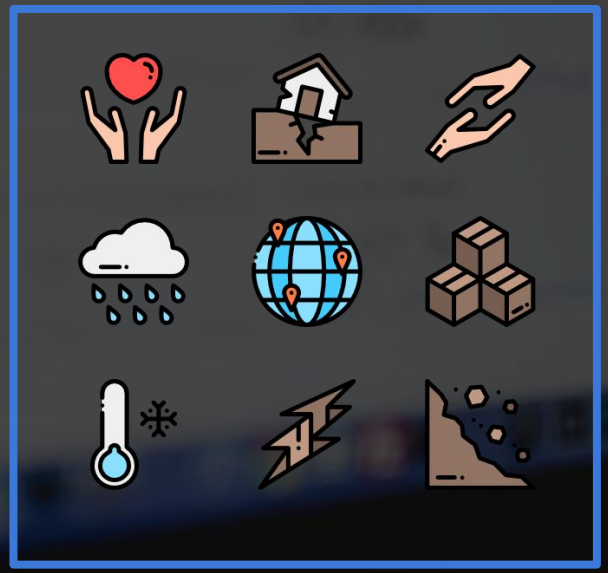
Speech & NLP

(TERM PROJECT)

Anjali Jha (16IM10032)
Shubham Mawa (16IM10033)
Bhargav D (16IM10034)
Prerit Jain (16IM10035)

PROBLEM STATEMENT



AUTOMATIC EXTRACTION OF EVENTS FROM NEWS DOCUMENTS.

(Events depicts the occurrence of any Disaster i.e natural or man-made)

PROBLEM STATEMENT

TASK 1

- Classification of Documents into predefined event types.
- The objective is to find whether the event has been discussed in the document.

TASK 2

- Detecting event trigger for each word vector
- The objective is to find whether an event is being associated with the word.

APPROACHES

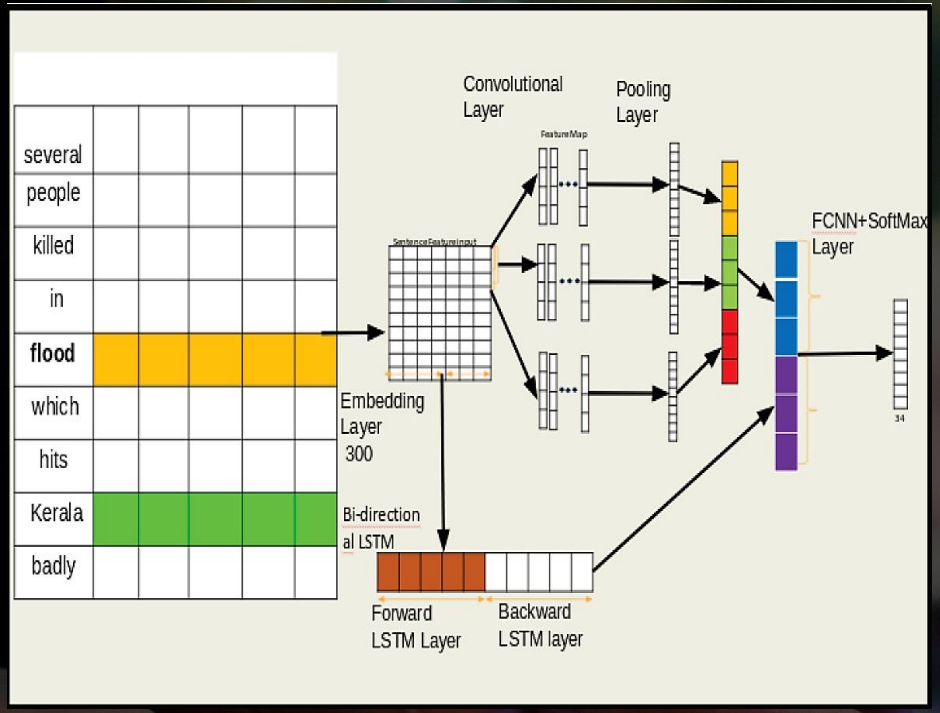
DENSE DOCUMENT EMBEDDINGS

- DATA EXTRACTION
  - Parsing the XML Document and converting it into txt format.
- DATA TRANSFORMATION
  - Creation of Dense Document embeddings using fastText.
  - Dimension Reduction.
- MACHINE LEARNING MODELS
  - Training multiple classifiers for different classes after splitting the dataset into Train and Test Sets.
  - Selection of appropriate classifier for the classification task.
- MODEL SELECTION & VALIDATION:
  - Based on the different model adequacy parameters, the best model is selected.
  - Hyper Parameter tuning.

APPROACHES

NEURAL NETWORK ARCHITECTURE

- DATA EXTRACTION:
  - Parsing the XML Document and converting it into sentences.
  - Creation of Vocabulary.
- PREPROCESSING STEPS:
  - Creation of Dense Word Embedding Matrix using fastText library for vocabulary.
  - Removing Punctuations in sentences.
- CREATION OF DATASET
  - Transformation of sentences into words with context. (Using appropriate window size).
  - Words are indexed by their position in the embedding matrix.
  - Corresponding event triggers for words are stored parallely.
  - The Event Triggers are numerically encoded.



CNN + Bi-LSTM ARCHITECTURE FOR EVENT TRIGGER CLASSIFICATION

NEURAL NETWORK ARCHITECTURE

- The two dimensional representation of each word is fed to a convolution layer followed by max-pooling layer.
- Parallel Bi-directional Long Short Term Memory(Bi-LSTM) for the same input
- The output of CNN and Bi-LSTM is concatenated.
- The representation vector is fed to a fully connected layer.
- Followed by a Softmax layer to get the proper event type of the current word.
- The gradients are calculated using back-propagation.
- Regularization is implemented by dropout.

DENSE DOCUMENT EMBEDDINGS

MODEL	Training Set Accuracy	Testing Set Accuracy
SVMs	95.31%	95.01%
Logistic Regression	95.16%	95.46%
Decision Tree	95.11%	95.61%

NEURAL NETWORK ARCHITECTURE

WORD	True Positive	False Positive	False Negative
Event Trigger	203	855	2348
NONE	103921	2285	792

CHALLENGES & LIMITATIONS

- Lack of Annotated Data in Hindi.
- Data Transformation was complex due to the structure of Data given.
  (XML Tree)
- Huge number of Parameters in the Neural Network
- Limited Vocabulary
- Multiclass Classification with insufficient examples for each class