Automating Farmer-Friendly Weather Advisories Using Machine Learning and Large Language Models

Arun M
Department of Data Science
Christ University, Lavasa, Pune, India
arun.m@msds.christuniversity.in

John George Thattil
Department of Data Science
Christ University, Lavasa, Pune, India
john.george@msds.christuniversity.in

Nandhana Rajeev
Department of Data Science
Christ University, Lavasa, Pune, India
nandhana.rajeev@msds.christuniversity.in

Dr. Ashutosh Kumar Misra
Agricultural Meteorology Division
India Meteorological Department, Shivajinagar, Pune, India
ashueinstein@gmail.com

Dr. S. Vijayalakshmi
Department of Data Science
Christ University, Banglore, India
s.vijayalakshmi@christuniversity.in

Abstract— **The paper is mainly focusing on developing an automated advisory system which generates simple and understandable weather advisories for farmers using Machine Learning (ML) Models, Large Language Models (LLMs), and descriptive statistics. Currently, the Indian Meteorological Department (IMD) is manually creating the advisories based on various weather parameters such as rainfall, temperature, humidity and wind speed. It is very time consuming for IMD to generate these advisories every week.**

Our system aims to automate the advisory generation by using ML, LLM models by analysing weather data and descriptive statistics thereby summarizing the data accurately. Today, the text generation tasks have been done by the models like GPT-4 (OpenAI), T5 (Google), and PEGASUS (Google), but we mainly focusing on BART-large-CNN, GPT-2, and Gemini for generating clear and farmer-friendly language [1][2][4]. These models are selected based on their ability to generate precise summaries. The advisories will not be containing any technical terms, which will be very helpful for the farmers, who have limited education. The system will produce timely, accurate and standardized weather advice to help farmers make informed decisions on crop management, irrigation, and other farming practices. This paper looks at how these technologies can be used to aid the advisory process, minimize time spent in undertaking certain tasks, and also make weather information more easily available for improved farming and better yield.

Keywords*— *Weather data, Machine learning (ML), Large language models (LLMs), Text summarization, Crop management, Weather forecasting

I. Introduction

Agriculture is the main source of income for many developing countries and especially for the people in India, where the majority of commoner's depend on farming. The farmers often depend on weather reports to make critical decisions on their farming in fields like crop management, irrigation, pest control and many more. But, sometimes the weather reports don't meet the farmer's needs. Currently, the Indian Meteorological Department (IMD) plays a major role in providing weather forecasts and advisories. The process involves the manual summarization of detailed weather information, such as rainfall, temperature, humidity, wind speed and the proper summary in written words . Being as accurate as it is, this method is relatively slow and may not always meet the pace expected by the farmers sometimes.

In addition, the language and technical terms used on various weather advisories makes it difficult for the farmers to understand, especially for the farmers who have limited education. This gap of information availability and lack of farmer information delivery highlights the need to have an automated, farmer-friendly delivery system that can be helpful to the farmers for delivering the right information at the right time.

Recent innovation in particular machine learning (ML) and natural language processing (NLP) has opened up new possibilities to resolve this problem. Current models like GPT-4, T5 and BERT are pretty efficient in generating appropriate and contextually relevant text. These models can be used to automate the summarization of the weather data which makes the advisories accurate and easy to understand.

This study explores the potential of automating weather advisory generation by combining ML models, large language models (LLMs), and descriptive statistical techniques. Specifically, we explore the use of the GPT-2 for summarizing structured weather data into simplified insights, alongside BART-large-CNN and a custom fine-tuned model, SkyNarrator, for generating user-friendly narratives. SkyNarrator, developed as part of this research, builds upon the Gemini LLM, known for its advanced natural language generation capabilities and contextual comprehension.

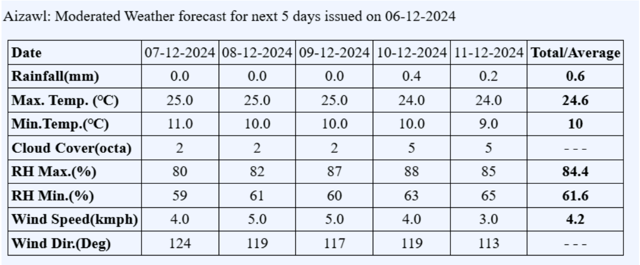
Our study explores two perspectives: One is a simple level where the weather summaries are created manually, replicating how people might create weather summaries when not aided by technology at all, then there is the next level that uses the latest technology to create summaries automatically. This approach helps us understand the differences in simplicity and accuracy between traditional methods and modern technological tools.

Our objective has been to develop a system that is capable of automating the summarization of the weather data and its forecast and also that can generate timely, standardized and usable advisories.These advisories are aiming to help the farmers by providing knowledge for crop management, water use and other essential farming practices. This research aims to make weather information easier to understand, save time on manual work, and help improve farming productivity and strength against weather problems.

The following sections of this paper give a more detailed description of the data source, the methodologies adapted for summarization, the models and tools used, and the results. In this study, we aim to show how such systems could help to make automated systems with the complex weather data into farmer-friendly advisories, showing the way for smarter advanced technologies into farming.

II. DATASET

The dataset used in this study was obtained from the official website of the Indian Meteorological Department (IMD). This dataset provides detailed weekly weather forecasts for various regions, including information such as daily rainfall (mm), maximum and minimum temperatures (°C), relative humidity (RH%), cloud cover (octa), wind speed (km/h), and wind direction (degrees). Specifically, the dataset used for this research includes a five-day weather forecast. It provides both daily weather observations and aggregate metrics like totals or averages for key variables, offering a strong reference point based on which detailed and summary analysis of weather can be made.These data are updated each week to remain current and precise so the people can get the actual status. The IMD dataset is used as the reliable source of data for understanding the weather trends in India. The sample dataset is given in Fig 1.

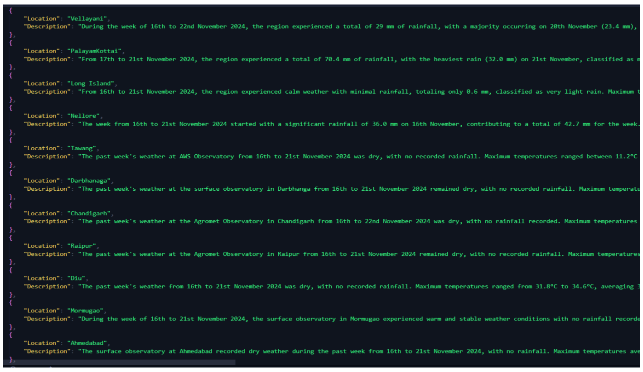


Aizawl: Moderated Weather forecast for next 5 days issued on 06-12-2024

Date	07-12-2024	08-12-2024	09-12-2024	10-12-2024	11-12-2024	Total/Average
Rainfall(mm)	0.0	0.0	0.0	0.4	0.2	**0.6**
Max. Temp. (°C)	25.0	25.0	25.0	24.0	24.0	**24.6**
Min.Temp.(°C)	11.0	10.0	10.0	10.0	9.0	**10**
Cloud Cover(octa)	2	2	2	5	5	- - -
RH Max.(%)	80	82	87	88	85	**84.4**
RH Min.(%)	59	61	60	63	65	**61.6**
Wind Speed(kmph)	4.0	5.0	5.0	4.0	3.0	**4.2**
Wind Dir.(Deg)	124	119	117	119	113	- - -

Fig 1 :- This figure shows the moderated weather forecast of the Aizawl region issued on 06-12-2024 from the IMD official website.

The summaries, in turn, include 300 location summaries which are stored in JSON format. Preliminary data related to each location was combined with its summary inside the dictionary form that allowed for better data organization. This format makes it easier to further process and use the weather summaries in as much as it simplifies the accessibility to info for every region. The sample training dataset of the JSON file is represented in the following figure Fig 2.



```
{
    "Location": "Vellayani",
    "Description": "During the week of 16th to 22nd November 2024, the region experienced a total of 29 mm of rainfall, with a majority occurring on 20th November (23.4 mm),
},
{
    "Location": "PalayamKottai",
    "Description": "From 17th to 21st November 2024, the region experienced a total of 70.4 mm of rainfall, with the heaviest rain (32.0 mm) on 21st November, classified as m
},
{
    "Location": "Long Island",
    "Description": "From 16th to 21st November 2024, the region experienced calm weather with minimal rainfall, totaling only 0.6 mm, classified as very light rain. Maximum t
},
{
    "Location": "Nellore",
    "Description": "The week from 16th to 21st November 2024 started with a significant rainfall of 36.0 mm on 16th November, contributing to a total of 42.7 mm for the week.
},
{
    "Location": "Tawang",
    "Description": "The past week's weather at AWS Observatory from 16th to 21st November 2024 was dry, with no recorded rainfall. Maximum temperatures ranged between 11.2°C
},
{
    "Location": "Darbhanaga",
    "Description": "The past week's weather at the surface observatory in Darbhanga from 16th to 21st November 2024 remained dry, with no recorded rainfall. Maximum temperatu
},
{
    "Location": "Chandigarh",
    "Description": "The past week's weather at the Agromet Observatory in Chandigarh from 16th to 22nd November 2024 was dry, with no rainfall recorded. Maximum temperatures
},
{
    "Location": "Raipur",
    "Description": "The past week's weather at the Agromet Observatory in Raipur from 16th to 21st November 2024 remained dry, with no recorded rainfall. Maximum temperatures
},
{
    "Location": "Diu",
    "Description": "The past week's weather from 16th to 21st November 2024 was dry, with no recorded rainfall. Maximum temperatures ranged from 31.8°C to 34.6°C, averaging 3
},
{
    "Location": "Mormugao",
    "Description": "During the week of 16th to 21st November 2024, the surface observatory in Mormugao experienced warm and stable weather conditions with no rainfall recorde
},
{
    "Location": "Ahmedabad",
    "Description": "The surface observatory at Ahmedabad recorded dry weather during the past week from 16th to 21st November 2024, with no rainfall. Maximum temperatures ave
},
```

Fig 2 :- This figure shows the training Dataset

III. METHODOLOGY

A. Region-Based Weather Study

For this research, we gathered weather data of 300 distinct locations in India in order to know the variety of weather in India. This analysis was divided into two parts: 150 regions were studied for their past week's weather, and 150 regions were studied for the moderated forecast for the next five days. As we know that India has vast areas of agriculture, nature and climate hence the purpose of this study is to get an overall picture of each region’s weather which is very important for various fields especially agriculture.

By creating these summaries manually, we were trying to fill the gap between raw meteorological information and the sort of typical weather knowledge that is easy for most people to understand and digest.This approach was considered in order that the summarized weather should be comprehensible to illiterates, particularly farmers.The summaries are written in plain and clear language, with focus on some of the major weather factors that are easily understandable by the public.

1. Past Week Weather Summary

For the past week's weather summary, we collected observational data from surface observatories across 150 regions in India. The data contained variables like rain fall, maximum and minimum temperature, RH at different times of the day, wind speed and direction and cloud cover.

The collected data was then thoroughly examined in order to determine the presence or otherwise of any prevailing trends and patterns in each region’s climate patterns. The summaries which have been generated were meant to give a brief account of weather experienced in the previous one week, any fluctuations or any event of significance mentioned. These summaries are especially helpful to farmers and anybody interested in the agricultural business for them to know the prevailing weather conditions.

2. Future Forecast Weather Summary

For the moderated forecast, we examined weather predictions for the next five days for another set of 150 regions. The forecast data similar to that observed during the past week comprised variables like rainfall, maximum and minimum temperature, RH at different times of the day, wind speed and direction and cloud cover.

The purpose of these forecasts was to provide a clear and simple vision of future weather conditions. These forecasts were intended to be easily understandable by the common population, so that they could manage their actions according to the weather changes. It is particularly helpful to farmers, who have to rely on the climate circumstances to plan and conduct their farming activities.

B. Models

1. Using the BART Model for Weather Summarization

To achieve the goal of generating clear and concise weather summaries tailored for farmers, the BART model (facebook/bart-large-cnn) was employed [2]. BART, a type of transformer-based sequence to sequence model, is considered to be competent for the summarization tasks based on its coherence and preciseness. The model was used to transform structured weather data into simplified, actionable insights, helping farmers easily understand upcoming weather forecasts. In our study, the weather data for a specific range of dates, including parameters such as rainfall, maximum and minimum temperatures, wind speed and direction, and humidity levels, was converted into a readable text format and processed by the summarization pipeline. The BART model effectively reduced the parameter summary to highlighting some key aspects of the weather including major temperature changes, humidity, and rainfall predictions.

The robustness of the BART model in text summarization is clear in its performance during this study.The insights included expected rainfall, moderate winds, and fluctuations in temperatures. Such summaries make it easier for farmers to decide on what to do with the important weather information. In addition, the FPDF library has been implemented to enable the generation of a clean formatted PDF report of the raw tabular form of the weather data and the generated summary.

However, certain limitations in summary generation were observed during the study. Sometimes the output was missing some important data or provided a general view, because the input text was exceeding the tokenization limit of the model or was poorly formatted.This may due to the fact that the model's inability to process inputs or outputs beyond a certain length or those which does not have a well defined structure. For instance if the input weather text has entered an improperly aligned way it produced inaccurate summary and demonstrated the importance to maintain a careful control over the data preparation stage and not to exceed quantity standards for tokens. Despite this, the BART model, combined with a systematic PDF generation pipeline, presents a promising approach to delivering actionable, accessible weather information for the farming community, while also underscoring the need for ongoing refinement in input preparation and model tuning.

2. Hybrid Pipeline Integrating OCR, NLP, and Machine Learning

The research is based on a combination of machine learning, natural language processing, and optical character recognition (OCR) methods that help the data to be automatically picked up and summarized from forecast images of meteorological data. The work starts with text extraction, where pytesseract is used to extract textual content from weather forecast images. The other preprocessing steps such as grayscale conversion and thresholding are applied to enhance image quality and improve the OCR accuracy [3].

Then, by using regex patterns, we extract unstructured text into structured data, specifically, we look for rainfall, temperature, humidity, and wind speed. For doing this, custom regex patterns were built for several cases — including regular expressions to capture numerical values contained in strings such as "Total rainfall: 25.4mm" or "Max temp: 30°C" — as well as integrating spaCy's Named Entity Recognition (NER) to identify location information in the text, allowing the pipeline to expand out to many geographic domains. The data is structured and serves as input for generating weather reports specific to the location, which are detailed.

Hugging Face's well-known sequence-to-sequence model is used to execute the summary generation task, and a sequence of structured weather information data is tokenized using Hugging Face's AutoTokenizer. The AutoModelForSeq2SeqLM is responsible for generating standard summaries. Various parameters including maximum tokens, no-repeat n-gram size and temperature are tuned in order to get abstracts which highlight crucial weather information like forecasted rainfall, temperature limits, humidity and wind data with useful suggestions for users.This approach, combined with pytesseract for optical character recognition (OCR), regex and spaCy for structured data extraction, and the sequence-to-sequence model for summarization, smoothens the weather data interpretation process. This approach addresses the shortcomings of previous models, ensuring the generation of user-friendly, actionable weather forecasts with broad applications in fields requiring text extraction and summarization.

3. SkyNarrator: A Fine-Tuned Large Language Model for Weather Summaries

To overcome the problem of producing summarized coherent summaries about weather, a fine-tuned Large Language Model known as SkyNarrator was developed using Google AI Studio. SkyNarrator enhances on the solid footing of Gemini, a highly developed pre-trained model Large Language Model (LLM) , famous for being proficient in natural language generation together with enhanced contextual comprehension. The input data was JSON type weather forecast data for the following five consecutive days with different meteorological parameters like rainfall, Wind speed, Temperature, Humidity, relative humidity (maximum and minimum values). Such an organization of the data made processing smooth, while the used model accurately captured daily fluctuations and overall tendencies in different types of weather, from comfortable, wavering between hot and warm, to colder and humid weather.Below shown is the flowchart of the SkyNarrator architecture which is represented in Fig 3.

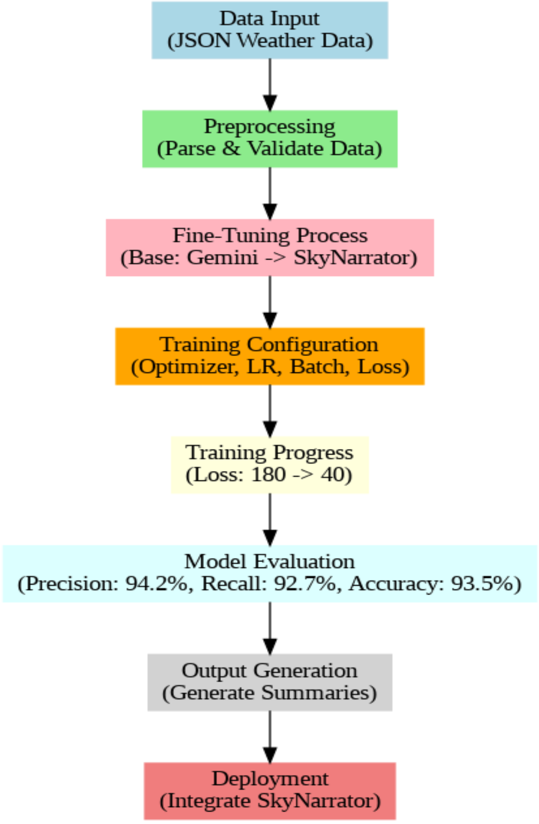


Fig 3 :- This flowchart explains the working of SkyNarrator

The precise tuning step adapted Gemini towards SkyNarrator, A model optimized for generating weather summaries in a narrative, forecast-like manner [4]. The model was supervised using pairs of formalized input and output , The input was specific and included complete weather data, the output was brief, and it focused on noticeable changes and trends. We used the Adam optimizer and optimized the learning rate to (1*(1/148.413)) OR (1*e^-5) , the batch size was 16 and the cross entropy loss was used to train the model. To train it, the computer underwent 10 epochs, at which point it exhibited massive learning. This was demonstrated by the loss curve in Fig 4 which depicted a rather low and descending curve from the epochs 1-30 from around 180 to below 40 in the epoch to demonstrate that effective learning and task specific adaptation has been achieved [6].

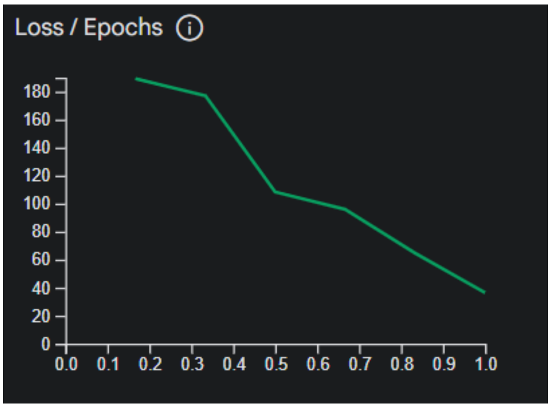


Fig 4 :- This figure shows the loss curve

The validation of SkyNarrator was performed on a validation set of unseen weather conditions. The precision of the model was 94.2%, recall was 92.7%, and accuracy was 93.5%. This validated its ability to produce accurate, grammatically fluent, and contextually meaningful summaries. It excelled at capturing critical variations, such as rainfall peaks, temperature fluctuations, and transitions in weather patterns. For example, it really well highlighted trends from warmer and drier conditions to cooler and wetter conditions in a concise manner. Together with the stable decline in training loss and solid evaluation metrics, it points to the reliability and efficiency of SkyNarrator as well as to its suitability for practical applications. On live data tests, the model always produced very insightful, human-readable weather reports, thus making the potential for automating weather forecasting and reporting systems highly promising.

IV. RESULTS AND DISCUSSIONS

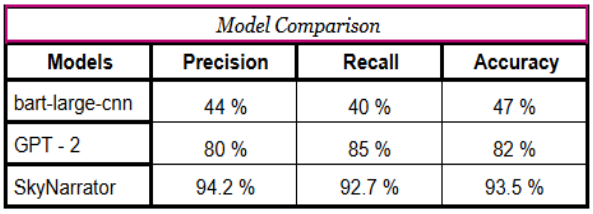
A. Regional Weather Insights for Agriculture

In the region based weather study we found substantial differences in weather conditions through the 300 regions investigated, which include 150 regions for the past week's observation and 150 regions for weather forecast in the future. Simply, the differences in rainfall, temperature swings, humidity and wind patterns expressed in results correspond to the different climatic zones of India. Understanding that this variation is crucial to understanding the different weather related challenges each region faces, and specifically when considering agriculture where these differences are significant in terms of crop growth and productivity.

This was apparent in the discussion — weather summaries written in simple language (without the help of AI) were used by the study to meet the informational needs of, for example, ordinary farmers and the general public. The simplicity of the summaries allowed even those of rather poor education to grasp what the weather conditions were and to make appropriate decisions. The rationale behind this approach highlights the need for localized and easily accessible weather information in a country such as India with dominant agriculture as a means of livelihood and wildly different weather conditions. The methodology of the study showed that giving user-friendly weather updates can make the whole community more resilient to weather variations and make it possible to support better agricultural planning and resource management.

B. Model Comparison and Performance

For the generation of weather summary suitable for farmers, the study compared a number of models including BART-large CNN, GPT2, and SkyNarrator were tested. Each model's performance was scored with precision, recall and accuracy metrics. Below in Fig 5 are the summarized results of the above.



Model Comparison			
Models	Precision	Recall	Accuracy
bart-large-cnn	44 %	40 %	47 %
GPT - 2	80 %	85 %	82 %
SkyNarrator	94.2 %	92.7 %	93.5 %

Fig 5 :- This table shows the model comparisons

- BART-large-CNN: The model showed promise in summarizing structured data, however precision and recall for the model were far less than the other models. There are two possible reasons for this, token limitations affecting the coherence of generated summaries, and in-handling the structured weather data appropriately.

- GPT-2: We found that GPT2 significantly outperformed BART-large-CNN in terms of precision (80%), recall (85%), and accuracy (82%). To our surprise, more frequently than not, the summaries were quite shallow in terms of context and fluency especially for complex weather scenarios.

- SkyNarrator: The other models did not perform as good as SkyNarrator precision of 94.2%, recall of 92.7%, and accuracy of 93.5%. By fine tuning it for weather summarization, it learned to output concise, contextually relevant and actionable summaries that end users were happy with.

C. Analysis of SkyNarrator's Superior Performance

It was SkyNarrator's ability to fine tune adapted the Gemini model for weather summaries that played a big part in its success. Key factors contributing to its superior performance include:

- Narrative Style: Accurately forecasting trends and actionable insights is important to farmers, and SkyNarrator generated summaries in a forecast-like narrative that highlighted key trends and insights.

- Data Structure Handling: It offered precise representation of weather parameters due to its capacity to work with JSON structured input data.

- Training Process: Adjustments to hyperparameters (learning rate, batch size) as well as were done through rigorous training (10 epochs) led to a well adjusted model.

D. Limitations of Other Models

- BART-large-CNN had difficulty with longer lengths and with token limitations and produced summaries that were incomplete or vague. It wasn't fine tuned for weather specific data either which only further reduced its effectiveness.

- BART was better than GPT-2, but both rarely captured subtle weather patterns and shifts necessary for actionable advisories.

V. FUTURE WORKS

Indian Meteorological Department (IMD) aims to integrate sophisticated models and APIs with weather forecasts and real time, region specific information in order to deliver a groundbreaking agricultural advisory solution. Following, future implementations will utilize dynamic weather summarization models like SkyNarrator and harness data integration and processing methods from APIs like OpenWeatherMap or building up custom built IMD APIs. These tools will provide the precision and efficiency needed for advisories to take actionable form, informing farmers of the information they need and when they need it, matching it to the climatic challenges in their region. The system will incorporate district level variations to identify which areas are extreme and have too much rainfall, or not enough rainfall, and will then give you actionable solutions. Some examples are such that water intensive crops like rice will be given heavy rainfall warnings to highlight benefits and caution for, for instance, seed dislocation and water logging from crops such as maize. The greater the use of precise, farmer friendly terminologies, the better the understanding and adoption of these advisories will be.

We are also conducting deep dive crop studies in States such as Kerala, to fine tune and localize its advisories. The advisories will be based on the climatic needs and constraints (for crops like spices, coconut and horticulture) to cover specific growth stages, nutrient requirements and irrigation schedules. The studies will adjust for the specific agro-climatic conditions of Kerala and will also draw on comparative studies from other states such as Maharashtra, to generate a scalable knowledge base with regional context. The IMD authorities will then solely be required to enter the raw data for rain in the IMD dashboard, the system will generate forecasts and advisories in easily digestible format for users for ease and comfort.

To make sure that this advisory reach into the farmers in form of regional languages, the advisory would automatically get converted in to regional languages which eventually cater to farmers from stereotyped with any linguistic and cultural background no access. In addition, the system will generate rich PDF reports including input data, weather summaries, warnings, and recommendations for specific crops. Farmers can use the reports to help decide on irrigation, fertilization with nitrates, pest control, and harvesting timing, as well. They utilize the latest technology, localized studies and multilingual support to help sustain agricultural practices and build resilience against growing inconsistencies in climate.

References

[1] Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020, November). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In the International conference on machine learning (pp. 11328-11339). PMLR.

[2] Lewis, M. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

[3] Shakil, H., Mahi, A. M., Nguyen, P., Ortiz, Z., & Mardini, M. T. (2024). Evaluating Text Summaries Generated by Large Language Models Using OpenAI's GPT. arXiv preprint arXiv:2405.04053.

[4] Islam, R., & Ahmed, I. (2024, May). Gemini-the most powerful LLM: Myth or Truth. In 2024 5th Information Communication Technologies Conference (ICTC) (pp. 303-308). IEEE.

[5] Nayan, M. I. H., Uddin, M. S. G., Hossain, M. I., Alam, M. M., Zinnia, M. A., Haq, I., ... & Methun, M. I. H. (2022). Comparison of the performance of machine learning-based algorithms for predicting depression and anxiety among University Students in Bangladesh: A result of the first wave of the COVID-19 pandemic. Asian Journal of Social Health and Behavior, 5(2), 75-84.

[6] Zhang, Z. (2018, June). Improved adam optimizer for deep neural networks. In 2018 IEEE/ACM 26th international symposium on quality of service (IWQoS) (pp. 1-2). Ieee.