# CS-613

Natural Language Processing

# AgriBot: Agriculture-Specific Question Answer System

Naman Jain, Pranjali Jain, Pratik Kayal, Sahit PJ, Soham Pachpande
Discipline of Computer Science, IIT Gandhinagar

**Keywords :** Entity Extraction, Web Scraping, Sentence Embeddings, Text Similarity

## Introduction

- India is an agro-based economy and proper information is key to optimal agricultural output. In this project we build a agricultural chatbot.
- According to our analysis, about 1.36 million calls were made to Kisan call centers in 2017 and increased to about 1.72 million calls in 2018. This is a **21% increase in calls** from 2017 to 2018. 92% percent of calls for state of Maharashtra were redundant in the year 2017. Also, for the entire country, only **5% unique** new queries were made in 2018 in comparison to 2017.
- Using NLP techniques, we design a system which would exploit the data reserve and redundancy to create a chatbot to cater maximum agricultural queries without human intervention.

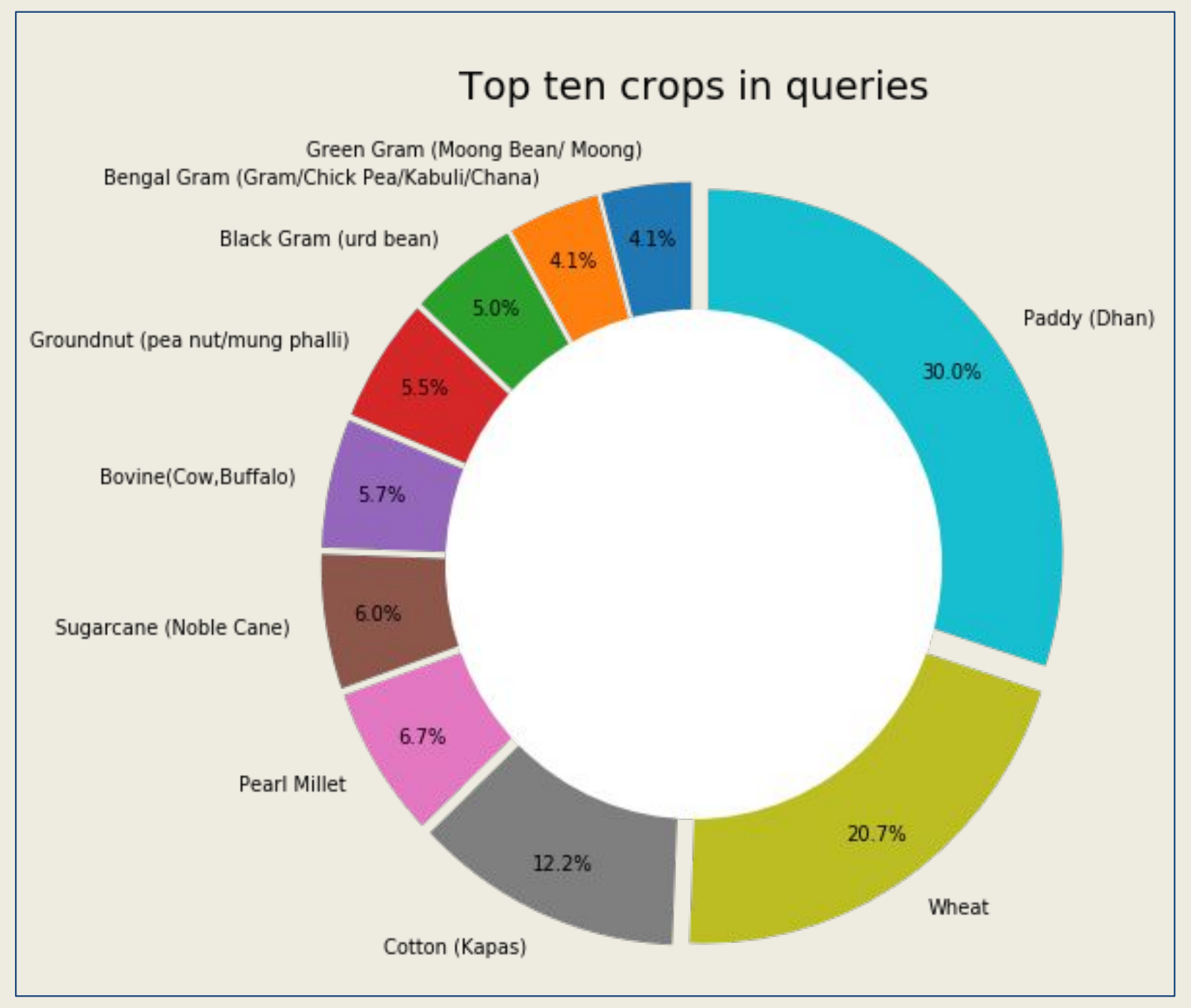
## Data

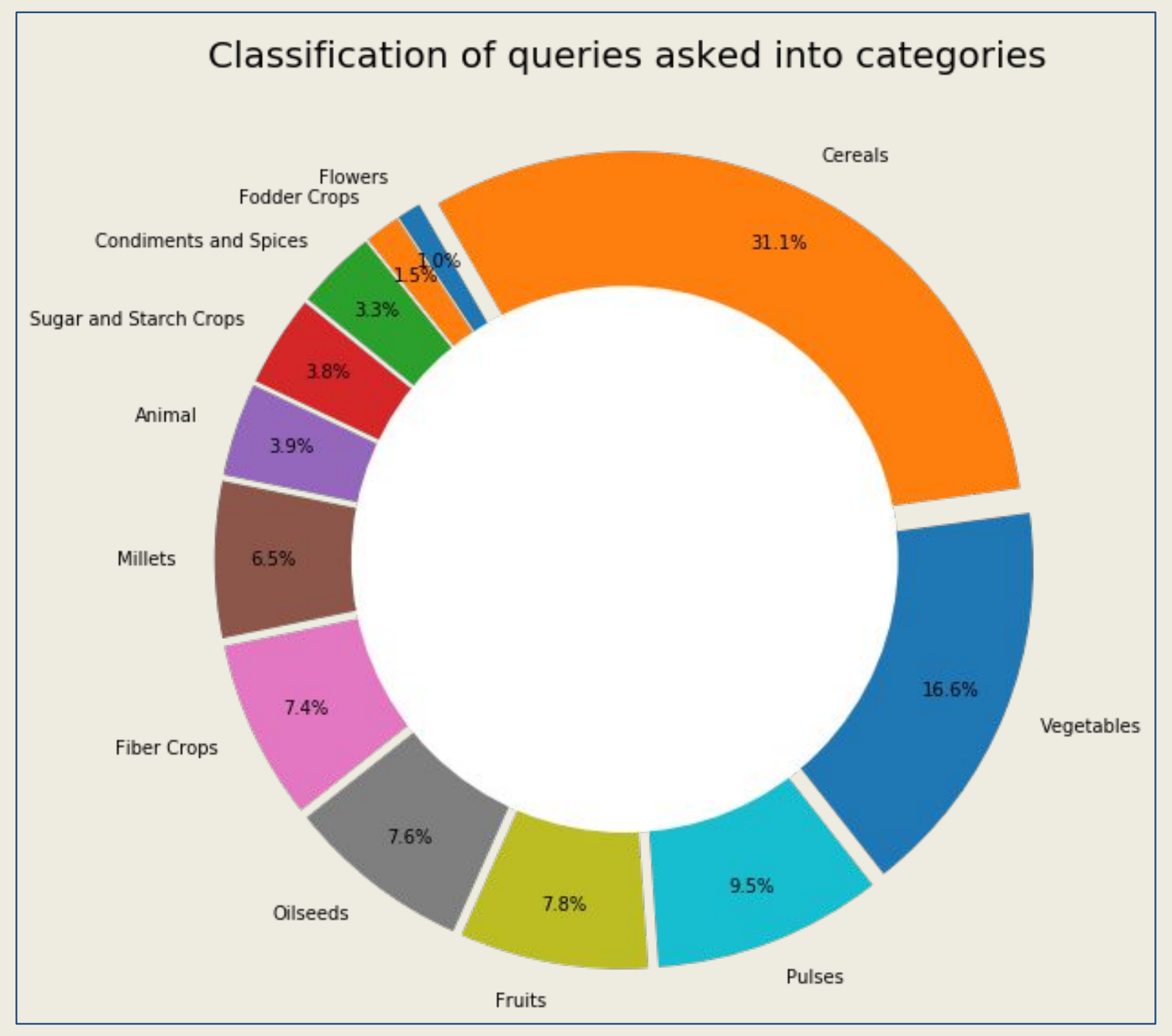
### Dataset

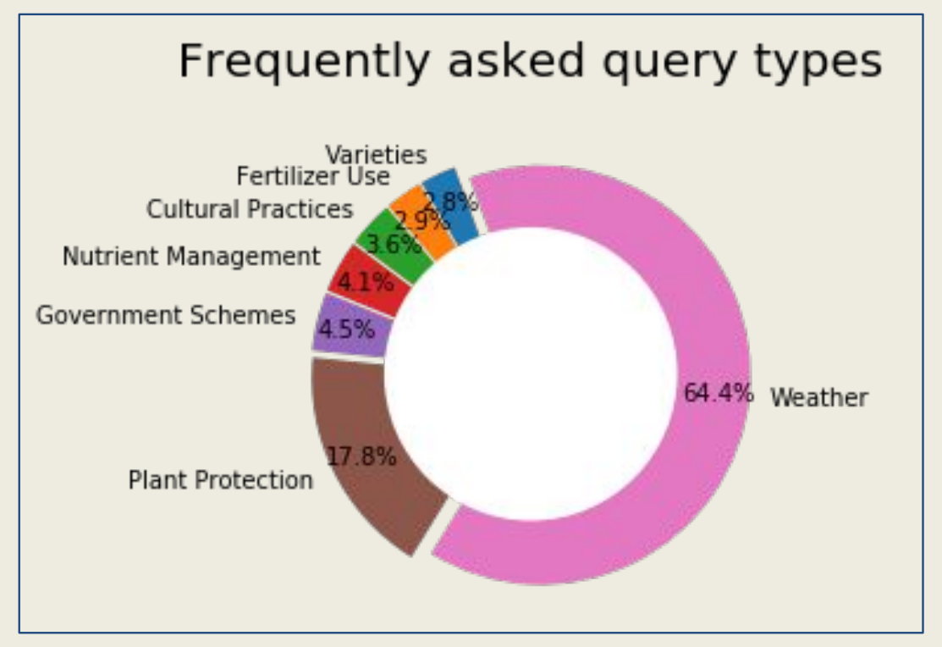
| | |
|---|---|
| Total Number of Pairs | 99,568,354 |
| Year 2017 | 13,613,465 |
| Year 2018 (Till October) | 17,265,241 |

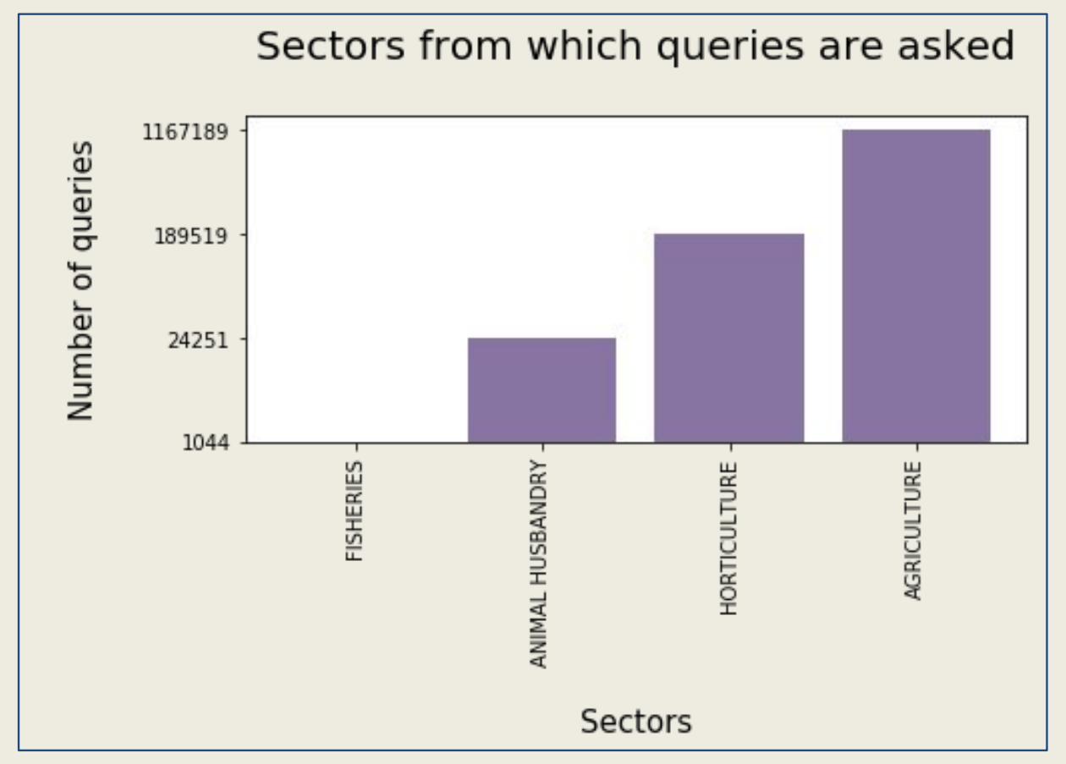
### Most Common Questions

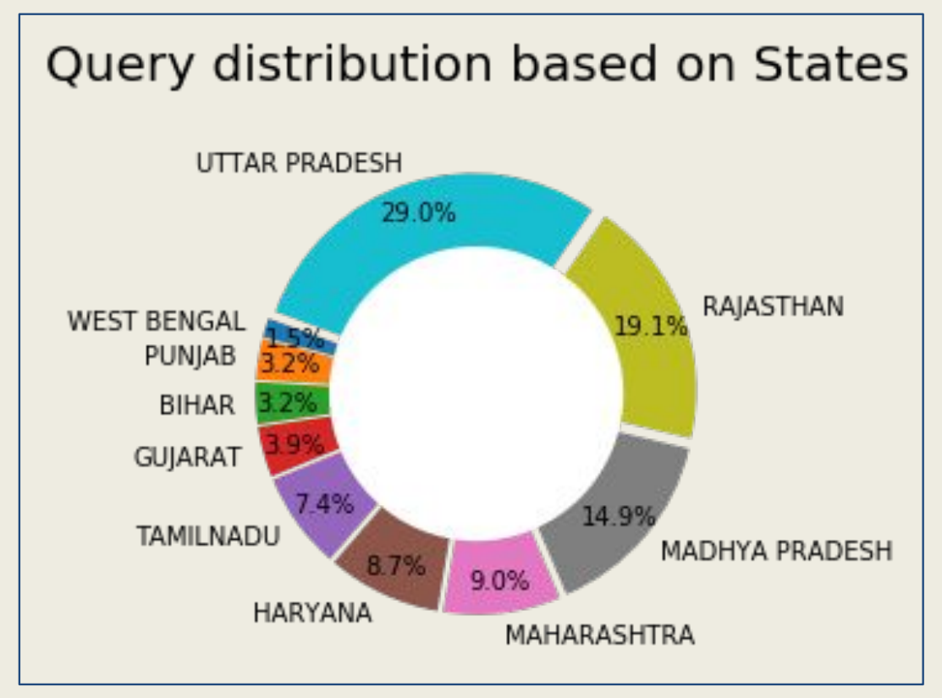
| | |
|---|---|
| What is the weather? | 79.4% |
| How to control of zinc deficiency in wheat? | 2.3% |
| Information regarding macro nutrient management? | 1.76% |

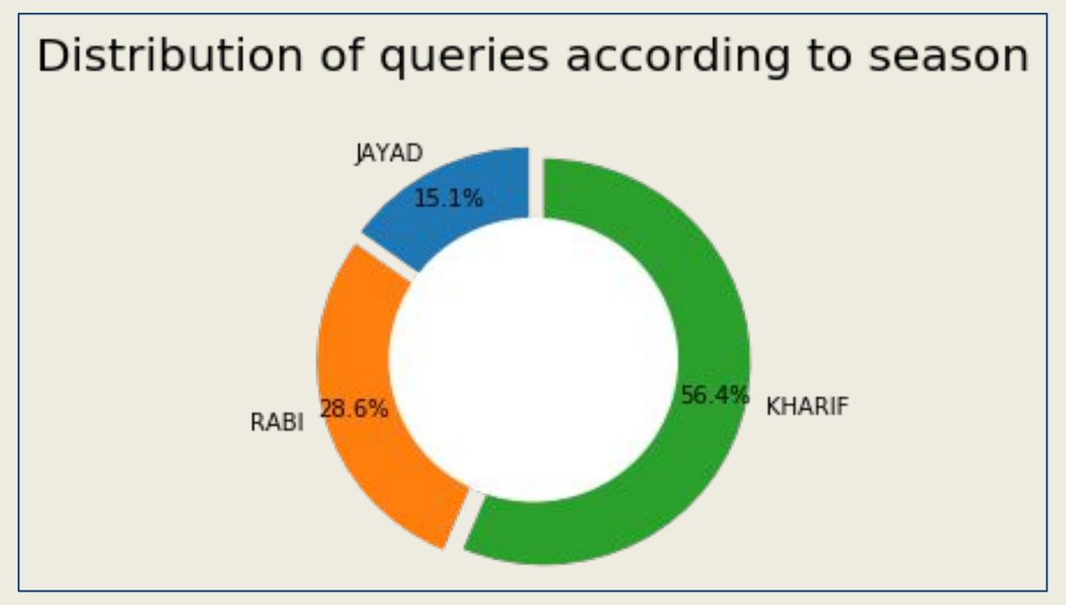

Top ten crops in queries


Classification of queries asked into categories

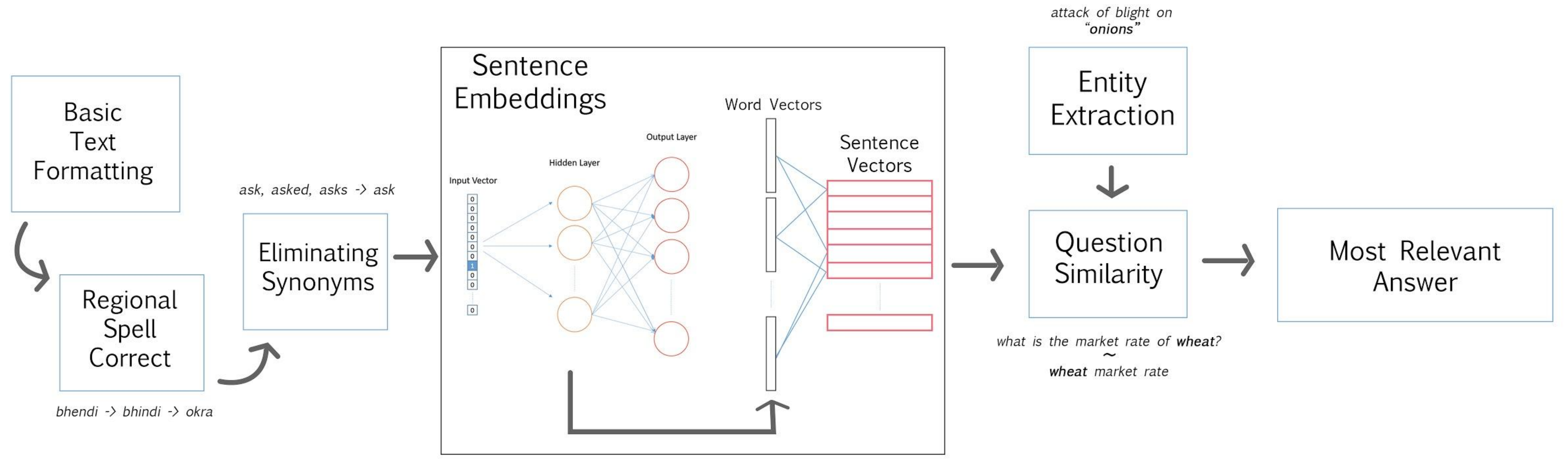

Frequently asked query types


Sectors from which queries are asked


Query distribution based on States


Distribution of queries according to season

## Proposed Approach



Basic Text Formatting

Regional Spell Correct

*bhendi -> bhindi -> okra*

Eliminating Synonyms

*ask, asked, asks -> ask*

Sentence Embeddings

Entity Extraction

*attack of blight on "onions"*

Question Similarity

*what is the market rate of wheat?*
~
***wheat** market rate*

Most Relevant Answer

## Challenges

- **Data Collection** - non-availability of any API/ easy download option to fetch all data.
- **Format of Data** - The lack of consistency in the format of the questions and answers. There were many redundant words and spelling errors.

Q - caterpiler on grem damage?
A - spray quinolphos 30 ml/15 1 water.

- **Preprocessing for various Languages** - given the number of languages in which we had data, we had to figure out a way of processing it.
- **Analysing Truth Value -** since our problem involves finding a relative truth w.r.t to the question, finding a suitable metric was a major task

## Metric

- **Modified Lesk Score for Sentence Similarity:**
  - Use glosses (gloss bag) of all senses of all words in the two sentences to find similarity between two sentences.

$$\text{LeskScore} = \frac{\text{count(gloss(known)} \cap \text{gloss(predicted))}}{\text{count(gloss(known))} + 1}$$

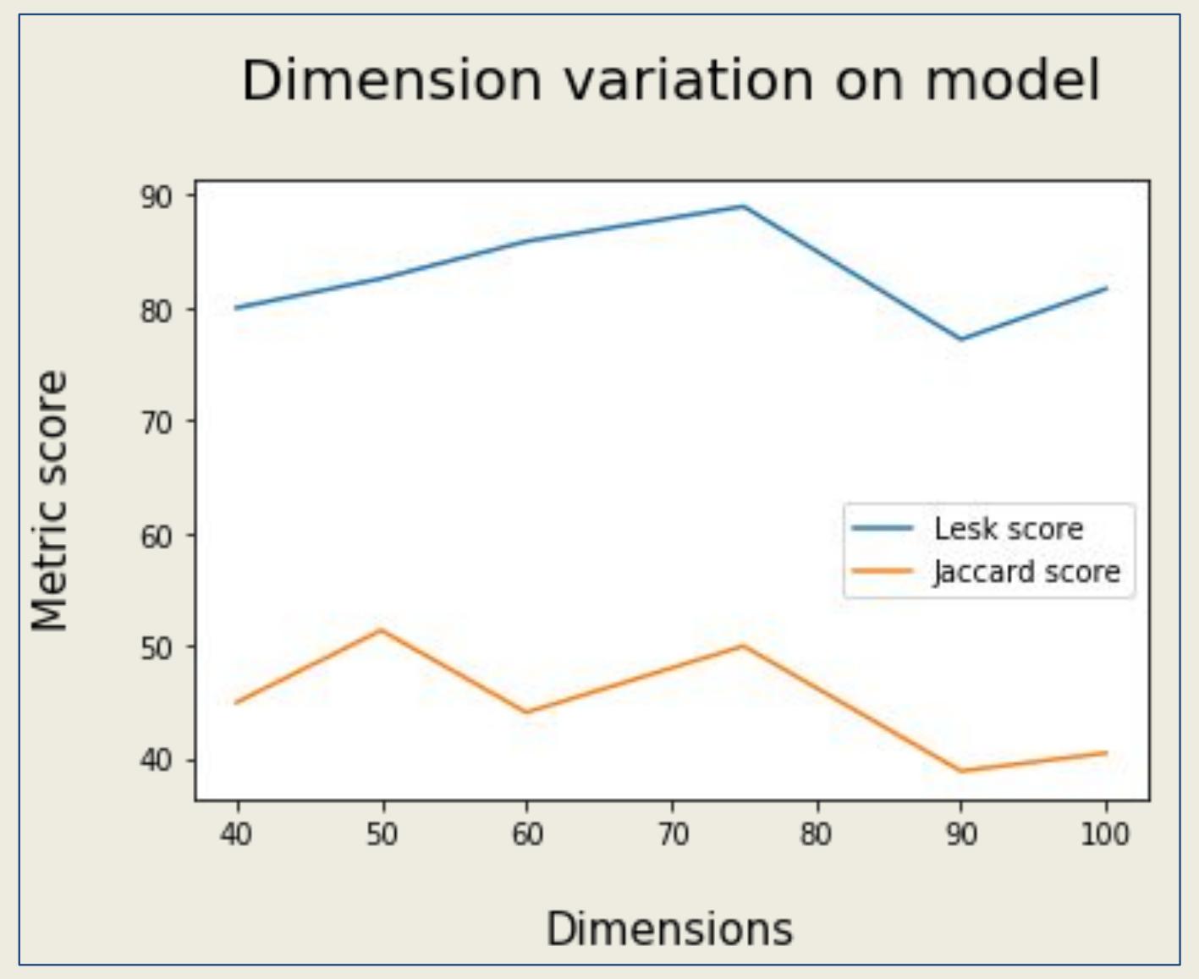
- **Modified Jaccard Similarity**
  - Use words themselves to compute the similarity

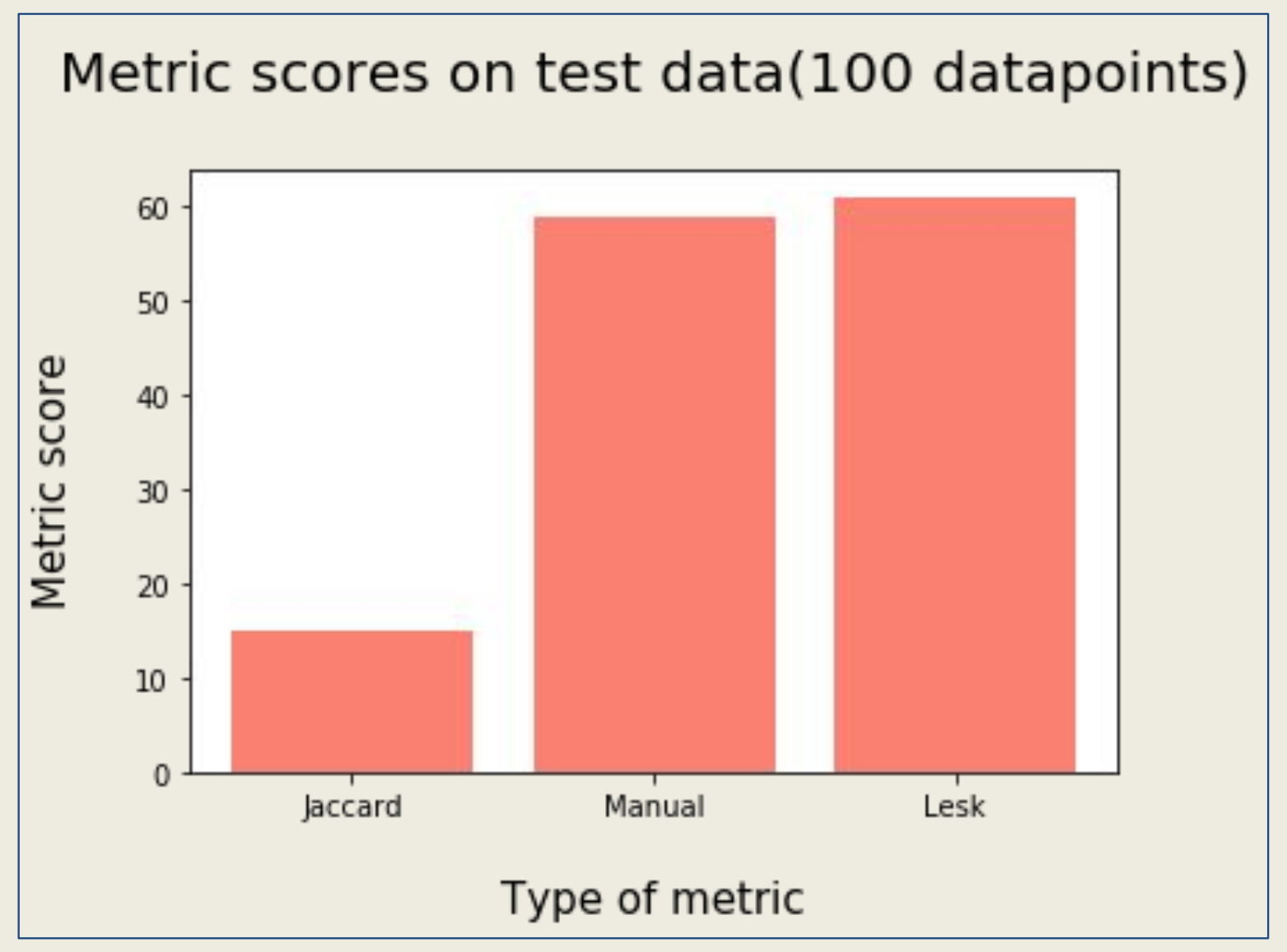
$$\text{Jaccard} = \frac{\text{count(knownSent} \cap \text{predictedSent)}}{\text{count(knownSent)} + 1}$$

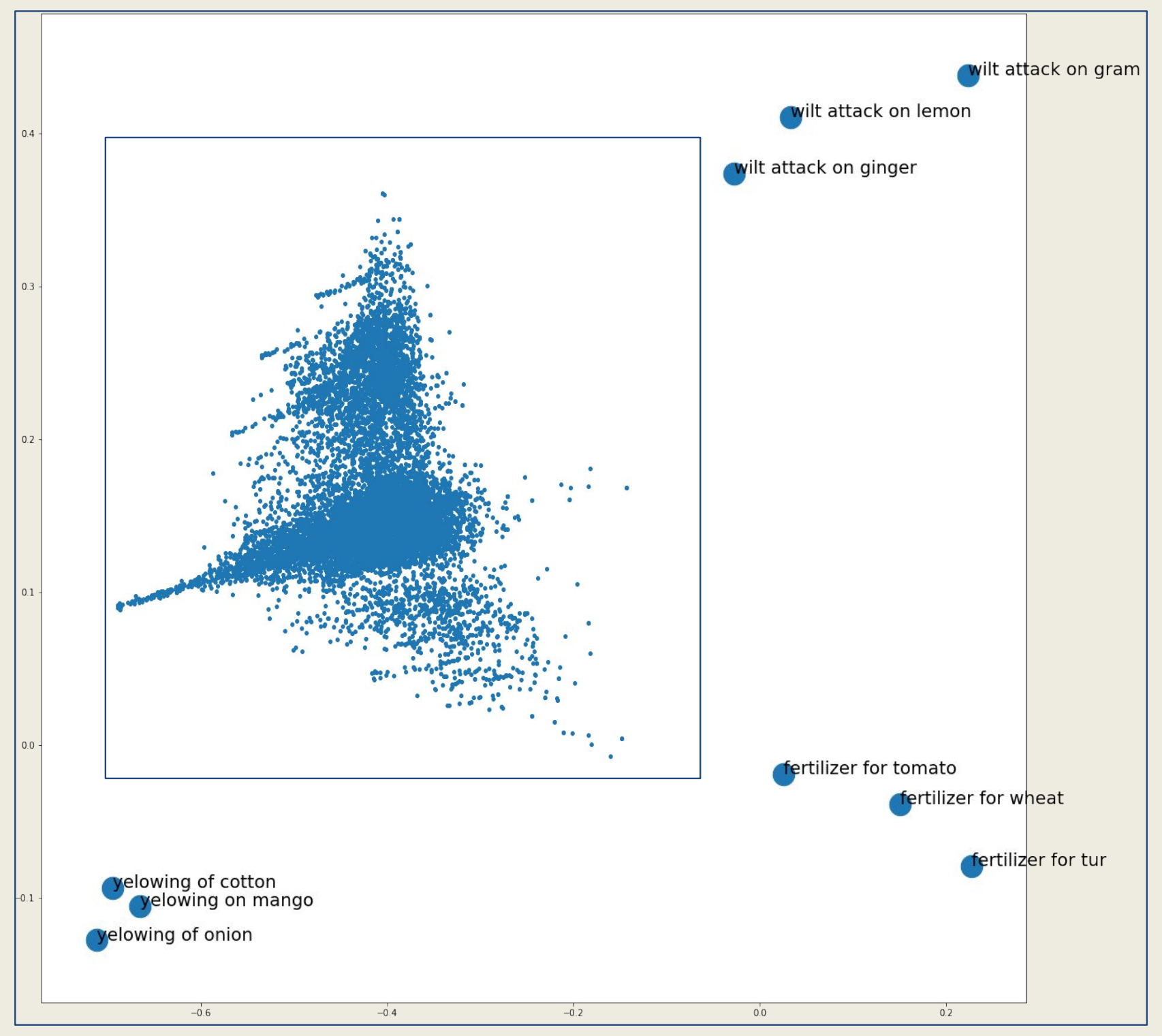
- The known entity would be the input or test data question. We divide by the number of terms in this known entity in order to normalise scores as the the predicted questions/answers are of variable lengths which would skew the scores.
- **Correlation with ground truth and Threshold:**
  - We manually labeled some 100 test data and found out lesk scores for test question and predicted question
  - Using the ground truth, we devised a threshold for Lesk score as well as Jaccard score, above which the sentences would be considered similar
  - We found Lesk Similarity score to perform better than Jaccard Similarity

## Evaluation Results

| | Jaccard | Lesk |
|---|---|---|
| Top - 1 | 64 % | 86 % |
| Top - 3 | 69 % | 89 % |
| Top - 5 | 70 % | 92 % |


Dimension variation on model


Metric scores on test data(100 datapoints)



*Q - What is the market rate of wheat?*
*sQ - wheat market rate*

*A - wheat market rate--1800---2200 rups pq*

*Q - attack of blight on tomato*
*sQ - attack of blight on tomato*

*A - spray saaf 20 gm / 15 litter of water for controlling blight on tomato*

*Q - What is the fertiliser for grape?*
*sQ - fertiliser for grape*

*A - drip madhun 00:00:50 1 kilo /ekar dyave*

## Future Work

- Improve the technique for answer ranking
- Implement multilingual support in the chatbot
- Entity extraction from answers for information generation

## Conclusion

- We think this system can benefit both the farmers and the Call Center employees by providing a faster and simpler interface to communicate 24x7 throughout the year. This work will help the workforce of Kisan Call Center as redundant queries would not have be answered by them. At the same time more unique queries can be added to the database.
- Through this work, we were also be able to survey the agricultural trends and problems across India. We believe this information helps better understand the farmer's mindset and gives an insight into the problems and opportunities in Agricultural sector.

## Reference

- Arora, Sanjeev, Yingyu Liang, and Tengyu Ma. "A simple but tough-to-beat baseline for sentence embeddings." (2016).
- Rong, Xin. "word2vec parameter learning explained." (2014).
- Banerjee, S., & Pedersen, T. (2002, February). "An adapted Lesk algorithm for word sense disambiguation using WordNet "
- Bikel, Daniel M., Richard Schwartz, and Ralph M. Weischedel. "An algorithm that learns what's in a name." (1999)