

# Transfer Learning for International Crisis Response

Challenges Track 28.01.2020

A red square logo containing the letters 'A M' in the top row and 'L D' in the bottom row, all in white capital letters.

A M  
L D

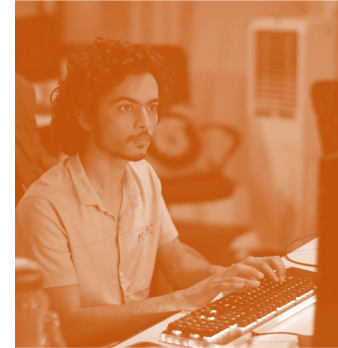
**Who are we**



**Navid Rekabsaz**  
Assistant Professor,  
Johannes Kepler  
University Linz



**Wilhelmina Welsch**  
Head of information  
management &  
innovation, Joint IDP  
Profiling Services  
(JIPS)



**Bibek Pandey**  
NLP Engineer  
DEEP



**Ewan Oglethorpe**  
Executive Director,  
Data Friendly Space



	A	B	C	D	F	G	H	I	J	K	L	M	
1	Severity	Reliability	ID	Publication date	Country	Region	Sector/Issues	Group	Type of source	Source	Title of document	Problem identified	Response
33	3	1	32	28-Feb	Libya	Benghazi	Food Security	Residents	Media	IRIN	food and medical supplies in Libya's rebel-held east.	Bread, vegetables. Prices have gone up by 75 percent.	
34	4	1	33	28-Feb	Libya	Eastern	Food Security	Residents	Media	IRIN	Unrest is disrupting food and medical supplies in Libya's rebel-held east.	Rebel-held eastern Libya will start to experience serious food and medical shortages within three weeks, a public health volunteer said on Monday.	
35	4	1	34	28-Feb	Libya	Benghazi	Health	Residents	RC/RC	British Red Cross		We hear that surgeons and orthopaedic specialists are needed in Benghazi's hospitals, as well as medicine for patients suffering from chronic illnesses. Our initial assessment is that there is no urgent need for food supplies. It's difficult to know, however, what the needs are outside the city."	The Egyptian R countries' borde crossing from L
36	0	2	35	28-Feb	Libya	Benghazi	Health	Residents	NGO	Voice of America			France says it is
37	1	1	36	28-Feb	Libya	Benghazi	Food Security	Residents	RC/RC	British Red Cross		Our initial assessment is that there is no urgent need for food supplies. It's difficult to know, however, what the needs are outside the city."	
38	0	1	36	28-Feb	Libya	Benghazi	Food Security	Residents	RC/RC	British Red Cross			The Egyptian R countries' borde crossing from L
39	4	1	37	28-Feb	Libya	All	Food Security	Residents	Government	USDOS		need to keep medical pipelines well stocked and intact, we are also concerned that the ongoing violence may disrupt distribution networks and lead to food shortages. the ongoing violence may disrupt distribution networks and lead to food shortages.	
40	0	1	38	28-Feb	Tunisia	Ra's Ajdir	Food Security	Refugees	UN	WFP			A WFP-chartere on Monday
41	4	1	39	28-Feb	Libya	Benghazi	Health	Residents	NGO	International medical corps		as well as the central medical supply warehouse for all of eastern Libya and a number of local pharmacies. Team idtenified a need for medicines and medical supplies as the supply chain from Tripol has been affected by the crisis. The team also heard reports of the need for surgeons.	Arrived in Bengl
42	1	1	40	28-Feb	Egypt	Salum	Health	Refugees	UN	WHO		However, the situation at the border is reported to be calm, with an occupancy rate of almost zero in the existing health facilities.	Health services Ministry of Health ambulances are on site to rec
	5	1										medical staff have left the country. Shortages of essential medicines and medical supplies including anaesthetic drugs, syringes, plasma expanders, blood transfusion bags and other items are being reported. However, most hospitals in Benghazi (the scene of some of the most intense fighting) are still operational, with more than 50 doctors providing specialized medical	However, most H fightnral are still

## PHASE I: 2015-2017

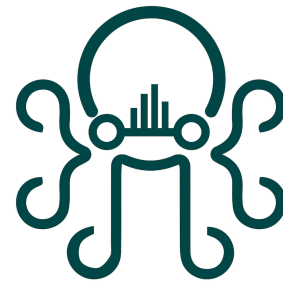
From manual analysis of unstructured secondary data (text) on excel to systematic information processing through tagging according to set frameworks

SIDR

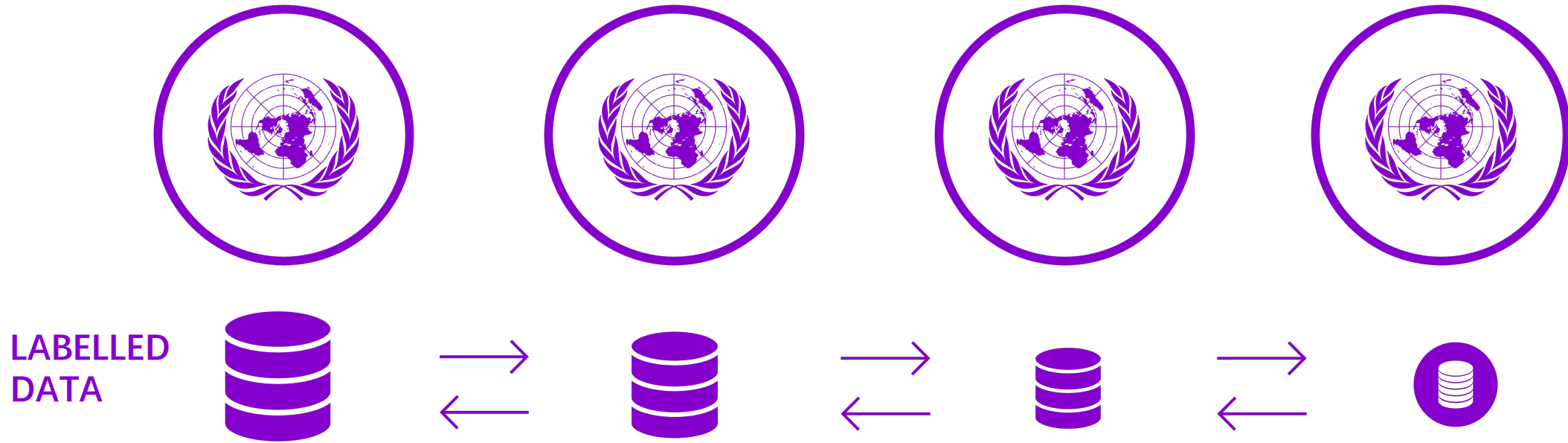


## PHASE II: 2017-2019

Improved user experience; introduction of assisted tagging processes as well as functionalities of collaborative information annotation; enhancement of assisted tagging through **Machine Learning (ML)** elements and in particular **Natural Language Processing (NLP)** functionalities



# TRANSFER LEARNING AND DATA SHARING– ACROSS INTERNATIONAL ORGANIZATIONS



HOW TO TRANSFER KNOWLEDGE ACROSS THEMATIC DOMAINS? HOW TO SHARE SEMANTICALLY SIMILAR CONCEPTS?

# Challenge Goal

The aim is to **learn novel text classification models, able to transfer knowledge across organizations, and specifically improve the classification effectiveness of the organizations with smaller amount of available training data.** Ideally, transfer and joint learning methods provide a robust solution for the lack of data in the data-sparse scenarios.

# Transfer Learning for International Crisis Response

Rohit Sharma





# Reverend Bayes and RoBERTa team up for DEEP

## Wang & Manning's G / D Variant<sup>1</sup>

For each document  $X_i$  with label  $y_i$ :

For each word  $w_{ij}$  in  $X_i$ :

count  $[w_{ij}][y_i] ++$

count  $[y_i] ++$

count  $++$

To classify a new  $X = w_1, \dots, w_n$ , pick  $y$  with top score:

$$\text{score}(y, w_1, \dots, w_n) = \log \frac{\text{count}[y]}{\text{count}} + \sum_{i=1}^n \log \frac{\text{count}[w_i][y] + mp}{\text{count}[y] + m}$$

## Extract more out of BERT<sup>2</sup>

1 - **Semi-supervised** training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

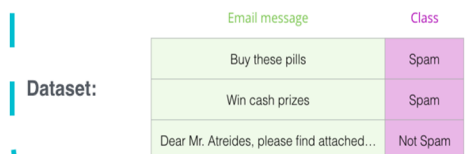
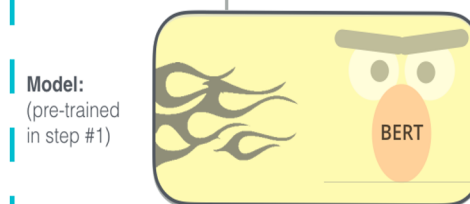
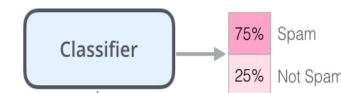
**Semi-supervised Learning Step**



**Objective:** Predict the masked word (language modeling)

2 - **Supervised** training on a specific task with a labeled dataset.

**Supervised Learning Step**



## References

1. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification: Wang and Manning.
2. RoBERTa: Robustly Optimized Pretraining Approach: Yinhan Liu, Myle Ott, Naman Goyal et al.
3. Jay Alamar's blog on the Illustrated BERT, ELMo & co.; Hailey Nguyen's blog on Naive Bayes





# **AMLD 2020 - Transfer Learning for International Crisis Response Challenge**

Rohit Midha

Shraddhaa Mohan

Sainath Prasanna



# Overview

1. Problem Statement
2. Competition Goal
3. Text Preprocessing
4. What didn't work?
5. Changes made
6. Mappings
7. Transformer Models
8. Ensembles



# Problem Statement

There is a lot of classified data available but different organizations want data to be classified differently.

For some organizations, especially the ones with their own custom frameworks, due to the lack of sufficiently tagged data, text classifiers show poor performance.

For these organizations, there is a cold-start challenge.



# Competition Goal

- Learn novel text classification models.
- Transfer knowledge across organizations.
- Improve the classification effectiveness of the organizations with smaller amount of available training data.
- **Understand the conceptual semantic linkages between the sectors of various organizations.**

# Text Preprocessing

1. Mapping unicode data
1. Mapping stylized letters
1. Mapping non english languages to a single character.

```
CUSTOM_TABLE = str.maketrans({"\xad": None, "\x7f": None,
"\ufeff": None, "\u200b": None, "\u200e": None,
"\u202a": None, "\u202c": None, "'": "'", "'": "'", "`": "`",
"“": "'", "”": "'", "«": "'", "»": "'", "G": "G", "I": "I",
"N": "N", "R": "R", "Y": "Y", "B": "B", "H": "H", "L": "L",
"F": "F", "A": "A", "C": "C", "D": "D", "E": "E", "J": "J",
"K": "K", "M": "M", "M": "M", "O": "O", "P": "P", "T": "T",
"U": "U", "W": "W", "V": "V", "K": "K", "B": "B", "M": "M",
"H": "H", "T": "T", "s": "S", "-": "-", "-": "-"})

NMS_TABLE = dict.fromkeys(
    i for i in range(sys.maxunicode + 1)
    if unicodedata.category(chr(i)) == "Mn"
)

HEBREW_TABLE = {i: "כ" for i in range(0x0590, 0x05FF)}
ARABIC_TABLE = {i: "ا" for i in range(0x0600, 0x06FF)}
CHINESE_TABLE = {i: "是" for i in range(0x4E00, 0x9FFF)}
KANJI_TABLE = {i: "ッ" for i in range(0x2E80, 0x2FD5)}
HIRAGANA_TABLE = {i: "ッ" for i in range(0x3041, 0x3096)}
KATAKANA_TABLE = {i: "ッ" for i in range(0x30A0, 0x30FF)}
```



# Additional Preprocessing Methods

1. Isolation of punctuations and certain other symbols. (eg. ?, /, °, <, ~, •, ≠, ™, ', ∪)
2. Deletion of various types of special symbols (eg. \n, \t, \r)
3. Contraction Mapping (eg. “doesn’t” : “does not”, “ain’t” : “is not”)
4. Removal of extra spaces.
5. Unicode Normalization





## What didn't work?

1. TF - IDF on `entry\_translated` + Logistic Regression gave a very poor mean of accuracies, around 53.2%.
2. Treating the problem as a multi-class classification problem where each text entry can map to only one of the 12 classes.
3. LSTMs were trained on both preprocess text and original text but the mean of accuracies was very poor.
4. Trying to perform Named Entity Recognition using Spacy to replace Names of Places with a "PLACE" tag.



## Changes Made

1. Concatenated the Org1, Org2 and Org3 training data.
2. For every row, we created 12 columns. Each of these 12 columns represented a class.
3. Treated the problem like a multi label problem.
4. 12 rows now represented the One Hot Encoded values.

# Data

## Before

	id	entry_original	language	entry_translated	labels
1	org1_12399	La emergencia humanitaria que en nombre de la ...	es	Humanitarian emergency on behalf of the Castro...	6;4

## After

:

	id	entry_original	language	entry_translated	labels	1	2	3	4	5	6	7	8	9	10	11	12
1	org1_12399	La emergencia humanitaria que en nombre de la ...	es	Humanitarian emergency on behalf of the Castro...	6;4	0	0	0	1	0	1	0	0	0	0	0	0

# Data Distribution: A Comparison

Org123

	category	number_of_comments
0	1	709
1	2	5823
2	3	1226
3	4	2461
4	5	4852
5	6	2213
6	7	816
7	8	622
8	9	900
9	10	5210
10	11	2502
11	12	2584

Org4

	category	number_of_comments
0	101	22
1	102	31
2	103	33
3	104	15
4	105	21
5	106	30
6	107	0
7	108	1
8	109	29
9	110	28
10	111	51
11	112	34

How do we add Org4 to this data?

# Mappings

## ORG4

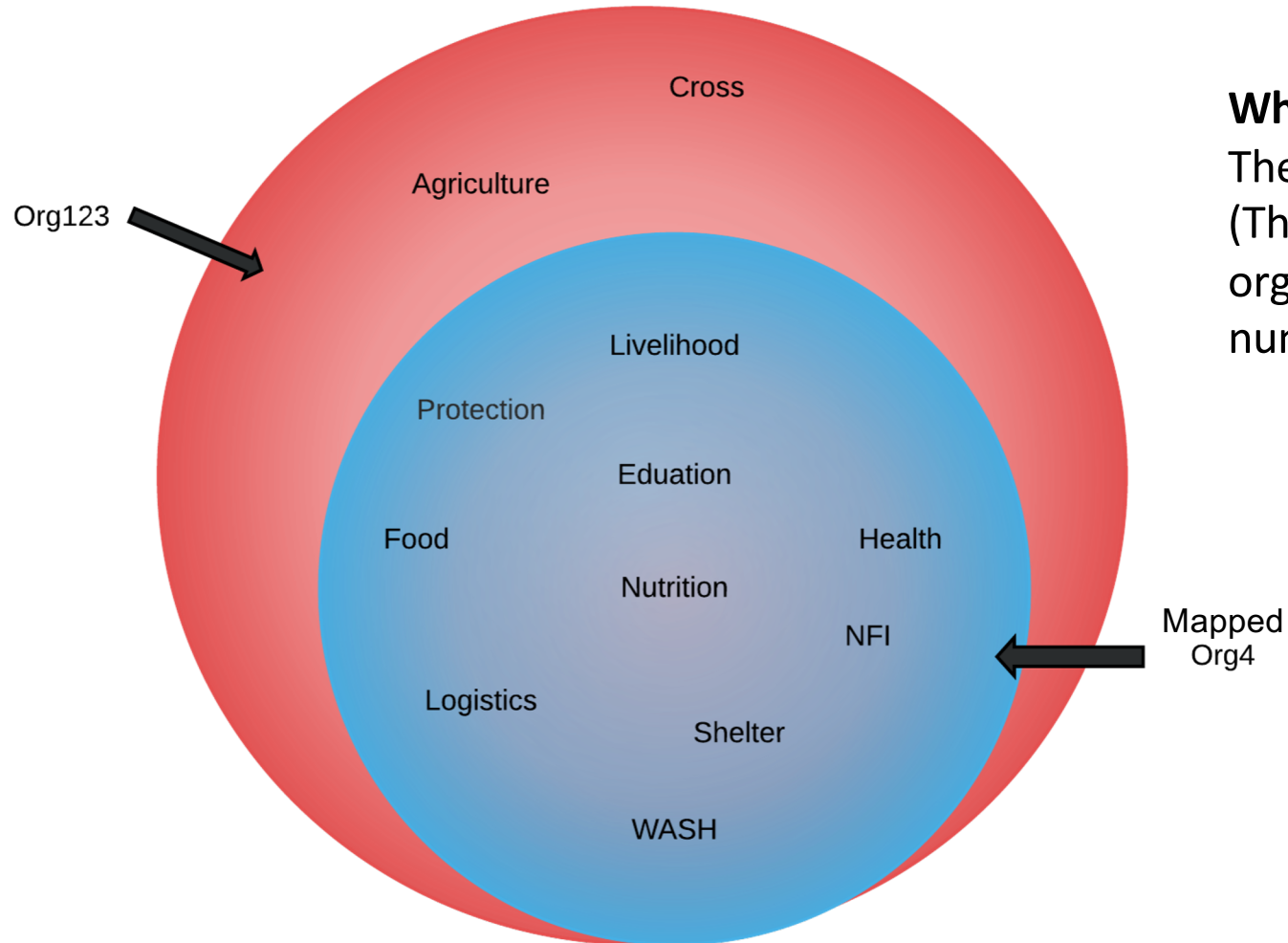
- (101) Child Protection
- (102) Early Recovery and Livelihoods
- (103) Education
- (104) Food
- (105) GBV: Gender Based Violence
- (106) Health
- (107) Logistics
- (108) Mine Action
- (109) Nutrition
- (110) Protection
- (111) Shelter and NFIs
- (112) WASH



## ORG123

- (10) Protection
- (6) Livelihood
- (3) Education
- (4) Food
- **X**
- (5) Health
- (7) Logistics
- **X**
- (9) Nutrition
- (10) Protection
- (11) Shelter + (8) NFI
- (12) WASH

# Intersection of classes




## What is the model trained on?

The Union of these two sets.

(This is extremely beneficial to org4 but also org123 as they also do get some extra number of samples.)

# Reverse Mapping (for org4)

When it comes to predicting on the test set for org4, the model outputs a probability array of size 12, giving the probability that a specific text belongs to a particular class [1-12]. Since it is required only to provide one label to an instance of text, we took the class corresponding to the one with maximum probability. After this is done, the predicted class is mapped onto the original org4 classes.

- (1) Agriculture
  - (2) Cross
  - (3) Education
  - (4) Food
  - (5) Health
  - (6) Livelihood
  - (7) Logistic
  - (8) Nfi
  - (9) Nutrition
  - (10) Protection
  - (11) Shelter
  - (12) WASH
- 
- X
  - X
  - (103) Education
  - (104) Food
  - (106) Health
  - (102) Early Recovery and Livelihood
  - (107) Logistic
  - (111) Shelter and NFIs
  - (109) Nutrition
  - (101) Protection
  - (111) Shelter and NFIs
  - (112) WASH

**How do we deal with the model predicting classes 1,2 for org4 ?**

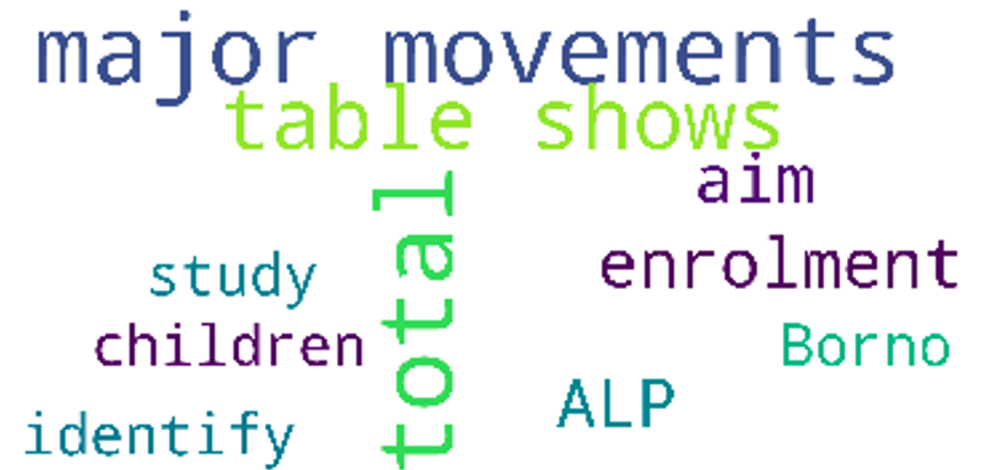
# Probabilities

Take the second most probable class predicted by the model.

# Heuristics

1. Analyse the data based on each specific label.
2. Find the prevalent words for each ID.
3. Search the text and if 2 or more of such prevalent words are present, change to that ID.

**Note : Not suggested until absolutely necessary.**



major movements  
table shows  
aim  
enrolment  
Borno  
ALP  
total  
study  
children  
identify

Prevalent words in text where ID : Education

**Why did this not work?**

Too many overlapping prevalent words.



# Transformer Models

---

<b>Model</b>	<b>Mean of Accuracies</b>	<b>Text Used</b>
bert-base-uncased	81.8	preprocessed text
bert-base-cased	81.4	preprocessed text
bert-base-multilingual-cased	80.1	entry_original
bert-large-cased	81.8	preprocessed text
roberta-base (original text)	81.9	entry_translated
roberta-base (preprocessed text)	81.2	preprocessed text
roberta-large	81.1	preprocessed text
distil-roberta	79.9	preprocessed text
xlm	79.8	preprocessed text

# Ensembles



## Additive Ensemble

1. Add all the probabilities for each class from each of the 9 final models row by row.
2. Take the class with the maximum value.

**Mean of Accuracies : 83.5**

## Voting Ensemble

1. For each of the 9 final models, for each row, take the class with maximum probability.
2. Across the 9 predictions for each row, take the class with maximum occurrences.

**Mean of Accuracies : 83.7**

# Example

Model	Text	Food	Shelter	NFI
Model 1	It is imperative to note that the top 3 immediate needs for HHs in Ran are: Food, NFIs and Shelter.	0.9	0.7	0.8
Model 2	It is imperative to note that the top 3 immediate needs for HHs in Ran are: Food, NFIs and Shelter.	0.8	0.7	0.85
Model 3	It is imperative to note that the top 3 immediate needs for HHs in Ran are: Food, NFIs and Shelter.	0.8	0.8	0.82

**Voting Ensemble :**

Model	Prediction
Model 1	Food
Model 2	NFI
Model 3	NFI

*Final : NFI*

**Additive Ensemble :**

Model	Food	Shelter	NFI
Model 1 + Model 2 + Model 3	2.5	2.2	2.47

*Final : Food*

**Final Model : Additive Ensemble for ORG1, ORG2, ORG3 and Voting Ensemble for ORG4**

**Mean of Accuracies : 84.1**



# Thank You

Questions? Contact us at:



Rohit Midha : [rohit.midha23@gmail.com](mailto:rohit.midha23@gmail.com)



Shraddhaa Mohan : [shraddhaa.mohan@gmail.com](mailto:shraddhaa.mohan@gmail.com)



Sainath Prasanna : [sainathprasanna@gmail.com](mailto:sainathprasanna@gmail.com)