Digitalization And Validation of Traditional Agricultural Knowledge on Seed Sowing Using OCR And NLP

Bhavishya M Rai[1], Megha Rai[2], Reshma M[3], Shruthali Gaapathi Sharma[4]
[1,2,3]*Department of AI & DS, SIT, Valachil, Mangalore – 574143, Karnataka, India*
[4]*Associate Professor, Department of AI & DS, SIT, Valachil, Mangalore – 574143, Karnataka, India*
doi.org/10.64643/IJIRTV12I12-206590-459

Abstract*—Agriculture has a major role in supporting livelihoods, especially in regions that depends heavily on rainfall. Traditional agricultural knowledge related to seed sowing and crop behavior during rainfall is often documented in regional languages like Kannada, making it difficult to access and analyze. This project aims to develop a system that extracts such information from scanned Kannada documents using Optical Character Recognition (OCR) and translates it into English using Natural Language Processing (NLP). Further than translation, the system also attempts to evaluate the relevance of traditional agricultural practices by comparing them with present day weather patterns and modern agricultural knowledge. This allows us to understand whether these practices are still applicable in changing climate conditions. The overall goal is to make regional knowledge easily accessible and support better decision-making for both farmers and researchers.

I. INTRODUCTION

Agriculture matters a lot for so many people, you know, especially in places where everything hinges on the rain for farming. Things like how the rain falls, the soil quality, and when to sow the seeds, all affect if the crops will grow right from the start.

Farmers back in the day just went by what they knew from experience, and they wrote it down in languages like Kannada. But getting to that stuff now is tough, and analyzing it is even harder since not everyone can read it easily. Lately, with the climate shifting and rain not being as reliable, it makes you wonder if those old ways still work. I think that's why this paper is interesting, it tries to pull out info about seeds and rain from those old Kannada papers using something called OCR to read the text, then turns it into English with NLP tools. Once you have that, you compare it to what the weather knowledge holds up in modern farming. It feels like. some parts might not fit anymore, but others could still be useful, I am not totally sure.

II. LITERATURE REVIEW

Traditional knowledge is based on ages of skills and environmental transformation. Includes rainwater management methods useful for farming choices. Concentration needs to documentation regional knowledge (like Kannada texts) [1]. Seeds are treated with cow dung, milk, buttermilk improves germination & disease tolerant. Covering soil after sowing prevents loss due to wind/rain and protects seeds [2]. Techniques like utera cropping use leftover soil moisture efficiently mixed cropping reduces threats of crop failure [3]. The OCR for printed text can be used to scan old, classical books and convert them into browsable text, making such literature easily available. Identification technology for printed text can also help in making books available for visually challenged individuals [4]. Optical Character Recognition (OCR) is used for automatic reading of text images and converting them into machine readable format. In OCR systems, the document image is segmented into lines, words, and characters before recognition. Preprocessing steps such as skew detection and noise removal are necessary to improve identification and accuracy [5]. OCR lets the conversion of text from images into machine-readable format, assisting further processing and storage.

The performance of Tesseract OCR can be upgraded using preprocessing techniques such as adaptive thresholding, median filtering, and edge detection. The system performance is evaluated using metrics such as Word Error Rate (WER) and Character Error Rate (CER) [6].

Digital images are increasing rapidly in many fields such as education, engineering, and healthcare. These images often contain useful text information. Extracting this text helps in understanding and using the data properly. In this work, we use Optical Character Recognition (OCR) to automatically extract text from images.
The process includes text detection, positioning, separation, and binary conversion. Finally, the extracted text is converted into readable format. This method is useful for applications like e-library, e-books, and image search systems [8]. BERT is a pretrained language model that get the context of words from both directions in a sentence, making it easier for opinion extraction. Unlike traditional methods such as TF-IDF and Bag of Words, BERT collects deeper expressive meaning and correlation between words.

Researchers further improve its accuracy by fine-tuning the model with additional layers and domain-specific data [7]. In the digital era, almost everything is automated, and information is stored and shared in digital forms. However, there are several situations where the data is not digitized, and it might become essential to extract text from those to store in digitized form [9]. OCR systems capable of recognizing characters have gained maximum concentration of researchers these days, especially when it comes to recognizing ancient documents. Digitization of these documents to make them readable and also searching from paper-based data becomes a great challenge. Literature shows that there are a lot of OCR systems that use different feature extraction and segmentation techniques to calculate the recognition accuracy [10]. Preserving historical documents is essential for safeguarding cultural heritage and making historical knowledge accessible to future generations.

Old digitization methods often fail to record and process corrupted or handwritten texts effectively, low visibility and usability. Existing Optical Character Recognition (OCR) techniques struggle with mistakes due to changes in handwriting styles, faded ink, and document damage, making it difficult to convert these texts into usable digital formats [11]. As the number of digitized historical documents has increased rapidly during the last a few decades, it is necessary to provide efficient methods of information retrieval and knowledge extraction to make the data accessible. Such methods are dependent on optical character recognition (OCR) which converts the document images into textual representations [12]. BERT is an NLP model that understands the meaning of words based on their context in a sentence. It performs better than older models like Word2Vec because it reads text in both directions.

The paper reviews BERT's use in tasks such as translation, question answering, and sentiment analysis. It also reviews ALBERT and Distil BERT that reduce memory and improve speed. [13] The study explains how seeds, seedlings, and saplings help in the regeneration of tropical rainforests. Seeds come from both local trees and outside sources through animal dispersal, which increases diversity. However, local species are more likely to survive and grow into mature plants. Many species are lost during early stages due to factors like low light, competition, and predation. Plants with larger seeds have a better chance of survival in shaded conditions.

Overall, forest regeneration depends on seed dispersal, plant characteristics, and environmental conditions [14]. With rise of digital age, there is an explosion of information in the form of news, articles, social media, and so on. Much of this data lies in unstructured form and manually managing and effectively making use of it is tedious, boring and labor intensive. This explosion of information and need for more sophisticated and efficient information handling tools gives rise to Information Extraction (IE) and Information Retrieval (IR) technology [15]. Seed treatment is an important agricultural practice where seeds are treated with physical, chemical, or biological agents before sowing to protect them from pests, diseases, and environmental stress. It helps improve seed germination, seedling growth, and overall crop yield while reducing dependence on excessive chemical pesticides. Various methods such as seed coating, pelleting, priming, and biological treatments are used to enhance seed performance [16]. Despite the fact that comprehension and validation seem to co-occur during many discourse experiences, validation has only recently attracted the attention of text comprehension researchers. This relative lack of interest may be due in part to the popularity of twostep models of comprehension and evaluation in

psychology [17]. As domain experts often do not feel confident in judging the correctness and completeness of process models that system analysts create, the validation often has to regress to a discourse using natural language. In order to support such a discourse appropriately, so-called verbalization techniques have been defined for different types of conceptual models [18].

The current study aims to contrast text- and knowledge-based monitoring to investigate their unique influences on processing and whether validation is passive. Therefore, we collected reading times in a self-paced experiment using expository texts containing information that conflicts with either the preceding text or readers' background knowledge [19].

III. PROPOSED SYSTEM

The suggested system will assist in extracting traditional knowledge about seed sowing practices from documents written in the Kannada language. Optical character recognition technology converts the scanned images of texts into digital data, which is then processed using NLP methods to identify significant knowledge such as rainfall patterns and seed sowing techniques. Afterward, the knowledge is analyzed to derive valuable insights that can be used in agriculture.

IV. OBJECTIVES

The primary goal of this project is to create a system which will help in retrieving traditional knowledge regarding seed sowing and the conditions of rainfall from the documents written in regional languages. The purpose of this initiative is to make this crucial knowledge available, accessible and usable in contemporary times.

The first goal is to use Optical Character Recognition (OCR) technology to transcribe unstructured agricultural data, like documents and scripts written in the local language Kannada, into machine readable text. OCR is a very helpful tool which would make it easy to handle the unstructured agricultural data. The next goal is to employ NLP methods to identify and extract information on rainfall conditions, sowing timing, and seeding from the textual agricultural data. It would help in converting the raw textual data into useful insights. The system will also be able to find out the relationship between the two, which is an important goal of the project.

Another goal is to digitize the indigenous agricultural knowledge in the form of structured information, which would help in preserving it for future generations and make it easily available.

V. SYSTEM ARCHITECTURE

The whole system starts off with these seed sowing books in Kannada, you know, like scanned pages or PDFs that have all this old farming info on crops and when to plant them. It feels like that's the main input, pulling in traditional knowledge about stuff like soil and seasons.

Then there's this OCR step to pull out the text from those images. Since it's in Kannada and not readable by machines yet, you have to convert it somehow. I think that's crucial because without it, nothing else works.

After that, the text gets cleaned up a bit. Things like fixing noise or wrong spellings, breaking it into words, getting rid of junk. It makes sure everything's ready for the next parts, I guess. Preprocessing seems messy but necessary.

Now, using NLP, the system picks out key details, like what crop, when to sow, soil types, all that. It turns the loose text into something structured, maybe lists or records. Some parts might overlap here, like farming methods tied to seeds.

Once it's structured, you can actually use it, generating recommendations on sowing times or crop choices from the old books. That part stands out, turning history into practical advice.

But then validation comes in, comparing this to real current data, like rainfall or actual farm records. It checks if the extracted stuff holds up, accuracy and all. I might be oversimplifying, but it seems like that's to make sure it's reliable.

Finally, the output wraps it up with verified info, reports on how accurate it is, and some insights for farmers. Helps preserve that traditional knowledge while making decisions easier today.

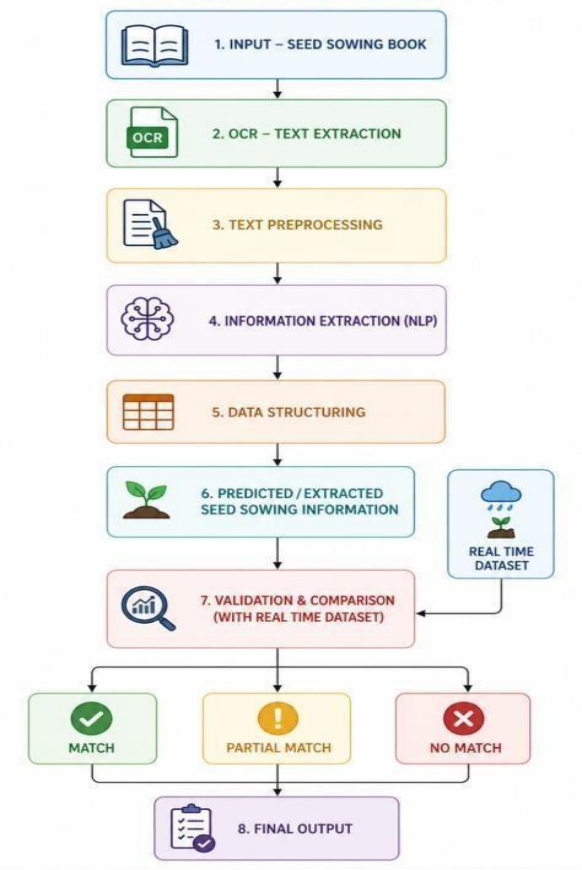

Fig (1). Flow Diagram of Digitalization and Validation

VI. TECHNOLOGY USED

1. Programming Language used Python

The solution is written in Python because of its ease of use, extensibility, and rich ecosystem of libraries for data handling and machine learning operations.

2. OCR Technology used Tesseract OCR

The process of OCR is done using Tesseract OCR – a powerful tool for recognizing handwritten and printed text from image files. In particular, this OCR recognizes text in multiple languages including Kannada.

3. Natural Language Processing (NLP)

NLP processes are performed by using tools such as NLTK and spaCy. These allow: Tokenization, Stop word removal, Keyword extraction, Information extraction.

Thanks to NLP, it is possible to extract entities related to agriculture, i.e., crop, time, seeds, etc.

4. Data Processing and Analysis

Pandas and NumPy are used to do necessary manipulations and analyses with the data. It helps organize the data, perform comparison, and draw conclusions.

5. Real-Time Dataset Integration

A real-time or current database consisting of agricultural-related data (such as rainfall, sowing, etc.) can be used in this case to make sure that extracted data is accurate.

6. Development Environment

Tools such as Jupyter Notebook or Visual Studio Code can be used to develop the solution effectively.

VII. CONCLUSION

The suggested model efficiently demonstrates the methodology of obtaining data regarding seed sowing from ancient farming records through the integration of OCR and NLP algorithms. The architecture supports the extraction of vital agricultural parameters including crop type, seed sowing period, and other procedures, which are further verified against contemporary data sets.

The inclusion of validation allows one to ensure that the obtained data is reliable and accurate. It helps to build confidence in the results and promotes the use of the proposed system for various agricultural purposes.

Apart from promoting easy access to data, the model plays an important role in preserving cultural and traditional agricultural knowledge by transforming it into digital and useful information. Comparing it with modern data shows the value of old farming practices in contemporary agricultural conditions.

All things considered, the project serves as a basis for further improvements of the system such as multilingual support, enhanced predictive models, and incorporation into live agricultural advisory services.

REFERENCES

[1] Mishra, S. R. K. Singh, and A. A. Raut, *Traditional Knowledge in Agriculture*. New Delhi, India: Division of Agricultural Extension, ICAR, 2020, p. 39.

[2] J. U. Duncombe, "Infrared Navigation—Part I: An Assessment of Feasibility," IEEE

Transactions on Electronic Devices, vol. ED-11, pp. 34-39, Jan. 1959.

[3] C. Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, and M. L. Miller, "Rotation, Scale, and Translation Resilient Public Watermarking for Images," IEEE Transactions on Image Processing, vol. 10, no. 5, pp. 767-782, May 2001.

[4] H. R. S. Kumar and A. G. Ramakrishnan, "Lipi Gnani: A Versatile OCR for Documents in Any Language Printed in Kannada Script," ACM Journal/Proceedings, doi: 10.1145/3387632.

[5] N. Anakpluek, W. Pasanta, L. Chantharasukha, P. Chokratansombat, P. Kanjanakaew, and T. Siriborvornratanakul, "Improved Tesseract Optical Character Recognition Performance on Thai Document Datasets," Big Data Research, vol. 39, Art. no. 100508, 2025, doi: 10.1016/j.bdr.2025.100508.

[6] J. P. K. K. M. V. M. A., "Text Recognition – Performance Comparison of Tesseract and Paddle OCR in OpenCV," International Journal of Emerging Technologies and Innovative Research, p. 161, 2026, doi: 10.48175/ijetir9227.

[7] D. Deepa, "Bidirectional Encoder Representations from Transformers (BERT) Language Model for Sentiment Analysis Task: Review," Turkish Journal of Computer and Mathematics Education, vol. 12, no. 7, pp. 1708-1721, 2021, doi: 10.17762/turcomat. v12i7.3055.

[8] C. Kaundilya, D. Chawla, and Y. Chopra, "Automated Text Extraction from Images Using OCR System," in Proc. Int. Conf. Computing for Sustainable Global Development (INDIACom), 2019, pp. 145-150.

[9] R. Mittal and A. Garg, "Text Extraction Using OCR: A Systematic Review," in Proc. 2nd Int. Conf. Inventive Research in Computing Applications (ICIRCA), 2020, pp. 357-362, doi: 10.1109/ICIRCA48905.2020.9183326.

[10] A. Moudgil, S. Singh, and V. Gautam, "An Overview of Recent Trends in OCR Systems for Manuscripts," in Advances in Intelligent Systems and Computing. Singapore: Springer, doi: 10.1007/978-981-16-4284-5_46.

[11] M. Askarov, A. Gafforov, A. Darmonova, M. Dadakhonova, T. Ismailov, and U. Qushnazarova, "Preserving Historical Documents Using OCR and Natural Language Processing (NLP)," in Proc. IEEE Int. Conf. Computational Intelligence and Information Engineering Systems (ICCIES), 2025, doi: 10.1109/ICCIES63851.2025.11032769.

[12] J. Martínek, L. Lenc, and P. Král, "Building an Efficient OCR System for Historical Documents with Little Training Data," Neural Computing and Applications, vol. 32, no. 23, pp. 17209-17227, 2020, doi: 10.1007/s00521-020-04910-x.

[13] S. Gardazi, N. M., A. Daud, M. K. Malik, A. Bukhari, T. Alsahfi, and B. Alshemaimri, "BERT Applications in Natural Language Processing: A Review," Artificial Intelligence Review, vol. 58, no. 6, 2025, doi: 10.1007/s10462-025-11162-5.

[14] M. Martinez-Ramos and A. Soto-Castro, "Seed Rain and Advanced Regeneration in a Tropical Rain Forest," Vegetatio, vol. 107-108, no. 1, pp. 299-318, 1993, doi: 10.1007/BF00052231.

[15] E. A. Olivetti, J. M. Cole, E. Kim, O. Kononova, G. Ceder, T. Y. Han, and A. M. Hiszpanski, "Data-Driven Materials Research Enabled by Natural Language Processing and Information Extraction," Applied Physics Reviews, vol. 7, no. 4, 2020, doi: 10.1063/5.0021106.

[16] K. Sharma, U. Singh, P. Sharma, A. Kumar, and L. Sharma, "Seed Treatments for Sustainable Agriculture—A Review," Journal of Applied and Natural Science, vol. 7, no. 1, pp. 521-539, 2015, doi: 10.31018/jans. v7i1.641.

[17] T. Richter and D. N. Rapp, "Comprehension and Validation of Text Information: Introduction to the Special Issue," Discourse Processes, 2013, doi: 10.1080/0163853X.2013.855533.

[18] "Supporting Process Model Validation Through Natural Language Generation," IEEE Transactions on Software Engineering, doi: 10.1109/TSE.2014.2327044.

[19] M. L. van Moort, A. Koornneef, and P. van den Broek, "Validation: Knowledge- and Text-Based Monitoring During Reading," Discourse *Processes*, doi: 10.1080/0163853X.2018.1426319.