Original Research Article
Understanding Farmers' Needs through Visual Analytics of Queries

ABSTRACT

Aims: To identify the trending queries raised in Kisan Call Centers by employinng visual analytics.
Study design: Analyzing secondary data on farmers' queries by employing various visual analytics techniques.
Methodology: The study was conducted using various visual analytics techniques. The choropleth map was used to visualize the district-wise spatial distribution of queries within Telangana state. The line graph was used to understand the temporal trend in queries raised during the study period. The word clouds, bigrams and network maps were used to identify the words that are often appeared in the query text. Before visual analytics, the raw data was processed using various text cleansing steps such asremoving all non-alphanumeric characters, punctuation, and extraneous whitespace, lowercasing and tokenizing. Several popular packages available in the R programming language were employed to carry out the analysis.
Results: From choropleth map, it was found out that most queries during the study period were raised in the eastern districts of Telangana, highest being raised at Mahabubabad (>17000). The average number of queries raised in a month was found to be around 8000 queries per month. The line graph indicated that the queries raised during the second wave of COVID-19 were less. The word cloud indicated that the word 'management' has appeared most number of times in the query text followed by 'weather' and 'paddy'.
Conclusion: From this study, we found that the maximum queries on farming in Telangana are raised in Mahabubabad district.It was also found that the queries on 'nutrient management in paddy' and 'pm kisansamman' were highest. This highlights the need for educating farm tele advisors on the trending queries for efficient delivery of information needed by the farmers.

Keywords:kisan call centre,management, text data, visual analytics, weather

1. INTRODUCTION

The Indian economy and society are anchored on agriculture, which is essential to the growth and survival of the country. Agriculture is the main industry in India for a large percentage of people, which means that industries related to it, such as horticulture, fisheries, and animal husbandry, create a lot of jobs. These businesses give farmers indirect employment opportunities in allied industries such as marketing, transportation, and agro-processing, in addition to direct employment. It is essential to India's food security and a major contributor to the country's GDP.The total cultivated area for food grain production expanded substantially, surging from 97.32 million hectares in 1950-51 to 129.34 million hectares by 2020-21 [1]. Innovation and technology are being embraced by the agriculture industry more and more. Utilizing precision farming, biotechnology, Information and Communication Tools (ICTs) and modern agricultural practices results in sustainable

agriculture, which can both prevent the effects of climate change and maintain the long-term health of the ecosystem.

Agriculture has played an important role in the growth of Telangana's economy.Telangana is India's youngest state, formed on June 2, 2014, and has a rich agricultural heritage. Telangana is divided into three main agro-climatic zones. Northern Telangana Zone is distinguished by black cotton soils that are ideal for crops such as cotton, maize, and soybeans. The Central Telangana Zone is dominated by red sandy soils that sustain crops like as millets, pulses, and oilseed. The Southern Telangana Zone is known for its red soils and ideal growing conditions for paddy, groundnut, and horticultural crops.Agriculture and related activities are a primary source of food and income for over 60% of the state's population [2].

Due to the increased application of ICT tools inagricultural sector, large amount of unstructured dataincluding text data are available. Such unstructuredtext data needs to be analysed for its content using textmining techniques to capture the information hiddenin them [3]. For instance, On January 21, 2004, the Ministry of Agriculture, Government of India, introduced the "Kisan Call Centers" (KCCs) program nationwide to provide the farming community with individualized extension and consulting services in an effort to promote farmer prosperity. This scheme's primary goal is to respond to farmers' phone calls in their native tongue, with responses provided in 22 regional languages. Through KCCs, farmers may better organize their agricultural activities by receiving weather forecasts and alarms in real time. KCCs assist farmers make educated decisions about when and where to sell their goods by giving them access to the most recent information on market prices. A lot of data about the various questions that farmers ask and the answers that Kisan Call Center operators provide are generated. This process has made available a large amount of text data which needs to be analysed to study the patterns in queries raised for better delivery of the farm advisories. An exploratory analysis of such query data in five south Indian states for the year 2017 indicated that >60% of the queries raised were for weather related Information [3]. The Indian government can also address the problems that farmers in various states confront, and it would encourage the development of laws that are advantageous to farmers.

Visual analytics is the method of studying data with visualization,to conduct descriptive analytics to shed light on the nature and dimensionsof data [4].Visual analytics facilitates the effective analysis and understandingof big datasets in real-time. It alsoenables the exploration of unforeseen and hidden patterns to gain insight to makeinformed decisions [5,6].Visual analytics integrates theanalytic capabilities of the computer and the abilities of the humananalyst [7]. Visual Analytics has been successfully employed to study data from various domains such as health [8], research trends [9] and crime [10].

This study employs the visual analytics to study the queries raised in Kisan Call Centres of Telangana thereby making an effort to explore the hidden patterns in farmers' queries. The results from this study are useful to policy makers to bring new schemes or policies which can address farmers' issues.

2. METHODOLOGY

2.1 DATA

The data for this study was gathered from the Kisan Call Centre in Telangana. The monthly district-wise data of queries raised in all 33 districts of Telangana in 3 years (2020, 2021, and 2022) is used for the study.The data comprised information related to the district and

block of the farmer and sector, category, query type, crop, query text and the answer provided by the farm tele advisor. Among them, the query text was chosen as the key criteria for text analysis.The dataset was obtained in raw text format, necessitating multiple stages of data pre-processing before carrying out visual analytics.

2.2DATAPREPROCESSING

Data pre-processing is the process of transforming raw data into a clean and usable format for analysis. It involves tasks such as cleaning to handle missing values and outliers, integrating data from multiple sources, transforming data through normalization and encoding, and selecting relevant features. Effective pre-processing enhances data quality and improves the performance of statistical analysis.According to Gomes et al. textpreprocessingcanboost theaccuracybymorethan20percent insentiment analysisof socialmediadata [11].

The text pre-processing of the raw query data involved several essential steps to prepare it for visual analytics. Initially, in the text cleaning phase, the raw data was thoroughly cleaned by removing all non-alphanumeric characters, punctuation, and extraneous whitespace, ensuring uniformity across the dataset. This step was crucial to eliminate any inconsistencies that could affect the analysis.Next, the lowercasing step was performed where all text data was converted to lowercase. This transformation was necessary to eliminate word duplication caused by case sensitivity, ensuring that words like "Paddy" and "paddy" are treated as the same entity.Following this, the cleaned text was subjected to tokenization. This process involved separating the text (query text in the form of sentences) into individual words or tokens, a necessary step for subsequent text analysis. Tokenization facilitates the handling of text data by breaking it down into manageable pieces.Another critical step was the removal of stop words. Commonly used words that carry little semantic meaning, such as 'and', 'the', and 'is', were deleted from the dataset. This helps to reduce noise and allows the analysis to focus on the most significant words in the query, enhancing the quality of the insights derived.

These pre-processing steps collectively ensured that the text data was in a clean, consistent, and analyzable format, ready for further text analysis. By systematically cleaning, lowercasing, tokenizing, and removing stop words, the dataset was refined to highlight the most relevant and meaningful information.

2.3VISUAL ANALYSIS

Visual analytics is concerned with providing data-driven insights to aid decision making [12] and this approach can be seen in popular web analytics and business analytics tools [13]. The Visual analysis in this study focuses on using various graphical and visual tools to investigate and comprehend data patterns and associations between words. Using a variety of visual representations, we tried to find underlying themes, trends, and linkages that typical data analysis tools may not instantly reveal. The following subsections describe the visual tools used in this study: Thematic map, Line graph, Word cloud, Bigram, and Network diagram.

2.3.1 CHOROPLETH MAP

Choropleth or thematic maps are used to show spatial changes in data across different geographical locations.Thematic maps show spatial distributions [14].Choropleth maps, which display data spatially, are a valuable tool for finding regional patterns and trends that

may influence the entire research.We employed choropleth maps to represent the distribution and intensity of queries raised in different districts of Telangana.

2.3.2 LINE GRAPH

Line graphs are used to illustrate changes in data over time, providing a clear visual picture of patterns and fluctuations. In this study, line graphs were used to examine temporal patterns in the queries raised. By showing data points over a time axis, we were able to identify variances and trends in the total number of queries raised over specific times.

2.3.3 WORD CLOUD

Word clouds, or tag clouds, are visual representations of text data that show the frequency of words based on their relative size. In this study, word clouds were created to analyze textual data and identify relevant themes and keywords.Bielenberg [15] proposed a circular layout to display word clouds where important words are placed closer to the center.The generated word clouds will give a clear and fast visual summary of the most important terms in the text corpus.Shaw [16] proposed to display tag clouds using a graph layout whose nodes represent tags and edges indicate the relations between tags.Cui et al [13] proposed a visualization method that couples a trend chart with word clouds to visually illustrate the content evolution.

2.3.4 BIGRAM

Bigrams, or two-word combinations, are used to uncover common pairs of words within a text corpus, revealing contextual relationships between terms. In this study, bigrams were extracted from the query data to identify frequently co-occurring word pairs.The bigram analysis provided insights into the syntactic and semantic structures of the query text, highlighting important associations and patterns that single-word analysis might miss.Tan et al found positive results utilizing bigrams on the Reuters and Yahoo! Science datasets [18].

2.3.4 NETWORK DIAGRAM

Network diagrams, also known as graph representations, are used to visually portray the links and interactions between elements in a dataset. It is a powerful tool in data visualization and analysis that allows for the representation of complex relationships between entities. By mapping these keywords as nodes and their co-occurrences as edges, the network diagram visually illustrates how terms are related and clustered within the data. This method provides a clear and comprehensive understanding of the underlying structure and patterns within the query data, enabling the identification of key themes and the relationships between different concepts.

All the analysis including data pre-processing was carried out in R Programming Language [19] by using various packages such as tidytext, ggplot2,wordcloud, igraph, ggraph etc.

3. RESULTS AND DISCUSSION

3.1 SPATIAL DISTRIBUTION OF QUERIES

The choropleth map in Figure 1 provides the spatial distribution of queries raised by farmers in Telangana during the study period. Figure indicates that Eastern districts whereas Central and Northern districts have registered lower number of queries.During the study period 265200 queries were raised in Telangana. Highest number of queries were raised in

Mahabubabad (17669 queries) followed by Khammam, Warangal rural, Kamareddy and Nalgonda districts. The least number of queries were raised from MedchalMalkajgiri (854), Hyderabad, KumaramBheemAsifabad, RajannaSircilla and YadadriBhuvanagiri districts.

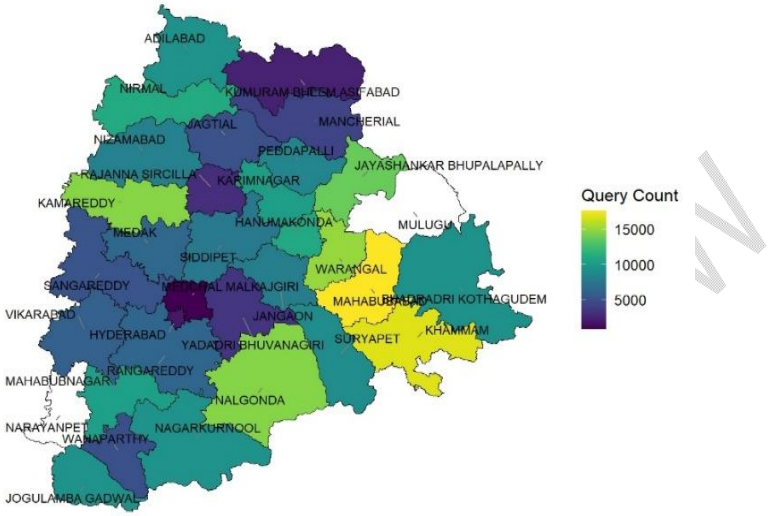

Figure 1 Spatial Distribution of Queries raised in Telangana during the study period.

3.2 TREND IN TOTAL MONTHLY QUERIES

Line graph in Figure 2 shows the trend in total number of queries raised in Telangana during the study period. From the figure it is clearly understood that the number of queries raised in each month of each year varied throughout the study period. The highest number of queries were raised in the month of January, 2021and the lowest number of queries were raised in June, 2021.The less number of queries during the March to July, 2021 may be attributed to the second wave of COVID-19. On an average, 8036 queries were raised in a month in entire Telangana.

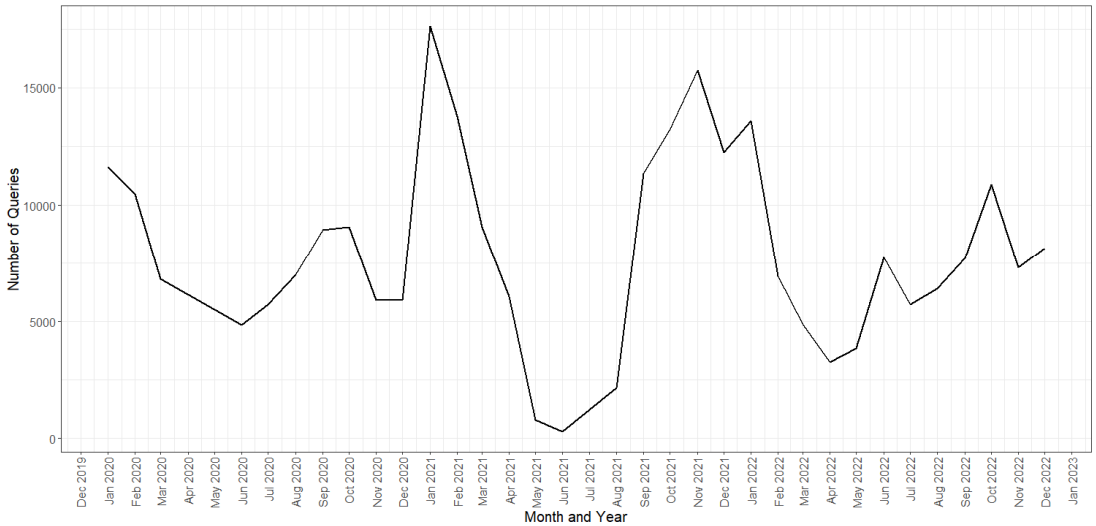

Fig2Trend in number of queries raised

3.3 FREQUENT WORDS USED IN QUERIES

By using Word Cloud, we can find out the words appearing most number of times in the data. A total of 7081 distinct words were identified in the text corpus excluding the stop words. Figure 3 showssome of the words that are present in the text corpus. Figure 4 shows the top 50 words that are frequently used by the farmers, where size of the word corresponds to the number of times it is appearing in the text corpus.

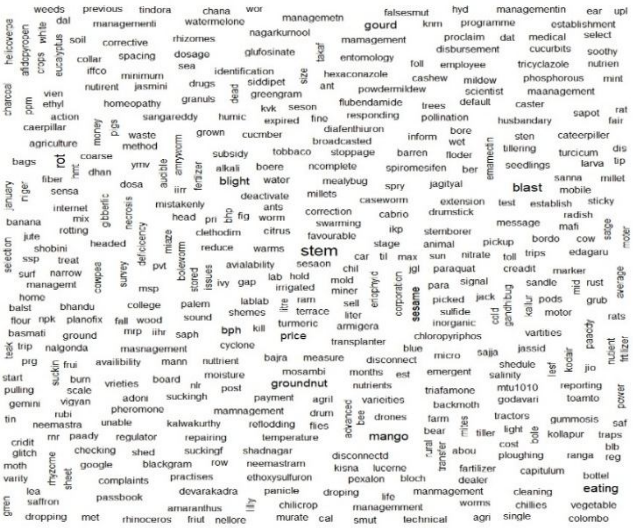

Fig3Glimpse of the words appearing in the farmers' queries

Fig4 Top 50 words appearing in the farmers' queries

From Figure 4, it is clear that the word 'management' is appearing most number of times in the query text corpus suggesting that most of the queries were related to management practices. 'weather' is the other word which appeared distant second. Among the crops, 'paddy', 'cotton' and 'chilli' are appearing more number of times which is obvious since these crops are vastly cultivated in the study area. Other frequent words are 'leaf', 'sucking' and 'nutrient' indicating that the queries regarding nutrient management and sucking pest management are quite common in the queries. However, it is not that clear from the word cloud that which management query is raised more number of times. This issue can be

addressed by using bigrams. Figure 5 provides the frequencies of top 20 bigrams from the text corpus.

Among all the bigrams, ‘nutrient management’ appeared most number of times (>23000) indicating that most number of queries are raised on nutrient management in different crops. Other queries related to management were on ‘pest management’, ‘weed management’, ‘borer management’, ‘fertilizer management’ which were raised more than 8000 times in the study period. Apart from bigrams related to management practices, the bigram ‘pm kisan’ appeared more number of time indicating that there were more queries on pm kisansammannidhiyojana.

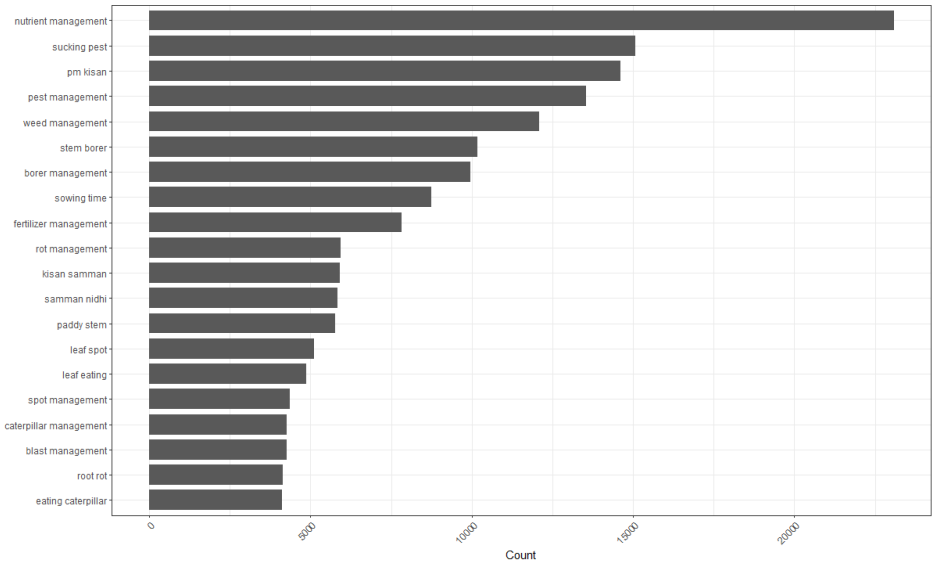

Fig5 Top 20 bigrams appearing in the farmers’ queries

3.4 NETWORK OF QUERY WORDS

Network Diagrams are helpful to visually portray the inter-connections or co-occurrence of words.The network of top 50 words which commonly occur together (highest correlation between words) was plotted to investigate the relationship between those words which is presented in Figure 6. In the network diagram, the dot indicates the preceding text and the arrow indicates succeeding text.The darkness of the arrow indicates the number of times the two words have co-occurred. Darker the arrow, more number of times the words have co-occurred.

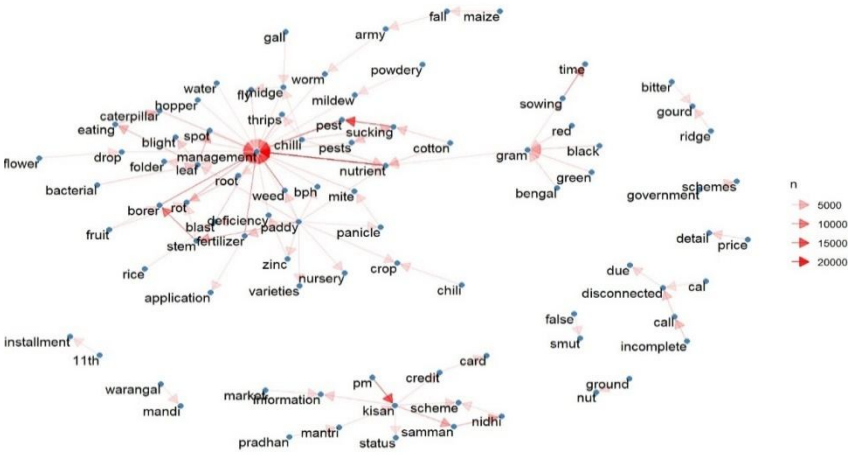

Fig6 Network diagram of 50 most frequent words

In the above Fig6, we can see the interconnection between the query words that clearly depicts how they related to each other. Here the arrow direction shows us how they inter related to each other. If we took the word maize from the above Fig6, we can observe that the arrow started from the maize and ended at the word fall and from there again the arrow started and ended at the word army and next from there it ended at the word worm which means the maize crop is infested with fall army worm. Likewise, with the help of above figure we can understand the interconnections between the query words. In that figure based upon the brightness of the arrow we can depict the frequency of those words.

4. CONCLUSION

The study investigated the trends and patterns that are hidden in the text corpusof farmers' queries using visual analytics techniques.Using choropleth map we found that eastern districts of Telangana registered more number of queries whereas northern districts registered less queries. Among 33 districts, highest number of queries were raised from Mahabubabad whereas Medchal-Malkajgiri registered least. The line graph indicated that the highest number of queries were registered in the month of January 2021 and least number of queries during the second wave of COVI-19. From the word cloud, we found that the top three words appearing most frequently are 'management', 'paddy' and 'weather'.The visual analytics of bigrams revealed that queries on nutrient management were highest among all management related queries. A good number of queries were also raised about the PM Kisan Samman scheme. The network diagram also reiterated the findings. Finally, we can conclude that, among all the queries, queries related to 'nutrient management in paddy' and 'weather information' were the highest.

REFERENCES

1. Jeyanthi T, KannanA. Analysis of growth and cropping pattern changes of Indian agriculture since independence: critical account on the sustainable developmentperspectives. International Journal of Social Science and Economic Research. 2024;9(3):650-674.
2. Guntukula R. Agricultural Performance of Telangana State: An Analysis. Asian Journal of Research in Social Sciences and Humanities. 2017;7(8):169-81.
3. Yashavanth BS, Sreekanth PD. Topic Modelling for Discovering Themes in the Queries Raised at Farmers' Call Center. Journal of the Indian Society of Agricultural Statistics. 2022;76(1):7-16.
4. Börner K, Bueckle A, &Ginda M. Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. Proceedings of the National Academy of Sciences, 2019;*116*(6):1857-1864.
5. Singh D, Singh B. Investigating the impact of data normalization on classification performance. Applied Soft Computing. 2020;97:105524.
6. Cao N, Koch S, Gotz D. ACM TIST Special Issue on Visual Analytics. ACM Transactions on Intelligent Systems and Technology. 2018;10:1–4.
7. Keim D, Kohlhammer J, Ellis G, Mansmann F. Mastering the information age solving problems with visual analytics. Eurographics Association. 2010.

8. Raghupathi W, Raghupathi V, &Saharia A. Analyzing health data breaches: a visual analytics approach. AppliedMath. 2023;*3*(1):175-199.
9. Chen, H., Sun, D., Yang, Y., Looi, C. K., & Jia, F. Detecting and visualizing research trends of blended learning: A bibliometric analysis of studies from 2013-2022. Eurasia Journal of Mathematics, Science and Technology Education. 2023;19(10): em2336.
10. Lettieri N, Guarino A, Malandrino D,& Zaccagnino R. The sight of Justice. Visual knowledge mining, legal data and computational crime analysis. In *2021 25th International Conference Information Visualisation (IV)* 2021:267-272. IEEE.
11. Gomes FB, Adán-Coello JM, Kintschner FE. Studying the effects of text preprocessing and ensemble methods on sentiment analysis of Brazilian Portuguese Tweets. In International Conference on Statistical Language and Speech Processing. Springer. 2018;167–177.
12. Thomas JJ, Cook KA. A visual analytics agenda. IEEE computer graphics and applications. 2006;26(1):10-13.
13. Chaudhuri S, Dayal U, Narasayya V. An overview of business intelligence technology. Communications of the ACM. 2011;54(8):88-98.
14. Tennekes M. tmap: Thematic Maps in R. Journal of Statistical Software. 2018;84:1-39.
15. Bielenberg K. Groups in social software: Utilizing tagging to integrate individual contexts for social navigation. Master's thesis, Universitat Bremen. 2005.
16. Shaw B. Utilizing Folksonomy: Similarity Metadata from the Del. icio. us System CS 6124 Project. 2005.
17. Cui W, Wu Y, Liu S, Wei F, Zhou MX, Qu H. Context preserving dynamic word cloud visualization. In2010 IEEE Pacific Visualization Symposium (PacificVis). 2010;121-128.
18. Tan CM, Wang YF, Lee CD. The use of bigrams to enhance text categorization. Information processing & management. 2002;38(4):529-546.
19. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, 2024, Vienna, Austria.