumukuvu	Uwepo wa kuvu hutegemeana na nini?	Uwepo wa kuvu hutegemea hali ya hewa ya mahali husika, hivyo, maeneo tofauti huweza kuwa na aina tofauti za kuvu.
4	3	Maana ya Sumukuvu	Sumukuvu ni nini?	sumukuvu ni sumu zinazozalishwa na kuvu wanapoota kwenye mazao ya chakula. Neno sumukuvu limetokana na muunganiko wa maneno mawili, yaani "sumu" na "kuvu".
5	4	Maana ya Sumukuvu	Kuna aina ngapi za kuvu?	kuna aina mbalimbali za kuvu wanaoweza kushambulia mazao lakini wanaohusika zaidi katika kuzalisha sumukuvu kwenye mazao ya chakula ni jamii ya Aspergillus, Penicillium na Fusarium.
6	5	Maana ya Sumukuvu	Je aina zote za sumukuvu husababisha madhara?	zipo aina mbalimbali za sumukuvu. Hata hivyo, aina za sumukuvu zinazozalishwa kwa wingi katika mazao ya chakula na zilizothibitishwa kupitia tafiti za kisayansi kuwa na madhara makubwa kiafya ni Aflatoxin, Fumonisin T-2/HT-2 toxins, Deoxynivelenol (DON) au vomitoxin, Ochratoxin A (OTA), Ergot toxins na Zearalenone.

Sheet4

Fig. 3. Sample Swahili dataset rows and columns.

4. Experimental Design, Materials and Methods

4.1. Literature Search

The researcher started with a thorough literature review of 127 publications on aflatoxin contamination knowledge. The selection of the used publication was based on the inclusion criteria including publications available in full text, publications in English, and publications from 2017 to 2022. These publications came from the Science Direct database, Research 4 Life databases, Google Scholar, Education Resources Information Centre (ERIC), and the PubMed database. The thorough literature review enabled the identification of the data collection areas, population, data collection method, and data analysis method [11,12].

4.2. Data Collection Process

4.2.1. Regions Involved

Data were collected from the Arusha, Dodoma, Iringa, Morogoro, Tabora, and Mtwara regions, each representing one agriculture zone from the seven zones in Tanzania [2]. Furthermore, the selected areas had maize and/or groundnut farming since they were the main crops focused on the data collection. These regions are shown in Fig. 4. Purposive sampling was used to select specific data collection areas in the regions.

4.2.2. Population and Sampling

The population included maize and groundnut farmers, traders, and consumers from Arusha, Dodoma, Iringa, Morogoro, Tabora, and Mtwara regions. Out of these 70 were randomly selected

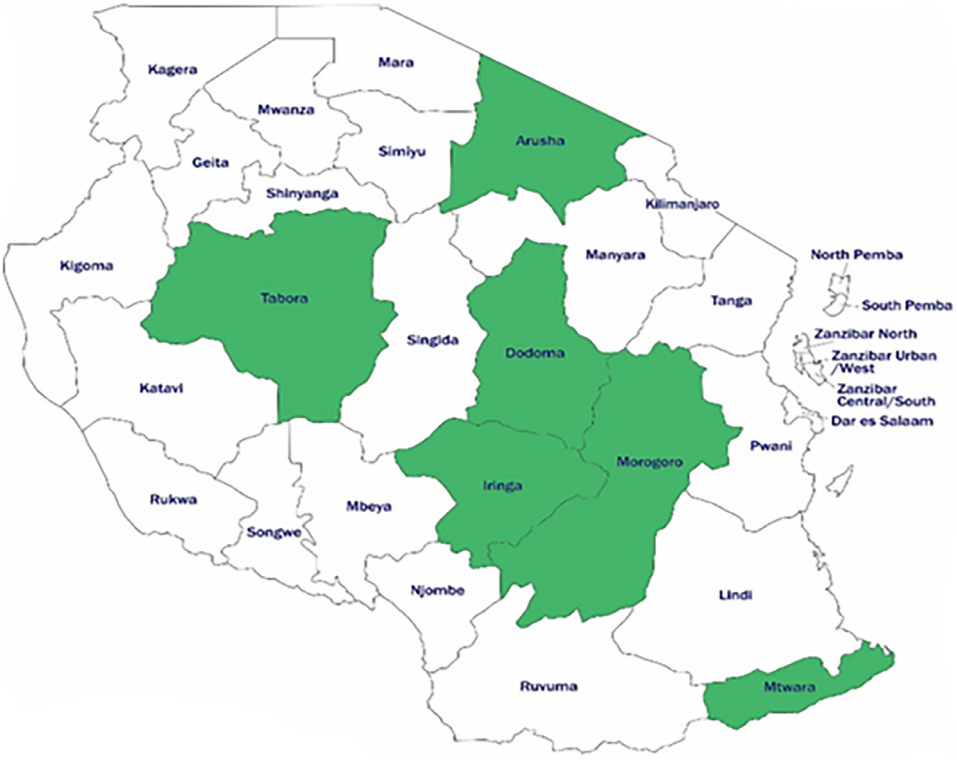


Fig. 4. Map of study areas.

to be included in the data collection process as respondents. The selection was based on the list of farmers, traders, and consumers provided by the community leaders from specific data collection areas in the regions.

4.2.3. Research Clearance

Research clearance was obtained from the Tanzanian President's Office – Regional Administration and Local Government Ministry. This enabled easy access to the study site with assistance from community leaders.

4.3. Data Collection Process

Focus group discussion (FDG) was the data collection method used. This was selected to allow an easy understanding of the phenomena through the interaction with the respondents. During its implementation, 70 selected respondents were contacted to be invited to the FGD sessions. Not all showed up, but all the sessions had between 3 to 10 respondents, making up 60 respondents [1]. The 7 FGD sessions were conducted with the researchers at the selected data collection sites using the simple pre-developed and tested FGD guide. The guide and the execution of the session were developed and conducted following FGD guide standards as stated in [13]. This included introductions to the data collection team, an explanation of the data collection process, and an explanation and signing of the consent form. Some of the sessions are as seen in Fig. 5.

Fig. 5. Focus group discussion and interview sessions.

The conversation was recorded through note-taking and audio recording by the research assistants. The language used was Swahili language to enable participants to understand and interact as the majority were not conversant with the English language.

4.4. Data Processing and Analysis

The Swahili data collected was transcribed using a summary transcription style that focuses on key points. Then the transcribed data were translated into the English language using Google Translate and then manual verification was done. The translated data were then coded by assigning labels to significant portions of text. Then thematic data analysis was done using the R Qualitative Data Analysis (RQDA) tool. From the analysis, six themes and sub-themes were identified. These were (1) Meaning of Aflatoxin (2) Effects of Aflatoxin (3) the Management of Contaminated crops (4) Laws and Regulations (5) Prevention and (6) Aflatoxin awareness creation [1]. All questions collected through the FGD sessions were organized into the identified themes. This brought the data collection process to the next stage as the questions needed answers.

4.5. Response to the Questions

An interview guide was developed and all the collected questions were given to 9 experts and researchers to give responses. These experts were from identified organizations dealing with aflatoxin and aflatoxin contamination issues. The organizations were the International Institute of Tropical Agriculture (IITA), the Tanzania Bureau of Standards (TBS), the Economic and Social Research Foundation (ESRF), the Tanzania Initiative for Preventing Aflatoxin Contamination (TANIPAC), and the Nelson Mandela African Institution of Science and Technology (NM-AIST) [1].

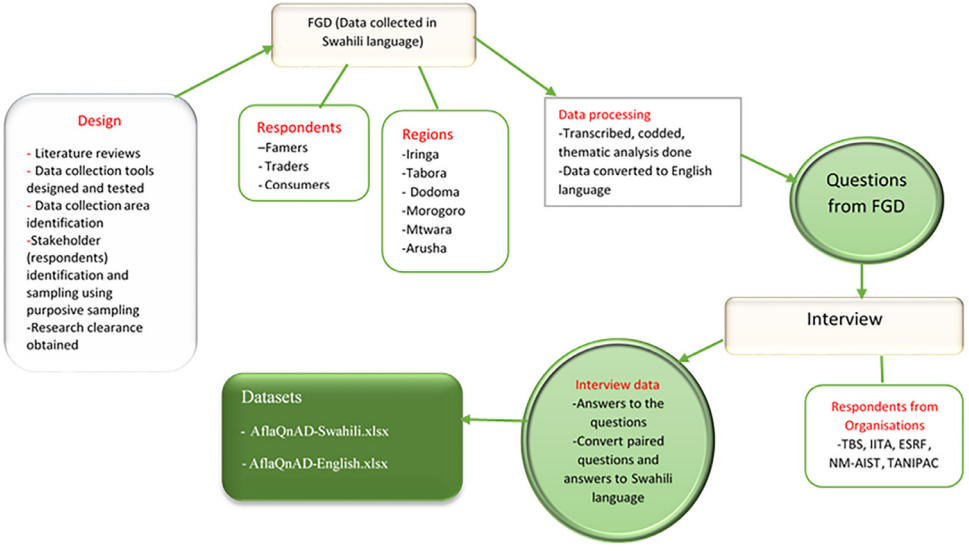


Fig. 6. Data collection flow diagram.

4.6. Dataset Development

Responses from all 9 interviewed experts were collected and recorded together. Then they were manually analyzed by comparing all answers to the same question, removing duplicates, and maintaining uniqueness from specific responses. This resulted in an English aflatoxin knowledge dataset of 221 paired questions and answers the number of pairs being similar to other available datasets [14]. This dataset was then translated to Swahili language using Google Translate and then manual verification was done to ensure that the semantic composition remained the same. Fig. 6 shows the data collection flow diagram.

Limitations

The dataset provided contains aflatoxin contamination knowledge limited to maize and groundnuts. The data was collected from a sample representing only 6 regions in Tanzania. The presented dataset is also limited in size. Future datasets can expand the contents to cover all crops affected by aflatoxin but also the sample size could be increased.

Ethics Statement

Participants consented by signing the consent form provided to them. Ethical clearance certificate with number KNCHREC00024/01/2024 was granted by KNCHREC and Research clearance was provided by the Tanzanian President's Office – Regional Administration and Local Government ministry.

Data Availability

Aflatoxin Question and Answer Swahili Dataset (AflaQnASD) (Original data) (Mendeley Data)

CRediT Author Statement

Pamela Chogo: Conceptualization, Methodology, Resources, Formal analysis, Data curation, Writing – original draft, Writing – review & editing; **Elizabeth Mkoba:** Conceptualization, Methodology, Data curation, Writing – review & editing, Supervision; **Neema Kassim:** Conceptualization, Methodology, Data curation, Writing – review & editing, Supervision, Validation.

Acknowledgments

The research was funded by the Tanzania Initiative for Preventing Aflatoxin Contamination (TANIPAC) project.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] P. Chogo, N. Kassim, E. Mkoba, Identification of aflatoxin knowledge areas for developing natural language processing chatbot datasets, in: Proceedings of the First International Conference on the Advancements of Artificial Intelligence in African Context (AAIAC). IEEE Xplore, 2024, doi:10.1109/AAIAC60008.2023.10465309.
[2] M. Mkonda, Agricultural sustainability and food security in agroecological zones of Tanzania. In: Lichtfouse, E. (eds) Sustainable Agriculture Reviews 52. Sustainable Agriculture Reviews, vol 52. Springer, Cham, (2021), 10.1007/978-3-030-73245-5_9
[3] M. Ajmal, W. Bedale, A. Akram, J. Yu. Comprehensive review of aflatoxin contamination, impact on health and food security, and management strategies in Pakistan. Toxins J.. 10.3390/toxins14120845
[4] N. Jain, P. Jayakrishna, P. Jain, S. Pachpande, M. Singh, P. Kayal, J. Choudhari, Agribot: agriculture-specific question answer system. Department of science and technology government of Gujarat, 2019 IndiaRxiv. doi: 10.35543/osf.io/3qp98
[5] N. Darapaneni, R. Tiwari, A. Paduri, S. Saurav, R. Chaoji, Farmer-bot: an interactive bot for farmers. arXiv preprint arXiv:2204.07032 (2022). 10.48550/arXiv.2204.07032
[6] L. Estes, A. Kehs, O. Alzubi, G. Owomugisha, A field-based recommender system for crop disease detection using machine learning, Front. Artif. Intell. 6 (2023) 1010804, doi:10.3389/frai.2023.1010804.
[7] N. Rachmawati, E. Yulianti, StatMetaQA: a dataset for closed domain question answering in Indonesian statistical metadata, Data Brief (2024), doi:10.1016/j.dib.2024.110816.
[8] C. Shikali, R. Mokhosi, Enhanced African low-resource languages: Swahili data for language modelling enhancing African low-resource languages: Swahili data for language modelling, Data Brief (2020), doi:10.1016/j.dib.2020.105951.
[9] B. Masua, N. Masasi, In the heart of Swahili: an exploration of data collection methods and corpus curation for natural language processing, Data Brief (2024), doi:10.1016/j.dib.2024.110751.
[10] P. Chogo, E. Mkoba, N. Kassim, Aflatoxin question and answer Swahili dataset (AflaQnASD)", Mendeley Data, V1, (2024), 10.17632/zhpxzpnjfs.1
[11] S. Boni, F. Beed, M. Kimanya, E. Koyano, O. Mponda, D. Mamiro, B. Kaoneka, R. Bandyopadhyay, S. Korie, G. Mahuku, Aflatoxin contamination in Tanzania: quantifying the problem in maize and groundnuts from rural households, World Mycotoxin J. Vol 14 (N0 4) (2021) 553–564 https://www.ingentaconnect.com/content/wagac/wmj/2021/00000014/00000004/art00008?crawler=true&mimetype=application/pdf.
[12] C. Mollay, N. Kassim, R. Stoltzfus, M. Kimanya, Childhood dietary exposure of aflatoxins and fumonisins in Tanzania: a review, Cogent Food Agric. Vol 6 (No 1) (2020) https://www.tandfonline.com/doi/full/10.1080/23311932.2020.1859047.
[13] M. Lauri, WASP (Write a scientific paper): collecting qualitative data using focus groups, Early Hum. Dev. 133 (2019) 65–68, doi:10.1016/j.earlhumdev.2019.03.015.
[14] Y. Sumikawa, M. Fujiyoshi, H. Hatakeyama, M. Nagai, An FAQ dataset for E-learning system used on a Japanese University, Data Brief (2019), doi:10.1016/j.dib.2019.104001.