Model card - CCI Cameroon

Model card

Model Details

Person or organisation developing model
Nesta in collaboration with a Data Science Fellow based in Nepal. As built for Nepal Red Cross support and guidance provided by them.

Model date
Models developed between Oct 2021 - July 2022

Model version
Only one version built / shared publicly

Model type

- Classification model
  - KNN classifier
  - Parameters: nearest neighbours = 5, p= 1, weights = distance
  - Above model and parameters were arrived at after testing different models and parameters using grid-search
  - Input: Word embeddings built from pre-trained transformer are created from comments which act as input to the model.
  - Features: Drawn from the word embeddings
- Clustering model
  - Igraph is generated using an adjacency matrix. The adjacency matrix contains 1 at position i,j where data point i is a nearest neighbour to data point j.
  - Igraph is fed into leiden community detection algorithm which generates clusters
  - Input: word embeddings built using pre-trained transformer model. FAISS index used to generate n-similar neighbours to each data point.

Paper or other resource for more information
Github project repository: https://github.com/nestauk/cci_cameroon

Intended Use

Primary intended uses

1. To provide volunteers with information on rumours they hear in the field
2. To help the Red Cross identify new incoming rumours

Primary intended users

1. Red Cross volunteers
2. Red Cross staff members

Out-of-scope use cases

- Cannot predict if an input is a rumour or not
- Would not be able to classify social media data
    - Embeddings formed french text would be clustered using the algorithm. Data in other forms (emojis, audio) would need other preprocessing steps/approaches to be handled.
- Can classify eight categories of COVID-19 related rumours (although the clustering step allows for new rumours to be categorised later)
- Community feedback not related to Covid-19. community feedback in a language other than French

Factors

Relevant factors

Classification model

- Trained on dataset from feedback collected over different mediums (eg forums, whatsapp, radio) during the period of March 2020 to July 2021. New data outside of this time period or by groups not represented in the dataset may have lower performing results.
- Input by different volunteers could affect the performance of the model. This is not something that we can measure as the dataset the model is trained on does not collect information about the volunteer.

Clustering model

- Terminologies used to express ideas can lead to different clustering outcomes.

Metrics

Model performance measures

Classification model
Main performance metric is the micro F1 score. The F1-score combines the precision and recall of a classifier into a single metric by taking their harmonic mean. The micro F1 score is used when you have a multi-label binary problem to assess the overall performance of all labels. It measures the F1-score of the aggregated contributions of all labels. The macro f1 score on the overhand, takes the mean off the computed f1 scores for each label. For this project we chose to go for the micro f1 score so the results won't be skewed by smaller labels.

Overall the f1 score was chosen for this project as a way to capture both the precision and recall in one metric. In addition to the F1 score we also looked at the individual confusion matrix plots for all labels.

Decision thresholds
Classification model
Probability threshold for predicting a class is 0.5. This is the default value for the sklearn classification model used. Some work was done to look at optimum thresholds for each label using the geometric mean and inspecting the ROC curves. However this was not implemented into the final code as, due to time restrictions of the project we were unable to complete the work to be confident of the results.

Evaluation Data

Details on the dataset(s) used for the quantitative analyses in the card.

Datasets
What datasets were used to evaluate the model?

Classification

- Cameroon rumours subset - IFRC dataset
    - This was the held out test used as the primary evaluation dataset for the model.
- DRC rumours - IFRC dataset
    - A dataset of rumours, beliefs and observations collected by the Red Cross from the Democratic Republic of the Congo. It was collected using the same fields / structure as the Cameroon dataset. This data was provided by the IFRC later on in the project.
- New rumours generated by CRC under each category
    - Generated via an activity designed as part of the project where Red Cross staff members were asked to generate new rumours under each of the eight codes.

Clustering

As an unsupervised approach there wasn't a dataset used as a form of evaluation. Instead the model was evaluated using manual inspections of the resulting clusters.

Motivation

Why were these datasets chosen?

- Cameroon rumours subset - IFRC dataset (test set)
  - The data is assigned to the same eight codes as the training set and is also collected by the Cameroon Red Cross.
- DRC rumours - IFRC dataset
  - This dataset was used as a way to test if the model could work on data collected from other French speaking countries.
- New rumours generated by CRC under each category
  - This was a small dataset produced by the Cameroon Red Cross staff as a way to test new rumour inputs.

Preprocessing

Pre-processing steps taken prior to input into both models:

Data cleaning

- Translation of non-french comments using the google translate python library
- Removal of duplicates, white spaces

Pre-processing

- Sentence transformer model applied to produce embeddings
- MultiLabelBinarizer function used to transform labels into binary values
  - This transformation was applied to the training set and then fit on the test set

Training Data

Datasets

What datasets were used to train the model?

Classification

Cameroon rumours subset - IFRC dataset: With the help of the International Federation of the Red Cross (IFRC), the CRC have been collecting feedback from communities on covid-19 including rumours, beliefs and observations. For each piece of feedback, a code is assigned that aims to group the feedback by theme e.g. 'beliefs about facemasks.' To train the model, a subset of this dataset was used which contains rumours, beliefs and observations from eight codes.

Clustering

Model developed using subsets of the IFRC dataset described above.

Motivation

Why was this dataset chosen?

At the start of the project different problem areas were highlighted as a focus where Collective Crisis Intelligence (CCI) could potentially help. One of the problem areas defined was handling rumours and miss-information. As the Red Cross already had a method of collecting feedback on covid-19 related rumours via this dataset, we wanted to build on this in this project and look to optimise the way volunteers can access information on rumours and how the Red Cross staff can better handle new incoming rumours.

Preprocessing

The same steps for pre-processing the data for model evaluation were applied to the training set.

Ethical Considerations

- Reported on the F1 score and accuracy across different sensitive characteristics that were captured in the IFRC dataset including - gender, age and location
- As the model is trained on feedback collected via Red Cross volunteers, we recognise that there is bias there in the information being recorded by them. As the dataset does not collect information on the volunteer, we are not able to capture the level of potential bias here.

Caveats and Recommendations

- The classification model has been tested on a subset of data from the Cameroon RC and an even smaller sample from the DRC Red Cross. We would not recommend using the model in other countries without first testing it's performance on labelled data
- The models are designed to be used with data collected via Red Cross volunteers. To handle different text formats e.g. data from social media, additional training data would need to be provided.
- The clustering model requires manual inspection of the clusters and should not be considered as a fully automated approach for detecting new rumour groups.