CS685 Group 01: KCC Query Analysis

Final Presentation

Group Members:

Aniket Sanghi 170110 sanghi@iitk.ac.in

Neil Rajiv Shirude 170429 neilrs@iitk.ac.in

Paramveer Raol 170459 paramvir@iitk.ac.in

Sarthak Singhal 170635 ssinghal@iitk.ac.in

Aman Kumar Thakur 170083 amankt@iitk.ac.in

Problem Statement

Analyzing Kisan Call Centre Query Dataset and working towards-

- Developing an **FAQ generator** and analyzing important FAQs
- Analyzing queries related to **government schemes** to investigate implementation issues
- Analyzing major **crop protection** problems in popular crops
- Finding major **weather-related issues** plaguing farmers in various states
- Implementing **classification algorithms** that assist KCC executives in documentation

Data Scraping and Pre-processing

Data Scraping [Automated Script]

- Iterate through webpages
- Ask for csv file for each dataset
- Fill the Authentication form
- Submit download request

Pre-processing

- Data Narrowing
- Inter-dataset duplication
- Intra-dataset duplication
- Outlier detection and removal
- Data formatting

FAQ Generation

Approach

- Find Unique Queries
- Generate Embeddings
- Cluster Similar Sentences Together
- Generate FAQs

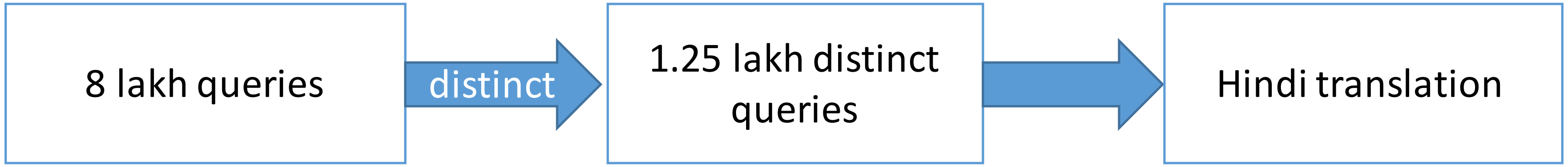
Major Cultivators, Major Problems

- Wheat – UP
  - weather information, weed control
- Paddy – West Bengal
  - brown leaf spot, stem borer, sheath rot
- Cotton – Gujarat
  - fertilizer management, sucking pests problem

Government Schemes Analysis

- very large number of hinglish queries
- scheme names had too many variants
- tagging quite difficult

Processing Pipeline:

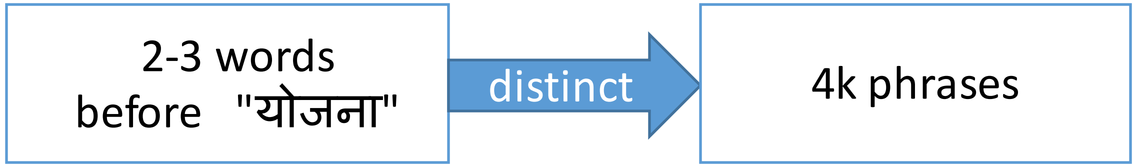


8 lakh queries

distinct

1.25 lakh distinct queries

Hindi translation

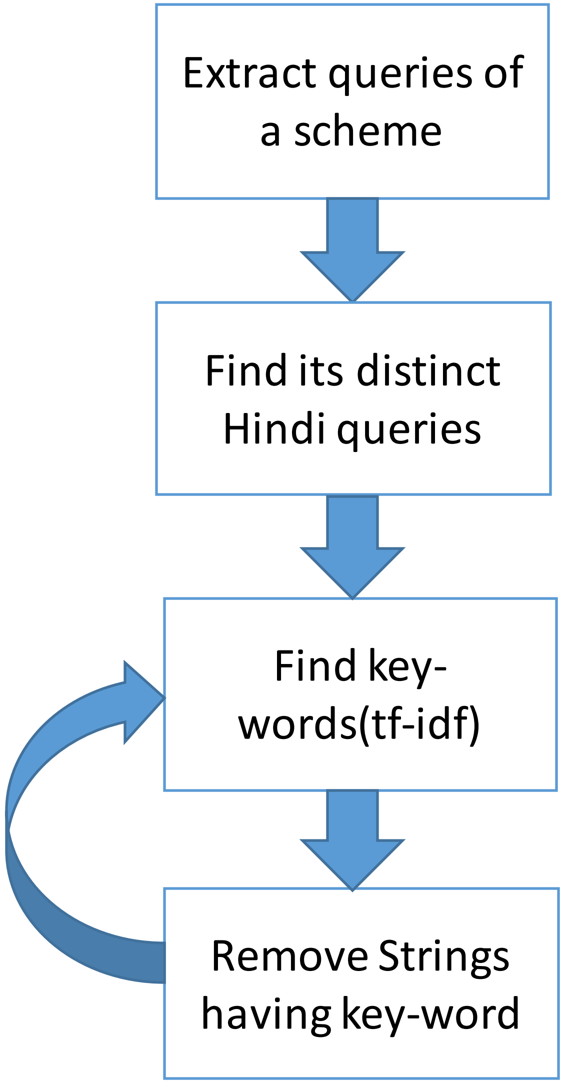


2-3 words before "योजना"

distinct

4k phrases

Government Schemes Analysis



Extract queries of a scheme

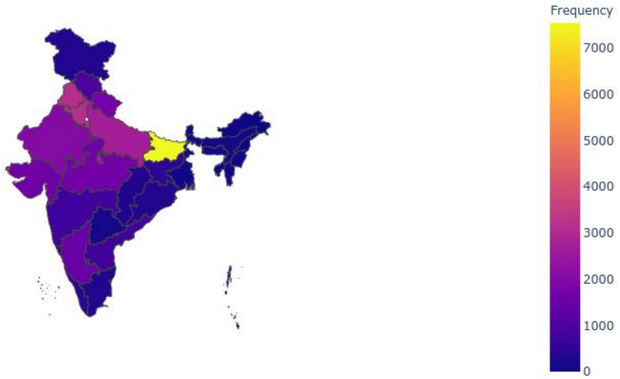
Find its distinct Hindi queries

Find key-words(tf-idf)

Remove Strings having key-word

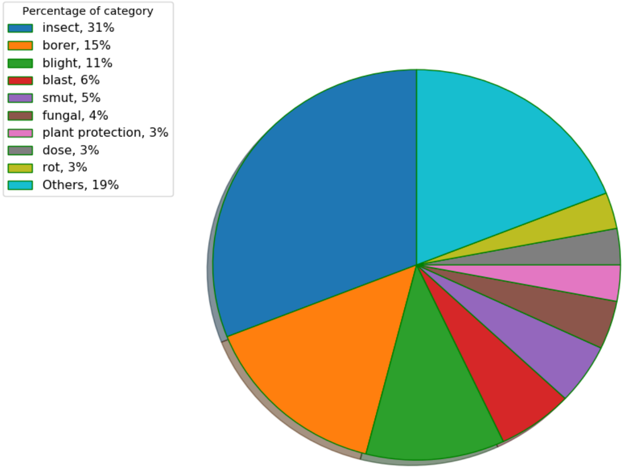
- Cluster the key-words found manually(<100)
- **Major categories**: registration, finance, general information etc.

- **Results** : 1. state-wise frequency of a scheme.
  2. state-wise frequency of query per-holding.
  3. Classification of queries of a scheme based on the key-words and the corresponding count.



Distribution of queries across India

Plant Protection Analysis



Major crop protection categories of Paddy

General analysis

- ➢ Dominance of Paddy and Cotton
- ➢ Major contributor: Uttar Pradesh

Per crop disease analysis

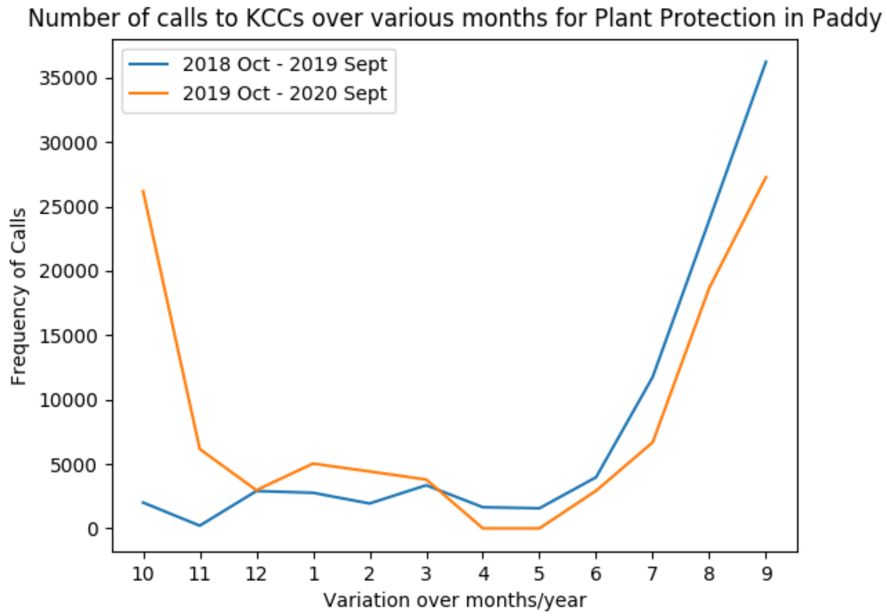
- ➢ Major proportion of insects and fungus
- ➢ Less fraction of queries related to fertilizers

Per crop state-wise analysis

- ➢ Major contributor for Paddy: Uttar Pradesh
- ➢ Major contributor for Cotton: Maharashtra

Per crop month-wise analysis

- ➢ Queries only during crop's season



Weather Analysis

Data

- Large number of queries with spelling errors
- 99 percent general queries

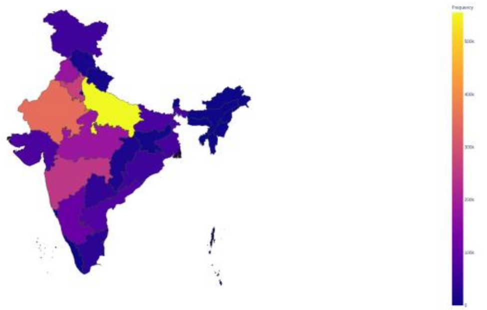
General Analysis

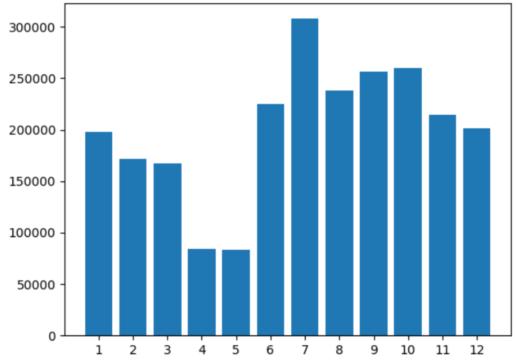
- Major contributor: UP
- Low queries: North east and east coast states
- Queries mostly during monsoon and winter
- Minimum correlation between monthly data and yearly data across state is 0.9

Specific Analysis

- Mostly queries: Rainfall
- Low queries : Cyclones, IMD

Total Queries across India





Distribution of Monthly Queries

Multi-Class Classification

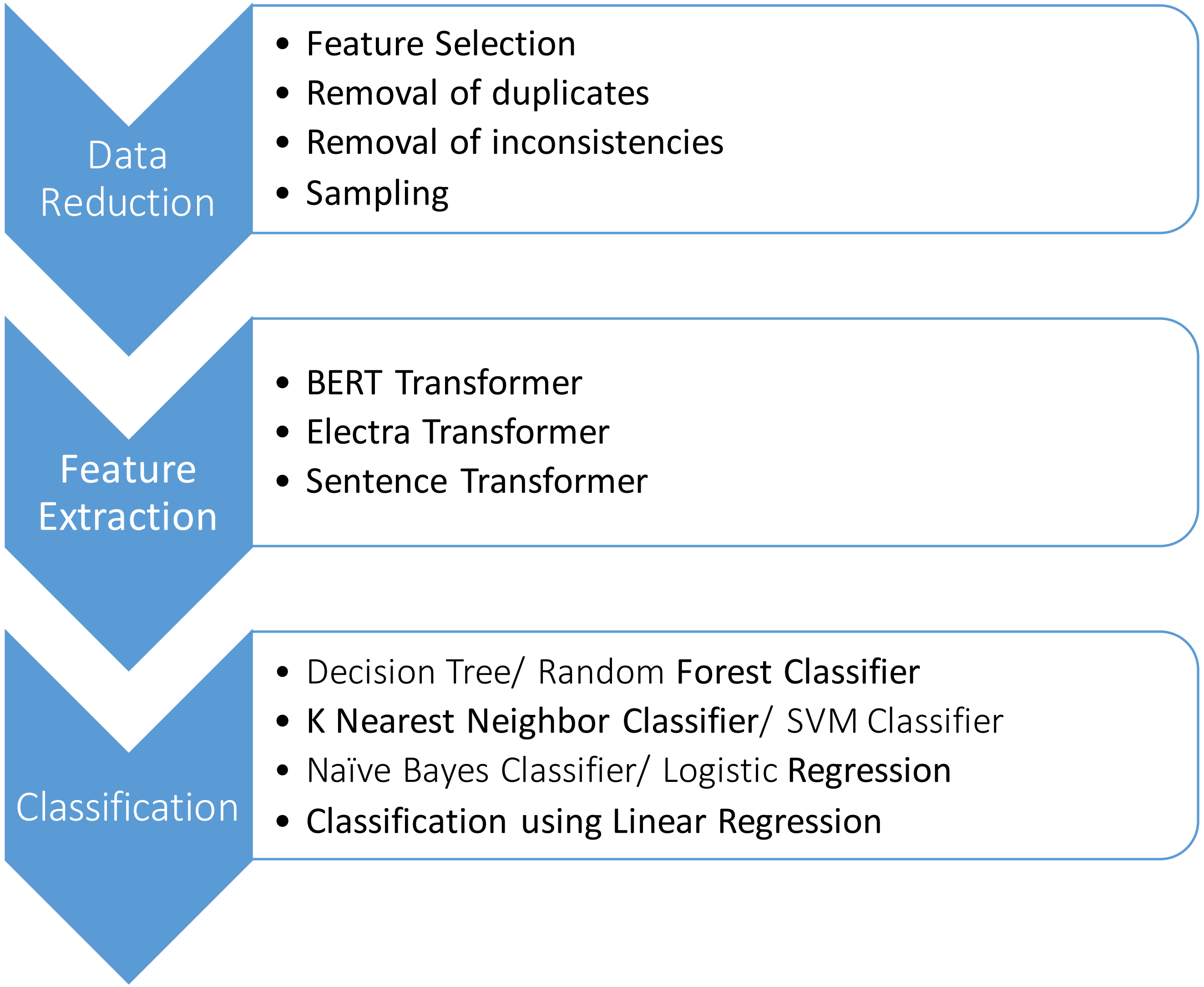
Tasks:

- Sector Classification
- Query Type Classification

Results:

- Sector Classification-
  - Best Accuracy: **0.89**
  - Logistic Regression with Sentence Transformer Embeddings
- Query Type Classification-
  - Best Accuracy: **0.67**
  - Linear Regression Technique with Sentence Transformer Embeddings

Classification Pipeline:



Data Reduction

- Feature Selection
- Removal of duplicates
- Removal of inconsistencies
- Sampling

Feature Extraction

- BERT Transformer
- Electra Transformer
- Sentence Transformer

Classification

- Decision Tree/ Random Forest Classifier
- K Nearest Neighbor Classifier/ SVM Classifier
- Naïve Bayes Classifier/ Logistic Regression
- Classification using Linear Regression

Thank You