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

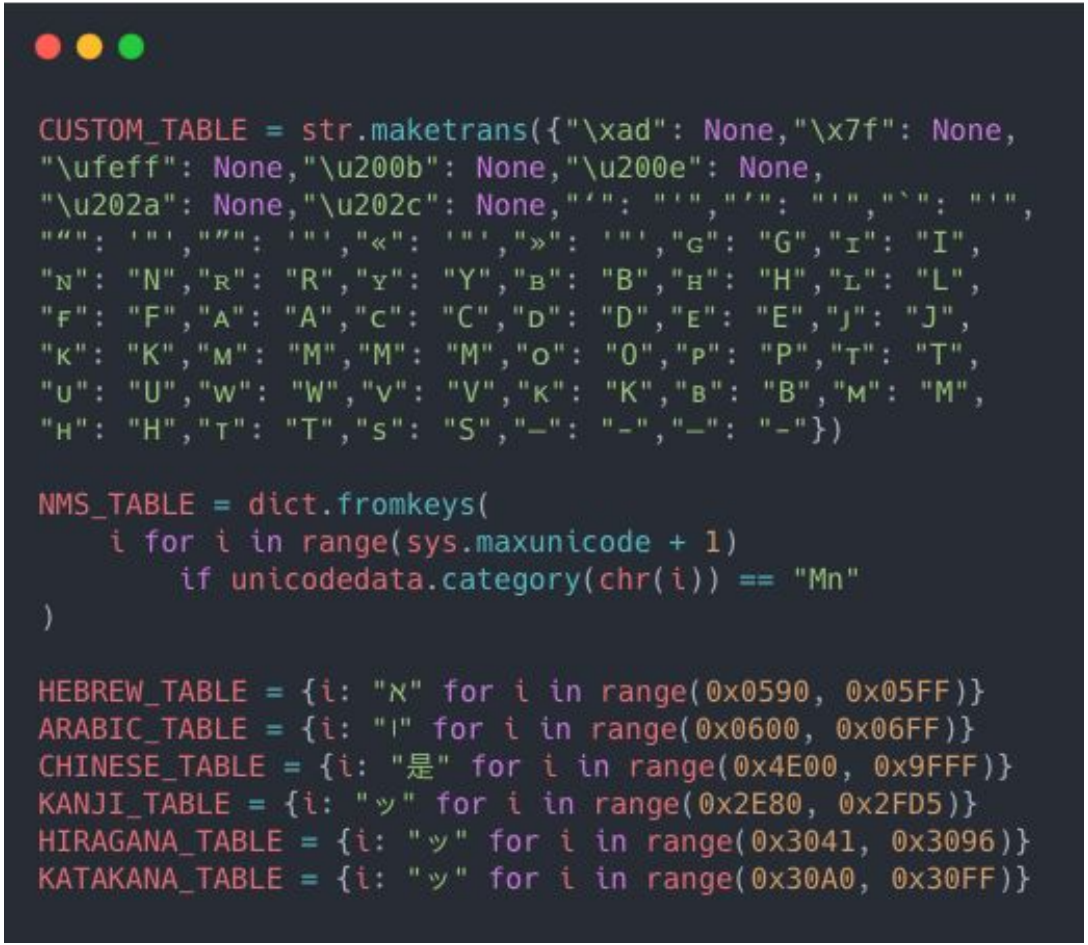
- Learn novel text classification models.
- Transfer knowledge across organizations.
- Improve the classification effectiveness of the organizations with smaller amount of available training data.
- **Understand the conceptual semantic linkages between the sectors of various organizations.**

Text Preprocessing

1. Mapping unicode data

2. Mapping stylized letters

3. 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,"‘": "'","’": "'","`": "'",
"“": '"',"”": '"',"«": '"',"»": '"',"ɢ": "G","ɪ": "I",
"ɴ": "N","ʀ": "R","ʏ": "Y","ʙ": "B","ʜ": "H","ʟ": "L",
"ꜰ": "F","ᴀ": "A","ᴄ": "C","ᴅ": "D","ᴇ": "E","ᴊ": "J",
"ᴋ": "K","ᴍ": "M","Μ": "M","ᴏ": "O","ᴘ": "P","ᴛ": "T",
"ᴜ": "U","ᴡ": "W","ᴠ": "V","ĸ": "K","в": "B","м": "M",
"н": "H","т": "T","ѕ": "S","—": "-","–": "-"})

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

HEBREW_TABLE = {i: "א" 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. ?, /, °, <, ~, •, ≠, ™, ', ʊ)
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.

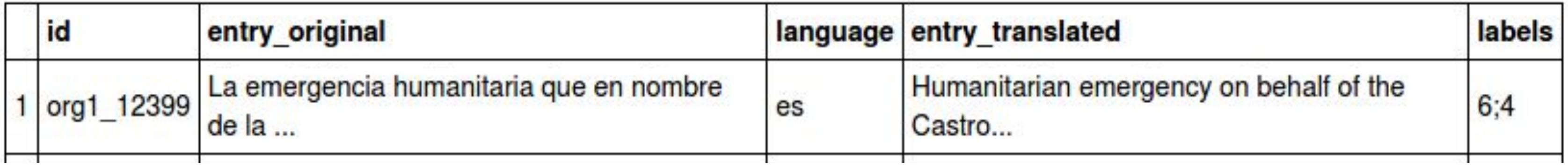
2. For every row, we created 12 columns. Each of these 12 columns represented a class.

3. Treated the problem like a multi label problem.

4. 12 rows now represented the One Hot Encoded values.

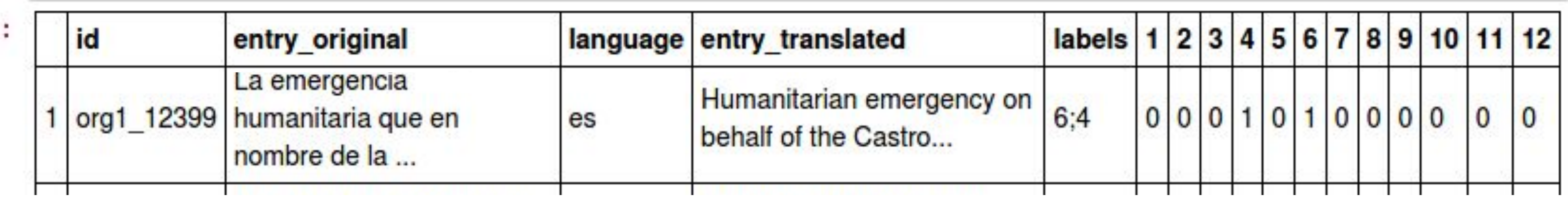
Data

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

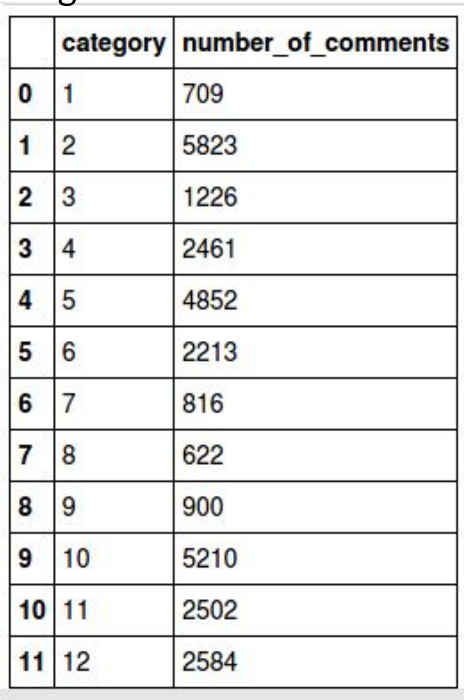
After



	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

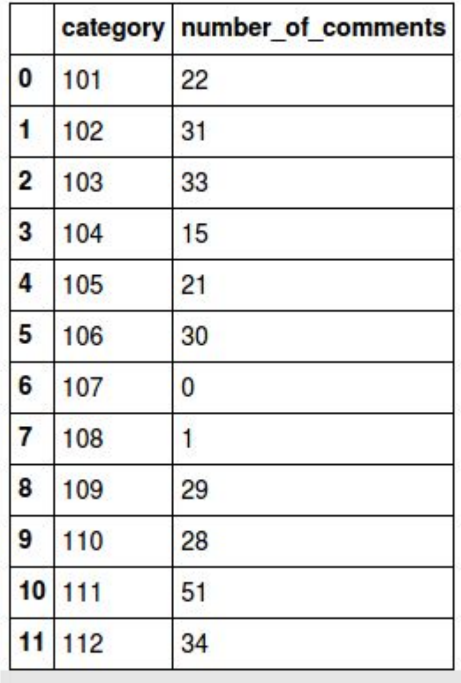
Data Distribution: A Comparison

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



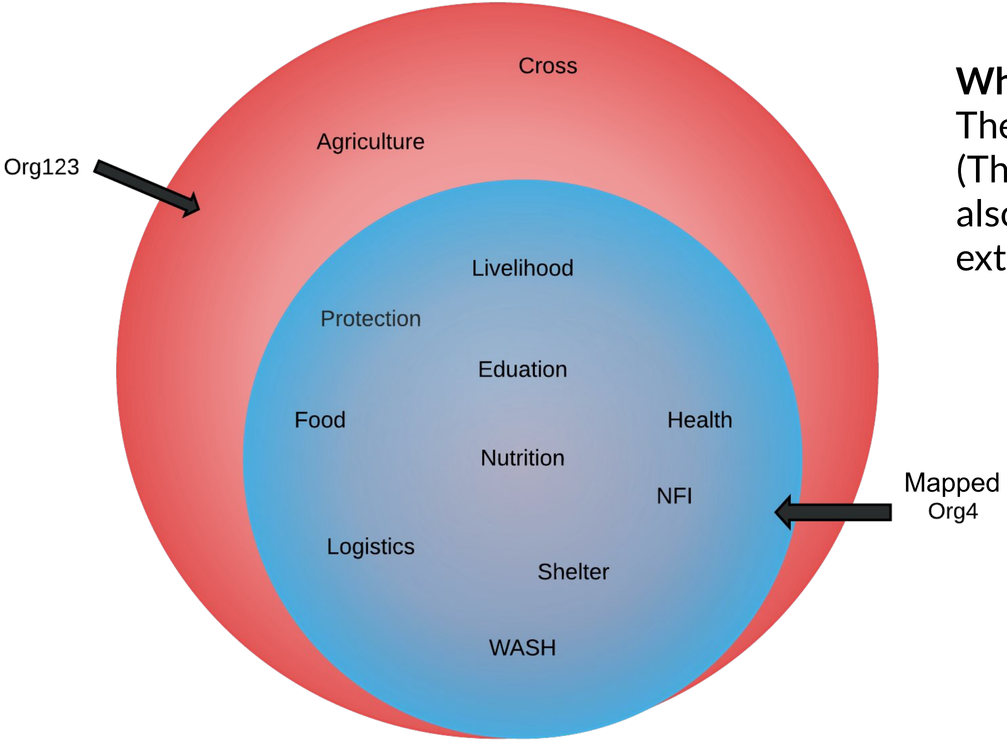
	category	number_of_comments
0	101	22
1	102	31
2	103	33
3	104	15
4	105	21
5	106	30
6	107	0
7	108	1
8	109	29
9	110	28
10	111	51
11	112	34

How do we add Org4 to this data?

Mappings

ORG4	ORG123
(101) Child Protection	(10) Protection
(102) Early Recovery and Livelihoods	(6) Livelihood
(103) Education	(3) Education
(104) Food	(4) Food
(105) GBV: Gender Based Violence	**X**
(106) Health	(5) Health
(107) Logistics	(7) Logistics
(108) Mine Action	**X**
(109) Nutrition	(9) Nutrition
(110) Protection	(10) Protection
(111) Shelter and NFIs	(11) Shelter + (8) NFI
(112) WASH	(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	X
(2) Cross	X
(3) Education	(103) Education
(4) Food	(104) Food
(5) Health	(106) Health
(6) Livelihood	(102) Early Recovery and Livelihood
(7) Logistic	(107) Logistic
(8) Nfi	(111) Shelter and NFIs
(9) Nutrition	(109) Nutrition
(10) Protection	(101) Protection
(11) Shelter	(111) Shelter and NFIs
(12) WASH	(112) WASH

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

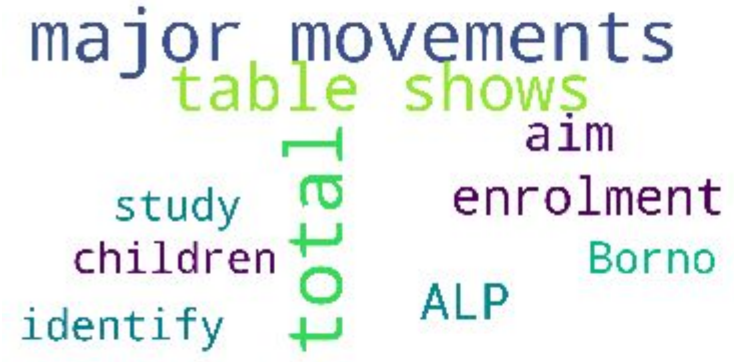
Probabilities

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.

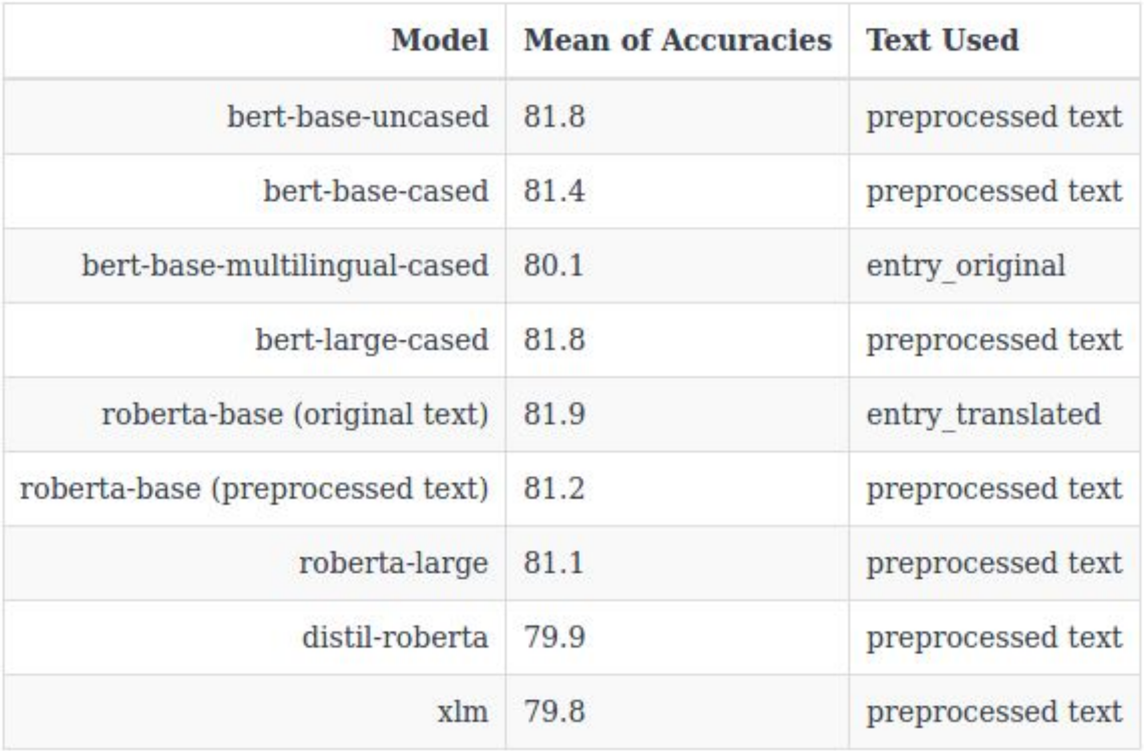


Prevalent words in text where ID : Education

Why did this not work?

Too many overlapping prevalent words.

Transformer Models



Model	Mean of Accuracies	Text Used
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.

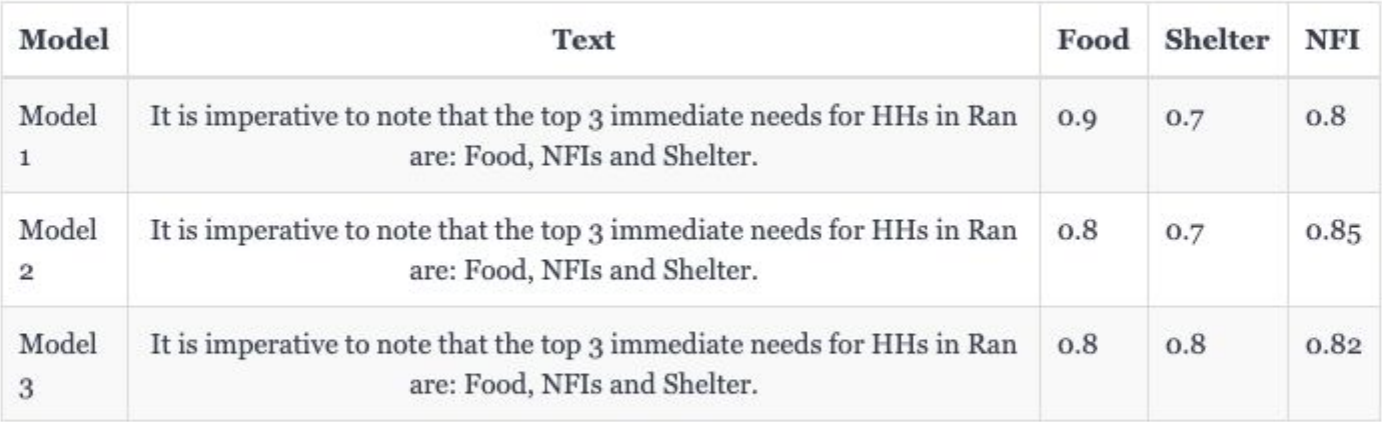
Mean of Accuracies : 83.5

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.

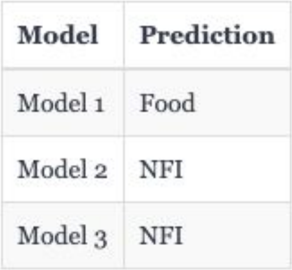
Mean of Accuracies : 83.7

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

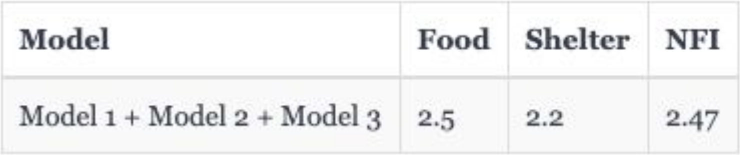
Voting Ensemble :



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

Shraddhaa Mohan : shraddhaa.mohan@gmail.com

Sainath Prasanna : sainathprasanna@gmail.com